ACCIDENT DATA REPORT

This report examines accident statistics recorded in the United Kingdom in 2020. Police collected data at the scene of a collision or, in some instances, reported by a member of the public at a police station or online, according to the Department for Transportation (2022). For each collision, 50 data points were collected by the police, including the time and location of the crash, the types of cars involved and what they were doing at the time of the collision, and information on the drivers and casualties. The dataset was retrieved from a SQL database and then transformed into a pandas dataframe before cleaning and analyzing.

This report aims to gather insights into accident patterns and trends to predict fatal injuries incurred on the road while informing and enhancing road safety measures. Due to the nature of the dataset, extensive data cleaning must be performed to effectively handle missing values and inaccurate data entries before analyses and insights can be achieved.

DATA CLEANING

As can be seen from the Jupyter Notebook, the following issues will be addressed:

- Missing values in the accident table
- Columns in accident, casualty and vehicle tables have -1 present.
- 99 in the police force column
- Columns in accident, casualty and vehicle tables have 0, 9 and 99 present.
- Inconsistent data entry in the vehicle table

Longitude, latitude, location_easting_osgr and location_northing_osgr missing values were fixed with the mode of the approximate value of the location nearest to them based on local_authority_ons_district. The local_authority_ons_district column, according to STATS20 Department for Transportation (2011), was used to address 99 in the police force column.

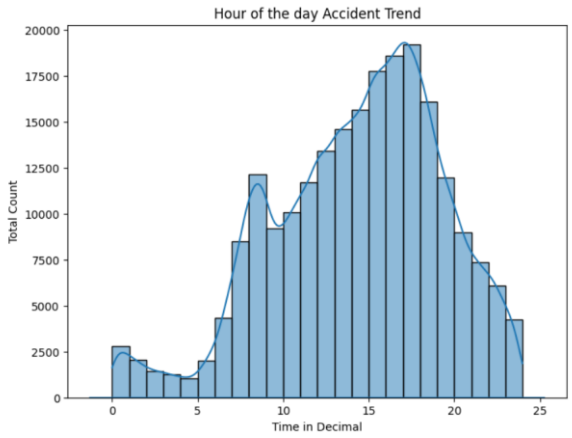
-1 and some 0, 9 and 99 values in the three tables were initially replaced with NaN. Columns with the class as unknown, depicting that the condition couldn't be ascertained, were used to replace the NaN values in them, e.g., second_road_class, light_conditions, weather_conditions and junction_control while kernel density estimation (KDE) and fillna using forward fill were used depending on the type of variable. The Lsoa_of_accident_location in Scotland and Wales were determined using the mode of the district_ons as described in Doogal (2021).

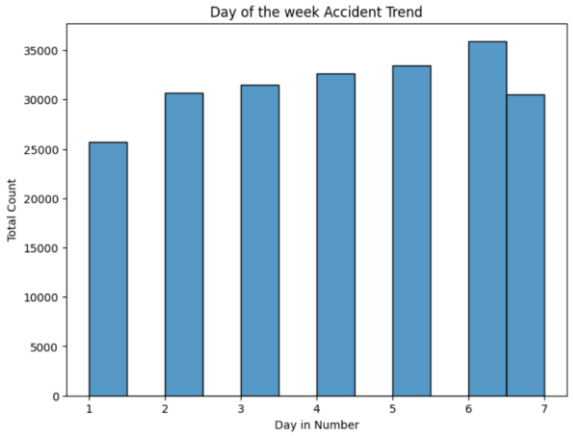
STATS20 Department for Transportation (2011) was a crucial deciding factor in interpreting the dataset and considerably aided in correcting these missing or out-of-range values and ensuring a cleaned dataset. Finally, using an inner join with accident_index as a reference, the cleaned tables were merged into a single dataset.

EXPLORATORY ANALYSIS

According to the histograms below, the hour of day and day of week with the most accidents were 17:00 and Fridays. According to STATS20 (2011), the week begins on Sunday and ends on Saturday; hence the numbers 1 and 6 on the plot represent Sunday and Friday, respectively.

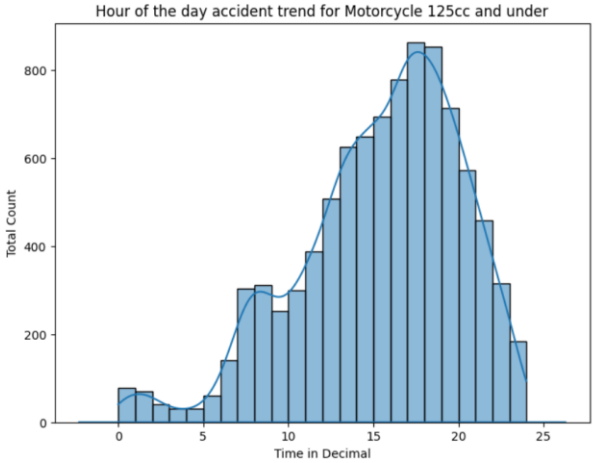
Furthermore, at least 1974 and 35938 accidents happened in 2020 at these times and days. Liverpool, Merseyside, had the most accidents at 17:00 on Fridays with daylight, fine without high winds, and dry road surface, accounting for 21 accidents in 2020 under the given conditions. In general, 422 accidents occurred at 17:00 hours and on a Friday.

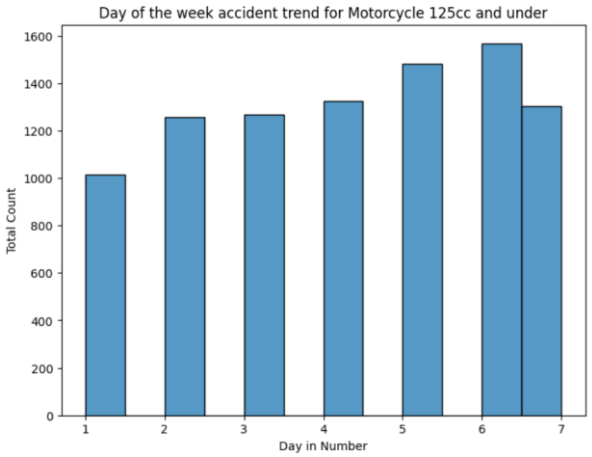




In the same year, nine accidents with slight injuries happened in Liverpool, Merseyside, at a speed limit of 30 mph, T, or staggered junction, and with male drivers, while nine occurred at the exact city and speed but at a slip road by unknown drivers.

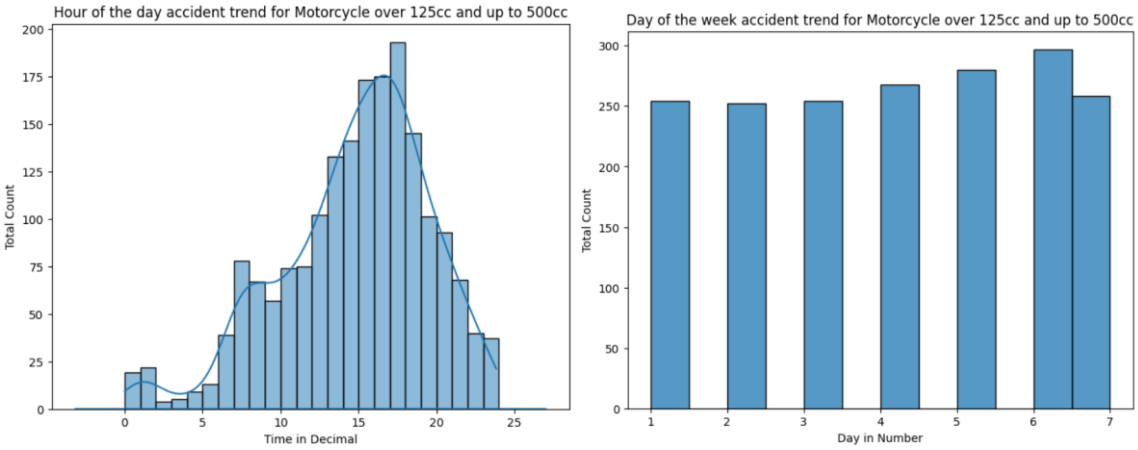
Motorcycles 125cc and under had the most accidents on Fridays and at 17:00hrs. At 17:00, 20 accidents occurred on Fridays. Northeast Derbyshire had the most accidents at 17:00 on Fridays with daylight, fine without high winds, and dry road surface, with two accidents.



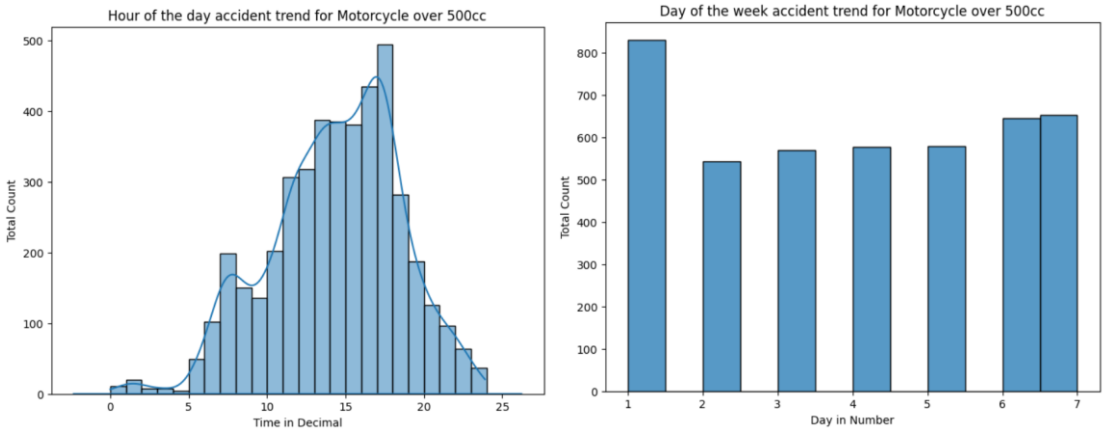


As seen below, motorcycles 125cc and up to 500cc accidents were most common at 17:00 and on Fridays. 26 and 297 accidents happened at 17:00 and on Fridays, respectively. The City of Bristol, Avon, and Somerset had one accident at 17:00 and on Fridays with daylight, fine without high winds, and dry road surface. One non-fatal accident occurred at a T or staggered junction in the city of Bristol, Avon, and Somerset, with male drivers traveling at 20 mph. Four accidents occurred at this time of the day and on this day of the week.

accident_severity	speed_limit	junction_detail	sex_of_driver	local_authority_ons_district	
3	20	3	1	E06000023	1
	30	0	1	E09000007	1
		1	1	E06000037	1
	40	3	1	E07000111	1



Most motorcycles over 500cc accidents occurred at 17:30 and on Sundays. At least 46 and 830 accidents happened during that time of day and on Sundays. Teignbridge, Devon, and Cornwall had the only accident at 17:30 on a Sunday when there was daylight, no severe winds, and a dry road surface. The accident was non-fatal, not at or within 20 metres of a junction, with a male driver traveling at 30 mph.



According to the histograms above, 197 of the 16,010 pedestrians most likely to be involved in an accident occurred between 15:30 and 2764 on Fridays. On Fridays, 17.3 percent of pedestrian accidents happened, of which 30 occurred by 15:30.

An association rule mining approach for discovering commonly used itemsets is employed to detect meaningful patterns between variables in the data. This rule studies the correlations between variables that result in the severity of an accident.

	support	itemsets
0	0.019194	(severity_1)
1	0.203362	(severity_2)
2	0.777445	(severity_3)
3	0.720126	(police_1)
4	0.194720	(police_2)
5	0.085154	(police_3)
6	0.638224	(driver_1)
7	0.260362	(driver_2)
8	0.101413	(driver_3)
9	0.060072	(impact_0)
10	0.506113	(impact_1)
11	0.169333	(impact_2)
12	0.133736	(impact_3)
13	0.130746	(impact_4)
14	0.436936	(junction_0)
15	0.075528	(junction_1)
16	0.266119	(junction_3)
17	0.029596	(junction_5)
18	0.093719	(junction_6)
19	0.020156	(junction_8)

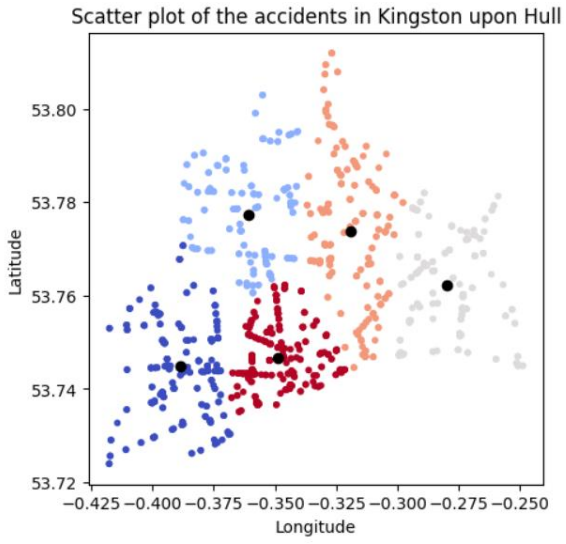
$$Rule: X \Rightarrow Y \rightarrow \begin{cases} Support = \dfrac{frq(X,Y)}{N} \\ Confidence = \dfrac{frq(X,Y)}{frq(X)} \\ Lift = \dfrac{Support}{Supp(X) \times Supp(Y)} \end{cases}$$

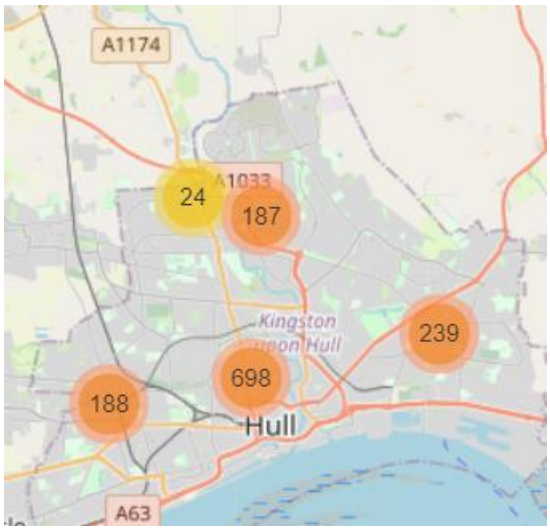
According to the table above, 77.7 percent of accidents were slight, 20.3 were serious and 0.02 were fatal, 63.8 percent of drivers were male, 50.6 percent were front collision, and 43.7 percent of accidents happened not at or within 20 metres of a junction.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
1	(police_1)	(severity_2)	0.720126	0.203362	0.170164	0.236297	1.161955	0.023718	1.043126	0.498014
4	(driver_1)	(severity_2)	0.638224	0.203362	0.142541	0.223340	1.098240	0.012751	1.025723	0.247259
7	(impact_1)	(severity_2)	0.506113	0.203362	0.113916	0.225080	1.106795	0.010992	1.028026	0.195369
8	(impact_3)	(severity_2)	0.133736	0.203362	0.028521	0.213263	1.048690	0.001324	1.012586	0.053597
9	(junction_0)	(severity_2)	0.436936	0.203362	0.098959	0.226484	1.113700	0.010103	1.029892	0.181315
13	(left_1)	(severity_2)	0.961835	0.203362	0.200531	0.208488	1.025208	0.004931	1.006477	0.644243
14	(pedestrian_0)	(severity_2)	0.794402	0.203362	0.170935	0.215174	1.058087	0.009384	1.015051	0.267015
17	(police_1)	(severity_3)	0.720126	0.777445	0.532756	0.739809	0.951590	-0.027102	0.855354	-0.153811
19	(police_2)	(severity_3)	0.194720	0.777445	0.164057	0.842532	1.083720	0.012674	1.413337	0.095932
20	(police_3)	(severity_3)	0.085154	0.777445	0.080631	0.946886	1.217947	0.014429	4.190158	0.195602
22	(driver_1)	(severity_3)	0.638224	0.777445	0.480686	0.753161	0.968765	-0.015498	0.901623	-0.081829
24	(driver_2)	(severity_3)	0.260362	0.777445	0.210475	0.808391	1.039806	0.008057	1.161510	0.051758
25	(driver_3)	(severity_3)	0.101413	0.777445	0.086284	0.850816	1.094376	0.007441	1.491822	0.095969
26	(impact_0)	(severity_3)	0.060072	0.777445	0.045419	0.756079	0.972518	-0.001283	0.912408	-0.029187
28	(impact_1)	(severity_3)	0.506113	0.777445	0.380475	0.751759	0.966962	-0.013000	0.896530	-0.064704
29	(impact_2)	(severity_3)	0.169333	0.777445	0.144083	0.850885	1.094464	0.012436	1.492513	0.103906
30	(impact_3)	(severity_3)	0.133736	0.777445	0.102579	0.767028	0.986602	-0.001393	0.955290	-0.015434
31	(impact_4)	(severity_3)	0.130746	0.777445	0.104888	0.802228	1.031877	0.003240	1.125310	0.035539
33	(junction_0)	(severity_3)	0.436936	0.777445	0.324962	0.743729	0.956633	-0.014732	0.868438	-0.074513
34	(junction_1)	(severity_3)	0.075528	0.777445	0.064055	0.848099	1.090880	0.005336	1.465134	0.090115

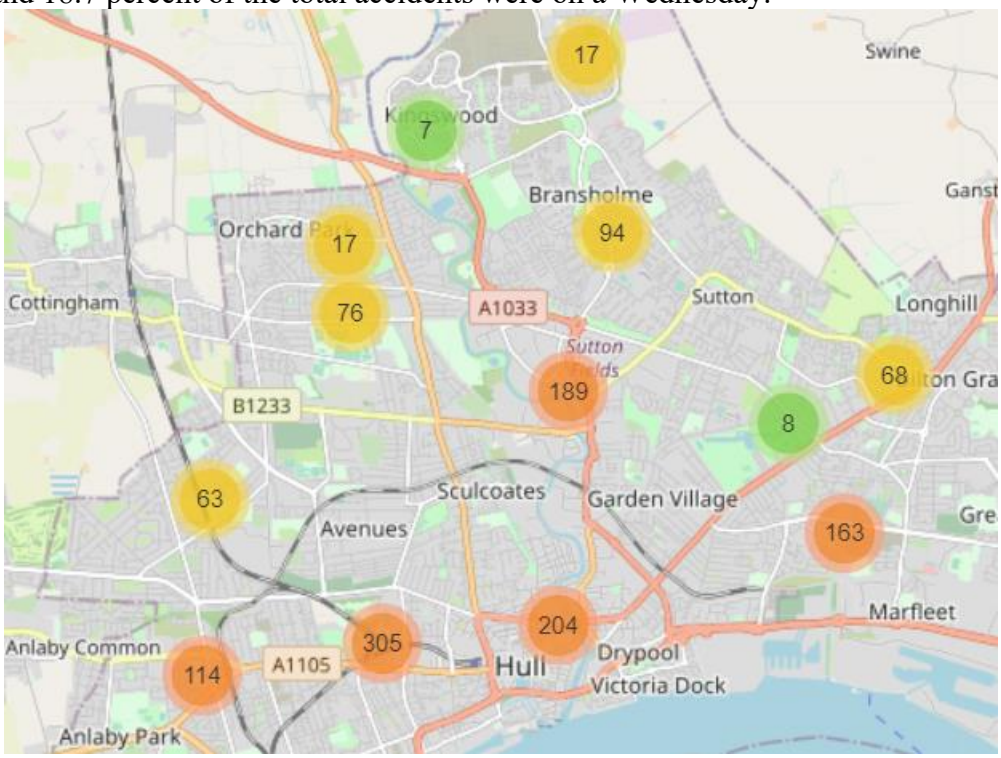
The antecedents are the conditions or events used to forecast the occurrence of the consequents. They serve as the foundation for predicting the presence or absence of the consequents. According to the above confidence table, 72 percent of the time a police officer visits an accident scene, 20.3 percent of the accident must be serious. Similarly, 63.8 percent of drivers are male, 50.6 percent collided with the front, 43.7 percent occur not at or within 20 metres of a junction, 96.2 percent were left-hand vehicles, and 79.4 percent of pedestrians have no physical crossing facilities within 50 metres.

Kingston upon Hull is a city located in Yorkshire's East Riding, England. Road accidents are a severe worry in Hull, as they are in every metropolitan region; thus, there is a need to analyze and obtain additional insights into road accidents that occur in Hull. The clusters below show the number of accidents in Hull.



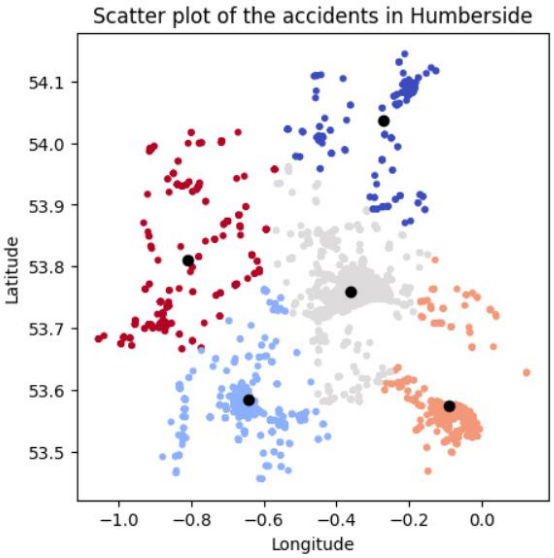


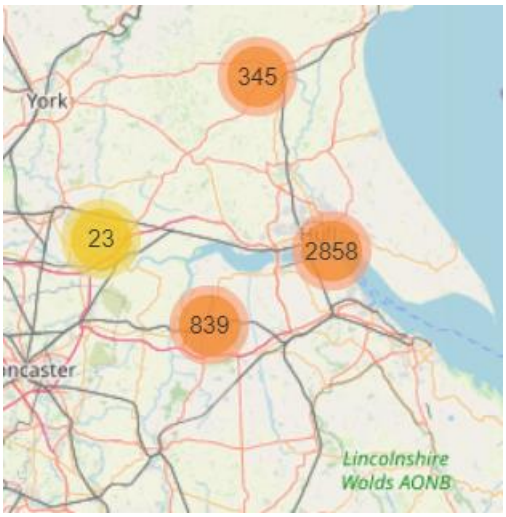
Hull was responsible for 0.61 percent of all accidents reported in the United Kingdom, with 52.2 percent happening between Hull Paragon Interchange and Anlaby Common, given that the Paragon Interchange is a transport interchange providing rail, bus, and coach services in Hull's City Center. A more detailed breakdown of the accidents in Kingston upon Hull can be found below. The most accidents occurred in Anlaby Common, Hull Paragon Interchange, and Sutton Fields, with 305, 204, and 189, respectively. Nineteen accidents in Hull occurred at 16:30, and 18.7 percent of the total accidents were on a Wednesday.



A MAP OF ACCIDENTS IN KINGSTON UPON HULL

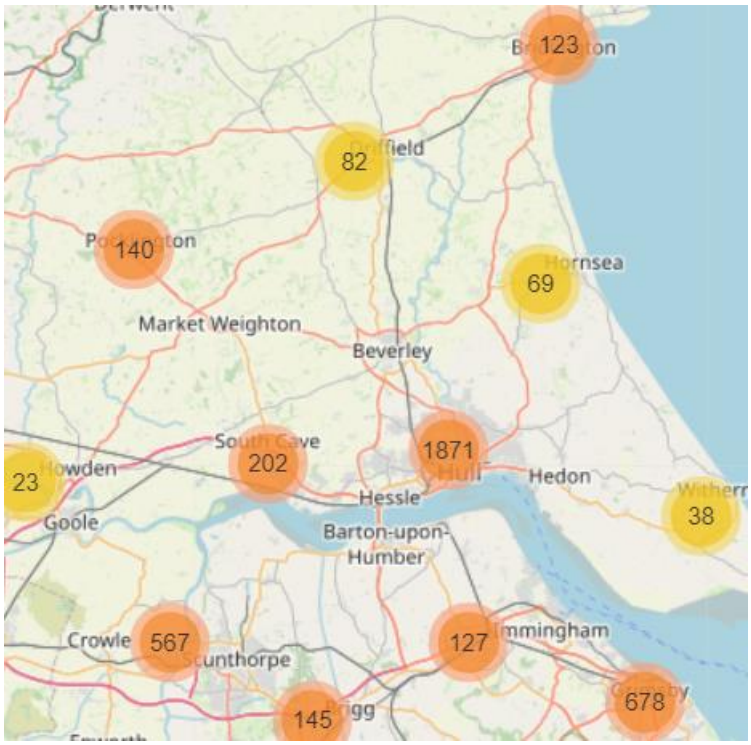
Humberside is a region in northern England that includes the areas surrounding the Humber estuary. It is the location of towns such as Goole, Beverley, Hull, Grimsby, Scunthorpe, Cleethorpes, and Bridlington. The clusters below show the number of accidents in the region.





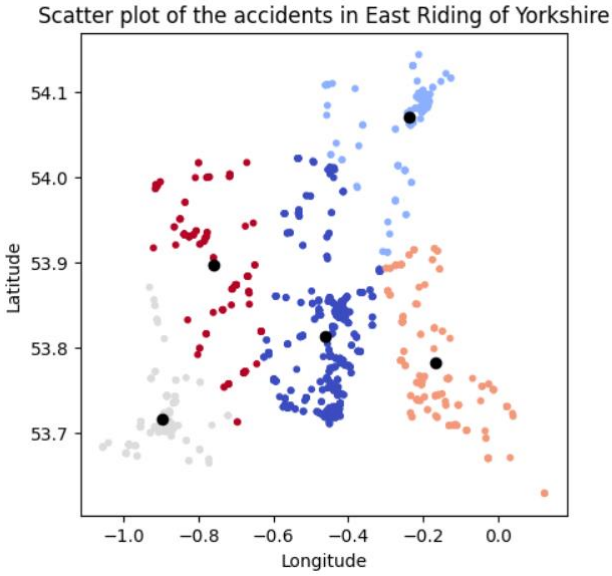
Due to the city's enormous population, Hull accounted for 32.87 percent of all accidents in Humberside. Hull, its environs, and Scunthorpe had the most accidents, with 2858 and 839, respectively. As seen in the map below, Hull, Grimsby, and Scunthorpe account for 76.7 percent of all accidents in Humberside, whereas Howden, Withernsea, and Hornsea had the fewest.

On average, 580 accidents occur on Humberside daily, with 45 accidents occurring at 16:30 hours and 655 on Fridays. As 74 of the 4065 recorded accidents were fatal, preventing these accidents through road safety recommendations is critical.

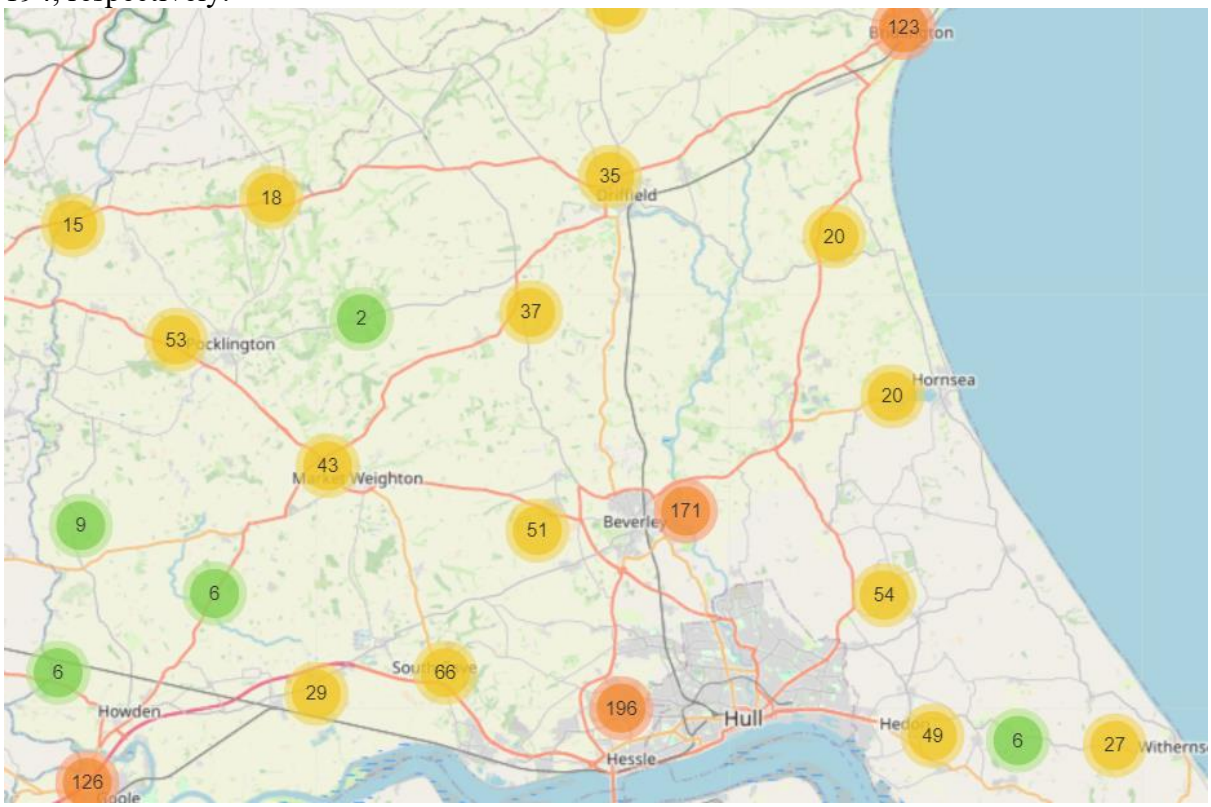


A MAP OF ACCIDENTS IN HUMBERSIDE

According to Wikipedia (2023), The East Riding of Yorkshire is a ceremonial county in England's Yorkshire and Humber region. The city of Kingston upon Hull is the most populous. The clusters below show the number of accidents in the region.

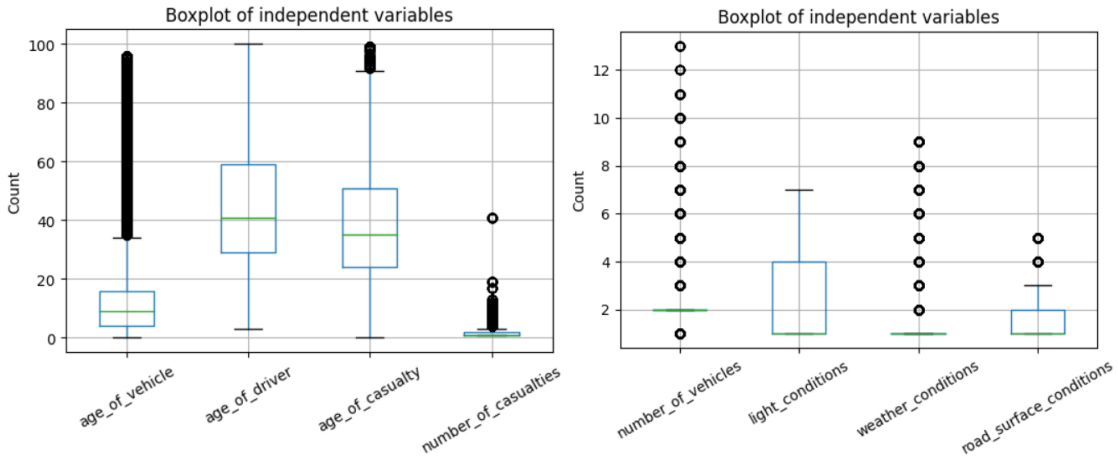


East Riding has an average of 171 accidents per day, with Hessle, Beverley, Bridlington, and Goole accounting for 51.3% of all accidents. Hull, the county's largest settlement with a population of 267,014, had no accident. East Riding had 1201 accidents, 25 of which were fatal. The most accidents in East Riding occurred at 11:30 and on Saturdays, totaling 26 and 194, respectively.



A MAP OF ACCIDENTS IN EAST RIDING OF YORKSHIRE

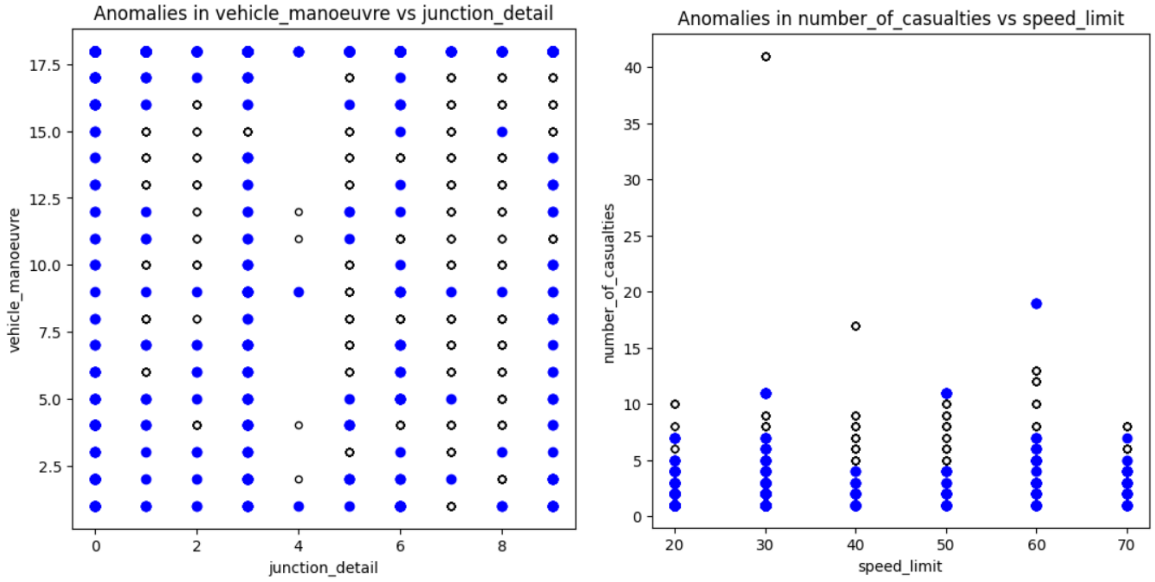
According to nist.co.uk (2023), it is vital to characterise standard observations before distinguishing abnormal ones. In other words, this approach leaves it up to the analyst to determine what is abnormal. As a result, an outlier is an observation that deviates from different values in a random population sampling. Outliers can be graphically represented by a box or scatter plot, as shown below.



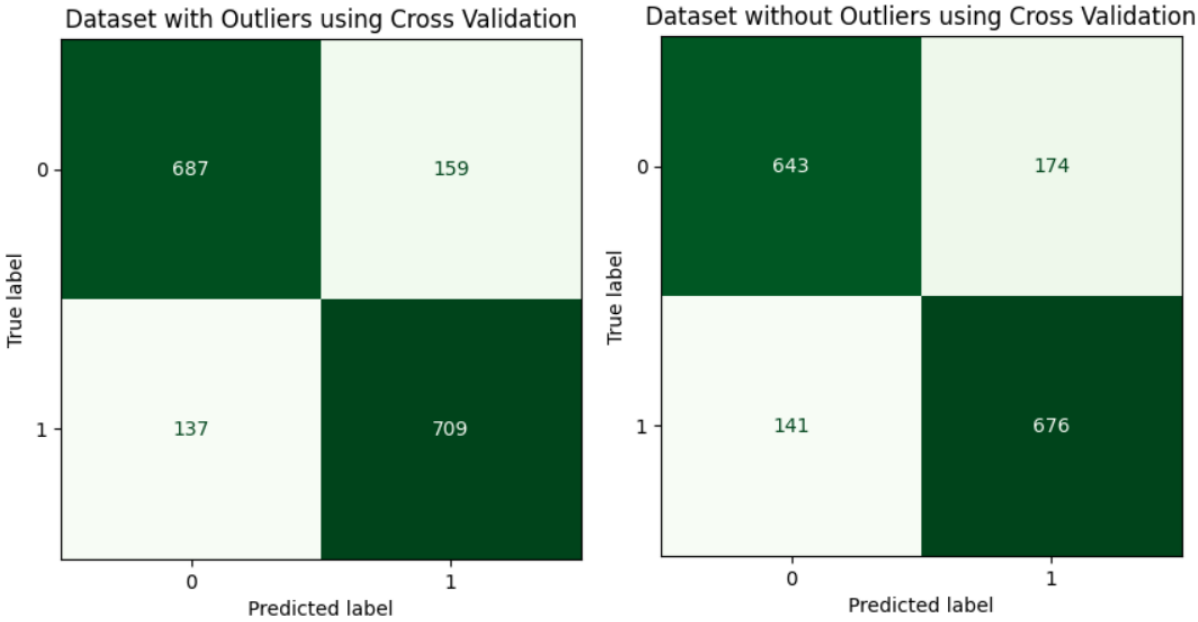
The above charts depict the distribution of few variables in the data. A local outlier in the number of casualties may exist, although this is also possible given that no one can truly

estimate the number of casualties during an accident. Aside from the suspicious age of vehicle, number of casualties, the interpretation of the box plots, reveals other outliers. Given the domain, labeling the above as outliers may be incorrect; hence, additional study and analysis will be performed to justify whether to keep, trim or remove them from the dataset.

Isolation Forest is employed because it can handle high-dimensional datasets successfully, isolates anomalies, and is very efficient and scalable to massive datasets.



Two approaches were used outliers should be kept in the data: dataset with and without outliers using cross-validation. The accuracy and performance of the different models were compared, and their confusion matrix was plotted to gain better insight.



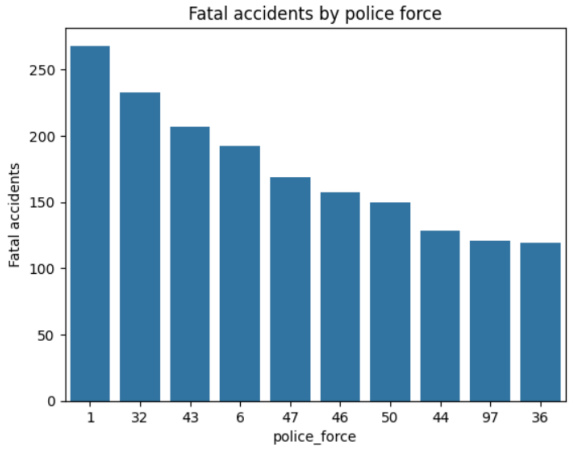
From the results for the dataset with outliers, 687 and 709 of 1692 accidents in 2020 were predicted to be non-fatal and fatal, respectively. The chances of a fatal accident occurring were 41.9%. Additionally, 137 and 159 accidents were incorrectly classified as fatal and non-fatal.

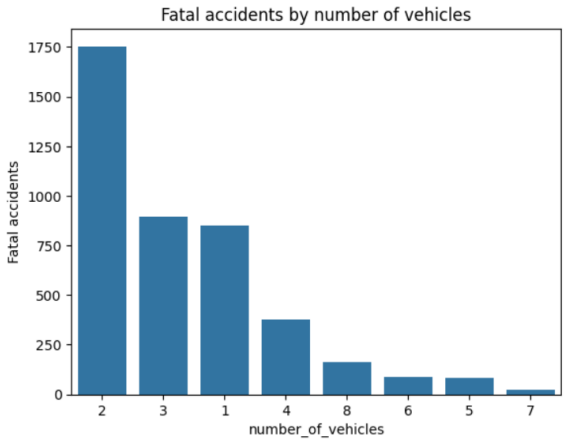
For every 1634 accidents in the dataset without outliers, 643 and 676 were predicted to be non-fatal and fatal, respectively. A fatal accident has a 41.4 percent chance of occurring. In addition, 141 and 174 accidents were incorrectly classified as fatal and non-fatal, respectively.

Furthermore, Random Forest Classifier outperformed other models before and after stacking, with 82 and 81% accuracies. In summary, maintaining outliers in the dataset has a 0.5% effect on the final prediction and, given the domain in which we are attempting to make predictions, they should be kept. Also, cleaning the dataset had a significant impact on the outcome. Removing data samples with -1 dropped the accuracy from 80 to 54%.

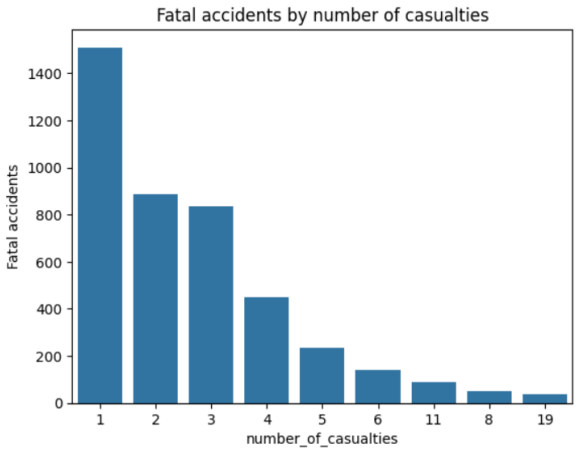
According to the classification model, 227 accidents out of 564 are fatal, implying that 40.9% of accidents were fatal. In 2020, around 92,369 fatal accidents occurred from all conceivable accidents.

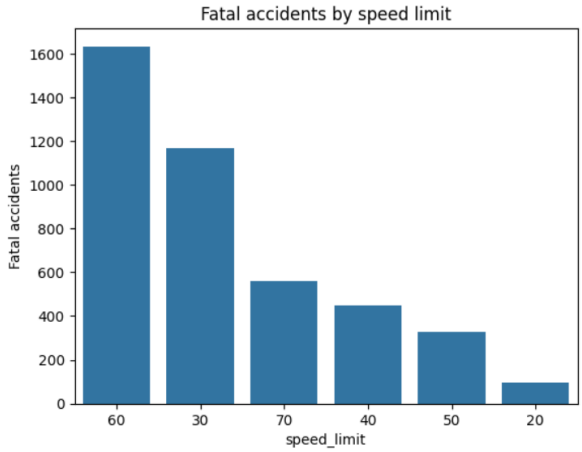
The number of vehicles, speed, vehicle manoeuvre, junction detail, first road class, area, driver sex, and casualties' number all have a si gnificant role in predicting the severity of an accident. The effects of the following can be seen in the plots below.



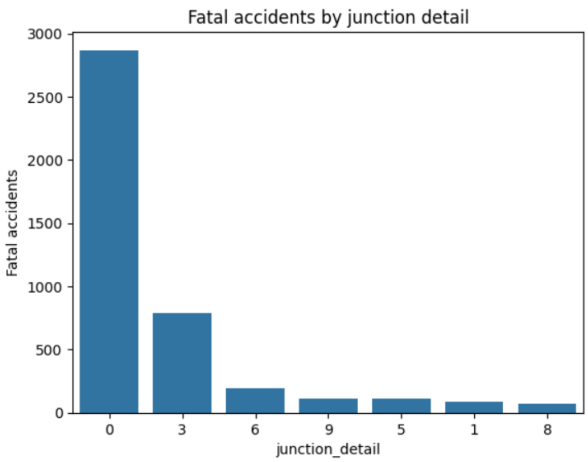


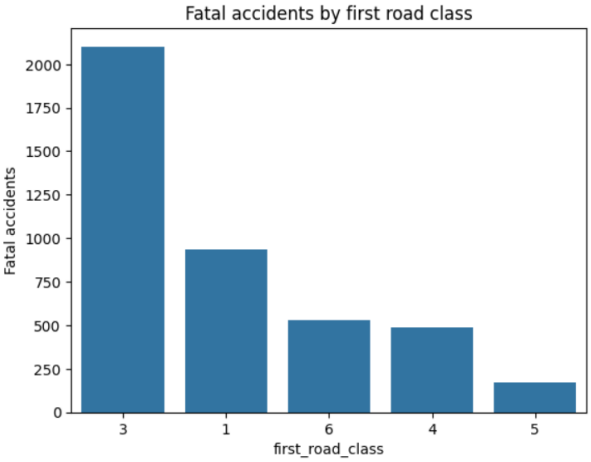
With at least 88 fatal accidents involving two vehicles, Metropolitan Police had the most significant number of fatal accidents. Two vehicles were involved in 1754 of the 4231 fatal accidents.



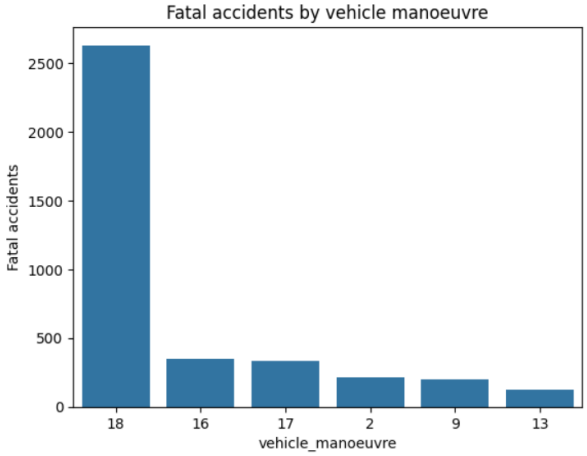


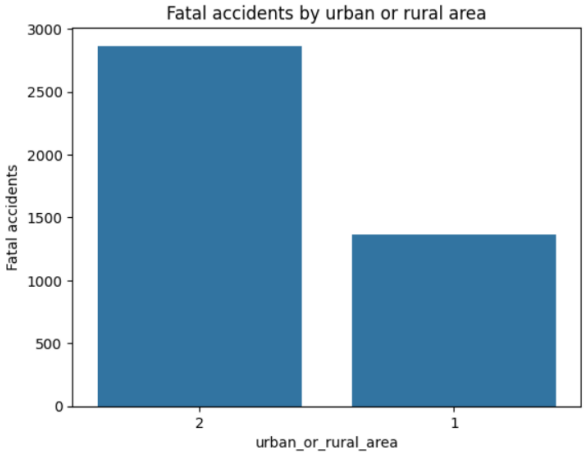
Most fatal accidents happened at 60mph, with only a casualty.



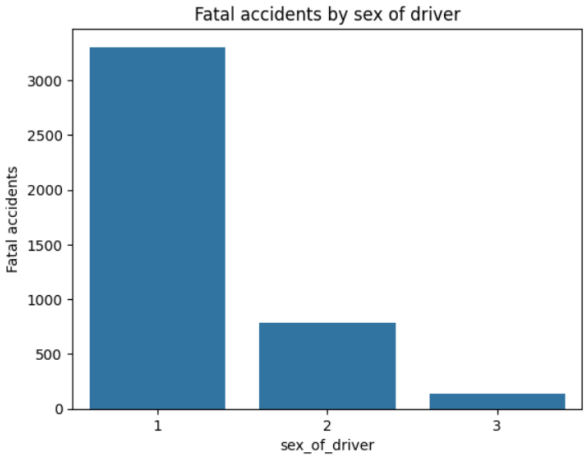


67.81% of fatal accidents occurred outside or within 20 metres of a junction, with 1397 occurring on A-class road. There were 2869 fatal accidents on A-class roads. According to Department for Transportation (2020), A-class roads are major roads intended to provide large-scale transport links within or between areas.





A driver going ahead of others caused 2629 fatal accidents, with 1746 occurring in rural area. In rural areas, 2867 accidents are fatal, accounting for 67.8% of all fatal accidents.



RECOMMENDATIONS

Following a detailed analysis of fatal accidents in the United Kingdom, implementing these recommendations will improve safety.

- Most people in fatal accidents drove at 60 mph, compared to 30 mph for non-fatal incidents. Hence, implementing a speed limit law of 40-50 mph will improve safety.
- Individuals aged 80 and over must not be permitted to drive, whether they have a medical condition or not. As shown above, most fatal incidents were by people over 80.
- More pedestrian crossing physical facilities within 50 metres should be created, as 90.17 percent of fatal accidents resulted from the absence of one, with Cornwall prioritized.
- Place a strict ban on drivers' going ahead of others, particularly in rural areas. 1746 of 2629 accounted for the most fatal accidents in this area.
- Place road signs/traffic signals on A-class roads not at or within 20 metres of a junction in rural areas. 1032 of 1219 fatal accidents were because of no road signs, and rural regions have narrow roads, with blind turns and brows, and few safe passing locations.

REFERENCES

Department for Transport (2020). *Guidance on road classification and the primary route network*. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/guidance-on-road-classification-and-the-primary-route-network/guidance-on-road-classification-and-the-primary-route-network [Accessed 10 Aug. 2023].

Department for Transport (2022). *Reported road casualties in Great Britain: notes, definitions, symbols, and conventions*. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-notes-and-definitions/reported-road-casualties-in-great-britain-notes-definitions-symbols-and-conventions#background-notes [Accessed 3 Aug. 2023].

Doogal.co.uk. (2021). *Postcode downloads*. [online] Available at: https://www.doogal.co.uk/PostcodeDownloads [Accessed 3 Aug. 2023].

Nist.gov. (2023). *7.1.6. What are outliers in the data?* [online] Available at: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm [Accessed 8 Aug. 2023].

Saedsayad.com. (2023). *Association Rules*. [online] Available at: https://www.saedsayad.com/association_rules.htm [Accessed 5 Aug. 2023].

STATS 20 Department for Transport Instructions for the Completion of Road Accident Reports from non-CRASH Sources. (2011). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995423/stats20-2011.pdf.

Wikipedia Contributors (2023). *East Riding of Yorkshire*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/East_Riding_of_Yorkshire [Accessed 5 Aug. 2023].