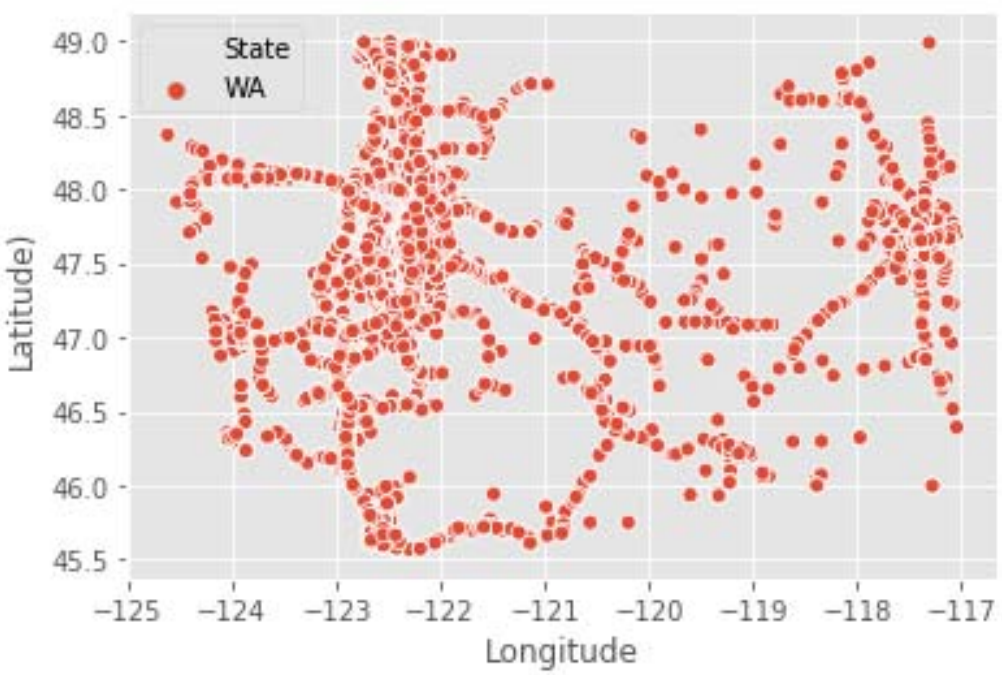
Capstone Project - Car accident severity (Week 2)

Background and Problem

- Road crash fatalities and disabilities recognised as a major public health issue.

- ~ 1.35 million people die and 30-50 million suffer non-fatal injuries in every year globally.
- Road accident data for Washington state has been analyzed.
- Machine learning techniques have been used on road accident data for King county.

Accident visualization - Washington



- The built model could be used for real time accident prediction.

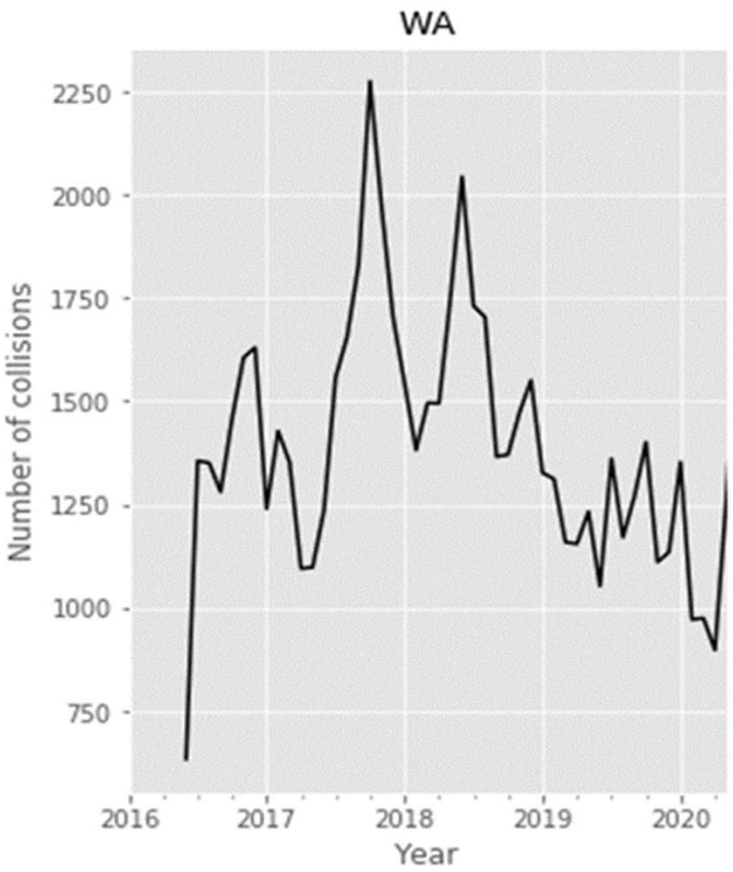
Data acquisition and cleaning

- The traffic collision dataset for 49 states of US was downloaded from Kaggle website (2016 to 2020)[1].
- The raw dataset consists of 3513617 rows and 49 columns.
- Data cleaning was performed.
- 34 features were selected for the improved accuracy of the prediction.

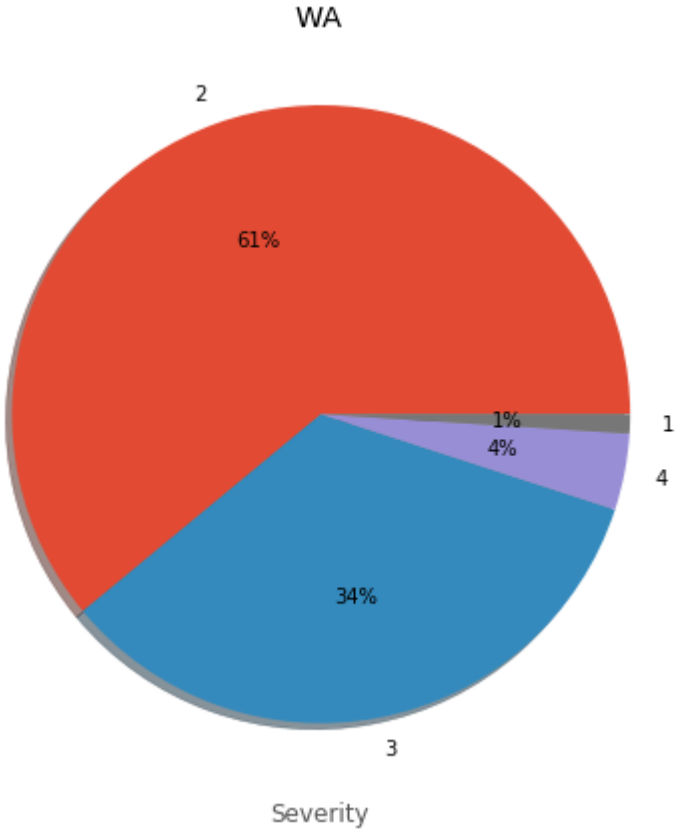
1. https://www.kaggle.com/sobhanmoosavi/us-accidents

Exploratory data analysis

Number of collisions per month



Severity of an accident

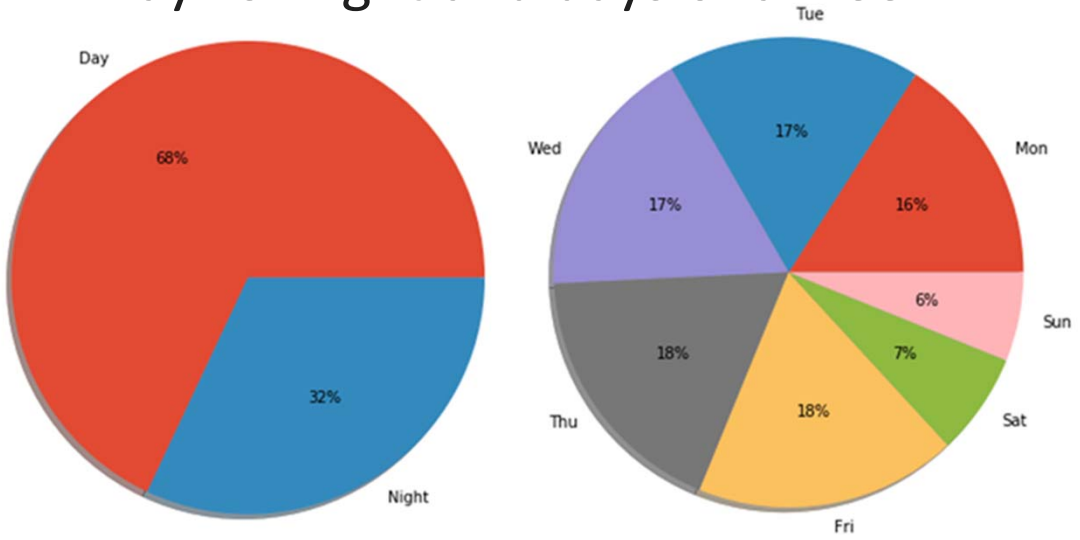


- Most of the accidents are in the severity level 2.

Exploratory data analysis

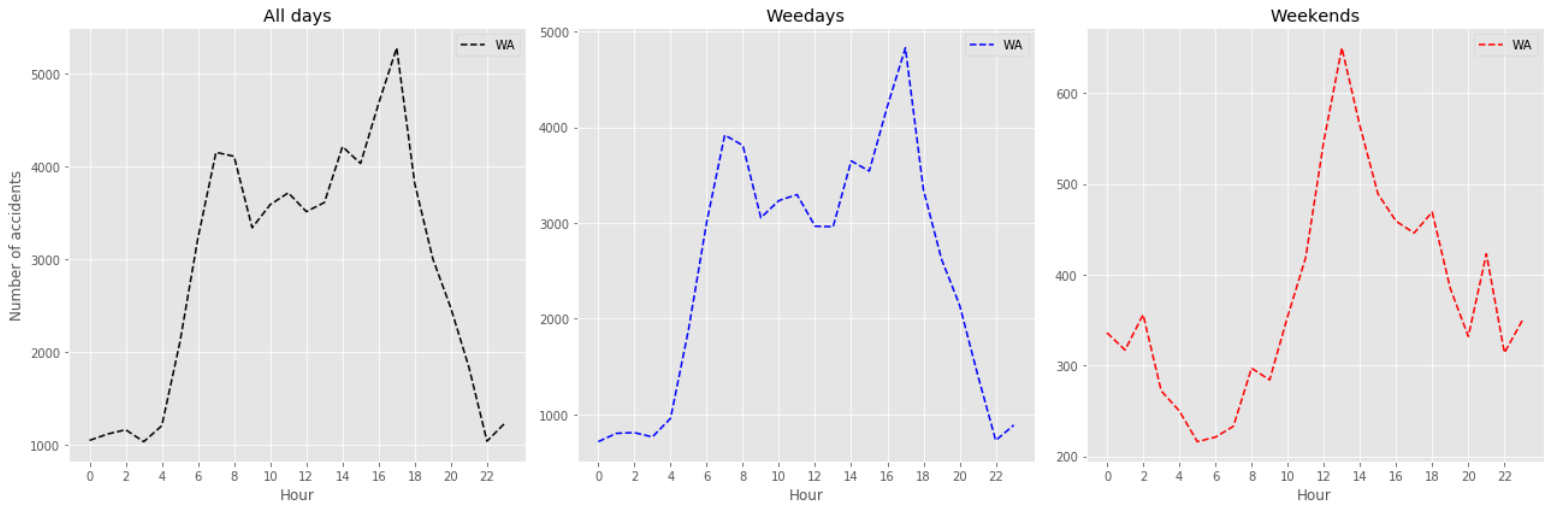
- A higher rate of accidents observed during daytime.
- More accidents are reported during weekdays than weekends.

Day vs. night and days of a week



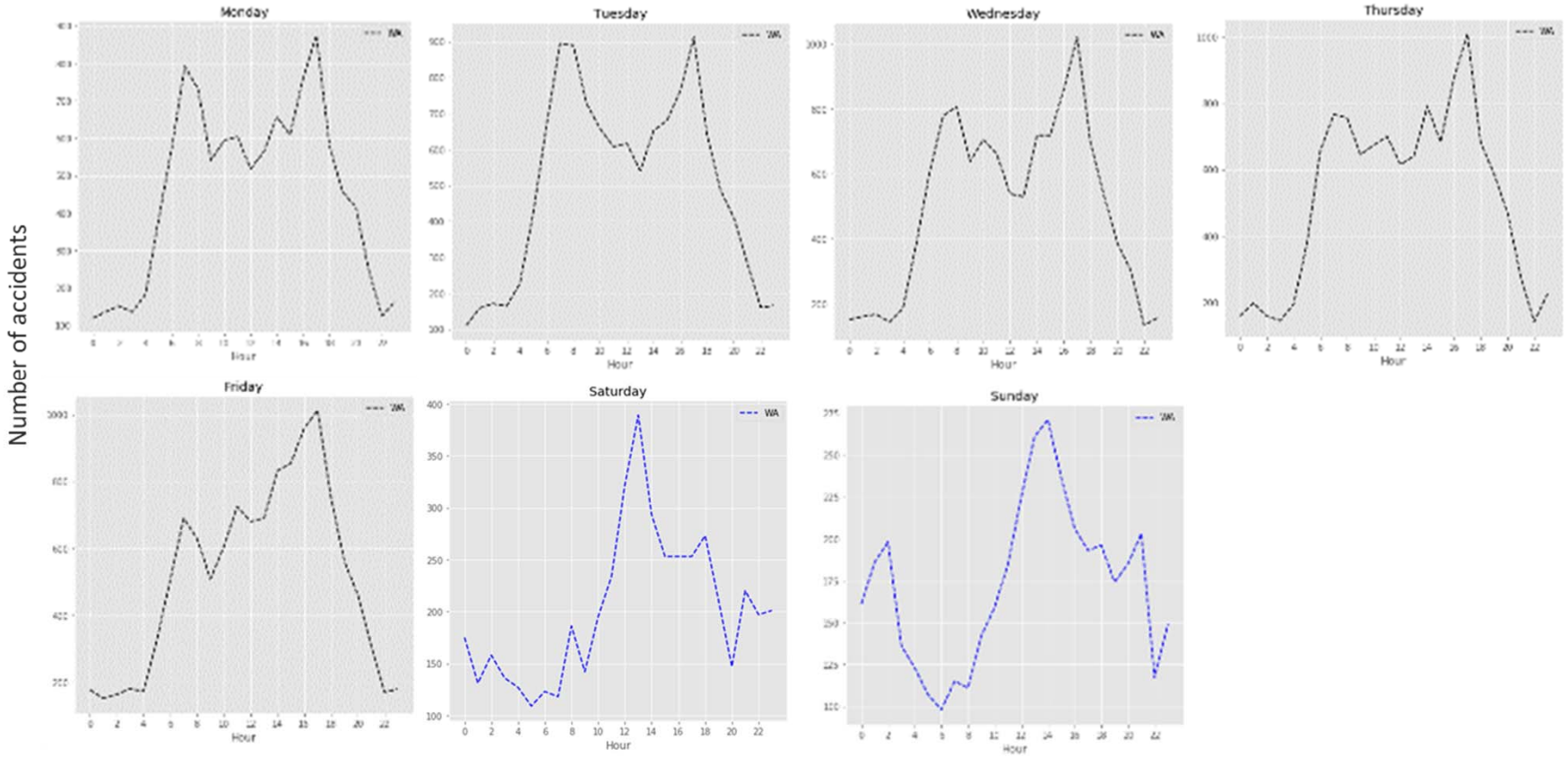
Sunrise_Sunset

Weekday/Weekend



Exploratory data analysis

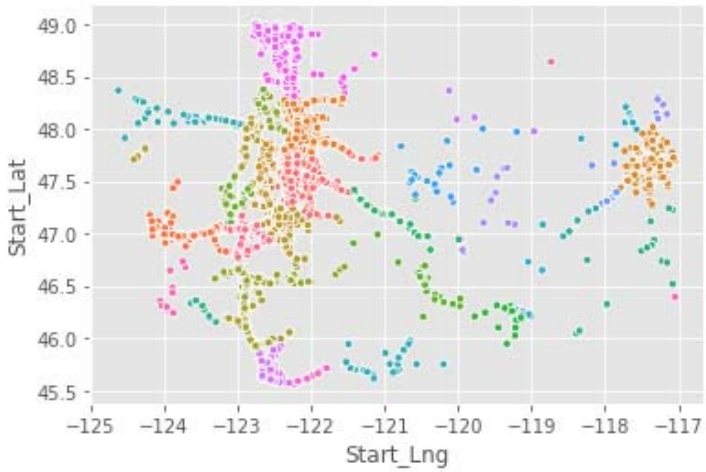
Hourly distribution of accidents

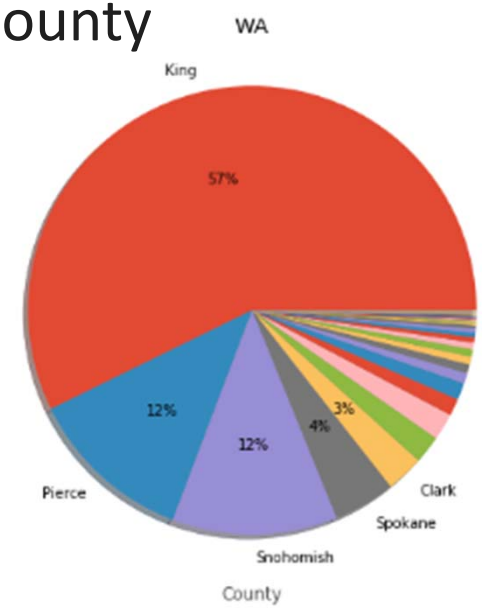


- It is not recommended to travel around 7-8 am and 4-5pm on weekdays.
- Similarly, early afternoon is not recommended for weekends.

Exploratory data analysis

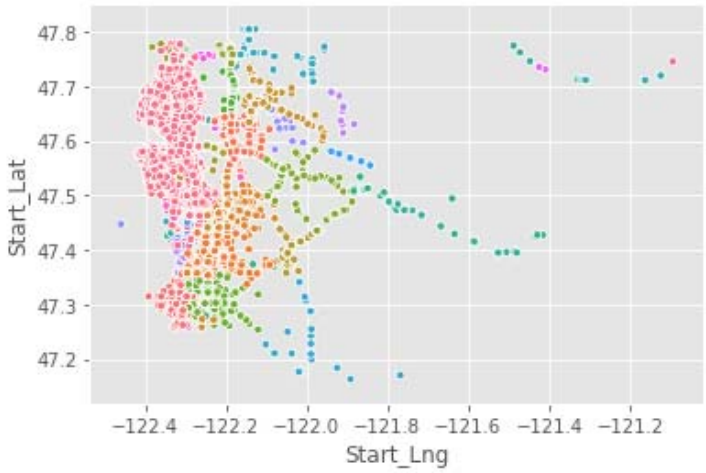
Collisions by county

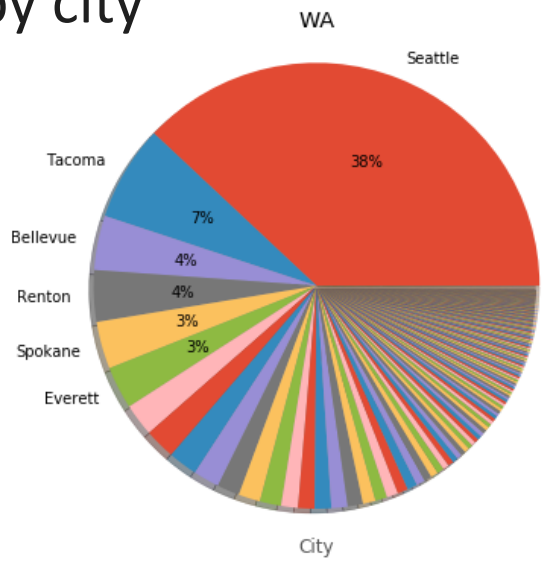




- Highest number of traffic collisions are reported from King county.

Collisions by city

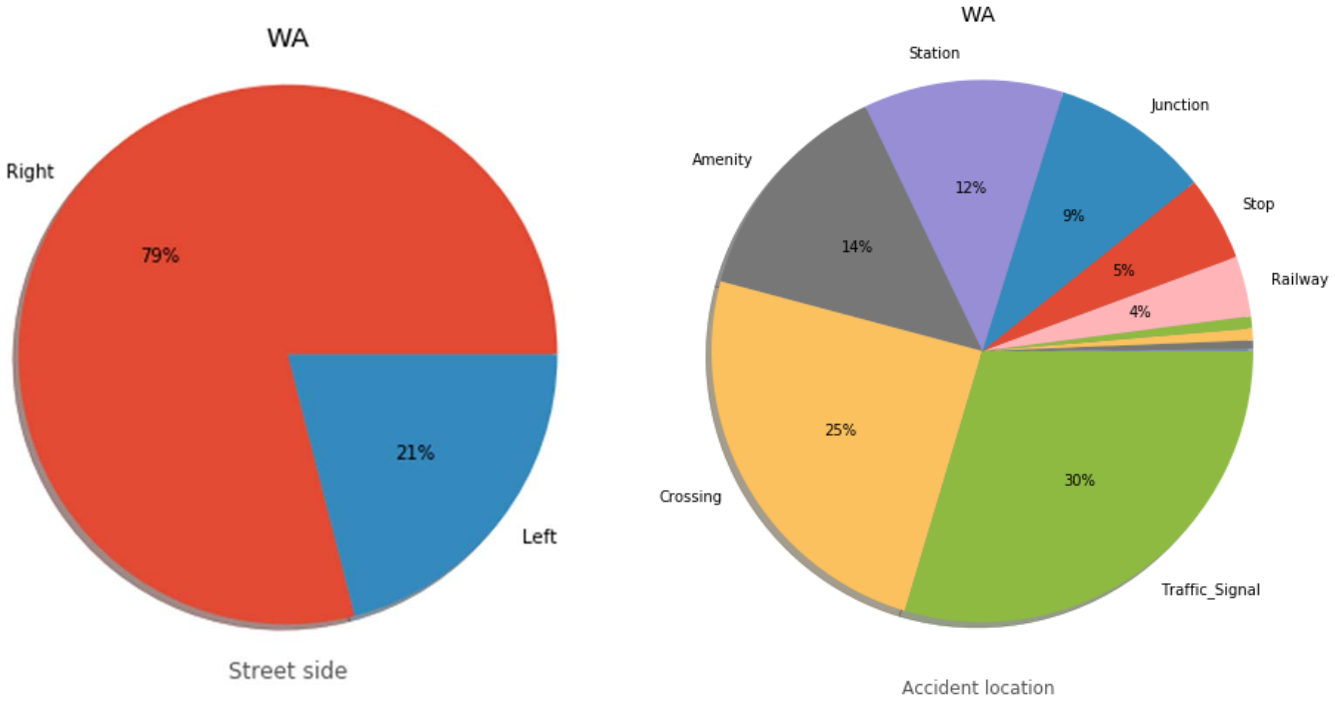




- Seattle is far ahead compared to other cities.

Exploratory data analysis

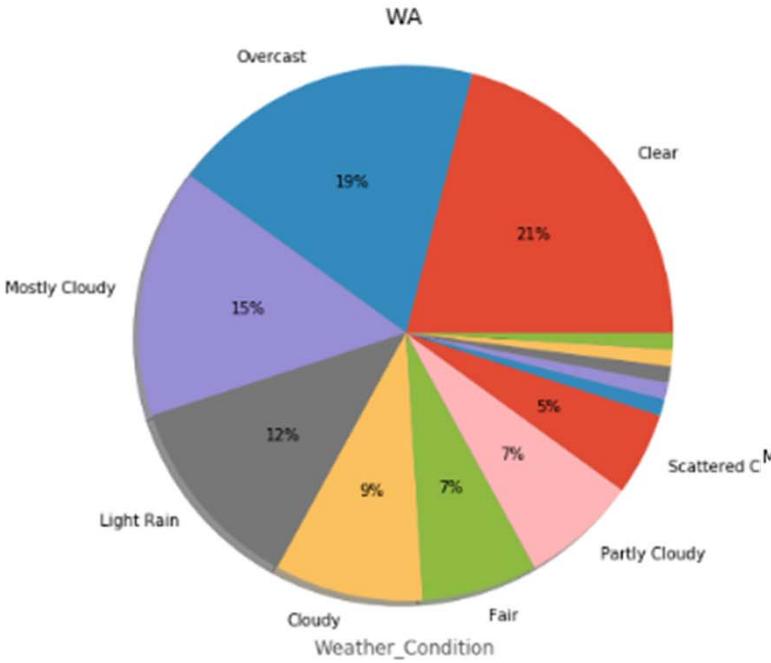
Percentage distribution by street side and location

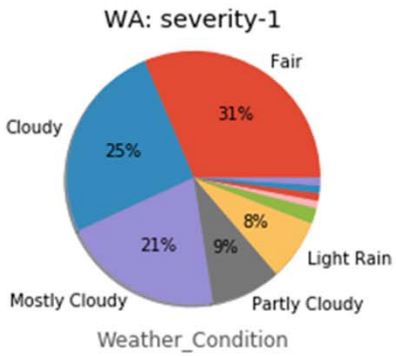


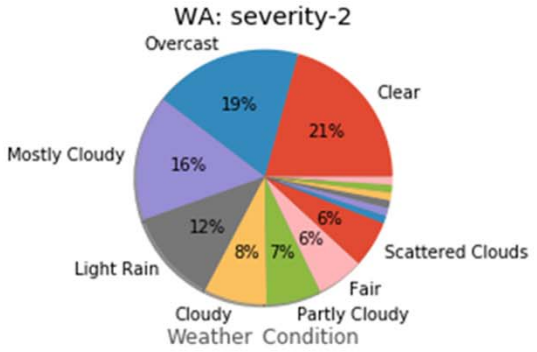
- Majority of the accidents occurred at the right side of the street.
- 30% of accidents are occurring at traffic signal, followed by crossing (25%).

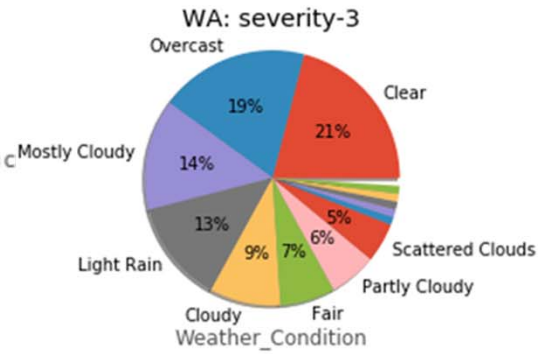
Exploratory data analysis

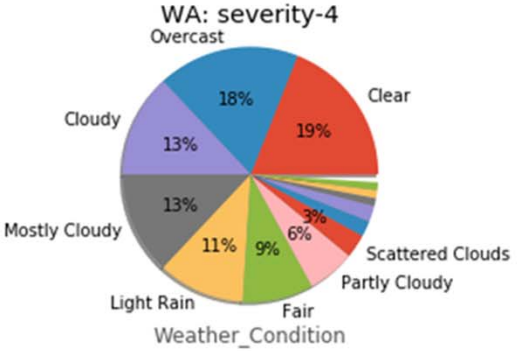
Weather conditions and severity











- 5 topmost weather conditions for accidents are clear, overcast, mostly cloudy, light rain, and cloudy.
- Overcast, mostly cloudy and light rain are realistic factors for accidents.

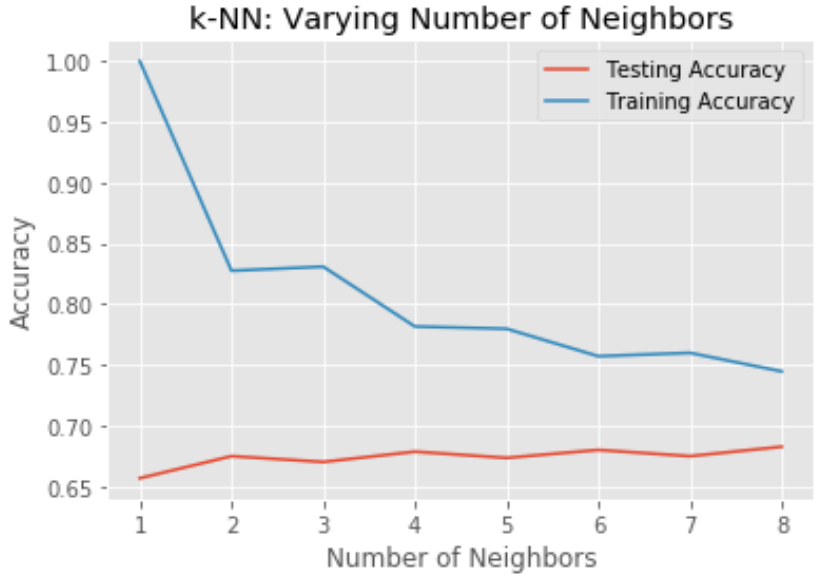
Machine learning algorithms and feature engineering

- Out of 49, 34 features were selected based on their impact on accidents.

feature_lst=['Source','TMC','Severity','Start_Lng','Start_Lat','Distance(mi)','Side','City','County','State','Timezone','Temperature(F)','Humidity(%)','Pressure(in)','Visibility(mi)','Wind_Direction','Weather_Condition','Amenity','Bump','Crossing','Give_Way','Junction','No_Exit','Railway','Roundabout','Station','Stop','Traffic_Calming','Traffic_Signal','Turning_Loop','Sunrise_Sunset','Hour','Weekday', 'Time_Duration(min)']

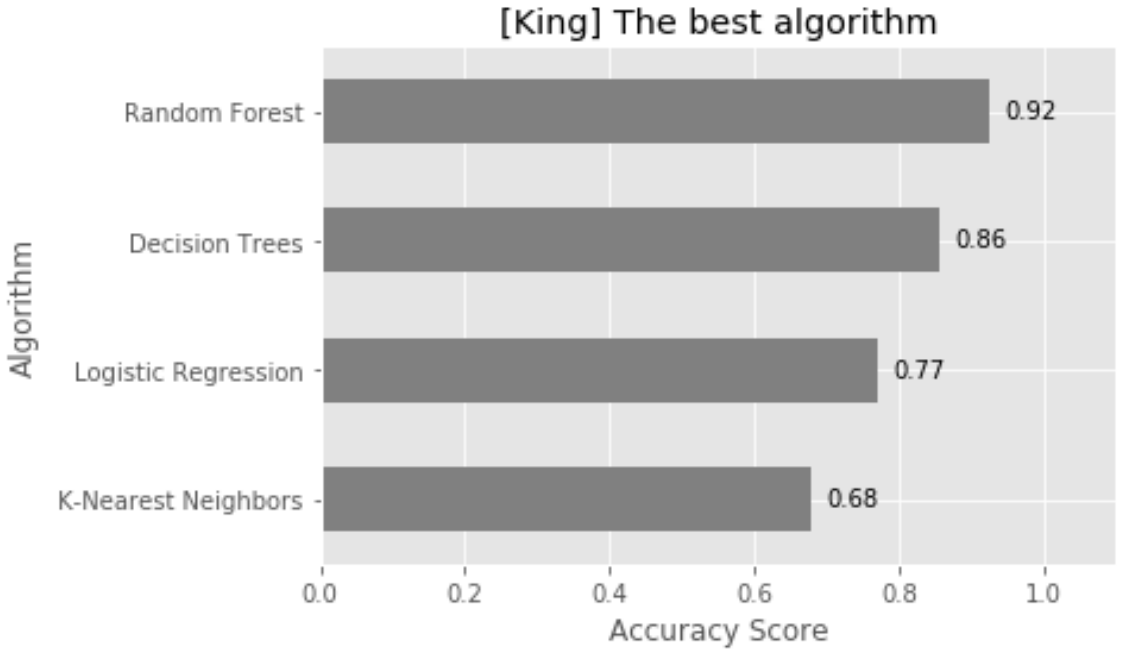
Supervised machine learning algorithms

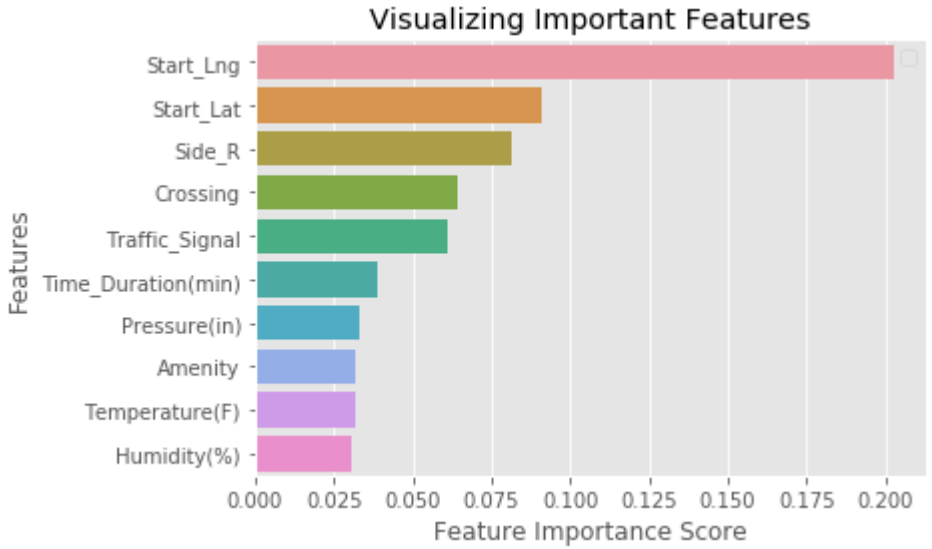
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest Classification



Machine learning algorithms

- Random Forest classification shows the highest accuracy with a score of 0.92.
- KNN shows the least accuracy (0.68).
- Top ten features for prediction of accident severity for King county was extracted using Random Forest classification model.





Conclusion

- Traffic collision data for WA state has been analyzed.
- Factors which could lead to road crash fatalities and disabilities were identified and listed.
- Machine learning algorithms were applied to predict the accident severity for King county and Random Forest classification shows the highest accuracy .
- There is a room for improvement for accuracy of the models .
- A full weather data could give a more realistic picture and also increase the accuracy of the models.

Acknowledgment

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.