**Project Report**

1. **INTRODUCTION** 
   1. Project Overview
   2. Purpose
2. **LITERATURE SURVEY**
   1. Existing Solution
   2. References
   3. Problem Statement Definition
3. **IDEATION & PROPOSED SOLUTION**
   1. Empathy Map Canvas
   2. Ideation & Brainstorming
   3. Proposed Solution
   4. Problem Solution fit
4. **REQUIREMENT ANALYSIS**
   1. Functional requirement
   2. Non-Functional requirements
5. **PROJECT DESIGN** 
   1. Data Flow Diagrams
   2. Solution & Technical Architecture
   3. User Stories
6. **PROJECT PLANNING & SCHEDULING** 
   1. Sprint Planning & Estimation
   2. Sprint Delivery Schedule
7. **WORKING WITH THE DATASET & DATA VISUALIZATION**

7.1 Understanding the Dataset

7.2 Loading the Dataset

7.3 Visualization Chart

8. **CREATING THE DASHBOARD**

9. **ADVANTAGES & DISADVANTAGES**

10. **CONCLUSION**

11. **FUTURE SCOPE**

12. **SOURCE CODE**

13. **GITHUB LINK**

1. **INTRODUCTION**
   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. **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

1. **LITERATURE SURVEY**
   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.

* 1. **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

* 1. **Problem statement Definition**

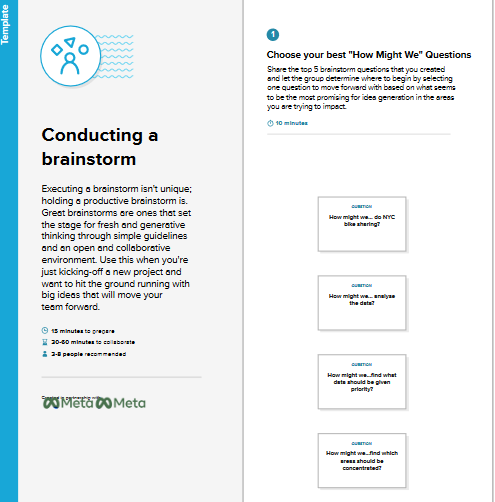
In busy cities like New York the people are facing difficulties in analyzing the demand for bikes during peak hours.

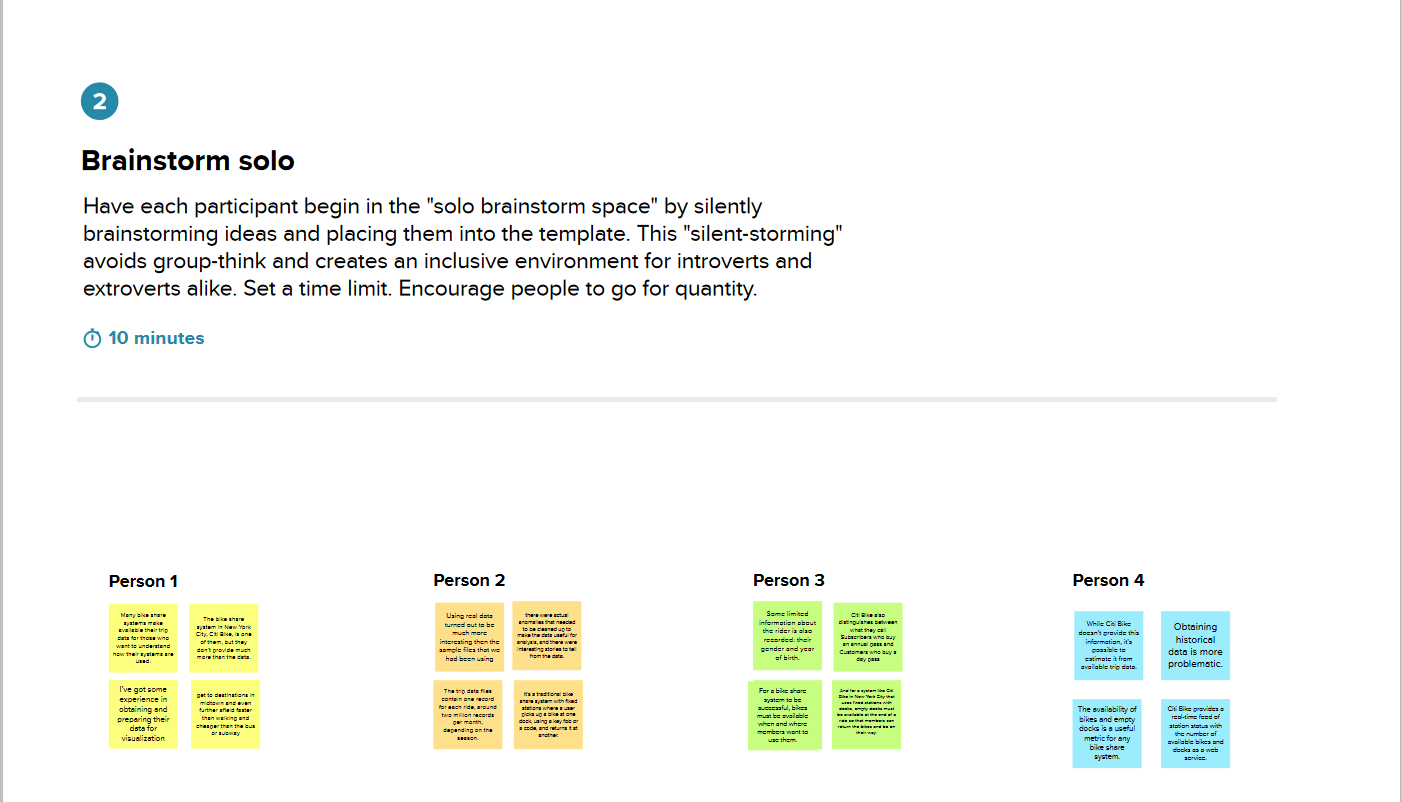
The main objective of this project is to predict bike patterns that will be extremely helpful for people to plan their travel.

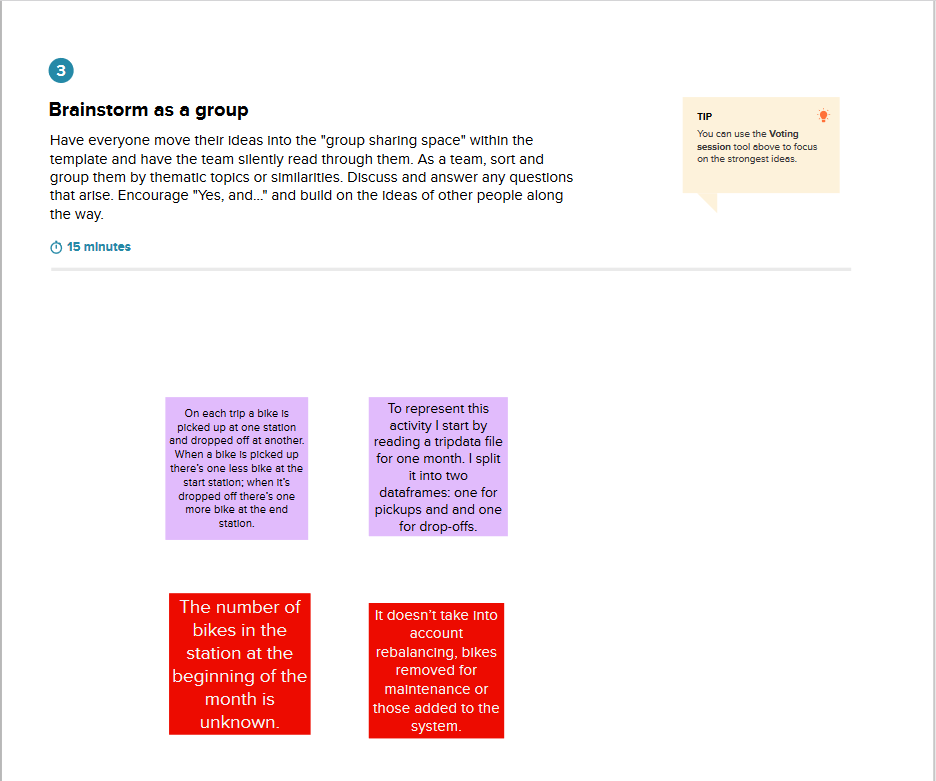
1. **IDEATION & PROPOSED SOLUTION**
   1. **Empathy Map Canvas**

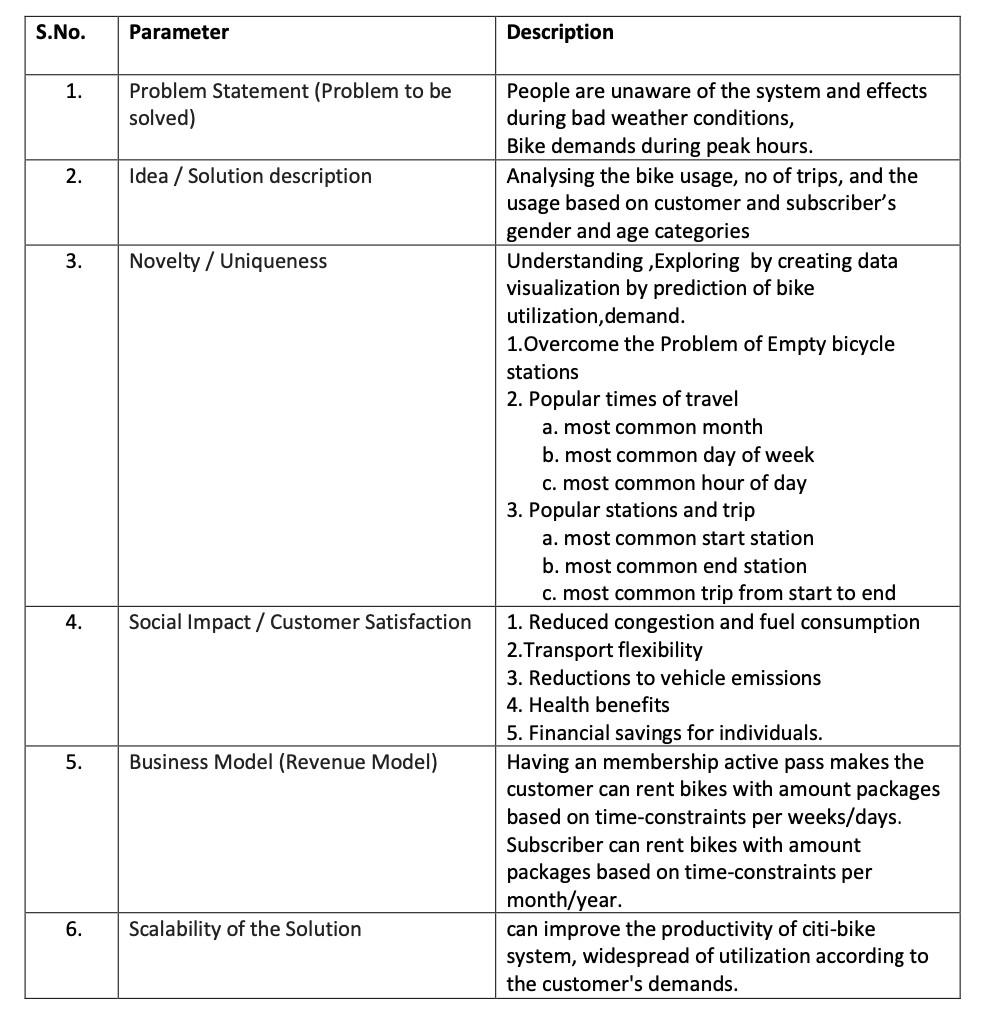


* 1. **Ideation and Brainstorming**

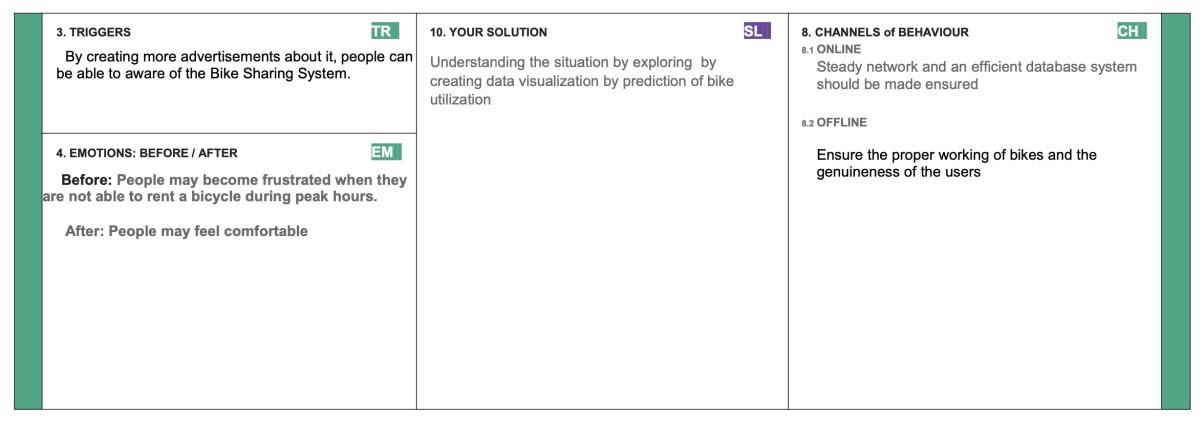




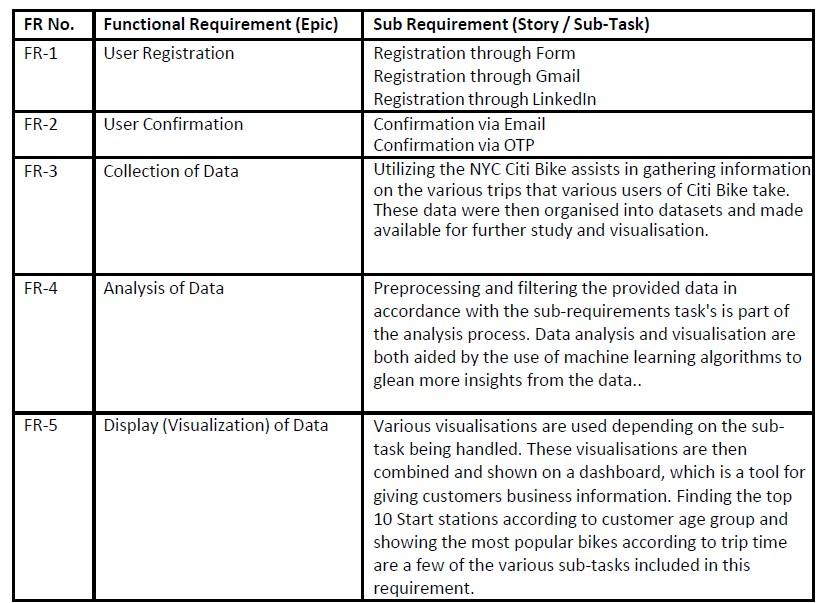


* 1. **Proposed Solution**
  2. **Problem Statement Fit**

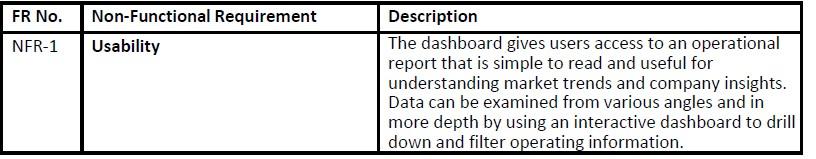


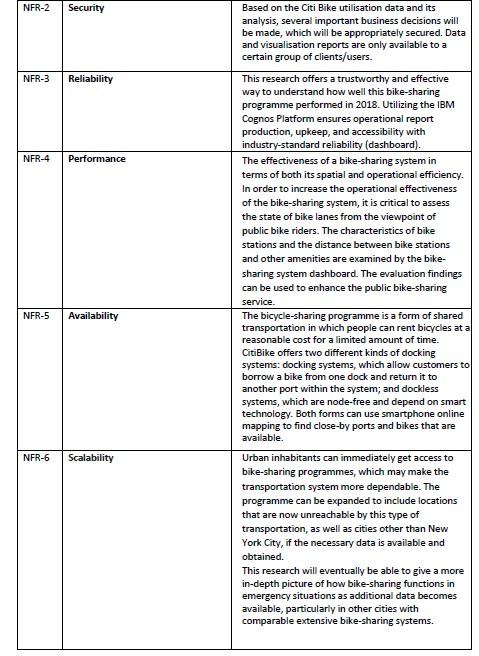


1. **REQUIREMENT ANALYSIS**
   1. **Functional Requirement**

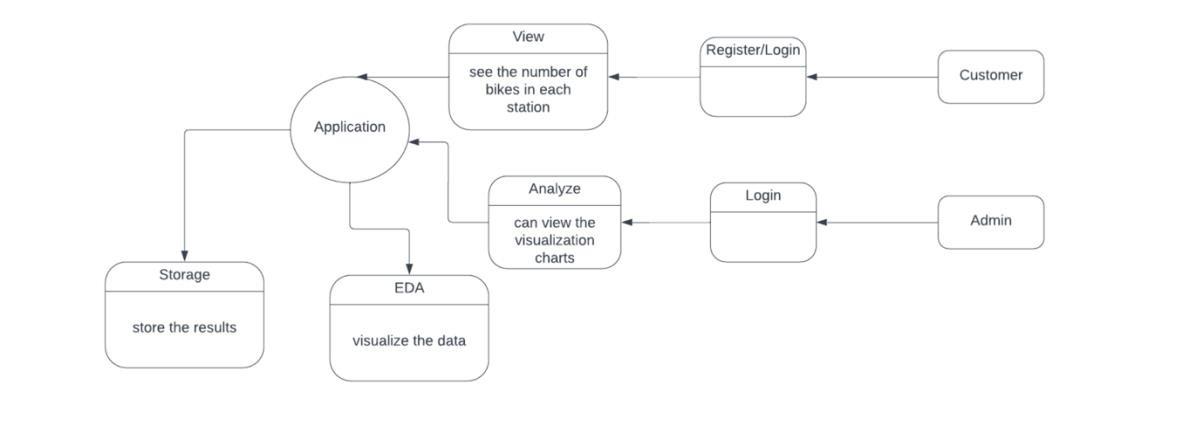


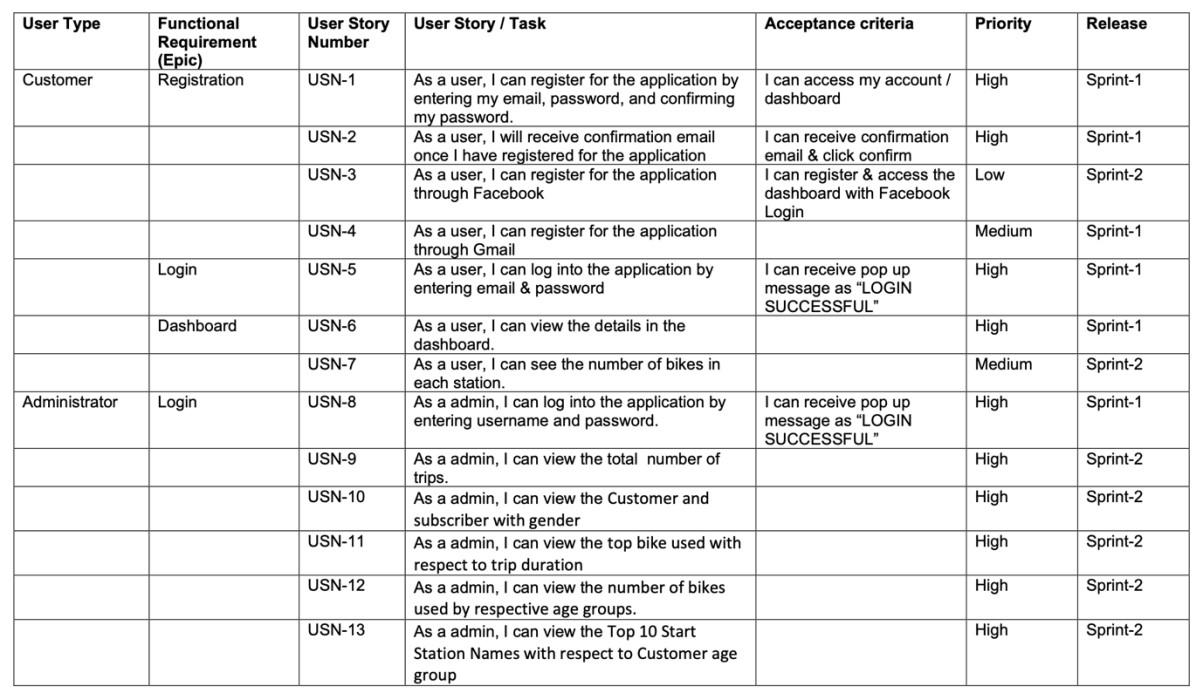
* 1. **Non-Functional Requirement**



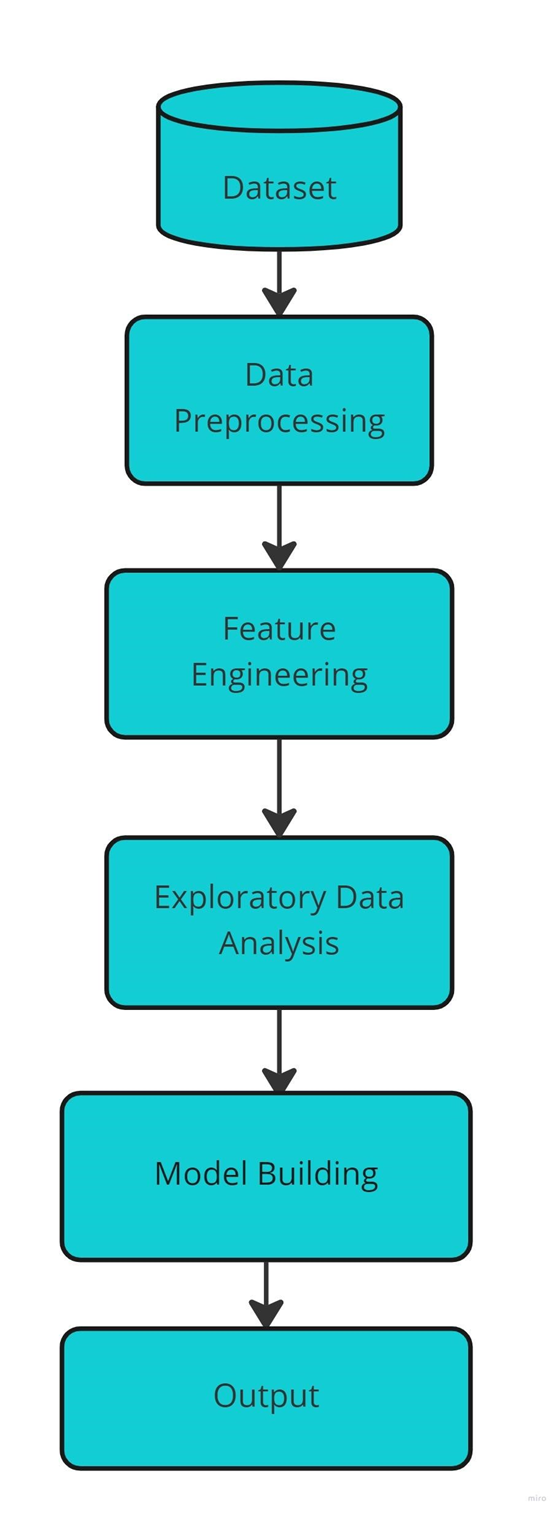


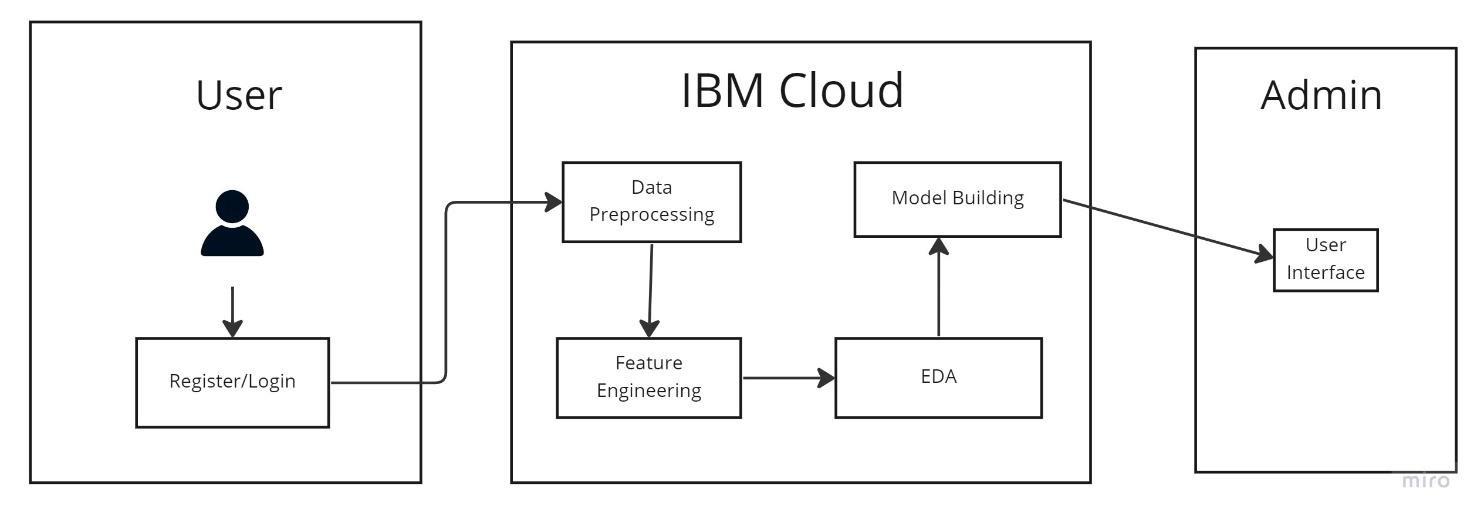
1. **PROJECT DESIGN**
   1. **Data Flow Diagram**

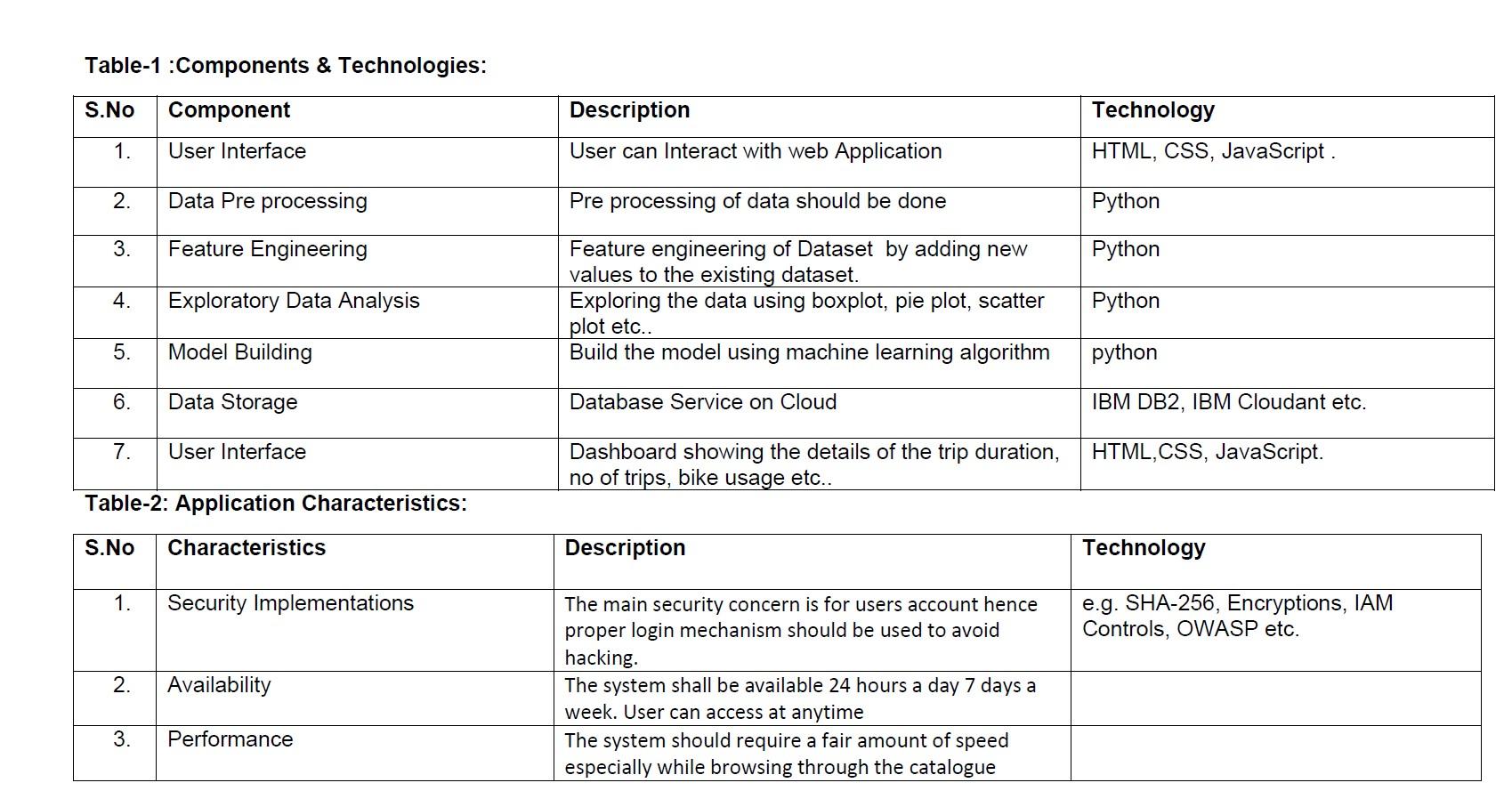




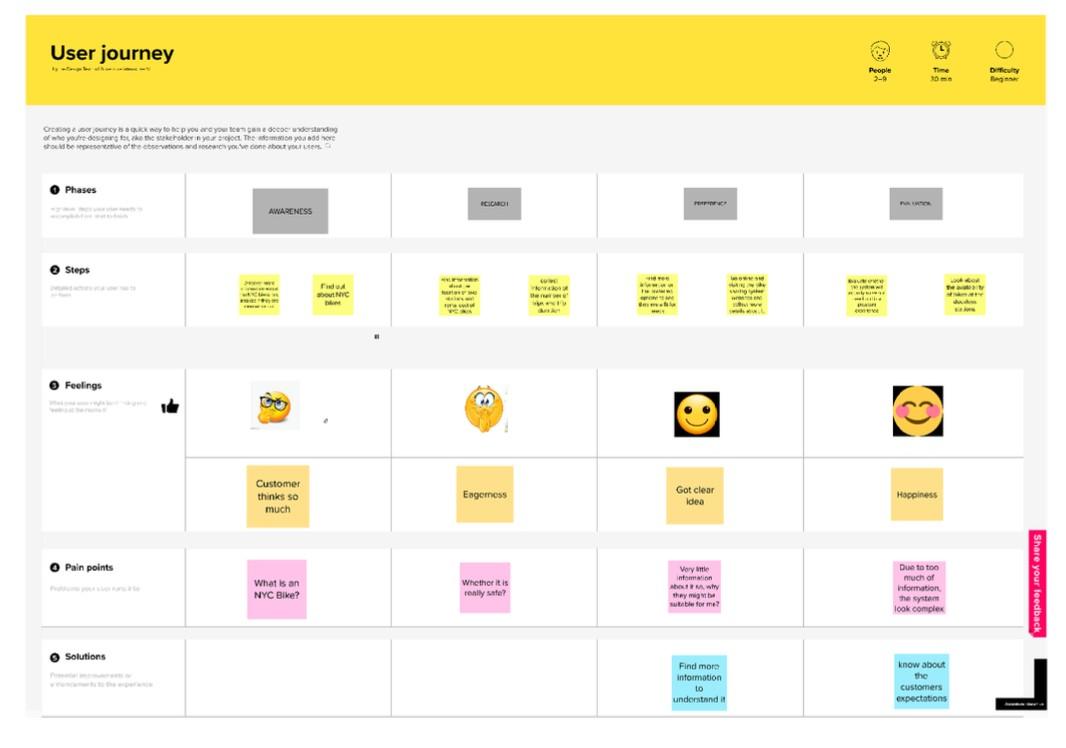
* 1. **Solution & Technical Architecture**

****

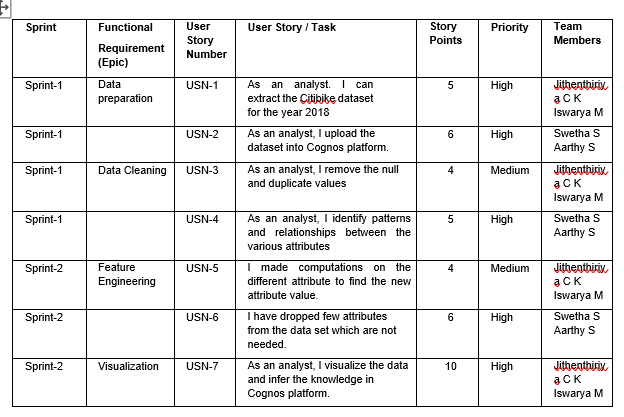


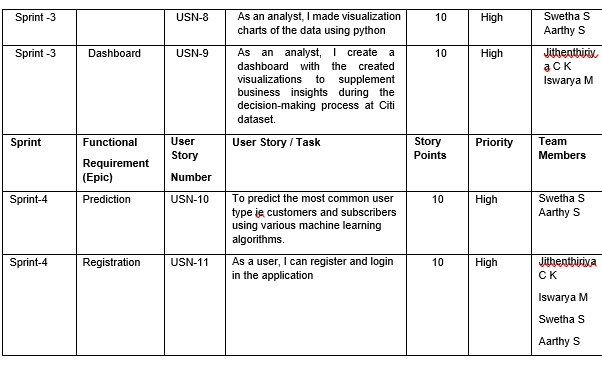


* 1. **User Stories**



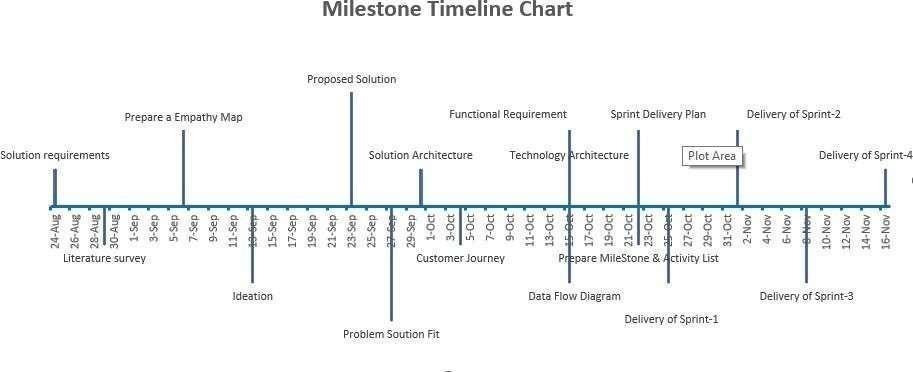
1. **PROJECT PLANNING & SCHEDULING**
   1. **Sprint Planning & Estimation**

****

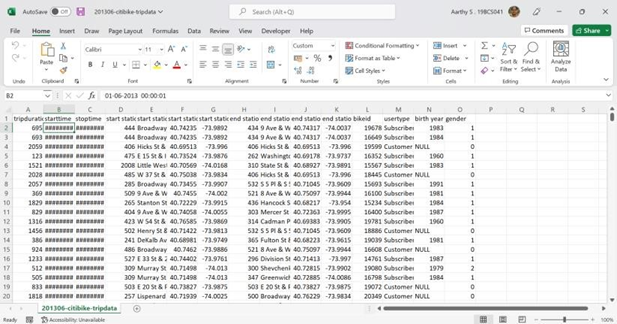
****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

* 1. **Sprint Delivery Schedule**



1. **WORKING WITH THE DATASET & DATA VISUALISATION** 
   1. **Understanding the dataset**

****

**Dataset Link:** [**Dataset**](https://github.com/IBM-EPBL/IBM-Project-10706-1659197931/blob/main/Project%20Design%20%26%20Planning/Project%20Development/Team%20Member%202%20-%20Iswarya%20M/Sprint%203/dataset.csv)

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

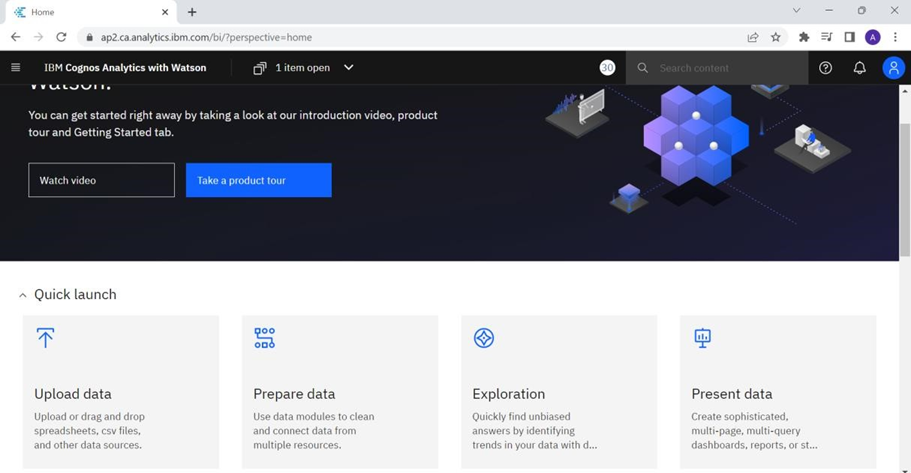
13. User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member): Customers are usually tourists, subscribers are usually NYC residents

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

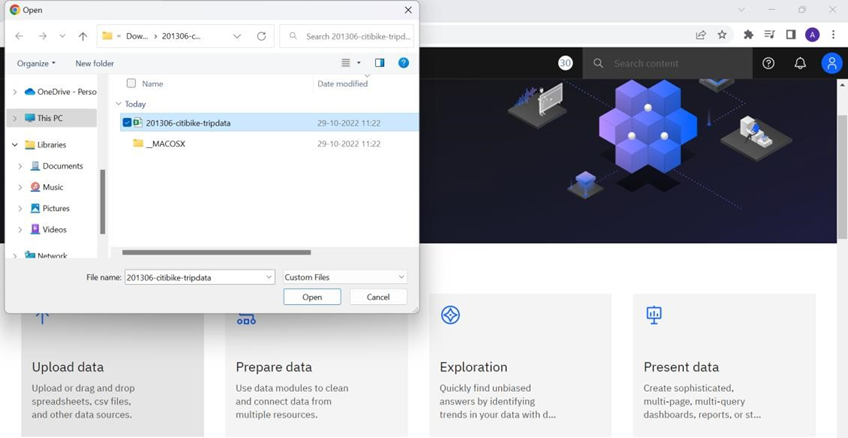
Gender (Zero=unknown; 1=male; 2=female): Usually unknown for customers since they often sign up at a kiosk

**7.2 Loading the dataset**

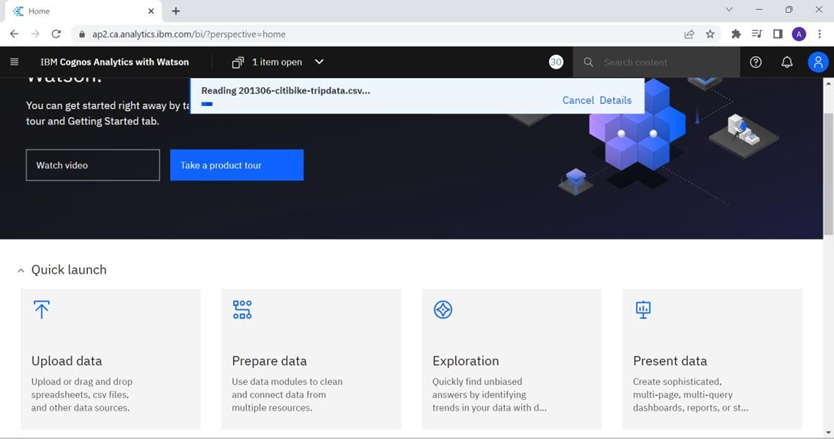
Open Cognos Analytics and click upload data



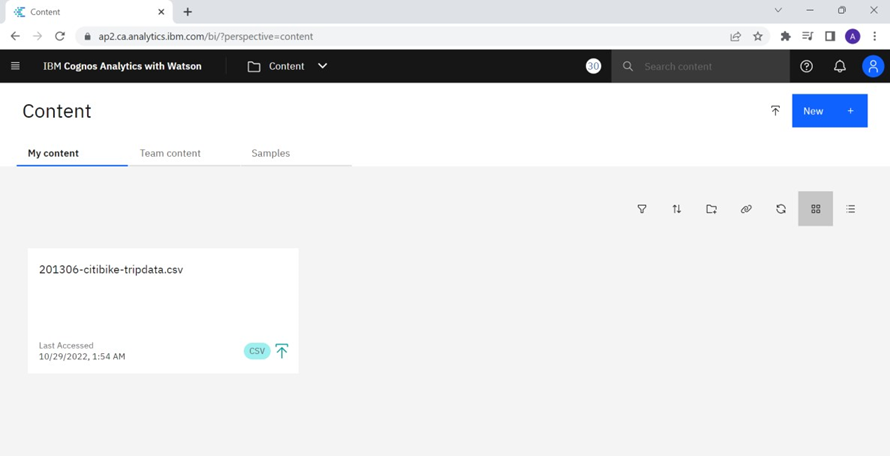
# Select the dataset to be uploaded



The excel file is getting uploaded in Cognos Analytics

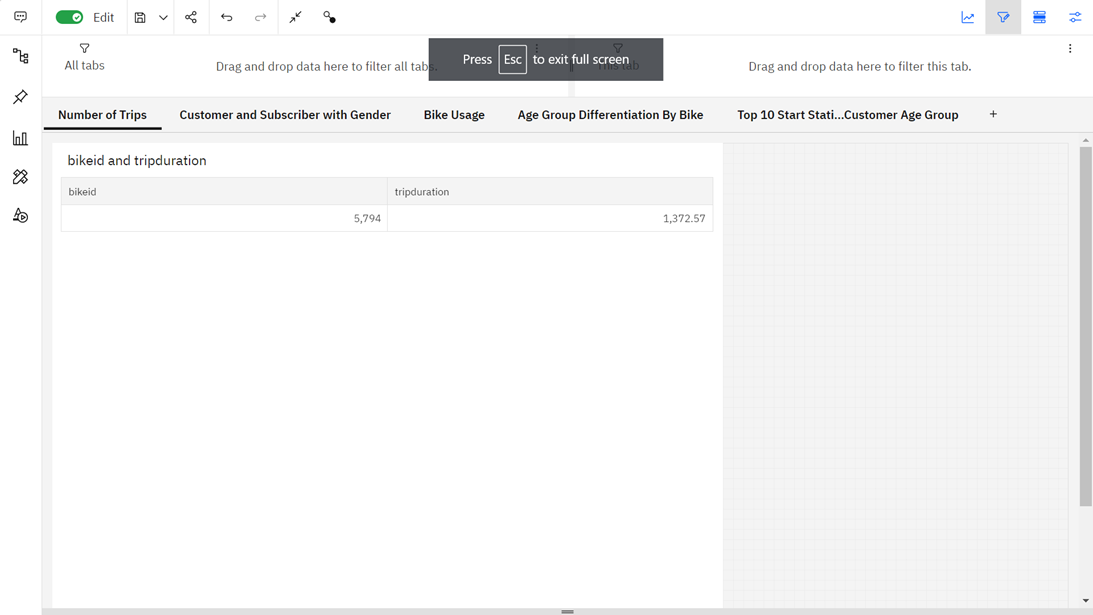
****

The dataset can be accessed in My Content in Cognos Analytics

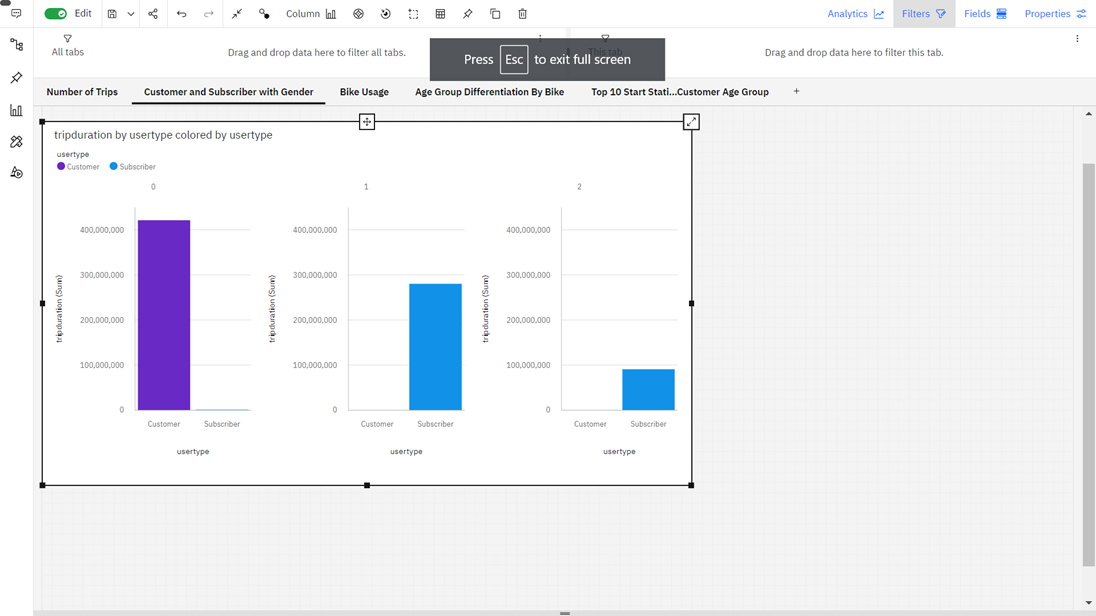


**7.3 Visualization charts**

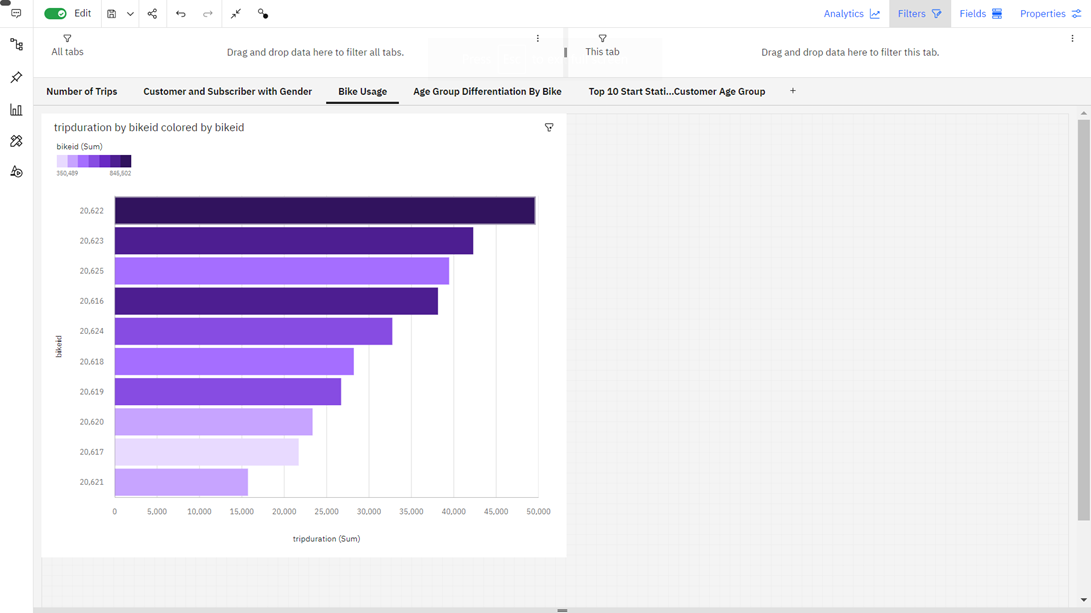
Number of Trips:



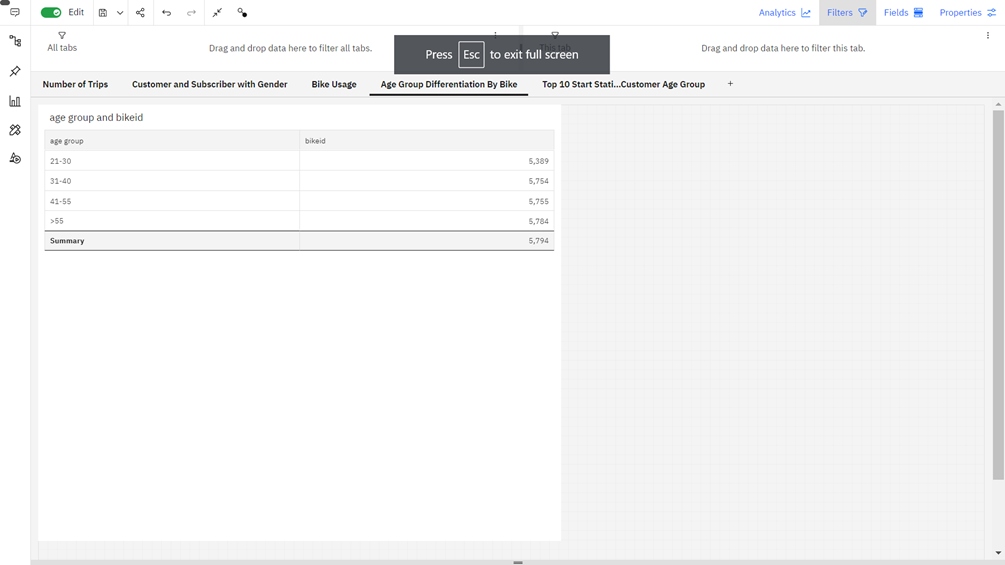
Customer and Subscriber with Gender:

****

Bike Usage:

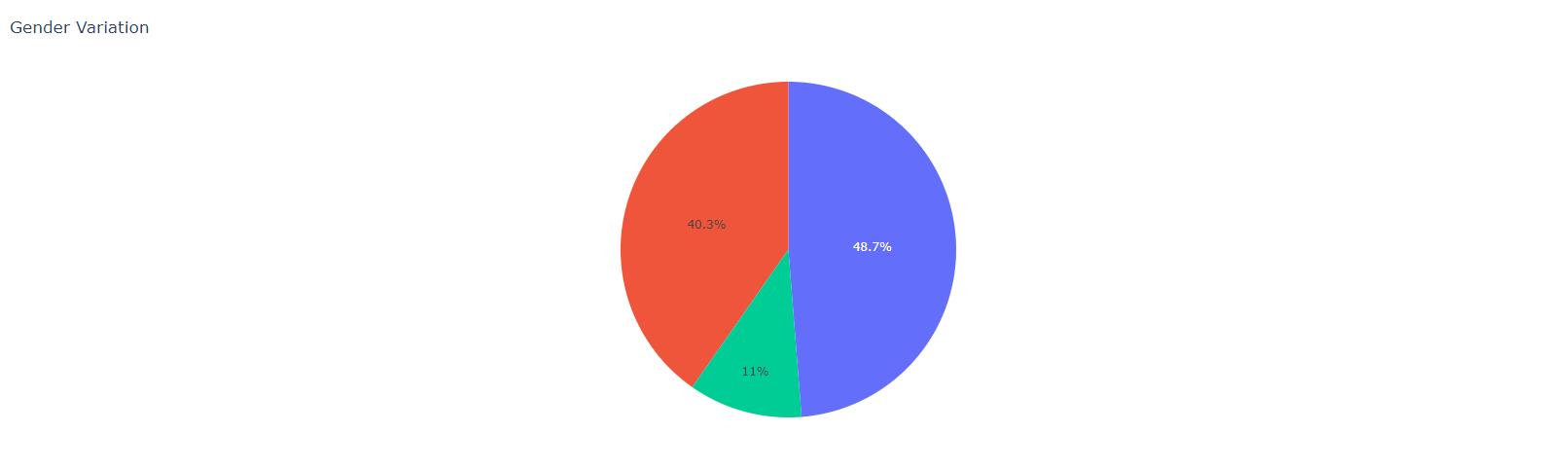


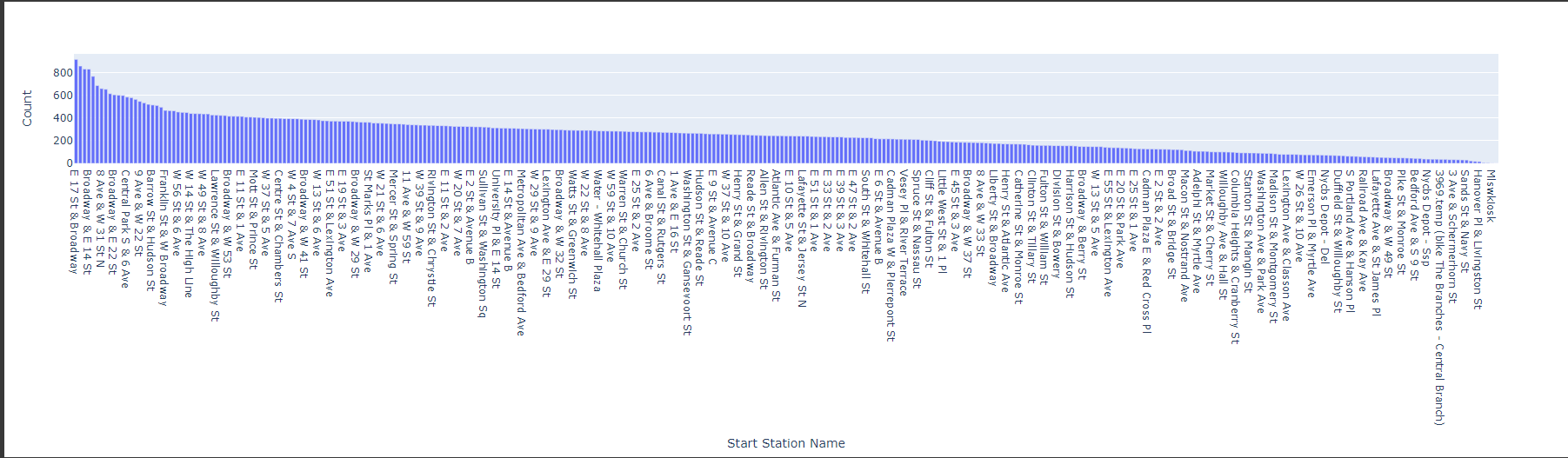
Age group differentiation by bike:

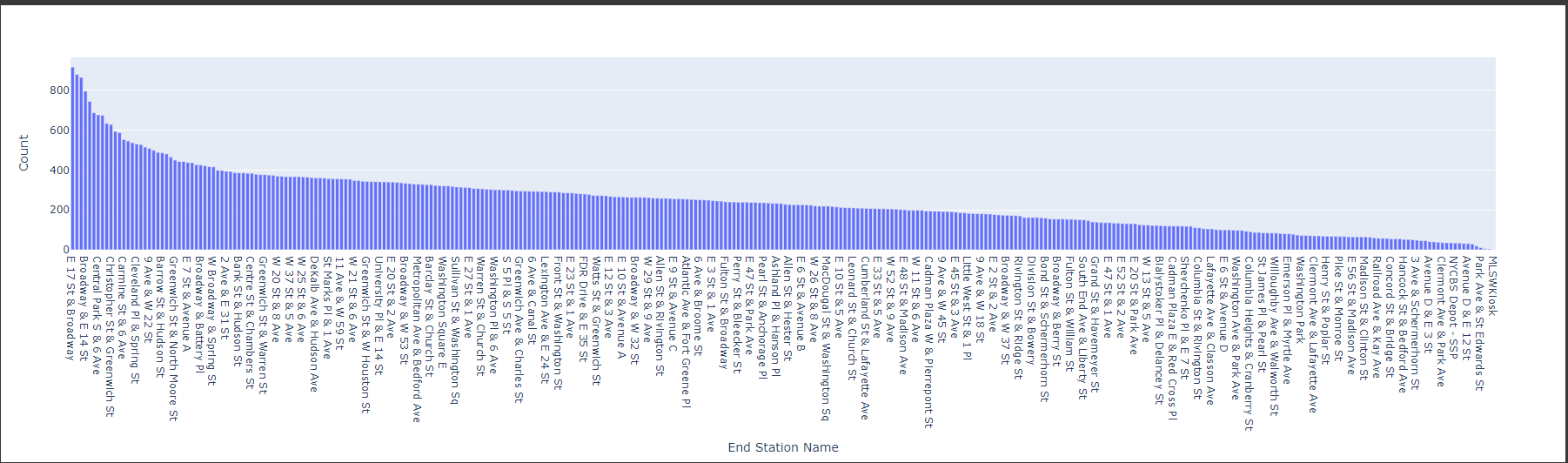


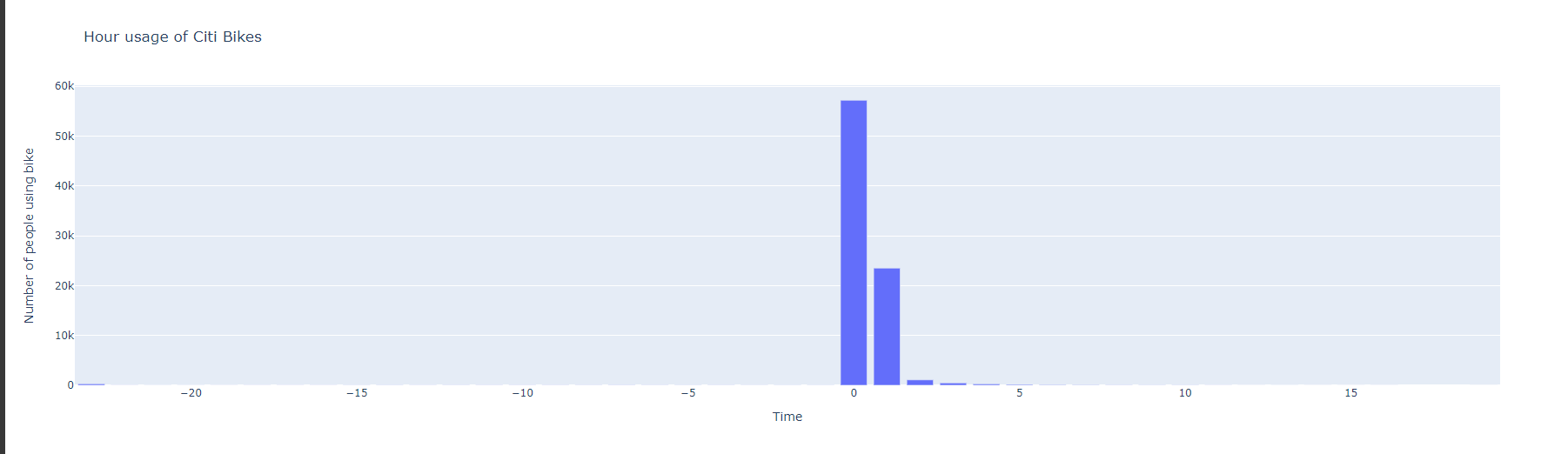
Top 10 Start Station Names with Respect to Customer Age Group:



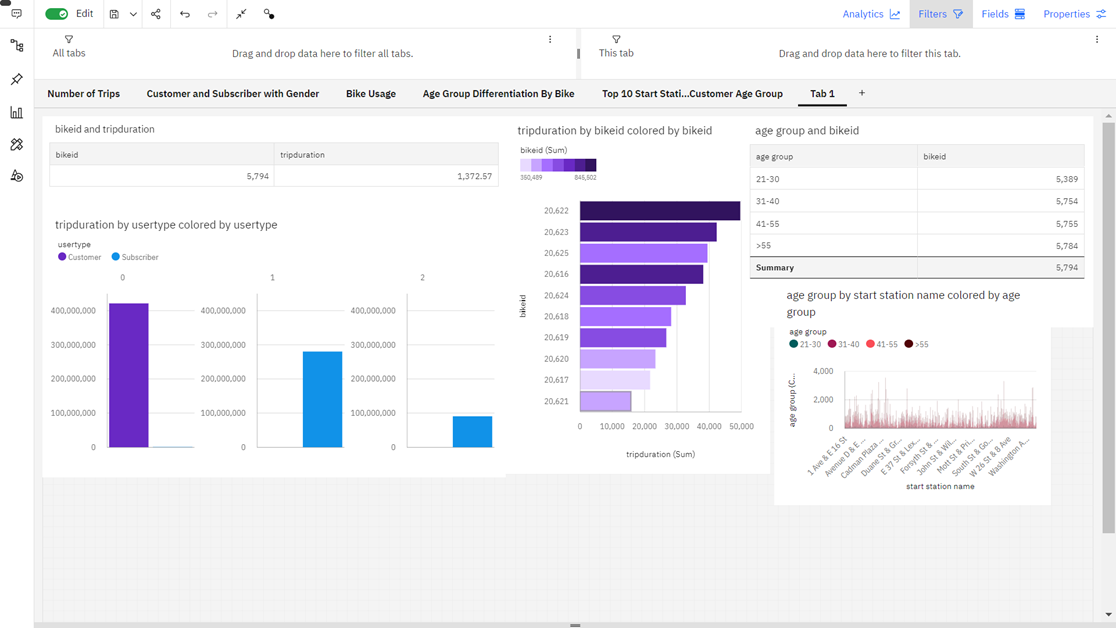








1. **CREATING THE DASHBOARD**

****

1. **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**.**

1. **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.

1. **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.

1. **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"})**

1. **GITHUB LINK**

[**https://github.com/IBM-EPBL/IBM-Project-10706-1659197931**](https://github.com/IBM-EPBL/IBM-Project-10706-1659197931)