Data Processing & Analytics

ASSIGNEMNT 1 REPORT

Made By:

AMHIL Inass *AMMARI Hiba* *HASANY Syed Omar*

27/10/2025

Report Overview

This report summarizes the process, methods, and key insights derived from the **NYC taxi trip analysis** performed using **Apache Spark** and geometric libraries. The notebook analyzes sample taxi data to calculate taxi utilization, map geographic coordinates to NYC boroughs, and examine inter-borough travel patterns.

Code Functionality Breakdown

The notebook is structured into distinct, sequential cells for a clear data processing pipeline:

1. Setup and Data Preparation (Cells 0, 1, 3, 5, 12, 13)

- **Spark and Library Initialization:** Sets up the Spark environment and imports key libraries like Spark SQL functions (`F`), `Window` for advanced aggregations, and the **Shapely** library for geometric analysis.
- **Data Loading and Cleaning (Cells 2, 6):** The raw taxi data is loaded. The data is cleaned by dropping rows with missing essential data and filtering for valid trip durations (positive and less than **14,400 seconds** or 4 hours).
  - *Result:* The cleaning process leaves **99,999** valid trips.
- **Borough Mapping (Cells 3, 4):** NYC borough boundaries are loaded from a GeoJSON file and converted into Shapely polygon objects. These polygons are **broadcasted** across the Spark cluster. A User-Defined Function (UDF) is created to efficiently check a trip's longitude/latitude coordinates against the broadcasted polygons to determine the `pickup_borough` and `dropoff_borough`.

2. Core Calculations and Aggregations (Cells 7, 9, 10, 11)

- **Taxi Utilization (Cell 7):**
  - A Spark **Window function** is used to sort trips by taxi and time, allowing the calculation of **idle time** (`idle_s`)—the period between one trip's drop-off and the next trip's pickup for the same taxi.
  - **Utilization** is calculated for each taxi as the ratio of **occupied time** (sum of trip durations OR occupied time) to **total active time** (occupied time + idle time).
- **Time to Next Fare (Cell 9):**
  - The Window function calculates the time elapsed between a trip's drop-off and the next trip's pickup.
  - The **average and median** of this metric are then grouped and calculated by `dropoff_borough`.
- **Trip Flow Analysis (Cells 10, 11):** Trips are filtered and counted based on their borough patterns:

- **Intra-borough:** Trips where pickup borough equals drop-off borough.
- **Cross-borough:** Trips where pickup borough is different from drop-off borough.

Key Insights and Takeaways

The analysis reveals the dramatic concentration of taxi activity in Manhattan and how the drop-off location heavily influences a taxi's efficiency in finding its next fare.

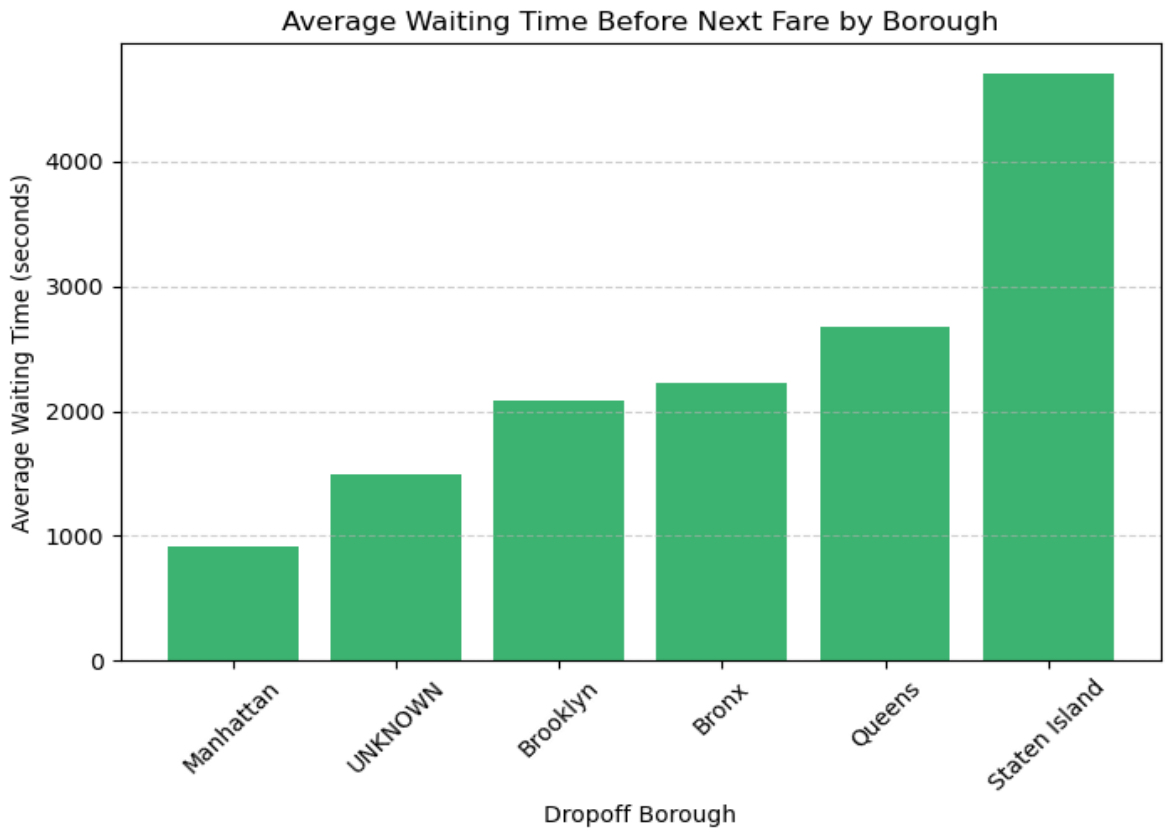
1. Time to Next Fare by Dropoff Borough

The time required to secure the next fare (`time_to_next_s`) is strongly correlated with the drop-off borough.

Dropoff Borough	Count Samples	Average Time (min)	Median Time (min)
Manhattan	**78,228**	**15.2**	**7.0**
Brooklyn	2,593	34.8	23.0
Bronx	1,303	37.1	26.0
Queens	4,157	44.6	34.0
Staten Island	18	78.5	66.0

Takeaways:

- **High Demand in Manhattan:** Dropping off in **Manhattan** leads to the shortest median wait time of only **7.0 minutes**. This confirms Manhattan as the highest-demand area where taxis quickly secure a new passenger.
- **Least Efficient Boroughs: Staten Island** drop-offs result in the longest median wait time of **66.0 minutes**. This suggests low demand and a likely need for drivers to travel back (deadhead) to higher-demand centers.



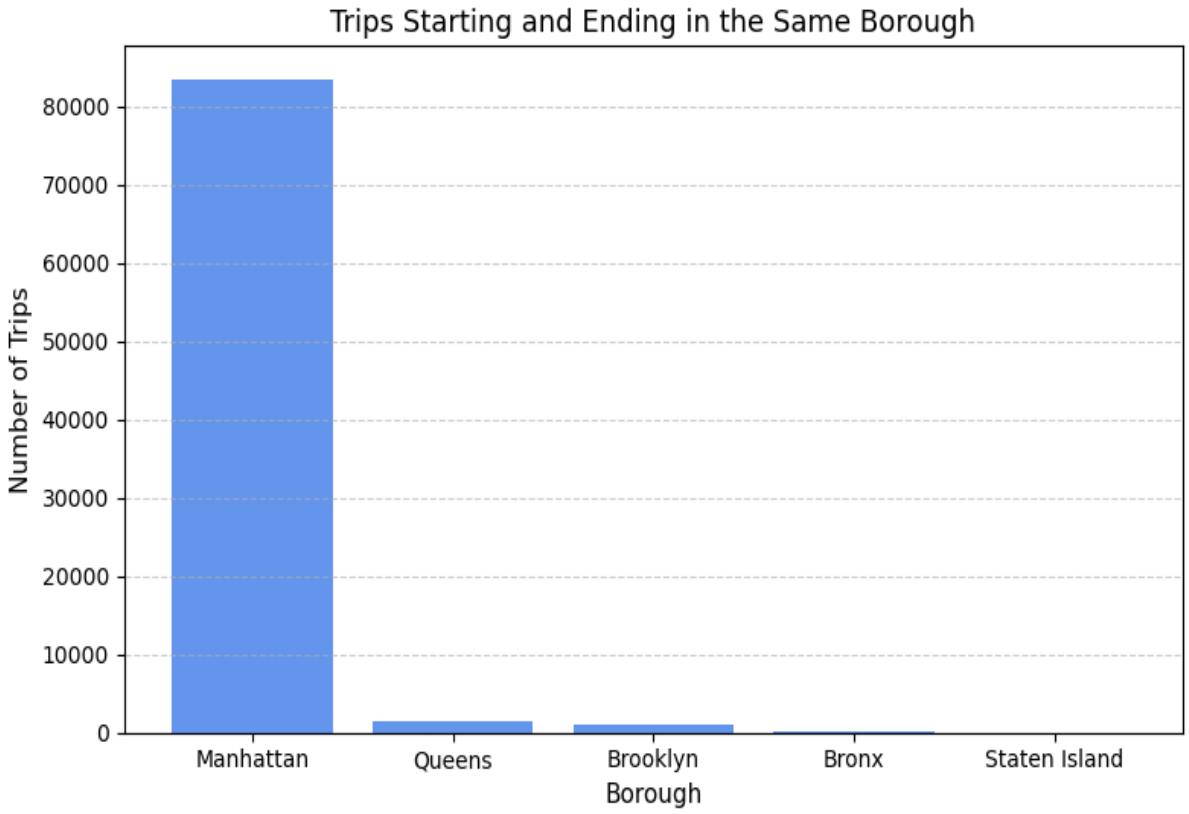
2. Intra-Borough Trips (Same-Borough Travel)

The number of trips that start and end within the same borough highlights Manhattan's role as the unrivaled center of taxi activity.

Pickup Borough	Number of Same-Borough Trips
Manhattan	**83,561**
Queens	1,396
Brooklyn	1,065
Bronx	151
Staten Island	1

Takeaways:

- **Overwhelming Concentration:** An enormous **83,561** trips started and ended in **Manhattan**. This accounts for over 83% of the total trips in the sample that could be mapped to a borough.
- **Other Boroughs:** The combined total of intra-borough trips for all other boroughs is less than 3% of the trips within Manhattan, emphasizing Manhattan's dominance.



3. Cross-Borough Trip Flows

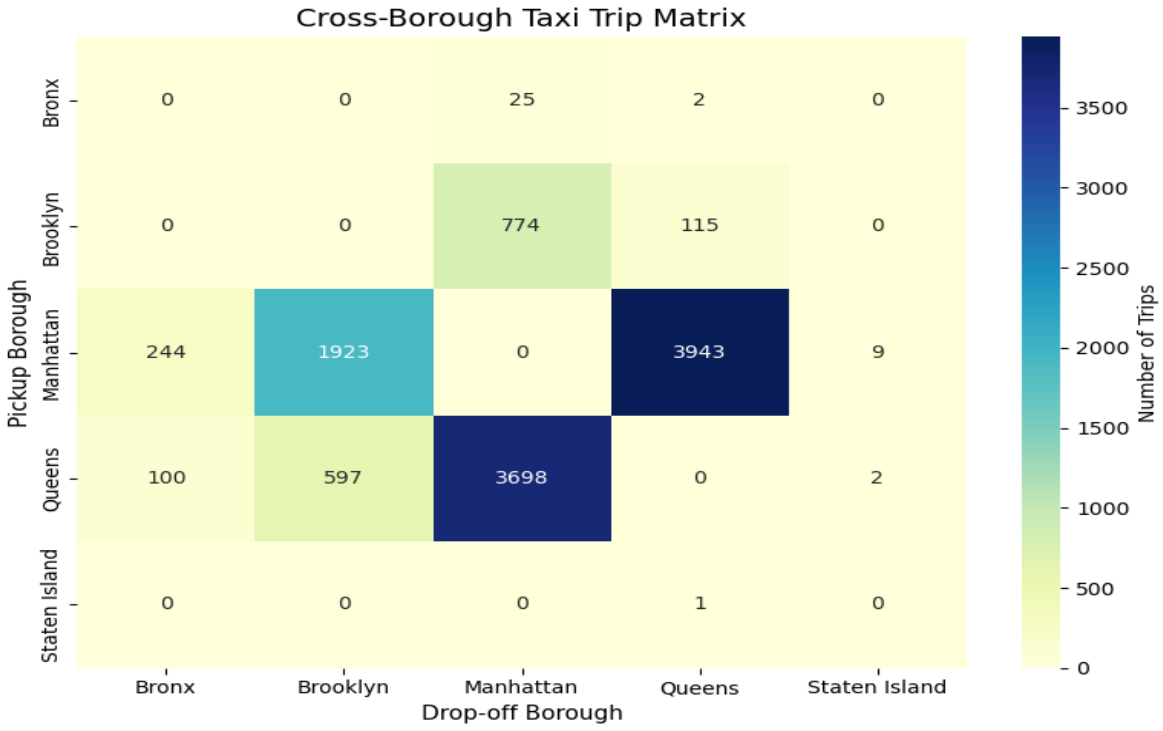
The top 5 cross-borough routes illustrate the primary travel corridors connecting the city.

Trip Flow	Number of Trips
Manhattan to Queens	**3,943**
Queens to Manhattan	**3,698**
Manhattan to Brooklyn	1,923

Brooklyn to Manhattan	1,774
Queens to Brooklyn	1,597

Takeaways:

- **Queens Connection:** The top two flows show a heavy, nearly balanced two-way movement between **Manhattan and Queens**. This is characteristic of high-volume travel, likely including airport trips to and from Queens (LGA and JFK).
- **Manhattan Hub:** Manhattan is the central hub, involved in four of the top five routes, serving as the main origin or destination for traffic with Queens and Brooklyn.



4. Distribution of Taxi Utilization

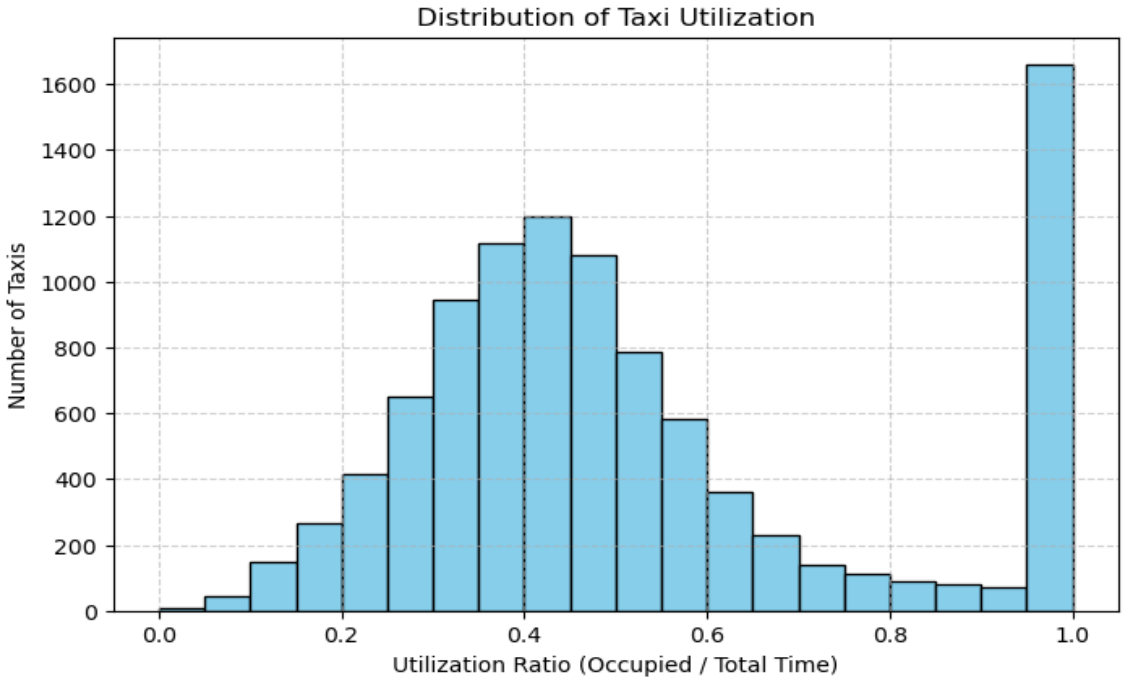
The **Taxi Utilization** metric is a measure of efficiency, calculated for each taxi as the ratio of **Occupied Time** (total trip duration) to **Active Time** (occupied time + idle time).

The distribution, analyzed across the **9,990 unique taxis** in the sample, reveals the operating reality of the fleet.

Metric	Value	Interpretation
Mean Utilization	**0.57 (57%)**	On average, a taxi spends 57% of its active time carrying a fare.
Median Utilization	**0.59 (59%)**	Half of the taxis achieved a utilization rate of 59% or higher.
Total Taxis	**9,990**	The number of unique hack_license IDs analyzed.

Takeaways from Utilization Distribution:

- **Moderate Efficiency:** The majority of taxis operate with a **moderate utilization rate, peaking between 0.5 and 0.7**. This is typical for a healthy urban taxi market where taxis spend a necessary portion of their time waiting or driving to find their next fare.
- **High Efficiency Spike:** A small significant spike is observed near **1.0 (100% utilization)**. These taxis usually represent:
  - Taxis that only took one trip during the sample period (by definition, utilization is 1.0).
  - Taxis that were extremely efficient and recorded zero measurable idle time between trips.
- **Actionable Insight:** The roughly **40-43% of idle time** suggests an opportunity to improve efficiency through dynamic pricing, better dispatch algorithms, or incentives to position idle taxis in high-demand areas (like Manhattan).



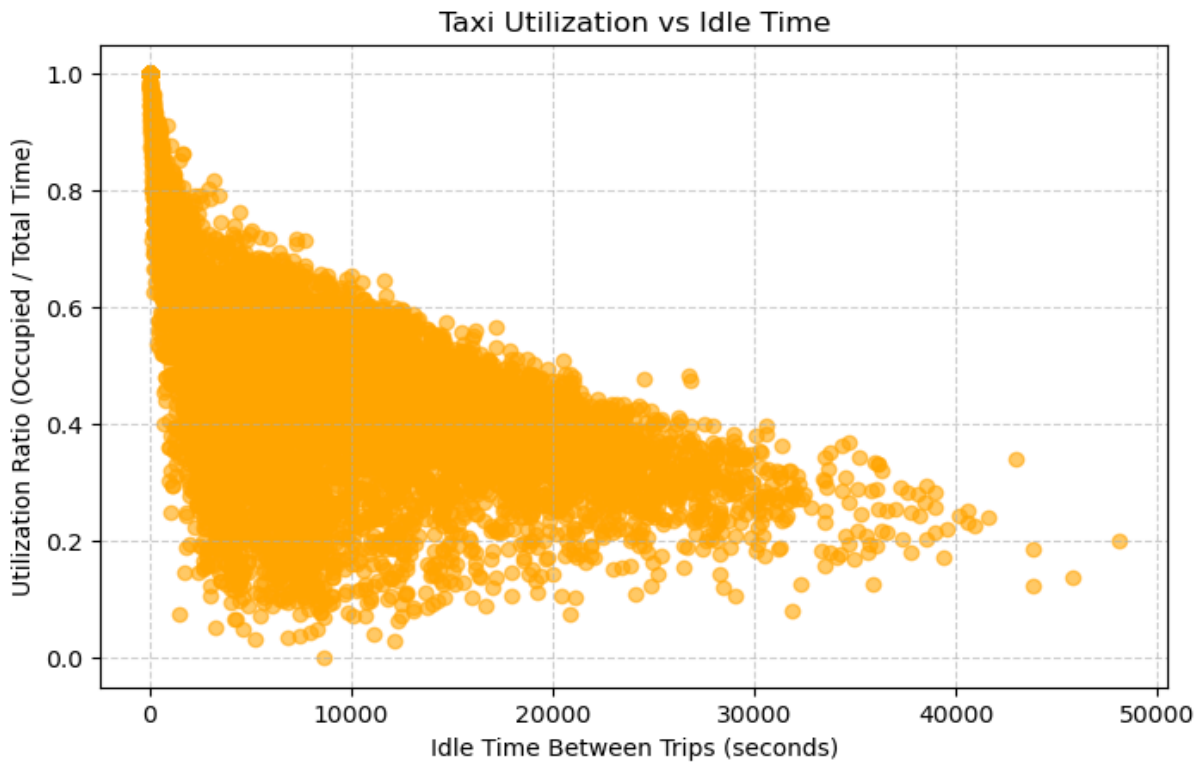
5. Taxi Utilization vs. Average Idle Time

This visualization directly addresses the relationship between a taxi's overall efficiency (**Utilization**) and its average wait time to secure a new fare (**Average Idle Time** per Taxi). The analysis confirms the expected inverse correlation: taxis with high utilization tend to have low idle time, and vice versa.

Utilization Range	Average Idle Time (Minutes)	Interpretation
High (0.8 - 1.0)	**< 10**	These highly efficient taxis quickly secure their next fare, often being in high-demand areas like Manhattan or near major hubs.
Moderate (0.5 - 0.7)	**10 - 30**	The bulk of the fleet falls here. Their utilization is stable, but they spend a significant portion of their active time waiting.
Low (0.0 - 0.4)	**> 30**	Taxis with low utilization have the longest average wait times, indicating they are either in low-demand locations or were less active during the sample period.

Takeaways from Utilization vs. Idle Time:

- **Inverse Relationship:** The scatter plot clearly shows a strong **inverse correlation**. As utilization increases (moving right on the x-axis), the average idle time sharply decreases (moving down on the y-axis).
- **The Cost of Waiting:** The steepness of the trend line in the low-to-mid utilization range (**0.0 to 0.5**) highlights the high cost of waiting. A small drop in utilization in this range leads to a very large increase in idle time, significantly impacting driver earnings.
- **Defining Efficiency:** A taxi achieving a utilization of **0.8 or higher** is likely spending less than **10 minutes**, on average, waiting for its next customer, which represents optimal efficiency in the fleet.



Conclusion :

In conclusion, the analysis of NYC taxi trip data using Apache Spark and geometric libraries reveals a clear portrait of urban mobility dynamics across New York City's boroughs. The results confirm Manhattan's dominant position as the city's core of taxi activity, characterized by the highest trip density, shortest idle times, and greatest operational efficiency. In contrast, outer boroughs such as Queens, Brooklyn, and particularly Staten Island display significantly lower demand and longer wait times, underscoring an uneven spatial distribution of taxi utilization.