1. INTRODUCTION

1.1 BACKGROUND

The world as a whole suffers a lot from car accidents and lots of people lost their lives.
America has the largest car market in the world. 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 U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, making it the 15th-largest in the United States. As the headquarters of Boeing, Amazon, Microsoft, Seattle attracted widespread attention as home to these many companies and their employees. The city has found itself "bursting at the seams", and with the country's sixth-worst rush hour traffic. According to a 2017 report published by Washington State Department of Transportation, **car accident occurs every 4 minutes** and a **person dies in car crash every 20 hours** in the Washington State alone.

1.2 PROBLEM

It is always rainy and windy in Seattle, and on the way, you always come across a terrible traffic jam on the other side of the highway, with long lines of cars barely moving.

It would be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

Luckily enough, The Seattle Police Department (SPD) has recorded all car collision accident from 2004 to present. Basing on those historical data (194,673 records), we can create a map and information chart to help us understand the high-risk areas, understand car injury factors to avoid accident, and plan our next trip to Seattle better.

1.3 STAKEHOLDERS

The reduction in severity of accidents can be beneficial to **the whole society,** including the Public Development Authority of Seattle which works towards improving those road factors, and the car drivers themselves who may take precaution to reduce the severity of accidents.

2. DATA ACQUISITION AND CLEANING

2.1DATA SOURCES

1. Collisions data from 2004 to present. Those data are provided by the Traffic Records Group in the SDOT Traffic Management Division from Seattle, WA. It includes all collisions (194,673 records) provided by the Seattle Police Department and recorded by the Traffic Record, displayed at the intersection or mid-block of a segment from 2004 to the present. Each record has 38 variables/attributes which contains information such as severity code, address type, location, collision type, weather, road condition, speeding, among others. Of the 194,673 records, only 58,188 records are injury collision, so this is **an unbalanced dataset for our research**.

2. Seattle Map data . This can be found at Github searching for "seattle-boundariesdata". And can also be accessed via a JSON API by using boundaries-api.seattle.io.

2.2 DATA CLEANING

2.2.1 Feature selection

As we want to analyze what factors will probably cause a car collision and the severity of the accident, we would drop those unrelated features and useless information, only keep 15 features/attributes of the original dataset. Those 23 features we dropped/deleted are:
'OBJECTID', 'INCKEY', 'COLDETKEY', ‘REPORTNO', 'STATUS', 'INTKEY', 'LOCATION', ‘EXCEPTRSNCODE', 'EXCEPTRSNDESC','PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', ‘VEHCOUNT', 'SDOT_COLCODE', ‘SDOT_COLDESC', 'PEDROWNOTGRNT', 'ST_COLCODE', ‘ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', ‘HITPARKEDCAR’.

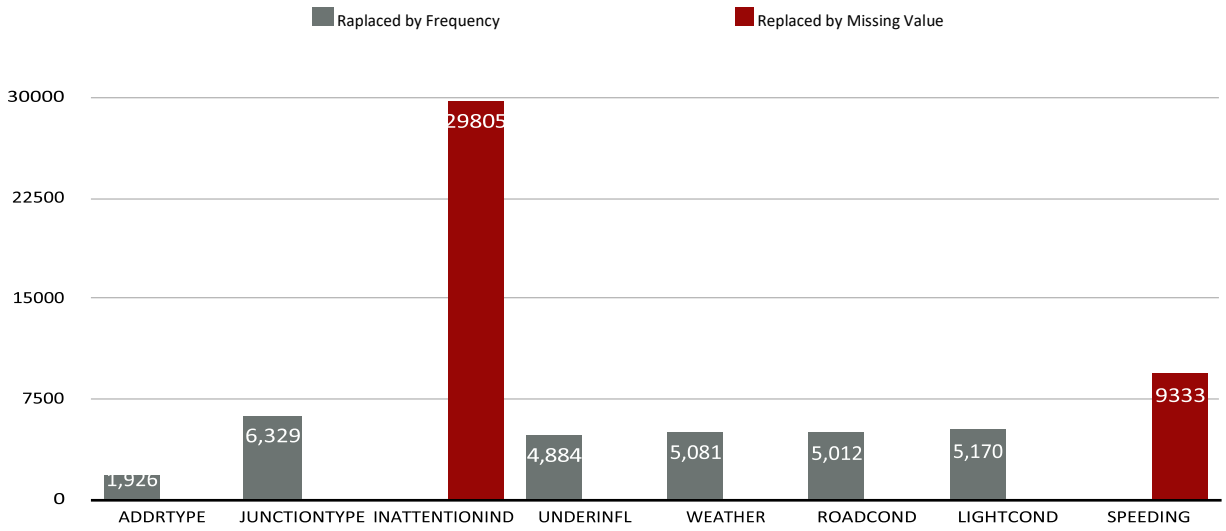
2.2.2 Handle missing values

- Convert "?" to NaN.

We replace "?" with NaN (Not a Number), which is Python's default missing value marker, for reasons of computational speed and convenience.

- Replace missing value by frequency/

Whole columns should be dropped only if most entries in the column are empty. In our dataset, none of the columns are empty enough to drop entirely. Basing on the character of the data features we choose, we mainly replace the missing value with the most frequent values.



However, feature “INATTENTIONIND” and “SPEEDING” only have “Y” value, thus we replace the missing value in those two columns with “Y”.

2.2.3 Correct data format

Two features (“INCDTTM” and “INCDATE”) should NOT be object type, thus we change those two columns with “datetime64” type.

And then use those data to calculate which hour (“hourofday”) and which weekday (“dayofweek”) those accidents happened.

2.2.4 Delete/Drop some rows

We want to use location data to map the accident, so have to drop more than 5000 records who don't have X and Y data.

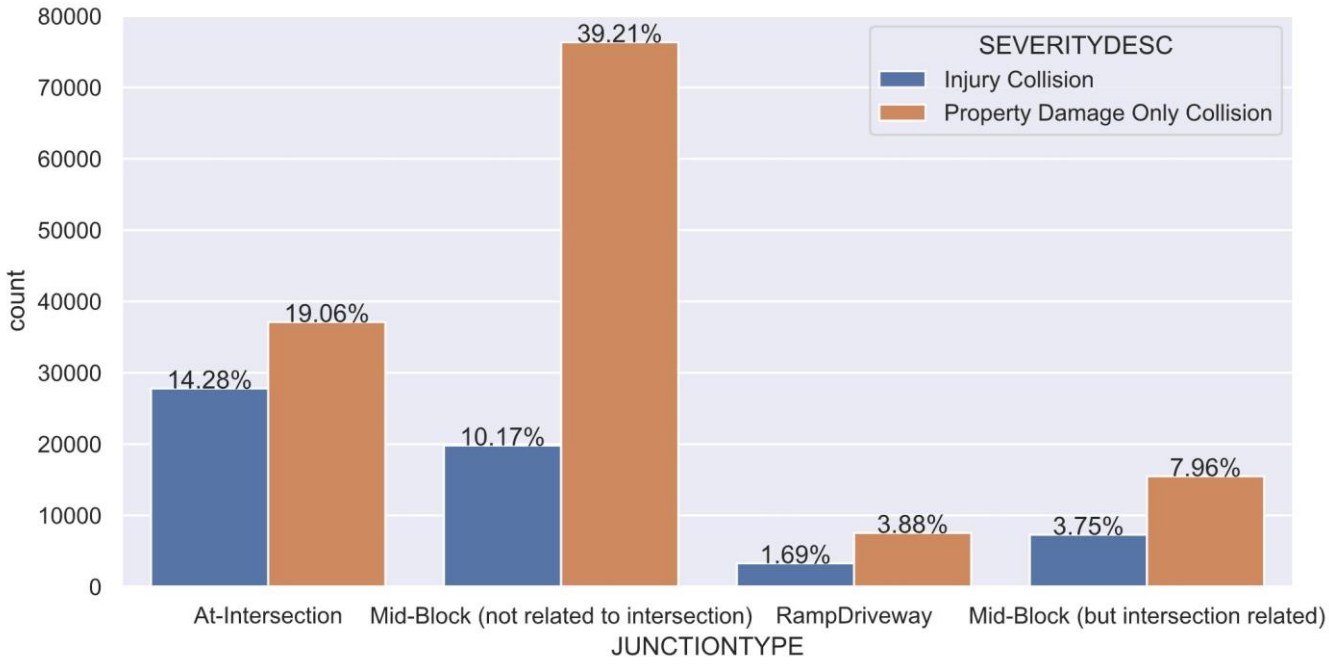
3. DATA ANALYSIS (ANALYZING INDIVIDUAL FEATURE PATTERNS USING VISUALIZATION)

As we want to know what factors will have impact on injury collision compared to property damage, I plot the count and percentage of each type of each feature/attribute. The main goal is to find whether **the relative ratio of each feature type differs**. For example, in the bar plot below, the relative ratio of "At-Intersection" = 19.06% / 14.28% = 1.33, but the relative ratio of "Mid-Block" = 39.21% / 10.17% = 3.86. This means that people will be **more likely to get injury "At-Intersection", but less likely at "Mid-Block"**.

3.1INDIVIDUAL FEATURE PATTERNS

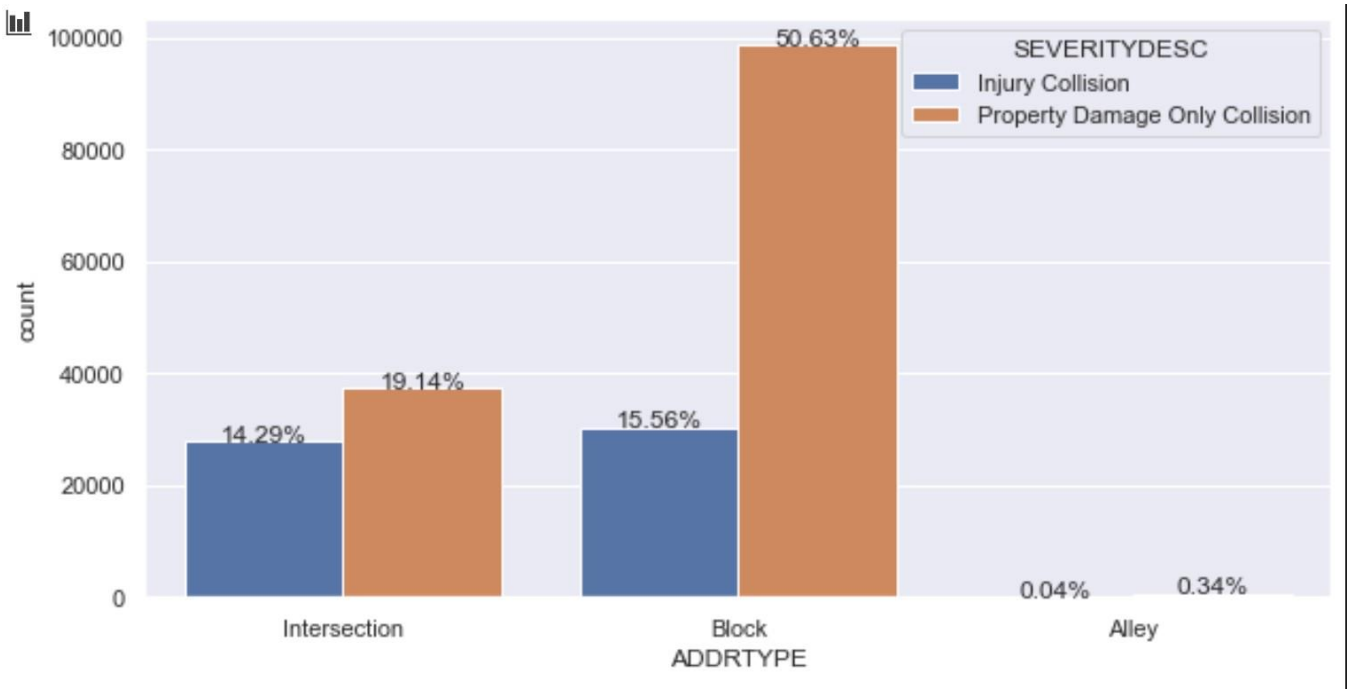
3.1.1 Junction Type have big impact on collision severity

From the plot above, we can see "At Intersection" is an important factor, where the accidents are more likely to involve injury, only a little less than the chances of property damage.



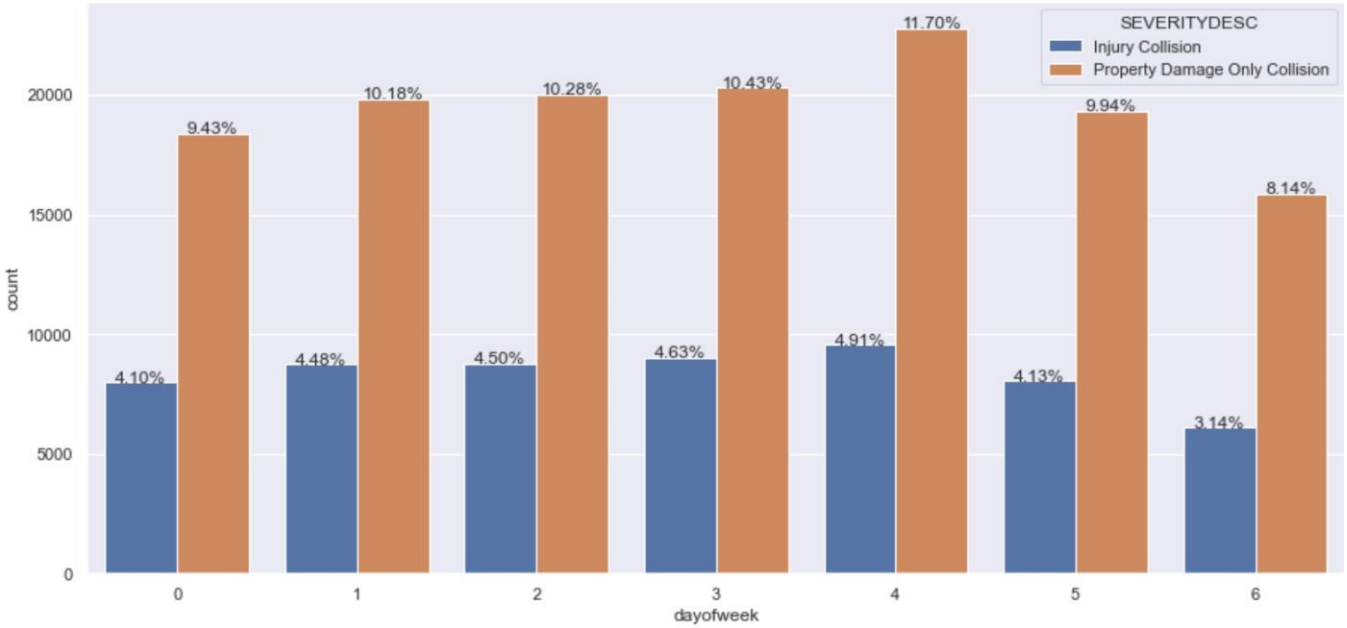
3.1.2 Address Type data also show that Intersection greatly influence the severity

We can see that an accident happened at Intersection will be more likely to have people injured. Car collision occurred at Alley are also different than other two types, but data are not big enough.

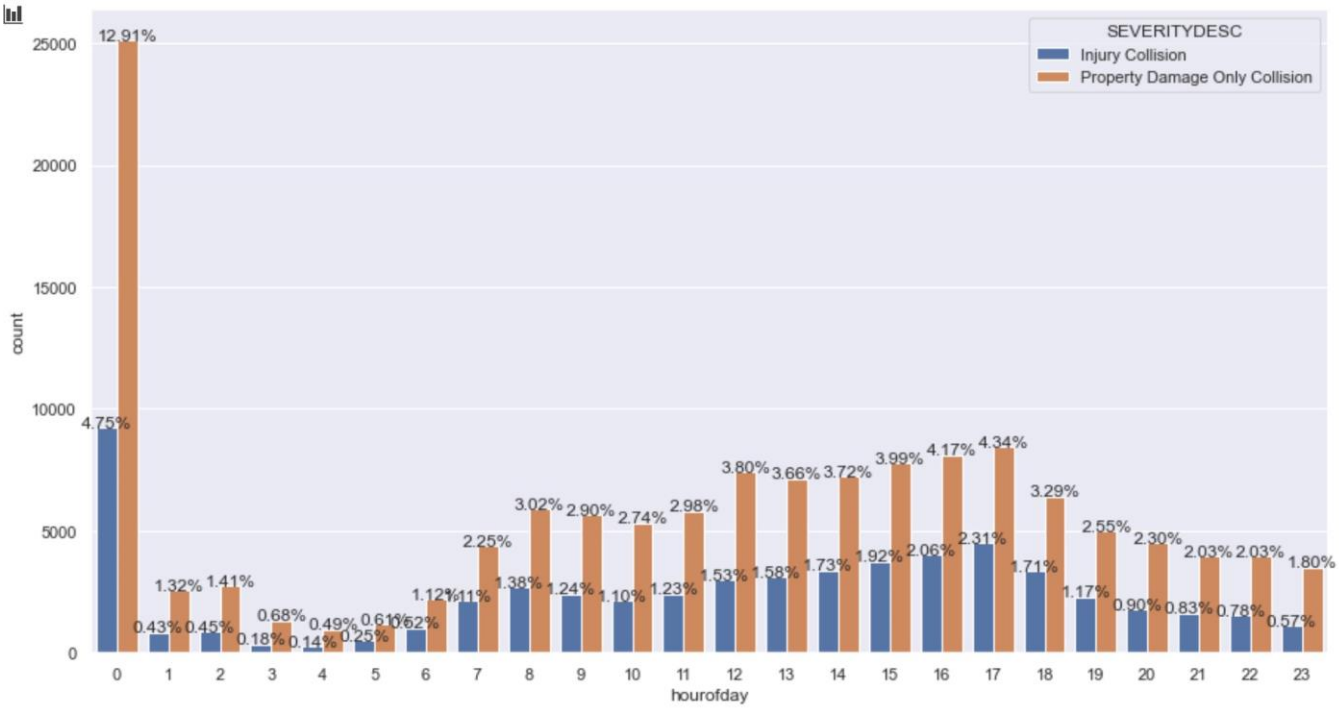


3.1.3 No evidence show that weekday plays a role in influencing severity

Previous I thought weekend will be more likely to have injury collision, but data shows that there is no clear evidence.

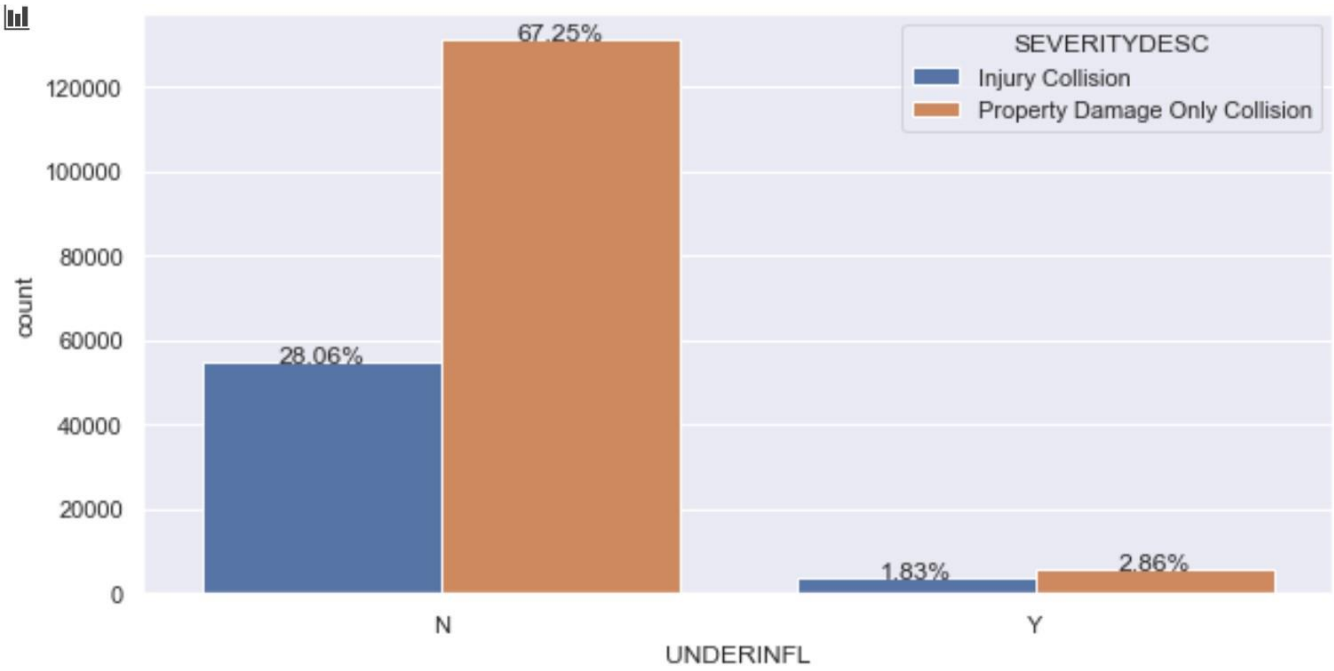


3.1.4 Hour of Day seems plays a role.

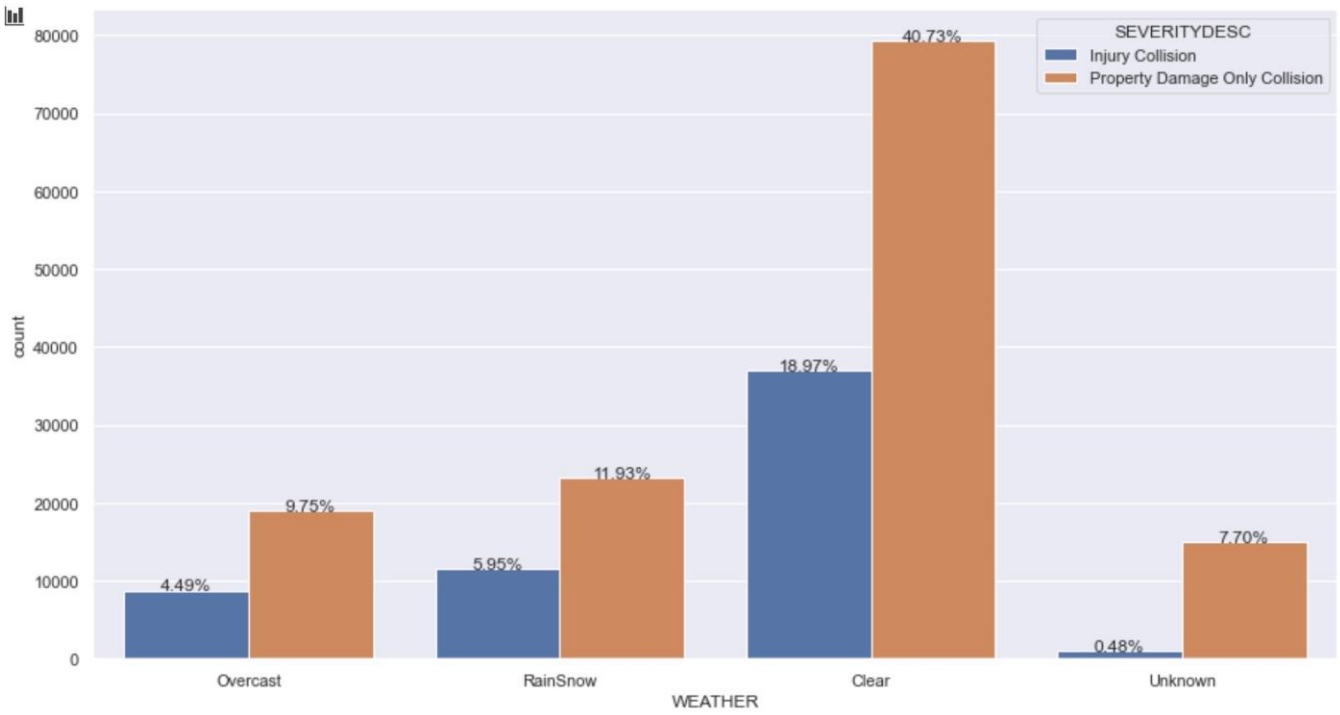


We can see that early morning will be less risky to drive and are less likely to have injury, and there are differences between the trend of day and night.

3.1.5 Drug and alcohol will greatly increase the possibility of injury collision

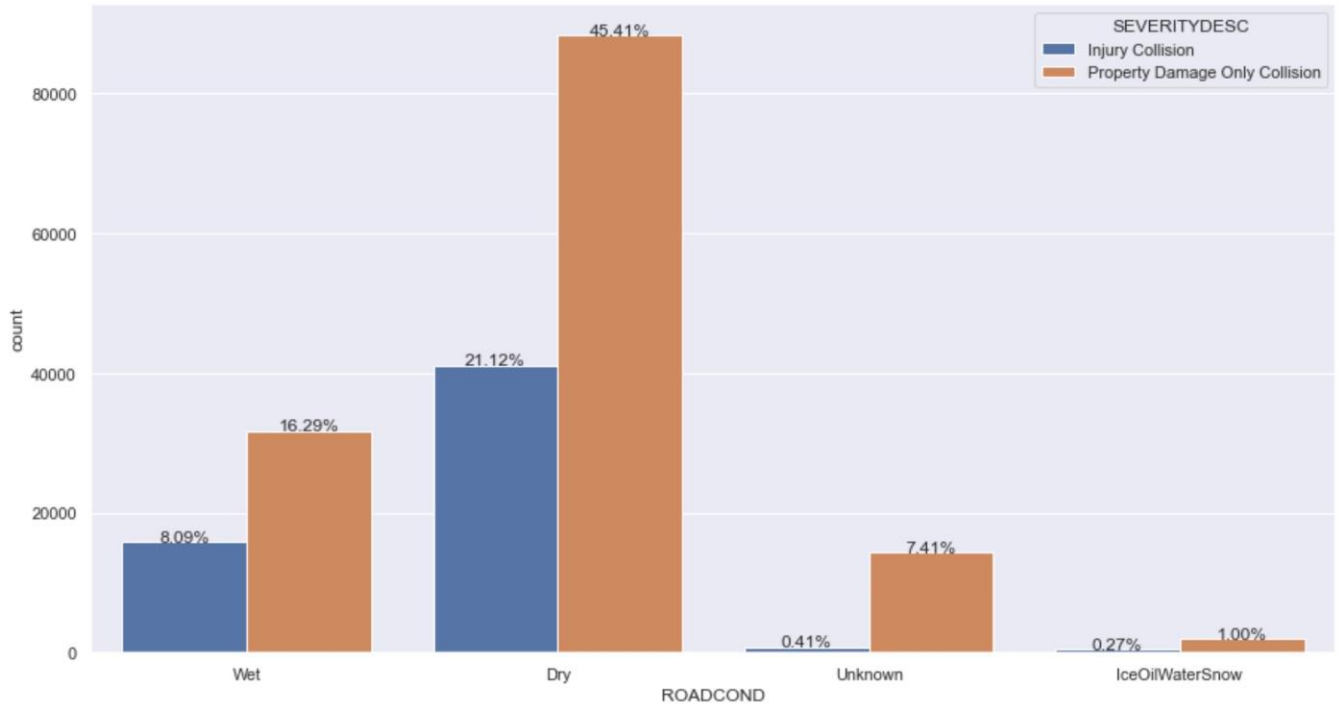


3.1.6 "Unknown" weather condition plays an important role

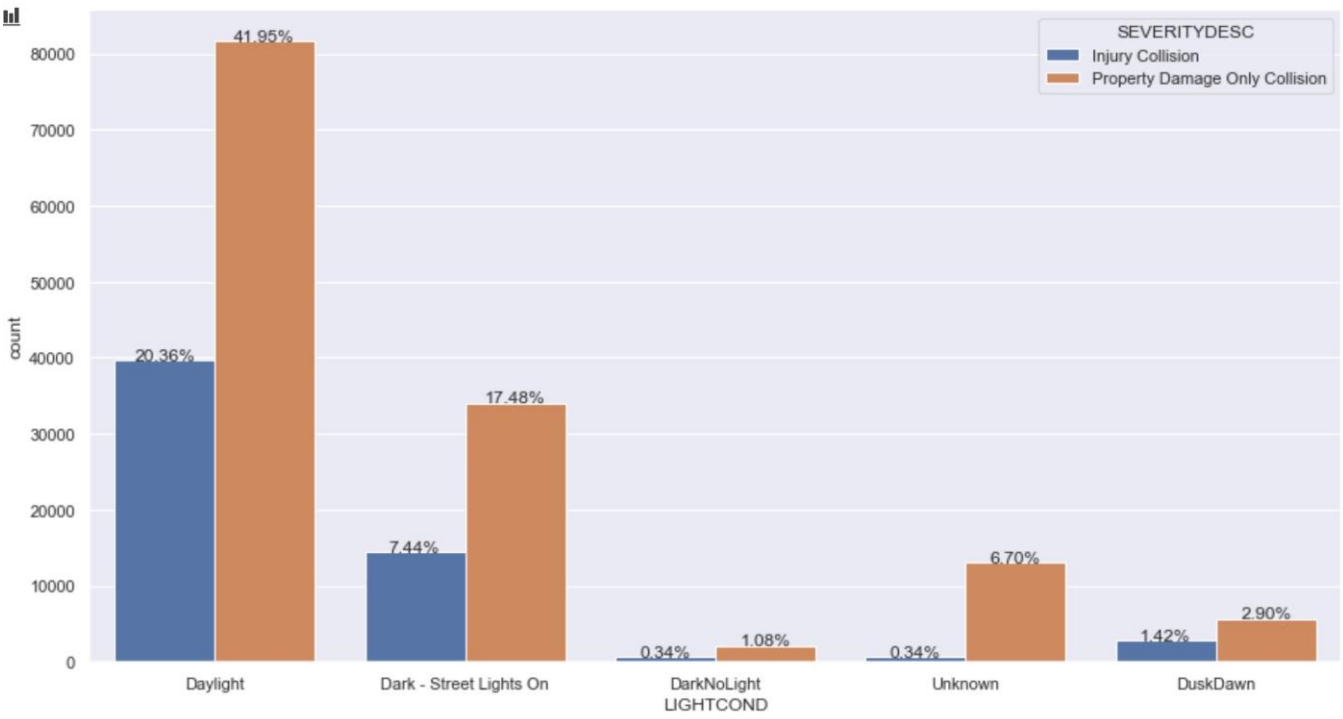


However, we can see the “unknown” weather condition differs hugely with other weather, we should get more about the “unknown” data to dig into and find out the reason.

3.1.7 “Unknown” Road condition has the same pattern with weather



3.1.8 Light condition is an important factor



3.2 FEATURE SELECTION

Basing on the individual feature analysis, we will only choose/keep those which showed significant difference. Thus we **choose 10 features** to begin our analysis, and those features are:

Feature variables	Description	Value
ADDRTYPE	Type of Address	'Intersection', 'Block', 'Alley'
JUNCTIONTYPE	Junction Type	'MidBlock (not related to intersection)','RampDriveway','MidBlock (but intersection related)','At-Intersection'
INATTENTIONIND	Whether or not the driver was inattentive	Y/N
UNDERINFL	Whether or not the driver was under the influence of Drug or Alcohol	Y/N
WEATHER	Weather condition during time of collision	'Overcast', 'RainSnow', 'Clear', 'Unknown'
ROADCOND	Road condition during the collision	'Wet', 'Dry', 'Unknown', 'IceOilWaterSnow'

LIGHTCOND	Light conditions during the collision	'Daylight', 'Dark-With-Light', 'Dark-No-Light', 'Unknown','DuskDawn'
SPEEDING	Whether the car was above the speed limit at the time of collision	Y/N
RiskTime	Whether the timing is risk	Medium'(5-16), 'Low'(1-4), 'High'(17-23)
weekend	Whether it is weekend	Y/N

4. METHODOLOGY

4.1 MACHINE LEARNING MODEL SELECTION

The machine learning models used are Logistic Regression, Decision Tree Analysis and kNearest Neighbor.

- Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.

- 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 neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance).

- Support Vector Machine (SVM) model is inaccurate for large data sets, while this data set has more than 190,000 rows filled with data. Furthermore, SVM works best with dataset filled with text and images.

4.2 UNBALANCED DATA ADJUSTMENT

Imbalanced classification involves developing predictive models on classification datasets that have a severe class imbalance. We can see that only 30% of the records (58,188) are "Injury Collision", which is also the most important type we want to analyze.

One approach to addressing imbalanced datasets is to oversample the minority class. This is a type of data augmentation for the minority class and is referred to as the **Synthetic Minority Oversampling Technique**, or **SMOTE** for short.

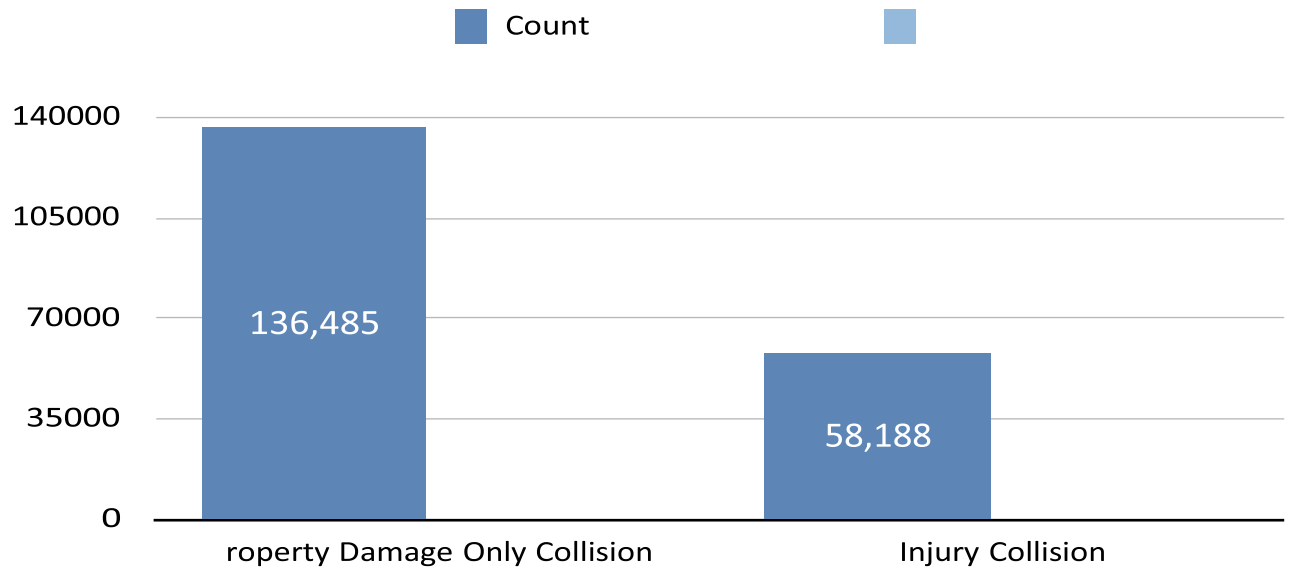
Using SMOTE method, my train dataset will increase from 116,803 (70%:30%) to 163,758 (50%:50%).

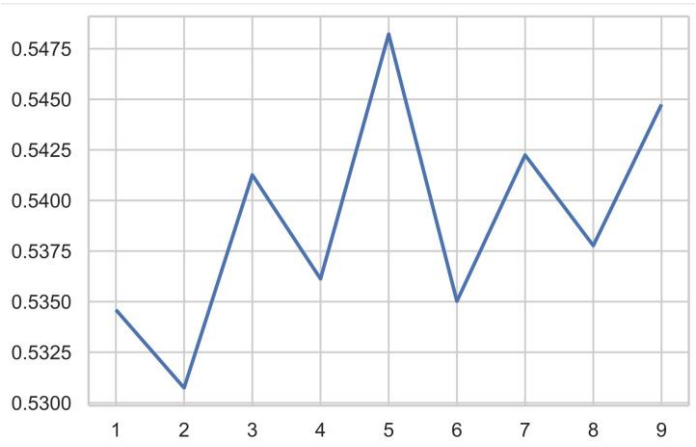
5. RESULT

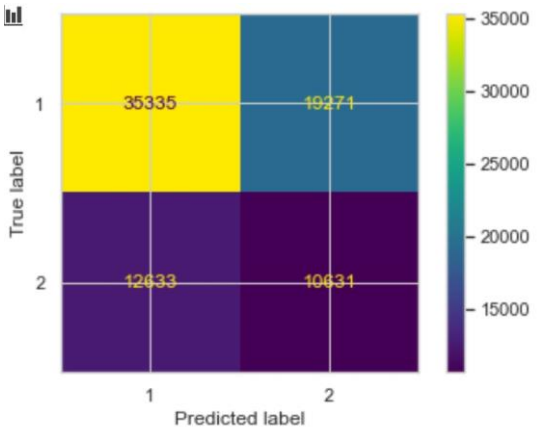
We would want our prediction to be **more precision about "Injury Collision" type**, NOT the "property damage only collision" type. This also means that we don't care too much about the overall accuracy, but want to **focus on the "Recall" ratio,** and can sacrifice some precision ratio to achieve this goal.

5.1K-NEAREST NEIGHBOR

The best K, as shown below, for the model where the highest elbow bend exists is at 5.





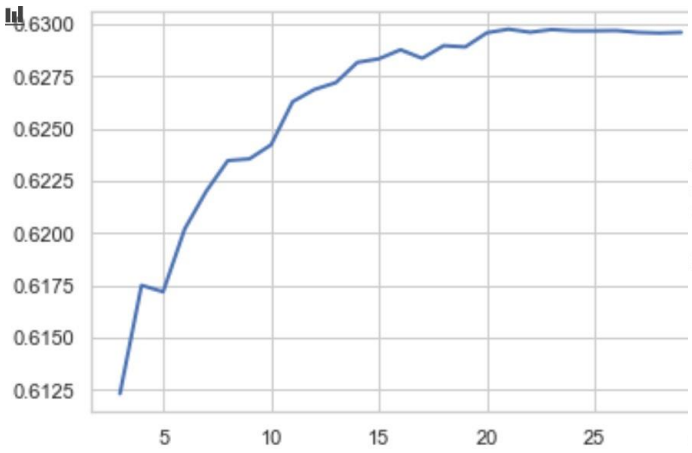


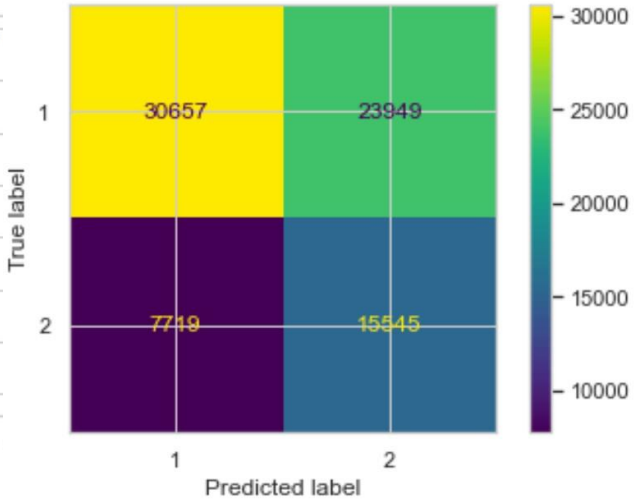
The Confusion Matrix shows that the KNN model have an accuracy of 59%, and the recall rate is 46%.

	precision	recall	f1-score	support
Property collision	0.74	0.65	0.69	54,606
Injury Collision	0.36	**0.46**	0.40	23,264
accuracy			0.59	77,870
macro avg	0.55	0.55	0.54	77,870
weighted avg	0.62	0.59	0.60	77,870

5.2 DECISION TREE ANALYSIS

The criterion chosen for the classifier was 'entropy' and the **max depth was 21** with **best accuracy of 63%.**





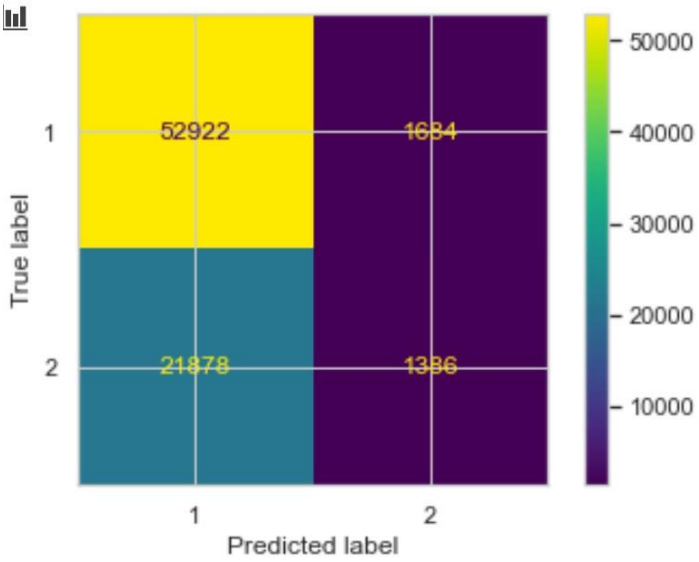
Confusion Matrix shows that the Decision Tree model have an accuracy of 59%, and the recall rate is 67%.

	precision	recall	f1-score	support
Property collision	0.80	0.56	0.66	54606
Injury Collision	0.39	**0.67**	0.50	23264
accuracy			0.59	77870
macro avg	0.60	0.61	0.58	77870
weighted avg	0.68	0.59	**0.61**	77870

5.3 LOGISTIC REGRESSION

We use GridSearchCV to search the best parameters.

The C used for regularization strength was '0.01' and penalty was "l2", whereas the solver used was 'liblinear'.



	precision	recall	f1-score	support
Property collision	0.71	0.97	0.82	54606
Injury Collision	0.45	0.06	0.11	23264
accuracy			0.70	77870
macro avg	0.58	0.51	0.46	77870
weighted avg	0.63	0.70	0.61	77870

Although the Logistic Regression model has an accuracy ratio of 70%, but the recall ratio is only 6%, which is not what we want to predict.

6. DISCUSSION

Alogorithm	Average F-1 Score	Type	Precision	Recall
Decision Tree	0.61	Property collision	0.80	0.56
		Injury Collision	0.39	**0.67**
k-Nearest Neighbor	0.60	Property collision	0.74	0.65
		Injury Collision	0.36	**0.46**
Logistic Regression	0.61	Property collision	0.71	0.97
		Injury Collision	0.45	**0.06**

6.1 AVERAGE F-1 SCORE

f1-score is a measure of accuracy of the model, which is the harmonic mean of the model's precision and recall. Perfect precision and recall is shown by the f1-score as 1, which is the highest value for the f1-score, whereas the lowest possible value is 0 which means that either precision or recall is 0.

Our result shows that the three models are almost the same (0.60-0.61). However, the **average f1-score doesn't depict the true picture of the models accuracy because of the different precision and recall of the model for both the elements of the target variable**. Hence, it is biased more towards the precision and recall of Property Damage due to its weightage in the model.

6.2 PRECISION

Precision refers to the percentage of results which are relevant, in simpler terms it can be seen as how many of the selected items from the model are relevant. Mathematically, it is calculated by dividing true positives by true positive and false positive.

Decision Tree model has the highest precision level (0.8) with the lowest being Logistic Regession Model (0.71). However, the precision level to predict the “Injury Collision” are all at a low level (0.39/0.36/0.45), which is definitely what we want to avoid. We would want an opposite result.

6.3 RECALL

Recall refers to the percentage of total relevant results correctly classified by the algorithm. In simpler terms, it tells how many relevant items were selected. It is calculated by dividing true positives by true positive and false negative.

In our case, we would want our prediction to be more precision about “Injury Collision” type, NOT the “property damage only collision” type. This also means that we don’t care too much about the overall accuracy, but want to focus on the “Recall” ratio, and can sacrifice some precision ratio to achieve this goal.

The three model we use differ a lot in this ratio. Decision Tree model has the best performance of recall ratio of 67%, compared with the other two model’s poor result of 46% and 6%.

Using Decision Tree model, we can predict 67% of the “Injury collision” car accidents.

7. CONCLUSION

In this study, I analyzed the factors which may lead to injury (severity of a car collision). I identified address type/intersection type, weather condition, light condition, whether the driver had drug/alcohol, among the most important features that affect the severity of a car collision. I built three classification models to predict what condition would cause injury collision. These models can be very useful in helping the society in a number of ways. For example:

- drivers can use this model to adjust their behavior to avoid injury;
- insurance company can use this model to adjust auto insurance premium level;
- Traffic management department can use the model to help decrease future injury collision by providing better light and road infrastructure.

8. FUTURE DIRECTIONS

I was able to achieve a 67% recall ratio using the Decision Tree Model. However, the other two model didn't do better and can only predict "Property collision" well. This may be due to unbalanced dataset. If we can have a better dataset, future performance of these two other models might be different.

We can see from individual feature analysis part that there are some "unknown" types of weather, road condition, etc. have a significant difference with other conditions. This means that there maybe some unrecognized factors which need us to dig deeper.

9. RECOMMENDATIONS TO CAR DRIVERS

The map below shows 5% of all the "Injury Collision" in Seattle area since 2004 to present, which can give you an idea of where you should drive more carefully than other places.