NYC Yellow Taxi: NYC Weather

Ragavi Pobbathi Ashok

University of North Texas

Toulouse Graduate School of Business

ADTA 5940 – Capstone

Dr. Denise Philpot

Fall 2024

Table of Contents

CHAPTER 1 - INTRODUCTION

Background

Taxis are an important mode of Transportation and are particularly helpful in places with fewer public transportation options since they offer door-to-door services. This makes them more convenient than buses or trains that follow set routes, especially in places where public transport is limited and not accessible (Ahmadreza Faghih-Imani, Sabreena Anowar, Eric J. Miller, Naveen Eluru, 2017).

As one of the most highly populated cities in the country, New York City mostly depends on yellow taxis as an easily accessible and practical means of transportation, particularly in places where other public transportation options are scarce or ineffective. The New York City Taxi and Limousine Commission (TLC), established in 1971, regulates and licenses the city's yellow taxis, which are integral to the transportation network throughout Manhattan and its surrounding boroughs. The TLC trip data offers taxi trip information such as pickup and drop-off times, fares, trip distances, passenger counts, tip amount, and more. Taxi trip information analysis can uncover trends in service efficiency, passenger demand, and fare structures, enhancing transportation services for both residents and visitors.

This capstone project aims to analyze the NYC Yellow Taxi Data with NYC Weather Data for January, May, and July 2024. By merging these datasets, the project seeks to explore insights into how weather factors influence taxi rides, as weather conditions such as temperature, precipitation, snowfall, and rain can play a significant role in taxi operations.

Research questions

With the influence of weather conditions on urban transportation, our project aims to explore the dynamics between the NYC Yellow Taxi Data and NYC Weather Data, concentrating on the winter month of January, along with the spring and summer months of May and July 2024. Our research will investigate and answer the following questions:

1. How do trip distance ranges relate to adverse weather conditions, and do taxis show a preference for specific distance ranges during these conditions?
2. Can yellow taxi fares be predicted based on trip characteristics and weather conditions?
3. How do pick-up counts over the three months correlate with weather conditions?
4. Can we predict if a taxi passenger will tip 20% more or less based on fare amounts, and what factors most influence tipping behavior?

CHAPTER 2 - LITERATURE REVIEW

The history of New York's yellow taxis dates back to 1897, when the first electric hansom cabs began operating. Over the years, the taxi industry evolved significantly, with the introduction of the iconic yellow color in 1907 by Harry N. Allen to make his cabs more visible (Backes, 2020) The New York Taxi and Limousine Commission (TLC), established in 1971, regulates the city's taxis, ensuring their safety and service quality, and has been a vital component of the city's transportation network (PERRY, 2023). In July 2024, Yellow Taxi had approximately 97,000 trips per day, spanning 9000 taxi cabs and making around $ 8.500 each day per taxi, which implies that it is a very sought-after mode of transportation(TLC, 2024).

Factors Affecting the Taxis Services

Weather: Weather is a significant factor that influences both the demand for taxis and their availability. Severe weather conditions, such as heavy rain and snow, often alter the demand for taxis and lead to a limited supply, as demonstrated by the study conducted (R. C. P. Wong, 2021) on the Taxi Market in Hong Kong, they observed that extreme weather events significantly reduce the availability of taxis while driving up prices.

(Abel Brodeur , Kerry Nield, 2018), also explored how weather impacts taxi services by providing insights into how location and time interact with weather conditions to create variability in taxi ridership and operations, they observed high demand in urban environments like New York City.

Economic Factors: Economic conditions significantly influence both

the availability and pricing of taxi services. In his 2005 study, (Farber, 2005) he examined how

factors like fuel prices and wage structures influence taxi driver labor supply in New York City. The findings showed that drivers don't stick to a set daily income goal. Instead, they change their working hours based on what they earn each hour and other factors like fuel costs. This flexibility, along with fare rules, can impact how many taxis are on the road. Higher costs or lower wages may result in fewer taxis being available.

(Zhang, Wenbo; Le, Tho V; Ukkusuri, Satish V; Li, Ruimin, 2020) Explores the influencing factors in traditional and app-based taxi systems, focusing on time and location differences. They found factors such as employment rates, vehicle ownership, infrastructure density (transit stations, parking, bike racks), and fuel prices influence traditional and app-based taxi systems.

Traffic and Urban Infrastructure Traffic congestion is another major factor influencing the availability of taxis, particularly in urban centers.

A study in (Zhenhua Chen, Yongjian Yang, Liping Huang, En Wang, Dawei Li, 2018) Beijing revealed how traffic patterns affect taxi services by utilizing GPS trajectory data. They found that during peak traffic hours, taxi availability decreases significantly, especially in city centers and high-traffic zones. This underscores how urban infrastructure and traffic congestion limit the operational efficiency of taxis.

Similarly, (Zhang Y, Sui X, Zhang S, 2024) analyzed the impact of traffic congestion on taxi services in Beijing using Data Field Theory and a Geographically Weighted Regression model. It found that road networks and how land is used in the city affect the availability of taxis. In areas with heavy traffic and dense road networks, fewer taxis were available, while areas with better urban planning had more taxis readily available.

The COVID-19 pandemic also influenced taxi services (Zhang Y, Sui X, Zhang S, 2024) research found that taxi demand dropped significantly due to reduced mobility and health concerns. Interestingly, while demand reduced in business districts, it remained more stable in residential areas. This change in behavior highlights how major disruptions, such as a pandemic, can reshape transportation needs and create new patterns in urban mobility.

Influence of Weather on Taxis

Weather is a critical factor that influences taxi services, diverse weather events such as rain, snow, and extreme temperatures tend to alter typical travel behavior, making it harder for passengers to hail taxis and for drivers to maintain regular service. Research conducted on Hong Kong taxi data during tropical cyclones illustrates this point, showing that up to 80% of taxis stopped operating during severe weather despite having high passenger demand. This shows how concerns for safety and operational challenges can drastically reduce taxi availability, leaving passengers stranded (R. C. P. Wong, 2021).

(Ruijie Bian, Chester G. Wilmot, Ling Wang, 2019) Researchers examined how extreme weather, particularly hurricanes, impacted taxi ridership in New York City. They discovered that taxi ridership dropped significantly about 24 hours before Hurricanes Irene and Sandy, with the steepest declines during weekends and nighttime. This was primarily due to less taxi availability, passenger safety concerns, road flooding, and poor visibility. Models developed to predict taxi demand in China found that temperature, weather conditions, and ozone levels are significant factors influencing the need for taxis, particularly in high-demand areas. During rainy weather in Wuhan, taxi travel decreased by over 4% on average, with weekends showing an even higher reduction in trips. However, nighttime rain on weekdays led to increased taxi demand,

suggesting that passengers adjust their travel preferences to avoid public transportation during poor weather (Zhizhen Liu, Hong Chen, Yan Li, Qi Zhang, 2020).

(Rong Chen, Lingjia Liu, Yongping Gao, 2024), The researchers studied how rainfall impacts taxi travel by using taxi route and weather data from Wuhan, China. They found that weekly taxi trips dropped by 4.16% due to rainfall, with a 3.96% decrease on workdays and a bigger 4.64% drop on weekends. Interestingly, they noticed that on weekdays, rain between 7 PM and 10 PM increased demand for taxis, showing that bad weather changes travel choices. On weekends, rain had a stronger effect by reducing trips to leisure spots and limiting movement. This study highlights how rain affects taxi availability, efficiency, traffic, and travel behavior in cities.

(M. Anil Yazici , Camille Kamga, Abhishek Singhal , 2016). Research in Shanghai demonstrated that rainfall during peak hours significantly reduces the quality of taxi service, making it harder for passengers to find a taxi (Jian Sun, He Dong, Guoyang Qin, Ye Tian, 2020).

A study was carried out to examine how bad weather affects tipping behavior in New York Taxis. Results indicated that passengers tend to tip more generously during poor weather, possibly as a way to make up for the inconvenience of traveling in such conditions. This shift in behavior highlights how weather influences both demand and the interactions between drivers and passengers ((Won Kyung Lee, So Young Sohn,, 2020)

Can future taxi prices be predicted?

In the prediction of taxi pricing and availability, several studies have explored the use of advanced techniques. (Bhawana Rathore, Pooja Sengupta, Baidyanath Biswas, Ajay Kumar , 2024) Developed AI models (clustering and ordinal regression) on the New York City Taxi dataset to predict taxi pricing across different zones. The model predicted pricing variations

based on real-time factors such as congestion, traffic conditions, and passenger demand. (Guo, Suiming and Chen, Chao and Wang, Jingyuan and Liu, Yaxiao and Xu, Ke and Zhang, Daqing and Chiu, Dah Ming, 2018) applied a similar approach in China, using data from ride-on-demand services with traditional taxis and public transportation systems. Their study also integrated weather conditions and urban maps to improve price predictions through a linear regression model. By analyzing these multiple sources, the researchers provided more information on pricing strategy and the relationship between ride-on-demand services and traditional transportation.

Moving beyond pricing, (Tianyi Li, Guo-Jun Qi, Raphael Stern, 2021)used taxi trip data to predict availability by using LSTM deep learning techniques. This model provided an efficient way to dispatch taxis to high-demand areas, improving availability and efficiency, particularly during the COVID-19 pandemic. This approach helped align (K. -F. Chu, A. Y. S. Lam and V. O. K. Li, 2020) the MultiConvLSTM model for predicting travel demand and origin-destination flows using a large dataset of New York City taxis. Their model outperformed prediction methods, allowing for better vehicle allocation and reduced wait times in high-demand areas, offering a major contribution to the prediction of taxi availability.

We will build our capstone stone project based on previous research by analyzing how taxi fares vary across different seasons and examining the impact of weather conditions on trip characteristics during the winter, spring, and summer of 2024 in New York City.

CHAPTER 3 - METHODOLOGY: DATA PREPARATION

Software

All data manipulation and analysis were performed using Python 3.9 within Jupyter Notebook for data exploration, cleaning, and modeling. Data handling was optimized using PySpark to manage large datasets efficiently. The dataset was stored in 'parquet' format for efficiency, easy access, and processing.

Data Collection

To answer our research questions data for this project was sourced from:

1. NYC Yellow Taxi Trip Data: Monthly data for January, May, and July 2024 was obtained from the NYC Taxi & Limousine Commission website, covering winter, spring, and summer months. This dataset includes various trip characteristics and is stored in Parquet format.
2. Taxi Zone Lookup Table: To provide geographic context, a Taxi Zone Lookup Table was obtained from the NYC Taxi & Limousine Commission. This table maps each location ID in the trip data to specific boroughs and zones.
3. Weather Data: Hourly weather data for January, May, and July 2024 was collected using API access from my Open Weather account based on the latitude and longitude of each pickup location borough. This provided real-time, borough-specific weather conditions, capturing hourly weather details for each borough across the three months. The dataset is stored in 'parquet' format.

Taxi Data Wrangling

Data Combining

The NYC Taxi trip data for January, May, and July 2024 was combined to create a single dataset representing the winter, spring, and summer months. Each monthly dataset was imported separately, with shapes as follows:

- January: 2,964,624 rows, 19 columns
- May: 3,723,833 rows, 19 columns
- July: 3,076,903 rows, 19 columns

After verifying the structure and completeness of each file, the datasets were merged into a single dataset named Taxi Combined, with a total shape of 9,765,360 rows and 19 variables. This combined dataset was saved in Parquet format, allowing for efficient storage and retrieval in subsequent processing and analysis steps. An example of a Taxi Combined dataset with a few variables data is shown in Fig.1 below.

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	pay
0	2	2024-01-01 00:57:55	2024-01-01 01:17:43	1.0	1.72	1.0	N	186	79	
1	1	2024-01-01 00:03:00	2024-01-01 00:09:36	1.0	1.80	1.0	N	140	236	
2	1	2024-01-01 00:17:06	2024-01-01 00:35:01	1.0	4.70	1.0	N	236	79	
3	1	2024-01-01 00:36:38	2024-01-01 00:44:56	1.0	1.40	1.0	N	79	211	
4	1	2024-01-01 00:46:51	2024-01-01 00:52:57	1.0	0.80	1.0	N	211	148	
5	1	2024-01-01 00:54:08	2024-01-01 01:26:31	1.0	4.70	1.0	N	148	141	
6	2	2024-01-01 00:49:44	2024-01-01 01:15:47	2.0	10.82	1.0	N	138	181	
7	1	2024-01-01 00:30:40	2024-01-01 00:58:40	0.0	3.00	1.0	N	246	231	
8	2	2024-01-01 00:26:01	2024-01-01 00:54:12	1.0	5.44	1.0	N	161	261	
9	2	2024-01-01 00:28:08	2024-01-01 00:29:16	1.0	0.04	1.0	N	113	113	
10	2	2024-01-01 00:35:22	2024-01-01 00:41:41	2.0	0.75	1.0	N	107	137	

Figure 1: taxi combined example

Data Filtering

Data filtering was applied to streamline the dataset and align it with the analytical goals of the project. This filtering focused on areas with the most data coverage, completed trips, and real-time entries, improving processing efficiency and keeping the analysis manageable.

The VendorID column identifies the taxi service provider for each trip. Initial analysis of data distribution by VendorID showed that VendorID 2 had the largest share, with 7,420,531 rows, and VendorID 1 had 2,344,344 rows. VendorID 6 had the least share, 485 rows. VendorID 2 had a large amount of data, which could cause performance issues, and VendorID 6 had too little data to be useful. Therefore, we filtered the Taxi Combined dataset to focus only on VendorID 1.

We created a trip distance range column to categorize trips by distance. Analyzing the trip distance distribution showed that most trips were within the 0–20-mile range, with a significant drop in counts beyond this range. We filtered the data to include only trips within this 0–20-mile range, focusing on the typical trip distances that make up most of the dataset. The visual representation is shown in the graph and Table 1 and Fig 2 below.

Trip Distance Range (miles)	Number of Trips
0–1	505,498
1–2	706,215
2–5	561,086
5–10	181,878
10–20	149,388
20–50	18,303
50–100	200

Table 1: Trip distance range and number of trips

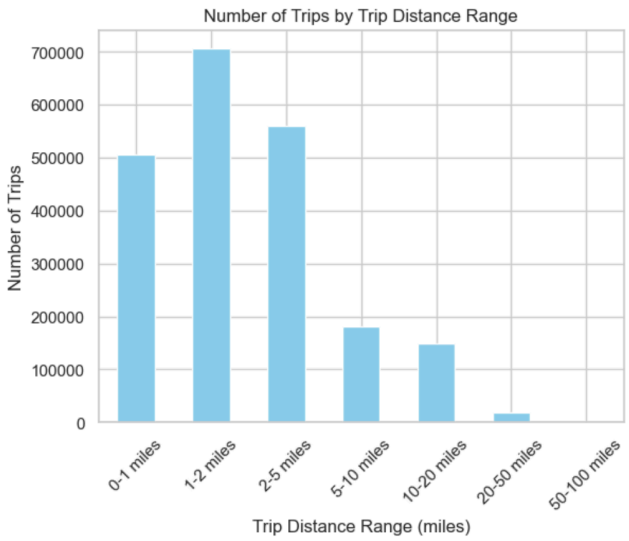

Figure 2: Number of trips versus Trip distance range(miles)

Additional filters were applied to ensure data quality:

- Payment Types: We kept only the commonly used payment types to focus on standard transactions and ensure consistency. The payment type column shows how the passenger paid, coded by numbers. We included:

  1 = Credit card

  2 = Cash

  Other codes, like 3 (No charge), 4 (Dispute), 5 (Unknown), and 6 (Voided trip) rows were filtered.

- Store and Forward Flag: This flag indicates whether a trip record was temporarily stored on the vehicle before being transmitted to the server. To align the taxi data with real-time weather information, we included only trips with real-time transmission (flag = 'N'). This ensured data accuracy, as stored trips (flag = 'Y') might not match real-time weather conditions due to delays in data transmission.

  After applying the filters, we have 2,064,608 rows and 20 columns in the dataset.

Data Cleaning

Handling Null Values

During the data cleaning process, we first checked for null values in each column of the taxi combined dataset. Most columns showed zero null values, indicating complete data. However, the trip distance range column had 34 null entries. To examine the NaN values in the trip distance range column, we plot a bar graph with the count of trips versus the trip distance range with the counts of trips in that range and the trips total to 34.

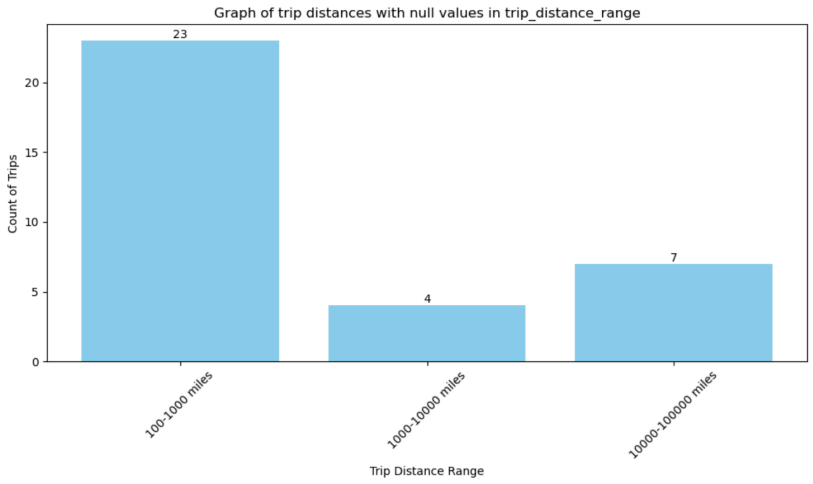

Figure 3: Trip distances column with Null values in Trip Distance Range

The bar graph in Fig 3 above visualizes the count of trips with exceptionally large trip distance values, reaching into the thousands and even tens of thousands of miles, beyond the scope of realistic NYC taxi trips.

By observing this pattern, we can infer that these extreme distances are likely data errors or outliers. Including them would skew our analysis, so we chose to remove these entries from the dataset to maintain data quality and accuracy, and it was less than 5% of our total data.

After filtering out rows with null values in the trip distance range column, we confirmed that no null values remained in the column. The Taxi Combined dataset was then verified by checking for any remaining null values across all columns, resulting in a fully complete dataset. After removing the rows with null values in the trip distance range column, the dataset was reduced to 2,064,574 rows and 20 columns.

Handling Outliers

As part of our data-wrangling process, we analyzed each column to identify and handle outliers that could impact the accuracy and reliability of our project. Outliers, defined as values that fall outside the expected range, can result from data entry errors, unusual cases, or anomalies and need to be managed carefully to ensure a clean dataset.

Passenger Count: Our initial focus was the passenger count column. Through our analysis, we determined that typical NYC taxi trips generally range from 1 to 5 passengers. Any values outside this range, specifically counts of 0 or greater than 5, were flagged as outliers, as these are unlikely scenarios for standard taxi trips. We found 95,655 rows with such outlier values, shown in Table 2 and Fig 4 below.

Passenger Count	Occurrences
0.0	93599
6.0	2054
7.0	1
9.0	1

Table 2: Anomalous Passenger Counts

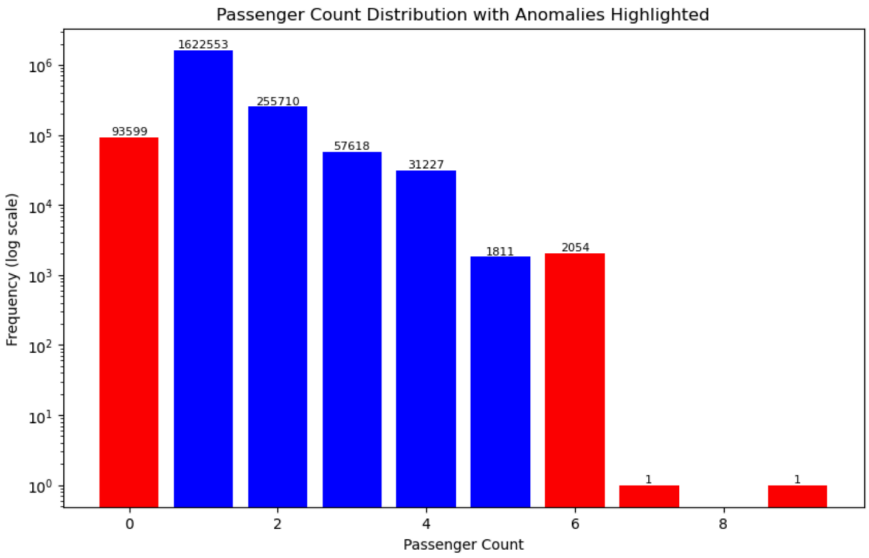

Figure 4: Passenger Count with Anomalies Highlighted

Since the anomalous passenger counts 4.63% of the dataset, which falls below the Practical Significance Threshold of 5%, we opted to remove these rows.

Trip Distance: Our next focus was the trip distance column. Through our analysis, we identified rows where the trip distance is 0.0, yet a fare was charged. Such entries are likely data errors, as a recorded fare should correspond to an actual trip distance. We found 25,426 rows with these anomalies represented in Fig 5.

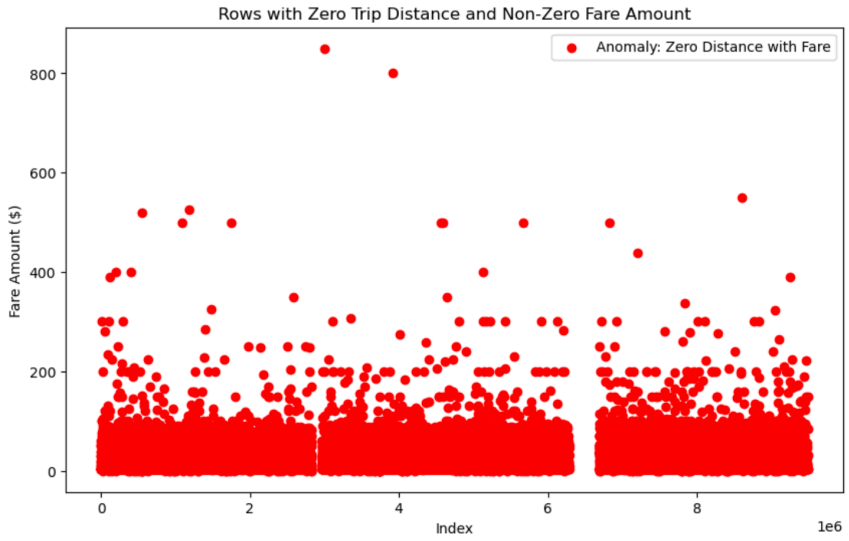

Figure 5: Zero distance with Non-Zero Fare amount.

This scatter plot shows rows where the trip distance column is 0.0, but a fare has been charged. The x-axis represents the row indices spread across the dataset, while the y-axis shows the fare amount column, which ranges widely, even reaching high values. These entries are likely data errors, as a fare should correspond to a traveled distance and represent 1.29% of the dataset. Removing these anomalies improves data accuracy.

Fare Amount: We identified trips with a fare amount of 0 despite having a non-zero trip distance. According to NYC taxi fare regulations, all trips should have at least a minimum fare of $2.50, making these zero-fare entries highly unlikely to be valid.

Descriptive Statistics: For trips with a zero fare amount, the average trip distance was 4.35 miles, with distances ranging from 0.1 to 18.4 miles. The breakdown of these trips is as follows:

- 25th Percentile: 0.5 miles
- Median (50th Percentile): 2.75 miles
- 75th Percentile: 5.45 miles

These distances show a variety of trip lengths but still record no fare, indicating a high likelihood of data error with 150 rows with a fare amount of 0, representing 0.01% of the total dataset. Since this is a very small percentage and likely due to data entry errors or anomalies, we decided to remove these rows.

We looked at trips with a trip distance of 20 miles or less to find any unusually high fares. For a 20-mile trip in NYC, the fare generally averages around $50, including the base fare, per-mile charges, and common surcharges. Even with extra fees, a fare over $100 for trips within 20 miles is unusual.

We set $100 as a limit because:

- Normal Fare Range: Fares rarely go over $100 for 20 miles, even with tolls and rush-hour surcharges.
- Realistic NYC Pricing: This limit helps flag errors without removing significantly high fares in rare cases.

Using this $100 threshold, we found 917 trips (0.05% of the dataset) with fares over this amount. Because these cases are rare and likely errors, we removed them to keep our data clean and focused on realistic trip costs.

We also identified 68 trips with fare amounts less than $2.50. Such values are inconsistent with the standard fare structure of NYC taxis, where the minimum fare should be $2.50. we decided to remove these entries from the dataset to ensure the accuracy and integrity of our analysis.

MTA Tax: In our analysis of the MTA tax column, we identified 11 rows with values of 3.25 and 4.00. The standard MTA tax for NYC taxis is a fixed $0.50 for applicable trips, and

it does not vary. These higher values do not align with the official fare structure and are likely data entry errors.

These 11 rows accounted for only 0.0006% of the dataset, representing rare occurrences. To maintain data consistency and accuracy, we removed these rows, as their impact on the overall analysis is minimal, as shown in Table 3 below.

MTA Tax ($)	Count
0.50	1,930,885
0.00	11,379
4.00	9
3.25	2

Table 3: MTA tax count

Toll Amounts: In our analysis of the tolls amount column, we identified eight trips with toll charges exceeding $50, which accounted for just 0.0004% of the dataset. According to the NYC taxi fare guidelines and typical toll rates, toll amounts above $50 are highly unusual for standard taxi trips.

The threshold of $50 was chosen based on common toll rates within NYC and surrounding areas. Even for trips that involve major toll crossings, such as the Verrazzano-Narrows Bridge (around $19) or multiple tolls through tunnels and bridges, typical tolls do not usually exceed this amount. Additionally, NYC taxis are required to use EZ Pass, which applies discounted rates for tolls, further limiting toll costs((NYC Taxi & Limousine Commission, 2022). By setting a $50 threshold, we ensure that the dataset remains representative of realistic trip costs while excluding rare, atypical cases or potential data errors.

Given the small percentage and the unlikely nature of these entries, we removed these eight trips from the dataset to maintain data accuracy and prevent outliers from skewing fare analysis.

Tip Amounts: In our analysis of the tip amount column, we found three trips with very high tips: $437.84, $228.00, and $300.00. These trips were linked to the following pickup locations:

- PULocationID 162: Upper East Side South, Manhattan
- PULocationID 68: Lincoln Square East, Manhattan
- PULocationID 193: JFK Airport, Queens

These locations are known for high demand and can lead to higher tips, especially for trips starting at JFK Airport.

Since these high tips represent a small portion of the dataset and are tied to specific, well-known locations, we decided to keep them in the dataset. This decision helps maintain the integrity of the data and ensures we capture realistic tipping behavior in NYC taxis.

Duplicate Rows Check: We checked the dataset for duplicate entries in the combined dataset. The results showed that there are 0 duplicate rows in the dataset. This means each record is unique, which is important for ensuring the accuracy of our analysis. When we attempted to display any duplicate rows, an empty DataFrame was returned, confirming that there were no duplicates present.

Spatial Data Integration

Location Data Reading and Combining

The location data was read from a parquet file containing taxi zone lookup information. An example of a location dataset with few variables data is shown in Fig.6 below.

We used the location dataset to map PULocationID and DOLocationID to their corresponding boroughs and zones.

	LocationID	Borough	Zone	service_zone
0	1	EWR	Newark Airport	EWR
1	2	Queens	Jamaica Bay	Boro Zone
2	3	Bronx	Allerton/Pelham Gardens	Boro Zone
3	4	Manhattan	Alphabet City	Yellow Zone
4	5	Staten Island	Arden Heights	Boro Zone

Figure 6: Location dataset example

We created dictionaries to map LocationID to their corresponding Borough and Zone. This mapping allows us to assign the appropriate borough and zone information to each trip based on its pickup and drop-off locations. After mapping, we analyzed the unique values for PULocation Borough and DOLocation Borough. This analysis confirmed that there were seven different boroughs present in the dataset, including Manhattan, Queens, Brooklyn, Bronx, Staten Island, and EWR.

The Taxi Combined dataset's few rows are displayed in Figure 7 below.

	tpep_pickup_datetime	PULocationID	PULocation_Borough	PULocation_Zone	tpep_dropoff_datetime	DOLocationID	DOLocation_Borough	DOLocation_Zone
1	2024-01-01 00:03:00	140	Manhattan	Lenox Hill East	2024-01-01 00:09:36	236	Manhattan	Upper East Side North
2	2024-01-01 00:17:06	236	Manhattan	Upper East Side North	2024-01-01 00:35:01	79	Manhattan	East Village
3	2024-01-01 00:36:38	79	Manhattan	East Village	2024-01-01 00:44:56	211	Manhattan	SoHo
4	2024-01-01 00:46:51	211	Manhattan	SoHo	2024-01-01 00:52:57	148	Manhattan	Lower East Side
5	2024-01-01 00:54:08	148	Manhattan	Lower East Side	2024-01-01 01:26:31	141	Manhattan	Lenox Hill West

Figure 7: Taxi Combined dataset

Data Filtering

We examined the unique values and their counts for the PULocation Borough column to understand the distribution of taxi pickups across different boroughs. The initial counts revealed the following distribution:

- Manhattan: 1,703,359 pickups
- Queens: 142,841 pickups
- Brooklyn: 46,265 pickups
- Bronx: 14,023 pickups
- EWR: 11 pickups
- Staten Island: 5 pickups

Given this distribution, we identified EWR (which refers to Newark Liberty International Airport) and Staten Island as locations with minimal data points relative to the overall dataset. To enhance the quality of our analysis and focus on more significant boroughs, we decided to filter rows where PULocation Borough was EWR or Staten Island. This filtering helps ensure that our analysis is concentrated on areas with substantial taxi activity, allowing for more reliable insights into the factors affecting taxi fares and availability. After this step, the revised counts for PULocation Borough were as follows:

- Manhattan: 1,703,359 pickups
- Queens: 142,841 pickups
- Brooklyn: 46,265 pickups
- Bronx: 14,023 pickups

Null Value Check

A check was conducted for any rows with null values in the taxi combined dataset. The results indicated no null values in either column, confirming that the data was complete for these fields.

Weather Data Wrangling

Data Extraction and Combining

For our analysis, we extracted historical weather data using the Open Weather API. We focused on specific locations within New York City, including Manhattan, Brooklyn, Queens, and the Bronx. The extraction covered the months of January, May, and July in 2024 to represent different seasons.

To facilitate the data retrieval, we defined the coordinates for each location and created a function to fetch the weather data. The function made HTTP requests to the Open Weather API, retrieving hourly weather information for each specified date and including temperature (in Fahrenheit), precipitation (in millimeters), and general weather conditions.

The gathered data was organized into a structured format, and after fetching the necessary data for all specified locations and dates, we compiled it into a data frame. This data Frame was then saved in '.parquet' format for efficient storage and future analysis. A few values of weather data are shown in the Fig 8 below.

	location	date	time	temp_f	precip_mm	condition
0	Brooklyn	2024-01-01	2024-01-01 00:00	41.6	0.08	Patchy rain possible
1	Brooklyn	2024-01-01	2024-01-01 01:00	41.5	0.01	Patchy rain possible
2	Brooklyn	2024-01-01	2024-01-01 02:00	41.7	0.00	Cloudy
3	Brooklyn	2024-01-01	2024-01-01 03:00	41.6	0.00	Cloudy
4	Brooklyn	2024-01-01	2024-01-01 04:00	41.4	0.00	Overcast

Figure 8: Weather dataset

Weather condition Analysis

The weather data we collected includes several unique conditions, which are listed below:

- Patchy rain possible
- Cloudy
- Overcast
- Partly cloudy
- Clear
- Sunny
- Moderate snow
- Light snow
- Light sleet
- Light rain
- Light drizzle
- Heavy snow
- Light snow showers
- Fog
- Mist
- Moderate or heavy snow showers
- Moderate rain
- Heavy rain
- Light rain shower
- Patchy light snow
- Patchy snow possible

- Blowing snow
- Moderate rain at times
- Light freezing rain
- Patchy light drizzle
- Moderate or heavy rain with thunder
- Patchy light rain
- Thundery outbreaks possible
- Moderate or heavy rain shower
- Patchy light rain with thunder
- Heavy rain at times

This list of unique weather conditions is essential for understanding how different weather scenarios may affect taxi usage and trip patterns in New York City.

Null Value Check

After the extraction, we conducted a thorough check for any null values within the weather data. The check confirmed that there were no null in all the columns. This indicates that our dataset is complete and ready for analysis.

Taxi And Weather Data Wrangling

NYC Taxi Weather Data Combining

To integrate the NYC taxi data with the corresponding weather data, we first converted the date and time columns in the weather dataset to DateTime objects. This conversion was essential for

ensuring accurate merging with the taxi data. Similarly, the tpep pickup datetime in the taxi combined dataset was converted to maintain consistency across both datasets.

Next, we extracted the hour from the tpep pickup datetime and created a new column called pickup hour. This step allowed us to match the taxi trip times with the hourly weather data, which is crucial for a proper analysis of how weather conditions affect taxi usage.

Following this, we merged the taxi combined dataset with the weather dataset based on the pickup borough and the extracted hour. We used a left join to ensure that all taxi records were retained, even if there was no corresponding weather data available for some hours, It was saved in the taxi weather dataset. A few variables of the combined dataset are shown in Fig 9 below.

	tpep_pickup_datetime	PULocation_Borough	PULocation_Zone	temperature	precip_mm	weather_condition
0	2024-01-01 00:03:00	Manhattan	Lenox Hill East	40.6	0.1	Patchy rain possible
1	2024-01-01 00:17:06	Manhattan	Upper East Side North	40.6	0.1	Patchy rain possible
2	2024-01-01 00:36:38	Manhattan	East Village	40.6	0.1	Patchy rain possible
3	2024-01-01 00:46:51	Manhattan	SoHo	40.6	0.1	Patchy rain possible
4	2024-01-01 00:54:08	Manhattan	Lower East Side	40.6	0.1	Patchy rain possible

Figure 9: NYC Taxi and Weather Dataset

To evaluate the impact of weather on ride pickups, four distinct weather conditions—Rain, Snow, Sunny, and Clear—were defined based on precipitation and temperature data. Each condition was established with exclusive criteria to ensure that only one condition applied at a given time.

- Rain: An hour was classified as Rain if there was precipitation greater than 0 mm and the temperature was above 32°F. This criterion identifies periods when precipitation likely occurs as rain, excluding snowfall.

- Snow: An hour was classified as Snow if there was precipitation greater than 0 mm and the temperature was 32°F or below, indicating conditions where precipitation is likely to fall as snow.
- Sunny: An hour was classified as Sunny if there was no precipitation (0 mm) and the temperature exceeded 59°F. This condition represents warmer, precipitation-free hours typically associated with sunny weather.
- Clear: An hour was classified as Clear if there was no precipitation (0 mm) and the temperature was 59°F or below, distinguishing clear, cooler periods from the warmer "Sunny" category.

This classification scheme was developed to ensure that each hour could be categorized under one specific weather type, avoiding overlaps and simplifying the analysis of weather impacts on ride pickups, as shown in Fig below.

	tpep_pickup_datetime	PULocation_Borough	PULocation_Zone	temperature	precip_mm	Rain	Snow	Sunny	Clear
0	2024-01-01 00:03:00	Manhattan	Lenox Hill East	40.6	0.1	True	False	False	False
1	2024-01-01 00:17:06	Manhattan	Upper East Side North	40.6	0.1	True	False	False	False
2	2024-01-01 00:36:38	Manhattan	East Village	40.6	0.1	True	False	False	False
3	2024-01-01 00:46:51	Manhattan	SoHo	40.6	0.1	True	False	False	False
4	2024-01-01 00:54:08	Manhattan	Lower East Side	40.6	0.1	True	False	False	False

Figure 10: Taxi Weather Combined Dataset

Null Value Check

After merging the NYC taxi data with the weather data, we conducted a thorough check for null values across all columns in the merged dataset. The results showed that there were no null values present in any of the columns, confirming the completeness and integrity of the data. The shape of the merged dataset was confirmed to be (1,906,488, 29), indicating a total of 1,906,488 rows and 29 columns.

CHAPTER 4 - METHODOLOGY: EXPLORATORY DATA ANALYSIS

Descriptive Statistics

Descriptive statistics of Numerical variables are shown in the table below:

Variable	Obs	Mean	Std. dev.	Min	Max
Passenger Count	1,906,488	1.24	0.59	1	5
Trip Distance (miles)	1,906,488	3.12	3.78	0.10	19.90
Fare Amount ($)	1,906,488	17.87	14.31	2.70	100.00
Tip Amount ($)	1,906,488	3.20	3.47	0.00	437.84
Tolls Amount ($)	1,906,488	0.48	1.89	0.00	46.14
Total Amount ($)	1,906,488	26.31	18.17	4.00	461.64
Temperature (°F)	1,906,488	60.47	19.69	14.60	96.80
Precipitation (mm)	1,906,488	0.15	0.57	0.00	11.21

Table 4: Descriptive Statistics of Numerical Variables

Variables Explanation:

- Passenger Count: This variable represents the number of passengers per trip. The average is approximately 1.24, indicating that most trips typically carry one passenger, while the maximum recorded is five.
- Trip Distance: This measures the length of each trip in miles, averaging around 3.12 miles. Distances vary significantly, with some trips as short as 0.1 miles and others nearly reaching 20 miles.
- Fare Amount: This indicates the total fare charged for each trip. The average fare is roughly $17.87, with the minimum fare set at $2.50 and a maximum fare of $100.
- Tip Amount: This reflects the gratuity passengers leave. On average, tips amount to about $3.20, with the highest tip recorded being $437.84.
- Tolls Amount: This represents the total toll fees incurred during a trip, averaging around $0.48, with many trips that do not incur any tolls.

- Total Amount: This is the overall cost to passengers, including the fare, tolls, and tips, with an average total of $26.31.
- Temperature: These variables record the temperature during each trip in degrees Fahrenheit, averaging about 60.47°F, which can influence passenger demand.
- Precipitation: This measures the amount of rainfall or moisture recorded in millimeters, with an average of 0.15 mm, suggesting that most trips occurred under dry conditions.

Categorical Variables Analysis

Payment Type Distribution by Pickup Borough:

In this section, we analyzed the distribution of payment types across different pickup boroughs to understand passenger behavior regarding fare payment methods. We calculated the total number of trips for each borough and determined the percentage of payments made in cash versus those made by credit card.

The results are summarized in the table below:

PU Location Borough	Total Trips	Cash Payments (%)	Credit Card Payments (%)
Manhattan	1,703,359	15.45%	84.55%
Queens	142,841	17.37%	82.63%
Brooklyn	46,265	2.75%	97.25%
Bronx	14,023	3.08%	96.92%

Table 5: Categorical Variables Analysis

- **Total Trips**: This column shows the total number of taxi trips recorded in each borough.
- **Cash Payments (%)**: This column indicates the percentage of trips that were paid for in cash.

- **Credit Card Payments (%)**: This column reflects the percentage of trips paid for using credit cards.

This analysis provides insights into payment preferences across different boroughs, highlighting trends in passenger behavior concerning fare payments.

Visualizations

Distribution of Trip Distances

The histogram below illustrates the distribution of trip distances for NYC taxi rides. The x-axis represents the trip distance in miles, while the y-axis indicates the frequency of trips within specific distance ranges.

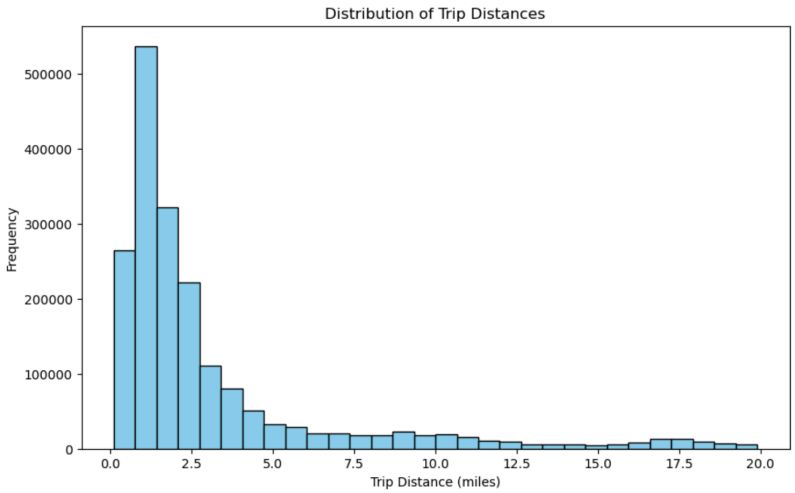

Figure 11: Distribution of Trip Distances

The histogram reveals that most taxi trips are short, with a peak between 0.0 and 5.0 miles, reflecting typical urban travel patterns. The right-skewed distribution indicates that while short trips are prevalent, there are fewer longer rides, particularly beyond 10 miles. Very few trips exceed 15 miles, highlighting the infrequency of long-distance taxi rides in the dataset.

Distribution of Total Amount

The histogram below shows the distribution of the total amount charged to passengers for NYC taxi rides. The x-axis represents the total amount in dollars, and the y-axis indicates the frequency of trips within specific total amount ranges.

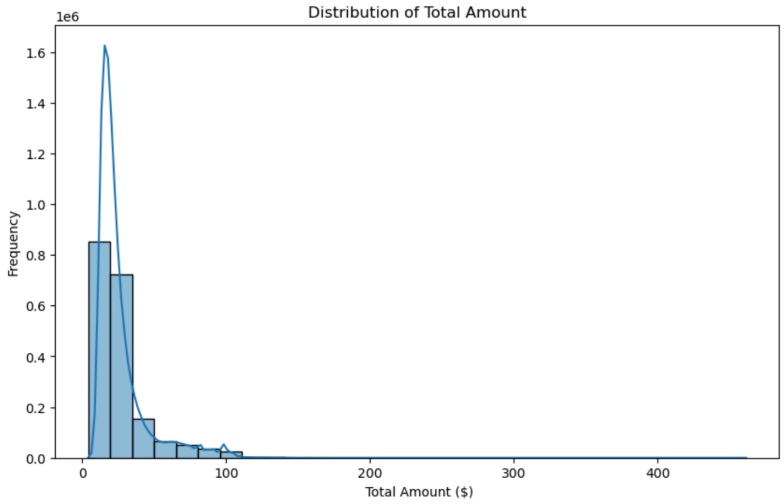

Figure 12: Distribution of Total Amount

The histogram shows that most taxi fares are primarily between $0 and $50, which is typical for urban travel. The distribution is right-skewed, indicating that while low fares are common, there are a few significantly higher fares, leading to a tail on the right side. Although the dataset includes fares that can reach up to $450, these occurrences are rare, as shown by the sharp drop-off in frequency at higher amounts. This visualization emphasizes the prevalence of lower fares while highlighting that extreme values are uncommon.

Average Total Amount per Trip Distance Range

The bar chart below illustrates the average total fare amount for taxi trips categorized by different trip distance ranges. The x-axis represents the trip distance range in miles, and the y-axis indicates the average total amount of trips in dollars.

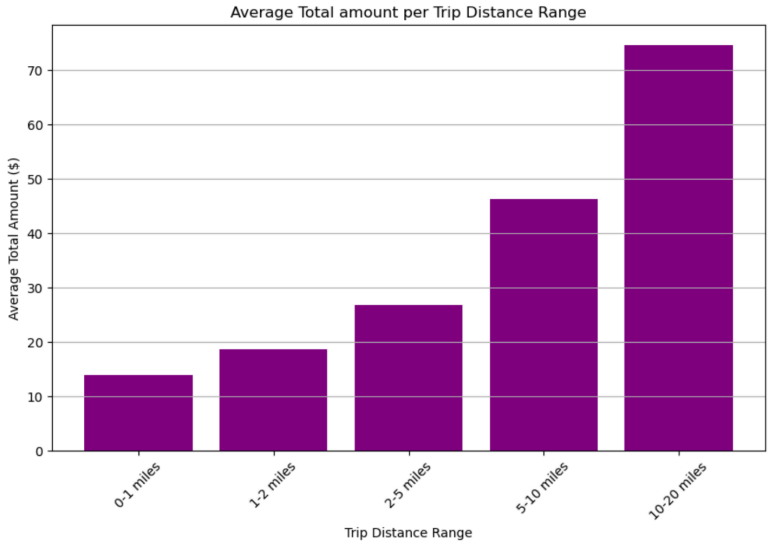

Figure 13: Average Total Amount vs Distance Range

Key observations reveal an increasing trend in average fare amounts with longer trip distances. For short trips (0-1 miles), fares are lowest, while the average fare rises gradually for distances of 1-2 miles and 2-5 miles. A significant increase in average fare occurs for trips between 5-10 miles, and the highest average fare is found in the 10-20 miles range, indicating these longer trips incur more charges. This analysis is vital for understanding pricing models in the NYC taxi service.

Average Total Amount by Pickup Location

The bar chart below illustrates the average total fare amount collected for taxi trips based on their pickup locations in New York City. The x-axis represents the pickup location borough, and the y-axis indicates the average total amount of trips in dollars.

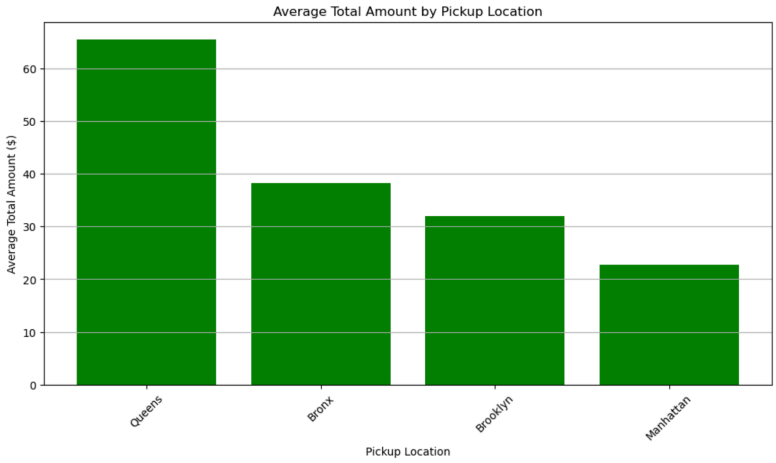

Figure 14: Total Amount vs Pickup Location Borough

Key observations from the graph show Queens has the highest average total fare, likely due to longer distances or heavier traffic in the area. The Bronx and Brooklyn have moderate average fares, indicating that trips from these boroughs can also be expensive but not as much as those from Queens. Manhattan, despite its high traffic, shows a lower average fare, suggesting that many trips originating here are shorter in distance. This analysis helps to understand fare dynamics in NYC's taxi services.

Average Tip Amount by Pickup Borough

The bar chart below represents the average tip amounts received by taxi drivers based on the pickup borough in New York City. Each bar indicates the average tip amount in dollars, allowing us to compare tipping behavior across different areas.

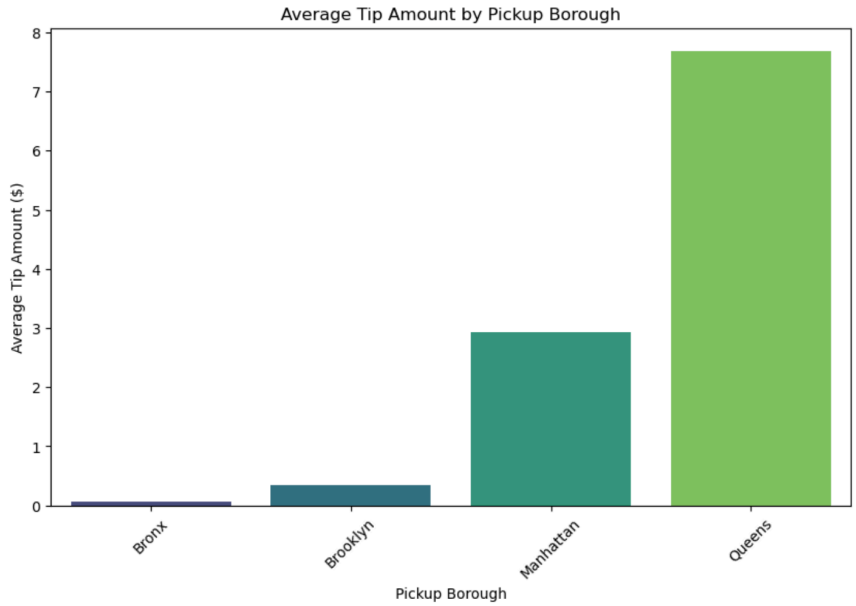

Figure 15: Average tip Amount vs Pickup Borough

Key insights from the graph show Queens have the highest average tip amount, and Manhattan follows closely. In contrast, Brooklyn and the Bronx show significantly lower average tip amounts. This analysis shows tipping trends across NYC's boroughs.

Distribution of Temperature on Trip Days

The histogram below shows the distribution of temperatures recorded on trip days. The x-axis represents the temperature in degrees Fahrenheit (°F), while the y-axis indicates the frequency of trips occurring at each temperature range.

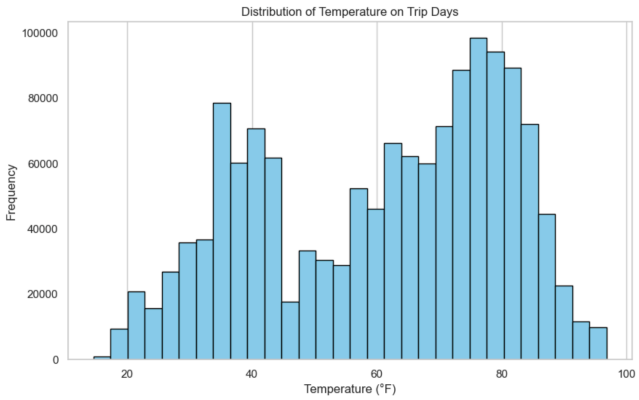

Figure 16: Distribution of Temperature

Key observations from the histogram show that there are more trips when temperatures are between 60°F and 80°F, suggesting that nice weather leads to more taxi use. When temperatures drop below 40°F or go above 80°F, the number of trips decreases, indicating that people may choose to stay indoors during extreme weather. This pattern suggests that weather affects how often people use taxis.

Fare Amount Variation by Pickup Hour

The box plot below displays the distribution of fare amounts across each hour of the day. It shows that fare amounts are higher between 4 AM and 7 AM, likely due to early morning trips such as airport rides or higher off-peak fares. During the day and evening hours, fare amounts stabilize with consistent median values and a narrower range. Outliers, most visible in the early morning, indicate occasional high fares, potentially from longer trips or peak demand. This chart provides a clear view of how fare prices fluctuate by time of day.

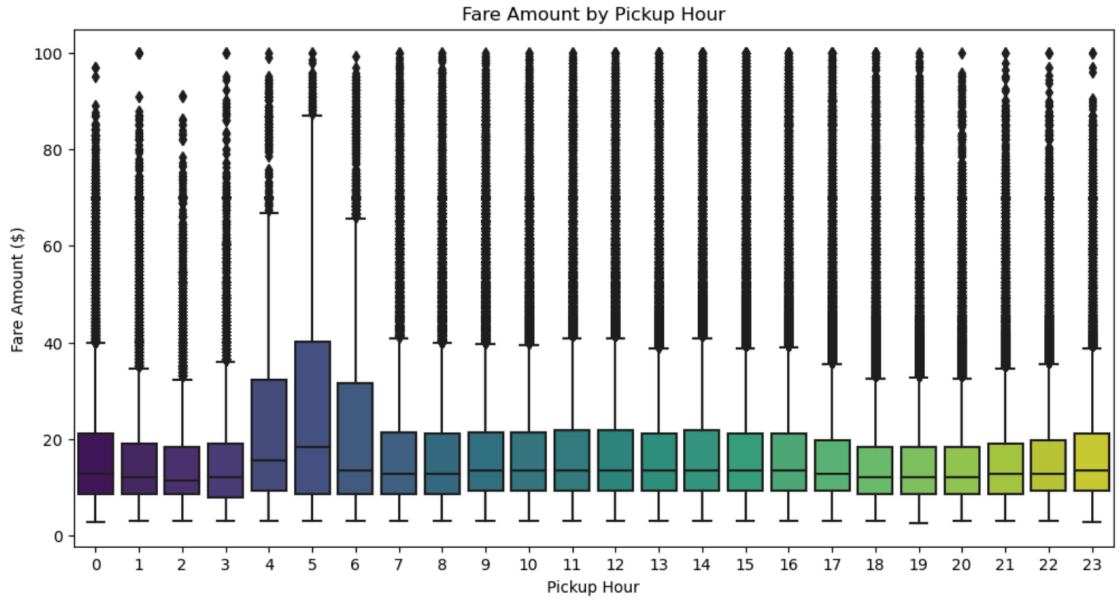

Figure 17: Box plot for Fare Amount by Pickup hour

Weighted Average Taxi Fare by Distance Range and Passenger Count in Manhattan

The bar below chart shows the weighted average taxi fare in Manhattan across different trip distance ranges, segmented by passenger count. Fares increase as the trip distance grows, with short trips (0-1 miles) averaging around $10, while longer trips (10-20 miles) average over $60. Across all distance ranges, the fare remains relatively consistent regardless of passenger count, suggesting that passenger count has little effect on fare rates within each distance category. This pattern highlights that fare pricing is primarily driven by distance rather than the number of passengers.

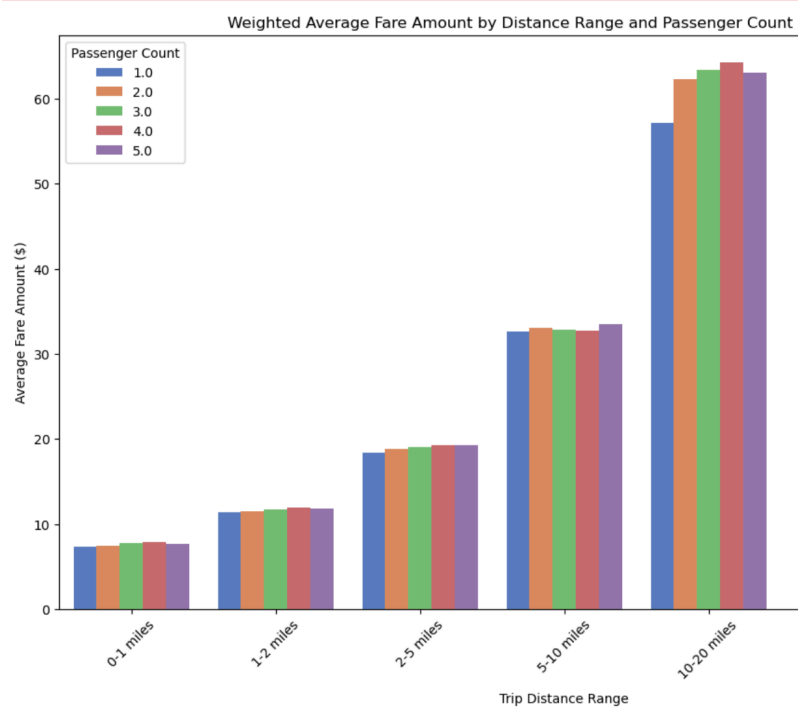

Figure 18: Average Fare Amount by Distance Range and Passenger Count

Impact of Rainy and Snowy Weather on Taxi Ride Distances

The table and bar chart below illustrate how rainy and snowy weather affects the distribution of taxi rides across various trip distances The table shows that under both rainy and snowy conditions, shorter trips are more frequent. For instance, during rainy weather, about 34% of rides fall within the 1-2 miles range, and 27% in the 2-5 miles range. Snowy conditions reveal a similar trend, with 36% of rides in the 1-2 miles range and 26% in the 2-5 miles range.

The bar chart focuses on the 1-2 miles and 2-5 miles ranges, making it easier to compare ride proportions between rainy and snowy days. Both conditions show a higher proportion of shorter rides, with snowy days leaning slightly more towards the 1-2 miles range. This pattern suggests that people are more inclined to take shorter trips in poor weather, or taxi drivers prefer small distance rides. The chart effectively complements the table, offering a clear visual comparison of ride distribution between rainy and snowy conditions.

Weather Condition	Trip Distance Range	Ride Proportion
Rain	0–1 miles	22.95%
Rain	1–2 miles	34.19%
Rain	2–5 miles	27.12%
Rain	5–10 miles	8.74%
Rain	10–20 miles	6.99%
Snow	0–1 miles	25.63%
Snow	1–2 miles	35.95%
Snow	2–5 miles	25.82%
Snow	5–10 miles	7.04%
Snow	10–20 miles	5.56%

Table 6: Rainy and Snowy Weather Distribution of Taxi Ride Distances

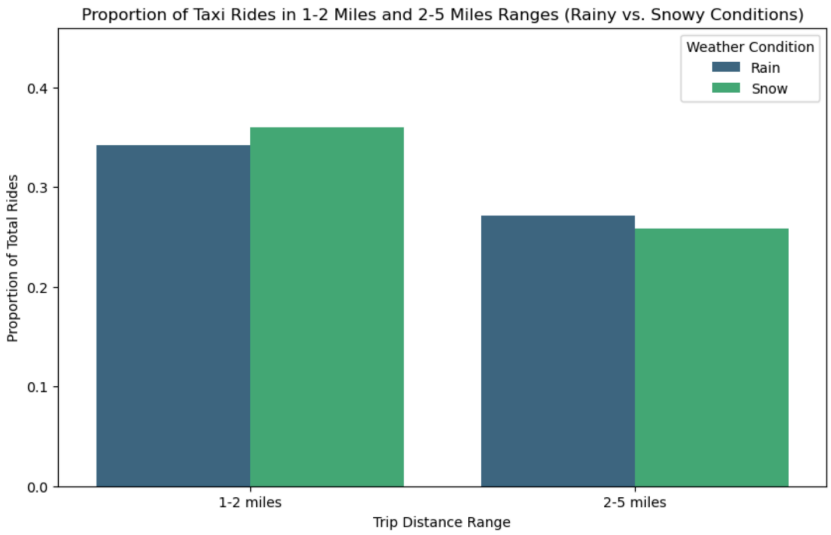

Figure 19: Proportion of Taxi Rides in miles

Daily Taxi Pickup Patterns Across Boroughs and Months

The Daily pickup counts for January, May, and July are shown below for Manhattan, Bronx, Brooklyn, and Queens. Manhattan has the highest pickup for all months, followed by Queens, Brooklyn, and Bronx. The X-axis represents the day of the month and Y-axis represents the Count in the below fig 20.

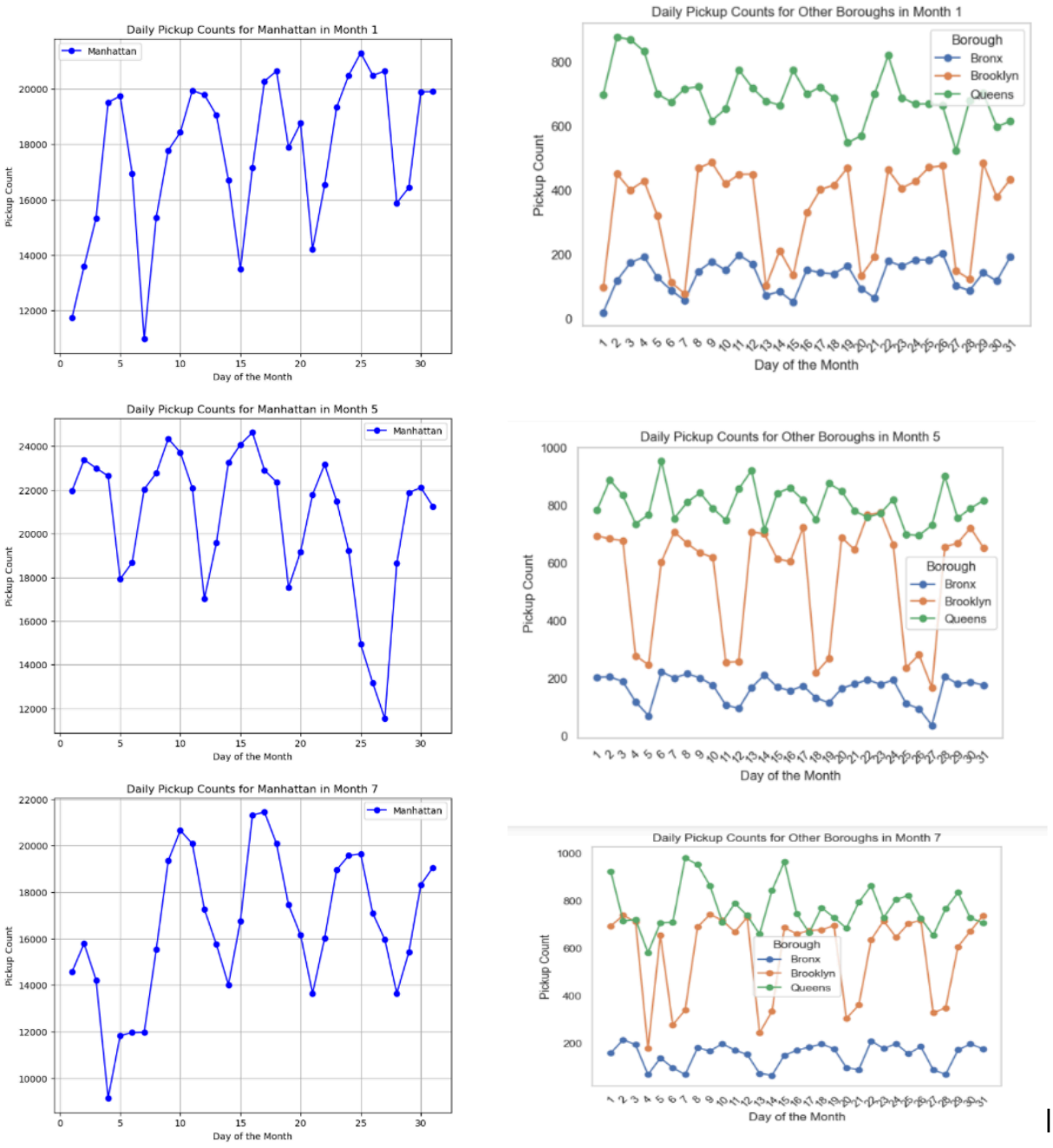

Figure 20: Pickup count of Borough

Monthly Correlations Between Weather Conditions and Daily Pickup Counts

To calculate the correlations between weather conditions and daily taxi pickup counts, we followed these steps:

1. Data Preparation:

- Converted the tpep pickup datetime column to a date format and extracted the month for monthly grouping.
- For each month, grouped the data by day to obtain daily totals for taxi pickups and counts for each weather condition: Rain, Snow, Sunny, and Clear.

2. Aggregation:
   - For each day, we calculate the total pickup count and the sum of hours that each weather condition (Rain, Snow, Sunny, Clear) occurred.
3. Correlation Calculation:
   - For each month, calculated the Pearson correlation between the daily pickup count and each weather condition's daily count. This resulted in a correlation coefficient for each condition, where a positive value indicates a direct relationship, and a negative value indicates an inverse relationship.

Month	Date Correlation	Rain Correlation	Snow Correlation	Sunny Correlation	Clear Correlation
1	0.46	0.15	0.1	NaN	0.21
5	-0.36	0.22	NaN	0.14	0.11
7	0.31	0.42	NaN	0.14	NaN

Table 7: Correlations Between Weather Conditions and Daily Pickup Counts

The analysis reveals how weather impacts taxi pickup counts across the three months:

- Rainy Days: Rain tends to increase taxi demand, especially in July, where the correlation with pickups is strongest (0.42). This suggests that during rainy weather, people prefer to take taxis rather than walk, indicating a heightened reliance on taxi services during inclement weather.
- Clear and Sunny Days: Clear conditions show a positive correlation with taxi pickups, especially in January (0.21), indicating that people are more likely to use taxis on clear

days. However, sunny days have a weak correlation overall, suggesting that the effect of sunny weather on taxi demand is less significant compared to rainy conditions.

- Cold and Snowy Weather: In January, snow has a slight positive correlation (0.10) with pickups, indicating that there might still be a demand for taxis despite the weather. However, in May and July, snowy conditions show no correlation, likely reflecting the absence of snow during these months.

Overall, the data indicates that adverse weather, particularly rain, has a notable effect on taxi counts, with people more likely to opt for taxi services during inclement weather. In contrast, clear and sunny days have a milder influence on pickup counts.

Weighted Average Tipping Percentage by Weather Type

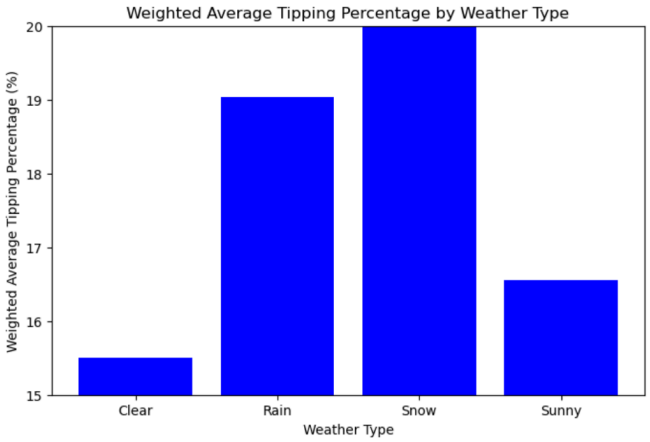

Figure 21: Weighted Average Tipping Percentage by Weather Type

This chart shows the weighted average tipping percentage (%) by weather type (Clear, Rain, Snow, Sunny) A weighted average accounts for each trip's contribution to overall tipping behavior, proportional to its fare amount, ensuring a fair comparison despite uneven trip counts. For our data, the weighted average was calculated as:

Weighted Average Tipping Percentage=(∑(Tip Amount)/ ∑(Fare Amount))×100

Observations:

- Snow (20.18%) has the highest weighted tipping percentage, indicating that passengers tip more generously during snow conditions.
- Rain (19.04%) also has a higher tipping percentage, likely due to passengers appreciating the service in adverse weather.
- Clear (15.5%) and Sunny (16.56%) have relatively lower tipping percentages, reflecting less perceived effort or inconvenience during favorable weather conditions.

Key Insight:

- Adverse weather conditions, such as Snow and Rain, result in higher tipping percentages, potentially due to greater driver effort or passenger gratitude during difficult driving conditions.

Weighted Average Tip Amount by Weather Type

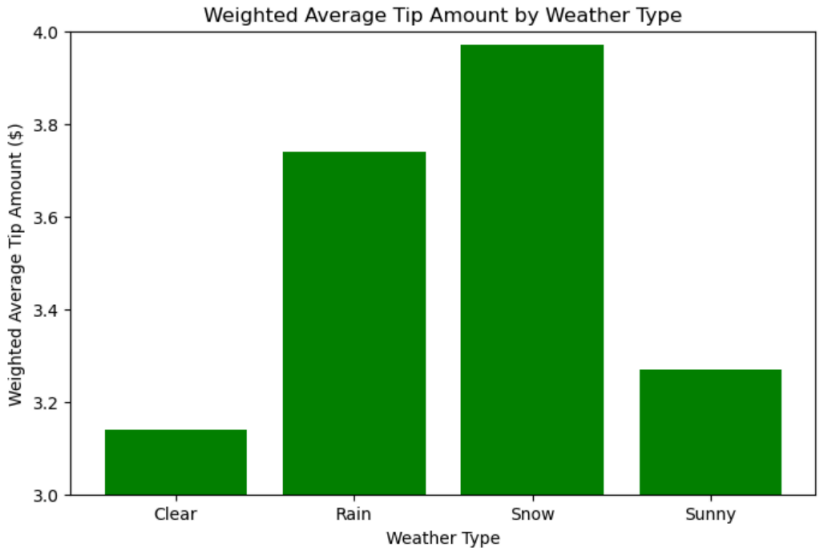

Figure 22: Weighted Average Tip Amount By Weather Type

This chart displays the weighted average tip amount ($) for each weather type (Clear, Rain, Snow, Sunny). A weighted average accounts for the fare revenue of each trip when calculating tip amounts, ensuring that weather types with more revenue (e.g., Sunny) or fewer trips (e.g., Snow) contribute proportionally to the overall result. The formula used is:

Weighted Average Tip Amount= Total Tip Amount Across All Trips/ Total Number of Trips

Observations

- Snow ($3.97) has the highest weighted average tip amount, consistent with the highest tipping percentage observed in the first chart.
- Rain ($3.74) follows closely, further emphasizing increased generosity during adverse weather.
- Sunny ($3.27) and Clear ($3.14) have the lowest weighted tip amounts, aligning with their lower tipping percentages.

Key Insight

Snowy and rainy conditions not only lead to higher tipping percentages but also result in higher tip amounts in absolute dollars. This highlights a clear correlation between adverse weather and passenger generosity.

Revenue and Trip Counts by Weather Type

This section explores how total revenue, average revenue per trip, and trip counts vary across weather conditions (Clear, Rain, Snow, and Sunny). These metrics provide insights into the financial performance of taxi trips under different weather scenarios

Weather Type	Total Revenue ($)	Average Revenue Per Trip ($)	Trip Count
Clear	13,987,888.16	27.90	501,298
Rain	16,273,037.63	28.52	570,563
Snow	1,157,432.75	27.52	42,060
Sunny	23,233,210.55	29.31	792,567

Table 8: Revenue and Trip Counts by Weather Type

Metrics Were Calculated

1. Total Revenue

   Total revenue represents the combined earnings from all trips, including fare amounts, tips, and additional charges. It was calculated using the formula:

   Total Revenue= ∑(fare amount+tip amount+extra charges)

2. Average Revenue Per Trip

   This was calculated by dividing the total revenue for each weather type by the total number of trips:

   Average Revenue Per Trip=Total Revenue /Trip Count

3. Trip Count

   The total number of trips recorded for each weather type was directly derived from the dataset.

Findings

1. Total Revenue
   - Sunny ($23,233,210.55) generated the highest total revenue, primarily due to its large trip count (792,567 trips).
   - Rain ($16,273,037.63) followed, while Clear ($13,987,888.16) ranked third.

- Snow ($1,157,432.75) produced the lowest total revenue, reflecting its significantly smaller trip count (42,060 trips).

2. Average Revenue Per Trip
    - Sunny ($29.31) had the highest average revenue per trip, followed closely by Rain ($28.52).
    - Clear ($27.90) and Snow ($27.52) had slightly lower averages, though the differences were minimal, indicating consistent fare structures across weather types.
3. Trip Count
    - Sunny and Rain had the most trips, with 792,567 and 570,563 trips, respectively.
    - Snow had the fewest trips (42,060), likely due to limited taxi availability or reduced passenger demand during snowy conditions.

CHAPTER 4 - METHODOLOGY: FEATURE ENGINEERING

New features were created from the original dataset to enhance the analysis and capture patterns in taxi demand. These features provide insights into timing, fare details, and pickup characteristics crucial for understanding taxi usage. Below is a summary of each engineered feature and the method used to calculate it:

1. **Rush Hour**:
    - We defined rush hours, typically 7-9 AM and 4-6 PM on weekdays, to capture high-demand periods. A binary variable was created (1 for rush hour, 0 for non-rush hour) to distinguish these times from regular hours.
2. **Pickup Hour**:
    - Extracted from the pickup timestamp (tpep pickup datetime), this feature indicates the exact hour of the trip, ranging from 0 (midnight) to 23 (11 PM), facilitating hourly demand analysis.
3. **Taxi Fare (Tip - Total Amount)**:
    - This feature represents the fare excluding the tip, calculated by subtracting the tip amount from the total fare. It provides a clearer view of the base fare and other charges.
4. **Pickup Day Name**:
    - Derived from the pickup timestamp, this categorical variable indicates the day of the week (e.g., Monday, Tuesday), allowing for analysis of daily demand patterns.
5. **Pickup Day**:
    - The numerical day of the month was extracted to capture the specific calendar day for each trip, which is useful for identifying trends on certain dates.

6. **Pickup Counts**:
    - Daily pickup counts were calculated by aggregating trips for each day. This provides a daily demand overview, highlighting peak days and monthly patterns.

	tpep_pickup_datetime	pickup_hour	pickup_day_name	pickup_day_name	pickup_date	rush_hour	taxi_fare
0	2024-01-01 00:03:00	0	Monday	1	1	No	15.0
1	2024-01-01 00:17:06	0	Monday	1	1	No	28.3
2	2024-01-01 00:36:38	0	Monday	1	1	No	15.0
3	2024-01-01 00:46:51	0	Monday	1	1	No	12.9
4	2024-01-01 00:54:08	0	Monday	1	1	No	34.6

Figure 23: Taxi Weather Combined Dataset

CHAPTER 5 - METHODOLOGY: MODELING

Correlation Matrix

The correlation matrix below is created to examine the relationships between features and taxi fares. The heatmap visually displays the strength of these correlations, with values from -1 to 1 indicating positive or negative relationships.

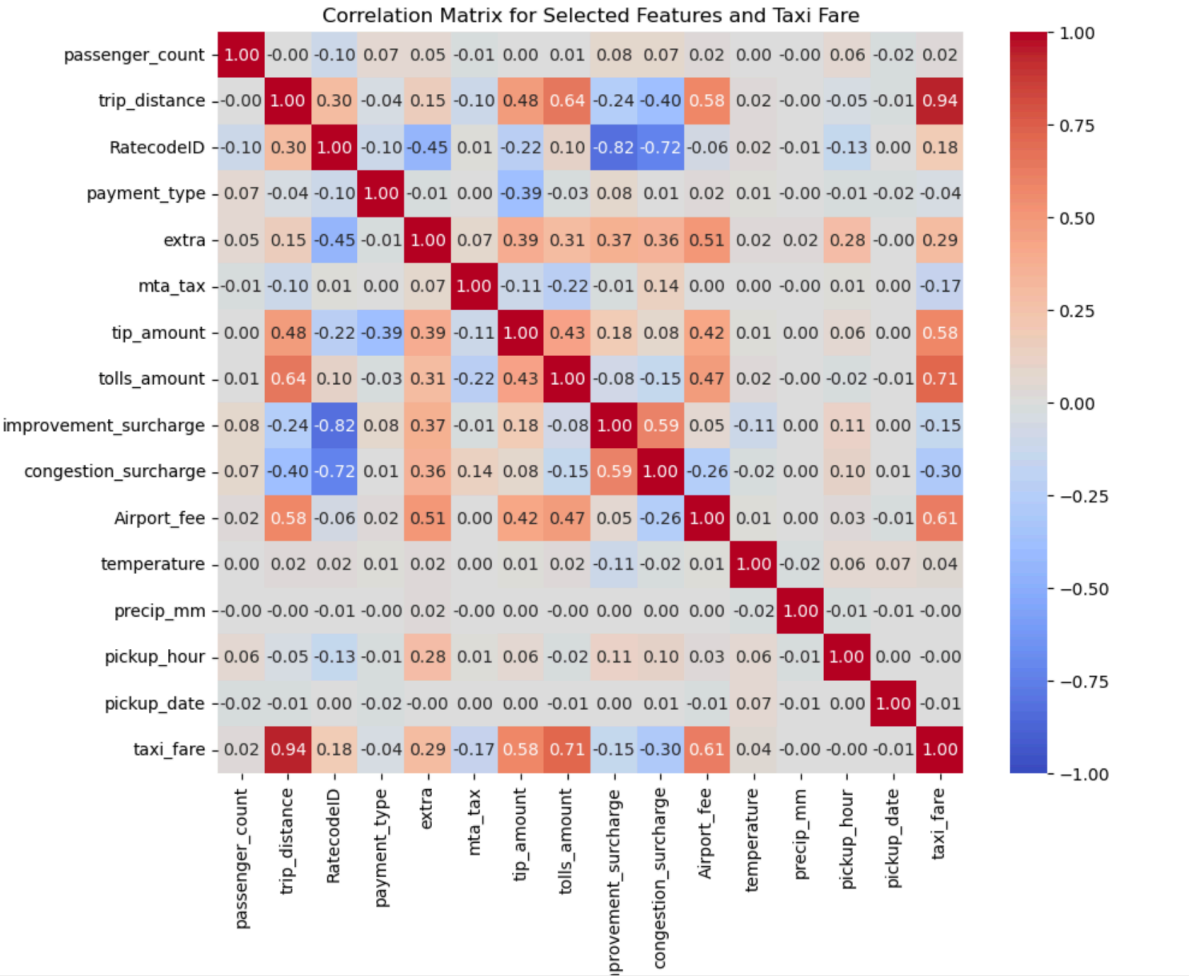

Figure 24: Correlation Matrix for Taxi Fare

The correlation matrix reveals that trip distance (0.94) and toll amount (0.71) have strong positive correlations with taxi fare, while tip amount (0.58) shows a moderate correlation. Other features like passenger count (0.18) and temperature (-0.02) exhibit weaker relationships, highlighting their minimal influence on taxi fare.

Linear Regression

Modeling Technique

Linear Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship and uses the equation $y=\beta0 + \beta1\ x1+ \beta2\ x2 +...+ \beta x\ xn + \epsilon$. where y is the predicted value of the dependent variable, $\beta0$ is the intercept, $\beta1$ represents the coefficients for each predictor x1, and ϵ is the error term.

In our analysis, the Ordinary Least Squares (OLS) method estimates the coefficients by minimizing the sum of squared residuals between the actual and predicted values. The OLS method finds the best-fitting line by ensuring that the differences between the observed values and the line are as small as possible.

Model Construction

Feature Selection and Preparation:

- Numerical Features: trip distance, passenger count, RatecodeID, PULocationID, pickup hour, temperature, precipitation in mm.
- Categorical Features: rush hour, PULocation Borough, pickup day name.
- Target Variable: taxi fare.

Consolidating Weather Data:

- A new categorical variable, weather type, was created by combining individual weather indicators (Rain, Snow, Sunny, and Clear) through a function that assigned a single condition.

- The original binary weather columns were then removed.

One-Hot Encoding:

- One-hot encoding was applied to categorical features (rush hour, PULocation Borough, pickup day name, and weather type) to convert each category into individual binary columns.
- One category per feature was dropped to prevent multicollinearity and enhance interpretability.

Model Training:

- The dataset was split into training and test sets with an 80-20 ratio, resulting in 1,525,190 rows in X_train.
- A constant term was added to both sets to account for the intercept in the Linear Regression model, and the model was fitted accordingly.

Model Results and Interpretation

The OLS summary is shown below in Fig 24

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              taxi_fare   R-squared:                       0.905
Model:                            OLS   Adj. R-squared:                  0.905
Method:                 Least Squares   F-statistic:                 6.935e+05
Date:                Sun, 03 Nov 2024   Prob (F-statistic):               0.00
Time:                        12:55:58   Log-Likelihood:            -4.5921e+06
No. Observations:             1525190   AIC:                         9.184e+06
Df Residuals:                 1525168   BIC:                         9.184e+06
Df Model:                          21
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            8.8195      0.059    150.090      0.000       8.704       8.935
trip_distance                    3.9678      0.001   2796.944      0.000       3.965       3.971
passenger_count                  0.3818      0.007     55.568      0.000       0.368       0.395
RatecodeID                      -0.0820      0.000   -317.431      0.000      -0.082      -0.081
PULocationID                    -0.0013   6.34e-05    -20.595      0.000      -0.001      -0.001
DOLocationID                    -0.0021   5.86e-05    -35.163      0.000      -0.002      -0.002
pickup_hour                      0.0737      0.001    100.066      0.000       0.072       0.075
temperature                      0.0083      0.000     24.367      0.000       0.008       0.009
precip_mm                       -0.1297      0.008    -16.962      0.000      -0.145      -0.115
rush_hour_Yes                    1.0368      0.009    121.315      0.000       1.020       1.054
PULocation_Borough_Brooklyn      2.4448      0.053     46.003      0.000       2.341       2.549
PULocation_Borough_Manhattan     0.1675      0.052      3.205      0.001       0.065       0.270
PULocation_Borough_Queens        3.5197      0.053     66.170      0.000       3.415       3.624
pickup_day_name_Monday          -0.3975      0.015    -26.280      0.000      -0.427      -0.368
pickup_day_name_Saturday        -0.8501      0.015    -54.929      0.000      -0.880      -0.820
pickup_day_name_Sunday          -1.3329      0.016    -81.974      0.000      -1.365      -1.301
pickup_day_name_Thursday         0.4743      0.015     32.574      0.000       0.446       0.503
pickup_day_name_Tuesday          0.3285      0.014     22.663      0.000       0.300       0.357
pickup_day_name_Wednesday        0.4761      0.014     33.673      0.000       0.448       0.504
weather_Rain                     0.4054      0.014     28.728      0.000       0.378       0.433
weather_Snow                     0.4923      0.028     17.281      0.000       0.436       0.548
weather_Sunny                    0.3568      0.016     21.993      0.000       0.325       0.389
==============================================================================
Omnibus:                   858582.824   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         41302953.526
Skew:                           2.018   Prob(JB):                         0.00
Kurtosis:                      28.172   Cond. No.                     6.35e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.35e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 25: OLS Regression result

Model Performance Metrics

- R-squared (0.905): The model explains 90.5% of the variance in taxi fares.
- Adjusted R-squared (0.905): The close alignment with R-squared confirms the stability of the model's explanatory power, even with multiple predictors.
- F-statistic: The low p-value of the F-statistic demonstrates that the model is statistically significant.

Feature Coefficients and Interpretations

- Trip Distance (3.9678): The fare increases by approximately $3.97 for each additional mile.
- Passenger Count (0.3818): Each additional passenger adds around $0.38 to the fare.

- RatecodeID (-0.0820): Specific rate codes are associated with a slight reduction in fare, as indicated by the negative coefficient.
- Pickup Hour (0.0737): Each additional hour is associated with a $0.07 increase in fare.
- Temperature (0.0083): Higher temperatures correspond to a small increase in fare.
- Precipitation (precip_mm, -0.1297): Increased precipitation is linked with a slight decrease in fare.

Impact of Rush Hour and Weather Conditions

- Rush Hour (1.0368): Fares increase by approximately $1.04 during rush hours.
- Weather Conditions:
    - Rain (0.4054): Rain conditions are associated with an increase in fare of around $0.41.
    - Snow (0.4923): Snow conditions correspond to a fare increase of about $0.49.
    - Sunny (0.3568): Sunny conditions are linked to an increase of approximately $0.36 in fare.

Day of the Week

- Weekend Days (Monday, Saturday, Sunday): Fares tend to be lower on these days.
- Mid-Weekdays (Tuesday, Wednesday, Thursday): These days show slightly higher fare amounts.

Model Evaluation

For evaluating the model, we used the test data, which consisted of 20% of the dataset. After training the Linear Regression model on the training data, we generated predictions on this test data and compared them to the actual taxi fares to assess model performance.

- R-squared (0.904): The model explains 90.4% of the variance in taxi fares, indicating a strong fit.
- Root Mean Squared Error (RMSE: 4.93): On average, the predicted fares deviate from actual fares by about $4.93.
- Mean Absolute Error (MAE: 3.07): The average absolute difference between predicted and actual fares is $3.07.

Random Forest

Modeling Technique

Random Forest is an ensemble machine learning method that constructs multiple decision trees during training and combines their predictions to improve accuracy. For regression tasks, the final prediction is the average of individual tree predictions. This approach helps to reduce overfitting and enhance generalization, making it effective for complex datasets, even when the data size is large.

Model Construction

We used RandomForestRegressor from sklearn. Ensemble and choose the following features:

- Numerical Features: trip distance, passenger count, RatecodeID, PULocationID, pickup hour, temperature, precipitation in mm.
- Categorical Features (OneHotEncoded): rush hour, PULocation Borough, pickup day name, and weather type
- Target Variable: taxi fare

The model was configured with n_estimators=20, meaning it constructs 20 decision trees and max_depth=10 to control tree complexity. Using 20 trees strikes a balance between

prediction accuracy and computational efficiency, suitable for larger datasets where too many trees can increase training time significantly.

The model was trained using the same training data split as the Linear Regression model, enabling direct performance comparisons between the two approaches.

Model Evaluation

For evaluating the model, we used the test data, which consisted of 20% of the dataset (381,298 rows set aside during the train-test split). After training the Random Forest model on the training data, we generated predictions on this test data and compared them to the actual taxi fares to assess model performance.

Performance Metrics:

- Root Mean Squared Error (RMSE): 3.61, indicating the average deviation of the predicted taxi fares from the actual fares in the test set.
- Mean Absolute Error (MAE): 2.29, showing the average absolute difference between predicted and actual fares.
- R-squared (R^2): 0.948, suggesting that 94.8% of the variance in taxi fares is explained by the Random Forest model.

Feature Importance:

- The feature with the highest importance is trip distance (0.948), making it the primary factor in predicting taxi fares.
- RatecodeID (0.033) and DOLocationID (0.0055) also contribute a little, reflecting the significance of trip-related factors.
- Time-related variables like pickup hour and rush hour Yes show moderate importance.

- Weather conditions (e.g., weather_Snow, weather_Rain, weather_Sunny) and day-of-week indicators have low importance values, indicating a smaller influence on fare prediction compared to trip distance and location factors. The below Fig 23show the importance of the feature

	Feature	Importance
0	trip_distance	0.948041
2	RatecodeID	0.033374
4	DOLocationID	0.005513
5	pickup_hour	0.005234
3	PULocationID	0.003342
11	PULocation_Borough_Queens	0.001411
14	pickup_day_name_Sunday	0.001069
13	pickup_day_name_Saturday	0.000739
8	rush_hour_Yes	0.000482
6	temperature	0.000298
10	PULocation_Borough_Manhattan	0.000169
7	precip_mm	0.000097
9	PULocation_Borough_Brooklyn	0.000051
1	passenger_count	0.000049
12	pickup_day_name_Monday	0.000040
17	pickup_day_name_Wednesday	0.000027
15	pickup_day_name_Thursday	0.000020
16	pickup_day_name_Tuesday	0.000014
19	weather_Snow	0.000013
18	weather_Rain	0.000009
20	weather_Sunny	0.000008

Figure 26: Feature Importance table

Visualization of a single Tree from Random Forest

This single tree represents one of many in the Random Forest, where multiple trees are averaged to improve accuracy.

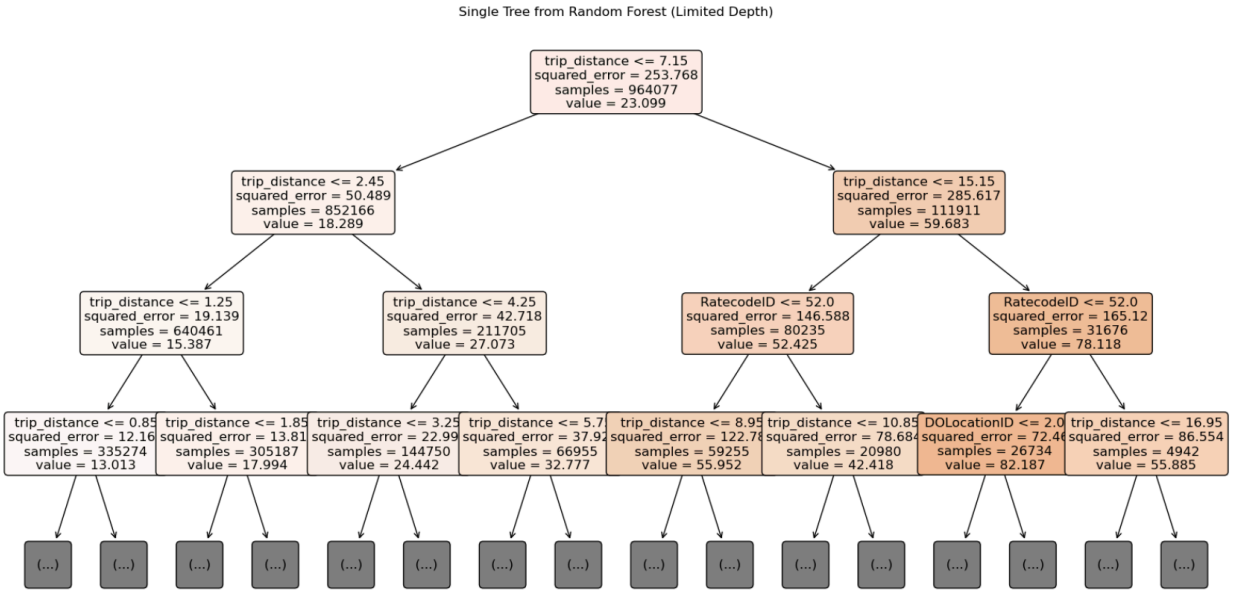

Figure 27: Single Tree from Random Forest

Model Comparison:

Random Forest and Linear Regression showed that yellow taxi fares can be predicted using trip characteristics and weather conditions.

The Linear Regression model gave an R-squared value of 0.904, which means that the model can explain 90.4% of the variance in taxi fares. The model's Root Mean Squared Error (RMSE) was 4.93, suggesting an average deviation of approximately $4.93 from actual fares. This model showed it utilized trip characteristics and weather conditions, such as temperature, precipitation, and specific weather events (rain and snow), providing valuable insights into how external factors influence fare predictions.

The random Forest model highlighted the importance of trip distance as an important predictor of taxi fare, with a feature importance score of 0.948. Other contributing features included RatecodeID and DOLocationID. It shows it only depended on Trip characteristics, and weather

conditions didn't play a significant role this shows Random Forest model did not effectively incorporate weather conditions compared to the Linear Regression model.

Cross-validation results for the Linear Regression model yielded RMSE scores of around 4.93 on average, showing consistent predictive performance across different subsets of data.

Both models are effective, we choose the Linear Regression model in this scenario for predicting yellow taxi fares as it shows some feature importance for both trip characteristics and weather conditions.

The model explains 90.5% of fare variance due to the inclusion of trip distance, which is directly tied to fare calculations. However, the objective was not only to predict fares accurately but also to explore the potential influence of weather conditions on fares. The analysis revealed that while weather does show some correlation with fare adjustments, its overall impact on fare prediction is minimal. This indicates that weather conditions contribute less significantly to fare variance compared to distance or other trip characteristics.

We are using Random Forest to answer the question**, can we predict if a taxi passenger will tip 20% more or less based on fare amounts, and what factors most influence tipping behavior?**

Model Construction

This section outlines the methodology for constructing a predictive model to estimate taxi tips based on weather and Trip features.

Feature Engineering

New features were created, including:

1. Rush Hour Interaction: This term refers to taxi pickups that occur during peak traffic times when demand for rides is typically higher.
2. Weather-Day Interactions: This involves examining how rain and snow affect taxi demand on specific days, especially on Mondays.
3. Fare Per Mile: This is calculated to evaluate how the fare charged per mile influences tipping behavior.
4. tip_20_percent feature is a binary variable derived from the total fare and the tip amount for each taxi ride. It is calculated as follows:

    - If the tip amount is greater than or equal to 20% of the total fare, the value of tip_20_percent is set to 1, indicating that the passenger tipped 20% or more.

    - If the tip amount is less than 20% of the total fare, the value of tip_20_percent is set to 0, indicating that the passenger tipped less than 20%.

Features

The features (X) include numerical variables such as trip distance, passenger count, pickup hour, temperature, and precipitation, as well as categorical variables like pickup location and interaction terms related to weather and rush hour and above-created columns, while the target variable (y) is defined as tip_20_percent, and the data is divided in 80% training and 20% testing.

Two Methods for Model Construction

The analysis utilized two methods for constructing the predictive model for tipping behavior.

Random Forest Classifier: The first method uses the Random Forest model to classify tipping behavior based on the defined features. This model was initialized with n_estimators=100, indicating that 100 decision trees will be created, max_depth=10, which limits the depth of each tree to prevent overfitting, and random_state=42, ensuring consistent results across different runs**.**

Random Forest with SMOTE and Standard Scaler: The Random Forest model uses the Synthetic Minority Over-sampling Technique (SMOTE) to address the class imbalance in the target variable, along with a Standard Scaler to standardize the numerical features. Like the first method, this model was initialized with the same parameters (n_estimators=100, max_depth=10, random_state=42).

Comparison of Random Forest Models
Evaluation matrix

Metric	Without SMOTE and StandardScaler	With SMOTE and StandardScaler
Confusion Matrix		
True Negatives (0)	278,369	237,832
False Positives	95	40,271
True Positives (1)	406	237,111
False Negatives	102,428	41,477
Precision for Tips ≥ 20%	0.81	0.85
Recall for Tips ≥ 20%	0.00	0.85
F1-Score for Tips ≥ 20%	0.01	0.85
Precision for Tips < 20%	0.73	0.85
Recall for Tips < 20%	1.00	0.86
F1-Score for Tips < 20%	0.84	0.85
Accuracy	0.73	0.85
ROC AUC	0.69	0.92

Table 9: Random Forest Model Comparisons

Feature Importance

Feature	Importance (Without SMOTE)	Importance (With SMOTE)
Top Feature 1	Fare Per Mile: 60.09%	Fare Per Mile: 34.45%
Top Feature 2	Trip Distance: 20.65%	Trip Distance: 24.93%
Top Feature 3	Passenger Count: 8.02%	Temperature: 14.07%
Bottom Feature 1	PULocationID: 4.47%	Weather Type - Snow: 0.06%
Bottom Feature 2	Weather Type - Rain: 0.19%	Weather Type - Rain: 0.19%

Table 10: Feature Importance

The evaluation of our model's performance using SMOTE and StandardScaler showed improvements over the model with these techniques. The accuracy increased to 0.85, and the ROC AUC rose to 0.92. While the recall and F1-Score for tips greater than or equal to 20% improved from 0.00 to 0.85, the model still did not fully identify all such instances. The feature importance analysis revealed that Fare Per Mile remained significant but decreased from 60.09% to 34.45%, while Temperature gained importance. However, there is still room for improvement, suggesting the need for further feature engineering and potentially additional data to refine our model further.

CHAPTER 6 – CONCLUSION

Discussion

The research questions were answered by the analysis.

1. How do trip distance ranges relate to adverse weather conditions, and do taxis show a preference for specific distance ranges during these conditions?

   The analysis revealed that shorter trip distances are more common during rain and snow weather conditions. For instance, 34% of rides during rainy weather fall within the 1-2 miles range, while snowy conditions show 36% in the same range. Additionally, 27% of rainy trips and 25% of snowy trips are within the 2-5 miles range. This indicates that people tend to take shorter trips in poor weather, as both drivers and passengers prefer to avoid long distances during rainy or snowy days. Overall, the findings suggest a clear preference for trips within 1 to 5 miles when faced with rain and snow weather conditions.

2. Can yellow taxi fares be predicted based on trip characteristics and weather conditions?

   Yes, yellow taxi fares can be predicted based on trip characteristics and weather conditions. The analysis revealed that the key trip characteristics that influence taxi fare are trip distance, passenger count, and pickup location. Specifically, trip distance was found to be the most significant factor, with an increase in fare of approximately $3.97 for each additional mile traveled. Additionally, each extra passenger contributes around $0.38 to the total fare.

   Incorporating weather conditions into the model demonstrated that extreme weather impacts taxi fares as well. The coefficients indicated fare increases of approximately $0.41 for rainy conditions, $0.49 for snowy conditions, and $0.36 for sunny days. This means that even

though taxi meters don't change prices based on the weather, higher demand during bad weather, longer trips because of traffic, and negotiated prices can result in higher overall fares.

The model explains 90.5% of fare variance due to the inclusion of trip distance, which is directly tied to fare calculations. However, the objective was not only to predict fares accurately but also to explore the potential influence of weather conditions on fares. The analysis revealed that while weather does show some correlation with fare adjustments, its overall impact on fare prediction is minimal. This indicates that weather conditions contribute less significantly to fare variance compared to distance or other trip characteristics.

Overall, the model successfully predicts taxi fares by accounting for both trip characteristics and weather conditions. However, to fully confirm these findings, more comprehensive data would be beneficial, particularly in varying weather conditions over a longer period and with additional trip characteristics.

3. How do pick-up counts over the three months correlate with weather conditions?

   The analysis showed that taxi pickups were slightly higher during rainy weather, with a correlation of 0.15 in January. Snowy days had little impact, with a correlation of 0.10. Clear days were preferred, showing a correlation of 0.21. In May, the overall demand decreased with a correlation of -0.36, but rainy days still had a positive effect with a correlation of 0.22. In July, rain significantly increased pickups, with a strong correlation of 0.42, while sunny weather contributed less, with a correlation of 0.14.

   Overall, the findings illustrate that while rain significantly boosts taxi pickups, snow, and

sunny conditions have varying effects, with snow showing minimal influence and sunny days contributing slightly to increased demand. Further data would be beneficial to draw more definitive conclusions regarding these relationships.

4. Can we predict if a taxi passenger will tip 20% more or less based on fare amounts, and what factors most influence tipping behavior?

   We were able to predict whether a taxi passenger would tip 20% or more based on fare amounts and other influencing factors. Our analysis revealed that features such as fare per mile, trip distance, and temperature of the day influence tipping behavior. However, the model did not fully capture all instances of passengers tipping 20% or more, primarily due to the limited amount of data available for this tipping category. With a larger dataset, particularly one that includes more examples of higher tipping behavior, we could improve the model's predictive accuracy and better understand the factors that drive tipping decisions. This suggests that expanding the dataset could lead to more robust insights and improve the ability to predict tipping behavior.

5. EDA Insights

- Revenue and Trip Counts by Weather Type

  Total Revenue: Sunny weather generated the highest total revenue ($23.23M) with 792,567 trips, while Snow had the lowest ($1.16M) from only 42,060 trips. This may be influenced by work-from-home policies during snow, reducing travel demand.

Average Revenue Per Trip: Sunny trips had the highest average revenue per trip ($29.31), followed by Rain ($28.52). Snow ($27.52) and Clear ($27.90) showed minimal variation, indicating consistent fare structures.

Trip Count**:** Sunny (792,567) and Rain (570,563) had the highest trip counts, while Snow had the fewest, likely due to reduced demand or taxi availability.

The differences in revenue across weather types can be attributed to passenger behavior and demand patterns. Sunny days generate the highest revenue due to a higher trip count, reflecting increased travel activity in favorable weather. In contrast, snowy conditions see significantly fewer trips, likely due to work-from-home policies, reduced passenger demand, or limited driver availability during extreme weather.

- Average Tipping Percentage by Weather Type

  Snow (20.18%) had the highest tipping percentage, reflecting greater generosity during challenging driving conditions.

  Rain (19.04%) also showed higher tipping, likely due to passenger appreciation of the service in adverse weather.

  Clear (15.5%) and Sunny (16.56%) tipping percentages were lower, indicating less perceived driver effort during favorable weather.

  The higher tipping percentages during Snow and Rain are likely due to passengers acknowledging the increased difficulty and effort drivers face in adverse weather conditions, such as poor visibility, icy roads, or heavy traffic.

Applications

The findings from this analysis provide important insights for the taxi and transportation industry. Taxi operators can consider implementing pricing strategies that adjust fares during bad weather. This could encourage more drivers to work when demand is high, making sure that passengers can find rides more easily. Also, understanding how trip distances relate to weather conditions can help taxi companies place their cars where they are needed most during storms or

heavy rain. Predicting fares based on trip characteristics and weather conditions can also lead to improved customer satisfaction through more accurate fare estimates.

In terms of the correlation between pickup counts and weather conditions, public transport authorities can adjust bus and train schedules during adverse weather to increase service frequency, helping passengers find alternative transportation when taxi availability decreases. Taxi services can also create targeted marketing campaigns based on weather, offering discounts to attract more customers. Taxi companies can develop strategies to enhance tipping by training drivers on factors that influence tips and encouraging practices that lead to higher customer satisfaction and increased gratuities. They can optimize fleet allocation for high-demand weather like sunny and rainy days to maximize revenue. During snowy conditions, they could offer incentives or promotions to encourage travel and increase trip count. Higher tipping during adverse weather suggests an opportunity to motivate drivers with bonuses or support to ensure service quality.

Limitations and Future Research

The analysis had limitations because it didn't capture enough ride information during snowy and rainy weather, as it was based on just three months of data. Even though this dataset had over 9 million rows, looking at a longer time frame with more months would take a lot of computing power. A regular laptop might struggle to run complex models on larger datasets efficiently. To manage the increased data volume and complexity, using a supercomputer or high-performance computing resources would be necessary. This would allow for deeper analysis without slowing down performance.

The dataset only included pickup information and did not have information on how many taxis were available at any given time. This missing information about taxi availability made it hard to

understand how supply demand is correlated with weather conditions. Future research would be beneficial if it has that information as it would help understand service levels and customers to get better insights on expected wait times. To evaluate how weather affects taxi operations, future research should include data over multiple years to observe changing trends over time. This approach would lead to a better understanding of how weather impacts taxi demand, fare rates, and tipping behavior. Additionally, considering other factors like local traffic events and economic changes would improve the analysis. By using better processing methods and resources, future studies can handle larger datasets more effectively and ensure accurate modeling and analysis.

REFERENCES

Abel Brodeur , Kerry Nield. (2018). An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC. *ScienceDirect*.

Ahmadreza Faghih-Imani, Sabreena Anowar, Eric J. Miller, Naveen Eluru. (2017). Hail a cab or ride a bike? A travel time comparison of taxi and bicycle-sharing systems in New York City. *sciencedirect*.

Backes, A. D. (2020). Retrieved from classicnewyorkhistory: **https://classicnewyorkhistory.com/history-of-new-yorks-yellow-taxi-cab/**

Bhawana Rathore, Pooja Sengupta, Baidyanath Biswas, Ajay Kumar . (2024). Predicting the price of taxicabs using Artificial Intelligence: A hybrid approach based on clustering and ordinal regression models. *Predicting the price of taxicabs using Artificial Intelligence: A hybrid approach based on clustering and ordinal regression models*.

Farber, H. S. (2005). Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers. *JSTOR*.

Government, N. (2022, Dec 19). Retrieved from NYC Taxi & Limousine Commission: **https://www.nyc.gov/assets/tlc/downloads/pdf/taxi_information.pdf**

Guo, Suiming and Chen, Chao and Wang, Jingyuan and Liu, Yaxiao and Xu, Ke and Zhang, Daqing and Chiu, Dah Ming. (2018). A Simple but Quantifiable Approach to Dynamic Price Prediction in Ride-on-demand Services Leveraging Multi-source Urban Data. *Association for Computing Machinery, New York*.

Jian Sun, He Dong, Guoyang Qin, Ye Tian. (2020). Quantifying the Impact of Rainfall on Taxi Hailing and Operation.

K. -F. Chu, A. Y. S. Lam and V. O. K. Li. (2020). Deep Multi-Scale Convolutional LSTM Network for Travel Demand and Origin-Destination Predictions. *IEEE Transactions on Intelligent Transportation Systems*.

M. Anil Yazici , Camille Kamga, Abhishek Singhal . (2016). Modeling taxi drivers' decisions for improving airport ground access: John F. Kennedy airport case. *Transportation Research Part A: Policy and Practice*.

PERRY, B. (2023). Why Are NYC Taxi Cabs Yellow?

R. C. P. Wong, P. L. (2021). Spatio-Temporal Influence of Extreme Weather on a Taxi Market. *Sage journal*.

Rong Chen, Lingjia Liu, Yongping Gao. (2024). The Association between Rainfall and Taxi Travel Activities: A Case Study from Wuhan, China. *Journal of Advanced Transportation*.

Ruijie Bian, Chester G. Wilmot, Ling Wang. (2019). Estimating spatio-temporal variations of taxi ridership Estimating spatio-temporal variations of taxi ridership caused by Hurricanes Irene and Sandy: A case study of New York City. *Transportation Research Part D*.

Tianyi Li, Guo-Jun Qi, Raphael Stern. (2021). Taxi Utilization Rate Maximization by Dynamic Demand Prediction: A Case Study in the City of Chicago. *Sage Journals*.

TLC. (2024). Retrieved from **https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/.**

Won Kyung Lee, So Young Sohn,. (2020). A large-scale data-based investigation on the relationship between bad weather and taxi tipping. *Journal of Environmental Psychology*.

Zhang Y, Sui X, Zhang S. (2024). Exploring spatio-temporal impact of COVID-19 on citywide taxi demand: A case study of New York City. *Google scholar*.

Zhang, M, Tian, D, Liu, J, Li, X. (2024). Analysis of taxi demand and traffic influencing factors in urban core area based on data field theory and GWR model: A case study of beijing. *ProQuest*.

Zhang, Wenbo; Le, Tho V; Ukkusuri, Satish V; Li, Ruimin. (2020). Influencing factors and heterogeneity in ridership of traditional and app-based taxi systems. *Google Scholar*.

Zhenhua Chen, Yongjian Yang, Liping Huang, En Wang, Dawei Li. (2018). Discovering Urban Traffic Congestion Propagation Patterns With Taxi Trajectory Data. *IEEE Access, 6*.

Zhizhen Liu, Hong Chen, Yan Li, Qi Zhang. (2020). Taxi Demand Prediction Based on a Combination Forecasting Model in Hotspots. *Journal of Advanced Transportation*.