Intro to Data Analytics

Project Preliminary Report

Danny Wong, Chi Nguyen, Josue Chigwira

29 November, 2021

**Abstract**

For more than a decade, social media has been one of the fastest growing industries in the world. Its use and potential is unparalleled by anything else in the world. In this project, we explored basic elements that define a user’s experience. We used general data analytic tactics, basic models, and our combined knowledge of programming to analyze three separate data sets. We found that there exists slight correlation between many of the aspects of social media (eg. Time, Likes, Followers) as one would likely predict.

**Table Of Contents**

Introduction……………………………………...……………………..…...3

Data Description……………………..………………………...……………3

Research Plan…………………………………………………………….….5

Exploratory Analysis…………………………………………………..…….6

Instagram………………..………………………………………………..….6

Facebook………………………………………………………………….….9

Twitter………….…………………………………………………….……..14

Models…………………………………………………………...………….16

Conclusion……………………………………………………………..……18

Future of the Project……………………………………………………..….18

**Introduction**

In this project, we will be taking data from popular social media apps/websites and using it to create visuals, make predictions, and analyze trends and relationships. The social media giants we will be using are Instagram, Twitter, and Facebook. We intend to use typical data analytic methods such as regression analysis and categorical predictions. The goal of this project is to learn about tendencies among social media users and share our findings.

**Data Description**

In this project, we will be utilizing 3 separate data sets, all of which regard social media. Each data set comes from the kaggle website: <https://www.kaggle.com>, which provides over 100,000 free datasets. Kaggle is an online community of data scientists and machine learning practitioners. Only one dataset has some missing values and all three have some repetitions due to the information that are in the datasets. (For example: In the Instagram dataset the number of likes, number of tags, number of comments, and years have some are repeated based on the people that interacted with the post and the year it was posted.)

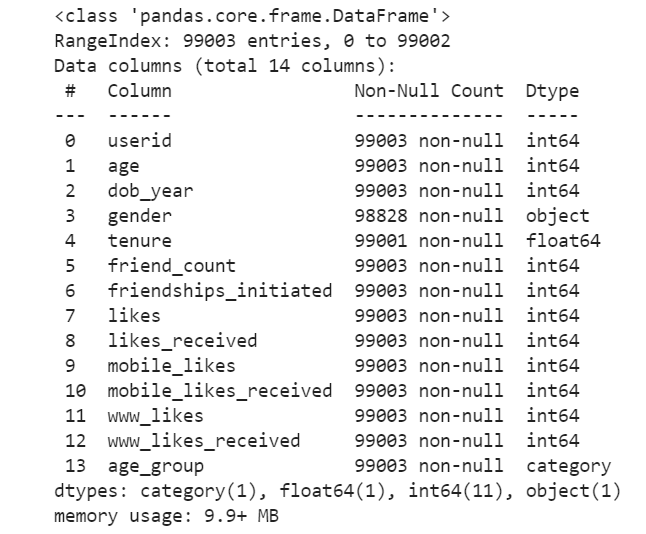
There are 13 columns in the Instagram dataset, each one represents a user, likes, days passed from post, like score, type, number of tags, number of comments, date posted, year, month, day, hour, and minute about a post. And in the twitter dataset, there are 25 columns that represent unit id, golden, unit state, trusted judgement, last judgement, gender, gender confidence, profile yn, profile yn confidence, created, description, fav number, gender gold, link color, name, profile yn gold, profile image, retweet count, sidebar color, text, tweet coordinate, tweet count, tweet created, tweet id, tweet location, and user timezone about the tweet. And the pseudo Facebook has the same data that represents a post. It has 14 columns. Each information that is entered in the Instagram database has about 200,000 columns with it. All the datasets have valuable information about the subject they represent but two of them are missing some information (value) which might be important. Apart from that, these datasets have enough information to find anything someone might want from it.

Instagram data information:

A screenshot of a computer

Description automatically generated with low confidence

Facebook data information:



**Twitter data information:**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable name | Data type | Variable name | Data type |
| \_unit\_id | numerical | link\_color | categorical |
| \_golden | categorical | name | categorical |
| \_unit\_state | categorical | profile\_yn\_gold | categorical |
| \_trusted\_judgments | numerical | profileimage | categorical |
| \_last\_judgment\_at | numerical | retweet\_count | numerical |
| gender | categorical | sidebar\_color | categorical |
| gender:confidence | numerical | text | categorical |
| profile\_yn | categorical | tweet\_coord | categorical |
| profile\_yn:confidence | numerical | tweet\_count | numerical |
| created | numerical | tweet\_created | numerical |
| description | categorical | tweet\_id | numerical |
| fav\_number | numerical | tweet\_location | categorical |
| gender\_gold | categorical | user\_timezone | categorical |

**Research Plan**

The first step in our research plan was to locate large .csv data files that contain enough rows and inputs to do a complete analysis. We decided to look at and analyze social media. We found three files, each from a separate social media website.

The next step in our project is to visually evaluate each column in all three data sets. We only want to work with columns that we deem useful in our analysis.

The third step in our analysis is to use the variables that we liked and begin fitting models. We aim to find, compare, and evaluate, relationships within each dataset, and maybe find patterns that fit over all three data sets.

The fourth step is to summarize our findings and compile a report. We want to share the trends in our data to help ourselves and others better understand social media.

**Exploratory Analysis / Visuals**

**Instagram Dataset**

In order to get a general idea of the spread among our features/inputs in our Instagram dataset, we use the built in .describe() function to print the summary statistics.

Summary statistics from Python

**Table

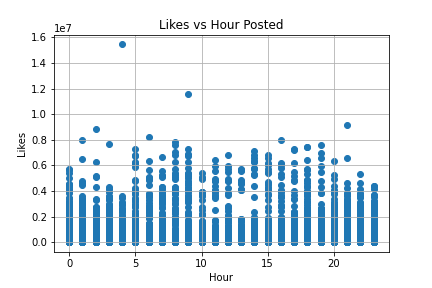
Description automatically generated**

Distribution of post type:

Chart, bar chart

Description automatically generated

**Task:** Can we correlate Time Posted with Number of Likes?

The relationship between Hour Posted and Likes:

The relationship between Day of the Month Posted and Likes:

Chart, scatter chart

Description automatically generated

There seems to be no clear correlation between what time of day a user decides to post or what day of the month a user decides to post and number of likes they recieve. Analysis of a larger dataset however might show evidence of “prime time”. Overall, the Instagram dataset isn’t able to tell us much. It contains only 13 different features, most of which are slightly redundant.

Model #1: OLS

Table

Description automatically generated

The OLS Regression Results show us the weights assigned to each Hour, Minute, and Day of a post as it correlates to total Likes.

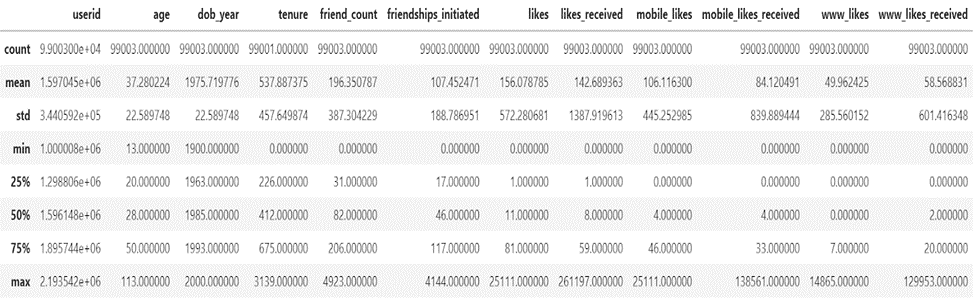
Models #2 and #3: Linear Regression and Random Forest

The Linear Regression model resulted in a Root Mean Square Error between 250,000 and 300,000. The Random Forest also resulted in a Root Mean Square Error between 250,000 and 300,000. As a result, we concluded that Hour, Minute, and Day are not effective at predicting the number of likes on an Instagram post. At least according to this dataset.

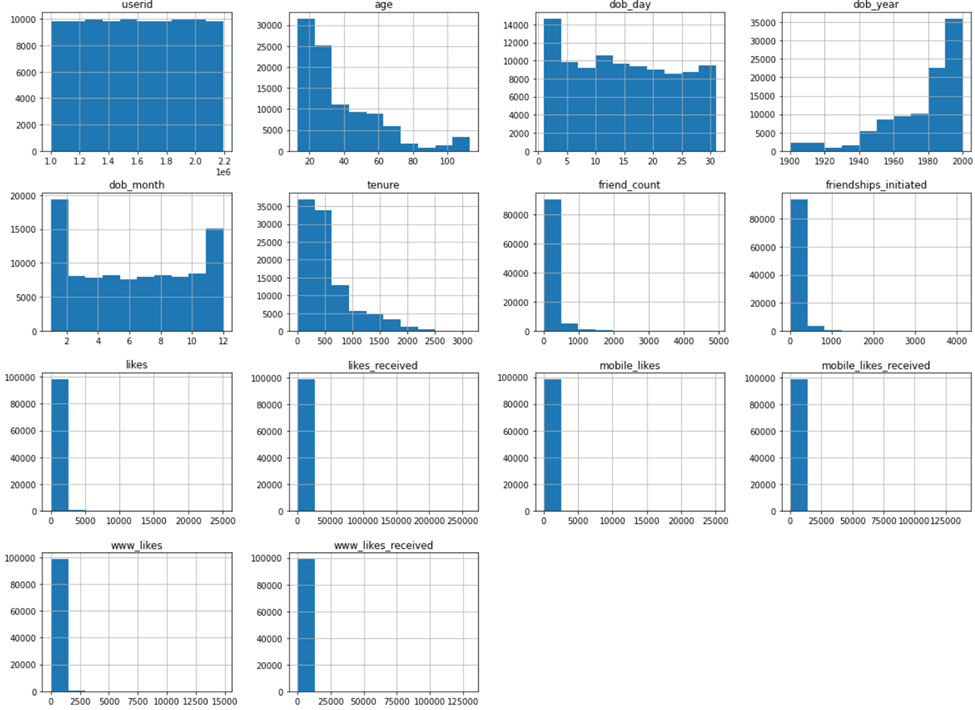
**Facebook dataset**

This summary statistic helps us summarize the set of observations in the Facebook dataset emphasizing the average and helps determine whether the data is skewed. In python, we use the describe function that returns the summary statistic for each numerical variable.

Facebook dataset summary statistic from python:



The summary statistic provides a lot of information regarding each numerical variable and shows the average value which will help to show whether the data is skewed. The following bar charts are going to show us if the data are right skewed or left skewed.



We can clearly see that most of the data are right-skewed which means mode is less than median and median is less than mean except the dob\_year is the only bar chart that is left-skewed which means its mean is less than the median and the median is less than the mode. We are going to use this bar chart to help visualize the gender that uses Facebook more than the other.

Graph of genders presented in the Facebook dataset:

Chart, bar chart

Description automatically generated

This bar chart shows us exactly how many genders are presented and the gender that use Facebook more than the other, the pie chart shows us which gender has a higher percentage and more likes in the dataset. From this chart we can see that male gender has more likes and uses Facebook more than the female gender.

The next chart is going to be used to find the age of the people who use Facebook more than others. To find the age of users that use Facebook more we will have to group users by their age.

Graph of Facebook users grouped by age:

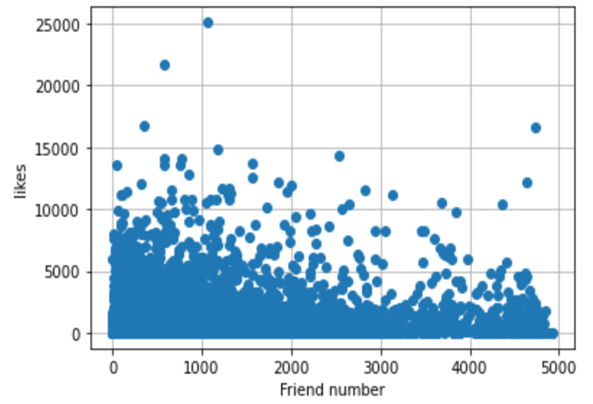
Chart, bar chart

Description automatically generated

Here, we can see that the age of users that use Facebook more than the other are the users that are age 21-30. Also, we can see that all the people that are 40-10 use Facebook more than every other user.

This scatter-plot chart is going to help us show the linear relationship between the number of likes and Friends number.

The relationship between Friends number and Number of likes on the Facebook dataset:



The graph shows that the number of likes and the number of friends has a negative correlation. the number of friends is increasing, the likes are decreasing. When the friend number is 4000 the likes are about 3000, and when the friend number is 0 the likes are about 14000 which explains the negative correlation coefficient.

Conclusion for the Facebook dataset:

In this part, we used our knowledge in Python language to highlight the basic property of the data. First, we analyzed the Facebook user by simply exploring the data and creating charts that helped visualize the data. Visualizing the data helped to show the trends and patterns in the data. Next, we analyzed the data and created a bar chart to show the gender that used Facebook more than the other. Later, we analyzed the data and created a chart that shows the age that used Facebook more than others. Lastly, we analyzed the correlation of the numerical variables by showing the example of linear correlation in the dataset.

We can see that male gander used Facebook more than the other gender and users from the age 10-40 used Facebook more than other ages. We also see that users do not get more likes based on the number of friends because the chart about the relationship between Friends number and Number of likes on the Facebook dataset showed that users with a smaller number of friends had more likes than the users with a big number of friends.

**Twitter dataset**

**Task:** What are the average, max, and min values (likes, likes score, number of tags, etc) for the Twitter dataset?

We want to summarize the statistics of each column. To solve this task, we use the describe() function to summarize the statistics for each numerical variable in Python.

Twitter dataset summary statistics from Python

Table

Description automatically generated

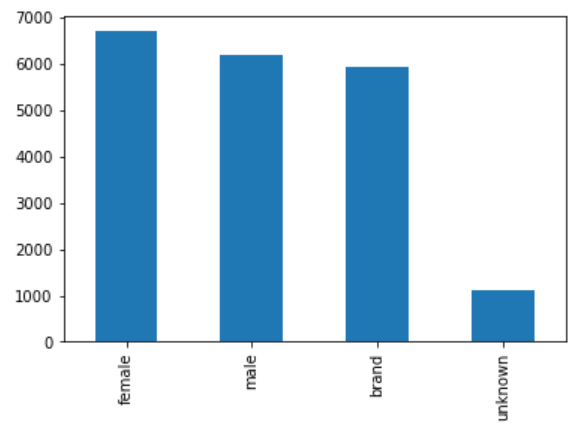
In this table, the average , max, and min values of each column are described at row 3, 5, and 9.

**Task:** How do men and women use Twitter differently?

To answer this question, we create frequency table and a bar chart based on gender column.

Graph of Twitter users gender:

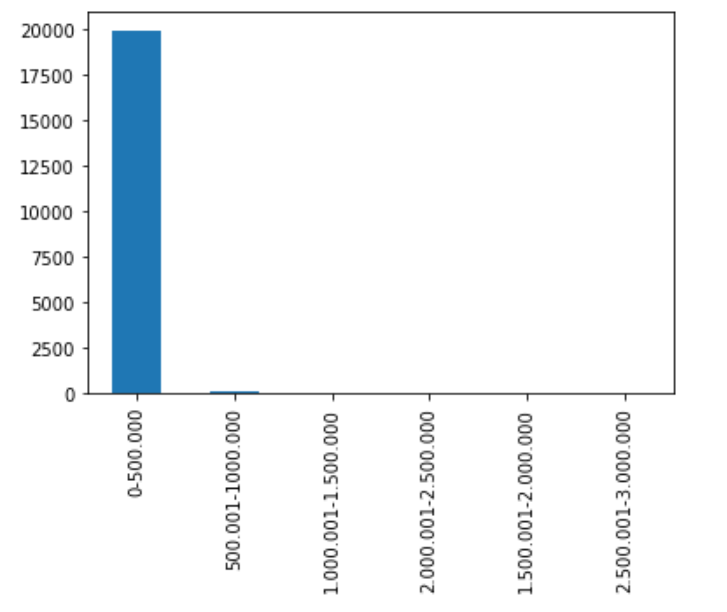
|  |  |
| --- | --- |
| female | 6700 users |
| male | 6194 users |
| brand | 5942 users |
| unknown | 1117 users |



The bar chart shows us the summary statistics of how gender use Twitter differently and how many people verify their gender in Twitter. From the bar chart and the table, we can see women are more active users than men with 6700 users, the second is men with 6194 users, and the third is brand with 5942 users. There are 1117 users who did not verify gender in Twitter.

**Task**: Based on the statistics table "Twitter dataset summary statistics from Python", evaluate the number of users' tweet count.

As we have summarized in the "Twitter dataset summary statistics from Python", the smallest tweet\_count is 1 and the largest is 2.680.199. We divide into subgroups according to the numbers 0 - 500.000, 500.001-1000.000, 1.000.001-1.500.000, 1.500.001-2.000.000, 2.000.001-2.500.000, and 2.500.001-3.000.000.

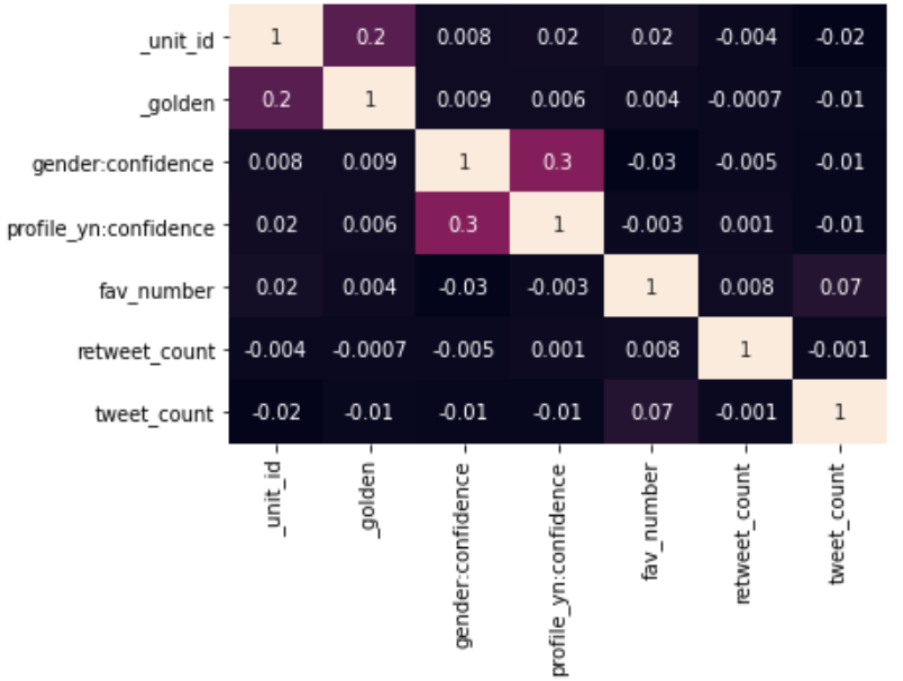


|  |  |
| --- | --- |
| 0-500.000 | 19904 user |
| 500.001-1000.000 | 89 user |
| 1.000.001-1.500.000 | 29 user |
| 2.000.001-2.500.000 | 25 user |
| 1.500.001-2.000.000 | 2 user |
| 2.500.001-3.000.000 | 1 user |

From the statistical results, we can see that the tweet\_count of the group 0 - 500.000 is the highest, which means that the users in this data table have the most tweets from 0 tweets to 500.000 tweets. The number of tweet\_count decreased rapidly to group 1.500.001-2.000000 with 2 users and increased to 25 users in group 2.000.001-2.500.000. And, reduced to 1 user in group 2.500.001-3.000.000.

**Task**: What is the correlation between the numerical variables of tweet\_count?

We generate a heat map of the variables to check the linear correlation between the variables in the data and correlation coefficients among the variables. Using the MatPlotLib library in Python:



**Models**

**K-Means Clustering:**

We want to build K-Means clustering of the users who have tweet\_count equal or greater than 200000. This process will require the following steps:

* As a first step, we will filter the information of users whose tweet\_count is greater than or equal to 200000.
* We will normalize the data using Z-Score Normalization. The K-Means process depends on the Euclidean distance and therefore all values must have the same scale to start the process.
* Then, we will perform Hierarchical Clustering. Because we have 539 rows, we will build on cluster 11.
* Next, with the information provided, we run K-Means and group observations with similar characteristics.

Chart, histogram

Description automatically generated

To describe this, we choose the height and go horizontally, each vertical line that we cross represents the moment when the objects join together into a cluster. When you look at our chart, it's a tough task to determine the appropriate number of clusters. There is not a clear number for us to consider.

**Conclusion**

We used 3 datasets about users of Facebook, Instagram, and Twitter in this project. The data we collected is user data which is information that users provide when interacting via websites, mobile applications, surveys, social media, etc. To make this information more objective, we used our knowledge of Python and R-Studio to build and evaluate the data. It is an extensive process that examines large data sets through various tools and processes to uncover unknown patterns, hidden correlations, meaningful trends, and other insights, to make data-driven decisions in the pursuit of better outcomes. These tools help us build relationships between columns by graphing, modeling, and mapping. We used them to build a model of gender, likes, age, etc. based on data about user behavior, by input or group of inputs; make reports based on criteria: age, gender, location of visitors..., helping to target the right advertising campaigns. This is information that can be used to identify an individual from which we can set up groups of data, and can help businesses increase their user data collection by investing in social media ads. Through the targeting capabilities of these platforms, you can understand the preferences and other characteristics of your customers.

Understanding user data will help us collect the necessary information to achieve a certain purpose. Best of all, you can enhance user experience, increase conversion rates, or even understand our users’ latent desires through data analysis.

**Future of the Project**

With this project, we are collecting and sharing information regarding trends among popular social media sites. This information will allow us to study and understand the common user. One possibility is to share our findings with our peers and let them use our predictions to their advantage. Additionally, we could send our results to these social media companies so they can use our models to improve their websites. After this course, we gained our knowledge about collecting, processing and consolidating discrete and fragmented data.