Trend and Prediction Analysis for Enhanced Road Safety

By:

Sumaira Shamim	**2009-NUST-BE-BICSE-212**
Muneza Naeem	**2009-NUST-BE-BICSE-198**
Hamza Shahzad	**2009-NUST-BE-BICSE-153**

A Project Report submitted in partial fulfillment of the requirement for the degree

of

Bachelors in Information and Communication Systems Engineering

School of Electrical Engineering & Computer Science

National University of Sciences & Technology

Islamabad, Pakistan

(May 2013)

CERTIFCATE OF APPROVAL

It is certified that the contents and form of Final Report entitled **“Trend and Prediction Analysis for Enhanced Road Safety”** submitted by *Sumaira Shamim* **(2009-NUST-BE-BICSE-212),** *Hamza Shahzad* **(2009-NUST-BE-BICSE-153)** and *Muneza Naeem* **(2009-NUST-BE-BICSE-198)** has been found satisfactory for the requirement of the degree.

Advisor: Dr. Sharifullah Khan

Signature: ___________________________

Co-Advisor: Dr. Ali Mustafa Qamar

Signature: ____________________________

Co-Advisor: Dr. Usman Younis

Signature: ____________________________

DEDICATION

To Allah Almighty

&

To our Parents and Faculty

ACKNOWLEDGEMENTS

We are extremely thankful to Allah for the timely completion of our project. We would also like to thank our family whose support and prayers led us through hard times.

We are also grateful to Dr. Sharifullah Khan, Dr. Ali Mustafa Qamar and Dr. Usman Younis for their support and guidance.

We would also offer our gratitude to our respected seniors Mr.Ahsan Rehman, Mr Zaid Bin Asif and Mr.Timsal Ahmed who worked tirelessly to help us with the milestones in our project. We would also like to specially thank Dr. Sarah Shafiq Khan and Mr. Moin-ud-Din for their valuable guidance in a very crucial time of our project. Our class fellows Wajahatullah Khan, Ammar Saleem and Saad Ahmed Qureshi also deserve a mention here for their selfless help in various disciplines.

We are extremely thankful to AIG Mr. Farooq Azam of National Highways and Motorway Police and SP Sajid Kiyani of Rescue 15 for making sure we were provided with the latest auto accidents data of Pakistan.

We would take this opportunity to thank a person whose presence was crucial to the technical hardships we faced in IBM Competency Centre, The respectable lab Incharge Mr Jamshed Khushi.

Table of Contents

List of Abbreviations:

- NH & MP: National Highways and Motorways Police
- ITP: Islamabad Traffic Police
- BI: Business Intelligence
- CSV: Comma Separated Variables
- WHO: World Health Organization
- OLAP: Online Analytical Processing

List of Figures:

List of Tables:

ABSTRACT

Most organizations are realizing the value of data; and Business Intelligence (BI) is the key to analyze data in today's information world. BI is used to investigate real world scenarios to identify potential problems. One such scenario is the ever increasing rate of auto-mobile accidents in Pakistan. Most of the countries are gaining benefits by analyzing the past road accidents data and forecasting a lot of information about possible future accidents. But unfortunately in Pakistan, data is collected to a limited extent, and the standard format of collecting data is not truly followed. There is no central department for accident prevention neither is there any data repository. Hence the number of accidents is steadily on the rise.

The aim of this project is to improve road safety and reduce accidents by identifying the root causes of these accidents. In order to achieve this goal, we have analyzed the traffic accidents data of Pakistan's Motor-ways and National highways. The data was collected from National Highways and Motorway Police, cleaned and modeled on different dimensions for analysis and prediction. The major recommendations concluded from the results are as follows.

1. Sunder (near Lahore) had the most number of accidents due to non-availability of pedestrian overhead bridge; and over speeding of drivers. A solution is to install overhead bridge and to have police patrol the area round the clock to keep drivers in check.
2. Trucks and trollies are involved in the most *dozing at the wheel* accidents; hence we need to plant road bumps on locations such as Sahiwal, where these accidents are more prominent.
3. Motorcycle riders have the most frequent fatal accidents among all vehicle types; the reason being disregard of safety by not wearing the helmet. A solution is to impose heavy fines on driving without helmet.

Chapter 1

INTRODUCTION

This chapter gives an overview of our project and explains the motivation, scope, problem statement and project domain.

Defining the Domain: Business Intelligence

The domain of the project is Data warehousing and Business intelligence.

> *"**Data warehouse** is a system that retrieves and consolidates data periodically from the source systems into a dimensional or normalized data store. It usually keeps years of history and is queried for business intelligence or other analytical activities*". [1]

> ***"Business intelligence** (**BI**) is a set of theories, methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information for business purposes. BI can handle large amounts of information to help identify and develop new opportunities. Making use of new opportunities and implementing an effective strategy can provide a competitive market advantage and long-term stability".* [2]

Data warehousing helps us to provide information on the techniques involved in designing, building, maintaining and retrieving information, from a data warehouse. [3] [4]

Business Intelligence is a method that extracts useful and meaningful information from the raw data which is used to enable more effective strategic, calculated, and operational insights and assist in decision-making. [2] [5]

Often data warehouse or data mart are basis to extract information using BI. The aim is to define and specify useful management reports from warehouse data, which will be used to make fast, accurate and intelligent business decisions. [5]

Problem Statement:

Traffic Accidents is the cause of a considerable number of casualities and injuries to such an extent that now the issue of Road security is being raised internationally. Accidents are being recorded according to the World Health Organizatin (WHO) standard internationally, which unluckily cannot be implemented in Pakistan as recording of accidents data is not followed comprehensively. So the need for devising intelligent system for road security by prediction analysis of traffic accidents in Pakistan has become more prevalent.

Thus, the problems statement of the project is:

> *"Reduce the severity i.e. fatalities caused by traffic accidents by analyzing the trends of auto-accidents in Pakistan".*

Project Motivation

- *BI and Dataware housing – A technology in Trend:*

  With the increasing data, there is a greater trend in making data warehouses and BI using them for especially in the area of predictive analytics, in which data is analyzed to predict future trends.

- *Traffic status in Pakistan:*

  In Pakistan traffic is increasing annually because the public transport available to general public is grossly insufficient. Hence the only outcome is for people to buy their own vehicles, which is leading to an alarming increase in traffic. Hence the need to better control traffic.

- *Lack of any concrete work on Road safety in Pakistan:*

  There has been some work done in individual capacity before to address the issue of road safety in Pakistan but nothing concrete has been achieved. Hence the motivation for us to explore this field and possibly come up with a solution that will benefit Pakistan and its people.

Project Description

The project focuses on the trend and prediction analysis of traffic accidents in Pakistan for which Traffic Accidents Data is collected from National Highway and Motorways Authority.

Central database of accidents is created using IBM DB2, which is acting as the primary data source for metadata modeling and reporting. Prior to creating the database, data cleaning was performed using Google Refine.

Data Modeling and Reporting was performed using IBM Cognos Framework Manager and IBM Cognos Report Studio. The aim is to provide better strategy making for traffic management. The analysis will provide for information regarding trends in traffic accidents and aid in reducing the severity and occurrences of auto-accidents in Pakistan.

Areas that are identified for having higher number of accidents are ranked as red zones. Those areas are marked on Pakistan Map using R – Language.

Further, using the same accidents data from National Highways and Motorways, a prediction model is made using IBM SPSS to predict the severity of accident cases on National Highways and Motorways.

Aims and Objectives

Overall project has been divided into several milestones over the timeline, which are listed below:

- Trend Analysis:

Identify the trends of traffic accidents on National Highways and Motorways.

- Predictions Analysis:

Predicting the severity of accidents and ranking them as fatal and non-fatal.

- Efficient access of information through dashboards:

Customized dashboards created using the reported trends provide ease of access of information.

- Mapping of locations of accidents on maps:

Zones having maximum number of accidents are identified so that steps can be taken for enhanced road security.

Project Plan

The project has been divided into different phases. The phases are incremental and sub-tasks can be dependent on each other. There are comprehensive requirements of setting up servers of IBM COGNOS BI and IBM SPSS that have been catered to meet and support the overall project plan

.

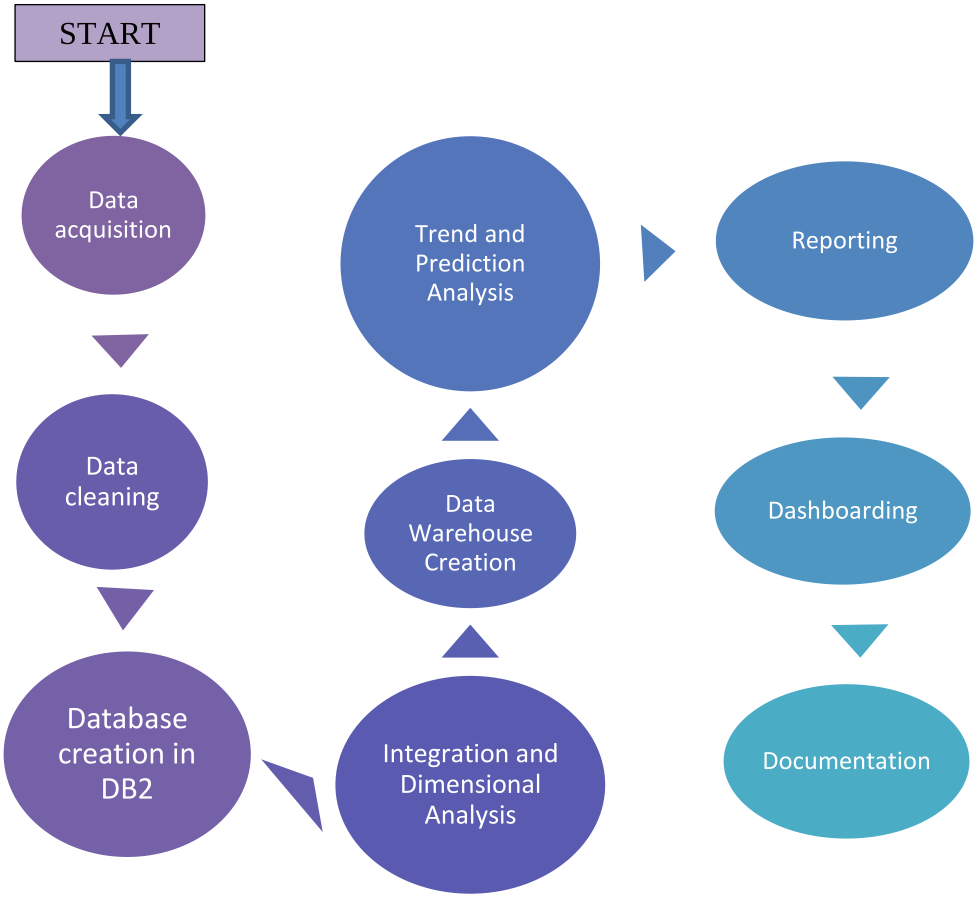


Figure 1: Work Flow of Project

Project Deliverables

Our project consists of the following deliverables:

1. Data acquisition.
2. Data Cleaning.
3. Integration and Dimensional Analysis.
4. Data Warehouse creation.
5. Dimensional Modeling
6. Trend and Prediction analysis.

7. Dash boarding.
8. Final Test case Reporting.
9. Documentation

Project Timeline

Table 1: Project Timeline

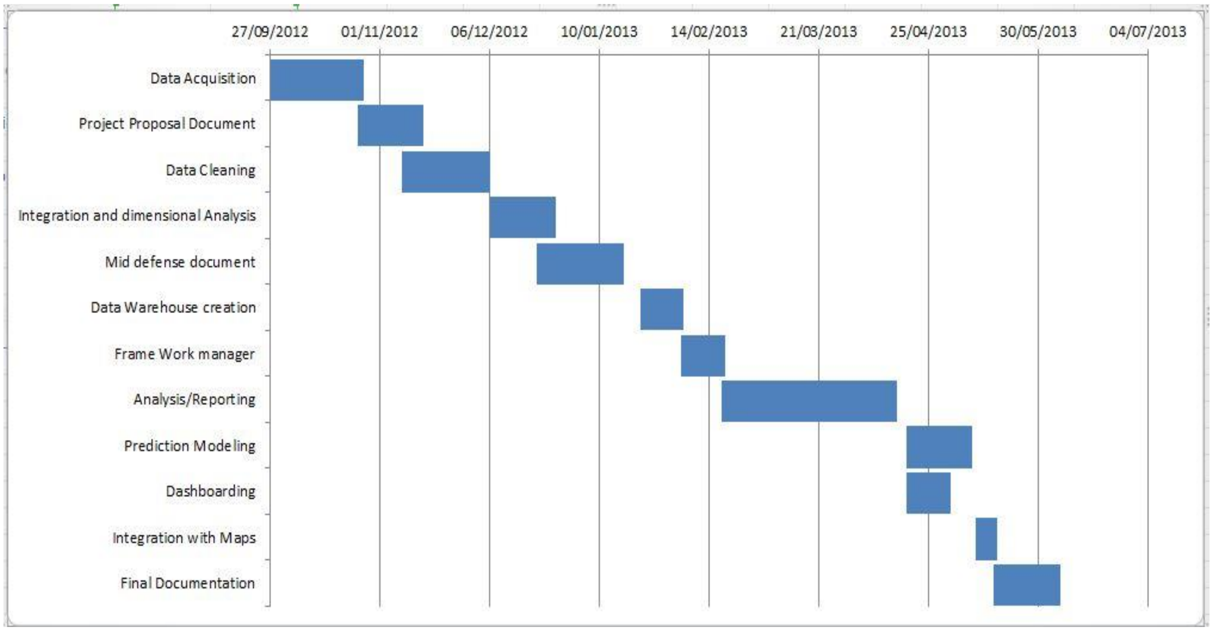


Figure 2: Gantt chart

Chapter 2

LITERATURE REVIEW

In this chapter the important topics and research related to this project will be highlighted.

Related Work:

This section contains review of the research papers, reports and projects related to Road safety.

a) **Intelligent System for Road Safety:**

This project was done in 2011-12, by Mr. Hamza Iftikhar and Mr. Abdul Ghaffar of BIT 10 SEECS NUST. The abstract of their project is as follows.

> *The problem of deaths and injuries as a result of road accidents is now acknowledged to be a Global phenomenon with authorities in virtually all countries of the world concerned about the growth in the number of people killed and seriously injured on their roads.*
>
> *World Health Organization has defined a standard for recording the road accidents data. Most of the countries follow that standard for recording their causalities and injuries caused by the accidents. The data is stored in bulk form, so that the government or the concerned authority can analyze the data and take proper measures in the future. Most of the countries are doing the same thing and using that data to reduce future accidents. But unfortunately in Pakistan, that data is collected to limited extent, but the standard format of surveying/collecting the data is not followed. So, the data is of no use for anyone.* [12]

b) Road Traffic Injury Research and Prevention Centre:

Ex-minister for health Dr. Jumma Khan started an awareness campaign in Karachi focused on road safety and accident prevention. He also collected data of accidents and related information like ambulance arrival time etc. His work in Karachi has been paramount in reducing the accidents there and creating awareness among people.[1]

Business Intelligence: Introduction and Benefits

Business intelligence is an assortment of applications that analyzes numerous aspects of data and facilitates decision making. It is a host of modules such as projecting, Online Analytical Processing (OLAP), predictive modeling, data management, data mining and optimization [6]. Using these tools, corporations can accurately judge trends in their organization and businesses and identify future trends to maximize profits. The various benefits of BI depend on the organization, but a number of general benefits are [7]

- Find out useful relationships between data events
- Provide visual representations of hypothetical scenarios, helping companies to see which option works best
- Meet or exceed customer expectations by using factual information
- Use multiple sources of data for strategic decisions
- Efficient distribution of statistics
- Central accessibility to departments and end users

[1] www.roadsafety.pk

Data Mining

The process of identifying and detecting patterns in the data is called data mining. Data mining techniques aim to provide insight that allows for a better understanding of data and its essential features. Companies and organizations can employ two basic types of data mining: [4]

- **Validation of Hypothesis**
  It is used to validate an idea or a hypothesis about a relation between data.
- **Knowledge Discovery:**
  It is used to find hidden and unknown relations to get statistical results which may be significant for the data elements.

Different tasks in data mining are

- **Classification:** A model classifies samples to a particular classified group
- **Clustering:** Groups samples that have common characteristics
- **Association:** Finds relations between events at a given time
- **Sequencing:** Similar to association. But relations are sorted according to time period
- **Regression:** Linear and nonlinear techniques to display information according to predictive value
- **Forecasting:** Prediction of future values based on history

Predictive Analysis

The success of an organization depends on the capability to analyze the future possibilities and understand the trends, so that correct action can be taken at the right time. Predictive analytics play an essential role to help and plan future actions according to risks and seize opportunities. The result of predictive analysis is used to identify patterns and trends and accuracy of the analysis depends on the complexity of interdependency of data [8]. Some predictive analysis methods are

- Linear regression

- Binomial Logistic regression
- Multinomial Logistic regression
- Decision trees
- Probity regression
- Neural networks

Data Analysis Tools:

There are many different tools available for data warehousing and Business Intelligence. Some of them are

- For ETL: Informatica and Kettle are available
- For reporting/analytics: Jasper, Pentaho and Cognos are available
- For Benchmarking: Informatic, Kettle and Cognos
- For enterprise Management: BO, Cognos and opensource tools

IBM Cognos for trend analysis:

It Provides services such as financial performace, strategy management, analytics application. [9]

IBM Cognos offers the following tools:

- ***IBM Cognos Metric Designer:***
  It is used to create extracts for use in IBM Cognos scorecarding applications. The created extracts are used to map and transfer information from existing metadata sources.
- ***IBM Cognos Query Studio:***
  Reporting tool for creating simple queries and reports in IBM Cognos Business Intelligence.
- ***IBM Cognos Framework Manager:***
  It is a metadata model development environment i.e. it assists in creating and managing business related metadata for use in BI analysis.

- ***IBM Cognos Connection:***

  It is a portal to IBM Cognos software and provides a single acces point to all corporate data available in IBM Cognos software.
- ***IBM Cognos Report Studio:***

  It is a report authoring tool used to create sophisticated and managed reports in short it is a full bloown report generatin tool.
- ***IBM Cognos Analysis Studio:***

  Wih analysis studio, it is possible to see trends and understand strange condition or situation or variances that may not be evident with other types of reporting.
- ***IBM Cognos Business Insight:***

  It is a web-based tool that allows you to use IBM Cognos content and external data sources like HTML/text sources to build interactive dashboards for better decision making.
- ***IBM Cognos for Microsoft Office:***

  It can be used to work with secure IBM Cognos BI content in Microsoft Office environment.

IBM SPSS for predictive analysis:

Given a dataset, many types of data analysis can be done. IBM SPSS is a powerful statistical tool that aids in building of accurate predictive models intuitively [10]. It enables the user to:

- Discover patterns and trends in data more easily using a visual interface and supported by advanced analytical techniques.
- The outcome of the model shows the influence of various factors on the considered problem. This helps the user to make better statistical decisions.

R-Language for statistical computing:

R is a free open source software programming language and a software environment for statistical computing and graphics. The R language is widely used in statistics and data mining for developing statistical software and data analysis. Polls and surveys of data miners are showing R's popularity has increased substantially in recent years. [11]

R language has a library called RgoogleMaps that can be used to do integration of your data set to Google maps using R.

Chapter 3

METHODOLGY AND IMPLEMENTATION

This chapter explains the detailed structure of the proposed solution along with the brief enlightenment.

Proposed Methodology

The methodology consists of the following mentioned points with their definitions.

- Data collection

  The most critical part of any project in the field of BI is the collection of authentic data for reporting. It is often the first stage, due to the fact that data is available in different formats and it is necessary to make a standard data set.
- Data cleaning

  It is possible that the gathered data is not up to the mark and contains some inconsistencies like duplication or null values, therefore data cleaning is an important task to make the data consistent.
- Database Creation

  This is the first proper stage of data warehousing. Database needs to be created so that the gathered data may be unified and arranged in a central location to offer ease of access.
- Metadata Modeling

  It is also important as this provides the crucial link between the raw data and the highly organized reports; it is a sample of data on which you model your design to help in the reporting process.
- Reporting and Analysis

  The most crucial component is BI is the reporting and analysis because in the end that is why business intelligence exists so that you may be able to make sense of data and answer important business questions, reporting and analysis is the key to that.

- Prediction Analysis

  The next stage in BI is to predict what happens next or what if this scenario occurs what then, therefore prediction is the usage of statistical models for successful answers of what if questions.
- Location Mapping

  To make BI more appealing visually to the customer mapping is used so that you get an easy to understand interface of your business.

Methodology Implementation

As we discussed in the proposed methodology the different concepts and stages of BI, let us now explain how they are implemented in our project.

Now, all the steps will be explained in detail.

Data collection:

It is a really hectic process as data collection of accidents in Pakistan is not up to the international standards on which effective reporting is done and further recommendations are made. We visited the following offices for data acquisition.

- National Highways & Motorway Police: Meeting with Mr. Farooq Azam (AIG-R&D)
- Rescue 15: Meeting with SP Mr. Sajid Kiyani
- Islamabad Traffic Police: Meeting with SSP Mr. Moin who provided us with statistical data.
- Road Traffic Injury Research and Prevention Centre Karachi: contacted Dr. Jooma Khan
- Kashmir: SP Mr. Sajid Kiyani from RESCUE 15 assisted us in getting statistical data from Kashmir.
- 1122: Meeting with Mr. Tauqir Khan (Assistant Director - Cares)

We were able get data from National Highways & Motorway Police with the assistance from Mr. Farooq Azam, who was kind enough to lend us the data from end-

2008 to mid-2012 with the hope to improve the Traffic System of NH & MP with our road safety recommendations.

We acquired Accidents data in the form of excel sheets.

1- Accident Wise

2- Vehicle Wise

Accident Wise:

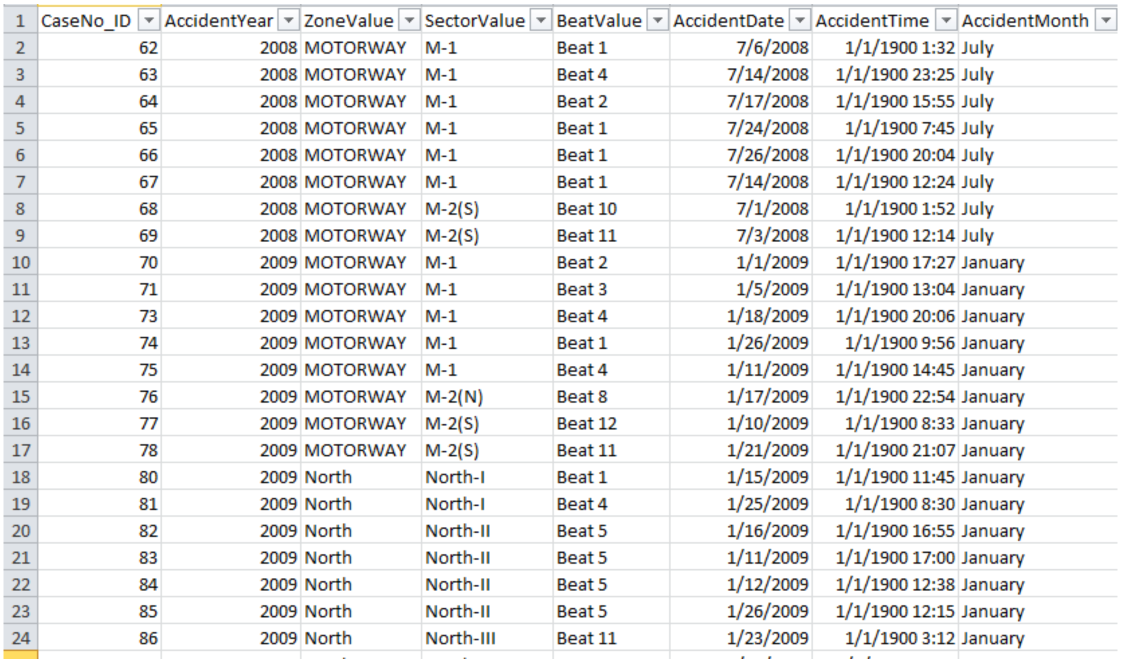
No of accident data entries = 2044

Column Names:

Table 2: Accident Wise Column names

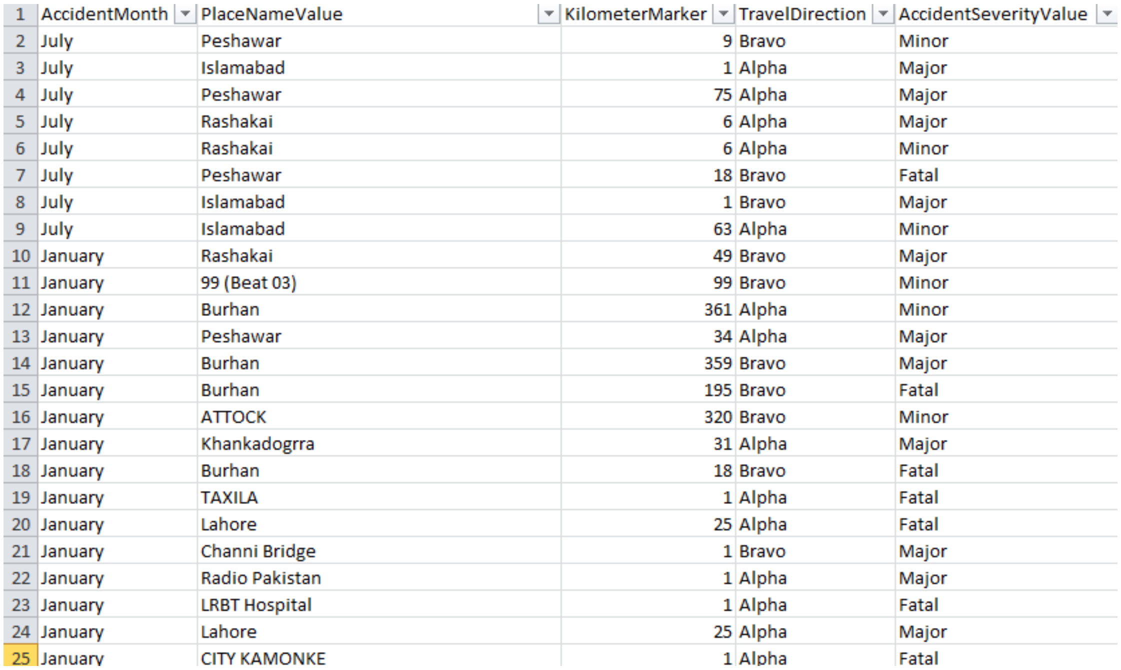
Column Names	Column Names	Column Names
CaseNo_ID	Accident Year	Zone Value
Sector Value	No Passenger Killed	No Passenger Injured
Beat Value	No Driver Injured	No Pedestrian Killed
Accident Month	No Driver Killed	No Pedestrian Injured
Place Name Value	Accident Severity Value	Collision Type Value
Accident Time	Travel Direction	Weather Value
Accident Date	Kilometer Marker	Light Condition Value
Road Geometry Value	Accident Cause Value	Accident Day
Network Value	Motorway Name Value	

Following are screenshots of Accident Wise data:



1	CaseNo_ID	AccidentYear	ZoneValue	SectorValue	BeatValue	AccidentDate	AccidentTime	AccidentMonth
2	62	2008	MOTORWAY	M-1	Beat 1	7/6/2008	1/1/1900 1:32	July
3	63	2008	MOTORWAY	M-1	Beat 4	7/14/2008	1/1/1900 23:25	July
4	64	2008	MOTORWAY	M-1	Beat 2	7/17/2008	1/1/1900 15:55	July
5	65	2008	MOTORWAY	M-1	Beat 1	7/24/2008	1/1/1900 7:45	July
6	66	2008	MOTORWAY	M-1	Beat 1	7/26/2008	1/1/1900 20:04	July
7	67	2008	MOTORWAY	M-1	Beat 1	7/14/2008	1/1/1900 12:24	July
8	68	2008	MOTORWAY	M-2(S)	Beat 10	7/1/2008	1/1/1900 1:52	July
9	69	2008	MOTORWAY	M-2(S)	Beat 11	7/3/2008	1/1/1900 12:14	July
10	70	2009	MOTORWAY	M-1	Beat 2	1/1/2009	1/1/1900 17:27	January
11	71	2009	MOTORWAY	M-1	Beat 3	1/5/2009	1/1/1900 13:04	January
12	73	2009	MOTORWAY	M-1	Beat 4	1/18/2009	1/1/1900 20:06	January
13	74	2009	MOTORWAY	M-1	Beat 1	1/26/2009	1/1/1900 9:56	January
14	75	2009	MOTORWAY	M-1	Beat 4	1/11/2009	1/1/1900 14:45	January
15	76	2009	MOTORWAY	M-2(N)	Beat 8	1/17/2009	1/1/1900 22:54	January
16	77	2009	MOTORWAY	M-2(S)	Beat 12	1/10/2009	1/1/1900 8:33	January
17	78	2009	MOTORWAY	M-2(S)	Beat 11	1/21/2009	1/1/1900 21:07	January
18	80	2009	North	North-I	Beat 1	1/15/2009	1/1/1900 11:45	January
19	81	2009	North	North-I	Beat 4	1/25/2009	1/1/1900 8:30	January
20	82	2009	North	North-II	Beat 5	1/16/2009	1/1/1900 16:55	January
21	83	2009	North	North-II	Beat 5	1/11/2009	1/1/1900 17:00	January
22	84	2009	North	North-II	Beat 5	1/12/2009	1/1/1900 12:38	January
23	85	2009	North	North-II	Beat 5	1/26/2009	1/1/1900 12:15	January
24	86	2009	North	North-III	Beat 11	1/23/2009	1/1/1900 3:12	January

Figure 3: Data



1	AccidentMonth	PlaceNameValue	KilometerMarker	TravelDirection	AccidentSeverityValue
2	July	Peshawar	9	Bravo	Minor
3	July	Islamabad	1	Alpha	Major
4	July	Peshawar	75	Alpha	Major
5	July	Rashakai	6	Alpha	Major
6	July	Rashakai	6	Alpha	Minor
7	July	Peshawar	18	Bravo	Fatal
8	July	Islamabad	1	Bravo	Major
9	July	Islamabad	63	Alpha	Minor
10	January	Rashakai	49	Bravo	Major
11	January	99 (Beat 03)	99	Bravo	Minor
12	January	Burhan	361	Alpha	Minor
13	January	Peshawar	34	Alpha	Major
14	January	Burhan	359	Bravo	Major
15	January	Burhan	195	Bravo	Fatal
16	January	ATTOCK	320	Bravo	Minor
17	January	Khankadogrra	31	Alpha	Major
18	January	Burhan	18	Bravo	Fatal
19	January	TAXILA	1	Alpha	Fatal
20	January	Lahore	25	Alpha	Fatal
21	January	Channi Bridge	1	Bravo	Major
22	January	Radio Pakistan	1	Alpha	Major
23	January	LRBT Hospital	1	Alpha	Fatal
24	January	Lahore	25	Alpha	Major
25	January	CITY KAMONKE	1	Alpha	Fatal

Figure 4: Data-2

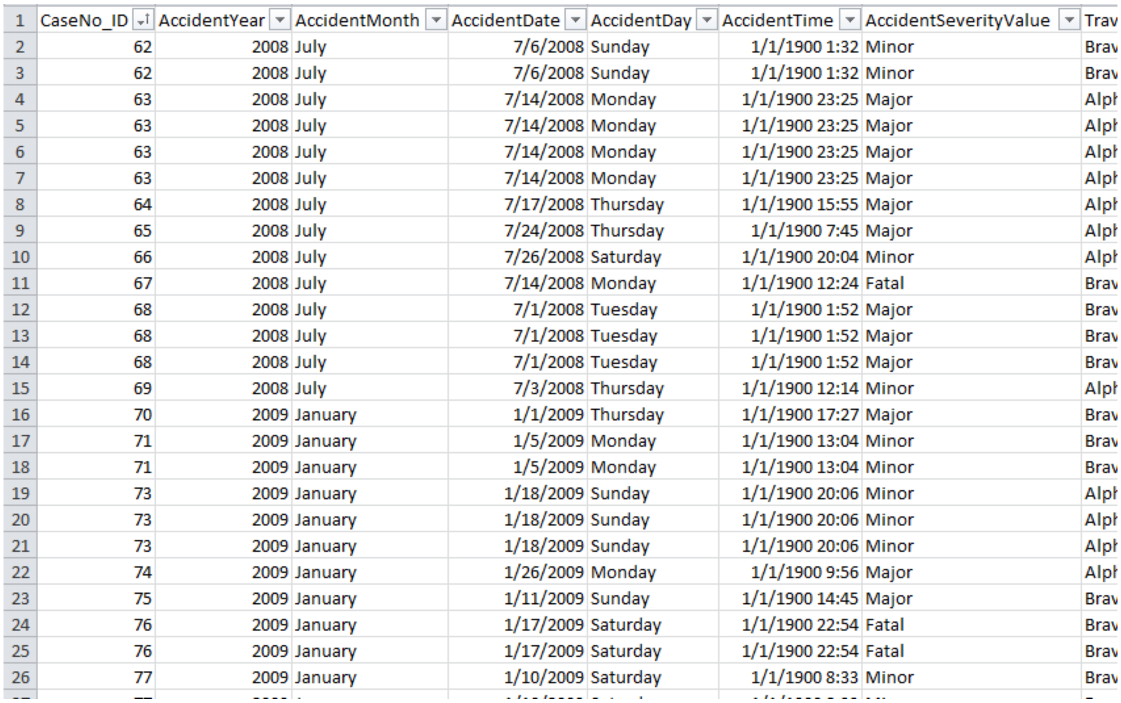
Vehicle Wise:
No of accident data entries = 3471 (i.e. more vehicles are involved in one accident)

Column Names:

Table 3: Vehicle Wise Column Names

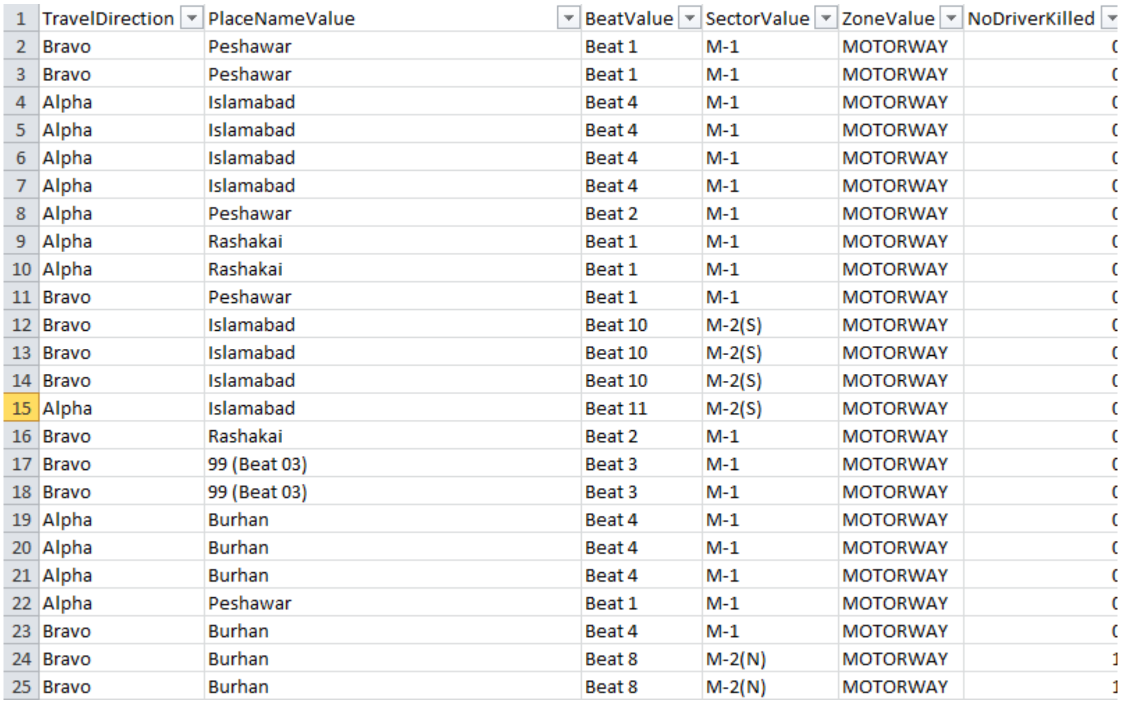
Column Names	Column Names	Column Names
CaseNo_ID	Accident Year	Zone Value
Sector Value	No Passenger Killed	No Passenger Injured
Beat Value	No Driver Injured	No Pedestrian Killed
Accident Month	No Driver Killed	No Pedestrian Injured
Place Name Value	Accident Severity Value	Collision Type Value
Accident Time	Travel Direction	Weather Value
Accident Date	Kilometer Marker	Light Condition Value
Road Geometry Value	Accident Cause Value	Accident Day
Vehicle Company Value	Motorway Name Value	Vehicle Type Value

Following are screenshots of Vehicle Wise data:



1	CaseNo_ID	AccidentYear	AccidentMonth	AccidentDate	AccidentDay	AccidentTime	AccidentSeverityValue	Trav
2	62	2008	July	7/6/2008	Sunday	1/1/1900 1:32	Minor	Brav
3	62	2008	July	7/6/2008	Sunday	1/1/1900 1:32	Minor	Brav
4	63	2008	July	7/14/2008	Monday	1/1/1900 23:25	Major	Alph
5	63	2008	July	7/14/2008	Monday	1/1/1900 23:25	Major	Alph
6	63	2008	July	7/14/2008	Monday	1/1/1900 23:25	Major	Alph
7	63	2008	July	7/14/2008	Monday	1/1/1900 23:25	Major	Alph
8	64	2008	July	7/17/2008	Thursday	1/1/1900 15:55	Major	Alph
9	65	2008	July	7/24/2008	Thursday	1/1/1900 7:45	Major	Alph
10	66	2008	July	7/26/2008	Saturday	1/1/1900 20:04	Minor	Alph
11	67	2008	July	7/14/2008	Monday	1/1/1900 12:24	Fatal	Brav
12	68	2008	July	7/1/2008	Tuesday	1/1/1900 1:52	Major	Brav
13	68	2008	July	7/1/2008	Tuesday	1/1/1900 1:52	Major	Brav
14	68	2008	July	7/1/2008	Tuesday	1/1/1900 1:52	Major	Brav
15	69	2008	July	7/3/2008	Thursday	1/1/1900 12:14	Minor	Alph
16	70	2009	January	1/1/2009	Thursday	1/1/1900 17:27	Major	Brav
17	71	2009	January	1/5/2009	Monday	1/1/1900 13:04	Minor	Brav
18	71	2009	January	1/5/2009	Monday	1/1/1900 13:04	Minor	Brav
19	73	2009	January	1/18/2009	Sunday	1/1/1900 20:06	Minor	Alph
20	73	2009	January	1/18/2009	Sunday	1/1/1900 20:06	Minor	Alph
21	73	2009	January	1/18/2009	Sunday	1/1/1900 20:06	Minor	Alph
22	74	2009	January	1/26/2009	Monday	1/1/1900 9:56	Major	Alph
23	75	2009	January	1/11/2009	Sunday	1/1/1900 14:45	Major	Brav
24	76	2009	January	1/17/2009	Saturday	1/1/1900 22:54	Fatal	Brav
25	76	2009	January	1/17/2009	Saturday	1/1/1900 22:54	Fatal	Brav
26	77	2009	January	1/10/2009	Saturday	1/1/1900 8:33	Minor	Brav

Figure 5: Vehicle wise data



1	TravelDirection	PlaceNameValue	BeatValue	SectorValue	ZoneValue	NoDriverKilled
2	Bravo	Peshawar	Beat 1	M-1	MOTORWAY	(
3	Bravo	Peshawar	Beat 1	M-1	MOTORWAY	(
4	Alpha	Islamabad	Beat 4	M-1	MOTORWAY	(
5	Alpha	Islamabad	Beat 4	M-1	MOTORWAY	(
6	Alpha	Islamabad	Beat 4	M-1	MOTORWAY	(
7	Alpha	Islamabad	Beat 4	M-1	MOTORWAY	(
8	Alpha	Peshawar	Beat 2	M-1	MOTORWAY	(
9	Alpha	Rashakai	Beat 1	M-1	MOTORWAY	(
10	Alpha	Rashakai	Beat 1	M-1	MOTORWAY	(
11	Bravo	Peshawar	Beat 1	M-1	MOTORWAY	(
12	Bravo	Islamabad	Beat 10	M-2(S)	MOTORWAY	(
13	Bravo	Islamabad	Beat 10	M-2(S)	MOTORWAY	(
14	Bravo	Islamabad	Beat 10	M-2(S)	MOTORWAY	(
15	Alpha	Islamabad	Beat 11	M-2(S)	MOTORWAY	(
16	Bravo	Rashakai	Beat 2	M-1	MOTORWAY	(
17	Bravo	99 (Beat 03)	Beat 3	M-1	MOTORWAY	(
18	Bravo	99 (Beat 03)	Beat 3	M-1	MOTORWAY	(
19	Alpha	Burhan	Beat 4	M-1	MOTORWAY	(
20	Alpha	Burhan	Beat 4	M-1	MOTORWAY	(
21	Alpha	Burhan	Beat 4	M-1	MOTORWAY	(
22	Alpha	Peshawar	Beat 1	M-1	MOTORWAY	(
23	Bravo	Burhan	Beat 4	M-1	MOTORWAY	(
24	Bravo	Burhan	Beat 8	M-2(N)	MOTORWAY	1
25	Bravo	Burhan	Beat 8	M-2(N)	MOTORWAY	1

Figure 6: Vehicle wise-2

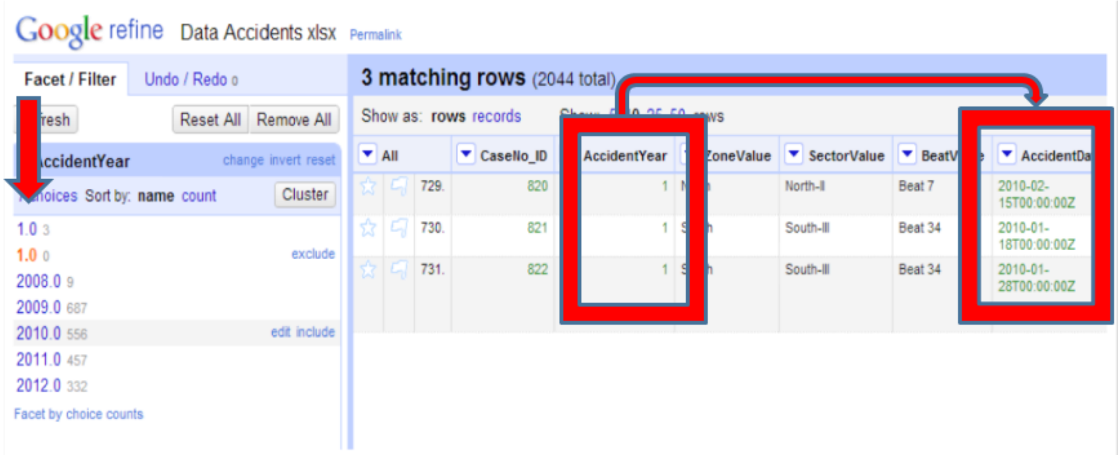
Data cleaning:

Google Refine was used as a tool for data cleaning.

Data Cleaning in Google Refine included

- **Finding/ Removing the inconsistencies in the data**

  An inconsistency was found in ‘Accident Year’ column where three rows had the value of 1.



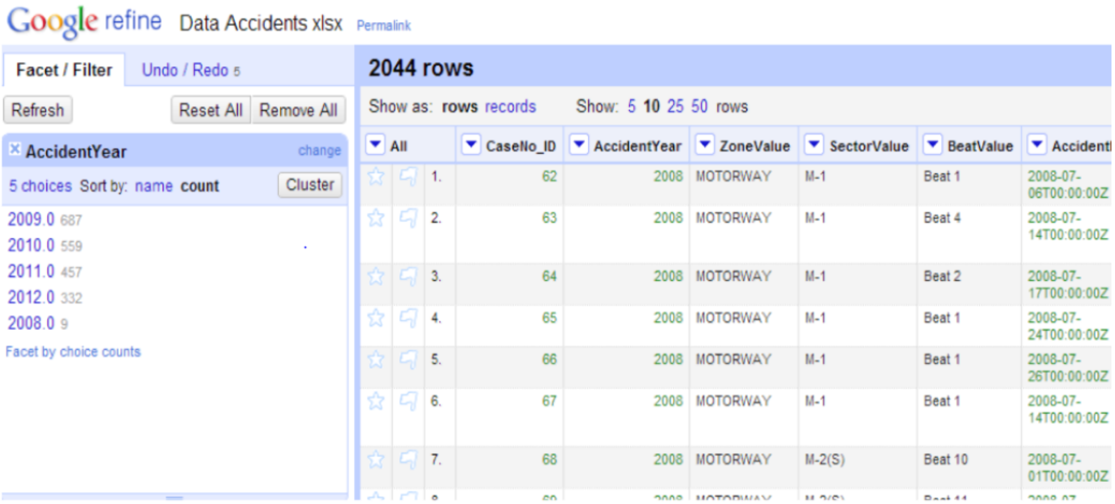
3 matching rows (2044 total)

1.0 3
1.0 0
2008.0 9
2009.0 687
2010.0 556
2011.0 457
2012.0 332

All	CaseNo_ID	AccidentYear	ZoneValue	SectorValue	BeatV e	AccidentDa
729.	820	1	N	North-II	Beat 7	2010-02-15T00:00:00Z
730.	821	1	S h	South-III	Beat 34	2010-01-18T00:00:00Z
731.	822	1	S h	South-III	Beat 34	2010-01-28T00:00:00Z

Figure 7: Google Refine

After the data was closely observed, it was found form the ‘Accident Date’ column that these entries belong to ‘Accident Year’ 2010. And so this inconsistency was removed.



2044 rows

AccidentYear

5 choices Sort by: name count

2009.0	687
2010.0	559
2011.0	457
2012.0	332
2008.0	9

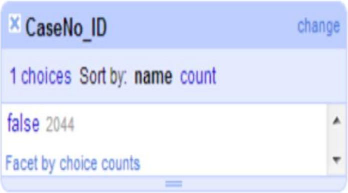
All	CaseNo_ID	AccidentYear	ZoneValue	SectorValue	BeatValue	Accident
1.	62	2008	MOTORWAY	M-1	Beat 1	2008-07-06T00:00:00Z
2.	63	2008	MOTORWAY	M-1	Beat 4	2008-07-14T00:00:00Z
3.	64	2008	MOTORWAY	M-1	Beat 2	2008-07-17T00:00:00Z
4.	65	2008	MOTORWAY	M-1	Beat 1	2008-07-24T00:00:00Z
5.	66	2008	MOTORWAY	M-1	Beat 1	2008-07-26T00:00:00Z
6.	67	2008	MOTORWAY	M-1	Beat 1	2008-07-14T00:00:00Z
7.	68	2008	MOTORWAY	M-2(S)	Beat 10	2008-07-01T00:00:00Z

Figure 8: Google Refine-2

- **Validation**

1. Duplicate Facet

It was found that data has 2044 unique entries by checking with Duplicate Facet.



CaseNo_ID

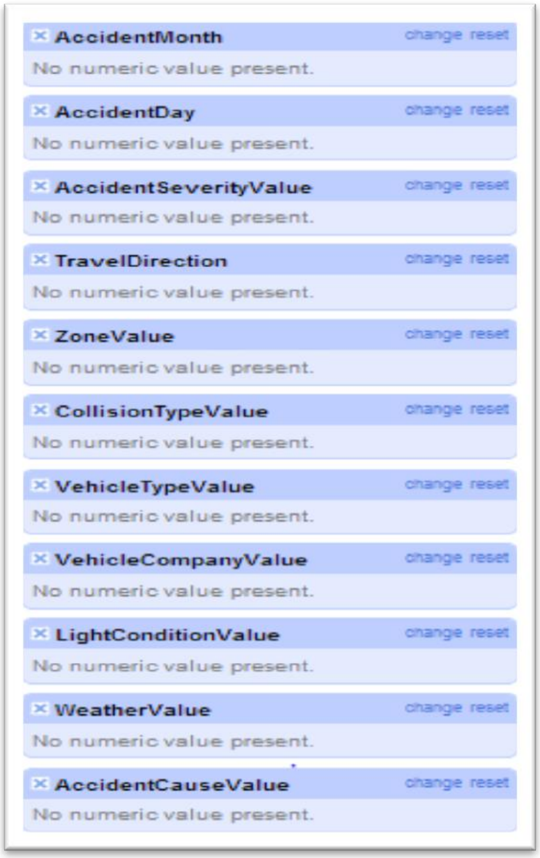
1 choices Sort by: name count

false 2044

Figure 9: Duplicate facet

2. Numerical Facet

The rows which should have text values present e.g. Accident Month, were checked with numeric facet to check that they do not have any text value present.



AccidentMonth
No numeric value present.

AccidentDay
No numeric value present.

AccidentSeverityValue
No numeric value present.

TravelDirection
No numeric value present.

ZoneValue
No numeric value present.

CollisionTypeValue
No numeric value present.

VehicleTypeValue
No numeric value present.

VehicleCompanyValue
No numeric value present.

LightConditionValue
No numeric value present.

WeatherValue
No numeric value present.

AccidentCauseValue
No numeric value present.

Figure 10: Numerical Facet

3. Blank Facets

The whole data i.e. for all 26 columns was verified for null values using Blank Facet. ‘False 2044’ indicates that none of the 2044 rows has any null value present.

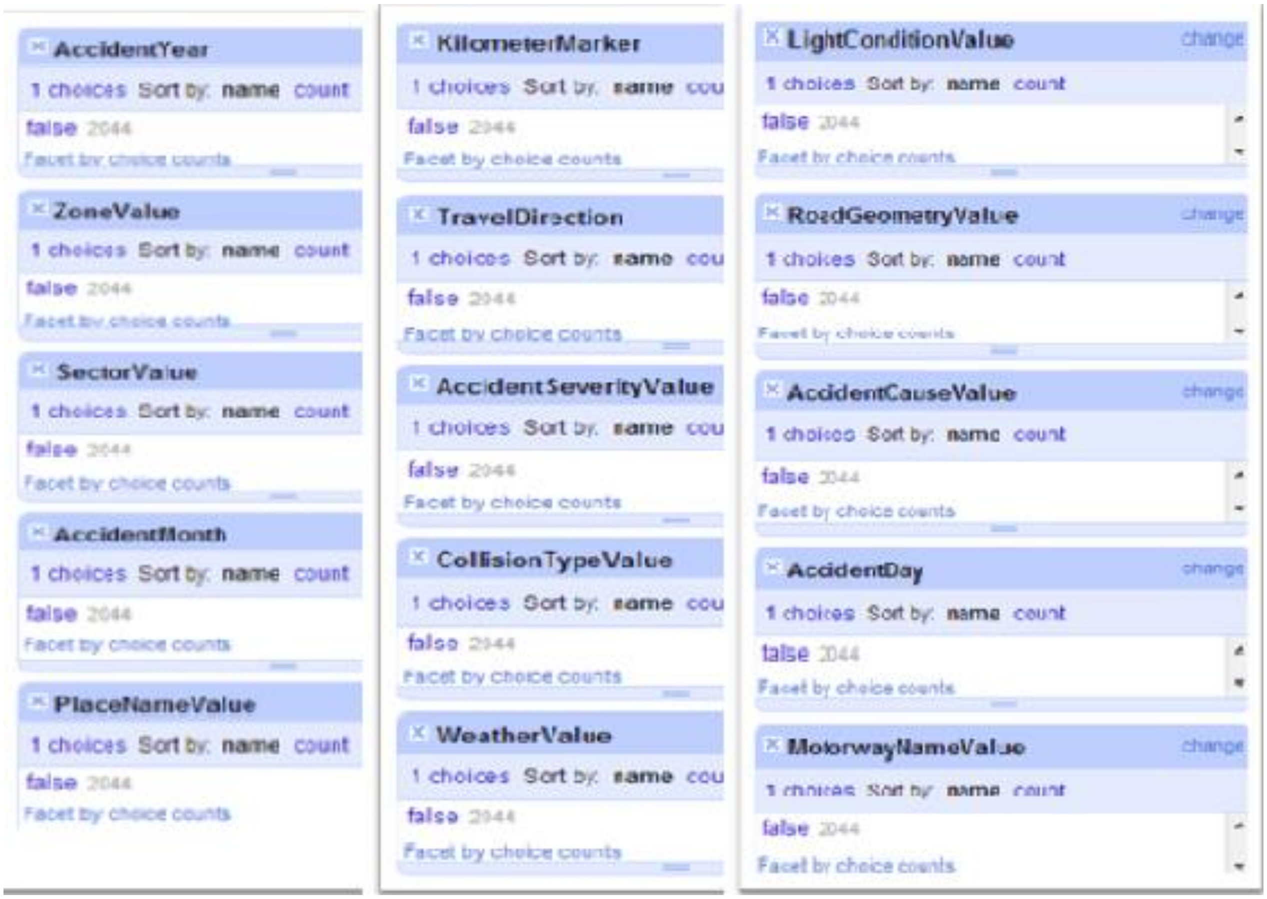
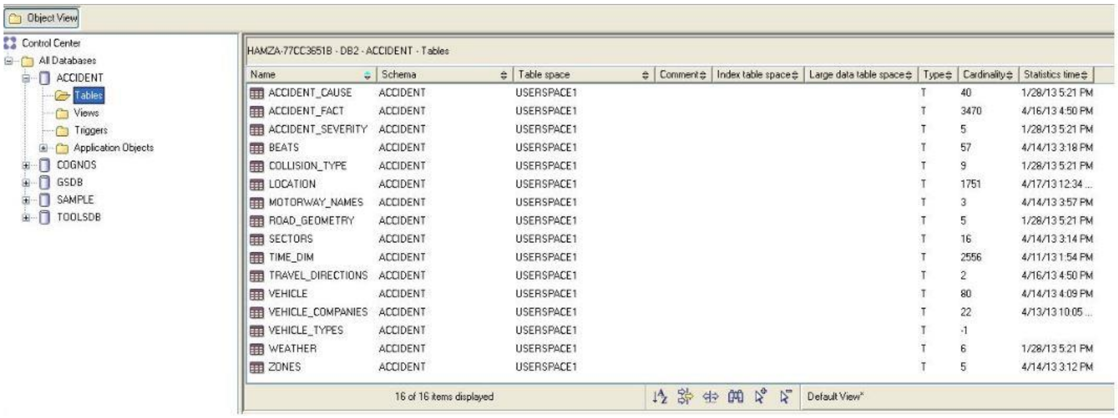


Figure 11: Blank facets

Database Creation:

IBM DB2 Enterprise Server was chosen as a database server. All the refined data was loaded into the database from csv files. A snapshot is as follows

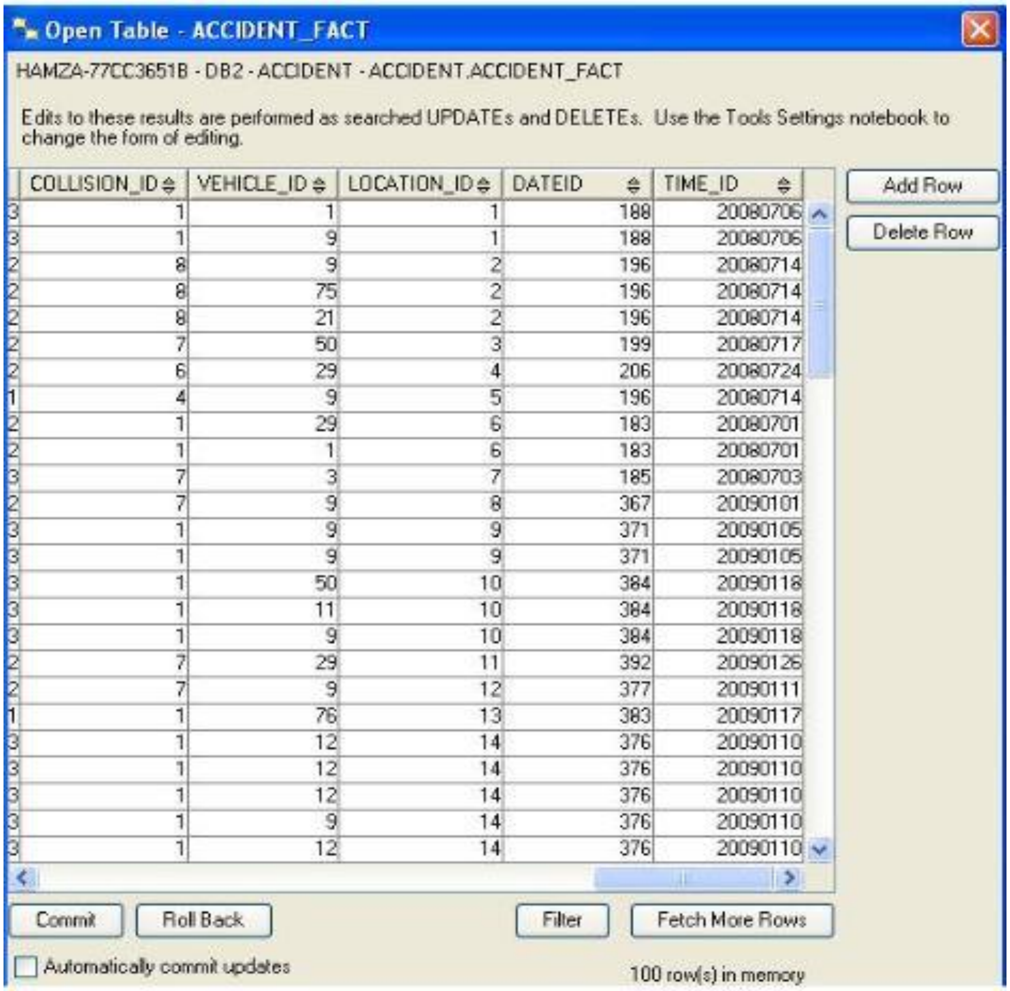


HAMZA-77CC3651B - DB2 - ACCIDENT - Tables

Name	Schema	Table space	Comment	Index table space	Large data table space	Type	Cardinality	Statistics time
ACCIDENT_CAUSE	ACCIDENT	USERSPACE1				T	40	1/28/13 5:21 PM
ACCIDENT_FACT	ACCIDENT	USERSPACE1				T	3470	4/16/13 4:50 PM
ACCIDENT_SEVERITY	ACCIDENT	USERSPACE1				T	5	1/28/13 5:21 PM
BEATS	ACCIDENT	USERSPACE1				T	57	4/14/13 3:18 PM
COLLISION_TYPE	ACCIDENT	USERSPACE1				T	9	1/28/13 5:21 PM
LOCATION	ACCIDENT	USERSPACE1				T	1751	4/17/13 12:34 ...
MOTORWAY_NAMES	ACCIDENT	USERSPACE1				T	3	4/14/13 3:57 PM
ROAD_GEOMETRY	ACCIDENT	USERSPACE1				T	5	1/28/13 5:21 PM
SECTORS	ACCIDENT	USERSPACE1				T	16	4/14/13 3:14 PM
TIME_DIM	ACCIDENT	USERSPACE1				T	2556	4/11/13 1:54 PM
TRAVEL_DIRECTIONS	ACCIDENT	USERSPACE1				T	2	4/16/13 4:50 PM
VEHICLE	ACCIDENT	USERSPACE1				T	80	4/14/13 4:09 PM
VEHICLE_COMPANIES	ACCIDENT	USERSPACE1				T	22	4/13/13 10:05 ...
VEHICLE_TYPES	ACCIDENT	USERSPACE1				T	-1	
WEATHER	ACCIDENT	USERSPACE1				T	6	1/28/13 5:21 PM
ZONES	ACCIDENT	USERSPACE1				T	5	4/14/13 3:12 PM

16 of 16 items displayed

Figure 12: IBM DB2



HAMZA-77CC3651B - DB2 - ACCIDENT - ACCIDENT.ACCIDENT_FACT

Edits to these results are performed as searched UPDATEs and DELETEs. Use the Tools Settings notebook to change the form of editing.

	COLLISION_ID	VEHICLE_ID	LOCATION_ID	DATEID	TIME_ID
3	1	1	1	188	20080706
3	1	9	1	188	20080706
2	8	9	2	196	20080714
2	8	75	2	196	20080714
2	8	21	2	196	20080714
2	7	50	3	199	20080717
2	6	29	4	206	20080724
1	4	9	5	196	20080714
2	1	29	6	183	20080701
2	1	1	6	183	20080701
3	7	3	7	185	20080703
2	7	9	8	367	20090101
3	1	9	9	371	20090105
3	1	9	9	371	20090105
3	1	50	10	384	20090118
3	1	11	10	384	20090118
3	1	9	10	384	20090118
2	7	29	11	392	20090126
2	7	9	12	377	20090111
1	1	76	13	383	20090117
3	1	12	14	376	20090110
3	1	12	14	376	20090110
3	1	12	14	376	20090110
3	1	9	14	376	20090110
3	1	12	14	376	20090110

□ Automatically commit updates

100 row(s) in memory

Figure 13: DB2

Installation of COGNOS BI:

After database creation, the next step was to configure an IBM Cognos BI server, in order to move to metadata model development and BI reporting and analysis.

We installed the IBM Cognos BI 10.1.1. The server was setup up using the standard procedure, available in the Official Installation Guide published by IBM Corporation, USA. The following components of IBM Cognos BI were available for use, after the successful deployment:

- Framework Manager
- Report Studio
- Query Studio

Framework Manager:

Metadata model was created for the purpose of reporting in which facts and Dimensions related to Auto Accidents were identified. The information given comes from the data mart.

Metadata model hides the structural complexity of the underlying data sources and makes reporting faster and easier. Advantages of framework manager include predictable results, an easy to use interface and a dimensional model in the form of Star Schema.

Star schema:

It is the dimensional view necessary for data warehousing and for intelligent reporting. Following are individual dimensions in the star schema followed by the star schema.

Accident Fact

- No of drivers killed
- No of drivers injured
- No of passengers killed
- No of passengers injured
- No of pedestrians killed
- No of pedestrians injured

Weather

- Weather ID
- Weather Type

Road Geometry

- Road_ID
- Road geometry Type

Collision

- Collision ID
- Collision Type

Accident Cause

- Cause ID
- Cause Type

Accident Severity

- Severity Id
- Severity Type

Location

- Location ID
- Place name
- Kilometer marker
- Beat ID
- Zone ID

Time

- Day Key
- Month Key
- Year Key
- Day of Week Key
- Quarter Key
- Current Date

Vehicle

- Vehicle ID
- Vehicle Type
- Vehicle Company

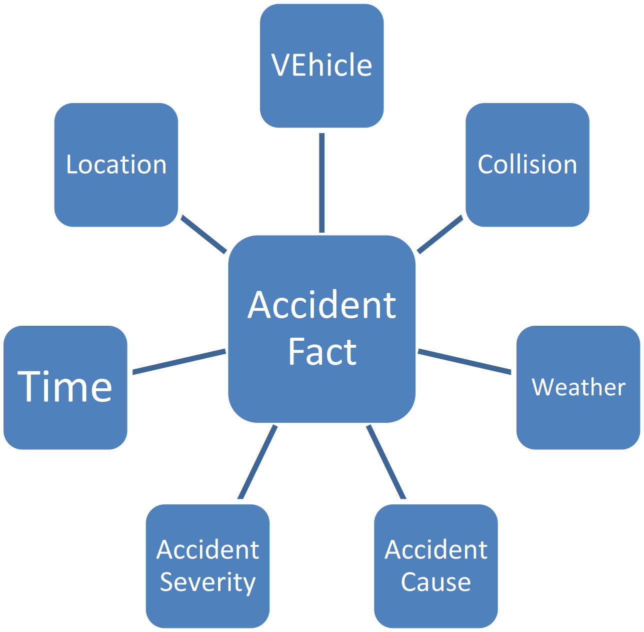
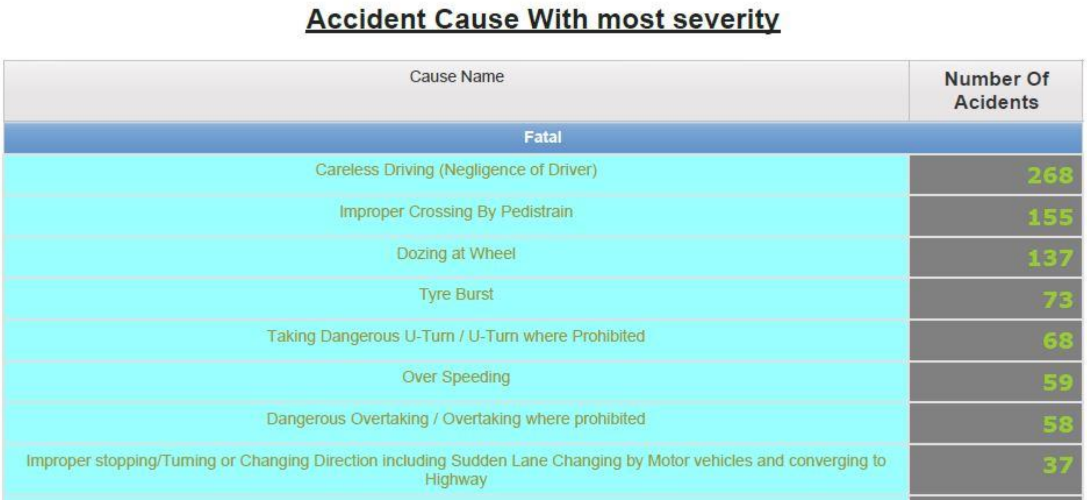


Figure 14:Star Schema

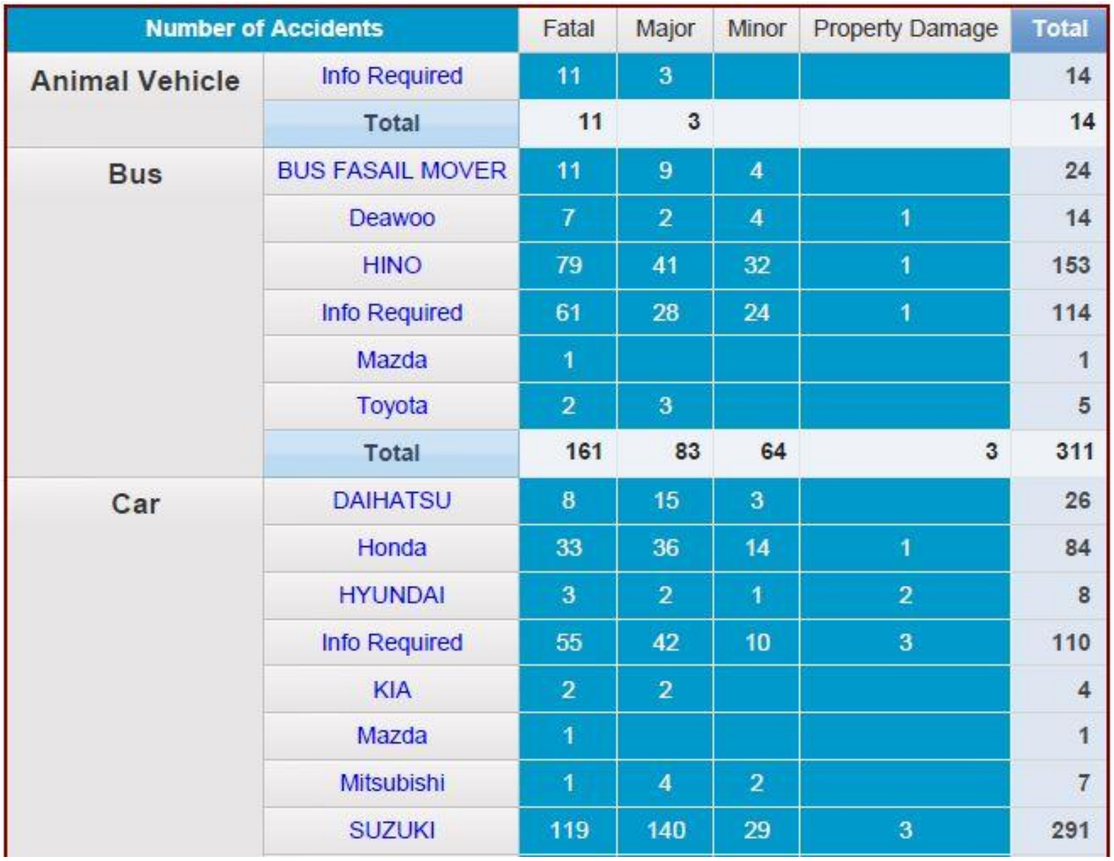
Reporting:
After the metadata model was completed the next step was to publish it on Cognos content so that we may form reports and perform analysis on the data. We made different reports which we will mention in detail in the next chapter. A couple of examples are as follows.



Accident Cause With most severity

Cause Name	Number Of Acidents
Fatal	
Careless Driving (Negligence of Driver)	268
Improper Crossing By Pedistrain	155
Dozing at Wheel	137
Tyre Burst	73
Taking Dangerous U-Turn / U-Turn where Prohibited	68
Over Speeding	59
Dangerous Overtaking / Overtaking where prohibited	58
Improper stopping/Turning or Changing Direction including Sudden Lane Changing by Motor vehicles and converging to Highway	37

Figure 15: Report Example



Number of Accidents		Fatal	Major	Minor	Property Damage	Total
Animal Vehicle	Info Required	11	3			14
	Total	11	3			14
Bus	BUS FASAIL MOVER	11	9	4		24
	Deawoo	7	2	4	1	14
	HINO	79	41	32	1	153
	Info Required	61	28	24	1	114
	Mazda	1				1
	Toyota	2	3			5
	Total	161	83	64	3	311
Car	DAIHATSU	8	15	3		26
	Honda	33	36	14	1	84
	HYUNDAI	3	2	1	2	8
	Info Required	55	42	10	3	110
	KIA	2	2			4
	Mazda	1				1
	Mitsubishi	1	4	2		7
	SUZUKI	119	140	29	3	291

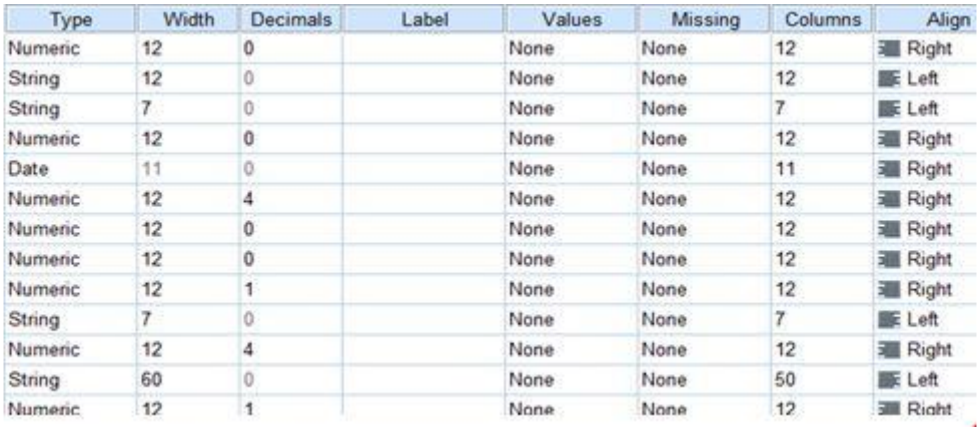
Figure 16: Report example 2

Statistical Analysis:

Most important step of the project is to statistically analyze the data to understand the impact of different variables that are affecting the particular data. For this purpose IBM SPSS statistics tool is used.

IBM SPSS:

The different files are loaded into IBM SPSS statistics tool and one comprehensive file containing all the required variables is saved in .sav format. A validation check is performed in the end to verify variable data types and measures.



Type	Width	Decimals	Label	Values	Missing	Columns	Align
Numeric	12	0		None	None	12	Right
String	12	0		None	None	12	Left
String	7	0		None	None	7	Left
Numeric	12	0		None	None	12	Right
Date	11	0		None	None	11	Right
Numeric	12	4		None	None	12	Right
Numeric	12	0		None	None	12	Right
Numeric	12	0		None	None	12	Right
Numeric	12	1		None	None	12	Right
String	7	0		None	None	7	Left
Numeric	12	4		None	None	12	Right
String	60	0		None	None	50	Left
Numeric	12	1		None	None	12	Right

Figure 17: SPSS Example

Solution Architecture

After the detailed analysis of the problem statement and deeper insight into the project, an approach to address the problem statement was devised.

A layered plan of action from the first step i.e. data sources to the last step i.e. Predictive Modeling, is adapted to achieve all the objectives smoothly and timely.

The diagram of the detailed 6-layered solution architecture of the project is given below.

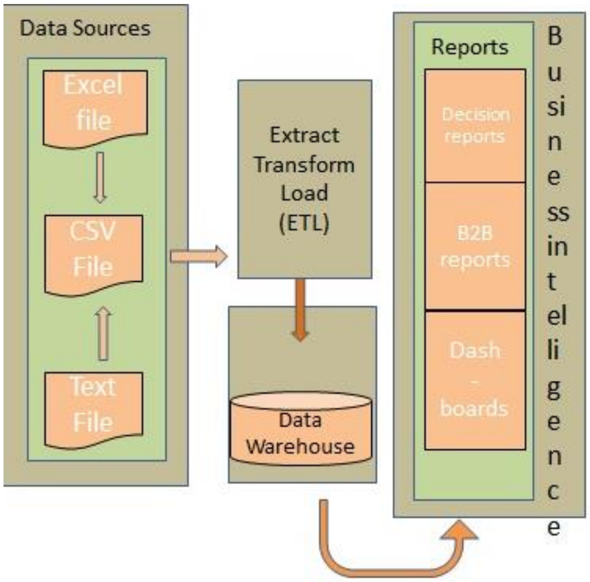


Figure 18: Architecture Diagram

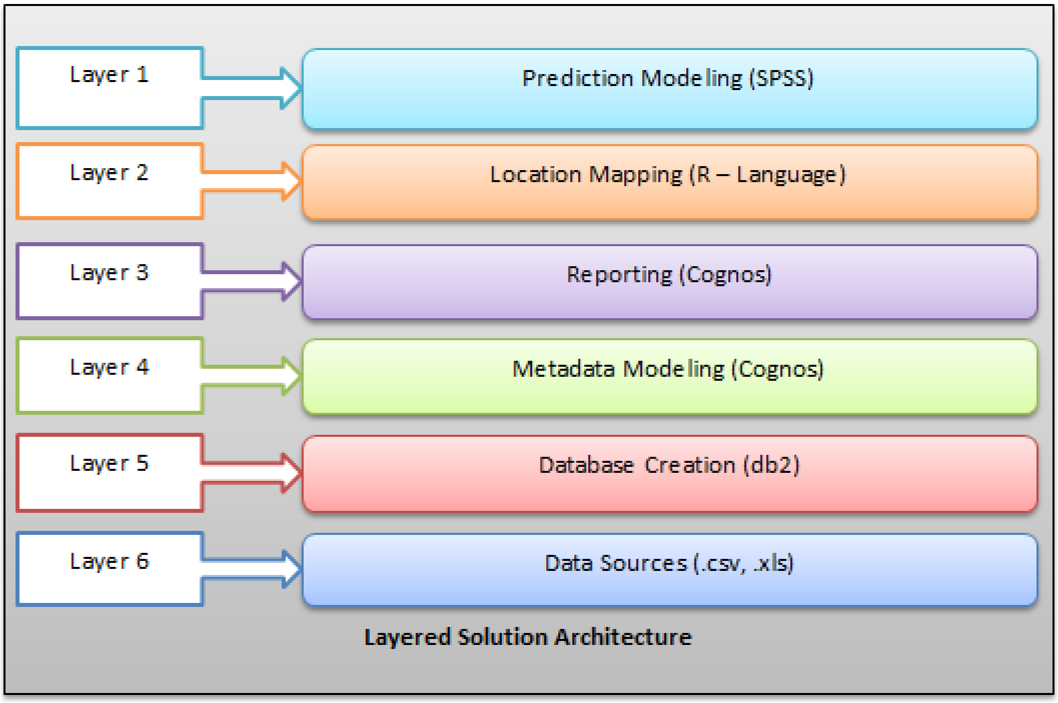

Layer 1
Prediction Modeling (SPSS)
Layer 2
Location Mapping (R – Language)
Layer 3
Reporting (Cognos)
Layer 4
Metadata Modeling (Cognos)
Layer 5
Database Creation (db2)
Layer 6
Data Sources (.csv, .xls)
Layered Solution Architecture

Figure 3: Layered Solution Architecture

From the figure, we can see that layer 6 and layer 5 are related to data and database. In these two layers the data has been loaded into the database after collection from different data sources and performing data cleaning. Database is created in IBM DB2.

In layer 4, a metadata model is created which would store the star schema groupings of the dimension tables and a fact table.

Now, metadata model is ready for business intelligence reports to be made upon it. Thus in layer 3, business intelligence reports are made with suggestion and recommendation alongside.

Also, the locations of accident spots are marked on map using R-language.

Finally, comes the part of predictive analytics in which the objective is to predict the severity category of the accidents depending upon various independent factors.

Tools and Technologies

The tools and technologies being used to achieve the stated objectives by following the proposed strategy are as follows:

Table 2: Tools and Technologies

	Tools	**Purposes**
Database	Google Refine	Data Cleaning and Refining
	IBM DB2 Enterprise Server Edition	Database Creation
BI Modeling and Reporting	Cognos: Framework Manager	Metadata Modeling for quick reporting and trend
	Cognos: Report Studio	For making intelligent reports and dashboards
Predictive Analysis	IBM SPSS Statistics	Prediction Modeling
Mapping	R-Language	Plotting locations on maps

Chapter 4

RESULTS

Our results are divided into two main categories, Analysis results and Prediction results. We will discuss analysis results first.

Analysis Results:

These results include the different reports we created to answer different questions about accidents, their causes and how to reduce them. These reports are divided into the following categories.

1. Drill reports.
2. Exploratory reports.
3. Descriptive Reports.
4. Active reports.

Drill reports:

Drill reports include those reports which involve drill through process, meaning that you can go from one report to the other just by clicking on it. We made two drill reports, as follows.

Year Wise Report:

This first drill report is basically a year wise analysis of accidents which is further drilled to quarters, then months and finally to days. As you can see, we cannot say that accidents are decreasing or increasing with the passage of time. For 2009 the accidents are large whereas for 2010 and 2011 they are less, for 2012 the data is till mid of august and if we extrapolate the analysis we see the accidents roughly equal to 2009.

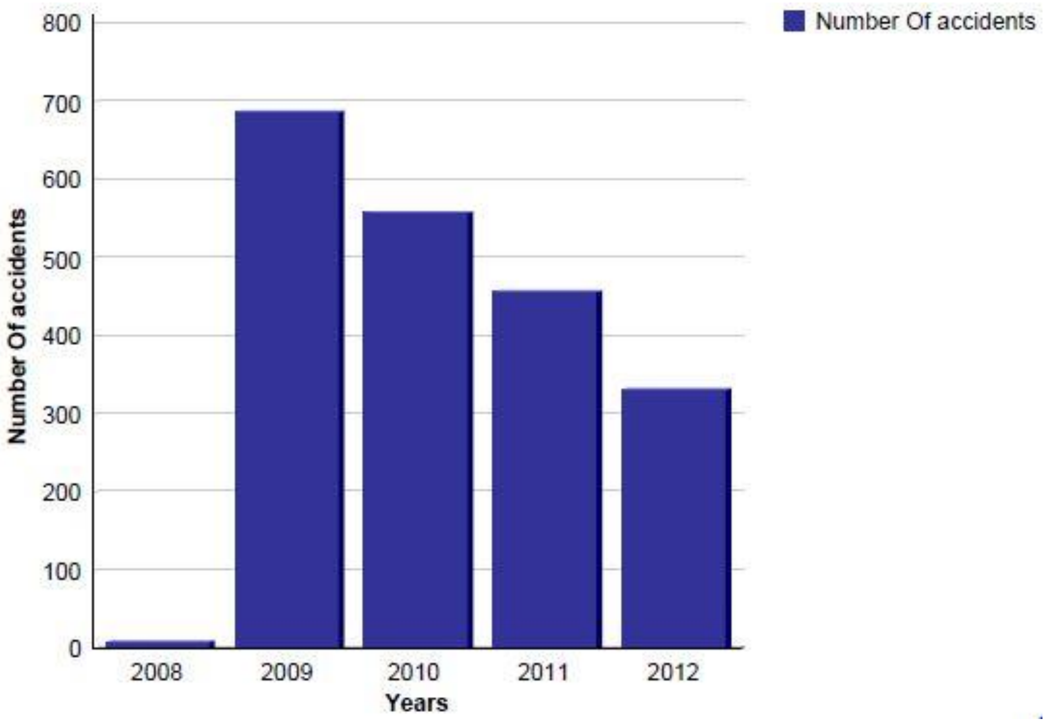


Figure 19: Year Wise

Location Wise Report:

This drill report is about locations, first off is the general distribution of accidents with respect to locations, which then drills down to zones and sectors. The location analysis tells us that the accidents on motorway are very less as compared to national highway this is because of certain extra precautions that the motorway implies while are absent on national highway.

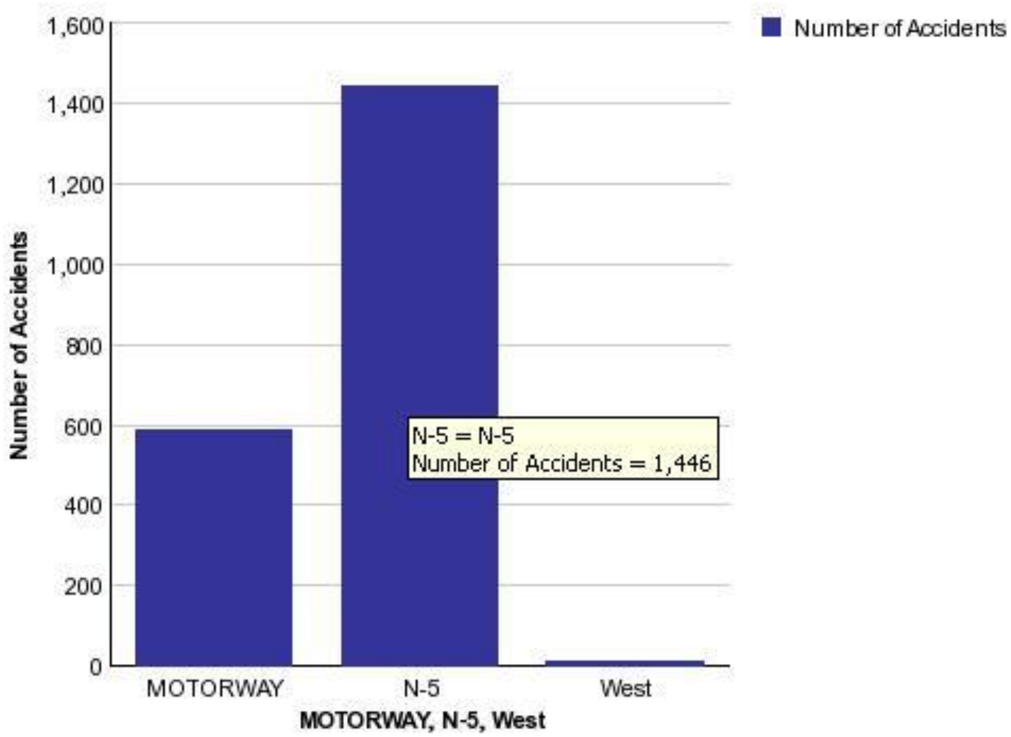


Figure 20: Location Wise

Exploratory Analysis:

It is used when one result lets us explore additional results and in the end it leads us to an answer to the problem. The following reports were made for this purpose.

Dozing at Wheel:

We see from the following report that when the road geometry is straight, dozing at wheel is contributing as an accident cause. This warranted further study and hence we further explored what role dozing at wheel plays.

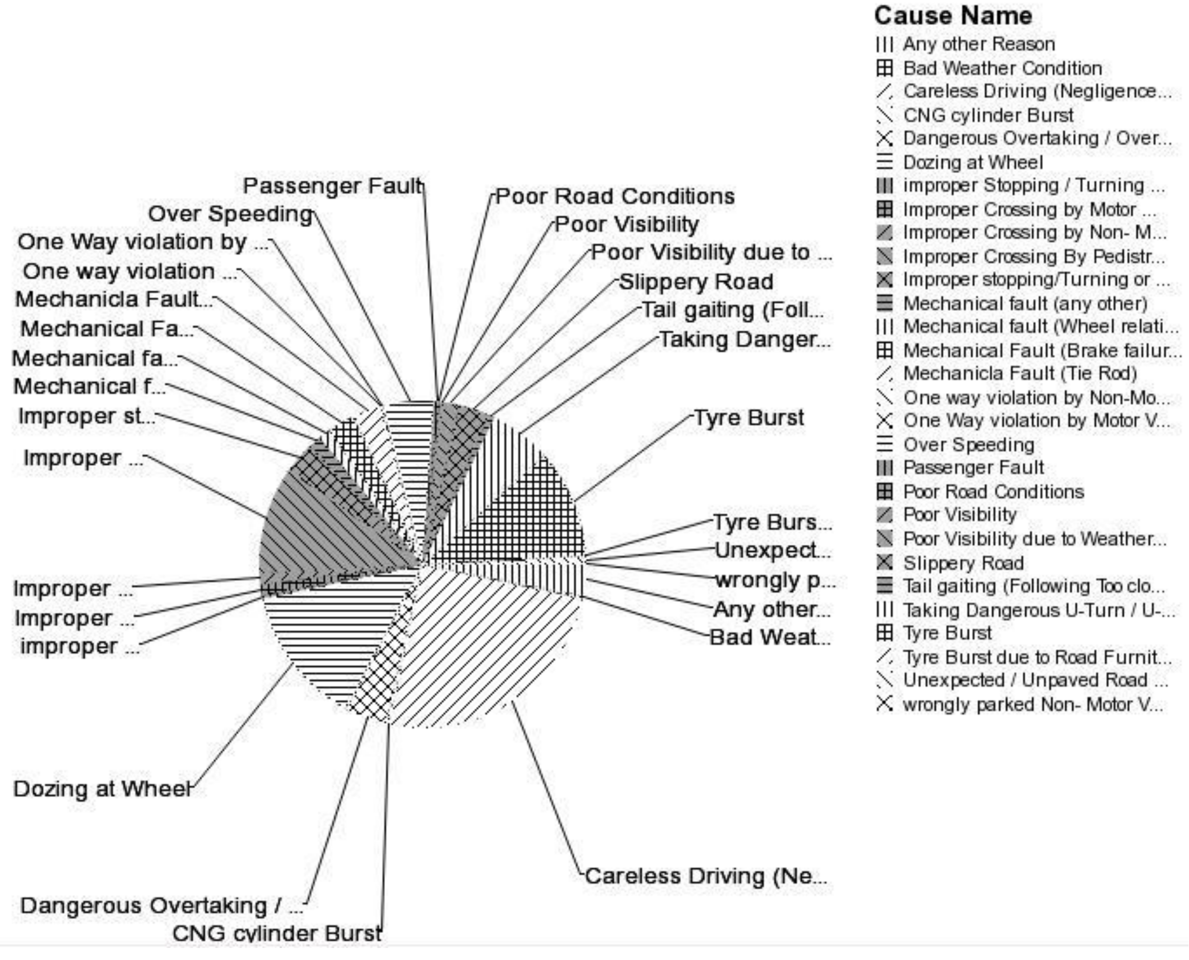


Figure 21: Road Geometry and Cause

Vehicle types and dozing at wheel:

The next stage of exploration was to see if different vehicle types had more dozing at wheel incidents than others, and we found out that heavy transport vehicles like trailers trucks etc. have more cases of dozing at wheel than other vehicle types. We then further explored to see which locations had a large number of accidents due to dozing at the wheel.

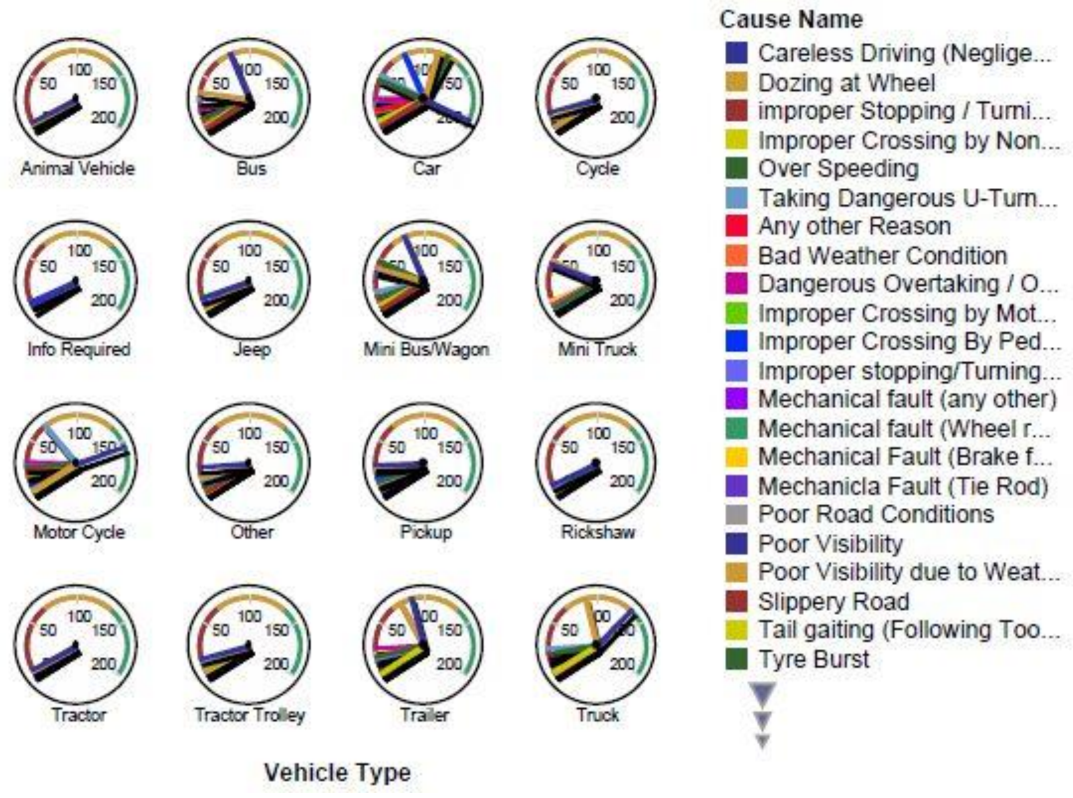
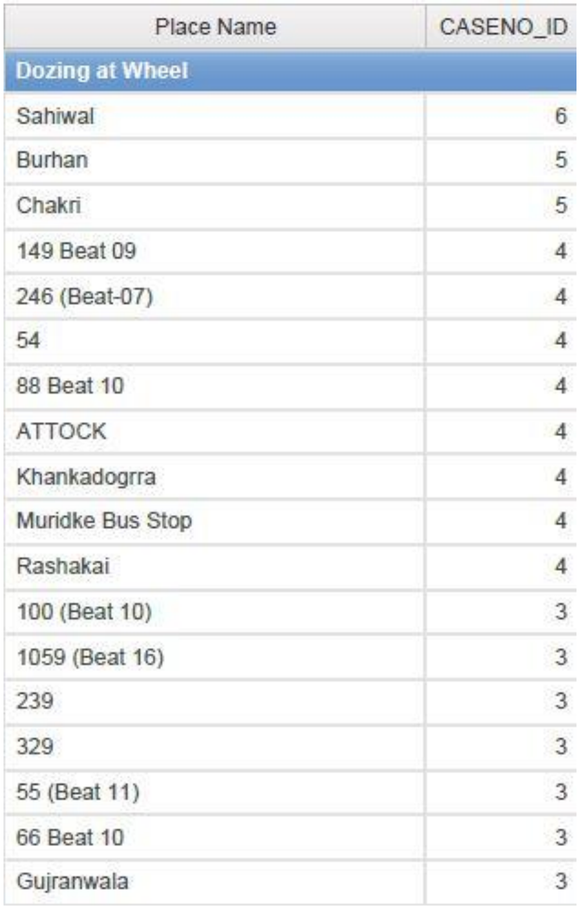


Figure 22: Vehicle Type and Cause

Locations and Dozing at Wheel:

The third report we made was location wise analysis of Dozing at Wheel accidents. So that we might have an idea about the locations which involve dozing at wheel accidents and we found out that a few places as seen from the image have more frequency of these accidents.



Place Name	CASENO_ID
Dozing at Wheel	
Sahiwal	6
Burhan	5
Chakri	5
149 Beat 09	4
246 (Beat-07)	4
54	4
88 Beat 10	4
ATTOCK	4
Khankadogrra	4
Muridke Bus Stop	4
Rashakai	4
100 (Beat 10)	3
1059 (Beat 16)	3
239	3
329	3
55 (Beat 11)	3
66 Beat 10	3
Gujranwala	3

Figure 23: Location Wise dozing

Mechanical Faults:

A report we made was on mechanical faults and the effects it has on different vehicles and if any vehicles have more faults than others. We found that brake failure was a major factor in some of the heavy transport vehicles and hence this warranted further exploration.

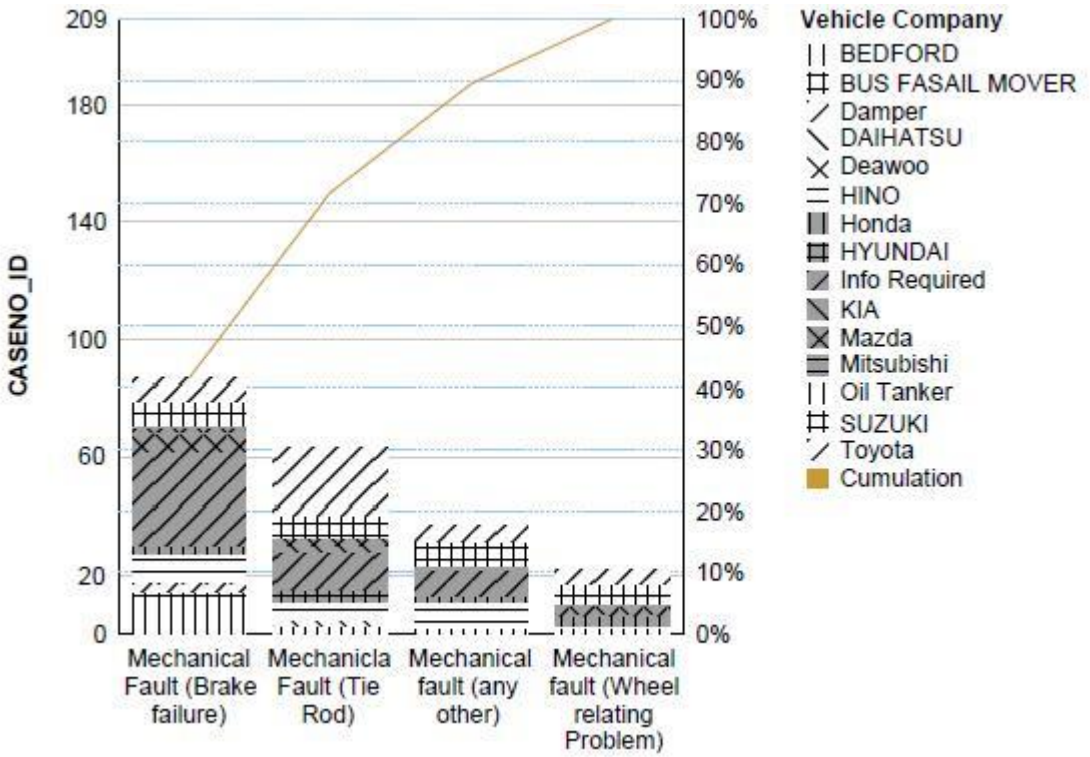
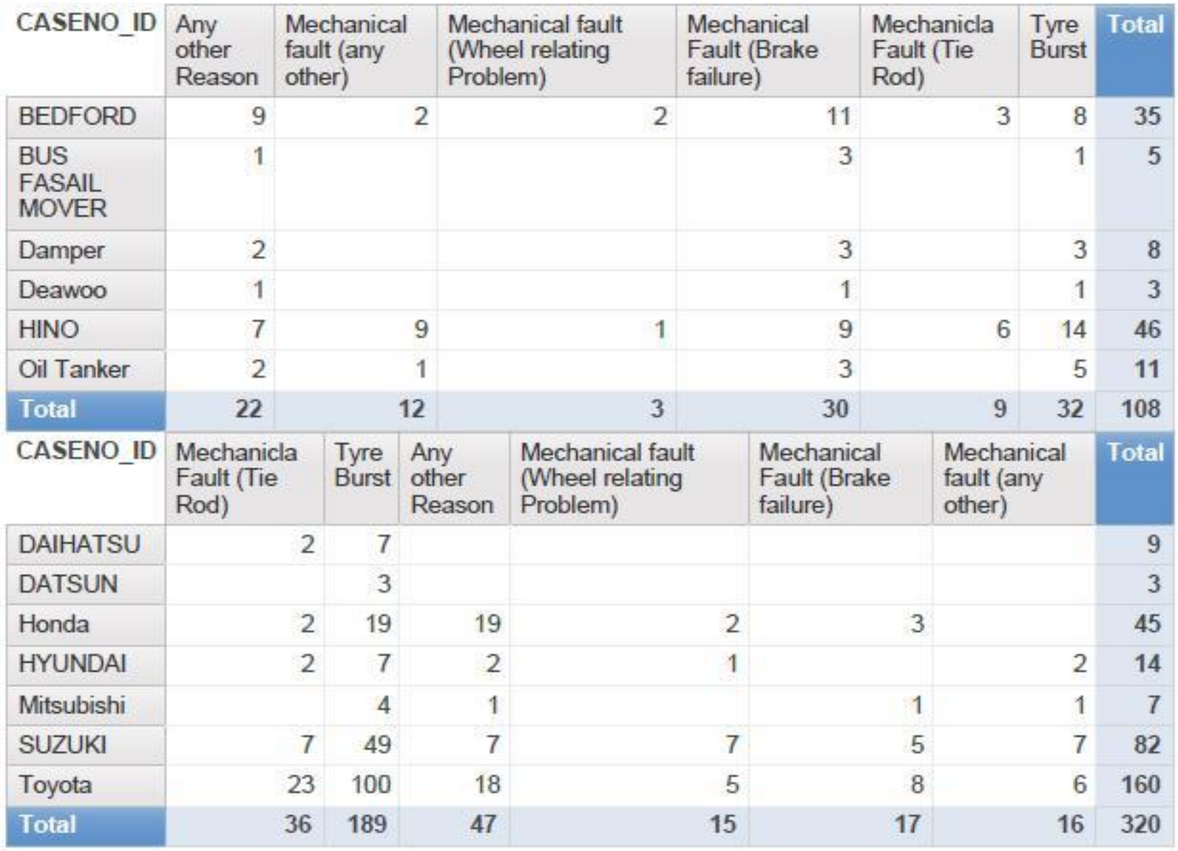


Figure 24: Mechanical Faults in Vehicle Types

HTV vs. LTV:

This report further clarified to us that mechanical faults are more prevalent in heavy vehicles than in light vehicles. The brake failure especially is greater in heavy vehicles hence we came to our second exploratory result.



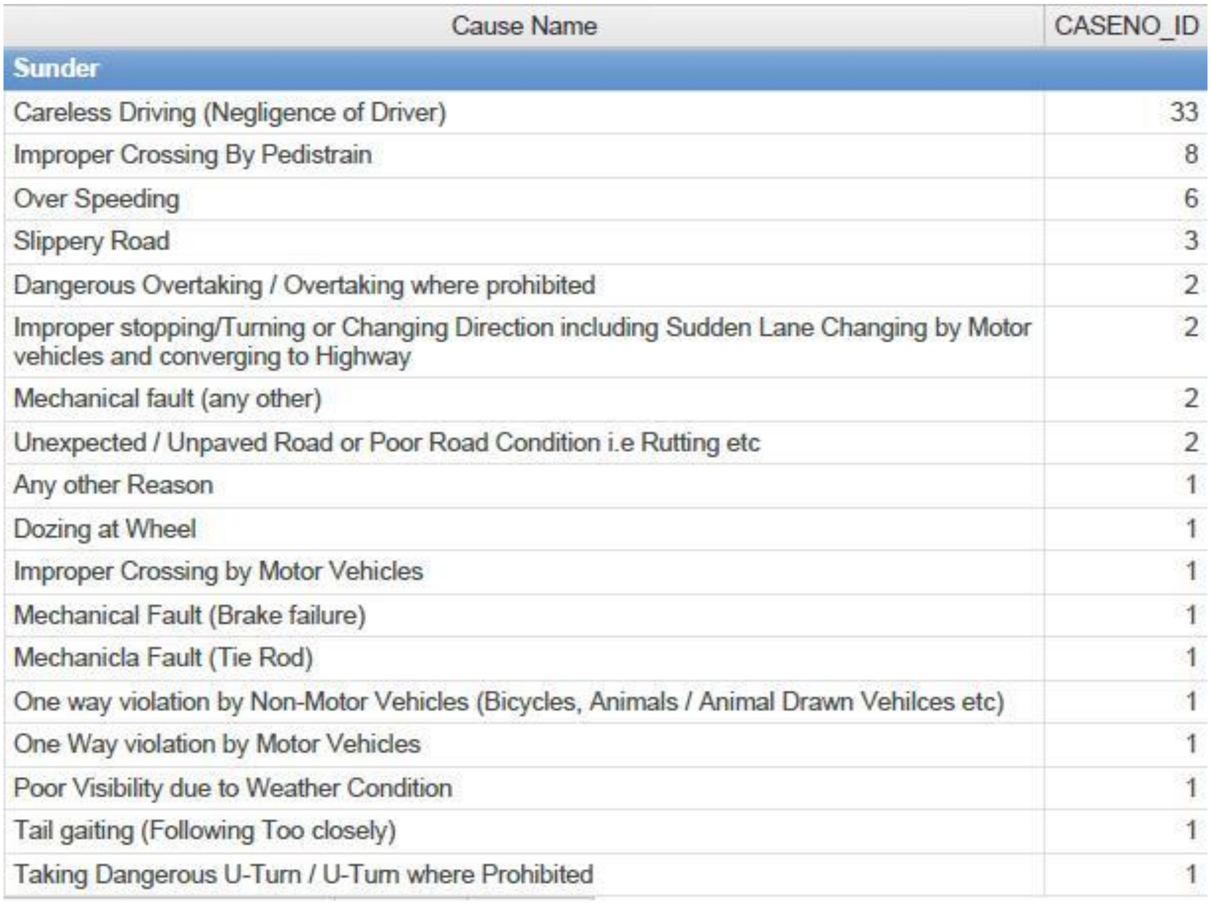
CASENO_ID	Any other Reason	Mechanical fault (any other)	Mechanical fault (Wheel relating Problem)	Mechanical Fault (Brake failure)	Mechanicla Fault (Tie Rod)	Tyre Burst	Total
BEDFORD	9	2	2	11	3	8	35
BUS FASAIL MOVER	1			3		1	5
Damper	2			3		3	8
Deawoo	1			1		1	3
HINO	7	9	1	9	6	14	46
Oil Tanker	2	1		3		5	11
Total	22	12	3	30	9	32	108

CASENO_ID	Mechanicla Fault (Tie Rod)	Tyre Burst	Any other Reason	Mechanical fault (Wheel relating Problem)	Mechanical Fault (Brake failure)	Mechanical fault (any other)	Total
DAIHATSU	2	7					9
DATSUN		3					3
Honda	2	19	19	2	3		45
HYUNDAI	2	7	2	1		2	14
Mitsubishi		4	1		1	1	7
SUZUKI	7	49	7	7	5	7	82
Toyota	23	100	18	5	8	6	160
Total	36	189	47	15	17	16	320

Figure 25: HTV vs LTV

Maximum accidents location wise:

The third exploratory report we made was on the maximum number of accidents to occur in a location. We came to this report via the map results which showed that Sunder has the most number of accidents.

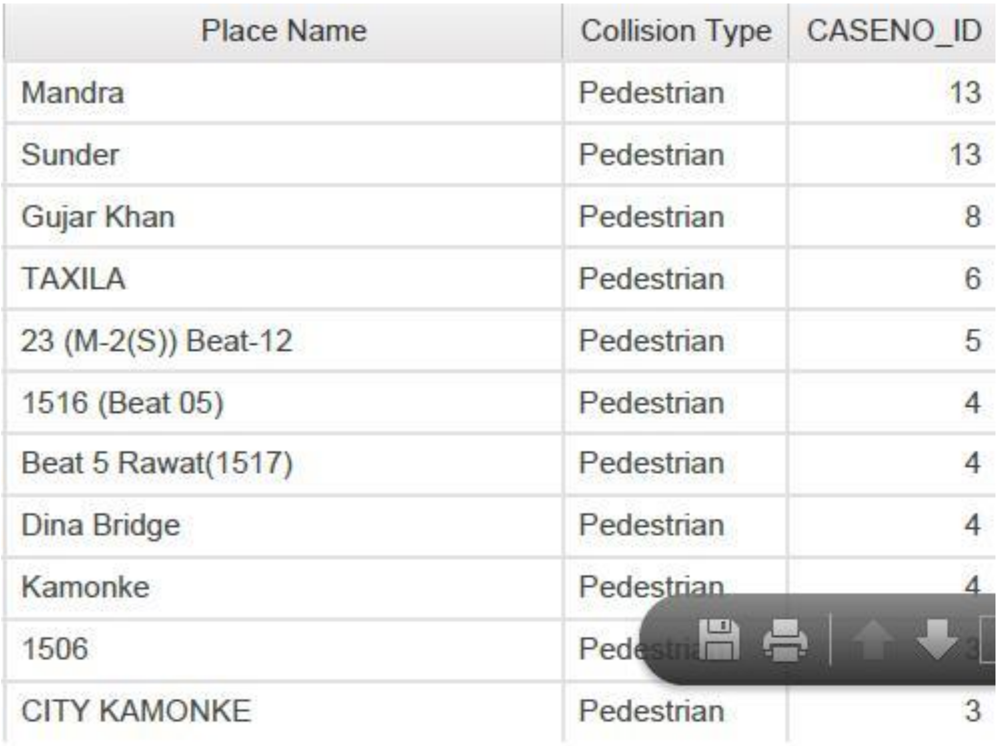


Cause Name	CASENO_ID
Sunder	
Careless Driving (Negligence of Driver)	33
Improper Crossing By Pedistrain	8
Over Speeding	6
Slippery Road	3
Dangerous Overtaking / Overtaking where prohibited	2
Improper stopping/Turning or Changing Direction including Sudden Lane Changing by Motor vehicles and converging to Highway	2
Mechanical fault (any other)	2
Unexpected / Unpaved Road or Poor Road Condition i.e Rutting etc	2
Any other Reason	1
Dozing at Wheel	1
Improper Crossing by Motor Vehicles	1
Mechanical Fault (Brake failure)	1
Mechanicla Fault (Tie Rod)	1
One way violation by Non-Motor Vehicles (Bicycles, Animals / Animal Drawn Vehilces etc)	1
One Way violation by Motor Vehicles	1
Poor Visibility due to Weather Condition	1
Tail gaiting (Following Too closely)	1
Taking Dangerous U-Turn / U-Turn where Prohibited	1

Figure 26: Accident Causes at Sunder

Pedestrian Accidents in Sunder:

Upon further exploration we found out that pedestrian related accidents were the most frequent in Sunder, the reason being that there is no overhead pedestrian bridge, hence a large number of pedestrians walk on the road.

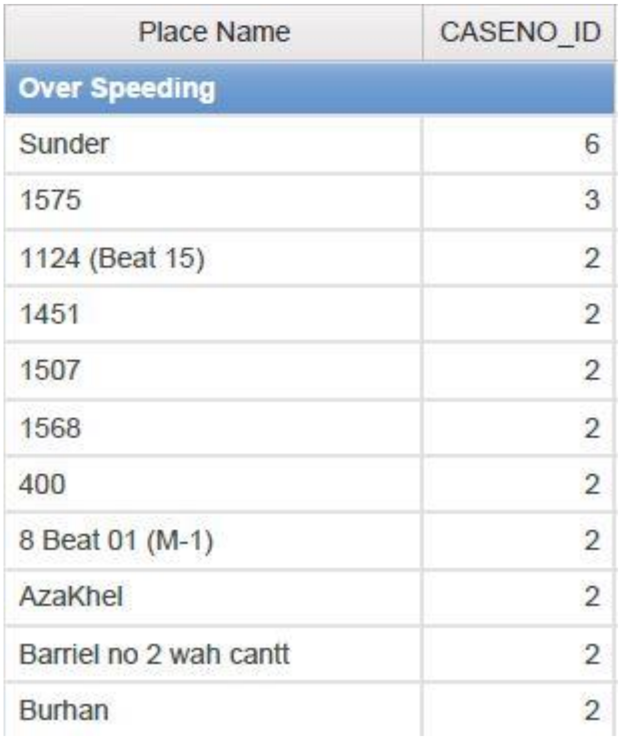


Place Name	Collision Type	CASENO_ID
Mandra	Pedestrian	13
Sunder	Pedestrian	13
Gujar Khan	Pedestrian	8
TAXILA	Pedestrian	6
23 (M-2(S)) Beat-12	Pedestrian	5
1516 (Beat 05)	Pedestrian	4
Beat 5 Rawat(1517)	Pedestrian	4
Dina Bridge	Pedestrian	4
Kamonke	Pedestrian	4
1506	Pedestria	3
CITY KAMONKE	Pedestrian	3

Figure 27: Pedestrian accidents vs Locations

Over speeding incidents in Sunder:

Another result we found out was that over speeding accidents were also the most frequent in Sunder, due to the vehicles there driving very fast. Hence a strict speed check should be enforced in that place to reduce the accidents.



Place Name	CASENO_ID
Over Speeding	
Sunder	6
1575	3
1124 (Beat 15)	2
1451	2
1507	2
1568	2
400	2
8 Beat 01 (M-1)	2
AzaKhel	2
Barriel no 2 wah cantt	2
Burhan	2

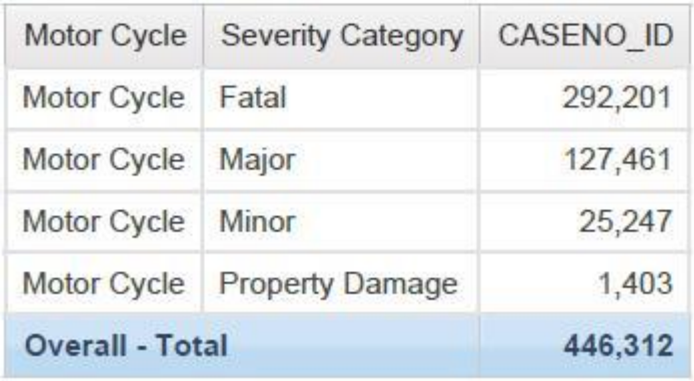
Figure 28: Overspeeding vs Location

Descriptive Reports:

These reports are the basic reports which basically answer a simple question or prove some fact. They can also be made in simple databases. Following are some descriptive reports.

Motorcycle fatality ratio:

During our research we found a survey that motorcycle riders have the most number of fatal accidents, using the data we thought to see whether this was in fact true or not. And we found out that it was true. Motorcycle riders have a fatality rate of 71% higher than any other transport type.

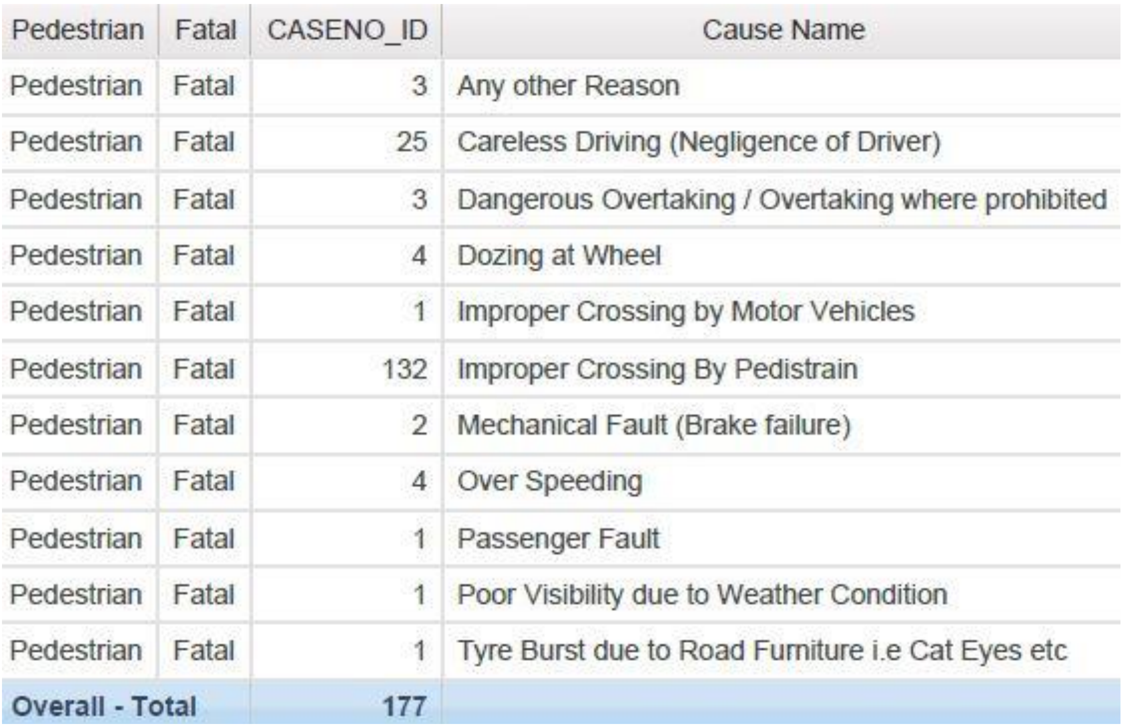


Motor Cycle	Severity Category	CASENO_ID
Motor Cycle	Fatal	292,201
Motor Cycle	Major	127,461
Motor Cycle	Minor	25,247
Motor Cycle	Property Damage	1,403
Overall - Total		**446,312**

Figure 29: Motorcycle fatalities

Pedestrian related accidents:

Another survey we read stated that pedestrians here are not aware of road crossing and safety rules and hence get involved in accidents. We checked this with our data and found out that yes indeed this was the case.



Pedestrian	Fatal	CASENO_ID	Cause Name
Pedestrian	Fatal	3	Any other Reason
Pedestrian	Fatal	25	Careless Driving (Negligence of Driver)
Pedestrian	Fatal	3	Dangerous Overtaking / Overtaking where prohibited
Pedestrian	Fatal	4	Dozing at Wheel
Pedestrian	Fatal	1	Improper Crossing by Motor Vehicles
Pedestrian	Fatal	132	Improper Crossing By Pedistrain
Pedestrian	Fatal	2	Mechanical Fault (Brake failure)
Pedestrian	Fatal	4	Over Speeding
Pedestrian	Fatal	1	Passenger Fault
Pedestrian	Fatal	1	Poor Visibility due to Weather Condition
Pedestrian	Fatal	1	Tyre Burst due to Road Furniture i.e Cat Eyes etc
Overall - Total		**177**	

Figure 30: Pedestrian Accidents Causes

Days of Week report:

Another descriptive report we made was about the comparison of whether the number of accidents change during weekdays or weekends. We found out that Thursday and Friday have the most accidents due to the fact that most travelers of highways are businessmen who get Friday off so they travel on these two days. Then Saturday and Sunday have greater accidents due to government employees traveling mostly on these days and Wednesday has the fewest.

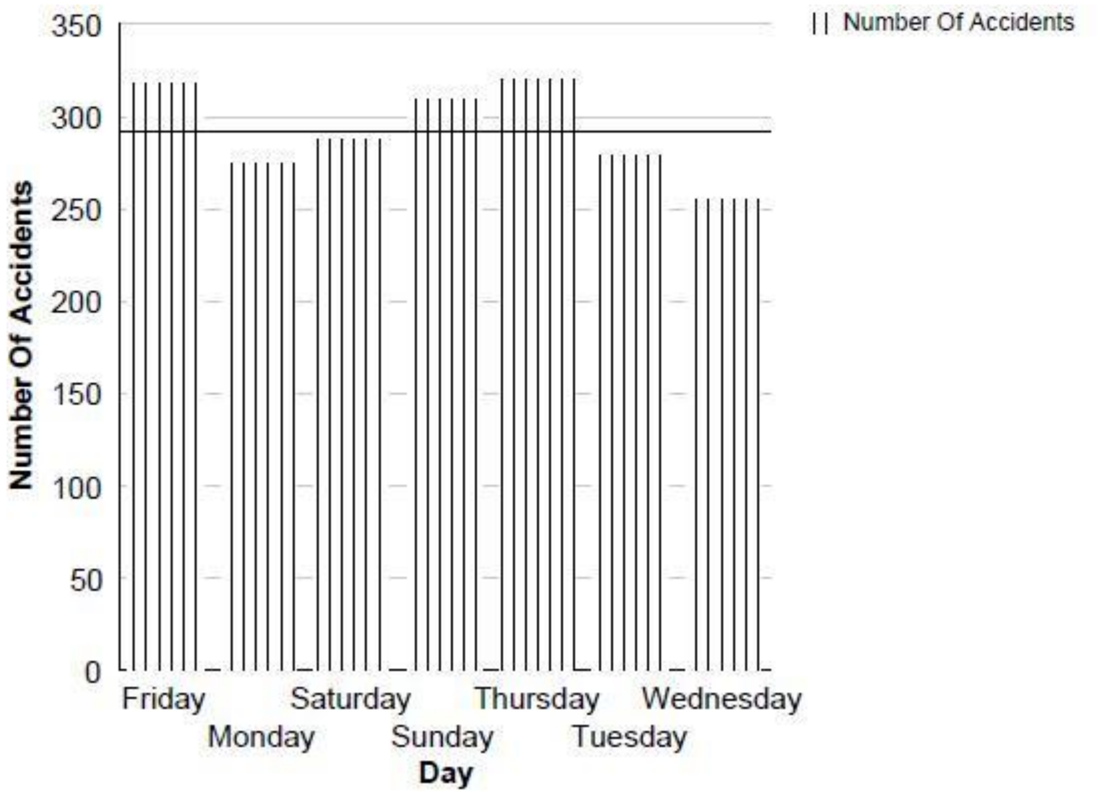
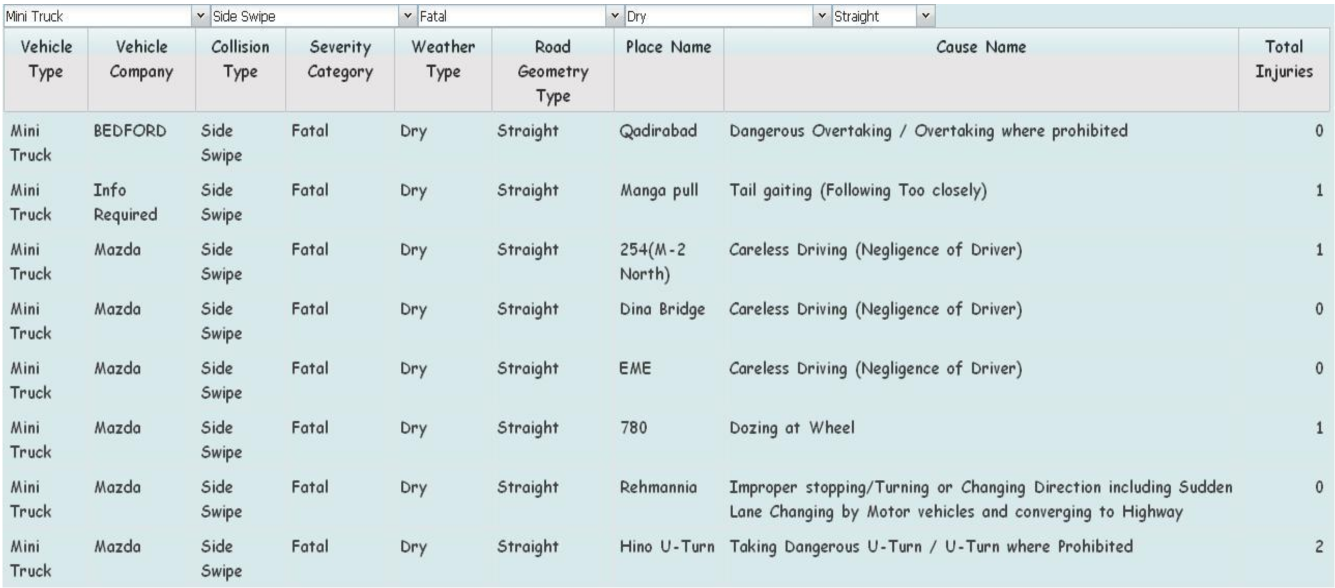


Figure 31: Day of Week Analysis

Active Reports:

These reports are called such because they are interactive, they allow the user to make changes to them and see their result in real time. This report can be used by people to show results and consequences in real time and to do what if analysis. Following report is an active report allowing the user to filter the search according to what he/she feels is best.



Mini Truck | Side Swipe | Fatal | Dry | Straight

Vehicle Type	Vehicle Company	Collision Type	Severity Category	Weather Type	Road Geometry Type	Place Name	Cause Name	Total Injuries
Mini Truck	BEDFORD	Side Swipe	Fatal	Dry	Straight	Qadirabad	Dangerous Overtaking / Overtaking where prohibited	0
Mini Truck	Info Required	Side Swipe	Fatal	Dry	Straight	Manga pull	Tail gaiting (Following Too closely)	1
Mini Truck	Mazda	Side Swipe	Fatal	Dry	Straight	254(M-2 North)	Careless Driving (Negligence of Driver)	1
Mini Truck	Mazda	Side Swipe	Fatal	Dry	Straight	Dina Bridge	Careless Driving (Negligence of Driver)	0
Mini Truck	Mazda	Side Swipe	Fatal	Dry	Straight	EME	Careless Driving (Negligence of Driver)	0
Mini Truck	Mazda	Side Swipe	Fatal	Dry	Straight	780	Dozing at Wheel	1
Mini Truck	Mazda	Side Swipe	Fatal	Dry	Straight	Rehmannia	Improper stopping/Turning or Changing Direction including Sudden Lane Changing by Motor vehicles and converging to Highway	0
Mini Truck	Mazda	Side Swipe	Fatal	Dry	Straight	Hino U-Turn	Taking Dangerous U-Turn / U-Turn where Prohibited	2

Figure 32: Active Report

Prediction Results:

This was done using binary logistic regression. The data we collected was totally categorical. Our dependent variable was categorical and independent variables were categorical as well. Therefore the only statistical model suitable for prediction was binary logistic regression.

Our dependent category was accident severity, our prediction was on whether an accident will be fatal or not. Initially it had multiple categories of fatal major minor property damage. This was giving us less accuracy in prediction so we grouped it into two major categories.

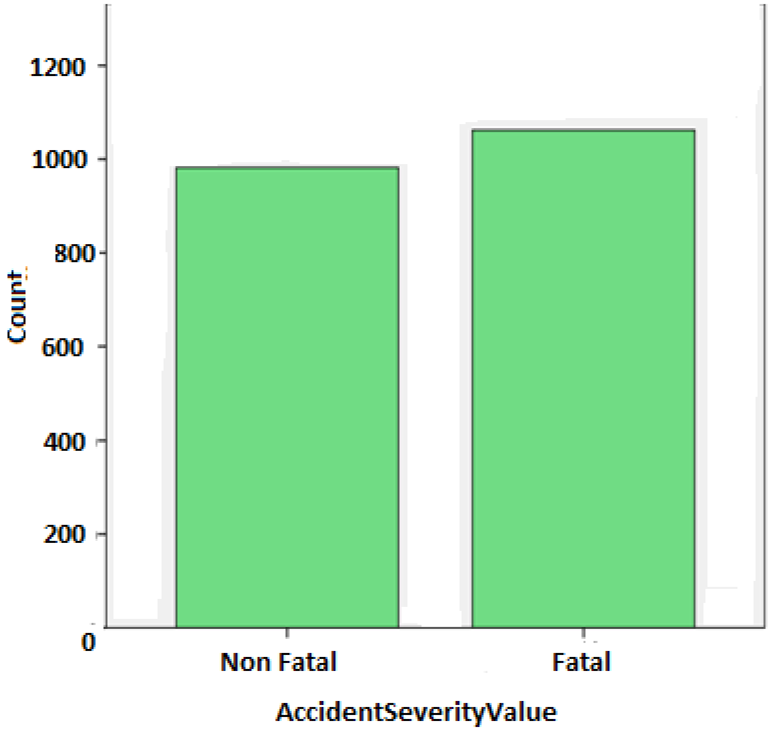


Figure 33: Accident Severity

We also had to group multi category variables like accident causes and collision type. We grouped them as follows.

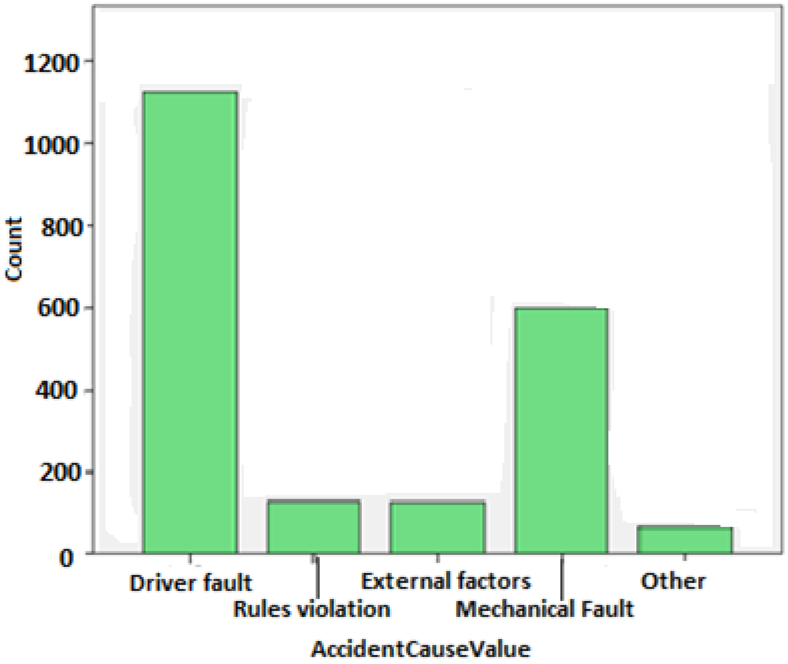


Figure 34: Accident Cause

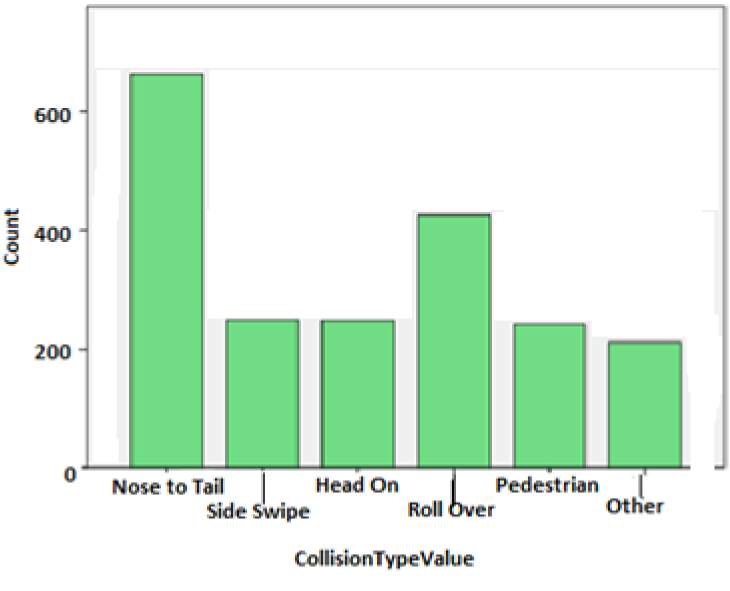


Figure 35: Collision Type

Independent Variable Impact on Accident Severity:

We then check and see whether these variables have any effect on accident severity or not.

Accident Cause Value:

We found that most of the accidents are caused due to the Driver Fault. Further, under the category of driver fault, we see that the bar for fatal accidents is a little higher than for non-fatal. Moreover there are fewer accidents caused due to external factors which include light conditions and weather factors.

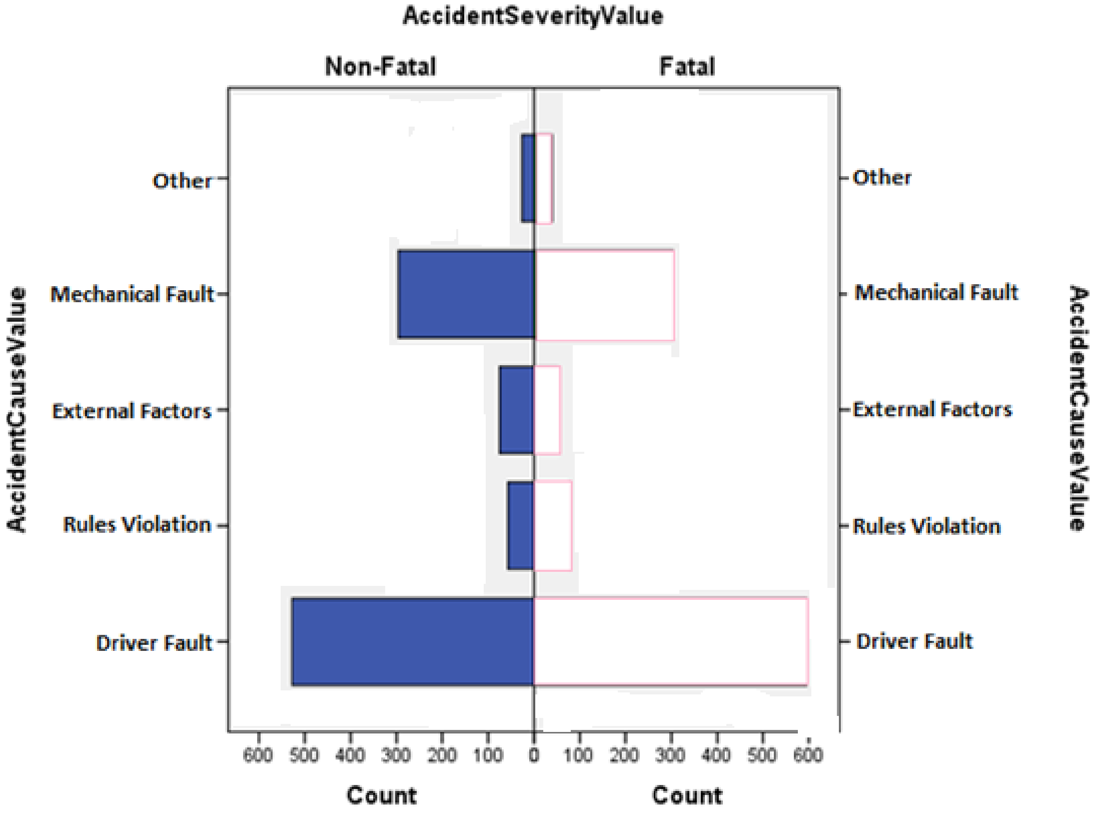


Figure 36: Cause Value

Collision Type Value:

We see that most of the accidents are caused due to the Nose to Tail Collision. Further, more Non-fatal accidents are caused by Roll over collision type than fatal. Additionally, when the vehicles hit the pedestrians, the result is mostly a fatal accident.

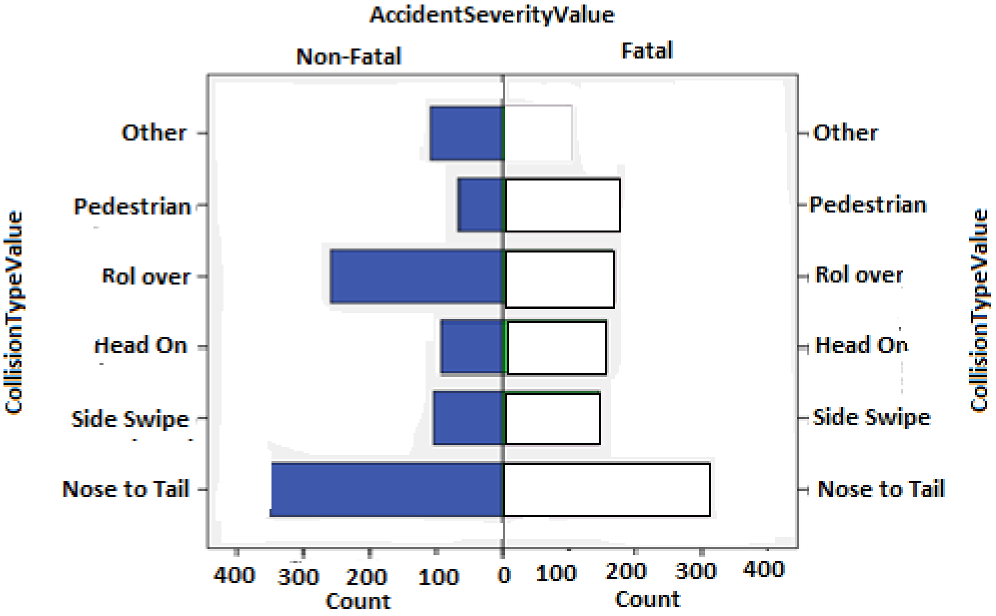


Figure 37: Collision value

Location Value:

We had three cases to consider, motorway name, sector or zone. We found out that motorway name was very generic while sector had too much information to be a good predictor hence we select zone value as a good predictor.

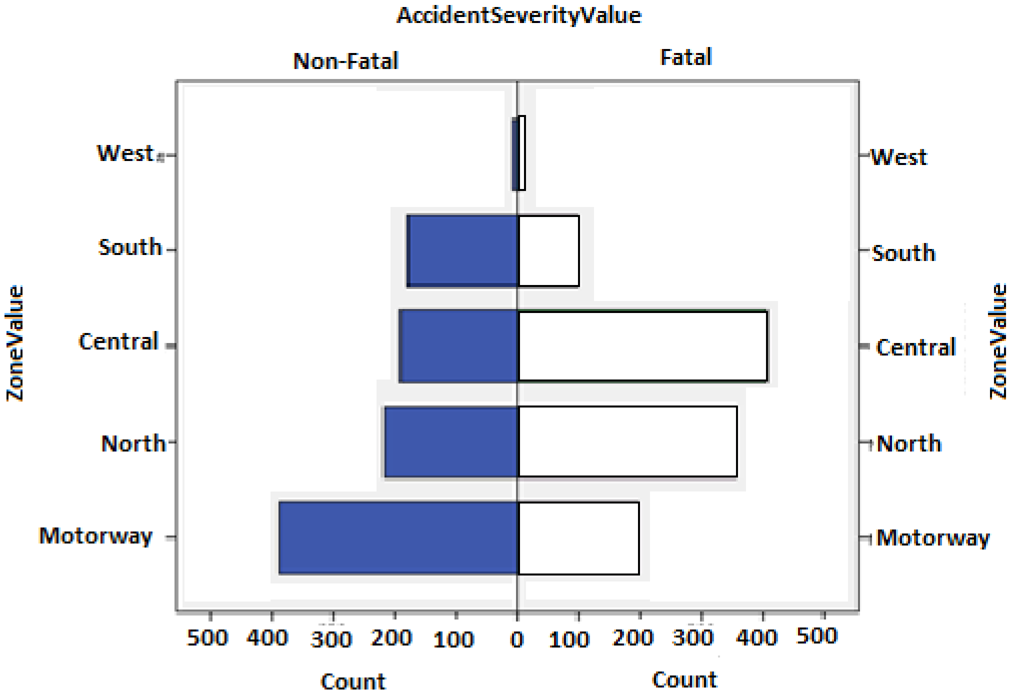
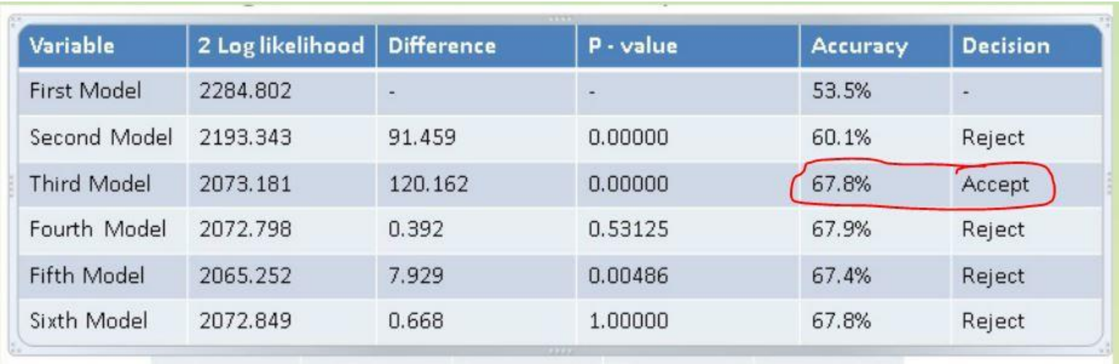


Figure 38: Zone Value

Training Model:

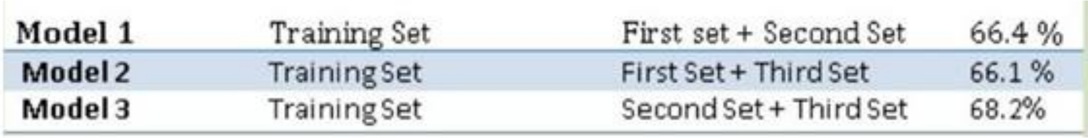
We then trained our model on 80% of data and came to this result.



Variable	2 Log likelihood	Difference	P - value	Accuracy	Decision
First Model	2284.802	-	-	53.5%	-
Second Model	2193.343	91.459	0.00000	60.1%	Reject
Third Model	2073.181	120.162	0.00000	67.8%	Accept
Fourth Model	2072.798	0.392	0.53125	67.9%	Reject
Fifth Model	2065.252	7.929	0.00486	67.4%	Reject
Sixth Model	2072.849	0.668	1.00000	67.8%	Reject

Figure 39: Training Set

We then verified it by using three test models.



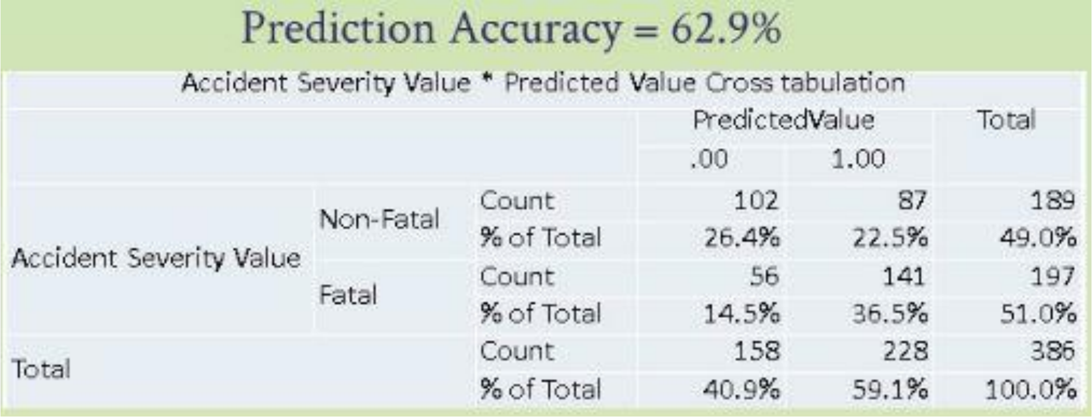
Model 1	Training Set	First set + Second Set	66.4 %
Model 2	Training Set	First Set + Third Set	66.1 %
Model 3	Training Set	Second Set + Third Set	68.2%

Average Accuracy of the Training Sets = 66.9%

Figure 40: Three Set

Model Testing:

We then tested our model for both the training sets. First is the result of 80 % sample model. We tested on 20 % remaining data.



Prediction Accuracy = 62.9%

Accident Severity Value * Predicted Value Cross tabulation

			PredictedValue		Total
			.00	1.00	
Accident Severity Value	Non-Fatal	Count	102	87	189
		% of Total	26.4%	22.5%	49.0%
	Fatal	Count	56	141	197
		% of Total	14.5%	36.5%	51.0%
Total		Count	158	228	386
		% of Total	40.9%	59.1%	100.0%

Figure 41: Prediction Accuracy-20% test

Below is the actual graph predicting about accident severity.

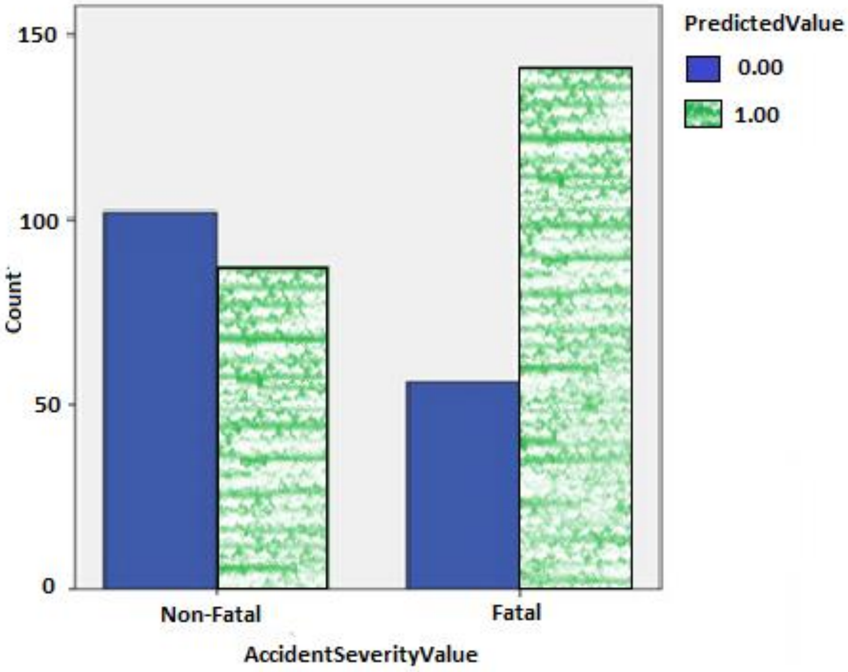


Figure 42: Accuracy of Prediction

Combination of three dataset model testing is just to verify that the results we got were in fact in approximation of these results.

Average accuracy of Test Sets = 65%

	Data Set		Percentage Accuracy
Test Model 1	Training Set	First Set + Second Set	66.4%
	Test Set	Third Set	64.1%
Test Model 2	Training Set	First Set + Third Set	66.1%
	Test Set	Second Set	66.9%
Test Model 3	Training Set	Second Set + Third Set	68.2%
	Test Set	First Set	64%

Chapter 5

CONCLUSIONS & RECOMMENDATIONS

Conclusion:

Automotive accidents are a global phenomenon which is increasing day by day. Pakistan is similarly affected by these accidents and a major portion of deaths are due to automobiles. Therefore a system needs to be in place to keep a check and balance on these accidents so that their frequency may decrease. We took data from National Highways and Motorway police and for the first time in Pakistan concluded reports and facts based on actual data instead of surveys or guesswork.

We identified places and locations where accidents are more frequent. We proved that motorcycle drivers are involved in the most number of fatal accidents, while a large portion of pedestrian accidents occur due to ignorance on road crossing and safety. In fact a large reason of accidents in Pakistan is the unawareness about road safety and driving etiquettes.

Our hope is that these analysis and reports will be a stepping stone in reducing the number of accidents in Pakistan. The scope of this project is unlimited. Our data was limited to motorway and national highway. This project can easily be extended to other parts of the country or cities where the numbers of accidents are even greater like Lahore, Rawalpindi etc.

There is no substitute for life and we should all try to work together in reducing the deaths caused by auto accidents.

Recommendations:

Following recommendations are made so that it might help people in the future.

Change data from categorical to Scale:

This will specially help in prediction as if the data collected is according to a scale then it will form a continuous distribution which will give even better results in prediction analysis.

A standard data gathering format:

World Health Organization has defined specific data gathering parameters which need to be filled when an accident takes place. It comprises of nearly 80 questions but unfortunately for Pakistan that measure cannot be implemented has a whole. So a compromise has to be made and custom based data gathering parameters must be defined which can be utilized here and which are acceptable to W.H.O as well.

Following are the recommendations we prepared for the motorway so that they may incorporate them and help reduce the number of accidents.

Pedestrian Crossings:

Overhead bridges at Mandra and Sunder need to be built and guard rails need to be implemented all along the road to stop pedestrians from crossing and only use the bridges. This would bring reduction in the pedestrian deaths.

Over speeding:

Speed cameras and at least one police car must be present in those locations where over speeding is frequent like Sunder. This would force drivers to be careful and reduce their speeds.

Dozing at Wheel:

Rumble Strips need to be implied on locations like Sahiwal where dozing at the wheel is a prominent cause of accidents. The strips will shake the vehicle strongly waking up the drivers and hopefully avoiding the accident.

Brake checks:

Heavy transport vehicles must be checked for any brake problems before being allowed to travel on an inclined road. If they check their brakes before climbing or descending then it would reduce the accidents.

Chapter 6

References

[1] [Online]. Available: http://www.datawarehousing-concepts.com/index.php/data-warehouse-introduction/definition.html.

[2] [Online]. Available: http://en.wikipedia.org/wiki/Business_intelligence.

[3] [Online]. Available: http://en.wikipedia.org/wiki/Data_warehouse.

[4] [Online]. Available: http://en.wikipedia.org/wiki/Data_mining.

[5] [Online]. Available: http://www.audit-commission.gov.uk/SiteCollectionDocuments/AuditCommissionReports/NationalStudies/Cranfield_Information_use_review.pdf.

[6] [Online]. Available: http://www.cubist-project.eu/index.php?id=438.

[7] [Online]. Available: http://en.wikipedia.org/wiki/Real-time_business_intelligence.

[8] [Online]. Available: http://en.wikipedia.org/wiki/Predictive_analytics.

[9] [Online]. Available: http://www-01.ibm.com/software/analytics/cognos/.

[10] [Online]. Available: http://www-01.ibm.com/software/analytics/spss/.

[11] [Online]. Available: http://en.wikipedia.org/wiki/R_(programming_language).

[12] Hamza Iftikhar, Abdul Ghaffar. Intelligent Systems For Road Safety. FYP Project, School Of Electrical Engineering And Computer Science, NUST, Islamabad, July 2012