Project Report – Data Science

1. Introduction

This report provides an in-depth analysis of ride-hailing trends in New York City (NYC) from 2019 to 2024, focusing on the major companies Lyft, Uber, and Via. The data spans monthly ride counts for each company over the specified period. This analysis explores the impact of regulatory changes, company exits, and recovery patterns post-COVID-19. Additionally, the unique dynamics of NYC's five boroughs are considered to contextualize the ride-hailing market trends.

2. Context and Background

Ride-Hailing Market Overview

- **Uber**: Maintained a dominant position in the NYC market throughout the observed period.
- **Lyft**: Consistently served as a strong competitor to Uber, with significant market share.
- **Via Transportation Inc.**: Ceased its ride-hailing services in NYC on December 20, 2021, due to operational challenges.
- **Juno**: Filed for bankruptcy and ceased operations in late 2019 due to stringent NYC wage regulations for drivers.

The Five Boroughs of NYC

- **Manhattan (New York County)**: Cultural and economic hub, featuring dense ride-hailing activity.
- **Brooklyn (Kings County)**: Diverse neighborhoods and a vibrant art scene contribute to ride-hailing demand.
- **Queens (Queens County)**: Proximity to major airports makes it a significant contributor to ride-hailing volumes.
- **The Bronx (Bronx County)**: Features attractions like Yankee Stadium and the Bronx Zoo, driving periodic ride demand.
- **Staten Island (Richmond County)**: Suburban characteristics result in comparatively lower ride-hailing activity.

3. Methodology

Data Cleaning

To ensure the integrity and usability of the dataset, a systematic data cleaning process was implemented using Python. The following steps were applied to handle missing values and optimize the dataset for analysis:

1. **Column-Specific Null Value Replacement:**
    - airport_fee Column: Null values were replaced with a default value of 0 to standardize records without an airport fee.

- wav_match_flag Column: Null values were replaced with 'N', indicating the absence of a match.
- congestion_surcharge Column: Null values were replaced with 0 to ensure consistency in surcharge data.

2. **Dropping Records with Critical Null Values:**
    - on_scene_datetime Column: Rows with null values were removed, as this timestamp is critical for accurate analysis.
    - dispatching_base_num Column: Records missing this identifier were dropped to maintain the reliability of dispatch data.
    - originating_base_num Column: Rows with null values were removed to ensure consistency in base information.
3. **File-Specific Cleaning:**
    - For each input file, a unique destination path was created to save the cleaned data. File paths were adjusted dynamically for organization and accessibility.

- **Output:**
    - The cleaned data was saved as a new Parquet file using PyArrow. This ensured efficient storage and faster data retrieval for subsequent analysis.

4. **Logging:**
    - Before and after each operation, the total number of records and null values were logged to monitor the impact of cleaning.

Code Details:

The Python function nulls_handling iteratively processed multiple files as follows:

- Input Files: Parquet files containing raw data.
- Output Files: Cleaned Parquet files saved in organized directories.
- Libraries Used:
    - Pandas: For data manipulation and null value handling.
    - PyArrow: For efficient Parquet file handling.

Code Snippet:

```
# Example of airport_fee null replacement

print("Replacing NUlls for airport_fee column")

data['airport_fee'].fillna(0, inplace=True)
```

This detailed and methodical cleaning ensured high-quality datasets, forming the basis for reliable insights presented in this report.

4. Yearly Analysis of Peak Days:

- In **2020**, Thursdays accounted for the highest number of trips (16.8%), indicating a mid-week peak for ride-hailing demand.
- In **2021**, Fridays saw the highest trip volume (18.0%), reflecting a shift in consumer behavior towards the end of the workweek.
- In **2022**, Saturdays recorded the highest percentage (18.9%), suggesting a weekend peak in demand, possibly due to leisure activities.
- In **2023**, Fridays regained the highest trip share (18.0%), demonstrating continued end-of-week ride-hailing reliance.
- In **2024**, Saturdays again showed the highest trip volumes (18.5%), highlighting consistent weekend activity.

1.Lowest Trip Days:

- Mondays consistently exhibited the lowest number of trips across all years, ranging from **10.7%** to **11.5%**, indicating reduced ride-hailing demand at the beginning of the week.

2.Trend Shifts Over the Years:

- There was a notable shift in consumer preferences from mid-week peaks (Thursdays) in 2020 to end-of-week or weekend peaks (Fridays and Saturdays) in subsequent years.
- The transition reflects changes in work-from-home policies, weekend leisure travel, and societal adjustments post-pandemic.

3.Weekend vs. Weekday Demand:

- Weekends (Friday to Sunday) consistently contributed to a higher percentage of trips compared to weekdays, highlighting the leisure-driven nature of ride-hailing services during this period.

4.Year-on-Year Growth in Weekend Usage:

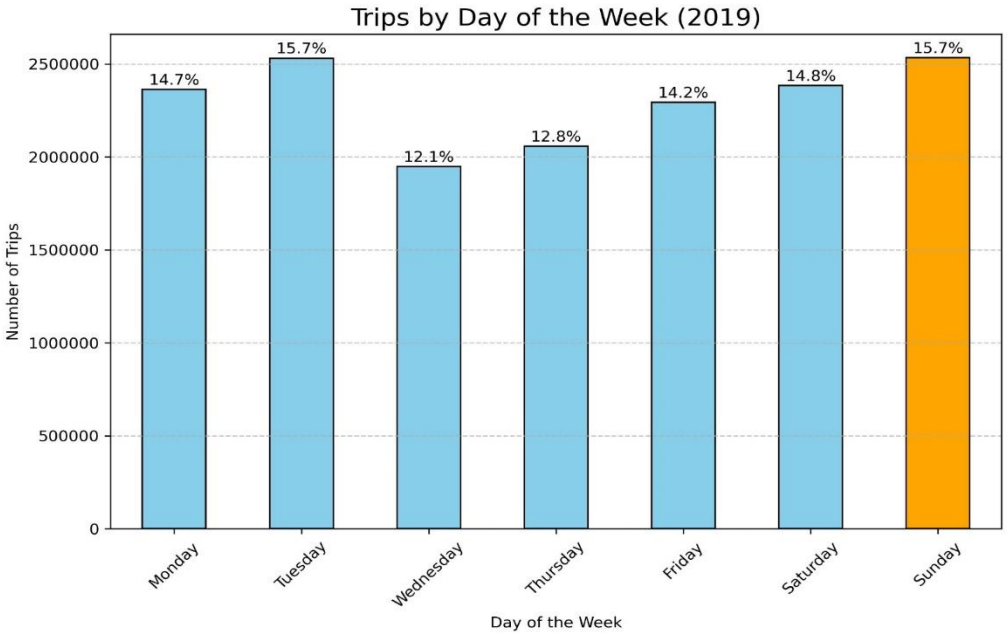
- From 2020 to 2024, the proportion of weekend trips increased, showcasing growing consumer reliance on ride-hailing services for leisure and non-work-related travel.

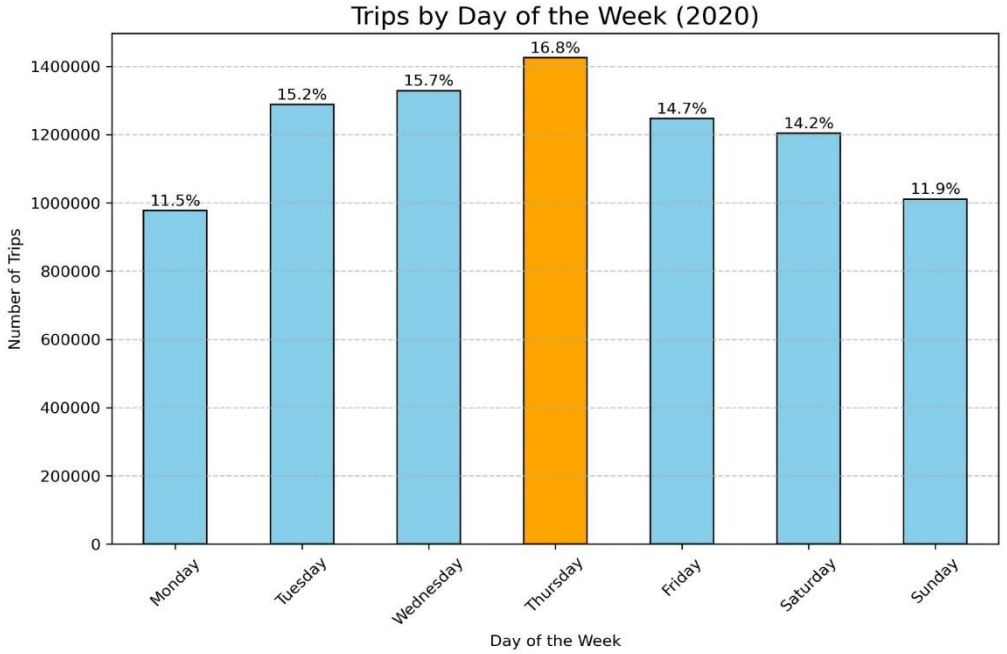
5.Seasonal and Behavioral Insights:

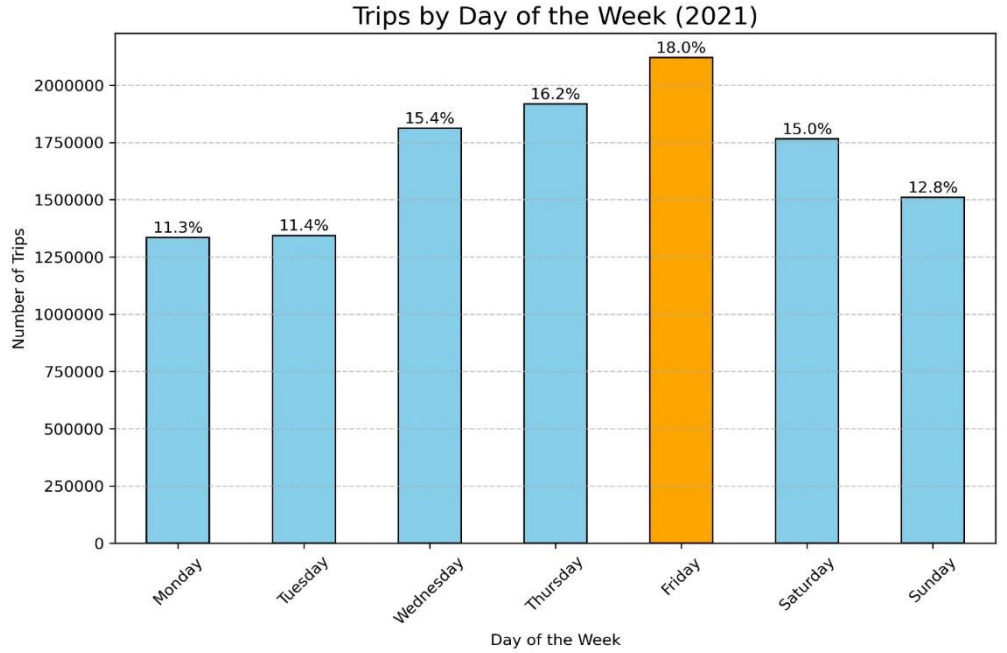
- The increase in Saturday trip volumes from 2022 onward may be attributed to the return of large-scale events, nightlife, and tourism recovery after the pandemic.
- The consistently low trip percentages on Mondays align with reduced travel needs after weekends, potentially reflecting remote work adoption and fewer commutes.

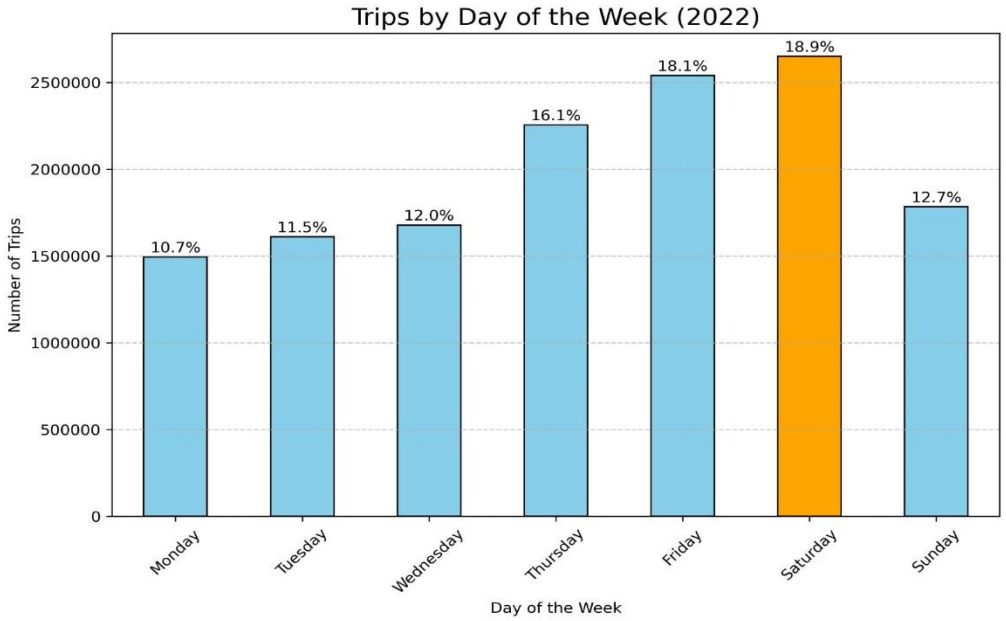
6.Variability in Mid-Week Demand:

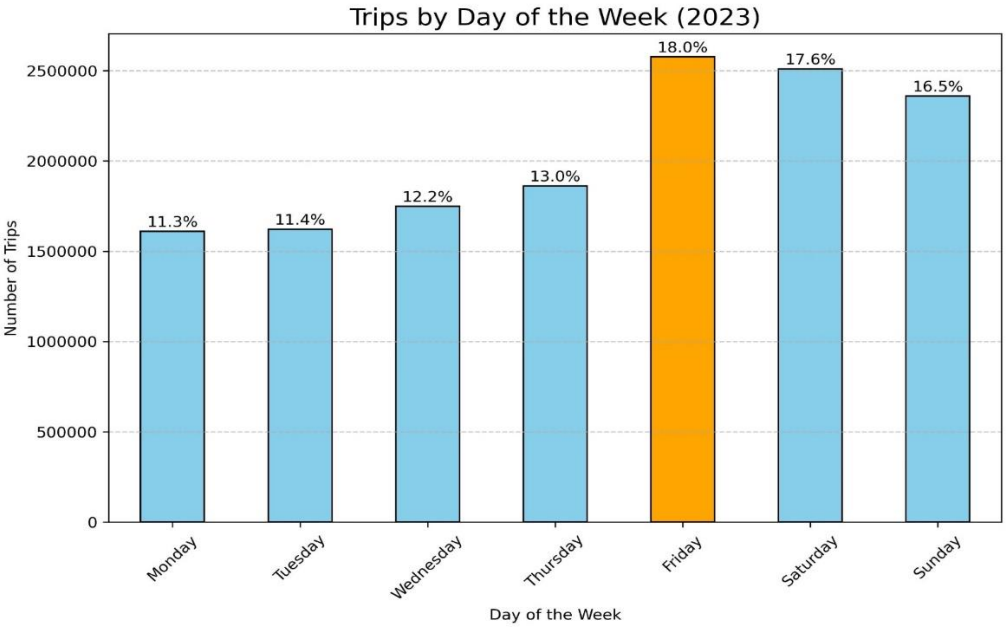
- Mid-week (Tuesday to Thursday) demand remained stable, although fluctuations suggest a balance between work-related travel and other activities.

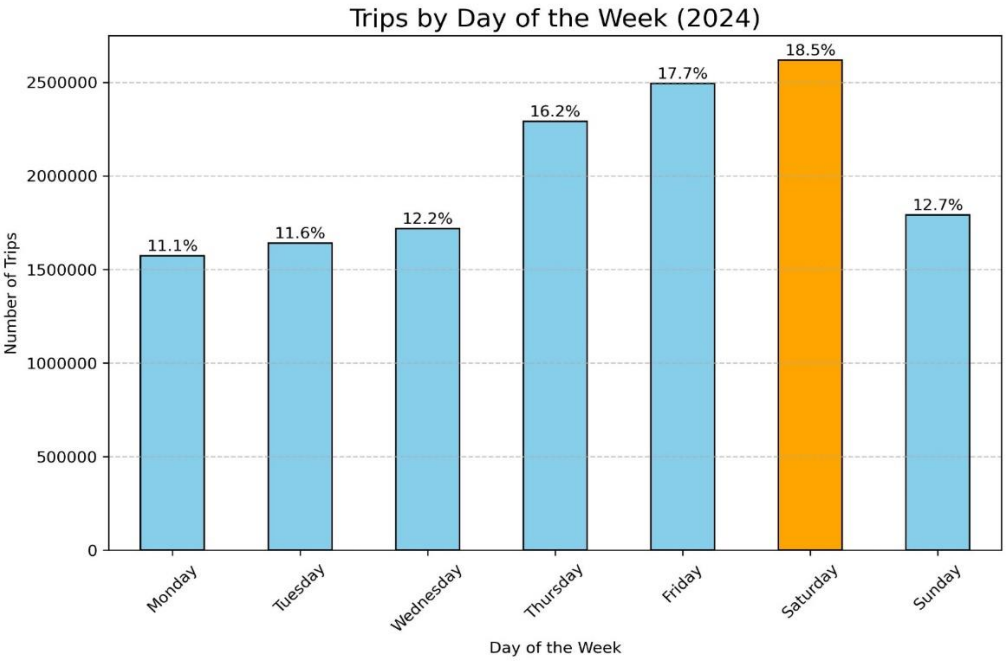

Trips by Day of the Week (2019)
Number of Trips
Day of the Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
14.7%
15.7%
12.1%
12.8%
14.2%
14.8%
15.7%


Trips by Day of the Week (2020)
Number of Trips
Day of the Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
11.5%
15.2%
15.7%
16.8%
14.7%
14.2%
11.9%


Trips by Day of the Week (2021)
Number of Trips
Day of the Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
11.3%
11.4%
15.4%
16.2%
18.0%
15.0%
12.8%


Trips by Day of the Week (2022)
Number of Trips
Day of the Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
10.7%
11.5%
12.0%
16.1%
18.1%
18.9%
12.7%


Trips by Day of the Week (2023)
Number of Trips
Day of the Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
11.3%
11.4%
12.2%
13.0%
18.0%
17.6%
16.5%

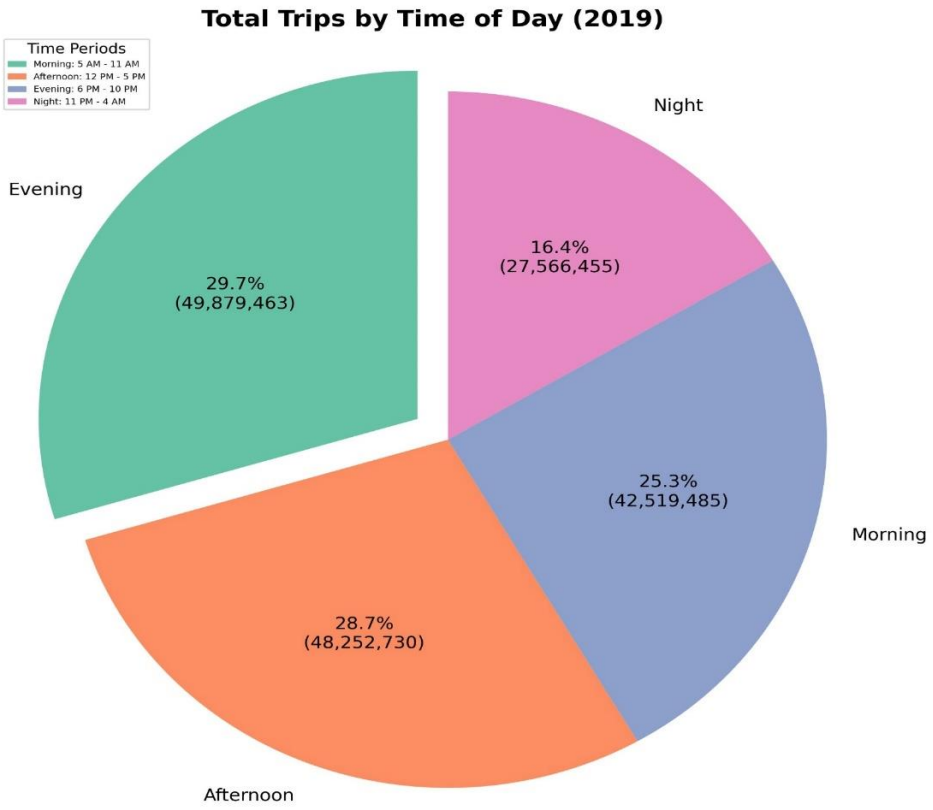

Trips by Day of the Week (2024)
Number of Trips
Day of the Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
11.1%
11.6%
12.2%
16.2%
17.7%
18.5%
12.7%

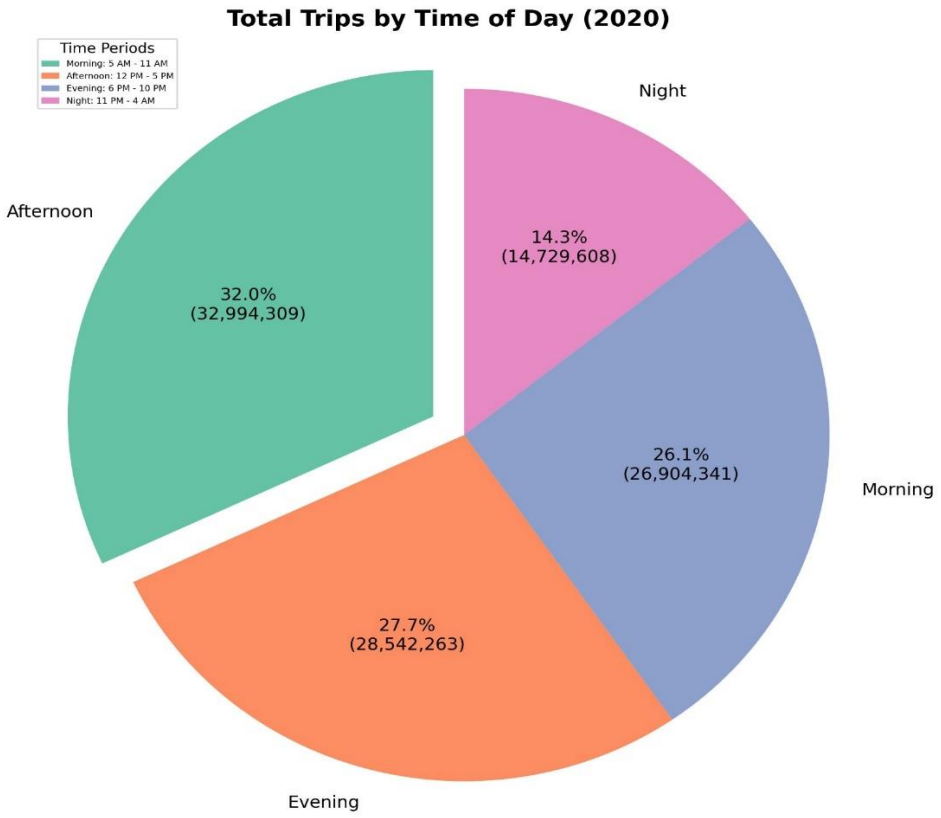
5.Analysis of Ride Distribution by Time of Day (2019–2024)

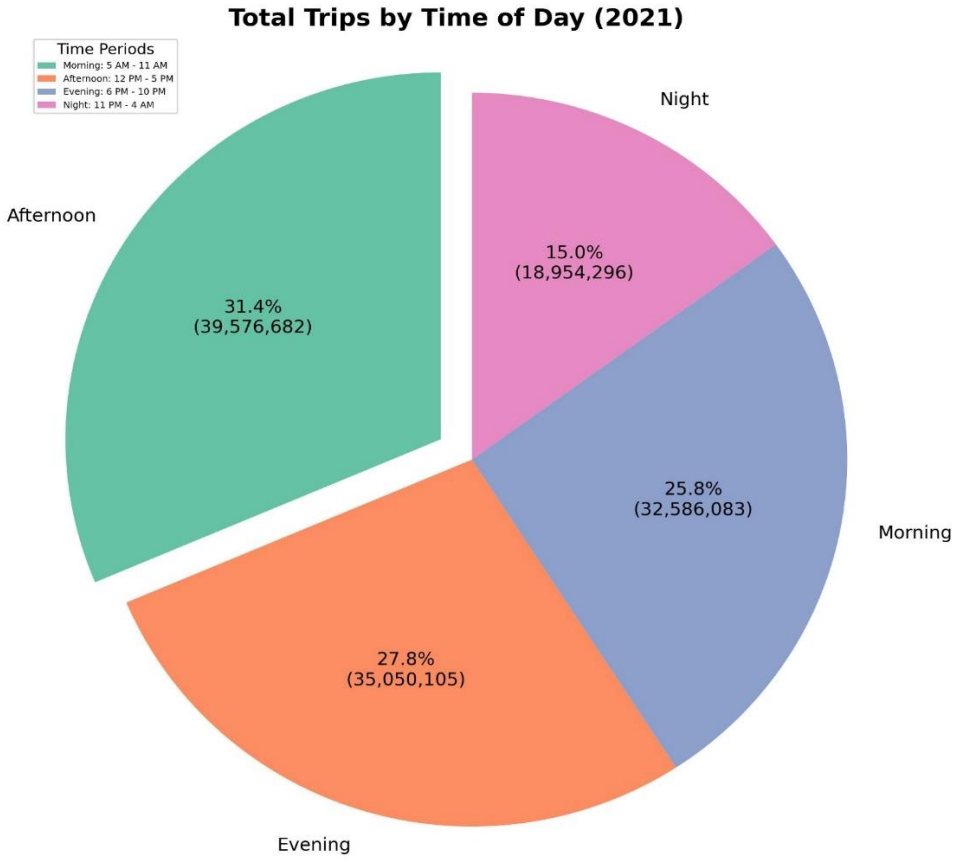
1. **General Observations**:
    - The dataset categorizes trips into four time periods: **Morning (5 AM–11 AM)**, **Afternoon (12 PM–5 PM)**, **Evening (6 PM–10 PM)**, and **Night (11 PM–4 AM)**.
    - Afternoon and Evening time slots consistently accounted for the majority of trips across all years, highlighting their dominance in ride-hailing demand.
2. **Afternoon Peak Usage**:
    - The **Afternoon (12 PM–5 PM)** period consistently had the highest share of trips in most years, with percentages ranging from **29.7%** (2019) to **32.0%** (2020).
    - The afternoon peak suggests heightened activity during this period, likely driven by work commutes, shopping, and leisure activities.
3. **Evening Surge**:
    - The **Evening (6 PM–10 PM)** period maintained the second-highest share, contributing between **27.7%** (2020) and **28.7%** (2019) of total trips.
    - This reflects the strong demand for ride-hailing services during post-work hours, coinciding with social events, dining, and other evening activities.
4. **Morning Trends**:
    - The **Morning (5 AM–11 AM)** time slot consistently accounted for around **25%–26%** of trips in all years.
    - This indicates significant demand for ride-hailing services during morning commutes and airport travel.
5. **Nighttime Demand**:
    - The **Night (11 PM–4 AM)** time slot consistently had the lowest share of trips, ranging between **14.3%** (2020) and **16.4%** (2019).
    - The relatively lower demand at night is expected due to fewer events and reduced commuting needs during these hours.
6. **Year-on-Year Trends**:
    - In **2020**, afternoon trips peaked at **32.0%**, potentially due to a shift in consumer behavior during the pandemic, where work-from-home policies may have increased midday travel.
    - By **2024**, afternoon usage stabilized at **30.1%**, reflecting a return to pre-pandemic behavior with balanced demand across time periods.
7. **Comparing Morning and Evening**:

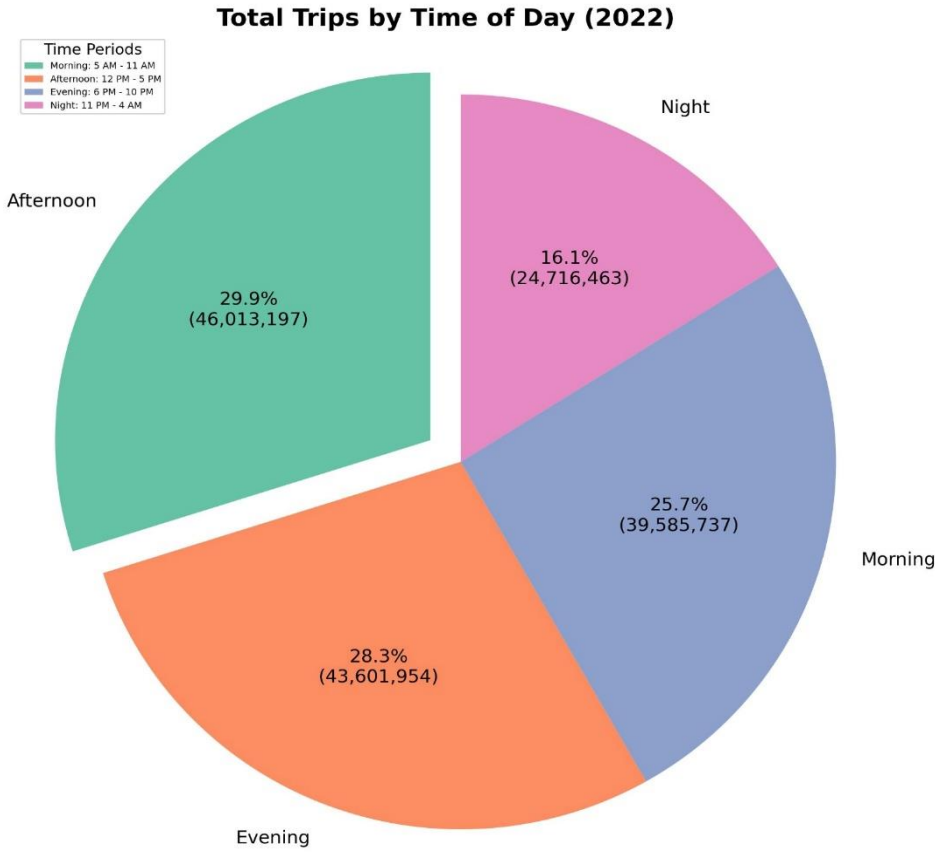
- The **Evening** consistently outperformed the **Morning**, likely driven by the variety of activities during the evening hours compared to the structured nature of morning commutes.

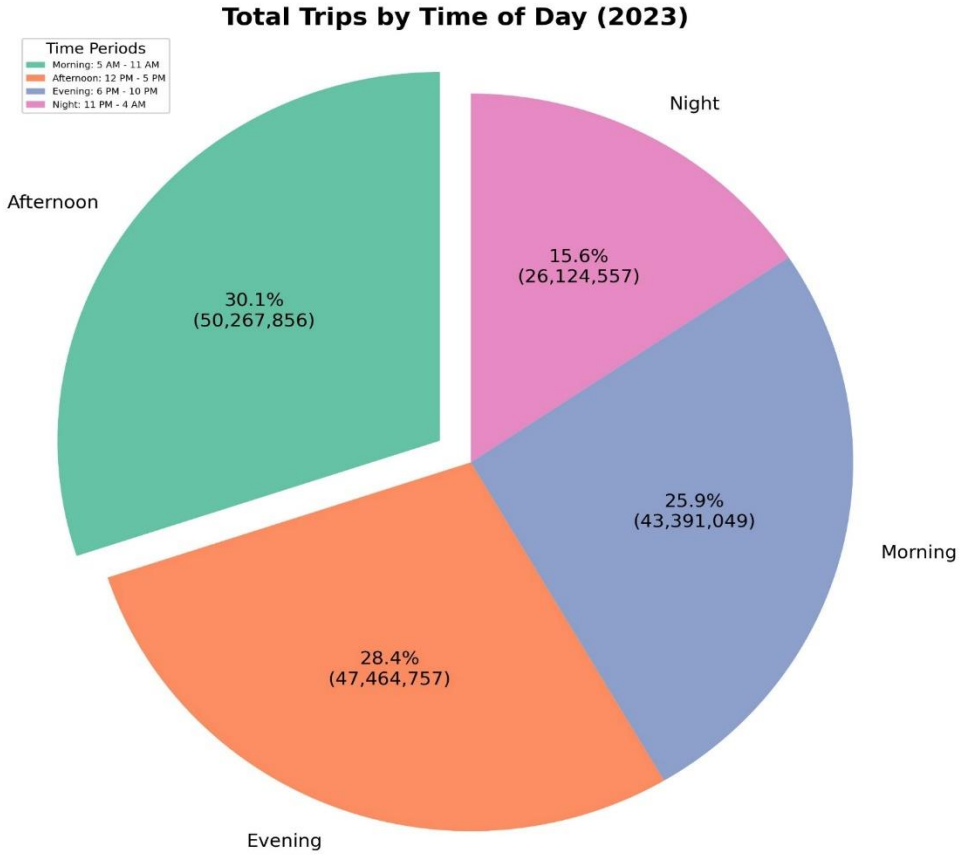
8. **Overall Patterns**:
   - The trip distribution remained relatively stable across the years, with small variations highlighting the adaptability of ride-hailing services to societal shifts like the pandemic and the recovery phase.
9. **Implications for Service Optimization**:
   - Ride-hailing companies could optimize driver availability during Afternoon and Evening periods to maximize efficiency and reduce wait times.
   - Additional incentives for drivers during nighttime hours may help address the relatively lower service coverage during this time.

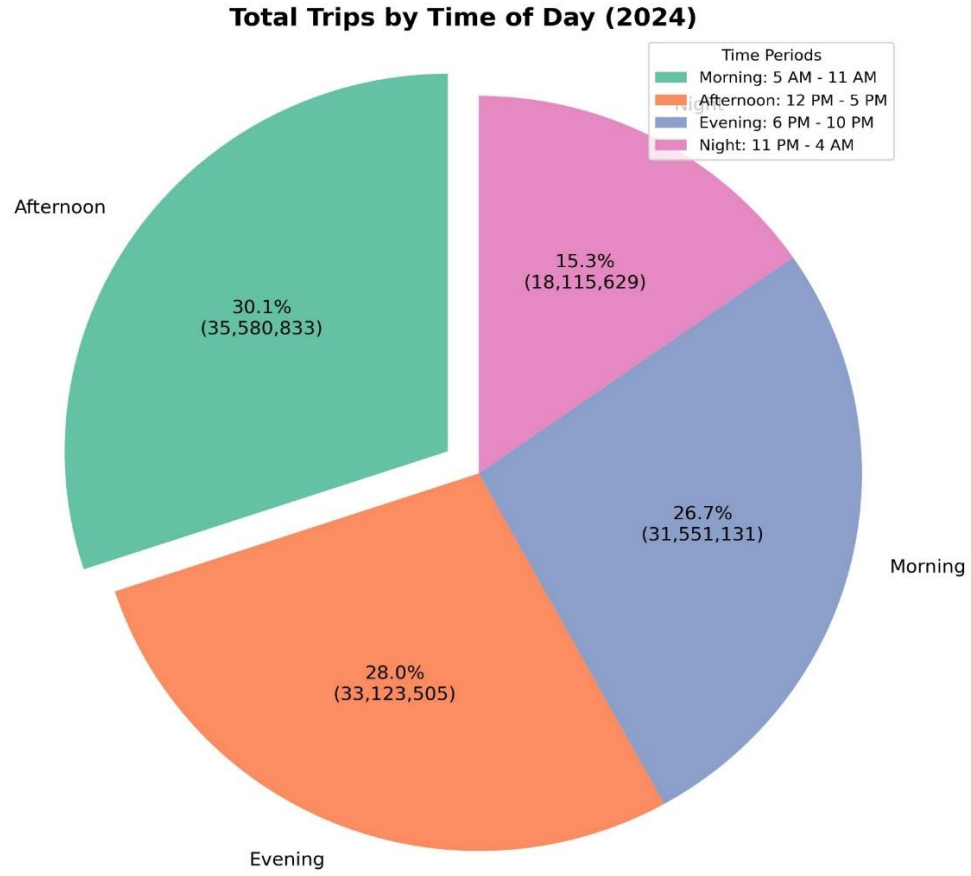



Total Trips by Time of Day (2020)
Time Periods
Morning: 5 AM - 11 AM
Afternoon: 12 PM - 5 PM
Evening: 6 PM - 10 PM
Night: 11 PM - 4 AM
Night
Afternoon
14.3%
(14,729,608)
32.0%
(32,994,309)
26.1%
(26,904,341)
Morning
27.7%
(28,542,263)
Evening


Total Trips by Time of Day (2021)
Time Periods
Morning: 5 AM - 11 AM
Afternoon: 12 PM - 5 PM
Evening: 6 PM - 10 PM
Night: 11 PM - 4 AM
Night
Afternoon
15.0%
(18,954,296)
31.4%
(39,576,682)
25.8%
(32,586,083)
Morning
27.8%
(35,050,105)
Evening


Total Trips by Time of Day (2022)
Time Periods
Morning: 5 AM - 11 AM
Afternoon: 12 PM - 5 PM
Evening: 6 PM - 10 PM
Night: 11 PM - 4 AM
Night
Afternoon
16.1%
(24,716,463)
29.9%
(46,013,197)
25.7%
(39,585,737)
Morning
28.3%
(43,601,954)
Evening


Total Trips by Time of Day (2023)
Time Periods
Morning: 5 AM - 11 AM
Afternoon: 12 PM - 5 PM
Evening: 6 PM - 10 PM
Night: 11 PM - 4 AM
Night
Afternoon
15.6%
(26,124,557)
30.1%
(50,267,856)
25.9%
(43,391,049)
Morning
28.4%
(47,464,757)
Evening



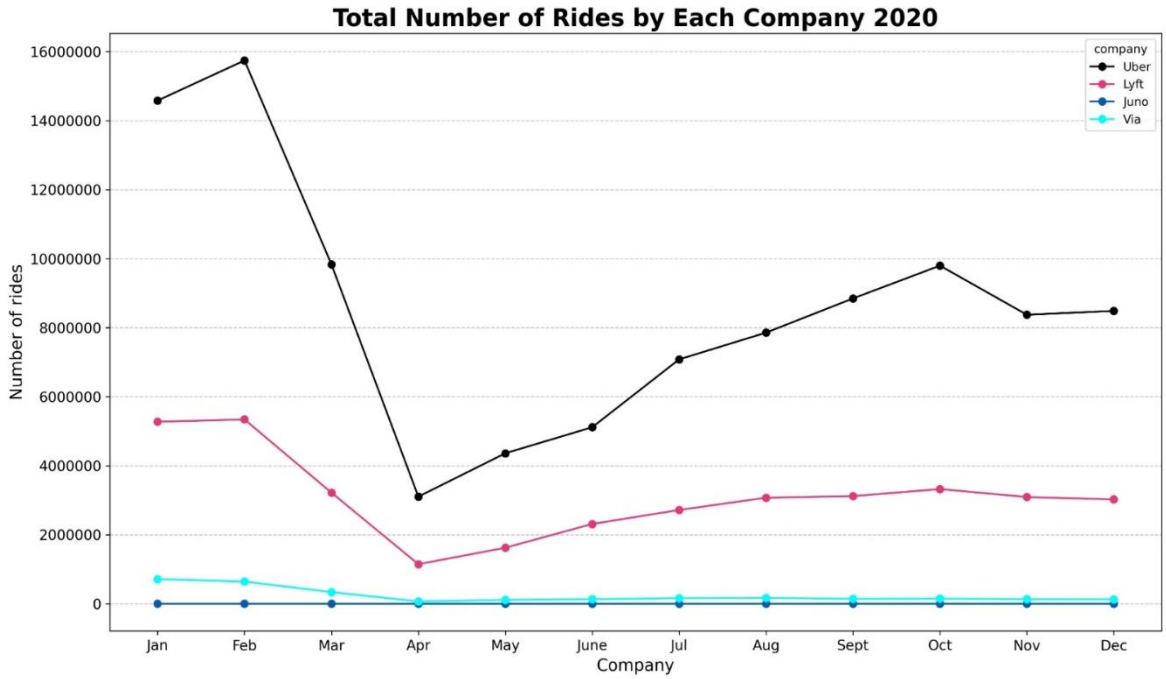
6.Analysis of Ride-Sharing Trends by Company

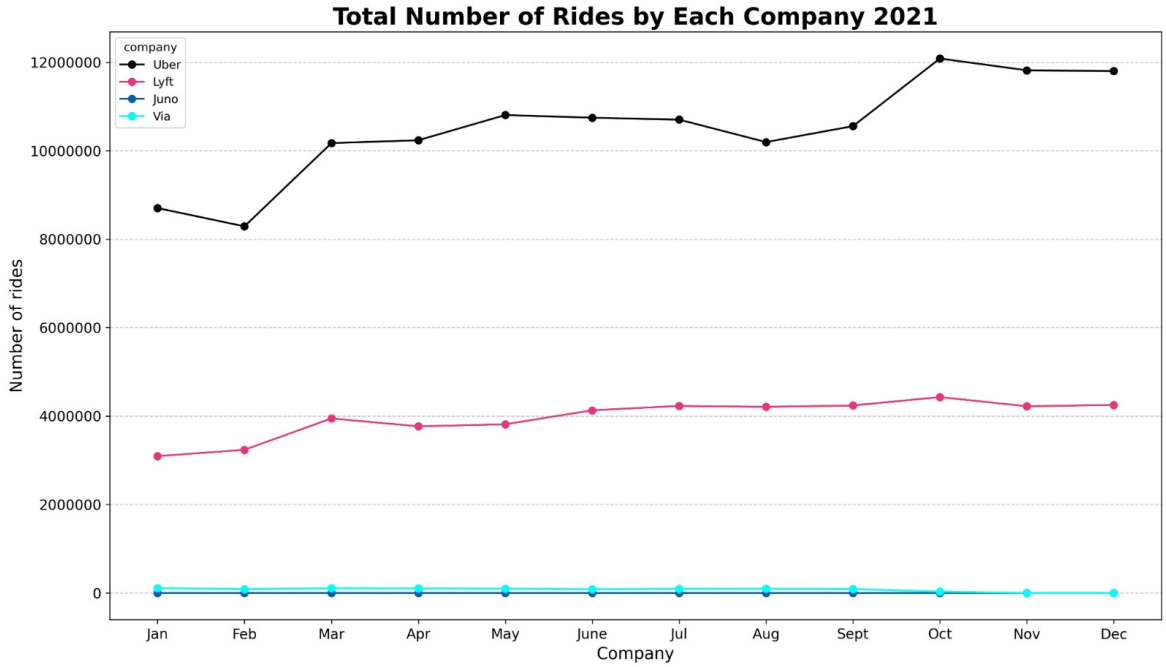
Key Points:

1. **Objective**:
    - To analyze monthly ride counts for Uber, Lyft, Juno, and Via over multiple years.
    - Visualize trends to identify dominant players, seasonal patterns, and yearly variations in ride-sharing demand.
2. **Methodology**:
    - Data from parquet files was processed using DuckDB to calculate the total number of rides for each company per month.
    - Companies were identified using the hvfhs_license_num field, and ride counts were grouped and aggregated for visualization.
    - Time-series line plots were generated to present the monthly trends for each company, with unique colors and markers for clarity.
3. **Findings**:
    - **Uber's Dominance**:
        - Uber consistently recorded the highest ride counts across all years, reflecting its market leadership.

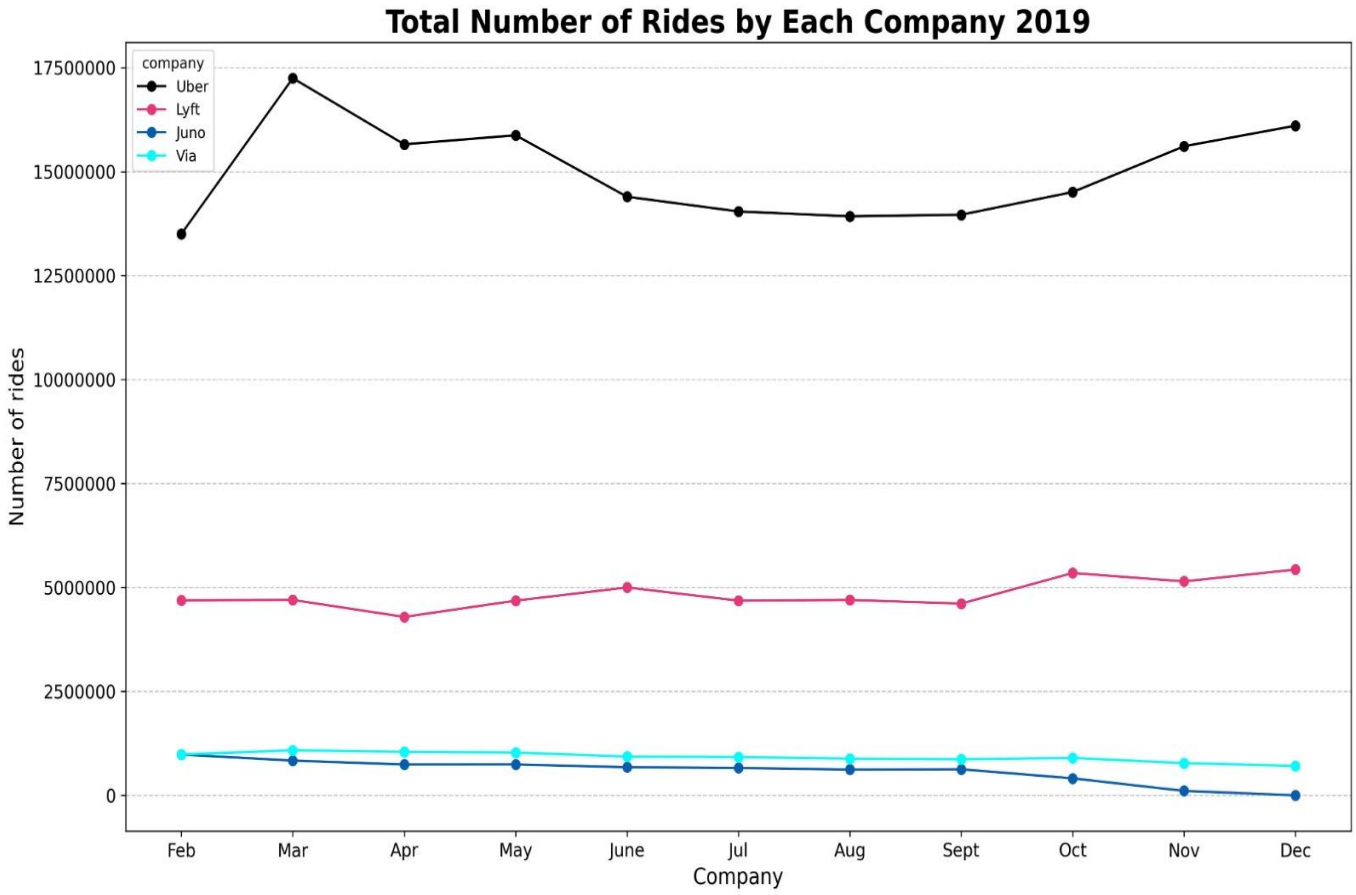
        - Monthly ride counts for Uber showed relatively stable trends, except for disruptions in 2020.
    - **Lyft's Secondary Position**:
        - Lyft maintained a stable but significantly smaller share compared to Uber.
        - Seasonal variations in Lyft's ride counts were observed, particularly in 2020 and 2023.
    - **Minimal Contributions from Juno and Via**:
        - Juno and Via had negligible ride counts across all years, indicating their limited market presence.
    - **Impact of COVID-19**:
        - 2020 showed a sharp decline in rides for all companies, with the lowest counts observed during the early months of the pandemic. Recovery began mid-year, though it remained slow.
4. **Image Explanations**:
    - **2019 Ride Trends**:
        - Uber's ride counts peaked early in the year and gradually stabilized. Lyft maintained consistent trends, while Juno and Via showed minimal activity.
    - **2020 Ride Trends**:
        - The sharp drop in rides for all companies during the pandemic is evident, particularly from March to May. Gradual recovery is seen later in the year, with Uber rebounding faster than Lyft.
    - **2021 to 2023 Ride Trends**:
        - Ride counts for Uber and Lyft regained stability, with Uber consistently outperforming. Lyft's trend shows mild seasonal fluctuations, while Juno and Via remained almost negligible.
    - **2024 Ride Trends (Up to August)**:
        - Uber maintained dominance, with ride counts showing stability. Lyft's trends were consistent, but its market share remained much smaller. Data for Juno and Via showed no significant activity.
5. **Visualization Insights**:
    - Each line plot highlights Uber's significant lead over other competitors in every month and year.
    - Lyft's performance, while secondary, shows steadiness and a moderate level of recovery post-2020.

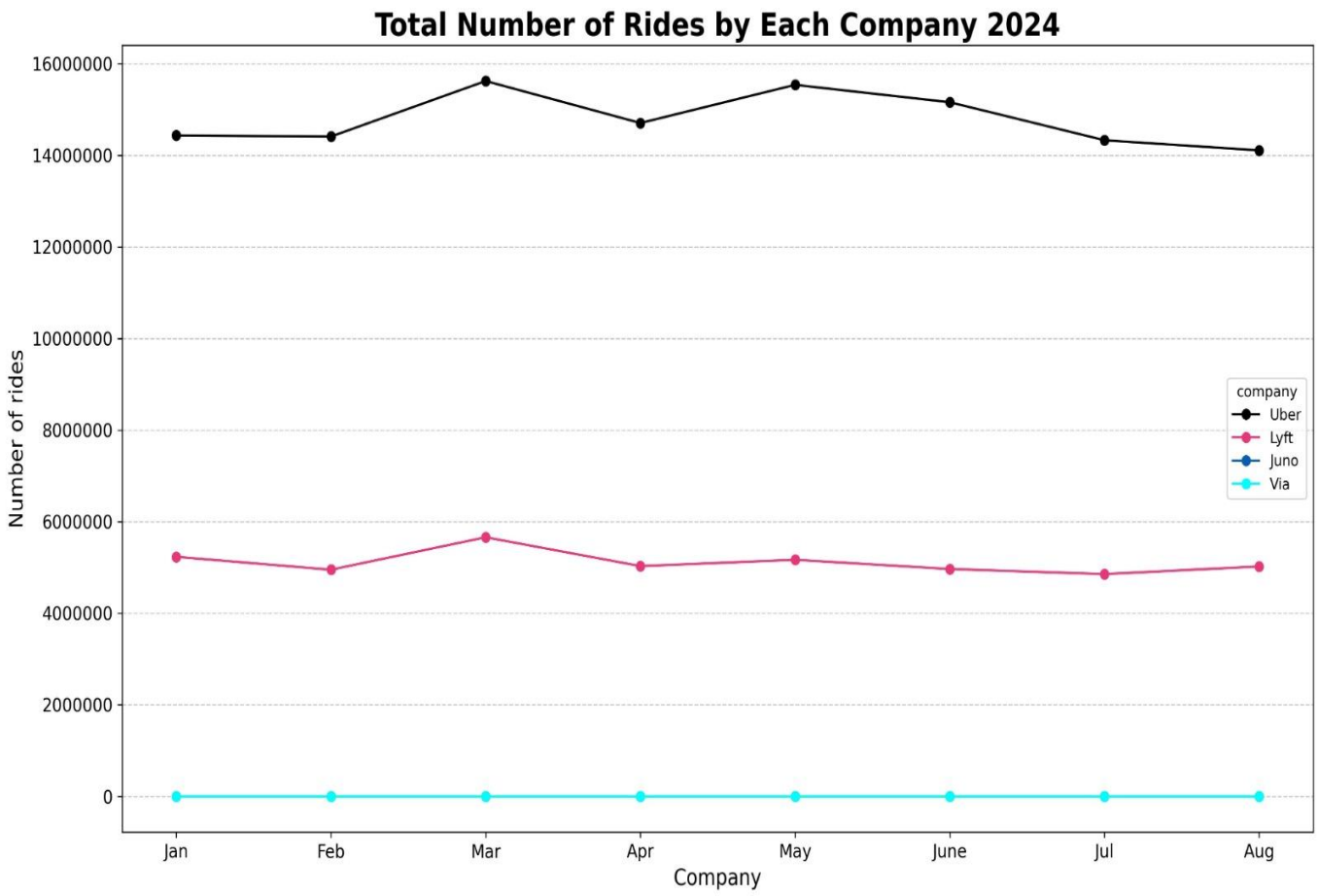
- The visualizations clearly capture the impact of COVID-19, illustrating the sharp decline and eventual recovery trends.

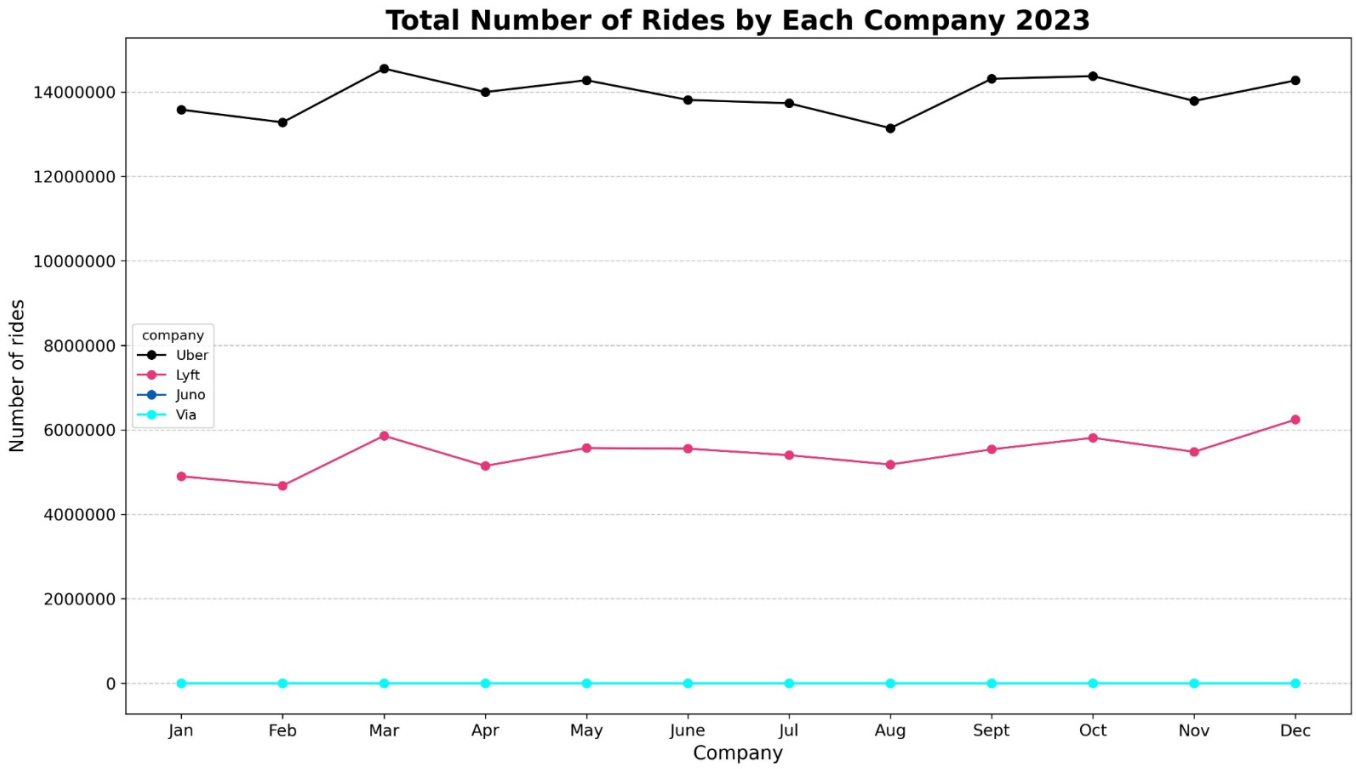
6. **Conclusion**:
   - Uber continues to dominate the ride-sharing market, with Lyft maintaining a smaller but stable presence.
   - The sharp decline in 2020 underscores the pandemic's effect on the ride-sharing industry, with gradual recovery visible in subsequent years.
   - 

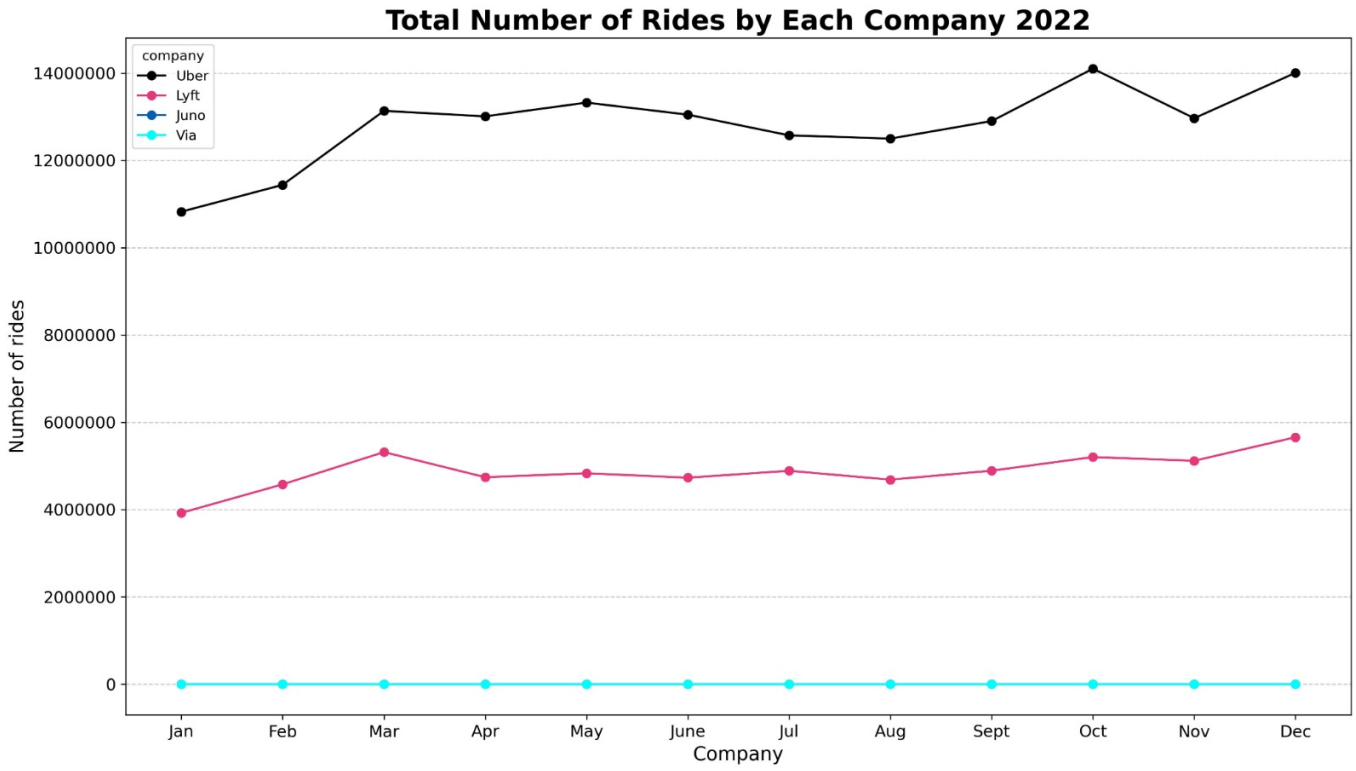
   - 

Juno


Total Number of Rides by Each Company 2019
company
Uber
Lyft
Juno
Via
17500000
15000000
12500000
10000000
7500000
5000000
2500000
0
Number of rides
Feb
Mar
Apr
May
June
Jul
Aug
Sept
Oct
Nov
Dec
Company


Total Number of Rides by Each Company 2024
16000000
14000000
12000000
10000000
8000000
6000000
4000000
2000000
0
Number of rides
company
Uber
Lyft
Juno
Via
Jan
Feb
Mar
Apr
May
June
Jul
Aug
Company





7.Explaining the Machine Learning Model Code:

1. **Data Preprocessing:**
    - Dropped rows with missing or invalid values in base_passenger_fare, pickup_datetime, and trip_miles.
    - Filtered out records with base_passenger_fare outside the range (0, 500) and non-positive values for trip_miles to remove outliers.

- Extracted time-related features:
  - Hour: Hour of the pickup.
  - Day of the Week: Numerical representation (0 for Monday, 6 for Sunday).
  - Month: Numerical representation (1 for January, 12 for December).

2. **Feature Selection:**
   - Chose trip_miles, hour, day_of_week, and month as independent features.
   - Used base_passenger_fare as the dependent variable.
3. **Dataset Splitting:**
   - Split the dataset into training (80%) and testing (20%) sets to train and evaluate the model.
4. **Model Training:**
   - Applied linear regression as the predictive model.
   - Trained the model using the training dataset.
5. **Evaluation Metrics:**
   - Calculated root mean squared error (RMSE) to measure the magnitude of prediction errors.
   - Calculated R^2 score to evaluate the proportion of variance explained by the model.
6. **Visualization:**
   - Created a scatter plot of actual vs. predicted fares on the test set.
   - Plotted a histogram to analyze the residuals (prediction errors) distribution.

Graph Explanation and Results

1. **Scatter Plot: Actual vs. Predicted Fares**
   - **Observation:**
     - The scatter plot shows predicted fares against actual fares.
     - A red dashed line represents perfect predictions (i.e., predicted = actual).
     - Most points cluster around the line, indicating that the model makes accurate predictions for a majority of cases.
   - **Interpretation:**
     - The model performs well for moderate fare values.
     - Deviations from the line, especially for higher fare values, indicate a slight decline in prediction accuracy for extreme cases.
2. **Histogram: Residuals Distribution**

- **Observation:**
  - The residuals (errors) are plotted, showing a near-normal distribution centered around zero.
  - The majority of residuals are close to zero, indicating minimal errors.
- **Interpretation:**
  - The model's errors are unbiased and symmetrically distributed, satisfying the assumptions of linear regression.
  - Larger residuals for some extreme cases might suggest the need for a more complex model to handle outliers effectively.

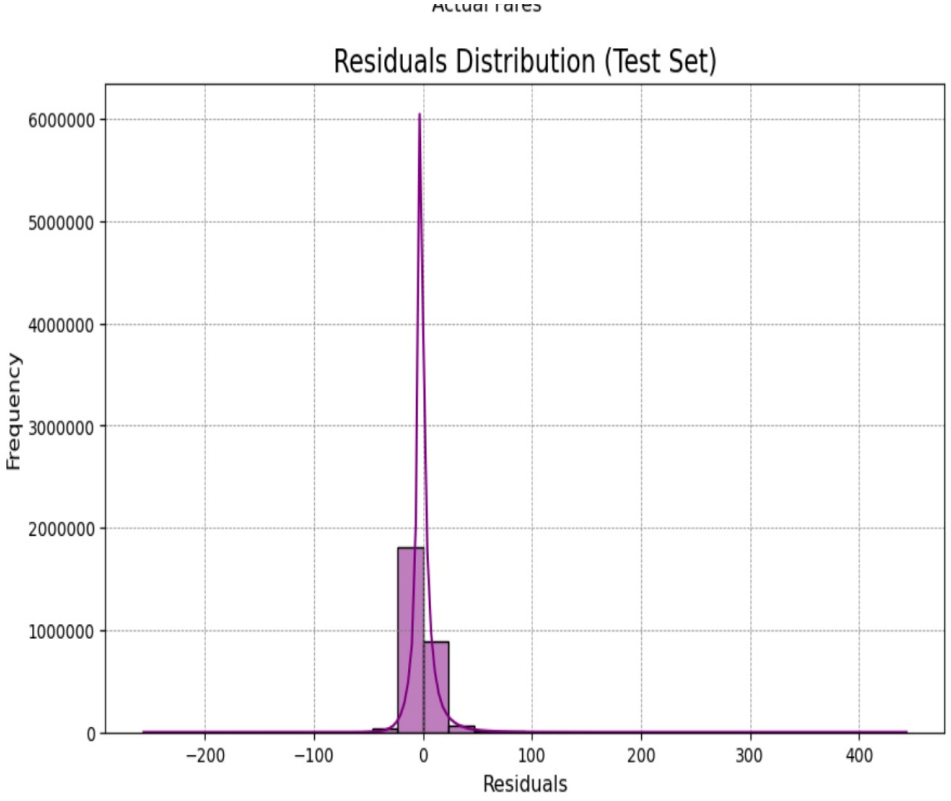
Key Results

- Training RMSE: 11.38
- Testing RMSE: 11.36
- Training R^2: 0.765
- Testing R^2: 0.765

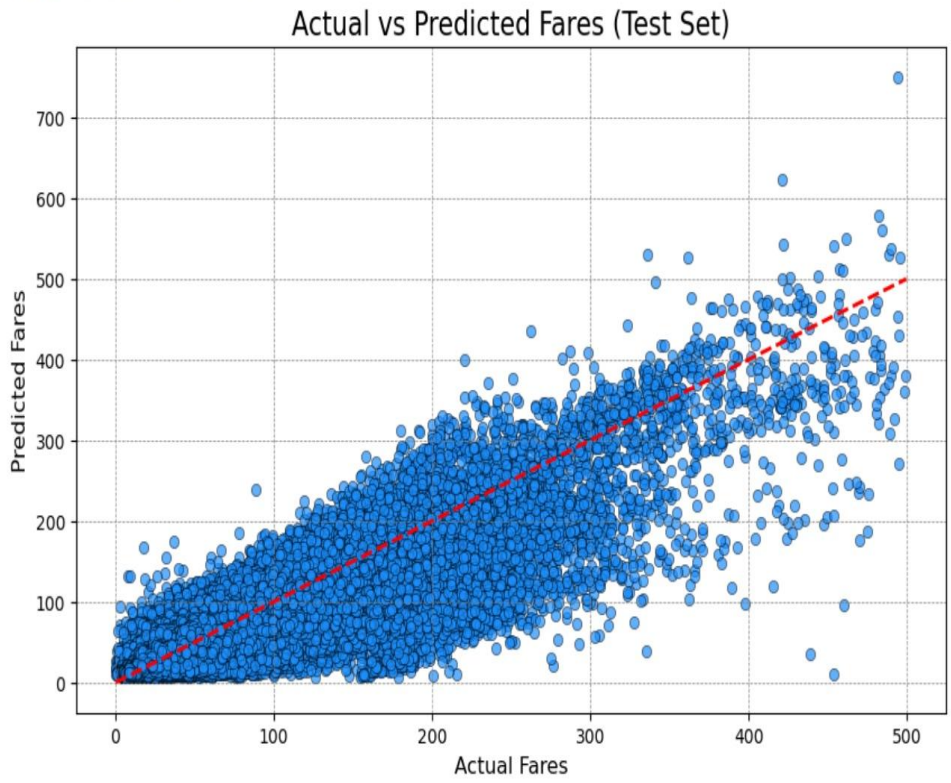
The model explains approximately 76.5% of the variance in base_passenger_fare.

What We Can Interpret

1. **Performance:**
   - The model performs consistently on both training and testing datasets, suggesting no overfitting.
   - The R^2 score indicates that a significant portion of the variability in fares is explained by the selected features.
2. **Errors:**
   - The normal distribution of residuals implies no systematic bias.
   - Slight inaccuracies in predicting extreme fare values highlight the limitations of linear regression.
3. **Improvement Opportunities:**
   - Consider adding more features such as pickup and drop-off locations or weather data.
   - Explore non-linear models like decision trees, random forests, or gradient boosting for better handling of complex relationships in the data.



Training RMSE: 11.379595956093752
Testing RMSE: 11.361645224662459
Training R²: 0.7654488809093806
Testing R²: 0.7650304284339348



8. Results and Discussion

8.1 Ride Volume Trends (2020-2024)

1. Uber:
    - Maintained growth year-over-year, particularly post-2021, with strong recovery post-pandemic.

- Annual ride volume exceeded 140 million by 2024, indicating sustained dominance.

2. Lyft:
    - Demonstrated consistent growth, with notable increases in 2023 and 2024, reflecting a 20% rise in ride volume from 2022 to 2024.
3. Via:
    - Ride volume gradually declined through 2021, ceasing entirely after December 2021.

8.2 Market Dynamics Post-Via Exit

- Via's exit resulted in redistributed demand, primarily absorbed by Uber and Lyft.
- Uber capitalized more effectively on this shift due to its established market leadership.

8.3 Borough-Level Analysis

1. Manhattan: Highest ride demand due to business and tourist activities.
2. Brooklyn: Steady growth in ride volume, attributed to expanding residential and commercial hubs.
3. Queens: Spikes in ride-hailing demand correlated with airport traffic.
4. The Bronx: Periodic demand tied to events and local attractions.
5. Staten Island: Lower demand compared to other boroughs, consistent with its suburban profile.

9. Analysis of Top Pickup and Drop-Off Zones in NYC

Image 1: Top 10 Most Popular Pickup Zones in NYC

- This visualization highlights the most frequent pickup zones in New York City.
- The top pickup locations include major transportation hubs and busy urban areas like **LaGuardia Airport**, **JFK Airport**, and **East Village**.
- Airports such as LaGuardia and JFK dominate the pickup data, reflecting high passenger traffic in these regions due to their significance in commuting and tourism.
- Other popular zones, such as **Crown Heights North** and **Times Square/Theatre District**, indicate their importance as residential and tourist hubs.
- The insights suggest that rideshare services are extensively used in locations with dense foot traffic, tourism, and residential activities.

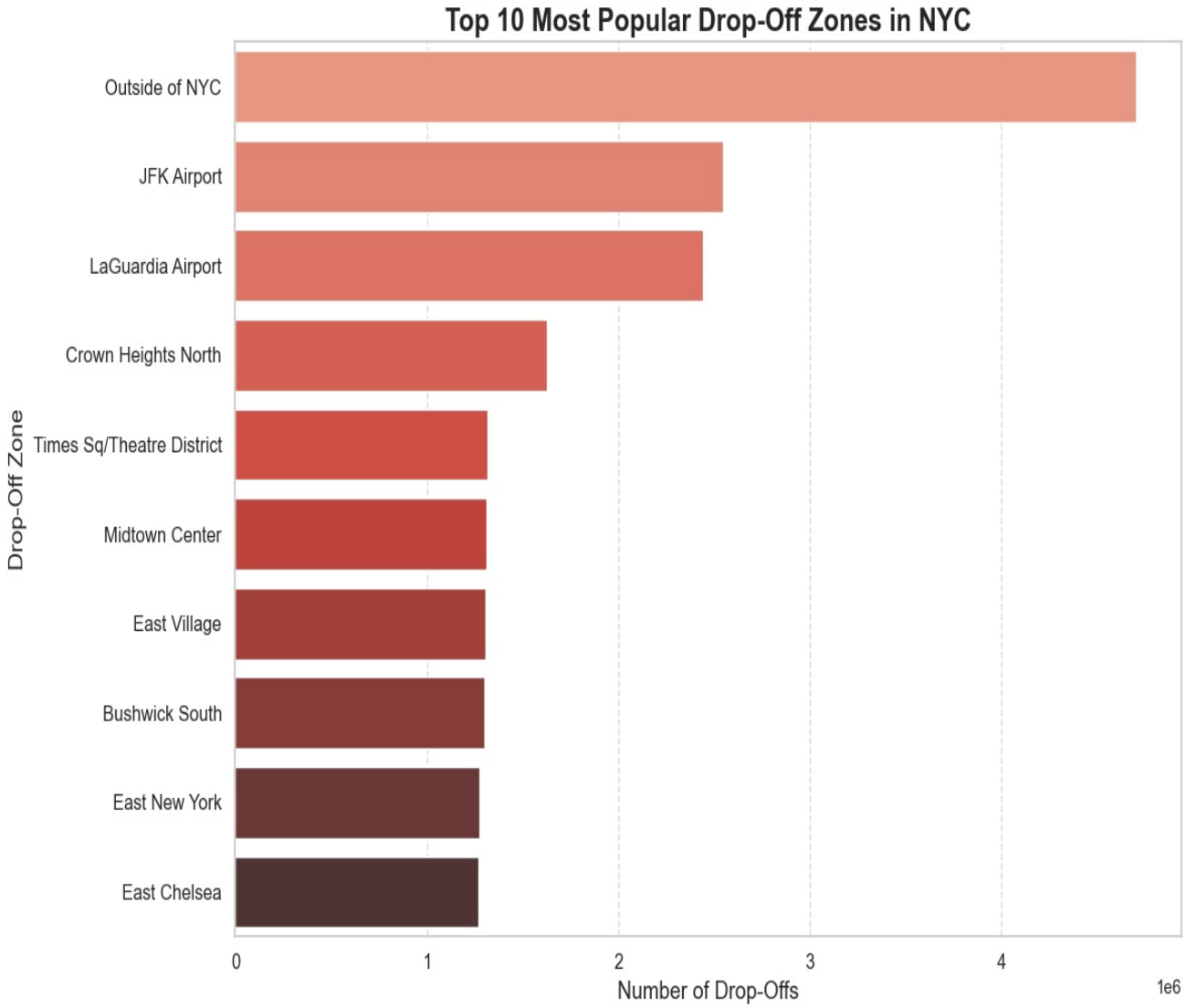
Image 2: Top 10 Most Popular Drop-Off Zones in NYC

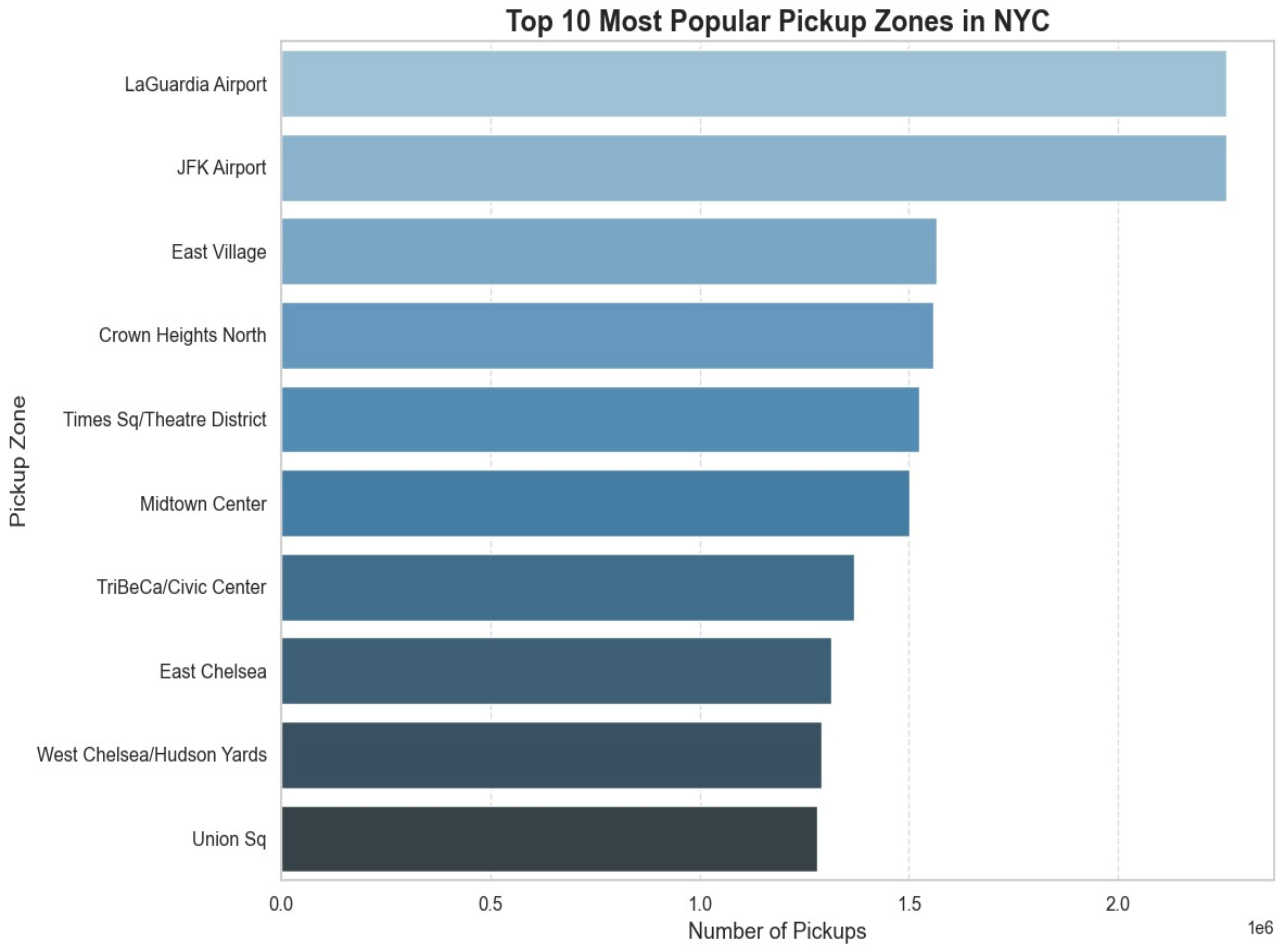
- This chart displays the most frequent drop-off zones in New York City.
- The **"Outside of NYC"** category emerges as the most common drop-off location, which might indicate intercity rides or long-distance travel.

- Airports like **JFK** and **LaGuardia** are also among the top drop-off locations, aligning with the heavy travel demand at these sites.
- Popular city areas such as **Crown Heights North**, **Times Square/Theatre District**, and **Midtown Center** also rank high, showing consistent urban activity in these zones.
- These trends underline the importance of rideshare services in connecting key business, tourist, and transportation hubs in NYC.

Overall Insights:

- The patterns in pickup and drop-off data provide valuable information for optimizing fleet management, improving customer service, and targeting high-demand zones for operational efficiency in rideshare services.





10. Conclusion

The analysis presented in this report highlights the evolution of the NYC ride-hailing market from 2019 to 2024, shaped by significant events such as the COVID-19 pandemic, regulatory changes, and market exits. Key findings include:

1. **Dominance of Uber and Lyft**: Uber consistently led the market with substantial ride volumes, while Lyft maintained a stable secondary position. Juno and Via had limited market impact, with their exit further solidifying Uber and Lyft's dominance.
2. **Resilience and Recovery**: The industry demonstrated remarkable resilience post-pandemic, with gradual recovery in ride volumes by 2021 and sustained growth thereafter. Consumer reliance on ride-hailing services for leisure and travel remained evident, particularly on weekends.
3. **Temporal and Spatial Patterns**:
    - Afternoon and evening trips were the most popular, driven by commuting, leisure, and social activities.
    - Morning trips were largely influenced by work and airport travel, while nighttime usage was minimal.
    - Manhattan consistently showed the highest demand, reflecting its business and tourist activity, while suburban boroughs like Staten Island had lower ride-hailing activity.

4. **Machine Learning Insights**: A linear regression model effectively predicted ride fares based on trip distance, time of day, and day of the week. The model achieved strong performance, with opportunities for improvement through additional features or advanced algorithms to address outliers.
5. **Future Opportunities**: Companies could further optimize service availability during peak demand periods and leverage advanced machine learning techniques to enhance operational efficiency. Additionally, emerging technologies like electric and autonomous vehicles present avenues for future growth and innovation.