### **DIVE Analysis Report: Operational Insights from NYC Taxi Data (2022)**

**Analyst:** Grace Bosma, Operational Excellence

**Date:** August 13, 2025

**Dataset:** bigquery-public-data.new\_york\_taxi\_trips.tlc\_yellow\_trips\_2022

**Objective:** Synthesize data from the Discover, Investigate, Validate and Extend phases to provide a holistic view of operational performance and identify key areas for improvement.

### **Discover: Operational Bottlenecks**

The first set of results from the refined Discover queries successfully identified key areas and times with the slowest average speeds.

#### **Key Findings:**

* **Morning Rush is a Major Bottleneck:** The lowest average speeds, between 8.3 and 8.9 mph, are consistently found during the morning commute hours of **7 AM to 10 AM**. This confirms that traffic congestion during this period is a primary factor affecting service speed.
* **Widespread Bottlenecks:** The issue is not confined to one location. Several high-volume pickup locations, including IDs **186** and **236**, are experiencing significant slowdowns. The sheer volume of trips (over 60,000 for location ID 186 and over 125,000 for location ID 236) indicates these are critical hubs where delays have a massive impact.
* **Afternoon Slowdowns:** A new insight from the refined query shows that low speeds also occur during afternoon hours (12 PM to 3 PM), suggesting that traffic is a persistent daily challenge.

### **Investigate: Revenue Drivers and High-Value Routes**

The Investigate query results reveal which routes generate the most revenue for the taxi service.

#### **Key Findings:**

* **Airport Runs are Top Revenue Drivers:** The top seven highest-revenue routes all originate from pickup\_id **132**. This location ID corresponds to John F. Kennedy International Airport (JFK).
* **Consistent High-Value Destinations:** The most profitable routes are consistently to drop-off locations **265** and **230**. These locations are major hubs, and the high revenue suggests these are long-distance trips.
* **Weekday and Weekend Consistency:** These high-value routes are not limited to one day of the week. They consistently generate high revenue on both weekdays and weekends (day\_of\_week 1, 2, 4, 5, 6), indicating a stable, high-value source of income that is crucial to protect.

### **Validate: Data Quality and Integrity**

The Validate query successfully identified critical issues in the raw data that would have compromised the entire analysis.

#### **Key Findings:**

* **Impossible Outliers:** The data contained numerous trips with impossible distances (hundreds of thousands of miles) but normal durations. These are clear data entry errors.
* **Validation is Essential:** This finding demonstrates the importance of the Validate phase. Without the data cleaning step, any calculation of average speed or revenue per mile would have been completely skewed by these erroneous records.

### **Extend: Areas of Improvement**

The DIVE analysis has provided a clear picture of operational strengths and weaknesses. The following recommendations are designed to capitalize on strengths and address identified bottlenecks:

1. **Prioritize Bottleneck Optimization:** Focus resources on analyzing the root cause of slow speeds in high-volume pickup locations like **186**, **236**, and **237**, particularly during rush hours.
2. **Protect High-Value Routes:** The routes originating from JFK Airport are a critical revenue stream. Monitor performance on these routes closely to ensure reliability and customer satisfaction, as any operational issue here would have a significant financial impact.
3. **Implement Automated Data Validation:** Moving forward, all analytical queries should include filters to exclude impossible data points, such as those found in the trip\_distance column, to ensure the integrity of future reports.
4. **Actionable Insights for the "Extend" Phase:** Use the **Extend** query to drill down into the demand patterns for high-volume and high-value locations to inform future strategies for AI-driven dispatch or predictive scheduling.

### **Appendix**

### **1. Discover: Slowest Zones and Times**

This query identifies the locations and hours with the lowest average speeds, filtering for statistically significant results and excluding erroneous data.

SQL

SELECT

pickup\_location\_id,

EXTRACT(HOUR FROM pickup\_datetime) AS pickup\_hour,

AVG(TIMESTAMP\_DIFF(dropoff\_datetime, pickup\_datetime, MINUTE)) AS avg\_trip\_duration\_minutes,

AVG(trip\_distance / TIMESTAMP\_DIFF(dropoff\_datetime, pickup\_datetime, SECOND) \* 3600) AS avg\_speed\_mph,

COUNT(\*) AS trip\_count

FROM

`bigquery-public-data.new\_york\_taxi\_trips.tlc\_yellow\_trips\_2022`

WHERE

data\_file\_year = 2022

AND trip\_distance > 0

AND trip\_distance < 200

AND TIMESTAMP\_DIFF(dropoff\_datetime, pickup\_datetime, SECOND) > 0

GROUP BY

pickup\_location\_id,

pickup\_hour

HAVING

trip\_count > 100 AND avg\_speed\_mph > 1

ORDER BY

avg\_speed\_mph ASC;

### **2. Investigate: Most Valuable Routes**

This query identifies the top revenue-generating routes based on total trip count and revenue.

SQL

SELECT

pickup\_location\_id AS pickup\_id,

dropoff\_location\_id AS dropoff\_id,

EXTRACT(DAYOFWEEK FROM pickup\_datetime) AS day\_of\_week,

COUNT(\*) AS trip\_count,

SUM(total\_amount) AS total\_revenue

FROM

`bigquery-public-data.new\_york\_taxi\_trips.tlc\_yellow\_trips\_2022`

WHERE

data\_file\_year = 2022

GROUP BY

pickup\_id,

dropoff\_id,

day\_of\_week

ORDER BY

total\_revenue DESC

LIMIT 10;

### **3. Validate: Data Outliers**

This query helps you find potential data errors by looking for unrealistic trip distances or durations.

SQL

SELECT

trip\_distance,

TIMESTAMP\_DIFF(dropoff\_datetime, pickup\_datetime, SECOND) AS trip\_duration\_seconds

FROM

`bigquery-public-data.new\_york\_taxi\_trips.tlc\_yellow\_trips\_2022`

WHERE

data\_file\_year = 2022

AND (trip\_distance > 200 OR TIMESTAMP\_DIFF(dropoff\_datetime, pickup\_datetime, SECOND) > 10800)

ORDER BY

trip\_distance DESC

LIMIT 10;

### **4. Extend: Demand Peaks**

This query identifies the pickup locations and times with the highest trip counts, which is valuable for demand forecasting and fleet management.

SQL

SELECT

pickup\_location\_id,

EXTRACT(DAYOFWEEK FROM pickup\_datetime) AS day\_of\_week,

EXTRACT(HOUR FROM pickup\_datetime) AS pickup\_hour,

COUNT(\*) AS trip\_count

FROM

`bigquery-public-data.new\_york\_taxi\_trips.tlc\_yellow\_trips\_2022`

WHERE

data\_file\_year = 2022

GROUP BY

pickup\_location\_id,

day\_of\_week,

pickup\_hour

ORDER BY

pickup\_location\_id,

day\_of\_week,

pickup\_hour;

**Key Gemini Prompts:**

“are you familiar with bigquery free datasources? I'm looking at the NYC yellow taxi data for 2022, dataset ID is bigquery-public-data.new\_york\_taxi\_trips and table is tlc\_yellow\_trips\_2022”

“I'm using that table, filtering everything to 2022. what do you need to know to answer this information: Operational Excellence Analyst: Grace Efficiency, processes, performance metrics…”

“consider a DIVE analysis based on the above info and give me SQL queries if you need me to find specific data”

“this is the actual list of column names:

vendor\_id

pickup\_datetime

dropoff\_datetime

passenger\_count

trip\_distance

rate\_code

store\_and\_fwd\_flag

payment\_type

fare\_amount

extra

mta\_tax

tip\_amount

tolls\_amount

imp\_surcharge

airport\_fee

total\_amount

pickup\_location\_id

dropoff\_location\_id

data\_file\_year

data\_file\_month”

\*also submitted query results received through Gemini recommended SQL\*

"I'm using that table, filtering everything to 2022. what do you need to know to answer this information: Operational Excellence Analyst: Grace Efficiency, processes, performance metrics...[followed by a detailed list of questions]"

"complete a DIVE analysis based on the above info and give me SQL queries if you need me to find specific data"

"please rewrite with different questions within each part of the DIVE if we can't answer it well with this table"

"for the first query Table bigquery-public-data:new\_york\_taxi\_trips.tlc\_taxi\_zones was not found"

"Not found: Table bigquery-public-data:new\_york\_taxi\_trips.tlc\_taxi\_zones was not found in location US"

"i have it set to automatic location"

"redo, this is the actual list of column names vendor\_id...[followed by a list of column names]"

"top results for SQL in Discover Q1: pickup\_location\_id...[followed by query results]"

"can you share those SQL queries again please"

"results for validate: trip\_distance...[followed by query results]"

"top few rows of new discover query: pickup\_location\_id...[followed by query results]"

"can you also ensure every query is filtering to 2022 data"

"New discover results: pickup\_location\_id...[followed by multiple sets of query results]"

"consider the DIVE analysis above, what points would you highlight in a powerpoint slide?"

"list top 4 things since it will all be in one slide"