Project report on

Analysis of the relationship between New york City Accidents and Weather

Carried out at

CENTER FOR DEVELOPMENT OF ADVANCED COMPUTING
ELECTRONIC CITY, BANGALORE

UNDER THE SUPERVISION OF

Mr. Abhishek Raju Chavana

C-DAC Bangalore

Submitted By

Jemima Joaquina Dias (220950125041)
Kajal Patel (22095015442)
Kamal Prakash Yadav(220950125043)
Nainesh Anil Khanjire (220950125053)
Nalin Pushp (220950125054)

Candidate's Declaration

We hereby certify that the work being presented in the report entitled An Analysis of the Relationship between Weather and Environment factors with Motor Vehicle Accidents, in partial fulfilment of the requirements for the award of PG Diploma Certificate and submitted in the department of PG-DBDA of the C-DAC Bangalore, is an authentic record of our work carried out during the period, 2nd January 2023 to 2nd March 2023 under the supervision of Mr. Abhishek Raju Chavan, C-DAC Bangalore. The matter presented in the report has not been submitted by us for the award of any degree of this or any other Institute/University.

(Name and Signature of Candidate)

Jemima Joaquina Dias (220950125041)

Kajal Patel (220950125042)

Kamal Prakash Yadav (220950125043)

Nainesh Anil Khanjire (220950125053)

Nalin Pushp (220950125054)

Counter Signed by

_____

ACKNOWLEDGEMENT

We take this opportunity to express our gratitude to all those people who have been directly and indirectly with us during the competition of this project.
We pay thanks to Mr. Abhishek Chavan who has given guidance and a light to us during this major project. His versatile knowledge about " An Analysis of the Relationship between Weather and Environment Factors with Motor Vehicle Accidents " has eased us in critical times during the span of this Final Project.

We acknowledge our debt to those who contributed significantly to one or more steps. We take full responsibility for any remaining sins of omission and commission.

Jemima Joaquina Dias
Kajal Patel
Kamal Prakash Yadav
Nainesh Anil Khanjire
Nalin Pushp

CERTIFICATE

This is to certify that the work titled An Analysis of the Relationship between Weather and Environment factors with Motor Vehicle Accidents, is carried out by Jemima Joaquina Dias (220950125041), Kajal Patel (220950125042), Kamal Prakash Yadav (220950125043), Nainesh Anil Khanjire (220950125053), Nalin Pushp (220950125054) the bona fide students of Post Graduate Diploma in Big Data Analytics of Centre for Development of Advanced Computing, Electronic City, Bangalore from 2nd January2023 - 2nd March 2023. The Course End Project work is carried out under my direct supervision and 100% completed.

Mr. Abhishek Chavan

Name of Supervisor

C-DAC #68, Electronic City,

Bangalore - 560100, India

ABSTRACT

Traffic crashes have a significant impact on the world economy and are a leading cause of death and injuries worldwide. More than one-half of all traffic deaths globally occur among people ages 15 to 44, their most productive earning years. This project is developed to understand feature correlations, patterns and analyse trends with the current weather conditions in the area of the accident. To develop this system end to end, first we need to extract vehicle crash dataset from the NYC open data website and web scrape weather data from wunderground website. Traffic crashes have a significant impact on the world economy and are a leading cause of death and injuries worldwide. More than one-half of all traffic deaths globally occur among people ages 15 to 44, their most productive earning years. This project is developed to understand feature correlations, patterns and analyse trends with the current weather conditions in the area of the accident. The data set contains about 2 million records. The next step involves transforming the data set by formatting date and time which is common in both the datasets so that we have familiar data about each incident. We also scraped weather data from wunderground for each weather station in New York and this data is then processed and mapped with the crash data set. This data is used for visualization and prediction analysis using machine learning models such logistic regression, SVM and Gradient Boosting, to predict the severity of a motor vehicle accident based on 4 classes such as very high, high, medium, low centred on weather and environmental factors such as temperature, precipitation rate, dew, wind speed and high-risk areas that can warn user based on historical data.

TABLE OF CONTENT

1. Introduction

The aim of this project is to analyse the motor vehicle accidents that have occurred in New York City over the past few years. The project involves exploring a large dataset consisting of more than 1.6 million records, each of which provides detailed information about the accident, including the date and time of the accident, the location, the type of vehicles involved, and the injuries sustained. In addition, the project aims to identify the types of injuries that are most common in motor vehicle accidents, such as injuries sustained by motorists, pedestrians, and cyclists. The analysis shows that motorist injuries are the most common, followed by pedestrian injuries, while cyclist injuries are the least common. The analysis of this data aims to provide insights into the factors that contribute to motor vehicle accidents in New York City, such as the time of day, the day of the week, the type of vehicle, and the location of the accident. One of the primary objectives of this project is to identify the boroughs of New York City that have the highest accident rates, as well as the types of vehicles and injuries that are most common in each borough. Another important aspect of the project is to identify the time periods during which accidents are most likely to occur. The analysis shows that accidents are most common during the peak travel hours, particularly between 4pm and 6pm when many workers are commuting home. This is often due to factors such as driver distraction and fatigue, which are more likely to occur at the end of the workday. The analysis also shows that certain types of vehicles are more likely to be involved in accidents than others. In addition, the project aims to identify the types of injuries that are most common in motor vehicle accidents, such as injuries sustained by motorists, pedestrians, and cyclists. The analysis shows that motorist injuries are the most common, followed by pedestrian injuries, while cyclist injuries are the least common.

Overall, this project aims to provide valuable insights into the factors that contribute to motor vehicle accidents in New York City, which can help policymakers and transportation officials to develop more effective strategies for reducing accident rates and improving road safety.

2.Software Requirement specification

2.1.1 Purpose:

To analyse traffic accidents in New York City and identify key trends and patterns in the data. The project aims to provide insights that can help policymakers, transportation planners, and other stakeholders improve road safety and reduce the number of accidents and injuries in the city. By analysing factors such as population density, car ownership, and time of day, the project seeks to shed light on the causes of accidents and suggest ways to prevent them. Ultimately, the goal of the project is to contribute to a safer and more sustainable transportation system in New York City.

2.1.2 Definitions, Acronyms, and abbreviations Acronyms

Definitions:

- Population density: the number of people living in a given area, usually measured as the number of people per square mile or square kilometre.
- Crash density: the number of crashes per unit of area, usually measured as the number of crashes per square mile or square kilometre.
- Car ownership: the percentage of households or individuals who own a car or other motor vehicle.
- Commute: the act of travelling to and from work or school.
- Peak travel hours: the times of day when the most people are commuting, typically the morning and evening rush hours.
- Driver inattention/distraction: a factor that contributes to many traffic accidents, often caused by using electronic devices, eating, or other distractions while driving.
- Fatigued/drowsy driving: a common cause of accidents, particularly during long commutes or driving late at night.

Acronyms and abbreviations:

- NYC: New York City
- SUV: Sport Utility Vehicle
- km/h: kilometre per hour
- mph: miles per hour
- %: percent (or percentage)
- EDA: Exploratory Data Analysis
- LR: Logistic Regression
- SVC: Support Vector Classifier
- GB: Gradient Boosting

2.1.4 Overview

The project is aimed at analysing and visualizing data related to road accidents in New York City. The analysis includes examining the factors that contribute to accidents, the types of vehicles involved, the time and location of accidents, and the severity of injuries and fatalities. The project aims to provide insights into the causes of accidents and to identify areas for improvement in road safety measures. The project will involve collecting and cleaning data from various sources, including the New York City Open Data Portal The data will then be analysed and visualized using various tools and techniques such as Python, Pandas, and Matplotlib. The project will also include the development of a web page that will provide interactive visualizations and allow to explore the data and gain insights into the factors contributing to road accidents in New York City. Overall, the project aims to provide a comprehensive analysis of road accidents in New York City and to use this analysis to inform policy decisions and improve road safety measures in the city.

List Of Functions:

S.no	Name	Function
1	Extract Weather Data	Web Scraping Weather Data
2	Extract Crash Data	Loading Crash Data
3	EDA of Weather Data	Analysing and finding trends in data
4	EDA of Crash Data	Analysing and finding trends in data
5	Merge data	Merge on basis of date and timestamp
6	Insertion in database	Pushing data into MongoDB Atlas
7	Machine Learning	Applying LR, SVC, GB classifiers
8	Visualisation	Creating HTML page with Power BI

2.2 Functional Description

- Extract Weather Data: This function involves web scraping the weather underground website using python and extracting and formatting data from JSON to CSV for the year 2016 to 2023 using BeautifulSoup. Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

- Extract Crash Data: This function involves the collection of data related to traffic accidents in the New York City NYC Open Data website from the year 2012 to 2023.

- EDA of Weather Data: The data needs to be analysed to identify patterns, trends, and insights related to weather attributes such as wind speed , visibility, precipitation rate, condition, temperature etc.

- EDA of Crash Data: The data needs to be analysed to identify patterns, trends, and insights related to traffic accidents in New York City with respect to injuries, kills, type of vehicle ,region, time of the day etc

- Merge data: This function deals with creating a single column by combining date and timestamp to create a unique key in both the data set and performing a join by using PySpark. PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time,

large-scale data processing

- Insertion in database: This function deals with creating the clusters on MongoDB Atlas, the Cloud-Native Document Database as a Service for Easiest Way to Deploy, Operate, and Scale MongoDB in the Cloud. Atlas simplifies deploying and managing your databases while offering the versatility and availability you need to build resilient and performant global applications on the cloud providers of your choice

- Machine Learning: This function deals with creating the target variable “severity level” , applying machine learning techniques to do classification tasks based on the created target variable. The formula for severity level is calculated on the basis of “['NUMBER OF MOTORIST KILLED']*3 +['NUMBER OF MOTORIST INJURED']*1” resulting in severity levels of low, medium, high, very high

- Visualisation: The analysed data needs to be presented in a visually appealing way using graphs, charts, and maps to make it easier for users to understand and draw insights from the data using a HTML page with PowerBI.

2.3 Specific Requirements

2.3.1 Software interfaces:

- python
- matplotlib
- sklearn
- pandas
- jupyterlab
- PyMongo
- chrome
- HTML
- PySpark
- Power BI
- MongoDB

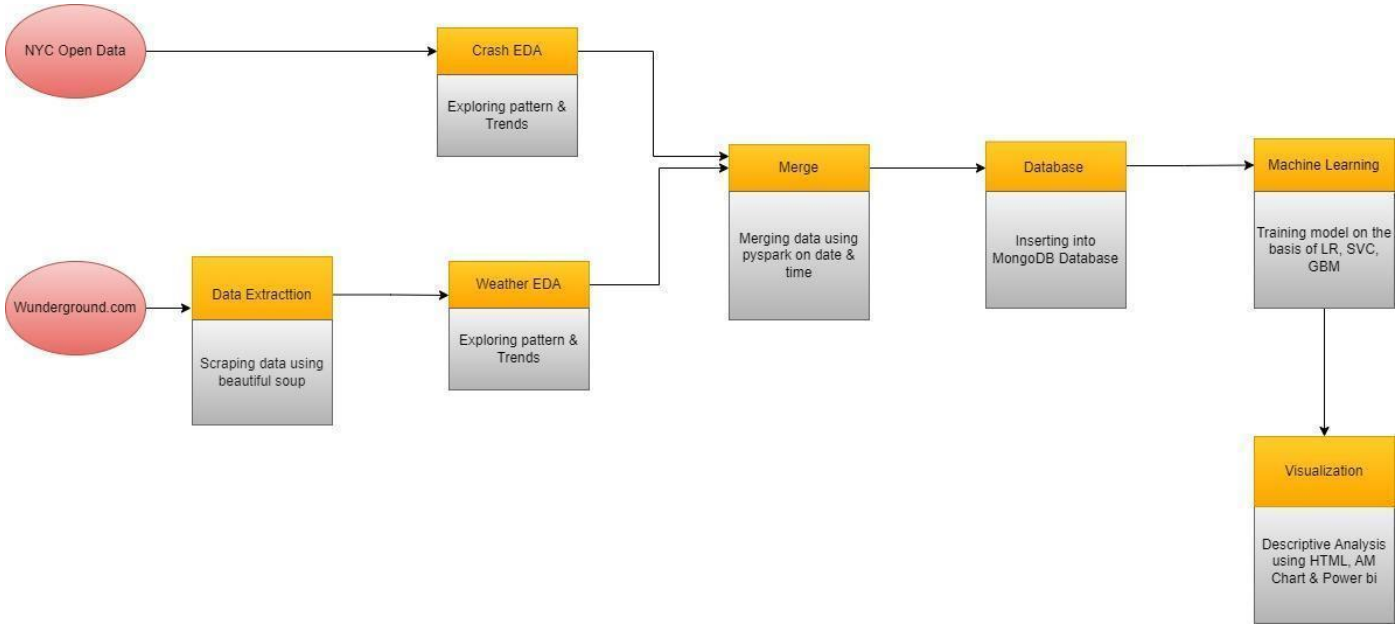
2.3.2 Hardware Interfaces

- 8GB RAM
- 512GB HDD
- Intel core i5 processor
- Intel Graphics card

3.Architecture

3.1 Flow diagram

The steps involve Scraping weather data from wunderground for NYC using BeautifulSoup. Doing an EDA on Weather Data and Crash Data and finding patterns. Processing and mapping the crash dataset with a weather dataset using PySpark. Using MongoDB Atlas for data storage. ML models like logistic regression, SVM and Gradient Boosting are evaluated to predict severity of a vehicle crash. For Visualization Power BI and AMCharts were used in a webpage.



4.Dataset Extraction

4.1 Weather Data

This function involves web scraping the weather underground website using python and extracting and formatting data from JSON to CSV for the year 2016 to 2023 using BeautifulSoup. Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.Total no of rows in the extracted dataset is 68030 and total no of columns are 17. Column names are: date, timestamp, time_of_day, temperature, condition, pressure, dew_point, humidity, visibility_in_miles, wind_chill_temp, wdir_cardinal, wind_gust, wind_speed, maximum_temp, minimum_temp, precipitation_rate, snow_rate.

4.2 Crash Data

This function involves the collection of data related to traffic accidents in the New York City NYC Open Data website from the year 2012 to 2023.
Total no of rows in the extracted dataset is 1972605 and total no of columns are 29. Column names are 'crash date', 'crash time', 'borough', 'zip code', 'latitude', 'longitude', 'location', 'on street name', 'cross street name','off street name', 'number of persons injured','number of persons killed', 'number of pedestrians injured', 'number of pedestrians killed', 'number of cyclist injured','number of cyclist killed', 'number of motorist injured','number of motorist killed', 'contributing factor vehicle 1','contributing factor vehicle 2', 'contributing factor vehicle 3','contributing factor vehicle 4', 'contributing factor vehicle 5','collision_id', 'vehicle type code 1', 'vehicle type code 2','vehicle type code 3', 'vehicle type code 4', 'vehicle type code 5'.

```python
def get_api_key(link):
    """Get the API key from wunderground
    Args:
        link (str): website link
    Returns:
        api_key: API key
    """
    response = requests.get(link)
    html_doc = response.text
    soup = BeautifulSoup(html_doc, 'html.parser')
    scrp = soup.find_all('script', id="app-root-state")
    str1 = 'SUN_API_KEY&q;:&q;'
    res1 = re.findall(str1+'(.*)', str(scrp))
    api_key = res1[0].split('&q;')[0]
    return api_key

d=get_api_key("https://www.wunderground.com/history/daily/us/ny/new-york-city/KJFK/")
print(d)
```

Python

5. Dataset Exploratory Data Analysis

EDA of Weather Data: The data needs to be analysed to identify patterns, trends, and insights related to weather attributes such as wind speed , visibility, precipitation rate, condition, temperature etc.Creating column for day name using “day_name()” function on “date” column. Creating column for month using “strftime()” function similarly extracting year ino a new column using the same function. “crash_time” column was formatted with the help of applying a lambda function to strip the last three characters so that it gives only hours instead of hours and min.

EDA of Crash Data: The data needs to be analysed to identify patterns, trends, and insights related to traffic accidents in New York City with respect to injuries, kills, type of vehicle ,region, time of the day etc. Using day_name() created a new column “Day_Name” to find distribution across seven days of the week. “crash_time” column was formatted with the help of applying a lambda function to strip the last three characters so that it gives only hours instead of hours and min. No of crashes were calculated by grouping “day_name” and “crash_time” Distribution of crashes over area were extracted by using value_counts() on
column “borough”.Distribution of crashes over vehicles were calculated using value_counts() on column “Vehicle_type”.Year has been extracted from column “Crash_date” and no of crashes per year were calculated

6. Dataset Merge

This process deals with creating a single column by combining date and timestamp to create a unique key in both the data set and performing a join by using PySpark. PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, large-scale data processing. Apache Spark is written in Scala programming language. PySpark has been released in order to support the collaboration of Apache Spark and Python, it actually is a Python API for Spark. In addition, PySpark, helps you interface with Resilient Distributed Datasets (RDDs) in Apache Spark and Python programming language.

7. Insertion in Database

To enable further processing of a merged dataset that combined data from two different sources, a scalable and flexible data storage and query solution was needed. MongoDB Atlas was selected for this purpose, with a cluster set up and data uploaded to a collection via PyMongo. To enhance the upload process, the PySpark dataframe was first converted to a dictionary format, which simplified interactions with MongoDB's document-based storage model.

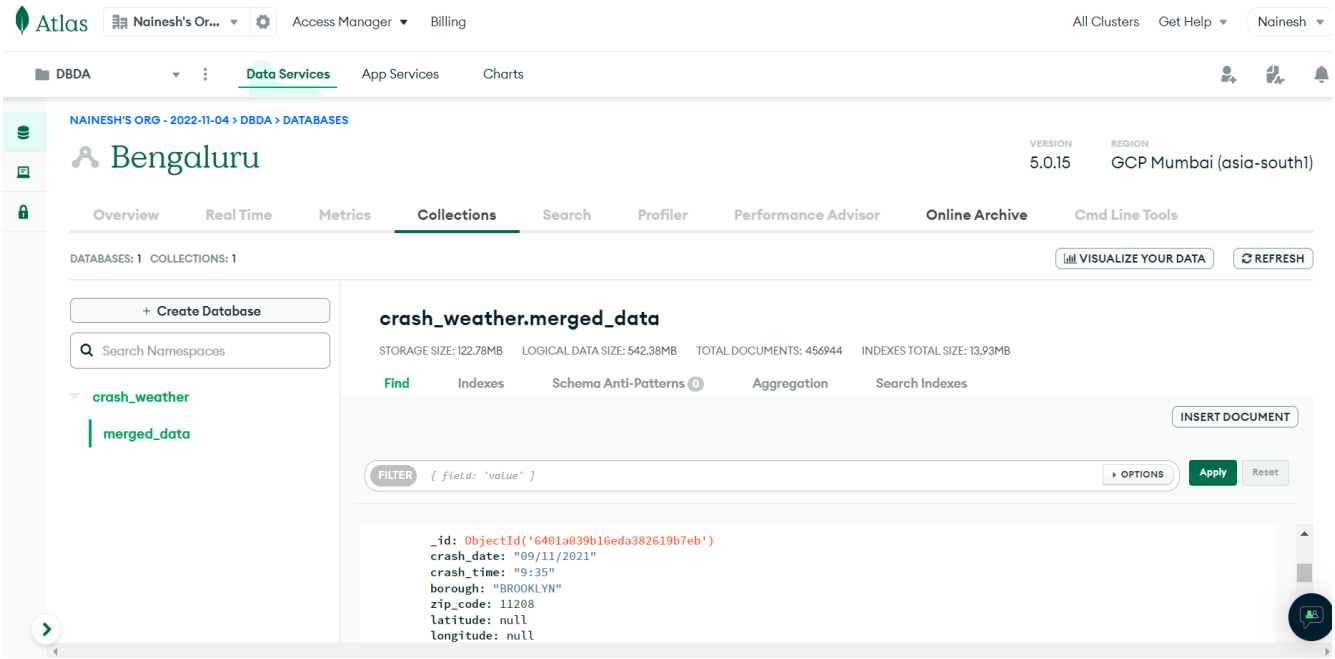
The uploaded data was utilized for machine learning purposes, including data transformations and feature engineering. The document model of MongoDB provided the flexibility to store and query complex data structures, which was essential for the project's data analysis and machine learning tasks.

By utilizing MongoDB and PySpark in together, large datasets can be effectively managed and analyzed without performance or scalability issues. The combination of these technologies enabled the project to take advantage of their strengths and optimize data analysis workflows.

here's how to push and pull data from the MongoDB Atlas cluster:
connect to instance :
client = pymongo.MongoClient("<link_to_cluster>")
db = client['db_name']
collection = db['collection_name']

to push data:
data = [row.asDict() for row in merged_data.collect()]
collection.insert_many(data)

to load data:
df = pd.DataFrame(list(collection.find({},{'_id':0})))



8.1 Target Variable

In our project we used following ML algorithms to build model which classifying severity level. The formula for severity level is calculated on the basis of "['NUMBER OF MOTORIST KILLED']*3 +['NUMBER OF MOTORIST INJURED']*1" resulting in severity levels of low, medium, high, very high

```python
def severity_calculate(x):
    if (x > 3):
        return 'Very High'
    elif (x > 2):
        return 'High'
    elif (x > 1):
        return 'Medium'
    else:
        return 'Low'

final_df['hazard_level'] = final_df['helper1'].apply(severity_calculate)
del final_df['helper1']
```

Python

Severity	Killed	Injured	Severity condition
Low	0	1	At least one injured
Medium	0	2	At least two injured
High	1	3	Either one killed or more than two injured
Very high	>1	>3	Either more than one killed or more than 3 injured

8.2 Machine Learning Algorithms

1) Logistic Regression- Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. By using hyperparameter tuning we set max_iter parameter, used the model parameters to minimize the logistic loss function using an optimization algorithm such as gradient descent.

2) Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. We did hyperparameter tuning using
learning rate, max depth, n_estimators and other parameters.

3) Support vector machine (SVM)- Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier detection. The advantages of support vector machines are: Effective in high dimensional spaces. Still effective in cases where the number of dimensions is greater than the number of samples. In which hyperparameter tuning we did using gamma, kernel, class weight parameters.

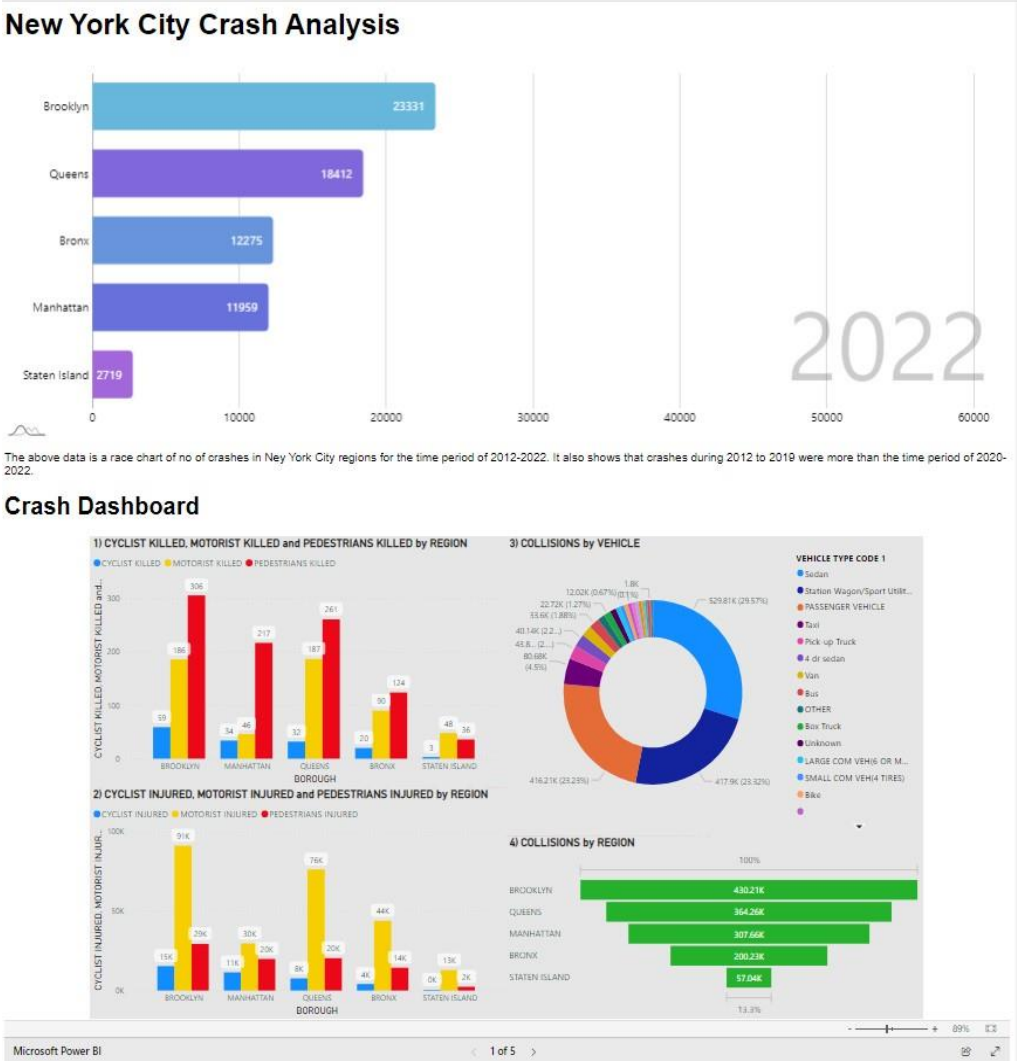
8.2.1 Machine learning metrics

Machine learning algorithms	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9590	0.9404	0.9590	0.9489
Gradient Boost	0.9568	0.9155	0.9568	0.9357
SVM	0.92			

9.Visualisation

Visualization lets you comprehend vast amounts of data at a glance and in a better way. It helps to understand the data better to measure its impact on the business and communicates the insight visually to internal and external audiences. The analysed data needs to be presented in a visually appealing way using graphs, charts and maps to make it easier for users to understand and draw insights from the data.

For this project we have created a frontend using HTML , Amcharts and PowerBI. HTML and CSS is used to create the main page. AM charts is used to create a race chart displaying 10 years of data from 2012 to 2022 across all five regions of New York City and a heat map for displaying distribution of no of crashes during the week across time. PowerBI has been embedded into HTML using an inline frame to display the dashboard of weather and crash data.



Frontend link: https://nyc-crash-analysis.netlify.app/

10.Conclusion

New York, often called New York City or NYC, is the most populous city in the United States. With a 2020 population of 8,804,190 distributed over 300.46 square miles (New York City is sometimes referred to collectively as the Five Boroughs. The names of the Boroughs(Regions) are Bronx, Brooklyn, Manhattan, Queens and Staten Island. From the descriptive analysis we come to find that crashes before 2019 were much higher than the time period after with Brooklyn being the most dangerous region as it has contributed the most no of crashes amongst all other regions for the last decade. Staten Island has the lowest no of crashes and can be regarded as the safest region amongst all the regions in NYC.

From the crash statistics we can see that more pedestrians are injured compared to motorists and cyclists whereas more motorists are killed than cyclists and pedestrians. Sedan and SUV type of vehicle contribute to almost 50% of the crashes. If we compare the injuries and kills because of a collision we came to know that injuries are more likely to happen in a crash than a kill. When we see the distribution of crashes over the week we can find the weekdays has more no of crashes than the weekends with the working time i.e. 8am to 6pm contributing more no of crashes than any other time of the day. On Weekends we can also see that more crashes occur from 2pm to 6 pm. On Weekdays most of the crashes have occurred at 5pm on Fridays in the afternoon and least of the crashes have occurred on Tuesdays early morning at 3am.

From the weather data we can see that there is inverse relationship between humidity and “visibility in miles” being afternoons very humid the most no of crashes also occurs in the afternoon thus this makes humidity one of the maj
or cause of concern for accidents. Seeing the distribution of months we can see July and August being most humid and these months can be more prone to accidents in the year. When observing the weather conditions we can see that “Mostly Cloudy ” and “Cloudy” weather conditions are a majority more than others. If we see the distribution of humidity across the given time period we can see that humidity has peaked at 2018 and from the crash data we can also see that crashes were more before 2019 .

11 References

11.1 Websites

- https://opendata.cityofnewyork.us/
- https://www.wunderground.com/
- https://learn.microsoft.com/
- https://www.amcharts.com/
- https://www.mongodb.com/docs/atlas/getting-started/
- https://www.mongodb.com/docs/drivers/python/

11.2 Books

- Python Crash Course: A Hands-On, Project-Based Introduction to Programming by Eric Matthes
- The Spark for Python Developers by Amit Nandi
- MongoDB Fundamentals: A hands-on guide to using MongoDB and Atlas in the real world
- Mastering Microsoft Power Bi: Expert techniques for effective data analytics and business intelligence