Accidents in Great Britain (2015)

Varun Agarwal, Ajinkya Khamkar, Jivitesh Poojary, Yatin Sharma

April 28, 2017

Introduction

Road accidents are a major source of human and material loss. The department of transport (Great Britain) states that
1) **There were a total of 186,209 casualties of varying severities**
2) **140,086 personal-injury road traffic accidents were reported to the police**
3) **Traffic volumes in 2015 rose by 1.6% compared with 2014**

Through this project, we try modelling the severity of an accident using demographic information about the drivers, vehicle specific information and regional statistics. We base our study on the severity of accidents caused by heavy vehicles. This enabled better visual inference and pattern identification.

Data Description

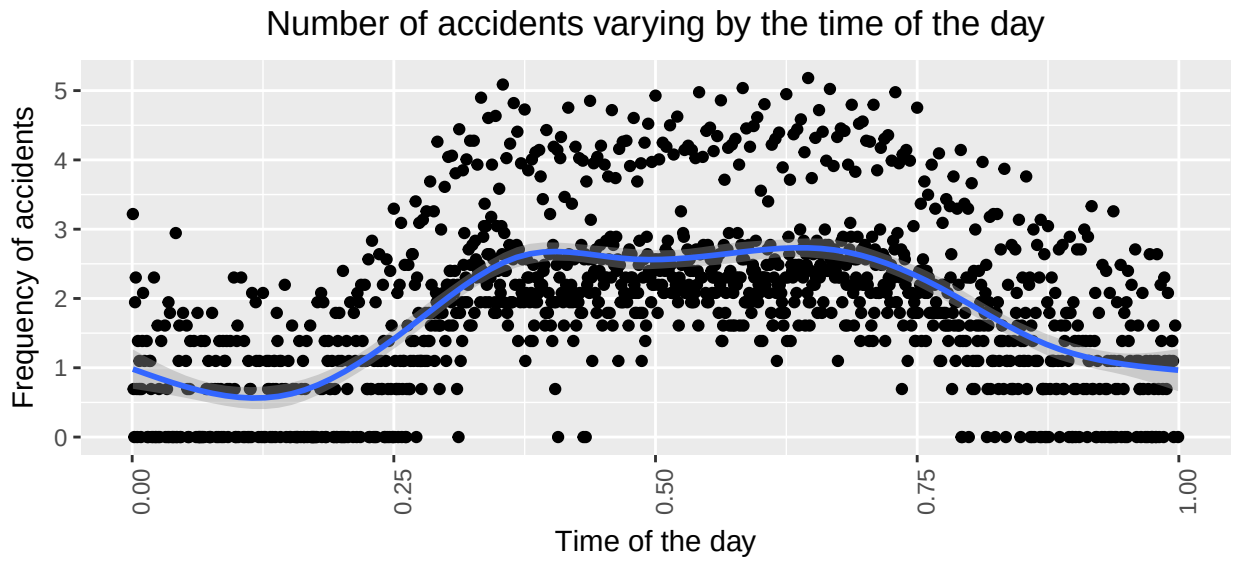
According to the dataset, there were a total of *18824* accidents reported across Great Britain in 2015. There were a total of *3260* reported fatal accidents and a total of *15564* non fatal accidents. We use the following variables to model accident severity

1) **Accident severity** = 1 indicates fatal accident, 3 indicates non fatal accident
2) **Vehicle Category** = Our dataset contains the following vehicle categories 2.1) Coaches
   2.2) MiniBus
   2.3) Agricultural Vehicle
   2.4) Vans
   2.5) Trucks over 3.5 tn
   2.6) Trucks over 7.5 tn
   2.7) Trucks over unknown weights
3) **Date** - Date on which the accident occurred
4) **Time** - Time of day when the accident occurred
5) **Latitude** - Latitude at which the accident occurred, used to cluster regions
6) **Longitude** - Longitude at which the accident occurred, used to cluster regions
7) **Urban or Rural area**= 1 indicates urban area, 2 indicates rural area
8) **Number of casualties unique to each accident** = Total number of casualties involved in accident. Range 1:38
9) **Age of vehicle**: Indicates the age of the vehicle
10) **Age of Driver**: Indicates the age of driver
11) **Gender**: Gender of driver

Exploring the dataset

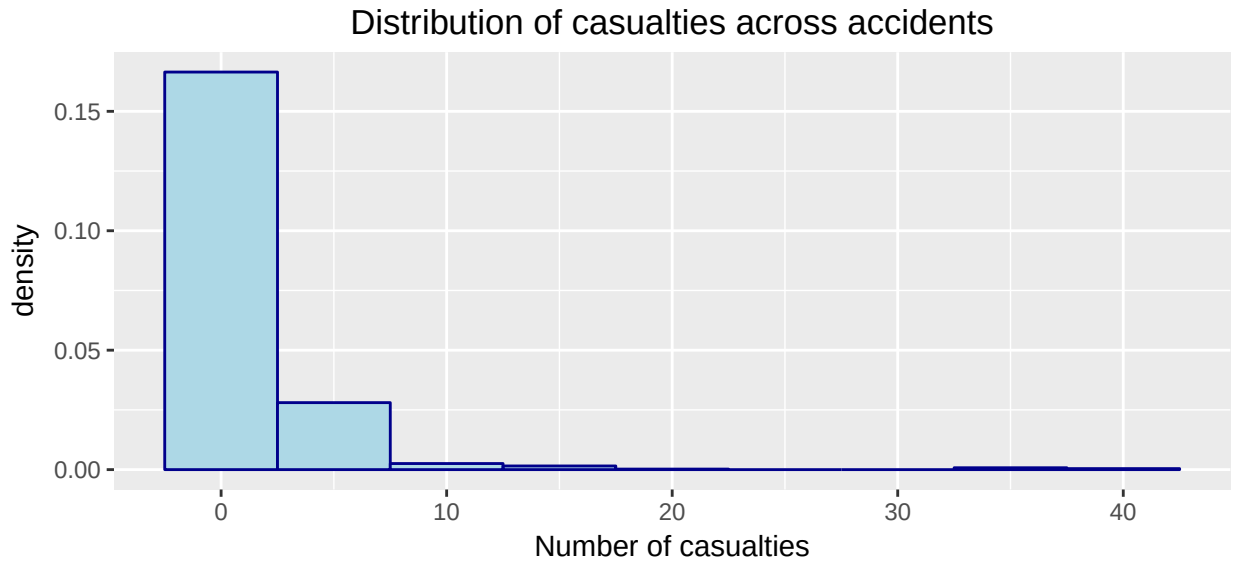
Before we proceed ahead, we would like to explore the data, and have basic overview of what the raw data tells us. We use univariate/bivariate plotting techniques, like histograms and density plots, for some of the variables which we think are interesting and worth looking at.

Let us have a look at how the frequency of accidents vary during different times of the day.



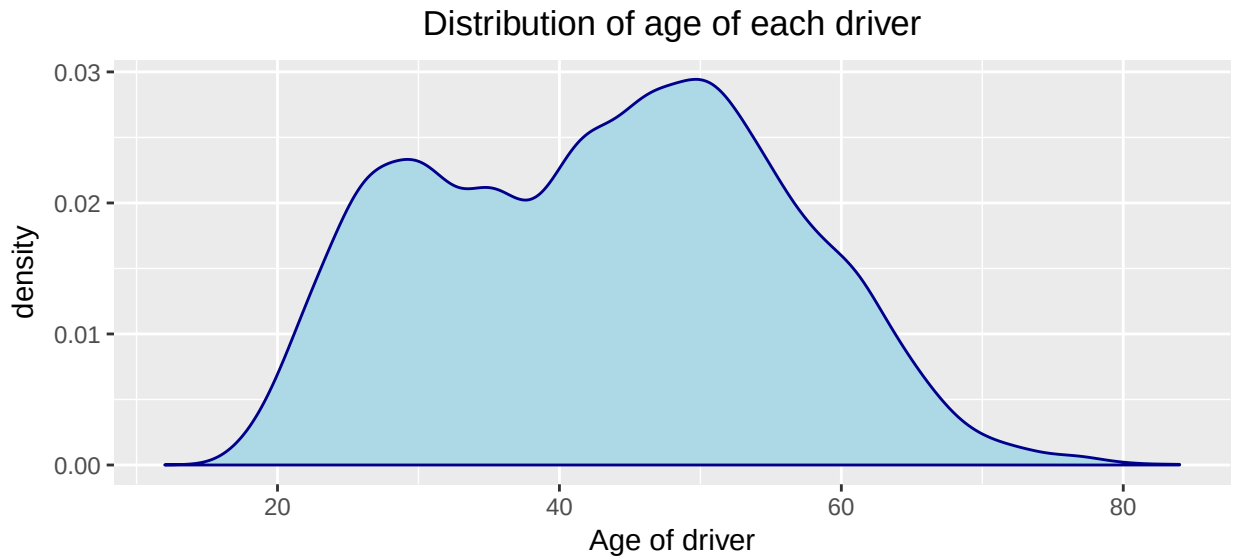
From the above plots, we can infer that the distribution of accidents is not constant throughout the day. Frequency of accidents is considerably lower at night, and gradually increases thereafter. It remains constant through the day, and we see a further dip in the night, as seen at the rightmost part in the graph. This is intuitive, as we would expect the number of vehicles out on the road to be higher during the day and lower during the night.

Let us also look at the distribution of the number of casualties involved in each accident across the whole dataset.



We see that accidents with lower number of casualties are more frequent than accidents with higher number of casualties. This makes sense, since we often come across minor accidents on the road.

Next, we look at the age distribution of the drivers involved in accidents.



We see that the density of accidents is highest for drivers between the age group of 40 - 60. This could simply be because heavy vehicles are probably driven by more experienced driver. Another probable reason could be that older drivers are more prone to accidents, owed to their age factor.

Let us look at the gender distribution of drivers getting into accidents while driving heavy vehicles.

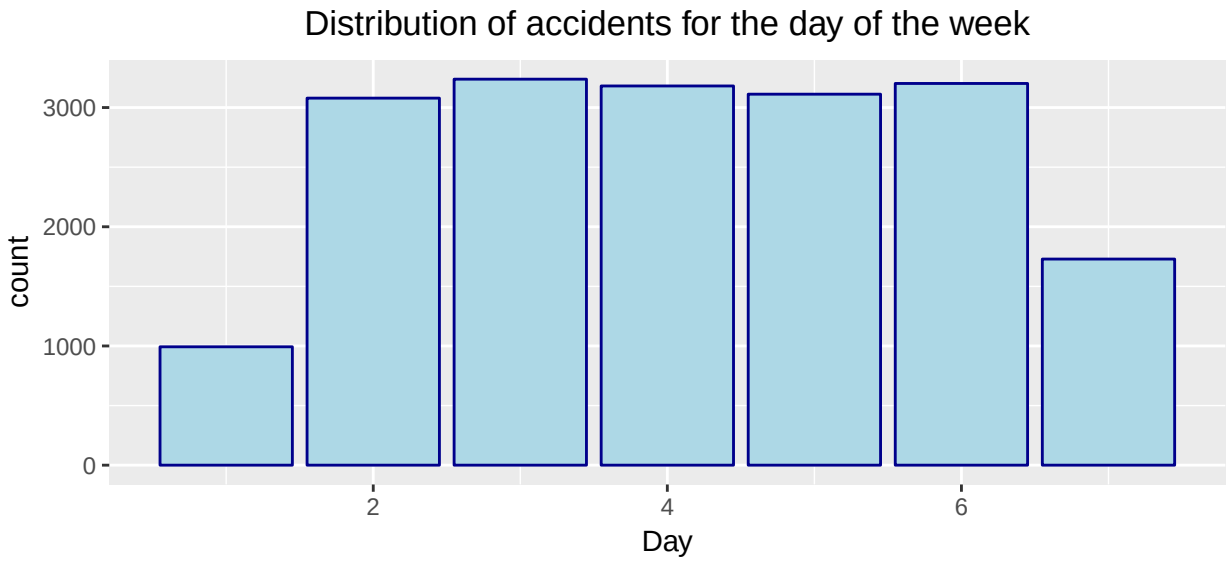
Female	Male	NA
896	17575	63

We see that our data is pretty skewed towards male drivers. This is mostly due to the reason that there are more male drivers on the road.

Below, we look at the distribution of various vehicle types that were involved in accidents across the whole dataset. We observe that maximum number of vehicles were Vans followed by Coaches and Trucks.

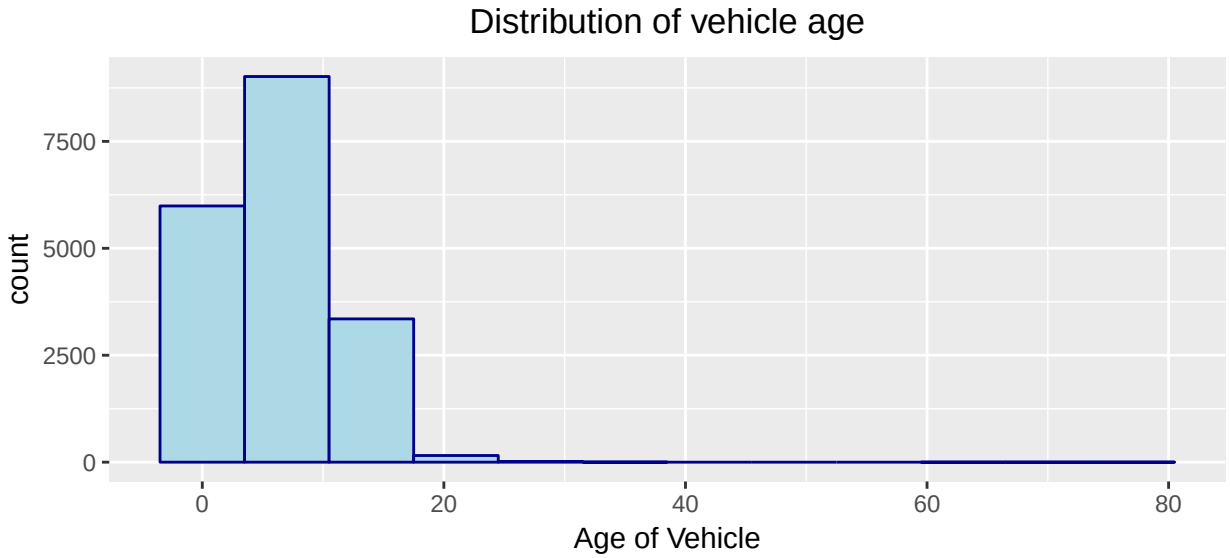
Coach	MiniBus	Trucks over 3.5 tn	Trucks over 7.5 tn	Trucks unknown weight	Vans
5235	303	701	2808	289	9198

Next, we would like to see if there is a trend in the distribution of accidents across all days of the week. We have summed the instances of accidents over the whole year across all the days.



We see that the number of accidents during the weekends is considerably lower than the number of accidents across the weekdays. Sundays are least likely days for accidents to occur on the roads. This should be because there are generally lower number of heavy vehicles on the road during non-business days.

Let us also see if the age of vehicle matters in the cases of accidents.



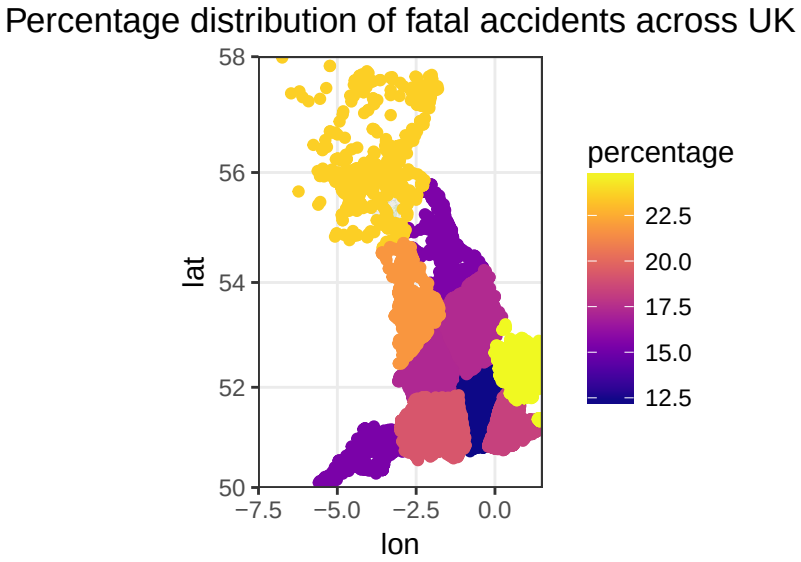
We see below that the distribution is skewed towards the right, with majority of the vehicles being less than 20 years of age. Vehicles with an age of more than 20 are less likely to be seen on roads, which is probably why we have such less instances for them. Other than that, we do see that the number of accidents caused by vehicles aged between 7-14 years is considerably higher. We can infer that aged vehicles are more likely to be involved in accidents.

Finally, let us look at how the occurrence of accidents vary in urban and rural areas. We see that occurrence of accidents in urban areas is more than rural areas. That is probably because there are more number of vehicles in urban areas than in rural areas.

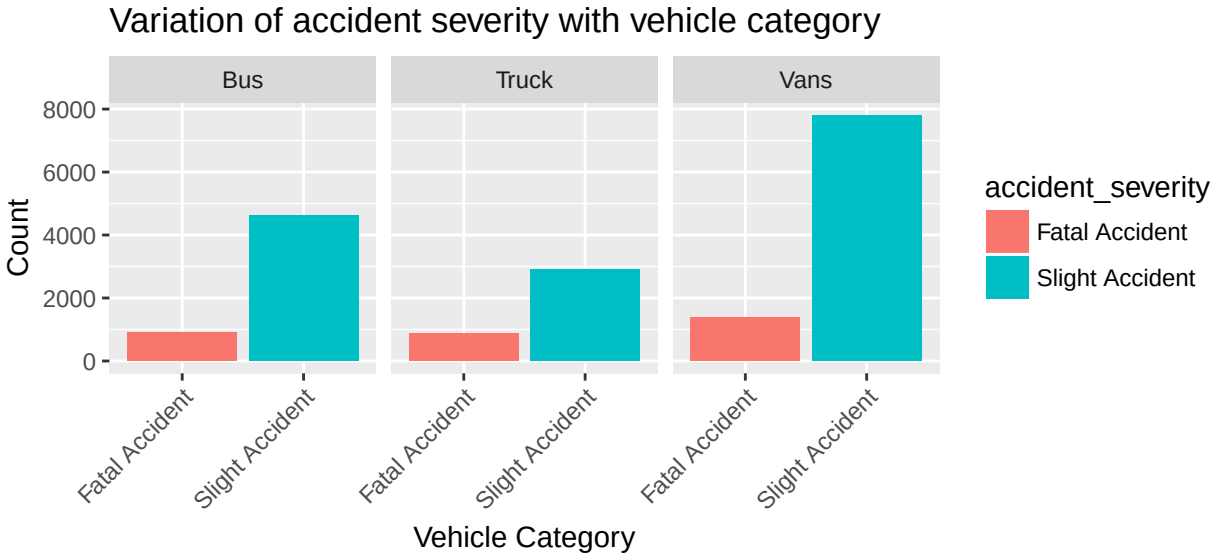
Rural	Urban
7566	10968

Exploring relationships between Accidents Severity and other variables

We believe the distribution of fatal accidents is not constant across United Kingdom. We clustered the longitudes and latitudes of accidents available in our dataset into 10 clusters. The following plot represents the percentage frequency of fatal accidents in each of the 10 regions.



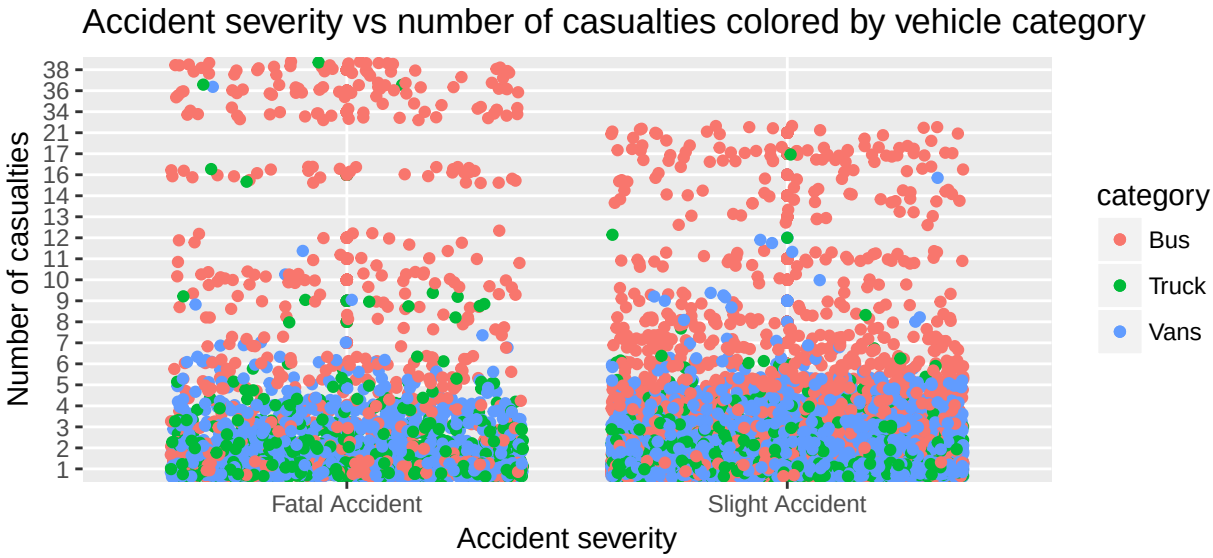
Let us see how accident severity varies with the category of the vehicle.



We can see from the above graph that the ratio of fatal accidents to slight accidents is not constant across all categories of vehicles. We can see that Trucks have a higher ratio of fatal accidents to slight accidents,

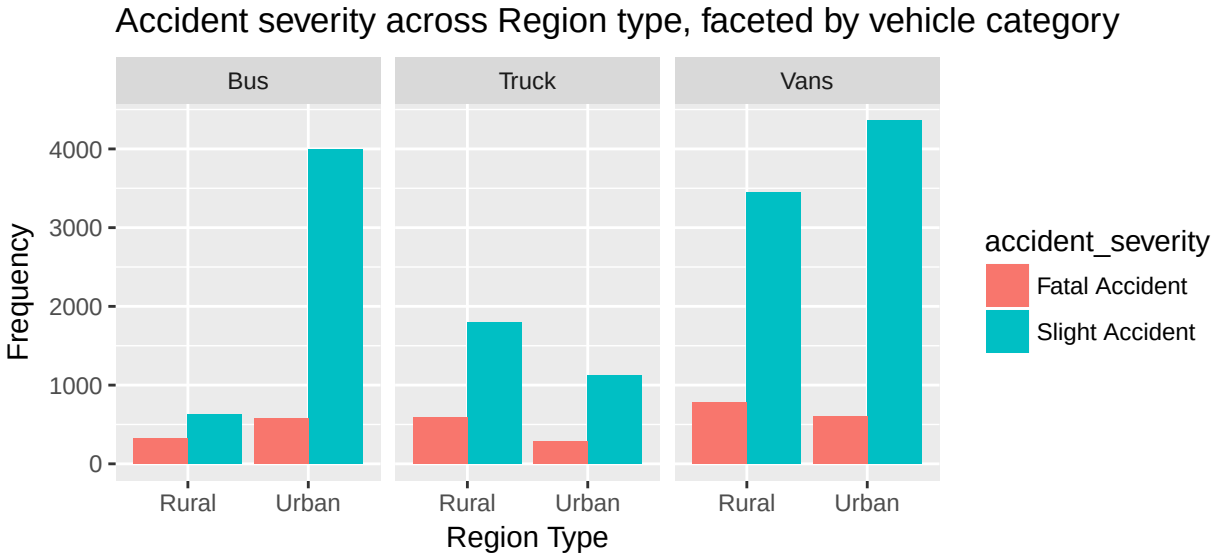
followed by buses and vans, in that order.

Now, let us plot the number of casualties against severity of accidents, and colouring it by the vehicle category, to see if we can find any trend.



We see that buses generally have a higher number of casualties during accidents, which is pretty intuitive since they accommodate higher number of passengers. We tried to see if we could find a possible interaction between the two variables, but we do not observe one that should be included in the model.

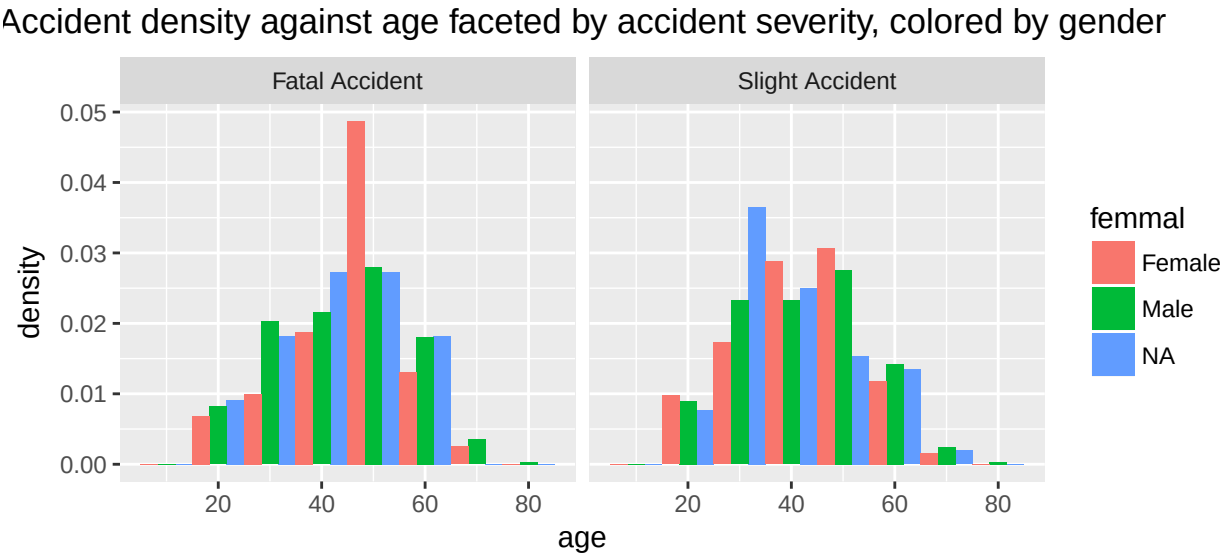
Below, we plot the severity of accidents for different types of regions, across all categories of vehicles to observe possible trends and interactions. However for slight accidents the proportion of vans is more than other categories. In the case of urban areas the density of slight accidents is significantly higher than for rural areas. These are some interesting things that can be observed in the plot. However it does not seem to be significant enough to be considered as an interaction.



We see that the ratio of fatal accidents to slight accidents in rural areas is generally higher than the ratio of fatal accidents to slight accidents in urban regions, across all vehicle categories. This seems like a good differentiator for predicting severity of accidents. We do also see a slight interaction between vehicle category

and region type, but we shall use this interaction only if it is absolutely necessary for the model.

Now, we shall see how the severity of accidents vary by the age and the gender of the drivers.



In the above graph, we see that drivers of both genders between the age of 40-50, are more prone to cause fatal accidents. Slighter accidents are generally higher for drivers between the age of 30-40, for both genders. Also, we see that the trends for fatal and severe accidents across both genders are not the same. So, we see a possible interaction between gender and age.

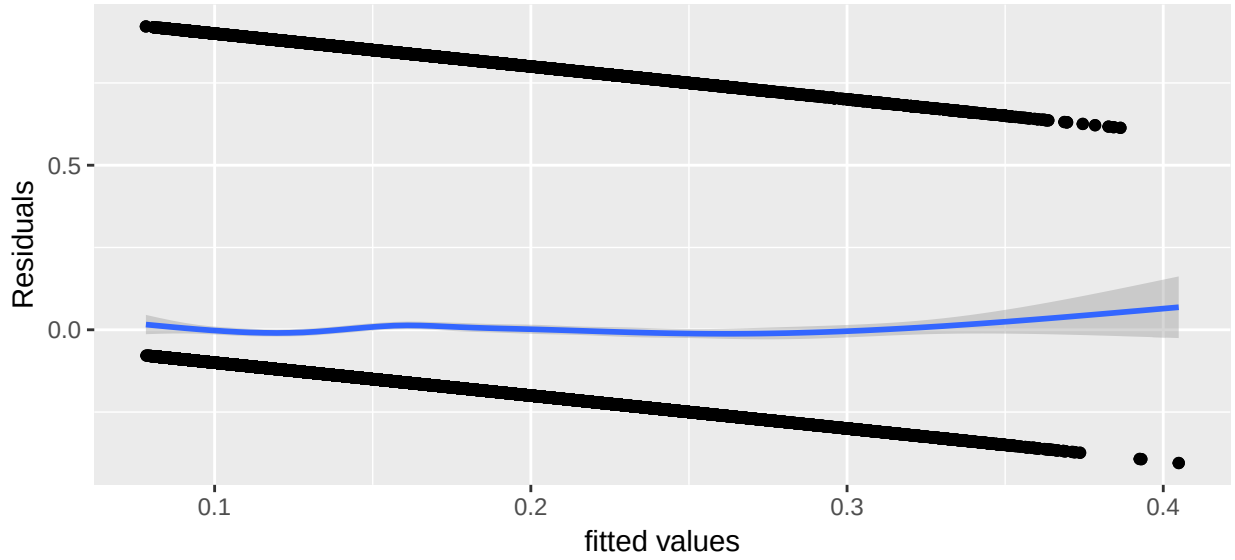
Modelling Accident Severity

We have explored our data and have a good understanding of the factors which would have an affect on Accident severity. Now, we will model accident severity against vehicle category, region type, age of driver, sex of driver, regions (differentiated by percentage of fatal accidents), and age of vehicle. We have not included any of the identified interactions in our initial model, just to see how our model performs without them. We have used multiple logistic regression to predict the probability of a fatal accident, given the predictor variables.

Model Equation

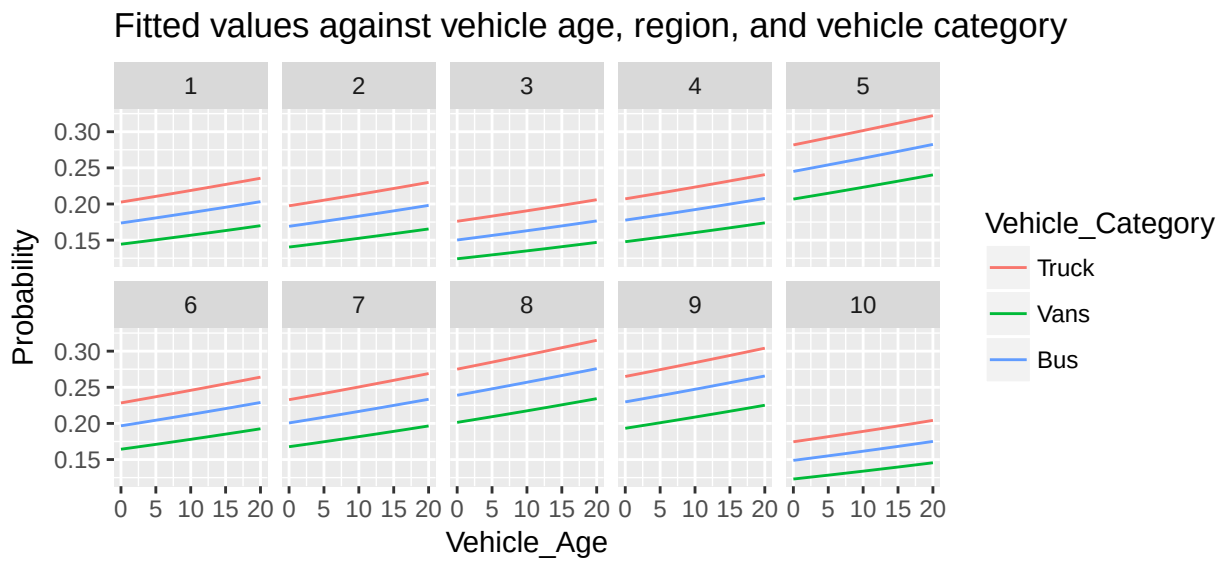
```
model2 = glm(factor(accident_severity, levels = c("Slight Accident", "Fatal Accident")) ~
    category + region_type + age_of_driver + femmal + region + age_of_vehicle,
    data = df1.hex.subset, family = binomial(link = "logit"))
```

Let us have a look at the fitted vs residuals plot to see the fit of the model.

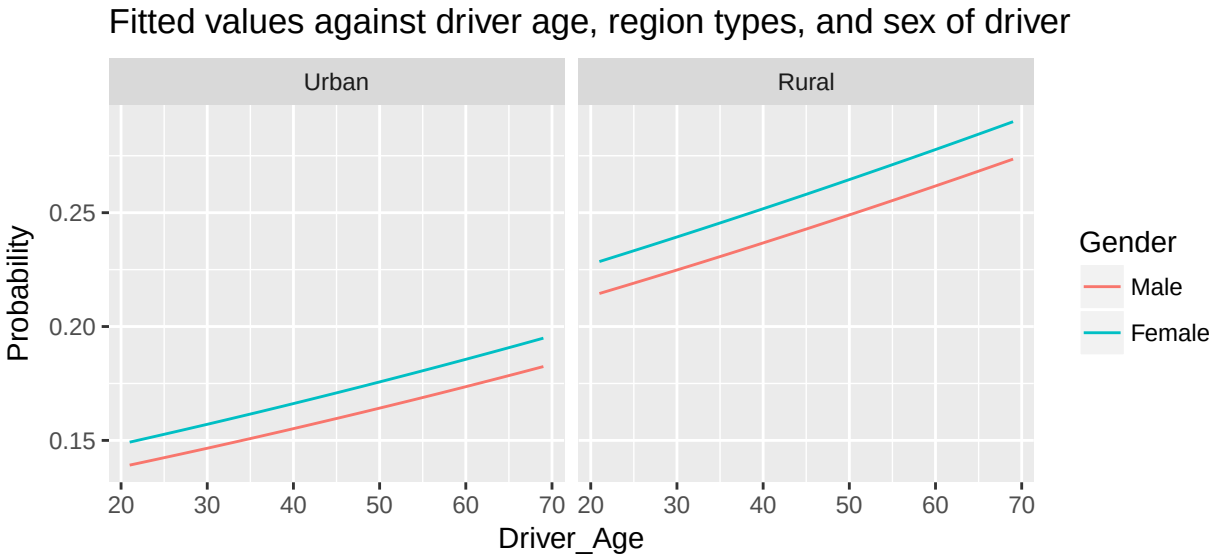


The fit is pretty good and consistent throughout, except for the very slight kink in the middle, and that it sways upwards from the zero line by a bit at the right hand side of the plot. The slight sway is probably due to lesser data points at that end. Overall, it seems like the categorical predictor variables used in the model predict the fatality of the accident pretty accurately. We also conclude, that there is no need of including the interactions in our model, since we have a good fit without them.

Let us see the trends our model presents across different predictor variables.



We see that trucks are generally more prone to cause fatal accidents, followed by vans and buses, in that order. We also see that the odds of fatal accidents increase with the increase in the age of the vehicles. These trends seems to be constant throughout all regions. The probability across different regions of the map is not constant. Region 8 seems to be most prone to fatal accidents, whereas region 5 is least prone to fatal accidents.



We see that rural areas are in general are more prone to fatal accidents than urban areas. We also see that female drivers for heavy vehicles are more prone to fatal accidents than male drivers, in both rural and urban areas. We also see that the odds of fatal accidents increase with the increase in the age of drivers. Somehow, these odds show a slightly more rapid rise for accidents in rural areas, than for accidents in urban areas. The reason for this is not very apparent, though.

Conclusion

In conclusion, we see that our model seems to explain severity of accidents accurately, and confirms with all the observations we had made on the unmodeled data. It performs reasonably well without the identified interactions, too.