Proposal: Traffic Accident Analysis in The UK

DATA 450 Capstone

[Ayush Mainali]

2/5/23

1 Introduction

Road traffic accidents involving cars are a serious problem nowadays and one of the main sources of death and disability worldwide. Numerous people are injured or killed in car accidents every day, as well as experience emotional distress and financial losses. The issue has mostly been present in developing nations, where a lack of adequate traffic management, insufficient road infrastructure, and low levels of driver education all contribute to high accident rates.

Although there are many different and intricate factors that can lead to car accidents, some of the more frequent ones are speeding, being distracted while driving, or driving while impaired by drugs or alcohol. For instance, reckless driving involves actions like blowing through red lights, making erroneous turns, and tailgating other cars. Meanwhile, driving while intoxicated makes it more difficult for a driver to respond quickly and make wise decisions, which raises the possibility of an accident. Speeding shortens a driver's reaction time and lengthens the distance needed to stop the car, which raises the probability of an accident. Accident risk can also rise as a result of distracted driving, such as using a cell phone while driving.

2 Dataset

The dataset that I will be using will be "The Great Britain Road Accident 2005-2016" from Kaggle. Here is the link to the the dataset Include a full citation, as specified [https://www.kaggle.com/datasets/nichaoku/gbaccident0516?select=acc2005_2016.csv]

The dataset of the accident has two more sub datasets of Vehicle Description and Casualty Reference. The data was obtained from the UK government website [http://data.dft.gov.uk/road-accidents-safety-data/Road-Accident-Safety-Data-Guide.xls]. As well as giving details of date, time and location, the accident file gives a summary of all reported vehicles and pedestrians involved in road accidents and the total number of casualties, by severity.

The variable that will be used in the analysis are as follows:

i) latitude and longitude: The geographic location of the data where the accident occured.

ii) Accident Severity: The casualty and damage severity of accident

iii) Accident Day: The day of the week most numbers of accident occured.

iv) Accident Time: The time of the day accident occured.

v) Speed zone: The speed limit roads most accidents we occuring.

vi) Vehicle: Type of vehicle resulting in most number of accidents- cars/SUVs/Trucks, Bus etc

vii) Road Type: Type of road that saw most number of the accidents

viii) Light Condition: The light outside when the accident occured. Morning sun, midday light, dawn, dusk etc

ix) The gender of license holders: male, female, or others

x) Age: The driver's age

xi) Officers as scene: Number of police offices and local administratives that arrived in scene after the accident arrived.

xii) Road Surface condition: Conditions such as wet, damp, dry etc

xiii) Rural/Urban: Where the accident occured. Highly populated city or less populated rural areas

3 Data Acquisition and Processing

The selected dataset has 1.7 million rows and 32 columns. There is a total of 7% null data and to fix this we will be using mean and media for quantitative and mode for qualitative data. Rows of absolute no use will be deleted when analysis is performed. The encoded values for variables such as light condition, road type will be mapped to their actual values instead of category numbers so that the visualization becomes more effective.

We have 9 variables of no use which we will be dropping. A plethora of variables have numerical symbols representing various conditions such as lighting, daytime, road conditions etc and they will be reverse encoded where numerical value will be changed. Impprtant data will be retrieved as efficiently as possibly.

4 Research Questions and Methodology

1. Which day of the week has the most accident? We can create a barchart for all the days in the week where each bar will represent distinct day (like Sunday) in order to find out most accident occurance day.
2. Which time of the day has the most accidents? We can plot a line graph for every 3hrs time interval where each bar will represent distinct time(like 6pm-9pm for evening) in order to find out in what time of the day do most accident occour. The MM-DD-YY will be filtered to just hours to plot the graph.
3. What age group was mostly involved in accidents? Using the year of birth by filtering the DOB category from license details, we can figure out what age group is mostly involved in accident which in my understanding could be the age between 18-25.
4. At what speed zones did most of the accidents occur ? Using histogram we can map continuous data and interpret the zones with most number of accidents.
5. At what exact location in The UK are most accidents occuring? Using dot plot from plotly we can view some hotspots for accidents at different major cities. With most dots clustered around the major coordinates of the area, we will be also be analyse the data and report number of severe casualties in that area
6. What road types are causing most number of accidents? Barchart with frequency for accident severity can be mapped based on road type. We can use stacked bar to visualize severity(levels) and roadtypes.
7. What light conditions are causing most number of accidents? We will have multiple barchart with different colors that map the frequency for accident severity in accorance to the light condition.
8. What is the corelation between accident factors? We will predict the severity of the accident based on features with high correlation and be using a heatmap for observing a correlation. What variable accounts to what kind of corelation can be figured using heat map.
9. What gender have caused most number of accidents? Use pie chart and analyze the proportions. We will catergorize into male, female and unidentified.
10. Boxplot the age group for different age severity. There will be 10 boxplots in total with a window of ten years for each group.

5 Work plan

Week 4 (2/6 - 2/12): [Just an example:

- Data tidying and recoding (4 hours)
- Question 1 and 2 (4 hours).]

Week 5 (2/13 - 2/19):

- Data Visualizations for research questions and methodology listed above.(7 hr)

Week 6 (2/20 - 2/26):

- Machine Learning - Logistic Regression, KNN and Random Forest (7 hr)

Week 7 (2/27 - 3/5):

- Presentation prep and practice (4 hours)

Week 8 (3/6 - 3/12): *Presentations in class on Thurs 3/9.*

- Presentation peer review (1.5 hours)

Week 9 (3/20 - 3/26):

- Poster prep (4 hours)

Week 10 (3/27 - 4/2): *Poster Draft 1 due Monday 3/27. Peer feedback due Thursday 3/30.*

- Peer feedback (2.5 hours)
- Poster revisions (2 hours)

Week 11 (4/3 - 4/9): *Poster Draft 2 due Monday 4/3. Final Poster due Saturday 4/8.*

- Poster revisions (1 hour).

Week 12 (4/10 - 4/16):

- Overall revision.

Week 13 (4/17 - 4/23): [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (5 hours).

Week 14 (4/24 - 4/30): *Blog post draft 1 due Monday 4/24. Peer feedback due Thursday 4/27. Blog post draft 2 due Sunday 4/30.*

- Peer feedback (2.5 hours)
- Blog post revisions (2 hours)
- [Do not schedule any other tasks for this week.]

Week 15 (5/1 - 5/7): *Final blog post due Tuesday 5/2.*

- Final presentation prep and practice.
- [Do not schedule any other tasks for this week.]

Final Exam Week (5/8): *Final Presentations during final exam slot, Monday May 9th 3:20-6:40pm.* [Do not schedule any other tasks for this week.]

5.1 Some cool Quarto stuff

6 References

https://data.gov.uk/dataset/road-accidents-safety-data

https://www.kaggle.com/datasets/silicon99/dft-accident-data

https://data.dft.gov.uk/road-accidents-safety-dataRoad-Accident-Safety-Data-Guide.xls