Final Report

for

TRAFFIC ACCIDENTS ANALYSIS

Version 1.0 approved

Prepared by GROUP

12 / 08 / 2019

TRAFFIC ACCIDENTS

By

GROUP

DEPARTMENT OF NETWORKS

SCHOOL OF COMPUTING AND INFORMATICS TECHNOLOGY

A Concept Paper submitted to the School of Computing and Informatics Technology

For the Study Leading to a Project Proposal in Partial Fulfillment of the

Requirements for the Award of the Degree of Bachelor of

Software Engineering of Makerere University.

Supervisor

MR. Grace Kamulegeya

Department of networks

School of Computing and Informatics Technology, Makerere University

[**gkamuleg@cit.ac.ug**](mailto:gkamuleg@cit.ac.ug)

July 2019

**1.0 Introduction**

Traffic accidents in particular road accidents are one of the major causes of death, injury and disability in both developed and developing countries. In addition, traffic accidents lead to global economic losses as estimated in road traffic injury costs of US$518 billion per year in the USA. Therefore, the road accident data is necessary not only for statistical analysis, but also in setting priority targets so as to increase the Gross National Products.

**2.0 Background to the Problem**

Road traffic accidents occur on all continents and in every country of the world. Every year they take lives of more than a million people and incapacitate many millions more. Pedestrians, users of non-motorized vehicles including bicycles, carts, and motorcyclists in low-income and middle-income countries carry a large proportion of the global burden of road traffic accidents. According to the World Health Organization’s infographics, nearly 1.25 million people die in road traffic accidents each year, on average 3,287 deaths a day. An additional 20-50 million are injured or disabled. More than half of all accident deaths occur among young adults ages 15-44.

**3.0 Problem Statement**

This project aims at analyzing existing traffic accident data to acquire useful information that can be used in making informed conclusions and supporting decision-making concerning the management of the rampant increase in road traffic accidents. While all decision makings lend their basis to the past information, a more appropriate decision with fewer flaws will be made if this information is more precise and knowledge-based.

**4.0 Objectives**

1. The main objective to analyze road traffic accident data from Denver state using data science techniques.
2. **Other Objectives**

* Visualize data inform of graphical charts, figures, and bars.
* Identify correlations in the data features.

**5.0 Methodology**

Because the objective of the current study is to discover the hidden patterns and present a behavioral model of traffic accident data for the roads of Isfahan Province, a constructive data pipeline methodology is required to be used. In this section, the steps from the beginning of data pipeline operations to the end have been succinctly elucidated.

**5.1 Data set**

The data needed for this project is related to the accident field, the main center for collecting such data is the Denver Police Department. The dataset to be used in the project is annual and related to the years 2013 to 2019. The key features are as shown below;

|  |  |
| --- | --- |
| **Variable** | **Description** |
| FIRST\_OCCURRENCE\_DATE | Date and time accident occurred |
| TOP\_TRAFFIC\_ACCIDENT\_OFFENSE | Type of accident |
| REPORTED\_DATE | Date and time accident were reported |
| INCIDENT\_ADDRESS | Address the accident occurred |
| ROAD\_LOCATION | Location of accident occurred on road |
| ROAD\_DESCRIPTION | Nature of road the accident occurred on |
| ROAD\_CONDITION | Wet/ Dry/ Icy/ Snowy |
| LIGHT\_CONDITION | Daylight/dark lighted/ dark unlighted/dawn or dusk |
| TU1\_VEHICLE\_TYPE | First Type of vehicle involved in the accident |
| TU2\_VEHICLE\_TYPE | Second type of vehicle involved in the accident |

**5.2 Preprocessing and Data Preparation**

This involves importing the libraries and the dataset into the data pipeline, checking out missing values, eliminating unwanted columns of identifying the key features of the data set to successfully manage the data and enhance accuracy of data to be used by the data pipeline.

**5.3 Data visualization and analysis**

Here, the data will be represented graphically in a visual way by use of graphs and charts, such that its easier for the Denver Police Department to see and understand the trends, outliers and patterns in the data. For example, a graph can be plotted to show the nature of accident that occurred the most in a particular year or month to aid the conclusion of which accidents occurred the most in that particular year or month, etc.

**A diagram showing data pipeline architecture**

**DATA**

**LOADING**

**CLEANING**

**VISUALIZE**

**ANALYZE**

**INSIGHT**

.

Data Preprocessing Data visualization and analysis

**6.0 Outcomes**

The obtained results in this project will be significant and can be considered by authorities as invaluable information to be used for decreasing the road accidents

**7.0 References**

<https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-traffic-accidents>

<https://www.asirt.org/safe-travel/road-safety-facts/>

**Software Requirements Specification**

**for**

**TRAFFIC ACCIDENTS ANALYSIS**

**Version 1.0 approved**

**Prepared by GROUP**

**15/07/2019**

Table of Contents

[1. Introduction 3](#_Toc16505354)

[2. Overall Description 3](#_Toc16505355)

[3. External Interface Requirements 6](#_Toc16505356)

[4. System Features 6](#_Toc16505357)

[5. Other Nonfunctional Requirements 6](#_Toc16505358)

1. Introduction

**1.1 Purpose**

The intended purpose of this document is to provide information about the requirements of the data pipeline that analyzes traffic accident data informed insights.

* 1. **Document Conventions**

This document consists of headers and sub headers of font type Calibri, font size 14 and 12 and are bold in nature. The body of this document is in Calibri font type and font size 14.

* 1. **Intended Audience and Reading Suggestions**

The document is designed as a guide for the developing team and the project supervisor for

tracking the ongoing progress.

* 1. **Product Scope**

He system uses attributes from the data on the traffic accidents (accident description, time of accident, nature of weather condition, etc.) to give an insight through analysis. The goal

is to portray understandable patterns and the relationship between the attributes in the

traffic accident data.

* 1. **References**

<https://www.denvergov.org>

2. Overall Description

**2.1. Product Perspective**

The data pipeline is designed to utilize data provided by Denver police department on traffic

accidents as input csv files.

**2.2User Classes and Characteristics**

The use case is to describe the entities of the system as well as the processes.

**1.Actors**

* **The developers**

They are responsible for developing the analytics system and also aid in giving the

interpretation of the obtained results to the police.

* **Denver police organization**

They provide traffic accident data and also receive the output of the data pipeline.

The information they receive from the system would help them make insightful better decisions for example when devising means to curb the crime accidents

**2.Use cases**

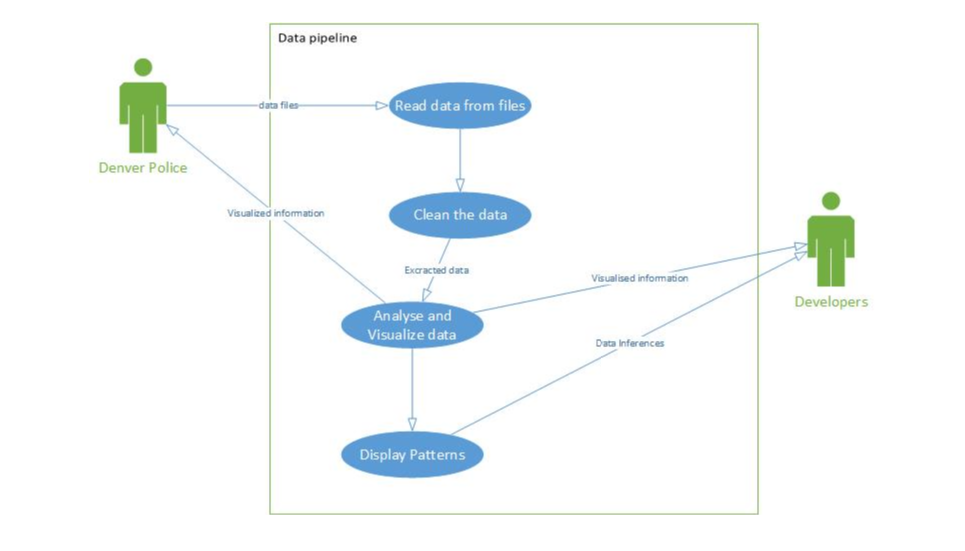
The following use cases are involved ibn our data analytics pipeline:

* **Reading of provide files.** This will involve lo0ading of the csv files containing traffic accident data from Denver police station such that a further processing for example cleaning can be done on the data from those files. Data is read from the csv files and imported into a data frame which is used in the cleaning process.
* **Clean the extracted data.** This is done to acquire the desired key features for example ROAD\_CONDITION, ROAD\_LOCATION by removing unnecessary columns, checking missing values.

This is will involve careful examination of every feature and a thorough understanding of its impact on the desired outcomes of the pipeline for example the impact of the FIRST\_OCCURRENCE\_DATE feature on the time that has more accidents to outcome.

* **Visualization of the cleaned dataset.** It will occur in form of graphs, features and charts i.e. a plot of a bar graph showing the month or the year with the highest number of accidents, which type of accidents occurred most in a particular weather condition and so many others.

**Figure1: Use Case Diagram for the system**

****

**2.3 Assumptions and Dependencies**

* The system uses anaconda tool for all its processing.
* To be able to create elegant and well graphically represented visuals, the system will use the matplotlib package in python

3. External Interface Requirements

**3.1. User Interface**

The system will only generate graphical representations from the input data, from which

meaning will then be relayed to the users.

4. System Features

**4.1. Read csv files.**

4.1.1 Functional Requirements

System needs local access to the necessary csv files.

* 1. **Create graphs from variables** 
     1. Functional requirements
* System needs numerical variables as input to be able to create visualizations.

5. Other Nonfunctional Requirements

**5.1 Performance Requirements**

The system will only extract and use attributers corresponding to the scope of analysis from

the dataset any attributes contained within the datasets do not fit the purpose of the

system will not be used.

Software Design Document

for

TRAFFIC ACCIDENTS ANALYSIS

Version 1.0 approved

Prepared by GROUP

26 / 07 / 2019

**TABLE OF CONTENTS**

1. **INTRODUCTION**
2. **SYSTEM ARCHITECTURE**
   1. Architectural Design
   2. Decomposition Description

List of figures

Figure 2: architecture diagram ……………………………………………………………1

Figure 2.1: Diagram showing the architectural framework of the data pipeline

1. **INTRODUCTION**

To get a better view of the traffic accident data of Denver state, we defined a few questions which we will answer during our data analytic project.

Here is a list of the questions:

1. How has the weather influenced the specific types of accidents?
2. How has the number of traffic accidents changed over the years?
3. Which accidents occur most frequently?
4. During which time has the greatest number of accidents?
5. **SYSTEM ARCHITECTURE**
   1. **Architectural Design**

Data Source

Visualization

Pattern Discovery

Import Libraries

Import Data Set

Data Cleaning

Analysis

Post Processing

Data Pre - Processing

**Figure 3: Diagram showing the components of the data pipeline**



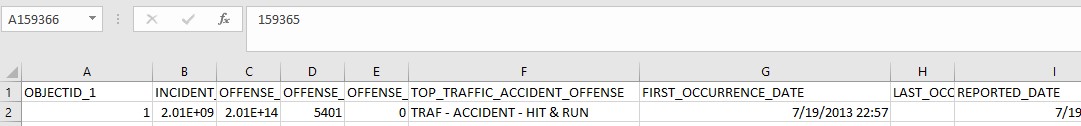
**Figure 3.1: Diagram showing the architectural framework of the data pipeline**

* 1. **Decomposition Description**
     1. **Data source**

This is where the data on traffic accidents that will be used by the data pipeline is located i.e. the Denver mile high city website. The data is downloaded from the website and comes in a CSV file format. The name of the downloaded file is ‘*traffic\_accidents.csv’.*

The traffic\_accidents CSV file contains 159366 rows and 47 features with a total

size of 73.6 megabytes



**3.2.2 Data Pre-Processing**

**Import Libraries**

At this stage, all the necessary libraries and packages in python needed to run the system will be imported using the syntax

Import *libraryname* or *packagename*

The various libraries imported will enable the carrying out of various functions like dataset importing, data visualization.

The libraries to be imported include

* *NumPy*

We are using this library due to its provision of high-performance manipulation of sequences of homogenous data items. The library will be represented as np in the code using the ‘*as’ function as shown below*

*‘Import numpy as np’*

* *Pandas*

It can present data in a way that is suitable for data analysis via its Series and Data Frame data structures. The package contains multiple methods for convenient data filtering.Pandas has a variety of utilities to perform Input/output operations seamlessly. The library will be represented as pd in the code using the ‘*as’ function as shown below*

*‘import pandas as pd’*

* *Seaborn*

This library will be used to carry out statistical data visualization

‘*import seaborn*

* *Matplotlib*

The library will be used because it is more efficient in terms of data visualization and easy to use. A sub-package of the library named pyplot will be used and is represented as plt as shown below

‘*import matplotlib. pyplot as plt’*

Language used:

**python**

**Import Dataset**

Here the data that is stored in the CSV file will be imported to the data frame using the imported pandas’ library. In the panda’s library, the ‘*read\_csv’* function is used to import the data in the CSV file into the data frame. Below is the line of code

‘*pd. read\_csv(‘filename’)*

**Data frame**

A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column. Variables in the data frame will come from the names of the columns in the dataset and rows will contain information in the rows of the dataset. The data frame will be stored in a variable. Operations on data occur on the data frame.

**Missing data and filter out variables**

At this stage, we check for the columns with missing data and delete those columns because the existence of rows with missing data can cause errors. After handling missing values, we identify the required variables, and the unwanted variables are dropped from the dataset.

* + 1. **Analysis**

*Pattern discovery*

Under this component of the pipeline, we suggest the path to the solution to the questions listed in the introduction.

* How has the weather influenced the specific types of accidents?

We use the TOP\_TRAFFIC\_ACCIDENT\_OFFENSE and ROAD\_CONDITION columns to find out the frequency of the specific type of accidents in a particular weather condition.

* How has the number of traffic accidents changed over the years?

We shall perform a count on the TOP\_TRAFFIC\_ACCIDENT\_OFFENSE column for every year in the data set.

* Which accidents occur most frequently?

We shall perform a count on the TOP\_TRAFFIC\_ACCIDENT\_OFFENSE column based on the different accident types present in the data set.

* During which time has the greatest number of accidents?

The FIRST\_OCCURRENCE\_DATE feature contains the time and data that will be used to extract the time ranges and obtain the frequency of accidents in those time ranges.

**3.2.3 Visualization of variables**

The filtered variables from the preprocessing step will be used to create better more meaningful forms of data. The relationship between variables will be established i.e. the relationship between two variables *year* and *accident-type* can be found by plotting to find out the correlation between these variables. These components shall then be presented on a dashboard.

Two libraries will be used in the visualization component i.e.

* Matplotlib.pyplot
* Seaborn – used for statistical visualization

Visual components to be used

* Scatter plot

It is the best way of showing the non-linear pattern in variables i.e. FIRST\_OCCURRENCE\_DATE and TOP\_TRAFFIC\_ACCIDENT\_OFFENSE

Scatter plots will be used to show the number of accidents that occurred in a given time range

* Bar charts and pie charts

We used bar charts for the following reasons i.e.

* Summarizes a large data set in a visually appealing format i.e. we can show the number of times each specific accident occurred throughout the years
* clarify trends in data variables
* Histograms

We choose to use a histogram because it allows the visualization of the distribution of data.

**Implementation Report**

**for**

**TRAFFIC ACCIDENTS ANALYSIS**

**Version 1.0 approved**

**Prepared by GROUP**

**10 / 08 / 2019**

Table of Contents

[Introduction 3](#_Toc16504384)

[How has the weather influenced the specific types of accidents? 3](#_Toc16504385)

[How has the number of traffic accidents changed over the years 4](#_Toc16504386)

[Which accidents occur most frequently? 5](#_Toc16504387)

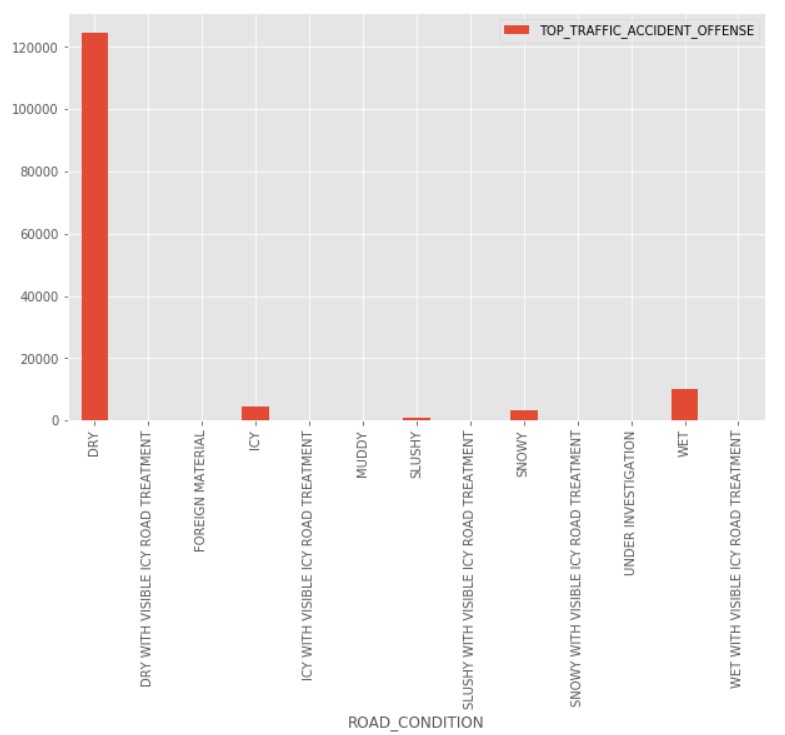
[Days 5](#_Toc16504388)

[During which time has the greatest number of accidents? 6](#_Toc16504389)

Introduction

To clearly understand the traffic accident data of Denver state, we answer the few questions we described in the software design document (SDD) with the help of visualizations.

How has the weather influenced the specific types of accidents?



According to the bar graph, it can be concluded that more accidents happen during the dry season than during any other weather condition, because many drivers would like to drive under clear conditions hence many cars on road and many accidents. It is also evident that drivers also drive under tricky conditions which may lead to accidents that may not have happened if they had used public means. This indicates that people don’t know how to drive under bad weather conditions. Therefore, we would suggest that the Denver police department puts up programs to train public drivers that will drive under bad weather conditions in order to eradicate accidents that happen during those conditions (ICY, SNOW, WET).

How has the number of traffic accidents changed over the years

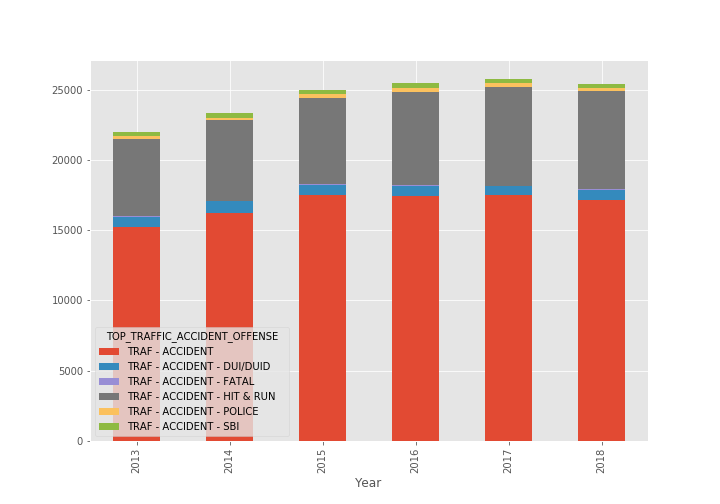
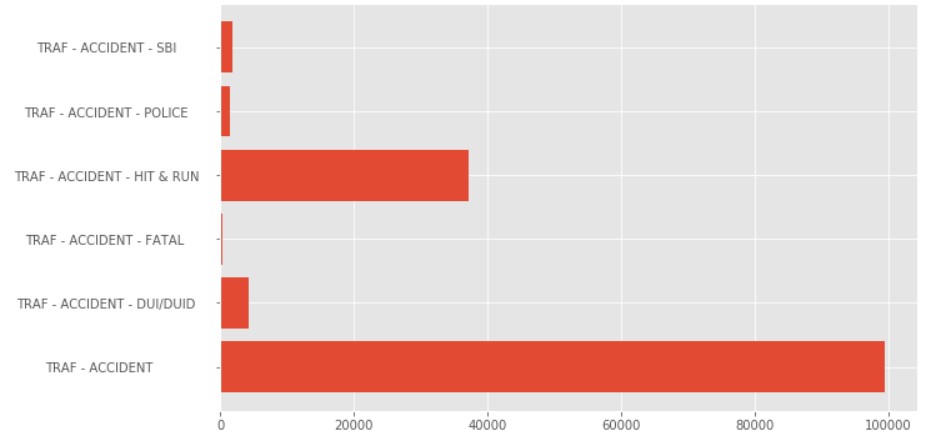


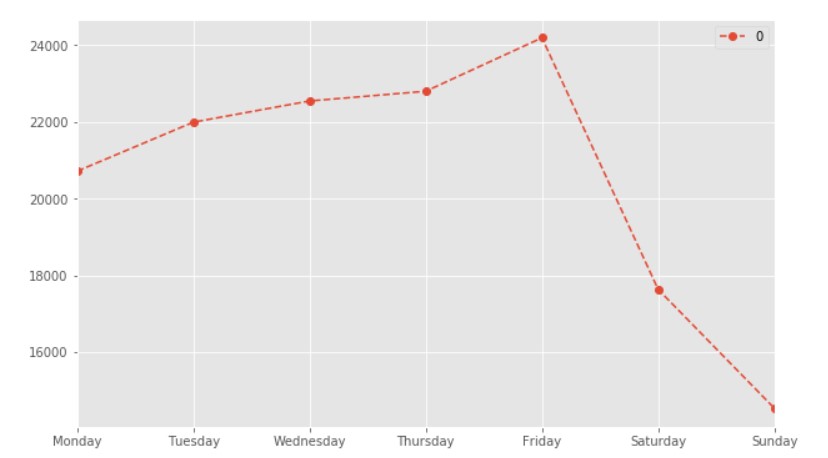
Figure above shows a stacked bar chart which explains the occurance of a particular accident type in a year. Taking an example of the bar representing the year 2013, the bar is divided into 6 colors with each color representing a particular accident type. The height each color fills up in bar, shows the number of accidents that have happened for that particular accident. The color red represents other traffic accidents, the color blue represents DUI/DUID, grey for hit & run, yellow for police and light green for SBI.

Which accidents occur most frequently?



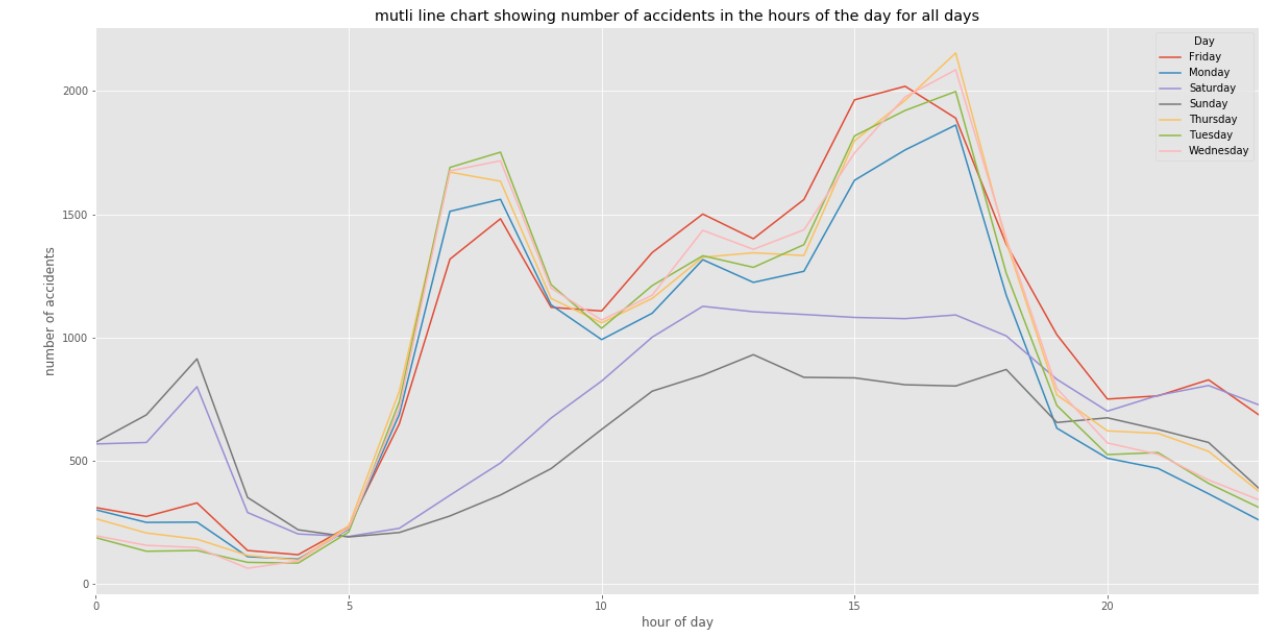
TRAF-ACCIDENT (small minor accidents) are recorded most frequently. It is therefore evident that Denver state has a lot of incompetent drivers which causes such many accidents, we therefore recommend that the traffic department of Denver state tightens the system used to issue out driving licenses to drivers.

Days



From the line graph it is evident that the number of accidents increase as the week progresses from Monday to Thursday and reach the peak on Friday then decreases through the weekend most probably because they drive a lot during the week going to work and also taking kids to the school. Friday has the greatest number of accidents most probably because drivers resort to drinking as a way of relieving stress they endured during the working days, and so may end up driving under the influence of alcohol leading to more accidents on Friday

During which time has the greatest number of accidents?

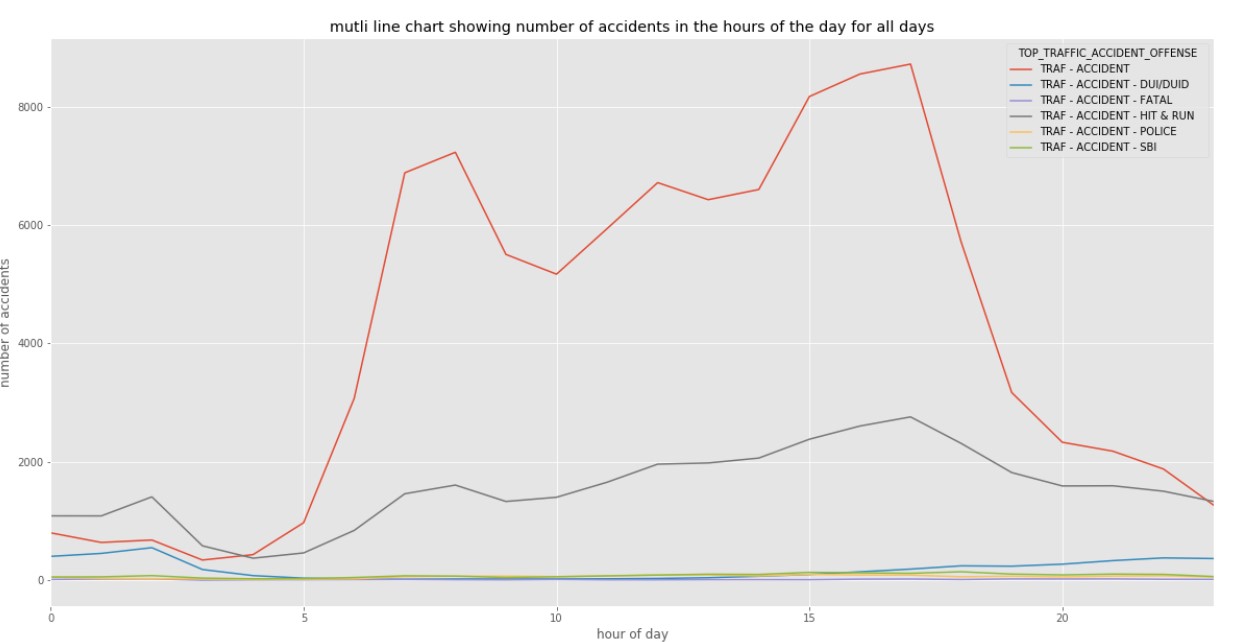


The numbers spike on Monday, Tuesday, Wednesday, Thursday and Friday between 3Pm and 5pm and drop between 8pm and 5am after which they rise. This indicates that more people drive between 3pm and 5pm which is attributed to the fact that people are going to work. The number of accidents on weekends Saturday and Sunday are less are the only spikes appear between 1am and 2am after which they decline gradually until they rise again at 5am.

This indicates that there are more drunkard drivers on the weekend between 1am and 2am, since most the people are not coming from work during that time period. This can be evidenced from the graph showing number of accidents vs time, there is a spike in the TRAF-ACCIDENT -DUI/DUID (drive under influence) in the time between 12am and 5am when most people are traveling from bars.

It would therefore be relevant if the Denver police department put patrol cars between 1am and 5am in order to reduce on the accidents that curb lives between those hours.

Hit and Run accidents are more evident in the morning hours when people are going to work and after work hours in the evening when people are leaving work, this indicates that these accidents are a result of traffic. Therefore, the police should find measures of reducing traffic in those hours in order to reduce on the number of accidents in those hours.



**

MAKERERE UNIVERSITY

COLLEGE OF COMPUTING AND

INFORMATION SCIENCES

**BSE 2301 PROFESSIONAL SOFTWARE ENGINEERING**

**MINI PRACTICAL**

**PROJECT II**

**SUPERVISOR** GRACE KAMULEGEYA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NAMES** | **Reg No** | **Stud No** | **Signature** |
| 1 | KAMASU PAUL | 17/U/4521/EVE | 217017362 |  |
| 2 | KIYINGI ALOYSIOUS | 17/U/5344/PS | 217002946 |  |
| 3 | NAKAYENGA VIOLA | 17/U/7565/PS | 217003981 |  |
| 4 | KAAYA MARVIN | 17/ U/4253/EVE | 217017391 |  |