Text

Description automatically generated with medium confidence

**John O’Brien**

Email:

[John.obrien@henkel.com](mailto:John.obrien@henkel.com)

[Jobrien2.irl@gmail.com](mailto:Jobrien2.irl@gmail.com)

Project Report GitHub URL (insert URL here)

<https://github.com/jobrien2/UCD-project.git>

**Abstract (Short overview of the entire project and features)**

The project looks at the road casualty statistics for the UK over the period of 2016 to 2020 inclusive. An analysis of 2020 data shows that the data is skewed due to the impact of Covid-19 and the national lockdown. Through much of the analysis, 2020 data was omitted.

The project combines datasets on road casualty statistics and vehicle statistics

**Introduction (Explain why you chose this project use case)**

Why did I use this data

**Dataset (Provide a description of your dataset and source. Also justify why you chose this source)**

The project uses data from gov.uk website <https://www.gov.uk/government/statistical-data-sets/> including the 2016, 2017, 2018, 2019 & 2020 road casualty statistics datasets.

Road Safety Data with vehicle statistics was sourced from [Road Safety Data - data.gov.uk](https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data) and was used to link road casualty statistics.

The total number of registered cars by make and model was obtained again from gov.uk. <https://www.gov.uk/government/statistical-data-sets/vehicle-licensing-statistics-data-files>

Only information on registered cars for end of quarter 2021 could be accessed from the website, and although the dataset is more recent than the casualty data, the dataset is used nonetheless to give an indication of the number of cars on the road and the comparative accidents by car make.

**Implementation Process (Describe your entire process in detail)**

Before bringing in datasets, the appropriate packages were imported to the shell. As the project developed, additional packages were imported and scripted in the same cell.

Pip install was used to install packages. Once an additional package was installed, the pip install statement was hashed out (#).

**Step one – Import data:**

The CSV files were imported with pandas function ‘pd.read\_csv( )’ to convert the CSV directly into a dataframe

BeautifulSoup package was used to scrape the uk.gov website for an additional CSV file

The first objective was to analyse car accidents as a proportion of the number of cars on the road by make. As I will be merging two datasets, to simplify the analysis, all car makes will be joined based on the first name. Regular expression (regex) was used to extract the first car name from the Make column. the data frame was grouped by the regex output make.

2019 data was used instead of the 2020 data as there was an obvious impact of accident data due to the Covid-19 lockdown which would obscure the analysis

Chart, bar chart, histogram

Description automatically generated

The 2019 data set was used to compare the accidents by number of cars on the road as a proportion.

The make of the car was gleaned using the same regex code. The two datasets were then merged on the regex identified make name.

The merged dataset was then analysed to look at patterns

**Join Data:**

I mostly joined the data sets with pd.join ensuring I matched on keys for both dataset. Left join was suitable for all events.

Pd.concat was used to concatenate onto the end of a datframe with a second dataframe.

**Create functions:**

I made two functions

1. A function to plot bar charts

Text

Description automatically generated

1. A function to slice the dataset guide in order to review the field and code data

Text

Description automatically generated

**Results (Include the charts and describe them)**

Chart

Description automatically generated

Pareto chart ranking the Car make by the accident rate proportionate to the number of cars of that make on the road.

Chart, scatter chart

Description automatically generated

The scatterplot above shows the incident rate of each car make relative to the number of cars on the road. The size of the dot on the chart represents the number of registered cars of that make plotted against the average age of the driver in the accident.

The graph, although busy, shows that certain makes are more likely to be in an accident. Wrightbus has an incident level of >6% however.

Chart, histogram

Description automatically generated

The graph plots the probability of a driver of a an age range to be in a fatal accident, not surprisingly the age range between 20 and 60 is higher with a likely decrease in number of drivers >70 years of age on the road, particularly commuting to work.

Graphical user interface, application

Description automatically generated with medium confidence

The median age of a male driver involved in a fatal accident (coded as number 1 ) is lower than that of a female driver.

The data was further sorted and the outcome of an accident was recoded. Fatal or Serious was coded as 0 and Slight accident was coded as 1.

Further analysis of the data set showed that women were less likely to be in an accident and less likely to be in a severe accident or fatal accident.

Chart, bar chart

Description automatically generated

Fatal accidents peak at 16:00

Chart, bar chart, histogram

Description automatically generated

Although accidents peak at 16:00, there is a spike in accidents at 08:00 which doesn’t correlate to fatal accidents in the same way accidents at 16:00 does.

Chart, histogram

Description automatically generated

Most accidents occur on a single carriageway followed by a Dual carriageway

Chart, bar chart

Description automatically generated

Accidents are most common when car is ‘going ahead other’ code 18. Head on collisions.

However code 9 accidents, turning right, is more frequent than code 7, turning left.

Chart

Description automatically generated

Accidents mostly occur in Fine weather

Chart, histogram

Description automatically generated

**Modelling the data**

A logregression model was attempted to predict data

As there was multiple outputs for many of the variables, i.e. more that 2 outputs, dummy variables were created to have only two values, 0 and 1.

Recursive Feature Elimination was used to identify the best performing features in the dataset.

The RFE identified columns were then selected and a logreg model was applied.

The accuracy of the logistic regression classifer was 0.78

The ROC curve was then plotted.

Chart, line chart

Description automatically generated

**Insights**

1. Road accidents reduced significantly in 2020 with a deep decline in March. This can be explained by the national lockdown in the UK during Covid
2. Accidents peak at 08:00 and from 15:00 to 18:00. These times coincide with school and work start and end times. Of interest is a fatal accident is less likely to occur in the morning. Does the strain of a full day lead to less attentive driving that has a higher consequence.
3. Dual carriageways are safer than single carriageways
4. A right hand turn is far more likely to lead to an accident then a left hand turn. This makes sense as typically a right hand turn requires crossing a lane where a left hand turn only requires a driver to join a lane.
5. Accidents are less likely to occur in poor weather conditions, drivers are maybe more attentive whilst driving in poor weather conditions.