IBM Data Science Certificate Program

Applied Data Science Capstone

Car Accident Severity Analysis

Zhao Zhang

Date: 10/19/2020

Contents

1. Introduction

1.1 Background

Motor vehicle crashes in the United States cost $242 billion in economic damage and 32,999 deaths in 2010. Furthermore, there were 3.9 million injuries and 24 million damaged vehicles (NHTSA, 2015). Property damage and injury-related traffic accidents are a socially and economically disruptive force that often sinks into the back of our minds. Factors such as medical expense, productivity loss, and administrative costs amount to approximately $871 billion damage in societal value (NHTSA, 2015). This astonishing figure is approximately twice the damage done by Hurricane Katrina, the most costly hurricane in U.S history. However, unlike Hurricane Katrina, road accidents are more easily managed through human effort. For instance, the degree of human error and traffic environment, which increase the likelihood of accident occurrence and the magnitude of damage, can be optimized through education and increased investment in infrastructure.

1.2 Understanding the Problem

This project examines the traffic collision data in Seattle, Washington from 2004 to the present. Through exploratory data analysis and machine learning techniques, it aims to provide some insight in aspects such as traffic management, road infrastructure, and importance of traffic safety. Specifically, the project focuses on two crucial aspects of accidents: human error and environment. Factors such as inattention, rule-breaking actions (e.g., exceeding speed limit) and substance abuse constitute the subjective reason of an accident. On the other hand, natural environment

(e.g., weather) and infrastructure (e.g., road and light conditions) form objective reasons for an accident. Gaining perspective on these issues can help the relevant stakeholders develop scientific solutions and minimize the harm done on the road.

1.3 Stakeholders

The main stakeholders of this report are governmental agencies in charge of city planning/civil engineering, transportation, traffic control, and road maintenance. Predictive machine learning models can help identify weak spots that are more susceptible to severe accidents based on real-time data. In turn, short-term tactics such as more stringent traffic regulations and lower speed limits can be placed on high-risk areas. Furthermore, in the long term, maintenance plans and infrastructure investments can be implemented in a more clinical, targeted, and effective approach. In addition to public organizations, individual drivers can also benefit from this report from an educational perspective. There are only benefits from further understanding the harm of irresponsible driving and the unnecessary risks of driving in harsh conditions.

2. Data

2.1 Data Source

The dataset contains weekly-updated collisions data in Seattle, Washington, USA from 2004 to Present (October 2020). The employed open-source government dataset is collected and published by Seattle Department of Transport (SDOT). The original dataset can be found here.The metadata information is released by the SDOT Traffic Management Division. Metadata information can be found here.

The raw dataset contains 194673 entries and 38 columns, representing 194673 recorded collisions and a wide range of attributes of the accidents. See **Appendix 1.** for the basic descriptive statistics and data information of the raw dataset.

2.2 Feature Selection

The selected feature set of variables are shown in Table 1.

Column Name	Description
ROADCOND	The road condition at the time of the accident – e.g., wet, dry
LIGHTCOND	The light condition at the time of the accident – e.g., dark, daylight
WEATHER	The weather at the time of the accident – e.g., clean, rain
UNDERINFL	Whether the driver is under influence – Yes or No
SPEEDING	Whether the driver was speeding – Yes or No
INATTENTIONIND	Whether the driver was paying attention – Yes or No
ADDRTYPE	The address type of the accident – e.g., alley, mid-block
JUNCTIONTYPE	The junction type at the accident location -e.g., mid-block
COLLISIONTYPE	The type of collision that took place – e.g., head-on, rear-ended
PEDCOUNT	The number of pedestrians in the accident -e.g., 3
PEDCYLCOUNT	The number of cycling pedestrians in the accident -e.g., 1
VEHCOUNT	The number of vehicles in the accident -e.g., 2
SEVERITYCODE	The type of damage done in the accident – i.e., property damage, injury

Table 1. Feature Selection and Description

The selected features can be categorized into surrounding environment variables (including physical location of the accident and natural environment) and driver discretion. Other attributes in the raw dataset are neglected for a variety of reasons. For instance, SDOTCOLNUM, a number assigned to the incident, is disregarded because it is irrelevant to the objectives of the project. The employed features provide universal factors in a car accident. Moreover, almost all of the information can be obtained through first-hand report, hence offering value to first responders and emergency service agencies. Therefore, the project aims to provide appliable insight appliable to other areas, even though the study is solely done on the city of Seattle.

3. Methodology

The study conducts basic data cleaning, restructuring, and feature engineering to prepare the dataset for exploratory analysis and predictive modelling. Our exploratory analysis section uses descriptive statistics and data visualization techniques to better understand the data and provide insight on the business problem. Furthermore, predictive models are developed using the Decision Tree, Random Forest, k-Nearest Neighbor and Logistic Regression techniques, to predict accident severity based on the selected features. See section 4 and 5 for the results.

3.1 Data Cleaning

The main data cleaning issues of this project include the presence of categorical variables, missing

variables, and data imbalance. The process is primarily conducted through python and its Pandas package. Some variables are also encoded and standardized to suit machine learning techniques. Please check the python notebook attached in my GitHub for the specifics as well as the coding.

3.2 Missing Variables

First, data entries with missing values in the target variable column, SEVERITYCODE, are dropped. Second, we also drop entries with missing values in the feature variables (i.e., Road Condition, Weather, Light Condition, Under Influence, Speeding, Inattention, Address Type, Collision Type, and Junction Type). The rationale is the small volume of missing data contained won't make a meaningful impact for this report. The data volume after dropping missing values sit at 188344 entries.

3.3 Feature Engineering and Pre-processing

Based on our feature selection, we disregard the columns in the original dataset that are not in Table 1. However, problems such as data format, consistency, and categorical values require remain. Therefore, the feature variables require further processing. For instance, we convert object Yes/No values to integer 1s and 0s, which better fit machine learning algorithms. In terms of consistency, I address those that have inconsistent data formats (e.g., the UNDERINFL column contains both 1s/0s and Y/N values) to a uniform data type (in this case, integer 1s and 0s). Lastly, for columns with categorical variables, such as WEATHER and ROADCOND, different methods are implemented for the exploratory analysis part and machine learning modelling part. I leave the variable in its original string format for exploratory analysis, which contains mostly data visualization. For machine learning models, these variables are encoded by creating dummy variables for each different value. I also convert

SEVERITYCODE values, 1(property damage) and 2(injury), into 0(property damage) and 1(injury). As a result, the final machine learning dataset is expanded into 44 columns. It is also notable that data entries with 'unknown' or 'other' categorical values are treated as null values and are therefore dropped. At this stage, there are 167310 entries in the dataset.

Another key issue in machine learning models is data imbalance. There are about 3 times as many property damage collisions in the dataset than injury-related collisions. If this issue is left unaddressed, machine learning techniques are therefore likely to be biased (for instance, if we project all the entries to be property damage related, we can still achieve an accuracy of 75%. Therefore, we down-size the final dataset so that there are an equal amount of property damage collisions and injury-related collisions for our target variable, SEVERITYCODE. Lastly, we standardize all variables using the scikit-learn package to fit machine learning

4. Exploratory Analysis Results

Basic exploratory analysis is employed to better understand the relationship between predictor variables and the target variable. Furthermore, it presents the impact of different factors the project seeks to understand and explore.

First, we look at the impact of driver irresponsibility (see Figure 1). Namely, I plot the count of accidents (represented by the number of data entries in the dataset) based on the severity level. It is notable that inattention-related accidents more than doubles the count of those related to speeding or

influence. While it is not surprising, this information is, at a degree, contrary to the common understanding that speeding and under-influence driving are much more dangerous activities. Therefore, it is reasonable to suggest that distracted driving should be considered a core part with regard to education and policy enactment.

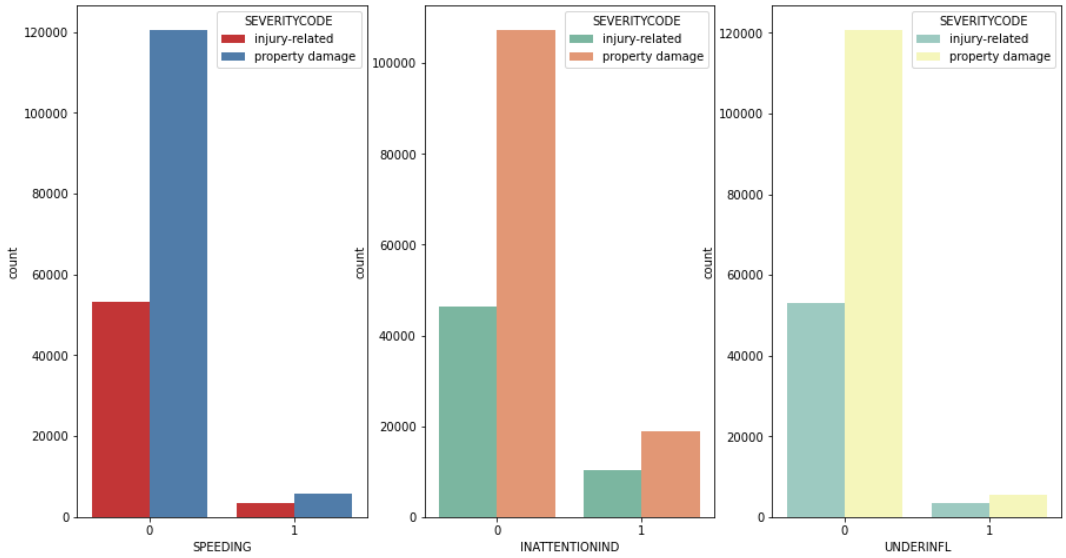


Figure 1. Driver Irresponsibility - Severity

Second, we explore the impact of the surrounding environment of an accident. Table 2. presents a summarized version of information illustrated through bar graphs and pie charts (see Appendix 2).

	Normal Weather Conditions	Adverse Weather Conditions	Normal Road Conditions	Adverse Road Conditions	Normal Light Conditions	Adverse Light Conditions
Count (Total)	107474	60366	119897	47943	119276	48564
Percentage (Total)	64.03%	35.97%	71.44%	28.56%	71.07%	28.93%
Percentage (Property Damage)	64%	36%	71.6%	28.4%	70%	30%
Percentage (Injury-related)	63.9%	36.1%	71.2%	28.8%	72.9%	27.1%

Table 2. Normal vs. Adverse Environment Conditions in Car Collisions

It is apparent that a significant amount of accidents happens under adverse weather, road, and lighting conditions. Notably, the percentage of accidents are significantly higher with adverse weather conditions than with other adverse conditions. This may be attributable to the fact that adverse weather conditions are more loosely defined. For light conditions, day light (including dusk and dawn) is considered to be normal and darkness (even with street lighting) is considered adverse. Therefore, interpretations may change depending on the categorization of "normal" and "adverse". Another significant point illustrated by the table is that the number of severe (injury-related) accidents is barely influenced by adverse environment conditions. The percentage make-up of each category is nearly identical across all three environmental factors. See Appendix 2. for the count and percentage make-up of each individual category of the environmental predictor variables.

5. Predictive Modelling Results

The scikit-learn package in python is used for machine learning analysis section of the study. A separate machine learning dataset is created with specific features mentioned above. The data is then normalized and balanced through the scikit-learn preprocessing and standard scaler function. Moreover, the SDOT dataset was split into train/test datasets with a 0.7/0.3 weight. Specifically, four standard machine learning techniques are employed, namely Decision Tree, Random Forest, Logistic Regression, and K-nearest Neighbor, to predict car accident severity. Standard accuracy-evaluation tools, such as Jaccard Similarity report, log loss, and Classification report are used where applicable, for the better evaluation of different developed models.

5.1 Decision Tree

First, a decision tree model is developed. A decision tree model uses an iterative process to classify the accident as property-damage or injury-related based on the selected features. The decision tree is constructed top-down and categorizes data based on value homogeneity. The criterion “entropy” is used for the calculation of homogeneity. For its criterion “max-depth”, I use an iterative selection process to select the most accurate model at max-depth = 9. Table 3. shows the classification report. The decision tree model has a F-1 score of 0.63, 0.72 for property damage and injury-related accident, respectively.

	Precision	Recall	F-1
0	0.75	0.55	0.63
1	0.65	0.81	0.72
Micro avg	0.68	0.68	0.68
Macro avg	0.70	0.68	0.68
Weighted avg	0.70	0.68	0.68

Table 3. Decision Tree Model Classification Report

5.2 Random Forest

Second, a random forest model is developed. A random forest model employs a collection of

randomly created decision tree models, trains the model different samples of the data, and eventually produces an aggregated predictive outcome. Random forest models are consistently used in data science studies for its ability to improve accuracy and enhance over-fitting problems. We set the estimator to 100, indicating that 100 decision tree models are used in the forest. The random forest model has an overall F-1 score of 0.65, 0.70 for property damage and injury-related predictions, respectively. We also use the *feature_importances_* feature of the python scikit-learn package to illustrate the usefulness of each predictor variable in the prediction of the target variable (see Figure 2). The predictors that relates to magnitude of the accident, such as pedestrian count, have the highest feature importance. Notably, it also echoes the results of exploratory analysis, which indicates that driver irresponsibility is more useful in the prediction of accident severity than the surrounding environment.

	Precision	Recall	F-1
0	0.71	0.59	0.65
1	0.65	0.76	0.70
Micro avg	0.68	0.68	0.68
Macro avg	0.68	0.68	0.68
Weighted avg	0.68	0.68	0.68

Table 4. Random Forest Model Classification Report

	variables	importance
0	Parked Car	0.161246
1	VEHCOUNT	0.115974
2	PEDCOUNT	0.075387
3	Rear Ended	0.060131
4	Pedestrian	0.056647
5	Sideswipe	0.054377
6	Cycles	0.050561
7	PEDCYLCOUNT	0.048511
8	INATTENTIONIND	0.032363
9	UNDERINFL	0.031749
10	SPEEDING	0.028382
11	Other	0.024663
12	Mid-Block (not related to intersection)	0.023298
13	Intersection	0.020580
14	At Intersection (intersection related)	0.018245
15	Daylight	0.014864
16	Dark - Street Lights On	0.014856
17	Overcast	0.014361
18	Block	0.013749
19	Clear	0.013336
20	Wet	0.011948
21	Angles	0.011860
22	Dry	0.011849
23	Raining	0.011451
24	Dusk	0.009311
25	Left Turn	0.008870
26	Mid-Block (but intersection related)	0.008803
27	Dark without lights	0.007598
28	Dawn	0.007463
29	Head On	0.007376
30	Right Turn	0.007303
31	Driveway Junction	0.006492
32	Other Poor Conditions	0.006124
33	Other Poor Road Conditions	0.006020
34	At Intersection (but not related to intersection)	0.002359
35	Ramp Junction	0.001287
36	Alley	0.000605

Figure 2. Feature Importance in the Random Forest Model

5.3 Logistic Regression

Since SEVERITYCODE is a binary target variable, a logistic regression model is also employed. Logistic regression predicts the probability of the binary variable (fitting into either class) based on feature variables. Similarly, I use the SCIKIT-LEARN package in python for model development and evaluation. The model uses regularization level of C = 0.01 and liblinear solver. The model result has an F-1 score of 0.64, 0.72 for property damage and injury-related accidents.

	Precision	Recall	F-1
0	0.74	0.56	0.64
1	0.65	0.80	0.72
Micro avg	0.68	0.68	0.68
Macro avg	0.69	0.68	0.68
Weighted avg	0.69	0.68	0.68

Table 5. Logistic Regression Model Classification Report

5.4 K-Nearest Neighbor

Lastly, a kNN model is developed. A kNN model is a standard classification technique that sets patterns of data in a hyperspace and classifies a datum based on its neighbors. I write a simple code to map and select the optimal k value for best model accuracy. The k value indicates how many neighbors (clusters of classified data) in the model. Due to computational capability and time, I test out 50 k values (from 1 to 50), and found an optimal value of k at 46 (see Figure 3). See Table 6. For

the classification report. The kNN model has an F-1 score of 0.65 and 0.69 for property damage and injury-related accidents, respectively.

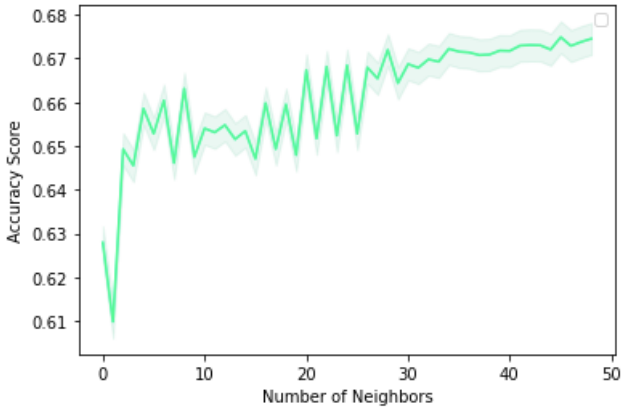


Figure 3. K value and Model Accuracy

	Precision	Recall	F-1
0	0.69	0.62	0.65
1	0.65	0.72	0.69
Micro avg	0.67	0.67	0.67
Macro avg	0.67	0.67	0.67
Weighted avg	0.67	0.67	0.67

Table 6. kNN Model Classification Report

6. Discussion

The discussion section focuses on the comparison between different modelling techniques. Table 7. Presents a summarized report on the performance of each model based on different evaluation metrics. The train/test set accuracy employs the *accuracy_score* function in SCIKIT-LEARN package. Furthermore, both F-1 and Jaccard Similarity are standard evaluation metrics on the accuracy between predicted result and actual result (for the test set of the data). Specifically, F-1 score to measure the harmonic mean of model's precision and recall and the Jaccard coefficient measure the similarity between the two sets of the data. Although the differences of each model are barely discernable, decision tree model and logistic regression model performs the best, with an equal level of F-1 score and Jaccard Similarity score. The kNN model performs the worst in each category.

	Train Set Accuracy	Test Set Accuracy	F-1 Score	Jaccard Similarity Score
Decision Tree	0.683	0.681	0.677	0.683
Random Forest	0.703	0.678	0.676	0.678
Logistic Regression	0.680	0.683	0.678	0.683
kNN	0.678	0.670	0.669	0.670

Table 7. Model Accuracy Comparison

7. Conclusion and Recommendations

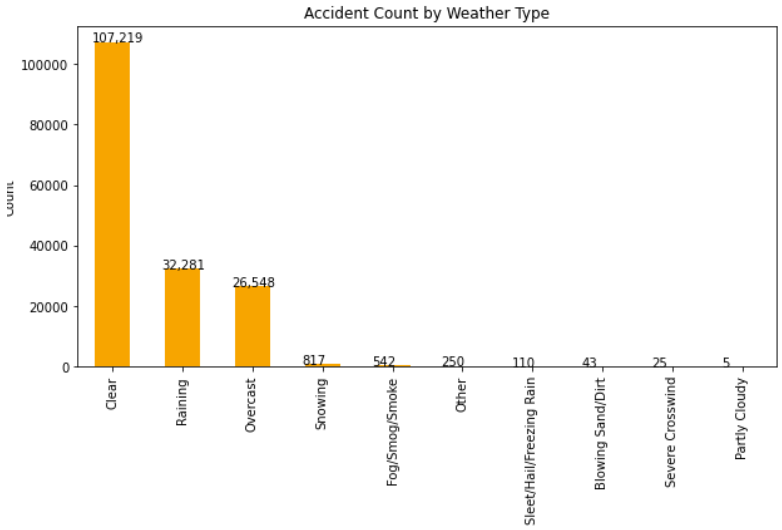
In sum, the exploratory analysis and predictive models provide valuable insights for city planners, emergency responders, and other stakeholders. Through visualization and feature importance analysis, we understand that driver irresponsibility is a significant factor in the occurrence and severity of an accident. The inattention factor is specifically highlighted, since it results in a higher volume of accidents compared to under-influence driving and speeding. Such information can be incorporated into the driver education and policy-making process. For instance, a more stringent policy against distracted driving (e.g., phone usage) can help avoid collisions. Our predictive models can also help policy-makers to understand the weak points with regard to civil engineering. For instance, they can employ the models to identify physical locations that are more susceptible to more severe collisions, based on its traffic flow (the susceptibility of pedestrian involvement), road and light conditions, etc. Moreover, in addition to Seattle, Washington, the models are applicable to other major metropolitan cities in the United States and around the world, given the selected features are universal and easily obtainable.

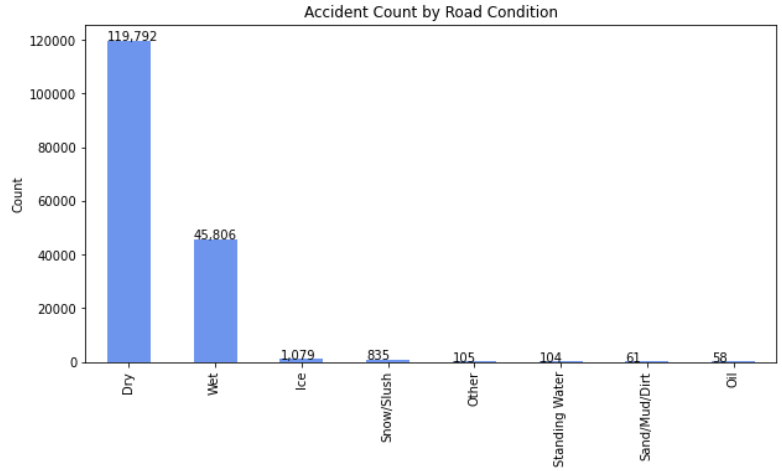
In terms of model evaluation, all four models perform at a similar level, but the accuracy of each model is not up to an optimal standard. I suspect that having a better feature selection, which includes the date and physical location of the accident, would provide a better prediction result. In addition, a more inclusive and complete dataset can possibility improve the model accuracy. The large volume of missing (NaNs, unknowns, and others) in categorical variables significantly reduced the size and quality of the dataset. Moreover, other useful information such as visibility and humidity, which are easily obtainable and recorded, should be incorporated into the data collection process for future use.

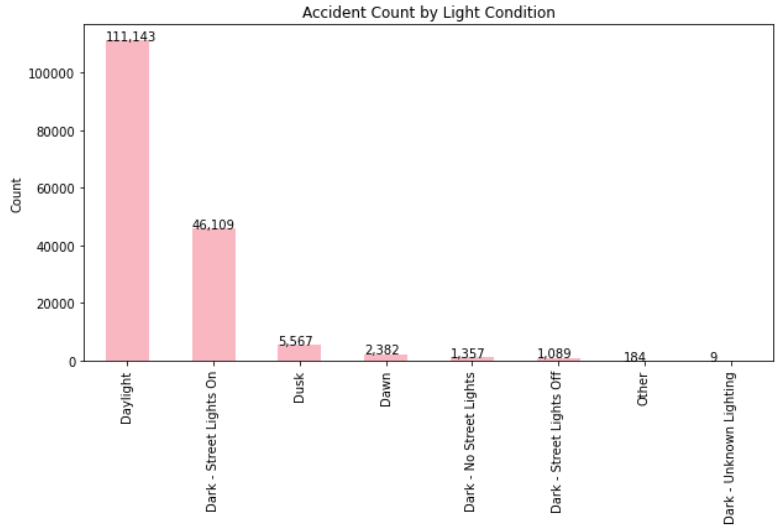
Appendix 1. Raw Data Information

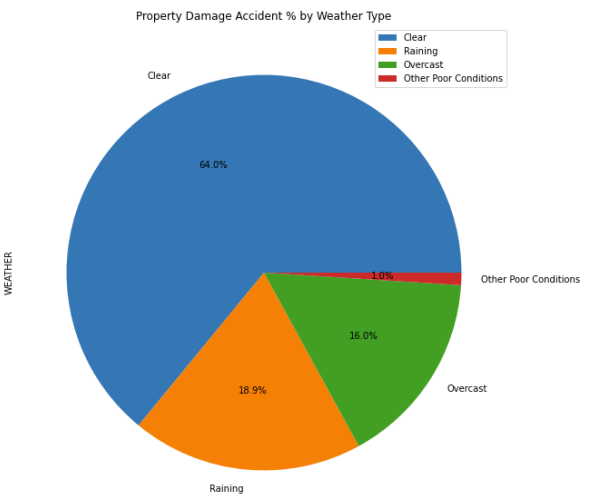
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 38 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   SEVERITYCODE    194673 non-null  int64
 1   X               189339 non-null  float64
 2   Y               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
 11  EXCEPTRSNCODE   84811 non-null   object
 12  EXCEPTRSNDESC   5638 non-null    object
 13  SEVERITYCODE.1  194673 non-null  int64
 14  SEVERITYDESC    194673 non-null  object
 15  COLLISIONTYPE   189769 non-null  object
 16  PERSONCOUNT     194673 non-null  int64
 17  PEDCOUNT        194673 non-null  int64
 18  PEDCYLCOUNT     194673 non-null  int64
 19  VEHCOUNT        194673 non-null  int64
 20  INCDATE         194673 non-null  object
 21  INCDTTM         194673 non-null  object
 22  JUNCTIONTYPE    188344 non-null  object
 23  SDOT_COLCODE    194673 non-null  int64
 24  SDOT_COLDESC    194673 non-null  object
 25  INATTENTIONIND  29805 non-null   object
 26  UNDERINFL       189789 non-null  object
 27  WEATHER         189592 non-null  object
 28  ROADCOND        189661 non-null  object
 29  LIGHTCOND       189503 non-null  object
 30  PEDROWNOTGRNT   4667 non-null    object
 31  SDOTCOLNUM      114936 non-null  float64
 32  SPEEDING        9333 non-null    object
 33  ST_COLCODE      194655 non-null  object
 34  ST_COLDESC      189769 non-null  object
 35  SEGLANEKEY      194673 non-null  int64
 36  CROSSWALKKEY    194673 non-null  int64
 37  HITPARKEDCAR    194673 non-null  object
dtypes: float64(4), int64(12), object(22)
memory usage: 56.4+ MB
```

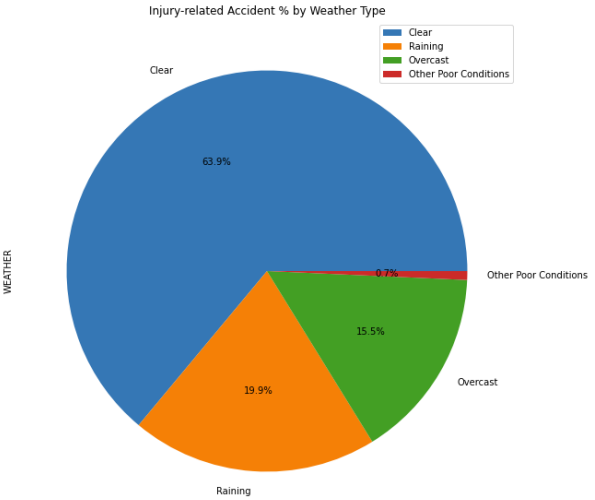
Appendix 2. Bar charts and Pie Charts

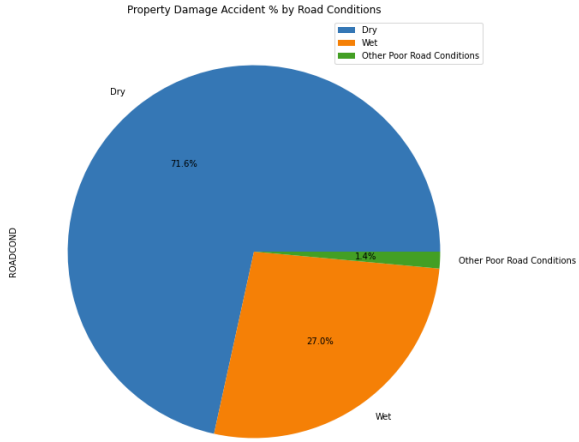


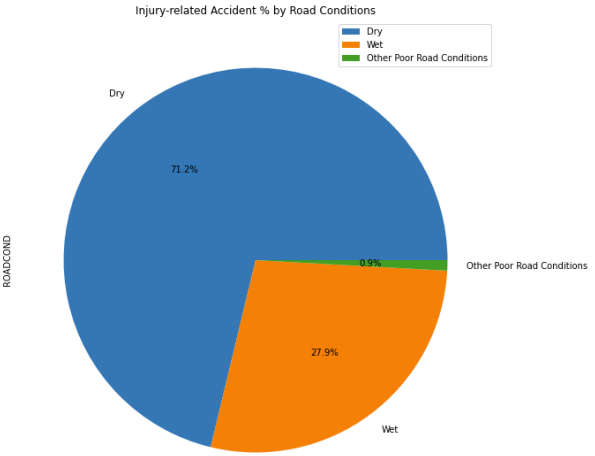


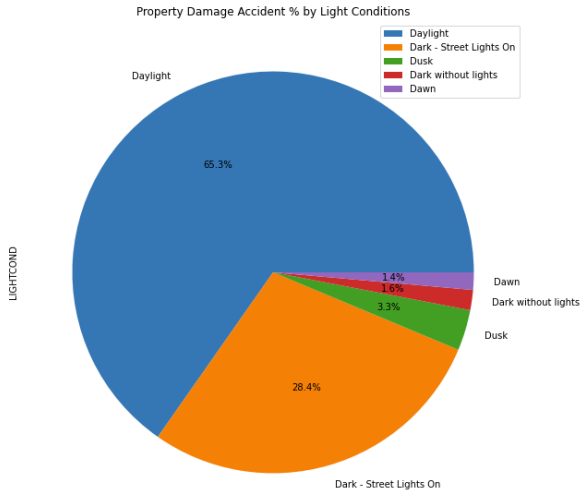


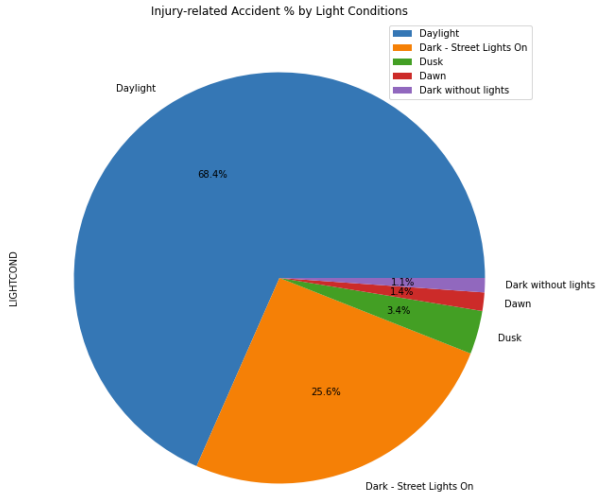

Property Damage Accident % by Weather Type
Clear
Raining
Overcast
Other Poor Conditions
WEATHER
Clear
64.0%
1.0%
Other Poor Conditions
16.0%
Overcast
18.9%
Raining


Injury-related Accident % by Weather Type
Clear
Raining
Overcast
Other Poor Conditions
WEATHER
Clear
63.9%
0.7%
Other Poor Conditions
15.5%
Overcast
19.9%
Raining


Property Damage Accident % by Road Conditions
Dry
Wet
Other Poor Road Conditions
ROADCOND
Dry
71.6%
1.4%
Other Poor Road Conditions
27.0%
Wet


Injury-related Accident % by Road Conditions
Dry
Wet
Other Poor Road Conditions
ROADCOND
Dry
71.2%
0.9%
Other Poor Road Conditions
27.9%
Wet


Property Damage Accident % by Light Conditions
Daylight
Dark - Street Lights On
Dusk
Dark without lights
Dawn
LIGHTCOND
Daylight
65.3%
1.4%
1.6%
3.3%
Dawn
Dark without lights
Dusk
28.4%
Dark - Street Lights On


Injury-related Accident % by Light Conditions
Daylight
Dark - Street Lights On
Dusk
Dawn
Dark without lights
LIGHTCOND
Daylight
68.4%
1.1%
1.4%
3.4%
Dark without lights
Dawn
Dusk
25.6%
Dark - Street Lights On

Reference:

https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013