B.M.S College of Engineering



**P.O. Box No.: 1908 Bull Temple Road, Bangalore-560 019**

**DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING**

**Course – Big Data Analytics**

**Course Code –20IS6PEDLG**

**AY 2021-2022**

**FACEBOOK DATA ANALYSIS**

Submitted to – Shubha V Rao (Asst. Professor)

Submitted by -

Omisha N USN: 1BM18IS063

Ritesh Pamadi USN:1BM18IS080

Aditya Kamat USN:1BM18IS003

**B.M.S College of Engineering**

**P.O. Box No.: 1908 Bull Temple Road**

**Bangalore-560019**



**CERTIFICATE**

Certified that the Project has been successfully presented at **B.M.S College Of Engineering** by, **Ritesh Pamadi** , Omisha N, Aditya Kamat bearing USN: **1BM18IS080, 1BM18IS063 and 1BM18IS003** in partial fulfillment of the requirements for the IV Semester degree in **Bachelor of Engineering in Information Science & Engineering** of **Visvesvaraya Technological University,**

**Belgaum** as a part of the course **Bigdata Analytics *(*course code)** during academic year 2020-2021.

**Faculty Name –**

**Designation –**

**Shubha V Rao**

**Assistant Professor**

**Department of ISE, BMSCE**

CONTENTS

|  |  |
| --- | --- |
| **TOPICS** | **PAGE NO.** |
| Abstract | 4 |
| Problem Statement | 4 |
| Introduction | 5 |
| Overview of the Project | 6 |
| High Level Design | 7 |
| Tools Used | 8 |
| Dataset Details | 9 |
| Analysis Perfromed | 10 |
| Implementation | 11-12 |
| Snapshots | 13-18 |
| References | 19 |

ABSTRACT

The concept of big data has been around for years; most organizations now understand that if they capture all the data that streams into their businesses, they can apply analytics and get significant value from it. But even in the 1950s, decades before anyone uttered the term “big data,” businesses were using basic analytics (essentially numbers in a spreadsheet that were manually examined) to uncover insights and trends. The new benefits that big data analytics brings to the table, however, are speed and efficiency. Whereas a few years ago a business would have gathered information, run analytics and unearthed information that could be used for future decisions, today that business can identify insights for immediate decisions. The ability to work faster – and stay agile – gives organizations a competitive edge they didn’t have before. Through our project we intend to carry out analysis on a preferably large dataset. So we have chosen the dataset obtained from several Facebook users. By carrying out certain operations, we intend to harness their data and use it to identify new opportunities.

PROBLEM STATEMENT

While [using Facebook for business](https://blog.hootsuite.com/facebook-marketing-tips/), people involved need to take a structured approach that ties their social media efforts to real [business goals](https://blog.hootsuite.com/smart-social-media-goals/). Analytics give a lot of valuable information that can help them track and measure your results so they can refine your strategy and [measure your return on investment](https://blog.hootsuite.com/measure-social-media-roi-business/).

Understanding how and when people interact with the content they post on Facebook is also an important way to make sure the [Facebook algorithm](https://blog.hootsuite.com/facebook-algorithm/) works for them, rather than against you.

This guide to Facebook Insights will help you understand how and why each metric is important to your overall social media strategy.

INTRODUCTION

A Social network is defined as a network of relationships or interactions, where the nodes consist of people or actor, and the edges or arcs consist of the relationships or interactions between these actors. Social networks and the techniques to analyse them existed since decades. There can be several type of social networks like email network, telephone network, collaboration network. But recently online social networks like Facebook, Twitter, LinkedIn, MySpace etc have been developed which gained popularity within very short amount of time and gathered large number of users. Facebook is said to have more than 500 million users in 2010. The field of social networks and their analysis has evolved from graph theory, statistics and sociology and it is used in several other fields like information science, business application, communication, economy etc. Analysing a social network is similar to the analysis of a graph because social networks form the topology of a graph. Graph analysis tools have been there for decades. But they are not designed for analysing a social network graph which has complex properties. An online social network graph may be very large. It may contain millions of nodes and edges. Social networks are dynamic i.e. there is continuous evolution and expansion. A node in social network usually has several attributes. There are small and large communities within the social graph. Old graph analysis tools are not designed to manage such large and complex social network graph. Facebook is a preferred social network by marketers, not only because of the sheer number of users represented but also because of its incredibly insightful analytics suite. It’s important to be able to analyze customers and their behavior on a micro level due to Facebook’s ever-changing algorithm, and the implications for our content and business. If we refuse to adapt our approach based on these insights, we’re doomed to obscurity on the news feed. A deep Facebook data analysis shouldn’t be a one and done situation. Ideally, we’ll be auditing our efforts every few months or so at most. This will help us predict the likings, and the general summary of many users as a whole.

OVERVIEW OF PROJECT

In this project, we use Exploratory Data Analysis (EDA) on the Facebook dataset. It is the process of understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. It often takes much time to explore the data. We implement EDA on the Facebook dataset using Google Colab. We plot many graphs explaining details of the Facebook dataset and showing different analysis.

This exploratory data analysis gives insights from Facebook dataset which consists of identifying users that can be focused more to increase the business.

HIGH LEVEL DESIGN

TOOLS USED

Google Colab- Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs.

Colab notebooks are stored in [Google Drive](https://drive.google.com/), or can be loaded from [GitHub](https://github.com/). Colab notebooks can be shared just as you would with Google Docs or Sheets.

DATASET DETAILS

This dataset contains 99903 entries with 15 columns.

Column names are well defined so that everyone can interpret easily.

(<https://www.kaggle.com/sheenabatra/facebook-data>)

It contains 15 columns which are-

* 1. UserID (primarykey)
  2. age- Age of user in years
  3. dob\_year- Year of Birth of user
  4. dob\_day- Day of Birth of user
  5. dob\_month- Month of Birth of user
  6. gender- Male or Female
  7. tenure- Days since user started using Facebook
  8. friend\_count- Number of friends
  9. friendships\_initiated- Number of friend requests sent
  10. likes- Number of likes given
  11. likes\_received- Number of likes received
  12. www\_likes- Number of likes given through website
  13. www\_likes\_received- Number of likes received through website
  14. mobile\_likes- Number of likes given through the app
  15. mobile\_likes\_received- Number of likes received through the app

ANALYSIS PERFORMED

In this project we have analyzed the dataset to:

1. Find out what age group uses Facebook the most
2. Find which gender has more contacts or friends on Facebook
3. Compare likes given and received between age groups
4. See whether web interface preferred more or mobile and which age group uses which interface
5. Find out number of users in each birthday month
6. Which gender uses fb more
7. Which age group has more friends
8. Compare likes given and received between age groups mobile interface

IMPLEMENTATION

import pandas as pd

import numpy as np

import seaborn as sns                       #visualisation

import matplotlib.pyplot as plt             #visualisation

%matplotlib inline

sns.set(color\_codes=True)

Loading the dataset into the frame

from google.colab import files

uploaded = files.upload

import pandas as pd

df=pd.read\_csv('pseudo\_facebook.csv')

df.head(5)

df.tail(5) # To display the botton 5 rows

Checking the types of data

df.dtypes

# computing number of rows

rows = len(df.axes[0])

# computing number of columns

cols = len(df.axes[1])

print(df)

print("Number of Rows: ", rows)

print("Number of Columns: ", cols)

df.shape

duplicate\_rows\_df = df[df.duplicated()]

print("Number of duplicate rows: ", duplicate\_rows\_df.shape)

df.count() # Used to count the number of rows

df.count()

Renaming

df = df.rename(columns={"likes": "Likes","userid": "User ID","age": "Age","gender": "Gender", "friend\_count": "Friends", "www\_likes": "Web Likes" })

df.head(5)

Dropping the missing values

print(df.isnull().sum())

df = df.dropna()    # Dropping the missing values.

df.count()

print(df.isnull().sum())   # After dropping the values

Detecting Outliers

plt.hist(x=df['tenure'])

plt.hist(x=df['Gender'])

plt.hist(x=df['Age'])

Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1

print(IQR)

df = df[~((df < (Q1 - 1.5 \* IQR)) |(df > (Q3 + 1.5 \* IQR))).any(axis=1)]

df.shape

Plots

df.Gender.value\_counts().nlargest(40).plot(kind='bar', figsize=(10,5))

plt.title("Gender")

plt.ylabel('Number of people')

plt.xlabel('Gender');

df.Age.value\_counts().nlargest(40).plot(kind='bar', figsize=(10,5))

plt.title("Age")

plt.ylabel('Age')

plt.xlabel('Number of people');

df.dob\_month.value\_counts().nlargest(40).plot(kind='bar', figsize=(10,5))

plt.title("Birthday Month")

plt.ylabel('dob\_month')

plt.xlabel('Number of people');

Heat Maps

plt.figure(figsize=(10,5))

c= df.corr()

sns.heatmap(c,cmap="BrBG",annot=True)

Scatterplot

fig, ax = plt.subplots(figsize=(10,6))

ax.scatter(df['Gender'], df['Friends'])

ax.set\_xlabel('Gender')

ax.set\_ylabel('Friends')

plt.show()

fig, ax = plt.subplots(figsize=(10,6))

ax.scatter(df['Age'], df['Friends'])

ax.set\_xlabel('Age')

ax.set\_ylabel('Friends')

plt.show()

fig, ax = plt.subplots(figsize=(10,6))

ax.scatter(df['Age'], df['mobile\_likes'])

ax.set\_xlabel('Age')

ax.set\_ylabel('Mobile Likes')

plt.show()

fig, ax = plt.subplots(figsize=(10,6))

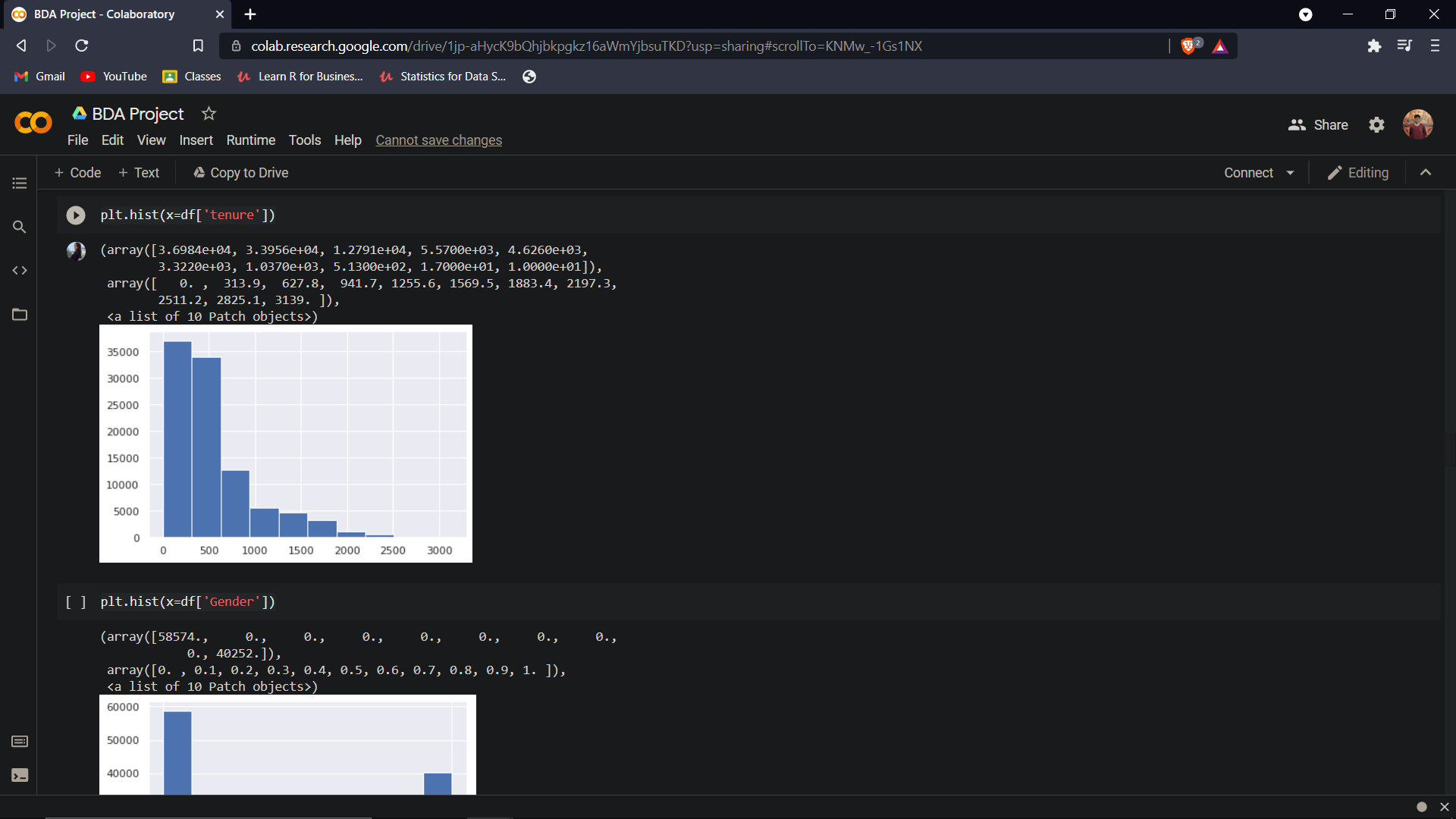
ax.scatter(df['Age'], df['Web Likes'])

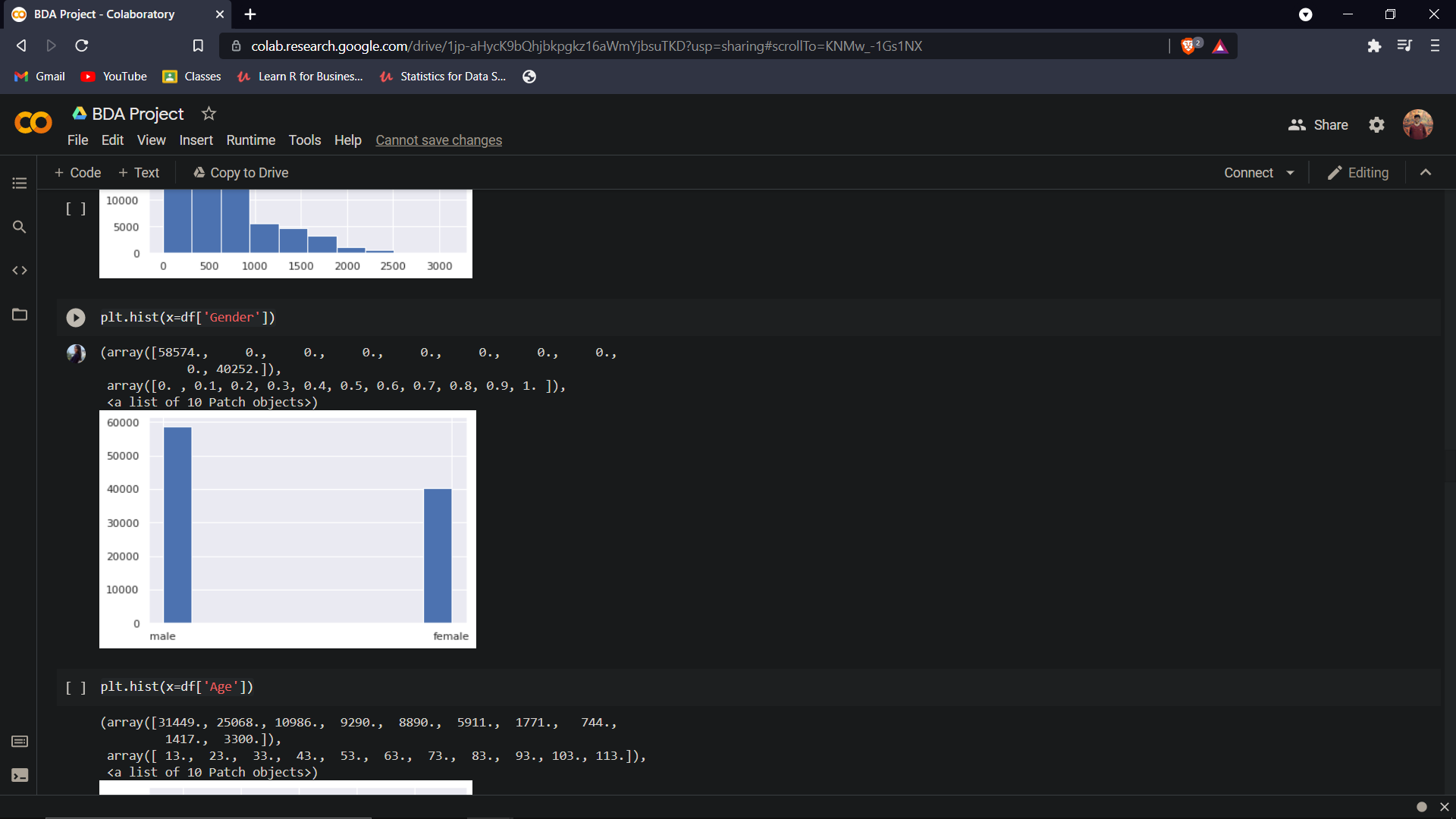
ax.set\_xlabel('Age')

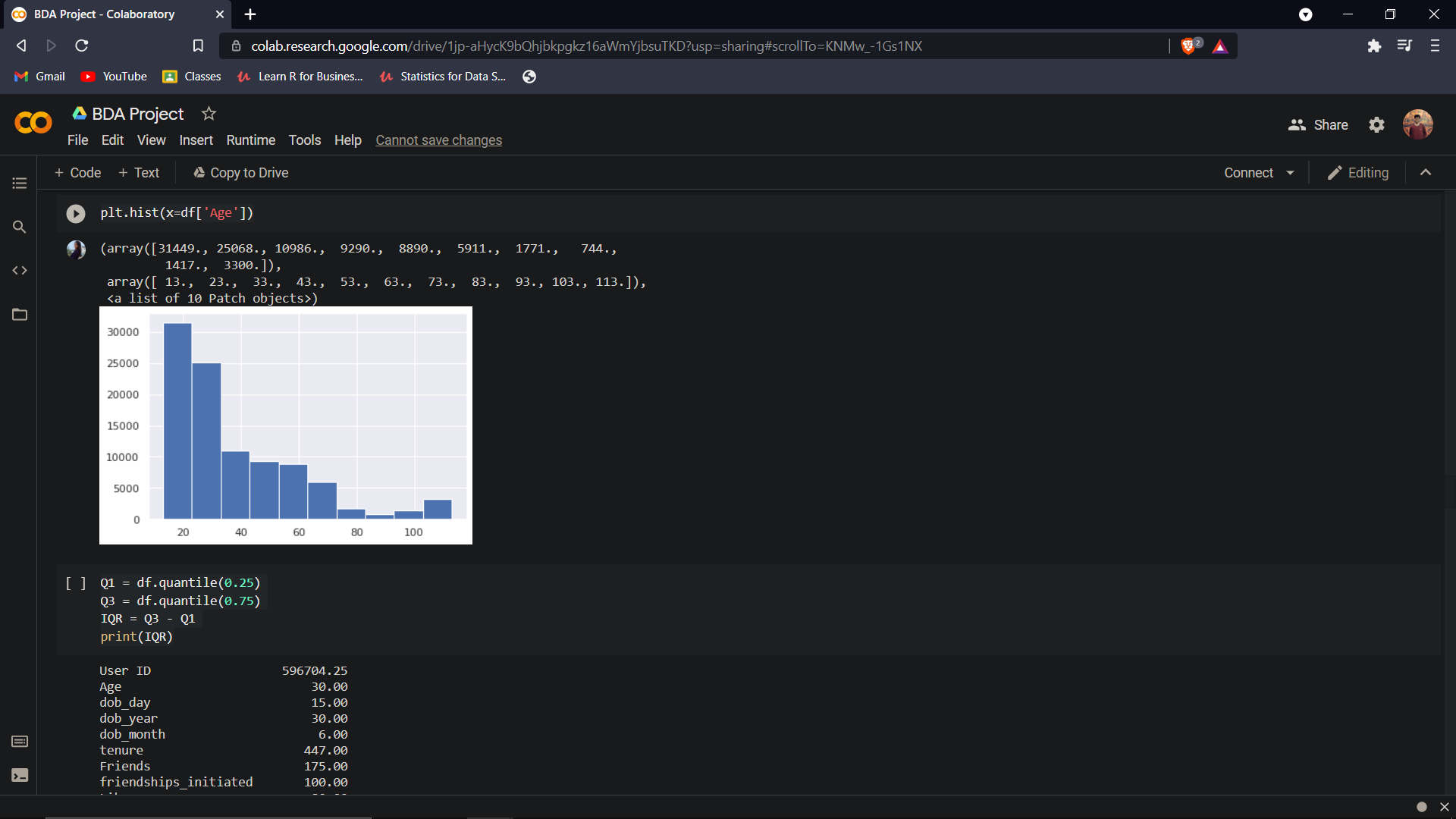
ax.set\_ylabel('Web Likes')

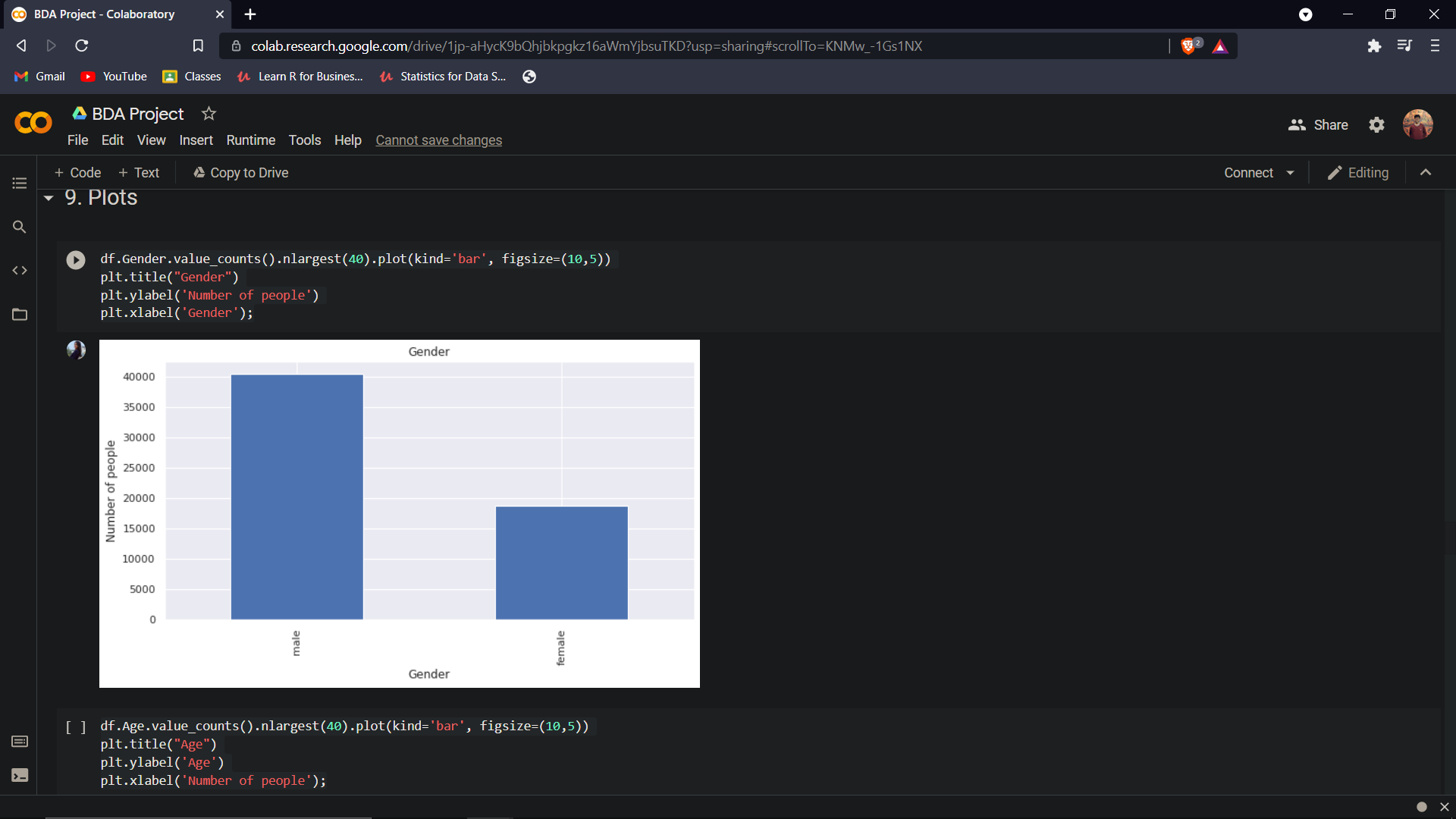
plt.show()

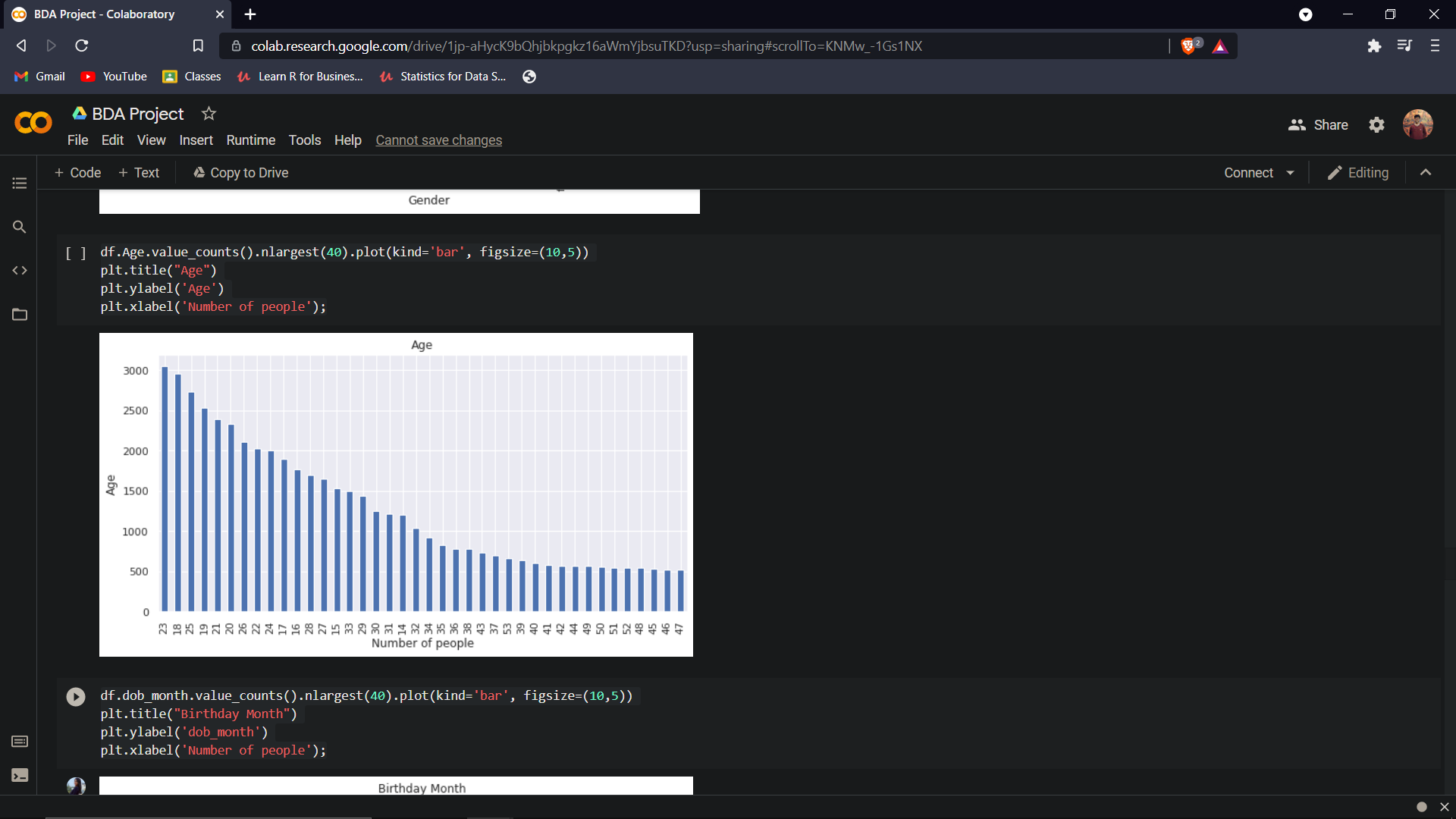
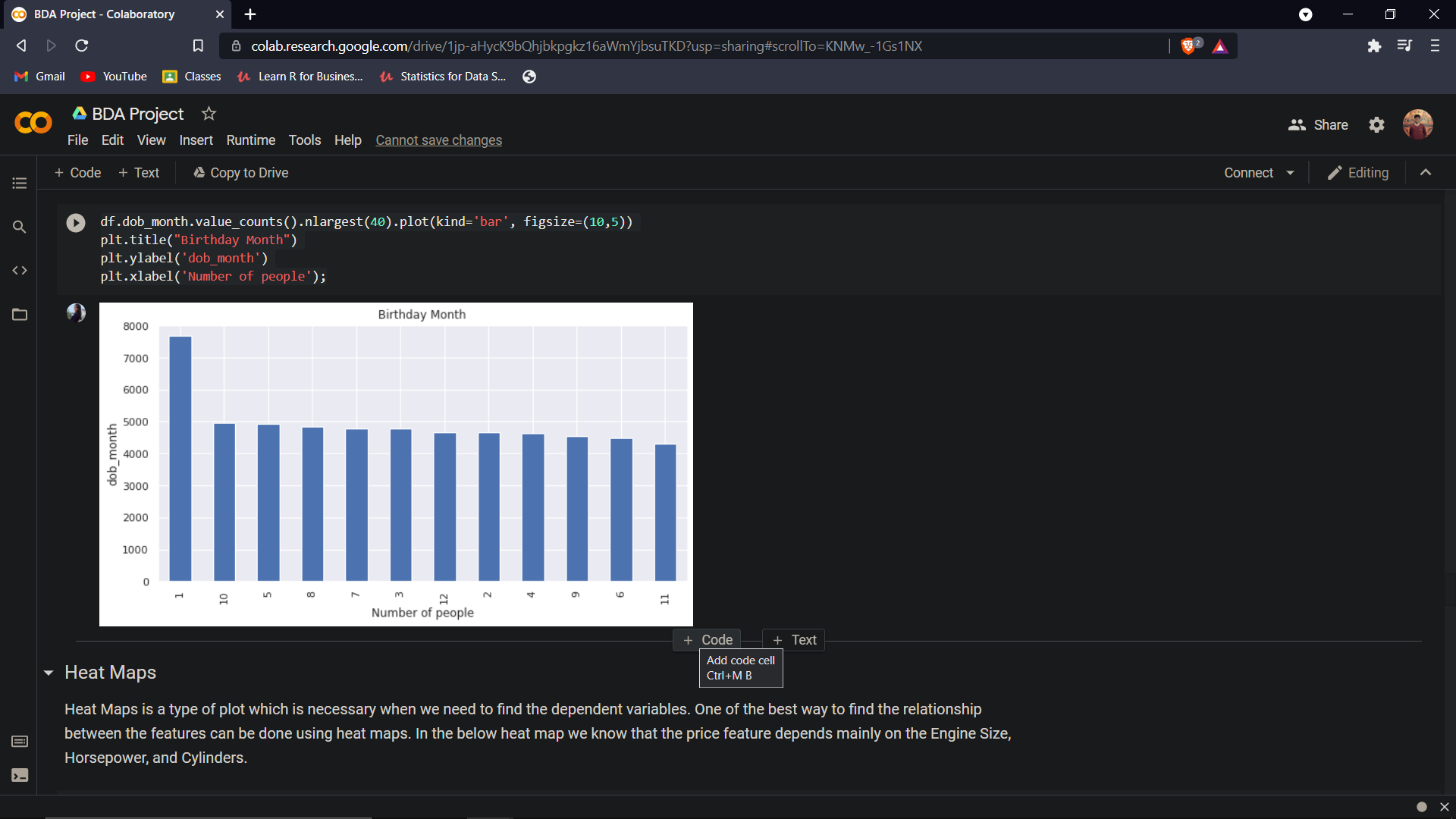
SNAPSHOTS

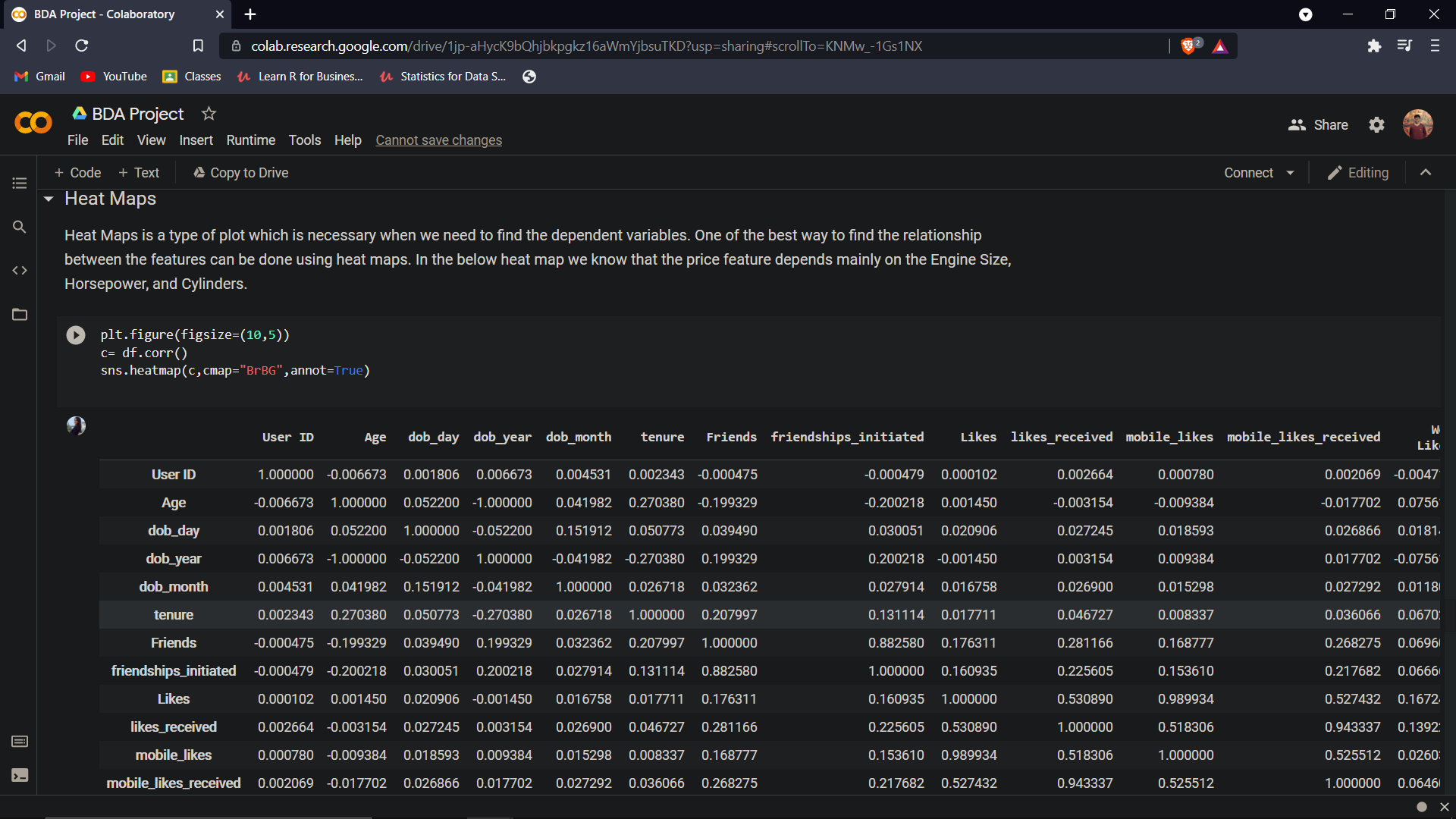


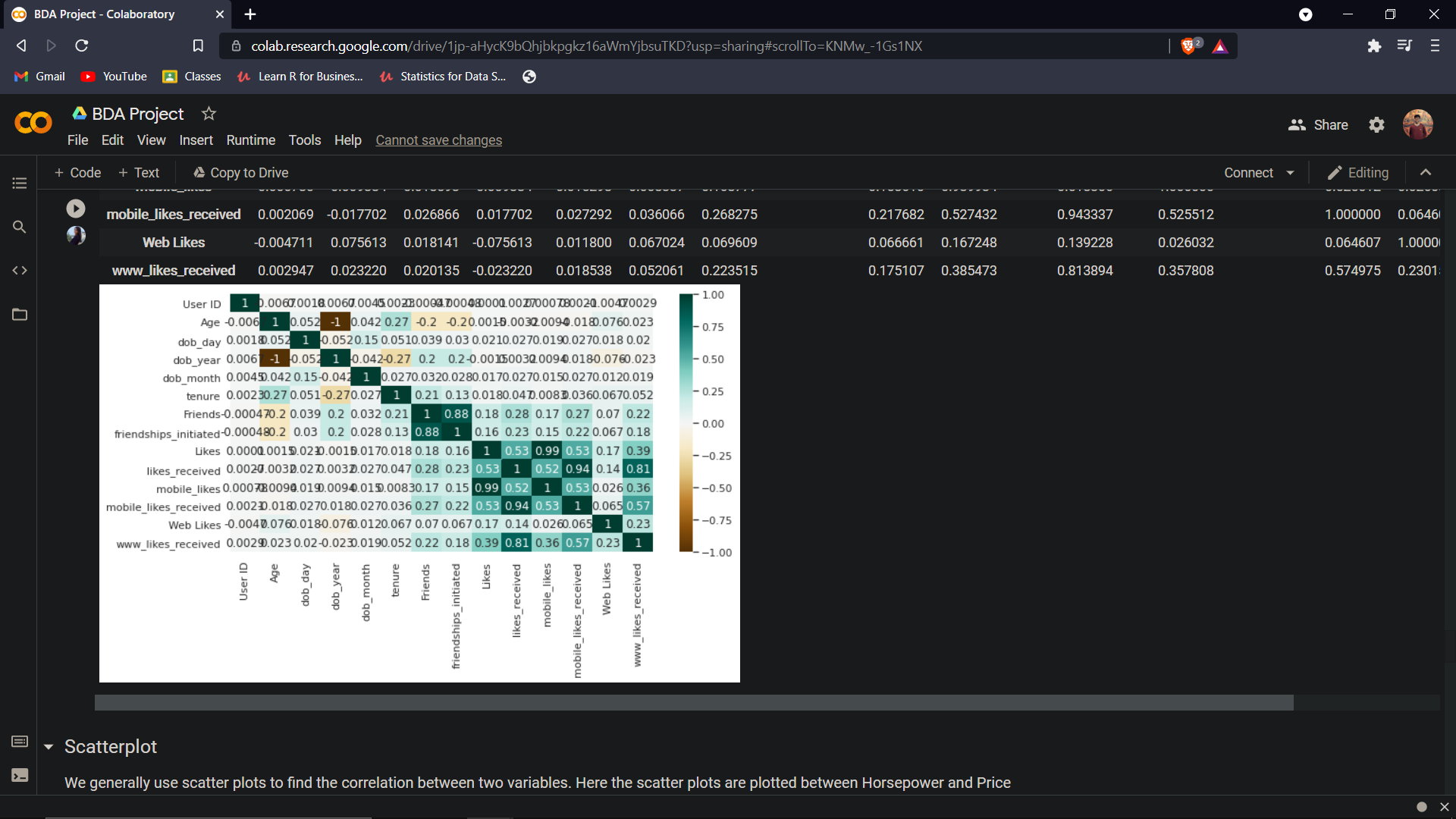


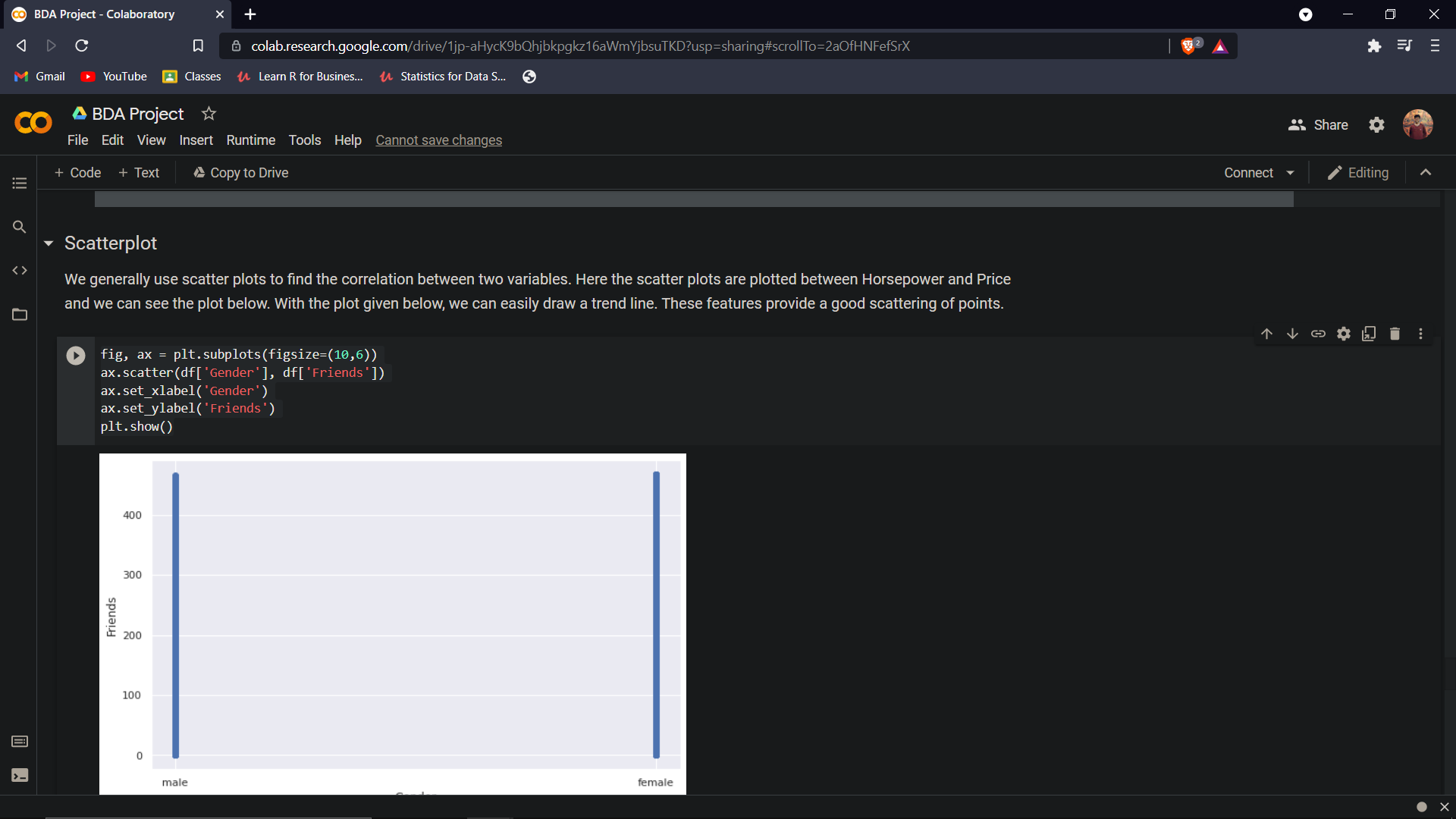
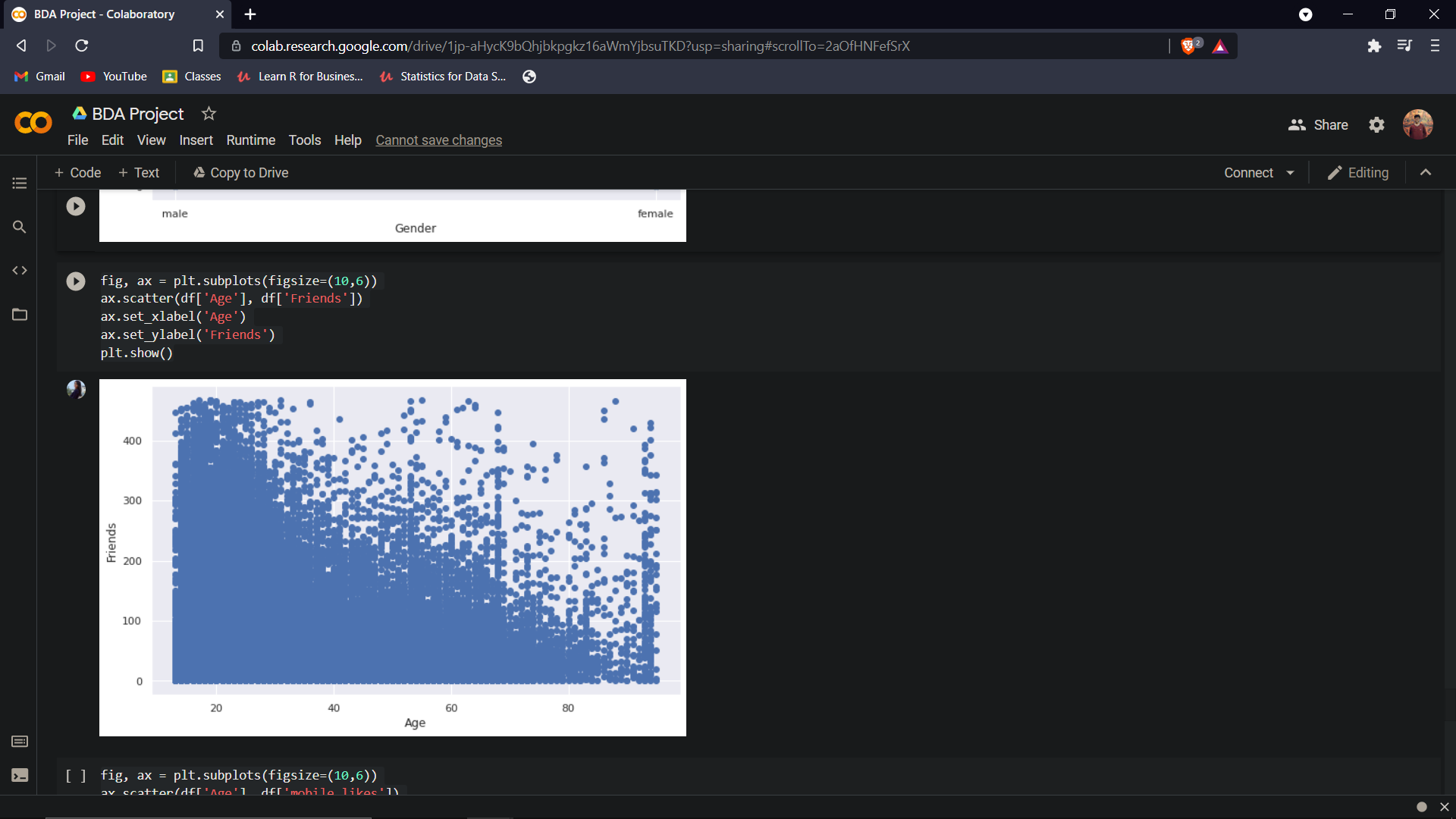


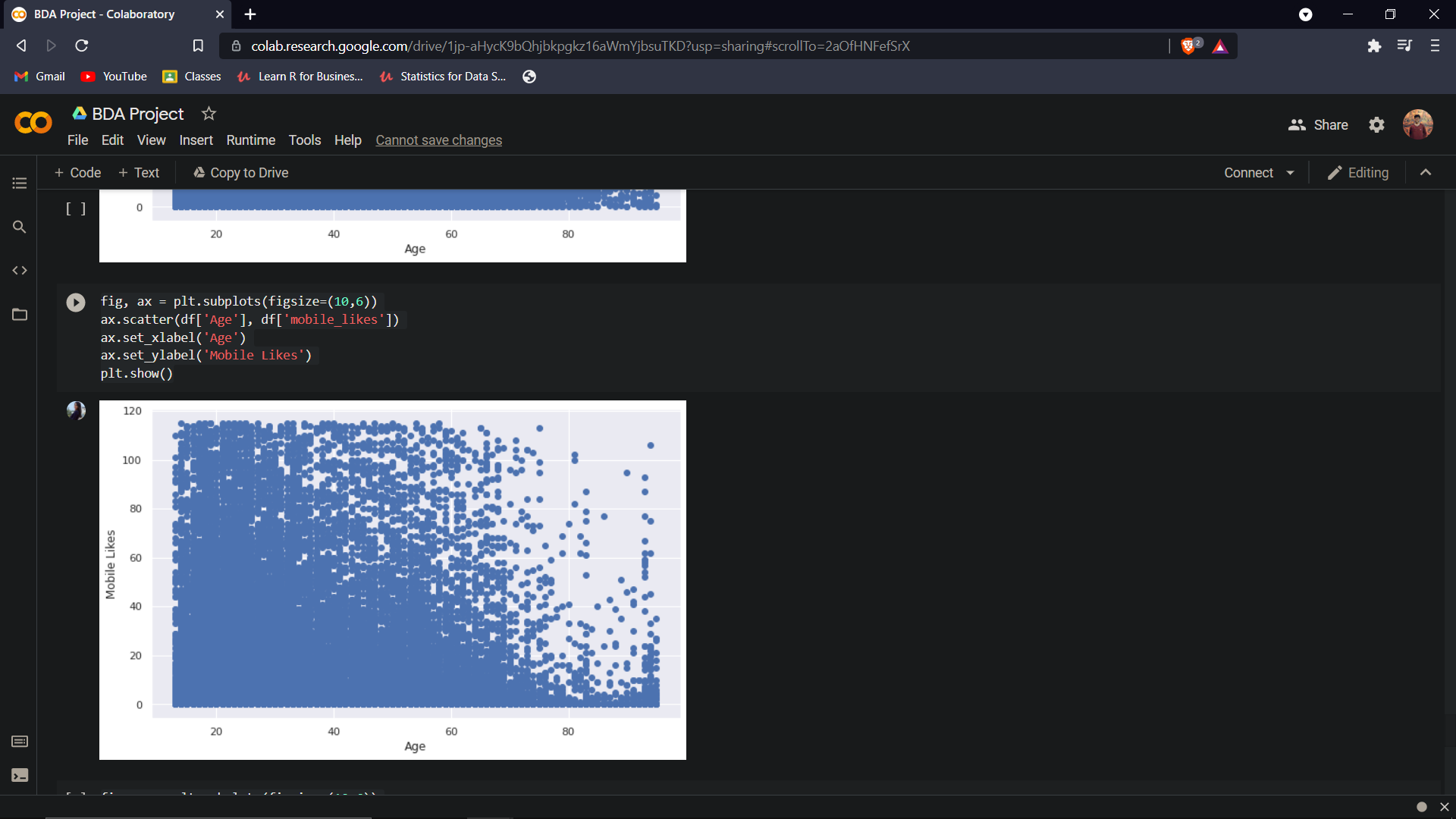


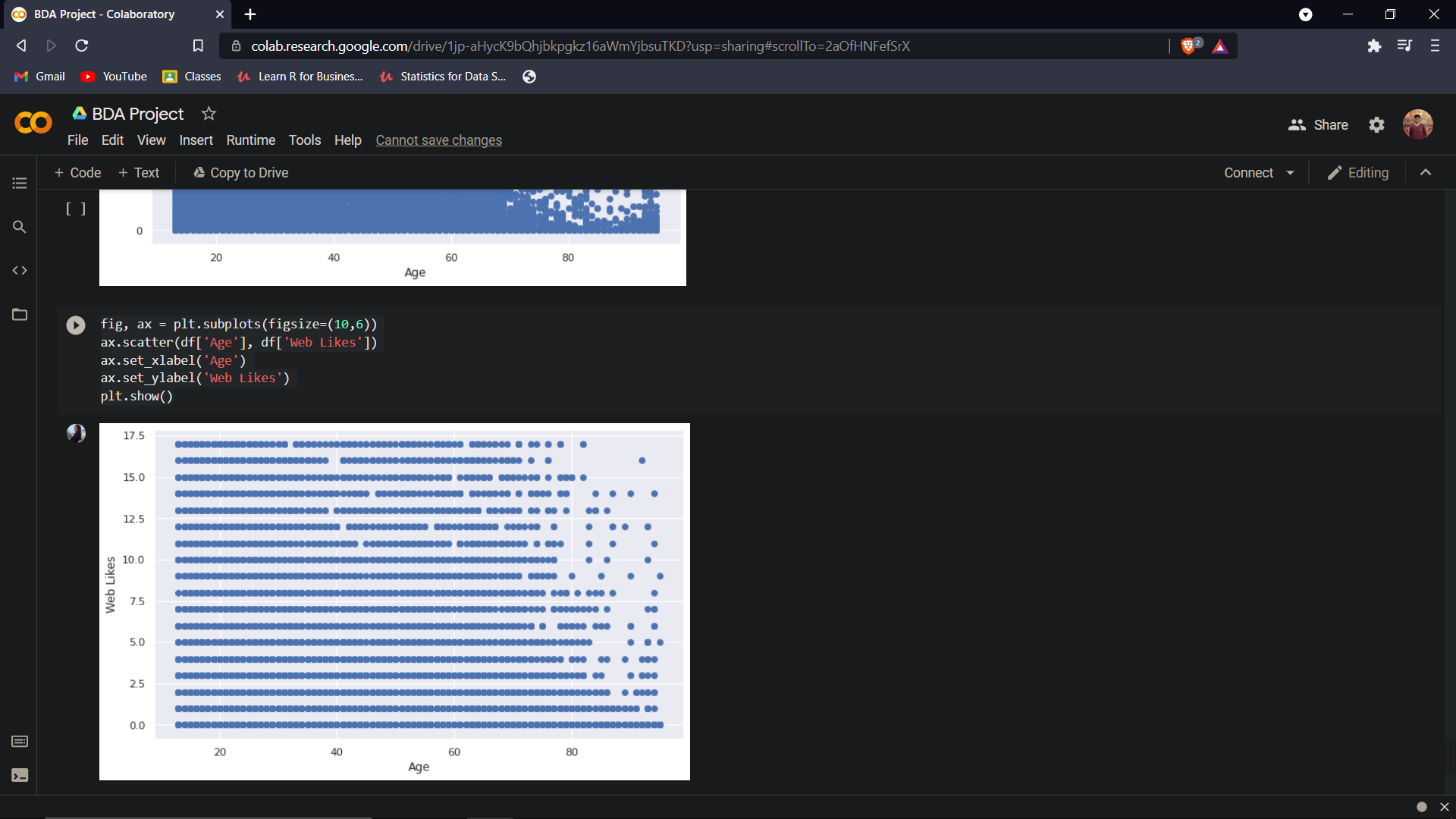
 









REFERENCES

1. https://www.simplilearn.com/how-facebook-is-using-big-data-article
2. towardsdatascience.com
3. medium.com
4. https://maxwellflitton.com/2018/03/23/basic-facebook-data-analysis-with-python/