Instagram Web Scraper and Analyzing Makeup Content Creators:

A Data Report

Intisar L. Muhammad

University of Maryland, College Park

INST 447: Data Sources and Manipulation

Professor Aditya Ravindra Bhat

May 15, 2021

Table of Contents

Introduction

Purpose

During the last few years of the social media age, the rise of instagram influencers has become a phenomenon. For every niche market, there are influencers who rely on instagram to make money or even just to share their content with a larger audience. There are many influencers who have made content creating their full time job. In the makeup industry, content creators and influencers have figured out how to succeed on Instagram and beat the algorithm so that their posts reach a high audience. Sponsors reach out to these creators with influential followings so that they can sell their products and in return, the influencers make their money that way. For this project I want to investigate how to get the best engagement on Instagram as a makeup influencer. Examining the patterns of these influencers will help myself and others understand the steps to take in order to reach higher audiences and what the characteristics of viral posts are. The goal of this project is to be able to find the average time most creators post, which hashtags produce the most likes, and the type of content that produces the most likes, for example, single photos, carousels, or videos.

Motivation

Over winter break 2021, I began planning out ways to become an instagram content creator in the makeup niche market. I made a new account and designated that I wanted it to be a business account so that I could look at the analytics of my profile. I knew that the Instagram algorithm has been changing recently and making it harder for small creators to show up on other users' timelines. This caused me to wonder how to go about starting a new content creating page and how to gain followers, cater to the right audience, and beat the algorithm. Since I haven't

made any posts yet, there's no way for me to view my analytics because I don't have any yet. I figured that other people were also having this same problem and might be wondering what time is best to post your content, which hashtags to use, and what type of media to post.

My final research question is: what are the posting patterns of Instagram makeup creators that help them beat the algorithm?

Data

Variable Name	**Data Type**
username	string
link	string
type	string
likes	integer
age	date/time
caption	string

Methods

Collection

Since the topic that I covered is very specific, I wasn't able to find a dataset online that exactly matched what I wanted to research. There were many datasets that collected data from big influencers and beauty was listed as one of the categories, but I was specifically researching makeup influencers and content creators. The API that I built allowed me to gather information from each post such as the username, unique link/url, number of likes, caption, date/time posted, and the type of post it is (photo, video, carousel). I manually picked 100 instagram accounts owned by makeup influencers in order to put them in a list. This step wasn't too difficult because

I already followed a large amount of makeup influencers. My API gathered data from their most recent 10 posts so that I had enough data to conduct reliable research. Most of the variables had strings stored in them except the time of the post. The time of the post was stored in date/time format. The variables are as follows: username, followers, link, type, likes, age , and caption. The variable type will represent if it's a picture or video.

API Specifics

For this project, I will be using Selenium to collect my data. Selenium is a python library that provides users with simple API and allows you to access a web browser through python. I will be accessing Instagram through Google Chrome. I then created a program that acts as an API that takes a username as a parameter and will grab the links of the last 10 posts from a user in one function. The second function I created will return the data from each post/link. In another python script, I imported those two functions and created a list of each makeup influencer's username to pass through the functions. The data will be appended into a list which created a DataFrame out of so that I could apply data analysis techniques to.

I also created a new variable called time_of_day. This variable categorized the time of the post into 3 categories: morning, afternoon, and evening. This helped me better understand if a user should be making early or late posts in addition to the time they should be posting. I did this by using the pd.cut() technique from class and creating bins for the times of the day. For example, 6-12 am is morning, 12-4 pm is afternoon, 5-9 pm is evening, and an additional category is late night which is any post made after 9 pm and before 6 am.

Data Cleaning

The data cleaning that I performed was extracting all of the hashtags in a caption and putting them into a new column named 'hashtags. I did this by using a regular expression which

was very helpful. I created a new function and used the regular expression: re.findall('#[A-Za-z]+', comment). I also used another regular expression function to retrieve only the time when the post was created which will help me with post statistics.

New Variables

The new variables I created are ‘time_posted’ and ‘hashtags’. I did this with pandas by extracting things from one column and putting them in a new one with a new variable name. Another variable I created was time_of_day which will group the posts by the time of day they were posted. To achieve this, I used the pd.cut() function.

Analysis

Analytic Steps

In order to analyze the data using pandas, I had to recode the age column so that it showed the time instead of how many days or weeks ago it was posted. I created a new column and called it time_posted. I did this so that I can figure out what time of the day is best to post on your Instagram. In pandas, I also used a regular expression in order to capture each hashtag from their caption so that I could analyze which ones are most used and produce the most likes as well. Since the process of data collection took some time, I split the usernames up just to make sure the process was being done correctly. In this case, I had to combine multiple datasets that were produced which wasn't hard to do because I didn't have to recode any variables. The final data set ended up looking like this:

	username	link	likes	age	caption	time_posted	hashtags	hours	time_of_day
0	meliysabel	https://www.instagram.com/p/CFNSCFsnjnp/	13713.0	2020-09-16T18:30:08.000Z	meliysabel\nI HAD to 🇲🇽😇✨Feliz día de la Inde...	18:30	[#jamescharlespallette, #mexicanmakeup, #hudab...	18.0	evening
1	meliysabel	https://www.instagram.com/p/CEaNztqH74-/	NaN	2020-08-27T22:37:11.000Z	meliysabel\nHeeeey.... how y'all doin 😁😁I kno...	22:37	[#jamescharlespalette, #dipbrow, #euphoriamake...	22.0	late night
2	meliysabel	https://www.instagram.com/p/CDNAAdYnCyo/	8302.0	2020-07-28T22:49:50.000Z	meliysabel\nHi it's me back with periwinkle lo...	22:49	[#pastelmakeup, #wakeupandmakeup, #hauslabs, #...	22.0	late night
3	meliysabel	https://www.instagram.com/p/CCbxKOlAFpv/	NaN	2020-07-09T19:57:57.000Z	meliysabel\nHere's how I do my brows, I haven'...	19:57	[#browtutorial, #makeuptutorial, #eyebrowtutor...	20.0	evening
4	meliysabel	https://www.instagram.com/p/CCHO_0HHxan/	11773.0	2020-07-01T20:33:59.000Z	meliysabel\nmy favorite looks to do 🍎✨(swipe ...	20:33	[#morphebrushes, #jamescharlespalette, #colour...	20.0	evening

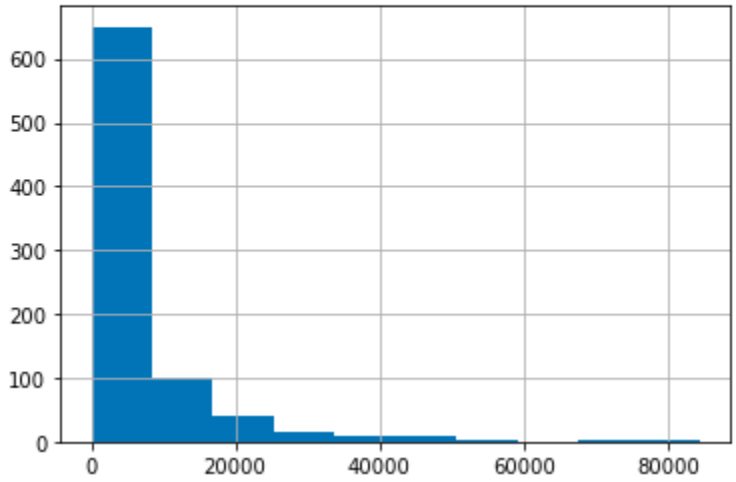
Summary Statistics

I have two float objects which are likes and hours. These variables are the amount of likes each post has and the time (by hour) that the post was created. For likes, there are 830 values and the mean is 5,816, the standard deviation is 9,854, the minimum is 11 likes, and the maximum is 84,388 likes. For hours, there are 885 values with the mean being 16.47, the standard deviation is 5.98, the minimum is 0, and the maximum is 24. This means that most of the posts were posted at around 4 p.m., with the 'earliest" in a 24 hour day being exactly midnight, and also the latest being right before midnight.

Likes histogram:

```
insta_scrape['likes'].hist()
```

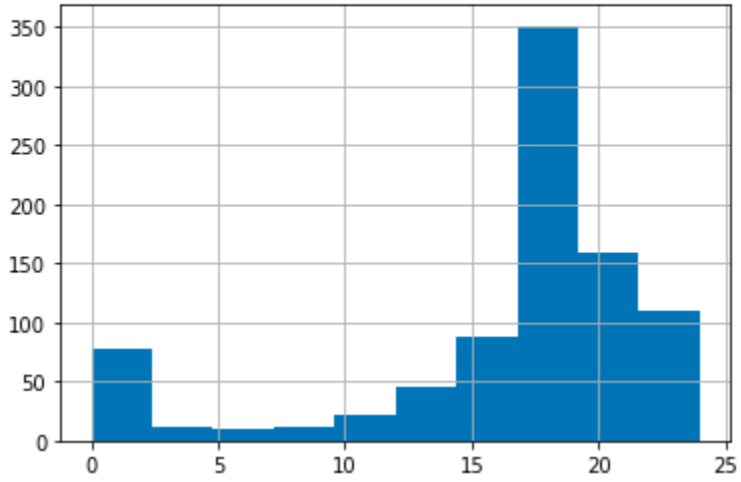
```
<AxesSubplot:>
```



Hours of the day histogram:

```
insta_scrape['hours'].hist()
```

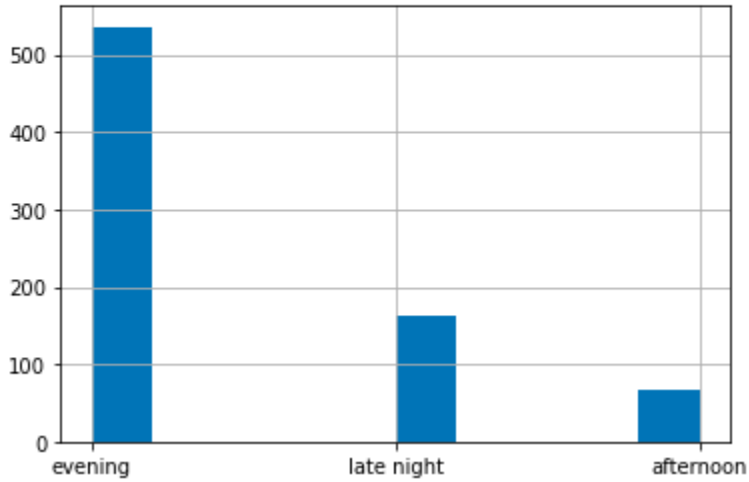
```
<AxesSubplot:>
```



Time_of_day histogram:

```
insta_scrape['time_of_day'].hist()
```

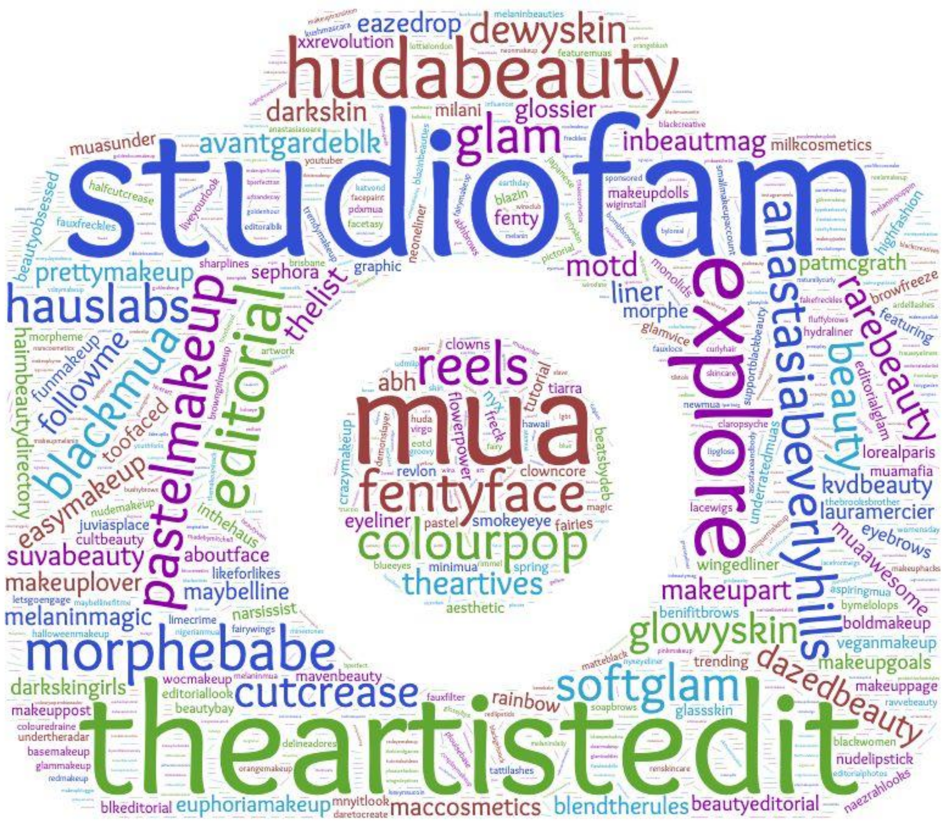
```
<AxesSubplot:>
```



Word Cloud

In order to figure out which were the most commonly used hashtags that makeup content creators used, I created a world cloud that displays the most used words as large text and the less used words as small text to get a visual. To do this, I appended each hashtag into one big list and

flattened the list to remove extra brackets by using a package called itertools. Then, I removed the ‘#’ with a regular expression that stripped the words. After that, I saved the list into a text file and uploaded it to wordclouds.com. This website made an image that shows the most used words, least used words, and everything in between. Here is the word cloud image in the shape of a camera:



Results

The information from the histograms revealed that most makeup content creators post their pictures during the evening and the mean of the hours column showed that the time is

usually around a little bit after 4.pm. Also, the amount of likes for each post is mostly under 10,000 with the bigger creators' most viral posts getting between 50,000 to 80,000 likes. The time of day histogram revealed that most creators post in the evening which is some time after 4 p.m. and before 9 p.m. This makes sense because a lot of people are coming home from work and school around that time which is when they start winding down and being more active on social media. The world cloud was the most interesting result of the project to me because the image shows the most used hashtags were mua, studiofam, theartistedit, explore, hudabeauty, etc. I also saw that a lot of creators hashtagged really famous makeup brands such as Fenty Beauty, ColourPop, Morphe, and more.

Conclusions

The posting pattern of Instagram makeup content creators that help them beat the algorithm seem to be posting in the evening between 4-9 p.m., using popular hashtags which you can find on other viral creator's posts to get inspiration, and if you're posting in the evening, try to do it closer to 4 p.m. Using all of these steps and being consistent, a content creator will hopefully be able to build a following and grow their account to become successful. In the future, if there was more time to work on this project, I would see how often each user was posting by comparing the dates that they posted to see how often a creator should be posting to achieve success.

Appendix

Extra Details

Something that will help a reader understand my project better is that for some postings that were solely videos, the likes could not be collected for it because likes were not displayed,

only views. Whenever I tried to extract the views, I kept getting an error saying the attribute couldn't be found.

References

In order to build my personalized web scraper, I had to conduct a lot of research including how to use Selenium and how to gather information from Instagram using Xpath and class names. To get started, I read an article by John Naujoks entitled “Tutorial: Web scraping Instagram’s Most Precious Resource- Corgis”. This helped me set up my Python program and get started with Selenium by showing me the basic steps and how to use it. This article can be found here: https://medium.com/swlh/tutorial-web-scraping-instagrams-most-precious-resource-corgis-235bf0389b0c