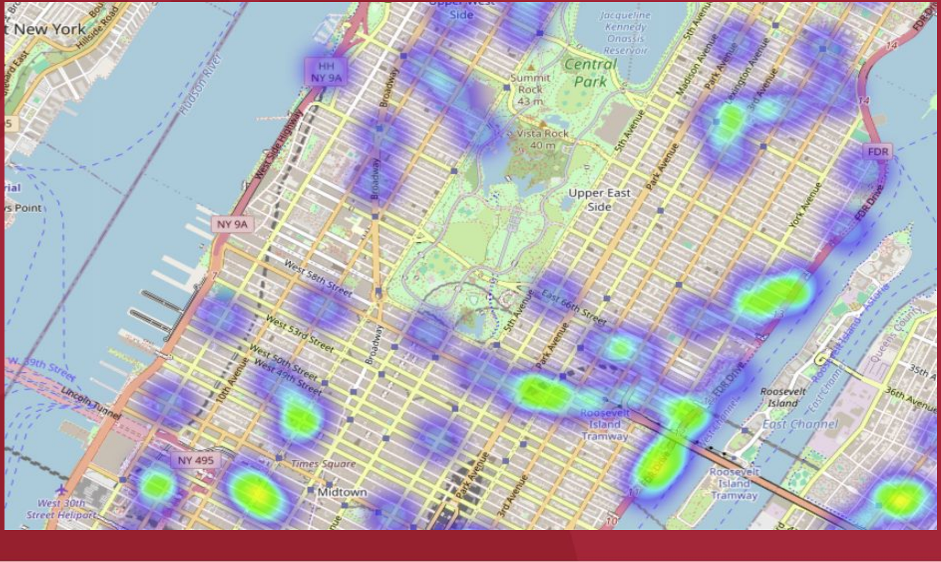
MA 541 - Statistical methods

STEVENS
INSTITUTE OF TECHNOLOGY
1870

STATISTICAL ANALYSIS OF VEHICLE COLLISIONS



AGENDA

1. Objective
2. Data Description
3. Data Pre-Processing
4. Data Visualization/ Descriptive Statistics
5. Inferential Statistics
   a. Maximum Likelihood Estimator
   b. Bayesian Approach
   c. ANOVA
   d. Tukey's HSD
   e. Chi Square Test
6. Linear Regression
7. Future work
8. Conclusion

OBJECTIVE

Project aims to provide a comprehensive understanding of motor vehicle collisions in NYC. Through rigorous statistical analysis, we hope to draw insights that can guide future safety protocols and interventions to reduce accidents and improve public safety.

- Identify areas (by geography and time) with a high frequency of collisions, helping in targeted interventions.
- Determine the leading causes of collisions and how they differ across various boroughs or vehicle types.
- Linear Regression analysis to predict the number of people injured.

DATA DESCRIPTION

- Source: NYC motor vehicle collisions database.

  https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes
- Data Duration - 2017 to 2021
- Data contains 2 million Data
- Features included -

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

DATA DESCRIPTION

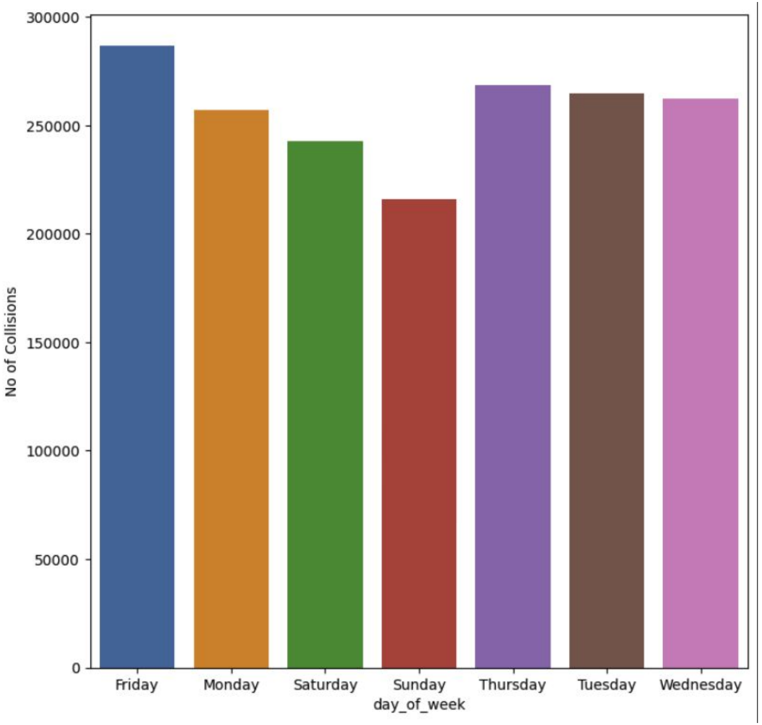
MODIFIED FEATURES :

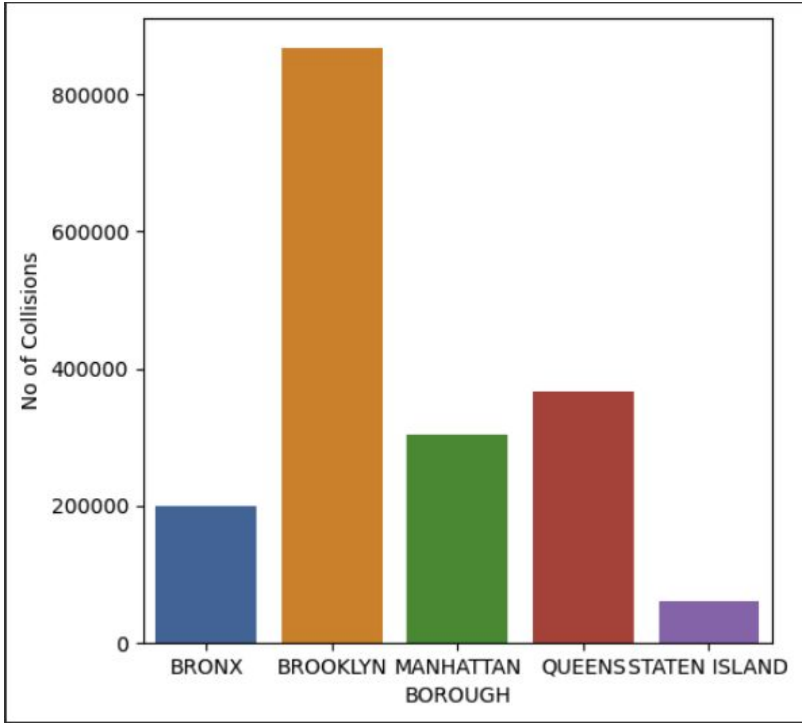
- NUMBER OF PERSONS INJURED = NUMBER OF PERSONS INJURED + NUMBER OF PEDESTRIANS INJURED +NUMBER OF CYCLIST INJURED + NUMBER OF MOTORIST INJURED
- NUMBER OF PERSONS KILLED = NUMBER OF PERSONS KILLED + NUMBER OF PEDESTRIANS KILLED +NUMBER OF CYCLIST KILLED + NUMBER OF MOTORIST KILLED
- Extracted Date , Hour, day of week, Year from Crash Date and Crash time columns.

DATA PRE-PROCESSING

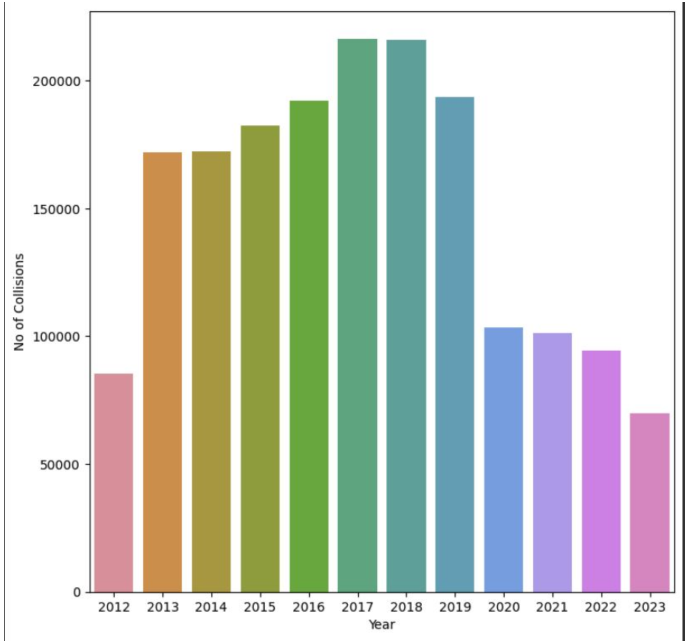
- Duplicate Removal
    - Ensured uniqueness of data by removing duplicate entries.
- Feature Completeness
    - Excluded data points with insufficient feature data to maintain data quality.
- Consistency Check
    - Eliminated data that did not align with the intended feature representations.
- Null Value Management
    - Dropped the rows where Location, latitude, longitude are null.
- Focused Data Selection
    - Dropped rows lacking essential information.

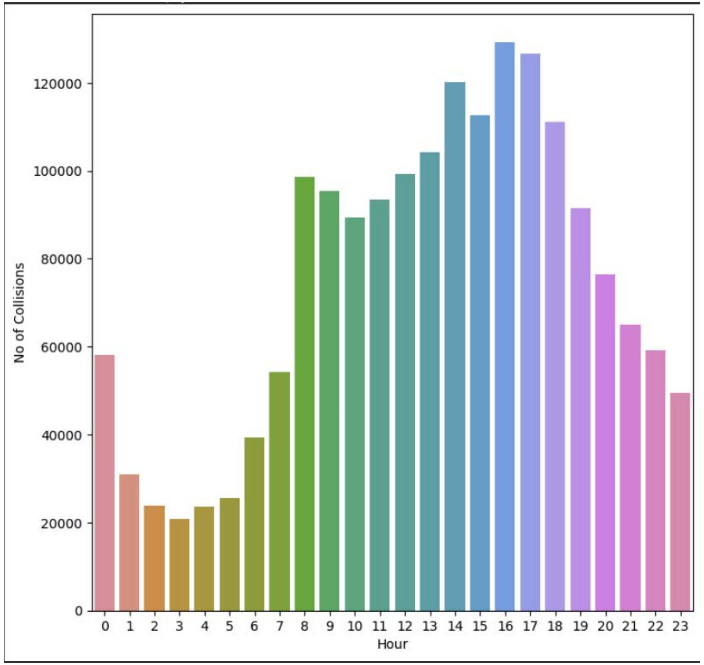
DATA VISUALIZATION



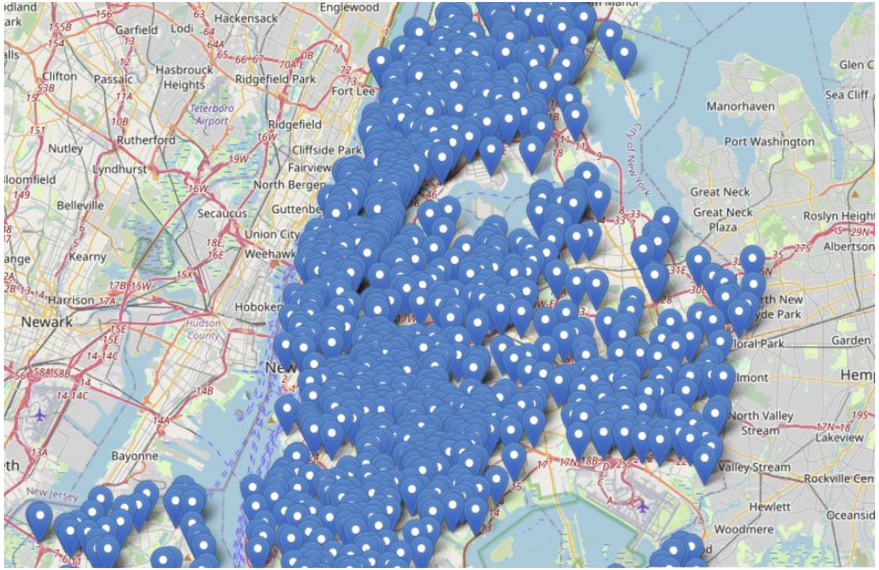


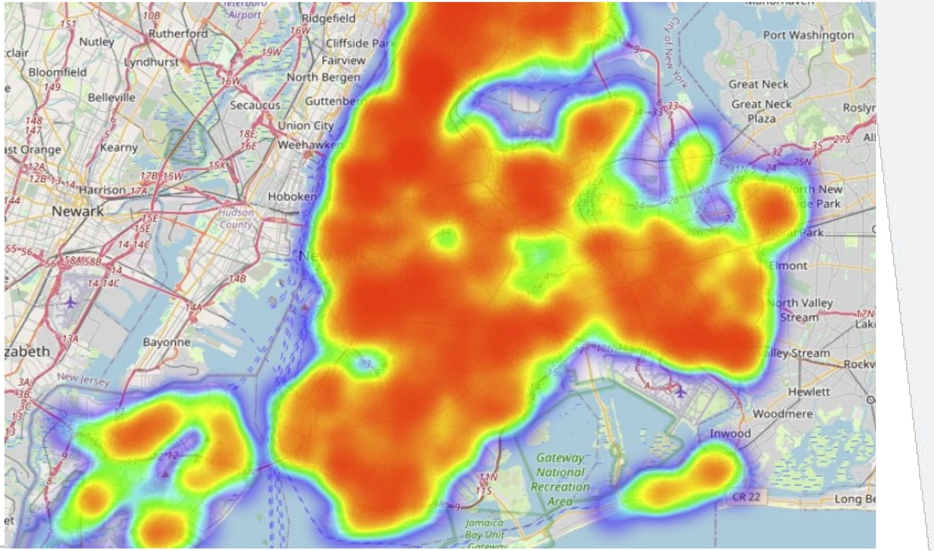
DATA VISUALIZATION



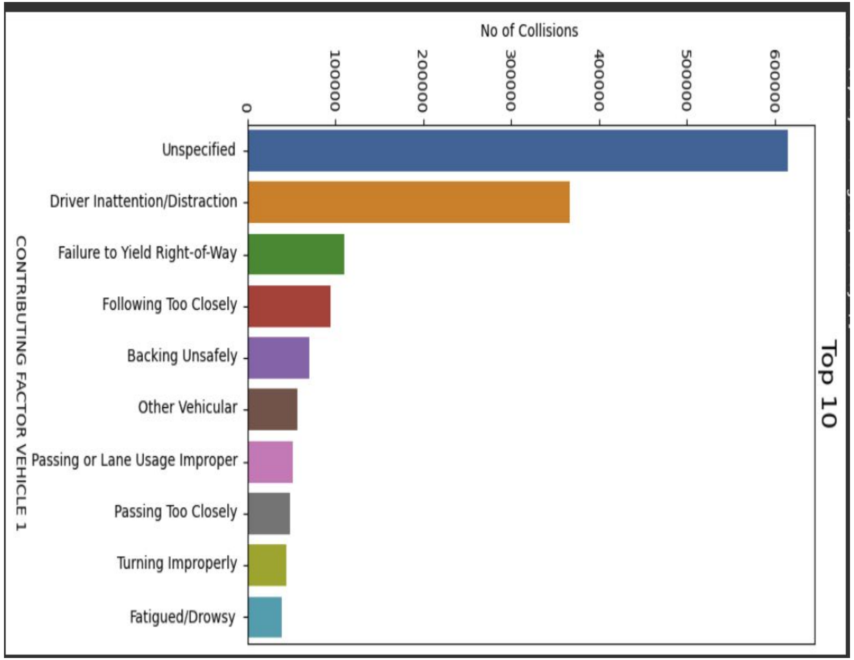


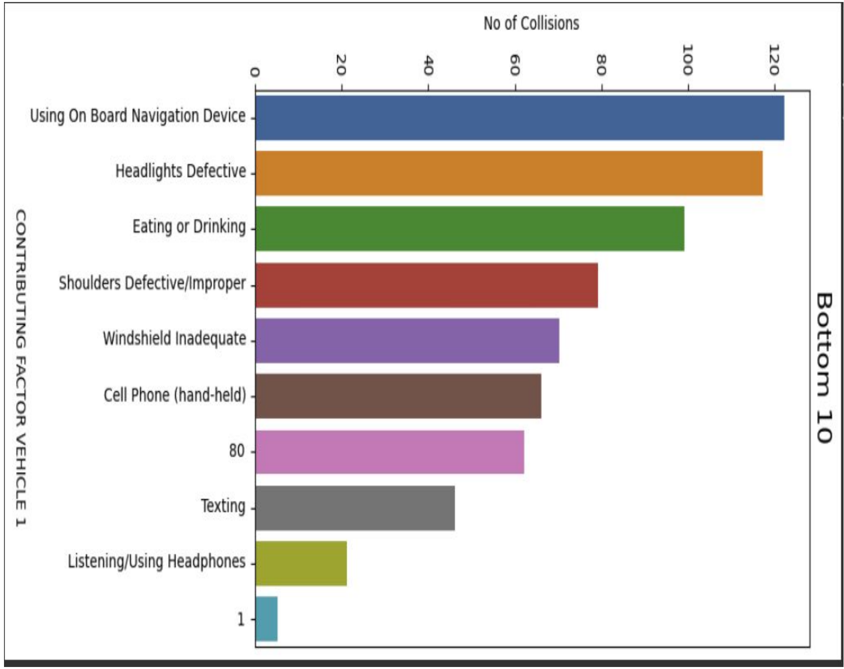
DATA VISUALIZATION



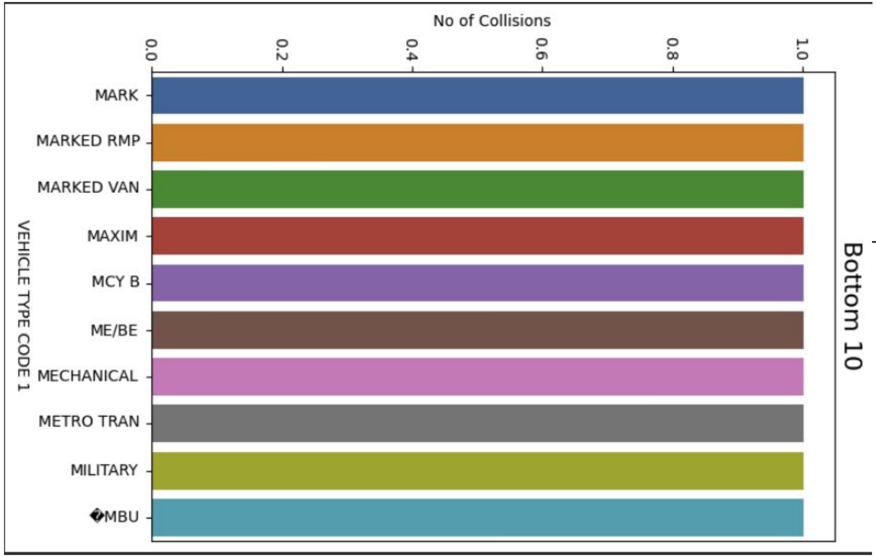


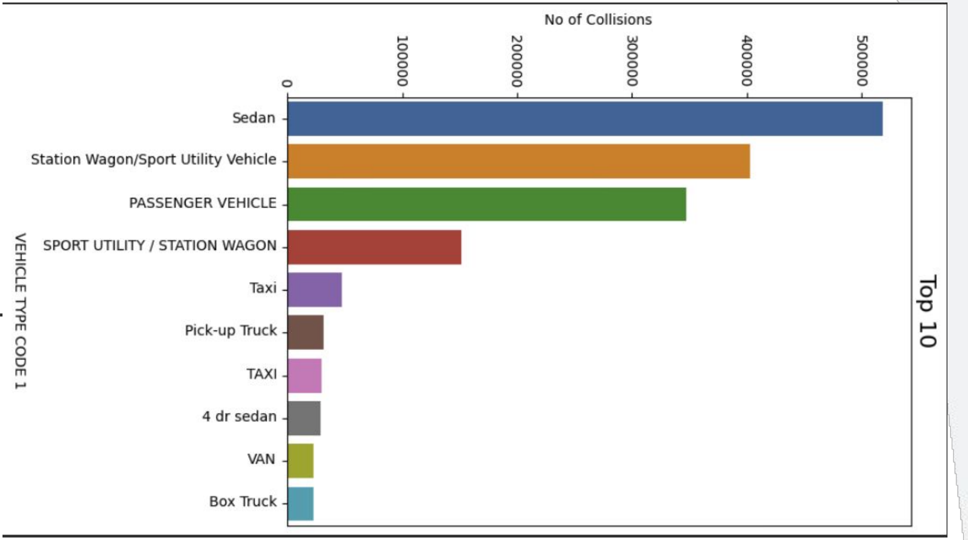
DATA VISUALIZATION





DATA VISUALIZATION





DATA VISUALIZATION

CONCLUSION :

- It seems the number of collisions are the highest on Friday and Least on a Sunday.
- Brooklyn seems to be most prone to collisions and Staten Island seems to be the safest.
- There was a significant rise in collisions in 2013. Then, it was steadily increasing until 2019. Then there was a sharp decrease in collisions in 2020. Most probably due to COVID restrictions.
- Most collisions occur between the time of 4:00 pm to 5:00 pm

INFERENTIAL STATISTICS

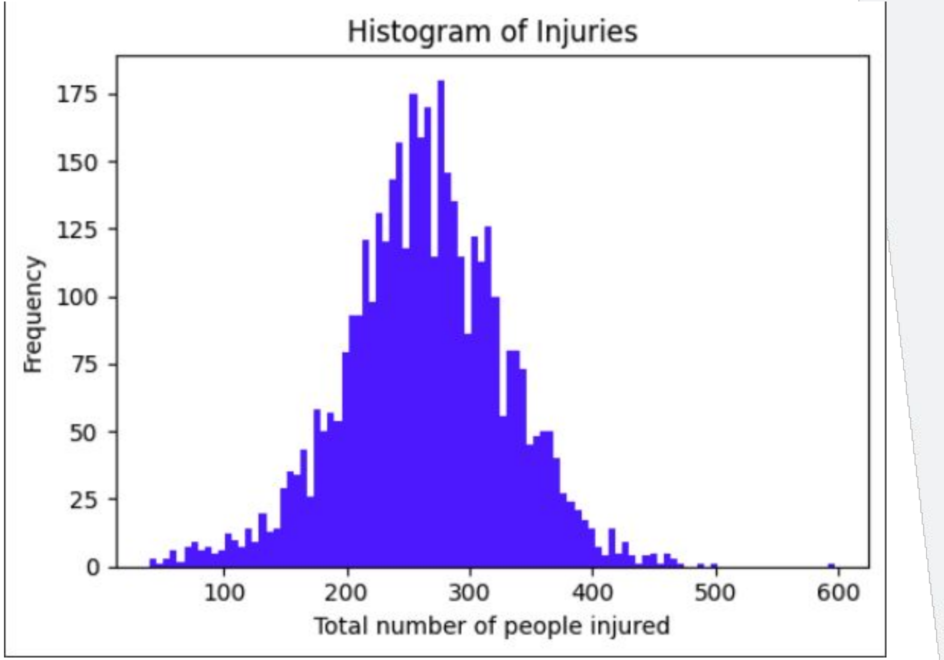
MAXIMUM LIKELIHOOD ESTIMATOR :

- **Daily Injury Estimate**: The MLE suggests that, on average, there are approximately 264 people injured every day due to motor vehicle collisions. This number provides a baseline for understanding the scale of traffic-related injuries.
- **Variability of Injuries**: The standard deviation from the MLE is about 65 injuries, indicating a considerable day-to-day variation in the number of injuries. Some days might have significantly more than 264 injuries, while others might have less.

INFERENTIAL STATISTICS

BAYESIAN APPROACH:

Normality in Bayesian Estimation: In large datasets, both the likelihood and prior can be assumed normal, leading to a normally distributed posterior, which aligns with the central limit theorem ensuring that sample means approximate a normal distribution, regardless of the original data's shape.



INFERENTIAL STATISTICS

BAYESIAN APPROACH:

1. **Confident Estimation of Daily Injuries**: The Bayesian approach has provided a posterior mean estimate of approximately 263.56 for the average daily number of people injured in motor vehicle collisions. This figure reflects a combination of prior knowledge and observed data, offering a confident estimate for stakeholders to consider when assessing the impact of road traffic injuries on public health resources.
2. **Low Variability in Daily Injury Counts**: The posterior standard deviation is approximately 0.99, indicating low variability around the mean estimate. This suggests that the number of daily injuries does not fluctuate widely, which allows for more reliable planning in healthcare resource allocation and emergency response services.
3. **Data-Driven Decision Making**: The close alignment between the prior belief and the sample data underscores the robustness of the current safety measures and strategies. It also highlights the potential for data-driven decision-making in public health and traffic safety to further reduce the incidence of injuries from motor vehicle accidents.

INFERENTIAL STATISTICS

ANALYSIS OF VARIANCE(ANOVA) :

1. **Null hypothesis** : Total number of people injured across boroughs is same.
2. **Alternative hypothesis** : Total number of people injured across boroughs is different.
3. 'Total_number_of_people_injured' - This is the continuous variable representing the count of people injured in motor vehicle collisions. Independent Variable (Factor): 'BOROUGH'
4. ANOVA results reject the null hypothesis, indicating significant differences in injury counts across boroughs.
5. Borough-specific variations call for tailored road safety and healthcare resource allocation.
6. Post hoc analysis is required to identify which boroughs differ significantly in injury rates.
7. Location-specific insights are crucial for targeted interventions and policy-making.

INFERENTIAL STATISTICS

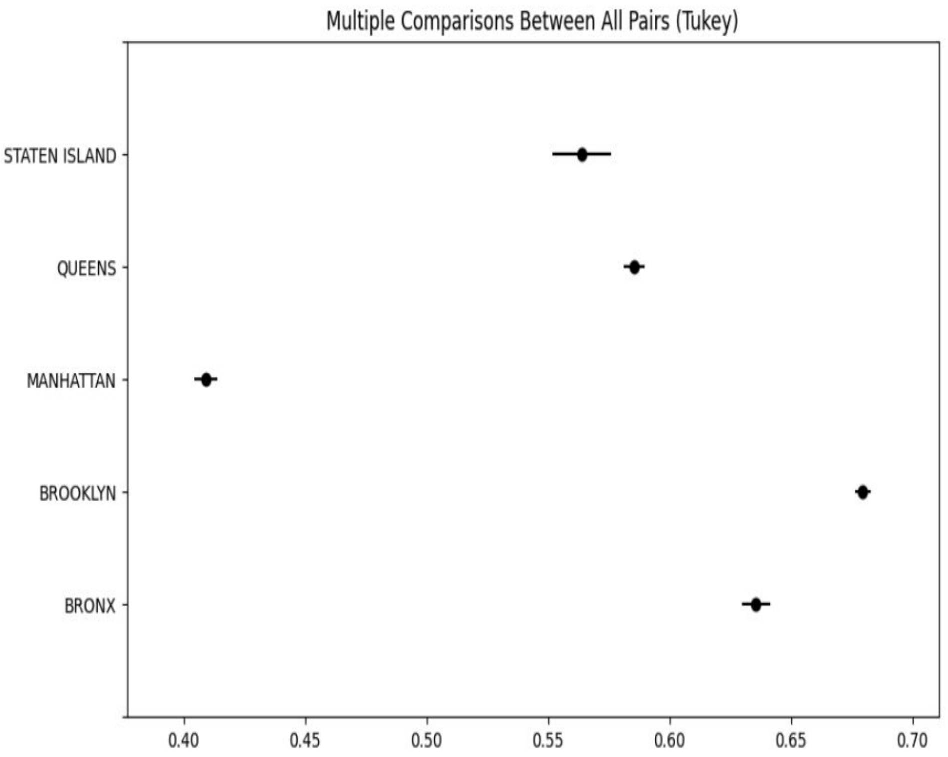
TUKEY's HSD

1. For each pair of boroughs compared by Tukey's HSD, the

null hypothesis : "The mean number of people injured in each borough are equal."

Alternate hypothesis : "The mean number of people injured in each borough are not equal.."

2. Tukey HSD confirms significant differences in injuries across boroughs, invalidating the hypothesis of uniform injury rates.
3. Location-specific factors are implicated in the variance of collision severity.
4. The Brooklyn shows a higher mean injury count, necessitating targeted safety measures.
5. Visual data representation underscores the need for borough-specific road safety policies.



INFERENTIAL STATISTICS

CHI-SQUARE TEST :

1. **Null hypothesis:** "There is no association between the boroughs in New York City and the contributing factors to vehicle collisions."
2. **Alternative hypothesis:** "There is an association between the boroughs in New York City and the contributing factors to vehicle collisions.”
3. With a Chi-square statistic of approximately 75325.13 and a p-value of 0.0, the results are highly significant. This allows us to reject the null hypothesis and conclude that there is a statistically significant association between the boroughs and the contributing factors to vehicle collisions.
4. The degrees of freedom (df) for the test is 240, likely reflecting a large number of categories within the two variables.

LINEAR REGRESSION

1. **Quantifying Relationships:** To quantify the relationship between the time of the incident (hour of the day, day of the week) and the location (borough) with the total number of people injured in collisions.
2. **Predictive Analysis:** To create a model that could potentially be used to predict the number of injuries that might occur at a given time and place, aiding in resource planning and emergency response.
3. **Influence Assessment:** To assess the influence of different times and locations on the severity of collisions as measured by the number of injuries.

LINEAR REGRESSION

Intercept: 0.6208861959480656
Slopes: [-0.04742112 -0.00723824 0.00483017]

Y = -0.047X1 -0.0072X2 + 0.0048X3 + 0.62088

Where Y is no.of people injured

X1 is Borough

X2 is Day of the week

X3 is Hour at the time of incident

The resulting regression coefficients tell us the expected change in the response variable (total number of people injured) for a one-unit change in each predictor variable, assuming all other variables are held constant. The intercept represents the baseline number of injuries when all other variables are zero, and the slopes reflect the changes associated with each borough, day of the week, and hour of the day.

FUTURE WORK

Time Series Analysis: Analyze the data as a time series to identify trends, seasonality, and cyclic patterns in the number of injuries over time. This could involve decomposing the series into its components and using models like ARIMA for forecasting.

Survival Analysis: For cases where the time until an event (such as time until a collision from a reference point) is of interest, survival analysis could be applied to study the time-to-event data.

Cluster Analysis: Perform cluster analysis to identify patterns and groupings among incidents without predefined categories, which might reveal hidden structures in the data.

Text Analysis: If the dataset includes narrative reports or textual data, natural language processing (NLP) techniques could be used to extract additional insights from unstructured data.

CONCLUSION

1. Analysis confirms significant impact of borough and time on collision injuries, suggesting targeted timing and location for interventions.
2. Contributing factors to collisions vary by borough and confirmed by Chi Square Test.
3. MLE and Bayesian methods effectively estimate daily injury counts, offering a data-driven approach to emergency planning.
4. The number of persons injured vary by Borough and proved by ANOVA and Tukey’s HSD tests.

STEVENS
INSTITUTE OF TECHNOLOGY
1870

THANK YOU

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030