Data Analysis and Prediction on Green Taxi data in New York City

GITHUB LINK:
https://github.com/JerryWu-code/IND5003_Group_Project
GROUP MEMBERS:
WU QILONG, XIANG XIAONENG, JIN YIXUAN, DONG XINYUE, QI SHUOLI

1. Task Description:

In the bustling metropolis of New York City, millions of citizens rely on taxis for daily commutes, which exposes taxi companies to the challenges of taxi services management. Consequently, our task here is to evaluate how taxi companies can improve operations using data analytics. Our project focuses on exploratory data analysis and prediction of taxi pickups within New York City.

2. Data Collection: TLC Trip Record Data

2.1 Green Taxi Dataset in New York City

We gathered information spanning from January 2019 to July 2023 concerning taxi services and their availability within the boroughs of New York City. These taxis respond to street hails but only operate in specified "green areas," which are located above West 110th Street and East 96th Street in Manhattan and in the boroughs.

2.2 Daily weather data in New York City

We Collected comprehensive daily weather data for New York City dating from Jan 2019 to Jul 2023 from NCEI (https://www.ncei.noaa.gov/access/search/index). Integrating this weather data with the taxi dataset, we aim to delve into the interplay between weather conditions and taxi demand, potentially revealing significant correlations or causative factors. The dataset provides in-depth weather metrics for each day, including Dew Point, Atmospheric Sea Level Pressure, Wet Bulb Temperature and Average Wind Speed.

2.3 Location zone data corresponding to ID

We collected data Taxi Zone Maps and Lookup Tables from NYC Government (https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page) and also get Whole US ZIP Code Tabulation Areas(shapefile: "~.shp") for geo-visualization from https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2022&layergroup=ZIP+Code+Tabulation+Areas. Accordingly, we could get the zone corresponding to each region ID.

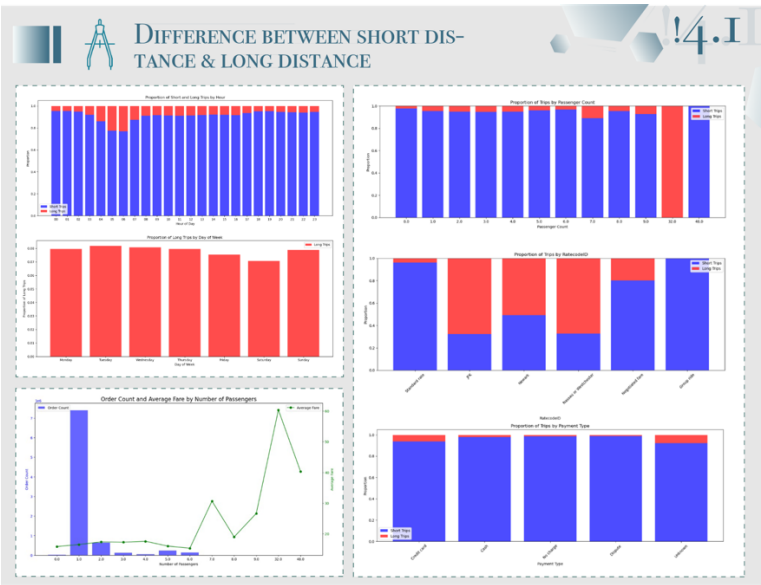
3. Data Cleaning

We use the Google Maps API to get more geographic information, including latitude and longitude (LatLong) and zip code (Zipcode), to make the data more complete. Next, we get weather data, including rainfall, maximum temperature, minimum temperature, and average temperature. We introduced two new features in the data table called "PU_Day_Count" and "PU_Day_Avg_Fare". These features provide information about the number of daily pick-up zip code observations and the average total dollar amount. We perform a missing value check on the data to remove missing values. Also, we processed latitude and longitude information to support subsequent geospatial analysis.

We split the data into two groups to meet different analysis needs. df1 is used for multivariate analysis that does not include NaN row information, while df2 is used for analysis that does not include NaN column information, such as studies related to price, distance, and region. Throughout the data pre-processing process, we adhere to high standards of data cleaning and quality control to ensure data reliability and accuracy. This allows us to understand and explore the information and trends in the data more deeply.

4. Exploratory Data Visualization

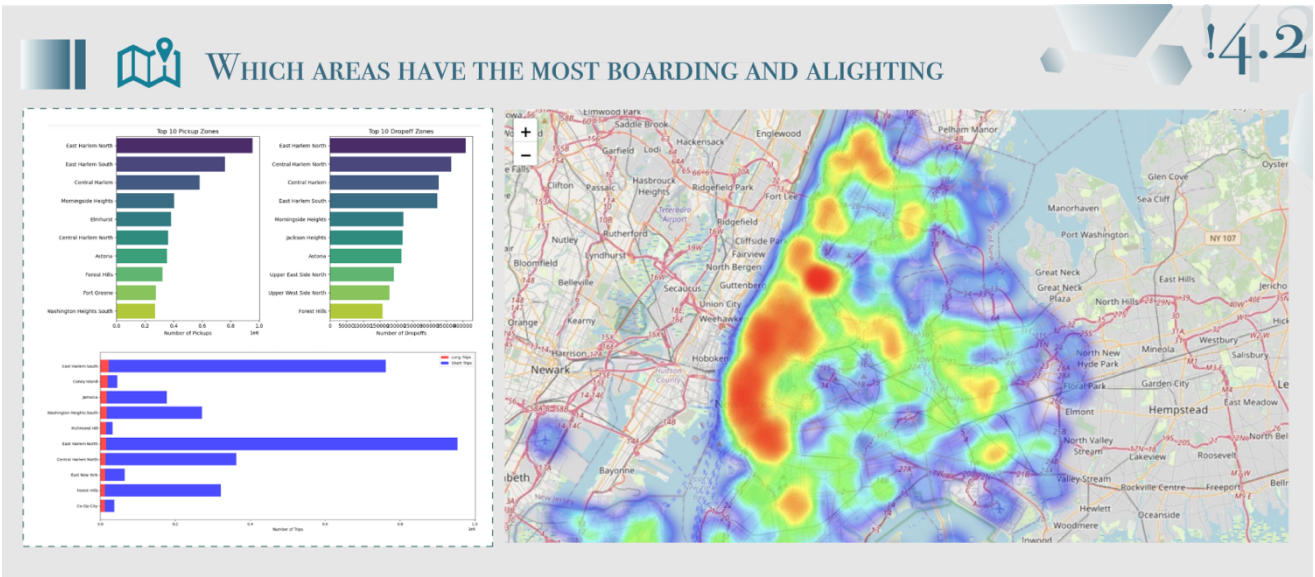
4.1 Analysis of difference between short distance and long distance:



A careful analysis of short- and long-haul taxi trips reveals patterns at different times of the day and on different days of the week that can be used to optimize services to meet user needs.

- **Short-haul and long-haul travel strategies:** Short-haul trips were more frequent across all hours, but long-haul trips were also present, while long-haul trips remained stable throughout the week without significant change. This allows us to adjust services and resources based on hourly patterns while ensuring that ongoing demand for long-distance services is met, guiding weekly strategy and resource allocation.
- **Geographical analysis** reveals the needs and preferences of different regions for short and long-distance trips. Strategies that optimize key areas, routes, and availability can improve service quality and customer experience.
- **Pick-up area priority:** Different regions have different needs for short and long distances, and key pick-up areas need to be considered to meet the demand efficiently.
- **Optimization of drop-off areas:** The main drop-off areas are analyzed, and service reliability and customer satisfaction can be improved through strategic planning.

4.2 Which areas have the most boarding and alighting:



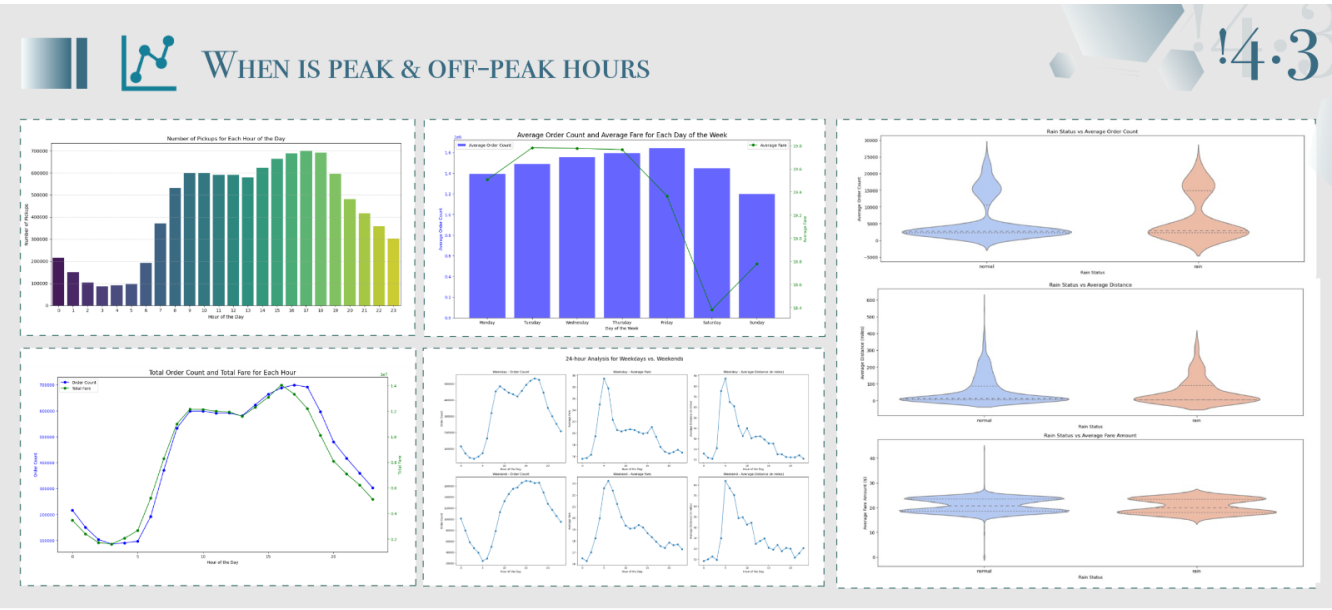
- **East Harlem North** is a critical transport hub for both pickups and drop-offs, implying a consistent demand for taxi services.
- Areas like **Central Harlem** (both North and South) are significant for taxi services, indicating the

need for a focused deployment of taxis in these neighborhoods.

- The presence of **Upper East** and **West Sides** points to a high residential and possibly tourism-driven demand.
- **LaGuardia Airport**'s high drop-offs volume highlights the taxi industry's reliance on airport traffic and indicates a potential for fixed-fare, scheduled rides for airport commuters.

Taxi operators are encouraged to leverage these insights for strategic decision-making: focusing on profitable areas with higher average fares and ensuring a robust presence in high-demand zones like Manhattan to maintain market presence. Dynamic pricing in Manhattan could optimize revenue during peak demand, while service expansion in the less-served boroughs like The Bronx and Staten Island could tap into untapped markets, potentially increasing the overall number of trips and market penetration.

4.3 Peak and Off-Peak Times

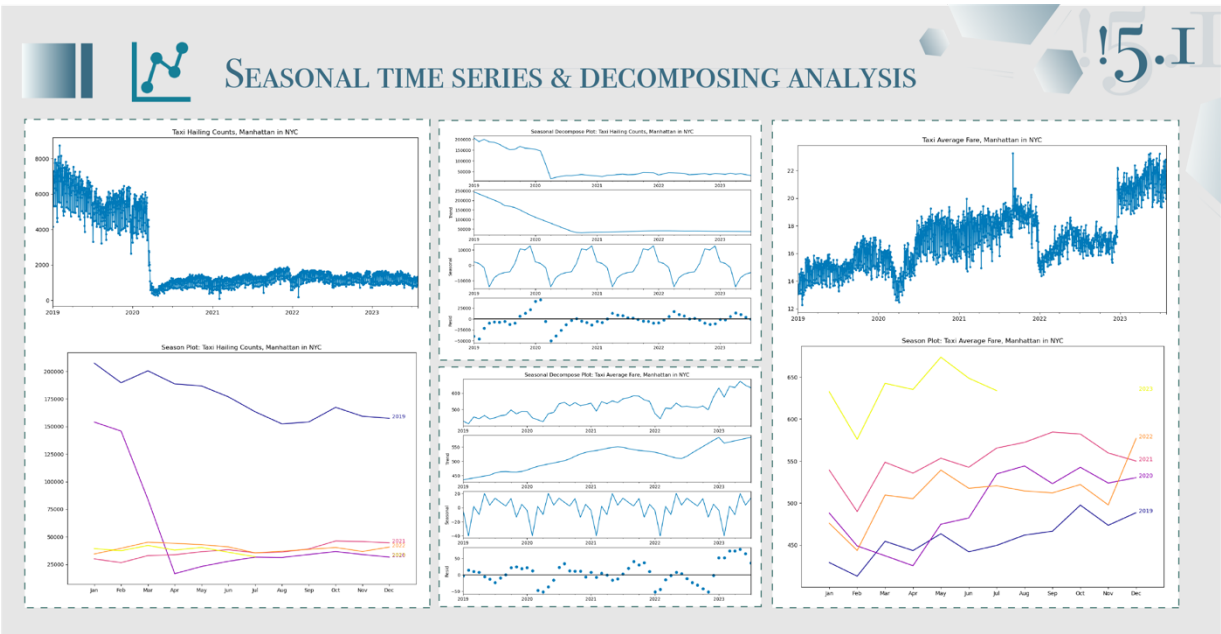


- **Hourly:** Taxi pickups peak during the evening rush hour, particularly around 6 PM, while the early morning hours from midnight to 5 AM represent the off-peak period with the fewest pickups.
- **Daily:** The number of taxi orders and the total fare collected both reach their highest levels in the evening, reflecting increased economic activity and possibly longer or more frequent rides during these hours.
- **Weekly:** Taxi usage is relatively stable across the weekdays but begins to taper off on Fridays, with the weekends, especially Sundays, seeing the least activity; concurrently, average fares incrementally increase as the week progresses, reaching the highest point on Fridays.
- **Weekday vs. Weekend:** On weekdays, taxi demand spikes during traditional commuting hours in the early mornings and evenings, whereas the weekend demand pattern shifts to later start times in the morning and sustained demand into the late evening.
- **Weather-Related:** Inclement weather, such as rain, has a clear impact on taxi usage, with a noticeable increase in both the frequency of rides and the distances travelled, resulting in higher average fares on rainy days compared to non-rainy days.

Peak hours for taxi services are in the evenings, particularly around 6 PM on weekdays, with a broader peak period on weekends. Off-peak hours occur during the early morning on all days and mid-afternoon on weekdays. Rain increases both the demand for taxis and the average fare, suggesting the potential for weather-adaptive pricing strategies. These insights could help taxi service providers in scheduling drivers, setting dynamic pricing, and understanding the influence of weather on taxi usage.

5. Time Series Analysis:

5.1 Seasonal Time Series Analysis & Decomposing Analysis



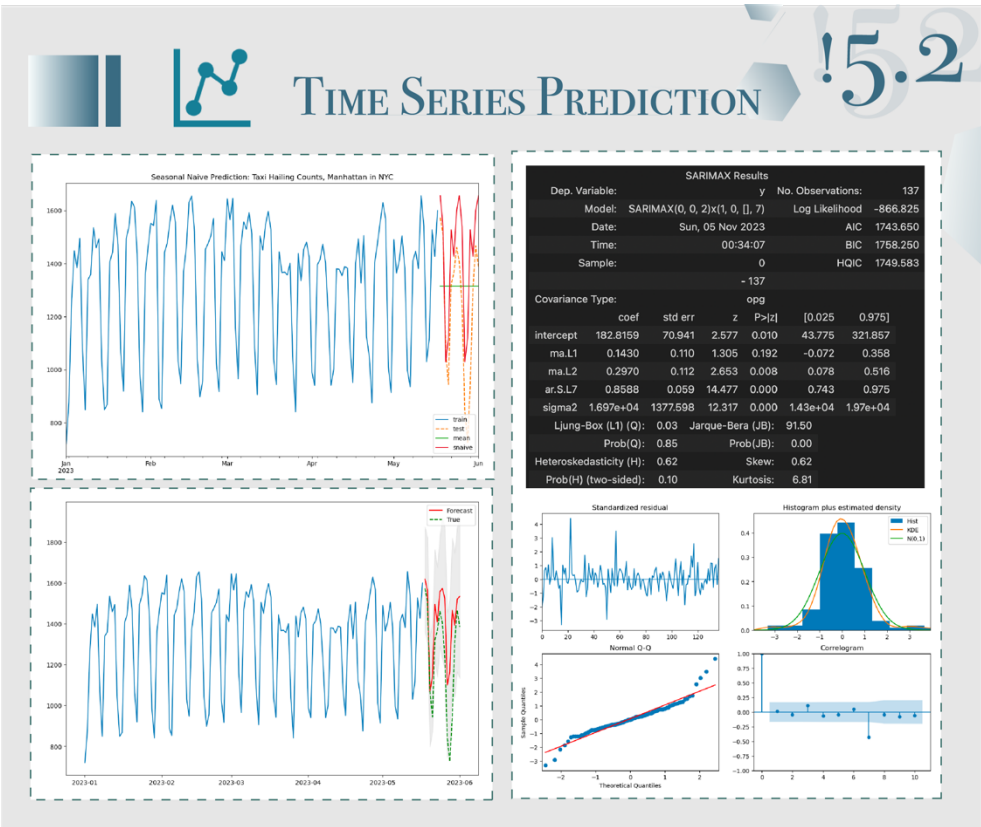
- **Seasonal Time Series Analysis:** Applications of seasonal time series analysis in taxi operations include identifying cyclical fluctuations in the number of orders over different time periods, such as weekends of the week or specific seasons, thereby helping taxi companies more efficiently allocate resources to meet high-demand seasonal demands, such as increasing the number of vehicles and drivers to meet peak hour orders.
- **Decomposing Analysis:** Decomposition analysis is a method that helps to break down time series data into three main components: trend, seasonality, and residual. In taxi operations, the trends section helps to understand the long-term trend of the number of orders, that is, whether the number of orders is gradually increasing or decreasing. The seasonal section reveals cyclical fluctuations in order volumes over different time periods, helping to understand peak and trough periods. The residuals represent fluctuations in the data that are not explained by trends and seasonal patterns and can be used to identify unexpected events or anomalies to better understand fluctuations and changes in the order data. This helps to make more precise strategies and decisions to adapt to market changes and improve the efficiency of the business.

Based on the results of Seasonal time series analysis and Decomposing Analysis, it can be concluded that the COVID-19 pandemic has had a significant impact on taxi companies, with a significant decrease in the number of orders but a significant increase in prices. This will help taxi companies maintain profitability in the face of declining order volumes and increase average revenue per order, thus meeting challenges during the pandemic. Companies can use this insight to adapt their operational strategies to market changes and take measures to improve efficiency, such as offering high value-added services or flexible pricing strategies to ensure sound operations.

5.2 Time Series Prediction

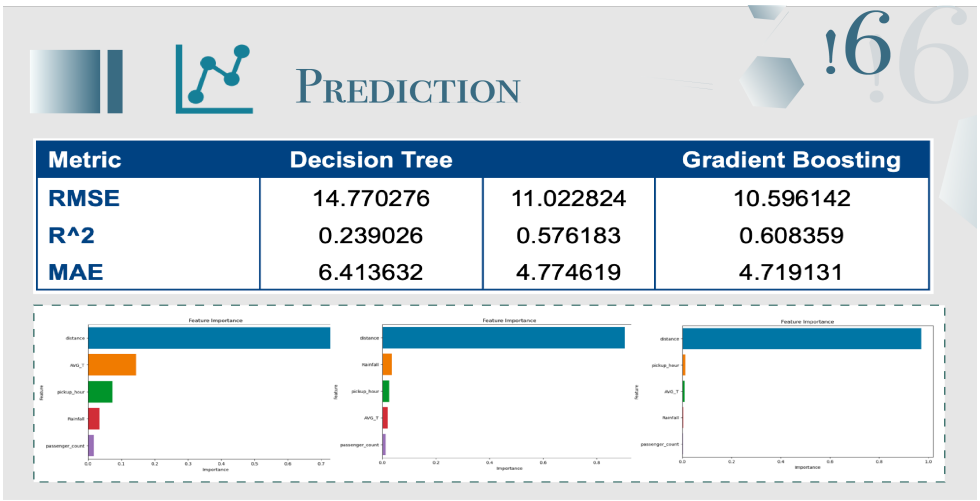
- **Seasonal Naive Prediction:** This technique was employed to establish a baseline forecast by assuming that the pattern observed in the past season (a week in this case) would repeat itself in the future. It's a simple method that relies on historical data patterns without considering trends or other factors. The graph suggested that this method was effective for short-term forecasting where the weekly pattern is consistent and significant. The visual alignment of predicted and actual counts underscored the model's utility for immediate operational decision-making.

- **ARIMA:** The Autoregressive Integrated Moving Average (ARIMA) model, with seasonal adjustments, was utilized to capture both the non-seasonal and seasonal elements in the data. This model is more sophisticated than the Seasonal Naive approach, as it involves identifying and estimating parameters that govern the autoregressive and moving average components, as well as differencing orders to account for trends and seasonality. Diagnostic graphs indicated that the ARIMA model residuals did not perfectly follow a normal distribution, as evidenced by the Q-Q plot and the slightly skewed Histogram. The correlogram showed that the residuals were largely uncorrelated, except for a few outside the confidence interval, hinting at a generally good model fit but with room for improvement.



By combining the simplicity of the Seasonal Naive Prediction model with the comprehensive nature of the ARIMA model, businesses can form a robust predictive framework. This integrated approach is instrumental for both tactical week-by-week planning and strategic long-term decision-making in the dynamic environment of taxi hailing services in Manhattan. It is, however, vital to continuously update and evaluate these models with new data and to remain cognizant of external factors that could influence demand patterns, such as regulatory changes, competitive services, and evolving consumer preferences.

6. Regression Analysis



According to performance matric, it can be seen that the Gradient Boosting model shows better performance in this problem because it has lower RMSE, higher R^2, and smaller MAE. This means they fit the data more accurately and provide better predictive power. The performance of the Decision Tree model is relatively poor, with high RMSE, low R^2 and large MAE, which means that its prediction ability is weak, and it may overfit the data.
Three models, Gradient Boosting, XGBoost and Decision Tree, were compared in regression analysis to study the relationship between them and expenses. By analyzing the Feature Importance of these models, taxi companies can better understand the relationship between fees and various factors, so as to optimize their business strategies, improve efficiency, provide better services and meet customer needs.

- **Cost and distance correlation:** Since distance has the largest weight in Feature Importance, this can mean that there is a strong positive correlation between expense and ride distance. This means that taxi companies can reasonably expect fares to increase accordingly as the distance travelled increases.
- **Impact of Other features:** Although distance is the most important, other features also have an impact on cost. For example, boarding times can have an impact on fares during peak and low peak hours, average temperatures can affect whether people are willing to take a taxi, rainfall can cause fares to rise during periods of high demand, and ridership can also affect fares as more riders may require a larger vehicle.
- **Pricing strategies:** Based on these analysis results, taxi companies can optimize their pricing strategies. If distance is a major cost factor, they might consider developing a more refined pricing strategy, setting prices based on distance traveled. At the same time, they can also adjust prices under different time, weather, and passenger flow conditions to better meet market demand.

7. Conclusion

Our comprehensive analysis of New York City's taxi data has illuminated critical insights into the industry's operating patterns and potential areas for optimization. By harnessing advanced predictive models, such as Gradient Boosting which outshines others in performance metrics, we can pinpoint the significant impact of travel distance on fares and delineate the influence of various other factors including time, weather conditions, and passenger flow.

The exploratory data analysis has revealed not only the temporal dynamics of taxi demand—highlighting evening peaks and early morning troughs—but also geographical hotspots for taxi activity, underscoring the importance of strategic fleet deployment. In addition, the influence of seasonal and weekly cycles has been captured, offering taxi companies a granular view of demand fluctuations.

As the taxi industry faces ongoing challenges, including those brought on by the pandemic, the use of geo-visualization and time series forecasting empowers service providers to make data-driven decisions that cater to changing customer behaviors. Moreover, regression analysis has facilitated a deeper understanding of fare structuring, guiding companies towards more dynamic and nuanced pricing strategies.

In conclusion, our analysis serves as a testament to the power of data in transforming operations and enhancing the efficiency of taxi services. With these insights, taxi companies are well-positioned to navigate the complex urban transportation landscape, adjusting to the ebb and flow of city life, and meeting the evolving needs of New Yorkers. The deployment of analytical models not only anticipates the industry's trajectory but also equips providers with the agility to sustain profitability and maintain competitive advantage in an ever-changing environment.