

Exploratory Data Analysis on Accidents-Casualty-Vehicle

**Subrata Das**

**P2681987**

**P2681987@my365.dmu.ac.uk**

Date: 30-03-2022

Contents

[Abstract 3](#_Toc99619484)

[Introduction 4](#_Toc99619485)

[Anatomy of data 4](#_Toc99619486)

[Focus Area 4](#_Toc99619487)

[Data scope, Range and Accuracy 5](#_Toc99619488)

[Data Format 5](#_Toc99619489)

[Primary data analysis 5](#_Toc99619490)

[First Step 5](#_Toc99619491)

[Data Exploration 6](#_Toc99619492)

[Data Pre-processing 8](#_Toc99619493)

[Joining the dataframes 8](#_Toc99619494)

[Initialize the joined\_allDF dataframe 8](#_Toc99619495)

[Data Cleaning 8](#_Toc99619496)

[Cleaning of ambiguous columns 9](#_Toc99619497)

[Cleaning unwanted columns 9](#_Toc99619498)

[Checking the null values in all the columns of joined dataframe 10](#_Toc99619499)

[Visualization & Analytics 12](#_Toc99619500)

[Accidents percentage in last 5 year 12](#_Toc99619501)

[Accidents count by gender over 5 years 13](#_Toc99619502)

[Accidents count group by the speed limit and casualty severity over 5 years 13](#_Toc99619503)

[Accidents count group by the speed limit and accident severity over 5 years 14](#_Toc99619504)

[Accidents count by light conditions over 5 years 14](#_Toc99619505)

[Weather conditions vs Accident count group by accident severity 15](#_Toc99619506)

[Accidents by the day 16](#_Toc99619507)

[Accidents count by road type 17](#_Toc99619508)

[Accidents based on urban or rural area 18](#_Toc99619509)

[Accidents by casualty class with age band of casualty 19](#_Toc99619510)

[Age band of the driver vs accident count group by accident severity 19](#_Toc99619511)

[Engine capacity histogram 20](#_Toc99619512)

[Purpose of journey 21](#_Toc99619513)

[1st point of impact vs accident count 21](#_Toc99619514)

[Density map of accidents 22](#_Toc99619515)

[Insights Summary 23](#_Toc99619516)

[Discovery 23](#_Toc99619517)

[Future research 24](#_Toc99619518)

[Conclusion 25](#_Toc99619519)

[Appendix A 25](#_Toc99619520)

[Configuration 25](#_Toc99619521)

[Codes 26](#_Toc99619522)

[Convert numerical column values into meaningful data as per metadata 26](#_Toc99619523)

[Matplotlib graph plot code sample 28](#_Toc99619524)

[Pixiedust table generation using dataframe 30](#_Toc99619525)

[References 31](#_Toc99619526)

Table of figures

[Figure 1: Accident percentage in last 5 years 12](#_Toc99619558)

[Figure 2 : Accidents count by gender over 5 years 13](#_Toc99619559)

[Figure 3: Accidents count group by the speed limit and casualty severity over 5 years 14](#_Toc99619560)

[Figure 4: Accidents count group by the speed limit and accident severity over 5 years 14](#_Toc99619561)

[Figure 5: Accidents count by light conditions over 5 years 15](#_Toc99619562)

[Figure 6: Weather conditions vs Accident count group by accident severity 16](#_Toc99619563)

[Figure 7: Accidents by the day 17](#_Toc99619564)

[Figure 8: Accidents count by road type 17](#_Toc99619565)

[Figure 9: Accidents based on urban or rural area 18](#_Toc99619566)

[Figure 10: Accidents by casualty class with age band of casualty 19](#_Toc99619567)

[Figure 11: Age band of the driver vs accident count group by accident severity 20](#_Toc99619568)

[Figure 12: Engine capacity histogram 20](#_Toc99619569)

[Figure 13: Purpose of journey 21](#_Toc99619570)

[Figure 14: 1st point of impact vs accident count 22](#_Toc99619571)

[Figure 15: Density map of accidents 22](#_Toc99619572)

[Figure 16: Discovery of highway accidents 24](#_Toc99619573)

# 

# Abstract

In this report, we are working with the UK Government data on accidents, casualties, and vehicle data published by the Department for Transport (UK gov department of transport, n.d.). The chosen data contains the last 5 years of records and the data is fairly accurate in terms of availability and consistency. After collecting the data, initially, it is analyzed for understanding, pre-processing, and cleansing. The cleaned data is used to create further visualizations and Exploratory Data Analysis (EDA). At the end of this report, a discussion of insights found in the data is discussed. Using Apache spark the whole datasets were managed swiftly and the underlying Hadoop storage is well capable of handling the Big data. There are 1.6 million records present in the combined dataset and by the analyses we ran through some insights obtained in this process.

# Introduction

## Anatomy of data

The data is publicly distributed by the Dept of Transportation Accidents, Casualties and Vehicles of UK on their website <https://data.gov.uk/dataset/road-accidents-safety-data>. There are various datasets present to explore and work with to extract meaningful insights but I have chosen 3 datasets which are below:

1. Accidents (597973 records)
2. Vehicle (1101591 records)
3. Casualty (781716 records)

These 3 datasets are related to each other and can be joined by the primary key present. The metadata is large and cannot be accommodated in this report but some of the important fields should be discussed. This data is for the whole UK but users can customize the data as per their needs also, such as region wise, city wise, time-specific, etc. I have chosen the date of the last 5 years (2016 to 2020) and this is across the whole UK.

Files contains string or number(int/float) data. String data are categorical and numbers have continuous or discrete values present in the dataset. Geolocations are also present in all the datasets which will help us to plot the map for some insights we are looking for.

## Focus Area

The dataset is rich and can be utilized to extract an enormous amount of insights from it. But as an analyst, my primary focus area is the relation between different situations such as road conditions, weather conditions, lighting conditions, etc with the accident count, severity.

Secondarily, geolocations are also available in this dataset. Accident density and region-wise analysis are also other prospects to dig in.

In the end, I decided to check some parameters which will be discussed thoroughly in the visualization and analysis section.

## Data scope, Range and Accuracy

Data contains accidents, casualties and vehicle details related to accidents are in the data set. The time range is customized as per users’ needs which is from 2016 to 2020. This data is fairly accurate later in Appendix A we can see there are no missing values in our targeted dataset which is fantastic in terms of data accuracy.

Note: There is a chance that a few of the accidents/casualties and vehicle details might be registered by different primary IDs but comparing the volume of the data set, the errors can be overlooked.

## Data Format

The data is enclosed in Comma-separated values (CSV) files, which means all the values are separated by a comma and the file contains a header row. Rows contain records and columns contain the variables.

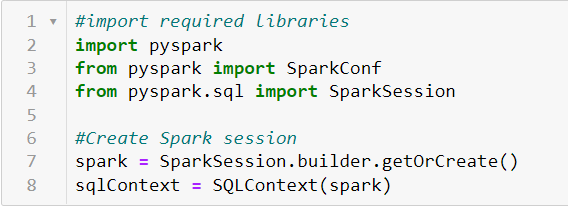
Three CSV files are joined together by using a primary key to create a combined dataframe to work with. In this way, all the entries in the three datasets are connected and no repeated entries are considered. The combined dataset has 39 columns but we will be using around 20 columns to gain the outcome. Most of the data is numeric but they represent some meaningful categorical values in real life. The proper transformation will be done in the data preparation stages to prepare the data for the outcome.

This dataset is vast and needs future attention for further exploration. Also, there are some discrepancies found in the data which for now can be ignored but should be suggested to UK Gov to rectify that.

## Primary data analysis

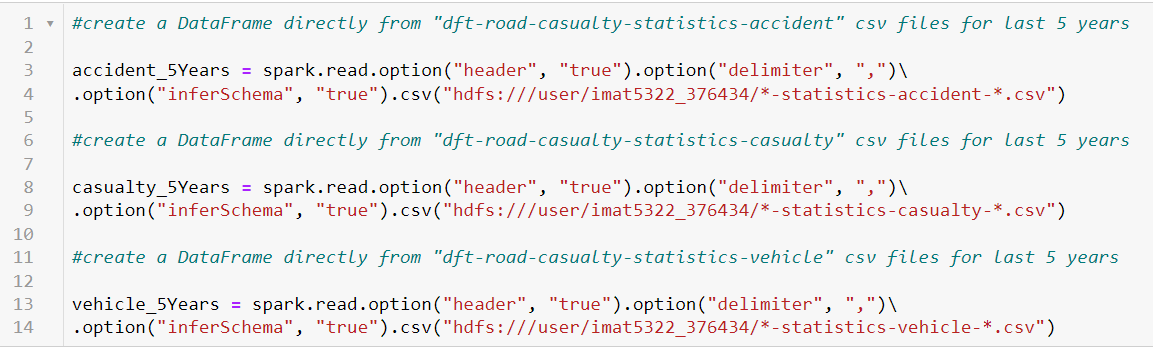
## First Step

At first, we create the Spark session to facilitate our entry to the spark (Apache Spark, n.d.) core with the below snippet:



Now there is 1 file for each dataset for each year. For five years of data, the number of total files is 15 (5\*3\*1=15). These files were individually downloaded and uploaded to spark local Linux storage using the Apache Spark console. Later it was uploaded to HDFS (hdfs:///user/imat5322\_376434/ ) file system to utilize the Hadoop storage benefits and faster processing. This step makes the files available for Apache Spark core to use the files and Hadoop storage facilities.

The files were in CSV format and therefore it was directly imported into data frames by using the Spark session and wildcard characters below:



Each statement combines 5 files into one file which contains 5 years of data for each dataset. Keeping the “header”,” true” we directly added all the data keeping one top row as header values.

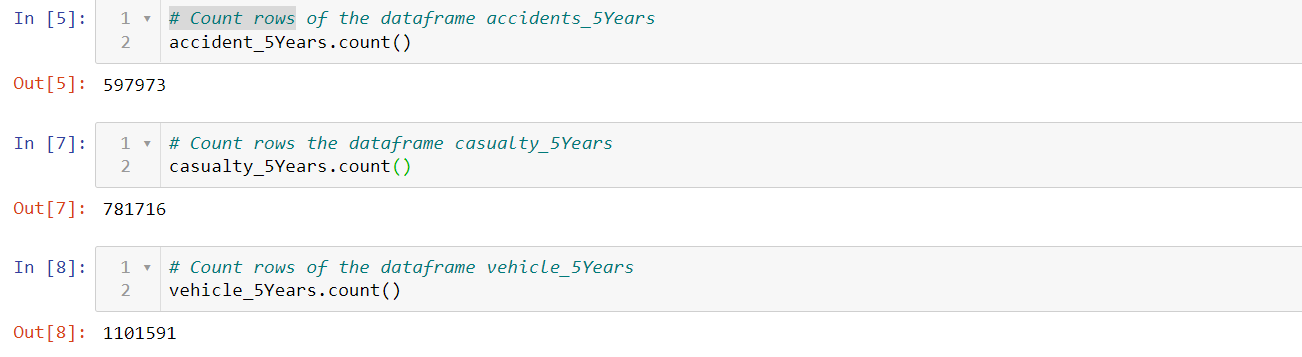
Also, the files were “,” delimited and automatically the schema was inferred using the arguments shown in the above snippet.

## Data Exploration

In the beginning, the checking of schema and variables should be done. As we can see schema has been inferred correctly otherwise explicit type casting was required. At this point, the number of variables is too high and it is not our concern. This will be dealt with later. This data can also be imported into Resilient Distributed Datasets (RDDs) but as our data is structured, we do not need to put extra effort into RDDs.



Checking the number of rows/records in each dataframe:



Initialize the dataframes:

The first step after importing the data into data frames should be to initialize the dataframes to check if there is an issue with importing. In this case, it is imported perfectly and 1st record in the dataframes is shown below:

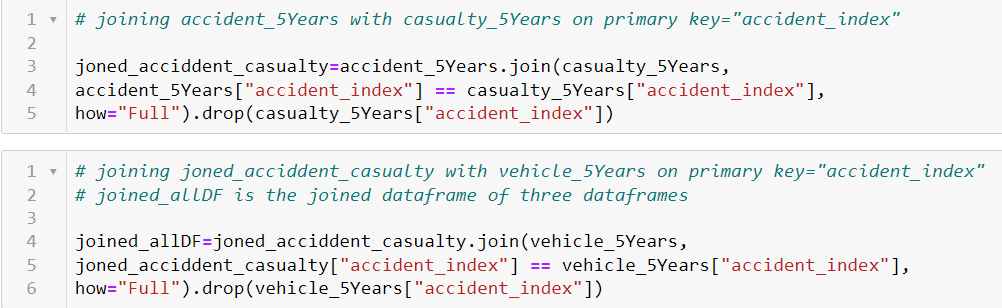


Now we move on to the data pre-processing step.

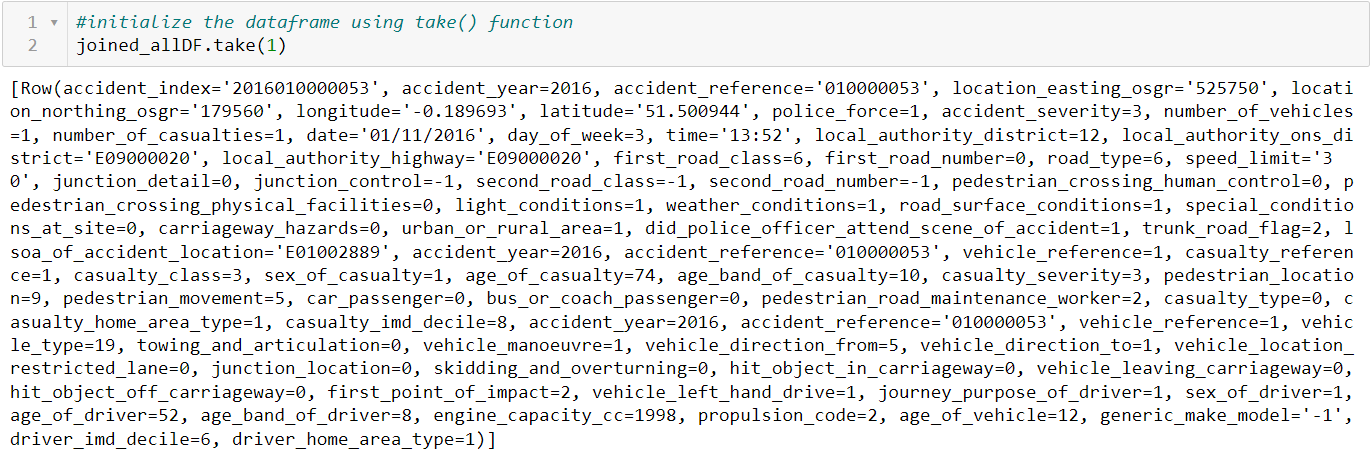
# Data Pre-processing

## Joining the dataframes

We have 3 combined dataframes and Accident\_index is the unique identifier mentioned in the metadata, so we join all three dataframes using the primary key=” Accident\_index”.



## Initialize the joined\_allDF dataframe

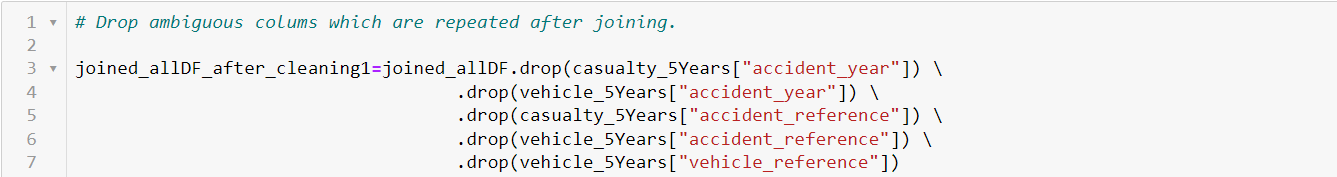


Still, we can see there are lots of columns associated with the joined dataframe. We will clean some of those in the next step.

## Data Cleaning

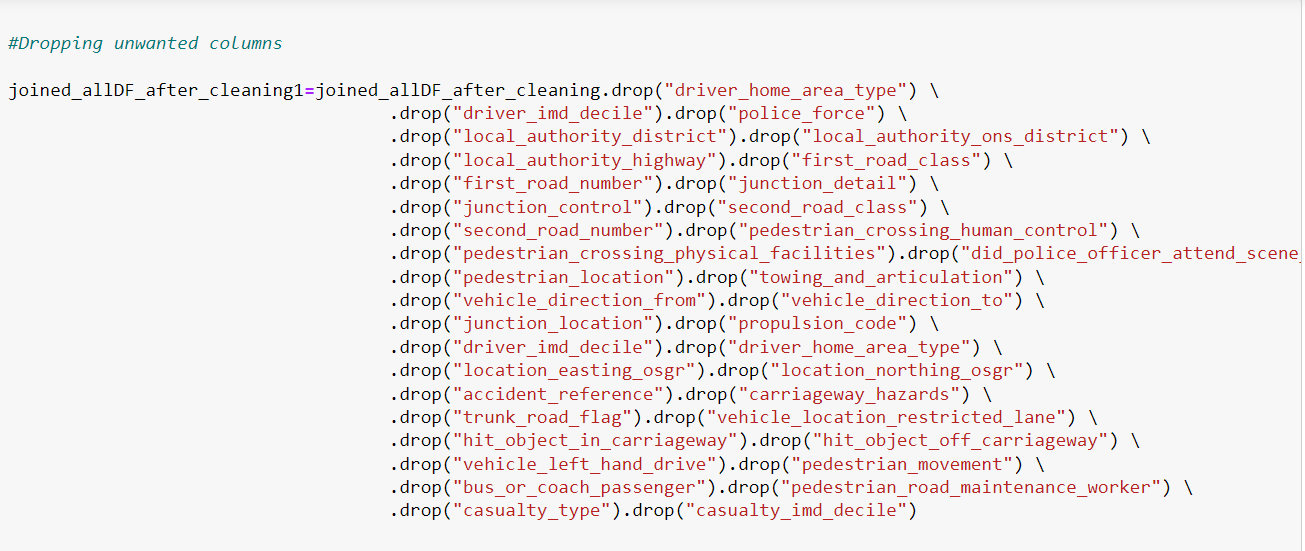
In the first step of data cleaning, we analyze the joined dataframe and identify the columns which are not required for our analysis. We can also see because of joining three dataframes there are some repeated columns present in the dataframe such as “accident\_year”, “accident\_reference” and “vehicle\_reference”.

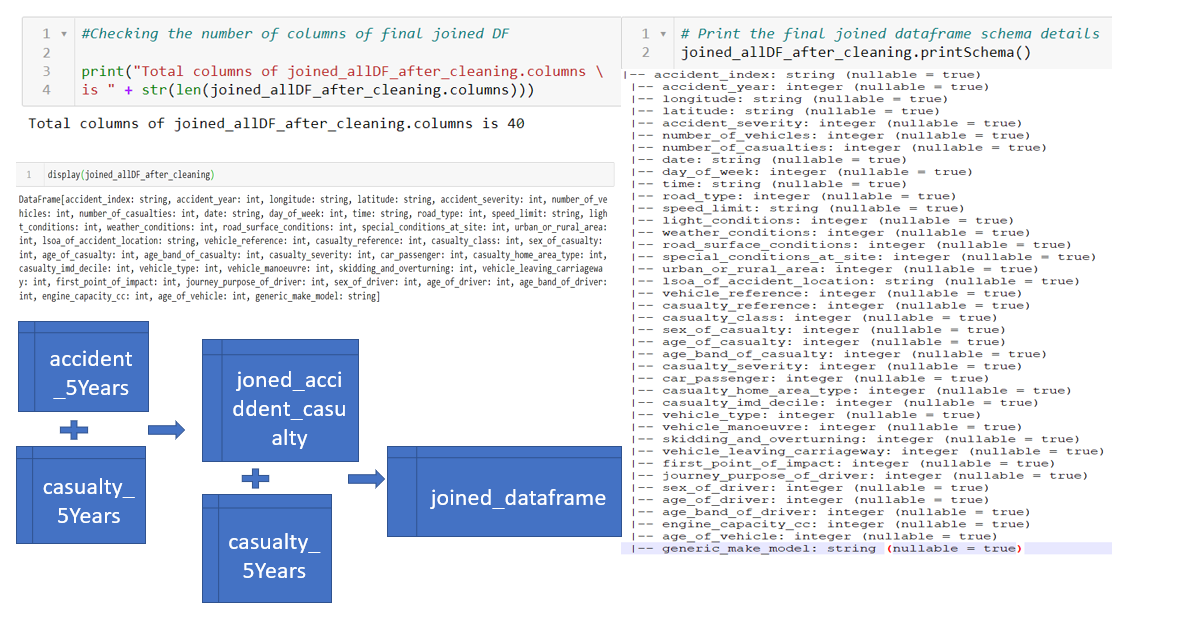
## Cleaning of ambiguous columns



## Cleaning unwanted columns

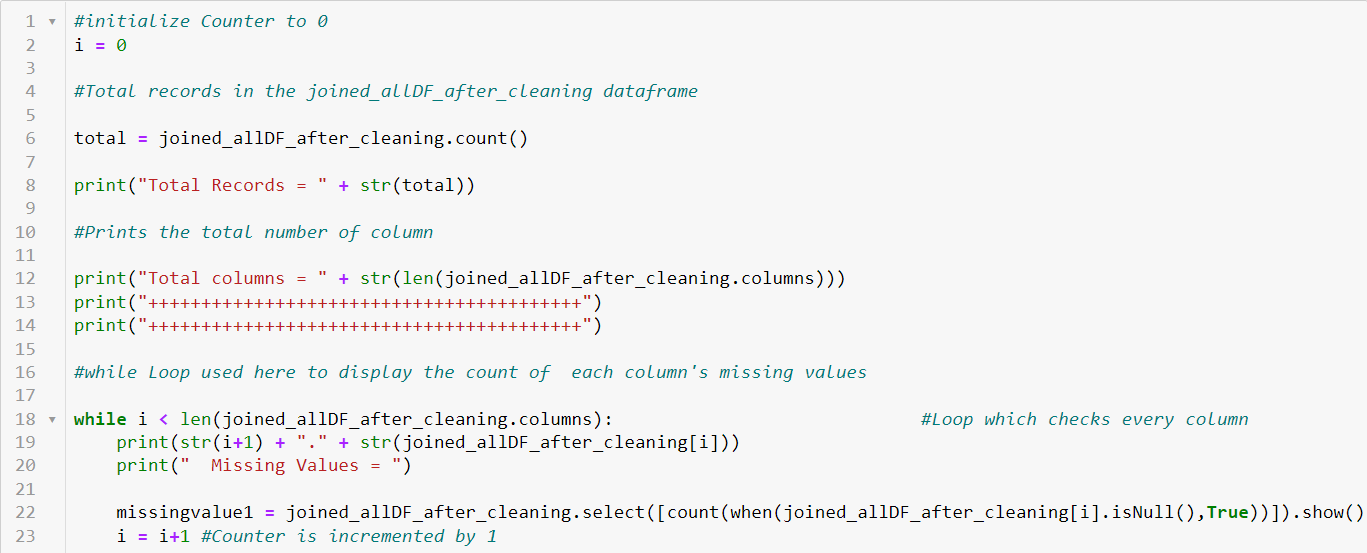
These columns can be useful for analysis but in this case, these are deleted because of keeping the goal of the analyst. These columns contain too many categories or are not required for analysis. For this reason, I am dropping these columns but these can help in future exploration of the dataset and be found handy.

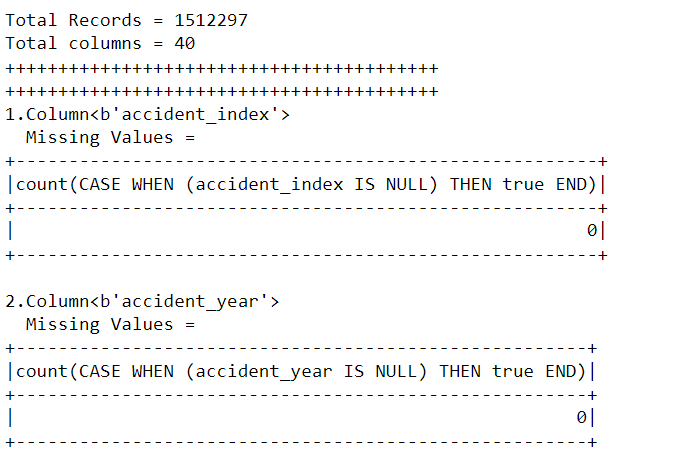


Finally, we have processed our final dataset which is shown below 

## Checking the null values in all the columns of joined dataframe

The below snippet is self-explanatory with proper comments. This piece of code is checking if there is any missing value in any of the records for each column. After successfully running the code there was no missing value found in this dataset which was mentioned earlier also. For the limited scope only a few columns output is shown below:





As there are no null values present in the dataset we are really good to go with the next part which is, replacing numerical values with the actual meaningful data from the metadata of the dataset which can be found here:

<https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>

This CSV file contains the data definition. Most of the column values in the dataset are categorical and they are represented with numbers that have their meaning defined in the above metadata. To understand and visualize with actual meaningful labels I converted the values into human-readable values.

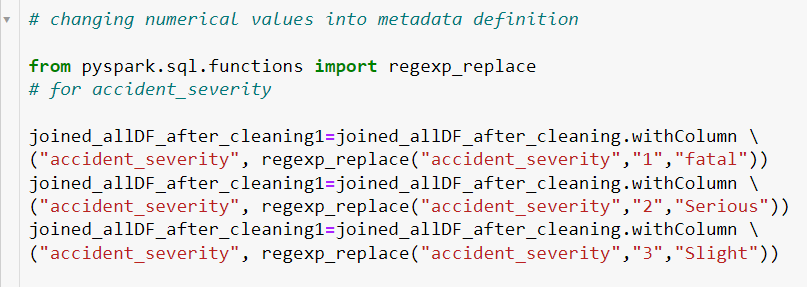
Below is one example where accident severity is categorized with 3 levels of measurements.

1 defines Fatal

2 defines Serious

3 defines Slight

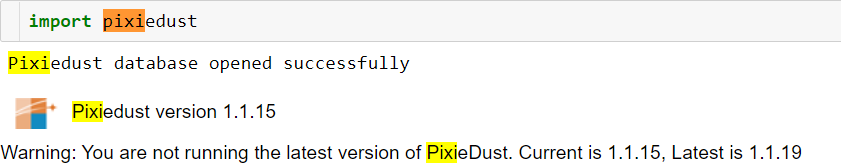
With the below code we can create a new RDD with these converted values into the respective columns:

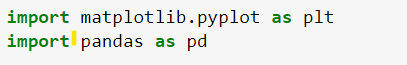


With the same approach, I have converted several other variables which can be found in the appendix in detail.

# Visualization & Analytics

We have come to a stage where we can start visualizing and find insights from this rich dataset. We have used here Pixiedust(Version 1.1.15) and Matplolib libraries to accomplish the visualization.





Note: In some of the visualizations Axis numerical values are mentioned as e^values.

## Accidents percentage in last 5 year

The very first visualization to start with is the accident percentage over the 5 years of data. The below pie chart clearly shows a significant dip in accidents count and we get a gradually decreasing trend.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2016 | 2017 | 2018 | 2019 | 2020 |
| 23% | 22% | 21% | 20% | 15% |



Figure 1: Accident percentage in last 5 years

We can exclude 2020 and the drastic dip in that year because this dip was caused by Covid-19 pandemic effects but apart from that, we can see there is on average a 1% dip every year which clearly shows the increase in traffic awareness.

## Accidents count by gender over 5 years

Below bar chart clearly shows that males are more prone to accidents in terms of all three severity categories. The possible reason for this is the total number of male drivers in the UK is much greater than female drivers. The ratio of all the different accident severity is proportional in all the genders.

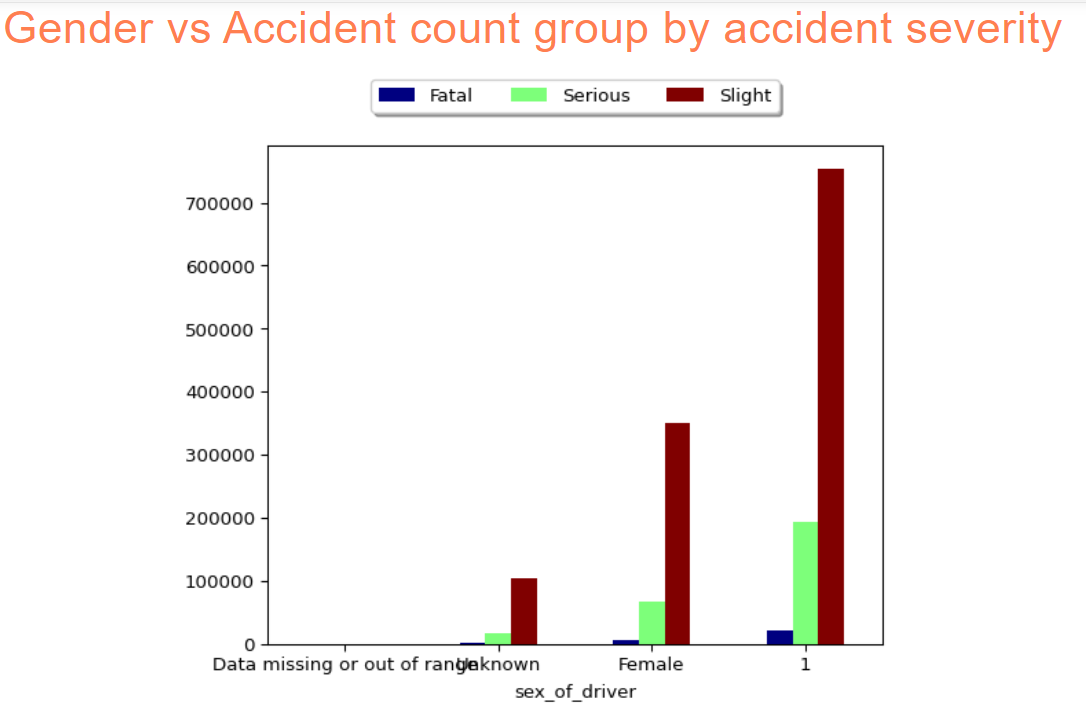


Figure 2 : Accidents count by gender over 5 years

## Accidents count group by the speed limit and casualty severity over 5 years

This clustered bar chart represents the accident counts grouped by casualty severity and speed limit. Interestingly pathways with 30 mph are most accident-prone whereas we can also see most of the fatal accidents also happened in the same speed limit zone. The possible reason could be less awareness due to low speed and most of the 30mph motorways are inside city limits. Heavy traffic can cause accidents with more numbers and low severity.



Figure 3: Accidents count group by the speed limit and casualty severity over 5 years

## Accidents count group by the speed limit and accident severity over 5 years

Similarly, this bar chart shows that the 30 Mph zone is more prone to accident severity in all three categories. As shown in the figure below it is safe for us to assume that 60 Mph is the second most accidental prone speed limit with fatalities being slighter when compared to 30 Mph.

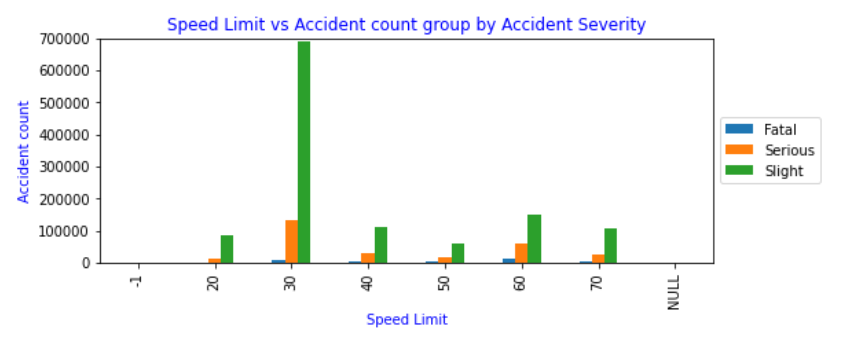


Figure 4: Accidents count group by the speed limit and accident severity over 5 years

## Accidents count by light conditions over 5 years

The bar chart above enables us to interpret that the greatest number of accidents that occurred over 5 years unfortunately happened while being in the best lighting conditions and one of the possible contributing factors to it might be a larger number of vehicles are on the road traveling when compared to night condition or worse lighting conditions.

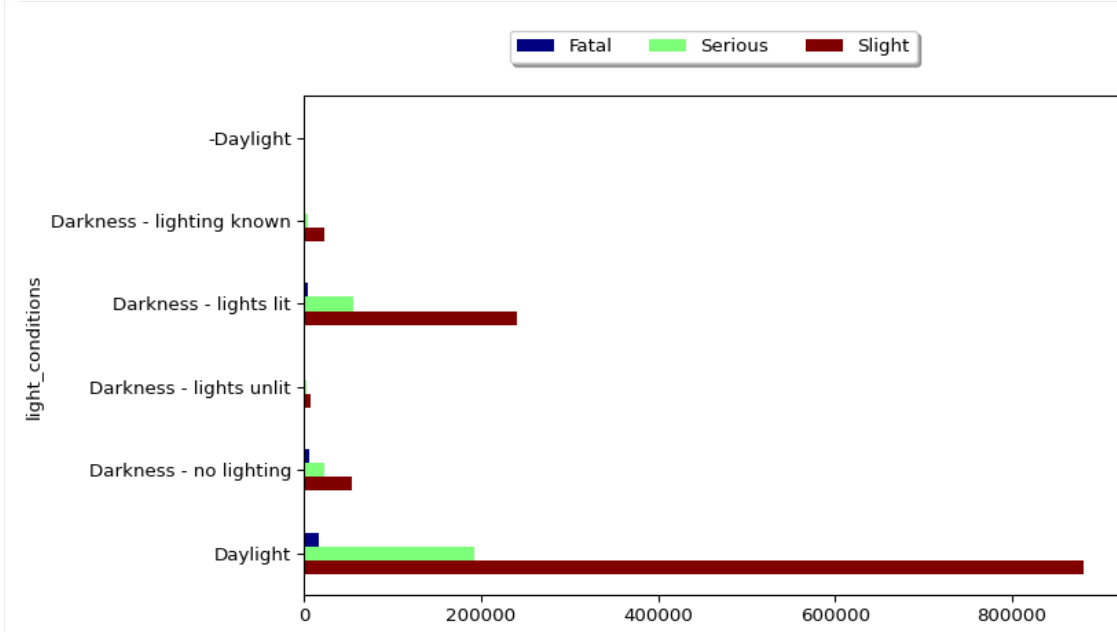


Figure 5: Accidents count by light conditions over 5 years

And similarly, as we can see the second most accidental prone area is darkness with the lights lit on the roads showing us around 600 thousand fewer accidents over 5 years when compared to accidents occurring in the daylight.

## Weather conditions vs Accident count group by accident severity

As we can see a strange anomaly happening in the accidents in the last 5 years when the factor that is taken under consideration is the weather. From the bar chart above it is safe for us to interpret that the highest amount of fender benders that occurred over 5 years, unfortunately, happened under fine weather conditions where the wind condition was better in comparison to high winds and worse weather conditions. One of the micro factors is people driving carelessly during the conditions being comfortable to them. And the second most accidents being in rain condition with no winds.

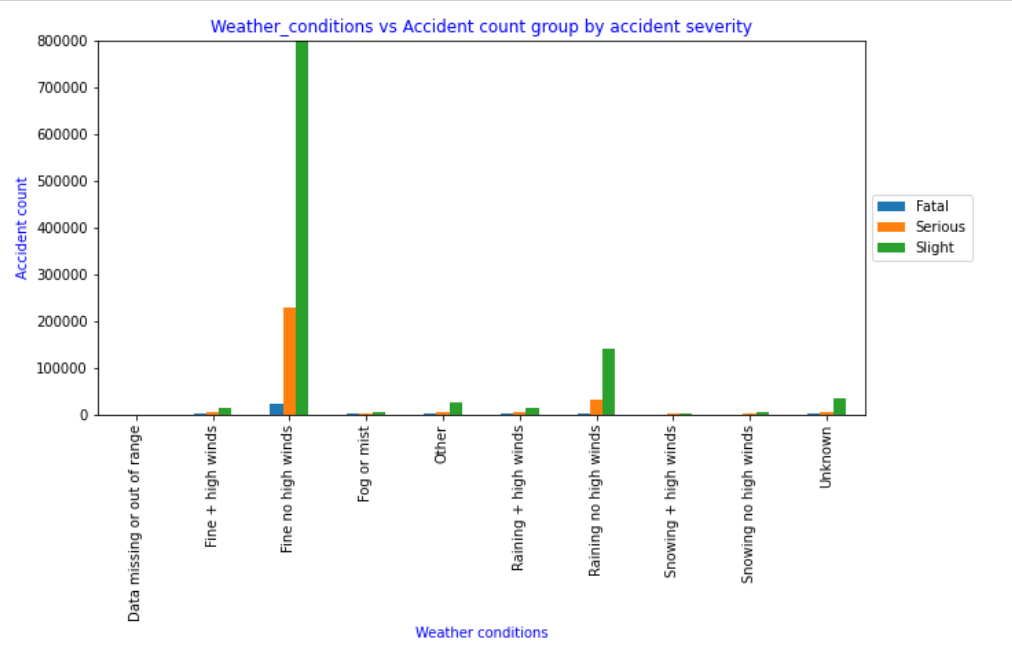


Figure 6: Weather conditions vs Accident count group by accident severity

## Accidents by the day

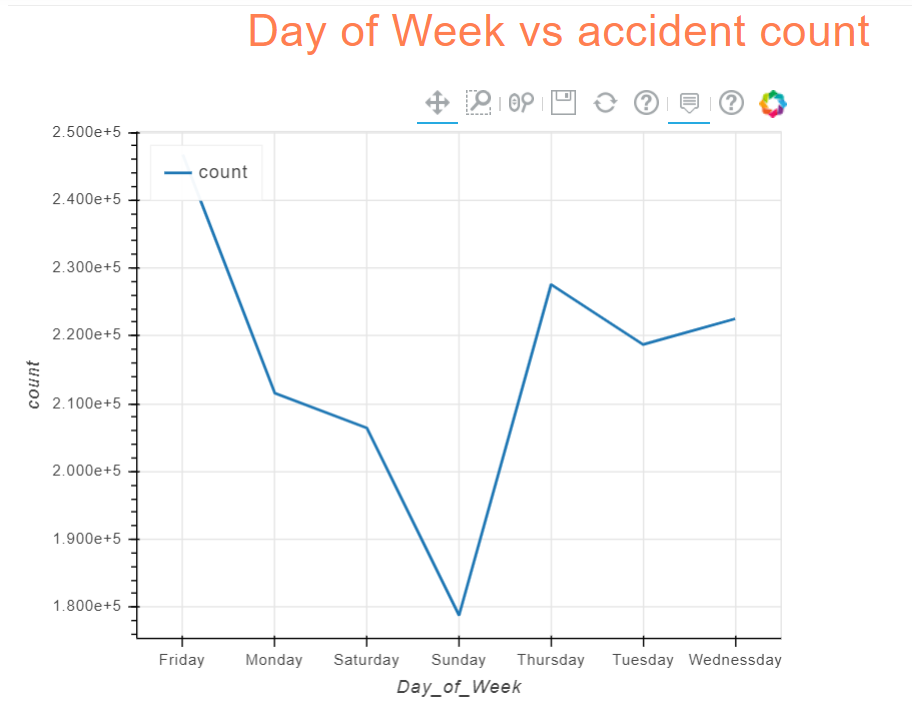
The line charts stated below shows us the uptrends and downtrends of the accident that are occurring on all days of the week. With the help of the figure below we can safely say that most amounts of the accidents that happened on a particular day over 5 years were on the weekends with weekends being one of the major factors contributing to the accidents. We can also see that the number of accidents drops exponentially with one of the reasons being people like to rest on Sundays after a tiring weekend. And strangely we can see an anomaly where the second the greatest number of accidents that are happening over 5 years is on Thursday.

Figure 7: Accidents by the day

## Accidents count by road type

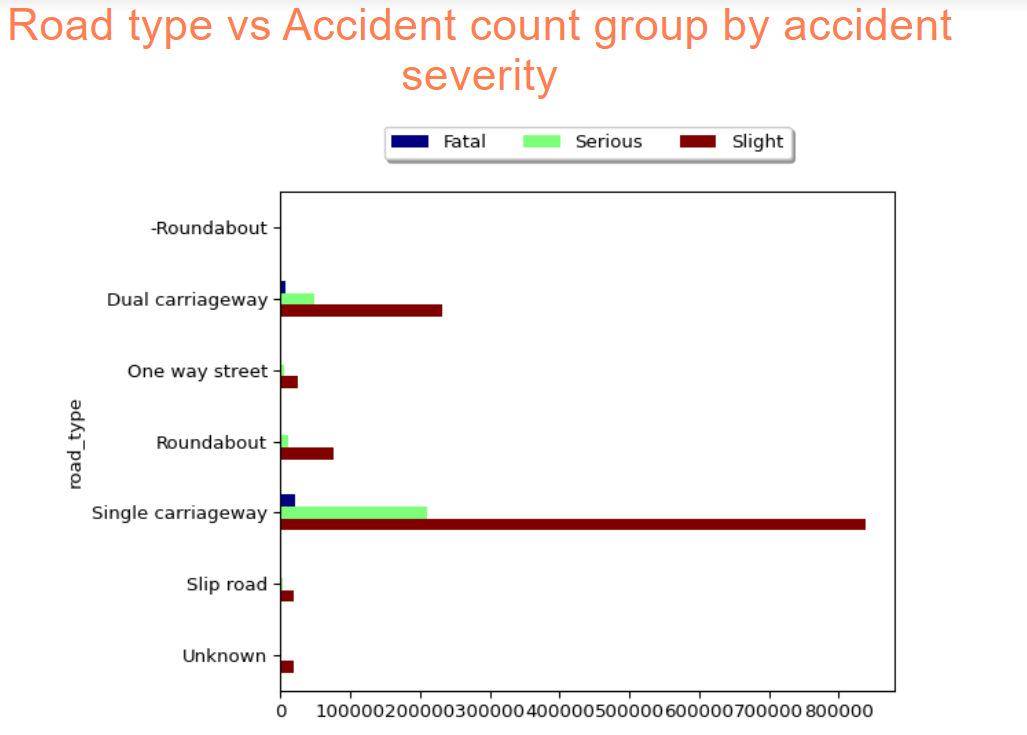


Figure 8: Accidents count by road type

## Accidents based on urban or rural area

This report shows the percentage of accidents in urban or rural areas and clearly, urban areas are performing worse than rural areas. The possible reason is heavy traffic and a greater number of cars and pedestrians present on the roads.

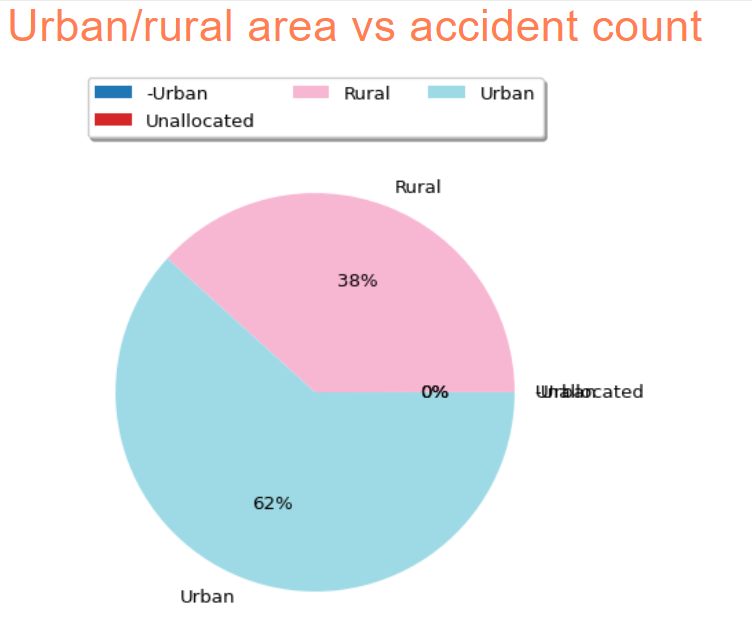


Figure 9: Accidents based on urban or rural area

## Accidents by casualty class with age band of casualty

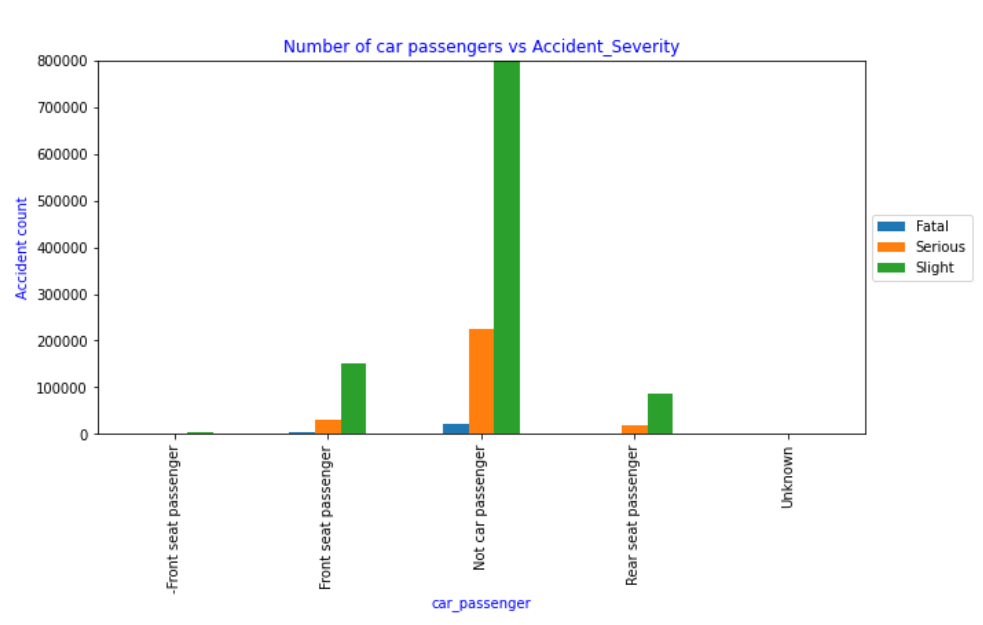


Figure 10: Accidents by casualty class with age band of casualty

## Age band of the driver vs accident count group by accident severity

Below data shows most of the accidents in all three accident severity categories are of age band 26-35 followed by 36-45. This is because the median of the distribution of the driver’s age is near these values and the age band is also 10 years compared to another age band present in the dataset.

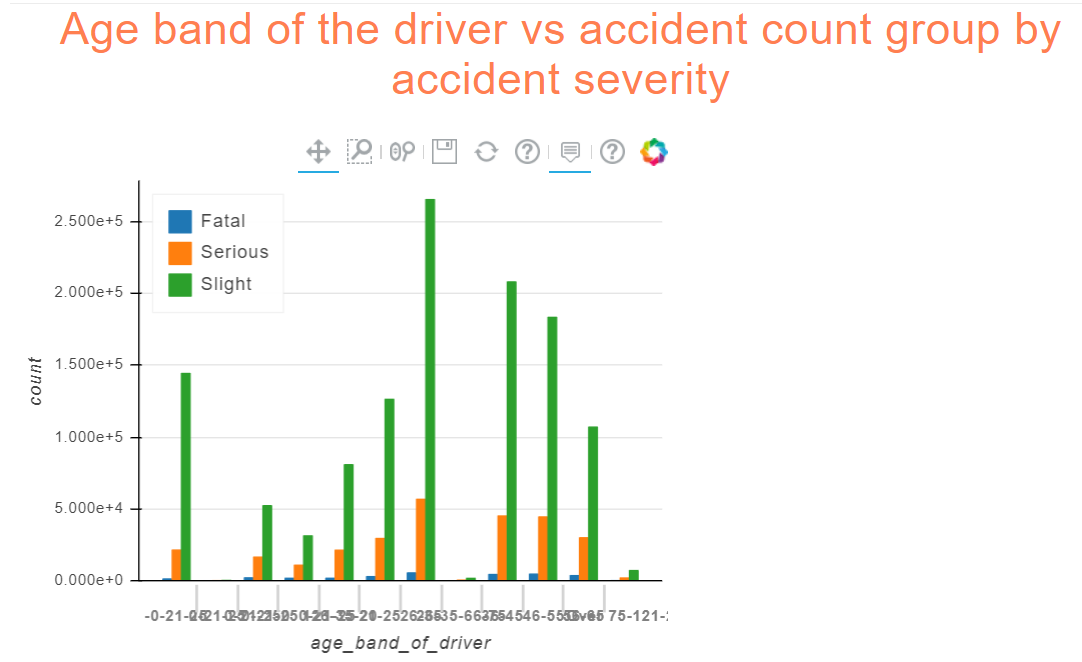


Figure 11: Age band of the driver vs accident count group by accident severity

## Engine capacity histogram

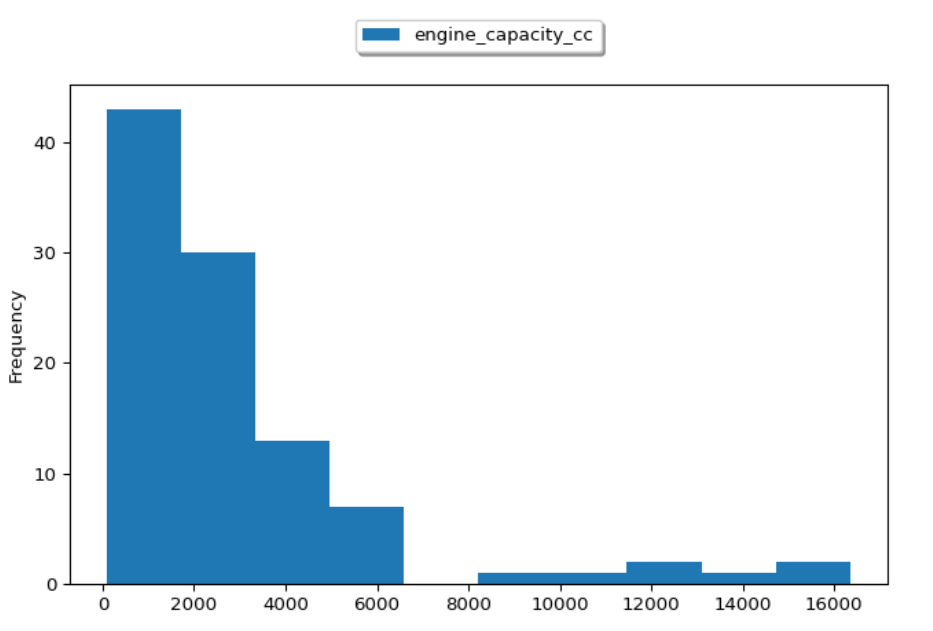
The engine capacity histogram shows the distribution of cars and it is clear that most cars are between 0-6000 cc. Apart from that in higher engine capacity 12000 and 16000 cc cars are more in the road that other categories.

Figure 12: Engine capacity histogram

## Purpose of journey

In this pie chart, the data insufficiency is depicted. 62% of the data is unknown which shows the lack of information collected. This can be suggested to UK Gov to rectify in future to increase the data quality. Apart from that Most of the collisions happened during driving at work, such as professional drivers, cabs, and other people who use cars in their profession which is normal in this case. Actual data is not clear here due to insufficient data.

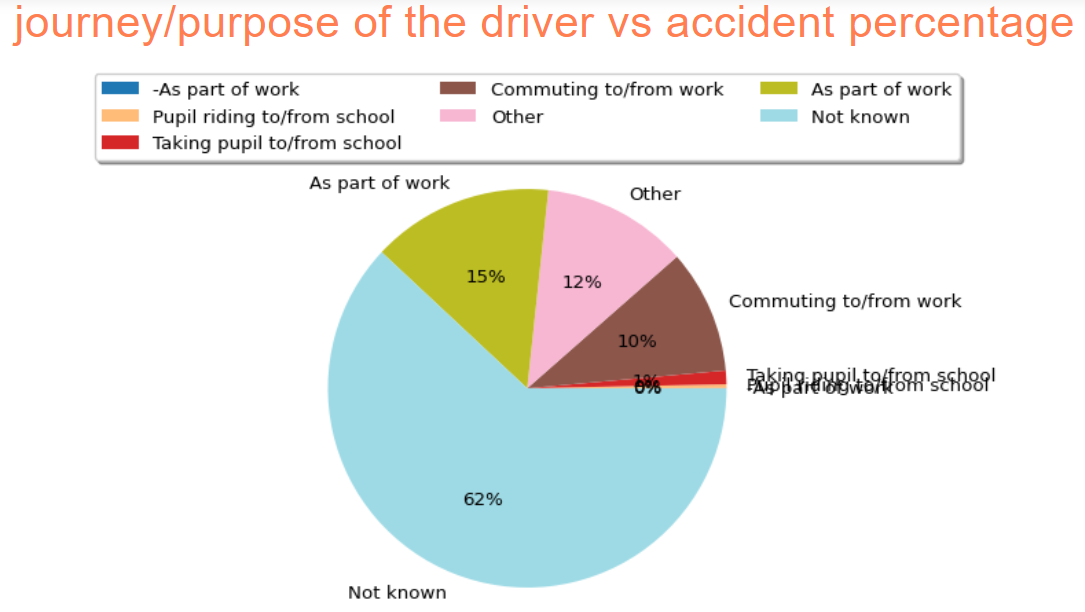


Figure 13: Purpose of journey

## 1st point of impact vs accident count

As below graph depicts that most of the impacts are from the front side of the car which implied a head-on collision. Most fatal accidents are also in the same category followed by backside and offside. Back and offside implies the blindside of the driver and it is quite natural in the distribution.

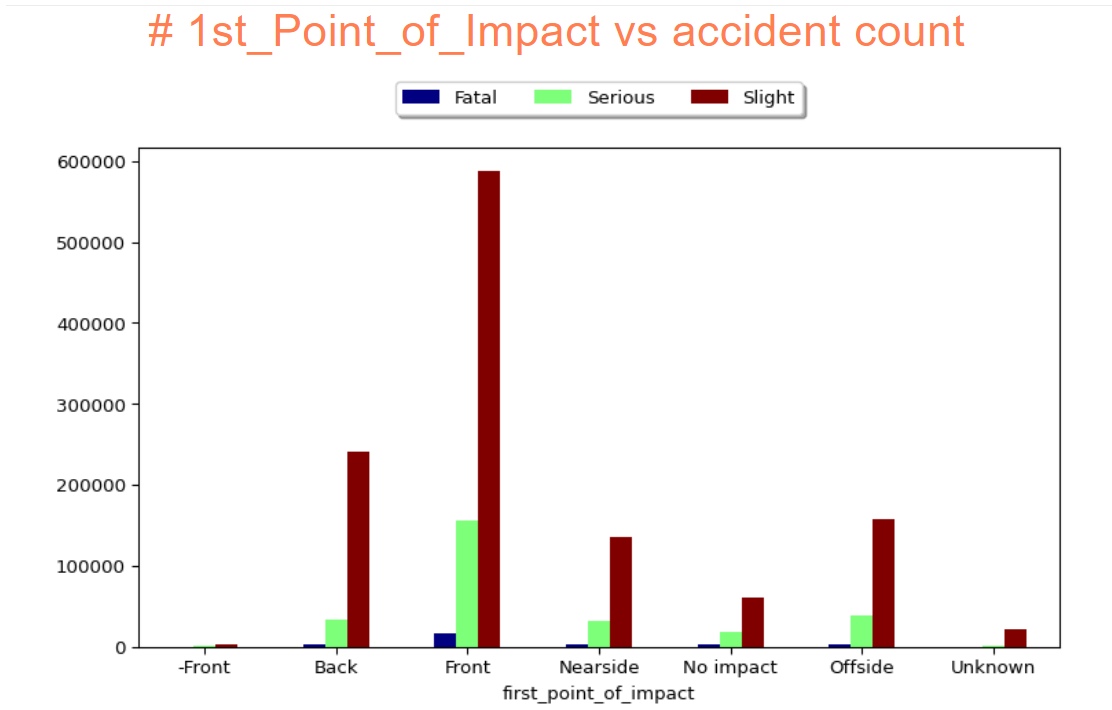


Figure 14: 1st point of impact vs accident count

## Density map of accidents

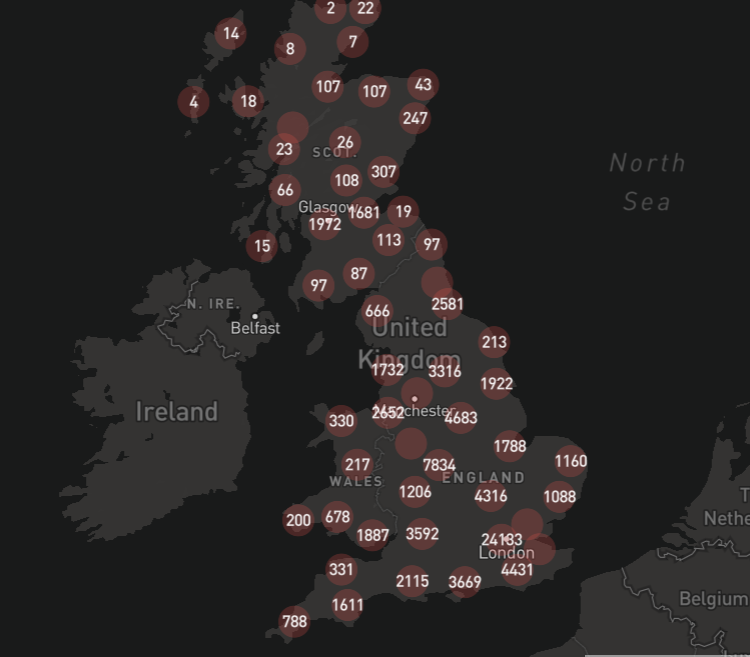


Figure 15: Density map of accidents

With the mapped plot we confidently say that the most amount of accidents that is happened in England over 5 years is in central London with a whopping 24 thousand accidents and then followed by east and west midlands like Birmingham and Leicester. We can see a general trend wherein the south region is more prone to accidents when compared to the north region like Scotland.

**Note: Due to low capacity of Pixiedust only 5,00,000 records are shown in the map. This represents a sample of the whole population.**

# Insights Summary

1. Accidents are gradually decreasing over the last 4 years by 1%. In 2020 it dipped by 5% because of the Covid-19 hit.
2. Most males are involved in accidents because the number of male drivers is more.
3. Most of the accidents are happening in clear weather and on a single carriageway. People are more reluctant during clear weather which is the reason for this. Also, a single carriageway allows people to reach higher speeds and overtaking tendencies.
4. Most accidents are happened because of head-on collisions and blindside collisions.
5. Journey of purpose data is not sufficient and it has 62% unknown reason.
6. Most cars are on 0-6000 cc engine capacity.
7. Passengers are the majority affected in the accidents.
8. Fridays are the worst day for driving as because of traveling lots of accidents are happening. The influence of alcohol can also be another reason.
9. London is highest with accident count followed by Birmingham.

# Discovery

**We have found that the below yellow lines are connecting most of the accident prone areas and they are highways connecting the major cities. Most of accident counts are higher in this highways.**

**Route 1: M25🡪M6🡪M60🡪M08**

**Route 2: M25🡪M62🡪 M60🡪M08**

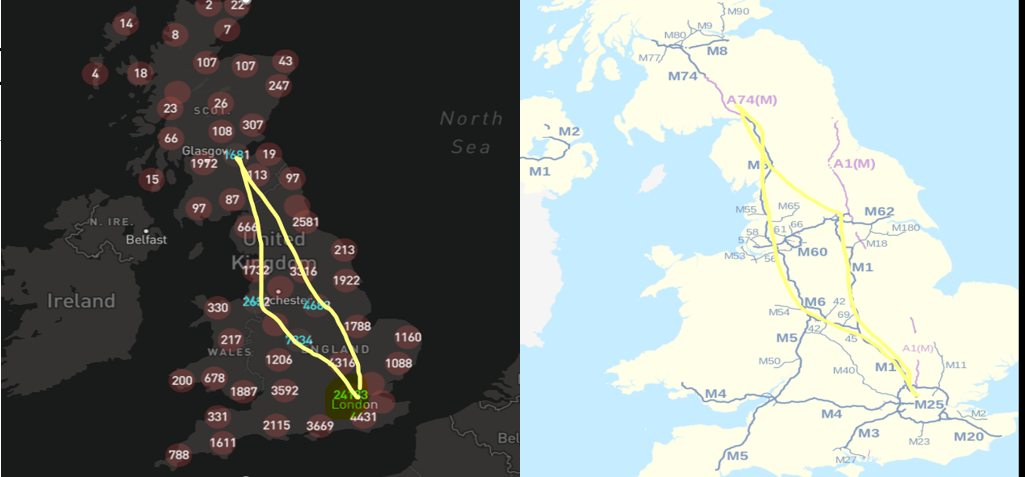


Figure 16: Discovery of highway accidents

# Future research

We have dropped a lot of columns to focus on our goal but those columns are important in different scenarios. In the future, we can also add different other datasets present on the website with the existing data to find a correlation between a lot of other variables. Data never stops up to amuse!!

Also, there are some discrepancies found in the data which reduces the data quality which can be considered in the future for improvement. Examples are below:

1. Missing data
2. Unknown data
3. Not reported categories.

Also, some clustering algorithms can be applied to check if there is any cluster present by accident density. We can fine some of the roads, highways or areas which are in high accident zone. In that way Transport department can put signs on those particular places.

# Conclusion

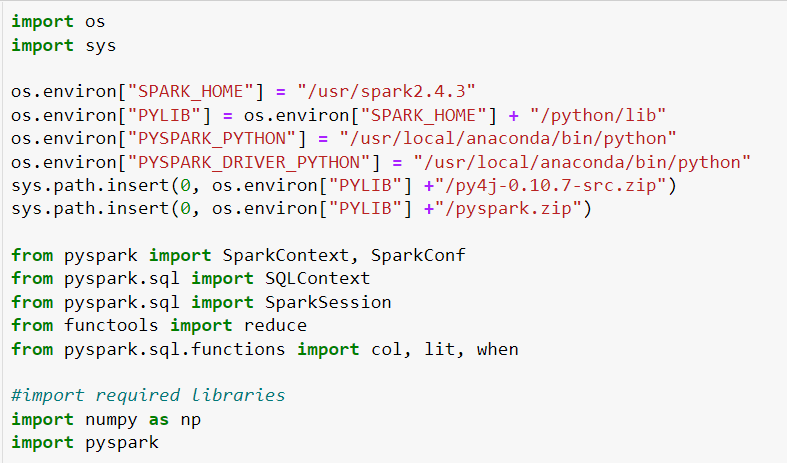
To conclude the department of UK transport has lots of datasets that can be integrated with the current datasets to gain more insights. The only concern is in some places the data is labelled as Unknown or missing which can also improve and help the community to research further.

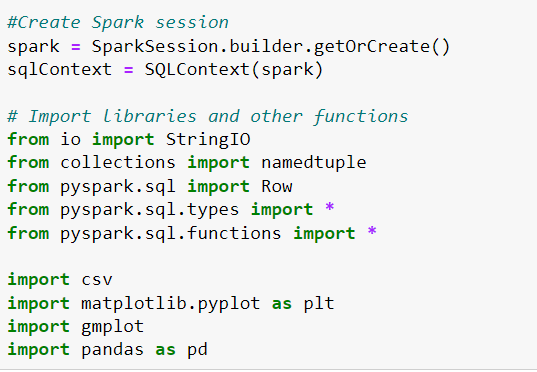
Insights drawn from this report are quite normal with expected assumptions. Only the Covid-19 hit has impacted the normal flow. This dataset has a lot of potential and can be further researched.

# Appendix A

## Configuration

Jupiter notebook configuration for setting up Spark core, Matplotlib, Numpy, SQLContext, Spark Session, etc.

****

****

## Codes

All the codes are attached in Turnitin. Kindly refer to the codes for detailed implementation.

## Convert numerical column values into meaningful data as per metadata

Below code translate all the numeric values into a meaningful human understandable format.

Metadata details can be found below the link:

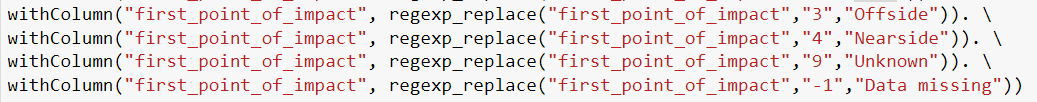
<https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>

regexp\_replace function has been used along with the column function which replaces any regular expression with the argument passed in the function.

Example:

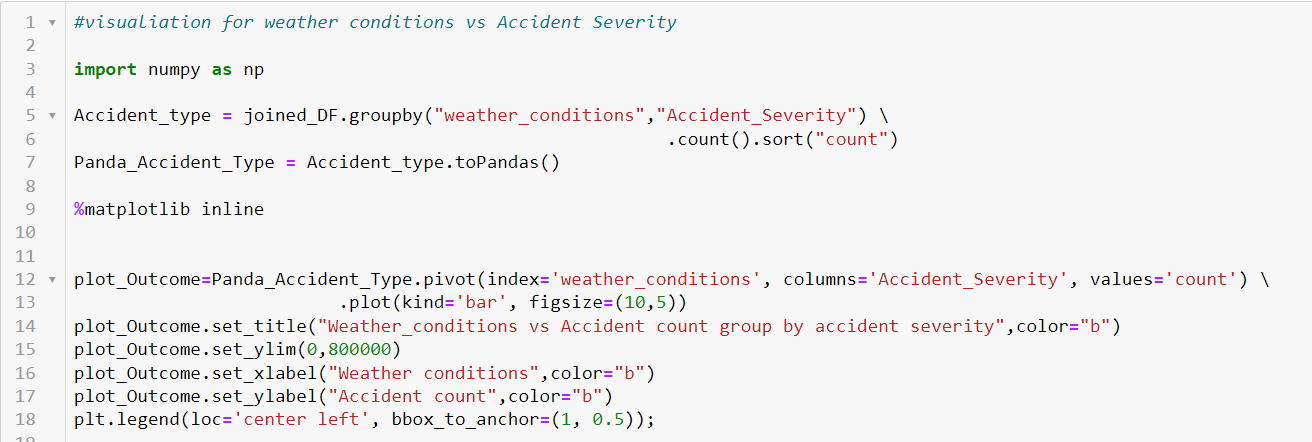
Your\_new\_dataframe = your\_dataframe.withColumn("Column\_name", \ regexp\_replace("column\_name" , "value\_to\_replace" ,"replaced\_value"))

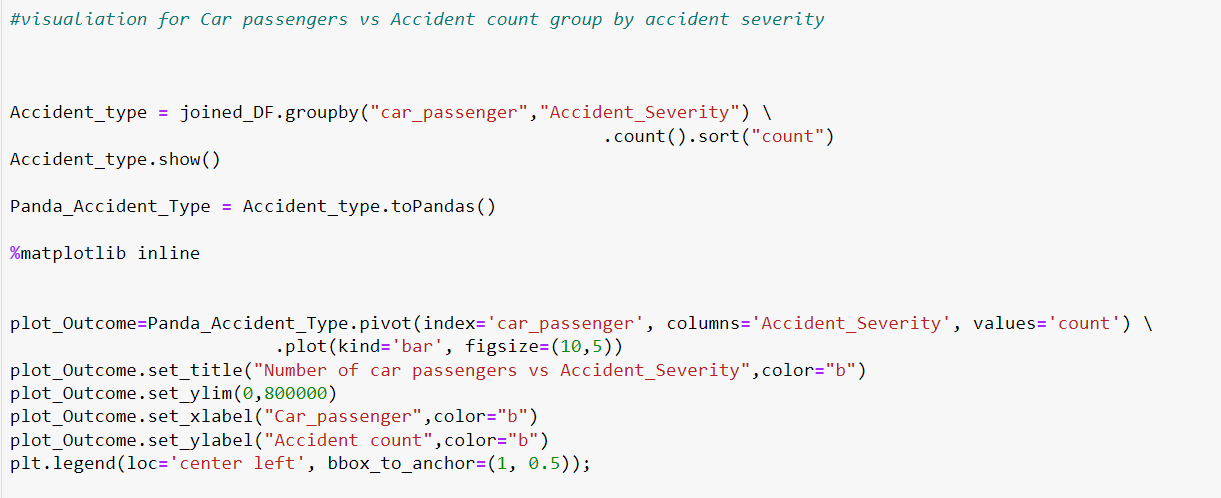
****

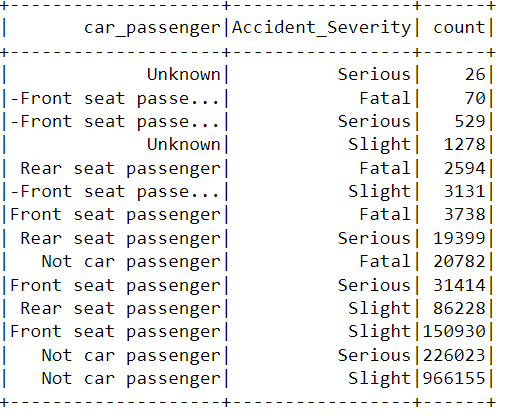
****

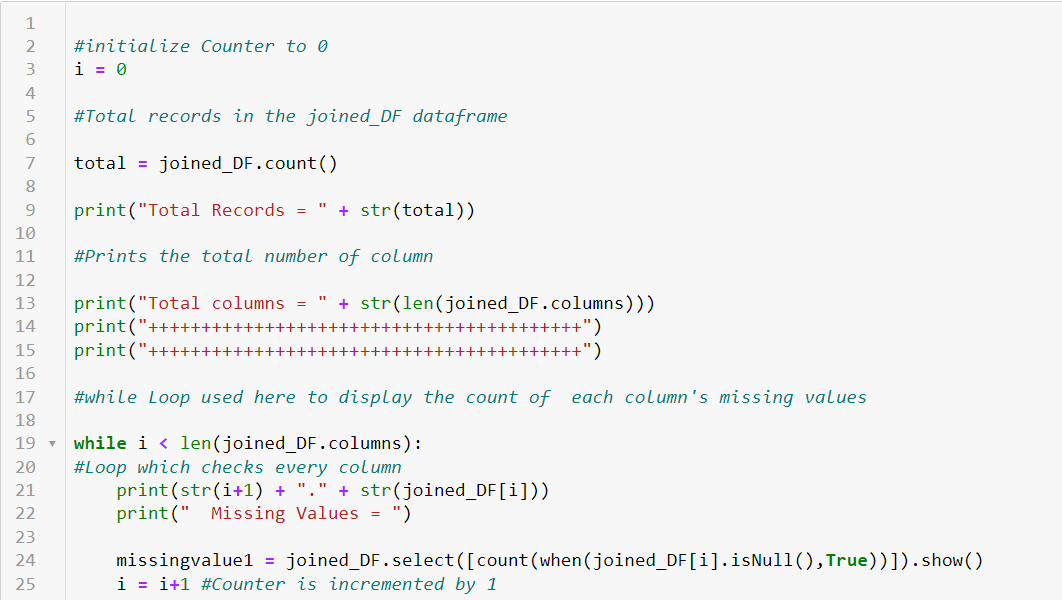
## Matplotlib graph plot code sample





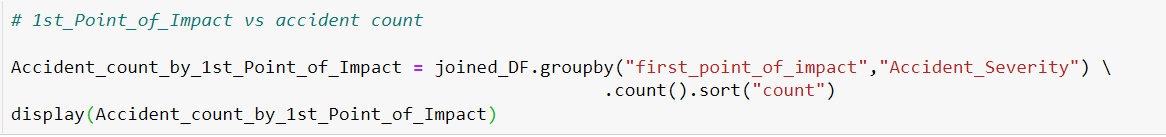


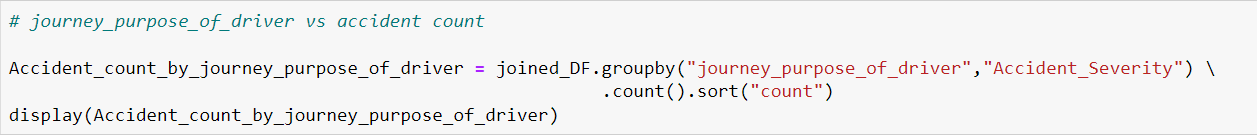


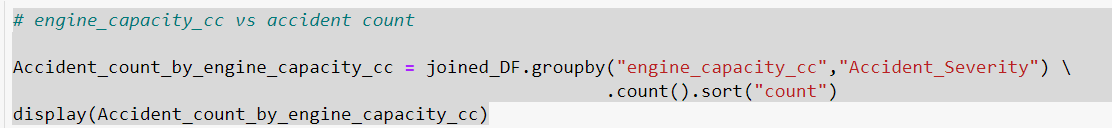


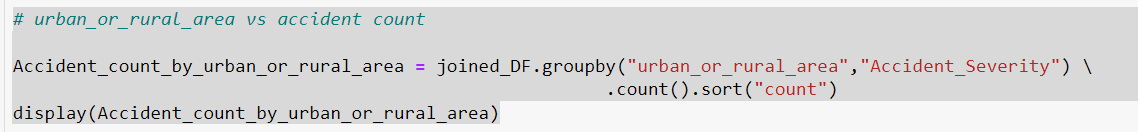
## Pixiedust table generation using dataframe

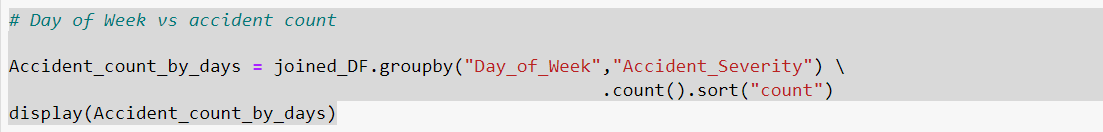
Below are some examples of the Pxiedust table preparation for plotting the bar, pie, histogram, and line charts. Detailed code is attached in Turnitin.



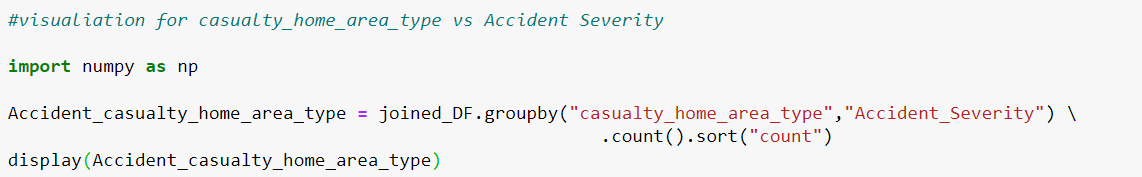












# References

Apache Spark, n.d. [Online]   
Available at: https://spark.apache.org/docs/latest/

UK gov department of transport, n.d. *UK gov department of transport.* [Online]   
Available at: https://www.gov.uk/government/organisations/department-for-transport