IBM Applied Data Science Capstone
Predicting Road Accidents in Seattle

Author: Ilan Gil

Date: 20/09/2020

Contents

1. Problem Description

Road accidents are one of the leading causes of death globally. In the USA alone, more than 38'000 people died in motor vehicle collisions last year. This particular project will focus on the American city of Seattle. Located on the West Coast of the USA, it is the largest city in the state of Washington with a population above 700'000. Naturally, this high concentration of people implies a high number of motor vehicles and the corresponding accidents usually associated with them.

These accidents often involve dramatic consequences. On top of direct health damages, people involved in such accidents may expect large medical bills, significant property damage, or missing revenue from their inability to work. Immediately after occurring, these accidents also require a very swift response from the authorities. Preparation and coordination between the paramedic team, the police department, and the nearest hospitals is key in order to ensure an appropriate response.

This project will therefore focus on the ''response'' side of these accidents. In other words, the upcoming analysis will strive to shine light on some factors that may indicate a higher likelihood that accidents will take place, as well as at the potential ways that authorities may prepare themselves in order to respond to these accidents.

Being able to anticipate and better prepare oneself to incoming accidents could prove very useful for a number of federal bodies. In detail:

- Hospitals could use those predictions in order to coordinate their staff strength with respect to the amount of accidents prone to occurring on a given day. For instance, should the data analysis indicate that accidents are more likely to occur on say Fridays, hospitals could respond by increasing their headcount on that specific day.
- Police departments could be interested in those predictions insofar as they would allow them to adapt their level of alertness if the probability of severe accidents turns out to be higher than usual on certain days.
- Governing authorities may also use those predictions, notably in order to warn people of a potentially dangerous day when it comes to road accidents.

2. Data

This analysis will use an extensive dataset from the Seattle Police Department, with over 190'000 observations since 2004.

The dataset includes a number of factors that could help us determine the likelihood of a road accident occurring. These factors pretty much describe the overall context in which an accident takes place. Some of the most important factors include:

- **SEVERITYCODE**: A code that corresponds to the severity of the collision
- **PERSONCOUNT**: The total number of people involved in the collision
- **VEHCOUNT**: The number of vehicles involved in the collision
- **WEATHER:** A description of the weather conditions during the time of the collision
- **LIGHTCOND:** The light conditions during the collision
- **ROADCOND:** The condition of the road during the collision
- **Etc.**

3. Methodology

Jupyter Notebook combined with Python 3 will be used in order to conduct this analysis. In addition, the following packages need to be imported into Python:

- Pandas
- Numpy
- Matplotlib
- Seaborn

```
# Importing the Relevant Packages

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

Prior to starting the analysis, the Seattle Police Department dataset also needs to imported using pandas. The function *.head()* allows us to gain a quick overview of the dataset.

```
# Importing the Dataset

df = pd.read_csv("Data-Collisions.csv", low_memory=False)
df.head()
```

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO
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

Once the dataset and the necessary packages have been imported, we may start taking a closer look at the dataset. The first step will be to use the *.describe()* and *.info()* function.

```
# Observing the Dataset

df.describe()
```

	SEVERITYCODE	X	Y
count	194673.000000	189339.000000	189339.000000
mean	1.298901	-122.330518	47.619543
std	0.457778	0.029976	0.056157
min	1.000000	-122.419091	47.495573
25%	1.000000	-122.348673	47.575956
50%	1.000000	-122.330224	47.615369
75%	2.000000	-122.311937	47.663664
max	2.000000	-122.238949	47.734142

```
# Observing the Dataset P2

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 38 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   SEVERITYCODE    194673 non-null  int64
 1   X               189339 non-null  float64
 2   Y               189339 non-null  float64
 3   OBJECTID        194673 non-null  int64
 4   INCKEY          194673 non-null  int64
 5   COLDETKEY       194673 non-null  int64
 6   REPORTNO        194673 non-null  object
 7   STATUS          194673 non-null  object
 8   ADDRTYPE        192747 non-null  object
 9   INTKEY          65070 non-null   float64
 10  LOCATION        191996 non-null  object
```

Upon a broad examination of the variables and their corresponding metrics and types, some variables distinguished themselves from the others. I have therefore decided to more carefully analyze those variables in the next section.

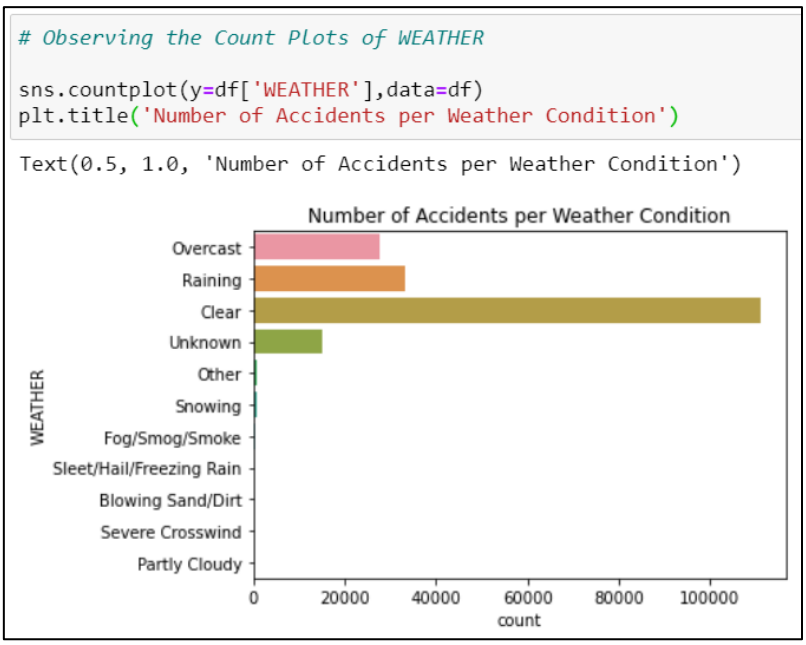
4. Results

First, I decided to take a closer look at the following attributes: WEATHER, LIGHTCOND, and ROADCOND. These attributes seemed especially pertinent with respect to my project's objective of anticipating and preventing road accidents. Using seaborn, I conducted the following graphical analysis:

```
# Observing the Count Plots of WEATHER

sns.countplot(y=df['WEATHER'],data=df)
plt.title('Number of Accidents per Weather Condition')
```

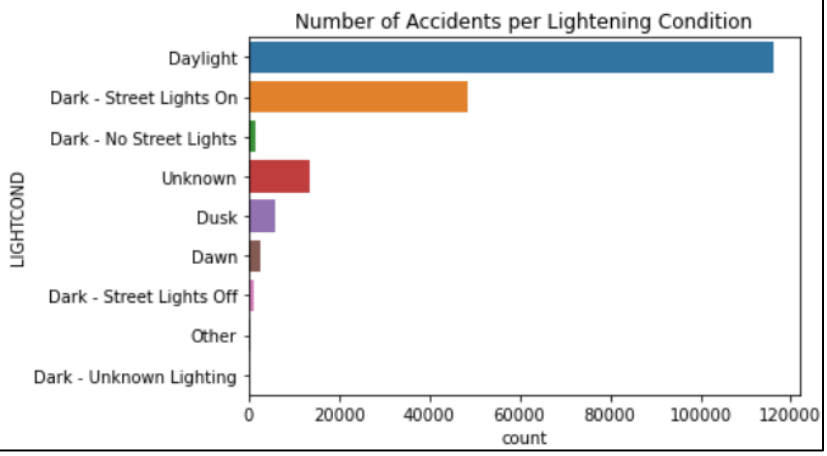
```
Text(0.5, 1.0, 'Number of Accidents per Weather Condition')
```



```
# Observing the Count Plots of LIGHTCOND

sns.countplot(y=df['LIGHTCOND'],data=df)
plt.title('Number of Accidents per Lightening Condition')
```

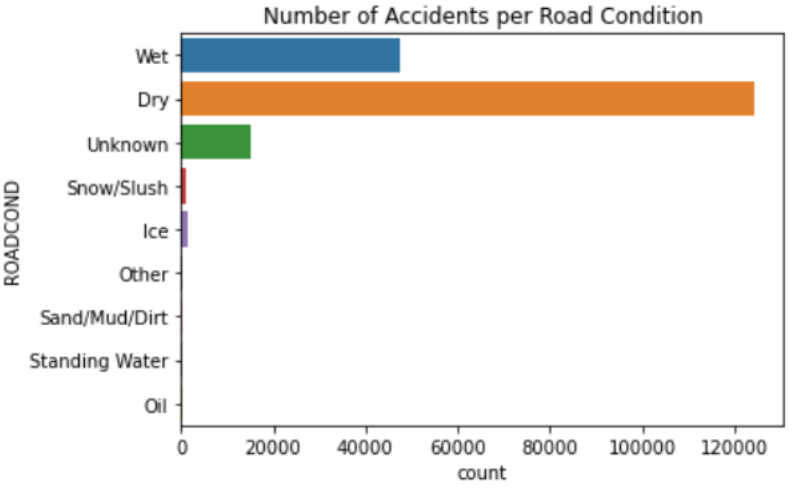
```
Text(0.5, 1.0, 'Number of Accidents per Lightening Condition')
```




```
# Observing the Count Plots of ROADCOND

sns.countplot(y=df['ROADCOND'],data=df)
plt.title('Number of Accidents per Road Condition')
```

```
Text(0.5, 1.0, 'Number of Accidents per Road Condition')
```



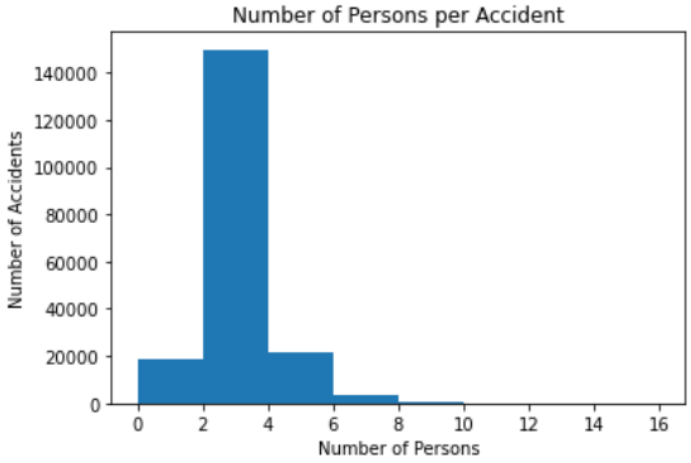
Based on these results, one may say that the overwhelming majority of accidents take place in (1) clear weather conditions, (2) during daylight, and (3) with dry road conditions. These results may seem surprising at first, though they most likely reflect the conditions in which most people are using motor vehicles, i.e., during the day on normal roads.

Secondly, using Matplotlib, I decided to check the number of persons and vehicles in most accidents:

```
# Observing the Number of Persons per Accident

bins = np.arange(df.PERSONCOUNT.min(),17,2)
plt.hist(df.PERSONCOUNT,bins=bins)
plt.xlabel('Number of Persons')
plt.ylabel('Number of Accidents')
plt.title('Number of Persons per Accident')
```

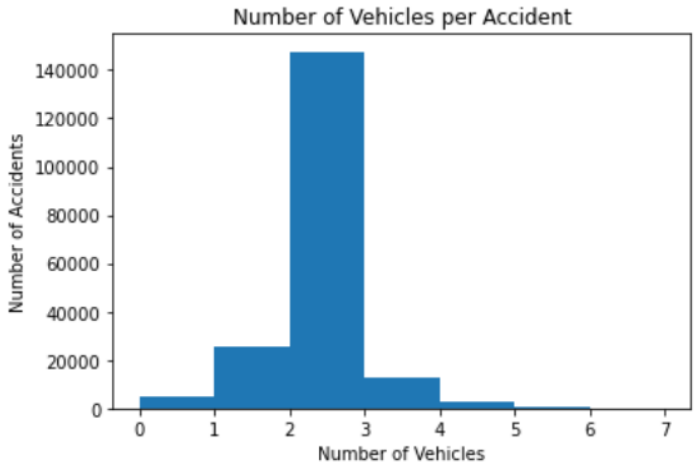
```
Text(0.5, 1.0, 'Number of Persons per Accident')
```




```
# Observing the Number of Vehicles per Accident

bins = np.arange(df.VEHCOUNT.min(),8,1)
plt.hist(df.VEHCOUNT,bins=bins)
plt.xlabel('Number of Vehicles')
plt.ylabel('Number of Accidents')
plt.title('Number of Vehicles per Accident')
```

```
Text(0.5, 1.0, 'Number of Vehicles per Accident')
```



As one may observe, most accidents included 2-4 persons situated in 2-3 vehicles. This implies that most accidents involve only one passenger per vehicle. This could perhaps be due to the fact that solo drivers are more prone to speeding or distractions.

5. Discussion

This analysis strived to examine some various factors that federal bodies could monitor in order to predict road accidents. From the analysis, it is clear that most accidents involve solo drivers on relatively normal weather, lightiening, and road conditions.

This information could be especially useful to the police department in understanding where to install more stop signs or cameras. For instance, they could focus on areas that are known to have a significant proportion of solo drivers, who probably commute for work purposes during the day.

6. Conclusion

Overall, although this analysis has provided some valuable insights, a closer inspection of other variables needs to be undertaken. In particular, further analysis would need to be undertaken in order to perhaps come up with a comprehensive model that would precisely predict the expected number of accidents based on a number of factors.

Nonetheless, this report's broad approach to this issue already enables to draw a number of conclusions. In particular, it seems like most accidents are minor and avoidable. Indeed, contrarily to what one may think, most accidents occur on good weather, lightening, and road conditions. This result goes against the pre-conceived idea that environmental factors are driving road accidents. On the contrary, this would rather point to the idea that the human element seems to stand at the core of the overwhelming majority of road accidents.