
DTU

Muhammad Numan Bashir s204465
Vaneeza Fatima Butt s205835
Mostafa Adnan Al-Gharifi s205425
Mikail Oguzhan Kocak s195205

Group 13

Introduction To Project 1 - NYC Taxi Rides

- **Brief overview of the project's purpose**
  This project aims to analyze and forecast taxi ride patterns in New York City to enhance service efficiency, inform urban planning, and improve passenger experience.
- **Importance of analyzing and forecasting taxi rides in NYC**
  Analyzing and forecasting NYC taxi rides is crucial for optimizing urban mobility, supporting the local economy, and guiding effective transportation policies.

Data Overview

- **Yellow and Green Taxi trip records**
  The project uses Yellow and Green Taxi trip records from New York City, which include details on trip times, locations, distances, fares, and passenger counts.
- **The types of data collected:**
  The data collected includes pick-up and drop-off times, locations, trip distances, fare amounts, types of rate, payment methods, and passenger counts for NYC's Yellow and Green Taxis.

Data Preparation

- **The data loading process:**
  The data loading process involves fetching Yellow and Green Taxi trip records for specified months in 2022 using Python and pandas, and then concatenating them into single dataframes for each taxi type. We did this using (df_yellow.head() and df_green.head()). (See notebook for table reference).

Exploratory Data Analysis (EDA) - Part 1

- **Data cleaning steps**
  1. Removed Negative Values: Excluded entries with negative fares and trip distances.
  2. Handled Outliers: Identified and removed extreme values in fares and distances using the Interquartile Range method.
  3. Consistency Checks: Ensured pick-up times were before drop-off times.

Basic descriptive statistics

Table: Yellow Taxi Data Descriptive Statistics

Statistic	Trip Distance	Fare Amount	Passenger Count
Count	7.821078e+06	7.821078e+06	7.578304e+06
Mean	1.875273e+00	1.032960e+01	1.393165e+00
Standard Dev.	1.167014e+00	4.838808e+00	9.618065e-01
Minimum	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.000000e+00	6.500000e+00	1.000000e+00
50% (Median)	1.600000e+00	9.300000e+00	1.000000e+00
75%	2.500000e+00	1.300000e+01	1.000000e+00
Maximum	5.510000e+00	3.200000e+01	9.000000e+00

Table: Green Taxi Data Descriptive Statistics

Statistic	Trip Distance	Fare Amount	Passenger Count
Count	181526.000000	181526.000000	165255.000000
Mean	2.134229	11.701958	1.287943
Standard Dev.	1.480277	5.911292	0.936957
Minimum	0.000000	0.000000	0.000000
25%	1.080000	7.500000	1.000000
50% (Median)	1.800000	10.500000	1.000000
75%	2.920000	15.000000	1.000000
Maximum	6.540000	34.020000	8.000000

EDA - Part 2 (Visualizations)

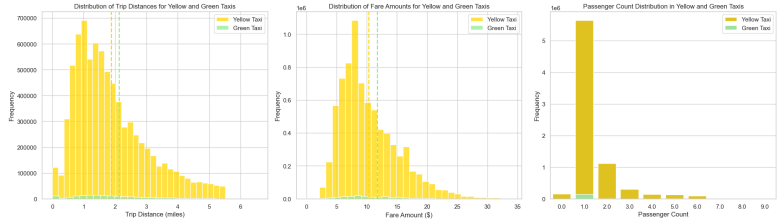

Figure: Histogram showing distribution of trip distances and fares.

The histograms show that most taxi trips are short, with fares usually under $20, most of them being around $8. Also each ride is mostly around 1-2 miles and they are also used mostly by single passengers. This indicates that taxis are primarily used for short, solo trips in the city.
Green taxis appears to be a little more expensive than yellow ones, and they are used far less than yellow taxis.

Spatial Analysis

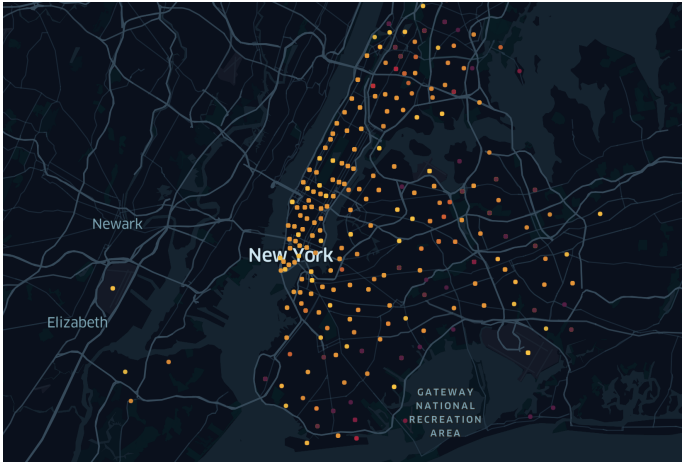

Figure: Spatial map over NYC taxi pickup and drop-off.

Spatial Analysis

The map displays clusters of taxi pickups and drop-offs in New York City, with dense areas indicating hotspots of high taxi demand, particularly in central Manhattan. This suggests that taxis are frequently used in busy commercial and residential districts, guiding efficient fleet distribution and service planning.

Temporal Analysis - Yellow Taxis

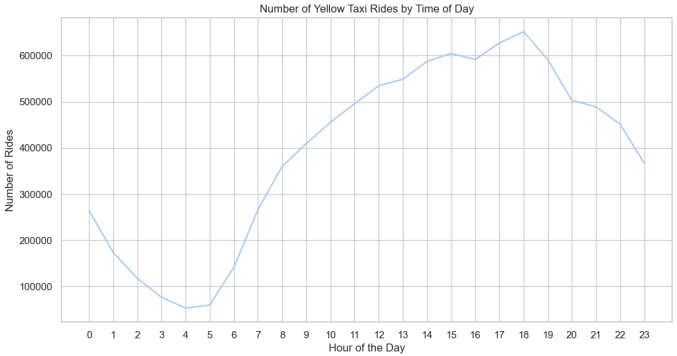

Figure: Yellow taxi: Graph for number of rides by time of day.

Temporal Analysis - Green Taxis

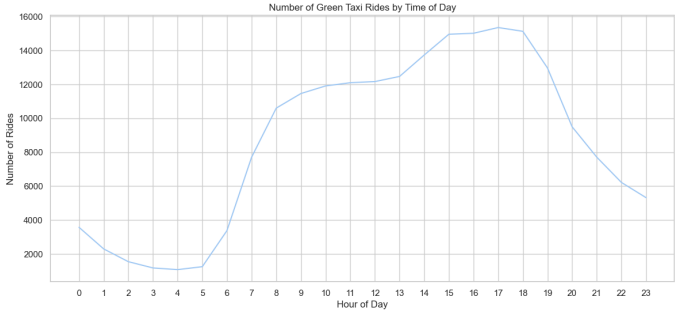

Figure: Green taxi: Graph for number of rides by time of day.

Temporal Analysis of Yellow and Green Taxis

In the Yellow Taxi graph, we see a clear spike in rides around evening time, hinting at a common trend of evening travel, possibly from work to home. On the other hand, the Green Taxi graph shows a steadier flow of rides throughout the day, with a slight increase in the evening, suggesting a more varied use of Green Taxis at different times of the day. As it showcases differences over time the display should reflect that and a line graph was therefore chosen with a light blue colour making it visually appealing.

Time-Series Forecasting - Model Building Yellow Taxis

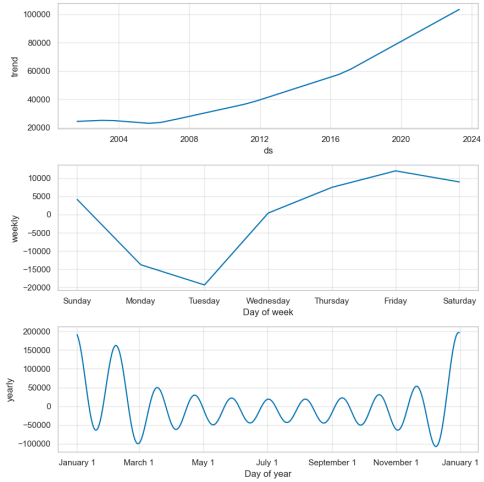

Figure: Forecast of yellow taxi.

Time-Series Forecasting - Model Building Green Taxis

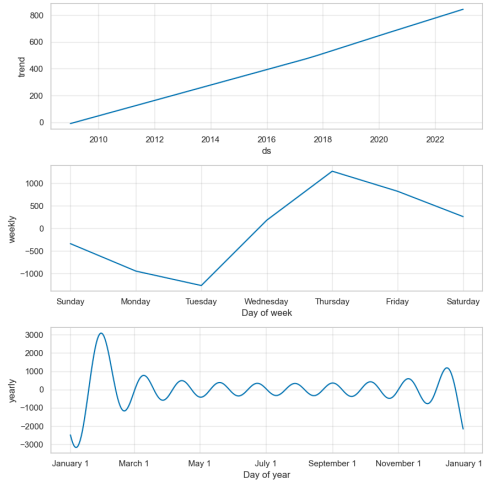

Figure: Forecast of green taxi.

Time-Series Forecasting Analysis

Trend: The first graph shows the long-term trend for each service. For Yellow Taxis, there's a clear upward trend, indicating increasing usage over the years. The Green Taxis also show an upward trend, but with less steepness compared to Yellow Taxis, suggesting a slower growth rate.
Weekly Seasonality: The second graph shows fluctuations in usage throughout the week. For both services, midweek days tend to be busier than weekends, indicating higher demand for taxis during workdays.

Time-Series Forecasting Analysis

Yearly Seasonality: The third graph depicts annual fluctuations, which could be due to seasonal changes, holidays, or other yearly events. For Yellow Taxis, there are significant peaks and troughs corresponding to specific times of the year, possibly reflecting tourist seasons or holidays. Green Taxis show less pronounced yearly seasonality, which might indicate a steadier demand throughout the year with some minor fluctuations. These graphs together reveal that while both taxi services are experiencing growth, Yellow Taxis have higher variability in demand throughout the year, likely due to factors like tourism and seasonal events in the areas they primarily serve. Green Taxis, while also growing, may serve a customer base with more consistent taxi usage habits.

Conclusion

Key Findings:

- Peak Times: Yellow Taxis are most used during evening peak hours, while Green Taxis have a more consistent demand throughout the day.
- Spatial Demand: High-demand hotspots are centered in Manhattan, indicating a concentration of taxi usage in busy urban areas.
- Growth Trends: Forecasting shows an overall growth trend in taxi usage, with Yellow Taxis having a higher rate of increase.

Conclusion

Implications:

- Resource Allocation: Taxi companies can optimize fleet distribution, especially during peak times in high-demand areas.
- Urban Planning: Insights into taxi usage can guide traffic management and infrastructure development, particularly in Manhattan.
- Service Adaptation: Understanding temporal and spatial patterns allows for tailored services, such as more taxis during peak hours and in bustling city zones.

Introduction To Project 2 - NASA Data

- **Brief overview of the project's purpose**
  This project aims to analyze Near Earth Objects (NEO) by checking if they are potentially hazardous or not based on their distance and velocity.
- **Importance of analyzing and forecasting NEOs**
  Analyzing NEOs is crucial for identifying potentially hazardous objects and calculating their risk for impacting our planet. This information is essential for developing strategies to mitigate the threat and protect Earth from a catastrophic impact.

Data Overview

- **NASA's API Service**
  The project collects data from NASA's API service about NEOs. It includes data such as closest approach date, velocity, and miss distance.

Data Preparation

- **The data loading process:**
  The data loading process involves fetching NEOs records for a whole year using Python, an API request and pandas, and then concatenating them into single dataframes for each NEO. We did this using (basic_df.head()).

- **Basic descriptive statistics**

Table: NEO Data Descriptive Statistics

Statistic	Value
Mean Closest Approach Distance	3.309654e+07
Median Closest Approach Distance	3.061032e+07
Standard Deviation Closest Approach Distance	4.496702e+07
Mean Is Hazardous	0.083333
Median Is Hazardous	0.055556
Mean Size	146.70796003049114
Median Size	54.14303752645
Standard Deviation Size	288.1746144777652
Correlation - Approach Distance	0.09154507447676018
Correlation - Size	0.26006749451733024

Data Analysis - Visualizations

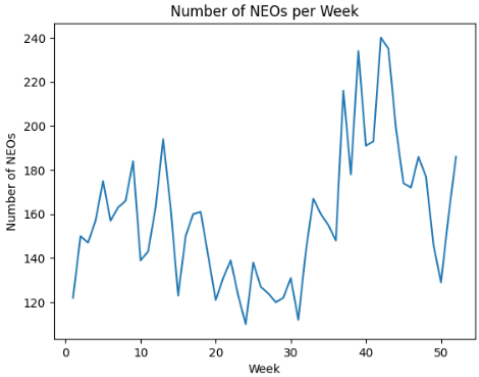

Figure: Number of NEOs pr. week.

Data Analysis - Part 2 (Continued)

The line plot diagram shows how many NEOs are there for each week. Here we can see that in week 40, there were over 225 NEOs identified, in contrast to week 25 where there were less than 25. As it showcases differences over time the display should reflect that and a line graph was therefore chosen with a light blue colour making it visually appealing and the weekly intervals seperated by 10 reduces clutter.

Data Analysis - Visualizations (Continued)

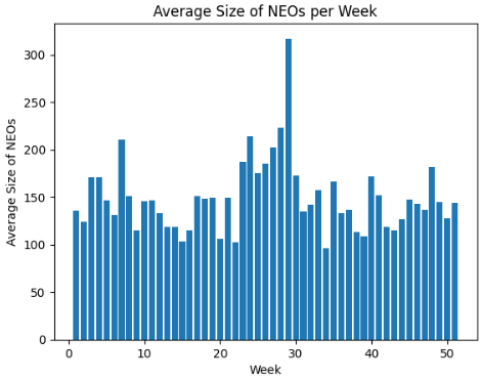

Figure: Average size of NEOs pr. week.

Data Analysis - Visualizations (Continued)

In this histogram, we can see and analyze the average size of NEOs for each week. We can read that in week 29, the largest NEOs were found and a significant overall average increase in Neo-size occured between week 22 and 30.In this case "size" would be easier displayed with a histogram as the bars provide a visual representation and draws attention to the largest ones (the middle part)

Data Analysis -Visualizations (Continued)

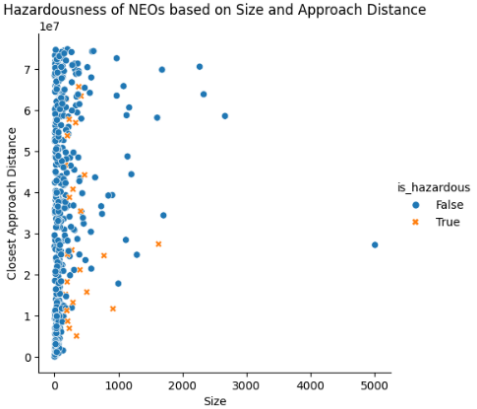

Figure: Hazardousness of NEOs based on size and approach distance.

Data Analysis -

By reading this scatter plot, we can see the hazardous NEOs and nonhazardous NEOs plotted in. The scatter plot visualizes the relationship between the size of Near-Earth Objects (NEOs), their closest approach distance, and their hazardousness. Non-hazardous NEOs are represented by blue circles and hazardous ones by orange circles. The plot showcases no correlation between a NEO being Hazardous and closest approach distance. The contrasting colours of blue and orange makes it easy to spot the hazardous vs non hazardous NEOs and the scatter plot itself showcases the sheer quantity of NEOs well without making it overwhelming visually.

Data Analysis -Visualizations

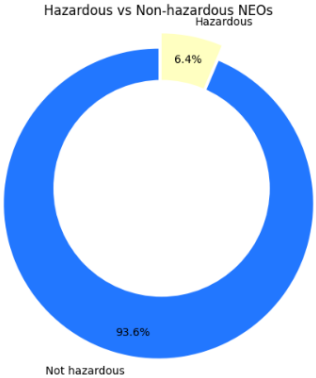

Figure: Percentage of hazardous vs Non-hazardous NEOSs. Showcases that the vast majority of NEOs are non hazardous. Visually appealing minimalistic display without clutter makes it easier to understand. Blue colour for non hazardous also emphasizes this

Data Analysis -Visualizations

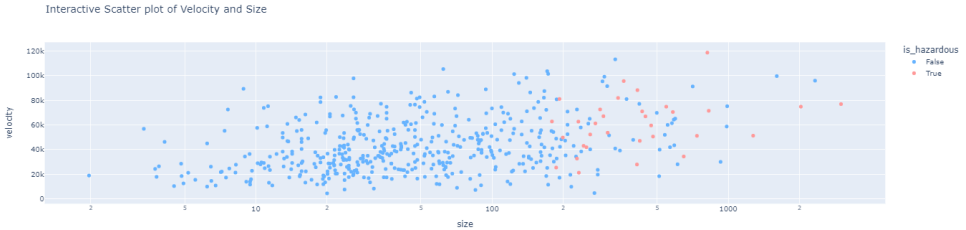

Figure: scatter plot showcasing velocity and size. We can see that there is a correlation between size and if the NEOs are hazardous. The bigger the size the higher the likelyhood for it being hazardous. The scatter plot display showcases the quantity well and the contrasting colours also makes it easy to comprehend.

Data Analysis -Visualizations

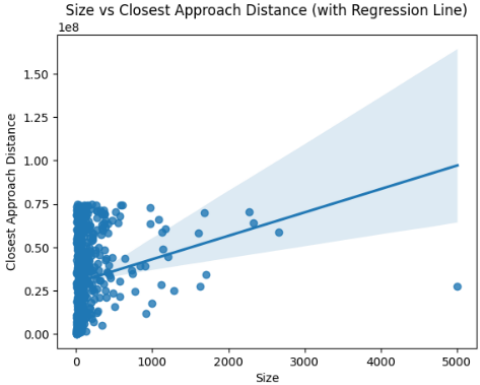

Figure: Here we see the size vs closest approach distance with a regression line making it clear if a relation between the two is to be found. In this case the size/CAD have no relation as the largest/smallest of the NEOs could have the same closest approach distance. The regression line draws attention to the non-relation

Data Analysis -Visualizations

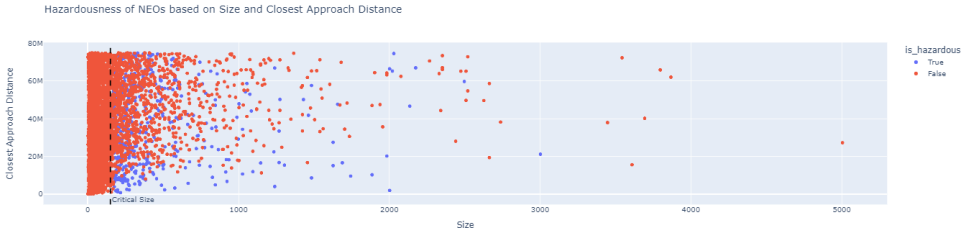

Figure: This scatter plot showcases size/Closest approach distance and if the NEOs are hazardous. .

Data Analysis - Overall Scatter plot

By reading this scatter plot, we can see the hazardous NEOs and nonhazardous NEOs plotted in. The scatter plot visualizes the relationship between the size of Near-Earth Objects (NEOs), their closest approach distance, and their hazardousness. Non-hazardous NEOs are represented by blue circles and hazardous ones by orange circles. The plot showcases no correlation between a NEO being Hazardous and closest approach distance. The contrasting colours of blue and orange makes it easy to spot the hazardous vs non hazardous NEOs and the scatter plot itself showcases the sheer quantity of NEOs well without making it overwhelming visually.

Predictions

Resource Allocation: Allocate more resources to track larger NEOs as they might pose a higher risk if they have close approaches to Earth.
Trend Analysis: Conduct further analysis on trends in the number and size of NEOs over time to predict potential increases in detection rates.
Risk Mitigation Strategies: Develop and implement targeted risk mitigation strategies based on the size and hazardous nature of detected NEOs.

Scientific Paper Recommendation:

Paper Title: "A Machine Learning Approach for NEO Classification Based on Orbital Characteristics"
This paper discusses a machine learning approach for classifying Near-Earth Objects (NEOs) based on their orbital characteristics.