# PROJECT ASSIGNMENT - 4

# TITLE: DATA ANALYTICS RESEARCH PROJECT

# DATASET: TOP 100 MOST FOLLOWED INSTAGRAM USERS

**Semester:** Fall 2022

**Course:** INFS 580 – DL2

**Name:** Sarayu Soma

**G Number:** G01380103

**Major:** MS in Information Systems

# TABLE OF CONTENTS

# Top 100 most followed Instagram Users

George Mason University
INFS – 580 – DL2
Sarayu Soma
ssoma@gmu.edu

## ABSTRACT

With the increase in demand for entertainment and networking, more than half of the population of mobile phone users has an account in one or more social media sites. Since then, there have been extensive researches on social networking platforms. This research seeks to determine the most well-liked Instagram users based on their followings and other characteristics. The user's average post likes, audience engagement rate, frequency of posts, etc. are all recorded in addition to total followers, and data on the top 100 influencers is contained in the dataset. Data analytics, statistical analysis, and visualizations are used in this study to gain insights about which user has the most followers. In this paper, we examine more closely at the significant ideas that emerged from the analysis, explain about the study's limitations, and present the significant benefits. The influencers may find this information helpful in luring new followers and expanding their fan base. On the basis of the dataset's data, research questions are developed, and different methodologies are used to seek answers. Python and R are applied for the exploratory analysis and visualization. Excel was employed to manually clean the data because the dataset had fewer rows. A few queries were also executed on the data to produce outcomes. The main purpose of this data collection is to investigate and better understand some of the elements that influencers believe to be helpful in growing their followers.

**Keywords –** influencer, followers, followees, engagement rate, average likes

# I.INTRODUCTION

A social media platform is any firm that offers a service to the general public for the purpose of disseminating communication, expression, information, or other content (usually content in the form of messages, videos, pictures, and/or sound files) [6] . YouTube, Facebook (formerly known as Meta), Twitter, NextDoor, LinkedIn, Instagram, Google, Reddit, Facebook Messenger, WeChat, TikTok, Weibo, Wikipedia, Snapchat, and Pinterest are just a few examples of "Social-Media Platforms."One such photo- and video-sharing social networking platform is Instagram, which is owned by the American business Meta Platforms. Users of the app can upload media that can be altered using filters, arranged by hashtags, and categorized by location. Public or pre-approved followers may share posts.

Instagram is one of the most widely used social media platforms globally, used by every teenager and majority of adults. This site offers users entertainment, and by using it, users can expand their networks as well. Since it first launched, this platform has undergone numerous adjustments, and multiple new features have been added to occasionally amuse users. This study aids influencers in comprehending the methods by which well-known people earn followers.

There are 11 columns in the dataset, and they contain all NOIR data kinds [4]. Since the title indicates that the top 100 most followed users are stated, the data contains a total of 100 records. Channel Information, Influencer Score, Follower Count, Average Likes, and other data are listed in the columns. Deep analysis of the influencer score column will provide us with insight into the User's behavioral patterns that appeal to the audience. The relationship between the amount of followers and interaction rate can be understood by examining the new post average likes and 60 day engagement rate columns. Analysis of the posts and the total likes column can reveal user engagement and popularity. The analysis's findings can be used to generate outcomes that assist influencers in gaining more fans.

# II.RESEARCH QUESTIONS & PROBLEM STATEMENT

**Problem Statement:**

The research problem is to analyze the behavioral patterns of top instagram followers containing users such as their number of followers, posts, average likes earned, activity in past 60 days etc. This study contributes to our understanding of the practices that aspiring influencers can use to increase their fan base. They benefit from knowing how to gain attention and how the Instagram users with the most followers are faring in order to keep their names at the top of the list.

**Research Questions:**

By utilizing the insights obtained by applying deep analysis on different columns, this paper will explore the following research questions

a. Most of the top influenced instagramers are from which country?
b. What is the relationship between 60_day_eng_rate and posts amongst top 20 followers?
c. What proportion of the influencers has their Education Level higher than High School?

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rank(Ordi | channel_info(N | influence_score( | posts(Ratio) | followers(Rati | avg_likes( | 60_day_eng_rate( | new_post_avg_like(I | total_likes | country(No | Education Level(Ordinal) | | |
| 2 | 1 | cristiano | 92 | 3.3k | 475.8m | 8.7m | 1.39% | 6.5m | 29.0b | Spain | Middle School | | |
| 3 | 2 | kyliejenner | 91 | 6.9k | 366.2m | 8.3m | 1.62% | 5.9m | 57.4b | United Stat | High School | | |
| 4 | 3 | leomessi | 90 | 0.89k | 357.3m | 6.8m | 1.24% | 4.4m | 6.0b | | High School | | |
| 5 | 4 | selenagomez | 93 | 1.8k | 342.7m | 6.2m | 0.97% | 3.3m | 11.5b | United Stat | High School | | |
| 6 | 5 | therock | 91 | 6.8k | 334.1m | 1.9m | 0.20% | 665.3k | 12.5b | United Stat | BA | | |
| 7 | 6 | kimkardashian | 91 | 5.6k | 329.2m | 3.5m | 0.88% | 2.9m | 19.9b | United Stat | High School | | |
| 8 | 7 | arianagrande | 92 | 5.0k | 327.7m | 3.7m | 1.20% | 3.9m | 18.4b | United Stat | High School | | |
| 9 | 8 | beyonce | 92 | 2.0k | 272.8m | 3.6m | 0.76% | 2.0m | 7.4b | United Stat | Middle School | | |
| 10 | 9 | khloekardashia | 89 | 4.1k | 268.3m | 2.4m | 0.35% | 926.9k | 9.8b | United Stat | High School | | |
| 11 | 10 | justinbieber | 91 | 7.4k | 254.5m | 1.9m | 0.59% | 1.5m | 13.9b | Canada | High School | | |
| 12 | 11 | kendalljenner | 90 | 0.66k | 254.0m | 5.5m | 2.04% | 5.1m | 3.7b | United Stat | High School | | |
| 13 | 12 | natgeo | 91 | 10.0k | 237.0m | 302.2k | 0.07% | 159.3k | 3.0b | United States | | | |
| 14 | 13 | nike | 90 | 0.95k | 234.1m | 329.0k | 0.08% | 181.8k | 313.6m | United States | | | |
| 15 | 14 | taylorswift | 91 | 0.53k | 222.2m | 2.4m | 1.01% | 2.3m | 1.3b | United Stat | High School | | |
| 16 | 15 | jlo | 89 | 3.2k | 220.4m | 1.7m | 0.62% | 1.4m | 5.3b | United Stat | High School | | |
| 17 | 16 | virat.kohli | 87 | 1.4k | 211.8m | 3.5m | 0.96% | 2.0m | 4.9b | India | Intermediate level | | |
| 18 | 17 | nickiminaj | 90 | 6.4k | 201.6m | 2.1m | 0.53% | 1.0m | 13.5b | United Stat | High School | | |
| 19 | 18 | kourtneykardas | 89 | 4.4k | 195.2m | 1.8m | 0.67% | 1.3m | 7.7b | United Stat | BA | | |
| 20 | | mileycyrus | 89 | 1.2k | 181.5m | 1.3m | 0.51% | 913.6k | 1.6b | | High School | | |
| 21 | 20 | neymarjr | 90 | 5.3k | 177.1m | 2.7m | 1.09% | 1.9m | 14.1b | Brazil | Middle School | | |
| 22 | 21 | katyperry | 92 | 2.0k | 170.3m | 715.0k | 0.16% | 265.1k | 1.5b | | High School | | |
| 23 | 22 | kevinhart4real | 88 | 8.2k | 152.0m | 522.0k | 0.08% | 115.2k | 4.3b | United Stat | High School | | |
| 24 | 23 | zendaya | 87 | 3.5k | 150.7m | 5.8m | 3.17% | 4.8m | 20.6b | United Stat | BA | | |
| 25 | 24 | iamcardib | 75 | 1.6k | 140.5m | 3.1m | 1.10% | 1.5m | 5.0b | United Stat | High School | | |

# III. LITERATURE REVIEW

The dataset that I chose is regarding the top followers list on Instagram. It includes the list of names of the celebrities who has most followers and few of their activities such as their weekly number of posts, average likes, number of followers etc. Instagram is a free Social Media platform that is popular now-a-days for sharing photos and videos. There are many more features in this app like Short videos, Reels, IGTV etc which made it much more entertaining for the users. The dataset is related to this app where the activities of the people having most followers are analyzed and constructed in the form of spreadsheet.

There have been a few studies on this subject. However, rather than focusing specifically on Instagram, the majority of studies are conducted on the subject of social media. So, a few studies that relate to the dataset obtained are listed below.

## 1. Research in the Instagram Context : Approaches and Methods

The research paper belongs to Chen Yang, Faculty of Humanities, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, International Education College, Henan Polytechnic University, Jiaozuo, China [1]. The growth of social media is first described in the journal, along with its definition, how it has occupied a significant portion of our lives, and the unique characteristics that make it appealing to so many people. There were a few studies that were conducted that explained why the percentage of individuals using the Internet everyday climbed so quickly in the years 2017, 2018, and 2019. The study then turns its attention to Instagram in the subsequent sections, where various facets of it are covered. It is initially discussed how Instagram got its start and how it evolved into Facebook. The author of this article conducted a search in Web of Sciences and discovered a total of 1226 publications in order to comprehend the number of studies conducted on Instagram.

**Figure-1.** Numbers of publications in "Instagram" research on WoS (2010-2019)

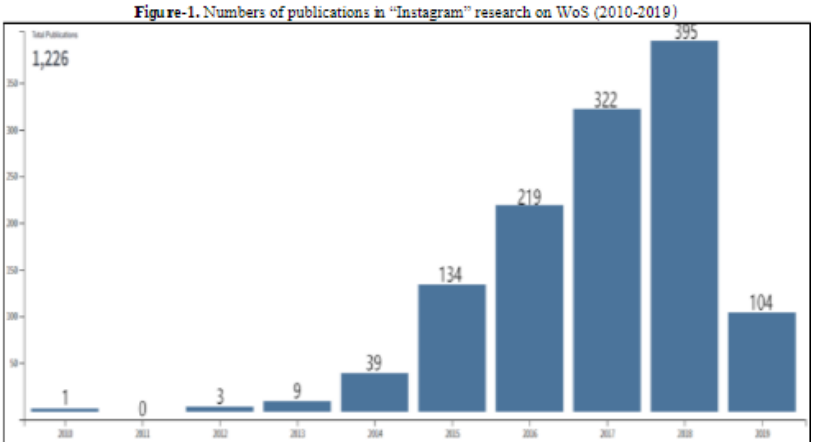


**Figure-2.** Proportion of "Instagram" research in different fields on WoS (2010-2019)

| Field: Web of Science Categories | Record Count | % of 1,226 |
|---|---|---|
| COMPUTER SCIENCE INFORMATION SYSTEMS | 202 | 16.476 % |
| COMPUTER SCIENCE THEORY METHODS | 158 | 12.887 % |
| COMMUNICATION | 156 | 12.724 % |
| ENGINEERING ELECTRICAL ELECTRONIC | 132 | 10.767 % |
| COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE | 98 | 7.993 % |
| BUSINESS | 79 | 6.444 % |
| COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS | 74 | 6.036 % |
| SOCIAL SCIENCES INTERDISCIPLINARY | 67 | 5.465 % |
| EDUCATION EDUCATIONAL RESEARCH | 62 | 5.057 % |
| PSYCHOLOGY MULTIDISCIPLINARY | 59 | 4.812 % |

Here the author argues that Instagram in considered not just as an application used to share videos and photos but also as a venue for promoting businesses, which will increase interest in the subject and the number of people conducting research on it. Even if various researchers examine the data in different ways, they all follow the same traditional quantitative and qualitative research methodologies to obtain the analysis. The procedures used and the outcomes are thoroughly disclosed in the next section [1].

Quantitative analysis makes use of the mathematical models in order to obtain the results related to statistical data. A few studies conducted by different writers are listed, and the steps they took to achieve their goals are described in detail. Additionally, researchers frequently employ questionnaire methods to gather information for quantitative analysis. Furthermore, qualitative analysis techniques are addressed, which rely on the analyst's judgment and experience rather than experiments or numerical results to determine the nature of the subject. It perfectly expresses the author's viewpoint on the subject. Although these approaches are specified separately, they are both written for a subject in order to provide a coherent explanation. One gives the other method closure. These techniques provide me a clear understanding of how to discover the answers to my research questions and what all the steps I need to do to acquire the results.

In this article, Quantitative and Qualitative analysis processes are explained and how they are used in order to retrieve the data from the applications is also mentioned in detail. The research questions also deals with the processes which are different as compared to the Instagram Dataset (My Dataset). So it can be discussed that the way of style in which the author looks into the data completely differs from individual to individual which results in the variation of research questions from article to article even though the topics are similar.

## 2. In a World of Social Media: A Case Study Analysis of Instagram

This Research article was drafted by Dr. Daryl D. Green1*, Dr. Richard Martinez1, Amalan Kadja2, Lauran Evenson2 Lisa MacManus2, Stephanie Dirlbeck2 1Professors, Oklahoma Baptist University, 500 W. Oklahoma's University Drive in Shawnee [2]. 2 Oklahoma Baptist University MBA student, 500 W. Oklahoma's University Drive in Shawnee. The definition of social networks and an explanation of their history and current state are provided in the introduction section. It is asserted that businesses must use digitalization strategies to maintain contact with clients and boost sales. Following the definition of Instagram, the significance of Instagram for business marketing tactics is investigated in the subsequent sections.

By analyzing the company's strategy, structure, and organizational culture, this study delves further into Instagram's infrastructure. A strategic analysis includes elements like the firm's competition and the benefits and drawbacks of the organization's operations. Using SWOT analysis and competitor analysis approaches, information about the competitor organization is gathered and examined in light of their strengths and shortcomings. Instagram was compared to Snapchat and Pinterest in this article, and the findings from this investigation are listed below [2].

**Table1.1.** *Competitor Analysis*

| Critical Elements | Instagram | Snapchat (Industry Competitor) | Pinterest (Industry Competitor) |
|---|---|---|---|
| **Unique Selling Proposition** | Promotes itself as an instant photo sharing/video application. | Promotes itself as a video/picture sharing messaging application. | Promotes itself as a web and mobile sharing application. |
| **Value to Prospective Customers** | Customers can view other people's lives at their fingertips. Can also shop and see the latest trends. | Customers can instant message pictures/ videos and send money in seconds. | Customers can search the world wide web from the application, utilizing images on a focused scale. |
| **Core Competencies** | Instagram specializes in advertising and picture sharing. | Snapchat specializes in video/picture sharing and has recently started with advertising. | Pinterest is a social media, bookmarking application that allows users to discover and share creative ideas. |
| **Positioning in the Market (i.e., top, middle, bottom)** | Top | Top | Middle |
| **Marketing Mix**<br>**Product**<br>**Price**<br>**Place**<br>**Promotion** | Product: Variety of services<br>Price: Free<br>Place: Online<br>Promotion: Ads | Product: Variety of services<br>Price: Free<br>Place: Online<br>Promotion: Ads | Product: Variety of services<br>Price: Free<br>Place: Online<br>Promotion: Ads |
| **Overall Impressions in the Market** | Instagram is considered an industry leader in social media. It allows people to see other's lives, and it is utilized by companies to promote their products. | Snapchat is considered an industry leader in messaging through picture sharing. | Pinterest is considered an industry leader in web searching. For example, it has allowed people to search recipes and DIYs conveniently. |

**Table1.2.** *SWOT Analysis*

| | Instagram | Snapchat | Pinterest |
|---|---|---|---|
| **Strengths** | Multiple filters available to enhance and customize pictures<br>Strong and supportive parent company<br>800+ million users<br>First social media app for pictures only | Constant refreshing of pictures<br>Growing application for users<br>Innovative platform for picture sharing<br>Does not store historical photos | Diverse user group<br>Virtual bulletin boards<br>User friendly |
| **Weaknesses** | Not available on PC operations; i.e., interact online like you can on your device<br>Required to follow to view content<br>Weakened privacy firewall | Lack of discreetness<br>Not a diverse product<br>Pictures are only available for a small amount of time | Possible gender biases<br>Susceptible to spam pictures<br>Based on user uploads |
| **Opportunity** | Advertisement opportunities<br>Technology development | Advertising opportunities<br>Growth and product enhancement | Growing business with opportunity for advertisement<br>Social networking<br>Linked with Facebook and Twitter |
| **Threats** | Faces strong competition<br>Possible issues with photo rights | Negative publicity<br>Legal ramifications | Copy-cat websites<br>Spam<br>Other social media sites adopting bookmarking features |

The conclusion is that businesses can enhance their sales by using Instagram into their marketing plans. It claims that Instagram has quickly gained a large following and elevated to the top spot among social media platforms. However, it also makes the case that it is still having trouble connecting with older generations.

This article's main concept is around the techniques used to extract data from applications like Instagram, Snapchat, and Pinterest. Because of this, the research questions also focus on the analysis that was utilized, the data that was gathered, etc., which is very different from the approach that I took for my research.

### 3. Marketing through Instagram Influencers : Impact of number of followers and product divergence on brand attitude

Marijke De Veirman, a Belgian student at Ghent University's Faculty of Political and Social Sciences, is the author of this journal [3]. It describes the tactic of promoting the company's brand value by using influencers' notoriety. It emphasizes on the idea that celebrities with large

followings are more popular and, as a result, are the target for businesses looking to draw in clients.

It makes the case that the number of followers an influencer has defines their popularity, which makes them more appealing to customers. However, it is noted that even while a celebrity has a larger following, this doesn't always assist the business promote its goods. It is crucial to determine whether the celebrity is regarded as a leader by the followers or if their content alone is what draws people in. The article focuses on how an organization should select an influencer in order to engage the widest possible audience and boost brand sales. Two studies were conducted to provide further detail, and the findings are discussed in this publication [3].

Study 1 examined into which Instagram influencer would be best for a company's promotion. The quantity of followers and followers determines this. An influencer with a larger following is thought to be more likeable. As a result, influencers with large followings are sought after by businesses. On the other hand, other researchers dispute this. Having a large following does not automatically imply that one is able to influence all of those followers. The influencer is popular if they have a large following, but it doesn't indicate that everyone like the content they post. Therefore, a company must pick an influencer who is not only well-liked but also whose opinions are respected and adopted by others. Another factor taken into account is the amount of followees. According to some "laws" in popular literature, the number of followers to followees has an effect on an influencer's popularity. People who have more followees than followers exhibit true attributes of opinion leadership because they can learn more about all the factors by following more people. The audience could get the impression that certain followers are false when there are many followers but few followees. Thus, it is claimed that no study has established this problem to date [3].

One of the research questions in this journal almost matches with the one that I wrote. The author asked the question in this article, "Is there any impact of number of followers and followees on product divergence?" This is comparable to the relationship between 60 day eng rate and postings that I saw in my study project. Other research questions, however, differ.

# IV. MATERIALS AND METHODS

Kaggle was used to collect this dataset [5]. It includes a variety of information, including the user's Instagram account information, number of followers, posts, average likes, engagement rate over the previous 60 days, and educational background. The dataset was in CSV format, however for ease of processing, it was changed to Excel format. The content in the dataset is drawn from what existed in May 2022. Since the top 100 users are stated, there are 100 rows in total. The dataset had a small number of records, making data analysis simple.

The data were cleaned and checked for errors and null values as the initial step in the investigation. Considering that there are less records, the clean - up is performed manually in Excel. To prevent receiving inaccurate results, rows with a high percentage of null values are eliminated. Numerous cells had excessive blank spaces, misspelled words, or lacking units beside data; all of these were fixed to prevent errors from occurring in frequency counts. The dataset is loaded into the platform after the data has been cleaned, and statistical summary findings are obtained using Python. R was used to create the visuals and find answers to the study questions. The dataset was thereafter subjected to a few queries, and results were produced using SQL. Visualizations were constructed using techniques in Excel as well.

# V. DATA PRE PROCESSING METHODS

## 5.1 Data Cleaning using Excel

Since the "Top 200 most followed Instagram users" dataset only has 200 rows, cleaning up the data directly in the excel sheet is simpler than loading it and running programs on it. The entire file was examined, and the following changes were applied to a selected few cells:

K23 | High School

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 23 | 22 | kevinhart4real | 88 | 8.2k | 152.0m | 522.0k | 0.08% | 115.2k | 4.3b | United States | High School |
| 24 | 23 | zendaya | 87 | 3.5k | 150.7m | 5.8m | 3.17% | 4.8m | 20.6b | United States | BA |
| 25 | 24 | iamcardib | 75 | 1.6k | 140.5m | 3.1m | 1.10% | 1.5m | 5.0b | United States | High School |
| 26 | 25 | ddlovato | 88 | 0.08k | 139.1m | 1.1m | 0.27% | 363.4k | 91.3m | United States | High School |
| 27 | 26 | badgalriri | 88 | 4.8k | 135.3m | 3.7m | 0.02% | 133.4k | 17.9b | United States | High School |
| 28 | 27 | kingjames | 86 | 2.3k | 130.9m | 2.1m | 0.92% | 1.2m | 4.9b | United States | High School |
| 29 | 28 | theellenshow | 87 | 10.0k | 125.1m | 420.5k | 0.07% | 81.7k | 4.2b | United States | High School |
| 30 | 29 | realmadrid | 90 | 6.9k | 123.4m | 996.2k | 0.48% | 588.3k | 6.8b | Spain | |
| 31 | 30 | champagnepapi | 85 | 5.2k | 119.6m | 1.7m | 1.10% | 1.3m | 9.0b | Netherlands | High School |
| 32 | 31 | chrisbrownoffic | 86 | 7.3k | 118.5m | 463.2k | 0.22% | 255.9k | 3.4b | United States | High School |
| 33 | 32 | fcbarcelona | 90 | 10.0k | 111.4m | 1.2m | 0.64% | 706.3k | 11.6b | United Kingdom | |
| 34 | 33 | billieeilish | 73 | 0.69k | 105.2m | 8.5m | 5.02% | 5.2m | 5.9b | United States | High School |
| 35 | 34 | dualipa | 74 | 1.3k | 85.9m | 2.1m | 1.26% | 1.1m | 2.6b | United Kingdom | High School |
| 36 | 35 | gal_gadot | 85 | 1.7k | 85.6m | 1.4m | 0.69% | 586.5k | 2.3b | United States | Middle School |
| 37 | 36 | vindiesel | 88 | 1.8k | 82.3m | 1.4m | 0.60% | 482.3k | 2.5b | United States | High School |
| 38 | 37 | nasa | 88 | 3.6k | 81.3m | 1.2m | 1.53% | 1.2m | 4.2b | United States | |
| 39 | 38 | priyankachopra | 85 | 3.6k | 81.1m | 1.6m | 1.00% | 802.9k | 5.6b | United States | BA |
| 40 | 39 | lalalalisa_m | 70 | 0.87k | 80.9m | 5.8m | 9.00% | 7.2m | 5.1b | South Korea | High School |
| 41 | 40 | Shakira | 88 | 2.0k | 76.1m | 975.1k | 0.41% | 304.7k | 1.9b | Columbia | Middle School |
| 42 | 41 | snoopdogg | 86 | 10.0k | 75.3m | 203.7k | 0.17% | 125.8k | 2.0b | United States | Secondary Schoc |
| 43 | 42 | gigihadid | 85 | 3.3k | 75.3m | 2.5m | 2.56% | 1.9m | 8.2b | United States | BA |
| 44 | 43 | davidbeckham | 86 | 1.5k | 74.9m | 1.2m | 0.46% | 340.9k | 1.9b | United States | High School |
| 45 | 44 | shraddhakapoo | 81 | 1.9k | 73.9m | 1.6m | 1.17% | 859.9k | 3.0b | India | BS |
| 46 | 45 | victoriassecret | 88 | 2.9k | 73.2m | 147.0k | 0.04% | 29.8k | 423.5m | United States | |
| 47 | 46 | k.mbappe | 86 | 1.2k | 72.7m | 2.5m | 2.16% | 1.6m | 2.8b | French | Middle School |

The dataset comprises users from the top 100 Instagram followers list, not just individuals but also fan pages, team pages, etc. Therefore, the Education level (Column K) for such an account is empty, and all of the empty cells in the Education Level column are replaced with a null value.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 81 | 80 | paulpogba | 80 | 1.3k | 55.2m | 1.4m | 0.85% | 462.6k | 1.8b | France | Secondary School | |
| 82 | 81 | iamzlatanibrahimov | 86 | 0.87k | 55.1m | 1.5m | 1.53% | 837.1k | 1.3b | United Kingdom | High School | |
| 83 | 82 | leonardodicaprio | 86 | 1.7k | 54.6m | 395.5k | 0.20% | 107.6k | 669.3m | United States | BS | |
| 84 | 83 | juventus | 87 | 10.0k | 54.5m | 194.8k | 0.25% | 135.5k | 1.9b | Spain | - | |
| 85 | 84 | zacefron | 86 | 0.66k | 54.5m | 2.3m | 8.18% | 4.4m | 1.5b | United States | High School | |
| 86 | 85 | bellahadid | 79 | 3.2k | 54.1m | 1.1m | 2.06% | 1.1m | 3.6b | United Kingdom | High School | |
| 87 | 86 | tatawerneck | 86 | 5.6k | 53.9m | 959.8k | 0.51% | 266.5k | 5.4b | Brazil | Middle School | |
| 88 | 87 | beingsalmankhan | 76 | 1.2k | 53.5m | 1.4m | 1.33% | 693.9k | 1.6b | India | Intermediate level | |
| 89 | 88 | robertdowneyjr | 86 | 0.42k | 53.4m | 3.0m | 2.05% | 1.1m | 1.3b | United States | High School | |
| 90 | 89 | sunnyleone | 84 | 4.6k | 53.4m | 764.0k | 0.57% | 301.3k | 3.5b | India | BA | |
| 91 | 90 | ladygaga | 83 | 3.6k | 53.2m | 1.4m | 1.02% | 531.6k | 5.1b | United States | High School | |
| 92 | 91 | dishapatani | 74 | 2.1k | 53.0m | 1.6m | 1.85% | 971.7k | 3.4b | United States | BA | |
| 93 | 92 | sergioramos | 87 | 2.2k | 52.8m | 1.1m | 1.06% | 557.9k | 2.5b | French | High School | |
| 94 | 93 | jbalvin | 87 | 10.0k | 52.8m | 878.0k | 0.66% | 340.5k | 8.8b | United States | Middle School | |
| 95 | 94 | mosalah | 22 | 0.84k | 52.5m | 1.8m | 2.18% | 1.1m | 1.5b | Italy | High School | |
| 96 | 95 | ayutingting92 | 85 | 10.0k | 52.4m | 147.3k | 0.11% | 56.3k | 1.5b | Indonesia | Middle School | |
| 97 | 96 | 433 | 79 | | 51.2m | | 1.52% | 774.2k | 8.9b | | | |
| 98 | 97 | hudabeauty | 82 | 2.4k | 50.8m | 186.4k | 0.04% | 16.8k | 453.6m | United States | High School | |
| 99 | 98 | adele | 84 | 0.42k | 50.7m | 4.7m | 3.82% | 1.9m | 2.0b | United States | High School | |
| 100 | 99 | michelleobama | 85 | 0.60k | 50.7m | 700.5k | 1.22% | 611.2k | 421.7m | United States | BA | |
| 101 | 100 | kritisanon | 76 | 2.7k | 50.2m | 897.2k | 1.21% | 604.4k | 2.4b | India | PhD | |

The 96th most followed user is represented by row 97, but this row has a lot of empty cells, and the user name is also mentioned incorrectly. The entire row is subsequently removed from the dataset.

E10 | 268.3

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rank(Or | channel_info(Nomi | influence_score(Inten | posts(Ratio) | followers(Ratio) | avg_likes(Ratio) | 60_day_eng_ | new_post | total_like: | country(Nominal) | Education Level(Ordinal) |
| 2 | 1 | cristiano | 92 | 3.3k | 475.8m | 8.7m | 1.39% | 6.5m | 29.0b | Spain | Middle School |
| 3 | 2 | kyliejenner | 91 | 6.9k | 366.2m | 8.3m | 1.62% | 5.9m | 57.4b | United States | High School |
| 4 | 3 | leomessi | 90 | 0.89k | 357.3m | 6.8m | 1.24% | 4.4m | 6.0b | Argentina | High School |
| 5 | 4 | selenagomez | 93 | 1.8k | 342.7m | 6.2m | 0.97% | 3.3m | 11.5b | United States | High School |
| 6 | 5 | therock | 91 | 6.8k | 334.1m | 1.9m | 0.20% | 665.3k | 12.5b | United States | BA |
| 7 | 6 | kimkardashian | 91 | 5.6k | 329.2m | 3.5m | 0.88% | 2.9m | 19.9b | United States | High School |
| 8 | 7 | arianagrande | 92 | 5.0k | 327.7m | 3.7m | 1.20% | 3.9m | 18.4b | United States | High School |
| 9 | 8 | beyonce | 92 | 2.0k | 272.8m | 3.6m | 0.76% | 2.0m | 7.4b | United States | Middle School |
| 10 | 9 | khloekardashian | 89 | 4.1k | 268.3 | 2.4m | 0.35% | 926.9k | 9.8b | United States | High School |
| 11 | 10 | justinbieber | 91 | 7.4k | 254.5m | 1.9m | 0.59% | 1.5m | 13.9b | Canada | High School |
| 12 | 11 | kendalljenner | 90 | 0.66k | 254.0m | 5.5m | 2.04% | 5.1m | 3.7b | United States | High School |
| 13 | 12 | natgeo | 91 | 10.0k | 237.0m | 302.2k | 0.07% | 159.3k | 3.0b | United States | - |
| 14 | 13 | nike | 90 | 0.95k | 234.1m | 329.0k | 0.08% | 181.8k | 313.6m | United States | - |
| 15 | 14 | taylorswift | 91 | 0.53k | 222.2m | 2.4m | 1.01% | 2.3m | 1.3b | United States | High School |
| 16 | 15 | jlo | 89 | 3.2k | 220.4m | 1.7m | 0.62% | 1.4m | 5.3b | United States | High School |
| 17 | 16 | virat.kohli | 87 | 1.4k | 211.8m | 3.5m | 0.96% | 2.0m | 4.9b | India | Intermediate level |
| 18 | 17 | nickiminaj | 90 | 6.4k | 201.6m | 2.1m | 0.53% | 1.0m | 13.5b | United States | High School |
| 19 | 18 | kourtneykardash | 89 | 4.4k | 195.2m | 1.8m | 0.67% | 1.3m | 7.7b | United States | BA |
| 20 | 19 | mileycyrus | 89 | 1.2k | 181.5m | 1.3m | 0.51% | 913.6k | 1.6b | United States | High School |
| 21 | 20 | neymarjr | 90 | 5.3k | 177.1m | 2.7m | 1.09% | 1.9m | 14.1b | Brazil | Middle School |
| 22 | 21 | katyperry | 92 | 2.0k | 170.3m | 715.0k | 0.16% | 265.1k | 1.5b | United States | High School |
| 23 | 22 | kevinhart4real | 88 | 8.2k | 152.0m | 522.0k | 0.08% | 115.2k | 4.3b | United States | High School |
| 24 | 23 | zendaya | 87 | 3.5k | 150.7m | 5.8m | 3.17% | 4.8m | 20.6b | United States | BA |
| 25 | 24 | iamcardib | 75 | 1.6k | 140.5m | 3.1m | 1.10% | 1.5m | 5.0b | United States | High School |

The number of followers corresponding to the user must be stated in millions or thousands in row 10, column E. Since a user cannot have followers in decimal values and 268.3 is a decimal number, it is treated as 268.3m and substituted for 268.3.

| A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|
| 31 | chrisbrownofficial | 86 | 7.3k | 118.5m | 463.2k | 0.22% | 255.9k | 3.4b | United States | High School |
| 32 | fcbarcelona | 90 | 10.0k | 111.4m | 1.2m | 0.64% | 706.3k | 11.6b | United Kingdom | - |
| 33 | billieeilish | 73 | 0.69k | 105.2m | 8.5m | 5.02% | 5.2m | 5.9b | United States | High School |
| 34 | dualipa | 74 | 1.3k | 85.9m | 2.1m | 1.26% | 1.1m | 2.6b | United Kingdom | High School |
| 35 | gal_gadot | 85 | 1.7k | 85.6m | 1.4m | 0.69% | 586.5k | 2.3b | United States | Middle School |
| 36 | vindiesel | 88 | 1.8k | 82.3m | 1.4m | 0.60% | 482.3k | 2.5b | United States | High School |
| 37 | nasa | 88 | 3.6k | 81.3m | 1.2m | 1.53% | 1.2m | 4.2b | United States | - |
| 38 | priyankachopra | 85 | 3.6k | 81.1m | 1.6m | 1.00% | 802.9k | 5.6b | United States | BA |
| 39 | lalalalisa_m | 70 | 0.87k | 80.9m | 5.8m | 9.00% | 7.2m | 5.1b | South Korea | High School |
| 40 | Shakira | 88 | 2.0k | 76.1m | 975.1k | 0.41% | 304.7k | 1.9b | Columbia | Middle School |
| 41 | snoopdogg | 86 | 10.0k | 75.3m | 203.7k | 0.17% | 125.8k | 2.0b | United States | Secondary School |
| 42 | gigihadid | 85 | 3.3k | 75.3m | 2.5m | 2.56% | 1.9m | 8.2b | United States | BA |
| 43 | davidbeckham | 86 | 1.5k | 74.9m | 1.2m | 0.46% | 340.9k | 1.9b | United States | High School |
| 44 | shraddhakapoor | 81 | 1.9k | 73.9m | 1.6m | 1.17% | 859.9k | 3.0b | Indian | BS |
| 45 | victoriassecret | 88 | 2.9k | 73.2m | 147.0k | 0.04% | 29.8k | 423.5m | United States | - |
| 46 | k.mbappe | 86 | 1.2k | 72.7m | 2.5m | 2.16% | 1.6m | 2.8b | French | Middle School |
| 47 | nehakakkar | 84 | 2.3k | 70.4m | 1.5m | 0.54% | 370.1k | 3.5b | India | BA |
| 48 | nba | 87 | 12.9k | 70.1m | 370.8k | 0.28% | 196.9k | 4.8b | United States | - |
| 49 | shawnmendes | 83 | 2.5k | 69.9m | 3.5m | 2.87% | 2.0m | 8.8b | Canada | High School |
| 50 | jennierubyjane | 76 | 0.86k | 68.9m | 5.1m | 8.36% | 5.7m | 4.4b | South Korea | BS |
| 51 | narendramodi | 85 | 0.54k | 68.9m | 2.9m | 3.01% | 2.0m | 1.6b | India | BE |
| 52 | aliaabhatt | 82 | 1.8k | 68.7m | 1.8m | 3.14% | 2.1m | 3.3b | India | Intermediate level |
| 53 | deepikapadukone | 83 | 0.26k | 68.4m | 1.6m | 2.23% | 1.5m | 419.0m | India | BA |
| 54 | tomholland2013 | 77 | 1.2k | 67.7m | 5.4m | 10.83% | 7.3m | 6.6b | United Kingdom | High School |
| 55 | ronaldinho | 78 | 2.9k | 67.7m | 872.9k | 0.49% | 325.7k | 2.6b | Brazil | Middle School |

TOP 100 INSTA FOLLOWERS | Sheet2 | Sheet3

The dataset's column J designates the user's country of origin. However, the user's nationality rather than their country is listed in row 44, column J. As a result, "India" in place of "Indian" is used.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rank(Or | channel_info(Nomi | influence_score(Inter | posts(Ratio) | followers(Ratio) | avg_likes(Ratio) | 60_day_eng_ | new_post | total_like: | country(Nominal) | Education Level(Ordinal) |
| 2 | 1 | cristiano | 92 | 3.3k | 475.8m | 8.7m | 1.39% | 6.5m | 29.0b | Spain | Middle School |
| 3 | 2 | kyliejenner | 91 | 6.9k | 366.2m | 8.3m | 1.62% | 5.9m | 57.4b | United States | High School |
| 4 | 3 | leomessi | 90 | 0.89k | 357.3m | 6.8m | 1.24% | 4.4m | 6.0b | Argentina | High School |
| 5 | 4 | selenagomez | 93 | 1.8k | 342.7m | 6.2m | 0.97% | 3.3m | 11.5b | United States | High School |
| 6 | 5 | therock | 91 | 6.8k | 334.1m | 1.9m | 0.20% | 665.3k | 12.5b | United States | BA |
| 7 | 6 | kimkardashian | 91 | 5.6k | 329.2m | 3.5m | 0.88% | 2.9m | 19.9b | United States | High School |
| 8 | 7 | arianagrande | 92 | 5.0k | 327.7m | 3.7m | 1.20% | 3.9m | 18.4b | United States | High School |
| 9 | 8 | beyonce | 92 | 2.0k | 272.8m | 3.6m | 0.76% | 2.0m | 7.4b | United States | Middle School |
| 10 | 9 | khloekardashian | 89 | 4.1k | 268.3m | 2.4m | 0.35% | 926.9k | 9.8b | United States | High School |
| 11 | 10 | justinbieber | 91 | 7.4k | 254.5m | 1.9m | 0.59% | 1.5m | 13.9b | Canada | 11th standard |
| 12 | 11 | kendalljenner | 90 | 0.66k | 254.0m | 5.5m | 2.04% | 5.1m | 3.7b | United States | High School |
| 13 | 12 | natgeo | 91 | 10.0k | 237.0m | 302.2k | 0.07% | 159.3k | 3.0b | United States | - |
| 14 | 13 | nike | 90 | 0.95k | 234.1m | 329.0k | 0.08% | 181.8k | 313.6m | United States | - |
| 15 | 14 | taylorswift | 91 | 0.53k | 222.2m | 2.4m | 1.01% | 2.3m | 1.3b | United States | High School |
| 16 | 15 | jlo | 89 | 3.2k | 220.4m | 1.7m | 0.62% | 1.4m | 5.3b | United States | High School |
| 17 | 16 | virat.kohli | 87 | 1.4k | 211.8m | 3.5m | 0.96% | 2.0m | 4.9b | India | Intermediate level |
| 18 | 17 | nickiminaj | 90 | 6.4k | 201.6m | 2.1m | 0.53% | 1.0m | 13.5b | United States | High School |
| 19 | 18 | kourtneykardash | 89 | 4.4k | 195.2m | 1.8m | 0.67% | 1.3m | 7.7b | United States | BA |
| 20 | 19 | mileycyrus | 89 | 1.2k | 181.5m | 1.3m | 0.51% | 913.6k | 1.6b | United States | High School |
| 21 | 20 | neymarjr | 90 | 5.3k | 177.1m | 2.7m | 1.09% | 1.9m | 14.1b | Brazil | Middle School |
| 22 | 21 | katyperry | 92 | 2.0k | 170.3m | 715.0k | 0.16% | 265.1k | 1.5b | United States | High School |
| 23 | 22 | kevinhart4real | 88 | 8.2k | 152.0m | 522.0k | 0.08% | 115.2k | 4.3b | United States | High School |
| 24 | 23 | zendaya | 87 | 3.5k | 150.7m | 5.8m | 3.17% | 4.8m | 20.6b | United States | BA |
| 25 | 24 | iamcardib | 75 | 1.6k | 140.5m | 3.1m | 1.10% | 1.5m | 5.0b | United States | High School |

TOP 100 INSTA FOLLOWERS / Sheet2 / Sheet3

Although the Education Level in row 10, column J, refers to standards, all the other cells in this column refer to elementary levels; as a result, "11th standard" is changed to "High School."

C35 | 74%

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 24 | iamcardib | 75 | 1.6k | 140.5m | 3.1m | 1.10% | 1.5m | 5.0b | United States | High School |
| 26 | 25 | ddlovato | 88 | 0.08k | 139.1m | 1.1m | 0.27% | 363.4k | 91.3m | United States | High School |
| 27 | 26 | badgalriri | 88 | 4.8k | 135.3m | 3.7m | 0.02% | 133.4k | 17.9b | United States | High School |
| 28 | 27 | kingjames | 86 | 2.3k | 130.9m | 2.1m | 0.92% | 1.2m | 4.9b | United States | High School |
| 29 | 28 | theellenshow | 87 | 10.0k | 125.1m | 420.5k | 0.07% | 81.7k | 4.2b | United States | High School |
| 30 | 29 | realmadrid | 90 | 6.9k | 123.4m | 996.2k | 0.48% | 588.3k | 6.8b | Spain | |
| 31 | 30 | champagnepapi | 85 | 5.2k | 119.6m | 1.7m | 1.10% | 1.3m | 9.0b | Netherlands | High School |
| 32 | 31 | chrisbrownoffic | 86 | 7.3k | 118.5m | 463.2k | 0.22% | 255.9k | 3.4b | United States | High School |
| 33 | 32 | fcbarcelona | 90 | 10.0k | 111.4m | 1.2m | 0.64% | 706.3k | 11.6b | United Kingdom | |
| 34 | 33 | billieeilish | 73 | 0.69k | 105.2m | 8.5m | 5.02% | 5.2m | 5.9b | United States | High School |
| 35 | 34 | dualipa | 74% | 1.3k | 85.9m | 2.1m | 1.26% | 1.1m | 2.6b | United Kingdom | High School |
| 36 | 35 | gal_gadot | 85 | 1.7k | 85.6m | 1.4m | 0.69% | 586.5k | 2.3b | United States | Middle School |
| 37 | 36 | vindiesel | 88 | 1.8k | 82.3m | 1.4m | 0.60% | 482.3k | 2.5b | United States | High School |
| 38 | 37 | nasa | 88 | 3.6k | 81.3m | 1.2m | 1.53% | 1.2m | 4.2b | United States | |
| 39 | 38 | priyankachopra | 85 | 3.6k | 81.1m | 1.6m | 1.00% | 802.9k | 5.6b | United States | BA |
| 40 | 39 | lalalalisa_m | 70 | 0.87k | 80.9m | 5.8m | 9.00% | 7.2m | 5.1b | South Korea | High School |
| 41 | 40 | Shakira | 88 | 2.0k | 76.1m | 975.1k | 0.41% | 304.7k | 1.9b | Columbia | Middle School |
| 42 | 41 | snoopdogg | 86 | 10.0k | 75.3m | 203.7k | 0.17% | 125.8k | 2.0b | United States | Secondary School |
| 43 | 42 | gigihadid | 85 | 3.3k | 75.3m | 2.5m | 2.56% | 1.9m | 8.2b | United States | BA |
| 44 | 43 | davidbeckham | 86 | 1.5k | 74.9m | 1.2m | 0.46% | 340.9k | 1.9b | United States | High School |
| 45 | 44 | shraddhakapoo | 81 | 1.9k | 73.9m | 1.6m | 1.17% | 859.9k | 3.0b | Indian | BS |
| 46 | 45 | victoriassecret | 88 | 2.9k | 73.2m | 147.0k | 0.04% | 29.8k | 423.5m | United States | |
| 47 | 46 | k.mbappe | 86 | 1.2k | 72.7m | 2.5m | 2.16% | 1.6m | 2.8b | French | Middle School |
| 48 | 47 | nehakakkar | 84 | 2.3k | 70.4m | 1.5m | 0.54% | 370.1k | 3.5b | India | BA |
| 49 | 48 | nba | 87 | 12.9k | 70.1m | 370.8k | 0.28% | 196.9k | 4.8b | United States | |

Column C represents the influencer score of the user. But the score is not mentioned in terms of percentage thus ‘74%’ is replaced by ‘74’.

### 5.2 Loading the dataset using python and getting Summary Statistics

```
# Loading Dataset

Import pandas as pd

info = pd.read_excel(r"C:\Users\sarayu\Downloads\top 100 insta
followers dataset.xlsx")
print(info)
```

**Output:**

```
ass3_code.py 1    top100_code.py 1
C: > Users > sarayu > Downloads > top100_code.py > ...
1  # Loading Dataset
2  import pandas as pd
3
4  info = pd.read_excel(r"C:\Users\sarayu\Downloads\top 100 insta followers dataset.xlsx")
5  # print(info)
6
7
```

```
PROBLEMS 2   OUTPUT   DEBUG CONSOLE   TERMINAL   JUPYTER                                   powershell
[notice] To update, run: C:\Users\sarayu\AppData\Local\Microsoft\WindowsApps\PythonSoftwareFoundation.Python.3.10_qbz5n2kfra8p0\python.exe -m pip install --upgrade pip
PS C:\Users\sarayu\Downloads> python top100_code.py
    rank(Ordinal) channel_info(Nominal)  influence_score(Interval) posts(Ratio)  ... new_post_avg_like(Ratio) total_likes(Ratio)  country(Nominal) Education Level(Ordinal)
0               1            cristiano                        92        3.3k  ...                   6.5m            29.0b            Spain           Middle School
1               2          kyliejenner                        91        6.9k  ...                   5.9m            57.4b    United States             High School
2               3             leomessi                        90       0.89k  ...                   4.4m             6.0b        Argentina             High School
3               4          selenagomez                        93        1.8k  ...                   3.3m            11.5b    United States             High School
4               5              therock                        91        6.8k  ...                 665.3k            12.5b    United States                      BA
..            ...                  ...                       ...         ...  ...                    ...              ...              ...                     ...
94             95        ayutingting92                        85       10.0k  ...                  56.3k             1.5b        Indonesia           Middle School
95             97           hudabeauty                        82        2.4k  ...                  16.8k           453.6m    United States             High School
96             98                 adele                        84       0.42k  ...                   1.9m             2.0b    United States             High School
97             99        michelleobama                        85       0.60k  ...                 611.2k           421.7m    United States                      BA
98            100           kritisanon                        76        2.7k  ...                 604.4k             2.4b            India                     PhD

[99 rows x 11 columns]
```

The rows and columns that are loaded from the dataset into visual studio workbench are shown in the screenshot up top. The output is displayed as an 11-column list with 99 rows. There are the following columns: rank, channel info, influencer score, posts, followers, average likes, 60 day engagement rate, average likes for new posts, total likes, nation, and education level. The excel file is loaded using the "pd.read excel" function, and the output is printed using the "print()" expression.

```
# Summary Statistics of rank,influence score, 60 days engagement rate

print("statistics of rank", info[["rank(Ordinal)"]].describe())
print("median of rank", info[["rank(Ordinal)"]].median())
print("mode of rank", info[["rank(Ordinal)"]].mode())
print("statistics of influence_score",
info[["influence_score(Interval)"]].describe())
```

```
print("median of influence_score",
info[["influence_score(Interval)"]].median())
print("mode of influence_score",
info[["influence_score(Interval)"]].mode())
print("statistics of
60_day_eng_rate",info[["60_day_eng_rate(Interval)"]].describe())
print("median of 60_day_eng_rate",
info[["60_day_eng_rate(Interval)"]].median())
print("mode of 60_day_eng_rate",
info[["60_day_eng_rate(Interval)"]].mode())
```

**Output:**

```
PROBLEMS 2   OUTPUT   DEBUG CONSOLE   TERMINAL   JUPYTER

statistics of rank        rank(Ordinal)
count      99.000000
mean       50.040404
std        28.790914
min         1.000000
25%        25.500000
50%        50.000000
75%        74.500000
max       100.000000
median of rank rank(Ordinal)    50.0
dtype: float64
mode of rank     rank(Ordinal)
0               1
1               2
2               3
3               4
4               5
..            ...
94             95
95             97
96             98
97             99
98            100

[99 rows x 1 columns]
```

The output of the rank column's summary statistics is displayed in the screenshot up top. It is followed by information about the column's median and mode. As can be seen, the column "rank" comprises 99 modes since it contains user rankings, which are individual to each user and never repeat, making all of the values in the column modes.

```
PROBLEMS 2   OUTPUT   DEBUG CONSOLE   TERMINAL   JUPYTER

statistics of influence_score        influence_score(Interval)
count                   99.000000
mean                    83.919192
std                      8.998500
min                     22.000000
25%                     82.000000
50%                     86.000000
75%                     88.000000
max                     93.000000
median of influence_score influence_score(Interval)    86.0
dtype: float64
mode of influence_score    influence_score(Interval)
0                          86
```

One of the columns in the dataset, influence score, shows the value a user has contributed through their influence. The more a user's impact over others, the higher their influence score. In the screenshot up above, the summary statistics for this column are shown. The mode of this column is displayed as 86, meaning that the majority of top influencers have an influence score of 86. Statistics indicate that this score might have a maximum value of 93.

```
PROBLEMS 2   OUTPUT   DEBUG CONSOLE   TERMINAL   JUPYTER

statistics of 60_day_eng_rate        60_day_eng_rate(Interval)
count                   99.000000
mean                     0.017896
std                      0.024747
min                      0.000100
25%                      0.004850
50%                      0.009700
75%                      0.020450
max                      0.108300
median of 60_day_eng_rate 60_day_eng_rate(Interval)    0.0097
dtype: float64
mode of 60_day_eng_rate    60_day_eng_rate(Interval)
0                     0.0051
```

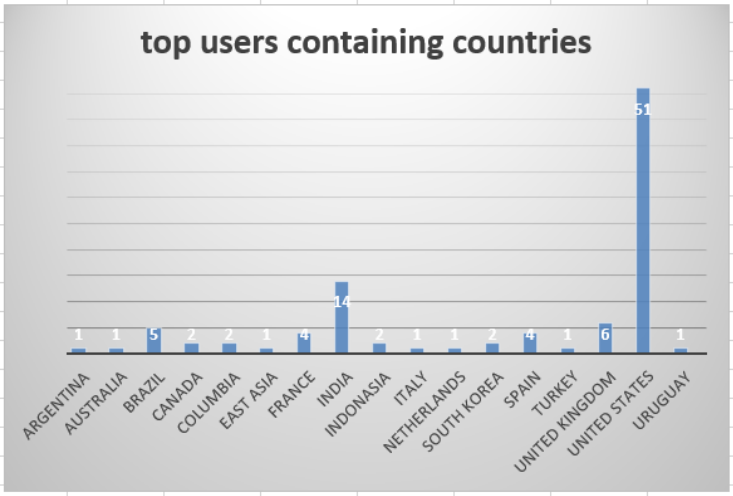
The 60-day engagement rate is the frequency with which a user interacts with the platform's audience by uploading a story, post, reel, etc. A percentage value is used to represent the 60-day engagement rate. The screenshot above shows the Summary Statistics for this column.

# VI. RESULTS AND ANALYSIS

a. Most of the top influenced instagramers are from which country?

The first study question, "Which country has the most top Instagram followers?" is answered by the below graph, which shows the country of origin of the top Instagram users. The graph is created using this table with the country name on the X-axis and the number of users on the Y-axis. The table next to the graph includes two columns: one with country of origin and the other with the number of users who are from that particular nation.

| COUNTRY | NUMBER OF USERS |
|---|---|
| Argentina | 1 |
| Australia | 1 |
| Brazil | 5 |
| Canada | 2 |
| Columbia | 2 |
| East Asia | 1 |
| France | 4 |
| India | 14 |
| Indonasia | 2 |
| Italy | 1 |
| Netherlands | 1 |
| South Korea | 2 |
| Spain | 4 |
| Turkey | 1 |
| United Kingdom | 6 |
| United States | 51 |
| Uruguay | 1 |




Thus, it can be seen from the graph displayed in the Excel spreadsheet that the country with the greatest number of top influencers is the United States, which has 51 of the top 100 influencers in its population. India is in second place with '14' followers, followed by 'United Kingdom' in third with '6' followers.

Not simply by studying from visuals, but also by writing code and displaying the output, the answer to the first research question can be found.

### Loading the dataset in R

```
# Loading Dataset
install.packages("expss")
library(expss)

install.packages("readxl")
library("readxl")

info <- read_excel("C:\\Users\\sarayu\\Downloads\\top 100 insta followers dataset.xlsx")
print(info)
```

```
> info <- read_excel("C:\\Users\\sarayu\\Downloads\\top 100 insta followers dataset.xlsx")
> print(info)
# A tibble: 99 × 11
    rank channel_info  influ…¹ posts follo…² avg_l…³ 60_da…⁴ new_p…⁵ total…⁶ country
   <dbl> <chr>            <dbl> <chr> <chr>   <chr>     <dbl> <chr>   <chr>   <chr>
 1     1 cristiano           92 3.3k  475.8m  8.7m     0.0139 6.5m    29.0b   Spain
 2     2 kyliejenner         91 6.9k  366.2m  8.3m     0.0162 5.9m    57.4b   United…
 3     3 leomessi            90 0.89k 357.3m  6.8m     0.0124 4.4m    6.0b    Argent…
 4     4 selenagomez         93 1.8k  342.7m  6.2m     0.0097 3.3m    11.5b   United…
 5     5 therock             91 6.8k  334.1m  1.9m     0.002  665.3k  12.5b   United…
 6     6 kimkardashian       91 5.6k  329.2m  3.5m     0.0088 2.9m    19.9b   United…
 7     7 arianagrande        92 5.0k  327.7m  3.7m     0.012  3.9m    18.4b   United…
 8     8 beyonce             92 2.0k  272.8m  3.6m     0.0076 2.0m    7.4b    United…
 9     9 khloekardash…       89 4.1k  268.3m  2.4m     0.0035 926.9k  9.8b    United…
10    10 justinbieber        91 7.4k  254.5m  1.9m     0.0059 1.5m    13.9b   Canada
# … with 89 more rows, 1 more variable: `Education Level` <chr>, and abbreviated
#   variable names ¹influence_score, ²followers, ³avg_likes, ⁴`60_day_eng_rate`,
#   ⁵new_post_avg_like, ⁶total_likes
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
>
```

The dataset is loaded into R workbench in the screenshot above. The output only displays the first 10 rows and 10 columns, however there are actually 89 additional rows and additional columns loaded. Given that the whole dataset has been loaded, analysis may now be done.

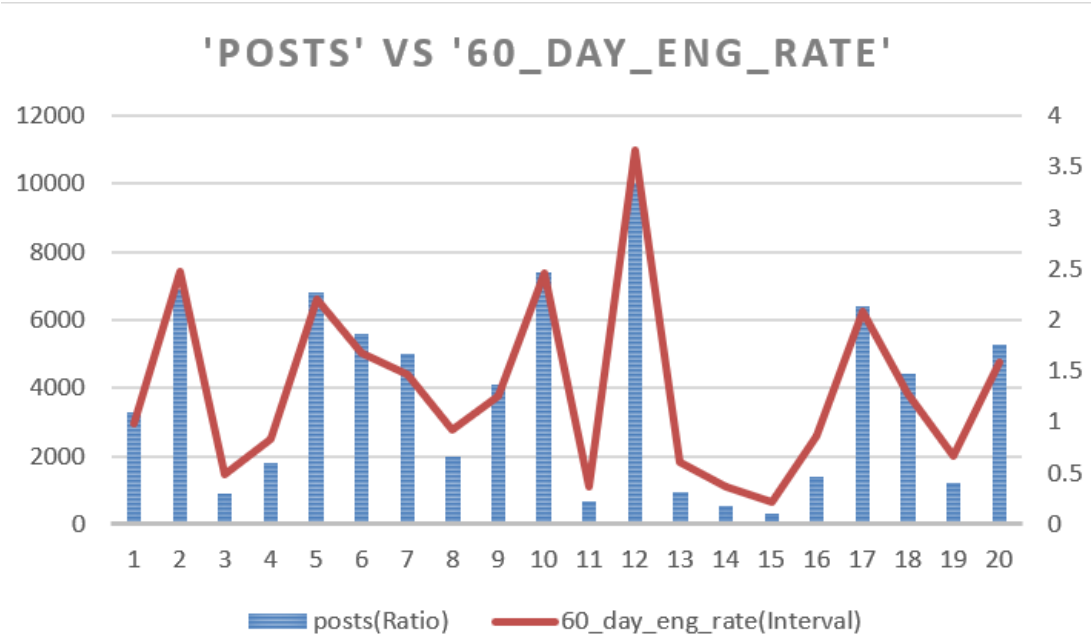
```
> country_table <- table(info$country)
> country_table

     Argentina      Australia         Brazil         Canada       Colombia       Columbia
             1              1              5              2              1              1
     East Asia         France         French          India      Indonesia          Italy
             1              2              2             14              2              1
   Netherlands    South Korea          Spain         Turkey United Kingdom  United States
             1              2              4              1              6             51
       Uruguay
             1
> country_max
United States
```

The output of the R Studio code is depicted in the screenshot up above. One can see that a table is displayed that lists all the countries along with the appropriate amount of users for each one. The output below displays "United States," the nation with the most users, according to the visualization.

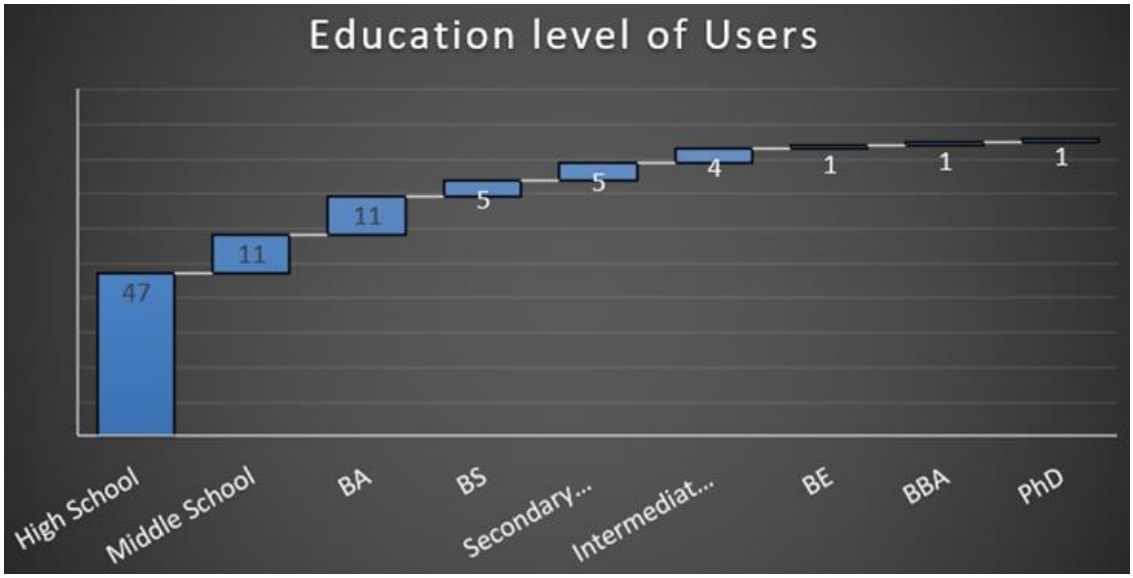
b. What is the relationship between 60_day_eng_rate and posts amongst top 20 followers?

The results of the second study question can be obtained by creating a graph between these 2 columns, which shows the relationship between 60 day eng rate and postings among the top 20 followers. The identical screenshot is shown below.



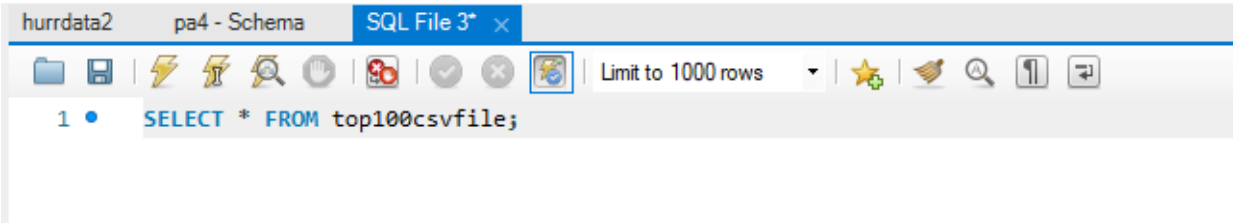

The values in the posts column are represented in thousands i.e., ‘k’ is mentioned beside the value. However, to obtain the graph the values are converted into numerical form for example 1000 in place of 1k. Posts are displayed with a "blue" bar chart, while 60 day eng rate is depicted with a "red" line. Instagram user rank is represented by the X-axis, and posts and the 60 day eng rate are represented by the Y-axis. The scale of the Y-axis is 0.5 units for the 60 day eng rate and 2000 units for the number of postings on the left side. The graph makes it quite clear that if there are more postings, there will also be more 60 day eng rate, and vice versa. The rate at which a user engages with their audience, as measured by postings, can be deduced to be 60 day eng rate. Posts and 60 day eng rate are therefore exactly proportionate, it can be said.

c. What proportion of the influencers has their Education Level higher than High School?




The bar plot created for the Education Level column is shown in the screenshot above. It is clear that there are various educational standards at various levels, and the number of users who have attained each degree of education is reflected. A few calculations from the graph must be conducted in order to get the answer to the third research question. The percentage of users with education levels above high school must be determined. The degrees above a high school diploma include those with a BA, BS, BE, BBA, PhD, intermediate level, and secondary level. The total number of users who met these education requirements is equal to 28, or 11+5+5+4+1+1+1. There are therefore 28 users with higher education levels than a high school diploma out of a total of 100 users. Thus, 28% of top 100 users have their qualification greater than High School.

**Loading dataset into SQL Workbench**



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| | rank | channel_info | influence_score | posts | followers | avg_likes | 60_day_eng_rate | new_post_avg_like | total_likes | country | Education Level |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ▶ | 1 | cristiano | 92 | 3.3k | 475.8m | 8.7m | 1.39% | 6.5m | 29.0b | Spain | Middle Schoo |
| | 2 | kyliejenner | 91 | 6.9k | 366.2m | 8.3m | 1.62% | 5.9m | 57.4b | United States | High School |
| | 3 | leomessi | 90 | 0.89k | 357.3m | 6.8m | 1.24% | 4.4m | 6.0b | Argentina | High School |
| | 4 | selenagomez | 93 | 1.8k | 342.7m | 6.2m | 0.97% | 3.3m | 11.5b | United States | High School |
| | 5 | therock | 91 | 6.8k | 334.1m | 1.9m | 0.20% | 665.3k | 12.5b | United States | BA |
| | 6 | kimkardashian | 91 | 5.6k | 329.2m | 3.5m | 0.88% | 2.9m | 19.9b | United States | High School |

Output

Action Output

| # | Time | Action | Message | Duration / Fetch |
|---|---|---|---|---|
| 23 | 23:57:37 | select * from <top100csvfile> | Error Code: 1064. You have an error in your SQL syntax; check the manual that corresponds ... | 0.000 sec |
| 24 | 23:58:53 | SELECT * FROM import_data LIMIT 0, 1000 | Error Code: 1146. Table 'pa4.import_data' doesn't exist | 0.000 sec |
| 25 | 23:59:26 | SELECT * FROM top100csvfile LIMIT 0, 1000 | 99 row(s) returned | 0.000 sec / 0.000 sec |
| 26 | 00:00:29 | SELECT * FROM top100csvfile LIMIT 0, 1000 | 99 row(s) returned | 0.000 sec / 0.000 sec |

The above screenshot shows the output of loaded dataset onto SQL Workbench. However, all the rows are not displayed, the 2nd screenshot shows that 99 rows are imported.

hurrdata2 | pa4 - Schema | SQL File 3*

Limit to 1000 rows

```
1 • SELECT * FROM top100csvfile;
2 • select * from top100csvfile where country = 'United States'
3
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| | rank | channel_info | influence_score | posts | followers | avg_likes | 60_day_eng_rate | new_post_avg_like | total_likes | country | Education Level |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ▶ | 2 | kyliejenner | 91 | 6.9k | 366.2m | 8.3m | 1.62% | 5.9m | 57.4b | United States | High School |
| | 4 | selenagomez | 93 | 1.8k | 342.7m | 6.2m | 0.97% | 3.3m | 11.5b | United States | High School |
| | 5 | therock | 91 | 6.8k | 334.1m | 1.9m | 0.20% | 665.3k | 12.5b | United States | BA |
| | 6 | kimkardashian | 91 | 5.6k | 329.2m | 3.5m | 0.88% | 2.9m | 19.9b | United States | High School |
| | 7 | arianagrande | 92 | 5.0k | 327.7m | 3.7m | 1.20% | 3.9m | 18.4b | United States | High School |
| | 8 | beyonce | 92 | 2.0k | 272.8m | 3.6m | 0.76% | 2.0m | 7.4b | United States | Middle Schoo |
| | 9 | khloekardashian | 89 | 4.1k | 268.3m | 2.4m | 0.35% | 926.9k | 9.8b | United States | High School |
| | 11 | kendalljenner | 90 | 0.66k | 254.0m | 5.5m | 2.04% | 5.1m | 3.7b | United States | High School |
| | 12 | natgeo | 91 | 10.0k | 237.0m | 302.2k | 0.07% | 159.3k | 3.0b | United States | - |
| | 13 | nike | 90 | 0.95k | 234.1m | 329.0k | 0.08% | 181.8k | 313.6m | United States | - |
| | 14 | taylorswift | 91 | 0.53k | 222.2m | 2.4m | 1.01% | 2.3m | 1.3b | United States | High School |

Result Grid | Form Editor | Field Types

top100csvfile 4 | Read Only | Context Help | Snippets

Output

Action Output

| # | Time | Action | Message | Duration / Fetch |
|---|---|---|---|---|
| 28 | 00:05:41 | SELECT * FROM top100csvfile LIMIT 0, 1000 | 99 row(s) returned | 0.000 sec / 0.000 sec |
| 29 | 00:05:57 | select * from top100csvfile where country = 'United States' LIMIT 0, 1000 | 51 row(s) returned | 0.000 sec / 0.000 sec |

The users whose country of origin is the United States are displayed in the screenshot above. Since it can be observed that 51 rows are returned at the bottom in the output section, there are 51 users in the top 100 list whose origin is the United States. The response to the first research question can be found with this query.

```
SELECT * FROM top100csvfile;
select * from top100csvfile where country = 'United States'
select * from top100csvfile order by influence_score
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| rank | channel_info | influence_score | posts | followers | avg_likes | 60_day_eng_rate | new_post_avg_like | total_likes | country | Education Level |
|---|---|---|---|---|---|---|---|---|---|---|
| 94 | mosalah | 22 | 0.84k | 52.5m | 1.8m | 2.18% | 1.1m | 1.5b | Italy | High Scho |
| 77 | karimbenzema | 42 | 2.0k | 56.9m | 878.5k | 1.62% | 918.9k | 1.8b | Spain | High Scho |
| 39 | lalalalisa_m | 70 | 0.87k | 80.9m | 5.8m | 9.00% | 7.2m | 5.1b | South Korea | High Scho |
| 33 | billieeilish | 73 | 0.69k | 105.2m | 8.5m | 5.02% | 5.2m | 5.9b | United States | High Scho |
| 34 | dualipa | 74 | 1.3k | 85.9m | 2.1m | 1.26% | 1.1m | 2.6b | United Kingdom | High Scho |
| 91 | dishapatani | 74 | 2.1k | 53.0m | 1.6m | 1.85% | 971.7k | 3.4b | United States | BA |
| 24 | iamcardib | 75 | 1.6k | 140.5m | 3.1m | 1.10% | 1.5m | 5.0b | United States | High Scho |
| 50 | jennierubyjane | 76 | 0.86k | 68.9m | 5.1m | 8.36% | 5.7m | 4.4b | South Korea | BS |
| 87 | beingsalmankhan | 76 | 1.2k | 53.5m | 1.4m | 1.33% | 693.9k | 1.6b | India | Intermedi |
| 100 | kritisanon | 76 | 2.7k | 50.2m | 897.2k | 1.21% | 604.4k | 2.4b | India | PhD |
| 61 | camila_cabello | 77 | 2.9k | 65.0m | 1.9m | 2.19% | 1.4m | 5.4b | United States | Secondar |

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| rank | channel_info | influence_score | posts | followers | avg_likes | 60_day_eng_rate | new_post_avg_like | total_likes | country | Education Level |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | kyliejenner | 91 | 6.9k | 366.2m | 8.3m | 1.62% | 5.9m | 57.4b | United States | High Scho |
| 5 | therock | 91 | 6.8k | 334.1m | 1.9m | 0.20% | 665.3k | 12.5b | United States | BA |
| 6 | kimkardashian | 91 | 5.6k | 329.2m | 3.5m | 0.88% | 2.9m | 19.9b | United States | High Scho |
| 14 | taylorswift | 91 | 0.53k | 222.2m | 2.4m | 1.01% | 2.3m | 1.3b | United States | High Scho |
| 12 | natgeo | 91 | 10.0k | 237.0m | 302.2k | 0.07% | 159.3k | 3.0b | United States | - |
| 10 | justinbieber | 91 | 7.4k | 254.5m | 1.9m | 0.59% | 1.5m | 13.9b | Canada | High Scho |
| 7 | arianagrande | 92 | 5.0k | 327.7m | 3.7m | 1.20% | 3.9m | 18.4b | United States | High Scho |
| 21 | katyperry | 92 | 2.0k | 170.3m | 715.0k | 0.16% | 265.1k | 1.5b | United States | High Scho |
| 1 | cristiano | 92 | 3.3k | 475.8m | 8.7m | 1.39% | 6.5m | 29.0b | Spain | Middle Sch |
| 8 | beyonce | 92 | 2.0k | 272.8m | 3.6m | 0.76% | 2.0m | 7.4b | United States | Middle Sch |
| 4 | selenagomez | 93 | 1.8k | 342.7m | 6.2m | 0.97% | 3.3m | 11.5b | United States | High Scho |

The output of the rows that are organized in accordance with the influence score is seen in the screenshot above. The output is obtained by ordering by function. The result shows that 93 is the greatest influencer score and that 22 is the least influencer score.

```
1  SELECT * FROM top100csvfile;
2  SELECT * FROM top100csvfile WHERE country = 'United States'
3  SELECT * FROM top100csvfile ORDER BY influence_score
4  SELECT country, COUNT(country) FROM top100csvfile GROUP BY country
5
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| country | COUNT(country) |
|---|---|
| Spain | 4 |
| United States | 51 |
| Argentina | 1 |
| Canada | 2 |
| India | 14 |
| Brazil | 5 |
| Netherlands | 1 |
| United Kingdom | 6 |
| South Korea | 2 |
| Columbia | 1 |
| French | 2 |
| Uruguay | 1 |
| Turkey | 1 |
| Indonesia | 2 |

By executing the line 4 query output is obtained which contains the list of how many users are present from each country from the top 100 list. It can be seen in the output that United States is the country which maximum number of users.

# VII. LIMITATIONS AND FUTURE RESEARCH

When conducting the study, few restrictions were encountered. Initially, there weren't many studies done on this subject. There have been numerous studies on the subject of social media, but few specifically on those that have the most Instagram followers. Even the research papers used in the literature review only broadly align with the topic at hand; they do not directly address it. Therefore, it becomes challenging to understand the

For the second study topic, the relationship between the columns 60 day eng rate and the total number of posts among the top 20 users is examined, and it is discovered that they are directly proportional. However, when the top 100 users are examined, the pattern does not continue to show this relationship. It is evident that there is no special relationship between these columns and the other top users.

Another crucial element is that the user's level of popularity will be useful in gaining a clear understanding of the behavior of the top users. If the user's level of popularity, or throughout how many countries, is known, then according to that, his or her in-depth insight can be ascertained. The presence of a popularity level column would be advantageous to other influencers if additional study were to be done to obtain more information from individuals with the majority of followers.

# VIII. DISCUSSIONS & CONCLUSION

The analysis on top 100 most followed instagram users gives insights which can be used to understand the behavioural pattern of the users. By exploring this data, the influencers who aim at increasing their followers count can know the strategies and techniques followed by these users and make use of the same. We can conclude that by having more engagement rate which is obtained by posting more stuff on the platform gives mor enumebr of followers, since as the user posts content frequently the people will follow them in order to look at their posts.

The visualizations help me conclude that the most number of users in top 100 list are from United States. Also it can be understood that 60_day_eng_rate and number of posts are directly related for top 20 users.

**Define/Explain terms:**

- **Influencer:** a person who has ability to influence others and make them follow their style.
- **Followers:** people who follow other users on a social media platform
- **Followee:** person is being followed by other people.
- **Engagement rate:** the rate at which a person is able to communicate with other users of the platform.
- **Average likes:** the number of likes that all the posts received divided by the number of posts that are posted by a user on instagram.

# REFERENCES

[1] Yang, C. (2021, February 8). *Research in the Instagram context: Approaches and methods*. Retrieved October 27, 2022, from https://www.researchgate.net/publication/349117428_Research_in_the_Instagram_Context_Approaches_and_Methods

[2] Dr. Daryl D. Green, Dr. Richard Martinez, et al. *In a world of social media: A case study analysis of instagram*. Retrieved October 28, 2022, from https://www.arjonline.org/papers/arjbm/v4-i1/12.pdf

[3] De Veirman, M. (n.d.). *Marketing through Instagram influencers: Impact of number of ... - core*. Retrieved October 29, 2022, from https://core.ac.uk/download/pdf/55691871.pdf

[4] Jha, S. (2022, August 13). *Top instagram influencers data (cleaned)*. Kaggle. Retrieved October 16, 2022, from https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned?utm_medium=social&utm_campaign=kaggle-dataset-share&utm_source=linkedin

[5] *Your machine learning and Data Science Community*. Kaggle. (n.d.). Retrieved October 162022, from https://www.kaggle.com/

[6] Lutkevich, B., & Wigmore, I. (2021, September 3). *What is social media?* WhatIs.com. Retrieved December 5, 2022, from https://www.techtarget.com/whatis/definition/social-media