COMS4060A/7056A: Assignment 2

Puseletso Nakana : 2341162
Ntando Ngobese : 2112256
Nthabiseng Thema : 2012016

Department of Computer Science
University of Witwatersrand

1 Question 1:

The analysis focused on three key metrics: trip duration, distance traveled, and average speed. For each metric, the interquartile range (IQR) method was used to identify outliers. This method considers values below Q1 - 1.5IQR or above Q3 + 1.5IQR as outliers, where Q1 and Q3 are the first and third quartiles respectively. The code calculated these thresholds for each metric and identified data points falling outside these ranges. Finally, all unique outliers from the three metrics were combined and removed from the dataset, resulting in a cleaned version of the data.

Number of outliers in trip duration: 17512
Number of outliers in distance: 33017
Number of outliers in speed: 14002
Number of unique outliers: 51322

Trip duration outliers (17,512): Extremely short or long trip durations could indicate:
Justification: Including these outliers could skew average trip times and impact analysis of typical travel patterns.

Distance outliers (33,017): Unusually short or long distances might represent: GPS errors
Incomplete trips

Justification: These outliers could distort average trip lengths and affect route analysis or fuel consumption estimates.

Justification for Removing Outliers Data Quality: Outliers often result from data entry errors or malfunctions in the taxi meters. Removing them helps in ensuring the quality and accuracy of the data.

Analysis Accuracy: Outliers can skew the results of statistical analyses and models. By removing them, you can obtain more accurate and reliable insights.

Realistic Insights: Outliers can distort the interpretation of patterns and trends. Removing them ensures that the insights you gain are more representative of typical trips.

2 Question 2

In this feature generation step, three new columns were added to the DataFrame:

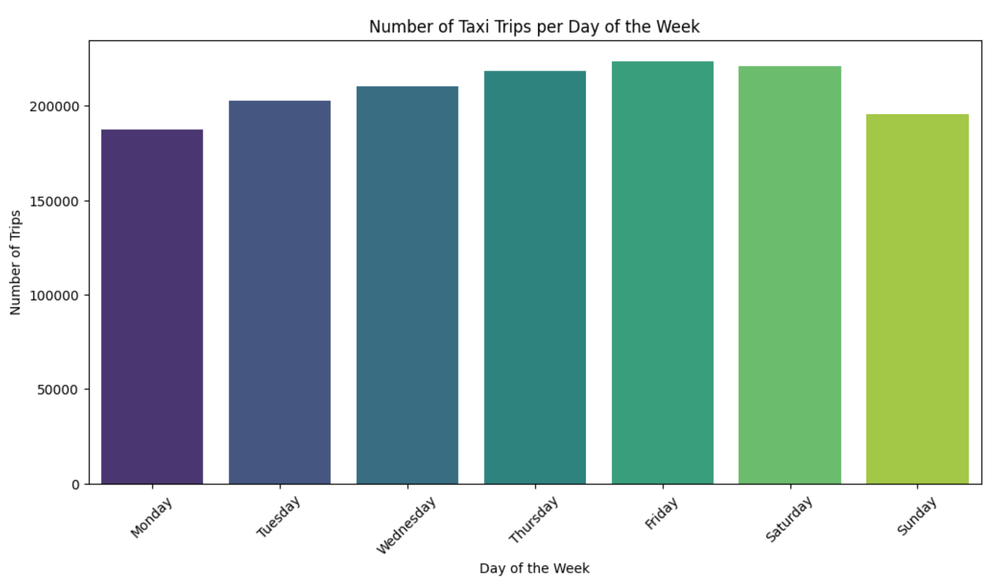
Day of week: Extracted from the 'pickup_datetime' column using the dt.day_name() method. This provides information about which day of the week each trip occurred.
Average speed: Calculated by dividing the 'distance_km' by the trip duration (converted from seconds to hours). This gives the average speed of each trip in km/h.

	distance_km	day_of_week	speed_kmh
0	1.497580	Monday	11.848984
1	1.804374	Sunday	9.797504
2	6.381090	Tuesday	10.815406
3	1.484566	Wednesday	12.457894
4	1.187842	Saturday	9.830418

3 Question 3

3.1 3.1

Day of Week Extraction:
The code extracts the day of the week from the 'pickup_datetime' column using the dt.day_name() method, creating a new 'day_of_week' column.
Trip Counting:
It then counts the number of trips for each day of the week using value_counts().
Reordering:
The days are reordered to follow the standard weekday sequence (Monday to Sunday) for better interpretation.



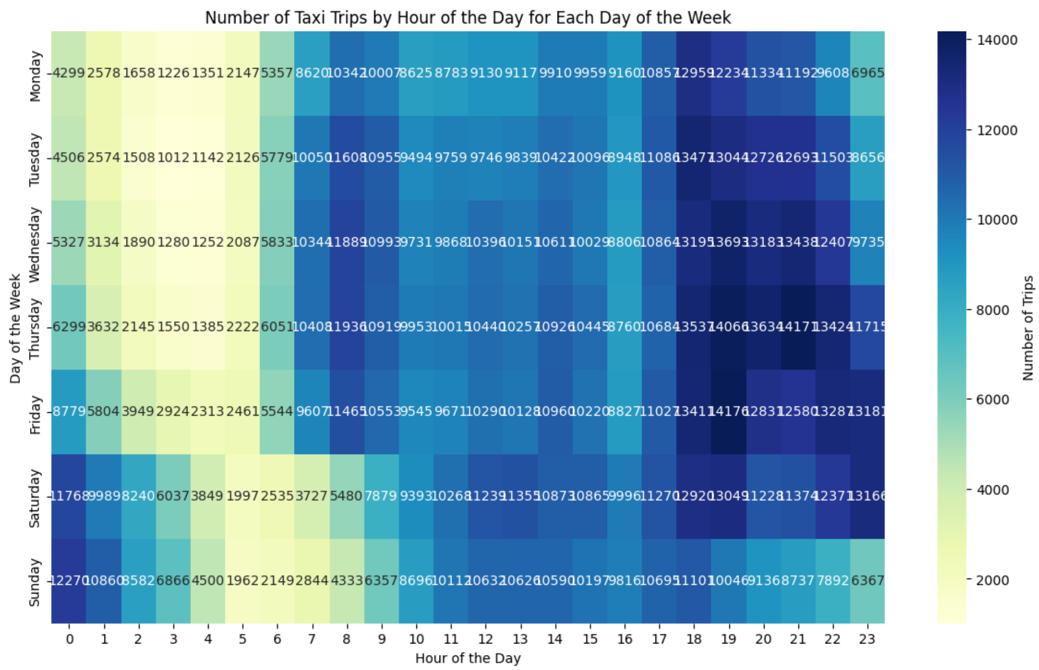
3.2
Data Processing:

The NYC taxi dataset was loaded, and two new columns were created: 'hour_of_day' and 'day_of_week'. The data was then aggregated to count the number of pickups for each hour of each day of the week.

Data Analysis:

The code identified the most popular hour (with the highest number of pickups) for each day of the week. A DataFrame 'most_popular_hours_df' was created to display these results, which is what we see in the image.

	Most Popular Hour	Count
day_of_week		
Monday	18	12959
Tuesday	18	13477
Wednesday	19	13693
Thursday	21	14171
Friday	19	14176
Saturday	23	13166
Sunday	0	12270



3.3

Observations and Explanations You will see higher numbers of pickups during early morning 7-9 AM and evening hours 5-10 PM on weekdays, corresponding to typical work commute times.

You will see higher numbers of pickups during 6 PM Saturday evening to 1 AM on Sunday morning, corresponding to late night outings on weekends.

Weekends vs. Weekdays: Weekends show different patterns, such as more late-night pickups compared to weekdays. This could be due to social activities or nightlife.

3.4

Data Preparation:

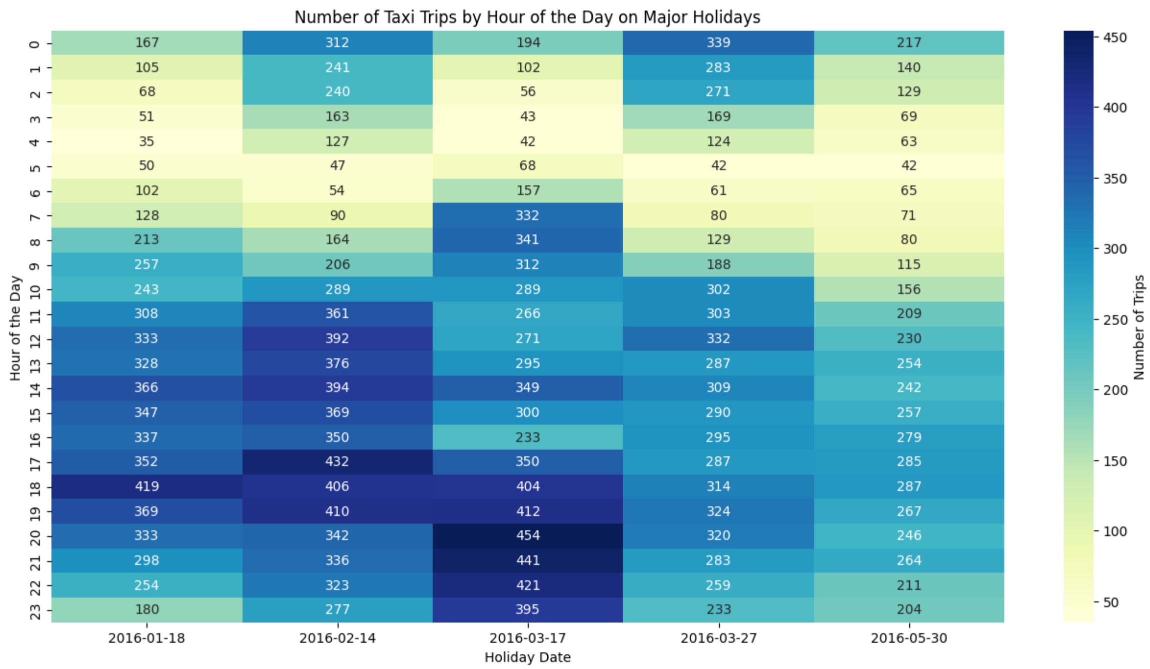
Extracts hour of the day, day of the week, and date from the pickup_datetime. Defines specific holiday dates for 2016.

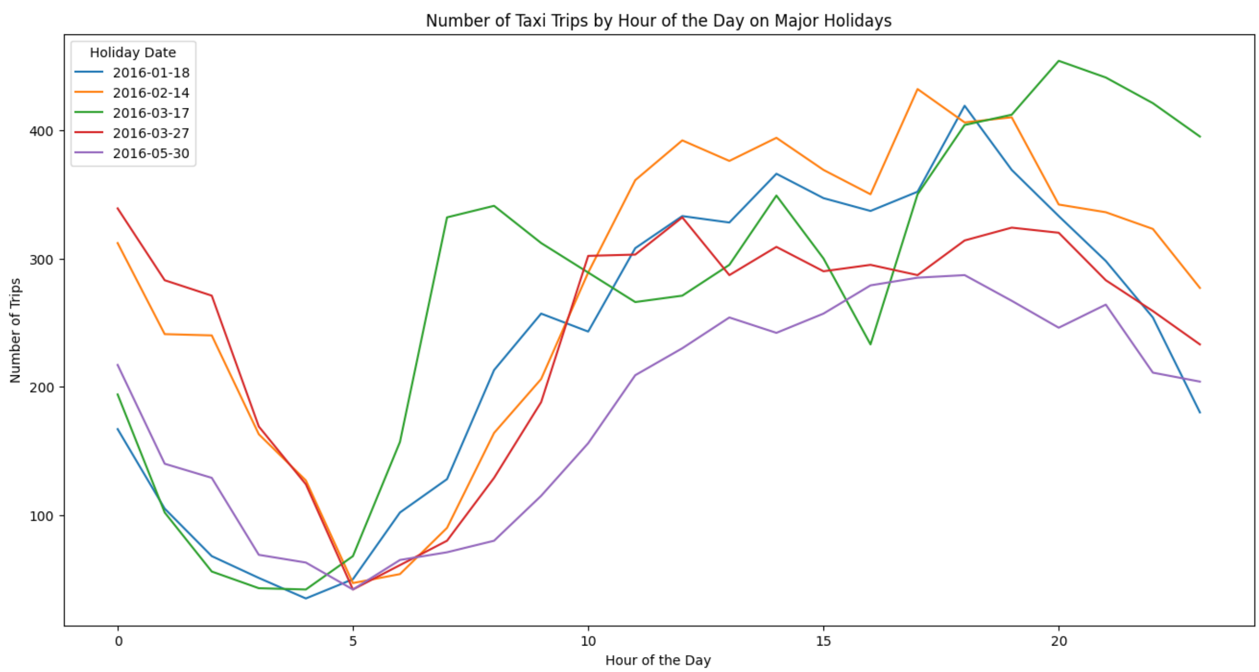
Holiday Data Check:
Prints the holiday dates. Checks if these holidays are present in the dataset. Filters the dataset for trips on holiday dates.

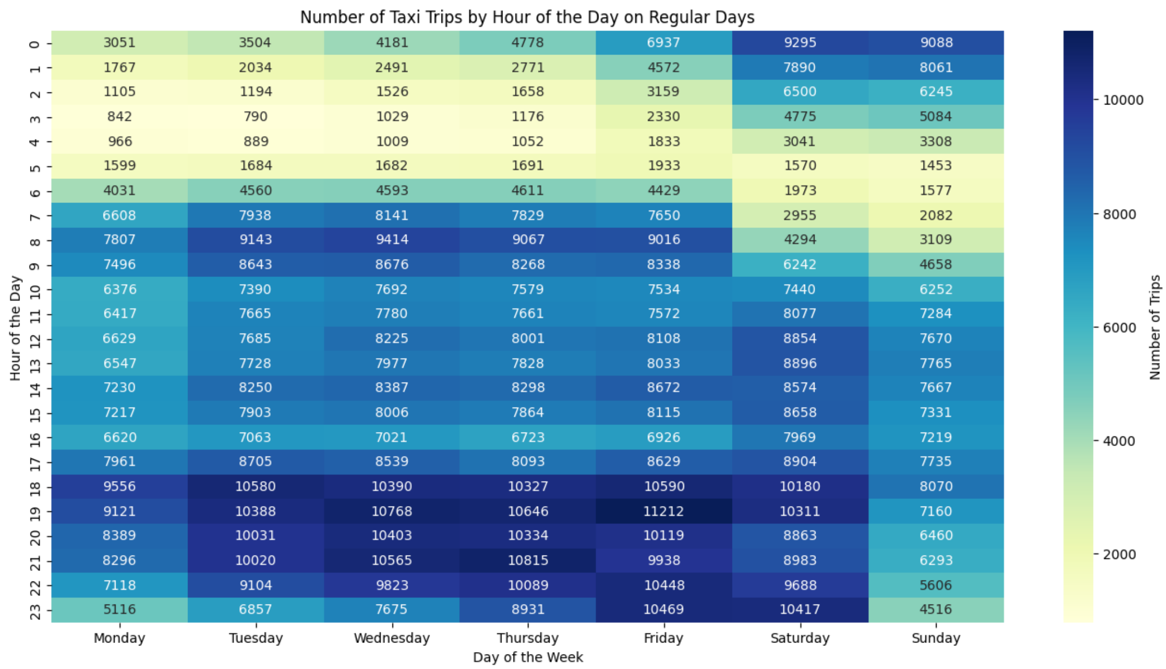
Holiday Analysis:
If holiday data is available:
Aggregates the number of pickups per hour for each holiday. Creates a heatmap to visualize the distribution of trips by hour on holidays.

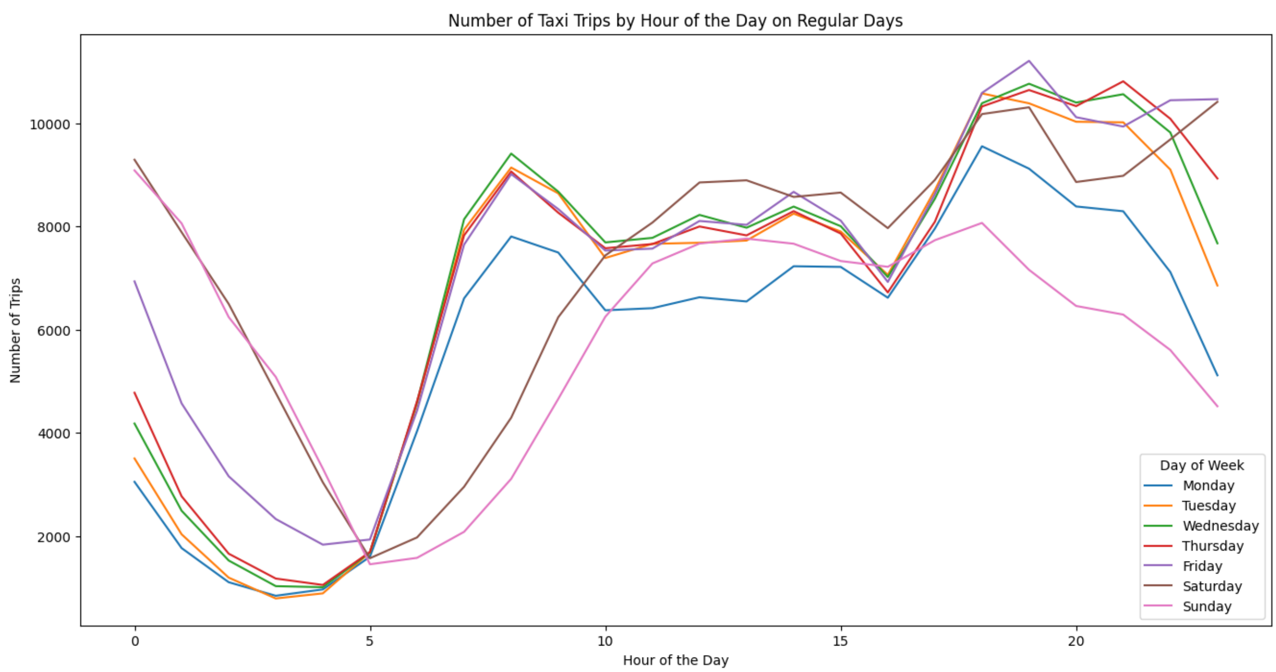
Regular Day Analysis:
Filters out holiday data to get regular days. Aggregates the number of pickups per hour for each day of the week. Reorders days of the week and ensures all hours are represented. Creates a heatmap to visualize the distribution of trips by hour on regular days.

```
Holiday dates: [datetime.date(2016, 1, 18) datetime.date(2016, 2, 14)
 datetime.date(2016, 3, 17) datetime.date(2016, 3, 27)
 datetime.date(2016, 5, 30)]
Are holidays present in the dataset?
date
False    1421832
True       36812
Name: count, dtype: int64
Filtered holiday data:
            id  vendor_id     pickup_datetime    dropoff_datetime  \
25   id0129640          2 2016-02-14 13:27:56  2016-02-14 13:49:19
121  id2648478          1 2016-01-18 11:13:59  2016-01-18 11:18:56
162  id0762989          2 2016-03-17 08:24:27  2016-03-17 08:26:11
277  id2437858          2 2016-03-27 11:55:02  2016-03-27 12:05:06
346  id0861216          2 2016-01-18 13:00:37  2016-01-18 13:10:57

     passenger_count  pickup_longitude  pickup_latitude  dropoff_longitude  \
25                 1        -73.956581        40.771358         -73.974968
121                1        -73.951576        40.766468         -73.960213
162                1        -73.977615        40.763573         -73.972572
277                5        -73.986832        40.761829         -73.977837
346                1        -73.990250        40.757286         -73.963982

     dropoff_latitude store_and_fwd_flag  trip_duration  hour_of_day  \
25          40.732792                  N           1283           13
121         40.760540                  N            297           11
162         40.765957                  N            104            8
277         40.792122                  N            604           11
346         40.756920                  N            620           13

    day_of_week        date
25       Sunday  2016-02-14
121      Monday  2016-01-18
162    Thursday  2016-03-17
277      Sunday  2016-03-27
346      Monday  2016-01-18
```


Number of Taxi Trips by Hour of the Day on Major Holidays
Hour of the Day
Holiday Date
2016-01-18
2016-02-14
2016-03-17
2016-03-27
2016-05-30
Number of Trips
0 167 312 194 339 217
1 105 241 102 283 140
2 68 240 56 271 129
3 51 163 43 169 69
4 35 127 42 124 63
5 50 47 68 42 42
6 102 54 157 61 65
7 128 90 332 80 71
8 213 164 341 129 80
9 257 206 312 188 115
10 243 289 289 302 156
11 308 361 266 303 209
12 333 392 271 332 230
13 328 376 295 287 254
14 366 394 349 309 242
15 347 369 300 290 257
16 337 350 233 295 279
17 352 432 350 287 285
18 419 406 404 314 287
19 369 410 412 324 267
20 333 342 454 320 246
21 298 336 441 283 264
22 254 323 421 259 211
23 180 277 395 233 204
50
100
150
200
250
300
350
400
450


Number of Taxi Trips by Hour of the Day on Major Holidays
Holiday Date
2016-01-18
2016-02-14
2016-03-17
2016-03-27
2016-05-30
Number of Trips
100
200
300
400
Hour of the Day
0
5
10
15
20





3.5

Data Preparation:

Creates a copy of the original dataframe to avoid warnings. Calculates the distance of each trip using the Haversine formula. Ensures datetime columns are in the correct format.

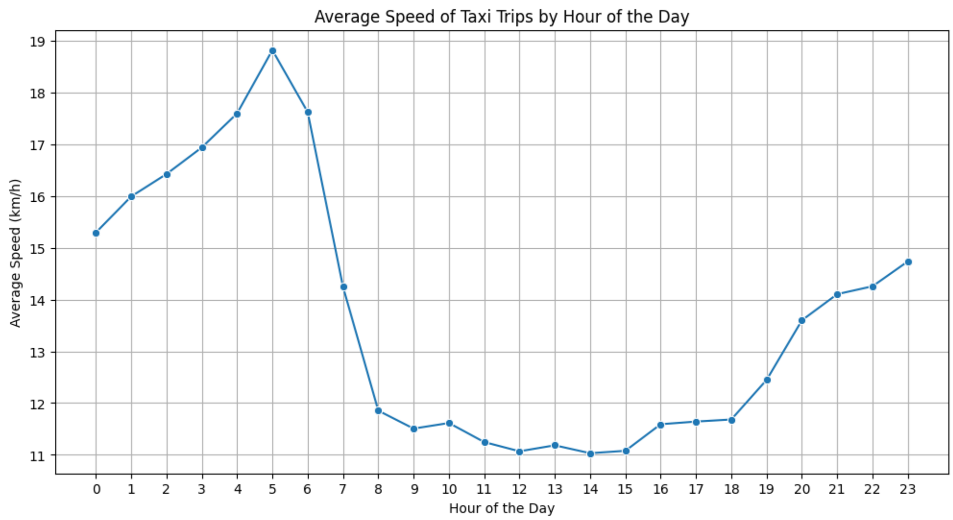
Feature Engineering:

Computes trip duration in seconds. Calculates speed in km/h for each trip. Extracts the hour of the day from the pickup datetime.

Data Cleaning:
Removes outliers identified earlier in the analysis.

Analysis:
Groups the data by hour of the day and calculates the mean speed for each hour.

Visualization:



The fastest time of day is 5:00 with an average speed of 18.82 km/h.

4 Question 4 : Location clusters

4.1 Heatmaps

To analyze the distribution of trip pickups, we generated a heatmap based on different time periods.

- **Weekday Rush Hours**: Clear peaks are visible on weekdays, with the highest number of trips occurring during evening rush hour (6-8 PM) and a secondary peak during morning rush hour (8-9 AM).
- **Weekend Late Night Activity**: Saturdays and Sundays show distinctly different patterns, with the busiest times occurring late night to early morning (11 PM to 3 AM), especially on Saturday night/Sunday morning.
- **Daily Cycle**: Across all days, trip numbers are lowest between 4-5 AM, then gradually increase throughout the day, peaking in the evening on weekdays and late night on weekends.
- **Friday Surge**: Friday consistently shows the highest overall activity, particularly in the evening and night hours, likely due to a combination of commuter traffic and the start of weekend social activities.
- **Weekday vs Weekend Contrast**: Weekdays display structured patterns aligned with typical work schedules, while weekends show a more evenly distributed trip pattern throughout the day and night, reflecting leisure and social behaviors.

4.2 Hotspots

To identify popular taxi hotspots, we used DBSCAN clustering on two different time frames: from 23:00 on a Friday evening to 02:00 on a Saturday morning, and from 17:00 to 20:00 on a Thursday

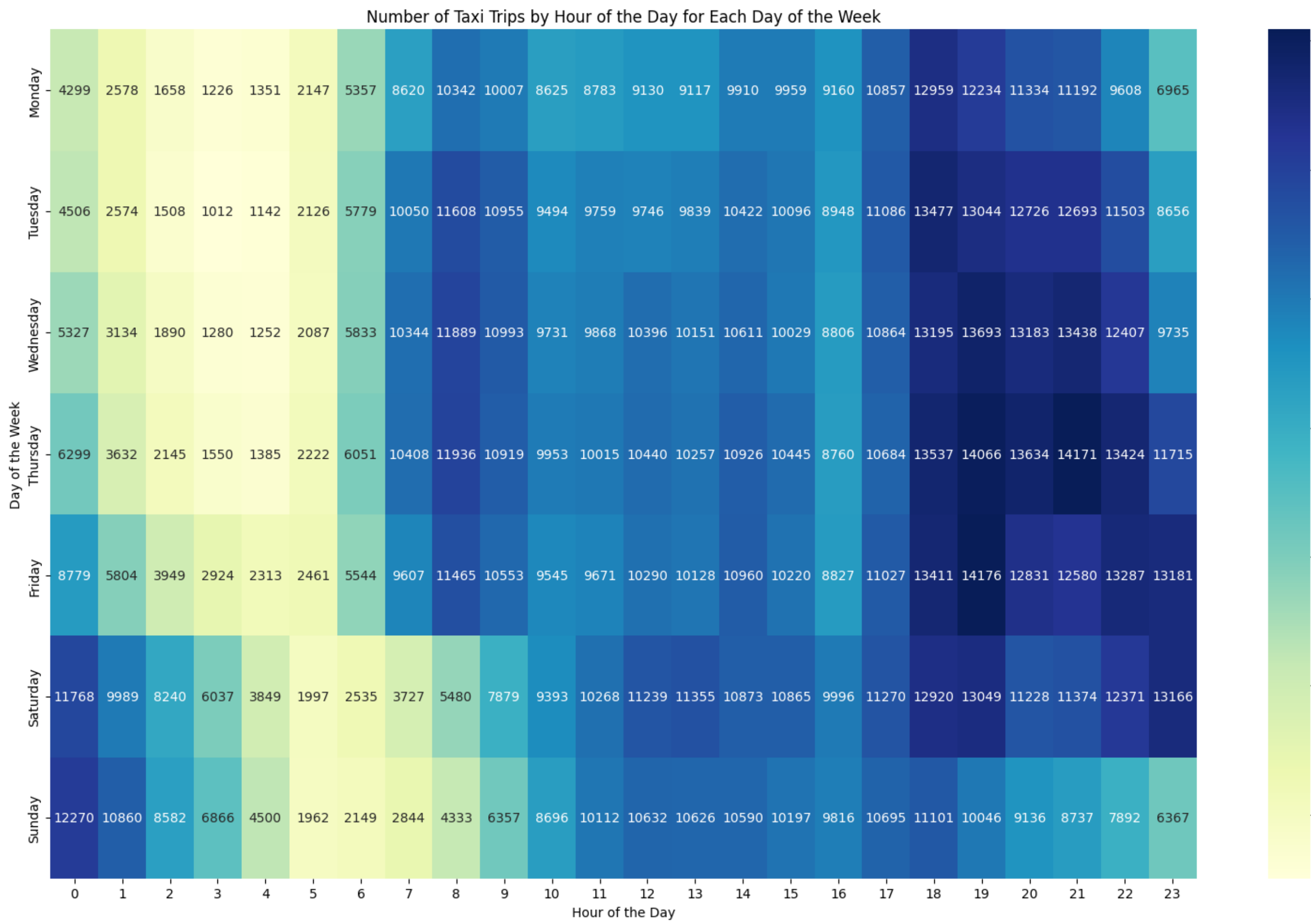


Figure 1: Heatmap of Trip Pickups by Time of Day and Hour

evening. We defined a hotspot as a cluster of at least 15 pickups within a maximum distance of 75 meters. The DBSCAN algorithm determined the number of clusters based on these parameters.

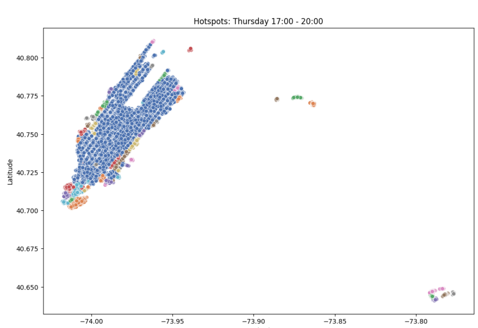


Figure 2: Hotspot Locations: Thursday 17:00 - 20:00

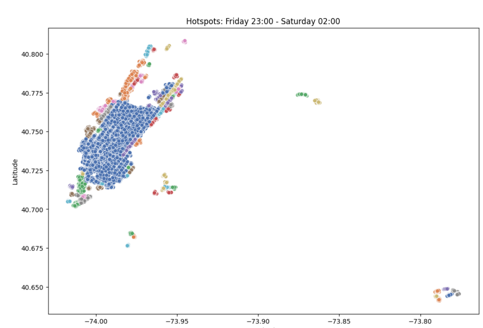


Figure 3: Hotspot Locations: Friday 23:00 - 02:00

Figure 4: Hotspot Locations for Different Time Periods

fig. 1 illustrates:

- **Cluster Comparison:** More hotspots were identified for Friday night/Saturday morning (87) than Thursday evening (71), suggesting more dispersed taxi activity during weekend nights.
- **Hotspot Variation:** Different cluster sizes and colors suggest varying levels of taxi demand across locations, with some areas showing more intense activity than others.

5 Question 5 : Airports

We examined the average travel time from the Empire State Building to JFK and Newark Airports. The radii for each location were set using the $75th$ percentile of the distance distribution to effectively cover typical travel distances. For the Empire State Building, a radius of 3.67 km was chosen. For JFK Airport, a radius of 21.49 km was used, and for Newark Airport, 19.58 km. This approach ensures the radii reflect the common travel distances while minimizing the influence of outliers.

Table 1: Coordinates and radii for the selected locations.

Location	Coordinates	Radius (km)
Empire State Building	(40.756724, -73.983806)	3.67
JFK Airport	(40.647929, -73.777813)	21.49
Newark Airport	(40.689442, -74.173242)	19.58

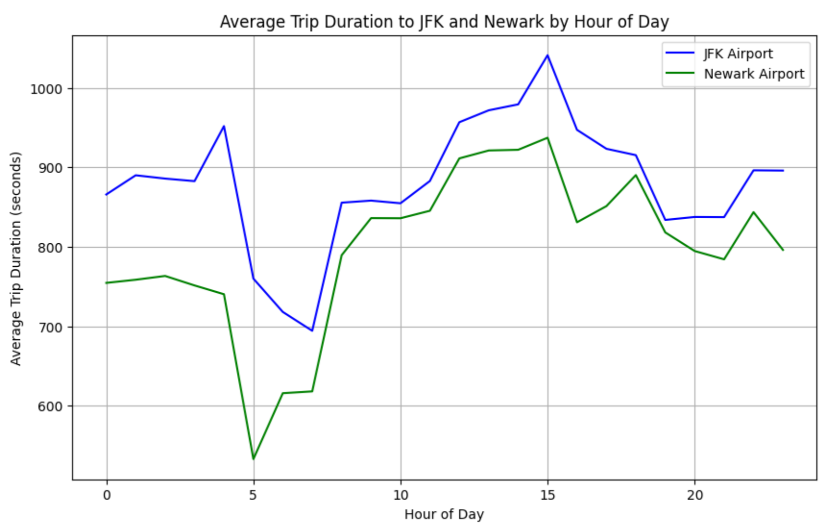


Figure 5: Average Trip Duration from the Empire State Building to JFK and Newark Airports

The plot in fig. 5 illustrates the average trip duration from the Empire State Building to JFK and Newark Airports throughout the day. The results indicate that trips to JFK Airport generally have longer average durations compared to trips to Newark Airport. This trend is observed consistently throughout different times of the day.The results reveal several key patterns:

- **Daily Variations:**
  - Both JFK and Newark Airport trip durations show significant fluctuations by the hour. Notably, average trip durations increase during the early morning hours, particularly between 6 and 7 AM. This increase is likely due to peak traffic times, as people commute to work or catch early flights.
  - After 3 PM, there is a general decrease in average trip durations to both airports. This decline may be attributed to reduced traffic congestion as the day progresses and rush hour subsides.
- **Geographical and Traffic Factors:**
  - The consistently higher average travel times to JFK compared to Newark are likely due to the greater geographical distance and potentially more congested or less direct routes to JFK Airport.

6 Question 6: Boroughs

6.1 Neighbourhoods

Using the provided shapefile, we identified the neighborhoods for the trip start and end locations using GeoPandas. The unique neighborhoods for the pickup locations and dropoff locations respectively are:

```
['Lincoln Square', 'Murray Hill-Kips Bay', 'Midtown-Midtown South',
 'SoHo-TriBeCa-Civic Center-Little Italy', 'Upper West Side', 'Gramercy', ...]
```

```
['Upper East Side-Carnegie Hill', 'West Village',
 'Battery Park City-Lower Manhattan',...,]
```

6.2 Chloropeth

The map in fig. 6 highlighted neighborhoods with varying counts of pickup and dropoff locations. Areas with more pickups or dropoff were shown in darker colors, while those with fewer pickups appeared lighter.

The choropleth maps for pickups and dropoffs revealed that the shading intensity was the same for both cases across neighborhoods. This observation indicates that the spatial distribution of pickups and dropoffs is consistent. In other words, neighborhoods with high pickup counts also tend to have high dropoff counts, and vice versa. Implying the following :

- **Similar Distribution:** The uniformity in shading suggests that the distribution of transportation activity is balanced across neighborhoods. Areas with high levels of both pickups and dropoffs are likely major transportation hubs or popular destinations, while those with low activity are consistently less active in both respects.
- **High Traffic Areas:** Neighborhoods showing intense colors in both maps are likely experiencing high volumes of both pickups and dropoffs, indicating significant transportation activity.
- **Uniform Patterns:** The consistent shading indicates that there is no significant difference in where trips start and end. This might be due to even distribution of trips or uniform service patterns across neighborhoods.

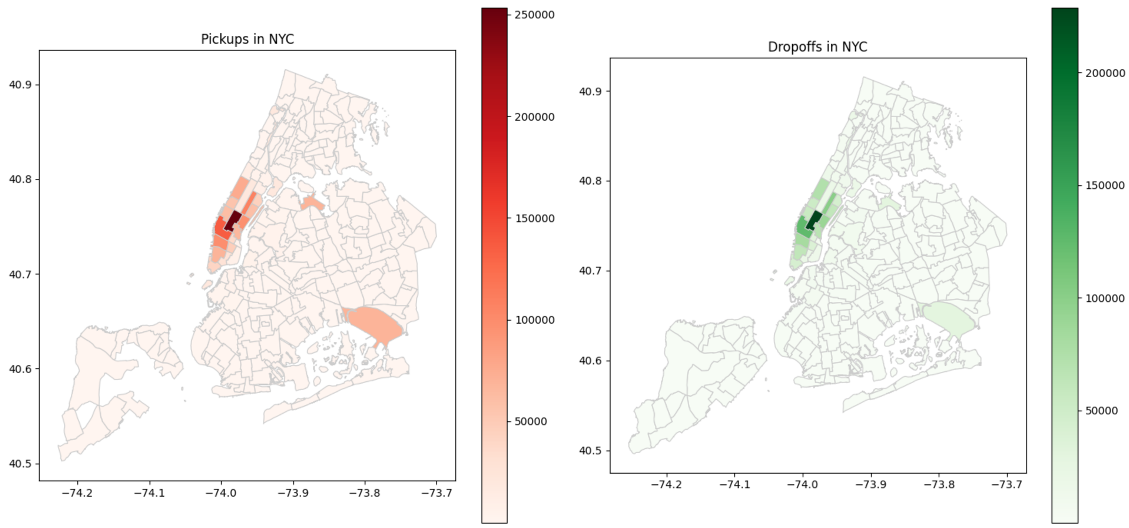


Figure 6: Cloropeth for Pickup and Dropoff trips in NYC

6.3 Quietest Neighborhoods Between Midnight and 5 AM

To determine the quietest neighborhood(s) between midnight and 5 AM, we filtered trips that started and ended within this time window and counted the number of trips in each neighborhood. We define "quiet" as having the fewest trips. The table below shows the top 3 neighborhoods with the fewest trips, indicating they are the least busy during this time period. The quietest neighborhood, with only a single trip, is Oakwood-Oakwood Beach.

Table 2: Top 3 quiet neighborhoods with the fewest trips between midnight and 5 AM.

Neighborhood	Trip Count
Oakwood-Oakwood Beach	1
Annadale-Huguenot-Prince's Bay-Eltingville	2
Old Town-Dongan Hills-South Beach	2

6.4 Busiest Neighborhoods Between Midnight and 5 AM

To determine the busiest neighborhood(s) between midnight and 5 AM, we filtered trips that started and ended within this time window and counted the number of trips in each neighborhood. We define "busiest" as having the most trips. The table below shows the top 3 neighborhoods with the most trips, indicating they are the busiest during this time period. The busiest neighborhood, with the highest number of trips, is Midtown-Midtown South.

Table 3: Top 3 busiest neighborhoods with the most trips between midnight and 5 AM.

Neighborhood	Trip Count
Midtown-Midtown South	14,507
Hudson Yards-Chelsea-Flatiron-Union Square	11,141
West Village	8,727