Introduction

The fact that car accidents cause serious loss of life and property today, where the rate of use of cars is increasing, encourages governments to take more various measures. At this point, determining the factors that may have an impact on survival in car accidents can be critical in terms of preventing car accidents by using methods that warn citizens about traffic accidents and the health system and the police.

As a result of a study conducted by the World Health Organization, it has been identified that car speed is an important risk factor for injuries. In addition, it was found that the speed of the driver depends on factors such as age, gender, alcohol level, surface quality and vehicle power.

The target audience of this project is the local Seattle government, police organization, health system and institutions such as car insurance. The model and its results will provide the government with meaningful data to reduce car accidents and related injuries in the city, and advise the target audience to make meaningful decisions.

Data

The data has been provided by the Seattle Police Department since 2004 and consists of 37 independent variables and 194.673 lines.

In order to find out which factors affected the injuries in accidents, I chose to use in the data set the columns "SEVERITYCODE" which is the dependent variable and defines the severity of the collision and "WEATHER" which defines the weather at the time of the collision, "ROADCOND", "LIGHTCOND" which describes the road condition and light conditions and "SPEEDING", which describes the speed of the driver.

"SEVERITYCODE" contains numbers from 0 to 4 corresponding to different severity levels of the accident as:

0: Little to no injured.

1: Very Low injured.

2: Low injured.

3: Mild injured.

4: High injured.

Methodology

There are too many variables in our original data set that we will not be using. First of all, we will need to separate the most important variables that will affect the dependent variable "SEVERITYCODE" from the data set and categorize these variables.

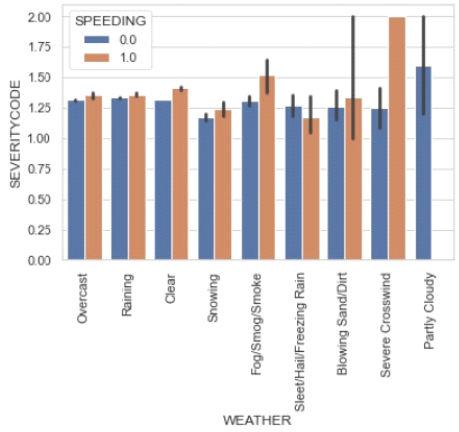
```
df_var = df[['WEATHER', 'ROADCOND', 'LIGHTCOND', 'SPEEDING', 'SEVERITYCODE']]
df_var.head()
```

	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	SEVERITYCODE
0	Overcast	Wet	Daylight	NaN	2
1	Raining	Wet	Dark - Street Lights On	NaN	1
2	Overcast	Dry	Daylight	NaN	1
3	Clear	Dry	Daylight	NaN	1
4	Raining	Wet	Daylight	NaN	2

Now let's do some visualization and look at the effects of weather and driver speed on "SEVERITYCODE".

```
ax = sns.barplot(x="WEATHER", y="SEVERITYCODE", hue="SPEEDING", data=df_var)
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
```

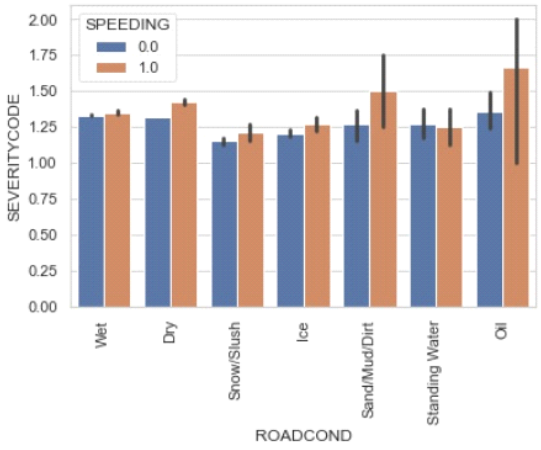
```
[Text(0, 0, 'Overcast'),
 Text(1, 0, 'Raining'),
 Text(2, 0, 'Clear'),
 Text(3, 0, 'Snowing'),
 Text(4, 0, 'Fog/Smog/Smoke'),
 Text(5, 0, 'Sleet/Hail/Freezing Rain'),
 Text(6, 0, 'Blowing Sand/Dirt'),
 Text(7, 0, 'Severe Crosswind'),
 Text(8, 0, 'Partly Cloudy')]
```



Likewise, let's see how much ground condition and driver speed affect "SEVERITYCODE".

```
ax = sns.barplot(x="ROADCOND", y="SEVERITYCODE", hue="SPEEDING", data=df_var)
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
```

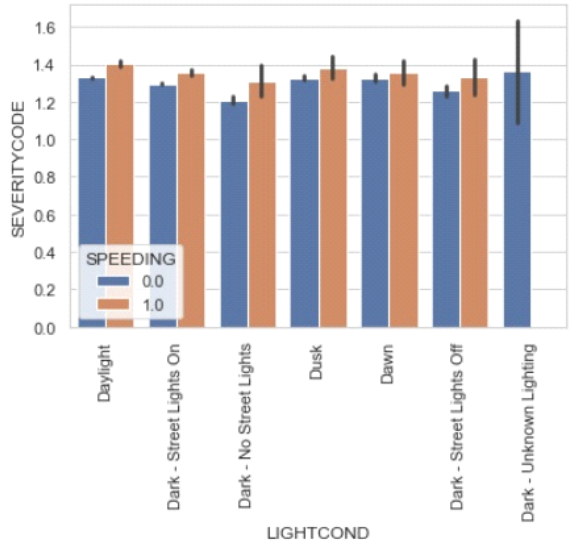
```
[Text(0, 0, 'Wet'),
 Text(1, 0, 'Dry'),
 Text(2, 0, 'Snow/Slush'),
 Text(3, 0, 'Ice'),
 Text(4, 0, 'Sand/Mud/Dirt'),
 Text(5, 0, 'Standing Water'),
 Text(6, 0, 'Oil')]
```



And likewise we can look at the effects of ambient light conditions and driver speed.

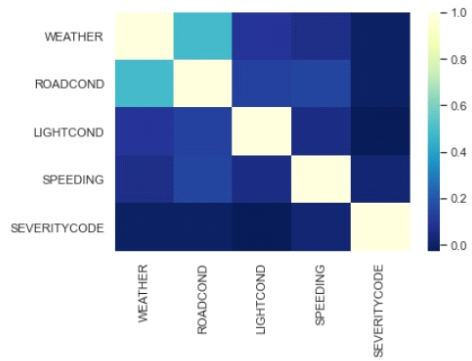
```
ax = sns.barplot(x="LIGHTCOND", y="SEVERITYCODE", hue="SPEEDING", data=df_var)
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
```

```
[Text(0, 0, 'Daylight'),
 Text(1, 0, 'Dark - Street Lights On'),
 Text(2, 0, 'Dark - No Street Lights'),
 Text(3, 0, 'Dusk'),
 Text(4, 0, 'Dawn'),
 Text(5, 0, 'Dark - Street Lights Off'),
 Text(6, 0, 'Dark - Unknown Lighting')]
```



```
sns.heatmap(df_var.corr(), cmap='YlGnBu_r')
```

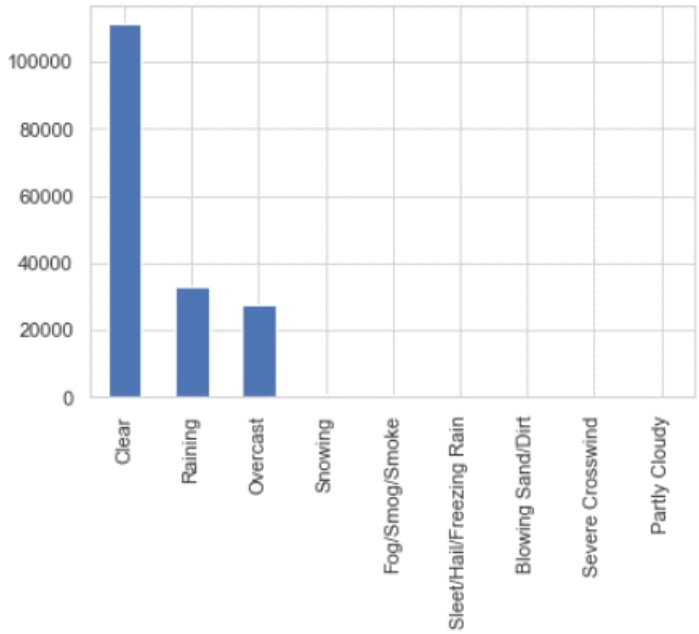
```
<AxesSubplot:>
```



As can be seen, using the seaborn library, we can visualize the effects of several variables on the dependent variable in this way and examine their effects. Apart from that, with the matplotlib library, we can find the reasons for the accidents by accessing the numbers of accidents that occurred in various conditions.

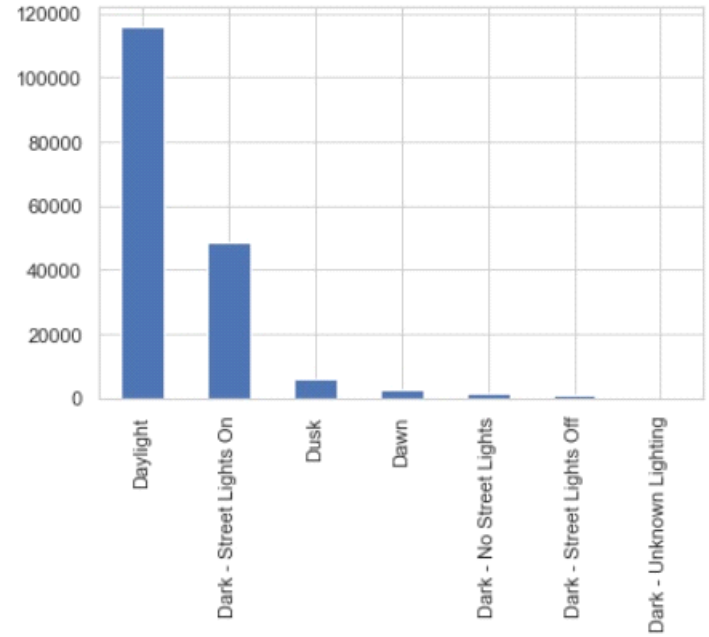
```
df_var["WEATHER"].value_counts().plot(kind = "bar")
```

```
<AxesSubplot:>
```



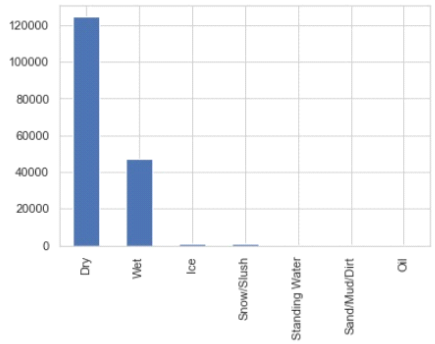
```
df_var["LIGHTCOND"].value_counts().plot(kind = "bar")
```

```
<AxesSubplot:>
```




```
df_var["ROADCOND"].value_counts().plot(kind = "bar")
```

```
<AxesSubplot:>
```



As a result, it would not be surprising that many of the accidents occur during the daytime, on dry ground and outdoors, when the use of vehicles is higher. At this point, looking at the second conditions where accidents occur most can help us to make a healthier decision, and we can say that most of the accidents happen in the rainy, wet and evening hours.

Before moving on to machine learning applications, arranging our data set to be suitable for work will help us get healthier results. Let's try to categorize variables and remove missing values from our data set.

```
df_var.dropna(inplace=True)
```

```
C:\Users\eevli\anaconda3\lib\site-packages\ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

```
weather_map = {"Clear": 0, "Raining": 1, "Overcast":2, "Snowing":3,
               "Fog/Smog/Smoke":4, "Sleet/Hail/Freezing Rain":5,
               "Blowing Sand/Dirt":6, "Severe Crosswind":7, "Partly Cloudy":8}
df_var["WEATHER"] = df_var["WEATHER"].map(weather_map)
df_var["WEATHER"] = df_var["WEATHER"].astype("int64")
```

After that, we can create our models.

```
from sklearn import preprocessing

x = df_var.drop(["SEVERITYCODE"], axis=1)
y = df_var[["SEVERITYCODE"]]
df_var_scaled = preprocessing.StandardScaler().fit(x).transform(x)
df_var_scaled[0:3]
```

```
array([[ 1.79168786,  1.30988802, -0.56684089, -0.23764147],
       [ 0.56237301,  1.30988802,  0.69194391, -0.23764147],
       [ 1.79168786, -0.57885829, -0.56684089, -0.23764147]])
```

```
from sklearn.metrics import classification_report, roc_auc_score
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(df_var_scaled, y,
                                                    test_size=0.2, random_state=42)
```

Results & Conclusion

	Model	Jaccard	F1
0	KNN	0.662439	0.419013
1	Tree	0.670933	0.401979
2	Log Reg	0.671011	0.401560

It can be concluded that certain classes, such as weather conditions, road condition, driver speed, have some effect on injury in vehicle accidents under certain conditions. Our data set originally contained some variables and classes as objects, and we categorized them to apply our model on it. After getting rid of our lost values, we applied three machine learning algorithms, KNN, Decision Tree, and Logistic regression, on our model. The evaluation criteria used to test the accuracy of our models were the jaccard index and the f-1 score.