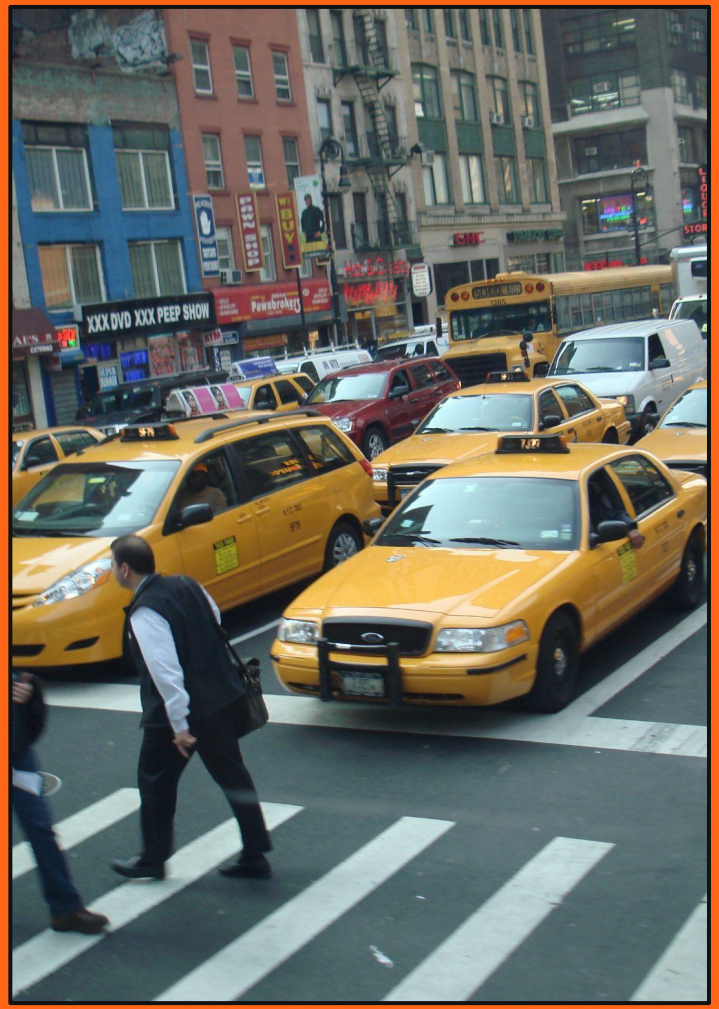
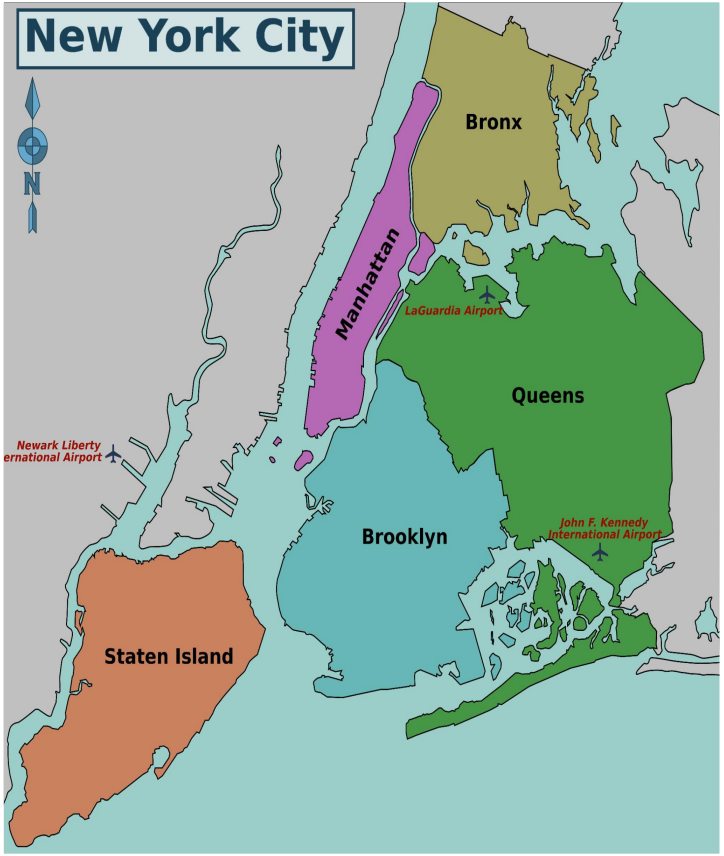
Statistics 405 Project



Connor Herbert, Zoë Weinstein, Jake Rottier,
Sakshi Shah, Zhongzhen Zhou
04.28.2025

Group 3

Project Report



Project Overview

- Project Explores Trends in Historical Taxi and For-Hire Rides in NYC since 2009
- Includes analysis on ride volume, rates, and differences between ride types

The Data

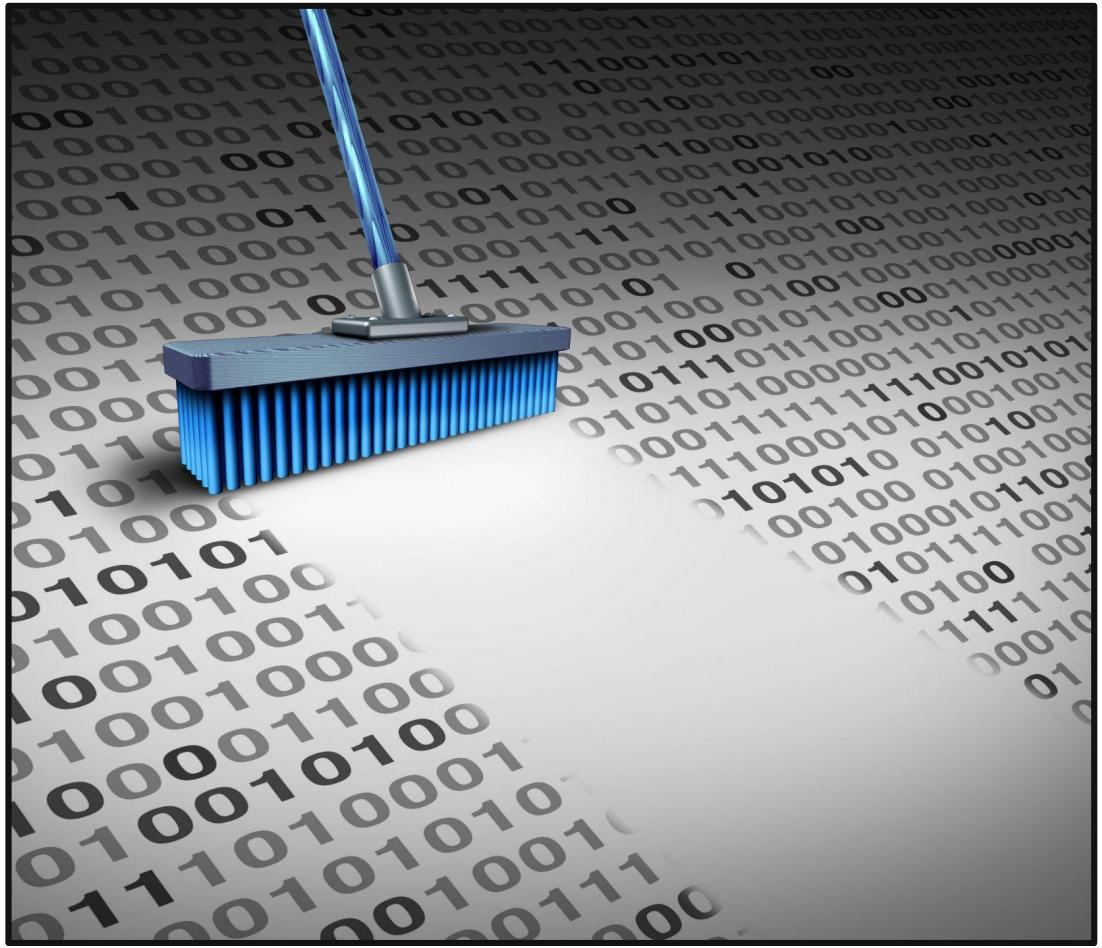
- Data comes from The New York City Taxi and Limousine Commission (TLC)
  - Has been received electronically since 2009
  - Added For-Hire Trip Data in 2015
    - Uber
    - Lyft
    - Via
    - Juno

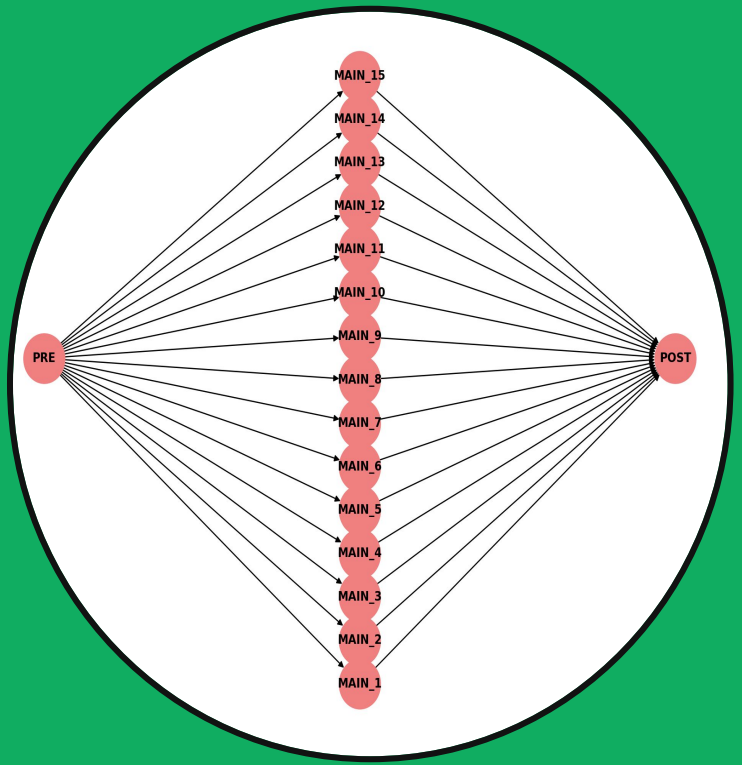
The Data

- Data contains directories for each year from 2009-2023
  - 473 Data Files
  - 68 GB of Data
  - Stored as .parquet files
    - Much smaller than CSV but not human readable

The Data

- Downloaded and Extracted Zip File
- Processed each year in parallel
- Converted .parquet files to .csv and combined with Python
- Removed unnecessary columns and rows with N/A





Statistical Computation

- 15 Parallel Jobs
  - Each Took About 15 Minutes
  - Required 32GB of RAM to process the 20+GB CSV Files
  - Had to Work in Staging so we had enough Disk Space (100GB+ Combined)

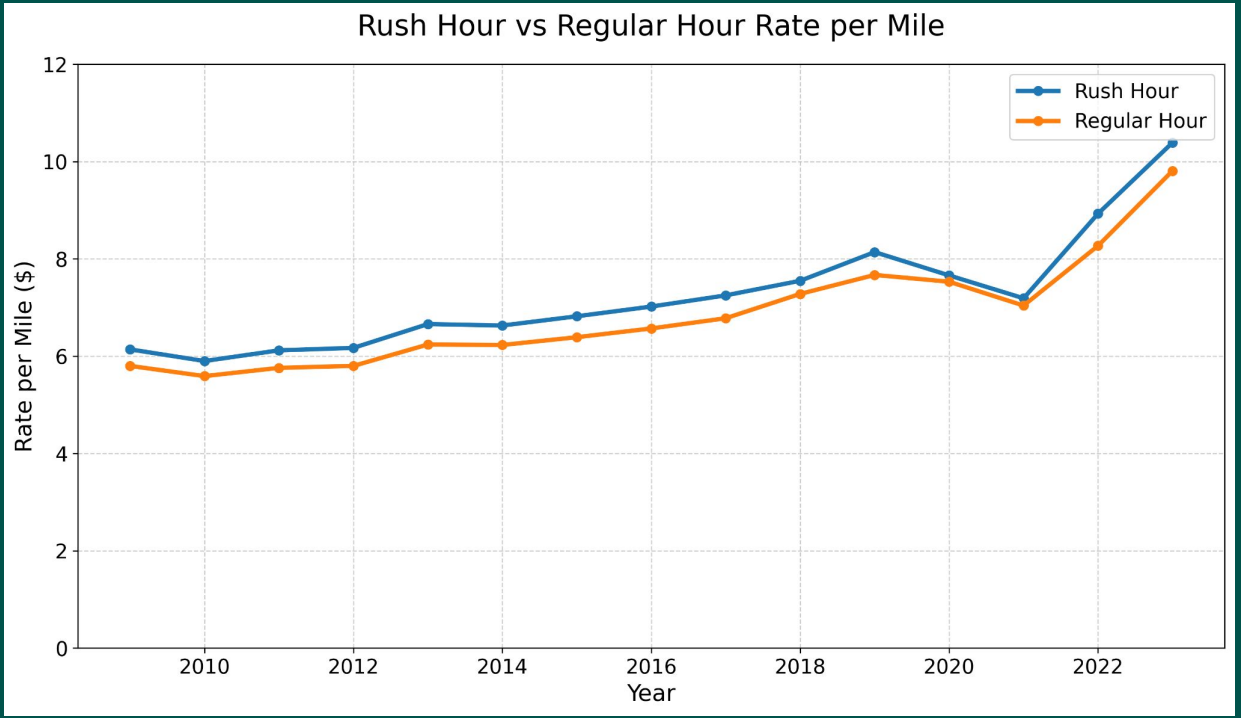
Statistical Computation

- Container Images
  - Custom Built Containers to Include Necessary Python Libraries
  - Required API keys from Kaggle to download the data

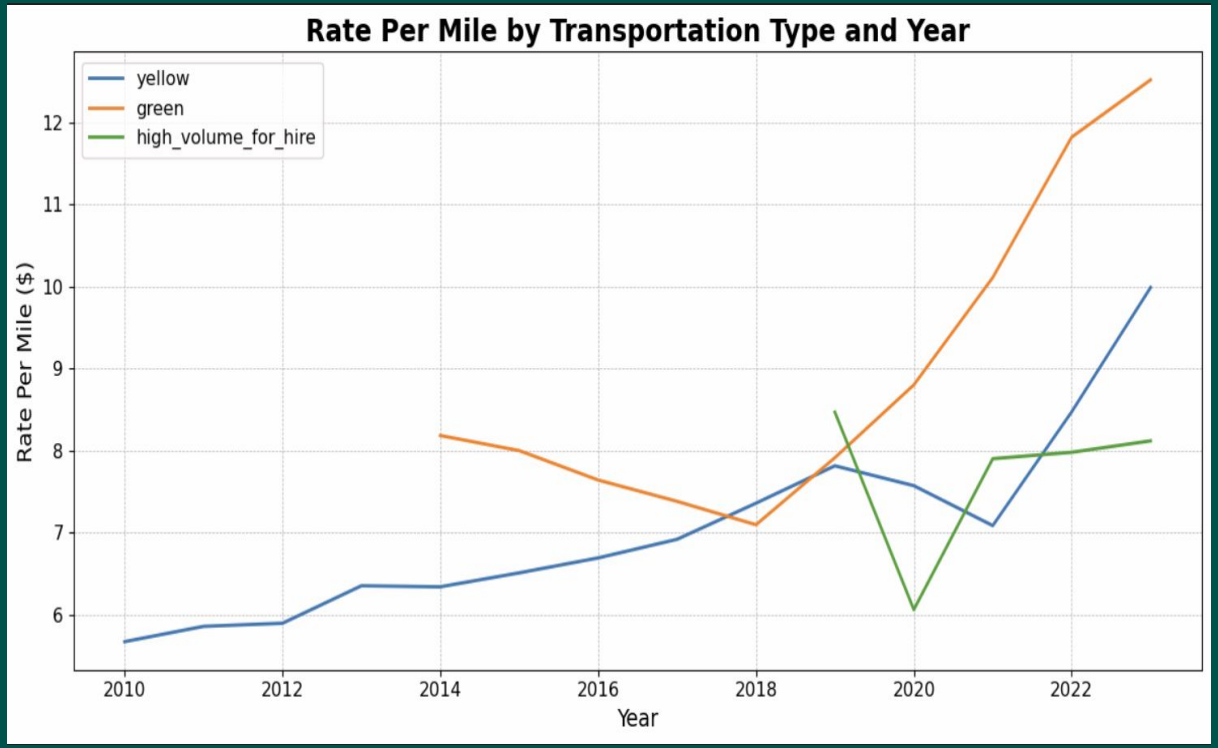
Analysis 1 - Peak vs Regular Rates

How do fares during peak commute times (7am-9am) and (4pm-6pm) differ from non commute times?

- During peak commute times, average rate/mile is higher (Avg. 70¢ Higher)
- Taxis likely up prices during rush hour
- Interestingly, average overall fare prices are higher during non commute times



Analysis 2 - Comparing Rate/Mile



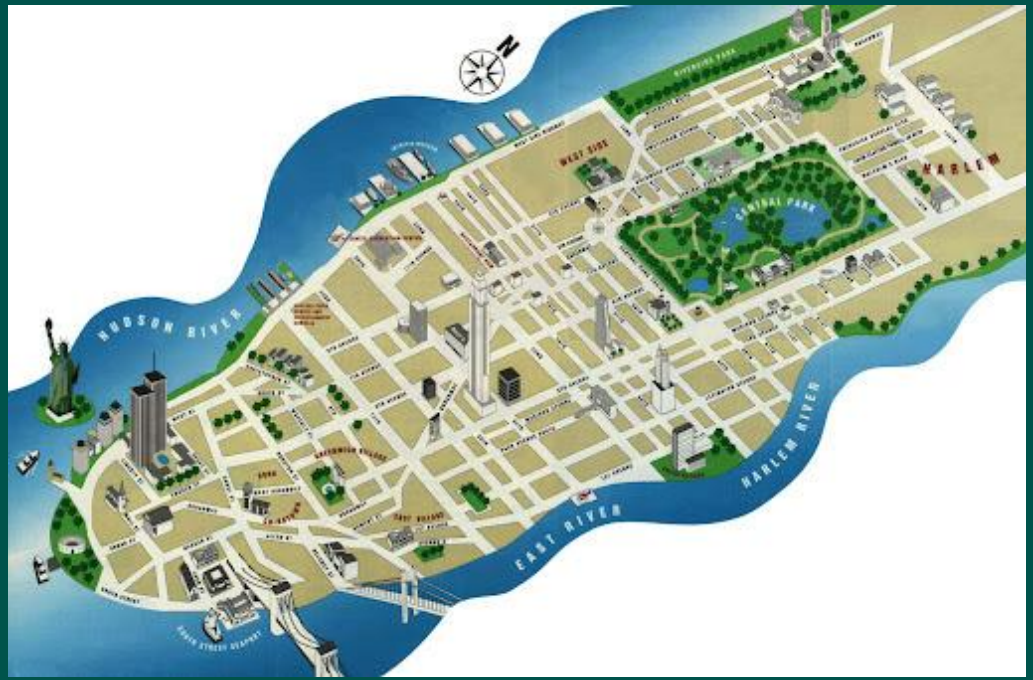
Is there a difference in the rate/mile based on transportation type?

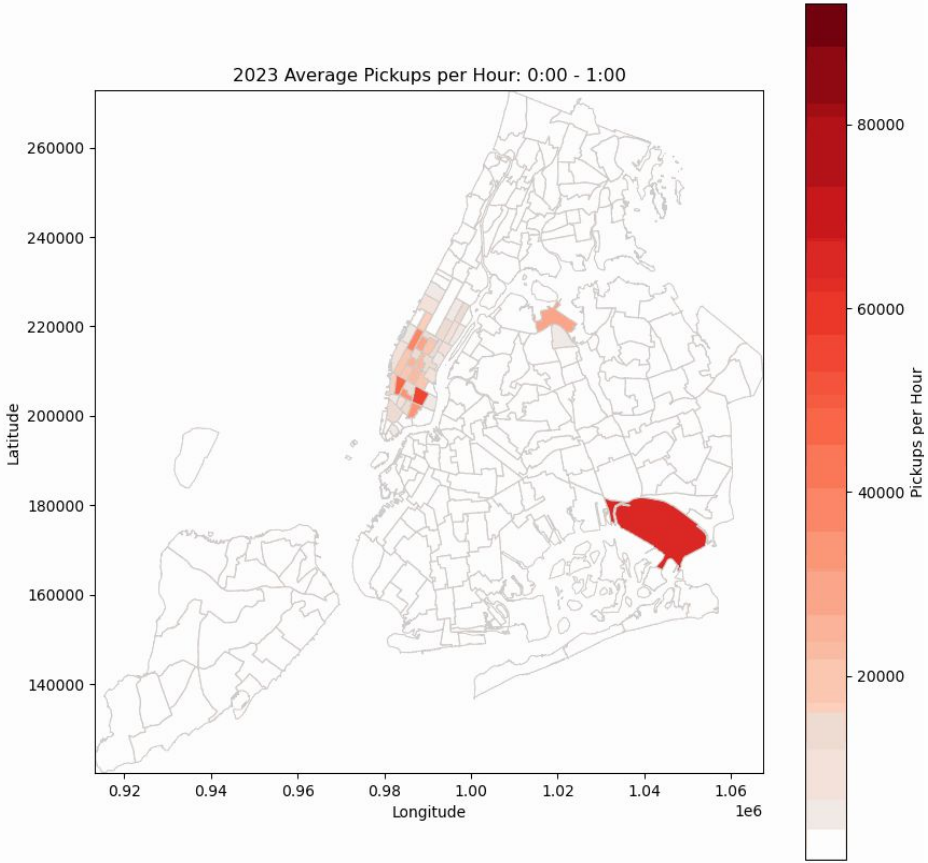
- Used Parallel Computing to process all the CSV files
- Some years did not have data for all types
- Rate / Mile has been increasing over time
- Yellow and Green Taxis have higher rates than Uber

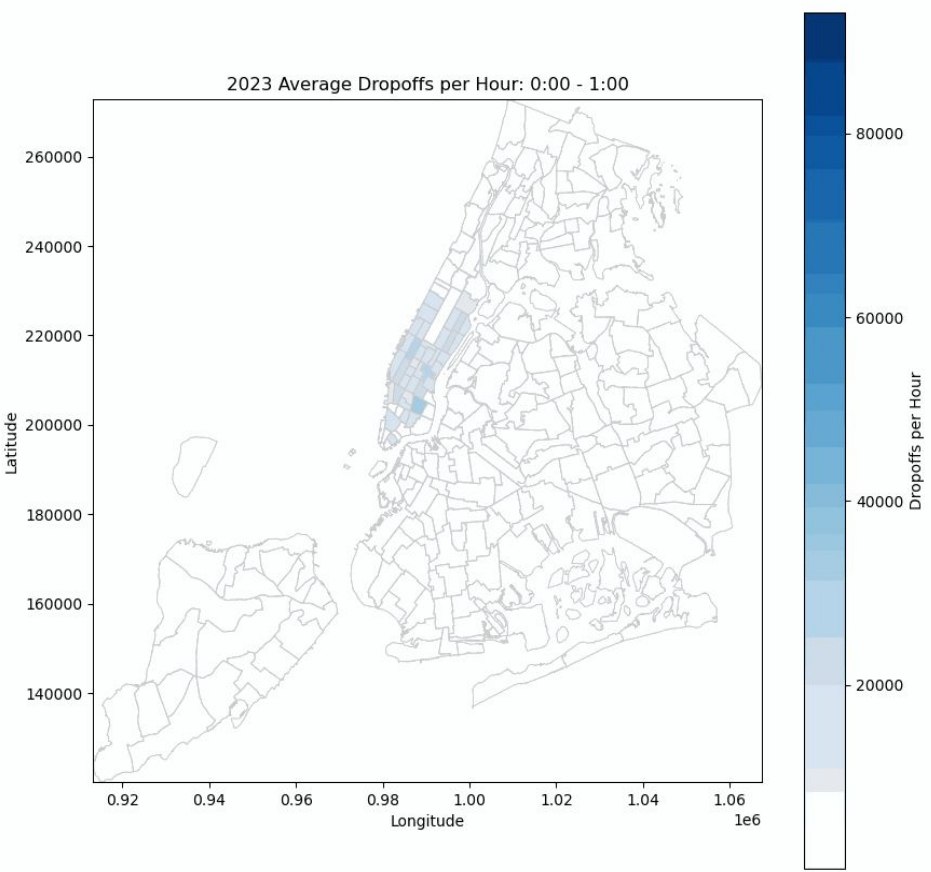
Analysis 3 - Mapping Pick ups/Drop Offs

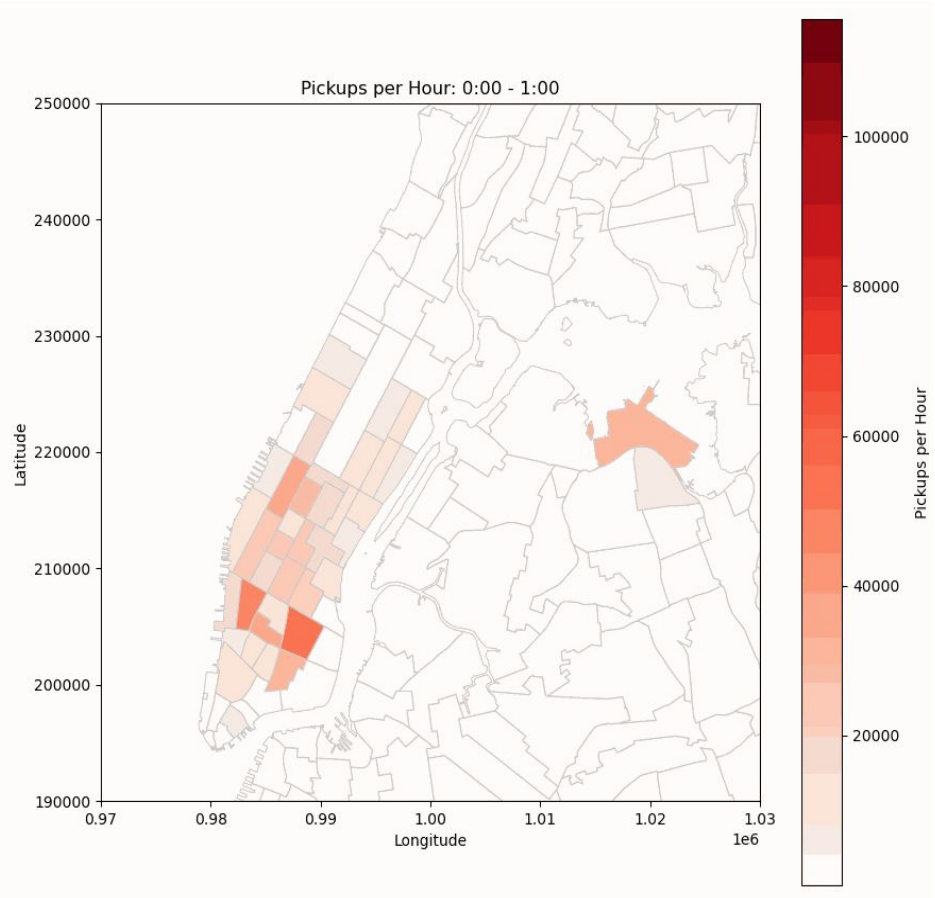
Where are the most popular locations (origin and destination) for taxi travel?

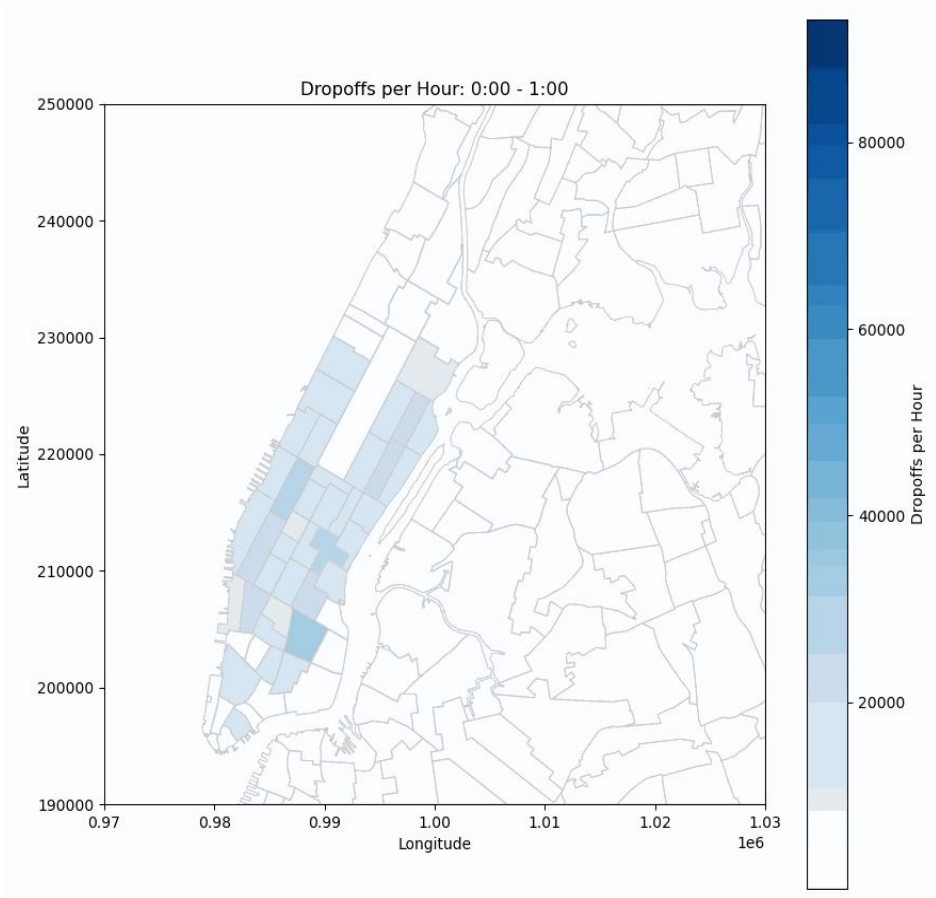
- Created heat maps by hour to show flows of people
- One graph for pick up locations, one for drop off locations




2023 Average Pickups per Hour: 0:00 - 1:00
Latitude
Longitude
1e6
Pickups per Hour


2023 Average Dropoffs per Hour: 0:00 - 1:00
Latitude
Longitude
1e6
Dropoffs per Hour


Pickups per Hour: 0:00 - 1:00
Latitude
Longitude
1e6
Pickups per Hour
250000
240000
230000
220000
210000
200000
190000
0.97
0.98
0.99
1.00
1.01
1.02
1.03
100000
80000
60000
40000
20000


Dropoffs per Hour: 0:00 - 1:00
Latitude
Longitude
1e6
Dropoffs per Hour
250000
240000
230000
220000
210000
200000
190000
0.97
0.98
0.99
1.00
1.01
1.02
1.03
80000
60000
40000
20000

Analysis 4 - Overall trend in fare rates

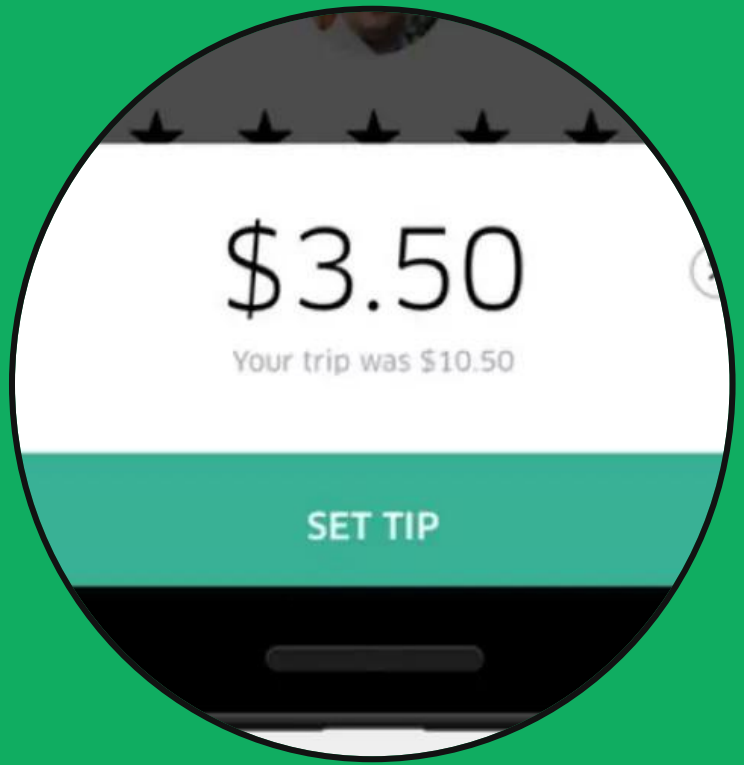
What Patterns have been observed in our dataset?

- 2009-2014 -> Stable with minor increases
- 2015 was the introduction of uber / lyft ride services *in our dataset
- 2015-2019 also stayed steady, with changes due to similar outside sources
- Dip in 2020 due to COVID-19 outbreak
- Post COVID dip, then spike

Analysis 5 - Compare fare prices with the day of the week

```python
import pandas as pd
import os
#current csv data folder
input_folder = "./"
years = list(range(2009, 2025))
all_data = []

for year in years:
    file_path = os.path.join(input_folder, f"yellow_taxi_combined_data_{year}.csv")
    try:
        print(f"Reading {file_path}...")
        df = pd.read_csv(file_path, usecols=["pickup_datetime", "fare_amount"])
        df['pickup_datetime'] = pd.to_datetime(df['pickup_datetime'])
        #drop bad rows
        df = df.dropna(subset=['pickup_datetime', 'fare_amount'])
        #keep valid rows
        df = df[df['fare_amount'] > 0]
        all_data.append(df)
    except Exception as e:
        print(f"Errors in {year}: {e}")

#merge them to a gaint dataframe
merged_df = pd.concat(all_data, ignore_index=True)

merged_df['day_of_week'] = merged_df['pickup_datetime'].dt.day_name()

# Group by day of the week and get the average fare of each day of the week
avg_fare_by_day = merged_df.groupby('day_of_week')['fare_amount'].mean().reset_index()

# Sort day order
day_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
avg_fare_by_day['day_of_week'] = pd.Categorical(avg_fare_by_day['day_of_week'], categories=day_order, ordered=True)
avg_fare_by_day = avg_fare_by_day.sort_values('day_of_week')

print(avg_fare_by_day)

avg_fare_by_day.to_csv("average_fare_by_day_all_years.csv", index=False)
```

- Read cvs files from 2009-2025 load pickup time and fare
- Clean data and merge all yearly data into one Dataframe
- Extract day of week and compute average fare per day then sort days from Monday to Sunday
- Save the final average fare table to CSV file



Potential Weaknesses

- Handling Outliers
  - Trips with high distances or rates
- Data Cleaning
  - Instead of removing data with missing values, try to understand why it is missing and fix it using data imputation
- Improved Fare Definitions for For-Hire Rates
  - Including Tips, Surcharges, and Fees