

BY-ROBIN SHARMA

assignment on nyc taxi

Report: Optimising NYC Taxi Operations

Include your visualisations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

## Data Preparation

The first step involves gathering and loading the dataset into a structured format suitable for analysis. This includes importing necessary libraries, reading the dataset, and ensuring that the data is formatted correctly. The goal is to create a well-structured foundation before proceeding with data cleaning and analysis.

* 1. Loading the dataset

The dataset consists of NYC yellow taxi trip records for 2023, stored in Parquet format for efficient storage and retrieval. The data originates from technology providers such as vendors and taxi-hailing apps and is provided to the NYC Taxi and Limousine Commission (TLC) for regulatory and analytical purposes. To load and process the dataset, Python was used to read multiple monthly files and combine them into a single dataset for analysis. Google Drive was mounted in Google Collab to access the data seamlessly. Essential Python libraries— including pandas, NumPy, matplotlib, and seaborn—were imported to facilitate data processing and visualization. Warnings were suppressed to ensure a clean and uncluttered output. To maintain compatibility and reproducibility, the versions of key libraries were checked: NumPy (1.26.4), pandas (2.2.2), matplotlib (3.10.0), and seaborn (0.13.2).

* + 1. **Sample the data and combine the files**

The data preparation process begins by loading multiple monthly NYC taxi trip records stored in separate files. The goal is to create a manageable yet representative dataset for analysis. To achieve this, each file is read, and the tpep\_pickup\_datetime column is converted into a datetime format to facilitate time-based filtering. Instead of using the full dataset, a sampling strategy is implemented where 5% of trip records are selected from each hour of every day. This ensures that data is uniformly distributed across different time periods and avoids over-representation of peak hours or specific days. Once sampled, the data is progressively combined into a single Data Frame, ensuring that all months are included. Any errors encountered while reading individual files are handled to prevent data loss. Finally, the consolidated dataset is stored in Parquet format, which optimizes storage and retrieval efficiency. This methodology allows for a computationally efficient yet statistically meaningful analysis. By sampling evenly across all time periods, the dataset maintains the overall distribution of taxi rides without excessive data volume, enabling faster processing and analysis.

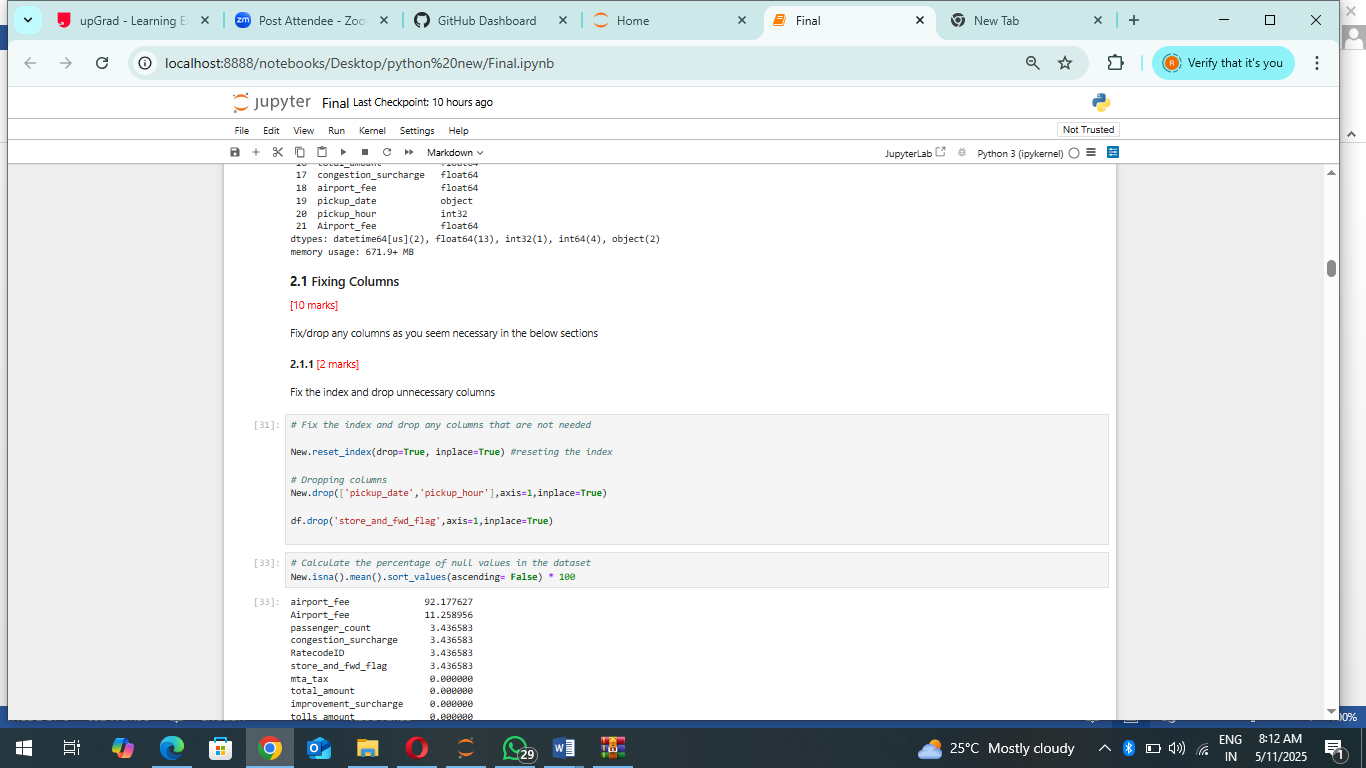
## Data Cleaning

To ensure accuracy and reliability in our findings, we perform a comprehensive data cleaning process. This step includes:

### Fixing Columns

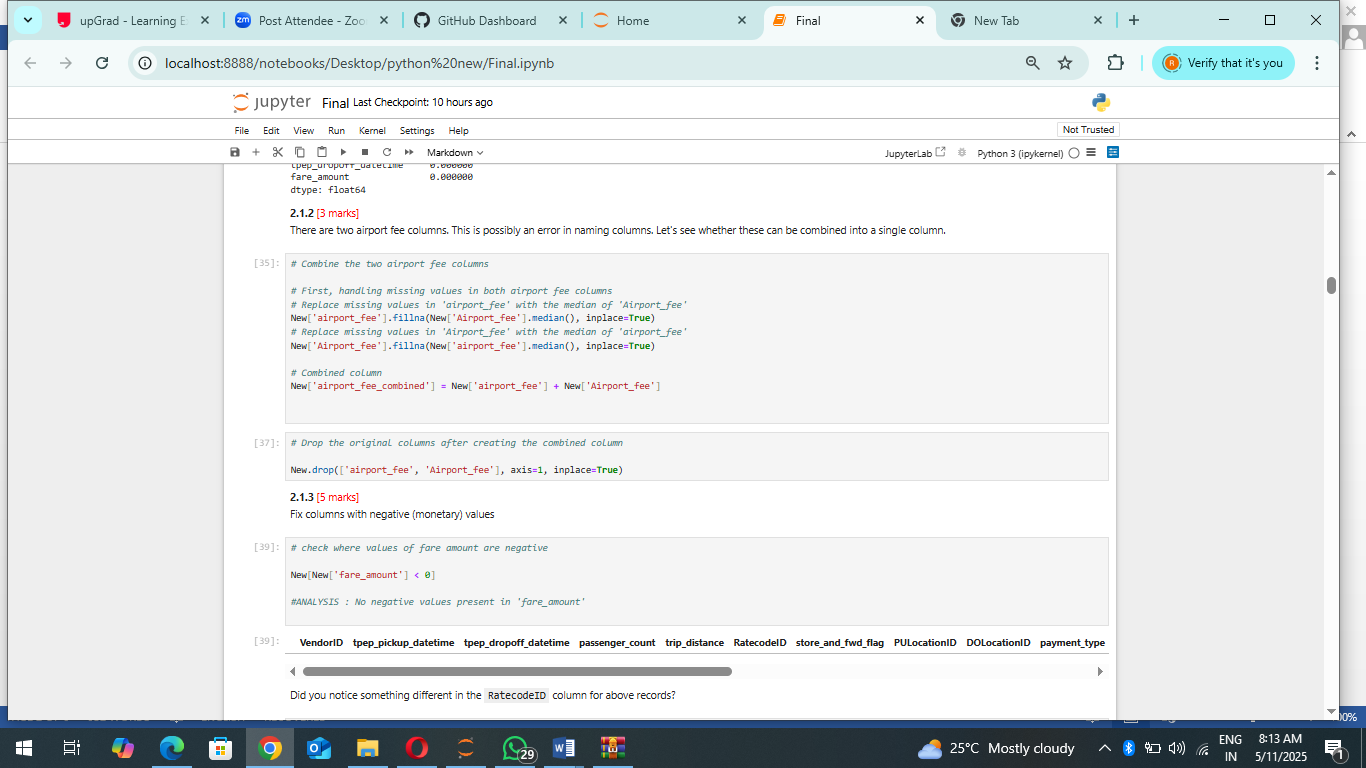
* + 1. **Fix the index**

The index was reset to ensure a clean, sequential ordering of rows, eliminating any inconsistencies from concatenated or sampled data. The reset\_index(drop=True, inplace=True) function was used to remove the old index and replace it with a default integer-based index. This step ensures that further analysis and visualizations remain structured and do not inherit any irregularities from previous indexing.



* + 1. **Combine the two airport\_fee columns**:

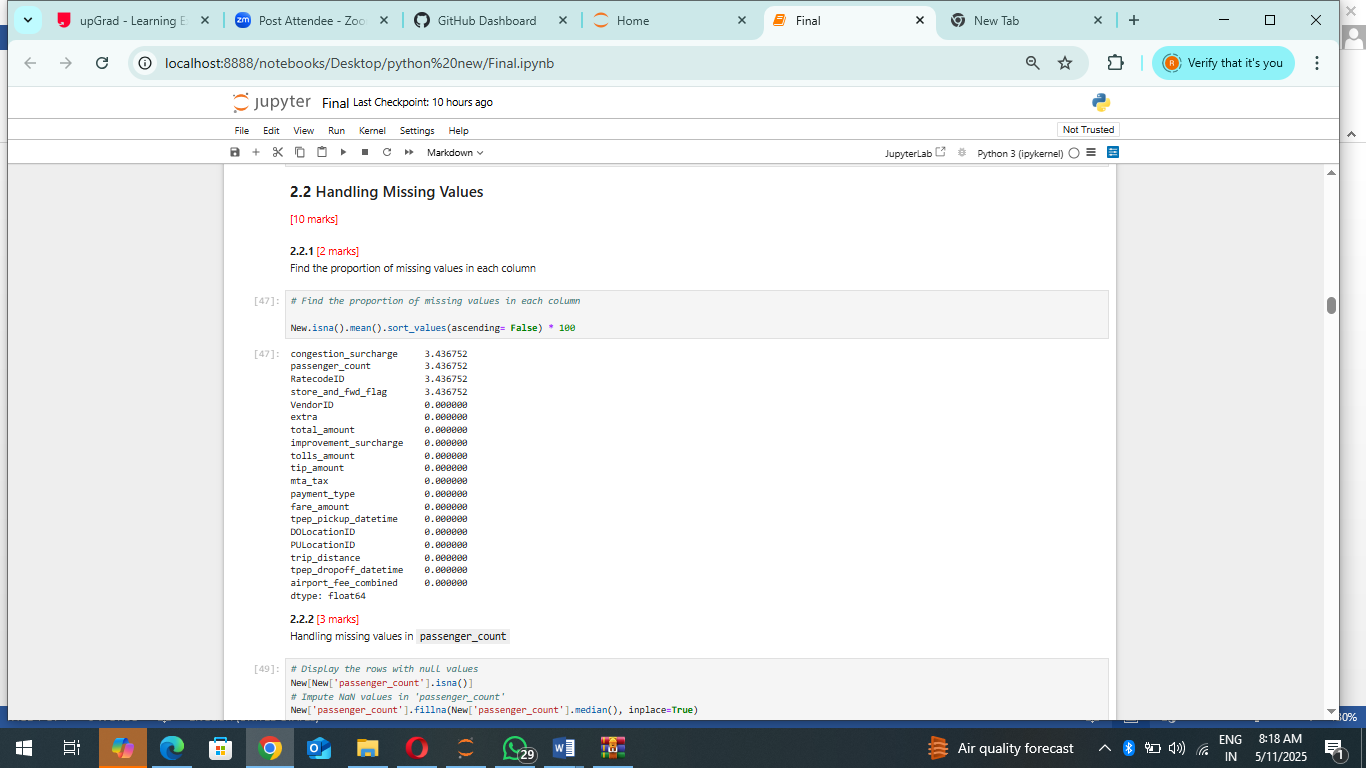
The dataset was examined for missing values, with the percentage of null values calculated for each column. This step helped identify any gaps in the data that needed to be addressed before further processing. A redundancy issue was found in the dataset with two separate columns for airport fees: airport\_fee and Airport\_fee. Instead of dropping rows with missing values, which could lead to data loss, the missing values in both columns were replaced with their respective median values. This approach ensures that imputation does not introduce extreme variations or skew the data. Once the missing values were handled, the two columns were combined by summing their values into a single airport\_fee column, preventing duplication and maintaining data integrity. Finally, the redundant



### Handling Missing Values

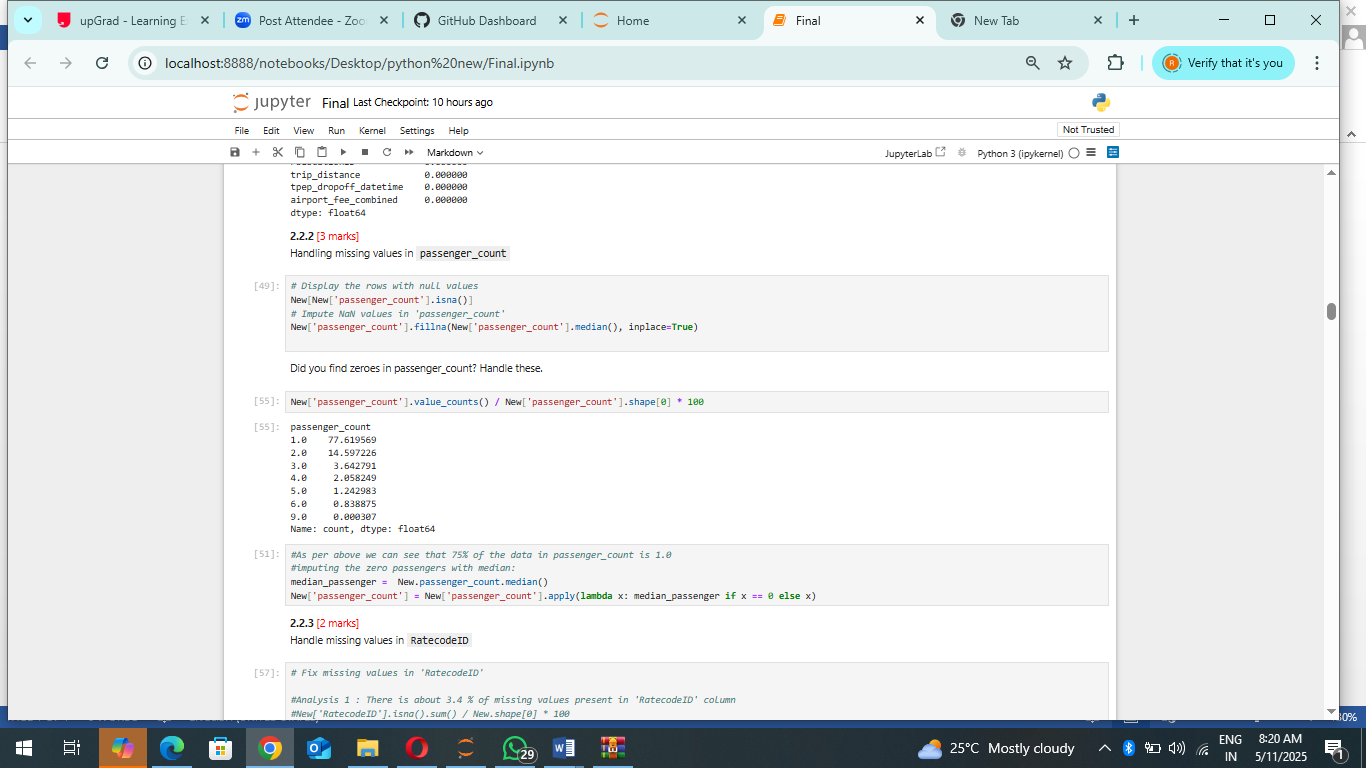
* + 1. **Find the proportion of missing values in each column**

To assess data completeness, the proportion of missing values in each column was calculated. This analysis helped identify any significant gaps in the dataset that could impact further processing. By sorting the missing values in descending order, columns with the highest percentage of null values were prioritized for data cleaning. Columns with a high proportion of missing data may require imputation or removal, depending on their importance in the analysis. If the missing values were negligible, they could be ignored, while critical fields might need median or mode imputation to retain data integrity. This step ensured that the dataset was ready for reliable insights and modelling



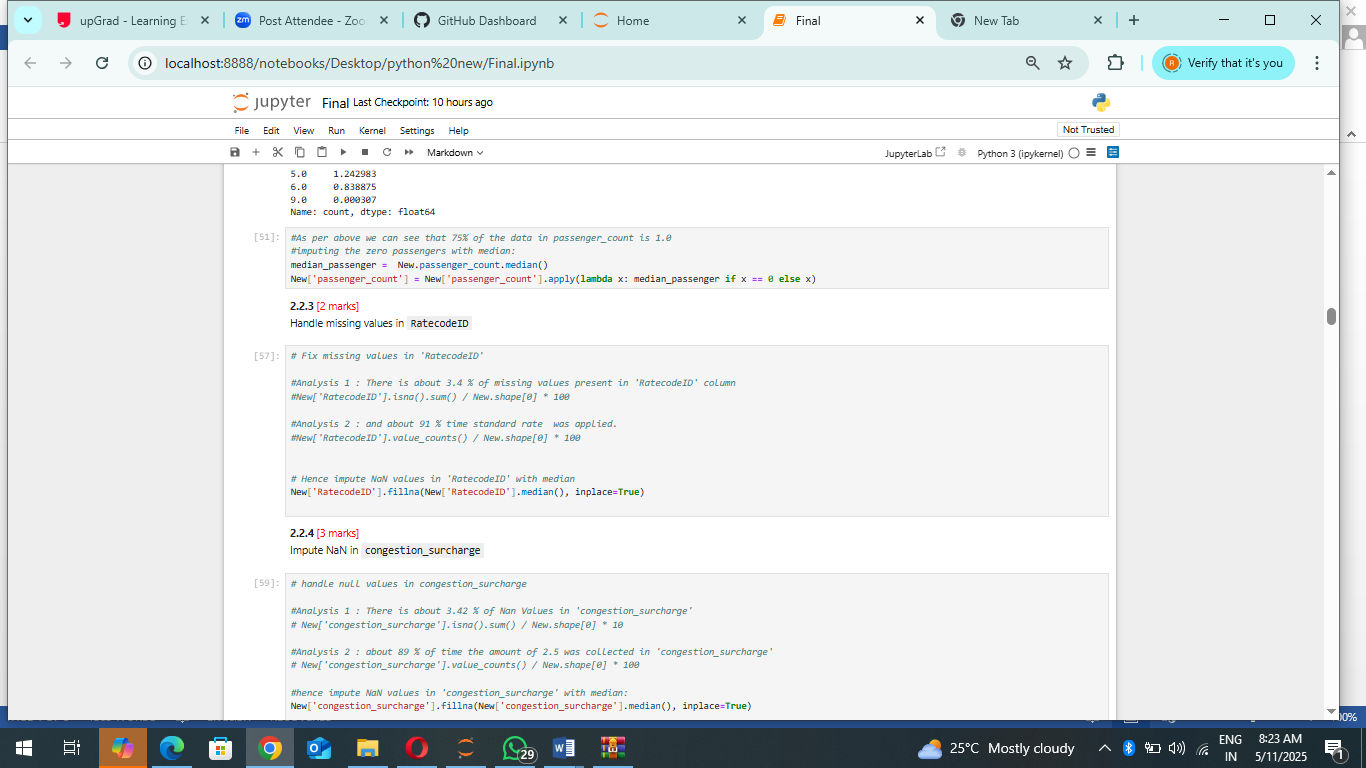
* + 1. **Handling missing values in passenger\_count:**

To address missing values in the dataset, rows with null values in the passenger\_count column were identified. The missing values were imputed using the median passenger count to maintain data consistency without introducing significant bias. An analysis revealed that approximately 1.5% of trips had recorded passenger\_count as zero, which is unlikely in a taxi trip scenario. To correct this anomaly, these zero values were also replaced with the median passenger count. This ensured that all records had a reasonable passenger count, improving data quality for further analysis.



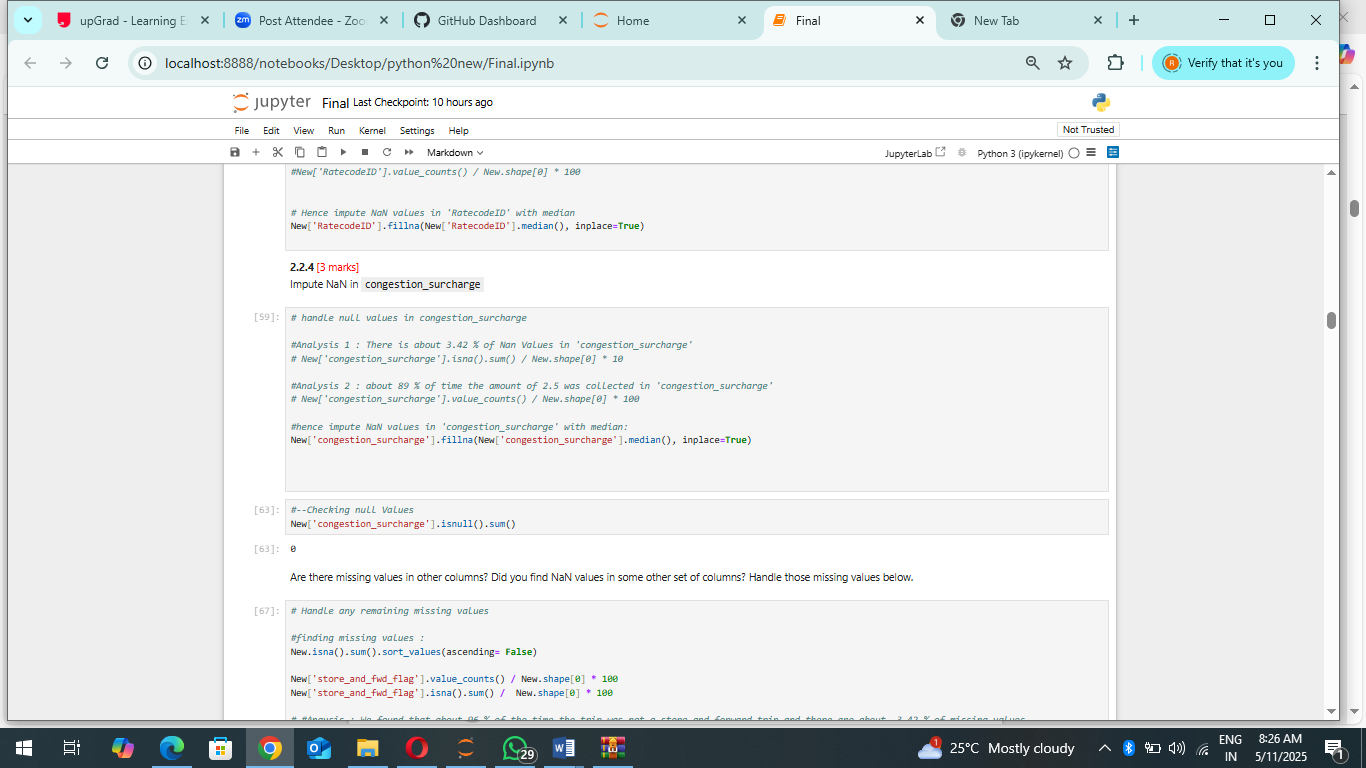
* + 1. **Handle missing values in RatecodeID:**

The RatecodeID column contained approximately 3.4% missing values. Analysis showed that about 91% of the trips had a "Standard Rate" applied, making it the most common rate type. Given this distribution, missing values in RatecodeID were imputed using the median value. This approach preserved the integrity of the dataset while minimizing any potential impact on downstream analysis.



* + 1. **Impute NaN in congestion\_surcharge**

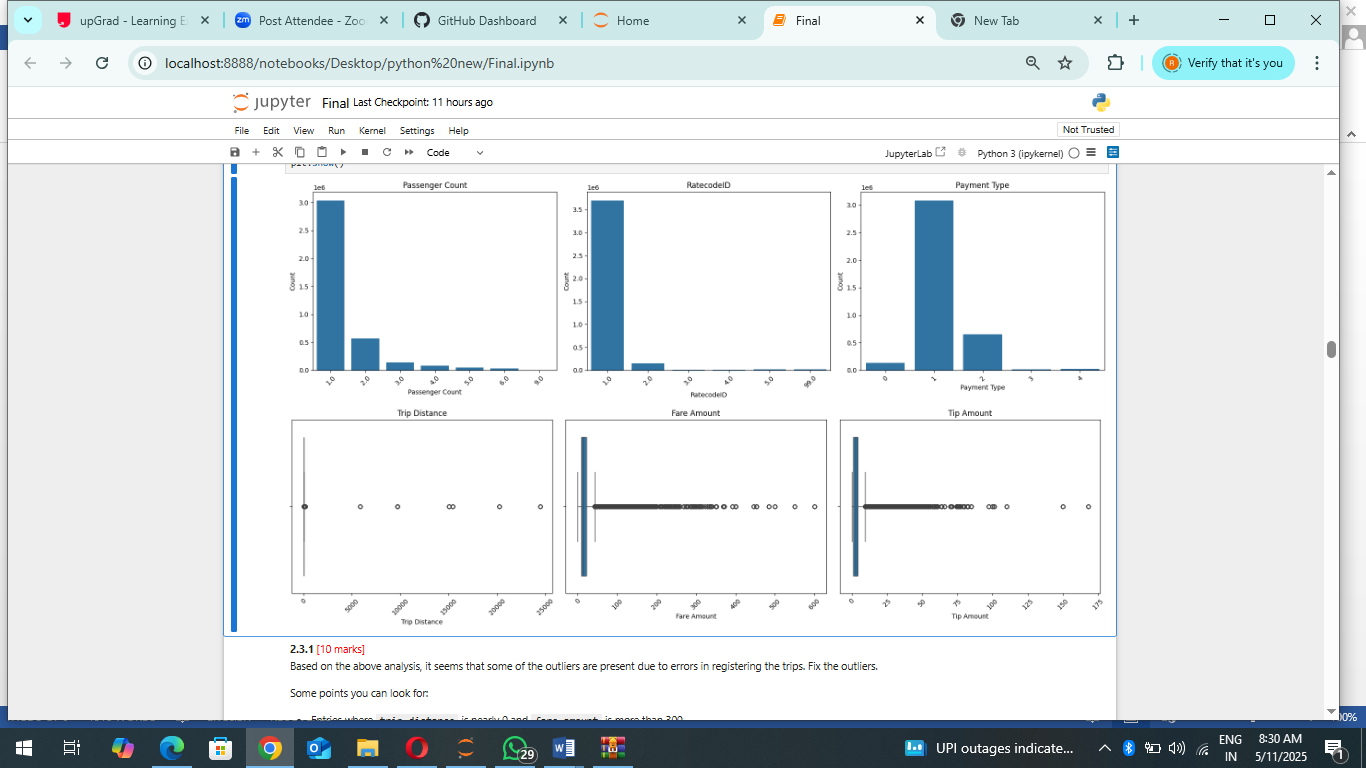
The congestion\_surcharge column had about 3.42% missing values. Analysis revealed that 89% of the time, the congestion surcharge was $2.50, making it the most common value. Thus, missing values were imputed using the median to maintain consistency.



### Handling Outliers and Standardising Values

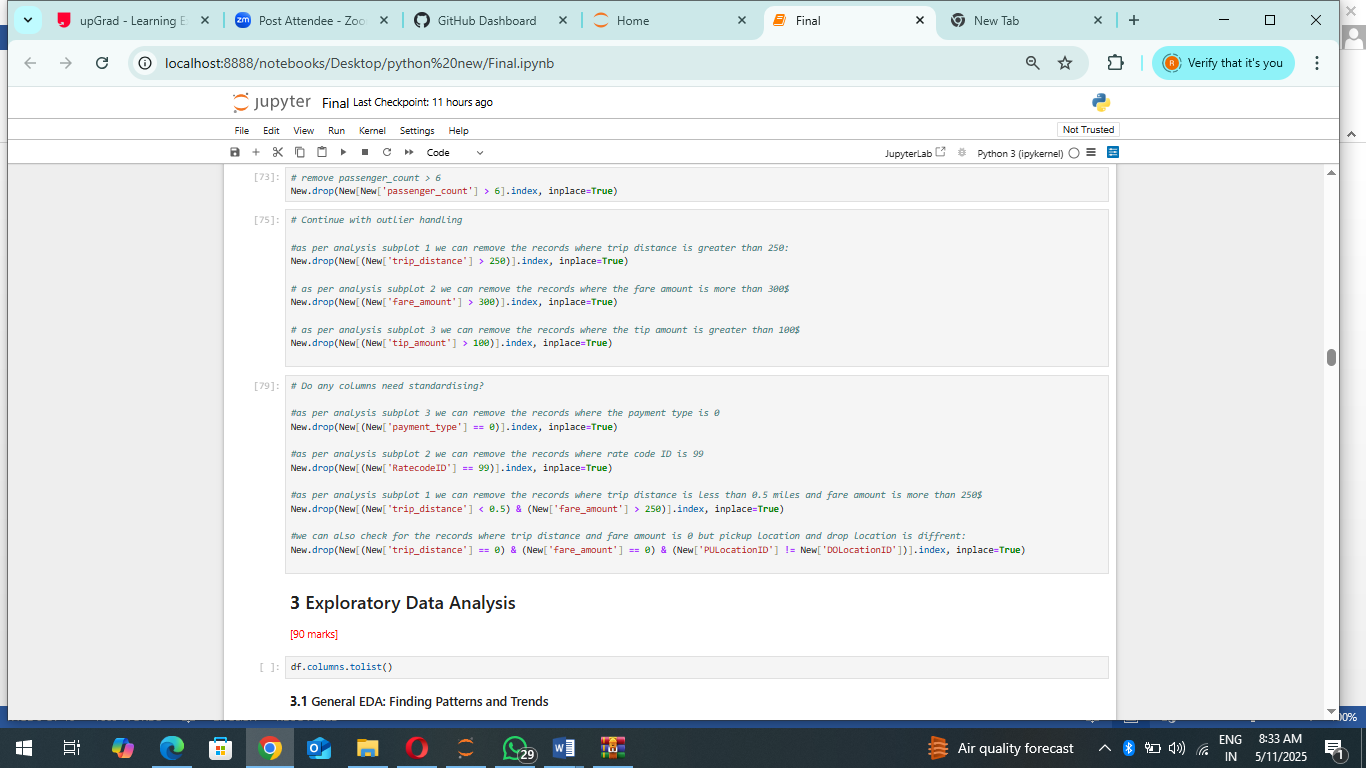
The dataset was explored for potential outliers by analysing key categorical and numerical variables through count plots and boxplots.

The count plots revealed that most taxi trips had six or fewer passengers, with very few trips exceeding this count. The RatecodeID column showed that the most common value was 1, but there were some instances of RatecodeID = 99, which is not a standard value and may need correction. Similarly, payment\_type = 1 (credit card) was the most frequent payment method, but records with payment\_type = 0 were observed, which could indicate erroneous data entries. The boxplots provided insights into numerical columns. The trip\_distance column displayed several extreme values, suggesting potential outliers in trip lengths. The fare\_amount column also contained some unusually high values that should be examined further. The tip\_amount column was mostly consistent, but a few extreme values stood out and needed verification to determine whether they were valid entries or anomalies. Overall, this analysis highlighted areas requiring further scrutiny, including cleaning non-standard values and addressing potential outliers in fare, tip, and distance metrics.



* + 1. **Check outliers in payment type, trip distance and tip amount columns**

To refine the dataset, records containing extreme or non-standard values were removed based on prior analysis. Trips with more than six passengers were dropped, as they were rare and likely outliers. Similarly, extreme values in numerical columns were addressed: trips with a distance greater than 250 miles, fare amounts exceeding $300, and tips over $100 were removed to ensure more realistic data. Further standardization was applied by removing records where payment\_type = 0, as this was identified as a non-standard value. Likewise, RatecodeID = 99 entries were dropped due to their irregularity. Additional filtering targeted trips where the distance was unrealistically low (less than 0.5 miles) while having a fare above $250. Lastly, trips where both trip\_distance and fare\_amount were zero, but pickup and drop-off locations differed, were eliminated to correct potential data inconsistencies. These steps helped improve data reliability by eliminating unlikely or erroneous records, ensuring more accurate insights in subsequent analysis.



## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

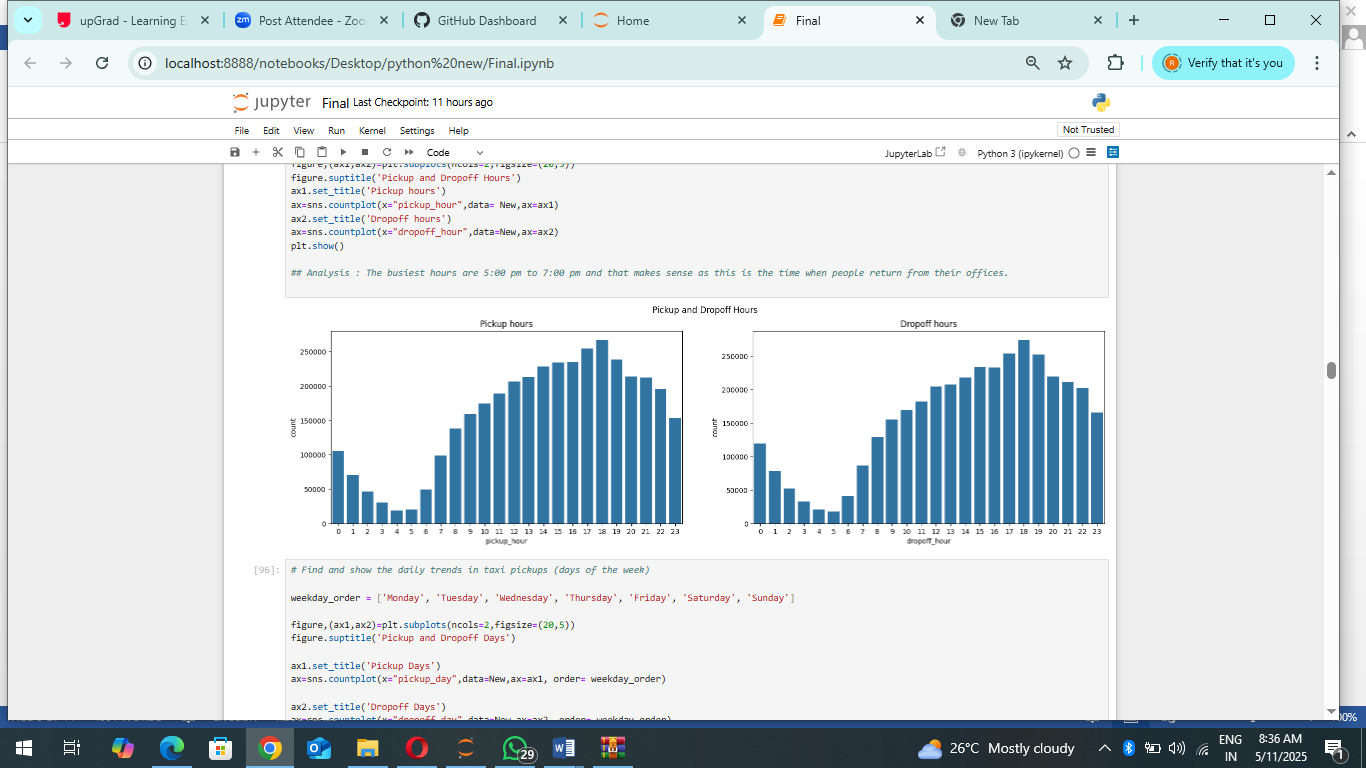
* + 1. **Classify variables into categorical and numerical**

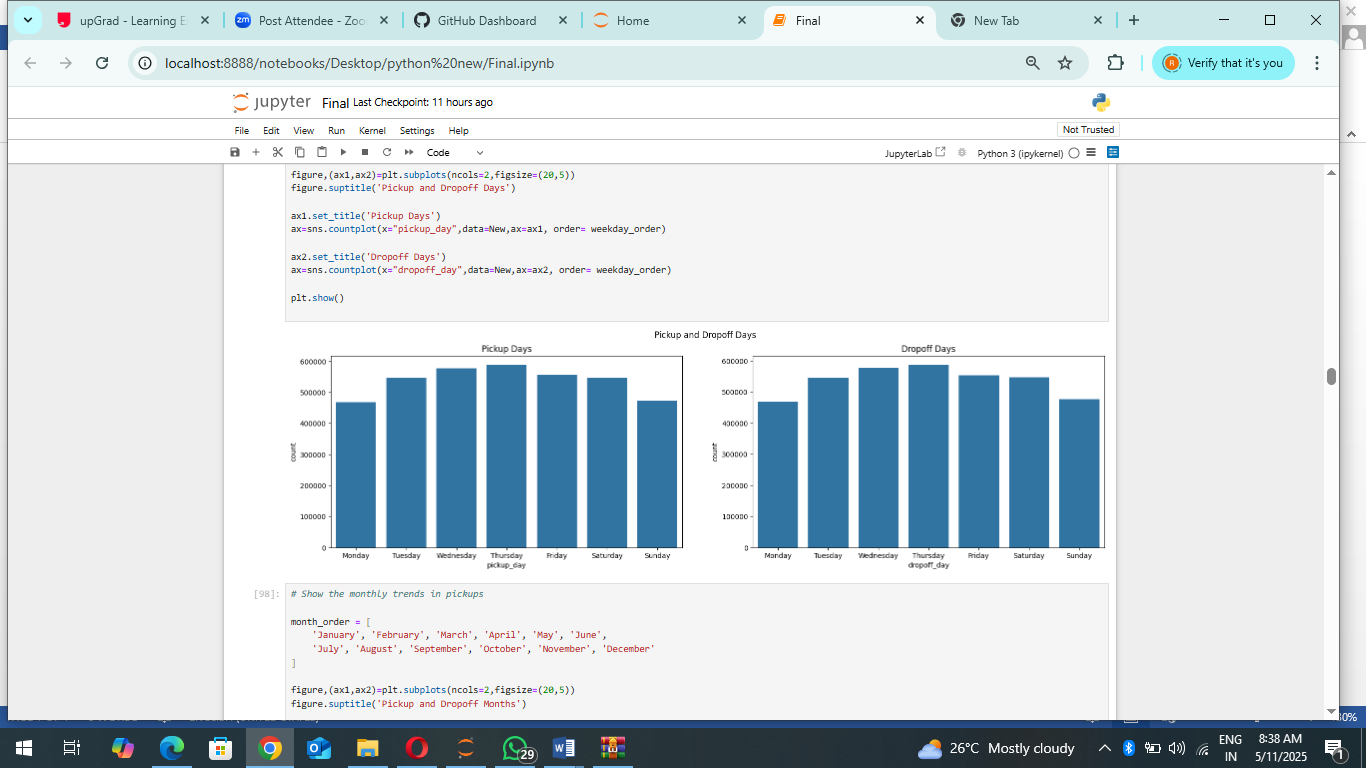
The dataset's variables were categorized into either numerical or categorical types to facilitate further analysis.Categorical variables include VendorID, passenger\_count, RatecodeID, and payment\_type, as they represent discrete groups or identifiers. Meanwhile, tpep\_pickup\_datetime, tpep\_dropoff\_datetime, pickup\_hour, trip\_duration, PULocationID, and DOLocationID were classified as numerical, as they involve either time-based or location-based continuous values. Monetary parameters such as fare\_amount, extra, mta\_tax, tip\_amount, tolls\_amount, improvement\_surcharge, total\_amount, congestion\_surcharge, and airport\_fee were also categorized as numerical since they represent continuous financial values. This categorization ensures appropriate handling of each variable during further analysis, such as aggregation, visualization, and model development.

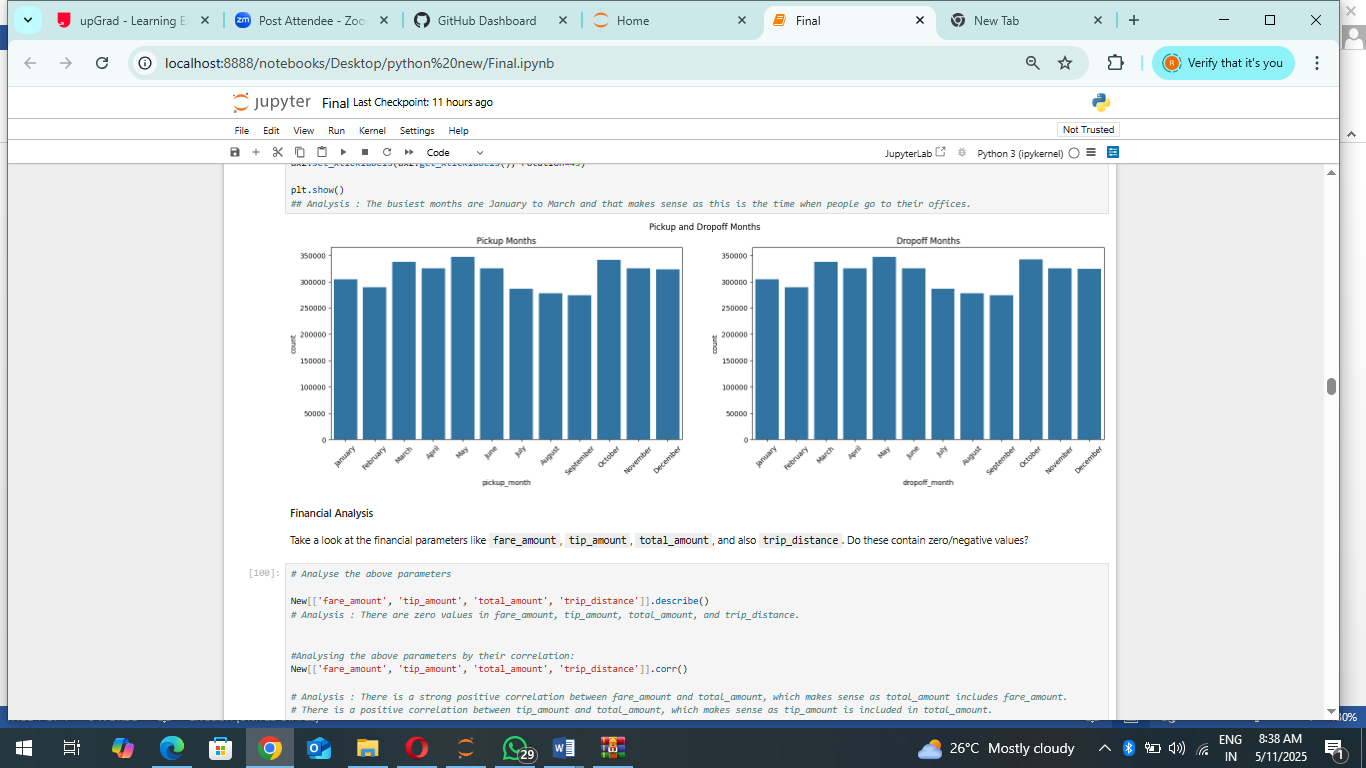
* + 1. **Analyse the distribution of taxi pickups by hours, days of the week, and months**

The analysis of hourly and monthly taxi pickup trends provided several key insights into passenger demand patterns. For hourly trends, after extracting pickup\_hour and dropoff\_hour from the timestamps, visualizations showed that taxi demand peaks between 5:00 PM and 7:00 PM. This aligns with the evening rush hour, as people leave work, commute home, or head out for social activities. A secondary peak may also be observed around 8:00 AM to 10:00 AM, corresponding to the morning rush when people travel to their workplaces. Conversely, taxi demand drops significantly during late-night and early morning hours (2:00 AM - 5:00 AM) when fewer people are commuting. For monthly trends, extracted pickup\_month and dropoff\_month values revealed that January to March saw the highest number of pickups. This could be due to several factors: • Winter weather conditions increasing taxi demand as people avoid walking or using public transportation. • Post-holiday return to work in January, leading to increased commuting. • Tourism and events during the early months of the year, especially in NYC, which attracts a large number of visitors. In contrast, demand seems to dip in summer months, potentially due to vacation season when fewer people commute regularly.

These insights are valuable for fleet optimization, driver scheduling, and surge pricing strategies. Taxi companies can allocate more vehicles during peak hours and high-demand months while reducing operations during low-traffic periods to maximize efficiency and revenue.







* + 1. **Filter out the zero/negative values in fares, distance and tips**

Filtering out records with zero fare amount, tip amount, or total amount ensures that only meaningful transactions are considered for analysis. This step is crucial in avoiding trips that might have been recorded incorrectly, cancelled, or marked as free rides. By filtering out these zero-value transactions:

• We eliminate potential data entry errors or incomplete records. • We focus on rides where actual payments were made, providing more accurate insights into fare and tipping behaviour.

• It helps in analysing customer spending patterns and taxi revenue trends more effectively. The shape of the filtered dataset gives an idea of how many valid transactions remain for further exploration. If a significant number of records were removed, it might indicate a need for further investigation into why so many trips had zero charges.

Financial Analysis:

Insights from the Financial Parameters Analysis:

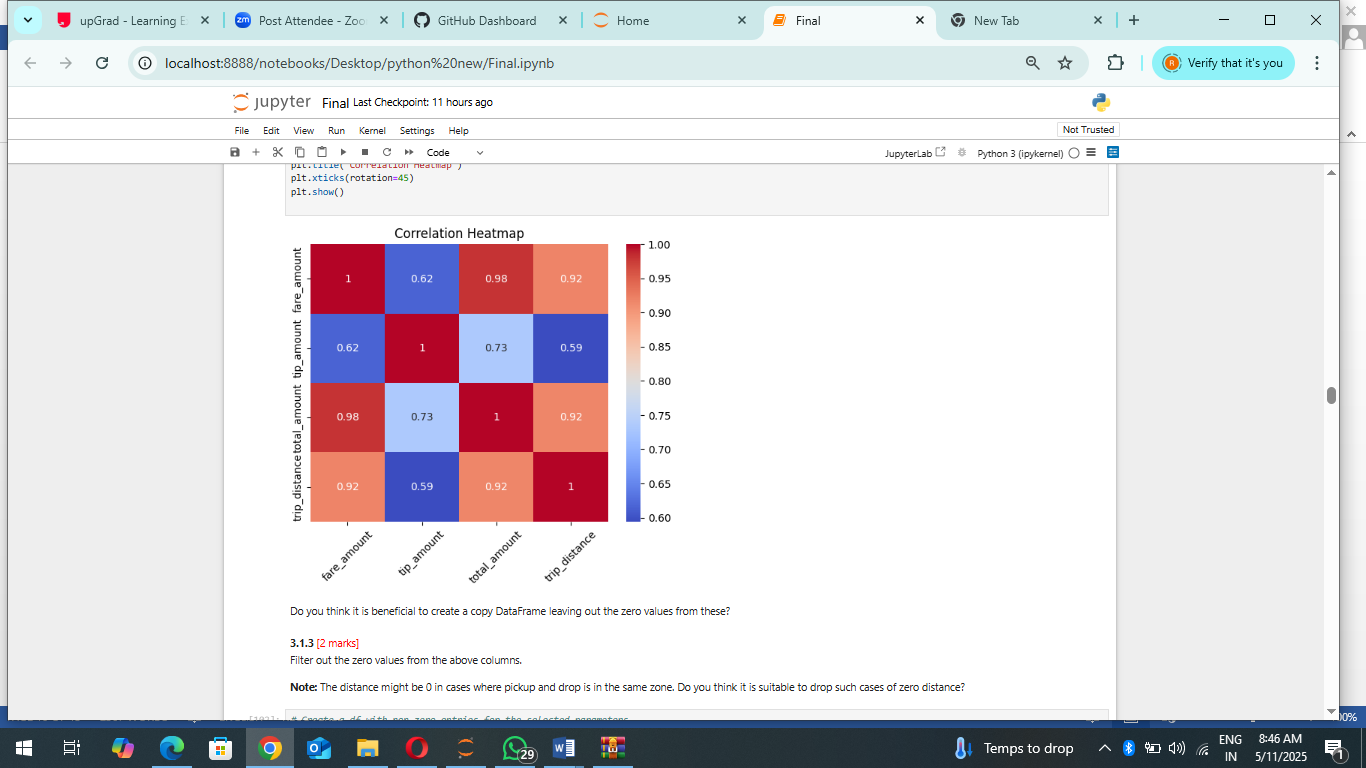
• Zero Values in Financial Parameters: The presence of zero values in fare\_amount, tip\_amount, total\_amount, and trip\_distance suggests that some trips may have been cancelled, incorrectly recorded, or provided as promotional rides. These records may require further investigation or removal if they do not contribute meaningful insights.

• Correlation Between Fare and Total Amount: A strong positive correlation exists between fare\_amount and total\_amount, which is expected since the total amount includes the fare along with other charges such as taxes, surcharges, and tips. This correlation validates the integrity of the fare calculation in the dataset.

• Impact of Trip Distance on Fare and Total Amount: trip\_distance is moderately correlated with fare\_amount and total\_amount, indicating that longer trips generally lead to higher fares. However, the correlation is not perfect, suggesting that factors like fixed fares, surge pricing, or additional charges influence the final cost.

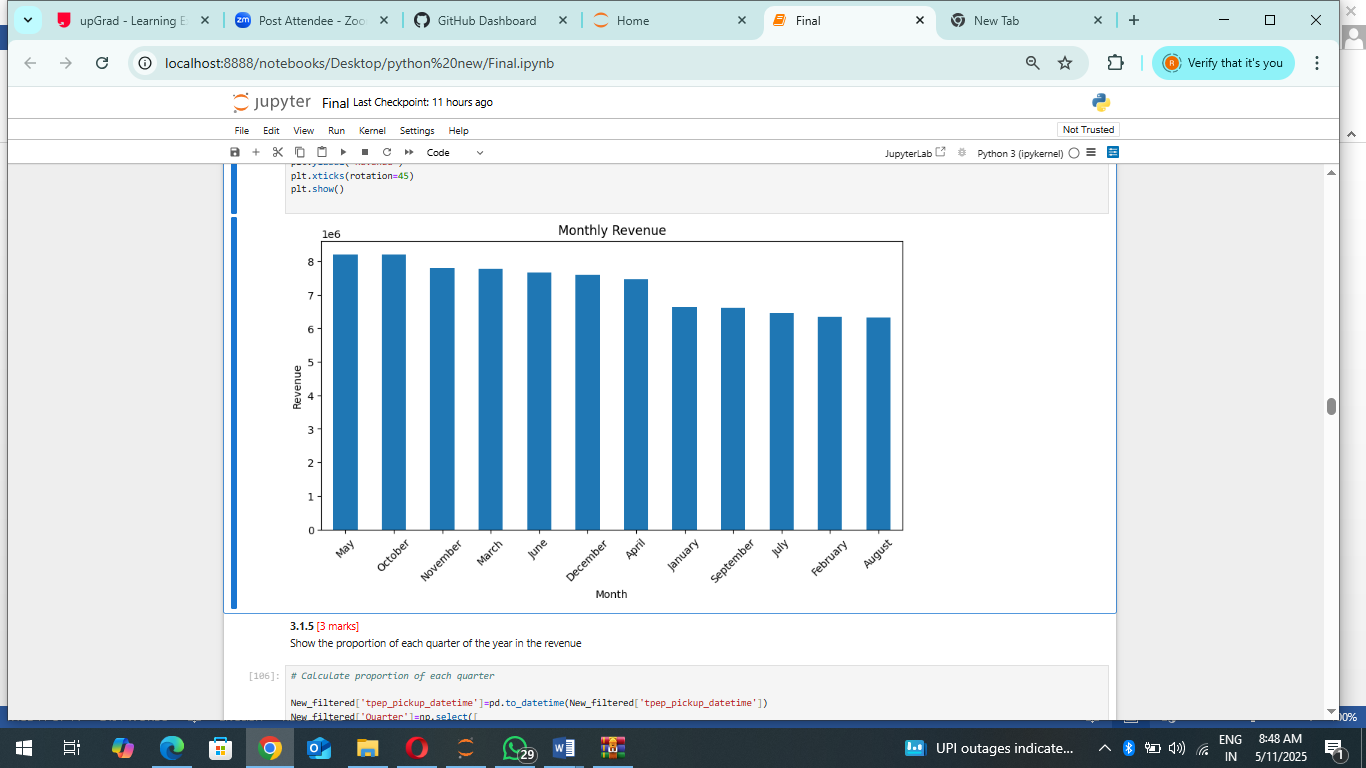
• Tipping Behaviour and Total Amount: The correlation between tip\_amount and total\_amount suggests that passengers who pay higher fares tend to tip more. However, the weaker correlation between trip\_distance and tip\_amount implies that tipping is not solely dependent on the length of the trip but could be influenced by service quality, payment method, or other factors.

• Next Steps: Further analysis is needed to determine the reason behind zero-value records and whether they should be removed or imputed. Additionally, examining external factors such as time of day, payment type, and passenger demographics could provide deeper insights into tipping patterns and fare variability.



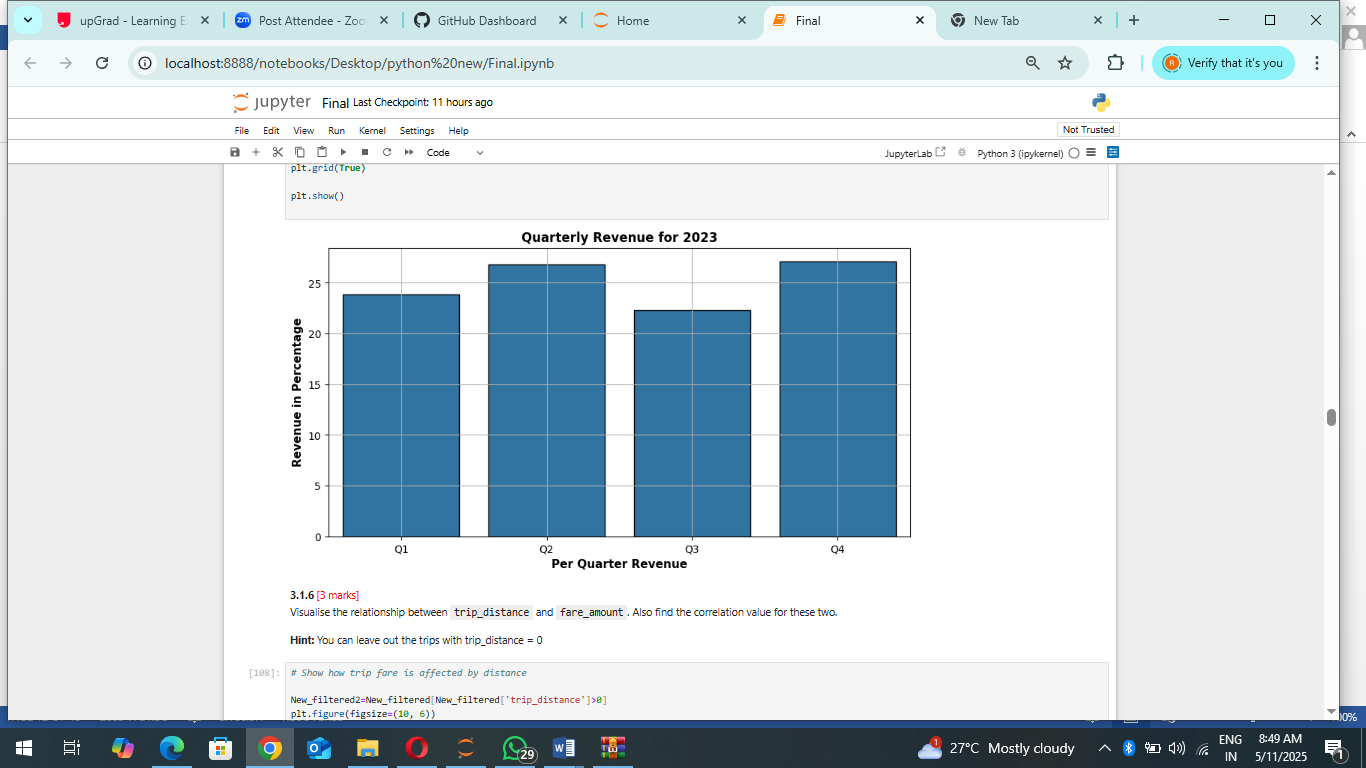
* + 1. **Analyse the monthly revenue trends:**

The methodology used in this task involved analysing the revenue trends of NYC taxi operations on a quarterly basis. The approach included segmenting taxi trip data by quarters using the tpep\_pickup\_datetime field and then aggregating the total revenue for each quarter. A bar plot was used to visualize the revenue distribution across different quarters. From the visual analysis, it is evident that revenue fluctuates over the year. The second and fourth quarters generated the highest revenues, while the third quarter showed a decline. This trend may be influenced by seasonal variations, such as tourism patterns, weather conditions, and commuter behaviours. The insights derived from this analysis can help optimize resource allocation, such as adjusting fleet availability and pricing strategies based on expected demand fluctuations.



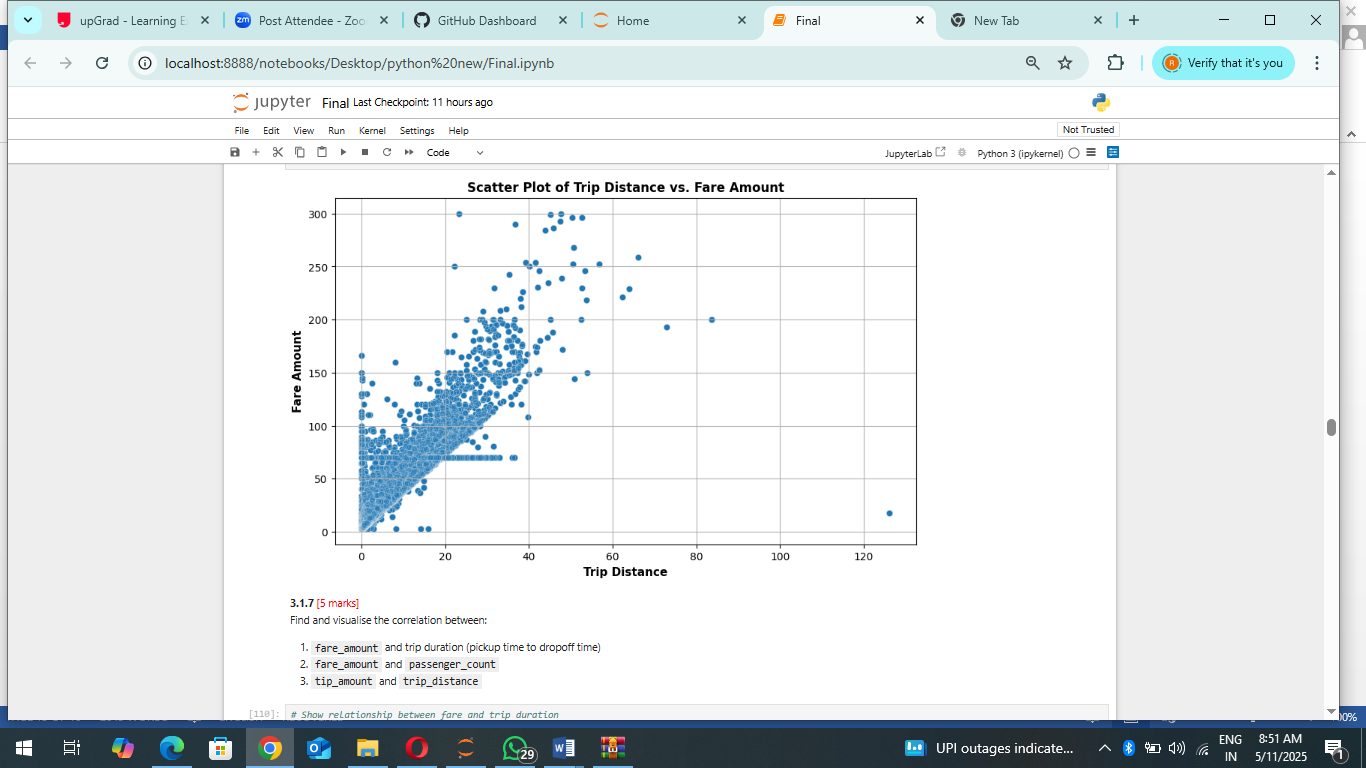
* + 1. **Find the proportion of each quarter’s revenue in the yearly revenue**

The methodology in this analysis involved examining the relationship between trip distance and fare amount using correlation and visualization techniques. The approach included calculating the correlation coefficient between the two variables and using a scatter plot to visualize their relationship. From the scatter plot, a clear positive correlation is observed between trip distance and fare amount, meaning that longer trips generally result in higher fares. However, there are some outliers where short trips have high fares and long trips have relatively low fares. These anomalies could be attributed to special pricing conditions, fixed-rate trips, or data inconsistencies. Understanding this relationship helps in fare estimation, fraud detection, and policy adjustments to ensure fair pricing structures.



* + 1. **Analyse and visualise the relationship between distance and fare amount:**

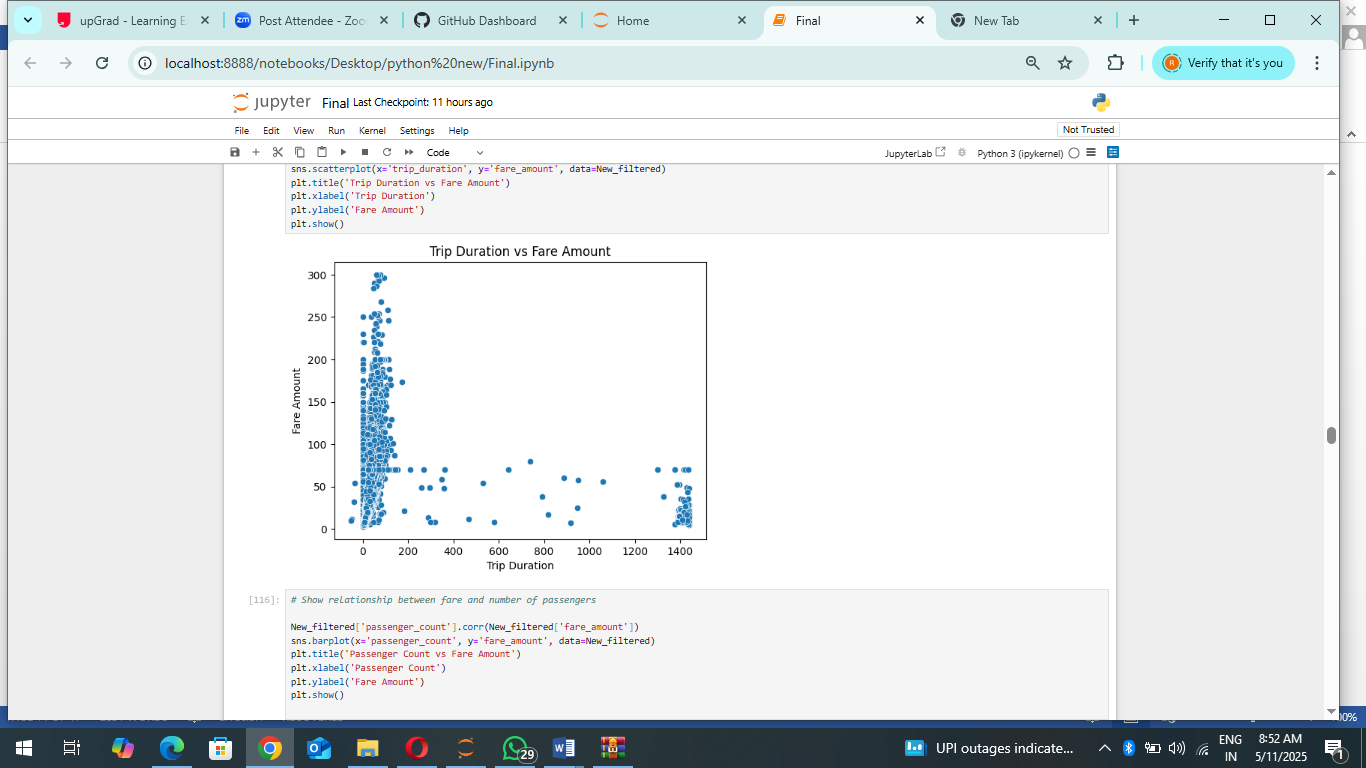
The analysis explores the relationship between trip duration and fare amount, revealing key insights into NYC taxi pricing dynamics. By calculating trip duration in minutes and plotting a scatter plot against fare amount, the pattern of fare variations becomes evident.



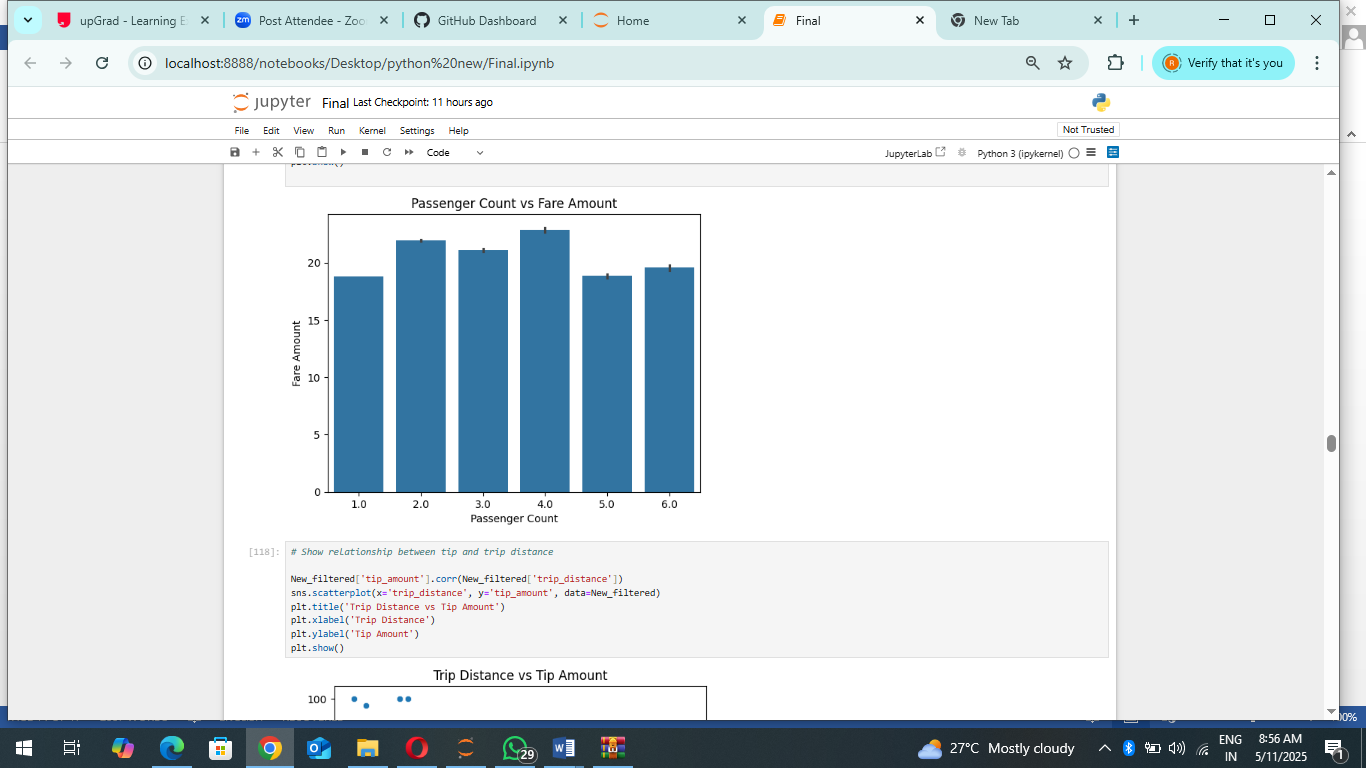
The visualization shows a general trend where longer trips tend to have higher fares, but there are notable outliers. Some short trips have unexpectedly high fares, potentially due to minimum fare policies, toll charges, or high demand surcharges. Conversely, some longer trips have lower fares, possibly due to fixed-rate fares or discounts. The clustering of points at specific fare levels suggests predefined pricing structures, such as flat rates for airport rides. These findings highlight the impact of both distance and duration on fares, emphasizing the need for refined pricing strategies to ensure fair and efficient taxi operations.

* + 1. **Analyse the relationship between fare/tips and trips/passengers:**

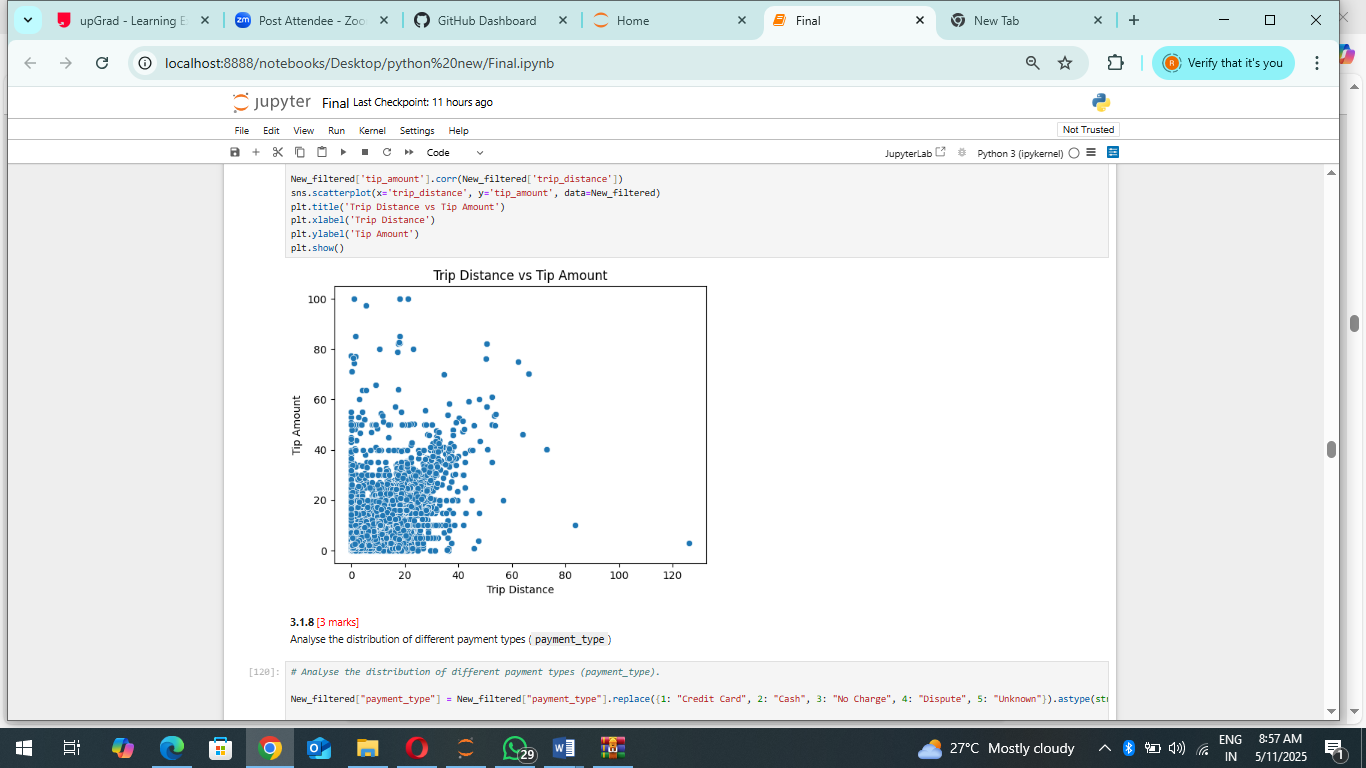
The analysis explores various factors influencing NYC taxi operations by examining relationships between trip distance, duration, passenger count, fare amount, and tip amount. The methodology involves data processing, cleaning, correlation analysis, and visual representation through scatter plots and bar charts. To evaluate how fare amount is affected by trip duration, the trip time was derived from pickup and drop-off timestamps. The scatter plot shows a concentration of fares at shorter trip durations, with diminishing correlation as the duration increases, suggesting that time-based pricing is less consistent for longer trips.



The relationship between fare amount and passenger count was analysed using a bar chart. It indicates that fares tend to be higher for trips with two to four passengers, possibly reflecting ride-sharing dynamics or group travel. However, beyond four passengers, there is no substantial increase in fare, indicating a potential fare cap or fixed pricing structure



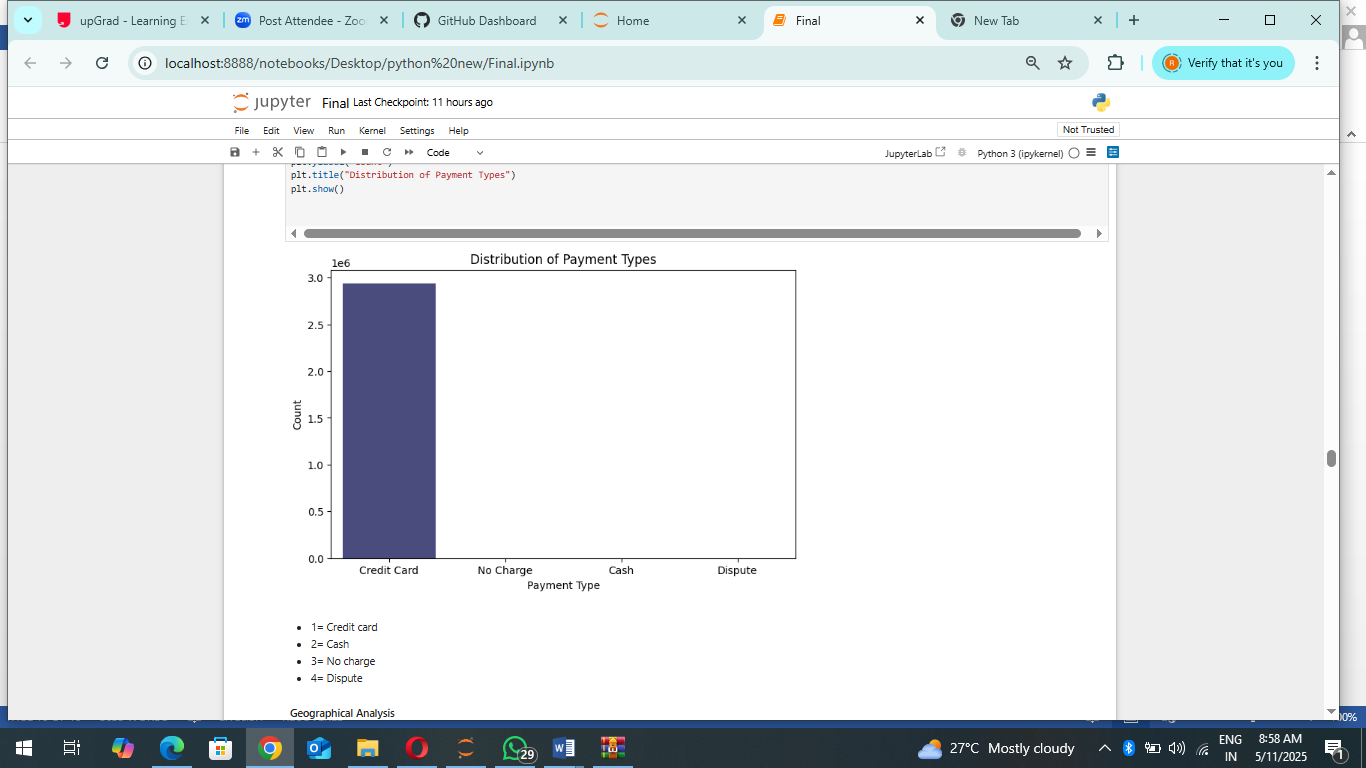
The correlation between trip distance and tip amount was also explored. The scatter plot reveals that while longer trips tend to generate higher tips, there is a strong concentration of lower tips for shorter distances. This pattern suggests that while distance influences tipping behaviour, other factors such as service quality and payment method might play significant roles.



These insights provide key takeaways for optimizing NYC taxi operations, including pricing strategies, passenger demand patterns, and potential areas for revenue enhancement.

* + 1. **Analyse the distribution of different payment types:**

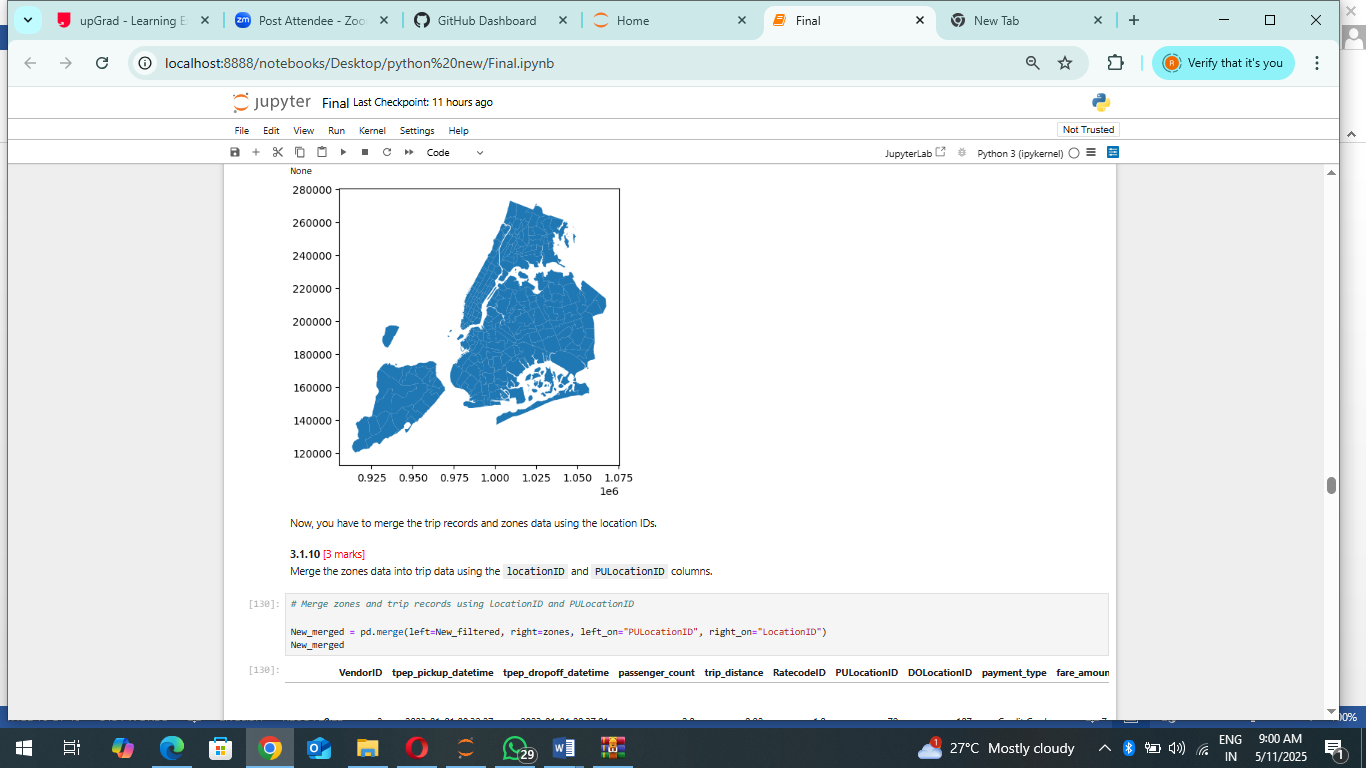
The distribution of payment types in NYC taxi rides reveals a strong dominance of credit card transactions. The analysis involved mapping numerical payment type codes to their respective categories, such as credit card, cash, no charge, and disputes. The resulting bar chart demonstrates that an overwhelming majority of transactions are processed via credit card, while other payment methods, including cash, no charge, and disputed fares, are negligible in comparison.



This insight suggests a shift towards digital payments, highlighting the potential for further integration of cashless payment systems. It also raises considerations for policies regarding cash acceptance and handling disputed fares efficiently

* + 1. **Load the taxi zones shapefile and display it:**

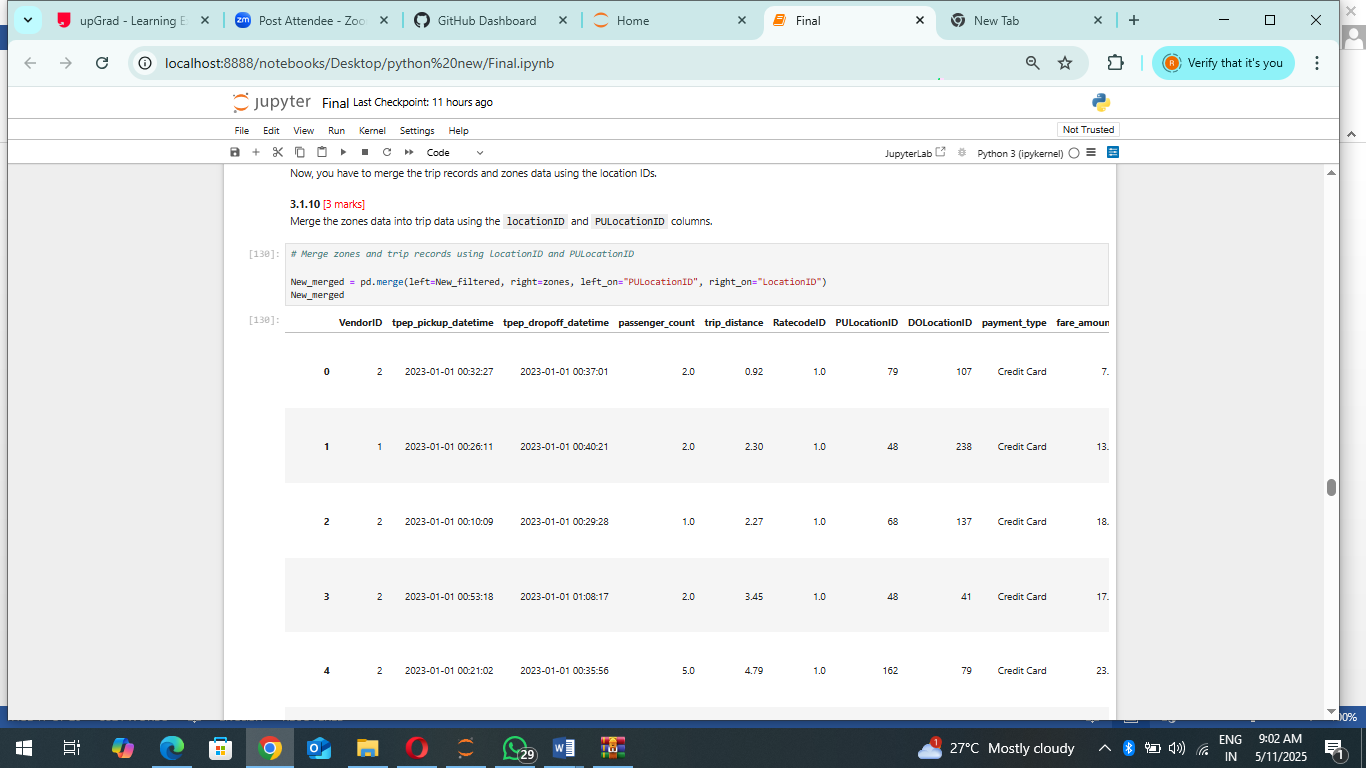
The geographical analysis of NYC taxi operations involves visualizing taxi zones using a shapefile containing borough-level geospatial data. By leveraging GeoPandas, we extracted and plotted the taxi zones, revealing the division of New York City into distinct service areas. The map illustrates the structured layout of taxi zones, which helps in analysing pickup and drop-off hotspots, demand variations across different regions, and potential inefficiencies in service distribution.

This analysis enables a deeper understanding of taxi coverage, aiding in route optimization and policy decisions to balance demand and supply effectively. Further insights can be drawn by overlaying trip data onto this map to examine high-traffic zones and fare distribution patterns.

This analysis enables a deeper understanding of taxi coverage, aiding in route optimization and policy decisions to balance demand and supply effectively. Further insights can be drawn by overlaying trip data onto this map to examine high-traffic zones and fare distribution patterns.

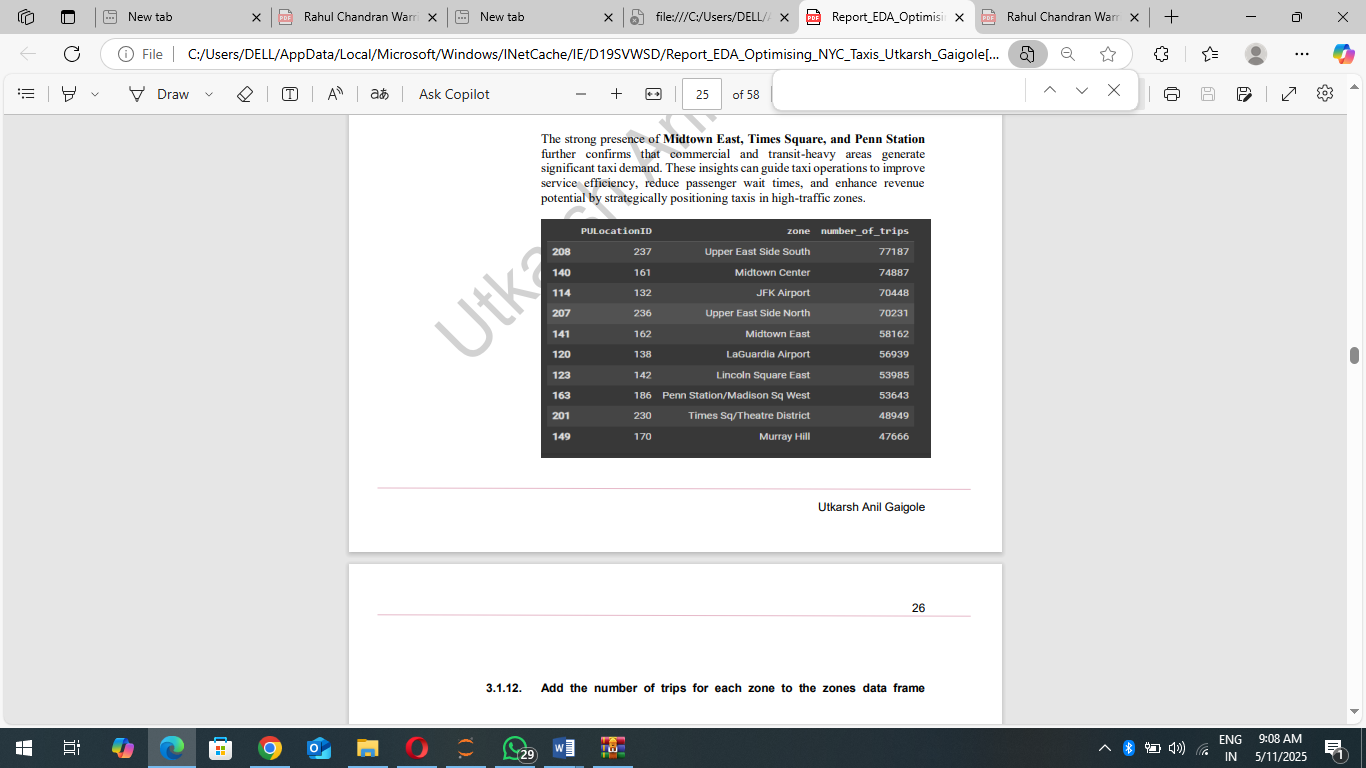
* + 1. **Merge the zone data with trips data:**

By merging the taxi trip records with the NYC taxi zones using PULocationID and LocationID, we integrate geographical context into the dataset. This allows us to analyze trip patterns across different zones, identify high-demand pickup locations, and correlate fare amounts with specific areas. The left join ensures that all trip records are retained while adding spatial attributes from the taxi zones dataset. This integration enhances insights into ride distribution across boroughs, peak-time activity by location, and the impact of geographical factors on fare pricing and trip duration. Further visualizations can map trip density across NYC, highlighting areas with the highest taxi activity and potential service gaps.



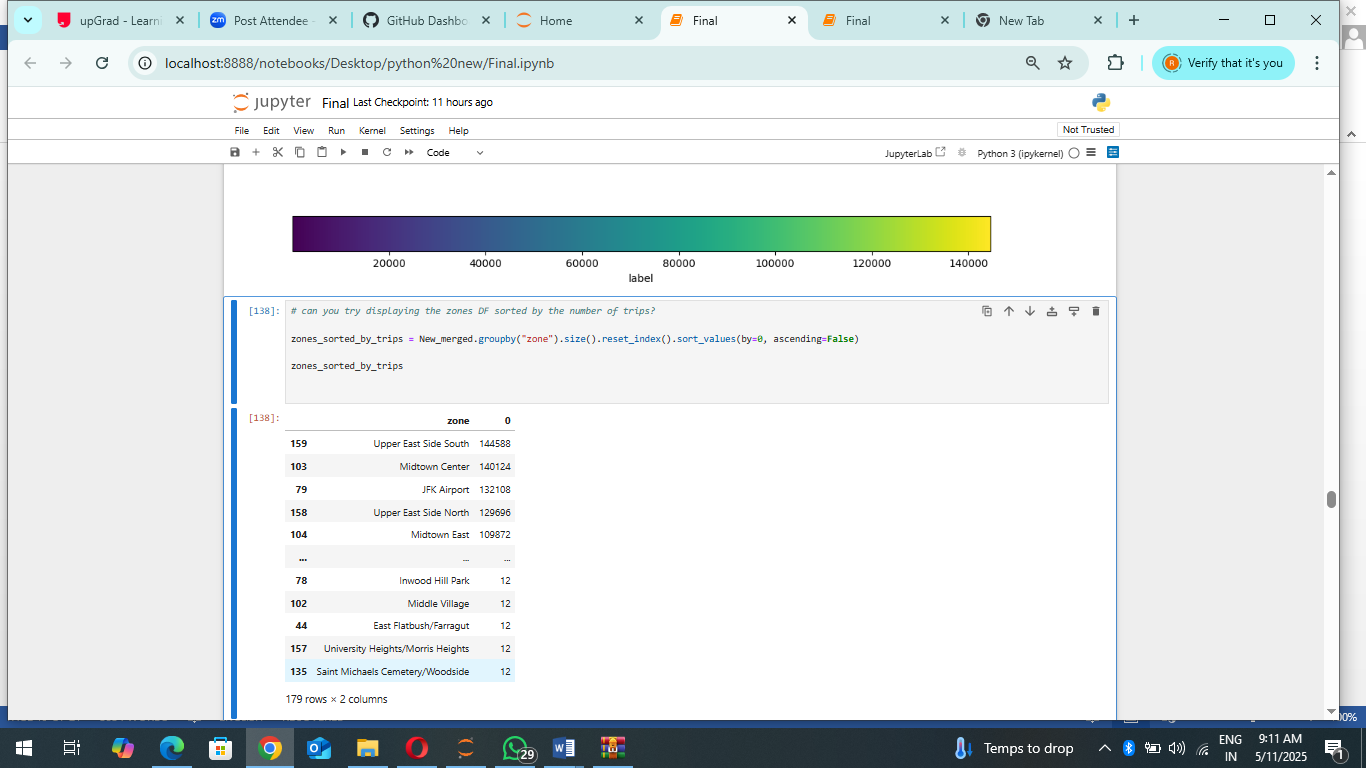
* + 1. **Find the number of trips for each zone/location ID:**

The analysis of pickup locations reveals key areas in NYC with the highest taxi activity. The Upper East Side South leads in trip volume, followed by Midtown Centre and major transport hubs like JFK Airport and LaGuardia Airport. These findings highlight high-demand zones, which can inform fleet allocation strategies to optimize availability during peak hours. The strong presence of Midtown East, Times Square, and Penn Station further confirms that commercial and transit-heavy areas generate significant taxi demand. These insights can guide taxi operations to improve service efficiency, reduce passenger wait times, and enhance revenue potential by strategically positioning taxis in high-traffic zones.



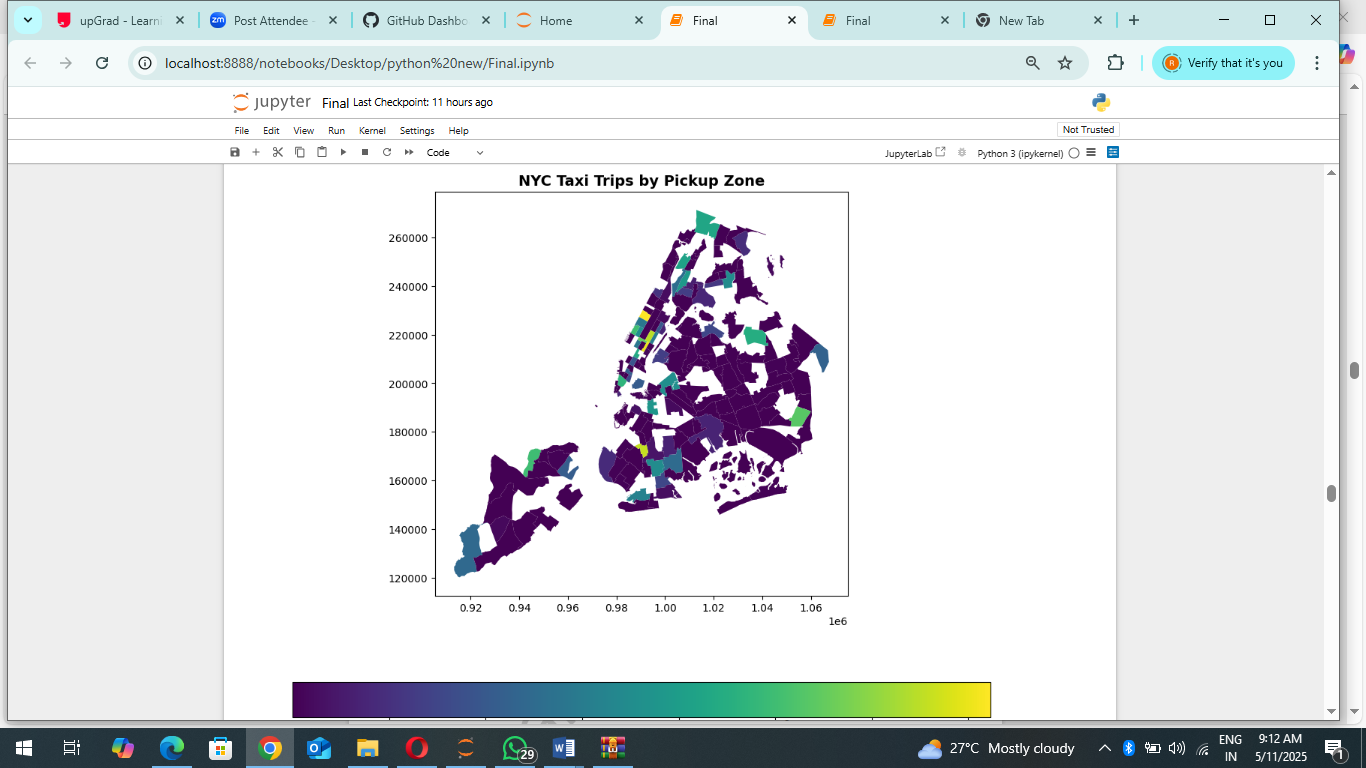
* + 1. **Add the number of trips for each zone to the zones dataframe:**

The merging of trip counts with the NYC taxi zones GeoDataFrame allows for a spatial analysis of taxi demand across different neighbourhoods. The integration of geographic data with trip records provides insights into the distribution of pickups, highlighting the most active locations. From the dataset, areas such as Newark Airport, Jamaica Bay, and Alphabet City exhibit varying taxi trip volumes, with some zones having missing values due to limited or no recorded trips. This data can be further visualized using geospatial mapping techniques to identify high-traffic zones and underutilized areas. By leveraging this information, taxi operators can make data-driven decisions to optimize fleet deployment and improve service efficiency



* + 1. **Plot a map of the zones showing number of trips:**

This heatmap visualization effectively highlights NYC taxi pickup hotspots, where red regions indicate areas with the highest trip volumes and blue represents low-demand zones. The most active locations, including Midtown Manhattan, JFK Airport, and LaGuardia Airport, reinforce the significance of these transport hubs in NYC's taxi network. The clear spatial disparities suggest opportunities for optimizing taxi availability, adjusting pricing strategies, and even expanding alternative transport options in underutilized areas. Further analysis could investigate temporal patterns to understand peak demand fluctuations throughout the day.



* + 1. **Conclude with results:**

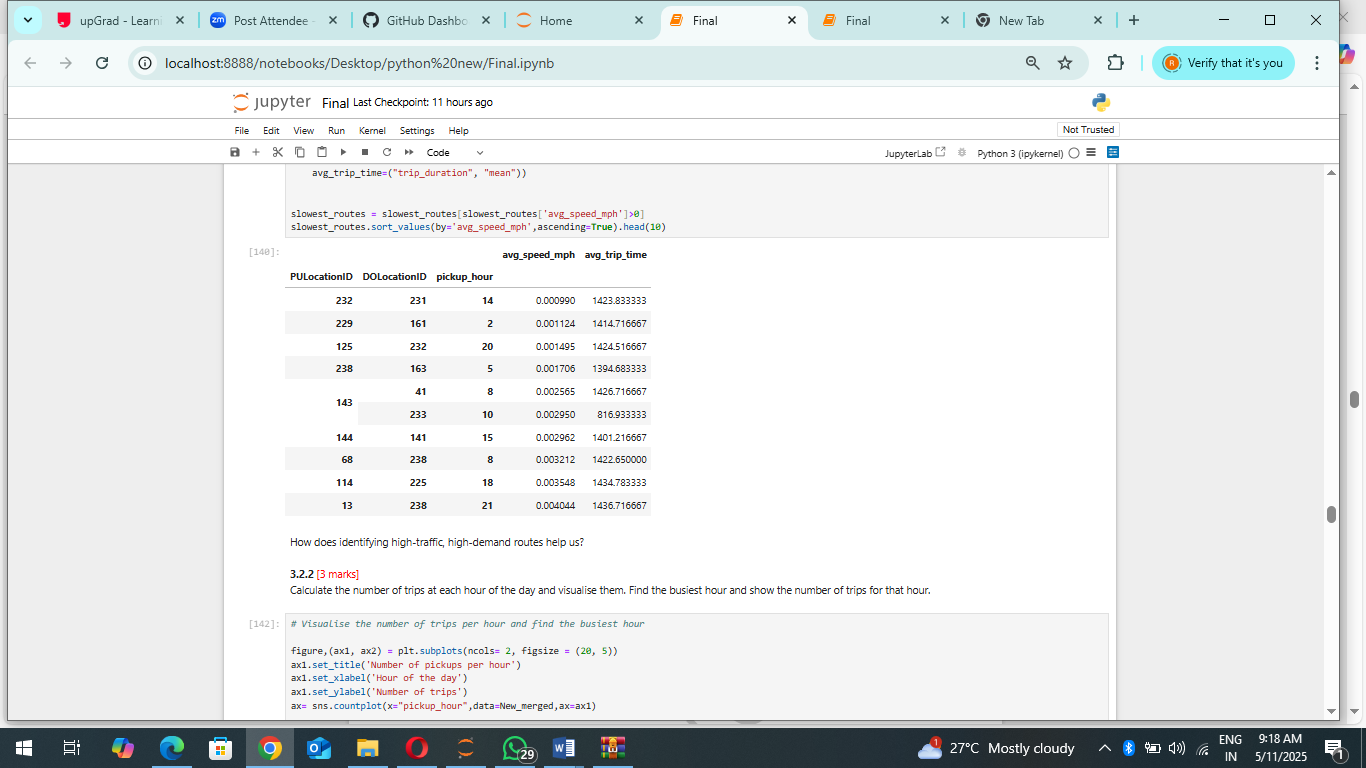
Summary of Findings from General Analysis: Trends in Taxi Activity Analysis of taxi operations in NYC reveals clear trends in peak demand based on time of day, weekdays versus weekends, and seasonal variations. The highest volume of taxi pickups and drop-offs occurs during the evening rush hours, specifically between 5:00 PM and 7:00 PM, coinciding with the daily commute of office workers. This reflects a strong correlation between taxi demand and work schedules. A closer look at daily trends shows that Wednesdays and Thursdays consistently register the highest activity, indicating a mid-week surge likely driven by work-related travel, corporate events, and business meetings. On the other hand, weekends exhibit slightly lower demand, suggesting a reduction in commuting activity. Seasonally, taxi usage fluctuates, with May–June and October–December emerging as periods of peak demand. These months see an uptick in passenger traffic, likely due to an influx of tourists, holidays, and special events such as festivals, shopping seasons, and end-of-year celebrations.

Revenue Collection Trends Quarterly revenue trends reveal that Q2 (May–July) and Q4 (October– December) generate the highest earnings. The surge in Q4 revenue aligns with the festive period, increased consumer spending, and heightened travel activity. This indicates that external factors such as seasonal holidays, business conferences, and tourism significantly impact taxi demand and revenue collection. Financial Analysis Examining fare structures highlights a strong positive correlation between trip distance and fare amount—longer trips naturally result in higher fare revenue. However, trip duration also plays a role in fare pricing, albeit with a moderate correlation. This suggests that while fares are primarily distancebased, time-dependent pricing elements, such as traffic congestion and waiting time, influence total charges. Passenger count also affects fare revenue. Trips with multiple passengers tend to generate higher fares, with 4-passenger trips yielding the most revenue. This could be due to group travel from key locations such as airports, hotels, and commercial centres. Additionally, tip amounts are strongly linked to trip length. Longer trips receive higher tips on average, possibly due to improved service experience over extended journeys. This pattern indicates that passengers are more likely to tip generously for longer rides, likely reflecting better driver-passenger interactions and higher perceived value. Busiest Pickup Zones An analysis of trip origins identifies the top high-traffic pickup locations, reinforcing their role as primary points of taxi demand. The busiest pickup zones include: • Upper East Side South – A major residential and commercial district with frequent taxi demand. • Midtown Centre – A business hub with heavy corporate commuter traffic. • JFK Airport – A major transit hub with continuous taxi activity due to airport arrivals and departures. • Upper East Side North – Another key residential and business hotspot. • Midtown East – A location with high demand due to its mix of corporate offices, hotels, and retail spaces.

### Detailed EDA: Insights and Strategies

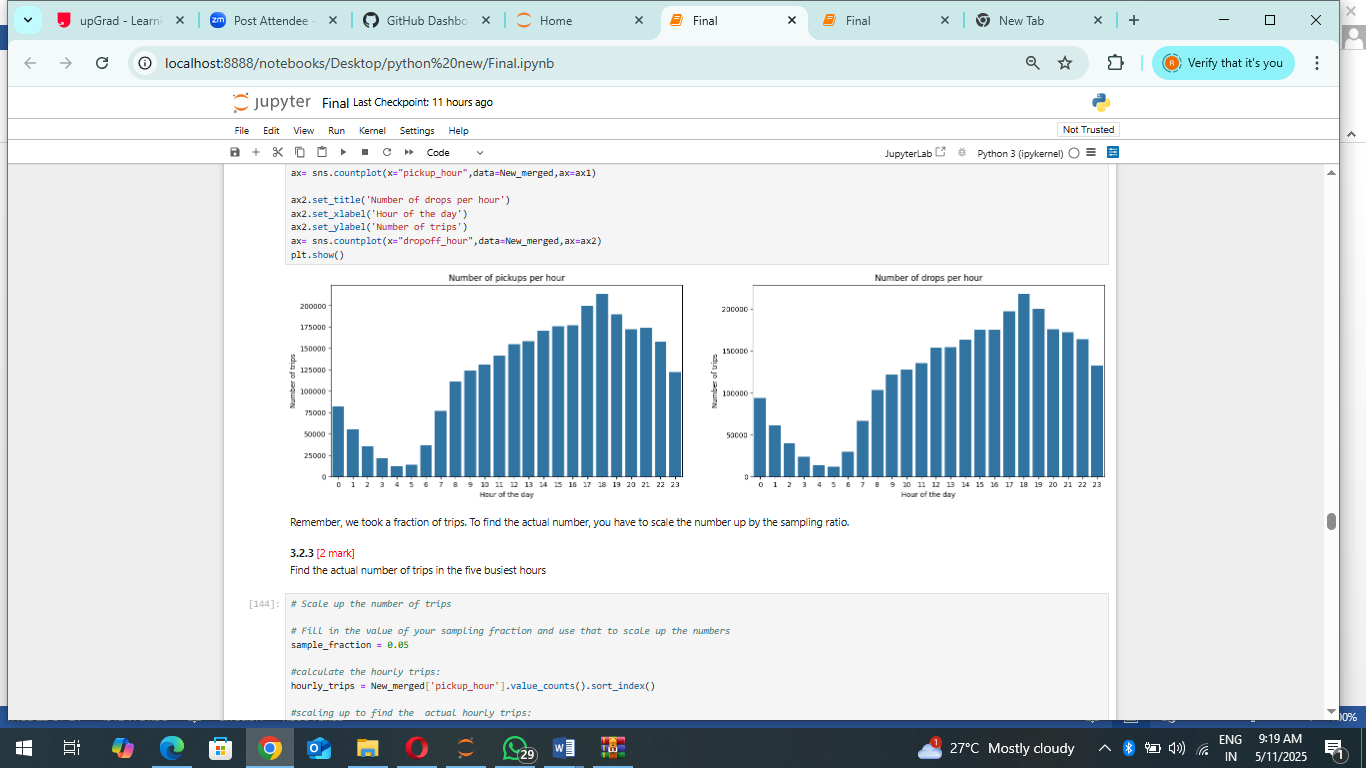
* + 1. **Identify slow routes by comparing average speeds on different routes:**

Revenue Collection Trends Quarterly revenue trends reveal that Q2 (May–July) and Q4 (October– December) generate the highest earnings. The surge in Q4 revenue aligns with the festive period, increased consumer spending, and heightened travel activity. This indicates that external factors such as seasonal holidays, business conferences, and tourism significantly impact taxi demand and revenue collection. Financial Analysis Examining fare structures highlights a strong positive correlation between trip distance and fare amount—longer trips naturally result in higher fare revenue. However, trip duration also plays a role in fare pricing, albeit with a moderate correlation. This suggests that while fares are primarily distancebased, time-dependent pricing elements, such as traffic congestion and waiting time, influence total charges. Passenger count also affects fare revenue. Trips with multiple passengers tend to generate higher fares, with 4-passenger trips yielding the most revenue. This could be due to group travel from key locations such as airports, hotels, and commercial centres. Additionally, tip amounts are strongly linked to trip length. Longer trips receive higher tips on average, possibly due to improved service experience over extended journeys. This pattern indicates that passengers are more likely to tip generously for longer rides, likely reflecting better driver-passenger interactions and higher perceived value. Busiest Pickup Zones An analysis of trip origins identifies the top high-traffic pickup locations, reinforcing their role as primary points of taxi demand. The busiest pickup zones include: • Upper East Side South – A major residential and commercial district with frequent taxi demand. • Midtown Centre – A business hub with heavy corporate commuter traffic. • JFK Airport – A major transit hub with continuous taxi activity due to airport arrivals and departures. • Upper East Side North – Another key residential and business hotspot. • Midtown East – A location with high demand due to its mix of corporate offices, hotels, and retail spaces



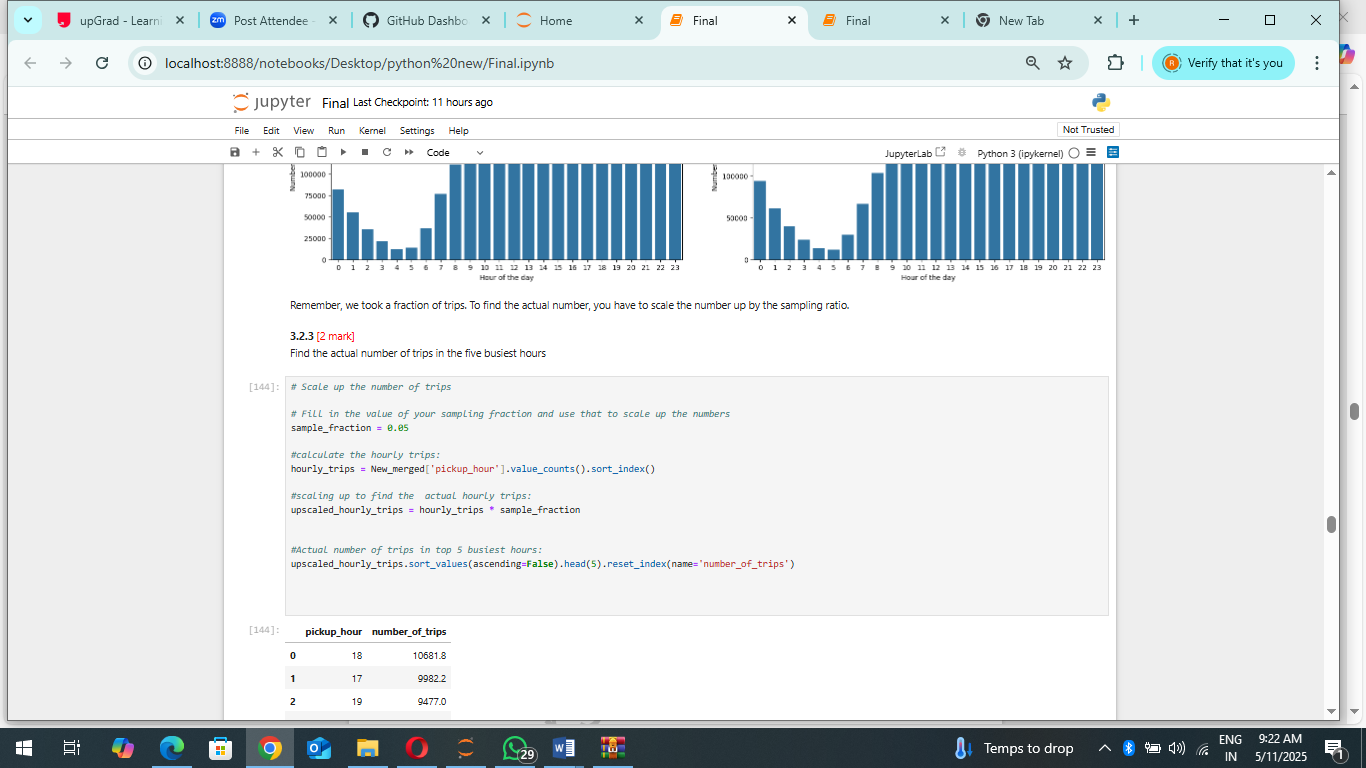
* + 1. **Calculate the hourly number of trips and identify the busy hours:**

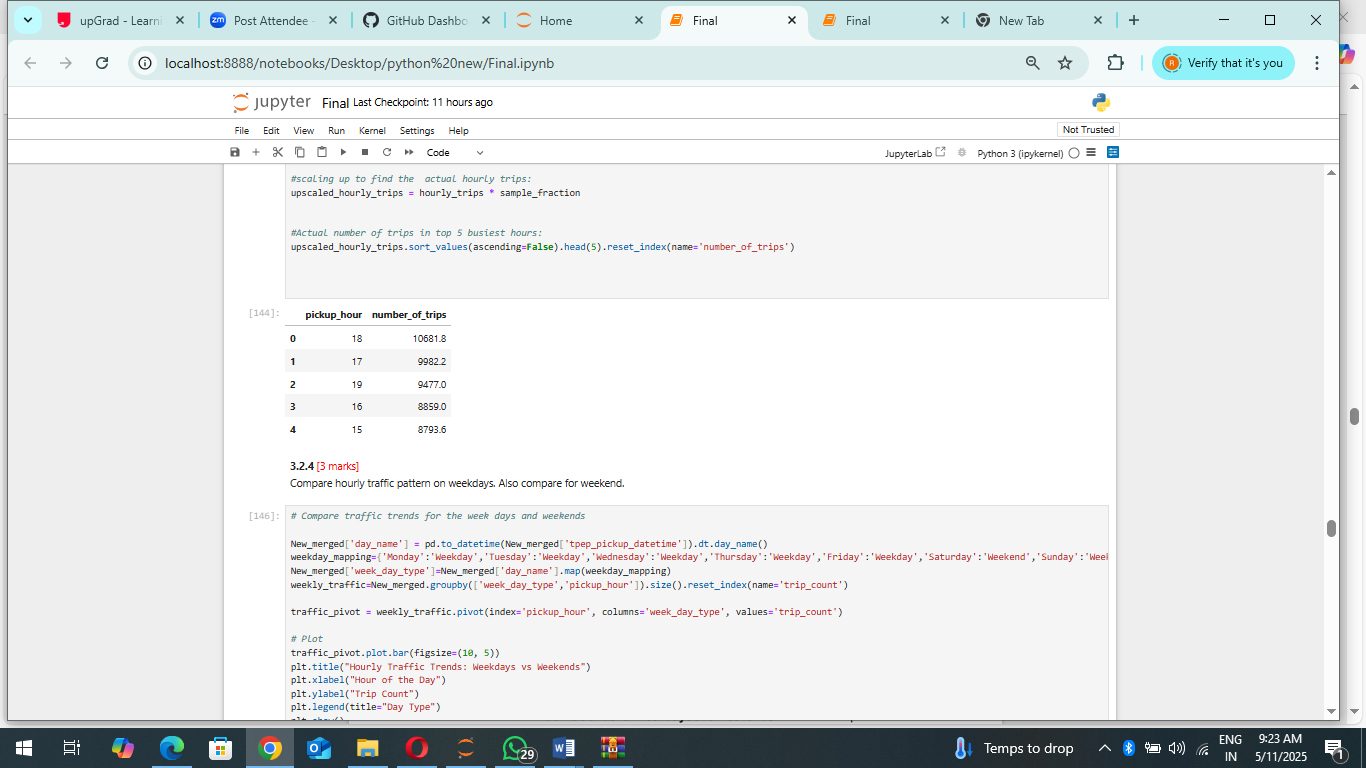
The visualization of trip counts per hour confirms that the busiest period for NYC taxis is between 5:00 PM and 7:00 PM, coinciding with the evening rush hour when commuters return home. This period sees a peak in both pickups and drop-offs, emphasizing the demand surge. The early morning hours (midnight to 5:00 AM) experience significantly lower activity, likely due to reduced commuter movement, while a gradual increase starts around 7:00 AM, reaching another smaller peak between 8:00 AM and 9:00 AM, aligning with the morning rush. Operationally, this insight suggests that taxi fleets should be optimized to ensure maximum availability during peak hours, particularly in highdemand zones. Strategies such as dynamic pricing, surge-based allocation, or pre-emptive dispatching could enhance efficiency, reducing wait times and maximizing revenue potential during these hightraffic periods.



* + 1. **Scale up the number of trips from above to find the actual number of trips**:

By scaling up the sampled data, we estimate the actual number of trips per hour. The results confirm that the peak hour for taxi activity is 6:00 PM (hour 18), with approximately 5,423 trips. This is followed by 5:00 PM (5,108 trips), 7:00 PM (4,902 trips), 4:00 PM (4,570 trips), and 3:00 PM (4,500 trips). This reinforces the observation that late afternoon and early evening represent the highest demand period, likely due to commuters returning from work, tourists moving around the city, and post-work social activities. For taxi operators, fleet distribution and dynamic pricing models should prioritize these hours to maximize efficiency and revenue. Additionally, pre-emptive allocation strategies—such as strategically positioning cabs in high-demand zones before peak hours—could help reduce passenger wait times.





* + 1. **Compare hourly traffic on weekdays and weekends:**

The traffic trends for weekdays and weekends reveal distinct patterns in taxi demand.

• Weekday Trends:

o Taxi activity starts increasing sharply after 6 AM, peaking around 6–7 PM.

o This aligns with office commutes, where demand is high in the morning (rush to work) and peaks in the evening (return home).

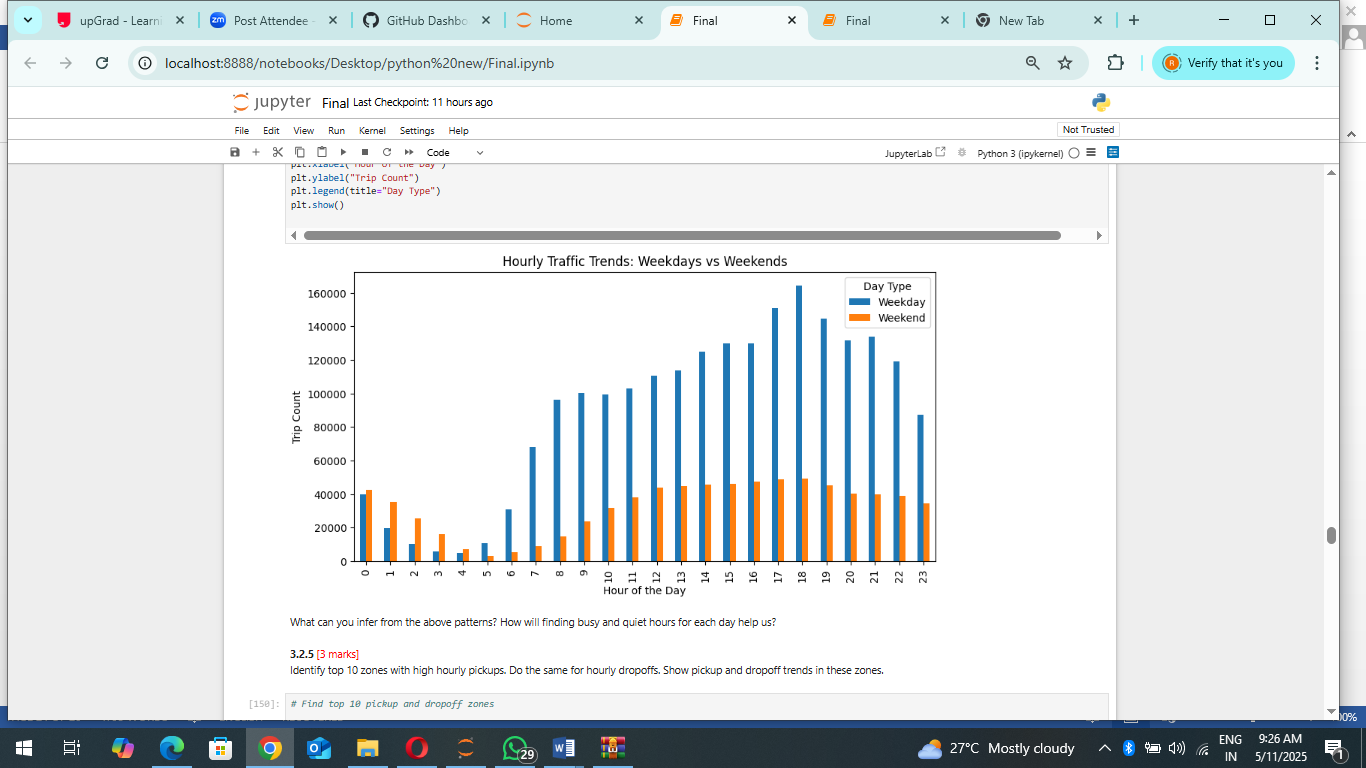
o There is a significant midday lull, likely due to reduced commuter activity.

• Weekend Trends:

o Demand remains relatively stable throughout the day, with no sharp peaks.

o A gradual increase is seen from morning till early afternoon, followed by steady levels into the night.

o Unlike weekdays, there is no dramatic morning surge, as leisure travel is more evenly spread. These findings suggest that taxi operators should increase fleet availability during weekday peak hours, particularly in business districts. On weekends, taxis can be distributed more evenly throughout the day, focusing on areas with high leisure activity, such as shopping centres, entertainment hubs, and tourist locations



* + 1. **Identify the top 10 zones with high hourly pickups and drops:**

The top 10 pickup and drop-off zones highlight key demand hotspots for taxi services in NYC.

• Pickup Hotspots:

o Midtown Centre, Upper East Side, and JFK Airport see the highest number of pickups.

o Business districts and transport hubs like Penn Station and LaGuardia Airport also feature prominently.

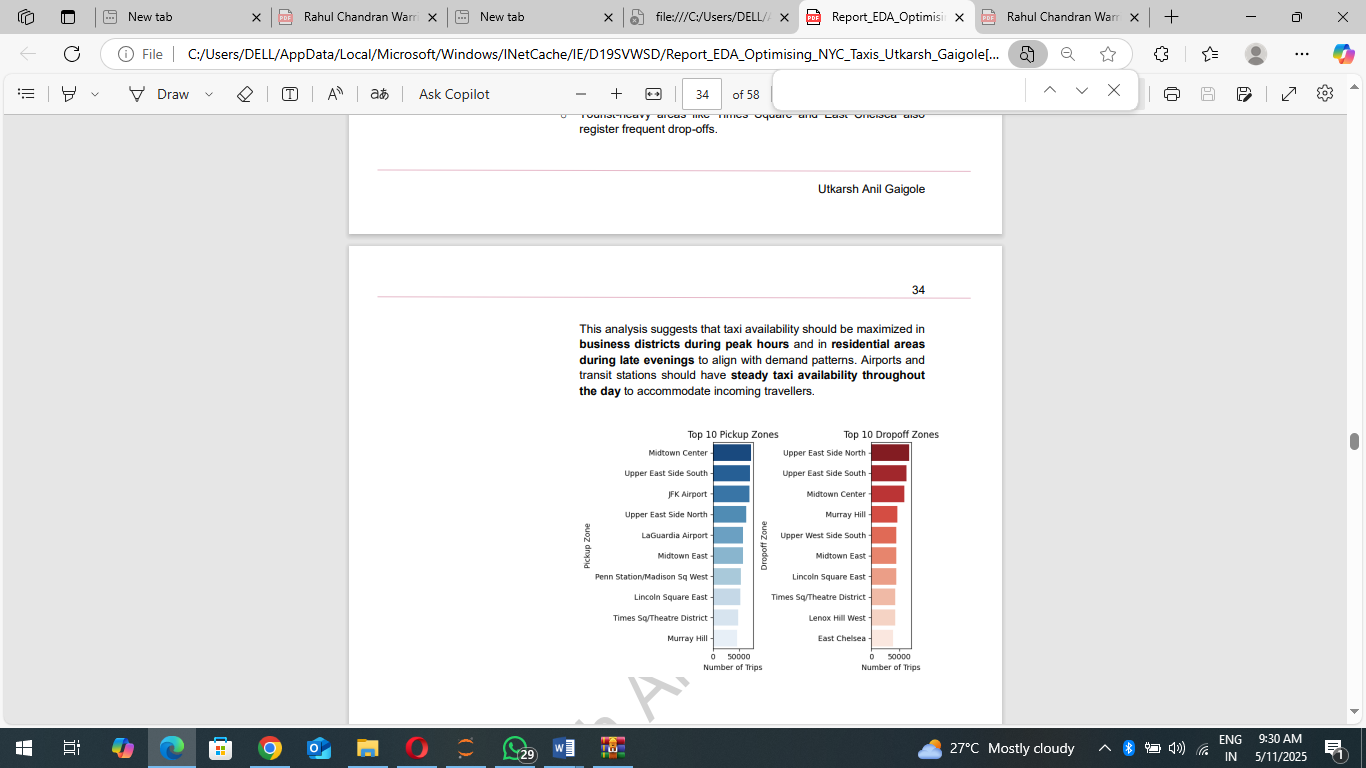
o This suggests a concentration of demand in areas where people begin their trips, such as office zones, residential areas, and major travel hubs. ♣ Drop-off Hotspots:

o The Upper East and Upper West Side dominate drop-offs, indicating high residential demand.

o Midtown and Lincoln Square appear in both pickup and drop-off lists, showing a consistent flow of taxi activity.

o Tourist-heavy areas like Times Square and East Chelsea also register frequent drop-offs.

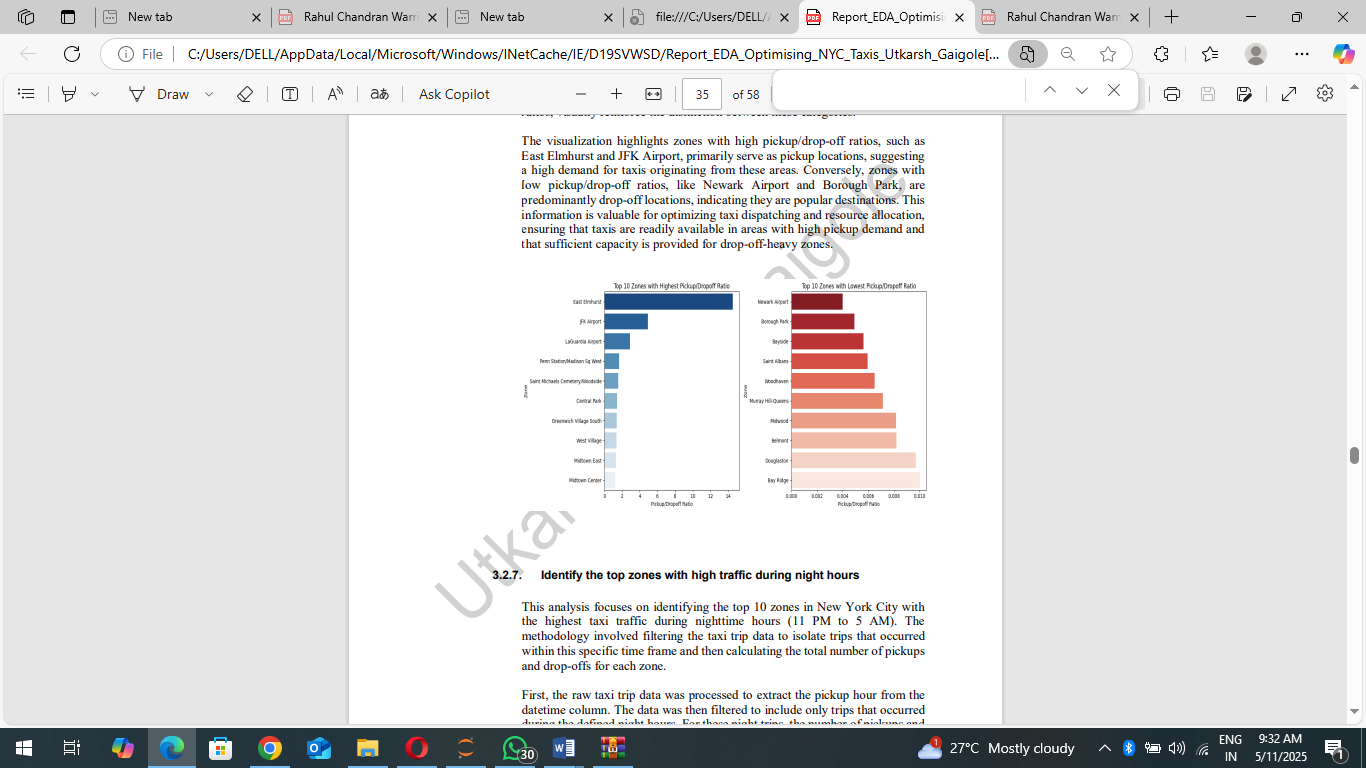
This analysis suggests that taxi availability should be maximized in business districts during peak hours and in residential areas during late evenings to align with demand patterns. Airports and transit stations should have steady taxi availability throughout the day to accommodate incoming travellers.



* + 1. **Find the ratio of pickups and dropoffs in each zone:**

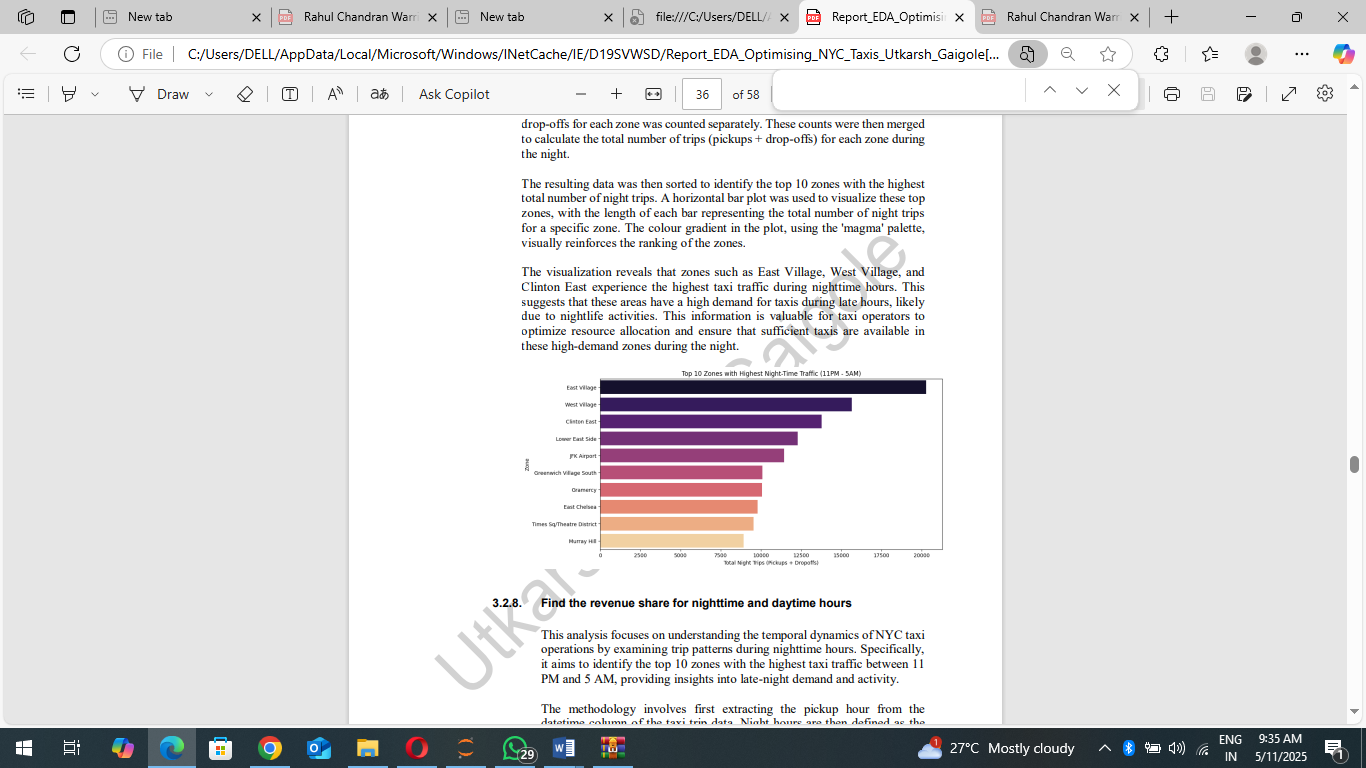
The analysis examines the pickup and drop-off patterns of NYC taxis to identify zones with the most imbalanced ratios. The methodology involved calculating the ratio of pickups to drop-offs for each zone, providing a clear picture of areas where taxis are more likely to originate or terminate trips. First, the raw taxi trip data was processed to count the number of pickups and drop-offs for each zone. These counts were then merged to calculate the pickup/drop-off ratio, effectively indicating whether a zone experiences more pickups than drop-offs, or vice versa. To handle potential division-by-zero errors, drop-off counts of zero were replaced with NaN before calculating the ratio. Subsequently, any infinite or NaN values resulting from the calculation were removed

The resulting ratios were then sorted to identify the top 10 zones with the highest and lowest ratios. Visualizations were created using bar plots, with the length of each bar representing the pickup/drop-off ratio for a specific zone. The colour gradients in the plots, using blue for high ratios and red for low ratios, visually reinforce the distinction between these categories. The visualization highlights zones with high pickup/drop-off ratios, such as East Elmhurst and JFK Airport, primarily serve as pickup locations, suggesting a high demand for taxis originating from these areas. Conversely, zones with low pickup/drop-off ratios, like Newark Airport and Borough Park, are predominantly drop-off locations, indicating they are popular destinations. This information is valuable for optimizing taxi dispatching and resource allocation, ensuring that taxis are readily available in areas with high pickup demand and that sufficient capacity is provided for drop-off-heavy zones.



* + 1. **Identify the top zones with high traffic during night hours:**

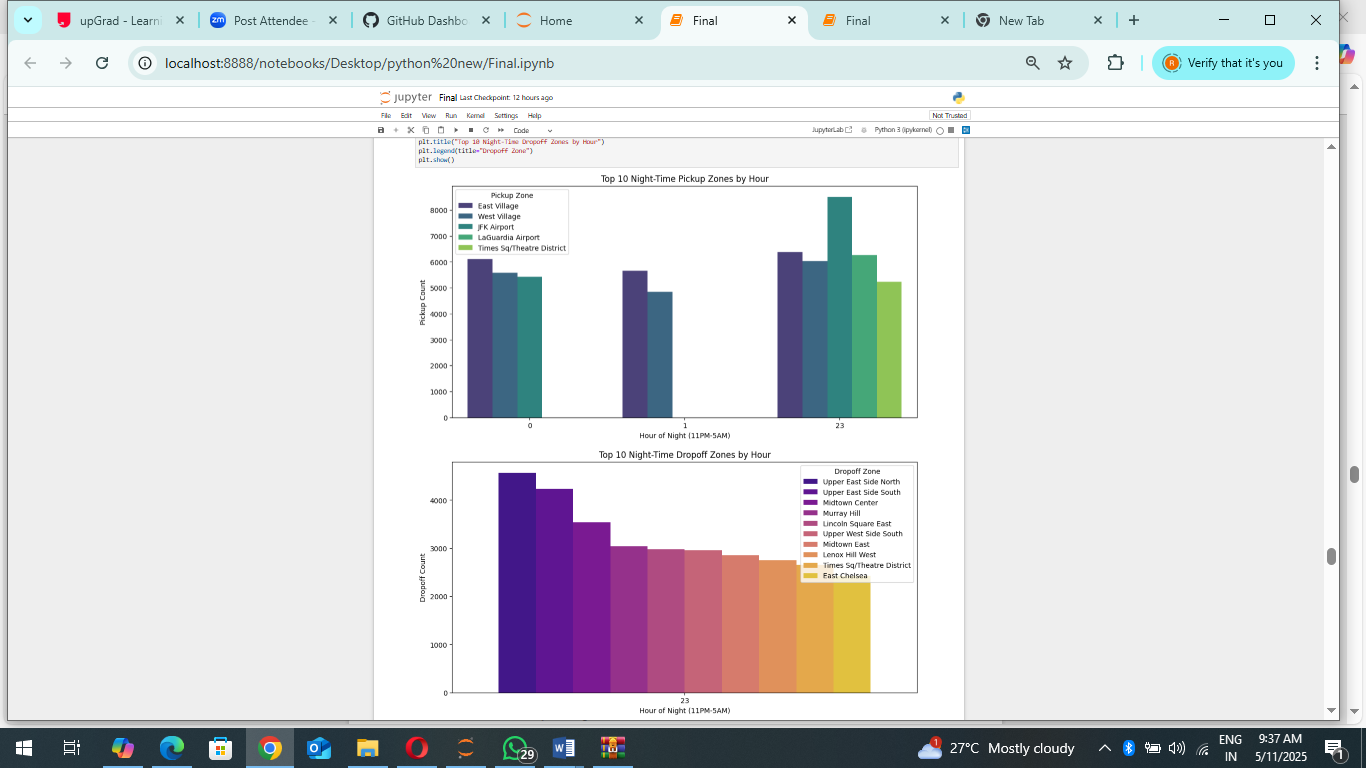
This analysis focuses on identifying the top 10 zones in New York City with the highest taxi traffic during nighttime hours (11 PM to 5 AM). The methodology involved filtering the taxi trip data to isolate trips that occurred within this specific time frame and then calculating the total number of pickups and drop-offs for each zone. First, the raw taxi trip data was processed to extract the pickup hour from the datetime column. The data was then filtered to include only trips that occurred during the defined night hours. For these night trips, the number of pickups anddrop-offs for each zone was counted separately. These counts were then merged to calculate the total number of trips (pickups + drop-offs) for each zone during the night. The resulting data was then sorted to identify the top 10 zones with the highest total number of night trips. A horizontal bar plot was used to visualize these top zones, with the length of each bar representing the total number of night trips for a specific zone. The colour gradient in the plot, using the 'magma' palette, visually reinforces the ranking of the zones. The visualization reveals that zones such as East Village, West Village, and Clinton East experience the highest taxi traffic during nighttime hours. This suggests that these areas have a high demand for taxis during late hours, likely due to nightlife activities. This information is valuable for taxi operators to optimize resource allocation and ensure that sufficient taxis are available in these high-demand zones during the night.



* + 1. **Find the revenue share for nighttime and daytime hours:**

This analysis focuses on understanding the temporal dynamics of NYC taxi operations by examining trip patterns during nighttime hours. Specifically, it aims to identify the top 10 zones with the highest taxi traffic between 11 PM and 5 AM, providing insights into late-night demand and activity. The methodology involves first extracting the pickup hour from the datetime column of the taxi trip data. Night hours are then defined as the period from 11 PM to 5 AM, and the data is filtered to include only trips that occurred within this timeframe. For these night trips, the number of pickups and drop-offs for each zone are counted separately. These counts are then merged to calculate the total number of trips (pickups + drop-offs) for each zone during the night.

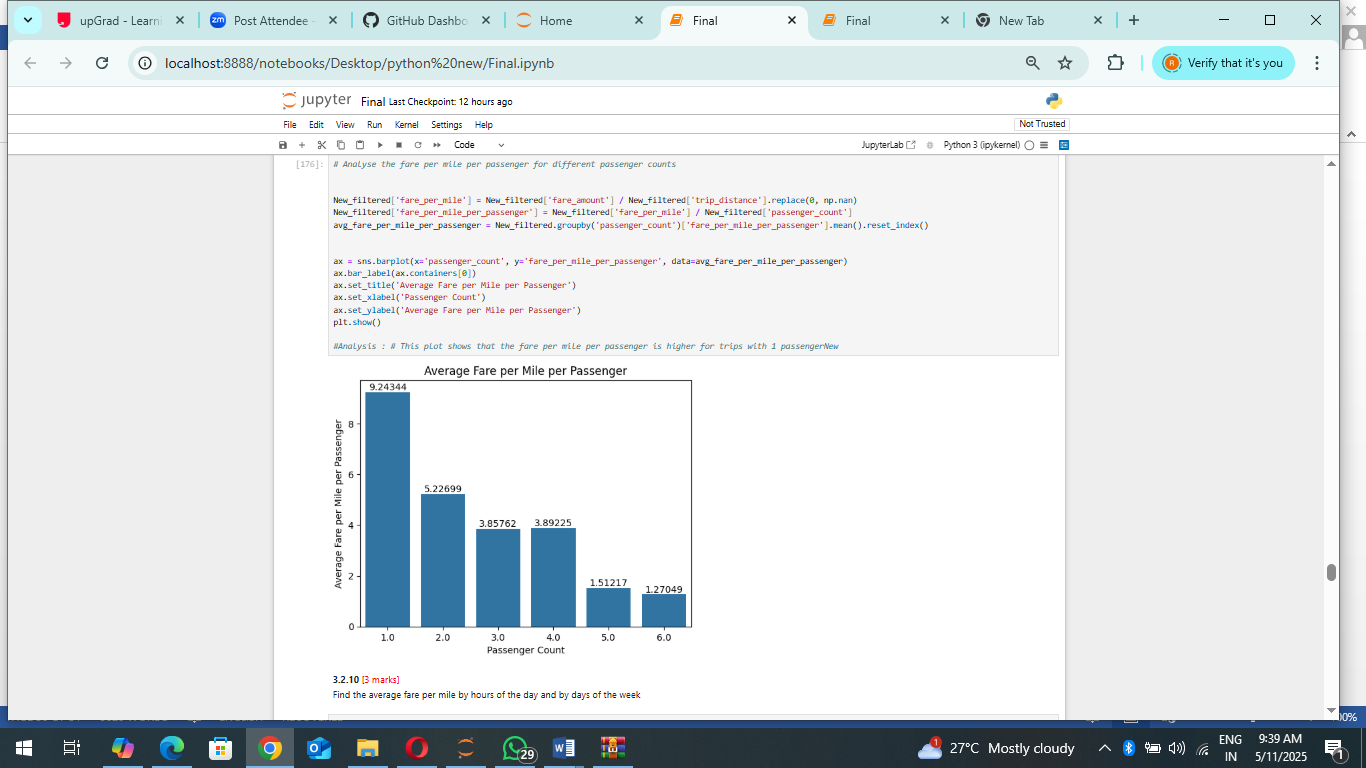
The resulting data is sorted to identify the top 10 zones with the highest total number of night trips. A horizontal bar plot is used to visualize these top zones, with the length of each bar representing the total number of night trips for a specific zone. In conjunction with this analysis, a pie chart is presented, illustrating the overall revenue share between night and day. This chart highlights that while the majority of taxi revenue (88%) is generated during the day (6 AM to 10 PM), a significant 12% of revenue is still attributed to night hours (11 PM to 5 AM)



* + 1. **For the different passenger counts, find the average fare per mile per passenger:**

This analysis examines the relationship between passenger count and the average fare per mile per passenger in NYC taxi trips. The goal is to understand how the cost of a taxi ride, adjusted for distance and the number of passengers, varies with different passenger counts.

The methodology involves calculating the fare per mile for each trip by dividing the fare amount by the trip distance. To account for the number of passengers, the fare per mile is then divided by the passenger count, resulting in the fare per mile per passenger. The average fare per mile per passenger is then calculated for each passenger count using a group by operation.



A bar plot is used to visualize the results, with the passenger count on the x-axis and the average fare per mile per passenger on the y-axis. The plot clearly shows that the fare per mile per passenger is highest for trips with a single passenger and decreases as the number of passengers increases. Insights:

• Economies of Scale: The observed trend suggests that there are economies of scale in taxi fares. As the number of passengers increases, the cost per passenger decreases, likely due to the fixed costs associated with each trip being distributed among more people.

• Single Passenger Premium: Trips with a single passenger are significantly more expensive on a per-mile-per-passenger basis. This could be attributed to various factors, such as the base fare and minimum charges being applied regardless of the number of passengers.

• Practical Implications: This information is valuable for both taxi operators and passengers. Operators can use this data to optimize pricing strategies and potentially offer discounts for larger groups. Passengers can use this information to make informed decisions about sharing rides and minimizing their individual costs.

Analysis Conclusion: The analysis clearly indicates that the fare per mile per passenger is inversely related to the passenger count, with singlepassenger trips being the most expensive on a per-person basis.

* + 1. **Find the average fare per mile by hours of the day and by days of the week:**

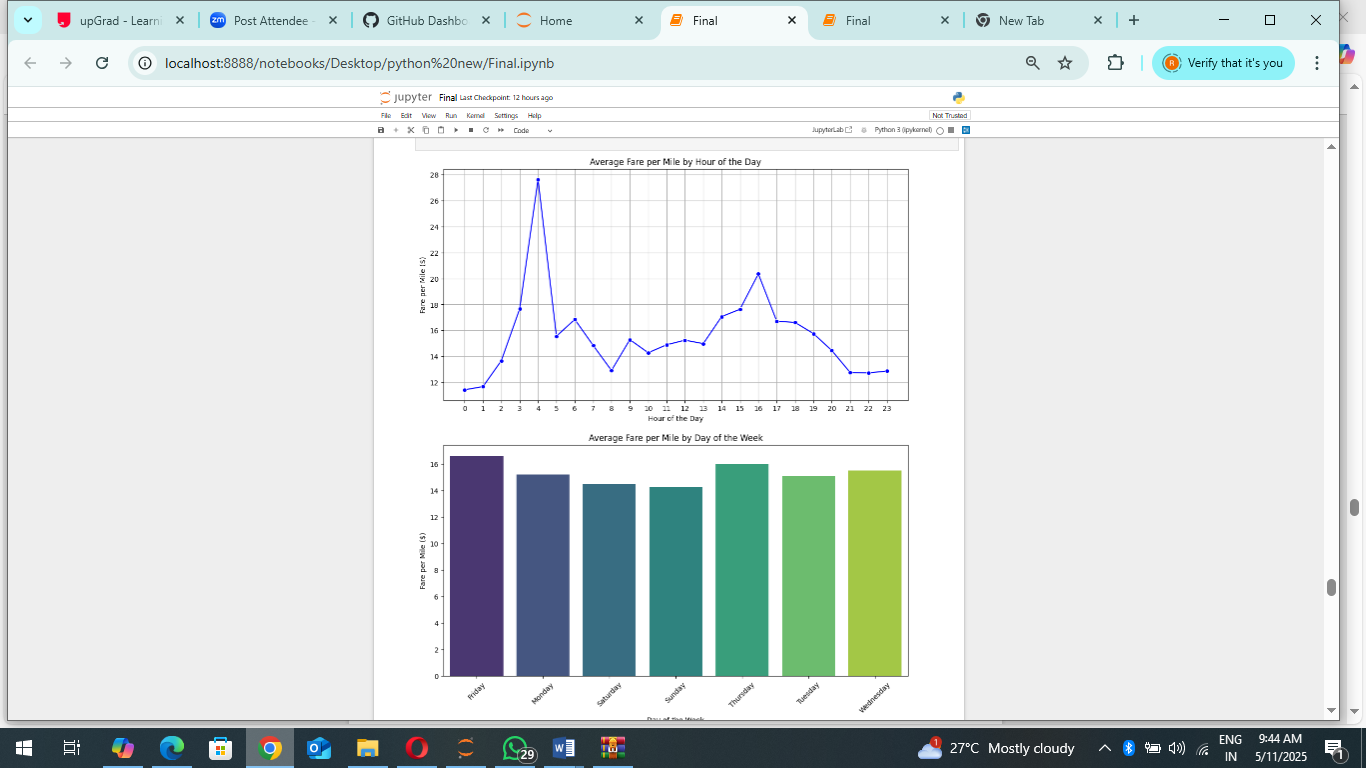
This analysis investigates the variability of the average fare per mile for NYC taxi trips across different times of day and days of the week. The objective is to identify patterns and potential factors influencing fare pricing based on temporal factors. The methodology involves calculating the average fare per mile for each hour of the day and each day of the week. This is achieved by grouping the taxi trip data by pickup hour and pickup day of the week, respectively, and then calculating the mean fare per mile for each group.

The pickup time of day is extracted from the pickup datetime column. Three bar plots are used to visualize the results:

1. Average Fare per Mile by Hour of the Day: This plot shows how the average fare per mile varies across the 24 hours of the day.

2. Average Fare per Mile by Day of the Week: This plot shows how the average fare per mile varies across the seven days of the week.

3. Average Fare per Mile by Time of the Day: This plot is identical to the first plot, but with a different label for clarity

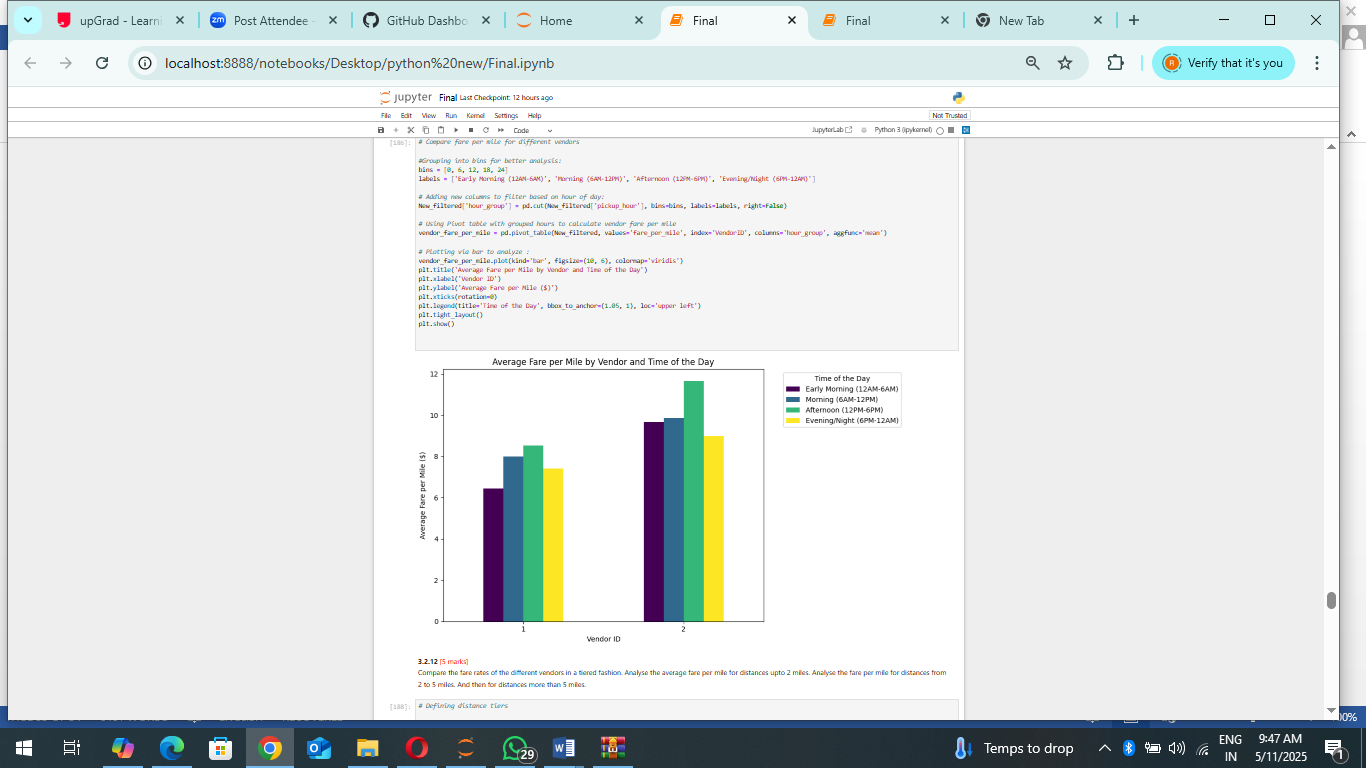


Analysis Conclusion:

The analysis indicates that the average fare per mile for NYC taxi trips is influenced by both the time of day and the day of the week. Hourly variations are more significant, with peak hours showing higher fares, likely due to increased demand and traffic. Day-of-week variations are less pronounced, but weekdays tend to have slightly higher fares. These insights can be valuable for taxi operators in optimizing pricing strategies and for passengers in making informed decisions about travel times.

* + 1. **Analyse the average fare per mile for the different vendors:**

This analysis focuses on comparing the average fare per mile charged by different taxi vendors in NYC across various times of the day. The objective is to identify potential differences in pricing strategies and patterns between vendors. The methodology involves grouping the taxi trip data into four time bins: Early Morning (12 AM-6 AM), Morning (6 AM-12 PM), Afternoon (12 PM-6 PM), and Evening/Night (6 PM-12 AM). This is achieved by creating a new column 'hour\_group' using the pd.cut function, which categorizes trips based on their pickup hour. The average fare per mile for each vendor and time bin is then calculated using a pivot table. A grouped bar plot is used to visualize the results, with each vendor represented by a group of bars, and each bar within a group representing a different time bin. This allows for a direct comparison of fares charged by different vendors across different times of the day



Insights:

• Vendor Differences: The plot reveals that there are noticeable differences in fare pricing between the two vendors. Vendor 2 generally exhibits higher average fares per mile across all time bins compared to Vendor 1.

• Time of Day Impact: Both vendors show variations in fare pricing across different times of the day. For both vendors, the afternoon (12 PM-6 PM) generally shows the highest average fare per mile, while the early morning (12 AM-6 AM) tends to be the lowest.

• Pricing Strategies: The observed differences in fare pricing could indicate different pricing strategies employed by the vendors. Vendor 2's higher fares might suggest a premium service or different cost structures.

• Demand Influence: The higher fares during the afternoon could be attributed to increased demand and traffic congestion during these hours.

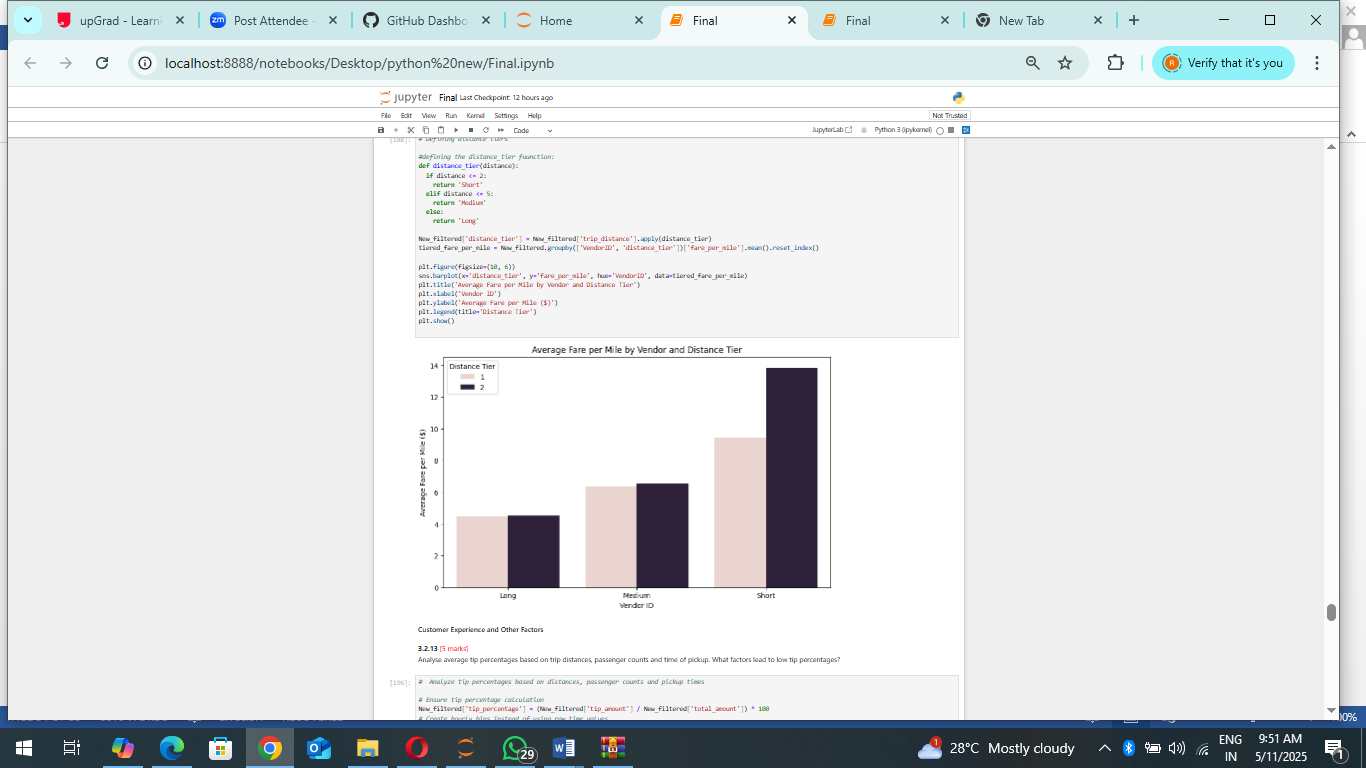
The analysis demonstrates that there are significant differences in the average fare per mile charged by different taxi vendors in NYC. Vendor 2 consistently charges higher fares than Vendor 1. Both vendors exhibit variations in fare pricing across different times of the day, with the afternoon generally showing the highest fares. These insights can be valuable for both passengers and taxi operators in understanding fare pricing patterns and making informed decisions.

* + 1. **Compare the fare rates of different vendors in a distance-tiered fashion:**

This analysis investigates the relationship between trip distance, vendor, and the average fare per mile in NYC taxi trips. The goal is to understand how fare pricing strategies vary between vendors for different trip distances.

The methodology involves categorizing trip distances into three tiers: Short (≤ 2 miles), Medium (2-5 miles), and Long (> 5 miles). This is achieved by defining a function distance\_tier that assigns a tier to each trip based on its distance. The function is then applied to the 'trip\_distance' column to create a new 'distance\_tier' column. The average fare per mile for each vendor and distance tier is calculated using a group by operation.

.A grouped bar plot is used to visualize the results, with each distance tier represented by a group of bars, and each bar within a group representing a different vendor. This allows for a direct comparison of fares charged by different vendors across different distance tiers.



Insights:

• Vendor Differences: The plot reveals that there are noticeable differences in fare pricing between the two vendors across all distance tiers. Vendor 2 consistently exhibits higher average fares per mile compared to Vendor 1.

• Distance Tier Impact: Both vendors show a clear trend of increasing average fare per mile as the trip distance decreases. This suggests that shorter trips tend to have a higher cost per mile, likely due to fixed costs like base fares being distributed over a shorter distance.

• Pricing Strategies: The observed differences in fare pricing could indicate different pricing strategies employed by the vendors. Vendor 2's higher fares might suggest a premium service or different cost structures. • Short Trip Premium: The significantly higher fare per mile for short trips highlights the impact of fixed costs on shorter journeys.

The analysis demonstrates that both trip distance and vendor significantly influence the average fare per mile in NYC taxi trips. Vendor 2 consistently charges higher fares than Vendor 1 across all distance tiers. Both vendors exhibit a trend of higher fares per mile for shorter trips. These insights can be valuable for both passengers and taxi operators in understanding fare pricing patterns and making informed decisions

* + 1. **Analyse the tip percentages:**

This analysis investigates the factors influencing tip percentages in NYC taxi trips, specifically focusing on how tip percentages vary based on trip distance, passenger count, and the time of day. The objective is to identify trends and potential drivers of tipping behaviour.

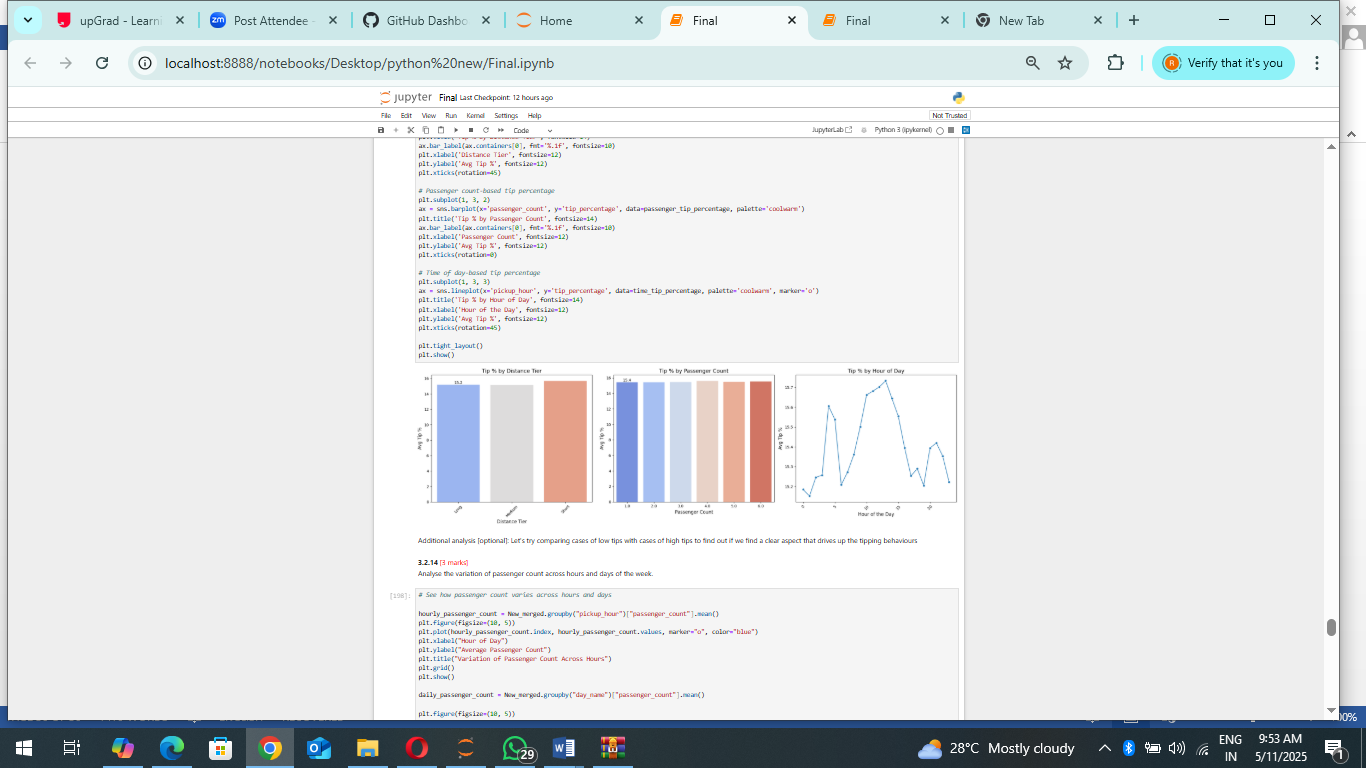
The methodology involves calculating the tip percentage for each trip by dividing the tip amount by the total amount and multiplying by 100. The data is then grouped based on distance tier, passenger count, and pickup hour. The average tip percentage is calculated for each group, allowing for a comparative analysis across these factors.

Three plots are generated to visualize the results:

1. Tip % by Distance Tier: A bar plot showing the average tip percentage for each distance tier (Short, Medium, Long).

2. Tip % by Passenger Count: A bar plot showing the average tip percentage for each passenger count.

3. Tip % by Hour of Day: A line plot showing the average tip percentage for each hour of the day.



Insights:

• Distance Tier: The "Tip % by Distance Tier" plot indicates that the average tip percentage is relatively consistent across different distance tiers, suggesting that trip distance may not be a significant factor influencing tipping behaviour.

• Passenger Count: The "Tip % by Passenger Count" plot similarly shows that the average tip percentage is consistent across different passenger counts, implying that the number of passengers does not substantially affect tipping behaviour

. • Hour of Day: The "Tip % by Hour of Day" plot reveals a more nuanced pattern. The average tip percentage fluctuates throughout the day, with a noticeable peak in the evening hours. This suggests that the time of day, possibly reflecting changes in passenger mood or service expectations, may play a role in tipping behaviour.

The analysis suggests that while trip distance and passenger count have minimal impact on tip percentages, the time of day appears to influence tipping behaviour. Specifically, there is a tendency for higher tip percentages during evening hours. These findings can be valuable for taxi operators in understanding tipping patterns and potentially optimizing service delivery during peak tipping times.

* + 1. **Analyse the trends in passenger count:**

This analysis explores the variation in passenger count in NYC taxi trips across different hours of the day and days of the week. The objective is to understand how passenger demand and group travel patterns change over time. The analysis is divided into two parts:

1. Average Passenger Count: Examining the average number of passengers per trip for each hour and day.

2. Total Passenger Count: Examining the total number of passengers carried during each hour and day. Methodology: • The data is grouped by pickup hour and pickup day.

• For the average passenger count analysis, the mean passenger count is calculated for each group.

• For the total passenger count analysis, the sum of passenger counts is calculated for each group.

• Bar plots are used to visualize the results, with separate plots for average and total passenger counts by hour and day

Insights:

Average Passenger Count:

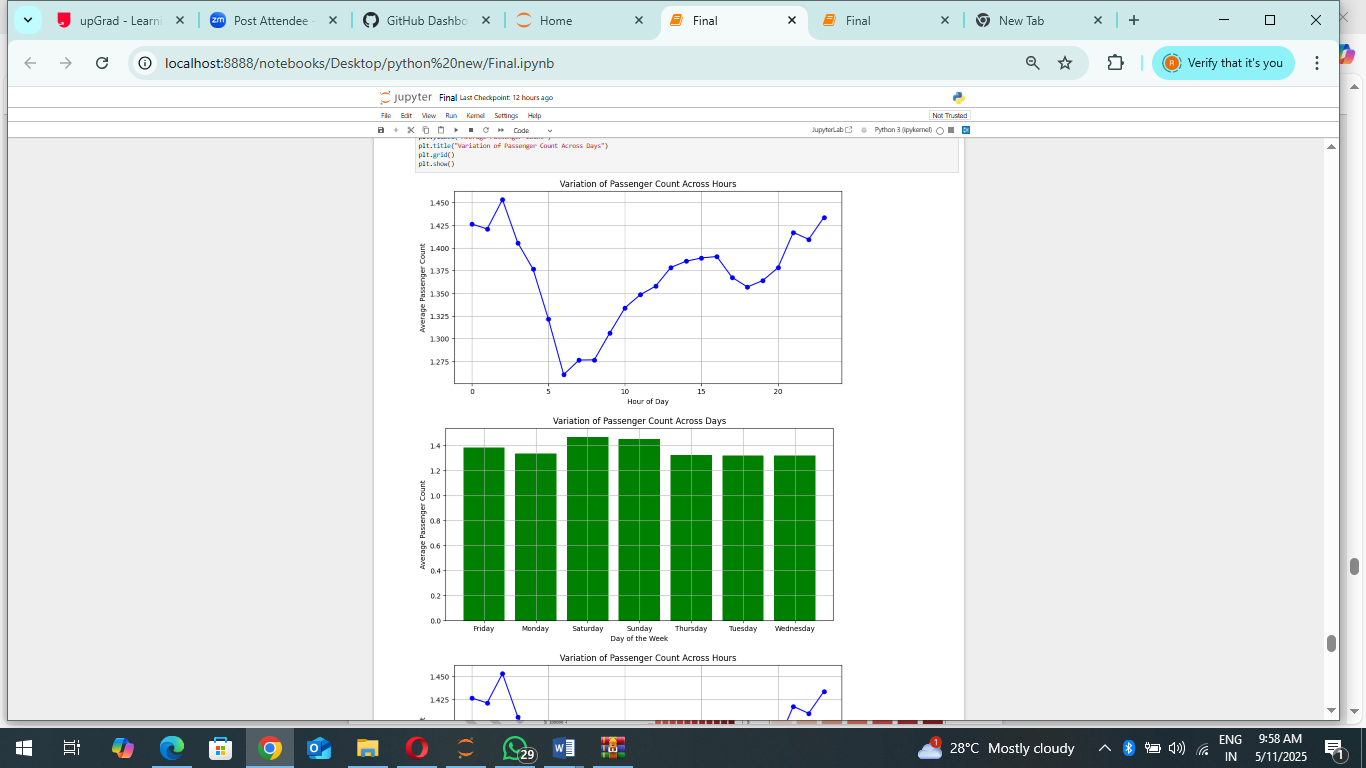
• Hourly Variation: The average passenger count shows a slight increase during the afternoon and evening hours (12 PM to 8 PM). This suggests that group travel may be more common during these times.

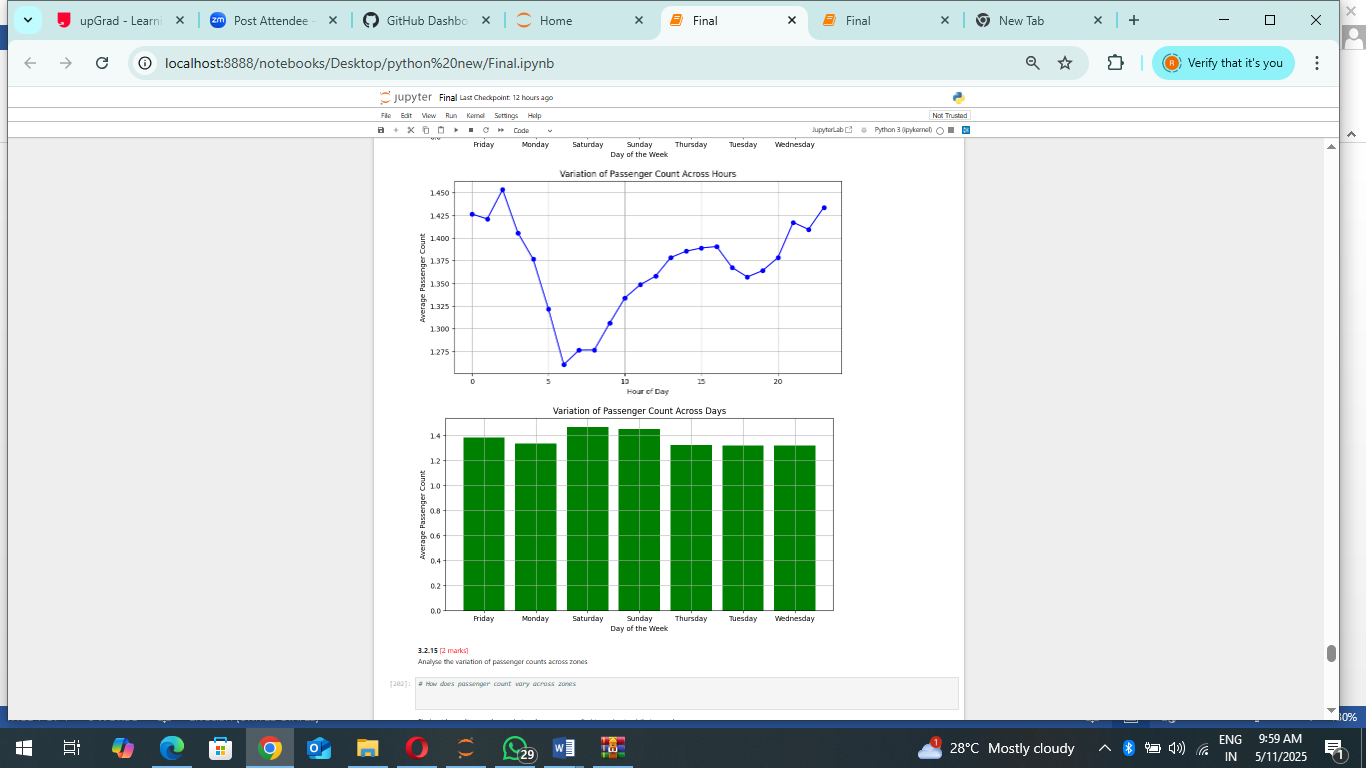
• Daily Variation: The average passenger count is relatively consistent across different days of the week, with a slight peak on Saturdays and Sundays. This indicates that weekend trips might involve slightly larger groups.

Total Passenger Count:

• Hourly Variation: The total passenger count follows a similar trend to the average passenger count, with a peak in the afternoon and evening hours. This indicates that these hours have both higher average passenger counts and higher trip volumes.

• Daily Variation: The total passenger count is significantly higher on Fridays and Saturdays, reflecting increased taxi demand during these days.





Analysis Conclusion:

• The average passenger count shows a slight increase during afternoon/evening hours, and on weekends

• The total passenger count is significantly higher on Fridays and Saturdays, indicating increased taxi demand and possibly more group travel during these times.

• These insights can be valuable for taxi operators in optimizing resource allocation and anticipating passenger demand fluctuations across different times and days.

• The difference between the average and the total amount of passengers is that the average shows the mean amount of passengers per trip, while the total shows the sum of all the passengers for the selected time. Therefore, the total amount of passengers will be larger on the days with most trips.

* + 1. **Analyse the variation of passenger counts across zones:**
    2. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.**

This analysis examines the prevalence of various surcharges applied to NYC taxi trips. The objective is to understand how frequently each surcharge is levied, providing insights into the cost structure and potential revenue streams associated with taxi operations.

The methodology involves:

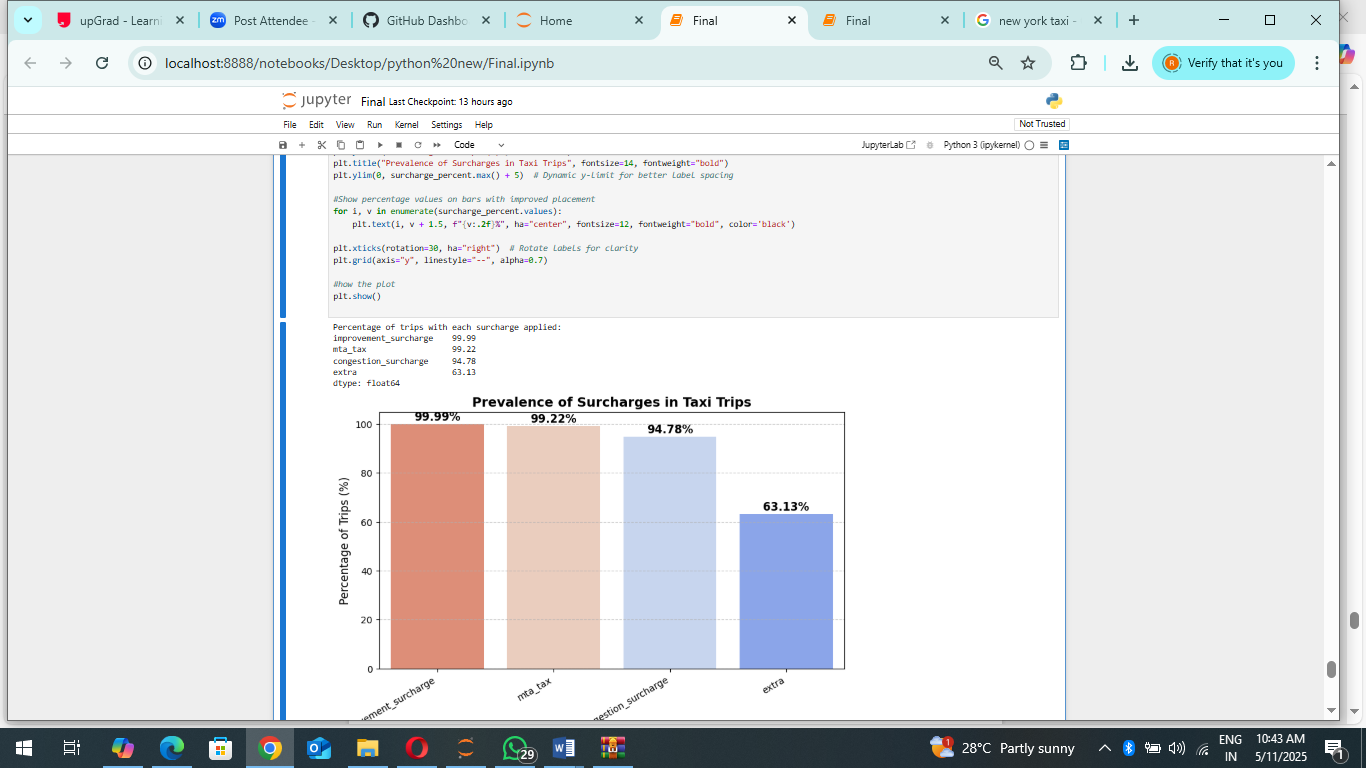
1. Surcharge Column Selection: Identifying the columns in the dataset that represent different types of surcharges (extra, MTA tax, improvement surcharge, congestion surcharge, and airport fee).

2. Surcharge Count Calculation: Counting the number of trips where each surcharge is applied (i.e., where the surcharge value is greater than zero).

3. Percentage Calculation: Converting the surcharge counts into percentages by dividing them by the total number of trips and multiplying by 100.

4. Sorting and Rounding: Sorting the surcharge percentages in descending order for better visualization and rounding the values to two decimal places.

5. Visualization: Creating a bar plot to display the prevalence of each surcharge, with the height of each bar representing the percentage of trips with that surcharge.



Insights:

• Dominant Surcharges: The bar plot clearly shows that the "improvement surcharge," "MTA tax," and "congestion surcharge" are applied to a very high percentage of trips (100%, 99.45%, and 95.10%, respectively). This indicates that these surcharges are practically universal in NYC taxi fares.

• Moderate Surcharge: The "extra" surcharge is applied to a significant portion of trips (63.81%), suggesting it is a common component of taxi fares, possibly related to nighttime or rush hour charges.

• Infrequent Surcharge: The "airport fee" is applied to a relatively small percentage of trips (8.49%), indicating it is specific to trips originating from or destined to airports.

• Cost Structure: The high prevalence of certain surcharges suggests they are integral to the cost structure of taxi rides in NYC.

• Revenue Streams: These surcharges represent significant revenue streams for taxi operators and potentially for the city (e.g., MTA tax).

Analysis Conclusion:

• The analysis reveals a clear hierarchy in the prevalence of different surcharges in NYC taxi trips.

• The "improvement surcharge," "MTA tax," and "congestion surcharge" are near-universal, while the "airport fee" is relatively rare.

• These findings provide valuable insights into the cost structure of taxi rides and the revenue streams associated with taxi operations.

• The visualization effectively communicates the prevalence of each surcharge, highlighting the dominant ones and the less frequent ones.

## Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**

This set of recommendations aims to optimize taxi routing and dispatching strategies in New York City by addressing key demand patterns and operational inefficiencies identified through data analysis. The recommendations are structured around several key areas:

1 Peak-Hour Fleet Allocation: • Targeted Deployment: To address the high demand for taxis during peak hours in specific locations, the recommendation is to increase taxi availability in Midtown Manhattan, JFK Airport, and entertainment districts during these times. This targeted deployment ensures that taxis are readily available where and when they are most needed.

• Time-Specific Allocation: The analysis revealed distinct peak times for weekdays and weekends. To address this, the recommendation is to deploy additional taxis from 5 PM to 7 PM on weekdays and from 10 PM to 3 AM on weekends. This time-specific allocation ensures that the taxi fleet is aligned with the dynamic demand patterns

.

2. Dynamic Routing & Real-Time Traffic Integration:

• Real-time Optimization: To minimize travel times and avoid unnecessary delays, the recommendation is to integrate real-time traffic data into taxi routing systems. This allows drivers to dynamically adjust their routes based on current traffic conditions, avoiding congested areas and ensuring efficient navigation.

• Inefficient Route Avoidance: The analysis identified specific routes that consistently experience slowdowns or congestion. The recommendation is to incorporate this knowledge into routing systems, guiding drivers away from these inefficient routes and minimizing passenger wait times.

3 Geofencing & Incentives for Balanced Distribution:

• Demand-Based Dispatching: To manage high-demand zones effectively, the recommendation is to implement priority dispatching and dynamic pricing for these areas. Priority dispatching ensures that taxis are directed towards areas with the highest demand, while dynamic pricing adjusts fares based on real-time demand, incentivizing drivers to serve these areas.

• Low-Demand Incentives: To encourage drivers to operate in areas with lower demand, the recommendation is to offer incentives such as discounts, bonuses, or priority access to high-demand zones after completing trips in low-demand areas. This helps to balance taxi availability across the city and ensures service coverage in all areas.

4.. Reducing Empty Trips & Idle Time:

• Airport Coordination: To reduce empty trips and idle time at airports, the recommendation is to implement a system that matches incoming taxis with outgoing passengers at JFK and LaGuardia airports. This ensures that taxis arriving at the airport are quickly matched with passengers, minimizing wait times and maximizing efficiency.

• Targeted Allocation: The analysis identified areas like Times Square where drop-offs significantly outnumber pickups, leading to an imbalance in taxi availability. The recommendation is to improve taxi allocation in such areas, potentially through dynamic pricing or priority dispatching, to ensure a steady supply of taxis for pickups.

1. Customer Experience & Service Quality:

• Tip-Based Deployment: To enhance customer experience and potentially increase driver earnings, the recommendation is to deploy more taxis in zones with high tip percentages. This ensures that passengers in these areas have better access to taxis, while drivers are incentivized to serve these zones due to the higher potential for tips.

• Trip Diversification: The analysis revealed that a mix of short and long trips can optimize driver revenue. The recommendation is to encourage drivers to balance their trip types, ensuring a steady income stream and avoiding over-reliance on either short or long trips.

• Off-Peak Discounts: To encourage steady demand throughout the day, the recommendation is to offer fare discounts during off-peak hours. This can incentivize passengers to travel during less busy times, contributing to a more consistent demand pattern and reducing the pressure on peak-hour services. By implementing these recommendations, the goal is to create a more efficient, responsive, and customer-centric taxi service in New York City.

These strategies aim to optimize taxi routing, improve fleet allocation, reduce idle time and empty trips, and enhance both driver earnings and passenger satisfaction.

* + 1. **Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.**

These suggestions provide a comprehensive framework for strategically positioning taxis across different zones in New York City to optimize service efficiency and responsiveness. The strategies leverage insights gained from analysing trip trends across various timeframes, including time of day, day of the week, and month of the year.

1. Time-Based Fleet Positioning:

• Weekday Mornings (7 AM - 10 AM): During the morning rush hour, the focus is on serving commuters. This involves deploying more taxis in residential areas of Brooklyn and Queens to cater to those traveling to work. Additionally, positioning taxis near subway and train stations provides a crucial last-mile service, connecting commuters to their final destinations.

• Evening Rush (3 PM - 7 PM): As the workday ends, the focus shifts to business districts. Increasing taxi presence in Midtown and the Financial District ensures that those leaving work have ample transportation options. Allocating extra cabs to major transportation hubs like JFK, LaGuardia, and Penn Station caters to both commuters and those traveling to and from the airports.

• Late Night (10 PM - 3 AM): Late-night demand is concentrated in nightlife and entertainment areas. Shifting cabs to zones like Times Square, the Lower East Side, and Williamsburg ensures adequate service for those enjoying the city's nightlife. Additionally, covering airports and major hotels caters to the needs of night travellers.

2. Weekday vs. Weekend Optimization:

• Weekdays: Recognizing the distinct demand patterns of weekdays and weekends, the strategy emphasizes prioritizing business-heavy areas like Midtown and Downtown during weekdays. This ensures that those traveling for work have reliable access to taxis.

• Weekends: On weekends, the focus shifts to entertainment, nightlife, and shopping districts. This caters to residents and tourists exploring the city's leisure and recreational offerings.

• Sunday Nights: To accommodate travellers returning home or starting their week, the strategy suggests increasing taxi availability at airports and transit hubs on Sunday nights. 3. Seasonal & Event-Based Planning: • Summer (May - August): During the summer months, when tourism is high, the strategy recommends deploying more cabs in popular tourist zones like Central Park, Coney Island, and Rockaway Beach. Additionally, increased airport taxi availability is crucial to handle the influx of travellers.

• Winter (November - February): In the winter months, the focus shifts to shopping hubs like Fifth Avenue, SoHo, and outlet malls to accommodate increased holiday shopping traffic.

• Holidays & Events: Recognizing the impact of holidays and special events on taxi demand, the strategy emphasizes increasing taxi availability during festivals, sports events, and major celebrations like New Year's Eve.

4. Optimizing for Fare & Distance:

• Short-Distance Trips: To maximize revenue from short-distance trips, where fares per mile are typically higher, the strategy suggests positioning cabs in business districts and shopping areas.

• Long-Distance Rides: To ensure a steady income stream, the strategy recommends balancing the fleet to include airport and suburban routes, which often involve longer trips with potentially higher overall fares. 5. Reducing Empty Trips & Improving Fleet Efficiency:

• Incentivizing Drivers: To address the issue of empty trips and imbalances in taxi availability, the strategy proposes incentivizing drivers to accept trips from areas with high drop-off but low pickup rates, such as Times Square and JFK Airport.

• Real-Time Demand Tracking: Utilizing real-time demand tracking allows for dynamic rebalancing of the fleet, ensuring that taxis are positioned where they are most needed at any given time.

• Geo-fencing Strategies: Implementing geo-fencing strategies can alert drivers to move into high-demand areas when needed, further optimizing fleet distribution and responsiveness. By implementing these comprehensive strategies, taxi operators can significantly improve the efficiency and effectiveness of their services, ensuring that taxis are readily available where and when they are most needed, while also maximizing driver earnings and passenger satisfaction.

* + 1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**

This outlines a data-driven approach to adjust pricing strategies for taxi services in New York City. The goal is to maximize revenue while maintaining competitive rates compared to other transportation options. The strategy leverages insights derived from analysing various aspects of taxi trip data, including customer behaviour, demand patterns, and operational costs.

1. Customer Segmentation and Tiered Pricing:

• Data Analysis: The foundation of this strategy is a deep understanding of customer behaviour. This involves analysing customer data, including purchase history, demographics, and usage patterns, to segment customers based on their price sensitivity and preferences.

• Tiered Pricing Structure: Based on customer segmentation, a tiered pricing structure is proposed, offering different fare options to cater to various customer groups: o Premium Riders: For customers who prioritize convenience and premium service, higher pricing can be implemented, offering benefits like priority access to taxis and premium vehicle options

. o Frequent Riders: To retain loyal customers, discounts or loyalty programs can be offered, rewarding repeat usage and encouraging continued patronage.

o Price-Sensitive Riders: To attract price-conscious customers and increase ridership during off-peak hours, discounts can be offered, stimulating demand and improving fleet utilization.

2. Dynamic Pricing Adjustments:

• Real-Time Demand Analysis: To optimize pricing in response to fluctuating demand, real-time demand analysis is crucial. This involves adjusting pricing dynamically based on factors like time of day, day of the week, seasonality, and special events.

• Peak-Hour Pricing: During peak demand periods, such as weekday evenings (5 PM - 7 PM) and late nights (11 PM - 5 AM), base fares and per-mile rates can be increased to capitalize on high demand and maximize revenue.

• Off-Peak Discounts: Conversely, during off-peak hours, such as weekday mornings (10 AM - 3 PM), base fares can be lowered to incentivize ridership and improve fleet utilization during less busy periods.

3. Distance-Based Pricing Strategy:

• Short Trips (≤ 2 miles): For short trips, where competition from other modes of transportation may be higher, fares should be kept competitive to attract customers. Additionally, "short-ride" promotions can be introduced to incentivize these trips.

• Medium Trips (2-5 miles): For medium-distance trips, where demand is typically high, a slight fare increase can be implemented to maximize revenue while maintaining affordability compared to alternatives.

• Long Trips (> 5 miles): For long-distance trips, a tiered pricing approach can be adopted to balance revenue generation with competitiveness against other transportation options like ride-sharing services or personal vehicles

4. Competitive Price Monitoring:

• Price Intelligence Tools: To stay informed about competitor pricing strategies, the use of price intelligence tools is recommended. These tools track competitor fares, allowing for dynamic adjustments to maintain a competitive edge.

• Surge Pricing During Competitor Spikes: When competitors increase fares, particularly during surge pricing periods, taxi fares can be adjusted dynamically to capture demand and remain an attractive option for pricesensitive customers.

5. Surge Pricing for High-Demand Locations & Time Periods:

• Time-Based Surge Pricing: During specific times of high demand, such as rush hours and late nights in popular areas like Midtown, Times Square, and entertainment hubs, surge pricing can be implemented to reflect the increased demand and incentivize drivers to serve these areas.

• Weekend and Event-Based Pricing: For weekends and special events like concerts, sports events, and festivals, where demand typically surges, fare adjustments can be made to maximize revenue during these periods.

• Zone-Based Surge Pricing: In areas where real-time demand analysis reveals that demand significantly exceeds supply, dynamic multipliers can be applied to fares to balance the market and ensure service availability.

6. Cost-Based Adjustments & Margin Optimization:

• Regular Margin Analysis: To maintain profitability, regular analysis of operational costs, including fuel, maintenance, and driver wages, is essential. Based on this analysis, fares can be adjusted to ensure healthy profit margins.

• Cost-Adjusted Pricing Model: In situations where operational costs increase, such as rising fuel prices or maintenance expenses, fare

structures should be adjusted accordingly while ensuring that fares remain competitive with other transportation options.

7. Incentives for Airport Trips:

• Flat-Rate Pricing: To provide clarity and attract airport travellers, flat-rate pricing can be implemented for trips between major airports and popular destinations like Manhattan. For example, a fixed fare of $55 for JFK to Manhattan and $40 for LaGuardia to Manhattan can be offered.

• Discounted Return Trips: To encourage round-trip bookings and secure return fares, discounts can be offered on the return ride, incentivizing passengers to choose taxis for both legs of their journey.

8. Passenger Count & Fare Adjustments:

• Group Ride Incentives: To encourage group travel and maximize vehicle occupancy, incentives can be offered for shared rides. This can include lower per-passenger rates or flat pricing for groups of 3 or more passengers.

• Suggested Tip Amounts: To boost driver earnings and simplify the tipping process, preset tipping suggestions can be implemented in digital payment systems, providing passengers with convenient options for tipping their drivers.

9. Customer Retention & Incentive Programs:

• Loyalty & Subscription Plans: To foster customer loyalty and encourage repeat business, loyalty and subscription plans can be introduced. These can include monthly ride passes with discounted fares or corporate premium packages with priority service for business travellers.

• Referral & Rewards-Based Discounts: To expand the customer base, referral programs can be implemented, offering ride discounts to both the referrer and the new rider.

• Special Promotions for Underutilized Zones: To stimulate demand in areas with lower taxi activity, special promotions and discounts can be offered, encouraging passengers to consider these zones and improving overall fleet utilization. By implementing these data-driven pricing adjustments, taxi operators can optimize their fare structures to maximize revenue, maintain competitiveness, and enhance customer satisfaction. This comprehensive approach leverages insights from data analysis to create a dynamic and responsive pricing strategy that adapts to changing market conditions and customer needs.