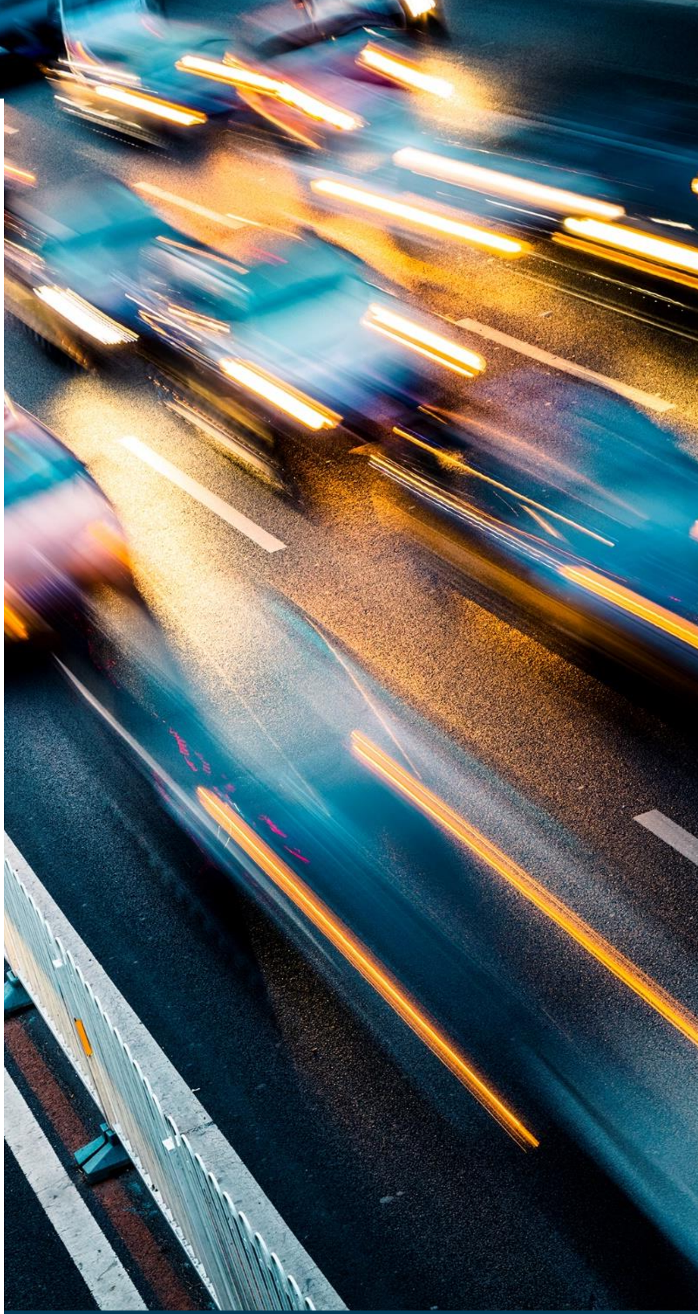


TRAFFIC ACCIDENTS IN THE UK 2021

Investigating traffic accident seriousness in the UK and the variables that are more likely to lead to a car collision.

06 NOVEMBER 2022

Ingrid Ionita, Hiu Ching (Clarissa) Lo, Michelle Obonyano, Ayomide Olarewaju

CODE FIRST GIRLS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

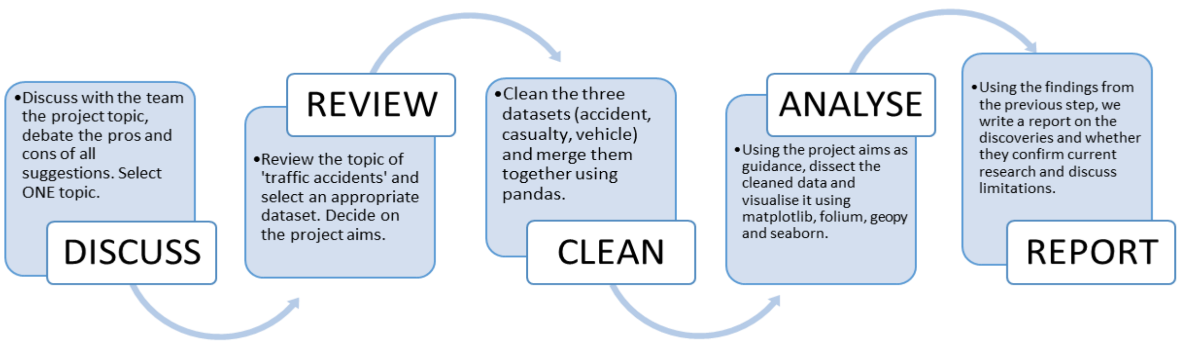
Aims and objectives

Annually, in the UK, there are approximately 150,000 people who are caught in a traffic accident. Every 16 minutes someone is killed or seriously injured in the UK (Brake, 2022), and that can be any of us as a passenger or a driver. From our own experiences with vehicles in the UK, we found it important to explore the factors that can lead to a traffic accident and consider potential solutions to reducing the risk of them happening. Our project aims are:

- Discover the circumstances in which traffic accidents are more likely to happen.
- Produce an analysis and visualisation that will highlight the answers to our project objective questions.
- Provide an insightful view into the traffic accidents that happened in 2021 and offer recommendations to the UK government in reducing the number of traffic accidents.

Report roadmap

Figure 1: *The project's process roadmap*



BACKGROUND

The project is relevant to the UK government as it has the ability to reduce traffic accidents through legislation or building new infrastructure to reduce the numbers of cars on the UK roads. Our project offers insights into the way a traffic accident can happen, based on environmental and social factors. This will give the audience a deeper insight into the dangerous collisions that happened in the UK in 2021 and allow for a possible solution to be explored. Our findings should be used to explore a possible solution that aims to reduce the number of traffic accidents in the UK that cost the government over £5 billion in human costs for serious and fatal accidents (Carlier, 2022). Therefore, our project objectives based on the aims are:

1. **Are certain drivers or casualties more likely to get involved in a severe car accident?** These drivers and casualties would be categorised by their age and gender.
2. **What dates are more likely to have a higher number of traffic accidents and cause more severe accidents?** Can this be linked to environmental factors such as the weather and the light conditions?
3. **What areas in the UK are more likely to have more severe traffic accidents?** This can be from the road type, junction location, county.
4. **Extension question: can machine learning be used to predict traffic accidents?**

STEP SPECIFICATIONS

Data gathering

We used sources from the UK government website, specifically the latest published dataset from 2021. This data will be in the form of a CSV file, to make it easier for Jupyter to handle. The datasets we will use are:

- **Vehicle:** information on the vehicle that was in an accident, from the model to the age of the car. It also shows the information about the driver, from age to gender.
- **Casualty**: this specifically looks at the casualties involved in the accident. It has information about their gender, age, and the gravity of the injuries.
- **Accidents:** combines information from the vehicle and casualty datasets but with the extra added information surrounding the accident. This is the time, the conditions in which the accident happened, the location, and how the police force dealt with the incident.

Data cleaning

The accident dataset contained 101,087 rows and 36 columns, the vehicle dataset contained 186,443 rows and 28 columns, and the casualty dataset contained 128,209 rows and 19 columns. Missing values (NaN) values were found, and the same were eliminated because the total figure had little to no bearing on the concatenated dataset. The dataset had no duplicates.

Furthermore, the dataset contained some missing or out-of-range data points represented by the value -1 and it was cleaned using the mean, mode, and median processes. For features such as junction control which had over 44% as negative figure, instead of using standard deviation techniques, they were combined with the unknown column and further transformed into a positive binary figure during model training. However, some columns were dropped as they had little or no impact in our analysis.

Data analysis and visualisation

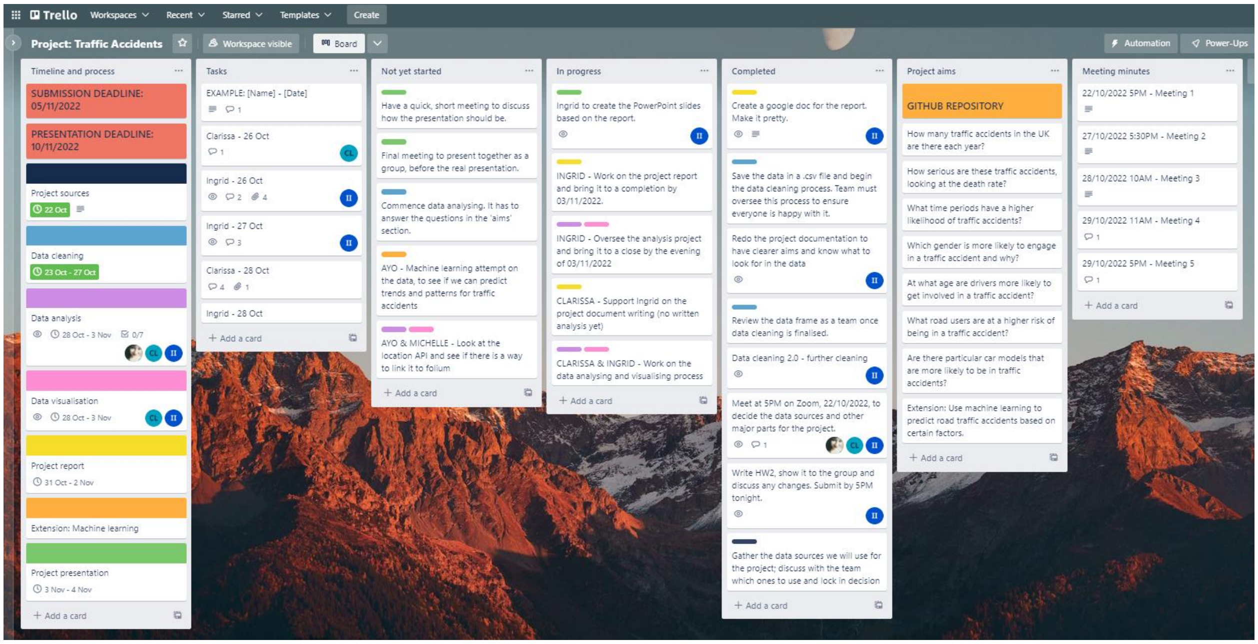
The team conducted an in-depth analysis, using a bivariate analysis to discover relationships between the independent variable (casualty severity and accident severity) with the dependent variables of our choice. We also used geocoding to better visualise the accident locations through geopy, Nominatim and folium. Our analysis used a range of visualisations from line plots to heatmaps to better understand the data and to visualise it accordingly.

IMPLEMENTATION AND EXECUTION

Team member roles and agile planning

Due to the time to work on this project, we found it imperative that we understood each other's strengths and weaknesses (Appendix A). This allowed the team to discuss the project topic, the team roles and the work distribution. We undertook an agile planning approach by having daily team meetings to discuss how the project was coming along, what needed to be done and to identify areas we needed to support. We used Trello to organise the team efforts and to visualise our tasks better.

Figure 2: *Trello board of the group project taken on 29/11/2022*



Furthermore, we assigned team member roles based on strengths to ensure that the team knew what it was doing (see Table 1).

Table 1: *Team member roles*

MEMBER NAME	ROLE
Ingrid Ionita	Scrum master, Main data cleaner, Main analyst
Hiu Ching (Clarissa) Lo	Lead visualisation, Main analyst
Ayomide Olarewaju	Lead data cleaner, Lead machine learning
Michelle Obonyano	Main analyst, Main visualisation

Tools and libraries

In order to work on this project, we used libraries taught through the CFGdegree course and some libraries that we discovered online, and thought would enrich our project. The table below shows the full libraries and tools we used and their methods:

Table 2: *Tools and libraries used by the team*

LIBRARY / TOOL	METHOD	USED FOR
Pandas	read_csv, to_csv, describe, info, value_counts, iloc, merge, sort_values, fillna	Data analysis, Data cleaning
NumPy	min, max, std, median	Data analysis
Matplotlib	Pyplot	Data visualisation
CSV	to_csv, read_csv	Data visualisation
Seaborn	catplot, heatmap, set, lineplot	Data visualisation
Functools	Reduce	Data visualisation

Folium	Map, Marker, HeatMap	Data visualisation
Geopy	Geocoders, Nominatim API, Pointers, RateLimiter, Reverse	Data visualisation
SciKit-Learn	Model_selection, linear_model, metrics, tree, feature_selection, model_selection, neighbors, ensemble, decomposition, over_sampling	Machine learning
Warnings	.filterwarnings	Machine learning
Trello		Project management
Google Drive		File management
GitHub		Project management
Zoom		Meetings
Google Meet		Meetings
Slack		Discussions

Limitations

Throughout the assignment, the team encountered limitations that added difficulty to the completion of the project. There are two main ones:

- **Communication:** The team faced difficulty at the beginning of the project, because there was a lack of leadership and unity, and an added big time zone difference. Each member of the team had responsibilities outside of the CFGdegree that made it hard to discuss the project as a group, which delayed the work by two weeks. The team acknowledges that the output of the project would have significantly improved if we united much sooner as a group.
- **Data used:** The data used for our project only looked at the year 2021, which limits the overall analysis to a time that was still during COVID-19 pandemic. The team acknowledges the external factors that impact the data, such as the lockdowns that resulted in an economic slow-down and the lack of travel. Therefore, in future research it would be beneficial to conduct a time analysis, comparing years to view the increase or decrease in traffic accidents and their severity.

RESULT REPORTING

QUESTION 1: Are certain drivers or casualties more likely to be in a severe car accident?

Total accidents and casualties by gender and age

To examine this question, we looked at the total accidents involved by driver and casualty gender and age. In terms of their gender, we found that male drivers (n = 47668) and casualties (n = 41381) are twice more likely to be involved in a traffic accident compared to female drivers (n = 20265) and casualties (n = 26552). Surprisingly, for males, there were more drivers than casualties involved. However, for females, the opposite pattern was shown.

In terms of age, there were also differences in total accidents involved between driver ages as well as between casualty ages (Table 3). Notably, for both drivers and casualties, the number of accidents peaked at between ages of 26-35, followed by the ages between 36-45 and 46-55.

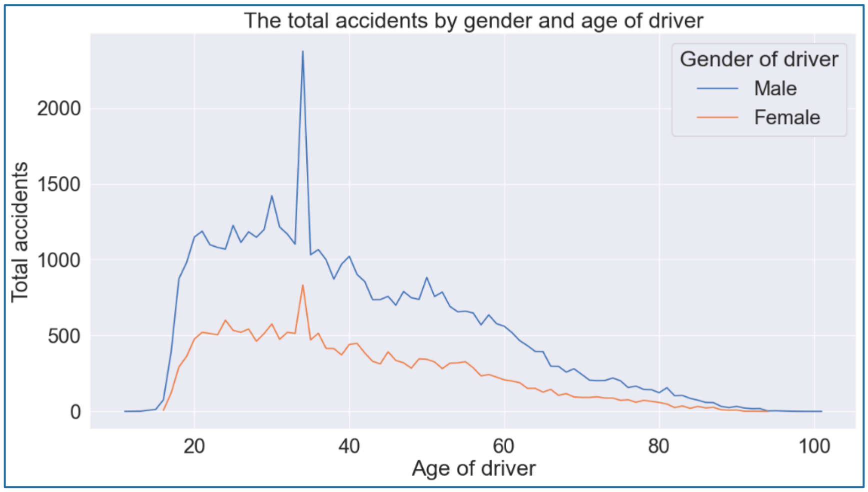
Table 3: *Total accidents by age*

	0-5	6-10	11-15	16-20	21-25	26-35	36-45	46-55	56-65	66-75	75+

Driver	0	0	28	4746	8343	16766	12963	10632	7238	3426	2156
Casualty	948	1258	1877	6462	8708	15791	11227	9549	6609	3199	2305

We plotted the total accidents of drivers by their gender and age (see Figure 3). Similar to the above, the line graph shows that male drivers had more accidents than female drivers at all ages. Interestingly, for both genders, the total accidents peaked at the age of 34 (males = 2374 accidents, females = 834 accidents).

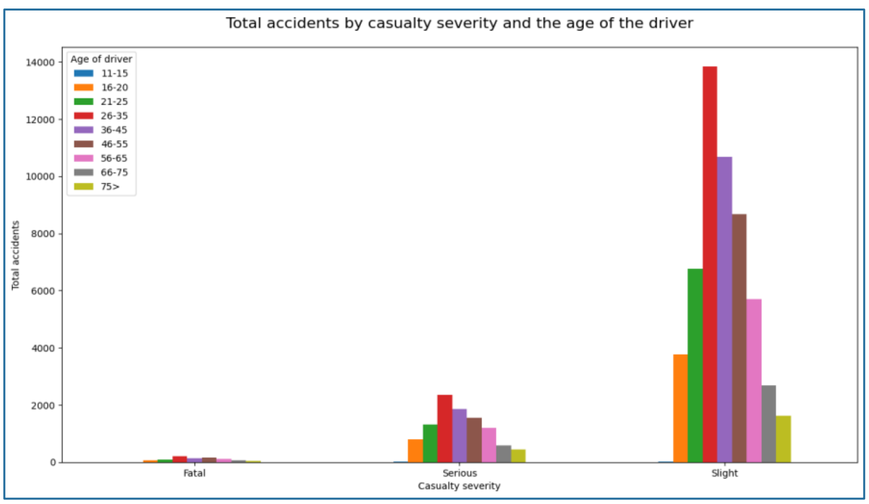
Figure 3: *Total accidents by gender and age of driver*



We also looked at the distribution of casualties through a line graph. Generally, the line graph illustrates that males were more likely to be involved in casualties than females, with the exception at the ages of 72 (male = 132, female = 141), 79 (male = 89, female = 113), 86 (male = 46, female = 54). The highest number of male casualties were at the age of 27 (*n* = 1152.0) and females at 22 years old (*n* = 715.0).

The distribution of casualty severity by gender and age

Figure 4: *Total accidents by casualty severity and age of driver*



We are also interested in whether casualty severity would differ according to gender and age of drivers. In terms of gender, male drivers cause more slight, fatal, and serious accidents than females. As for age, as plotted in the bar chart in Figure 4, drivers at 26-35 years old are the most likely to cause accidents at all casualty severity levels, followed by those aged at 36-45 and 46-55. In all age groups, the majority of accidents are slight.

Possible explanations for the gender and age distribution of total accidents

Across all analyses, more male drivers were involved in traffic accidents. In general, there were also a higher number of male casualties. According to past research, these findings may be because men are more likely than women to drive (Fergusson *et.al.*, 2003). Moreover, men pose a higher risk to road users in comparison to women (Aldred *et al.*, 2019) as they are more likely to engage in riskier driving such as speeding, drunk driving and not using a seatbelt (Insurance Institute for Highway Safety, 2020). Given that, they may be more likely to be involved and be injured in traffic accidents as backed by past studies suggesting that traffic accidents involving men are often more severe than those involving women (Li *et al.*, 1998).

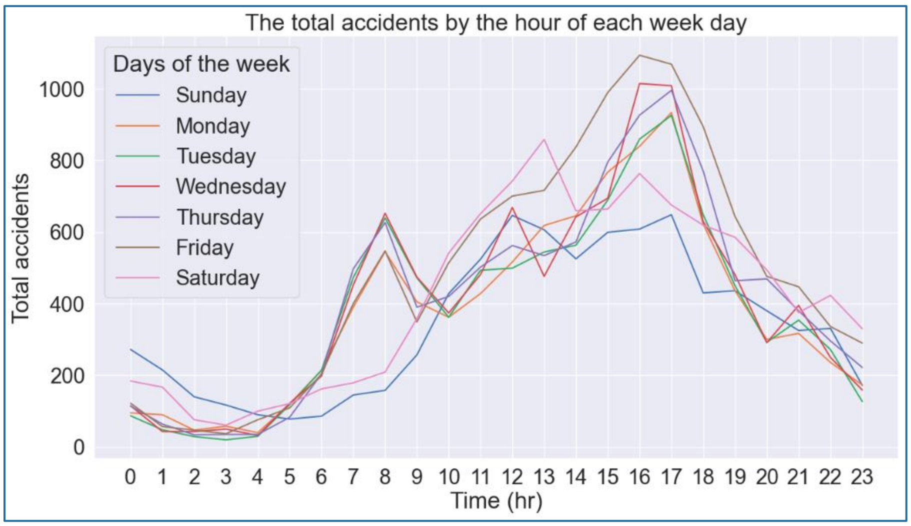
We also found that most drivers and casualties were involved in accidents at the age of 26-35. Compared to teenagers, people at this age group may be more likely to own a car (Thakuriah *et al*., 2010). However, compared to older people, they may be less experienced on the road and tend to underestimate driving risks, resulting in higher risk of accidents (Kahane, 2013).

Yet, it is unclear the reason behind the significantly high number of total accidents at male and female drivers aged 34, but this finding is supported by previous research (Regev *et al*., 2018). This pattern is likely to be caused by a combination of the reasons explained above, including that males and people at this age are more likely to take risks while driving (Thakuriah *et al.,* 2010; Kahane, 2013).

QUESTION 2: What dates are more likely to have a higher number of traffic accidents?

The distribution of traffic accidents by month, day and hour

Figure 5: *Total accidents by the hour of each weekday*



To investigate the second question, we looked at the hour, day, and the month variables to discover the times a traffic accident is more likely to take place. Looking at the hour of each day of the week (Figure 5), we discovered that the total traffic accidents peaks at 8AM and 4PM - 5PM for the weekday and 12PM - 1PM and 4PM - 5PM for a weekend.

Possible explanations for the month, day and hour difference

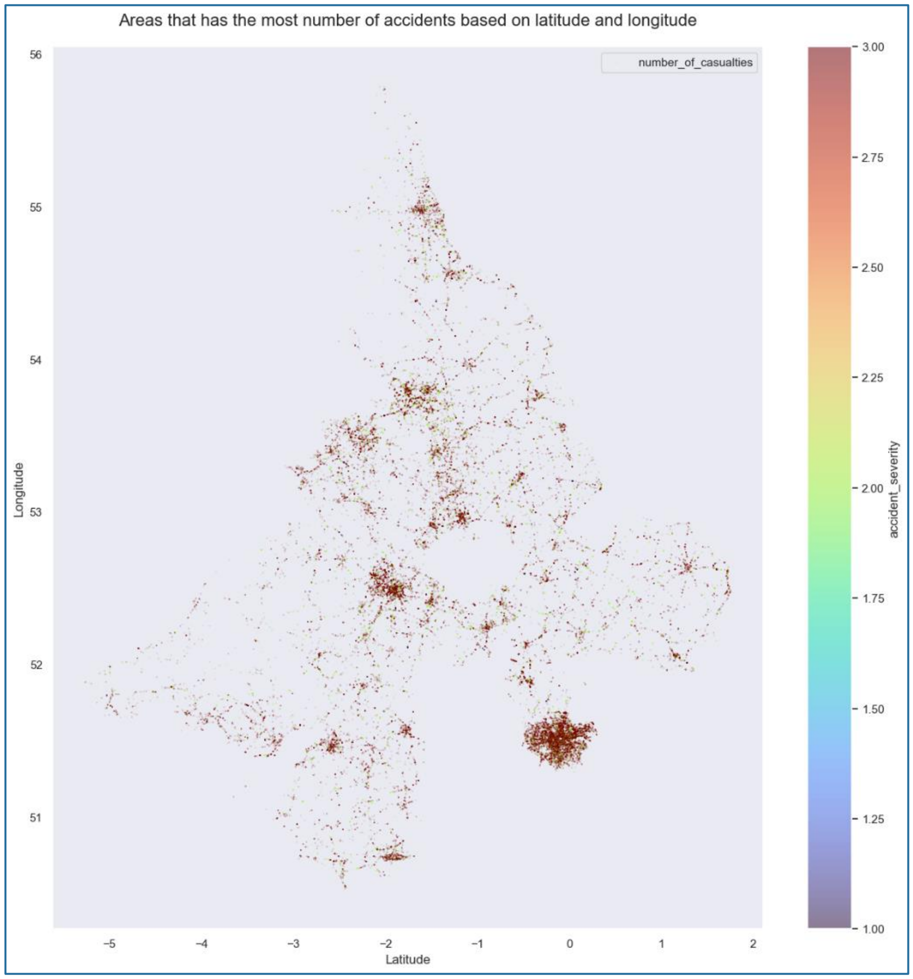
According to YouGov (2018), 37% of Britons work 8 to 4, 21% work 7 to 3 and 16% work 9 to 5. The pattern in Figure 5 highlights how the work pattern can affect the number of traffic accidents. We discovered that Friday (n = 22,533), Wednesday (n = 19,493), and Thursday (n = 19,977) have the highest number of traffic accidents, which suggests that the workday fatigue may affect the driving quality; this is in line with current research. Driver fatigue is a major safety issue, which is a result of poor sleep quality and increased fatigue from sleep deprivation (Caponecchia & Williamson, 2018). Research suggests Friday, Wednesday and Thursday are the most tiring for men and women (Sabau & Niculescu, 2020) and there is also a linear increase in evening fatigue across weekdays predicted by sleep quality and after-work fatigue (Elfering *et al.,* 2021).

Furthermore, research suggests that summer months are more likely to have car collisions due to the increased risk of tire blowouts, overheating engines and other equipment malfunctions, which lead to poor vehicle control. Our findings (Appendix B) are in line with this research, as we found that May (n = 13,121), June (n = 12,190), July (n = 12,661) and August (n = 12,135) had the highest number of traffic accidents in 2021.

QUESTION 3: What areas in the UK are most likely to have traffic accidents?

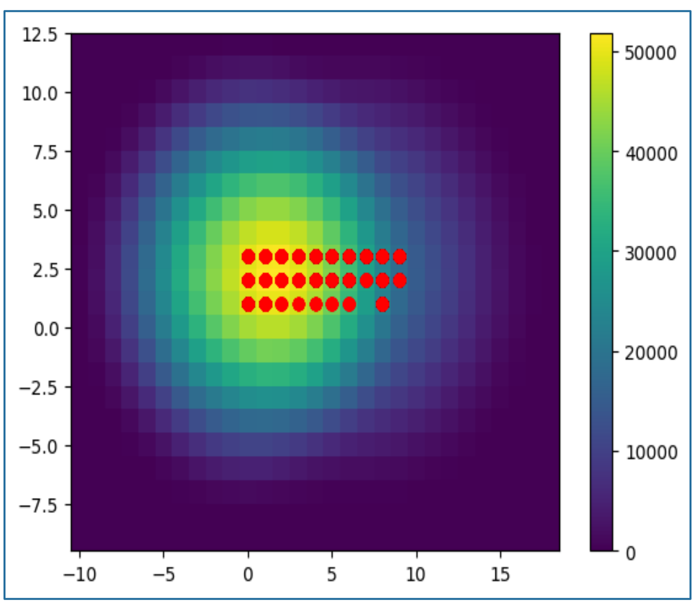
The distribution of traffic accidents across the UK

Figure 6: *Location of the traffic accidents in the UK*



We investigated the relationship between the traffic accident severity and the UK areas by using scatter colormaps and heatmaps. The variables considered were latitude, longitude, accident severity, accident casualty and junction location. Figure 6 determines the areas in the UK with the highest number of accidents; the x-axis represents the longitude data and y-axis represents the latitude data. Therefore, the highest number of accidents centres around longitudes -1 and +1, and latitudes 50 and 51, which is the South, South-West, and South-East of the UK. We used the Sklearn module and imported Kmeans followed by clusters. Figure 6 shows the area that has the highest number of traffic accidents, which is London (n = 49,386). This is also visible in the Folium heatmap (Appendix C) that shows that the big cities, including London, have the higher concentration of traffic accidents which we attribute to the high population density in the area.

Figure 7: *Junction location impact on accident severity*



We also plotted a heatmap plot to show how junction locations impact the accident severity (Figure 7). While junction location 0-9 has quite a high level of impact on accident severity, there is a higher level of impact on accident severity with the addition of junction location 10 (i.e., 0-10). Most traffic accidents happen not at or within 20 metres of the junction (n = 26,725), approaching a junction (n = 14,094) or mid-junction (n = 10,280), with the highest number of traffic accidents being seen on single carriageways (n = 41,604) and dual carriageways (n= 15,521). Based on our analysis, findings indicate that London has the highest number of road accidents and junction location numbers between 0 and 10 have the highest-level accident severity.

Possible explanations for traffic accident volume across the UK

Fuelgine (2022), while quoting the Royal Society for Prevention of Accidents, asserted that the common causes of road traffic accidents are human error, environmental issues, and mechanical faults. London is one of the most populated cities in the world, the fifth city out of the top ten that is prone to traffic accidents Maguire (2022). Mortality tends to be higher in rural areas in comparison to urban areas (Clark, 2003) and that the number of accidents in urban areas depends on population size super linearly (Cabrera-Arnau *et al*., 2020). Junctions require a higher level of attention from both the driver and the pedestrians, and failure to give way and inappropriate manoeuvres are one of the main factors in junction collisions (Nitsche *et al.,* 2017). We found that a major number of collisions on the UK roads happen at a junction or nearby it.

QUESTION 4: Can machine learning be used to predict traffic accidents?

Machine learning experimental setup

Our dataset comprises 69 features that are almost independent of one another, indicating that it is extremely complex and that not all elements are significant for accuracy prediction. The concatenated dataset was separated into 80% train sample size, 20% test sample size, and the random state as 99 using the Sklearn train test split. On the train sample, four classifiers were used which are Decision Tree, Random Forest, K-NN and Logistic Regression. Oversampling technique, known as SMOTE, was used to balance the dataset's imbalanced class 'accident_severity', which raised the number of minority class samples (serious, fatal) and the same algorithms were applied.

Performance metrics for classification

The performance of the classifier was evaluated using the confusion matrix, which is known to evaluate the correctness of a classification (Sokolova & Laplame 2009), and its associated metrics which are accuracy, precision, and recall. Elements of the confusion matrix are:

- True Positives (TP) : for event values that were successfully predicted
- False Positive (FP) : for event values that were mistakenly predicted.
- True Negative (TN) : for successfully anticipated no-event values.
- False Negative(FN) : for no-event values that were mistakenly predicted.

Table 4: *Result - unbalanced dataset*

MODELS	Accuracy	Precision	Recall	F1-score
Random Forest	96.4	96.4	96.4	96.4
Decision Tree	93.8	94.0	93.8	93.8
K-NN	94.2	94.3	94.2	93.7
Logistic Regression	84.5	83.0	85.0	83.0

Table 5: *Balanced (SMOTE) dataset*

MODELS	Accuracy	Precision	Recall	F1-score
Random Forest	98.0	98.0	98.0	98.0
Decision Tree	96.8	96.8	96.8	96.8
K-NN	80.4	80.2	80.4	80.0
Logistic Regression	93.0	93.0	93.0	93.0

With a 98% accuracy rate, the balanced dataset of the Random Forest model is the best classifier for predicting accidents, as shown in Table 4. However, accuracy is not the only performance metric, and the same model also demonstrated superior predictive abilities for identifying the minority class in the confusion matrix when compared to algorithms trained on the unbalanced dataset.

CONCLUSION

In conclusion, our research presents a snippet into the UK drivers: who they are based on their age and gender, the times throughout the year in which a traffic accident may occur, and the location of the traffic accidents. We discovered that there may be a link between fatigue and traffic accidents based on the day and the hour, which could be attributed to work tiredness. This should be further investigated in future studies, to determine whether the current working pattern affects the increase or decrease in traffic accidents. Junctions on dual and single carriageways are more likely to have a traffic accident and our current data does not fully explain the reason behind it. We recommend that this should be investigated further to discover a method of reducing car collision numbers.

We further investigated machine learning and discovered a way that it can be used to predict a traffic accident happening before it does. Because traffic accident severity prediction is vital for accident management, we can conclude that machine learning is a potential method for predicting and ultimately avoiding road traffic accident severity. However, to accurately predict higher-severity incidents, a balanced dataset is required. Furthermore, based on the machine learning's model we also recommend the following:

- **Traffic congestion:** In urban areas it should be reduced through a variety of methods and initiatives, including improved bus service, workplace parking fees, and existing rail networks.
- **Speed limit:** Installing speed and highway cameras to catch vehicles exceeding speed limits, particularly on single carriageways, which are infamous for accidents, and enforcing speeding fines.
- **Safety education:** Proper road safety education is critical, so it's important to promote knowledge about road safety protocols and best practices.
- **Reporting channels:** Information on how to report dangerous drivers as well as penalties for reckless driving should be readily available.
- **Proper traffic management:** The majority of accidents seem to have happened at uncontrolled junctions due to the absence of proper traffic management mechanisms like traffic signals and bumps, these should be considered by the government.

Lastly, our recommendations based on our research and the current existing research is to invest in better public transport infrastructure, which will reduce the pollution levels and the number of cars on the UK roads. Furthermore, for every 100 million increase in traffic investment, the number of road traffic casualties is reduced which shows that economic development can improve the traffic accident number (Sun *et al.,* 2019).

REFERENCE LIST

Aldred, R., Johnson, R., Jackson, C. & Woodcock, J., 2021. How does mode of travel affect risks posed to other road users? An analysis of English road fatality data, incorporating gender and road type. *Injury Prevention,* Volume 27, pp. 71-76.

Baker, C., 2022. *Constituency Data: Road traffic accidents - house of commons library.* [Online]
Available at: https://commonslibrary.parliament.uk/constituency-data-traffic-accidents/
[Accessed 4 11 2022].

Brake, 2022. *UK collision and casualty statistics: How many people are killed and hurt on our roads?.* [Online]
Available at: https://www.brake.org.uk/get-involved/take-action/mybrake/knowledge-centre/uk-road-safety
[Accessed 25 10 2022].

Cabrera-Arnau, C., Prieto Curiel, R. & Bishop, S. R., 2020. Uncovering the behaviour of road accidents in urban areas. *Royal Society Open Science,* 7(4), pp. 1-12.

Caponecchia, C. & Williamson, A., 2018. Drowsiness and driving performance on commuter trips. *Journal of Safety Research,* Volume 66, pp. 179-186.

Carlier, M., 2022. *Average cost of road casualties and accidents by severity in Great Britain in 2020.* [Online]
Available at: https://www.statista.com/statistics/322862/average-cost-of-road-accidents-and-casualties-in-great-britain-uk/
[Accessed 25 10 2022].

Clark, D. E., 2003. Effect of population density on mortality after motor vehicle collisions. *Accident: Analysis and Prevention,* 35(6), pp. 965-71.

Elfering, A., Kottwitz, M. U. & Semmer, N. K., 2021. Battery discharge from Monday to Friday: Background social stress at work is associated with more rapid accumulation of fatigue. *Sleep and Vigilance,* Volume 5, pp. 49-60.

Fergusson, D., Swain-Campbell, N. & Horwood, J., 2003. Risky driving behaviour in young people: prevalence, personal characteristics and traffic accidents. *Australian and New Zealand Journal of Public Health,* 27(3), pp. 337-42.

Fuelgenie, 2022. *The most common causes of road accidents, and how to avoid them.* [Online]
Available at: https://www.fuelgenie.co.uk/news/the-most-common-causes-of-road-accidents/
[Accessed 4 11 2022].

Gan, J., Li, L., Zhang, D., Yi, Z. & Xiang, Q., 2020. An alternative method for traffic accident severity prediction: Using deep forests algorithm. *Journal of Advanced Transportation,* pp. 1-13.

Li, G., Baker, S. P., Langlois, J. A. & Kelen, G., 1998. Are female drivers safer? An application of the decomposition method. *Epidemiology,* Volume 9, pp. 379-384.

Maguire, J., 2022. *UK's most accident-prone cities revealed.* [Online]
Available at: https://www.fleetnews.co.uk/news/car-industry-news/2022/01/24/uk-s-most-accident-prone-cities-revealed
[Accessed 4 11 2022].

Nitsche, P., Thomas, P., Stuetz, R. & Welsh, R., 2017. Pre-crash scenarios at road junctions: A clustering method for car crash data. *Accident Analysis and Prevention,* Volume 107, pp. 137-151.

Regev, S., Rolison, J. J. & Moutari, S., 2018. Crash risk by driver age, gender, and time of day using a new exposure methodology. *Journal of Safety Research,* Volume 66, pp. 131-140.

Sabau, E. & Niculescu, G., 2020. Perceptions of fatigue in young genders. *Journal of Sport and Kinetic Movement,* 36(2), pp. 60-65.

Sun, L. L., Liu, D., Chen, T. & He, M. T., 2019. Road traffic safety: An analysis of the cross-effects of economic, road and population factors. *Chinese Journal of Traumatology,* 22(5), pp. 290-295.

Thakuriah, P., Menchu, S. & Tang, L., 2010. Car ownership among young adults: Generational and period-specific perspective. *Transportation Research Record,* 2156(1), pp. 1-8.

YouGov, 2018. *Over nine in ten not working the usual 9-5 week.* [Online]
Available at: https://yougov.co.uk/topics/economy/articles-reports/2018/08/24/over-nine-ten-not-working-usual-9-5-week
[Accessed 31 10 2022].

APPENDICES

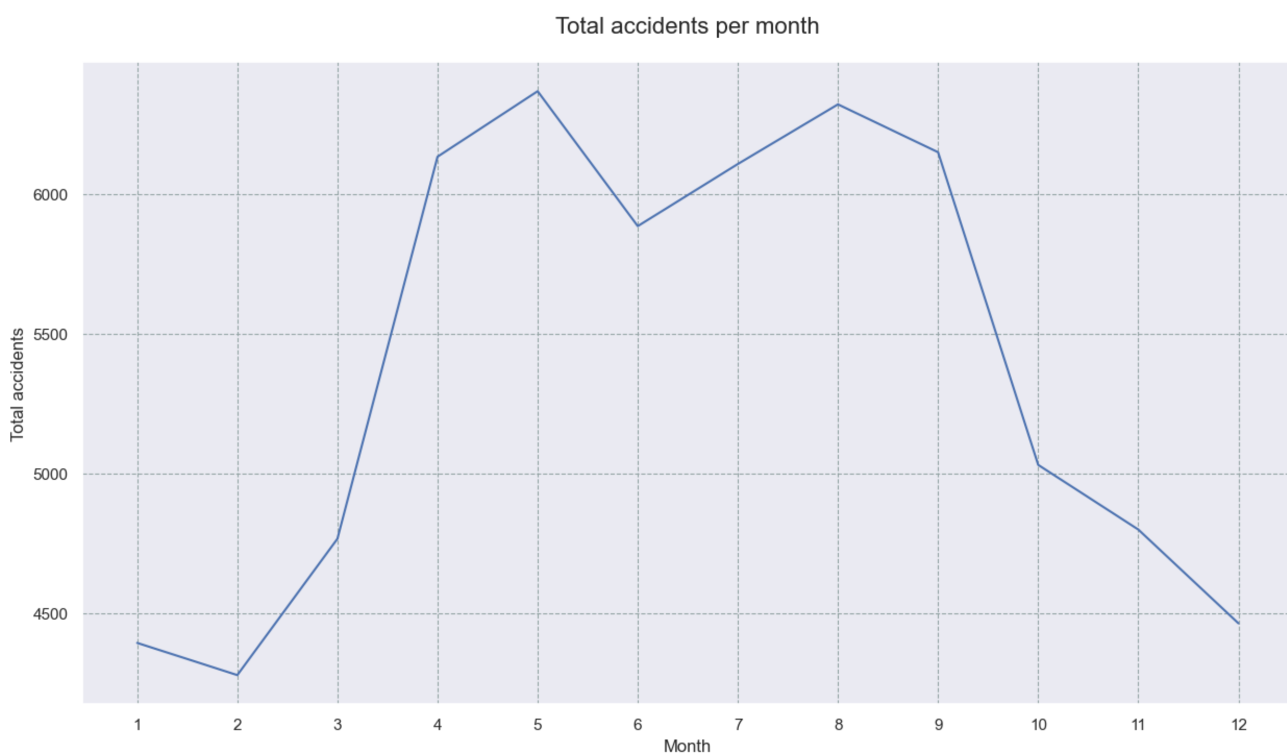
Appendix A:

SWOT team analysis created at the beginning of the project, in week 1 of the specialization.

STRENGTHS	WEAKNESS
• Academic writing • Presenting • Python in general • Marketing and advertising • Psychology • Data analysis	• SQL • Databases • Financial terms/banking • Marketing/ business/ banking • APIs
OPPORTUNITY	**THREAT**
• N/A	• University deadlines at the same time as the CFGdegree ones • Return to HongKong, different time zone

Appendix B:

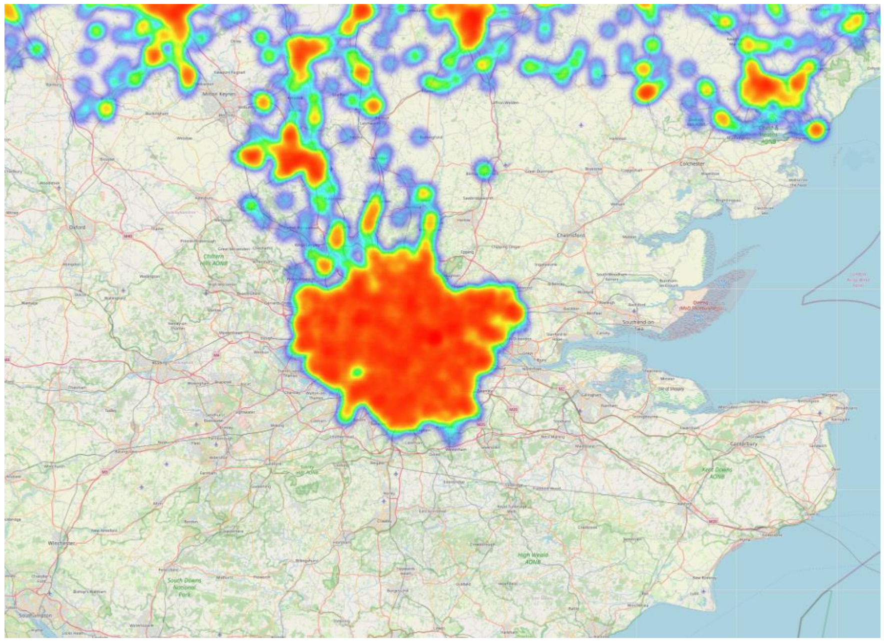
The months with the highest number of accidents.



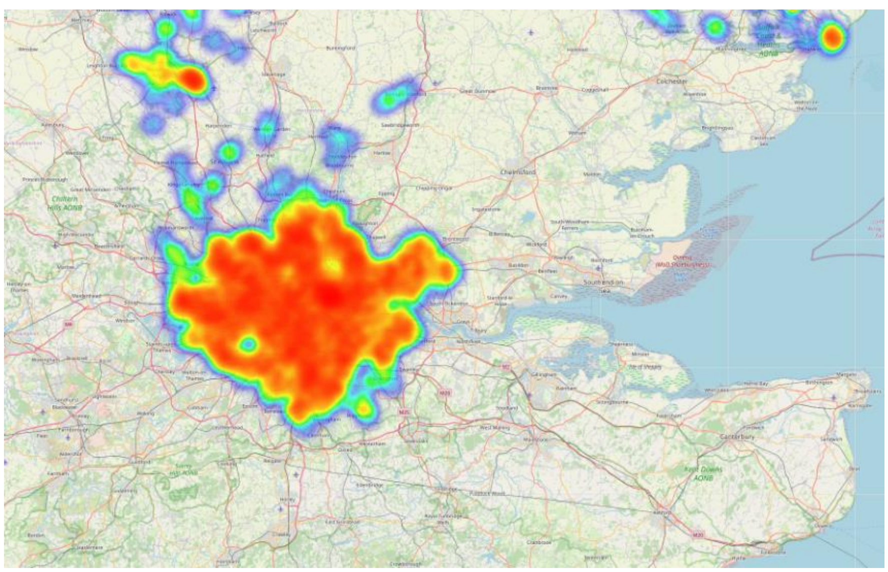
Appendix C:

Folium heatmaps and markers showing the number of traffic accidents based on their severity.

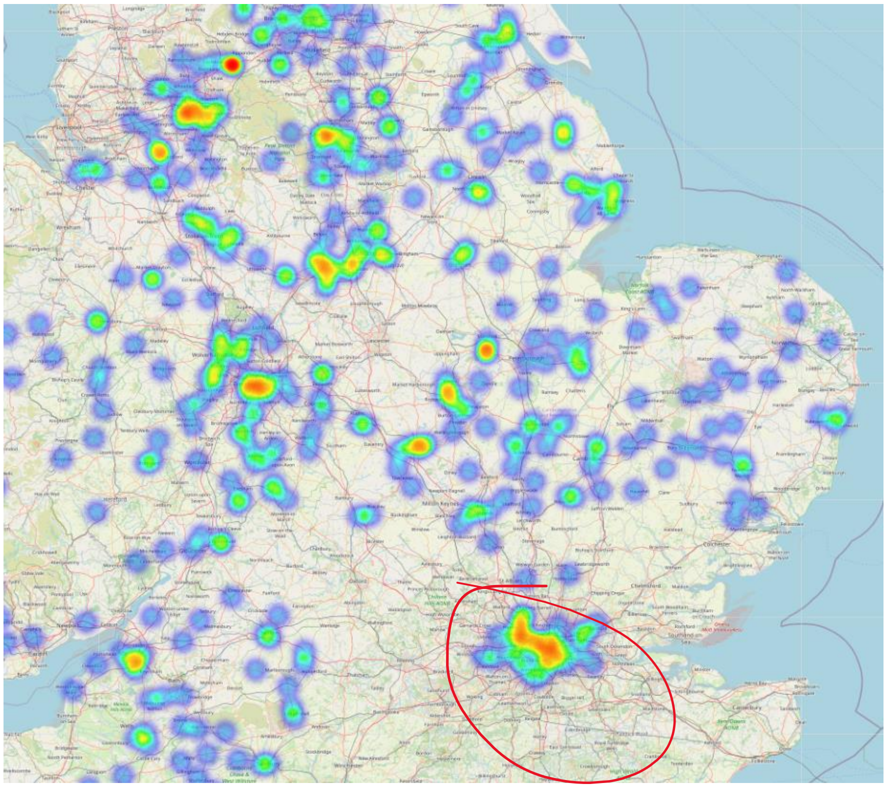
1. Heatmap showing the slight accidents in London.



2. Heatmap showing the serious accidents in London.



3. Heatmap showing the fatal accidents in London.



4. Map with markers for the fatal accidents in London.