**Road Accidents Data Analysis: Report**

The 2016 datasets for road traffic accidents, casualties and vehicles used in this case study are from government data. They contain many variables giving detail about each accident, vehicle and casualty, indexed by an Accident Index reference number. Variables include the age and sex of drivers and casualties, speed limit of the road, location (by Longitude and Latitude) and the police force that attended. The purpose of this case study is to analyse the given data and investigate possible relationships between variables.

Using R it was possible to import the data and plot each accident on a map of Great Britain, coloured by the severity of the accident (Fatal, Serious or Slight). From the graph in Figure 1, we can see that most accidents are only slight accidents, and that accidents tend to occur more frequently in areas with more roads and a greater population density, as you would expect (around Greater London and Birmingham have far more accidents than in the Scottish Highlands, for example).

Map

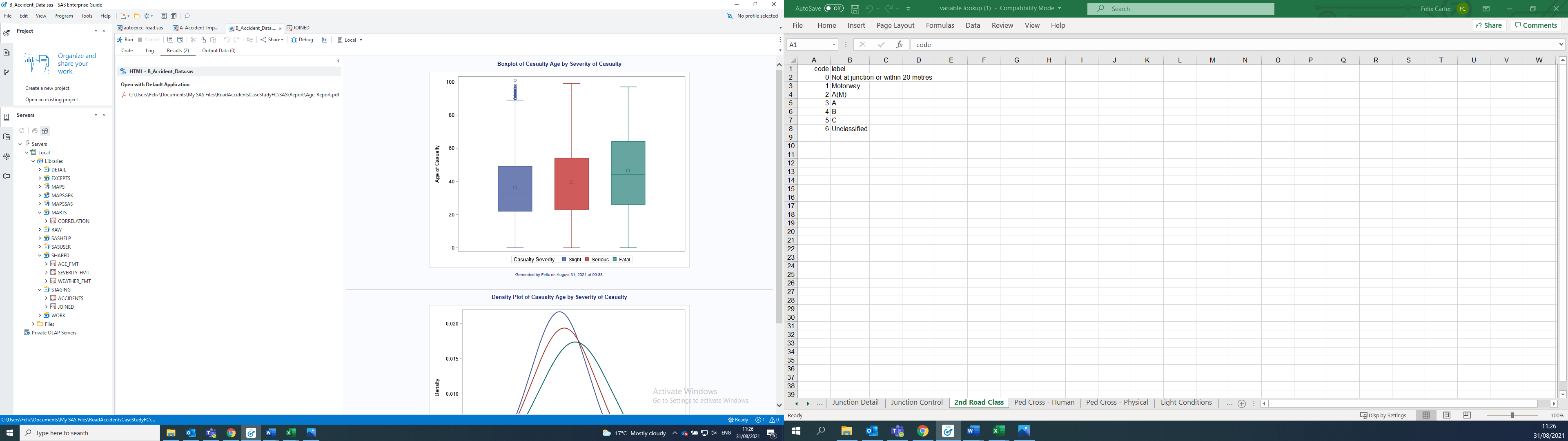
Description automatically generated

**Figure 1:** A map showing the road accidents in Great Britain in 2016, coloured by the accident severity.

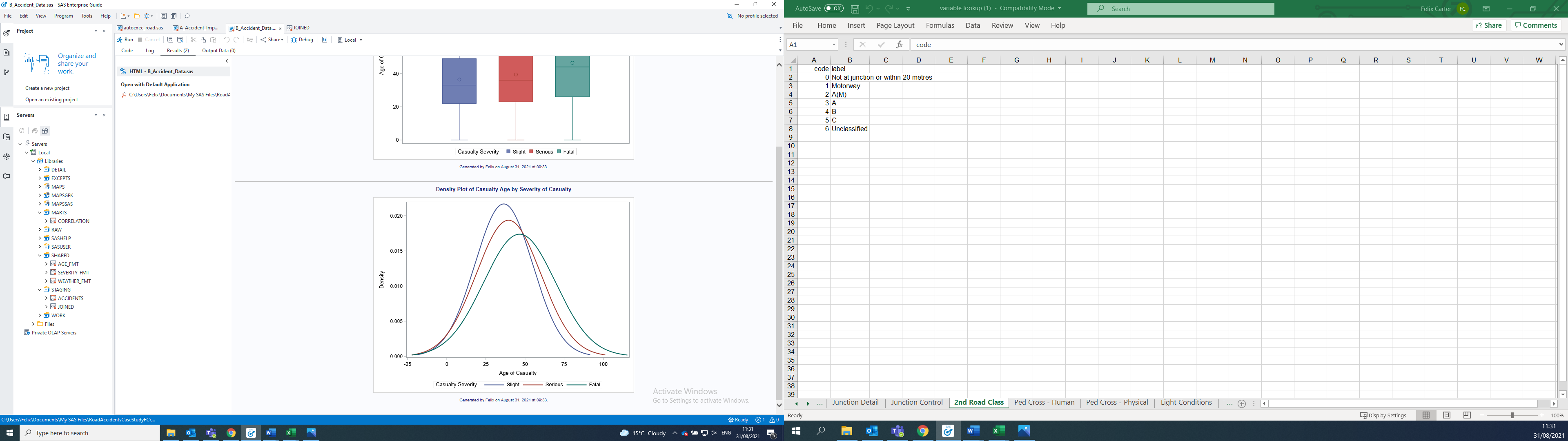
We can see one possible error on this map: off the West coast of Cornwall, there seems to be a serious road accident in the sea – this observation could have had its Longitude/Latitude incorrectly entered in this dataset.

After this, we imported the accidents, casualties and vehicles datasets into SAS and joined them on their Accident Index, applying formats to selected variables that may be used in analysis. We then created a correlation matrix of each numeric variable in the dataset, to see if we could pick out any correlations between variables. Obviously, some correlations would be expected, such as between Longitude and the OSGR Location Easting, as these both measure how easternly a location is. Notable correlations (with a correlation coefficient of over 0.6) include the relationship between the speed limit and whether a location is urban or rural (0.670) and the relationship between the junction detail and the 2nd road class (0.722).

We then used SAS to investigate the distribution of casualty age for each type of casualty (slight, serious or fatal).



**Figure 2:** Boxplots of casualty age by casualty severity.

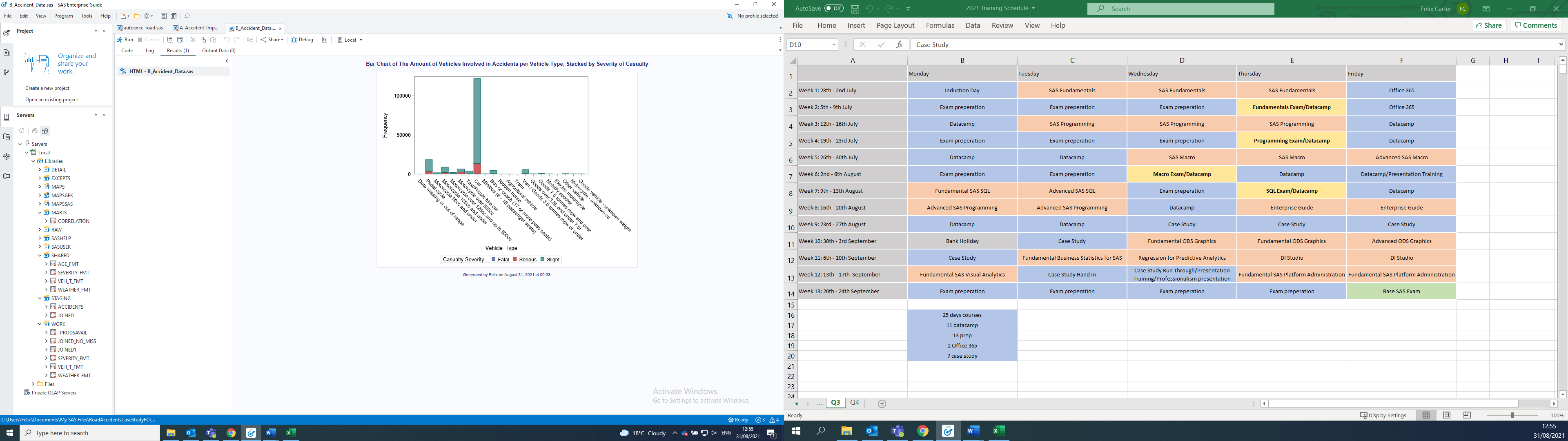


**Figure 3:** Density plot of casualty age by casualty severity.

Although the three casualty severities have very similar age distributions, fatal injuries seem to occur to slightly older people than slight or serious injuries. This could suggest that older people are more likely to die from road accidents. Looking at the distributions of casualty ages, the interquartile range spans from early twenties to approximately the age of fifty for slight and serious accidents, and from mid twenties to early sixties for fatal accidents. However, for each casualty type the range covers all ages.

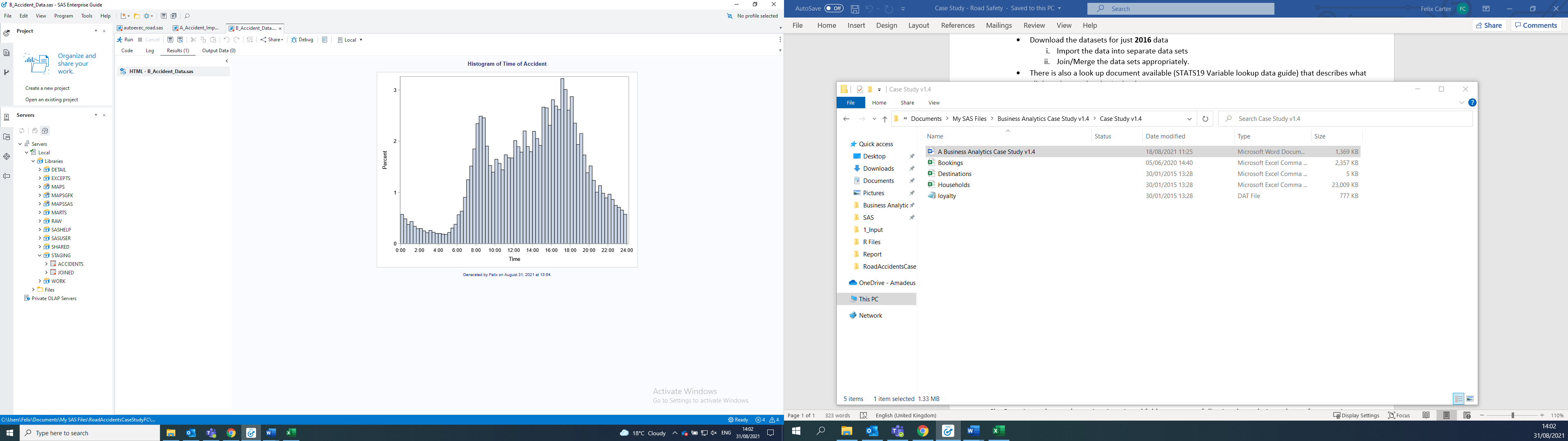
Next, we created a stacked bar chart showing the frequency count of each vehicle involved in a road accident, stacked by severity of casualty hit by that vehicle. Unsurprising, cars were by far the most common type of vehicle involved in a road accident, followed by bicycles and motorbikes under 125cc.

The plot in Fig. 4 also shows that the vast majority of accidents only lead to slight injuries for each casualty.



**Figure 4:** A bar chart showing the frequency count of each vehicle involved in a road accident, stacked by severity of casualty hit by that vehicle.

We then looked at what time accidents tended to occur by creating a histogram showing the percent of accidents occurring in each 20-min period throughout the day, as seen in Fig. 5. Accidents occur most frequently around the ‘rush hour’ times of 9am and 5pm, as seen by the two peaks surrounding these times. Accidents occur at the lowest rate in the early hours of the day (there are far less cars on the road at this time).



**Figure 5:** A density plot of the number of accidents occurring at each 20-min time interval throughout the day.