SHAZAD LADHA IBM CAPSTONE PROJECT

Road Traffic Accidents in Leeds 2019

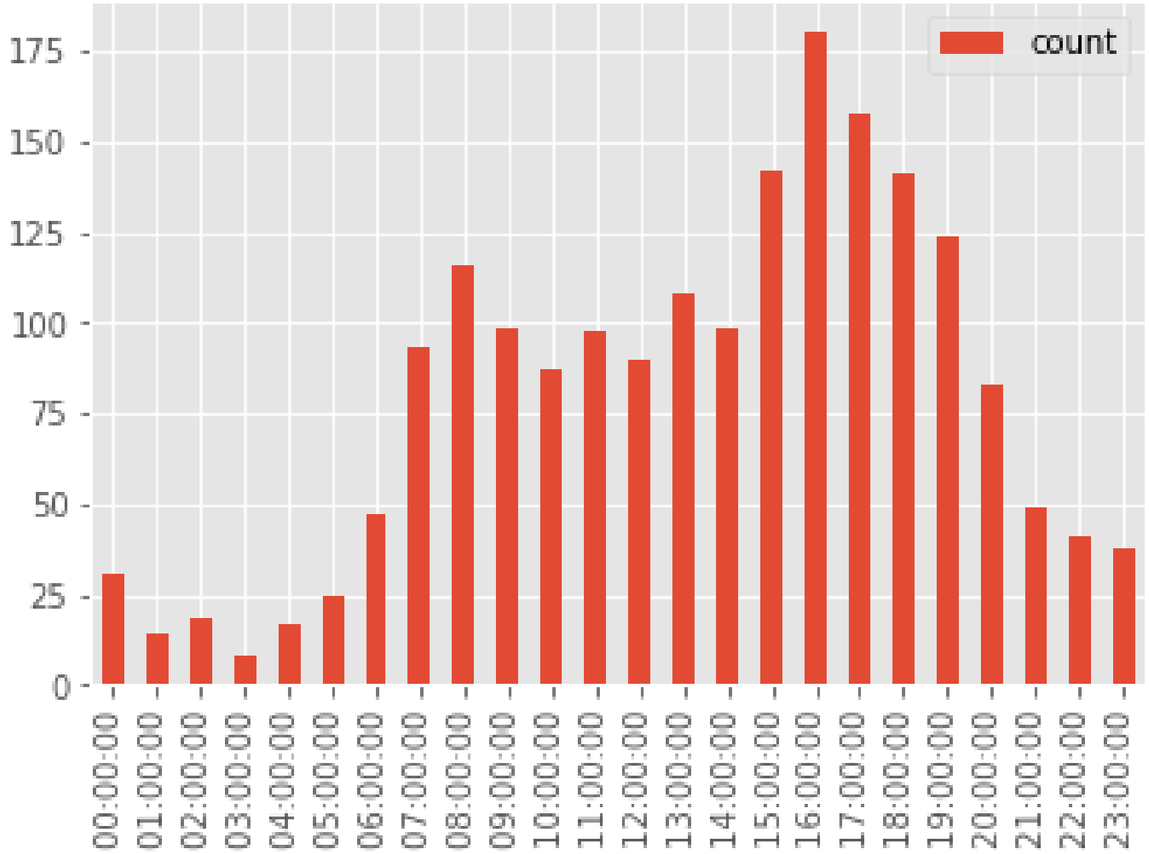
described through data

Presentation Outline

TODAY'S DISCUSSION

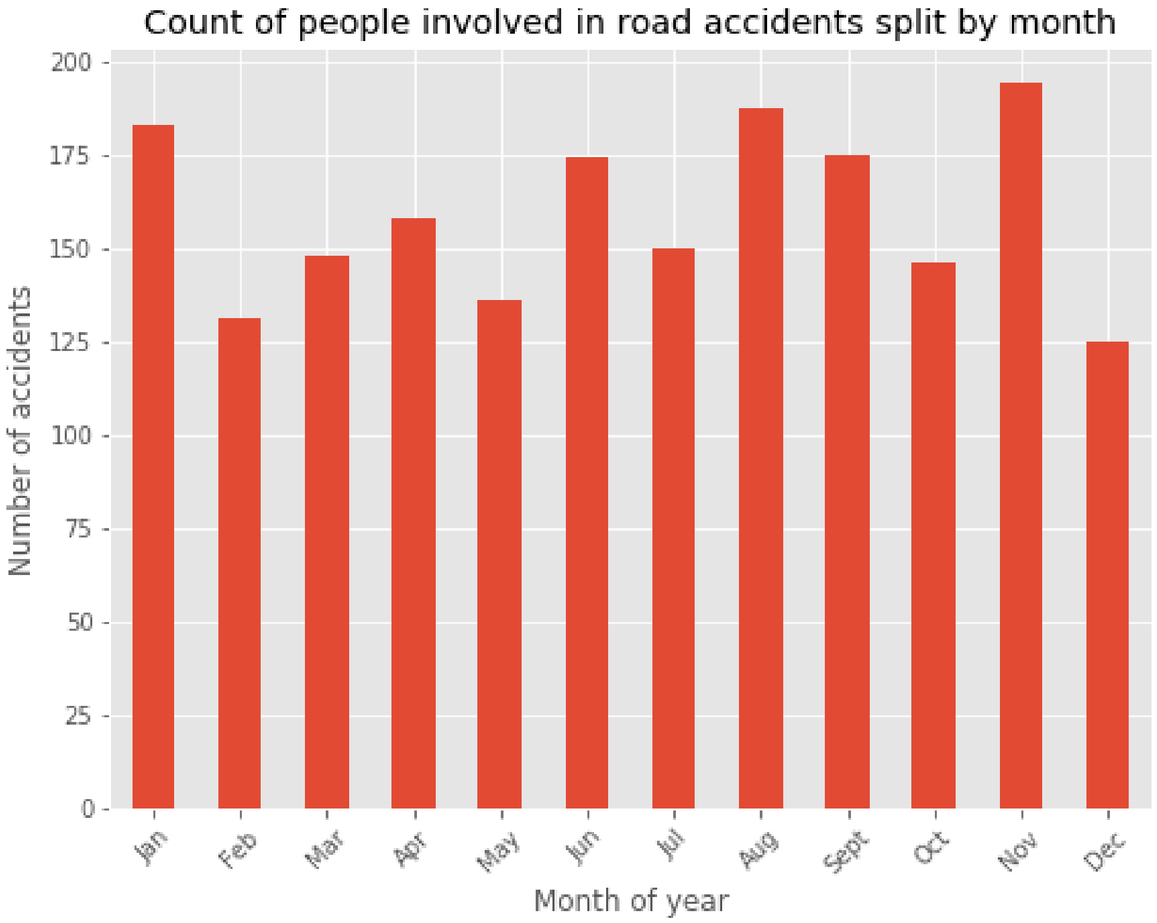
- Trends and features of road accidents
- Predicting the severity of an accident

Time of accident



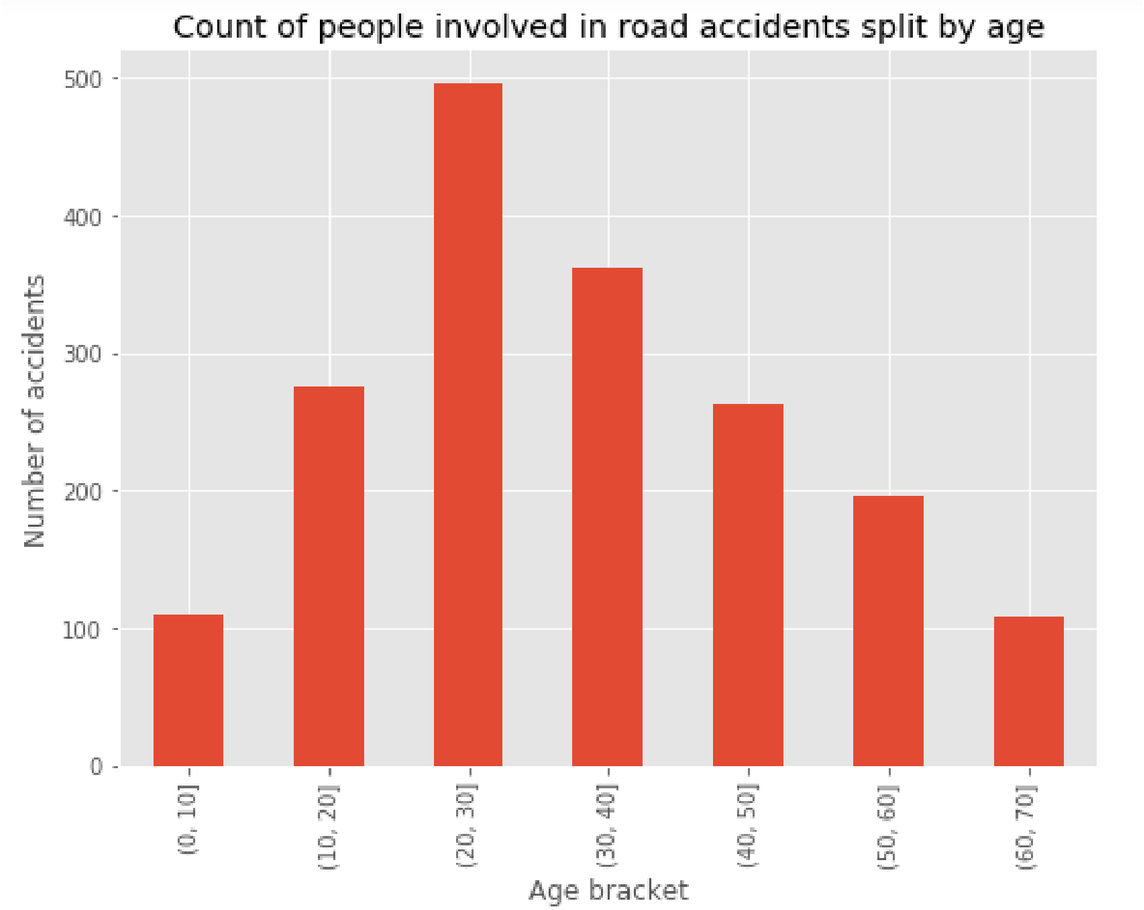
Our natural intuition assumes that most accidents will happen late night/early morning as we match car accidents with sleepless or drunk drivers. This heuristic is called out with data as we see most accidents happen between 4-6pm.

Month of accident



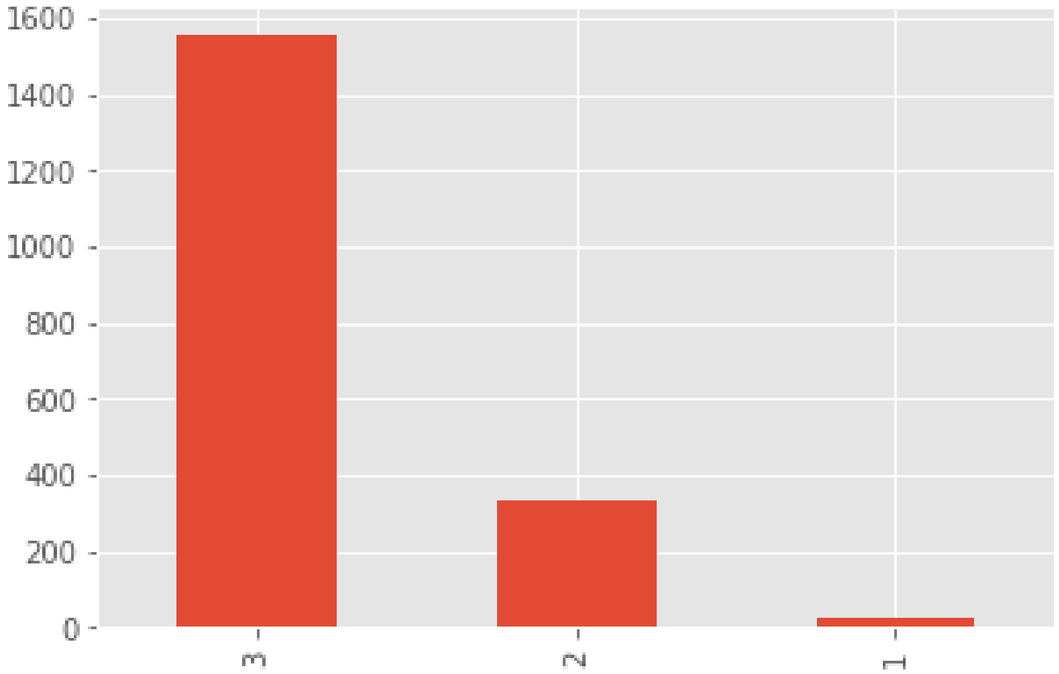
As we can see, November seems to be when most accidents occur, closely followed by August and January. I don't think there's much intuition that can help provide insights into why that is. I can see that December has the lowest incidents, this may be because of people taking time off for christmas and new year which means overall less commuting.

Age of people involved in a crash



Most incidents occur with people in their 20-30s. This explains the high insurance premiums for this age group. the 0-20 age groups are listed as they must be passengers which have been logged in the dataset.

Severity of accidents



With 3 being the least severe accident type, we can see that most accidents are not severe. This will mean that the dataset may not be usable for any ML model given the lack of severe(1) accidents (which total 22)

Using ML Models to predict Severity

A CLASSIFICATION PROBLEM

ML Models that I will use

KNN	DECISION TREE	SVM	LOGISTIC REGRESSION
KNN algorithms use data and classify new data points based on similarity measures	DTs go from observations about an item to conclusions about the item's target value.	SVM is a supervised model with associated learning algorithms that analyze data used for classification analysis.	Logistic regression is a model that in its basic form uses a logistic function to model a binary dependent variable

Model evaluation

KNN	DECISION TREE	SVM	LOGISTIC REGRESSION
Jaccard index=0.8 F1-score=0.71	Jaccard index=0.79 F1-score=0.71	Jaccard index=0.8 F1-score=0.71	Jaccard index=0.8 F1-score=0.71

Conclusion

- All models did performed similarly
- They all performed with quite a high f1-score and jaccard index which indicates that the models do quite well in predicting the outcome of the test data.
- However, given the low count of severe accident data (22 incidents) there is not enough data to prove that these models are reliable. Please look at the confusion matrix for the Decision Tree model to prove this case: