# 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)
```

```
HV0003
                                          B02682
                                                                                           B02682
                                                                                                                                           2021-01-01 00:2
8:09 \mid 2021-01-01 \quad 00:31:42 \mid 2021-01-01 \quad 00:33:44 \mid 2021-01-01 \quad 00:49:07 \mid 230
                              5.26
                                                       923
                                                                               22.28
                                                                                                                             0.0 | 0.67 | 1.98
2.75
                                                                          0.0 | 14.99
                                              null
                                                                                                                N
                                                                                                                                                              N
                                                                                   N
                                            N
HV0003
                                          B02682
                                                                                           B02682
                                                                                                                                           2021-01-01 00:4
5:56 \mid 2021-01-01 \quad 00:55:19 \mid 2021-01-01 \quad 00:55:19 \mid 2021-01-01 \quad 01:18:21 \mid 152
                              3.65
                                                       1382
                                                                               18.36
                                                                                                                             0.0
                                                                                                                                        0.55 1.63
0.0
                                                                          |0.0 |17.06
                                              null
                                                                                                                N
                                                                                   N
                                            N
HV0003
                                          B02764
                                                                                          B02764
                                                                                                                                           2021-01-01 00:2
1:15 \mid 2021 - 01 - 01 \quad 00:22:41 \mid 2021 - 01 - 01 \quad 00:23:56 \mid 2021 - 01 - 01 \quad 00:38:05 \mid 233 \mid 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2021 - 2
                                                                               14.05
                              3.51
                                                     849
                                                                                                                             0.0 | 0.48 | 1.25
2.75
                                              null
                                                                          0.94 12.98
                                                                                                                N
                                            N
                                                                                   N
                                                                                          B02764
HV0003
                                          B02764
                                                                                                                                           2021-01-01 00:3
9:12 \mid 2021-01-01 \quad 00:42:37 \mid 2021-01-01 \quad 00:42:51 \mid 2021-01-01 \quad 00:45:50 \mid 142
143
                              0.74
                                                     |179
                                                                              7.91
                                                                                                                             0.0 | 0.24 | 0.7
2.75
                                                                          0.0 | 7.41
                                              null
                                                                                                               N
                                                                                                                                                              N
                                            N
                                                                                    N
HV0003
                                          B02764
                                                                                          B02764
                                                                                                                                           2021-01-01 00:4
6:11 \mid 2021 - 01 - 01 \quad 00:47:17 \mid 2021 - 01 - 01 \quad 00:48:14 \mid 2021 - 01 - 01 \quad 01:08:42 \mid 143
                                                       1228
                                                                              27.11
                                                                                                                             0.0 | 0.81 | 2.41
                                                                          0.0 | 22.44
2.75
                                                                                                                N
                                                                                                                                                              | N
                                              null
                                            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
                                                                                                                             0.0 | 0.84 | 2.49
42
                                                                               28.11
2.75
                                                                          0.0 | 28.9
                                                                                                                N
                                                                                                                                                              N
                                              null
l N
                                            N
                                                                                   | N
HV0005
                                          B02510
                                                                                                                                           2021-01-01 00:4
                                                                                          null
                                                        |2021-01-01 00:50:00|2021-01-01 01:04:57|42
0:06 | null
151
                                                       897
                                                                              25.03
                                                                                                                             0.0 | 0.75 | 2.22
0.0
                                                                          0.0 | 15.01
                                              null
                                                                                                                N
                                                                                                                                                              N
N
                                            N
                                                                                   N
HV0003
                                          B02764
                                                                                          B02764
                                                                                                                                           2021-01-01 00:1
0:36 \,|\, 2021 - 01 - 01 \quad 00:12:28 \,|\, 2021 - 01 - 01 \quad 00:14:30 \,|\, 2021 - 01 - 01 \quad 00:50:27 \,|\, 71
                                                     2157
226
                              13.53
                                                                               29.67
                                                                                                                             0.0 | 1.04 | 3.08
0.0
                                              null
                                                                          0.0 | 34.2
                                                                                                                N
                                                                                                                                                              l N
                                            N
                                                                                          B02875
                                                                                                                                           2021-01-01 00:2
HV0003
                                          B02875
1:17 \mid 2021-01-01 \quad 00:22:25 \mid 2021-01-01 \quad 00:22:54 \mid 2021-01-01 \quad 00:30:20 \mid 112
                              1.6
                                                                                                                             0.0 | 0.21 | 0.61
255
                                                       446
                                                                               6.89
0.0
                                              null
                                                                          0.0 | 6.26
                                                                                                                N
                                                                                                                                                              N
                                            N
                                                                                   lΝ
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
                                                                          |2.82|10.99
                                                                                                                                                              N
                                              null
                                                                                                                N
                                            | N
                                                                                    N
                                                                                          B02875
HV0003
                                          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
```

```
|5.74
                        1257
                                   17.18
                                                        0.0 | 0.52 | 1.52
198
2.75
                     null
                                 0.0 | 17.61
                                                  N
                    N
                                     N
                   B02835
                                        B02835
                                                               2021-01-01 00:2
HV0003
2:58 \mid 2021-01-01 \quad 00:27:01 \mid 2021-01-01 \quad 00:29:04 \mid 2021-01-01 \quad 00:36:27 \mid 113
                                   8.18
                                                        |0.0 |0.25|0.73
             1.8
                        443
                                 0.0 | 6.12
2.75
                                                  N
                     null
                    N
                                     N
HV0003
                   B02835
                                        B02835
                                                               2021-01-01 00:4
6:44 \mid 2021-01-01 \quad 00:47:49 \mid 2021-01-01 \quad 00:48:56 \mid 2021-01-01 \quad 00:59:12 \mid 239
175
             12.9
                         616
                                   13.1
                                                        0.0 | 0.45 | 1.17
2.75
                                 0.94 | 8.77
                     null
                                                  N
                    N
                                     N
HV0004
                   B02800
                                                               2021-01-01 00:1
                                         null
2:50 | null
                         |2021-01-01 00:15:24|2021-01-01 00:38:31|181
237
              9.66
                        1387
                               32.95
                                                        10.0 | 0.0 | 2.34
2.75
                                 0.0 | 21.1
                                                  N
                     null
N
                    N
                                     N
                   B02800
                                                               2021-01-01 00:3
HV0004
                                         null
5:32 | null
                         |2021-01-01 00:45:00|2021-01-01 01:06:45|236
68
                        1305
                                   22.91
                                                        0.0 | 0.0 | 1.63
             4.38
                                 |3.43|15.82
2.75
                                                  N
                                                                       | N
                     null
l 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
             2.03 | 373
                                   7.84
                                                        0.0 | 0.42 | 0.7
2.75
                                 |2.82|6.93
                     null
                                                  N
                                                                       N
                                     N
                    N
                                         B02682
HV0003
                   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
             3.08 | 789
                                   13.2
                                                        0.0 | 0.4 | 1.17
                                 0.0 | 11.54
2.75
                     null
                                                  N
                                     N
                    N
                                         B02682
                                                               2021-01-01 00:4
HV0003
                   B02682
4:56 \mid 2021-01-01 \quad 00:49:55 \mid 2021-01-01 \quad 00:50:49 \mid 2021-01-01 \quad 00:55:59 \mid 17
             |1.17 | 310
                                  7.91
                                                        0.0 | 0.24 | 0.7
0.0
                                 0.0 | 6.94
                                                  N
                     null
                                                                       N
                                     N
HV0005
                   B02510
                                                               2021-01-01 00:0
                                         null
5:04|null
                         |2021-01-01 00:08:40|2021-01-01 00:39:39|62
                                  31.18
29
              10.852
                        1859
                                                        0.0 | 0.94 | 2.77
0.0
                                  0.0 | 27.61
                                                                       N
                     null
                                                  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
                                  28.67
             3.52
                      1072
                                                         0.0 | 0.86 | 2.54
0.0
                                  |0.0 | 17.64
                     null
                                                  N
                                      l N
```

# 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
        #combined dataframe with rows from all the 12 datasets
In [0]:
        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
                                          B02682
                     HV0003
                                                                B02682 | 2021-01-01 00:2
        8:09 \mid 2021-01-01 \quad 00:31:42 \mid 2021-01-01 \quad 00:33:44 \mid 2021-01-01 \quad 00:49:07 \mid
                                                                                    23
                           5.26
                    166
                                        923
                                                            22.28
                                                                    0.0 | 0.67 |
                                                                                   1.98
                          2.75
                                      null| 0.0|
                                                      14.99
                                                                              N
        N \mid
                                             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 |
                                                            18.36
                    167
                            3.65
                                        1382
                                                                    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
        14.05
                                                                    0.0 | 0.48 |
                                                                                   1.25
                         3.51
                                         849
                          2.75
                                      null|0.94|
                                                                              N
                                                      12.98
        N
                                             N
                                                             N
                                                                B02764 | 2021-01-01 00:3
                    HV0003
                                          B02764
        9:12 | 2021-01-01 | 00:42:37 | 2021-01-01 | 00:42:51 | 2021-01-01 | 00:45:50 |
                                                                                    14
                              0.74
                                         179
                                                             7.91
                                                                                    0.7
        2 |
                                      null| 0.0|
                          2.75
                                                       7.41
                                                                              N
        N
                                             N
                                                             N \mid
                                          B02764
                                                                B02764 | 2021-01-01 00:4
                    HV0003
        6:11 \mid 2021-01-01 \quad 00:47:17 \mid 2021-01-01 \quad 00:48:14 \mid 2021-01-01 \quad 01:08:42 \mid
                                                                                    14
        3 |
                               9.2
                                        1228
                                                            27.11
                                                                    0.0|0.81|
                                      null| 0.0|
                          2.75
                                                      22.44
                                                                              N
                                             N
        N
                     HV0005 |
                                          B02510
                                                                  null | 2021-01-01 00:0
                             null|2021-01-01 00:06:59|2021-01-01 00:43:01|
        4:00
        8 |
                                                            28.11 | 0.0 | 0.84 |
                     42
                             9.725
                                        2162
                                                                                   2.49
                          2.75
                                      null| 0.0|
                                                       28.9
                                                                              N
        N
                            N
                                             N
                     HV0005
                                          B02510
                                                                  null 2021-01-01 00:4
                             null|2021-01-01 00:50:00|2021-01-01 01:04:57|
        0:06
```

```
25.03 | 0.0 | 0.75 |
        151
              2.469
                      897|
                                                         2.22
              0.0| null| 0.0|
                                  15.01
               N
                      N l
                                        N
N
                   B02764|
        HV0003|
                                        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
        226 | 13.53 | 2157 | 29.67 | 0.0 | 1.04 |
              0.0| null| 0.0|
                                  34.2
                                                     N
        | N| N|
HV0003| B02875| B02875|2021-01-01 00:2
N \mid
1:17 | 2021-01-01 00:22:25 | 2021-01-01 00:22:54 | 2021-01-01 00:30:20 | 11
        255 | 1.6 | 446 |
                                        6.89 | 0.0 | 0.21 |
              0.0 | null | 0.0 |
                                  6.26
Νl
                          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
        232 | 3.2 | 800 | 11.51 | 0.0 | 0.53 |
             2.75 | null|2.82|
                                   10.99
N
                       N
        HV0003 | B02875 | B02875 | 2021-01-01 00:5
3:31 \mid 2021-01-01 \quad 00:56:21 \mid 2021-01-01 \quad 00:56:45 \mid 2021-01-01 \quad 01:17:42 \mid 23
       198 | 5.74 | 1257 | 17.18 | 0.0 | 0.52 |
             2.75 | null | 0.0 | 17.61 |
         | N| N|
HV0003| B02835| B02835|2021-01-01 00:2
N
2:58 \mid 2021-01-01 \mid 00:27:01 \mid 2021-01-01 \mid 00:29:04 \mid 2021-01-01 \mid 00:36:27 \mid 11
         48 | 1.8 | 443 | 8.18 | 0.0 | 0.25 |
             2.75 | null | 0.0 | 6.12 |
         | N| N|
HV0003| B02835| B02835|2021-01-01 00:4
Νl
6:44|2021-01-01 00:47:49|2021-01-01 00:48:56|2021-01-01 00:59:12|
         75 | 2.9 | 616 | 13.1 | 0.0 | 0.45 |
             2.75 | null | 0.94 |
                                   8.77
               | N| N|
| B02800| null|2021-01-01 00:1
         HV0004
              null|2021-01-01 00:15:24|2021-01-01 00:38:31| 18
         237
              9.66 | 1387 | 32.95 | 0.0 | 0.0 |
1 |
                                                        2.34
             2.75 | null| 0.0|
                                    21.1
                         N| N|
B02800| null|2021-01-01 00:3
                      N
         HV0004|
              null|2021-01-01 00:45:00|2021-01-01 01:06:45| 23
5:32
              4.38 | 1305 | 22.91 | 0.0 | 0.0 | 1.63
         68
             2.75 | null|3.43|
                                   15.82
               N | N | N | B02682 | 2021-01-01 00:1
         HV0003
0:22 | 2021-01-01 | 00:11:03 | 2021-01-01 | 00:11:53 | 2021-01-01 | 00:18:06 |
         148 | 2.03 | 373 |
                                        7.84 | 0.0 | 0.42 |
                                                          0.7
                                   6.93
             2.75 | null|2.82|
                                                      N
                         N |
B02682 |
         HV0003|
                                          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|
        80 | 3.08 | 789 |
                                        13.2 | 0.0 | 0.4 | 1.17
             2.75 | null | 0.0 |
                                   11.54
                                                      N
                          N |
                         B02682|
         HV0003|
                                           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|
        217 | 1.17 | 310 | 7.91 | 0.0 | 0.24 |
                                                          0.7
                      null | 0.0 | 6.94 |
```

```
N
                                                    N
                                                                     N
                       HV0005
                                                                           null | 2021-01-01 00:0
                                                B02510
                                 null 2021-01-01 00:08:40 2021-01-01 00:39:39
         5:04
         2
                       29
                                10.852
                                             1859
                                                                    31.18
                                                                             0.0 | 0.94 |
                                                                                              2.77
                               0.0
                                           null | 0.0|
                                                             27.61
         N
                                N
                                                    N
                                                                     N
                       HV0003
                                                                         B02836 2021-01-01 00:4
                                                B02836
         0:44 \, | \, 2021 - 01 - 01 \quad 00:53:34 \, | \, 2021 - 01 - 01 \quad 00:53:48 \, | \, 2021 - 01 - 01 \quad 01:11:40 \, |
         2
                       22
                                  3.52
                                              1072
                                                                    28.67
                                                                             0.0 | 0.86 |
                                                                                              2.54
                               0.0
                                           null | 0.0|
                                                             17.64
         N
                                                    N
                                                                     Ν
         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|
                                                       B03404 | 2022-01-01 00:0
                                 B03404
5:31 | 2022-01-01 00:05:40 | 2022-01-01 00:07:24 | 2022-01-01 00:18:28 |
           161
                 1.18
                              664
                                                    24.9 | 0.0 | 0.75 |
                                                                          2.21
                              0.0 | 0.0 |
                 2.75
                                             23.03
ИΙ
                                    ИΙ
            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 |
           161 0.82
                               460
                                                   11.97 | 0.0 | 0.36 |
                              0.0 | 0.0 |
                 2.75
                                             12.32
                                                                     N
                                    Ν
Νl
            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 |
                                                   29.82 | 0.0 | 0.89 |
           161 1.18
                              595
                 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
           229 | 1.65 |
                               303
                                                    7.91 | 0.0 | 0.24 |
                                                                           0.7
                              0.0| 0.0|
                                               6.3
                 2.75
N \mid
                                    N
                                                       B03404 | 2022-01-01 00:2
            HV0003
                                 B03404
5:45|2022-01-01 00:26:01|2022-01-01 00:28:01|2022-01-01 00:35:42|
           141
                     1.65
                               461
                                                    9.44 | 0.0 | 0.28 |
                                                                          0.84
                 2.75
                              0.0 | 0.0 |
                                              7.44
                                                                     N
                                    Ν
Νl
            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|
                                                   17.67 | 0.0 | 0.53 |
            79
                               762
                     4.51
                                                                          1.57
                 2.75
                              0.0 | 0.0 |
                                             12.25
                                                                     N \mid
                                    N \mid
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|
                                                   16.68 | 0.0 | 0.5 |
           140
                               931
                     3.68
                                                                          1.48
                              0.0 | 0.0 |
                 2.75
                                             12.75
                                                                     N \mid
                                    N \mid
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|
1 |
            75
                     2.77
                               843|
                                                   14.41 | 0.0 | 0.43 |
                  0.0
                              0.0 | 4.0 |
                                             11.47
                                    Ν
Νl
                                                       B03404 | 2022-01-01 00:2
            HV0003
                                B03404
7:54|2022-01-01 00:30:26|2022-01-01 00:32:25|2022-01-01 00:44:15|
                                                                         0.94
           229
                     2.04
                               710|
                                                   10.64 | 0.0 | 0.32 |
                              0.0 | 0.0 |
                                              9.55
                                                                     N \mid
                 2.75
N
                                    N
            HV0003|
                                 B03404
                                                       B03404 | 2022-01-01 00:4
4:59 \mid 2022-01-01 \quad 00:48:23 \mid 2022-01-01 \quad 00:50:23 \mid 2022-01-01 \quad 01:15:30 \mid
           169 | 8.79 |
                              1507|
                                                  107.56 | 0.0 | 0.83 |
                                                                         2.45
                 2.75
                              0.0 | 0.0 |
                                             23.67
                                                                     N \mid
N
                                    N
            HV0003|
                                 B03404
                                                       B03404 | 2022-01-01 00:1
```

```
3:49 \mid 2022-01-01 \quad 00:16:15 \mid 2022-01-01 \quad 00:17:02 \mid 2022-01-01 \quad 00:40:09 \mid
                                                                            26
           223
                    11.29
                                1387
                                                    34.9 | 0.0 | 1.05 |
                                                                           3.1
                 2.75
                               0.0 | 0.0 |
                                             25.17
                                                                      N \mid
N
                                     N
                                                    N
                                  B03404
            HV0003
                                                       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
                                 251
                                                    7.91 | 0.0 | 0.24 |
                                                                           0.7
           223
                     0.87
                               0.0 | 0.0 |
                  0.0
                                              6.51
N
                                     N \mid
                                                    N
                                                       B03404 | 2022-01-01 00:4
            HV0003
                                  B03404
5:50|2022-01-01 00:52:15|2022-01-01 00:52:29|2022-01-01 01:01:48|
                                                                            22
                                                    9.71 | 0.0 | 0.29 |
             7 |
                     1.89
                                 559
                  0.0
                               0.0 | 0.0 |
                                              7.89
                                     N
                                                    N \mid
N
            HV0005|
                                  B03406
                                                         null|2022-01-01 00:3
6:54
                    null|2022-01-01 00:45:34|2022-01-01 00:54:11|
           148
                    3.585
                                 810
                                                   27.02 | 0.0 | 0.81 |
                 2.75
                               0.0 | 0.0 |
                                             21.08
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
                                                   30.37 | 0.0 | 0.91 |
                                                                           2.7
           243
                    9.2
                               1384
                 2.75
                               0.0 | 0.0 |
                                             22.69
N
                                     N
                                                    N
            HV0003
                                  B03404
                                                       B03404 2022-01-01 00:5
3:32 \mid 2022-01-01 \quad 00:58:14 \mid 2022-01-01 \quad 00:58:26 \mid 2022-01-01 \quad 01:07:23 \mid
                                                   15.02 | 0.0 | 0.45 |
           127
                     1.88
                                 537
                                                                          1.33
                  0.0
                               0.0 | 2.0 |
                                             10.02
                                                                      N
N \mid
                                     N
                                                    N \mid
            HV0005|
                                  B03406
                                                         null 2022-01-01 00:2
                    null|2022-01-01 00:34:59|2022-01-01 00:50:15|
8:07
                                                   20.47 | 0.0 | 0.61 |
           170
                    2.699
                                 916
                                                                          1.82
                               0.0 | 0.0 |
                 2.75
                                             10.66
                   N
                                     N
Νl
                                                    N
            HV0005
                                  B03406
                                                         null 2022-01-01 00:4
                    null|2022-01-01 00:58:09|2022-01-01 01:28:23|
                                                                            17
                    8.037
                               1904
                                                   24.63 | 20.0 | 1.34 |
                                                                           0.0
0 |
           265
                  0.0
                               0.0 0.0
                                             26.63
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 |
                                                                          0.87
           179
                     1.69
                                 637
                                                    9.81 | 0.0 | 0.29 |
3 |
                  0.0
                               0.0 | 0.0 |
                                              9.55
                                                                      N
ИΙ
                                     N \mid
                                                    N \mid
            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
```

```
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]:
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\_datetime', 'timestamp'), ('PULocationID', 'bigint'), ('DOLocationID', 'bigint'), ('trip\_miles', 'double'), ('trip\_time', 'bigint'), ('base\_passenger\_fare', 'double'), ('tolls', 'double'), ('bcf', 'double'), ('sales\_tax', 'double'), ('c ongestion\_surcharge', 'double'), ('airport\_fee', 'double'), ('tips', 'double'), ('driver\_pay', 'double'), ('shared\_request\_flag', 'string'), ('shared\_match\_flag', 'string'), ('access\_a\_ride\_flag', 'string'), ('wav\_request\_flag')

```
g', 'string'), ('wav_match_flag', 'string')]
|hvfhs license num|dispatching base num|originating base num|
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
                                                            B02682 | 2021-01-01 00:2
                                    B02682
8:09 \mid 2021-01-01 \quad 00:31:42 \mid 2021-01-01 \quad 00:33:44 \mid 2021-01-01 \quad 00:49:07 \mid
                                                                                  23
                       5.26
                                   923
                                                       22.28 | 0.0 | 0.67 |
                   2.75
                                null| 0.0|
                                                14.99
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 |
                                 1382
                                                       18.36 | 0.0 | 0.55 |
                       3.65
                                null| 0.0|
                    0.0
                                                17.06
N \mid
                                        N |
                                                           B02764 | 2021-01-01 00:2
             HV0003
                                    B02764
1:15 \mid 2021-01-01 \quad 00:22:41 \mid 2021-01-01 \quad 00:23:56 \mid 2021-01-01 \quad 00:38:05 \mid
3 |
            142
                       3.51
                                   849
                                                       14.05 | 0.0 | 0.48 |
                                                                                1.25
                                null | 0.94 |
                                                12.98
                   2.75
                                                                           N
ИΙ
                                       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 |
            143
                                                        7.91 | 0.0 | 0.24 |
                                                                                 0.7
                       0.74
                                   179
                   2.75
                                null| 0.0|
                                                  7.41
                                                                           N
N
                                       N
                                                        N
                                    B02764
                                                           B02764 2021-01-01 00:4
             HV0003
6:11 | 2021-01-01 00:47:17 | 2021-01-01 00:48:14 | 2021-01-01 01:08:42 |
             78
                                  1228
                                                       27.11 | 0.0 | 0.81 |
                        9.2
                                                                                2.41
                                null| 0.0|
                   2.75
                                                22.44
                                                                           N
N
                                       N \mid
             HV0005|
                                    B02510
                                                              null | 2021-01-01 00:0
                      null 2021-01-01 00:06:59 2021-01-01 00:43:01
4:00
8 |
             42
                      9.725
                                  2162
                                                       28.11 | 0.0 | 0.84 |
                                                                                2.49
                   2.75
                                null| 0.0|
                                                  28.9
                                                                           N \mid
Νl
                                       N
             HV0005
                                    B02510
                                                              null 2021-01-01 00:4
                      null|2021-01-01 00:50:00|2021-01-01 01:04:57|
0:06
                                   897|
                      2.469
                                                       25.03 | 0.0 | 0.75 |
                                                                               2.22
2
            151
                                null| 0.0|
                                                                           N |
                    0.0
                                                15.01
N
                                        N
             HV0003
                                    B02764
                                                           B02764 2021-01-01 00:1
0:36 \mid 2021-01-01 \quad 00:12:28 \mid 2021-01-01 \quad 00:14:30 \mid 2021-01-01 \quad 00:50:27 \mid
                                                       29.67 | 0.0 | 1.04 |
            226
                      13.53
                                  2157
                                                                                3.08
                    0.0
                                null| 0.0|
                                                  34.2
                                                                           N \mid
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 |
           255
                       1.6
                                 446
                                                     6.89 | 0.0 | 0.21 |
                                                                            0.61
                              null| 0.0|
                   0.0
                                               6.26
                                                                       N \mid
N
                                     N
                                                     N
                                  B02875
            HV0003
                                                        B02875 | 2021-01-01 00:3
6:57 \mid 2021-01-01 \quad 00:38:09 \mid 2021-01-01 \quad 00:40:12 \mid 2021-01-01 \quad 00:53:31 \mid
                                                                              25
                                 800
                                                    11.51 | 0.0 | 0.53 |
                                                                            1.03
           232
                       3.2
                              null|2.82|
                  2.75
                                              10.99
                                     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
                                                    17.18 | 0.0 | 0.52 |
                    5.74
                                1257
                  2.75
                              null| 0.0|
                                              17.61
                                     N \mid
                                                     N \mid
N
                                                        B02835 | 2021-01-01 00:2
            HV0003
                                  B02835
11
                    1.8
                                 443
                                                     8.18 | 0.0 | 0.25 |
            48
                                               6.12
                  2.75
                              null| 0.0|
N
                                     N
            HV0003
                                  B02835
                                                         B02835 | 2021-01-01 00:4
6:44 \mid 2021-01-01 \quad 00:47:49 \mid 2021-01-01 \quad 00:48:56 \mid 2021-01-01 \quad 00:59:12 \mid
                                                                              23
                     2.9
                                                     13.1 | 0.0 | 0.45 |
            75
                                 616
                                                                            1.17
                  2.75
                              null|0.94|
                                               8.77
N
                                     N
                                                     N
            HV0004
                                  B02800
                                                          null 2021-01-01 00:1
                    null|2021-01-01 00:15:24|2021-01-01 00:38:31|
2:50
                                                    32.95 | 0.0 | 0.0 |
1 |
           237
                     9.66
                                1387
                                                                            2.34
                  2.75
                              null| 0.0|
                                               21.1
                    N
                                     N
                                                     N \mid
N
            HV0004
                                  B02800
                                                          null 2021-01-01 00:3
                    null 2021-01-01 00:45:00 2021-01-01 01:06:45
                                                    22.91 | 0.0 | 0.0 |
            68
                    4.38
                                1305
                                                                            1.63
                              null|3.43|
                  2.75
                                              15.82
                    N
                                     N
Νl
            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|
           148
                                 373
                                                     7.84 | 0.0 | 0.42 |
                                                                             0.7
                   2.03
                  2.75
                              null | 2.82 |
                                               6.93
N
                                     N
            HV0003
                                  B02682
                                                        B02682 2021-01-01 00:2
5:00 \mid 2021-01-01 \quad 00:26:31 \mid 2021-01-01 \quad 00:28:31 \mid 2021-01-01 \quad 00:41:40 \mid
                                                                              7
9 |
            80
                   3.08
                                 789
                                                     13.2 | 0.0 | 0.4 |
                                                                            1.17
                              null| 0.0|
                  2.75
                                              11.54
                                                                       ИΙ
ИΙ
                                     ΝΙ
                                                     N \mid
            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|
                                                     7.91 | 0.0 | 0.24 |
7 |
           217
                      1.17
                                 310
                                                                             0.7
                   0.0
                              null| 0.0|
                                               6.94
                                                                       N
N
                                     N
                                                     N
            HV0005
                                  B02510
                                                         null 2021-01-01 00:0
                    null|2021-01-01 00:08:40|2021-01-01 00:39:39|
5:04
                                                                               6
2 |
            29
                    10.852
                                1859|
                                                     31.18 | 0.0 | 0.94 |
                              null| 0.0|
                   0.0
                                              27.61
                                                                       N
                    N
N
                                     N
                                                     N
            HV0003
                                 B02836
                                                        B02836 | 2021-01-01 00:4
0:44 \mid 2021-01-01 \quad 00:53:34 \mid 2021-01-01 \quad 00:53:48 \mid 2021-01-01 \quad 01:11:40 \mid
2 |
            22
                      3.52
                               1072
                                                   28.67 | 0.0 | 0.86 |
```

```
ИΙ
                                          N \mid
        -----+
        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]: 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)
```

null| 0.0|

17.64

N

0.0

```
-- 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|
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 \mid 2021-01-01 \quad 00:31:42 \mid 2021-01-01 \quad 00:33:44 \mid 2021-01-01 \quad 00:49:07 \mid
                                                                          23
           166
                            923.0
                                                  22.28 | 0.0 | 0.67 |
                     5.26
                                                                        1.98
                 2.75
                             null| 0.0|
                                            14.99
                                                                false
false
                   null
                                    false
                                                   false
           HV0003
                                                      B02682 2021-01-01 00:4
                                 B02682
5:56 | 2021-01-01 00:55:19 | 2021-01-01 00:55:19 | 2021-01-01 01:18:21 |
                                                                          15
                           1382.0
                                                  18.36 | 0.0 | 0.55 |
           167
                    3.65
                                                                        1.63
                             null| 0.0|
                                            17.06
                  0.0
                                                                false
false
                    null
                                    false
                                                   false
           HV0003
                                 B02764
                                                      B02764 2021-01-01 00:2
23
           142
                    3.51
                              849.0
                                                  14.05 | 0.0 | 0.48 |
                                                                        1.25
                             null|0.94|
                                            12.98
                 2.75
                                                                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 |
                              179.0
           143
                    0.74
                                                   7.91
                                                          0.0|0.24|
                                                                         0.7
                 2.75
                             null| 0.0|
                                             7.41
                                                                false
                    null
                                    false
                                                   false
false
           HV0003
                                                      B02764 | 2021-01-01 00:4
                                 B02764
6:11 | 2021-01-01 | 00:47:17 | 2021-01-01 | 00:48:14 | 2021-01-01 | 01:08:42 |
                                                                          14
                           1228.0
            78
                      9.2
                                                  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
                    null | 2021-01-01 00:06:59 | 2021-01-01 00:43:01 |
4:00
                                                                           8
            42
                    9.725
                             2162.0
                                                  28.11
                                                          0.0 | 0.84 |
                                                                        2.49
8 |
                 2.75
                            null| 0.0|
                                             28.9
                                                                false
```

```
false
                             false| false|
false
         HV0005|
                          B02510
                                              null|2021-01-01 00:4
               null|2021-01-01 00:50:00|2021-01-01 01:04:57|
0:06
         151
                2.469 | 897.0 | 25.03 | 0.0 | 0.75 |
                        null| 0.0|
               0.0
                                    15.01
                             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
         226
                13.53 | 2157.0 | 29.67 | 0.0 | 1.04 |
                                                            3.08
                        null| 0.0|
               0.0
                                     34.2
                             false|
false
                null
                                          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 |
         255
               1.6 446.0
                                         6.89 | 0.0 | 0.21 |
                                                            0.61
                        null| 0.0|
               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|
         232 | 3.2 | 800.0 | 11.51 | 0.0 | 0.53 |
              2.75|
                        null|2.82|
                                    10.99
                                                     false
                        false|
B02875|
false
                null
                                          false
         HV0003|
                                           B02875 | 2021-01-01 00:5
3:31 \mid 2021-01-01 \quad 00:56:21 \mid 2021-01-01 \quad 00:56:45 \mid 2021-01-01 \quad 01:17:42 \mid
         198 | 5.74 | 1257.0 | 17.18 | 0.0 | 0.52 |
              2.75
                        null | 0.0 | 17.61 |
                                                     false
                        false|
B02835|
                null
                                          false
false
                                           B02835 | 2021-01-01 00:2
         HV0003|
2:58|2021-01-01 00:27:01|2021-01-01 00:29:04|2021-01-01 00:36:27|
         48 | 1.8 | 443.0 |
                                         8.18 | 0.0 | 0.25 |
                                                            0.73
                        null | 0.0 | 6.12 |
              2.75
                                                     false
                        false|
B02835|
false
                null
                                          false
                                          B02835 | 2021-01-01 00:4
         HV0003
6:44|2021-01-01 00:47:49|2021-01-01 00:48:56|2021-01-01 00:59:12|
         75 | 2.9 |
                                          13.1 | 0.0 | 0.45 |
                       616.0
                                                            1.17
                        null|0.94|
                                     8.77|
              2.75
                                                    false
                        false|
B02800|
false
                null
                                          false
         HV0004|
                                          null|2021-01-01 00:1
               null|2021-01-01 00:15:24|2021-01-01 00:38:31|
2:50
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
                                             null|2021-01-01 00:3
         HV0004|
                           B02800
               null|2021-01-01 00:45:00|2021-01-01 01:06:45|
               4.38
                        1305.0|
                                         22.91 | 0.0 | 0.0 |
          68
                        null|3.43|
              2.75
                                    15.82
                                                    false
                                          false
               false
                              false
false
                                    B02682 | 2021-01-01 00:1
         HV0003|
                        B02682
0:22|2021-01-01 00:11:03|2021-01-01 00:11:53|2021-01-01 00:18:06|
                                          7.84 | 0.0 | 0.42 |
         148 | 2.03 | 373.0 |
                                                            0.7
              2.75
                        null|2.82|
                                     6.93
                                                   false
         null| false|
HV0003| B02682|
false
                                          false
                                            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|
         80 | 3.08 | 789.0 |
                                          13.2 | 0.0 | 0.4 |
                                                           1.17
                       null| 0.0|
              2.75
                                    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 |
        7 |
                                                                                    0.7
                    217
                              1.17
                                       310.0
                                                              7.91 | 0.0 | 0.24 |
                           0.0
                                      null| 0.0|
                                                       6.94
                                                                           false
        false
                             null
                                              false
                                                             false
                                                                   null | 2021-01-01 00:0
                     HV0005
                                           B02510
        5:04
                             null|2021-01-01 00:08:40|2021-01-01 00:39:39|
        2
                     29
                                      1859.0
                                                             31.18 | 0.0 | 0.94 |
                            10.852
                                                                                   2.77
                           0.0
                                      null | 0.0|
                                                      27.61
                                                                           false
        false
                            false
                                                             false
                                              false
                     HV0003
                                           B02836
                                                                 B02836 2021-01-01 00:4
        0:44 \mid 2021-01-01 \quad 00:53:34 \mid 2021-01-01 \quad 00:53:48 \mid 2021-01-01 \quad 01:11:40 \mid
        2
                              3.52
                                      1072.0
                                                            28.67 | 0.0 | 0.86 |
                                                                                   2.54
                           0.0
                                      null| 0.0|
                                                      17.64
                             null|
        false
                                              false
                                                             false
        only showing top 20 rows
In [0]: 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", "original
         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)
```

1

```
-- 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 \mid 2022-01-01 \quad 00:05:40 \mid 2022-01-01 \quad 00:07:24 \mid 2022-01-01 \quad 00:18:28 \mid 2022-01-01 \quad 00:05:40 \mid 2022-01-01 \quad 00:05:40
                                                                                                                                                                      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 2022-01-01 00:1
                                                                         B03404
0.82
                                                                                                                11.97 | 0.0 | 0.36 |
7 |
                        161
                                                                   460.0
                                                                                                                                                                  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 |
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
                                                                                                       6.3
                                      2.75
                                                                   0.0 | 0.0 |
                                                                                                                                               false
false
                                            null
                                                                                 false
                                                                                                                  false
                          HV0003|
                                                                                                                        B03404 | 2022-01-01 00:2
                                                                       B03404
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
```

-- dropoff datetime: timestamp (nullable = true)

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

# Removing Outliers and Duplicates from 2021 and 2022

```
In [0]: 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) <= upper

# 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()</pre>
```

```
In [0]: # 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]:
                  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
                                                                                     800.0
                  5
                                           232
                                                                   3.2
                                                                                                                                 11.51
                                                                                                                                                   0.0 | 0.53 |
                                                                                                                                                                                  1.03
                                                        2.75
                                                                                   null | 2.82 |
                                                                                                                    10.99
                                                                                                                                                                false
                  false
                                                              null
                                                                                                  false
                                                                                                                                   false
                                             HV0003
                                                                                                                                          B02764 | 2021-01-01 00:1
                                                                                           B02764
                  0:36 \,|\, 2021 - 01 - 01 \,|\, 00:12:28 \,|\, 2021 - 01 - 01 \,|\, 00:14:30 \,|\, 2021 - 01 - 01 \,|\, 00:50:27 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\, 00:36 \,|\,
                                                                                                                                                                                         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|
                                                                                                                                                   0.0 | 0.59 |
                                           191
                                                               5.864
                                                                                     839.0
                                                                                                                                 19.82
                  2 |
                                                                                                                                                                                  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 00:48:14 | 2021-01-01 01:08:42 |
                                                                                                                                                                                       14
                                             78
                                                                   9.2
                                                                                   1228.0
                                                                                                                                 27.11
                                                                                                                                                   0.0 | 0.81 |
                                                                                   null| 0.0|
                                                                                                                                                                false|
                                                        2.75
                                                                                                                    22.44
                                                              null|
                                                                                                  false
                                                                                                                                   false
                  false
```

```
HV0003 | B02875 | B02875 | 2021-01-01 00:2
255 | 1.6 | 446.0 |
                                         6.89 | 0.0 | 0.21 |
                                                           0.61
               0.0
                      null | 0.0 | 6.26 |
                                                     false
                             false
false
                null
                                          false
                                             B02835 | 2021-01-01 00:2
          HV0003|
                          B02835
2:58|2021-01-01 00:27:01|2021-01-01 00:29:04|2021-01-01 00:36:27|
         48 | 1.8 | 443.0 |
                                          8.18 | 0.0 | 0.25 |
                                                           0.73
              2.75 | null| 0.0 | 6.12 |
                                                     false
                             false
false
                null
                                          false
          HV0005|
                           B02510
                                              null|2021-01-01 00:4
               null|2021-01-01 00:50:00|2021-01-01 01:04:57|
0:06
                2.469 | 897.0 |
                                          25.03 | 0.0 | 0.75 |
                                                            2.22
         151
                        null| 0.0|
               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|
         143 | 0.74 | 179.0 |
                                                             0.7
                                          7.91 | 0.0 | 0.24 |
              2.75 | null | 0.0 |
                                     7.41
                                                     false
false
                null
                              false
                                          false
          HV0003|
                           B02682
                                             B02682 2021-01-01 00:1
0:22 \mid 2021-01-01 \quad 00:11:03 \mid 2021-01-01 \quad 00:11:53 \mid 2021-01-01 \quad 00:18:06 \mid
         148 | 2.03 | 373.0 |
                                         7.84 | 0.0 | 0.42 |
                                                             0.7
                                                    false
              2.75 | null|2.82|
                                     6.93
                              false
false
                null
                                          false
          HV0005
                           B02510
                                            null 2021-01-01 00:0
                null | 2021-01-01 00:08:40 | 2021-01-01 00:39:39 |
5:04
               10.852 | 1859.0 | 31.18 | 0.0 | 0.94 |
2
          29|
                                                            2.77
                        null | 0.0|
               0.0
                                     27.61
                                                     false
               false
                              false
                                          false
false
          HV0005
                           B02510
                                            null 2021-01-01 00:0
               null|2021-01-01 00:06:59|2021-01-01 00:43:01|
4:00
                        2162.0 | 28.11 | 0.0 | 0.84 |
          42
               9.725
                                                            2.49
                        null| 0.0|
              2.75
                                     28.9
                                                     false
false
               false
                              false
                                          false
                                            B02682 2021-01-01 00:2
          HV0003
                          B02682
5:00 | 2021-01-01 | 00:26:31 | 2021-01-01 | 00:28:31 | 2021-01-01 | 00:41:40 |
                                                            7
          80 | 3.08 | 789.0 |
                                          13.2 | 0.0 | 0.4 |
                        null | 0.0 | 11.54 |
              2.75
                                                     false
                              false
                null
                                          false
false
                           B02764
                                             B02764 2021-01-01 00:2
          HV0003
1:15 | 2021-01-01 00:22:41 | 2021-01-01 00:23:56 | 2021-01-01 00:38:05 |
                                                             23
         142 | 3.51 | 849.0 | 14.05 | 0.0 | 0.48 |
                        null|0.94|
                                                   false
              2.75
                                    12.98
false
                null
                              false
                                          false
                           B02836
                                             B02836 2021-01-01 00:4
          HV0003
0:44 \mid 2021-01-01 \quad 00:53:34 \mid 2021-01-01 \quad 00:53:48 \mid 2021-01-01 \quad 01:11:40 \mid
                      1072.0
                                          28.67 | 0.0 | 0.86 |
          22
                 3.52
                                                            2.54
                        null | 0.0 | 17.64 |
               0.0
                                                    false
                null
                              false
                                        false
false
          HV0004
                           B02800
                                             null|2021-01-01 00:1
                null|2021-01-01 00:15:24|2021-01-01 00:38:31|
2:50
                                                            18
                9.66
                      1387.0
                                          32.95 | 0.0 | 0.0 |
                                                            2.34
         237
                        null| 0.0|
              2.75
                                     21.1
                                                    false
               false
                              false
                                          false
false
                           B02682
                                             B02682 2021-01-01 00:4
         HV0003
5:56|2021-01-01 00:55:19|2021-01-01 00:55:19|2021-01-01 01:18:21| 15
```

```
false
                          null
                                         false
                                                       false
                                                          B02682 | 2021-01-01 00:2
                   HV0003
                                      B02682
        8:09 \mid 2021-01-01 \quad 00:31:42 \mid 2021-01-01 \quad 00:33:44 \mid 2021-01-01 \quad 00:49:07 \mid
        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
                                                            null | 2021-01-01 00:3
                                      B02800
                          null|2021-01-01 00:45:00|2021-01-01 01:06:45|
        5:32
                                                                             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 | 2021-01-01 00:4
                                      B02835
        6:44|2021-01-01 00:47:49|2021-01-01 00:48:56|2021-01-01 00:59:12|
                                                                             23
                                   616.0
                                                       13.1
                                                              0.0 | 0.45 |
                                                                           1.17
                   75 l
                            2.9
                       2.75
                                   null|0.94|
                                                  8.77
                                                                   false
        false
                          null
                                         false
                                                       false
                                                          B02875 | 2021-01-01 00:5
                   HV0003
                                      B02875
        3:31 | 2021-01-01 00:56:21 | 2021-01-01 00:56:45 | 2021-01-01 01:17:42 |
                           5.74
                                   1257.0
                                                      17.18
                                                              0.0 | 0.52 |
                  198
                                                                           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]:
       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|
       time on scene datetime
                                   pickup datetime
                                                    dropoff datetime | PULocation I
        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|
```

2 |

167

3.65

0.0

1382.0

null| 0.0|

18.36

17.06

0.0 | 0.55 |

false

1.63

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

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: string, 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, DOLocationID: 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("pickudf = df_2021_cleaned.withColumn("dropoff_datetime", unix_timestamp(col("dropoff = df.withColumn("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 |
                  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 | 1183.564444444445
         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

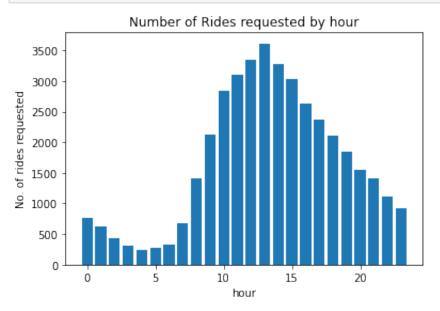
```
In [0]: # Your code goes here
    from matplotlib import pyplot as plt
    # group the rows by hour and count the number of rows in each hour
    count_df = df_2021_cleaned.groupBy(hour("request_datetime").alias("hour")).a

# 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()
```



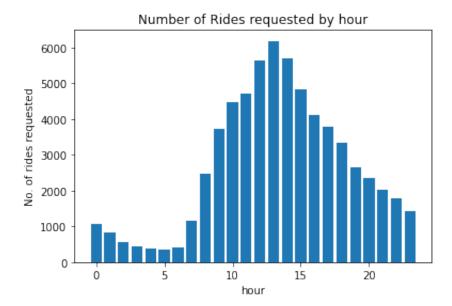
```
In [0]: from matplotlib import pyplot as plt
    # group the rows by hour and count the number of rows in each hour
    count_df = df_2022_cleaned.groupBy(hour("request_datetime").alias("hour")).a

# 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()

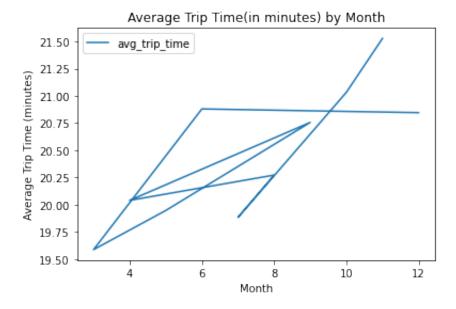
    pandas_df['avg_trip_time'] = (pandas_df['avg_trip_time']/60)

# 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()

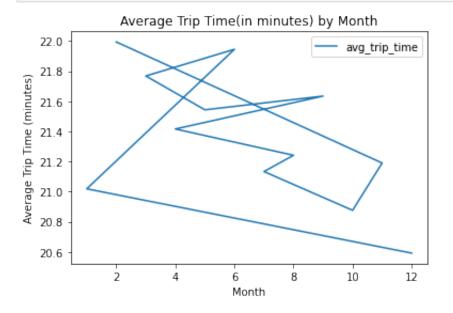
pandas_df['avg_trip_time']= (pandas_df['avg_trip_time']/60)

# 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]: # Identify high-demand pickup locations
        demand_analysis = (
            df 2021 cleaned.groupBy("PULocationID")
            .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()
```

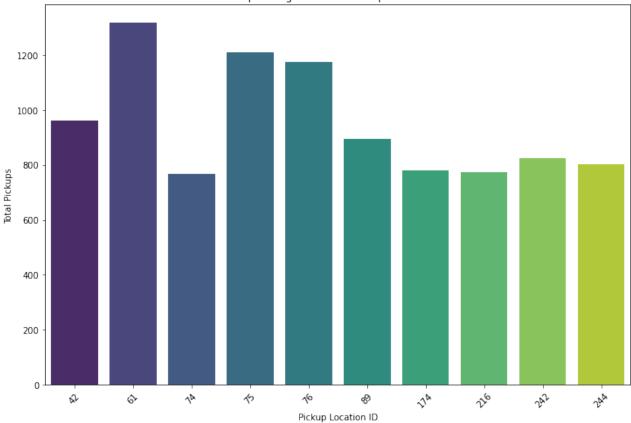
Pickup Location ID

ŵ

ô

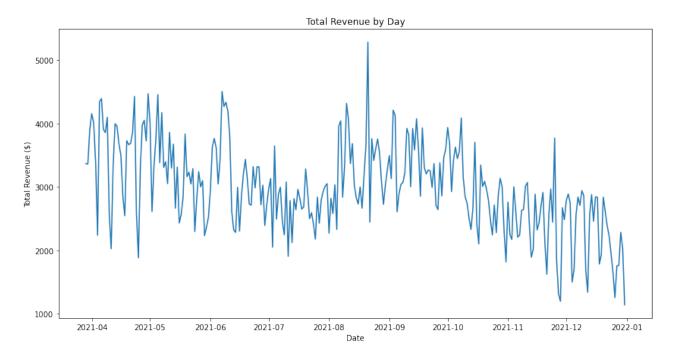
```
In [0]: # Identify high-demand pickup locations
        demand analysis = (
            df_2022_cleaned.groupBy("PULocationID")
            .count()
            .withColumnRenamed("count", "total pickups")
            .orderBy("total_pickups", ascending=False)
        import matplotlib.pyplot as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



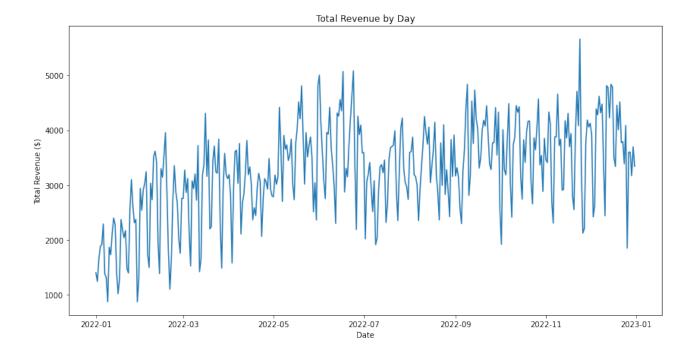


#### Total Revenue for 2021 and 2022

```
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()
```



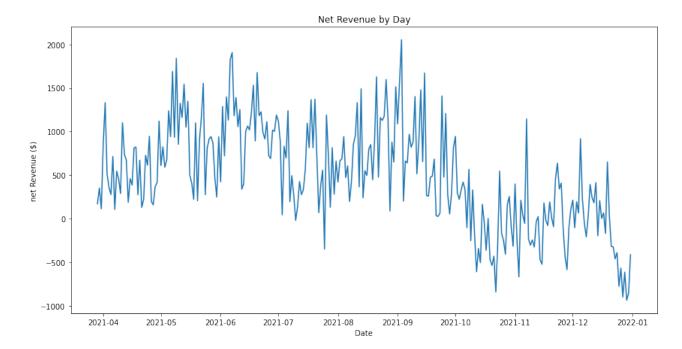
```
In [0]: df_2022_cleaned = df_2022_cleaned.withColumn('total_revenue',col('base_passe
# Group by date and sum the total revenue
revenue_by_day = df_2022_cleaned.groupBy(F.to_date("Pickup_datetime").alias(
# Convert to Pandas DataFrame for plotting (if you want to use matplotlib)
revenue_by_day_pd = 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

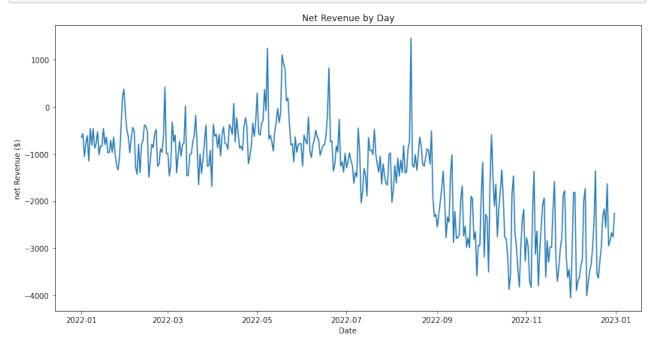
```
In [0]: df_2021_cleaned = df_2021_cleaned.withColumn('net_revenue',col('base_passeng
# Group by date and sum the total revenue
profit_by_day = df_2021_cleaned.groupBy(F.to_date("Pickup_datetime").alias('
# Convert to Pandas DataFrame for plotting (if you want to use matplotlib)
profit_by_day_pd = 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

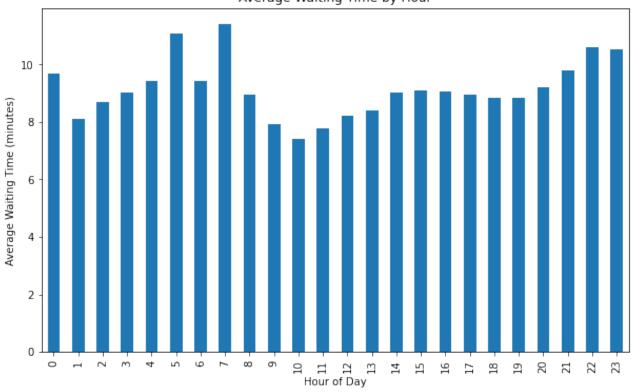
# Visualization with Pandas and Matplotlib
plt.figure(figsize=(14, 7))
profit\_by\_day\_pd['net\_revenue'].plot(title='Net Revenue by Day')
plt.xlabel('Date')
plt.ylabel('net Revenue (\$)')
plt.show()



# 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()
```





```
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()
```



