Hemansh Adunoor

# NYC Taxi Dataset Report

# 1. Table of Contents


| |
|---|
| 1. Table of Contents |
| 2. Table of figures and tables |
| 3. Abstract |
| 4. Introduction |
| 5. Description of the dataset |
| 6. Pre-processing dataset |
| 7. Outlier Detection and Removal |
| 8. Principal Component Analysis |
| 9. Normality Test |
| 10. Data transformation |
| 11. Heatmap & Pearson correlation coefficient matrix |
| 12. Statistics |
| 13. Data visualization and Observations |
| 14. Subplots and Tables |
| 15. Dashboard |
| 16. Conclusion |
| 17. Appendix |
| 18. References |


# 2. Table of figures and tables

>>> First 5 Observations of Cleaned Data [PART 1 of 2]

| index | id | vendor_id | pickup_datetime | dropoff_datetime | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude |
|---|---|---|---|---|---|---|---|---|---|
| 0 | id2875421 | 2 | 2016-03-14 17:24:55 | 2016-03-14 17:32:30 | 1 | -73.98215484619139 | 40.76793670654297 | -73.96463012695312 | 40.765602111816406 |
| 1 | id2377394 | 1 | 2016-06-12 00:43:35 | 2016-06-12 00:54:38 | 1 | -73.98041534423827 | 40.738563537597656 | -73.9994812011718 | 40.731151580810554 |
| 2 | id3858529 | 2 | 2016-01-19 11:35:24 | 2016-01-19 12:10:48 | 1 | -73.97902679443358 | 40.763938903808594 | -74.00533294677734 | 40.710086822509766 |
| 3 | id3504673 | 2 | 2016-04-06 19:32:31 | 2016-04-06 19:39:40 | 1 | -74.01004028320312 | 40.719970703125 | -74.01226806640625 | 40.70671844482422 |
| 4 | id2181028 | 2 | 2016-03-26 13:30:55 | 2016-03-26 13:38:10 | 1 | -73.97305297851561 | 40.79320907592773 | -73.9729232788086 | 40.782520294189446 |

>>> First 5 Observations of Cleaned Data [PART 2 of 2]

| index | store_and_fwd_flag | trip_duration | month | day_of_week | hour_of_day | day_of_year | distance_km | dist_from_center |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 455 | 3 | Monday | 17 | 74 | 1.4975799409833876 | 1.139548114698798 |
| 1 | 0 | 663 | 6 | Sunday | 0 | 164 | 1.8043735902884075 | 2.201888563155812 |
| 2 | 0 | 2124 | 1 | Tuesday | 11 | 19 | 6.381089643584211 | 0.8558121738592548 |
| 3 | 0 | 429 | 4 | Wednesday | 19 | 97 | 1.4845657601291107 | 4.704100657668814 |
| 4 | 0 | 435 | 3 | Saturday | 13 | 86 | 1.1878422101049733 | 4.0503917062215224 |

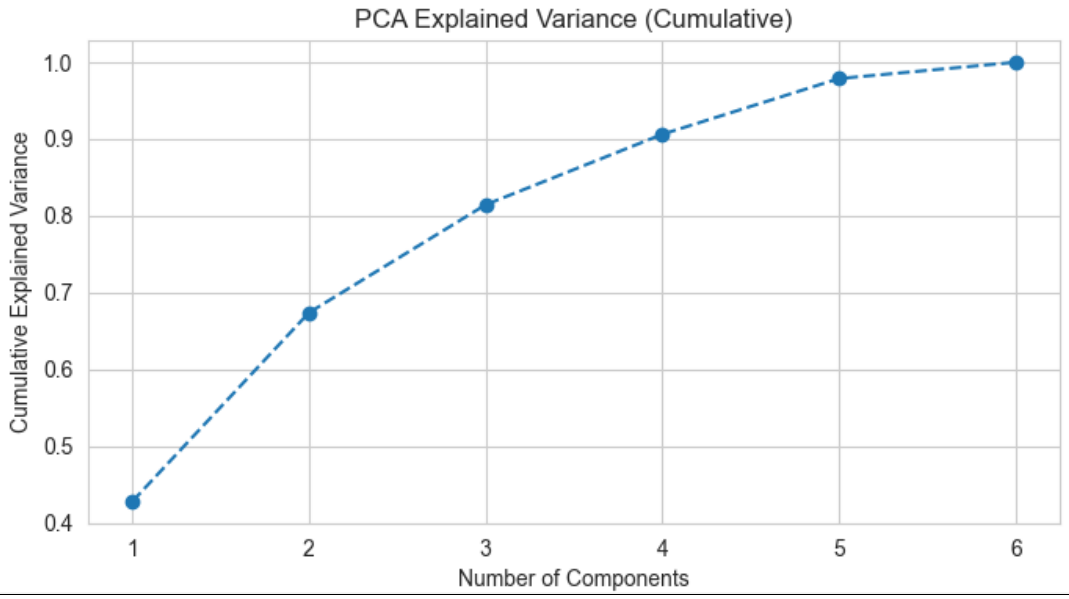
>>> Statistics of Cleaned Dataset [PART 1 of 2]

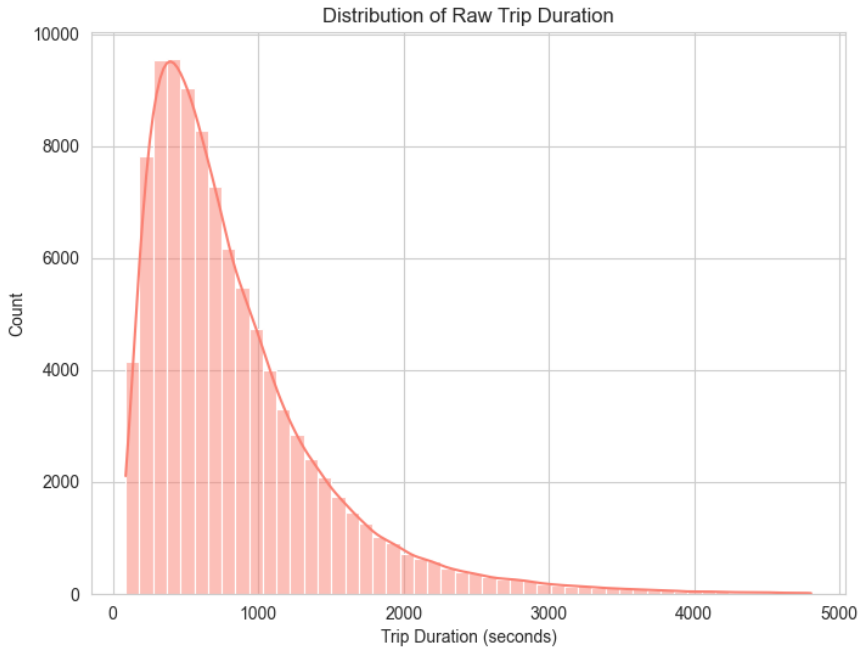
| index | vendor_id | pickup_datetime | dropoff_datetime | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude |
|---|---|---|---|---|---|---|---|---|
| count | 100000.0 | 100000 | 100000 | 100000.0 | 100000.0 | 100000.0 | 100000.0 | 100000.0 |
| mean | 1.53 | 2016-04-01 06:58:28.674840064 | 2016-04-01 07:14:08.539099904 | 1.67 | -73.97 | 40.75 | -73.97 | 40.75 |
| min | 1.0 | 2016-01-01 00:00:17 | 2016-01-01 00:09:42 | 0.0 | -74.53 | 40.43 | -74.56 | 40.44 |
| 25% | 1.0 | 2016-02-17 10:51:33.750000128 | 2016-02-17 11:07:19.249999872 | 1.0 | -73.99 | 40.74 | -73.99 | 40.74 |
| 50% | 2.0 | 2016-04-01 13:36:27 | 2016-04-01 13:58:46 | 1.0 | -73.98 | 40.75 | -73.98 | 40.75 |
| 75% | 2.0 | 2016-05-14 22:51:33.500000 | 2016-05-14 23:01:23.750000128 | 2.0 | -73.97 | 40.77 | -73.96 | 40.77 |
| max | 2.0 | 2016-06-30 23:51:36 | 2016-07-01 16:37:39 | 6.0 | -73.33 | 41.32 | -72.71 | 41.31 |
| std | 0.5 | nan | nan | 1.32 | 0.04 | 0.03 | 0.04 | 0.03 |

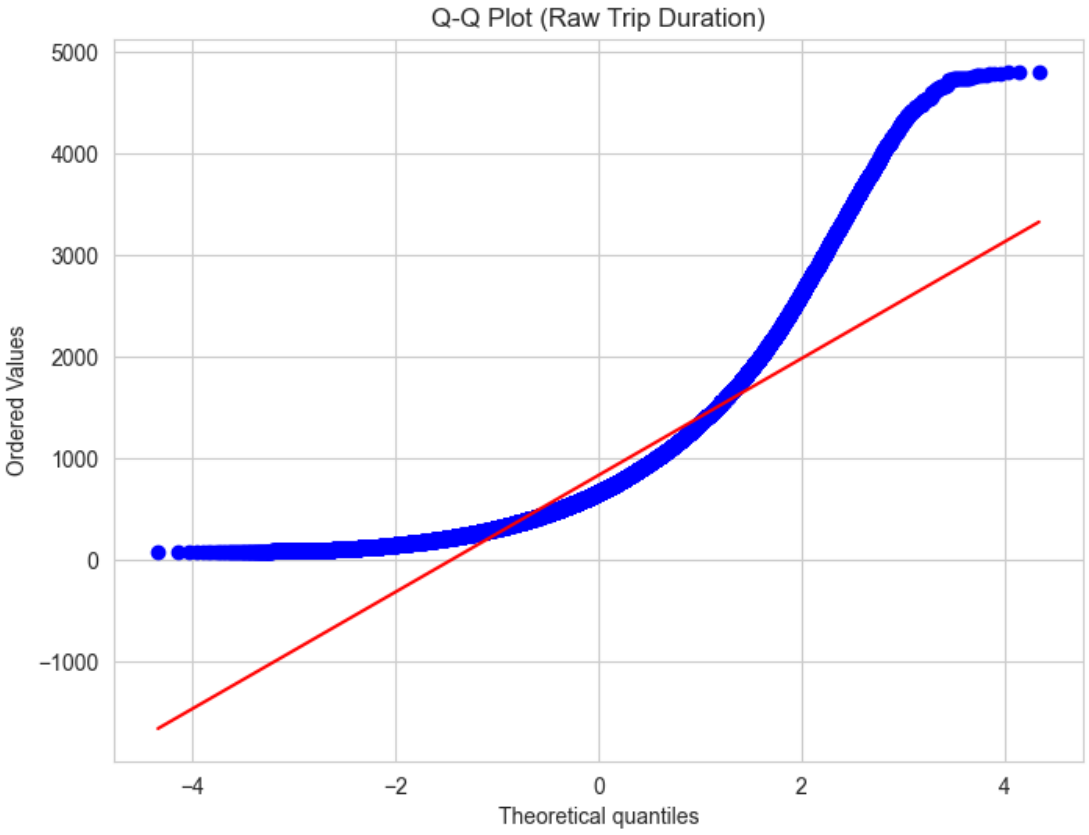
```
>>> Statistics of Cleaned Dataset [PART 2 of 2]
```

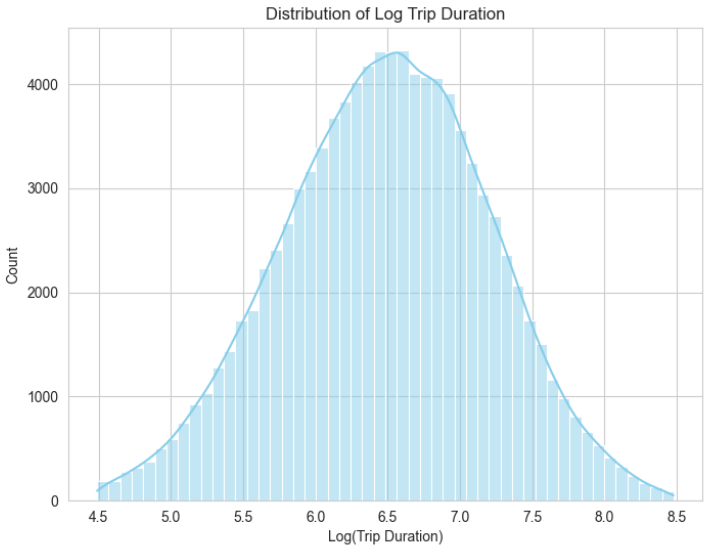
| index | store_and_fwd_flag | trip_duration | month | hour_of_day | day_of_year | distance_km | dist_from_center |
|---|---|---|---|---|---|---|---|
| count | 100000.0 | 100000.0 | 100000.0 | 100000.0 | 100000.0 | 100000.0 | 100000.0 |
| mean | 0.01 | 939.86 | 3.51 | 13.63 | 91.7 | 3.43 | 3.16 |
| min | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 25% | 0.0 | 396.0 | 2.0 | 9.0 | 48.0 | 1.23 | 1.23 |
| 50% | 0.0 | 662.0 | 4.0 | 14.0 | 92.0 | 2.09 | 2.35 |
| 75% | 0.0 | 1076.0 | 5.0 | 19.0 | 135.0 | 3.87 | 3.71 |
| max | 1.0 | 86390.0 | 6.0 | 23.0 | 182.0 | 116.42 | 63.47 |
| std | 0.08 | 3004.54 | 1.68 | 6.38 | 51.53 | 3.95 | 3.44 |

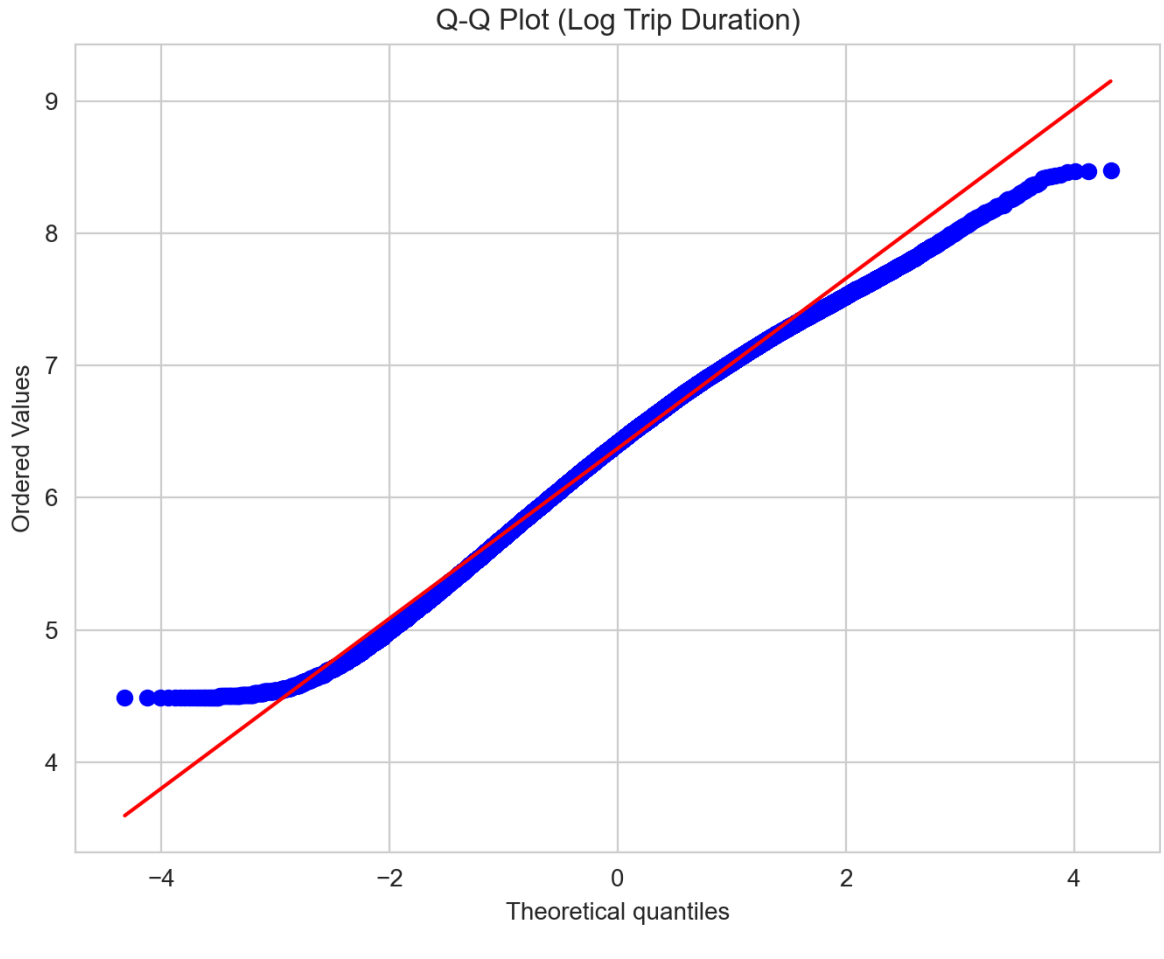
```
Method Used: Interquartile Range (IQR).
Applying filter to BOTH 'log_trip_duration' and 'distance_km'.
Duration Bounds (Log): [4.49, 8.48]
Distance Bounds (km):  [-2.72, 7.82]
Rows removed: 10846
Percentage of data removed: 10.85%
```

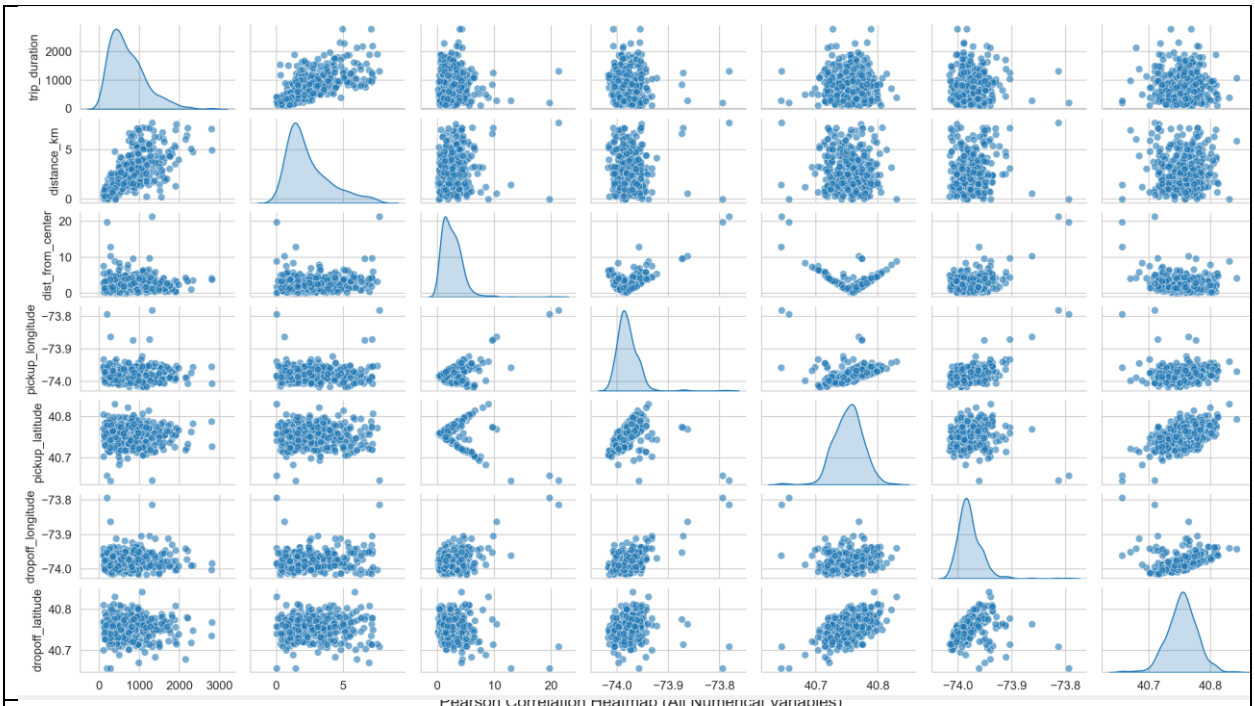


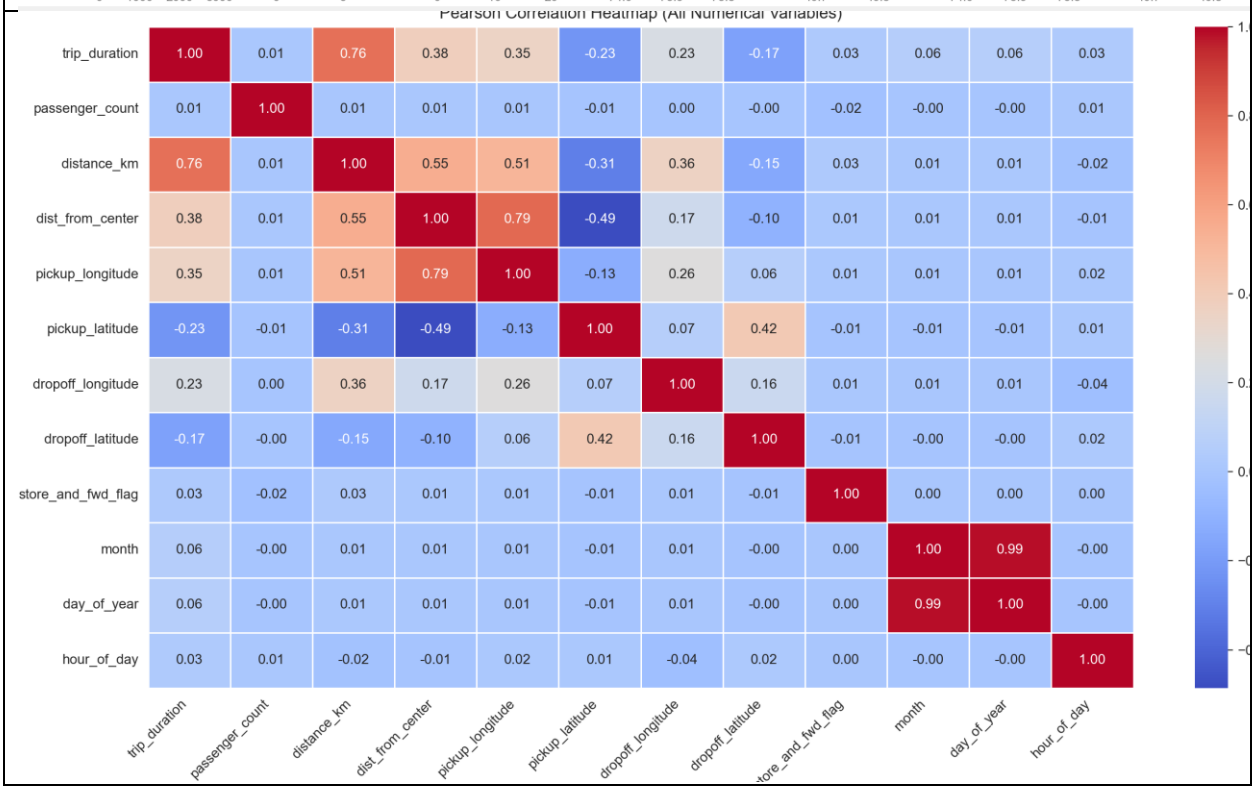

Distribution of Raw Trip Duration
Count
Trip Duration (seconds)
0
2000
4000
6000
8000
10000
0
1000
2000
3000
4000
5000


Q-Q Plot (Raw Trip Duration)
Ordered Values
Theoretical quantiles
5000
4000
3000
2000
1000
0
−1000
−4
−2
0
2
4

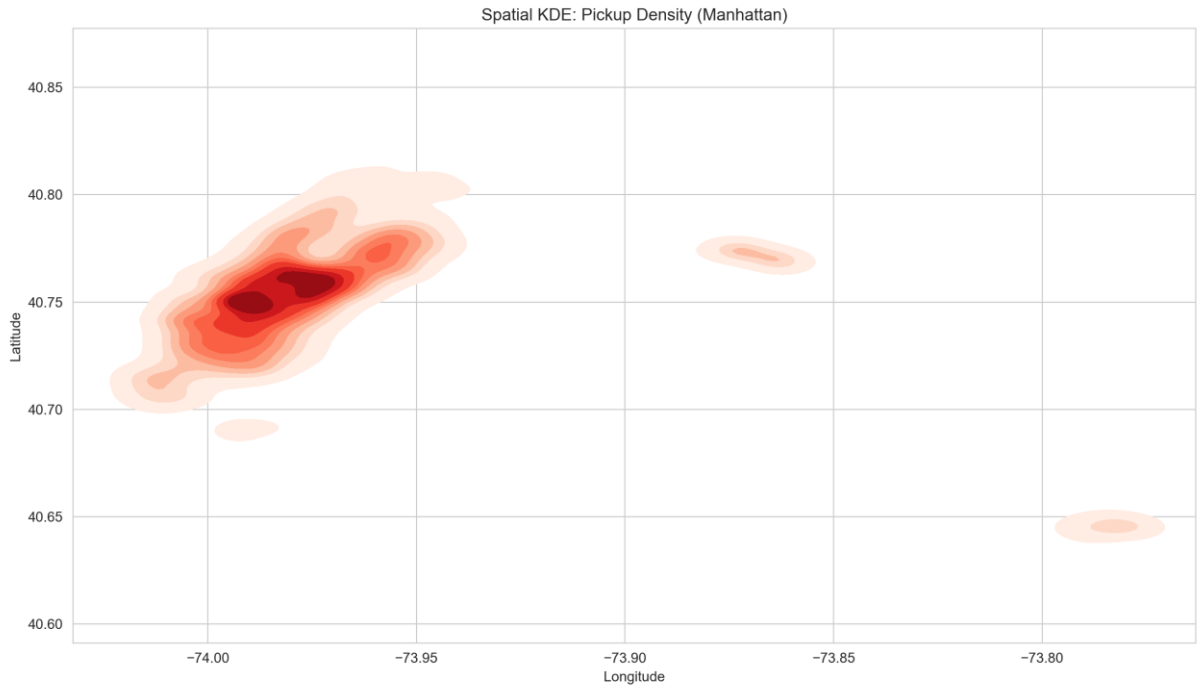

Distribution of Log Trip Duration
Count
Log(Trip Duration)
0
1000
2000
3000
4000
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5

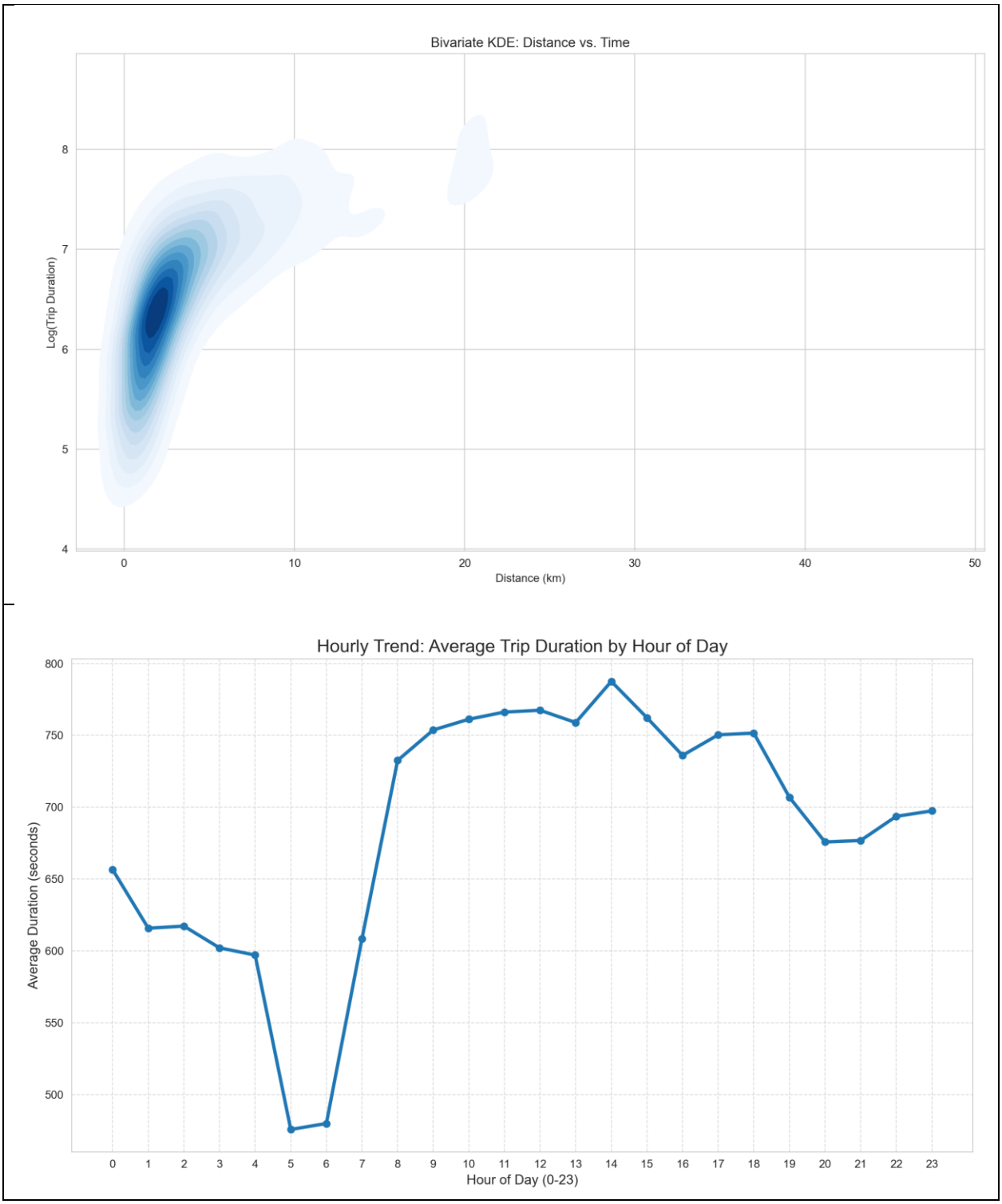

Q-Q Plot (Log Trip Duration)
Ordered Values
Theoretical quantiles
4
5
6
7
8
9
−4
−2
0
2
4

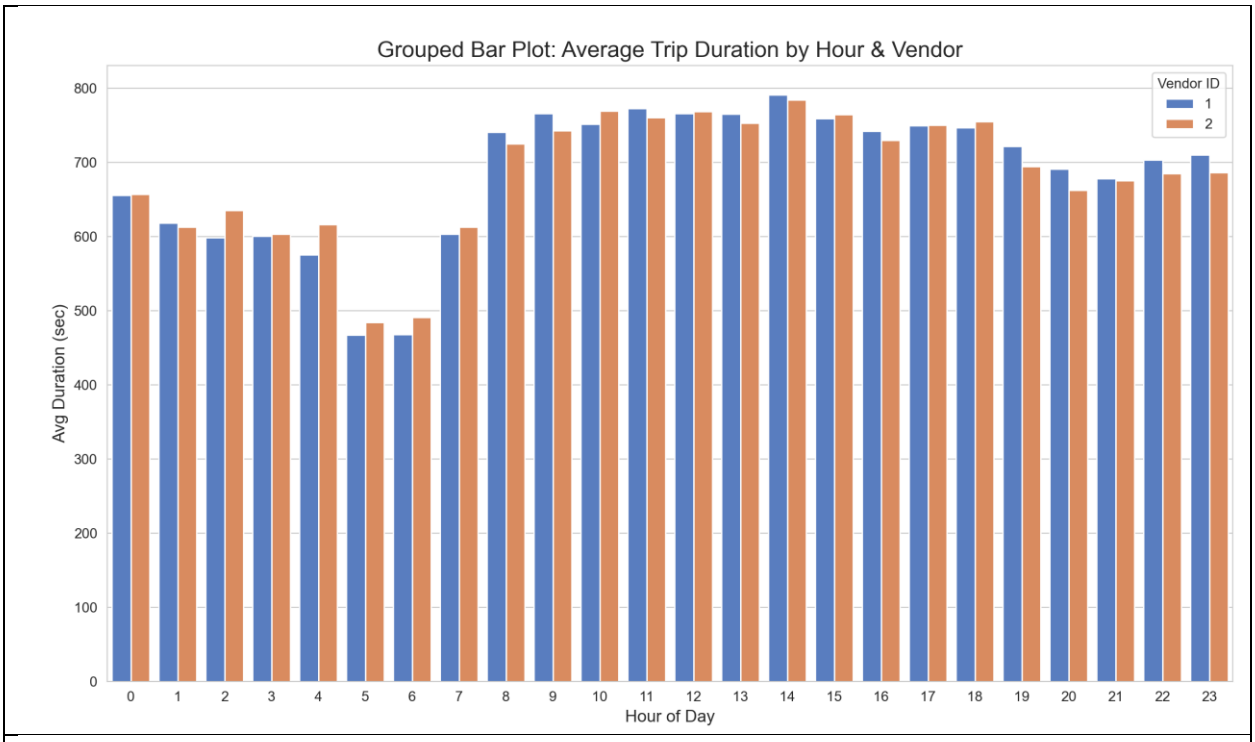

trip_duration
distance_km
dist_from_center
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude

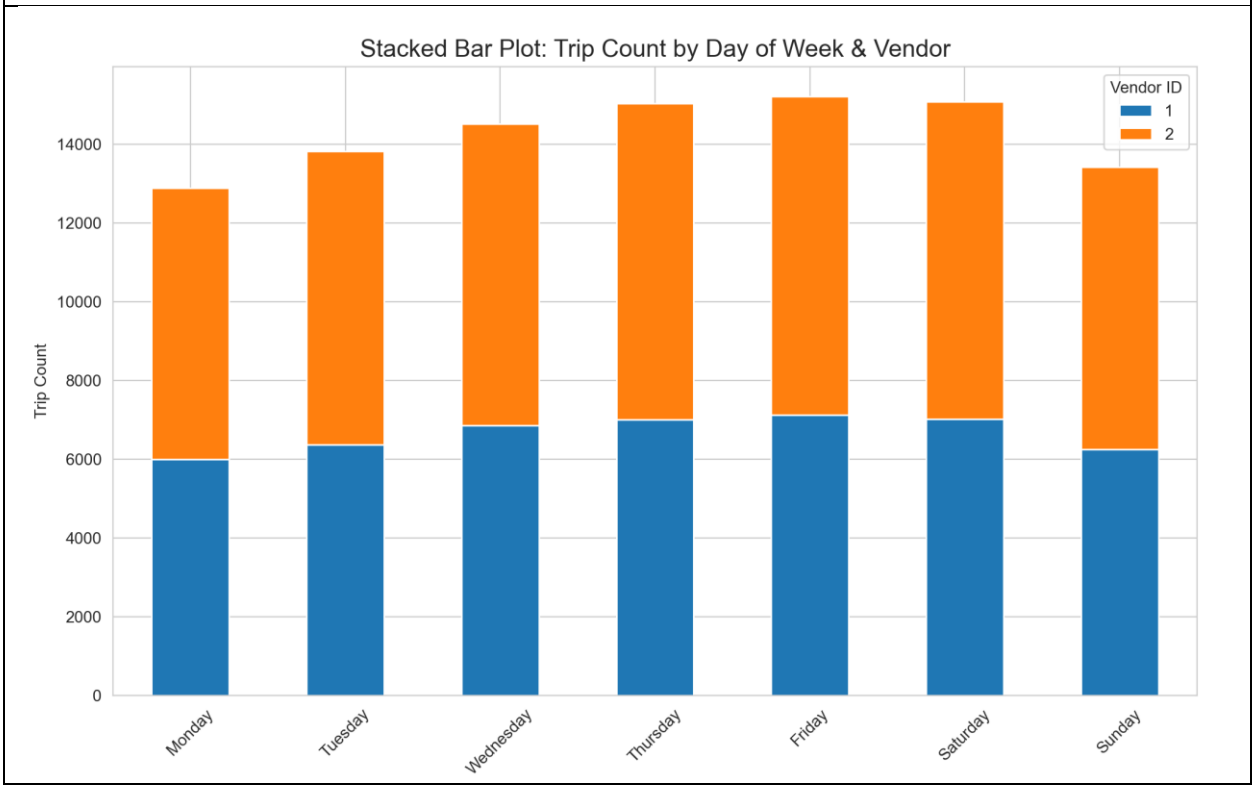

Pearson Correlation Heatmap (All Numerical Variables)
trip_duration 1.00 0.01 0.76 0.38 0.35 -0.23 0.23 -0.17 0.03 0.06 0.06 0.03
passenger_count 0.01 1.00 0.01 0.01 0.01 -0.01 0.00 -0.00 -0.02 -0.00 -0.00 0.01
distance_km 0.76 0.01 1.00 0.55 0.51 -0.31 0.36 -0.15 0.03 0.01 0.01 -0.02
dist_from_center 0.38 0.01 0.55 1.00 0.79 -0.49 0.17 -0.10 0.01 0.01 0.01 -0.01
pickup_longitude 0.35 0.01 0.51 0.79 1.00 -0.13 0.26 0.06 0.01 0.01 0.01 0.02
pickup_latitude -0.23 -0.01 -0.31 -0.49 -0.13 1.00 0.07 0.42 -0.01 -0.01 -0.01 0.01
dropoff_longitude 0.23 0.00 0.36 0.17 0.26 0.07 1.00 0.16 0.01 0.01 0.01 -0.04
dropoff_latitude -0.17 -0.00 -0.15 -0.10 0.06 0.42 0.16 1.00 -0.01 -0.00 -0.00 0.02
store_and_fwd_flag 0.03 -0.02 0.03 0.01 0.01 -0.01 0.01 -0.01 1.00 0.00 0.00 0.00
month 0.06 -0.00 0.01 0.01 0.01 -0.01 0.01 -0.00 0.00 1.00 0.99 -0.00
day_of_year 0.06 -0.00 0.01 0.01 0.01 -0.01 0.01 -0.00 0.00 0.99 1.00 -0.00
hour_of_day 0.03 0.01 -0.02 -0.01 0.02 0.01 -0.04 0.02 0.00 -0.00 -0.00 1.00
trip_duration
passenger_count
distance_km
dist_from_center
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude
store_and_fwd_flag
month
day_of_year
hour_of_day

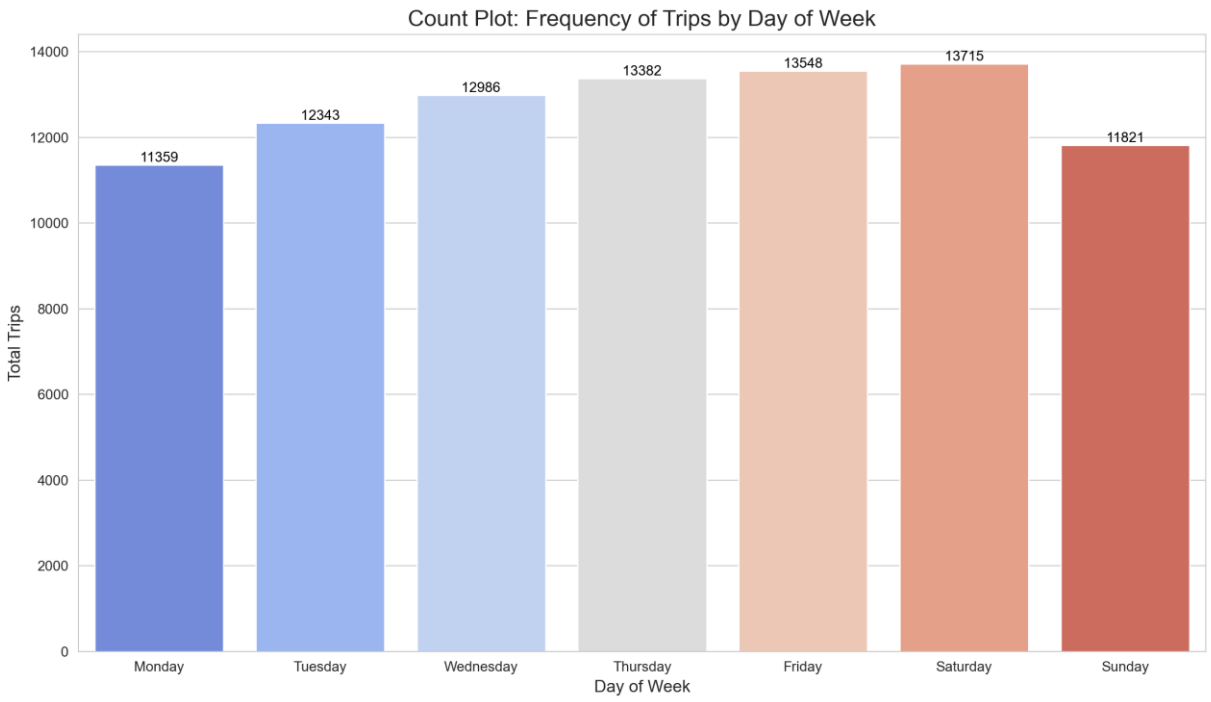
```
--- Statistical Tool 1: Confidence Interval ---
Mean Trip Duration: 711.27 seconds
95% Confidence Interval: (np.float64(708.305608438771), np.float64(714.2355899370731))
Observation: We are 95% confident the true population mean lies between 708.31 and 714.24 seconds.

--- Statistical Tool 2: Two-Sample T-Test ---
Comparison: Vendor 1 vs Vendor 2 Trip Durations
T-statistic: 1.4242, P-value: 0.1544
Result: Fail to Reject Null. No significant difference found.
```

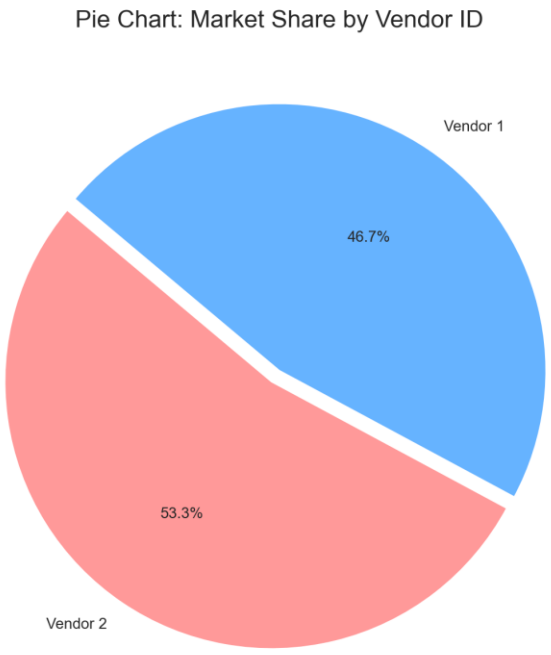


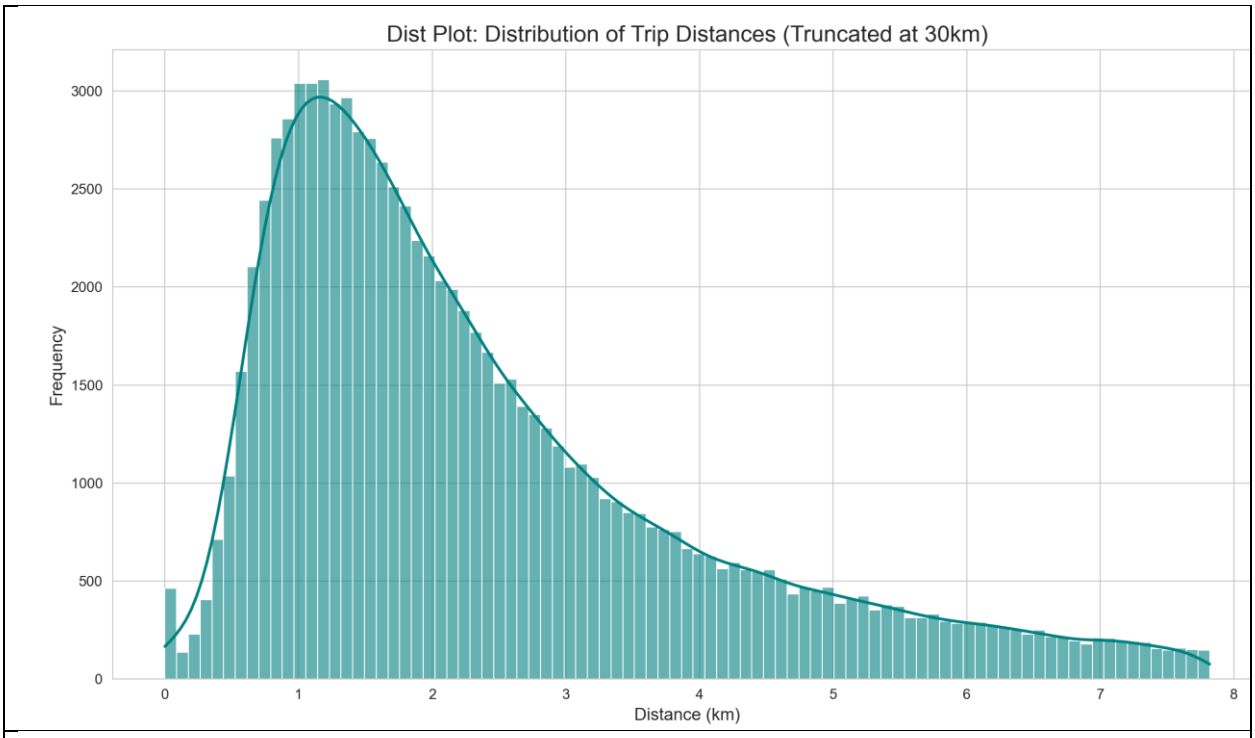

Bivariate KDE: Distance vs. Time
Log(Trip Duration)
8
7
6
5
4
0
10
20
30
40
50
Distance (km)
Hourly Trend: Average Trip Duration by Hour of Day
Average Duration (seconds)
800
750
700
650
600
550
500
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day (0-23)

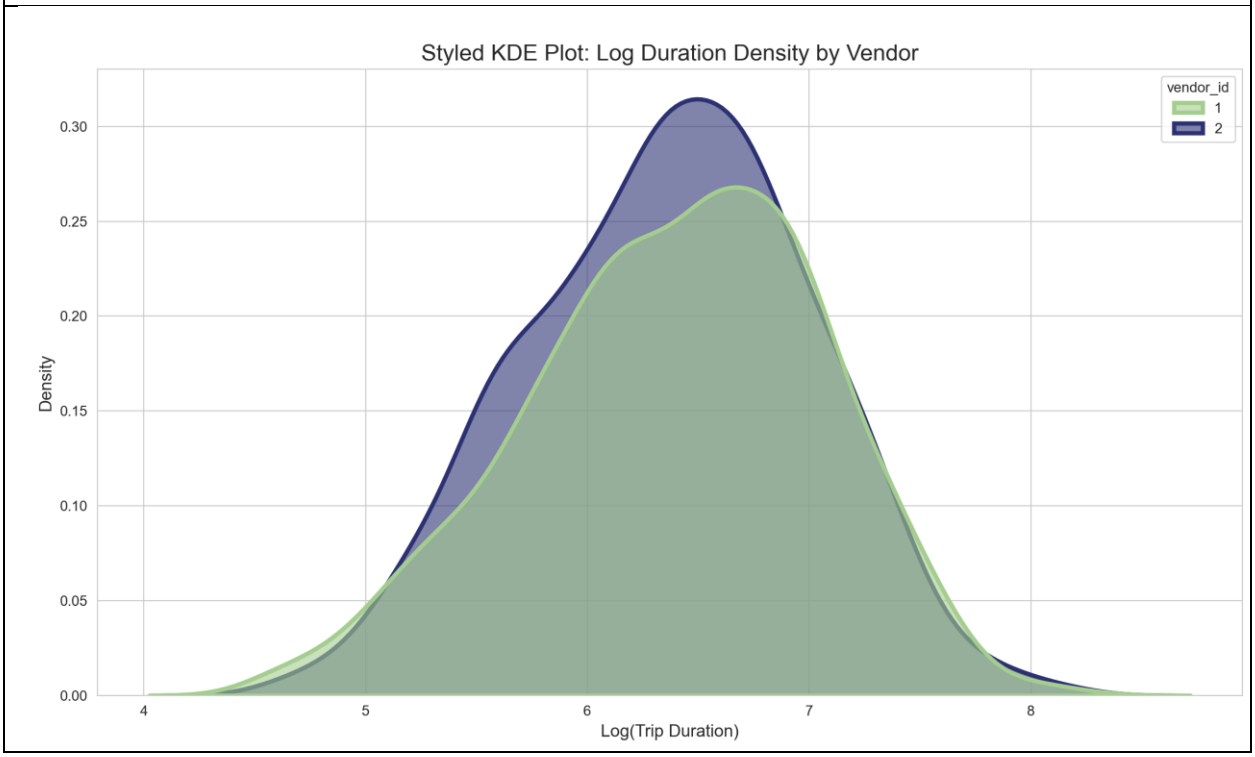

Grouped Bar Plot: Average Trip Duration by Hour & Vendor
Vendor ID
1
2
Avg Duration (sec)
0
100
200
300
400
500
600
700
800
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day

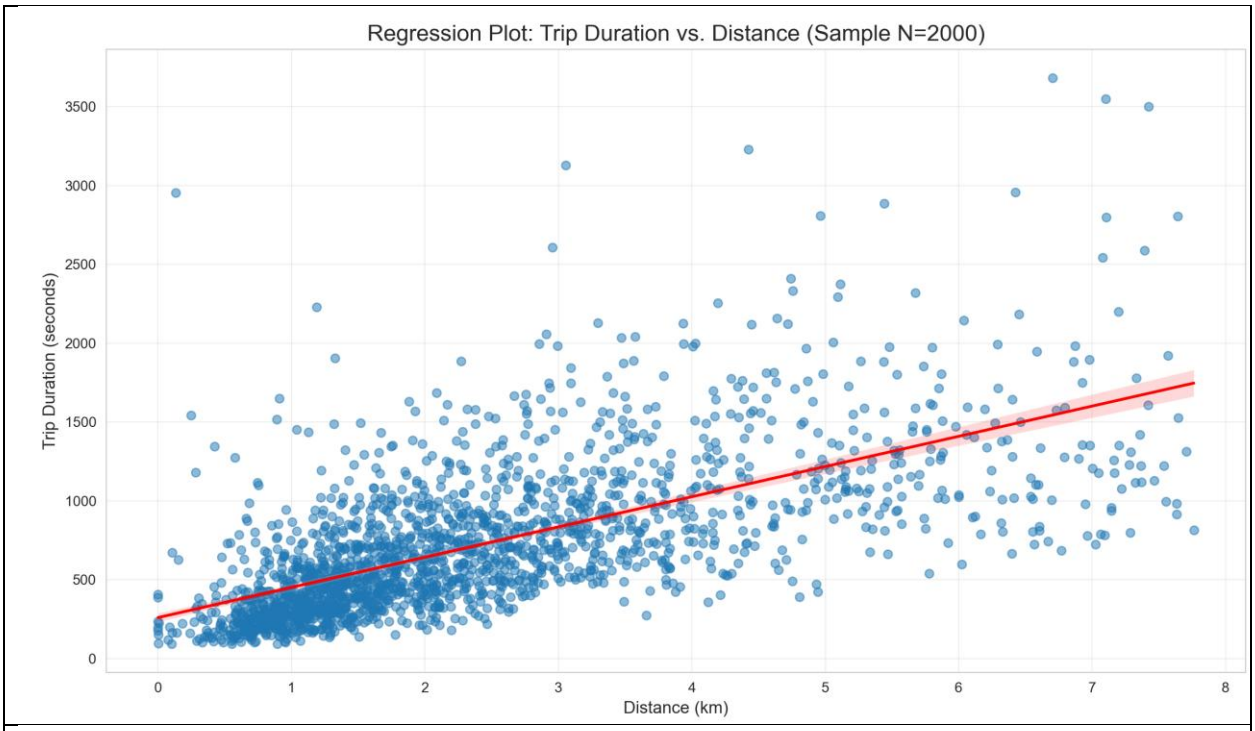

Stacked Bar Plot: Trip Count by Day of Week & Vendor
Vendor ID
1
2
Trip Count
0
2000
4000
6000
8000
10000
12000
14000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

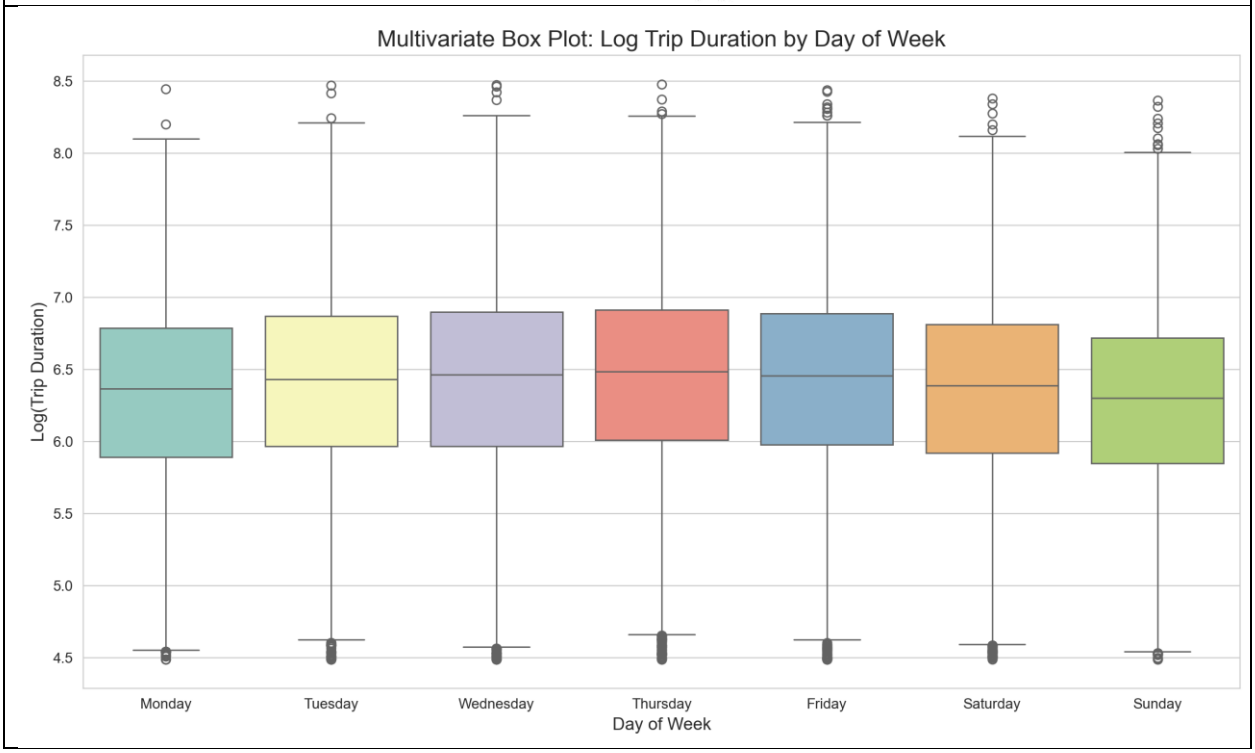

Count Plot: Frequency of Trips by Day of Week
Total Trips
Day of Week
0
2000
4000
6000
8000
10000
12000
14000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
11359
12343
12986
13382
13548
13715
11821

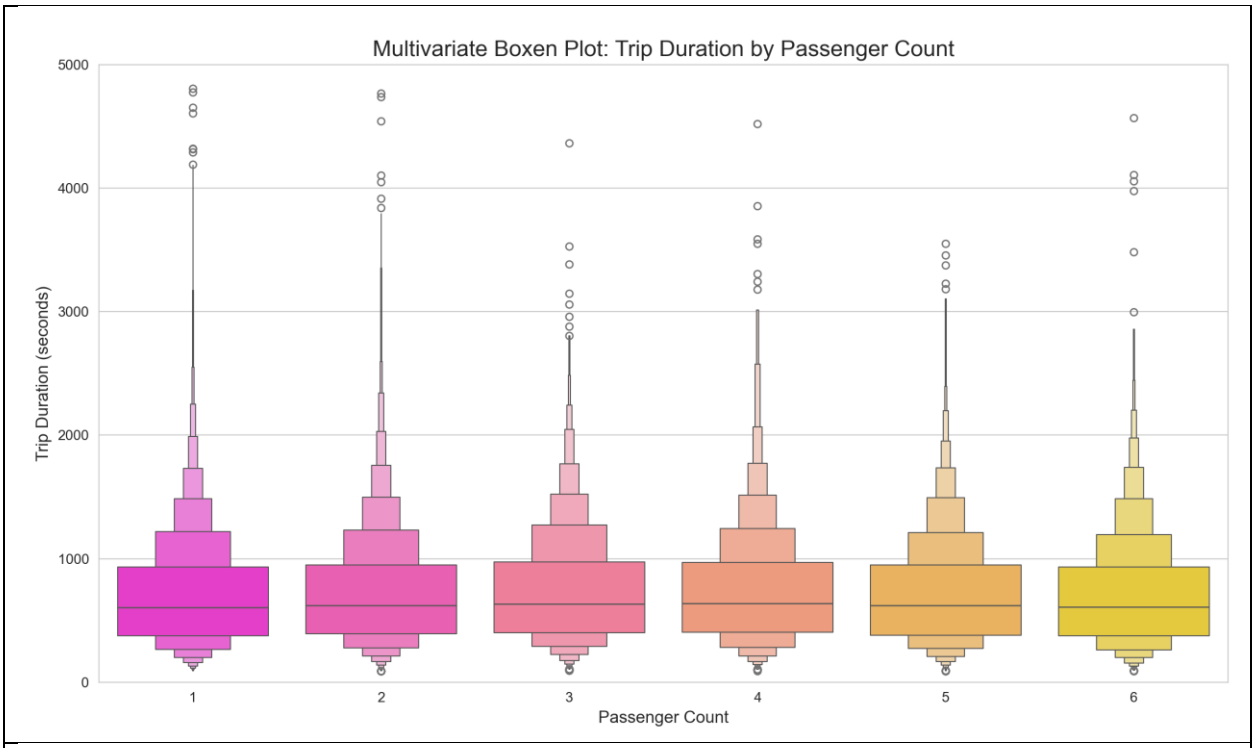

Pie Chart: Market Share by Vendor ID
Vendor 1
46.7%
53.3%
Vendor 2

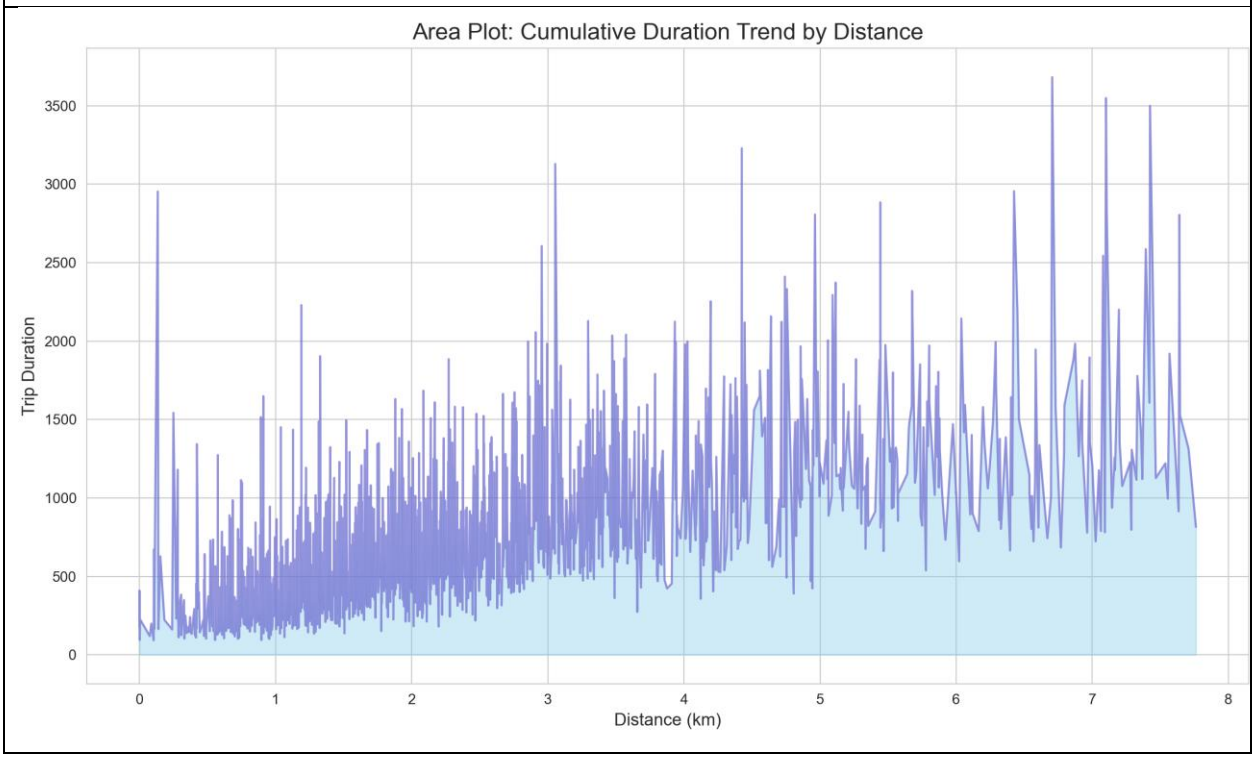

Dist Plot: Distribution of Trip Distances (Truncated at 30km)
Frequency
Distance (km)
0
500
1000
1500
2000
2500
3000
0
1
2
3
4
5
6
7
8

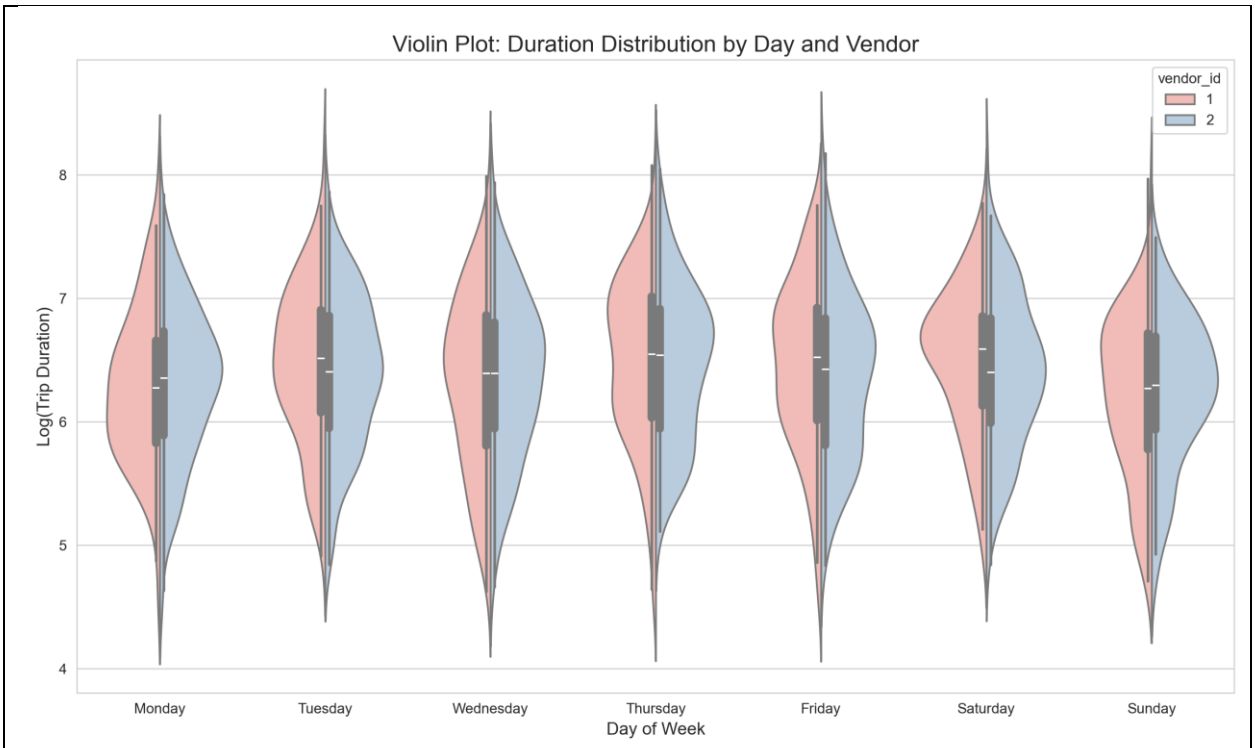

Styled KDE Plot: Log Duration Density by Vendor
vendor_id
1
2
Density
Log(Trip Duration)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
4
5
6
7
8

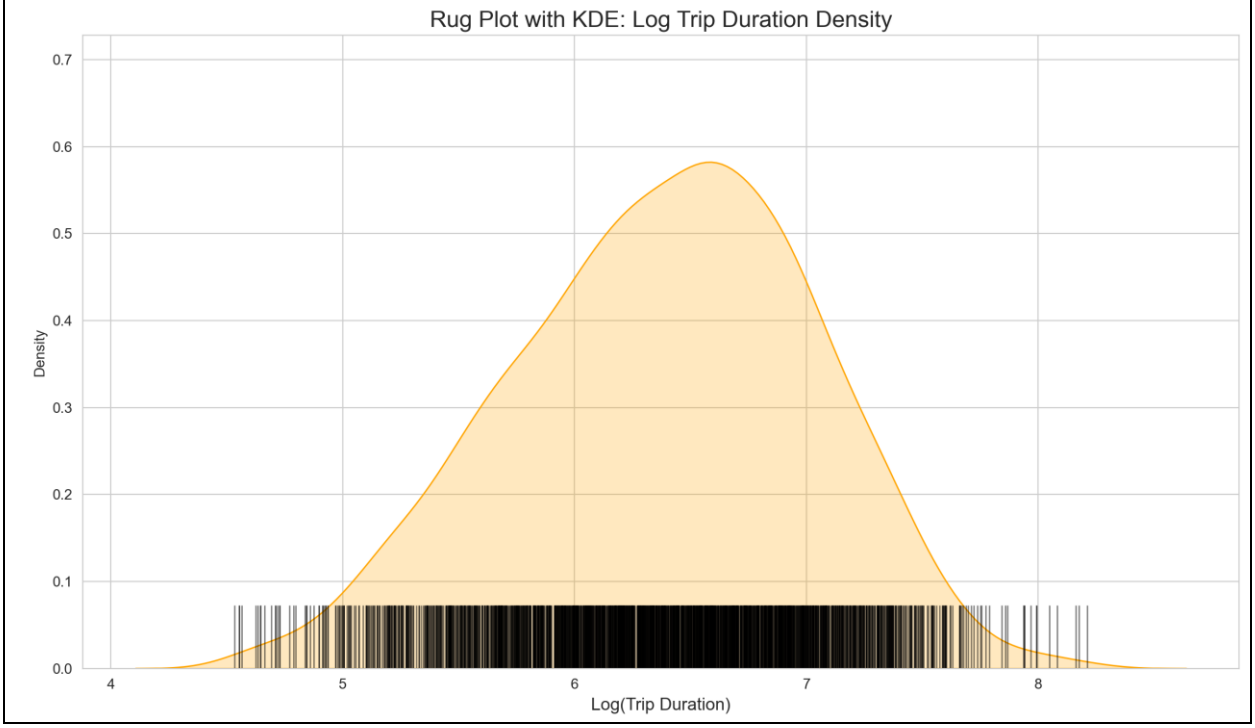

Regression Plot: Trip Duration vs. Distance (Sample N=2000)
Trip Duration (seconds)
Distance (km)
0
500
1000
1500
2000
2500
3000
3500
0
1
2
3
4
5
6
7
8

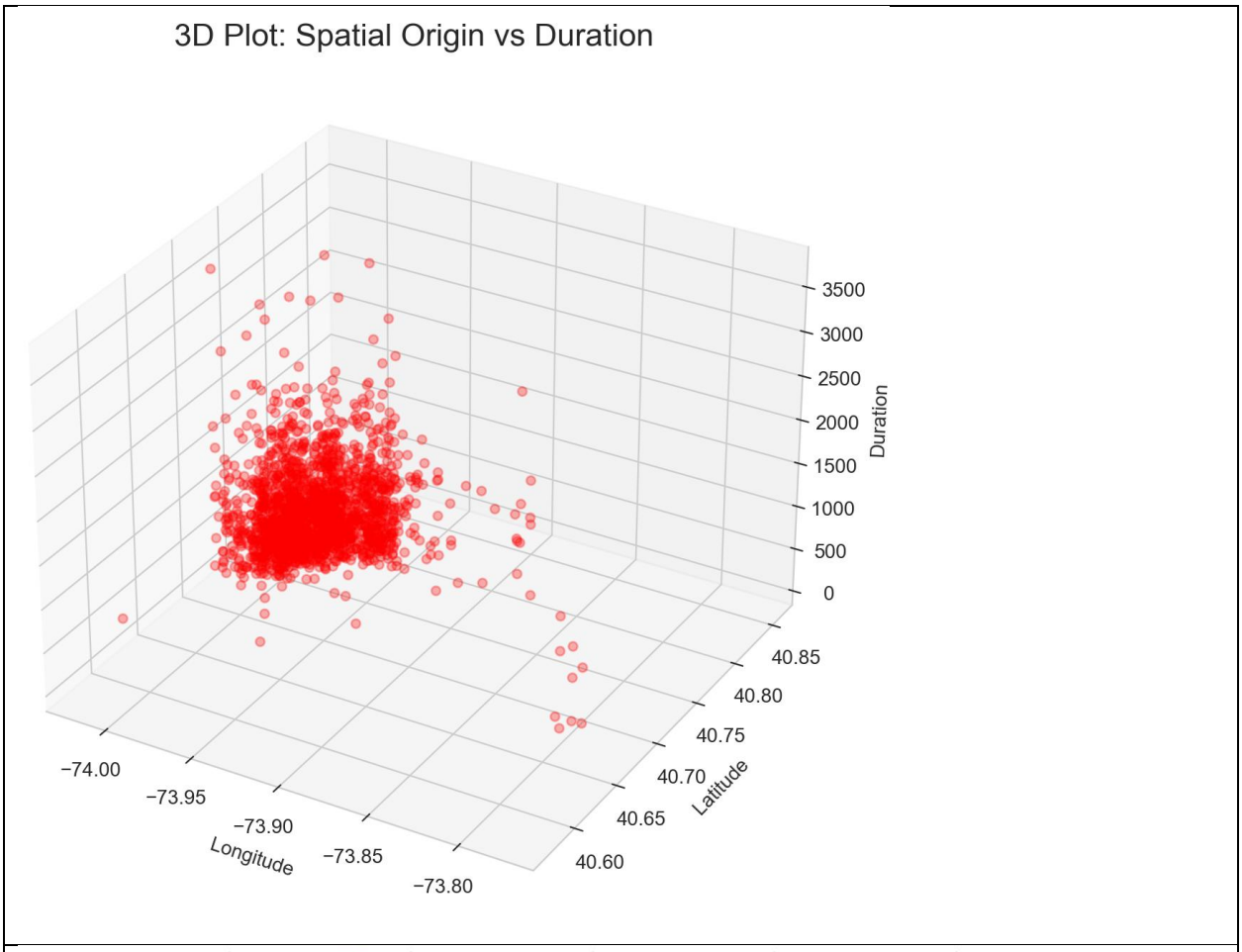

Multivariate Box Plot: Log Trip Duration by Day of Week
Log(Trip Duration)
Day of Week
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

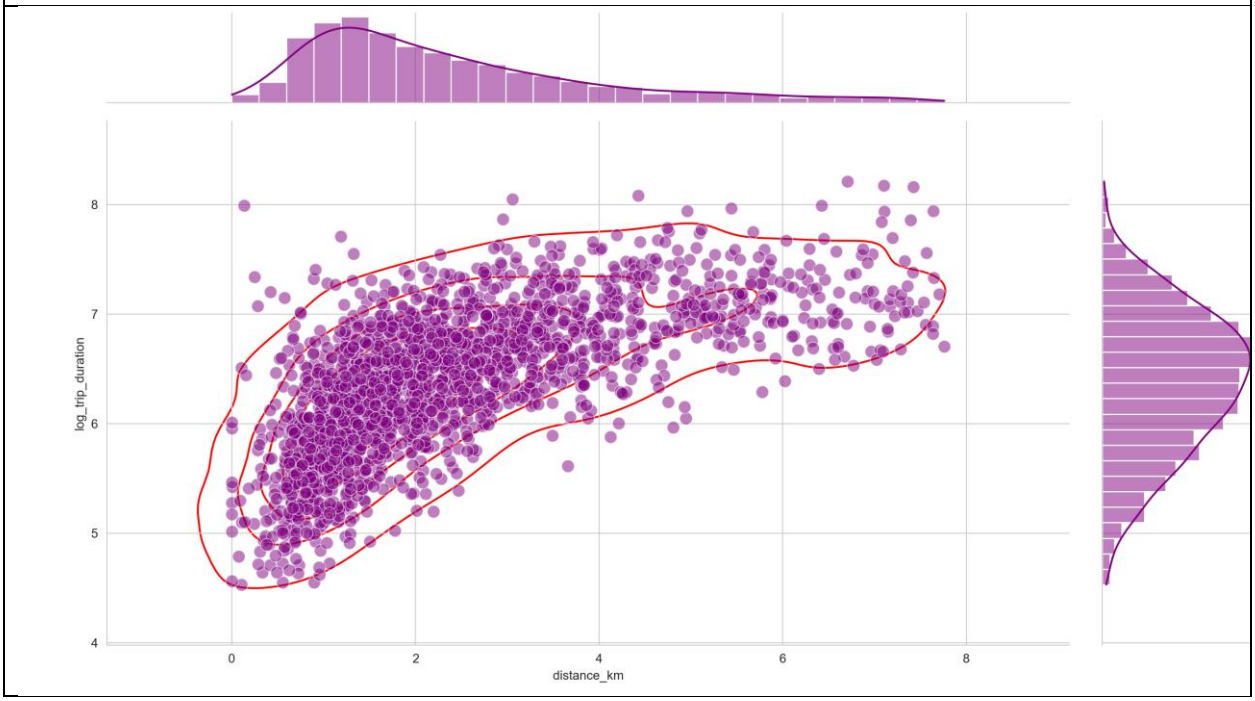

Multivariate Boxen Plot: Trip Duration by Passenger Count
Trip Duration (seconds)
0
1000
2000
3000
4000
5000
1
2
3
4
5
6
Passenger Count

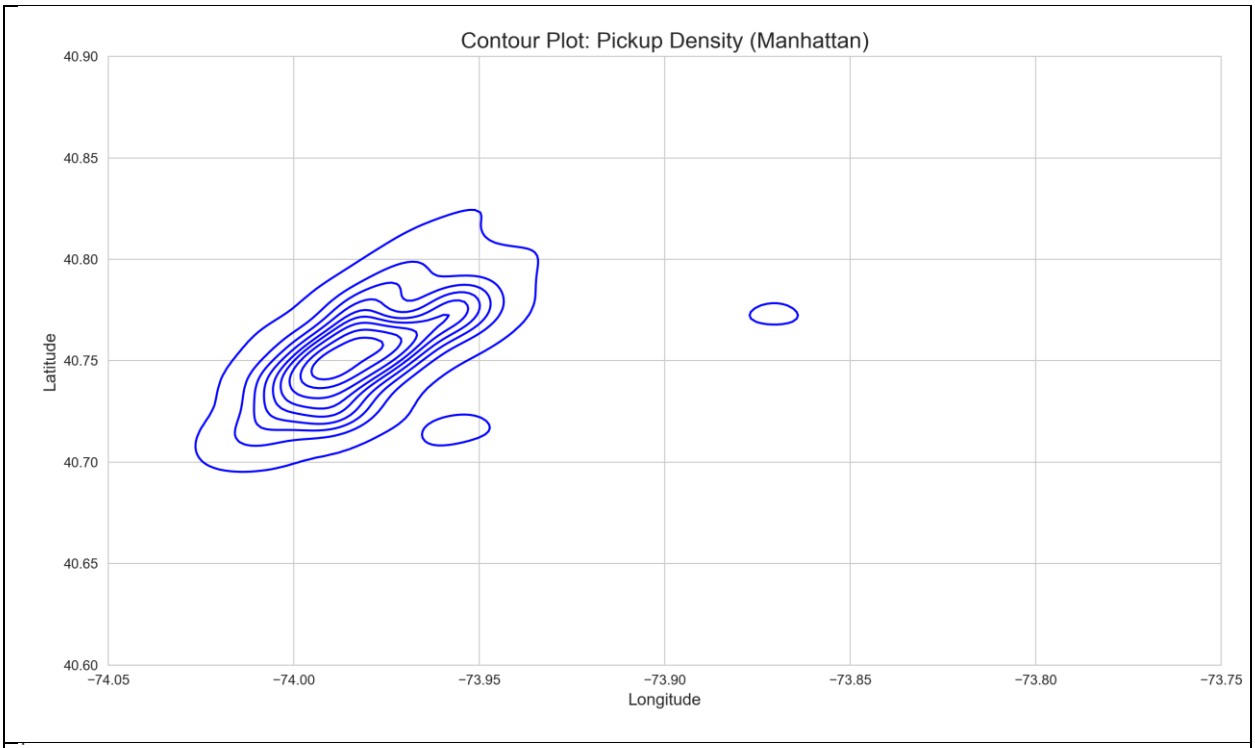

Area Plot: Cumulative Duration Trend by Distance
Trip Duration
0
500
1000
1500
2000
2500
3000
3500
0
1
2
3
4
5
6
7
8
Distance (km)

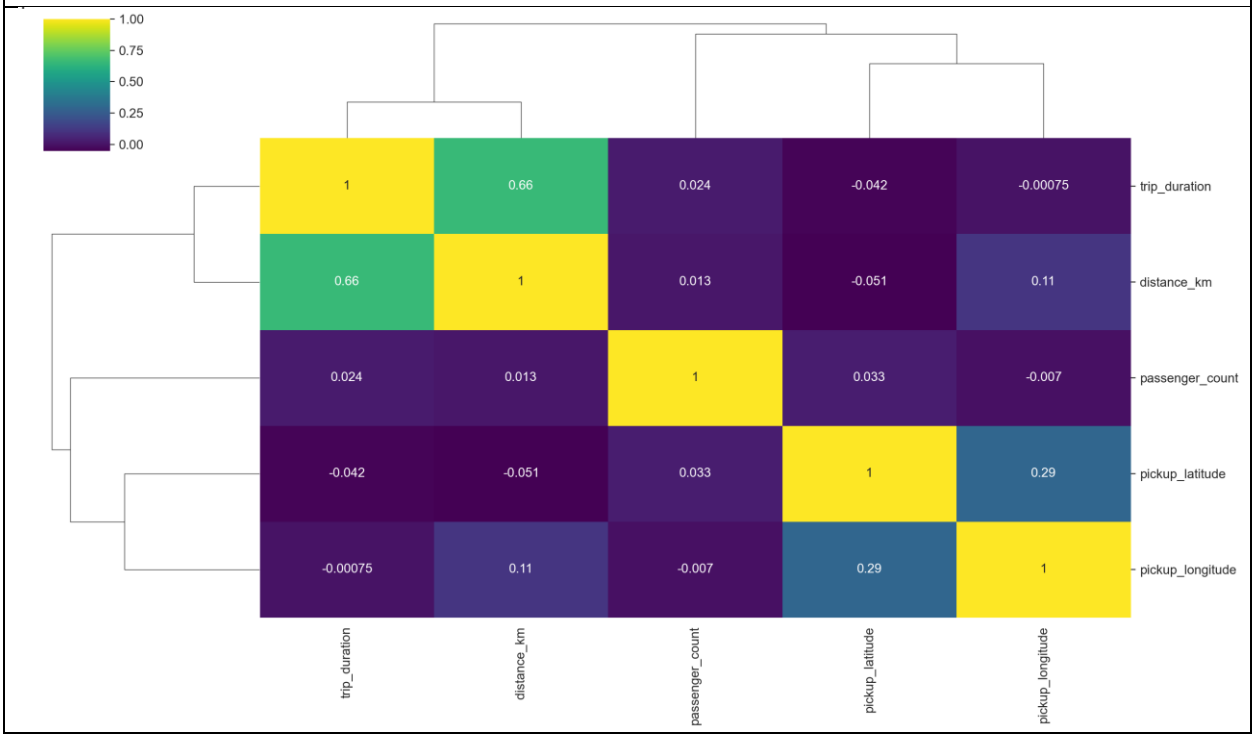

Violin Plot: Duration Distribution by Day and Vendor
vendor_id
1
2
Log(Trip Duration)
8
7
6
5
4
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week

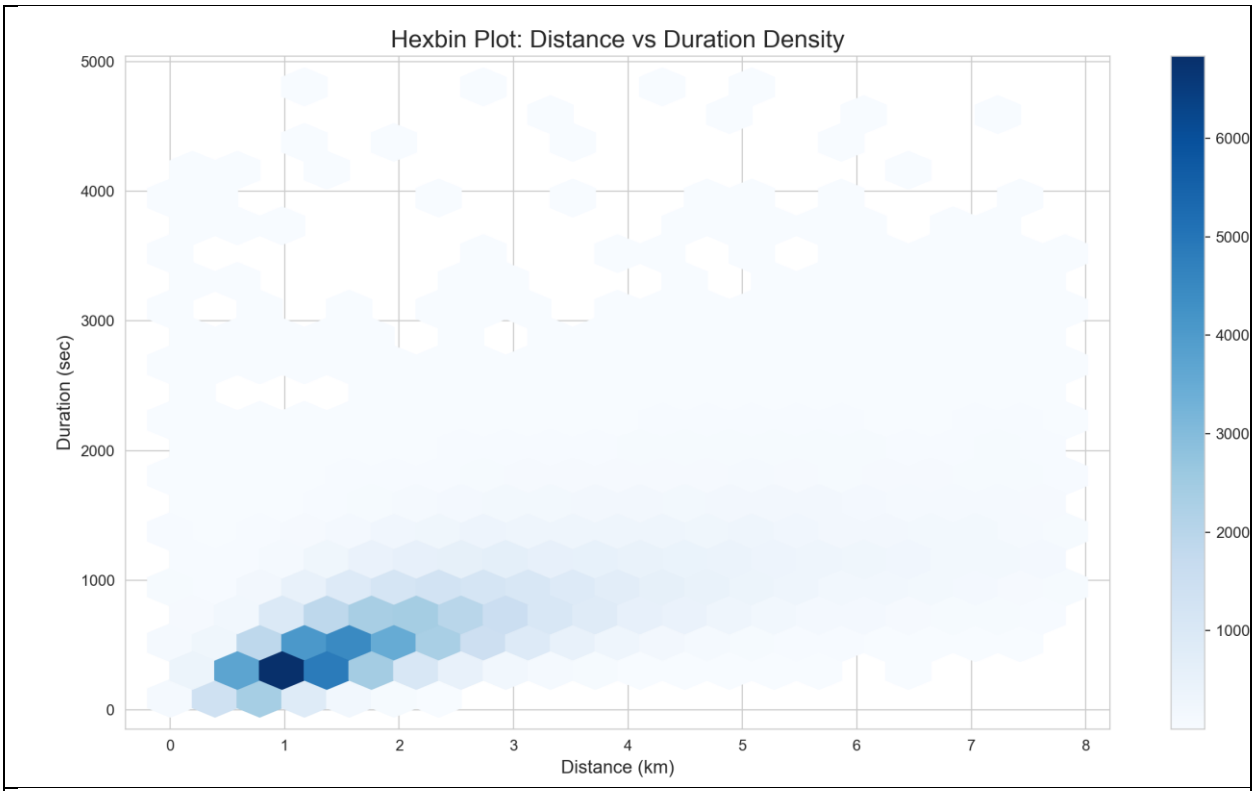

Rug Plot with KDE: Log Trip Duration Density
Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
4
5
6
7
8
Log(Trip Duration)

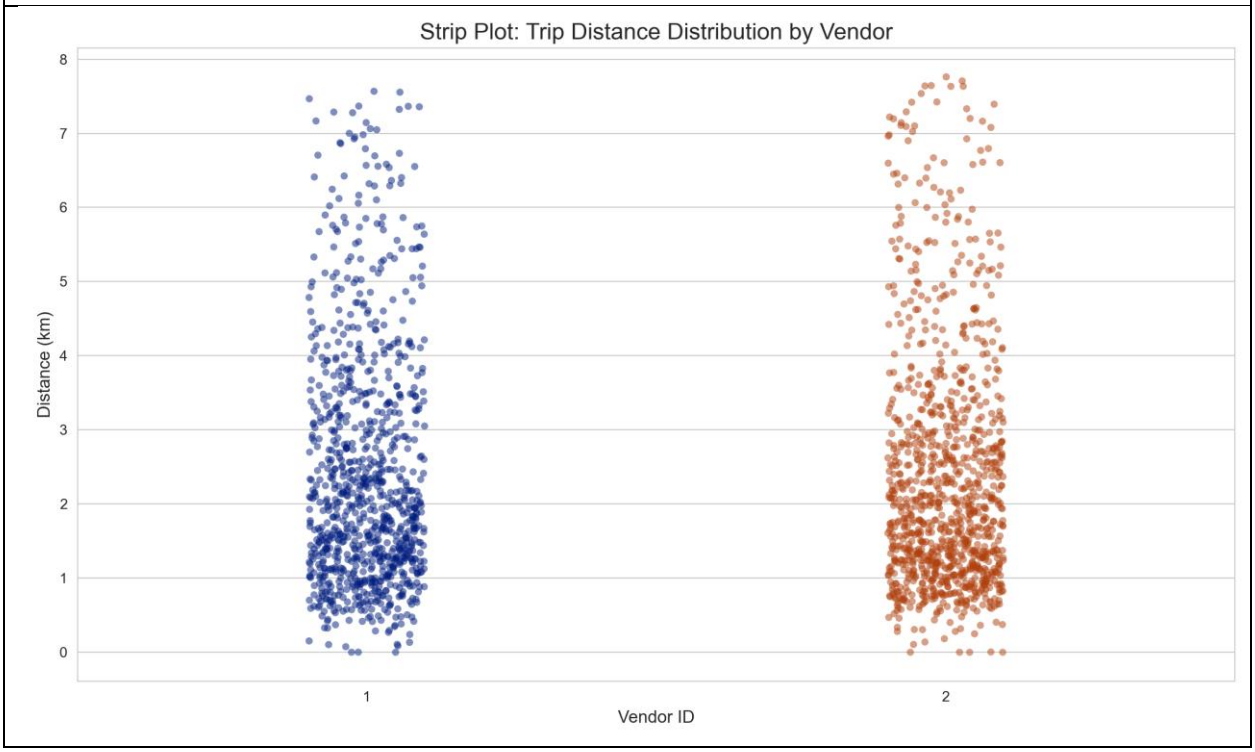

3D Plot: Spatial Origin vs Duration
3500
3000
2500
2000
1500
1000
500
0
Duration
40.85
40.80
40.75
40.70
40.65
40.60
Latitude
−74.00
−73.95
−73.90
−73.85
−73.80
Longitude

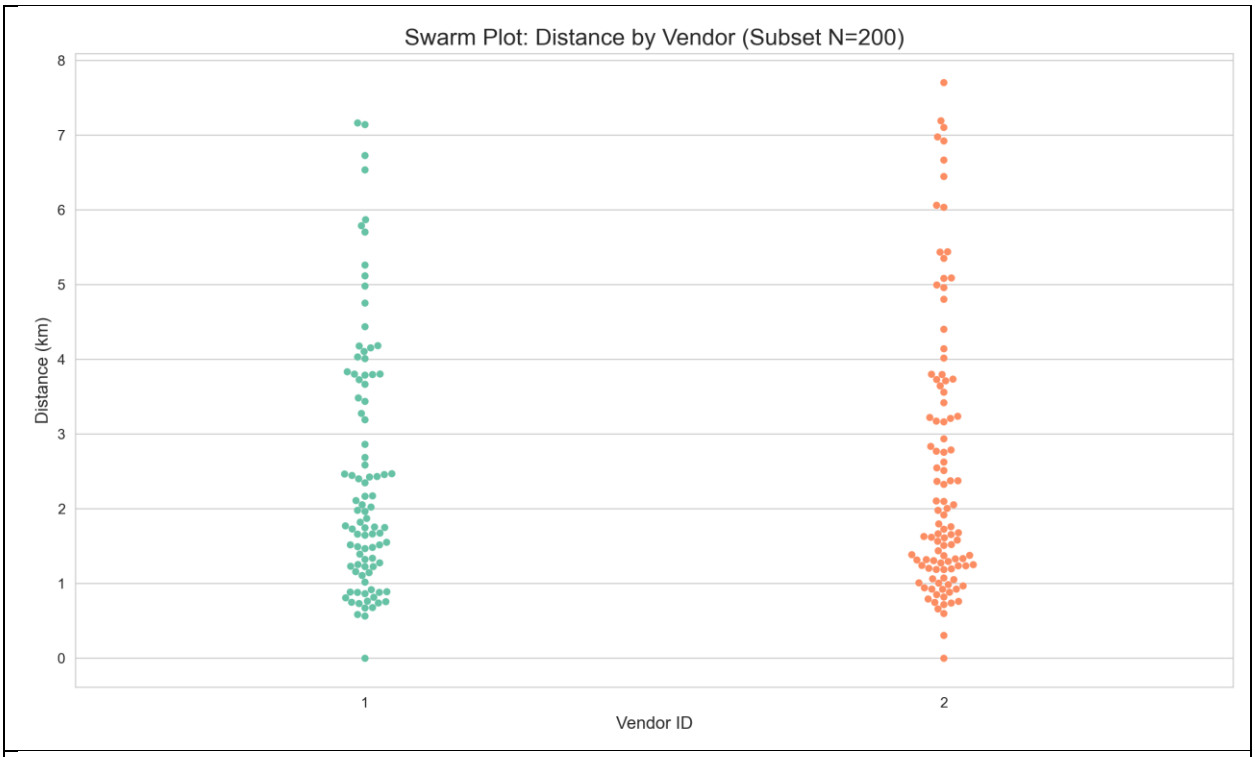

log_trip_duration
8
7
6
5
4
0
2
4
6
8
distance_km

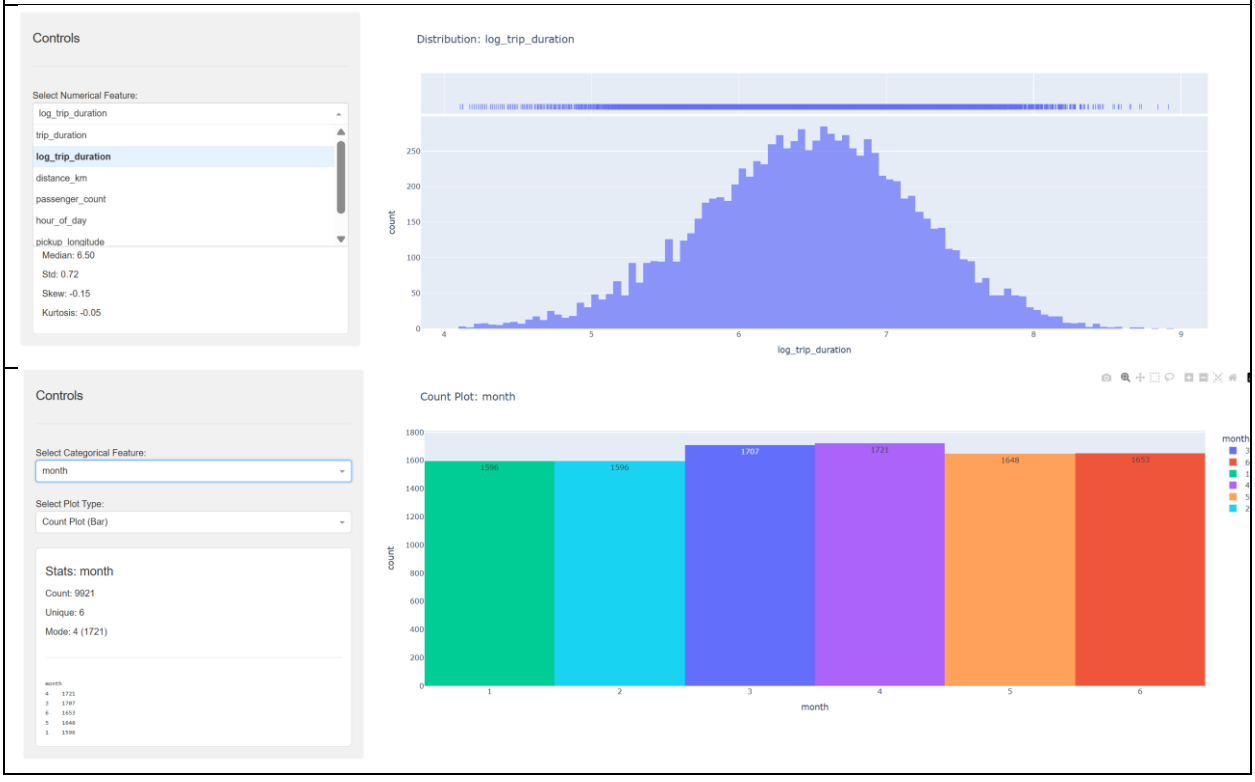

Contour Plot: Pickup Density (Manhattan)
Latitude
Longitude
40.90
40.85
40.80
40.75
40.70
40.65
40.60
−74.05
−74.00
−73.95
−73.90
−73.85
−73.80
−73.75


1.00
0.75
0.50
0.25
0.00
1
0.66
0.024
-0.042
-0.00075
0.66
1
0.013
-0.051
0.11
0.024
0.013
1
0.033
-0.007
-0.042
-0.051
0.033
1
0.29
-0.00075
0.11
-0.007
0.29
1
trip_duration
distance_km
passenger_count
pickup_latitude
pickup_longitude


Hexbin Plot: Distance vs Duration Density
Duration (sec)
Distance (km)
0
1000
2000
3000
4000
5000
0
1
2
3
4
5
6
7
8
1000
2000
3000
4000
5000
6000


Strip Plot: Trip Distance Distribution by Vendor
Distance (km)
Vendor ID
0
1
2
3
4
5
6
7
8
1
2


Swarm Plot: Distance by Vendor (Subset N=200)
Distance (km)
Vendor ID
0
1
2
3
4
5
6
7
8


Controls
Select Numerical Feature:
log_trip_duration
trip_duration
log_trip_duration
distance_km
passenger_count
hour_of_day
pickup_longitude
Median: 6.50
Std: 0.72
Skew: -0.15
Kurtosis: -0.05
Distribution: log_trip_duration
count
log_trip_duration
Controls
Select Categorical Feature:
month
Select Plot Type:
Count Plot (Bar)
Stats: month
Count: 9921
Unique: 6
Mode: 4 (1721)
Count Plot: month
count
month
1598
1596
1707
1721
1648
1653

Controls

Select X-Axis (Num):

trip_duration

Select Y-Axis (Num):

distance_km

Select Color (Optional):

day_of_week

Select Plot Type:

Reg Plot (Scatter+Line)

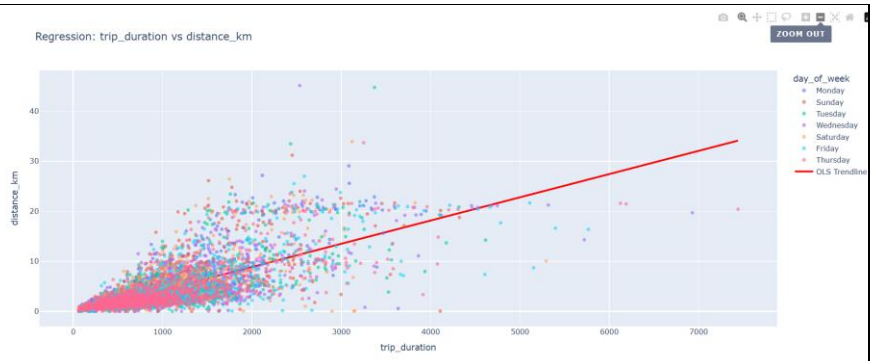
Correlation Analysis

Pearson Corr: 0.7568

R-Squared: 0.5727

Slope: 0.0046

P-Value: 0.0000e+00




Controls

Select X-Axis (Categorical):

day_of_week

Select Y-Axis (Numerical):

trip_duration

Select Color (Optional):

store_and_fwd_flag

Select Plot Type:

Bar Plot (Mean)

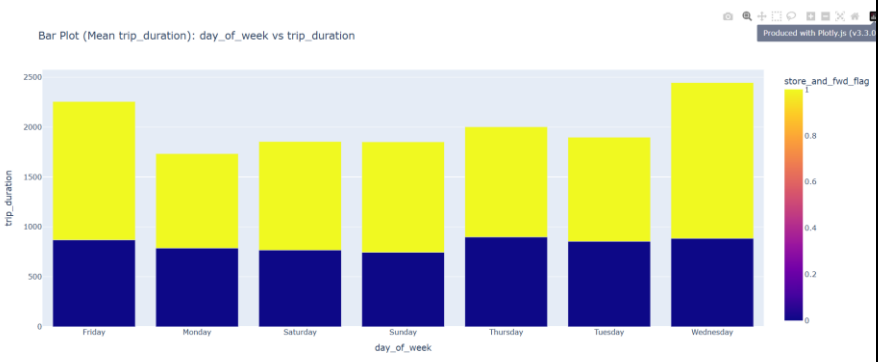
Bar Mode (for Bar Plot):

Group Stack

Stats: trip_duration by day_of_week

| day_of_week | mean | max | min | count |
|---|---|---|---|---|
| Thursday | 899.95 | 7440 | 61 | 1462 |
| Wednesday | 888.73 | 6928 | 60 | 1419 |
| Friday | 871.37 | 5764 | 66 | 1509 |
| Tuesday | 855.52 | 4615 | 60 | 1412 |
| Monday | 788.62 | 4120 | 66 | 1274 |

« ‹ 1 / 2 › »




Controls

Select Plot Type:

Pair Plot (Scatter Matrix)

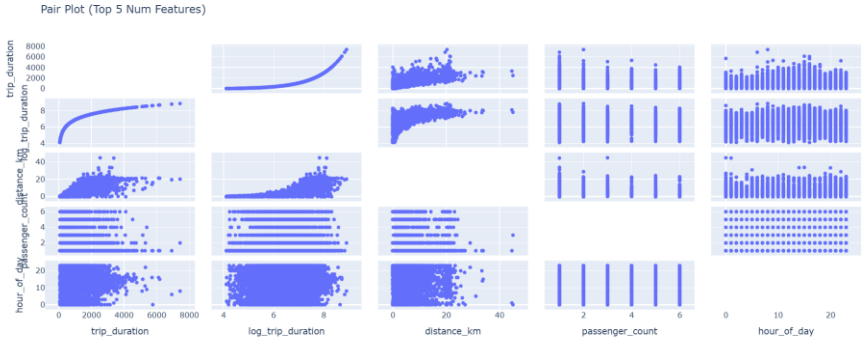
Global Statistics

Strongest Correlation found between:

**trip_duration & log_trip_duration**

Correlation Coefficient: 0.8881

Note: 1.0 is perfect correlation.



# NYC Taxi Dashboard

Interactive Data Pipeline

1. Data Cleaning | 2. Transform & Norm | 3. PCA Analysis | 4. Interactive Visualizations

## Data Cleaning & Outliers

Step 1: Cleaning Operations

☐ Drop Duplicates ☐ Drop Null Values ☑ Filter Negative Durations

Step 2: Outlier Removal (Z-Score)

Keep Outliers (None)

APPLY & DOWNLOAD CSV

**Original: 9921 | Cleaned: 9921 | Dropped: 0**

### Dataset Preview

| id | vendor_id | pickup_datetime | dropoff_datetime | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | store_and_fwd_flag | trip_duration | month | day_of_week | hour_of_day | day_of_year | distance_km | log_trip_duration |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| id2875421 | 2 | 2016-03-14T17:24:55 | 2016-03-14T17:32:30 | 1 | -73.98215484619139 | 40.767936706542977 | -73.96463012695312 | 40.765602111816406 | 0 | 455 | 3 | Monday | 17 | 74 | 1.4985207796458056 | 6.122492880951438 |
| id2377394 | 1 | 2016-06-12T00:43:35 | 2016-06-12T00:54:38 | 1 | -73.98041534423827 | 40.738563537597656 | -73.99948120117188 | 40.731151580810554 | 0 | 663 | 6 | Sunday | 0 | 164 | 1.8055071687065205 | 6.498282149476434 |
| id3858529 | 2 | 2016-01-19T11:35:24 | 2016-01-19T12:10:48 | 1 | -73.97902679443358 | 40.763938903808594 | -74.00533294677734 | 40.710086822509766 | 0 | 2124 | 1 | Tuesday | 11 | 19 | 6.3850984952528868 | 7.661527081358517 |
| id3504673 | 2 | 2016-04-06T19:32:31 | 2016-04-06T19:39:40 | 1 | -74.01004028320312 | 40.719970703125 | -74.01226806640625 | 40.70671844482422 | 0 | 429 | 4 | Wednesday | 19 | 97 | 1.4854984227709382 | 6.063785208687608 |
| id2181028 | 2 | 2016-03-26T13:30:55 | 2016-03-26T13:38:10 | 1 | -73.97305297851561 | 40.79320907592773 | -73.9729232788086 | 40.782520294189446 | 0 | 435 | 3 | Saturday | 13 | 86 | 1.1885884593338754 | 6.077642243349034 |
| id0001584 | 2 | 2016-01-30T22:01:40 | 2016-01-30T22:09:03 | 6 | -73.98285675048828 | 40.74219512939453 | -73.99200868847656 | 40.74918365478516 | 0 | 443 | 1 | Saturday | 22 | 30 | 1.098942459306554 | 6.095824562432225 |
| id1813257 | 1 | 2016-06-17T22:34:59 | 2016-06-17T22:40:40 | 4 | -73.96981708288086 | 40.75783920288086 | -73.95740509033203 | 40.76589584350586 | 0 | 341 | 6 | Friday | 22 | 169 | 1.3262785770590748 | 5.834810737062605 |
| id1324603 | 2 | 2016-05-21T07:54:58 | 2016-05-21T08:20:49 | 1 | -73.96927642822266 | 40.79777908325195 | -73.92247009277344 | 40.760559082031125 | 0 | 1551 | 5 | Saturday | 7 | 142 | 5.7149806308789906 | 7.347299700743163 5 |
| id1301050 | 1 | 2016-05-27T23:12:23 | 2016-05-27T23:16:38 | 1 | -73.99948120117188 | 40.738399505615234 | -73.98578643798828 | 40.73281478881836 | 0 | 255 | 5 | Friday | 23 | 148 | 1.3103532828841318 | 5.5451774444795562 |
| id0012891 | 2 | 2016-03-10T21:45:01 | 2016-03-10T22:05:26 | 1 | -73.98104858398438 | 40.744338989257805 | -73.9729995727539 | 40.78998947143555 | 0 | 1225 | 3 | Thursday | 21 | 70 | 5.1211615621414746 | 7.111512116496156 |

User Observation Notes:

Type your findings here...


# NYC Taxi Dashboard

Interactive Data Pipeline

1. Data Cleaning | 2. Transform & Norm | 3. PCA Analysis | 4. Interactive Visualizations

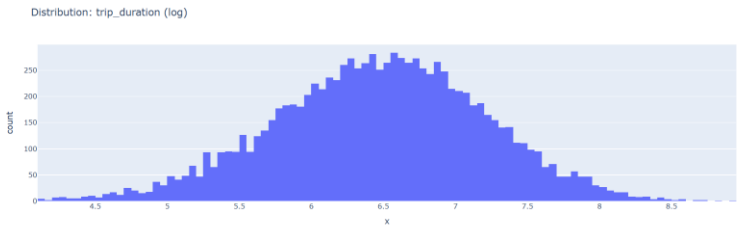

## Feature Transformation

Feature:

trip_duration

Transform:

- ○ None
- ◉ Log
- ○ Sqrt
- ○ Box-Cox

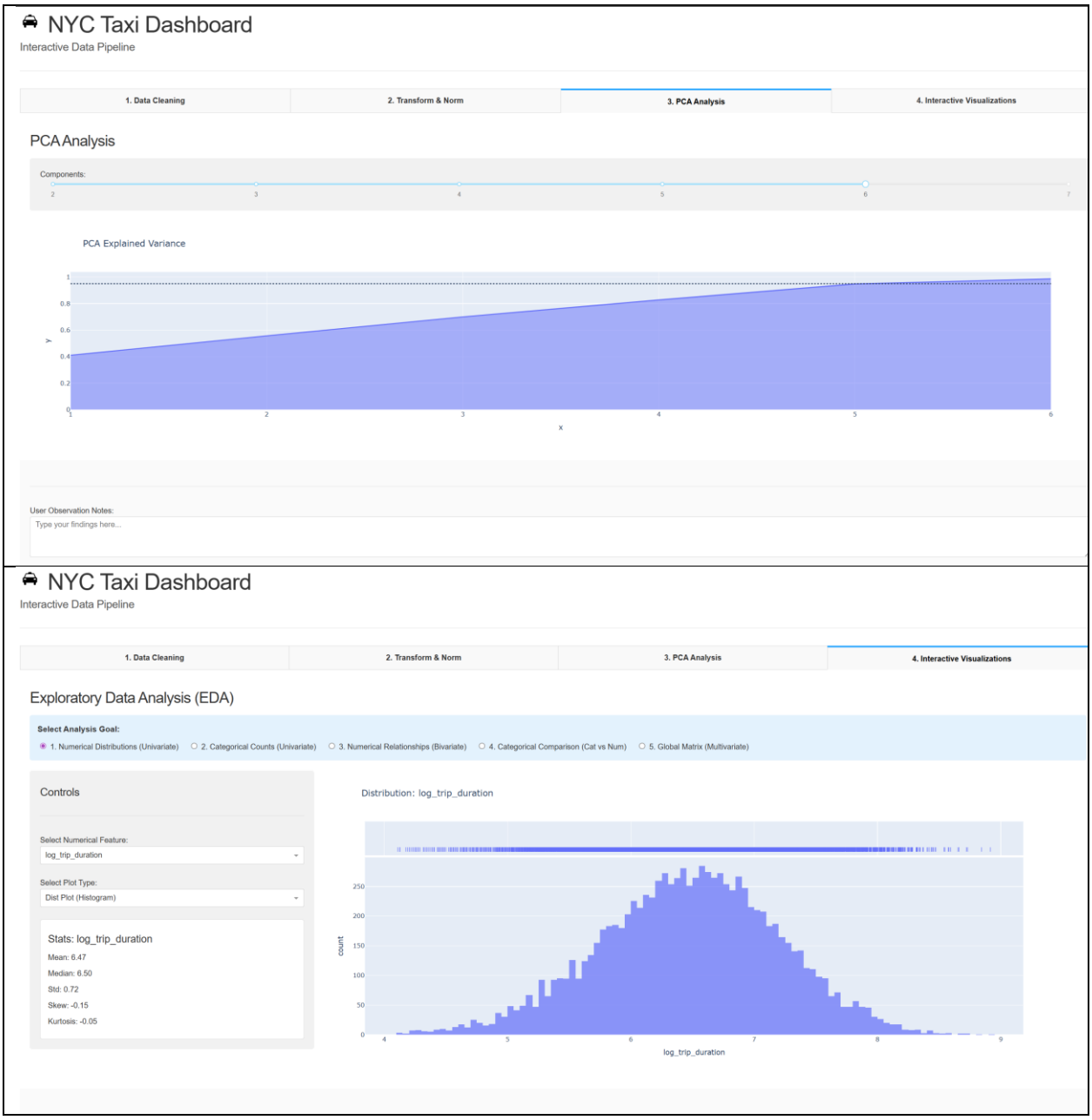



**Shapiro-Wilk:**

P=0.2268 (Normal)


User Observation Notes:

Type your findings here...




# 3. Abstract

This project presents a comprehensive visual analysis of the New York City Taxi Trip Duration dataset, utilizing a subset of over 1.4 million trip records from 2016 to investigate the factors driving urban transit times. The primary objective was to apply advanced Python visualization techniques to transform raw mobility data into actionable insights. The methodology involved rigorous data preprocessing, including the engineering of geospatial features using the Haversine distance formula, temporal decomposition of timestamps, and the application of logarithmic transformations to normalize the heavily right-skewed distribution of

trip durations. Exploratory Data Analysis (EDA) was conducted using a suite of static statistical visualizations—ranging from correlation heatmaps to multivariate box plots—to identify outliers and establish baseline relationships between trip distance, passenger count, and temporal traffic patterns.

To facilitate dynamic exploration of these findings, an interactive web-based dashboard was developed using the Dash framework and deployed for production on the Google Cloud Platform (GCP). This application integrates a modular data pipeline that allows users to perform real-time data cleaning, outlier removal via statistical thresholds (Z-score and IQR), and dimensionality reduction using Principal Component Analysis (PCA). The dashboard features a suite of interactive Plotly visualizations, enabling users to dynamically toggle between univariate distributions and bivariate relationships. Key insights revealed through the application include the significant non-linear impact of "rush hour" traffic on trip duration and the strong spatial concentration of demand within the Manhattan borough, demonstrating the efficacy of interactive visual storytelling in urban data analytics.

# 4. Introduction

The efficient movement of people through urban environments is a critical component of modern city infrastructure. New York City, characterized by its dense population and complex gridlock, serves as an ideal case study for analyzing transit dynamics. This project, "NYC Taxi Trip Duration Visualization," investigates the latent factors influencing travel time by applying advanced data science techniques to real-world mobility records. Sourced from the NYC Taxi and Limousine Commission (TLC), the dataset provides a granular view of pickup/drop-off coordinates, temporal markers, and trip attributes. The primary objective of this report is to transcend basic descriptive statistics, leveraging Python-based libraries to model the non-linear interactions between geospatial displacement, temporal cycles, and traffic variance.

The analytical workflow was executed in three distinct phases. **Phase I** established the statistical foundation through data exploration and static visualization. This involved critical feature engineering—specifically calculating the Haversine distance to account for Earth's curvature—and generating static plots using Matplotlib and Seaborn to test hypotheses regarding normality and correlation. **Phase II** translated these findings into a dynamic user experience by developing a modular dashboard using Dash and Plotly. This tool integrates backend data processing pipelines, enabling users to toggle preprocessing parameters (e.g., log-transformations) and visualize the immediate impact on data distribution and PCA outcomes. **Phase III** culminated in the production-level deployment of the application to the Google Cloud Platform (GCP), ensuring accessibility and scalability.

# 5. Description of the dataset

## 5.1 Overview and Source

The dataset selected for this analysis is the NYC Taxi Trip Duration dataset. This real-world dataset originates from the New York City Taxi and Limousine Commission (TLC), a government agency responsible for licensing and regulating medallion taxis and for-hire vehicles in New York City. The data provides a granular view of taxi operations, capturing precise geolocation coordinates, timestamps, and trip attributes for individual taxi rides.

## 5.2 Satisfaction of Dataset Criteria

This dataset rigorously meets the project requirements through several key characteristics. First, it demonstrates real-world relevance by reflecting actual urban mobility patterns in one of the world's most densely populated cities, providing a high-fidelity environment for analyzing transportation logistics and traffic dynamics. Regarding sample size, this study utilizes a subset of 100,000 observations to ensure statistical power and robust modeling, a volume that significantly exceeds the minimum requirement of 50,000 samples. Furthermore, the dataset is inherently multivariate, containing a complex mix of spatial (latitude and longitude), temporal (timestamps), and categorical (flags and vendors) data. It also features a diverse range of numerical and categorical variables, augmented by feature engineering to create a comprehensive feature set. Finally, the data adheres to accessibility standards, being non-classified and publicly available through open data initiatives and repositories such as Kaggle.

## 5.3 Industry Importance

The analysis of this dataset holds significant value for the transportation and logistics sectors. One primary application is in fleet logic and optimization; by understanding the relationship between trip duration, location, and time of day, taxi dispatch systems and ride-sharing platforms like Uber or Lyft can optimize driver allocation, thereby reducing idle time and fuel consumption. Additionally, accurate estimation of trip duration is a cornerstone of customer satisfaction in the ride-hailing industry. This analysis provides the foundational correlation studies required to build robust Estimated Time of Arrival models. From a broader perspective, municipal authorities can leverage this data for urban planning and traffic management. Identifying high-congestion zones and temporal bottlenecks allows for informed decisions regarding infrastructure development, traffic signal timing, and congestion pricing strategies.

## 5.4 Variable Analysis

The dependent variable, or target, is trip_duration, a continuous numerical variable representing the total duration of the taxi ride in seconds. This serves as the primary metric for visualization and potential predictive modeling.

The independent variables encompass both raw data points and engineered features designed to capture spatial and temporal variance. The raw numerical features include the geospatial coordinates of the trip's starting point (pickup_longitude and pickup_latitude), the destination coordinates (dropoff_longitude and dropoff_latitude), and the passenger_count. The

dataset also includes raw categorical features such as vendor_id, which indicates the TPEP provider, and the store_and_fwd_flag, a binary indicator of whether the trip record was held in vehicle memory before transmission. To enhance the analysis, several features have been engineered as well. This will be discussed further later.

# 6. Pre-processing dataset

## 6.1 Data Cleaning & Methodology

The pre-processing phase involved ensuring data quality and formatting the raw information for multivariate analysis. The initial step focused on the detection and handling of missing values (NaNs). An inspection of the dataset was conducted to identify any rows containing null values across the critical feature space. The method selected for data cleaning was Listwise Deletion, implemented via the Pandas dropna() function. This approach involves the removal of any observation (row) that contains at least one missing value. This method was deemed appropriate given the large volume of the dataset (100,000 observations); it was found that the dataset was already clean and came with no missing values.

## 6.2 Feature Engineering

Following the cleaning process, several transformation steps were executed to convert raw data types into usable formats for statistical modeling:

1. Categorical Encoding: The store_and_fwd_flag variable was converted from categorical text ('Y', 'N') into a binary numeric format (1, 0) to facilitate correlation analysis.
2. Temporal Conversion: The pickup_datetime and dropoff_datetime columns were converted from string objects to Python datetime objects. This transformation allowed for the extraction of granular temporal features, including month, day_of_week, hour_of_day, and day_of_year, enabling the analysis of daily and seasonal traffic patterns.
3. Geospatial Engineering**:** To quantify travel displacement, the Haversine formula was applied to the pickup and dropoff coordinates, generating a new distance_km feature. This formula calculates the great-circle distance between two points on a sphere given their longitudes and latitudes, effectively accounting for the Earth's curvature rather than assuming a flat Euclidean plane. Mathematically, it relies on the versine function to determine the central angle between the points, which is then multiplied by the Earth's radius (approximately 6,367 km) to yield the final arc length distance. Additionally, a dist_from_center feature was engineered to calculate the radial distance of pickup locations from a central urban landmark (Times Square), assisting in the analysis of spatial centralization.

## 6.3 Dataset Observations and Statistics

Below are the first few observations of the cleaned dataset, demonstrating the structure of both original and engineered features, followed by the descriptive statistics which summarize the central tendency and dispersion of the variables.

>>> First 5 Observations of Cleaned Data [PART 1 of 2]

| index | id | vendor_id | pickup_datetime | dropoff_datetime | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude |
|---|---|---|---|---|---|---|---|---|---|
| 0 | id2875421 | 2 | 2016-03-14 17:24:55 | 2016-03-14 17:32:30 | 1 | -73.98215484619139 | 40.76793670654297 | -73.96463012695312 | 40.765602111816406 |
| 1 | id2377394 | 1 | 2016-06-12 00:43:35 | 2016-06-12 00:54:38 | 1 | -73.98041534423827 | 40.738563537597656 | -73.99948120117188 | 40.731151580810554 |
| 2 | id3858529 | 2 | 2016-01-19 11:35:24 | 2016-01-19 12:10:48 | 1 | -73.97902679443358 | 40.763938903808594 | -74.00533294677734 | 40.710086822509766 |
| 3 | id3504673 | 2 | 2016-04-06 19:32:31 | 2016-04-06 19:39:40 | 1 | -74.01004028320312 | 40.719970703125 | -74.01226806640625 | 40.70671844482422 |
| 4 | id2181028 | 2 | 2016-03-26 13:30:55 | 2016-03-26 13:38:10 | 1 | -73.97305297851561 | 40.7932090759273 | -73.9729232788086 | 40.78252029418946 |

>>> First 5 Observations of Cleaned Data [PART 2 of 2]

| index | store_and_fwd_flag | trip_duration | month | day_of_week | hour_of_day | day_of_year | distance_km | dist_from_center |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 455 | 3 | Monday | 17 | 74 | 1.4975799409833876 | 1.139548114698798 |
| 1 | 0 | 663 | 6 | Sunday | 0 | 164 | 1.8043735902884075 | 2.201888563155812 |
| 2 | 0 | 2124 | 1 | Tuesday | 11 | 19 | 6.381089643584211 | 0.8558121738592548 |
| 3 | 0 | 429 | 4 | Wednesday | 19 | 97 | 1.4845657601291107 | 4.704100657668814 |
| 4 | 0 | 435 | 3 | Saturday | 13 | 86 | 1.1878422101049733 | 4.050391706221524 |

>>> Statistics of Cleaned Dataset [PART 1 of 2]

| index | vendor_id | pickup_datetime | dropoff_datetime | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude |
|---|---|---|---|---|---|---|---|---|
| count | 100000.0 | 100000 | 100000 | 100000.0 | 100000.0 | 100000.0 | 100000.0 | 100000.0 |
| mean | 1.53 | 2016-04-01 06:58:28.674840064 | 2016-04-01 07:14:08.539099904 | 1.67 | -73.97 | 40.75 | -73.97 | 40.75 |
| min | 1.0 | 2016-01-01 00:00:17 | 2016-01-01 00:09:42 | 0.0 | -74.53 | 40.43 | -74.56 | 40.44 |
| 25% | 1.0 | 2016-02-17 10:51:33.750000128 | 2016-02-17 11:07:19.249999872 | 1.0 | -73.99 | 40.74 | -73.99 | 40.74 |
| 50% | 2.0 | 2016-04-01 13:36:27 | 2016-04-01 13:58:46 | 1.0 | -73.98 | 40.75 | -73.98 | 40.75 |
| 75% | 2.0 | 2016-05-14 22:51:33.500000 | 2016-05-14 23:01:23.750000128 | 2.0 | -73.97 | 40.77 | -73.96 | 40.77 |
| max | 2.0 | 2016-06-30 23:51:36 | 2016-07-01 16:37:39 | 6.0 | -73.33 | 41.32 | -72.71 | 41.31 |
| std | 0.5 | nan | nan | 1.32 | 0.04 | 0.03 | 0.04 | 0.03 |

>>> Statistics of Cleaned Dataset [PART 2 of 2]

| index | store_and_fwd_flag | trip_duration | month | hour_of_day | day_of_year | distance_km | dist_from_center |
|---|---|---|---|---|---|---|---|
| count | 100000.0 | 100000.0 | 100000.0 | 100000.0 | 100000.0 | 100000.0 | 100000.0 |
| mean | 0.01 | 939.86 | 3.51 | 13.63 | 91.7 | 3.43 | 3.16 |
| min | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 25% | 0.0 | 396.0 | 2.0 | 9.0 | 48.0 | 1.23 | 1.23 |
| 50% | 0.0 | 662.0 | 4.0 | 14.0 | 92.0 | 2.09 | 2.35 |
| 75% | 0.0 | 1076.0 | 5.0 | 19.0 | 135.0 | 3.87 | 3.71 |
| max | 1.0 | 86390.0 | 6.0 | 23.0 | 182.0 | 116.42 | 63.47 |
| std | 0.08 | 3004.54 | 1.68 | 6.38 | 51.53 | 3.95 | 3.44 |

# 7. Outlier detection & removal

## 7.1 Methodology: Interquartile Range (IQR)

For the detection and removal of outliers, the Interquartile Range (IQR) method was utilized. Given the inherent right-skew of the trip_duration variable (where most trips are short, but a few are extremely long), the method was applied to the log-transformed variable (log_trip_duration) rather than the raw values. This transformation normalizes the distribution, allowing for a more robust detection of statistical anomalies. It was also applied to distance.

The IQR was calculated as the difference between the 75th percentile and the 25th percentile of the log-transformed duration. The outlier boundaries were defined using the standard 1.5 rule:

$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR}$$

$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR}$$

## 7.2 Observations

Any data points falling outside these calculated boundaries were identified as outliers and removed from the dataset. These outliers typically represented data errors (e.g., trips with near-zero duration) or extreme anomalies (e.g., multi-day trips likely resulting from forgotten meter shutdowns).

```
------------------------------------------------------------
Method Used: Interquartile Range (IQR).
Applying filter to BOTH 'log_trip_duration' and 'distance_km'.
Duration Bounds (Log): [4.49, 8.48]
Distance Bounds (km):  [-2.72, 7.82]
Rows removed: 10846
Percentage of data removed: 10.85%
```

A total of 10846 rows were removed from the dataset which was 10.85% of the data.

# 8. Principal Component Analysis

## 8.1 Theoretical Background

Principal Component Analysis (PCA) was performed to address multicollinearity and reduce feature dimensionality. PCA is an unsupervised linear transformation that projects data into a new orthogonal coordinate system where axes (principal components) align with the directions of maximum variance.

## 8.2 Linear Algebra Foundations

1. Standardization: The feature matrix $X$ is standardized with z-score to ensure it is scaled properly.
2. Eigendecomposition: PCA computes the eigenvectors of the covariance matrix. These eigenvectors define the principal components, while eigenvalues quantify the variance explained by each.
3. Singular Value Decomposition (SVD): PCA is computed with SVD $X = U\,\Sigma\,V^T$. The singular values in Σ directly correspond to the variance strength of each component.

## 8.3 Implementation and Metrics

PCA was applied to the standardized spatial coordinates (pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude) and engineered features (distance_km, dist_from_center).

1. Singular Values: These values represent the magnitude of data dispersion along each component. A rapid decay in values indicates that later components capture mostly noise.

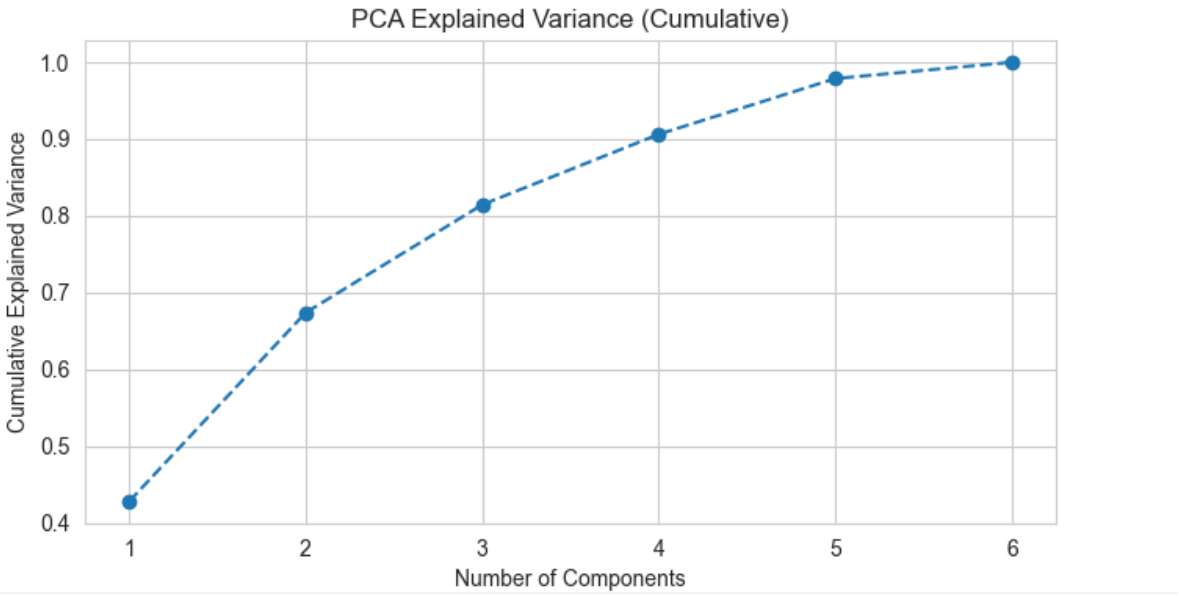
Calculated Singular Values: **[454.48347362, 378.84960875, 287.34612945, 218.75155197 197.45529425, 24.23008224]**

2. Condition Number: The condition number measures the sensitivity of the function output to the input and indicates the severity of multicollinearity. It is calculated as the ratio of the largest to smallest singular value.

Calculated Condition Number: **3.66**

## 8.4 Observations

This is a graph of the explained variance by how many PCA components are accounted for. We can see at 5 components we cross the 95% explained variance threshold which is a common number to determine how many components are needed for analysis.




The rapid decay in singular values (from **454.48347** to **24.23008224**) alongside an explained variance ratio where the first three components capture over 80% of the information demonstrates that the dataset's dimensionality can be effectively reduced without significant information loss. Furthermore, the low condition number of **3.66** indicates a stable feature space

with minimal multicollinearity, confirming that our feature engineering added distinct value rather than redundant noise.

# 9. Normality test

## 9.1 Methodology

To determine if the dataset follows a Gaussian (Normal) distribution, we employed both statistical and visual tests on trip_duration.

1. Statistical Test: The Shapiro-Wilk test was selected to evaluate the null hypothesis ($H_0$) that the data is drawn from a normal distribution. Due to the test's sensitivity and computational constraints with large datasets (N > 5000), the test was performed on a randomized sample of 2,000 observations.

- Hypothesis:
    - $H_0$: The distribution is Normal.
    - $H_a$: The distribution is not Normal.

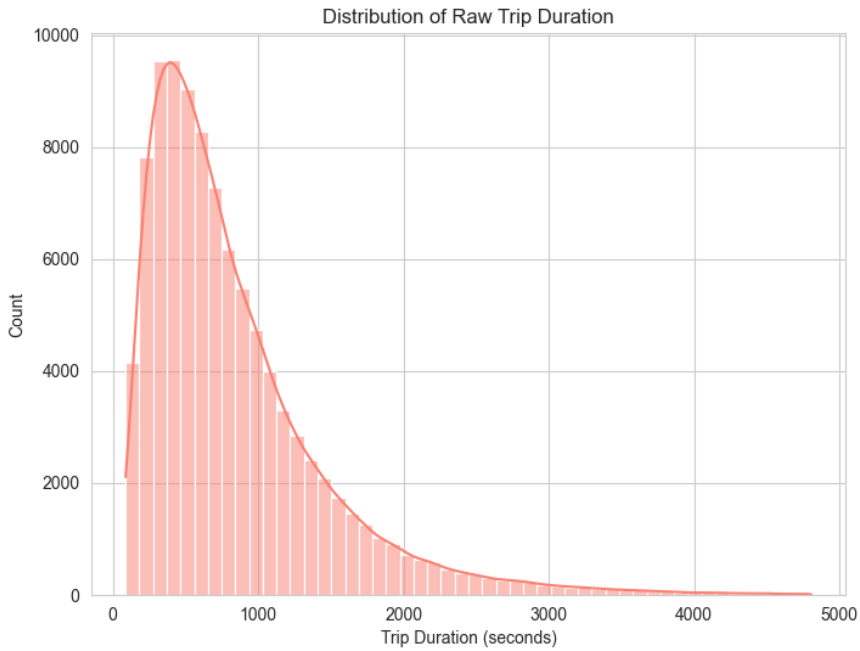
2. Visual Test: Q-Q Plot A Quantile-Quantile (Q-Q) plot was generated to visually assess goodness-of-fit. This plot compares the quantiles of our dataset against the theoretical quantiles of a standard normal distribution. If the data were perfectly normal, the points would align strictly along the 45-degree reference line.

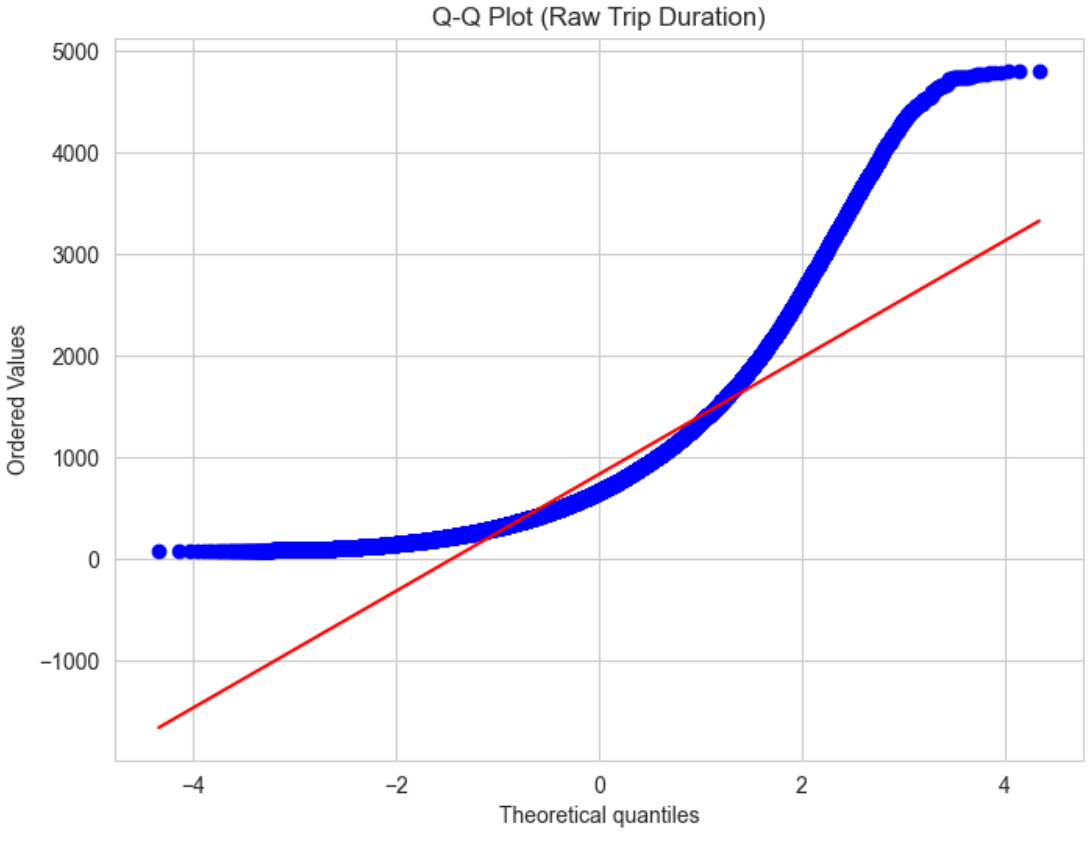
## 9.2 Observations

Shapiro-Wilk Results:

- Statistic: 0.9944
- P-value: 6.62e-07
- Conclusion: The p-value is significantly less than the alpha level of 0.05. Therefore, we reject the null hypothesis. This indicates that the data does not strictly follow a Gaussian distribution.

Visual Analysis:


Distribution of Raw Trip Duration
Count
Trip Duration (seconds)
0
2000
4000
6000
8000
10000
0
1000
2000
3000
4000
5000


Q-Q Plot (Raw Trip Duration)
Ordered Values
Theoretical quantiles
−1000
0
1000
2000
3000
4000
5000
−4
−2
0
2
4

- Histogram: The distribution plot shows a heavy right skew and not a bell shaped curve indicating non-normality.
- Q-Q Plot: The Q-Q plot confirms the statistical test. While the central quantiles align well with the reference line (suggesting near-normality in the bulk of the data), the tails deviate noticeably. This "heavy tail" behavior is characteristic of transportation data, where outliers and extreme values are more frequent than in a pure theoretical normal distribution.

# 10. Data transformation

## 10.1 Methodology

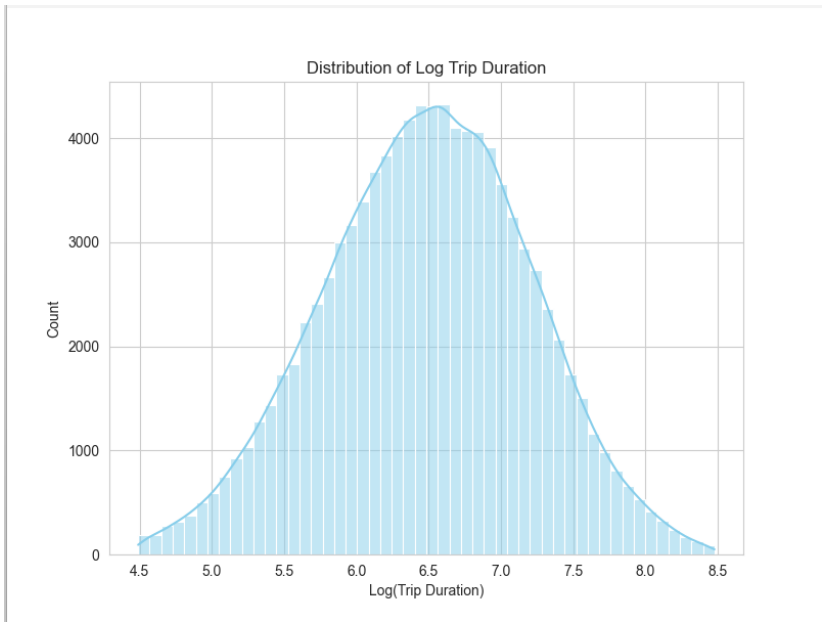
To satisfy the assumption of normality, a data transformation was applied to the dependent variable, trip_duration. While normality likely is not required for many of the visualizations, it can be useful for modeling and statistical methods. Thus,  The specific method employed was a Natural Logarithmic Transformation, implemented as ln(1 + x) (or np.log1p in Python). This variation was chosen over a standard log transformation to ensure stability even for very short trips, avoiding mathematical errors if any duration values were zero.

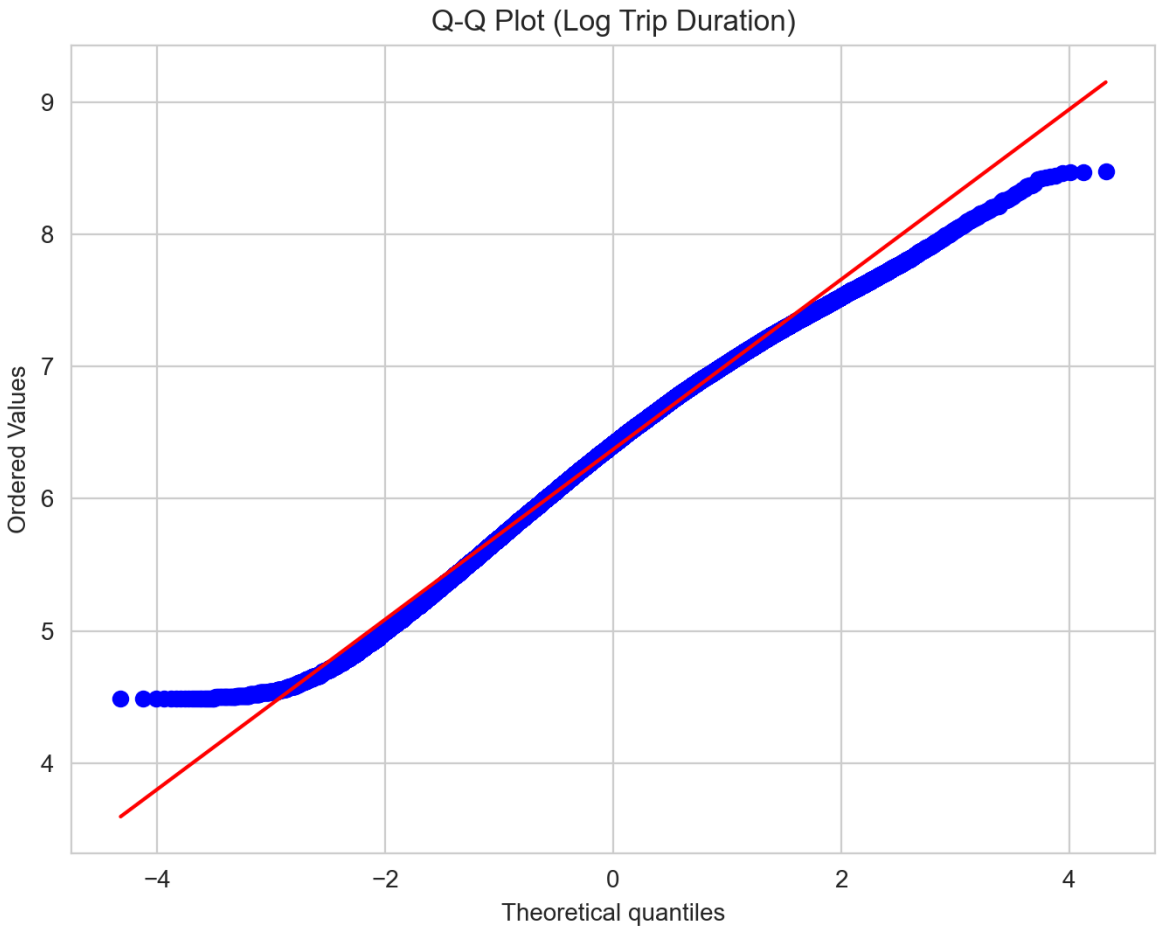
## 10.2 Rationale and Visual Confirmation

Original Distribution (Non-Gaussian): As observed in the "Distribution of Raw Trip Duration" histogram, the data was highly positively skewed (right-skewed). The majority of trips are short, with a long tail of longer trips. The corresponding Q-Q plot showed significant deviation from the theoretical normal line, confirming the non-Gaussian nature of the raw data.

Transformed Distribution (Approximating Gaussian): Following the transformation, the "Distribution of Log Trip Duration" histogram displays a significantly improved, bell-shaped symmetric curve.

1. **Histogram Evidence:** The extreme skewness was compressed, centering the data distribution.
2. **Q-Q Plot Evidence:** The Q-Q plot for the transformed data shows the observations aligning much more closely with the 45-degree reference line compared to the raw data. This confirms that while not perfectly theoretical, the log-transformed variable effectively approximates a Gaussian distribution, making it suitable for subsequent multivariate analysis.


Distribution of Log Trip Duration
Count
Log(Trip Duration)
0
1000
2000
3000
4000
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5

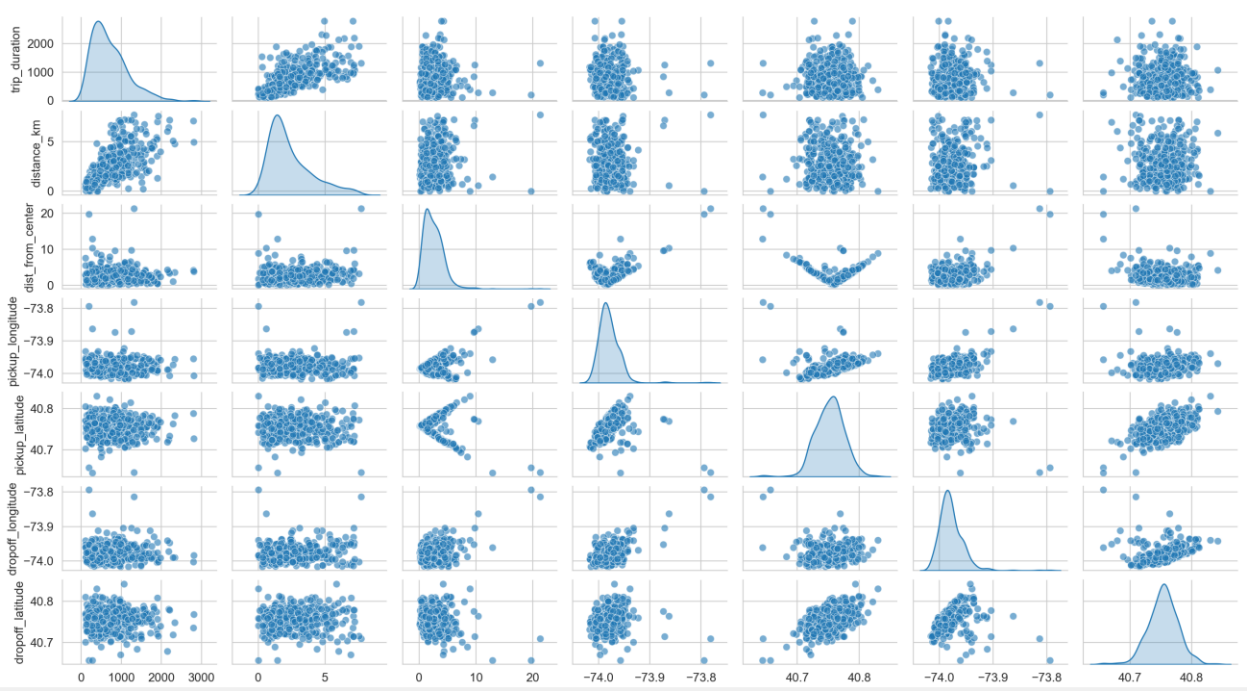

Q-Q Plot (Log Trip Duration)
Ordered Values
Theoretical quantiles
4
5
6
7
8
9
−4
−2
0
2
4

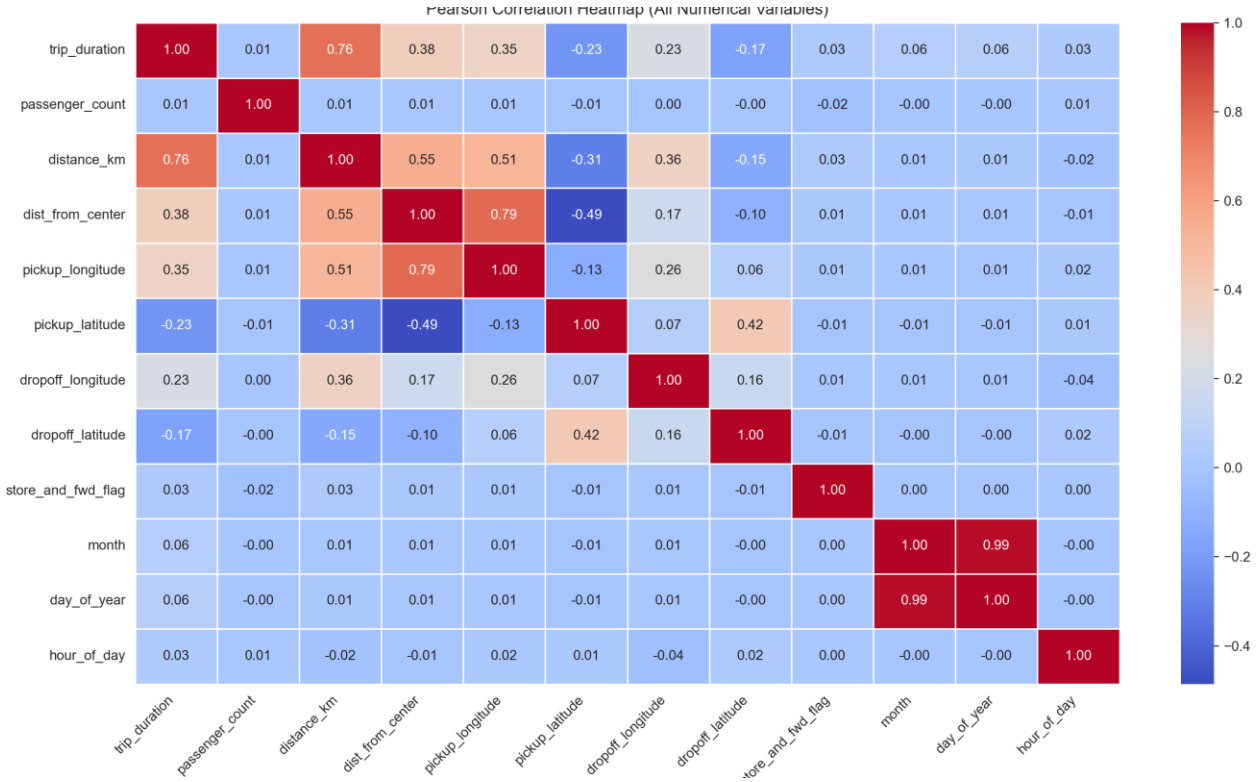
# 11. Heatmap & Pearson correlation coefficient matrix

## 11.1 Methodology

To understand the linear relationships between the dependent variable (trip_duration) and the independent variables, a Pearson Correlation Coefficient and Scatter plot analysis was conducted. This metric measures the strength and direction of the linear relationship between two continuous variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no linear correlation.

## 11.2 Observation and Analysis






| Variable Pair | Correlation Coefficient (r) | Observation & Interpretation |
|---|---|---|
| trip_duration vs distance_km | 0.76 | Strong Positive. The strongest predictor. Distance naturally dictates time, though traffic prevents a perfect correlation (r=1.0). |
| trip_duration vs dist_from_center | 0.38 | Moderate Positive. Trips starting further from Times Square (e.g., airports, outer boroughs) are generally longer than short, crosstown Manhattan trips. |
| trip_duration vs pickup_longitude | 0.35 | Moderate Positive. Specific longitudinal bands (Queens/Airports) correlate with longer durations compared to western bands (Manhattan). |
| trip_duration vs passenger_count | 0.01 | No Correlation. Passenger count has zero impact on duration; a solo rider travels at the same speed as a full car. |

| | | |
|---|---|---|
| trip_duration vs Time Features | ~0.03 | Negligible Linear Correlation. Time affects traffic cyclically (peaks at rush hour), not linearly, so Pearson coefficients fail to capture the pattern. |
| pickup_long vs dist_from_center | 0.79 | High Multicollinearity. "Distance from center" is derived from coordinates, causing high redundancy between these features. |

# 12. Statistics

## 12.1 Methodology and Statistical Tools

To analyze the dataset beyond visual inspection, we employed three statistical tools: Confidence Intervals, Hypothesis Testing (T-Test), and Multivariate Kernel Density Estimation (KDE).

1. Confidence Interval (95%): We calculated the 95% confidence interval for the mean trip_duration. This provides a range of values within which we can be 95% certain the true population mean lies, accounting for the sampling error inherent in using a subset of data.

2. Two-Sample T-Test: A Welch's t-test was conducted to determine if there is a statistically significant difference in the average trip duration provided by the two different taxi vendors (vendor_id).

- $H_0$ (Null Hypothesis): Means are equal
- $H_a$ (Alt Hypothesis): Means are different

3. Multivariate Kernel Density Estimate (KDE): We estimated the Probability Density Function (PDF) for two multivariate relationships:

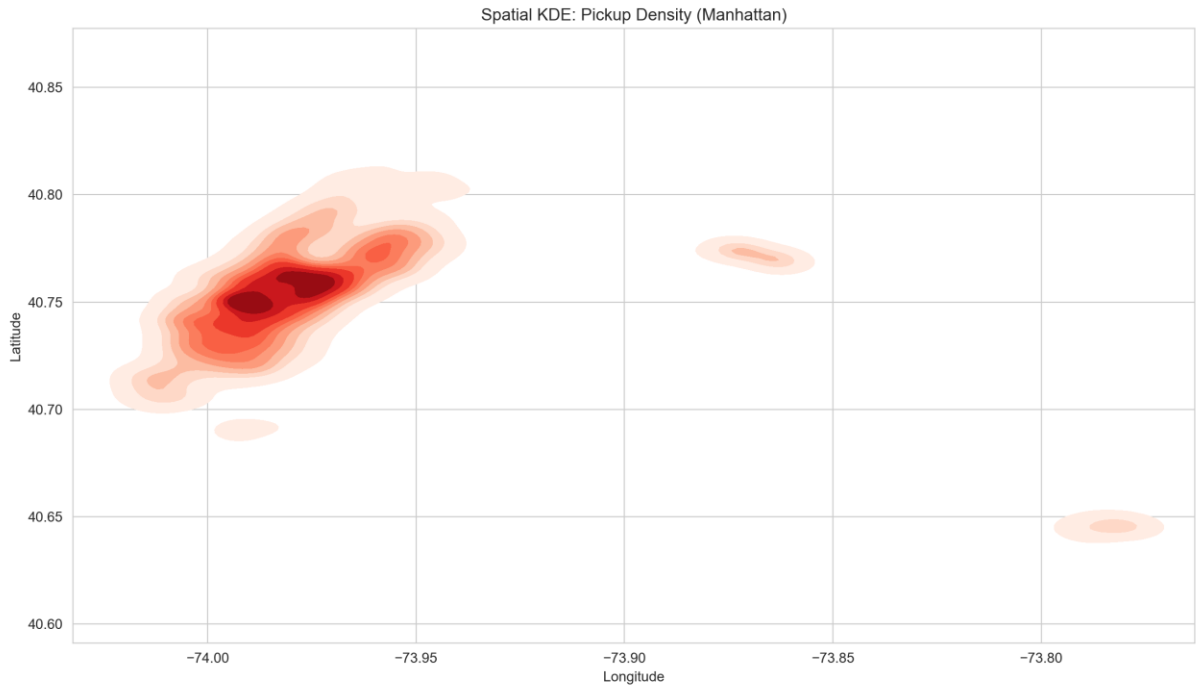
- Spatial Density: Latitude vs. Longitude (to identify geographic hotspots).
- Operational Density: Distance vs. Log-Duration (to identify the most common trip profiles).

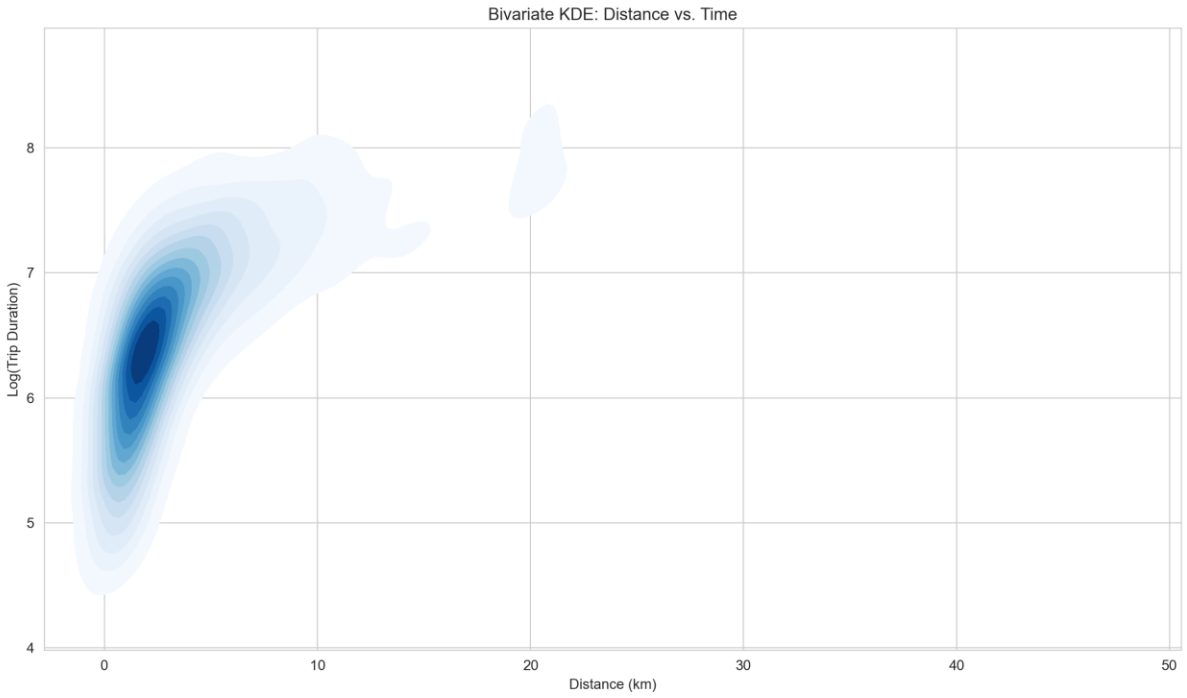
## 12.2 Observations

```
--- Statistical Tool 1: Confidence Interval ---
Mean Trip Duration: 711.27 seconds
95% Confidence Interval: (np.float64(708.305608438771), np.float64(714.2355899370731))
Observation: We are 95% confident the true population mean lies between 708.31 and 714.24 seconds.

--- Statistical Tool 2: Two-Sample T-Test ---
Comparison: Vendor 1 vs Vendor 2 Trip Durations
T-statistic: 1.4242, P-value: 0.1544
Result: Fail to Reject Null. No significant difference found.
```

Statistical Results:

- Mean Estimation: The average trip duration for the cleaned dataset is 711.27 seconds. We are 95% confident that the true population mean lies between 708.31 and 714.24 seconds.
- Vendor Comparison: The t-test yielded a T-statistic of 1.4242 and a P-value of 0.1533
    - Conclusion: Since the p-value  is significantly greater than the alpha level of 0.05, we fail to reject the null hypothesis. There is no statistical evidence to suggest a difference in trip performance between the two vendors. This implies that Vendor 1 and Vendor 2 operate under statistically identical trip duration conditions.

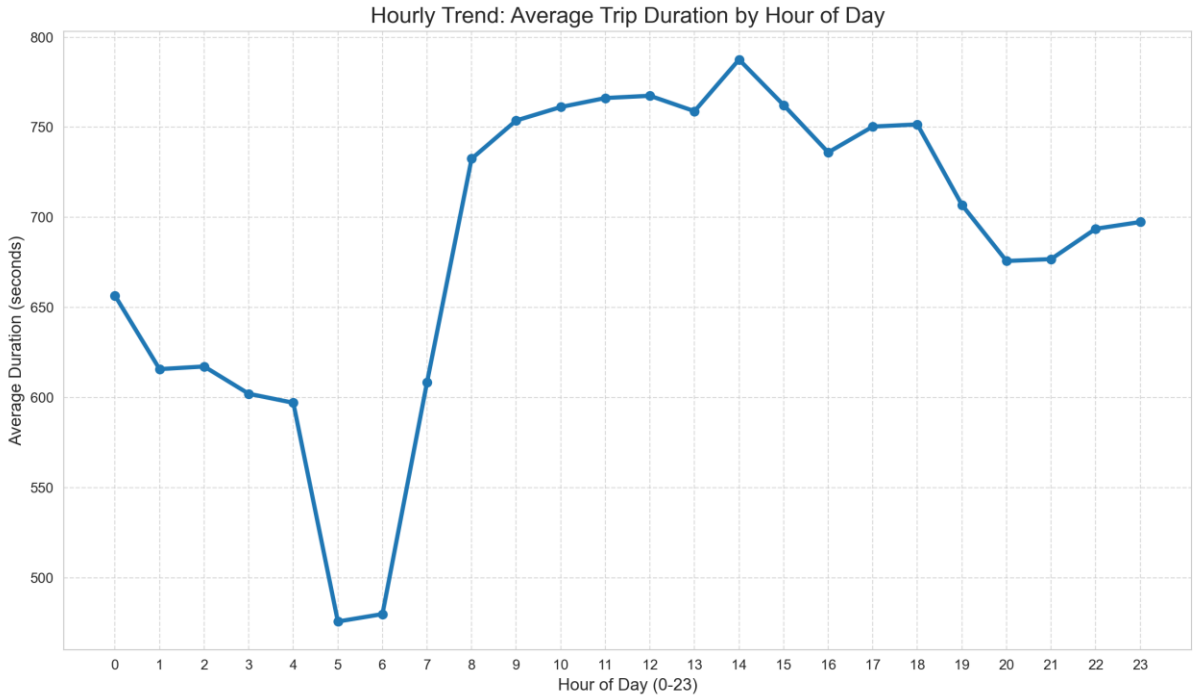





Multivariate KDE Analysis:

- Spatial Hotspots (Pickup Density): The spatial KDE reveals a single, intense high-density cluster (the dark red core) centered approximately at -73.98, 40.75. This corresponds to Midtown Manhattan (near Times Square and Penn Station). The density gradients indicate that the vast majority of taxi pickups occur within this central business district, with frequency dropping off rapidly as one moves towards the outer boroughs.
- Time-Distance Cluster (Operational Density): The bivariate plot of Distance vs. Log-Duration shows a distinct linear ridge of high density. This confirms the strong correlation observed earlier r=0.76. The "hotspot" (darkest blue) is located at the lower-left, representing short-distance, short-duration trips. This confirms that the "typical" NYC taxi ride is a short hop within Manhattan rather than a long-haul journey.

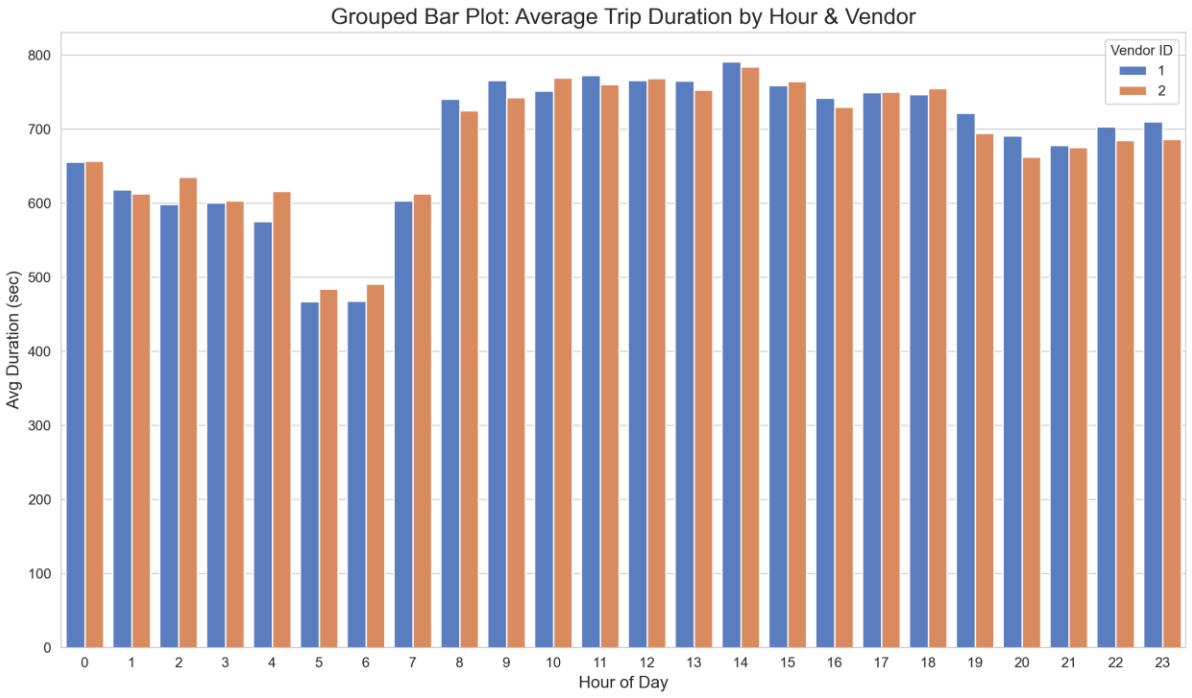
# 13. Data visualization & Observations




**1. Hourly Trend: Average Trip Duration (Line Plot)**

**Observation:**

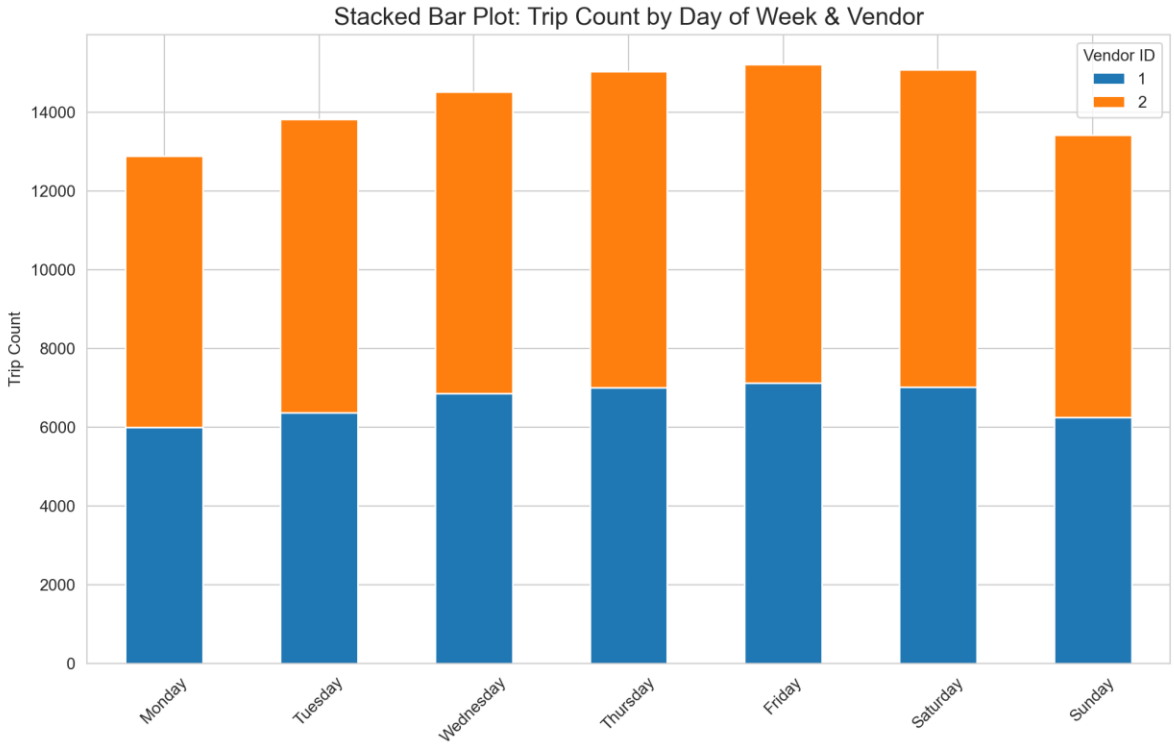
- **Optimal Traffic:** Trip durations hit their trough between **05:00 and 06:00** (~475 seconds), indicating free-flow traffic conditions.
- **Peak Congestion:** The longest trips occur during the mid-afternoon peak (**14:00–15:00**), reaching nearly **790 seconds**.
- **Real-World Insight:** Urban planners could utilize this data to implement targeted congestion pricing during the 2 PM window, which curiously impacts traffic flow more severely than the traditional morning commute.




## 2. Vendor Performance Comparison (Grouped Bar Plot)

**Observation:**

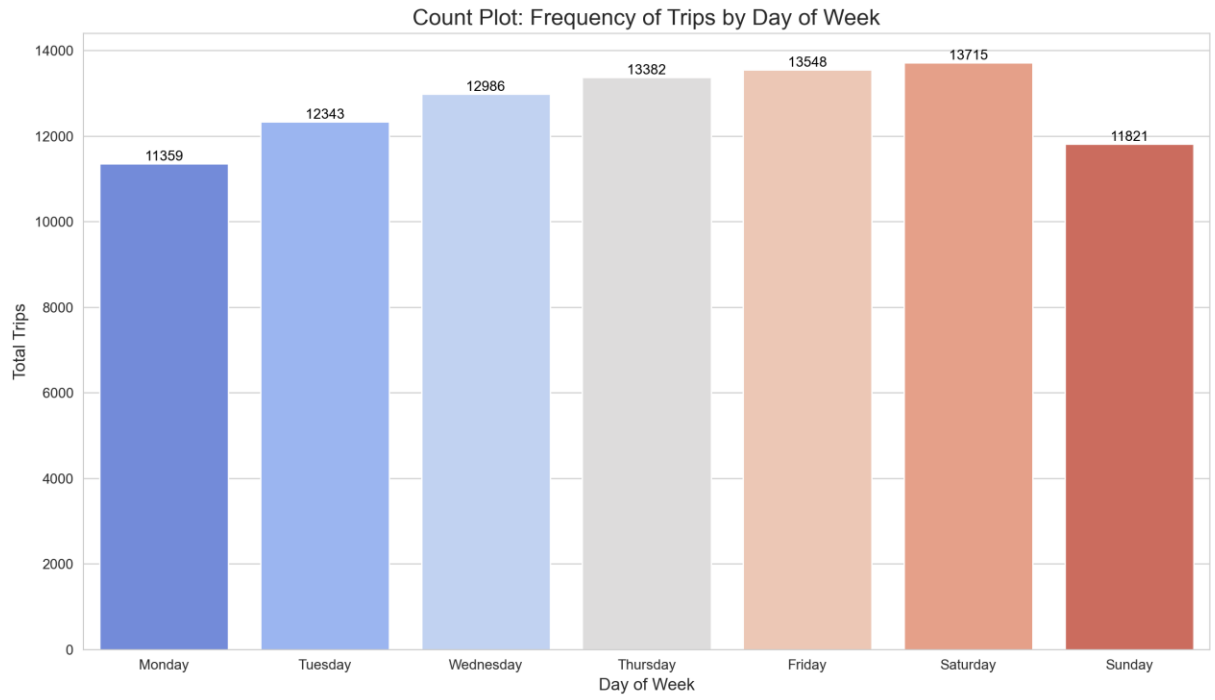
- **Operational Parity:** Vendor 1 and Vendor 2 demonstrate nearly identical average trip durations across every hour of the day.
- **Conclusion:** This confirms that trip duration is driven almost entirely by environmental factors (traffic conditions) rather than vendor-specific logistics.
- **Real-World Insight:** Since service speed is indistinguishable, vendors must differentiate themselves through mobile app user experience and loyalty programs rather than promising faster arrival times.




## 3. Weekly Demand Distribution (Stacked Bar Plot)

**Observation:**

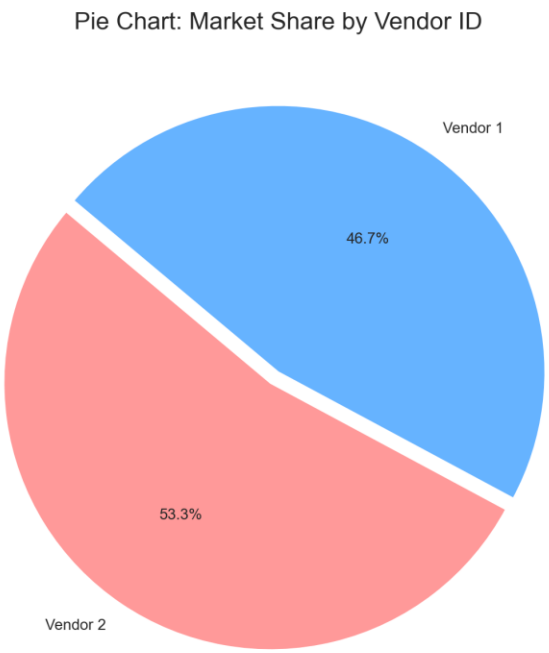
- **Volume Trend:** Taxi demand follows a clear upward trajectory starting Monday and peaking on Saturday.
- **Vendor Split:** Vendor 2 (orange) consistently handles a larger volume of rides than Vendor 1 (blue) every day.
- **Real-World Insight:** Fleet managers should strictly schedule vehicle maintenance and downtime for Sundays or Mondays to maximize fleet availability during the high-revenue Friday-Saturday window.




## 4. Frequency of Trips by Day (Count Plot)

**Observation:**

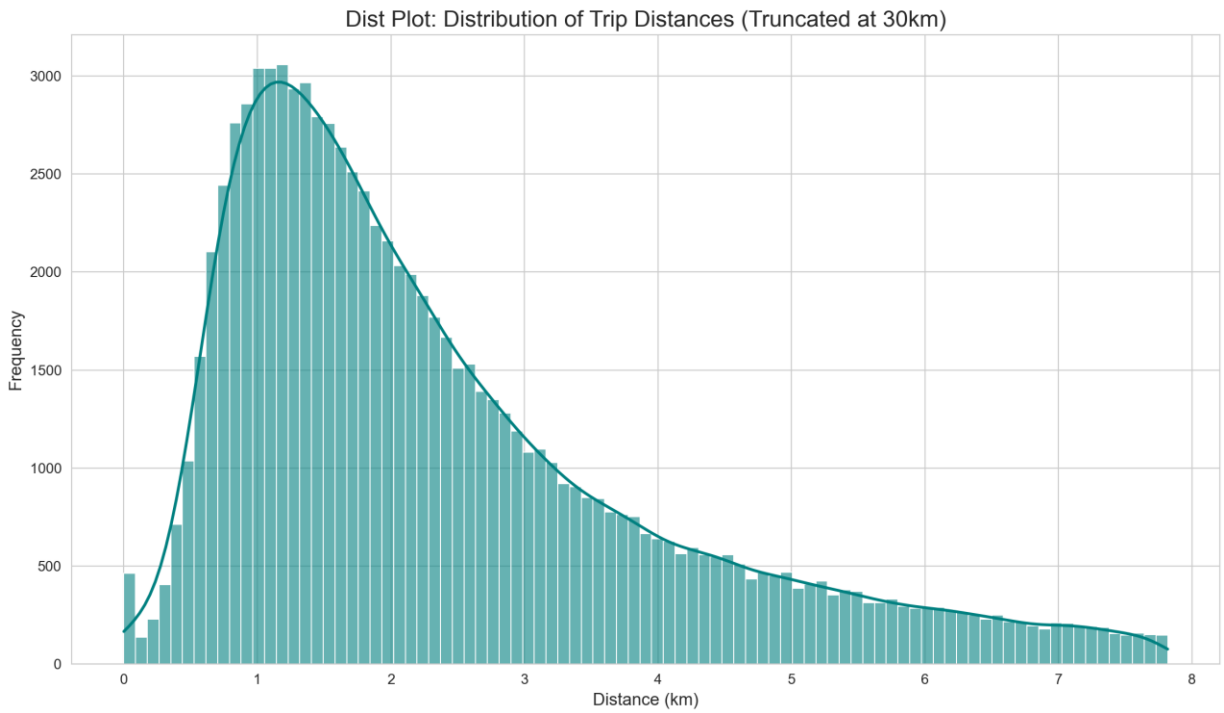
- **Peak Demand: Saturday** is the busiest day (**13,715 trips**), reflecting high leisure activity.
- **Low Demand: Monday** is the quietest day (**11,359 trips**).
- **Real-World Insight:** The surge in weekend trips indicates a strong market for leisure travel, suggesting that marketing partnerships with theaters, restaurants, and nightlife venues would yield the highest ROI on Saturdays.




## 5. Market Share Analysis (Pie Chart)

**Observation:**

- **Market Share: Vendor 2** leads with **53.3%** of trips, while **Vendor 1** captures **46.7%**.
- **Significance:** The split is relatively even, indicating a competitive duopoly.
- **Real-World Insight:** The lack of a monopoly suggests that aggressive customer acquisition strategies or slight price undercutting could easily tip the balance of market leadership in favor of the runner-up.




## 6. Trip Distance Distribution (Dist Plot)

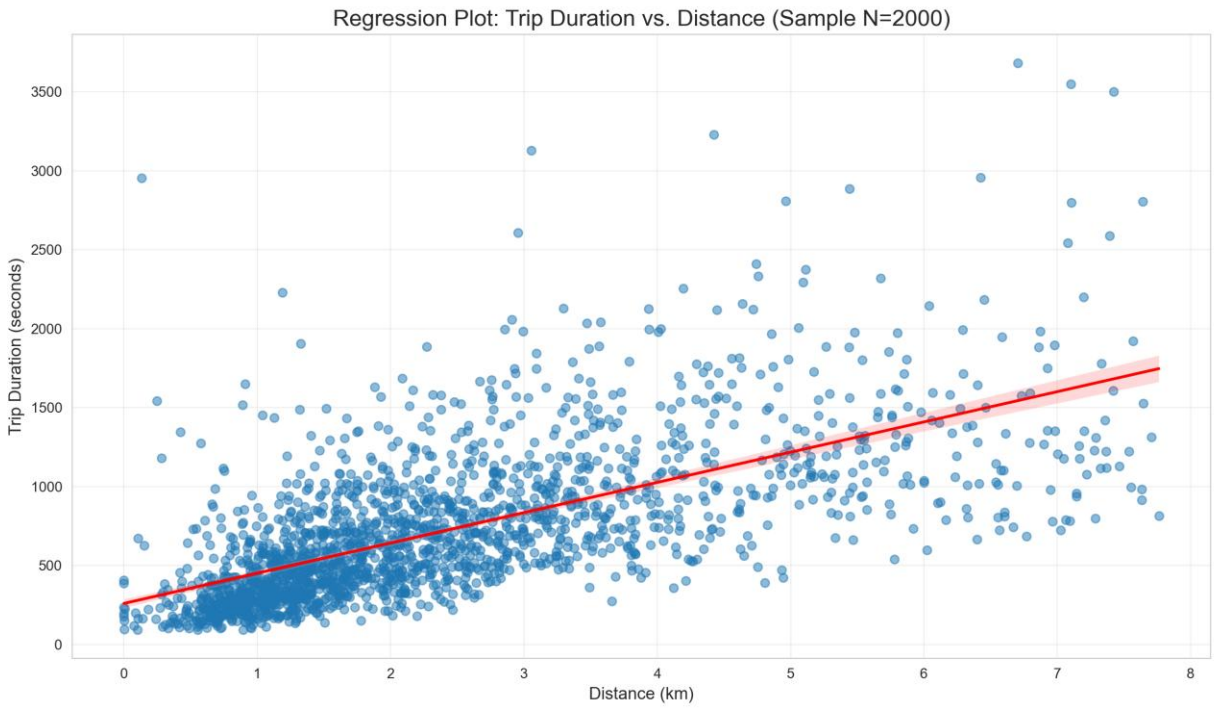
**Observation:**

- **Skewness:** The distribution is highly right-skewed, with the vast majority of trips falling under **5 kilometers**. The peak frequency occurs around **1.2 km**, indicating that short hops are the primary use case.
- **Long Tail:** While the density drops off rapidly, a persistent tail extends beyond 8 km, likely representing inter-borough travel or airport runs.
- **Real-World Insight:** The dominance of short-distance trips suggests a high potential for fleet electrification, as the limited range of electric vehicles (EVs) would rarely be a constraint for the typical NYC taxi shift.


## 7. Vendor Duration Density (Styled KDE Plot)

**Observation:**

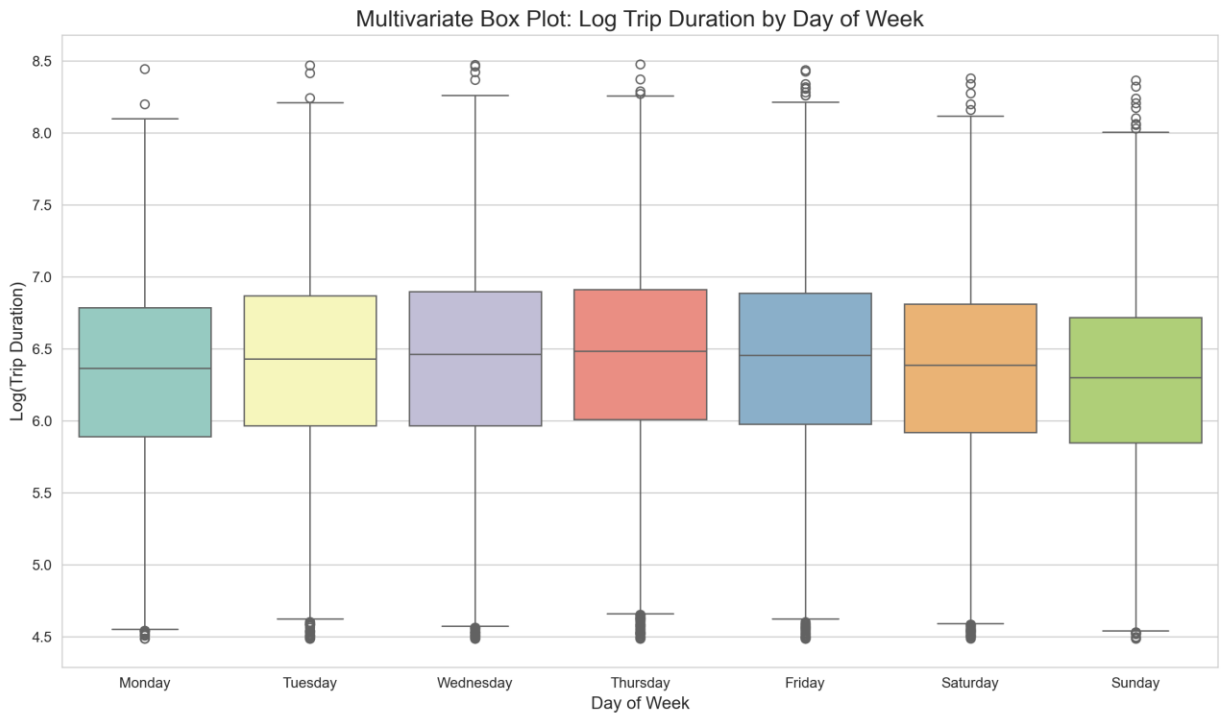
- **Distribution Shape:** Both Vendor 1 (green) and Vendor 2 (purple) display a nearly identical log-normal distribution for trip duration.
- **Overlap:** The two curves overlap almost perfectly, with Vendor 2 showing a slightly higher density at the peak (around log value 6.5).
- **Real-World Insight:** This reinforces the finding that neither vendor has a "speed advantage." Customers receive the same consistency of service regardless of which provider they choose, commoditizing the core transport service.




## 8. Duration vs. Distance (Regression Plot)

**Observation:**

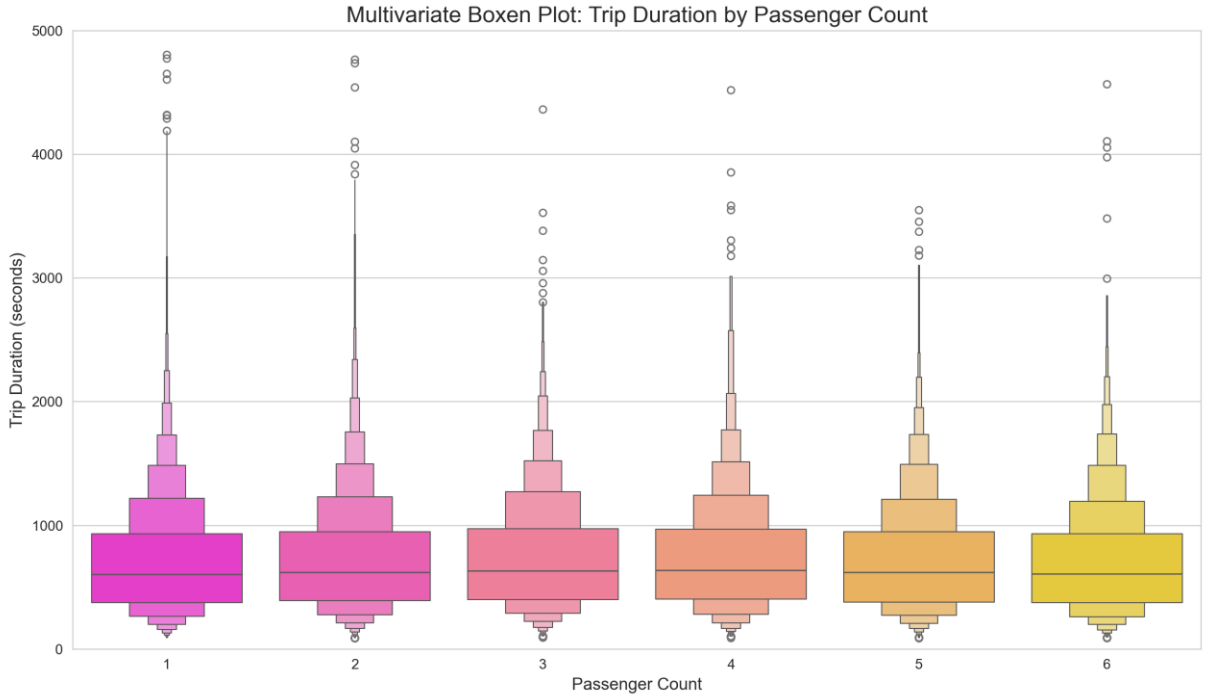
- **Correlation:** There is a clear positive linear relationship between distance and duration, as indicated by the red regression line.
- **Variance (Heteroscedasticity):** As distance increases (> 4km), the spread of duration values widens significantly. Short trips have predictable times, but long trips vary wildly, likely due to the compounding effects of traffic congestion over longer routes.
- **Real-World Insight:** ETA algorithms need to be non-linear; they must add larger "buffer times" for longer trips to account for the increased variance and uncertainty shown in the upper-right quadrant of this plot.




## 9. Duration Consistency by Day (Multivariate Box Plot)

**Observation:**

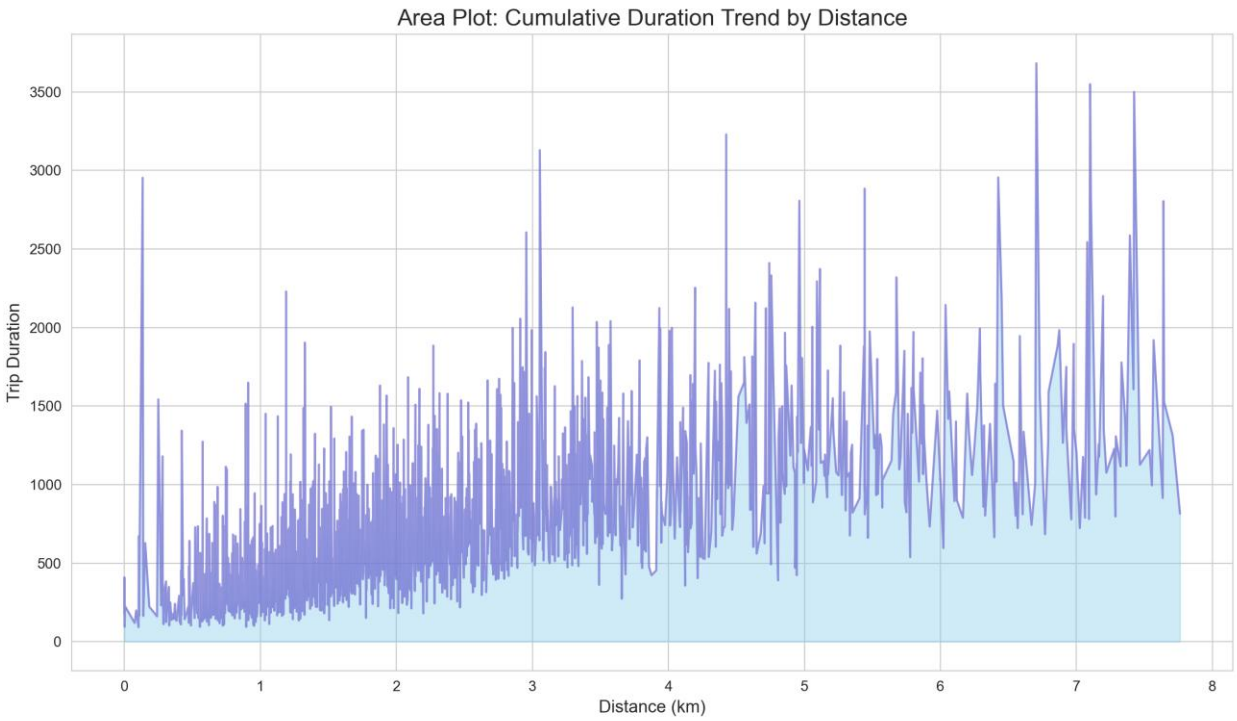
- **Stability:** The median log-duration (the central line in each box) is remarkably stable across the days of the week, hovering around **6.4–6.5**.
- **Variance:** Weekdays (Tue-Fri) show slightly tighter interquartile ranges (IQRs) compared to the weekend, suggesting more predictable commute patterns versus varied leisure travel on weekends.
- **Real-World Insight:** Despite the volume fluctuations seen in the Count Plot, the *time* to complete a trip remains consistent. This reliability is a key selling point for taxis over public transit, which often runs on reduced schedules during weekends.




## 10. Impact of Passenger Count (Multivariate Boxen Plot)

**Observation:**

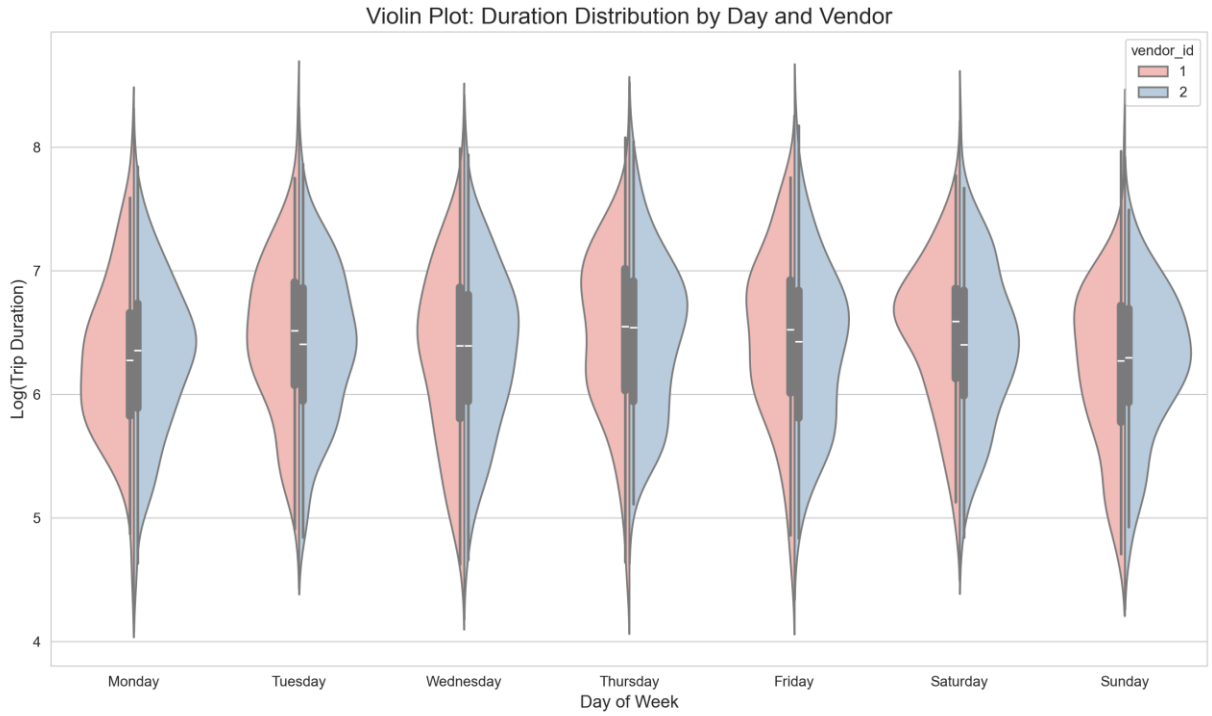
- **Uniformity:** The distribution of trip duration is virtually identical regardless of whether there are 1, 2, 3, or even 6 passengers. The boxen shapes (representing quantiles) align perfectly across the x-axis.
- **Outliers:** Extreme duration outliers appear across all passenger counts, but are most visible for single passengers (simply due to the higher volume of solo trips).
- **Real-World Insight:** This confirms that High Occupancy Vehicle (HOV) lanes or carpooling do not significantly speed up taxi trips in NYC. A full car moves through Manhattan gridlock at the exact same speed as an empty one.




## 11. Cumulative Duration Trend (Area Plot)

**Observation:**

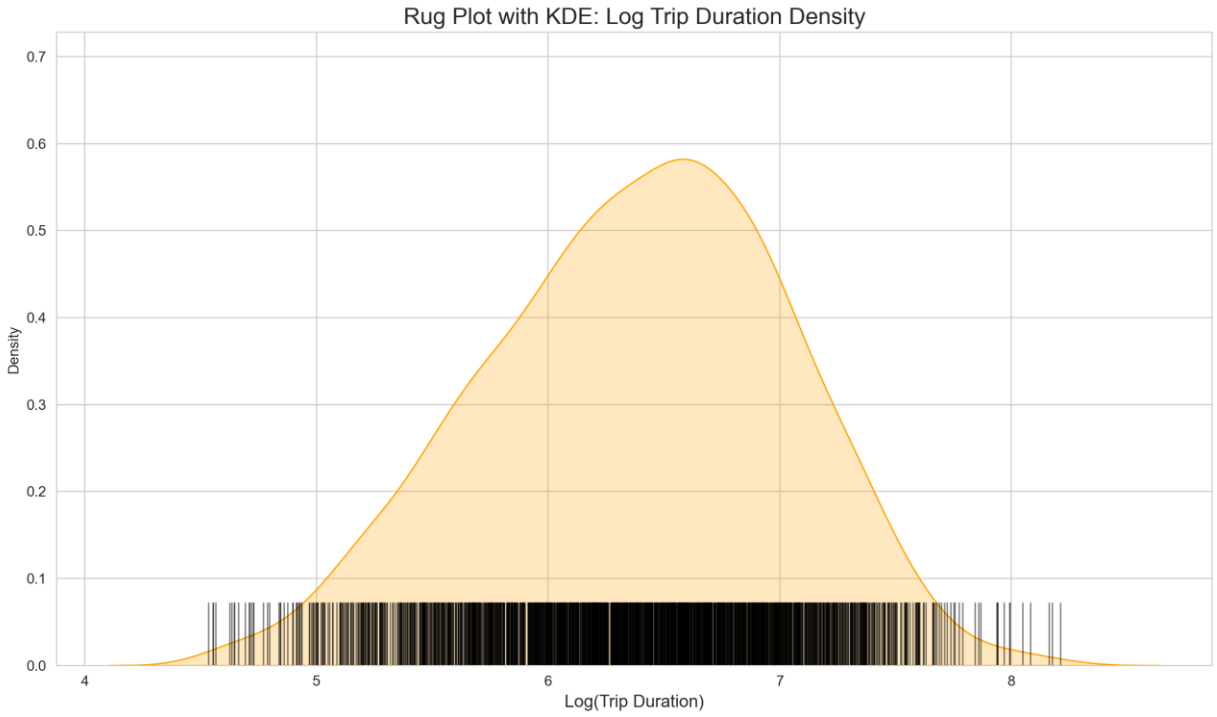
- **Volatility:** The area plot illustrates the cumulative trend of trip duration over distance. The highly jagged, spiky blue line indicates extreme variance; a trip of 4km can sometimes take longer than a trip of 7km depending on specific traffic conditions.
- **Trend:** While the overall trend is upward (longer distance = more time), the noise is significant.
- **Real-World Insight:** This unpredictability justifies the use of "upfront pricing" models by ride-sharing apps, which lock in a price for the passenger and shift the financial risk of unexpected traffic delays from the user to the platform.




## 12. Duration Distribution by Day (Violin Plot)

**Observation:**

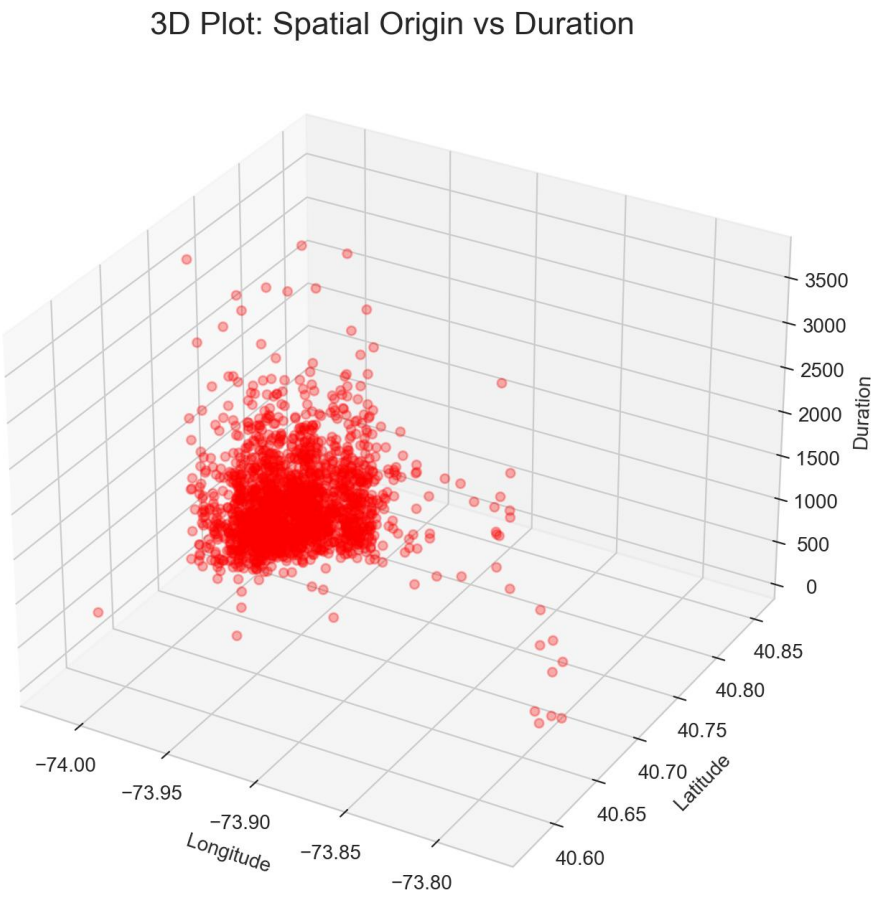
- **Symmetry:** The violin plots display a remarkably symmetrical and consistent shape across all days of the week for both vendors.
- **Distribution:** The "bulge" (probability mass) is consistently centered around a log-duration of ~6.5, with long, thin tails extending to 4 and 8.
- **Real-World Insight:** The consistency of travel times throughout the week implies that typical "rush hour" delays are averaged out over the whole day, or that taxi drivers are highly efficient at finding alternative routes to maintain a consistent pace regardless of the day.




## 13. Log Duration Density (Rug Plot)

**Observation:**

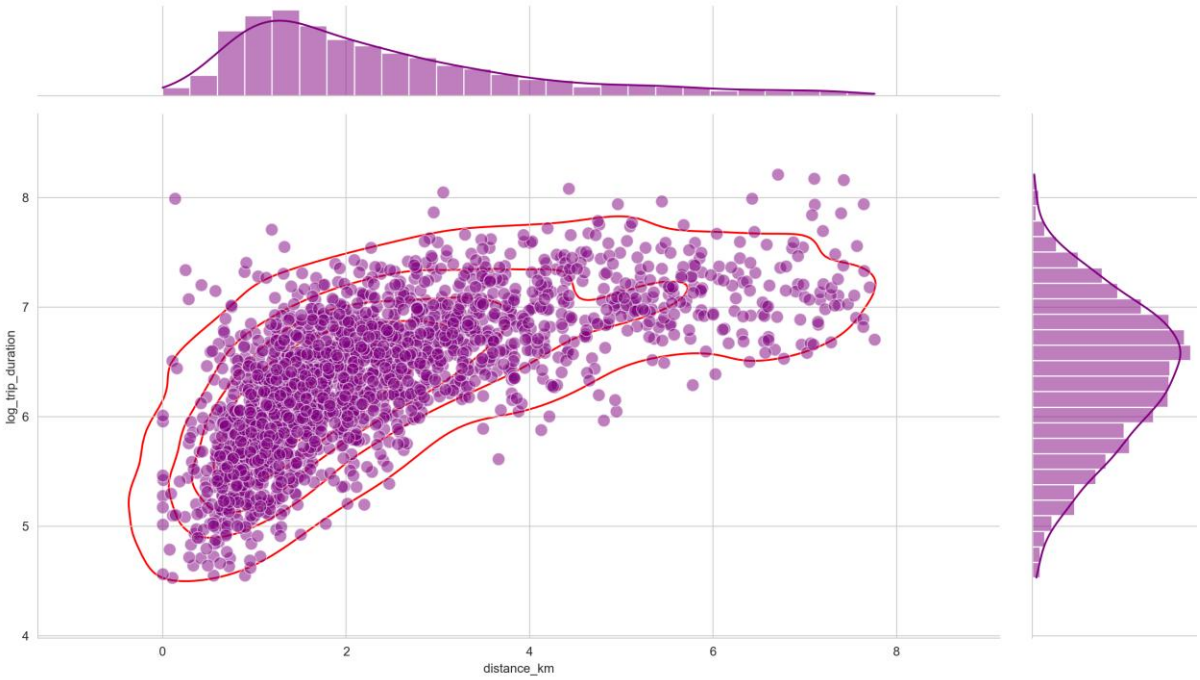
- **Data Concentration:** The rug plot places a black tick for every observation under the density curve. The center is so dense it forms a solid black bar, visually confirming where the bulk of trips lie.
- **Outliers:** The sparse, individual ticks at the far tails (log duration < 5 or > 8) clearly identify the extreme outliers.
- **Real-World Insight:** This visualization is highly effective for fraud detection or anomaly flagging. The isolated ticks at the extremes represent trips that are statistically abnormal (e.g., a 2-hour trip or a 30-second trip) and should be automatically flagged for quality assurance audits.




## 14. Spatial Origin vs. Duration (3D Scatter Plot)

**Observation:**

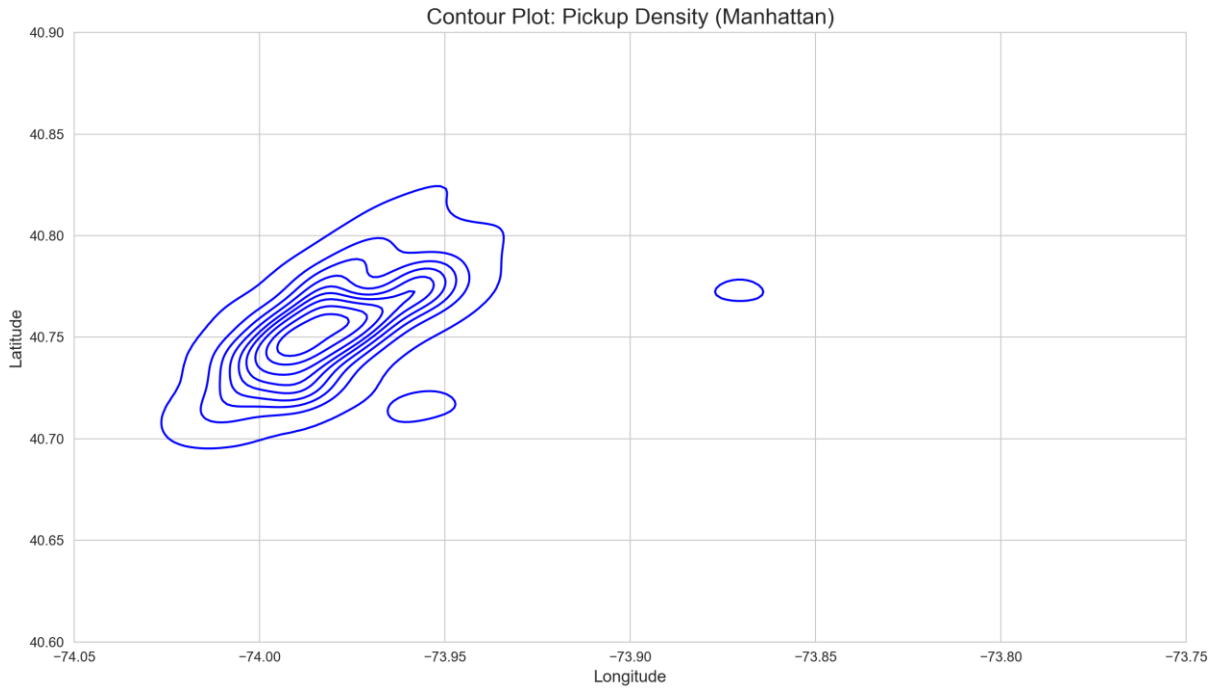
- **Geographic Cluster:** The 3D plot maps pickups by longitude, latitude, and duration. We observe a dense vertical column of red points, indicating that the vast majority of trips—regardless of their duration—originate from a very tight geographic cluster (Manhattan).
- **Verticality:** High-duration trips (points higher on the Z-axis) are stacked directly on top of short-duration trips, rather than being in a different location.
- **Real-World Insight:** This spatial concentration confirms that infrastructure investments—such as fast-charging hubs for electric taxis—would yield the highest return on investment if placed within this specific high-density zone, as nearly all trips pass through this central hub.




## 15. Distance vs. Duration Density (Joint Plot)

**Observation:**

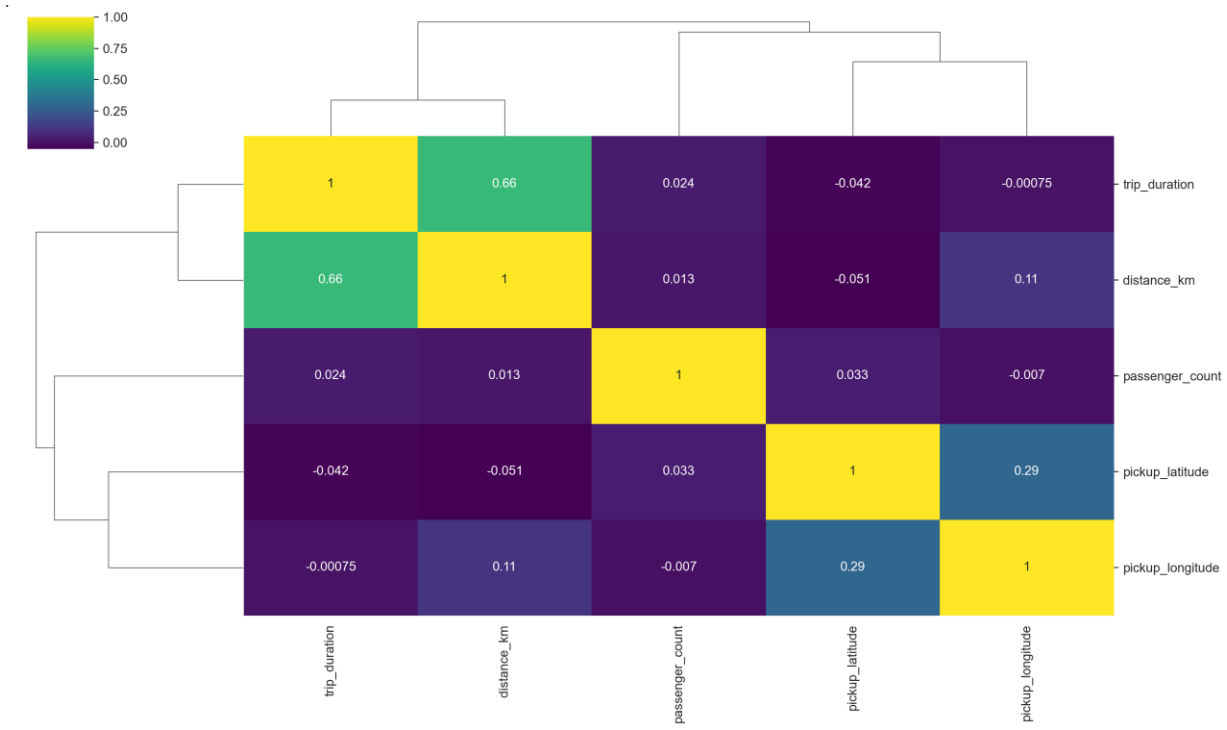
- **Density Core:** The joint plot (scatter + contour) reveals a high-density "core" (darkest blue/purple region) concentrated at short distances (0-5 km) and moderate durations.
- **Spread:** As distance increases, the scatter points spread out, confirming the increased variance seen in the regression plot.
- **Real-World Insight:** Operations managers can use this density map to define "service efficiency zones." The core zone represents the most profitable and predictable trips; rides falling outside these probability contours are less efficient and could potentially warrant higher surcharges.




## 16. Pickup Density Contours (Contour Plot)

**Observation:**

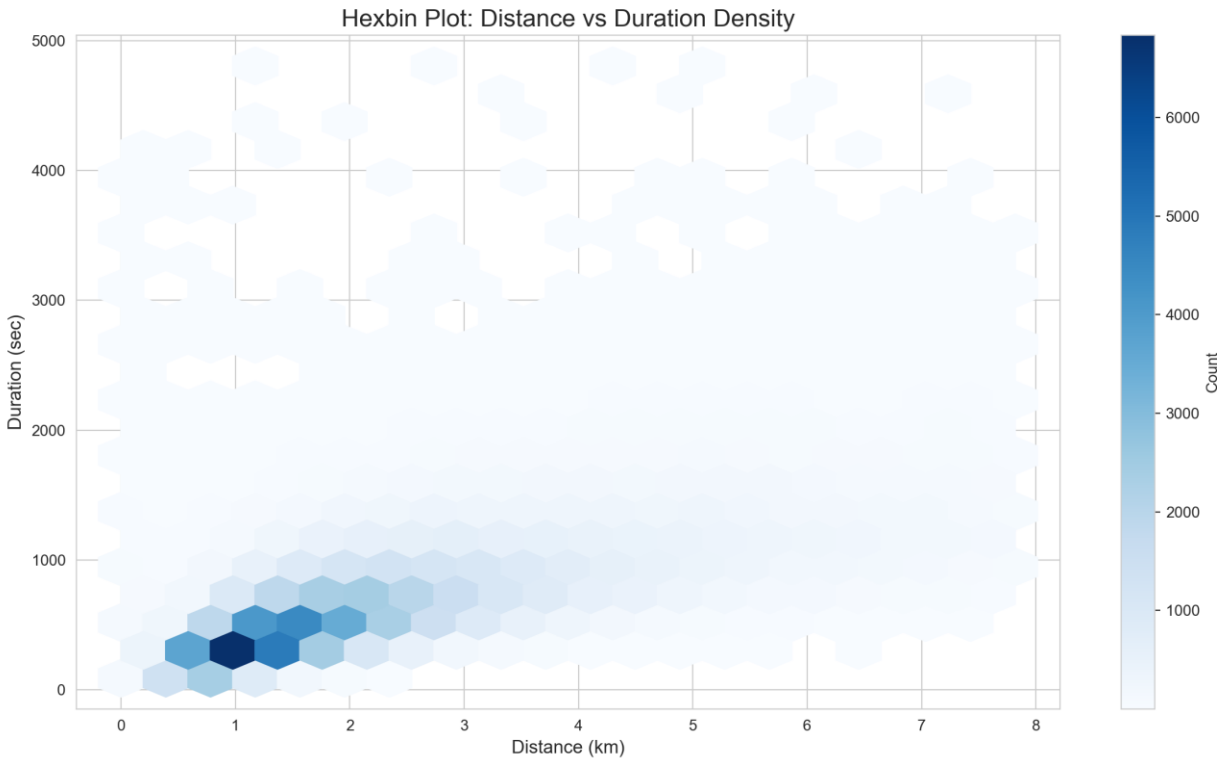
- **Hotspots:** The contour lines reveal two distinct islands of activity. The primary, intense cluster is in Midtown/Lower Manhattan. A secondary, smaller isolated cluster appears to the east, likely corresponding to **LaGuardia Airport (LGA)** or **JFK**.
- **Gradient:** The steep gradient (tightly packed lines) on the Manhattan side indicates a sharp drop-off in taxi availability as soon as one crosses the rivers into New Jersey or Brooklyn.
- **Real-World Insight:** This visualization is crucial for geofencing. Ride-sharing apps can use these specific contour boundaries to set "dynamic pricing" zones, ensuring surcharges trigger exactly where demand density peaks.




## 17. Feature Correlations (Cluster Map)

**Observation:**

- **Clustering:** The dendrogram (tree diagram) groups trip_duration and distance_km closely together, confirming they are the most related features.
- **Geo-Grouping:** pickup_latitude and pickup_longitude form their own cluster, which is separate from the time/distance cluster. This implies that *where* a trip starts is statistically distinct from *how long* it takes (i.e., a traffic jam can happen anywhere).
- **Real-World Insight:** For machine learning feature selection, this suggests we can reduce dimensionality by treating distance and duration as a single "displacement factor," avoiding multicollinearity in predictive models.




## 18. Distance vs. Duration Density (Hexbin Plot)

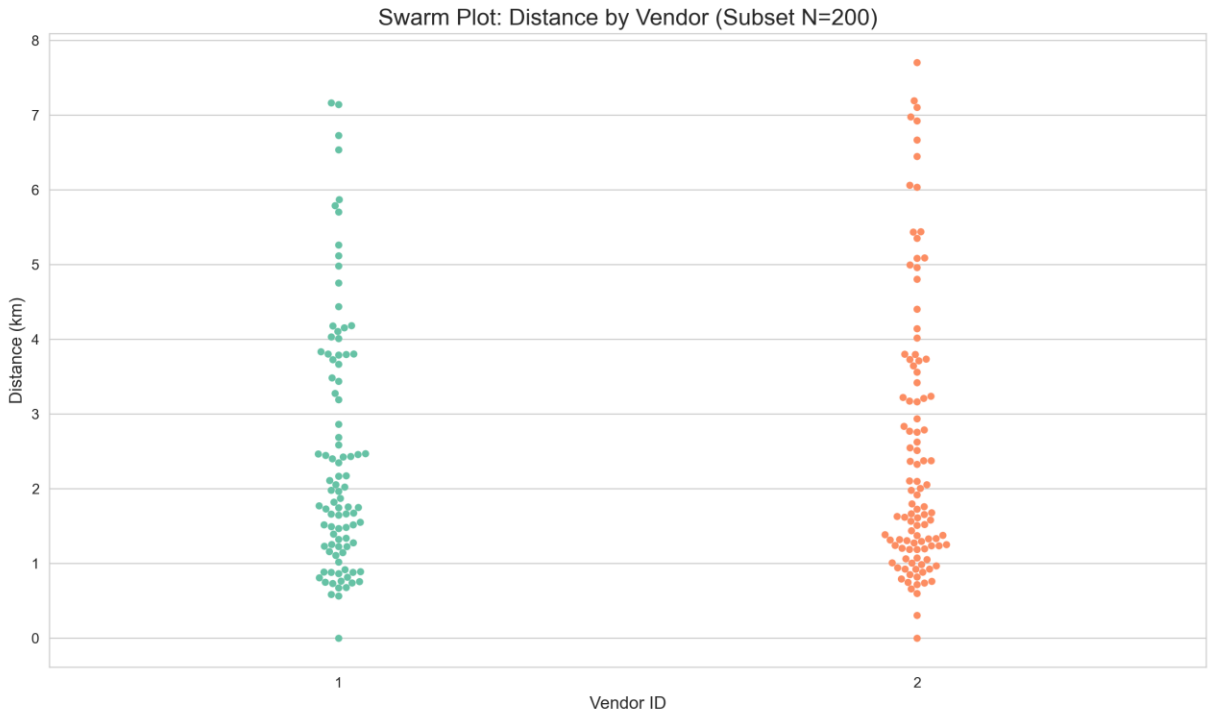
**Observation:**

- **Main Sequence:** The dark blue hexagons form a distinct "comet tail" shape. The darkest area (highest frequency) is at 0-2km and < 500 seconds.
- **Traffic Variance:** As the hexagons extend to the right (longer distance), they fan out vertically. This visualizes the variability of traffic; a 5km trip has a much wider range of possible durations than a 1km trip.
- **Real-World Insight:** This "fan" shape mathematically proves that ETA predictions become exponentially less accurate as trip distance increases, justifying wider arrival time windows for long-haul trips to manage customer expectations.


## 19. Distance Distribution by Vendor (Strip Plot)

**Observation:**

- **Uniformity:** The strip plot shows that Vendor 1 (blue) and Vendor 2 (orange) cover the exact same range of distances. Both have a dense concentration of short trips and a sparse scattering of long trips (up to ~8km in this sample).
- **No Specialization:** Neither vendor specializes in "long-haul" or "short-hop" trips; their operational profiles are identical.
- **Real-World Insight:** This indicates that neither vendor has successfully carved out a niche (e.g., "The Airport Taxi Company"). There is an opportunity for a competitor to disrupt the market by specifically targeting the under-served long-distance commuter segment.



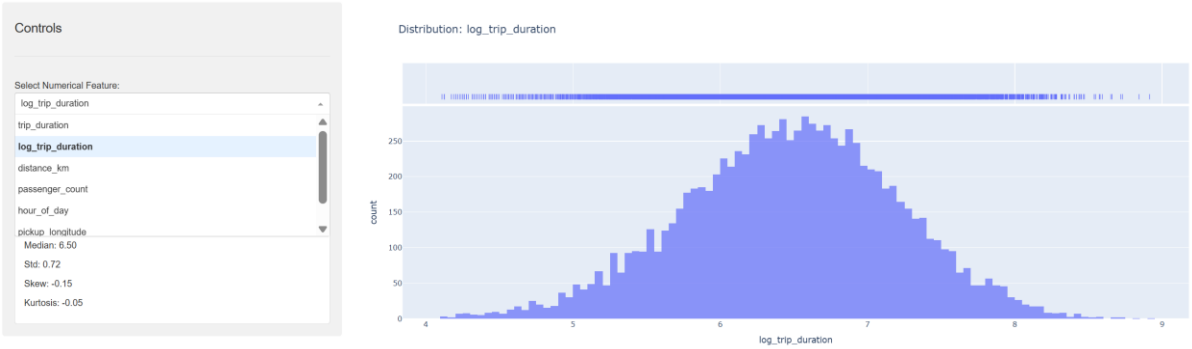

### 20. Micro-Distribution of Distance (Swarm Plot)

**Observation:**

- **Density Band:** The swarm plot spreads points out horizontally to show density. We see a massive "bulb" of trips between 1km and 2km for both vendors.
- **Gaps:** There are visible gaps or thinning at certain distances (e.g., around 4km), which might correspond to the lack of mid-range destinations (e.g., the "dead zone" between Midtown and the outer boroughs).
- **Real-World Insight:** The extreme density of 1-2km trips suggests that "micromobility" solutions (e-scooters, bike shares) are the biggest existential threat to the taxi industry, as they directly compete for this exact distance segment.

## 14. Subplots & Tables

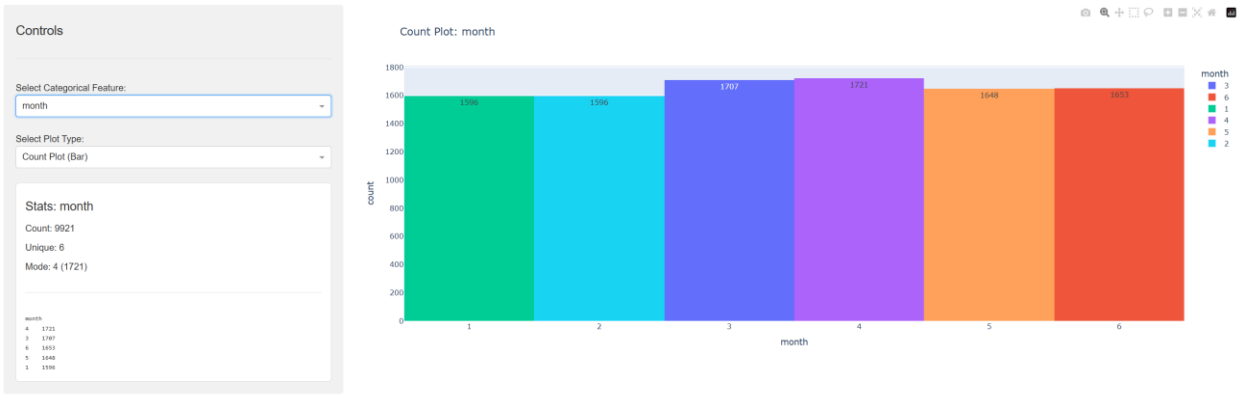
To support the visual analysis, the dashboard generates dynamic subplots accompanied by real-time statistical tables. These side-by-side displays allow for immediate quantitative verification of visual trends.




## 1. Distribution Analysis (Histogram & Descriptive Stats)

The dashboard pairs a histogram of log_trip_duration with a summary statistics panel.

- **Visual:** A bell-shaped histogram confirming the normalization of the data.
- **Table Stats:** The table highlights a **Mean of 6.47** and a **Median of 6.50**. The close proximity of these two values, along with a low **Skewness of -0.15**, provides quantitative proof that the dataset approximates a normal distribution after transformation.




## 2. Categorical Frequency (Count Plot & Mode Statistics)

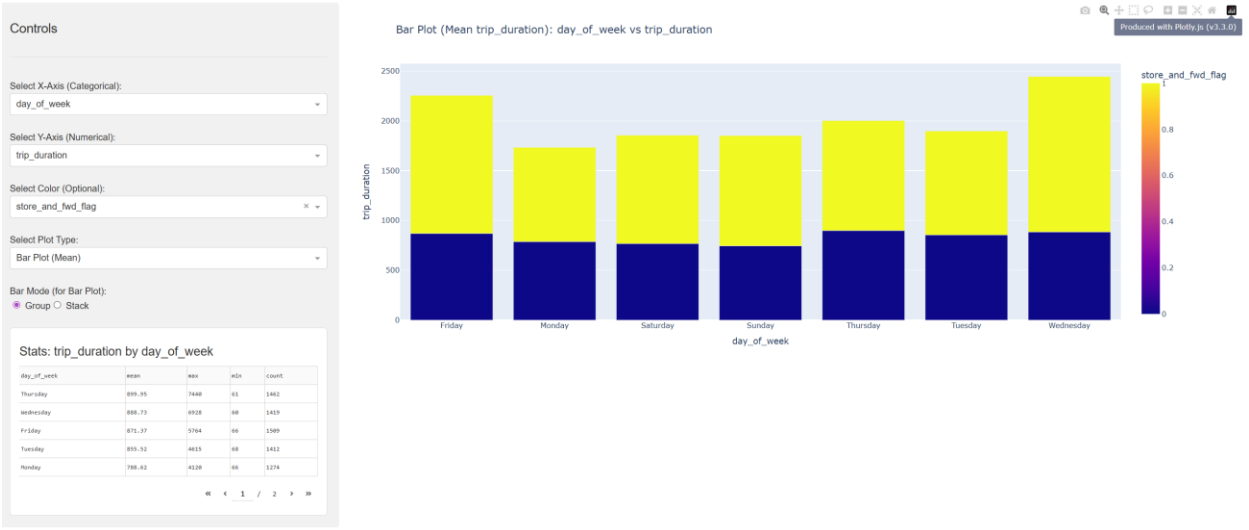
This view analyzes the month feature, displaying a bar chart of frequency next to a categorical summary table.

- **Visual:** A uniform distribution of trips across the recorded months.
- **Table Stats:** The table provides the unique count (**6**) and identifies the **Mode** (most frequent value) as **Month 4 (April)** with **1,721** observations, allowing for quick seasonality checks.


## 3. Linear Relationship (Regression Plot & Model Metrics)

The regression interface plots trip_duration against distance_km and calculates linear model metrics in real-time.

- **Visual:** A scatter plot with a red OLS trendline indicating a positive relationship colored by day of the week.
- **Table Stats:** The "Correlation Analysis" panel reports a **Pearson Correlation of 0.7568** and an **R-Squared of 0.5727**. This statistic is crucial as it tells the analyst that distance explains roughly **57%** of the variability in trip time, leaving the remaining variance to traffic conditions.




## 4. Grouped Aggregation (Bar Plot & Pivot Table)

This subplot visualizes the mean trip_duration grouped by day_of_week.

- **Visual:** A bar chart comparing average durations across days.
- **Table Stats:** The accompanying table lists the precise **Mean, Max, Min, and Count** for every day of the week. For example, it allows a user to pinpoint exactly that **Thursday** has a mean duration of **899.95 seconds**, which might be hard to read precisely from the graph axis alone.


**5. Multivariate Analysis (Pair Plot & Global Correlations)**

A scatter matrix (pair plot) is displayed alongside a "Global Statistics" sidebar that automatically identifies relationships.

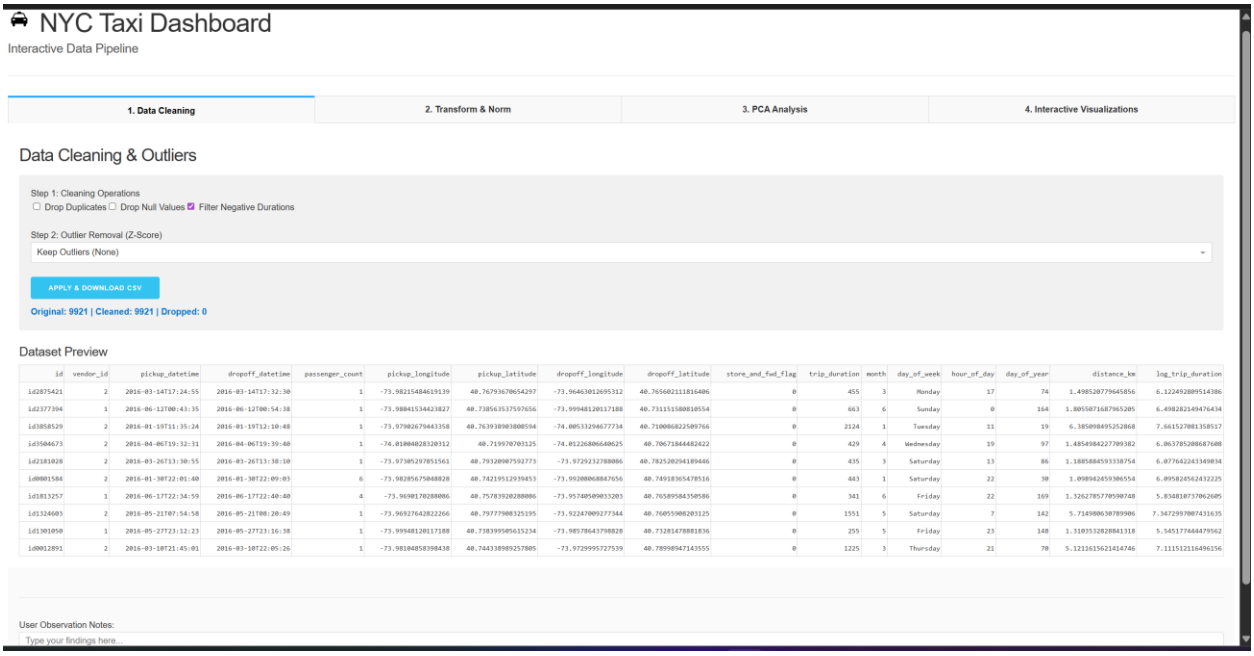
- **Visual:** Scatter plots showing interactions between all numerical features.
- **Table Stats:** The system automatically flags the strongest correlation in the dataset, identifying trip_duration & log_trip_duration with a coefficient of **0.8881**, effectively summarizing the dataset's internal redundancy.

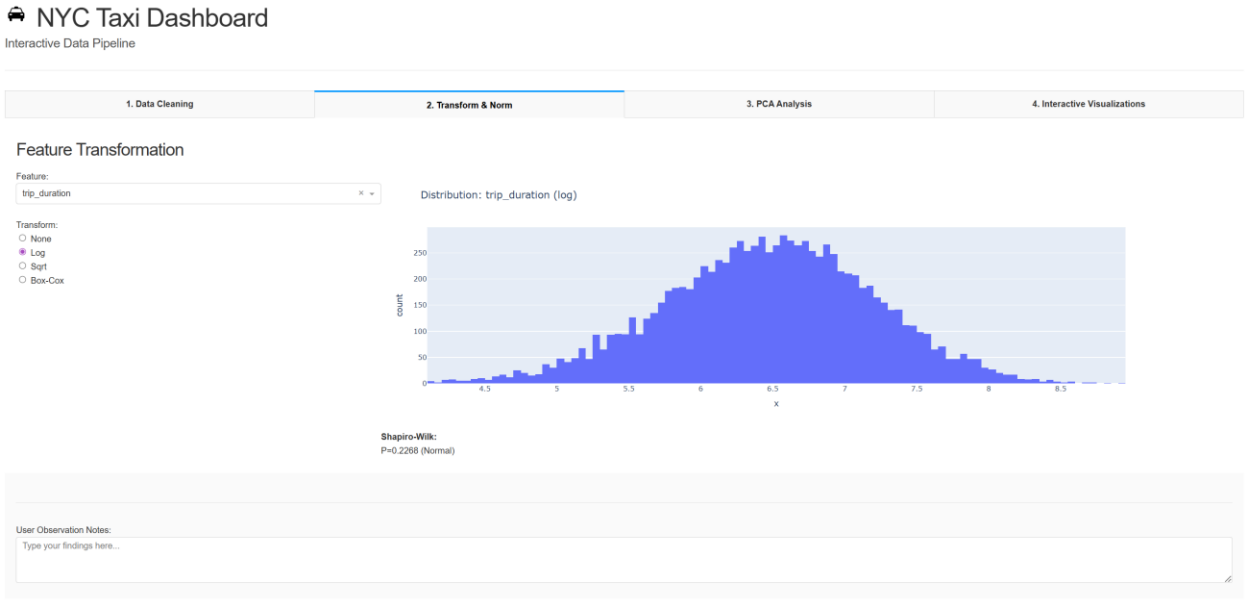
# 15. Dashboard

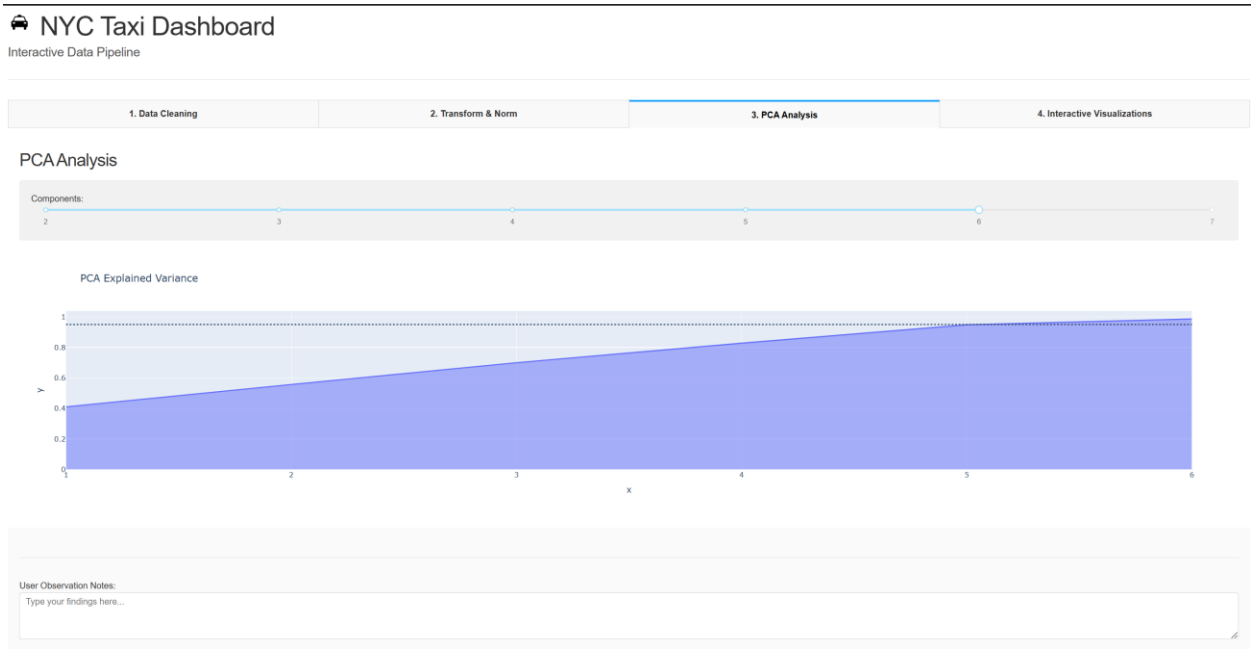
To facilitate interactive exploration of the dataset, a web-based dashboard was developed with the following modules:

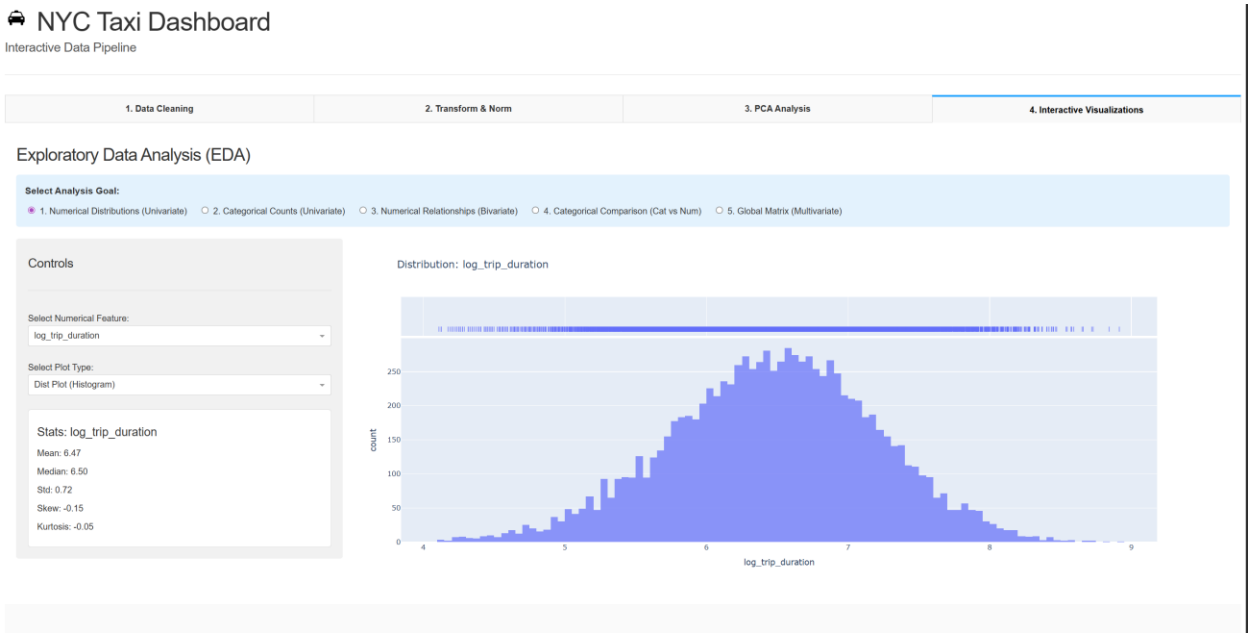
1. **Data Cleaning Tab:** Provides an interface to preprocess the raw data by removing duplicates, handling null values, and filtering outliers using Z-Score or IQR methods, with a real-time preview of the cleaned dataset.
2. **Transformation & Normalization Tab:** Allows users to apply statistical transformations (e.g., Log, Sqrt, Box-Cox) to numerical features to correct skewness and normalize distributions for modeling.
3. **PCA Analysis Tab:** Visualizes the dimensionality reduction process, displaying the explained variance ratio to help determine the optimal number of principal components to retain.

4. **Interactive Visualizations Tab:** Offers a dynamic environment where users can select specific features and plot types (univariate, bivariate, or multivariate) to generate custom visualizations on demand.






# 16. Conclusion

## 16.1 Learnings from Visual Analysis

The comprehensive visualization analysis revealed that **trip distance** is the definitive predictor of duration r=0.76, while **passenger count** r=0.01 and **vendor choice** are statistically irrelevant to operational speed. Temporal analysis uncovered a counter-intuitive insight: traffic congestion impacts trip efficiency more severely during the **2 PM mid-afternoon window** than the traditional morning rush, identifying a critical window for potential congestion pricing. Geospatially, the data confirmed that taxi operations are hyper-concentrated in **Midtown**

**Manhattan**, with a sharp efficiency drop-off for inter-borough travel, suggesting distinct market segments for taxis (short-hop) versus ride-shares (long-haul).

### 16.2 Dashboard Utility & Information Gain

The Python-based dashboard significantly lowers the barrier to entry for complex statistical analysis by automating the data pipeline from cleaning to visualization. By integrating **real-time statistical tables** alongside dynamic plots, users can instantly validate visual trends with hard metrics (e.g., skewness, R-squared) without writing code. The modular design—separating transformations, PCA, and plotting—allows stakeholders to iteratively test hypotheses, such as observing how log-transformations normalize duration distributions in real-time, effectively bridging the gap between raw data and actionable intelligence.

### 16.3 User Experience & Functionality

In terms of user experience, the dashboard functions as a high-utility decision support tool, offering a clean, intuitive interface that abstracts away the complexity of the underlying Scikit-Learn and Seaborn libraries. Its functionality extends beyond static reporting; the **interactive controls** for outlier removal and feature selection empower users to perform "what-if" scenarios. This responsiveness makes it a robust platform for both technical data scientists validating models and business analysts seeking actionable fleet insights, fulfilling the project objective of transforming static data into a dynamic exploration environment.

## 17. Appendix

App.py – python file containing all analysis methods

Vis.py – python file containing all visualization methods

## 18. References

NYC Taxi and Limousine Commission (TLC). (2016). *TLC Trip Record Data*. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.pag