# CAR ACCIDENT SEVERITY ANALYSIS: Seattle Washington

DATA SCIENCE CAPSTONE PROJECT

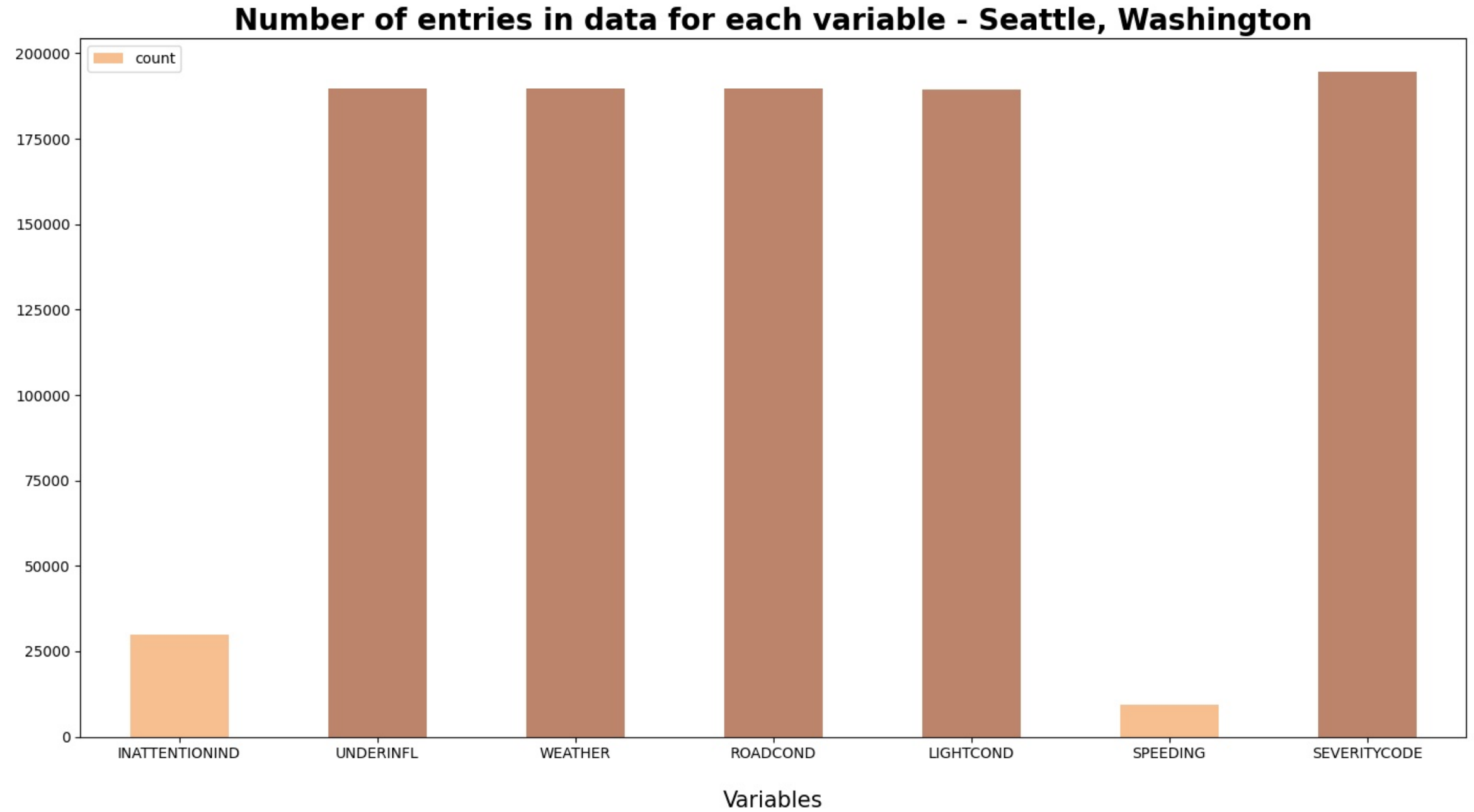
Tanvi

# Introduction

- The world as a whole suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to *$871 billion* in a single year. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours in the state of Washington while Fatal crashes went from *508* in 2016 to *525* in 2017, resulting in the death of *555* people. The project aims to predict how severity of accidents can be reduced based on a few factors.
- **Stakeholders:**
- Public Development Authority of Seattle
- Car Drivers

# Data

- The dataset used for this project is based on car accidents which have taken place within the city of *Seattle, Washington* from the year *2004* to *2020*. This data is regarding the *severity of each car accidents* along with the time and conditions under which each accident occurred.


Number of entries in data for each variable - Seattle, Washington
count
200000
175000
150000
125000
100000
75000
50000
25000
0
INATTENTIONIND
UNDERINFL
WEATHER
ROADCOND
LIGHTCOND
SPEEDING
SEVERITYCODE
Variables

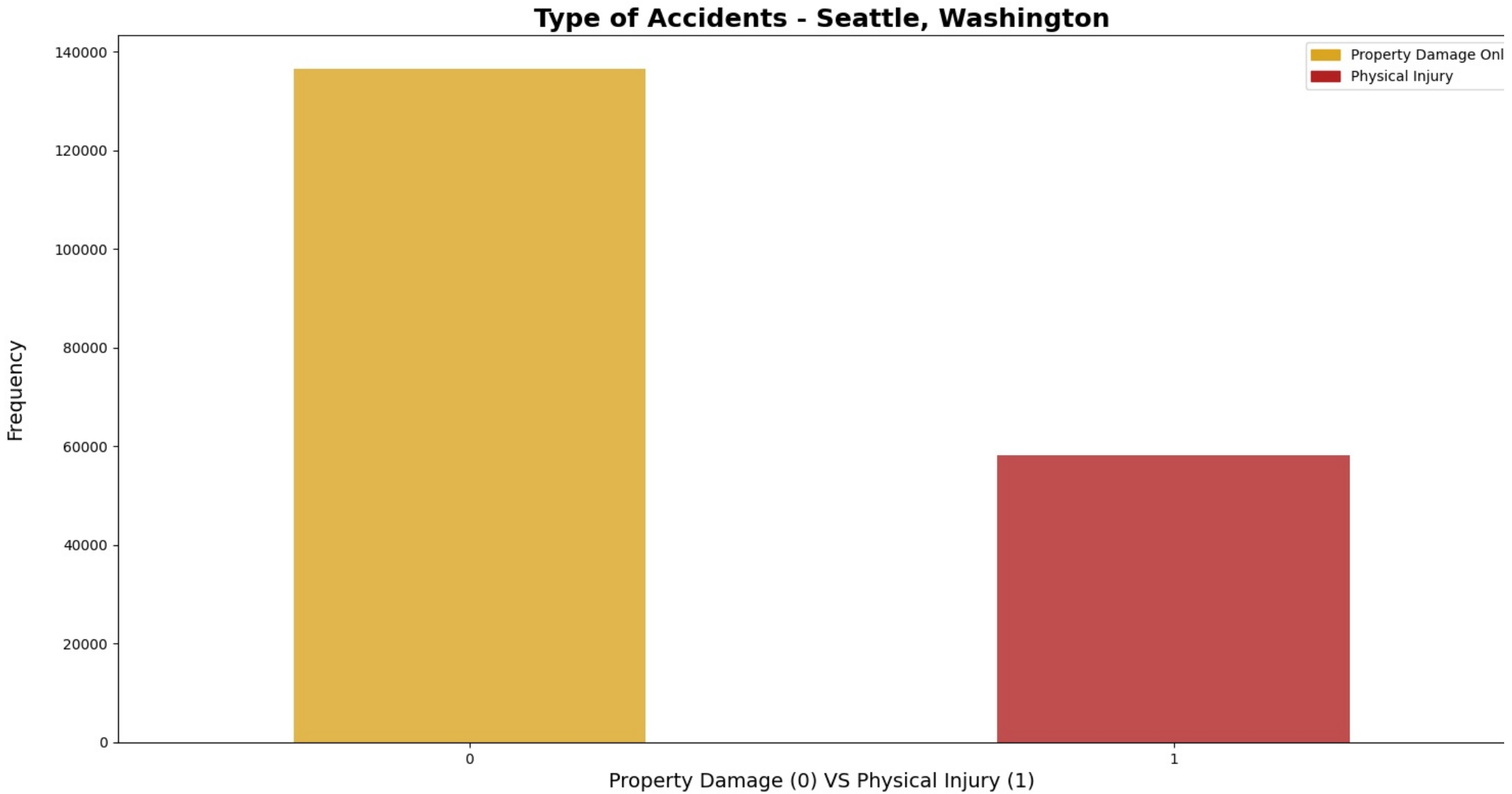
# Feature Selection

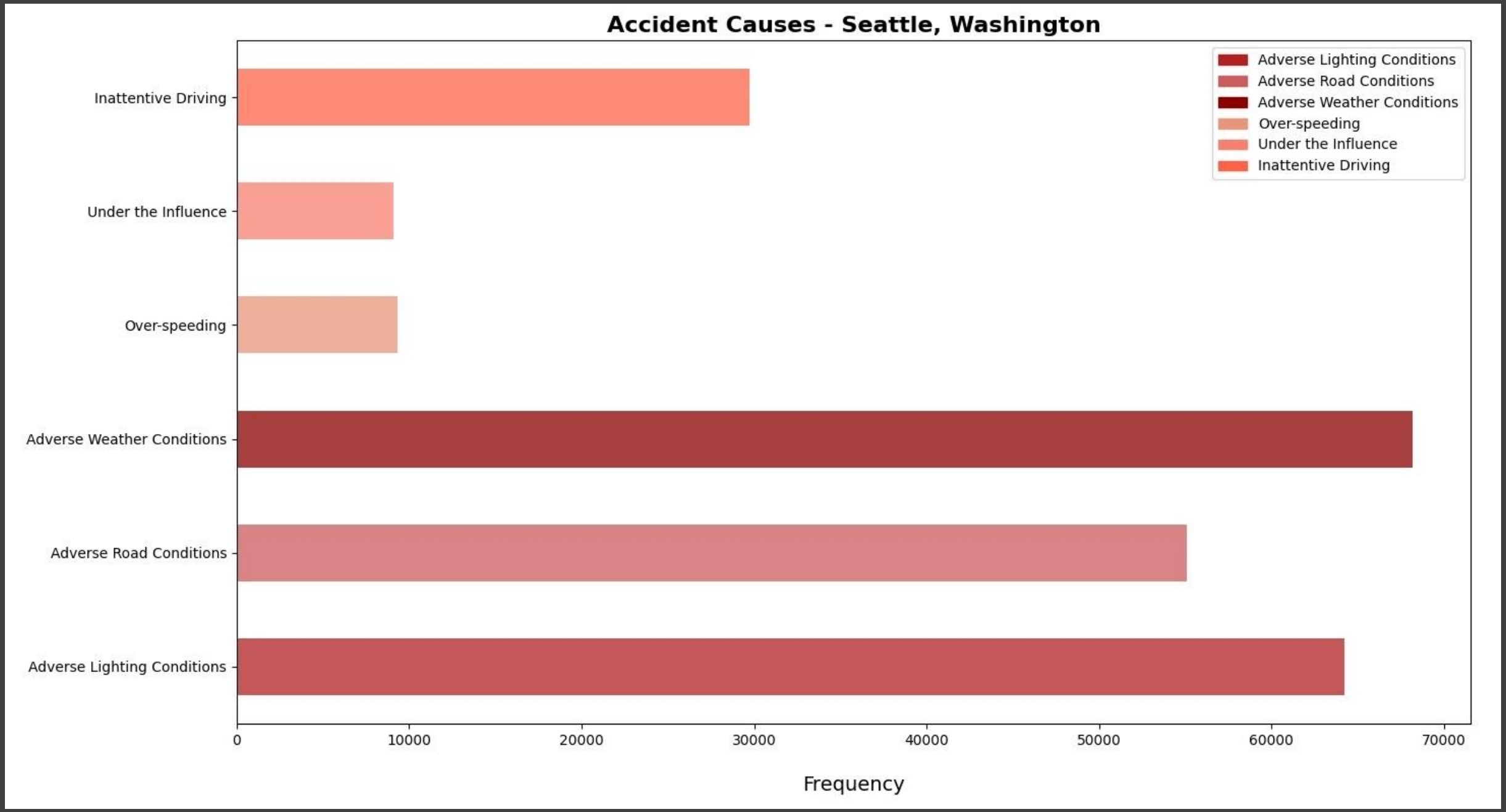
| Feature Variables | Description |
| --- | --- |
| INATTENTIONIND | Whether or not the driver was inattentive (Y/N) |
| UNDERINFL | Whether or not the driver was under the influence (Y/N) |
| WEATHER | Weather condition during time of collision (Overcast/Rain/Clea |
| ROADCOND | Road condition during the collision (Wet/Dry..) |
| LIGHTCOND | Light conditions during the collision (Lights On/Dark with l |
| SPEEDING | Whether the car was above the speed limit at the time of c |

# Methodology

# Exploratory Analysis

- Considering that the feature set and the target variable are categorical variables with the likes of weather, road condition and light condition being an above *level 2* categorical variables whose values are limited and usually based on a particular finite group whose correlation might depict a different image then what it actually is. Generally, considering the effect of these variables in car accidents are important hence these variables were selected. A few pictorial depictions of the dataset were made in order to better understand the data.


Type of Accidents - Seattle, Washington
Property Damage Onl
Physical Injury
Frequency
140000
120000
100000
80000
60000
40000
20000
0
0
1
Property Damage (0) VS Physical Injury (1)


Accident Causes - Seattle, Washington
Adverse Lighting Conditions
Adverse Road Conditions
Adverse Weather Conditions
Over-speeding
Under the Influence
Inattentive Driving
Inattentive Driving
Under the Influence
Over-speeding
Adverse Weather Conditions
Adverse Road Conditions
Adverse Lighting Conditions
0
10000
20000
30000
40000
50000
60000
70000
Frequency

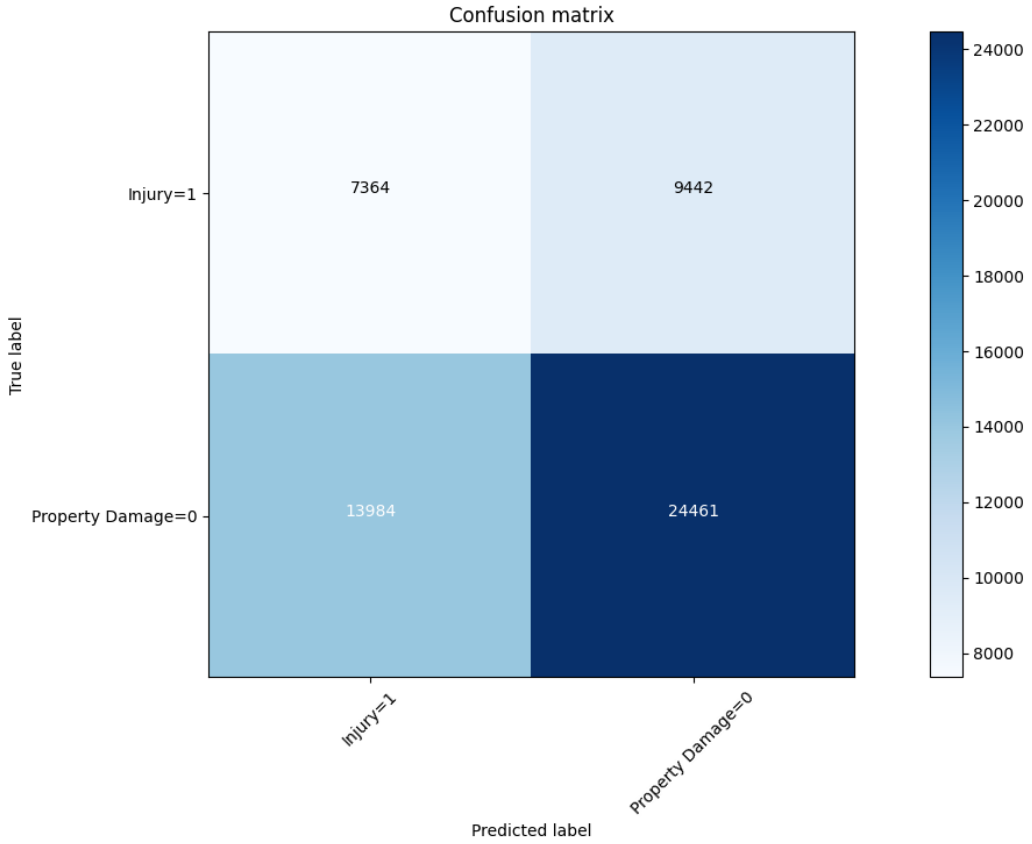
- **Machine Learning Models chosen**
- **Logistic Regression:** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable
- **Decision Tree Analysis:** The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.
- **k-Nearest Neighbor:** K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance)

# RESULTS

# Decision Tree Classification Report

| | Precision | Recall | f1-score |
|---|---|---|---|
| **0** | 0.64 | 0.72 | 0.68 |
| **1** | 0.44 | 0.34 | 0.39 |
| **Accuracy** | 0.58 | | |
| **Macro Avg** | 0.54 | 0.53 | 0.53 |
| **Weighted Avg** | 0.56 | 0.58 | 0.56 |

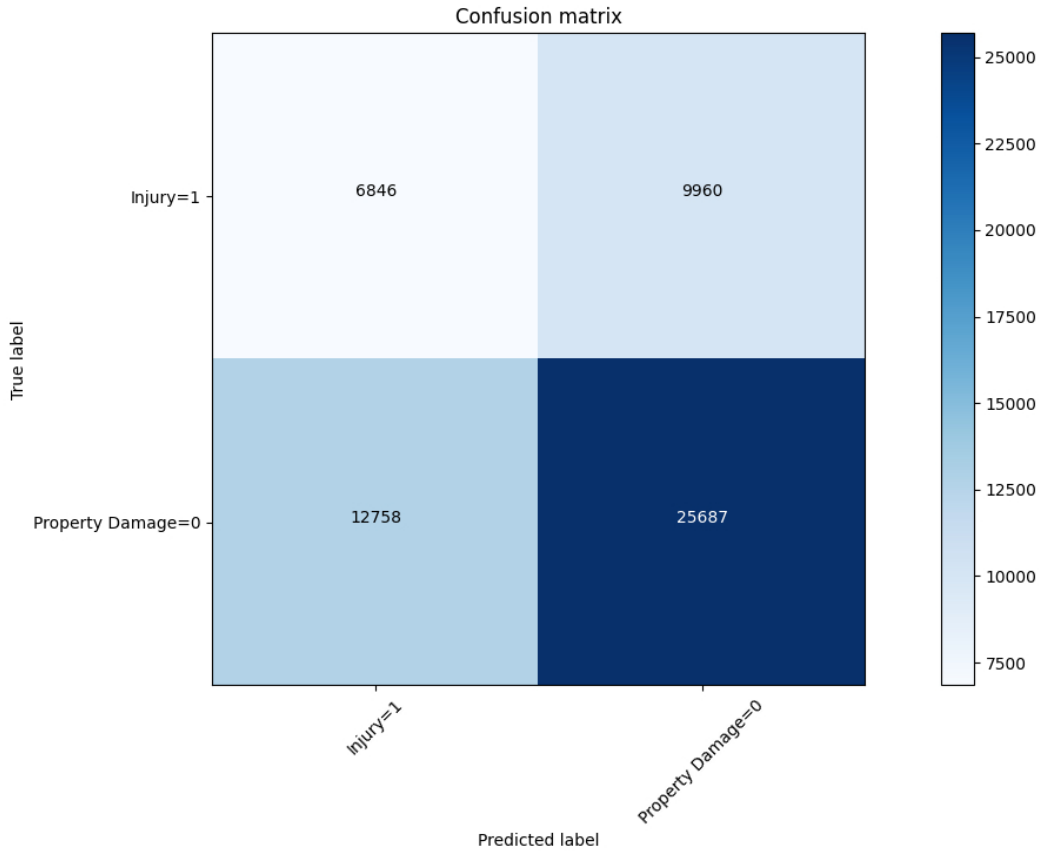
# DT Confusion matrix



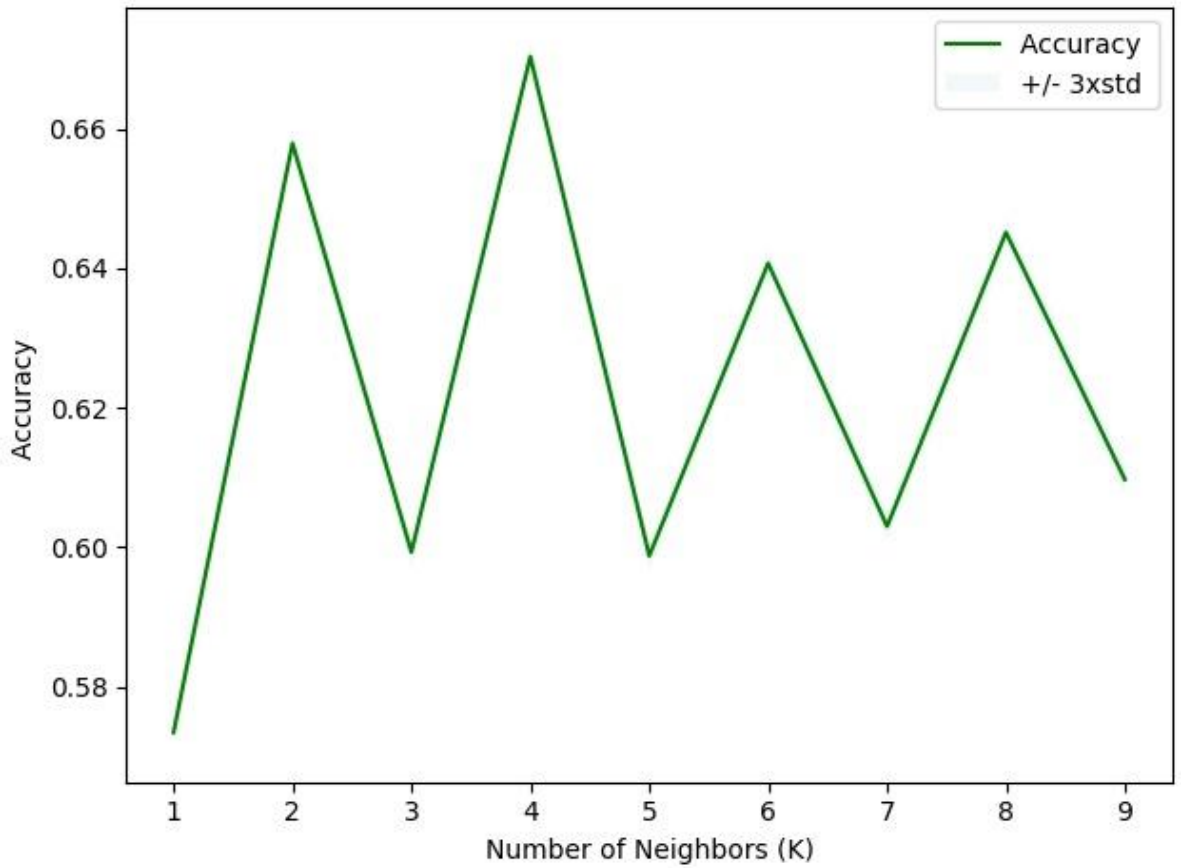
# Logistic Regression

| | Precision | Recall | f1-score |
|---|---|---|---|
| **0** | 0.72 | 0.67 | 0.69 |
| **1** | 0.35 | 0.41 | 0.38 |
| **Accuracy** | 0.59 | | |
| **Macro Avg** | 0.53 | 0.54 | 0.53 |
| **Weighted Avg** | 0.61 | 0.59 | 0.60 |
| **Log Loss** | 0.68 | | |

# LR Confusion Matrix



# KNN K value


Accuracy
+/- 3xstd
0.66
0.64
0.62
0.60
0.58
Accuracy
1
2
3
4
5
6
7
8
9
Number of Neighbors (K)

# k-Nearest Neighbor Classification Report

| | Precision | Recall | f1-score |
|---|---|---|---|
| **0** | 0.93 | 0.70 | 0.80 |
| **1** | 0.08 | 0.32 | 0.13 |
| **Accuracy** | 0.67 | | |
| **Macro Avg** | 0.50 | 0.51 | 0.46 |
| **Weighted Avg** | 0.86 | 0.67 | 0.75 |

# Model Accuracy

| Algorithm | Average f1-Score | Property Damage (0) vs Injury (1) | Precision | Recall |
|---|---|---|---|---|
| Decision Tree | 0.56 | 0 | 0.64 | 0.72 |
| | | 1 | 0.44 | 0.34 |
| Logistic Regression | 0.60 | 0 | 0.72 | 0.67 |
| | | 1 | 0.35 | 0.41 |
| k-Nearest Neighbor | 0.75 | 0 | 0.93 | 0.70 |
| | | 1 | 0.08 | 0.32 |

# Conclusion

- When comparing these scores, we can see that the f1-score is highest for k-Nearest Neighbor at *0.75*. However, later when we compare the precision and recall for each of the model, we can see that the k-Nearest Neighbor model performs poorly in the precision of *1* at *0.08*. The variance is too high for the model to be selected as a viable option. When looking at the other two models, we can see that the Decision Tree has a more balanced precision for *0* and *1*. Whereas, the Logistic Regression is more balanced when it comes to recall of *0* and *1*. Furthermore, the average f1-score of the two models are very close but for the Logistic Regression it is higher by *0.04*. It can be concluded that the both the models can be used side by side for the best performance.