IBM Capstone Project Report
Analysis of Seattle Traffic Accidents Data
Law Lap Wun, Natalie

1. Introduction and Understand the problem

   Traffic accidents occur in everywhere around the world with very high frequency. There are many factors that cause a traffic accident to happen. For example, poor weather condition, difficult road condition, insufficient light intensity and irresponsible drivers.

   In order to reduce the number of traffic accidents, we need to understand the causes of common traffic accidents so as to remind people to be aware of these traffic accidents and minimize the chances of inducing traffic accidents.

   Therefore, we deeply hope that by doing this analysis, we can educate people and give some useful insights to related authority like transport department.

2. Data Sources

   The data contributing in this capstone project come from IBM Data Science Capstone Course. The data is about all accidents or collisions that happened in Seattle. From these data, we will be able to understand the causes of traffic accidents and what makes the accidents more serious.

3. Summary of data set

- Title: Collisions—All Years
- Abstract: All collisions provided by SPD and recorded by Traffic Records.
- Description: This includes all types of collisions. Collisions will display at the intersection or
- mid-block of a segment. Timeframe: 2004 to Present.
- Update Frequency: Weekly
- Keyword(s): SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle
- Organization: SDOT Traffic Management Division, Traffic Records Group
- Contact Person: SDOT GIS Analyst
- Contact Email: DOT_IT_GIS@seattle.gov

4. Attribute information

Attribute	Data Type, Length	Description
OBJECTID	Long	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
ADDRTYPE	Text, 12	Collision address type: Alley, Block, Intersection
INTKEY	Double	Key that corresponds to the intersection associated with a collision
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: 3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision Type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

5. Data Cleaning

Since there are missing values, incorrect data types and useless columns and features, therefore, I will drop the unneeded columns, columns with missing values, columns that have the same value in all columns. Also, I will use the desc column as well.

```
df.drop(['SEVERITYCODE','X','Y','OBJECTID','INCKEY','COLDETKEY',
         'REPORTNO','STATUS','INTKEY','LOCATION','EXCEPTRSNCODE','EXCEPTRSNDESC',
         'SEVERITYCODE.1','SDOT_COLCODE','INATTENTIONIND','PEDROWNOTGRNT',
         'SDOTCOLNUM','SPEEDING','SEGLANEKEY','CROSSWALKKEY'],axis=1, inplace=True)
```

Next, I will fix ST_COLCODE data because there are duplicated values and empty data.

Before:

```
array(['10', '11', '32', '23', '5', '22', '14', '30', ' ', '28', '51',
       '13', '50', '12', '45', '0', '20', '21', '1', '52', '16', '15',
       '74', '81', '26', '19', '2', '66', '71', '3', '24', '40', '57',
       '6', '83', '25', '27', '4', '72', '29', '56', '73', '41', '17',
       '65', '82', '67', '49', '84', '31', '43', '42', '48', '64', '53',
       32, 50, 15, 10, 14, 20, 13, 22, 51, 11, 28, 12, 52, 21, 0, 19, 30,
       16, 40, 26, 27, 83, 2, 45, 65, 23, 24, 71, 1, 29, 81, 25, 4, 73,
       74, 72, 3, 84, 64, 57, 42, 41, 48, 66, 56, 31, 82, 67, '54', '60',
       53, 43, 87, 54, '87', nan, '7', '8', '85', '88', '18'],
      dtype=object)
```

After:

```
array([10., 11., 32., 23.,  5., 22., 14., 30., nan, 28., 51., 13., 50.,
       12., 45.,  0., 20., 21.,  1., 52., 16., 15., 74., 81., 26., 19.,
        2., 66., 71.,  3., 24., 40., 57.,  6., 83., 25., 27.,  4., 72.,
       29., 56., 73., 41., 17., 65., 82., 67., 49., 84., 31., 43., 42.,
       48., 64., 53., 54., 60., 87.,  7.,  8., 85., 88., 18.])
```

6. Exploratory Data Analysis

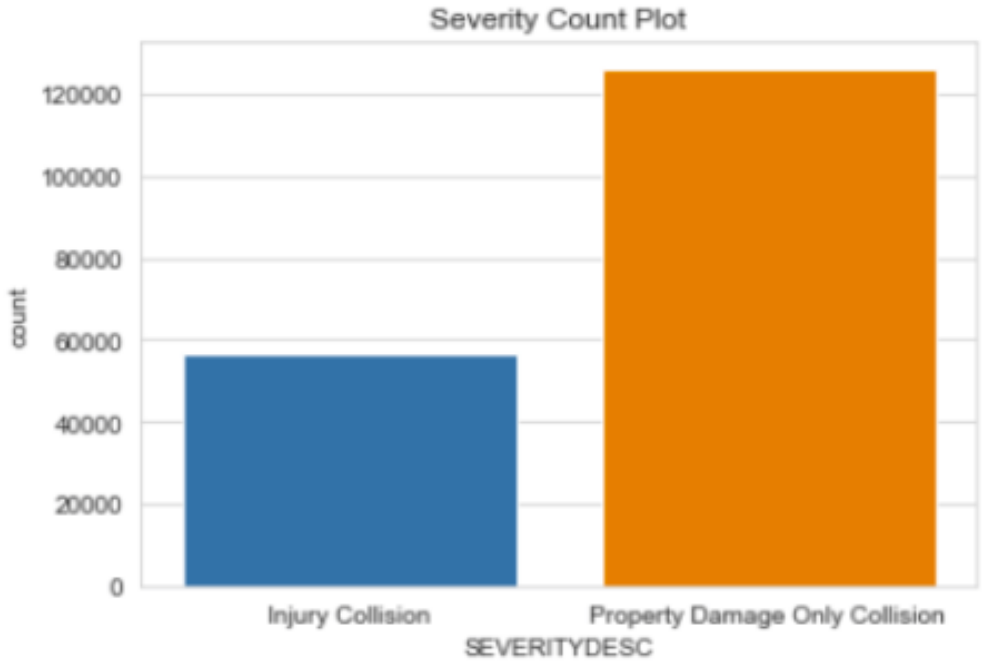
By using df.describe(), I can check the distribution of numerical columns.

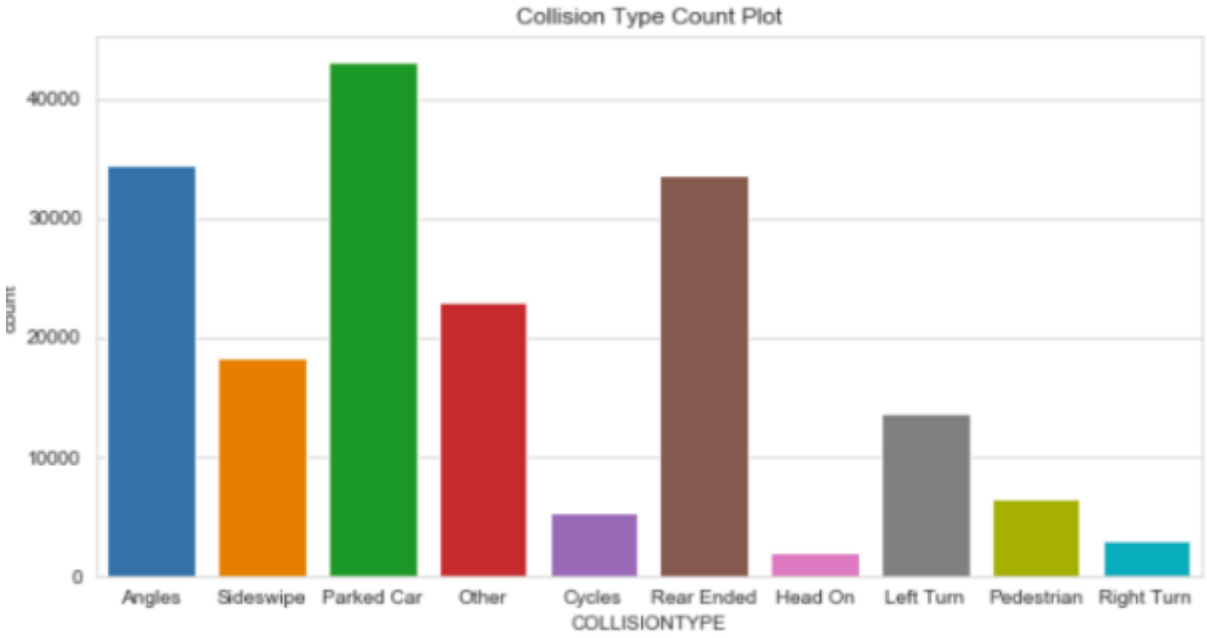
	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	UNDERINFL	ST_COLCODE
count	182895.000000	182895.000000	182895.000000	182895.000000	182895.000000	182895.000000
mean	2.476268	0.038995	0.029831	1.971984	0.049192	22.583411
std	1.370912	0.202960	0.171435	0.563237	0.216270	14.575891
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	0.000000	0.000000	2.000000	0.000000	11.000000
50%	2.000000	0.000000	0.000000	2.000000	0.000000	15.000000
75%	3.000000	0.000000	0.000000	2.000000	0.000000	32.000000
max	81.000000	6.000000	2.000000	12.000000	1.000000	88.000000

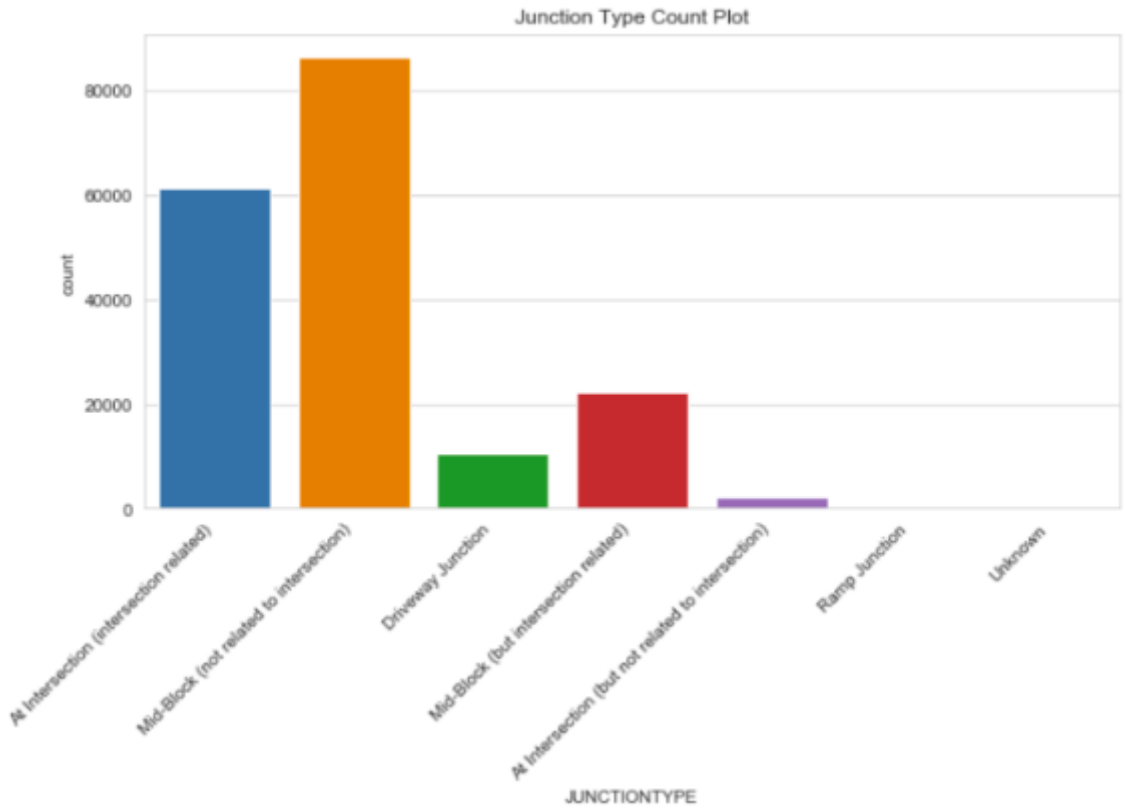
Furthermore, I checked the distribution of the categorical columns.

	ADDRTYPE	SEVERITYDESC	COLLISIONTYPE	JUNCTIONTYPE	SDOT_COLDESC	WEATHER	ROADCOND	LIGHTCOND	ST_COLDESC	HITPARKEDCAR
count	182895	182895	182895	182895	182895	182895	182895	182895	182895	182895
unique	3	2	10	7	39	11	9	9	62	2
top	Block	Property Damage Only Collision	Parked Car	Mid-Block (not related to intersection)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	Clear	Dry	Daylight	One parked--one moving	N
freq	119362	126270	43119	86609	83024	109059	122153	113837	39619	177205

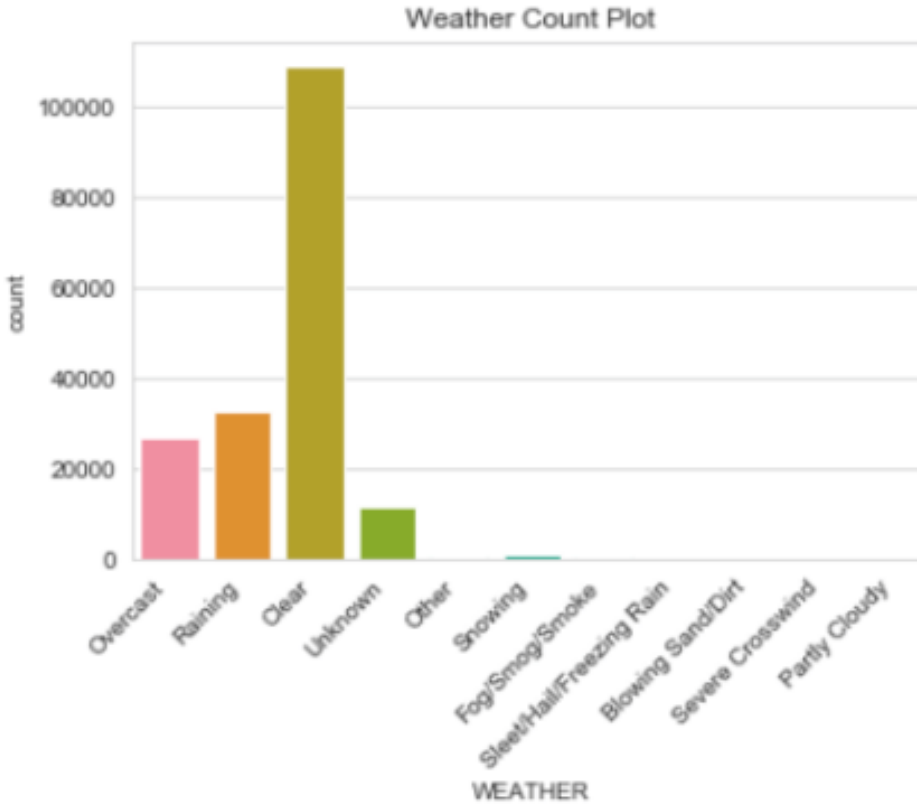
Here is the distribution of the Severitydesc column, Collisiontype data and Junctiontype column respectively.

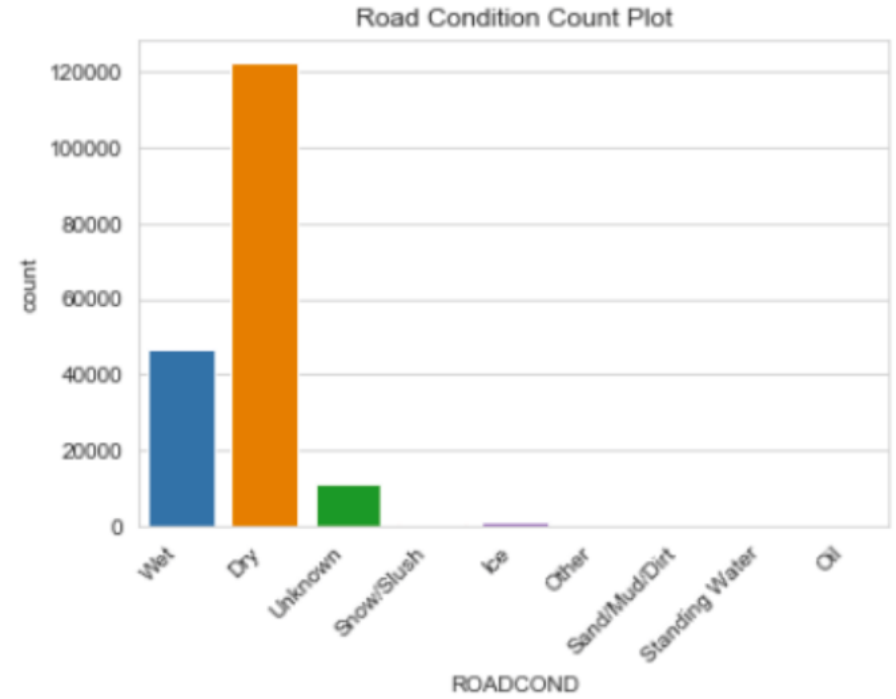


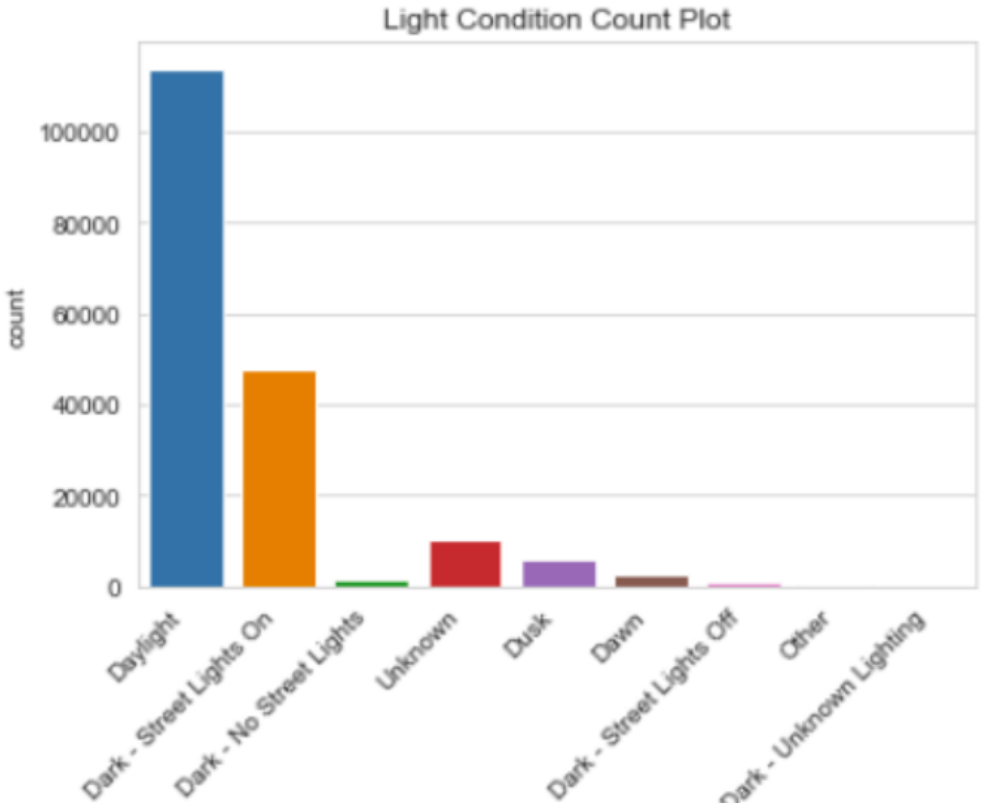




Here is the distributions of other categorical columns.

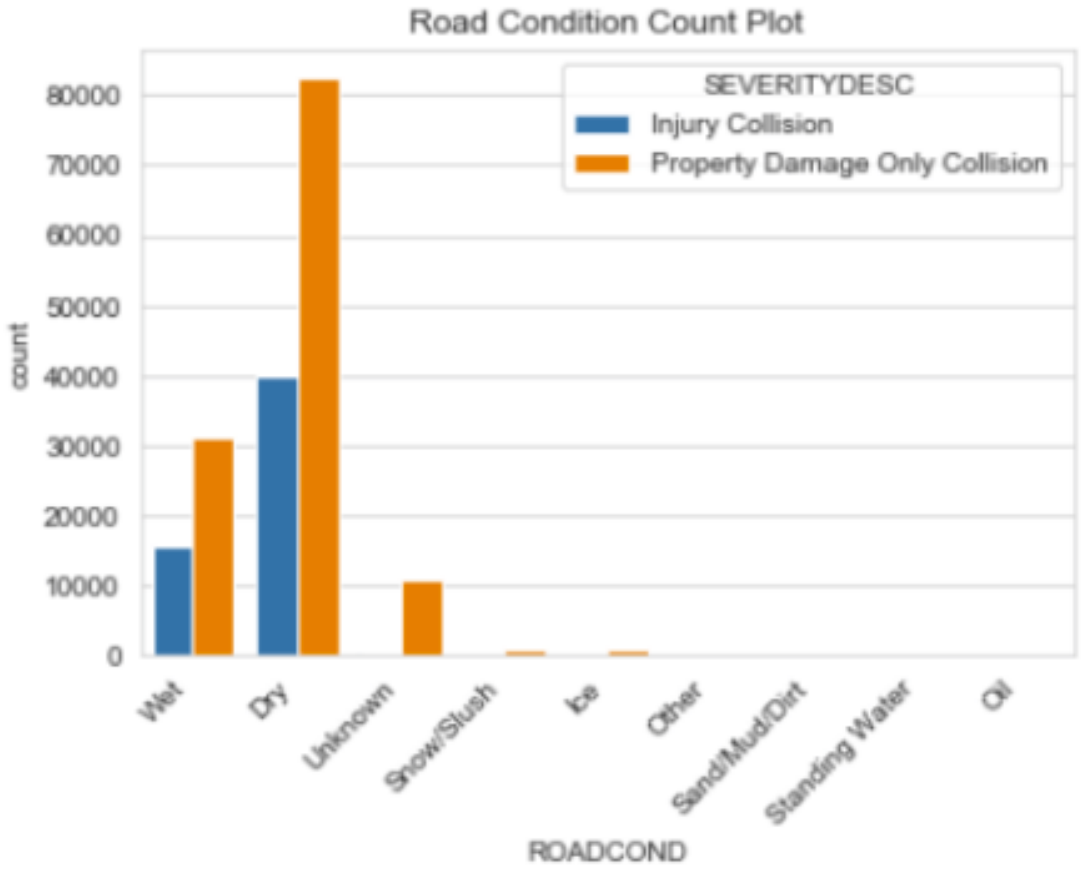




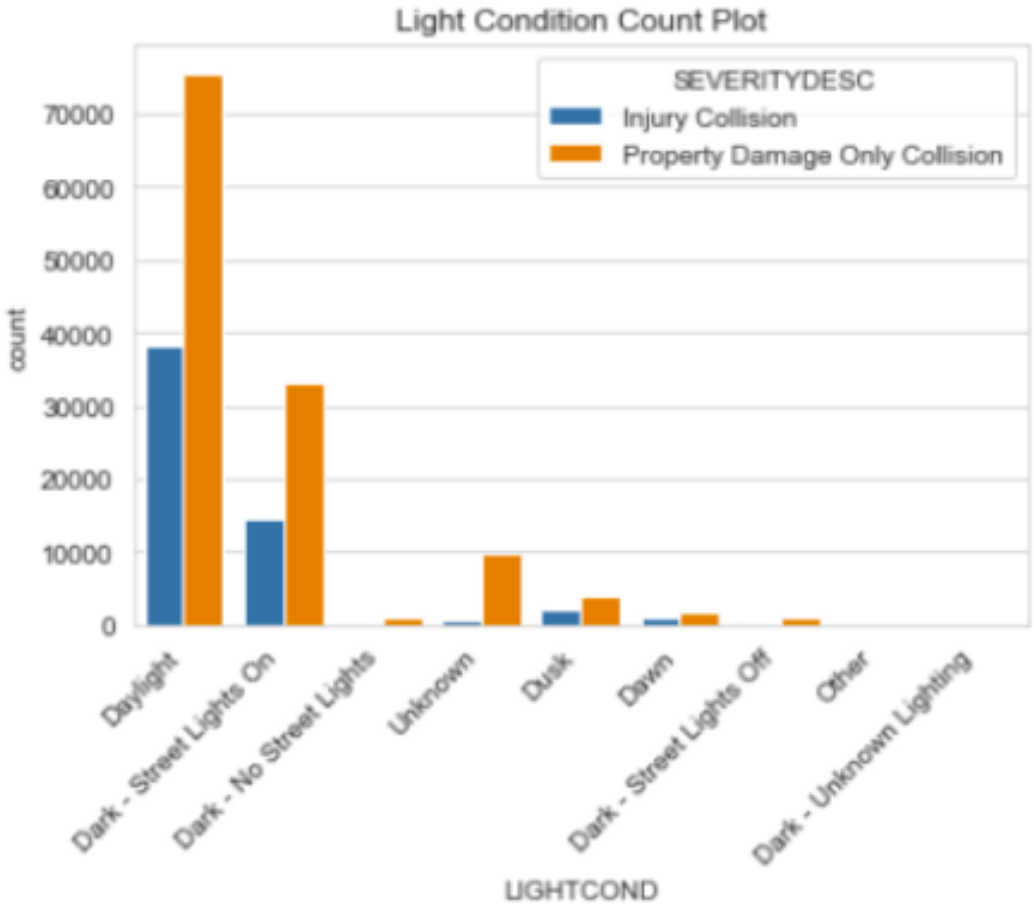


Then, I compare the ratio between Severitydesc injury collision and property damage only collision in different conditions.

SEVERITYDESC WEATHER	Injury Collision	Property Damage Only Collision
Partly Cloudy	0.600000	0.400000
Raining	0.339532	0.660468
Fog/Smog/Smoke	0.334532	0.665468
Clear	0.326273	0.673727
Overcast	0.318986	0.681014
Severe Crosswind	0.280000	0.720000
Blowing Sand/Dirt	0.265306	0.734694
Sleet/Hail/Freezing Rain	0.241071	0.758929
Snowing	0.189557	0.810443
Other	0.152815	0.847185
Unknown	0.066254	0.933746

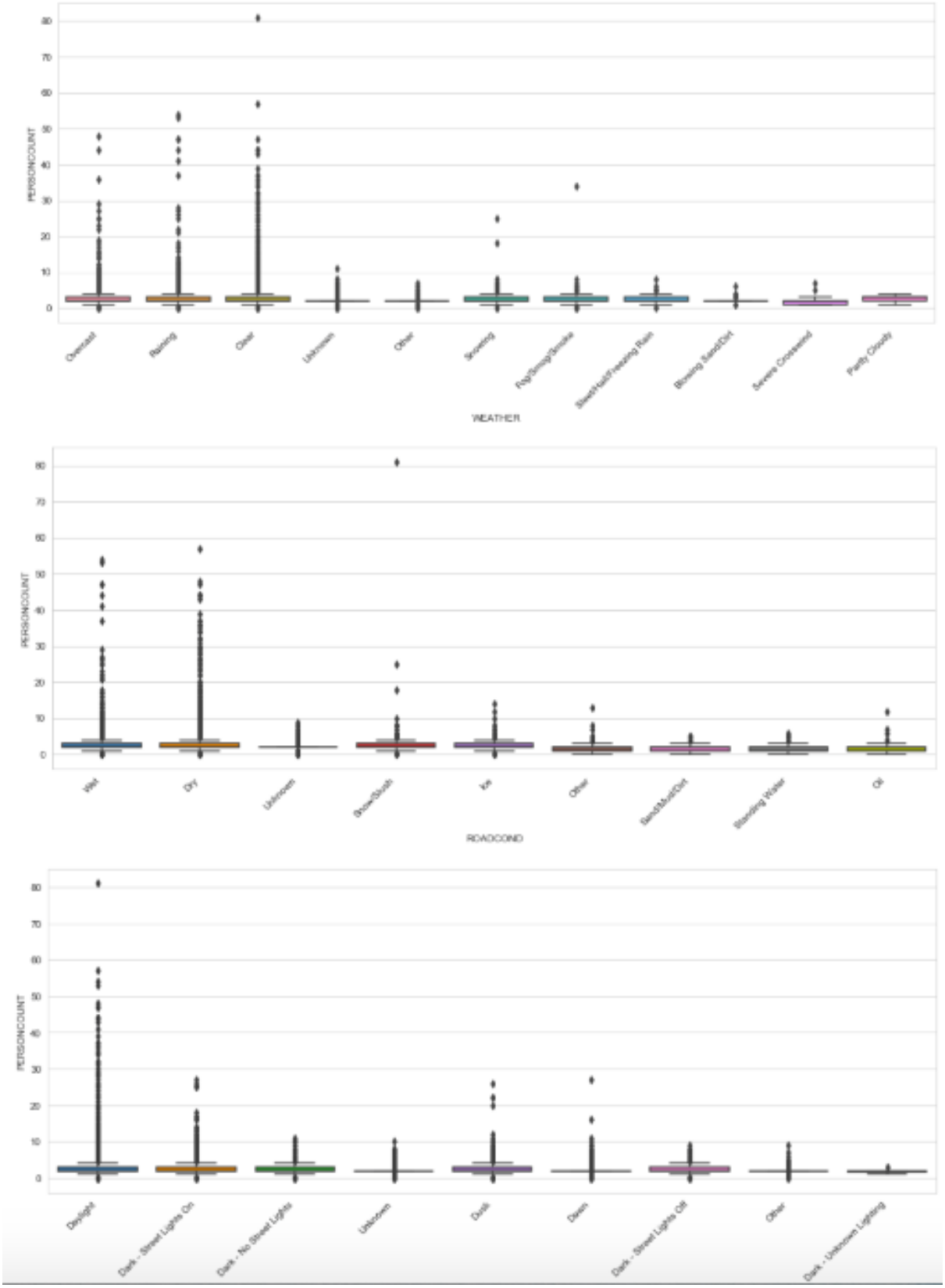


SEVERITYDESC ROADCOND	Injury Collision	Property Damage Only Collision
Oil	0.400000	0.600000
Other	0.341463	0.658537
Wet	0.334618	0.665382
Sand/Mud/Dirt	0.328358	0.671642
Dry	0.325322	0.674678
Standing Water	0.268519	0.731481
Ice	0.226848	0.773152
Snow/Slush	0.168712	0.831288
Unknown	0.061377	0.938623

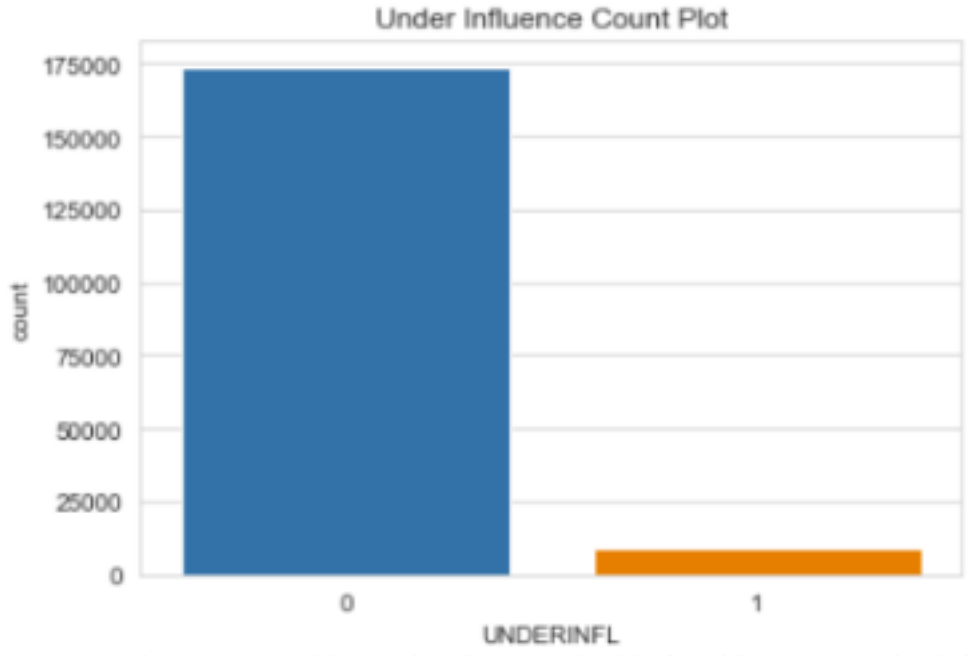


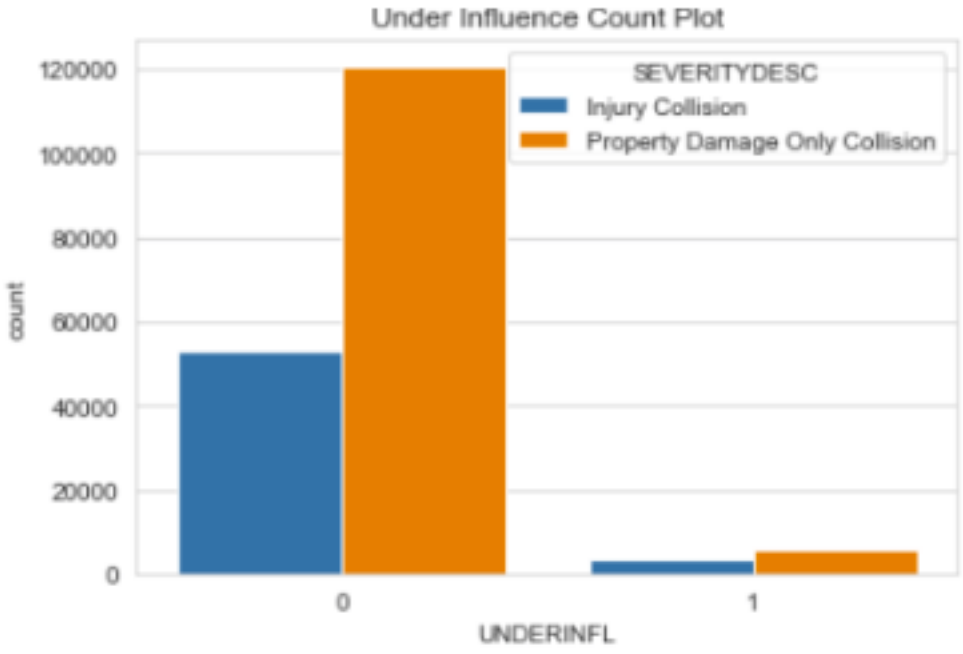
SEVERITYDESC LIGHTCOND	Injury Collision	Property Damage Only Collision
Dark - Unknown Lighting	0.363636	0.636364
Daylight	0.336059	0.663941
Dawn	0.333877	0.666123
Dusk	0.333738	0.666262
Dark - Street Lights On	0.301828	0.698172
Dark - Street Lights Off	0.270527	0.729473
Other	0.247619	0.752381
Dark - No Street Lights	0.224504	0.775496
Unknown	0.055130	0.944870

Next, I check the impact on total number of people involved in collision due to weather, road and light condition.



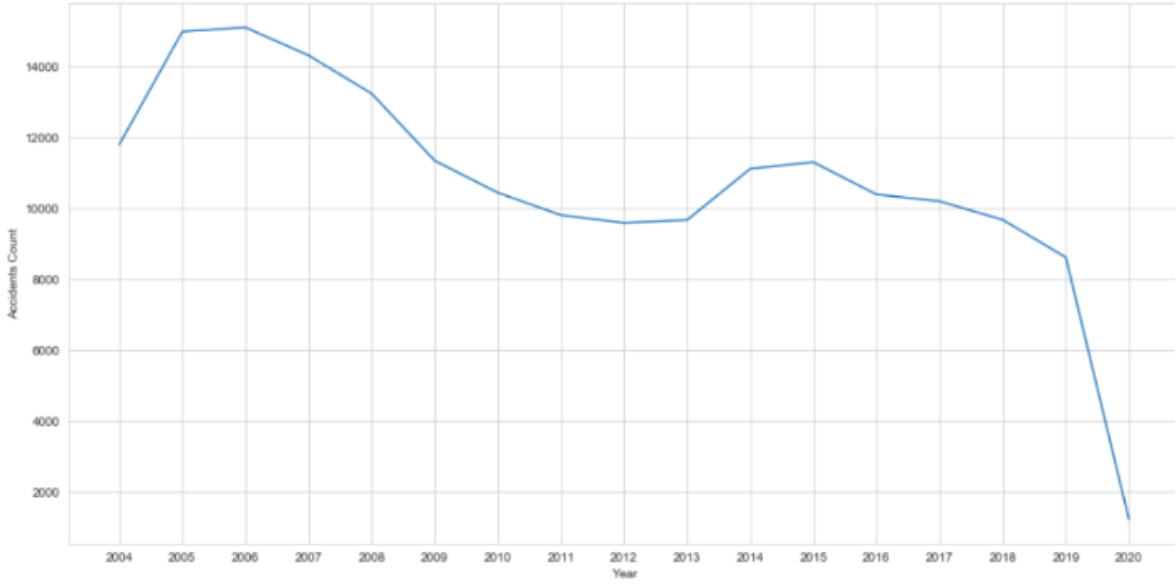
Then, I will see if the traffic accidents happened due to the fact that the driver took drugs or alcohol.





SEVERITYDESC UNDERINFL	Injury Collision	Property Damage Only Collision
1	0.392131	0.607869
0	0.305334	0.694666

Next, I would like to see the accident trend from 2004 to 2020.



7. Prediction modelling

I used Random Forest Classifier Algorithm.

```
rfc = RandomForestClassifier(n_estimators=100, random_state=101,
                             max_depth=10, class_weight={'Injury Collision':2, 'Property Damage Only Collision': 1 })
rfc.fit(X_train, y_train)
```

Here is the evaluation.

	precision	recall	f1-score	support
Injury Collision	0.34	0.79	0.47	16597
Property Damage Only Collision	0.70	0.25	0.37	33757
accuracy			0.43	50354
macro avg	0.52	0.52	0.42	50354
weighted avg	0.58	0.43	0.40	50354

8. Result

The performance model did make sense. However, it is difficult to determine what type of accidents or collisions that we cannot predict under what circumstances the accidents will happen. It makes senses because accidents are what we cannot predict.

9. Conclusion

Although the data only recorded the traffic accidents in Seattle, not other places, it can still be used as reference in other places all over the world.

Also, it is safer for drivers not to take any drugs and drink any alcohol before driving.

Last but not least, though there is no direct relation in different conditions in weather, road and light etc, we should be more aware of poor weather condition , difficult road condition and insufficient light intensity before and during driving.