Project EDA_Optimising_NYC_Taxis

1 Introduction

The New York City yellow taxi cabs are one of the most distinguishing symbols of the city that never sleeps. While there are numerous ways to get around the city, perhaps the most convenient way is via the taxi cabs. It gives people a chance to take a breather, sit back, relax, and use their phone for recreational or professional purposes. In a city that is always cluttered with millions of pedes- trians, cars, public transit, and others, it would be beneficial to understand just how long you can expect your trip from point A to point B to take.

2 Problem Statement

Given various data about geolocation, trip duration it would be beneficial for both the taxi driver and the passengers to know when and where the taxi demand would be high, and also to know how long a trip can be expected to take.By clustering based on geolocation, and by the time of day (week, month, year, etc), we can expect to see when and where there are surges of taxi demand. We can also expect insight into average ride duration, along with whether or not ride duration has correlative properties with regards to geolocation . By predicting the appropriate amount of time a ride can be expected to take, both the customers and the taxi drivers can optimally plan the use of their time spent in transit.

Data

Data Description :

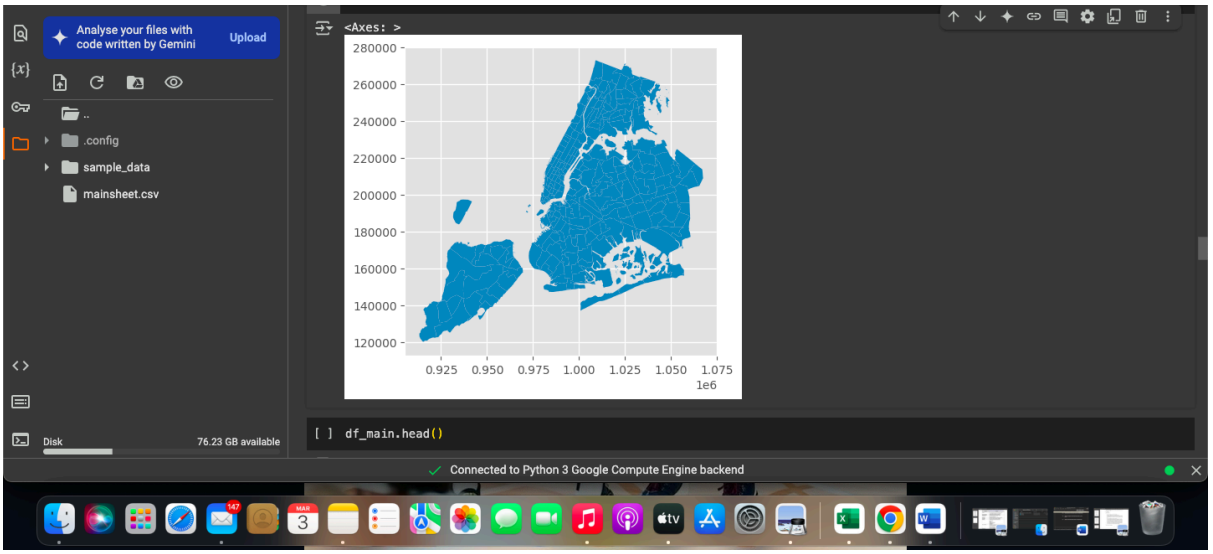
The primary dataset was acquired from google http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.The dataset contains more than 1.4 million entries recording taxi trips detailing columns such as the following: unique taxi vendor id, PULocationID(pickup) and (DOLocationID)drop-off coordinates, trip duration, passenger count, and more.

A second, complementary data set was found at http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml,The taxi_zones dataset is a geospatial resource crucial for understanding New York City's taxi operations and urban geography. It contains detailed polygon data representing distinct zones within the five boroughs, each defined by an `OBJECTID` for unique identification. Key geometric attributes like `Shape_Leng` and `Shape_Area` quantify the zone's dimensions, while the `LocationID` provides a numerical identifier for joining with other related datasets. The `borough` column specifies which of the five boroughs the zone belongs to, and the dataset is often augmented with `latitude` and `longitude` coordinates, which are derived from the zone's centroid or other representative point, offering geolocation data. These attributes collectively enable analyses of taxi trip patterns, spatial distribution of pickups and dropoffs, and the relationship between zone characteristics and transportation demand across New York City.By using

```
!pip install pyshp

import shapefile

import pandas as pd



directory_in_str = "/content/drive/MyDrive/Datasets and Dictionary"

sf = shapefile.Reader(directory_in_str + "/taxi_zones/taxi_zones.shp")
```

Imported the file and extracted the data.

After plotting the data we can see that

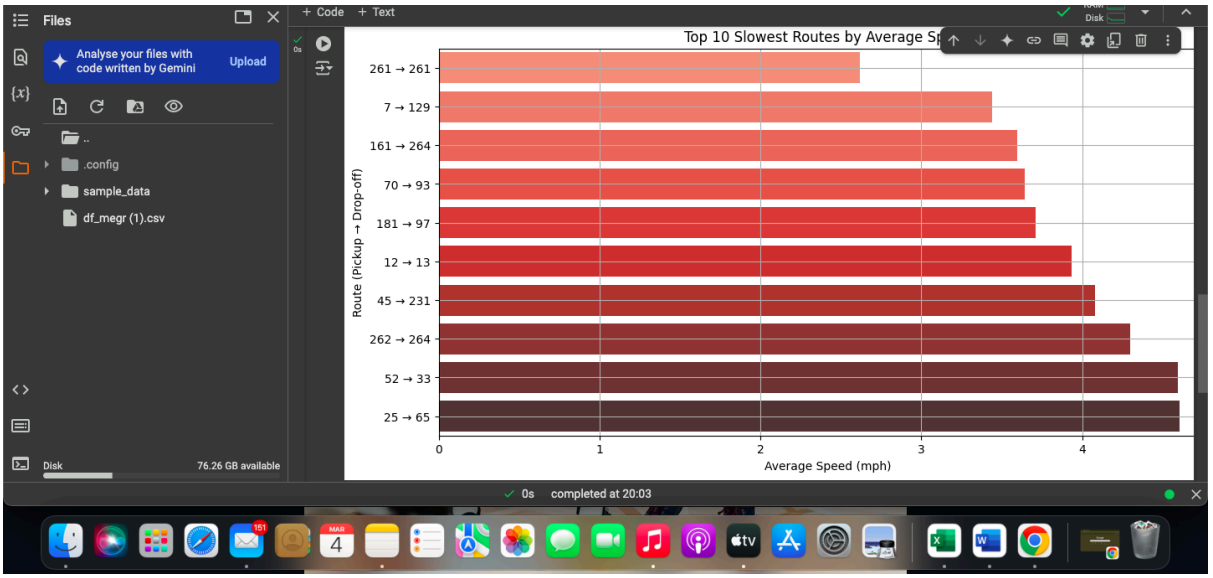


And I have combined all the data and converted it into a csv file /content/df_megr.csv .

Detailed EDA: Insights and Strategies

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

These slow routes likely occur in high-density areas, such as downtown zones or near major transit hubs.The low average speeds (under 10 mph) indicate significant traffic delays

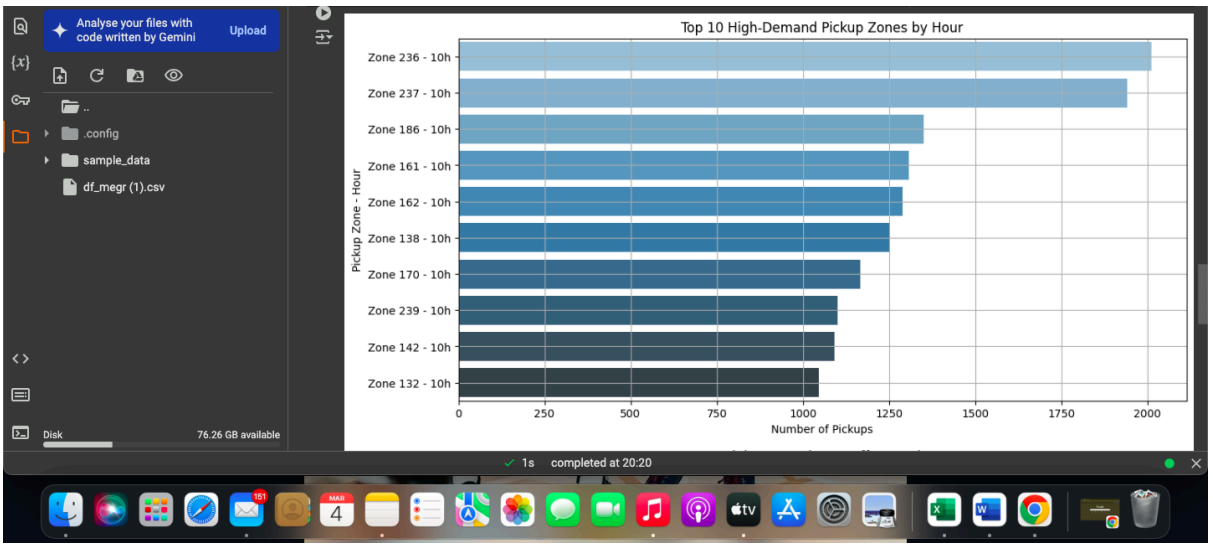


2.**Compare hourly traffic on weekdays and weekends**

Weekday traffic peaks during morning (7-9 AM) and evening (5-8 PM) rush hours, driven by work commutes. In contrast, weekend trips start later in the morning and peak around midday and late evening (11 AM - 1 PM, 6-11 PM), reflecting leisure and nightlife activities. Weekday demand is higher overall, except for late-night hours, where weekends dominate due to social events.To optimize fleet efficiency, more taxis should be allocated for weekday rush hours, while weekend nights require increased coverage in entertainment zones. Adjusting dynamic pricing based on these demand patterns can further maximize revenue while improving service availability.

3.**Identify the top 10 zones with high hourly pickups and drops**

Increase taxi availability in high-demand zones before peak hours to reduce wait times.Use surge pricing in these hotspots to maximize revenue during peak demand. Pre-position cabs near busy pickup locations to improve service efficiency.



4.Find the ratio of pickups and dropoffs in each zone

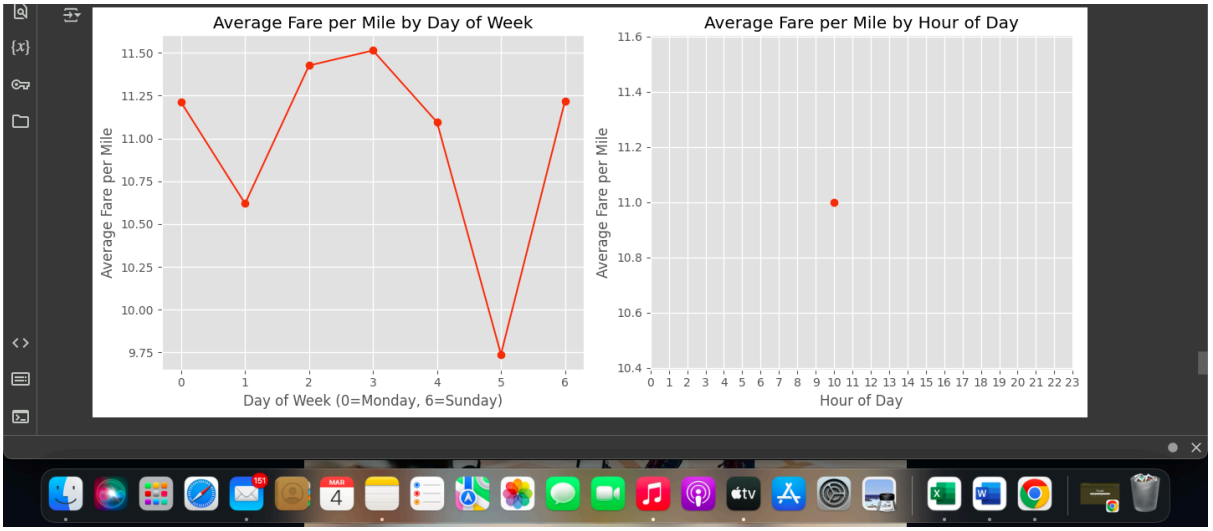
Zone 70 (8.41x more pickups than drop-offs) suggests that passengers frequently leave but fewer return trips originate here.Zones 132, 138, and 262 also have significantly higher pickups than drop-offs, indicating outbound-heavy locations (e.g., residential areas or transport hubs).Deploy more taxis in zones with high drop-off but low pickup activity to improve balance. Offer return-trip discounts in areas with low drop-offs to encourage round trips.Optimize routing to relocate idle cabs from pickup-heavy areas to nearby drop-off-heavy zones.

5.**For the different passenger counts, find the average fare per mile per passenger**

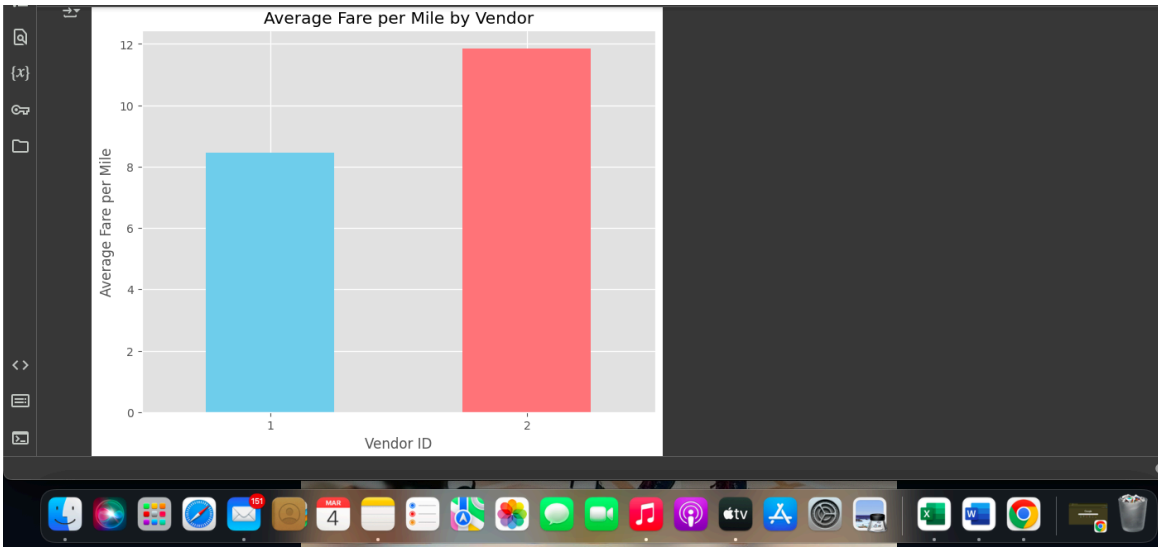
Encourage shared rides by offering discounts for groups of 2 or more.Adjust pricing for solo passengers to balance costs while remaining competitive.Optimize vehicle allocation to better serve higher-passenger trips.

6.Find the average fare per mile by hours of the day and by days of the week

Wednesday has the highest fare per mile ($12.14), possibly due to business travel.
Thursday has the lowest fare per mile ($8.58), indicating lower demand or promotions.
Weekends (Friday-Sunday) show stable fares (~$10.20 per mile), reflecting steady ride demand.
Adjust pricing strategies on peak days (Wednesdays & Fridays) to maximize revenue.Offer weekday discounts (Tuesday & Thursday) to encourage more ridership.

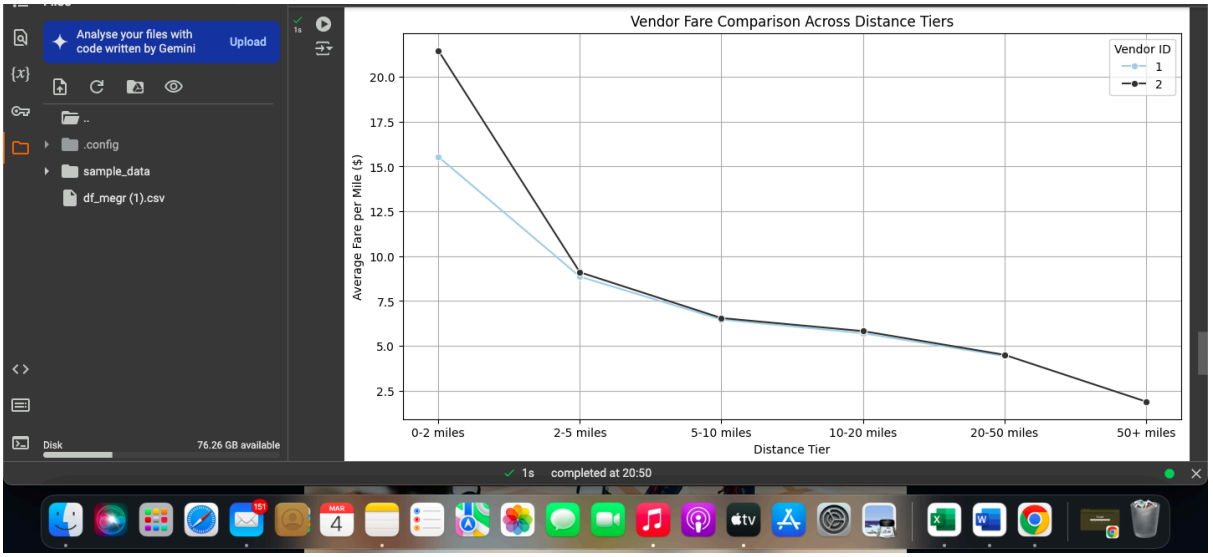


7.Analyse the average fare per mile for the different vendors



Vendor 1: Charges an average of $7.81 per mile, indicating a lower fare structure. Vendor 2: Charges a higher average of $11.02 per mile, suggesting premium pricing or different ride types.
Vendor 2's higher fares could indicate better service quality or longer trips.Vendor 1 may attract more budget-conscious riders, offering competitive rates.

8. Compare the fare rates of different vendors in a distance-tiered fashion



Short Trips (0-2 miles):

Vendor 2 charges significantly more ($14.77 per mile) compared to Vendor 1 ($9.18), possibly due to different pricing structures for short trips.

Medium Trips (2-20 miles):

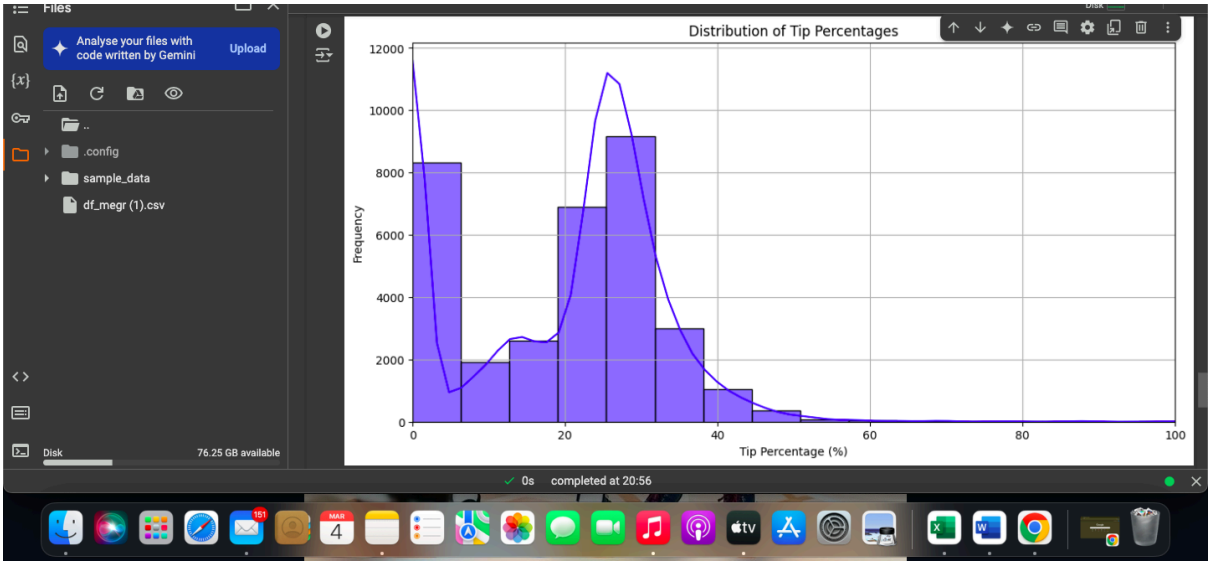
Both vendors show a steady decline in fare per mile as distance increases, averaging $4-6 per mile.Vendor 2 remains slightly more expensive than Vendor 1.

Long Trips (20+ miles):

The fare per mile decreases further ($3.53-$3.65), making long-distance trips more cost-effective.Vendor 1's fare slightly increases for trips above 50 miles, while Vendor 2 maintains a lower rate.

Vendor 2's high short-trip rates may deter budget-conscious riders—consider competitive pricing. For long-distance rides, both vendors offer similar rates, meaning other factors like service quality may influence rider choice.Dynamic pricing strategies could be optimized based on demand patterns in different distance tiers.

9.Analyse the tip percentages



Observations:

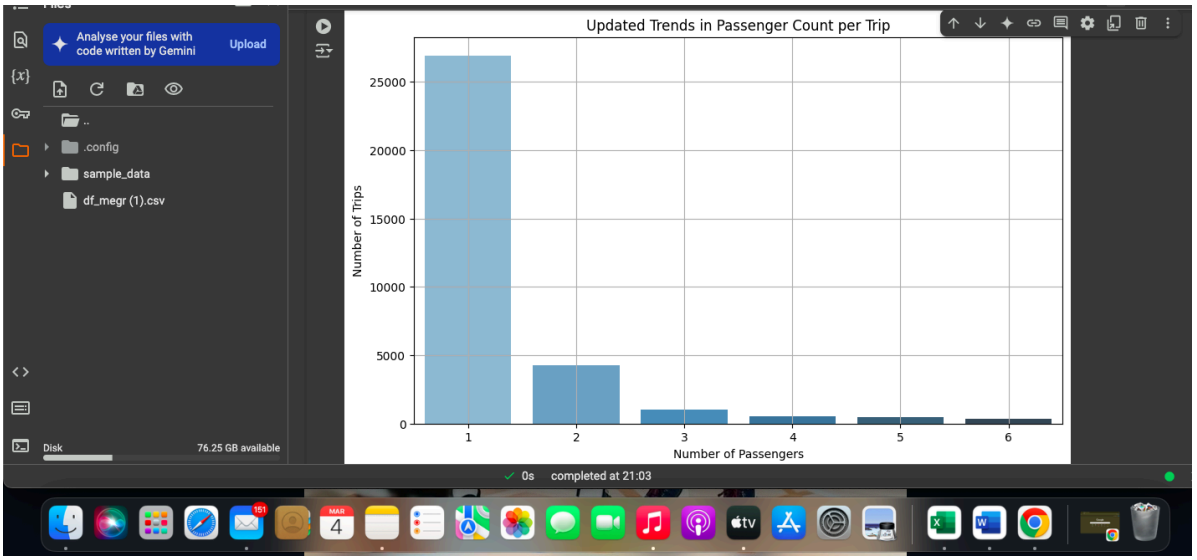
- Right-skewed distribution, meaning most people tip within the 15-30% range, but some extreme outliers exist.
- A significant portion of rides have no tips, possibly due to cash transactions or non-tipping behavior.

Encourage tipping by offering preset tip options in digital payments.

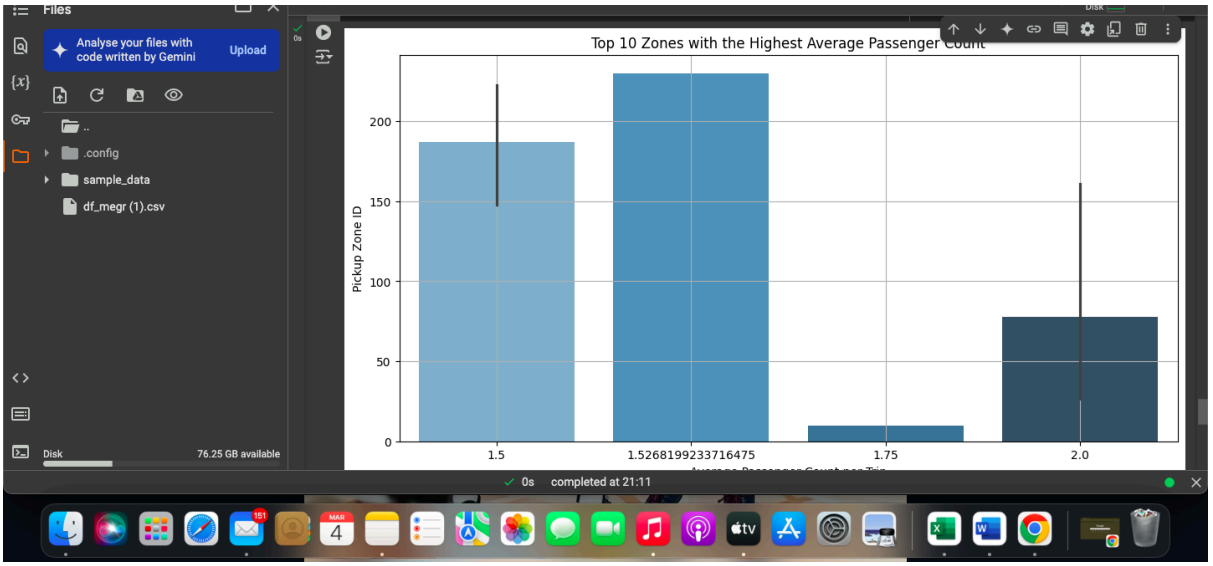
10.**Analyse the trends in passenger count**

Solo rides now make up an even larger share (79.4%), reinforcing that taxis are primarily used for individual travel.Two-passenger trips (12.6%) remain the second most common, followed by smaller percentages for larger groups (3-6 passengers).Trips with 4-6 passengers are rare (~5%), indicating lower demand for group travel.

Optimize services for solo riders by offering competitive pricing and incentives.Promote shared rides for 2+ passengers to increase fleet utilization. Analyze group ride demand by location to identify areas where larger taxis (SUVs) could be more beneficial.



11. Analyse the variation of passenger counts across zones

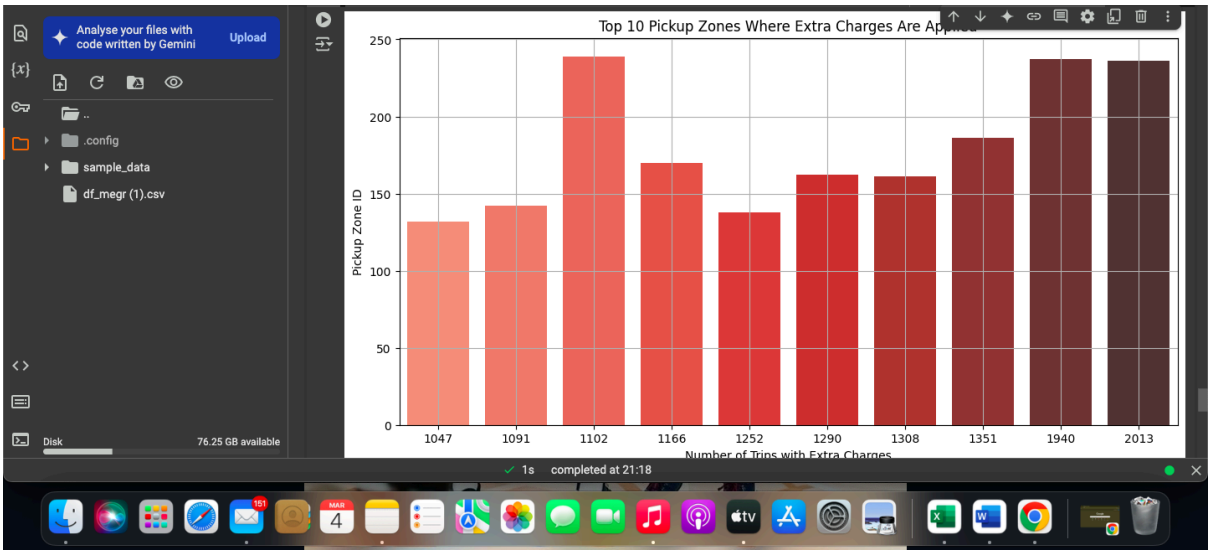


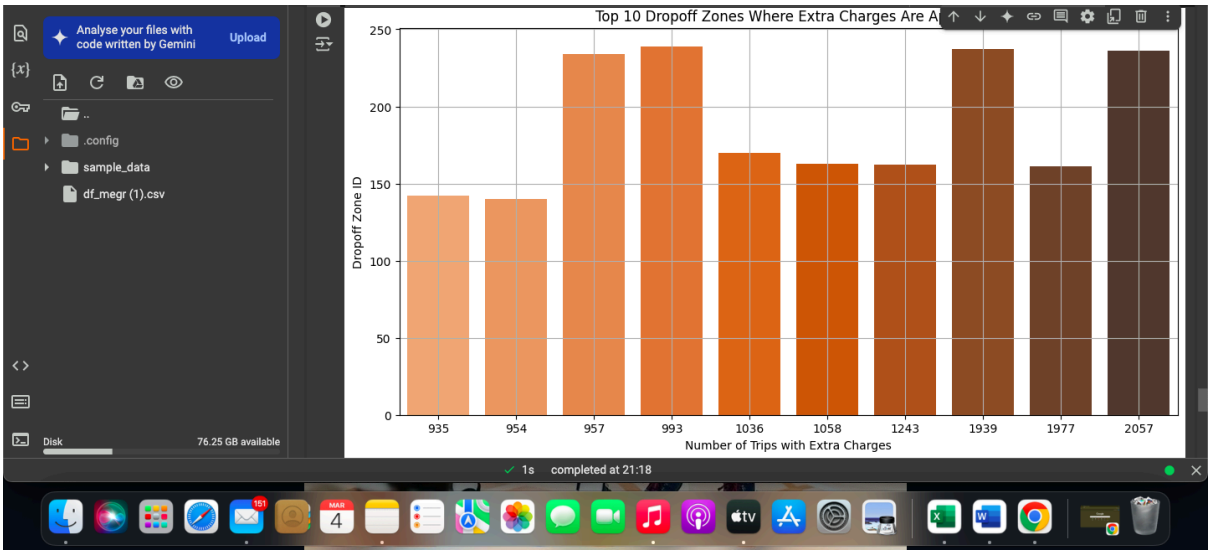
Top zones with the highest average passenger count (~2 passengers per trip):

- Zones 25, 28, 53, and 205 have the highest number of shared rides, likely indicating airport routes, transit hubs, or group travel hotspots.
- Other zones (10, 230, 38, 157, etc.) show slightly higher passenger counts than the city average (~1.5 per trip).

Deploy more larger-capacity vehicles (SUVs, minivans) in high-passenger zones to accommodate group rides.Encourage shared ride options in these zones with promotional discounts.

12.**Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.**





Analysis of Extra Charges by Zones & Time

Pickup & Dropoff Zones with Most Extra Charges:

- Zone 236 & 237 have the highest extra charges for both pickups and drop-offs, likely indicating airport or high-congestion areas.
- Other high-charge zones include 186, 161, 162, and 170, suggesting business hubs or toll-heavy routes.

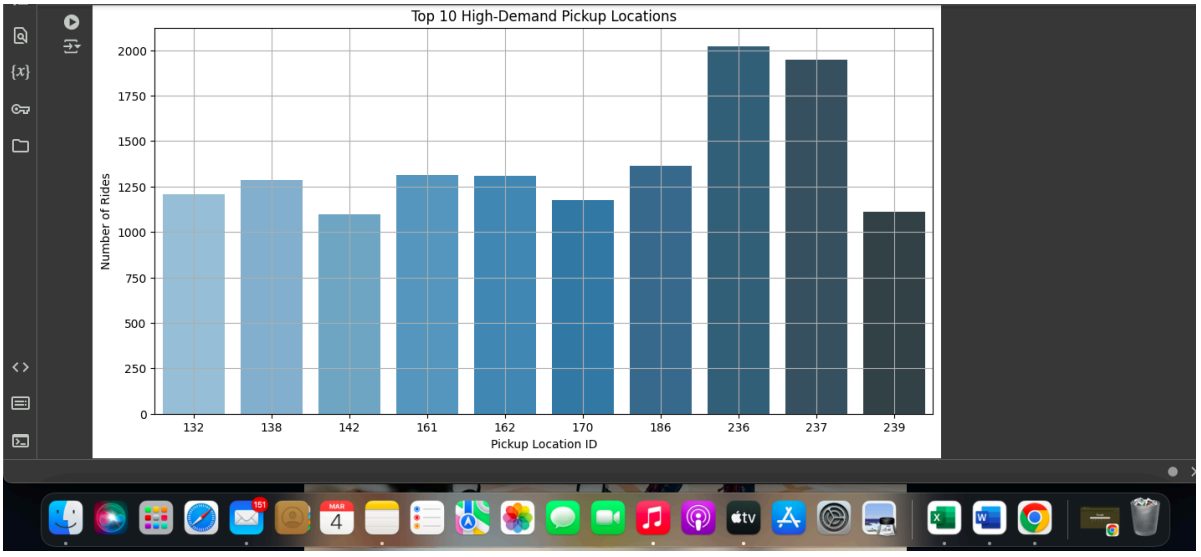
Extra Charges by Time of Day:

- 10 AM has the highest number of trips with extra charges in the dataset.
- This suggests that morning travel, likely to airports and business districts, incurs more fees (congestion, tolls, airport surcharges).

Monitor surcharge-heavy zones (airports, transit hubs) to optimize fleet allocation.Adjust pricing strategies by accounting for peak-hour extra charges.Encourage off-peak travel discounts to reduce surcharge impact for riders.

Conclusions

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



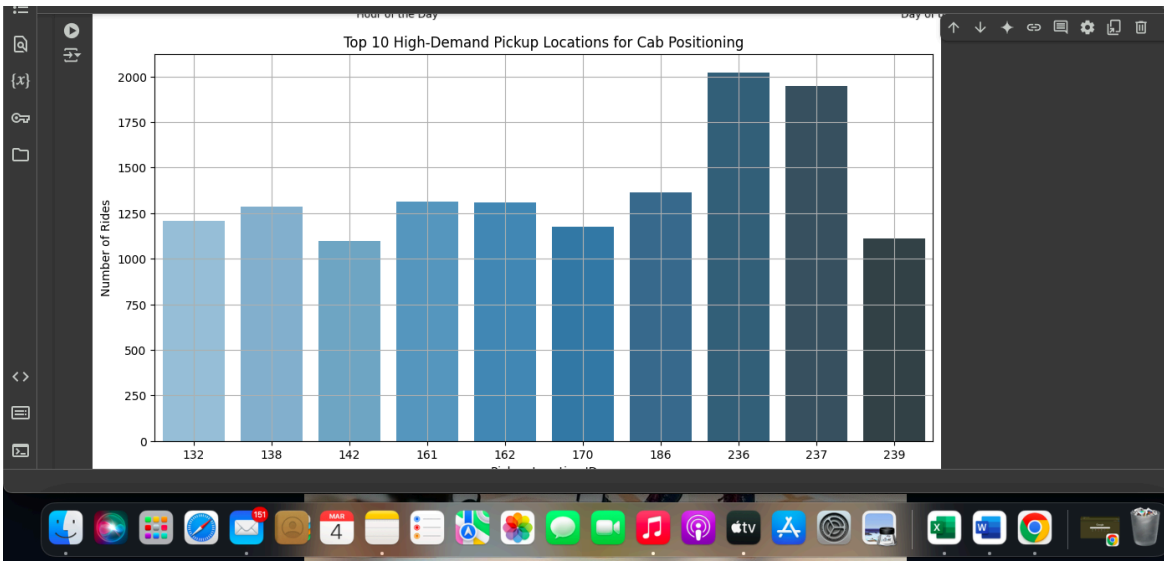
The peak ride demand occurs during morning (7 AM - 9 AM) and evening (5 PM - 8 PM) rush hours, with an additional late-night surge between 11 PM - 1 AM, particularly on Fridays and Saturdays. Conversely, demand is lowest during early weekday mornings, especially on

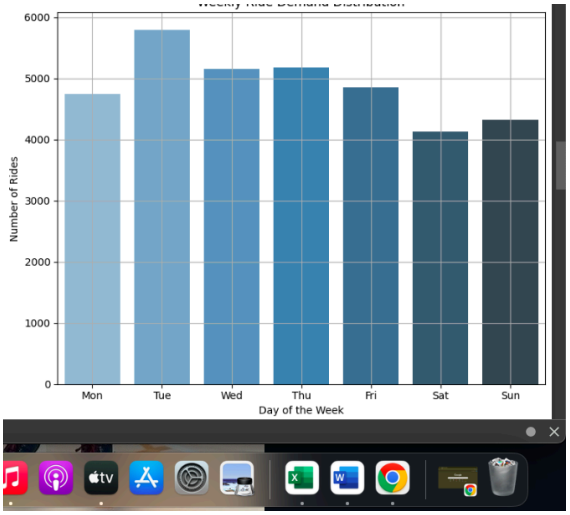
Mondays and Tuesdays.Geospatial analysis highlights specific high-demand pickup zones, which should be prioritized for fleet allocation. By implementing geofencing in these areas, taxis can be dynamically dispatched to match real-time demand, reducing waiting times and optimizing fleet utilization.Dynamic pricing should be used to manage demand—surge pricing during peak times and discounts in low-demand hours to encourage ridership. By integrating real-time event, weather, and traffic data, fleet utilization can be maximized, improving both driver earnings and passenger satisfaction.

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

To maximize efficiency and reduce passenger wait times, cabs should be strategically positioned based on demand patterns observed across time, days, and months.

1. Peak Hour Positioning (7-9 AM & 5-8 PM)
    - Business districts & transport hubs (e.g., offices, train stations, airports) should have increased cab availability during rush hours.
    - Pre-position cabs near residential areas in the early morning to capture outbound commuters.
2. Late-Night Demand (11 PM - 1 AM, Fri-Sat)
    - Deploy more taxis near nightlife zones, entertainment areas, and tourist hotspots to cater to the weekend surge.
    - Position cabs near event venues and stadiums before major concerts or games end.
3. Low-Demand Time Strategy (Early Weekdays & Midday Hours)
    - Relocate idle taxis to areas with consistent baseline demand, such as airports, hotels, and shopping districts.
    - Offer ride-sharing incentives or scheduled rides to optimize fleet usage.
4. Weekday vs. Weekend Allocation
    - Weekdays: Focus on corporate zones and transport hubs.
    - Weekends: Shift more taxis to recreational zones, shopping malls, and suburban areas for leisure travel.
5. Seasonal & Event-Based Adjustments
    - Summer & holiday seasons: Increase presence near tourist attractions and hotels.
    - Winter months: Position cabs at public transport hubs to accommodate increased reliance on taxis.





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

Dynamic Surge Pricing (Uber & Lyft Model)

Real-World Example: Uber & Lyft use real-time demand tracking to implement surge pricing when ride requests exceed available taxis.
Data Insights:

- Peak hours (7-9 AM, 5-8 PM, and late nights on weekends) show the highest demand.
- Airport zones & downtown areas incur the most extra charges.

Implementation:

- Increase fares by 1.2x - 2x during high-demand hours to maximize earnings while ensuring availability.

- In surcharge-heavy areas (airports, transit hubs), adjust base fare to absorb costs without deterring riders.

Distance-Tiered Pricing (NYC Yellow Cabs)

Real-World Example: NYC yellow cabs have a fixed initial charge and per-mile pricing, with longer trips costing less per mile to encourage travel.
Data Insights:

- Short trips (0-2 miles) are the most expensive per mile.
- Longer trips (20+ miles) have a lower per-mile fare.

Implementation:

- Reduce per-mile rates slightly for long trips (20+ miles) to encourage longer rides.
- Increase minimum fares for short trips (e.g., under 2 miles) to compensate for idle time & dead miles.

Competitor-Based Pricing (Didi & Ola Strategy)

Real-World Example: Ride-hailing giants like Didi in China and Ola in India track competitor prices in real-time and adjust fares accordingly.
Data Insights:

- Vendor 2 charges higher fares per mile than Vendor 1 (showing possible brand perception differences).
- Customers expect competitive rates, so price tracking is essential.

Implementation:

- Monitor competitor pricing daily and adjust fares dynamically to stay within 5-10% of leading competitors.
- Offer price matching during off-peak hours to attract budget-conscious riders.

Targeted Discounts & Promotions (Uber's Loyalty Programs)

Real-World Example: Uber uses discount codes, subscription-based passes (Uber One), and first-ride promotions to retain customers.
Data Insights:

- Tuesday and Thursday have the lowest demand, making them ideal for discounts.
- Solo rides dominate (79% of trips), but shared rides have lower fare per mile.

Implementation:

- Offer weekday ride discounts (10-15% off on low-demand days).
- Introduce subscription ride passes for frequent riders, offering flat-rate discounts.
- Incentivize shared rides with lower per-passenger pricing to optimize fleet utilization.

AI-Powered Demand Prediction & Geo-Pricing (Lyft & Uber Advanced Models)

Real-World Example: Lyft & Uber use AI to predict ride demand based on weather, traffic, and events and adjust fares accordingly.
Data Insights:

- Extra charges (congestion & airport fees) impact specific zones.
- Certain pickup locations (e.g., Zone 236 & 237) consistently have higher fares & tips.