SEATTLE TRAFFIC COLLISIONS 2004 - 2020

Coursera IBM Professional Data Science Certificate
Capstone
Guilherme Werner

The Data

- The data was provided by Coursera and is a sample of the data collected by the Seattle Police Department over the period of (2004 – 2020).
- Python *pandas* was used for the processing of the data as DataFrames.
- The original dataset was processed and cleaned using *pandas,* and trends were identified with the help of the *sklearn machine learning library algorithms,* in particular, the DecisionTreeClassifier, RandomForestClassifier, train_test_split, and LabelEncoder Functionalities.
- The image above is a sample DataFrame (.head(10)), containing 10 data entries, the overall frame contained 194,673 points.

Index	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE	EXCEPTRSNDESC	SEVERITYCODE.1	SEVERITYDESC	COLLISIONTYPE
0	2	-122…	47.7…	1	1307	1307	3502005	Matched	Intersect…	37475	5TH AVE …		nan	2	Injury Collision	Angles
1	1	-122…	47.6…	2	52200	52200	2607959	Matched	Block	nan	AURORA B…	nan	nan	1	Property Dam…	Sideswipe
2	1	-122…	47.6…	3	26700	26700	1482393	Matched	Block	nan	4TH AVE …	nan	nan	1	Property Dam…	Parked Car
3	1	-122…	47.6…	4	1144	1144	3503937	Matched	Block	nan	2ND AVE …		nan	1	Property Dam…	Other
4	2	-122…	47.5…	5	17700	17700	1807429	Matched	Intersect…	34387	SWIFT AV…	nan	nan	2	Injury Collision	Angles
5	1	-122…	47.6…	6	320840	322340	E919477	Matched	Intersect…	36974	24TH AVE…		nan	1	Property Dam…	Angles
6	1	-122…	47.6…	7	83300	83300	3282542	Matched	Intersect…	29510	DENNY WA…	nan	nan	1	Property Dam…	Angles
7	2	-122…	47.6…	9	330897	332397	EA30304	Matched	Intersect…	29745	BROADWAY…		nan	2	Injury Collision	Cycles
8	1	-122…	47.6…	10	63400	63400	2071243	Matched	Block	nan	PINE ST …	nan	nan	1	Property Dam…	Parked Car
9	2	-122…	47.5…	12	58600	58600	2072105	Matched	Intersect…	34679	41ST AVE…	nan	nan	2	Injury Collision	Angles
10	1	nan	nan	14	48900	48900	2024040	Matched	Alley	nan	nan	nan	nan	1	Property Dam…	Other

Data Cleaning and Trend Analysis

- Following the acquisition of the data, the data was cleared of empty entries, data such as the police incident indicators, sidewalk identifier, incident numerical code, and other variables which did not provide insight on traffic correlations.

- A machine learning algorithm was run, creating a decision tree of the data, which allowed for the identification of the variables with the greatest impacts on the collision correlations. An example of such a decision tree can be found on the next page.
- Additional trends were obtained through the use of the *RandomForrestClassifier,* a machine learning algorithm which generates several decision trees, as opposed to a single tree, and ranks variables (features) in accordance with their influence on the parameter we seek to investigate.

Decision Tree

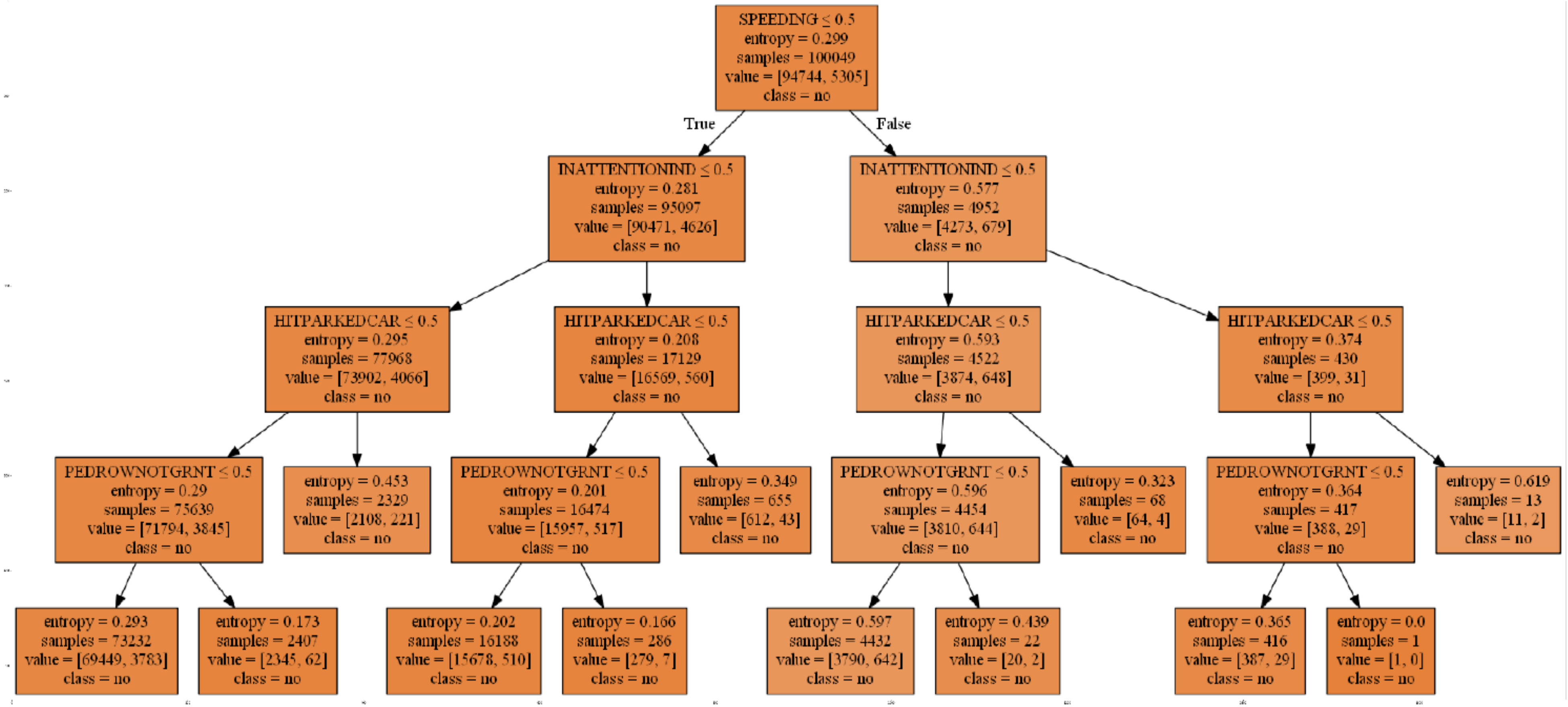

Figure 1 – *DecisionTreeClassifier* algorithm, correlating , Inattention, Pedestrian Right of Passage, Speeding, Hit Parked Car, accuracy score of 0.94762005, max_depth was not defined.

RandomForestClassifier Algorithm

Weight	Feature
0.0641 ± 0.0150	ST_COLCODE
0.0251 ± 0.0132	INATTENTIONIND
0.0200 ± 0.0101	PERSONCOUNT
0.0190 ± 0.0126	JUNCTIONTYPE_CODE
0.0106 ± 0.0100	LIGHT_CODE
0.0023 ± 0.0040	SDOT_COLCODE
0.0008 ± 0.0010	VEHCOUNT
0.0005 ± 0.0014	SPEEDING
0.0002 ± 0.0012	PEDCOUNT
0 ± 0.0000	PEDCYLCOUNT
0 ± 0.0000	COLLISIONTYPE_CODE
0 ± 0.0000	HITPARKEDCAR
-0.0029 ± 0.0069	SEVERITYCODE.1
-0.0029 ± 0.0023	SEVERITYCODE
-0.0041 ± 0.0144	LOCATION_CODE
-0.0060 ± 0.0128	ADDRTYPE_CODE
-0.0175 ± 0.0096	ROADCOND_CODE
-0.0190 ± 0.0091	WEATHER_CODE

Figure 2 - Correlation between pedestrian right of passage, for collisions involving vehicles and pedestrians. Features in green indicate decision variables which have share the strongest (weight) correlation between vehicle-pedestrian collisions.

Weight	Feature
0.0030 ± 0.0020	INATTENTIONIND
0.0019 ± 0.0032	PERSONCOUNT
0.0005 ± 0.0022	LIGHT_CODE
0.0005 ± 0.0023	ROADCOND_CODE
0 ± 0.0000	COLLISIONTYPE_CODE
0 ± 0.0000	HITPARKEDCAR
0 ± 0.0000	PEDCYLCOUNT
0 ± 0.0000	VEHCOUNT
-0.0005 ± 0.0009	SPEEDING
-0.0006 ± 0.0011	SDOT_COLCODE
-0.0006 ± 0.0080	PEDROWNOTGRNT
-0.0010 ± 0.0022	SEVERITYCODE.1
-0.0010 ± 0.0018	SEVERITYCODE
-0.0013 ± 0.0009	PEDCOUNT
-0.0023 ± 0.0042	WEATHER_CODE
-0.0056 ± 0.0080	LOCATION_CODE
-0.0084 ± 0.0046	ADDRTYPE_CODE
-0.0090 ± 0.0033	ST_COLCODE
-0.0114 ± 0.0011	JUNCTIONTYPE_CODE

Figure 3 – Random forest classifier for vehicle collisions involving alcohol and drug consumption.

Geospatial Representations Using Folium

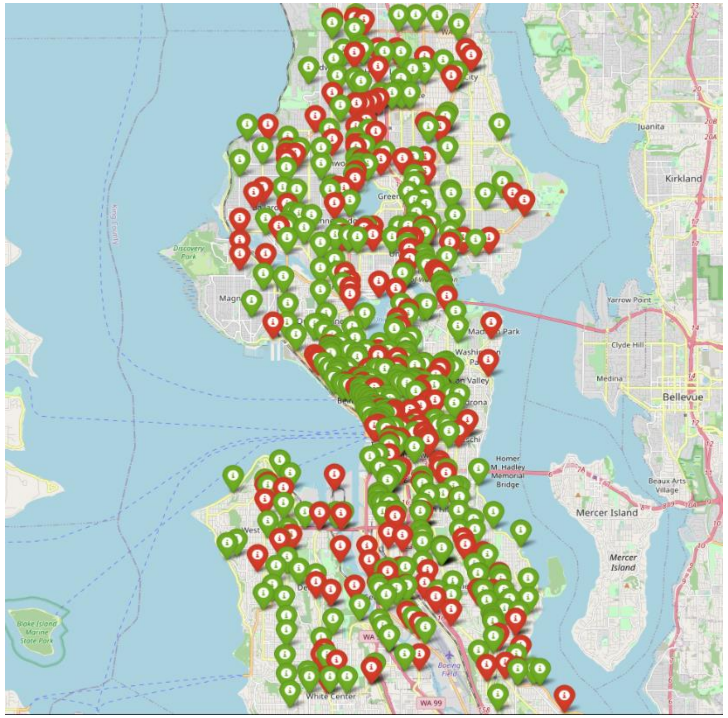

Figure 4 - Geospatial folium representing all collision types, where red indicates collisions with injuries and green indicates collisions with property damage only. A random sample of 600 of the databases 194,673 data points was plotted.

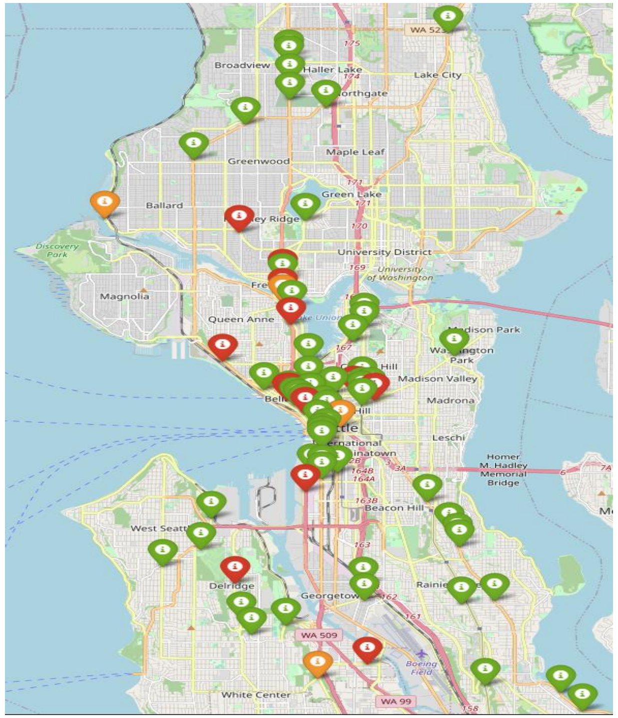

Figure 5 - Geospatial folium representing all locations where there were more than 5 collisions involving Drugs and Alcohol over the period from 2004-2020, green indicates between 6-7 collisions, orange indicates between 8-10 collisions and red indicates

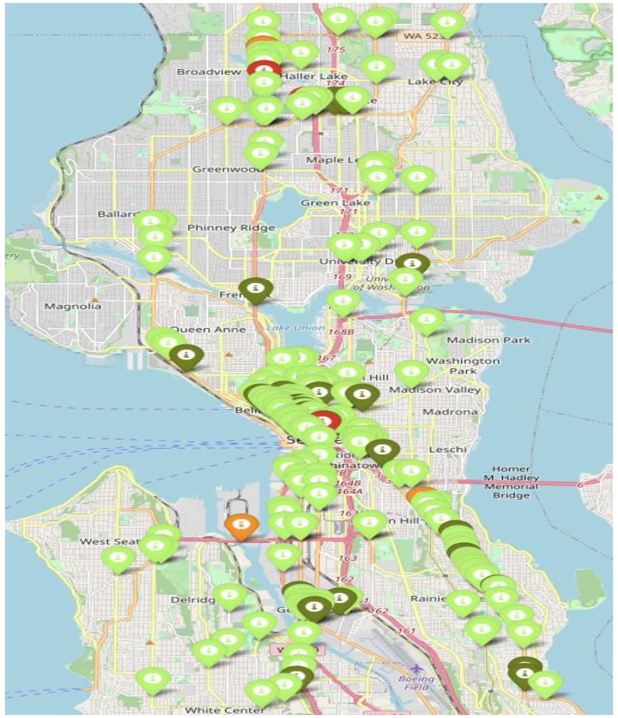

Figure 6 – Geospatial folium representing all locations where more than 60 Pedestrians were involved in vehicle Collisions between 2004-2020, where light-green represents between 60-99 pedestrians, dark-green represents 100-149, orange 150-200, and red more than 200, up to a maximum observed cases of 217.

Analysis of Collisions with Speeding

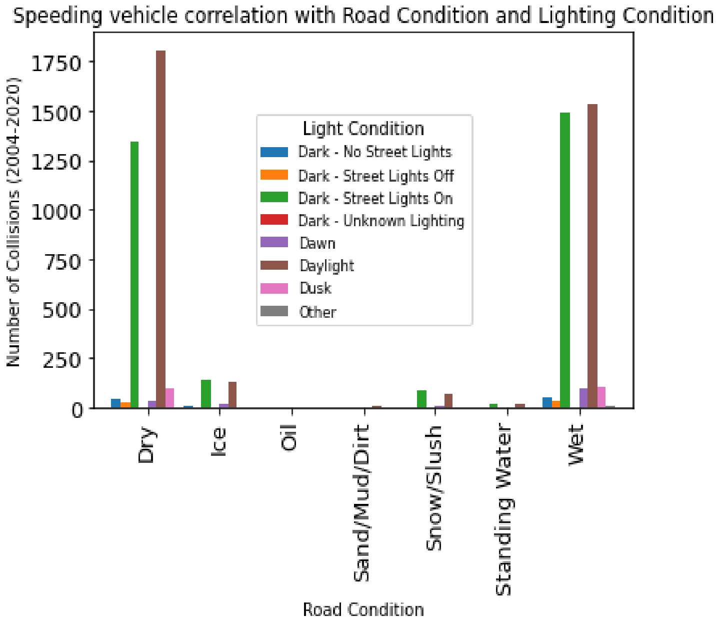

Figure 7 - Correlation between Speeding, Lighting and Road conditions (2004-2020).

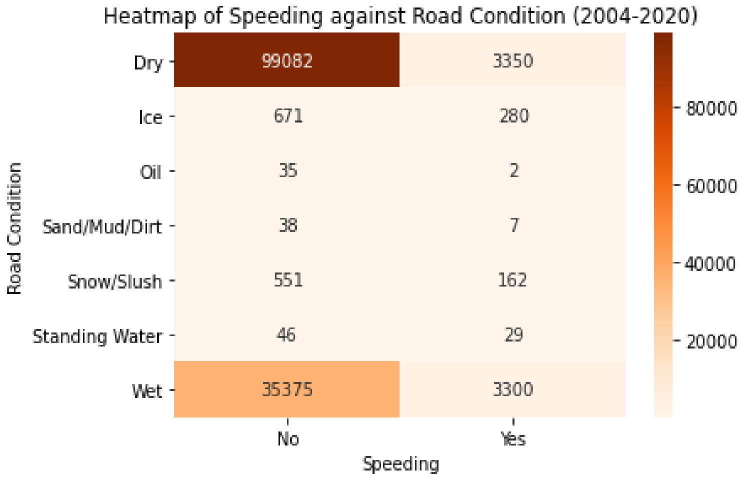

Figure 8 - Co-Occurrence Matrix, correlating speeding vehicles and road condition (2004-2020).

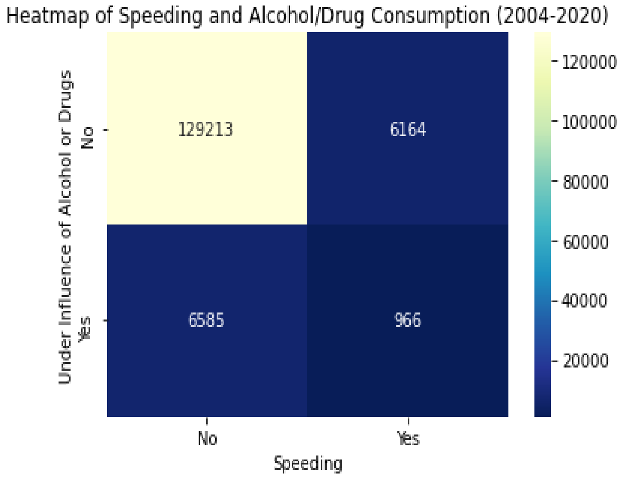

Figure 9 – Co-occurrence matrix of two binary data sources, correlating Speeding collisions with Alcohol and Drug consumption collisions.

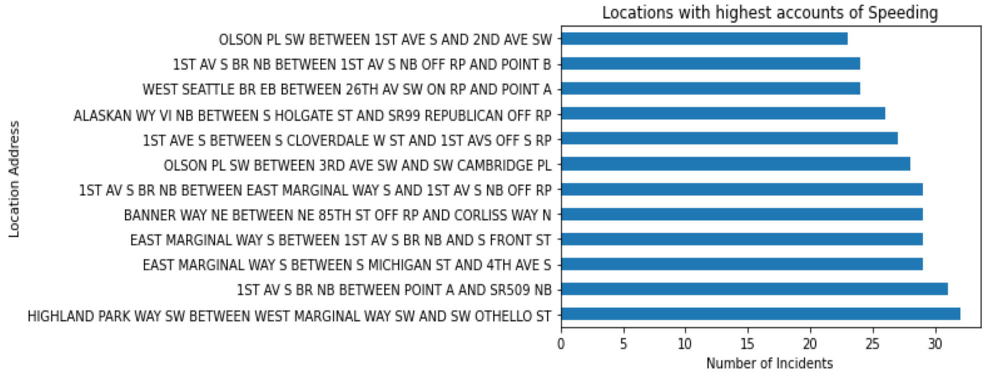

Figure 10- Locations with the highest rate of speeding infractions between (2004-2020).

Time of Day and Collision Analysis

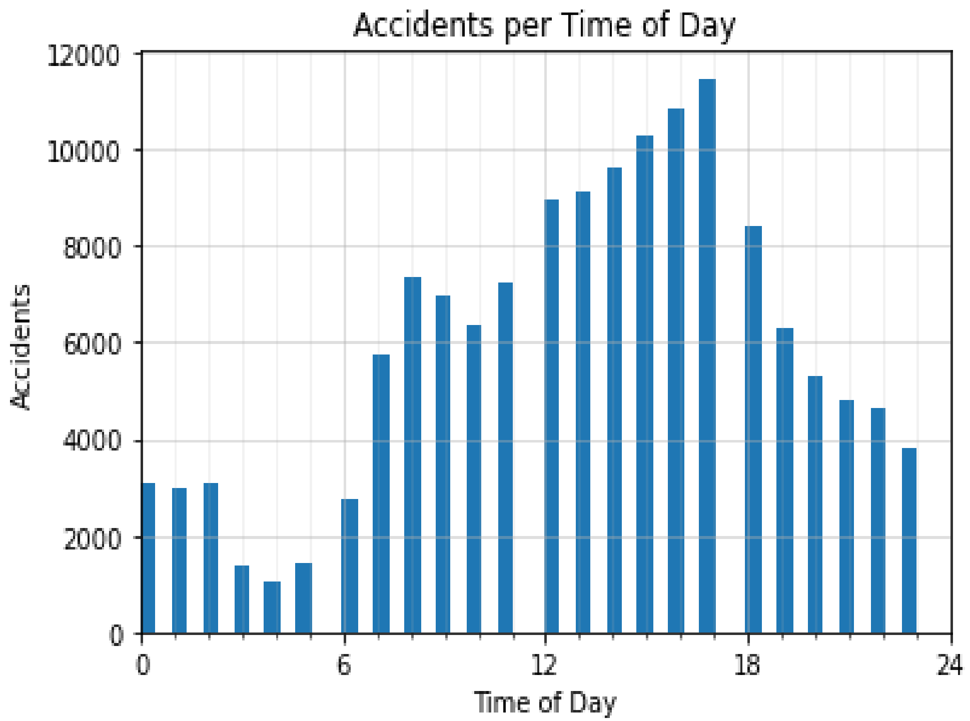

Figure 11 – Histogram of the distribution of accidents per time of day (2004-2020)

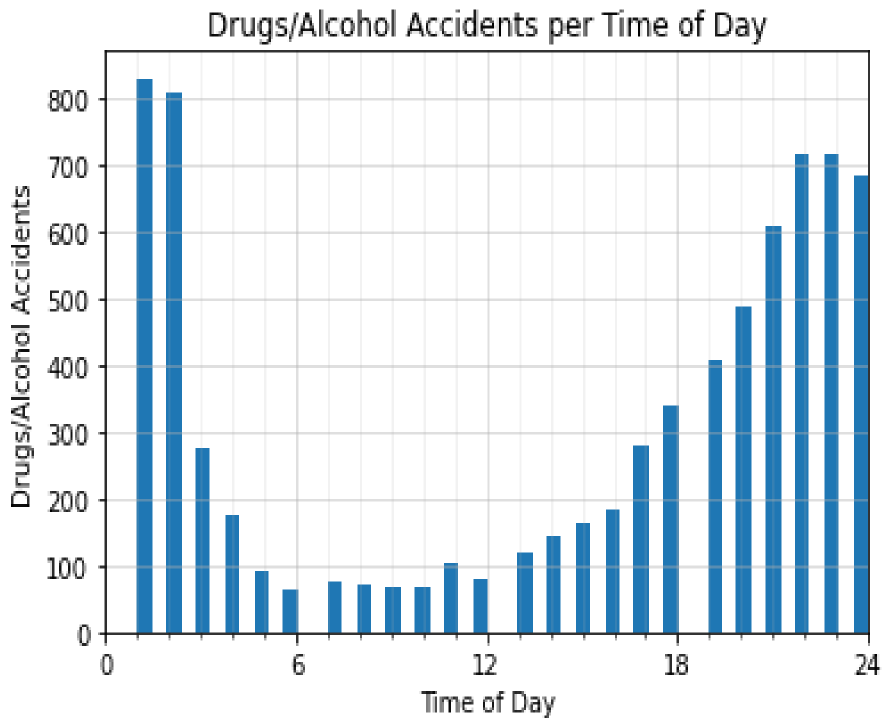

Figure 12 – Histogram of the distribution of Drug Related Accidents per time of Day

Alcohol and Drug Related Collisions

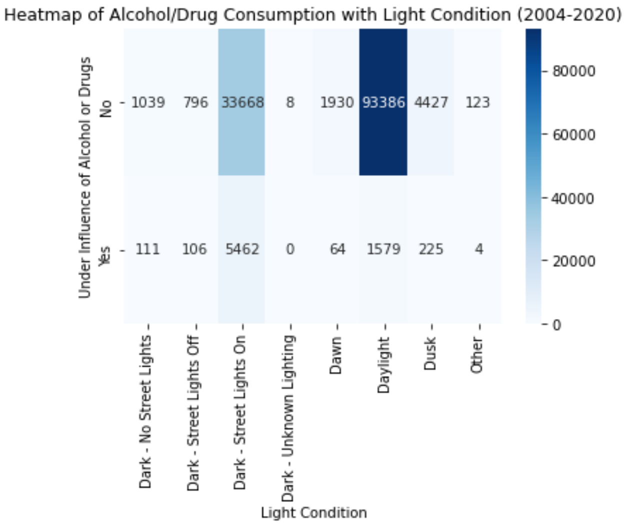

Figure 13 – Correlation between alcohol and drug consumption and light conditions of the road at the time of the collision.

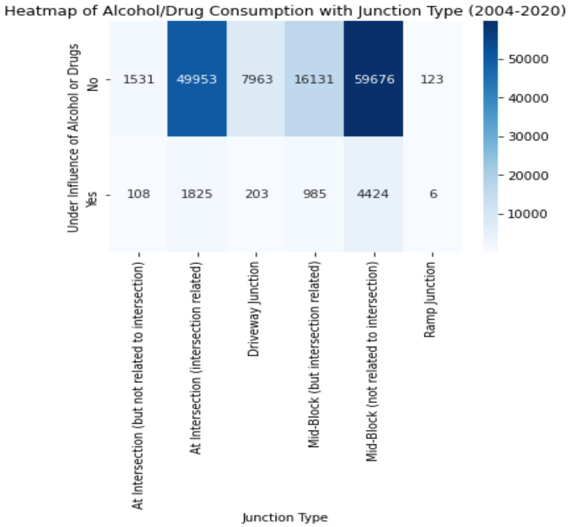

Figure 14 – Correlation between alcohol and drug consumption and the junction type of the collision.

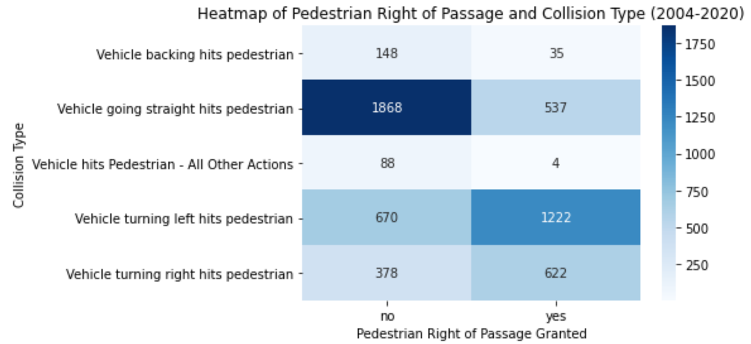

Figure 15 - Co-occurrence matrix correlating collisions according to collision type and pedestrian right of passage. This was the highest correlated parameter given by the Random Forest Classifier.

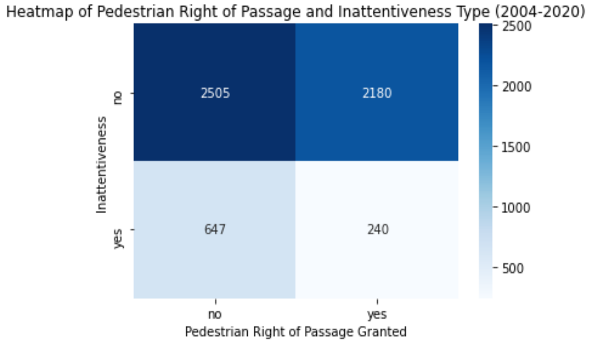

Figure 16 - Pedestrian Right of Passage and Inattentiveness. This was the second highest correlated parameter given by the Random Forest Classifier.

Further Information

- For further information regarding data preparation, data analysis, machine learning algorithms, and on the report, please refer to the final report on the GitHub repository.