**1. Reducing Traffic Mortality in the USA**

**1.1 Introduction**

Road accidents kill thousands of people each month in the United States, and thousands each week in the world. They injure more thousands each week. The majority involve a car but there are also many bus, truck  and bicycle accidents. Aircraft crashes, ship sinking's and other transport accidents also attract much attention.

A number of factors contribute to the risk of collision, including [vehicle design](https://en.wikipedia.org/wiki/Vehicle_design), speed of operation, [road design](https://en.wikipedia.org/wiki/Road_design), road environment, and driver skill, [impairment due to alcohol or drugs](https://en.wikipedia.org/wiki/Driving_under_the_influence), and behavior, notably distracted driving, [speeding](https://en.wikipedia.org/wiki/Speed_limit_enforcement) and [street racing](https://en.wikipedia.org/wiki/Street_racing). Worldwide, [motor vehicle](https://en.wikipedia.org/wiki/Motor_vehicle) collisions lead to death and disability as well as financial costs to both society and the individuals involved.

**1.2 Objective of Research**

While the rate of fatal road accidents has been decreasing steadily since the 80's,the past ten years have seen a stagnation in this reduction. Coupled with the increase in number of miles driven in the nation, the total number of traffic related fatalities has now reached a ten year high and is rapidly increasing. More than one-half of all road traffic deaths globally occur among people ages 15 to 44—their most productive earning years. Moreover, the disability burden for this age group accounts for 60 percent of all DALYs lost because of road traffic accidents.

**1.3 Problem Statement**

By looking at the demographics of traffic accident victims for each US state, we find that there is a lot of variation between the states. Now we want to understand if there are patterns in this variation in order to derive suggestions for a policy action plan. In particular, instead of implementing a costly nation-wide plan we want to focus on groups of states with similar profiles.

**1.4 Industry Profile**

Industry profiles are in-depth documents that give insight into an industry, where it came from, and where it appears to be going. A typical report looks at the industry leaders, forces affecting the industry and financial data for the industry.

**2. Review of Literature**

Road traffic accidents-the leading cause of death by injury and the tenth-leading cause of all deaths globally now make up a surprisingly significant portion of the worldwide burden of ill-health. An estimated 1.2 million people are killed in road crashes each year, and as many as 50 million are injured, occupying 30 percent to 70 percent of orthopedic beds in developing countries hospitals. And if present trends continue, road traffic injuries are predicted to be the third-leading contributor to the global burden of disease and injury by 2020.

In developed countries, road traffic death rates have decreased since the 1960s because of successful interventions such as seat belt safety laws, enforcement of speed limits, warnings about the dangers of mixing alcohol consumption with driving, and safer design and use of roads and vehicles. For example, road traffic fatalities declined by 27 percent in the United States and by 63 percent in Canada from 1975 to 1988. But traffic fatalities increased in developing countries during the same period—by 44 percent in Malaysia and 243 percent in China, for instance.

As in developed countries, driver impairment is an important component of road traffic accidents in developing countries. Driving at excess speeds, while under the influence of alcohol or drugs, while sleepy or tired, when visibility is compromised, or without protective gear for all vehicle occupants are major factors in crashes, deaths, and serious injuries

**3. Data Collection**

Per request of the US Department of Transportation, we are currently investigating how to derive a strategy to reduce the incidence of road accidents across the nation. By looking at the demographics of traffic accident victims for each US state, we find that there is a lot of variation between states. Now we want to understand if there are patterns in this variation in order to derive suggestions for a policy action plan. In particular, instead of implementing a costly nation-wide plan we want to focus on groups of states with similar profiles. How can we find such groups in a statistically sound way and communicate the result effectively?

To accomplish these tasks, we will make use of data wrangling, plotting, dimensionality reduction, and unsupervised clustering.

The data given to us was originally collected by the National Highway Traffic Safety Administration and the National Association of Insurance Commissioners. This particular dataset was compiled and released as a CSV-file by Five Thirty Eight under the CC-BY4.0 license

**4.** **Methodology**

**4.1 Exploratory Data Analysis**

We now have an idea of what the dataset looks like. To further familiarize ourselves with this data, we will calculate summary statistics and produce a graphical overview of the data. The graphical overview is good to get a sense for the distribution of variables within the data and could consist of one histogram per column. It is often a good idea to also explore the pair-wise relationship between all columns in the data set by using a using pair-wise scatter plots (sometimes referred to as a "scatter plot matrix").

4.1.1 Figures and table

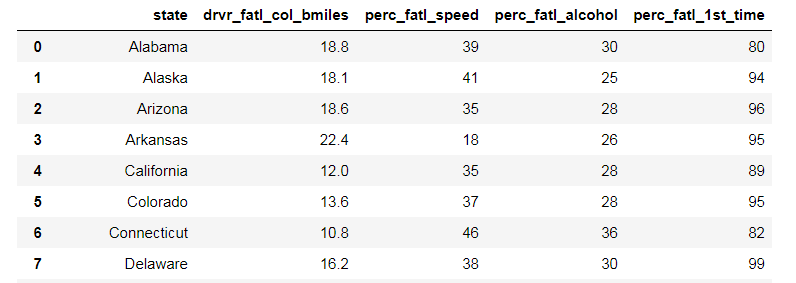


FIG 4.1.1.1 : Tables showing the road accidents occurring in particular due to different reasons like over speeding, consumption of alcohol.

|  |  |  |  |
| --- | --- | --- | --- |
| drvr\_fatl\_col\_bmiles = Number of drivers involved in fatal collisions per billion miles (2011) |  |  |  |
| perc\_fatl\_speed = Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding (2009) |  |  |  |
| perc\_fatl\_alcohol = Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired (2011) | |  |  |
| perc\_fatl\_1st\_time = Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In  any Previous Accidents (2011)  C:\Users\LAB USER\Pictures\new table 2.PNG  FIG 4.1.1.2: Table showing the correlation between different parameters.  C:\Users\LAB USER\Pictures\plot1.PNG | | | |
| FIG 4.1.1.3: Scatter plot showing accidents v/s causes | | | |
|  | | | |

**4.2 Statistical techniques and visualization**

Data is preprocesses before it is used for analysis and prediction. The steps in Data preprocessing are,

* Problem Framing: Requires the use of exploratory data analysis and data mining.
* Data Understanding: Requires the use of summary statistics and data visualization.
* Data Cleaning. Requires the use of outlier detection, imputation and more.
* Data Selection. Requires the use of data sampling and feature selection methods.
* Data Preparation. Requires the use of data transforms, scaling, encoding and much more. Model Evaluation. Requires experimental design and re sampling methods.
* Model Configuration. Requires the use of statistical hypothesis tests and estimation statistics. Model Selection. Requires the use of statistical hypothesis tests and estimation statistics. Model Presentation. Requires the use of estimation statistics such as confidence intervals. Model Predictions. Requires the use of estimation statistics such as prediction intervals.

Data visualization is viewed by many disciplines as a modern equivalent of  visual communication It involves the creation and study of the visual representation of data.

To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information Graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message.Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

The statistical Technique here we used is Scatter plot. A scatter plot (also called a scatter plot, scatter graph, scatter chart, scatter gram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian Coordinates to display values for typically two variables for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis

NumPy:

NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Since, arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem.

NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation. Numpy can be imported into the notebook using

|  |  |
| --- | --- |
| 1 | >>> import numpy as np |

Pandas:

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe. It is like a spreadsheet with column names and row labels.

Hence, with 2d tables, pandas is capable of providing many additional functionalities like creating pivot tables, computing columns based on other columns and plotting graphs. Pandas can be imported into Python using:

|  |  |
| --- | --- |
| 1 | >>> import pandas as pd |

Some commonly used data structures in pandas are:

**1. Series objects**: 1D array, similar to a column in a spreadsheet

**2.DataFrame objects:** 2D table, similar to a spreadsheet

**3.Panel objects:** Dictionary of DataFrames, similar to sheet in MS Excel

Pandas Series object is created using pd.Series function. Each row is provided with an index and by defaults is assigned numerical values starting from 0. Like NumPy, Pandas also provide the basic mathematical functionalities like addition, subtraction and conditional operations and broadcasting.

Matplotlib:

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits.

matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Majority of plotting commands in pyplot have MATLAB

**4.3 Data Modelling and visualization**

Data modelling is a process used to define and analyze data requirements needed to support the business process within the scope of corresponding information systems in organizations. Therefore, the process of data modelling involves professional data modellers working closely with business stakeholders, as well as potential users of the information system.

In this project we used Multiple Linear Regression. Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

*yi*​=*β*0​+*β*1​*xi*1​+*β*2​*xi*2​+...+*βp*​*xip*​+*ϵ*

where, for*i*=*n*observations:

*yi*​=dependent variable

*xi*​=explanatory variables

*β*0​=y-intercept (constant term)

*βp*​=slope coefficients for each explanatory variable

*ϵ*=the model’s error term (also known as the residuals)​

Imported libraries are numpy,pandas,matplotlib. NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits.

A library is essentially a collection of modules that can be called and used. A lot of the things in the programming world do not need to be written explicitly ever time they are required. There are functions for them, which can simply be invoked. This is a list for most popular Python libraries for Data Science.

A lot of datasets come in CSV formats. We will need to locate the directory of the CSV file at first (it’s more efficient to keep the dataset in the same directory as your program) and read it using a method called read\_csv which can be found in the library.

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.

Sometimes our data is in qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.

Now we need to split our dataset into two sets — a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import test\_train\_split from model\_selection library of scikit.

The final step of data pre-processing is to apply the very important feature scaling. Itis a method used to standardize the range of independent variables or features of data.

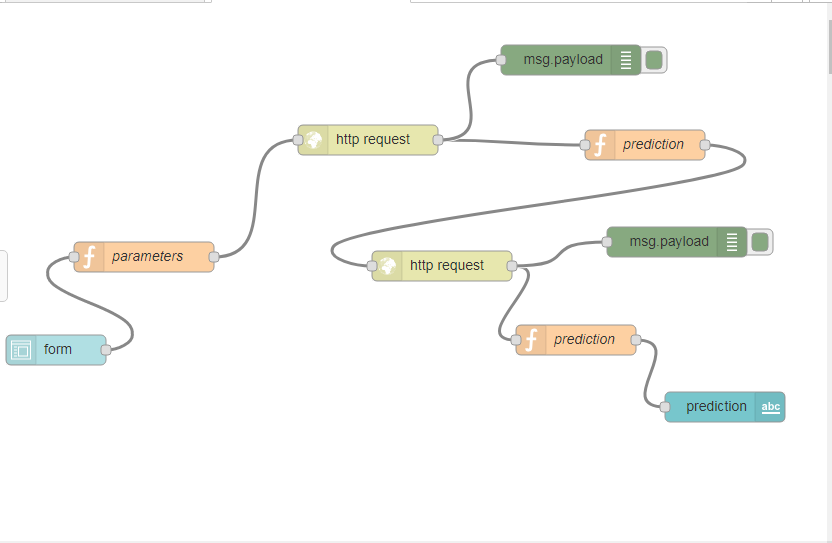


FIG 4.3.1: Representation of dataflow using Node-Red

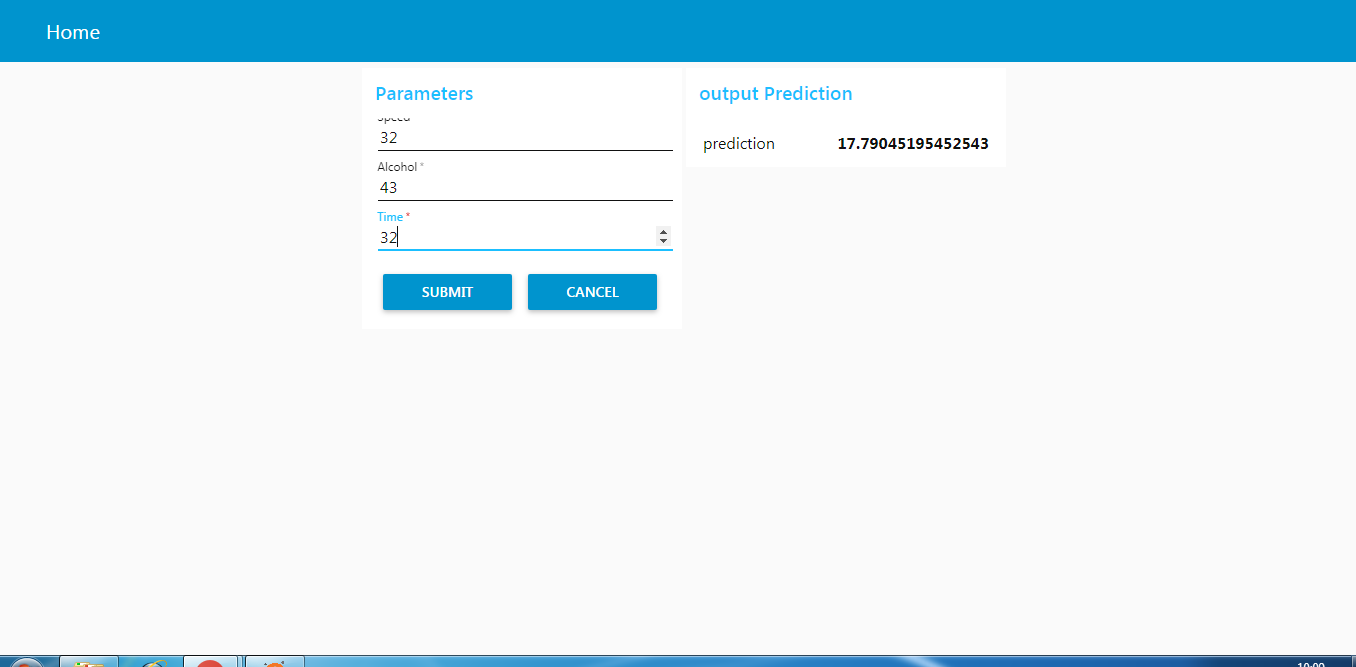


FIG 4.3.2: prediction results based on the data analysed

**5. References**

<https://www.datacamp.com/projects/462>

<https://cloudxlab.com/blog/numpy-pandas-introduction/>

<https://github.com/humzaali123/Reducing-Traffic-Mortality-in-the-USA>

**6. Conclusion**

This report attempts to contribute to the body of knowledge on road safety. It is hoped that it will inspire and facilitate increased cooperation, innovation and commitment to preventing road traffic crashes around the world. Road traffic crashes are predictable and therefore preventable. In order to combat the problem, though, there needs to be close coordination and collaboration, using a holistic and integrated approach, across many sectors and many disciplines. While there are many interventions that can save lives and limbs, political will and commitment are essential and without them little can be achieved. The time to act is now. Road users everywhere deserve better and safer road travel.