Untitled1

June 3, 2018

In [1]:
```
import pandas as pd
```

In [61]:
```
passCount=pd.io.gbq.read_gbq("""
SELECT passenger_count, count(passenger_count)
FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
group by passenger_count
order by passenger_count
""", project_id='nyc-data-205818')
passCount

# This shows the number of trips by passenger count.
```

Out[61]:

	passenger_count	f0_
0	0	40853
1	1	102991045
2	2	20901372
3	3	6135107
4	4	2981071
5	5	7939001
6	6	5123951
7	7	239
8	8	181
9	9	169

In [53]:
```
full=pd.io.gbq.read_gbq("""
SELECT vendor_id AS Vendor_ID,
    MIN(pickup_datetime) AS Data_Begin_Date,
    MAX(dropoff_datetime) AS Data_End_Date,
    COUNT(*) AS Total_Trips
FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
GROUP BY vendor_id
Limit 1000
""", project_id='nyc-data-205818')
full

# This shows the number of trips by Vendor_ID.
```

Out[53]:

	Vendor_ID	Data_Begin_Date	Data_End_Date	Total_Trips
0	2	2015-01-01	2016-01-01 23:51:32	76658633
1	1	2015-01-01	2253-08-23 07:56:38	69454356

In [38]:
```
full=pd.io.gbq.read_gbq("""
SELECT *
FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
limit 10
""", project_id='nyc-data-205818')
full

# Dataset for personal perusal.
```

Out[38]:

	vendor_id	pickup_datetime	dropoff_datetime	passenger_count
0	1	2015-10-28 14:21:34	2015-10-28 14:56:38	1
1	2	2015-06-27 20:28:34	2015-06-27 20:56:25	5
2	2	2015-11-04 14:42:54	2015-11-04 15:26:17	6
3	1	2015-12-02 14:33:36	2015-12-02 15:03:33	1
4	2	2015-05-30 20:02:28	2015-05-30 20:32:10	1
5	2	2015-03-15 03:03:55	2015-03-15 03:30:59	1
6	1	2015-01-15 11:04:28	2015-01-15 11:41:33	1
7	2	2015-08-17 13:09:54	2015-08-17 13:37:22	5
8	1	2015-08-02 11:00:59	2015-08-02 11:12:43	1
9	1	2015-08-25 10:59:12	2015-08-25 11:17:16	1

	trip_distance	pickup_longitude	pickup_latitude	rate_code
0	9.20	-73.974998	40.756504	1
1	6.18	-74.001587	40.741020	None
2	10.78	-73.975739	40.762390	1
3	8.70	-73.954071	40.766953	1
4	10.37	-73.863098	40.769184	None
5	7.59	-73.995331	40.725002	None
6	19.20	-73.953484	40.772774	None
7	8.95	-73.994232	40.751041	1
8	1.60	-73.992668	40.750549	1
9	6.80	-73.782318	40.644608	4

	store_and_fwd_flag	dropoff_longitude	dropoff_latitude	payment_type
0	N	-73.872536	40.774345	1
1	N	-73.955109	40.685692	1
2	N	-73.861626	40.768303	1
3	N	-74.009018	40.731213	1
4	N	-73.964119	40.679508	1
5	N	-73.932930	40.665352	1
6	N	-73.776382	40.645233	1
7	N	-73.942871	40.850689	1
8	N	-73.981712	40.736671	1
9	N	-73.747185	40.629021	1

	fare_amount	extra	mta_tax	tip_amount	tolls_amount	imp_surcharge
0	31.0	0.0	0.5	7.45	5.54	0.3
1	22.5	0.5	0.5	4.76	0.00	0.3

```
2          38.0    0.0       0.5          8.87          5.54           0.3
3          31.0    0.0       0.5          6.36          0.00           0.3
4          31.0    0.5       0.5          4.00          0.00           0.3
5          26.0    0.5       0.5          5.46          0.00           0.3
6          55.0    0.0       0.5         12.22          5.33           0.3
7          30.0    0.0       0.5          9.24          0.00           0.3
8           2.5    0.0       0.5          0.66          0.00           0.3
9          27.5    0.0       0.5          5.66          0.00           0.3

   total_amount
0         44.79
1         28.56
2         53.21
3         38.16
4         36.30
5         32.76
6         73.35
7         40.04
8          3.96
9         33.96
```

```
In [40]: avgFare=pd.io.gbq.read_gbq("""
         SELECT passenger_count AS Total_Passengers,
             AVG(total_amount) AS Average_Fare
         FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
         GROUP BY Total_Passengers
         ORDER BY Total_Passengers ASC
         """, project_id='nyc-data-205818')
         avgFare

         # This shows the average fare by number of passengers in the cab.
```

```
Out[40]:    Total_Passengers  Average_Fare
         0                 0     16.050098
         1                 1     15.909998
         2                 2     16.847977
         3                 3     16.403699
         4                 4     16.577590
         5                 5     16.251514
         6                 6     15.880795
         7                 7     47.339833
         8                 8     55.257624
         9                 9     59.443491
```

```
In [48]: pickUp=pd.io.gbq.read_gbq("""
         SELECT HOUR(pickup_datetime) AS Pickup_hour,
              COUNT(*) AS Total_Trips
         FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
```

```
GROUP BY 1
ORDER BY 2 desc
""", project_id='nyc-data-205818')
pickUp

# This shows the number of trips by the pickup hour.
```

Out[48]:

	Pickup_hour	Total_Trips
0	19	9022002
1	18	8752363
2	20	8519738
3	21	8473070
4	22	8159782
5	14	7409760
6	17	7327544
7	12	7213591
8	13	7184575
9	23	7140065
10	15	7079931
11	11	6901937
12	9	6734087
13	10	6641324
14	8	6566727
15	16	6179820
16	0	5587235
17	7	5410009
18	1	4126459
19	6	3263655
20	2	3030364
21	3	2207835
22	4	1648389
23	5	1532727

In [49]:
```
dropOff=pd.io.gbq.read_gbq("""
SELECT HOUR(dropoff_datetime) AS Dropoff_hour,
     COUNT(*) AS Total_Trips
FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
GROUP BY 1
ORDER BY 2 desc
""", project_id='nyc-data-205818')
dropOff

# This shows the number of trips by the dropoff hour.
```

Out[49]:

	Dropoff_hour	Total_Trips
0	19	9236353
1	20	8665025
2	18	8643958

```
3              21      8466107
4              22      8227757
5              23      7417141
6              14      7269638
7              12      7205975
8              15      7198346
9              13      7120431
10             17      6880229
11             11      6752983
12              9      6744163
13             10      6633361
14             16      6388589
15              8      6228940
16              0      6028081
17              7      4902369
18              1      4453086
19              2      3238975
20              6      2856517
21              3      2323086
22              4      1781610
23              5      1450269
```

In [59]:
```
goodTip = pd.io.gbq.read_gbq("""
SELECT tip_amount, tip_amount/total_amount AS goodTip,
FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
order by tip_amount desc
limit 10
""", project_id='nyc-data-205818')
goodTip

# This shows the highest tip paid by the customer as a percentage of the total amount
```

Out[59]:

	tip_amount	goodTip
0	3950588.80	0.999994
1	1603.05	0.166666
2	1200.80	0.130435
3	980.91	0.990108
4	969.69	0.984967
5	950.00	0.947347
6	950.00	0.947347
7	950.00	0.947347
8	910.05	0.500000
9	905.00	0.230761

In [60]:
```
pickUp=pd.io.gbq.read_gbq("""
SELECT MONTH(pickup_datetime) AS Pickup_month,
     COUNT(*) AS Total_Trips
FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
```

```
GROUP BY 1
ORDER BY 2 desc
""", project_id='nyc-data-205818')
pickUp

# This shows the most trips by month of the year.
```

Out[60]:

	Pickup_month	Total_Trips
0	3	13351609
1	5	13158262
2	4	13071789
3	1	12748986
4	2	12450521
5	6	12324935
6	10	12315488
7	7	11562783
8	12	11460573
9	11	11312676
10	9	11225063
11	8	11130304

In [77]:
```
avgSpeed=pd.io.gbq.read_gbq("""
SELECT HOUR(pickup_datetime) AS Pickup_hour,
     ROUND(AVG(trip_distance / DATEDIFF(dropoff_datetime,
        pickup_datetime)))
FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
GROUP BY 1
ORDER BY 1
""", project_id='nyc-data-205818')
avgSpeed

# This shows the average speed of trips by the hour of the day.
```

Out[77]:

	Pickup_hour	f0_
0	0	3.0
1	1	4.0
2	2	4.0
3	3	4.0
4	4	5.0
5	5	6.0
6	6	5.0
7	7	4.0
8	8	4.0
9	9	4.0
10	10	4.0
11	11	4.0
12	12	4.0
13	13	4.0

```
14           14   4.0
15           15   4.0
16           16   4.0
17           17   4.0
18           18   4.0
19           19   4.0
20           20   4.0
21           21   4.0
22           22   8.0
23           23  28.0
```

In [82]:
```
avgSpeed=pd.io.gbq.read_gbq("""
SELECT DAYOFWEEK(pickup_datetime) AS Pickup_day,
     ROUND(AVG(trip_distance / DATEDIFF(dropoff_datetime,
        pickup_datetime)))
FROM [bigquery-public-data.new_york.tlc_yellow_trips_2015]
GROUP BY 1
ORDER BY 1
""", project_id='nyc-data-205818')
avgSpeed

# This shows the average speed of trips by the day of the week.
```

Out[82]:

	Pickup_day	f0_
0	1	53.0
1	2	31.0
2	3	24.0
3	4	6.0
4	5	43.0
5	6	31.0
6	7	7.0