Text

Description automatically generated

Logo

Description automatically generated

Rambhupal Payyavula

ALY6140- Analytics System Technology, 80554, Spring 2022.

Capstone Project: New York Motor Vehicle Collisions Analysis

[Payyavula.r@northeastern.edu](mailto:Payyavula.r@northeastern.edu)

**Introduction:**

The Motor Vehicle Collisions crash dataset (Motor Vehicle Collisions - Crashes, n.d.) **c**ontains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables have information from all police-reported motor vehicle collisions in NYC.

This project aims to analyze the victims of motor vehicle collisions in New York. Based on the killed and injured in different years. The road collision accidents were recorded from July 2012 to April 2022. Based on this available data, I present my findings regarding the collision incidents.

**Questions that need to be answered from the dataset:**

1. Which year is evident for the highest number of injured people?
2. Which year is clear for the highest number of killed people?
3. What is the distribution of different kinds of people injured?
4. What is the distribution of different types of people dead in collisions?
5. A total number of people injured/killed based on the Borough?
6. Which place is safer between Manhattan and Queens?
7. Is there any relation between the people wounded and people killed?
8. Is there any relation between the people hurt and people injured due to motorcycle collisions?
9. Variation in the number of people been wounded along the available time?

**Methods**

General data extractions and cleaning. Dropping the unwanted fields from the data will decrease the execution time. I am visualizing different charts with the help of python matplotlib and seaborn. Moreover, in this project, I will use three other models as t-tests to find which borough has more injuries, linear regression to locate the relation between the persons killed and injured, and Time series analysis to forecast the number of injuries.

**Dataset Dimensions:**

This dataset comprises of 1.83Million records and 29 attributes.

**Attributes Description:**

CRASH DATE - Occurrence date of collision - Date & Time

BOROUGH - Borough where the collision occurred - Plain Text

ZIP CODE - Postal code of incident occurrence - Plain Text

LATITUDE - Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) - Number

LONGITUDE - Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) – Number

LOCATION - Latitude, Longitude pair - Location

ON STREET NAME - Street on which the collision occurred - Plain Text

CROSS STREET NAME - Nearest cross street to the crash - Plain Text

OFF-STREET NAME - Street address if known - Plain Text

NUMBER OF PERSONS INJURED - Number of persons injured - Number

NUMBER OF PERSONS KILLED - Number of persons killed - Number

NUMBER OF PEDESTRIANS INJURED - Number of pedestrians injured - Number

NUMBER OF PEDESTRIANS KILLED - Number of dead pedestrians - Number

NUMBER OF CYCLIST INJURED - Number of cyclists injured - Number

NUMBER OF cyclists KILLED - Number of cyclists killed - Number

NUMBER OF MOTORIST INJURED - Number of vehicle occupants injured - Number

NUMBER OF motorists KILLED - Number of vehicle occupants killed - Number

CONTRIBUTING FACTOR VEHICLE 1 - Factors contributing to the collision for designated vehicle - Plain Text

CONTRIBUTING FACTOR VEHICLE 2 - Factors contributing to the crash for selected vehicle - Plain Text

CONTRIBUTING FACTOR VEHICLE 3 - Factors contributing to the collision for designated vehicle - Plain Text

CONTRIBUTING FACTOR VEHICLE 4 - Factors contributing to the crash for the designated car - Plain Text

CONTRIBUTING FACTOR VEHICLE 5 - Factors contributing to the collision for selected vehicle - Plain Text

COLLISION\_ID - Unique record code generated by the system. Primary Key for Crash table. - Number VEHICLE TYPE CODE 1 - Type of vehicle based on the selected vehicle category - Plain Text VEHICLE TYPE CODE 2 - Type of vehicle-based on the selected vehicle category - Plain Text VEHICLE TYPE CODE 3 - Type of vehicle-based on the selected vehicle category - Plain Text VEHICLE TYPE CODE 4 - Type of vehicle-based on the selected vehicle category - Plain Text VEHICLE TYPE CODE 5 - Type of vehicle-based on the selected vehicle category - Plain Text

**Exploratory Data Analysis:**

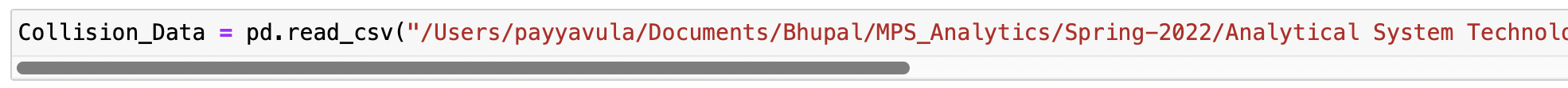
**Data Extraction:**

All the necessary packages have been installed to perform the analysis.

**Text, letter

Description automatically generated**

Loading the collision data into pandas’ data frame



**Validating the data.**

Table

Description automatically generated with medium confidence

From the below screenshot, it is clear that the dataset comprises 1.8 million observations and 29 columns.

Table

Description automatically generated with low confidence

From the below screenshot, we can see that the maximum number of people injured in the collisions is 43, and the minimum is 0, with a mean of 0.2. The maximum number of people killed is eight, and the minimum is 0.

Table

Description automatically generated

**Data Cleaning:**

Removing the unwanted columns and renaming the existing columns:

Text

Description automatically generated

Fourteen columns, which we will not use, have been dropped, and ten columns have been renamed, as listed in the above screenshot.

Graphical user interface, application, Word

Description automatically generated

Borough data have missing records, around ~583k records, which are more than the ~30% of data, which is significant, and we can't disregard them. I'll assign NYC to the missing Borough records. It will be five boroughs and NYC to collect the null values.

We can also see a few null values in total injured and killed; the null values can be removed as they are significantly less in number.

Graphical user interface, text

Description automatically generated with medium confidence

Graphical user interface, application, Word

Description automatically generated

**Data Visualizations:**

Since we have data from July 2012 to April 2022, we cannot conclude the entire year based on the half data in 2012 and 2022. So, data for 2012 and 2022 has been filtered and eliminated

Visualizing the different groups of people injured in vehicle collisions from 2013 to 2021:

**Graphical user interface, chart, bar chart

Description automatically generated**

The above screenshot shows that most people were injured in 2018, and an almost identical number of people were injured in 2016, 2017, 2018, and 2020. The highest number of cyclists were injured in 2020. The highest number of pedestrians were injured in 2013.

Visualizing the different groups of people killed in vehicle collisions from 2013 to 2021:

**Graphical user interface, chart, application, bar chart

Description automatically generated**

The above visualization shows that the most significant number of people died in 2013 and 2021.The highest number of motorists were injured in 2020 and 2021. The highest number of pedestrians were killed in 2013. The highest number of cyclists were killed in 2019 and 2020.

**Visualizing the total victims in different boroughs:**

**Chart, pie chart

Description automatically generated**

NYC is unknown values; we can see that 31 percent of victims' place is unknown. The most significant victims were from Brooklyn, Queens, Manhattan, Bronx, and Staten Island.

**Visualizing the total number**

**Chart, histogram

Description automatically generated**

The above graphs show that most people were injured/ killed in Brooklyn, followed by queens. The exact number of people were wounded in queens and Manhattan, but most people died in Manhattan.

**Predictive Models:**

**Two sample t-tests:**

The two-sample (unpaired or independent) t-test analyzes the means of two separate groups to see if they are similar or different. We compute the sample means from two groups and deduce the conclusion for the population's means (unknown means) from which two groups are taken in a two-sample t-test.

I want to perform a t-test to compare the mean of total injured between the Manhattan and Queens.

Before performing the t-test, let's visualize the available data of total injured people in Queens and Manhattan.

**A picture containing text

Description automatically generated**

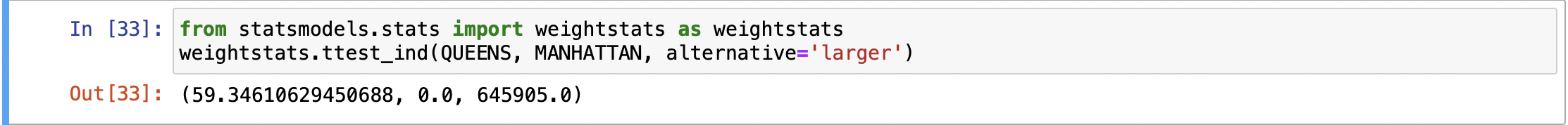
From the above graph, it is clear that the total number of injured people in Manhattan is greater than the Queens. We can validate this with the help of a two-sampled t-test.

Let’s start the two sampled t-tests by stating the null and alternative hypotheses.

**Null hypothesis:** Mean of injured people in Queens is more significant than in Manhattan.

**Alternative** hypothesis**:** Mean of injured people in Queens is not greater than in Manhattan.

**Significance level**: 0.01



From the two sampled t-tests, we obtained a p-value of 0, which is less than 0.01. As P-vale is less than the significance, we have enough evidence to reject the null hypothesis.

We can conclude that the Mean of injured people in Manhattan is not greater than in Queens.

**Linear regression:**

By fitting a linear equation to observed data, linear regression seeks to model the connection between two variables. One variable is regarded as an independent variable, while the other is considered a dependent variable. This will be useful in finding the relation between the types of vehicles and accidents.

.

**Regression model 1:**

Let’s find if any relationship exists between the total injured and killed with the help of linear regression.  
Total injured is an independent variable and killed is a dependent variable.

Table

Description automatically generated

From the above screenshot, the R-squared value of 0. So, we can say that no relationship exists between the total injured and the total killed.

**Regression model 2:**

In this model, we will find if any relationship exists between the total injured and motorists injured with the help of linear regression.  
Total injured is a dependent variable, and motorists injured is an independent variable.

Table

Description automatically generated

We can see that the R-squared value is 0.826, where we can say that there is a relationship between the total injured and motorists injured.

**Time series analysis:**

Time series analysis is a method for studying a collection of data points over some time. Instead of sporadically or arbitrarily, time-series analyzers capture data points at constant intervals over a predetermined time.

In this Analysis, we will generate time series data based on the existing data of total injured and the crash date, and then we will split the data into train and test data. We will predict test data values based on the train data.

Graphical user interface, text, application, email

Description automatically generated

I am plotting the graph between the crash date and the total injured.

**Chart

Description automatically generated**

We can see from the graph that there is no consistent pattern, but we can see a decrease in the number of injured from 2013 to 2022.

**Splitting the train and test data and visualizing it:**

**Graphical user interface, application

Description automatically generated**

Data is split into train and test based on the year. I consider less than November 2020 as train data and more significant than November 2020 as test data.

Graphical user interface, text, application

Description automatically generated

The above graph visualizes the predictions on the test data. We can see in the forecast that the total number of injured decreased by the end.

**Conclusion:**

Considering all the Visualizations, most people were injured in 2018, and an almost identical number of people were injured in 2016, 2017, 2018, and 2020. The highest number of cyclists were injured in 2020. The highest number of pedestrians were injured in 2013. the most significant number of people died in 2013 and 2021.The highest number of motorists were injured in 2020 and 2021. The highest number of pedestrians were killed in 2013. The highest number of cyclists were killed in 2019 and 2020.

Most people were injured/ killed in Brooklyn, followed by queens. The exact number of people were wounded in queens and Manhattan, but most people died in Manhattan.

**References:**

s*eaborn-line-plot*. (n.d.). Retrieved from pythonbasics.org: https://pythonbasics.org/seaborn-line-plot/

*How to Interpret the F-test of Overall Significance in Regression Analysis*. (n.d.). Retrieved from statisticsbyjim.com: https://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/

*Motor Vehicle Collisions - Crashes*. (n.d.). Retrieved from data.cityofnewyork.us: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

*Pythonic Data Cleaning With Pandas and NumPy*. (n.d.). Retrieved from realpython.com: https://realpython.com/python-data-cleaning-numpy-pandas/

*time-series-forecasting-python*. (n.d.). Retrieved from builtin.com: https://builtin.com/data-science/time-series-forecasting-python