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.

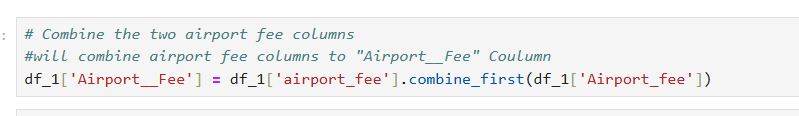
## Data Preparation

* 1. Loading the dataset
     1. **Sample the data and combine the files**  
        Loading of sample data was achieved by iterating through each parquet file and sampling out 5% of data based on date and hour of date.

## Data Cleaning

### Fixing Columns

* + 1. **Fix the index**  
       df.reset index used to fix the index issue after combining sampled data from each parquet  
       A white background with black text

       AI-generated content may be incorrect.
    2. **Combine the two airport\_fee columns**
    3. ***Fix columns with negative (monetary) values***A screenshot of a computer program

       AI-generated content may be incorrect.

### Handling Missing Values

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

After handling of -ve values itself I have noticed there no null values for data sets.

A screenshot of a computer

AI-generated content may be incorrect.

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

A screenshot of a computer

AI-generated content may be incorrect.

* + 1. **Handle missing values in RatecodeID**  
       A white rectangular object with colorful text

       AI-generated content may be incorrect.
    2. **Impute NaN in congestion\_surcharge**  
       A white rectangular object with text

       AI-generated content may be incorrect.

### Handling Outliers and Standardising Values

* + 1. **Check outliers in payment type, trip distance and trip amount columns**#just noticed my data set contains 1 crore record hence I had to reduce it again to bring to operational size. Did again sampling on the df to get data set of 2lakh 70k records.

Noticed that fare\_amount close to 300 are very regular trips hence trimmed all the trips above 300 $ payment charged. Which in turn reduced some records in my data set.

AA graph of a taxi

AI-generated content may be incorrect.

After removing

A graph of a taxi fare

AI-generated content may be incorrect.

There were 3338 rows with trip distance 0 , removed this row as reduction in this count will not impact the analysis ( small %)

notice only few trips above 60 we can remove that as well proper analysis

A screenshot of a computer

AI-generated content may be incorrect.

Post that I had remaining 266863 records.

A screenshot of a graph

AI-generated content may be incorrect.

As noticed from the above graph there are few fare\_amount columns and total amount column I had to cleanup as they were having 0 values.

## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical** A screenshot of a computer

       AI-generated content may be incorrect.
    2. **Analyse the distribution of taxi pickups by hours, days of the week, and months**

**Inferences :**

**The peak hour starts from 2PM to 9 PM considering the number of pickups recorded close to min 15000+ and maximum(17500+)**

**If you take the days all days are having above 30000 pickups , to highlight we can consider Wednesday and Thursday which recorded 40k+ pickups**

**Apart from AUG and SEP moths all months we have good pickup counts, topers being May and OCT**

A graph with a line going up

AI-generated content may be incorrect.

A graph showing the number of blue bars

AI-generated content may be incorrect.

A graph of blue lines

AI-generated content may be incorrect.

* + 1. **Filter out the zero/negative values in fares, distance and tips**A screenshot of a computer program

       AI-generated content may be incorrect.
    2. **Analyse the monthly revenue trends**As we can see the pickup counts also high for MAY and OCT , they are the top notch revenue generating months as well

A graph with green lines and dots

AI-generated content may be incorrect.

* + 1. **Find the proportion of each quarter’s revenue in the yearly revenue**A pie chart with numbers and a number on it

       AI-generated content may be incorrect.
    2. **Analyse and visualise the relationship between distance and fare amount**A graph with blue dots

       AI-generated content may be incorrect.
    3. **Analyse the relationship between fare/tips and trips/passengers**

A graph with blue dots

AI-generated content may be incorrect.

* + 1. **Analyse the distribution of different payment types  
       most of the payments are through credit card and next liked payment method is cash.**A graph of a distribution of payment types

       AI-generated content may be incorrect.
    2. **Load the taxi zones shapefile and display it**A map of the united states

       AI-generated content may be incorrect.
    3. **Merge the zone data with trips data**A screenshot of a computer program

       AI-generated content may be incorrect.
    4. **Find the number of trips for each zone/location ID**A screenshot of a computer screen

       AI-generated content may be incorrect.
    5. **Add the number of trips for each zone to the zones dataframe**We have merged the zones with location\_id and PulocationID
    6. **Plot a map of the zones showing number of trips**A map of a pickup truck

       AI-generated content may be incorrect.
    7. **Conclude with results**As per the above chart most of the pickups are from the location ID 132 which is JFK Airport , as expected it is the entry point to America and most bookings will be from this location expected Similarly midtown center, upper east side north/south are the next pickup locations

### Detailed EDA: Insights and Strategies

* + 1. **Identify slow routes by comparing average speeds on different routes**A screenshot of a computer

       AI-generated content may be incorrect.High traffic areas have high trip duration and by identifying them (per hour per day) we can choose a less traffic route to the destination.

Above trip duration gives a picture of most delayed routes per hour.

By finding out high-demand pickup locations per hour per day, we can allot a greater number of cabs to these locations.

* + 1. **Calculate the hourly number of trips and identify the busy hours**A graph of blue bars

       AI-generated content may be incorrect.
    2. **Scale up the number of trips from above to find the actual number of trips**A graph of blue bars

       AI-generated content may be incorrect.

\* The peak hour (here 3pm to 7pm above 1 lakh requests) indicates high demand for rides , most rush hour being 6PM.

\* Off-peak hours (here midnight- 01 Am- 6AM nearly 20k trips only) have fewer trips.

\* As mentioned above this helps us with better taxi allocation and pricing strategies

* + 1. **Compare hourly traffic on weekdays and weekends**A graph with a line

       AI-generated content may be incorrect.

Weekday peak hours: Likely Evening (4PM - 7.30PM) rushing to home after office hours

Weekend peak hours: Might shift later in the day (afternoon/evening due to family outing or gathering).

Off-peak hours: Late-night & early morning (12 AM - 6 AM) show low trip counts.

* + 1. **Identify the top 10 zones with high hourly pickups and drops**A screenshot of a graph

       AI-generated content may be incorrect.
    2. **Find the ratio of pickups and dropoffs in each zone**A screenshot of a computer

       AI-generated content may be incorrect.
    3. **Identify the top zones with high traffic during night hours**A screenshot of a computer program

       AI-generated content may be incorrect.

**Its visible that top drop off zones and top pick up zones near, which may indicate these are locations of night clubs, or party locations. Aligning more taxi to these location during night time will help to increase the revenue.**

* + 1. **Find the revenue share for nighttime and daytime hours**A screenshot of a computer code

       AI-generated content may be incorrect.it indicates a strong daytime demand, all business men rush hour commuting, airport trips, or business areas generate more revenue
    2. **For the different passenger counts, find the average fare per mile per passenger**A graph of a passenger count

       AI-generated content may be incorrect.

This clearly shows Higher passenger counts Lower fare per passenger (cost-sharing).

If fare per mile is high for low passenger counts, it might indicate shorter trips with minimum fare applied.

for single passenger price is high indicates no sharing

* + 1. **Find the average fare per mile by hours of the day and by days of the week**A graph and chart of a graph

       AI-generated content may be incorrect.

Higher fares per mile is comstant through out the week days with slight up and and down , means a regular comuter strength

Peak fare per mile in early morning (midnight - 4 AM) → Late-night trips may be more expensive.

Price is lower for peek hours indicate that business strategy for more trips

* + 1. **Analyse the average fare per mile for the different vendors**A graph with a line graph

       AI-generated content may be incorrect.Vendor two is charging more for late night trips were as vendor is keeping stable pricing throughout day and night Vendor 1s pricing strategy need to be reconsidered, vendor two is having good performance during peak hours as well.
    2. **Compare the fare rates of different vendors in a distance-tiered fashion**A graph of a number of people

       AI-generated content may be incorrect.

**For shorter distance vendor 2 is charging**

* + 1. **Analyse the tip percentages**A graph of green bars

       AI-generated content may be incorrect.

**The more distance the more tips will be given**

**If the number of passengers is 1 or more than = 5 then the tip amount is increasing**

**Also, during rush hours of 4 - 6 pm tip amount is high**

* + 1. **Analyse the trends in passenger count**A graph showing the average passenger

       AI-generated content may be incorrect.

There is high number of passenger count in early morning -- may be returning after party or bar etc

then from 7am till 12 am there is continues increase

* + 1. **Analyse the variation of passenger counts across zones**A map of a passenger

       AI-generated content may be incorrect.

A map of the united states

AI-generated content may be incorrect.

* + 1. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.**–A graph with a red line

       AI-generated content may be incorrect.  
       Higher extra charges at night (08 PM - 11 pM) & peak hours (4-8 PM).

Comparatively low extra charges early morning 3am to 7am

## Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**
* Strategic Cab Placement Near High-Pickup Zones
  + High-demand pickup spots: JFK Airport, Upper East Side, Midtown Center, Midtown East.
  + Cabs should be pre-positioned near these locations to reduce waiting times.
  + Use predictive dispatching to position more cabs before demanding spikes.
  + Increase cab availability at airports & business districts in peak hours.
* Dropoff-Based Cab Reallocation
  + Most drop-offs happen in East Village, Clinton East, Murray Hill East, Chelsea.
  + These areas have high rider availability, reducing cab downtime
  + Reduce cab dispatching to these areas (self-balancing effect).
  + Instead, allocate more cabs to pick up-dominant areas.
  + Encourage drivers to reposition themselves to pick up hotspots after drop-offs.
    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.**
* High Passenger Zones for Maximum Coverage
  + Position cabs near high-demand zones (IDs 263, 261, 262) before peak hours.
  + Deploy more cabs in bar/nightlife zones (early morning rush).
  + Increase cabs in business districts from 7 AM onward for commuter demand.
* Late-Night High Fare Per Mile Areas
  + Deploy more cabs in nightlife areas at night (bars, clubs, entertainment zones).
  + Reduce cab allocation in business districts after midnight (low demand).
  + Encourage late-night drivers by increasing surge pricing & incentives.
* Cab Distribution by Time of Day
  + Increase cab presence in residential areas before 7 AM (morning rush).
  + Maintain balanced distribution in business hubs during work hours.
  + Keep cabs available near transit hubs (subways, train stations) during rush hours.
    1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**
* Adjust Pricing Based on Passenger Count
  + Higher passenger counts → Lower fare per passenger (cost-sharing effect)
  + Solo passengers pay a higher fare per mile (no shared cost)
  + Trips with 1 or 5+ passengers receive higher tips
    - Proposal
      * Encourage shared rides: Offer discounts for 2+ passengers
      * Increase base fare slightly for solo trips to compensate for non-shared rides
      * Introduce a per-passenger fare cap beyond a certain number of riders
* . Fare Adjustments During Rush Hours
  + Peak demand: 3 PM - 7 PM (1 lakh+ requests) = opportunity for premium pricing
  + High tips during 4-6 PM → Business professionals, corporate rides, airports.
  + Off-peak (Midnight - 6 AM) sees low demand (~20K trips).
    - Proposal
      * Increase fares by 10-15% during peak hours (3 PM - 7 PM) but limit excessive surge pricing to stay competitive.
      * Offer discounted off-peak fares (Midnight - 6 AM) to attract more riders
      * Introduce flat-rate airport pricing to simplify fare calculation.
* Implement Vendor 2’s Dynamic Pricing Model for Vendor 1
  + Vendor 2 is making higher profits by adjusting pricing based on travel hours.
  + Vendor 1 could implement similar flexible pricing.
  + Use time-based fare adjustments (e.g., peak vs. non-peak rates).
  + Charge premium fares during high-demand periods (4-6 PM, late-night rides).
  + Lower fares during off-peak hours to attract more customers.
* Optimize Route Selection to Reduce Trip Duration & Fuel Costs
  + High-traffic zones cause long trip durations, increasing operational costs
  + Choosing less congested routes reduces travel time & fuel consumption.
    - Proposal
      * Charge extra for high-traffic zone pickups to compensate for delays.
      * Prioritize drivers on optimized routes with minimal congestion.