Coursera Capstone Project Report

Ganesh Murdeshwar
Oct 12, 2020

Project Objective

This project will use collision severity data from Seattle in order to better understand what influences the severity of collisions.

In particular, the data will be used to understand what influences collision severity *at intersections*.

The goal will be to provide a deeper understanding of the potential risk factors at intersections, in order to potentially help improve current intersections and/or better design future intersections for safety.

Target Audience

The primary target audience for this project is city council, city planners, and traffic planners and engineers.

The goal as previously stated is for the model outputs to be used in influencing future intersection design to make for safer intersections. Additionally, the model could be used by that audience to address particularly dangerous existing intersections, although specific intersections are not assessed in this particular study.

The Machine Learning Problem

The goal of the associated machine learning problem is to take the available data, clean and select atributes, and generate a predictive model for collision severity.

The model will be a *classification* model.

Since the goal is to identify the most important factors, the model will need to have a way to assess importance of specific attributes to the final classification.

The available data was retrieved here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The detailed description and metadata of the dataset can be found here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Project Plan

1. Assess the state and completeness of the data

2. Identify the key attributes which may potentially influence collision severity
2a. Clean and transform the data as necessary

3. Build several machine learning models and assess for accuracy in predicting collision severity

4. Use attribute importance within the most accurate model to prioritize factors that most strongly influence collision severity. As collision severity is a classification problem, this will direct the choices of algorithms.

5. Make preliminary investigative recommendations on possible intersection modifications (current and future)

Data Extract Summary

The dataset for this project is collision data from the City of Seattle, was downloaded from: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Detailed information and metadata regarding this dataset is here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

There are 38 columns, three of which represent the collision severity (code, code.1, and description). This leaves 35 columns of potential predictive attributes. There are 194,673 rows of data. The data ranges in date from Jan 1 2004 to May 2020.

When looking only at data on collisions at intersections, the dataset reduces in size to 65070 rows.

The dataset has collision severity listed as one of: 3—fatality; 2b—serious injury; 2—injury; 1—property damage; 0—unknown.

However, in this particular dataset, only severity levels of 1 and 2 are to be found. This simplifies the classification task as single classification can be used. However, in future datasets, the additional (potential) complication of a more complex multi-class situation needs to be accounted for.

Preliminary Data Analysis

The data has the following columns: SEVERITYCODE X Y OBJECTID INCKEY COLDETKEY REPORTNO STATUS ADDRTYPE INTKEY LOCATION EXCEPTRSNCODE EXCEPTRSNDESC SEVERITYCODE.1 SEVERITYDESC COLLISIONTYPE PERSONCOUNT PEDCOUNT PEDCYLCOUNT VEHCOUNT INCDATE INCDTTM JUNCTIONTYPE SDOT_COLCODE SDOT_COLDESC INATTENTIONIND UNDERINFL WEATHER ROADCOND LIGHTCOND PEDROWNOTGRNT SDOTCOLNUM SPEEDING ST_COLCODE ST_COLDESC SEGLANEKEY CROSSWALKKEY HITPARKEDCAR

The majority of this data is categorical in nature.

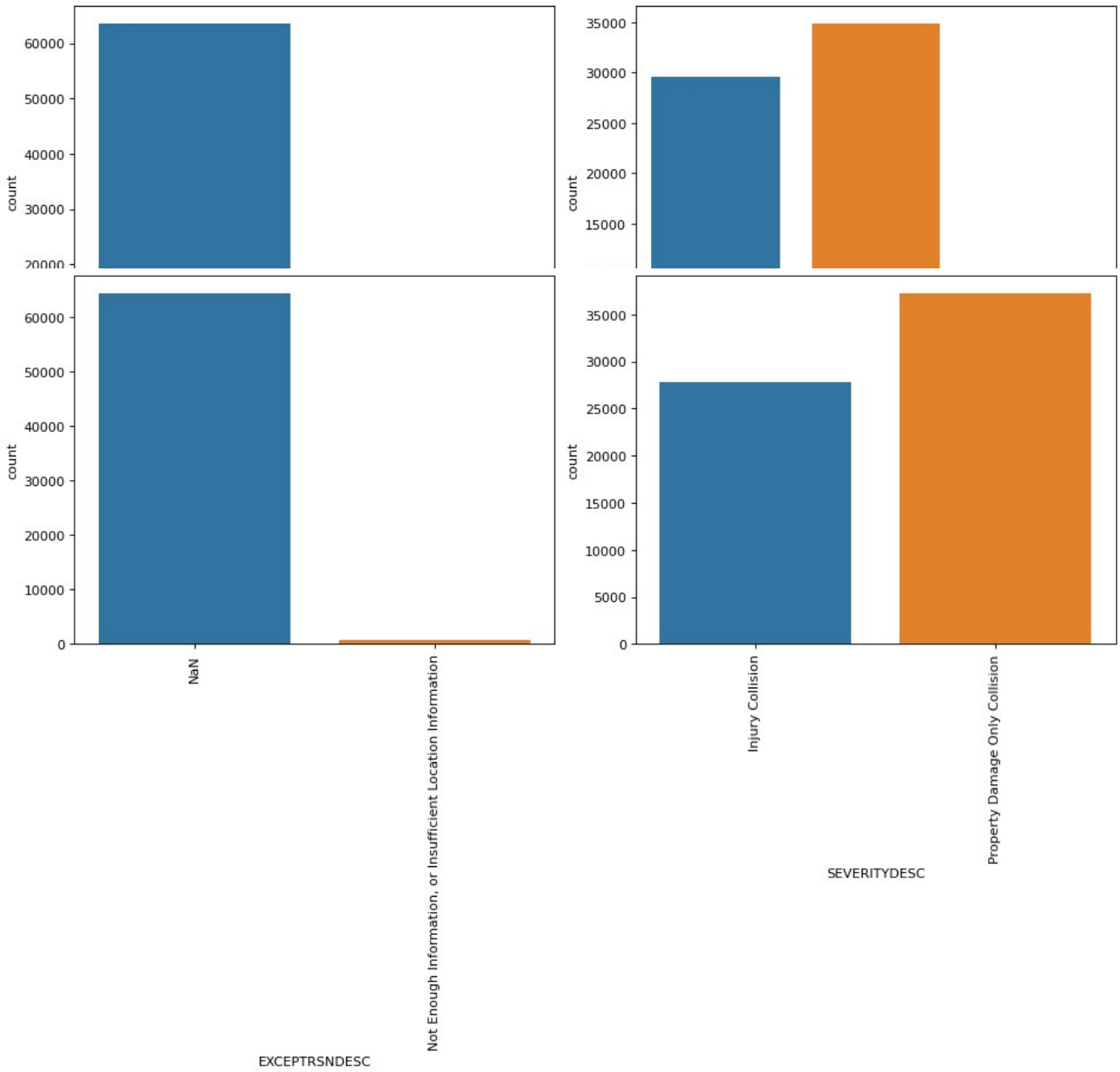
A number of the fields are administrative, such as OBJECTID, INCKEY, and related items.

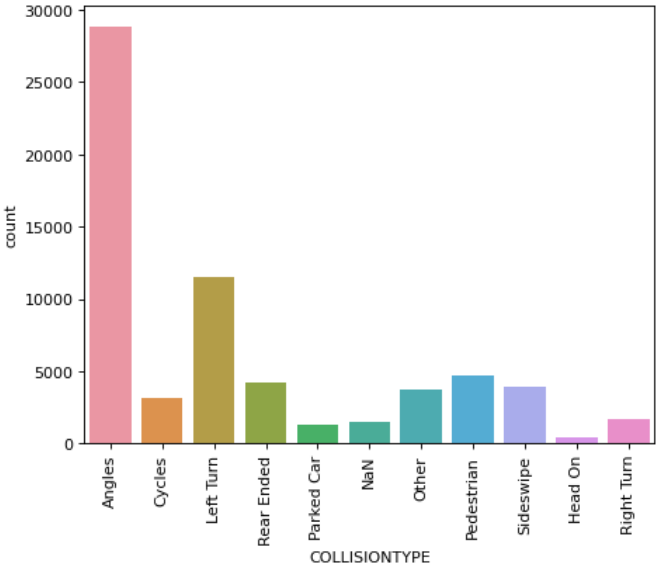
For the purposes of this analysis, only attributes likely to contribute to *intersection design* as it relates to collision severity need be considered. So for example the date of the collision and location of the intersection are not relevant.

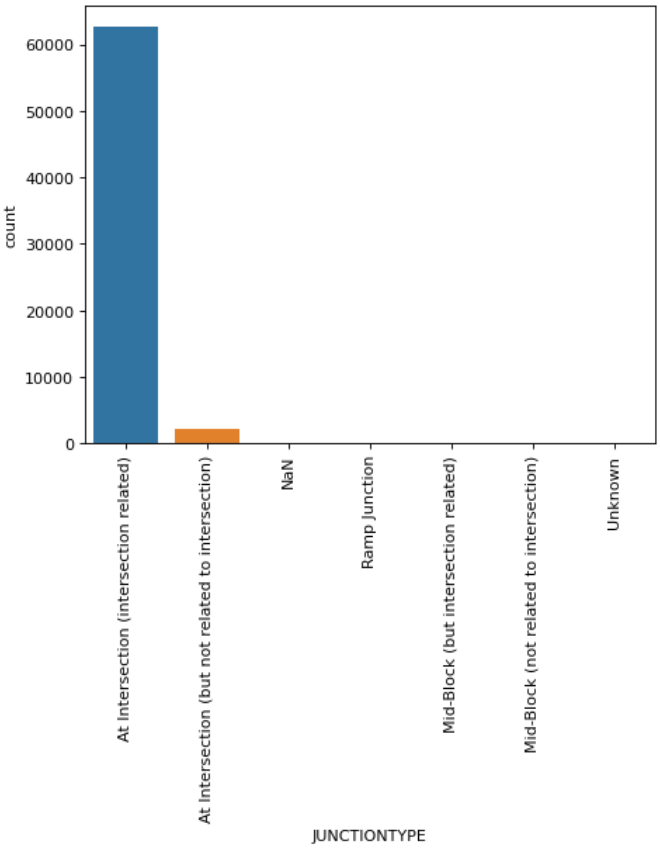
Methodology: Data Exploration

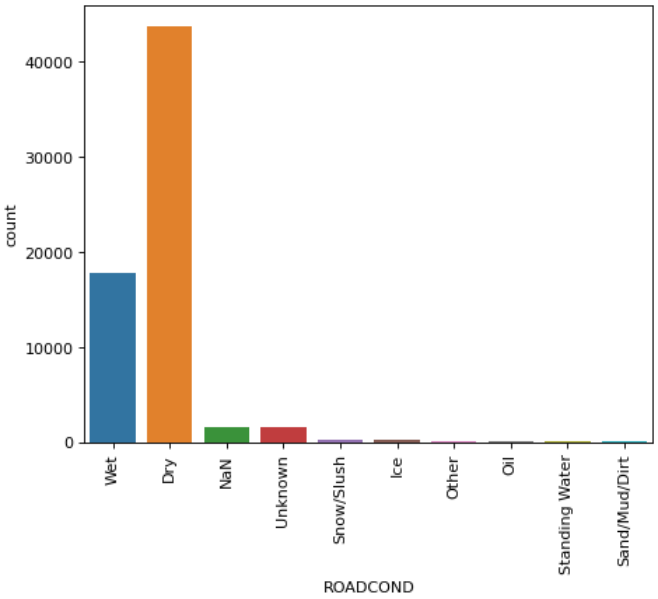
The following charts are generated by the code associated with this project.

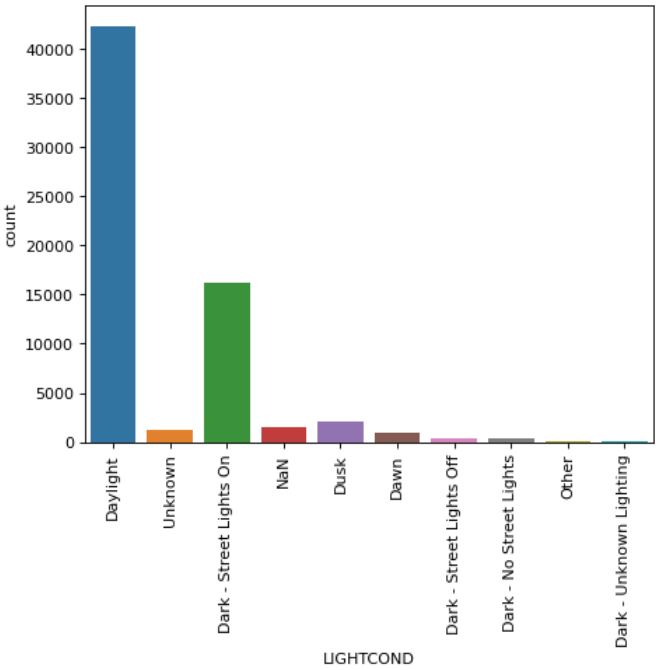
The categorical attributes are quite ‘messy’, and many are “polluted” with an empty (nan) field. The following are a number of representative graphs showing the distribution of values within the attributes. (This is not a complete set of graphs, just the first eight attributes; the full set is generated and shown within the Jupyter notebook).




count
COLLISIONTYPE
Angles
Cycles
Left Turn
Rear Ended
Parked Car
NaN
Other
Pedestrian
Sideswipe
Head On
Right Turn


count
JUNCTIONTYPE
At Intersection (intersection related)
At Intersection (but not related to intersection)
NaN
Ramp Junction
Mid-Block (but intersection related)
Mid-Block (not related to intersection)
Unknown


count
ROADCOND
Wet
Dry
NaN
Unknown
Snow/Slush
Ice
Other
Oil
Standing Water
Sand/Mud/Dirt

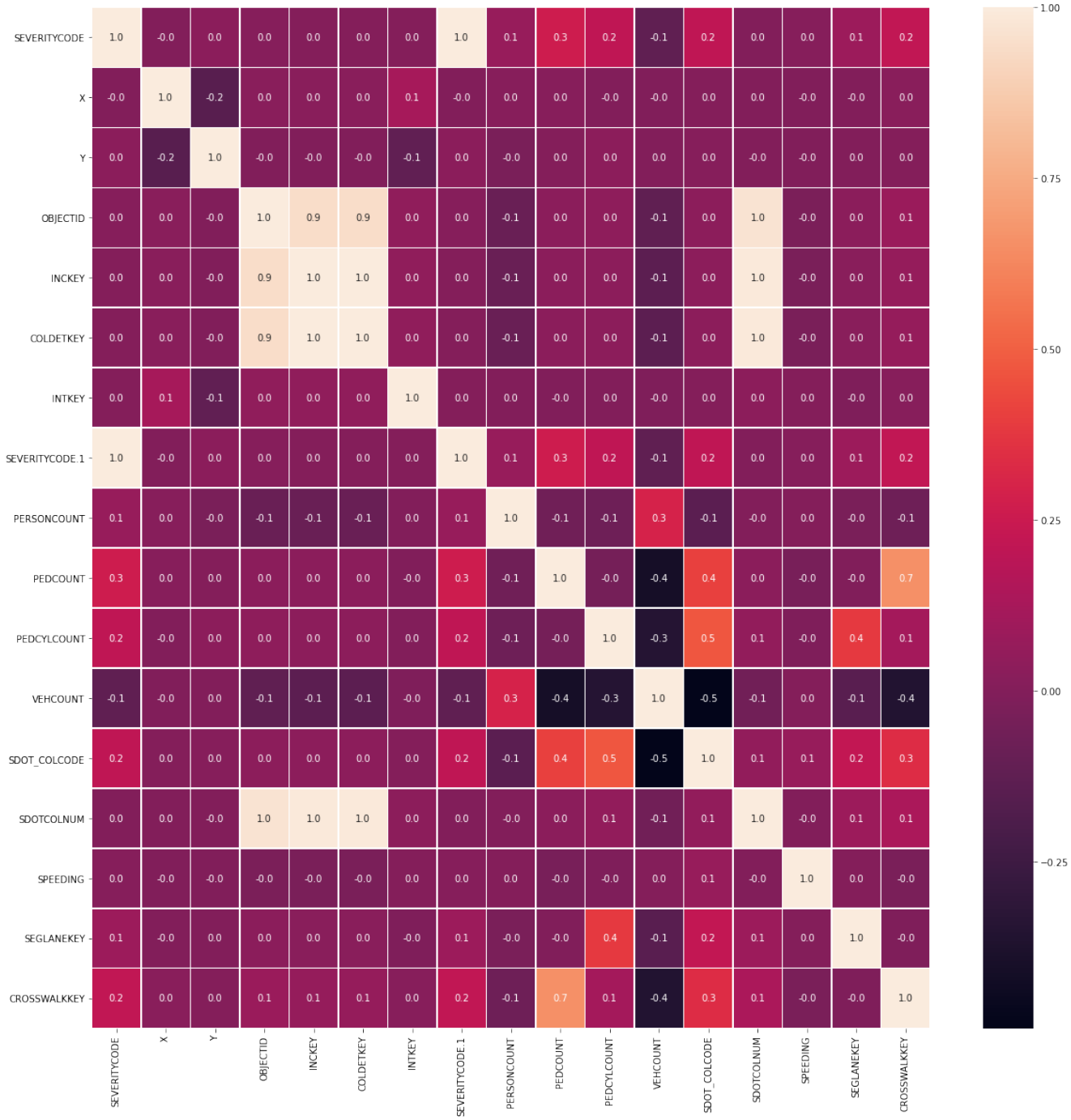

count
LIGHTCOND
Daylight
Unknown
Dark - Street Lights On
NaN
Dusk
Dawn
Dark - Street Lights Off
Dark - No Street Lights
Other
Dark - Unknown Lighting

For the numeric fields, the (partial) numeric description as produced by Pandas is as follows:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	INTKEY	SEVERITYCODE.1	PERSONCOUNT
count	65070.000000	64748.000000	64748.000000	65070.000000	65070.000000	65070.000000	65070.000000	65070.000000	65070.000000
mean	1.427524	-122.330274	47.622226	112007.559751	146575.658475	146812.308821	37558.450576	1.427524	2.571984
std	0.494723	0.028317	0.055691	63519.834795	89063.280106	89445.065377	51745.990273	0.494723	1.434564
min	1.000000	-122.419091	47.495573	1.000000	1001.000000	1001.000000	23807.000000	1.000000	0.000000
25%	1.000000	-122.347374	47.583674	56714.500000	73215.500000	73215.500000	28667.000000	1.000000	2.000000
50%	1.000000	-122.330229	47.616491	111426.000000	127954.500000	127954.500000	29973.000000	1.000000	2.000000
75%	2.000000	-122.312515	47.665035	169099.250000	215489.250000	215749.250000	33973.000000	2.000000	3.000000
max	2.000000	-122.241082	47.734141	219546.000000	331453.000000	332953.000000	757580.000000	2.000000	81.000000

Because of the heavy incidence of categorical data and as most of the numerical data is not useful for intersection analysis, this raw data description report is not very edifying. Although it is worth noting that the SEVERITYCODE is 1 or 2, and can be usefully normalized to indicate less vs more severe as 0 or 1, just by subtracting 1.

The correlation matrix showing which numeric fields are strongly correlated with others is shown below. Most of the fields for consideration in this analysis are not so strongly correlated as to represent a problem.



Methodology: Data Cleanup

1. Each of the categorical data attributes needs cleaning, primarily as a result of inconsistent nomenclature (e.g. using both 'Y' and '1'), or nan and 'Y'

2. Generally speaking, the data is reasonably well distributed among the various labels, albeit with some heavy tendencies, e.g. towards clear / dry / daylight conditions, reflecting traffic conditions

3. Other than a few key fields like VEHCOUNT, most of the numeric fields are not relevant to this analysis.

4. In fact, clearly, many of the data attributes are not useful for predictive purposes. Much of it is administrative record keeping.

5. As noted in the earlier data description, looking into the dataset description found here:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

and using this metadata, it's clear that most of the columns are not useful for our purposes, which is looking at intersection safety. So one important step to make the data more useful early on is to cut down to just the columns that are likely to influence intersection design.

6. Of the full list of columns, using both the metadata and the visualizations, data attributes that can be used on a (potential) predictive basis related to collisions is:

PEDCOUNT PEDCYLCOUNT VEHCOUNT JUNCTIONTYPE INATTENTIONIND UNDERINFL WEATHER ROADCOND LIGHTCOND PEDROWNOTGRNT SPEEDING CROSSWALKKEY HITPARKEDCAR.

Most of these 13 attributes are text based and will need to be converted to be used flexibly within machine learning models.

7. Multiple fields are text label style categoricals and must be converted to ‘one hot’ column style attributes: WEATHERCOND, ROADCOND, LIGHTTYPE, and COLLISIONTYPE.

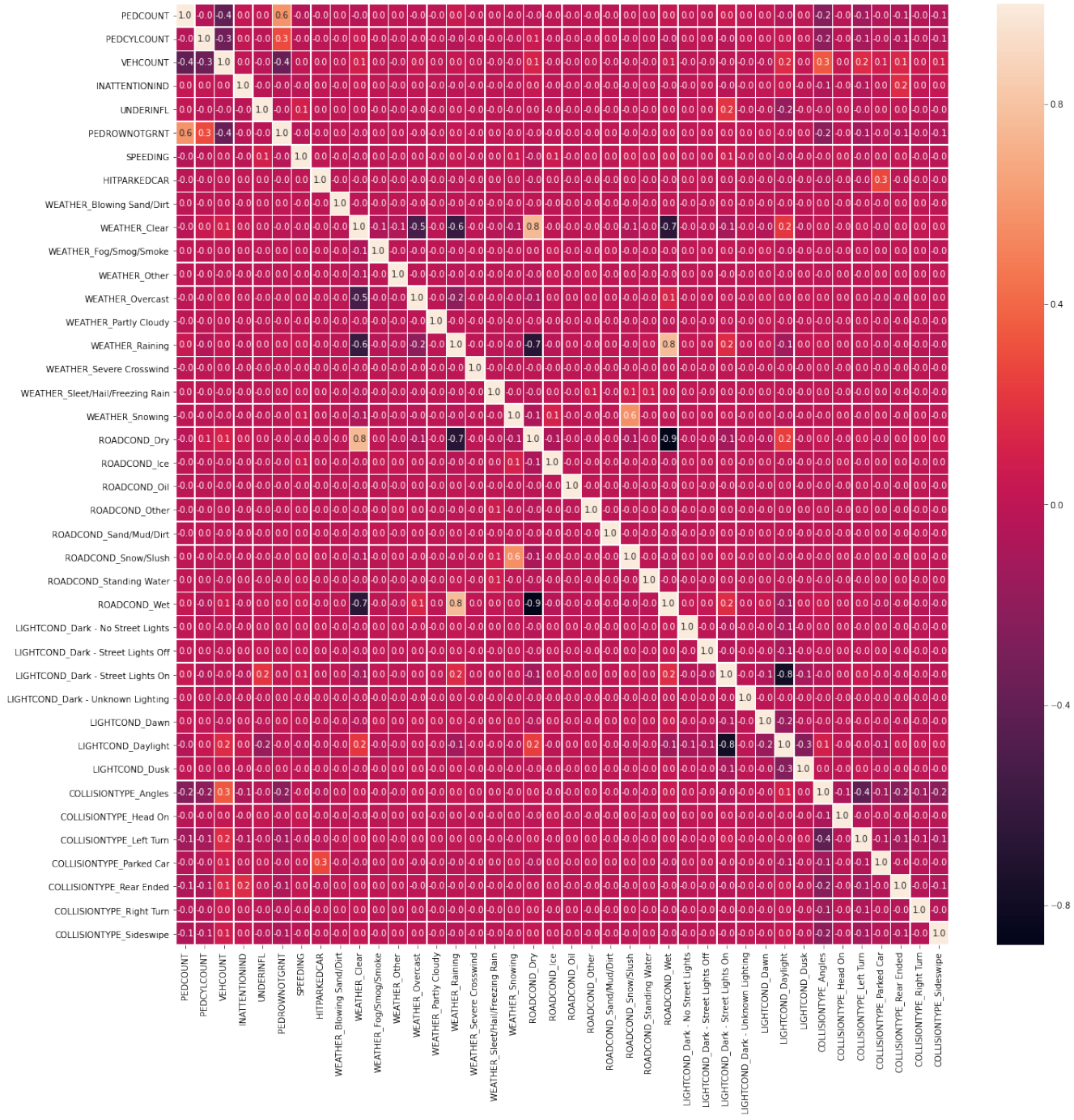
Cleaned Data

After application of the above cleaning steps, the data looks as follows (partial image):

PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INATTENTIONIND	UNDERINFL	PEDROWNOTGRNT	SPEEDING	HITPARKEDCAR	WEATHER_Blowing Sand/Dirt	WEATHER_Clear	..
0	0	2	0.0	0.0	0.0	0.0	0	0	0	..
0	0	2	0.0	0.0	0.0	0.0	0	0	0	..
0	0	2	0.0	0.0	0.0	0.0	0	0	1	..
0	0	2	0.0	0.0	0.0	0.0	0	0	0	..
0	1	1	0.0	0.0	0.0	0.0	0	0	1	..

Notice the one hot columns for Weather on the right hand side, and that all of the columns have been cleaned of labelled data and converted to 1/0 fields.

With this conversion, the data can be looked at entirely within a correlation heatmap, which now looks as follows:



Methodology: Machine Learning Models

To build the models, the data was split into train/test datasets, with 20% reserved for testing.

The models chosen for this initial analysis are Decision Trees, Random Forests, and Logistic Regression, which represent three very different approaches to modeling. This should provide a good selection of outcomes.

For most of the algorithms, the scikit-learn default parameters were used.

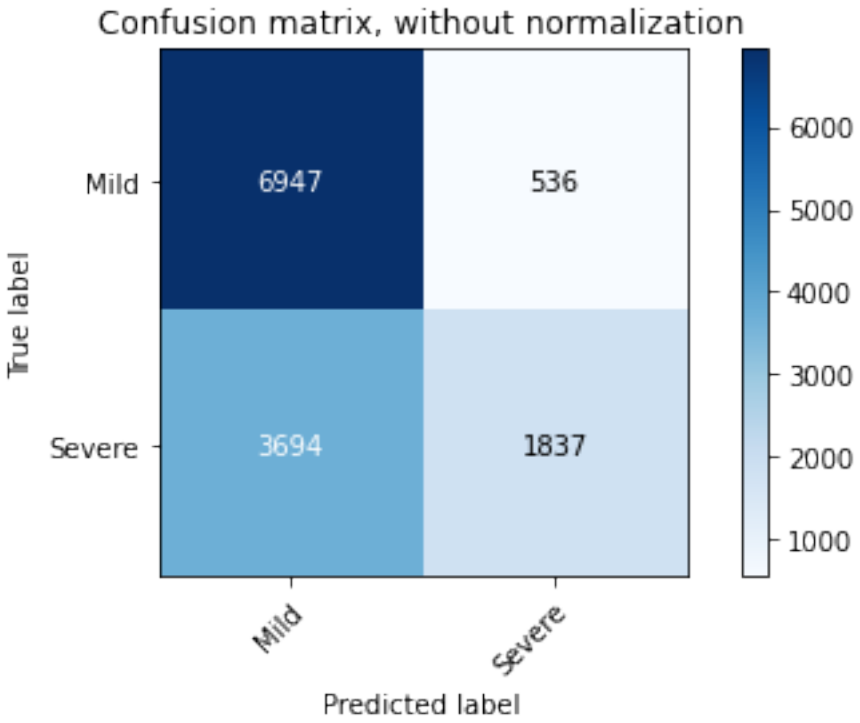
Decision Tree

After training on the training data, the decision tree produced the following results on the test data (the weighted F1 score is highlighted):

```
Jaccard score is 0.3
Accuracy score is 67.5%
              precision    recall  f1-score   support

           0       0.65      0.93      0.77      7483
           1       0.77      0.33      0.46      5531

    accuracy                           0.67     13014
   macro avg       0.71      0.63      0.62     13014
weighted avg       0.70      0.67      0.64     13014
```



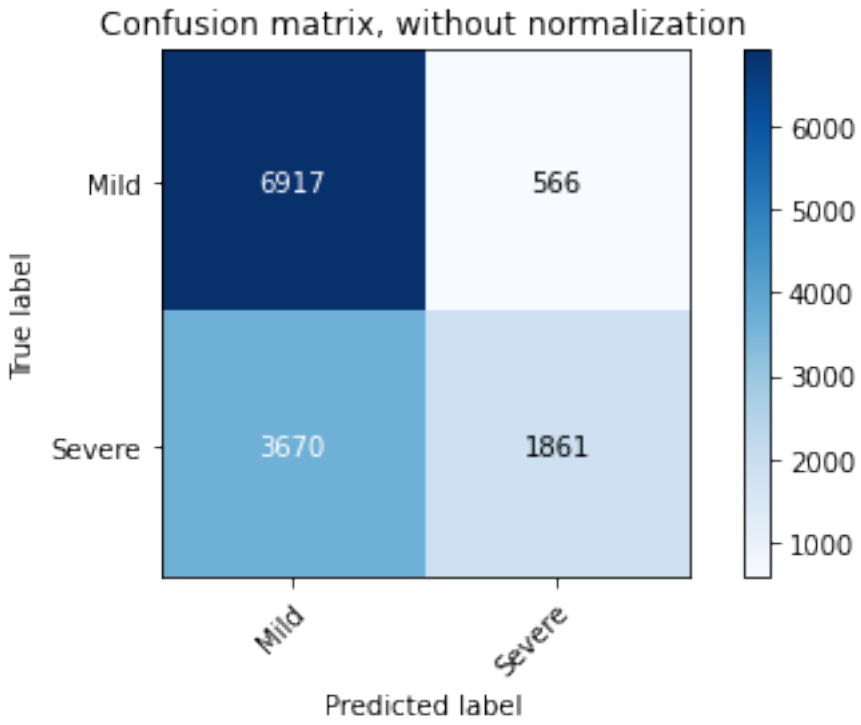
Random Forest

The Random Forest algorithm produced the following results:

```
Jaccard score is 0.31
Accuracy score is 67.5%
              precision    recall  f1-score   support
```

```
              0       0.65      0.92      0.77      7483
              1       0.77      0.34      0.47      5531

    accuracy                            0.67     13014
   macro avg       0.71      0.63      0.62     13014
weighted avg       0.70      0.67      0.64     13014
```



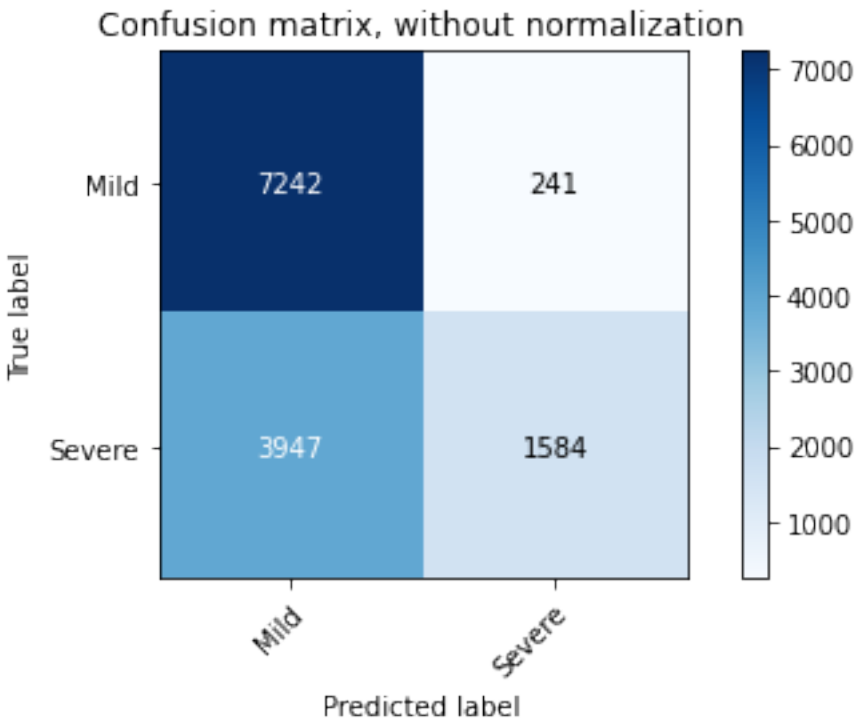
Logistic Regression

The Logistic Regression algorithm produced the following results:

```
Jaccard score is 0.27
Accuracy score is 67.8%
              precision    recall  f1-score   support

           0       0.65      0.97      0.78      7483
           1       0.87      0.29      0.43      5531

    accuracy                           0.68     13014
   macro avg       0.76      0.63      0.60     13014
weighted avg       0.74      0.68      0.63     13014
```



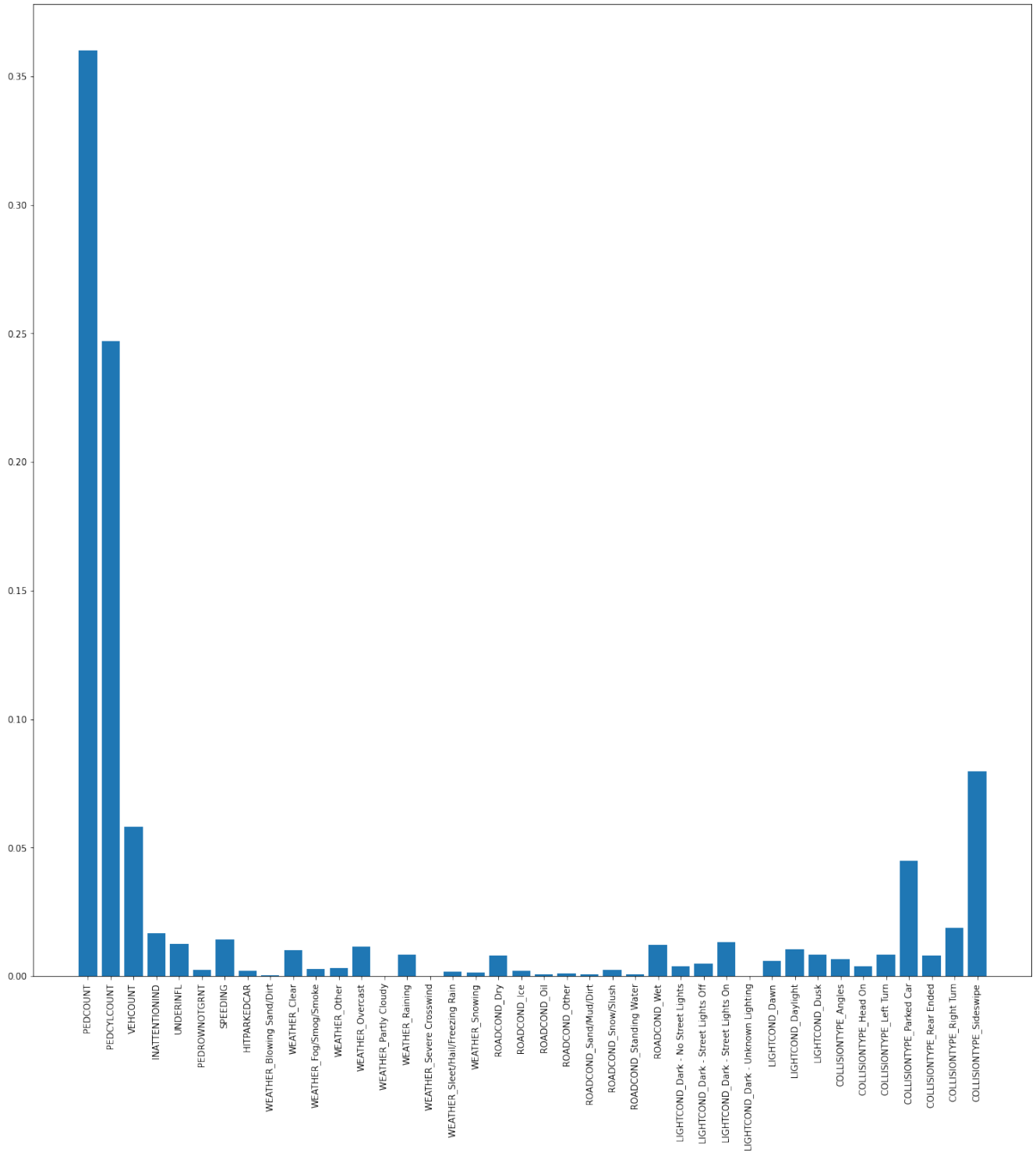
Results

All three algorithms performed very similarly, with extremely minor variations in performance. Logistic regression for example had the best prediction performance for mild severity and the highest accuracy, but the lowest Jaccard and F1 scores - all by a small margin.

However, it is important to bear in mind that predictive performance isn't the goal with this analysis. Rather, the goal is to assess the individual factors that drive severity of collisions.

To that end, given the similar performance, the Decision Tree model was used as it provides high explainability and automatically sorts for feature importance.

Using this capability, the following bar chart shows the relative importance of the features for prediction as determined by the Decision Tree algorithm:



We can conclude that there are five features that stand out for helping predict collision severity at intersections:

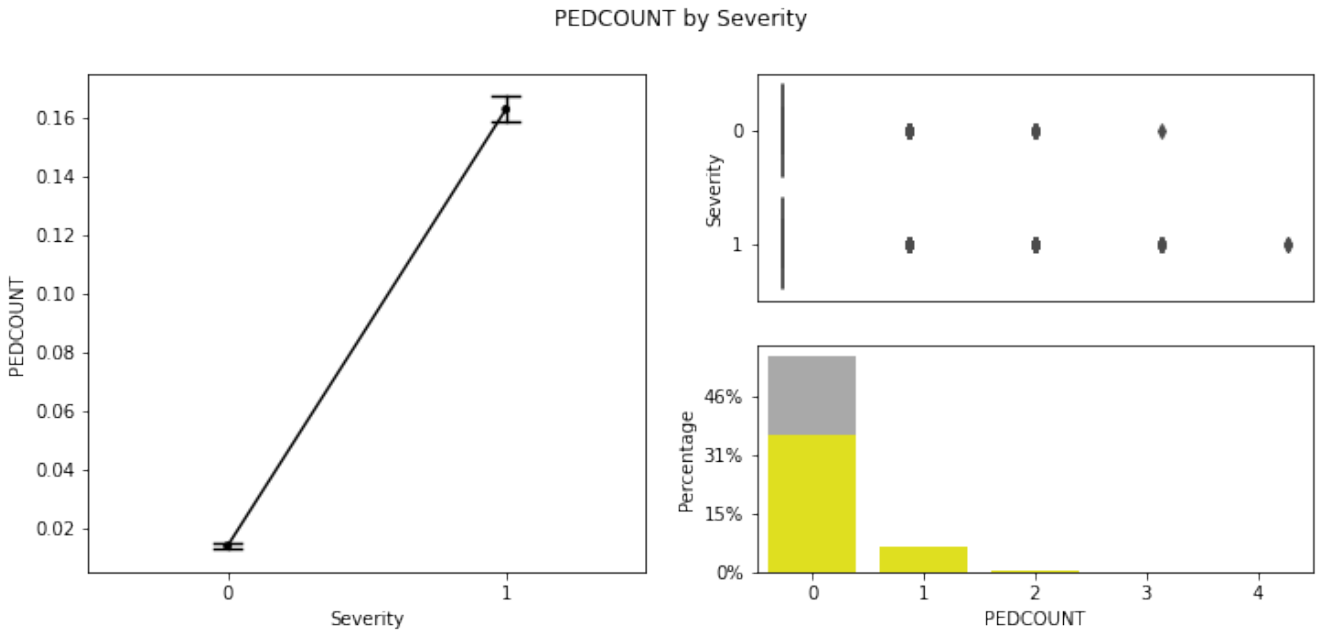
- Pedestrian(s) present
- Bicyclist(s) present
- Number of vehicles involved

- Presence of parked cars
- Sideswipe style collisions

The following charts show the influence of each of these factors on collisions severity:

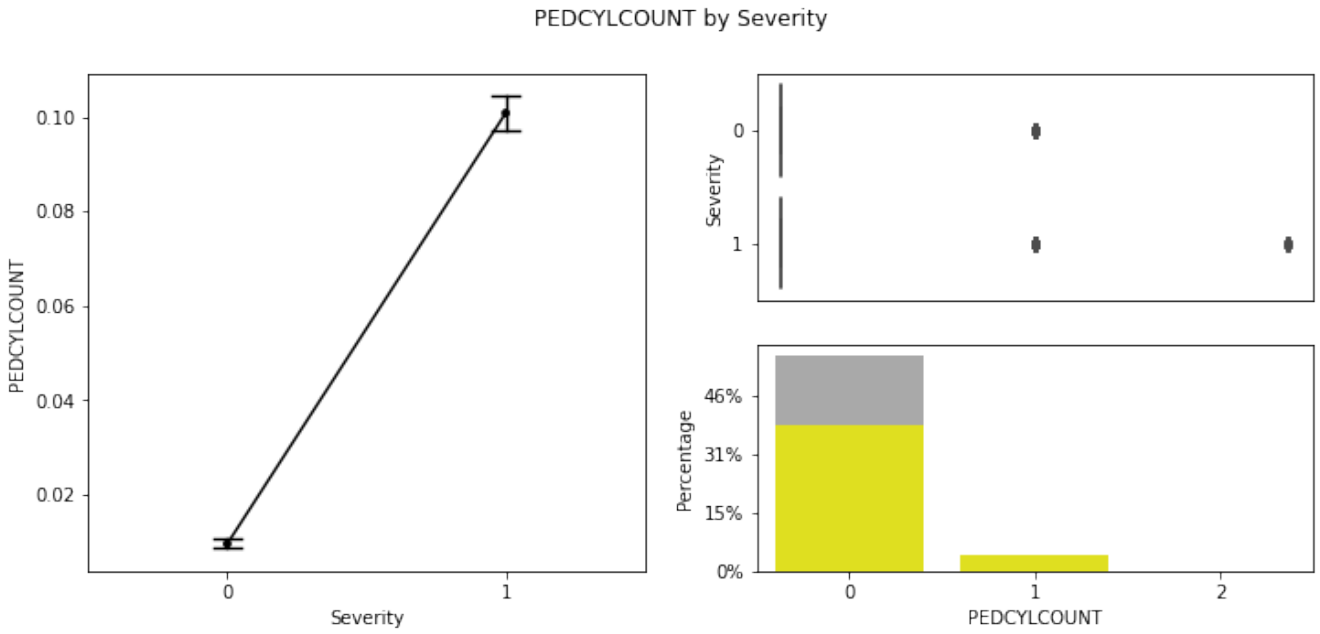
Pedestrian Count

Severity increases with increasing pedestrian count, though data is scarce (thankfully) beyond one.



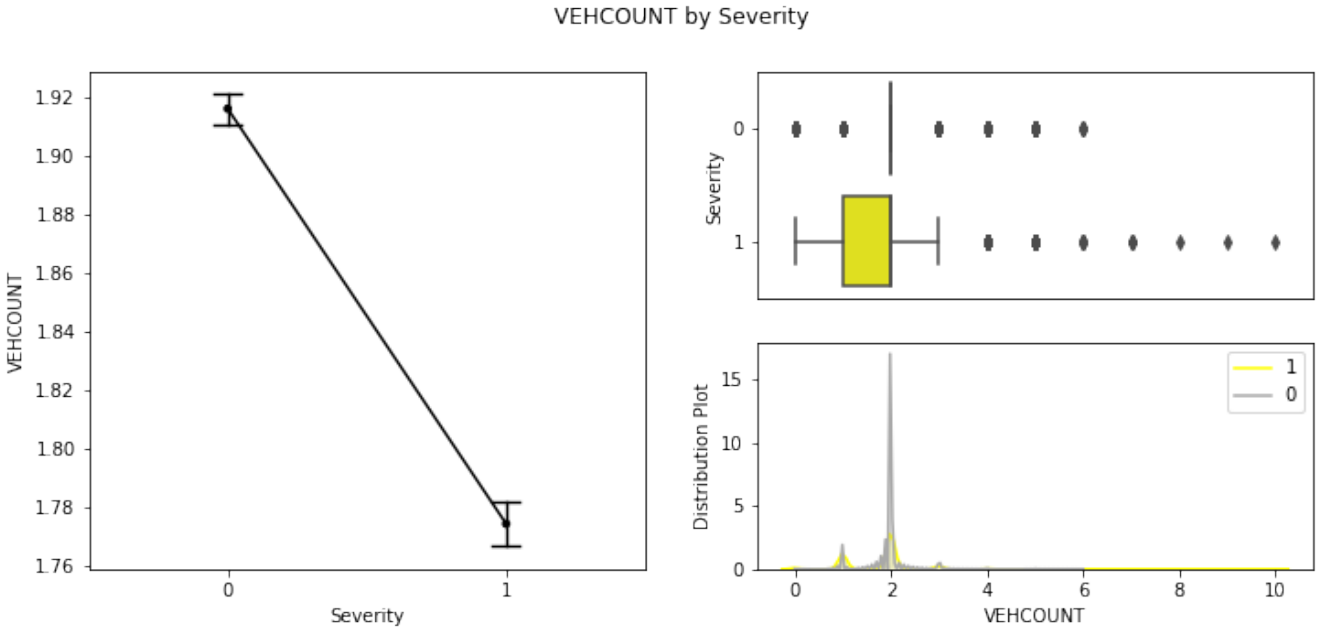
Bicycle Count

Severity increases with the presence of bicycles. Unsurprisingly, there is no data beyond two bicycles.



Vehicle Count

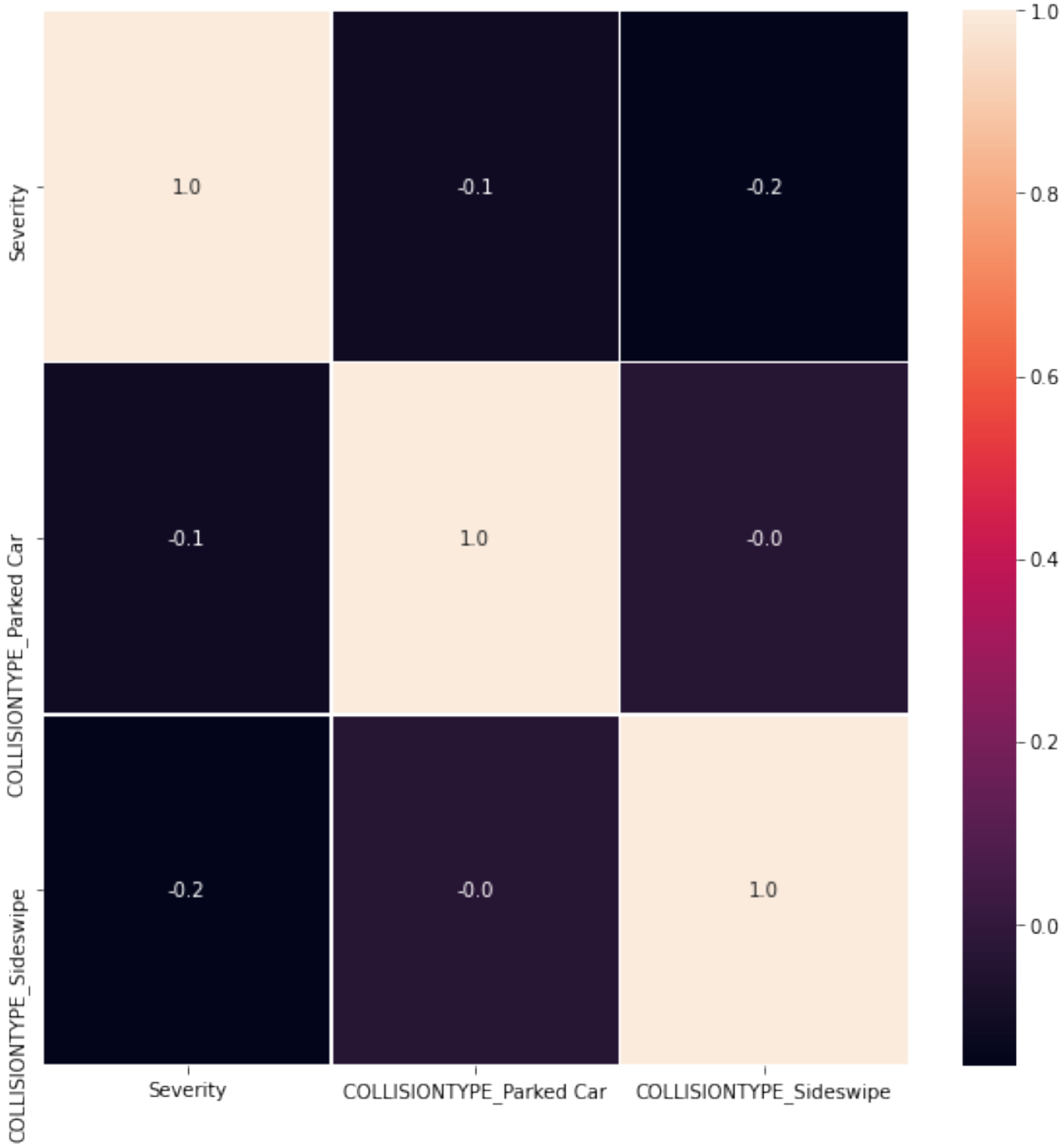
Collision severity *decreases* with increased vehicle count. At first glance this might be surprising, but in context of the previous two factors, it is likely that having another vehicle involved decreases likelihood of pedestrian or bicycle involvement.



Parked Car and Sideswipe Correlation

Parked car and sideswipe collisions are both negatively correlated with accident severity.

As with vehicle count, likely these collisions involve no pedestrians and no bicycles, so severity decreases.



Discussion

The end result of this analysis produces some straightforward results:

1. Severity of collision is most significantly impacted by pedestrian and bicycle involvement. These factors are by far the largest and most important, likely overwhelming all other factors.

2. Vehicle involvement, parked cars, and sideswipes all decrease collision severity. However, this effect may not be independent of the pedestrian effect.

As such, the number one priority for reducing collision severity is to increase protections for pedestrians and bicycles. This might involve signage, visibility, speed restrictions, etc.

To specifically reduce *vehicle-vehicle* collisions, more analysis is required, removing collisions that involve pedestrians and bicycles and focusing strictly on situations involving vehicles.

Additional future work could also include more detailed modeling (for example, using grid search, cross validation, and more algorithms) to increase predictive accuracy.

One enormous blind spot for this analysis is that it has data that only looks at collisions. More data on traffic at intersections with **no** collisions would be very helpful, as it would potentially be just as helpful to look at what makes a safe intersection as what makes an unsafe intersection.

Conclusion

This analysis used several decades of collision data in Seattle to look for factors that influence collision severity, using machine learning models to isolate these factors.

There are several opportunities to improve the analysis in future, but from this analysis, five factors were identified as particularly significant. By far the most important factors were pedestrians and to a lesser extent bicycles.

Traffic planners and engineers can incorporate this information as they feel appropriate and in keeping with their own expertise in assessing current and future intersections.