



# TRAFFIC ANALYSIS AND SEVERITY PREDICTION

Thinkful Final Capstone  
Genesis Taylor

# Background

- The UK government collects and publishes (usually on an annual basis) detailed information about traffic accidents across the country. This information includes, but is not limited to, geographical locations, weather conditions, type of vehicles, number of casualties and vehicle maneuvers, making this a very interesting and comprehensive dataset for analysis and research.
- The data for this project is available on Kaggle as [UK Road Safety: Traffic Accidents and Vehicles](#)

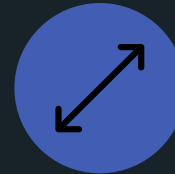
# Importance of Analyzing Traffic Accidents



Accurately analyzing accidents can help governments to better the safety of their roads and highways.



Identifying high areas of accidents and high areas of accident severity can highlight areas of concern.



It can also be beneficial to insurance companies looking to changer their rates in different areas.

# Agenda

The goal of this project is to investigate and determine what causes accidents and what attributes to their level of severity.

Through visualizations and machine learning algorithms, areas of concern will be highlighted, and the seriousness of accidents will be predicted as accurately as possible.

# Data Introduction Details

- The data for this project is obtained from a user on [Kaggle](#) and was composed from information on the [United Kingdom's Government Open Data](#) website.
- It consists of two different datasets that contain information from 2005-2017 that were combined on a common field (accident\_index).
  - Vehicle\_Information.csv: A file containing information about the vehicles, point of impact, maneuvers made, driver information, etc.
  - Accident\_Information.csv: A file containing details about the accident that include location, junction details, date and time, accident severity, etc.

# Research Questions

Who does this project benefit?

When do/did the most accidents happen?

How do the available factors contribute to accident seriousness?

Can we create a machine learning algorithm that correctly predicts the severity of accidents?

Can we forecast the number of accidents in upcoming years based on the information available?

What are the limitations of the current data?

What things would help this research to be more accurate?

# Who does this project benefit?



## Government Departments of Transportation

Create safer highways

Prevent fatalities

Reduce severity

Educate the public

# When do/did the most accidents happen?



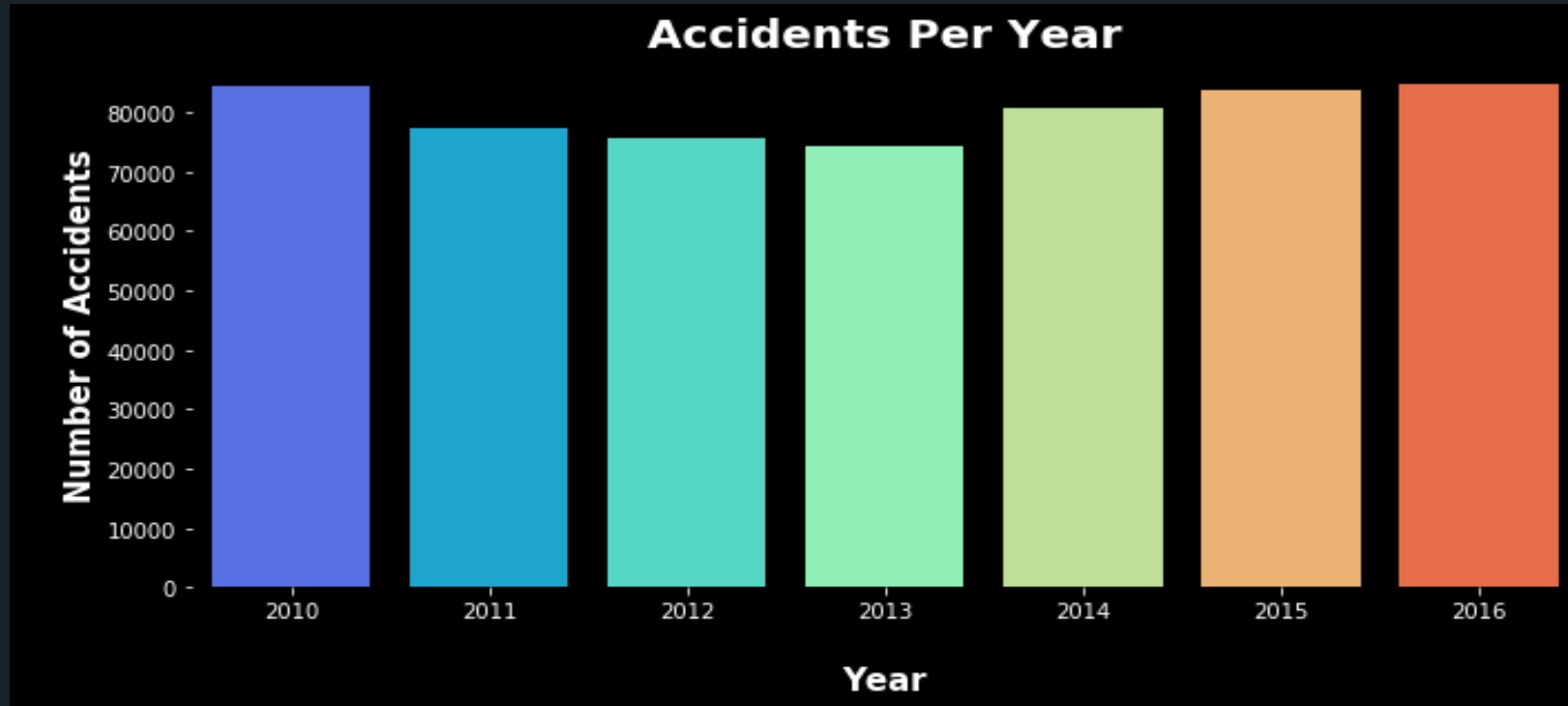
THE FOLLOWING CHARTS DISPLAY THE  
NUMBER OF ACCIDENTS BY YEAR,  
MONTH, SEASON AND DAY OF THE WEEK.



THE TRENDS FOR EACH TIME PERIOD  
ARE DISPLAYED FOR ANALYSIS AND  
REVIEWS.

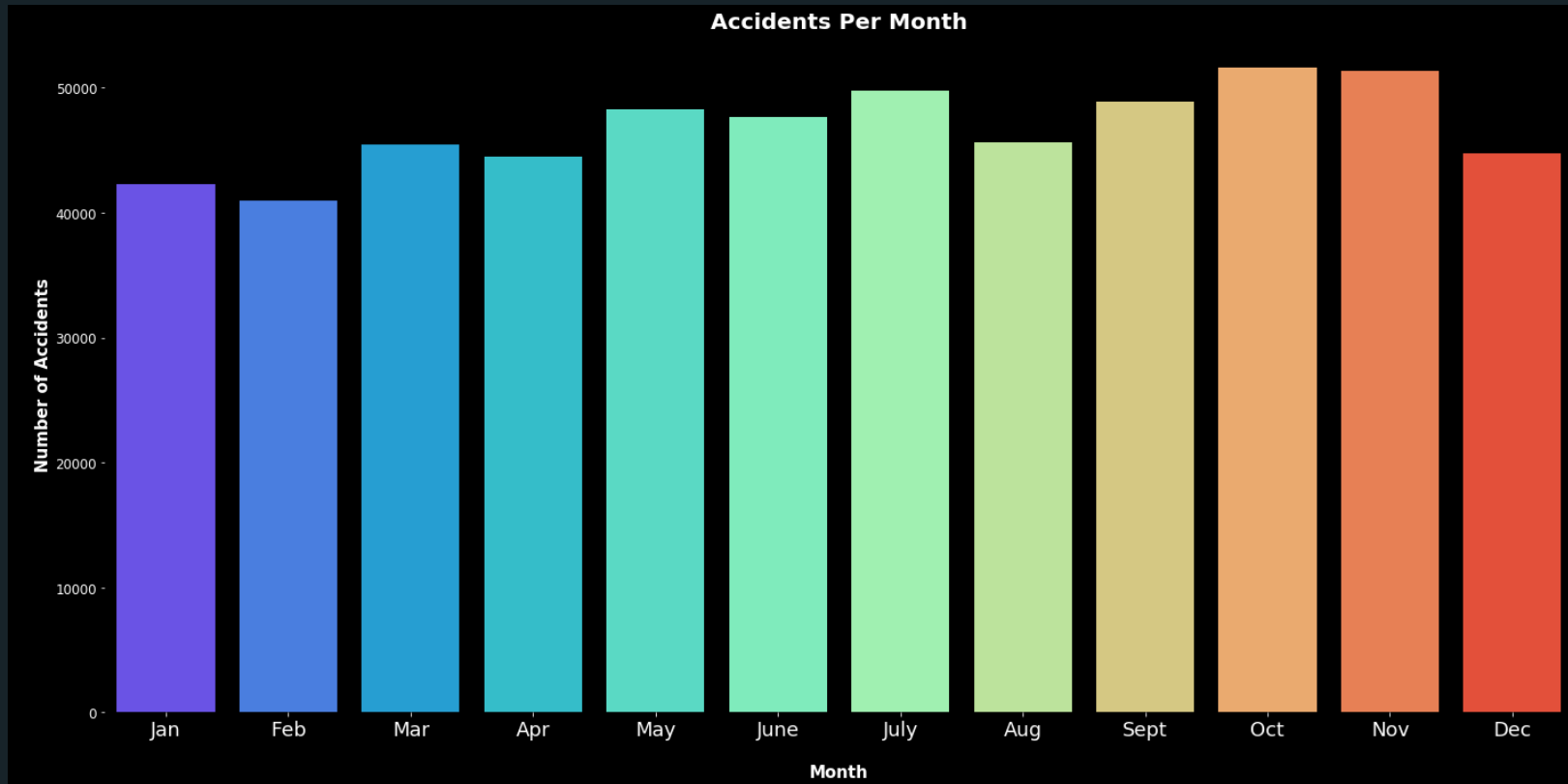


# Accidents per Year



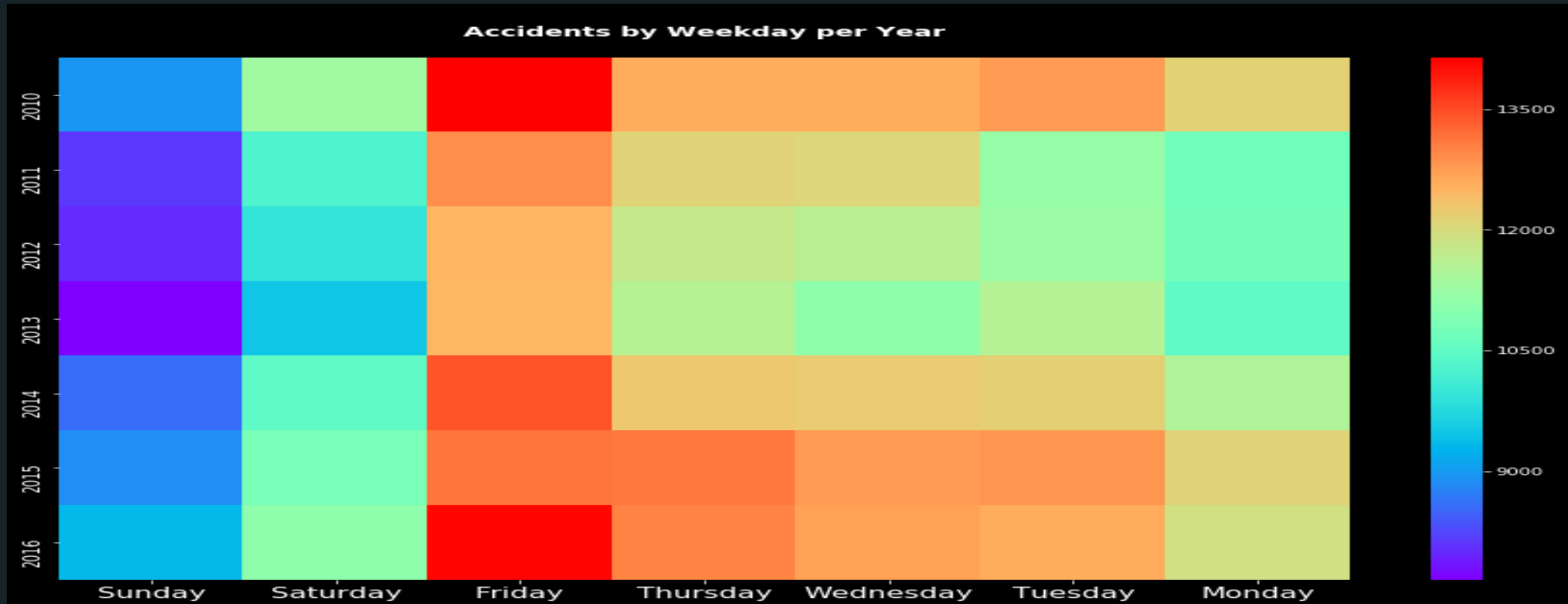
According to the graph above, from 2010 to 2013, there was a decrease in the number of accidents, however over the past few years there has been an increase in accidents, with 2016 being close to the 2010 numbers.

# Accidents per Month



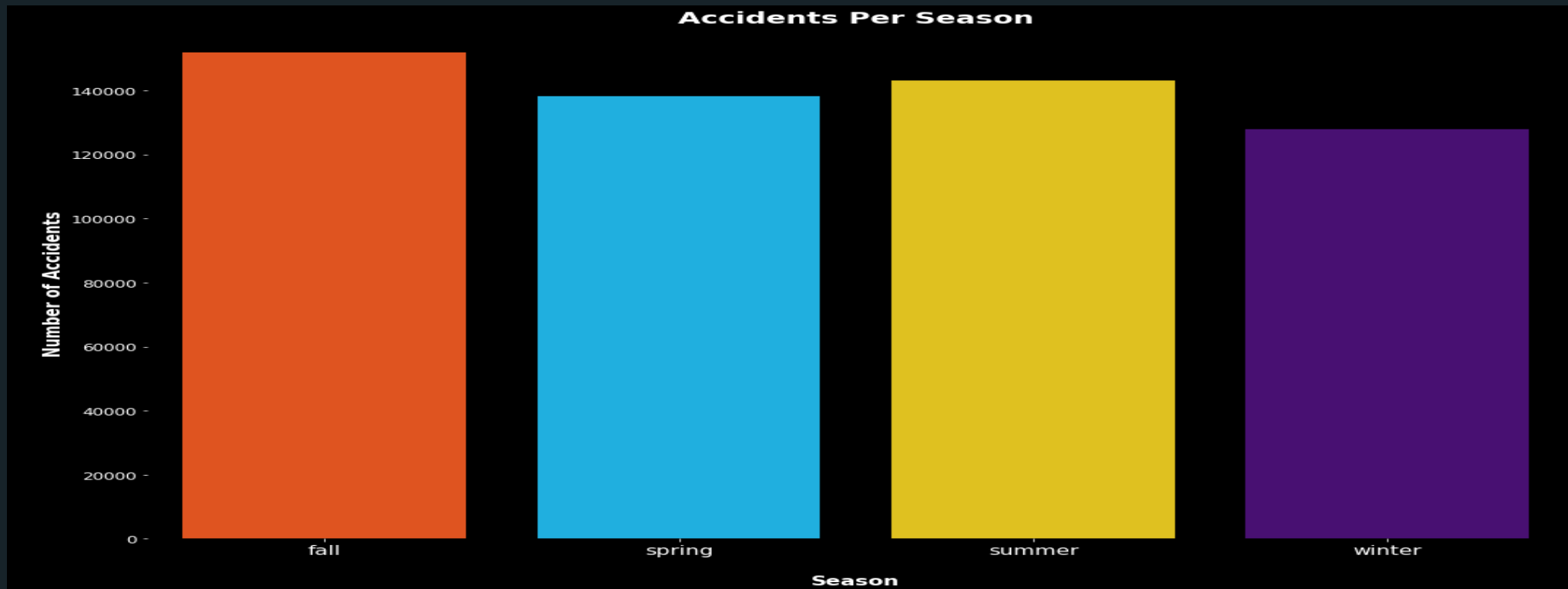
According to the graph above, the majority of accidents happen between May and July and September and November.

# Accidents Per Weekday Per Year



Fridays are the day of the week where the most accidents occur in each year. More information can help to explain this occurrence better.

# Accidents Per Season



Summer and Fall have the highest number of accidents. These match with the accidents per month and are something for governments to look at closer and compare to whatever events are in the area.

# How do the available factors contribute to accident seriousness?

did_police_officer_attend_scene_of_accident	x1st_point_of_impact	number_of_vehicles	speed_limit	urban_or_rural_area	skidding_and_overturning
vehicle_leaving_carriageway	sex_of_driver	vehicle_type	vehicle_manoeuvre	engine_capacity_cc	number_of_casualties
driver_home_area_type	age_band_of_driver	junction_control	hit_object_off_carriageway	hit_object_in_carriageway	driver_imd_decile
	junction_detail	junction_location	propulsion_code	year	

These features were found to have the highest relation to accident seriousness. While they may all have an impact, not every feature will be discussed in the following slides. The features discussed will be based on the findings from their visualization. For visual reasons, two separate dataframes were created, for not serious and serious accidents. I wanted to better scale the data and for me, this was the simplest way of doing so.

# How do the available factors contribute to accident seriousness?

```
class ChiSquare:
    def __init__(self, dataframe):
        self.df = dataframe
        self.p = None #P-Value
        self.chi2 = None #Chi Test Statistic
        self.dof = None

        self.dfObserved = None
        self.dfExpected = None

    def _print_chisquare_result(self, colX, alpha):
        result = ""
        if self.p < alpha:
            result += "The column {} is IMPORTANT for Prediction".format(colX)
        else:
            result += "The column {} is NOT an important predictor. (Discard {} from model)".format(colX)

        print(result)

    def TestIndependence(self, colX, colY, alpha=0.05):
        X = self.df[colX].astype(str)
        Y = self.df[colY].astype(str)

        self.dfObserved = pd.crosstab(Y, X)
        chi2, p, dof, expected = stats.chi2_contingency(self.dfObserved.values)
        self.p = p
        self.chi2 = chi2
        self.dof = dof

        self.dfExpected = pd.DataFrame(expected, columns=self.dfObserved.columns,
                                       index = self.dfObserved.index)

        self._print_chisquare_result(colX, alpha)

#Initialize ChiSquare Class
c1 = ChiSquare(df)
```

## Feature Selection

```
testColumns = ['accident_index', '1st_road_class', '1st_road_number', '2nd_road_number',
               'carriageway_hazards', 'date', 'day_of_week',
               'did_police_officer_attend_scene_of_accident', 'junction_control',
               'junction_detail', 'latitude', 'light_conditions', 'local_authority_district',
               'local_authority_highway', 'longitude', 'lsoa_of_accident_location',
               'number_of_casualties', 'number_of_vehicles', 'pedestrian_crossing-human_control',
               'pedestrian_crossing-physical_facilities', 'police_force', 'road_surface_conditions',
               'road_type', 'special_conditions_at_site', 'speed_limit', 'time',
               'urban_or_rural_area', 'weather_conditions', 'year', 'wscotland',
               'age_band_of_driver', 'age_of_vehicle', 'driver_home_area_type',
               'driver_ind_decile', 'engine_capacity_cc', 'hit_object_in_carriageway',
               'hit_object_off_carriageway', 'journey_purpose_of_driver', 'junction_location',
               'make', 'model', 'propulsion_code', 'sex_of_driver', 'skidding_and_overturning',
               'towing_and_articulation', 'vehicle_leaving_carriageway',
               'vehicle_locationrestricted_lane', 'vehicle_manoeuvre', 'vehicle_reference',
               'vehicle_type', 'was_vehicle_left_hand_drive', 'x1st_point_of_impact', 'month',
               'weekend', 'hour', 'time_of_day', 'season', 'engine_capacity_cc_size']
```

for var in testColumns:

```
c1.TestIndependence(colX=var, colY='accident_seriousness')
```

The column accident\_index is IMPORTANT for Prediction  
The column 1st\_road\_class is IMPORTANT for Prediction  
The column 1st\_road\_number is IMPORTANT for Prediction  
The column 2nd\_road\_number is IMPORTANT for Prediction  
The column carriageway\_hazards is IMPORTANT for Prediction  
The column date is IMPORTANT for Prediction  
The column day\_of\_week is IMPORTANT for Prediction  
The column did\_police\_officer\_attend\_scene\_of\_accident is IMPORTANT for Prediction  
The column junction\_control is IMPORTANT for Prediction  
The column junction\_detail is IMPORTANT for Prediction  
The column latitude is IMPORTANT for Prediction  
The column light\_conditions is IMPORTANT for Prediction  
The column local\_authority\_district is IMPORTANT for Prediction  
The column local\_authority\_highway is IMPORTANT for Prediction  
The column longitude is IMPORTANT for Prediction  
The column lsoa\_of\_accident\_location is IMPORTANT for Prediction  
The column number\_of\_casualties is IMPORTANT for Prediction  
The column number\_of\_vehicles is IMPORTANT for Prediction  
The column pedestrian\_crossing-human\_control is IMPORTANT for Prediction  
The column pedestrian\_crossing-physical\_facilities is IMPORTANT for Prediction  
The column police\_force is IMPORTANT for Prediction  
The column road\_surface\_conditions is IMPORTANT for Prediction  
The column speed\_limit is IMPORTANT for Prediction  
The column time is IMPORTANT for Prediction  
The column urban\_or\_rural\_area is IMPORTANT for Prediction  
The column weather\_conditions is IMPORTANT for Prediction  
The column year is IMPORTANT for Prediction  
The column wscotland is IMPORTANT for Prediction  
The column age\_band\_of\_driver is IMPORTANT for Prediction  
The column age\_of\_vehicle is IMPORTANT for Prediction  
The column driver\_home\_area\_type is IMPORTANT for Prediction  
The column driver\_ind\_decile is IMPORTANT for Prediction  
The column engine\_capacity\_cc is IMPORTANT for Prediction  
The column hit\_object\_in\_carriageway is IMPORTANT for Prediction  
The column hit\_object\_off\_carriageway is IMPORTANT for Prediction  
The column journey\_purpose\_of\_driver is IMPORTANT for Prediction  
The column junction\_location is IMPORTANT for Prediction

The column make is IMPORTANT for Prediction

The column model is IMPORTANT for Prediction

The column propulsion\_code is IMPORTANT for Prediction

The column sex\_of\_driver is IMPORTANT for Prediction

The column skidding\_and\_overturning is IMPORTANT for Prediction

The column towing\_and\_articulation is IMPORTANT for Prediction

The column vehicle\_leaving\_carriageway is IMPORTANT for Prediction

The column vehicle\_locationrestricted\_lane is IMPORTANT for Prediction

The column vehicle\_manoeuvre is IMPORTANT for Prediction

The column vehicle\_reference is IMPORTANT for Prediction

The column vehicle\_type is IMPORTANT for Prediction

The column was\_vehicle\_left\_hand\_drive is IMPORTANT for Prediction

The column x1st\_point\_of\_impact is IMPORTANT for Prediction

The column month is IMPORTANT for Prediction

The column weekend is IMPORTANT for Prediction

The column hour is IMPORTANT for Prediction

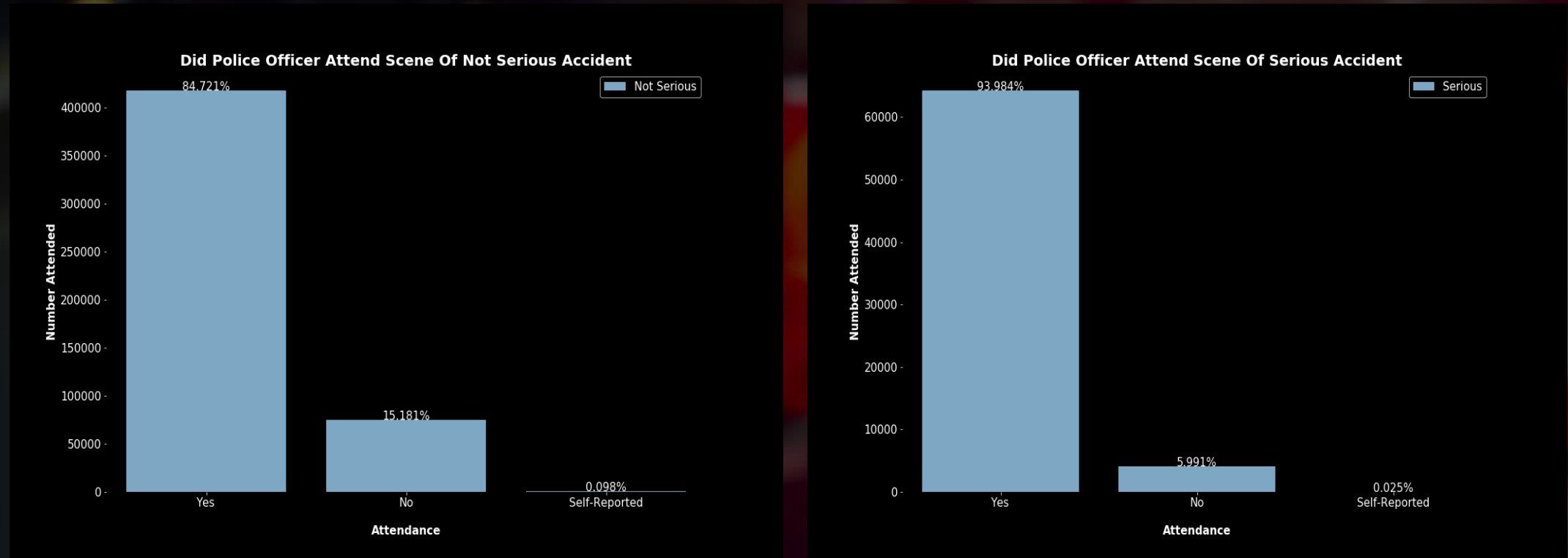
The column time\_of\_day is IMPORTANT for Prediction

The column season is IMPORTANT for Prediction

The column engine\_capacity\_cc\_size is IMPORTANT for Prediction

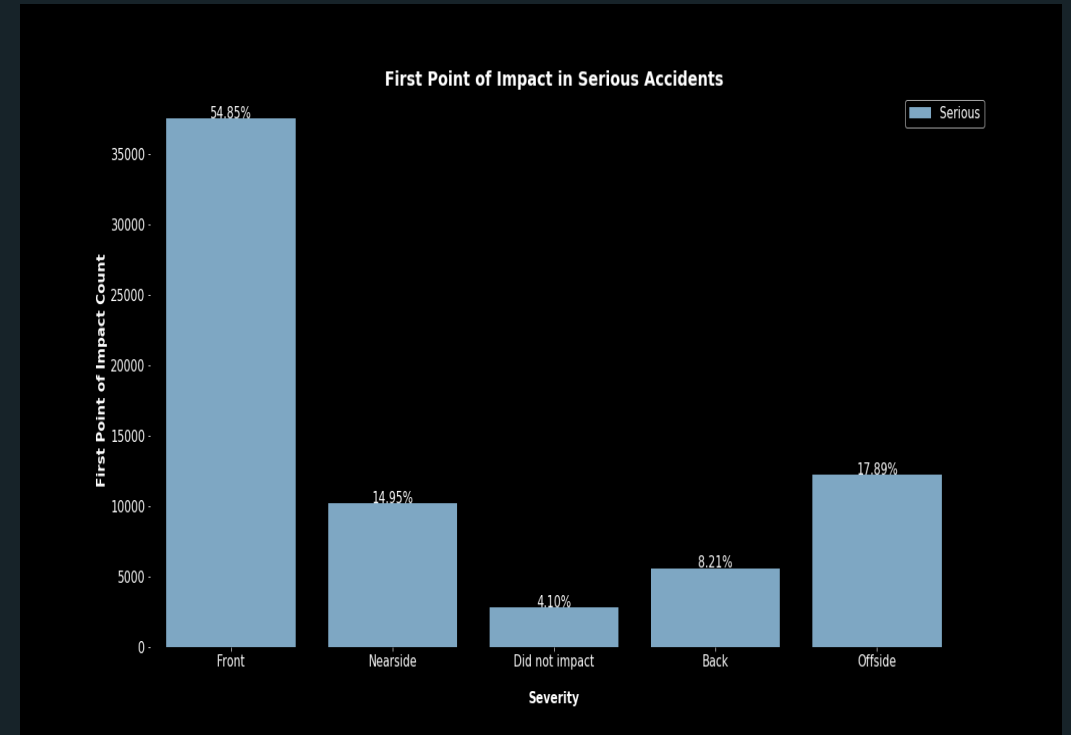
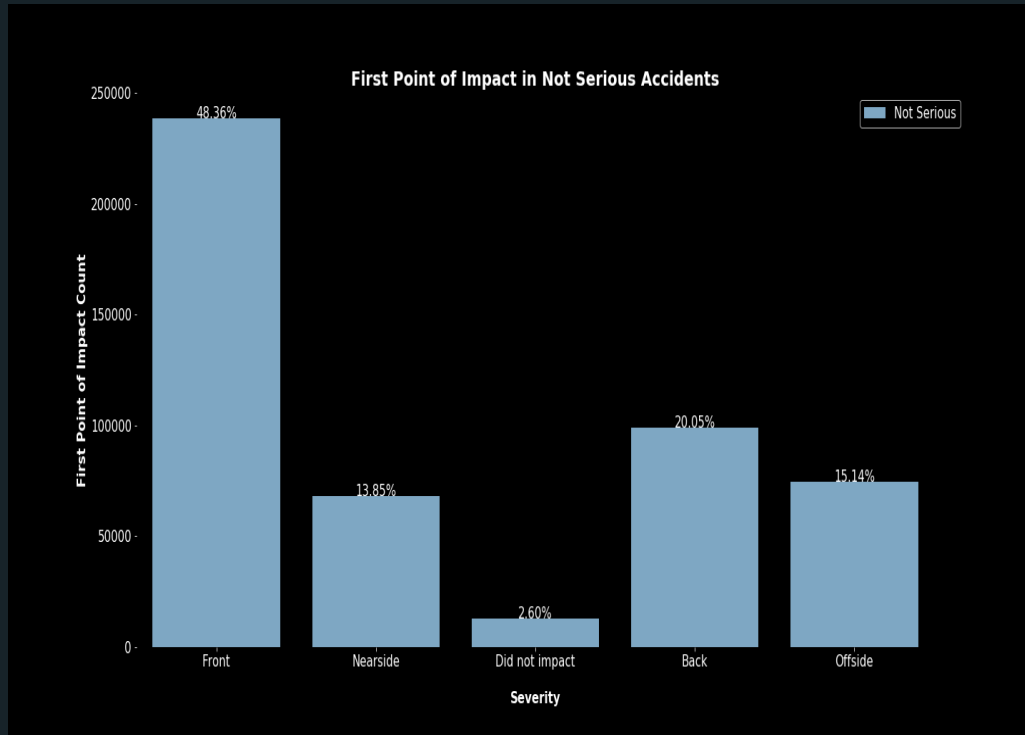
After getting these correlations, I ran them through a Chi-Squared test to check for relevance. Above are screenshots of the coding, and the results. With a requirement of  $p < 0.05$ , all features were deemed important enough for prediction, so I continued with my visualization comparisons.

# Did Police Officer Attend Scene Of Accident



Police attended most accidents but were less likely to NOT be called in serious accidents.

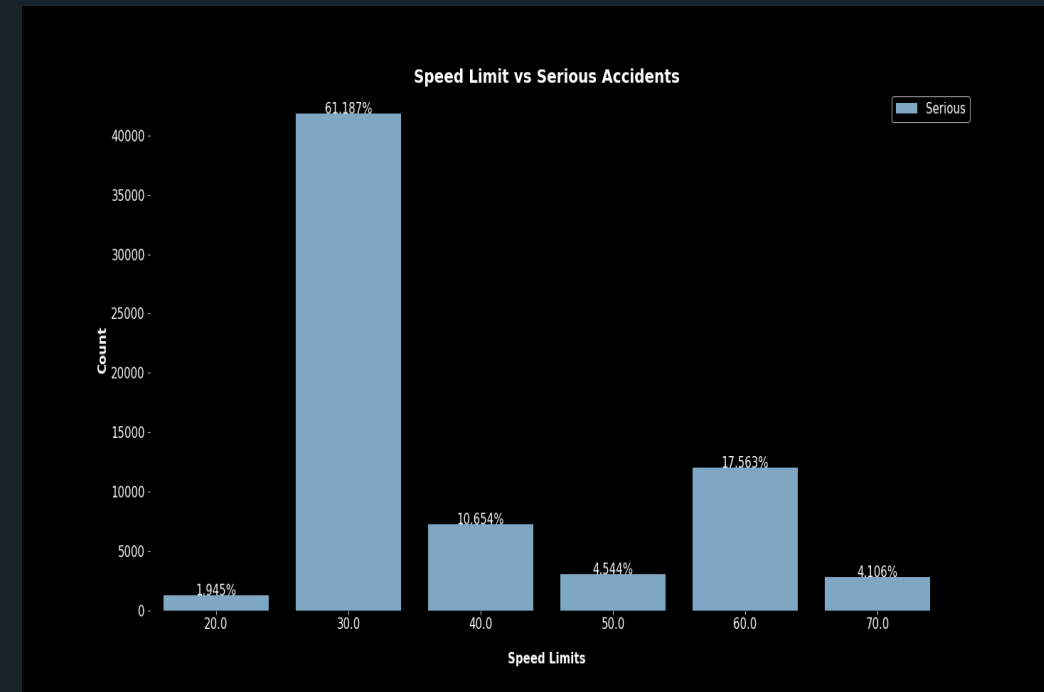
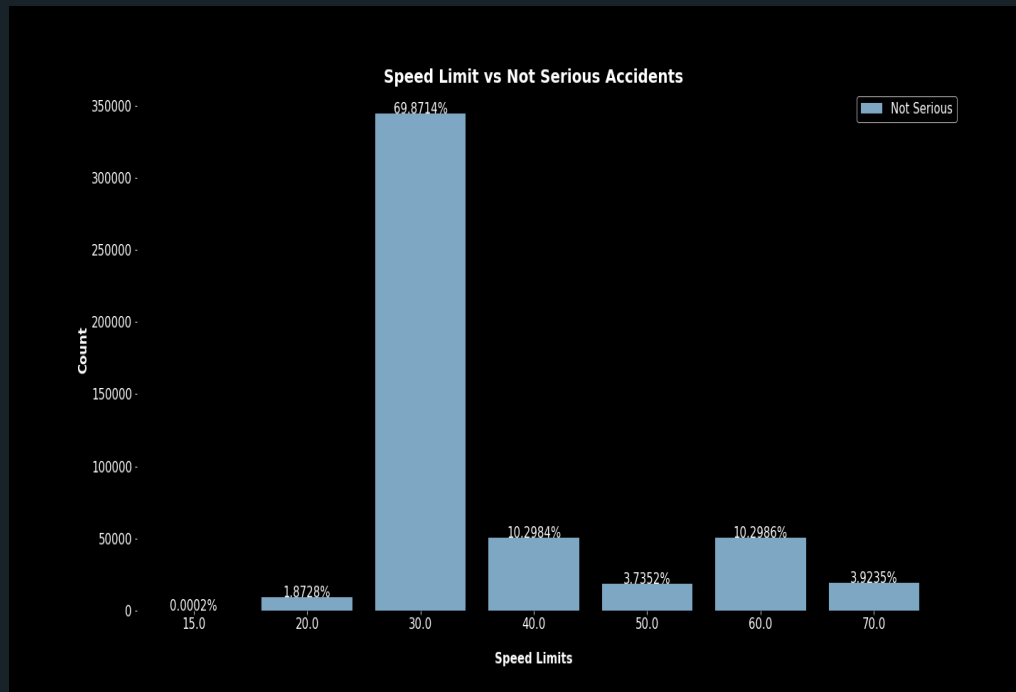
# First Point of Impact



Majority of accidents were front impacted as the first point of impact. Not serious accidents had a higher percentage of Back impact accidents than serious accidents. Serious accidents had higher percentages of Offside and Nearside accidents.

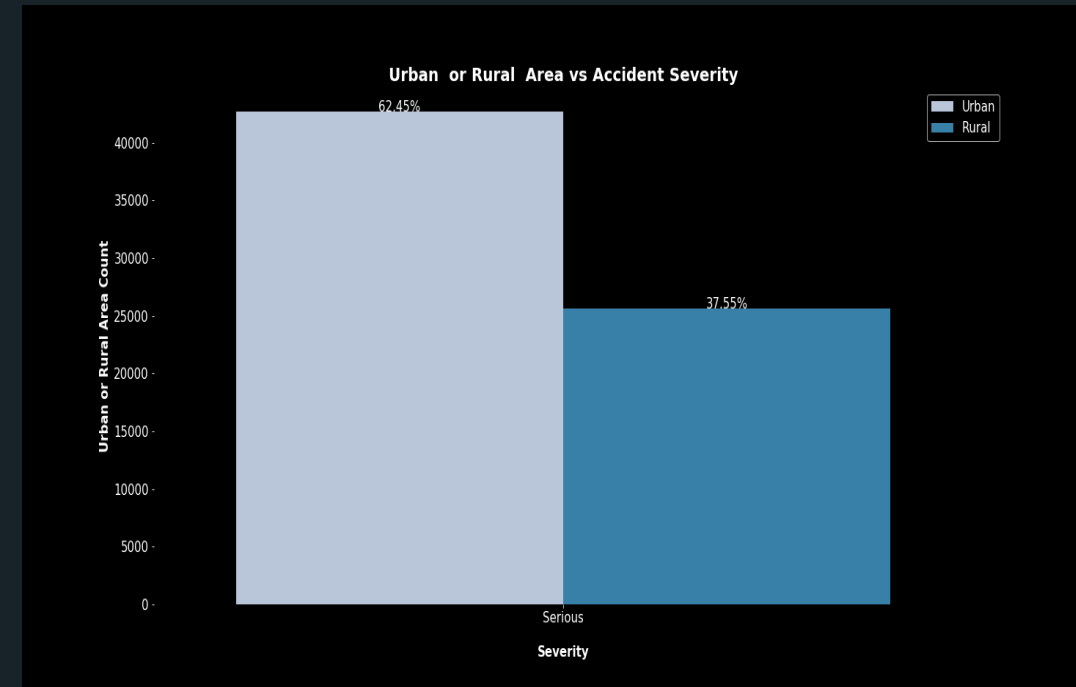
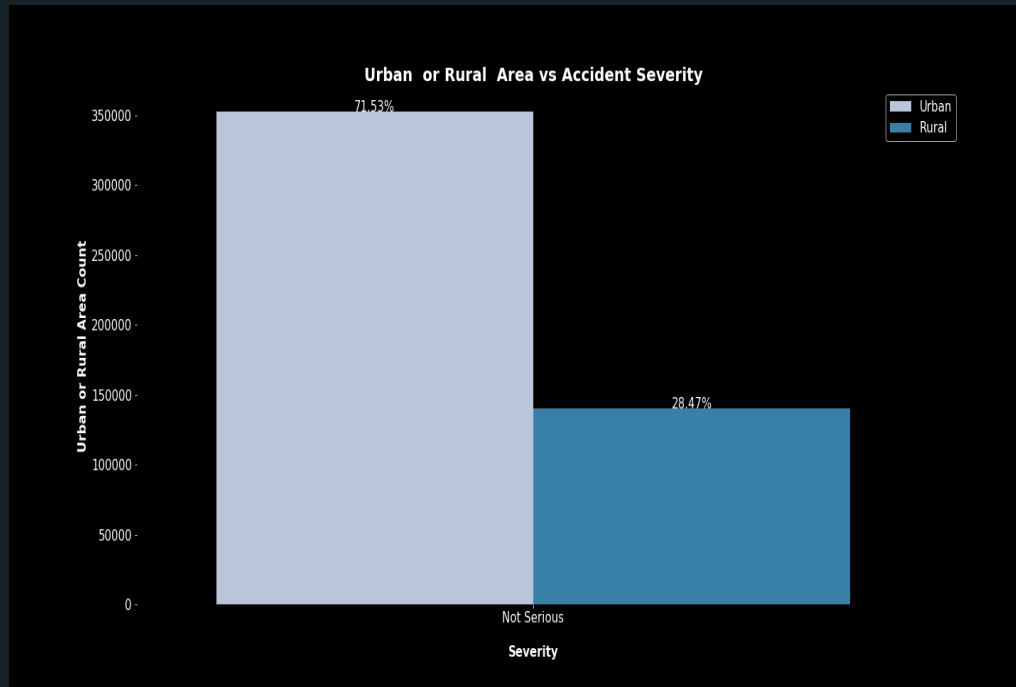


# Speed Limit



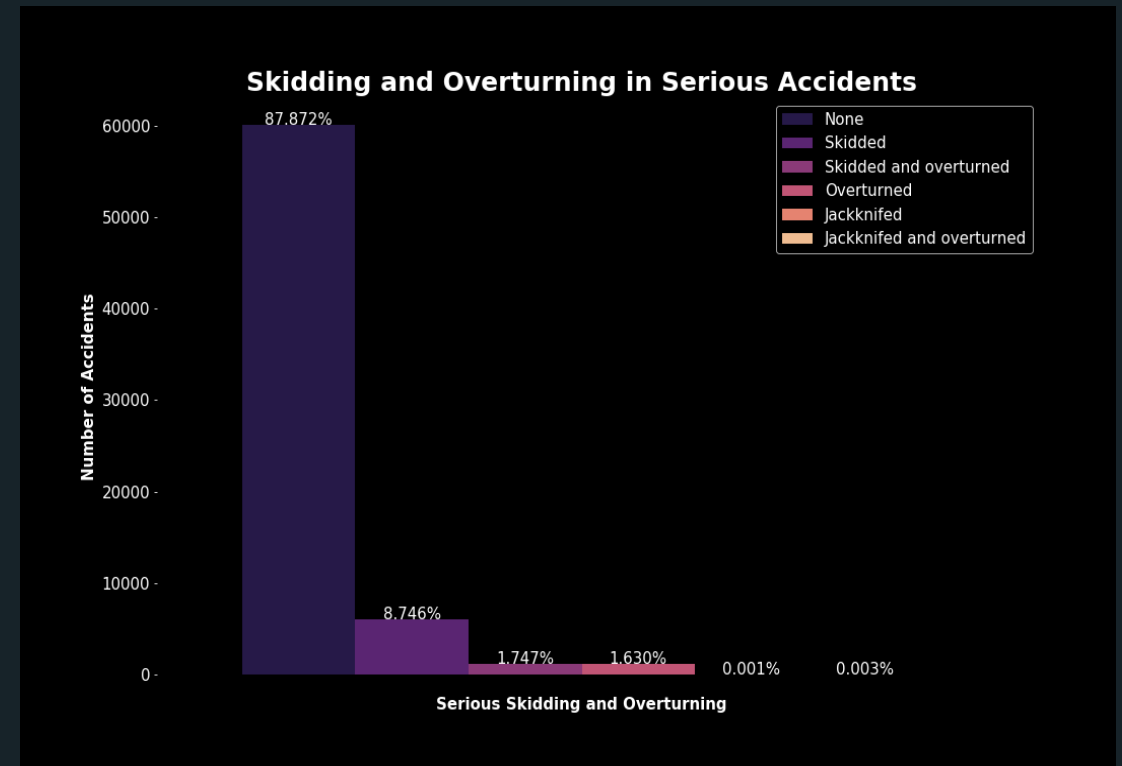
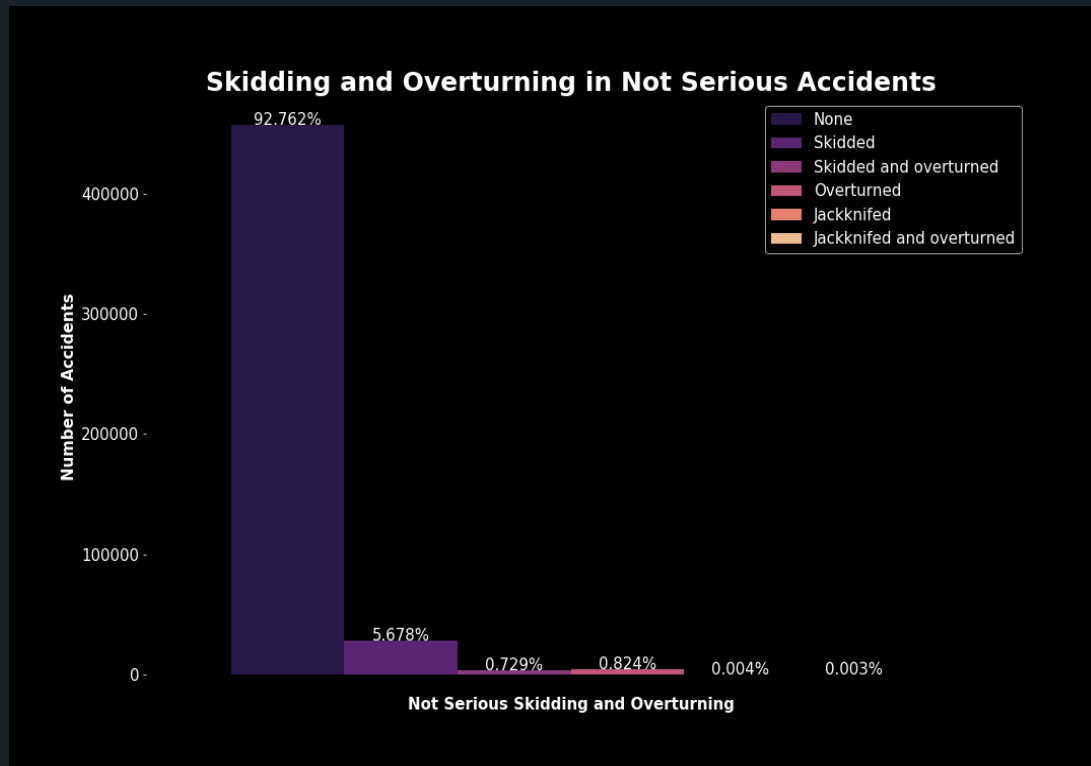
Majority of accidents occurred in 30 speed limit zones. It would have been beneficial to have actual data on the speeds of the vehicles involved or at least if they were speeding.

# Urban or Rural Areas



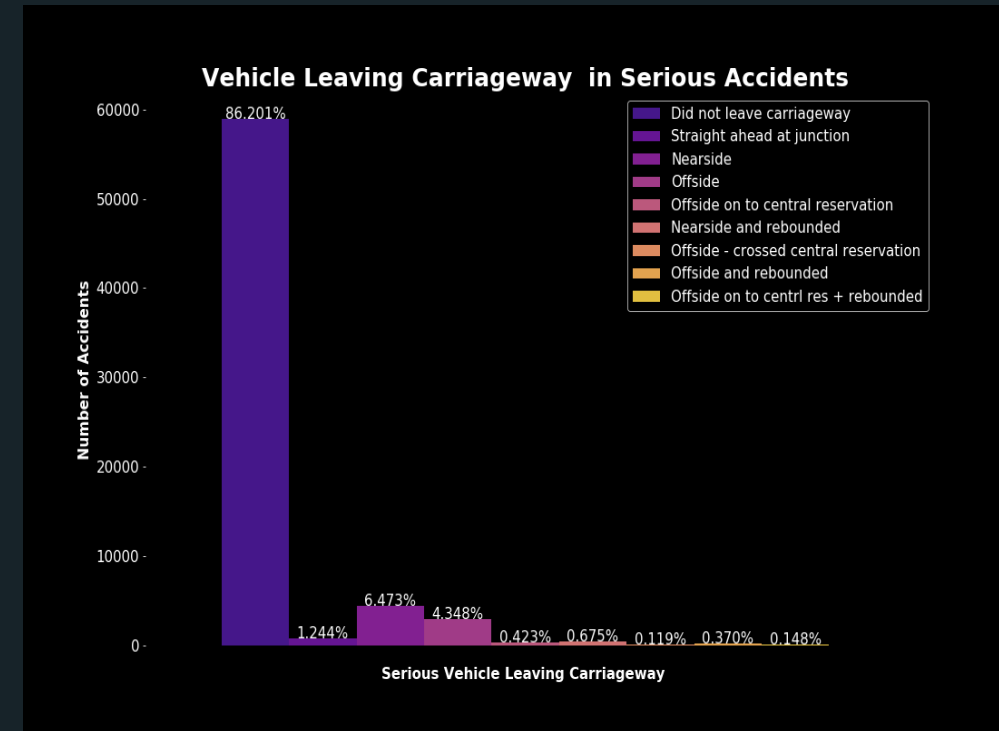
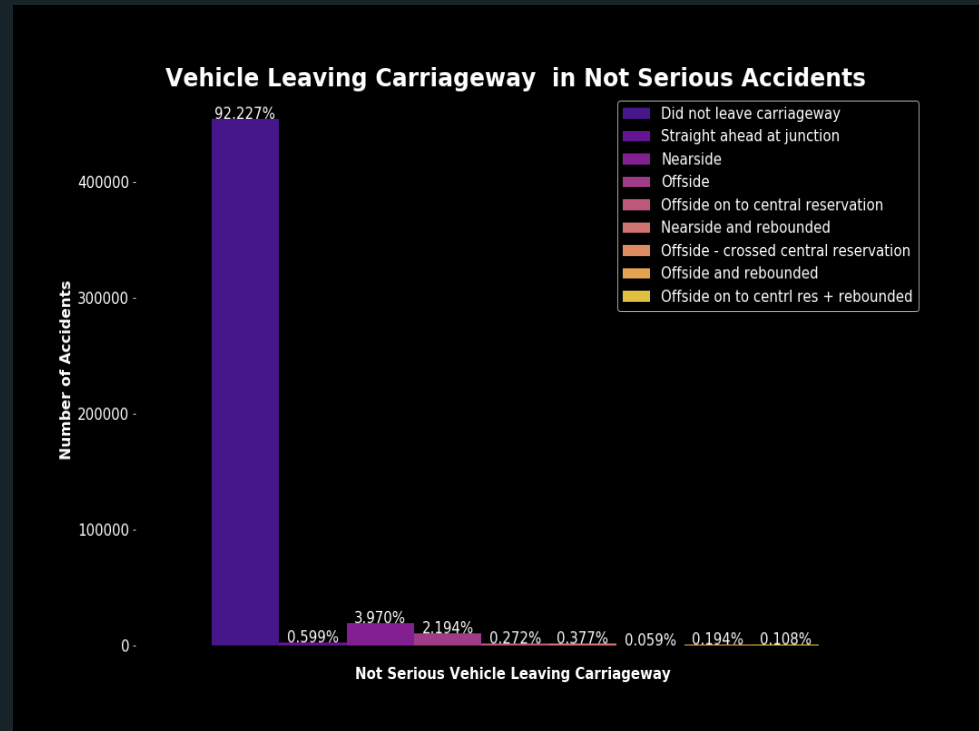
Rural areas had a higher percentage of serious accidents. This may relate to hospital locations or emergency vehicle arriving to the scene of the accident, both of which are not available through this data.

# Skidding or Overturning



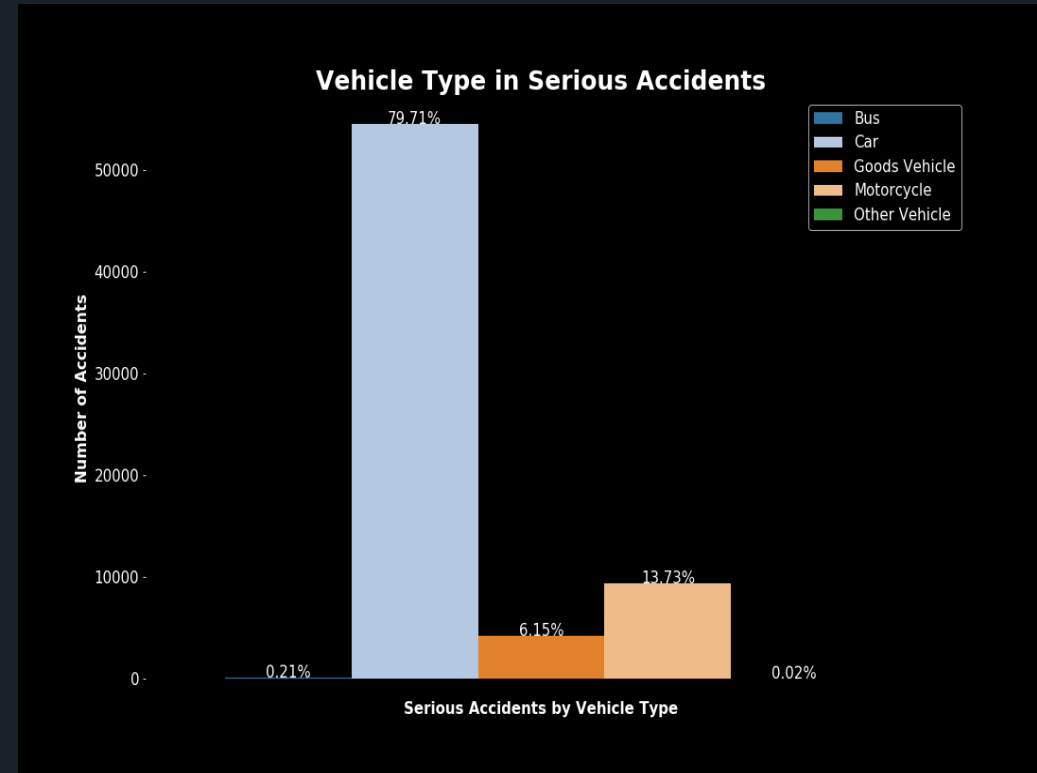
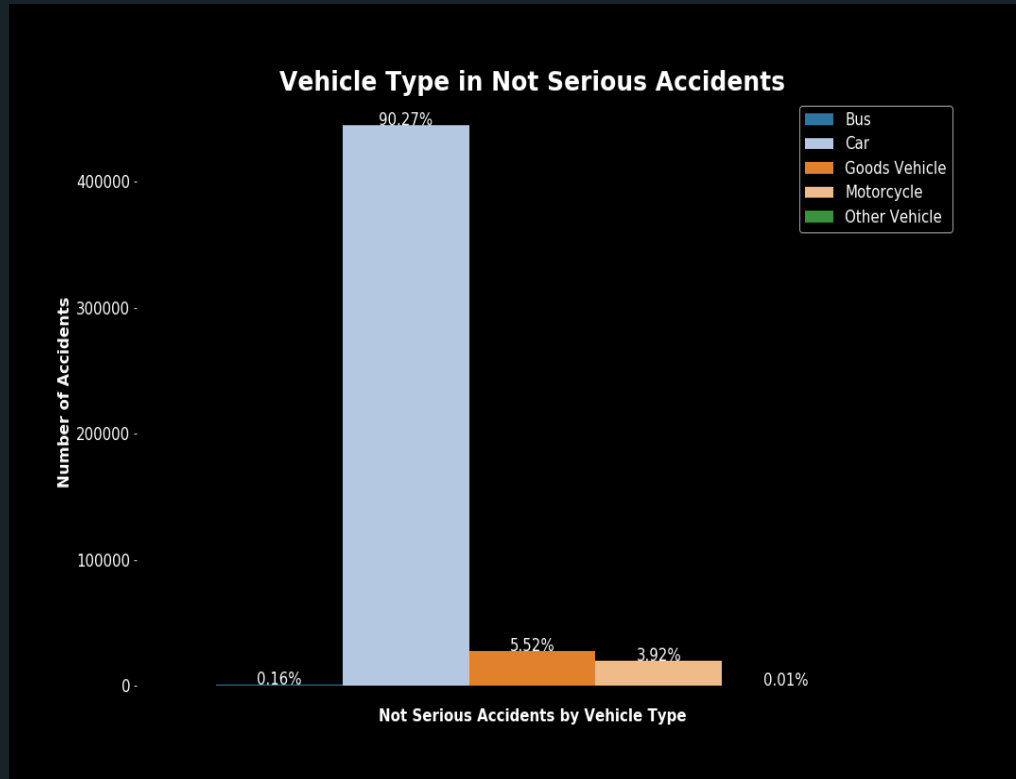
Higher percentages of serious accidents involved skidding, jackknifing or overturning.

# Vehicle Leaving Carriageway



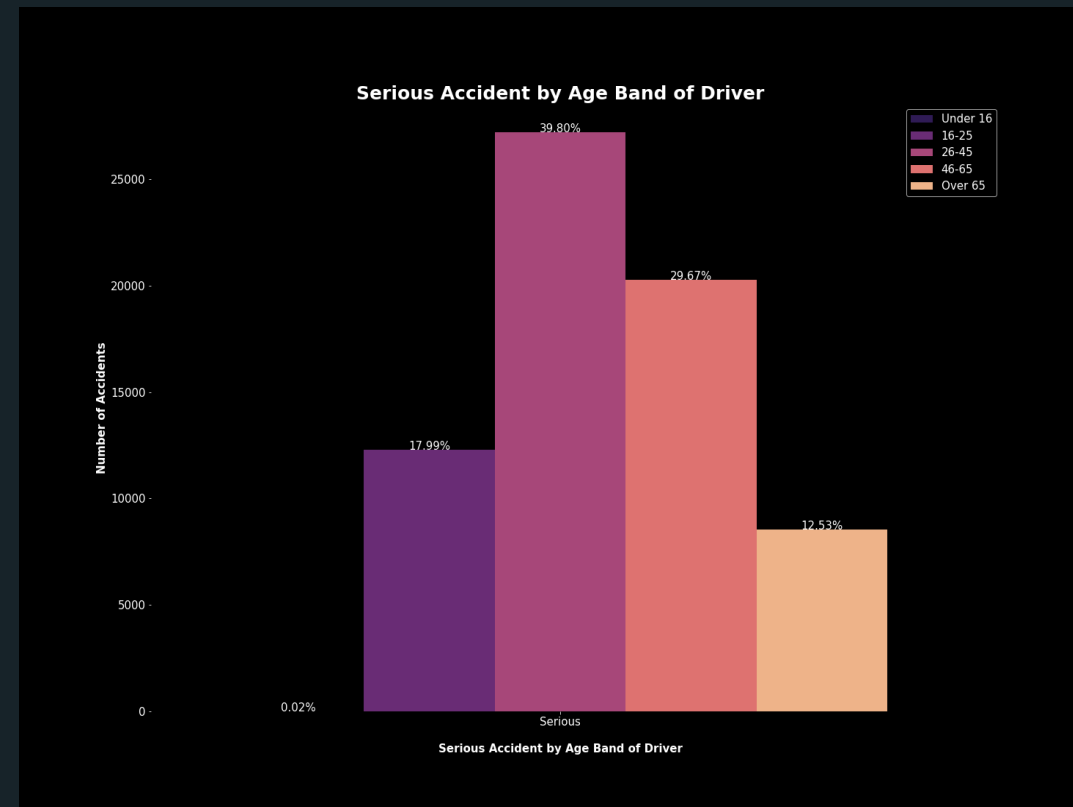
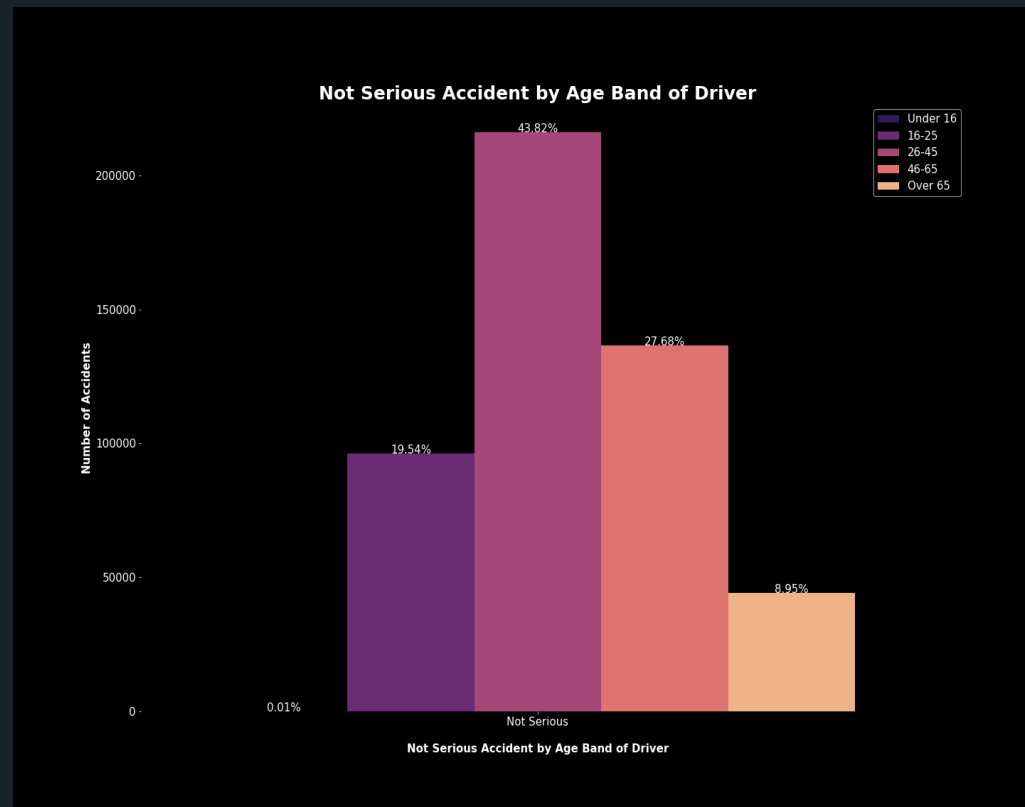
Most vehicles did not leave the carriageway in either type of accident, however serious accidents had higher percentages of those that did leave the carriageway.

# Vehicle Type



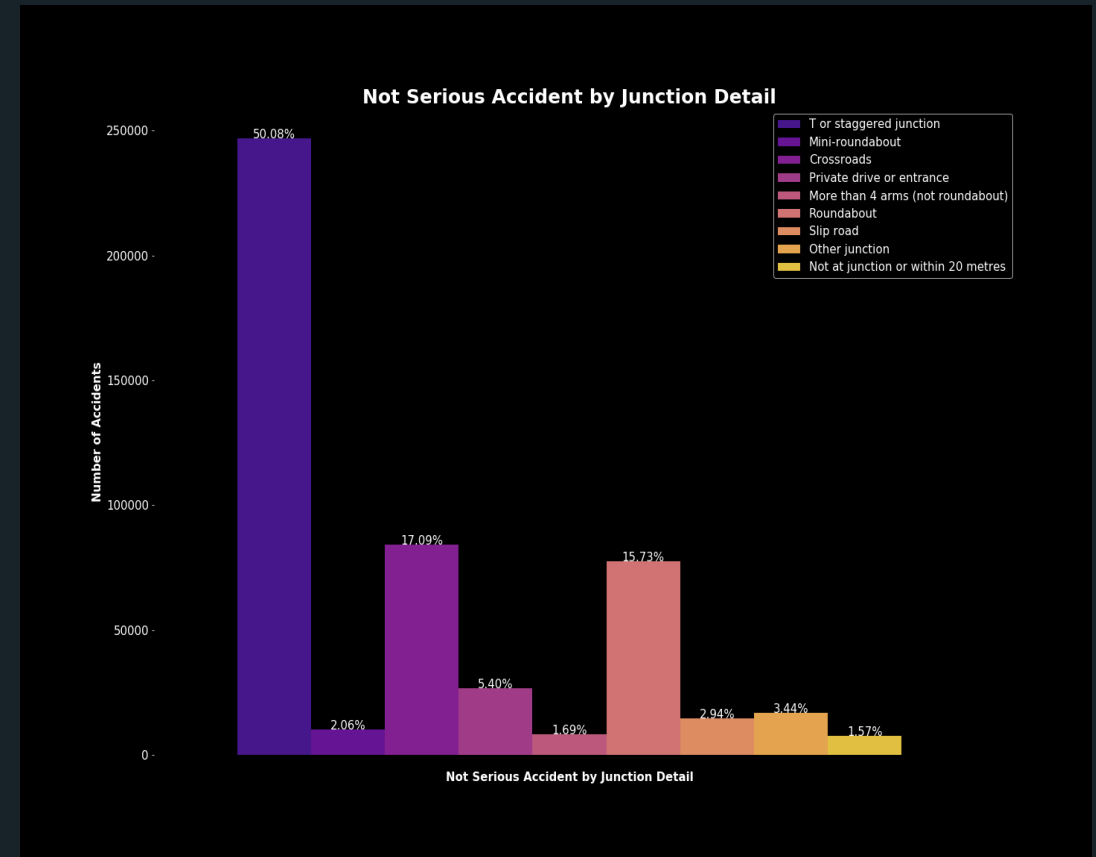
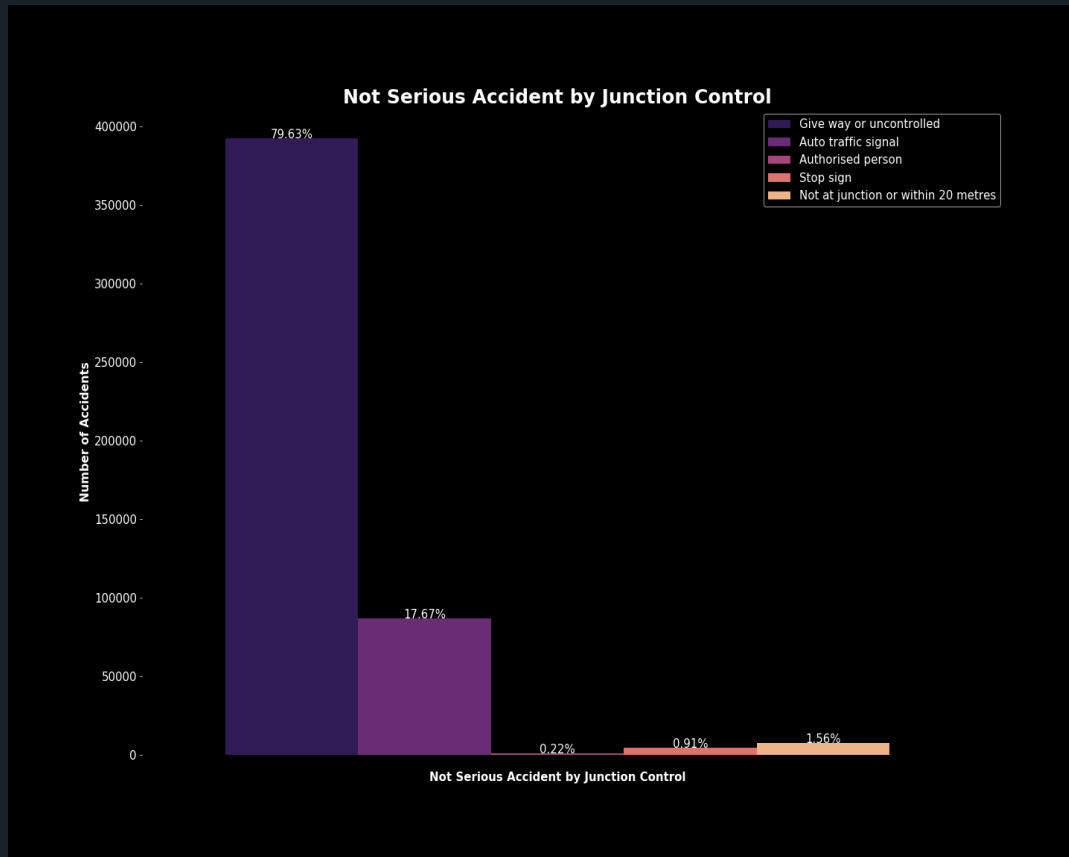
Motorcycles were involved in a significantly higher percentage of serious accidents than not serious accidents.

# Age Band of Driver



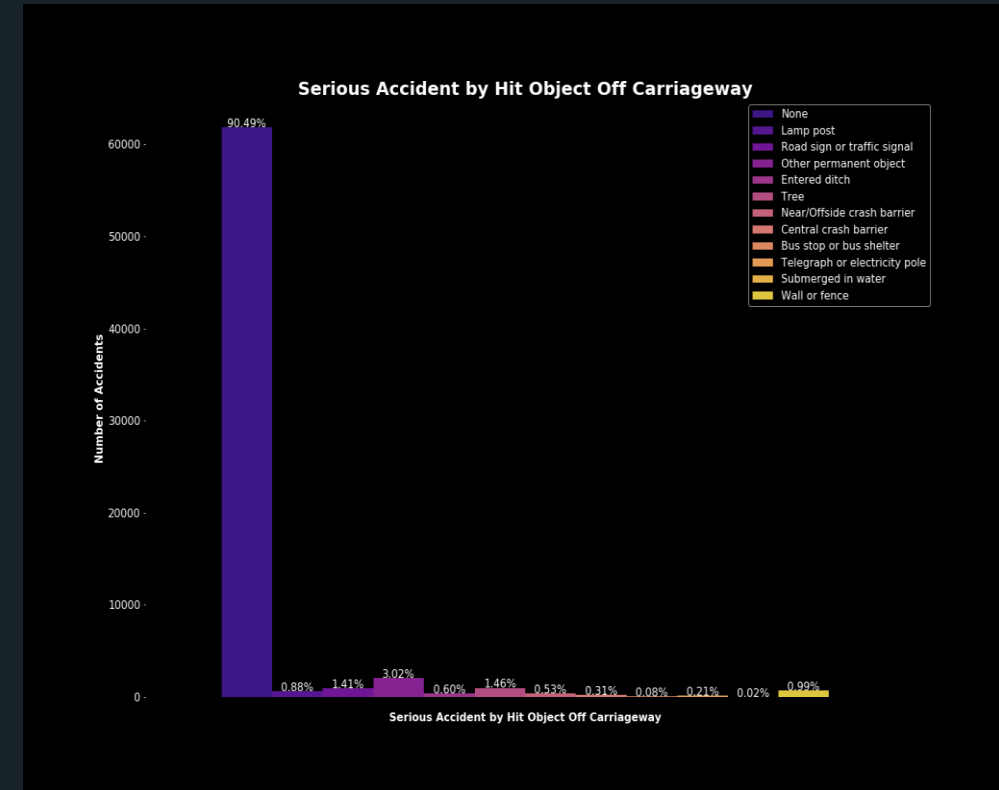
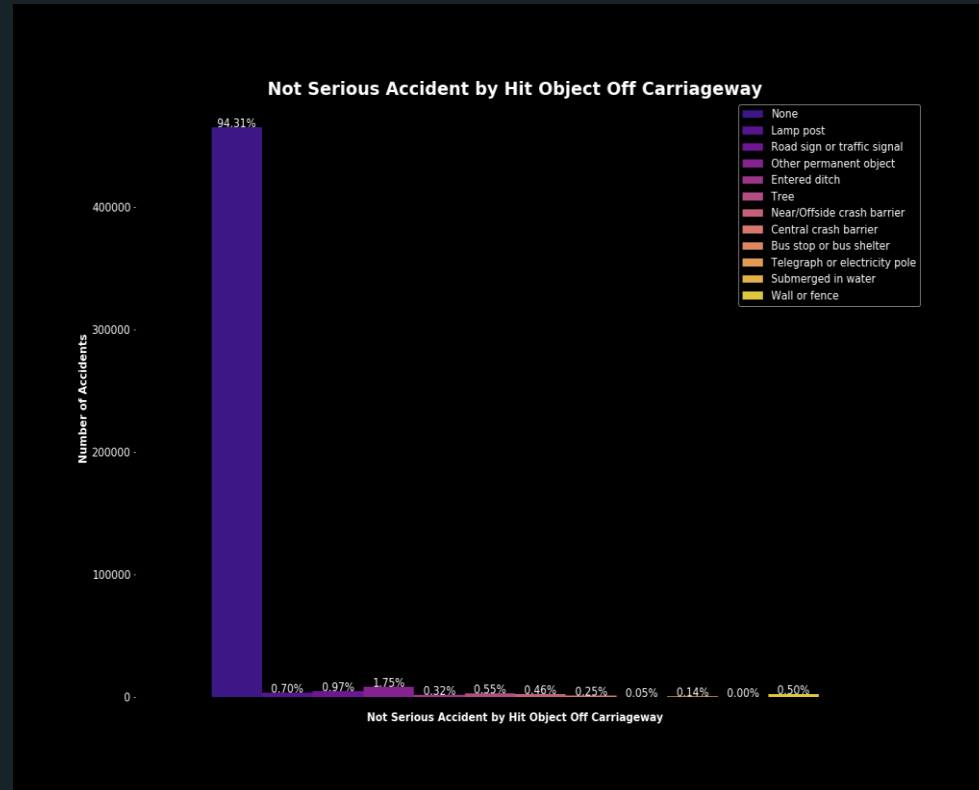
The age groups over the age of 25 had a higher percentage of serious accidents than not serious.

# Junction Control



Most areas with accidents were uncontrolled.

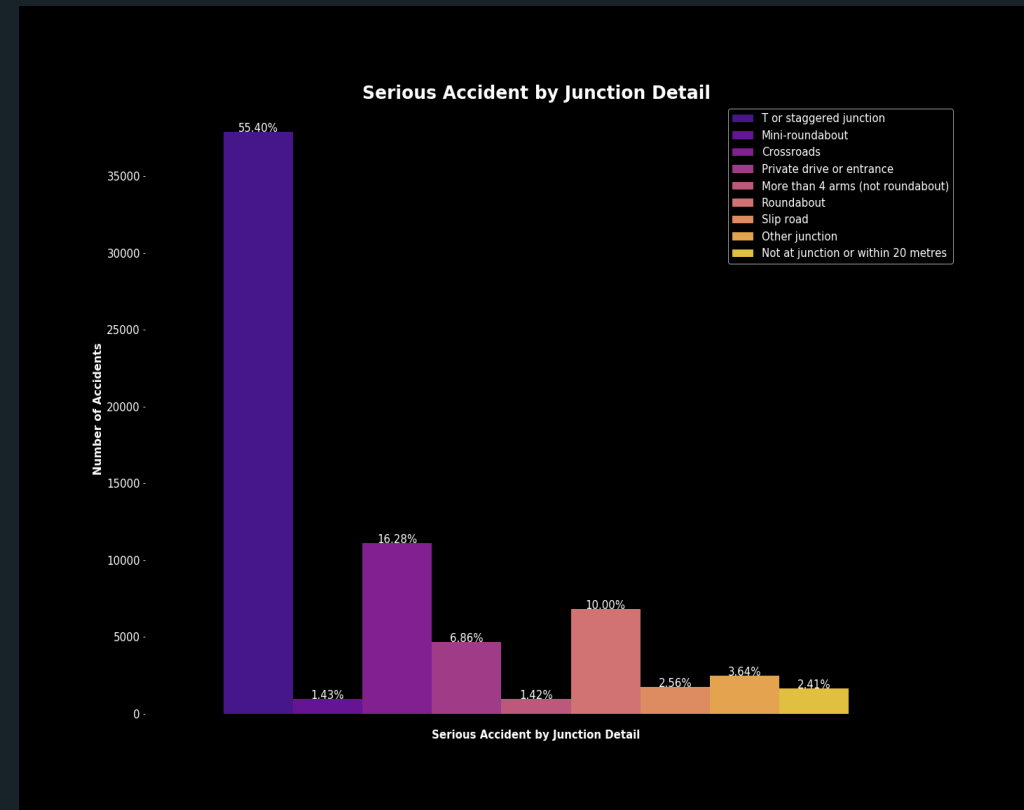
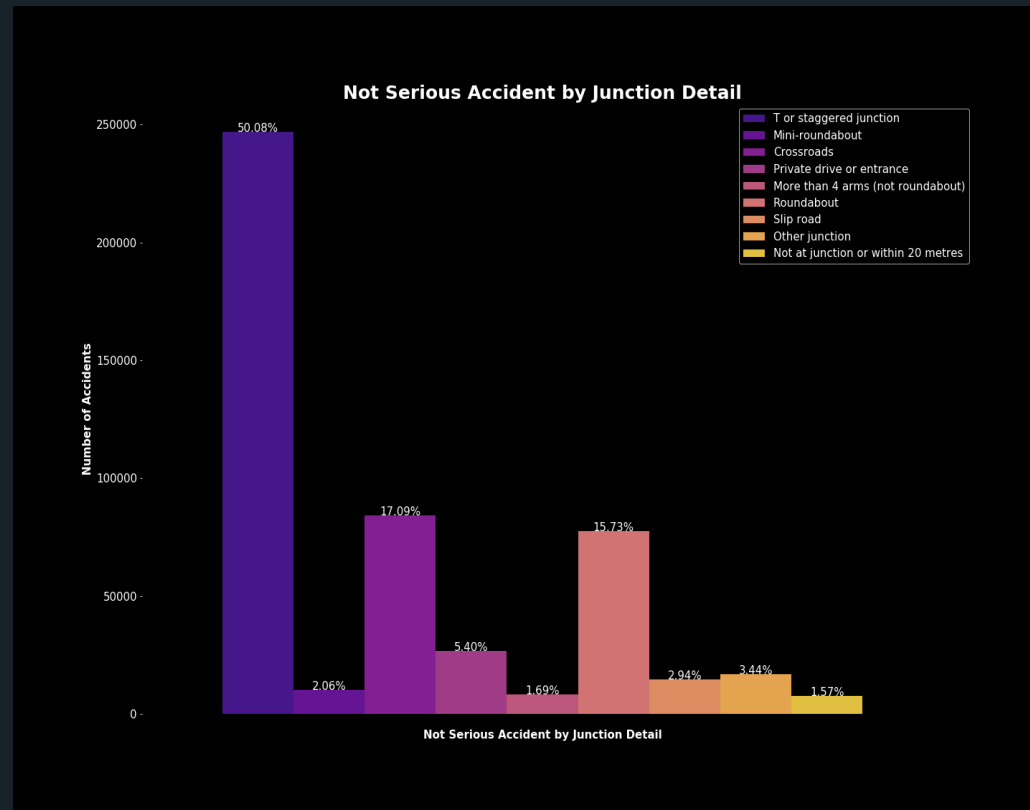
# Hit Object Off Carriageway



Most accidents did not involve objects being hit off the carriageway, however serious accidents had higher percentages of accidents that did involve hitting an object off the carriageway.

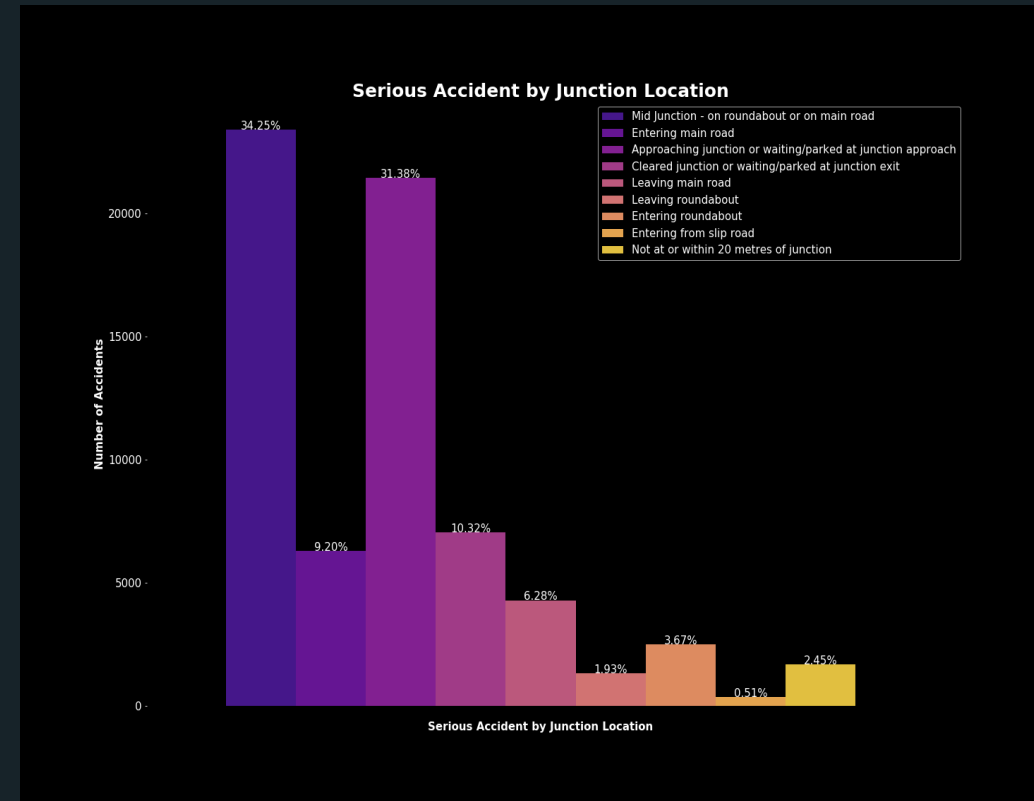
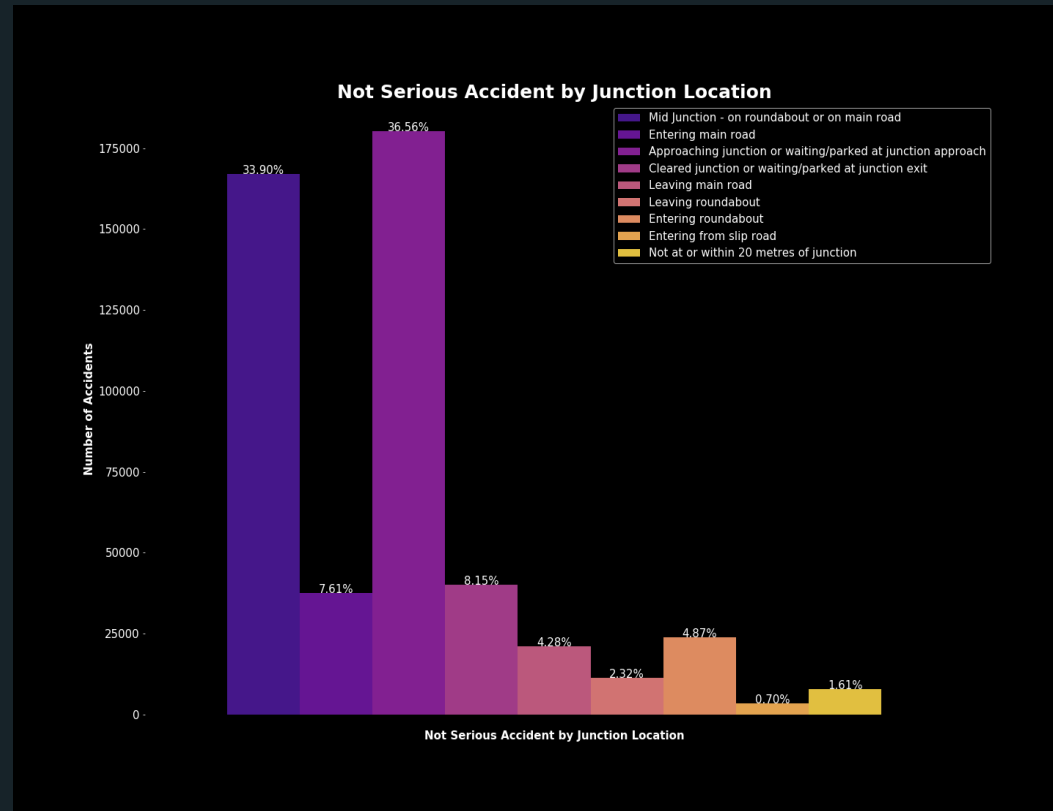


# Junction Detail



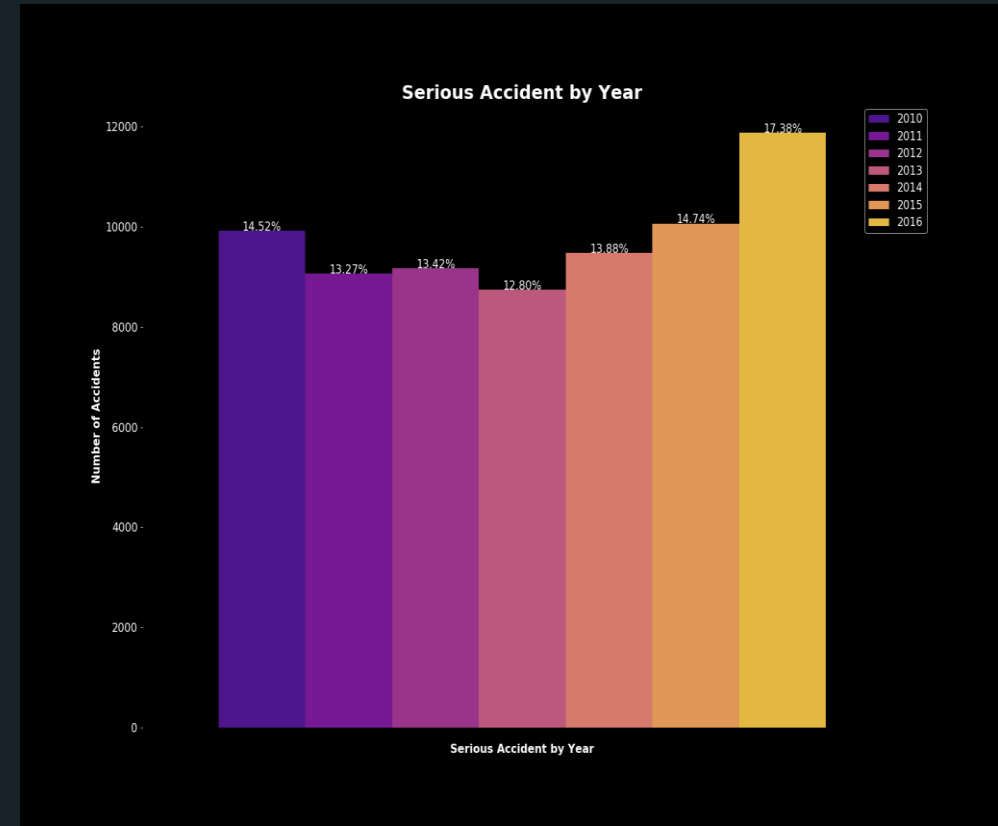
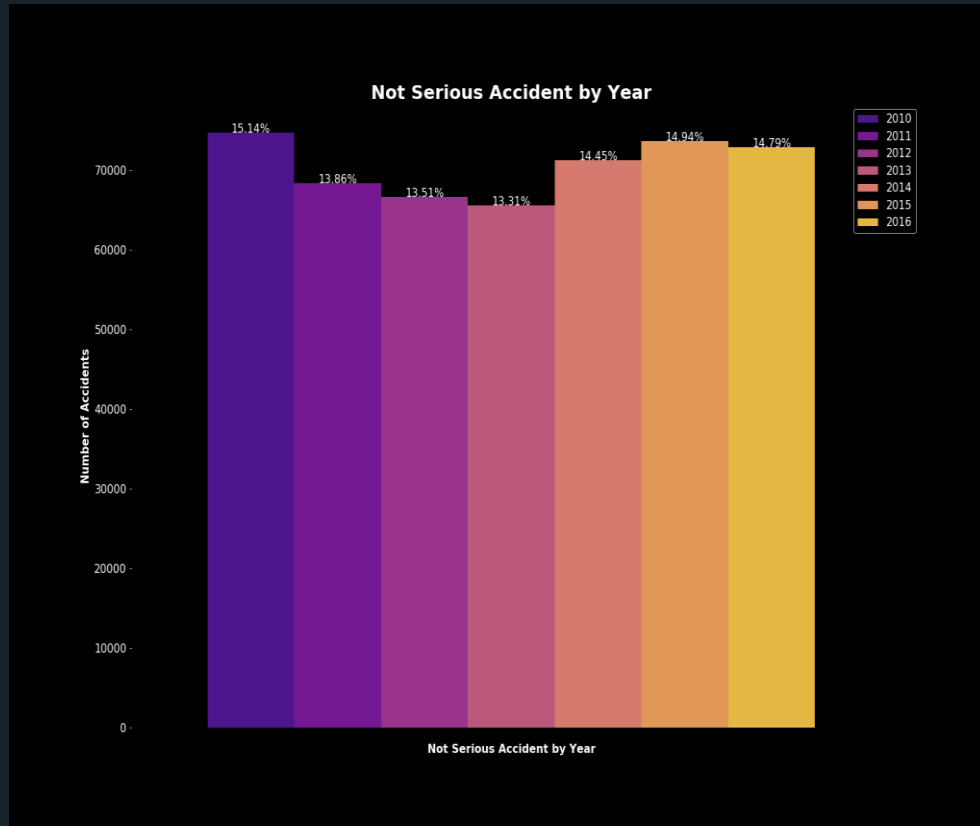
T or staggered junctions were where most of the accidents occurred.

# Junction Location



Most accidents seem to have occurred in the “Mid Junction - on roundabout or on main road” or situations where the driver was “approaching the junction or waiting/parked at junction approach”.

# Year



There has been a spike in percentage of serious accidents over the years. However, the percentage of not serious accidents has remained somewhat consistent.

# Visualizations Summary

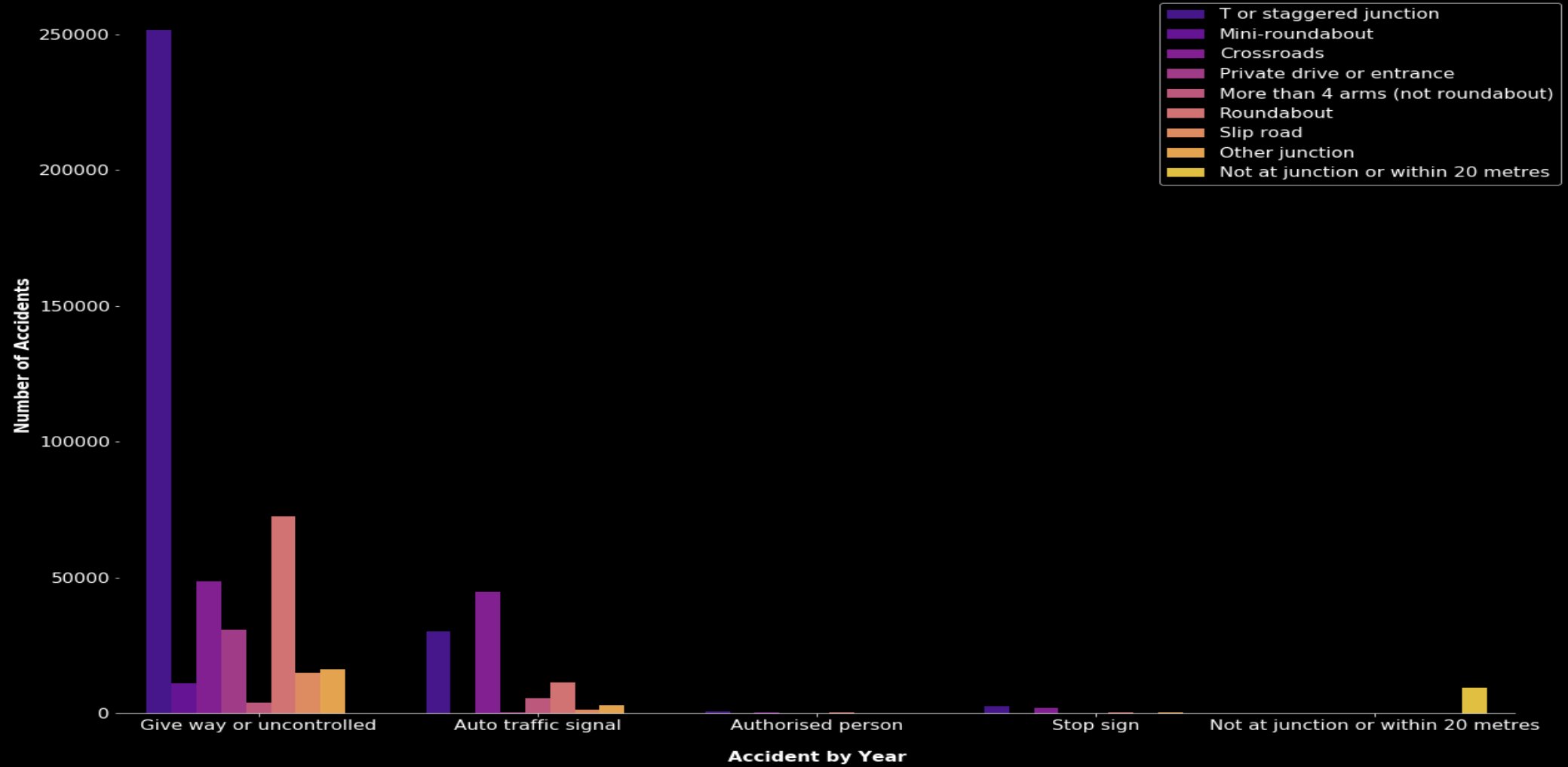
- `did_police_officer_attend_scene_of_accident`: Police attended most accidents but were less likely to NOT be called in serious accidents.
- `x1st_point_of_impact`: Majority of accidents were front impacted as the first point of impact. Not serious accidents had a higher percentage of Back impact accidents than serious accidents. Serious accidents had higher percentages of Offside and Nearside accidents.
- `number_of_vehicles`: Nothing significant.
- `speed_limit`: Majority of accidents occurred in 30 speed limit zones. It would have been beneficial to have actual data on the speeds of the vehicles involved or at least if they were speeding.
- `urban_or_rural_area`: Rural areas had a higher percentage of serious accidents. This may relate to hospital locations or emergency vehicle arrival data which was not available.
- `skidding_and_overturning`: Higher percentages of serious accidents involved skidding, jackknifing or overturning.
- `vehicle_leaving_carriageway`: Most vehicles did not leave the carriageway in either type of accident, however serious accidents had higher percentages of those that did leave the carriageway.
- `sex_of_driver`: Men were more involved in both serious and not serious accidents, however according to [racfoundation.org](https://www.racfoundation.org), there are only 355 of female privately registered cars on UK roads.
- `vehicle_type`: Motorcycles were involved in a significantly higher percentage of serious accidents than not serious accidents
- `vehicle_manoeuvre`: Nothing significant.
- `driver_home_area_type`: Rural and Small Towns has higher percentages of serious accidents. This may relate to hospital locations or emergency vehicle arrival data which was not available.
- `age_band_of_driver`: The age bands over the age of 25 had a higher percentage of serious accidents than not serious.
- `junction_control`: Most areas with accidents were uncontrolled.
- `hit_object_off_carriageway`: The majority of accidents did not involve objects being hit off the carriageway, however serious accidents had higher percentages of accidents that did involve hitting an object off the carriageway.
- `hit_object_in_carriageway`: Most accidents did not involve objects being hit in the carriageway; however serious accidents had higher percentages of accidents that did involve hitting an object off the carriageway.
- `driver_imd_decile`: Nothing significant. Most accidents occurred in areas that were Less deprived 20-30%
- `junction_detail`: T or staggered junctions were where most of the accidents occurred.
- `junction_location`: Nothing that separates the two serious types. However, most accidents seem to have occurred in Mid Junction - on roundabout or on main road or situations where the driver was approaching junction or waiting/parked at junction approach.
- `propulsion_code`: Diesel, Fuel cells, New fuel technology, vehicles were not recorded as a part of serious accidents.
- `year`: There has been a spike in percentage of serious accidents over the years. However, the percentage of not serious accidents has remained somewhat consistent

# Other Visualizations

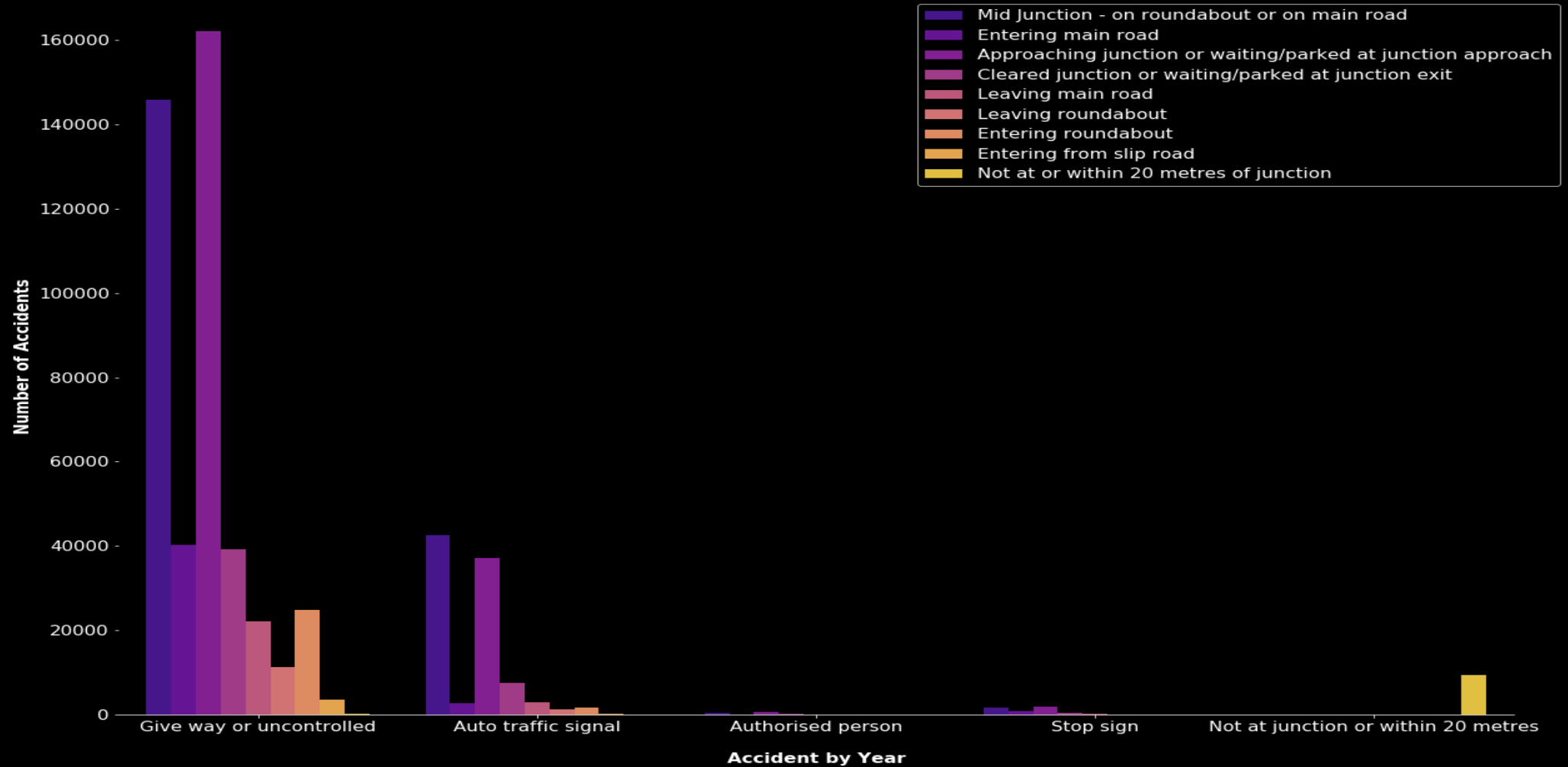
- Junction Control by Junction Detail
- Junction Control by Junction Location
- First point of Impact by Junction Detail
- First point of Impact by Junction Location
- Junction Control and First Point of Impact

Due to the previous visualization a comparison of certain variables was desired to see more correlations. The comparisons listed above will be displayed in the slides to follow.

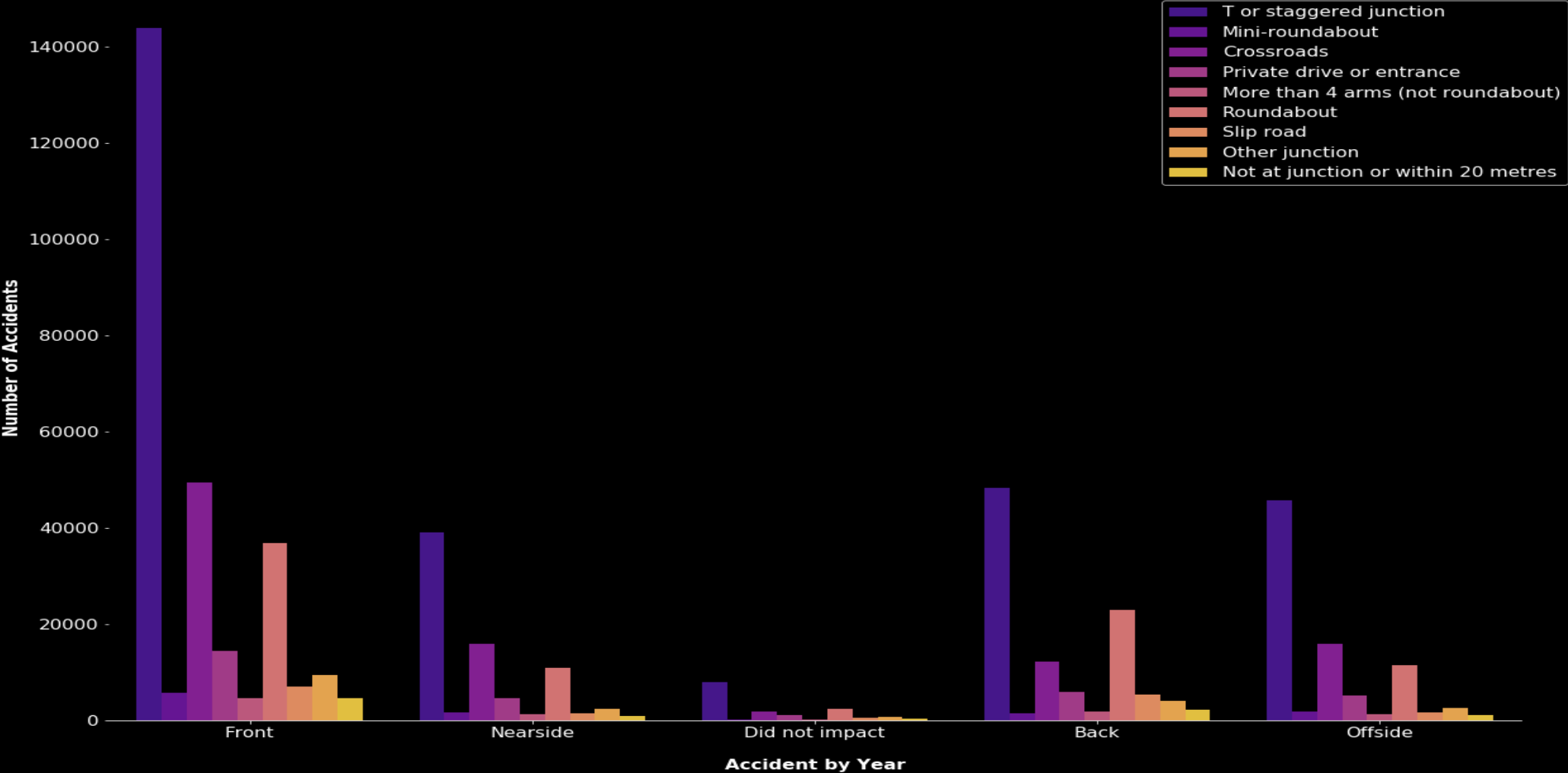
## Junction Control by Junction Detail



### Junction Control by Junction Location in Accidents

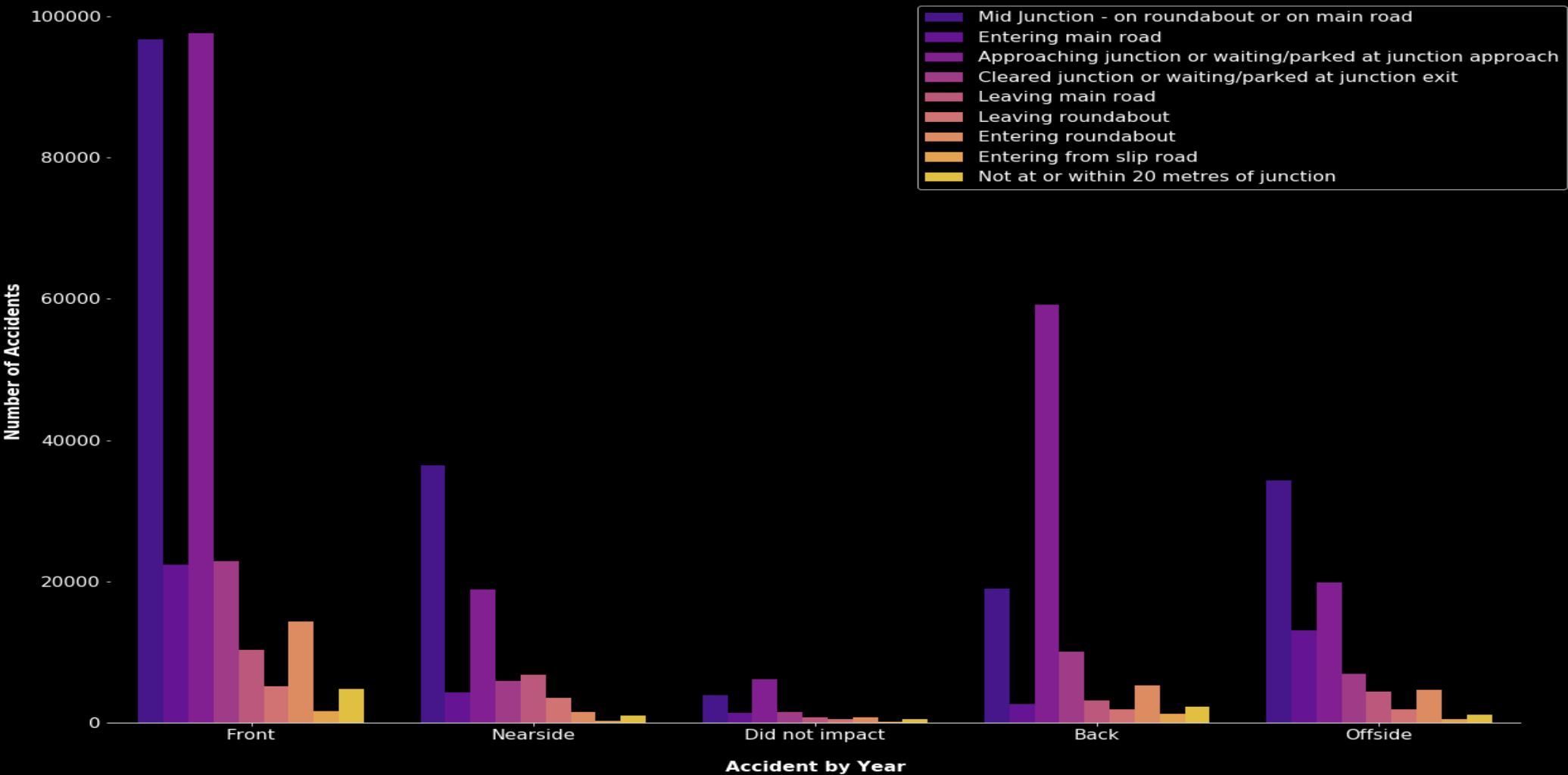


First point of Impact by Junction Detail

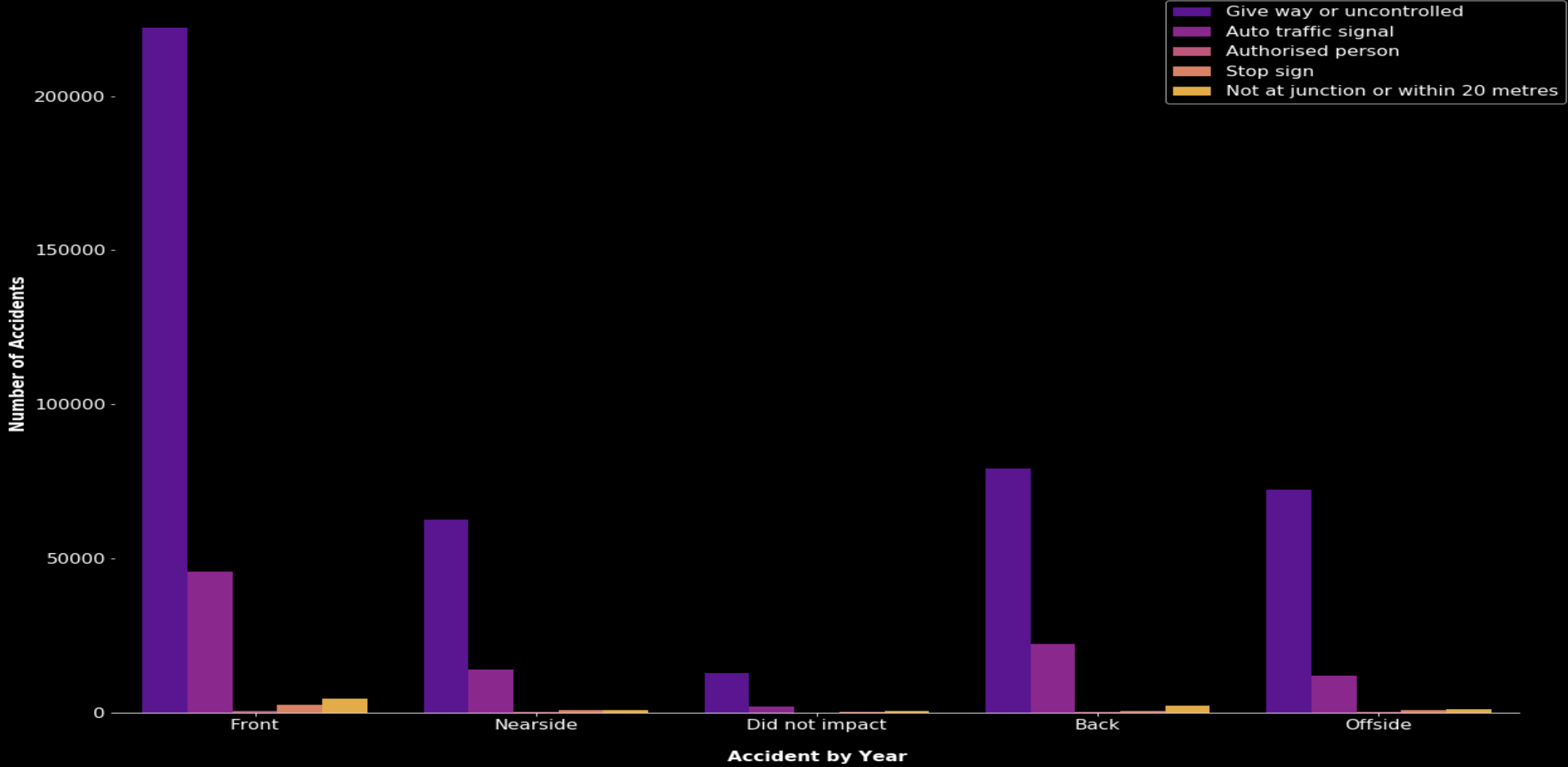




First point of Impact by Junction Location



First point of Impact by Junction Control



# Other Visualizations Summary

No matter the situation above, the most accidents were involving areas that were uncontrolled. One of the main areas where this happened was in the Junction Detail T or staggered junction.



Other areas of concern include accident locations that included Mid Junctions on roundabouts or main roads.



No matter the location, detail, or location of impact the common denominator seems to be a lack of signage or control in junction areas.

# Possible Solution






- From the data above more controlled areas would be beneficial. Maybe signs alerting drivers of the upcoming junctions, traffic lights, or stop signs would help in some of these areas where they are feasible.





For example, this is a staggered junction, the main junction detail in accidents. One can understand how a situation such as these can lead to numerous accidents especially if proper signage is not available. Perhaps traffic lights, stop signs, or warnings indicating that they are approaching certain junctions would help reduce accidents.

# Signage Options

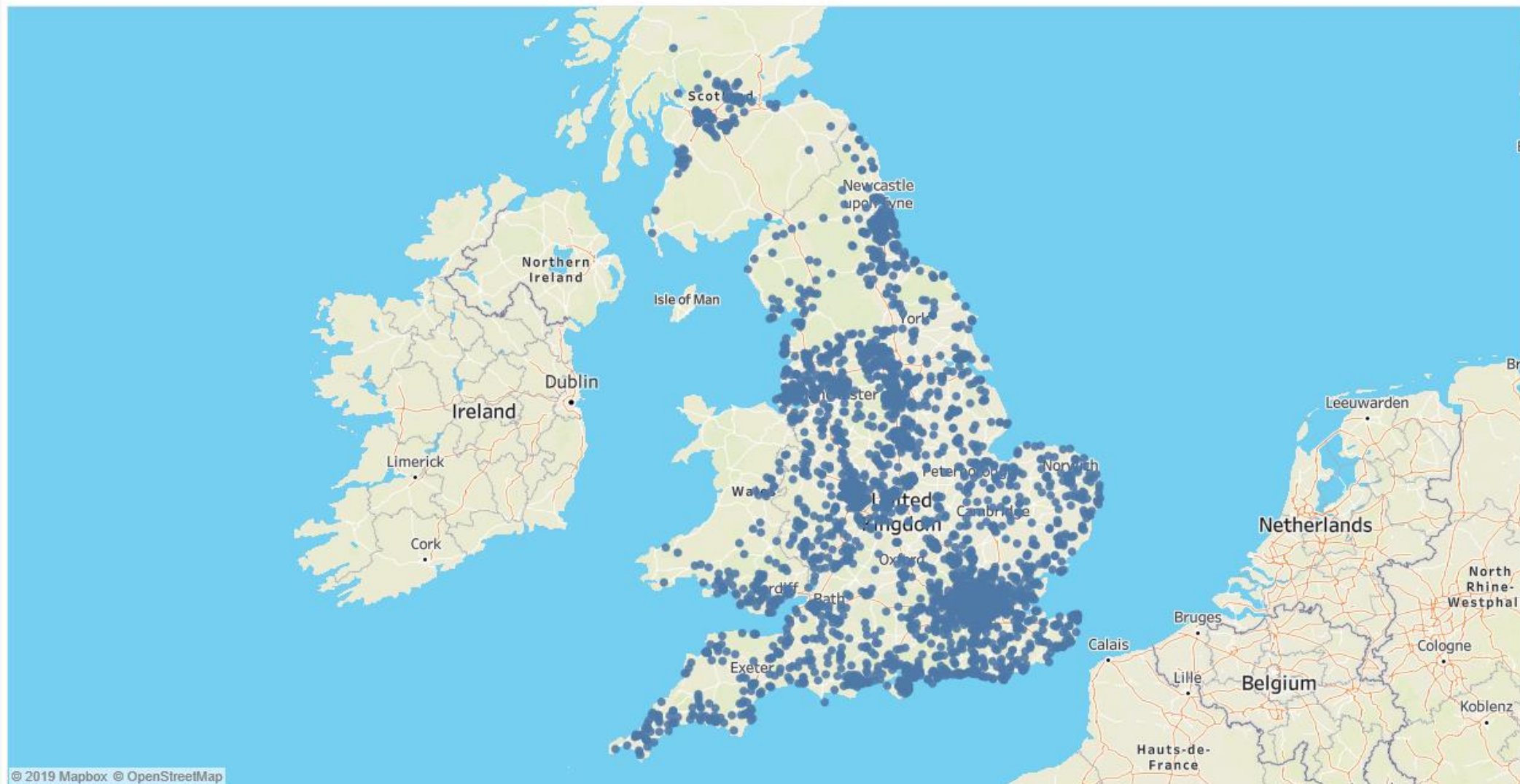
The following images were obtained through web scraping of the website [Learner Driving Centres](https://www.learnerdrivingcentres.co.uk/) which contains information on road signs in the UK.

	sign	sign description
5		Stop and give way
6		Give way to traffic on major road
20		Give priority to vehicles from opposite direction
30		Mini-roundabout (roundabout circulation - give way to vehicles from the immediate right)
39		Distance to 'Give Way' line ahead

	Sign	Sign Description
41		Junction on bend ahead
42		T-junction with priority over vehicles from the right
43		Staggered junction

	sign	sign description
30		Mini-roundabout (roundabout circulation - give way to vehicles from the immediate right)
47		Roundabout

## Accidents in Areas with High Deprivation and No Signage at T or Staggered Junctions in 2016



© 2019 Mapbox © OpenStreetMap

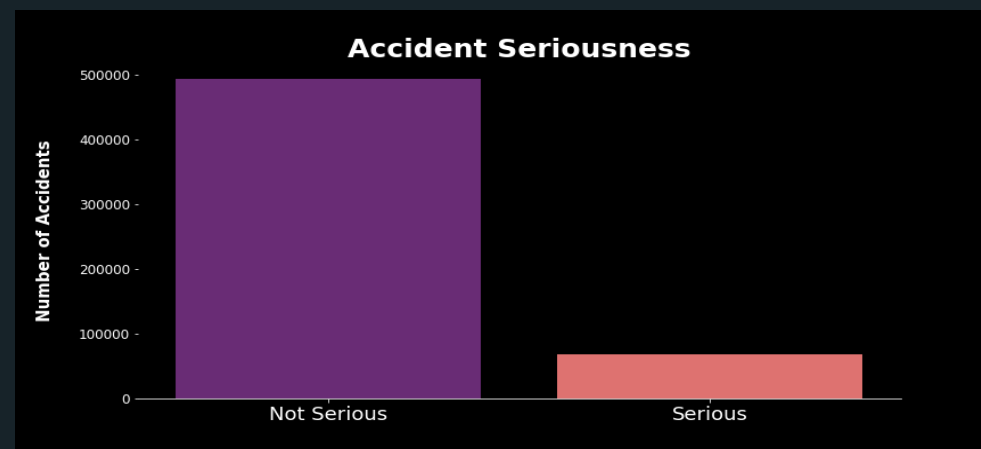
Source: [https://public.tableau.com/shared/43FDWZTYG?:display\\_count=yes&:origin=viz\\_share\\_link](https://public.tableau.com/shared/43FDWZTYG?:display_count=yes&:origin=viz_share_link)

Web Viewer [Terms](#) | [Privacy & Cookies](#)

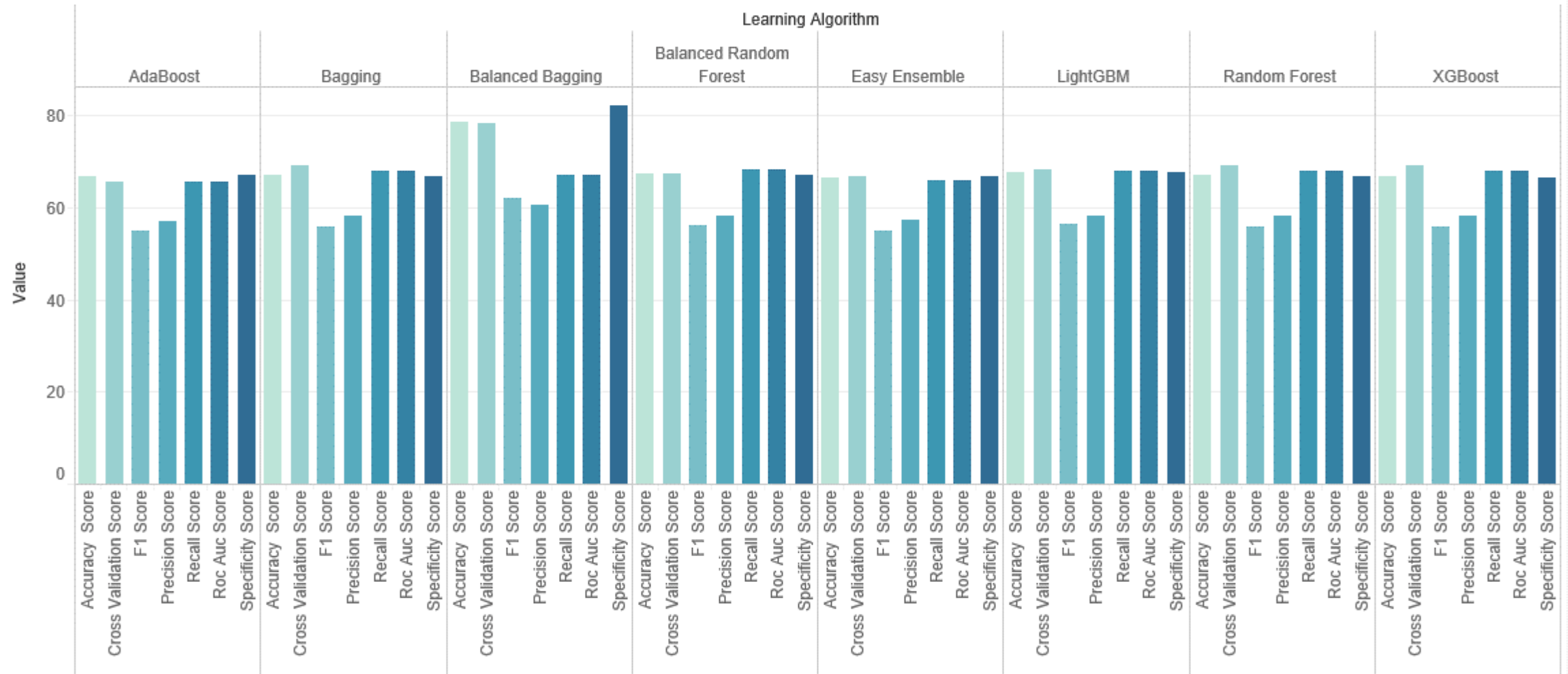
Edit

# Can we create a machine learning algorithm that correctly predicts the severity of accidents?

- The data in this dataset is extremely imbalanced for what we are trying to predict (see graph). We resampled the data as undersampling, where we reduce the number of majority (Not Serious Accidents) samples.
- The machine learning classifier algorithms that we are going to use are as follows:
  - Bagging Classifier (sklearn)
  - AdaBoost Classifier (sklearn)
  - Random Forest Classifier (sklearn)
  - LightGBM Classifier (LightGBM)
  - XGBoost Classifier (xgboost)
  - Balanced Bagging Classifier(imblearn)
  - Easy Ensemble Classifier (imblearn)
  - Balanced Random Forest Classifier (imblearn)



## Learning Algorithms Scores



+ a b | e a u

Source:

<https://public.tableau.com/view/LearningAlgorithmResults/LearningAlgorithmsScores2.comb>

Web Viewer [Terms](#) | [Privacy & Cookies](#)

Edit



# Learning

E. F. E. F.  
AIGUOPILITIS

## Measure Names

■ Error Rate

False Positive Rate



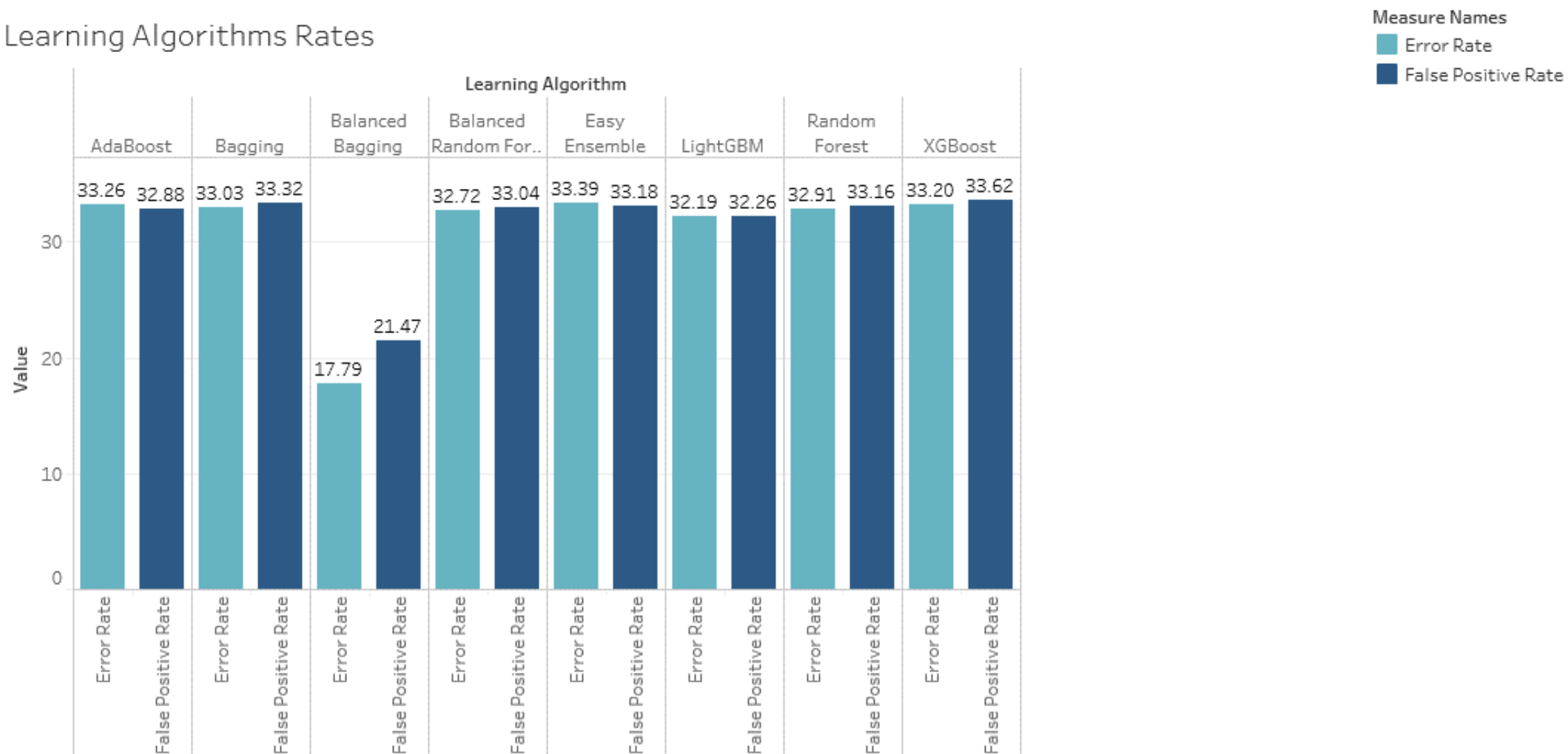
**Source:**

<https://public.tables...>

Edit

[Web Viewer Terms](#) | [Privacy & Cookies](#)

## Learning Algorithms Rates



tableau

Source:

<https://public.tableau.com/views/LearningAlgorithmResults/LearningAlgorithmsRates2>

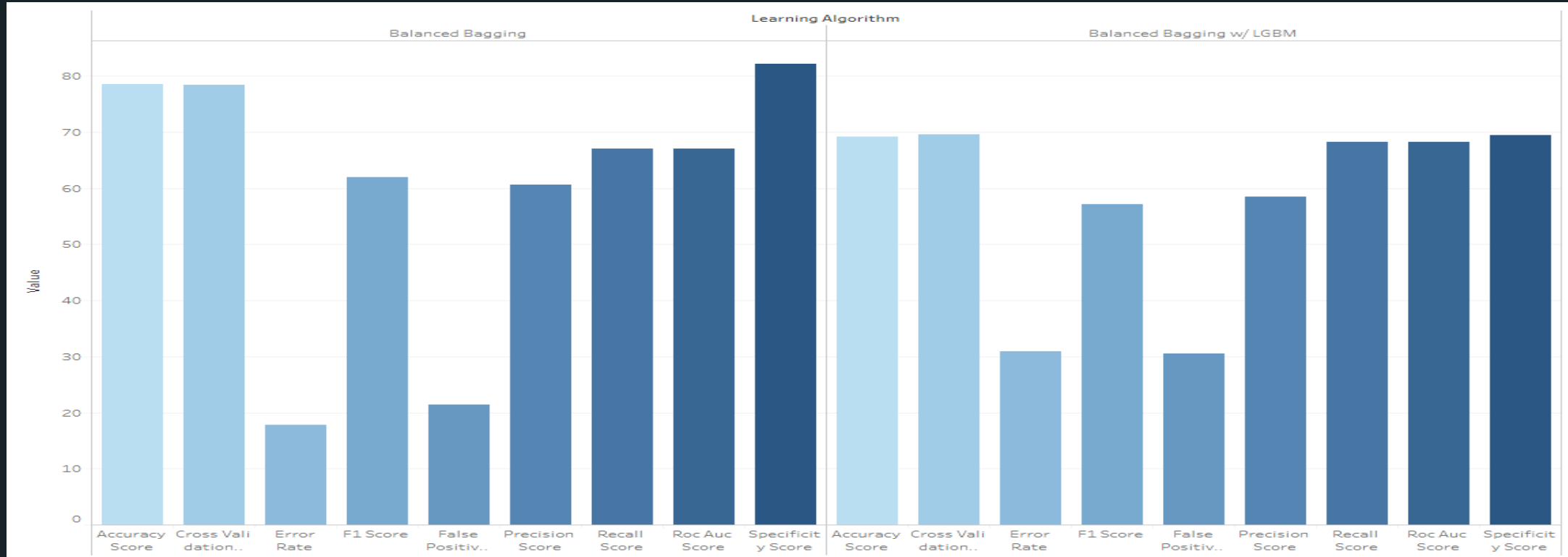
Web Viewer [Terms](#) | [Privacy & Cookies](#)

Edit

# Balanced Bagging Classifier

- Based on the previous visualizations, Balanced Bagging Classifier from imblearn is the algorithm of choice for this data. While some of the scores may have been close, Balanced Bagging Classifier had higher scores in Accuracy, Cross Validation, and Specificity. The algorithm also had the lower Error Rate and False Positive Rates of the group.
- Balanced Bagging Classifier performed the best of the classifiers, however, I was not comfortable with how close its predictions were for Serious Accidents in the confusion matrix. Due to this, I decided to combine Balanced Bagging Classifier with the second highest performing algorithm, LightGBM to see what results I would get.

# Balanced Bagging Classifiers Comparison



The results were better than the other learning algorithms but lower accuracy wise than the previous Balanced Bagging Algorithm. It also took longer than any other algorithm used. Taking all of that into consideration, I have decided that depending on what was the goal, either Balanced Bagging Classifier algorithm could be used. If I were more concerned with overall accuracy, the regular Balanced Bagging Classifier would be used. If I were more concerned with making sure "Serious" predictions were achieved, Balanced Bagging Classifier with LightGBM would be used.

# Expectations vs Reality

## Expectation

- Overall I thought there would be certain features that had a high impact on the severity of accidents.
  - skidding\_and\_overturning
  - time\_of\_day
  - weather\_conditions
  - day\_of\_week

## Reality

- There were very low correlations among features and accident severity. The highest correlation was vehicle\_type at 0.134.

# Limitations

- Not able to obtain accuracy over 70% without causing other issues such as overfitting and bias.
- The data was extremely imbalanced. The majority of accidents were not serious and while this is not a real life problem, it was a problem for this model. Undersampling was done to improve overall scoring.
- More factors surrounding accidents should be included in this data.
  - While there was information on the speed limit in certain areas, there was no information on whether the driver was speeding.
  - No information on cellphone usage of drivers
  - Rural areas had a higher rate of serious accidents which could be correlation to emergency vehicle arrival or distances from hospitals, but this information was also not available
  - No time of arrival for emergency units
  - No info on passengers
- Low correlations