Road Safety Data (UK) - Exploratory Data Analysis

1. Introduction

This report is aimed to analyze the road safety data in UK (https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data), these files provide detailed road safety data about the circumstances of personal injury road accidents, the types of vehicles involved, and the consequential casualties. The statistics relate only to personal injury accidents on public roads that are reported to the police and subsequently recorded, using the STATS19 accident reporting form. We use the files span from 2010 to 2019.

The purpose of the analysis including (1) summarize the main structure and characteristics of data with visual aids, and obtain interesting facts that are worth highlighting (2) Identity the relationship between variables and the casualties as well as the ratio of fatal casualties (3) provide some insight of how to predict at which condition the accident happen most, and which accident would be fatal.

The codes for this report can be found in …

2. File and data preprocessing

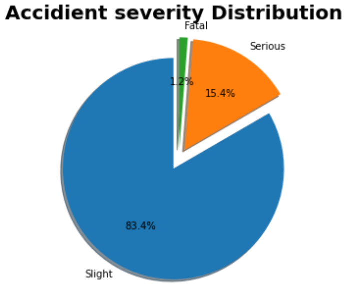
There are three different types of files in the dataset. The first type is the *Accident Circumstances*, which contain the records of the accident with its information such as the date, the environment of the accident. The second type is *Casualty*, which contains the records of the accident with the information about the victim such as the age, sex of transportation method. Finally, the *Vehicle* contains the records of the accident with the information about the driver and vehicle that cause the accident.

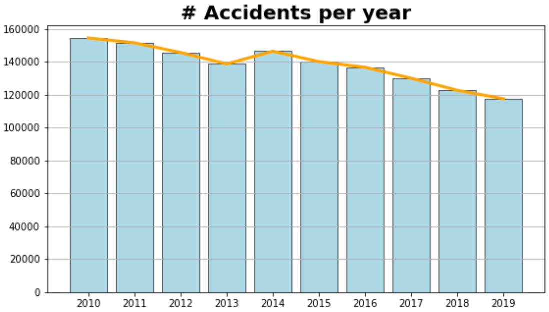
I merge the same file across different years, and rename some variable aliased by in different year (e.g., *"Act_Index"* and *"Accident_Index"*). Here I found some missing value, largely due to the record format is a little different in a different year (e.g., *"Casualty_IMD_Decile"* column appear in some years but not in others). Instead of removing the entire row which contains a missing value, I choose to leave them and remove them when the analysis is related to these columns.

3. EDA

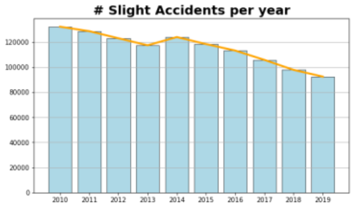
3.1 Accident Circumstances

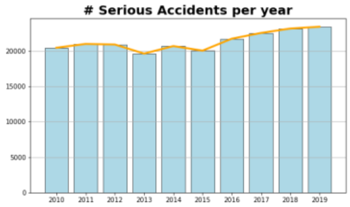
3.1.1 The trend of the number of accidents in 2010 to 2019

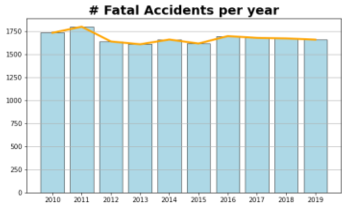




First, we look at the trend of the number of accidents over the last few years, and we can find that the number of accidents has continued to decrease for the last 6 years (146k in 2014 down to 117k in 2019). In addition, the accident was a mostly slight accident (83.4%).

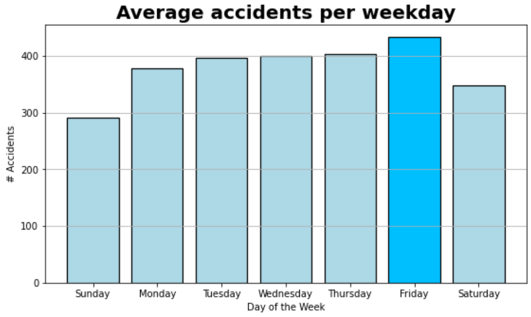


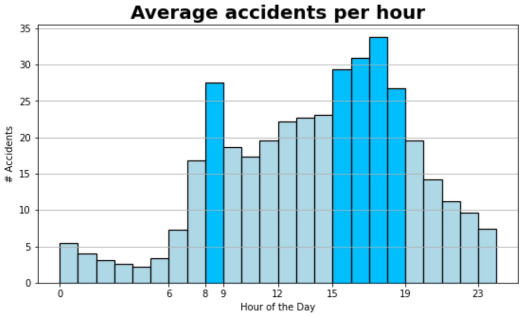




Although the total number of accidents has decreased a lot, the trend of different severity might be different. Therefore, we further view the trend for different accident severity, and we find that when the slight accident decreases a lot, the fatal accident does not decrease, and serious accident even increases in the few past years.

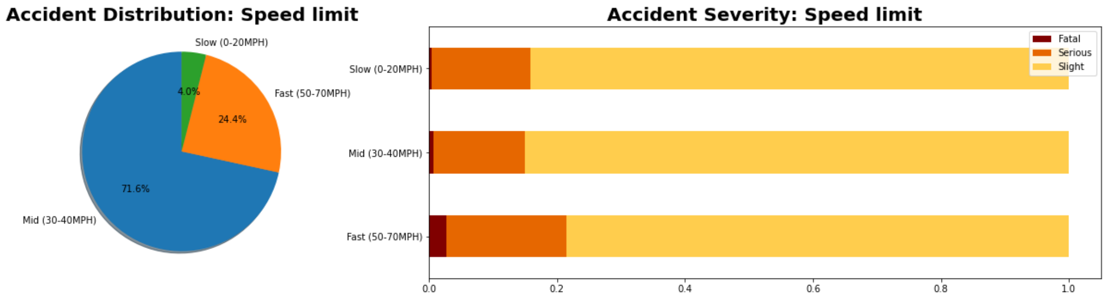
3.1.2 When do the accidents happen most?



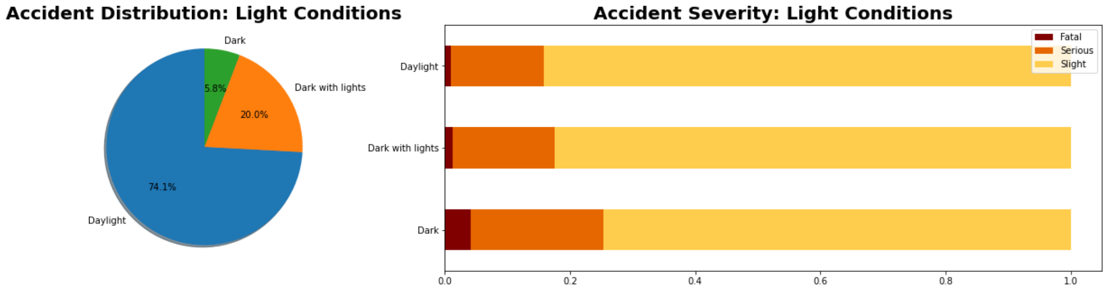


Next, we examine how the accident distributed through the weekday and the hour of the day. From the left figure, we can see that accidents happen most frequently on Friday, and least frequently on Saturday and Sunday. From the right figure, we can observe that the accident happened most frequently in 8-9 AM and 3-7 PM.

3.1.3 The relationship between accident and environment



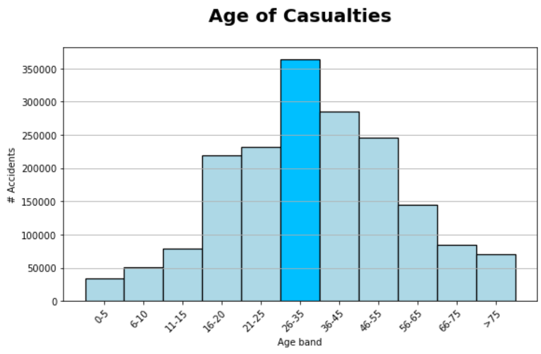
Here we analyze the relationship between the speed limit of the road and the accident. From the pie chart, we can see most accidents happened on the road with a speed limit of 30-40 MPH (71.6%). From the bar chart, we can further find that the accident that happened on the road with a higher speed limit would more possibly result in fatal casualty.

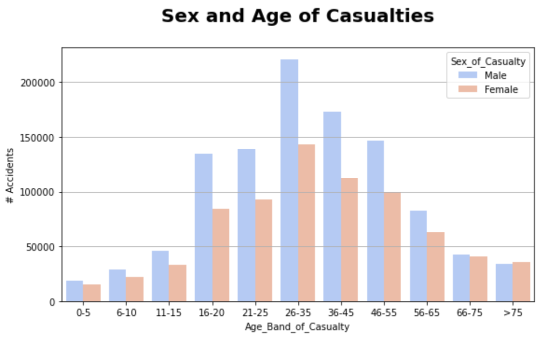


We also examine the relation between light condition and accident. From the pie chart, we can see that most accidents happen in the day light environment (74.1%). However, when the accident happened in the dark without a lighting environment, it is more possibly result in fatal casualty.

3.2 Casualties of accident

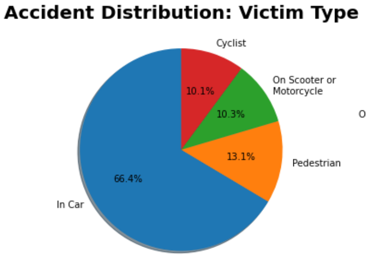
3.2.1 Age and sex of casualty

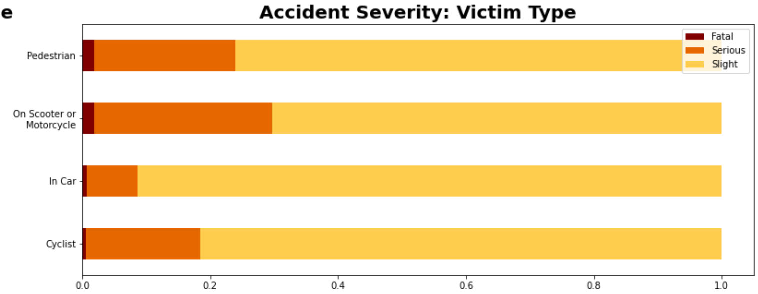




For casualties of accidents, we first look at how the casualties were distributed through different age band and sex. From the left figure, we can observe that accidents happen most with the young and middle people (especially 26-35), and least happened with the child or older people. From the right figure, we can further find that the male casualties are more than female casualties in every age band.

3.2.2 The relationship between accident and victim type

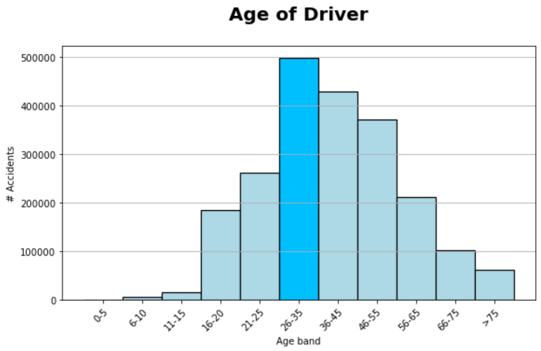


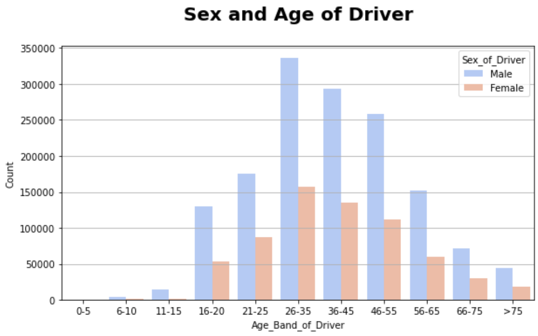


Next, we examine the type of transportation method for the victim, and we can find that most casualties are in the vehicle (66.4%), while other types nearly percentage (around 10%). However, when an accident happened, the casualties in the car were least likely to result in serious damage or death than the casualties not in the car.

3.3 Vehicles and drivers that cause accident

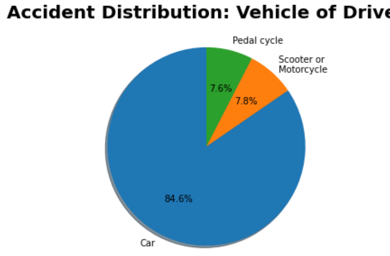
3.3.1 Age and sex of driver

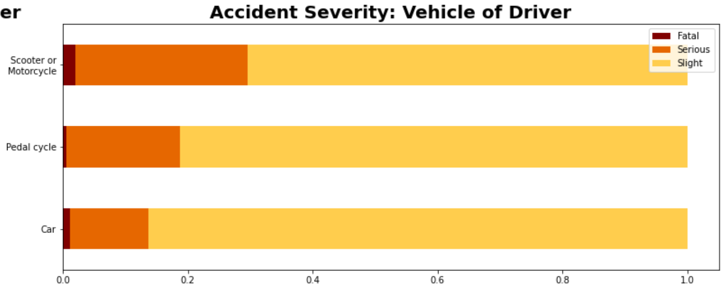




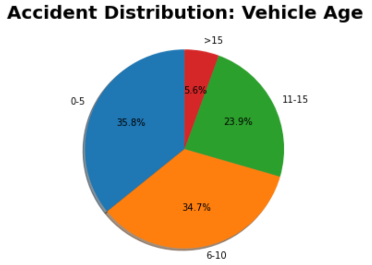
For drivers that cause an accident, we first look at how the casualties were distributed through different age band and sex. From the left figure, we can observe that accidents happen most with the young and middle people (especially 26-35), and least happened with the child or older people. From the right figure, we can further find that the male casualties are more than female casualties in every age band. The result here is quite similar to the result of *Age and sex of casualty*, except for that there are fewer drivers whose age is under 15.

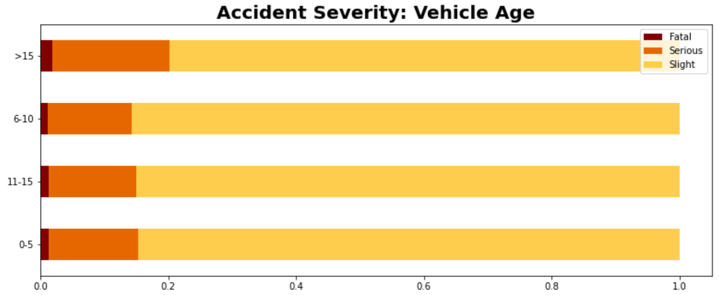
3.3.2 The relationship between accident and vehicle


Accident Distribution: Vehicle of Driver


Accident Severity: Vehicle of Driver

Next, we examine the factors related to accident. We first view the type of vehicle of driver. From the pie chart, we can know that in most accident, the vehicle of driver is car (about 85%). However, the accidents that caused by a motorcycle driver would more likely lead to serious and fatal casualties.


Accident Distribution: Vehicle Age


Accident Severity: Vehicle Age

We also examine the relationship between vehicle age and the accident. For the pie chart and the bar chart, we can observe that although accident with the vehicle age above 15 is not common (only 5.6%), but it is more likely to result in serious and fatal casualties.

4. Conclusion

This report analyzes the road safety data in the UK, and provide main characteristics with visual aids from three different aspects of an accident, including the circumstances of accidents, the victims of accidents, and the vehicles and drivers that cause the accidents.

In addition, many interesting facts were discovered during the analysis, such as when accident most happen with casualties in a car, the accident happened in a car are less likely to result in serious or fatal casualties than those casualties not in a car.

I believe this report is helpful to understand the characteristics and structure of the dataset, and it can be used to further research, such as analysis of the hotspot of accidents, or building a predictive model to predict whether a casualty is fatal.

5. Preliminary thoughts of how to build a predictive model

Since I don’t have enough time to build a predictive model, I could only propose some initial ideas of how I would build a predictive model. To build the predictive model, the first thing is to do data cleaning, such as remove invalid or missing data or replace those data with mean, median, or the majority of valid data. Next, I will split the dataset into training, validation, and test set with a ratio of 80%, 10%, and 10%. Meanwhile, I would keep the ratio of positive and negative data is equal across the different sets. In addition, we know that the fatal casualties only occupy a small part (around 1%) of total casualties from the EDA. Therefore, when building a predictive model, it would face the problem of an unbalanced dataset. To deal with the unbalanced problem, we can use the undersampling or oversampling technique for the training set. For the undersampling method, I would consider using the ENN (Edited Nearest Neighbor) approach. For the oversampling method, I would first try the classical approach called SMOTE (Synthetic Minority Oversampling Technique), and try some new approach based on GAN (Generative Adversarial Network). For the classifier, I would start from random forest and AdaBoost classifier, then turning and selecting the best model to have the highest f1-score on the validation set. Finally, I would test the performance of that model on the test set.