COURSERA IBM DATA SCIENCE CAPSTONE

Guilherme Werner

Table of Contents

1. Introduction

1.1. Background

With an ever-increasing arsenal of technologies aimed towards enhancing vehicle safety, the most notable of which being: self-driving autonomous vehicles, increased crash-test regulations and vehicle chassis stability, alongside state-of-the-art vehicle collision detection systems; the automotive industry of today is progressively working towards increasing vehicle occupant safety and decreasing the likelihood and severity of traffic accidents.

However, society is still far from achieving universal road safety, as shown by the annual American *National Highway Traffic Safety Administration (NHTSA)* report, which recorded 36,560 traffic fatalities over the year of 2018, although representing a decrease of 913 fatalities, a large number of individuals still perish annually due to road accidents.

Such rates are equivalent to 1.13 fatalities per 100 million vehicles miles (160 million km), which are low rates when we consider the vast number of motor vehicles (1), however numerous lives are still lost annually due to road accidents, and data science has a central role to play when identifying the hazards and mitigating the risks of vehicle collisions.

The present data-science investigates analyses an extensive collision dataset, collected by the *Seattle Police Department (SPD)*, over the period of 2004-2020, with the goal of identifying the variables which contribute towards traffic accidents, alongside the changing incident trends over recent years, identifying potential areas for improvement and proposing solutions to enhance road-safety across the streets of Seattle.

1.2. Business Proposal

The stakeholders identified for the present project are the *Seattle Police Department (SPD)*, the *Seattle Transportation Office*, vehicle insurance providers, politicians, active community residents, and city residents. The aim of the study shall be to propose solutions which can be implemented and acted upon to ensure a safer road network across the city of Seattle.

The investigation provides an overview of the vehicle collision database for Seattle and identifies a few potential areas for improvement. This report was written as part of the *Coursera IBM Professional Data Science Certificate* and uses the dataset provided by the *Coursera* team.

1.3. Problem

The city of Seattle is home to a massive fleet of vehicles, according to a *2017 Seattle Times Report*, there are around 637 cars for every 1000 residents, with the total vehicle count estimated at around 435,000 (2). The cities over-reliance on vehicles for daily commute and transportation brings with it growing challenges for traffic authorities, residents, and drivers alike. Over the period of 2004-2020 a total of 194,673 collisions were recorded within the city, with an annual mean of 17,638 collisions.

2. Data

2.1. Data Source

The dataset was obtained from a *Seattle Government* data repository and consists of 194,673 vehicle collision instances over the course of a 16-year period from 2004-2020. The data was provided as a *.csv* file by Coursera. Each data entry consists of 33 unique variables, which detail the accident, containing information on the location, date and time, collision type, junction type, lighting, and road conditions, amongst several others. Below is an outline of the original data frame, imported to python as a .csv file.

Index	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION
0	2	-122.323	47.7031	1	1307	1307	3502005	Matched	Intersection	37475	5TH AVE NE AND NE 103RD ST
1	1	-122.347	47.6472	2	52200	52200	2607959	Matched	Block	nan	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N
2	1	-122.335	47.6079	3	26700	26700	1482393	Matched	Block	nan	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST
3	1	-122.335	47.6048	4	1144	1144	3503937	Matched	Block	nan	2ND AVE BETWEEN MARION ST AND MADISON ST
4	2	-122.306	47.5457	5	17700	17700	1807429	Matched	Intersection	34387	SWIFT AVE S AND SWIFT AV OFF RP
5	1	-122.388	47.6906	6	320840	322340	E919477	Matched	Intersection	36974	24TH AVE NW AND NW 85TH ST
6	1	-122.338	47.6185	7	83300	83300	3282542	Matched	Intersection	29510	DENNY WAY AND WESTLAKE AVE
7	2	-122.321	47.6141	9	330897	332397	EA30304	Matched	Intersection	29745	BROADWAY AND E PIKE ST
8	1	-122.336	47.6119	10	63400	63400	2071243	Matched	Block	nan	PINE ST BETWEEN 5TH AVE AND 6TH AVE
9	2	-122.385	47.5285	12	58600	58600	2072105	Matched	Intersection	34679	41ST AVE SW AND SW THISTLE ST
10	1	nan	nan	14	48900	48900	2024040	Matched	Alley	nan	nan

Figure 1 - Sample of original *Pandas* Data Frame (*.csv*)

The data contains both numerical (quantitative) entries, but also several categorical (non-numerical, qualitative), data entries are were also included in the dataset. Overall, the data possesses numerous variables which allow for interesting correlation analysis.

2.2. Data Preparation

The data was downloaded as a .csv file from the *Seattle Police Department (SPD)* data repository (3). The data was imported to *Python* and a *Metadata* information document was used to provide insight on the structure of the Data Frame and the sort of Data Provided. The data set and respective information can be found on the GitHub repository for this project.

The spreadsheet contains data across the entirety of the municipality of Seattle and refers to collision involving vehicles and other modes of transportation, such as bicycles, trucks, construction machinery, trains, alongside collisions between vehicles and pedestrians.

An initial overview and cleaning of the data was required, contained on the table below is the decision making behind the initial cleaning process.

Data Attribute / Feature Discarded	Reason for Discarding	Type of Data
EXCEPTRSNDESC, EXCEPTRSNCODE, OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, EXCEPTRSNDESC, EXCEPTRSNCODE, INCDATE, SDOTCOLNUM, SEGLANEKEY, CROSSWALKKEY	Discarded as the Data does not allow for analysis or is redundant in nature.	Police and Seattle Government identifiers for the collision registration or repeated data.

Moreover, the format of the date and time of the collision, posed a challenge as not all data was recorded following the same methodology, this was made evident as some of the incidents only had a year, or only had a date, or only had a time. These instances had to be removed from the dataset when analysing the date/time relationships of the dataset.

In addition, several data categories possessed inconsistent data entries, for example the classifier for if an individual was under the influence of drugs or alcohol, *UNDERINFL,* was recorded as both 1/0 and Y/N (yes/no) respectively, such values had to be replaced to ensure a consistent dataset.

In addition, the data was also split into the respective vehicle accident categories, *ST_COLCODE,* the data set attribute which described the vehicle collision type, this allowed for data to be split into categorically relevant sub-sets, for example, pedestrian and vehicle collisions, train and vehicle collisions, and collisions where the driver was over the speed limit.

For the use of *Machine Learning* algorithms and libraries within python, all instances of categorical data had to be indexed as numeric, this meant associating a numerical value for each

respective categorical value. This was done using the *sklearn* library machine learning library, with the *LabelEncoder* attribute, which automated the process. Below is a sample of the final processed *pandas* Data Frame, with all data attributes recorded as *dummy* numerical values.

Index	SEVERITYCODE	SEVERITYCODE.1	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	SDOT_COLCODE
0	2	2	2	0	0	2	11
1	1	1	2	0	0	2	16
2	1	1	4	0	0	3	14
3	1	1	3	0	0	3	11
4	2	2	2	0	0	2	11
5	1	1	2	0	0	2	11
7	2	2	3	0	1	1	51
8	1	1	2	0	0	2	11
9	2	2	2	0	0	2	11

Figure 2 - Using *sklearn LabelEncoder* to prepare data for Machine Learning Algorithms

Once the dataset was processed and cleaned, all of the *np.nan* (empty values), were dropped from the DataFrame, in order to allow for the optimal data analysis, the final post-processed data contained a total of 142,928 data entries, a 26% reduction from the initial pre-processed dataset.

2.3. Data Correlations and Analysis

2.3.1. Machine Learning Algorithms

In order to identify the correlations which exist within the extensive *pandas* Data Frame, the *Python sklearn* machine learning library was used. The algorithms provided by the library allow for the creation of models, using training data from the Data Frame, and the testing and validation of the models through the use of testing data, both the training and testing data were obtained using the *sklearn train_test_split*. Within the algorithm library, the *DecisionTreeClassifier* and later the *RandomForestClassifier* models were used.

The *DecisionTreeClassifier* algorithm was used when analysing the correlation between the binary *alcohol and drug consumption* data and the several variables within the Data Frame, these being, the collision location, number of vehicles involved, weather, light conditions, amongst others. The aim of the Decision Tree algorithm was to build a model to decide upon which conditions are most likely to predict a collision when a driver is under the influence of a drug or alcohol.

The structure of a Decision Tree consists of nodes, which contain a category of the data, in the case of the drugs or alcohol consumption decision tree, the algorithm identified the *light condition, junction type* and *collision type* categories as the ones with the greatest influence on the

drugs and alcohol consumption data set, these node categories are referred to as the *decision variables*.

A decision tree is read according to its entropy, whereby a higher entropy signifies increased disorder (randomness, impurity) of the data, and a lower entropy represents less impurity, and a data set which is more effective at predicting the model. The aim of a decision tree is to model the data, and in doing so, indicate the feature which allows for the greatest information gain (4).

The machine learning algorithm was run subject solely to the constraint of maximum tree depth of 4 branches (children), the algorithm selected the most predictive feature, this being the lighting condition of the street. The primary challenge regarding this methodology was the fact that the data was categorical. To run the algorithm, the data had to be converted to a numerical value using the *sklearn LabelEncoder*. Below are three examples of decision trees on the alcohol and drug collisions data set.

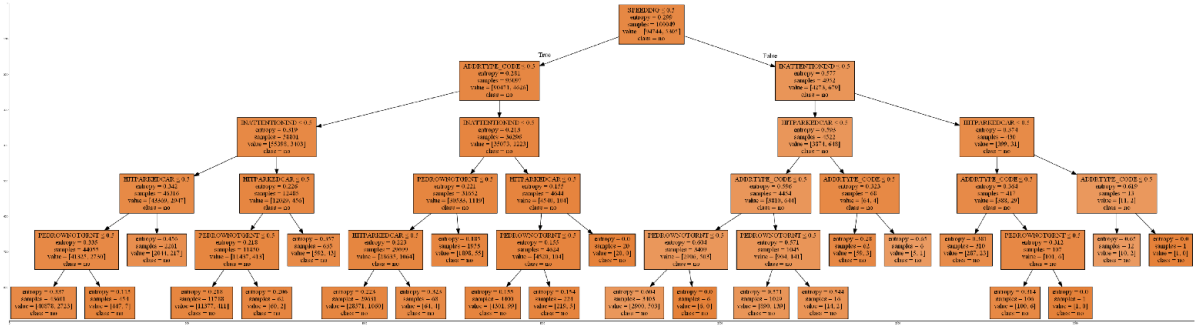

Figure 3 - Decision Tree Classifier, using binary data only, Inattention, Pedestrian Right of Passage, Speeding, Hit Parked Car, Address Code (only Junction and Block collisions considered), accuracy score of 0.94762005, max_depth = 5.

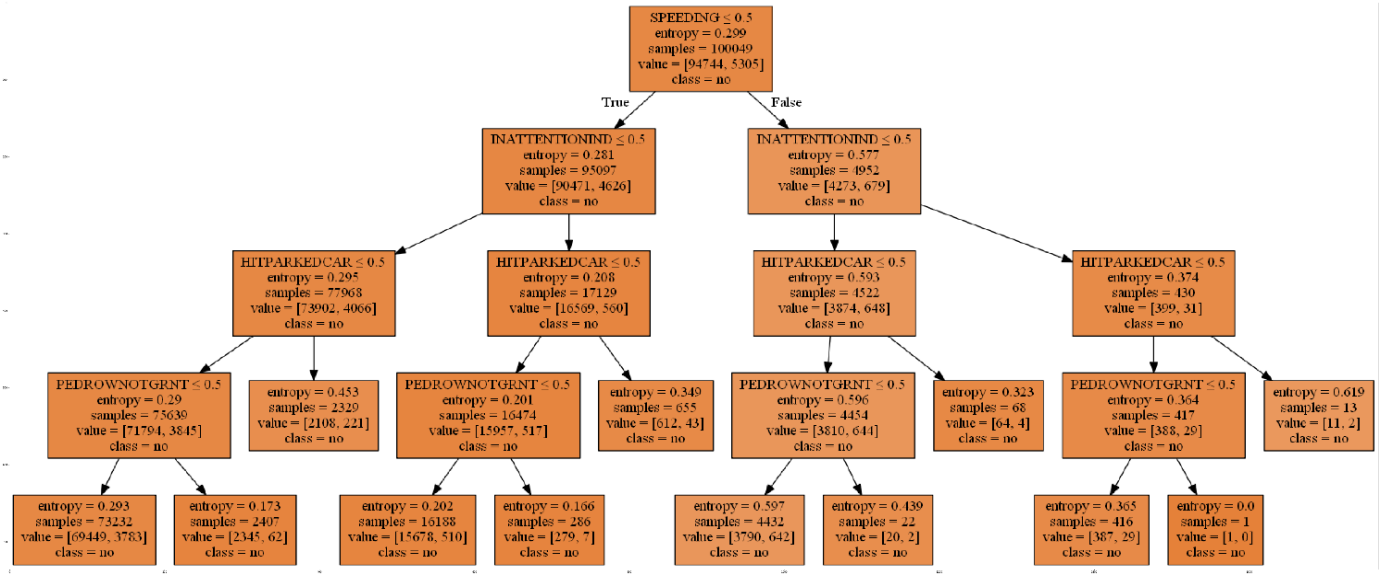

Figure 4 – *DecisionTreeClassifier* algorithm, correlating , Inattention, Pedestrian Right of Passage, Speeding, Hit Parked Car, accuracy score of 0.94762005, max_depth = 4.

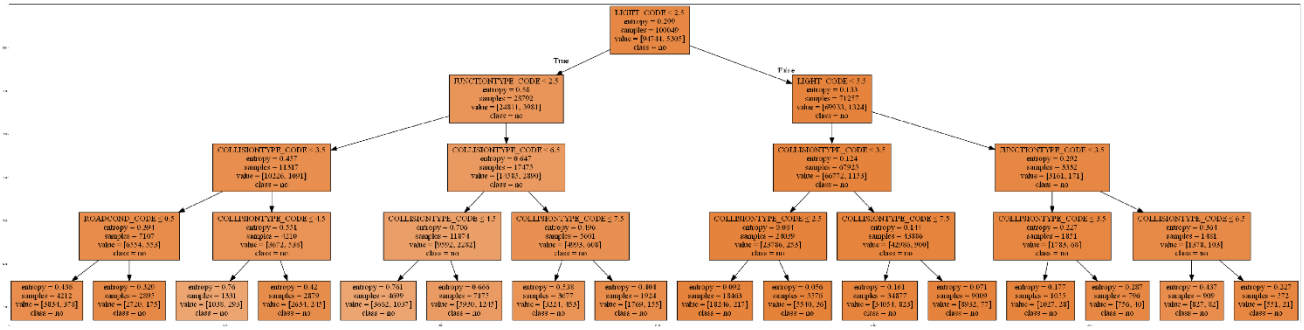

Figure 5 - Decision Tree containing non-binary data, such as Lighting Condition, Collision Type, Junction Type, Weather Conditions, Road Conditions, Address Type Conditions accuracy of 0.94762005, max_depth = 4

From the above decision trees, we can calculate the information gain from the splits, which represents the certainty after a split takes place, we seek to maximise the information gain, which is expressed by the below equation (5). A decreasing entropy, increasing homogeneity of the data, leads to an increased information gain. The below example compares the information gain from the first split of the decision tree in (*Figure 5*) with that of (*Figure 4*).

$$Information\ Gain = (entropy\ before\ split) - (weighted\ entropy\ after\ split)$$

$$Light\ Condition\ information\ gain\ (Figure\ 5) = 0.299 - \left(\frac{28792}{100049} \cdot 0.58 + \frac{71257}{100049} \cdot 0.133\right) = 0.0374$$

$$Speeding\ information\ gain\ (Figure\ 4) = 0.299 - \left(\frac{95097}{100049} \cdot 0.281 + \frac{4952}{100049} \cdot 0.577\right) = 0.000335$$

Thus, from the decision trees, it becomes evident that for the same data set, the Light Condition criterion has a lower weighted entropy and consequently, a higher information gain, for this reason, light condition is a better *attribute* than speeding when analysing collisions involving drug and alcohol consumption.

Overall however, a single decision tree is prone to *overfitting* (deep decision tree with too many levels) or *underfitting* (shallow decision tree with too few levels) of data. This challenge can be overcome through the use of another machine learning algorithm, the *RandomForestClassifier*.

Contrary to the previous classifier, the *RandomForestClassifier* creates several tree models, hence the names reference to a forest of trees, obtaining an averaged weighting of the entropy after the split, classifying the best indicator variable (*Feature*), using a *Weight* (voting) system, as indicated below.

The *mean_absolute_error* algorithm can be used for comparing a single decision tree with max_depth = 4 and the *RandomForestClassifier*, the below is the sample *RandomForestClassifier* for the Pedestrian Right of Passage, what we observe is that the idea indicator variables are inattentiveness, light condition, person count, and speeding. Overall, the *RandomForestClassifier,* serves as a better predictor than a single decision tree (6).

Weight	Feature
0641 ± 0.0150	ST_COLCODE
0251 ± 0.0132	INATTENTIONIND
0200 ± 0.0101	PERSONCOUNT
0190 ± 0.0126	JUNCTIONTYPE_CODE
0106 ± 0.0100	LIGHT_CODE
0023 ± 0.0040	SDOT_COLCODE
0008 ± 0.0010	VEHCOUNT
0005 ± 0.0014	SPEEDING
0002 ± 0.0012	PEDCOUNT
0 ± 0.0000	PEDCYLCOUNT
0 ± 0.0000	COLLISIONTYPE_CODE
0 ± 0.0000	HITPARKEDCAR
0029 ± 0.0069	SEVERITYCODE.1
0029 ± 0.0023	SEVERITYCODE
0041 ± 0.0144	LOCATION_CODE
0060 ± 0.0128	ADDRTYPE_CODE
0175 ± 0.0096	ROADCOND_CODE
0190 ± 0.0091	WEATHER_CODE

Figure 6 - Correlation between pedestrian right of passage, for collisions involving vehicles and pedestrians. Features in green indicate decision variables which have the strongest (weight) correlation, between the parameters.

Weight	Feature
0030 ± 0.0020	INATTENTIONIND
0019 ± 0.0032	PERSONCOUNT
0005 ± 0.0022	LIGHT_CODE
0005 ± 0.0023	ROADCOND_CODE
0 ± 0.0000	COLLISIONTYPE_CODE
0 ± 0.0000	HITPARKEDCAR
0 ± 0.0000	PEDCYLCOUNT
0 ± 0.0000	VEHCOUNT
0005 ± 0.0009	SPEEDING
.0006 ± 0.0011	SDOT_COLCODE
0006 ± 0.0080	PEDROWNOTGRNT
0010 ± 0.0022	SEVERITYCODE.1
0010 ± 0.0018	SEVERITYCODE
0013 ± 0.0009	PEDCOUNT
0023 ± 0.0042	WEATHER_CODE
0056 ± 0.0080	LOCATION_CODE
0084 ± 0.0046	ADDRTYPE_CODE
0090 ± 0.0033	ST_COLCODE
.0114 ± 0.0011	JUNCTIONTYPE_CODE

Figure 7 – Random forest classifier for vehicle collisions involving alcohol and drug consumption.

What we observe is that by running a single tree, *light condition* was shown to have the greatest influence on *Drug or Alcohol Consumption Collisions,* however, the random forest classifier identifies that driver *Inattention* should provide a better correlation.

2.3.2. Geospatial Plots using *Folium*

Geospatial representation of data, plotted with the aid of Python's *folium* library, allows for a better understanding of the geographical distribution of incidents, throughout the Seattle municipality. However, particular challenges were met when analysing the data.

Firstly, the issue of the sheer amount of data, even with a heat map, when plotting 142,928 data points in the diagram, the image became cluttered and challenging to interpret, for this reason a sample with 600 randomly collected data points was chosen, as shown (*Figure 7*).

One limitation of the Geospatial plot was that a few locations, particularly Tunnels, did not have their coordinates provided, and a few collisions did not have their coordinates specified, which led to an inaccurate representation of the dataset.

Overall, the *folium* allows for interesting analysis and interpretations of the distribution of collisions across Seattle, allowing the Seattle Police Department (SPD), Seattle Government, and traffic agencies, to identify the locations where accidents are predominantly occurring and to implement measures for minimising collisions within the areas with high accident frequencies.

Figure 8 - 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.

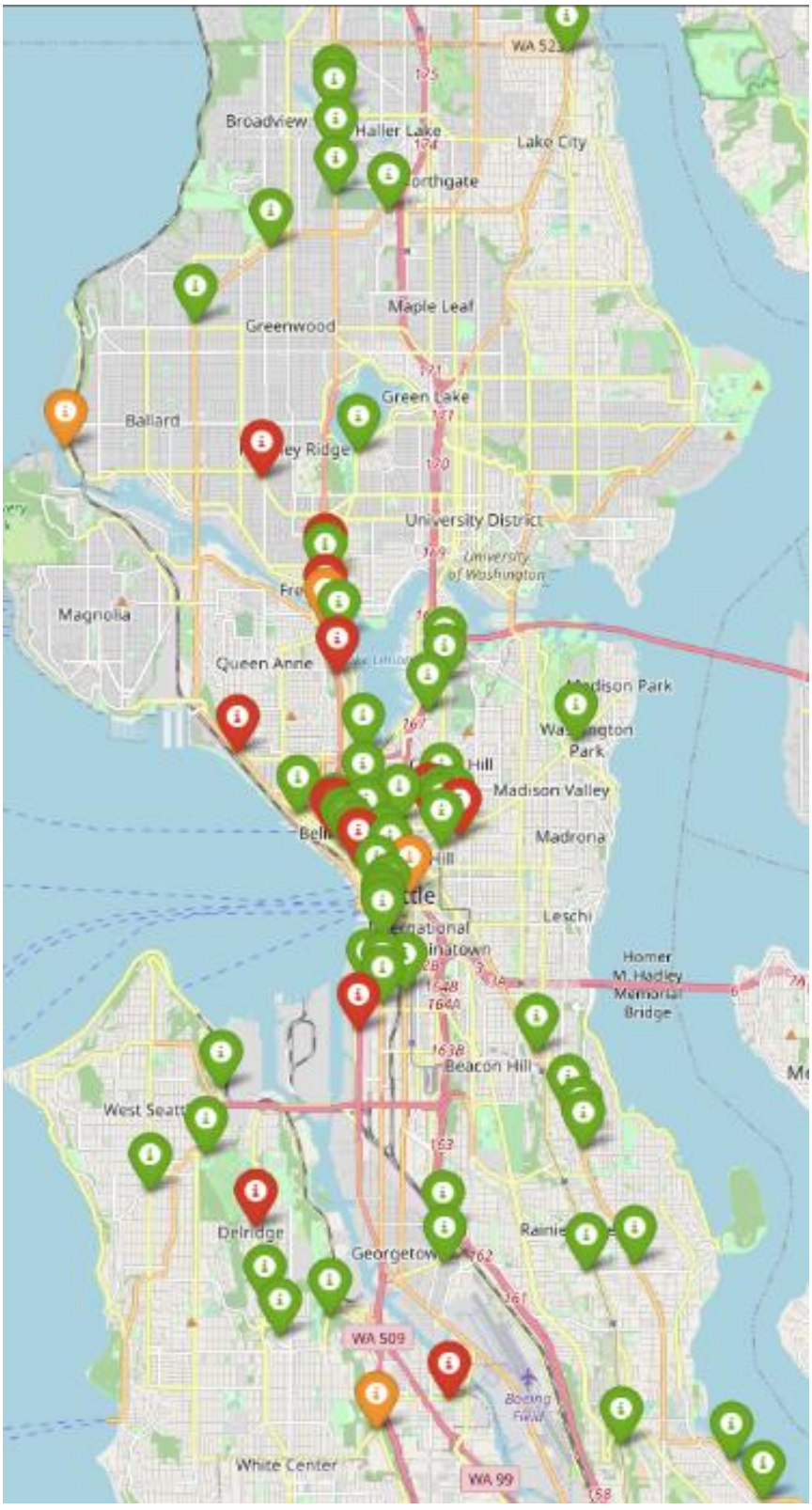

Figure 9 - Geospatial folium representing all locations where there were more than 5 collisions involving Drugs and Alcohol over the period from 2004-2020, green indicates between 6-7 collisions, orange indicates between 8-10 collisions and red indicates

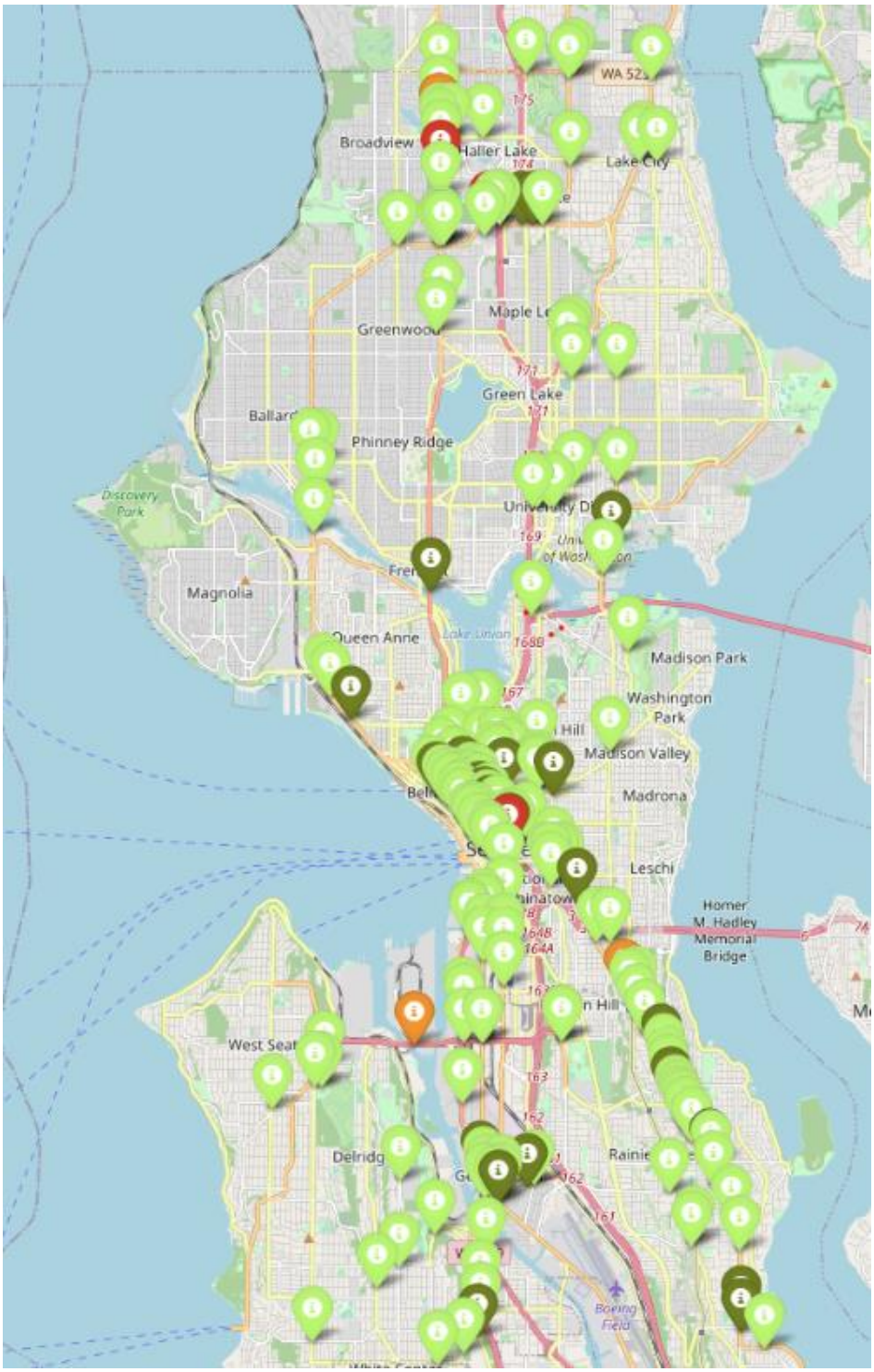

Figure 10 - Geospatial folium representing all locations where more than 60 Pedestrians were involved in vehicle Collisions occurred between 2004-2020, where light-green represents between 60-99 pedestrians, dark-green represents 100-149, orange 150-200, and red more than 200, up to a maximum number of 217

Overall, numerous plotting tools were used, amongst which were box plots, bar charts, co-occurrence matrices, geospatial representations, and scatter plots.

3. Methodology

The decision tree and machine learning algorithms provided an initial outline of the strongest correlations within the data set, these correlations were then used to plot visual data representations. Non-numerical (categorical) data was plotted using the data frequency, this can be done using either bar charts, histograms, or co-occurrence matrices, which allows for plotting the frequency of two categorical events. Both the *seaborn* and *matplotlib* libraries were used for the visual representation.

4. Data Analysis and Result Discussion

From the analysis of the data analysis, although vehicle collisions have followed a decreasing trend, from their all-time high of 2015, significant improvements is still required, as annually, a mean of 17,638 collisions occur within the municipality of Seattle. Such rates of collision have significant consequences for healthcare facilities, vehicle insurance providers, consequences to pedestrians and local residents, and ends lives prematurely.

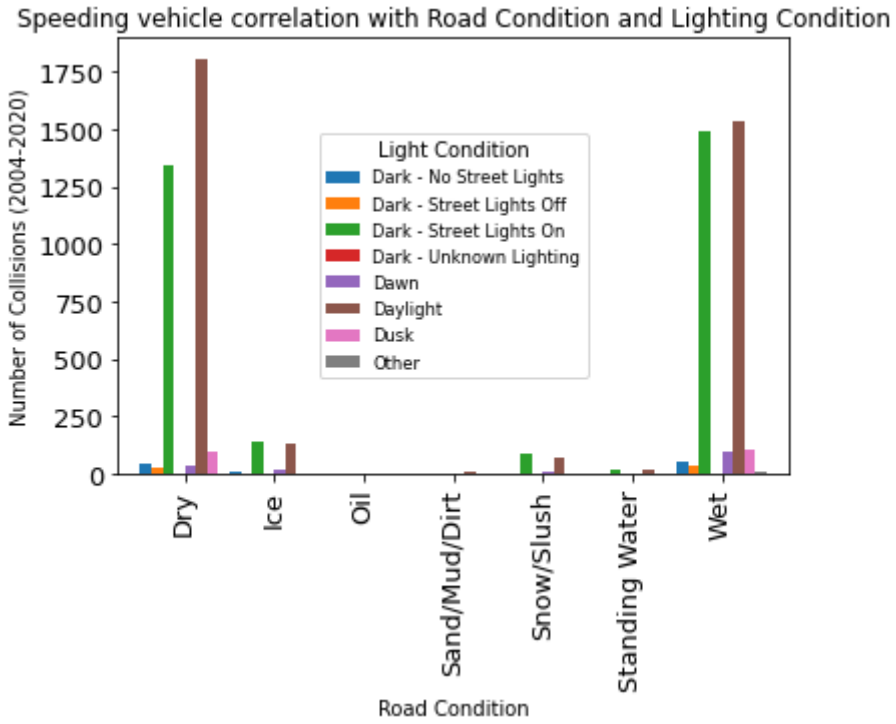

Figure 11 - Correlation between Speeding, Lighting and Road conditions (2004-2020).

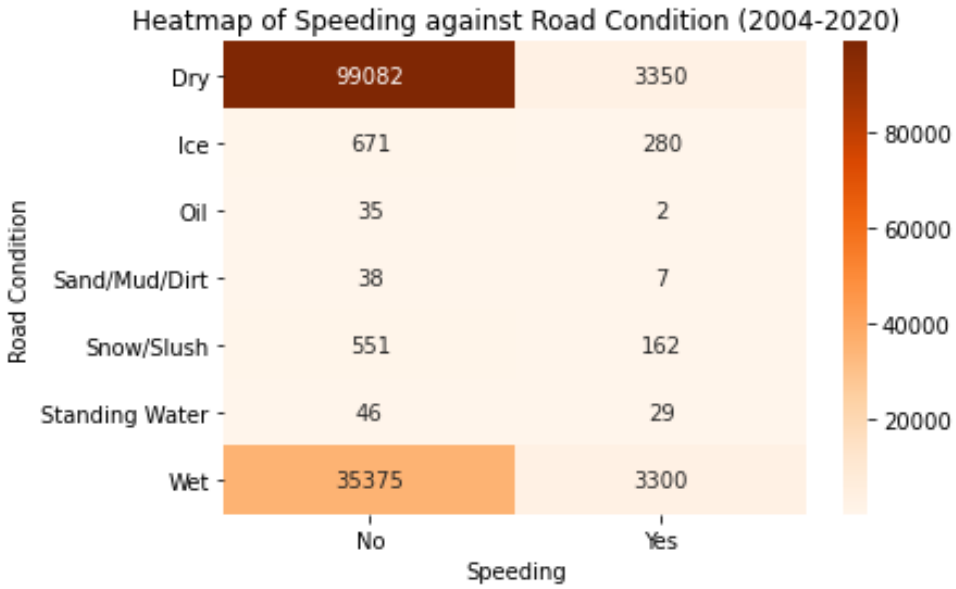

Figure 12 - Co-Occurrence Matrix, correlating speeding vehicles and road condition (2004-2020).

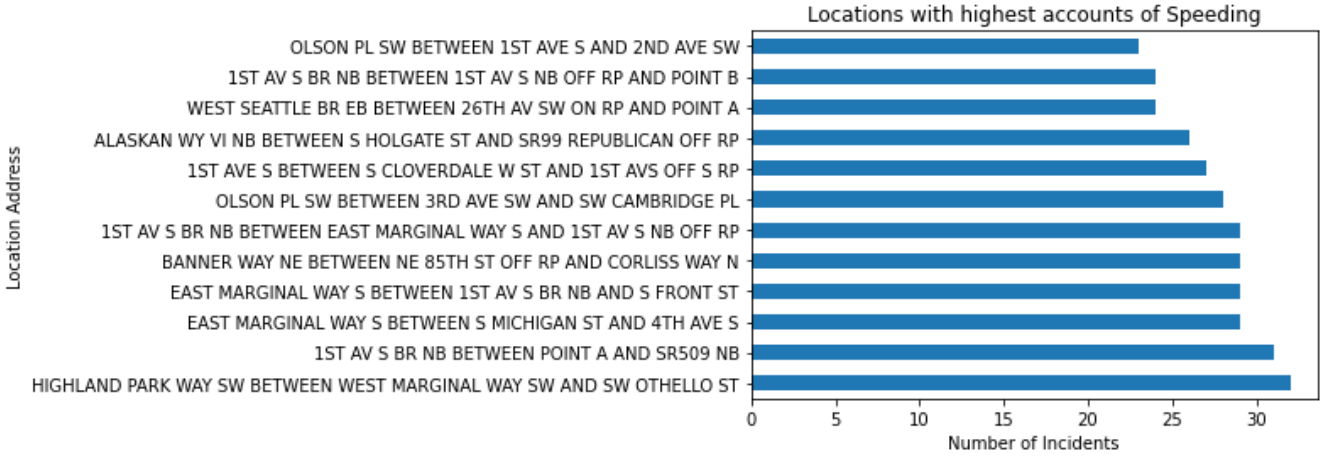

Figure 13 - Locations with the highest rate of speeding infractions (2004-2020).

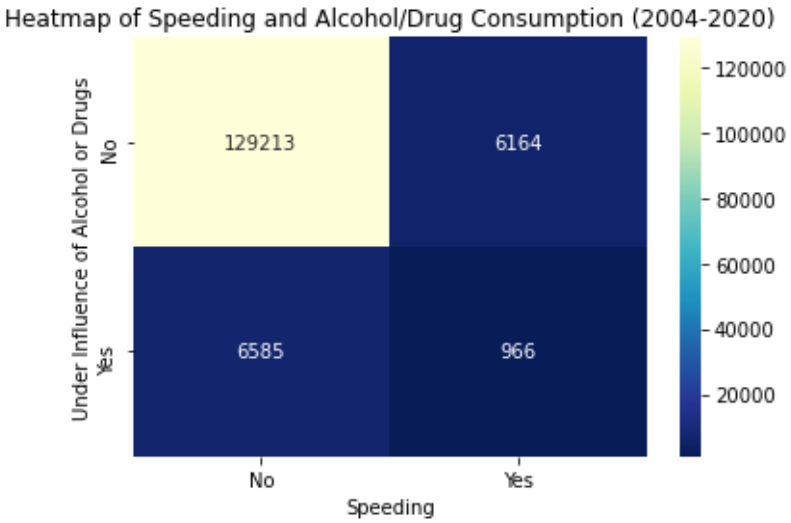

Figure 14 – Co-occurrence matrix of two binary data sources, correlating Speeding collisions with Alcohol and Drug consumption collisions.

The number of collisions where vehicles were above the speed limit totalled 7,130 over the period of 2004-2020. When analysing collisions in which speeding infractions were observed, one can identify that for both wet and dry road conditions an equivalent number of vehicles collided, with 1,750 and 1,500 vehicles, respectively (*Figure 11*).

In addition, collisions were far more likely to occur in daylight lighting conditions and on streets with public lighting at night, indicating that drivers are far more likely to commit speeding infractions when light conditions are bright and favourable, this could be a consequence of more confidence when driving through bright streets. What is also really interesting is that an equivalent number of collisions occurred in both wet and dry road conditions with public lighting. This further alludes to the greater confidence of drivers, who favour lighting over road conditions, when speeding .

Both speeding and non-speeding vehicle collisions primarily occurred in dry and wet road surfaces, which is expected, in addition a lower number of collisions in ice and snow conditions was observed, as the city receives regular snow during the winter months. What should be noted is that the threshold velocity for speeding varies in accordance with the road conditions, as lower speed limits would be observed during the winter months (*Figure 12*).

Another interesting aspect for analysis is the co-occurrence matrix correlating speeding and alcohol or drug consumption (*Figure 14*). We observe that motorists under the influence of drugs or alcohol in their majority 6,585, were within the speed limit, compared to 966 under the influence of drugs and above the speed limit. However, 12.8% of motorists who consumed drugs committed speeding infractions, compared to a speeding infraction rate of only 4.6% for motorists who were not under the influence of alcohol or drugs, demonstrating that alcohol and drugs make individuals more susceptible to speeding.

In addition, when analysing the time of day of the collisions (*Figure 15-16*), there is a clear distinction between collisions where the driver was under the influence of drugs or alcohol, and collisions where alcohol was not a factor. In particular, non-alcohol related collisions occur during peak traffic hours and in the later period of the afternoon, 14:00 – 17:00, a time associated with peak hours, when individuals are returning from work during week-days, with a significantly lower number of collisions during the night time. The opposite is observed with collisions for drivers under the influence of alcohol.

When the data for time of day (*Figure 15-16*) where collisions with drugs and alcohol are most prevalent, is coupled with the Geospatial *folium* (*Figure 9*), with the locations of where drugs and alcohol collisions occurred most frequently over the period of 2004-2020, police may use this

information to conduct alcohol testing and, closer surveillance, aimed towards preventing collisions in these locations.

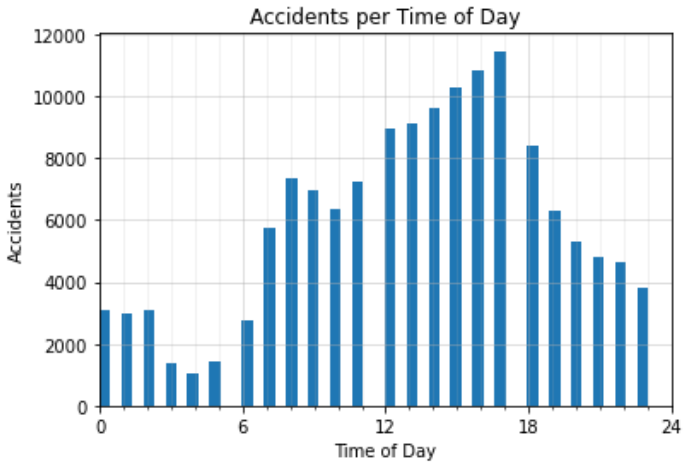

Figure 15 – Histogram of the distribution of accidents per time of day (2004-2020)

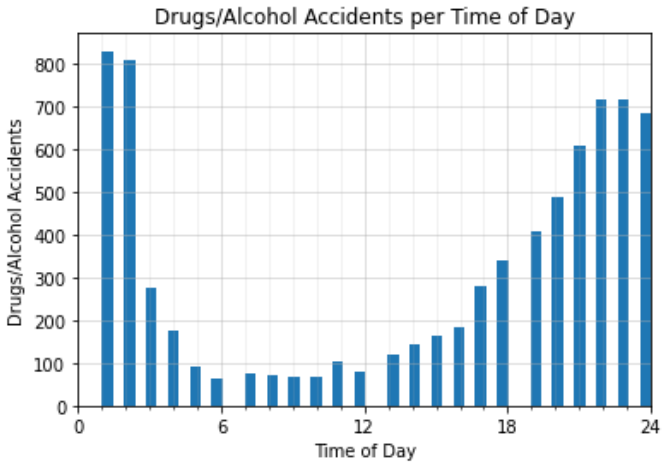

Figure 16 – Histogram of the distribution of Drug Related Accidents per time of Day (2004-2020)

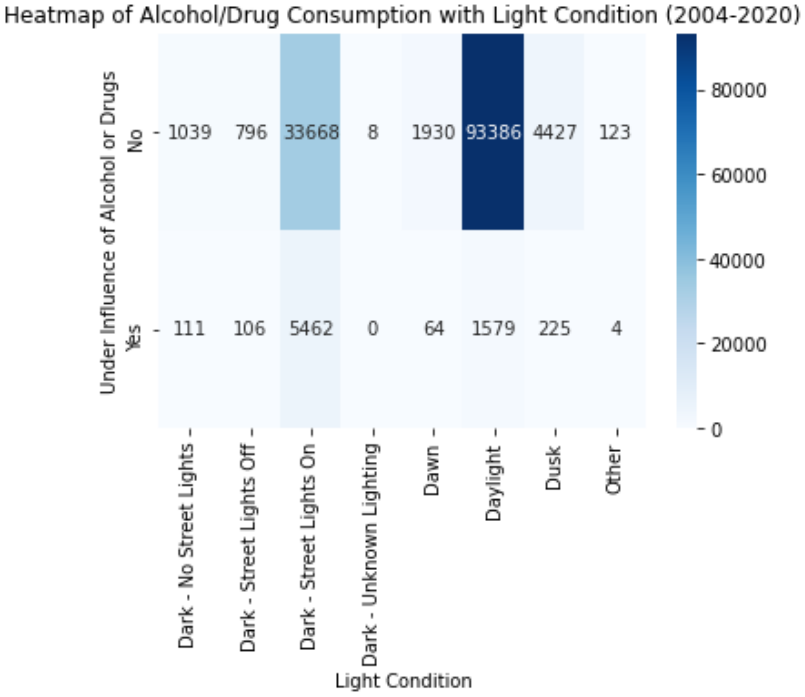

Figure 17 – Correlation between alcohol and drug consumption and light conditions of the road at the time of the collision.

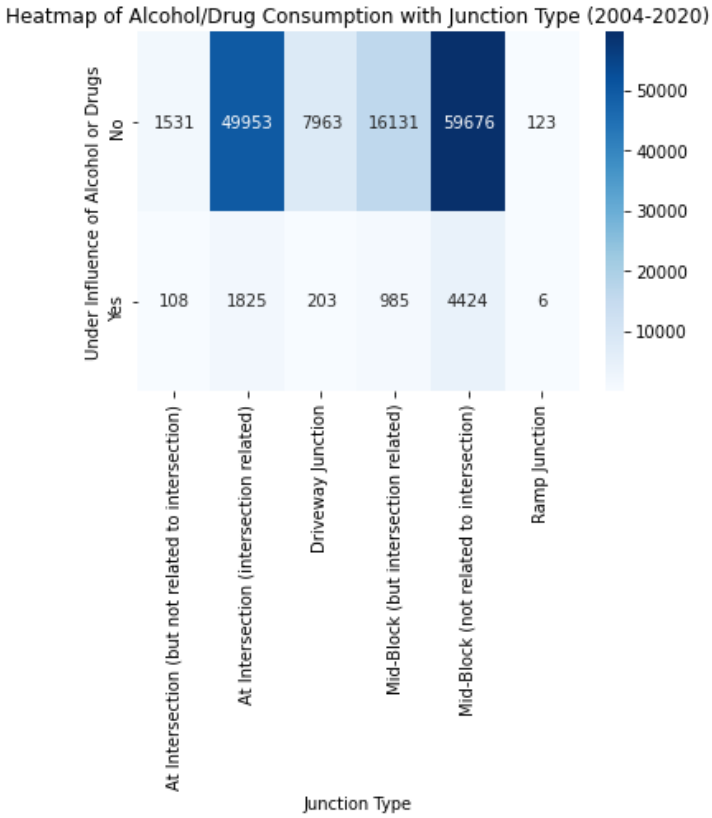

Figure 18 – Correlation between alcohol and drug consumption and the junction type of the collision.

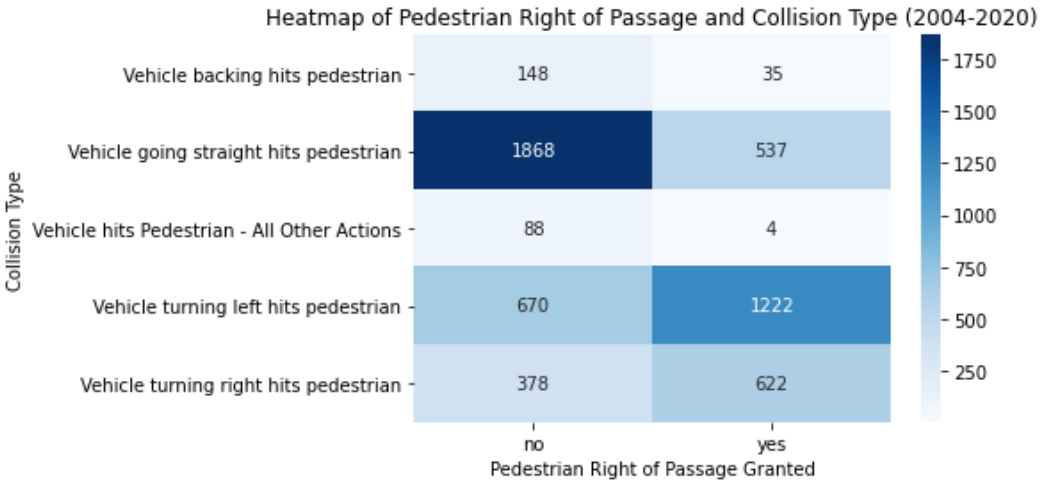

Figure 19 - Co-occurrence matrix correlating collisions according to collision type and pedestrian right of passage. This was the highest correlated parameter given by the Random Forest Classifier.

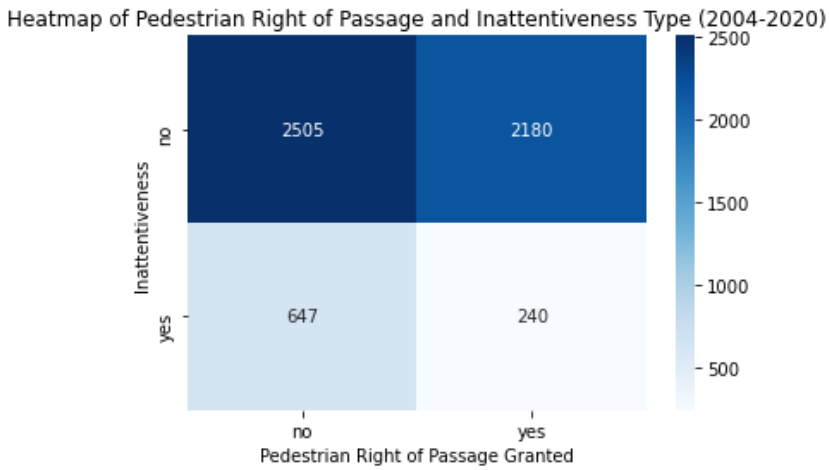

Figure 20 - Pedestrian Right of Passage and Inattentiveness. This was the second highest correlated parameter given by the Random Forest Classifier.

Several correlations exist in addition to the above mentioned, in particular with regards to pedestrian collisions (*Figure 10* & *19-20*). The machine learning algorithm identified inattentiveness as one of the variables most closely correlated to pedestrian collisions. Which is a key factor when we consider pedestrian zebra crossings without a pedestrian-specific traffic light, as is shown by the co-occurrence matrix (*Figure 20*) which indicates how, for pedestrians where the right of passage was not granted, inattentiveness contributed to 640 collisions, as opposed to collisions, where the pedestrian right of passage was granted, indicating the presence of a traffic light, where inattentiveness contributed to 240 collisions.

In addition, we see how vehicles turning left (1,222) and right (622) (*Figure 19*) are particularly prone to colliding with pedestrians when their right of passage is granted. This could be a consequence of vehicles not properly indicating their intention to turn. What is particularly interesting is that vehicles following straight are highly likely to hit pedestrians, particularly when their right of passage is not granted, this could be a consequence of inattentiveness when travelling down a street, or the fact that pedestrians cross outside their designated crossing points. The *folium* (*Figure 10*) can also greatly contribute towards better understanding the locations where pedestrians are particularly at risk, allowing authorities to address the safety at each of these locations.

The inattention of pedestrians can also be identified geospatially using *folium,* this would allow for the identification of areas where motorists are most likely to be using their telephones or are distracted. Implementing the appropriate steps towards reducing collisions in these areas.

Lastly, interesting correlations were identified with regards to Tunnel collisions, whereby the Battery Street Tunnel, between Alaskan Way and Aurora Avenue, ranked as the location with the greatest number of vehicle collisions on the entire DataFrame, with a total of 547 collisions over the period of 2004-2019, the tunnel was decommissioned by the city on the 1st of February 2019 (7).

The Severity of the Collisions within the tunnel was also interesting, as the majority of collisions only involved Property Damage (383), representing 70% of all of the collisions which occurred within the tunnel, of which all of the 383 collisions were related to hitting a “fixed object”, which could refer either to a stationary car or debris within the tunnel, a particularly interesting f.

5. Conclusion

The extensive data collection conducted by the Seattle Government and Police Department provide a rich archive of information over the course of 16-years, an invaluable data set for the city to ensure the development of its road network, improve road safety, and initiate new public policies and initiatives aimed at creating a safer city for pedestrians, cyclists, and motorists alike.

Machine learning provided the tools which allowed for trends to be visualised within the dataset and was crucial for both the data processing and analysis stages of the report. The data itself provided a rich portfolio of trends, particularly with regards to pedestrians, alcohol and drug

consumption, speeding, and time of day analysis. Additional trends could also be explored in future, in particular regarding bicycle collisions and road construction works.

Inconsistent data collection hindered the testing of correlations, although cleaning the data can easily be automated using *SQL* or *Python pandas,* the final clean dataset for the report was 26.5% smaller than the original DataFrame, this was a consequence of missing data and information, which could be avoided in future through the standardisation of the field data collection.

With regards to the DataFrame, the depth of data provided could be improved, in particular with regards to the levels of detail and information related to alcohol and drug consumption, which ideally should be input separately, alongside information on the toxicology test result, indicating the level of alcohol consumption, provided by the breath alcohol test conducted by the police officer at the time of the collision, this would allow potentially new findings and observations.

Likewise, the percentage at which the speed limit was exceeded by a speeding vehicle per location, alongside a better classification of *Inattentiveness,* by indicating the reason for the drivers distraction, such as a telephone, GPS, radio or some other device at the time of the collision, would all help towards identifying new correlations.

6. Future Discussion

A particularly interesting field for analysis would be to explore the potential for developing bicycle lanes or public transport infraestructure in regions where pedestrian collisions and vehicle collisions are particularly prevanlant, decreasing the hazards associated to the locations, improving saftey, and decrease collision rates, ensuring a more reliable tranporation network.

7. Additional Information

The *Python* programming was used for the cleaning, processing, and analysis of the data. The compiler and interface used for programming and running the programme was *Anaconda Spyder.* The capstone project relied upon the use of *Python's* in-built libraries, these being, *sklearn* (machine learning library), *numpy* (array library), *pandas* (data manipulation library), *seaborn* and *matplotlib* (data visualisation libraries), and the *folium* (geospatial visualisation library).

Please refer to the GitHub repository for access to the code written for the data analysis.

8. Works Cited

1. National Highway Traffic Saftey Administration (NHTSA). [Online] U.S. Transportation Secretary Elaine L. Chao Announces Further Decreases in Roadway Fatalities, 22 October 2019. [Cited: 11 September 2020.] https://www.nhtsa.gov/press-releases/roadway-fatalities-2018-fars.

2. Lloyd, Sarah Anne. Seattle's car population is growing just as fast as its human population. [Online] Curbed Seattle, 10 August 2017. [Cited: 11 September 2020.] https://seattle.curbed.com/2017/8/10/16127958/seattle-population-growth-cars-transit.

3. Data, Seattle Police Department (SPD) Collisions. [Online] https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

4. Klein, Bernd. What are Decision Trees? [Online] Python Machine, 2011. [Cited: 2020 September 11.] https://www.python-course.eu/Decision_Trees.php.

5. *Coursera IBM Data Science Professional Certificate, Machine Learning with Python, Week 3.*

6. Kaggle. 6 Random Forests. [Online] [Cited: 11 September 2020.] https://www.kaggle.com/dansbecker/random-forests.

7. Poyner, Fred. Seattle's Battery Street Tunnel is decommissioned on February 1, 2019. [Online] History Link.org, 12 October 2019. [Cited: 12 September 2020.] https://www.historylink.org/File/20937.