A New Hint to Transportation-Analysis of the NYC Bike Share System

IBM-39037-1662629538

NALAIYA THIRAN PROJECT BASED LEARNING ON PROFESSIONAL READLINESS FOR INNOVATION, EMPLOYNMENT AND ENTERPRENEURSHIP

A PROJECT REPORT

ANEESH AHMED A	**722819104007**
ANIRUTHRAM K P	**722819104008**
DEEPAN KUMAR S	**722819104026**
FAYAZ A	**722819104028**

BACHELOR OF ENGINEERING

COMPUTER SCIENCE AND ENGINEERING
SRI ESHWAR COLLEGE OF ENGINEERING
COIMBATORE-641202

1.INTRODUCTION

1.1 Project Overview

Bike share programs have risen in popularity in recent yearsandhavebeen promoted as a lower carbon alternative to other forms of transit.Interest in bicycle sharing has been growing exponentiallyoverthepastdecade, resulting in a proliferation of bike share systems in712citiesacross the world, encompassing 806,000 bicycles and 37,500stations.This can be largely attributed to the successful incorporationofinformation technology in docking stations and mobile devicesaswellas improved logistics such as bicycle rebalancing to ensureresponsivesupply management. Cities often hope bike sharing will bringmanybenefits such as extending the reach of transit, substitutingmotorizedtrips, and encouraging non-cyclists to try cycling. The premise of bicycle sharing is that it is a short-termbikerentalsystem, based on varying timed memberships. Members of the bike share network have access to stations, consisting of a pay-stationandmultiple bike docks, across the system where bikes canbe checkedoutfrom one station and returned to another nearest to their destination.The appeal of membership is 24/7 access to an automatedbikerentalnetwork and utility of bikes in completing “last-kilometer connections”without the worry of storage or maintenance. The price systemissettoencourage shorter trips (less than 30 minutes in time), withadditionalfees for any time used over that maximum.

There is evidence that bike share users switch to bike share frommotorizedtransport, such as bus and auto, creating the potential for significantreductions in transportation related greenhouse gas or CO2e emissions.However, there is significant heterogeneity between different cities, showingthat there is not a guaranteed CO2e reduction benefit frominstitutingbikeshare, especially if the trips would not have been made otherwiseoraresubstituting walking and private bicycle trips.

1.2 Purpose

The purpose of this analysis is to create an operating report of CitiBikefor the year 2018. From this analysis, the following datavisualizationswill 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 fromindividuals or bike rental shops. Zagster - Life is better on a bike! They are bringing bike sharetocommunities across the USA. Motivate International - Motivate is a global full-service bikeshareoperator and technology innovator. Spin - Spin is a stationless bike and electric scooter sharingservice.

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 systemusers?"

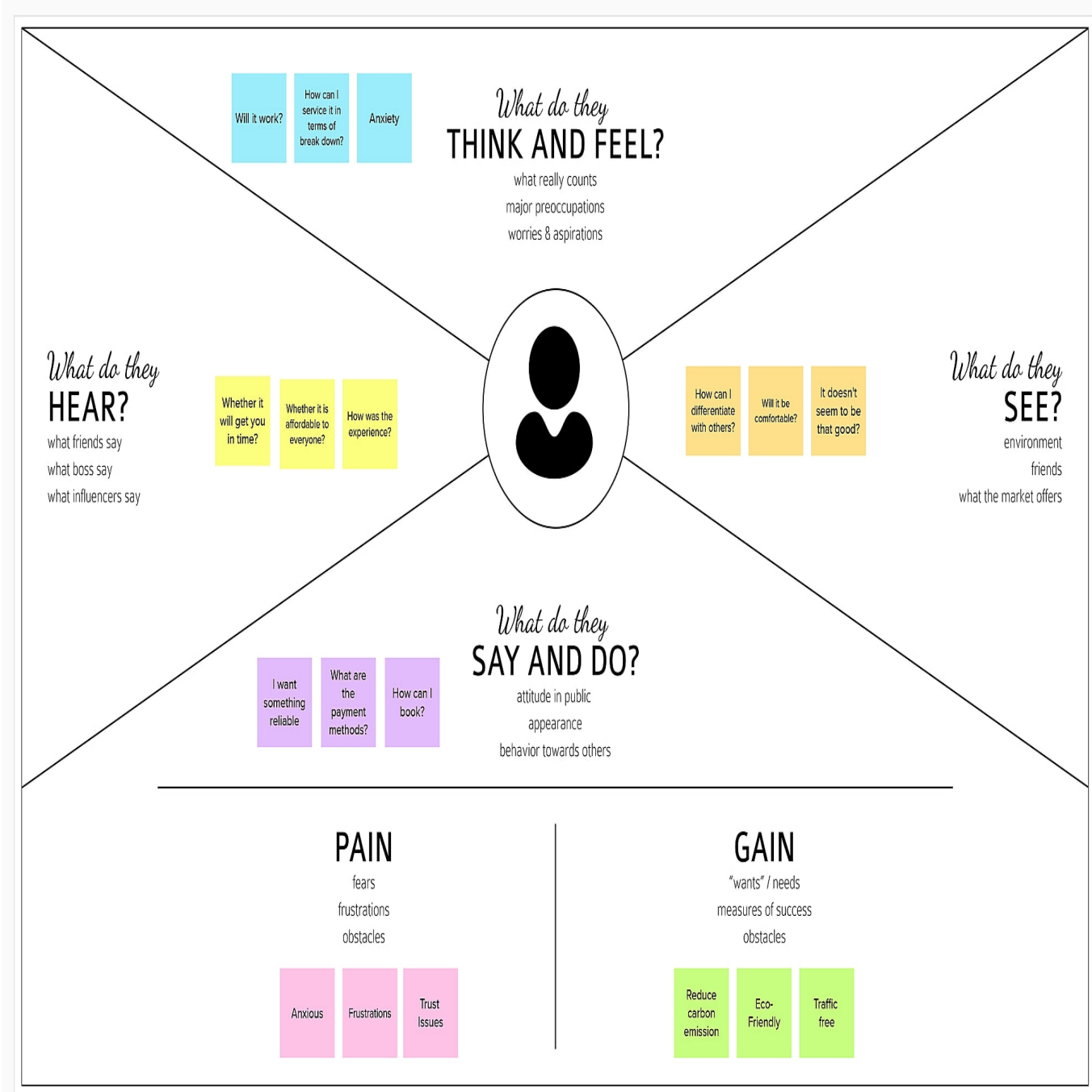
FRANCESCO et al.,IEEE Access 2020"Bike Sharing and UrbanMobilityina Post-Pandemic" "A long-term perspective on the COVID-19: The bike sharingsystem resilience under the epidemic environment"Journal of Transport&Health ,2021 Nguyen ThiHoai Thu, Chu Thi Phuong Dung, Vietnam2017InternationalConference on Advanced Technologies for Communications - Multi-sourceData Analysis for Bike Sharing Systems

2.3 Problem statement

Definition In busy cities like New York the people are facing difficultiesinanalyzing the demand for bikes during peak hours. The main objective of this project is to predict bike patterns thatwillbe extremely helpful for people to plan their travel.

3.IDEATION & PROPOSED SOLUTION

3.1 EMPATHY MAP


Will it work?
How can I service it in terms of break down?
Anxiety
What do they
THINK AND FEEL?
what really counts
major preoccupations
worries & aspirations
What do they
HEAR?
what friends say
what boss say
what influencers say
Whether it will get you in time?
Whether it is affordable to everyone?
How was the experience?
How can I differentiate with others?
Will it be comfortable?
It doesn't seem to be that good?
What do they
SEE?
environment
friends
what the market offers
I want something reliable
What are the payment methods?
How can I book?
What do they
SAY AND DO?
attitude in public
appearance
behavior towards others
PAIN
fears
frustrations
obstacles
Anxious
Frustrations
Trust Issues
GAIN
"wants" / needs
measures of success
obstacles
Reduce carbon emission
Eco-Friendly
Traffic free

3.2 Proposed Solution

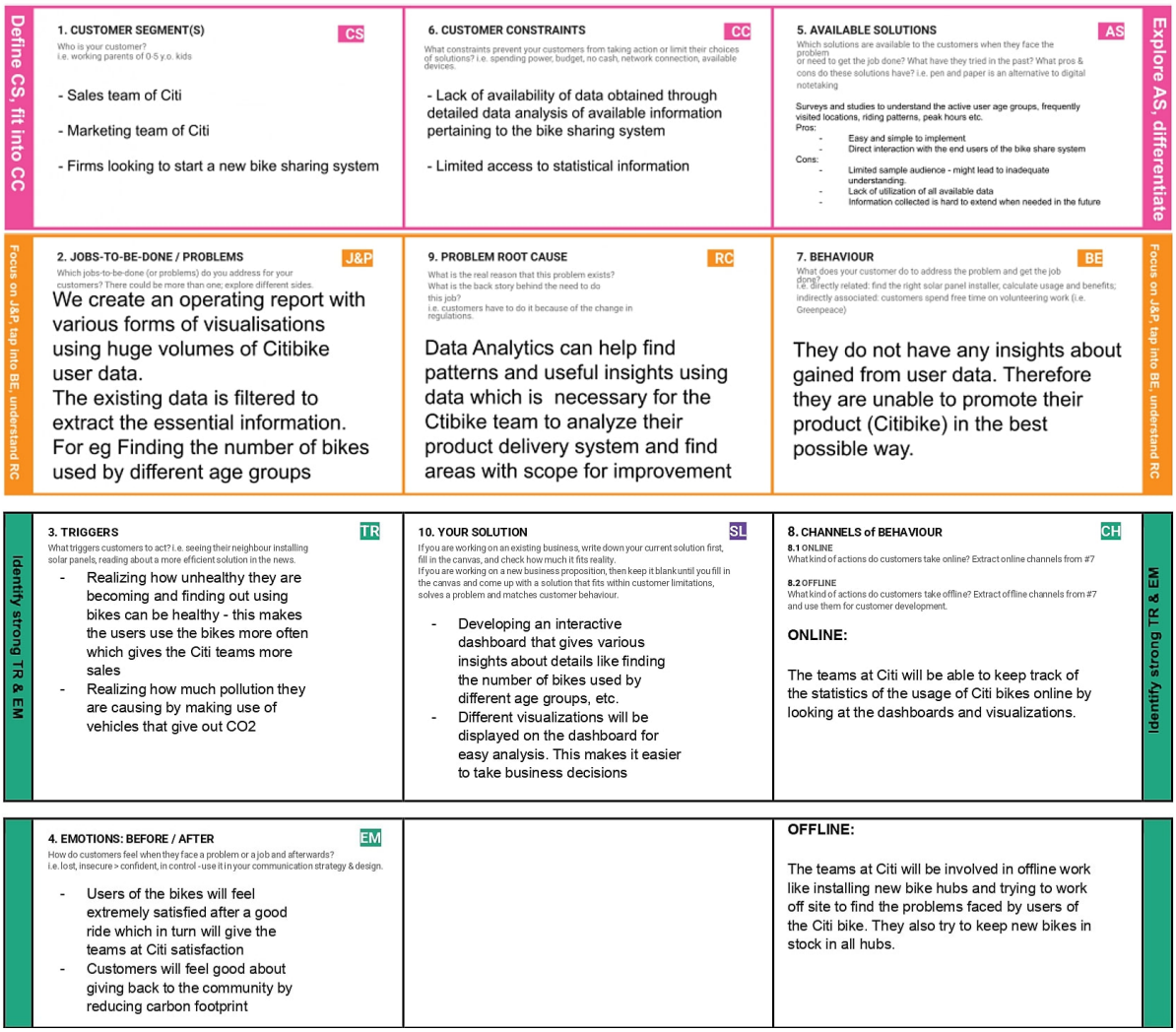
Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	The goal of this analysis is to create an operating report of Citi Bike for the year 2018. To create data visualizations to understand 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.	Idea / Solution description	To effectively process and visualize data, we use IBM Cognos because there are number of data and it is difficult to process it with some conventional methods, a web-based integrated business intelligence suite by IBM. Using IBM Cognos, we aim to create an operating report, provide useful insights and present them in the form of a dashboard.
3.	Novelty / Uniqueness	Our solution reduces maintenance due to complete report coverage, The conversion rate of customers (who could be non-tourists) to subscribers and also provides fast results.
4.	Social Impact / Customer Satisfaction	Bike share engages riders in physical activity, beneficial to health. In addition, it promotes green mobility and contributes to carbon neutrality. This analysis will help in understanding the association between bike share usage and the environment which is essential for system management and urban transportation planning.
5.	Business Model (Revenue Model)	Revenue is generated on renting the bike with the analyzed data we can alter the pricing according to the demand of the customer.
6.	Scalability of the Solution	The creation of the operating report (solution) involves an extended analysis of data presented for the year 2018. Our solution offers high scalability, as not only can it be

3.3Problem Solution Fit

Project Title: A New Hint To Transportation-Analysis Of The NYC Bike Share System **Project Design Phase-I - Solution Fit** **Team ID:** ***PNT2022TMID19163***



4.REQUIREMENT ANALYSIS

4.1 Functional Requirement

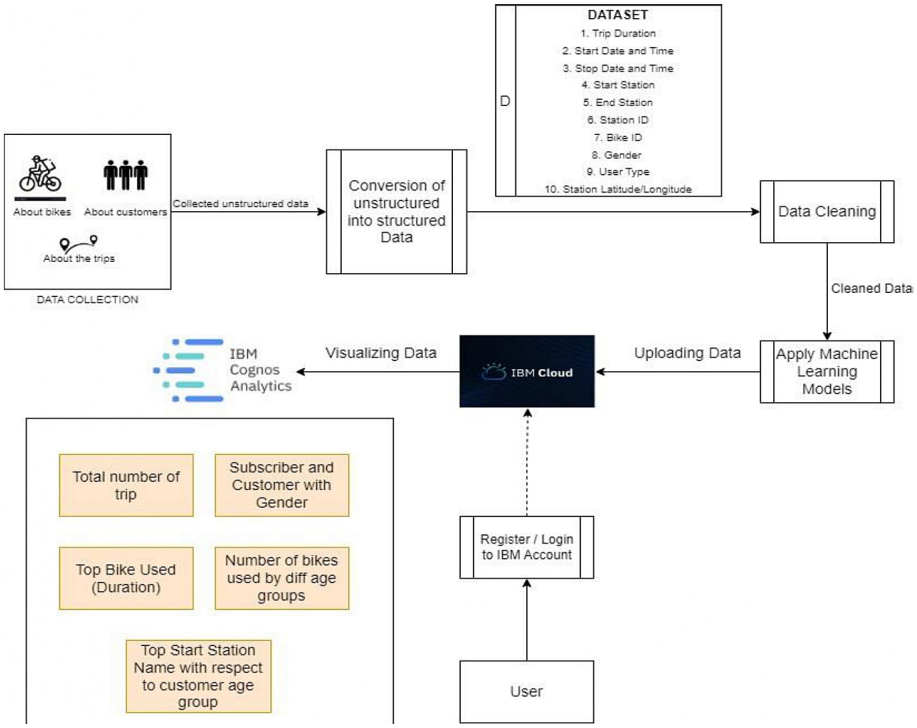
FR No. Functional Requirement (Epic) Sub Requirement (Story / Sub-Task)FR-1 User Registration Registration through FormRegistrationthrough Gmail RegistrationthroughLinkedIn FR-2 User Confirmation Confirmation via Email Confirmationvia OTP FR-3 Collection of Data Usage of the NYCCiti Bike helpsgenerate data regarding thedifferenttrips taken by different peopleusingCiti Bike. These data were

thencategorized and providedas datasets,onwhich further analysis andvisualizationare to be carried out FR-4 Analysis of Data Analysis of the given dataincludescarrying out preprocessing&filteringthe data as per the requirement posedby the sub-task. The usage of MachineLearning techniques to gainfurtherinsights into the data alsocontributestothe analysis, and as a result visualization of data. FR-5 Display (Visualization) of Data Different visualizations are carriedoutdepending on the sub-taskdealt with.These visualizations are thenpooledand displayed on a dashboard-whichserves as a tool to provide businessinsights to customers. Someof thedifferent sub-tasks involvedinthisrequirement include findingthetop10Start station names withrespect tocustomer age group, displayingthetopbikes used with respect totripduration etc.

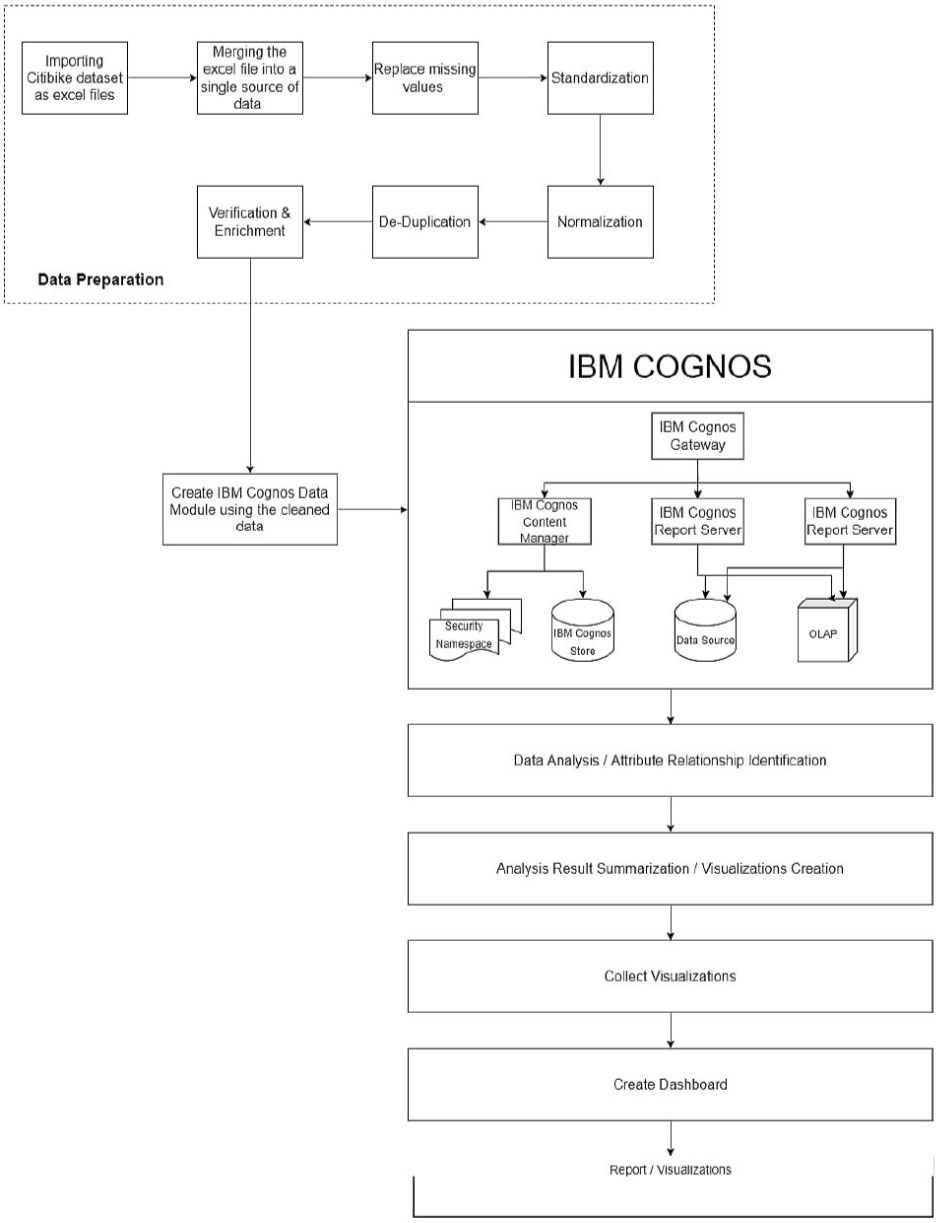
5.PROJECT DESIGN

5.1 Data Flow Diagram

A Data Flow Diagram (DFD) is a traditional visual representationoftheinformation flows within a system. A neat and clear DFDcandepict theright amount of the system requirement graphically. It shows howdataenters and leaves the system, what changes the information, and where data is stored.



5.2 Solution Architecture Diagram:



5.3 User Stories:

User Stories

Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer,Analysts, Government	Registration	USN-1	As a user, I should be able to register to seethe dashboard as a new user	Successful Registration	High	Sprint-1
Customer,Analysts, Government	Login	USN-2	As a user I should be able to login to see the dashboard with the correct credentials	Succesful Login with correct credentials	High	Sprint-1
Customer,Analysts, Government	Accessing the dashboard	USN-3	As a user, I should be able to view the visualizations displayed	Should be able to view the following analysis among others : 1. Total number of trips 2. Subscriber and Customer with gender 3. Top Bike used with respect to duration 4. Number of bikes used by different age groups 5. Top start station name with respect to customer age group	High	Sprint-1
Customer,Analysts,Government	Manipulating the data	USN-4	As a user I should be able to apply some modifications to the data to see how theresultant visualizations change	I should have the permission to manipulate the data	High	Sprint-2
Customer,Analysts,Government	Collection of data	USN-5	Lyft citi bike website provides data for analyse,visualization etc.	Accessing the data in Lyft citi bike	High	Sprint -1
Customer,Analysts,Government	Analysing the data	USN-6	Analysed data is used to create different types visualization and charts.	Can view the analysis	High	Sprint-1

6.PROJECT PLANNING &SCHEDULING

6.1 Sprint Planning & Estimation

Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	5	High	Aniruthram ,Deepan Kumar,Aneesh,Fayaz
Sprint-1	Registration	USN-2	As a user, I will receive confirmation email once I have registered for the application	2	High	Aneesh,Fayaz
Sprint-1	Registration	USN-3	As a user, I can register for the application through Gmail or phone number	1	Medium	Aniruthram ,Deepan Kumar
Sprint-2	Login	USN-4	As a user, I can log into the application by entering email & password	7	High	Deepan Kumar,Aneesh,Fayaz
Sprint-2	Collection of user data	USN-5	As a user, I can enter my information which includes bank details for easy online payment and my mobile number for OTP confirmation	10	High	Aniruthram ,Deepan Kumar,Aneesh
Sprint -3	Analysis of User data	USN -6	As a analyst, I can access the Lyft Citi bike database to access the bike sharing system data of the user	8	High	Aniruthram ,Deepan Kumar,Aneesh,Fayaz
Sprint-3	Analysis of User data	USN-7	The data can be used as input for creating various types of visualizations and analysis.	5	High	Deepan Kumar,Aneesh,Fayaz
Sprint -4	Visualization	USN-8	As an analyst, I create various visualizations using IBM Cognos based on the knowledge obtained at theend of the EDA process.	7	High	Aniruthram ,Deepan Kumar

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-4	Visualization	USN -8	As an analyst, I create a dashboard with the created visualizations to supplement business insights during the decision-making process at Citi. As an analyst, I apply predictive analytics and additional features to enhance visualizations	8	High	Aniruthram ,Deepan Kumar,Aneesh,Fayaz
Sprint-4	Dashboard	USN - 9	I can create and access the dashboard based on the user data	5	High	Aniruthram ,Deepan Kumar,Aneesh,Fayaz

Project Tracker, Velocity & Burndown Chart: (4 Marks)

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	8	6 Days	24 Oct 2022	29 Oct 2022	8	29 Oct 2022
Sprint-2	17	6 Days	31 Oct 2022	05 Nov 2022	25	
Sprint-3	13	6 Days	07 Nov 2022	12 Nov 2022	38	
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	48	

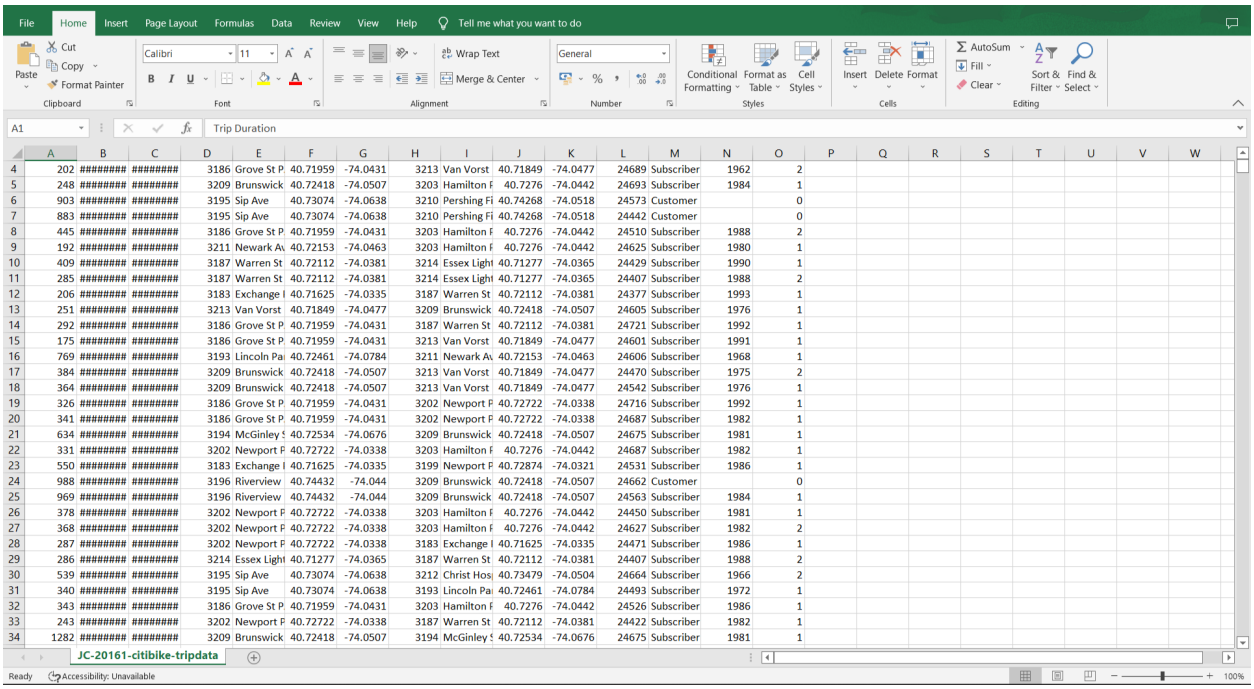
6.2 Sprint Delivery Schedule

MILESTONE AND ACTIVITY LIST:

TITLE	DESCRIPTION	DATE
Literature Survey & Information Gathering	Literature survey on the selected project & gatheringinformation by referring thetechnical papers, research publications, journals etc.	19 SEPTEMBER 2022
Prepare Empathy Map	Prepare Empathy Map Canvasto capture the user Pains & Gains, prepare list of problemStatements that are to be solved by this project.	19 SEPTEMBER 2022
Ideation	List the ideas by organizinga brin storming session and prioritize the top 3 ideas based on the feasibility & importance.	19 SEPTEMBER 2022
Proposed Solution	Prepare the proposed solution document, which includes novelty, feasibility ofidea, revenue model, social impact, scalability of solution, etc.	19 SEPTEMBER 2022
Problem Solution Fit	Prepare problem - solution fit document.	19 SEPTEMBER 2022
Solution Architecture	Prepare solution architecture document.	19 SEPTEMBER 2022
Customer Journey	Prepare the customer journey maps to understand the user interactions & experienceswith the application (entry to exit).	3 OCTOBER 2022
Functional Requirement	Prepare the functional requirement document.	3 OCTOBER 2022
Data Flow Diagrams	Draw the data flow diagrams and submitfor review.	3 OCTOBER 2022
Technology Architecture	Prepare the technology architecture diagram.	3 OCTOBER 20222
Prepare Milestone & ActivityList	Prepare the milestones& activity list of the project.	18 OCTOBER 2022
Project Development - Deliveryof Sprint-1, 2, 3 & 4	Develop & submit the developed code by testing it.	IN PROGRESS

7.WORKING WITH THE DATASET&DATAVISUALISATION

7.1 Understanding the dataset



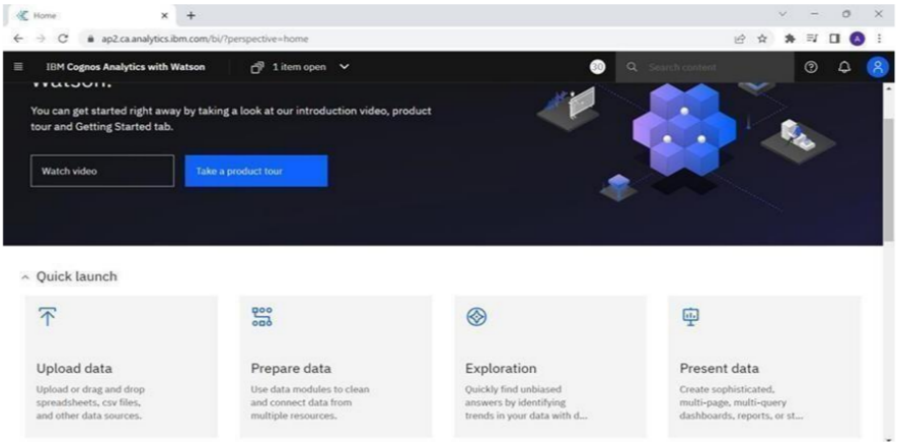
Dataset Link: Dataset

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 areusuallyNYC residents
14. Year of Birth: Self-entered, not validated by an IDGender (Zero=unknown; 1=male; 2=female): Usually unknown for customerssincethey often sign up at a kiosk

7.2Loading the dataset

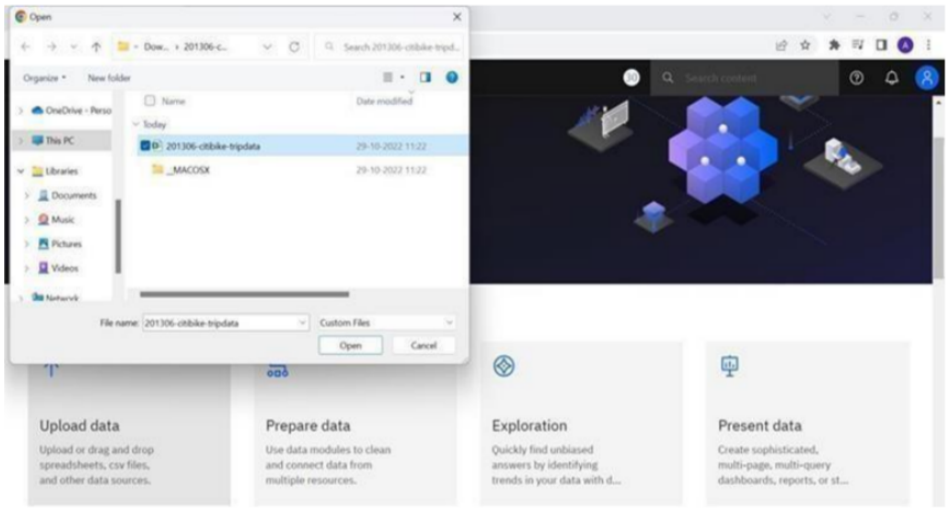
Open Cognos Analytics and click upload data

Open Cognos Analytics and click upload data

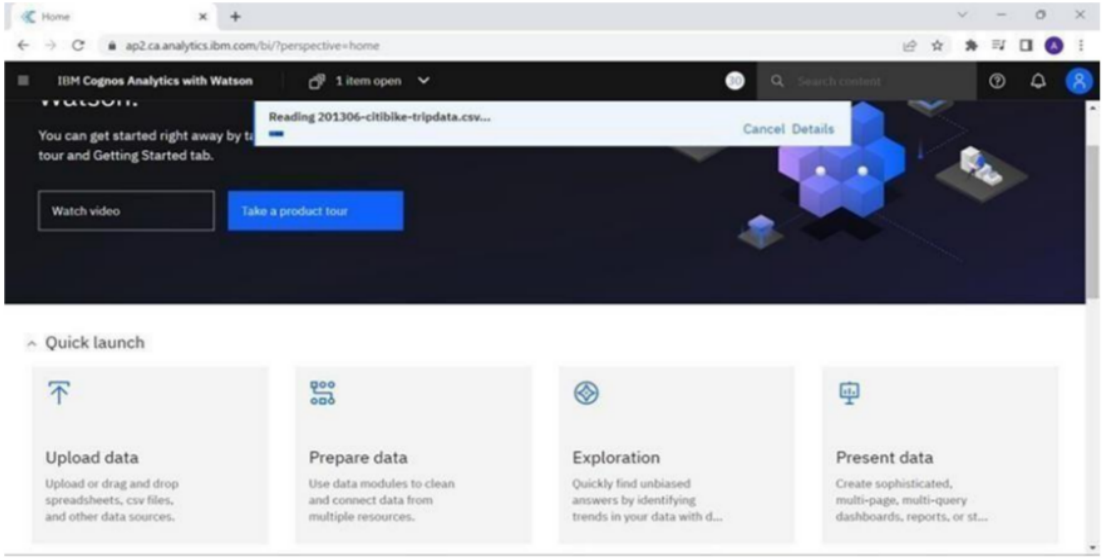


Select the dataset to be uploaded

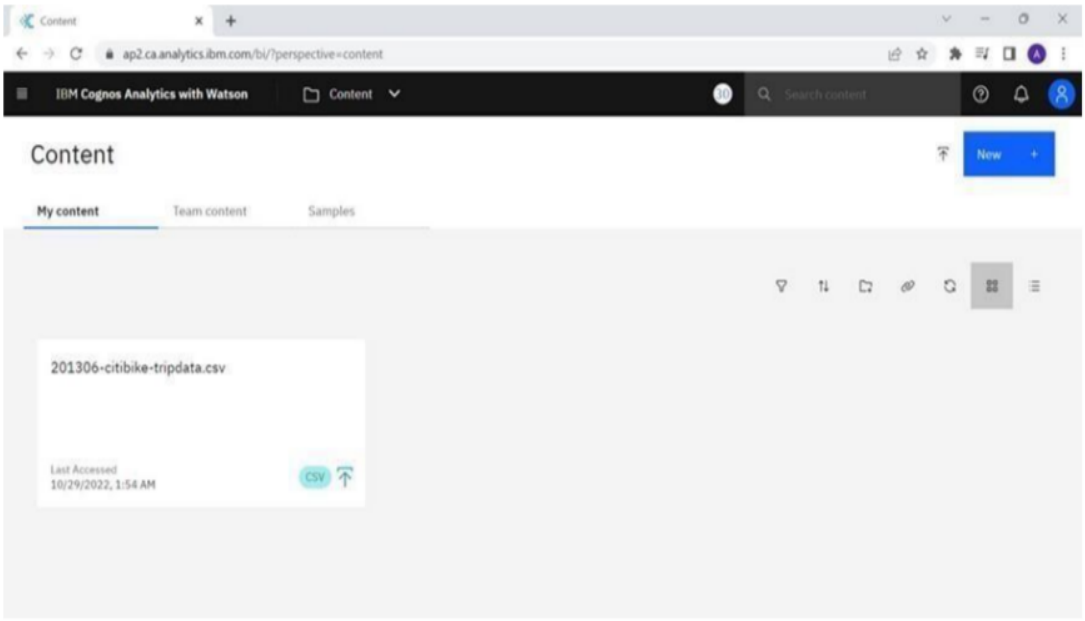
Select the dataset to be uploaded



The excel file is getting uploaded in Cognos Analytics

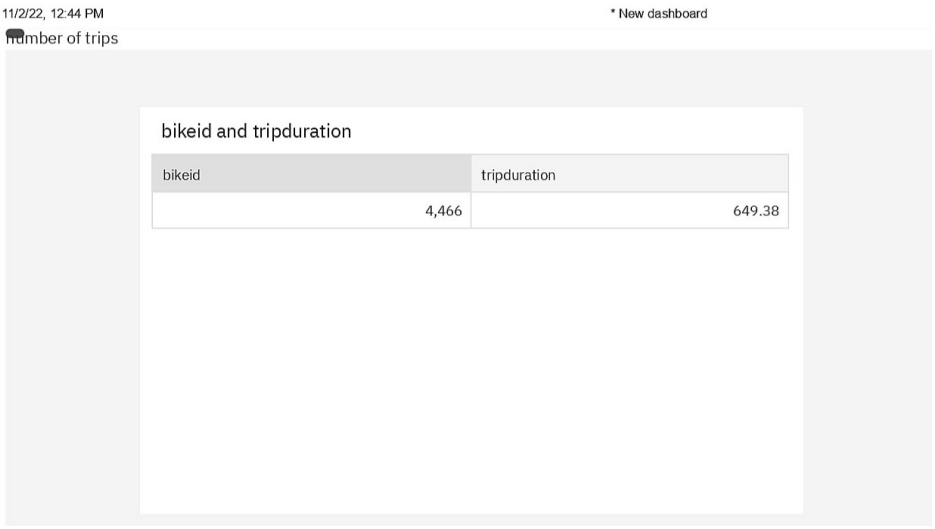


The dataset can be accessed in My Content in Cognos Analytics

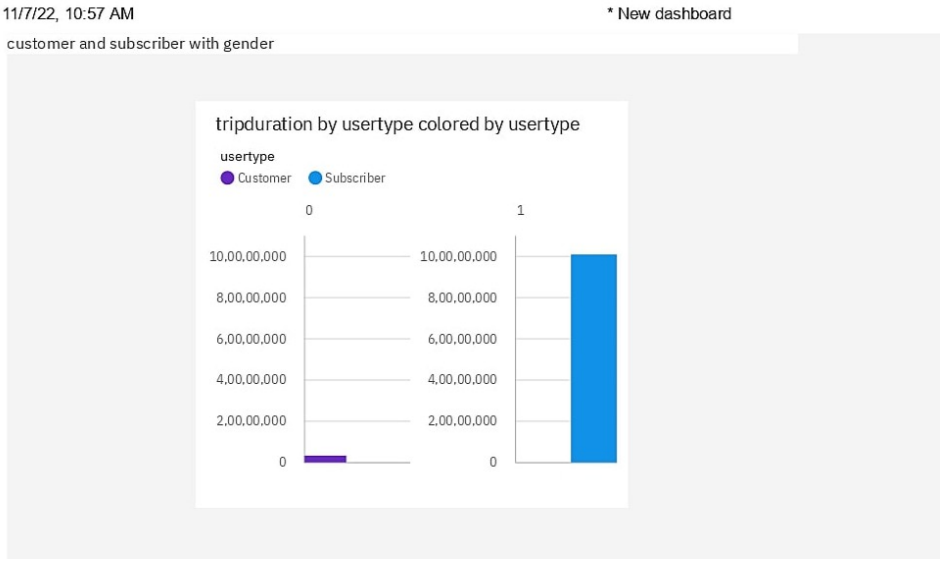


7.3 Data Visualization charts

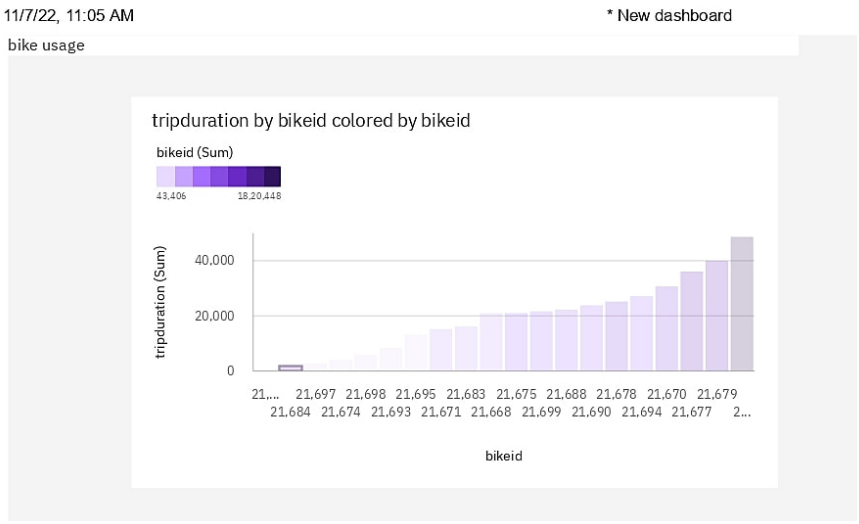
Number of Trips:



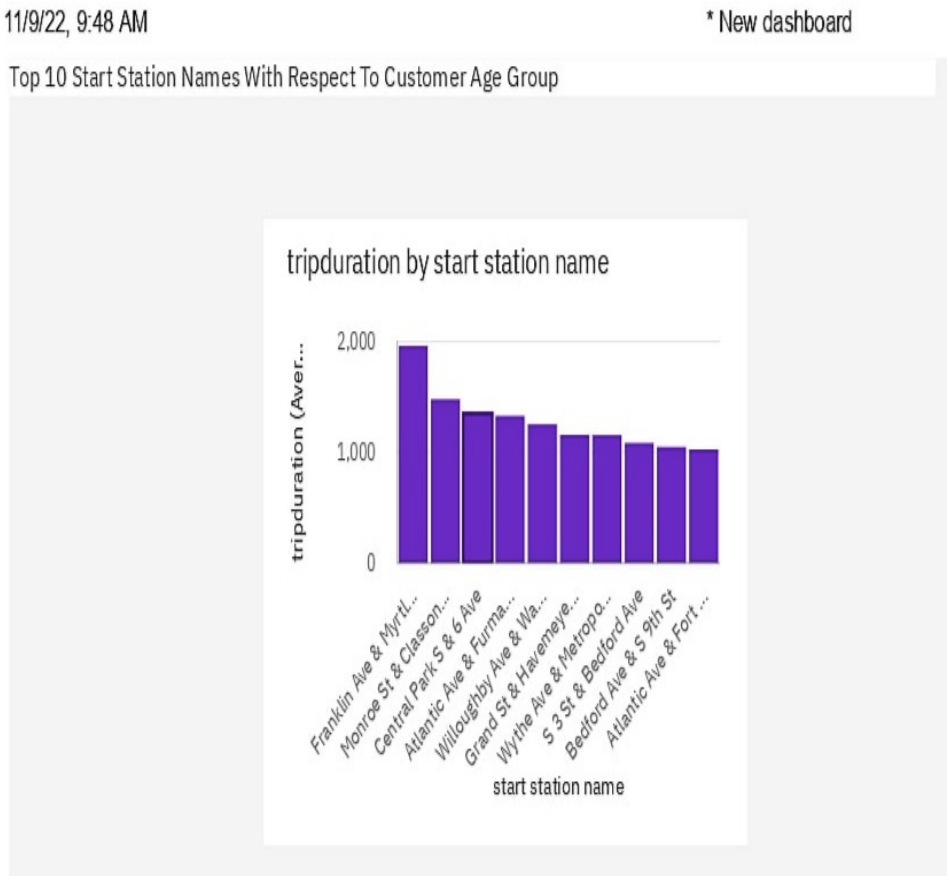
Customer and Subscriber with Gender:



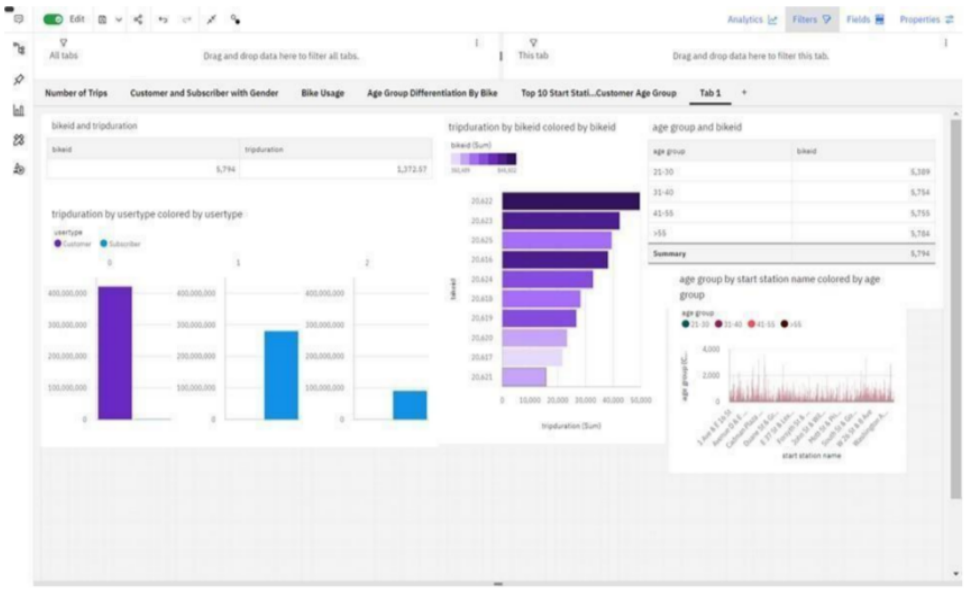
Bike Usage:



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



8.CREATING THE DASHBOARD



9.ADVANTAGES AND DISADVANTAGES

The benefits of bike sharing schemes include transport flexibility, reductions to vehicle emissions, health benefits, reducedcongestionandfuel consumption, and financial savings for individuals. One can easily analyze and understand trends in bike sharingpatternswith the created dashboard. With no prior skills and knowledgeaboutthe tools that we use for analysis, anyone (literate or illiterate) caneasily infer the knowledge that we represent in various chartsorgraphs or maps. So that it would be helpful to users andcompaniestomake appropriate decisions in the future.

10.CONCLUSION

Based on the quantitative as well as visual analysis of the NewYorkbike 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 andminimumusagein winter. This was initially hypothesized because of the harshnessofNew York's harsh winters and the treacherous riding conditionsthatexist during that time. However, despite the adverse weatherconditions, there is a strong core demographic that consistentlyusesthe system. This conclusion is based on that fact that evenduringthemonths of January and February which are the peakwintermonths,there are more than two hundred thousand trips in the system. New

York has a strong public transit system, and the bike sharesystemseems to complement it quite well with a majority of the highestusedstations located either close to subway lines or the commuterrail stations in the city.

Based on the locations of the stations and the durationof trips, itcanbe hypothesized that bike shares are replacing last mile tripsthatwould otherwise be done either on foot or on public transit. Thisisparticularly true in case of New York where a combinationof densepublic transit network, the road congestion during peakhoursandtheaverage trip distance as calculated create a situation wheretheonlypotential trips that the bike share systemis replacing currentlyarethose that would otherwise have been undertaken either onfootoronpublic bus.

11.FUTURE SCOPE

NYC is a very crowded and happening place whichleadstolotsof pollution. And in this busy world people are alwaysworriedabout transportation this bike sharing systemreducesthatstress. With increase in population pollutionalsoincreases.Soit is in our hands to reduce pollution and to makeabetterfuture for our younger generations. We can analyzewhichstation needs more bikes and any area needs newstationtobeinstalled. The survey outcomes indicates the needs forimproved techniques in bike sharing analytics. Thereexistsalot of scope in this research area.

12.SOURCE CODE

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
```

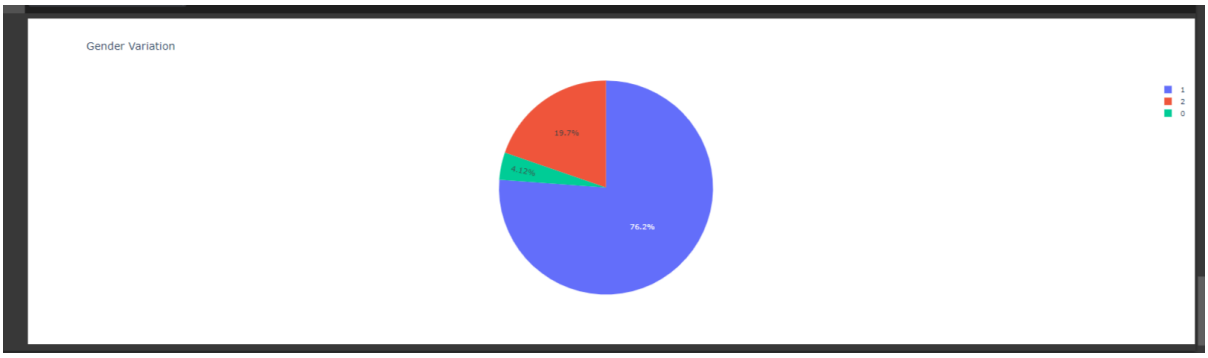
```
from oauth2client.client import GoogleCredentials
from google.colab import files
files.upload()
import pandas as pd
df = pd.read_csv('JC-202102-citibike-tripdata.csv')
df.head()
df.describe()
df.info()
df.isnull().sum()
df[df['started_at'].isnull()]
df[df['ended_at'].isnull()]
df = df[:-1]
print(type(df["start_lat"][0]))
print(df["start_lat"][0])
df['start_lat'].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.info()
df.shape
df.to_csv('New_dataset.csv', index=False)
```

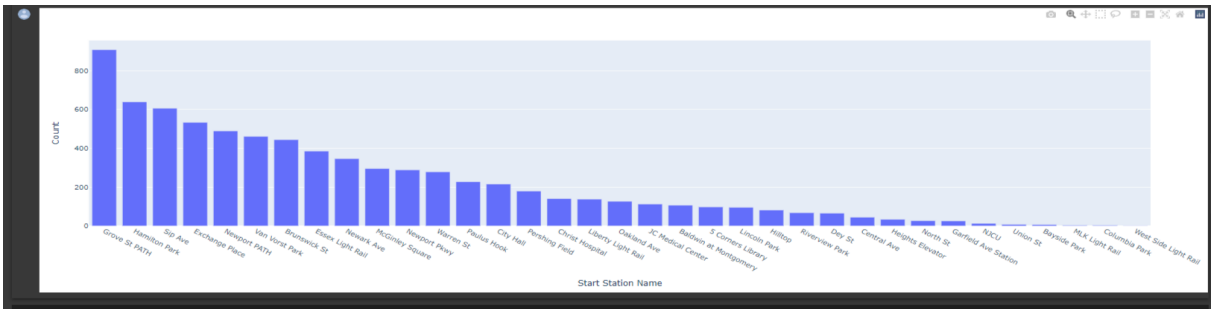
Sprint 4:

```
from google.colab import files
files.upload()
```

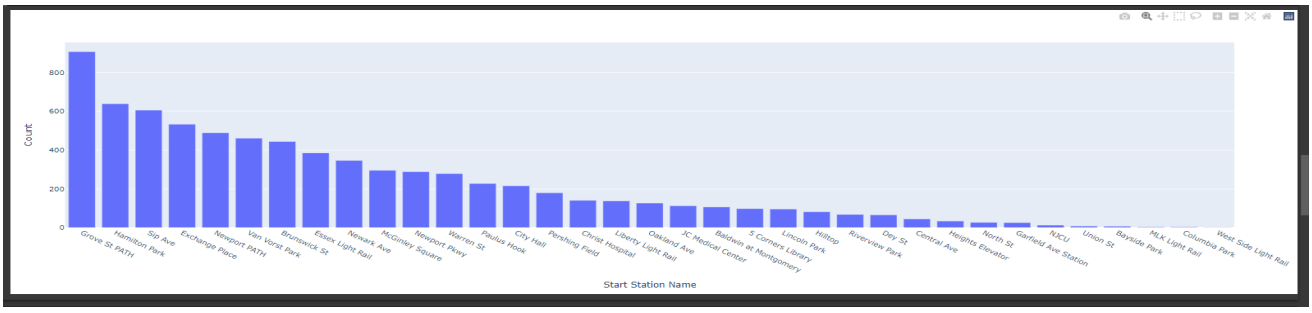
```
import pandas as pd
df = pd.read_csv('JC-20161-citibike-tripdata.csv')
df.head()
temp = df
temp.head()
temp.describe()
temp.info()
temp["Start Time"] = pd.to_datetime(temp["Start Time"])
temp["Stop Time"] = pd.to_datetime(temp["Stop Time"])
temp.info()
temp["Hour"] = temp["Stop Time"].dt.hour - temp["Start Time"].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")
```



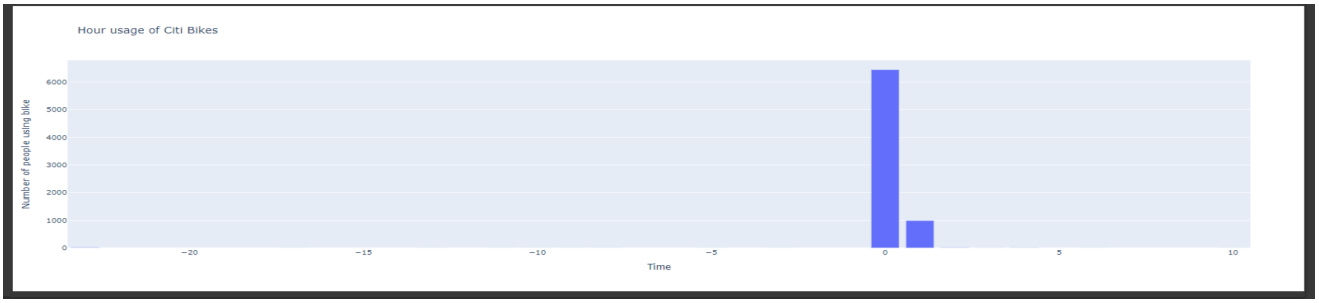

```
px.bar(x=temp["Start Station Name"].value_counts().index,
y=temp["Start Station Name"].value_counts().values,
labels={'x':'Start Station Name','y":"Count"})
```




```
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-39037-1660389703