Credit Hours System
CMPN451
Data Mining, Big Data and Data Analytics

Cairo University
Faculty of Engineering

Big Data
Project Proposal

Submitted to:

Eng. Hussein Fadl
Eng. Ahmed Youssry

Submitted by:

Ahmed Mohammed Khalifa
Omar Abdelaleem
Omar Osama
Mina Ashraf Louis

1. Idea

Traffic Crashes:

Road accident among others occurs frequently that result in injury, death and property damage. The main risk factors for collision including vehicle design speed of operation, road design, road environment. An accident is an unintentional injury, incidental and unplanned event that are caused by the transfer of energy between the human body and the environment. It involves a minimum of one road vehicle, occurring on a road vehicle during which at least a minimum of one person is dislocated or killed, and Motor vehicle crashes are the number one cause of death for persons age 6 to 33 and account for more than 90% of all transportation-related fatalities.

2. Dataset

Traffic Crashes:

https://www.kaggle.com/isadoraamorim/trafficcrasheschicago

https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables

3. Insights we could extract:

1. How does the weather impact the number or severity of an accident?
2. Does driver age have an effect on the number of accidents?
3. What is the relation between hour, day, week, month with the number of fatal accidents?
4. Are certain car models safer than others?
5. Is the social class of a casualty dependent on the accident severity?
6. Relation between pedestrian age and casualties/accidents
7. Light conditions vs Weather conditions which is more critical when it comes to accidents
8. Rural areas vs Urban areas? Which one is more common and is there a relation to external factors such as Special condition.

Prediction

1. Predict if an accident is fatal
2. Forecast the future daily/weekly/monthly accidents
3. Forecast the reason behind an accident

Data Wrangling

Use Pandas to merge different files and use lookup tables
We could make a correlation matrix to find out which variables are most related to one another. We could also use Random forest feature importance, to find out which features matter the most when predicting an accident

Data Visualizations:

Use Seaborn (Heat map) to view things such as correlation matrix, Bar charts of Age groups and number of casualties, Male vs Female when it comes to casualties and who is most likely to be driver when an accident happens, also stacked bar charts can be used if we need to extract some data form a combination of attributes.

Use Scatterplot to see the relationship between two variables. For example the relation between the road defect and the crash type.

Use Histogram to know the frequency of occurrences of an attribute in the dataset.

Area Chart can be used to show continuity across an attribute in dataset. It is commonly used for time series plots. And it is used to plot continuous variables and analyze the underlying trends.

.