.



MAKERERE UNIVERSITY

COLLEGE OF COMPUTING AND

INFORMATION SCIENCES

**BSE 2301 PROFESSIONAL SOFTWARE ENGINEERING MINI PRACTICAL**

**PROJECT II**

**SOFTWARE DESIGN SPECIFICATION DOCUMENT**

**SUPERVISOR** GRACE KAMULEGEYA

|  |  |  |  |
| --- | --- | --- | --- |
|  | **NAMES** | **Reg No** | **Stud No** |
| 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 |

**Traffic Accidents Analysis**

# Software Design Document

Date: (26th/July/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

### 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?

### 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**

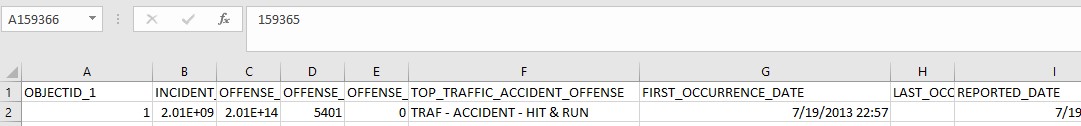
## Decomposition Description

## 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.