Predicting severe occurrences of automobile collisions in Seattle, WA

# Introduction

## Description of Problem

There are over 200 million licensed drivers in America. With that many people on the road, accidents are going to happen. The United States Department of Transportation estimates about 7 million car accidents happen nationwide each year. No matter how safe a driver you are, there’s a good chance you’ll get into at least one accident in your lifetime.

As for the Emerald City, Seattle, it ranks as one of the nation's worst cities to drive in and own a car, according to a newly released study. The report, by personal finance website [WalletHub](https://wallethub.com/edu/best-worst-cities-to-drive-in/13964/), found that Seattle is the 10th worst U.S. city for drivers.

## Background

The purpose of this report is to predict the severity of automobile collisions in Seattle with the application of machine learning models. These models should assist the [Public Development Authority of Seattle](https://library.municode.com/wa/seattle/codes/municipal_code?nodeId=TIT11VETR) focus limited resources towards the variables that have the highest impact on reducing the number of collisions and improve public safety.

# Data Description

## Data Source & Attributes

The data used for this project originated from the [City of Seattle Open Data Portal](https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d) website, which consists of vehicle collision incidents reported in the city of Seattle from 2004 to September 22, 2020 (the date dataset was downloaded). This input dataset is .csv format and consists of 221,526 incidents with 40 variables. Metadata regarding the variables is available from the [portal](https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf) website.

## Pre-Processing

The pre-processing of the dataset is to prepare it for exploratory data analysis (EDA). Although this dataset is downloaded from a single source, it was probably concatenated from multiple sources.

### Variable Redundancy

There are multiple variables that contain the same or similar information, which make them redundant and will be removed from the dataset.

OBJECTID has a unique value for each incident, which corresponds with the dataset index value.

INCKEY, COLDETKEY, and REPORTNO each have unique values for each collision incident.

SEVERITYCODE and SEVERITYDESC are duplicate information. Therefore, SEVERITYCODE will be dropped since SEVERITYDESC has more information.

INCDATE and INCDTTM both provide the date of the incident. Only INCDTTM is required since it includes both the time and date and will be used for time-series analysis.

LOCATION is a catagorical field that can be substituted by latitude and longitude variables (X and Y variables respectively).

SDOT\_COLCODE and SDOT\_COLDESC correspond to the same information. SDOT\_COLCODE is the type of collision, and SDOT\_COLDESC has the description of each type of collision. Therefore, we can drop SDOT\_COLCODE since SDOT\_COLDESC has more information.

### Missing Values

The dataset contains 221,525 incidents (rows) and 40 variables (columns). It contains multiple variables with a significant amount of missing data. For example:

INTKEY (71,936); EXCEPTRSNCODE (101,122); EXCEPTRSNDESC (11,779); INATTENTIONIND (30,188); PEDROWNOTGRNT (5,192); SPEEDING (9,929)

A large number of missing values could cause noise and bias in the results, therefore will need evaluated.

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 X 214050 non-null float64

1 Y 214050 non-null float64

2 OBJECTID 221525 non-null int64

3 INCKEY 221525 non-null int64

4 COLDETKEY 221525 non-null int64

5 REPORTNO 221525 non-null object

6 STATUS 221525 non-null object

7 ADDRTYPE 217813 non-null object

8 INTKEY 71936 non-null float64

9 LOCATION 216935 non-null object

10 EXCEPTRSNCODE 101122 non-null object

11 EXCEPTRSNDESC 11779 non-null object

12 SEVERITYCODE 221524 non-null object

13 SEVERITYDESC 221525 non-null object

14 COLLISIONTYPE 195212 non-null object

15 PERSONCOUNT 221525 non-null int64

16 PEDCOUNT 221525 non-null int64

17 PEDCYLCOUNT 221525 non-null int64

18 VEHCOUNT 221525 non-null int64

19 INJURIES 221525 non-null int64

20 SERIOUSINJURIES 221525 non-null int64

21 FATALITIES 221525 non-null int64

22 INCDATE 221525 non-null object

23 INCDTTM 221525 non-null object

24 JUNCTIONTYPE 209551 non-null object

25 SDOT\_COLCODE 221524 non-null float64

26 SDOT\_COLDESC 221524 non-null object

27 INATTENTIONIND 30188 non-null object

28 UNDERINFL 195232 non-null object

29 WEATHER 195022 non-null object

30 ROADCOND 195103 non-null object

31 LIGHTCOND 194933 non-null object

32 PEDROWNOTGRNT 5195 non-null object

33 SDOTCOLNUM 127205 non-null float64

34 SPEEDING 9929 non-null object

35 ST\_COLCODE 212112 non-null object

36 ST\_COLDESC 195212 non-null object

37 SEGLANEKEY 221525 non-null int64

38 CROSSWALKKEY 221525 non-null int64

39 HITPARKEDCAR 221525 non-null object

Table 1 The number of events located in each variable and its format.

INTKEY has too many missing entries to be used.

EXCEPTRSNCODE and EXCEPTRSNDESC will be dropped since they have too many missing values to be used in a model.

INATTENTIONIND has a significant amount of missing data with only 30,188 values, which all have a value of 'Y' and pertains to the driver not paying attention while driving and should be dropped.

PEDROWNOTGRNT also has a significant amount of missing data with only 5,195 values, which all have a value of 'Y' and pertain to pedestrian right of way was not granted.

SPEEDING has a significant amount of missing data with only 9929 values, which all have a value of 'Y' and pertian to whether or not speeding was a factor in the collision.

### Variable Reformat and Category Merging

To conduct time series analysis, INCDTTM needs to be converted from a string format to the pandas data-time format.

### Variables with Skewed Distribution

*SEGLANEKEY contains 2,101 unique values with the value '0' dominating with 218,489 of the observations. The distribution of unique values is highly skewed indicating that SEGLANEKEY should be dropped.*

CROSSWALKKEY contains 2,343 unique values with the value '0' dominating with 217,283 of the observations. The distribution of unique values is highly skewed indicating that CROSSWALKKEY should be dropped.

### Missing Values Matix and Heat Map

Using the missingno library to plot and identify where the missing values are located in each column and correlations between missing values across different columns (white lines in the figure 1). Note the high correlation with variables COLLISIONTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, and ST\_COLDESC.

A picture containing table

Description automatically generated

**Figure 1.** Missing Value Matrix. Visualization of missing values across all variables in the dataset.

The heat map confirms the correlation of the missing data for the same rows along with an approximation of its correlation. For example, variable JUNCTIONTYPE has about a 30% correlation with variables COLLISIONTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, and ST\_COLDESC.

Chart, histogram

Description automatically generated

**Figure 2.** Missing Value Heat Map. Visualization of missing values across all variables in the dataset.

Another strong positive correlation is ADDRTYPE with the X and Y variables (Latitude and Longitude respectively) for about 70% of the rows.

After identifying which variables have missing values it needs to be determined if the variable should be dropped completely, remove the incidents that are empty, or to save for possible upscaling.

For X and Y, approximately 3 percent of the dataset had no values so they were dropped. This also correlated directly with the missing data in ADDRTYPE, such that there are no longer any missing values.

### Variable Reformat and Category Merging

UNDERINFL is an example of a variable that originated from multiple sources. This is a binary variable that has two different binary designations. First is 'N' or 'Y'. The second is '0' or '1'. The '0' corresponds to 'N' and '1' is 'Y'. Therefore, replace all '0' values with 'N' and '1' with 'Y'.

INCDTTM (date and time). There will be many factors to analyze with regards to the date and time of events. Therefore, the INCDTTM timestamp were parsed into new fields year, month, day, hour, minute and weekday.

Merging categories within some variables were necessary to simplify the information for the models. For example, WEATHER has 10 catagories (including ‘Other’ and ‘Unknown’). Combining similar weather conditions the number of categories was reduced to six.

Clear 112508

Unknown 38212

Raining 32898

Overcast 27962

Snowing 906

Other 801

Fog/Smog/Smoke 561

Sleet/Hail/Freezing Rain 115

Blowing Sand/Dirt 50

Severe Crosswind 25

Partly Cloudy 10

Blowing Snow 1

Figure 3. WEATHER before combining categories.

Clear or Partly Cloudy 112518

Unknown 39013

Raining 32898

Overcast 27962

Snowing 906

Severe Conditions 75

Figure 3. WEATHER after combining categories.

This was also performed on variables ROADCOND, LIGHTCOND, and SEVERITYDESC

## Exploratory Data Analysis

At this stage many variables have been identified as not required for the model, missing values resolved, reformatted, and some had their categories simplified. Now, with the SEVERITYDESC as the dependent variable, it is compared to each independent variable to identify relationships that would be deemed beneficial for the models.

The final step is to convert categorical variables to discrete integers for the model algorithms.