OUTPUT SCREENSHOTS OF EACH TASK OF NYC RIDE SHARE DATA SET

TASK 1: Merging Datasets

#1}. Load rideshare_data.csv and taxi_zone_lookup.csv:

```
+-------+-------------+---------------+---------------+----------+---------------+
|business|pickup_location|dropoff_location|trip_length|request_to_pickup|total_ride_time|on_scene_to_pickup|on_scene_to_dropoff|time_of_day|
   date|passenger_fare|driver_total_pay|rideshare_profit|hourly_rate|dollars_per_mile|
+-------+-------------+---------------+---------------+----------+---------------+------------------+-------------------+-----------+---
-------+-------------+---------------+---------------+----------+---------------+
|    Uber|          151|            244|           4.98|     226.0|          761.0|              19.0|              780.0|    morning|168
4713600|        22.82|          13.69|           9.13|     63.18|           2.75|
|    Uber|          244|             78|           4.35|     197.0|         1423.0|             120.0|             1543.0|    morning|168
4713600|        24.27|           19.1|           5.17|     44.56|           4.39|
|    Uber|          151|            138|           8.82|     171.0|         1527.0|              12.0|             1539.0|    morning|168
4713600|        47.67|          25.94|          21.73|     60.68|           2.94|
|    Uber|          138|            151|           8.72|     260.0|         1761.0|              44.0|             1805.0|    morning|168
4713600|        45.67|          28.01|          17.66|     55.86|           3.21|
|    Uber|           36|            129|           5.05|     208.0|         1762.0|              37.0|             1799.0|    morning|168
4713600|        33.49|          26.47|           7.02|     52.97|           5.24|
|    Uber|          138|             88|          12.64|     230.0|         2504.0|              29.0|             2533.0|    morning|168
4713600|        69.15|          40.15|           29.0|     57.06|           3.18|
|    Uber|          200|            138|           14.3|     337.0|         1871.0|             120.0|             1991.0|    morning|168
4713600|        62.35|          37.48|          24.87|     67.77|           2.62|
|    Uber|          182|            242|           1.05|     177.0|          323.0|              30.0|              353.0|    morning|168
4713600|         10.3|           6.54|           4.76|      66.7|           6.23|
|    Uber|          248|            242|           0.57|     195.0|          169.0|               2.0|              171.0|    morning|168
4713600|          8.1|           5.54|           2.56|    116.63|           9.72|
|    Uber|          242|             20|           2.08|     308.0|          822.0|             120.0|              942.0|    morning|168
4713600|         14.8|          10.61|           4.19|     40.55|            5.1|
+-------+-------------+---------------+---------------+----------+---------------+------------------+-------------------+-----------+---
-------+-------------+---------------+---------------+----------+---------------+
only showing top 10 rows

root
 |-- business: string (nullable = true)
 |-- pickup_location: string (nullable = true)
 |-- dropoff_location: string (nullable = true)
 |-- trip_length: string (nullable = true)
 |-- request_to_pickup: string (nullable = true)
 |-- total_ride_time: string (nullable = true)
 |-- on_scene_to_pickup: string (nullable = true)
 |-- on_scene_to_dropoff: string (nullable = true)
 |-- time_of_day: string (nullable = true)
 |-- date: string (nullable = true)
 |-- passenger_fare: string (nullable = true)
 |-- driver_total_pay: string (nullable = true)
 |-- rideshare_profit: string (nullable = true)
 |-- hourly_rate: string (nullable = true)
```

```
+----------+-------------+--------------------+------------+
|LocationID|      Borough|                Zone|service_zone|
+----------+-------------+--------------------+------------+
|         1|          EWR|      Newark Airport|         EWR|
|         2|       Queens|         Jamaica Bay|   Boro Zone|
|         3|        Bronx|Allerton/Pelham G...|   Boro Zone|
|         4|    Manhattan|       Alphabet City| Yellow Zone|
|         5|Staten Island|       Arden Heights|   Boro Zone|
|         6|Staten Island|Arrochar/Fort Wad...|   Boro Zone|
|         7|       Queens|             Astoria|   Boro Zone|
|         8|       Queens|        Astoria Park|   Boro Zone|
|         9|       Queens|          Auburndale|   Boro Zone|
|        10|       Queens|        Baisley Park|   Boro Zone|
+----------+-------------+--------------------+------------+
only showing top 10 rows

root
 |-- LocationID: string (nullable = true)
 |-- Borough: string (nullable = true)
 |-- Zone: string (nullable = true)
 |-- service_zone: string (nullable = true)
```

OUTPUT SCREENSHOTS OF EACH TASK OF NYC RIDE SHARE DATA SET

#2} Joining the Datasets:

```
2024-04-03 17:42:20,126 INFO codegen.CodeGenerator: Code generated in 23.330603 ms
+--------+---------------+----------------+-----------+-----------------+---------------+------------------+---------
---------+-----------+----------+--------------+-----------------+----------------+-----------+-----------------+-------
-------+-----------+------------------+---------------+------------+--------------------+
|business|pickup_location|dropoff_location|trip_length|request_to_pickup|total_ride_time|on_scene_to_pickup|on_scene_t
o_dropoff|time_of_day|      date|passenger_fare|driver_total_pay|rideshare_profit|hourly_rate|dollars_per_mile|Pickup_
Borough|Pickup_Zone|Pickup_service_zone|Dropoff_Borough|Dropoff_Zone|Dropoff_service_zone|
+--------+---------------+----------------+-----------+-----------------+---------------+------------------+---------
---------+-----------+----------+--------------+-----------------+----------------+-----------+-----------------+-------
-------+-----------+------------------+---------------+------------+--------------------+
|    Uber|            169|             125|      12.52|             71.0|         1832.0|              17.0|
   1849.0|    morning|1677974400|         43.77|           39.64|            4.13|      77.18|             3.17|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|      12.75|            250.0|         1487.0|             121.0|
   1608.0|      night|1684281600|          53.7|           30.74|           22.96|      68.82|             2.41|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|      12.79|            270.0|         1754.0|              35.0|
   1789.0|    morning|1675987200|         53.57|            32.1|           21.47|      64.59|             2.51|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|       12.5|            170.0|         1961.0|              14.0|
   1975.0|    morning|1675468800|         42.87|           33.68|            9.19|      61.39|             2.69|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|      12.55|            164.0|         1988.0|              25.0|
   2013.0|    morning|1680912000|         43.72|           35.51|            8.21|      63.51|             2.83|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|      11.67|            781.0|         1758.0|             120.0|
   1878.0|    morning|1680566400|         46.19|           31.85|           14.34|      61.05|             2.73|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|      12.45|            326.0|         3980.0|              11.0|
   3991.0|    morning|1683676800|         56.53|           53.77|            2.76|       48.5|             4.32|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|      12.28|            346.0|         3559.0|             121.0|
   3680.0|    morning|1683676800|          54.3|            49.6|             4.7|      48.52|             4.04|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|      13.17|            146.0|         2678.0|               6.0|
   2684.0|    morning|1680307200|         59.07|           43.22|           15.85|      57.97|             3.28|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
|    Uber|            169|             125|      13.33|            279.0|         2904.0|             121.0|
   3025.0|  afternoon|1680307200|         52.28|           44.81|            7.47|      53.33|             3.36|
  Bronx| Mount Hope|          Boro Zone|      Manhattan|  Hudson Sq|         Yellow Zone|
+--------+---------------+----------------+-----------+-----------------+---------------+------------------+---------
```

#3} Convert the UNIX timestamp to the "yyyy-MM-dd" format:

```
ok 0.625438 s
2024-04-03 17:34:39,119 INFO codegen.CodeGenerator: Code generated in 16.83016 ms
+----------+
|      date|
+----------+
|2023-05-13|
|2023-05-13|
|2023-05-13|
|2023-05-14|
|2023-05-14|
|2023-01-27|
|2023-01-27|
|2023-01-27|
|2023-01-27|
|2023-01-27|
+----------+
only showing top 10 rows

2024-04-03 17:34:39,146 INFO server.AbstractConnector: Stopped Spark@48fad223{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-04-03 17:34:39,149 INFO ui.SparkUI: Stopped Spark web UI at http://task1-spark-app-cc81848ea5000bc8-driver-svc.da
ta-science-ec23863.svc:4040
```

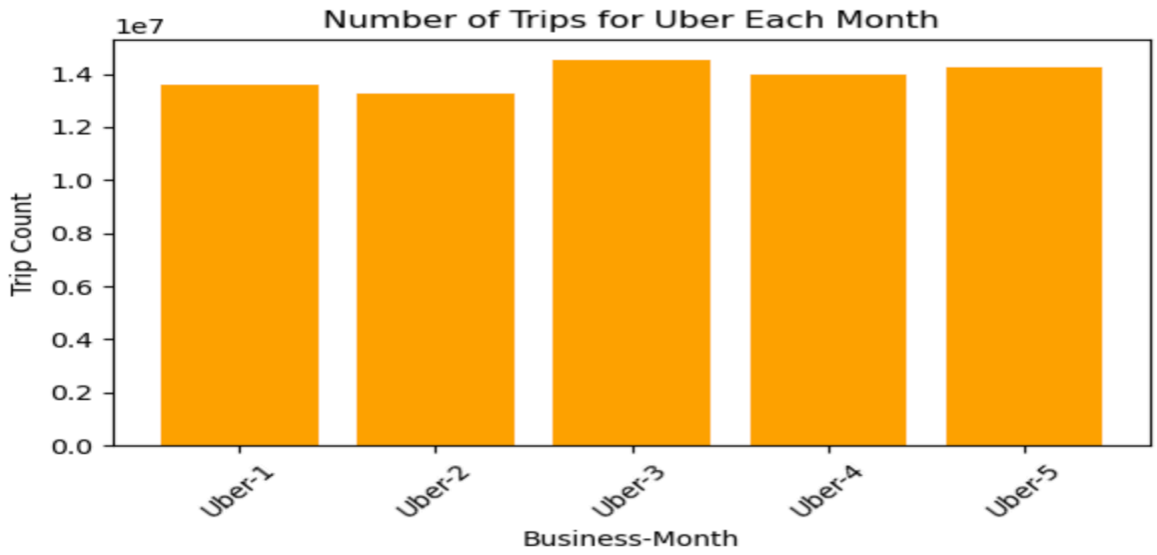
#4} Number of rows and Schema:

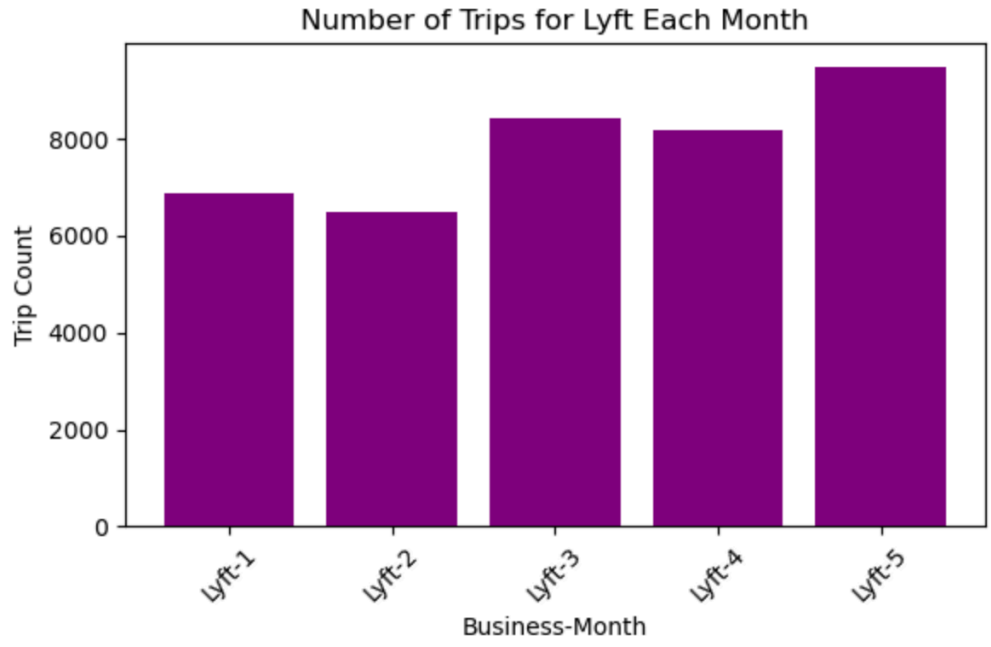
```
6.629486 s
Number of rows: 69725864
root
 |-- business: string (nullable = true)
 |-- pickup_location: string (nullable = true)
 |-- dropoff_location: string (nullable = true)
 |-- trip_length: string (nullable = true)
 |-- request_to_pickup: string (nullable = true)
 |-- total_ride_time: string (nullable = true)
 |-- on_scene_to_pickup: string (nullable = true)
 |-- on_scene_to_dropoff: string (nullable = true)
 |-- time_of_day: string (nullable = true)
 |-- date: date (nullable = true)
 |-- passenger_fare: string (nullable = true)
 |-- driver_total_pay: string (nullable = true)
 |-- rideshare_profit: string (nullable = true)
 |-- hourly_rate: string (nullable = true)
 |-- dollars_per_mile: string (nullable = true)
 |-- Pickup_Borough: string (nullable = true)
 |-- Pickup_Zone: string (nullable = true)
 |-- Pickup_service_zone: string (nullable = true)
 |-- Dropoff_Borough: string (nullable = true)
 |-- Dropoff_Zone: string (nullable = true)
 |-- Dropoff_service_zone: string (nullable = true)

2024-04-03 17:25:32,436 INFO server.AbstractConnector: Stopped Spark@711520cd{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-04-03 17:25:32,437 INFO ui.SparkUI: Stopped Spark web UI at http://task1-spark-app-d9d4e08ea4f1edce-driver-svc.da
ta-science-ec23863.svc:4040
```

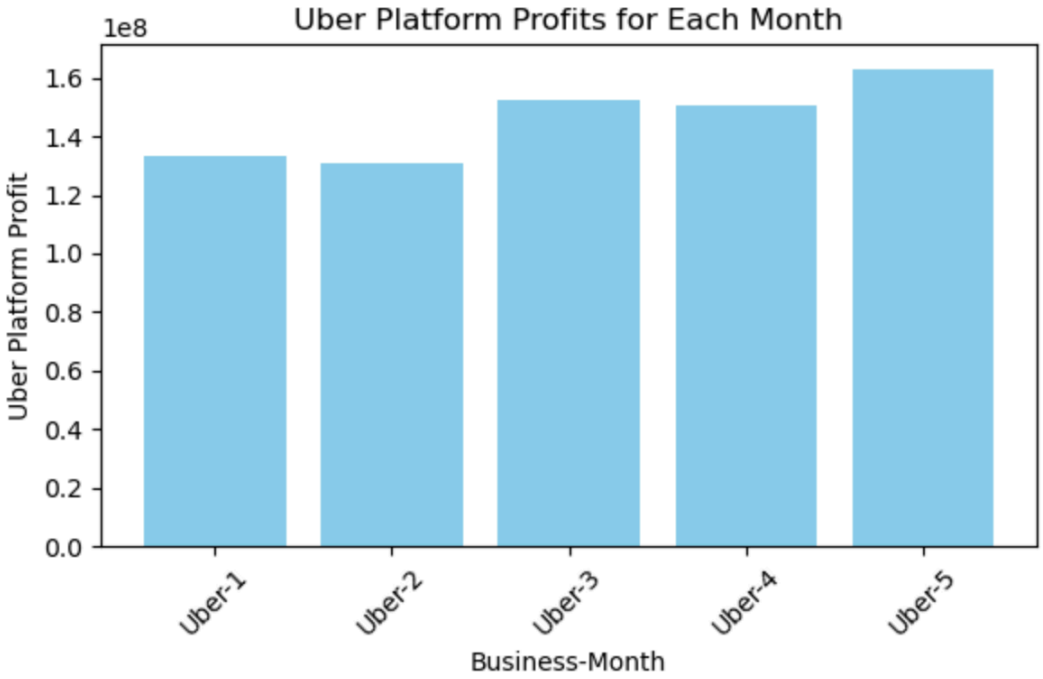
TASK 2: Aggregation of Data

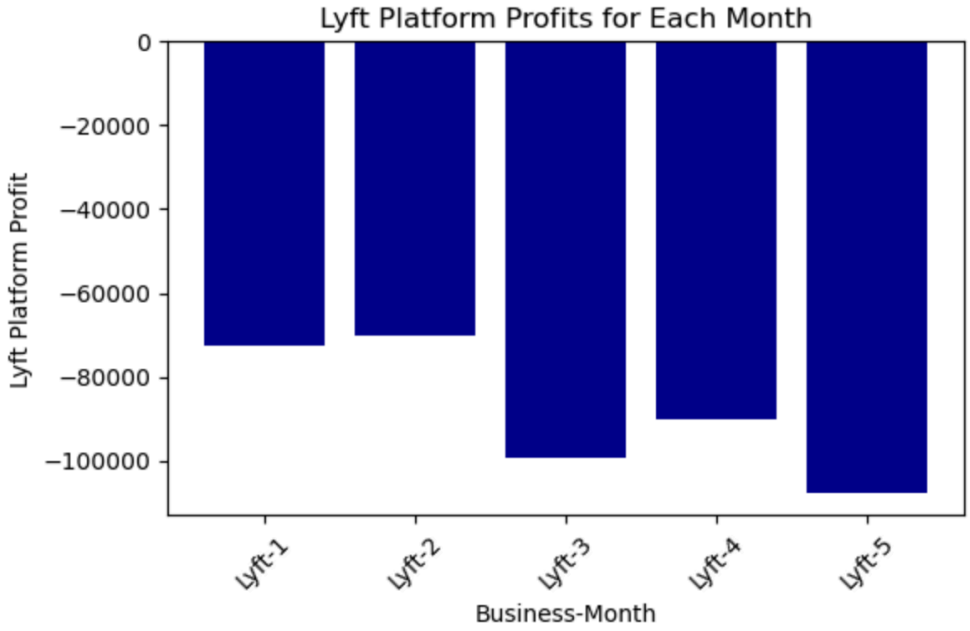
#1} Counting the number of trips for each business in each month



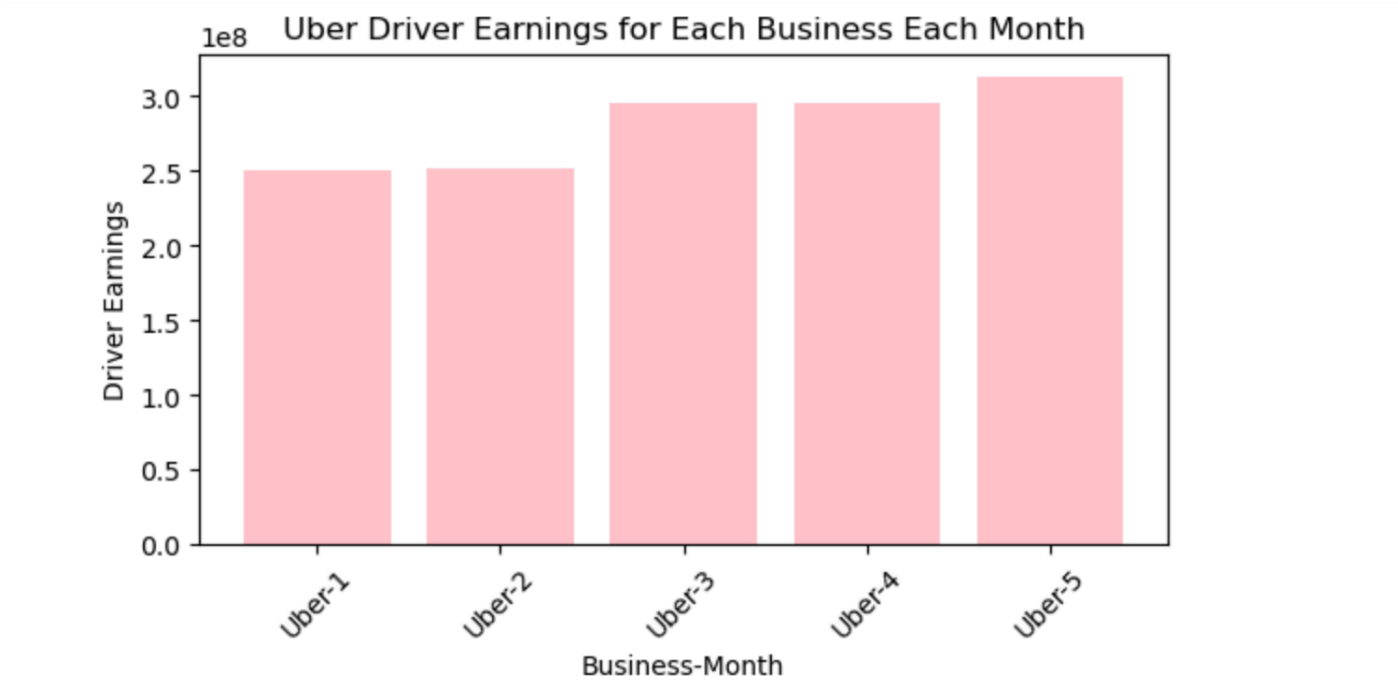


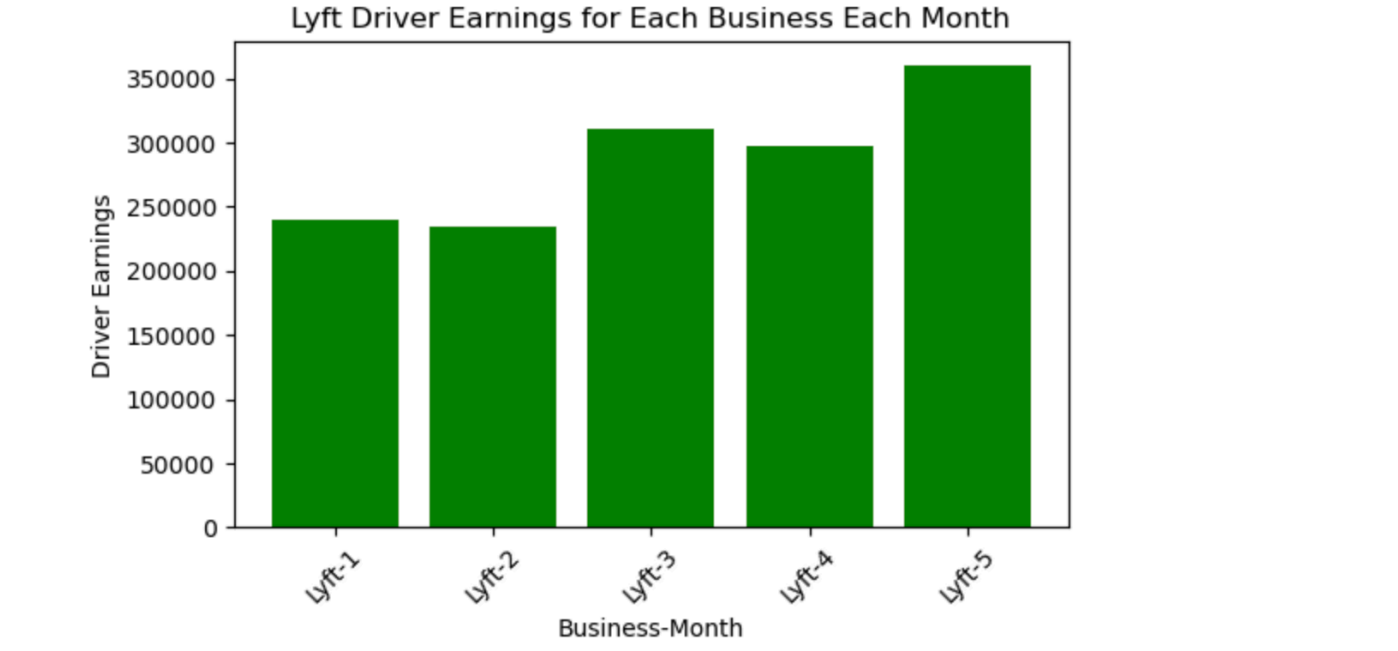
#2} Calculate the platform's profits (rideshare_profit field) for each business in each month





#3} Calculate the driver's earnings (driver_total_pay field) for each business in each month





TASK 3: Top-K Processing

#1}: Top 5 Popular Pickup Boroughs Each Month

```
2024-03-28 17:37:22,472 INFO codegen.CodeGenerator: Code generated in 17.193323 ms
2024-03-28 17:37:22,508 INFO codegen.CodeGenerator: Code generated in 18.438018 ms
+---------------+-----+-----------+
|Pickup_Borough|month|trip_count|
+---------------+-----+-----------+
|      Manhattan|    1|    5854818|
|       Brooklyn|    1|    3360373|
|         Queens|    1|    2589034|
|          Bronx|    1|    1607789|
| Staten Island|    1|     173354|
|      Manhattan|    2|    5808244|
|       Brooklyn|    2|    3283003|
|         Queens|    2|    2447213|
|          Bronx|    2|    1581889|
| Staten Island|    2|     166328|
|      Manhattan|    3|    6194298|
|       Brooklyn|    3|    3632776|
|         Queens|    3|    2757895|
|          Bronx|    3|    1785166|
| Staten Island|    3|     191935|
|      Manhattan|    4|    6002714|
|       Brooklyn|    4|    3481220|
|         Queens|    4|    2666671|
|          Bronx|    4|    1677435|
| Staten Island|    4|     175356|
|      Manhattan|    5|    5965594|
|       Brooklyn|    5|    3586009|
|         Queens|    5|    2826599|
|          Bronx|    5|    1717137|
| Staten Island|    5|     189924|
+---------------+-----+-----------+
```

#2} Top 5 Popular Dropoff Boroughs Each Month

```
2024-03-28 17:49:53,380 INFO codegen.CodeGenerator: Code generated in 13.133583 ms
2024-03-28 17:49:53,401 INFO codegen.CodeGenerator: Code generated in 10.008569 ms
+---------------+-----+----------+
|Dropoff_Borough|month|trip_count|
+---------------+-----+----------+
|      Manhattan|    1|   5444345|
|       Brooklyn|    1|   3337415|
|         Queens|    1|   2480080|
|          Bronx|    1|   1525137|
|        Unknown|    1|    535610|
|      Manhattan|    2|   5381696|
|       Brooklyn|    2|   3251795|
|         Queens|    2|   2390783|
|          Bronx|    2|   1511014|
|        Unknown|    2|    497525|
|      Manhattan|    3|   5671301|
|       Brooklyn|    3|   3608960|
|         Queens|    3|   2713748|
|          Bronx|    3|   1706802|
|        Unknown|    3|    566798|
|      Manhattan|    4|   5530417|
|       Brooklyn|    4|   3448225|
|         Queens|    4|   2605086|
|          Bronx|    4|   1596505|
|        Unknown|    4|    551857|
|      Manhattan|    5|   5428986|
|       Brooklyn|    5|   3560322|
|         Queens|    5|   2780011|
|          Bronx|    5|   1639180|
|        Unknown|    5|    578549|
+---------------+-----+----------+

2024-03-28 17:49:53,430 INFO server.AbstractConnector: Stopped Spark@d37a266{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-03-28 17:49:53,432 INFO ui.SparkUI: Stopped Spark web UI at http://task3-spark-app-c641f18e86279a5e-driver-svc.
40
```

3} Top 30 Earnest Routes

2024-04-03 15:14:28,854 INFO scheduler.DAGScheduler: Job 4 finished: showString at NativeMethodAccessorImpl.java:0, took 765.215750 s
2024-04-03 15:14:28,889 INFO codegen.CodeGenerator: Code generated in 14.056486 ms

Route	total_profit
Manhattan to Manhattan	3.338577255269214E8
Brooklyn to Brooklyn	1.7394472146560934E8
Queens to Queens	1.1470684718672623E8
Manhattan to Queens	1.0173842820749661E8
Queens to Manhattan	8.60354002623074E7
Manhattan to Unknown	8.010710241910338E7
Bronx to Bronx	7.41462257518282E7
Manhattan to Brooklyn	6.799047559133713E7
Brooklyn to Manhattan	6.31761610487396E7
Brooklyn to Queens	5.045416242985292E7
Queens to Brooklyn	4.729286535949615E7
Queens to Unknown	4.629299989943378E7
Bronx to Manhattan	3.2486325168083E7
Manhattan to Bronx	3.1978763449171744E7
Manhattan to EWR	2.375088861989542E7
Brooklyn to Unknown	1.0848827571691632E7
Bronx to Unknown	1.046480021000817 4E7
Bronx to Queens	1.0292266499867737E7
Queens to Bronx	1.0182898730611693E7
Staten Island to Staten Island	9686862.448563514
Brooklyn to Bronx	5848822.56057135
Bronx to Brooklyn	5629874.409598887
Brooklyn to EWR	3292761.709862232
Brooklyn to Staten Island	2417853.819514513
Staten Island to Brooklyn	2265856.459864326
Manhattan to Staten Island	2223727.3698619604
Staten Island to Manhattan	1612227.7201343775
Queens to EWR	1192758.6599292755
Staten Island to Unknown	891285.8100587726
Queens to Staten Island	865603.3800287247

2024-04-03 15:14:28,918 INFO server.AbstractConnector: Stopped Spark@6d9ce99f{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-04-03 15:14:28,920 INFO ui.SparkUI: Stopped Spark web UI at http://task3-spark-app-e8fa668ea47b201f-driver-svc.data-science-ec23

TASK 4: Average of Data

#1} Highest Average 'driver_total_pay' by Time of Day:

```
2024-04-03 15:31:22,307 INFO codegen.CodeGenerator: Code generated in 11.878683 ms
+-----------+-----------------------+
|time_of_day|average_drive_total_pay|
+-----------+-----------------------+
|  afternoon|      21.21242875569636|
|      night|      20.08743800270718|
|    evening|     19.777427701749232|
|    morning|     19.633332792748213|
+-----------+-----------------------+
```

#2} Highest Average 'trip_length' by Time of Day

```
2024-04-03 15:41:56,420 INFO codegen.CodeGenerator: Code generated in 18.0392 ms
2024-04-03 15:41:56,446 INFO codegen.CodeGenerator: Code generated in 10.512229 ms
+-----------+--------------------+
|time_of_day|average_trip_length|
+-----------+--------------------+
|      night|  5.323984802300154|
|    morning| 4.9273718666272845|
|  afternoon|   4.86141052588458|
|    evening|  4.484750367647451|
+-----------+--------------------+
```

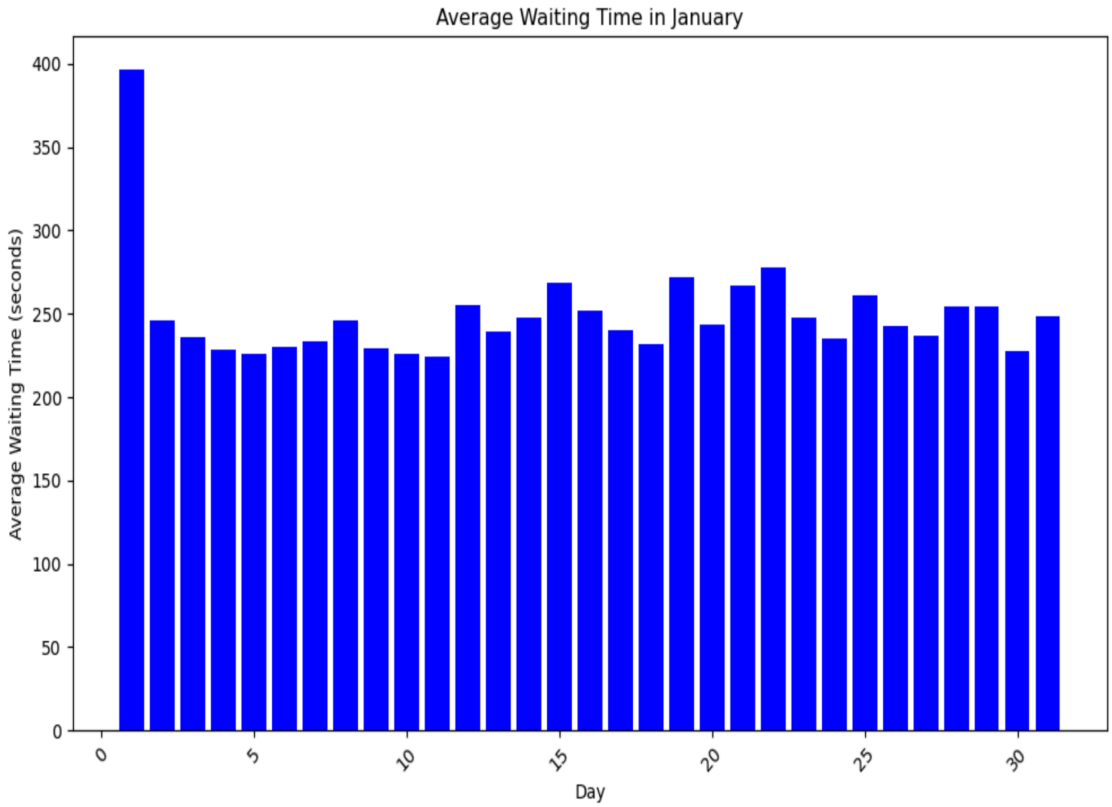
#3}: Calculating Average Earnings Per Mile by Time of Day

```
2024-04-03 15:59:34,434 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 48: Stage finished
2024-04-03 15:59:34,434 INFO scheduler.DAGScheduler: Job 9 finished: showString at NativeMethodAccessorImpl.java:0, took 0.101860 s
+----------+----------------------+
|time_of_day|average_earning_per_mile|
+----------+----------------------+
|afternoon  |4.363430869034982     |
|night      |3.773008141200681     |
|morning    |3.984544565374343     |
|evening    |4.4099283305536146    |
+----------+----------------------+

2024-04-03 15:59:34,499 INFO server.AbstractConnector: Stopped Spark@711520cd{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-04-03 15:59:34,501 INFO storage.BlockManagerInfo: Removed broadcast_14_piece0 on task4-spark-app-e229cc8ea4a5f222-driver-svc.data-science-ec23863.svc:7079 in memory (size: 26.2 KiB, free: 2004.2 MiB)
```

TASK 5: Finding Anomalies

#1}: Average Waiting Time in January



#2}: Identifying Days with Waiting Time Over 300 Seconds

```
2024-03-28 13:47:05,089 INFO codegen.CodeGenerator: Code generated in 15.821134 ms
+---+------------------+
|day|average_waiting_time|
+---+------------------+
|  1|  396.5318744409635|
+---+------------------+

2024-03-28 13:47:05,107 INFO server.AbstractConnector: Stopped Spark@c3c5d20{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-03-28 13:47:05,108 INFO ui.SparkUI: Stopped Spark web UI at http://task5-spark-app-9646978e8549a6d8-driver-svc.d
40
```

TASK 6: Filtering Data

#1} Finding trip counts greater than 0 and less than 1000 for different 'Pickup_Borough' at different 'time_of_day'

```
2024-03-28 13:32:07,923 INFO codegen.CodeGenerator: Code generated in 9.88417 ms
+--------------+-----------+----------+
|Pickup_Borough|time_of_day|trip_count|
+--------------+-----------+----------+
|EWR           |night      |3         |
|EWR           |afternoon  |2         |
|Unknown       |morning    |892       |
|Unknown       |afternoon  |908       |
|Unknown       |evening    |488       |
|EWR           |morning    |5         |
|Unknown       |night      |792       |
+--------------+-----------+----------+

2024-03-28 13:32:08,118 INFO codegen.CodeGenerator: Code generated in 14.120896 ms
```

#2} Calculate the number of trips for each 'Pickup_Borough' in the evening time.

```
2024-03-28 13:35:23,007 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 89: Stage finished
2024-03-28 13:35:23,008 INFO scheduler.DAGScheduler: Job 18 finished: showString at NativeMethodAccessorImpl.java:0, took 0.063474 s
+-------------+----------+---------+
|Pickup_Borough|time_of_day|trip_count|
+-------------+----------+---------+
|Bronx        |evening   |1380355  |
|Queens       |evening   |2223003  |
|Manhattan    |evening   |5724796  |
|Staten Island|evening   |151276   |
|Brooklyn     |evening   |3075616  |
|Unknown      |evening   |488      |
+-------------+----------+---------+

2024-03-28 13:35:23,033 INFO server.AbstractConnector: Stopped Spark@3568868{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-03-28 13:35:23,035 INFO ui.SparkUI: Stopped Spark web UI at http://task6-spark-app-a680a78e8539309e-driver-svc.data-science-ec238
```

#3} The number of trips that started in Brooklyn (Pickup_Borough field) and ended in Staten Island (Dropoff_Borough field)

```
2024-03-28 13:12:37,148 INFO scheduler.DAGScheduler: Job 7 finished: showString at NativeMetho
2024-03-28 13:12:37,182 INFO codegen.CodeGenerator: Code generated in 13.555765 ms
+--------------+---------------+--------------------+
|Pickup_Borough|Dropoff_Borough|         Pickup_Zone|
+--------------+---------------+--------------------+
|      Brooklyn|  Staten Island|     Columbia Street|
|      Brooklyn|  Staten Island|     Columbia Street|
|      Brooklyn|  Staten Island|     Columbia Street|
|      Brooklyn|  Staten Island|     Columbia Street|
|      Brooklyn|  Staten Island|     Columbia Street|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
+--------------+---------------+--------------------+
only showing top 10 rows

2024-03-28 13:12:37,435 INFO codegen.CodeGenerator: Code generated in 9.445144 ms
```

```
2024-03-28 13:26:04,172 INFO scheduler.DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.ja
2024-03-28 13:26:04,206 INFO codegen.CodeGenerator: Code generated in 14.281588 ms
+--------------+---------------+--------------------+
|Pickup_Borough|Dropoff_Borough|         Pickup_Zone|
+--------------+---------------+--------------------+
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
|      Brooklyn|  Staten Island|Marine Park/Mill ...|
+--------------+---------------+--------------------+
only showing top 10 rows

2024-03-28 13:26:04,472 INFO codegen.CodeGenerator: Code generated in 9.490304 ms
2024-03-28 13:26:04,491 INFO codegen.CodeGenerator: Code generated in 14.891281 ms
```

```
2024-03-28 13:16:25,117 INFO scheduler.DAGScheduler: Job 8 finished: count at NativeMethodAccessorImpl.java:0, took 227.54
Count of trips from Brooklyn to Staten Island:  69437
2024-03-28 13:16:25,159 INFO server.AbstractConnector: Stopped Spark@3568868{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
```

TASK 7: Route Analysis

Comparing the volume of trips between Uber and Lyft across different routes in New York City, as defined by pickup and dropoff service zones

```
2024-04-05 10:17:34,511 INFO codegen.CodeGenerator: Code generated in 12.050992 ms
```

Route	uber_count	lyft_count	total_count
JFK Airport to NA	253211	46	253257
East New York to East New York	202719	184	202903
Borough Park to Borough Park	155803	78	155881
LaGuardia Airport to NA	151521	41	151562
Canarsie to Canarsie	126253	26	126279
South Ozone Park to JFK Airport	107392	1770	109162
Crown Heights North to Crown Heights North	98591	100	98691
Bay Ridge to Bay Ridge	98274	300	98574
Astoria to Astoria	90692	75	90767
Jackson Heights to Jackson Heights	89652	19	89671

```
2024-04-05 10:17:34,539 INFO server.AbstractConnector: Stopped Spark@2989a773{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-04-05 10:17:34,542 INFO ui.SparkUI: Stopped Spark web UI at http://new-spark-app-2fcc3c8eadb590ed-driver-svc.data-science-ec23863.svc
```

TASK 8: Graphs

#1}Defining the StructType for vertexSchema and edgeSchema:

```
#1} Define the StructType of vertexSchema and edgeSchema.
    |
    # Schema for vertices dataframe based on the taxi zone lookup data
    vertexSchema = StructType([
        StructField("id", IntegerType(), False),  # False for nullable indicates this field cannot be null
        StructField("Borough", StringType(), True),
        StructField("Zone", StringType(), True),
        StructField("service_zone", StringType(), True)
    ])

    # Schema for edges dataframe based on the rideshare data
    edgeSchema = StructType([
        StructField("src", IntegerType(), False),
        StructField("dst", IntegerType(), False)
    ])
```

#2) Constructing edges and vertices DataFrames:

```
2024-04-03 09:46:52,782 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 6: Stage finished
2024-04-03 09:46:52,784 INFO scheduler.DAGScheduler: Job 6 finished: showString at NativeMethodAccessorImpl.java:0, took 0.942926 s
+---+-------------+--------------------+------------+
| id|      Borough|                Zone|service_zone|
+---+-------------+--------------------+------------+
|  1|          EWR|      Newark Airport|         EWR|
|  2|       Queens|         Jamaica Bay|   Boro Zone|
|  3|        Bronx|Allerton/Pelham G...|   Boro Zone|
|  4|    Manhattan|       Alphabet City| Yellow Zone|
|  5|Staten Island|       Arden Heights|   Boro Zone|
|  6|Staten Island|Arrochar/Fort Wad...|   Boro Zone|
|  7|       Queens|             Astoria|   Boro Zone|
|  8|       Queens|        Astoria Park|   Boro Zone|
|  9|       Queens|          Auburndale|   Boro Zone|
| 10|       Queens|        Baisley Park|   Boro Zone|
+---+-------------+--------------------+------------+
only showing top 10 rows

2024-04-03 09:46:52,854 INFO codegen.CodeGenerator: Code generated in 11.224382 ms
2024-04-03 09:46:52,860 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
2024-04-03 09:46:52,861 INFO scheduler.DAGScheduler: Got job 7 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
```

```
2024-04-03 09:46:53,025 INFO codegen.CodeGenerator: Code generated in 12.746302 ms
+---+---+
|src|dst|
+---+---+
|151|244|
|244| 78|
|151|138|
|138|151|
| 36|129|
|138| 88|
|200|138|
|182|242|
|248|242|
|242| 20|
+---+---+
only showing top 10 rows
```

#3) Creating a graph using the vertices and edges:

Printing the graphframe using show() method:

```
2024-04-03 18:02:42,084 INFO codegen.CodeGenerator: Code generated in 11.290808 ms
+----------------------------------+----------+---------------------------------------+
|src                               |edge      |dst                                    |
+----------------------------------+----------+---------------------------------------+
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
|[169, Bronx, Mount Hope, Boro Zone]|[169, 125]|[125, Manhattan, Hudson Sq, Yellow Zone]|
+----------------------------------+----------+---------------------------------------+
only showing top 10 rows

2024-04-03 18:02:42,107 INFO server.AbstractConnector: Stopped Spark@1578e326{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-04-03 18:02:42,109 INFO ui.SparkUI: Stopped Spark web UI at http://task8-spark-app-7a70f08ea51c0e08-driver-svc.c
ta-science-ec23863.svc:4040
```

OUTPUT SCREENSHOTS OF EACH TASK OF NYC RIDE SHARE DATA SET

Printing the output distinct() method to show unique triplet combinations:

```
2024-04-05 09:40:41,172 INFO codegen.CodeGenerator: Code generated in 15.548746 ms
+--------------------------------------------------+---------+--------------------------------------------------+
|src                                               |edge     |dst                                               |
+--------------------------------------------------+---------+--------------------------------------------------+
|[66, Brooklyn, DUMBO/Vinegar Hill, Boro Zone]     |[66, 124]|[124, Queens, Howard Beach, Boro Zone]            |
|[133, Brooklyn, Kensington, Boro Zone]            |[133, 124]|[124, Queens, Howard Beach, Boro Zone]           |
|[65, Brooklyn, Downtown Brooklyn/MetroTech, Boro Zone]|[65, 124]|[124, Queens, Howard Beach, Boro Zone]          |
|[133, Brooklyn, Kensington, Boro Zone]            |[133, 7] |[7, Queens, Astoria, Boro Zone]                   |
|[93, Queens, Flushing Meadows-Corona Park, Boro Zone]|[93, 7]  |[7, Queens, Astoria, Boro Zone]                   |
|[256, Brooklyn, Williamsburg (South Side), Boro Zone]|[256, 234]|[234, Manhattan, Union Sq, Yellow Zone]           |
|[34, Brooklyn, Brooklyn Navy Yard, Boro Zone]     |[34, 234]|[234, Manhattan, Union Sq, Yellow Zone]           |
|[223, Queens, Steinway, Boro Zone]                |[223, 200]|[200, Bronx, Riverdale/North Riverdale/Fieldston, Boro Zone]|
|[47, Bronx, Claremont/Bathgate, Boro Zone]        |[47, 200]|[200, Bronx, Riverdale/North Riverdale/Fieldston, Boro Zone]|
|[230, Manhattan, Times Sq/Theatre District, Yellow Zone]|[230, 200]|[200, Bronx, Riverdale/North Riverdale/Fieldston, Boro Zone]|
+--------------------------------------------------+---------+--------------------------------------------------+
only showing top 10 rows

2024-04-05 09:40:41,210 INFO server.AbstractConnector: Stopped Spark@7cb4aaf1{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-04-05 09:40:41,213 INFO ui.SparkUI: Stopped Spark web UI at http://task8-spark-app-2b581c8ead94c404-driver-svc.data-science-ec23863.svc:
40
```

#4) Counting connected vertices within the same Borough and service_zone:

Printing the graphframe using show() method:

```
272.523033 s
2024-04-04 13:07:07,440 INFO codegen.CodeGenerator: Code generated in 15.837765 ms
+---+---+---------+------------+
|id |id |Borough  |service_zone|
+---+---+---------+------------+
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
|125|249|Manhattan|Yellow Zone |
+---+---+---------+------------+
only showing top 10 rows

2024-04-04 13:07:07,465 INFO server.AbstractConnector: Stopped Spark@76a8489f{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2024-04-04 13:07:07,467 INFO ui.SparkUI: Stopped Spark web UI at http://task8-spark-app-2acac18ea9337ba7-driver-svc.dat
science-ec23863.svc:4040
2024-04-04 13:07:07,471 INFO k8s.KubernetesClusterSchedulerBackend: Shutting down all executors
```

OUTPUT SCREENSHOTS OF EACH TASK OF NYC RIDE SHARE DATA SET

Printing the output distinct() method to show unique combinations:

```
2024-04-05 09:52:44,041 INFO codegen.CodeGenerator: Code generated in 13.018027 ms
+---+---+-------------+------------+
|id |id |Borough      |service_zone|
+---+---+-------------+------------+
|252|19 |Queens       |Boro Zone   |
|206|245|Staten Island|Boro Zone   |
|131|207|Queens       |Boro Zone   |
|111|178|Brooklyn     |Boro Zone   |
|186|90 |Manhattan    |Yellow Zone |
|64 |95 |Queens       |Boro Zone   |
|121|95 |Queens       |Boro Zone   |
|144|158|Manhattan    |Yellow Zone |
|37 |65 |Brooklyn     |Boro Zone   |
|164|68 |Manhattan    |Yellow Zone |
+---+---+-------------+------------+
only showing top 10 rows

2024-04-05 09:52:44,213 INFO codegen.CodeGenerator: Code generated in 9.495648 ms
2024-04-05 09:52:44,232 INFO codegen.CodeGenerator: Code generated in 13.520469 ms
2024-04-05 09:52:44,301 INFO spark.SparkContext: Starting job: count at NativeMethodAccessorImpl.java:0
2024-04-05 09:52:44,302 INFO scheduler.DAGScheduler: Registering RDD 40 (count at NativeMethodAccessorImpl.java:0) as input to shuffle
```

Total connected Vertices Output:

```
Total connected vertices with the same Borough and service zone: 4688699
```

#5) Performing page ranking on the graph DataFrame:

```
2024-04-04 18:00:59,349 INFO scheduler.DAGScheduler: Job 40 finished: showString at NativeMethodAccessorImpl.java:0, toc
2024-04-04 18:00:59,366 INFO codegen.CodeGenerator: Code generated in 11.955213 ms
+---+------------------+
|id |pagerank          |
+---+------------------+
|265|11.105433344107194|
|1  |5.4718454249167205|
|132|4.551132572067087 |
|138|3.5683223416564713|
|61 |2.6763973653412996|
+---+------------------+
only showing top 5 rows
```