Project Development Phase – Sprint 2

Date	31 October 2022
Team ID	PNT2022TMID50853
Project Name	A new hint to transportation – Analysis of the NYC bike share system.
Maximum Marks	20 Marks

Feature Engineering:

calculating Age from birth year

```
from datetime import datetime, date

age=2018-df['birth_year']

df['Age']=age

df.head()
```

	tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype	birth_year	gender	tripduration_bins	Age
0	11.583333	2013-06-01 00:00:01	2013-06-01 00:11:36	444	Broadway & W 24 St	40.742354	-73.989151	434.0	9 Ave & W 18 St	40.743174	-74.003664	19678	Subscriber	1983.0	1	(0.0, 30.0]	35.0
1	11.550000	2013-06-01 00:00:08	2013-06-01 00:11:41	444	Broadway & W 24 St	40.742354	-73.989151	434.0	9 Ave & W 18 St	40.743174	-74.003664	16649	Subscriber	1984.0	1	(0.0, 30.0]	34.0
3	2.050000	2013-06-01 00:01:04	2013-06-01 00:03:07	475	E 15 St & Irving Pl	40.735243	-73.987586	262.0	Washington Park	40.691782	-73.973730	16352	Subscriber	1960.0	1	(0.0, 30.0]	58.0
4	25.350000	2013-06-01 00:01:22	2013-06-01 00:26:43	2008	Little West St & 1 Pl	40.705693	-74.016777	310.0	State St & Smith St	40.689269	-73.989129	15567	Subscriber	1983.0	1	(0.0, 30.0]	35.0
6	34.283333	2013-06-01 00:02:33	2013-06-01 00:36:50	285	Broadway & E 14 St	40.734546	-73.990741	532.0	S 5 Pl & S 5 St	40.710451	-73.960876	15693	Subscriber	1991.0	1	(30.0, 60.0]	27.0

calculating age group from age

```
max_limit = df['Age'].max()

max_limit

bins = [0,20,40,60,max_limit]

agegroup = pd.cut(df['Age'], bins=bins).value_counts()

Agegroup
```

```
(20.0, 40.0]     161563
(40.0, 60.0]     148805
(60.0, 119.0]     27014
(0.0, 20.0]           0
Name: Age, dtype: int64
```

calculating hour

peak_hour['Start Date'] = pd.to_datetime(df['starttime'])

peak_hour['Stop Date'] = pd.to_datetime(df['stoptime'])

peak_hour['year'] =peak_hour["Start Date"].dt.year

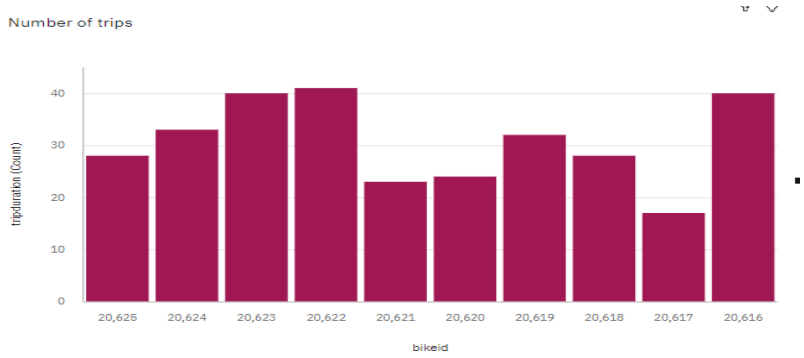
peak_hour["Hour"] = peak_hour["Start Date"].dt.hour

	Start Date	Stop Date	year	Hour	bikeid
0	2013-06-01 00:00:01	2013-06-01 00:11:36	2013	0	19678
1	2013-06-01 00:00:08	2013-06-01 00:11:41	2013	0	16649
3	2013-06-01 00:01:04	2013-06-01 00:03:07	2013	0	16352
4	2013-06-01 00:01:22	2013-06-01 00:26:43	2013	0	15567
6	2013-06-01 00:02:33	2013-06-01 00:36:50	2013	0	15693
...	...	...	...	...	...
577687	2013-06-30 23:58:09	2013-07-01 00:05:25	2013	23	19454
577689	2013-06-30 23:57:52	2013-07-01 00:00:57	2013	23	16746
577690	2013-06-30 23:58:39	2013-07-01 00:08:34	2013	23	19290
577698	2013-06-30 23:59:27	2013-07-01 00:14:52	2013	23	15250
577700	2013-06-30 23:59:33	2013-07-01 00:02:14	2013	23	18910

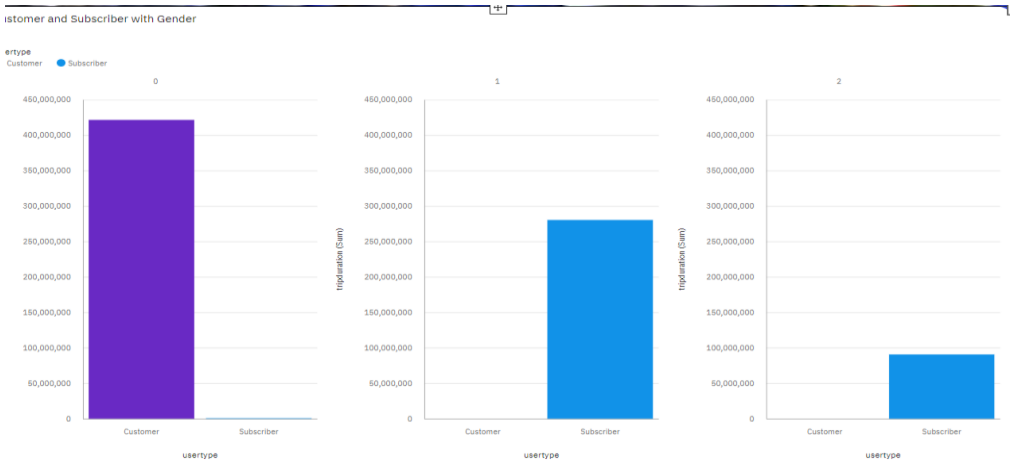
337382 rows × 5 columns

Visualization of the dataset in COGNOS Platform:

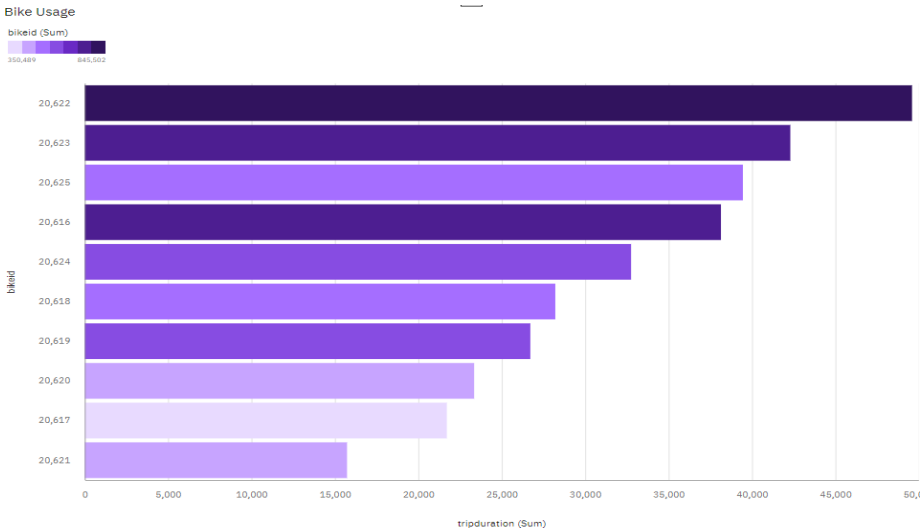
Finding the number of trips per each bike:



Finding the percentage of customers and subscribers



Bike Usage - Bike Id Vs Trip Duration:



Age Group Differentiation by BikeId:

Calculation:

if(age<=20) then

('<20')

else if(age>=21 and age<=30) then

('21-30')

else if(age>=31 and age<=40) then

('31-40')

else if(age>=41 and age<=55) then

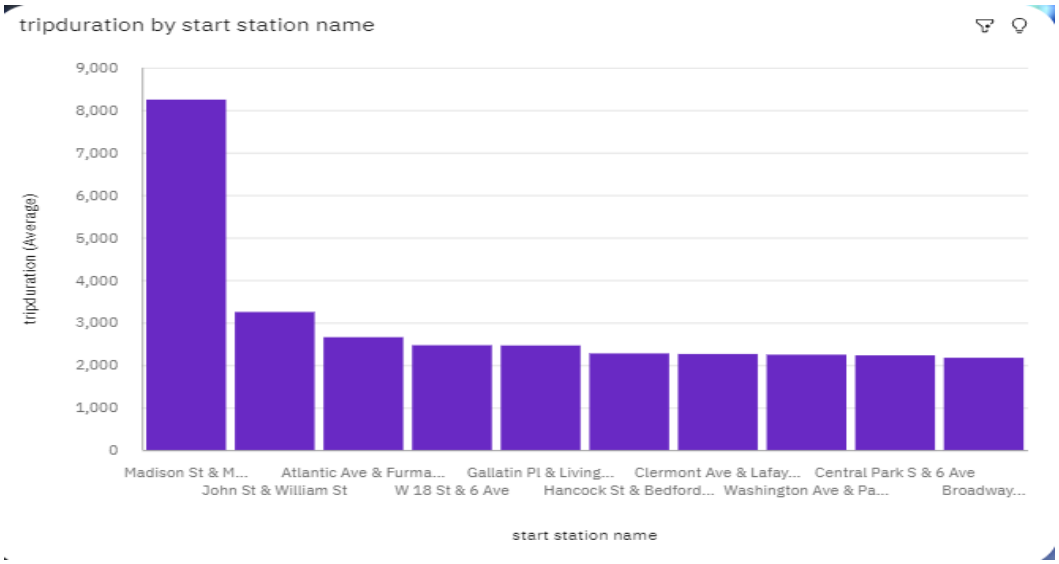
('41-55')

else('>55')

bikeid and Age_Group

Age_Group	bikeid
21-30	5,721
31-40	5,749
41=55	5,741
<20	1,525
>55	5,781
Summary	5,794

Finding the top 10 start stations with customer age group:


tripduration by start station name
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
tripduration (Average)
Madison St & M...
John St & William St
Atlantic Ave & Furma...
W 18 St & 6 Ave
Gallatin Pl & Living...
Hancock St & Bedford...
Clermont Ave & Lafay...
Washington Ave & Pa...
Central Park S & 6 Ave
Broadway...
start station name