December 2015

Study of Machine Learning Methods in Intelligent Transportation Systems

Vishal Jha
University of Nevada, Las Vegas

STUDY OF MACHINE LEARNING METHODS IN INTELLIGENT TRANSPORTATION SYSTEMS

By

Vishal Jha

Bachelor of Engineering (Hons.) - Mechanical Engineering
Birla Institute of Technology & Science
2005

A thesis submitted in partial fulfillment
of the requirements for the

Masters of Science in Engineering – Electrical Engineering

Department of Electrical and Computer Engineering
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
December 2015

UNLV | GRADUATE COLLEGE

Thesis Approval

The Graduate College
The University of Nevada, Las Vegas

November 12, 2015

This thesis prepared by

Vishal Jha

entitled

Study of Machine Learning Methods in Intelligent Transportation Systems

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Engineering – Electrical Engineering
Department of Electrical and Computer Engineering

Pushkin Kachroo, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

Emma Regentova, Ph.D.
Examination Committee Member

Ebrahim Saberinia, Ph.D.
Examination Committee Member

Haroon Stephen, Ph.D.
Graduate College Faculty Representative

ABSTRACT

Study of Machine Learning Methods in Intelligent Transportation Systems

by

Vishal Jha
Dr. Pushkin Kachroo, Examination Committee Chair
Lincy Professor of Electrical Engineering
University of Nevada, Las Vegas

Machine learning and data mining are currently hot topics of research and are applied in database, artificial intelligence, statistics, and so on to discover valuable knowledge and the patterns in big data available to users. Data mining is predominantly about processing unstructured data and extracting meaningful information from them for end users to help take business decisions. Machine learning techniques use mathematical algorithms to find a pattern or extract meaning out from big data. The popularity of such techniques in analyzing business problems has been enhanced by the arrival of big data.

The main objective of this thesis is to study the importance of big data and machine learning and their impact on transportation industry. This thesis is primarily a review of the important machine learning algorithms and their applications in the field of big data. The author has tried to showcase the need to extract meaningful information from the vast amount of big data in the form of traffic data available in today's world and also listed different machine learning techniques that can be used to extract this knowledge required in order to facilitate better decision making for transportation applications.

The analysis is done by using five different multivariate analysis and machine learning techniques in data mining namely cluster analysis, multivariate linear regression, hierarchical multiple regression, factor analysis and discriminant analysis in two different software packages namely SPSS and R. As part of the analysis, the author has tried to explain how knowledge extracted from random traffic data containing variables such as

age of the driver, sex of the driver, the day of the week, atmospheric condition and blood alcohol content of the driver can play an important role in predicting the traffic crash. The data taken into account is accident data, which was obtained from Fatality Analysis Reporting System (FARS) ranging from the year 1999 to 2009. It is concluded that traffic accidents were mostly impacted by the atmospheric conditions, blood alcohol content followed by the day of the week.

ACKNOWLEDGEMENTS

I want to thank my advisor Prof. Pushkin Kachroo for the constant, support and guidance throughout my time at the University of Nevada, Las Vegas (UNLV). He has been very patient with my shortcomings and mistakes and has always encouraged me. I have learnt a lot from him in electrical engineering and in my professional life. I think I would never have decided to stay in the program if I was not working with him. He never pressurized me or forced his decisions on me. I am immensely influenced by his dedication to the science and his focus to the research work. I would definitely try to incorporate his work ethics in my professional life if ever I become anything useful in my own life.

I would also like to thank the Transportation Research Corporation (TRC) for the kind of scientific atmosphere they provide to the graduate students. It is certainly an insight to the life of different researchers working here. It inspires students to have a vision and imagine and see the bigger picture of life.

I am also grateful to my thesis committee would like to thank Prof. Ebrahim Saberinia, Prof. Emma Regentova and Prof. Haroon Stephen for their constructive comments and encouragements.

I am indebted to my friends Sumant Jha and Shaurya Agarwal who have been of immense help in editing the thesis and making my work more presentable.

Lastly, I would thank my parents and my sister for offering constant support and back up in good and bad times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In Nevada, the population distribution and fatal crashes has changed over the years. According to the Fatality Analysis Reporting System (FARS), 290 traffic fatalities occurred on Nevada roads in 2014, a 9% year-on-year increase or an increase of 24 deaths compared to the previous year. However, Nevada has made progress in relation to number of fatalities per 100M Vehicle Miles Traveled in NV (VMT) each year. In 2012, that rate was 1.08, dropping to 1.06 in 2013 (2014 VMT not yet available) [1].

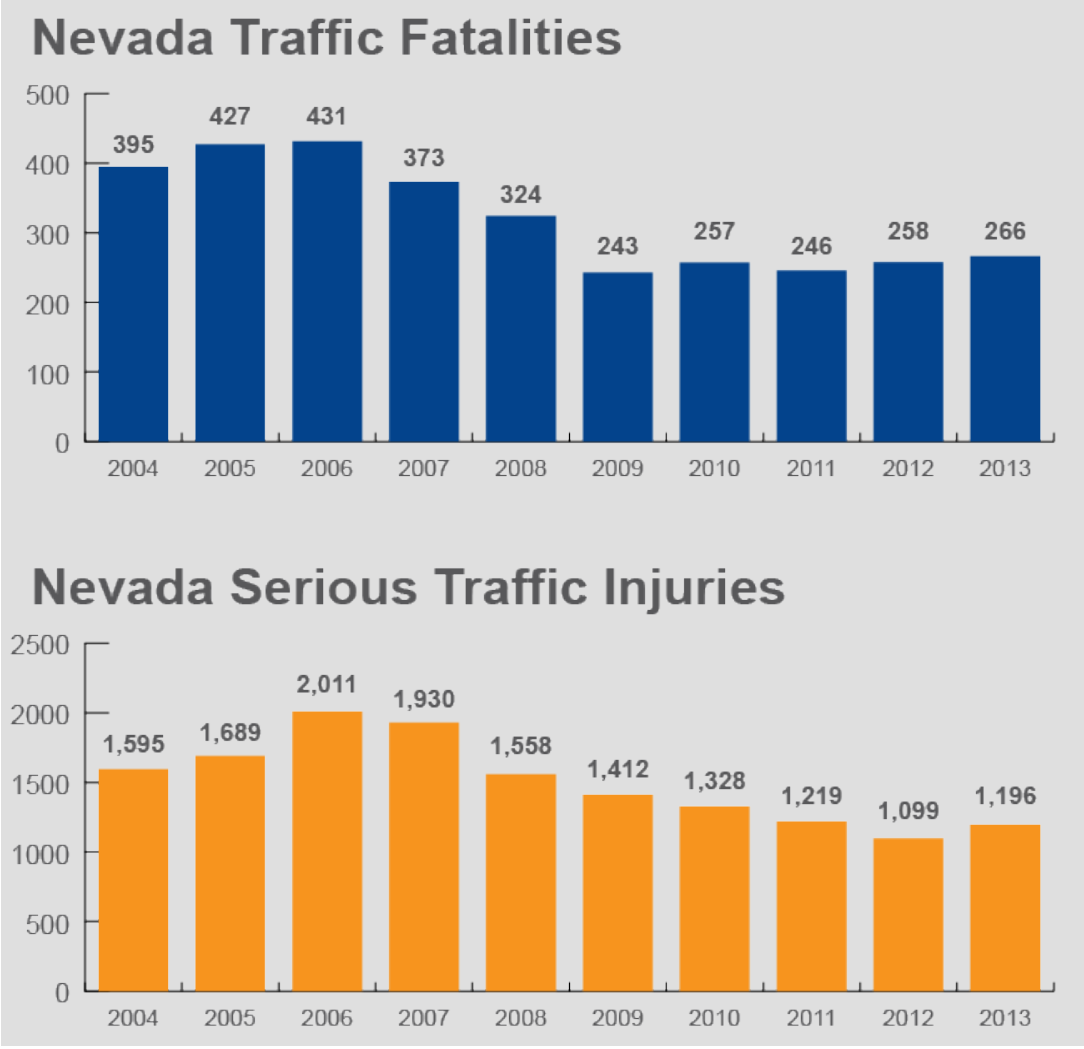

Figure 1.1: Nevada traffic fatalities and Nevada serious traffic injuries, 2004-2013 [1]

In 2011, Nevada launched Zero Fatalities program, aimed towards reducing deaths due to crashes to an absolute zero. Such a goal requires analysis at a microscopic level to reduce inconsistencies inside the system, along with visualization for an individual to understand how their decision catapults into usage of tax dollars. Thus a thrust in this direction is required and can be achieved by implementing an Intelligent Transportation System (ITS).

ITS encompasses a broad range of wireless and wire line communications based information and electronics technologies and is one of the most promising methods in dealing with the traffic problems. While some key information is readily obtained using traditional query operations of traffic information, the analysis and management of traffic information remains one of the key issues and deeper information are still difficult to discover. This deeper level information usually contains large volume of data known as big data and our inability to harness important information from them brings us to an important technique used universally, named as the data mining.

Data mining and machine intelligence are currently a hot topic research area and are applied in database, artificial intelligence, statistics and so on to discover valuable knowledge and the patterns in big data available to users. Data mining is predominantly about processing unstructured data and extracting meaningful information from them for end users to help take business decisions. Data mining techniques use mathematical algorithms and machine intelligence techniques. The popularity of such techniques in analyzing business problems has been enhanced by the arrival of big data [2]. In this thesis, data mining tools such as SPSS and R has been used to arrive at predictive models by analyzing past and present traffic data.

Data mining has become one of the most important tools for extracting and handling data and for establishing patterns to produce useful information for decision-making. Lately, there have

been a lot of breakthroughs in data collection technology, such as bar-code scanners in commercial domains and sensors in scientific and industrial sectors, which has led to the generation of huge amounts of data. This remarkable growth in big data and databases has produced a huge demand for new techniques and tools that can transform big data into useful information.

1.1 PROBLEM STATEMENT

The most important problem to address in analyzing big data is the choice of machine-learning technique to be used. It mainly depends on the jurisdiction of the analyst or the researcher. Therefore, the author has come up with five different techniques namely factor analysis, principal component analysis, discriminant analysis, cluster analysis and hierarchical clustering to analyze the given traffic data set. The goal is to help the analysts in understanding the different machine-learning techniques in order to find the best-fit solution for their particular problems.

1.2 CONTRIBUTION OF THE THESIS

This thesis gives a direction as to which are the different machine-learning techniques one should use in order to solve predictive big data problems. This thesis also helps to understand some of the data preprocessing tools that should be incorporated to a given data set to get an insight into the type and nature of data set being used. This thesis uses a unique data set to assess the performances of these five different predictive machine-learning techniques.

1.3 STRUCTURE OF THE THESIS

Chapter one is the introduction of the thesis. It deals with the meaning of data mining and some areas where this tool is used or needed. It also defines the problem statement and the contributions of this thesis.

Chapter two includes an overview of big data and its impact on the transportation industry. It also includes a case study on how big data analysis helped improve Dublin Transit System.

Chapter three comprises of a review on data mining, its major predictive techniques, applications and survey of the comparative analysis by other researchers and the criteria to be used for model comparison in this work. It also describes the methodology employed in this thesis and an introduction of the data sets used in the analysis.

Chapter four describes machine learning and its similarities with data mining.

Chapter five consists of the methodology and analysis of different machine learning techniques used to analyze big data.

Chapter six provides an analysis of the FARS dataset using statistical computing software called R. R is a software environment and programming language used for statistical computing and graphics. The R language is extensively used among mathematicians and data miners for developing statistical software and data analysis.

Chapter seven lists out the results of all the different machine learning techniques used in the course of this thesis and also compares the results.

Chapter eight is the last chapter in the thesis and it draws all the conclusions the author deduced from the analysis of the of the FARS data. This chapter also includes some possible areas for further research.

CHAPTER 2: BIG DATA

Big data is a catchphrase that describes large amount of data, which is both structured and unstructured in nature. These huge volumes of data are difficult to process using traditional databases and software systems. Big data is constantly being produced by numerous sources like software systems, traffic sensors, mobile devices and social media, at a very high volume, velocity and variety. An example of big data can be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data that contains billions and trillions of records of millions of people from different sources such as web, sales, customer contact center, social media, mobile data and so on. The data is normally unstructured or roughly structured data and is often unfinished and inaccessible. One needs to incorporate statistical analysis and processing tools to extract meaningful information from big data [3].

While the term big data may appear to reference the volume of information, that isn't generally the case. Big data, particularly when utilized by vendors, may allude to the innovation (which incorporates devices and procedures) that an association requires in managing large amounts of information and data. The term big data is ought to have come into being with web search organizations that wanted to query huge accumulation of unstructured data.

Big data is transforming the way employees work within an enterprise. It is creating an atmosphere within which organizations and software professionals must work hand-in-hand to extract value from all sorts of data. Big data insights can help professionals in better decision making and taking advantage of newer revenue sources.

2.1 ADVANTAGES OF BIG DATA

The three major advantages of big data are:

- Competitive advantage: Big data is evolving as the newest source for competitive advantage in the present world scenario.
- Decision making: Big data is enabling more mid-level and lower-level employees to participate in the decision making process of the organization.
- Value of data: As the data becomes more valuable, there is a need for more sophisticated systems to extract meaningful information.

2.2 CHARACTERISTICS OF BIG DATA

Big Data can be characterized into six different categories [4]:

- Volume: One of the major characteristics of big data is volume. The name big data itself suggests large volumes of data. The size of the data determines the potential and importance of the data, which is under consideration and whether it can be classified as big data.
- Variety: This is one important characteristic of big data that the data analysts must know in order to analyze and reproduce the data in a manner that is understandable and advantageous to them.
- Velocity: In context of big data, velocity is defined as the speed at which the data is generated and then process to help fuel organizational growth and development.
- Variability: Variability refers to the discrepancies present in the data that has been generated. This characteristic is a major roadblock in the effective management of big data.
- Veracity: This characteristic describes the accuracy of the collected data. The greater the veracity the more accurate is the final analysis of the data.

- Complexity: Big data can be very complex in nature as large volumes of data are collected from multiple resources. The collected data must be linked to extract information that is meaningful to the user or the organization.

2.3 BIG DATA IN TRANSPORTATION

Big data offers innovative ways for gathering information regarding transport infrastructure from passenger and automobile movements. For example, certain GPS systems empower users to inform others in case of traffic incidents. This information in turn is conveyed to network operators in real-time that permit immediate responses to disruptions.

2.4 HOW BIG DATA CAN CREATE A SMARTER TRANSPORTATION INDUSTRY

- **By upgrading cargo developments and directing**

Merging shipments and improving cargo development for extensive logistics can empower same day provincial conveyance. Knowing precisely which items are placed in which distribution centers can help organizations like Amazon to convey the right item at the opportune time to the right client within 24 hours. Uprooting store network waste and investigating exchange level item will guarantee proficient and more brilliant transportation of cargo. [5]

Utilizing satellite navigation and sensors, trucks, ships or planes can be followed in real time. The paths diverse trucks, ships or planes need to take can be streamlined utilizing a great deal of open information, for example, street conditions, car influxes, climate conditions, conveyance addresses, area of corner stores (in the instances of trucks) and so forth. At whatever time a change in location is reported from head office, it can be pushed to the driver or commander in real time. The framework consequently computes and advances the perfect and least expensive new path to the new destination.

Sensors in trucks, ships or planes can likewise give continuous data about how the truck, ship or plane is performing, how quick it is going, to what extent it is on the go, to what extent it is stopping and so forth. With this information, consolidated with sensors that screen the strength of the motor and hardware, lapses can be anticipated and support can be arranged without losing much time. It is even conceivable to consequently book support at the area that requires the slightest downtime for the transportation organization, while the specialist in a flash comprehends what the issue is and how it can be fathomed.

Huge logistic associations can have hundreds of trucks. In the event that their utilization is not streamlined, an organization can lose a considerable measure of cash. With sensor information it gets to be known where all trucks are at any minute in time, what their stock is and their destination. This data can help the transport organizations to enhance their taskforce and expand proficiency.

- **By determining inventory on hand**

Stock that is in-travel is still piece of the stock of an association, despite the fact that it physically left a distribution center. It is vital to know the definite stock at all times, particularly if changes need to be made in the final hour. At the point when all items contain sensors they can be followed in real-time and changes and/or stock tallying turns out to be extremely basic.

Stock administration investigation can be utilized to bring together a stage that offers associations an itemized diagram of flight and landing times, request cuts and also the likelihood to furnish clients with point by point data on their cargo.

- **By improving the end-to-end customer experience**

Clients need to know precisely where in the process their cargo is and where it is situated and in addition when the expected delivery time is. With a more brilliant transportation framework,

cargo shippers and clients are given the data and instruments to choose for themselves. The most ideal approach is to get their item from origin to destination, across diverse methods of transport, considering cost, time and convenience. A package can utilize numerous methods of transport and within a savvy transportation framework; clients can decide how their cargo goes from point A to B. This will empower the client to better deal with their store network plus costs.

- **By reducing environmental impact and increasing safety**

Fuel utilization can be lessened through a number of ways. Above all else, sensors can screen the motor and upgrade fuel information in light of the need of the motor and what the truck, ship or plane is doing. With the best enhanced routing, made possible by considering climate conditions, driving conduct, street conditions, area and so forth, a ton of fuel can be spared.

Sensors can likewise screen how quick the driver is driving, what is his position and whether the driver is adhering to the traffic rules. It can be observed if the driver is in the driver's seat too long or if his breaks are too long. It can keep the driver alert and reduce traffic accidents, while keeping the drive responsible.

More urban communities around the globe are trying different things with brilliant transportation frameworks that will diminish contamination and improve traffic safety. The city of Brisbane has built up a complete, continuous review of the city's vehicle system, which gives a stage to create and test new methodologies in a steady and ongoing virtual environment. This stage empowers the city to foresee and enhance traffic bottlenecks, bringing about more satisfied commuters and shippers, while diminishing discharges of poisonous gases. With the expanding demands of clients to have their cargo delivered as quick as possible and at a cheaper cost, transportation organizations confront a challenge that fortunately can be solved with big data.

2.5 CASE STUDY: BIG DATA HELPED REDUCE CONGESTION IN DUBLIN PUBLIC TRANSIT SYSTEM

Background

The Dublin Public Transit System (DPTS) project began in the year 2010 for 3+ years and was valued at €66 million which was jointly funded by IBM and Industrial Development Agency of Ireland [6].

Issue

Traffic bottleneck in public transport network mainly buses, all over the city of Dublin.

Objectives

- Minimize congestion and improve flow of traffic
- Enhanced mobility for commuters

Remedies

- Innovative analysis on data collected from each bus's trip: In collaboration with IBM researchers, DBTS gathered Big Data from multiple sources such as bus timetables, inductive-loop traffic detectors, closed-circuit television cameras and GPS updates.
- Better reporting and monitoring: Big Data collected was combined to build a digital map of the Dublin city covered with the real-time positions of Dublin's buses using stream computing and geospatial data.

Benefits

- Dublin City council releases and updates the journey information for each bus every minute, which empowers commuters to find out the quickest route to their destination online.

- Dublin city riding on the back of the improvement in reporting, can now identify optimal solutions to reduce traffic bottlenecks which in turn reduces congestion and also the best place(s) to add additional bus lanes and bus-only traffic systems.

CHAPTER 3: DATA MINING

Data mining leads us to identifying new, useful, and understandable correlations and patterns in existing data [7]. Different communities have different names for the same process of finding useful information (data) (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing) [8]. Statisticians, database researchers, and the MIS and Business communities started using the term "data mining" initially for extracting useful information from big data. One of the processes involving data mining is known as Knowledge Discovery in Databases (KDD), which is used for discovering useful knowledge from data. KDD involves data preparation, selection, cleaning and proper interpretation of results of the data mining process to ensure useful information is gleaned from the data. Data mining differs from traditional data analysis and statistical approaches in that it uses analytical techniques from several disciplines, for e.g. numerical analysis, pattern matching and areas of artificial intelligence such as machine learning, and neural networks and genetic algorithms [8][9][10].

Several data mining tasks utilize a traditional, hypothesis-driven data analysis approach. It is also very commonplace to use an opportunistic, data driven approach which aids pattern detection algorithm to find trends, patters and relationships which can then be used in the decision making process. These two data mining approaches differ in the outcome – a model or a pattern. The model approach is similar to the conventional exploratory statistical methods, except for the problems inherent from the large size of data sets. The main motive here is to summarize a set of data for identification and description of prominent features of distribution [11]. One of such models includes cluster analysis partition of a set of data, prediction using regression model, and tree-based classification rule. While creating a model, sometimes empirical and mechanistic models are treated

differentially [12]. Empirical (also called operational) models seek to establish relationships without any bias from any underlying theory. Mechanistic models (also called substantive or phenomenological models) are based on theory/mechanism for data generating process. Hence, data mining, by definition, is primarily concerned with operational.

Yet another type of data mining approach, pattern detection, seeks to identify small (and probably, important) departures from the norm, which helps in detecting unusual behavioral pattern, e.g. fraud detection for credit cards by monitoring unusual spending patterns, sporadic waveforms in EEG traces, etc. The notion that data mining seeks "nuggets" of information from a huge data repository, derived from this class of method. However, with the business database, it is not so easy to extract patterns, because of the complexity of data arising from anomalies as discontinuity, noise, ambiguity and incompleteness [13]. The predictive power of such mining algorithms might decrease with increase in number of anomalies [14].

One of the most important pre-processing steps in data mining is the construction of a data warehouse that involves data cleaning and integration. However, it is an optional step in most data mining process, as in case of larger data warehouse containing data from multiple sources – it becomes enormous task, taking huge amount of time running into years and costing millions of dollars [15]. Alternatives to data warehouses are available operational or transactional database, or data marts, which can be either logical or a physical subset of data warehouse.

Amalgam of artificial intelligence and statistics-related technique leads to KDD systems, which aids in finding associations, sequences, classifications, clusters and forecasts. Operational warehouse acts as the entry point for most of the data, which is then "cleaned" and moved into

warehouse. After a certain amount of time, these data are either purged, or summarized (along with other information) or archived.

There are three typical components in data warehouse architecture:

- The back end (data acquisition software), which is needed to extract data from legacy systems and external sources, and loading them into the warehouse after consolidating and summarizing the data.
- The data and associated database software, often referred to as "target database".
- The front-end (client) software, which enables users and applications to access and analyze data in the warehouse.

3.1 DATA MINING IN PERSPECTIVE

The term data mining is used to refer a specific set of activity, which involve extracting meaningful new information from data. It is not a new term to statisticians and is synonymous with data dredging or data snooping in hope of being able to find patterns. This occurs when a dataset is used more than once for inference or model selection [16]. As we can see, this "snooping" is a derogatory term for an exhaustive search can throw up a pattern of some kind. While many of these pattern can be a product of random fluctuation and do not represent any underlying issue, which conflicts with the purpose of data analysis – to model the underlying structure giving rise to consistent and replicable patterns.

Hence, data mining is a tool that helps organizations to focus on most important information available in their existing database. This does not eliminate the need to know the business, the available data or the underlying analytical methods in use. Here, it must be noted, that the predictive method resulting from data mining may not be the cause of an action/behavior. Such

causal inferences are subject to several error sources like latent variability, sample selection bias and model equivalence of data population, or population drift [17]. Furthermore, it should also be noted, that relationships gleaned from data mining process, does not indicate the value of pattern to an organization. Such a pattern should must be verified and validated in an appropriate context.

Several industries benefit from using data mining and have witnessed increased profits by reducing costs and raising revenues. There are several ways to do this, a few of which are as follows:

- Helping in reducing costs during the beginning of product life cycle in the research and development phase.
- Automated manufacturing processes use bounds for statistical processes, which can be derived from data mining.
- Reducing mailing costs by avoiding mailing to customers who do not respond to offers.
- Facilitating one-to-one marketing and mass customization opportunities in customer relationship management, etc.

Several organizations use data mining to acquire new customers, increase revenue from existing customers, and retaining good customers, which is almost the whole customer life cycle. Using customer profiling, these companies can target prospective customers with similar traits and also focus attention on those customers who have not bought similar products (cross-selling).

While this sort of profiling enables a company to study the consumer behavior and attract new customers, it also enables them to retain customers who are considered at risk of leaving (called reducing churn or attrition). It is usually far less expensive to retain a customer than acquire a new one [18].

Data mining is such a ubiquitous subject that it can make contribution in almost every stream of business. A few examples, where data mining can make contribution are:

- Telecommunication and credit card companies - these are two of the leaders in application of data mining to detect fraud.
- Insurance companies are yet another industry which uses data mining to reduce fraud.
- Medical industry uses data mining to predict effectiveness of surgical procedures, medical tests, and medications.
- Financial firms use data mining to determine market and industry characteristics as well as to predict individual company and stock performance.
- Retailers can make decision about products, which are popular, and items to stock in particular stores (even when and where to place them) and also access the effectiveness of promotions and coupons.
- Several Pharmaceutical firms mine large databases for chemical compounds and genetic materials to discover substances, which might be a potential candidate as new agents for treatment of a disease.

3.2 DATA MINING AND STATISTICS

Statistics and Data Mining both have a common goal to find some sort of correlation and structure in data. The goals are so similar that some people regard data mining as a subset of statistic, which is not a realistic assessment. Data Mining makes use of data but also, ideas, tools, and methods from other areas - particularly database technology and machine learning and focuses on subject other than those that statisticians fancy. That being said, statistic does play an important and

major role in data mining, most importantly in the process of development and assessment of models.

The learning algorithm uses statistical tests when constructing rules or trees and also for correcting models that are over fitted. These tests are also used in validating machine learning models and evaluating machine-learning algorithms. Commonly used statistical analysis techniques are discussed below. The reader is referred to Johnson and Wicheren, 1998 [19] for an extensive review of classical statistical algorithm. Descriptive and visualization techniques include simple descriptive statistics such as:

- Averages and measures of variation,
- Counts and percentages, and
- Cross-tabs and simple correlations

These descriptive and visualization techniques are useful for understanding data structure, which are primarily a discovery technique and useful for interpreting large amount of data and understanding the underlying structure. The tools used for this method include histograms, box plots, scatter diagrams, and multidimensional surface plots [20] [Tegarden, 1999].

Similarly, yet another method called Cluster Analysis, seeks to organize information about variables, which can then be used to identify relatively homogenous groups, or "clusters". These clusters should be highly internally homogenous (members similar to one another) and externally heterogeneous (members not like members of other clusters).

Correlation Analysis statistically measures the relationship between two variables that yields correlation output. This output shows the effect of change of one variable to another. The ultimate goal of correlation between two variables is to see if a change in independent variable will cause a

change in dependent variable. This information is helpful in analyzing the predictive ability of independent variable. The findings obtained from correlation analysis are useful in analyzing the causal relationships, but do not establish causal patterns.

Discriminant Analysis can be used to predict membership in two or more mutually exclusive groups from a set of predictors. This method is used when there is no natural ordering on the groups. This method can also be seen as the inverse of a one-way multivariate analysis of variance (MANOVA) in that the levels of the independent variable (or factor) for MANOVA become the categories of the dependent variable for discriminant analysis, and the dependent variables of the MANOVA become the predictors for discriminant analysis

Factor Analysis is a method that comes handy in understanding the underlying reasons for correlation among a group of variables. Factor analysis technique is mainly used to reduce the number of variables and to detect the structure in the relationships among variables. Factor analysis can be applied as a data reduction tool as well as for detecting structures. The goal, in exploratory factor analysis, is to explore or search for a factor structure. On the other hand, confirmatory factor analysis has the objective of empirically verifying or confirming that the assumed factor structure is correct, given the assumption that factor structure is known apriori.

Regression Analysis is another statistical tool, which makes use of relation between two or more quantitative variable so that the dependent variable can be predicted from the independent variable. In this method, no cause-effect pattern is implied. There are several types of regression analysis including simple linear, multiple linear, curvilinear, multiple curvilinear as well as logistic regression models.

Logistic Regression, as mentioned above, comes handy when the response variable is a binary or qualitative outcome. Linear and Logistic regression, though both find a "best fitting" equation, are different in their underlying principle. It uses a maximum likelihood method, which maximizes the probability of obtaining the observed results given fitted regression coefficient. Logistic regression is more robust as it does not make any assumption about distribution of independent variables.

3.3 DATA MINING AND TRANSPORTATION

With millions of vehicles on roads every day, identification of critical data through data

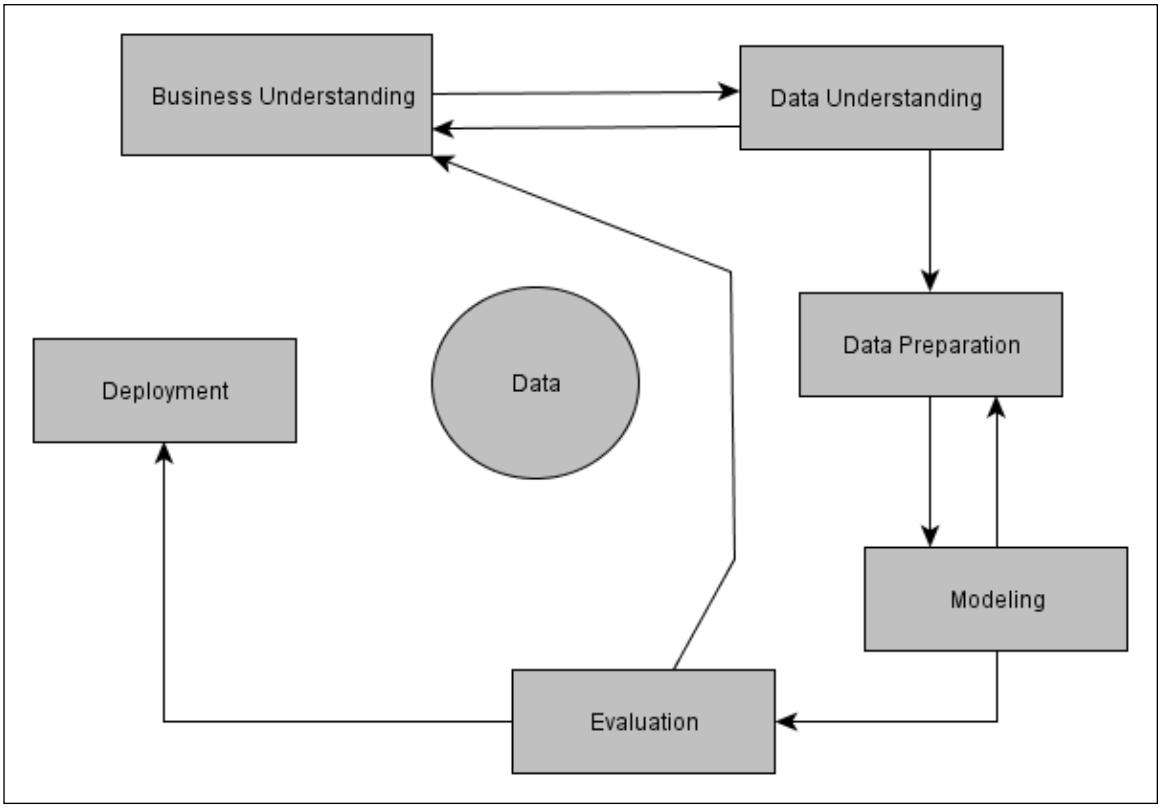

Figure 3.1: Data mining processes

mining is very useful. The identified data can be of value in 1) data driven discoveries to complement model driven one, 2) hypothesis generation to complement hypothesis testing, 3) computational scalability, 4) conceptual scalability – models of GPS tracks.

Potential value of transportation to data mining is 1) expose limitations, e.g. independence assumption, 2) new challenges: e.g. spatio-temporal networks, 3) new pattern families. Data mining can be defined as the non-trivial extraction of implicit, previously unknown, yet potentially useful information from data, and may be defined as the science of extracting useful information from large data sets or databases. With the help of data mining, derived knowledge, relationships and conclusions are often represented as models or patterns. For example, any data cluster, tree structure, or set of rules, etc., can form a model or a pattern. The whole process is sometimes referred to as knowledge discovery in databases (KDD). Data mining only implies modeling or an analytical method in this application sense and is considered to be a part of the KDD process. One standard, named CRISP-DM (Cross-Industry Standard Process for Data Mining), describes this process step by step. It develops each phase of the KDD process and, in addition, helps to avoid common mistakes.

3.4 DATA MINING PROCESSES

Data mining is a promising and relatively new technology. Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence (AI) and statistical (Figure 3.1).

3.4.1 BUSINESS UNDERSTANDING

This is the first phase of data mining where one tries to understand project objective and requirement from business point of view. The knowledge, hence, obtained is then converted to data mining problem definition, which results in development of a preliminary plan to achieve objectives.

3.4.2 DATA UNDERSTANDING

Data understanding begins with collecting relevant data and familiarizing with them to identify data quality problems, to get insight into data, or to detect subsets to form hypotheses for hidden information.

3.4.3 DATA PREPARATION

All and any activity used to construct final dataset, which will be used in modeling, from the initial raw data comes under data preparation. These activities are likely to be performed multiple times, not in any specific order and also during the modeling process. The activities/tasks include table, record, and attribute selection, and transformation and cleaning of data for modeling tools.

3.4.4 DATA MODELING

This is an important phase, where several modeling techniques are selected and applied. Usually, one applies several techniques on the same dataset, as each technique has their own specific requirement on the form of data. Which is why, during this phase, the database is also continuously being modified as in the data preparation phase.

3.4.5 DATA EVALUATION

From the perspective of data analysis, the model built in the previous stage, is of high quality. However, before deploying the model, which is the final stage, it needs to be checked thoroughly including reviewing the steps executed during model construction, and ensuring that the

model achieves its business objectives. During this review, one of the objectives is to check if there are any important business issues that have not been considered in sufficient details. It is expected that at the end of model evaluation, a decision regarding the usability of results have been achieved.

3.4.6 DATA DEPLOYMENT

The knowledge obtained from the model is almost never organized and presented in a way, which is useful to the end user. Additionally, depending on the requirement, the deployment can have varied complexity ranging from as less as generating a report to as much as implementing a repeatable data mining process. In general, clients or end users are the people, who are responsible for deployment, which necessitates the need for the client to understand the actions needed to carry out so as to use the model, hence, created.

CHAPTER 4: MACHINE LEARNING

Machine learning deals with a field of study, design and development of different types of algorithms that provides the computers the ability to learn from the data without giving clear instructions. Machine learning algorithms are able to extract information automatically without any human help. Machine learning is a subfield of computer science and artificial intelligence. It also has close connection with fields such as statistics and optimization. Kalman filter, optimal character recognition, etc. can be termed as an example of a machine learning algorithm [21].

4.1 MACHINE LEARNING ALGORITHM TYPES

Machine learning algorithms can be organized into the following taxonomy based on the outcome that is desired:

- Supervised Learning: These types of algorithm are the ones that are trained on examples called labeled examples where the inputs are provided with the desired output already known.
- Unsupervised Learning: These types of algorithms are the ones that are trained on examples called unlabeled examples where the inputs are provided without the desired output being known.
- Semi-supervised Learning: These types of algorithms operate on both labeled and unlabeled examples in order to produce a desired function.
- Transduction: These types of algorithms try to generate new outputs that are based on fixed test cases from observed training cases.
- Reinforcement Learning: Reinforcement learning is an area of machine learning that is concerned with the way software agents combine in an environment to maximize the reward or outcome.

- Multi-tasked Learning: Multi-tasked learning is an area of machine learning that is trying to generate algorithms that learn a problem simultaneously with related problems taking into account a shared representation that leads to produce a better model for the initial task. This often results in a better model because it helps the learner to take out the best out of both the tasks involved.
- Developmental Learning: Developmental learning also known as robot learning wherein the machine learns on its own based on human interactions and self-explorations and taking the help of guidance tools such as active learning, maturation, etc.

4.2 MACHINE LEARNING AND DATA MINING

Machine learning and data mining are two terms that are commonly confused, as they often employ the same methods and overlap significantly. They can be roughly defined as follows:

The difference between machine learning is that machine learning concentrates on predicting the outcome by learning from the data whereas in data mining the main aspect is to concentrate on finding the unknown traits in data set. It's also called the analysis step of knowledge discovery in databases. Both machine learning and data mining are closely correlated with data mining using a lot of machine learning methods wherein the desired outcome is slightly different. Also machine learning uses a lot of data mining techniques such as unsupervised learning in order to improve learner accuracy.

CHAPTER 5: METHODOLOGY AND ANALYSIS

This chapter evaluates the various predictive machine learning techniques using the chosen data set.

5.1 HYPOTHESIS

The hypothesis is based on the fact that, irrespective of the sex of driver, a traffic crash (which may or may not lead to fatality) is due to young people driving drunk at certain day of the week and at certain hour. It is worthwhile to analyze, the hypothesis that young people, driving drunk, most probably on Fridays (can be any day for that matter), between, say 11:00 PM – 3:00 AM, are more prone to having a traffic accident than rest of the drivers on the road.

5.2 CLUSTER ANALYSIS

The data taken into account is the accident data obtained from FARS. The variables included are:

- Age – age of the driver.
- Alcohol – the amount of alcohol consumed by the driver.
- Travel speed – the speed of the vehicle at the time of the accident.
- Atmospheric conditions – the atmospheric conditions present at the time of the accident, for example - rainy, cloudy, sunny, etc.
- Day of the week – denotes the day of the week on which the accident occurred. It has been noticed that the number of accidents increases on weekends.
- Sex – sex is inferred biological sex from physical appearance.

5.2.1 CONDUCTING THE ANALYSIS

We started by bringing the data file into SPSS. Before bringing the data in SPSS, the county cases were labeled numerically. 100 cases were chosen at random to minimize any inherent bias in the data. To perform the cluster analysis, numerical county label was chosen as the variable by which cases were labeled with alcohol, travel speed, atmospheric conditions, day of the week, sex as the variables. We indicate that we want to cluster cases rather than variables and want to display both statistics and plots.

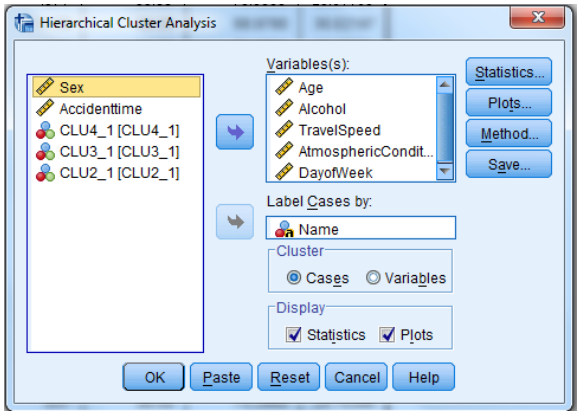

Figure 5.1: Hierarchical cluster analysis – main dialogue box (variables are decided on the basis of assumption)

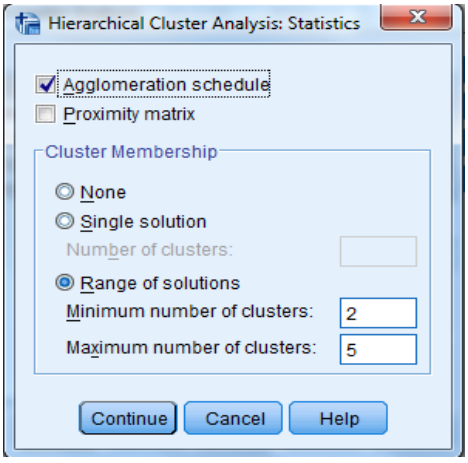

Figure 5.2: Setting agglomeration schedule and cluster membership options

In order to see an agglomeration schedule, we specify the same in the statistics option. For identifying clusters visually, we also specify in the same dialogue box, that we want to see a dendogram and a vertical icicle plot with 2, 3 and 4 cluster solution.

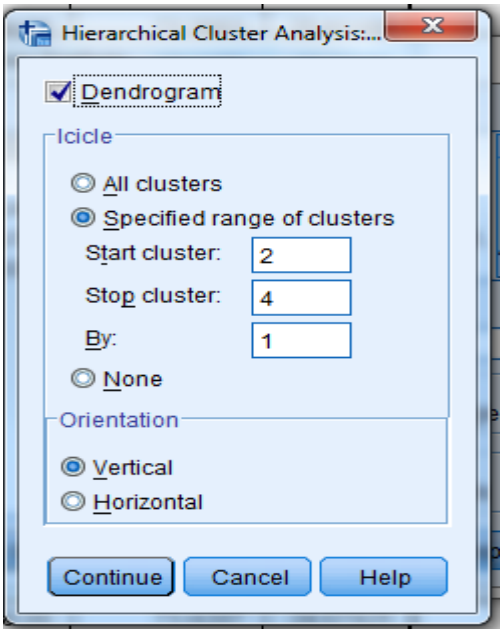

Figure 5.3: Specifying types of plot needed

For the variables to be able to contribute equally, we standardize them to Z scores. We use the Between-group linkage method of clustering, and squared Euclidian distances as our preferred method (Figure 5.4)

We save, for each case; the cluster for the case is assigned from 2, 3, and 4 cluster solution (Figure 5.5).

On initiating the processing, SPSS starts by standardizing all variables to mean 0 and variance 1, which has the benefit of having all variables on same scale and being equally weighted. SPSS does

this analysis in several steps. In the first step, SPSS computes for each pair of cases – the squared Euclidian distance between the cases.

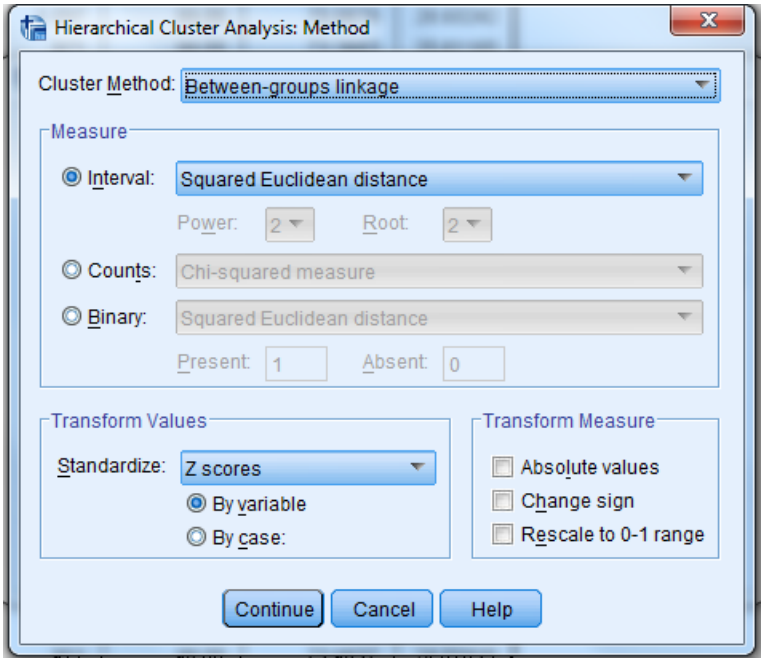

Figure 5.4: Methods used for hierarchical cluster analysis

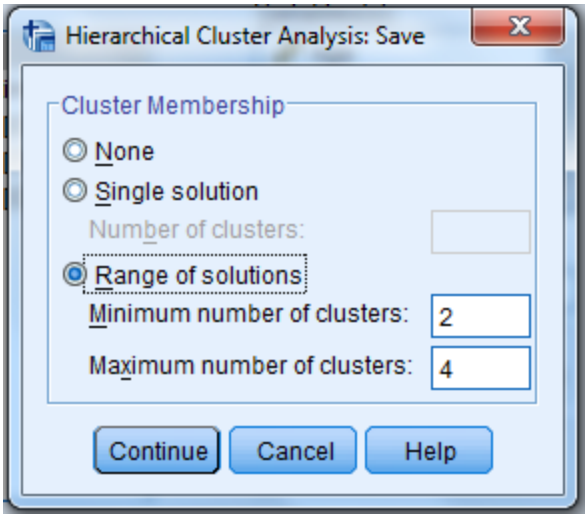

Figure 5.5: Saving the clusters

This is given as:

$$\sum_{i=1}^{v} \left(X_i - Y_i \right)^2$$

which is the sum across variables (from i=1 to V) of the squared difference between the score on variable i for the one case (X_i) and the score on variable I for the other case (Y_i). The smallest Euclidian separation identifies the two cases that will then be classified together into first cluster, thus making the first cluster with two cases in it.

In the second step SPSS re-computes the squared Euclidian distances between each entity (case or cluster) and every other entity. SPSS computes average squared Euclidian distance between members of one entity and members of other, whenever one or both of the compared entity is a cluster. As with previous step, the two entities with smallest squared Euclidian distance are classified together. The second step is repeated over and over again, until all of the cases have been clustered in one big cluster.

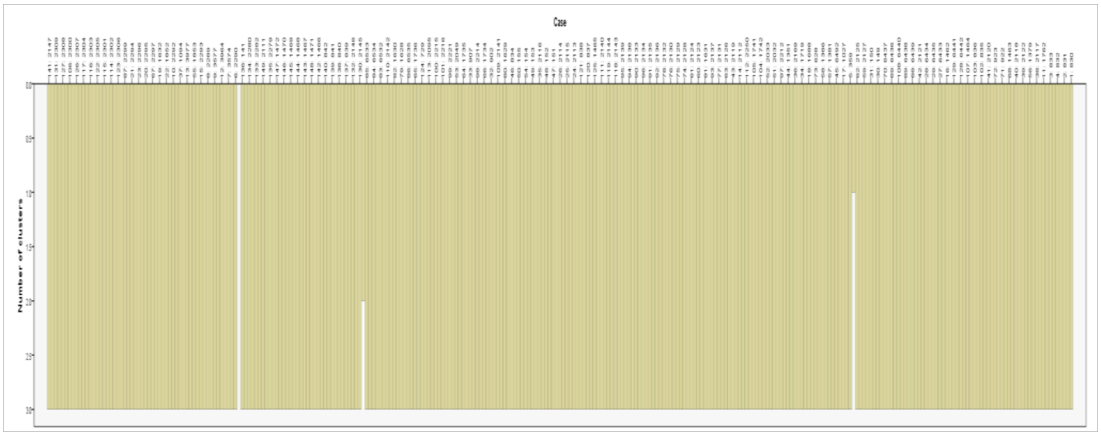

Figure 5.6: Vertical icicle diagram of cluster

Based on the agglomeration schedule (Appendix 1, Table 1), we see that on the first step, SPSS clustered case 97 with 99. The squared Euclidian distance between these two cases is 0.000. In subsequent cases, SPSS creates more clusters each containing two cases and keeps adding more cases to the cluster. By the 99th stage all cases have been clustered into one entity.

It is evident from the vertical icicle (Figure 5.6) that for the two cluster solution – one cluster consists of 35 cases, followed by a column with no X's. The atmospheric condition for these cases consisted of visibility impairing conditions like smog, snow, etc and hence, was re-classified as having "Poor Visibility". Second cluster consisted of rest of the cases, which was considered as having "Good Visibility", as the atmospheric conditions were mostly clear. The visibility condition is inherent in the data in form of atmospheric condition and has a good impact on driving conditions. It is understood from common knowledge that the chances of accident increases as the visibility decreases.

For the three-cluster solution, it can be observed from the same icicle plot that the cluster of good and average visibility is split into two. For the four-cluster solution, there are a total of 116 cases with visibilities ranging from good to average to poor.

The dendogram (Figure 5.7) shows essentially the same information as the agglomeration table – but in graphic format.

The three cluster solutions have been added to the datasheet and labeled as CLU2_1, CLU3_1 and CLU4_1 for two, three and four cluster solutions. For better understanding of the data, we remove the variable labels and then label the values for CLU2_1 and CLU3_1.

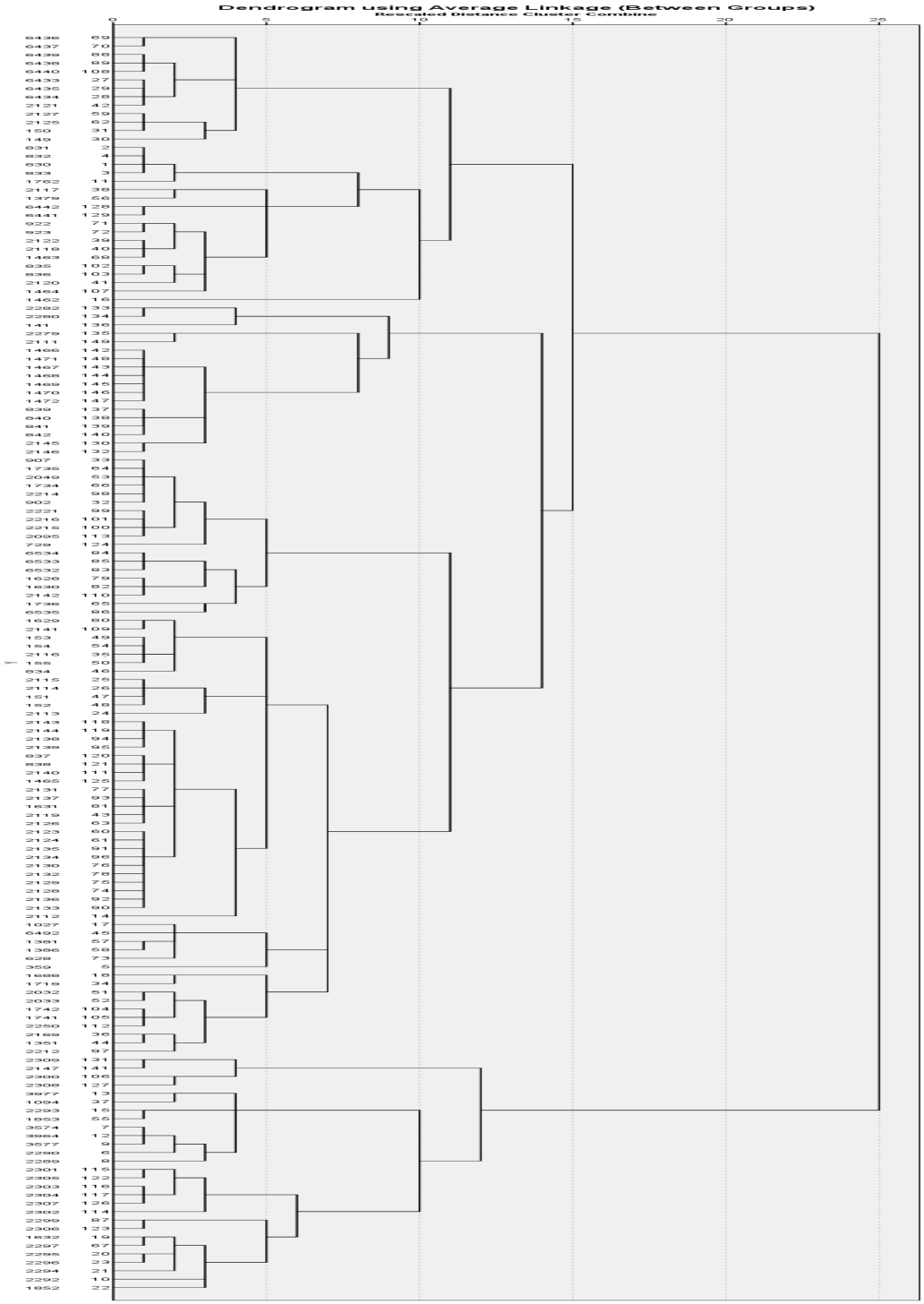

Figure 5.7: Dendogram diagram of agglomeration schedule

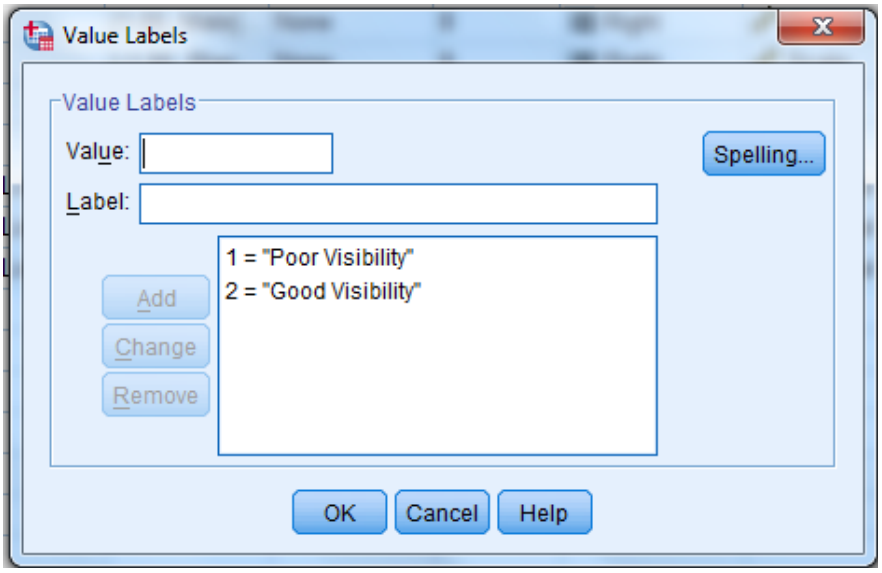

Figure 5.8: Value labels for CLU2_1

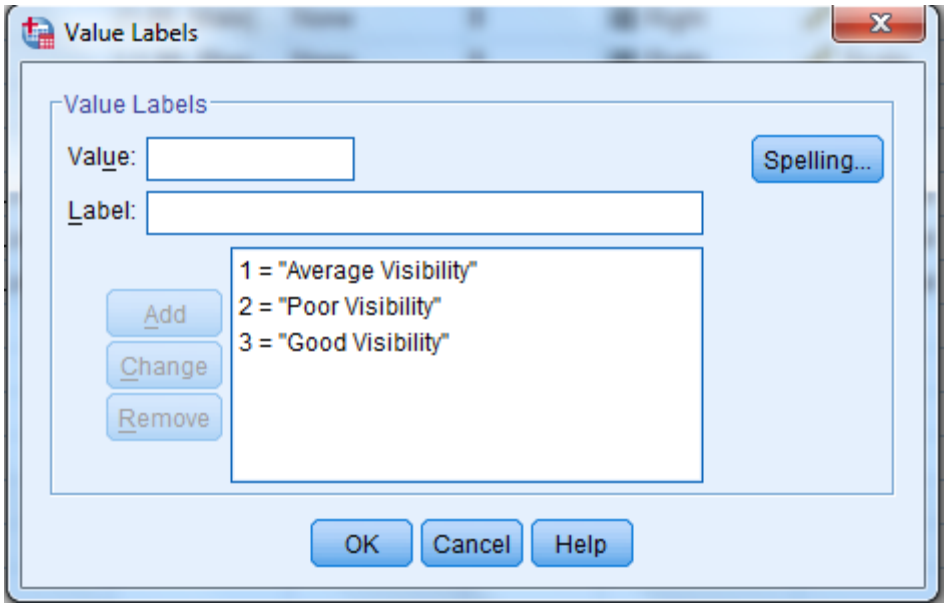

Figure 5.9: Value labels for CLU3_1

We further analyze the data by comparing means for Independent – Samples T test and find out how the two clusters in the two-cluster solution differ from one another on the variable used to cluster them.

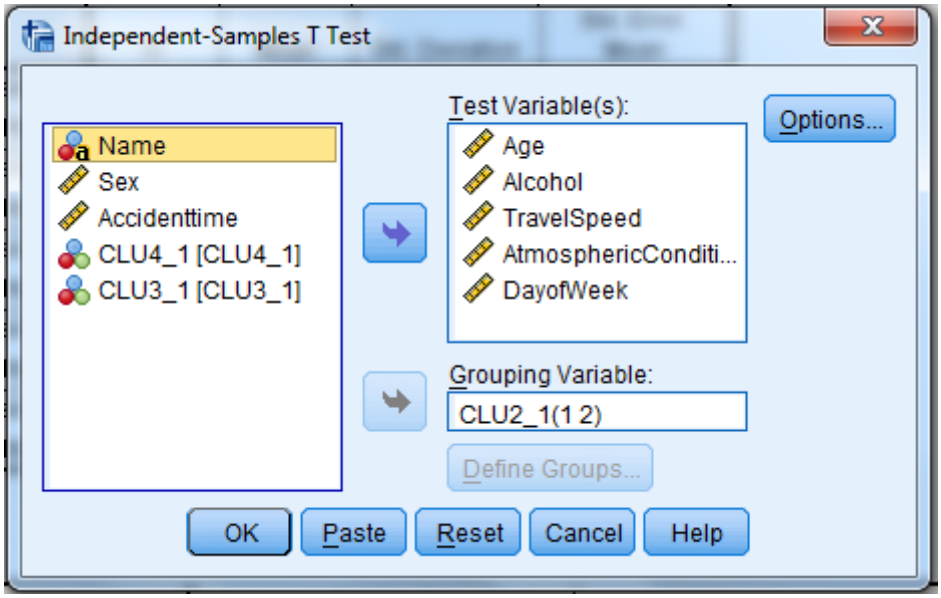

Figure 5.10: Independent samples T test

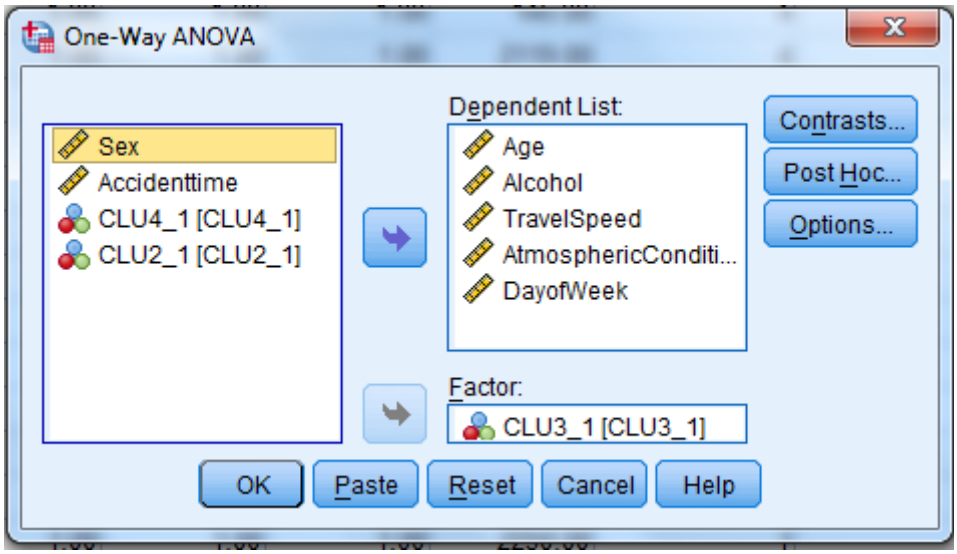

Figure 5.11: One-Way ANOVA

The output (Table 5.1) suggests that cluster "poor visibility" has higher age and travel speed, with lower alcohol content.

Next we compare the three clusters from the three-cluster solution using One-way ANOVAs and generating the plots of group means.

We see that the plot of means (Fig 5.12 a-e) show nicely the difference between the clusters. Additionally, it also gives us a fair idea about how the variables are related and behaving with respect to each other.

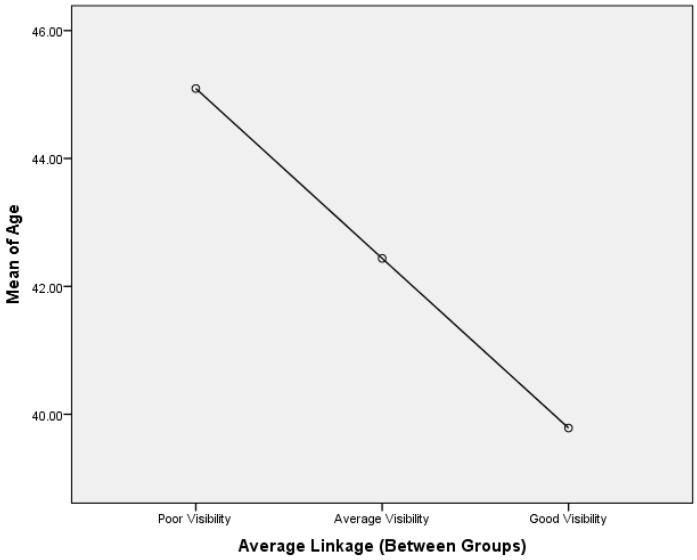

Figure 5.12 a: Mean plot of age vs visibility for three cluster solution

5.2.2 PREDICTING TRAVEL SPEED FROM AGE, ALCOHOL, ATMOSPHERIC CONDITION AND DAY OF THE WEEK

Our hypothesis is that accidents take place due to high travel speed. Travel speed is dependent on several factors, of which the human and climatic factors have been analyzed in this thesis. The premise of the data selection is that young drivers (lower age) are prone to drive faster and in more irresponsible manner than older drivers. Yet another premise is that in better atmospheric conditions with good visibility, people tend to drive faster, as they can see farther.

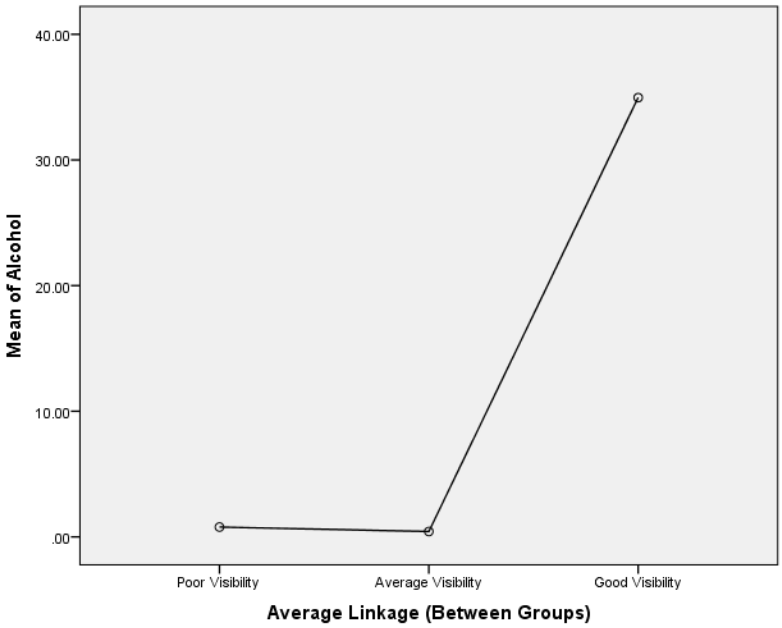

Figure 5.12 b: Mean plot of blood alcohol content vs visibility for three cluster solution

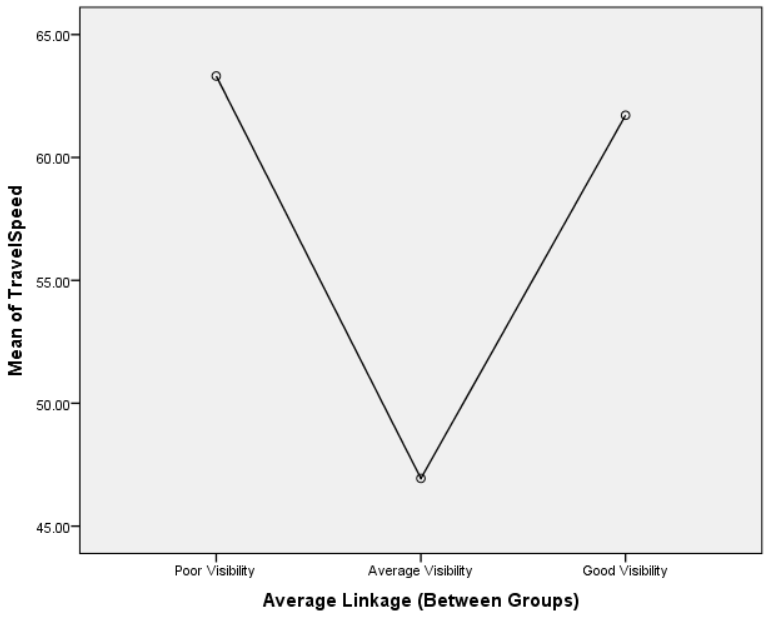

Figure 5.12 c: Mean plot of travel speed vs visibility for three cluster solution

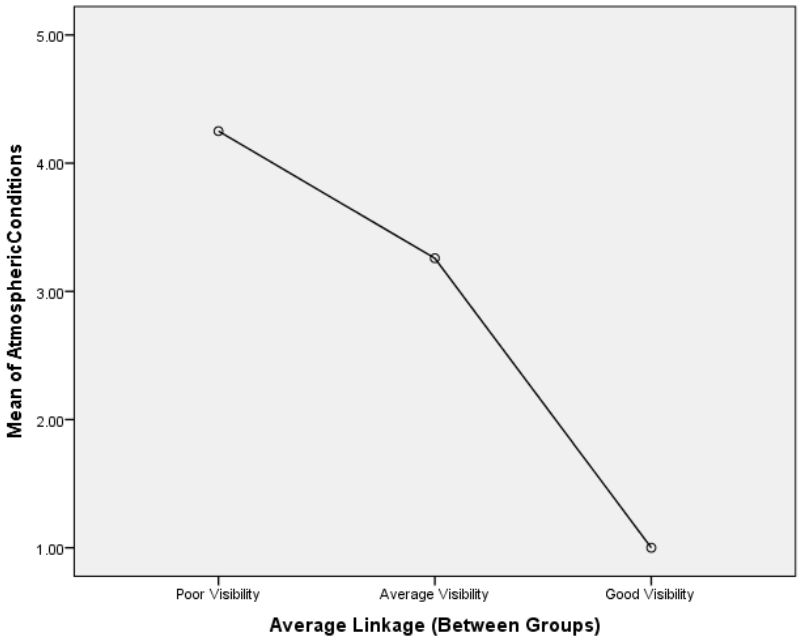

Figure 5.12 d: Mean plot of atmospheric condition vs visibility for three cluster solution

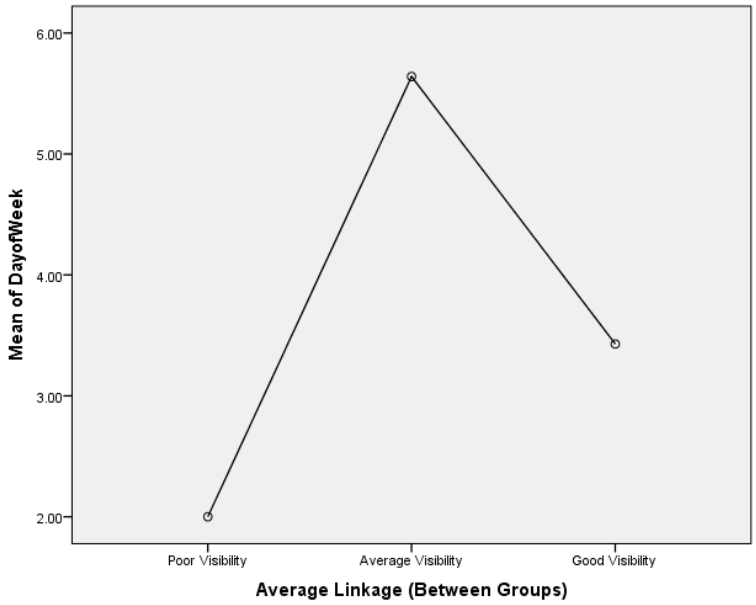

Figure 5.12 e: Mean plot of day of the week vs visibility for three cluster solution

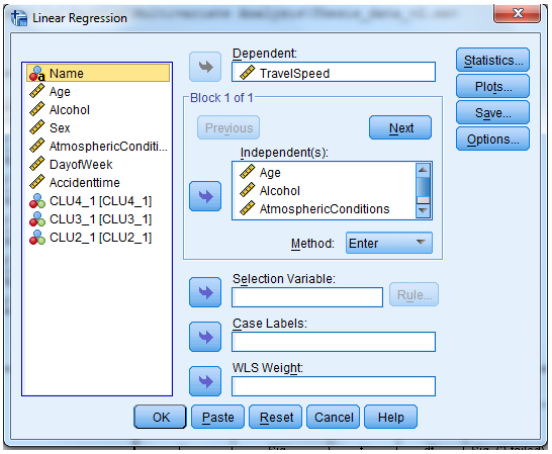

Figure 5.13: Linear multivariate regression for predicting travel speed based on age, alcohol content, atmospheric conditions and day of the week

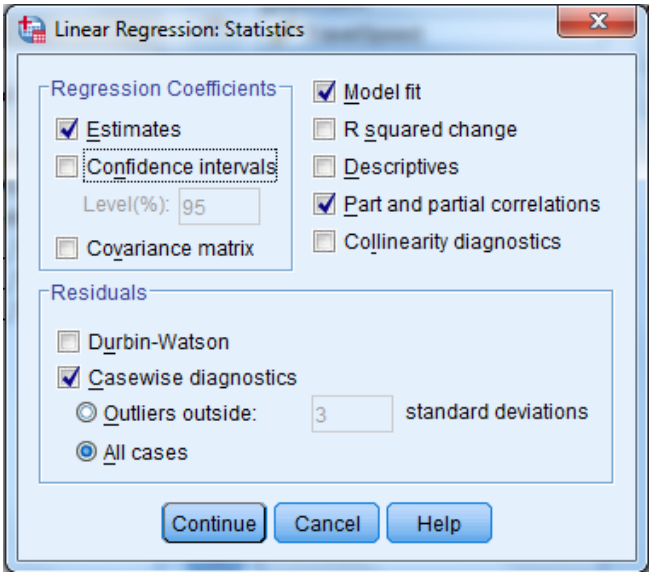

Figure 5.14: Specifying statistical parameters for multivariate regression

Alcohol also plays an important role as it not only impairs judgment of a driver, but also reduces reaction time. So, in case of higher travel speed with higher blood alcohol content, the reduced reaction time will result in more instances of crashes and hence fatalities. Day of the week is also considered as a factor of higher travel speed as in some of the places, Monday mornings and Friday evenings are times of high traffic volumes and more instance of crashes. So, to find out the relation

between all these variables, we use multiple regressions. We want to see how travel speeds are related to Age, Alcohol, Atmospheric Condition and Day of the week.

While doing the analysis, we also ask for part and partial correlation and for case wise diagnostic for all cases. The output shows that each of our predictors, except alcohol has negative zero-order correlation with travel speed. Out of the inputs, alcohol, atmospheric conditions and accident time show significant impact on travel speed as seen from the output.

In the case wise diagnostic table, standardized residual for each case is given, along with the actual travel speed and the travel speed predicted by model and the difference between the two. Those standardized residuals, whose absolute values exceed 1, are worthy of reconsideration. The current model is far from perfect, but it does show some pretty good fits.

To decide which drivers are safer, we split the data by sex, and repeat the regression analysis. It is interesting to note that for female drivers, the partial effect of alcohol and day of the week is more negative than their male counterparts, which suggests that alcohol and day of the week has more effect on men than women. For both drivers the time of accident has less partial effect on travel speed.

5.3 HIERARCHICAL MULTIPLE REGRESSION

If we have any information, which allows us to justify entering the predictor variables in any particular order, a hierarchical analysis may be appropriate. In our analysis, we hypothesize, that age and alcohol content have more effect on travel speed, than the Atmospheric conditions than day of the week than the time of the day. To find out if this assumption is correct, the best way to find out is to check out, how the different variables affect the regression. To do that, we check out the linear regression option as we did in the previous step and apply the variables as discussed above (Figure.

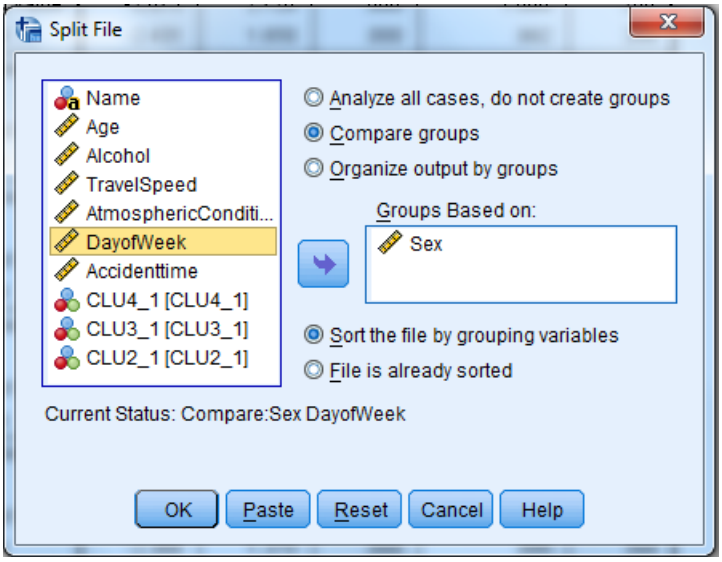

Figure 5.15: Splitting the data based on sex of drivers

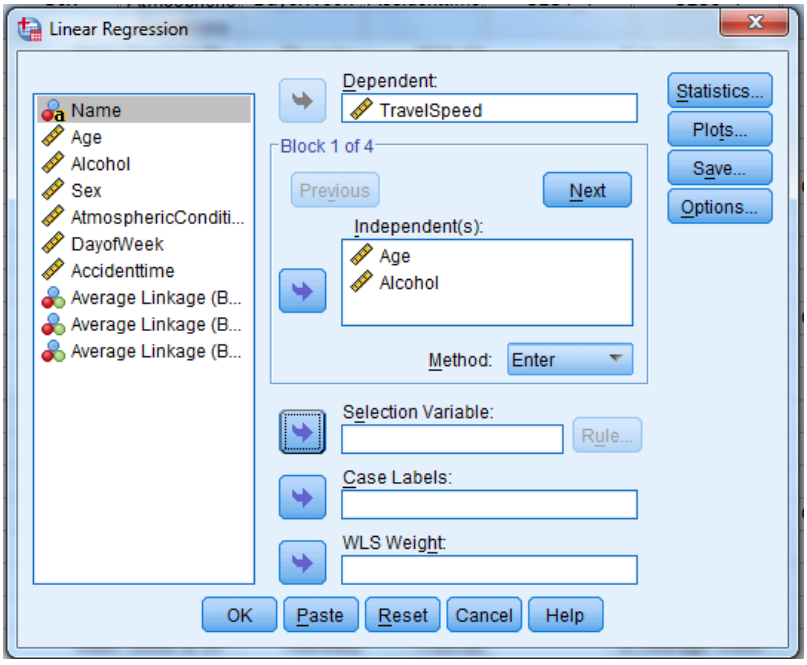

Figure 5.16 a: Hierarchical linear regression main dialogue box

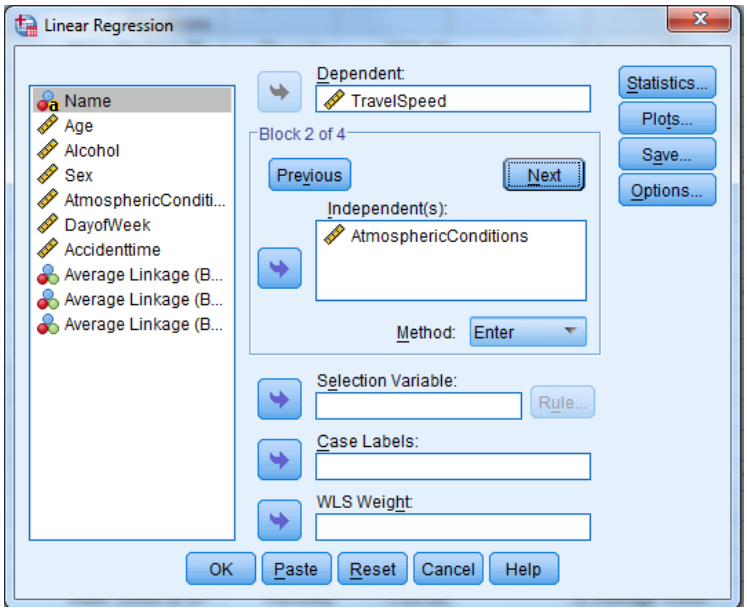

Figure 5.16 b: Entering independents for 2nd block

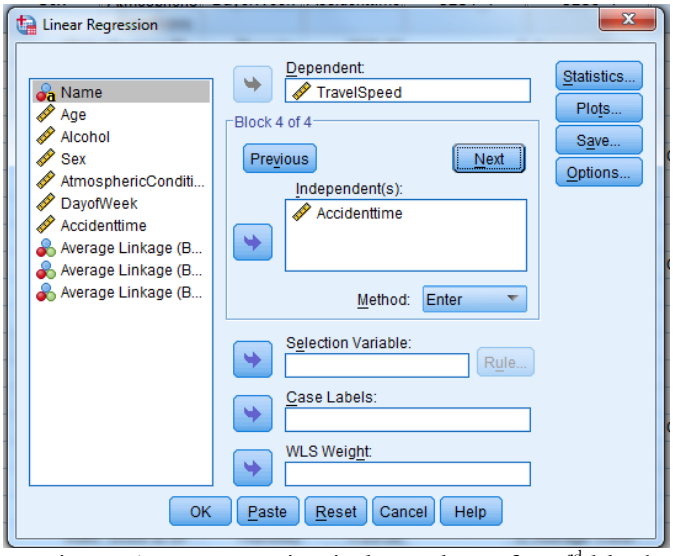

Figure 5.16 c: Entering independents for 3rd block

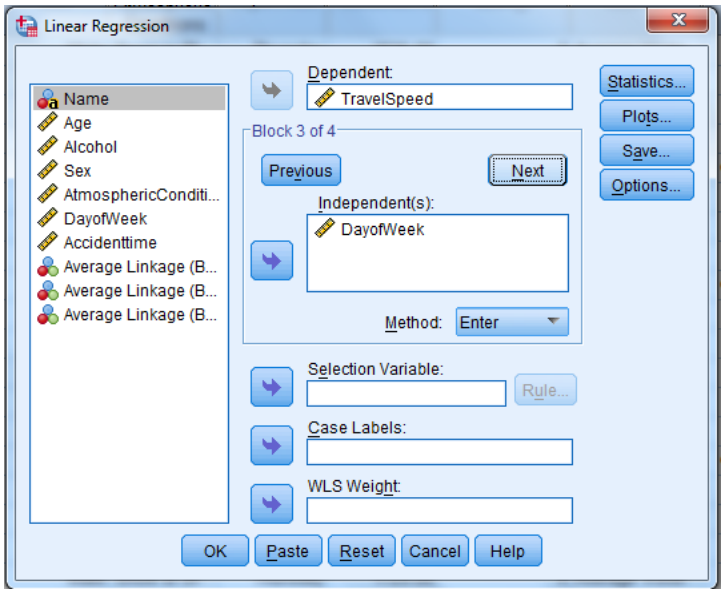

Figure 5.16 d: Entering independents for 4th block

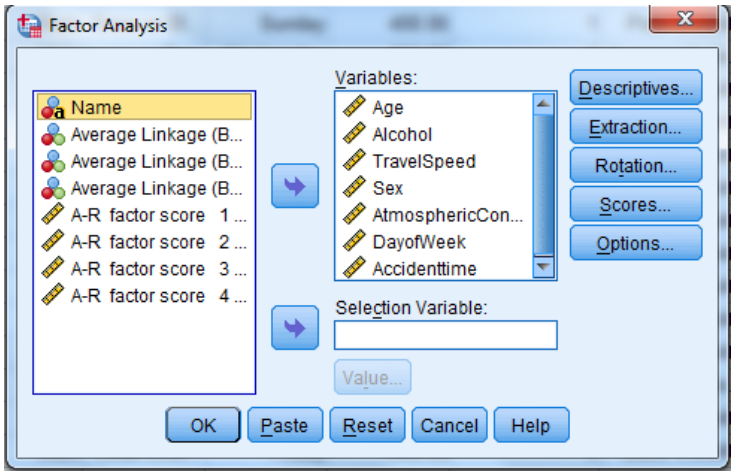

Figure 5.17 a: Factor analysis main window

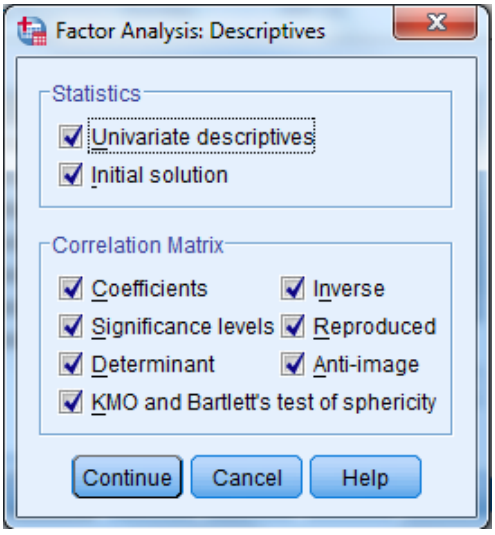

Figure 5.17 b: Factor analysis, descriptive option

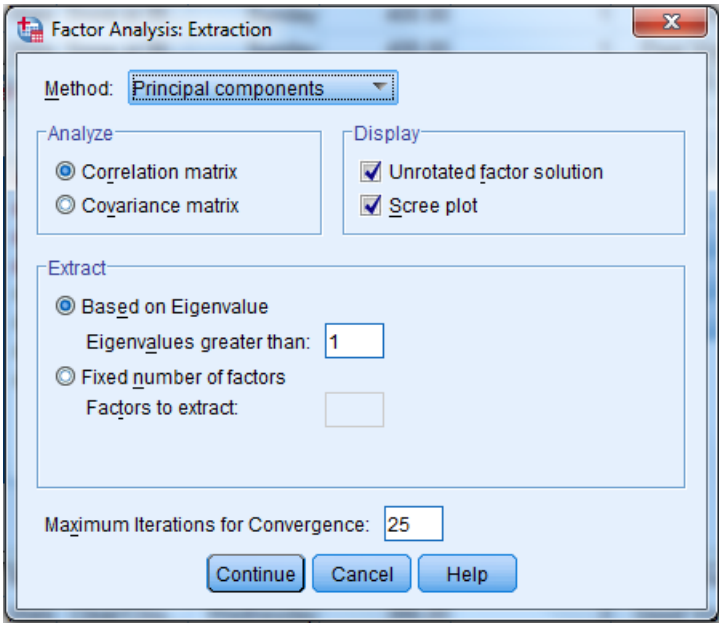

Figure 5.17 c: Factor analysis extraction option

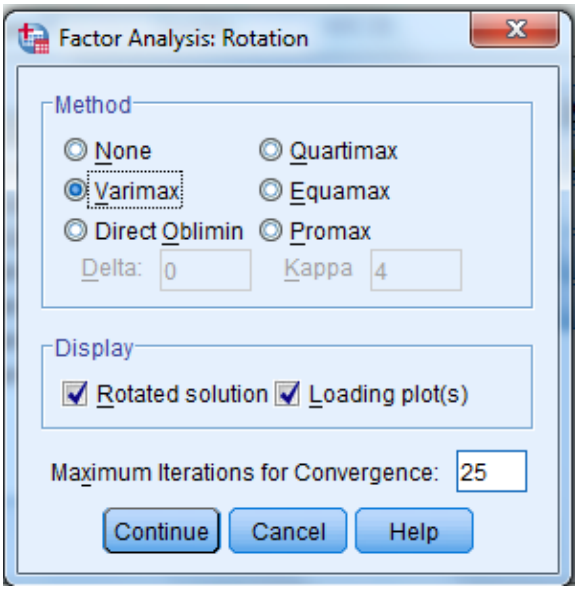

Figure 5.17 d: Factor analysis rotation option

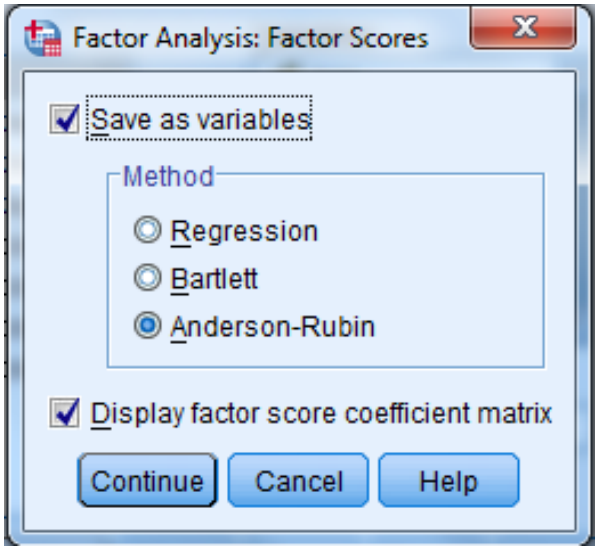

Figure 5.17 e: Factor analysis factor scores

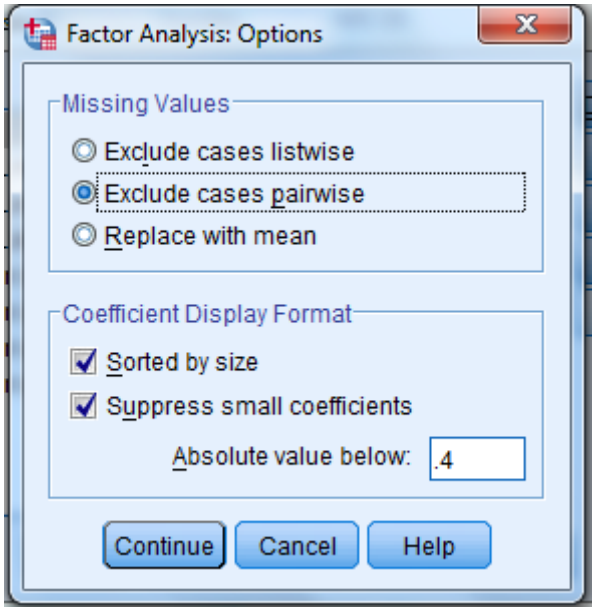

Figure 5.17 f: Factor analysis options

5.16 a –d). Based on the results (Table 5.8 – 5.10), we find, as expected, that alcohol, in all cases leads to higher travel speed and hence results in more crashes. This deduction, by all means, is not something that was not known earlier, but it is one more nail in coffin for drunk driving. In addition to alcohol content, which is a significant contributor to accidents, the time of accident, age of the driver, atmospheric condition and lastly the day of the week were a major component of contributors. Steps involved in Hierarchical regression analysis have been shown in figure 5.16 a – d.

5.4 FACTOR ANALYSIS

The user accesses the main dialog box (Figure 5.17 a) by using the menu path shown below:

Analyze → Data Reduction → Factor. The user simply selects all the variables that have been selected for the analysis. Since, we have already identified and removed any variables/data, we deemed problematic; we do not have to exclude anything. The step-by-step instruction to analyze the dataset is given in the figures 5.17 a – f.

SPSS nearly always finds a factor solution for a set of variables. However, if the variables are not sensible, evidently, the solutions will make no sense. This is why, the first thing to do when conducting a factor analysis is to look at the inter-correlation between variables. If the test question measures the same underlying dimension, then we would expect them to correlate with each other, since they are all measuring the same thing. If the variables do not correlate with each other, then these variables should be excluded before factor analysis is run.

5.4.1 PRELIMINARY ANALYSIS

SPSS output 1 shows an abridged version of the R-matrix. The top half of this table contains the Pearson correlation coefficient between all pairs of questions whereas the bottom half contains the one-tailed significance of these coefficients. We can use this correlation matrix to check the pattern of relationships. The value of determinant, given at the bottom of the output (Table 5.11) – is significantly greater than the necessary value of 0.00001, which suggests that multicollinearity will not be an issue with our dataset and that the data being used is good, which is to sum up, none of the correlation are typically large, which will warrant a second look at our data.

The next part of output shows the Bartlett test of Sphericity, which tests the null hypothesis, that the correlation matrix is an identity matrix. For factor analysis to work we need some relationships between variables and if the R-matrix were an identity matrix then all correlation coefficients would be zero. This suggests that we would need the test to be significant and hence, the significance value has to be less than 0.05. This test tells us, that R-matrix is not an identity matrix, and hence there is some relationship between the variables, which we can hope to include in the analysis. For our data, Bartlett's test is highly significant (sig. < 0.001), and therefore factor analysis is appropriate.

The above output also shows Kaiser-Meyer-Olkin test of sampling adequacy, which varies between 0 and 1. A value closer to 1, indicates correlations are relatively compact and so factor analysis should yield distinct and reliable factors. The usual recommendation is to accept values greater than 0.5, however in our data, which has been randomly sampled over 10 years, we will accept the current KMO value at 0.457, since it is close to the acceptable limit, and also in part due to random nature of data selection to keep any bias at minimum (Table 5.12).

5.4.2 FACTOR EXTRACTION

The third output (Table 5.13), shows the eigen values associated with each factor before and after extraction and after rotation. Before extraction, SPSS identified 7 linear components within the data set, which is same as the number of variables. It should be noted that in all cases, there should be as many eigenvectors as there are variables, and so there will be as many factors as variables. As we can see, the total cumulative variance is 100% and the amount of variance is higher for first few variables than the rest. Since, only 4 variables have variance higher than 1, SPSS extracts them for the next stage (variables for "after extraction" phase). These are labeled as "Extraction Sums of Squared Loadings", where the values of these are same as before extraction, except the values of discarded variables are ignored and the sum of total variances is less than 100%. The last part of the table shows the "Rotation Sums of Squared Loadings", which is actually optimizing the factor structure and thereby, equalizing the relative importance of the four factors. Before rotation, factor 1 accounted for ~27% variance, which changed to ~25% after rotation.

The next output (Table 5.14) shows the table of communalities before and after extraction. Principal component analysis works on the initial assumption that all variance is common; hence the communalities before extraction for all variables are 1. The communalities in the column labeled

"extraction" reflect the common variance in the data structure. So, for example, we can say that >90% of the variance associated with atmospheric conditions is common or shared variance. At this stage, SPSS has extracted four factors, which based on Kaiser's criterion, is accurate if the sample size is greater than 250 (not in our case) and average communality after extraction are greater than 0.6, and the number of variables are less than 30 (which is true in our case). Based on these, Kaiser's rule applies to our situation.

5.4.3 FACTOR ROTATION

The next output shows us the rotated component matrix, which is a matrix of the factor loadings for each variable onto each factor. This matrix shows us the same information as before, but after rotation. In this matrix, any factor loading below 0.4 have not been displayed, as we asked them to be suppressed. However, all the factors show loadings above 0.4 and hence, we can see the effects of all these variables. Also, these variables are listed in the order of their size of factor loadings as we specified the output to be "sorted by size". Comparing this output with the unrotated solution, we see that most variables were loaded highly on the first factor and the remaining factors were less looked into. Rotation of factors changed things considerably: there are four factors and variables load very highly onto only one factor.

Based on the factors, we can label the first component as "Incident based on atmospheric condition and alcohol content", second component as "Incident based on day of the week and travel speed" and third as "Incidents based on age of driver and time of the day" (Figure 5.18). This analysis reveals that the initial questions that we are asking in this research are: Do adverse driving conditions like weather and alcohol result in more incidences of traffic accidents? Does a particular

day have more incidences of traffic accidents than the rest? Does drivers of a given age group, driving at a particular time of day are more prone to accidents as compared to others?

However, the factor analysis, does not indicate, which of these might be true.

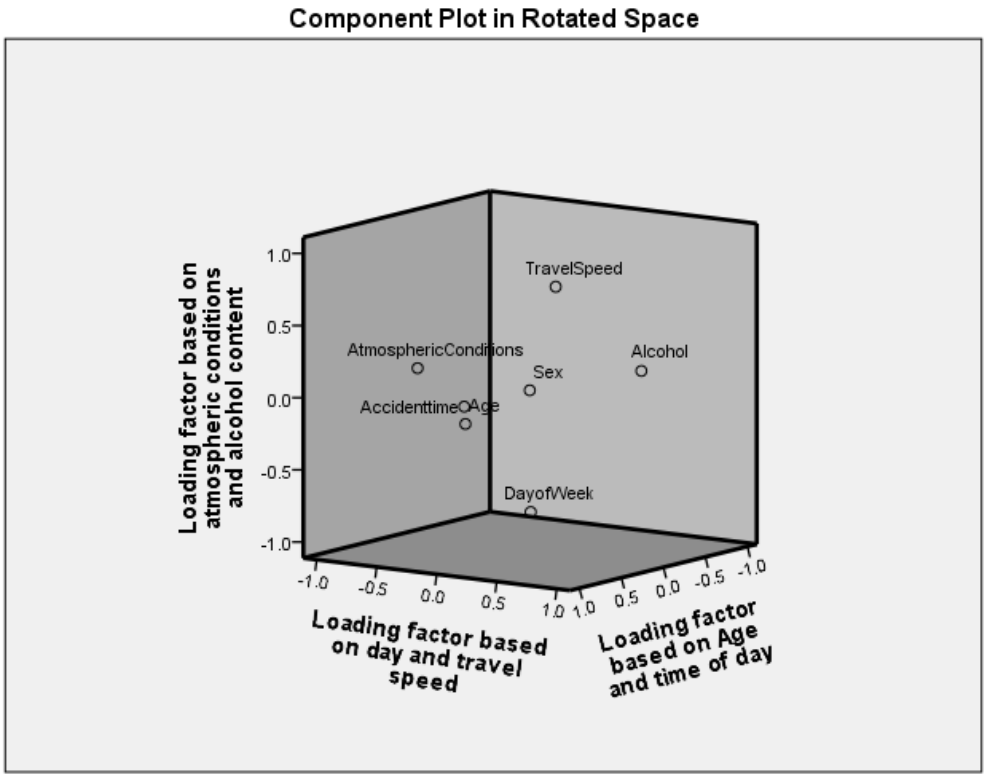

Figure 5.18: Component plot based on component matrix

5.5 DISCRIMINANT ANALYSIS

This method does the same task as multiple linear regressions by predicting an outcome. However, multiple linear regressions is limited to cases where the dependent variable on Y axis is an interval variable so that the combination of predictors will, through the regression equation, produce estimated mean population numerical Y values for given values of combination of X values.

Discriminant analysis involved determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is:

$D=v_1X_1 + v_2X_2 + v_3X_3 = \ldots\ldots v_iX_i + a$

where, D = discriminant function

v = the discriminant coefficient or weight for that variable

X = score for the variable

a = a constant

i = number of predictor variables

The major underlying assumptions of DA are:

- The observations are a random sample;
- Each predictor variable is normally distributed;
- Each of the allocations for the dependent categories in the initial classification are correctly classified;
- There must be at least two groups or categories, with each case belonging to only one group so that the groups are mutually exclusive and collectively exhaustive (all cases can be placed in a group);
- Each group or category must be well defined, clearly differentiated from any other group(s) and natural. Putting a median split on an attitude scale is not a natural way to form groups. Partitioning quantitative variables is only justifiable if there are easily identifiable gaps at the points of division;

- For instance, three groups taking three available levels of amounts of housing loan;
- The groups or categories should be defined before collecting the data;
- The attribute(s) used to separate the groups should discriminate quite clearly between the groups so that group or category overlap is clearly non-existent or minimal;
- Group sizes of the dependent should not be grossly different and should be at least five times the number of independent variables.

5.5.1 METHOD

Discriminant Analysis is used primarily to predict membership in two or more mutually exclusive groups. The procedure to perform this analysis using SPSS has been outlined in the following figures. The menu selection opens the dialogue box shown in Figure 5.19 a. The first step here is to enter the grouping variable. We had tested the drivers by sex and we will use the same here as grouping variable. Then we define the lowest and highest coded value for the grouping variable by clicking on "Define Range". As our variable category has only two levels, we can enter 1 and 2 in the boxes (Figure 5.19 b). After we are done with this, we select the independent variables (age, travel speed, atmospheric conditions etc.) in the "independents" box.

Now we will start specifying statistical method we are going to use in this analysis. The option of analysis by this method uses means, univariate ANOVAs, Box's M and Unstandardized Function Coefficients (Figure 5.19 c). We use:

- Means: means and standard deviations for each variables for each group (the two sexes in this case), and for the entire sample

- Univariate ANOVAs: Compares the mean values for each group for each variable to see if there are significant differences between means
- Box's M: test for the equality of the group covariance matrices. For sufficiently large samples, a non-significant p value means there is insufficient evidence that the matrices differ. This test is sensitive to departure from multivariate normality.
- Unstandardized function coefficients: Unstandardized coefficients of the discriminant equation based on the raw scores of discriminating variables.

We now choose the classify option (Figure 5.19 d). Here we have several options to select from, like prior probabilities and plots, summary table etc.

Finally, we save the output (Figure 5.19 e) as new variables: Predicted group membership, discriminant scores and probabilities of group membership.

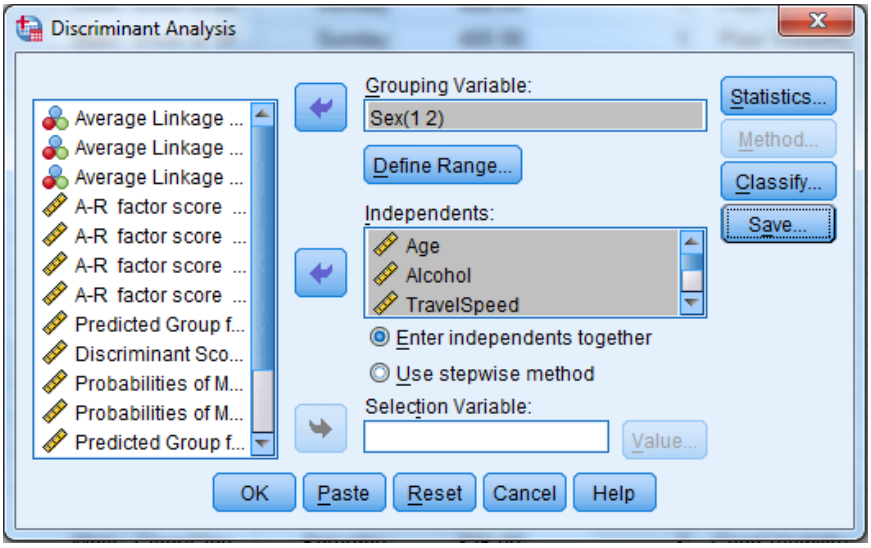

Figure 5.19 a: Discriminant analysis main box

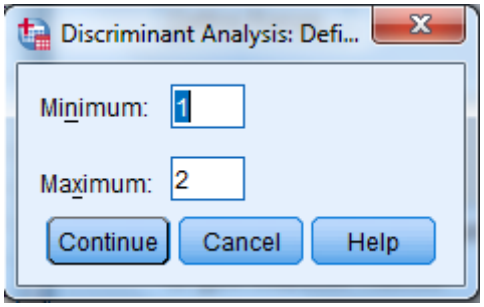

Figure 5.19 b: Defining discriminant analysis grouping variable range

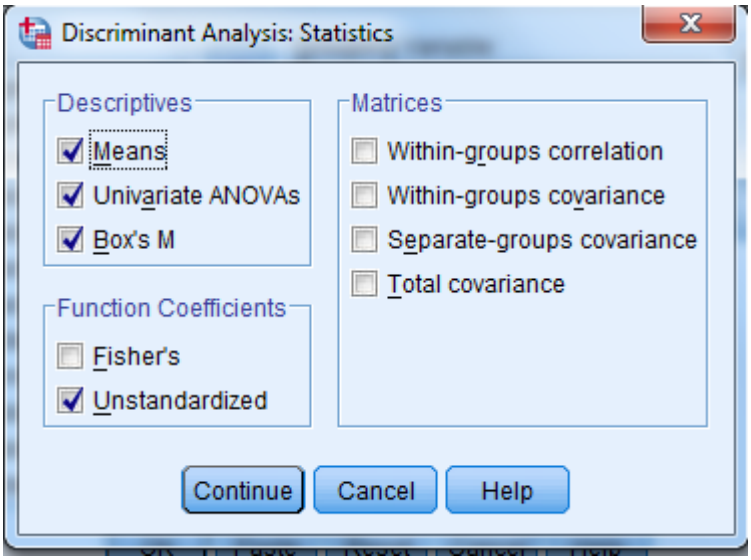

Figure 5.19 c: Discriminant analysis statistics option

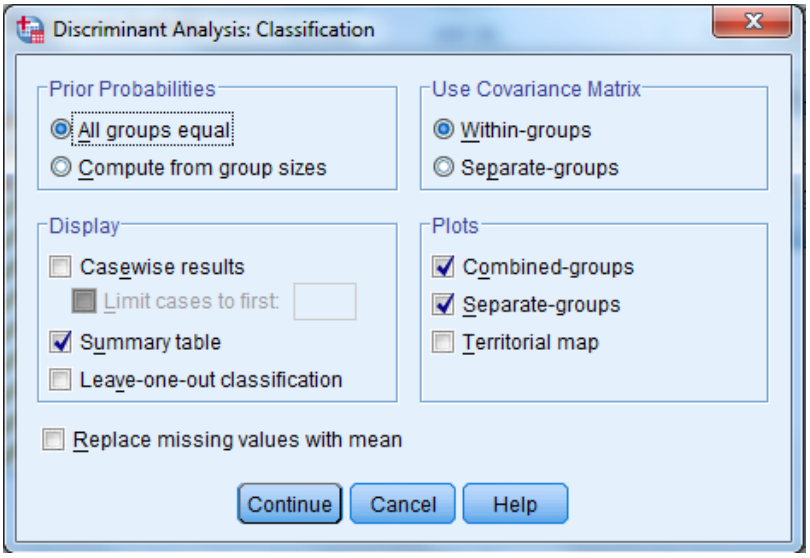

Figure 5.19 d : Discriminant analysis classification option

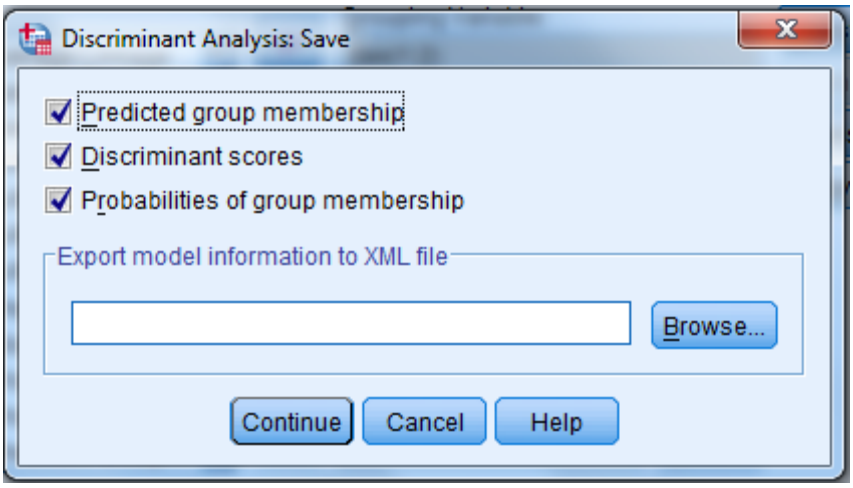

Figure 5.19 e: Discriminant analysis save option

5.5.2 RESULTS

In discriminant analysis, we are trying to predict a group membership, so we examine whether there are any significant differences between groups on each of the independent variables

using group means and ANOVA results data. The group statistics (Table 5.16) and test of equality of group means (Table 5.17) provide this information. If there are no significant group differences, it is not worthwhile proceeding any further with the analysis. A general idea about the important variables can be obtained by inspecting the group means and standard deviations.

For example, the mean differences between the accident time and the age of the driver suggest that these may be good discriminators as separations are large. The test of equality of group means table provides us with strong statistical evidence of significant differences between male and female group for all independent variables with age and travel speed producing very high value F's. The pooled within-group matrices (Table 5.18) also support using these independent variables as the intercorrelations are low.

5.5.3 LOG DETERMINANTS AND BOX'S M TABLES

An underlying assumption in ANOVA is that the variances were equivalent for each group, but in DA the assumption is that variance-co-variance matrices are equivalent. Box's M hypothesis tests the null hypothesis that the covariance matrices do not differ between groups formed by the dependent. We want this test to not be significant, so that the null hypothesis that the groups do not differ can be retained.

For this assumption to be valid, the log determinants should be equal. When tested by Box's M, we are looking for a non-significant M to show similarity and lack of significant differences. In this case, log determinants appear similar and Box's M is 16.226 with F as 0.729 which is significant at $p < 0.808$. In this case, the number of samples are large and there are more than three groups,

which is why we have $p = 0.808$. In such cases, a significant result is not regarded as important and groups with small log determinants are deleted from the analysis.

CHAPTER 6: ANALYSIS IN R

R is an open source programming language used for statistical analysis and data manipulation. R has powerful graphics abilities which are strongly linked with its analytical abilities. Simple analysis and calculations can be handled easily by R. As R is free, it's widely used among researchers and data analysts and for the same reason, R has been added as a part of this thesis.

6.1 BARTLETT'S TEST AND KAISER-MEYER-OLKIN (KMO) INDEX

Before any analysis in principal components, it remains essential to carry out a test of sphericity (test of Bartlett), in order to test the worthless assumption with knowing all the correlations of the variables are equal to zero. Using the function Bartlett.test in FactoMineR package, we find out the results given in Table 6.1:

We may conclude that we reject the null hypothesis and by consequent the variable are indeed correlated and globally dependent. The variables are all factorable. The next step consists on calculating the individual KMO values and the global KMO value. The measurement of KMO is an index of the adequacy of the factorial solution; it makes it possible to check the coherence of the variables selected. A high KMO indicates that there exists a statistically acceptable factorial solution that represents the relations between the variables. For this purpose, we use the KMO function available in psych package. The results obtained are shown in Table 6.2.

The Overall KMO increased after eliminating the Day of week variables, it becomes superior to 0.5, we may conclude the adequacy of the factorial solution and we may start now our analysis. In fact, the entire set of variables has a KMO value superior or equal to 0.5. It is thus concluded that an analysis in principal components is possible, and that all the variables must be used.

6.2 PRINCIPLE COMPONENT ANALYSIS

- **Principle**

The objective of the analysis in principal components is to provide a representation that makes it possible in only one glance to quickly seize the whole of the components presented and to highlight the links of correlations between the variables and the similarities between the individuals.

- **Application**

The KMO provide negative results, some variables are not such pertinent and should be eliminated from the principal component analysis. We delete the variable Day of week with the lowest KMO value and we recomputed the KMO of the other variables (Table 6.3). We use for this purpose, the PCA function available in the FactoMineR package. The Eigenvalues are given in Table 6.4. The four components explain 83.54% of the total variance that means these four axis may represent all the available information of our variables.

Each component adds supplementary information and enables to represent well, at least one of our variables. However, in this analysis, we'll only consider 3 components to make the analysis easier and more understandable. Table 6.5 presents the coordinate of each variable according the three components.

We will now represent the individuals and also the variables using the three principles component analysis. The first on the first plan (axis (1, 2)), the second plan (axis (1, 3)) and the final plan (axis (2, 3)). The principles are:

- The link, or in other term the correlation between two variables, is interpreted in term of the angle which the two variables form where 2 axis carrying the 2 variable points. Thus, it is necessary to not interpret the variables in terms of points but in terms of variable axis.

- The proximity between 2 individuals means similar behavior of these 2 individuals with respect to the whole of the variables. The point's individuals that are in the center of the graph either admit an average behavior or are points badly represented from this point of view.

Let's now interpret each factorial plan independently:

The variable's and individual's representations (axis (1; 2)):

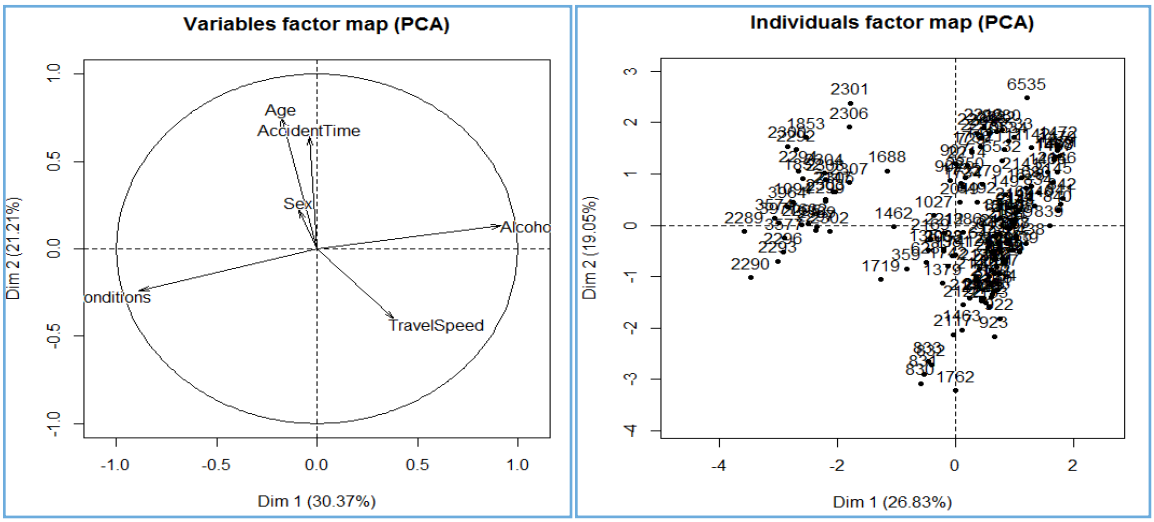

Figure 6.1: The graphic representations of the variables and the individuals (axis (1, 2))

- The variables Age, Sex and Accident time are highly and positively correlated, the three variables may be represented by one synthetic variable representing them all.
- The high and negative correlation between alcohol and atmospheric conditions.
- The negative correlation between the previous synthetic variable and the travel speed variable.
- Weak correlation between alcohol and travel speed.

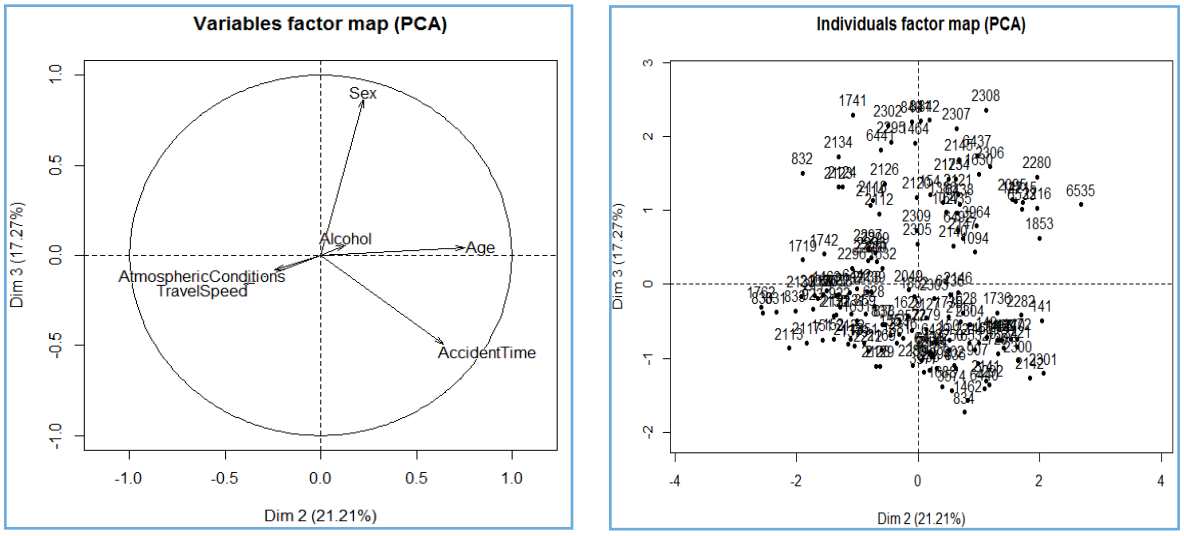

Figure 6.2: The graphic representations of the variables and the individuals (axis (2, 3))

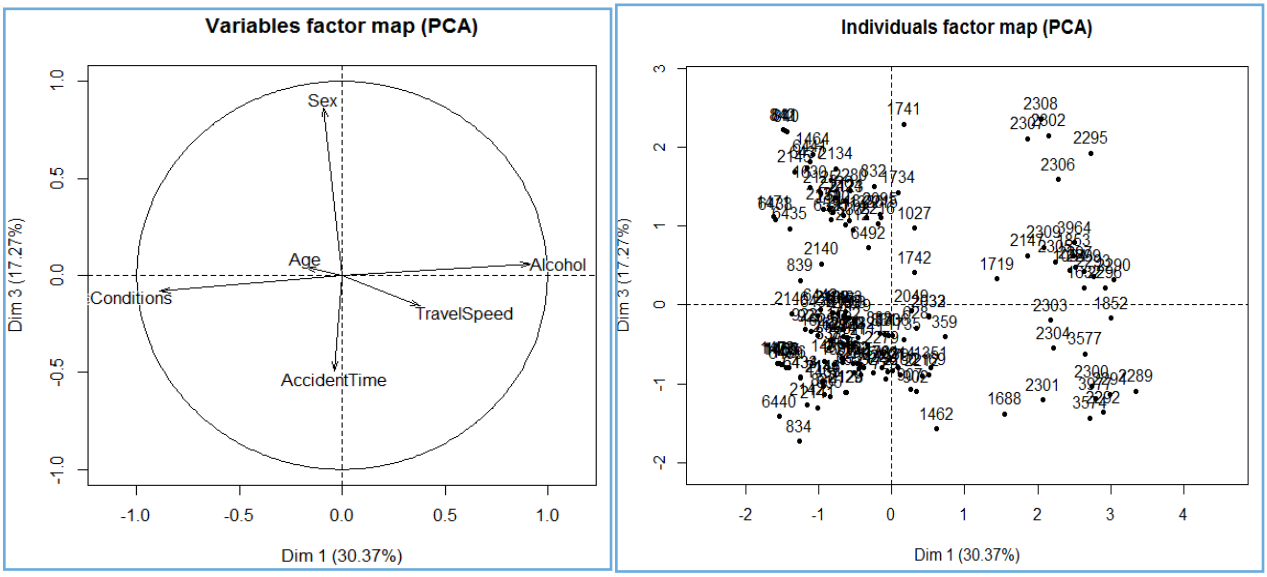

Figure 6.3: The graphic representations of the variables and the individuals (axis (1, 3)

The variable's and individual's representations (axis (2 3)):

- High and positive correlation between atmospheric conditions and travel speed.
- The variable alcohol still badly represented in this factorial plan.
- The variables age, accident time and sex are weakly and positively correlated.

The variable's and individuals representations (axis (1; 3)):

- The variables age, sex, accident time, travel speed, alcohol and conditions are well dispersed according to the second factorial plan (axis (1, 3)).
- The individuals with high percent alcohol have an accident in bad atmospheric conditions.
- The individuals with high percent of alcohol coincide with a high travel speed.
- The male individuals are more prone to accidents.

6.3 HIERARCHICAL CLUSTERING

- **Principle**

The principle is to seek the two closest points according to the distance considered and one gathers them in a cluster. The points are replaced thereafter by their center. Then one seeks the closest points or clusters again to gather them into only one cluster, and this in an iterative way. We used for the calculation of the distances a method called Euclidean distance. This iterative procedure is stopped when one does not have any more but one cluster.

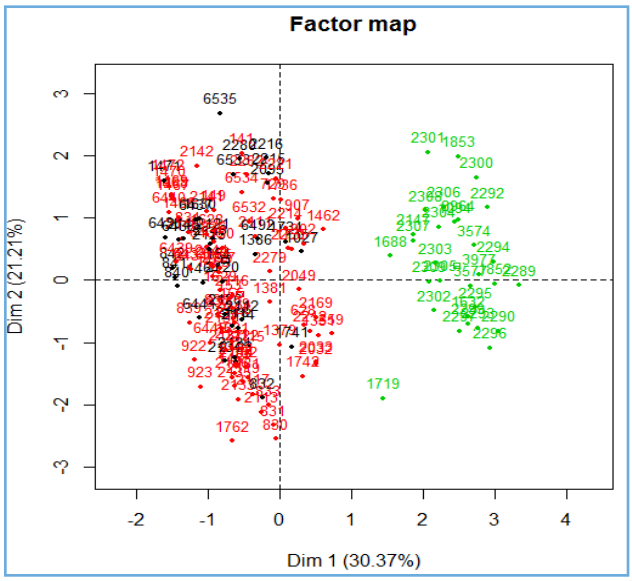

Figure 6.4: The hierarchical clustering

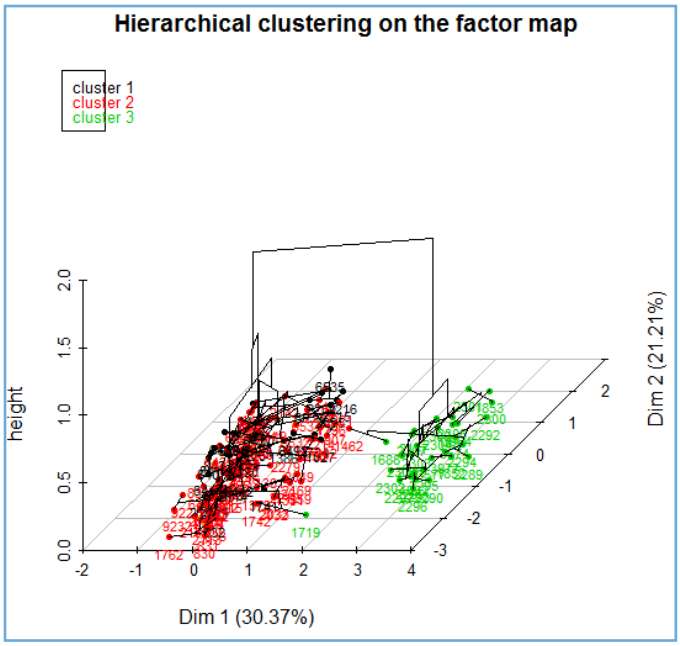

Figure 6.5: The 3D hierarchical clustering

- **Application**

We use function HCPC in the FactoMineR package, the function makes it possible to establish hierarchical clustering or a hierarchical classification of the results obtained during the analysis carried out in principal components and provides important results. It makes it possible to obtain compact clusters in spherical form (the distance considered between two individuals is the Euclidean distance). The results are shown in figure 6.4 and 6.5.

CHAPTER 7: RESULTS AND COMPARISONS

This chapter summarizes the results from each process based on the analysis conducted in the previous chapters.

7.1 CLUSTER ANALYSIS

All variables were standardized to mean 0 and variance 1.The data set was subdivided into 3 clusters, based on the atmospheric conditions. Based on this analysis method, the atmospheric conditions were grouped under three visibility conditions - poor, average and good visibilities. The poor visibility group consisted of snow, sleet and fog/smog/smoke conditions; average visibility consisted of rain and light snowy conditions while good visibility consisted of clear/cloudy condition (no adverse condition). The results suggest that in case of such atmospheric conditions, when we have blowing snow or fog or sleet - people in higher age group tend to lose control of their vehicle more often and hence have higher speed and accident rates. This group of people has lower blood alcohol content, and hence, the external factor contributing to such incidences is the atmospheric condition.

Based on ANOVA, we see from the figures 5.12 a - e, that changing pattern of variables with atmospheric conditions the number of incidence of accidents are less in case of good visibility and lower age, as compared to poor visibility/higher age. Sober drivers tend to overcome driving conditions (poor/average), which consist of low snow and rain. But as the alcohol content goes up, the number of incidence goes up, even in cases of good visibility. Drivers at higher speed tend to have more accidents than those at lower speed.

Based on atmospheric conditions, it is evident that the accident rate decreases with better visibility. Comparison of day of the week and visibility suggests that most accidents took place on/around weekend (Friday), even though the visibility was not poor, but not good.

7.2 MULTIVARIATE REGRESSION ANALYSIS

Based on this analysis, we see that, time of the day and atmospheric conditions have the highest significance on incidence of accident. These are followed by alcohol content, age of driver and lastly day of the week. This result, based on experience, seems pretty correct as the traffic load (yet another parameter, not analyzed in this thesis) is higher on weekends and adverse atmospheric conditions have led to higher incidences of accidents.

Further grouping the database by sex to decide which group is safer, it is interesting to note that male drivers outnumber in accidents that their female counterparts, as the regression tend towards a positive trend for male than female. Male drivers, it seems are more prone to drive drunk on high traffic days. The time of accident has less partial effect for both groups.

However, we also see that incidence among the female group is more significant when we use other predictors like accident time, day of the week, age, atmospheric conditions and alcohol. Regression analysis of both groups also shows that atmospheric condition has higher effect on the number of incidence for both groups. For males, time of the day is bigger killer. For female group, alcohol content and age are the next significant factors.

7.3 HIERARCHICAL MULTIPLE REGRESSION

We decided to use age and alcohol as the priority in hierarchy, based on above results which all show that alcohol have significant effect on the number of incidences. The results obtained with this method suggest, that alcohol, age and atmospheric condition are more significant factors in

contributing to the number accidents, which were caused by high travel speed. Among these three, atmospheric condition plays a bigger role, with high value of significance.

The partial effects of all the factors being used here on the dependent variable, travel speed, suggests irrespective of number of factors used, atmospheric condition continues to be a deciding factor.

7.4 FACTOR ANALYSIS

This test provides us a way to measure the consistency of the dataset being used here. We find that our dataset has no multicollinearity issue as the determinant is greater than 0.00001. Based on the sphericity test, we find that there are inherent relationships between the variables, with Bartlett's test being highly significant. We have accepted the dataset, without any further modification, even though the KMO test value is at 0.457. This value could have been improved by incorporating more data. Since, the available data is huge and the data selection is random, it is difficult to predict, if incorporating more data will be beneficial or not.

The factor analysis produces rotated results, where the factors have been rotated and show the effect of each variable. We see that most variables were loaded highly on the first variable and remaining factors were less looked into. However, after rotation, we see that there are four factors and variables load very highly onto the first factor.

Based on this analysis, we frame certain scenarios as stated earlier and the method is not able to confirm if any of those scenarios hold true.

7.5 DISCRIMINANT ANALYSIS

This analysis suggests that accident time, and age of the driver may be good discriminator as the separations are large. The results obtained in this method suggest that our data holds good and

hence, is adequate for analysis of the hypothesis stated above, as there are significant difference between male and female group for all independents.

CHAPTER 8: CONCLUSIONS

Based on our analysis of the data from 1999 to 2009, for all counties in Nevada using different methods we come to following conclusions:

- Adverse atmospheric conditions has more effect and is probably bigger causal factor in accidents than other factors.
- Alcohol content is next contributing factor as sober drivers tend to overcome poor to average atmospheric conditions.
- Accident rate decreases with visibility
- Accident rates increased as the week came to an end.
- Male drivers outnumber in accident incidences than their female counterparts.
- For male drivers, time of the day has more effect.
- For Female drivers alcohol content and age are the next significant factors after the time of the day.

We have also tested our data for reliability and we can say the following about our data:

- Data being used has no multicollinearity issue.
- Our assumption that the various factors being used in this research has some kind of effect on travel speed is correct based on sphericity test.
- KMO test value is lower than 0.5, which is the general acceptance level for data. However, it is very close to 0.5, and hence we have continued with the dataset without any alteration. A possible future work could be incorporating more data and testing it initially so that the KMO value is significant, before proceeding with analysis.

APPENDIX – I

APPENDIX – II

R Code:

« Loadings of the two packages: FactoMineR and psych »

```
library (FactoMineR)
library (psych)
```

« Data import: format .csv »

```
data=read.table("C:/Users/Admin/Desktop/ThesisDataCSV", sep=";", dec=",",  header=TRUE,
row.names=1)
```

« Bartlett test of sphericity and KMO index»

```
Bartett.test(data)
KMO(data)
```

« Principal component Analysis »

```
PCA1<- PCA(data)
PCA$eig : to extract the eigen values of each component
PCA"$var:  results for the variables"
PCA"$var$coord: coord. for the variables"
PCA"$var$cor : correlations variables - dimensions
PCA"$var$cos2: cos2 for the variables"
PCA"$var$contrib: contributions of the variables"
PCA"$ind: results for the individuals"
PCA"$ind$coord: coord. for the individuals"
PCA"$ind$cos2: cos2 for the individuals"
PCA"$ind$contrib: contributions of the individuals"
PCA"$call: summary statistics"
PCA"$call$centre : mean of the variables"
PCA"$call$ecart.type "standard error of the variables"
PCA"$call$row.w      "weights for the individuals"
PCA"$call$col.w      "weights for the variables"
```

« Hierarchical Clustering based on the PCA's results »

Clustering<- HCPC(ACP)

Clustering $data.clust

Clustering $desc.var

Clustering $desc.var$quanti.var

Clustering $desc.var$quanti

Clustering $desc.axes

Clustering $desc.axes$quanti.var

Clustering $desc.axes$quanti

Clustering $desc.ind

Clustering $desc.ind$para

Clustering $desc.ind$dist

Clustering $call

Clustering $call$t

APPENDIX - III

Table 5.1: Independent T-test output

Independent Samples Test

		Equality of Variances		t-test for Equality of Means						
									Interval of the	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Age	Equal variances assumed	17.081	.000	.861	147	.390	3.35478	3.89415	-4.34096	11.05053
	Equal variances not assumed			1.261	80.704	.211	3.35478	2.66042	-1.93892	8.64849
Alcohol	Equal variances assumed	10.377	.002	-48.281	147	.000	-34.44362	.71340	-35.85348	-33.03377
	Equal variances not assumed			-44.753	37.423	.000	-34.44362	.76963	-36.00246	-32.88479
TravelSpeed	Equal variances assumed	2.316	.130	-2.096	147	.038	-10.44156	4.98240	-20.28794	-.59518
	Equal variances not assumed			-1.971	37.951	.056	-10.44156	5.29678	-21.16478	.28167
Atmospheric Conditions	Equal variances assumed	80.085	.000	13.882	147	.000	2.52066	.18157	2.16183	2.87949
	Equal variances not assumed			28.934	120.000	.000	2.52066	.08712	2.34818	2.69315
DayofWeek	Equal variances assumed	.361	.549	3.037	147	.003	1.24911	.41124	.43640	2.06182
	Equal variances not assumed			2.982	39.650	.005	1.24911	.41884	.40238	2.09585

Table 5.2: Model summary multivariate regression

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.319[a]	.102	.070	23.17124

a. Predictors: (Constant), Accidenttime, Alcohol, Age, DayofWeek, AtmosphericConditions

b. Dependent Variable: TravelSpeed

Table 5.3: ANOVA for multivariate regression

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8677.189	5	1735.438	3.232	.009[b]
	Residual	76777.590	143	536.906		
	Total	85454.779	148			

a. Dependent Variable: TravelSpeed

b. Predictors: (Constant), Accidenttime, Alcohol, Age, DayofWeek, AtmosphericConditions

Table 5.4: Coefficient estimate for multivariable regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	74.737	12.320		6.066	.000			
	Age	-.165	.106	-.127	-1.561	.121	-.152	-.129	-.124
	Alcohol	.162	.225	.094	.722	.472	.195	.060	.057
	AtmosphericConditions	-.969	2.318	-.053	-.418	.676	-.114	-.035	-.033
	DayofWeek	-2.682	1.028	-.225	-2.608	.010	-.253	-.213	-.207
	Accidenttime	-.001	.003	-.017	-.207	.836	-.056	-.017	-.016

a. Dependent Variable: TravelSpeed

Table 5.5: Multivariate regression for data split by sex of drivers

Model Summary[b]

Sex		R	R Square	Adjusted R Square	Std. Error of the Estimate
Male	1	.337[a]	.114	.069	23.82395
Female	1	.475[c]	.226	.121	21.03302

a. Predictors: (Constant), Accidenttime, AtmosphericConditions, DayofWeek, Age, Alcohol

b. Dependent Variable: TravelSpeed

c. Predictors: (Constant), Accidenttime, DayofWeek, AtmosphericConditions, Age, Alcohol

Table 5.6: ANOVA for Multivariate regression for data split by sex of drivers

ANOVA[a]

Sex			Sum of Squares	df	Mean Square	F	Sig.
Male	1	Regression	7286.164	5	1457.233	2.567	.031[b]
		Residual	56758.072	100	567.581		
		Total	64044.236	105			
Female	1	Regression	4776.807	5	955.361	2.160	.080[c]
		Residual	16368.356	37	442.388		
		Total	21145.163	42			

a. Dependent Variable: TravelSpeed

b. Predictors: (Constant), Accidenttime, AtmosphericConditions, DayofWeek, Age, Alcohol

c. Predictors: (Constant), Accidenttime, DayofWeek, AtmosphericConditions, Age, Alcohol

Table 5.7: Coefficients of multivariate regression for data split by sex of drivers

Coefficients[a]

Sex			Unstandardized Coefficients		Standardized Coefficients			Correlations		
			B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
Male	1	(Constant)	71.888	14.638		4.911	.000			
		Age	-.169	.131	-.125	-1.288	.201	-.158	-.128	-.121
		Alcohol	.338	.265	.193	1.273	.206	.235	.126	.120
		AtmosphericConditions	.082	2.735	.004	.030	.976	-.136	.003	.003
		DayofWeek	-1.855	1.291	-.148	-1.436	.154	-.208	-.142	-.135
		Accidenttime	-.005	.004	-.113	-1.164	.247	-.164	-.116	-.110
Female	1	(Constant)	73.164	23.445		3.121	.003			
		Age	-.145	.178	-.125	-.813	.422	-.122	-.132	-.118
		Alcohol	-.270	.425	-.162	-.634	.530	.076	-.104	-.092
		AtmosphericConditions	-2.458	4.422	-.140	-.556	.582	-.050	-.091	-.080
		DayofWeek	-4.059	1.667	-.386	-2.436	.020	-.381	-.372	-.352
		Accidenttime	.010	.006	.260	1.714	.095	.265	.271	.248

a. Dependent Variable: TravelSpeed

Table 5.8: Model summary for hierarchical multivariate linear regression

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.238[a]	.057	.044	23.49587
2	.241[b]	.058	.039	23.56061
3	.318[c]	.101	.076	23.09410
4	.319[d]	.102	.070	23.17124

a. Predictors: (Constant), Alcohol, Age

b. Predictors: (Constant), Alcohol, Age, AtmosphericConditions

c. Predictors: (Constant), Alcohol, Age, AtmosphericConditions, DayofWeek

d. Predictors: (Constant), Alcohol, Age, AtmosphericConditions, DayofWeek, Accidenttime

e. Dependent Variable: TravelSpeed

Table 5.9: ANOVA for hierarchical multivariate linear regression

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4854.633	2	2427.317	4.397	.014[b]
	Residual	80600.145	146	552.056		
	Total	85454.779	148			
2	Regression	4964.960	3	1654.987	2.981	.033[c]
	Residual	80489.818	145	555.102		
	Total	85454.779	148			
3	Regression	8654.193	4	2163.548	4.057	.004[d]
	Residual	76800.585	144	533.337		
	Total	85454.779	148			
4	Regression	8677.189	5	1735.438	3.232	.009[e]
	Residual	76777.590	143	536.906		
	Total	85454.779	148			

a. Dependent Variable: TravelSpeed

b. Predictors: (Constant), Alcohol, Age

c. Predictors: (Constant), Alcohol, Age, AtmosphericConditions

d. Predictors: (Constant), Alcohol, Age, AtmosphericConditions, DayofWeek

e. Predictors: (Constant), Alcohol, Age, AtmosphericConditions, DayofWeek, Accidenttime

Table 5.10: Coefficients of hierarchical multivariate linear regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	58.576	5.003		11.708	.000			
	Age	-.178	.104	-.137	-1.704	.090	-.152	-.140	-.137
	Alcohol	.318	.139	.184	2.286	.024	.195	.186	.184
2	(Constant)	54.882	9.687		5.666	.000			
	Age	-.174	.105	-.134	-1.655	.100	-.152	-.136	-.133
	Alcohol	.389	.211	.225	1.843	.067	.195	.151	.149
	AtmosphericConditions	.994	2.231	.054	.446	.656	-.114	.037	.036
3	(Constant)	74.201	12.005		6.181	.000			
	Age	-.169	.103	-.131	-1.643	.102	-.152	-.136	-.130
	Alcohol	.162	.224	.094	.722	.471	.195	.060	.057
	AtmosphericConditions	-.964	2.310	-.053	-.417	.677	-.114	-.035	-.033
	DayofWeek	-2.692	1.024	-.226	-2.630	.009	-.253	-.214	-.208
4	(Constant)	74.737	12.320		6.066	.000			
	Age	-.165	.106	-.127	-1.561	.121	-.152	-.129	-.124
	Alcohol	.162	.225	.094	.722	.472	.195	.060	.057
	AtmosphericConditions	-.969	2.318	-.053	-.418	.676	-.114	-.035	-.033
	DayofWeek	-2.682	1.028	-.225	-2.608	.010	-.253	-.213	-.207
	Accidenttime	-.001	.003	-.017	-.207	.836	-.056	-.017	-.016

a. Dependent Variable: TravelSpeed

Table 5.11: Correlation matrix for factor analysis

Correlation Matrix[a]

		Age	Alcohol	TravelSpeed	Sex	AtmosphericConditions	DayofWeek	Accidenttime
Correlation	Age	1.000	-.078	-.152	.097	.002	.061	.207
	Alcohol	-.078	1.000	.195	-.030	-.747	-.229	-.002
	TravelSpeed	-.152	.195	1.000	-.056	-.114	-.253	-.056
	Sex	.097	-.030	-.056	1.000	.011	-.037	-.043
	AtmosphericConditions	.002	-.747	-.114	.011	1.000	-.039	-.028
	DayofWeek	.061	-.229	-.253	-.037	-.039	1.000	.061
	Accidenttime	.207	-.002	-.056	-.043	-.028	.061	1.000
Sig. (1-tailed)	Age		.172	.032	.120	.493	.231	.006
	Alcohol	.172		.009	.360	.000	.002	.489
	TravelSpeed	.032	.009		.250	.083	.001	.251
	Sex	.120	.360	.250		.447	.326	.300
	AtmosphericConditions	.493	.000	.083	.447		.320	.368
	DayofWeek	.231	.002	.001	.326	.320		.231
	Accidenttime	.006	.489	.251	.300	.368	.231	

a. Determinant = .310

Table 5.12: Kaiser-Meyer-Olkin test of sampling adequacy and Bartlett test of sphericity

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.457
Bartlett's Test of Sphericity	Approx. Chi-Square	169.610
	df	21
	Sig.	.000

Table 5.13: Factor extraction – total variance explained

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.878	26.828	26.828	1.878	26.828	26.828	1.748	24.974	24.974
2	1.333	19.049	45.877	1.333	19.049	45.877	1.295	18.502	43.476
3	1.073	15.332	61.209	1.073	15.332	61.209	1.209	17.269	60.745
4	1.028	14.683	75.892	1.028	14.683	75.892	1.060	15.147	75.892
5	.763	10.899	86.791						
6	.720	10.291	97.083						
7	.204	2.917	100.000						

Extraction Method: Principal Component Analysis.

Table 5.14: Factor analysis - communalities

Communalities

	Initial	Extraction
Age	1.000	.637
Alcohol	1.000	.880
TravelSpeed	1.000	.577
Sex	1.000	.854
AtmosphericConditions	1.000	.905
DayofWeek	1.000	.737
Accidenttime	1.000	.723

Extraction Method: Principal Component Analysis.

Table 5.15: Component matrix identifying common themes

Component Matrix[a]

	Component			
	1	2	3	4
Alcohol	-.909			
AtmosphericConditions	.817	-.423		
Age		.612	.442	
TravelSpeed	-.455	-.457		
Sex			.666	-.628
DayofWeek		.448	-.586	
Accidenttime		.559		.616

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Table 5.16: Discriminant analysis – group statistics

Group Statistics

Sex		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
Male	Age	41.3679	18.19541	106	106.000
	Alcohol	7.2547	14.13679	106	106.000
	TravelSpeed	54.0849	24.69707	106	106.000
	AtmosphericConditions	3.0377	1.33041	106	106.000
	DayofWeek	4.4906	1.97241	106	106.000
	Accidenttime	1138.7075	615.65290	106	106.000
Female	Age	45.3256	19.33504	43	43.000
	Alcohol	6.3488	13.51169	43	43.000
	TravelSpeed	51.1395	22.43783	43	43.000
	AtmosphericConditions	3.0698	1.27979	43	43.000
	DayofWeek	4.3256	2.13498	43	43.000
	Accidenttime	1081.9302	555.74173	43	43.000
Total	Age	42.5101	18.55293	149	149.000
	Alcohol	6.9933	13.91989	149	149.000
	TravelSpeed	53.2349	24.02909	149	149.000
	AtmosphericConditions	3.0470	1.31177	149	149.000
	DayofWeek	4.4430	2.01475	149	149.000
	Accidenttime	1122.3221	597.67714	149	149.000

Table 5.17: Discriminant analysis – test of equality of means

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Age	.991	1.396	1	147	.239
Alcohol	.999	.129	1	147	.720
TravelSpeed	.997	.458	1	147	.500
AtmosphericConditions	1.000	.018	1	147	.893
DayofWeek	.999	.204	1	147	.652
Accidenttime	.998	.275	1	147	.601

Table 5.18: Discriminant analysis – pooled within group matrices

Pooled Within-Groups Matrices

		Age	Alcohol	TravelSpeed	AtmosphericConditions	DayofWeek	Accidenttime
Correlation	Age	1.000	-.076	-.147	.000	.065	.212
	Alcohol	-.076	1.000	.194	-.747	-.231	-.004
	TravelSpeed	-.147	.194	1.000	-.114	-.256	-.058
	AtmosphericConditions	.000	-.747	-.114	1.000	-.038	-.027
	DayofWeek	.065	-.231	-.256	-.038	1.000	.059
	Accidenttime	.212	-.004	-.058	-.027	.059	1.000

Table 5.19: Discriminant analysis – Log Determinants

Log Determinants

Sex	Rank	Log Determinant
Male	6	31.147
Female	6	30.477
Pooled within-groups	6	31.066

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Table 5.20: Discriminant analysis – Box's M and F-test results

Test Results

Box's M		16.226
F	Approx.	.729
	df1	21
	df2	25141.681
	Sig.	.808

Tests null hypothesis of equal population covariance matrices.

Table 6.1: Bartlett's test results

Bartlett's K-Squared = 6684.8	**df = 6**	**p-value < 2.2e-16**

Table 6.2: KMO values of the variables

Overall KMO=0.46							
	Age	**Alcohol**	**Travel Speed**	**Sex**	**Atmospheric Conditions**	**Day of Week**	**Accident Time**
KMO	0.52	0.47	0.67	0.43	0.45	0.31	0.52

Table 6.3: New KMO values of the variables

Overall KMO=0.51						
	Age	**Alcohol**	**Travel Speed**	**Sex**	**Atmospheric Conditions**	**Accident Time**
KMO	0.51	0.51	0.66	0.50	0.50	0.50

Table 6.4: The Eigen values

	Eigen Value	**Percentage**	**Cumulative Percentage**
1	1.8224535	30.374225	30.37423
2	1.2723827	21.206378	51.58060
3	1.0364056	17.273426	68.85403
4	0.8815111	14.691851	83.54588
5	0.7418298	12.363831	95.90971
6	0.2454173	4.090288	100.00000

Table 6.5: Coordinates of the variable considering each component

	Dim.1	**Dim.2**	**Dim.3**
Age	-0.18018188	0.7546710	0.04027645
Alcohol	0.91878481	0.1307247	0.06015485
Travel Speed	0.38322641	-0.4026263	-0.16167796
Sex	-0.09021345	0.2249386	0.86795497
Atmospheric Conditions	-0.88842947	-0.2425710	-0.08215345
Accident Time	-0.03891795	0.6435988	-0.49490404

Table 6.6: Quality representation of the variable in each component

	Dim.1	**Dim.2**	**Dim.3**
Age	-0.18018188	0.7546710	0.04027645
Alcohol	0.91878481	0.1307247	0.06015485
Travel Speed	0.38322641	-0.4026263	-0.16167796
Atmospheric Conditions	-0.88842947	-0.2425710	-0.08215345
Sex	-0.09021345	0.2249386	0.86795497
Accident Time	-0.03891795	0.6435988	-0.49490404

Table 6.7: The Clusters by Individuals

Individuals	Cluster	Individuals	Cluster	Individuals	Cluster	Individuals	Cluster
830	2	1379	2	2303	3	2125	1
831	2	1381	2	2304	3	2126	1
833	2	2127	2	2143	2	1734	1
359	2	1735	2	2144	2	6437	1
2290	3	1736	2	837	2	1630	1
3574	3	2297	3	838	2	6533	1
2289	3	1463	2	2305	3	6535	1
3577	3	6436	2	729	2	6438	1
2292	3	922	2	1465	2	2134	1
1762	2	923	2	6442	2	2215	1
3977	3	628	2	2309	3	2216	1
2293	3	2128	2	2146	2	1741	1
1462	2	2129	2	2282	2	1464	1
1688	3	2130	2	2279	2	2140	1
1632	3	2131	2	141	2	2095	1
2294	3	2132	2	839	2	2302	3
1852	3	1628	2	2147	3	2306	3
2296	3	1629	2	1466	2	2307	3
2113	2	1631	2	1467	2	2308	3
2115	2	6532	2	1468	2	6441	1
6433	2	6534	2	1469	2	2145	1
6434	2	2299	3	1470	2	2280	1
149	2	6439	2	1472	2	840	1
150	2	2133	2	2111	2	841	1
902	2	2135	2	832	1	842	1
907	2	2136	2	3964	3	1471	1
1719	3	2137	2	2112	1		
2116	2	2138	2	1027	1		
2169	2	2139	2	2295	3		
2117	2	2212	2	2114	1		
2122	2	2214	2	6435	1		
2119	2	2221	2	1094	3		
1351	2	835	2	2118	1		
834	2	836	2	2120	1		
151	2	1742	2	2121	1		
152	2	2300	3	6492	1		
153	2	6440	2	154	1		
155	2	2141	2	1853	3		
2032	2	2142	2	1386	1		
2033	2	2250	2	2123	1		
2049	2	2301	3	2124	1		

REFERENCES

1. Nevada department of public safety (2015). "Fatality Statistics" [Online]. Available: http://www.zerofatalitiesnv.com/stats-current-year/

2. M. Brown (2012, December 11). "Data Mining Techniques" [Online]. Available: http://www.ibm.com/developerworks/library/ba-data-mining-techniques/

3. V. Beal. "Big data" [Online]. Available: http://www.webopedia.com/TERM/B/big_data.html

4. Wikipedia, (2015), "Big data" [Online]. Available: https://en.wikipedia.org/wiki/Big_data

5. M. van Rijmenam (2015, January 14). "How big data can create smarter transporations industry" [Online]. Available: https://datafloq.com/read/big-data-create-smarter-transportation-industry/119

6. International transport forum, (2013, October). "Big data and transport" [Online]. Available: http://www.openskydata.com/assets/media/downloads/BigDataOct2013.pdf

7. H. M. Chung and P. Gray, "Special Section: Data Mining", Journal of Management Information Systems, vol. 16, pp. 11-17, 1999.

8. U. Fayyad et al., "From data mining to knowledge discovery in databases", AI Magazine, vol. 17, pp. 37-54, 1996.

9. P. R. Peacock, "Data mining in marketing: Part 1", Marketing Management, pp. 9-18, 1998.

10. J. Han and M. Kamber, "Data mining: concepts and techniques", Morgan-Kaufmann Academic Press, San Francisco, 2001.

11. D. J. Hand, "Data mining: statistics and more?", The American Statistician, vol. 52, pp. 112-118, May 1998.

12. G. E. P. Box and W. G. HUNTER, "Sequential design of experiments for non linear models", in Proc. IBM Scientific Computing Symposium on Statistics, New York, pp. 113-137, 1965.

13. U. Fayyad et al., "The KDD process for extracting useful knowledge from volumes of data," Communications of the ACM, vol. 39, pp. 27-34, 1996.

14. B. Rajagopalan and R. Krovi, "Benchmarking data mining algorithms", Journal of Database Management, vol. 13, pp. 25-36, Jan-Mar, 2002.

15. P. Gray and H. J. Watson, "Professional Briefings...Present and future directions in data warehousing", Database for Adv. in Info. Sys., vol. 29, pp. 83-90, 1998.

16. H. White, "A Reality Check for Data Snooping", Econometrica, vol. 68, pp. 1097-1126, 2000.

17. C. Glymour et al., "Statistical Inference and Data Mining". Communications of the ACM, vol. 39, pp. 35-41, 1996.

18. M. J. Berry and G. S. Linoff, "Mastering Data Mining: The art and science of customer relationship management". Wiley Computer Publishing, New York, 2000.

19. R. Johnson and D. W. Wicheren, "Applied Multivariate Statistical Analysis". Prentice Hall, New York, 1998.

20. D. J. Tegarden, "Business Information Visualization". Communications of AIS, vol. 1, 1999.

21. Wikipedia, (2015). "Machine Learning" [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning

Curriculum Vitae

VISHAL JHA

37 103rd Ave NE, Apt 217 Bellevue 98004, WA
Phone: 7753035872
Email: vshljha@gmail.com

EDUCATIONAL QUALIFICATIONS

- B.E (Hons.) Mechanical Engineering, 2001-2005, BITS, Pilani. 6.3/10.0 GPA
- M.S. Electrical Engineering (pursuing), 2011-2015, University of Nevada Las Vegas (UNLV). 3.4/4.0 GPA. My thesis is based on performing multivariate analysis on traffic data using statistical software SPSS and R.

PROFESSIONAL EXPERIENCE

ETL Developer, EBAY, BELLEVUE, WA **Feb 2015 – till date**

- Involved in STEP (Ebay-Paypal Separation) project.
- Using Teradata 14.10.04.13b version.
- Developing Unix Shell Scripts for automation of ETL Processes, Error handling and Auditing.
- Extract data from various source systems like Oracle and flat files (e.g. .txt file, .dat file, .csv file) as per the requirements and loading into another flat file using different ETL logic.
- Loading data from various data sources and legacy systems into Teradata production and development warehouse using BTEQ, FASTEXPORT, MULTI LOAD and FASTLOAD.
- Writing queries using SQL.
- Performance tuning, including collecting statistics, analyzing and determining which tables needed statistics. Increased performance by 35-40% in some situations.
- Monitoring and scheduling jobs using UC4 scheduler tool.
- Unit testing of the developed workflows.
- Creating BTEQ scripts to load data from Teradata stage table to working table and from working table to final tables.
- Release engineering process (Production Deployment).
- Ensuring smooth running of the application and facilitating improvisation as and when required.
- Respond to the requests received from clients about enhancements/modifications in the environment.
- Monitor the changes carried out in the application as a part of maintenance.
- Provide QA support.

- Data coming from different types of flat file and presenting into another flat file using different ETL logic.

Business Systems Analyst, WELLS FARGO ADVISORS, ST. LOUIS, MO Sep 2013 – Jan 2014

- Provide technical knowledge to business partners and drive requirements based on business process. Support enhancements to the New Account Opening and Account Maintenance Applications.
- Develop clear and detailed Functional System Design (FSD) for technical teams. Partner with customer and account services teams to enable data exchange using Service-oriented Architecture (SOA).
- Help the business come up with the Business Requirement Document (BRD), providing estimates to the IT manager(s).
- Create High Level Test Scenarios (HLTS), partner with Quality Assurance resources, and provide support for defect identification and resolution. High level test scenarios spreadsheet covers all the functional test scenarios for the project. It calls out the negative scenarios, regression and in-flight scenarios that are required for the project.
- Coordinate with developers to validate the code changes in DEV.
- Work with the testing team closely to set up the inflight tests before the code is deployed to the test environment.
- Update Documentum with changes (if any) to the FSD, Appendices and Module specifications after the testing phase ends.

Graduate Research Assistant, UNIVERSITY OF NEVADA, LAS VEGAS Jan 2011 – Jan 2013

- Working for Transportation Research Center at UNLV.
- Ongoing research on Analysis Traffic Flow based on Sensor Data for various Construction Events. Thesis: Data Mining On Transportation Data Using Multivariate Statistical Analysis

Research Analyst, FROST & SULLIVAN, CHENNAI Apr 2010 – Dec 2010

- Monitoring technology market, and providing insightful quantitative and strategic analysis to clients through market sizing reports.
- Conducting quantitative and qualitative business research for Industrial Automation and Process Control Domain.
- Generating reports on the North American Motors and Drives market.
- Undertaking B2B research, analyze industry trends and data, and generate client-focused reports, briefings, presentations and other content

Research Analyst, CROSS-TAB MARKETING SERVICES, BANGALORE Jun 2007 – Feb 2009

- Conducting primary and secondary research on innovative and evolving businesses, markets and demographics prominently in the Software and Telecom Domain
- Questionnaire designing and Blog mining
- Documenting reports on the basis of various web researches. The report includes introduction about the business, market, competitors, prerequisites, execution plan, benefits, risks etc.
- Experience of using various Software Databases like IDC, Gartner , Forrester & others

Research Analyst, MANTHAN SERVICES, BANGALORE Oct 2006 – May 2007

- Conducted extensive Primary and Secondary research
- Conducted Quantitative and Qualitative business research
- Handled databases such as Factiva and Trial Trove
- Conducted competitive intelligence studies for global clients
- Assisted in the final presentation of project reports

Software Engineer, INFOSYS TECHNOLOGIES LTD., BANGALORE Aug 2005 – Sep 2006

I was a part of two projects namely Sourcing Workbench and Delhaize Vendor Portal with Sony Ericsson (Sweden) and Delhaize (Belgium**)** as our clients respectively. My main responsibilities included – coding and developing web applications using ASP.Net, SQL and Oracle.

Trainee, GENERAL MOTORS, INDIA

(BITS Practice School II Training) Jan 2005 – Jul 2005 Title of Project – Lower Floor Project on Chevrolet Tavera

As a part of the Product Engineering Department, I handled the lower floor project on the Chevrolet Tavera vehicle.

SKILL SET

- **Market Research:** Secondary Research (Quantitative & Qualitative), Primary Research (Qualitative), Competitive Analysis, Market Sizing, Modeling and Forecasting.
- **Business Research:** Requirement Gathering, Functional Specification Documentation, Use Case Development.
- **Languages:** C, C#, ASP.NET, SQL
- **Software Tools**: R, SPSS, Teradata
- **Database:** IDC, Gartner, Forrester, Factiva, Trial Trove, LexisNexis.

- Excellent Oral and written communication skills
- Proficient with Microsoft Office Suite (Word, PowerPoint and Excel)

ACHIEVEMENTS

- Gold medal, UNLV Badminton doubles, UNLV Intra College Competition.
- Silver medal, UNLV Badminton singles, UNLV Intra College Competition.
- General Secretary, Indian Students Association, UNLV
- Silver medal, Hockey team event in BITS Open Sports Meet (BOSM) 2003-04 & 2004-05.
- Member, Mechanical Engineering Association and Maurya Vihar in BITS, Pilani.
- Winner of many quiz competitions during school and college days.