Road Traffic Accidents- UK, 2019

771762 – Big Data and Data Mining PROJECT

Ruth Oni April 2022 202122441

Contents

Introduction

Since 1979, the UK Department of Transport publishes alongside a report on annual trends, data of traffic accidents reported to the police (Staines, 2018). This project investigates specific trends on smaller scales, including weekly and across days, the relationship between different data features and the rate of crash incidences. Finally, predictive models are developed to forecast the potential time, location and severity of accidents.

The data comprises three datasets.

1. **Accidents:** 32 variables, detailing location, time, date, lighting, weather, road conditions and other variables. The unique accident index identifies each observation and make up one of 117,536 collisions.
2. **Vehicles:** Contains details of vehicles involved in the accidents.
3. **Casualties:** 16 columns with information about casualties involved in accidents.

Data Cleaning

- **Filling missing Values:** Missing values (not unknown values) were handled with forward and backward fill (Tamboli, 2021).
- The date and time columns were converted into DateTime objects for the analysis.

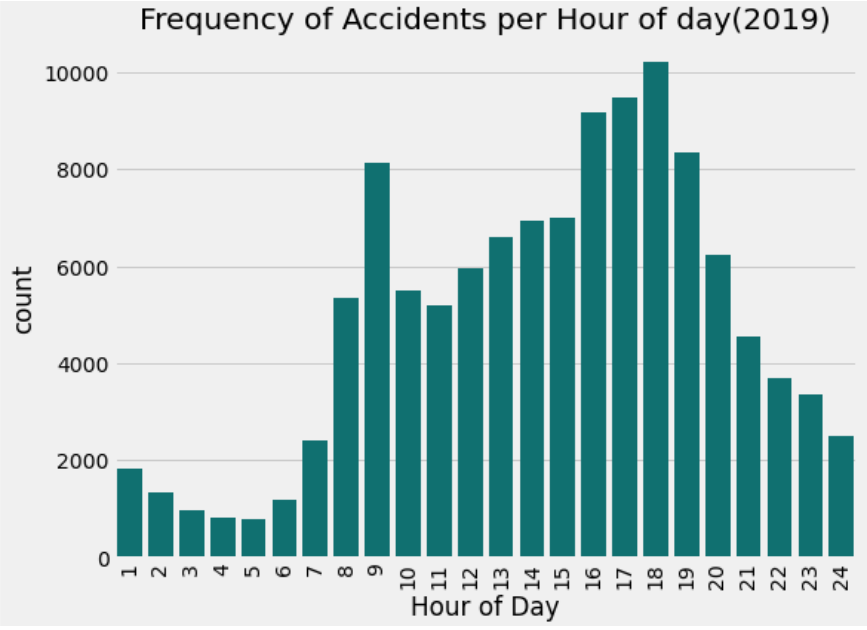
Analysis

Significant periods of all road traffic accidents.

Across all categories (vehicular or pedestrian), There are indeed significant hours of day and days of week when accidents are most likely to occur according to the data.

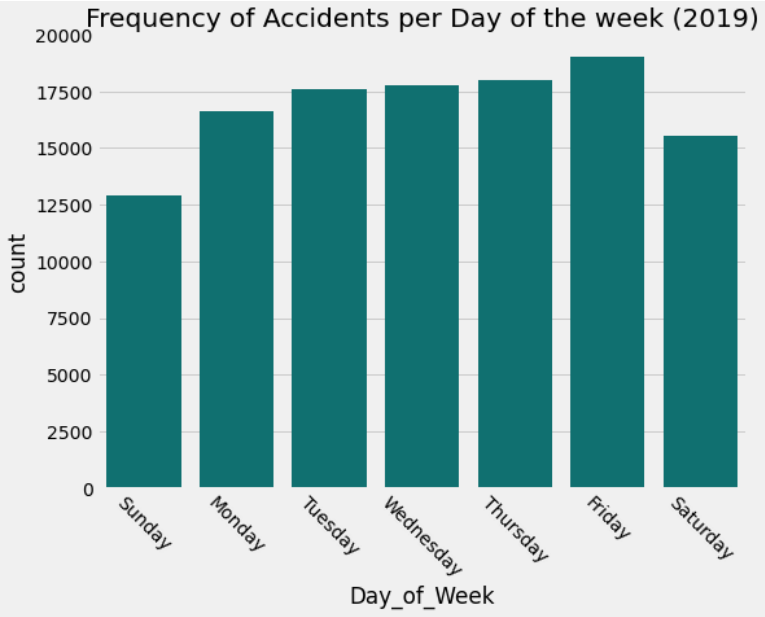
Hours of day

The **18th** hour of the day (between **17:00 and 18:00**) recorded the highest number of accidents all year with accidents within this period increasing by **108%** (**28/hr)** when compared to the total hourly average(**13/hr**). In addition, the frequency distribution plot of accidents across hours of the day reveals highly significant occurrences during daytime, with major spikes between **07:00 and 08:00hrs**, **15:00 and 18:00hrs**.



A report by (UKDrivingSkills, 2021) places the peculiar morning and afternoon periods within the "rush hour" of a weekday where traffic is busiest as a result of commute to and, from socio-economic activities (BBC, 2018).

Days of Week

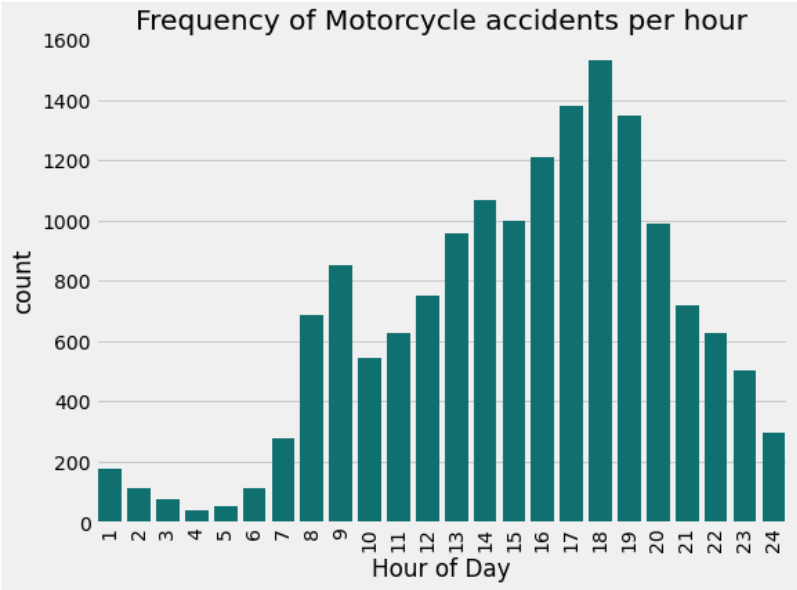


More accidents happen on a working day than on weekends with **Friday** recording the highest number of occurrences. When compared with the daily average (**322**), Fridays witnessed just about **13.8% (366)** more incidences on average.

We can ascribe this to people commuting from socio-economic activities to Friday festivities and weekend travel at the end of the business week. Weekends, Saturdays, and Sundays have fewer incidents, which could be ascribed to less traffic as a result of fewer activities that require the use of main highways.

Significant periods of motorcycle accidents.

Hours of day

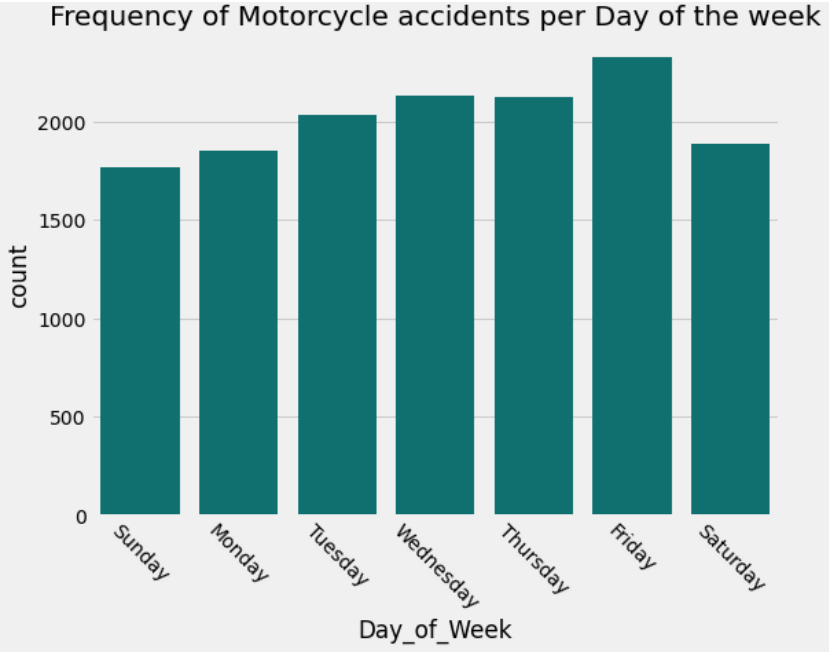


According to the distribution plot, the majority of motorcycle accidents occurred in daytime, with the highest incidences occurring at consistent times across all road incidents. When compared to the hourly average **(2/hr)**, the 18th hour of the day **(17:00-18:00)** recorded the largest number of crashes all year, with crashes increasing by **130% (3.7/hr)**. **12%** of accidents that happened in 2019 involved motorcycles. Accidents within the **18th hr** alone made up **(9.6%)** of daily occurrences.

We may accrue the observations to our 'rush hour' theory and a concentration of motorcycle use for courier services within cities.

Days of Week

The ascending trend of occurrences across the week reveals that more accidents happen as the business week progresses. Friday recording the highest numbers may be attributed to busier traffic, as a product of the rush hour on the weekend start. Accidents on Friday constituted **16.1**% of weekly occurrences with the rate (**39**) going up by **15**.**8**% compared to the daily average of **44**/day**.**

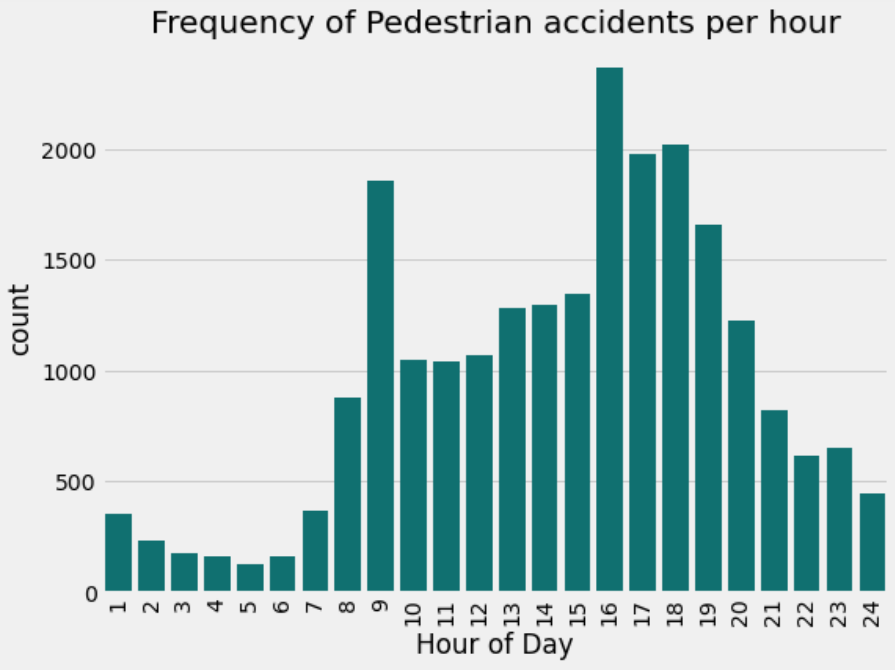


However, the general distribution is nearly uniform across all weeks, with between **1800** and **2500** motorcycle accidents occurring on each day. The frequency distribution of motorcycle accidents does not give sufficient information to support the hypothesis that certain days have a considerably higher probability of accident occurrence.

Based on our courier services theory, we believe that motorcycle-based courier services, such as food delivery, operate every day (business days and weekends), as such services are relatively constantly in demand. Apart from courier services, young men aged 17 to 29 make up the majority of motorbike owners and drivers in the UK. (Department of Transport, 2016).

Significant periods of accidents involving pedestrians.

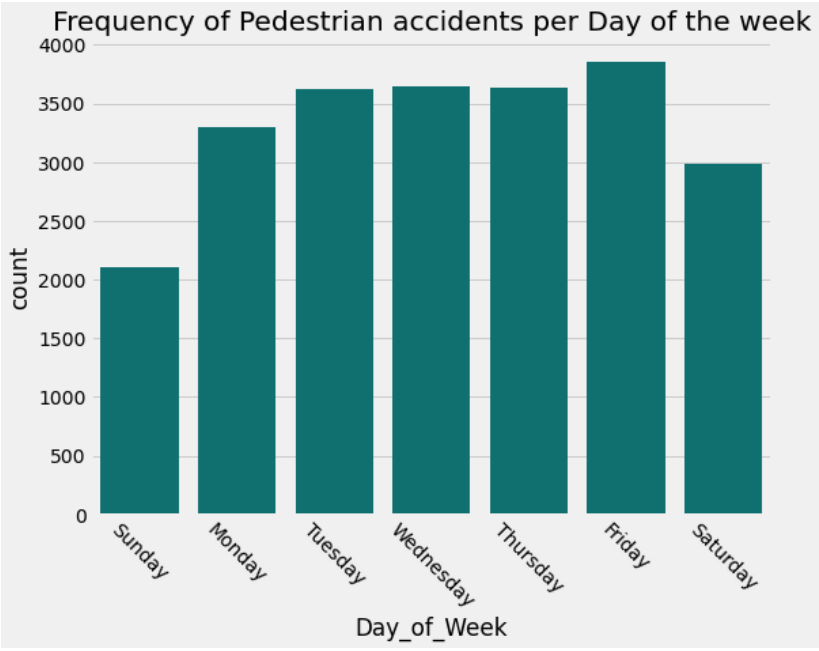
Hours of day



Pedestrian accidents were more common throughout the day (between **7:00** and **19:00 hours**) than at other times.

The distribution shows a surge in the **9th** hour (**8:00-9:00hrs**) of the day compared to other periods of the morning with other significant hours between **15:00** and **18:00hrs**. However, unlike observed in other scenarios, pedestrians were most involved in accidents during the 15th hour, when the rate was **145%** higher (**6.5/hr)** compared to the hourly average of **3**. These occurrences are **10.2**% of the daily numbers involving pedestrians.

Since major significant periods of pedestrian traffic are within the same range as automobile traffic, we presume that the same theories apply to both categories. Socio-economic activities have a significant impact on traffic concentration at various times of the day and heavy traffic directly implies a higher risk of accidents, especially those involving pedestrians.



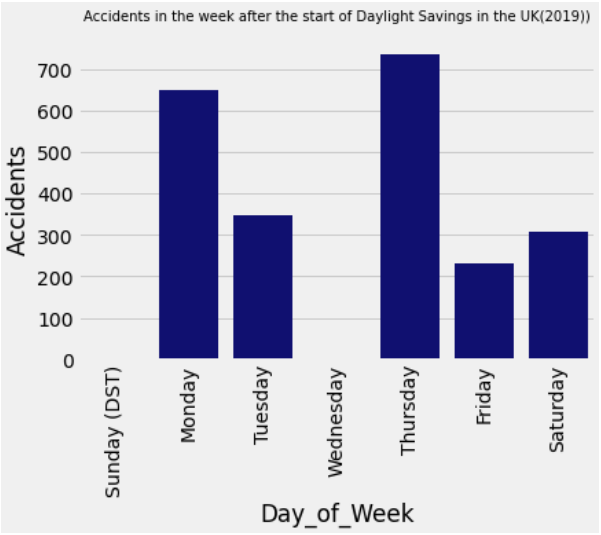
Days of Week

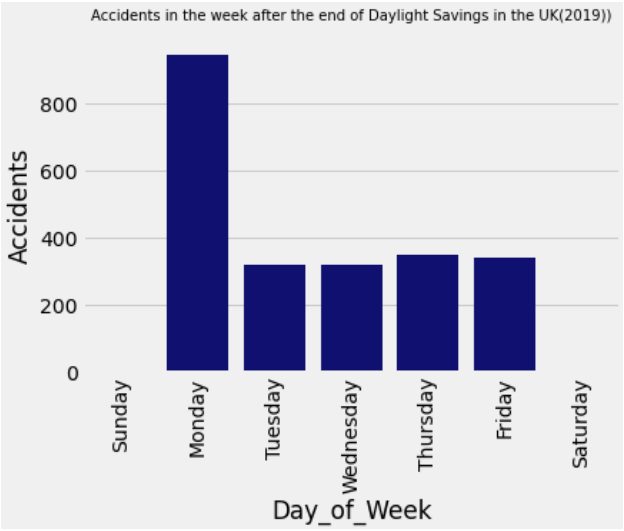
The plot of the distribution shows that accidents involving pedestrians are fairly evenly distributed across the working days of the week (Monday - Friday). However, Fridays witnessed the most of these accidents, accounting for **16.7%** of all pedestrian accidents per week. On Fridays, the rate increased by **17%** (**74/hr**) compared to the daily average of **63**.

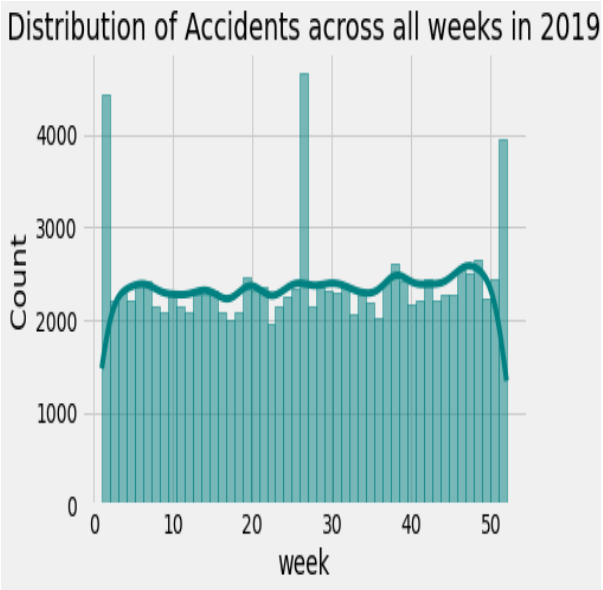
Friday, as the transition of the week into weekends, and the associated commute, is responsible for the greater frequency of accidents on that day when compared to the rest of the week, similar to the distribution of all accidents. Because there are less activities necessitating the use of main roadways on weekends, there are fewer accidents.

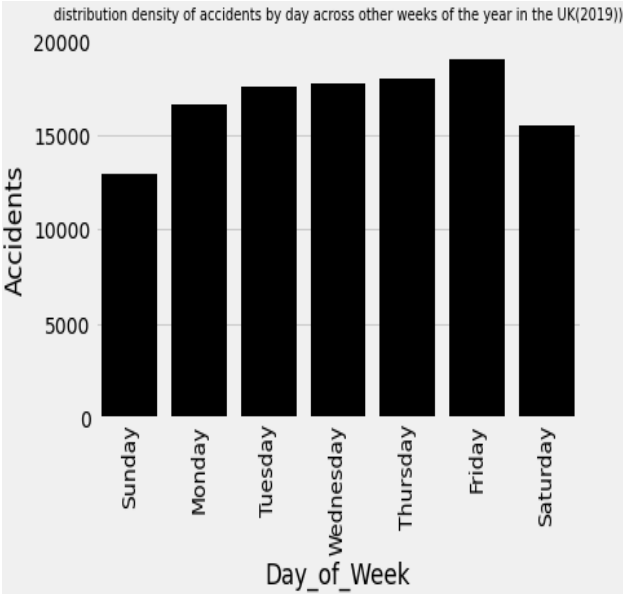
Impact of daylight savings (DS) on the frequency of accidents.

Distributions of accidents across all concerned periods were tested for normality and compared using appropriate statistical tests to confirm if DS impacted accident frequency. The analysis confirms no significant variation in the distribution of accidents across days in both periods.









The observed difference is mostly insignificant, with a weekly average reduction of **13** accidents in the week after DS began and an increase of **15** in the week after it ended. See table 1

Table 1 Daylight savings and statistics of accidents frequency UK(2019)

Period	Weekly avg (no of accidents)	Daily avg (no of accidents)	Peak day of week (% of period total)	
Week after start of DS	**2273**	**325**	Thursday	**32.4**
Other weeks of the year	**2260**	**322**	Friday	**16.2**
Week after end of DS	**2273**	**325**	Tuesday	**41.2**

Impacts of sunrise and sunset on the occurrence of accidents.

Sunrise and sunset have an impact on the occurrence of accidents. Sunrise and sunset times vary across the year in the UK (WorldData.info, 2022). The darkest months (with shorter daylight hours) October to February with less than 11 hours of daylight on average. The longest days are between March and September (WorldData.info, 2022).

Table 2 Sunrise, Sunset and accidents UK (2019)

Period		Time **(WorldData.info, 2022)**	% rate of period (bright/dark)	Rate / hour
Darker Months	Sunrise	**4:40 – 6:33**	**2.1**	**3.7**
	Sunset	**15:53 – 18:09**	**20.8**	**29.9**
	Daylight Hours	**6:34 – 15:52**	**40.2**	**17.1**
Brighter Months	Sunrise	**7:12 – 7:22**	**0.3**	**9.2**
	Sunset	**18:06 – 21:21**	**16.1**	**27.3**
	Daylight Hours	**7:23 – 18:05**	**2.1**	**34.7**

During the **darker** months, accidents happened less around sunrise when compared with sunset and normal daylight hours. See table 2. This may be due to reduced traffic on the roadways in the early morning hours. After all, road accidents only occur on the road when people are driving, not in homes.

We expect sunset to have higher impact on the occurrence of accidents as it is almost completely dark at the end of business days in these months. Darkness implies reduced visibility and increased risk of accidents.

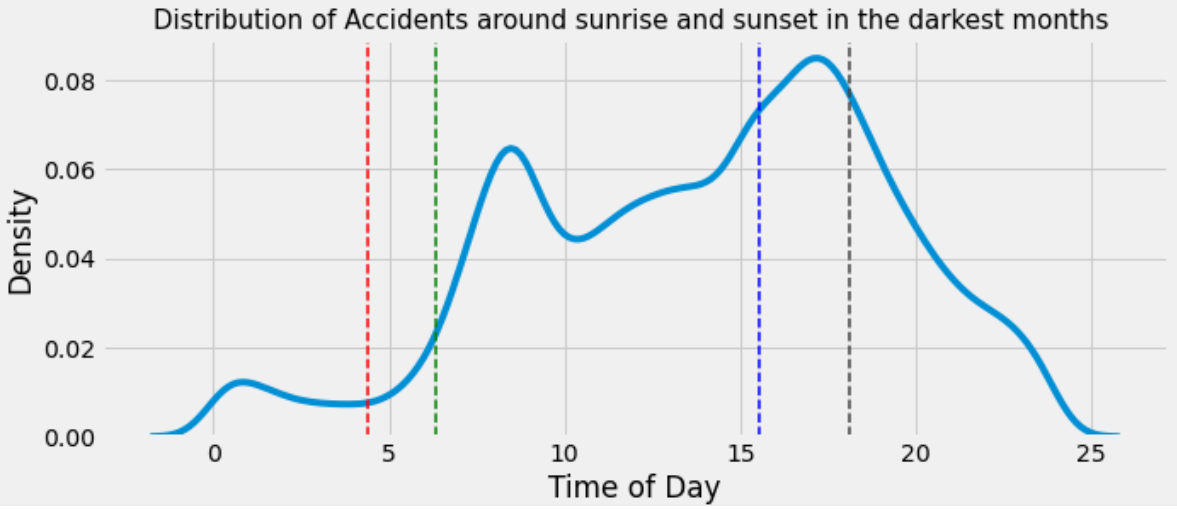


Figure 1 Accidents during the darkest months. sunrise between the red and green dotted lines. Sunset between the blue and black dotted lines.

In contrast, accidents occur less frequently around sunset during the brightest months. Because the sun sets well after rush hour during these months (see '**significant periods of road traffic accidents**'), traffic is less congested around dusk, lessening the likelihood of accidents.

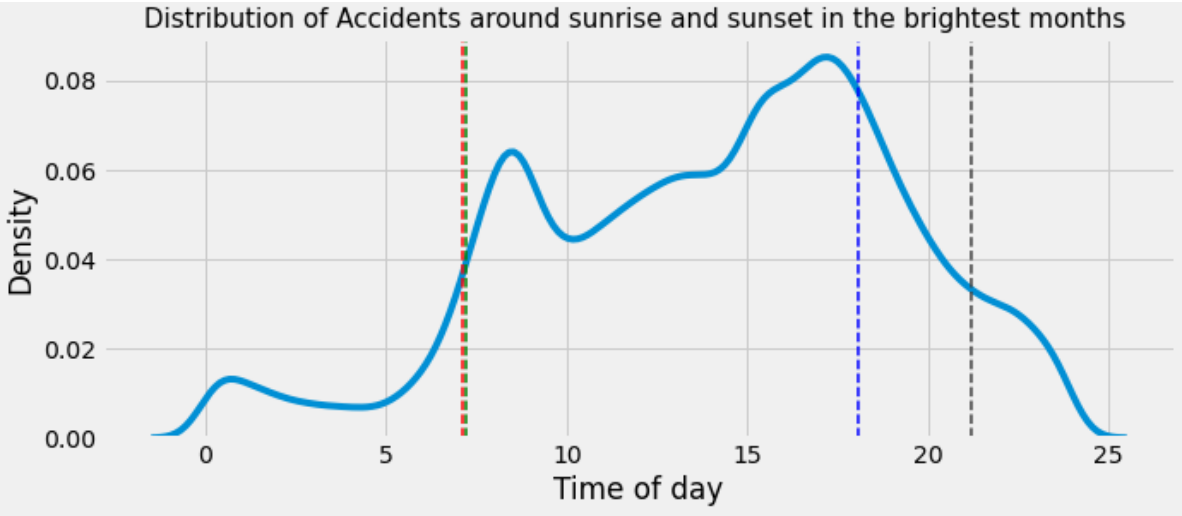
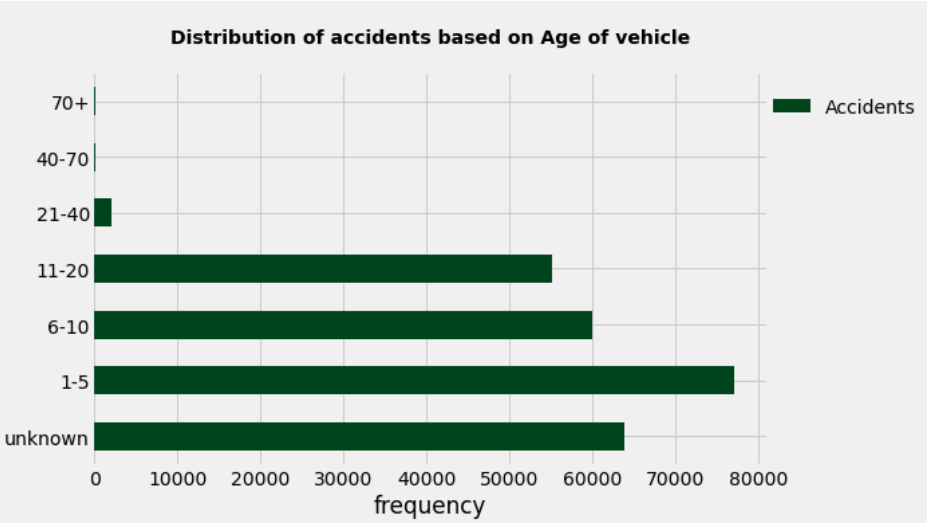


Figure 2 Accidents during the brightest months. sunrise between the red and green dotted lines. Sunset between the blue and black dotted lines

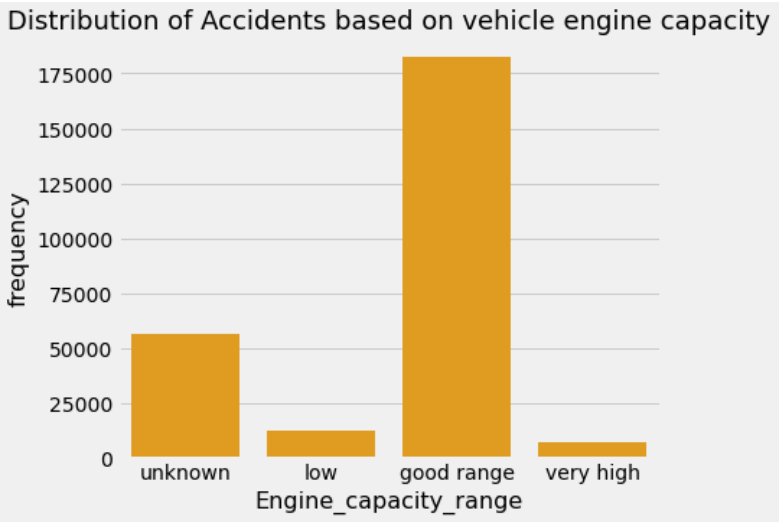
Vehicle features and frequency of accidents

Age of vehicle

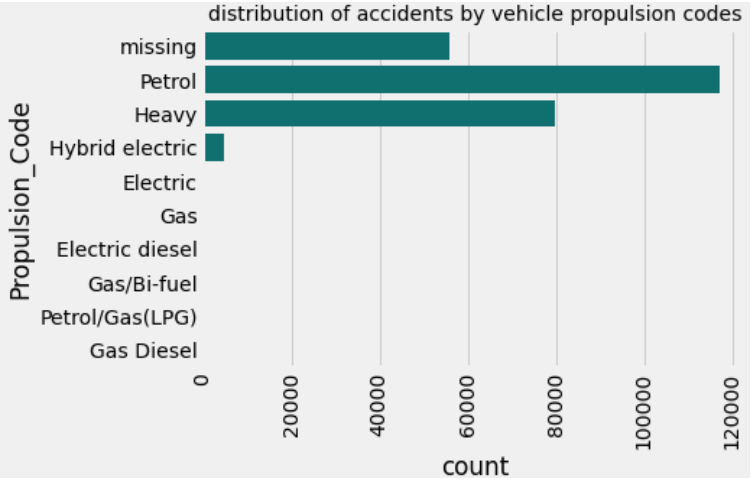


Vehicles 5 years old or younger were more frequently (**29%**) involved in accidents than vehicles in other age ranges. Vehicles between **6** and **20** years old also largely contributed (**24.6%** combined) to crashes in the year 2019. However, the data shows that vehicles with no information regarding their age constitute the second largest group (**24**%). Making a conclusion therefore, about the question will be reductive.

Engine Capacity(cc)



The larger a vehicle's cubic capacity, the more power and, utltimately, speed it can achieve. Vehicles with good cc were more frequently involved in accidents, according to the data. This is likely due to the fact that most motorbikes and other vehicles in the UK have engines larger than 50cc and 1000cc, respectively. (Statista, 2020). This result is however not conclusive as cc for **21.8%** of vehicles is unaccounted for.



Propulsion Code

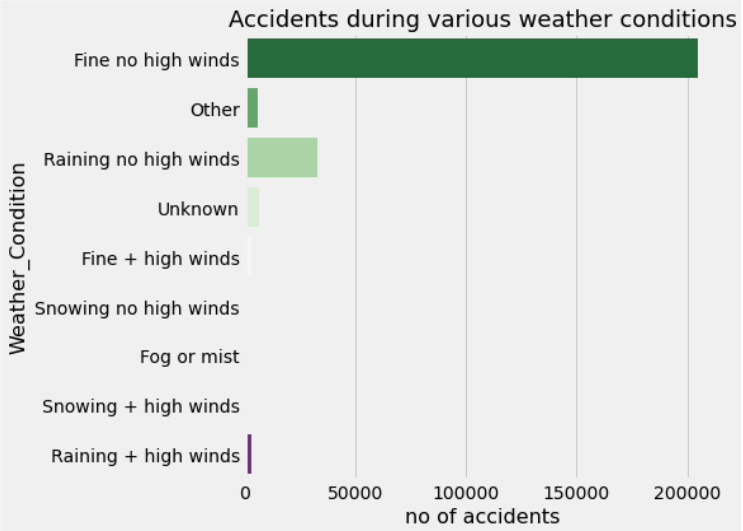
According to the findings, vehicles with the propulsion code **Petrol** were most frequently involved in accidents, accounting for **45.41%** of all collisions. In the UK, petrol is the most widely used percentage type. (Statista, 2022).

NB. The available data indicates a problem with identifying and collecting vehicle feature information for vehicles involved in road traffic accidents.

Environmental conditions and frequency of accidents

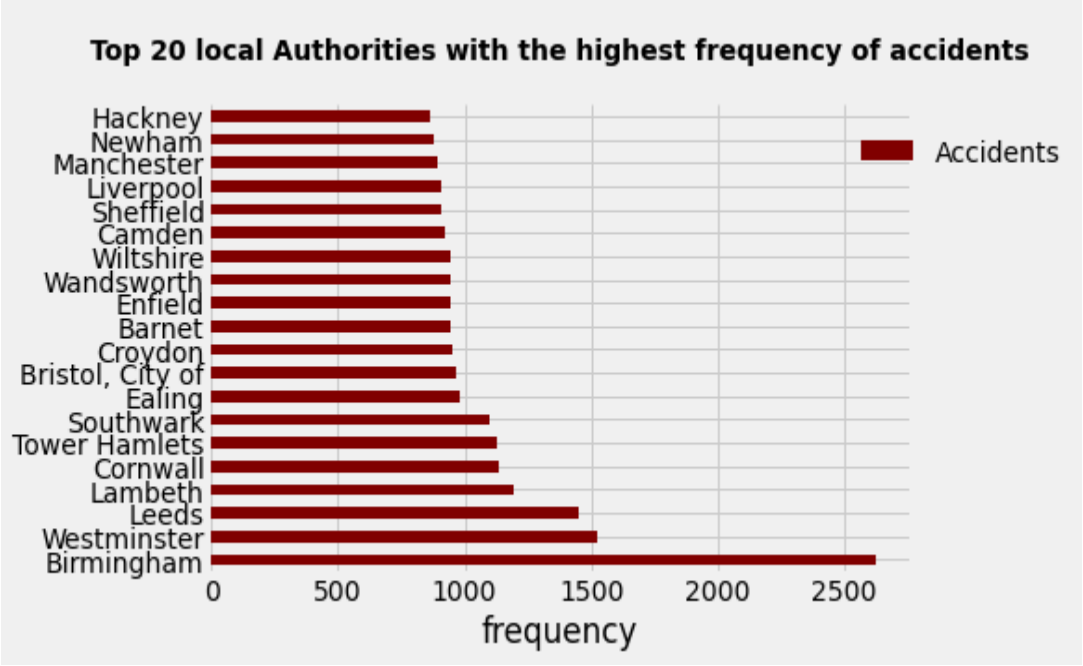
Weather conditions

Majority of accidents occurred when the weather was fine with no high winds, accounting for **78.5**% of all accidents. Poor weather conditions make up the remaining **20.5**% implying that they have little impact on the frequency of traffic accidents. According to a report by (Cooper, n.d.) the huge disparity in rates is because most UK drivers avoid driving in bad weather directly reducing the risk of accidents.

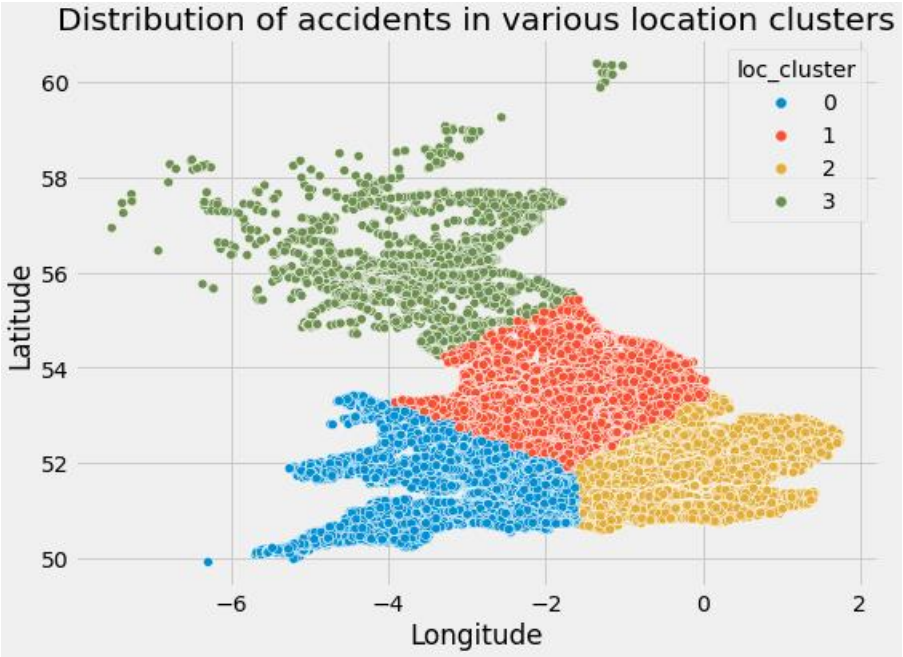


Geographic Location

Birmingham witnessed the highest number of accidents in 2019

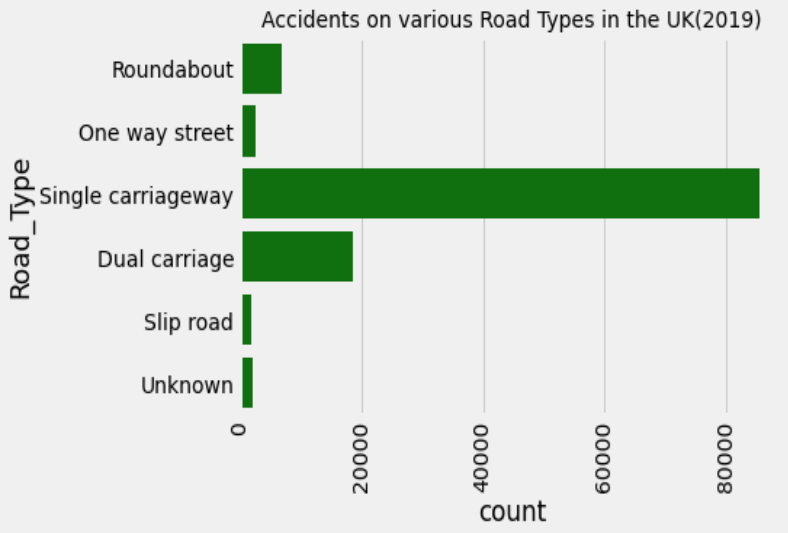


Regions within cluster 2 (including the South-East, London and East England) witnessed more accidents (48.1%) than other regions.



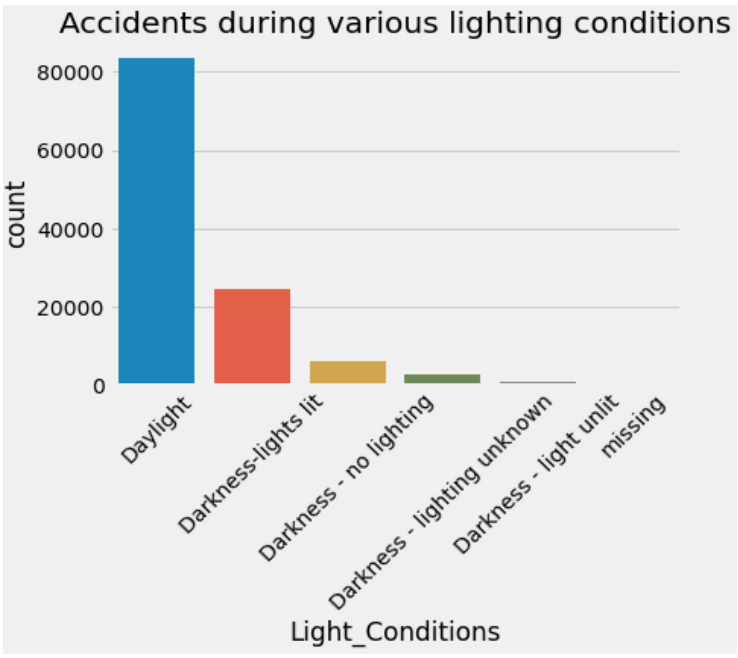
Road type

85320 accidents occurred on Single carriageways accounting for **72.6**% of all accidents. The lack of lane dividers on single carriageways could explain the greater risk on such roads.



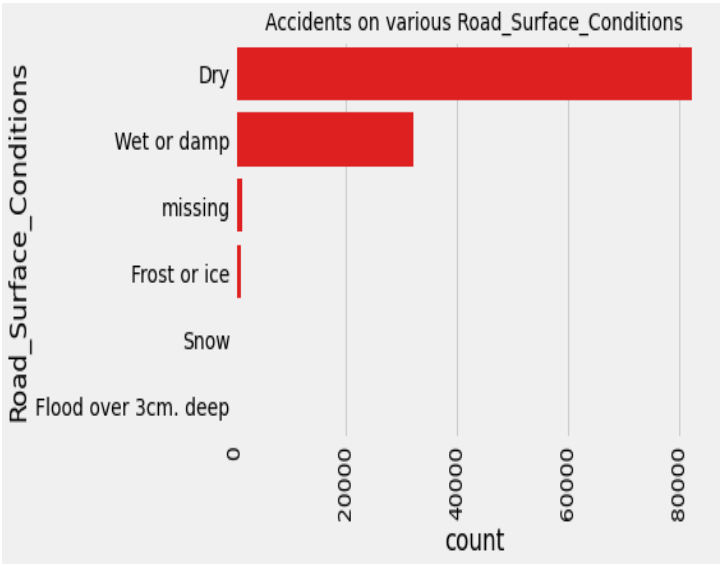
Light condition

71.1% of all accidents happened in broad daylight. Traffic is busier during daytime than at night with more people driving in daylight, directly increasing the risk of accidents.



Road surface

70% of all accidents happen on dry roads. Fewer people drive in severe weather, especially rain as previously discussed explaining this observation.

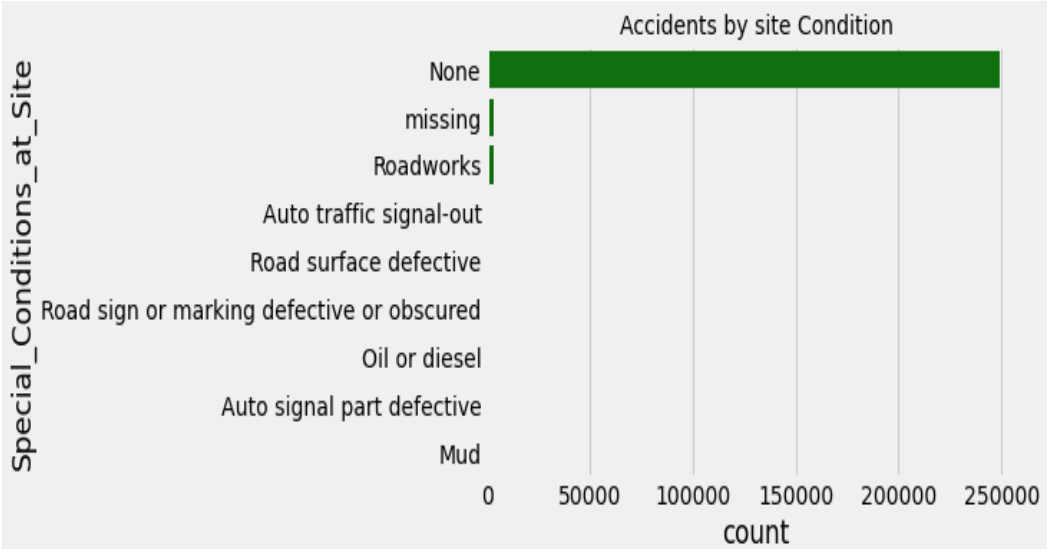


by traffic than the present road condition.

Special conditions

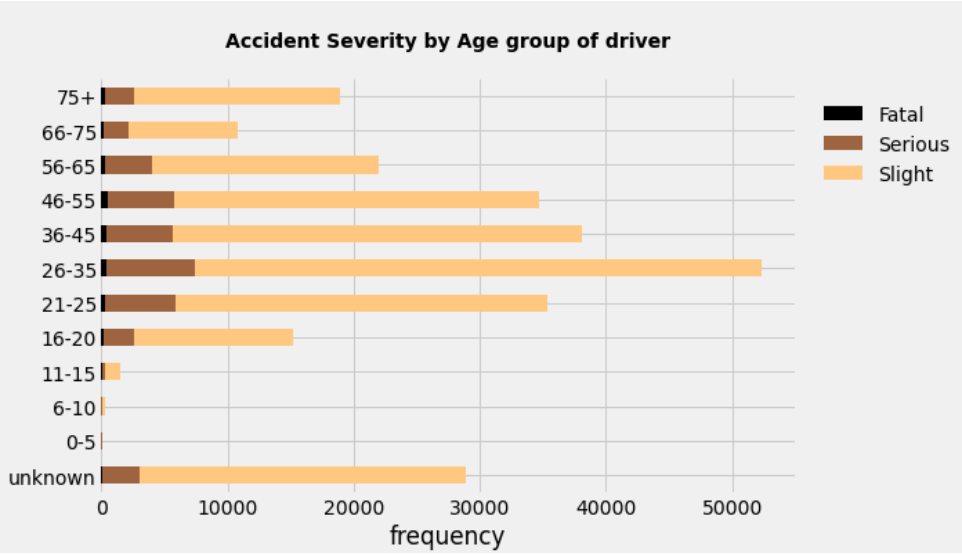
Fewer accidents (**3.7**%) happen at sites with special Conditions than they do regularly. Special conditions like defects on roads are not a norm and so do not contribute significantly to the probability of accidents when compared to normal roads.

According to the data, accidents are caused more



Driver related variables and Accident severity

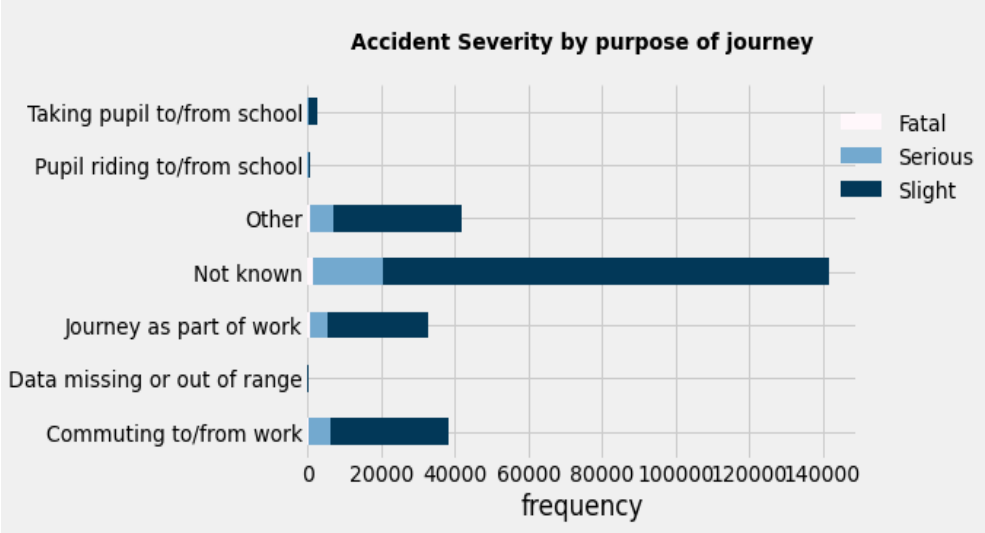
Age of the driver



With a support of **0.1** and confidence of **0.9**, the analysis demonstrates an association between the age group of drivers and accident severity. This is however irrelevant because the relationship is for drivers of undetermined ages. However, the plot reveals that drivers between the ages of 26 and 35 were the most common victims.

Purpose of the Journey

The purpose of journey had little bearing on the severity of accidents. Analysis reveals strong associations between work-related journeys and slight casualty severity. However, the strongest association with support of **0.47** and confidence of **0.85** is between journeys with no known purpose and slight casualty severity.



Predictions

Predicting Time, Location and Severity of accidents.

Table 3.0 below presents results of modelling the exact time of accidents. With the stacking model predicting with a final r2 score of 98%. Table 3.1 presents results for modelling accidents severity.

	Model	RMSE	r2_score
0	LinearRegression	3.265061	0.597664
1	KNeighborsRegressor	1.106786	0.953769
2	RandomForestRegressor	0.716096	0.980647
3	GradientBoostingRegressor	2.031173	0.844296
4	stacked_model	0.711515	0.980894

Table 3.0 Regressors for time

	Model	Accuracy	Precision
0	XGBClassifier	0.848157	0.782998
1	KNeighborsClassifier	0.828001	0.762680
2	RandomForestClassifier	0.845570	0.787343
3	GradientBoostingClassifier	0.848235	0.802602
4	stacked model	0.849396	0.804295

Table 3.1 Classifiers for severity

Table 4.0 Classifiers for Location clusters

	Model	Accuracy	Precision
0	XGBClassifier	0.999086	0.999086
1	KNeighborsClassifier	0.988055	0.988022
2	RandomForestClassifier	0.998947	0.998946
3	GradientBoostingClassifier	0.980743	0.980794
4	stacked_model	0.999194	0.999194

Table 4.1 Regressors for Location area

	Model	RMSE	r2_score
0	LinearRegression	83.071309	0.894996
1	KNeighborsRegressor	6.514577	0.999354
2	RandomForestRegressor	4.305939	0.999718
3	GradientBoostingRegressor	30.369323	0.985966
4	StackingRegressor	4.178288	0.999734

Tables 2.0 and 2.1 above present results of modelling the location cluster and area of accidents respectively.

When compared with the government baselines of **91%** and **67%,** the stacking classifier achieved better accuracy and precision of **95%** and **90%** respectively in predicting the severity of accidents to be serious or slight.

Recommendations

1. Since high traffic volume increases the risk of accidents, innovative regulation policies like the 'London Congestion Charge' which reduced traffic volumes by 30% in its first year (C40 Knowledge, 2022) should be introduced across the UK's central business areas.
2. Increased public awareness and policy formulation to influence the marketing of alternative fuel vehicles will affect the frequency and severity of accidents, since data shows that vehicles that utilize petrol are the most vulnerable.
3. Single carriageways experience less speeding traffic than dual carriageways creating a false sense of reduced risk to drivers. They have been reported to be up to 7 times more at risk (Eastern Daily Press, 2020). Cross hatching such roads as an alternative to simple markings might mitigate the associated risks. Awareness must also be made to help drivers adapt better to driving on such roads.
4. Separate lanes and paths for vulnerable road users such as Pedestrians and Cyclists must be implemented in the UK since they are involved in almost 50% of all accidents.
5. Improved enforcement of licensing regulations to bridge the data gap on vehicles involved in crashes.

Acknowledgments

For code, methods, or insights gleaned,

1. https://towardsdatascience.com/car-crashes-and-the-weather-an-exploratory-analysis-of-environmental-conditions-impact-on-traffic-12bcb7f9afed
2. https://www.scribbr.com/statistics/statistical-tests/
3. https://discuss.analyticsvidhya.com/t/how-to-find-percentage-of-total-with-groupby-pandas/66092/4
4. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
5. https://stackoverflow.com/questions/69409255/how-to-get-city-state-and-country-from-a-list-of-latitude-and-longitude-coordi
6. https://realpython.com/k-means-clustering-python/
7. https://stackoverflow.com/questions/34579213/dbscan-for-clustering-of-geographic-location-data
8. https://geoffboeing.com/2014/08/clustering-to-reduce-spatial-data-set-size/
9. https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/
10. https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

Bibliography

BBC, 2018. *UK rush hour drivers spend more than a day in traffic.* [Online]
Available at: https://www.bbc.co.uk/news/uk-england-42917201
[Accessed 24 April 2022].

C40 Knowledge, 2022. *How road pricing is transforming London.* [Online]
Available at: https://www.c40knowledgehub.org/s/article/How-road-pricing-is-transforming-London-and-what-your-city-can-learn?language=en_US
[Accessed 3 May 2022].

Cooper, M., n.d. *UK motorists unprepared for driving in bad weather.* [Online]
Available at: https://advancedmotoring.co.uk/uk-motorists-unprepared-for-driving-in-bad-weather/
[Accessed 22 April 2022].

Department of Transport, 2016. *National Travel Survey - Motorcycle use in England.* [Online]
Available at:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/694965/motorcycle-use-in-england.pdf
[Accessed 24 April 2022].

Eastern Daily Press, 2020. *Single-carriageway A-roads have 'seven times risk of motorways'.* [Online]
Available at: https://www.edp24.co.uk/news/traffic/single-carriageway-a-roads-have-seven-times-risk-of-motorways-608560
[Accessed 29 April 2022].

Staines, T., 2018. *Car Crashes and the Weather: An Exploratory Analysis of Environmental Conditions' Impact on Traffic Accident Rates.* [Online]
Available at: https://towardsdatascience.com/car-crashes-and-the-weather-an-exploratory-analysis-of-environmental-conditions-impact-on-traffic-12bcb7f9afed
[Accessed 21 April 2022].

Statista, 2020. *Licensed motorcycle numbers in the United Kingdom in 2020, by engine size.* [Online]
Available at: https://www.statista.com/statistics/314002/distribution-of-motorcycle-engine-sizes-in-great-britain-united-kingdom/
[Accessed 10 May 2022].

Statista, 2022. *Which of these propulsion types does your primarily used car use?.* [Online]
Available at: https://www.statista.com/forecasts/997807/car-propulsion-systems-by-type-in-the-uk
[Accessed 10 May 2022].

Tamboli, N., 2021. *All You Need To Know About Different Types Of Missing Data Values And How To Handle It.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/
[Accessed 10 May 2022].

UKDrivingSkills, 2021. *Avoiding Accidents in Rush Hour.* [Online]
Available at: https://www.ukdrivingskills.co.uk/avoiding-accidents-in-rush-hour/
[Accessed 22 April 2022].

WorldData.info, 2022. *Sunrise and sunset in the United Kingdom.* [Online]
Available at: https://www.worlddata.info/europe/united-kingdom/sunset.php#:~:text=The%20times%20of%20sunrise%20and,long%20and%20short%20in%20winter.
[Accessed 24 April 2022].