Capstone Project: Full Report: Car Accident Severity

# Introduction

Road Accidents are the most undesirable and unexpected thing to occur to a road user, though they happen quite often. Unfortunately, we can see a minatory rise of road accidents, conspicuously high road accidents over the past few years. It has a massive impact on society as well as in the economy of the country as there is an immense cost of fatalities and injuries. According to a recent report, annually on an average 9,000 lives have been taken by road accidents and lead to almost 28,000 injuries. This record indicates the idea to leverage the accident data that you found to predict the different accidents' severity and come up with solutions which can be beneficial to the government. Besides this, according to WHO, the economic cost of road accidents to a developing country like us is 2-3% of GDP, which is a significant loss for a country like ours. Moreover, reducing this loss has become a great matter of concern for our country now.

Traffic collisions continue to be a serious problem. According to the annual collision report for 2015 released by the Washington State Dept. of Transportation (WSDOT), there were 117,053 collisions and approximately 59.7 billion miles driven. In 2015, there were

* 5,576,586 Licensed drivers
* 6,252,554 Registered vehicles
* 416,699 Speeding citations
* 33,697 Cell phone and texting citations

# Problem Statement

Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police cars start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to. Well, this is exactly what you will be working on in this course. So the problem statement here would be to come up with a project which can predict the severity of an accident.

TRAFFIC REPORT is based on data from 2004 to Present. The report states that "Traffic volumes, speeds, and reported collisions are the three cardinal pieces of data traffic engineers and planners use to evaluate changes to Seattle streets.

The ultimate project goals include:

* Identify dangerous locations
* Identify predictors of accidents (e.g., physical characteristics of the location, road condition, DUI, weather)
* Examine increase or decrease in number of accidents over time
* Identify predictors of increase or decrease in the number of accidents
* Recommend improvements on dangerous locations

However, as you explore the dataset, other questions are likely to pop up in your head.

We recommend making a note of those questions as they may guide you through your project. You are welcome to answer any questions you are interested in using this dataset.

# Data Description

For the accurate prediction of the severity of accidents, a considerable number of traffic accident records with full information is required to train by using the proposed approaches. In this research work, the authors have collected a dataset from the Traffic Bureau that consists of total 37,885 traffic accidents recorded from the year 2007-2017. The entire dataset will split into two parts- Training Dataset and Test Dataset. 70% of the whole dataset has been chosen randomly by using a python library as a training data set and the remaining 30% has been used as our test dataset. We have used the 70-30 ratio for splitting dataset because of its proven accuracy.

## About the original column variables¶

X - longitude, the GPS values moving left to right (East and West) along the X axis

Y - latitude, represented by horizontal lines, which go up and down (North and South)

OBJECTID - ESRI unique identifier

INCKEY - a unique key for the incident, variable type: Long

COLDETKEY - a secondary key for the incident, variable type: Long

REPORTNO - unknown

STATUS - unknown

ADDRTYPE - Collision address location type, variable type: text, 12 VARCHAR, e.g. Alley, Block, Intersection

INTKEY - a key that corresponds to the intersection associated with a collision, variable type: Double

LOCATION - a text description of location, e.g. TERRY AVE BETWEEN JAMES ST AND CHERRY ST

EXCEPTRSNCODE - unknown

EXCEPTRSNDESC - unknown

## **SEVERITYCODE - a code that corresponds to the severity of the collision:**

**3 — fatality**

**2b — serious injury**

**2 — injury**

**1 — prop damage**

**0 — unknown**

SEVERITYDESC - a description of the collision, e.g. Property Damage Only Collision, Injury Collision

COLLISIONTYPE - a description of the collision type, e.g. Parked Car, Rear Ended, Sideswipe

PERSONCOUNT - the total number of people involved

PEDCOUNT - the total number of pedestrians involved

PEDCYLCOUNT - the total number of cyclists involved

VEHCOUNT - the total number of vehicles involved

INJURIES - the total number of injuries other than fatal or disabling at the scene, including broken fingers or toes, abrasions, etc.

SERIOUSINJURIES - total number of injuries that result in at least a temporary impairment, e.g. a broken limb. It does not mean that the collision resulted in a permanent disability

FATALITIES - includes the total number of persons who died at the scene of the collisions, were dead on arrival at the hospital, or died within 30 days of the collision from collision-related injuries

INCDATE - incident date

INCDTTM - date and time of the incident, variable type: text, 30 VARCHAR

JUNCTIONTYPE - category of the junction where the collision took place

SDOT\_COLCODE - the SDOT collision code

SDOT\_COLDESC - a description of the collision corresponding to the collision code

INATTENTIONIND - whether or not collision was due to inattention. (Y/N)

UNDERINFL - whether or not the driver was under the influence of alcohol or drugs

WEATHER - a description of the weather, e.g. Raining, Clear

ROADCOND - a description of the road conditions, e.g. Dry, Wet

LIGHTCOND - a description of the light conditions, e.g. Dark - No Street Lights, Daylight

PEDROWNOTGRNT - whether or not the pedestrian right of way was not granted. (Y/N)

SDOTCOLNUM - unknown

SPEEDING - whether or not the driver was speeding

\*ST\_COLCODE - code provided by the state that describes the collision, for example:

0 - Vehicle Going Straight Hits Pedestrian

1 - Vehicle Turning Right Hits Pedestrian

2 - Vehicle Turning Left Hits Pedestrian

3 - Vehicle Backing Hits Pedestrian

4 - Vehicle Hits Pedestrian - All Other Actions

5 - Vehicle Hits Pedestrian - Actions Not Stated

10 - Entering At Angle

11 - From Same Direction -Both Going Straight-Both Moving- Sideswipe

12 - From Same Direction -Both Going Straight-One Stopped- Sideswipe

13 - From Same Direction - Both Going Straight - Both Moving - Rear End

14 - From Same Direction - Both Going Straight - One Stopped - Rear End

15 - From Same Direction - One Left Turn - One Straight

16 - From Same Direction - One Right Turn - One Straight

19 - One Car Entering Parked Position

20 - One Car Leaving Parked Position

21 - One Car Entering Driveway Access

22 - One Car Leaving Driveway Access

23 - From Same Direction - All Others

24 - From Opposite Direction - Both Moving - Head On

25 - From Opposite Direction - One Stopped - Head On

26 - From Opposite Direction - Both Going Straight - sideswipe

27 - From Opposite Direction - Both Going Straight - One Stopped - sideswipe

28 - From Opposite Direction - One Left Turn - One Straight

29 - From Opposite Direction - One Left Turn - One Right Turn

30 - From Opposite Direction - All Others

31 - Not Stated

32 - One Parked - One Moving

40 - Train Struck Moving Vehicle

41 - Train Struck Stopped or Stalled Vehicle

42 - Vehicle Struck Moving Train

43 - Vehicle Struck Stopped Train

44 - Unicycle

45 - Bicycle

46 - Tricycle

47 - Domestic Animal (horse, cow, sheep, etc)

48 - Domestic Animal Other (Cat, Dog etc)

49 - Non Domestic Animal (deer, bear, elk, etc)

50 - Struck Fixed Object 51 - Struck Other Object

52 - Vehicle Overturned

53 - Person Fell, Jumped, or was Pushed From Vehicle

54 - Fire Started In Vehicle

55 - Accidently Overcame By Carbon Monoxide Poison

56 - Breakage Of Any Part Of the Vehicle Resulting In Injury or in Further Property Damage

57 - All Other Non-Collisions

60 - Vehicle Hits State Road or Construction Machinery

61 - Vehicle Struck By State Road or Construction Machinery

62 - Vehicle Hits County Road or Construction Machinery

63 - Vehicle Struck By County Road or Construction Machinery

64 - Vehicle Hits City Road or Construction Machinery

65 - Vehicle Struck By City Road or Construction Machinery

66 - Vehicle Hits Other Road or Construction Machinery

67 - Vehicle Struck by Other Road or Construction Machinery

71 - Same Direction - Both Turning Right - Both Moving - Sideswipe

72 - Same Direction - Both Turning Right - One Stopped - Sideswipe

73 - Same Direction - Both Turning Right - Both Moving - Rear End

74 - Same Direction - Both Turning Right - One Stopped - Rear End

81 - Same Direction - Both Turning Left - Both Moving - Sideswipe

82 - Same Direction - Both Turning Left - One Stopped - Sideswipe

83 - Same Direction - Both Turning Left - Both Moving - Rear End

84 - Same Direction - Both Turning Left - One Stopped - Rear End \*

ST\_COLDESC - a description that corresponds to the state’s coding designation

SEGLANEKEY - a key for the lane segment in which the collision occurred

CROSSWALKKEY - a key for the crosswalk at which the collision occurred

# **Kind of variables:**

## Numeric variables¶

Although there are many columns that contain numerical data, many of those columns use numbers for labels or identification (aka key values).

These numerical variables provide information about what each accident looked like:

PERSONCOUNT (# of people invovled) PEDCOUNT (# of pedestirans) PEDCYLCOUNT (# of cyclists) VEHCOUNT (# of vehicles) INJURIES (# of injuries) SERIOUSINJURIES (# of serious injuries) FATALITIES (# of deaths)

## Non-numeric columns of interest¶

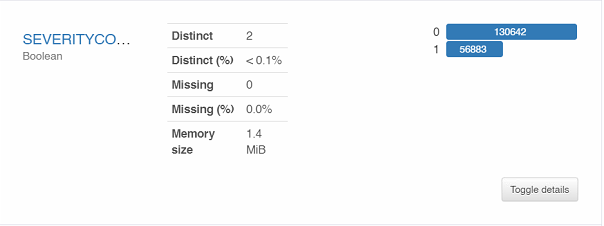
These string and categorical variables provide information about the situation or environment in which each accident occurred and could be used as predictors of accidents.

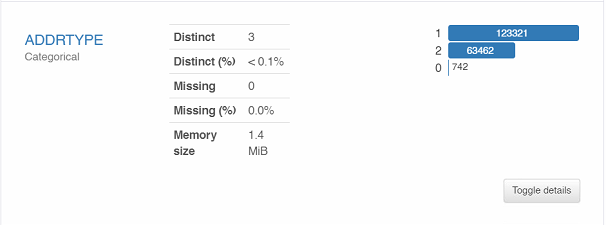
External factors (i.e., not dependent on the characteristics of the collision site)¶ LIGHTCOND (light condition) WEATHER (description of the weather conditions during the time of the collision)

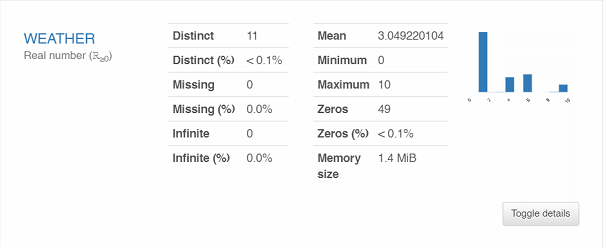
# **Profile: Overview**

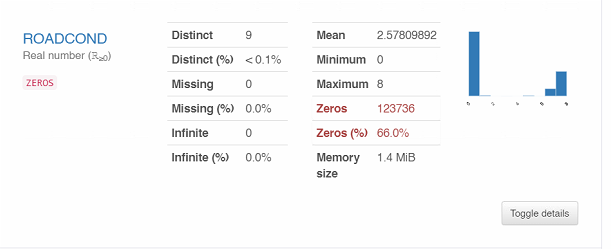
|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset statistics** | | **Variable types** | |
| Number of variables | 5 | NUM | 3 |
| Number of observations | 187525 | CAT | 1 |
| Missing cells | 0 | BOOL | 1 |
| Missing cells (%) | 0.0% |  |  |
| Duplicate rows | 186580 |  |  |
| Duplicate rows (%) | 99.5% |  |  |
| Total size in memory | 7.2 MiB |  |  |
| Average record size in memory | 40.0 B |  |  |

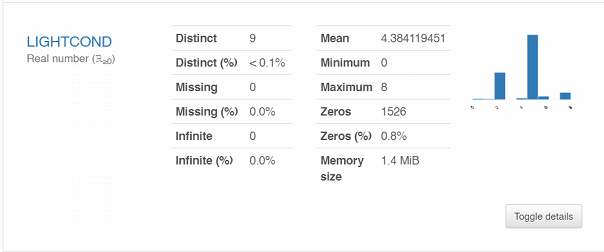
## Variable

****

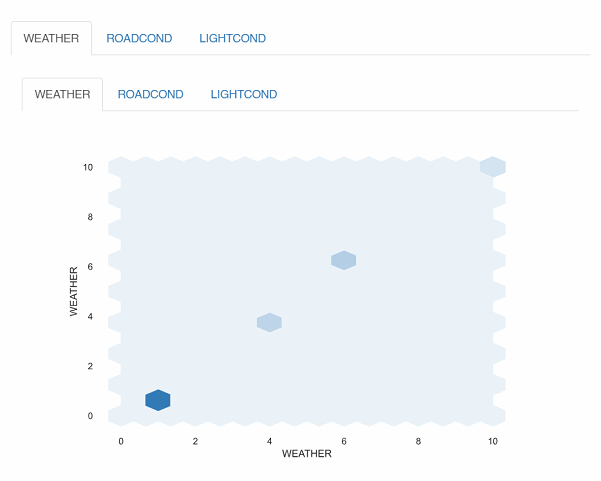
****

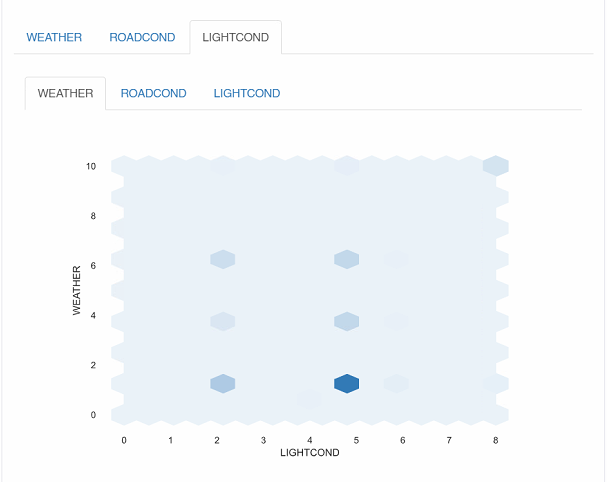
****

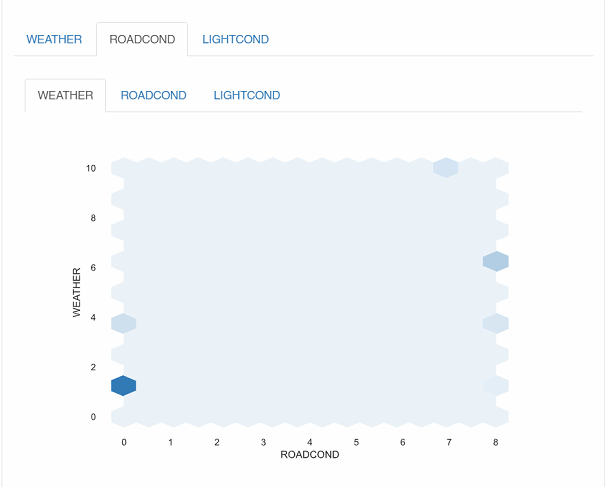
****

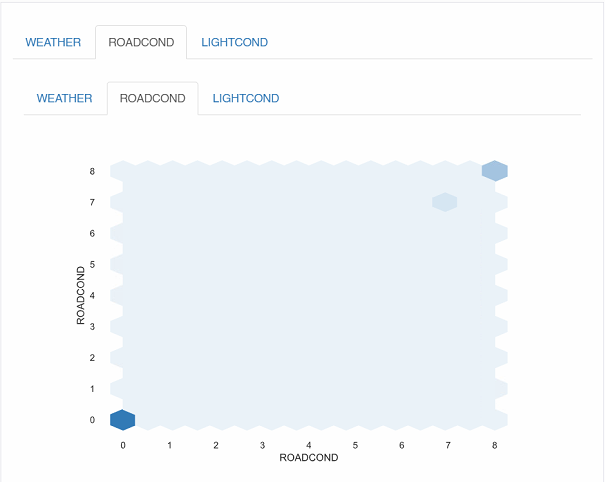
****

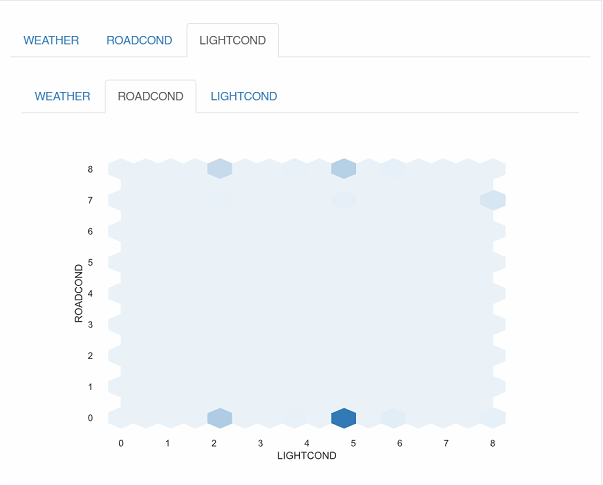
## Interactions:

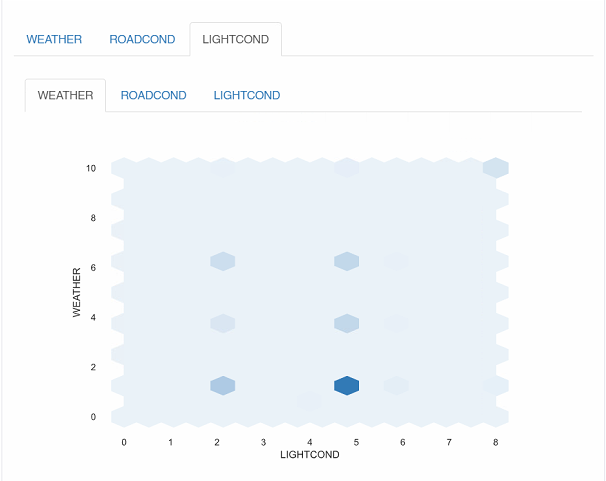
****

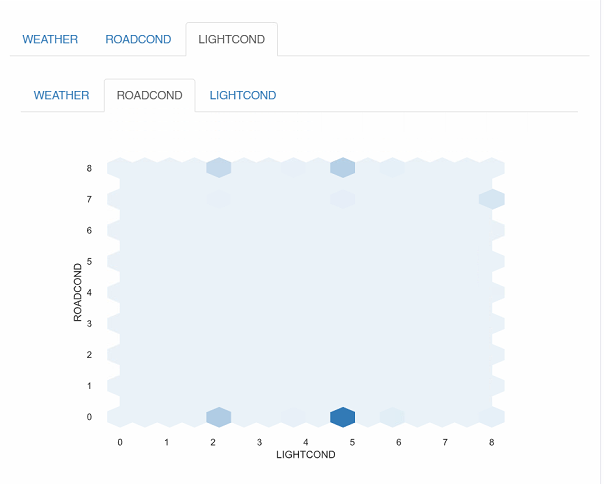
****

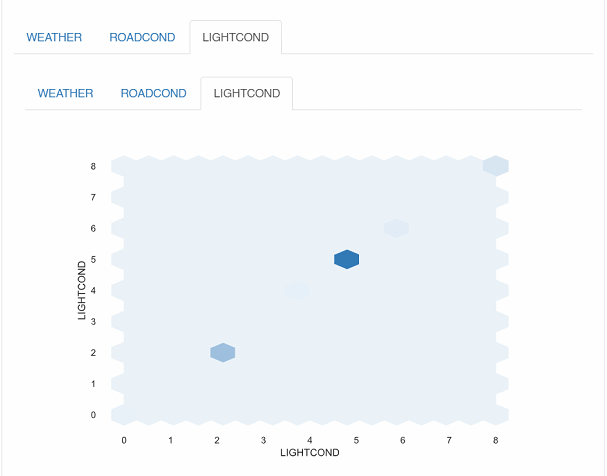
****

****

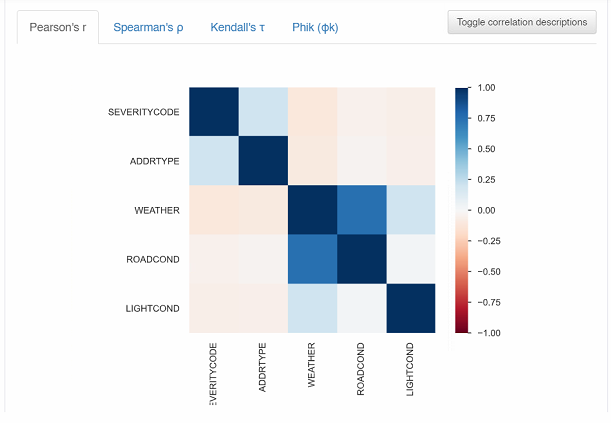
****

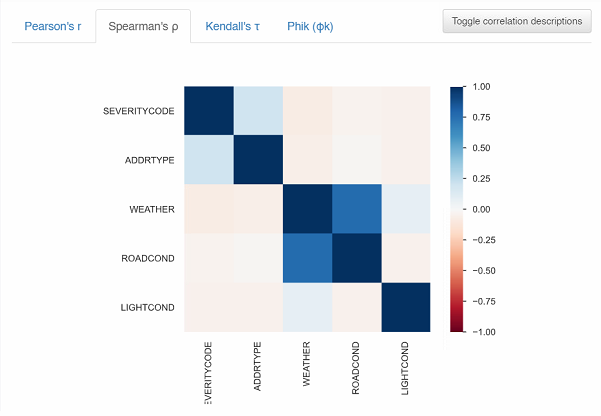
****

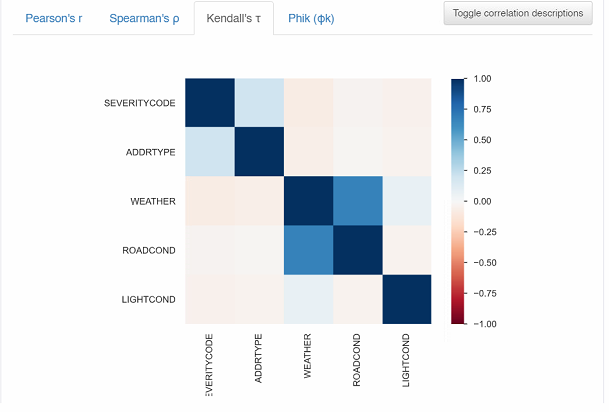
****

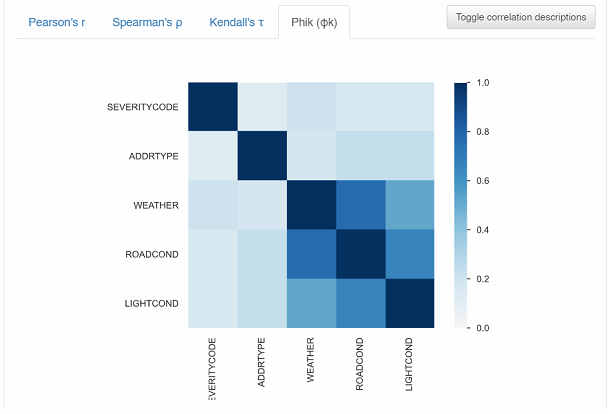
****

## Correlations:

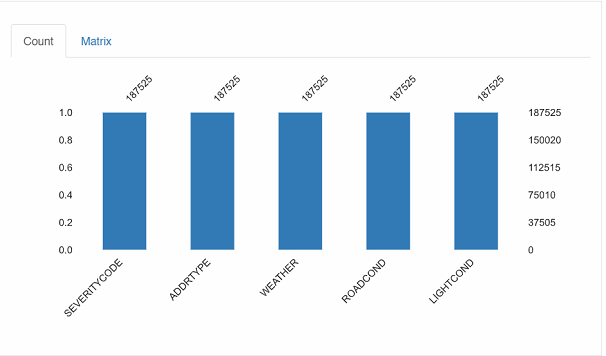
****

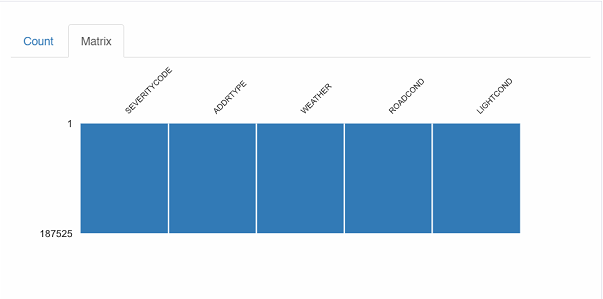
****

****

****

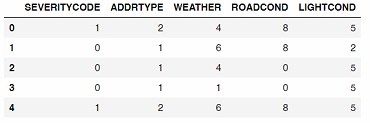
## Missing values:

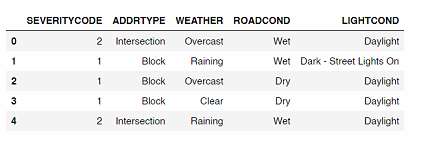
****

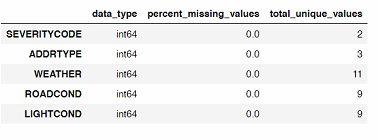
****

## Df.shape

## Overall data 194673 with 38 column

****

****

****

# **Understand the dataframe in relation to the questions at hand**

## Numeric variables

Although there are many columns that contain numerical data, many of those columns use numbers for labels or identification (aka key values).

These numerical variables provide information about what each accident looked like:

PERSONCOUNT (# of people invovled)

PEDCOUNT (# of pedestirans)

PEDCYLCOUNT (# of cyclists)

VEHCOUNT (# of vehicles)

INJURIES (# of injuries)

SERIOUSINJURIES (# of serious injuries)

FATALITIES (# of deaths)

## Non-numeric columns of interest

These string and categorical variables provide information about the situation or environment in which each accident occurred and could be used as predictors of accidents.

External factors (i.e., not dependent on the characteristics of the collision site)

LIGHTCOND (light condition)

WEATHER (description of the weather conditions during the time of the collision)

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 187525 entries, 0 to 187524

Data columns (total 5 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 SEVERITYCODE 187525 non-null int64

1 ADDRTYPE 187525 non-null object

2 WEATHER 187525 non-null object

3 ROADCOND 187525 non-null object

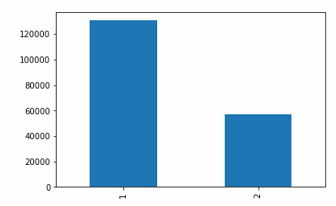
4 LIGHTCOND 187525 non-null object

dtypes: int64(1), object(4)

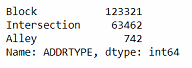
memory usage: 7.2+ MB

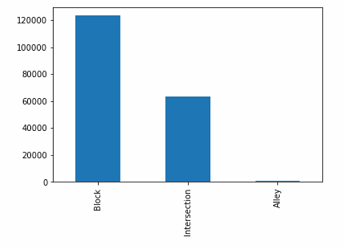
Out (13)



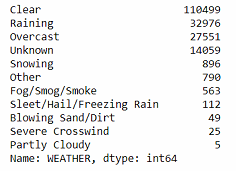


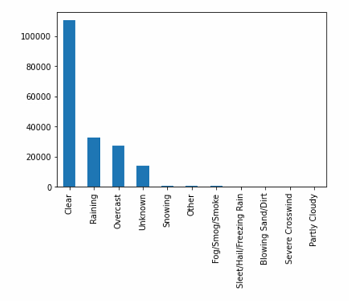
Out (16)



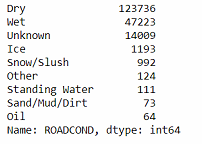


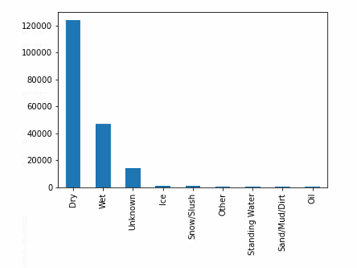
Out (19)



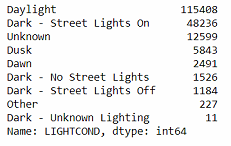


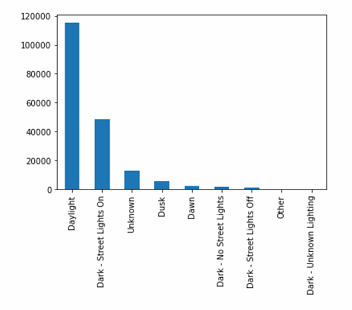
Out (22)



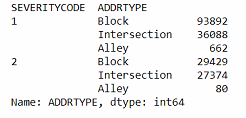


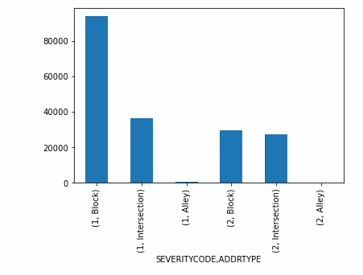
Out (25)



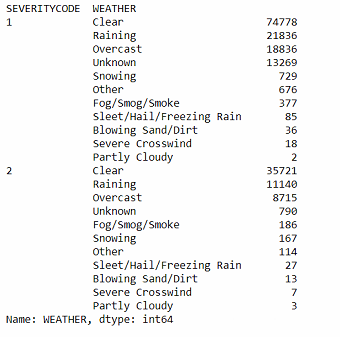


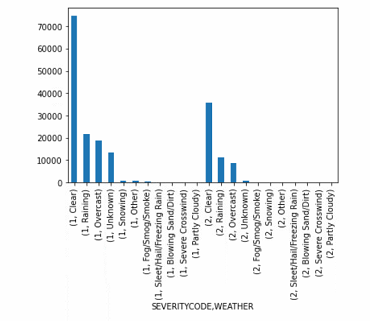
Out (28)



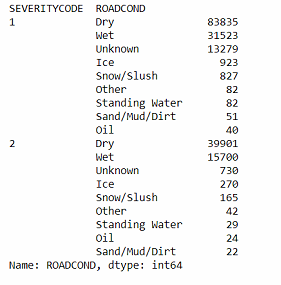


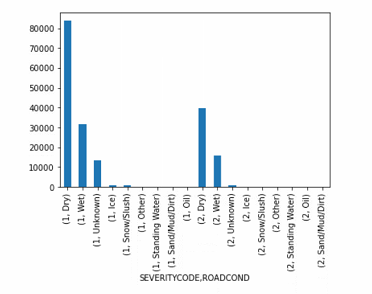
Out (30)



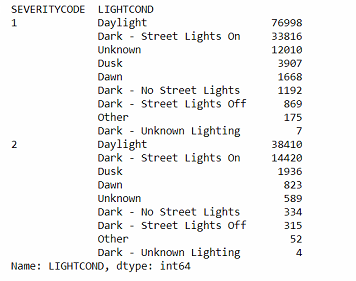


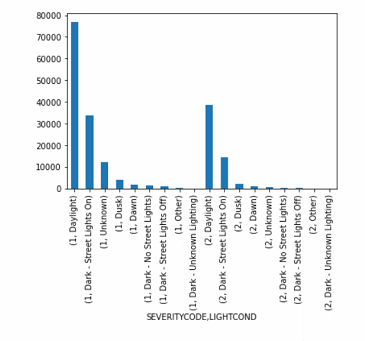
Out (32)





Out (34)





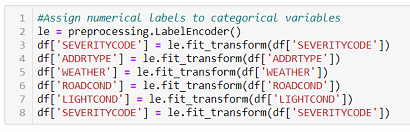
# **Methodology**

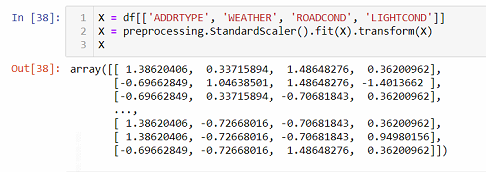
In the first step we properly loaded the data and kept only the variables we need for the analysis, which are the variable we are trying to predict (SEVERITYCODE) and the predictor variables (ADDRTYPE, WEATHER, ROADCOND, LIGHTCOND). We then explored the different types of values within each variable and plotted them on a bar graph. After removing unknown entries in the data, we then plotted each predictor variable value and grouped them by severity code. This gives a visual idea of what factors seemed to be more common in the different types of accidents.

In the second step we will prepare our data for model building. Since all the variables are categorical, we will label encode them to produce numerical labels. We will then randomly split our data into training and test sets for our model.

In the third and final step we will fit our data into different models and evaluate them to see which produces the highest accuracy. Since we want to use machine learning models that can be used to predict a certain class or group based on given conditions, the models we fit will be SVM (Support Vector Machines), K-Nearest Neighbors, and Decision Trees. After fitting each model, we will calculate its accuracy using various methods such as f1-score, jaccard similarity score, and classification report.

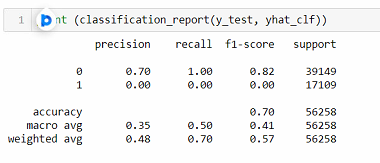
**In (64)**

****

****

**Modeling & Evaluation**

**In (56)**

****

# **Results and Discussion**

Through our exploratory analysis we see that the majority of accidents from this dataset resulted in prop damage. In all accident groups, it appears that the accident occurred the most during daylight, when the weather was clear, and road condition was dry. The second most likely conditions that cause accidents are rainy weather, wet road conditions, and being in the dark with street lights on. Most of the accidents that resulted in prop damage occurred at a block, while accidents that resulted in injury occurred almost equally at blocks and intersections. It seems that accidents that resulted in fatality only occurred at blocks and intersections.

The four models we built are all very similar in terms of prediction and accuracy. However, overall it seems that the model accuracy for all models can be greatly improved. The highest prediction accuracy is only around 70%. The classification reports for all the models (except K-Nearest Neighbors) also show that the model could only predict accidents with prop damage. This could likely be due to eliminating many variables that could be more significant predictors or having significantly more data for prop damage accidents than other accidents. We only kept the "environmental" variables so our model had very limited data and predictors. It is pretty difficult to choose the most accurate model since the test set accuracy is the same for the SVM, and Decision Tree models. However, the Decision Tree train set accuracy is slightly higher so we can say that this model is more accurate.

In terms of recommendations, based on our observations and analysis, we should pay more attention on the road when faced with conditions that do not seem to bring much risk. Road signs and warnings should be put up to caution drivers and pedestrians, especially at blocks and intersections. More caution and speed limits should be enforced during rainy and wet conditions since they are the second lead cause of car accidents (in this model). Installing signs and lights that light up in the dark will be very helpful in cautioning drivers at night. If we want to predict the severity of car accidents solely based on these environmental and road factors, the decision tree would be a good model to use.

# **Conclusion:**

In this project, I focused on finding the major environmental factors and road conditions that affect car accidents, as well as building a model that can help predict the severity of car accidents based on these conditions. I cleaned our data and prepared it for exploratory data analysis and model building. I fit four machine learning models on our data and determined which produces the most accurate predictions. Based on our analysis and results, we made some recommendations to improve the safety of drivers and those on the road during certain road and weather conditions. Finally, we suggested a model that produces the best results for further analysis of car accident severity based on the same predictors.