NYC Transportation Pattern Analysis

Hrithik Dhoka (hhd2023), Anuva Sehgal (as18774)

Abstract

This project analyzes New York City's transportation patterns by examining subway and taxi ridership data across different boroughs, times, and days in 2023. Through comprehensive analysis of ridership volumes, revenue patterns, and temporal trends, the study reveals critical insights about urban mobility preferences and identifies key patterns in transportation choice behavior. The analysis focuses particularly on the complementary nature of subway and taxi services, highlighting how these two modes of transport serve different needs throughout the day and across different areas of the city.

Introduction

New York City's transportation system, a complex network of subways and taxis, serves as the lifeline for millions of daily commuters, tourists, and residents. This study delves into the intricate patterns of how these two primary modes of transportation interact and complement each other across different times, locations, and circumstances. By analyzing comprehensive data from both the Metropolitan Transportation Authority (MTA) and TLC yellow taxi services, we aim to uncover patterns that can inform better transit planning and policy decisions.

The analysis focuses on several key aspects: temporal patterns (both daily and hourly), geographical variations across boroughs, revenue comparisons, and special circumstance scenarios such as late-night transportation preferences. We examine how factors such as time of day, location, and external circumstances influence citizens' choices between subway and taxi services. The study pays particular attention to anomalies in usual patterns, such as instances where taxi ridership exceeds subway usage, or where revenue patterns deviate from typical expectations.

Through this analysis, we seek to understand not just the raw numbers, but the underlying patterns that drive transportation choices in different contexts. This understanding is crucial for optimizing the city's transportation resources and improving service delivery to meet diverse needs across different boroughs and time periods.

Motivation

In the dynamic landscape of New York City's transportation system, understanding the intricate relationship between subway and taxi services is crucial for optimal urban mobility. This analysis aims to provide valuable insights for multiple stakeholders in the transportation ecosystem. Transit planners and policy makers can utilize these findings to optimize service schedules and resource allocation, while transportation operators can better understand demand patterns to improve service delivery. The business community, particularly in transit-heavy areas, can leverage this information for operational planning.

Being NYC residents and regular users of these services ourselves, there was also a personal motivation to undertake this analysis, as it allows for intuitive validation of the insights. This ensures not only the analytical rigor but also the practical relevance and veracity of the findings.

Most importantly, this analysis benefits NYC residents, commuters, and visitors by highlighting how different transportation modes complement each other, potentially leading to improved service integration and accessibility. The findings are particularly valuable for late-night workers and businesses in the entertainment sector, as they reveal critical patterns in off-peak transportation needs. Furthermore, understanding these patterns is essential for urban development planning and enhancing the overall efficiency of the city's transportation network, ultimately contributing to a more sustainable and accessible urban environment.

Data Sources

Yellow Taxi trip data (2023, size: 4 GB):

Description: Yellow taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts(as shown on the left below). The data was from January (01) to December (12) in `.parquet` format. It was first converted to `.csv` using Python to make it consumable for the MapReduce job. Additionally, the data needed to be joined with a lookup table containing borough and zone information for location IDs to enhance its usability. The final joined table included the columns listed above.

pickup_location_id	pickup_datetime	passenger_count	payment_type	tip_amount	total_amount	borough	zone
113	2023-02-11 13:30:14.0	1.0	1	5.88	35.28	Manhattan	Greenwich Village North
113	2023-02-12 01:14:39.0	1.0	1	5.8	34.8	Manhattan	Greenwich Village North
113	2023-02-10 16:01:05.0	1.0	1	3.2	19.3	Manhattan	Greenwich Village North
113	2023-02-07 14:15:25.0	1.0	1	2.98	14.88	Manhattan	Greenwich Village North
113	2023-02-07 14:34:29.0	1.0	1	1.0	18.5	Manhattan	Greenwich Village North

The joining was done later, in hive and the lookup table was as shown in the images below.

This data dictionary describes yellow taxi trip data. For a dictionary describing green taxi data, or a map of the TLC Taxi Zones, please visit http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. **1= Standard rate** **2=JFK** **3=Newark** **4=Nassau or Westchester** **5=Negotiated fare** **6=Group ride**
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**
Payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.
Congestion_Surcharge	Total amount collected in trip for NYS congestion surcharge.
Airport_fee	$1.25 for pick up only at LaGuardia and John F. Kennedy Airports

taxi_zone_lookup

LocationID	Borough	Zone	service_zone
1	EWR	Newark Airport	EWR
2	Queens	Jamaica Bay	Boro Zone
3	Bronx	Allerton/Pelham Gardens	Boro Zone
4	Manhattan	Alphabet City	Yellow Zone
5	Staten Island	Arden Heights	Boro Zone
6	Staten Island	Arrochar/Fort Wadsworth	Boro Zone
7	Queens	Astoria	Boro Zone
8	Queens	Astoria Park	Boro Zone
9	Queens	Auburndale	Boro Zone
10	Queens	Baisley Park	Boro Zone
11	Brooklyn	Bath Beach	Boro Zone
12	Manhattan	Battery Park	Yellow Zone
13	Manhattan	Battery Park City	Yellow Zone
14	Brooklyn	Bay Ridge	Boro Zone
15	Queens	Bay Terrace/Fort Totten	Boro Zone
16	Queens	Bayside	Boro Zone
17	Brooklyn	Bedford	Boro Zone
18	Bronx	Bedford Park	Boro Zone
19	Queens	Bellerose	Boro Zone
20	Bronx	Belmont	Boro Zone
21	Brooklyn	Bensonhurst East	Boro Zone
22	Brooklyn	Bensonhurst West	Boro Zone
23	Staten Island	Bloomfield/Emerson Hill	Boro Zone
24	Manhattan	Bloomingdale	Yellow Zone
25	Brooklyn	Boerum Hill	Boro Zone
26	Brooklyn	Borough Park	Boro Zone
27	Queens	Breezy Point/Fort Tilden/Riis Beach	Boro Zone
28	Queens	Briarwood/Jamaica Hills	Boro Zone
29	Brooklyn	Brighton Beach	Boro Zone
30	Queens	Broad Channel	Boro Zone
31	Bronx	Bronx Park	Boro Zone

The remaining columns were discarded either due to irrelevance or because subway stations lack drop-off information, making drop-off-related columns unnecessary for this analysis.
The MTA subway data and yellow taxi data had inconsistent zone definitions. The subway data used streets and avenues to mark zones, while the taxi data employed a different system. As a result, we had to drop the zone and location ID from the dataset. Additionally, to align the data with the MTA data, we split the datetime into separate date and hour components, enabling a more accurate comparison.

MapReduce and Hive changes of Taxi Data:

The Mapper process followed the usual steps of cleaning and dropping irrelevant columns. Since each month had a separate CSV file, the input to the Mapper was a directory containing all the monthly CSV files. During processing, rows with negative fare amounts and empty

passenger counts were encountered, requiring special handling. Rows with missing passenger counts were filled with a value of 1.0, while rows with negative fare amounts were dropped.

The Mapper logic did not change to accommodate Hive integration. However, the Reducer process before uploading to Hive initially outputted the pickup location ID as the key, with a list of values containing a string for each entry. The string included the pickup datetime, passenger count, payment type, tip amount, and total amount. Initially, the code used three reducers, but for simplicity, it was combined into a single reducer.

This data format was then modified to integrate with Hive. Instead of the usual "***key, list<values>***" format, it was modified to ***"key1, value1 \n key1, value2 \n key1, value3,..."*** with each key-value pair on a new line. Consequently, the reducer's key output was made to be a NullWritable and the combiner was removed in the final iteration. This output was then uploaded to a Hive table, where SQL joins and additional cleaning and profiling were performed as required to prepare the data for merging with the MTA data.

In Hive, due to discrepancies in zone data between the datasets, a finer-grained analysis based on zones was not feasible. Consequently, the analysis began with a high-level borough-wise comparison of passenger volume and revenue for both subway and taxi data. This provided a foundational understanding of overall trends. Once these broad patterns were established, the focus shifted to detailed insights from the annual data, examining dimensions such as month, hour, and day of the week. Combining these aspects with borough-level information offered a robust baseline for comparison.

To align the datasets, the taxi data was split into separate date and time columns, and timestamps were further truncated to hours to match the hourly granularity of the MTA data. With these foundational columns prepared, the analysis centered around key SQL joins, groupings, and aggregations to address important questions such as:

- Which boroughs show a stronger preference for taxis over the subway?
- What times of day see a higher preference for taxis over the subway?
- During which months do more people prefer taxis to the subway?
- How does the revenue correlate with passenger volume in taxi data, and what might explain these patterns?

This structured approach ensured that the insights derived were meaningful and addressed the core objectives of the project.

MTA Subway Hourly Ridership (2023)

Description: This dataset provides subway ridership estimates on an hourly basis by subway station complex and class of fare payment.

Data Size: 3.56 GB

Data Sample:

transit_timestamp	transit_mode	station_complex_id	station_complex	borough	payment_method	fare_class_category	ridership
05/17/2024 02:00:00 AM	subway	120	Bedford Av (L)	Brooklyn	metrocard	Metrocard - Fair Fare	2
05/17/2024 05:00:00 AM	subway	120	Bedford Av (L)	Brooklyn	metrocard	Metrocard - Fair Fare	3
05/17/2024 09:00:00 AM	subway	120	Bedford Av (L)	Brooklyn	metrocard	Metrocard - Fair Fare	11
05/17/2024 02:00:00 PM	subway	122	Graham Av (L)	Brooklyn	metrocard	Metrocard - Fair Fare	12
05/17/2024 12:00:00 AM	subway	123	Grand St (L)	Brooklyn	metrocard	Metrocard - Unlimited 30-Day	2
05/17/2024 07:00:00 PM	subway	119	1 Av (L)	Manhattan	omny	OMNY - Seniors & Disability	4
05/17/2024 08:00:00 PM	subway	120	Bedford Av (L)	Brooklyn	metrocard	Metrocard - Fair Fare	20
05/17/2024 09:00:00 PM	subway	122	Graham Av (L)	Brooklyn	omny	OMNY - Full Fare	207
05/17/2024 10:00:00 AM	subway	123	Grand St (L)	Brooklyn	metrocard	Metrocard - Students	5
05/17/2024 05:00:00 AM	subway	123	Grand St (L)	Brooklyn	metrocard	Metrocard - Unlimited 30-Day	6
05/17/2024 05:00:00 AM	subway	169	Canal St (A,C,E)	Manhattan	metrocard	Metrocard - Full Fare	8

MTA Subway Hourly Ridership

Data Dictionary

Data Label	Data Type	Data Description
transit_timestamp	DATE	Timestamp payment took place in local time. All transactions here are rounded **down** to the nearest hour. For example, a swipe that took place at 1:37pm will be reported as having taken place at 1pm.
transit_mode	TEXT	Distinguishes between the subway, Staten Island Railway, and the Roosevelt Island Tram
station_complex_id	ALPHANUMERIC	A unique identifier for station complexes
station_complex	TEXT	The subway complex where an entry swipe or tap took place. Large subway complexes, such as Times Square and Fulton Center, may contain multiple subway lines. The subway complex name includes the routes that stop at the complex in parenthesis, such as Zerega Av (6).
borough	TEXT	Represents one of the boroughs of New York City serviced by the subway system (Bronx, Brooklyn, Manhattan, Queens).
payment_method	TEXT	Specifies whether the payment method used to enter was from OMNY or MetroCard.
fare_class_category	TEXT	The class of fare payment used for the trip. The consolidated categories are: • MetroCard – Fair Fare • MetroCard – Full Fare • MetroCard – Other • MetroCard – Senior & Disability • MetroCard – Students • MetroCard – Unlimited 30-Day • MetroCard – Unlimited 7-Day • OMNY – Full Fare • OMNY – Other • OMNY – Seniors & Disabilities
ridership	NUMERIC	Total number of riders that entered a subway complex via OMNY or MetroCard at the specific hour and for that specific fare type.
transfers	NUMERIC	Number of individuals who entered a subway complex via a free bus-to-subway, or free out-of-network transfer. This represents a **subset** of total ridership, meaning that these transfers are already included in the preceding ridership column. Transfers that take place within a subway complex (e.g., individuals transferring from the 2 to the 4 train within Atlantic Avenue) are not captured here.
latitude	DECIMAL	Latitude for the specified subway complex
longitude	DECIMAL	Longitude for the specified subway complex

MapReduce and Hive changes of Subway Data:

The MapReduce and Hive implementation focused on processing and analyzing the MTA subway ridership data across New York City's boroughs. The Mapper process handled the initial data cleaning and transformation, processing input files containing hourly ridership counts, fare collections, and station-specific information.

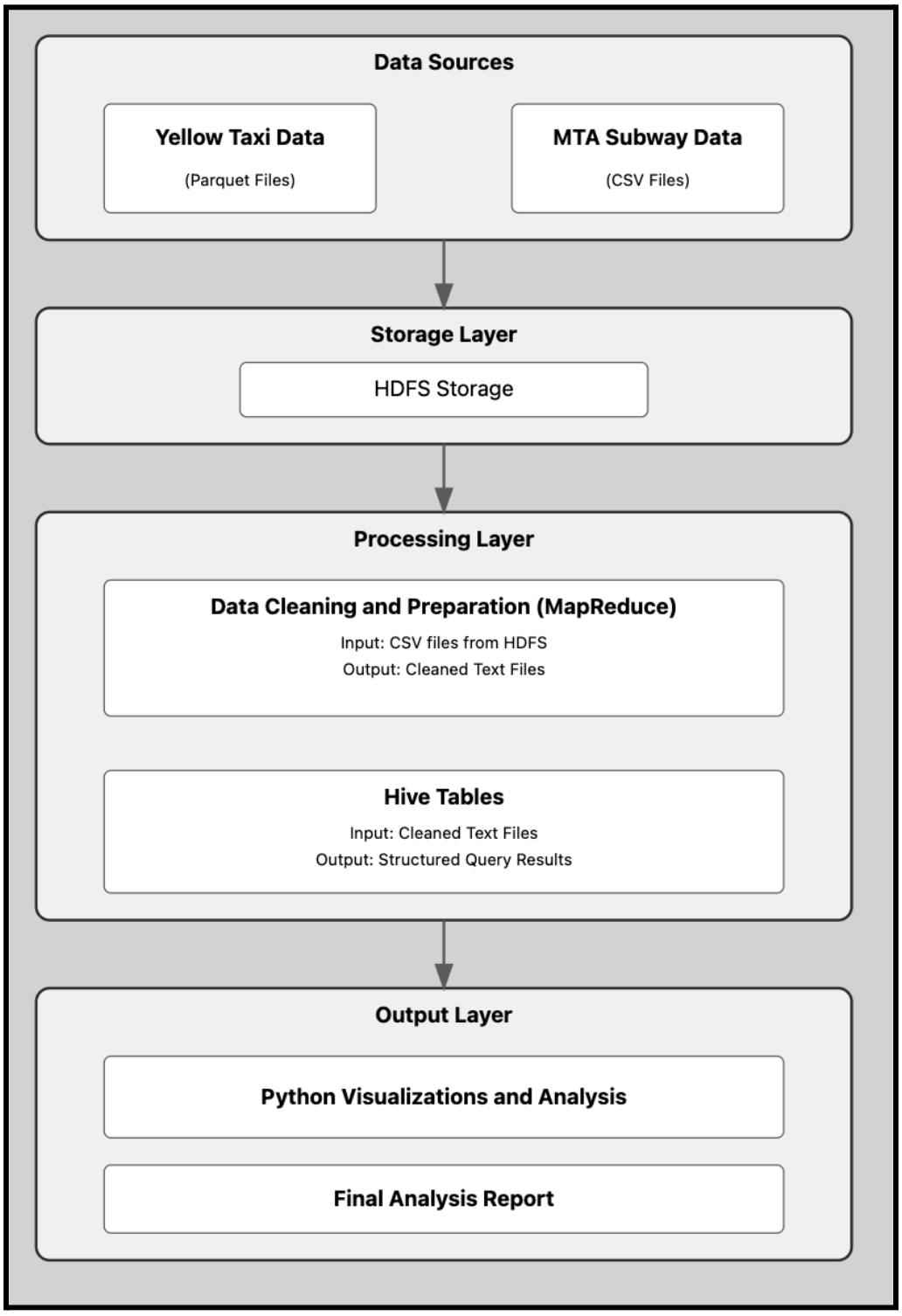
The Mapper implementation carefully handled data quality issues, particularly focusing on station complex identifiers and borough assignments. Records with missing borough information or invalid station complexes were flagged for removal. The timestamp data required special processing to standardize the format and extract hour-level granularity for consistent analysis. The Mapper emitted key-value pairs with borough and timestamp combinations as keys, along with ridership counts and revenue calculations as values.

The Reducer process initially aggregated the data by borough and hour, computing total ridership and revenue metrics. This was later enhanced to include daily and weekly patterns by incorporating day-of-week calculations. The revenue calculations in the reducer applied the standard MTA fare structure, accounting for different fare types and transfer patterns. The reducer output was structured to facilitate direct loading into Hive tables, with each record containing the standardized date, hour, borough, ridership count, and calculated revenue fields.

In Hive, the data underwent further transformation to enable comprehensive temporal analysis. The implementation focused on creating summary tables that aggregated ridership and revenue patterns at various granularities - hourly, daily, and monthly. These tables formed the foundation for analyzing peak usage patterns and revenue generation across different boroughs. The analysis particularly focused on understanding how ridership patterns varied across different times of day and days of the week, providing insights into commuter behavior and system utilization.

The integration of borough-level analysis with temporal patterns revealed important trends in subway usage across New York City's geography. This structured approach to data processing and analysis ensured that the insights derived were both statistically sound and operationally relevant, contributing to a deeper understanding of public transportation patterns in the city.

Design Diagram



Challenges Encountered

In developing this comprehensive analysis of NYC's transportation patterns, several significant technical and data-related challenges were encountered. The fundamental challenge lay in the heterogeneous nature of our primary data sources. The taxi data, sourced from the Taxi data, was structured with zone-level information and provided in Parquet file format with separate monthly files. In contrast, the MTA data contained street-station level information, with each dataset utilizing entirely different identification systems. This disparity in geographical referencing made direct spatial comparisons particularly challenging.

The data standardization process was complex for both datasets. The taxi data required conversion from Parquet to CSV format and temporal aggregation of multiple monthly files. Similarly, the MTA data needed extensive preprocessing to standardize station information and align it with broader geographical zones. The absence of a common geographical identifier between street-stations and taxi zones necessitated aggregation to the borough level, sacrificing granular spatial insights in favor of maintaining analytical validity. This mismatch in spatial granularity between street-level stations and taxi zones prevented us from conducting more detailed neighborhood-level analysis that could have revealed important micro-level transportation patterns.

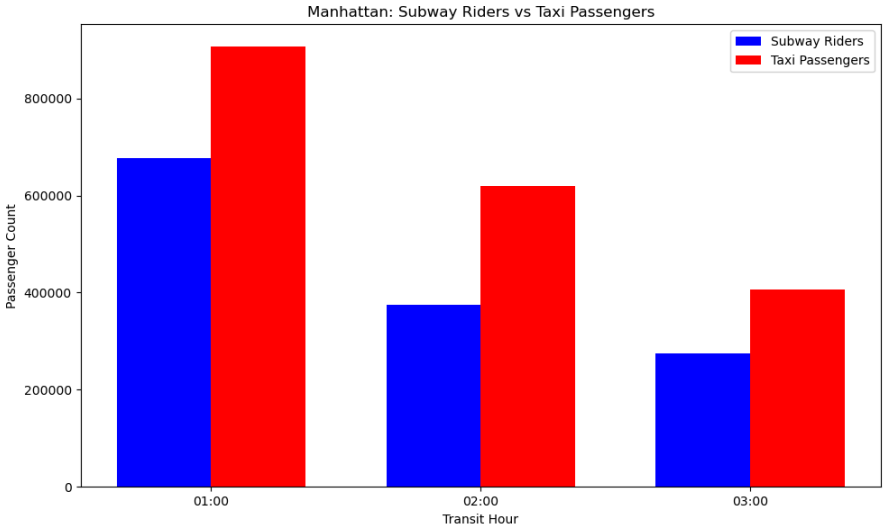
To facilitate meaningful comparative analysis, several data transformations were necessary. Temporal aggregations were particularly challenging, requiring careful handling of timestamp data from both sources to ensure consistent hourly and daily patterns. For revenue analysis, we implemented a derived calculation for the subway system by multiplying individual rides by the standard fare of $2.90, as direct revenue data was not available in the MTA dataset. Additionally, both datasets required modification of our Mapper and Reducer components to ensure compatibility with Hive's data storage requirements, adding another layer of complexity to the data preparation process.

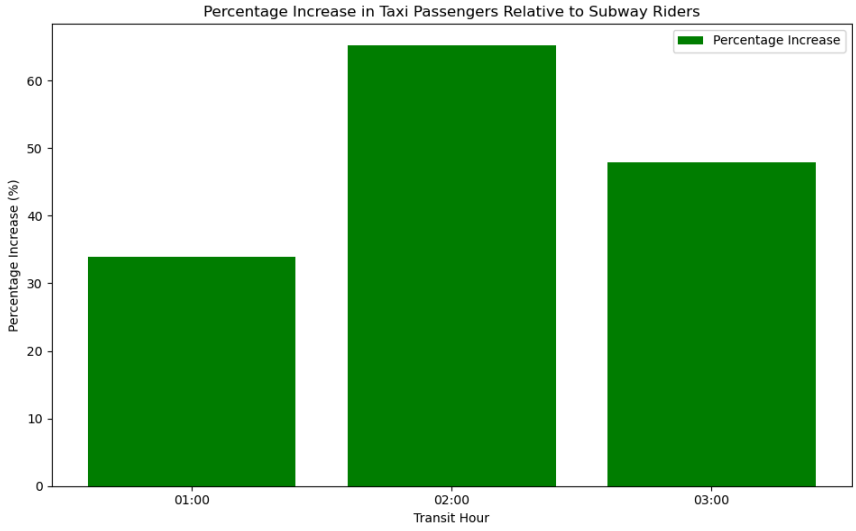
Despite these complications, we were able to derive meaningful insights by carefully structuring our analysis around the available data points and implementing appropriate data transformation and standardization procedures. The addition of derived fields such as day of the week and calculated revenue metrics helped bridge some of the gaps in the raw data, enabling us to construct a comprehensive view of NYC's transportation dynamics at the borough level.

Results

Insight Number 1:

Aggregating data across all hours and boroughs, it appears that only in Manhattan, during the early hours of 1 AM to 3 AM, do people prefer taking taxis over riding the subway. This trend is observed consistently across all dates in 2023. The preference for taxis during these hours may indicate that subway services are less frequent or accessible during the late night to early morning period, prompting commuters to opt for taxis instead. Additionally, this could reflect a shift in rider behavior during off-peak hours, where the safety, convenience and availability of taxis outweigh the typically lower cost and more extensive reach of the subway system.

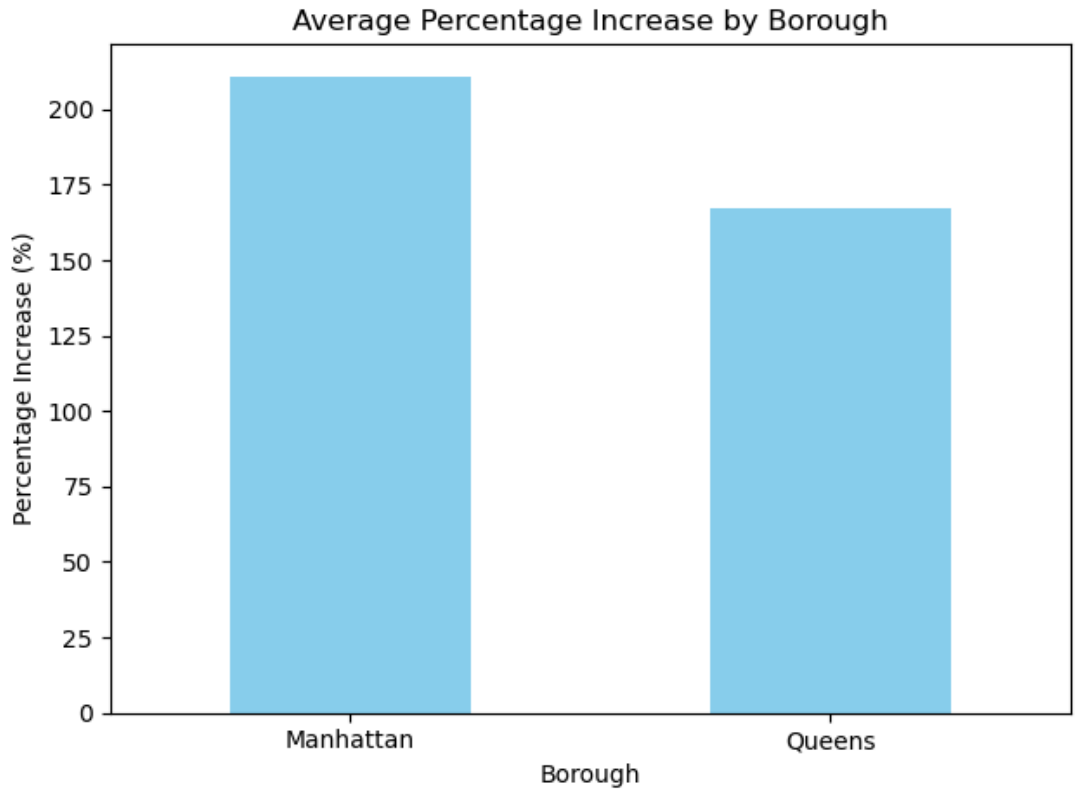


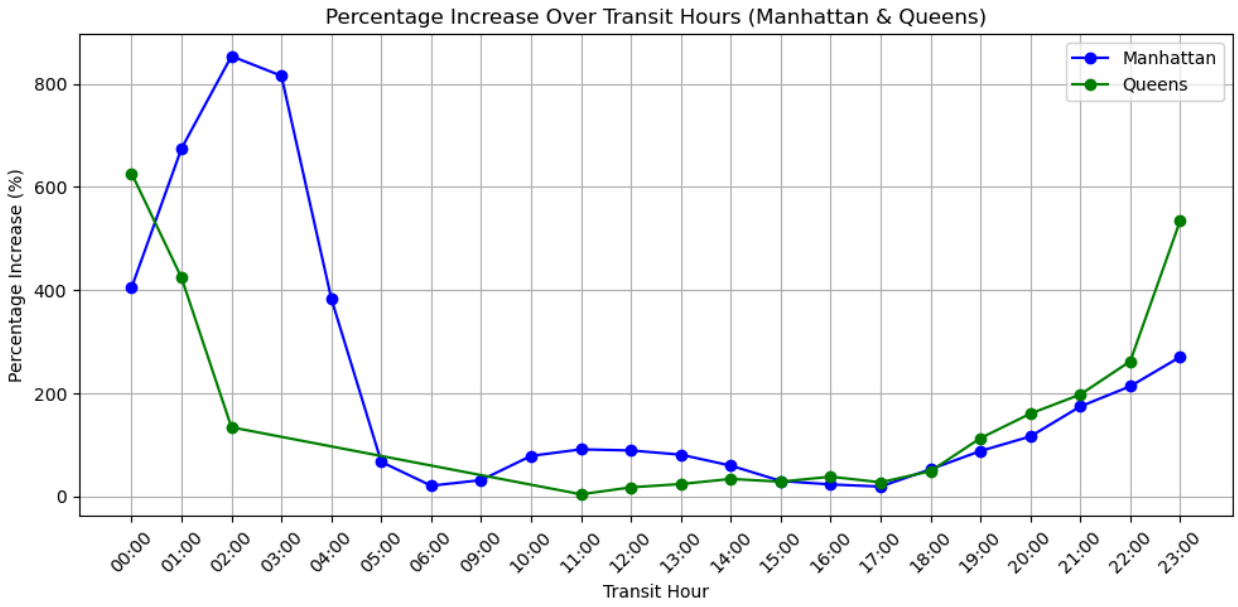


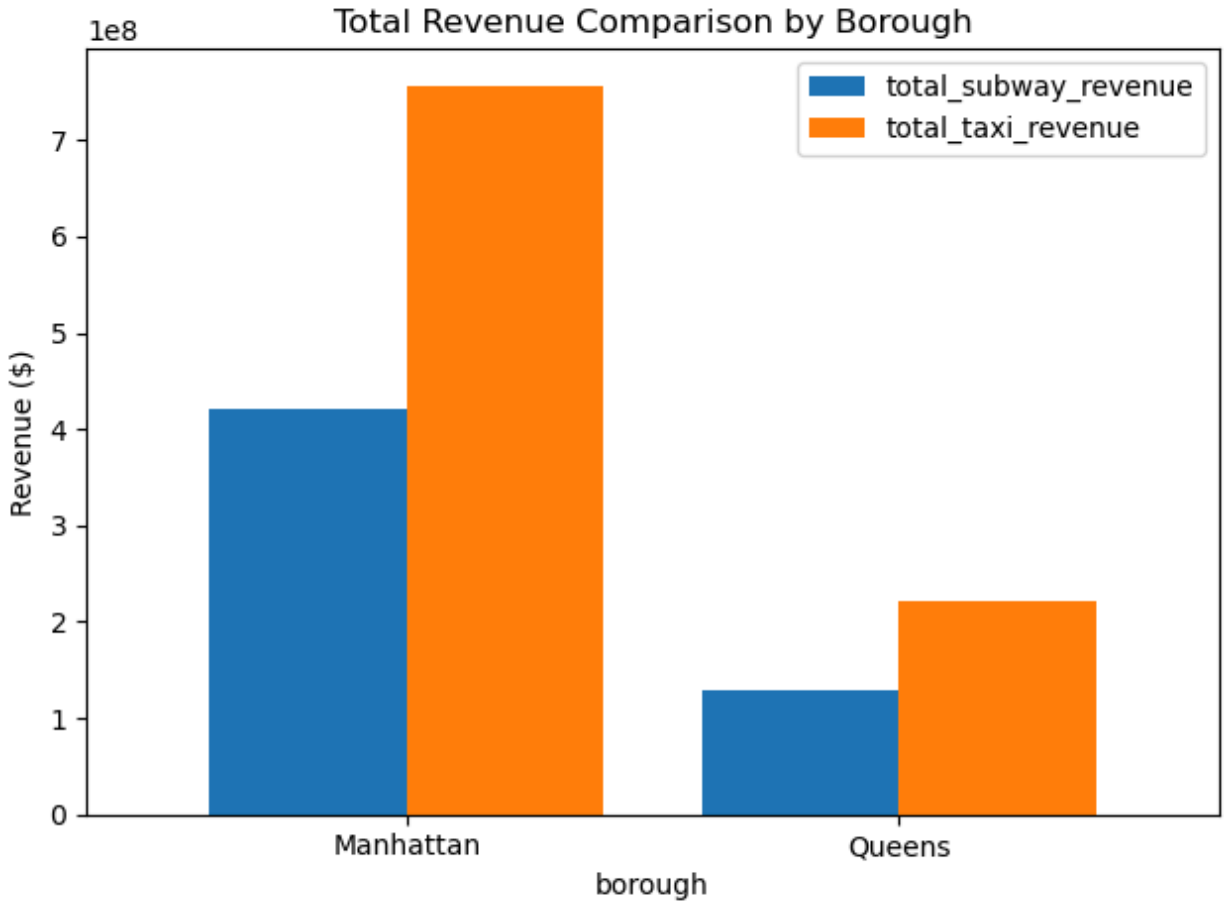
Insight Number 2:

Based on the aggregated data by transit hour and brough, we observe that in Manhattan, during a significant portion of the day, the total taxi revenue consistently exceeds the subway revenue. This includes both the late-night hours (such as from midnight to early morning) as well as the busier daytime hours. The trend is particularly pronounced in the early morning hours (from 00:00 to 06:00), where taxi fares significantly outperform subway revenue. This could indicate that, during these hours, more people are relying on taxis due to factors like reduced subway service frequency or convenience, especially in the less crowded, off-peak hours. The higher taxi revenue relative to subway revenue might also reflect a shift in commuter behavior, driven by factors such as comfort, convenience, and safety concerns, especially in Manhattan, where people may prefer the perceived security and privacy of taxis over the subway during late-night hours.

In comparison, while Queens also exhibits some instances of higher taxi revenue than subway revenue, the trend is not as consistent as in Manhattan. However, the higher taxi revenue in Queens could be attributed to the presence of LaGuardia Airport (LGA), JFK, which likely contributes to a significant amount of taxi traffic, particularly in the early morning and daytime hours when airport transfers are common. The proximity to a major airport may drive up the demand for taxi services, contributing to the overall higher revenue in the borough.




Percentage Increase Over Transit Hours (Manhattan & Queens)
Manhattan
Queens
Percentage Increase (%)
0
200
400
600
800
00:00
01:00
02:00
03:00
04:00
05:00
06:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Transit Hour


Total Revenue Comparison by Borough
1e8
total_subway_revenue
total_taxi_revenue
Revenue ($)
0
1
2
3
4
5
6
7
Manhattan
Queens
borough

Insight Number 3:

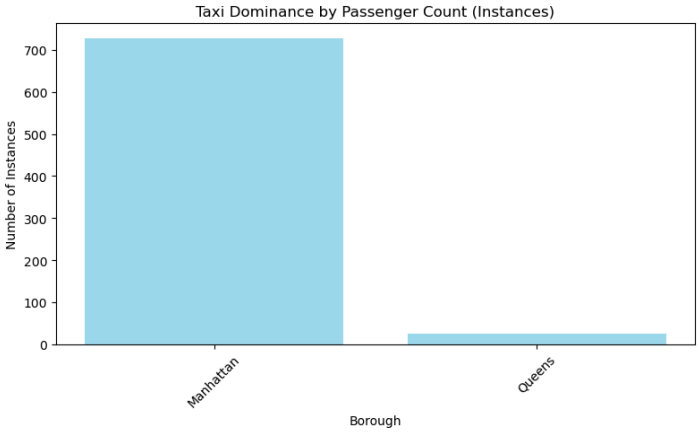
1. **Passenger Count higher in taxis vs subways (By borough):**
   The Query for this focused on instances where taxi passengers exceeded subway riders. It showed a significant preference for taxis in Manhattan (727 instances) and a much smaller but noticeable preference in Queens (26 instances). This reflects that, for specific time periods or conditions, people in these boroughs are more likely to take taxis than the subway based on passenger volume.
2. **Revenue collected higher in taxis vs subway (By borough):**
   This query highlights instances where taxi revenue surpassed subway revenue. The results indicate a much broader dominance of taxi revenue, with **Manhattan (8020 instances)** and **Queens (5448 instances)** leading by a large margin. Even Staten Island (84 instances) and Brooklyn (1 instance) appear, showing that taxis often generate more revenue than the subway across most boroughs.

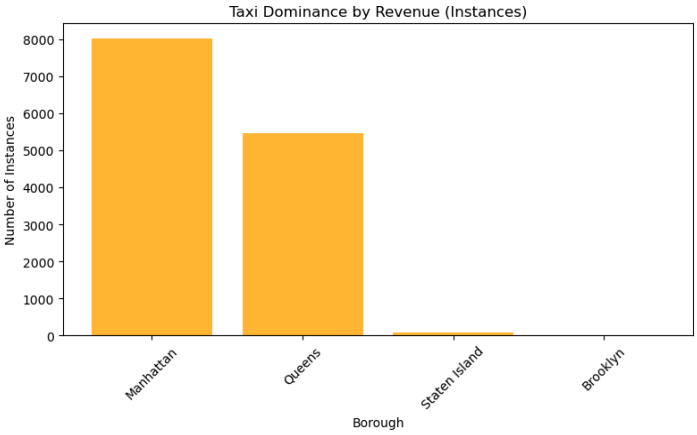
Combined Insight:

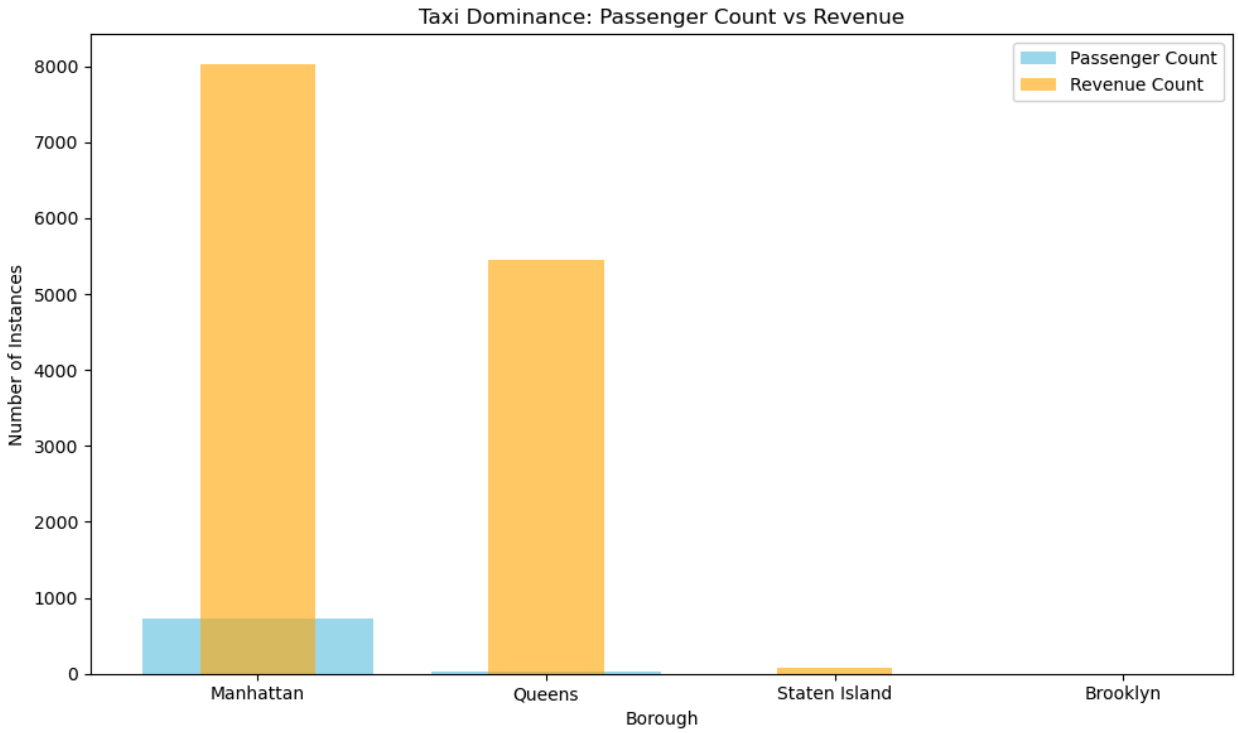
- The **discrepancy between revenue and passenger count** suggests that taxi rides are generally more expensive than subway trips, which is expected given the fare structure.
- In Manhattan, both the passenger count and revenue dominance of taxis emphasize the borough's reliance on taxis, potentially due to convenience, safety, or a culture of preferring taxis for short or late-night trips.
- In Queens, although passenger count instances were lower (26), taxi revenue dominance (5448 instances) highlights higher fares in Queens, possibly due to longer trip distances, like to/from airports (e.g., LaGuardia or JFK).
- For Staten Island and Brooklyn, the dominance in revenue but minimal passenger count suggests limited subway coverage in Staten Island and rare instances in Brooklyn where high taxi fares outweigh subway usage.

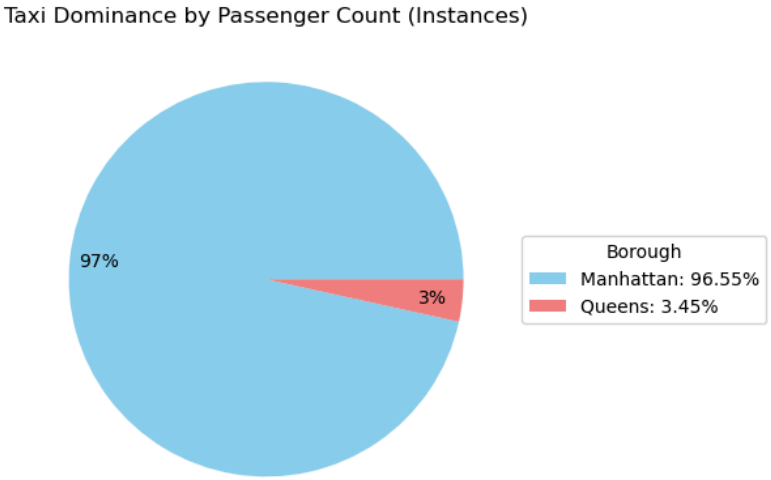
Conclusion:

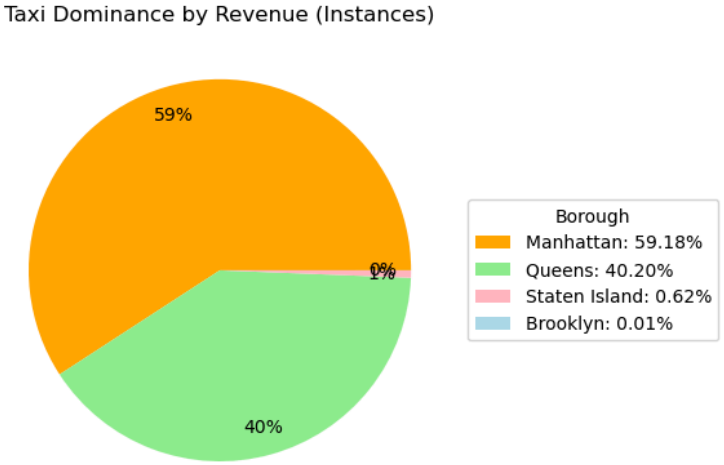
While Manhattan shows a strong preference for taxis in both passenger volume and revenue, Queens sees a stark dominance in revenue, likely influenced by longer or costlier trips. This indicates differing transit dynamics in each borough, shaped by geography, infrastructure, and trip patterns.





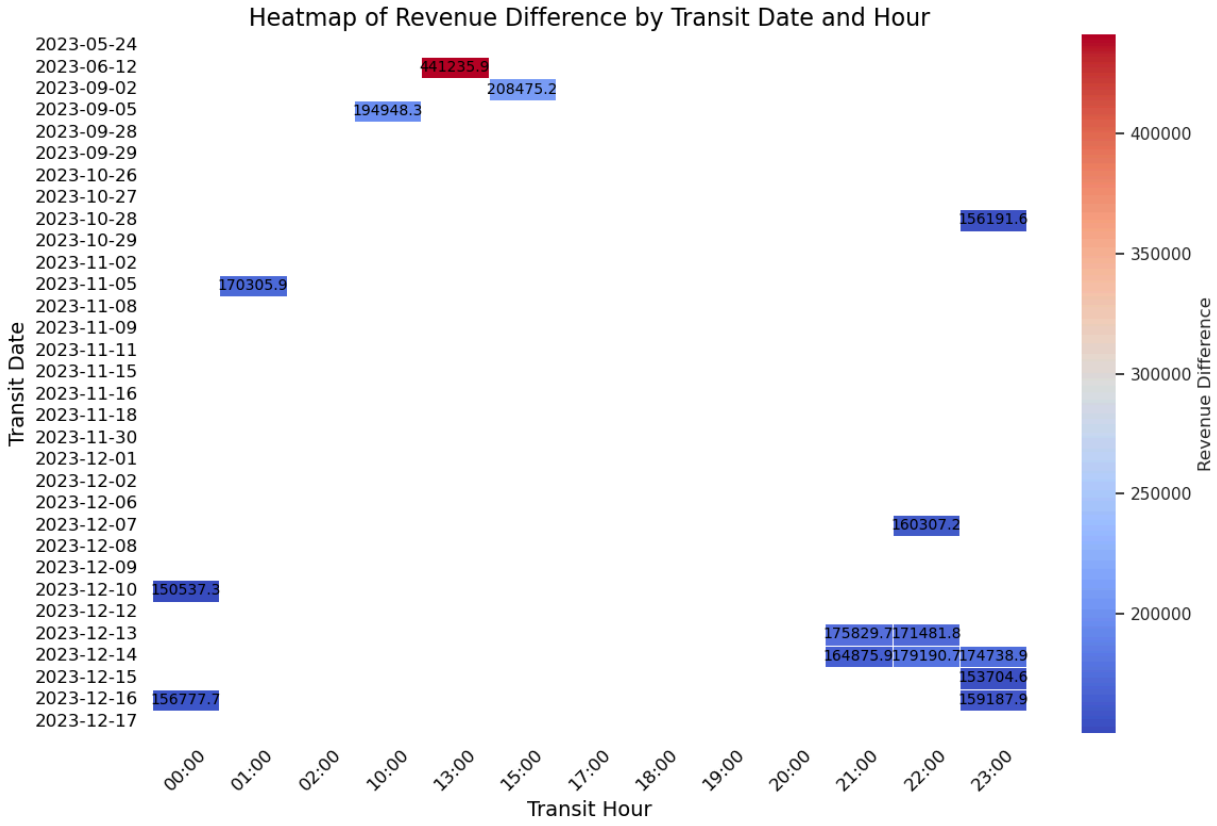

Taxi Dominance: Passenger Count vs Revenue
Passenger Count
Revenue Count
Number of Instances
0
1000
2000
3000
4000
5000
6000
7000
8000
Manhattan
Queens
Staten Island
Brooklyn
Borough

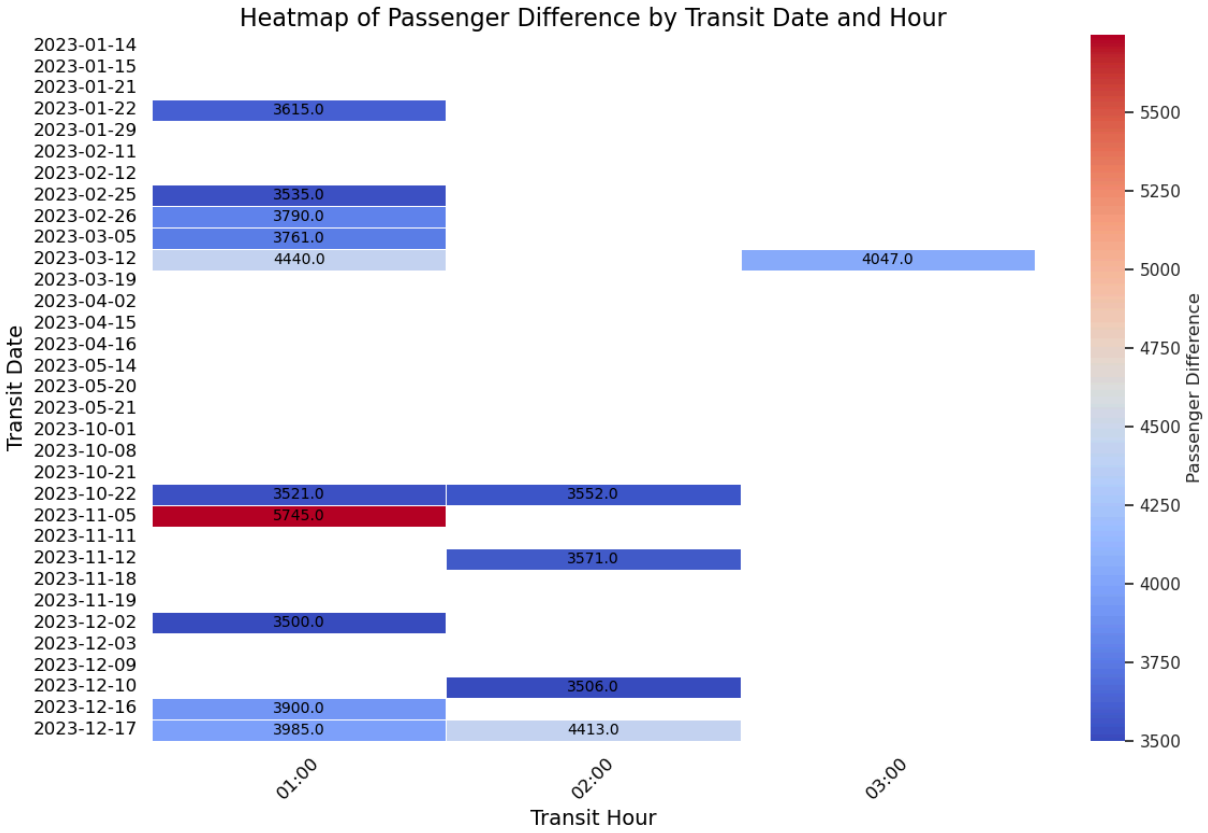

Taxi Dominance by Passenger Count (Instances)
97%
3%
Borough
Manhattan: 96.55%
Queens: 3.45%


Taxi Dominance by Revenue (Instances)
59%
40%
Borough
Manhattan: 59.18%
Queens: 40.20%
Staten Island: 0.62%
Brooklyn: 0.01%

Insight number 4:

Just to dig a little deeper for the above query, we now try to aggregate by borough and transit_hour to get finer grained analysis. We took the **top 50** records for both passenger volume and revenue grouped by borough and transit_hour and we found that Manhattan consistently experiences a higher taxi passenger count compared to subway riders, particularly between 1 AM and 3 AM. Notably, this is the case across multiple dates and hours, with significant differences in passenger volume (e.g., taxi passengers often outnumber subway riders by a large margin, as seen in records like those from 2023-11-05 01:00).





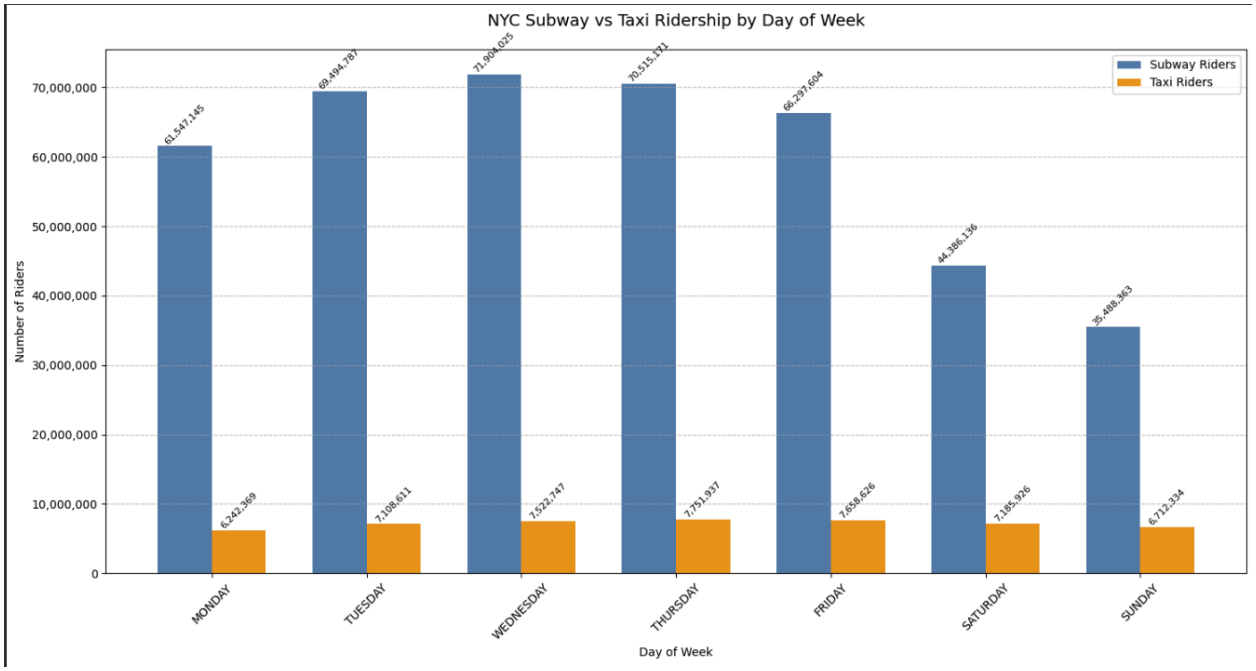
Key observations:

1. **Passenger Volume**: The taxi passenger counts far exceed subway ridership in the majority of the top 50 records. For instance, at 1:00 AM on 2023-11-05, there were 11,584 taxi passengers compared to just 5,839 subway riders, showing a difference of 5,745 passengers.
2. **Revenue Difference**: The records with the largest revenue differences (taxi fare exceeding subway revenue) generally correlate with higher taxi passenger numbers, suggesting that higher ridership in taxis leads to a substantial increase in revenue.
3. **Time of Day**: The period between 1 AM and 3 AM is a key window where taxi ridership exceeds subway ridership. This is consistent across multiple dates, indicating that during late-night hours, taxis might be a preferred choice due to lower subway frequency or availability.
4. **Borough**: The results exclusively feature Manhattan, highlighting that the data points in the top 50 are only from this borough. This may suggest that Manhattan is the main area where the difference between taxi passenger counts and subway riders is most pronounced. Other boroughs may not have shown as large a disparity in this time period, or the data for those areas might not have met the filtering criteria (e.g., having a higher taxi passenger count than subway ridership).

5. **Revenue and Ridership Correlation**: The correlation between the taxi passenger counts and revenue difference suggests that increased ridership in taxis contributes to a significant increase in taxi fare revenue, particularly when subway ridership is lower.
6. **Peak Hour Analysis**: Subway ridership shows a clear morning peak at 8:00 AM (39.5M riders) while taxi volume peaks at 6:00 PM (3.6M riders). This shift suggests that while morning commuters predominantly choose subways due to reliability and cost, evening travelers display more varied preferences, possibly influenced by factors like post-work activities and schedule flexibility. The gradual increase in taxi-to-subway ratio throughout the day, particularly during evening hours, indicates a clear pattern in how New Yorkers' transportation choices evolve from primarily subway-dependent mornings to a more balanced modal split by evening.

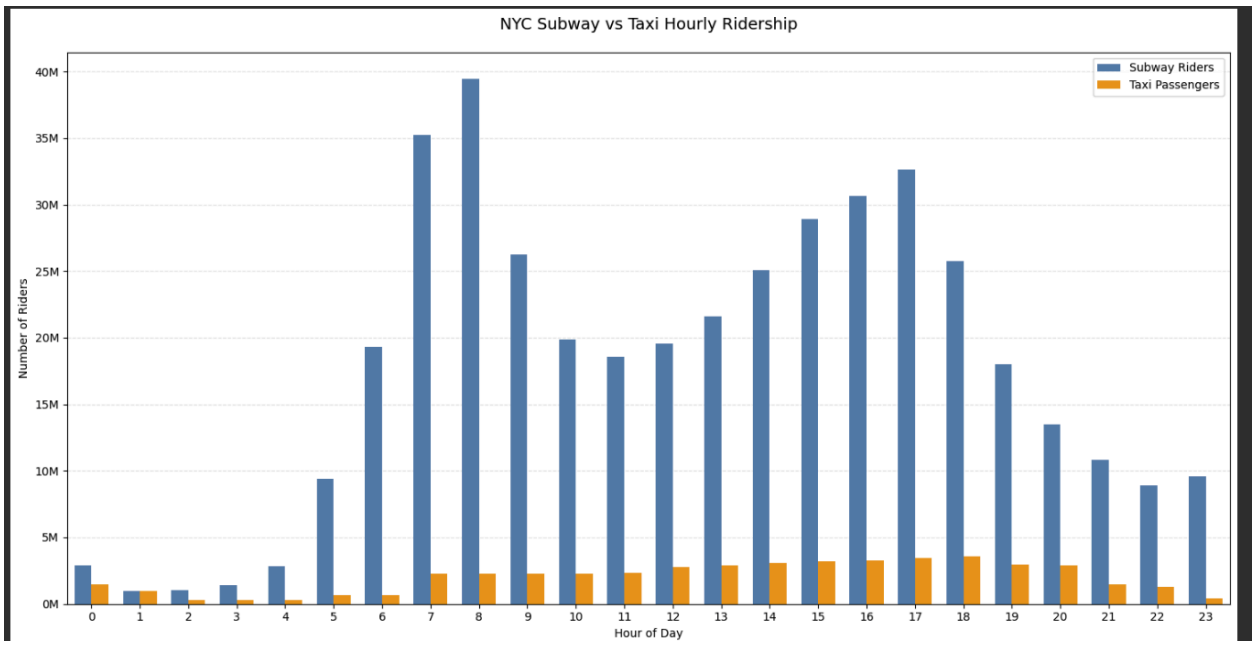
Manhattan sees a substantial increase in taxi passengers over subway riders, especially in late-night hours, which in turn results in a higher revenue difference. This could be indicative of late-night commuters or passengers who prefer taxis due to factors like convenience, safety, or availability, reinforcing the importance of analyzing transit hour-specific ridership trends for revenue and operational planning. Hence, the analysis done in this report confirms that taxi services could be particularly important during the late-night hours, possibly informing operational strategies for transit authorities and service providers.

Insight 5:



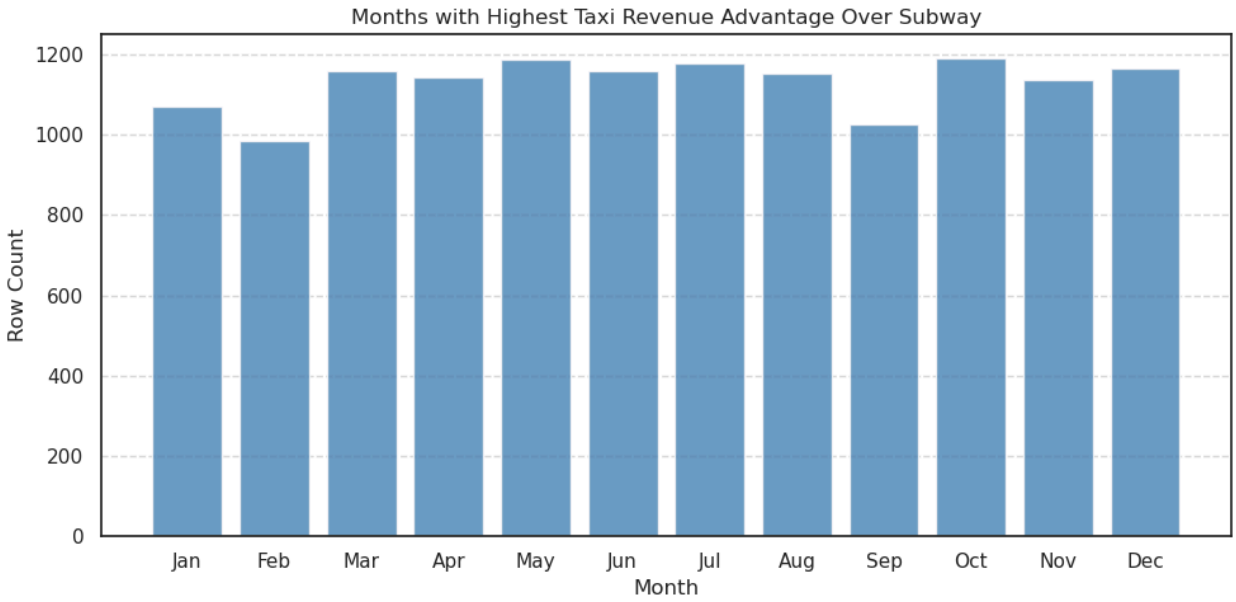
In terms of daily patterns, subway ridership demonstrates a clear peak during mid-week, with Wednesday showing the highest volume at 71.9 million riders, followed closely by Thursday at 70.5 million riders. This contrasts notably with taxi ridership, which maintains a more consistent pattern throughout the weekdays, peaking at 7.75 million riders on Thursday. The disparity between weekend and weekday ridership is particularly pronounced for subway service, with Sunday ridership dropping to 35.5 million (approximately 50% of peak), while taxi service shows more resilience, maintaining about 6.7 million riders (roughly 85% of peak levels). This trend suggests that while subways predominantly serve the commuter population, taxis provide more consistent service across all days, becoming particularly important during weekends when subway service is reduced.

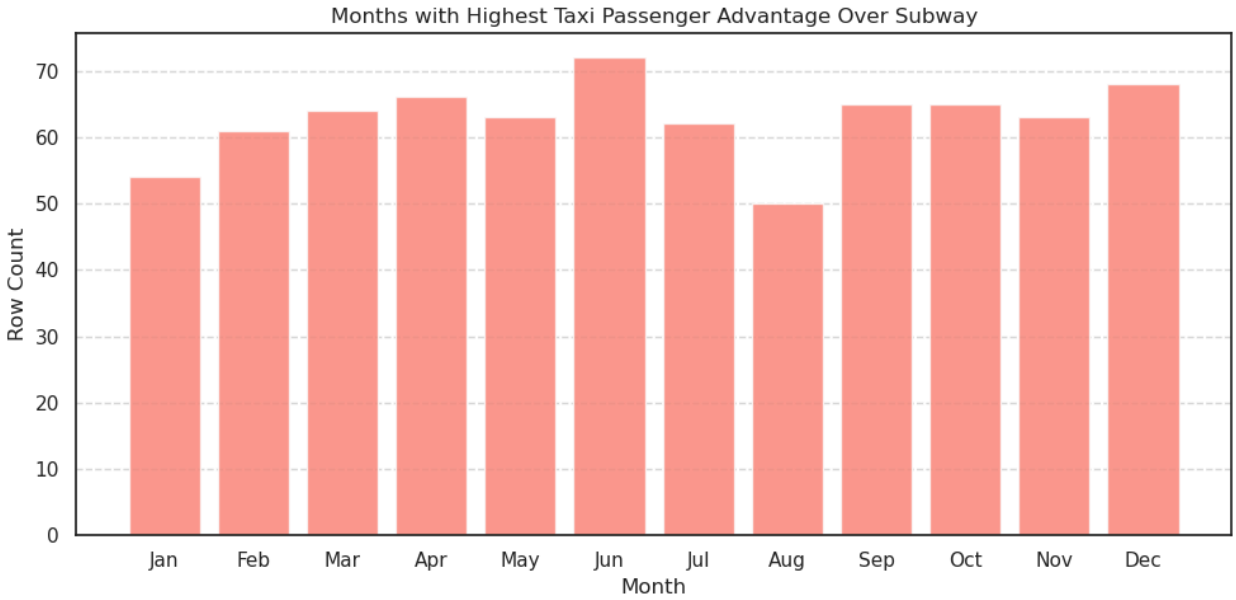
Insight Number 6:



When examining hourly patterns, the data reveals a compelling complementary relationship between the two modes of transport. Subway ridership shows a pronounced morning peak at 8:00 AM with 39.5 million riders, following a pattern that clearly aligns with traditional commuting hours. In contrast, taxi services reach their highest volume at 6:00 PM with 3.6 million riders, slightly offset from the subway's evening peak. This temporal difference in peak hours suggests that while morning commuters heavily favor subway transportation, evening travelers show more diverse transportation preferences. The most striking pattern emerges during late-night hours (1 AM to 3 AM), where taxi services maintain a relatively stronger presence compared to subway service. During these hours, while subway ridership drops to less than 1 million riders, taxi services continue to handle between 280,000 to 992,000 riders, indicating their crucial role in maintaining urban mobility during off-peak hours.

Insight Number 7:





Looking at month level analysis of passenger volume and revenue, the data confirms several key trends about taxi and subway usage patterns:

1. **Revenue Patterns**: Months with the highest counts for taxi revenue surpassing subway revenue include October, May, July, and December. These are likely driven by increased demand due to seasonal events, tourism, and holidays.
2. **Passenger Patterns**: Months where taxi passenger counts surpass subway riders include June, December, and April, with slightly less dominance compared to revenue

trends. This indicates that while revenue is consistently higher in certain months, passenger volume trends show a smaller margin, possibly due to fare structures or ride-sharing practices.

3. **Alignment with Tourist Influx**: Both data points confirm that tourism-driven demand during holiday seasons and summer months plays a significant role.

These patterns align well with the hypothesis that tourist activity significantly influences transit dynamics in New York City.

Conclusion

From the **revenue and passenger trends** , several observations can be made based on the heatmap patterns of revenue difference by date and hour:

1. **Peak Hours in Revenue**:
    a. High revenue differences tend to occur in the late evening and early night hours, particularly **between 9 PM and 12 AM**. This aligns with busy travel periods, such as tourists returning to hotels, late-night dining, or entertainment activities.
2. **Seasonal Impact**:
    a. Revenue surges are noticeable during specific **seasonal windows**, such as **holidays or tourist-heavy months**:
        i. **End-of-year months (November and December)**: This correlates with holidays like Thanksgiving, Christmas, and New Year's Eve, attracting more tourists and boosting transit services.
        ii. **Mid-year spikes (June)**: Potentially linked to summer vacation travel and people coming in for graduations etc.
3. **Further observations**:
    a. The passenger volume for taxis is consistently higher during off-peak hours from 1am to 3am due to different factors such as safety, convenience and accessibility.
    b. While the **passenger volume** data highlights that **beginning and end months of the year** see a significant increase in taxi ridership due to tourist influx, the **revenue patterns** during these hours remain relatively subdued. This may indicate:
        i. **Shorter trips** at night resulting in lower fares despite higher ridership.
        ii. **Promotional fares or discounts** offered during these hours to attract passengers.
4. **Tourism-Driven Revenue**:
    a. Influx of tourists, as evident from the passenger data trends, boosts ridership. However, the **per-trip revenue might be smaller** because:
        i. Tourists may prefer shared rides or shorter intra-city trips.
        ii. Discounts and capped fares for specific transit routes or zones could lower average revenue.
5. **Revenue Plateaus**:

a. Some dates with significant tourist activity (e.g., December) show **revenue plateaus** across multiple hours. This could indicate consistently high demand spread throughout the evening.

Combined Insight:

While the **1-3 AM time frame shows high passenger volumes**—primarily due to tourist activity in the **beginning and end months of the year**—the **revenue heat map suggests that the most profitable hours are earlier in the evening (9 PM-12 AM)**. This implies a disconnect between passenger volume and revenue, possibly influenced by:

- **Fare structure** differences during late-night hours.
- **Behavioral patterns** (e.g., tourists taking shorter trips at night or opting for cheaper transit options).

To optimize revenue during the 1-3 AM period, transit services could:

- Adjust fare pricing for late-night hours.
- Introduce premium or surge pricing for popular tourist destinations during peak months.
- Offer bundled ride packages for groups or tourists.

References

1. NYC Taxi & Limousine Commission Trip Record Data

[https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page]

2. MTA Subway Ridership Data

[https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-Beginning-July-2020/wujg-7c2s/about_data]

3. NYU Dataproc Cluster [https://dataproc.hpc.nyu.edu/]