National College of Ireland

Project Submission Sheet

Student Name: Mustafa Sayin & Mustafa Karaburun……………………………………………………

Student ID: x23174773 & x23216158……………………………………………………………………

Programme: Msc in Data Analytics………………………… **Year:** 2024……………

Module: Database & Analytics Programming …………………………………………………

Lecturer: Athanasios (Thanos) Staikopoulos……………………………………………………

Submission Due Date: 02/05/2024…………………………………………………………………………………

Project Title: Daily Traffic Count & Collisions in New York City……………………………

Word Count: 2669……………………………………………………………………………………………

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Mustafa Sayin & Mustafa Karaburun…………………………………………………

Date: 01/05/2024…………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Daily Traffic Count & Collisions in New York City

*Note: Sub-titles are not captured in Xplore and should not be used

Mustafa Karaburun
School of Computing
National College of Ireland
Dublin, Ireland
x23215158@student.ncirl.ie

Mustafa Sayin
School of Computing
National College of Ireland
Dublin, Ireland
x23174743@student.ncirl.ie

***Abstract*—Traffic collisions are important events that need to be investigated to learn the cause of the accidents. Therefore, researchers felt the need to examine whether most collisions occurred during peak traffic times. New York City was taken as a basis in 2019. This study uses two different datasets to examine factors of traffic congestion and visualization over time and the number of injuries/deaths. Traffic density and number of injured are considered. Visualization of the analysed data is shown in the report. In this study, we used data.ny.gov and data.cityofnewyork.us data. the result obtained from this study is Accidents occur more in places where traffic is dense**

Keywords—Traffic, Collision, New York, Python, MongoDB, PostgreSQL

I. Introduction (*Heading 1*)

In this research, we used New York traffic count and accident datasets as JSON and CSV files. Traffic density is increasing day by day in the world. There are many factors behind the increase in density. It might be said that the reasons for the increase in traffic accidents are drivers and pedestrians who do not comply with the traffic rules. It was aimed to investigate what the factors affecting traffic density might be and in which vehicle types accidents occur more frequently. Data preparation was carried out after the datasets we transferred to MongoDB using the Python programming language. Cleaned datasets were imported into PostgreSQL. Visualizations were made for the users of various libraries.

II. Research Questions

1. Which New York region has the highest number of injured people?
2. Which vehicle was involved in more accidents?
3. Is there a relationship between road length and traffic density?
4. What is the traffic density by road type?

III. Literature Review

Car crashes, sometimes referred to as traffic collisions, are a serious public safety issue that impacts millions of Americans. The National Highway Traffic Safety Administration predicts that there will be 38.680 motor vehicle crash deaths in 2020 alone, and that motor vehicle crashes will cost $242 billion in economic damages in 2019.[3]

Various strategies have been implemented to prevent traffic accidents and reduce their impact these approaches encompass diverse tactics such as enhancements to road infrastructure, advancements in vehicle safety systems and public awareness campaigns to promote safe driving habits. One notable example is the “Click it or Ticket ” campaign, which urges drivers and passengers to buckle up while on the road. This initiative has been credited with substantially raising seat belt usage in the United States from 58% in 1994 to exceeding 90% in recent times. (National Highway Traffic Safety Administration,2021) [4]

New York City is one of the most traffic-congested cities in the US and the world. Provides a comprehensive analysis of the total number of people injured and killed in road crashes in NYC. Night-time showed the highest number of people killed each day in all years. By type of people injured and killed, cyclists the lowest and persons (drivers and passengers) were the highest.[5]

New York City is one of the most crowded cities in the world. That`s why a lot of study has been done to reduce traffic count. The focus of the studies is to reduce traffic count and minimize the damage to the enviroment and people. Among the studies carried out for this purpose are the promotion of public transportation and the intensive use of cameras in traffic.

IV. Methodology

A. Dataset Selection

It contains the following information about each dataset we used in the project:

NAME	**TYPE**	**FORMAT**	**LINK**
2019 collisions NYC	CSV	CSV	CSV
2019 AADT - NYC	JSON	JSON	JSON

B. Dataset Description

1) Annual Average Daily Traffic (AADT): based on 2019, The Annual Average Daily Traffic(AADT) serves as an approximation of the typical daily traffic volume on a specific stretch of road. It is derived from short-term traffic counts conducted on the same segment, which are then adjusted to generate the AADT estimate. Consequently, the latest AADT data for any particular roadway will pertain to the preceding year. Comprehensive AADT information is accessible for all New York State Routes and roads incorporated into the federal Aid system. A few of the columns in the dataset are below[1]

Attribute	Description
Year	the year for which the AADT
County	where the count station exists
State Route	an interstate or NY route number
Road Name	the name of the road

2) 2019 collisions in NYC, The crash table within the Motor Vehicle Collisions dataset provides comprehensive details regarding each collision event. Every row in the table corresponds to a distinct crash incident. This dataset encompasses information from all Motor Vehicle collisions reported to the police in New York City. Specifically, the police report from (MN104-AN) must be completed for collisions resulting in injuries and fatalities, or those involving property damage totalling at least $1000. The dataset has 221k rows and 29 columns. Some of the columns' names are below[2]

Attribute	Description
Borough	where collision occurred
Latitude	coordinate
On Street Name	street on which the collision occurred
Number of Persons Injured	The count of injured people

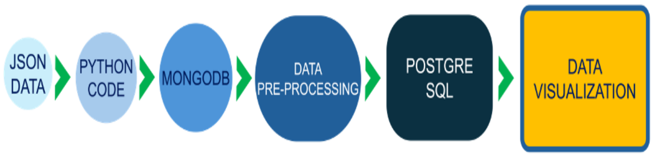
C. *Libraries and Tools*

This section discusses the methods for extracting, transforming, and loading data to analyse each dataset.

Processing the data involves importing the data in JSON format using the Python programming language and storing it in cloud Mongo DB instances, which are best used to store JSON data.

The dataset provided in JSON format, which contains information about traffic density and accidents in 2019, can be used to find where the accidents and density occur most, to analyse vehicle types and numbers, and to investigate the relationship between variables such as bridge-ramp and traffic density.

After this, each dataset was extracted from the Mongo DB database and converted into a data table in PostgreSQL to facilitate clean extraction and visualization of the data.



In this project, Python was chosen as the primary programming language for data processing and analysis. We also used a variety of libraries to assist with data processing, visualization, and pipeline management, including psycopg2, matplotlib, seaborn, pandas, NumPy, pymongo, and dagster. We chose MongoDB to store Traffic and collision data due to its scalability, flexibility, and availability. We used the PyMongo package to communicate with the MongoDB database.

Data Cleaning: We used various libraries to clean the datasets. For example, we handled missing values and removed unnecessary columns using the pandas library. We filled in other missing values with appropriate values.

Data Visualization: We leveraged several libraries for data visualization, including Matplotlib and Seaborn. These libraries offer a variety of tools to create various types of graphs and charts. We used Matplotlib and Seaborn to make visualizations such as scatter plots, bar charts and pie charts.

D. *Preprocessing*

1) CSV

A separate Python code was created to transfer the data from the MongoBD, the next task was to prepare the data and transform the data and another file was created for this task. Finally, Python code was created to load data into PostgreSQL and the file was loaded into it.

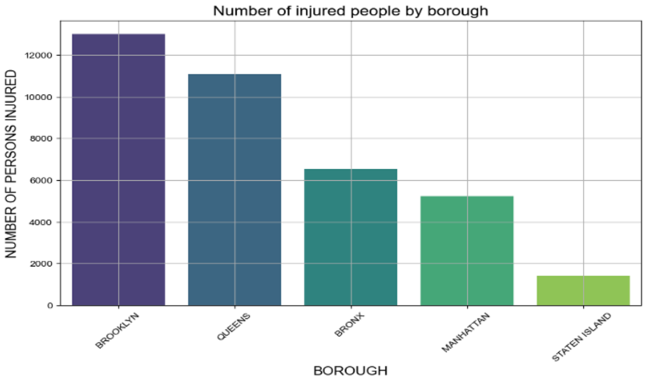
2) JSON

MongoDB was used to extract the json file more efficiently and conveniently. The entire data set was stored programmatically in MongoDB using the Python programming language. Access to MongoDB was achieved with the Pymongo library. Additionally, the Json library was used to read the json file in python programming language. While storing data into MongoDB, a key variable in the json file was selected to ensure that each data was stored uniquely. The processed data set was uploaded to the PostgreSQL database. Just like in MongoDB, duplication of existing data was prevented by selecting unique key variable.

V. RESULT AND EVALUATION

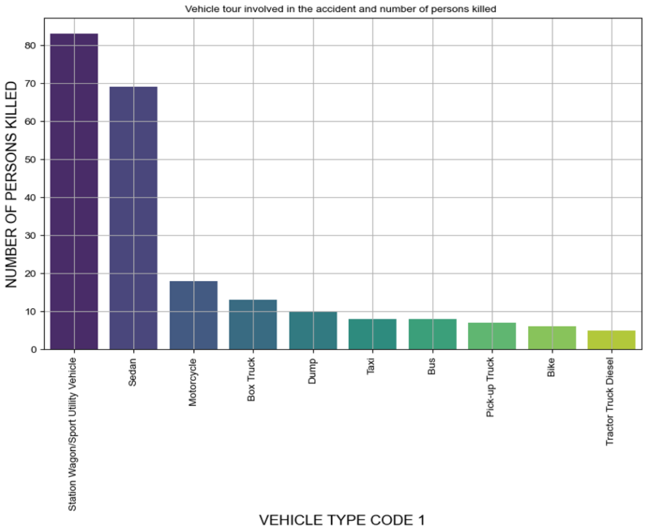
A. *CSV*

1) The graph named "Number of injured people by borough" is a bar chart that offers a visual image of injuries throughout five boroughs of New York City: Brooklyn, Bronx, Manhattan, and Staten Island. Brooklyn reports the highest number of injuries, showing nearly 12,000 cases. This suggests a higher rate of incidents leading to injuries or a larger population where such incidents are more frequent.



Queens follows Brooklyn closely with just under 12,000 injuries, indicative of similar issues as Brooklyn concerning safety or population destiny. Bronx and Manhattan exhibit notably lower figures, with the Bronx showing slightly over 6,000 injuries and Manhattan even less, at approximately 5,000 injuries. These figures might reflect differences in public safety measures, urban planning, or demographic factors.

2) Station Wagon/Sport Utility Vehicle and Sedans show significantly higher fatalities compared to other vehicle types, with fatalities nearing 80 and over 60 respectively. These types might be the most used, potentially explaining the higher number of incidents leading to deaths. Motorcycles, Box Truck and Dump Truck have a moderate number of fatalities, each contributing between approximately 10 and 25 deaths. The relatively higher fatalities in motorcycles could be due to the inherent lack of physical protection compared to enclosed vehicles.

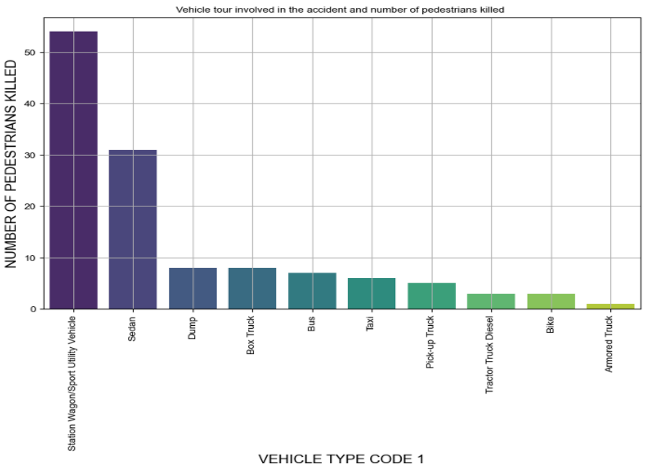


Vehicle Size and Safety Larger vehicles like buses and trucks, despite their potential for severe accidents, have lower fatalities, possibly due to stricter safety regulations and robust build. Bikers face significant risks, yet the relatively low fatality rate could be influenced by non-fatal injury accidents or successful helmet and road safety campaigns.

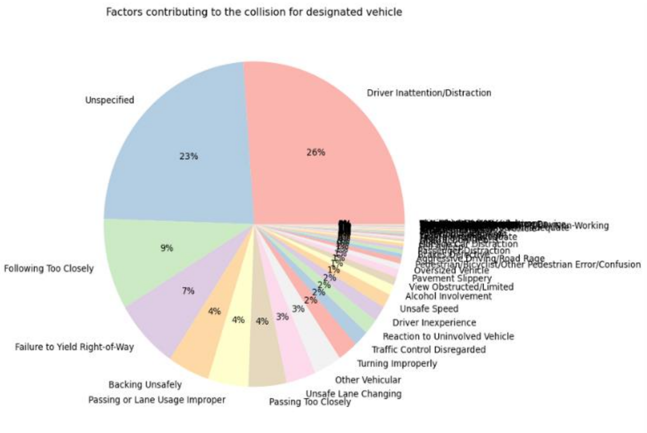
3) The graph titled "Vehicle type involved in the accident and number of pedestrians killed" shows us the count of pedestrian fatalities linked to various types of vehicles involved in accidents.

Station Wagon/Sport Utility Vehicle shows the highest number of pedestrian fatalities, close to 50 deaths, followed by Sedan with around 30 deaths. These vehicles, typically used frequently in urban and suburban settings, indicate a higher risk to pedestrians, possibly due to their numbers on the roads.

Dump Trucks and Box Trucks cause approximately 15 and 20 pedestrian deaths respectively; This could indicate potential visibility issues or difficulty maneuvering these larger vehicles in areas frequented by pedestrians.



4) The graph named "Factor contributing to the collision for designated vehicle" is a pie chart illustrating the different factors that contribute to vehicle collisions. Driver Inattention/Distraction is the largest segment at 26%, indicating that this is the most common cause of collisions. This may include scenarios where the driver was not fully focused on the road due to activities like using a mobile phone, adjusting the radio or other distractions.



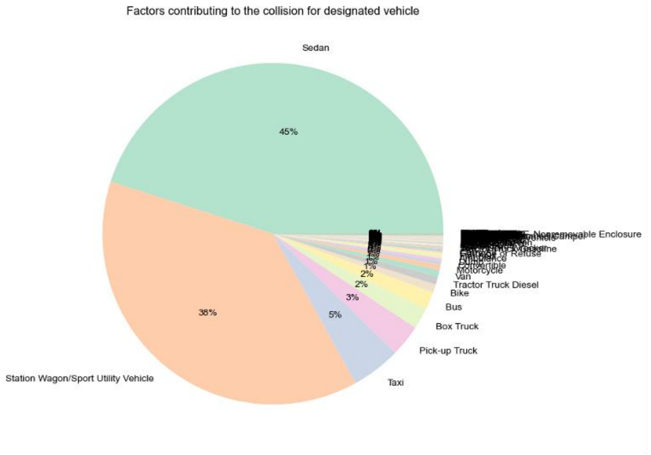
Unspecified is the second-largest category, accounting for 23%. This segment represents collisions where the exact contributing factor was not specified or remains unknown.

Following too closely is the third category that factor accounts for 9% of the collisions, highlighting issues with drivers not maintaining a safe following distance from the vehicle in front.

5) Station Wagon/Sport Utility Vehicle this category dominates the chart with 45% of collisions, suggesting that these vehicles are commonly involved in accidents, potentially due to their prevalence on the road.

Sedan is in the second largest segment at 38%, also indicating a high involvement in collisions, aligning with their frequent use in both urban and suburban environments.

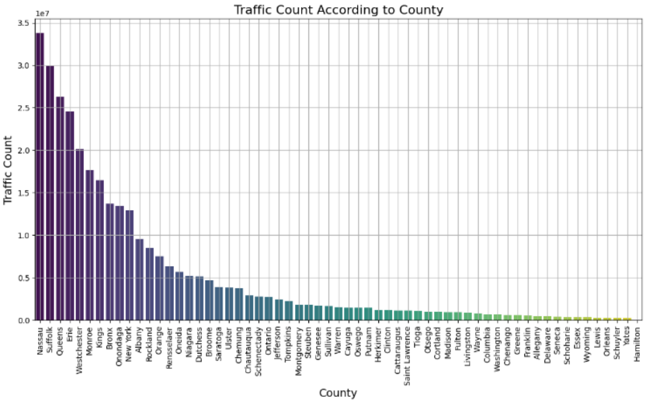
Taxis contribute to 5% of the collisions, notable since taxis are heavily used in urban areas, their proportion in collisions could reflect both high usage and the challenging driving conditions they often face.



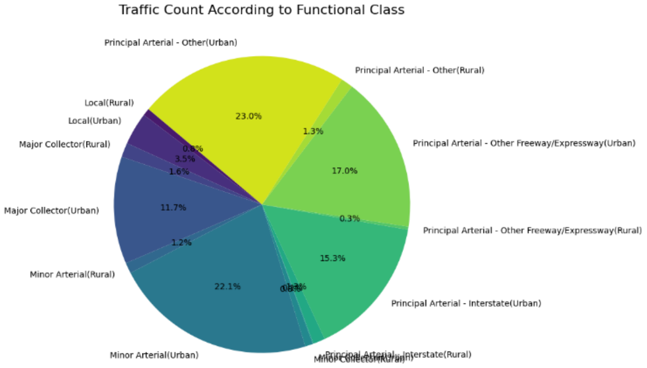
The high percentages of station wagons/SUVs and sedans in crashes underscore the importance of personal vehicles in tragic events. It probably has to do with their abundance in the vehicle population. Despite their constant presence in high-traffic areas and their operational demands, in terms of taxi safety, taxis account for a relatively modest share of crash statistics; this may reflect professional driving training or stronger regulatory compliance.

B. JSON

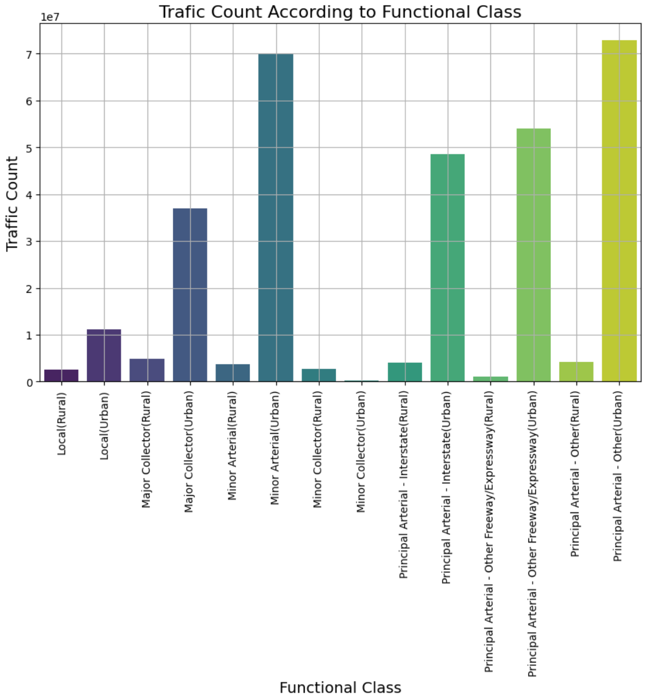
1) In this bar chart, the counties in New York City are represented on the x-axis, while the traffic count in the counties is represented on the y-axis. The traffic count is concentrated in certain areas. This density is even higher, especially in counties close to the city centre. Such as, Nassau, Suffolk, Queens, Westchester. These indicators indicate the high population density and the high number of vehicles in these counties.



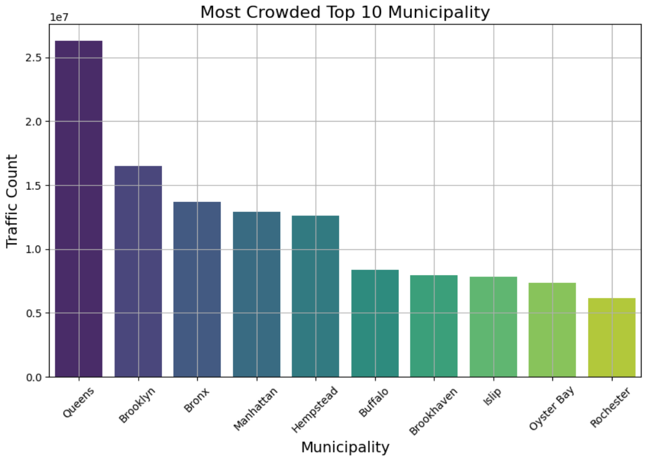
2) This pie chart shows traffic counts by the functional class. The slices in the chart are expressed as percentages and the slice with the highest percentage is Principal Arterial – Other(Urban) with 23 percentage. The second largest segment is the Minor Arterial(Urban) class with 22.1 percentage. This graph shows that almost fifty percentage of the traffic count is concentrated on these two types of roads



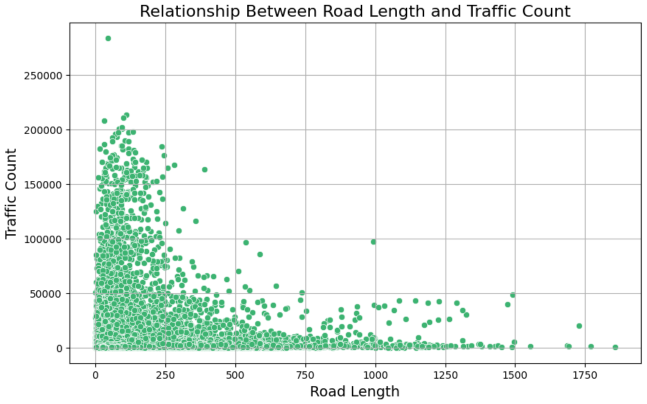
3) This Bar plot also shows the same thing as pie chart. The difference is the traffic count is showed with the numbers not percentage. The traffic counts on the roads in the urban areas are quite high. Traffic counts in rural areas are also relatively high, but not as much as on urban areas.



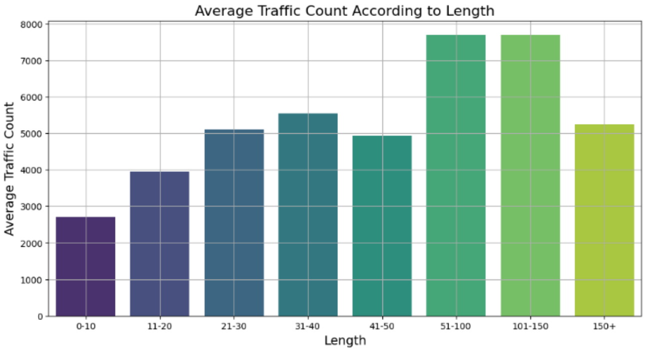
4) This bar chart shows the 10 most crowded municipalities by traffic count. The municipality with the highest traffic count in the chart is Queens. Queens is followed by the Bronx and Manhattan. It can be said that the traffic count in these municipalities is high due to the dense of population and proximity to financial centers.



5) This scatter plot shows the relationship between road length and traffic count. In the graph, the x-axis shows the road length and the y-axis shows the traffic count. From this graph, it can seen that traffic count decreases as the road length increases. They have negative correlation.



6) This bar chart shows the average traffic count by grouped road lengths. Road lengths are on the x-axis and average traffic counts are on the y-axis. As can be seen from this graph, as road lengths increase, the average traffic count also increases. However, on longer routes the average traffic count decreases. The reason can be interpreted as the infrequency of long-distance trips. Additionally, we can see positive correlation between road length and traffic count because these road lengths are very short according to previous graph.



VI. CONCLUSIONS AND FUTURE WORK

A. CSV

After examining the CSV file, it was concluded that the district where people were injured the most was Brooklyn and Queens is the second borough. Station Wagon/sports Utility vehicles and Sedans show significantly higher fatalities compared to other vehicle types, with fatalities nearing 80 and over 60 respectively. The factor contributing to the collision for a designated vehicle is the largest segment with Driver Inattention/Inattention at 26%, indicating that it is the most common cause of crashes.

B. JSON

When the Json file was examined, it was observed that the counties and municipalities with highest traffic count in New York City in 2019 were Nassau, Suffolk, Queens, and Brooklyn. Additionally, the traffic count of roads in urban areas was higher than in rural areas. Another result was that road lengths were found to be negatively correlated with traffic counts. Within the scope of the information obtained from these analyses, it has been determined in which regions the infrastructure and road maintenance works of the New

York city should be carried out more, in which regions the emergency centers should be established, and which regions the traffic reduction incentives or public transportation should be increased.

VII. References

[1] Annual Average Daily Traffic (AADT): based on 2019 online: https://data.ny.gov/Transportation/Annual-Average-Daily-Traffic-AADT-Beginning-1977/6amx-2pbv/about_data

[2] 2019 collisions in NYC online: https://data.cityofnewyork.us/Public-Safety/2019-collisions/rjnm-88k9/about_data

[3] L. L. Wolf *et al.*, "Factors Associated with Pediatric Mortality from Motor Vehicle Crashes in the United States: A State-Based Analysis," *The Journal of Pediatrics*, vol. 187, pp. 295-302.e3, Aug. 2017, doi: 10.1016/j.jpeds.2017.04.044.

[4] Khaled Shaaban. Mohamed Ibrahim. "Analysis and Identification of Contributing Factors of Traffic Crashes in New York City" Transportation Research Procedia 55 (2021) 1696-1703

[5] NYC Traffic Trends, Street Safety and Public Health [Online], Available at: https://news.climate.columbia.edu/2022/09/27/nyc-traffic-street-safety-public-health/

[6] MongoDb Tutorial [Online], Available at: https://www.datacamp.com/tutorial/mongodb-tutorial-how-to-set-up-and-query-mongodb-databases

[7] Seaborn Tutorial [Online], Available at: https://www.datacamp.com/tutorial/seaborn-python-tutorial

[8] PostgreSQL Tutorial [Online], Available at: https://www.datacamp.com/tutorial/beginners-introduction-postgresql

[9] Evaluating the Traffic and Emissions Impacts of Congestion Pricing in New York City [Online], Available at: https://www.mdpi.com/2071-1050/12/9/3655