Data Science Capstone Project

In this capstone project for the IBM Certificate course offered by Coursera, I was given a large dataset that focused on various data aspects regarding traffic accidents. The primary goal of this project will be to pinpoint the main conditions that would cause traffic accidents. I would like to reduce the frequency of collisions y taking a further look into how weather, road, and lighting conditions affect driving. The dataset included various factors and information such as severity, weather, road, and light conditions. I decided to compare these factors since I believe that they are the most important factors currently affecting driving conditions.

The source of this traffic collision data was provided by Coursera. This data was collected by a police department and included over a decade's worth of information. With this dataset, I would like to take a close look at its severity in relation to weather, lighting, and road conditions. In order to fully use this data, I had to clean it first by removing the unnecessary data points that I will not be using. There are many columns that I will not be using for this project.

For this project, I will be using many different libraries such as pandas, numpy, and matplotlib to complete my data science needs. The first steps to this analysis was to load and read the dataset that was provided to me.

```
import itertools
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.ticker import NullFormatter
import pandas as pd
from sklearn import preprocessing
%matplotlib inline
```

```
df = pd.read_csv("https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2
df.head(10)
```

After doing this, I needed to understand the types of data I was dealing with so I called a function to display out the data types of each column. This would help me understand which columns contained objects , floats, or ints. Upon reviewing this information, I decided to use panda's drop function to remove all the unnecessary data columns that I will not be using for this project to avoid confusion and clutter.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 9 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   SEVERITYCODE    194673 non-null  int64 
 1   ADDRTYPE        192747 non-null  object
 2   SEVERITYCODE.1  194673 non-null  int64 
 3   SEVERITYDESC    194673 non-null  object
 4   COLLISIONTYPE   189769 non-null  object
 5   JUNCTIONTYPE    188344 non-null  object
 6   WEATHER         189592 non-null  object
 7   ROADCOND        189661 non-null  object
 8   LIGHTCOND       189503 non-null  object
dtypes: int64(2), object(7)
memory usage: 13.4+ MB
```

In order to understand the root causes for traffic collisions, I will be comparing three separate factors to collision severity: weather, lighting, and road conditions. In the next section, I will be creating bar graphs to compare these severities and factors on driving conditions.

```
b = df.groupby(['SEVERITYCODE','ROADCOND'
b
```

SEVERITYCODE	ROADCOND	
1	Dry	84446
	Ice	936
	Oil	40
	Other	89
	Sand/Mud/Dirt	52
	Snow/Slush	837
	Standing Water	85
	Unknown	14329
	Wet	31719
2	Dry	40064
	Ice	273
	Oil	24
	Other	43
	Sand/Mud/Dirt	23
	Snow/Slush	167
	Standing Water	30
	Unknown	749
	Wet	15755

dtype: int64

```
c = df.groupby(['SEVERITYCODE','LIGHTCOND']).siz
c
```

SEVERITYCODE	LIGHTCOND	
1	Dark - No Street Lights	1203
	Dark - Street Lights Off	883
	Dark - Street Lights On	34032
	Dark - Unknown Lighting	7
	Dawn	1678
	Daylight	77593
	Dusk	3958
	Other	183
	Unknown	12868
2	Dark - No Street Lights	334
	Dark - Street Lights Off	316
	Dark - Street Lights On	14475
	Dark - Unknown Lighting	4
	Dawn	824
	Daylight	38544
	Dusk	1944
	Other	52
	Unknown	605

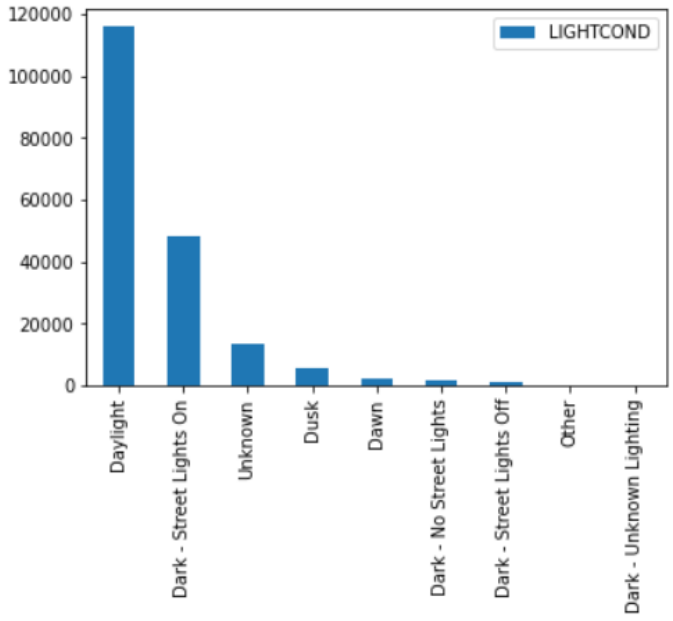
dtype: int64

```
a = df.groupby(['SEVERITYCODE','WEATHER']).size()
a
```

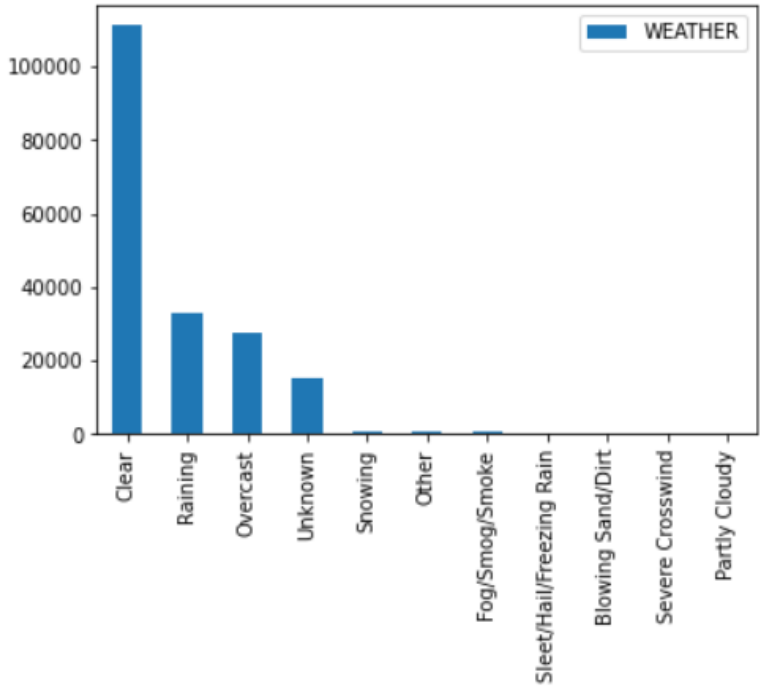
SEVERITYCODE	WEATHER	
1	Blowing Sand/Dirt	41
	Clear	75295
	Fog/Smog/Smoke	382
	Other	716
	Overcast	18969
	Partly Cloudy	2
	Raining	21969
	Severe Crosswind	18
	Sleet/Hail/Freezing Rain	85
	Snowing	736
	Unknown	14275
2	Blowing Sand/Dirt	15
	Clear	35840
	Fog/Smog/Smoke	187
	Other	116
	Overcast	8745
	Partly Cloudy	3
	Raining	11176
	Severe Crosswind	7
	Sleet/Hail/Freezing Rain	28
	Snowing	171
	Unknown	816

dtype: int64

For the first bar graph I created using matplotlib, we can see that the most common condition in which traffic collisions happen is during the daylight but this isn't a significant factor since this is the most common scenario in which people are driving. Taking another look at this graph, we can see that the second most common condition is during the dark with the street lights on. In this case, we would like to advise drivers to be more cautious when driving in the dark.

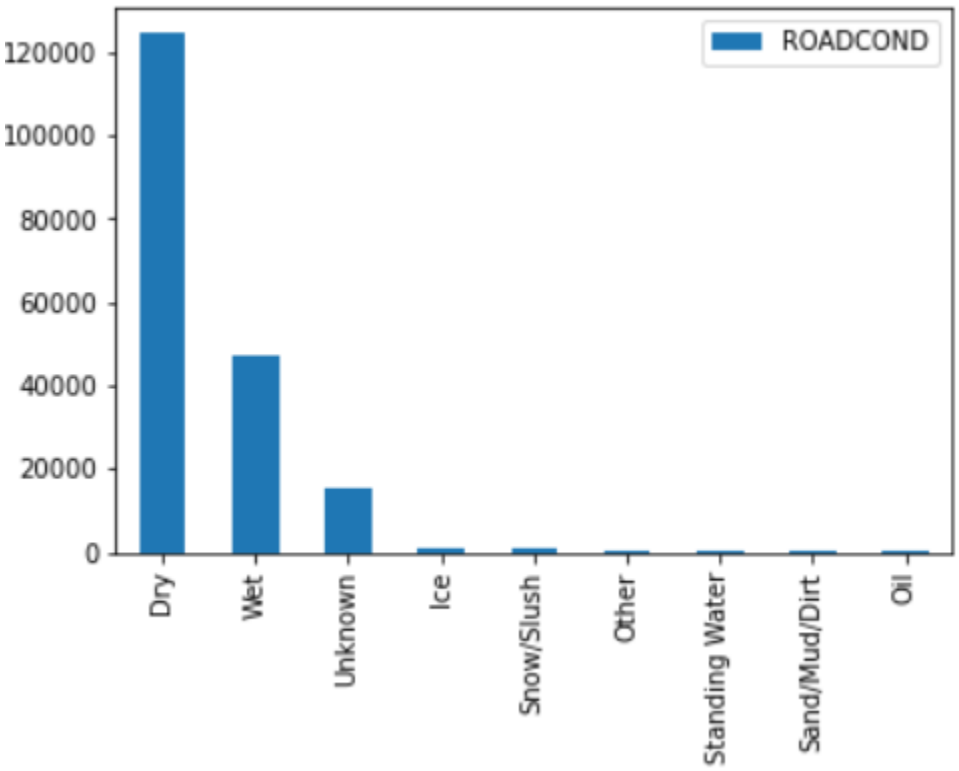


In this next graphical analysis, we compare weather with traffic collisions and we could see that the most common factor is clear weather but this doesn't display a significant impact in traffic collisions since this is also the most common weather when people are driving so there are external factors other than weather affecting these collisions during clear weather. What seems to be more significant in affecting driving conditions would be raining and overcast weather. This is most likely due to the fact that the roads are more slippery and the weather can be visually impairing for some people. Therefore, more collisions would happen.



In my last bar chart analyzing traffic collisions, we take a further look at how road conditions affect driving. The most common road conditions when these incidents are happening is dry and wet roads. We cannot say for sure that dry roads are the cause of these traffic collisions since that is the most common road condition but for the wet roads, it is plausible that the slippery roads cause drivers to be involved in more traffic collisions. What we advise for the

people using this model is to recommend drivers to be cautious of slippery roads so more warning signs could place in roads that cause the most accidents when it is wet.



In conclusion, by comparing these three factors with traffic collisions we can see the most common conditions in which these accidents occur. To recap the most common conditions during traffic collisions, we can see that the most common conditions were accidents during daylight, night time, clear weather, raining weather, overcast, dry roads, and wet roads.