Project Report

Project ID: PNT2022TMID16722

Project Name: A NEW HINT TO TRANSPORTATION-ANALYSIS OF THE NYC BIKE SHARE SYSTEM

TEAM MEMBERS:

Ramnath K	- 111919205033
Vignesh J	- 111919205049
Jeyanth J	- 111919205014
Vasantha Kumar S	- 111919205046

1. INTRODUCTION

1.1 Project Overview

Bike share programs have risen in popularity in recent years and have been promoted as a lower carbon alternative to other forms of transit. Interest in bicycle sharing has been growing exponentially over the past decade, resulting in a proliferation of bike share systems in 712 cities across the world, encompassing 806,000 bicycles and 37,500 stations. This can be largely attributed to the successful incorporation of information technology in docking stations and mobile devices as well as improved logistics such as bicycle rebalancing to ensure responsive supply management. Cities often hope bike sharing will bring many benefits such as extending the reach of transit, substituting motorized trips, and encouraging non-cyclists to try cycling.

The premise of bicycle sharing is that it is a short-term bike rental system, based on varying timed memberships. Members of the bike share network have access to stations, consisting of a pay-station and multiple bike docks, across the system where bikes can be checked out from one station and returned to another nearest to their destination. The appeal of membership is 24/7 access to an automated bike rental network and utility of bikes in completing “last-kilometer connections” without the worry of storage or maintenance. The price system is set to encourage shorter trips (less than 30 minutes in time), with additional fees for any time used over that maximum.

There is evidence that bike share users switch to bike share from motorized transport, such as bus and auto, creating the potential for significant reductions in transportation related greenhouse gas or CO2e emissions. However, there is significant heterogeneity between different cities, showing that there is not a guaranteed CO2e reduction benefit from instituting bike share, especially if the trips would not have been made otherwise or are substituting walking and private bicycle trips.

1.2 Purpose

The purpose of this analysis is to create an operating report of Citi Bike for the year 2018. From this analysis, the following data visualizations will be created.

1.Total Number of Trips

2.What is Customer and subscriber with gender

3.Find the top bike used with respect to trip duration?

4.Calculating the number of bikes used by respective age groups.

5.Top 10 Start Station Names with respect to Customer age group

2. LITERATURE SURVEY

2.1 Existing Problem

Spinlister - Spinlister is an online hub for renting bikes from individuals or bike rental shops.

Zagster - Life is better on a bike! They are bringing bike share to communities across the USA**.**
Motivate International - Motivate is a global full-service bike share operator and technology innovator.

Spin - Spin is a stationless bike and electric scooter sharing service.

2.2 References

https://craft.co/citi-bike/competitors

Ines et al.,ScienceDirect-Social and Behavioral Sciences 111 (2014) 518 – 527
“ Bicycle sharing systems demand”

Elias et al.,ScienceDirect Journal of Transport Geography 91 (2021)
102971”What do trip data reveal about bike-sharing system users? “

FRANCESCO et al.,IEEE Access 2020"Bike Sharing and Urban Mobility in a Post-Pandemic"

"A long-term perspective on the COVID-19: The bike sharing system resilience under the epidemic environment"Journal of Transport & Health ,2021

Nguyen Thi Hoai Thu, Chu Thi Phuong Dung, Vietnam 2017 International Conference on Advanced Technologies for Communications - Multi-source Data Analysis for Bike Sharing Systems

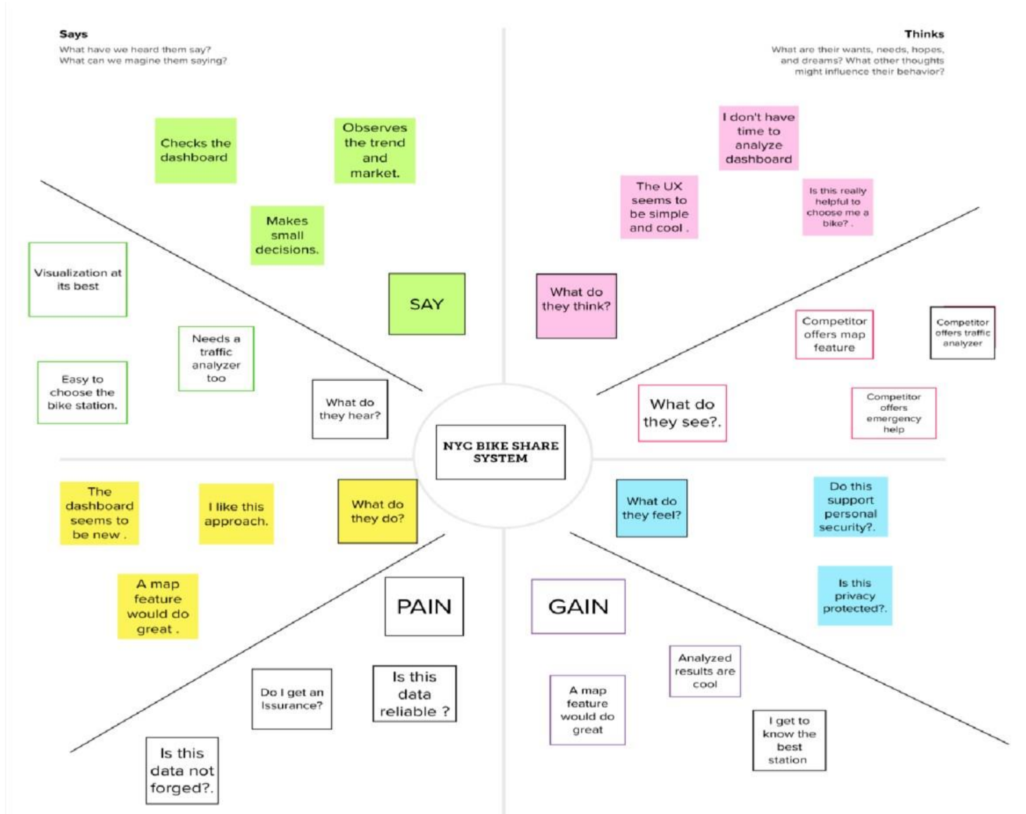
2.3 Problem statement Definition

In busy cities like New York the people are facing difficulties in analyzing the demand for bikes during peak hours.

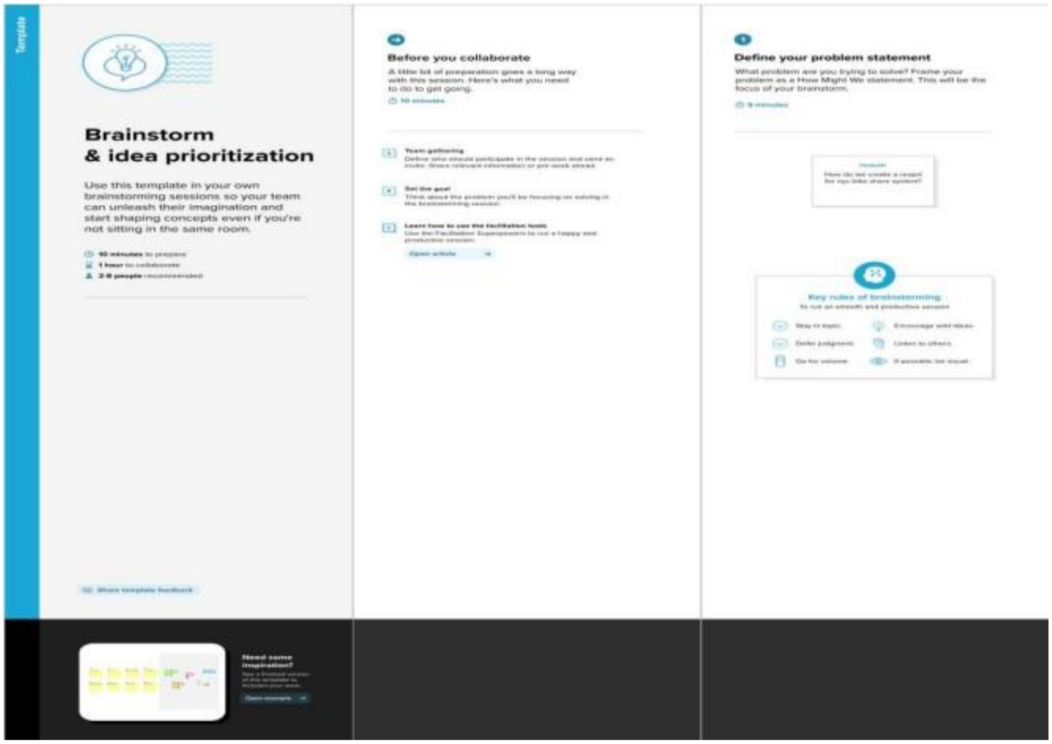
The main objective of this project is to predict bike patterns that will be extremely helpful for people to plan their travel.

3. IDEATION & PROPOSED SOLUTION

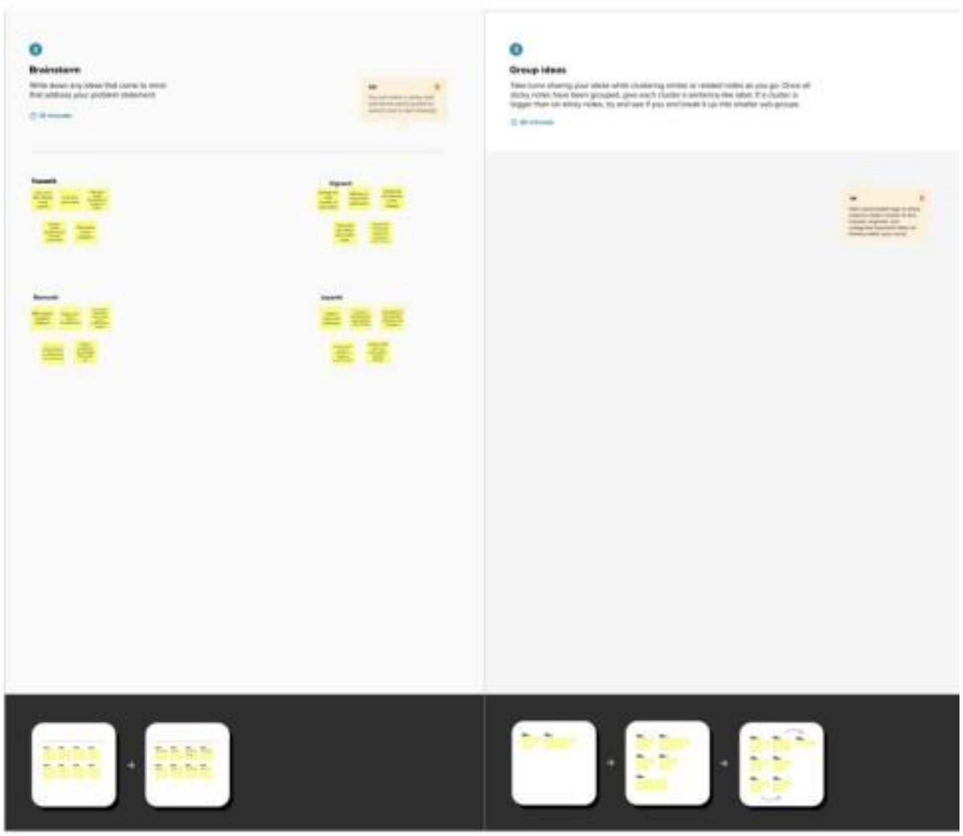
3.1 Empathy Map Canvas



3.2 Ideation and Brainstorming



Step-2: Brainstorm, Idea Listing and Grouping



3

Brainstorm as a group

Have everyone move their ideas into the "group sharing space" within the template and have the team silently read through them. As a team, sort and group them by thematic topics or similarities. Discuss and answer any questions that arise. Encourage "Yes, and..." and build on the ideas of other people along the way.

15 minutes

TIP

You can use the **Voting session** tool above to focus on the strongest ideas.

On each trip a bike is picked up at one station and dropped off at another. When a bike is picked up there's one less bike at the start station; when it's dropped off there's one more bike at the end station.

To represent this activity I start by reading a tripdata file for one month. I split it into two dataframes: one for pickups and and one for drop-offs.

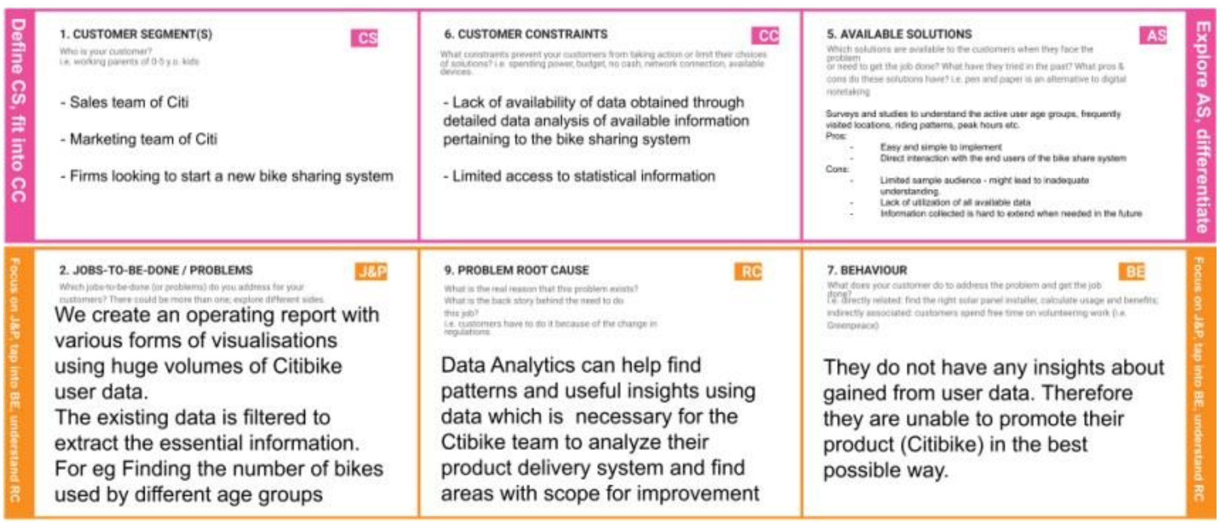
The number of bikes in the station at the beginning of the month is unknown.

It doesn't take into account rebalancing, bikes removed for maintenance or those added to the system.

3.3 Proposed Solution

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	People are unaware of the system and effects during bad weather conditions, Bike demands during peak hours.
2.	Idea / Solution description	Analysing the bike usage, no of trips, and the usage based on customer and subscriber's gender and age categories
3.	Novelty / Uniqueness	Understanding ,Exploring by creating data visualization by prediction of bike utilization,demand. 1.Overcome the Problem of Empty bicycle stations 2. Popular times of travel a. most common month b. most common day of week c. most common hour of day 3. Popular stations and trip a. most common start station b. most common end station c. most common trip from start to end
4.	Social Impact / Customer Satisfaction	1. Reduced congestion and fuel consumption 2.Transport flexibility 3. Reductions to vehicle emissions 4. Health benefits 5. Financial savings for individuals.
5.	Business Model (Revenue Model)	Having an membership active pass makes the customer can rent bikes with amount packages based on time-constraints per weeks/days. Subscriber can rent bikes with amount packages based on time-constraints per month/year.
6.	Scalability of the Solution	can improve the productivity of citi-bike system, widespread of utilization according to the customer's demands.

3.4 Problem Statement Fit



Identify strong TR & EM	3. TRIGGERS TR What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news. - Realizing how unhealthy they are becoming and finding out using bikes can be healthy - this makes the users use the bikes more often which gives the Citi teams more sales - Realizing how much pollution they are causing by making use of vehicles that give out CO_2	10. YOUR SOLUTION SL If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality. If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour. - Developing an interactive dashboard that gives various insights about details like finding the number of bikes used by different age groups, etc. - Different visualizations will be displayed on the dashboard for easy analysis. This makes it easier to take business decisions	8. CHANNELS of BEHAVIOUR CH 8.1 ONLINE What kind of actions do customers take online? Extract online channels from #7 8.2 OFFLINE What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development. **ONLINE:** The teams at Citi will be able to keep track of the statistics of the usage of Citi bikes online by looking at the dashboards and visualizations.	Identify strong TR & EM
	4. EMOTIONS: BEFORE / AFTER EM How do customers feel when they face a problem or a job and afterwards? i.e. lost, insecure > confident, in control - use it in your communication strategy & design. - Users of the bikes will feel extremely satisfied after a good ride which in turn will give the teams at Citi satisfaction - Customers will feel good about giving back to the community by reducing carbon footprint		**OFFLINE:** The teams at Citi will be involved in offline work like installing new bike hubs and trying to work off site to find the problems faced by users of the Citi bike. They also try to keep new bikes in stock in all hubs.	

4. REQUIREMENT ANALYSIS

4.1 Functional Requirement

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIn
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Collection of Data	Utilizing the NYC Citi Bike assists in gathering information on the various trips that various users of Citi Bike take. These data were then organised into datasets and made available for further study and visualisation.
FR-4	Analysis of Data	Preprocessing and filtering the provided data in accordance with the sub-requirements task's is part of the analysis process. Data analysis and visualisation are both aided by the use of machine learning algorithms to glean more insights from the data..
FR-5	Display (Visualization) of Data	Various visualisations are used depending on the sub-task being handled. These visualisations are then combined and shown on a dashboard, which is a tool for giving customers business information. Finding the top 10 Start stations according to customer age group and showing the most popular bikes according to trip time are a few of the various sub-tasks included in this requirement.

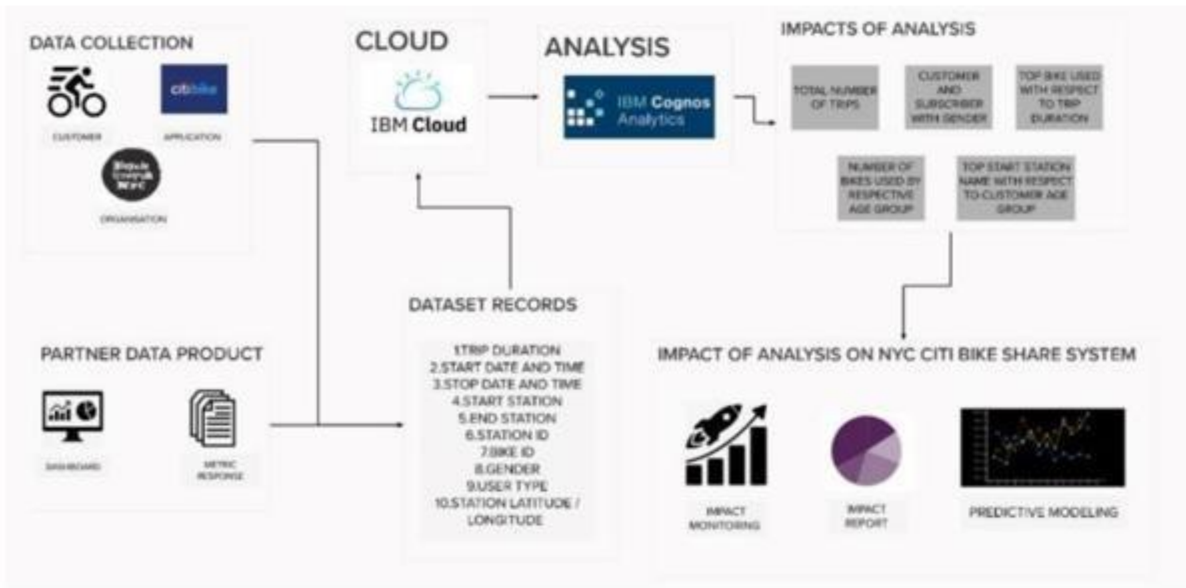
4.2 Non-Functional Requirement

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The dashboard gives users access to an operational report that is simple to read and useful for understanding market trends and company insights. Data can be examined from various angles and in more depth by using an interactive dashboard to drill down and filter operating information.

NFR-2	**Security**	Based on the Citi Bike utilisation data and its analysis, several important business decisions will be made, which will be appropriately secured. Data and visualisation reports are only available to a certain group of clients/users.
NFR-3	**Reliability**	This research offers a trustworthy and effective way to understand how well this bike-sharing programme performed in 2018. Utilizing the IBM Cognos Platform ensures operational report production, upkeep, and accessibility with industry-standard reliability (dashboard).
NFR-4	**Performance**	The effectiveness of a bike-sharing system in terms of both its spatial and operational efficiency. In order to increase the operational effectiveness of the bike-sharing system, it is critical to assess the state of bike lanes from the viewpoint of public bike riders. The characteristics of bike stations and the distance between bike stations and other amenities are examined by the bike-sharing system dashboard. The evaluation findings can be used to enhance the public bike-sharing service.
NFR-5	**Availability**	The bicycle-sharing programme is a form of shared transportation in which people can rent bicycles at a reasonable cost for a limited amount of time. CitiBike offers two different kinds of docking systems: docking systems, which allow customers to borrow a bike from one dock and return it to another port within the system; and dockless systems, which are node-free and depend on smart technology. Both forms can use smartphone online mapping to find close-by ports and bikes that are available.
NFR-6	**Scalability**	Urban inhabitants can immediately get access to bike-sharing programmes, which may make the transportation system more dependable. The programme can be expanded to include locations that are now unreachable by this type of transportation, as well as cities other than New York City, if the necessary data is available and obtained. This research will eventually be able to give a more in-depth picture of how bike-sharing functions in emergency situations as additional data becomes available, particularly in other cities with comparable extensive bike-sharing systems.

5. PROJECT DESIGN

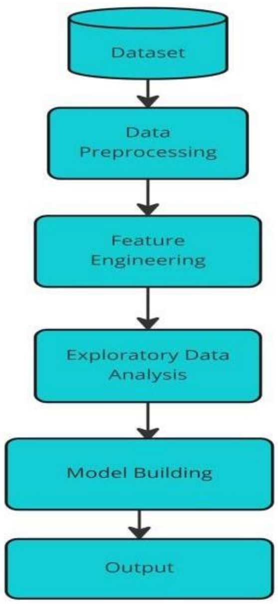
5.1 Data Flow Diagram



User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer, Analysts, Organizations, Government	Collection of user data	USN-1	Lyft citi bike's official website provides the data to help with analysis, development, visualization etc. Data is collected from these published files.	I can access the data on Lyft citi bike's official website	High	Sprint-1
Customer, Analysts, Organizations, Government	Analysing the user data	USN-2	This data is used as input for creating various types of visualizations and analysis	I can view the analysis of the citi bike	High	Sprint-1

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer, Analysts, Organizations, Government	Collection of user data	USN-1	Lyft citi bike's official website provides the data to help with analysis, development, visualization etc. Data is collected from these published files.	I can access the data on Lyft citi bike's official website	High	Sprint-1
Customer, Analysts, Organizations, Government	Analysing the user data	USN-2	This data is used as input for creating various types of visualizations and analysis	I can view the analysis of the citi bike	High	Sprint-1

5.2 Solution & Technical Architecture



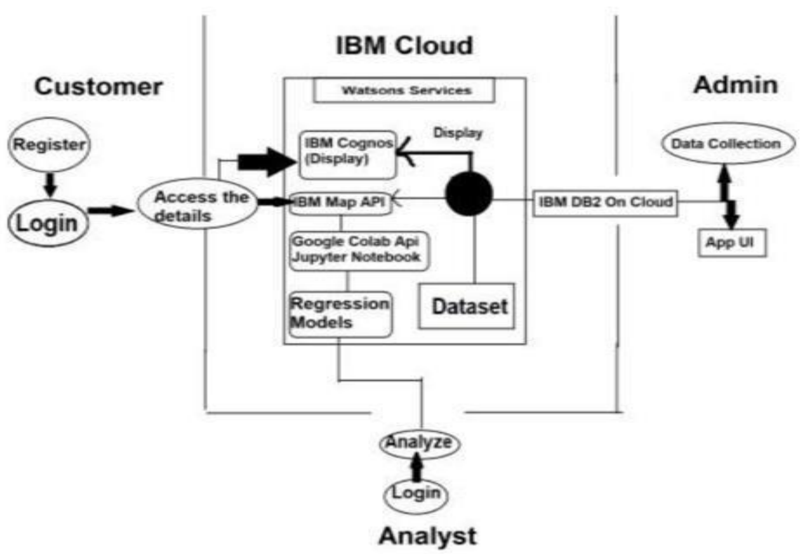


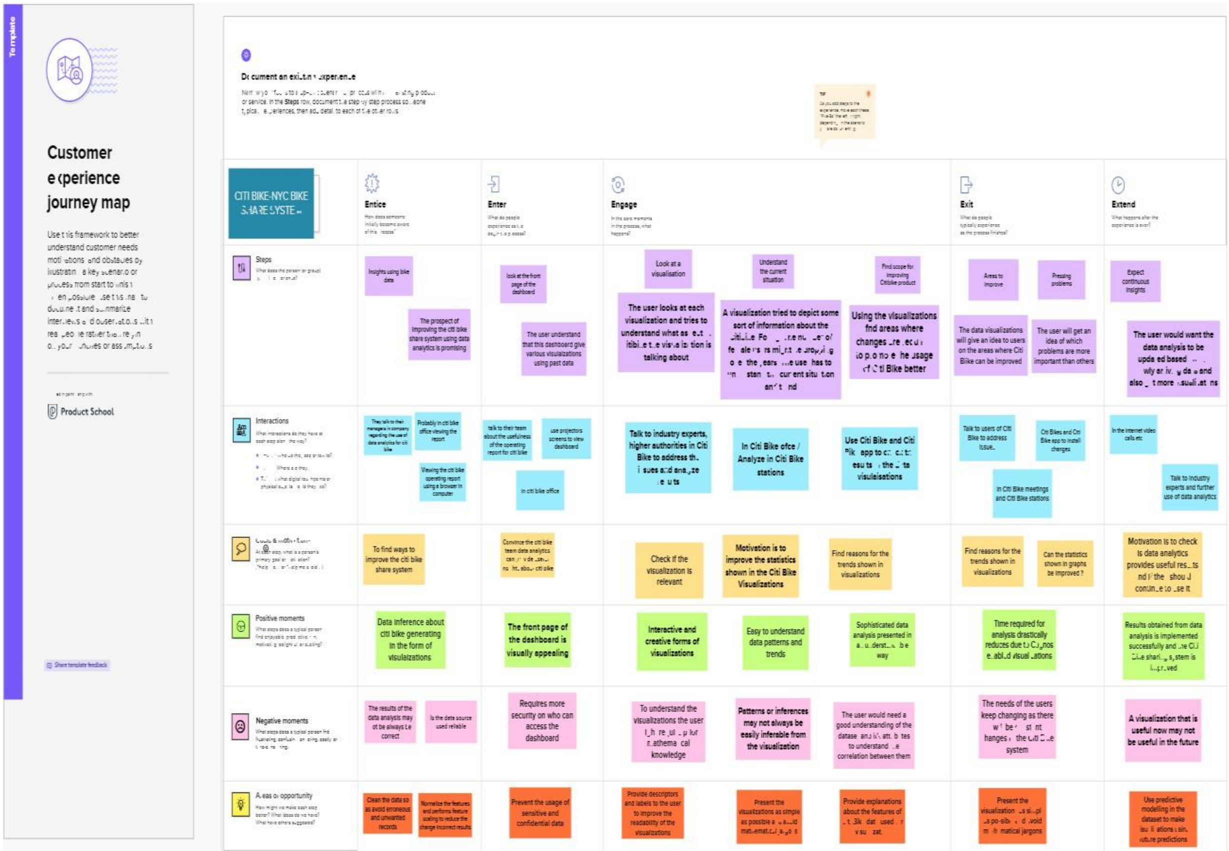
Table-1 :Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	User can Interact with web Application	HTML, CSS, JavaScript .
2.	Data Pre processing	Pre processing of data should be done	Python
3.	Feature Engineering	Feature engineering of Dataset by adding new values to the existing dataset.	Python
4.	Exploratory Data Analysis	Exploring the data using boxplot, pie plot, scatter plot etc..	Python
5.	Model Building	Build the model using machine learning algorithm	python
6.	Data Storage	Database Service on Cloud	IBM DB2, IBM Cloudant etc.
7.	User Interface	Dashboard showing the details of the trip duration, no of trips, bike usage etc..	HTML,CSS, JavaScript.

Table-2: Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Security Implementations	The main security concern is for users account hence proper login mechanism should be used to avoid hacking.	e.g. SHA-256, Encryptions, IAM Controls, OWASP etc.
2.	Availability	The system shall be available 24 hours a day 7 days a week. User can access at anytime	
3.	Performance	The system should require a fair amount of speed especially while browsing through the catalogue	

5.3 User Stories


Template
Customer
journey map
Use this framework to better
understand customer needs
Product School
CITI BIKE-NYC BIKE
Entice
Enter
Engage
Exit
Extend
Steps
Insights using bike data
The prospect of improving the citi bike share system using data analytics is promising
look at the front page of the dashboard
The user understand that this dashboard give various visualizations using past data
Look at a visualisation
Understand the current situation
Find scope for improving Citibike product
The user looks at each visualization and tries to
talking about
A visualization tried to depict some sort of information about the
Using the visualizations find areas where
Areas to improve
Pressing problems
The data visualizations will give an idea to users on the areas where Citi Biike can be improved
The user will get an idea of which problems are more important than others
Expect continuous insights
The user would want the data analysis to be
Interactions
They talk to their managers in company regarding the use of data analytics for citi bike
Probably in citi bike office viewing the report
Viewing the citi bike operating report using a browser in computer
talk to their team about the usefulness of the operating report for citi bike
use projectors screens to view dashboard
In citi bike office
Talk to industry experts, higher authorities in Citi Bike to address th.
In Citi Bike ofce / Analyze in Citi Bike stations
Use Citi Bike and Citi
visulaisations
Talk to users of Citi Bike to address
Citi Bikes and Citi Bike app to install changes
In Citi Bike meetings and Citi Bike stations
In the internet video calls etc
Talk to industry experts and further use of data analytics
To find ways to improve the citi bike share system
Convince the citi bike team data analytics
Check if the visualization is relevant
Motivation is to improve the statistics shown in the Citi Bike Visualizations
Find reasons for the trends shown in visualizations
Find reasons for the trends shown in visualizations
Can the statistics shown in graphs be improved ?
Motivation is to check is data analytics provides useful
Positive moments
Data inference about citi bike generating in the form of visulaizations
The front page of the dashboard is visually appealing
Interactive and creative forms of visualizations
Easy to understand data patterns and trends
Sophisticated data analysis presented in
Time required for analysis drastically
Results obtained from data analysis is implemented successfully and
Negative moments
The results of the data analysis may
correct
Is the data source used reliable
Requires more security on who can access the dashboard
To understand the visualizations the user
knowledge
Patterns or inferences may not always be easily inferable from the visualization
The user would need a good understanding of the
correlation between them
The needs of the users keep changing as there
system
A visualization that is useful now may not be useful in the future
Clean the data so as avoid erroneous and unwanted records
Normalize the features and performs feature scaling to reduce the change incorrect results
Prevent the usage of sensitive and confidential data
Provide descriptors and labels to the user to improve the readability of the visualizations
Present the visualizations as simple as possible
Provide explanations about the features of
Present the visualization
Use predictive modelling in the dataset to make

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data preparation	USN-1	As an analyst. I can extract the Citi bike dataset for the year 2018.	5	High	Ramnath, Jeyanth
Sprint-1		USN-2	As an analyst, I upload the dataset into Cognos platform.	6	High	Vignesh Vasanth
Sprint-1	Data Cleaning	USN-3	As an analyst, I remove the null and duplicate values.	4	Medium	Jeyanth, Vignesh
Sprint-1		USN-4	As an analyst, I identify patterns and relationships between the various attributes	5	High	Ramnath, Vasanth
Sprint-2	Feature Engineering	USN-5	I made computations on the different attribute to find the new attribute value.	4	Medium	Ramnath, Jeyanth
Sprint-2		USN-6	I have dropped few attributes from the data set which are not needed.	6	High	Vignesh, Vasanth
Sprint-2	Visualization	USN-7	As an analyst, I visualize the data and infer the knowledge in Cognos platform.	10	High	Ramnath, Vasanth
Sprint-3		USN-8	As an analyst, I made visualization charts of the data using python	10	High	Vignesh, Jeyanth
Sprint-3	Dashboard	USN-9	As an analyst, I create a dashboard with the created visualizations to supplement business insights during the decision-making process at Citi dataset.	10	High	Ramnath, Jeyanth
Sprint-4	Prediction	USN-10	To predict the most common user type ie customers and subscribers using various machine learning algorithms.	10	High	Vignesh, Vasanth
Sprint-4	Registration	USN-11	As a user, I can register and login in the application.	10	High	Ramnath, Vasanth

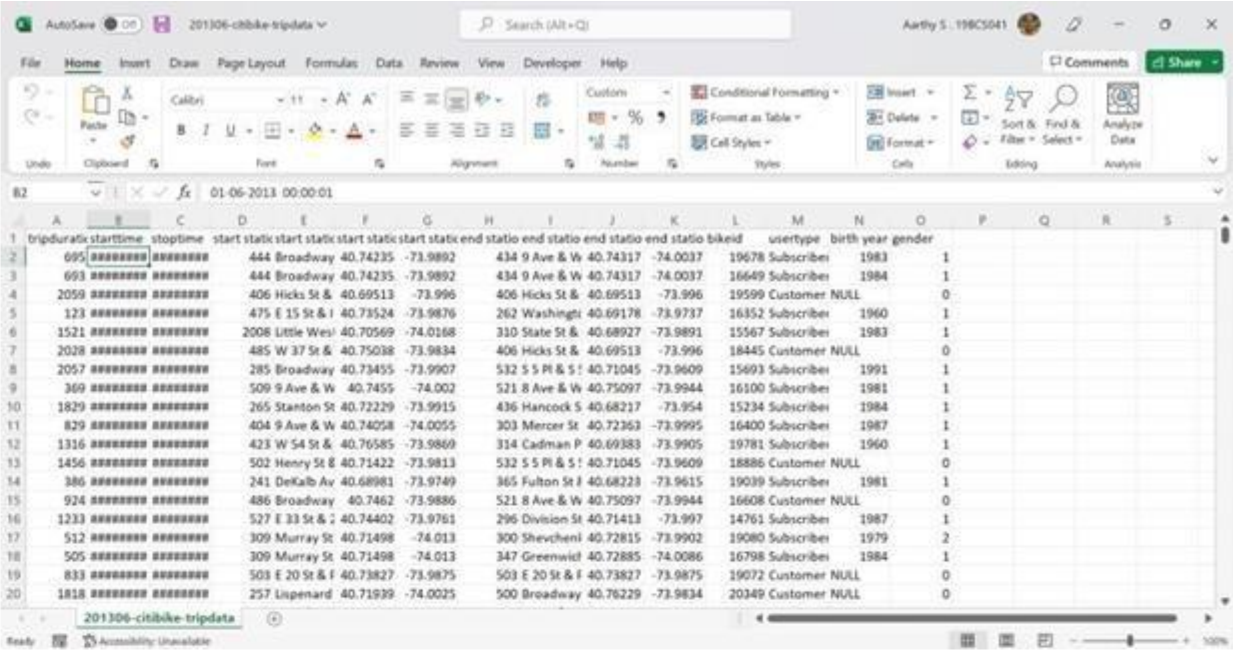
Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)

Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

6.2 Sprint Delivery Schedule

7. WORKING WITH THE DATASET & DATA VISUALISATION

7.1 Understanding the dataset



tripduratic	starttime	stoptime	start static	start static	start static	start static	end statio	end statio	end statio	end statio	bikeid	usertype	birth year	gender
695	########	########	444	Broadway	40.74235	-73.9892	434	9 Ave & W	40.74317	-74.0037	19678	Subscriber	1983	1
693	########	########	444	Broadway	40.74235	-73.9892	434	9 Ave & W	40.74317	-74.0037	16649	Subscriber	1984	1
2059	########	########	406	Hicks St &	40.69513	-73.996	406	Hicks St &	40.69513	-73.996	19599	Customer	NULL	0
123	########	########	475	E 15 St & I	40.73524	-73.9876	262	Washingt	40.69178	-73.9737	16352	Subscriber	1960	1
1521	########	########	2008	Little Wes	40.70569	-74.0168	310	State St &	40.68927	-73.9891	15567	Subscriber	1983	1
2028	########	########	485	W 37 St &	40.75038	-73.9834	406	Hicks St &	40.69513	-73.996	18445	Customer	NULL	0
2057	########	########	285	Broadway	40.73455	-73.9907	532	S 5 Pl & S	40.71045	-73.9609	15693	Subscriber	1991	1
369	########	########	509	9 Ave & W	40.7455	-74.002	521	8 Ave & W	40.75097	-73.9944	16100	Subscriber	1981	1
1829	########	########	265	Stanton St	40.72229	-73.9915	436	Hancock S	40.68217	-73.954	15234	Subscriber	1984	1
829	########	########	404	9 Ave & W	40.74058	-74.0055	303	Mercer St	40.72363	-73.9995	16400	Subscriber	1987	1
1316	########	########	423	W 54 St &	40.76585	-73.9869	314	Cadman P	40.69383	-73.9905	19781	Subscriber	1960	1
1456	########	########	502	Henry St 8	40.71422	-73.9813	532	S 5 Pl & S	40.71045	-73.9609	18886	Customer	NULL	0
386	########	########	241	DeKalb Av	40.68981	-73.9749	365	Fulton St	40.68223	-73.9615	19039	Subscriber	1981	1
924	########	########	486	Broadway	40.7462	-73.9886	521	8 Ave & W	40.75097	-73.9944	16608	Customer	NULL	0
1233	########	########	527	E 33 St &	40.74402	-73.9761	296	Division St	40.71413	-73.997	14761	Subscriber	1987	1
512	########	########	309	Murray St	40.71498	-74.013	300	Shevchenl	40.72815	-73.9902	19080	Subscriber	1979	2
505	########	########	309	Murray St	40.71498	-74.013	347	Greenwicl	40.72885	-74.0086	16798	Subscriber	1984	1
833	########	########	503	E 20 St & F	40.73827	-73.9875	503	E 20 St & F	40.73827	-73.9875	19072	Customer	NULL	0
1818	########	########	257	Lispenard	40.71939	-74.0025	500	Broadway	40.76229	-73.9834	20349	Customer	NULL	0

Dataset Link: [Dataset]

1. Trip Duration: How long a trip lasted in seconds

2. Start Date and Time: EX->01-06-2013 00:00:01

3. Stop Date and Time: EX->01-06-2013 00:11:36

4. Start Station ID: Unique identifier for each station

5. Start Station Name

6. Start Station Latitude: Coordinates

7. Start Station Longitude: Coordinates

8. End Station ID: Unique identifier for each station

9. End Station Name

10. End Station Latitude

11. End Station Longitude

12. Bike ID: Unique identifier for each bike

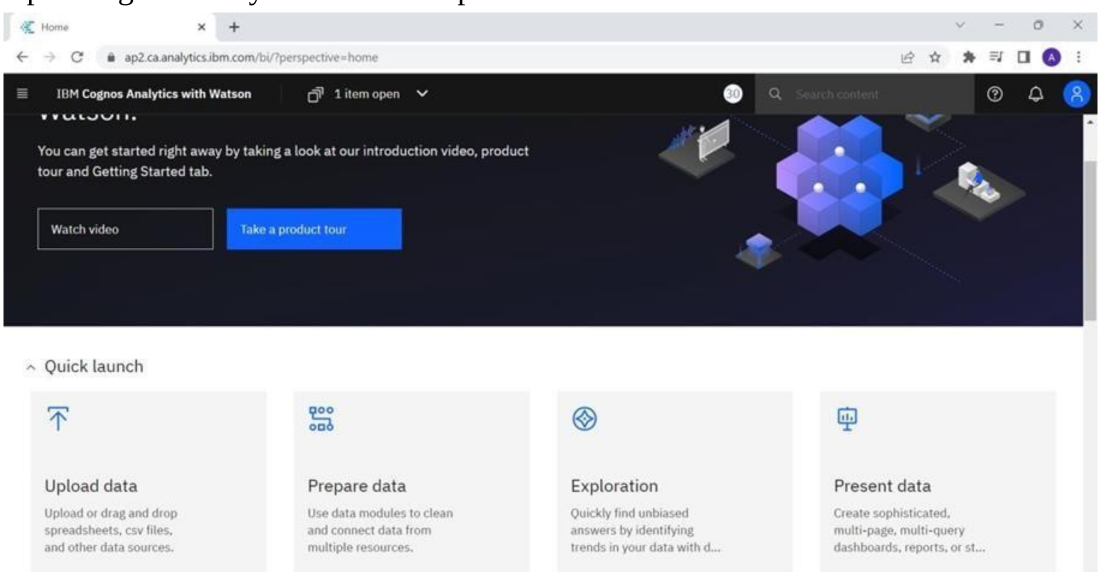
13. User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member): Customers are usually tourists, subscribers are usually NYC residents

14. Year of Birth: Self-entered, not validated by an ID

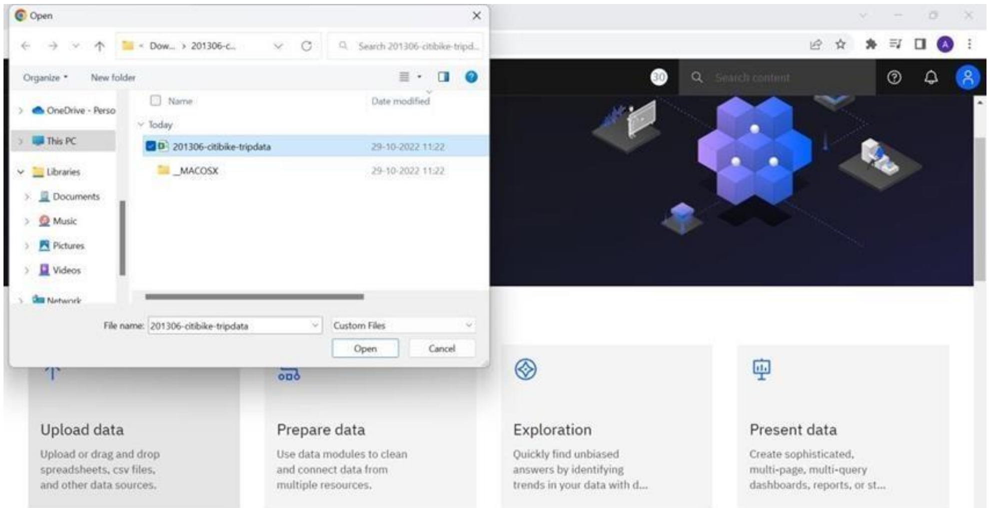
Gender (Zero=unknown; 1=male; 2=female): Usually unknown for customers since they often sign up at a kiosk

7.2 Loading the dataset

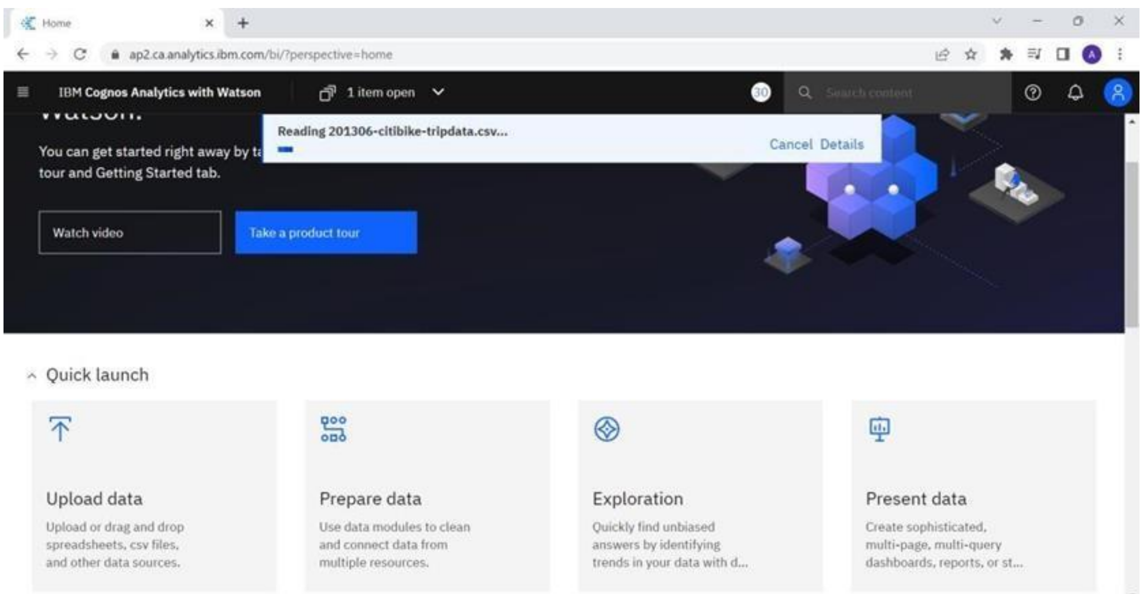
Open Cognos Analytics and click upload data



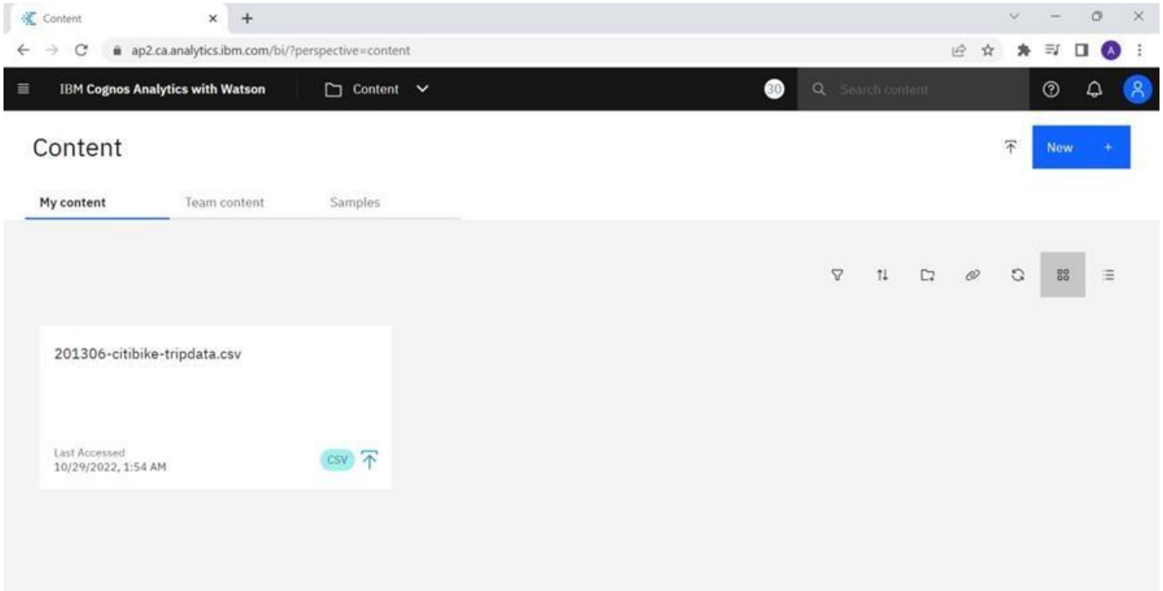
Select the dataset to be uploaded



The excel file is getting uploaded in Cognos Analytics

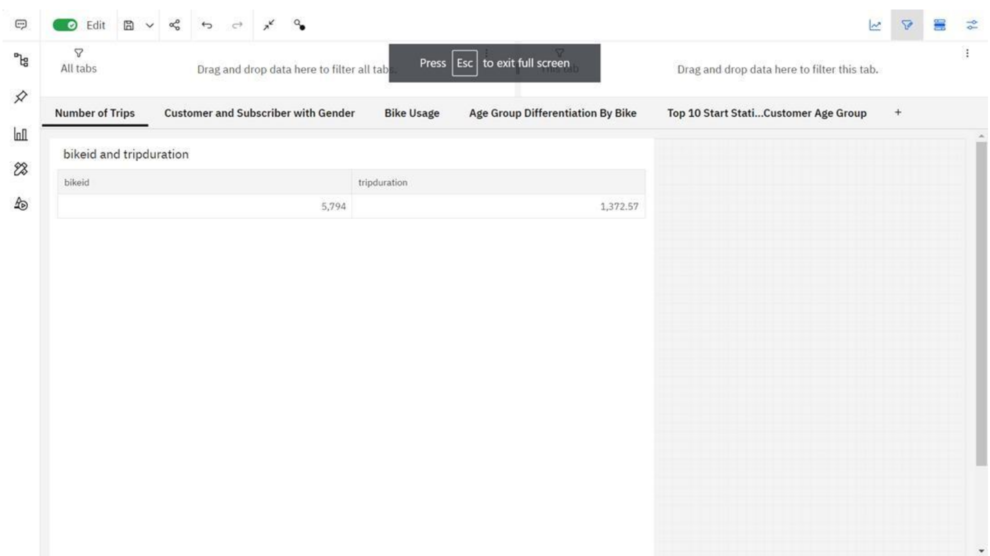


The dataset can be accessed in My Content in Cognos Analytics

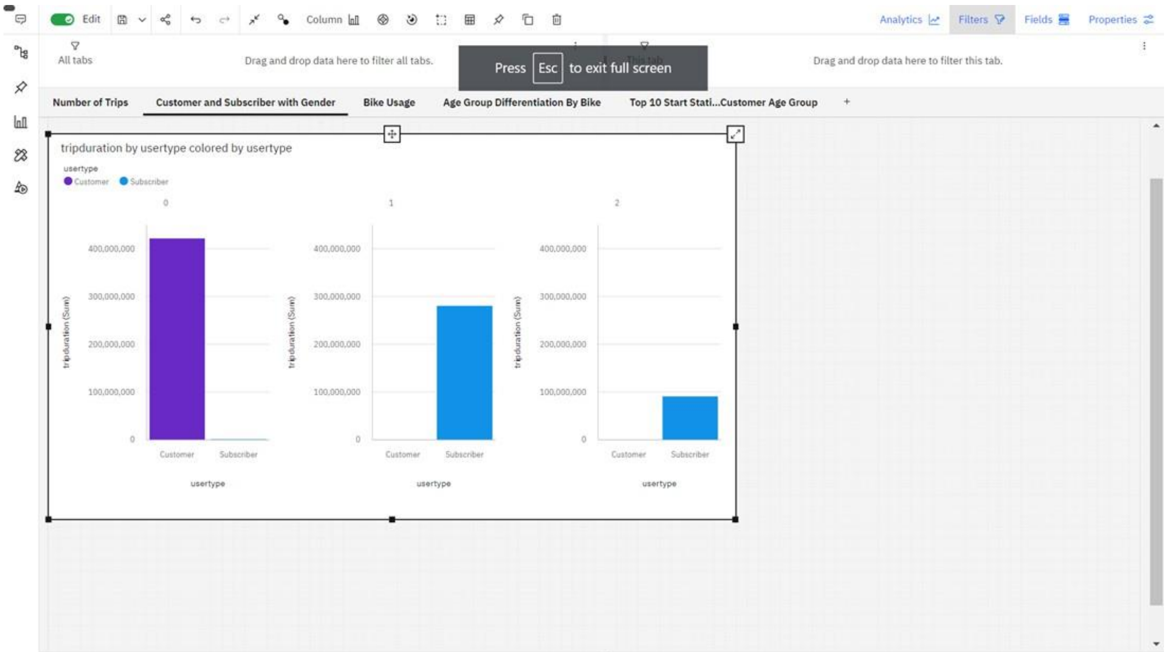


7.3 Visualization charts

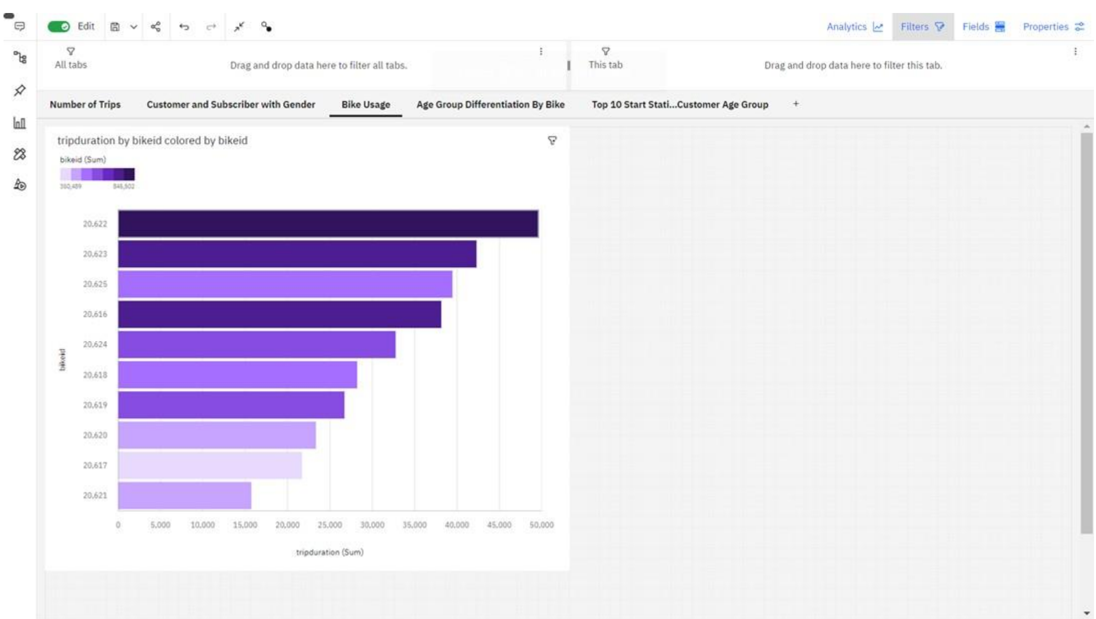
Number of Trips:



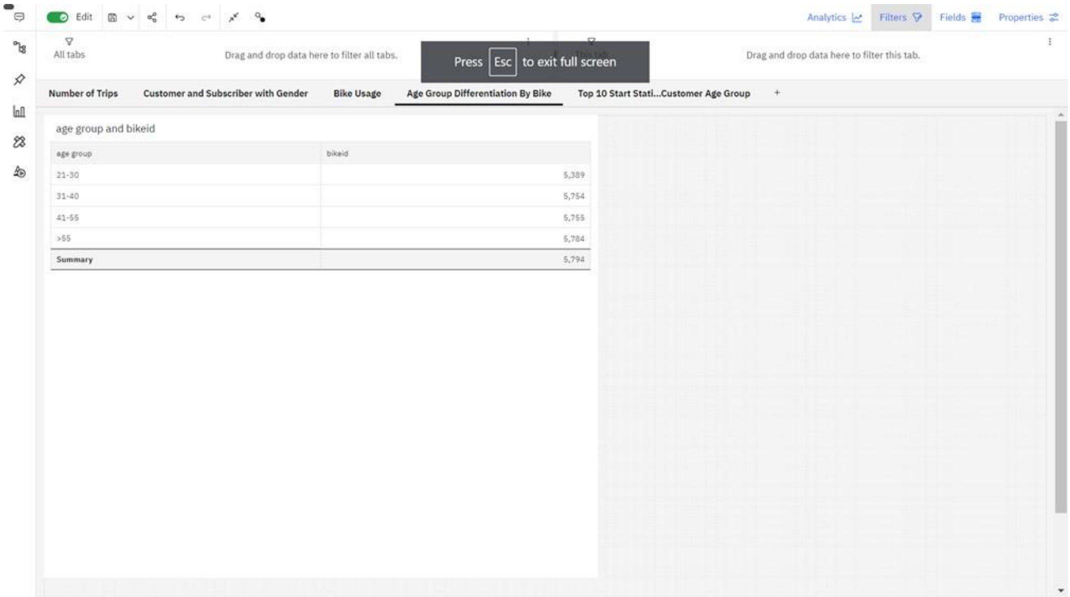
Customer and Subscriber with Gender:



Bike Usage:

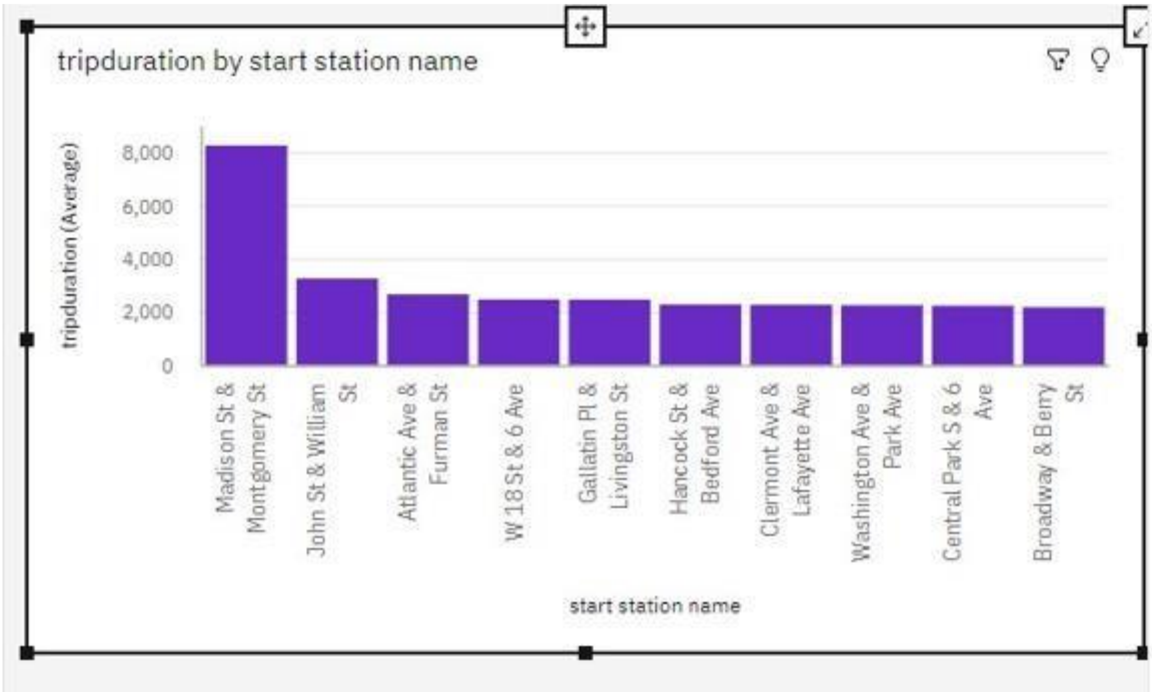


Age group differentiation by bike:

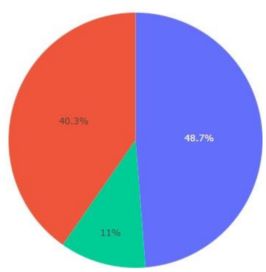


age group	bikeid
21-30	5,389
31-40	5,754
41-55	5,755
>55	5,784
Summary	5,794

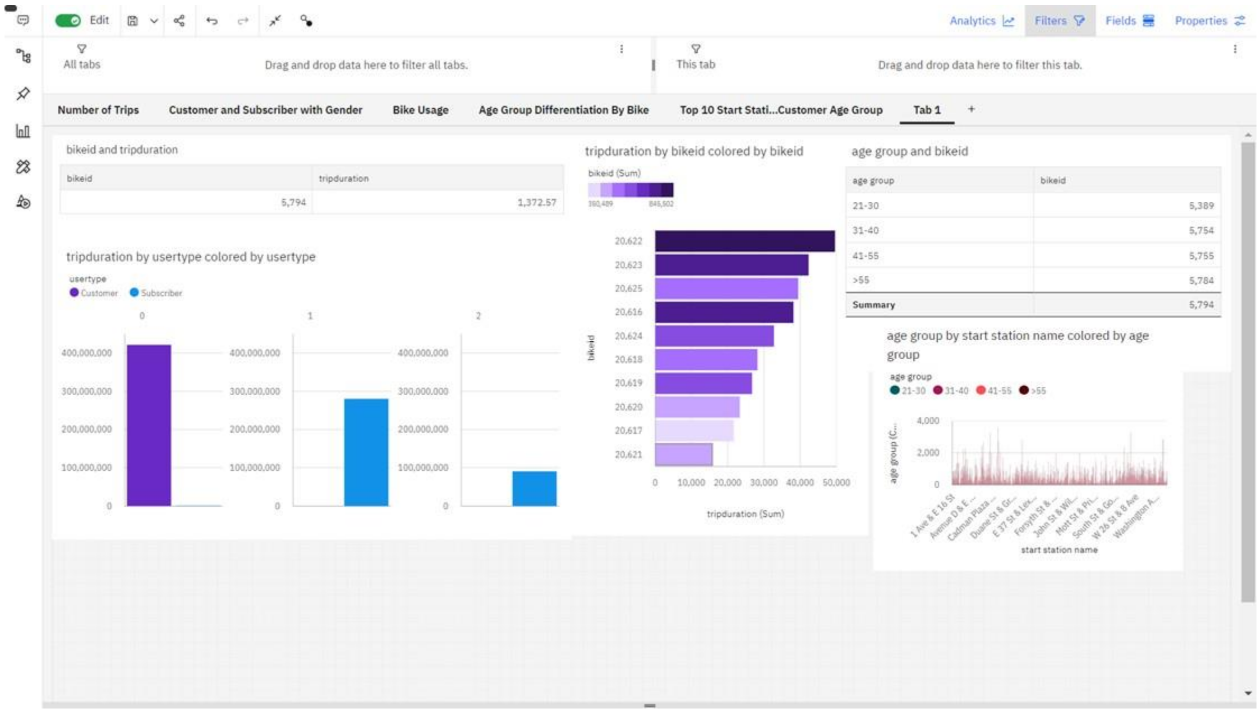
Top 10 Start Station Names with Respect to Customer Age Group:



Gender Variation



8. CREATING THE DASHBOARD



9. ADVANTAGES AND DISADVANTAGES

The benefits of bike sharing schemes include transport flexibility, reductions to vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals.

One can easily analyze and understand trends in bike sharing patterns with the created dashboard. With no prior skills and knowledge about the tools that we use for analysis, anyone (literate or illiterate) can easily infer the knowledge that we represent in various charts or graphs or maps. So that it would be helpful to users and companies to make appropriate decisions in the future**.**

10. CONCLUSION

Based on the quantitative as well as visual analysis of the New York bike share system, a number of interesting insights were gained.

One obvious conclusion was that there is a strong seasonal variation in the system usage with maximum usage in summer and minimum usage in winter. This was initially hypothesized because of the harshness of New York's harsh winters and the treacherous riding conditions that exist during that time. However, despite the adverse weather conditions, there is a strong core demographic that consistently uses the system. This conclusion is based on that fact that even during the months of January and February which are the peak winter months, there are more than two hundred thousand trips in the system.

New York has a strong public transit system, and the bike share system seems to complement it quite well with a majority of the highest used stations located either close to subway lines or the commuter rail stations in the city.

Based on the locations of the stations and the duration of trips, it can be hypothesized that bike shares are replacing last mile trips that would otherwise be done either on foot or on public transit. This is particularly true in case of New York where a combination of dense public transit network, the road congestion during peak hours and the average trip distance as calculated create a situation where the only potential trips that the bike share system is replacing currently are those that would otherwise have been undertaken either on foot or on public bus.

11. FUTURE SCOPE

NYC is a very crowded and happening place which leads to lots of pollution. And in this busy world people are always worried about transportation this bike sharing system reduces that stress. With increase in population pollution also increases. So it is in our hands to reduce pollution and to make a better future for our younger generations. We can analyze which station needs more bikes and any area needs new station to be installed. The survey outcomes indicates the needs for improved techniques in bike sharing analytics. There exists a lot of scope in this research area.

12. SOURCE CODE

```
#%% md

# SPRINT **3**

#%%

import pandas as pd
import numpy as np
import seaborn as sns
```

```
import matplotlib.pyplot as plt
import plotly.express as px
from datetime import datetime
from pprint import pprint

from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials


#%%

path = "/content/dataset.csv"
df = pd.read_csv(path)
print(df)




#%%

df.head()

#%%

df.describe()

#%%

df.info()

#%%

df.isnull().sum()

#%%

df[df['starttime'].isnull()]

#%%
```

```
df[df['stoptime'].isnull()]

#%%

df = df[:-1]

#%%

df.isnull().sum()

#%%

print(type(df["start station latitude"][0]))
print(df["start station latitude"][0])

#%%

df['start station name'].unique()

#%%

def camel_case(city):
   try:
       city = city.split(' ')
       city = ' '.join([x.lower().capitalize() for x in
city])
       if city == 'Unknown':
           return np.nan
       else:
           return city
   except:
       return np.nan

# Apply camel_case function to City column
df['start station name'] = df['start station
name'].apply(camel_case)
df['start station name'].value_counts()

#%%
```

```
df.count()

#%%

df["tripduration"] = pd.to_numeric(df["tripduration"])
res = df.iloc[52323]
print(res["tripduration"])

#%%

df_filtered = df[df['tripduration'] != "tripduration"]
df_filtered["tripduration"] =
pd.to_numeric(df_filtered["tripduration"])
df = df_filtered
type(df["tripduration"][0])

#%%

type(df["start station latitude"][0])

#%%

type(df["end station longitude"][0])

#%%

type(df["bikeid"][0])

#%%

type(df["birth year"][0])

#%%

type(df["gender"][0])

#%%

type(df["starttime"][0])

#%%
```

```
df["starttime"] = pd.to_datetime(df["starttime"])
df["stoptime"] = pd.to_datetime(df["stoptime"])
type(df["starttime"][0])

#%%

df["starttime"][0] < df["stoptime"][0]

#%%

df.info()

#%%

def find_outliers_IQR(df):
  q1=df.quantile(0.25)
  q3=df.quantile(0.75)
  IQR=q3-q1
  outliers = df[((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR)))]
  return outliers
outliers = find_outliers_IQR(df["birth year"])
print("number of outliers: " + str(len(outliers)))
print("max outlier value: " + str(outliers.max()))
print("min outlier value: " + str(outliers.min()))

#%%

df["gender"].value_counts()

#%%

temp_df = df[df["birth year"] <= 1957]
temp_df["gender"].value_counts()

#%%

df.shape

#%%
```

```
df.to_csv('cleaned_dataset.csv', index=False)

#%% md

# **SPRINT 4**

#%%

path = "/content/cleaned_dataset.csv"
edadf = pd.read_csv(path)
print(edadf)




#%%

temp = edadf

#%%

temp.head()

#%%

temp.describe()

#%%

temp.info()

#%%

temp["starttime"] = pd.to_datetime(temp["starttime"])
temp["stoptime"] = pd.to_datetime(temp["stoptime"])
temp.info()
temp["Hour"] = temp["stoptime"].dt.hour -
temp["starttime"].dt.hour
temp.head()

#%%
```

```
temp.shape

#%%

temp['Age'] = 2022 - temp['birth year']
temp.head()

#%%

Age_Groups = ["<20", "20-29", "30-39", "40-49", "50-59",
"60+"]
Age_Groups_Limits = [0, 20, 30, 40, 50, 60, np.inf]
Age_Min = 0
Age_Max = 100
temp["Age_group"] = pd.cut(temp["Age"], Age_Groups_Limits,
labels=Age_Groups)
temp.head()

#%%

trips_df = pd.DataFrame()
trips_df = temp.groupby(['start station name','end station
name']).size().reset_index(name = 'Number of Trips')
trips_df = trips_df.sort_values('Number of Trips',ascending
= False)
trips_df["start station name"] = trips_df["start station
name"].astype(str)
trips_df["end station name"] = trips_df["end station
name"].astype(str)
trips_df["Routes"] = trips_df["start station name"] + " to
" + trips_df["end station name"]
trips_df  =  trips_df[:50]
trips_df = trips_df.reset_index()
trips_df

#%%

px.pie(values = temp['gender'].value_counts(),
      names =temp['gender'].value_counts().index,
      title ="Gender Variation")
```

```
#%%

px.bar(x=temp["start station name"].value_counts().index,
       y=temp["start station name"].value_counts().values,
       labels={'x':'Start Station Name',"y":"Count"})

#%%

px.bar(x=temp["end station name"].value_counts().index,
       y=temp["end station name"].value_counts().values,
       labels={'x':'End Station Name',"y":"Count"})

#%%

px.bar(x=temp["Hour"].value_counts().index,
       y=temp["Hour"].value_counts().values,
       title = "Hour usage of Citi Bikes",
       labels={'x':'Time',"y":"Number of people using
bike"})
```

13. GITHUB LINK

https://github.com/IBM-EPBL/IBM-Project-15888-1659605911

14. DEMO LINK:

https://drive.google.com/file/d/1m2bC4Y2mtFK5re2nBt_5jKwhV8fh23EZ/view?usp=share_link