Social Media Analysis using Data Analytics methods

Team Members:

19BDS0159 (ADRIJA MUKHOPADHYAY)

19BDS0162 (PARITALA GOURI)

19BDS0165 (MAGADUM HRISHIKESH PANDURANG)

Report submitted for the

Final Project Review

Course Code: CSE3045

Predictive Analysis

Slot: A1 Slot

Professor: Dr. Ilanthenral Kandasamy

1. Introduction:

Social media, which is extensively used across the world, plays an important part in the everyday lives of the majority of people. Individuals can use social media platforms to discover and learn new information, exchange ideas, and engage with new individuals and organizations. It has transformed the way people live today and made communication much easier. It facilitates the sharing of user-generated material such as data, images, and videos. From social media sites, there is a wide range of high-speed, high-volume data generated on a daily basis. Greater accessibility of internet connections, improved software tools, sturdy PCs, and mobile devices may all be credited with the increasing availability and use of social media.

Today, social media platforms such as Facebook, Twitter, Instagram and WhatsApp have millions of users and have had a significant influence on people's lives. Not only have people shared photos and information, but trades and businesses have thrived as well. Businesses use social media to reach their potential customers easily. Furthermore, it is mostly utilized by educators and students which helps in contributing a lot for teaching and learning purposes.

Social media became a significant location to engage during a period of social distance and minimal contact with others. Internet memes have been utilized to find amusement and distraction from the epidemic through social networking platforms. Social media was also used to spread important information. One example would be during the second wave in India. During this time, people reached out for help and got help with the help of social media.

However, social isolation has led many people to adjust their lifestyles, putting a burden on their mental health. Many social media-based online counseling services were launched, and they immediately gained popularity since they could safely connect mental health professionals with those who needed them.

Social media is also a critical instrument for bringing about social change. Until the advent of social media, we had never had immediate, real-time communication. It has dismantled communication boundaries and enabled individuals to openly express themselves. Furthermore, social media has become a platform for debating social issues, exchanging ideas, scheduling meetings, and promoting causes.

The project aims to predict a person's overall social media consumption based on how much time they spend on social media sites like Whatsapp, Instagram, and Facebook. The algorithm will forecast if a user has a high or low engagement activity level based on their Instagram followers and amount of posts.

2. Literature Review Summary Table

Authors and Year (Reference)	*Title (Study)*	*Concept / Theoretical model/ Framework*	*Methodology used/ Implementation*	*Dataset details/ Analysis*	*Relevant Finding*	*Limitations/ Future Research/ Gaps identified*
[1] Kiran Chaudhary, Mansaf Alam, Mabrook S. Al-Rakhami and Abdu Gumaei (2021)	Machine learning-based mathematical modeling for prediction of social media consumer behavior using big data analytics	Big data technology was employed to handle and analyze data in order to forecast customer behavior on social media in this paper. Based on a set of factors and criteria, we looked at customer behavior on social media platforms. We looked at how people think about social media and how they feel about it.	They've suggested a mathematics and machine learning-based predictive model for determining customer behavior on social media platforms. Polynomial regression is used to create a random degree polynomial relationship between YouTube, Facebook, LinkedIn, and Twitter sets of data points using a linear predictor function.	They gathered consumer information from a variety of social media sites. Because the data they collected from social media networks was unclean.. A total of 5279 records are included in the dataset. 3962 are clean records. Agency, platform, URL, sampled date, and Likes/Followers/Visits/Downloads are	The biggest customer divergence from one social media to another is 99.51 percent, while the lowest is 12.22 percent. Among all, the biggest root mean square error is 156556.4529, while the lowest is 20691.787. All have a maximum accuracy of 0.9829 and a minimum accuracy of 0.0223\.	The model's shortcoming is that it will not operate with customer data collected on a daily basis. If this model is applied to everyday data, the results will be disastrous.

				the four attributes in the dataset.		
[2] Said A. Salloum , Nafla Mahdi Nasser AlAhbabi, Mohammed Habes ,Ahmad Aburayya , and Iman Akour (2021)	Predicting the Intention to Use Social Media Sites: A Hybrid SEM - Machine Learning Approach	The study's goal is to develop a conceptual model for calculating students' acceptance of social media in the classroom and the elements that influence it. The research is carried out by including perceived fun and social influence into the Technology Acceptance Model (TAM). In addition, the data is analyzed using Machine Learning (ML) techniques and partial least squares structural equation modeling (PLS-SEM)	This study's research instrument is divided into two parts. The first half is concerned with getting demographic data from participants, while the second part is tasked with gathering responses to components in the conceptual model. The items in the second part are measured using a 5-point Likert scale. PEOU and PU were measured using items derived from a prior study. Items taken from another study paper were used to measure "Perceived	Between September and October 2020, self-administered surveys were employed to collect data. The participants consented to do the survey on the condition that they would not be compensated. The data for this study is collected using a convenience sampling method. The poll was completed by 369 students out of 400, yielding a response rate of 92 percent. There were 170 boys and 199 females	The fundamental purpose of this study is to look at the factors that influence students' acceptance of social networks in colleges. To attain this purpose, TAM was extended by "Perceived Playfulness" and "Social Influence." PLS-SEM and machine learning methodologies were used to validate the proposed model, and 369 students from a well-known university in the UAE completed legitimate	Because the data was taken from a single private university in the UAE, the results may not be indicative of the general population of other higher educational institutions in the UAE. To address this restriction, further research on government students should be conducted in order to draw a comparison between government and private students in terms of the parameters explored.

			Playfulness" and "Social Influence."	among the students. 63 percent of the participants were between the ages of 18 and 29. Participants with a bachelor's degree made up 61%, those with a master's degree made up 23%, students with a Ph.D. made up 10%, and those with a diploma made up the remaining 6%.	questionnaire surveys. According to the findings of this study, "perceived playfulness," "social influence," "perceived usefulness," and "perceived ease of use" have a substantial impact on students' intention to use social media networks for learning. These findings highlighted the importance of students' potential and dependence on social networks for educational objectives, corroborating the findings of prior social network	

					acceptance studies.	
[3] Kristo Radion Purba, David Asirvatham, and Raja Kumar Murugesan (2020)	Instagram Post Popularity Trend Analysis and Prediction using Hashtag, Image Assessment, and User History Features	Using a global dataset, this study assessed multiple regression models for predicting the Engagement Rate (ER) of postings. In comparison to previous research, the prediction model, when combined with the results of the popularity trend study, will be more useful to a bigger audience. Hashtags, picture analysis, and user history were used to extract the features.	There were four phases in this research, i.e., data collection, data filtration, analysis, and popularity prediction.In the feature extraction phase Hashtag features,Post features ,Image assessment features and User Histroy features were extracted.The output of the model was to predict the engagement rate using Linear Regression (LR), Random Forest Regressor (RF) and Support Vector Regressor (SVR) models.	Top-Hashtags provided 2,000 top hashtags, which were used to begin data gathering. The Instagram Application Programming Interface was used to collect posts from the hashtags list (API). The data was gathered in two stages. The first phase was utilized to get all of the posts listed under a hashtag that was popular at the time. Videos and posts that were more than 30 days old were removed. The	EG is utilized as a comparator to account for the lower ER of users with more followers. Because ER is more readable, it was used as the output in the prediction. Image quality, posting day and time, user tags, and image kind were shown to be the most relevant factors in increasing EG. The user's past data is crucial when it comes to anticipating (or forecasting) ER. In the global dataset, the	To eliminate subjectivity, the manual assessment values in this study could be replaced with similar automated values in future research. Other aspects, such as user history, can still be tweaked to improve outcomes. To discriminate between popular and less popular postings, text analysis tools such as sentiment analysis and concept semantic similarity can be added

				second data collection period was used to collect data from those posts that were uploaded exactly 30 days after the first. As a result, a scraper programme was created to check the lifetime of each post available from period 1 on a regular basis, and then re-scrape the data of a post once it had been 30 days since it was posted.	history was used to reduce the substantial variability of ERs between users. With all features, prediction accuracy can reach 73.1 percent, and 64.8 percent without manual picture assessment, according to R.	
[4] B. Senthil Arasu,B Jonath Backia Seelan, N.Tham	A machine learning-based approach to enhancing social media	The goal of this aper is to look at ocial media data analytics using achine learning chnologies. The Waikato Environment for Knowledge	ML integrated social media marketing (ML-SMM) is our proposed approach. The steps of the process involved in the	Not much information about the dataset was mentioned in the research paper by the authors.	The proposed work on ML-SMM mechanisms overs the ideas f social media arketing and achine learning, as well as	his tool can also be used to investigate the goals, quirements, and preferences of MM campaigns other business fields such as

araiselvan (2020)	marketing	ıalysis is used in ıis novel method) design a social ıedia marketing rategy (WEKA). WEKA is ɔmpared to other algorithms of terest and shown to outperform ıem, particularly in terms of ɪrecision, recall, ınd F-measure, ɛmonstrating that 'EKA is superior other techniques.	proposed ML-SMM approach are as follows: (i) Text mining, (ii) Machine learning integrated with social media marketing, and (iii) ML-SMM analysis using WEKA.		including the 'EKA machine ɛarning tool to orecast online consumer behavior for effective ıarketing. The VEKA tool is ısed to collect and analyze ımple datasets, ıd the findings reveal that it erforms better ıan other tools. IL-SMM with WEKA outperforms other tools in rms of applying ırious types of mining techniques, business ɪplications, and data analysis ıethodologies, espite the fact hat numerous tools are ailable for this task. This mbination also improves reporting capabilities, ıich overcomes	ıline education, ealth care, and music.

					limits of other technologies.	
[5] M. Savci, A. Tekin, and J. D. Elhai (2020)	Prediction of problematic social media use (PSU) using machine learning approaches.	Artificial neural networks (ANN) and support vector machines were used to represent problematic social media usage (PSU) in this study (SVM). In order to predict PSU, fifteen predictor factors were investigated. Then they utilized forward selection processes to figure out which variables were the most important. They also discovered that the five most relevant characteristics in relation with PSU severity were frequency of daily social media use, frequency of	PSU modeling, ANN, and SVM were used in this work. The data was analyzed using ANN and SVM. Machine learning prediction models of data mining include ANN and SVM. First, the results of ANN and SVM with fifteen predictor variables are reported in this part. Then, as a subset selection approach for picking predictors with a stopping criteria, forward selection analysis was used. Finally, using the five most relevant predictor variables	The participants were university students from various departments. Participants were recruited using a convenience sample strategy in this study. All data was gathered from Firat University students (Turkey). After eliminating 23 participants due to missing or inaccurate data, the data set included 309 university students (208 females and 101 males) who had used social media	Prediction was done using k-folds (k = 5)cross validation in both ANN and SVM. The data generated by the ANN model and the real test data had a 0.62 correlation. The same prediction was then made using SVM, which revealed a 0.63 correlation between the data generated by SVM and the real test data. They discovered that a key predictor of PSU severity was a desire to be liked. They also	Their studies show that the features should be addressed in clinical contexts. However, this study relied on self-report measuring instruments, which is a drawback. Furthermore, a convenience sample strategy was used to obtain data. Finally, there were more women in the sample than men. Instead of self-report scales, measuring methods that give more objective data can be employed in future investigations.

		monitoring social media accounts, want to be liked, exhibitionism, and FOMO. Finally, the article looked at how well these five factors predicted PSU, discovering that the five-variable estimate had a better coefficient of estimation than the fifteen-variable estimate.	established by forward selection, the results of ANN and SVM were computed. The output variable, PSU, was predicted using our 15 predictor factors in this study. The analysis was carried out using data from 309 pupils	for at least one year and had at least one social media account.	discovered that FOMO was a strong predictor of PSU severity, confirming previous findings. FOMO is an intra-personal quality that motivates people to keep up with what other people are doing, particularly on social media platforms.	

3. Objective of the project:

The main purpose of our project is to predict the total social media usage of a particular individual based on the time spent on social media in various platforms such as whatsapp, instagram and facebook. Our aim is to predict which platform is being used the most by an individual based on time spent on weekends versus time spent on week-days. Using instagram followers and number of posts, the algorithm will predict whether the user has a high or low interaction activity status.

4. Innovation component in the project:

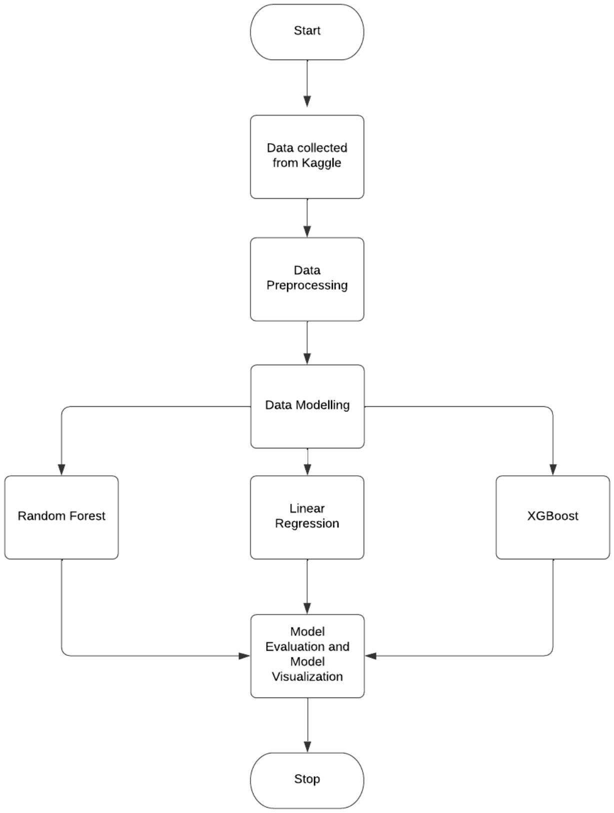
The innovative part of our project is that we are using the parameters like number of followers, number of posts and time spent by an individual on social media platforms like whatsapp, facebook and instagram to predict the social media usage of an individual as well as to predict which platform is used the most by people based on the time spent on each platform.

5. Work done and implementation

a. Methodology adapted:

- The data was collected from the Kaggle website. The dataset was called Social Media Usage analytics.
- Then the data-preprocessing starts.
- The data set we used, we first checked for the presence of any null values. Our data set did not include any null values.
- The column names were renamed for convenience
- Unwanted or irrelevant columns which had practically 0 contribution to our prediction were dropped.
- The numerical and string data types were in object format and we had to convert them into string and int data types respectively because mathematical functions cannot be applied to object data types.
- The last but the most important part of the preprocessing was converting string numeric values to integer values. Columns such as "Total Instagram Usage " had values stored as a string so we had to process it by removing the ',' and then converting it from Object to String and then finally to Integer.
- We also performed a few basic visualizations on the dataset.
- Machine Learning models used are; Random Forest, Linear Regression and XGBoost.
- Linear regression model was used to predict the total social media usage based on the time spent on each platform "last week". The train:test ratio was 70:30.
- We used cross- validation to overcome the overfitting issue in the model.
- Linear Regression model was used to predict the total social media usage based on the time spent on each platform "last weekend". The train:test ratio was 70:30.
- XGBoost model is used to improve the performance for total social media usage in the last week for all three platforms. The train:test ratio was 70:30.
- Random Forest model was used to predict which platform is used the most by an individual based on the time spent on week and time spent on weekends. The test size used is 25 percent.
- Linear regression model was used to predict the activity status of an individual based on number of instagram posts and number of instagram followers.
- Finally, model evaluation such as the mean absolute error, r-square score, mean standard error, the coefficient matrix and model visualization was performed for the models used..

Here is the flowchart of the methodology;



b. Dataset used:

The data set was collected from the Kaggle website. The dataset was named as - Social Media Usage Analysis. The dataset contains 26 variables (columns) and 1628 records (rows). The datatypes in the dataset were of object data type initially, which we preprocessed to fit into the models and to get a better accuracy performance of the models used.

Below is the following description of each variable in the dataset.

Column Name	***Dataset ColumnDescription***
Age	Age of the individual.
Current Status	Is he a working professional or a student?
Do you own multiple profiles on Instagram?	Does the individual own multiple profiles on instagram.

Gender	What's the gender of an individual?
Highest Education	What is the education qualification of the individual ?
Location (City Airport Code)	What's the location of an individual?
Phone OS	What type of operating system does the individual mobile use ?
State	Which state is the individual from ?
Zone	Which zone does the state belong to that he is from ?
How many followers do you have on Instagram?	How many followers does an individual have ?
How many posts do you have on Instagram?	How many posts does an individual have in an account ?
Latitude	The latitude coordinates of where the individual lives.
Longitude	The longitude coordinates of where the individual lives.
Time Spent on Facebook last week (in minutes)	How many minutes did an individual spend on facebook last week ?
Time Spent on Instagram in Last week (in minutes)	How many minutes did an individual spend on instagram last week ?
Time Spent on Instagram last weekend (in minutes)	How many minutes did an individual spend on instagram last weekend ?
Time Spent on WhatsApp in Last week (in minutes)	How many minutes did an individual spend on whatsApp last week ?
Total Facebook Usage	Total individual facebook usage.
Total Instagram Usage	Total individual instagram usage.
Total Social Media Usage	Total individual Social Media usage of whats app , instagram , facebook.
Total Week Usage	Total individual Social Media usage per week.

Total Weekend Usage	Total individual Social Media usage per weekend.
Total WhatsApp Usage	Total individual WhatsApp usage.

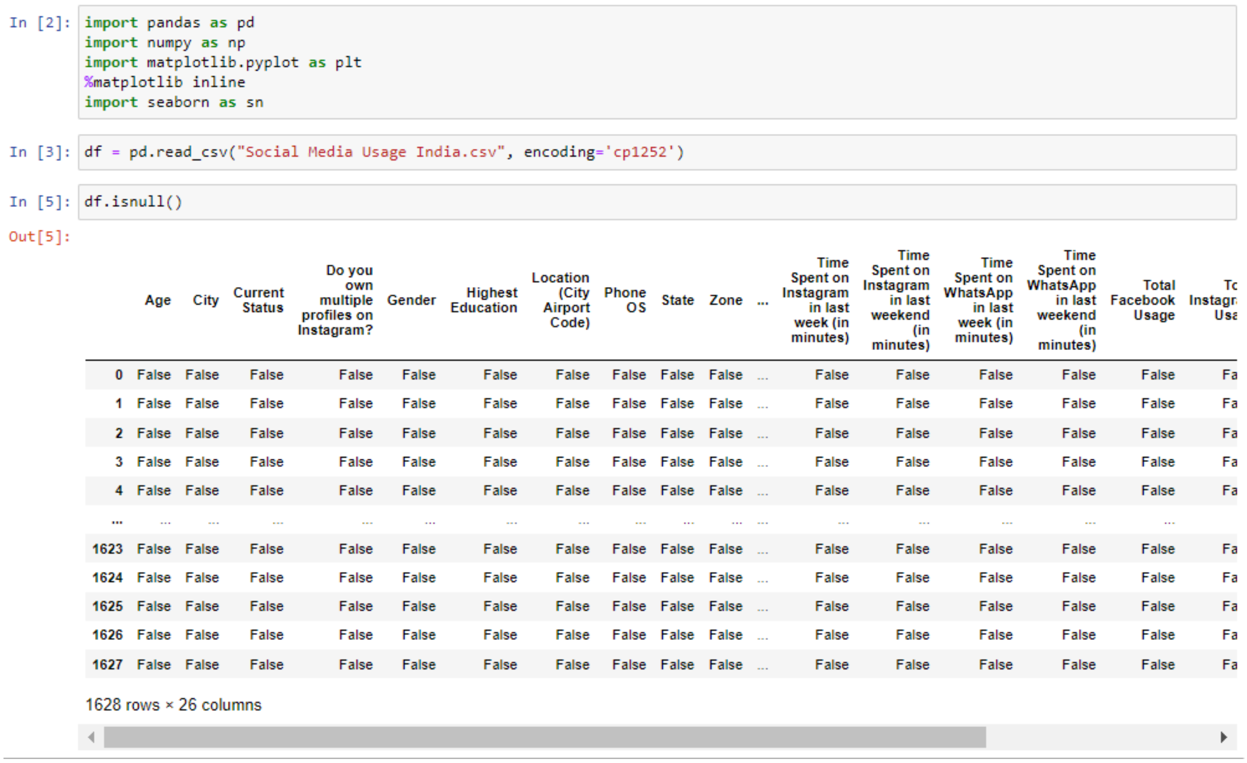
c. Tools used:

- Jupyter Notebook
- Python Programming Language..
- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scikit-Learn

d. Screenshot and Demo along with Visualization: (Preprocessing)

The preprocessing steps involves-

- Cleaning
- Instance selection
- Normalization
- One hot encoding
- Transformation
- Feature extraction
- Selection



In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sn
```

In [3]:
```
df = pd.read_csv("Social Media Usage India.csv", encoding='cp1252')
```

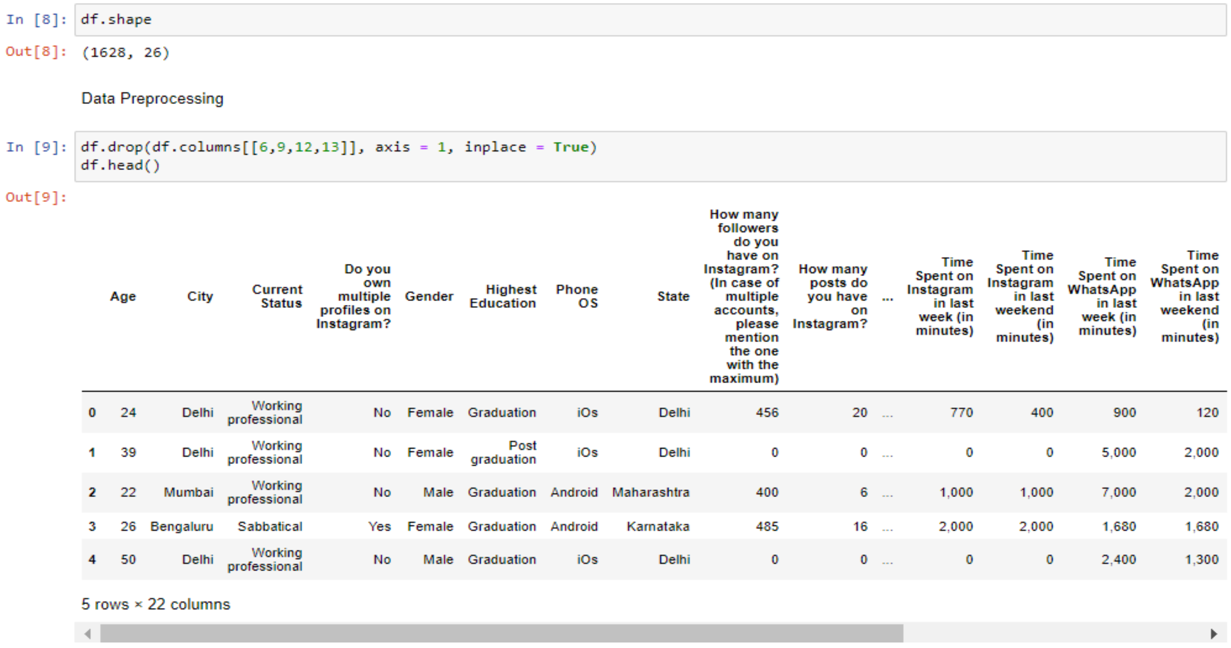
In [5]:
```
df.isnull()
```

Out[5]:

	Age	City	Current Status	Do you own multiple profiles on Instagram?	Gender	Highest Education	Location (City Airport Code)	Phone OS	State	Zone	...	Time Spent on Instagram in last week (in minutes)	Time Spent on Instagram in last weekend (in minutes)	Time Spent on WhatsApp in last week (in minutes)	Time Spent on WhatsApp in last weekend (in minutes)	Total Facebook Usage	Tc Instagr Usa
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1623	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
1624	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
1625	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
1626	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa
1627	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	Fa

1628 rows × 26 columns

Importing the necessary libraries and reading the data into the notebook.



In [8]:
```
df.shape
```

Out[8]:
```
(1628, 26)
```

Data Preprocessing

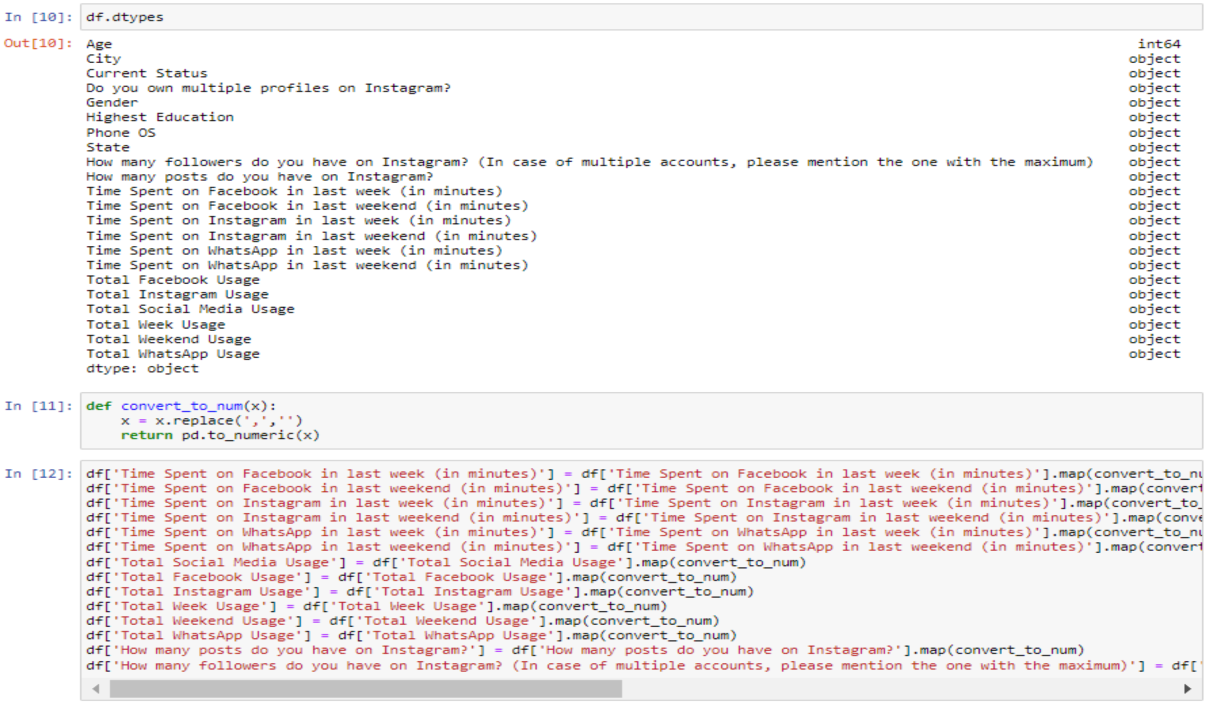
In [9]:
```
df.drop(df.columns[[6,9,12,13]], axis = 1, inplace = True)
df.head()
```

Out[9]:

	Age	City	Current Status	Do you own multiple profiles on Instagram?	Gender	Highest Education	Phone OS	State	How many followers do you have on Instagram? (In case of multiple accounts, please mention the one with the maximum)	How many posts do you have on Instagram?	...	Time Spent on Instagram in last week (in minutes)	Time Spent on Instagram in last weekend (in minutes)	Time Spent on WhatsApp in last week (in minutes)	Time Spent on WhatsApp in last weekend (in minutes)
0	24	Delhi	Working professional	No	Female	Graduation	iOs	Delhi	456	20	...	770	400	900	120
1	39	Delhi	Working professional	No	Female	Post graduation	iOs	Delhi	0	0	...	0	0	5,000	2,000
2	22	Mumbai	Working professional	No	Male	Graduation	Android	Maharashtra	400	6	...	1,000	1,000	7,000	2,000
3	26	Bengaluru	Sabbatical	Yes	Female	Graduation	Android	Karnataka	485	16	...	2,000	2,000	1,680	1,680
4	50	Delhi	Working professional	No	Male	Graduation	iOs	Delhi	0	0	...	0	0	2,400	1,300

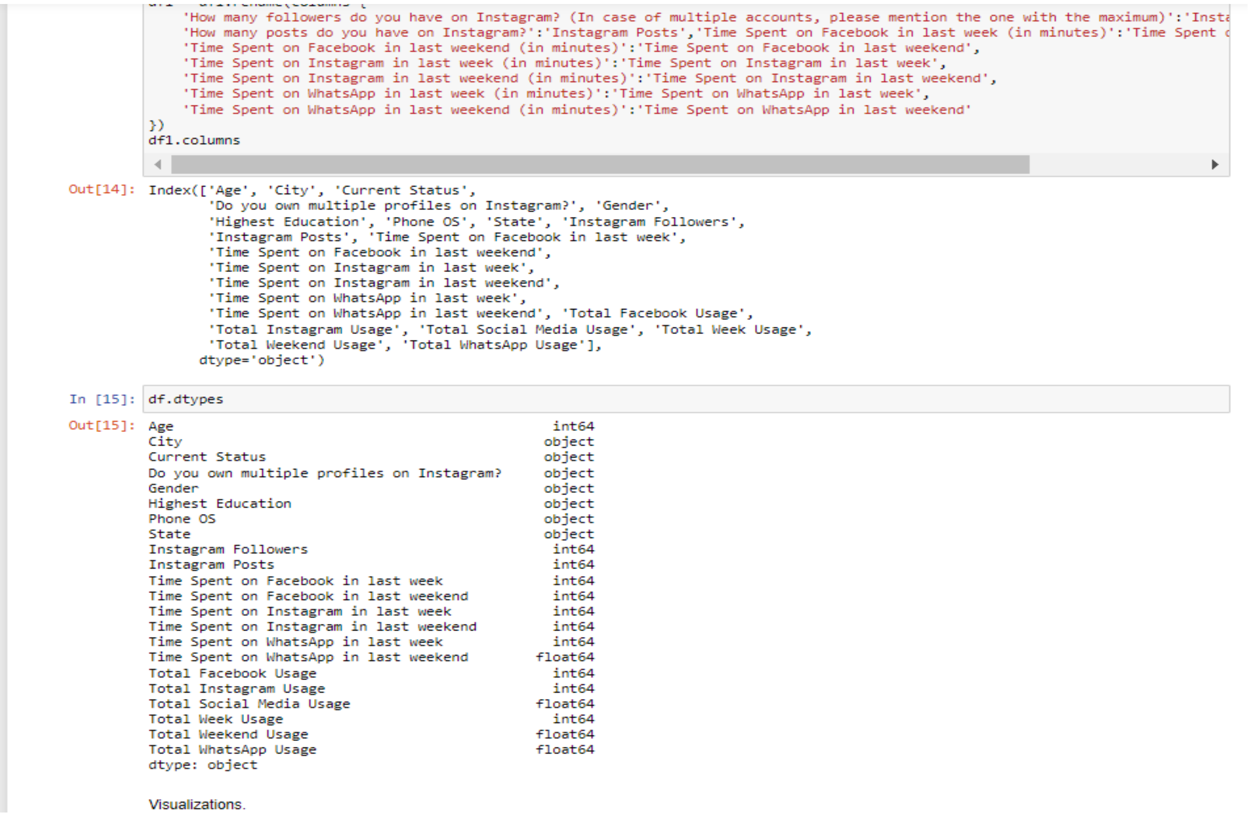
5 rows × 22 columns

Dropping unnecessary columns



```
In [10]: df.dtypes

Out[10]: Age                                                                                                         int64
         City                                                                                                       object
         Current Status                                                                                             object
         Do you own multiple profiles on Instagram?                                                                 object
         Gender                                                                                                     object
         Highest Education                                                                                          object
         Phone OS                                                                                                   object
         State                                                                                                      object
         How many followers do you have on Instagram? (In case of multiple accounts, please mention the one with the maximum)   object
         How many posts do you have on Instagram?                                                                   object
         Time Spent on Facebook in last week (in minutes)                                                           object
         Time Spent on Facebook in last weekend (in minutes)                                                        object
         Time Spent on Instagram in last week (in minutes)                                                          object
         Time Spent on Instagram in last weekend (in minutes)                                                       object
         Time Spent on WhatsApp in last week (in minutes)                                                           object
         Time Spent on WhatsApp in last weekend (in minutes)                                                        object
         Total Facebook Usage                                                                                       object
         Total Instagram Usage                                                                                      object
         Total Social Media Usage                                                                                   object
         Total Week Usage                                                                                           object
         Total Weekend Usage                                                                                        object
         Total WhatsApp Usage                                                                                       object
         dtype: object

In [11]: def convert_to_num(x):
             x = x.replace(',','')
             return pd.to_numeric(x)

In [12]: df['Time Spent on Facebook in last week (in minutes)'] = df['Time Spent on Facebook in last week (in minutes)'].map(convert_to_nu
         df['Time Spent on Facebook in last weekend (in minutes)'] = df['Time Spent on Facebook in last weekend (in minutes)'].map(convert
         df['Time Spent on Instagram in last week (in minutes)'] = df['Time Spent on Instagram in last week (in minutes)'].map(convert_to_
         df['Time Spent on Instagram in last weekend (in minutes)'] = df['Time Spent on Instagram in last weekend (in minutes)'].map(conve
         df['Time Spent on WhatsApp in last week (in minutes)'] = df['Time Spent on WhatsApp in last week (in minutes)'].map(convert_to_nu
         df['Time Spent on WhatsApp in last weekend (in minutes)'] = df['Time Spent on WhatsApp in last weekend (in minutes)'].map(convert
         df['Total Social Media Usage'] = df['Total Social Media Usage'].map(convert_to_num)
         df['Total Facebook Usage'] = df['Total Facebook Usage'].map(convert_to_num)
         df['Total Instagram Usage'] = df['Total Instagram Usage'].map(convert_to_num)
         df['Total Week Usage'] = df['Total Week Usage'].map(convert_to_num)
         df['Total Weekend Usage'] = df['Total Weekend Usage'].map(convert_to_num)
         df['Total WhatsApp Usage'] = df['Total WhatsApp Usage'].map(convert_to_num)
         df['How many posts do you have on Instagram?'] = df['How many posts do you have on Instagram?'].map(convert_to_num)
         df['How many followers do you have on Instagram? (In case of multiple accounts, please mention the one with the maximum)'] = df['
```

Renaming columns and converting columns into appropriate data types



```
    'How many followers do you have on Instagram? (In case of multiple accounts, please mention the one with the maximum)':'Insta
    'How many posts do you have on Instagram?':'Instagram Posts','Time Spent on Facebook in last week (in minutes)':'Time Spent c
    'Time Spent on Facebook in last weekend (in minutes)':'Time Spent on Facebook in last weekend',
    'Time Spent on Instagram in last week (in minutes)':'Time Spent on Instagram in last week',
    'Time Spent on Instagram in last weekend (in minutes)':'Time Spent on Instagram in last weekend',
    'Time Spent on WhatsApp in last week (in minutes)':'Time Spent on WhatsApp in last week',
    'Time Spent on WhatsApp in last weekend (in minutes)':'Time Spent on WhatsApp in last weekend'
})
df1.columns

Out[14]: Index(['Age', 'City', 'Current Status',
                'Do you own multiple profiles on Instagram?', 'Gender',
                'Highest Education', 'Phone OS', 'State', 'Instagram Followers',
                'Instagram Posts', 'Time Spent on Facebook in last week',
                'Time Spent on Facebook in last weekend',
                'Time Spent on Instagram in last week',
                'Time Spent on Instagram in last weekend',
                'Time Spent on WhatsApp in last week',
                'Time Spent on WhatsApp in last weekend', 'Total Facebook Usage',
                'Total Instagram Usage', 'Total Social Media Usage', 'Total Week Usage',
                'Total Weekend Usage', 'Total WhatsApp Usage'],
               dtype='object')

In [15]: df.dtypes

Out[15]: Age                                              int64
         City                                            object
         Current Status                                  object
         Do you own multiple profiles on Instagram?      object
         Gender                                          object
         Highest Education                               object
         Phone OS                                        object
         State                                           object
         Instagram Followers                              int64
         Instagram Posts                                  int64
         Time Spent on Facebook in last week              int64
         Time Spent on Facebook in last weekend           int64
         Time Spent on Instagram in last week             int64
         Time Spent on Instagram in last weekend          int64
         Time Spent on WhatsApp in last week              int64
         Time Spent on WhatsApp in last weekend         float64
         Total Facebook Usage                             int64
         Total Instagram Usage                            int64
         Total Social Media Usage                       float64
         Total Week Usage                                 int64
         Total Weekend Usage                            float64
         Total WhatsApp Usage                           float64
         dtype: object
```

Visualizations.

Final preprocessed data.

e. Models used:

Random Forest:

Random Forest is a well-known machine learning algorithm that uses the supervised learning method. In machine learning, it may be utilized for both classification and regression issues. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complicated issue and increase the model's performance. Ensemble employs two sorts of techniques. One is Bagging, which produces a distinct training subset with replacement from sample training data, and the final output is based on majority vote. Take, for instance, Random Forest. The other one is Boosting, which turns weak learners into good students by generating sequential models with the maximum accuracy possible. For instance, ADABOOST and XG-BOOST.

The random forest method has the following steps:

- n random records are chosen at random from a data collection of k records.
- For each sample, individual decision trees are built.
- Each decision tree produces a result.
- for classification and regression, the final output is based on Majority Voting or Averaging, accordingly.

Linear Regression:

Linear regression analysis is a statistical technique for predicting the value of one variable based on the value of another. The dependent variable is the variable you wish to forecast. The independent variable is the one you're using to forecast the value of the other variable. Linear regression creates a straight line or surface that reduces the difference between expected and actual output values. Simple linear regression calculators that employ the "least squares" approach to get the best-fit line for a set of paired data are available. You then use Y to calculate the value of X (dependent variable) (independent variable).A classic slope-intercept form is used to generate best-fit line linear regression.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

Here; y is a Dependent Variable, x is Independent Variable, a0 is intercept of the line, a1 is Linear regression coefficient. The cost function assists in determining the optimal values for a0 and a1, resulting in the best fit line for the data points. The accuracy of the mapping function that translates the input variable to the output variable is determined using the cost function. The Hypothesis function is another name for this mapping function. The Mean Squared Error (MSE) cost function is used in Linear Regression, which is the average of the squared error that occurred between the predicted and actual values.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Advantages of Linear Regression:

- For linearly separable data, linear regression works remarkably well.
- It's easier to implement, evaluate, and train with.
- It uses dimensionality reduction methods, regularization, and cross-validation to effectively tackle overfitting.
- Extrapolation beyond a certain data collection is another benefit.

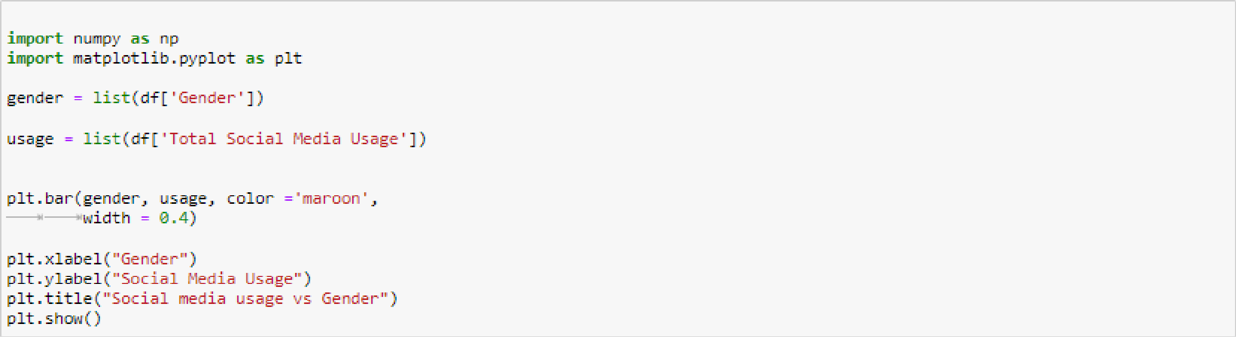
XGBoost:

XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. It has demonstrated exceptional results in a variety of applications, including motion detection, stock sales forecasting, virus classification, consumer behavior analysis, and many more. The two reasons to use XGBoost are also the project's two goals: Model Performance and Execution Speed. Gradient boosting, multiple additive regression trees, stochastic gradient boosting, and gradient boosting machines are all terms used to describe this approach. Gradient boosting is a technique that involves creating new models that forecast the residuals or mistakes of previous models, which are then combined to form the final prediction. Gradient boosting gets its name from the fact that it employs a gradient descent approach to minimize loss while adding new models.

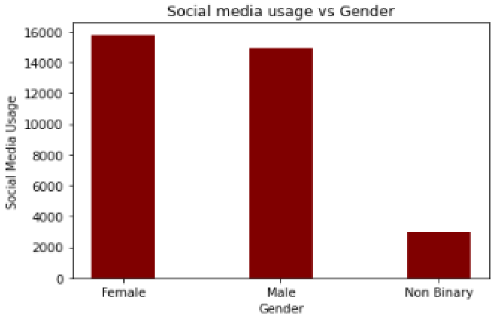
Advantages of XGBoost:

- It is quite adaptable.
- It makes use of parallel processing's power.
- It's a lot quicker than Gradient Boosting.
- It encourages regularization.
- With its built-in capabilities, it is intended to deal with missing data.
- After each cycle, the user can do a cross-validation.
- It is effective in small to larger datasets.

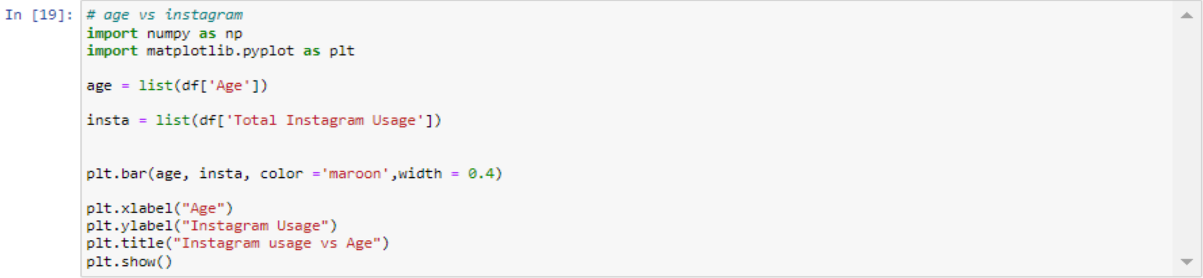
f. Screenshot and Demo along with Visualization (For results):



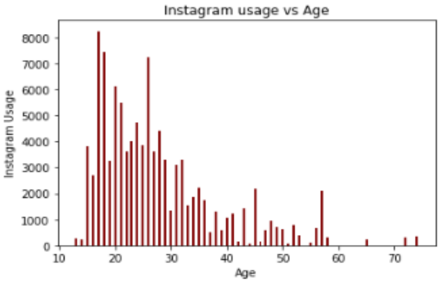
```
import numpy as np
import matplotlib.pyplot as plt

gender = list(df['Gender'])

usage = list(df['Total Social Media Usage'])


plt.bar(gender, usage, color ='maroon',
        width = 0.4)

plt.xlabel("Gender")
plt.ylabel("Social Media Usage")
plt.title("Social media usage vs Gender")
plt.show()
```



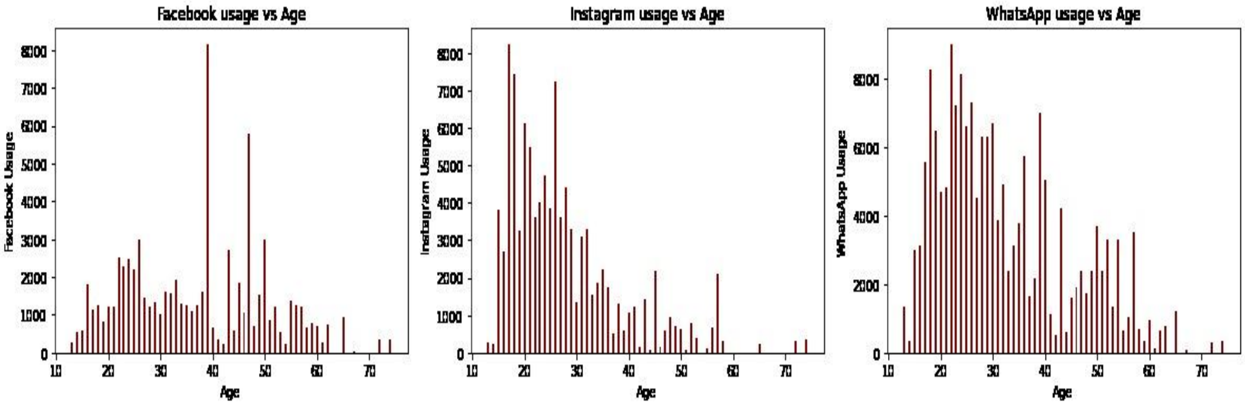
Social media usage plotted against gender



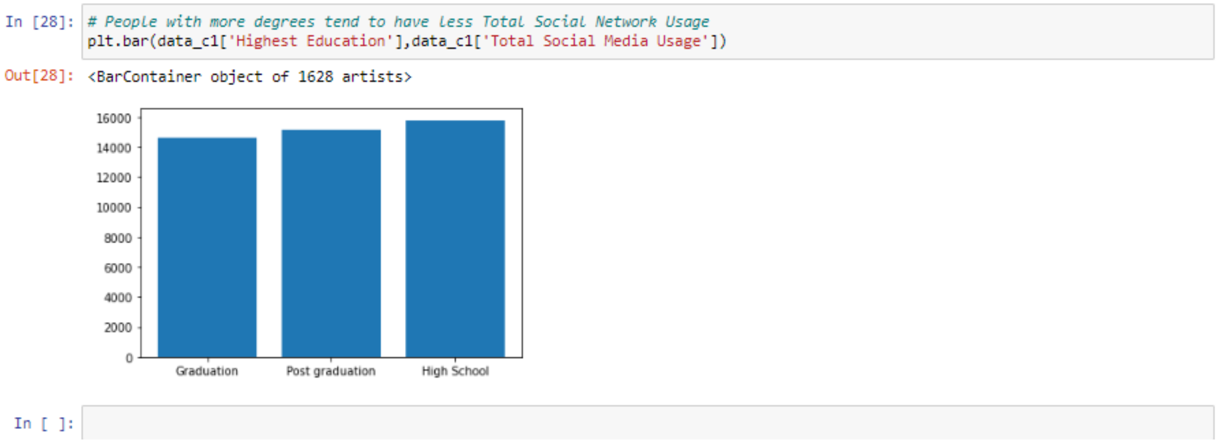
```
In [19]: # age vs instagram
         import numpy as np
         import matplotlib.pyplot as plt

         age = list(df['Age'])

         insta = list(df['Total Instagram Usage'])


         plt.bar(age, insta, color ='maroon',width = 0.4)

         plt.xlabel("Age")
         plt.ylabel("Instagram Usage")
         plt.title("Instagram usage vs Age")
         plt.show()
```



Plotting Instagram usage against age metric.

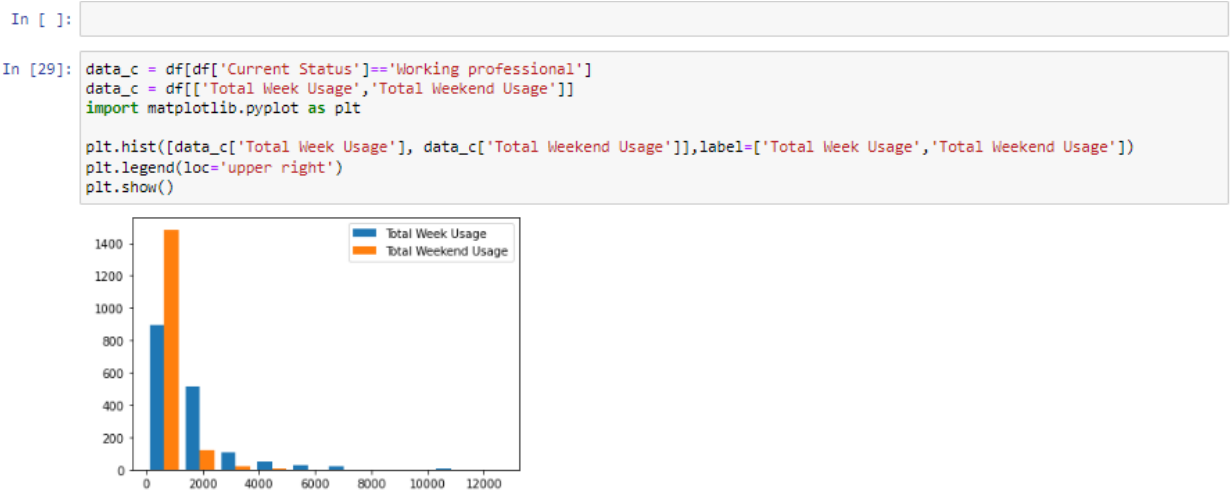


Comparison of Facebook,Instagram and WhatsApp usage against age.



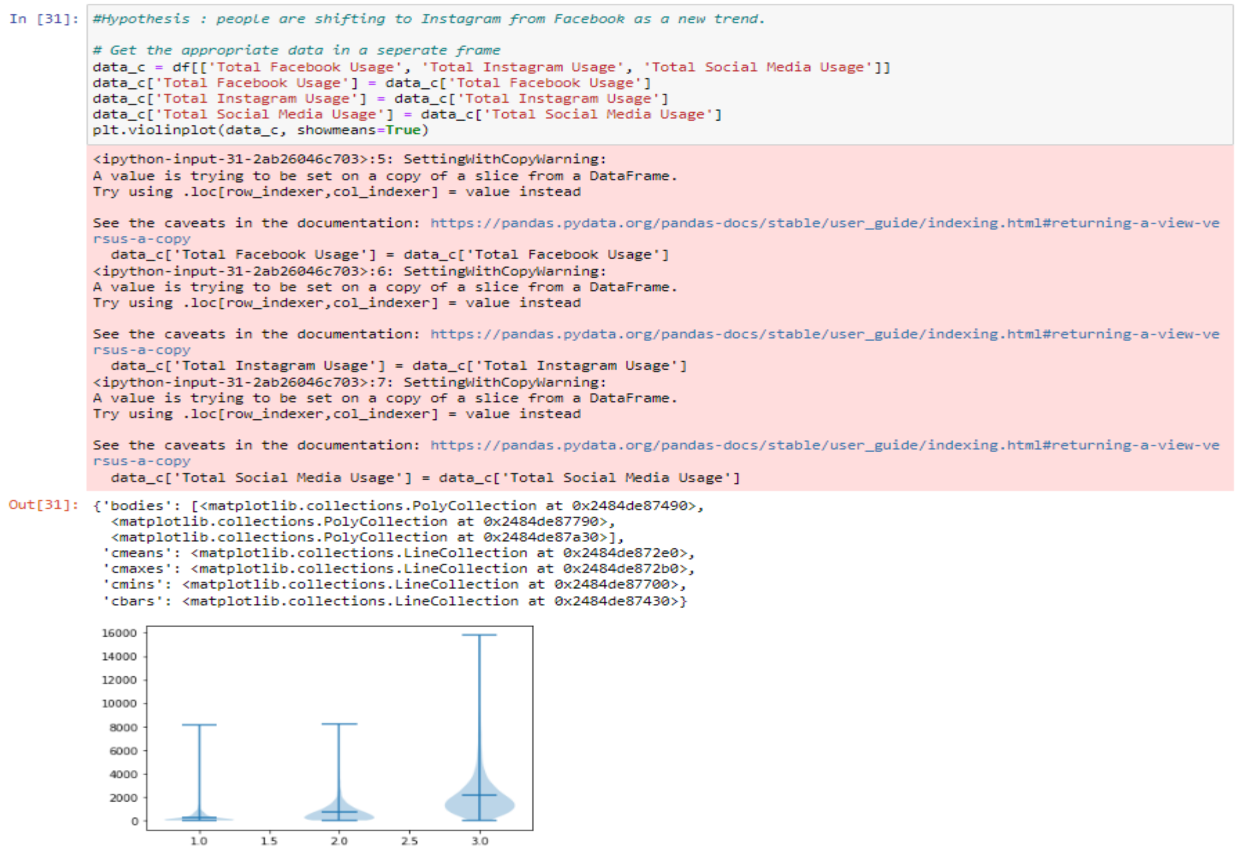
```
In [28]: # People with more degrees tend to have less Total Social Network Usage
         plt.bar(data_c1['Highest Education'],data_c1['Total Social Media Usage'])

Out[28]: <BarContainer object of 1628 artists>
```

Effect of education on social media usage plot



```
In [ ]:

In [29]: data_c = df[df['Current Status']=='Working professional']
         data_c = df[['Total Week Usage','Total Weekend Usage']]
         import matplotlib.pyplot as plt

         plt.hist([data_c['Total Week Usage'], data_c['Total Weekend Usage']],label=['Total Week Usage','Total Weekend Usage'])
         plt.legend(loc='upper right')
         plt.show()
```

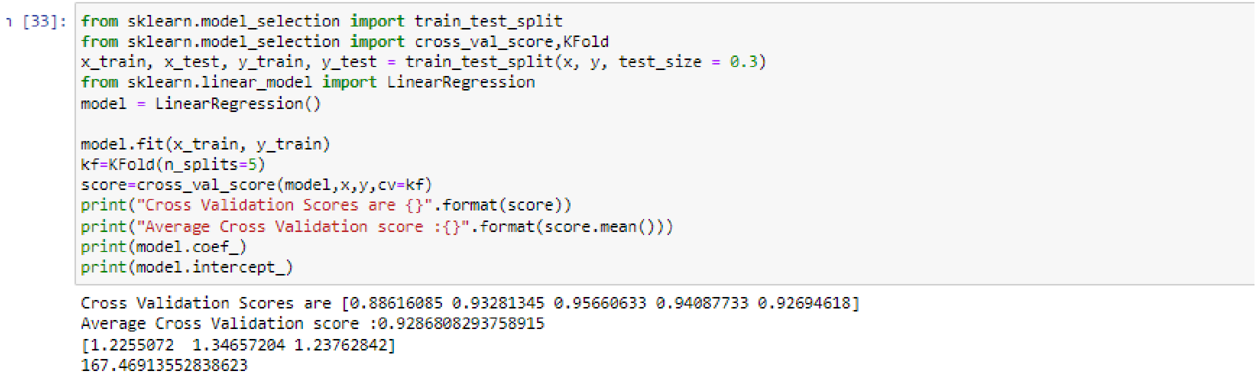
Comparing total week usage and total weekend usage



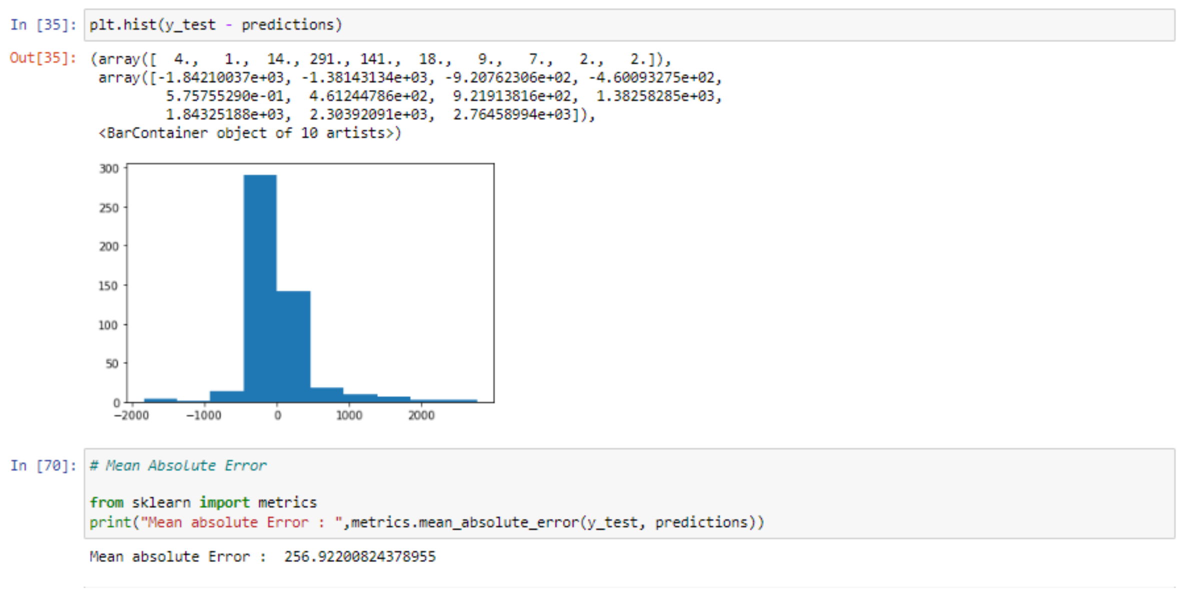
```
In [31]: #Hypothesis : people are shifting to Instagram from Facebook as a new trend.

         # Get the appropriate data in a seperate frame
         data_c = df[['Total Facebook Usage', 'Total Instagram Usage', 'Total Social Media Usage']]
         data_c['Total Facebook Usage'] = data_c['Total Facebook Usage']
         data_c['Total Instagram Usage'] = data_c['Total Instagram Usage']
         data_c['Total Social Media Usage'] = data_c['Total Social Media Usage']
         plt.violinplot(data_c, showmeans=True)

         <ipython-input-31-2ab26046c703>:5: SettingWithCopyWarning:
         A value is trying to be set on a copy of a slice from a DataFrame.
         Try using .loc[row_indexer,col_indexer] = value instead

         See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
         rsus-a-copy
           data_c['Total Facebook Usage'] = data_c['Total Facebook Usage']
         <ipython-input-31-2ab26046c703>:6: SettingWithCopyWarning:
         A value is trying to be set on a copy of a slice from a DataFrame.
         Try using .loc[row_indexer,col_indexer] = value instead

         See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
         rsus-a-copy
           data_c['Total Instagram Usage'] = data_c['Total Instagram Usage']
         <ipython-input-31-2ab26046c703>:7: SettingWithCopyWarning:
         A value is trying to be set on a copy of a slice from a DataFrame.
         Try using .loc[row_indexer,col_indexer] = value instead

         See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
         rsus-a-copy
           data_c['Total Social Media Usage'] = data_c['Total Social Media Usage']

Out[31]: {'bodies': [<matplotlib.collections.PolyCollection at 0x2484de87490>,
           <matplotlib.collections.PolyCollection at 0x2484de87790>,
           <matplotlib.collections.PolyCollection at 0x2484de87a30>],
          'cmeans': <matplotlib.collections.LineCollection at 0x2484de872e0>,
          'cmaxes': <matplotlib.collections.LineCollection at 0x2484de872b0>,
          'cmins': <matplotlib.collections.LineCollection at 0x2484de87700>,
          'cbars': <matplotlib.collections.LineCollection at 0x2484de87430>}
```

Using violin plot to conclude the trend of shift to Instagram from Facebook.

a. Linear Regression is used to predict the total social media usage in Last Week for all the three platforms.



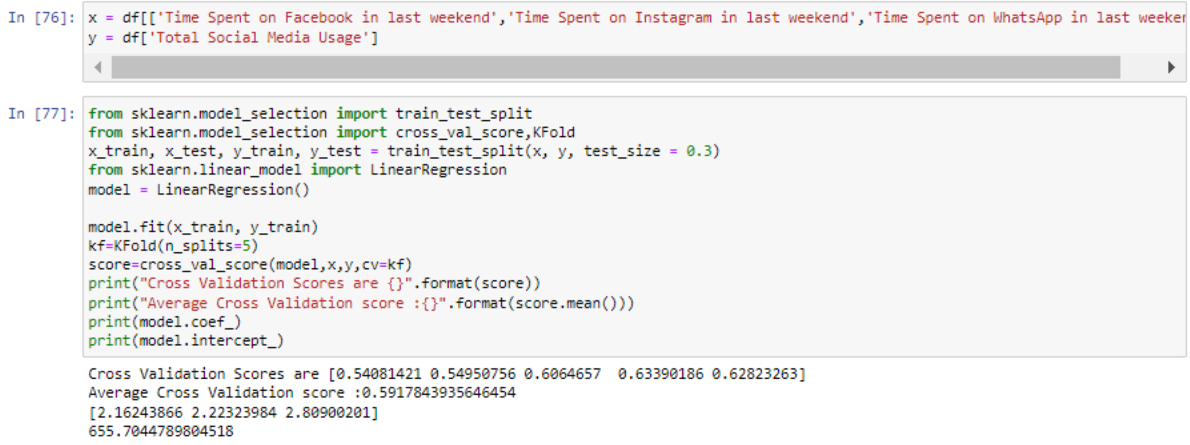
```
ı [33]: from sklearn.model_selection import train_test_split
        from sklearn.model_selection import cross_val_score,KFold
        x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3)
        from sklearn.linear_model import LinearRegression
        model = LinearRegression()

        model.fit(x_train, y_train)
        kf=KFold(n_splits=5)
        score=cross_val_score(model,x,y,cv=kf)
        print("Cross Validation Scores are {}".format(score))
        print("Average Cross Validation score :{}".format(score.mean()))
        print(model.coef_)
        print(model.intercept_)

        Cross Validation Scores are [0.88616085 0.93281345 0.95660633 0.94087733 0.92694618]
        Average Cross Validation score :0.9286808293758915
        [1.2255072  1.34657204 1.23762842]
        167.46913552838623
```



```
In [35]: plt.hist(y_test - predictions)

Out[35]: (array([  4.,   1.,  14., 291., 141.,  18.,   9.,   7.,   2.,   2.]),
          array([-1.84210037e+03, -1.38143134e+03, -9.20762306e+02, -4.60093275e+02,
                  5.75755290e-01,  4.61244786e+02,  9.21913816e+02,  1.38258285e+03,
                  1.84325188e+03,  2.30392091e+03,  2.76458994e+03]),
          <BarContainer object of 10 artists>)
```

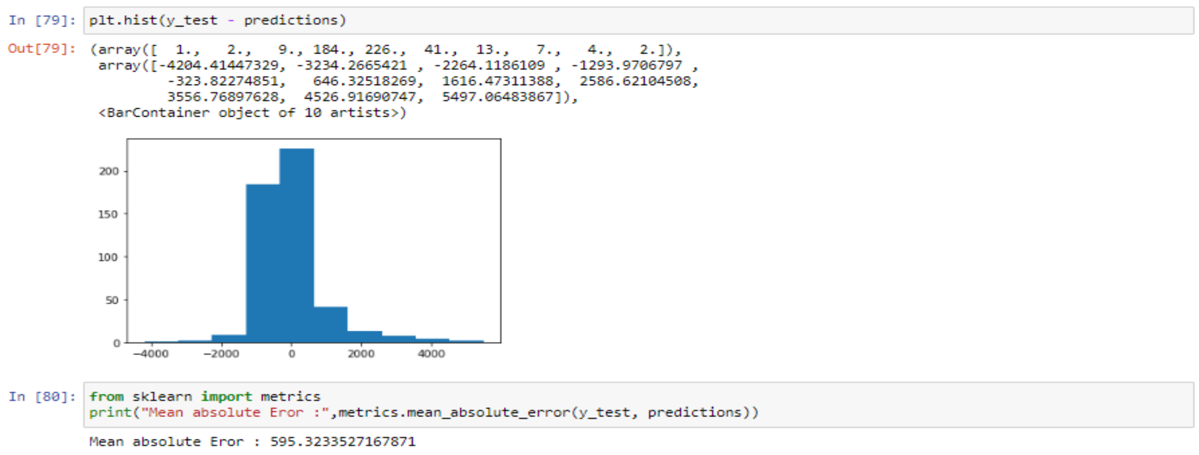
```
In [70]: # Mean Absolute Error

         from sklearn import metrics
         print("Mean absolute Error : ",metrics.mean_absolute_error(y_test, predictions))

         Mean absolute Error :  256.92200824378955
```

b. Linear Regression to predict the total social media usage in Last Weekend for all the three platforms.



```
In [76]: x = df[['Time Spent on Facebook in last weekend','Time Spent on Instagram in last weekend','Time Spent on WhatsApp in last weeker
         y = df['Total Social Media Usage']
```

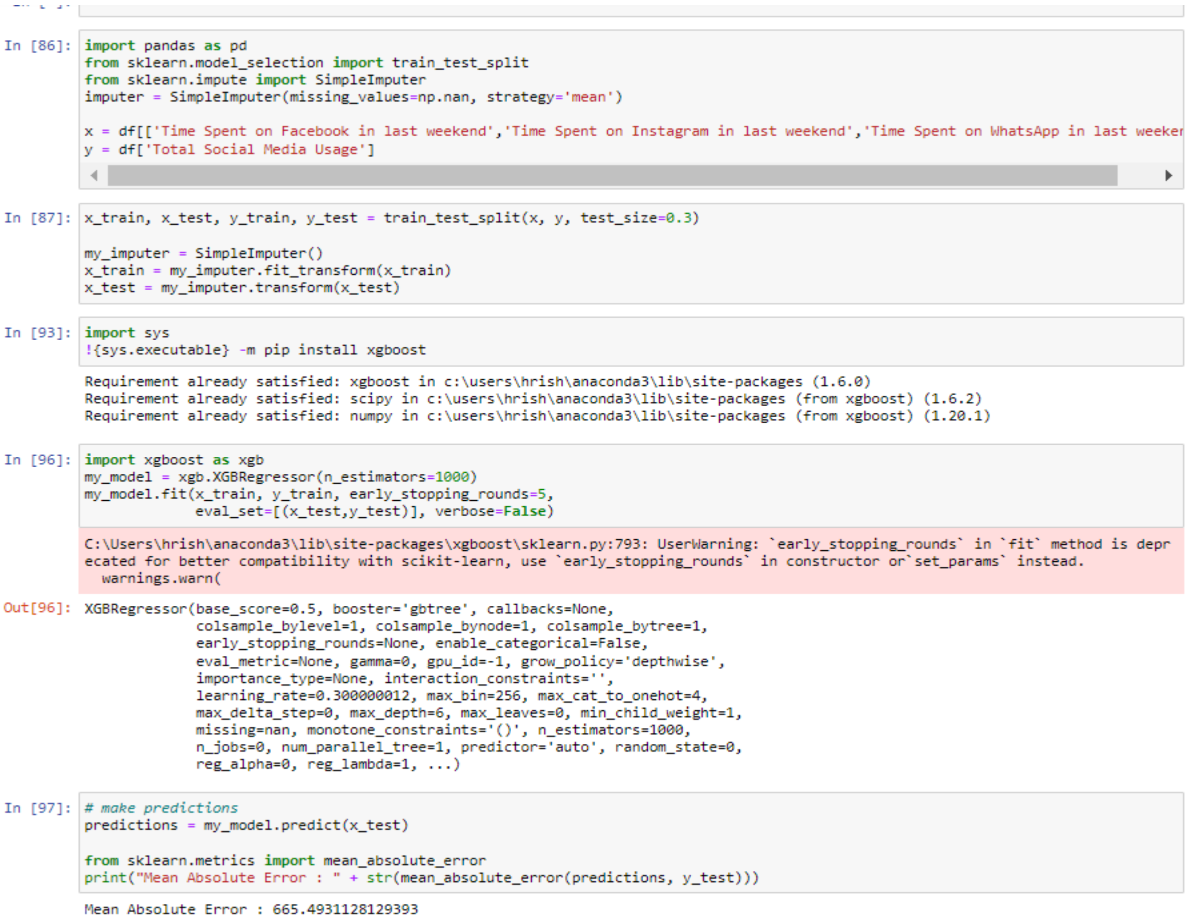
```
In [77]: from sklearn.model_selection import train_test_split
         from sklearn.model_selection import cross_val_score,KFold
         x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3)
         from sklearn.linear_model import LinearRegression
         model = LinearRegression()

         model.fit(x_train, y_train)
         kf=KFold(n_splits=5)
         score=cross_val_score(model,x,y,cv=kf)
         print("Cross Validation Scores are {}".format(score))
         print("Average Cross Validation score :{}".format(score.mean()))
         print(model.coef_)
         print(model.intercept_)

         Cross Validation Scores are [0.54081421 0.54950756 0.6064657  0.63390186 0.62823263]
         Average Cross Validation score :0.5917843935646454
         [2.16243866 2.22323984 2.80900201]
         655.7044789804518
```



```
In [79]: plt.hist(y_test - predictions)

Out[79]: (array([  1.,   2.,   9., 184., 226.,  41.,  13.,   7.,   4.,   2.]),
          array([-4204.41447329, -3234.2665421 , -2264.1186109 , -1293.9706797 ,
                  -323.82274851,   646.32518269,  1616.47311388,  2586.62104508,
                  3556.76897628,  4526.91690747,  5497.06483867]),
          <BarContainer object of 10 artists>)
```

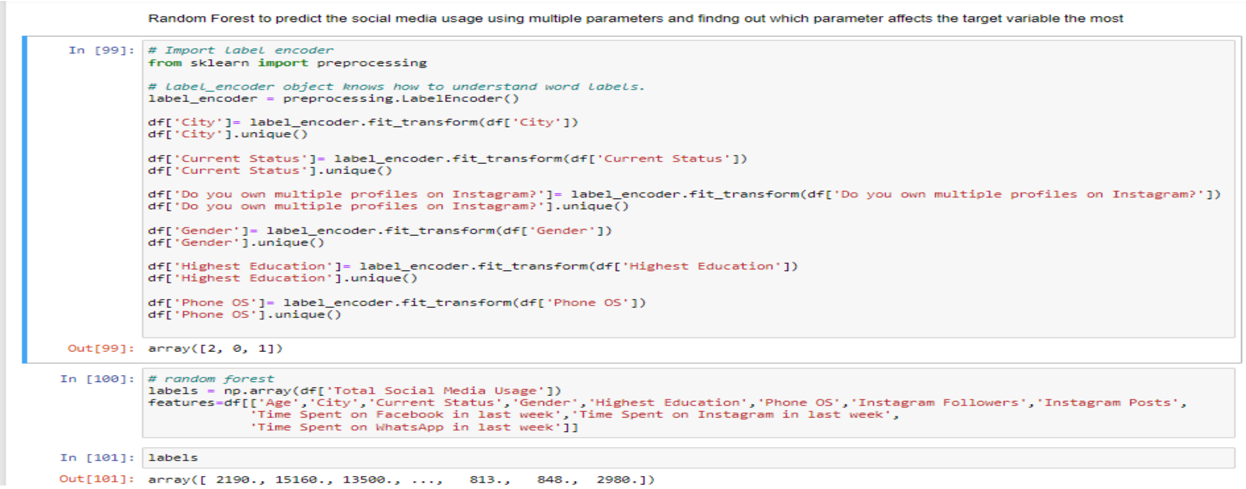
```
In [80]: from sklearn import metrics
         print("Mean absolute Eror :",metrics.mean_absolute_error(y_test, predictions))

         Mean absolute Eror : 595.3233527167871
```

c. XGBoost model is used to improve the performance for total social media usage in the last week for all 3 platforms.



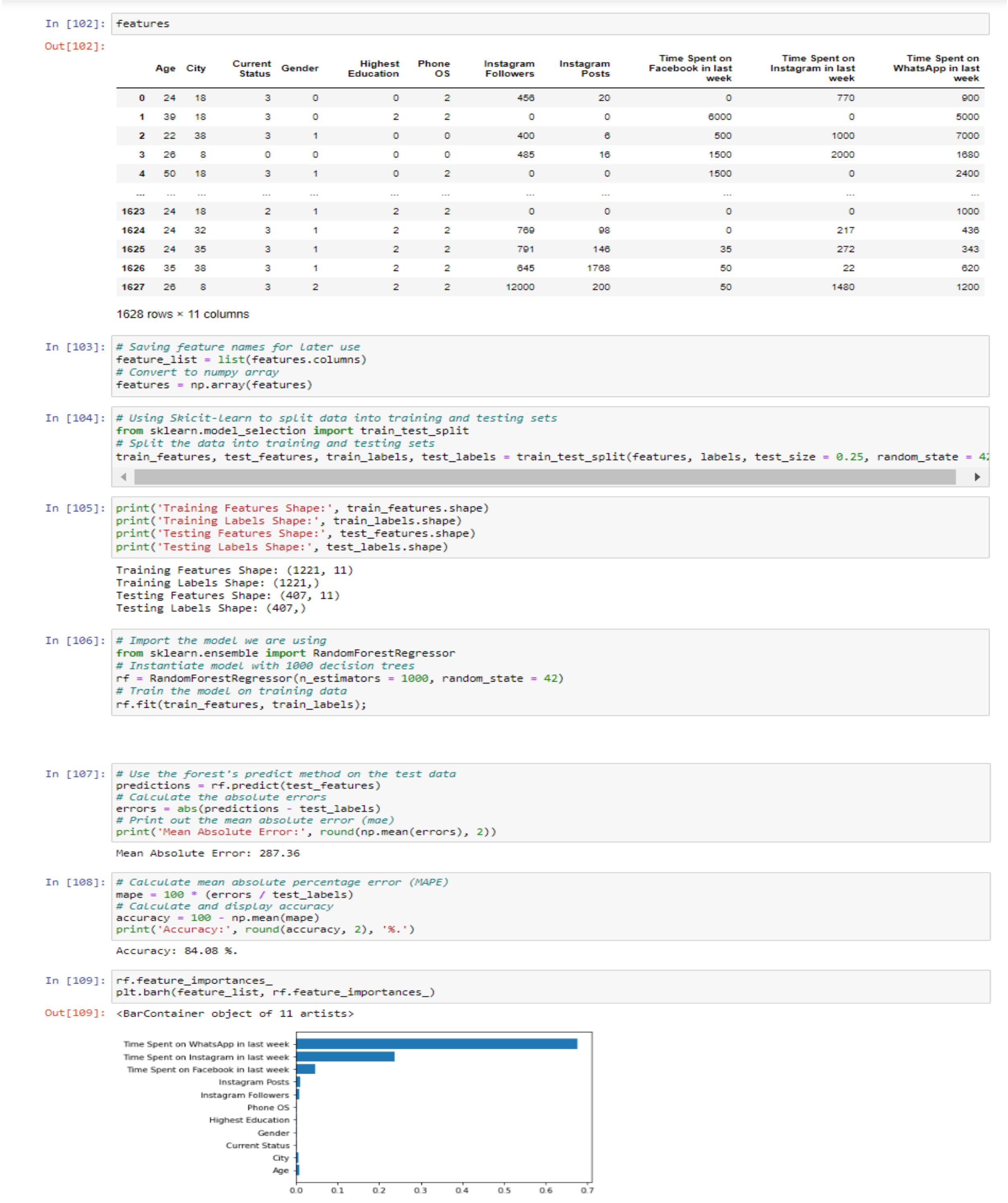
```
In [86]: import pandas as pd
         from sklearn.model_selection import train_test_split
         from sklearn.impute import SimpleImputer
         imputer = SimpleImputer(missing_values=np.nan, strategy='mean')

         x = df[['Time Spent on Facebook in last weekend','Time Spent on Instagram in last weekend','Time Spent on WhatsApp in last weeker
         y = df['Total Social Media Usage']

In [87]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3)

         my_imputer = SimpleImputer()
         x_train = my_imputer.fit_transform(x_train)
         x_test = my_imputer.transform(x_test)

In [93]: import sys
         !{sys.executable} -m pip install xgboost

         Requirement already satisfied: xgboost in c:\users\hrish\anaconda3\lib\site-packages (1.6.0)
         Requirement already satisfied: scipy in c:\users\hrish\anaconda3\lib\site-packages (from xgboost) (1.6.2)
         Requirement already satisfied: numpy in c:\users\hrish\anaconda3\lib\site-packages (from xgboost) (1.20.1)

In [96]: import xgboost as xgb
         my_model = xgb.XGBRegressor(n_estimators=1000)
         my_model.fit(x_train, y_train, early_stopping_rounds=5,
                      eval_set=[(x_test,y_test)], verbose=False)

         C:\Users\hrish\anaconda3\lib\site-packages\xgboost\sklearn.py:793: UserWarning: `early_stopping_rounds` in `fit` method is depr
         ecated for better compatibility with scikit-learn, use `early_stopping_rounds` in constructor or`set_params` instead.
           warnings.warn(

Out[96]: XGBRegressor(base_score=0.5, booster='gbtree', callbacks=None,
                      colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
                      early_stopping_rounds=None, enable_categorical=False,
                      eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
                      importance_type=None, interaction_constraints='',
                      learning_rate=0.300000012, max_bin=256, max_cat_to_onehot=4,
                      max_delta_step=0, max_depth=6, max_leaves=0, min_child_weight=1,
                      missing=nan, monotone_constraints='()', n_estimators=1000,
                      n_jobs=0, num_parallel_tree=1, predictor='auto', random_state=0,
                      reg_alpha=0, reg_lambda=1, ...)

In [97]: # make predictions
         predictions = my_model.predict(x_test)

         from sklearn.metrics import mean_absolute_error
         print("Mean Absolute Error : " + str(mean_absolute_error(predictions, y_test)))

         Mean Absolute Error : 665.4931128129393
```

d. Random Forest is used to to predict the social media usage as well as to predict which parameter affects the social media usage the most.



Random Forest to predict the social media usage using multiple parameters and findng out which parameter affects the target variable the most

```
In [99]: # Import label encoder
         from sklearn import preprocessing

         # label_encoder object knows how to understand word labels.
         label_encoder = preprocessing.LabelEncoder()

         df['City']= label_encoder.fit_transform(df['City'])
         df['City'].unique()

         df['Current Status']= label_encoder.fit_transform(df['Current Status'])
         df['Current Status'].unique()

         df['Do you own multiple profiles on Instagram?']= label_encoder.fit_transform(df['Do you own multiple profiles on Instagram?'])
         df['Do you own multiple profiles on Instagram?'].unique()

         df['Gender']= label_encoder.fit_transform(df['Gender'])
         df['Gender'].unique()

         df['Highest Education']= label_encoder.fit_transform(df['Highest Education'])
         df['Highest Education'].unique()

         df['Phone OS']= label_encoder.fit_transform(df['Phone OS'])
         df['Phone OS'].unique()

Out[99]: array([2, 0, 1])

In [100]: # random forest
          labels = np.array(df['Total Social Media Usage'])
          features=df[['Age','City','Current Status','Gender','Highest Education','Phone OS','Instagram Followers','Instagram Posts',
                       'Time Spent on Facebook in last week','Time Spent on Instagram in last week',
                       'Time Spent on WhatsApp in last week']]

In [101]: labels

Out[101]: array([ 2190., 15160., 13500., ...,   813.,   848.,  2980.])
```



```
In [102]: features
```

Out[102]:

	Age	City	Current Status	Gender	Highest Education	Phone OS	Instagram Followers	Instagram Posts	Time Spent on Facebook in last week	Time Spent on Instagram in last week	Time Spent on WhatsApp in last week
0	24	18	3	0	0	2	456	20	0	770	900
1	39	18	3	0	2	2	0	0	6000	0	5000
2	22	38	3	1	0	0	400	6	500	1000	7000
3	26	8	0	0	0	0	485	16	1500	2000	1680
4	50	18	3	1	0	2	0	0	1500	0	2400
...	...	...	...	...	...	...	...	...	...	...	...
1623	24	18	2	1	2	2	0	0	0	0	1000
1624	24	32	3	1	2	2	769	98	0	217	436
1625	24	35	3	1	2	2	791	146	35	272	343
1626	35	38	3	1	2	2	645	1768	50	22	620
1627	26	8	3	2	2	2	12000	200	50	1480	1200

1628 rows × 11 columns

```
In [103]: # Saving feature names for later use
          feature_list = list(features.columns)
          # Convert to numpy array
          features = np.array(features)
```

```
In [104]: # Using Skicit-learn to split data into training and testing sets
          from sklearn.model_selection import train_test_split
          # Split the data into training and testing sets
          train_features, test_features, train_labels, test_labels = train_test_split(features, labels, test_size = 0.25, random_state = 42
```

```
In [105]: print('Training Features Shape:', train_features.shape)
          print('Training Labels Shape:', train_labels.shape)
          print('Testing Features Shape:', test_features.shape)
          print('Testing Labels Shape:', test_labels.shape)

          Training Features Shape: (1221, 11)
          Training Labels Shape: (1221,)
          Testing Features Shape: (407, 11)
          Testing Labels Shape: (407,)
```

```
In [106]: # Import the model we are using
          from sklearn.ensemble import RandomForestRegressor
          # Instantiate model with 1000 decision trees
          rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)
          # Train the model on training data
          rf.fit(train_features, train_labels);
```

```
In [107]: # Use the forest's predict method on the test data
          predictions = rf.predict(test_features)
          # Calculate the absolute errors
          errors = abs(predictions - test_labels)
          # Print out the mean absolute error (mae)
          print('Mean Absolute Error:', round(np.mean(errors), 2))

          Mean Absolute Error: 287.36
```

```
In [108]: # Calculate mean absolute percentage error (MAPE)
          mape = 100 * (errors / test_labels)
          # Calculate and display accuracy
          accuracy = 100 - np.mean(mape)
          print('Accuracy:', round(accuracy, 2), '%.')

          Accuracy: 84.08 %.
```

```
In [109]: rf.feature_importances_
          plt.barh(feature_list, rf.feature_importances_)

Out[109]: <BarContainer object of 11 artists>
```

e. Predicting the Total Instagram usage based on number of followers, posts and age using Linear Regression.

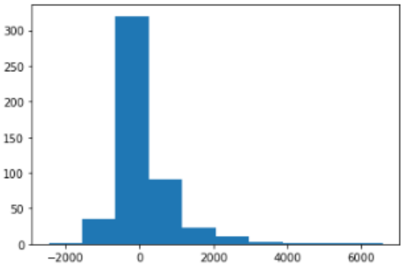
Instagram followers and instagram posts - using Linear Regression

```
In [136]: x = df[['Instagram Followers','Instagram Posts','Age']]
          y = df['Total Instagram Usage']
```

```
In [137]: from sklearn.model_selection import train_test_split
          from sklearn.model_selection import cross_val_score,KFold
          x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3)
          from sklearn.linear_model import LinearRegression
          model = LinearRegression()

          model.fit(x_train, y_train)
          kf=KFold(n_splits=5)
          score=cross_val_score(model,x,y,cv=kf)
          print("Cross Validation Scores are {}".format(score))
          print("Average Cross Validation score :{}".format(score.mean()))
          print(model.coef_)
          print(model.intercept_)

          Cross Validation Scores are [0.02968191 0.11225103 0.05582004 0.05934594 0.02536578]
          Average Cross Validation score :0.05649294096335773
          [-7.91535919e-04  7.35857560e-01 -2.16923283e+01]
          1202.111647779611
```

```
In [138]: predictions = model.predict(x_test)

          print(predictions)
```

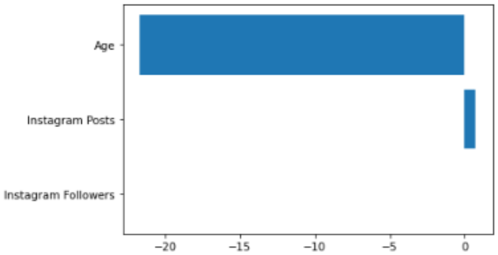
```
 In [139]: plt.hist(y_test - predictions)

Out[139]: (array([  1.,  36., 320.,  91.,  23.,  10.,   3.,   2.,   2.,   1.]),
           array([-2463.63422188, -1559.5802949 ,  -655.52636792,   248.52755906,
                   1152.58148604,  2056.63541302,  2960.68934   ,  3864.74326698,
                   4768.79719396,  5672.85112093,  6576.90504791]),
           <BarContainer object of 10 artists>)
```



```
 In [140]: from sklearn import metrics
           metrics.mean_absolute_error(y_test, predictions)

Out[140]: 542.3768019698879
```

```
In [153]: print(model.coef_)
          importance=model.coef_

          [-7.91535919e-04  7.35857560e-01 -2.16923283e+01]
```

```
 In [155]: plt.barh(x.columns, importance)

Out[155]: <BarContainer object of 3 artists>
```



```
In [ ]:
```

6. Comparison, Results and discussion along with Visualization

LR (1) : Linear Regression model to predict the Total Social Media Usage of the last week across all 3 platforms.

LR (2): Linear Regression model to predict the Total Social Media Usage of the last weekend across all 3 platforms.

LR (3) : Linear Regression model to predict the Total Instagram Usage based on parameters such as Age, number of followers and number of posts.

Model	MAE	MSE	R2 Score	RMSE	Adjusted R2	Accuracy
LR (1)	256.92	188499.940	0.9473	434.165	0.947	85.15 %
LR (2)	595.32	859042.615	0.7324	926.845	0.732	51.35 %.
XGBoost	665.49	1448536.50	0.6013	1203.55	0.600	55.04 %
RF	287.36	239622.08	0.9228	489.512	0.922	84.08 %.
LR (3)	542.37	1176261.75	0.0596	1084.556	0.675	58.03%

- We can clearly observe that people are shifting to Instagram from Facebook as a trend.
- Linear Regression model was found to have the highest accuracy among the other models.
- XGBoost model increased the overall accuracy and performance of LR (2) model by 4 % which is significant in model selection .
- Random Forest model obtained the highest accuracy indicating that this model can be used in future predictions since it had the least value of MAE.
- The RF model also predicts which features affect social media usage. It can be concluded that in India WhatsApp usage is the most followed by Instagram and then Facebook. This

also confirms our hypothesis that there is a shift in the social media platforms used by people as people are shifting from Facebook to Instagram.

- The parameters “city” and “Age” play a minor role in the prediction of Total Social Media Usage while the type of phone a user has does not matter the social media usage by an individual.
- For the prediction of Instagram usage of an individual the parameter “Age” had a negative coefficient while the number of Instagram followers has almost 0 impact on the target variable. It is worthwhile noting that the number of Instagram posts has a positive coefficient which concludes that the quantity of Instagram posts governs your Total Instagram activity.
- From the visualizations we can see that the weekend usage of social media platforms is more than the week usage. This point can be of strategic importance to various brands who use social media to promote their products. They can target specific social media platforms and to be more specific they can advertise their products during peak user activity to get maximum customer attention.
- Due to the limitations of the dataset we had very few parameters to predict the social media usage. Parameters such as “number of likes” , “no_of_posts_scrolled”, “number of views” etc can help to improve the performance and make the model learn better.

To conclude, we can say that predictive analysis of Social media platforms can be of great importance from a business point of view for the specific social media site and of marketing importance to the various brands who wish to sell their products via social media sites in order to increase their reach. More specific and insights data about the users can be very effective in predicting various trends as well as user behavior.

7. References

[1] S. A. Salloum, N. M. N. AlAhbabi, M. Habes, A. Aburayya, and I. Akour, “Predicting the intention to use social media sites: A hybrid SEM - machine learning approach,” in *Advances in Intelligent Systems and Computing*, Cham: Springer International Publishing, 2021, pp. 324–334.

[2] K. Chaudhary, M. Alam, M. S. Al-Rakhami, and A. Gumaei, “Machine learning-based mathematical modeling for prediction of social media consumer behavior using big data analytics,” *J. Big Data*, vol. 8, no. 1, 2021.

[3] S. De, A. Maity, V. Goel, S. Shitole, and A. Bhattacharya, “Predicting the popularity of instagram posts for a lifestyle magazine using deep learning,” in *2017 2nd International Conference on Communication Systems, Computing and IT Applications*

(CSCITA), 2017, pp. 174–177.

[4] R. Kristo, D. Purba, and R. K. Asirvatham, “Instagram post popularity trend analysis and prediction using hashtag, image assessment, and user history features,” *int. Arab j. inf. technol.*, vol. 18, no. 1, 2020.

[5] B. S. Arasu, B. J. B. Seelan, and N. Thamaraiselvan, “A machine learning-based approach to enhancing social media marketing,” *Comput. Electr. Eng.*, vol. 86, no. 106723, p. 106723, 2020.

[6] M. Savci, A. Tekin, and J. D. Elhai, “Prediction of problematic social media use (PSU) using machine learning approaches,” *Curr. Psychol.*, 2020.

[7] C. Liu, F. Li, and L. Li, “Research on Gender Prediction for Social Media User Profiling by machine learning method,” in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2021, pp. 831–836.