A PROJECT REPORT ON

CITI BIKE DATA ANALYSIS USING IBM CLOUD COGNO'S DASHBOARD

Submitted to Smart Bridge

By:

S. SOWMYA	**19R11A05J9**
V. VARUN	**19R11A05K0**
V. MYTHILI	**19R11A05K1**
Y. NANDINI	**19R11A05K2**
M. MADHURI	**20R15A0521**

TABLE OF CONTENTS:

INTRODUCTION :

New City is the most populous city in the us.In fact, it has more people than 40 states. Furthermore the NY City residents, a considerable number of people are commuting into the city from neighboring counties and states. New York has one of the largest Commuter-Adjusted Daytime Population according and although it has also one of the most developed public transportation system in world.

Since 2013, an alternate commuting option is being offered in New York city; a paying bike sharing system “Citibike”.Riders can rent bike at various docking stations throughout the city and returned them to another docking station. There are 2 main forms of payment; “pay as you go” meaning per ride or “Annual Subscription” meaning pay a flat fee for the year with unlimited rides and higher cap on the ride. There is a time limit on how long the bike can be in use per ride; 30 minutes for non-subscribers and 45 minutes for subscribers. Financial penalties are applied in the cases the ride exceed these limits.

In this study, we consider New York City bike sharing data. The details of every rides by a bike in the bike sharing system is recorded by the docking stations, cleansed, centralized, and made available to the public. Our particular focus for this study is all rides for the year 2015 taken by “subscribers” (riders that pay on a yearly basis and for whom additional demographic information is tagged with ride information captured by the docking stations. This additional information, mainly gender and date of birth is captured at the time of registration by the subscribers and is provided by the subscriber.

This collection represents our population of interest. In this study we would like to explore the possible relationship between the age of rider and the ride duration in minutes, both are numerical. In addition, we would like to explore other variables, such as gender , start station details, number of trips and many more. These additional variables are categorical in nature. For our study, the ride duration is the response variable and the other sited are considered explanatory variables. The study is observational, we are basing our analysis on actual observations. There was no interference when collecting the data on how the data came to be. The data collection is done automatically.

OVERVIEW:

The goal of this analysis is to create an operating report of Citi Bike data for the year of 2015.The Mayor of New York City, needs a better understanding of Citi Bike ridership. So the data is visualized in the form of charts to provide a better understanding. To create the required visualizations IBM Cloud Cognos Analytics is used.These visualizations enable the Mayor to get a better understanding about the Citibike data.

OBJECTIVES :

To create data visualizations to

- Find total number of trips recorded during the given time period.
- Find count of Customers and Subscribers with respect to gender.
- Find Top bike used by trip duration.
- Differentiate number of bikes used by respective age groups.
- Find top 10 start station names with customer age group.
- Find top 10 start stations by number of starts.
- Calculate average trip duration by age and gender.
- Locate start stations with respect to trip duration and user type.

ENVIRONMENT:

The visualizations for the operating report of citibike data analysis are created using IBM Cloud Cogno's Analytics Dashboard.

IBM CLOUD COGNOS ANALYTICS:

IBM Cognos Analytics is an Artificial Intelligence-based business intelligence platform that supports data analytics among other things. One can visualize as well as analyze the data and share actionable insights .It is a fast, flexible, and complete business intelligence and analytics solution that organizations use to improve decision quality and accelerate decision making.It provides a tool set for reporting, analytics, score carding, and monitoring of events and metrics.

In IBM Cognos Analytics, dashboards provide data discovery capabilities. On a dashboard, one can visually explore and interact with the data to identify the

key insights for improving data-driven decisions. The cognos dashboard can be used to crate sophisticated visualizations in an analytics project to identify patterns in the data.

The data can be dragged onto the canvas and various visualizations are used to communicate comparisons or understand relationships and trends in the data.

VISUALIZATIONS:

A dashboard, graph, infographics, map, chart, video, slide, etc. all these mediums can be used for visualizing and understanding data. It enables the user to analyze the data more effectively. Data visualization enables the users to gain insight in to vast amounts of data and to recognize patterns and errors in the data.

The following are the data visualizations created to analyze the citibike data.

MISSING VALUES :

	Total	Percent
Trip duration	0	0.000000
Start time	0	0.000000
Stop time	0	0.000000
Start Station id	0	0.000000
Start station name	0	0.000000
Start station latitude	0	0.000000
Start station longitude	0	0.000000
end station id	0	0.000000
end station name	0	0.000000
end station latitude	0	0.000000
end station longitude	0	0.000000

Bikeid	0	0.000000
Usertype	0	0.000000
birth year	2267	0.011512
Gender	0	0.000000

TOTAL NUMBER OF TRIPS:

- As we observe in the dataset, there are no null values for the trip duration field. So the data can be visualized without any trouble.
- To get the total number of trips, a text table visualization is used.
- It is calculated by using trip duration and bike id.

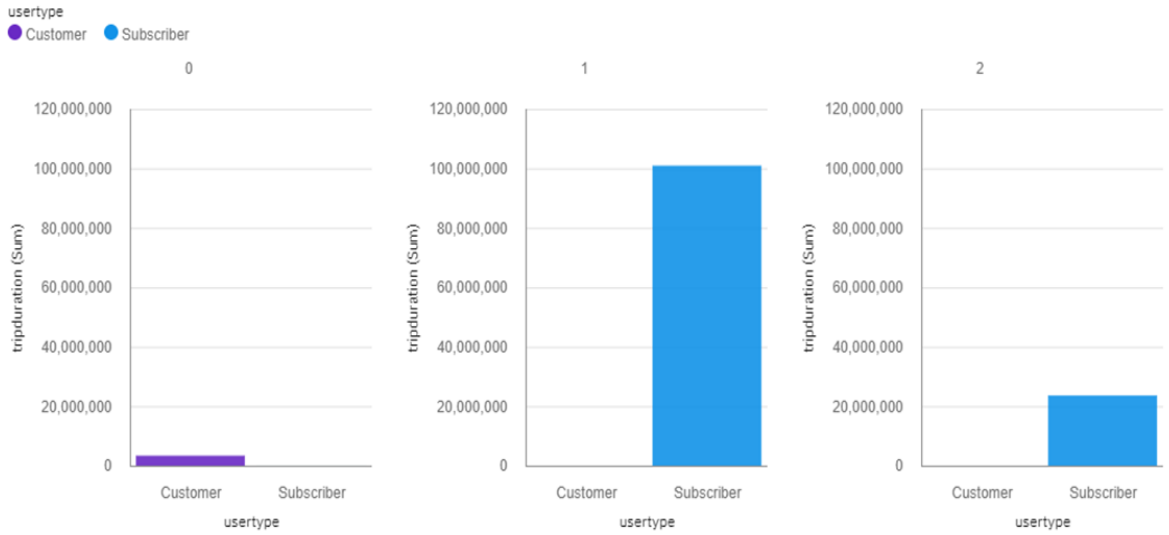
tripduration	bikeid
649.38	4,466

- The total number of trips are 197K .

CUSTOMER AND SUBSCRIBER WITH GENDER:

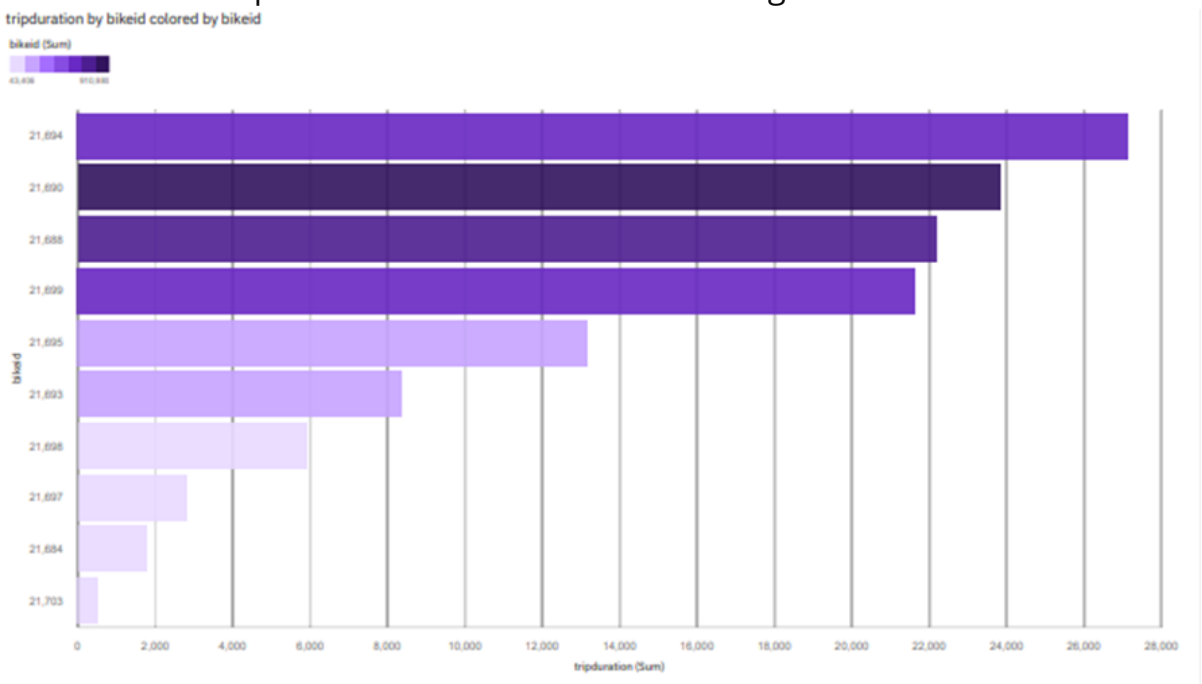
- In order to create this visualization, a bar chart was chosen.
- As we can see there are no null values in the user type field . So the data can be used for the analysis directly.
- A bar graph is plotted with user type on the x-axis and trip duration on the y-axis.

- The user type is categorized as customer and subscriber.
- There are 3 graphs that have been created to get a better understanding of the content.



TOP BIKE USAGE:

- As per the given data, there are no null values for either bike id field or the trip duration field. So the data can be processed.
- To create this visualization, a bar chart is used.
- The trip duration is taken on the x-axis and the bike id is taken on the y-axis.
- To get the accurate results, the bike id field is taken as the top count and the trip duration is set in the descending order.



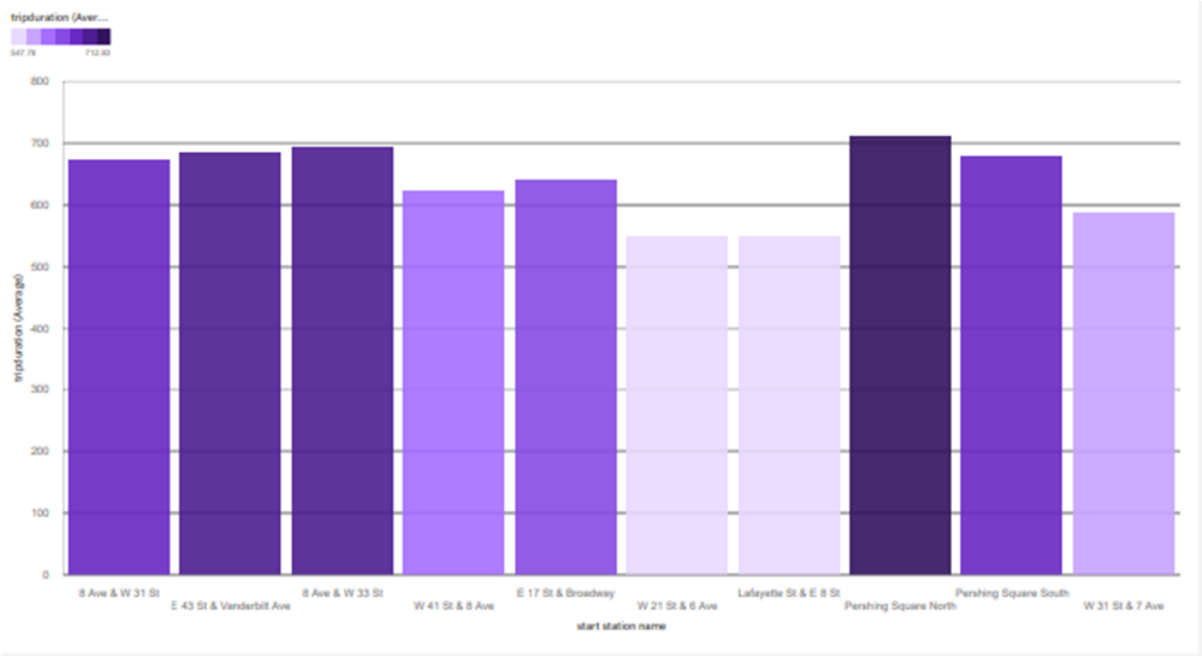
AGE GROUP DIFFERENTIATION BY BIKE:

- The number of bikes can be calculated based on their age groups.
- In the given data, the birth year field has null values. Since the percent of the missing values is meagre, it can be neglected while analyzing the data.
- The age of the person is calculated by subtracting his/her birth year from the current year to get accurate results.
- As we have collected 2015 data , the age is calculated by subtracting birth year from year 2015.
- The results obtained are then differentiated into age groups.
- For this visualization, a text table is used to differentiate the age groups based on bike id of the users.

age_group	bikeid
21-30	4,301
31-40	4,365
41-55	4,359
<20	1,572
>55	4,082
Summary	4,466

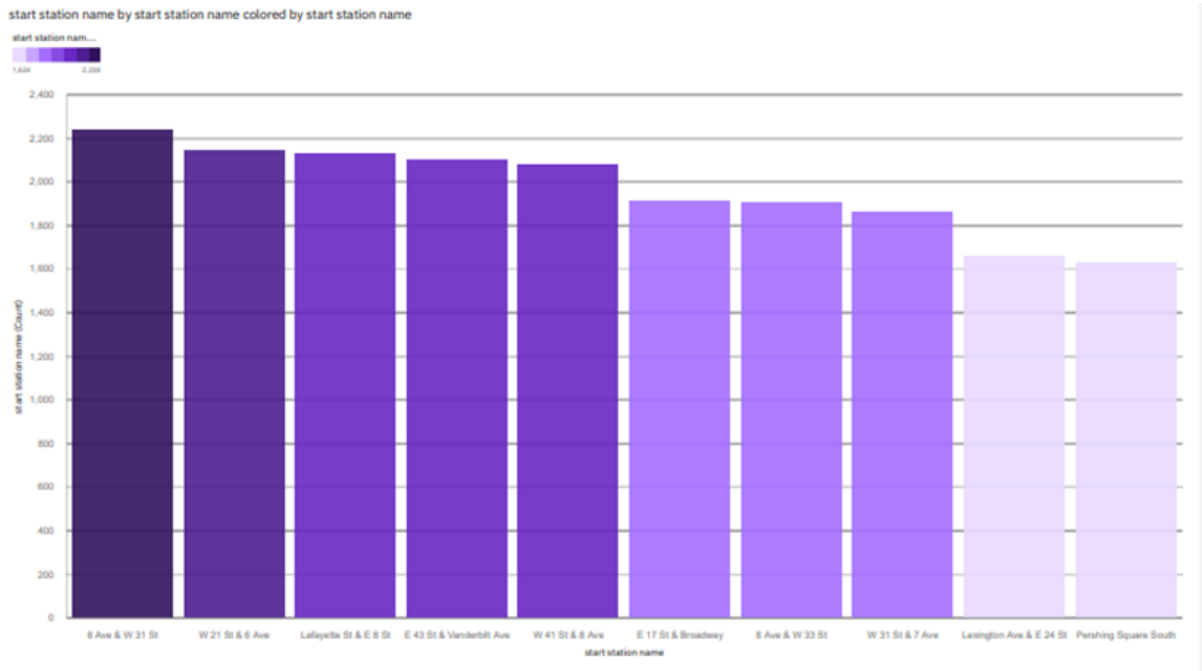
TOP 10 START STATION NAMES WITH RESPECT TO CUSTOMER AGE GROUP AND TRIP DURATION :

- In order to create this visualization , a horizontal bar chart is used.
- The customer age group is taken from the previous analysis. The trip duration is taken on the y-axis and the start station names are taken on the x-axis.
- The graph is created in such a way that the top 10 start station names can be identified.



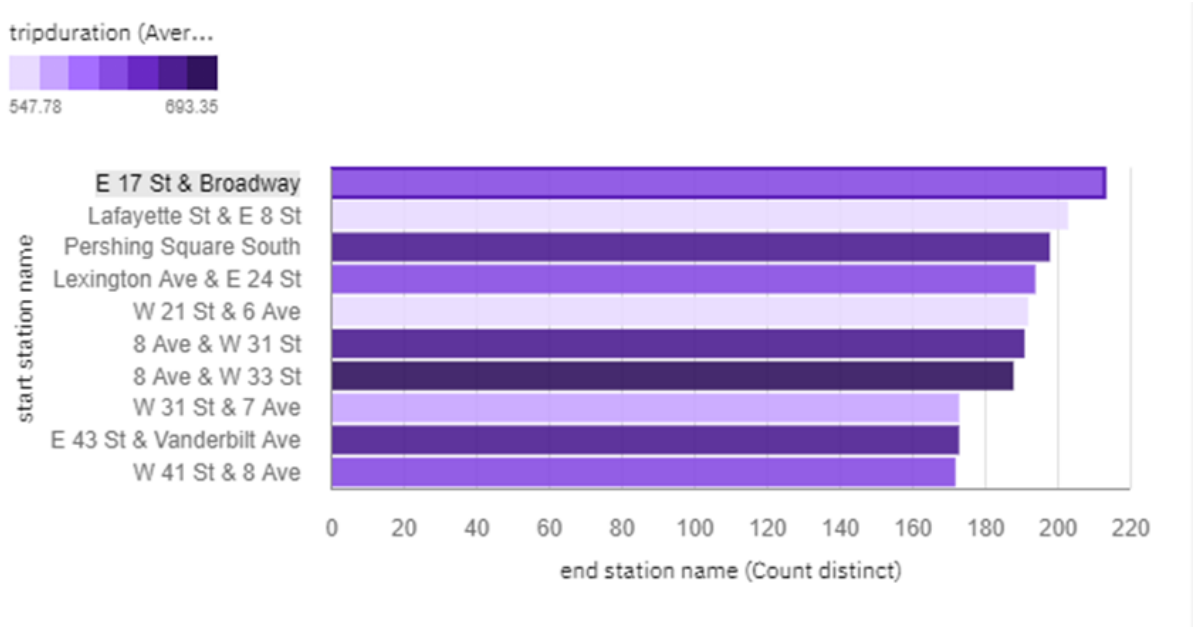
TOP 10 START STATIONS BY NUMBER OF STARTS:

- The data fields required for this analysis do not contain any null or missing values. So the data can be analyzed directly.
- To create this visualization , a bar chart is used where the start station names are plotted with respect to the specific count of number of starts at that particular station.



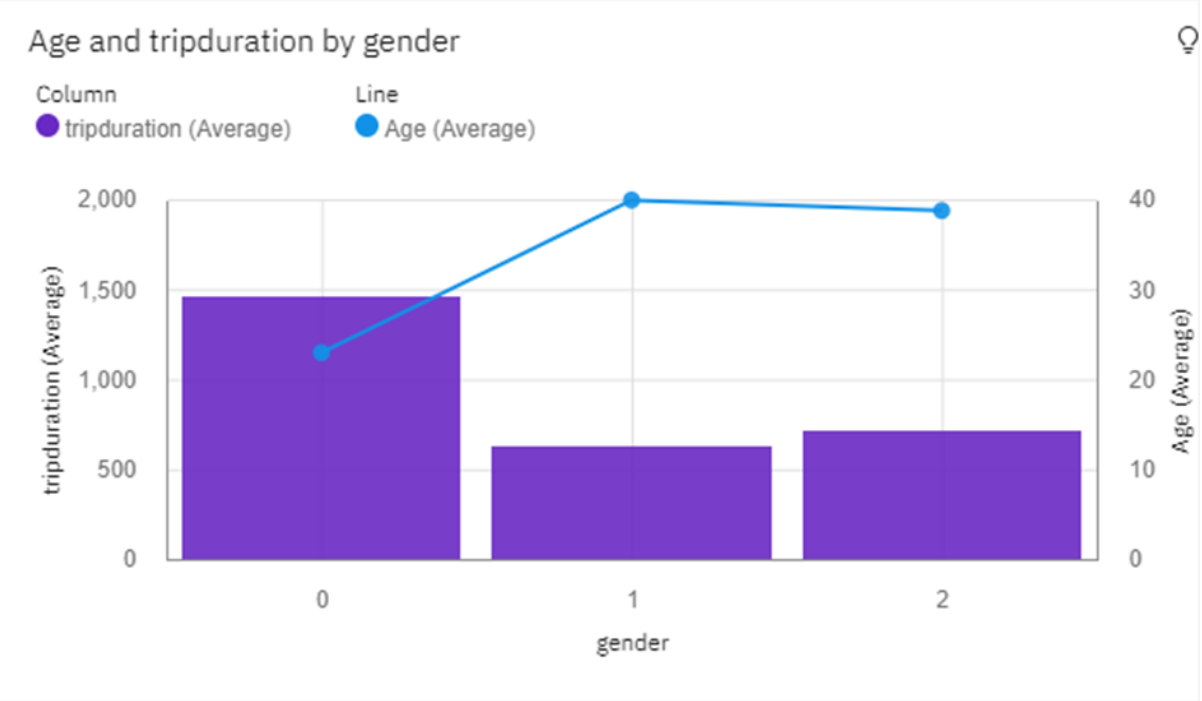
MOST POPULAR TRIPS:

- To find the most popular trips among all the trips, start station and the end station details are required.
- As the required data fields do not contain missing values, the data can be analyzed directly.
- A horizontal bar graph is used to create the visualization to find the most popular trips by plotting start station name and the distinct count of trips at the end station.



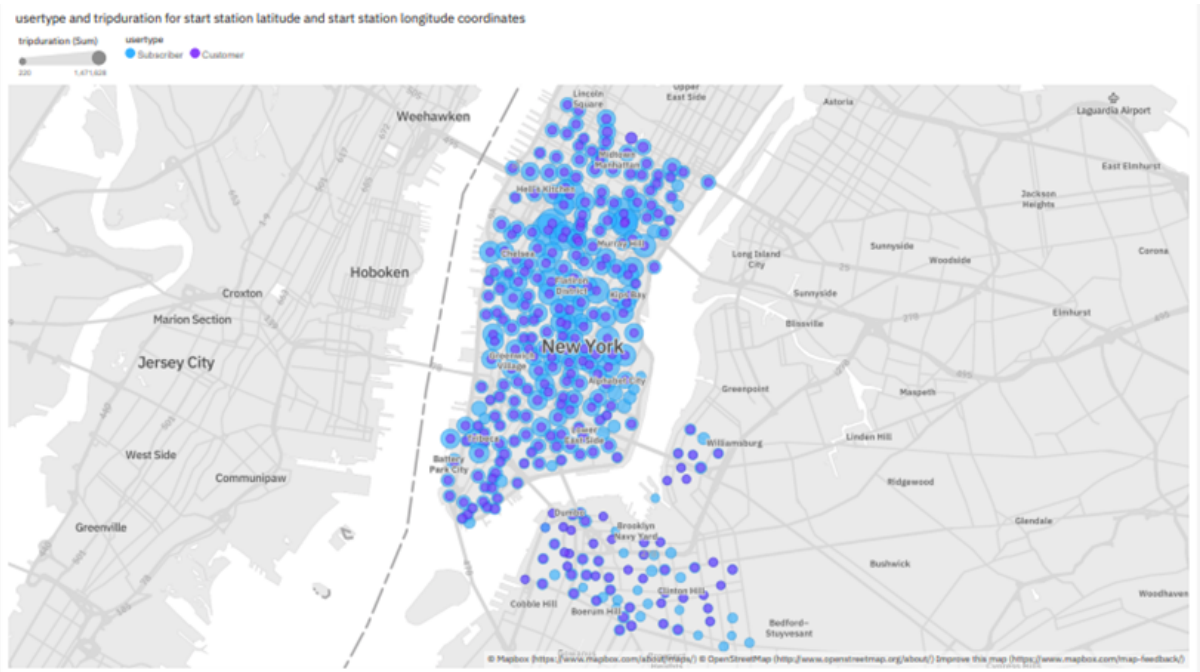
AVERAGE TRIP DURATION BY AGE AND GENDER:

- The average trip duration is calculated using the age and gender of the riders.
- The required data does not contain missing values. So the data is further analyzed.
- The average trip duration is calculated using a column and line graph where the trip duration and gender are plotted along the column graph
- The age is taken as a line and is plotted accordingly on to the column graph.
- The trip duration is taken as average on the y-axis, gender is taken as count on the x-axis and the age is also considered as average values.



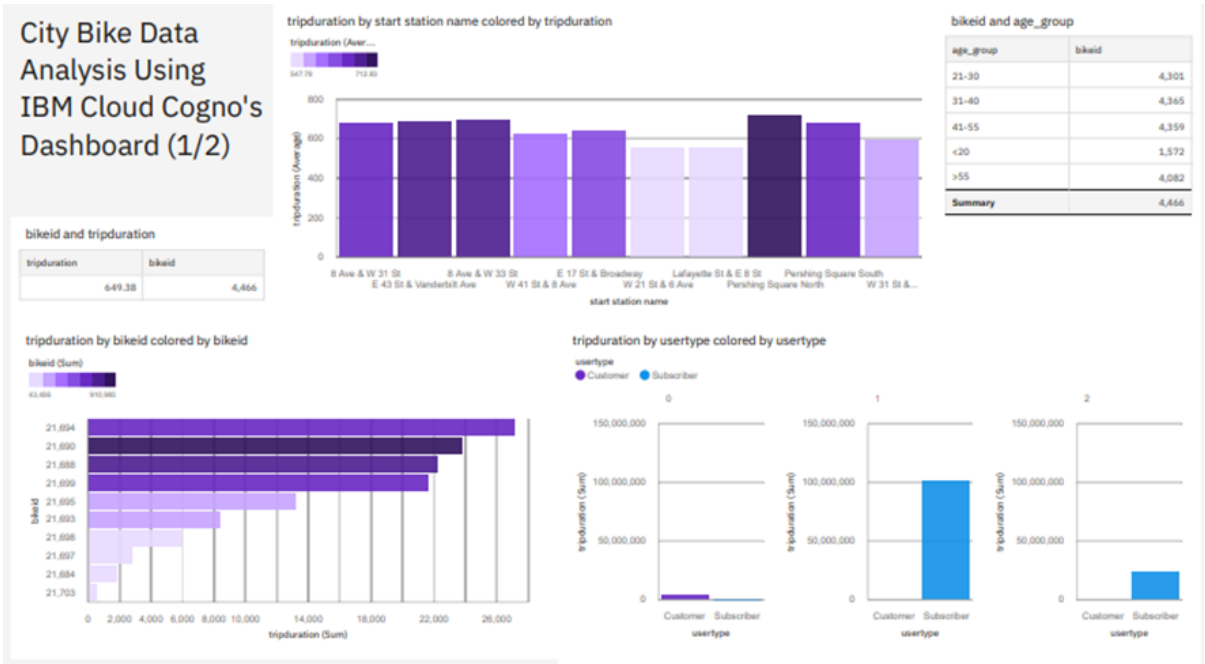
LOCATING START STATIONS :

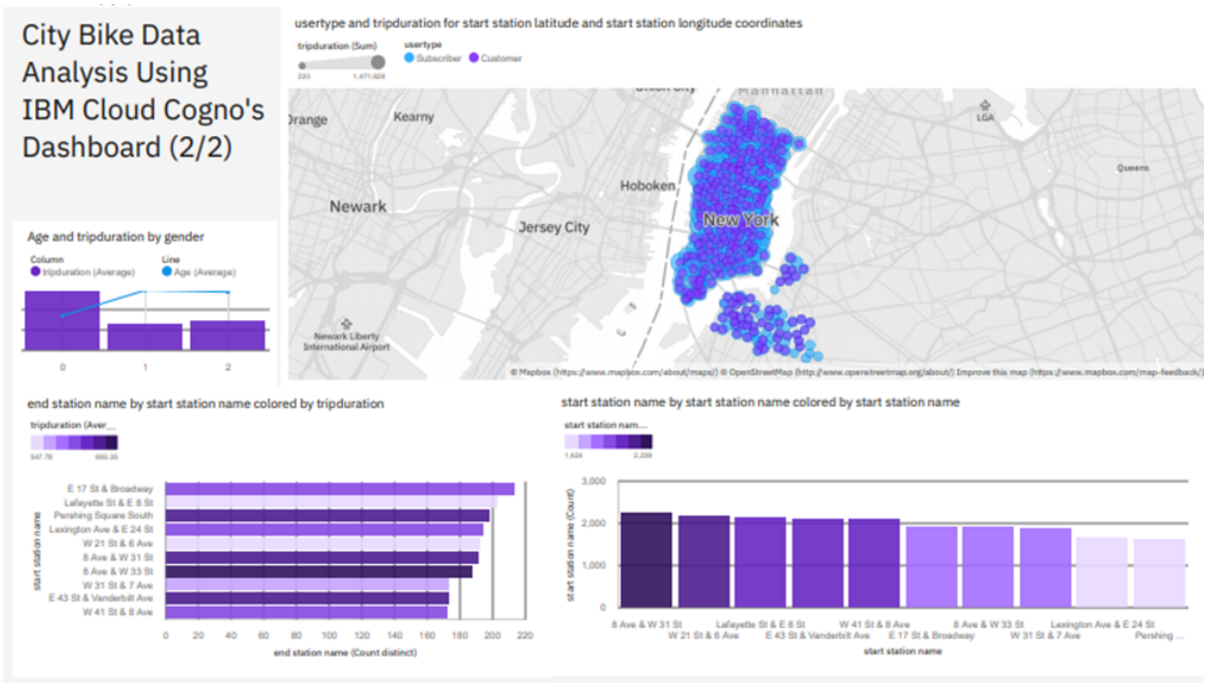
- The start stations can be located geographically using a map visualization.
- In this visualization, the start stations are plotted with respect to the trip duration and user type.
- The geographically representation gives a clear image of the locations of the start stations present across the NY city.



RESULT:

The resultant dashboard is the compilation of all the visualizations created. A Dashboard is an API-based solution that lets developers easily add end-to-end data visualization capabilities to their applications.It enables the viewer to get a clear understanding about the specific problem statements by viewing the visualizations.Also, filters can be added to the visualizations to filter the reports rather than selecting all the data in the application.





CONCLUSION:

This study could be used to analyze the relation between trip duration and various other fields. The results of the analysis indicate that the total number of trips is around 197K. It also indicates that the highest average trip duration is travelled by male riders around 40 years of age.It can also be incurred from above results that the majority citibike users belong to 30-40 years of age.
The most popular trip is through the E17 St & Broadway station.
The study could be useful for generating an operating report of the citibike program being conducted across the New York city. This analysis can help the mayor to get better understanding of the citibike ridership.
The study could be useful for a holistic understanding of the efficiency of the Citi Bike deployment for different types of economic, social and environmental stakeholders, including city agencies, as well as for informing future decisions on bike share deployments in other areas.

APPENDIX:

Citi Bike Trip History is a public dataset produced by Citi Bike, which is New York City's current bike share system. The year 2015 dataset is downloaded and the data is extracted to be analyzed further.

The data set includes :

Name	Description
trip duration	length of ride in second
star time	Start time of ride
stop time	Date and time when the ride ends
start.station.id	unique identifier for ride starting station
start.station.name	name of ride starting station
start.station.latitude	latitude of ride starting station
start.station.longitude	longitude of ride starting station
end.station.id	unique identifier for ride ending station
end.station.name	name of ride ending station

end.station.latitude	latitude of ride ending station
end.station.longitude	longitude of ride ending station
bike id	unique identifier for every bike in system
user type	type of user; “Customer”, “Subscriber”
birth.year	date of birth of “Subscriber” user
gender	gender of “Subscriber” user; 0=Unknown,1=Male, 2=Female