**Problem understanding:**

Every year there is 53k number of accidents on the British roads out of which 44% is severe. Hospital directors that plan resources may want to predict risk and severity of accidents on streets to help people that were victims of car accidents. Heads of police departments want to know how to protect citizens and patrol the streets so to limit dangerous behaviour on the streets. Aim of this project is to predict the severity of road accidents so to limit the number of incidences on the way and optimize resources to protect victims. To understand what influence the severity of the accident we may want to look at things like: weather conditions, light conditions, road conditions or place location.

**Data source:**

To this project I have used data from British department of transportation on car accidents that happened in years 2010 to 2018 <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data> The raw data is made of two files: vehicle details and accident circumstances. Vehicle details include facts such as: vehicle type, junction location, sex or age band age of driver, vehicle propulsion code and more. Accident circumstances include facts such as: police force division that attended the accident, road type, speed limit, weather conditions, urban or rural area and more. For full details on columns included please refer to the appendix.

Looking at the data I have noticed that the accident with maximum number of vehicles was not on the same data as accident with maximum number of casualities. The maximum number of casualities recorded during car accident was 93 on date 20/10/2014 and the maximum number of vehicles recorded during car accident was 67 on date 05/09/2013. I found information on the second accident in the press so it will allow me to validate the accuracy of the data.

**Data Cleaning:**

The data comes from multiple years, each year in separate csv file. Comparing columns across years we see that Age\_of\_Driver and Vehicle\_IMD\_Decile are not in all files across the years so we remove the variables in files that the columns exists, that’s to uniform the data. We normalize the column names across the files. Files from the same category: vehicle details or accident details we append on top of each other to have one dataset for each year and then we join together both complete files on accident index column. To better understand the data types and what columns say about the incident I recode the data using variable lookup file provided. Some of variables are in too granular level so that the ratio across different options is too big. For this reason some columns are changed to Yes|No type columns such as: Vehicle\_Leaving\_Carriageway, Hit\_Object\_off\_Carriageway, Carriageway\_Hazards and more. Finally we have 37 columns of type string, 11 of type float and 5 of type integer making a total of 53 columns and nearly 3mlns observations.

Some of the missing values are left blank, some are recorded as -1 and some are recorded as unclassified or else. To better understand the missing values I calculate proportion of all missing values cases per column. For some columns such as Was\_Vehicle\_Left\_Hand\_Drive?, Propulsion\_Code, 2nd\_Road\_Class, Junction\_Control, Age\_of\_Vehicle I imput missing values by the most popular option. Time column I imput using forwardfill as I noticed that the data is in chronical order. For column Age\_Band\_of\_Driver the missing values I change on Unknown. I noticed that the data has some records for driver in the age 1-5, 5-10, 10-15 since this is illegal to drive in this age and that doesn’t make sense I simply make this as unknown as well. For columns such as Engine\_Capacity(CC) or Driver\_Home\_Area\_Type the missing values I imput using mode grouping by Vehicle\_Type or Accident\_Severity. Since I want to visualise the data on a map with accident severity but I don’t need the columns Longitude and Latitude in the general data I create temporary data set with Accident\_Index, Accident\_Severity, Longitude and Latitude and I remove those two geographical columns from the main dataset. In total, in the initial cleaning stage, I have removed 10 columns either because the number of missing values was more than 50% or because they would not give good insights. For example, Location\_Easting\_OSGR and Location\_Northing\_OSGR are geographical variables. We already stored the longitude and latitude and in the main data we have column Police\_Force that gives information what area of Britain the Police\_Force operate in. Local\_Authority\_(District), 1st\_Road\_Number, 2nd\_Road\_Number are too granular. We don’t have information on LSOA\_of\_Accident\_Location or Driver\_IMD\_Decile to know what does it mean.

I have noticed that for each vehicle taking part in the accident there is separate record so all columns with small proportion of missing values first I tried to get the missing values from the accident index as they were taking part in the same accident. In total I managed to get 84% of the total data cleaned.

The first view on data type shows that there are 26 columns as float, 22 columns as integer and 5 columns as string. However, that’s not true as categorical data is coded in numbers. After recoding we find out that 37 columns are strings, 11 are floats and 5 are integers.

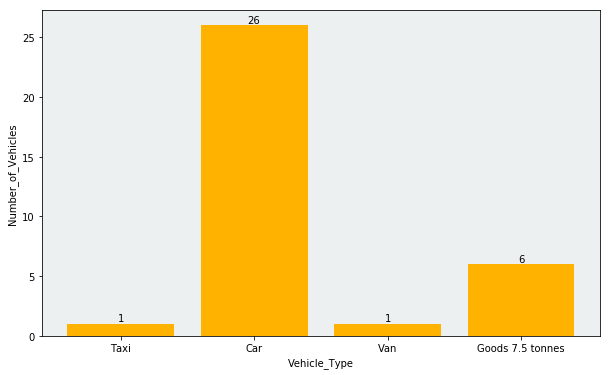
**Big accidents examples:**

**04/11/2011**

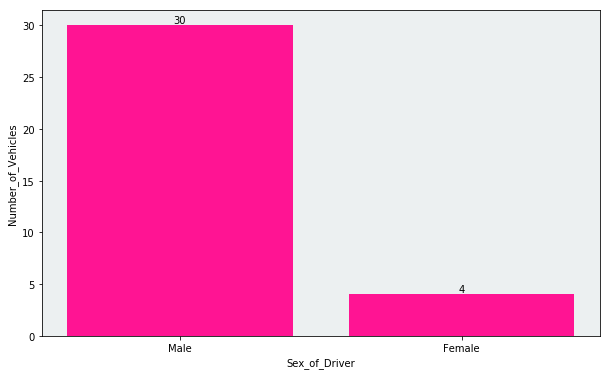
I decided to have a close look at 2 accidents. One from 4th of November 2011 with 34 number of vehicles taking part in the accident and another from 5th of September 2013 wth 67 vehicles. That’s to validate the data, how accurate the accidents were recorded and to understand columns better.

On 4th of November 2011, Friday 8:23pm was mass of crashed vehicles described by emergency worker as “the worst road traffic collision anoyone can remember”. 7 fatal victims, 51 injured and 34 cars taking part in this collision close to junction 25 of the M5’s northbound carriageway. Drivers reported a white wall of fog. People said that smoke from fireworks at the rugby club next to the motorway might have contributed to the poor visibility. There was also huge fireworks display that could distract the drivers. Cars that were driving slow managed to escape but cars speeding up with 60-70mph were just colliding one onto another. 3 lorries collide including one full of deodorants and cans. People were trapped in their vehicles and were screaming. <https://www.youtube.com/watch?v=1ZdsxShKJe0>

The data shows only 5 fatal casualities so 5 reported on the scene and 2 passed in the hospital. We learn that the Local\_Authority\_(District) is Taunton Deane. That was on Motorway, dual carriageway with speed limit 70. Not on junction or not junction within 20 meters. No pedestrians involved. In the data it was reported as daylight, which can't be true as that was November after 8pm. Weather condition from report as 9: Unknown but we know from press and reports that was fog, 100% humidity and dark outside, Road\_Surface\_Conditions: reported as 1 - dry That is not true as it was fog with humidity 100%. 6 Goods vehicles 7.5 tonnes mgw and over, 1 Van / Goods 3.5 tonnes mgw or under and rest normal cars



7. 4 female drivers and rest male drivers.



After this accident analytics we can identify that:

1. Important the severity is from 1 - 3 with 1 the highest

2. As that's interesting during data exploration this level is too granular and we won't include Local\_Authority\_(District) for our model

3. We can notice that we have multiple records for each accident as the details for vehicles are specific for particular vehicle.

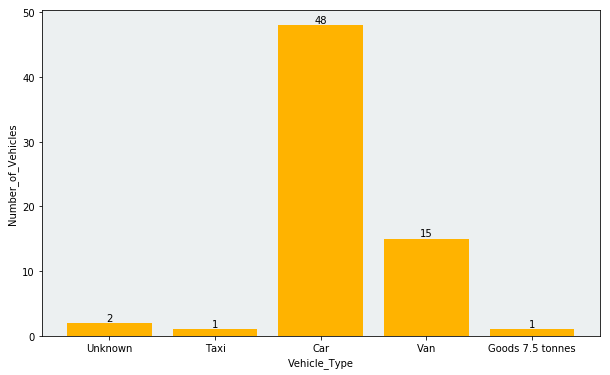
4. We can remove the 1st\_Road\_Number and 2nd\_Road\_Number as does not say something relevant for the model

**05/09/2013**:

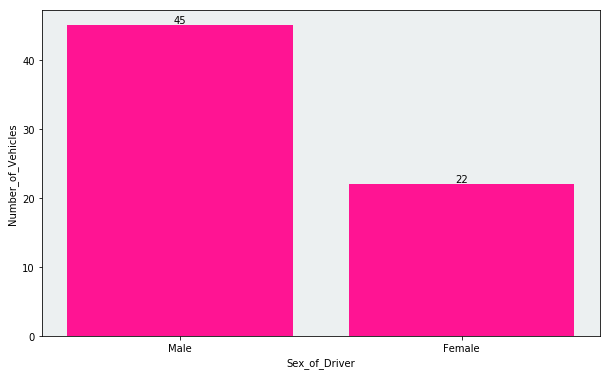
Another big accident: Carmageddon recorded on Thursday. Hundred car pile-up on foggy Sheppey crossing bridge on A249 in Kent at 7.15am. The crash happened as parts of Kent, Essex and parts of London suffered from some areas of dense, patchy fog between 4am and 8.30am. It was forecasted to be mostly dry, sunny and very warm over central, southern and eastern England once the early fog had cleared. After incident there was a call for a safety review on the A249 Sheppey Crossing to look at the speed limit, lack of matrix warning signs and lighting.’ Gordon Henderson, Conservative MP for Sittingbourne and Sheppey, said:‘I have had concerns in the past, particularly about the level of lighting on the bridge.’

Eight people seriously injured and another 200 with minor injuries - 33 people taken to hospital but no fatalities after the incident. A fleet of 30 ambulances and other response vehicles treated some casualties at the roadside. Firefighters used hydraulic cutting equipment to free those who were trapped. Stranded motorists sit on road for up to eight hours in 30C heat but left area by 3pm. The four-lane bridge was shut in both directions. Bridge reopened at 5.30pm. AA chief sayed crash may have been caused by 'stupid driving' - specifically tailgating and not using fog lights.: [More informtaion you can find in the Guardian article](https://www.dailymail.co.uk/news/article-2412099/Sheppey-crash-chaos-Kent-200-injured-100-vehicle-pile-up.html)

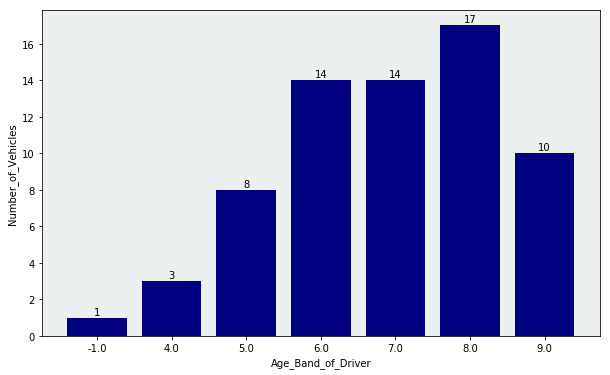
Data shows 67 vehicles and 70 casualities out of which 2 are serious, 33 taken to hospital rest were injured but not hospitalisation needed. Similarly like the previous one it was on dual carriageway with speed limit 70. Light conditions was recorded as daylight, weather conditions: Fog or mist. Road surface: dry.In the accident took part: 1 Taxi, 15 Vans, 1 Goods 7.5 tonnes rest cars:



45 male + 22 female drivers:



3 drivers in age 16 - 20, 8 in Age 21 - 25, 14 in age 26 - 36, 14 in age 36 - 45, 17 in age 46 - 55 and 10 in age 56 - 65



**Missing values imputation:**

The only columns that don’t have missing values are: Accident\_Index, Police\_Force, Accident\_Severity, Number\_of\_Vehicles, Number\_of\_Casualties, Date, Day\_of\_Week, Time, Local\_Authority\_(Highway), 1st\_Road\_Class, Was\_Vehicle\_Left\_Hand\_Drive?, Propulsion\_Code,

I have imputed missing numbers by most popular option in columns: Was\_Vehicle\_Left\_Hand\_Drive?, Propulsion\_Code, 2nd\_Road\_Class, Junction\_Control, Vehicle\_Type,

For column Age\_Band\_of\_Drvier I filled missing values with Unknown,

In Time column since I know don’t data comes in chronological order I filled missing values with forward fill.

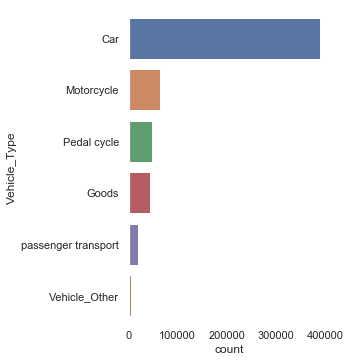
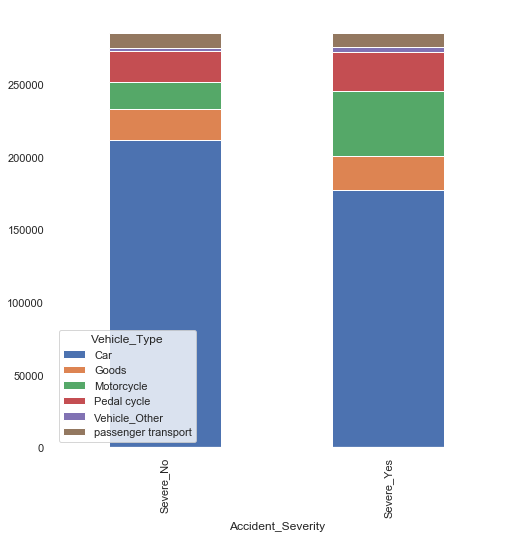
Engine\_Capacity\_(CC) I filled with most popular option by vehicle type, driver home area type by accident severity.

**Imbalanced dataset:**

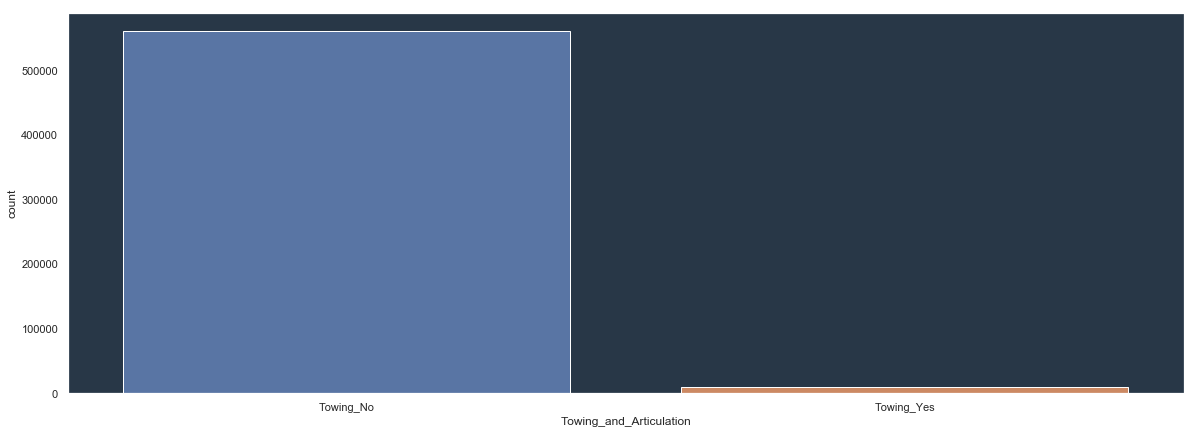
Data imbalance usually reflects an unequal distribution of classes within a dataset, for us it is Accidents severity. On British road majority of accidents is not sever but we are interested in the sever accidents. We want to see how different things impact severity of accidents. To remove this problem we need to normalize the data. To do that first we shuffle randomly the dataset then we separate sever and non severe classes and take the sizes of samples for each class. We finally concatenate the data sets together.

In the dataset we have variables that are unique for each car taking part in the accident such as age of vehicle or journey purpose of driver and we have columns that are uniform for all such as weather or road conditions.

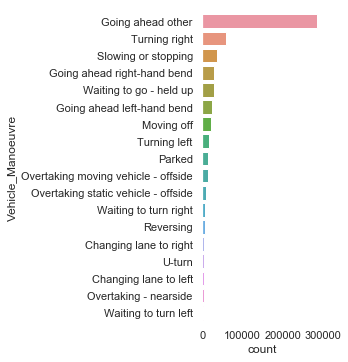
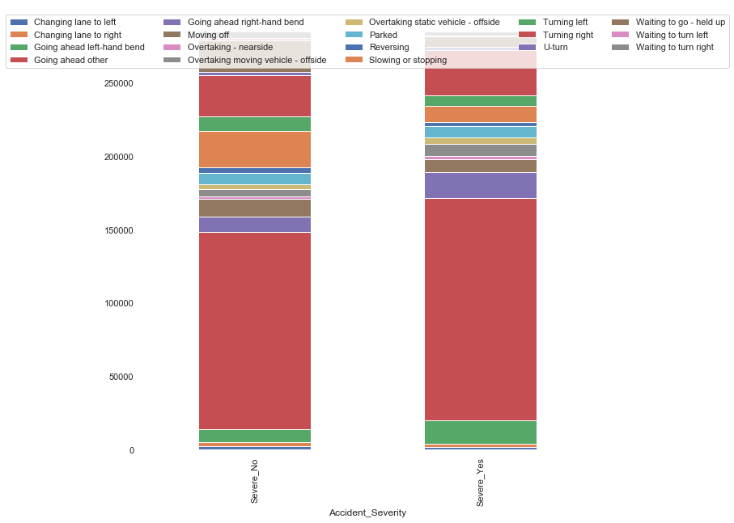
The distribution of vehicle types shows that majority of accidents with motorcyles are severe where majority of cars ccidents are non severe.

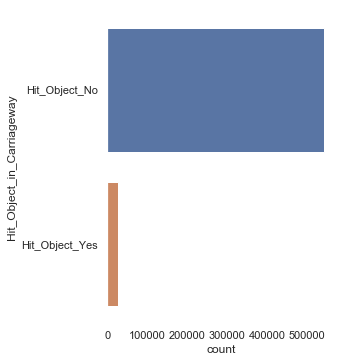
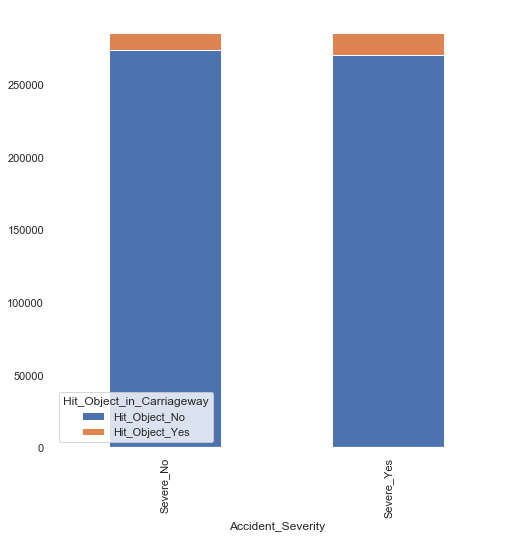
We can see that Towing and articulation distribution is not balanced, majority of the accidents had no Towing. I would recommend to remove this variable:



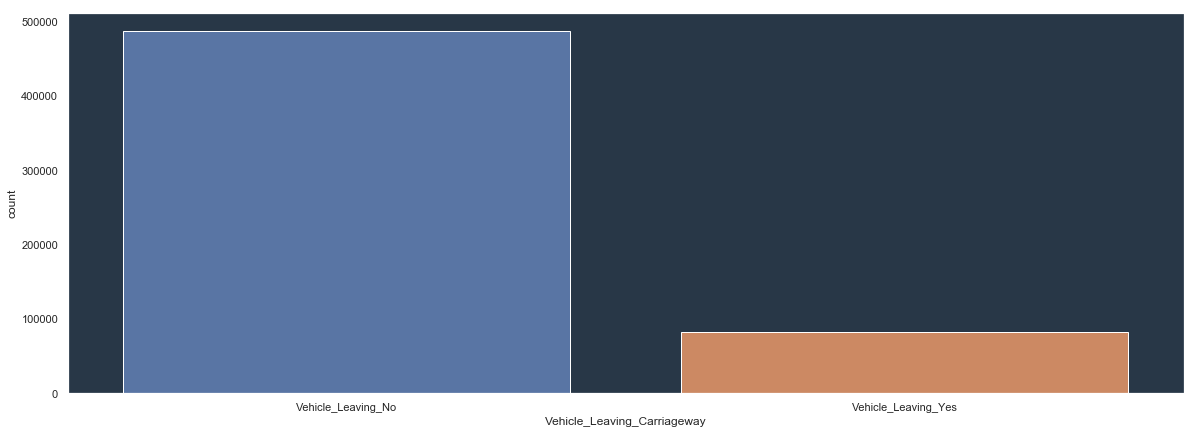
Majority of accidents are with vehicle going ahead other but accidents when turning left are more severe than those were turning right.

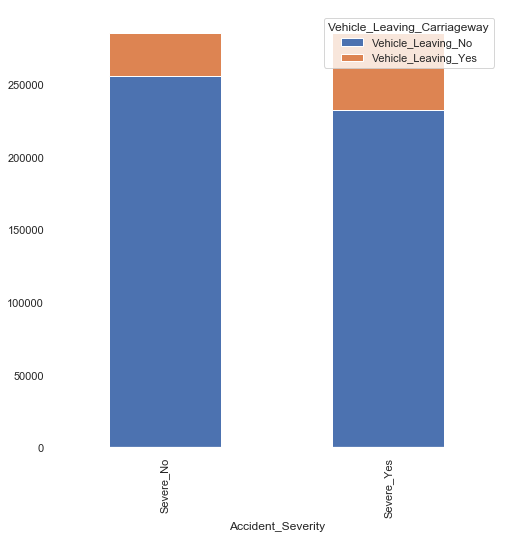
 

Majority of accidents did not hit object in carriage way and we can’t really see big difference in distribution between sever and non severe.

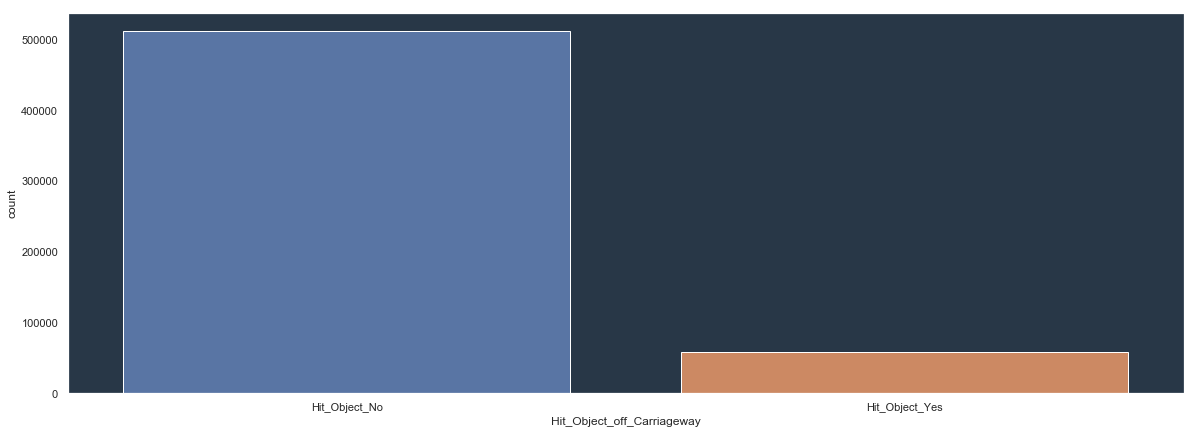
 

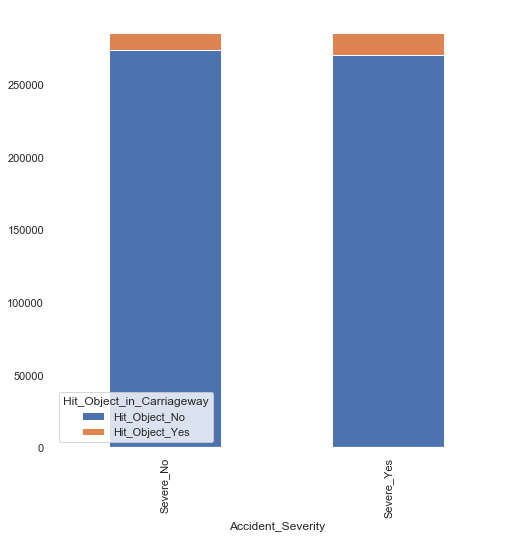
Majority of accidents were when vehicle was not leaving carriageway but more cases when it happened was with severe accidents rather than not.



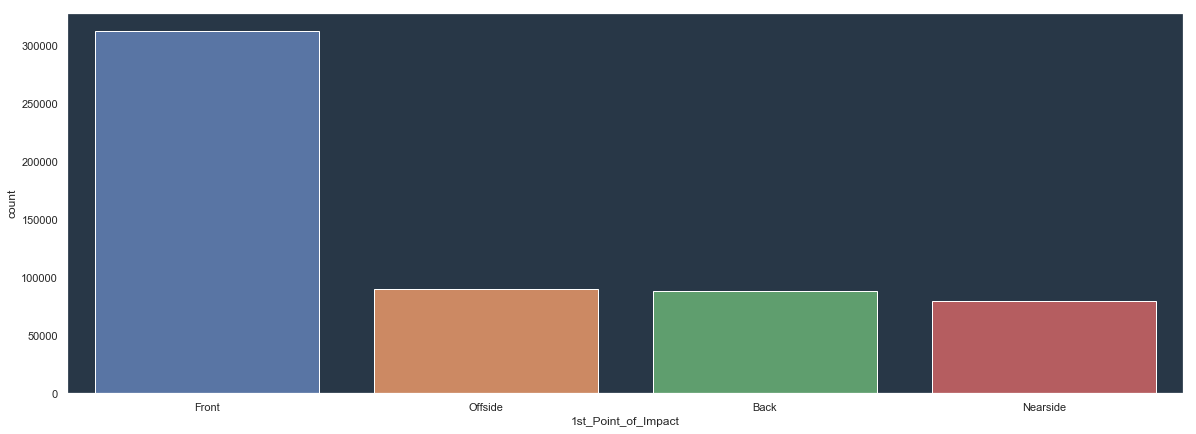


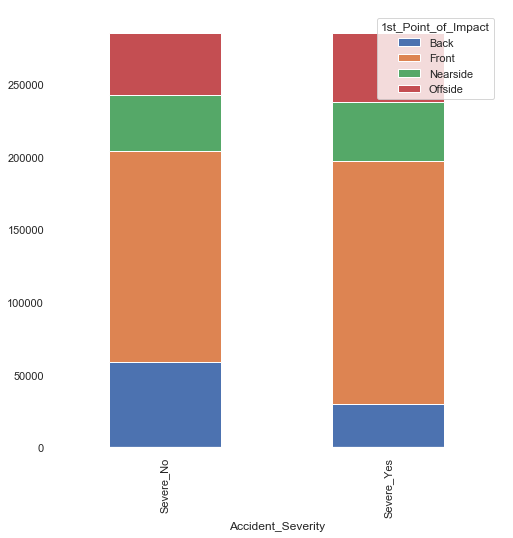
Majority of accidents did not hit objects off carriageway and we do not see big different in distribution between those when the accident was severe or not



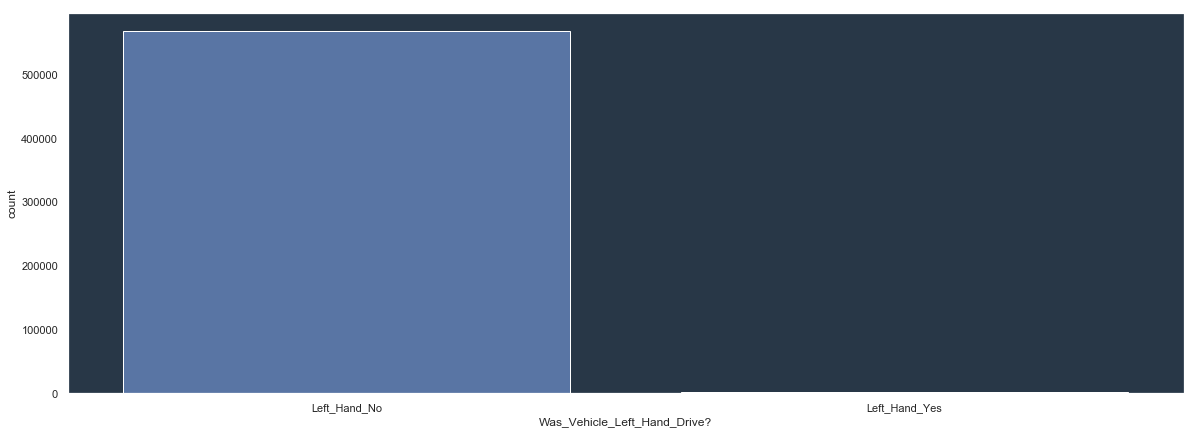


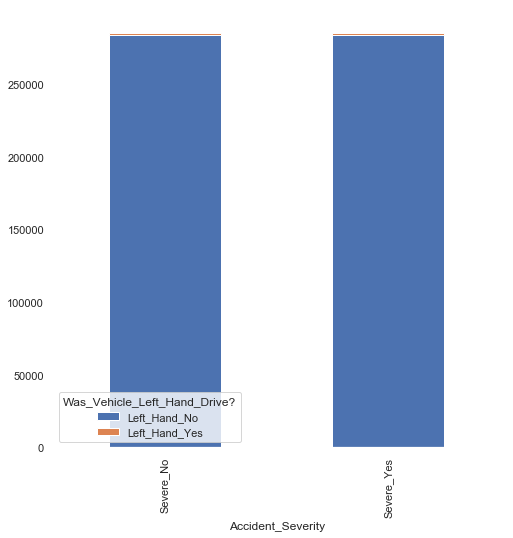
Majority of incidence 1st point of impact was front and more often the accident is severe when it hits in front in comparison to when it hits at the back.



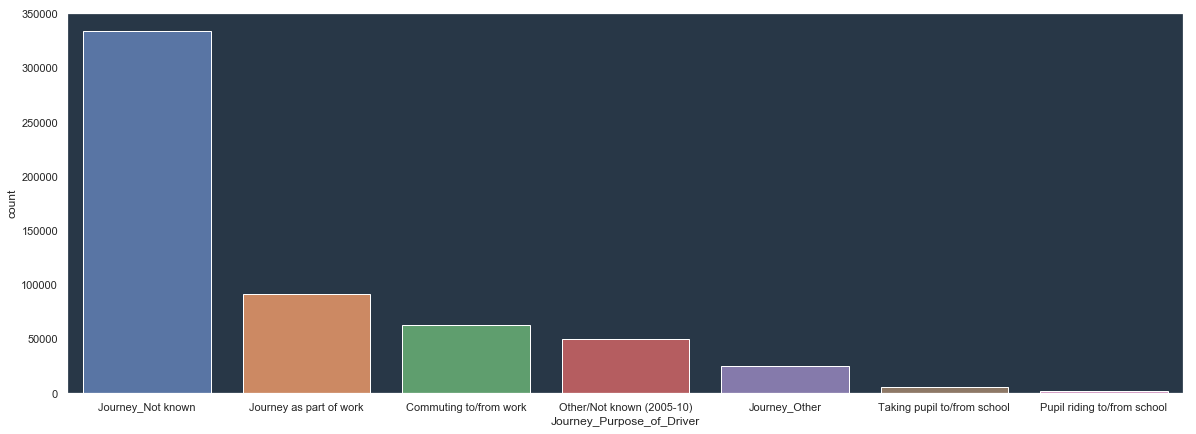


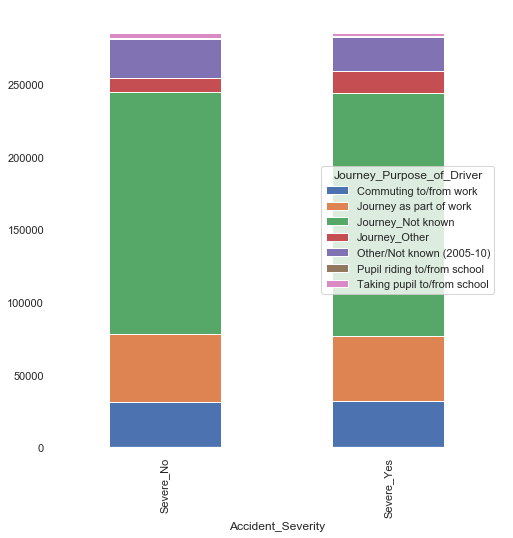
Majority of accidents where not left hand drive and I would remove the variable as the proportion is too big.



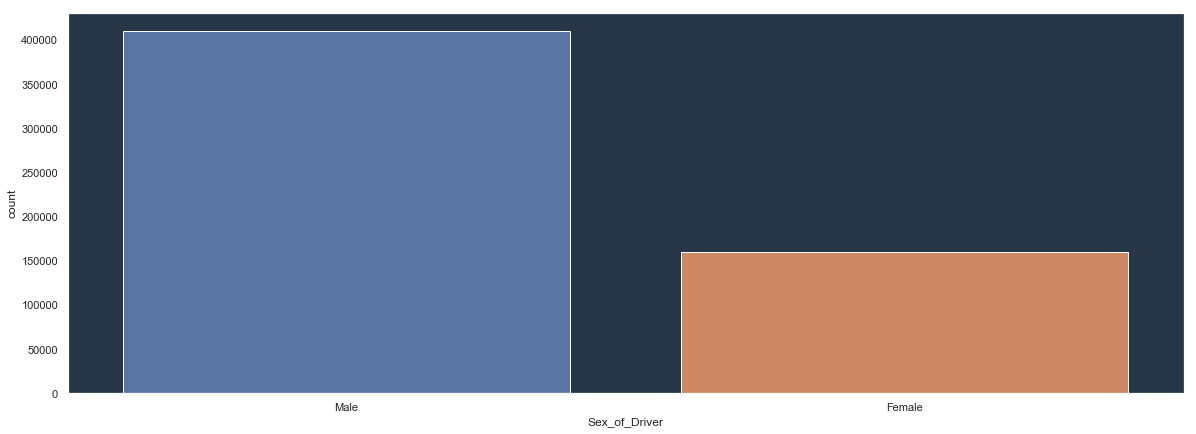


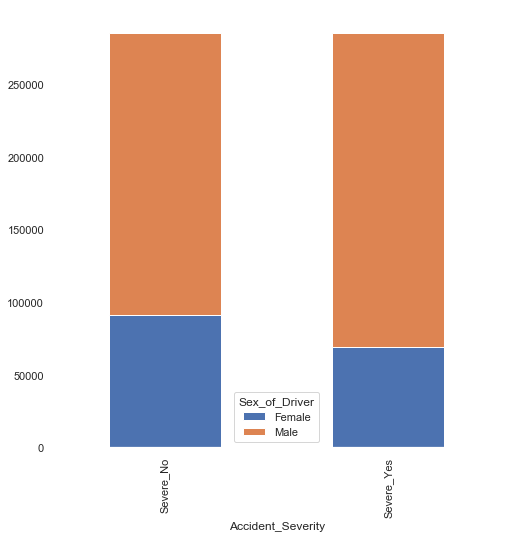
Majority of records are with unknown journey purpose and we don’t really see a difference in distribution



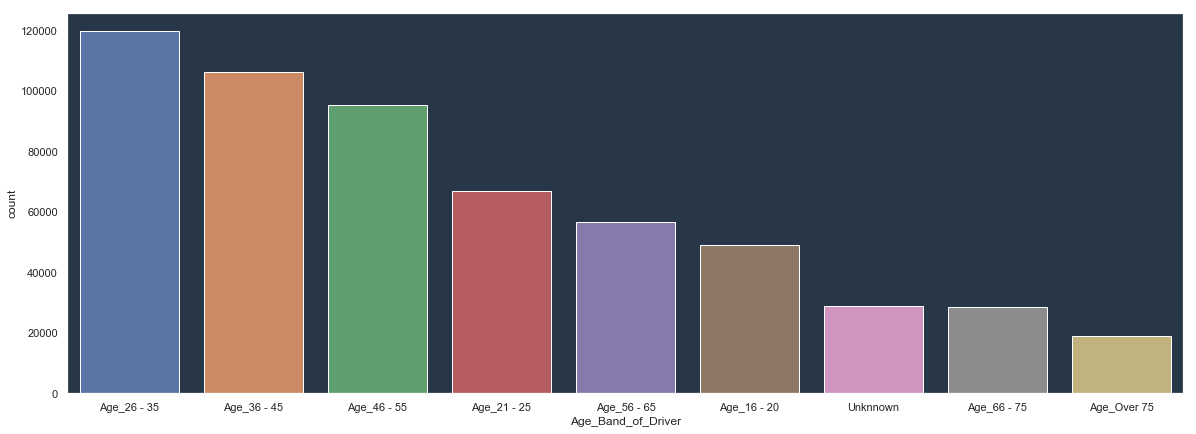


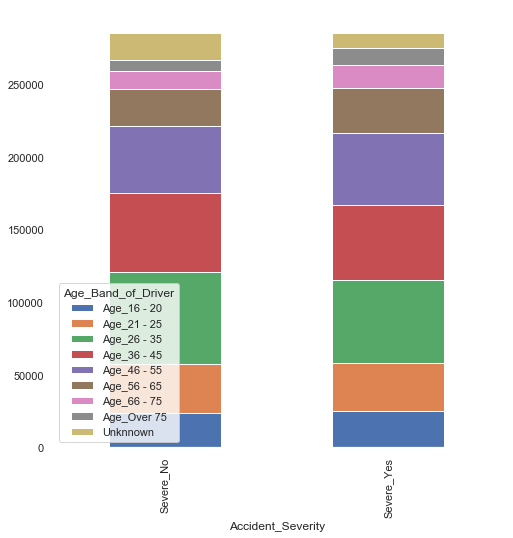
Majority of Accident records are with men and accidents when men is a driver are more severe when it’s women.



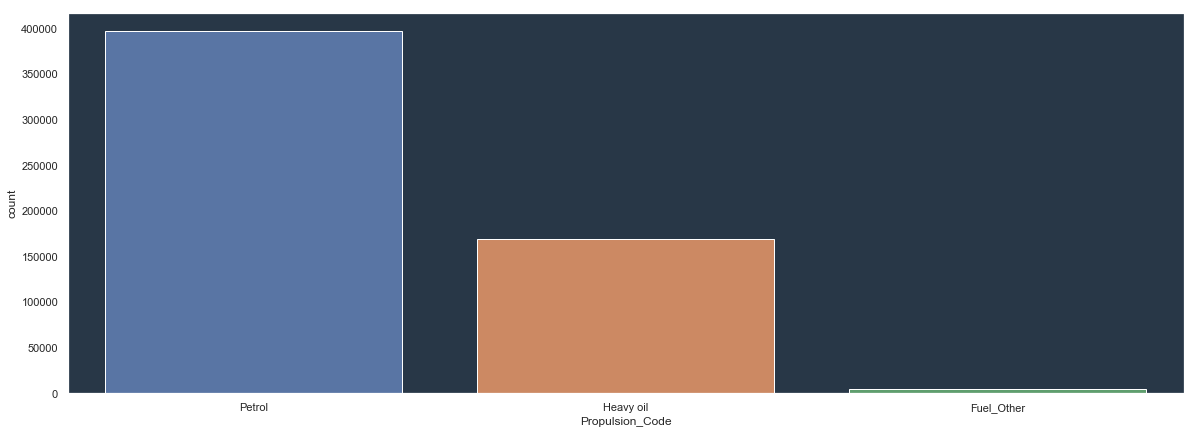


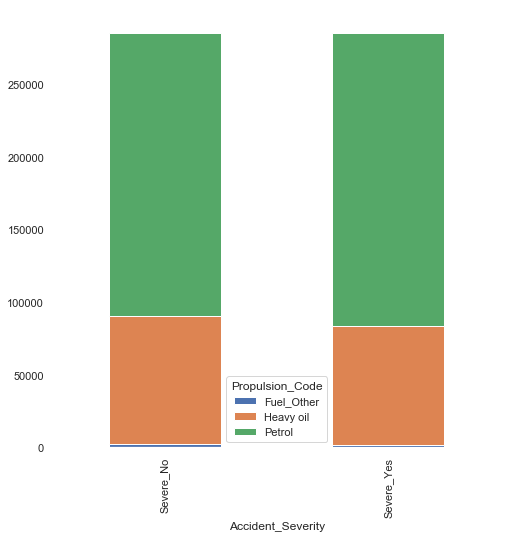
Majority of drivers are in age 26 – 35 but we don’t see big difference in distribution





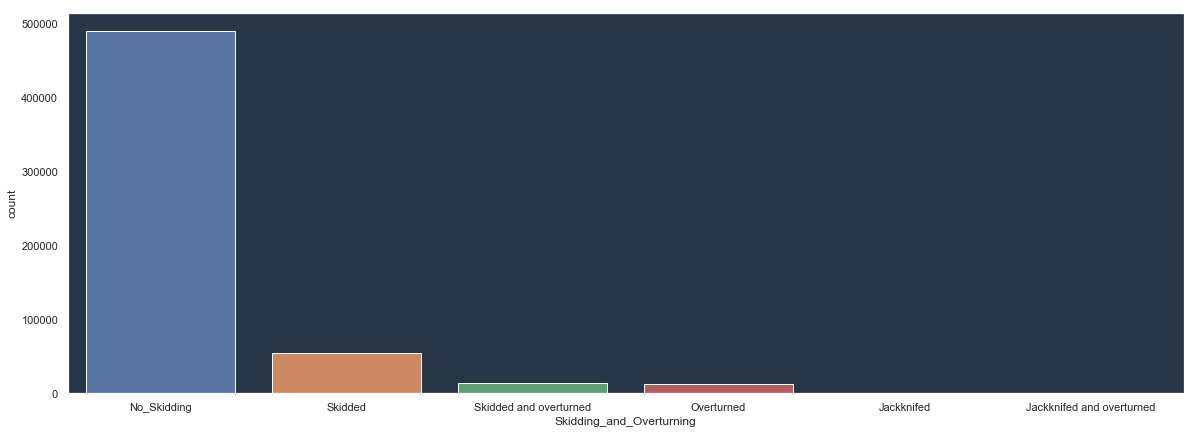
Majority of cars on British Roads are Petrol and we don’t see difference in distribution between severe and non severe accidents for the type of fuel.

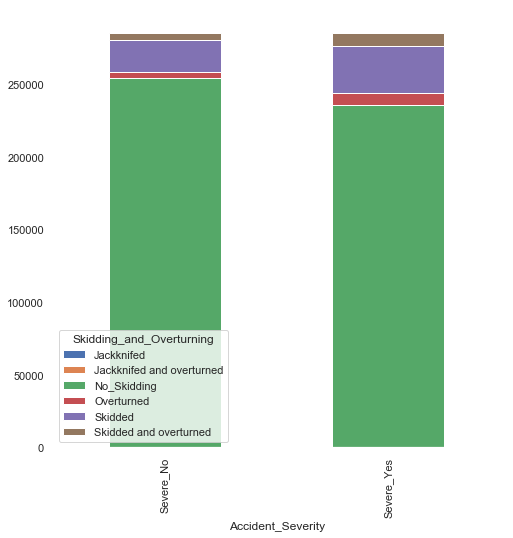




Majority of accidents are with drivers from 1 Home Area type. We don’t have information what does it mean so we remove this variable later.

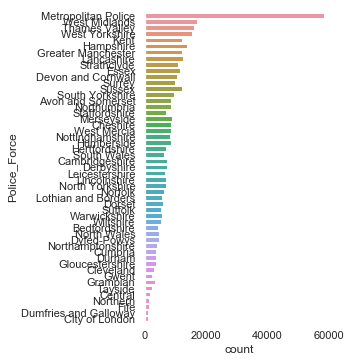
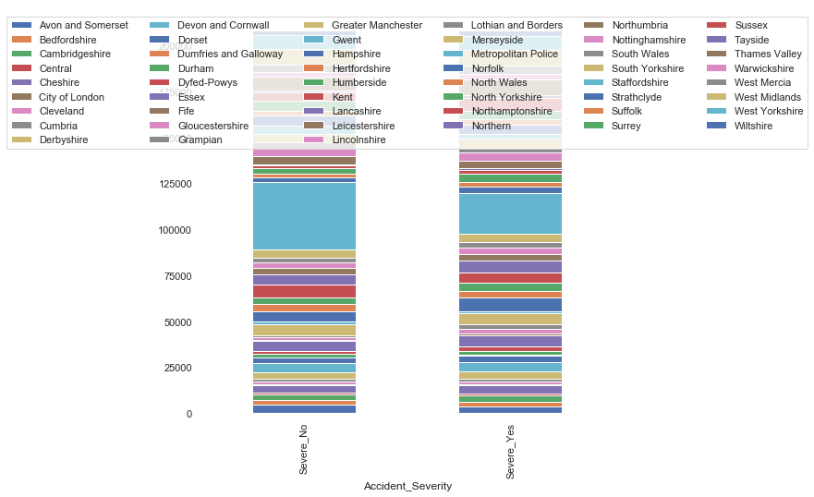
Majority of accidents are when there was no skidding and more severe accidents are when there is skidding rather than not.



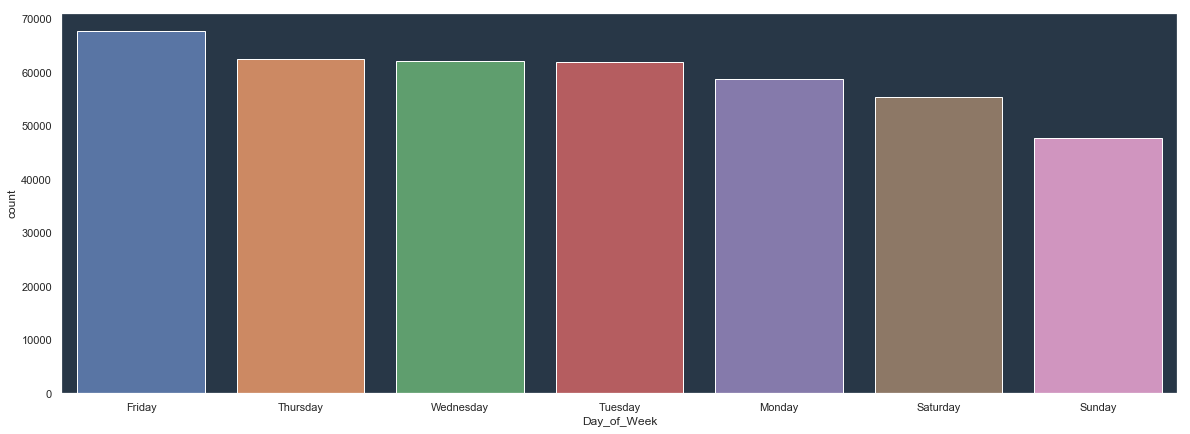


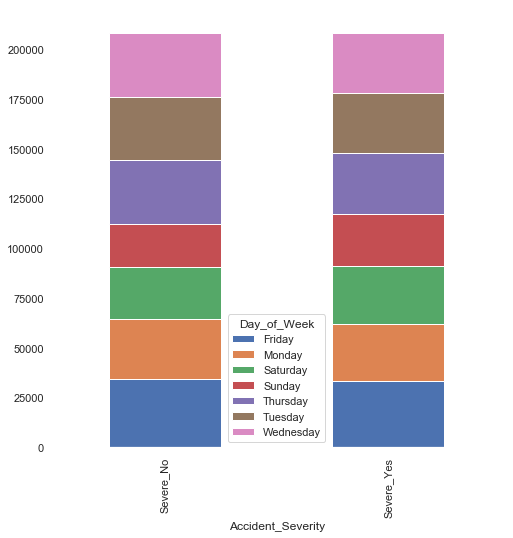
After that data is reduced by taking one observation per accident. The variables unique for individual cars are pivoted so that accident index is a row and different levels of the categorical variable becomes columns for the data. For the columns that are unique for the accident but all vehicles got the same observation e.g. speed limit on the road or weather condition duplicates are removed. So that out of 570,644 observations data is reduced to 474,174 Variable Time is first transformed on categorical: early morning, morning, afternoon, evening, night and then pivoted. Similarly Vehicle Age is transformed to new\_car, average\_car, old\_car, antique\_car. Due to grouping we need to normalized the dataset again so that to have equal number of severe and non severe cases of accidents.

The majority of data comes from Metropolitan Police district but majority of accidents from there are not severe

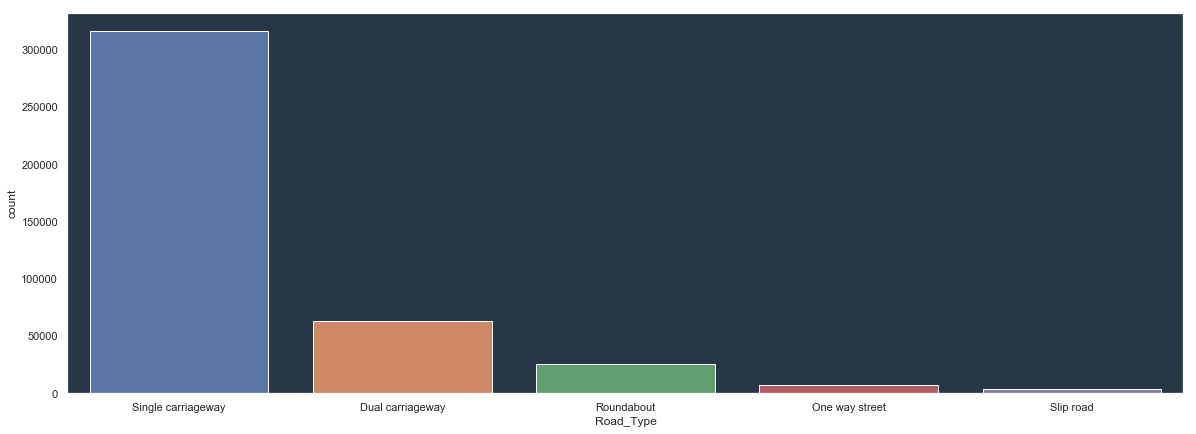
 

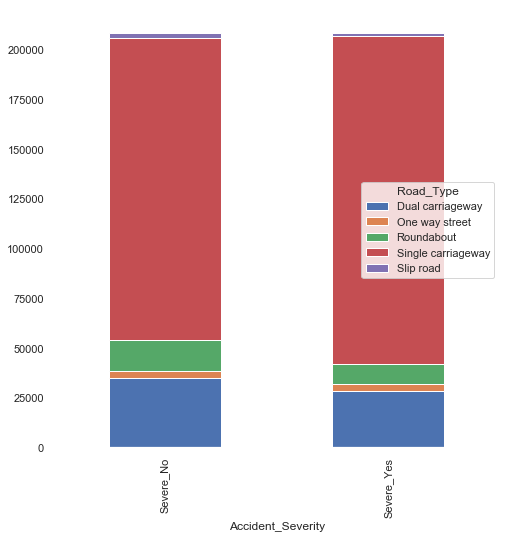
The biggest number of accidents recorded on Friday when Sunday is the safest day to travel



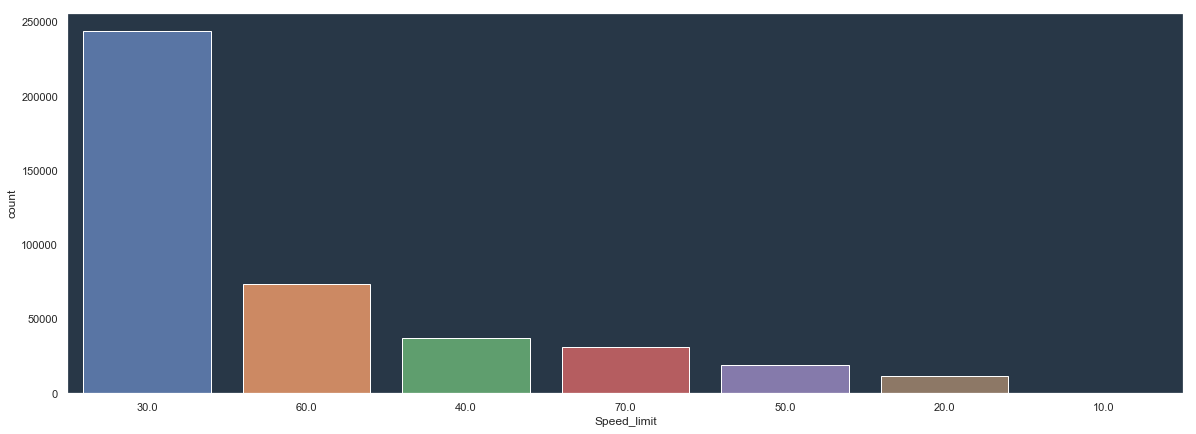


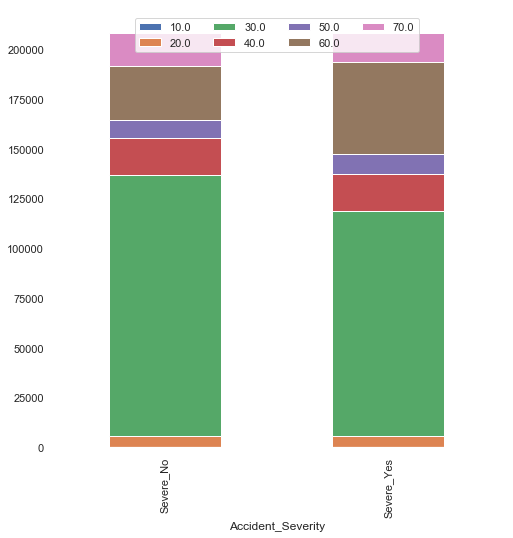
The most popular Road Type in Britain is Single Carriageway. Majority of accidents on roundabouts and dual carriageways are not severe.



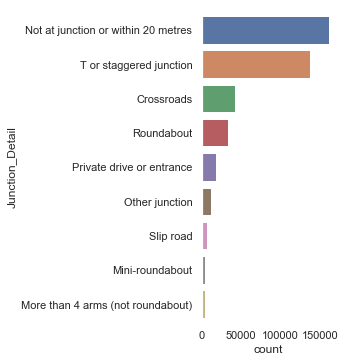


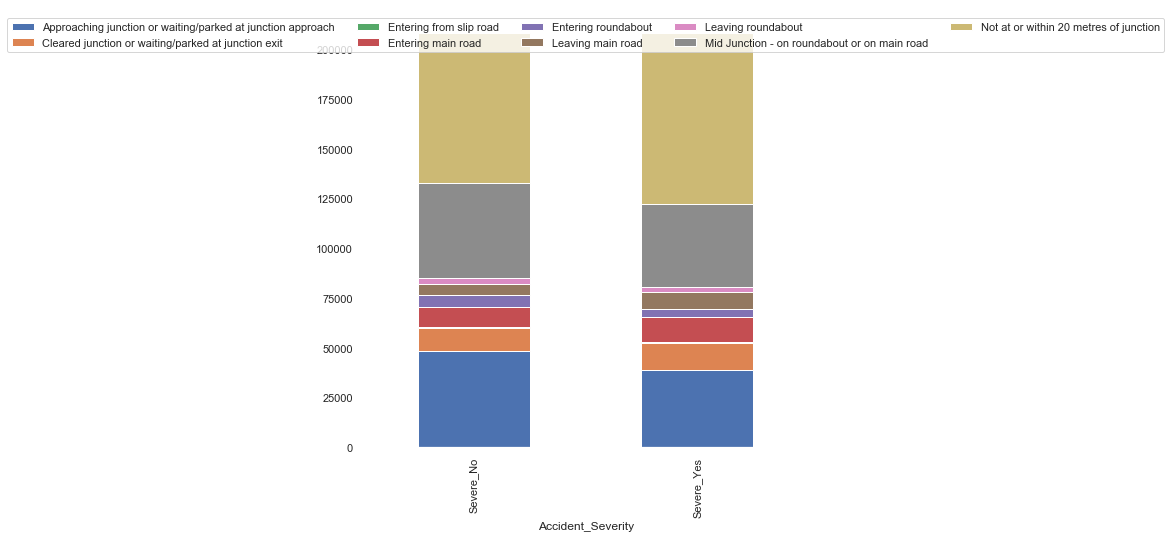
Majority of roads are with speed limit 30 where 2 times more accidents on roads with speed limit 60 is severe than not.



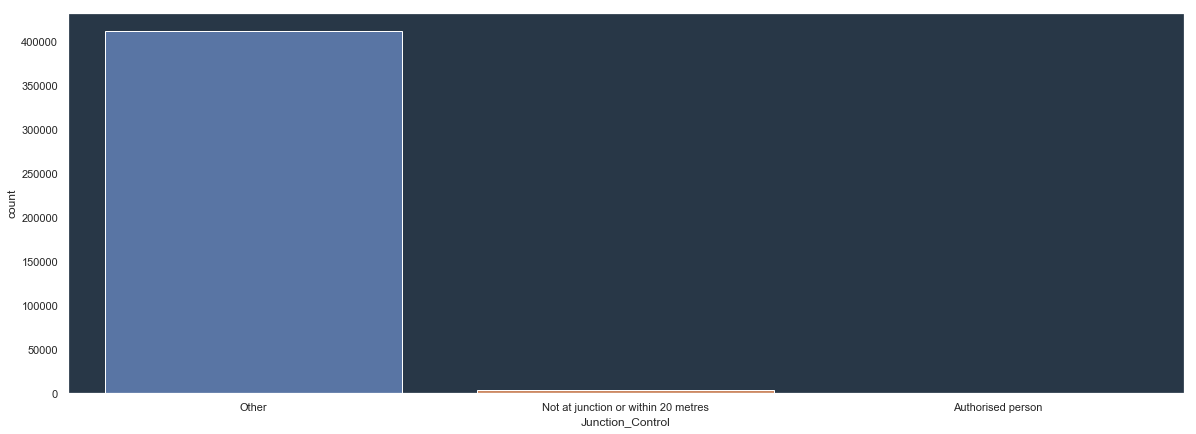


Majority of accidents happened not on junction or 20 miles away from there and on T or staggered junctions. And again 2 times more accidents that happened on roundabouts are not severe.

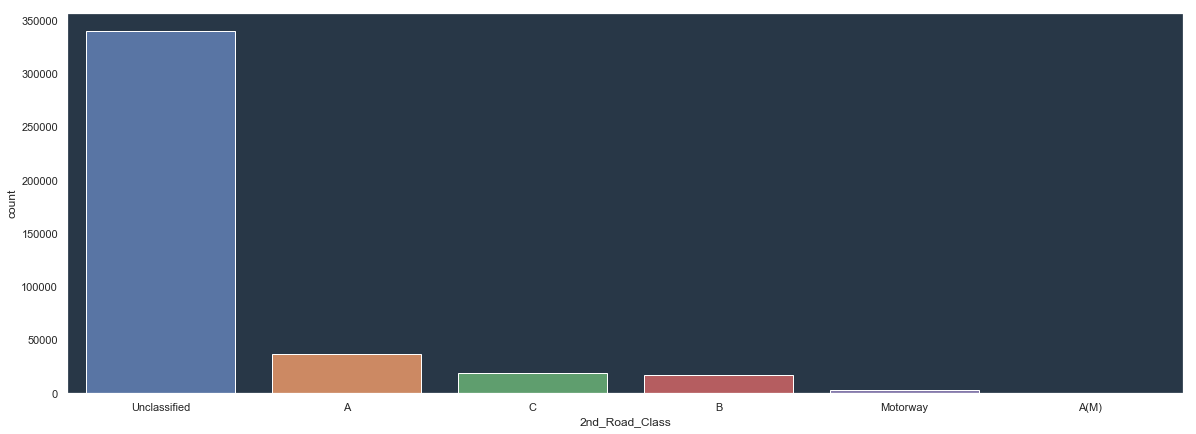


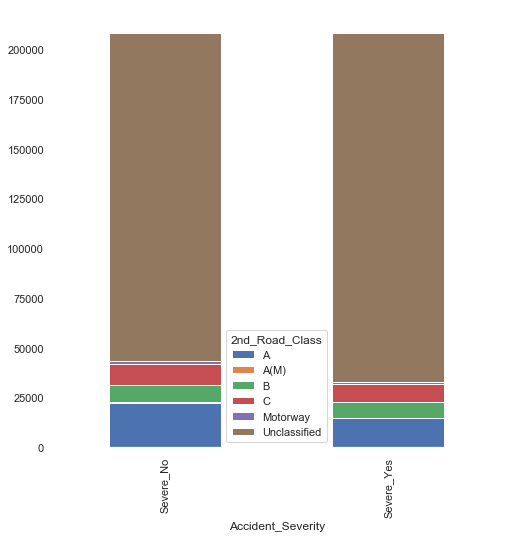


Nearly all cases are classified as Junction Control other. I would recommend removing this variables from the model

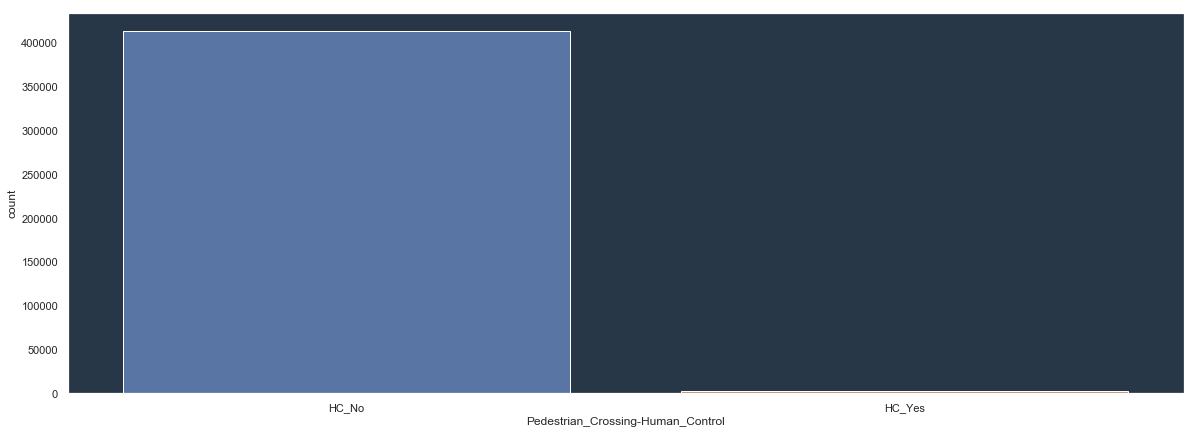


Majority of Roads has 2nd Road Class Unclassified. And 2 times more accidents on raods category A are not severe.

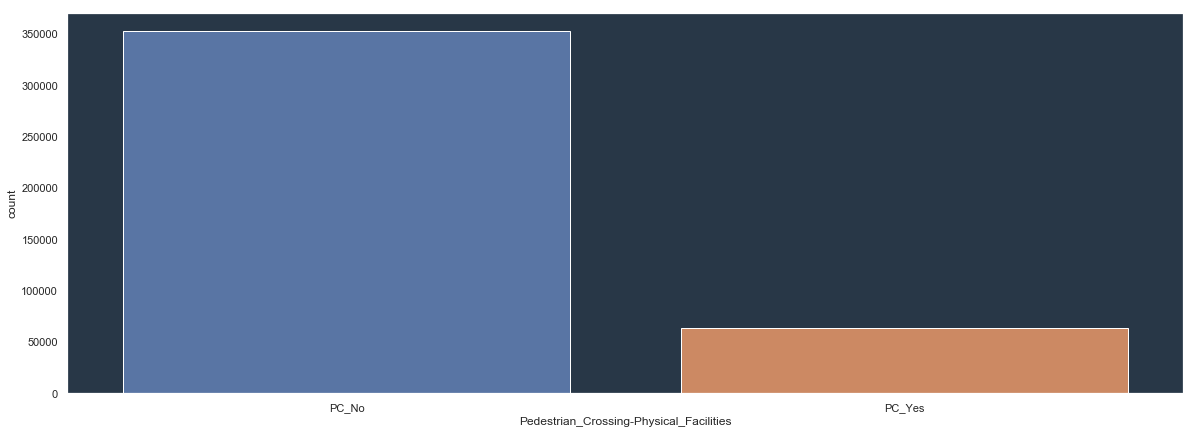


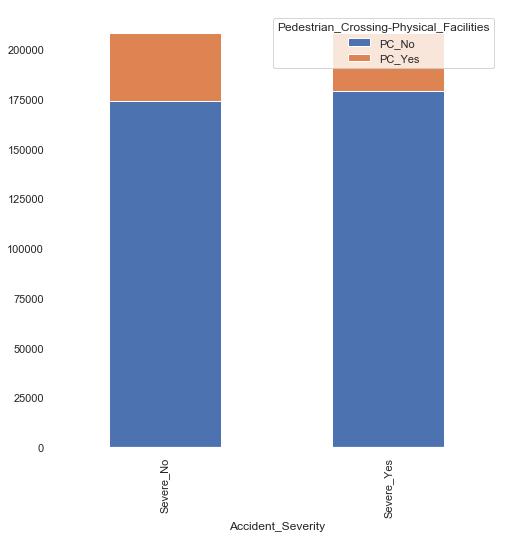


Nearly all of the cases has not Pedestrian\_Crossing-Human\_Control I would recommend to remove this variable

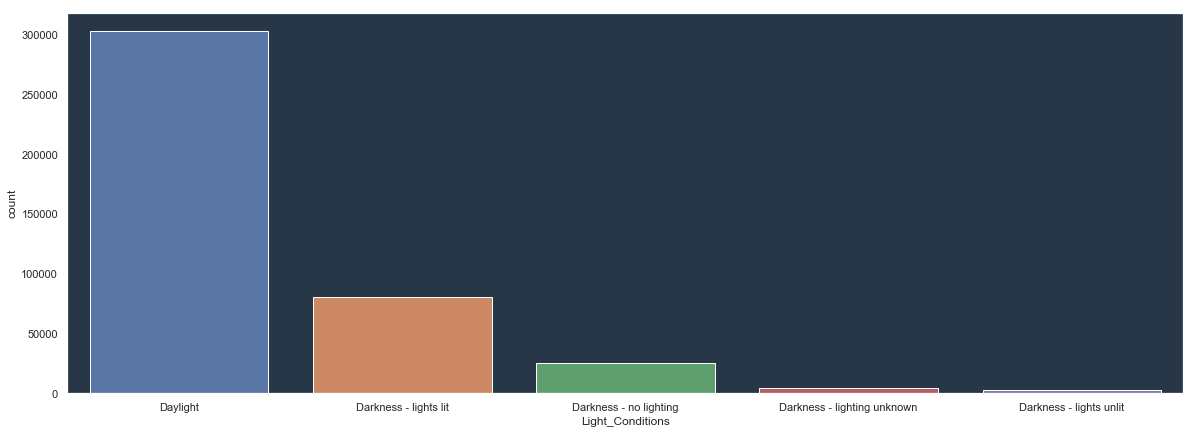


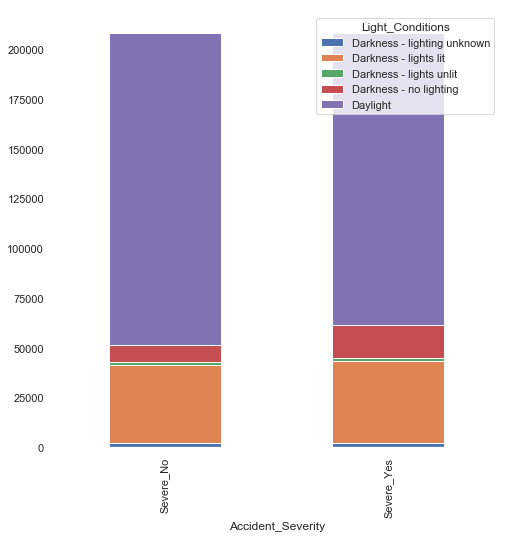
Majority of cases has not Pedestrian\_Crossing-Physical\_Facilities but less accidents that do have it are severe:



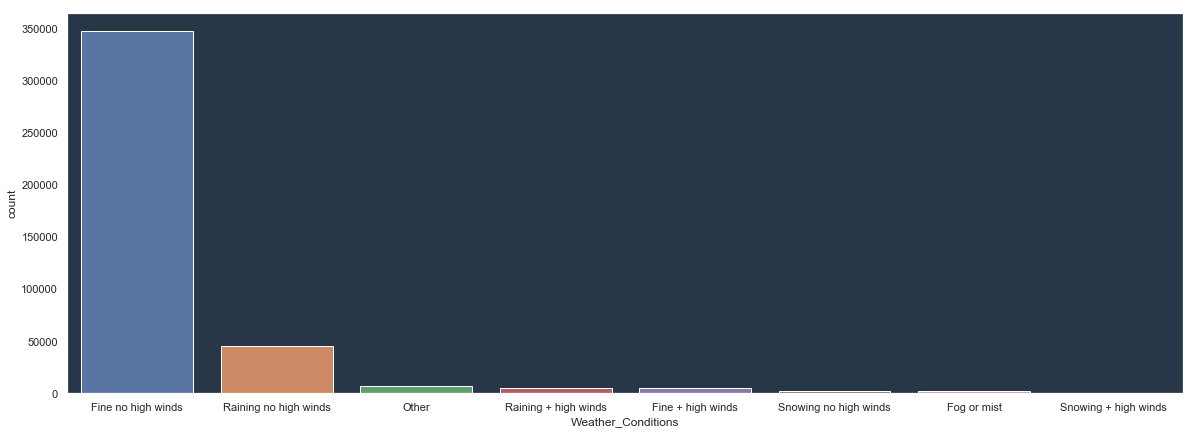


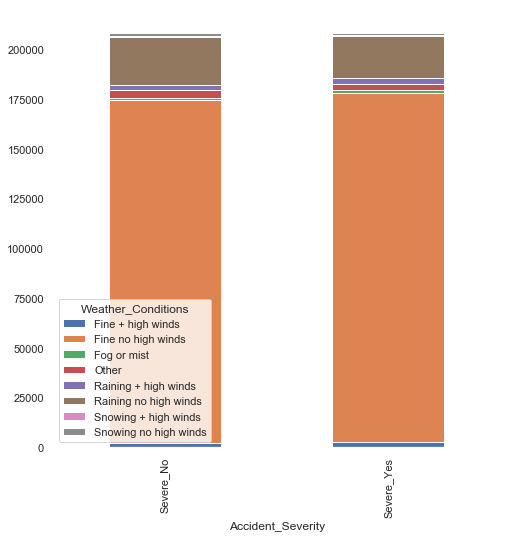
Majority of cases happened when was daylight but two times more accidents where severe when there was darkness – no lightning



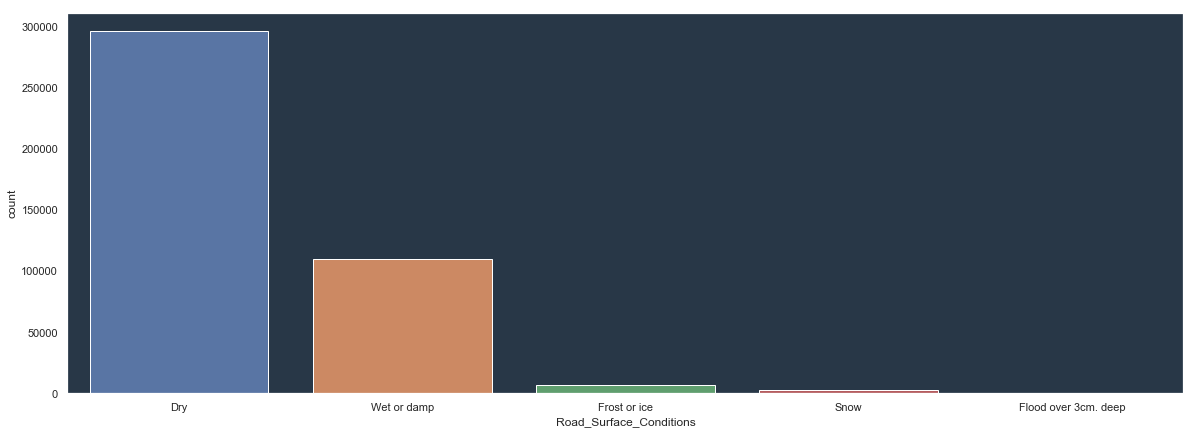


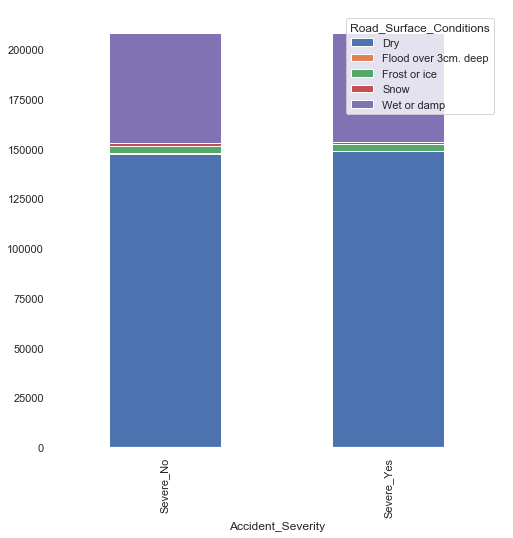
Majority of accidents were when there was good weather with no high winds



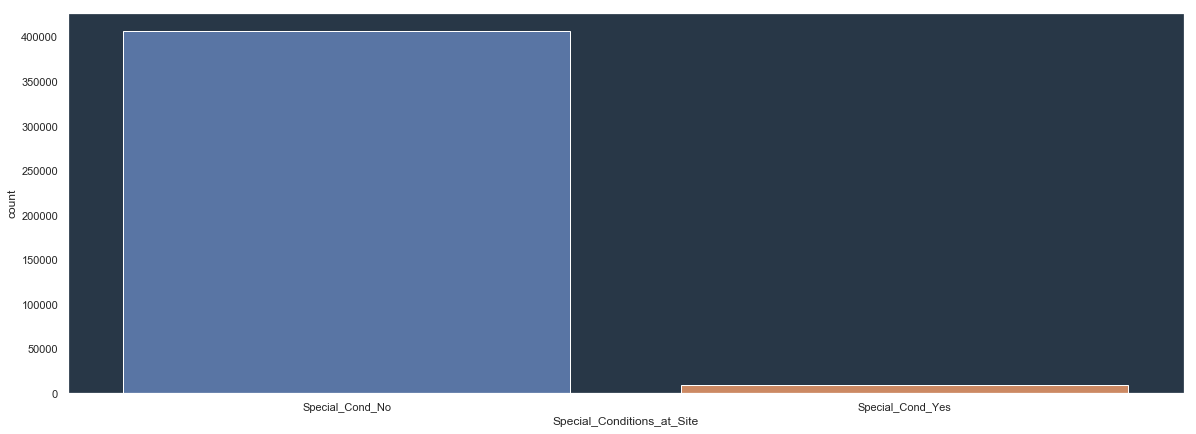


Majority of cases when there was Dry Road surface

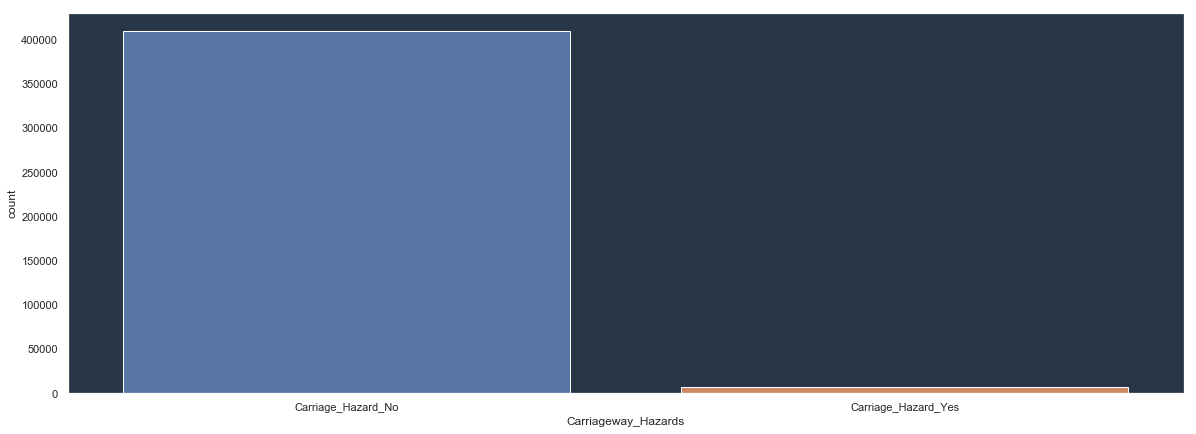




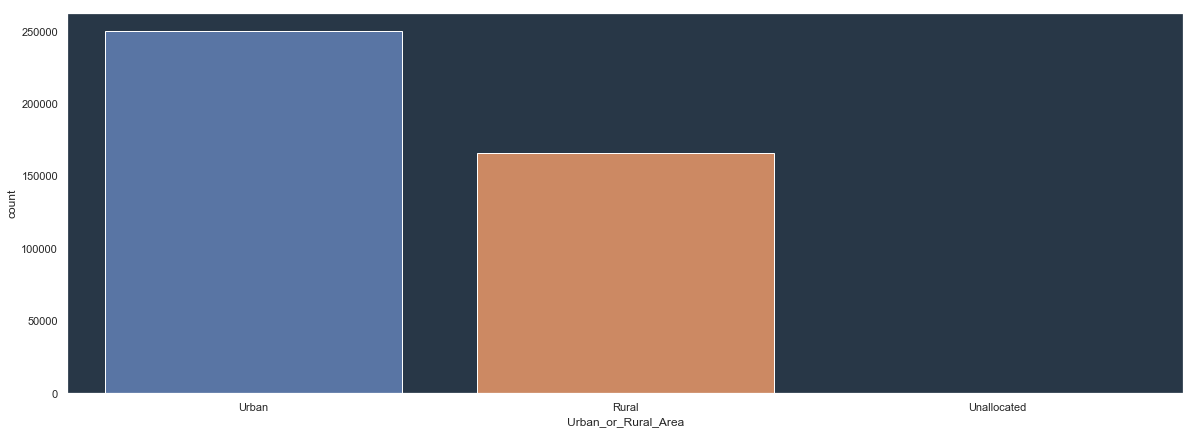
Nearly all accidents didn’t have special control on road side. I would recommend to remove this variable

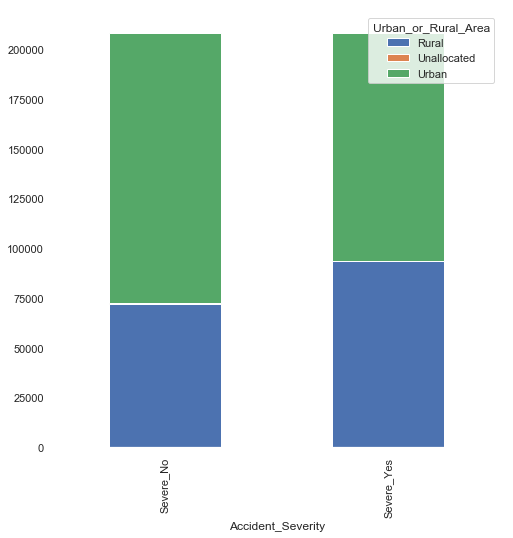


Nearly all cases didn’t have carriage hazards. I would recommend to remove this variable

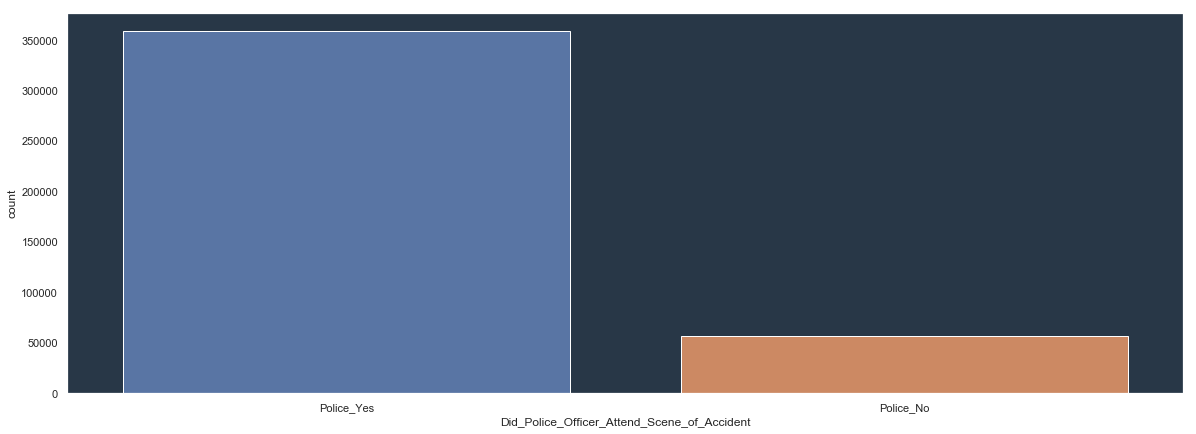


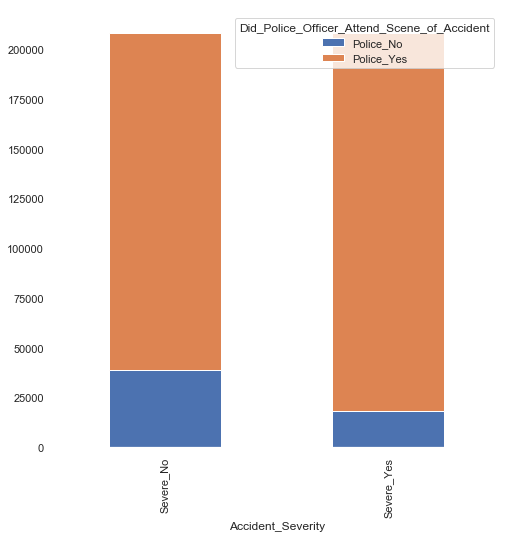
Majority of cases happened on Urban areas but more cases are severe when it happens on rural rather than not.



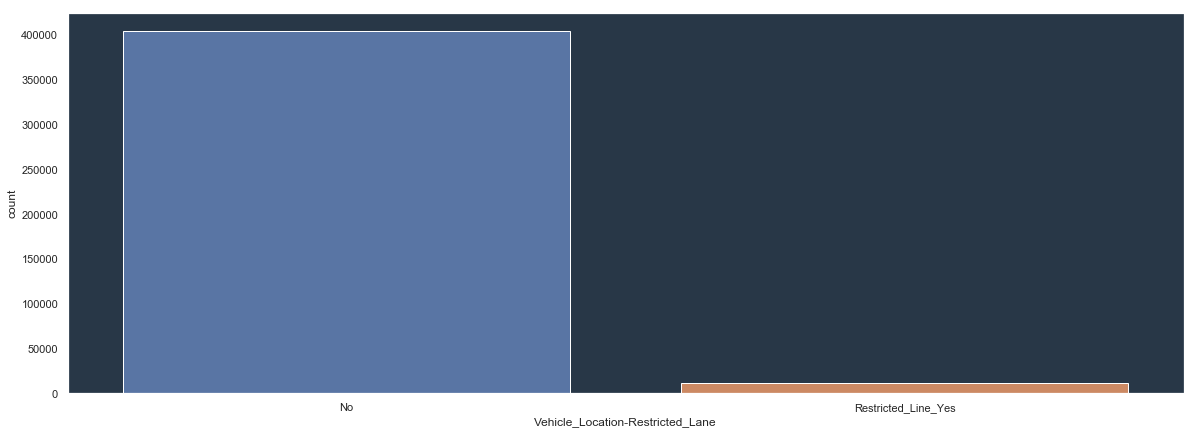


For majority of cases did attend police officer

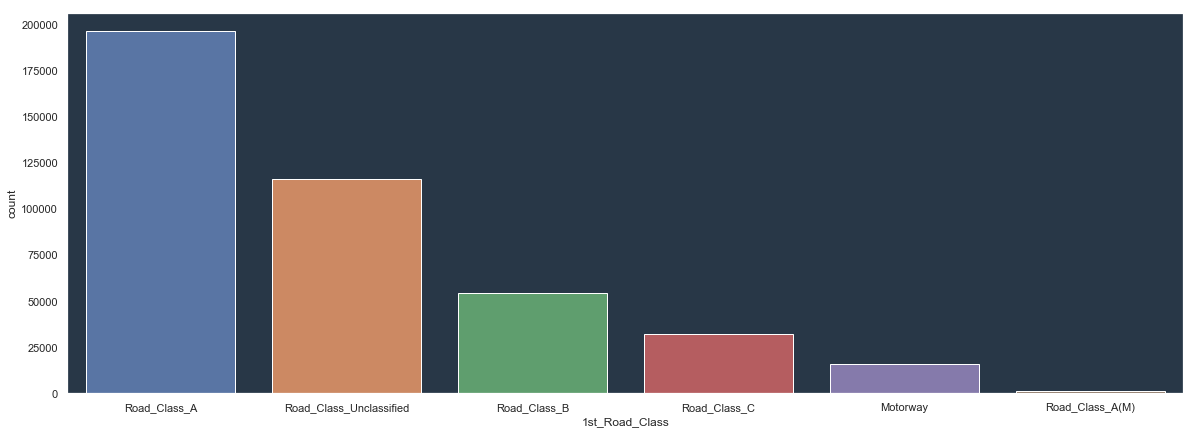


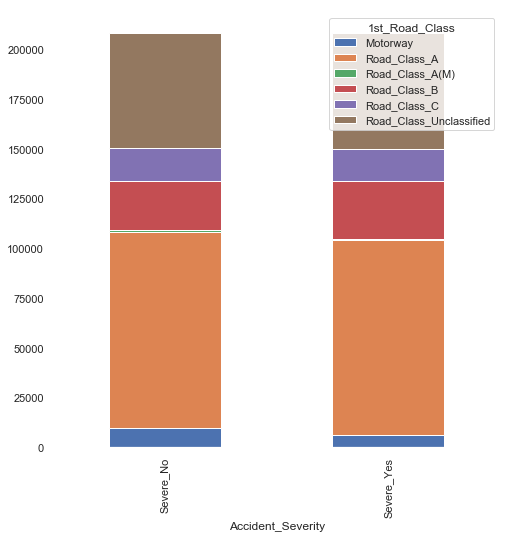


Majority of cases didn’t have restricted line. I would recommend to remove this variable



Majority of cases has 1st road class A



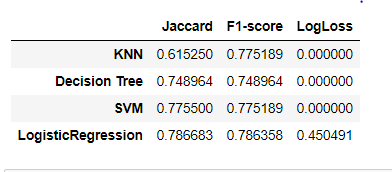


Summary of the most interesting points:

* Majority of accidents are with vehicle going ahead other but accidents when turning left are more severe than those were turning right.
* Majority of accidents were when vehicle was not leaving carriageway but more cases when it happened was with severe accidents rather than not.
* Majority of incidence 1st point of impact was front and more often the accident is severe when it hits in front in comparison to when it hits at the back.
* Majority of Accident records are with men and accidents when men is a driver are more severe when it’s women.
* Majority of accidents are when there was no skidding and more severe accidents are when there is skidding rather than not.
* The majority of data comes from Metropolitan Police district but majority of accidents from there are not severe
* The biggest number of accidents recorded on Friday when Sunday is the safest day to travel
* The most popular Road Type in Britain is Single Carriageway. Majority of accidents on roundabouts and dual carriageways are not severe.
* Majority of roads are with speed limit 30 where 2 times more accidents on roads with speed limit 60 is severe than not.
* Majority of accidents happened not on junction or 20 miles away from there and on T or staggered junctions. And again 2 times more accidents that happened on roundabouts are not severe.
* Majority of Roads has 2nd Road Class Unclassified. And 2 times more accidents on raods category A are not severe.
* Majority of cases happened when was daylight but two times more accidents where severe when there was darkness – no lightning
* Majority of accidents were when there was good weather with no high winds
* Majority of cases when there was Dry Road surface
* Majority of cases happened on Urban areas but more cases are severe when it happens on rural rather than not.

**Modelling:**

In final step I have changed all data on numeric except class labels to compare different classification models. For each of the model I found Jaccard score, F1-score and for logistic regression I found the LogLoss. Table below shows summary of the models:



1. KNN Classifier:

I found out that this method took the longest time to run the model, hence I took just sample of the data. This method compares distance across the observations (similarity) to segregate them into different groups. To better understand how big the K should be – number of groups, I have run KNN classifier for K from 1 to 10. We can see that for K = 7 it has the maximum accuracy and the best accuracy of the classifier is: 0.63

1. Decision Tree

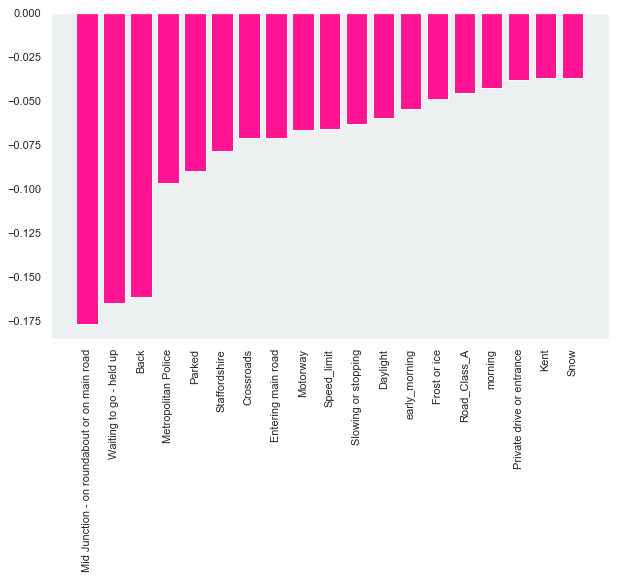
This algorithm worked very fast and I used full data for the model. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.From the table we can see that it had higher Jaccard score than for KNN but it had the lowest F1-score.

1. Support Vector Machine:

This method took very long time to run as well so I used the same sample as for KNN. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. In the table above we can see that it has higher Jaccard and F1-score than KNN. It has accuracy of 78%.

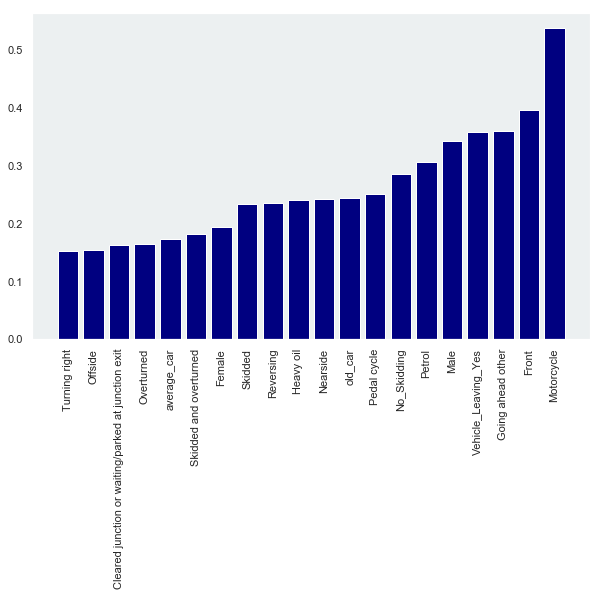
1. Logistic Regression had the highest Jaccard and F1 scores. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary/categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

I have visualised important features in the logistic regression algorithm. The first one represents features that has negative impact on severity, so when those features appear the accidents are less severe.



We have confirmation of a couple of things from the exploratory data analysis.

* When the accident is on roundabout or main road then it’s less severe. Similarly, for crossroads or private drive or entrance.
* Majority f accidents from Metropolitan Police were less severe so this feature is important in building the model. Less severe accidents are also in Staffordshire or Kent.
* When the car waits to go then the accident is less severe. We have seen that with increasing speed the accidents are more severe so that seems like obvious for us. Similarly for parked cars or when the car is stopping or slowing down.
* Less severe accidents are when is daylight and early in the morning, between 5 and 10 am.



This graph shows features that increase severity of accidents.

* As we have seen in EDA stage majority of motorcycle accidents are severe. Cyclists are also in danger or roads but less than motorcycles.
* When car hits at the front is severe. Completely opposite from the fact when the car hits at the back.
* When the driver of vehicle is men the accidents are more severe.
* Accidents with old cars are more severe than with average age cars.
* When cars are going ahead of other vehicle is leaving the junctions accidents are more severe.

Concluded the analytics with Random Forest classifier on a full data and this classifier had the highest accuracy: 81%. It has confirmed the feature selection performed after logistic regression. Random Forest identified features:

'Speed\_limit',

'Car',

'Motorcycle',

'Pedal cycle',

'Going ahead other',

'Turning right',

'Waiting to go - held up',

'No\_Skidding',

'Skidded',

'Back',

'Front',

'Nearside',

'Offside',

'Commuting to/from work',

'Journey as part of work',

'Journey\_Not known',

'Age\_16 - 20',

'Age\_21 - 25',

'Age\_26 - 35',

'Age\_46 - 55',

'Age\_56 - 65', 'Female',

'Male',

'Heavy oil',

'Petrol',

'new\_car',

'old\_car',

'average\_car',

'Engine\_Capacity\_(CC)',

'afternoon',

'early\_morning',

'morning',

'Metropolitan Police',

'Friday',

'Monday',

'Saturday',

'Sunday',

'Thursday',

'Tuesday',

'Wednesday',

'Road\_Class\_A',

'Road\_Class\_B',

'Single carriageway',

'Not at junction or within 20 metres',

'Daylight',

'Fine no high winds',

'Dry',

'Wet or damp',

'Mid Junction - on roundabout or on main road',

'Not at or within 20 metres of junction',

'Vehicle\_Leaving\_Yes'