# PREDICTING COLLISION SEVERITY

Data Science Capstone

Julian Ariza

# PREDICTING THE SEVERITY OF A COLLISION ACCIDENT

- Approximately 1.35 million people die each year as a result of road traffic crashes.
- Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.
- Road traffic injuries cause considerable economic losses to individuals, their families, and to nations.
- Road traffic crashes cost most countries 3% of their gross domestic product.
- The 2030 Agenda for Sustainable Development has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2020.

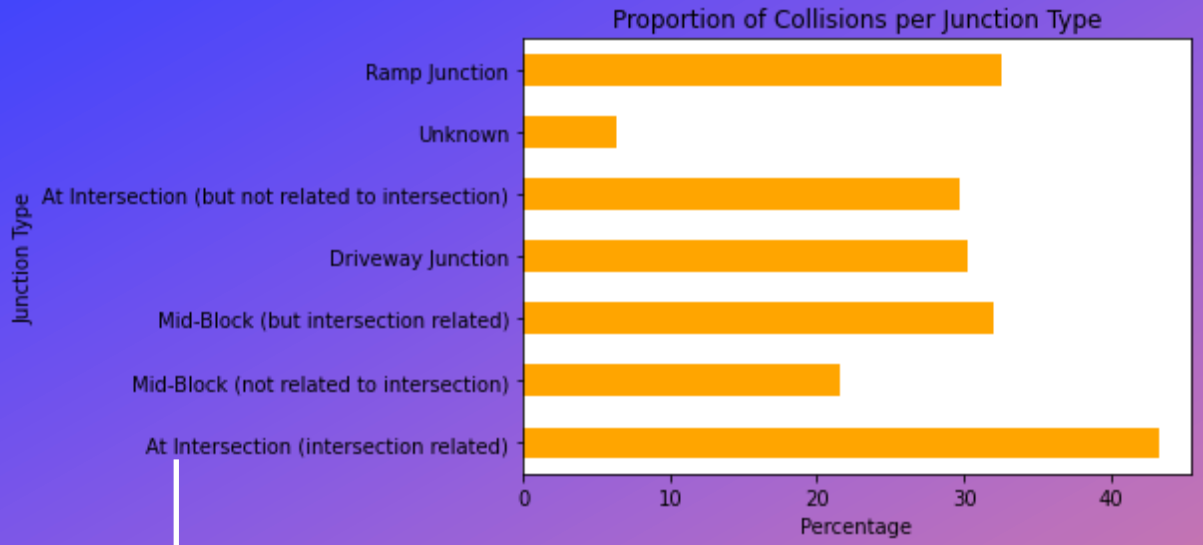
- World Health Organization

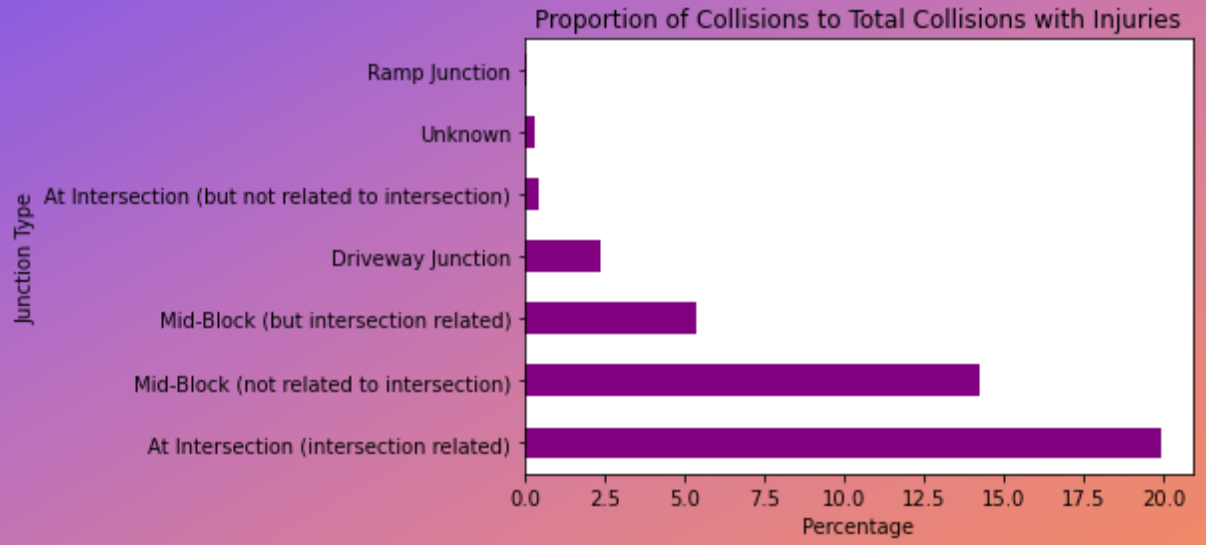
# DATA

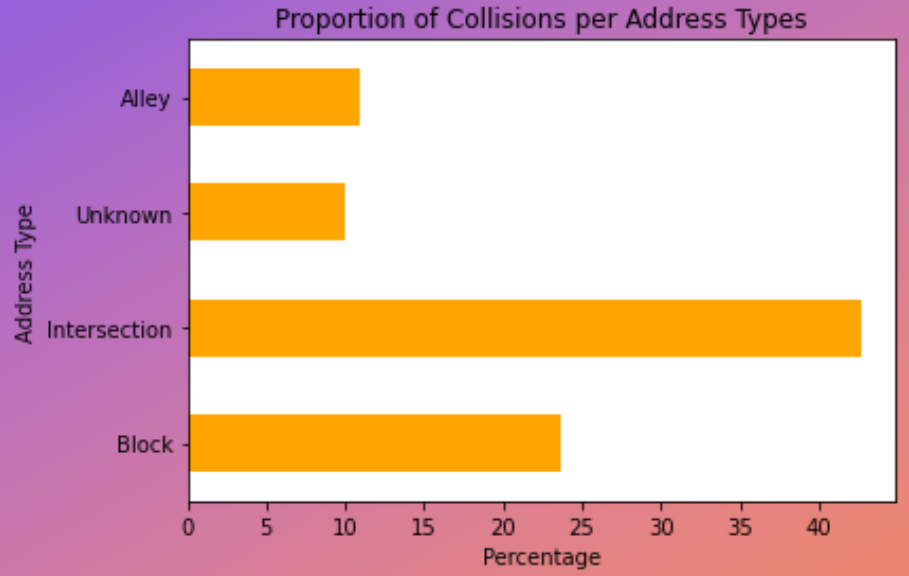
- The shared data is provided by the SDOT Traffic Management Division and Traffic Records Group from Seattle.
- It is a .csv file that contains 194.673 collisions registered from 2004 to 2019.
- Each collision provides 37 characteristics plus a given severity classification, from which we are going to obtain the features for the machine learning model.
- Many of the characteristics have the purpose of collision identification for the government entities, these are values that are unique for each collision, because of this, no pattern applies.
- Other characteristics provide further description of numerically categorized characteristics, making them redundant.
- Finally, other characteristics provide details that will not be available until someone is in place taking the report not being available at the moment of the prediction, therefore, we discard them.

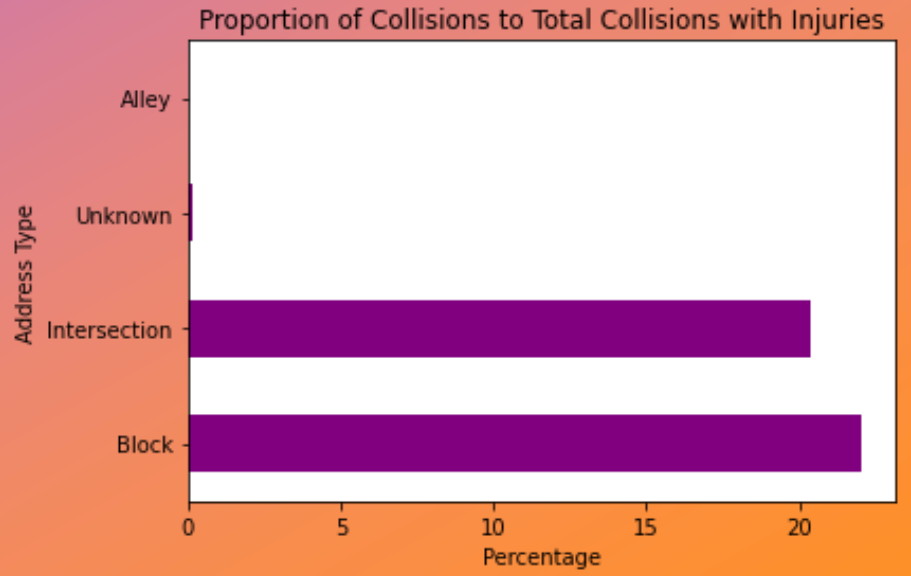
'https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv'

# PROPORTION OF COLISSIONS WITH INJURED FOR LOCATION FEATURES

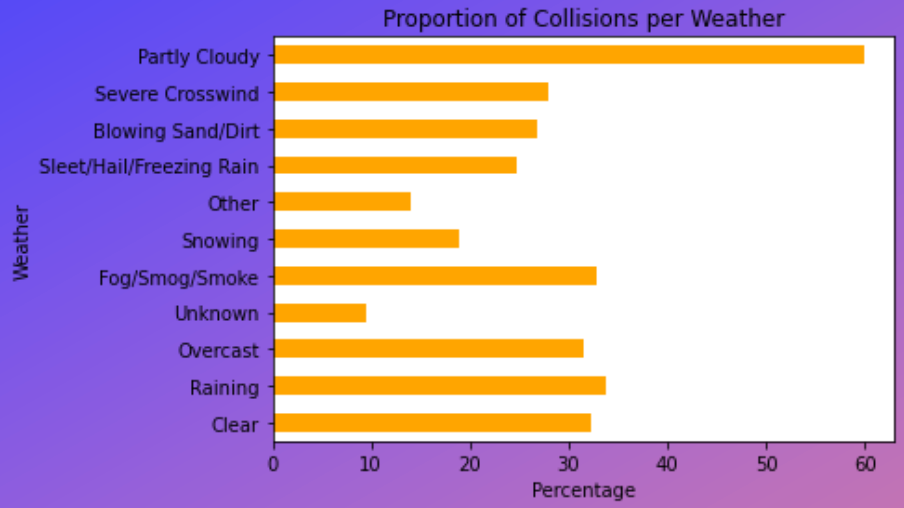


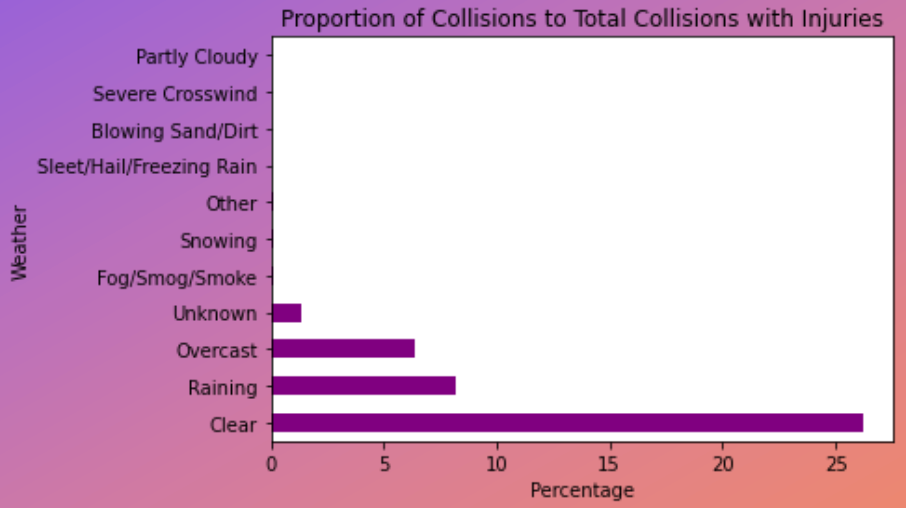



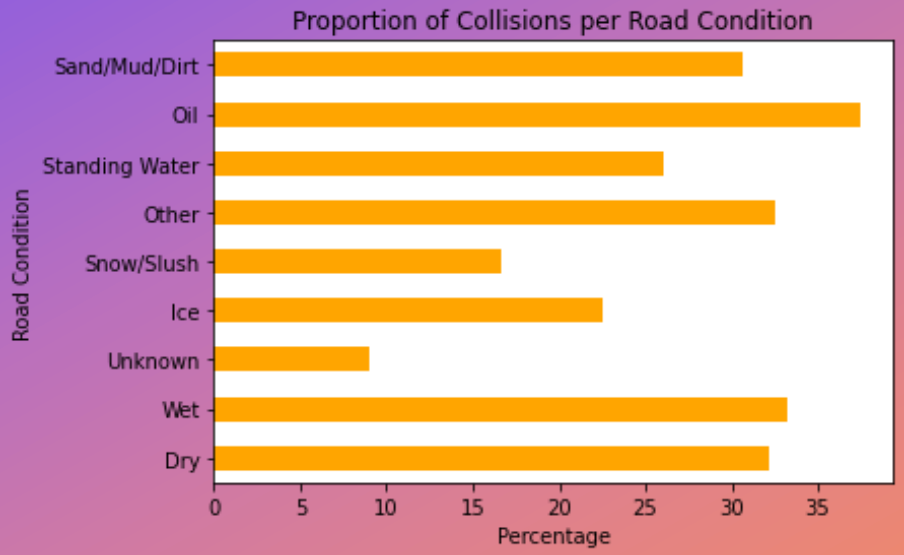



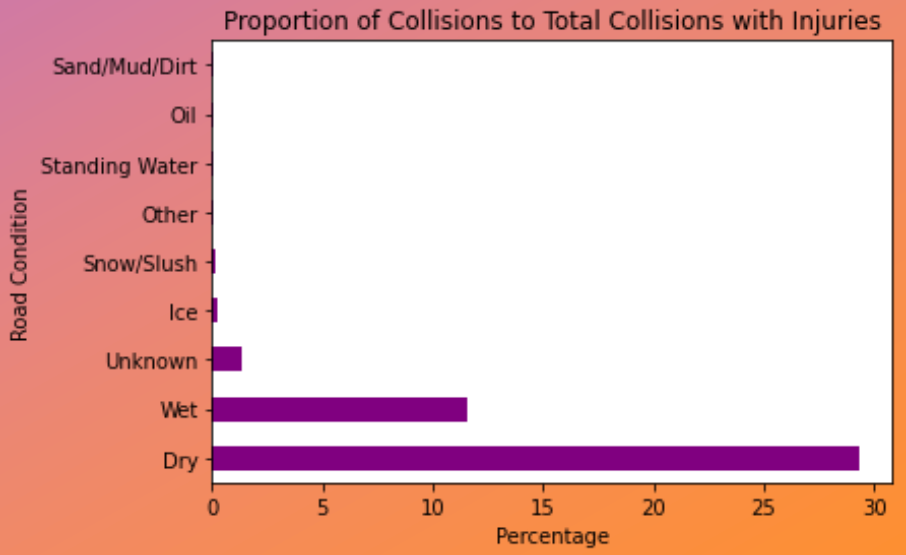



# PROPORTION OF COLISSIONS WITH INJURED FOR ENVIROMENT FEATURES

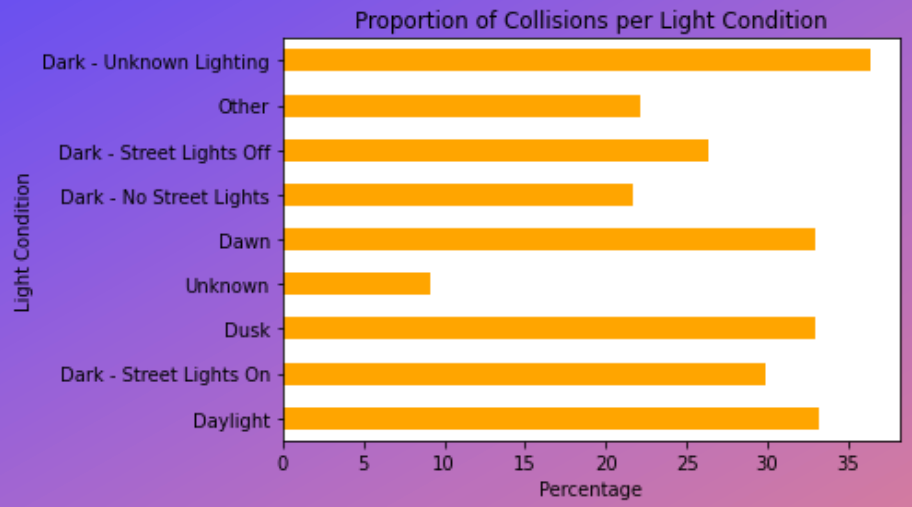


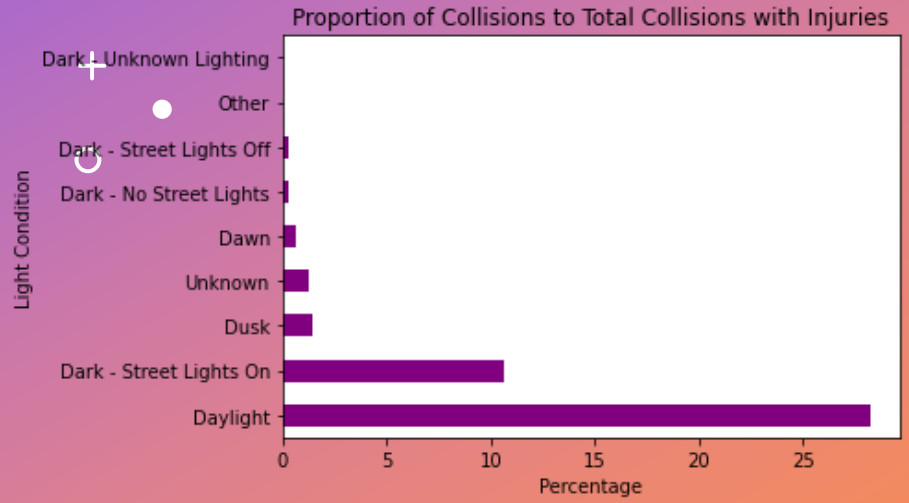



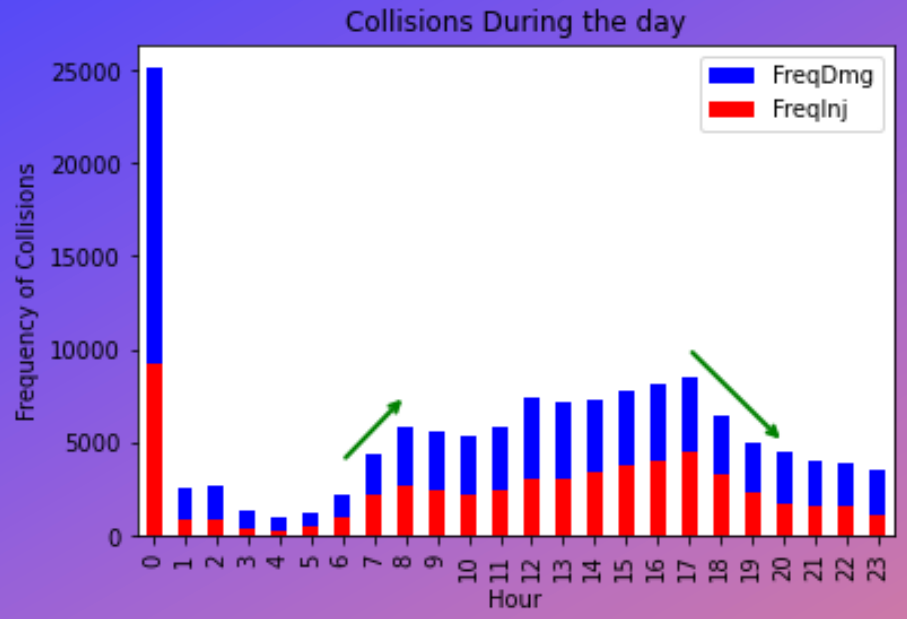



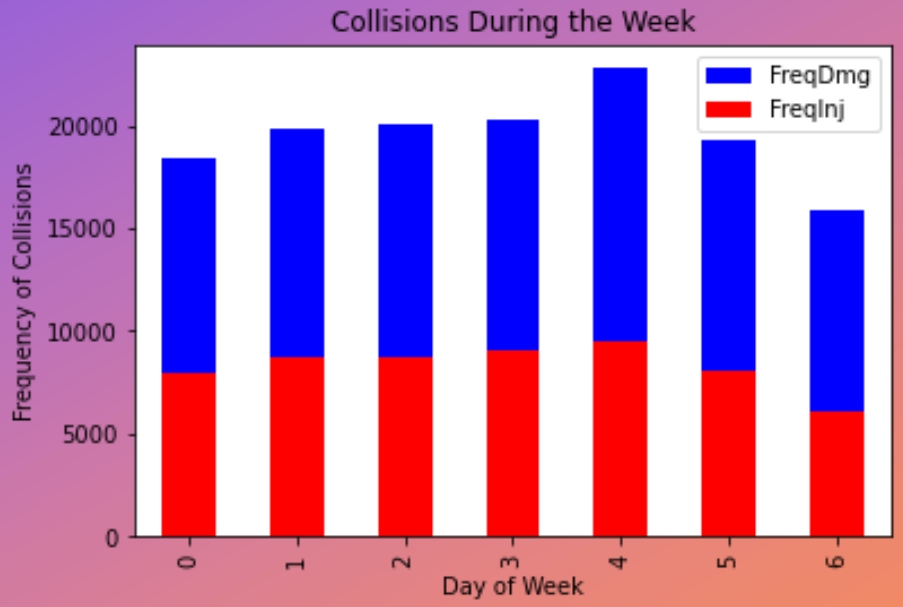
# PROPORTION OF COLISSIONS WITH INJURED FOR ENVIROMENT FEATURES



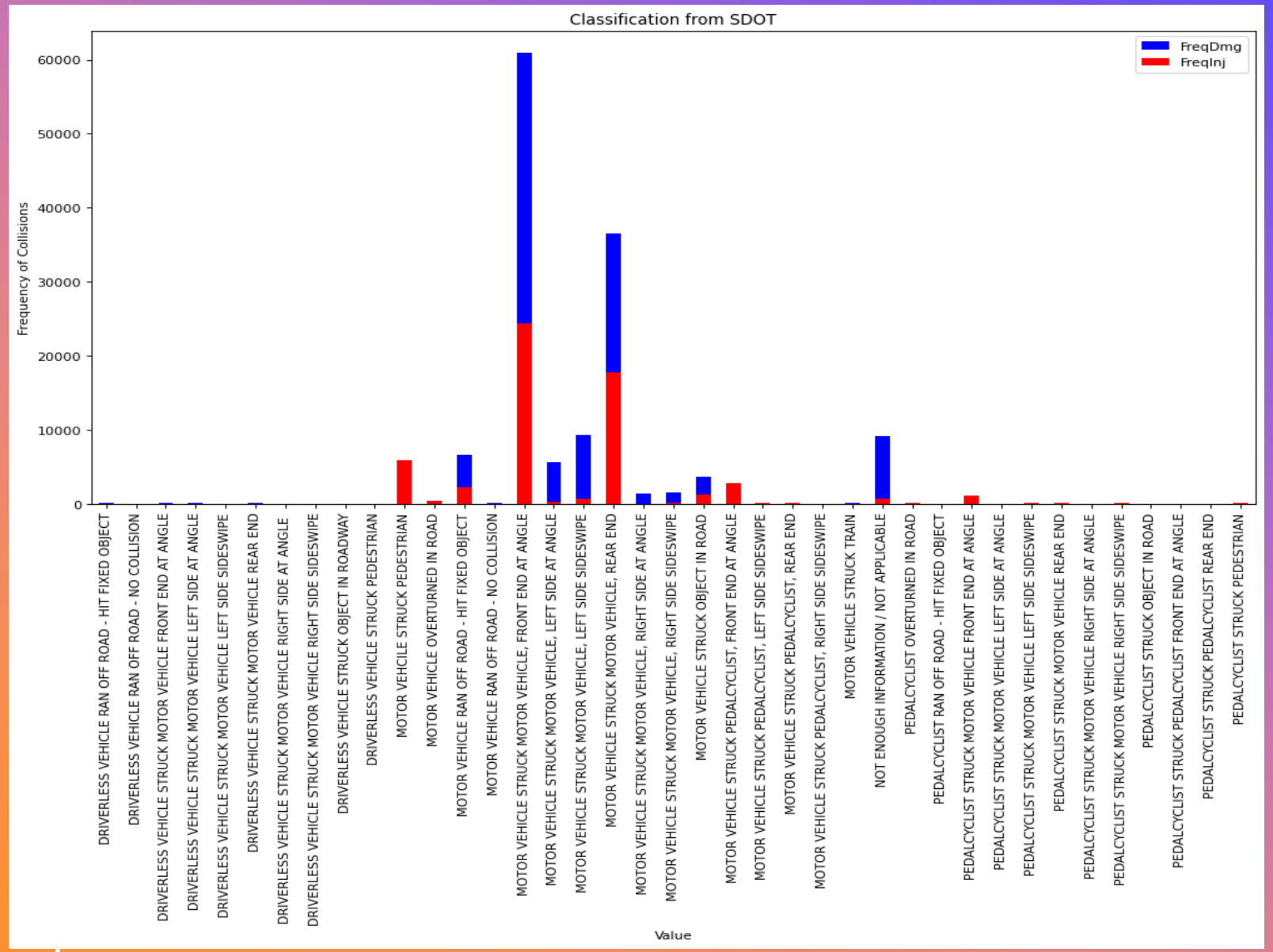
# FREQUENCY OF COLISSIONS FOR TIME FEATURES



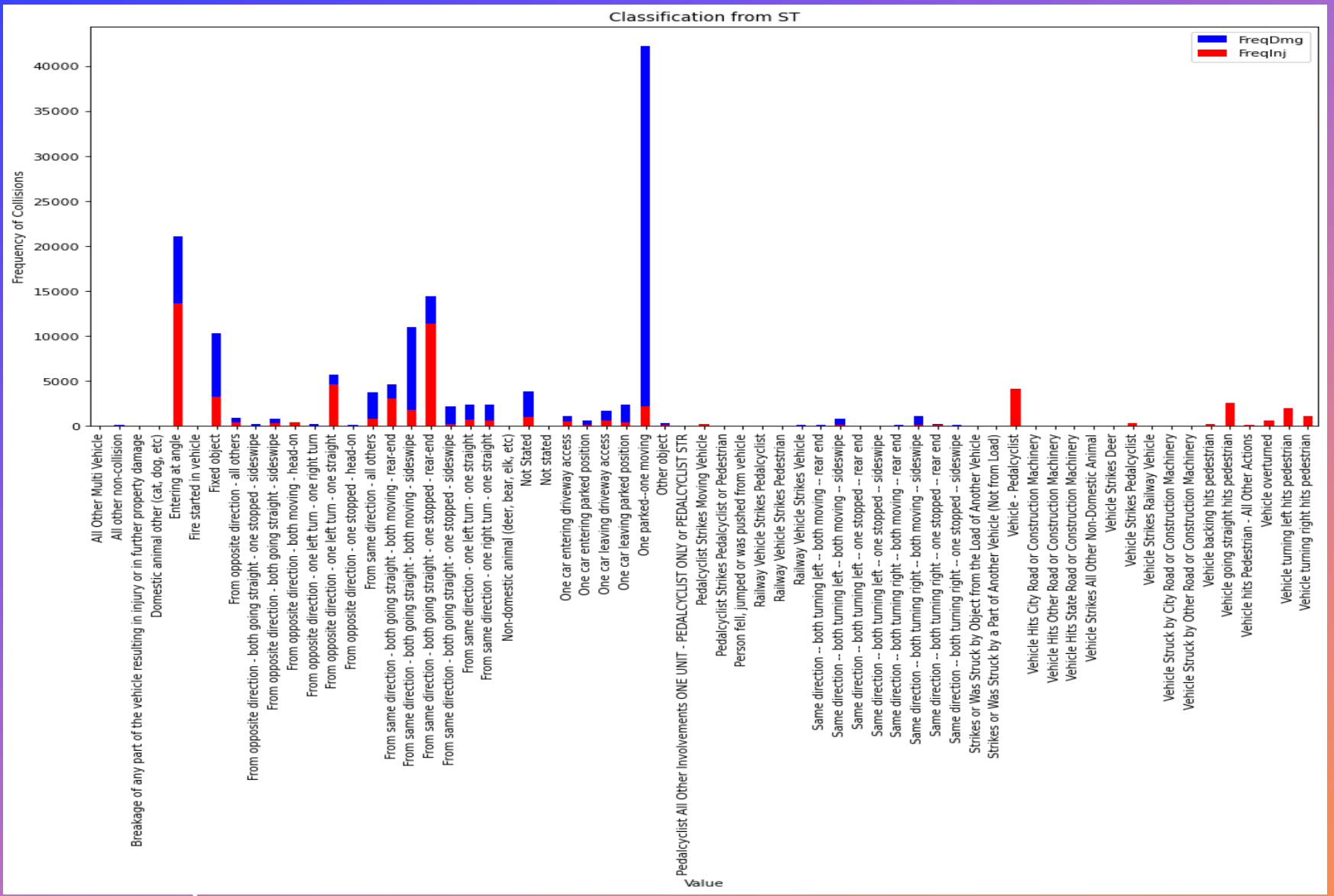

- We can se two peaks related to the US business hours.
- We can se a reduction of collisions on weekends.

# FREQUENCY OF COLISSIONS CAUSES ACCORDING TO SDOT



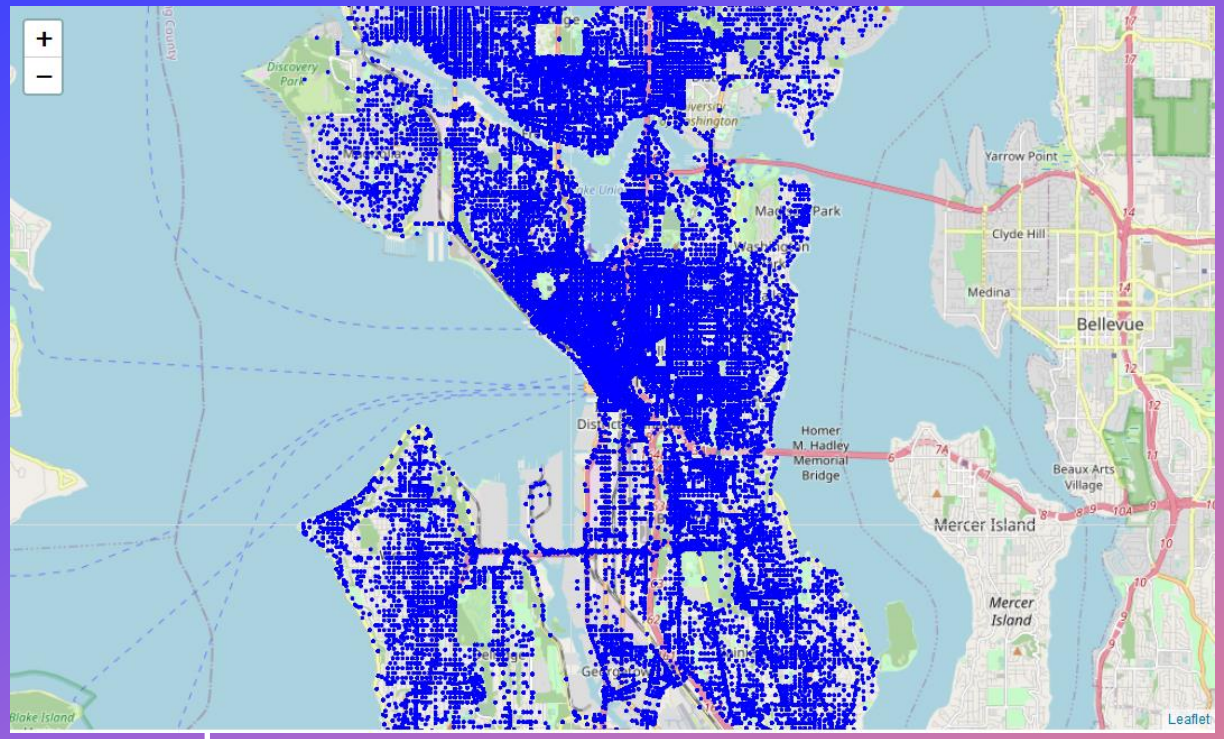

- Collisions that involve pedestrians or pedal cyclists usually are collisions with injured.

# FREQUENCY OF COLISSIONS CAUSES ACCORDING TO ST



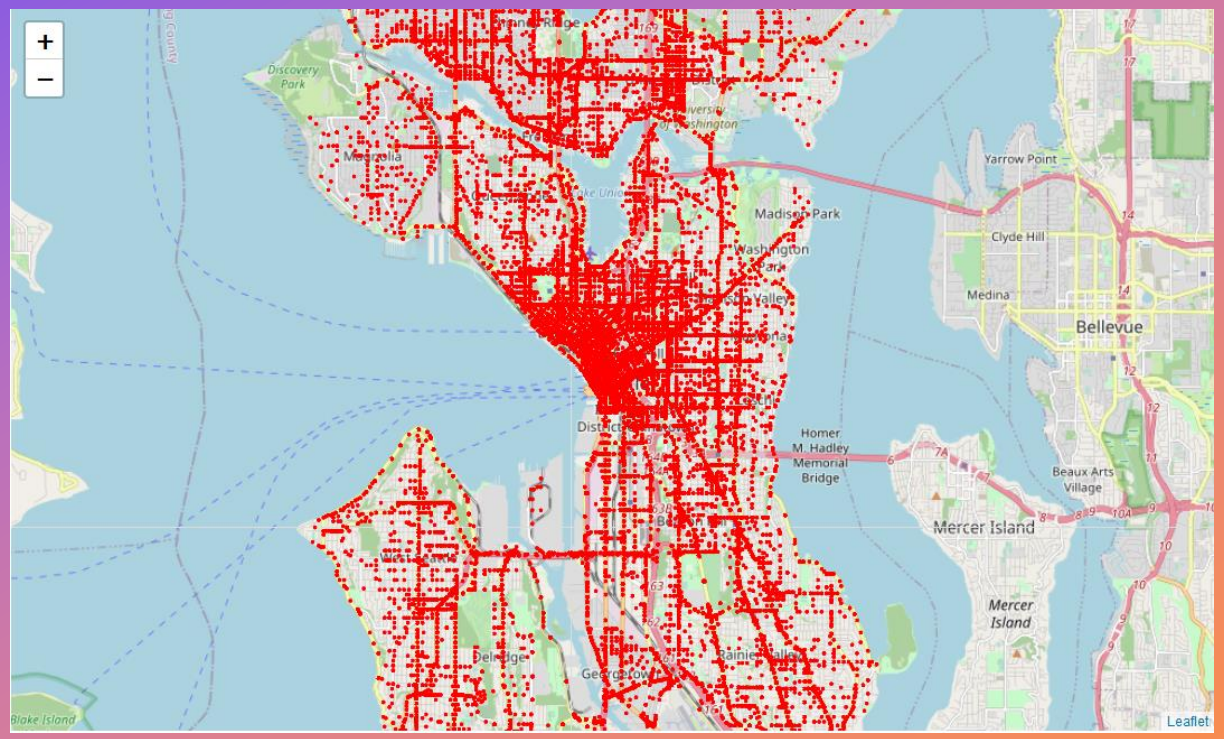

- Collisions that involve a parked vehicle, usually are collisions with property damage only.

# LOCATION OF COLISSIONS ON SEATTLE




Property Damage Only

They are dispersed around the city




With injured

They are more dense on specific areas or avenues.

# DATA PREPARATION

Overall the data for building the model has the following modifications:

1. The columns to be fed will be 'X', 'Y', 'ADDRTYPE', 'JUNCTIONTYPE', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SPEEDING', 'HOURDAY', 'DAYWEEK', 'MONTH', ‘PEDCOUNT’ and 'PEDCYLCOUNT’.
2. ‘PEDCOUNT’ and ‘PEDCYLCOUNT’ are joined to a new column ‘PEDNUM’ with a value ‘1’ when there is, at least, one pedestrian or one pedal cyclist involved on the collision. For this we sum up the columns and give a threshold with zero or greater than 0. Previous columns are dropped.
3. For 'ADDRTYPE', 'JUNCTIONTYPE', 'WEATHER', 'ROADCOND' and 'LIGHTCOND' dummies are created.
4. For ‘X’ and ‘Y’, they are normalized, the mean is subtracted, and they are divided by their standard deviation.
5. For ‘HOURDAY’, ‘DAYWEEK’, and ‘MONTH’ are normalized being divided by their maximum value.

# MODELING AND EVALUATION

- I used four different machine learning algorithms:
  - K Nearest Neighbors.
  - Decision Tree Classifier.
  - Support Vector Machine.
  - Logistic Regression.
- For their evaluation it is used:
  - Jaccard Index.
  - F1 Score.
  - Log Loss when possible.

1. Training data is 60%.
2. Cross Validation data is 20% and is used for tuning the algorithms.
3. Test data is 20% and is used as unseen data for final evaluation.

| Algorithm | Jaccard | F1-Score | Log Loss |
|---|---|---|---|
| KNN | 0.27 | 0.65 | - |
| Decision Tree | 0.27 | 0.65 | - |
| SVM | 0.18 | 0.68 | - |
| Logistic Regression | 0.19 | 0.69 | 0.53 |

# CONCLUSION

- We obtained over 60% on F1 Score for the Machine Learning Models. This is a promising result.
- The model has room for improvement including:
  - Since the data is updated weekly, real time data collection could improve the results adding complexity.
  - Since injuring collisions have locations patterns, information from other entities could explain and provide information.