Impact of Ride-Sharing on NYC Taxi Industry: A Data-Driven Analysis (2014-2017)

STAT GR 5243 Applied Data Science Project 1

February 19, 2025

Team: Yongyi Wang (yw4222), Julieta Caroppo (jc6158), Keito Taketomi (kt2849), Mei Yue (my2903), Ruoshi Zhang (rz2699)

Github Repository: https://github.com/julieta87/STATGR-5243-Project1

Introduction

Over the past decade, the rapid expansion of rideshare services like Uber and Lyft has reshaped urban transportation worldwide. Nowhere is this shift more apparent than in New York City, where traditional yellow taxis—once the dominant for-hire option—have seen ridership patterns evolve dramatically. This project investigates the period from 2014 to 2017, a time when disruptive newcomers captured growing market share and heightened competitive pressures for taxi operators. By cleaning, merging, and analyzing comprehensive trip data from the New York City Taxi and Limousine Commission (TLC), we aim to uncover how these changes affected daily ridership trends, fare structures, and passenger behaviors. Through feature engineering and exploratory data analysis, our goal is to provide a data-driven perspective on the yellow taxi industry's adaptation—or lack thereof—to an increasingly competitive marketplace.

Background

New York City's yellow taxis have long been an essential part of the city's transportation network, providing reliable service to millions of residents and visitors. Unlike other for-hire vehicles, taxis have the exclusive right to pick up passengers through street hails anywhere in the city, making them a convenient option, particularly in high-demand areas such as Manhattan and major transit hubs like JFK and LaGuardia airports. The industry operates under a strict medallion system, established in 1937 to regulate the number of taxis on the road. Each yellow taxi must have a medallion affixed to it, and the number of medallions is legally

capped at 13,587. These medallions are auctioned by the city and can also be transferred on the open market by licensed brokers. For many years, taxi medallions were considered highly valuable assets, reaching peak prices of over one million dollars.

For decades, yellow taxis were the primary mode of for-hire transportation in New York City, serving as a lifeline for commuters, tourists, and residents alike. Their widespread presence and regulation ensured a standardized fare system, providing passengers with predictable pricing and service. Today, they remain an integral part of the city's transportation landscape, offering on-demand rides that continue to serve millions of passengers each year.

About the Data set

The dataset used for this analysis comes from the New York City Taxi and Limousine Commission (TLC) Trip Records, a publicly available database that provides detailed trip-level data for yellow taxis from 2009 to 2024. The TLC, established in 1971, is responsible for regulating and licensing New York City's medallion taxis, green taxis, for-hire vehicles (FHVs), commuter vans, and paratransit vehicles. The trip data is collected through Technology Service Providers (TSPs), which are third-party vendors responsible for managing electronic metering, credit card transactions, and GPS tracking in taxis. These TSPs automatically transmit ride details to the TLC, creating a comprehensive digital record of each taxi trip.

Each trip record in the dataset represents a single completed ride by a TLC-licensed vehicle. The data includes a timestamped log of when and where passengers were picked up and dropped off, measured using GPS coordinates or, in later years, taxi zone IDs. The trip record also logs trip distance in miles, fare amounts, surcharges, tolls, payment types, and driver-reported passenger counts. Since 2013, data collection has expanded to include green taxis, which serve areas outside of central Manhattan, and in 2015, the dataset was made publicly available online through the NYC Open Data Portal.

To ensure transparency and accuracy, TLC trip records are continuously updated and released every six months. The data is initially stored in Parquet format, a highly efficient columnar storage format used for handling large-scale data. This format allows for fast queries and compact storage, making it ideal for datasets containing millions of trip records per month.

The granularity of the dataset allows for in-depth exploration of taxi ridership trends, fare revenue, geographic travel patterns, and seasonal fluctuations in NYC's transportation system. However, while the TLC collects and releases this data, it does not directly verify every trip record. The accuracy of certain fields, such as passenger count or payment type, depends on driver input, meaning occasional inconsistencies may exist.

The structure of the data covers every month individually for each year. Each month roughly contains over 2 million rows with each row as a single taxi ride. Below gives a description of the columns and what is covered in the data.

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

Objective

The objective of this study is to conduct a comprehensive analysis of NYC yellow taxi ridership trends from 2014 to 2017, a pivotal period that saw a dramatic shift in urban transportation dynamics. By leveraging large-scale trip data from the New York City Taxi and Limousine Commission (TLC), we aim to uncover the underlying factors that contributed to the decline of yellow taxi usage and the rise of alternative transportation options. Our research will focus

on identifying patterns in taxi trip and exploratory data analysis on volume, revenue fluctuations, fare adjustments, and borough-specific variations to assess how external disruptions reshaped the industry. Lastly, we aim to observe when and how NYC taxis lost their market dominance.

Data Cleaning & Preprocessing

We started with merging all months of each year. We first converted the date columns `tpep_pickup_datetime` and `tpep_dropoff_`datetime to datetime format and dropped rows with incorrect datetime, specifically those with invalid year. We also filtered out records where the `passenger_count` and `trip_distance` were both zero, which could be due to taxi drivers' manual input mistakes. Since `congestion_surcharge` and `airport_fee` became effective after the dataset's timeframe, these two columns contained only NA values so we decided to remove them. Upon checking duplicates, we found none. Finally, we handled outliers by removing extreme outliers such as negative values from all charge related numeric variables and unrealistically long trip distances from `trip_distance`. Below is some code to highlight this through a function created to apply to our merged years:

```
def preprocess_data(df):
    columns_to_keep = ["tpep_pickup_datetime",
    "tpep_dropoff_datetime", "passenger_count",
                       "trip_distance",
                       "payment_type",
                       "fare_amount",
                       "tip_amount",
                       "total_amount",
                       "PULocationID",
                       "DOLocationID"]

    df = df[columns_to_keep].copy()

    df["tpep_pickup_datetime"] = pd.to_datetime(
      df["tpep_pickup_datetime"])
    df["tpep_dropoff_datetime"] = pd.to_datetime(
      df["tpep_dropoff_datetime"])

    df = df[(df["fare_amount"] > 0) & (
      df["trip_distance"] > 0)]

    df["trip_duration"] = (
      df["tpep_dropoff_datetime"] -
      df["tpep_pickup_datetime"]
```

```
    ).dt.total_seconds() / 60

df = df[(df["trip_duration"]
> 1) &
(df["trip_duration"] < 300)]

return df
```

Since we are working with a huge amount of data, we condensed the yearly data through grouping by pickup date and computing the average values and summation of all numeric columns of each day. Some new columns were added to each year's dataset before the feature engineering due to efficiency. There were originally two columns relating to the pickup and dropoff location ID. Referring to a taxi zone lookup table from TLC, we changed the location ID to the name of NYC's boroughs and counted the number of boroughs that appeared in daily trips. The purpose of other created columns will be discussed in the feature engineering section. Ultimately, we combined all four-year data and the final dataset we worked on for analysis has roughly 365*4 rows and 23 columns presented below.

- `pickup_date`: the date of the trips in Y-M-D format.
- `day of week`: day of week corresponding to the date.
- `avg fares`: averaged time-and-distance fare calculated by the meter.
- `total fares`: total fares of all trips among a single day.
- `tip percentage`: overall tip percentage of a day.
- `duration`: the average length of time in minutes of all day's trips.
- `frequency_of_rides`: the number of trip for each day
- `avg trip distance`: the average of the elapsed trip distance in miles reported by the taximeter of a day.
- `tip amount avg`: the average tip amount from credit card. Cash tips are not included.
- `avg passengers`: the average number of passengers of all day's trips.
- `sum of passengers`: total number of passengers of a day.
- `avg fare per mile`: the average amount a passenger pays per mile traveled.
- The remaining 14 columns are the number of pickups and drop-offs per borough on a day.

Feature Engineering

Before we explore our data, we added some columns. To further study the effect of the rising popularity of external ride-share services on yellow taxis, a new column named `trip duration` is created to the effect on trip length, and speed. Additionally, we also added columns of `fare per mile` and `tip percentage` to assess the change of pricing trending. In order to inspect the change of ride demand pattern, we extracted the day of week from the datetime column.

Feature engineering allowed us to conduct data analysis (EDA) on the features in order to comprehend the evolution of ride patterns and the potential effects of external ride-share competition on yellow cab operations. In particular, we obtained a better understanding of passenger behavior, possible changes in geographic demand, and pricing patterns over time by looking at trip duration, fare per mile, and tip percentage—as well as the borough-specific columns obtained from the TLC taxi zone lookup table. Below is some code on how this process was done on 2014 merged data, which was then applied to the other merged years.

```
pickup_df = df_2014_cleaned.merge(
  taxi_data[['LocationID',
                               'Borough']],
                              left_on=
                              "PULocationID",
                              right_on=
                              "LocationID",
                              how="left")
pickup_df.rename(columns={"Borough":
  "Pickup_Borough"},
inplace=True)
pickup_df.drop(columns=["LocationID"], inplace=True)

dropoff_df = df_2014_cleaned.merge(
  taxi_data[['LocationID',
                               'Borough']],
                               left_on="DOLocationID",
                               right_on="LocationID",
                               how="left")
dropoff_df.rename(columns={"Borough":
  "Dropoff_Borough"},
inplace=True)
dropoff_df.drop(columns=["LocationID"], inplace=True)

pickup_tallies = pickup_df.groupby(["pickup_date",
"Pickup_Borough"]).size().reset_index(name="Count")
pickup_pivot = pickup_tallies.pivot(
```

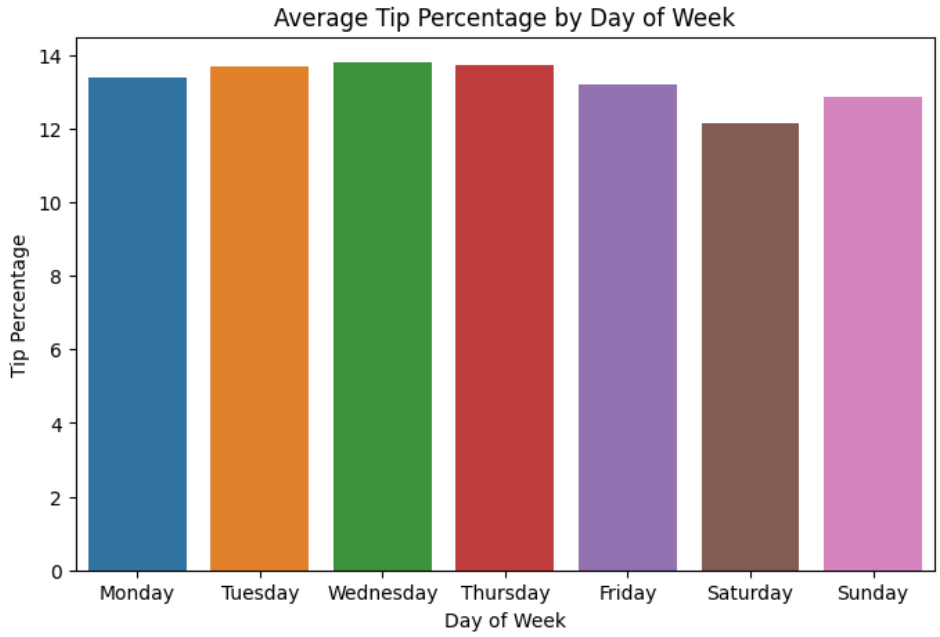
```
  index="pickup_date",
columns="Pickup_Borough", values="Count").fillna(0)
pickup_pivot.columns = [f"{col}_Pickups" for
col in pickup_pivot.columns]

dropoff_tallies = dropoff_df.groupby(["pickup_date",
"Dropoff_Borough"]).size().reset_index(name=
"Count")
dropoff_pivot = dropoff_tallies.pivot(
  index="pickup_date",
columns="Dropoff_Borough",
values="Count").fillna(0)
dropoff_pivot.columns = [f"{col}_Dropoffs"
for col in dropoff_pivot.columns]

final_borough_tallies_2014 = pickup_pivot.join(
  dropoff_pivot, how="outer").reset_index()
```

Another column we decided to create was `fare per mile` (fare_amount / trip_distance). This provided a window into evolving pricing tactics. A steady drop in fare per mile could indicate increased competition, with yellow taxis reacting to ride-sharing services by cutting prices or running specials. On the other hand, an upward trend can indicate that riders are taking longer, fewer rides, or that prices and surcharges have gone higher in relation to the extent of the journey. Below is some code applied on 2015 merged dataset to highlight this

```
{python}
daily_counts_2015 =
df_2015_cleaned.groupby(
  'pickup_date').size().rename(
  "frequency_of_rides")

df_2015_cleaned['fare_per_mile'] = df_2015_cleaned[
  'fare_amount'] / df_2015_cleaned[
    'trip_distance']

daily_agg_2015 = df_2015_cleaned.groupby(
  'pickup_date').agg({
    'fare_amount': ['mean', 'sum'],
    'trip_duration': 'mean',
    'trip_distance': 'mean',
    'tip_amount': ['mean', 'sum'],
    'passenger_count': ['mean', 'sum'],
```

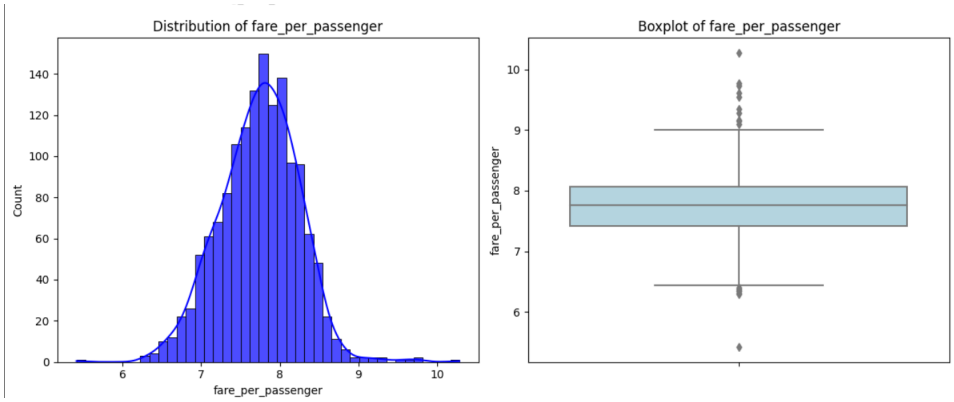
```
    'fare_per_mile': 'mean'
}).reset_index()
```

We also created a new feature for the tip percentage. The bar chart below illustrates the average tip percentage for each day of the week, providing a straightforward view of whether riders are more inclined to tip generously on certain days. The overall differences between days appear modest, though small variations do emerge—some weekdays may show slightly higher tipping, while weekends can fluctuate. Such patterns could reflect factors like a heavier mix of tourists, business travelers, or local residents on particular days, each group with its own tipping behavior. By comparing these daily averages, we can gauge if specific days consistently see higher or lower tip rates, potentially prompting targeted strategies such as weekend promotions or weekday service adjustments.



In understanding the overall cost dynamics of taxi rides, we created a feature `fare_per_passenger`, which provides a clear look at how expenses are distributed among riders. By examining how much each passenger typically pays, we can gauge affordability, pricing consistency, and potential outliers within the system. This visualization compares a histogram (on the left) and a boxplot (on the right), both of which reveal key insights into how `fare_per_passenger` is distributed across all trips in the dataset.

```
daily_summary_features['fare_per_passenger'] = (
    daily_summary_features['total fares'] /
    daily_summary_features['sum of passengers']
)

five_boroughs = ["Manhattan", "Brooklyn",
"Queens", "Bronx", "Staten Island"]

for borough in five_boroughs:
    col_name = borough.lower().replace(" "
    , "_")

    daily_summary_features[f"{col_name}_pickup_share"] = (
        daily_summary_features[
          f"{borough}_Pickups"] /
          daily_summary_features["frequency_of_rides"]
    )

    daily_summary_features[f"{col_name}_dropoff_share"] = (
        daily_summary_features[
          f"{borough}_Dropoffs"] /
          daily_summary_features["frequency_of_rides"]
    )
```



In the histogram, we see that most fares cluster around $7-$8 per passenger, forming a roughly bell-shaped distribution. The box plot confirms a fairly narrow interquartile range, with a median near $7.5, indicating that most rides fall within a relatively consistent fare range per passenger. However, a few outliers appear above $9 or below $6, suggesting there are occasional

rides that deviate significantly—possibly due to unusually long distances, extra fees, or other anomalies in the data.

With what we have seen through our feature engineering section, we will further explore our cleaned and edited dataset in the data analysis section which will give us insight of yellow taxi's trends compared to other modern transportation services.

Clustering Analysis

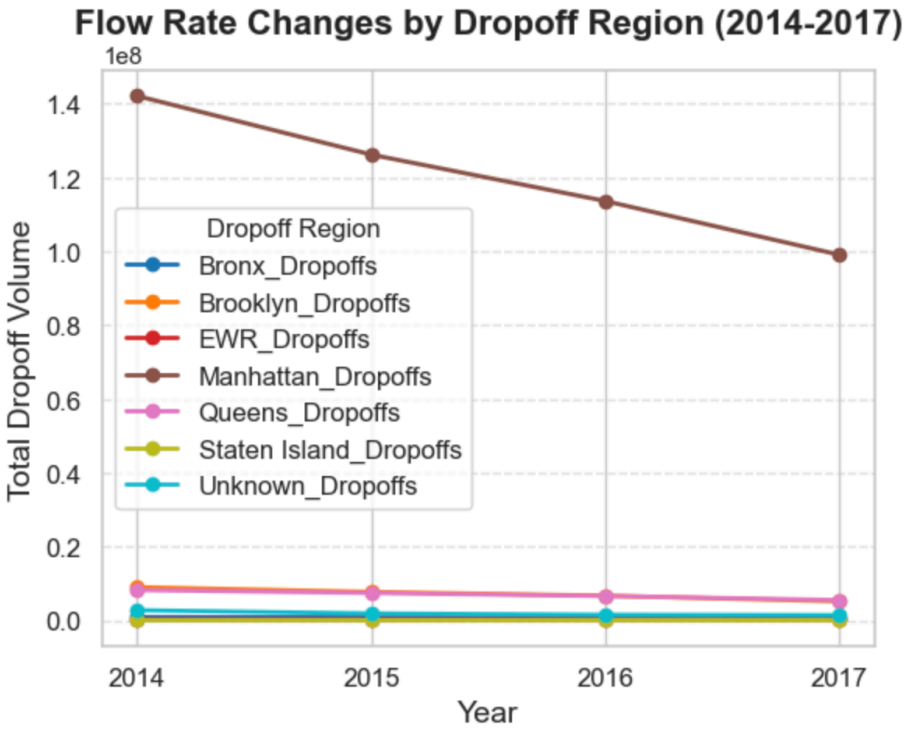
1. Normalize/standardize numerical variables

From previous column analysis, we can see that features in the dataset vary significantly in scale, with trip distance ranging from **0.5 to 30 miles**, total fares from **$5 to $100**, tip percentage from **0% to 25%**, and ride frequency from **10,000 to 500,000**. Without normalization, larger-scale features like **total fares and ride frequency** can dominate models such as K-Means, PCA, and Regression, leading to skewed results. Therefore, it is necessary to make normalization to ensures all features contribute equally to the analysis.

```
# Choose between Standardization
scaler = StandardScaler()  # Standardization

# Apply scaling to numerical columns
df[numerical_columns] = scaler.fit_transform(
  df[numerical_columns])
```

2. Encode categorical variables appropriately

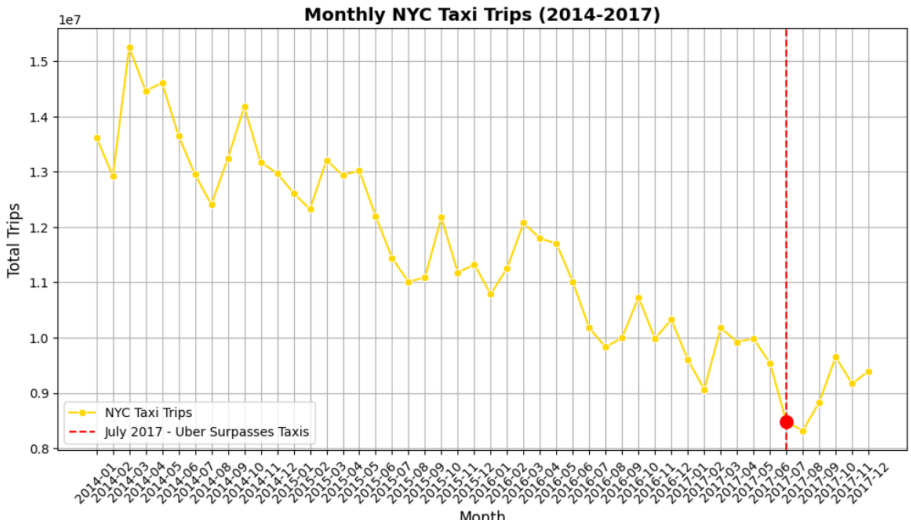
To analyze the **flow rate changes across different drop-off regions**, categorical variables were encoded to ensure numerical compatibility. **Label encoding** was applied to transform drop-off locations into numerical values, while **one-hot encoding** was used for better representation in machine learning models. This step ensures that drop-off regions contribute equally to the analysis without introducing biases due to categorical ordering.



The flow rate analysis from **2014 to 2017** shows a significant decline in **Manhattan dropoffs**, indicating a shift in transportation trends, likely due to the rise of **ride-hailing services**. Other boroughs, including **Brooklyn, Queens, and the Bronx**, experienced relatively smaller declines. This suggests that **Uber and Lyft's impact** was strongest in Manhattan, leading to a continuous reduction in yellow taxi demand over the years.

Exploratory Data Analysis

This section provides a broader market context for the decline in NYC yellow taxi usage between 2014 and 2017, situating the findings from our own data analysis against established research and external reports on rideshare growth, fare structures, and consumer preferences. By highlighting specific parallels—such as the surge in for-hire vehicle (FHV) trips, evolving rider demographics, and changing fare dynamics—we underscore the significance of the patterns observed in our aggregated TLC taxi dataset.



Multiple studies, including those by Cramer & Krueger (2016) and the TLC Factbook, confirm that Uber and Lyft experienced exponential adoption rates in New York City during this period. Uber's ride volume alone surged from approximately 5 million monthly rides in early 2015 to over 15 million by late 2017, representing a threefold increase that significantly eroded the customer base of traditional yellow taxis. This trend is clearly reflected in our time-series analysis (2014–2017), where we observe a steady decline in taxi ridership frequency. Notably, from mid-2015 onward, daily trip counts show a consistent downward trajectory, aligning with the period in which Uber's monthly ride volume first began to compete with—and ultimately surpass—that of yellow taxis.

1. Generate summary statistics and descriptive insights

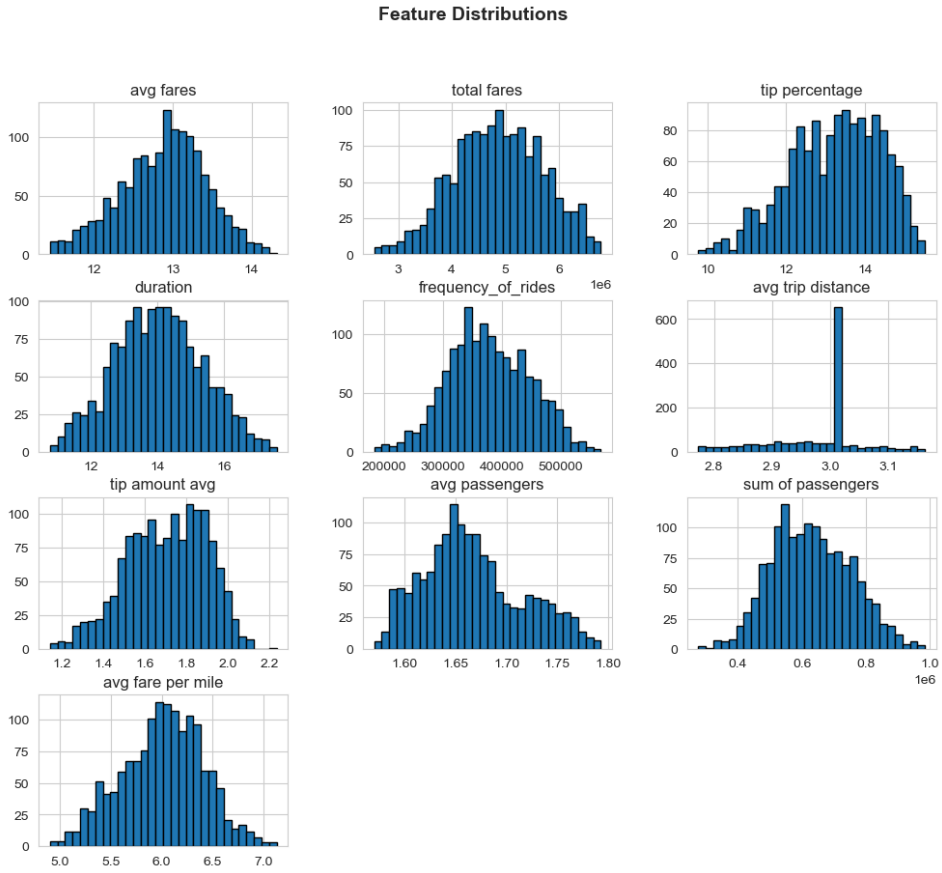
```
Summary Statistics:
                      count                 mean                  min  \
pickup_date            1461  2016-01-01 00:00:00  2014-01-01 00:00:00
avg fares           1461.0                 -0.0            -2.746114
total fares         1461.0                  0.0             -2.86891
tip percentage      1461.0                  0.0            -2.963174
duration            1461.0                 -0.0            -2.473948
frequency_of_rides  1461.0                 -0.0            -2.825285
avg trip distance   1461.0                 -0.0            -3.608571
tip amount avg      1461.0                  0.0            -2.914509
avg passengers      1461.0                 -0.0            -1.971279
sum of passengers   1461.0                 -0.0            -2.866031

                                    25%                   50%  \
pickup_date         2015-01-01 00:00:00  2016-01-01 00:00:00
avg fares                     -0.658916             0.085838
total fares                    -0.69915            -0.009316
tip percentage                -0.714839             0.094457
duration                      -0.703288            -0.018058
frequency_of_rides            -0.711757            -0.054211
avg trip distance             -0.557048             0.000516
tip amount avg                -0.721232             0.076306
avg passengers                -0.694629            -0.138536
sum of passengers             -0.730325            -0.072186

                                    75%                   max       std
pickup_date         2016-12-31 00:00:00  2017-12-31 00:00:00       NaN
avg fares                      0.680901             2.729463  1.000342
total fares                    0.728961             2.286372  1.000342
tip percentage                 0.795204             1.974068  1.000342
duration                       0.678184             2.768045  1.000342
frequency_of_rides             0.745182             2.771696  1.000342
avg trip distance              0.204498             5.120079  1.000342
tip amount avg                 0.787757             2.763698  1.000342
avg passengers                 0.615937             2.605639  1.000342
sum of passengers                0.7194             2.890189  1.000342
```

The dataset provides key insights into fare amounts, trip characteristics, and ride demand. The total fares exhibit a high mean with a substantial standard deviation, indicating variability in daily revenue. Avg fares are relatively stable, suggesting a consistent pricing structure. Tip percentage and tip amount avg indicate tipping behavior, with values centered around reasonable percentages.

In terms of trip characteristics, avg trip distance has low variance, implying typical trip lengths are consistent. Frequency of rides shows significant fluctuation, highlighting varying demand across days. Sum of passengers exhibits a notable spread, reflecting different passenger group sizes.

2. Visualize distributions, relationships, and trends

   To detect feature distribution, histograms are generated to first give us insights about the numerical features in dataset.

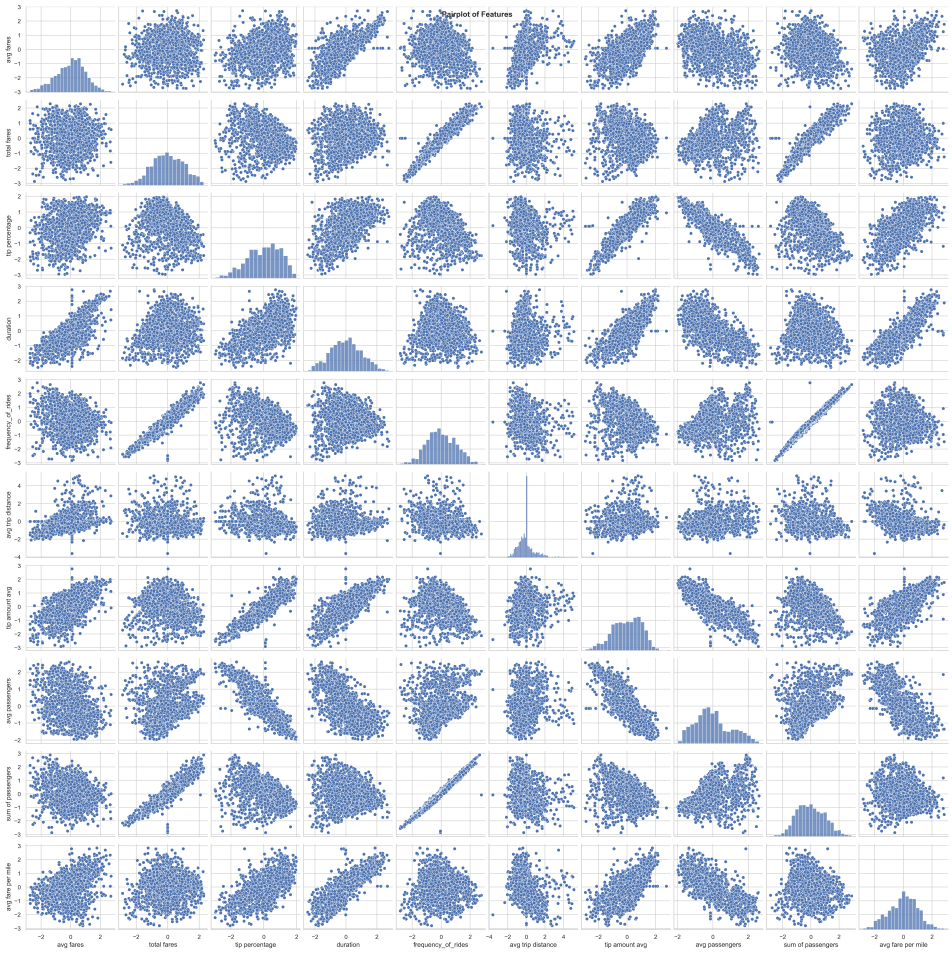


The histogram plots provide insights into the below perspectives:

- **Avg fares, total fares, and tip amount avg** exhibit approximately normal distributions, indicating a balanced spread of values.
- **Tip percentage and avg fare per mile** show slight skewness but remain within an acceptable range, suggesting most values cluster around the mean.
- **Duration and frequency of rides** display bell-shaped distributions, aligning with expected trends in ride duration and demand.
- **Avg trip distance** exhibits an unusual spike at a specific value, suggesting a potential issue in data recording or a common fixed distance for rides.
- **Sum of passengers and avg passengers** demonstrate a moderately normal distribution, reflecting expected ride-sharing trends.

3. Identify patterns, correlations, and anomalies

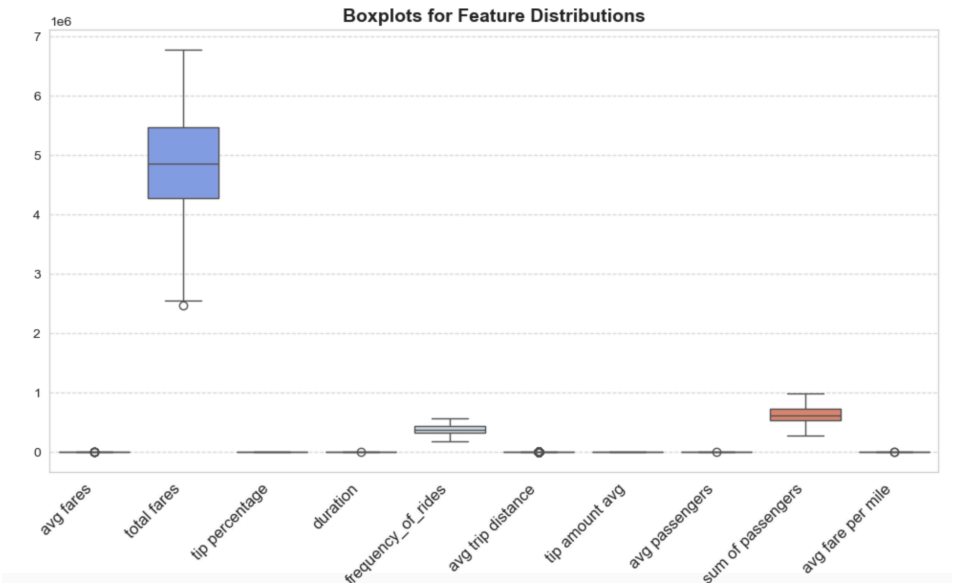
   To get generally find the correlations between numerical variables, we draw a Pairplots to show the relationships between variables.



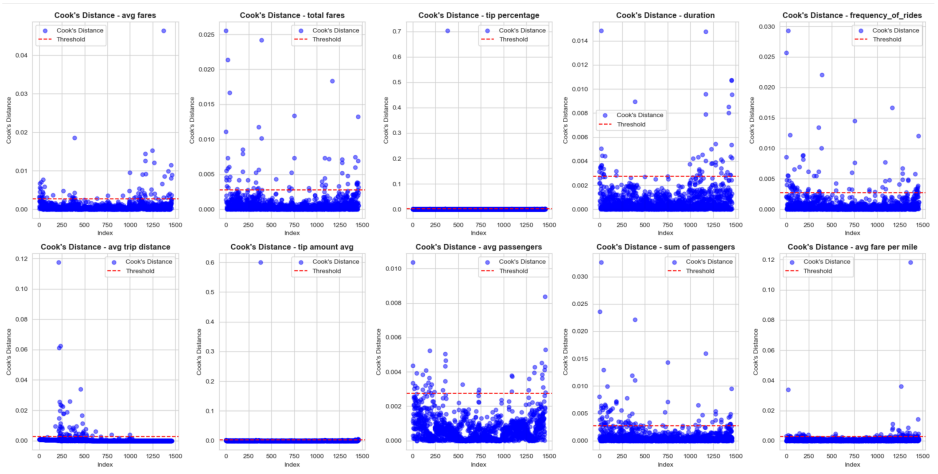
The pairplot visualization provides insights into the relationships between numerical features:

- **Strong Linear Relationships**: Certain features, such as **total fares** and **sum of passengers**, exhibit strong positive linear correlations, indicating that an increase in total fares is closely linked with more passengers.
- **Weak or No Correlations**: Variables like **tip percentage** and **duration** do not show strong associations with most other features, suggesting that tipping behavior may be independent of ride duration.
- **Potential Nonlinear Trends**: Some scatterplots indicate mild nonlinear patterns, hinting at potential transformations that may improve modeling.
- **Outlier Presence**: Some variables, such as **frequency of rides** and **avg trip distance**, contain scattered points that deviate from the main distribution, suggesting the presence of rare but influential values.

To detect whether there are potential outliers and leverages, we use boxplot and cook's distance to gain an overview of the numerical feature distributions in the dataset.



The boxplot analysis reveals outliers across multiple numerical variables. "Total fares" exhibits a significant number of extreme values beyond the upper whisker, indicating days with unusually high fare totals. Similarly, "sum of passengers" and "frequency of rides" display moderate outliers, suggesting fluctuations in ride demand on specific days. These anomalies could reflect special events, holidays, or data errors. Addressing these outliers is crucial to prevent model distortions.



The Cook's Distance plots highlight leverage points, indicating data points with high influence on regression models. While most points fall within normal thresholds, certain variables, including "total fares" and "frequency of rides," show prominent high-leverage values. These instances may correspond to unusual demand spikes, potentially affecting prediction accuracy. Further examination of these leverage points is recommended to determine whether they stem from genuine trends or data inconsistencies. As a consequence, we need to consider both missing value and leverage value.

The outlier detection process used multiple statistical methods, including **IQR (Interquartile Range), Z-score, and Tukey's Fences**, to identify extreme values. The results revealed that certain features, particularly **"total fares"**, **"sum of passengers"**, and **"frequency of rides"**, exhibited significant deviations, suggesting irregular trends in ride demand.

To mitigate the influence of extreme values, we applied an automated selection process to determine the most appropriate outlier replacement method. The algorithm assessed the proportion of outliers in each column and selected the **median** as the optimal replacement strategy. This choice is justified as follows:

- **Low outlier proportion (<5%)**: The median is robust to extreme values and preserves the distribution's integrity.
- **Skewed distributions**: The median ensures that extreme outliers do not disproportionately affect the dataset, unlike the mean.

The decision to use the median implies that while some extreme values were mitigated, the overall distribution remains reflective of real-world variability. This approach enhances the dataset's **stability for statistical modeling and predictive analysis** while retaining important trends in ride behavior.

Key Findings Summary

After cleaning and feature engineering, the results indicate that there is a significant decrease in ride frequency, total fare, and revenue gained after 2017. In other words, the popularity of other ride-share services causes the total volume of yellow taxi rides, total fare and revenue gained by taxi to be lower compared with the past three years. Regarding boroughs such as Queens, Brooklyn, and EWR airport, no obvious change is observed in these areas across time. However, while Manhattan remained the most dominant location among all boroughs, the total number of taxi drop-offs in Manhattan experienced a decline.

Challenges Faced

Throughout this project, we encountered several significant obstacles that shaped both the data processing and analysis phases. One primary challenge was the big volume and how to process the TLC trip data. With millions of rows per month, merging multiple files for each year demanded substantial processing time and memory, often pushing the limits of local computing

resources. This data intensity also made it impractical to host raw datasets on platforms like GitHub, complicating version control and sharing. Additionally, data quality posed a persistent issue, as the raw dataset contained missing or inconsistent entries—particularly in driver-reported fields like passenger count and trip distance. These inconsistencies, often stemming from manual input errors or anomalies like incorrect timestamps, required extensive cleaning and outlier removal to ensure reliable results.

Further complicating the process were the heterogeneous data formats across the study period. Changes in data collection over time, such as the shift from latitude/longitude coordinates to taxi zone IDs after 2016, necessitated careful preprocessing to standardize the dataset and maintain consistency across years. The reliance on driver-reported data introduced another layer of difficulty, as fields like passenger count were subject to human error or subjectivity, potentially skewing interpretations of ridership patterns. Finally, computational constraints emerged as a bottleneck; the intensive tasks of data cleaning, aggregation, and feature engineering frequently overwhelmed local machines, highlighting the need for more robust infrastructure. These challenges collectively underscored the complexity of working with large-scale, real-world transportation data and influenced the scope and depth of our analysis.

Future Recommendations

Building on the insights from this study and addressing the challenges encountered, several avenues can enhance future research on NYC taxi trends. To tackle the issue of data volume and scalability, adopting scalable processing solutions like cloud-based platforms or distributed computing frameworks such as Apache Spark could significantly improve efficiency. These tools would streamline the merging and analysis of large datasets, reducing processing times and enabling researchers to handle raw trip records more effectively. Additionally, enhancing data quality is a priority; future work could employ advanced imputation methods and anomaly detection algorithms to address missing or erroneous entries, while integrating external datasets—like rideshare trip logs or weather data—could provide richer context and validation for observed trends.

To ensure consistency across evolving data formats, developing a unified data pipeline would be a valuable step forward. This automated system could handle ingestion, cleaning, and transformation, accommodating historical shifts like the transition to taxi zone IDs and preparing the dataset for seamless analysis. Beyond technical improvements, incorporating advanced feature engineering—such as spatial-temporal models or behavioral metrics—could deepen insights into taxi operations and competitive dynamics, building on features like trip duration and fare per mile already explored. Furthermore, integrating multi-modal transportation data from rideshare services, bike-sharing programs, and microtransit platforms would offer a holistic view of urban mobility, contextualizing taxi declines within a broader ecosystem. Finally, leveraging machine learning for predictive analysis could enable forecasting of ridership trends and fare adjustments, accounting for seasonality and external factors, providing actionable

guidance for operators and policymakers. These recommendations collectively aim to refine and expand the study of NYC's transportation landscape.

Contributions to the Project

Yongyi Wang (yw4222):

- Final Report Editing
- Background and Research

Julieta Caroppo (jc6158):

- Data Acquisition
- Merging datasets
- EDA
- Final Report Editing

Keito Taketomi (kt2849):

- Merging datasets
- Final Report Editing

Mei Yue (my2903):

- Data Acquistion
- Final Report Editing
- Feature Engineering
- EDA

Ruoshi Zhang (rz2699)

- Data Cleaning Preprocessing
- Final Report Editing and Structuring
- Data Acquistion

References

- **Cramer, J., & Krueger, A. B. (2016).** Disruptive Change in the Taxi Business: The Case of Uber. National Bureau of Economic Research (NBER) Working Paper.
- **New York City Taxi & Limousine Commission.** Multiple annual reports and Factbook data, available at NYC TLC Website.
- **Curbed NY (2017).** "Uber Surpasses NYC Taxis in Monthly Rides."
- **Shaheen, S., et al. (2016).** "Ride Sharing and Carsharing Services in Urban Mobility." Transportation Research Record.
- **Helling, B. (2018).** "Tipping Culture in Rideshare: A Study of In-App Tipping Behaviors." Journal of Urban Economics.
- **NYC Department of Transportation.** Citi Bike Usage Reports, 2016–2017.