Importing Necessary Packages

In [0]:
```
from pyspark.sql.functions import hour, avg
from pyspark.sql.functions import month, count, round, sum
from pyspark.sql.window import Window
```

2021 Files

In [0]:
```
# File location and type
file1 = "/FileStore/tables/fhvhv_tripdata_2021_01.parquet"
file2 = "/FileStore/tables/fhvhv_tripdata_2021_02.parquet"
file3 = "/FileStore/tables/fhvhv_tripdata_2021_03.parquet"
file4 = "/FileStore/tables/fhvhv_tripdata_2021_04.parquet"
file5 = "/FileStore/tables/fhvhv_tripdata_2021_05.parquet"
file6 = "/FileStore/tables/fhvhv_tripdata_2021_06.parquet"
file7 = "/FileStore/tables/fhvhv_tripdata_2021_07.parquet"
file8 = "/FileStore/tables/fhvhv_tripdata_2021_08.parquet"
file9 = "/FileStore/tables/fhvhv_tripdata_2021_09.parquet"
file10 = "/FileStore/tables/fhvhv_tripdata_2021_10.parquet"
file11 = "/FileStore/tables/fhvhv_tripdata_2021_11.parquet"
file12 = "/FileStore/tables/fhvhv_tripdata_2021_12.parquet"



file_type='parquet'
df1=spark.read.parquet(file1, header=True, inferSchema=True)
df2=spark.read.parquet(file2, header=True, inferSchema=True)
df3=spark.read.parquet(file3, header=True, inferSchema=True)
df4=spark.read.parquet(file4, header=True, inferSchema=True)
df5=spark.read.parquet(file5, header=True, inferSchema=True)
df6=spark.read.parquet(file6, header=True, inferSchema=True)
df7=spark.read.parquet(file7, header=True, inferSchema=True)
df8=spark.read.parquet(file8, header=True, inferSchema=True)
df9=spark.read.parquet(file9, header=True, inferSchema=True)
df10=spark.read.parquet(file10, header=True, inferSchema=True)
df11=spark.read.parquet(file11, header=True, inferSchema=True)
df12=spark.read.parquet(file12, header=True, inferSchema=True)
```

In [0]:
```
df1.show(truncate=False)
```

```
+-----------------+-------------------+-------------------+---------------
----+-------------------+-------------------+-------------------+----------
-+-----------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+---------------+--------------+
|hvfhs_license_num|dispatching_base_num|originating_base_num|request_datetim
e    |on_scene_datetime   |pickup_datetime     |dropoff_datetime    |PULocationI
D|DOLocationID|trip_miles|trip_time|base_passenger_fare|tolls|bcf  |sales_tax
|congestion_surcharge|airport_fee|tips|driver_pay|shared_request_flag|shared
_match_flag|access_a_ride_flag|wav_request_flag|wav_match_flag|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|HV0003           |B02682              |B02682              |2021-01-01 00:2
8:09|2021-01-01 00:31:42|2021-01-01 00:33:44|2021-01-01 00:49:07|230
|166         |5.26      |923      |22.28              |0.0  |0.67|1.98     |
2.75                |null       |0.0 |14.99     |N                  |N
|                  |N                |N              |
|HV0003           |B02682              |B02682              |2021-01-01 00:4
5:56|2021-01-01 00:55:19|2021-01-01 00:55:19|2021-01-01 01:18:21|152
|167         |3.65      |1382     |18.36              |0.0  |0.55|1.63     |
0.0                 |null       |0.0 |17.06     |N                  |N
|                  |N                |N              |
|HV0003           |B02764              |B02764              |2021-01-01 00:2
1:15|2021-01-01 00:22:41|2021-01-01 00:23:56|2021-01-01 00:38:05|233
|142         |3.51      |849      |14.05              |0.0  |0.48|1.25     |
2.75                |null       |0.94|12.98     |N                  |N
|                  |N                |N              |
|HV0003           |B02764              |B02764              |2021-01-01 00:3
9:12|2021-01-01 00:42:37|2021-01-01 00:42:51|2021-01-01 00:45:50|142
|143         |0.74      |179      |7.91               |0.0  |0.24|0.7      |
2.75                |null       |0.0 |7.41      |N                  |N
|                  |N                |N              |
|HV0003           |B02764              |B02764              |2021-01-01 00:4
6:11|2021-01-01 00:47:17|2021-01-01 00:48:14|2021-01-01 01:08:42|143
|78          |9.2       |1228     |27.11              |0.0  |0.81|2.41     |
2.75                |null       |0.0 |22.44     |N                  |N
|                  |N                |N              |
|HV0005           |B02510              |null                |2021-01-01 00:0
4:00|null               |2021-01-01 00:06:59|2021-01-01 00:43:01|88
|42          |9.725     |2162     |28.11              |0.0  |0.84|2.49     |
2.75                |null       |0.0 |28.9      |N                  |N
|N                 |N                |N              |
|HV0005           |B02510              |null                |2021-01-01 00:4
0:06|null               |2021-01-01 00:50:00|2021-01-01 01:04:57|42
|151         |2.469     |897      |25.03              |0.0  |0.75|2.22     |
0.0                 |null       |0.0 |15.01     |N                  |N
|N                 |N                |N              |
|HV0003           |B02764              |B02764              |2021-01-01 00:1
0:36|2021-01-01 00:12:28|2021-01-01 00:14:30|2021-01-01 00:50:27|71
|226         |13.53     |2157     |29.67              |0.0  |1.04|3.08     |
0.0                 |null       |0.0 |34.2      |N                  |N
|                  |N                |N              |
|HV0003           |B02875              |B02875              |2021-01-01 00:2
1:17|2021-01-01 00:22:25|2021-01-01 00:22:54|2021-01-01 00:30:20|112
|255         |1.6       |446      |6.89               |0.0  |0.21|0.61     |
0.0                 |null       |0.0 |6.26      |N                  |N
|                  |N                |N              |
|HV0003           |B02875              |B02875              |2021-01-01 00:3
6:57|2021-01-01 00:38:09|2021-01-01 00:40:12|2021-01-01 00:53:31|255
|232         |3.2       |800      |11.51              |0.0  |0.53|1.03     |
2.75                |null       |2.82|10.99     |N                  |N
|                  |N                |N              |
|HV0003           |B02875              |B02875              |2021-01-01 00:5
3:31|2021-01-01 00:56:21|2021-01-01 00:56:45|2021-01-01 01:17:42|232
```

```
|198          |5.74       |1257      |17.18              |0.0  |0.52|1.52     |
2.75              |null       |0.0 |17.61     |N                   |N
|                  |N                |N              |
|HV0003           |B02835              |B02835              |2021-01-01 00:2
2:58|2021-01-01 00:27:01|2021-01-01 00:29:04|2021-01-01 00:36:27|113
|48           |1.8        |443       |8.18               |0.0  |0.25|0.73     |
2.75              |null       |0.0 |6.12      |N                   |N
|                  |N                |N              |
|HV0003           |B02835              |B02835              |2021-01-01 00:4
6:44|2021-01-01 00:47:49|2021-01-01 00:48:56|2021-01-01 00:59:12|239
|75           |2.9        |616       |13.1               |0.0  |0.45|1.17     |
2.75              |null       |0.94|8.77      |N                   |N
|                  |N                |N              |
|HV0004           |B02800              |null                |2021-01-01 00:1
2:50|null               |2021-01-01 00:15:24|2021-01-01 00:38:31|181
|237          |9.66       |1387      |32.95              |0.0  |0.0 |2.34     |
2.75              |null       |0.0 |21.1      |N                   |N
|N                 |N                |N              |
|HV0004           |B02800              |null                |2021-01-01 00:3
5:32|null               |2021-01-01 00:45:00|2021-01-01 01:06:45|236
|68           |4.38       |1305      |22.91              |0.0  |0.0 |1.63     |
2.75              |null       |3.43|15.82     |N                   |N
|N                 |N                |N              |
|HV0003           |B02682              |B02682              |2021-01-01 00:1
0:22|2021-01-01 00:11:03|2021-01-01 00:11:53|2021-01-01 00:18:06|256
|148          |2.03       |373       |7.84               |0.0  |0.42|0.7      |
2.75              |null       |2.82|6.93      |N                   |N
|                  |N                |N              |
|HV0003           |B02682              |B02682              |2021-01-01 00:2
5:00|2021-01-01 00:26:31|2021-01-01 00:28:31|2021-01-01 00:41:40|79
|80           |3.08       |789       |13.2               |0.0  |0.4 |1.17     |
2.75              |null       |0.0 |11.54     |N                   |N
|                  |N                |N              |
|HV0003           |B02682              |B02682              |2021-01-01 00:4
4:56|2021-01-01 00:49:55|2021-01-01 00:50:49|2021-01-01 00:55:59|17
|217          |1.17       |310       |7.91               |0.0  |0.24|0.7      |
0.0               |null       |0.0 |6.94      |N                   |N
|                  |N                |N              |
|HV0005           |B02510              |null                |2021-01-01 00:0
5:04|null               |2021-01-01 00:08:40|2021-01-01 00:39:39|62
|29           |10.852     |1859      |31.18              |0.0  |0.94|2.77     |
0.0               |null       |0.0 |27.61     |N                   |N
|N                 |N                |N              |
|HV0003           |B02836              |B02836              |2021-01-01 00:4
0:44|2021-01-01 00:53:34|2021-01-01 00:53:48|2021-01-01 01:11:40|22
|22           |3.52       |1072      |28.67              |0.0  |0.86|2.54     |
0.0               |null       |0.0 |17.64     |N                   |N
|                  |N                |N              |
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
only showing top 20 rows
```

Joining 2021 12 Months Files

In [0]:
```
#combining all the 12 datasets into one dataframe
df_2021=df1.union(df2).union(df3).union(df4).union(df5).union(df6).union(df7
```

In [0]:
```
#combined dataframe with rows from all the 12 datasets
df_2021.show()
```

```
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|hvfhs_license_num|dispatching_base_num|originating_base_num|   request_date
time|  on_scene_datetime|    pickup_datetime|   dropoff_datetime|PULocationI
D|DOLocationID|trip_miles|trip_time|base_passenger_fare|tolls| bcf|sales_tax
|congestion_surcharge|airport_fee|tips|driver_pay|shared_request_flag|shared
_match_flag|access_a_ride_flag|wav_request_flag|wav_match_flag|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|           HV0003|              B02682|              B02682|2021-01-01 00:2
8:09|2021-01-01 00:31:42|2021-01-01 00:33:44|2021-01-01 00:49:07|         23
0|         166|      5.26|      923|              22.28|  0.0|0.67|     1.98
|                2.75|       null| 0.0|     14.99|                  N|
N|                  |               N|             N|
|           HV0003|              B02682|              B02682|2021-01-01 00:4
5:56|2021-01-01 00:55:19|2021-01-01 00:55:19|2021-01-01 01:18:21|         15
2|         167|      3.65|     1382|              18.36|  0.0|0.55|     1.63
|                 0.0|       null| 0.0|     17.06|                  N|
N|                  |               N|             N|
|           HV0003|              B02764|              B02764|2021-01-01 00:2
1:15|2021-01-01 00:22:41|2021-01-01 00:23:56|2021-01-01 00:38:05|         23
3|         142|      3.51|      849|              14.05|  0.0|0.48|     1.25
|                2.75|       null|0.94|     12.98|                  N|
N|                  |               N|             N|
|           HV0003|              B02764|              B02764|2021-01-01 00:3
9:12|2021-01-01 00:42:37|2021-01-01 00:42:51|2021-01-01 00:45:50|         14
2|         143|      0.74|      179|               7.91|  0.0|0.24|      0.7
|                2.75|       null| 0.0|      7.41|                  N|
N|                  |               N|             N|
|           HV0003|              B02764|              B02764|2021-01-01 00:4
6:11|2021-01-01 00:47:17|2021-01-01 00:48:14|2021-01-01 01:08:42|         14
3|          78|       9.2|     1228|              27.11|  0.0|0.81|     2.41
|                2.75|       null| 0.0|     22.44|                  N|
N|                  |               N|             N|
|           HV0005|              B02510|                null|2021-01-01 00:0
4:00|               null|2021-01-01 00:06:59|2021-01-01 00:43:01|          8
8|          42|     9.725|     2162|              28.11|  0.0|0.84|     2.49
|                2.75|       null| 0.0|      28.9|                  N|
N|                 N|               N|             N|
|           HV0005|              B02510|                null|2021-01-01 00:4
0:06|               null|2021-01-01 00:50:00|2021-01-01 01:04:57|          4
```

```
2|         151|     2.469|      897|               25.03|  0.0|0.75|     2.22
|                0.0|       null| 0.0|     15.01|                 N|
N|                 N|              N|             N|
|           HV0003|              B02764|              B02764|2021-01-01 00:1
0:36|2021-01-01 00:12:28|2021-01-01 00:14:30|2021-01-01 00:50:27|         7
1|         226|     13.53|     2157|               29.67|  0.0|1.04|     3.08
|                0.0|       null| 0.0|      34.2|                 N|
N|                  |              N|             N|
|           HV0003|              B02875|              B02875|2021-01-01 00:2
1:17|2021-01-01 00:22:25|2021-01-01 00:22:54|2021-01-01 00:30:20|        11
2|         255|       1.6|      446|                6.89|  0.0|0.21|     0.61
|                0.0|       null| 0.0|      6.26|                 N|
N|                  |              N|             N|
|           HV0003|              B02875|              B02875|2021-01-01 00:3
6:57|2021-01-01 00:38:09|2021-01-01 00:40:12|2021-01-01 00:53:31|        25
5|         232|       3.2|      800|               11.51|  0.0|0.53|     1.03
|               2.75|       null|2.82|     10.99|                 N|
N|                  |              N|             N|
|           HV0003|              B02875|              B02875|2021-01-01 00:5
3:31|2021-01-01 00:56:21|2021-01-01 00:56:45|2021-01-01 01:17:42|        23
2|         198|      5.74|     1257|               17.18|  0.0|0.52|     1.52
|               2.75|       null| 0.0|     17.61|                 N|
N|                  |              N|             N|
|           HV0003|              B02835|              B02835|2021-01-01 00:2
2:58|2021-01-01 00:27:01|2021-01-01 00:29:04|2021-01-01 00:36:27|        11
3|          48|       1.8|      443|                8.18|  0.0|0.25|     0.73
|               2.75|       null| 0.0|      6.12|                 N|
N|                  |              N|             N|
|           HV0003|              B02835|              B02835|2021-01-01 00:4
6:44|2021-01-01 00:47:49|2021-01-01 00:48:56|2021-01-01 00:59:12|        23
9|          75|       2.9|      616|                13.1|  0.0|0.45|     1.17
|               2.75|       null|0.94|      8.77|                 N|
N|                  |              N|             N|
|           HV0004|              B02800|                null|2021-01-01 00:1
2:50|                null|2021-01-01 00:15:24|2021-01-01 00:38:31|        18
1|         237|      9.66|     1387|               32.95|  0.0| 0.0|     2.34
|               2.75|       null| 0.0|      21.1|                 N|
N|                 N|              N|             N|
|           HV0004|              B02800|                null|2021-01-01 00:3
5:32|                null|2021-01-01 00:45:00|2021-01-01 01:06:45|        23
6|          68|      4.38|     1305|               22.91|  0.0| 0.0|     1.63
|               2.75|       null|3.43|     15.82|                 N|
N|                 N|              N|             N|
|           HV0003|              B02682|              B02682|2021-01-01 00:1
0:22|2021-01-01 00:11:03|2021-01-01 00:11:53|2021-01-01 00:18:06|        25
6|         148|      2.03|      373|                7.84|  0.0|0.42|      0.7
|               2.75|       null|2.82|      6.93|                 N|
N|                  |              N|             N|
|           HV0003|              B02682|              B02682|2021-01-01 00:2
5:00|2021-01-01 00:26:31|2021-01-01 00:28:31|2021-01-01 00:41:40|         7
9|          80|      3.08|      789|                13.2|  0.0| 0.4|     1.17
|               2.75|       null| 0.0|     11.54|                 N|
N|                  |              N|             N|
|           HV0003|              B02682|              B02682|2021-01-01 00:4
4:56|2021-01-01 00:49:55|2021-01-01 00:50:49|2021-01-01 00:55:59|         1
7|         217|      1.17|      310|                7.91|  0.0|0.24|      0.7
|                0.0|       null| 0.0|      6.94|                 N|
N|                  |               N|              N|
|           HV0005|              B02510|                null|2021-01-01 00:0
5:04|              null|2021-01-01 00:08:40|2021-01-01 00:39:39|            6
2|          29|    10.852|     1859|              31.18|  0.0|0.94|      2.77
|                0.0|       null| 0.0|     27.61|                 N|
N|                N|               N|              N|
|           HV0003|              B02836|              B02836|2021-01-01 00:4
0:44|2021-01-01 00:53:34|2021-01-01 00:53:48|2021-01-01 01:11:40|            2
2|          22|      3.52|     1072|              28.67|  0.0|0.86|      2.54
|                0.0|       null| 0.0|     17.64|                 N|
N|                  |               N|              N|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
only showing top 20 rows
```

In [0]:
```
print((df_2021.count(), len(df_2021.columns)))
```

```
(174596652, 24)
```

Columns Summary

In [0]:
```
df_2021.printSchema()
```

```
root
 |-- hvfhs_license_num: string (nullable = true)
 |-- dispatching_base_num: string (nullable = true)
 |-- originating_base_num: string (nullable = true)
 |-- request_datetime: timestamp (nullable = true)
 |-- on_scene_datetime: timestamp (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- PULocationID: long (nullable = true)
 |-- DOLocationID: long (nullable = true)
 |-- trip_miles: double (nullable = true)
 |-- trip_time: long (nullable = true)
 |-- base_passenger_fare: double (nullable = true)
 |-- tolls: double (nullable = true)
 |-- bcf: double (nullable = true)
 |-- sales_tax: double (nullable = true)
 |-- congestion_surcharge: double (nullable = true)
 |-- airport_fee: double (nullable = true)
 |-- tips: double (nullable = true)
 |-- driver_pay: double (nullable = true)
 |-- shared_request_flag: string (nullable = true)
 |-- shared_match_flag: string (nullable = true)
 |-- access_a_ride_flag: string (nullable = true)
 |-- wav_request_flag: string (nullable = true)
 |-- wav_match_flag: string (nullable = true)
```

2022 Files

In [0]:
```
# File locations and type
file2_1  = "/FileStore/tables/fhvhv_tripdata_2022_01.parquet"
file2_2  = "/FileStore/tables/fhvhv_tripdata_2022_02.parquet"
file2_3  = "/FileStore/tables/fhvhv_tripdata_2022_03.parquet"
file2_4  = "/FileStore/tables/fhvhv_tripdata_2022_04.parquet"
file2_5  = "/FileStore/tables/fhvhv_tripdata_2022_05.parquet"
file2_6  = "/FileStore/tables/fhvhv_tripdata_2022_06.parquet"
file2_7  = "/FileStore/tables/fhvhv_tripdata_2022_07.parquet"
file2_8  = "/FileStore/tables/fhvhv_tripdata_2022_08.parquet"
file2_9  = "/FileStore/tables/fhvhv_tripdata_2022_09.parquet"
file2_10  = "/FileStore/tables/fhvhv_tripdata_2022_10.parquet"
file2_11 = "/FileStore/tables/fhvhv_tripdata_2022_11.parquet"
file2_12 = "/FileStore/tables/fhvhv_tripdata_2022_12.parquet"
file_type = 'parquet'

# Read Parquet files into DataFrames
df1 = spark.read.parquet(file2_1, header=True, inferSchema=True)
df2 = spark.read.parquet(file2_2, header=True, inferSchema=True)
df3 = spark.read.parquet(file2_3, header=True, inferSchema=True)
df4 = spark.read.parquet(file2_4, header=True, inferSchema=True)
df5 = spark.read.parquet(file2_5, header=True, inferSchema=True)
df6 = spark.read.parquet(file2_6, header=True, inferSchema=True)
df7 = spark.read.parquet(file2_7, header=True, inferSchema=True)
df8 = spark.read.parquet(file2_8, header=True, inferSchema=True)
df9 = spark.read.parquet(file2_9, header=True, inferSchema=True)
df10 = spark.read.parquet(file2_10, header=True, inferSchema=True)
df11 = spark.read.parquet(file2_11, header=True, inferSchema=True)
df12 = spark.read.parquet(file2_12, header=True, inferSchema=True)
```

Combining all 2022 12 Months and Columns Summary

In [0]:
```
#combining all the 12 datasets into one dataframe
df_2022=df1.union(df2).union(df3).union(df4).union(df5).union(df6).union(df7

#combined dataframe with rows from all the 12 datasets
df_2022.show()

# shape of first dataframe is 11908468 rows and 24 columns.
print((df_2022.count(), len(df_2022.columns)))

df_2022.printSchema()
```

```
+------------------+--------------------+--------------------+----------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+-------------------+----------------+---------------+
|hvfhs_license_num|dispatching_base_num|originating_base_num|    request_date
time|   on_scene_datetime|     pickup_datetime|    dropoff_datetime|PULocationI
D|DOLocationID|trip_miles|trip_time|base_passenger_fare|tolls|  bcf|sales_tax
|congestion_surcharge|airport_fee|tips|driver_pay|shared_request_flag|shared
```

```
_match_flag|access_a_ride_flag|wav_request_flag|wav_match_flag|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|           HV0003|              B03404|              B03404|2022-01-01 00:0
5:31|2022-01-01 00:05:40|2022-01-01 00:07:24|2022-01-01 00:18:28|         17
0|         161|      1.18|      664|               24.9|  0.0|0.75|     2.21
|                2.75|        0.0| 0.0|     23.03|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:1
9:27|2022-01-01 00:22:08|2022-01-01 00:22:32|2022-01-01 00:30:12|         23
7|         161|      0.82|      460|              11.97|  0.0|0.36|     1.06
|                2.75|        0.0| 0.0|     12.32|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
3:53|2022-01-01 00:57:37|2022-01-01 00:57:37|2022-01-01 01:07:32|         23
7|         161|      1.18|      595|              29.82|  0.0|0.89|     2.65
|                2.75|        0.0| 0.0|      23.3|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:1
5:36|2022-01-01 00:17:08|2022-01-01 00:18:02|2022-01-01 00:23:05|         26
2|         229|      1.65|      303|               7.91|  0.0|0.24|      0.7
|                2.75|        0.0| 0.0|       6.3|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:2
5:45|2022-01-01 00:26:01|2022-01-01 00:28:01|2022-01-01 00:35:42|         22
9|         141|      1.65|      461|               9.44|  0.0|0.28|     0.84
|                2.75|        0.0| 0.0|      7.44|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:3
4:44|2022-01-01 00:36:52|2022-01-01 00:38:50|2022-01-01 00:51:32|         26
3|          79|      4.51|      762|              17.67|  0.0|0.53|     1.57
|                2.75|        0.0| 0.0|     12.25|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
7:51|2022-01-01 00:52:00|2022-01-01 00:53:25|2022-01-01 01:08:56|         11
3|         140|      3.68|      931|              16.68|  0.0| 0.5|     1.48
|                2.75|        0.0| 0.0|     12.75|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:0
6:21|2022-01-01 00:06:58|2022-01-01 00:08:58|2022-01-01 00:23:01|         15
1|          75|      2.77|      843|              14.41|  0.0|0.43|     1.28
|                 0.0|        0.0| 4.0|     11.47|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:2
7:54|2022-01-01 00:30:26|2022-01-01 00:32:25|2022-01-01 00:44:15|         26
3|         229|      2.04|      710|              10.64|  0.0|0.32|     0.94
|                2.75|        0.0| 0.0|      9.55|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
4:59|2022-01-01 00:48:23|2022-01-01 00:50:23|2022-01-01 01:15:30|         23
7|         169|      8.79|     1507|             107.56|  0.0|0.83|     2.45
|                2.75|        0.0| 0.0|     23.67|                  N|
N|                  |               N|             N|
|           HV0003|              B03404|              B03404|2022-01-01 00:1
```

```
3:49|2022-01-01 00:16:15|2022-01-01 00:17:02|2022-01-01 00:40:09|        26
1|         223|     11.29|     1387|                34.9|  0.0|1.05|      3.1
|               2.75|          0.0| 0.0|      25.17|                 N|
N|                   |              N|             N|
|           HV0003|             B03404|                B03404|2022-01-01 00:3
9:10|2022-01-01 00:42:59|2022-01-01 00:43:20|2022-01-01 00:47:31|        22
3|         223|      0.87|      251|                7.91|  0.0|0.24|      0.7
|                0.0|          0.0| 0.0|       6.51|                 N|
N|                   |              N|             N|
|           HV0003|             B03404|                B03404|2022-01-01 00:4
5:50|2022-01-01 00:52:15|2022-01-01 00:52:29|2022-01-01 01:01:48|        22
3|           7|      1.89|      559|                9.71|  0.0|0.29|     0.86
|                0.0|          0.0| 0.0|       7.89|                 N|
N|                   |              N|             N|
|           HV0005|             B03406|                  null|2022-01-01 00:3
6:54|              null|2022-01-01 00:45:34|2022-01-01 00:54:11|         8
8|         148|     3.585|      810|               27.02|  0.0|0.81|      2.4
|               2.75|          0.0| 0.0|      21.08|                 N|
N|                  N|              N|             N|
|           HV0003|             B03404|                B03404|2022-01-01 00:0
7:13|2022-01-01 00:12:03|2022-01-01 00:12:03|2022-01-01 00:35:07|        24
6|         243|       9.2|     1384|               30.37|  0.0|0.91|      2.7
|               2.75|          0.0| 0.0|      22.69|                 N|
N|                   |              N|             N|
|           HV0003|             B03404|                B03404|2022-01-01 00:5
3:32|2022-01-01 00:58:14|2022-01-01 00:58:26|2022-01-01 01:07:23|        24
3|         127|      1.88|      537|               15.02|  0.0|0.45|     1.33
|                0.0|          0.0| 2.0|      10.02|                 N|
N|                   |              N|             N|
|           HV0005|             B03406|                  null|2022-01-01 00:2
8:07|              null|2022-01-01 00:34:59|2022-01-01 00:50:15|        23
9|         170|     2.699|      916|               20.47|  0.0|0.61|     1.82
|               2.75|          0.0| 0.0|      10.66|                 N|
N|                  N|              N|             N|
|           HV0005|             B03406|                  null|2022-01-01 00:4
5:44|              null|2022-01-01 00:58:09|2022-01-01 01:28:23|        17
0|         265|     8.037|     1904|               24.63| 20.0|1.34|      0.0
|                0.0|          0.0| 0.0|      26.63|                 N|
N|                  N|              N|             N|
|           HV0003|             B03404|                B03404|2022-01-01 00:2
1:54|2022-01-01 00:26:15|2022-01-01 00:28:15|2022-01-01 00:38:52|        22
3|         179|      1.69|      637|                9.81|  0.0|0.29|     0.87
|                0.0|          0.0| 0.0|       9.55|                 N|
N|                   |              N|             N|
|           HV0003|             B03404|                B03404|2022-01-01 00:3
5:08|2022-01-01 00:43:55|2022-01-01 00:45:55|2022-01-01 00:53:47|        17
9|           7|      0.87|      472|               10.53|  0.0|0.32|     0.93
|                0.0|          0.0| 0.0|       9.22|                 N|
N|                   |              N|             N|
+-----------------+-------------------+-------------------+---------------
----+-------------------+-------------------+-------------------+----------
-+------------+----------+---------+-------------------+-----+----+---------
+-------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
only showing top 20 rows

(212416083, 24)
```

```
root
 |-- hvfhs_license_num: string (nullable = true)
 |-- dispatching_base_num: string (nullable = true)
 |-- originating_base_num: string (nullable = true)
 |-- request_datetime: timestamp (nullable = true)
 |-- on_scene_datetime: timestamp (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- PULocationID: long (nullable = true)
 |-- DOLocationID: long (nullable = true)
 |-- trip_miles: double (nullable = true)
 |-- trip_time: long (nullable = true)
 |-- base_passenger_fare: double (nullable = true)
 |-- tolls: double (nullable = true)
 |-- bcf: double (nullable = true)
 |-- sales_tax: double (nullable = true)
 |-- congestion_surcharge: double (nullable = true)
 |-- airport_fee: double (nullable = true)
 |-- tips: double (nullable = true)
 |-- driver_pay: double (nullable = true)
 |-- shared_request_flag: string (nullable = true)
 |-- shared_match_flag: string (nullable = true)
 |-- access_a_ride_flag: string (nullable = true)
 |-- wav_request_flag: string (nullable = true)
 |-- wav_match_flag: string (nullable = true)
```

Data Preprocessing

In [0]:

In [0]:

```python
import pandas as pd

# Check the data types of each variable for 2021
print(df_2021.dtypes)
df_2021.show()

# Get descriptive statistics for numerical variables for 2021
print(df_2021[['trip_miles', 'trip_time', 'base_passenger_fare', 'tolls', 'b

# Get descriptive statistics for numerical variables for 2022
print(df_2022[['trip_miles', 'trip_time', 'base_passenger_fare', 'tolls', 'b

#df_2021.show()
```

```
[('hvfhs_license_num', 'string'), ('dispatching_base_num', 'string'), ('orig
inating_base_num', 'string'), ('request_datetime', 'timestamp'), ('on_scene_
datetime', 'timestamp'), ('pickup_datetime', 'timestamp'), ('dropoff_datetim
e', 'timestamp'), ('PULocationID', 'bigint'), ('DOLocationID', 'bigint'), ('
trip_miles', 'double'), ('trip_time', 'bigint'), ('base_passenger_fare', 'do
uble'), ('tolls', 'double'), ('bcf', 'double'), ('sales_tax', 'double'), ('c
ongestion_surcharge', 'double'), ('airport_fee', 'double'), ('tips', 'doubl
e'), ('driver_pay', 'double'), ('shared_request_flag', 'string'), ('shared_m
atch_flag', 'string'), ('access_a_ride_flag', 'string'), ('wav_request_fla
g', 'string'), ('wav_match_flag', 'string')]
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|hvfhs_license_num|dispatching_base_num|originating_base_num|   request_date
time|  on_scene_datetime|    pickup_datetime|   dropoff_datetime|PULocationI
D|DOLocationID|trip_miles|trip_time|base_passenger_fare|tolls| bcf|sales_tax
|congestion_surcharge|airport_fee|tips|driver_pay|shared_request_flag|shared
_match_flag|access_a_ride_flag|wav_request_flag|wav_match_flag|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|           HV0003|              B02682|              B02682|2021-01-01 00:2
8:09|2021-01-01 00:31:42|2021-01-01 00:33:44|2021-01-01 00:49:07|         23
0|         166|      5.26|      923|              22.28|  0.0|0.67|     1.98
|                2.75|       null| 0.0|     14.99|                  N|
N|                  |               N|             N|
|           HV0003|              B02682|              B02682|2021-01-01 00:4
5:56|2021-01-01 00:55:19|2021-01-01 00:55:19|2021-01-01 01:18:21|         15
2|         167|      3.65|     1382|              18.36|  0.0|0.55|     1.63
|                 0.0|       null| 0.0|     17.06|                  N|
N|                  |               N|             N|
|           HV0003|              B02764|              B02764|2021-01-01 00:2
1:15|2021-01-01 00:22:41|2021-01-01 00:23:56|2021-01-01 00:38:05|         23
3|         142|      3.51|      849|              14.05|  0.0|0.48|     1.25
|                2.75|       null|0.94|     12.98|                  N|
N|                  |               N|             N|
|           HV0003|              B02764|              B02764|2021-01-01 00:3
9:12|2021-01-01 00:42:37|2021-01-01 00:42:51|2021-01-01 00:45:50|         14
2|         143|      0.74|      179|               7.91|  0.0|0.24|      0.7
|                2.75|       null| 0.0|      7.41|                  N|
N|                  |               N|             N|
|           HV0003|              B02764|              B02764|2021-01-01 00:4
6:11|2021-01-01 00:47:17|2021-01-01 00:48:14|2021-01-01 01:08:42|         14
3|          78|       9.2|     1228|              27.11|  0.0|0.81|     2.41
|                2.75|       null| 0.0|     22.44|                  N|
N|                  |               N|             N|
|           HV0005|              B02510|                null|2021-01-01 00:0
4:00|               null|2021-01-01 00:06:59|2021-01-01 00:43:01|          8
8|          42|     9.725|     2162|              28.11|  0.0|0.84|     2.49
|                2.75|       null| 0.0|      28.9|                  N|
N|                 N|               N|             N|
|           HV0005|              B02510|                null|2021-01-01 00:4
0:06|               null|2021-01-01 00:50:00|2021-01-01 01:04:57|          4
2|         151|     2.469|      897|              25.03|  0.0|0.75|     2.22
|                 0.0|       null| 0.0|     15.01|                  N|
N|                 N|               N|             N|
|           HV0003|              B02764|              B02764|2021-01-01 00:1
0:36|2021-01-01 00:12:28|2021-01-01 00:14:30|2021-01-01 00:50:27|          7
1|         226|     13.53|     2157|              29.67|  0.0|1.04|     3.08
|                 0.0|       null| 0.0|      34.2|                  N|
N|                  |               N|             N|
|           HV0003|              B02875|              B02875|2021-01-01 00:2
```

```
1:17|2021-01-01 00:22:25|2021-01-01 00:22:54|2021-01-01 00:30:20|         11
2|         255|       1.6|      446|               6.89|  0.0|0.21|     0.61
|                 0.0|       null| 0.0|      6.26|                  N|
N|                  |               N|             N|
|           HV0003|              B02875|              B02875|2021-01-01 00:3
6:57|2021-01-01 00:38:09|2021-01-01 00:40:12|2021-01-01 00:53:31|         25
5|         232|       3.2|      800|              11.51|  0.0|0.53|     1.03
|                2.75|       null|2.82|     10.99|                  N|
N|                  |               N|             N|
|           HV0003|              B02875|              B02875|2021-01-01 00:5
3:31|2021-01-01 00:56:21|2021-01-01 00:56:45|2021-01-01 01:17:42|         23
2|         198|      5.74|     1257|              17.18|  0.0|0.52|     1.52
|                2.75|       null| 0.0|     17.61|                  N|
N|                  |               N|             N|
|           HV0003|              B02835|              B02835|2021-01-01 00:2
2:58|2021-01-01 00:27:01|2021-01-01 00:29:04|2021-01-01 00:36:27|         11
3|          48|       1.8|      443|               8.18|  0.0|0.25|     0.73
|                2.75|       null| 0.0|      6.12|                  N|
N|                  |               N|             N|
|           HV0003|              B02835|              B02835|2021-01-01 00:4
6:44|2021-01-01 00:47:49|2021-01-01 00:48:56|2021-01-01 00:59:12|         23
9|          75|       2.9|      616|               13.1|  0.0|0.45|     1.17
|                2.75|       null|0.94|      8.77|                  N|
N|                  |               N|             N|
|           HV0004|              B02800|                null|2021-01-01 00:1
2:50|               null|2021-01-01 00:15:24|2021-01-01 00:38:31|         18
1|         237|      9.66|     1387|              32.95|  0.0| 0.0|     2.34
|                2.75|       null| 0.0|      21.1|                  N|
N|                 N|               N|             N|
|           HV0004|              B02800|                null|2021-01-01 00:3
5:32|               null|2021-01-01 00:45:00|2021-01-01 01:06:45|         23
6|          68|      4.38|     1305|              22.91|  0.0| 0.0|     1.63
|                2.75|       null|3.43|     15.82|                  N|
N|                 N|               N|             N|
|           HV0003|              B02682|              B02682|2021-01-01 00:1
0:22|2021-01-01 00:11:03|2021-01-01 00:11:53|2021-01-01 00:18:06|         25
6|         148|      2.03|      373|               7.84|  0.0|0.42|      0.7
|                2.75|       null|2.82|      6.93|                  N|
N|                  |               N|             N|
|           HV0003|              B02682|              B02682|2021-01-01 00:2
5:00|2021-01-01 00:26:31|2021-01-01 00:28:31|2021-01-01 00:41:40|          7
9|          80|      3.08|      789|               13.2|  0.0| 0.4|     1.17
|                2.75|       null| 0.0|     11.54|                  N|
N|                  |               N|             N|
|           HV0003|              B02682|              B02682|2021-01-01 00:4
4:56|2021-01-01 00:49:55|2021-01-01 00:50:49|2021-01-01 00:55:59|          1
7|         217|      1.17|      310|               7.91|  0.0|0.24|      0.7
|                 0.0|       null| 0.0|      6.94|                  N|
N|                  |               N|             N|
|           HV0005|              B02510|                null|2021-01-01 00:0
5:04|               null|2021-01-01 00:08:40|2021-01-01 00:39:39|          6
2|          29|    10.852|     1859|              31.18|  0.0|0.94|     2.77
|                 0.0|       null| 0.0|     27.61|                  N|
N|                 N|               N|             N|
|           HV0003|              B02836|              B02836|2021-01-01 00:4
0:44|2021-01-01 00:53:34|2021-01-01 00:53:48|2021-01-01 01:11:40|          2
2|          22|      3.52|     1072|              28.67|  0.0|0.86|     2.54
```

```
|                  0.0|          null| 0.0|      17.64|                   N|
N|                  |             N|            N|
+-----------------+--------------------+--------------------+---------------
----+--------------------+--------------------+--------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
only showing top 20 rows

DataFrame[summary: string, trip_miles: string, trip_time: string, base_passe
nger_fare: string, tolls: string, bcf: string, sales_tax: string, congestion
_surcharge: string, airport_fee: string, tips: string, driver_pay: string]
DataFrame[summary: string, trip_miles: string, trip_time: string, base_passe
nger_fare: string, tolls: string, bcf: string, sales_tax: string, congestion
_surcharge: string, airport_fee: string, tips: string, driver_pay: string]
```

In [0]:

```python
from pyspark.sql import functions as F
from pyspark.sql.types import StringType, BooleanType, IntegerType, DoubleTy

# Convert columns to appropriate data types
df_2021 = df_2021.withColumn("request_datetime", F.col("request_datetime").c
df_2021 = df_2021.withColumn("trip_miles", F.col("trip_miles").cast("double"
df_2021 = df_2021.withColumn("trip_time", F.col("trip_time").cast("double"))

# Convert boolean columns to BooleanType
boolean_columns = ["shared_request_flag", "shared_match_flag", "access_a_rid
for col_name in boolean_columns:
    df_2021 = df_2021.withColumn(col_name, F.col(col_name).cast(BooleanType(

# Convert integer columns to IntegerType
integer_columns = ["PULocationID", "DOLocationID"]
for col_name in integer_columns:
    df_2021 = df_2021.withColumn(col_name, F.col(col_name).cast(IntegerType(

# Convert other categorical columns to StringType
categorical_columns = ["hvfhs_license_num", "dispatching_base_num", "origina
for col_name in categorical_columns:
    df_2021 = df_2021.withColumn(col_name, F.col(col_name).cast(StringType()

# Print the updated schema
df_2021.printSchema()

df_2021.show()
```

```
root
 |-- hvfhs_license_num: string (nullable = true)
 |-- dispatching_base_num: string (nullable = true)
 |-- originating_base_num: string (nullable = true)
 |-- request_datetime: timestamp (nullable = true)
 |-- on_scene_datetime: timestamp (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- PULocationID: integer (nullable = true)
 |-- DOLocationID: integer (nullable = true)
 |-- trip_miles: double (nullable = true)
 |-- trip_time: double (nullable = true)
 |-- base_passenger_fare: double (nullable = true)
 |-- tolls: double (nullable = true)
 |-- bcf: double (nullable = true)
 |-- sales_tax: double (nullable = true)
 |-- congestion_surcharge: double (nullable = true)
 |-- airport_fee: double (nullable = true)
 |-- tips: double (nullable = true)
 |-- driver_pay: double (nullable = true)
 |-- shared_request_flag: boolean (nullable = true)
 |-- shared_match_flag: boolean (nullable = true)
 |-- access_a_ride_flag: boolean (nullable = true)
 |-- wav_request_flag: boolean (nullable = true)
 |-- wav_match_flag: boolean (nullable = true)

+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|hvfhs_license_num|dispatching_base_num|originating_base_num|   request_date
time|  on_scene_datetime|    pickup_datetime|   dropoff_datetime|PULocationI
D|DOLocationID|trip_miles|trip_time|base_passenger_fare|tolls| bcf|sales_tax
|congestion_surcharge|airport_fee|tips|driver_pay|shared_request_flag|shared
_match_flag|access_a_ride_flag|wav_request_flag|wav_match_flag|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|           HV0003|              B02682|              B02682|2021-01-01 00:2
8:09|2021-01-01 00:31:42|2021-01-01 00:33:44|2021-01-01 00:49:07|         23
0|         166|      5.26|    923.0|              22.28|  0.0|0.67|     1.98
|                2.75|       null| 0.0|     14.99|              false|
false|             null|            false|         false|
|           HV0003|              B02682|              B02682|2021-01-01 00:4
5:56|2021-01-01 00:55:19|2021-01-01 00:55:19|2021-01-01 01:18:21|         15
2|         167|      3.65|   1382.0|              18.36|  0.0|0.55|     1.63
|                 0.0|       null| 0.0|     17.06|              false|
false|             null|            false|         false|
|           HV0003|              B02764|              B02764|2021-01-01 00:2
1:15|2021-01-01 00:22:41|2021-01-01 00:23:56|2021-01-01 00:38:05|         23
3|         142|      3.51|    849.0|              14.05|  0.0|0.48|     1.25
|                2.75|       null|0.94|     12.98|              false|
false|             null|            false|         false|
|           HV0003|              B02764|              B02764|2021-01-01 00:3
9:12|2021-01-01 00:42:37|2021-01-01 00:42:51|2021-01-01 00:45:50|         14
2|         143|      0.74|    179.0|               7.91|  0.0|0.24|      0.7
|                2.75|       null| 0.0|      7.41|              false|
false|             null|            false|         false|
|           HV0003|              B02764|              B02764|2021-01-01 00:4
6:11|2021-01-01 00:47:17|2021-01-01 00:48:14|2021-01-01 01:08:42|         14
3|          78|       9.2|   1228.0|              27.11|  0.0|0.81|     2.41
|                2.75|       null| 0.0|     22.44|              false|
false|             null|            false|         false|
|           HV0005|              B02510|                null|2021-01-01 00:0
4:00|               null|2021-01-01 00:06:59|2021-01-01 00:43:01|          8
8|          42|     9.725|   2162.0|              28.11|  0.0|0.84|     2.49
|                2.75|       null| 0.0|      28.9|              false|
false|             false|           false|        false|
|          HV0005|             B02510|                 null|2021-01-01 00:4
0:06|              null|2021-01-01 00:50:00|2021-01-01 01:04:57|         4
2|         151|     2.469|    897.0|              25.03|  0.0|0.75|     2.22
|              0.0|        null| 0.0|     15.01|              false|
false|             false|           false|        false|
|          HV0003|             B02764|               B02764|2021-01-01 00:1
0:36|2021-01-01 00:12:28|2021-01-01 00:14:30|2021-01-01 00:50:27|         7
1|         226|     13.53|   2157.0|              29.67|  0.0|1.04|     3.08
|              0.0|        null| 0.0|      34.2|              false|
false|              null|           false|        false|
|          HV0003|             B02875|               B02875|2021-01-01 00:2
1:17|2021-01-01 00:22:25|2021-01-01 00:22:54|2021-01-01 00:30:20|        11
2|         255|       1.6|    446.0|               6.89|  0.0|0.21|     0.61
|              0.0|        null| 0.0|      6.26|              false|
false|              null|           false|        false|
|          HV0003|             B02875|               B02875|2021-01-01 00:3
6:57|2021-01-01 00:38:09|2021-01-01 00:40:12|2021-01-01 00:53:31|        25
5|         232|       3.2|    800.0|              11.51|  0.0|0.53|     1.03
|             2.75|        null|2.82|     10.99|              false|
false|              null|           false|        false|
|          HV0003|             B02875|               B02875|2021-01-01 00:5
3:31|2021-01-01 00:56:21|2021-01-01 00:56:45|2021-01-01 01:17:42|        23
2|         198|      5.74|   1257.0|              17.18|  0.0|0.52|     1.52
|             2.75|        null| 0.0|     17.61|              false|
false|              null|           false|        false|
|          HV0003|             B02835|               B02835|2021-01-01 00:2
2:58|2021-01-01 00:27:01|2021-01-01 00:29:04|2021-01-01 00:36:27|        11
3|          48|       1.8|    443.0|               8.18|  0.0|0.25|     0.73
|             2.75|        null| 0.0|      6.12|              false|
false|              null|           false|        false|
|          HV0003|             B02835|               B02835|2021-01-01 00:4
6:44|2021-01-01 00:47:49|2021-01-01 00:48:56|2021-01-01 00:59:12|        23
9|          75|       2.9|    616.0|               13.1|  0.0|0.45|     1.17
|             2.75|        null|0.94|      8.77|              false|
false|              null|           false|        false|
|          HV0004|             B02800|                 null|2021-01-01 00:1
2:50|              null|2021-01-01 00:15:24|2021-01-01 00:38:31|        18
1|         237|      9.66|   1387.0|              32.95|  0.0| 0.0|     2.34
|             2.75|        null| 0.0|      21.1|              false|
false|             false|           false|        false|
|          HV0004|             B02800|                 null|2021-01-01 00:3
5:32|              null|2021-01-01 00:45:00|2021-01-01 01:06:45|        23
6|          68|      4.38|   1305.0|              22.91|  0.0| 0.0|     1.63
|             2.75|        null|3.43|     15.82|              false|
false|             false|           false|        false|
|          HV0003|             B02682|               B02682|2021-01-01 00:1
0:22|2021-01-01 00:11:03|2021-01-01 00:11:53|2021-01-01 00:18:06|        25
6|         148|      2.03|    373.0|               7.84|  0.0|0.42|      0.7
|             2.75|        null|2.82|      6.93|              false|
false|              null|           false|        false|
|          HV0003|             B02682|               B02682|2021-01-01 00:2
5:00|2021-01-01 00:26:31|2021-01-01 00:28:31|2021-01-01 00:41:40|         7
9|          80|      3.08|    789.0|               13.2|  0.0| 0.4|     1.17
|             2.75|        null| 0.0|     11.54|              false|
false|              null|           false|        false|
|          HV0003|             B02682|               B02682|2021-01-01 00:4
```

```
4:56|2021-01-01 00:49:55|2021-01-01 00:50:49|2021-01-01 00:55:59|          1
7|          217|       1.17|    310.0|                   7.91|  0.0|0.24|      0.7
|                 0.0|        null| 0.0|       6.94|                  false|
false|              null|              false|          false|
|           HV0005|              B02510|                   null|2021-01-01 00:0
5:04|              null|2021-01-01 00:08:40|2021-01-01 00:39:39|          6
2|           29|     10.852|   1859.0|                  31.18|  0.0|0.94|     2.77
|                 0.0|        null| 0.0|      27.61|                  false|
false|             false|              false|          false|
|           HV0003|              B02836|                 B02836|2021-01-01 00:4
0:44|2021-01-01 00:53:34|2021-01-01 00:53:48|2021-01-01 01:11:40|          2
2|           22|       3.52|   1072.0|                  28.67|  0.0|0.86|     2.54
|                 0.0|        null| 0.0|      17.64|                  false|
false|              null|              false|          false|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+-----------------+----------------+--------------+
only showing top 20 rows
```

In [0]:

```python
from pyspark.sql import functions as F
from pyspark.sql.types import StringType, BooleanType, IntegerType, DoubleTy

# Convert columns to appropriate data types
df_2022 = df_2022.withColumn("request_datetime", F.col("request_datetime").c
df_2022 = df_2022.withColumn("trip_miles", F.col("trip_miles").cast("double"
df_2022 = df_2022.withColumn("trip_time", F.col("trip_time").cast("double"))

# Convert boolean columns to BooleanType
boolean_columns = ["shared_request_flag", "shared_match_flag", "access_a_rid
for col_name in boolean_columns:
    df_2022 = df_2022.withColumn(col_name, F.col(col_name).cast(BooleanType(

# Convert integer columns to IntegerType
integer_columns = ["PULocationID", "DOLocationID"]
for col_name in integer_columns:
    df_2022 = df_2022.withColumn(col_name, F.col(col_name).cast(IntegerType(

# Convert other categorical columns to StringType
categorical_columns = ["hvfhs_license_num", "dispatching_base_num", "origina
for col_name in categorical_columns:
    df_2022 = df_2022.withColumn(col_name, F.col(col_name).cast(StringType()

# Print the updated schema
df_2022.printSchema()

df_2022.show()
```

```
root
 |-- hvfhs_license_num: string (nullable = true)
 |-- dispatching_base_num: string (nullable = true)
 |-- originating_base_num: string (nullable = true)
 |-- request_datetime: timestamp (nullable = true)
 |-- on_scene_datetime: timestamp (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- PULocationID: integer (nullable = true)
 |-- DOLocationID: integer (nullable = true)
 |-- trip_miles: double (nullable = true)
 |-- trip_time: double (nullable = true)
 |-- base_passenger_fare: double (nullable = true)
 |-- tolls: double (nullable = true)
 |-- bcf: double (nullable = true)
 |-- sales_tax: double (nullable = true)
 |-- congestion_surcharge: double (nullable = true)
 |-- airport_fee: double (nullable = true)
 |-- tips: double (nullable = true)
 |-- driver_pay: double (nullable = true)
 |-- shared_request_flag: boolean (nullable = true)
 |-- shared_match_flag: boolean (nullable = true)
 |-- access_a_ride_flag: boolean (nullable = true)
 |-- wav_request_flag: boolean (nullable = true)
 |-- wav_match_flag: boolean (nullable = true)

+-----------------+-------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|hvfhs_license_num|dispatching_base_num|originating_base_num|   request_date
time|  on_scene_datetime|    pickup_datetime|   dropoff_datetime|PULocationI
D|DOLocationID|trip_miles|trip_time|base_passenger_fare|tolls| bcf|sales_tax
|congestion_surcharge|airport_fee|tips|driver_pay|shared_request_flag|shared
_match_flag|access_a_ride_flag|wav_request_flag|wav_match_flag|
+-----------------+-------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|           HV0003|              B03404|              B03404|2022-01-01 00:0
5:31|2022-01-01 00:05:40|2022-01-01 00:07:24|2022-01-01 00:18:28|         17
0|         161|      1.18|    664.0|               24.9|  0.0|0.75|     2.21
|                2.75|        0.0| 0.0|     23.03|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:1
9:27|2022-01-01 00:22:08|2022-01-01 00:22:32|2022-01-01 00:30:12|         23
7|         161|      0.82|    460.0|              11.97|  0.0|0.36|     1.06
|                2.75|        0.0| 0.0|     12.32|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
3:53|2022-01-01 00:57:37|2022-01-01 00:57:37|2022-01-01 01:07:32|         23
7|         161|      1.18|    595.0|              29.82|  0.0|0.89|     2.65
|                2.75|        0.0| 0.0|      23.3|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:1
5:36|2022-01-01 00:17:08|2022-01-01 00:18:02|2022-01-01 00:23:05|         26
2|         229|      1.65|    303.0|               7.91|  0.0|0.24|      0.7
|                2.75|        0.0| 0.0|       6.3|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:2
5:45|2022-01-01 00:26:01|2022-01-01 00:28:01|2022-01-01 00:35:42|         22
9|         141|      1.65|    461.0|               9.44|  0.0|0.28|     0.84
```

```
|                2.75|         0.0| 0.0|      7.44|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:3
4:44|2022-01-01 00:36:52|2022-01-01 00:38:50|2022-01-01 00:51:32|         26
3|          79|      4.51|    762.0|              17.67|  0.0|0.53|     1.57
|                2.75|         0.0| 0.0|     12.25|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
7:51|2022-01-01 00:52:00|2022-01-01 00:53:25|2022-01-01 01:08:56|         11
3|         140|      3.68|    931.0|              16.68|  0.0| 0.5|     1.48
|                2.75|         0.0| 0.0|     12.75|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:0
6:21|2022-01-01 00:06:58|2022-01-01 00:08:58|2022-01-01 00:23:01|         15
1|          75|      2.77|    843.0|              14.41|  0.0|0.43|     1.28
|                 0.0|         0.0| 4.0|     11.47|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:2
7:54|2022-01-01 00:30:26|2022-01-01 00:32:25|2022-01-01 00:44:15|         26
3|         229|      2.04|    710.0|              10.64|  0.0|0.32|     0.94
|                2.75|         0.0| 0.0|      9.55|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
4:59|2022-01-01 00:48:23|2022-01-01 00:50:23|2022-01-01 01:15:30|         23
7|         169|      8.79|   1507.0|             107.56|  0.0|0.83|     2.45
|                2.75|         0.0| 0.0|     23.67|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:1
3:49|2022-01-01 00:16:15|2022-01-01 00:17:02|2022-01-01 00:40:09|         26
1|         223|     11.29|   1387.0|               34.9|  0.0|1.05|      3.1
|                2.75|         0.0| 0.0|     25.17|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:3
9:10|2022-01-01 00:42:59|2022-01-01 00:43:20|2022-01-01 00:47:31|         22
3|         223|      0.87|    251.0|               7.91|  0.0|0.24|      0.7
|                 0.0|         0.0| 0.0|      6.51|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
5:50|2022-01-01 00:52:15|2022-01-01 00:52:29|2022-01-01 01:01:48|         22
3|           7|      1.89|    559.0|               9.71|  0.0|0.29|     0.86
|                 0.0|         0.0| 0.0|      7.89|               false|
false|               null|           false|         false|
|           HV0005|              B03406|                null|2022-01-01 00:3
6:54|               null|2022-01-01 00:45:34|2022-01-01 00:54:11|          8
8|         148|     3.585|    810.0|              27.02|  0.0|0.81|      2.4
|                2.75|         0.0| 0.0|     21.08|               false|
false|              false|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:0
7:13|2022-01-01 00:12:03|2022-01-01 00:12:03|2022-01-01 00:35:07|         24
6|         243|       9.2|   1384.0|              30.37|  0.0|0.91|      2.7
|                2.75|         0.0| 0.0|     22.69|               false|
false|               null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:5
3:32|2022-01-01 00:58:14|2022-01-01 00:58:26|2022-01-01 01:07:23|         24
3|         127|      1.88|    537.0|              15.02|  0.0|0.45|     1.33
|                 0.0|         0.0| 2.0|     10.02|               false|
false|               null|           false|         false|
|           HV0005|                B03406|                null|2022-01-01 00:2
8:07|              null|2022-01-01 00:34:59|2022-01-01 00:50:15|         23
9|         170|      2.699|    916.0|                20.47|  0.0|0.61|      1.82
|                2.75|          0.0| 0.0|      10.66|              false|
false|              false|               false|          false|
|           HV0005|                B03406|                null|2022-01-01 00:4
5:44|              null|2022-01-01 00:58:09|2022-01-01 01:28:23|         17
0|         265|      8.037|   1904.0|                24.63| 20.0|1.34|       0.0
|                 0.0|          0.0| 0.0|      26.63|              false|
false|              false|               false|          false|
|           HV0003|                B03404|              B03404|2022-01-01 00:2
1:54|2022-01-01 00:26:15|2022-01-01 00:28:15|2022-01-01 00:38:52|         22
3|         179|       1.69|    637.0|                 9.81|  0.0|0.29|      0.87
|                 0.0|          0.0| 0.0|       9.55|              false|
false|               null|               false|          false|
|           HV0003|                B03404|              B03404|2022-01-01 00:3
5:08|2022-01-01 00:43:55|2022-01-01 00:45:55|2022-01-01 00:53:47|         17
9|           7|       0.87|    472.0|                10.53|  0.0|0.32|      0.93
|                 0.0|          0.0| 0.0|       9.22|              false|
false|               null|               false|          false|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+--------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
only showing top 20 rows
```

Removing Outliers and Duplicates from 2021 and 2022

In [0]:
```python
from pyspark.sql.functions import col

# Define a function to remove outliers from a column using IQR
def remove_outliers(df, col_name):
    q1, q3 = df.approxQuantile(col_name, [0.25, 0.75], 0.05)
    iqr = q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    return df.filter((col(col_name) >= lower_bound) & (col(col_name) <= uppe

# Remove outliers from specific columns
columns_to_remove_outliers = ["trip_miles", "trip_time"]
for col_name in columns_to_remove_outliers:
    df_2021_no_outliers = remove_outliers(df_2021, col_name)

# Remove duplicate rows
df_2021_no_duplicates = df_2021_no_outliers.dropDuplicates()
```

In [0]:
```python
# Remove outliers from specific columns
columns_to_remove_outliers = ["trip_miles", "trip_time"]
for col_name in columns_to_remove_outliers:
    df_2022_no_outliers = remove_outliers(df_2022, col_name)

# Remove duplicate rows
df_2022_no_duplicates = df_2022_no_outliers.dropDuplicates()
```

Dropping Unimportant Columns in 2021 and 2022

In [0]:
```python
from pyspark.sql.functions import count, col, when

df_2021_Sample = df_2021_no_duplicates
#.limit(20000)
df_2021_Sample.show()

# Drop rows with any missing values
df_2021_cleaned = df_2021_Sample.dropna()

#removing unimportant columns from the dataframe
df_2021_cleaned=df_2021_cleaned.drop("tolls","bcf","sales_tax","congestion_s

'''# Impute missing values with mean or other strategies
from pyspark.ml.feature import Imputer

imputer = Imputer(inputCols=df_2021_cleaned.columns, outputCols=df_2021_clea
df_2021_imputed = imputer.fit(df_2021_cleaned).transform(df_2021_cleaned)'''
```

```
+-----------------+--------------------+--------------------+---------------
----+--------------------+--------------------+--------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+--------------------+------
-----------+------------------+----------------+--------------+
|hvfhs_license_num|dispatching_base_num|originating_base_num|   request_date
time|   on_scene_datetime|     pickup_datetime|    dropoff_datetime|PULocationI
D|DOLocationID|trip_miles|trip_time|base_passenger_fare|tolls| bcf|sales_tax
|congestion_surcharge|airport_fee|tips|driver_pay|shared_request_flag|shared
_match_flag|access_a_ride_flag|wav_request_flag|wav_match_flag|
+-----------------+--------------------+--------------------+---------------
----+--------------------+--------------------+--------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+--------------------+------
-----------+------------------+----------------+--------------+
|           HV0003|              B02875|              B02875|2021-01-01 00:3
6:57|2021-01-01 00:38:09|2021-01-01 00:40:12|2021-01-01 00:53:31|         25
5|         232|       3.2|    800.0|              11.51|  0.0|0.53|     1.03
|                2.75|       null|2.82|     10.99|                  false|
false|              null|            false|         false|
|           HV0003|              B02764|              B02764|2021-01-01 00:1
0:36|2021-01-01 00:12:28|2021-01-01 00:14:30|2021-01-01 00:50:27|          7
1|         226|     13.53|   2157.0|              29.67|  0.0|1.04|     3.08
|                 0.0|       null| 0.0|      34.2|                  false|
false|              null|            false|         false|
|           HV0005|              B02510|                null|2021-01-01 00:1
1:57|                null|2021-01-01 00:22:24|2021-01-01 00:36:23|          9
2|         191|     5.864|    839.0|              19.82|  0.0|0.59|     1.76
|                 0.0|       null| 0.0|     13.54|                  false|
false|             false|            false|         false|
|           HV0003|              B02764|              B02764|2021-01-01 00:4
6:11|2021-01-01 00:47:17|2021-01-01 01:08:42|         14
3|          78|       9.2|   1228.0|              27.11|  0.0|0.81|     2.41
|                2.75|       null| 0.0|     22.44|                  false|
false|              null|            false|         false|
|           HV0003|              B02875|              B02875|2021-01-01 00:2
1:17|2021-01-01 00:22:25|2021-01-01 00:22:54|2021-01-01 00:30:20|          11
2|          255|        1.6|    446.0|              6.89|  0.0|0.21|      0.61
|                0.0|         null| 0.0|       6.26|               false|
false|             null|             false|          false|
|           HV0003|              B02835|              B02835|2021-01-01 00:2
2:58|2021-01-01 00:27:01|2021-01-01 00:29:04|2021-01-01 00:36:27|          11
3|           48|        1.8|    443.0|              8.18|  0.0|0.25|      0.73
|               2.75|         null| 0.0|       6.12|               false|
false|             null|             false|          false|
|           HV0005|              B02510|                null|2021-01-01 00:4
0:06|              null|2021-01-01 00:50:00|2021-01-01 01:04:57|           4
2|          151|      2.469|    897.0|             25.03|  0.0|0.75|      2.22
|                0.0|         null| 0.0|      15.01|               false|
false|            false|             false|          false|
|           HV0003|              B02764|              B02764|2021-01-01 00:3
9:12|2021-01-01 00:42:37|2021-01-01 00:42:51|2021-01-01 00:45:50|          14
2|          143|       0.74|    179.0|              7.91|  0.0|0.24|       0.7
|               2.75|         null| 0.0|       7.41|               false|
false|             null|             false|          false|
|           HV0003|              B02682|              B02682|2021-01-01 00:1
0:22|2021-01-01 00:11:03|2021-01-01 00:11:53|2021-01-01 00:18:06|          25
6|          148|       2.03|    373.0|              7.84|  0.0|0.42|       0.7
|               2.75|         null|2.82|       6.93|               false|
false|             null|             false|          false|
|           HV0005|              B02510|                null|2021-01-01 00:0
5:04|              null|2021-01-01 00:08:40|2021-01-01 00:39:39|           6
2|           29|     10.852|   1859.0|             31.18|  0.0|0.94|      2.77
|                0.0|         null| 0.0|      27.61|               false|
false|            false|             false|          false|
|           HV0005|              B02510|                null|2021-01-01 00:0
4:00|              null|2021-01-01 00:06:59|2021-01-01 00:43:01|           8
8|           42|      9.725|   2162.0|             28.11|  0.0|0.84|      2.49
|               2.75|         null| 0.0|       28.9|               false|
false|            false|             false|          false|
|           HV0003|              B02682|              B02682|2021-01-01 00:2
5:00|2021-01-01 00:26:31|2021-01-01 00:28:31|2021-01-01 00:41:40|           7
9|           80|       3.08|    789.0|              13.2|  0.0| 0.4|      1.17
|               2.75|         null| 0.0|      11.54|               false|
false|             null|             false|          false|
|           HV0003|              B02764|              B02764|2021-01-01 00:2
1:15|2021-01-01 00:22:41|2021-01-01 00:23:56|2021-01-01 00:38:05|          23
3|          142|       3.51|    849.0|             14.05|  0.0|0.48|      1.25
|               2.75|         null|0.94|      12.98|               false|
false|             null|             false|          false|
|           HV0003|              B02836|              B02836|2021-01-01 00:4
0:44|2021-01-01 00:53:34|2021-01-01 00:53:48|2021-01-01 01:11:40|           2
2|           22|       3.52|   1072.0|             28.67|  0.0|0.86|      2.54
|                0.0|         null| 0.0|      17.64|               false|
false|             null|             false|          false|
|           HV0004|              B02800|                null|2021-01-01 00:1
2:50|              null|2021-01-01 00:15:24|2021-01-01 00:38:31|          18
1|          237|       9.66|   1387.0|             32.95|  0.0| 0.0|      2.34
|               2.75|         null| 0.0|       21.1|               false|
false|            false|             false|          false|
|           HV0003|              B02682|              B02682|2021-01-01 00:4
5:56|2021-01-01 00:55:19|2021-01-01 00:55:19|2021-01-01 01:18:21|          15
```

```
2|         167|       3.65|   1382.0|                18.36|  0.0|0.55|     1.63
|                 0.0|        null| 0.0|     17.06|                false|
false|               null|            false|          false|
|          HV0003|              B02682|               B02682|2021-01-01 00:2
8:09|2021-01-01 00:31:42|2021-01-01 00:33:44|2021-01-01 00:49:07|         23
0|         166|       5.26|    923.0|                22.28|  0.0|0.67|     1.98
|                2.75|        null| 0.0|     14.99|                false|
false|               null|            false|          false|
|          HV0004|              B02800|                 null|2021-01-01 00:3
5:32|               null|2021-01-01 00:45:00|2021-01-01 01:06:45|         23
6|          68|       4.38|   1305.0|                22.91|  0.0| 0.0|     1.63
|                2.75|        null|3.43|     15.82|                false|
false|              false|            false|          false|
|          HV0003|              B02835|               B02835|2021-01-01 00:4
6:44|2021-01-01 00:47:49|2021-01-01 00:48:56|2021-01-01 00:59:12|         23
9|          75|        2.9|    616.0|                 13.1|  0.0|0.45|     1.17
|                2.75|        null|0.94|      8.77|                false|
false|               null|            false|          false|
|          HV0003|              B02875|               B02875|2021-01-01 00:5
3:31|2021-01-01 00:56:21|2021-01-01 00:56:45|2021-01-01 01:17:42|         23
2|         198|       5.74|   1257.0|                17.18|  0.0|0.52|     1.52
|                2.75|        null| 0.0|     17.61|                false|
false|               null|            false|          false|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
only showing top 20 rows

Out[97]: '# Impute missing values with mean or other strategies\nfrom pyspar
k.ml.feature import Imputer\n\nimputer = Imputer(inputCols=df_2021_cleaned.c
olumns, outputCols=df_2021_cleaned.columns)\ndf_2021_imputed = imputer.fit(d
f_2021_cleaned).transform(df_2021_cleaned)'
```

In [0]:

```python
df_2022_Sample = df_2022_no_duplicates
df_2022_Sample.show()

# Drop rows with any missing values
df_2022_cleaned = df_2022_Sample.dropna()

#removing unimportant columns from the dataframe
df_2022_cleaned=df_2022_cleaned.drop("tolls","bcf","sales_tax","congestion_s
```

```
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|hvfhs_license_num|dispatching_base_num|originating_base_num|   request_date
time|  on_scene_datetime|    pickup_datetime|   dropoff_datetime|PULocationI
D|DOLocationID|trip_miles|trip_time|base_passenger_fare|tolls| bcf|sales_tax
|congestion_surcharge|airport_fee|tips|driver_pay|shared_request_flag|shared
_match_flag|access_a_ride_flag|wav_request_flag|wav_match_flag|
+-----------------+--------------------+--------------------+---------------
----+-------------------+-------------------+-------------------+-----------
-+------------+----------+---------+-------------------+-----+----+---------
```

```
+--------------------+-----------+----+----------+-------------------+------
-----------+------------------+----------------+--------------+
|           HV0005|              B03406|                null|2022-01-01 00:4
5:44|               null|2022-01-01 00:58:09|2022-01-01 01:28:23|         17
0|         265|     8.037|   1904.0|              24.63|20.0|1.34|      0.0
|                 0.0|        0.0| 0.0|     26.63|              false|
false|             false|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:2
1:54|2022-01-01 00:26:15|2022-01-01 00:28:15|2022-01-01 00:38:52|         22
3|         179|      1.69|    637.0|               9.81| 0.0|0.29|     0.87
|                 0.0|        0.0| 0.0|      9.55|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:3
4:44|2022-01-01 00:36:52|2022-01-01 00:38:50|2022-01-01 00:51:32|         26
3|          79|      4.51|    762.0|              17.67| 0.0|0.53|     1.57
|                2.75|        0.0| 0.0|     12.25|              false|
false|              null|           false|         false|
|           HV0005|              B03406|                null|2022-01-01 00:2
8:07|               null|2022-01-01 00:34:59|2022-01-01 00:50:15|         23
9|         170|     2.699|    916.0|              20.47| 0.0|0.61|     1.82
|                2.75|        0.0| 0.0|     10.66|              false|
false|             false|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
4:59|2022-01-01 00:48:23|2022-01-01 00:50:23|2022-01-01 01:15:30|         23
7|         169|      8.79|   1507.0|             107.56| 0.0|0.83|     2.45
|                2.75|        0.0| 0.0|     23.67|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:0
7:13|2022-01-01 00:12:03|2022-01-01 00:12:03|2022-01-01 00:35:07|         24
6|         243|       9.2|   1384.0|              30.37| 0.0|0.91|      2.7
|                2.75|        0.0| 0.0|     22.69|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:3
9:10|2022-01-01 00:42:59|2022-01-01 00:43:20|2022-01-01 00:47:31|         22
3|         223|      0.87|    251.0|               7.91| 0.0|0.24|      0.7
|                 0.0|        0.0| 0.0|      6.51|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:2
7:54|2022-01-01 00:30:26|2022-01-01 00:32:25|2022-01-01 00:44:15|         26
3|         229|      2.04|    710.0|              10.64| 0.0|0.32|     0.94
|                2.75|        0.0| 0.0|      9.55|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:1
9:27|2022-01-01 00:22:08|2022-01-01 00:22:32|2022-01-01 00:30:12|         23
7|         161|      0.82|    460.0|              11.97| 0.0|0.36|     1.06
|                2.75|        0.0| 0.0|     12.32|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
5:50|2022-01-01 00:52:15|2022-01-01 00:52:29|2022-01-01 01:01:48|         22
3|           7|      1.89|    559.0|               9.71| 0.0|0.29|     0.86
|                 0.0|        0.0| 0.0|      7.89|              false|
false|              null|           false|         false|
|           HV0003|              B03404|              B03404|2022-01-01 00:4
3:53|2022-01-01 00:57:37|2022-01-01 00:57:37|2022-01-01 01:07:32|         23
7|         161|      1.18|    595.0|              29.82| 0.0|0.89|     2.65
|                2.75|        0.0| 0.0|      23.3|              false|
false|              null|           false|         false|
|            HV0003|              B03404|               B03404|2022-01-01 00:2
5:45|2022-01-01 00:26:01|2022-01-01 00:28:01|2022-01-01 00:35:42|        22
9|         141|      1.65|    461.0|                9.44|  0.0|0.28|     0.84
|                2.75|         0.0| 0.0|      7.44|              false|
false|              null|            false|         false|
|            HV0003|              B03404|               B03404|2022-01-01 00:0
5:31|2022-01-01 00:05:40|2022-01-01 00:07:24|2022-01-01 00:18:28|        17
0|         161|      1.18|    664.0|                24.9|  0.0|0.75|     2.21
|                2.75|         0.0| 0.0|     23.03|              false|
false|              null|            false|         false|
|            HV0003|              B03404|               B03404|2022-01-01 00:0
6:21|2022-01-01 00:06:58|2022-01-01 00:08:58|2022-01-01 00:23:01|        15
1|          75|      2.77|    843.0|               14.41|  0.0|0.43|     1.28
|                 0.0|         0.0| 4.0|     11.47|              false|
false|              null|            false|         false|
|            HV0003|              B03404|               B03404|2022-01-01 00:1
3:49|2022-01-01 00:16:15|2022-01-01 00:17:02|2022-01-01 00:40:09|        26
1|         223|     11.29|   1387.0|                34.9|  0.0|1.05|      3.1
|                2.75|         0.0| 0.0|     25.17|              false|
false|              null|            false|         false|
|            HV0003|              B03404|               B03404|2022-01-01 00:4
8:14|2022-01-01 00:55:03|2022-01-01 00:55:30|2022-01-01 01:03:26|
7|          82|      2.59|    476.0|               14.87|  0.0|0.45|     1.32
|                 0.0|         0.0| 0.0|      8.49|              false|
false|              null|            false|         false|
|            HV0003|              B03404|               B03404|2022-01-01 00:1
5:36|2022-01-01 00:17:08|2022-01-01 00:18:02|2022-01-01 00:23:05|        26
2|         229|      1.65|    303.0|                7.91|  0.0|0.24|      0.7
|                2.75|         0.0| 0.0|       6.3|              false|
false|              null|            false|         false|
|            HV0003|              B03404|               B03404|2022-01-01 00:5
3:32|2022-01-01 00:58:14|2022-01-01 00:58:26|2022-01-01 01:07:23|        24
3|         127|      1.88|    537.0|               15.02|  0.0|0.45|     1.33
|                 0.0|         0.0| 2.0|     10.02|              false|
false|              null|            false|         false|
|            HV0005|              B03406|                 null|2022-01-01 00:3
6:54|              null|2022-01-01 00:45:34|2022-01-01 00:54:11|         8
8|         148|     3.585|    810.0|               27.02|  0.0|0.81|      2.4
|                2.75|         0.0| 0.0|     21.08|              false|
false|             false|            false|         false|
|            HV0003|              B03404|               B03404|2022-01-01 00:4
7:51|2022-01-01 00:52:00|2022-01-01 00:53:25|2022-01-01 01:08:56|        11
3|         140|      3.68|    931.0|               16.68|  0.0| 0.5|     1.48
|                2.75|         0.0| 0.0|     12.75|              false|
false|              null|            false|         false|
+------------------+--------------------+--------------------+----------------
----+--------------------+--------------------+--------------------+-----------
-+------------+----------+---------+--------------------+-----+----+---------
+--------------------+-----------+----+----------+-------------------+------
------------+------------------+----------------+--------------+
only showing top 20 rows
```

In [0]:

```
df_2021_cleaned.summary()
```

```
Out[99]: DataFrame[summary: string, hvfhs_license_num: string, dispatching_b
ase_num: string, originating_base_num: string, PULocationID: string, DOLocat
ionID: string, trip_miles: string, trip_time: string, base_passenger_fare: s
tring, tips: string, driver_pay: string]
```

In [0]:
```
df_2022_cleaned.summary()
```

```
Out[100]: DataFrame[summary: string, hvfhs_license_num: string, dispatching_
base_num: string, originating_base_num: string, PULocationID: string, DOLoca
tionID: string, trip_miles: string, trip_time: string, base_passenger_fare:
string, tips: string, driver_pay: string]
```

Average Trip Duration for each pick up for 2021 and 2021

In [0]:
```
from pyspark.sql.functions import hour, col, unix_timestamp, avg

# Assuming your timestamps are in a string format
# Convert them to timestamp type
df = df_2021_cleaned.withColumn("pickup_datetime", unix_timestamp(col("picku
df = df_2021_cleaned.withColumn("dropoff_datetime", unix_timestamp(col("drop

# Extract hour from pickup_datetime for temporal analysis
df = df.withColumn("pickup_hour", hour("pickup_datetime"))

# Group by pickup_hour for peak and off-peak analysis
peak_offpeak_analysis = df.groupBy("pickup_hour").agg(avg("trip_time").alias

# Show the results
peak_offpeak_analysis.show()
```

```
+-----------+------------------+
|pickup_hour| avg_trip_duration|
+-----------+------------------+
|         12|1275.7314731473148|
|         22|1148.7966985230235|
|          1| 1032.119266055046|
|         13|1266.1343532684284|
|         16|1279.5437753971187|
|          6|1264.2215568862275|
|          3|1049.4830769230769|
|         20|  1163.93631778058|
|          5|1202.9431818181818|
|         19|1178.2867058195409|
|         15|1285.0228013029316|
|         17|1239.9241744802282|
|          9|1255.0997023809523|
|          4|     1158.74609375|
|          8|1295.5418950665623|
|         23|1146.3377823408625|
|          7| 1262.549019607843|
|         10|1261.8618541590326|
|         21|1115.3519434628975|
|         11|1254.6858174655285|
+-----------+------------------+
only showing top 20 rows
```

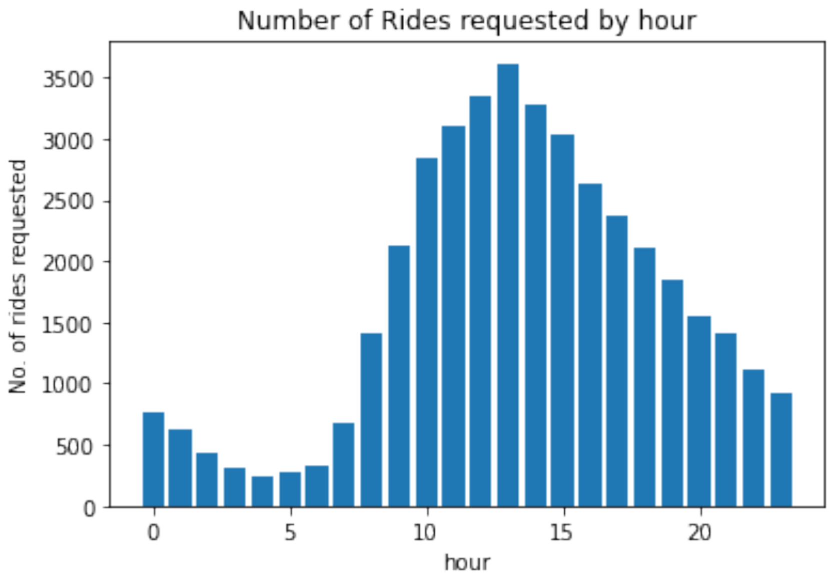
In [0]:
```
from pyspark.sql.functions import hour, col, unix_timestamp, avg

# Assuming your timestamps are in a string format
# Convert them to timestamp type
df = df_2022_cleaned.withColumn("pickup_datetime", unix_timestamp(col("picku
df = df_2022_cleaned.withColumn("dropoff_datetime", unix_timestamp(col("drop

# Extract hour from pickup_datetime for temporal analysis
df = df.withColumn("pickup_hour", hour("pickup_datetime"))

# Group by pickup_hour for peak and off-peak analysis
peak_offpeak_analysis = df.groupBy("pickup_hour").agg(avg("trip_time").alias

# Show the results
peak_offpeak_analysis.show()
```

```
+-----------+------------------+
|pickup_hour| avg_trip_duration|
+-----------+------------------+
|         12|1322.0442397977608|
|         22|1183.3990903922684|
|          1| 1084.404052443385|
|         13|1303.4444979919679|
|          6|1160.1767676767677|
|         16|1343.4551937247445|
|          3|1183.5644444444445|
|         20|1170.7235668789808|
|          5|1183.7664835164835|
|         19| 1209.151106111736|
|         15|1317.7156398104266|
|         17|1314.6259925886714|
|          9| 1290.034966887417|
|          4|1179.5340050377833|
|          8| 1373.143855322647|
|         23|1183.6850828729282|
|          7|1303.4048913043478|
|         10|1322.7816014394962|
|         21| 1163.602847324497|
|         11|1312.1053302187302|
+-----------+------------------+
only showing top 20 rows
```

Number of Rides requested per hour for 2021 and 2022

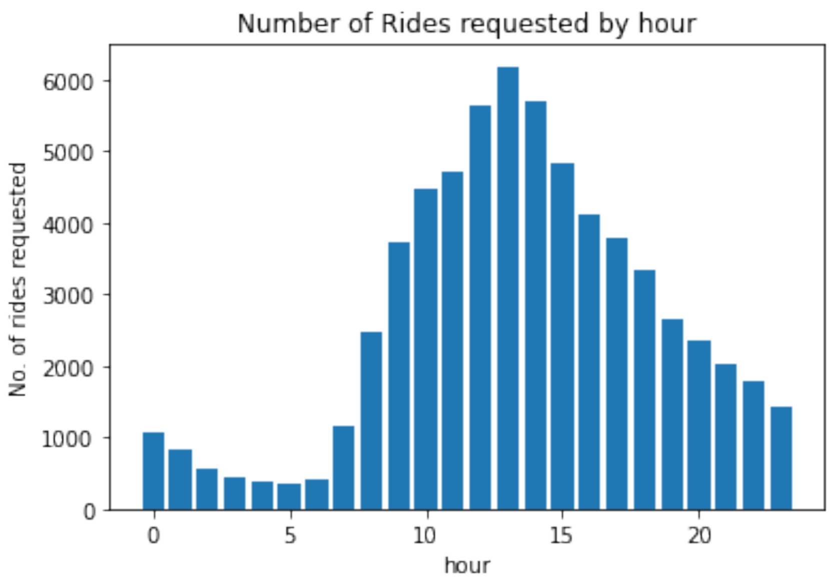
In [0]:

```
# Your code goes here
from matplotlib import pyplot as plt
# group the rows by hour and count the number of rows in each hour
count_df = df_2021_cleaned.groupBy(hour("request_datetime").alias("hour")).a

# convert the PySpark DataFrame to a Pandas DataFrame for plotting
pandas_df = count_df.toPandas()

# plot the bar chart
plt.bar(pandas_df["hour"], pandas_df["count"])

# add chart title and labels
plt.title("Number of Rides requested by hour")
plt.xlabel("hour")
plt.ylabel("No. of rides requested")

# show the chart
plt.show()
```



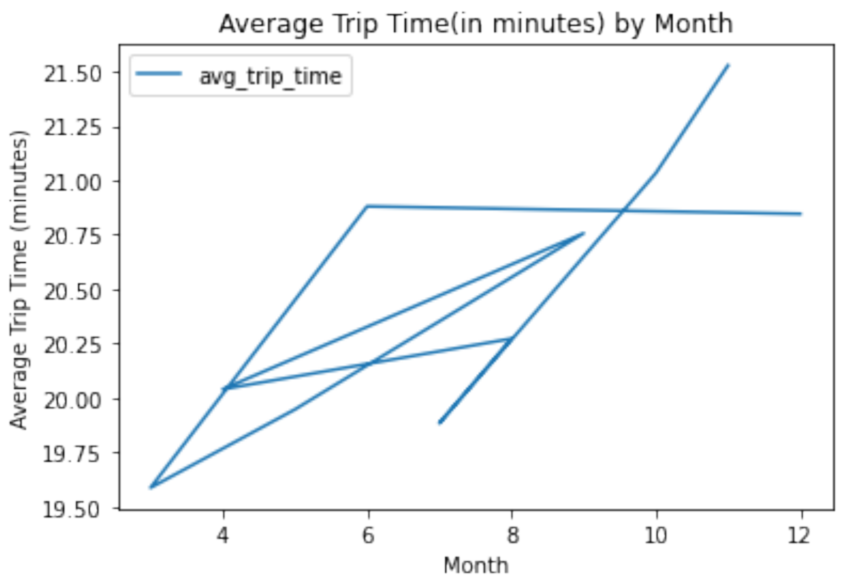
In [0]:

```
from matplotlib import pyplot as plt
# group the rows by hour and count the number of rows in each hour
count_df = df_2022_cleaned.groupBy(hour("request_datetime").alias("hour")).a

# convert the PySpark DataFrame to a Pandas DataFrame for plotting
pandas_df = count_df.toPandas()

# plot the bar chart
plt.bar(pandas_df["hour"], pandas_df["count"])

# add chart title and labels
plt.title("Number of Rides requested by hour")
plt.xlabel("hour")
plt.ylabel("No. of rides requested")

# show the chart
plt.show()
```

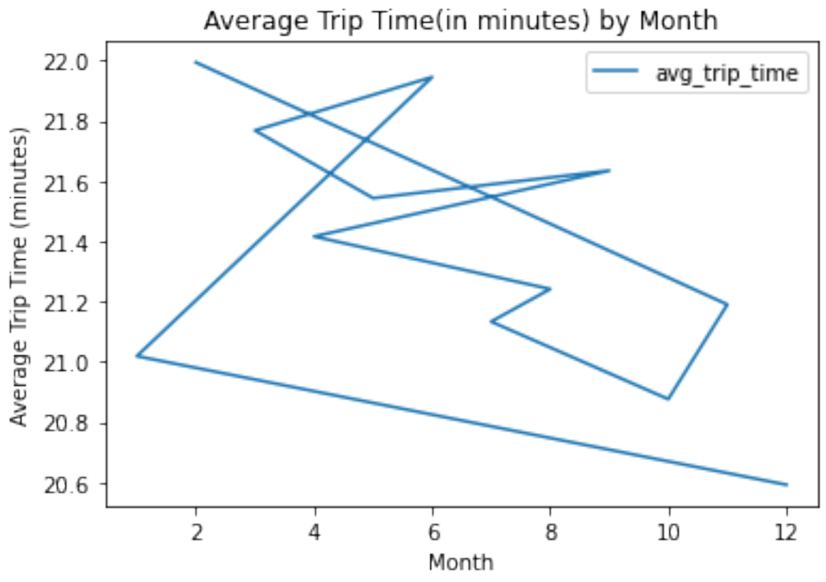


Average Trip Time(in minutes) by Month for 2021 and 2022

In [0]:
```
# group the rows by month and calculate the average trip time
avg_time_df = df_2021_cleaned.groupBy(month("request_datetime").alias("month

# convert the PySpark DataFrame to a Pandas DataFrame for plotting
pandas_df = avg_time_df.toPandas()

pandas_df['avg_trip_time']= (pandas_df['avg_trip_time']/60)

# plot the line graph
ax = pandas_df.plot(x="month", y="avg_trip_time", kind="line")

# set x and y labels
ax.set_xlabel("Month")
ax.set_ylabel("Average Trip Time (minutes)")

# set the plot title
ax.set_title("Average Trip Time(in minutes) by Month")

# show the plot
plt.show()
```

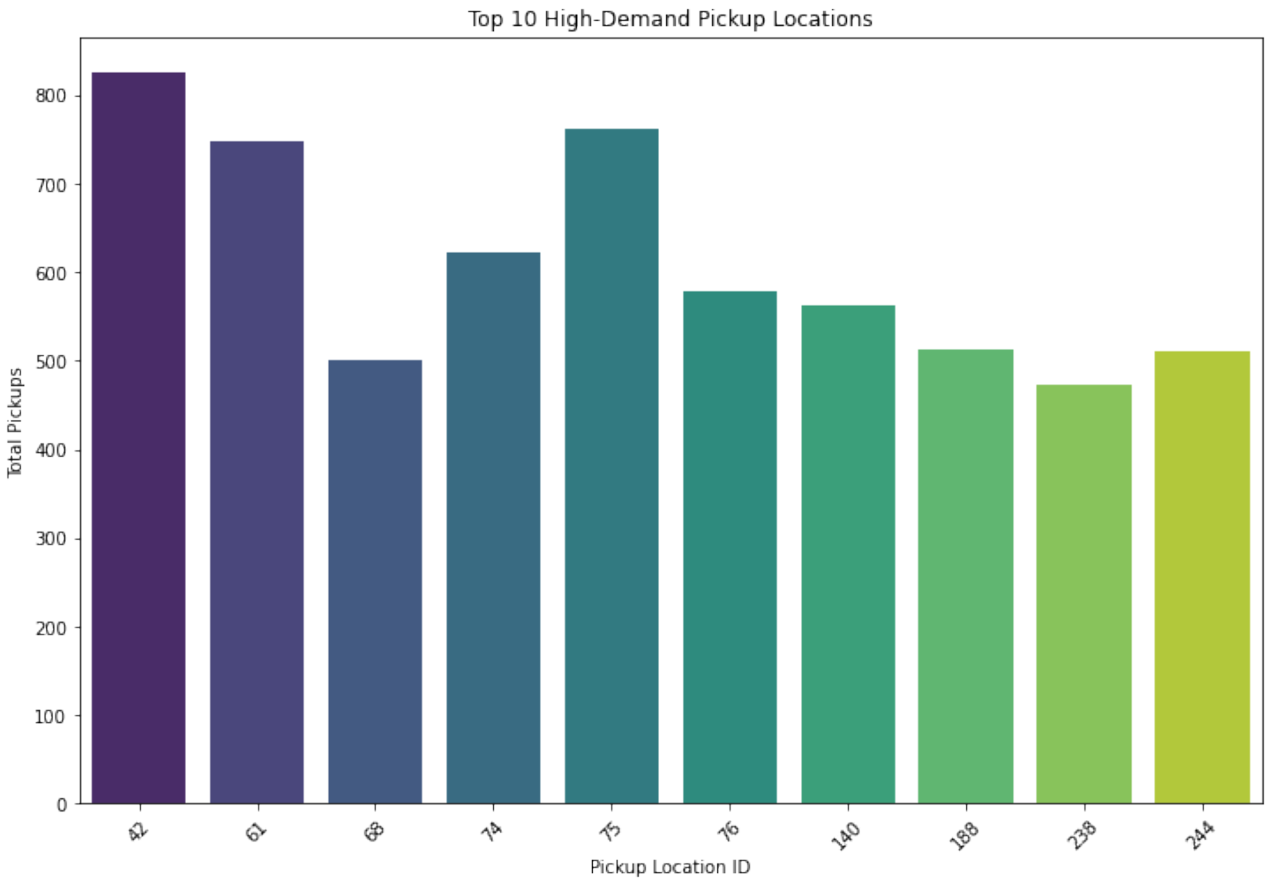


```
In [0]: # group the rows by month and calculate the average trip time
        avg_time_df = df_2022_cleaned.groupBy(month("request_datetime").alias("month

        # convert the PySpark DataFrame to a Pandas DataFrame for plotting
        pandas_df = avg_time_df.toPandas()

        pandas_df['avg_trip_time']= (pandas_df['avg_trip_time']/60)

        # plot the line graph
        ax = pandas_df.plot(x="month", y="avg_trip_time", kind="line")

        # set x and y labels
        ax.set_xlabel("Month")
        ax.set_ylabel("Average Trip Time (minutes)")

        # set the plot title
        ax.set_title("Average Trip Time(in minutes) by Month")

        # show the plot
        plt.show()
```

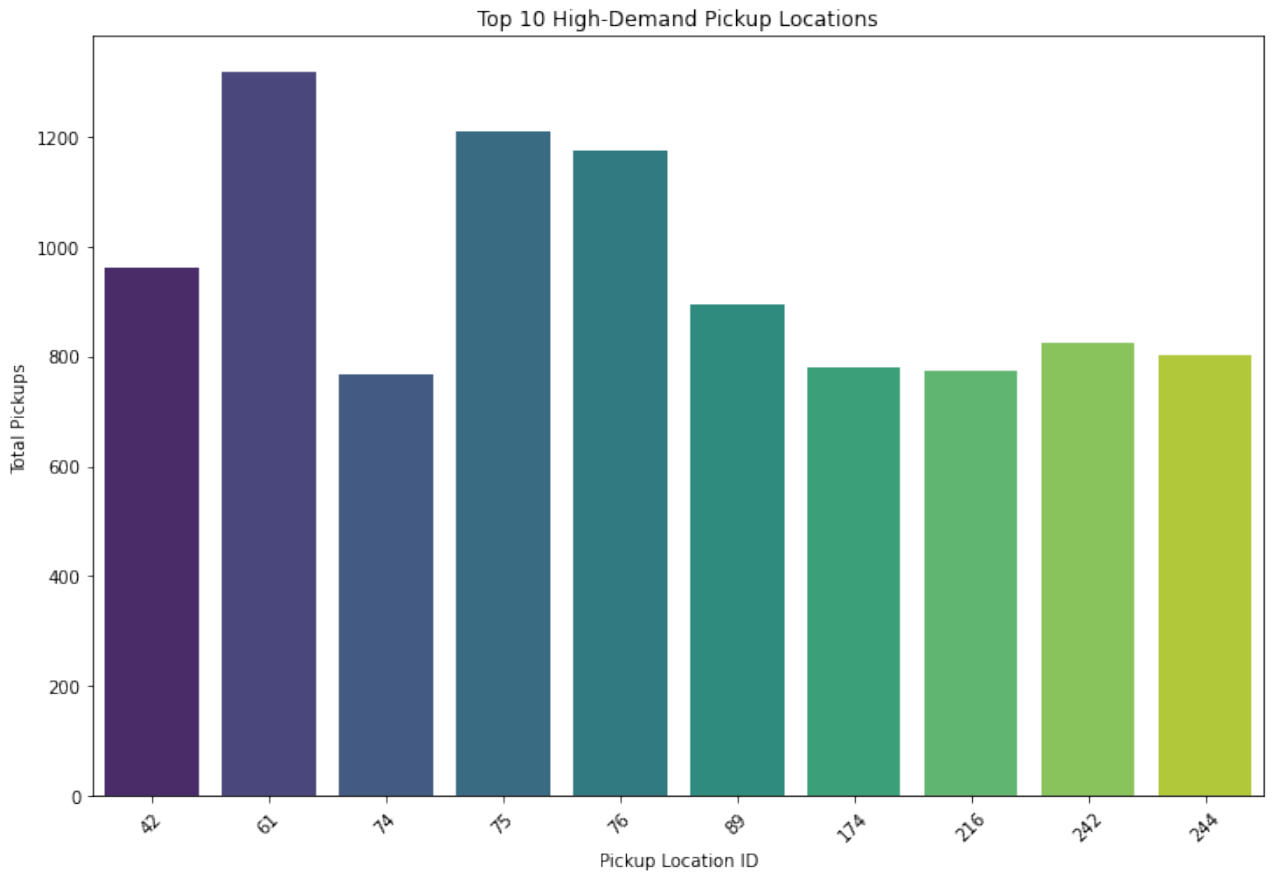


Top 10 High-Demand Pickup Locations for 2021 and 2022

In [0]:
```python
# Identify high-demand pickup locations
demand_analysis = (
    df_2021_cleaned.groupBy("PULocationID")
    .count()
    .withColumnRenamed("count", "total_pickups")
    .orderBy("total_pickups", ascending=False)
)
import matplotlib.pyplot as plt
import seaborn as sns

#PySpark DataFrame is sorted
demand_analysis = demand_analysis.orderBy("total_pickups", ascending=False)

# Convert to Pandas DataFrame and reset the index
demand_analysis_pd = demand_analysis.toPandas().reset_index(drop=True)

# Select the top 10 high-demand pickup locations
top_demand = demand_analysis_pd.head(10)

# Normalize the 'total_pickups' if needed, otherwise skip this step
# top_demand['total_pickups'] = top_demand['total_pickups'] / top_demand['to

# Plotting with seaborn
plt.figure(figsize=(12, 8))
sns.barplot(
    x='PULocationID',
    y='total_pickups',
    data=top_demand,
    palette='viridis'
)

plt.title('Top 10 High-Demand Pickup Locations')
plt.xlabel('Pickup Location ID')
plt.ylabel('Total Pickups')
plt.xticks(rotation=45)
plt.show()
```

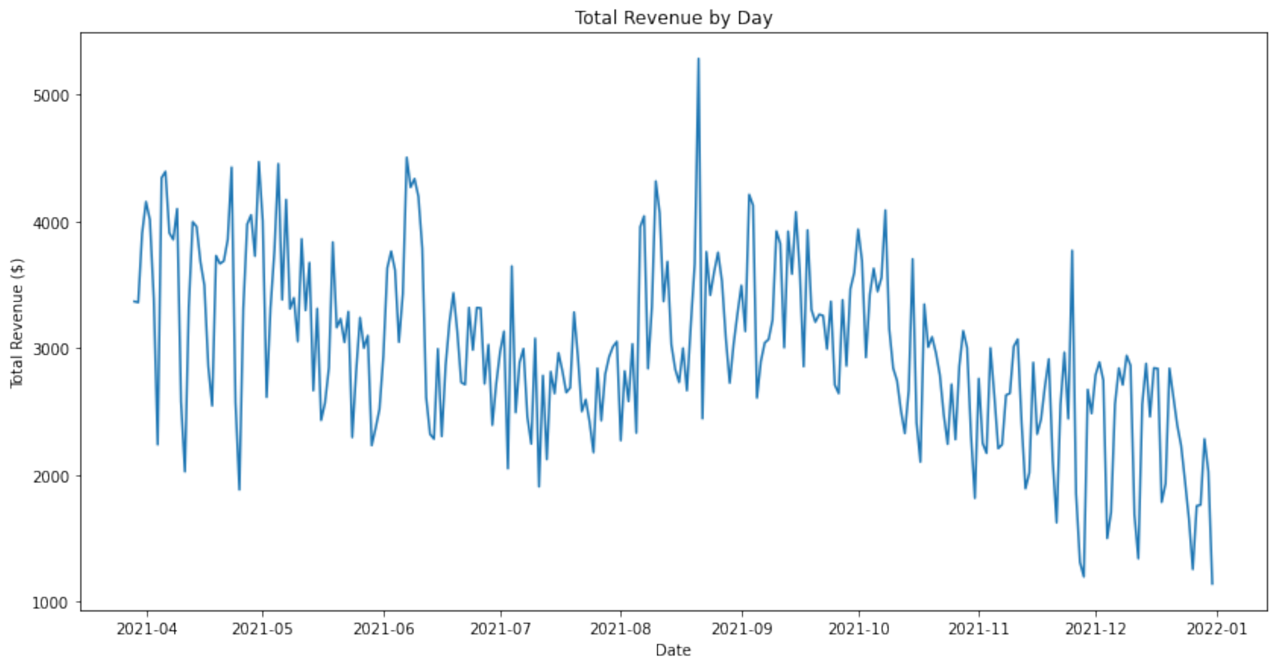

Top 10 High-Demand Pickup Locations
Total Pickups
0
100
200
300
400
500
600
700
800
42
61
68
74
75
76
140
188
238
244
Pickup Location ID

In [0]:
```python
# Identify high-demand pickup locations
demand_analysis = (
    df_2022_cleaned.groupBy("PULocationID")
    .count()
    .withColumnRenamed("count", "total_pickups")
    .orderBy("total_pickups", ascending=False)
)
import matplotlib.pyplot as plt
import seaborn as sns

#PySpark DataFrame is sorted
demand_analysis = demand_analysis.orderBy("total_pickups", ascending=False)

# Convert to Pandas DataFrame and reset the index
demand_analysis_pd = demand_analysis.toPandas().reset_index(drop=True)

# Select the top 10 high-demand pickup locations
top_demand = demand_analysis_pd.head(10)

# Normalize the 'total_pickups' if needed, otherwise skip this step
# top_demand['total_pickups'] = top_demand['total_pickups'] / top_demand['to

# Plotting with seaborn
plt.figure(figsize=(12, 8))
sns.barplot(
    x='PULocationID',
    y='total_pickups',
    data=top_demand,
    palette='viridis'
)

plt.title('Top 10 High-Demand Pickup Locations')
plt.xlabel('Pickup Location ID')
plt.ylabel('Total Pickups')
plt.xticks(rotation=45)
plt.show()
```

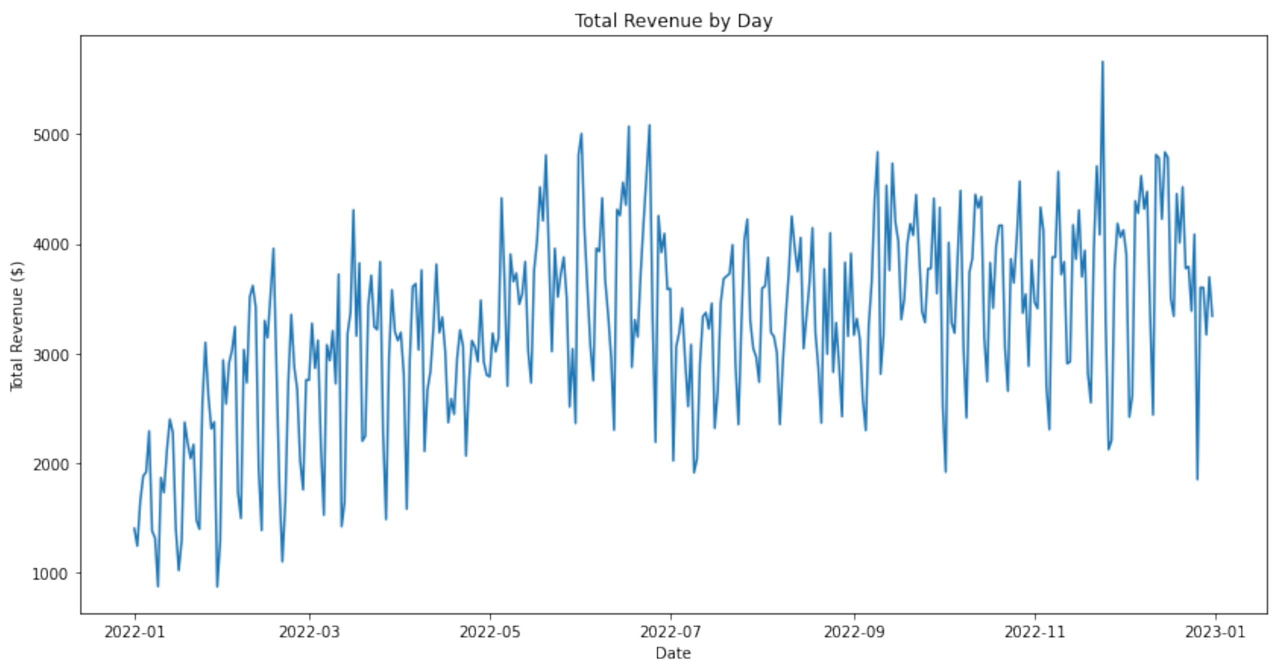


Total Revenue for 2021 and 2022

In [0]:
```
# Ensure df_2021_cleaned is a standalone DataFrame and not a view or copy
#df_2021_cleaned = df_2021_cleaned.copy()

# Convert Pickup_datetime to datetime if it's not already
#df_2021_cleaned['pickup_datetime'] = pd.to_datetime(df_2021_cleaned['pickup

# Create the total_revenue column
#df_2021_cleaned['total_revenue'] = df_2021_cleaned['base_passenger_fare'] +
df_2021_cleaned = df_2021_cleaned.withColumn('total_revenue',col('base_passe

# Group by date and sum the total revenue
revenue_by_day = df_2021_cleaned.groupBy(F.to_date("Pickup_datetime").alias(

# Convert to Pandas DataFrame for plotting (if you want to use matplotlib)
revenue_by_day_pd = revenue_by_day.toPandas()
revenue_by_day_pd = revenue_by_day_pd.set_index('date')
revenue_by_day_pd = revenue_by_day_pd.sort_index()  # Sorting by date if not

# Visualization with Pandas and Matplotlib
plt.figure(figsize=(14, 7))
revenue_by_day_pd['total_revenue'].plot(title='Total Revenue by Day')
plt.xlabel('Date')
plt.ylabel('Total Revenue ($)')
plt.show()
```

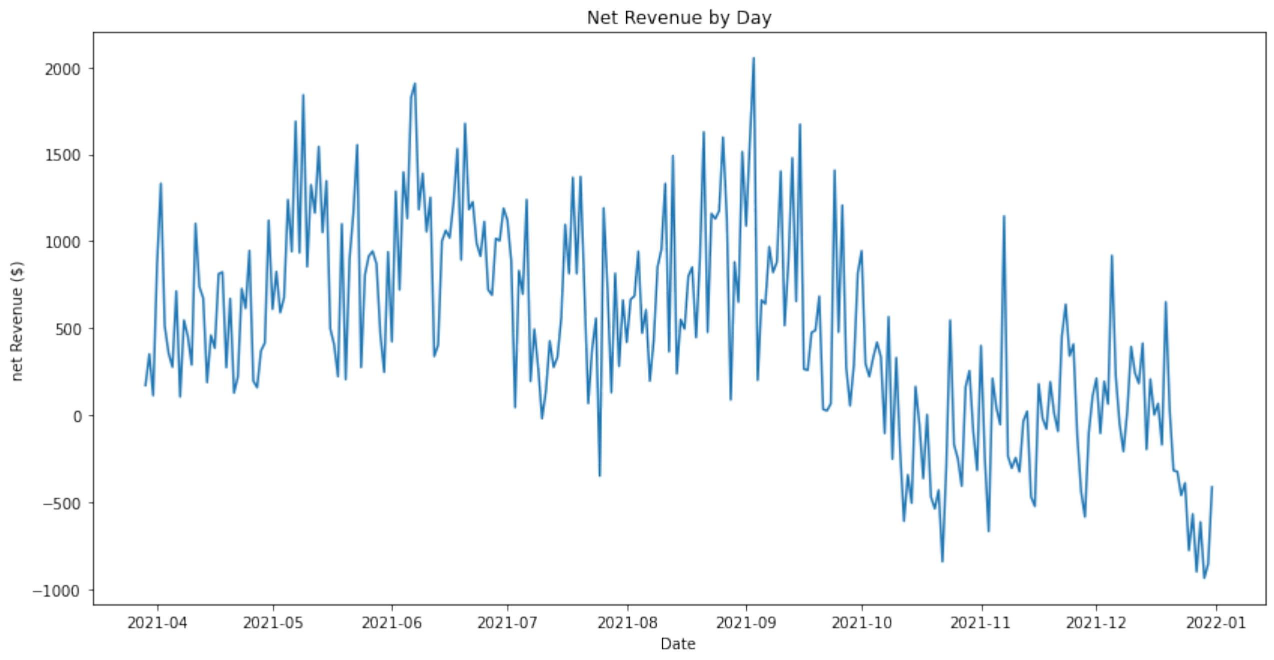


```
In [0]: df_2022_cleaned = df_2022_cleaned.withColumn('total_revenue',col('base_passe

        # Group by date and sum the total revenue
        revenue_by_day = df_2022_cleaned.groupBy(F.to_date("Pickup_datetime").alias(

        # Convert to Pandas DataFrame for plotting (if you want to use matplotlib)
        revenue_by_day_pd = revenue_by_day.toPandas()
        revenue_by_day_pd = revenue_by_day_pd.set_index('date')
        revenue_by_day_pd = revenue_by_day_pd.sort_index()  # Sorting by date if not

        # Visualization with Pandas and Matplotlib
        plt.figure(figsize=(14, 7))
        revenue_by_day_pd['total_revenue'].plot(title='Total Revenue by Day')
        plt.xlabel('Date')
        plt.ylabel('Total Revenue ($)')
        plt.show()
```

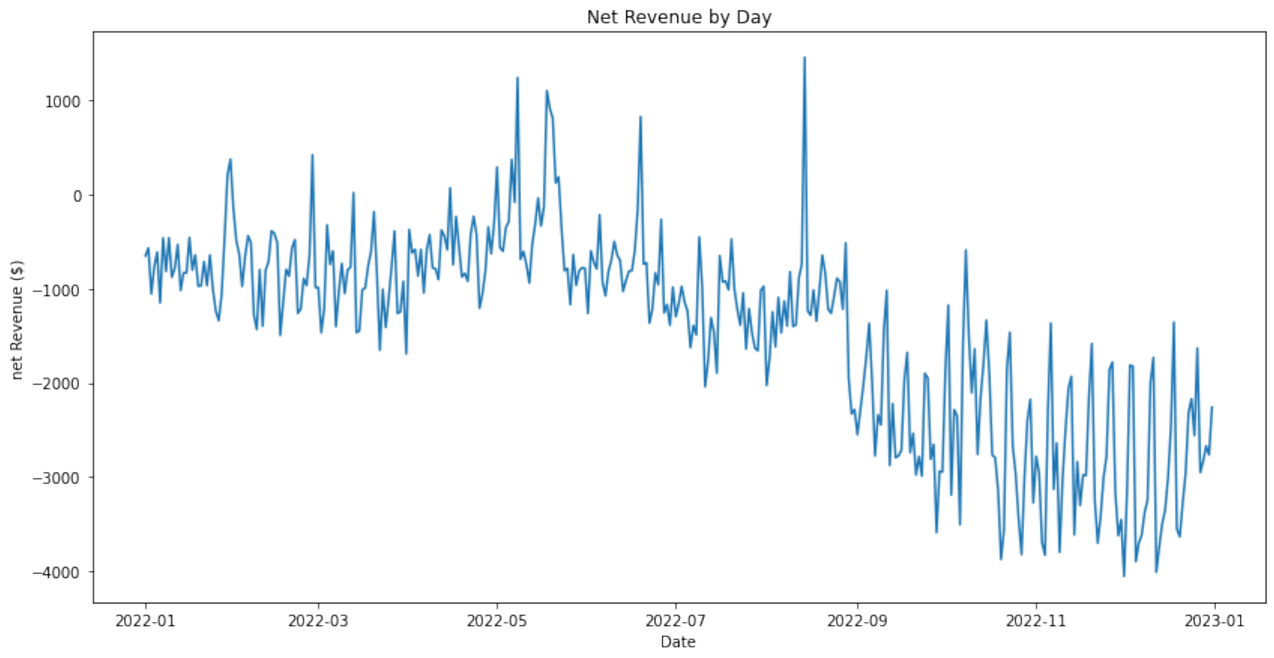


Net Revenue fro 2021 and 2022

In [0]:
```
df_2021_cleaned = df_2021_cleaned.withColumn('net_revenue',col('base_passeng
# Group by date and sum the total revenue
profit_by_day = df_2021_cleaned.groupBy(F.to_date("Pickup_datetime").alias('
# Convert to Pandas DataFrame for plotting (if you want to use matplotlib)
profit_by_day_pd = profit_by_day.toPandas()
profit_by_day_pd = profit_by_day_pd.set_index('date')
profit_by_day_pd = profit_by_day_pd.sort_index()  # Sorting by date if not a

# Visualization with Pandas and Matplotlib
plt.figure(figsize=(14, 7))
profit_by_day_pd['net_revenue'].plot(title='Net Revenue by Day')
plt.xlabel('Date')
plt.ylabel('net Revenue ($)')
plt.show()
```



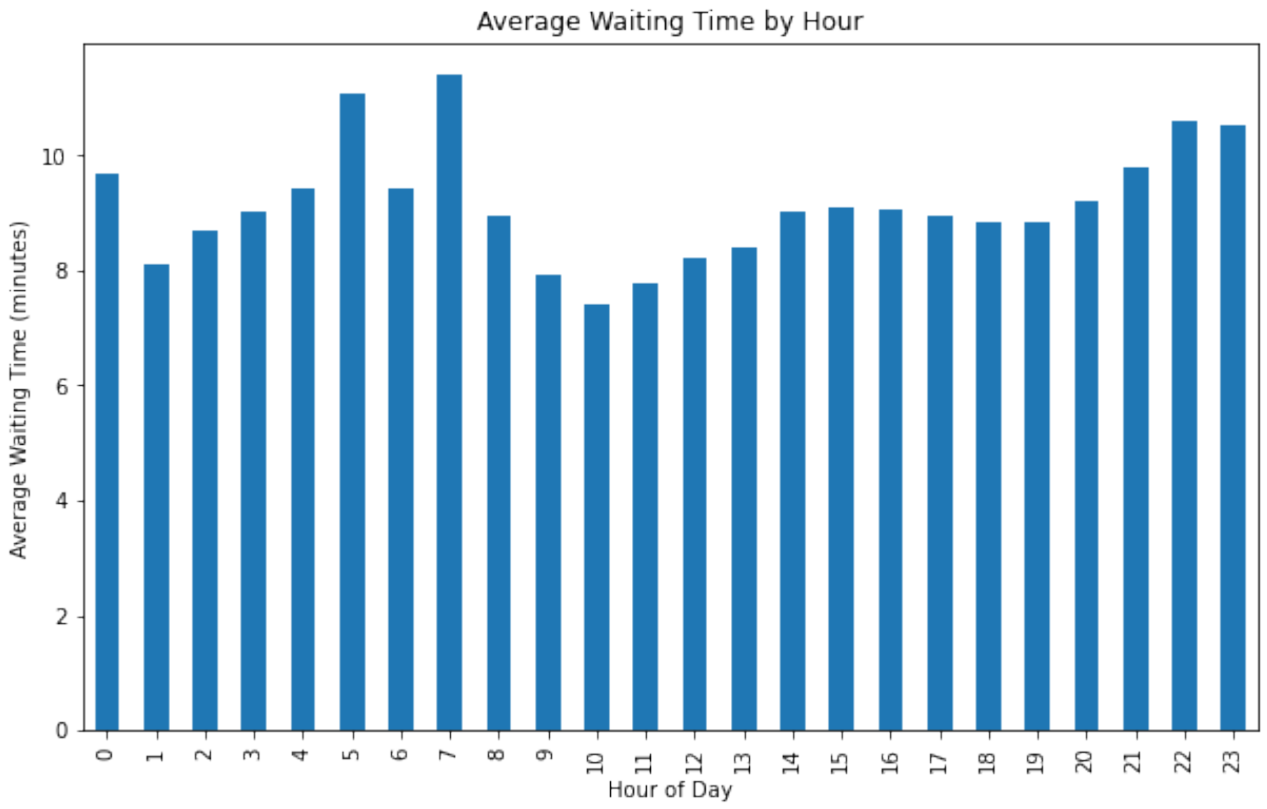
```
In [0]: df_2022_cleaned = df_2022_cleaned.withColumn('net_revenue',col('base_passeng
        # Group by date and sum the total revenue
        profit_by_day = df_2022_cleaned.groupBy(F.to_date("Pickup_datetime").alias('
        # Convert to Pandas DataFrame for plotting (if you want to use matplotlib)
        profit_by_day_pd = profit_by_day.toPandas()
        profit_by_day_pd = profit_by_day_pd.set_index('date')
        profit_by_day_pd = profit_by_day_pd.sort_index()  # Sorting by date if not a

        # Visualization with Pandas and Matplotlib
        plt.figure(figsize=(14, 7))
        profit_by_day_pd['net_revenue'].plot(title='Net Revenue by Day')
        plt.xlabel('Date')
        plt.ylabel('net Revenue ($)')
        plt.show()
```



Average Waiting Time (minutes) for 2021 and 2022

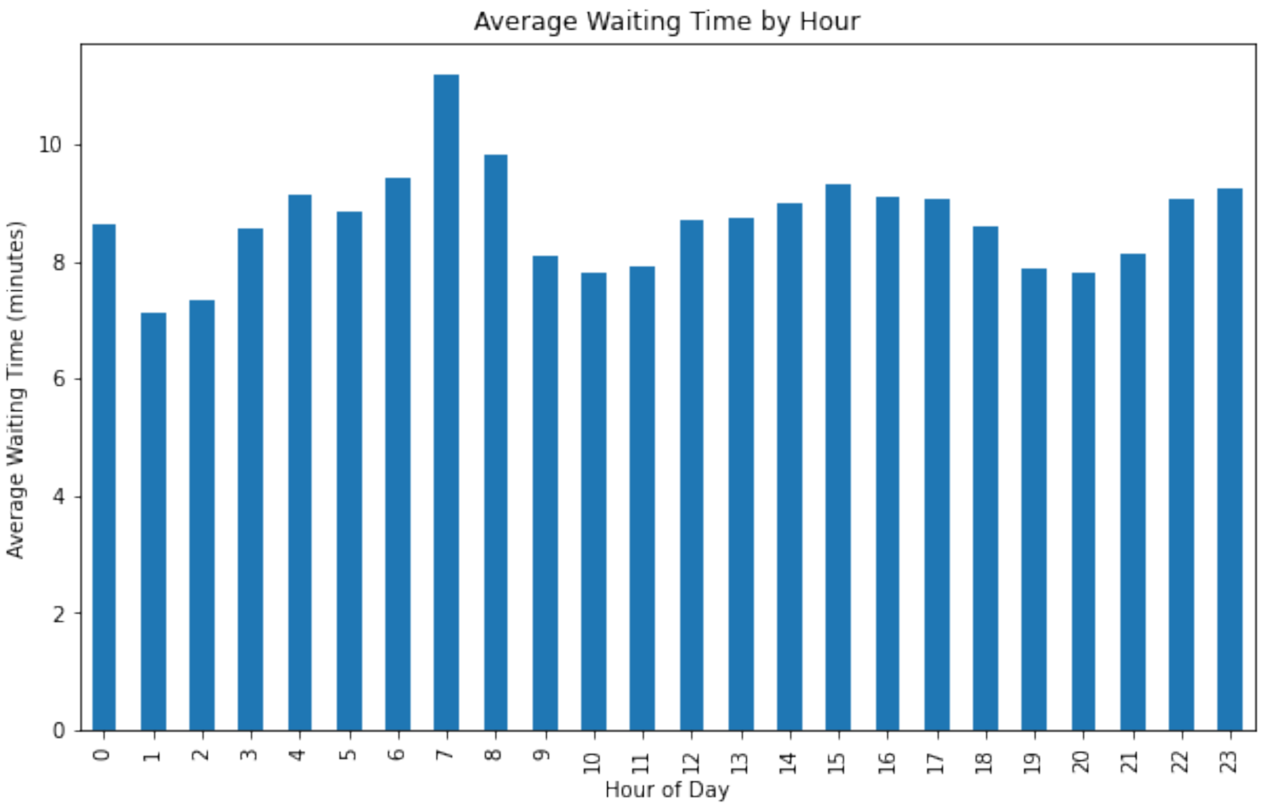
In [0]:

```
# Convert to timestamp if 'request_datetime' and 'on_scene_datetime' are not
df_2021_cleaned = df_2021_cleaned.withColumn(
    "request_datetime",
    F.to_timestamp("request_datetime", "yyyy-MM-dd HH:mm:ss")
)
df_2021_cleaned = df_2021_cleaned.withColumn(
    "on_scene_datetime",
    F.to_timestamp("on_scene_datetime", "yyyy-MM-dd HH:mm:ss")
)

# Calculate waiting time in minutes
df_2021_cleaned = df_2021_cleaned.withColumn(
    "waiting_time",
    (F.unix_timestamp("on_scene_datetime") - F.unix_timestamp("request_datet
)

# Extract hour from 'request_datetime' for grouping
df_2021_cleaned = df_2021_cleaned.withColumn(
    "hour_of_request",
    F.hour("request_datetime")
)

# Group by hour and calculate average waiting time
average_waiting_time_by_hour = df_2021_cleaned.groupBy("hour_of_request").ag
    F.avg("waiting_time").alias("average_waiting_time")
)

# To visualize the data, you would collect the data and use a plotting libra
# However, PySpark DataFrames cannot be directly plotted using Matplotlib wi

# Collect the data into a Pandas DataFrame for plotting
average_waiting_time_by_hour_pd = average_waiting_time_by_hour.toPandas()
average_waiting_time_by_hour_pd = average_waiting_time_by_hour_pd.set_index(
average_waiting_time_by_hour_pd = average_waiting_time_by_hour_pd.sort_index

plt.figure(figsize=(10, 6))
average_waiting_time_by_hour_pd['average_waiting_time'].plot(kind='bar', tit
plt.xlabel('Hour of Day')
plt.ylabel('Average Waiting Time (minutes)')
plt.show()
```


Average Waiting Time by Hour
Average Waiting Time (minutes)
Hour of Day
0
2
4
6
8
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

In [0]:

```
# Convert to timestamp if 'request_datetime' and 'on_scene_datetime' are not
df_2022_cleaned = df_2022_cleaned.withColumn(
    "request_datetime",
    F.to_timestamp("request_datetime", "yyyy-MM-dd HH:mm:ss")
)
df_2022_cleaned = df_2022_cleaned.withColumn(
    "on_scene_datetime",
    F.to_timestamp("on_scene_datetime", "yyyy-MM-dd HH:mm:ss")
)

# Calculate waiting time in minutes
df_2022_cleaned = df_2022_cleaned.withColumn(
    "waiting_time",
    (F.unix_timestamp("on_scene_datetime") - F.unix_timestamp("request_datet
)

# Extract hour from 'request_datetime' for grouping
df_2022_cleaned = df_2022_cleaned.withColumn(
    "hour_of_request",
    F.hour("request_datetime")
)

# Group by hour and calculate average waiting time
average_waiting_time_by_hour = df_2022_cleaned.groupBy("hour_of_request").ag
    F.avg("waiting_time").alias("average_waiting_time")
)

# Add a filter to only include rows where the waiting time is non-negative
df_2022_cleaned = df_2022_cleaned.filter(F.col("waiting_time") >= 0)

# Collect the data into a Pandas DataFrame for plotting
average_waiting_time_by_hour_pd = average_waiting_time_by_hour.toPandas()
average_waiting_time_by_hour_pd = average_waiting_time_by_hour_pd.set_index(
average_waiting_time_by_hour_pd = average_waiting_time_by_hour_pd.sort_index

plt.figure(figsize=(10, 6))
average_waiting_time_by_hour_pd['average_waiting_time'].plot(kind='bar', tit
plt.xlabel('Hour of Day')
plt.ylabel('Average Waiting Time (minutes)')
plt.show()
```


Average Waiting Time by Hour
Average Waiting Time (minutes)
Hour of Day
0
2
4
6
8
10
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23