Applied Data Science Capstone: Coursera

Abhineet Sharma

This notebook is meant for applied data science capstone project at Coursera.

The data set is available at the path: `data-set/Data-Collisions.csv`.

```
!pip3 install folium
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import folium
from folium.plugins import HeatMap
print('Hello Capstone Project Course!')

%matplotlib inline
```

```
Requirement already satisfied: folium in /usr/local/lib/python3.7/site-packages (0.11.0)
Requirement already satisfied: requests in /usr/local/lib/python3.7/site-packages (from folium) (2.22.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/site-packages (from folium) (1.17.4)
Requirement already satisfied: jinja2>=2.9 in /usr/local/lib/python3.7/site-packages (from folium) (2.11.1)
Requirement already satisfied: branca>=0.3.0 in /usr/local/lib/python3.7/site-packages (from folium) (0.4.1)
Requirement already satisfied: idna<2.9,>=2.5 in /usr/local/lib/python3.7/site-packages (from requests->foliu
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/site-packages (from requests->f
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/site-packa
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /usr/local/lib/python3.7/site-packages (from requests
Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.7/site-packages (from jinja2>=2.9->
Hello Capstone Project Course!
```

Business Understanding

We are interested in finding out on a given day and a location in Seattle what is the possibility of an accident happening and how severe that could be. For example, if we are planning on visiting a certain place and we know the road, weather and light conditions, is it possible to know what is the possibility of having an accident so that we could be better prepared and take steps to avoid that. The stakeholders for this project includes the motor vehicle department and the residents. After successful completion of this project, DMV could be able to issue advisories to the commuters commuting to certain areas.

This translates to a multi-class classification problem where we intend to classify given data into different classes representing severity of the accident. A likelihood measure assigned to each classification would provide additional metric to advise commuters accordingly. We will analyze Seattle DMV's accident data and train various classification models and select the best fitting model. The model can then be used to predict possibility of an accident on a given day and location based on certain parameters. Let us explore the available data in the next section to build a better understanding of data and chalk out the necessary steps to clean and prepare data for training models.

Data Understanding

The data set is available in CSV format at `data-set/Data-Collisions.csv`. Let us first load the data and check out the basic attributes like shape, types of parameters, missing values etc.

```
df = pd.read_csv('data-set/Data-Collisions.csv')
print(f'Data Shape: {df.shape[0]} rows x {df.shape[1]} columns')
df.head()
```

```
Data Shape: 194673 rows x 38 columns


/usr/local/lib/python3.7/site-packages/IPython/core/interactiveshell.py:3063: DtypeWarning: Columns (33) have
  interactivity=interactivity, compiler=compiler, result=result)
```

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTN
0	2	-122.323148	47.703140	1	1307	1307	3502005
1	1	-122.347294	47.647172	2	52200	52200	2607959
2	1	-122.334540	47.607871	3	26700	26700	1482393
3	1	-122.334803	47.604803	4	1144	1144	3503937
4	2	-122.306426	47.545739	5	17700	17700	1807429

5 rows × 38 columns

We have 194673 rows with 38 parameters. From the sample data, it is evident that `SEVERITYCODE` is our target variable. Let us look at all the variables in the dataset and identify the ones relevant to the problem that could be explored further.

```
print(df.dtypes)
```

```
SEVERITYCODE        int64
X                 float64
Y                 float64
OBJECTID            int64
INCKEY              int64
COLDETKEY           int64
REPORTNO           object
STATUS             object
ADDRTYPE           object
INTKEY            float64
LOCATION           object
EXCEPTRSNCODE      object
EXCEPTRSNDESC      object
SEVERITYCODE.1      int64
SEVERITYDESC       object
COLLISIONTYPE      object
PERSONCOUNT         int64
PEDCOUNT            int64
PEDCYLCOUNT         int64
VEHCOUNT            int64
INCDATE            object
INCDTTM            object
JUNCTIONTYPE       object
SDOT_COLCODE        int64
SDOT_COLDESC       object
INATTENTIONIND     object
UNDERINFL          object
WEATHER            object
```

```
ROADCOND           object
LIGHTCOND          object
PEDROWNOTGRNT      object
SDOTCOLNUM        float64
SPEEDING           object
ST_COLCODE         object
ST_COLDESC         object
SEGLANEKEY          int64
CROSSWALKKEY        int64
HITPARKEDCAR       object
dtype: object
```

Refering to the data set's metadata, we identify the following columns useful for our analysis.

- `SEVERITYCODE` : Target variable(Categorical)

  - `0` : Unknown
  - `1` : Property Damage
  - `2` : Injury
  - `2b` : Serious Injury
  - `3` : Fatality

- `X` & `Y` : Refer to the location of the incident.

- `OBJECTID` : ESRI Unique identifier. Can be dropped.

- `INCKEY` , `COLDETKEY` & `REPORTNO` are ids not relevant to our analysis and can be dropped.

- `STATUS` is a categorical variable with two values:

  - `Matched`
  - `UnMatched`
    It seems like it is reported post accident and might not be very useful for our case. So this can be dropped as well.

- `ADDRTYPE` : Collision address type(Categorical)

  - `ALLEY`
  - `BLOCK`
  - `INTERSECTION`

- `INTKEY` : Key corresponding to the intersectio associated with the collision.

- `LOCATION` : Actual location of the accident. This information is encoded into the `X` and `Y` columns. Thus, can be dropped.

- `EXCEPTRSNCODE` & `EXCEPTRSNDESC` : There is no information avaiable in metadata about these columns annd they seem to be sparsely populated in the data set. They can be dropped.

- `SEVERITYCODE.1` & `SEVERITYDESC` can also be dropped. As we already have `SEVERITYCODE` column. Description also won't help us much in our predictio. Beore dropping `SEVERITYCODE.1` it might be worthwhile to look if it is actually duplicate or if there are some discrepancies in `SEVERITYCODE` and `SEVERITYCODE.1` column.

- `COLLISIONTYPE` : Denotes the type of collisio. It is reported post accident, so can be dropped from our analysis.

- `PERSONCOUNT` , `PEDCOUNT` , `PEDCYLCOUNT` , `VEHCOUNT` are the counts of persons, pedestrians, bicycles and vehicles involved in the accident. This is also reported post accident but these columns contain the information of the types of vehicles involved and might be helpful in making predictions. For example, at a particular location there are more bike accidents than involving vehicles. May be you would feel safe going in a vehicle instead on a bicycle,.

- `INCDATE` & `INCTM` : Incident date and time. Although this data directly won't help up in making prediciton. But it contains informationn about the season and the time of the day an accident occured. There is a possibility that there were more accidents reported in winters at a particular location. We may want to infer that information from these columns to see if it helps in predictions.

- `JUNCTIONTYPE` : Category of junction where accident took place. This information might not be very handy in making predictions and might be dropped.

- `SDOT_COLCODE` & `SDOT_COLDESC` are the codes and descriptions, respectively, assigned to the collision. This is reported post collision and might as well be dropped.

- `INATTENTIONIND` : Whether or not the collision is due to inattention. Categorical variable with values `Y` and `N` .

- `UNDERINFL` : Whether or not the accident happened under influence of drugs or alcohol. This attribute is also reported post accident. There is no way of determining it before hand for our problem. Therefore, this can be dropped.

- `WEATHER` : Weather conditions during accident

- `ROADCOND` : Road conditions during accident

- `LIGHTCOND` : Light conditions during accident

- `PEDROWNOTGRNT` : Whether or not pedestrian right of way was granted or not. This is a post accident variable not useful for our prediction and can be dropped.

- `SDOTCOLNUM` : A number given to colision by SDOT. Can be dropped.

- `SPEEDING` : Whether or not speeding was factor in collision. This is also a post accident variable and not useful for our prediction. It can be dropped.

- `ST_COLCODE` & `ST_COLDESC` : Code and description from state defined codes. Can be dropped.

- `SEGLANEKEY` : A key for the lane in which collision occurred. Not very useful to us. Can be dropped.

- `CROSSWALKKEY` : A key for the crosswalk where collision occurred. Not very useful to us. Can be dropped.

- `HTPARKEDCAR` : Whether or not a parked car was hit.

Thus, from the metadata, following are the candidates to be dropped from our analysis as not being useful

- `INCKEY` , `COLDETKEY` & `REPORTNO`
- `OBJECTID`
- `STATUS`
- `EXCEPTRSNCODE` & `EXCEPTRSNDESC`
- `SEVERITYCODE.1` & `SEVERITYDESC`
- `COLLISIONTYPE`
- `JUNCTIONTYPE`
- `SDOTCOLCODE` , `SDOTCOLNUM` & `SDOTCOLDESC`
- `ST_COLCODE` & `ST_COLDESC`
- `INTKEY`
- `LOCATION`
- `UNDERINFL`
- `PEDROWNOTGRNT`
- `SEGLANEKEY`
- `CROSSWALKKEY`
- `SPEEDING`

Let us firt drop the above identified columns from the data frame. Before doing that, let us also check if `SEVERITYCODE` and `SEVERITYCODE.1` are actually duplicates.

```
total_unmatching_values = (df["SEVERITYCODE"] != df["SEVERITYCODE.1"]).sum()
print(f'Total unmatching values: {total_unmatching_values}')
```

```
Total unmatching values: 0
```

Since there are no unmatching values in `SEVERITYCODE` and `SEVERITYCODE.1` columns. They can be considered duplicate and one of them can be dropped.

```
df.drop(columns=["INCKEY", "COLDETKEY", "REPORTNO", "OBJECTID", "STATUS", "EXCEPTRSNCODE", "EXCEPTRSNDESC",
                "SEVERITYCODE.1", "SEVERITYDESC", "COLLISIONTYPE", "JUNCTIONTYPE", "SDOT_COLCODE",
                "SDOTCOLNUM", "SDOT_COLDESC", "ST_COLCODE", "ST_COLDESC", "INTKEY", "LOCATION", "UNDERINFL",
                "PEDROWNOTGRNT", "CROSSWALKKEY", "SEGLANEKEY", "SPEEDING"], inplace=True)
total_records = df.shape[0]
independent_variables = df.shape[1] - 1 # subtract one for the target variable.
print(f'Total of {total_records} records with {independent_variables} independent variables')
```

```
Total of 194673 records with 14 independent variables
```

This leaves us with 22 columns. Let us now analyze each column to understand the distribution better and see how we can use the information in our prediction.

Severity Code

```
df["SEVERITYCODE"].value_counts(normalize=True, dropna=False)
```

```
1    0.701099
2    0.298901
Name: SEVERITYCODE, dtype: float64
```

Around 70.1% of the accidents are of severity `1` in the data and the rest 29.9% of the accidents are of severity `2`.

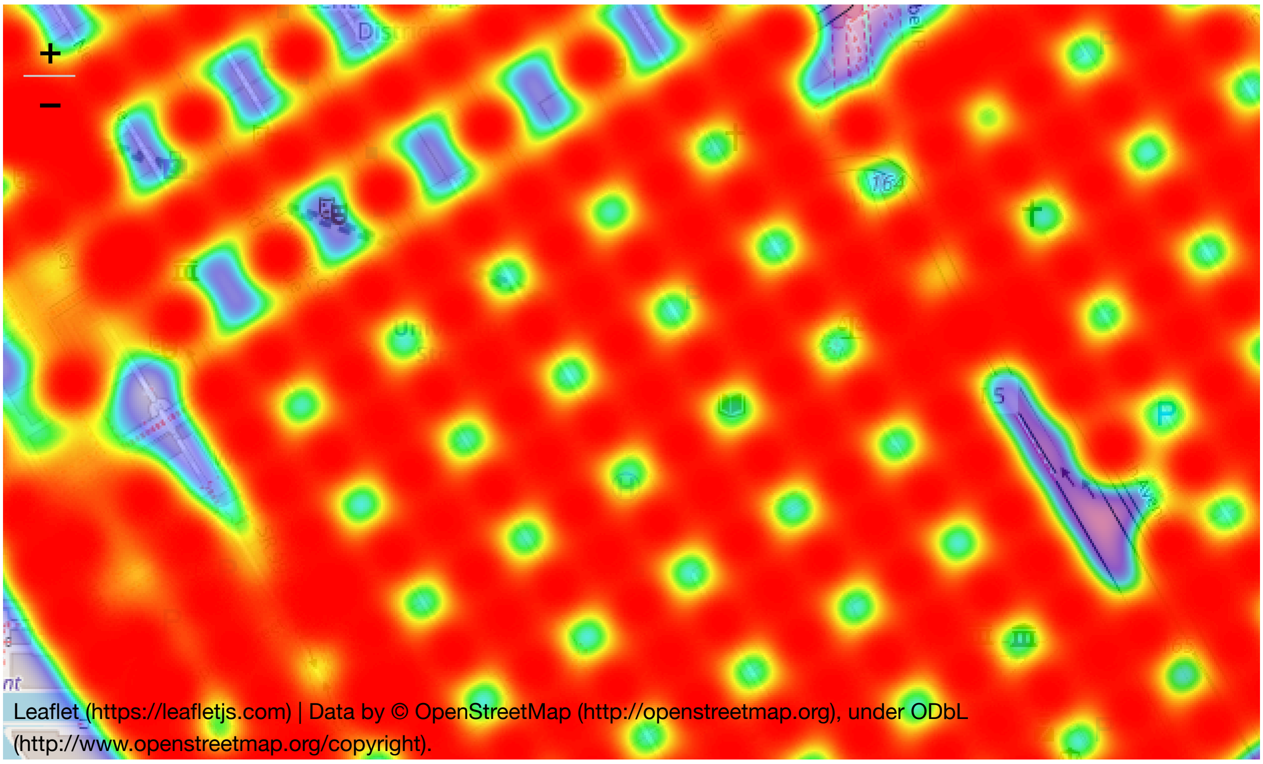
Location(`X` & `Y`)

Since `X` & `Y` are lattitudes and longitudes. So merely describing them won't help us much in understanding the distribution of the data. So let us try and plot them on Seattle's map. We plot a heat map to look at the distribution of accidents across Seattle. We shall also analyze any missing values of `X` and `Y`.

Note: `X` is longitude and `Y` is lattitude.

```
total_missing_values = df[["X", "Y"]].isna().sum().sum()
print(f'There are total of {total_missing_values} missing values of X and Y')
heat_df_sev = df[["X", "Y"]].dropna()
print(f'After dropping na values for heat map we have {heat_df_sev.shape[0]} rows x {heat_df_sev.shape[1]} co
There are total of 10668 missing values of X and Y
After dropping na values for heat map we have 189339 rows x 2 columns
```

```
map_seattle = folium.Map(location=[47.6062, -122.3321], zoom_start=16)
# X is longitude and Y is lattitude.
heat_data = [[row['Y'],row['X']] for index, row in heat_df_sev.iterrows()]
HeatMap(heat_data).add_to(map_seattle)
map_seattle
```


+
−
Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

From the heat map it is pretty evident that we have good coverage over all of Seattle.

ADDRTYPE

```
total_missing_addrtype = df["ADDRTYPE"].isna().sum()
print(f'There are {total_missing_addrtype} valus of ADDRTYPE column')
df["ADDRTYPE"].dropna().value_counts(normalize=True)
```

```
There are 1926 valus of ADDRTYPE column




Block           0.658511
Intersection    0.337593
Alley           0.003896
Name: ADDRTYPE, dtype: float64
```

We have 1926 missing values of `ADDRTYPE` column. `ADDRTYPE` is a categorical column that takes on `3` values:

- Block (65.85% of records)
- Intersection (33.76% of records)

- Alley (0.39% of records)

We have very few accidents reported in alley and mostly on blocks with a significant number on intersection as well. Address type information is encoded within the location attributes, therefore we could as well drop this column.

`PERSONCOUNT` , `PEDCOUNT` , `PEDCYLCOUNT` & `VEHCOUNT`

```
count_df = df[["PERSONCOUNT", "PEDCOUNT", "PEDCYLCOUNT", "VEHCOUNT"]]
missing_vals = count_df.isna().sum(axis=0)
print(f'There are {missing_vals.sum()} missing values for either of the counts')
count_df.describe()
```

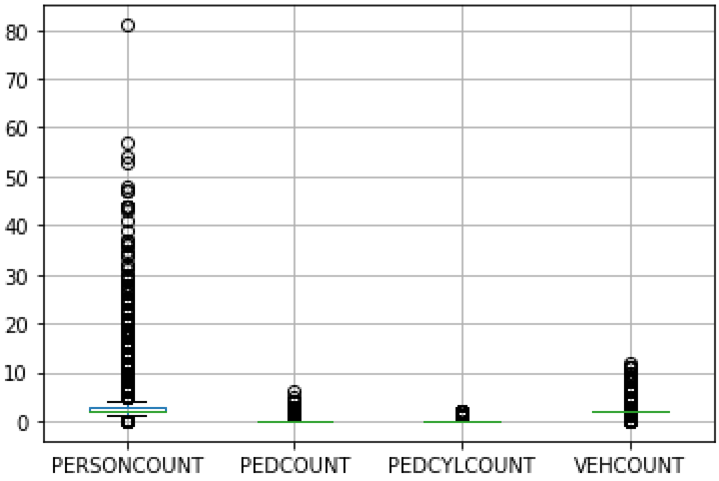
```
There are 0 missing values for either of the counts
```

	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
count	194673.000000	194673.000000	194673.000000	194673.000000
mean	2.444427	0.037139	0.028391	1.920780
std	1.345929	0.198150	0.167413	0.631047
min	0.000000	0.000000	0.000000	0.000000
25%	2.000000	0.000000	0.000000	2.000000
50%	2.000000	0.000000	0.000000	2.000000
75%	3.000000	0.000000	0.000000	2.000000
max	81.000000	6.000000	2.000000	12.000000

We can say that on average 2 persons are involved in an accident in Seattle and around 2 vehicles involved independently of each other. We have an accident reported with 81 people involved and 12 vehicles involved. They seem like outliers but it could be one huge accident chain. Let us draw a box plot to see the distribution.

```
count_df.boxplot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x11c8e2b90>
```



Most of the data is involved around less than 5 persons invloded and 2 vehicles. There seem to be a lot of outliers. However, the counts reported are post metric from an accident and we would not need it directly in our analysis. Rather, we could convert it into a boolean field whether or not a person, pedestrian, cyclist or a vehicle was involved in the accident. Then depending a dependent variable `mode_of_transport` could be used in prediction. We will address this in our `Feature Engineering` section.

Conditions: Weather, Road, Light

```
conditions_df = df.loc[:, ["WEATHER", "ROADCOND", "LIGHTCOND"]]
print(f'There are a total of {conditions_df.shape[0]} rows x {conditions_df.shape[1]} columns')
actual_rows = conditions_df.shape[0]
missing_vals = conditions_df.isna().sum()
print(f'Missing Value Counts:\n{missing_vals}')
conditions_df.dropna(inplace=True)
print(f'After dropping na columns, we are left with {conditions_df.shape[0]} rows x {conditions_df.shape[1]}
print(f'Number of rows dropped: {actual_rows - conditions_df.shape[0]}')
```

```
There are a total of 194673 rows x 3 columns
Missing Value Counts:
WEATHER      5081
ROADCOND     5012
LIGHTCOND    5170
dtype: int64
After dropping na columns, we are left with 189337 rows x 3 columns
Number of rows dropped: 5336
```

```
print(f'----------- WEATHER value counts -----------')
print(conditions_df["WEATHER"].value_counts(normalize=True))

print(f'\n----------- ROADCOND value counts -----------')
print(conditions_df['ROADCOND'].value_counts(normalize=True))

print(f'\n----------- LIGHTCOND value counts: -----------')
print(conditions_df['LIGHTCOND'].value_counts(normalize=True))
```

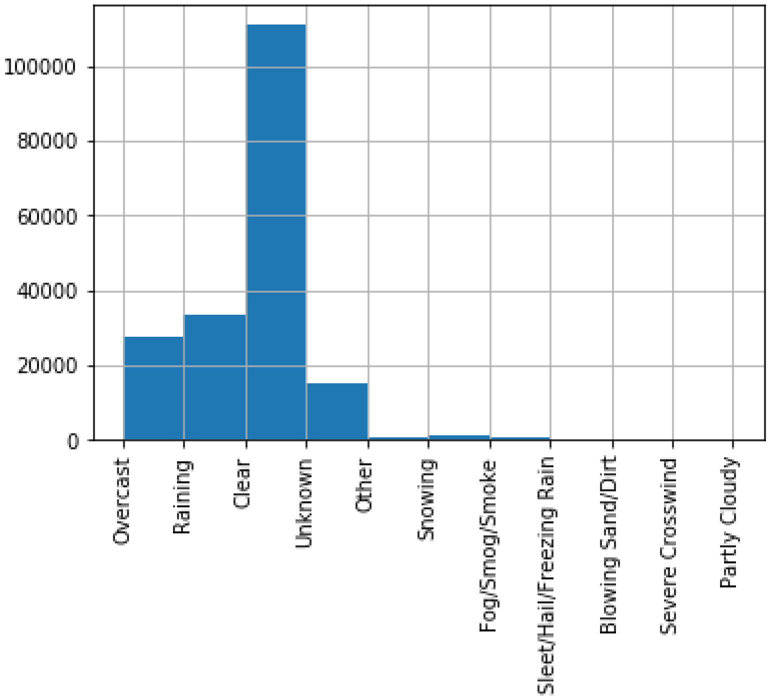
```
----------- WEATHER value counts -----------
Clear                       0.586299
Raining                     0.174910
Overcast                    0.146200
Unknown                     0.079430
Snowing                     0.004759
Other                       0.004352
```

```
Fog/Smog/Smoke              0.003005
Sleet/Hail/Freezing Rain    0.000597
Blowing Sand/Dirt           0.000290
Severe Crosswind            0.000132
Partly Cloudy               0.000026
Name: WEATHER, dtype: float64

----------- ROADCOND value counts -----------
Dry               0.656501
Wet               0.250437
Unknown           0.079388
Ice               0.006370
Snow/Slush        0.005276
Other             0.000692
Standing Water    0.000607
Sand/Mud/Dirt     0.000391
Oil               0.000338
Name: ROADCOND, dtype: float64

----------- LIGHTCOND value counts: -----------
Daylight                    0.613071
Dark - Street Lights On     0.255840
Unknown                     0.071069
Dusk                        0.031103
Dawn                        0.013215
Dark - No Street Lights     0.008107
Dark - Street Lights Off    0.006296
Other                       0.001241
Dark - Unknown Lighting     0.000058
Name: LIGHTCOND, dtype: float64
```

7.9% of the values are `Unknown` , they can also be dropped. For light conditions, there are 3 different categorizations in `Dark` . They can all be combined into single as `Dark` . Let us also look at the histogram of the values.

```
conditions_df["WEATHER"].hist(xrot=90)
```

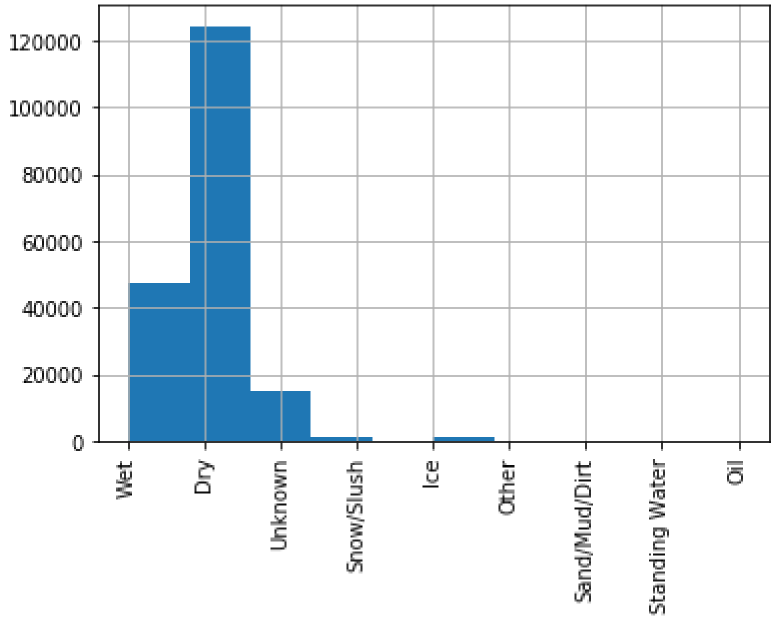
```
<matplotlib.axes._subplots.AxesSubplot at 0x11ea9a1d0>
```



Around 58.6% of the values are for `Clear` weather followed by `Raining` and `Overcast`.

```
conditions_df["ROADCOND"].hist(xrot=90)
```

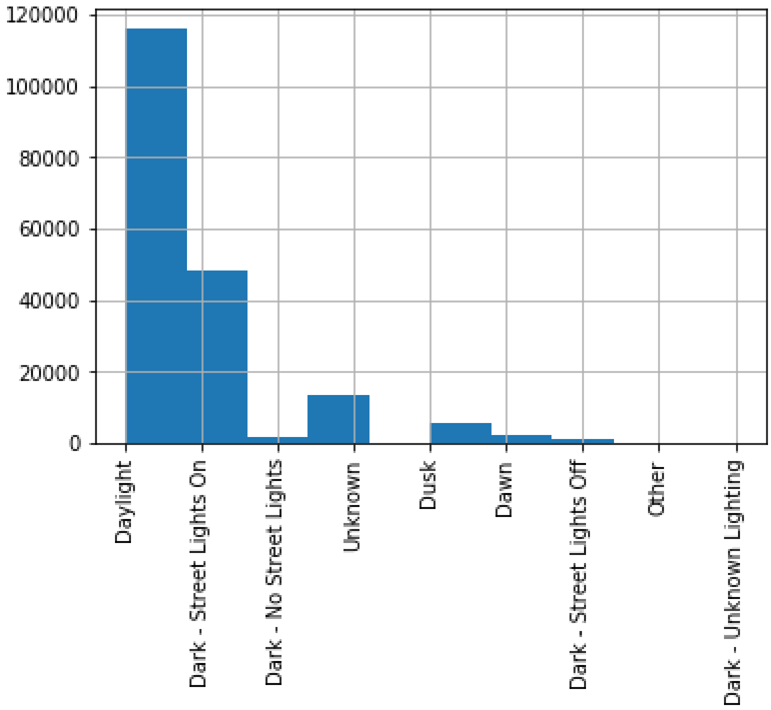
```
<matplotlib.axes._subplots.AxesSubplot at 0x11eb31d50>
```



Around 65% of the values are for `Dry` weather followed by `Wet` and `Unknown`.

```
conditions_df["LIGHTCOND"].hist(xrot=90)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x115997550>
```



Around 61.3% value are for Daylight followed by `Dark: Street Lights On` and `Unknown`.

HITPARKEDCAR

Let us look more into `HITPARKEDCAR` distribution to see if we can utilize this variable as whether or not a commuter is planning to park a car on road.

```
missing_vals = df["HITPARKEDCAR"].isna().sum()
print(f'Total missing values: {missing_vals}')
df["HITPARKEDCAR"].value_counts(normalize=True)
```

```
Total missing values: 0





N    0.962933
Y    0.037067
Name: HITPARKEDCAR, dtype: float64
```

The distribution is pretty much skewed with 96% of the records towards not hitting a parked car. So, we can drop the records with `HITPARKEDCAR='Y'` for the sake of convenience and simplicity.

Key Takeaways

From the exploratory analysis of data, we have identified the following to be used for our prediction problem:

- `SEVERITYCODE`
- `X` & `Y`
- `PEDCOUNT`, `PEDCYLCOUNT`, `VEHCOUNT`
- `WEATHER`, `ROADCOND`, `LIGHTCOND`

However, we would need to do some kind of feature engineering to be able to utilize the above mentioned columns for our predicition problem. We discuss that in the next section called *Feature Engineering*.

Data Cleaning Takeaways

The following values need to be dropped:

- `NA` or missing values for all the columns
- `Unknown` for `WEATHER`, `ROADCOND`, `LIGHTCOND`
- Values where `HITPARKEDCAR` is `Y`.

Let us check if we have sufficient data for our analysis following the clean up.

```
# wdf: working data frame
# WE keep HITPARKEDCAR for cleaning purposes and drop it later on.
wdf = df.loc[:, ["SEVERITYCODE", "X", "Y", "WEATHER", "ROADCOND", "LIGHTCOND", "HITPARKEDCAR", "PEDCOUNT", "P
wdf.dropna(inplace=True)
wdf = wdf[wdf["HITPARKEDCAR"]=='N']
wdf.drop(columns=["HITPARKEDCAR"], inplace=True)
wdf = wdf[wdf["WEATHER"] != 'Unknown']
wdf = wdf[wdf["ROADCOND"] != 'Unknown']
wdf = wdf[wdf["LIGHTCOND"] != 'Unknown']
print(f'After clean up, we have {wdf.shape[0]} rows x {wdf.shape[1]} columns')
```

```
After clean up, we have 161913 rows x 9 columns
```

```
print('--------- SEVERITYCODE value counts: ---------')
print(wdf["SEVERITYCODE"].value_counts(normalize=True))

print('\n--------- WEATHER value counts: ---------')
print(wdf["WEATHER"].value_counts(normalize=True))

print('\n--------- ROADCOND value counts: ---------')
print(wdf["ROADCOND"].value_counts(normalize=True))

print('\n--------- LIGHTCOND value counts: ---------')
print(wdf["LIGHTCOND"].value_counts(normalize=True))
```

```
--------- SEVERITYCODE value counts: ---------
1    0.663677
2    0.336323
Name: SEVERITYCODE, dtype: float64

--------- WEATHER value counts: ---------
Clear                       0.639170
Raining                     0.191294
Overcast                    0.159042
Snowing                     0.004768
Fog/Smog/Smoke              0.003205
Other                       0.001445
Sleet/Hail/Freezing Rain    0.000661
Blowing Sand/Dirt           0.000247
Severe Crosswind            0.000142
Partly Cloudy               0.000025
Name: WEATHER, dtype: float64

--------- ROADCOND value counts: ---------
Dry               0.714680
Wet               0.272214
Ice               0.006429
Snow/Slush        0.004861
```

```
Standing Water    0.000587
Other             0.000574
Sand/Mud/Dirt     0.000346
Oil               0.000309
Name: ROADCOND, dtype: float64

--------- LIGHTCOND value counts: ---------
Daylight                    0.665716
Dark - Street Lights On     0.271905
Dusk                        0.033357
Dawn                        0.014038
Dark - No Street Lights     0.007720
Dark - Street Lights Off    0.006281
Other                       0.000939
Dark - Unknown Lighting     0.000043
Name: LIGHTCOND, dtype: float64
```

```
wdf.head()
```

	SEVERITYCODE	X	Y	WEATHER	ROADCOND	LIGHTCOND	PEDC
0	2	-122.323148	47.703140	Overcast	Wet	Daylight	0
1	1	-122.347294	47.647172	Raining	Wet	Dark - Street Lights On	0
2	1	-122.334540	47.607871	Overcast	Dry	Daylight	0
3	1	-122.334803	47.604803	Clear	Dry	Daylight	0
4	2	-122.306426	47.545739	Raining	Wet	Daylight	0

Feature Engineering

In the light of the data understanding, Let us formally define our prediction problem,

Given the area of commute, weather, road and light conditions and the mode of transport we would like to predict the severity of the accident.

For our problem, we would need to perform certain actions so that the data is consumable for our prediction models.

- Location: `X` & `Y` categorized into Area
  Considering the ease of use, we would like to calculate severity of an accident in a certain area as compared to a certain point on the earth. So, we would need to categorize the `X` and `Y` coordinates into certain areas we can train our models on. This could be administrative area or a smaller part of the city. This problem will be addressed in the data preparation section.
- Dummy columns for categorical columns: `WEATHER`, `ROADCOND`, `LIGHTCOND`
- Convert the count columns as boolean values with a `TRUE` assigned to `value > 0` and `FALSE` assigned to `value == 0`.

Data Preparation

Modelling

Evaluation