**CSCE - 5320**

**Project Increment - 2**

## **Project Title:**

**Visualising the Social Media Sales Prediction & Ad-Campaign Analysis**

## **Team Members:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Email\_Id** | **Student\_Id** |
| **Nagasai Gummadi** | **nagasaigummadi@my.unt.edu** | **11540905** |
| **Divya Sri Vakkala** | **divyasrivakkala@my.unt.edu** | **11590101** |
| **Harsha Buddana** | **harshavardhanabuddana@my.unt.edu** | **11521994** |
| **Eswara Reddy Thimmapuram** | **eswarareddythimmapuram@my.unt.edu** | **11506566** |

**Increment - 2**

As part of Increment-2, Our main focus was to deploy the web application using Streamlit cloud, The initial Insights we have provided in the Increment -1, We converted the important metrics as primary attributes and created the web application

**Deployed Web-app URL** - <https://hbuddana-sdvproj-app-d9bhfj.streamlit.app>

**• Domain**

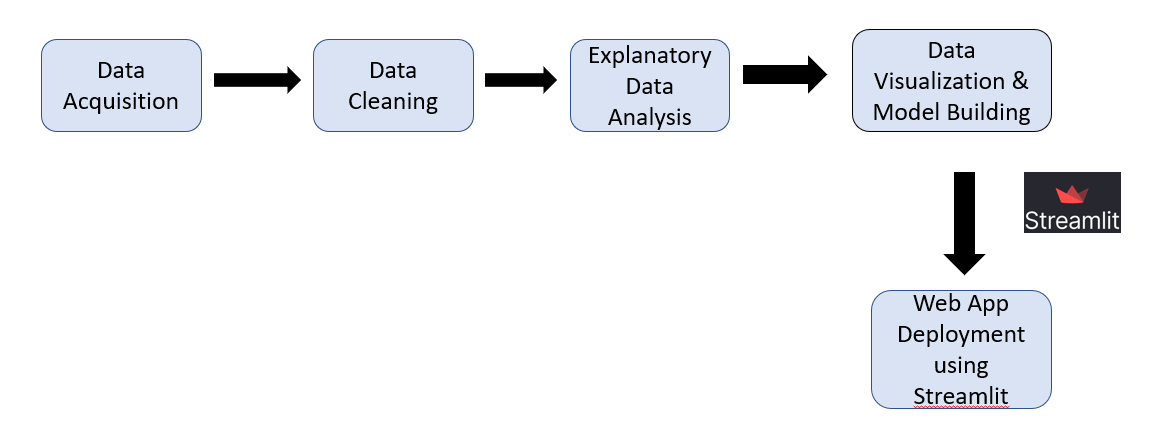
Our Project comes under the domain of **Digital Marketing.** It is a broad domain and deals with use of digital channels such as Search engines, Social Media, Email marketing and more. This domain majorly relies on a 3-step process. The steps are

1. **Targeting** : Where we first look who is the customer base we want to target for the product. This could be done by a variety of factors such as age, interest, demographics, gender and more.
2. **Reaching Out** : Process of reaching the Target audience and the tools required or means of communication we use to reach the defined Target audience
3. **Conversion Optimization** : Once we are done with reaching out, Now we need to optimise the user experience of the user in such a way that the ad we sent should convert into a Sale.

By following the above defined process, Digital marketers can turn ads into sales and drive real valuable results to their business.



**WorkFlow Diagram -**



In this project, first we acquire the required data and then do the basic data cleaning and then do the Exploratory Data analysis and find the insights and the correlations in the dataset and then using the Matplotlib library of the python we have done the basic data visualisations and the we have developed a python code using streamlit to develop the web application once the code is done we pushed it to GitHub and then connected the GitHub repo to the Streamlit Cloud to deploy the app.

**Related Work -**

The background work of the article with title [5] Social Media Advertisement and Its Effects in Sales Prediction- An Analysis by subha explores the effect of social media advertising on predicting sales. The primary goal of this paper is to explore the impact of social media advertising as a marketing tactic and how it affects forecasting sales. It is important for us to keep in mind that social media has become a more and more popular medium for businesses for marketing their products and services if we want to understand the study's background. Companies can target specific age groups and reach many people at a low cost by using social media advertising. Social media platforms can give companies useful information about consumer behaviour and preferences, which can be used to estimate sales and improve marketing plans. In this study, the author studies data from a few social media platforms to look at the effect of social media advertising on prediction of sales. The article gives details on social media advertising's efficiency and its potential to affect sales forecasting. Businesses must comprehend the link between social media advertising and sales forecasting in order to create effective promotional strategies and maximise their return on investment.

The background work of the article with title [6] (When) Can Social Media Buzz Data Replace Traditional Surveys for Sales Forecasting?" by Shi, Yuying, Karniouchina, Explores the effect of social media buzz as an alternative option for traditional surveys in sales forecasting. Nowadays social media has become a popular source for sales forecasting. The word social media buzz data means how much that product or that services interact in social media platforms. The data can be analysed to predict user behaviour on sales trends of products. Previously the traditional methods of sales forecasting used are gathering customer information and their preferences, but this type of methods are time consumption and the data collected is not accurate. This article determines that this social media survey is an alternative for traditional surveys in order to estimate sales. In this article the author compares accuracy of sales based on social media and traditional survey and he concluded that social media sales forecasting is best while doing sales prediction.

**• Data Abstraction**

**Dataset (Type and Attributes):**

For this project, we have downloaded the dataset from Kaggle. This dataset is taken from Facebook’s social media ad campaign.

This dataset is a CSV file. A table format with 11 columns and 1143 data points for each attribute.

The dataset consists of the following columns,

1.) Ad\_id: Each ad has a special ID.

2.) xyzcampaignid: an ID linked to every advertising campaign run by XYZ firm.

3.) fbcampaignid, a unique identifier used by Facebook to track each campaign.

4.) age: The age of the viewer of the advertisement.

5.) Gender: The gender of the viewer of the advertisement

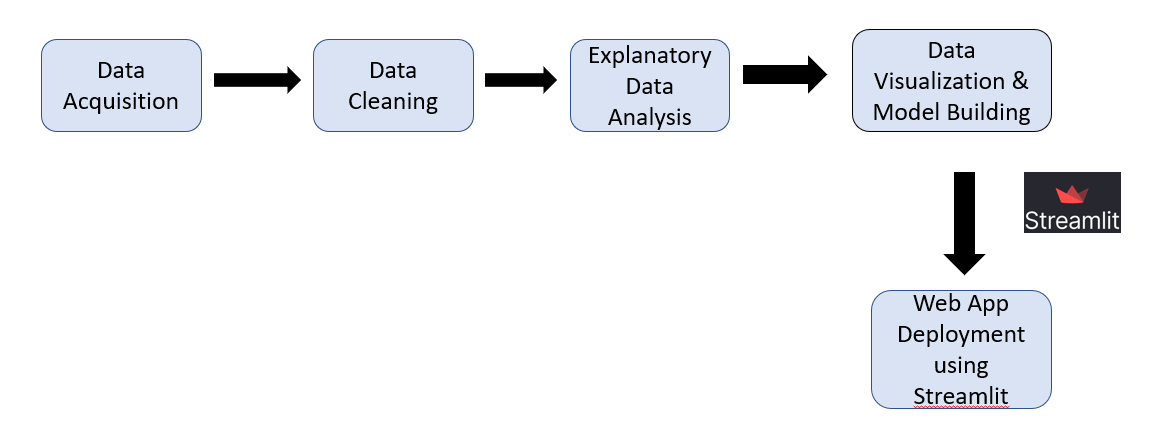
6.)Interest: A code that identifies the group to which a person's interests belong (interests are listed on a person's public Facebook profile).

7.)Impressions: the quantity of times the advertisement was viewed.

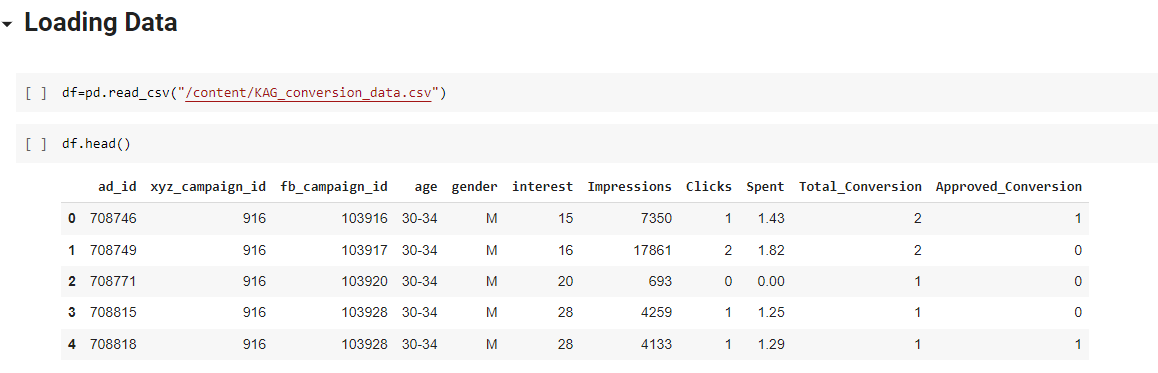
8.) Clicks: The quantity of times that ad was clicked.

9.) Spent: The sum that company xyz paid Facebook to display that advertisement.

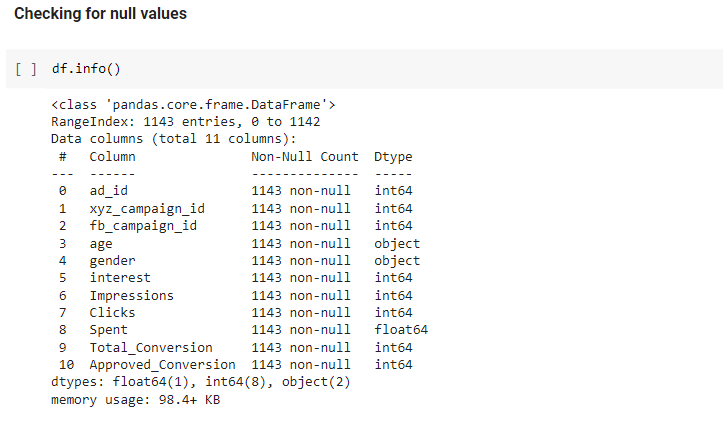
10.) Total conversion: The total amount of individuals that contacted the company about the product after viewing the ad



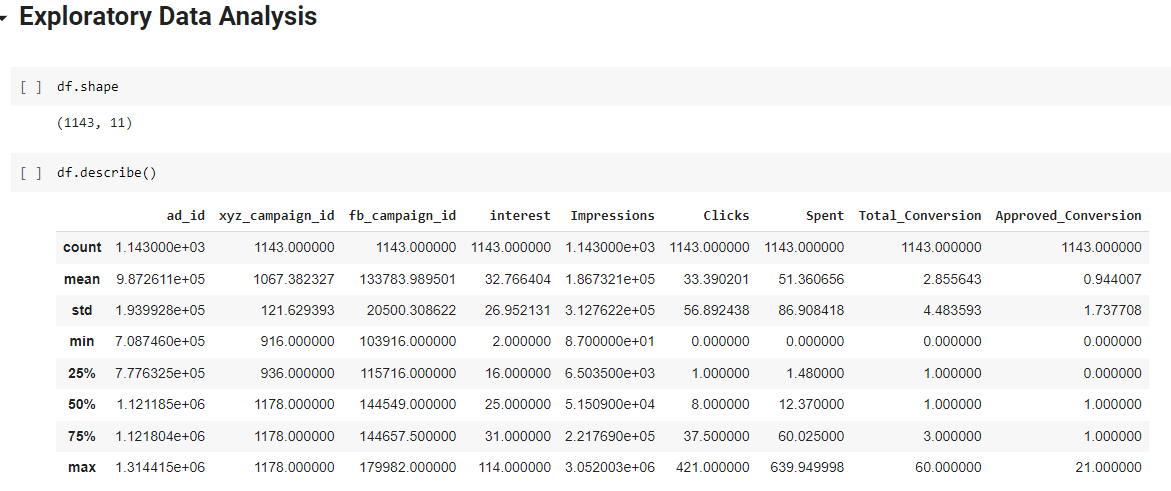
We can also look at a part of the dataset below.



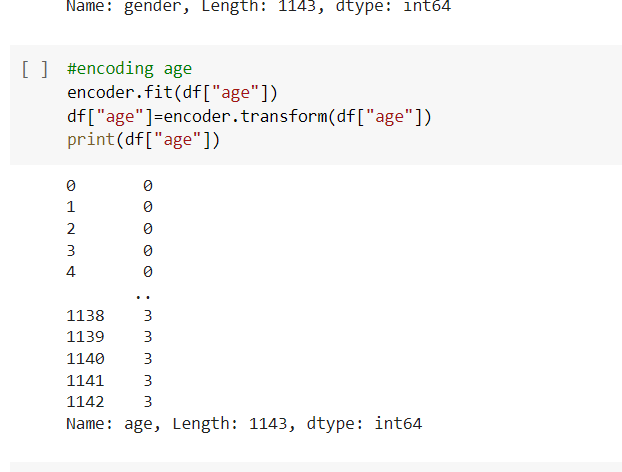
We have also made sure there are no null values in the dataset.



We can have a look at the shape and description of the dataset.

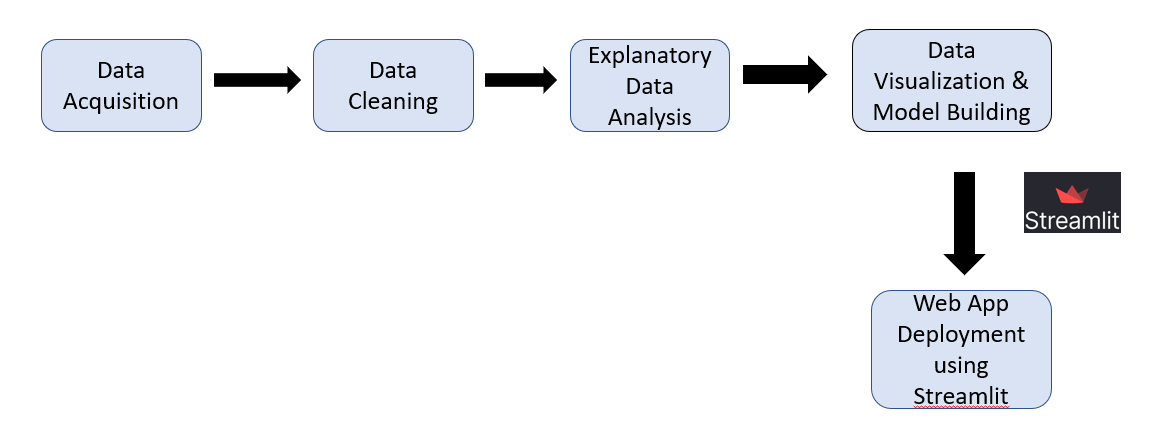


**Data Transformation **

****

Along with the above data transformations, we have also encoded the campaign names below.

**• Task Abstraction:**



The main task of this project is to perform analysis of Facebook ad campaign data and optimise sales conversion. We then use these campaign results to predict future sales. The objective is to analyse the factors that influence sales such as ad impressions, number of clicks and cost for each click etc and create a predictive model that can be applied to optimise future ad campaigns. The project involves tasks like loading the dataset, performing preprocessing, exploratory data analysis and developing machine learning models for sales prediction and then deploying the web application.

**Task (Target and Actions):**

The target of this project is to perform sales prediction based on the results of analysis of Facebook ad campaign dataset.

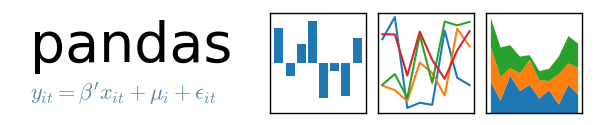
This task involves the following actions:

* Loading and Preprocessing the dataset
* Performing exploratory data analysis
* Training machine learning models for predicting sales
* Performance evaluation of models using different metrics
* Enhance the model based on the obtained results and gain insights that lead to sales prediction

**• Implementation using tools**

**Loading dataset and Exploratory data analysis - Pandas library**

* We have used the pandas library for loading the dataset and performing exploratory data analysis. Pandas is a popular python library which is commonly used for data analysis and manipulation. It is also used for tasks like data cleaning and data preprocessing. There are many tools and functions offered by pandas for data manipulation such as dataframes and series. It is used by many users due to its high performance and productivity.



**Data visualisation - Matplotlib and Seaborn**

* We have used Matplotlib and Seaborn libraries for the purpose of data visualisation. Matplotlib is a popular python library used for creating different kinds of charts, plots and graphs. It offers various visualisation tools making it easier to present the data to the users.
* Seaborn is also a popular data visualisation library built on top of matplotlib. It offers a high level interface for producing attractive and advanced visualisations. It can also be utilised for data exploration. Seaborn offers various functions for visualising different types of data like distribution plots, matrix plots and regression plots.



**Machine learning modelling - Scikit-learn**

* We have used the scikit-learn library for training the machine learning models to predict sales. Scikit-learn is a machine learning library which offers various tools for machine learning tasks such as clustering, classification and regression. It is built on top of Scipy, Numpy and Matplotlib. It also provides tools for data preprocessing, model selection and feature selection.



**Web Deployment - Streamlit**

* Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. In just a few minutes you can build and deploy powerful data apps.

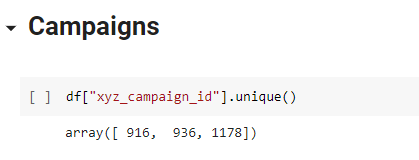


**• Preliminary Results for Analysis**

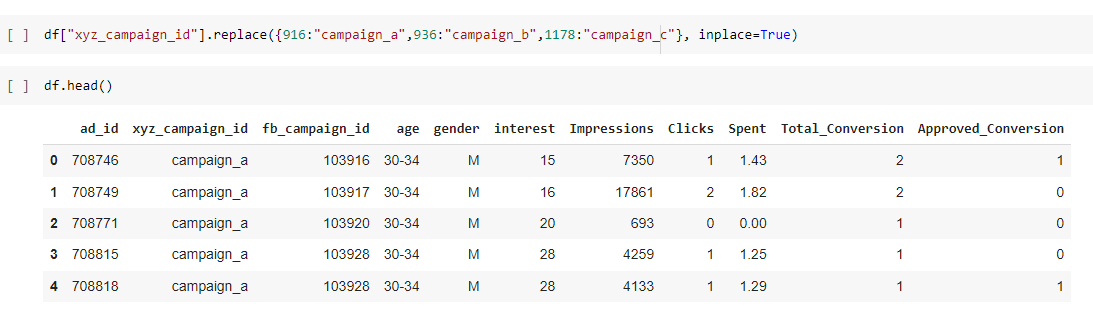
In this part of the project we majorly focus on performing Exploratory Data Analysis on the taken dataset. We have used the above mentioned tools to get the insights from the data. We plan to convert this project into a Web application using Streamlit.

Using the Pandas library and Matplotlib we have generated the following maps.

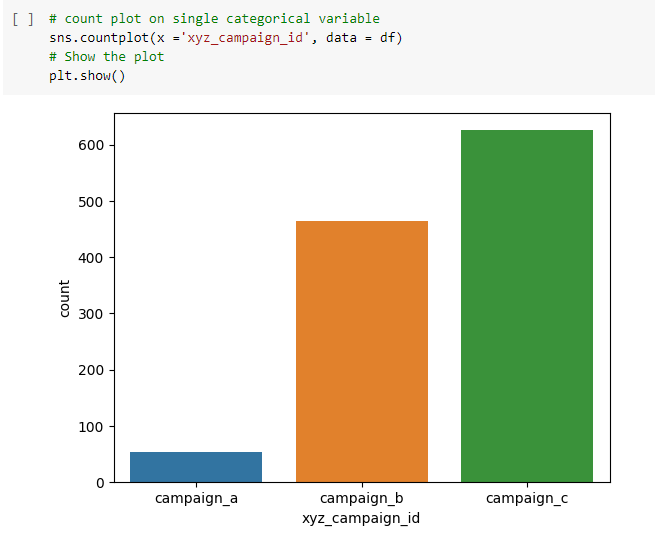
Firstly we analysed the no. of campaigns.



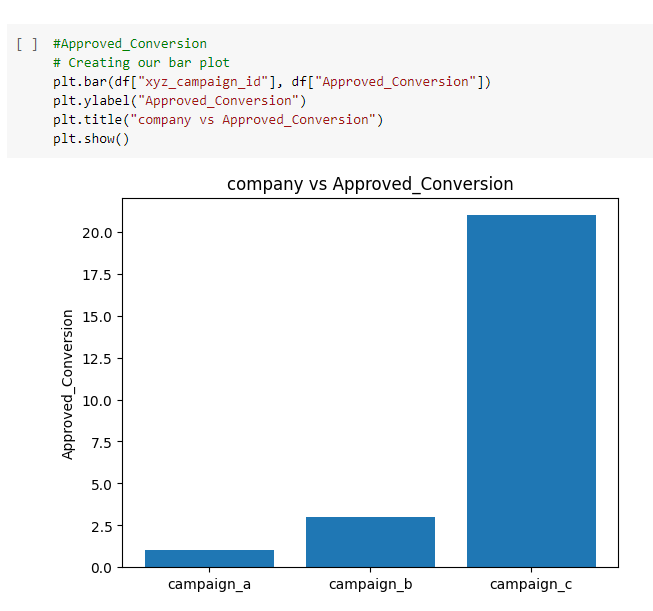
We can observe that the xyz corporation has three separate advertising campaigns here. For better visualisation, we'll now change their names to a,b,c.



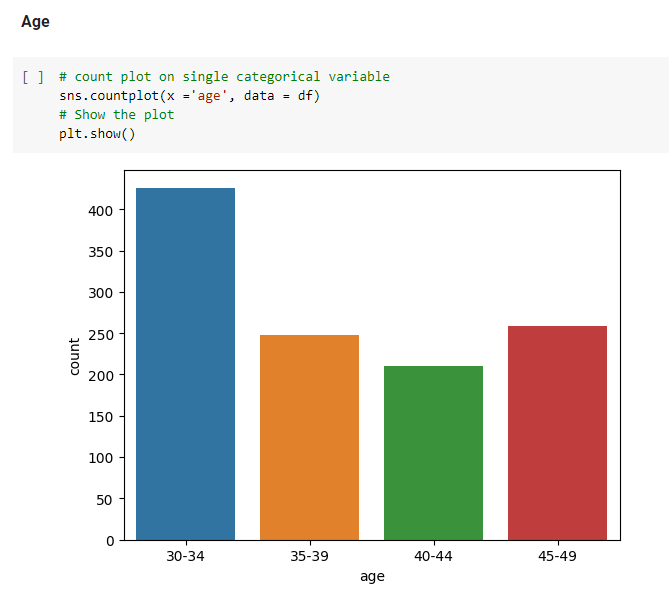
This is how the rendered dataframe looks after we replace the campaign names.



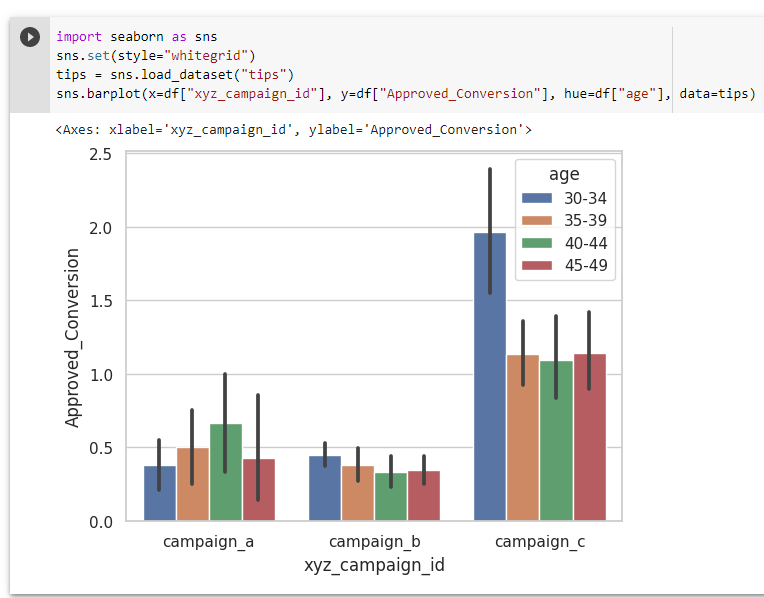
A simple count plot to plot the no of ad\_campaings per each campaign. From the above plot we can infer that Campaign\_c has most no\_of ads.



We tried to plot a countplot for no. Approved\_conversion in each campaign. From the above plot we can deduce that Campaign\_c has the best conversion rate i.e, most products were bought in campaign-c.

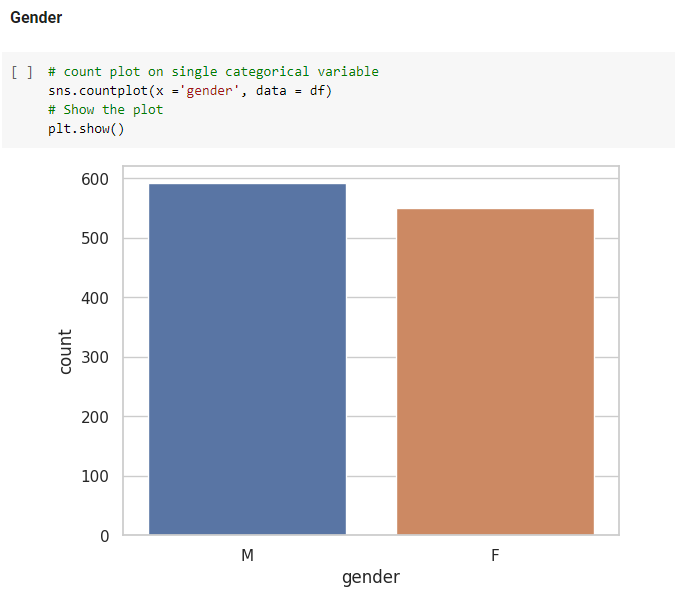


We tried to plot the most targeted age-group from all the campaigns. We can conclude from the above countplot that the age group 30-34 is the most target age-group.

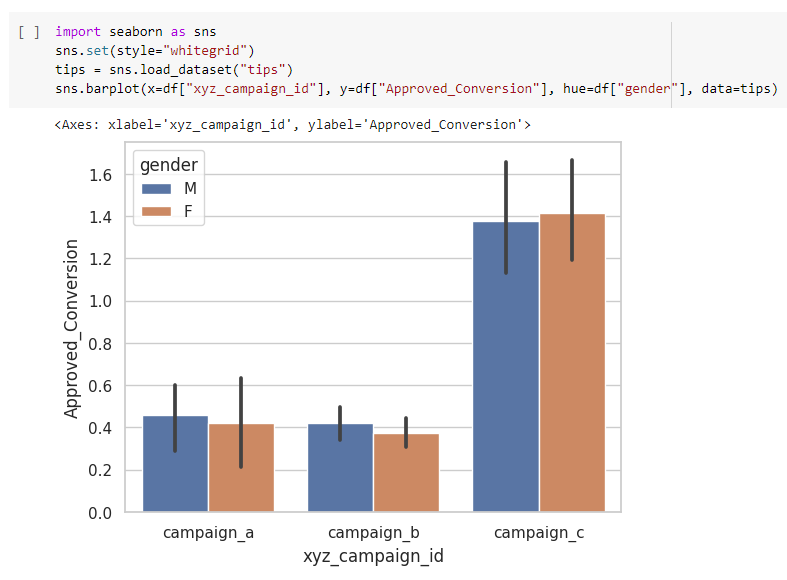


We used seaborn’s barplot function to create a grouped bar plot that shows the relationship between campaign ID, approved conversions, and age.

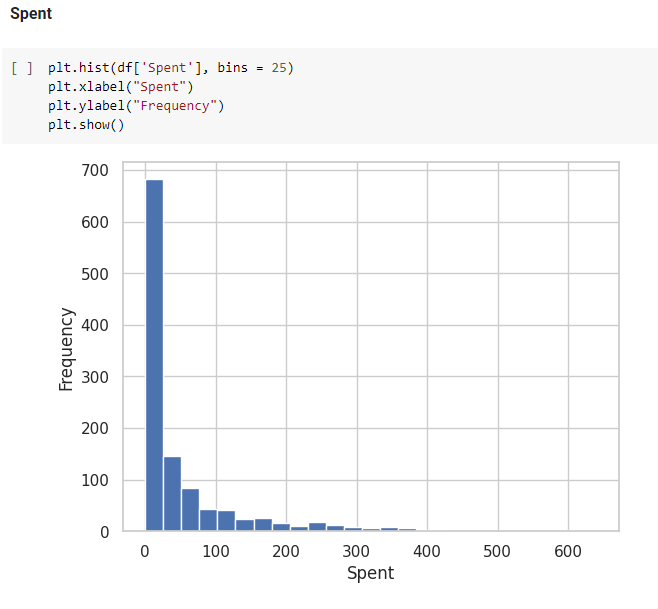
It's interesting to note that in campaign\_c and campaign\_b, the age group of 30-34 shows more interest, whereas in campaign\_a the age group of 40-44 shows more interest.



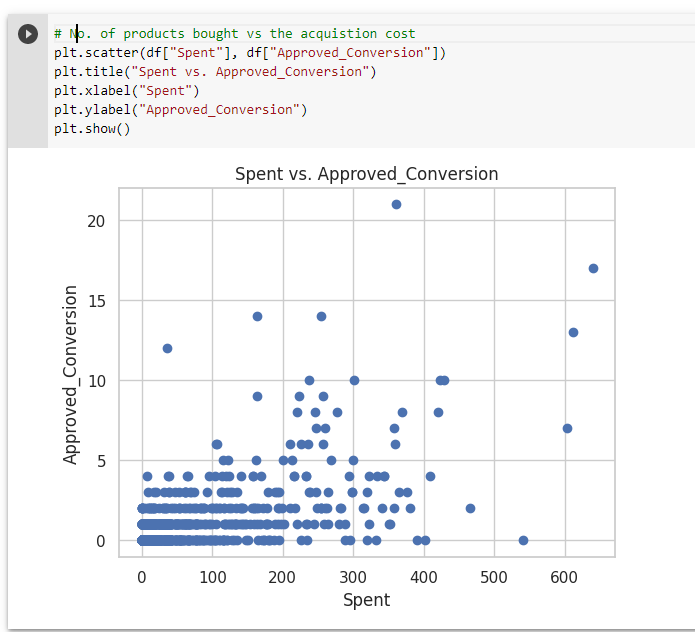
We tried to create a countplot between the gender and we can infer from the graph that there are more male audiences than the female audience.



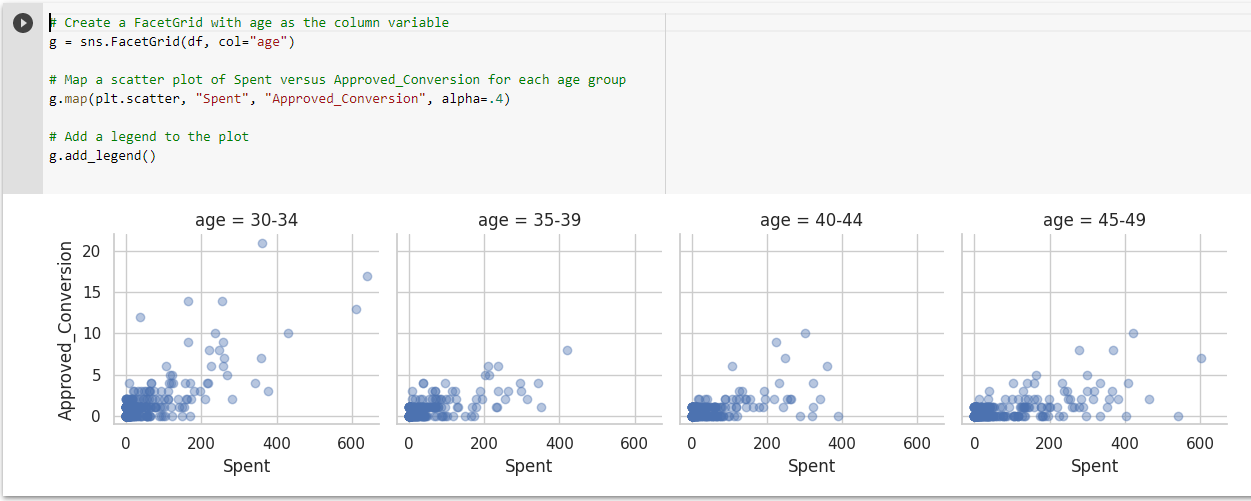
We created a grouped bar plot that shows the relationship between campaign ID, approved conversions, and gender. This allows us to see the gender distribution along each campaign. Both genders show almost similar interest in each campaign.



We tried to plot a histogram, to see the number of data points in the campaigns spent at each spend rate.



We can see, as the amount of money spent increases, the number of products bought increases.

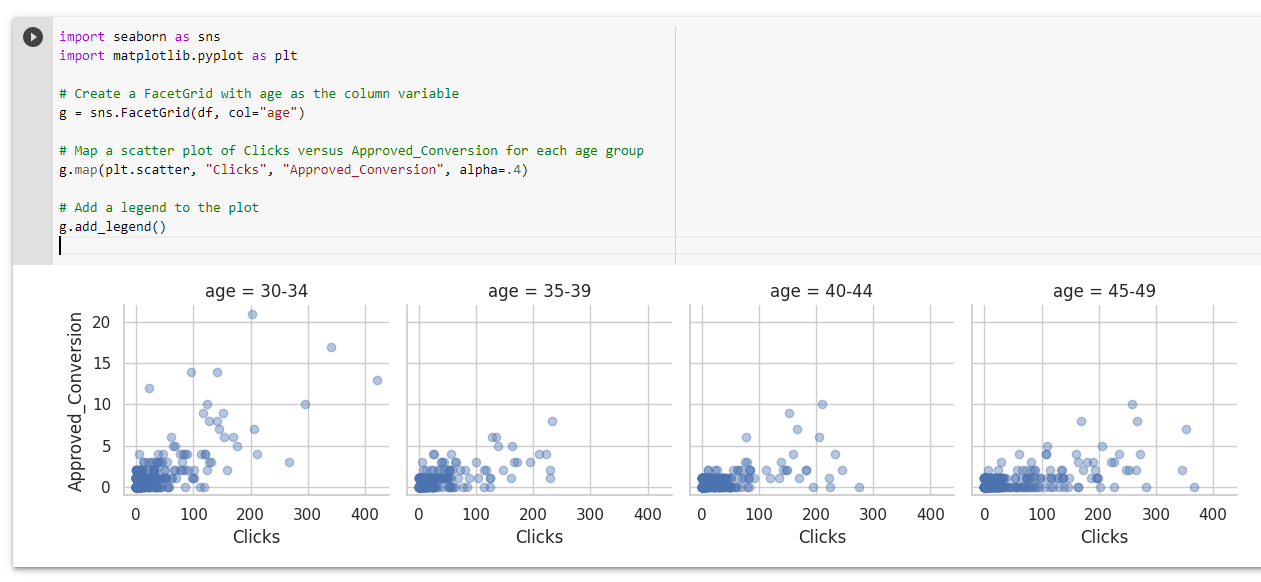


We created a grid of scatter plots(facet grid), with each plot showing the relationship between "Spent" and "Approved\_Conversion" for all age groups.

Now moving on to the vital attributes, “People who actually bought the product”.



Men appear to click on ads more often than women, but after clicking on the ad, women tend to make more purchases.



We created a grid of scatter plots, with each plot showing the relationship between "Clicks" and "Approved\_Conversion" for a different age group.

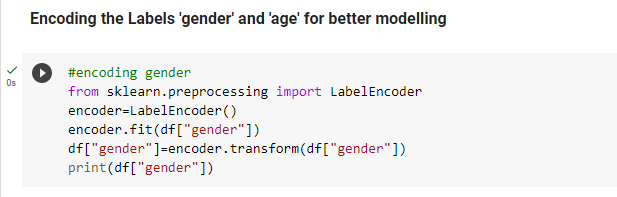
People in the 30-34 age group have a greater tendency to buy products after clicking the ad.

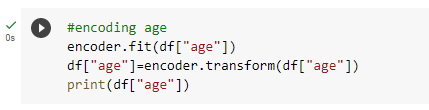


**Final Increment**

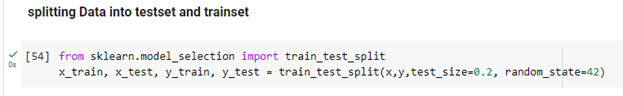
**Regression model and analysis**

Before doing the Regression model we are replacing **xyz\_campaign\_ids** withactual id’s and encoding the labels for age and gender.

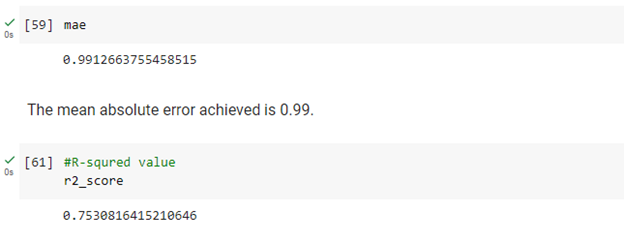
****

****

Splitting the data into train and test with proportion 80:20



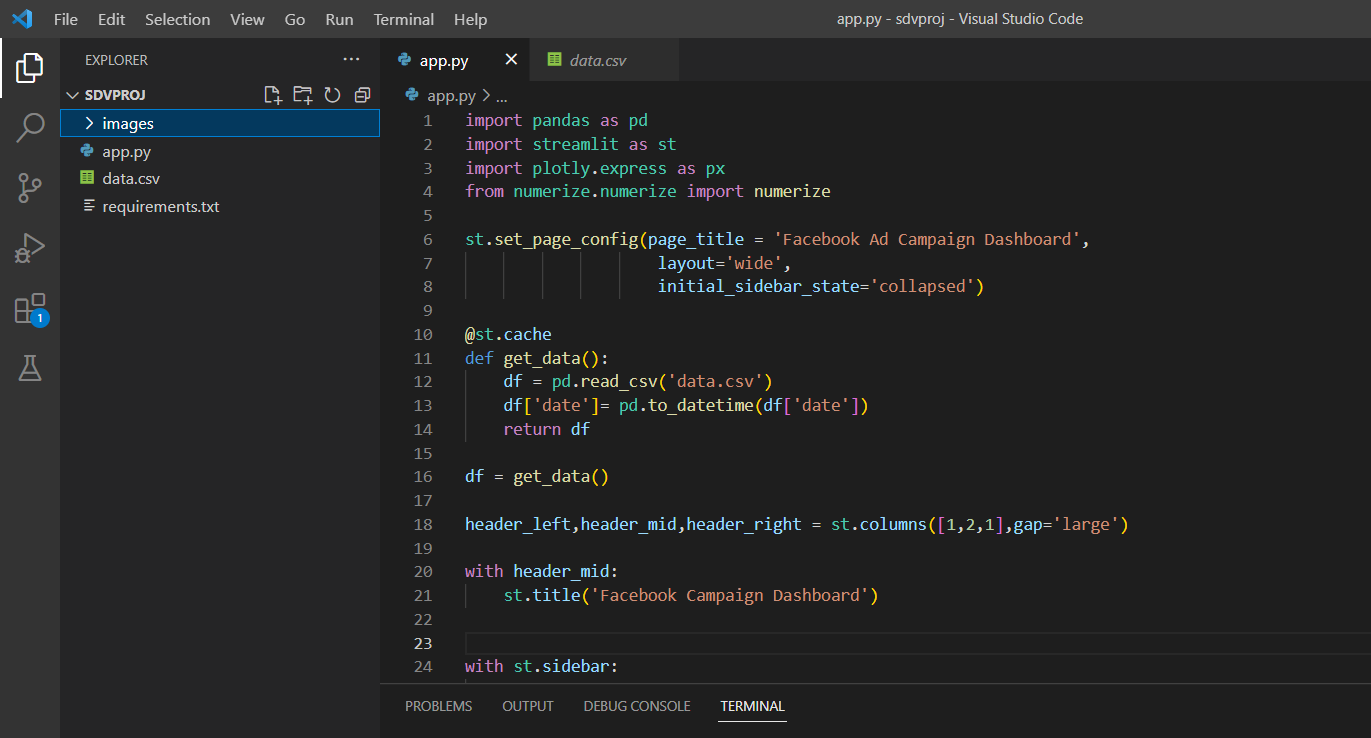
We are predicting Total conversion by using random forest regression. After the evaluation we got the mean absolute error is 0.99 and 75.3% of data fit into the regression model.



**Implementation**

* **Deployment of the Web App using Streamlit.**

1. Firstly, We have created a Folder on the Desktop by the name “sdvproj”.
2. We created a python file in VS-Code by the name “app.py”.
3. We have also created a Subfolder by the name “images”, which contains all the required icons that we want to represent on the webapp.
4. We also created a copy of our dataset, which is a .csv file.

****

This image showcases the file system and all the files that we have discussed above.

**CODE EXPLANATION**

Facebook Ad Campaign Dashboard is developed using the Python script. Streamlit framework and Plotly are used for data visualisation.

The first block of the code is related to the importing of the required libraries. The libraries used for this project that are required for creating the Facebook Ad Campaign Dashboard are pandas, streamlit, and plotly express. For formatting the numerical values, a numerize library (this library will format the numerical values) is imported.

The next block of the code is related to get\_data() function. Get\_data() will read the data file of Facebook Ad Campaign. The ‘date’ column is converted to a datetime format and the pandas dataframe results will be returned.

Get\_data() function is then called by the script and in a ‘df’ variable the data frame results are stored.

In the next section, the Streamlit framework is used for creating the layout of the Ad Campaign Dashboard. Layout, Page Title, and Initial Side Bar is set in this section. St.columns() function defines three header columns and the dashboard title is set in the middle column.

St.sidebar object creates the sidebar. This sidebar consists of filters such as the campaign, age group and gender group as well. These are multi-select filters.

Further, the ‘df’ dataframe is queried depending on the filters that are selected by the use of the .query() function.

Then by using the st.metric() function and five columns the total metrics are displayed. The .sum() function is used to calculate the total impressions, clicks, spent, conversions, and approved conversion of the filtered dataframe. Then it is formatted by using the numerize library.

St.plotly\_chart() function creates two sections of plots. The click-through rate(CTR) by campaign bar chart is shown in the first section of the plot. Daily impressions by the campaign are shown in the second plot in the form of a line chart. The bar chart and line chart are created using plotly express. Various update\_layout() and update\_xaxes() functions are used to customise the charts.

At last, another two sections of plot which shows the spend by gender and cost per conversation by age demographic are created. Similar to the first section of charts, these charts are also created by using plotly express and various update\_layout() and update\_xaxes() functions are used to customise the charts.

After developing the code. Now we can deploy the Web App locally by using this command

* “sdvproj> streamlit run app.py”

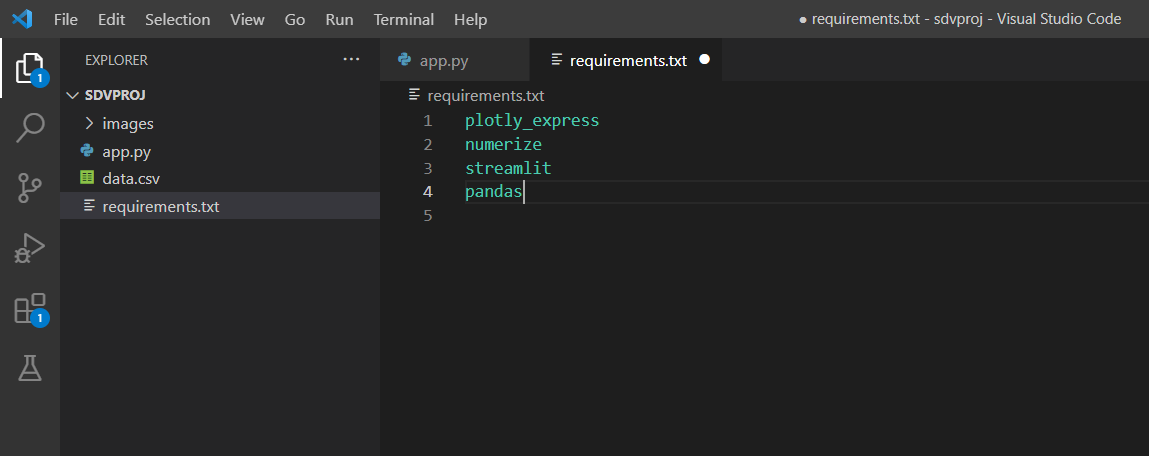
When we run this command in the terminal, The terminal automatically binds the application and opens the default web browser and deploys the application.

But, the limitation of this is the application is only visible locally on the machine it is run on, rather than the website being available globally. So, to solve this issue we need to use Streamlit Cloud to deploy our app globally.

**DEPLOYMENT TO THE PUBLIC CLOUD**

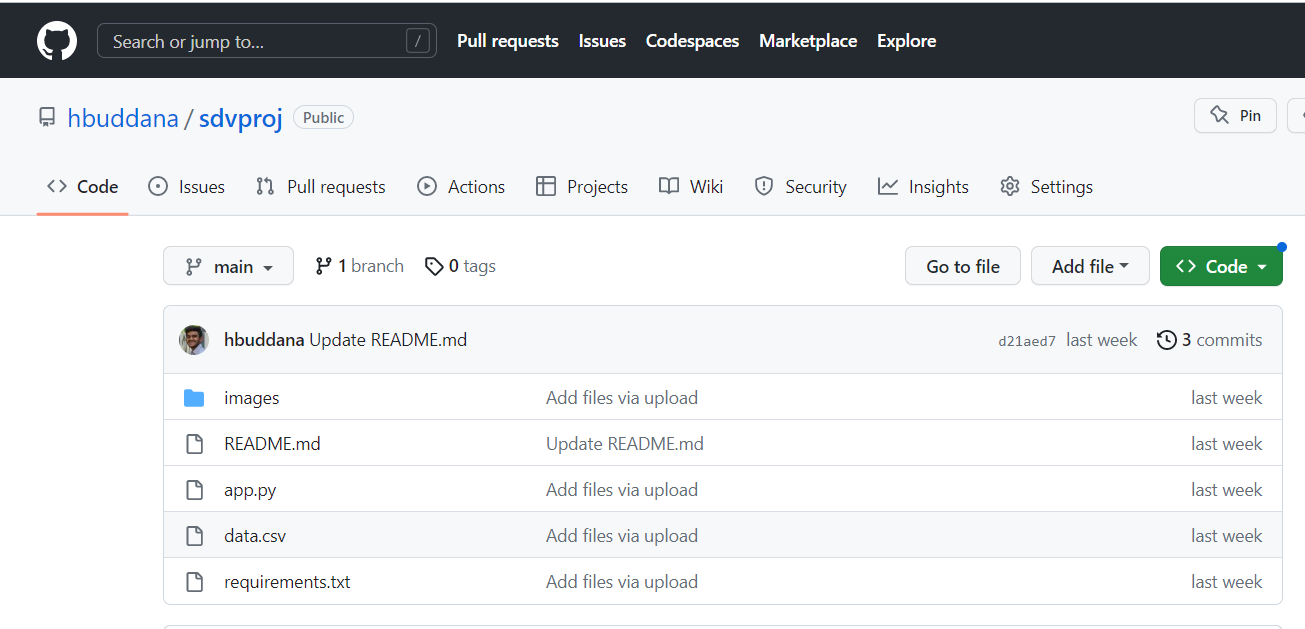
* **Step 1** - Requirements.txt

To deploy the web app globally, firstly we need to create a “requirements.txt” file in our folder and have all the imported libraries in our python code listed in it.



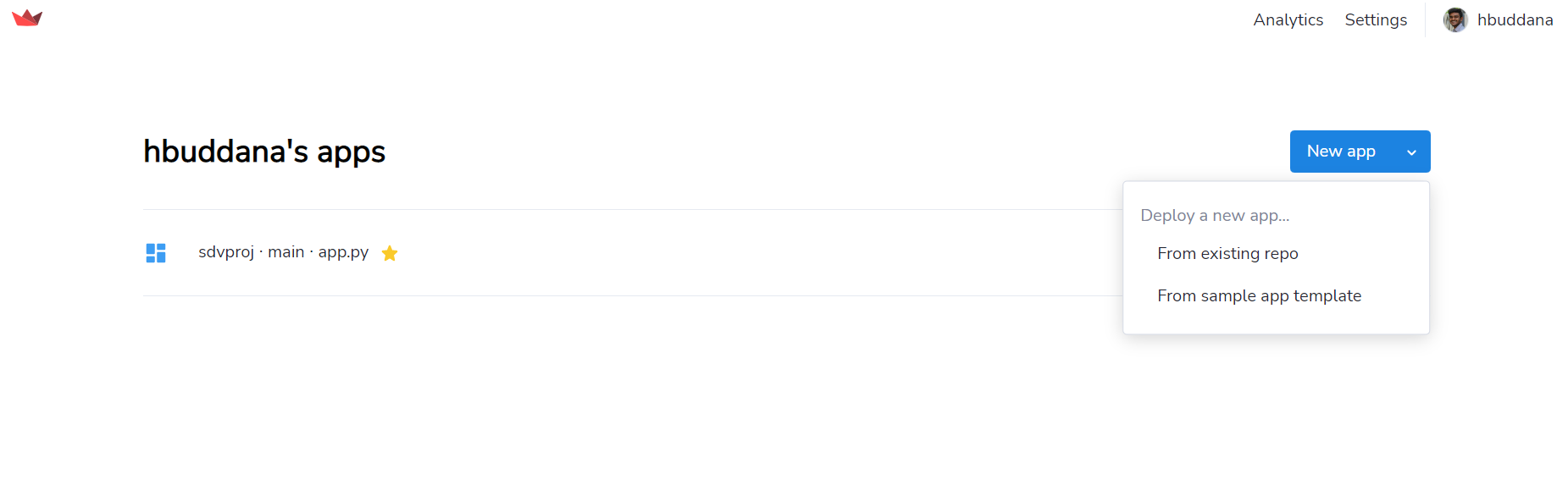
* **Step 2 -** Push the Code to GitHub

We have created a new repository by the name “sdvproj” and pushed the folder we have created on our desktop to this repository.



* **Step 3 -** Connect GitHub repo to Streamlit Cloud

Firstly, we need to create an account on [Streamlit Cloud](https://share.streamlit.io). After signing in, we can see the following screen.



We need to tap on > New app > Existing Repo and just give the repo name and the repo and Streamlit cloud are connected. I have already made that, hence we can see the “sdvproj-main-app.py” on the screen. So when we connect the repo to the cloud it does take a considerable amount of time to build the Global Website.



Deployed Web-app URL - <https://hbuddana-sdvproj-app-d9bhfj.streamlit.app>

**ANALYSIS**

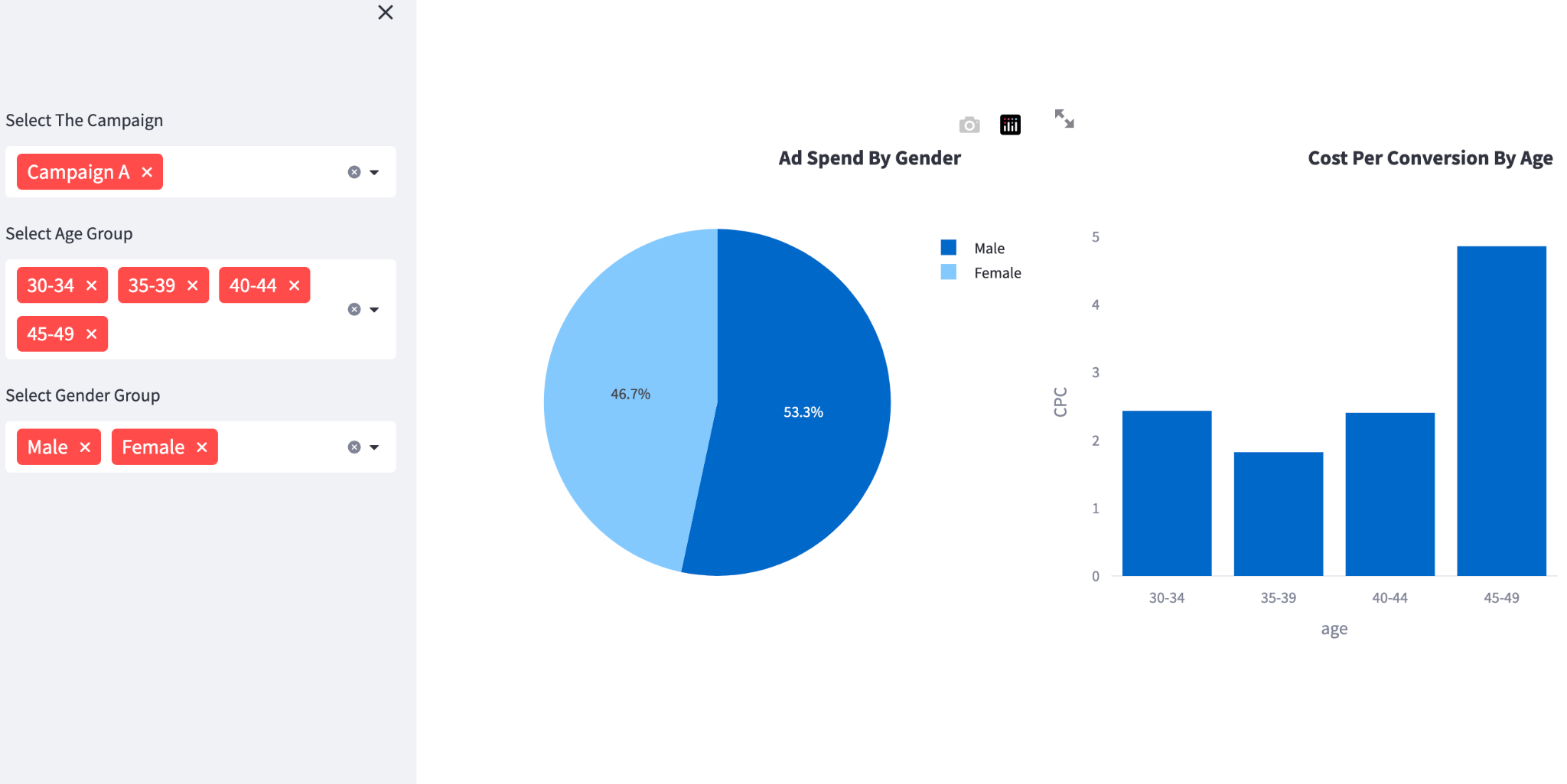
Firstly, we are going to analyse individual performance of each campaign before making mixed comparisons.

The analysis of campaign A is as follows:



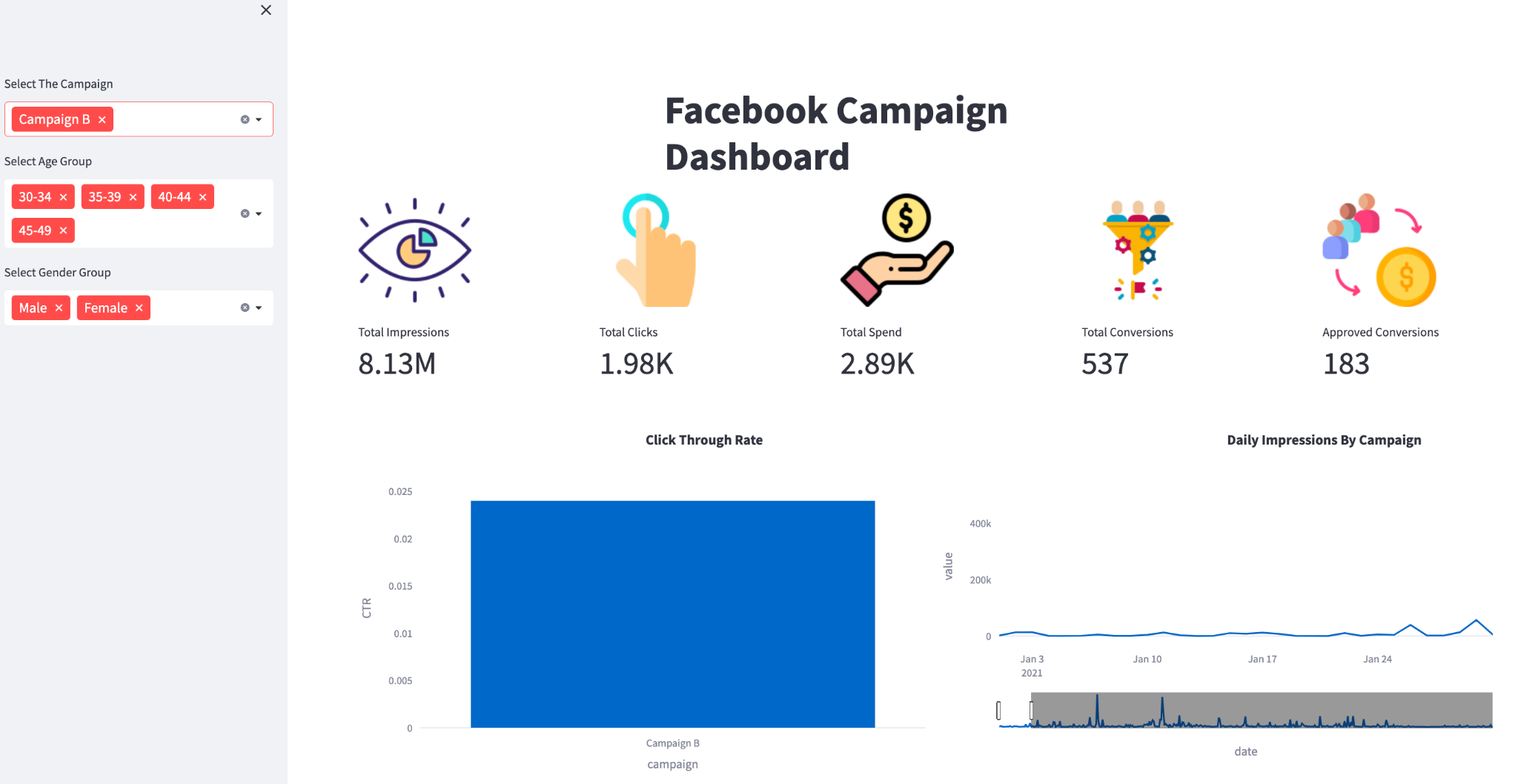
We can see that the total number of impressions for campaign A are 482.92K, the total number of clicks are 113, total spend is 149.71, total conversions are 58 and approved conversions are 24.

The click through rate for campaign A stands at 0.023. We can also see daily impressions by campaign where campaign A performs well in the month of January when compared to other months.



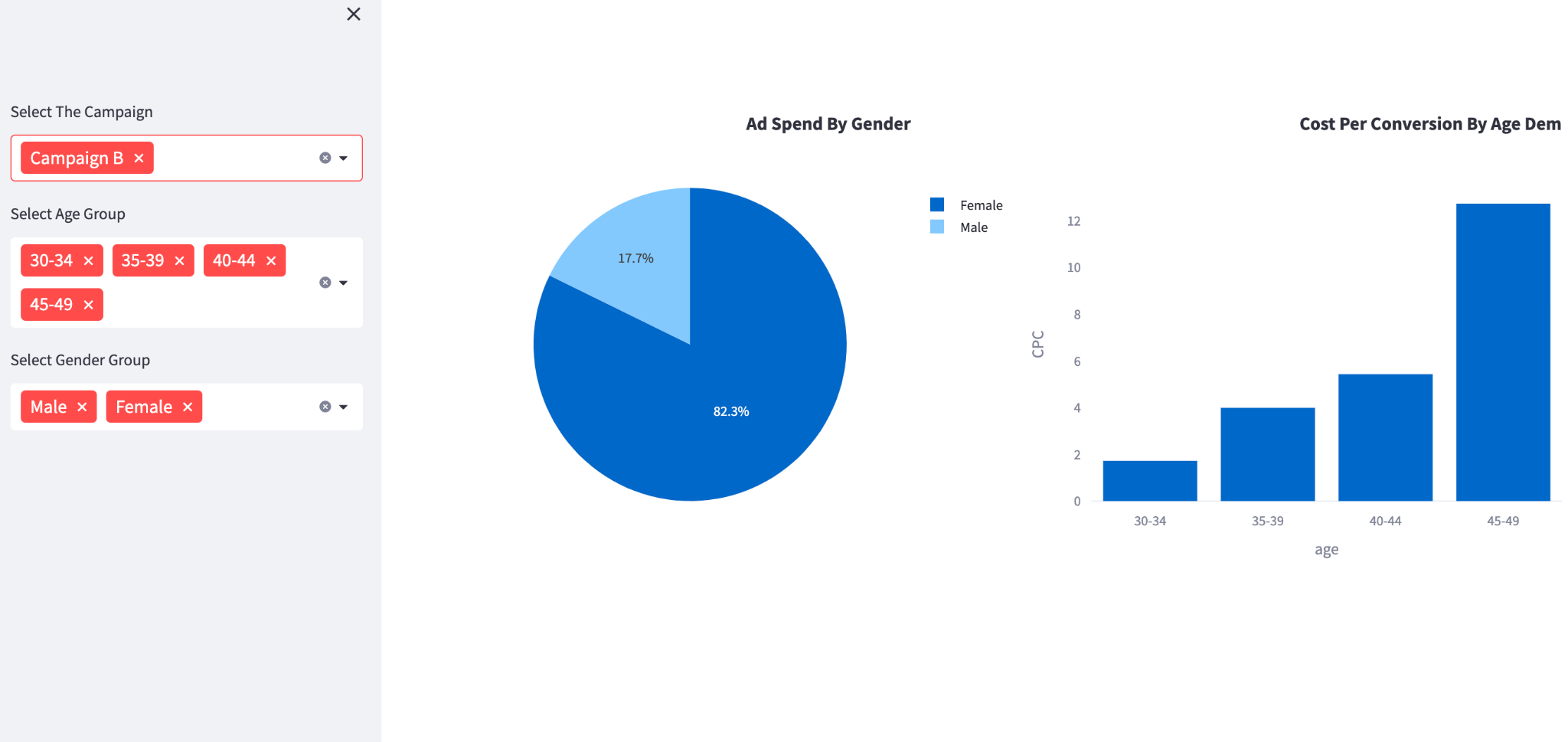
When coming to the ad spend which refers to the amount of money spent on advertising targeted to specific gender, campaign A has more ad spend of 53.3% towards males than females with an ad spend of 46.7%. Then we compare cost per conversion by age demographics where we can see that the age group of 45-49 have the highest cost per conversion rate of approx 5 and the age group of 35-39 has the lowest cost per conversion rate of approx 2 among all the age groups participated in the campaign.

The analysis of campaign B is as follows:



We can see that the total number of impressions for campaign B are 8.13 million, the total number of clicks are 1.98k, total spend is 2.89k, total conversions are 537 and approved conversions are 183.

The click through rate for campaign B stands at 0.024 almost same as campaign A. We can also see daily impressions by campaign where campaign B performs well in between months of March and July when compared to other months.



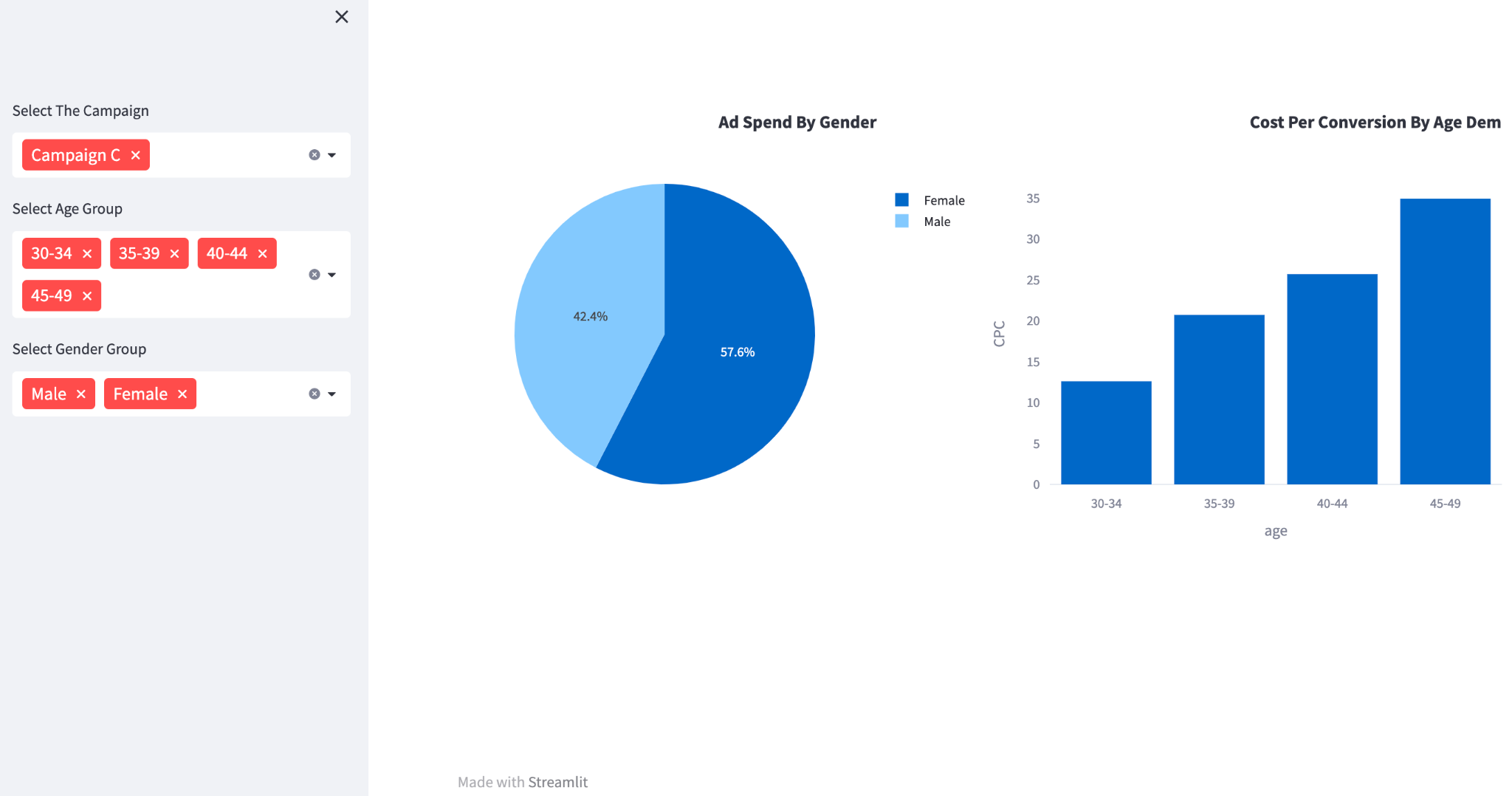
When coming to the ad spend which refers to amount of money spent on advertising targeted to specific gender, campaign B has more ad spend of 82.3% towards females than males with an ad spend of 17.7%. Then we compare cost per conversion by age demographics where we can see that the age group of 45-49 have the highest cost per conversion rate of approx 12 and the age group of 30-34 has the lowest cost per conversion rate of approx 2 among all the age groups participated in the campaign.

The analysis of campaign C is as follows:



We can see that the total number of impressions for campaign C are at a peak of 204.82 million, the total number of clicks are 36.07k, total spend is 55.66k, total conversions are 2.67k and approved conversions are 872.

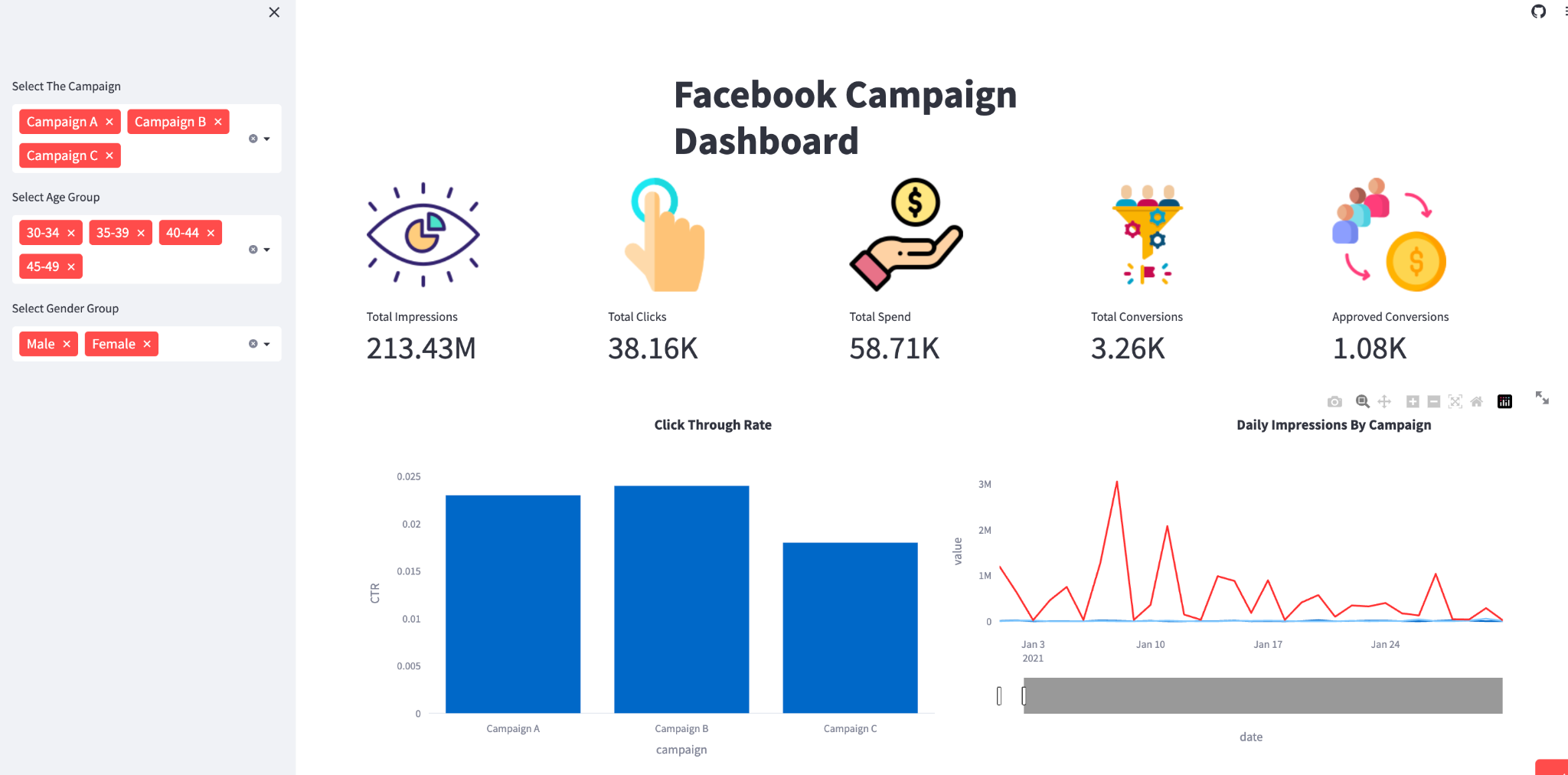
The click through rate for campaign C stands at 0.018 which is less than the other campaigns . We can also see daily impressions by campaign where campaign C performs well in the month of January when compared to other months.



When coming to the ad spend which refers to the amount of money spent on advertising targeted to specific gender, campaign C has more ad spend of 57.6% towards females than males with an ad spend of 42.4%. Then we compare cost per conversion by age demographics where we can see that the age group of 45-49 have the highest cost per conversion rate of approx 35 and the age group of 30-34 has the lowest cost per conversion rate of approx 14 among all the age groups participated in the campaign.

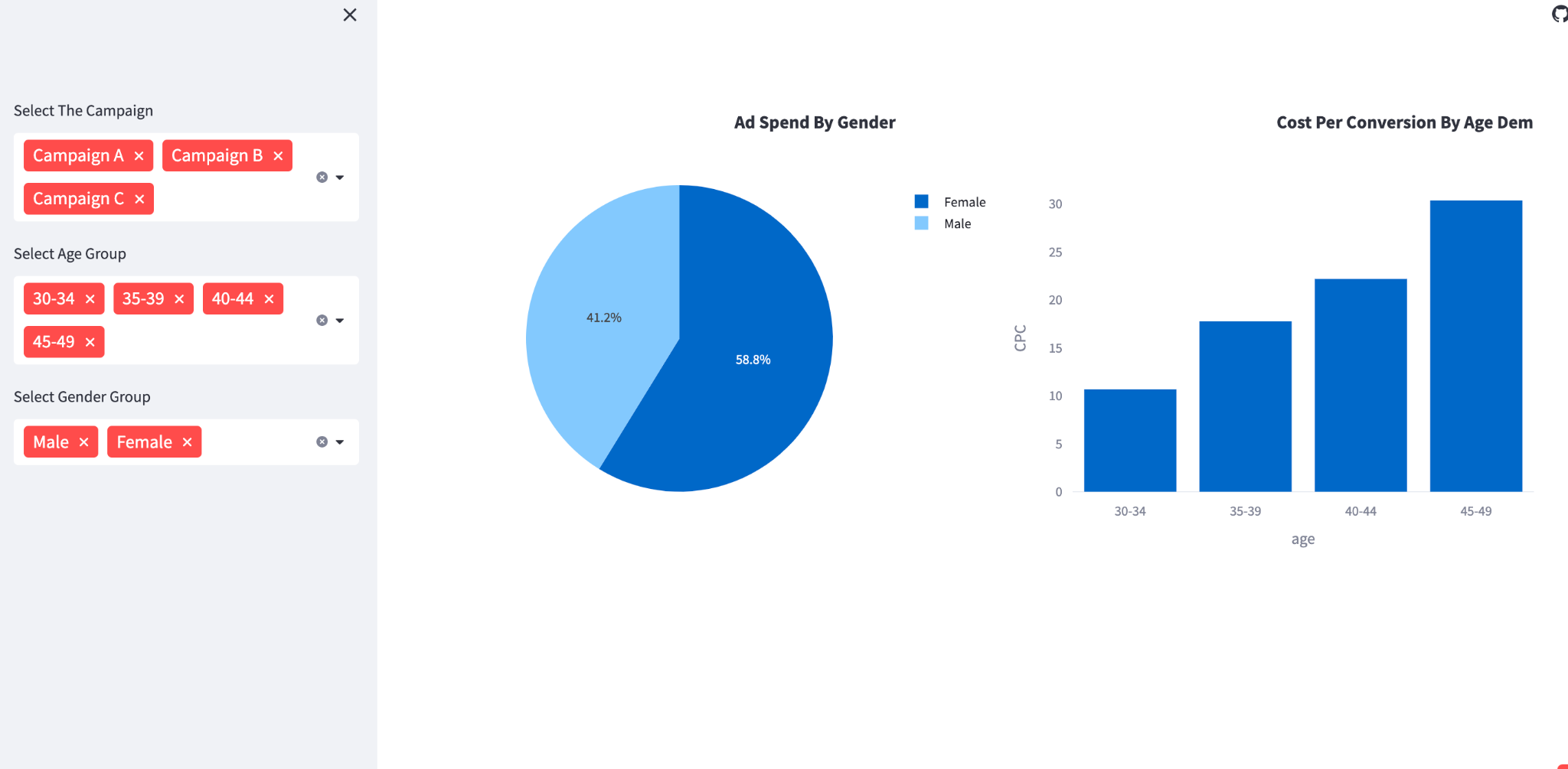
Of all the three campaigns, Campaign C stands top in terms of impressions, clicks, total spend, total conversion and Approved conversion.

Now we are going to compare all the three campaigns A, B and C which helps us to get a clear overview of the results.



We can see that the total number of impressions for all the three campaigns A,B and C are at a peak of 213.43 million, the total number of clicks are 38.16k, total spend is 58.71k, total conversions are 3.26kk and approved conversions are 1.08k.

While comparing the click through rate among the three campaigns, we can see that Campaign B has the highest click through rate of 0.024 and Campaign C has the lowest click through rate of approx 0.018. From this analysis, we can say that Campaign B is resonating with target audience and driving engagement.



When coming to the ad spend of all the three campaigns, the total ad spend for females stands at 58.8% than males with an ad spend of 41.2%. This shows us that the advertiser is trying to reach more female customers. Then we compare cost per conversion of all three campaigns by age demographics where we can see that the age group of 45-49 have the highest cost per conversion rate of approx 30 and the age group of 30-34 has the lowest cost per conversion rate of approx 11 among all the age groups participating in the campaign. This indicates that the targeting for that age group of 45-49 is not resonating well with the audience resulting in less engagement and conversion rates. On the other hand, the age group of 30-34 is resonating well with the audience, resulting in more engagement and conversion rates.

**Video Link** - [Final\_project.mp4](https://myunt-my.sharepoint.com/:v:/g/personal/harshavardhanabuddana_my_unt_edu/EfBbE_XMpwNIvUEEFTx0cUcBxQ8qj5lQm3pmxh5IusNaMw?e=5rf5Tv)

**• Project Management**

**Implementation status report**

**Work completed:**

**• Description**

Creating a Regression model and deployment of web app using streamlit.

**• Responsibility (Task, Person)**

Creating a Regression model - Nagasai Gummadi, Divya Sri Vakkala

Deployment of Web App - Harsha Buddana, Eswara Reddy Thimmapuram

Harsha Buddana - 25%

Nagasai Gummadi - 25%

Divya Sri Vakkala - 25%

Eswara Reddy Thimmapuram - 25%

**References/Bibliography:**

1. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.
2. VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.
3. Raschka, S., & Mirjalili, V. (2017). Python Machine Learning. Packt Publishing.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed.). Springer.
5. Subha, B. “Social Media Advertisement and Its Effect in Sales Prediction - An Analysis.” Shanlax International Journal of Management, vol. 8, no. 2, 2020, pp. 40–44.
6. Shi, Yuying and Karniouchina, Ekaterina and Uslay, Can, (When) Can Social Media Buzz Data Replace Traditional Surveys for Sales Forecasting? (April 1, 2020). Rutgers Business Review, Vol. 5, No. 1, 2020, pp.43-60.
7. ilad Dehghani, Mojtaba Khorram Niaki, Iman Ramezani, Rasoul Sali,Evaluating the influence of YouTube advertising for attraction of young customers,Computers in Human Behavior,Volume 59,2016,Pages 165-172,ISSN 0747-5632.
8. Jong Uk Kim, Woong Jin Kim, Sang Cheol Park,Consumer perceptions on web advertisements and motivation factors to purchase in the online shopping,Computers in Human Behavior,Volume 26, Issue 5,2010,Pages 1208-1222,ISSN 0747-5632.
9. Milad Dehghani, Mustafa Tumer,A research on effectiveness of Facebook advertising on enhancing purchase intention of consumers,Computers in Human Behavior,Volume 49,2015,Pages 597-600,ISSN 0747-5632.
10. Hung-Pin Shih,An empirical study on predicting user acceptance of e-shopping on the Web,Information & Management,Volume 41, Issue 3,2004,Pages 351-368,ISSN 0378-7206.
11. https://docs.streamlit.io