Predicting Car Accident Severity in Seattle

BASED ON DATA COLLECTED FROM 2004 TO MID-2020

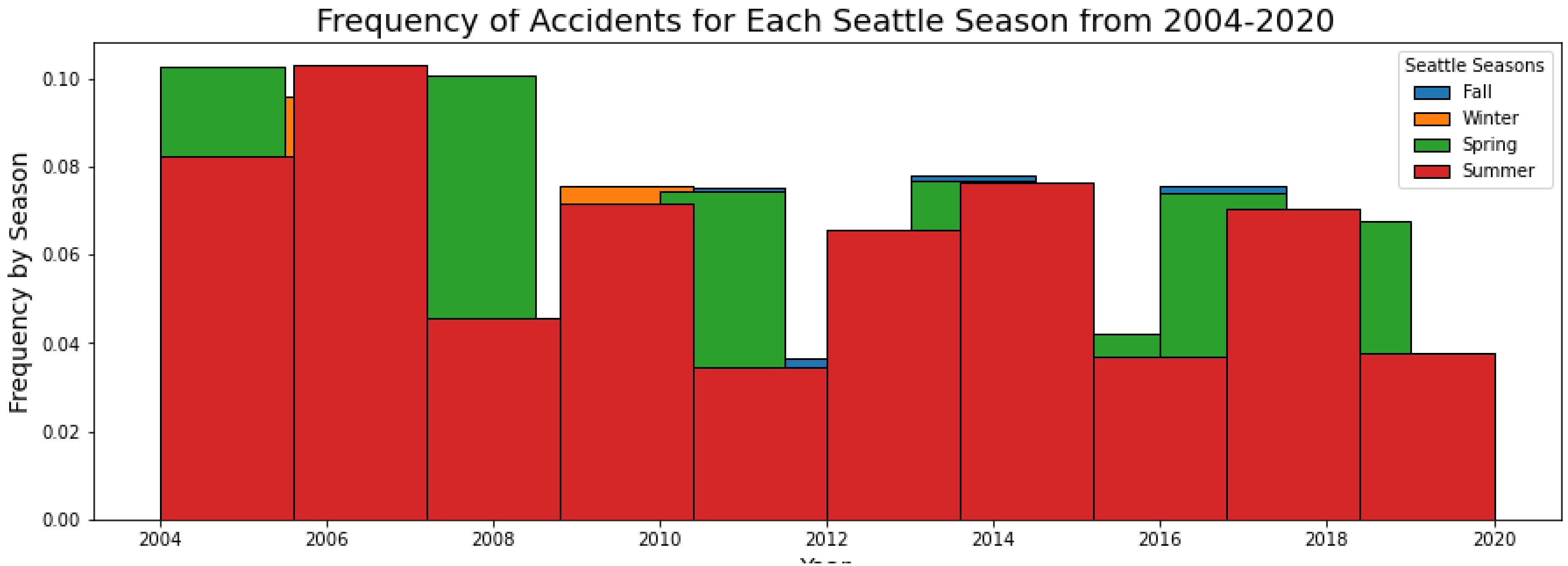
DOUG SMITH

Overview

There were 194,674 reported collisions between 2004 and 2020 in the Seattle Metropolitan Area. There is an opportunity to review common characteristics, including the time of day, weather, road and lighting conditions, geographic location, types of vehicles involved, presence of impairment by drugs or alcohol of individuals involved, among other factors to determine the severity of bodily harm of the individuals involved in the associated collisions via a predictive model.

Source data and associated code for analysis - https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/c18c2989-82e6-47e8-a8e1-cd8439e5a605/view?access_token=9aa75c4df9440c04c755e078047d02fe295c4c270b82ba6872b084e40acabae2

Observations – Numbers on the decline



Warmer months seen in Spring and Summer have the highest rate of reported accidents. This could be based on increased travel and the number of visitors to the region.

What data we should explore

Reviewing available data for predicting a severity code the following variables were selected:

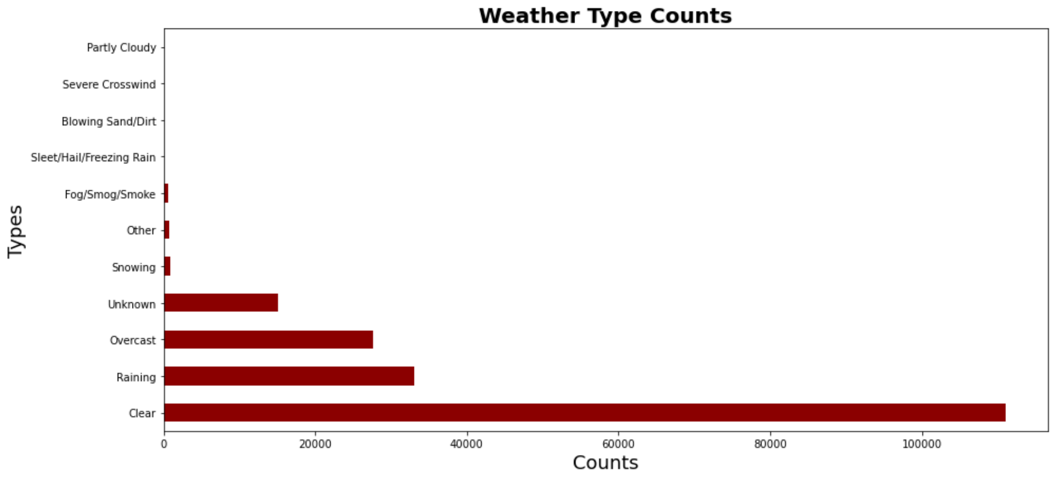
1. Weather Condition
2. Road Condition
3. Light Condition
4. Road Junction type

These environment variables were selected as the variability of each could predict the severity of a potential accident

Weather Type – Clear conditions are present for most accidents

```
weather = accidents.WEATHER.value_counts()
print(weather)
create_barh_plot(weather, 'Weather Type Counts', 'weather_counts', 'darkred')
```

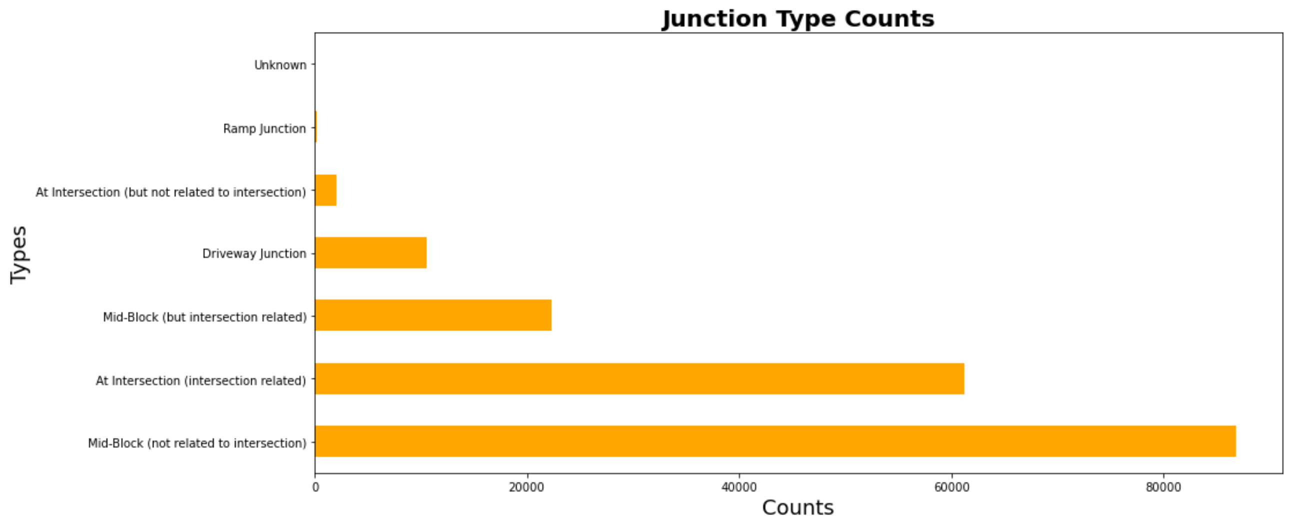
```
Clear                      111008
Raining                     33117
Overcast                    27681
Unknown                     15039
Snowing                       901
Other                         824
Fog/Smog/Smoke                569
Sleet/Hail/Freezing Rain      113
Blowing Sand/Dirt              55
Severe Crosswind               25
Partly Cloudy                   5
Name: WEATHER, dtype: int64
```



Junction Types – Mid-block accidents are common areas of accidents

In [51]:
```
JUNCTIONTYPE = accidents.JUNCTIONTYPE.value_counts()
print(JUNCTIONTYPE)
create_barh_plot(JUNCTIONTYPE, 'Junction Type Counts', 'Junction_counts', 'orange')
```

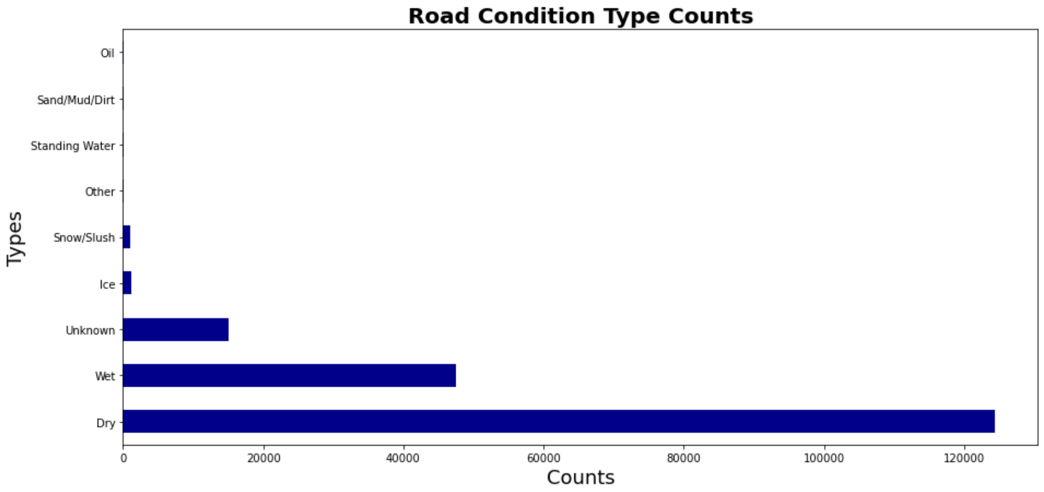
```
Mid-Block (not related to intersection)            86856
At Intersection (intersection related)             61241
Mid-Block (but intersection related)               22353
Driveway Junction                                  10520
At Intersection (but not related to intersection)   2057
Ramp Junction                                        162
Unknown                                                7
Name: JUNCTIONTYPE, dtype: int64
```



Road Conditions – Dry Conditions are most present

```
In [14]: roadcond = accidents.ROADCOND.value_counts()
print(roadcond)
create_barh_plot(roadcond, 'Road Condition Type Counts', 'roadcond_counts', 'darkblue')
```

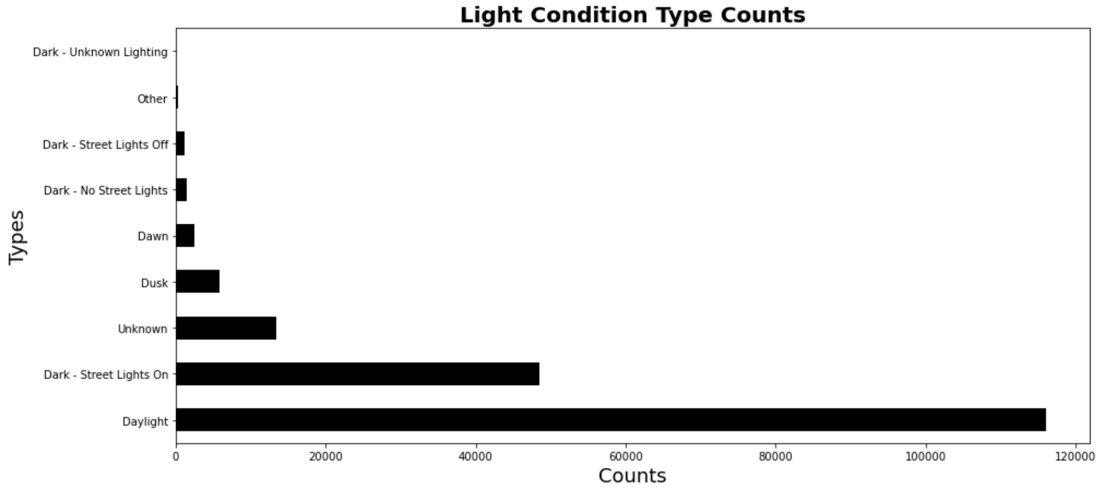
```
Dry               124300
Wet                47417
Unknown            15031
Ice                 1206
Snow/Slush           999
Other                131
Standing Water       115
Sand/Mud/Dirt         74
Oil                   64
Name: ROADCOND, dtype: int64
```



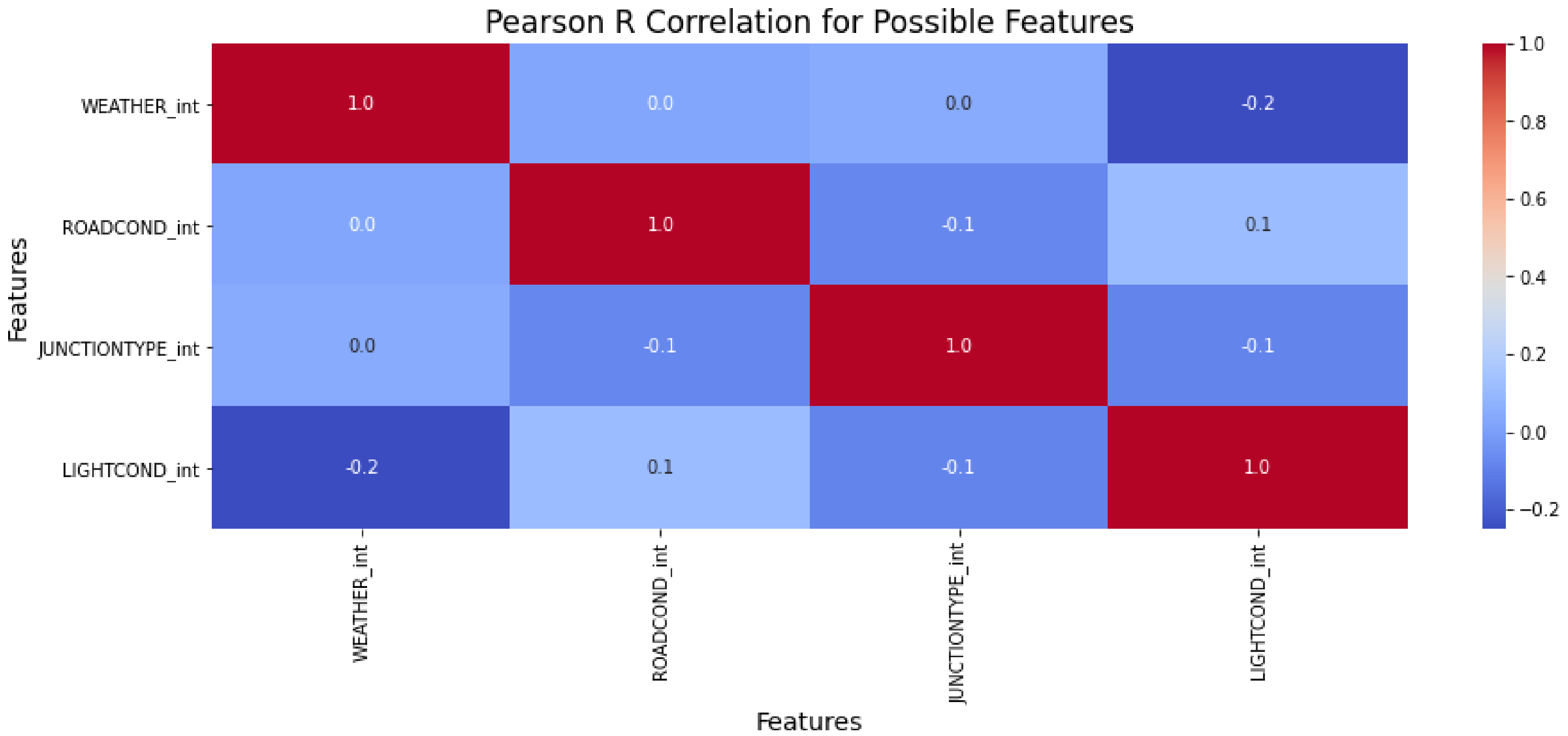
Light Conditions – Most accidents occur in broad daylight

```
lightcond = accidents.LIGHTCOND.value_counts()
print(lightcond)
create_barh_plot(lightcond, 'Light Condition Type Counts', 'lightcond_counts', 'black')
```

```
Daylight                    116077
Dark - Street Lights On      48440
Unknown                      13456
Dusk                          5889
Dawn                          2502
Dark - No Street Lights       1535
Dark - Street Lights Off      1192
Other                          235
Dark - Unknown Lighting         11
Name: LIGHTCOND, dtype: int64
```



How does it all relate?



After classifying variable outcomes by rank order into integers. Running a Pearson R Correction of the variables highlights a small to no strength of association amongst variables

What doe models tell us? – A fair amount, but not all

Decision Tree

	precision	recall	f1-score	support
1	0.69	1.00	0.82	25330
2	0.35	0.00	0.01	11310
accuracy			0.69	36640
macro avg	0.52	0.50	0.41	36640
weighted avg	0.58	0.69	0.57	36640

K- Nearest Neighbor

	precision	recall	f1-score	support
1	0.70	0.83	0.76	25330
2	0.34	0.20	0.25	11310
accuracy			0.64	36640
macro avg	0.52	0.52	0.51	36640
weighted avg	0.59	0.64	0.60	36640

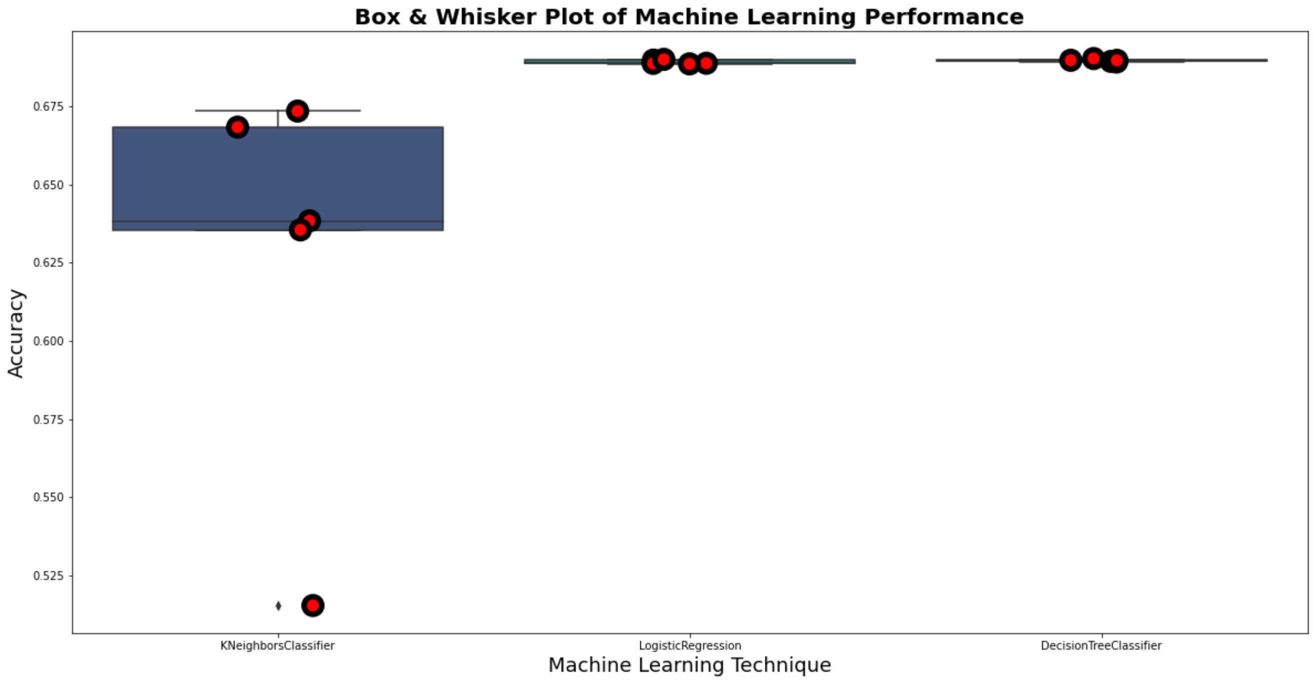
Logistic Regression

	precision	recall	f1-score	support
1	0.69	1.00	0.82	25330
2	0.33	0.00	0.01	11310
accuracy			0.69	36640
macro avg	0.51	0.50	0.41	36640
weighted avg	0.58	0.69	0.57	36640

- **Precision** quantifies the number of positive class predictions that actually belong to the positive class.
- **Recall** quantifies the number of positive class predictions made out of all positive examples in the dataset.
- **F-Measure** provides a single score that balances both the concerns of precision and recall in one number.

The Decision Tree and Logistic Regression score similarly, but the Decision Tree has a slight edge at with a F-1 Score of.567 vs .566

Visualizing the models



The variability of K-nearest nearest is notable when compared with the Logistic and Decision Tree Models.

Average Accuracy by Model

Decision Tree Classifier 0.698437

Logistic Regression 0.697983

K neighbors Classifier 0.685582

Takeaways

Opportunities exist to improve the accuracy of the models. A few other areas of research could be:

1. Geographic location for “clusters” of accidents
2. Appending the congestion or traffic of a roadway
3. Review of the days of the week that correlate with accidents

The existing model however, could be a useful addition to real-time GPS direction applications on phones and within cars to notify a driver when the presence of adverse conditions are present to increase their caution and alertness while driving to mitigate being a victim of a severe accident.