Data description

To evaluate the objectives of this project, a dataset by SDOT Traffic Management Divisions, Traffic Records Group in Seattle was downloaded using this link. This data is updated weekly and captures all types of collisions since 2004 to present. A record of severity of the collision, "SEVERITYCODE" is provided alongside other 39 variables related to a given collision. The dataset has a record of 221266 collisions.

Below is a snapshot of the first 5 rows of the dataset and the number of rows and columns at the end.

```
In [2]: df_collisions.head()
```

Out[2]:

	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING
0	-122.386772	47.564720	1	326234	327734	E984735	Matched	Intersection	31893.0	CALIFORNIA AVE SW AND SW GENESEE ST	...	Dry	Daylight	Y	NaN	NaN
1	-122.341806	47.686934	2	326246	327746	E985430	Matched	Intersection	24228.0	STONE AVE N AND N 80TH ST	...	Wet	Dark - Street Lights On	NaN	NaN	NaN
2	-122.374899	47.668666	3	329254	330754	EA16720	Matched	Block	NaN	NW MARKET ST BETWEEN 14TH AVE NW AND 15TH AVE NW	...	Dry	Daylight	NaN	NaN	NaN
3	-122.300758	47.683047	4	21200	21200	1227970	Matched	Intersection	24661.0	25TH AVE NE AND NE 75TH ST	...	Wet	Dark - Street Lights On	NaN	4160038.0	NaN
4	-122.313053	47.567241	5	17000	17000	1793348	Unmatched	Block	NaN	S DAKOTA ST BETWEEN 15TH AVE S AND 16TH AVE S	...	NaN	NaN	NaN	4289025.0	NaN

5 rows × 40 columns

```
In [3]: df_collisions.shape
Out[3]: (221266, 40)
```

Figure 1: First 5 rows and shape of data

See below also a list of the various names of columns provided.

```
In [10]: print(list(df_collisions))
['X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCOD
E', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES', 'INCDATE',
'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOT
COLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR']
```

Figure 2: List of names of columns

Columns that are not providing much information on the model such as ID, KEY etc will be dropped. For the purpose of this project, we will explore the various variables and select features that we would use in the prediction based on their importance in predicting severity of an accident. Through feature engineering more variables will be obtained to ensure good prediction of our model.

The target variable is '**SEVERITYCODE',** which captures the various categories of severity of the accident as follows; 0 – Unknown, 1 – prop damage, 2 – injury, 2b – severe injury and 3 – fatality. See below the description of the number and percentage of collisions in each category where we observe significance imbalance in the data. For instance the number of

collisions with code 1 (prop damage) were the highest (137485), contributing to 62% of all the collisions.

```
In [11]: #number of collisions in each severity category
         df_collisions.SEVERITYCODE.value_counts()

Out[11]: 1     137485
         2      58698
         0      21635
         2b      3098
         3        349
         Name: SEVERITYCODE, dtype: int64

In [8]: #percentage of collisions in each severity category
        df_collisions.SEVERITYCODE.value_counts() / len(df_collisions)

Out[8]: 1     0.621356
        2     0.265283
        0     0.097778
        2b    0.014001
        3     0.001577
        Name: SEVERITYCODE, dtype: float64
```

Figure 3: Categories of target variable

Below is also the number of missing values for each column which will guide us in the process of selection of variables. We observe that in our target variable "**SEVERITYCODE**" we have one missing value. We would drop this row going forward. For the other columns a decision will be made on whether to impute or drop the missing values.

```
In [9]: df_collisions.isnull().sum()

Out[9]: X                   7469
        Y                   7469
        OBJECTID               0
        INCKEY                 0
        COLDETKEY              0
        REPORTNO               0
        STATUS                 0
        ADDRTYPE            3712
        INTKEY            149443
        LOCATION            4586
        EXCEPTRSNCODE     120403
        EXCEPTRSNDESC     209491
        SEVERITYCODE           1
        SEVERITYDESC           0
        COLLISIONTYPE      26499
        PERSONCOUNT            0
        PEDCOUNT               0
        PEDCYLCOUNT            0
        VEHCOUNT               0
        INJURIES               0
        SERIOUSINJURIES        0
        FATALITIES             0
        INCDATE                0
        INCDTTM                0
        JUNCTIONTYPE       11967
        SDOT_COLCODE           1
        SDOT_COLDESC           1
        INATTENTIONIND    191078
        UNDERINFL          26479
        WEATHER            26688
        ROADCOND           26608
        LIGHTCOND          26776
        PEDROWNOTGRNT     216078
        SDOTCOLNUM         94061
        SPEEDING          211353
        ST_COLCODE          9413
        ST_COLDESC         26499
        SEGLANEKEY             0
        CROSSWALKKEY           0
        HITPARKEDCAR           0
```

Figure 4: Missing values in each column

After the data has been balanced, exploratory data analysis will be carried out to identify relationships between the various characteristics in the dataset with the target variable, severity of accident. Missing data will also be investigated and corrected. A set of predictions will also be identified to use in the prediction model. Since our data is labeled, supervised machine learning techniques will be used. The data will be split into training and

test sets for training and validation. Since our target variable is likely to be binary after balancing the data, classification models will be evaluated to select the best classifier. The following classification models will be evaluated; K – nearest neighbor, Naïve Bayes, Random Forest, Logistic Regression, Gradient Boosting, XGBoost and Support Vector Machines. Various measures of accuracy such as Jaccard index, F1 score and log loss will be used to select the best classification model.