jdb393, sy629
14 May 2019
IS 2950

INFO 2950 Final Project Datasheet:
Instagram Post Data

Our goal was to analyze the data of Instagram posts. We developed our dataset based off the information explicitly presented on these posts: the tags used on the posts, the total number of posts containing each tag used on all of Instagram, tagged accounts on posts, the number of followers for the tagged accounts, and the number of followers the posting user has. Our plan was to measure how these variables correlate with post engagement, how many likes and comments posts receive by other users. Our initial hypothesis was that the presence of popular tags and tagged users with large followings would lead to more engagement, meaning more likes and comments on posts. We also hypothesized that accounts with more followers would have higher engagement on their posts.

We developed code using Python, Chromedriver, Selenium, and BeautifulSoup to send requests to the Instagram website and scrape this data from Instagram accounts, whose usernames we collected from a dataset from HypeAuditor. We First attempted to access the Instagram HTML using BeautifulSoup, but the Instagram website's HTML is generated using JavaScript, so using BeautifulSoup to search for certain data was unsuccessful because the HTML that we were analyzing was primarily just <script> tags that did not contain the data we were looking for. We resolved this issue by integrating Selenium and ChromeDriver into our code to create a Chrome browser environment in which we used BeautifulSoup to obtain the actual HTML without the <script> tags.

However, one constraint we encountered during the data collection process was that the HTML for the Instagram webpage for each user is generated dynamically using JavaScript, even when using Chromedriver. Only code for 12 of the most recent posts for each Instagram user is generated when a request to an Instagram user's webpage is made, and the code for the other posts is generated when the page is scrolled down. Because of this, we were only able to scrape data about the 12 most recent posts for each user.

By scraping the data from these Instagram posts, we divided the information that we have collected into 2 datasets: "data" and "user data". We created the "user data" sheet from an online dataset from HypeAuditor that provided us with the account usernames of the 50 of the highest followed users and we supplemented that information with the number of followers each person had. We created the "data" sheet based off multiple posts from the 50 users we found from the "user data" sheet and scraped the data off every post, which included the variables mentioned in the introduction.

After, creating our datasheets and their conversion into CSV files, we created a new Notebook for our analysis of the data that we have collected and organized. We imported the dataset into the Notebook as a Pandas dataframe for easier manipulation of all the information of each variable for analysis. We have also expanded the "user data" dataset as we wanted to use some of the variables that were specific to each post. We appended the number of comments, the number of likes, as well as the number of followers that the user who posted had for each of the posts. This data was manually added to the dataframe but was the exact same data generated by our data collection code.

We first created histograms comparing different variables from each post to the different variables that signified interaction: likes and comments. This created a visual diagram that allowed us to see the type of distribution of each of the relationships between the variables we had chosen. Then, we began to divide the posts into two categories, posts that had tags and posts that did not. We began to run graphic representations and statistical analyses of the same variables but with the addition of the division of the two categories. We first created dual scatterplots of each of the relationships of the different variables based off the two categories, then created one larger scatterplot for each one that placed the two dual scatterplots side by side for better comparison. We decided to not have ticks on the x-axis as a visual choice as when having such a large dataset, the labels of each point creates way too many x-labels and with 50 different users, there wasn’t a large correlation between each of the points in terms of their location on the x-axis. However, for each of the graphical representations of the dual scatterplots, we have also created the same scatterplot with the points in numerical order from least to greatest in terms of the y-axis.

After our visual representations were created, we ran linear regression tests for these different relationships and calculated different correlation and linear coefficients to see if any of the relationships were statistically significant. We first calculated the correlation coefficient and efficient of determination (r & r^2) and we also found the Pearson and Kendall coefficient for each of the relationships. We also added further information between each of the two variables by creating a method for the estimated slope and the intercept.

Based on the statistical analyses that we calculated from the variables that reside within the Instagram posts, we drew several conclusions.

1. There are only three noticeable correlations between the variables that we measured (tags used, total posts for tags used, tagged accounts, number of followers of tagged posts, posting user followers) and post engagement (number of likes and comments). One correlation was between user followers and likes, which had a Pearson Correlation Coefficient of 0.1369. We expected to see this correlation. Another correlation was between user followers and comments which had a Pearson Correlation Coefficient of 0.1939. This was also expected. The third correlation was between comments and likes which had a Pearson Correlation Coefficient of 0.5282. This was the strongest correlation, but this does not necessarily mean that one influences the other. We believe

this to be more of a natural occurrence of similar levels of engagement of both likes and comments on a single post.

2. Factors such as tagged user followers and tagged post totals do not have a noticeable effect on post engagement in the posts included in our dataset. From this conclusion, we developed the hypothesis that this is because all of these posts are from accounts with large followings (millions to hundreds of millions of followers) and the increase in engagement from these variables may not have a large effect on posts by accounts that would have high levels of engagement regardless.

Based on the data analysis, these are some of the things we would do differently.

1. We would create subsets of data with smaller bins of follower counts (i.e. posts from accounts with 0-5M followers, 5-10M followers, etc). By separating the data between ranges of followers, we could better visualize the change in these different variables for each user, without being affected by the mass variation of followers and likes that come with 50 users.
2. We would make better use of qualitative variables (tags, tagged users) and collect data on more posts so that the dataset contains many posts with the same tags and tagged users (as opposed to only one or two posts that use the same tag or tag the same user). By creating a dataset that shares similar variables, it would be much easier to see the relationships between the variables that have a greater impact, like the number of comments and followers.
3. We would develop a way to collect data on all of the posts for one user so that we can compare posts by a single user against each other rather than against posts by many other users. This approach may have aided us in seeing less variation and higher correlation coefficients within the scatterplots as then the data would not be skewed by other users that vary largely with these different variables.
4. We would collect data on accounts with smaller followings that may be actively utilizing tags and tagging users to help grow their following and engagement. We would potentially choose small companies or online personalities so their growth within the past months or year would be much more easily visualized.