VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

Social Media Analytics(CSE4069)

J Component report

Influencer Analysis

Programme: MTech. Integrated (Business Analytics)
Professor: Dr. R Priyadharshini
Students:
Nithya Sharma – 19mia1028
Annugraha S – 19mia1059
Mogalapu Sri Sai Vivek – 19mia1070

ABSTRACT

Influencer marketing on social media has been rapidly getting popular. Nowadays, there are so many people who identify themselves as influencers that it gets harder and harder to identify fake and inauthentic accounts. At the same time, influencer marketing is swiftly moving up the list of priorities for marketing plans. And with good reason too. More than 90% of marketers believe that this kind of marketing is successful. But how can you stay away from dealing with fraudulent accounts? What are some effective ways to employ influencer marketing for your business? You analyze influencers. In this study, we will analyze top influencers using Gephi tool, Power BI and machine learning models like decision tree regressor, random forest regressor, XGB Regressor.

INTRODUCTION

The role of social media in our lives keeps growing. Almost half of all the people on Earth use social media on a daily basis. So it's hardly surprising that brands want to take advantage of this trend and attract more customers using social platforms. Influencer marketing has been one of the most attractive opportunities for brands over the past few years, especially on Instagram.. The number of Instagram influencers grows day by day. Marketers get increasingly interested in partnering with them to raise brand awareness and promote their products. It's not all positive, however. While the opportunities both for content creators and brands are great, there is no lack of pitfalls. Lots of content creators want recognition and brand deals but not every one of them is prepared to do what it takes to build an engaged and organic following. Some use shortcuts by buying fake followers, likes, and Stories views. Needless to say, content quality and interaction with followers are not a priority in this case. At the same time, marketers need authentic content and highly engaged audiences they simply won't get from a social media user with a large but fake subscriber base. And that's not the only issue here. Even if your social media user is completely authentic with real followers it's not a guarantee they are good for your brand. Their following might not include your target audience. They might be located in a region that has little to no interest in your product. They might not have any idea of what they are going to promote and how. All this doesn't help you reach your marketing campaign goals in the slightest and as a result, you will spend time and money for nothing. Leave alone the fact that partnering with wrong influencers might actually do quite a lot of damage to your brand reputation. This is why it's crucially important to perform a thorough analysis of every influencer you would potentially like to collaborate

with. In this study , we will analyse top influencers using Gephi tool , Power BI and machine learning models like decision tree regressor, random forest regressor, XGB Regressor.

LITERATURE REVIEW

Influencer analysis is a popular topic in marketing and social media research, as it focuses on identifying individuals who have the power to impact consumer behavior and purchasing decisions. The following literature review highlights some of the key findings and trends in influencer analysis research.

Influencers can be classified into different categories based on their follower count, content, and engagement. Macro-influencers are individuals who have more than 100,000 followers, while micro-influencers have fewer than 100,000 followers. Mega-influencers have millions of followers, and nano-influencers have only a few hundred followers. Studies have shown that micro and nano-influencers are more effective in driving engagement and conversions compared to macro and mega-influencers (Chen et al., 2019).

One of the challenges in influencer marketing is selecting the right influencers to work with. Researchers have proposed various methods for selecting influencers, including network analysis, content analysis, and sentiment analysis. Network analysis involves examining the social network structure of the influencers, while content analysis involves analyzing the type and quality of content they produce. Sentiment analysis focuses on the emotional tone of their content, which can be positive, negative, or neutral (Khamitov et al., 2018).

Measuring the impact of influencer marketing campaigns is crucial to determine their effectiveness. Researchers have proposed various metrics for measuring the impact of influencers, including reach, engagement, sentiment, and conversion rates. Reach refers to the number of people who have seen the influencer's content, while engagement measures the level of interaction with the content, such as likes, comments, and shares. Sentiment analysis measures the emotional tone of the content, while conversion rates measure the percentage of people who take action, such as making a purchase, after seeing the influencer's content (Zhang et al., 2019).

Influencer fraud, such as buying followers or engagement, is a common issue in influencer marketing. Researchers have proposed various methods for detecting influencer fraud, including analyzing the follower growth rate, engagement rate, and content quality. Some studies have also used machine learning algorithms to detect fake followers and engagement (Khan et al., 2020).

The use of influencers in marketing raises ethical issues, such as transparency and disclosure. Studies have shown that consumers are more likely to trust influencers

who disclose their relationship with the brand and are transparent about their sponsored content. The Federal Trade Commission (FTC) in the United States has guidelines for influencer marketing, which require influencers to disclose their sponsored content (Phua et al., 2017).
Overall, influencer analysis is an important area of research in marketing and social media. Researchers continue to explore new methods for selecting, measuring, and evaluating influencers, as well as addressing ethical issues and fraud in influencer marketing

DATASET DESCRIPTION

Instagram is an American photo and video sharing social networking service founded in 2010 by Kevin Systrom and Mike Krieger, and later acquired by Facebook Inc.. The app allows users to upload media that can be edited with filters and organized by hashtags and geographical tagging. Posts can be shared publicly or with preapproved followers. Users can browse other users' content by tag and location, view trending content, like photos, and follow other users to add their content to a personal feed. Instagram network is very much used to influence people (the users followers) in a particular way for a specific issue - which can impact the order in some ways.

The dataset contains information regarding the top influencers .In this dataset, basically there are 10 attributes. It has been ordered on basis of the rank which has been decided on basis of "followers".

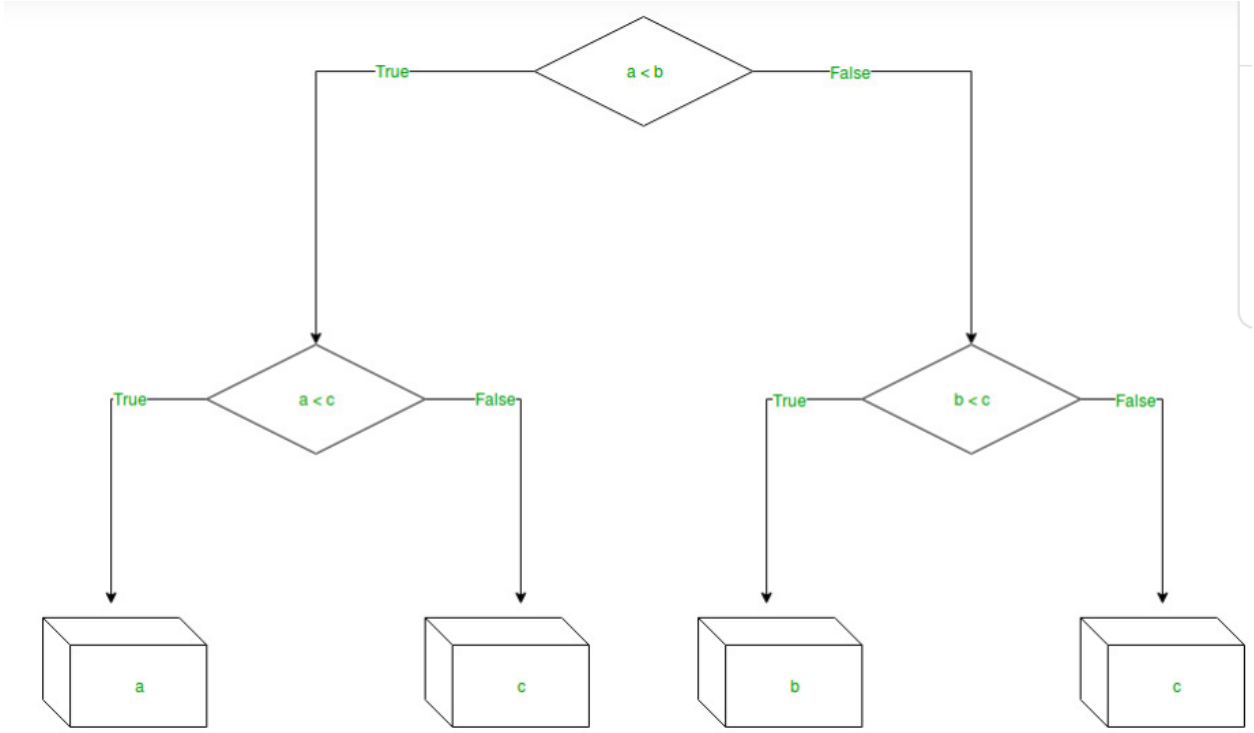
- rank: Rank of the Influencer on basis of number of followers they have
- channel_info: Username of the Instagrammer
- influence score: Influence score of the users. It is calculated on basis of mentions, importance and popularity
- posts: Number of posts they have made so far
- followers: Number of followers of the user
- avg_likes: Average likes on instagrammer posts (total likes/ total posts)
- 60_day_eng_rate: Last 60 days engagement rate of instagrammer as faction of engagements they have done so far
- new_post_avg_like: Average likes they have on new posts
- total Likes: Total likes the user has got on their posts. (in Billion)
- country: Country or region of origin of the user.

METHODOLOGY

In this study we analyzed top influencers data using gephi tool and created some visualizations in Power BI. We also analyzed influencer's data in python with some of the machine learning models as follows:

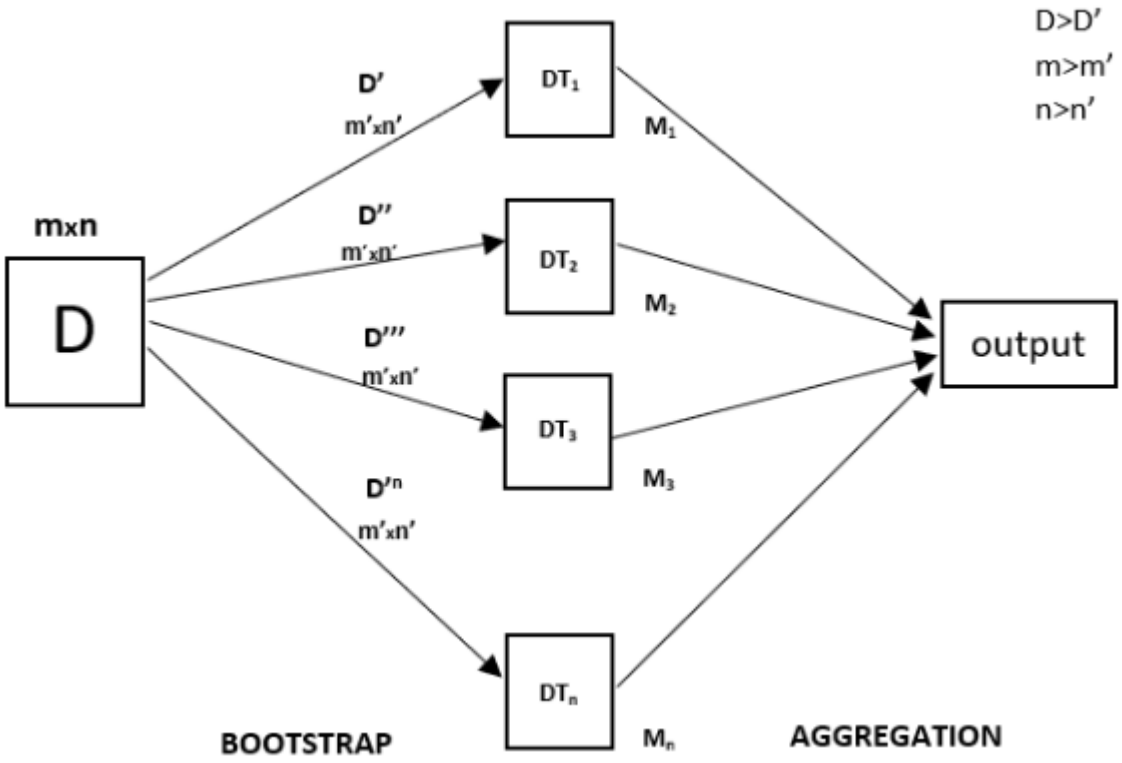
Decision tree Regressor:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.
Discrete output example: A weather prediction model that predicts whether or not there'll be rain on a particular day.
Continuous output example: A profit prediction model that states the probable profit that can be generated from the sale of a product.
Here, continuous values are predicted with the help of a decision tree regression model.



Random forest regressor:

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called Aggregation.

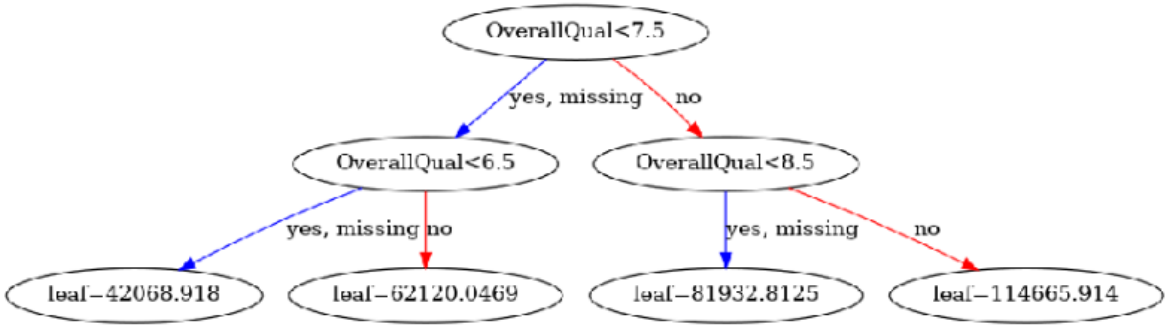


Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

XGB Regressor:

XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e. how far the model results are from the real values. The most common loss functions in XGBoost for regression problems are reg: linear, and that for binary classification is reg: logistics. Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction and XGBoost is one of the ensemble learning methods. XGBoost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancels out and better one sum up to form final good predictions.



RESULTS AND DISCUSSION

Data types:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   rank               200 non-null    int64
 1   channel_info       200 non-null    object
 2   influence_score    200 non-null    int64
 3   posts              200 non-null    object
 4   followers          200 non-null    object
 5   avg_likes          200 non-null    object
 6   60_day_eng_rate    200 non-null    object
 7   new_post_avg_like  200 non-null    object
 8   total_likes        200 non-null    object
 9   country            138 non-null    object
dtypes: int64(2), object(8)
memory usage: 15.8+ KB
Index(['rank', 'channel_info', 'influence_score', 'posts', 'followers',
       'avg_likes', '60_day_eng_rate', 'new_post_avg_like', 'total_likes',
       'country'],
      dtype='object')
```

Dataset:

rank	channel_info	influence_score	posts	followers	avg_likes	60_day_eng_rate	new_post_avg_like	total_likes	country
1	cristiano	92	3.3k	475.8m	8.7m	1.39%	6.5m	29.0b	Spain
2	kyliejenner	91	6.9k	366.2m	8.3m	1.62%	5.9m	57.4b	United States
3	leomessi	90	0.89k	357.3m	6.8m	1.24%	4.4m	6.0b	NaN
4	selenagomez	93	1.8k	342.7m	6.2m	0.97%	3.3m	11.5b	United States
5	therock	91	6.8k	334.1m	1.9m	0.20%	665.3k	12.5b	United States

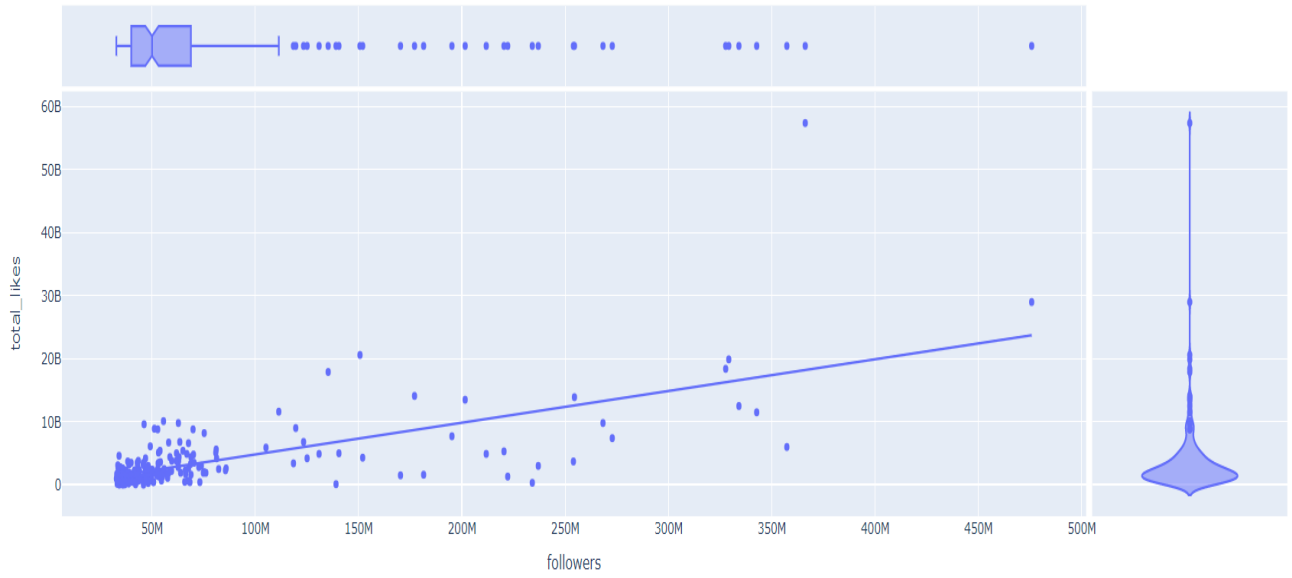
Change of Data types:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 200 entries, 1 to 200
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   channel_info       200 non-null    object 
 1   influence_score    200 non-null    int64  
 2   posts              200 non-null    float64
 3   followers          200 non-null    float64
 4   avg_likes          200 non-null    float64
 5   60_day_eng_rate    199 non-null    float64
 6   new_post_avg_like  200 non-null    float64
 7   total_likes        200 non-null    float64
 8   country            138 non-null    object 
dtypes: float64(6), int64(1), object(2)
memory usage: 15.6+ KB
```

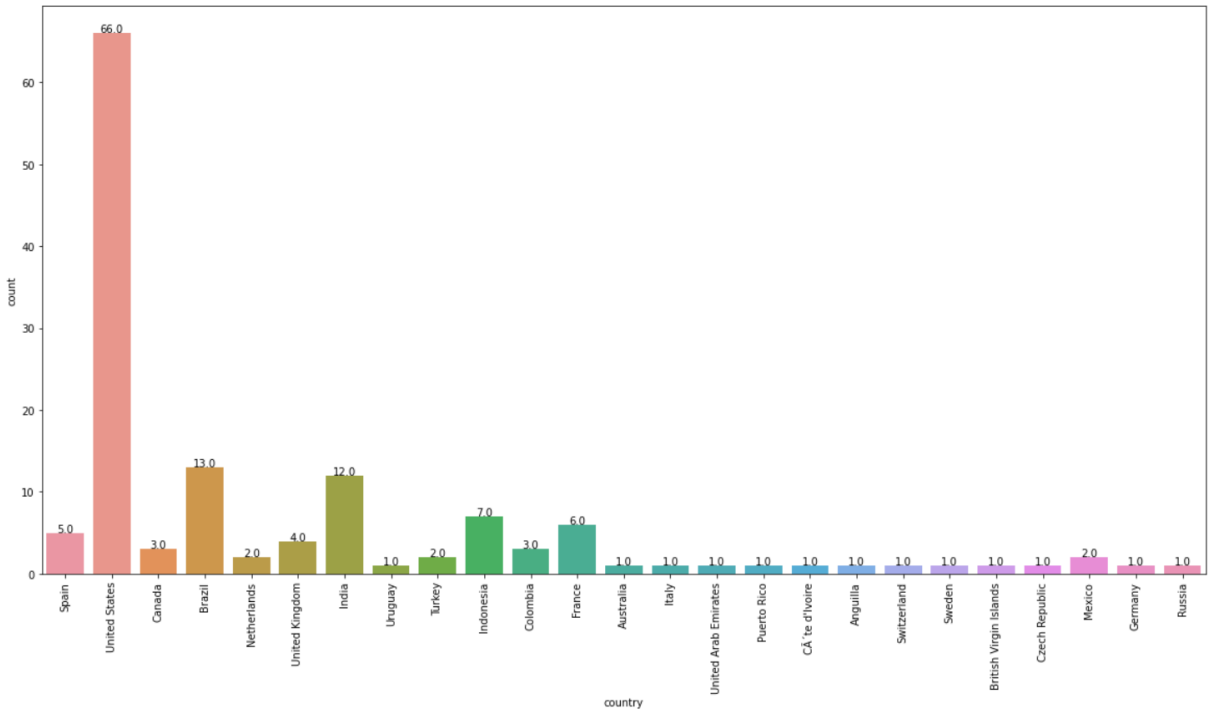
Statistical values of the data :

	influence_score	posts	followers	avg_likes	60_day_eng_rate	new_post_avg_like	total_likes
count	200.000000	200.000000	2.000000e+02	2.000000e+02	199.000000	2.000000e+02	2.000000e+02
mean	81.820000	3499.850000	7.740950e+07	1.787104e+06	1.902010	1.208132e+06	3.658112e+09
std	8.878159	3475.828158	7.368727e+07	2.193359e+06	3.329719	1.858322e+06	5.561939e+09
min	22.000000	10.000000	3.280000e+07	6.510000e+04	0.010000	0.000000e+00	1.830000e+07
25%	80.000000	947.500000	4.000000e+07	5.044000e+05	0.410000	1.957500e+05	9.968500e+08
50%	84.000000	2100.000000	5.005000e+07	1.100000e+06	0.880000	5.321500e+05	2.000000e+09
75%	86.000000	5025.000000	6.890000e+07	2.100000e+06	2.035000	1.325000e+06	3.900000e+09
max	93.000000	17500.000000	4.758000e+08	1.540000e+07	26.410000	1.260000e+07	5.740000e+10

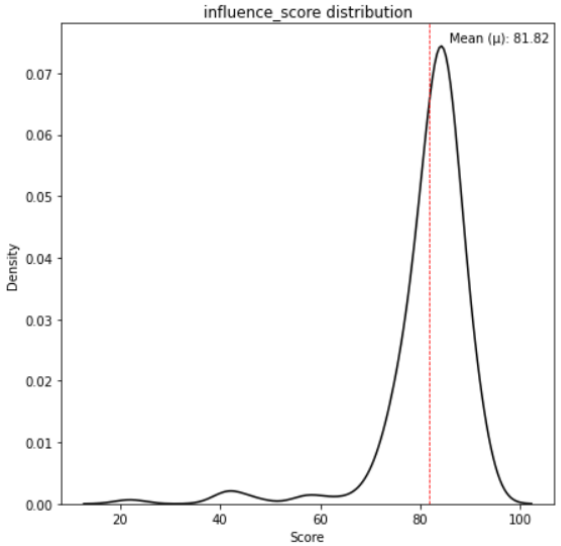
Distribution of total likes w.r.t engagement rate

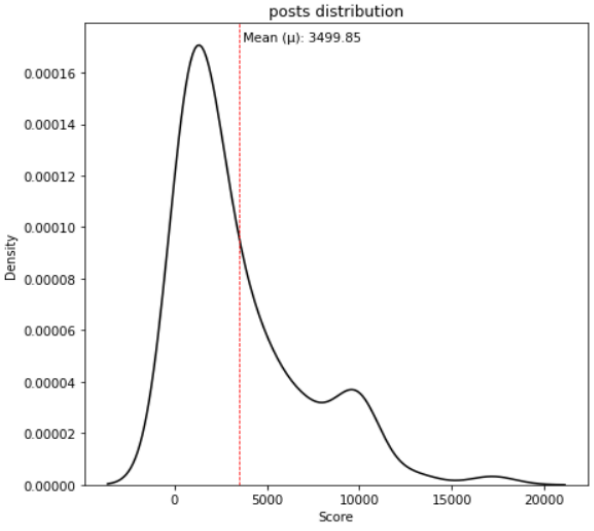


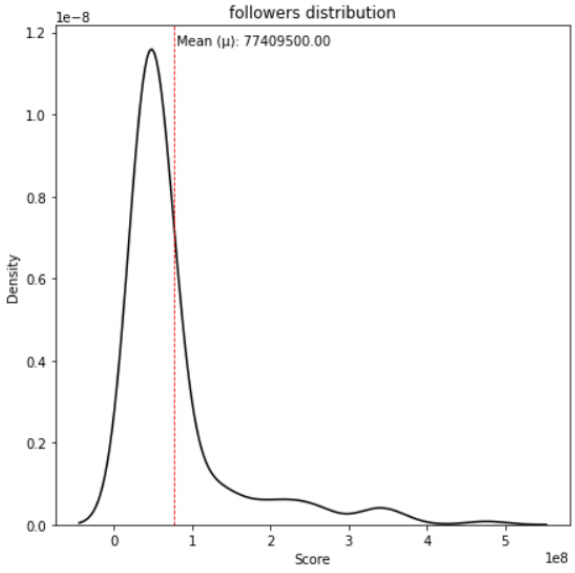
Distribution of countries

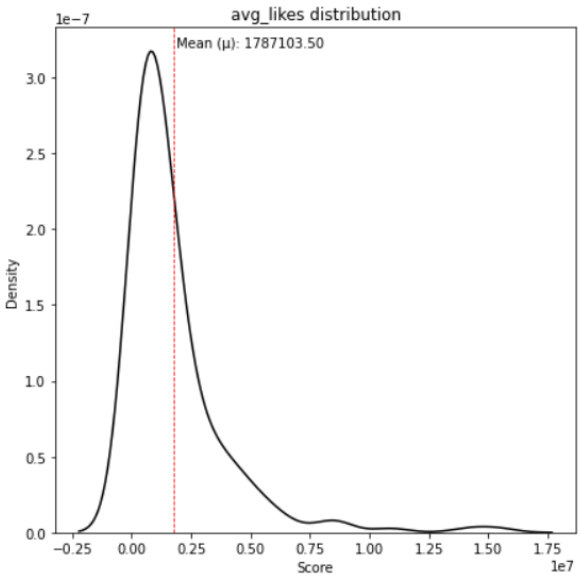


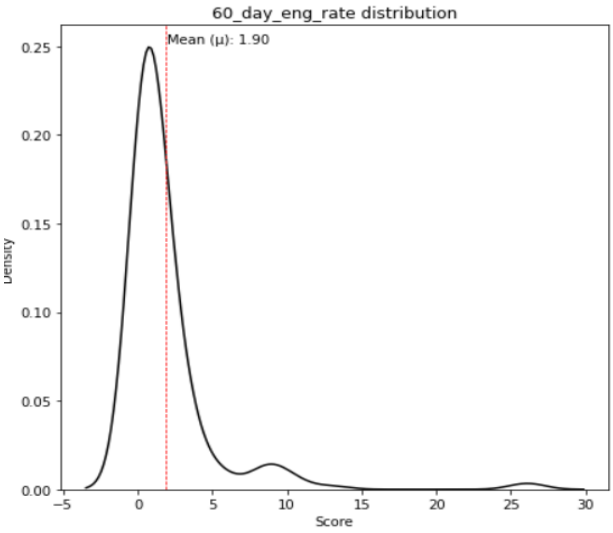
Distribution of number features

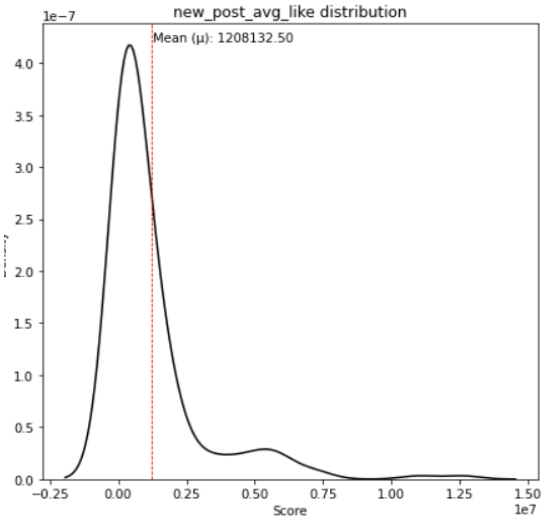


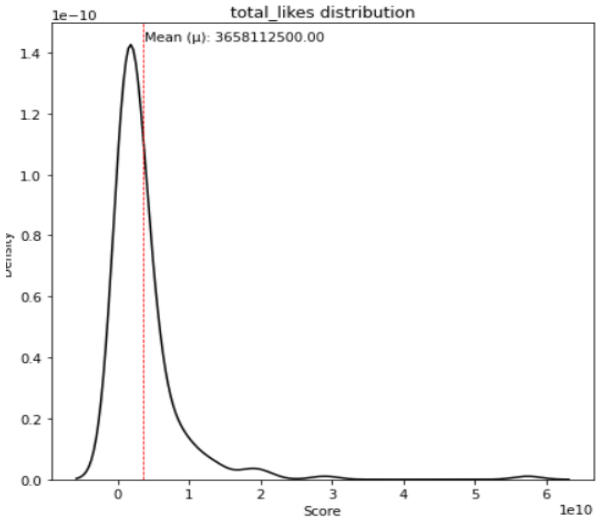



followers distribution
1e−8
Mean (μ): 77409500.00
Density
Score
1e8

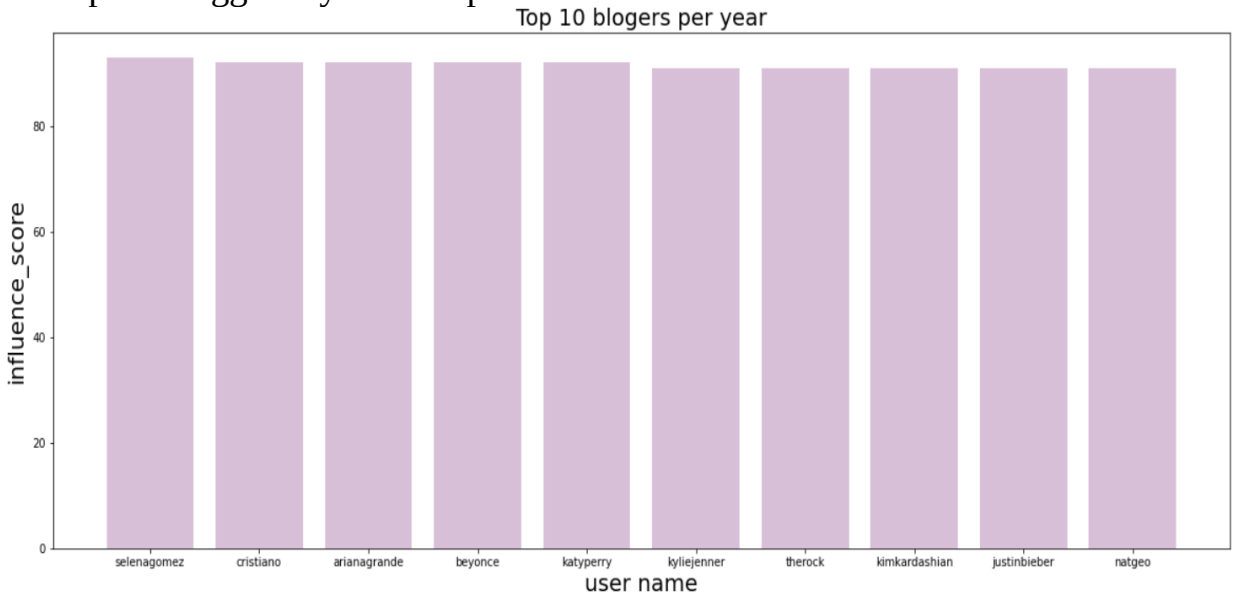

avg_likes distribution
1e−7
Mean (μ): 1787103.50
Density
Score
1e7

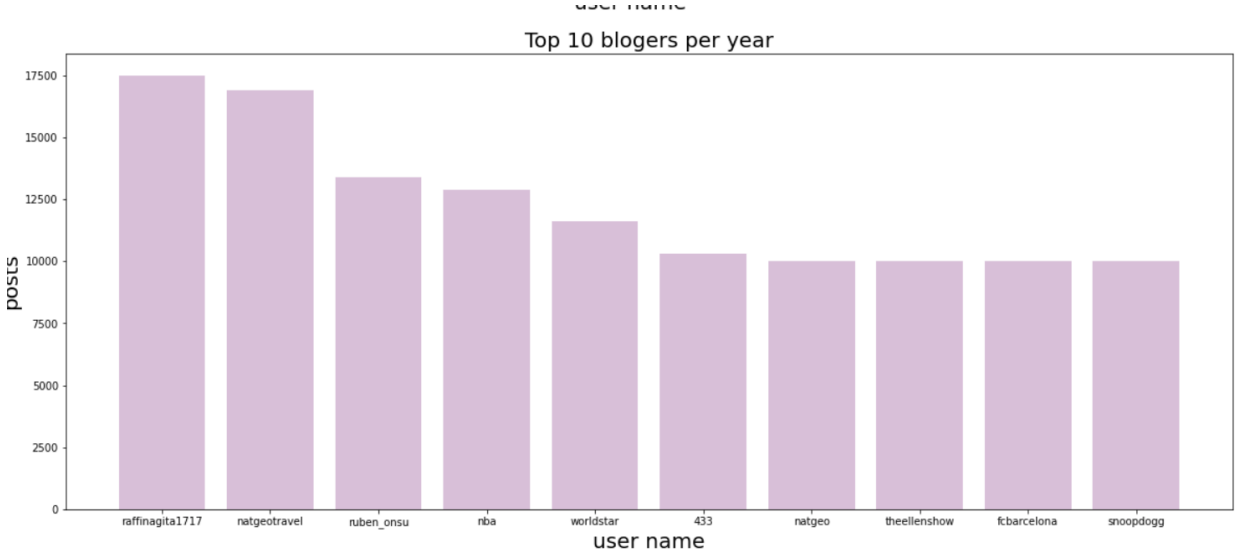

60_day_eng_rate distribution
Mean (μ): 1.90
Density
Score

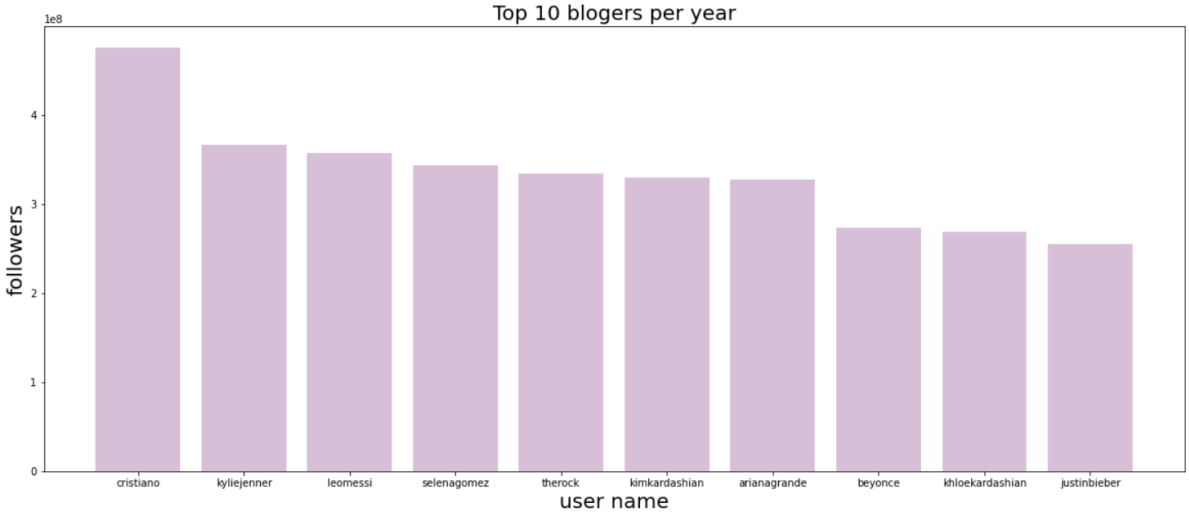

new_post_avg_like distribution
1e−7
Mean (μ): 1208132.50
Score
1e7

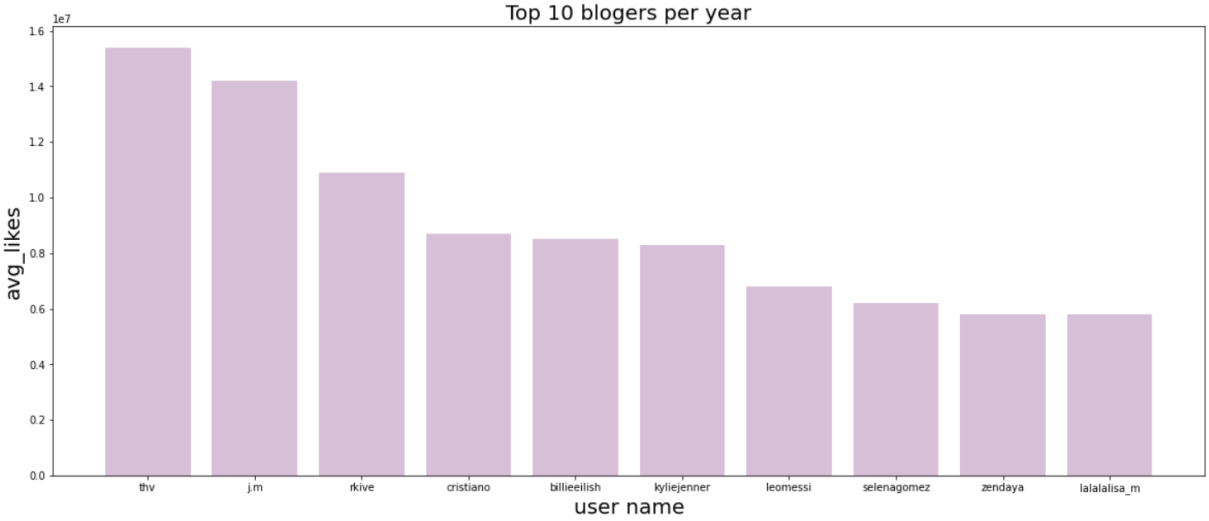


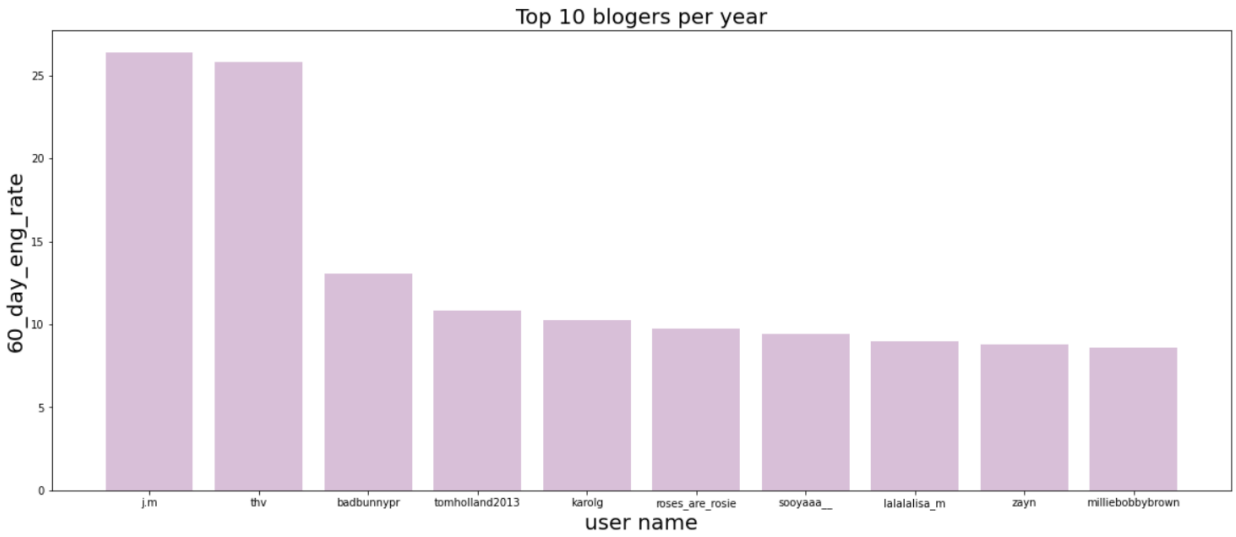
The top 10 bloggers by various parameters

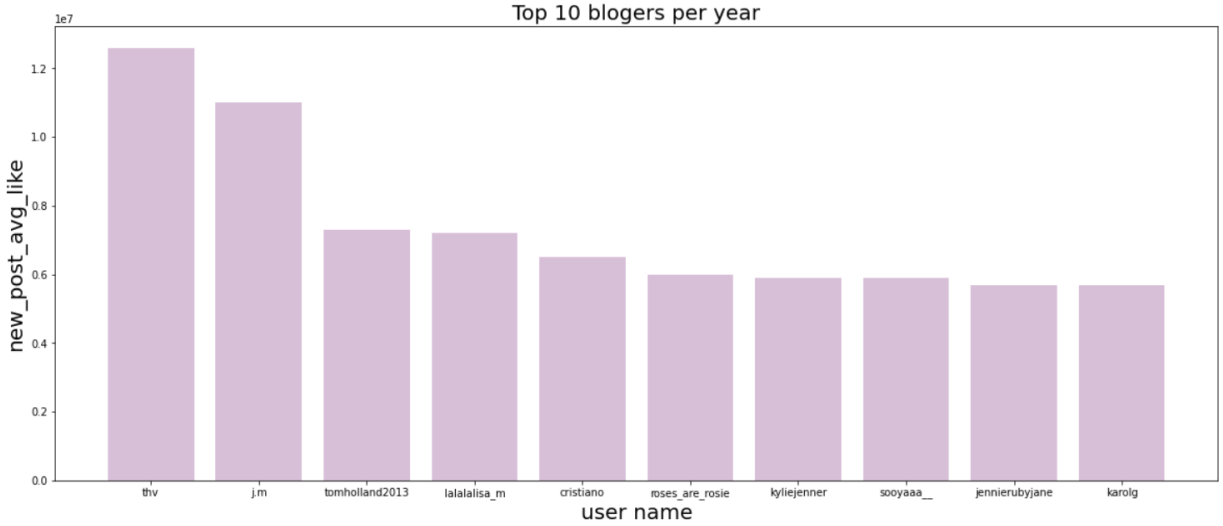


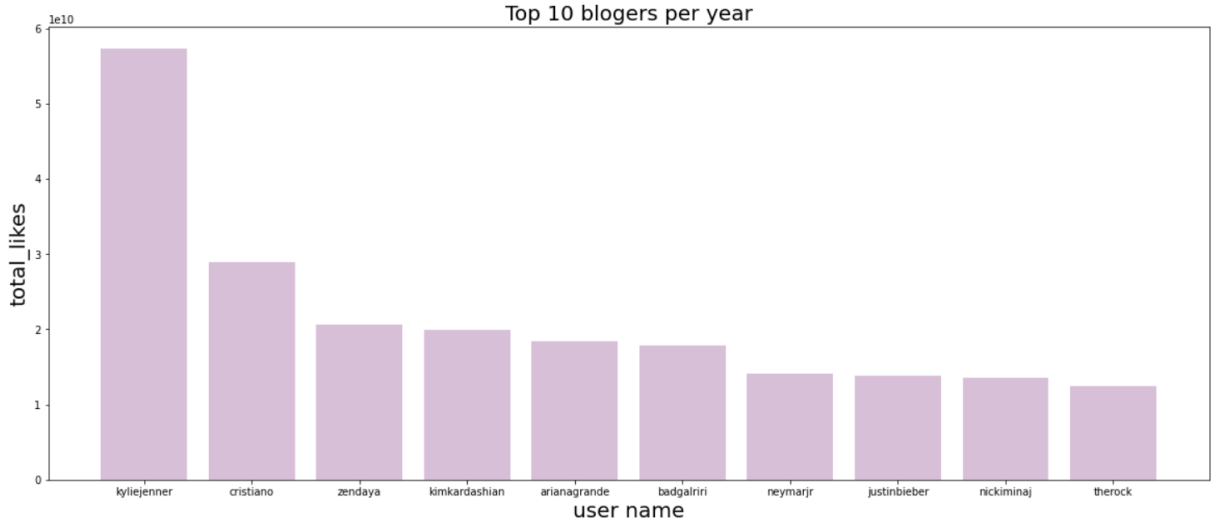

Top 10 blogers per year
posts
user name
raffinagita1717
natgeotravel
ruben_onsu
nba
worldstar
433
natgeo
theellenshow
fcbarcelona
snoopdogg


Top 10 blogers per year
1e8
followers
user name
cristiano
kyliejenner
leomessi
selenagomez
therock
kimkardashian
arianagrande
beyonce
khloekardashian
justinbieber

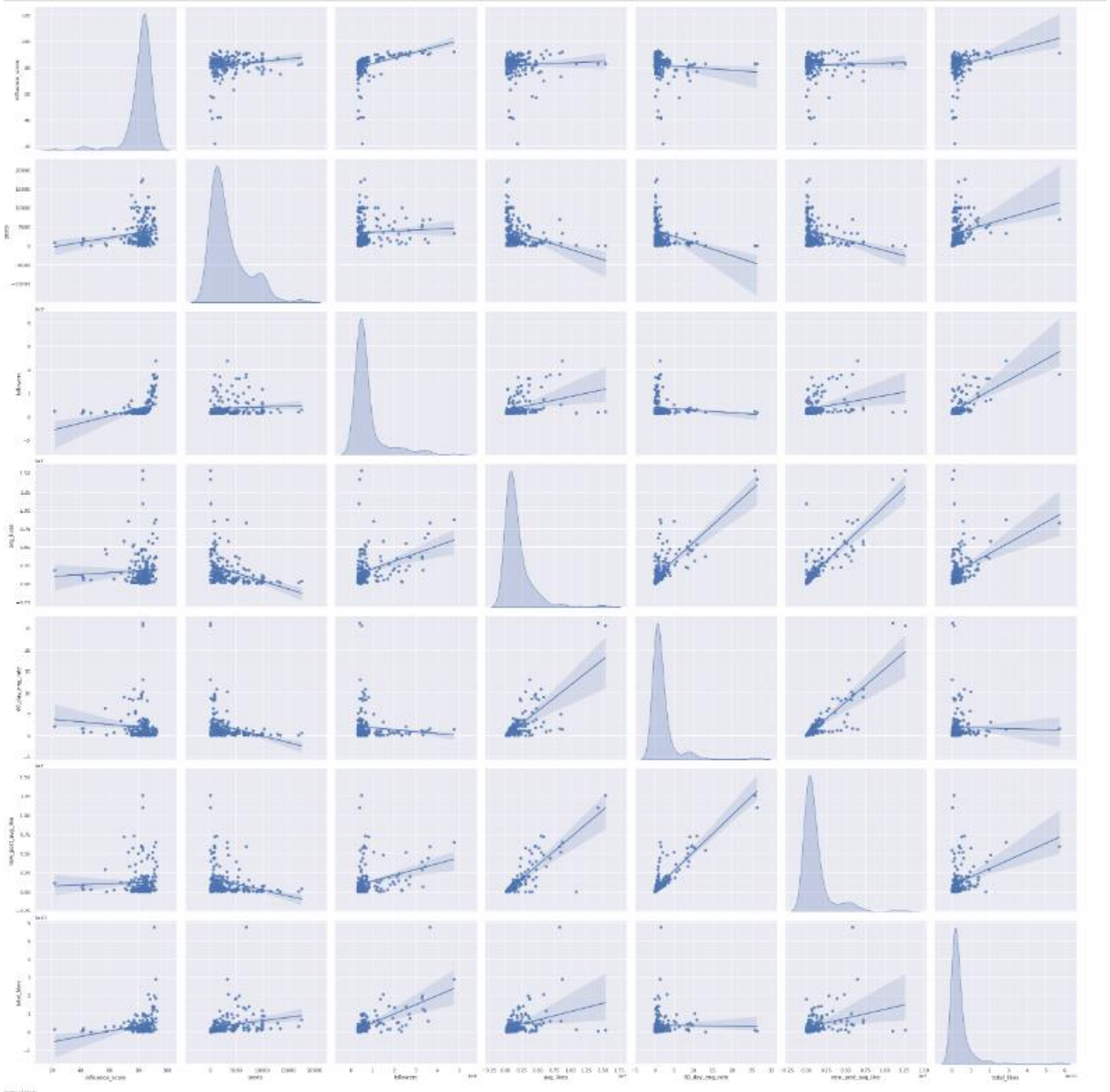

Top 10 blogers per year
1e7
avg_likes
user name
thv
j.m
rkive
cristiano
billieeilish
kyliejenner
leomessi
selenagomez
zendaya
lalalalisa_m


Top 10 blogers per year
60_day_eng_rate
0
5
10
15
20
25
j.m
thv
badbunnypr
tomholland2013
karolg
roses_are_rosie
sooyaaa__
lalalalisa_m
zayn
milliebobbybrown
user name


Top 10 blogers per year
1e7
new_post_avg_like
0.0
0.2
0.4
0.6
0.8
1.0
1.2
thv
j.m
tomholland2013
lalalalisa_m
cristiano
roses_are_rosie
kyliejenner
sooyaaa__
jennierubyjane
karolg
user name


Top 10 blogers per year
1e10
total_likes
0
1
2
3
4
5
6
kyliejenner
cristiano
zendaya
kimkardashian
arianagrande
badgalriri
neymarjr
justinbieber
nickiminaj
therock
user name

The relationship between the various features



Statiscal tests:

```
Statistics=0.674, p=0.000
Sample does not look Gaussian (reject H0)
Statistics=0.588, p=0.000
Sample does not look Gaussian (reject H0)
Statistics=0.517, p=0.000
Sample does not look Gaussian (reject H0)
Statistics=0.832, p=0.000
Sample does not look Gaussian (reject H0)
```

Correlation between features

Decision tree regressor

```
Making predictions for the following 5 influencers:
      influence_score    posts   followers  new_post_avg_like
rank
151                81   4500.0  40000000.0            87400.0
152                75  10000.0  39900000.0           162900.0
153                80   2300.0  39900000.0           181100.0
154                78  11600.0  39200000.0            84400.0
155                84   1200.0  39200000.0           227700.0
The predictions are
[1.00e+09 3.80e+09 1.70e+09 1.50e+09 7.84e+08]
```

```
array([1.000e+09, 3.800e+09, 1.700e+09, 1.500e+09, 7.840e+08, 4.520e+08,
       2.800e+09, 1.800e+09, 1.600e+09, 1.500e+09, 7.107e+08, 4.235e+08,
       9.800e+09, 7.107e+08, 2.900e+09, 5.807e+08, 7.840e+08, 4.520e+08,
       7.107e+08, 7.107e+08, 1.600e+09, 7.670e+08, 1.500e+09, 7.670e+08,
       3.681e+08, 7.670e+08, 1.800e+09, 2.800e+09, 7.107e+08, 7.107e+08,
       1.600e+09, 2.400e+09, 8.630e+08, 7.107e+08, 1.500e+09, 3.500e+09,
       4.520e+08, 4.520e+08, 2.400e+09, 3.500e+09, 1.000e+09, 3.000e+09,
       1.600e+09, 7.107e+08, 1.600e+09, 1.800e+09, 1.600e+09, 7.107e+08,
       1.800e+09, 1.000e+09])
```

```
The desicion tree regressor MAE is:
675290000.0
```

Random forest Regressor:

```
Random Forest Regressor MAE is:
676570280.0
```

```
Making predictions for the following 5 influencers:
      influence_score    posts   followers  new_post_avg_like
rank
151                81   4500.0  40000000.0            87400.0
152                75  10000.0  39900000.0           162900.0
153                80   2300.0  39900000.0           181100.0
154                78  11600.0  39200000.0            84400.0
155                84   1200.0  39200000.0           227700.0
The predictions are
[1.948633e+09 3.001415e+09 1.403513e+09 2.295820e+09 9.218200e+08]
```

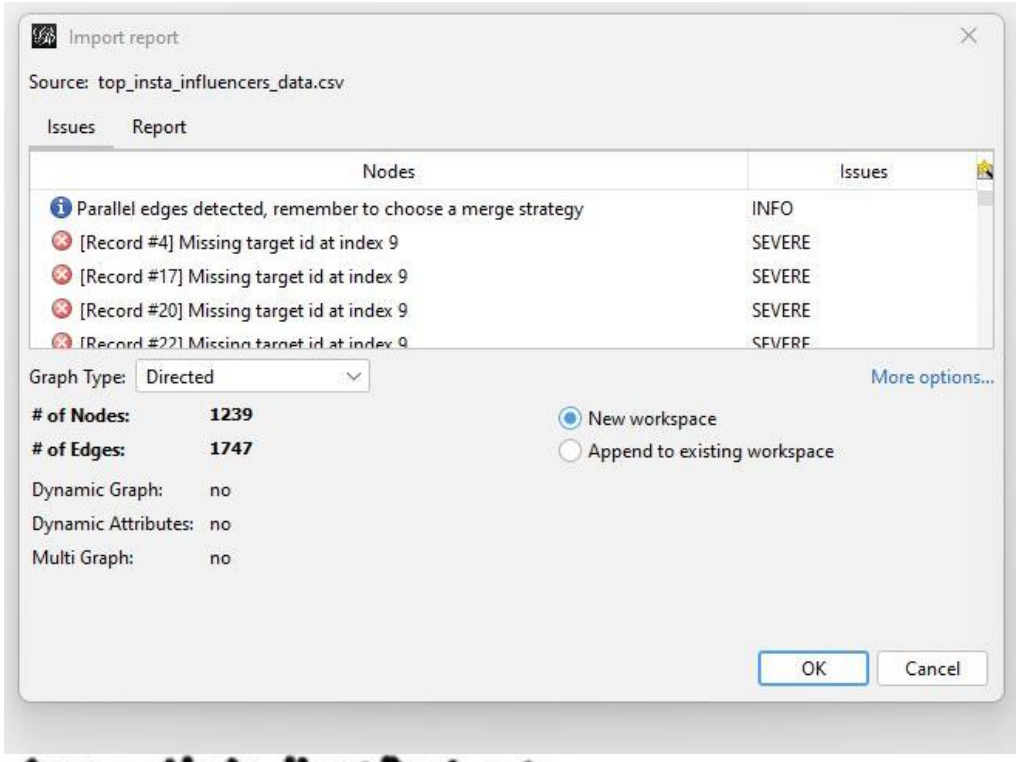
```
array([1.000e+09, 3.800e+09, 1.700e+09, 1.500e+09, 7.840e+08, 4.520e+08,
       2.800e+09, 1.800e+09, 1.600e+09, 1.500e+09, 7.107e+08, 4.235e+08,
       9.800e+09, 7.107e+08, 2.900e+09, 5.807e+08, 7.840e+08, 4.520e+08,
       7.107e+08, 7.107e+08, 1.600e+09, 7.670e+08, 1.500e+09, 7.670e+08,
       3.681e+08, 7.670e+08, 1.800e+09, 2.800e+09, 7.107e+08, 7.107e+08,
       1.600e+09, 2.400e+09, 8.630e+08, 7.107e+08, 1.500e+09, 3.500e+09,
       4.520e+08, 4.520e+08, 2.400e+09, 3.500e+09, 1.000e+09, 3.000e+09,
       1.600e+09, 7.107e+08, 1.600e+09, 1.800e+09, 1.600e+09, 7.107e+08,
       1.800e+09, 1.000e+09])
```

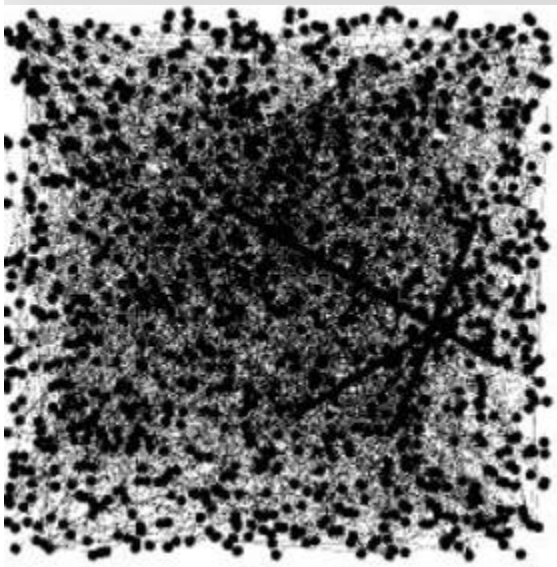
XGB Regressor

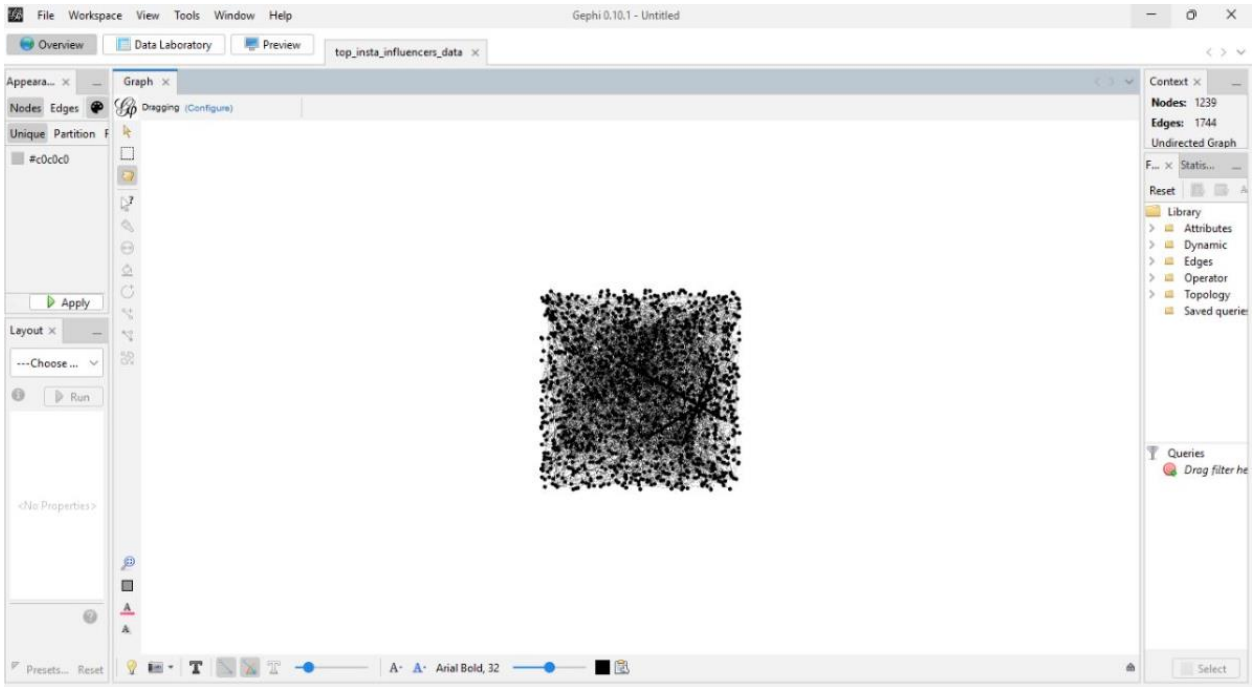
```
XGBRegressor MAE is:
525249457.92
```

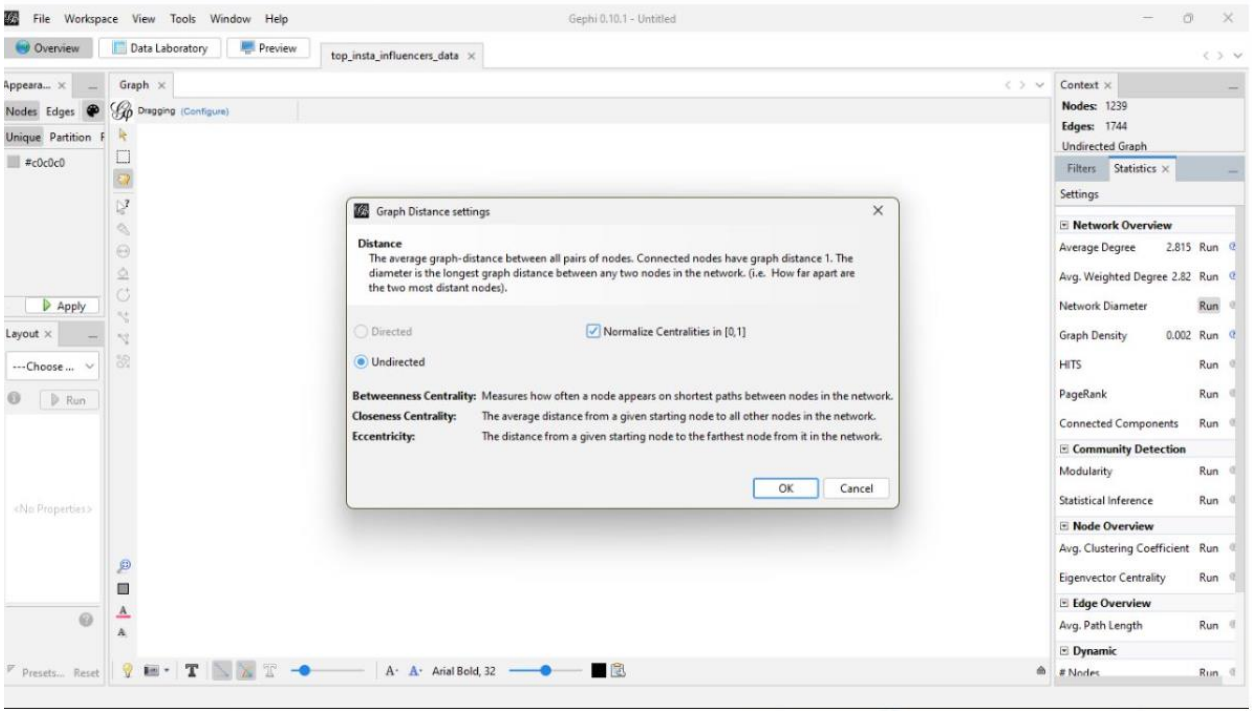
```
array([1.000e+09, 3.800e+09, 1.700e+09, 1.500e+09, 7.840e+08, 4.520e+08,
       2.800e+09, 1.800e+09, 1.600e+09, 1.500e+09, 7.107e+08, 4.235e+08,
       9.800e+09, 7.107e+08, 2.900e+09, 5.807e+08, 7.840e+08, 4.520e+08,
       7.107e+08, 7.107e+08, 1.600e+09, 7.670e+08, 1.500e+09, 7.670e+08,
       3.681e+08, 7.670e+08, 1.800e+09, 2.800e+09, 7.107e+08, 7.107e+08,
       1.600e+09, 2.400e+09, 8.630e+08, 7.107e+08, 1.500e+09, 3.500e+09,
       4.520e+08, 4.520e+08, 2.400e+09, 3.500e+09, 1.000e+09, 3.000e+09,
       1.600e+09, 7.107e+08, 1.600e+09, 1.800e+09, 1.600e+09, 7.107e+08,
       1.800e+09, 1.000e+09])
```

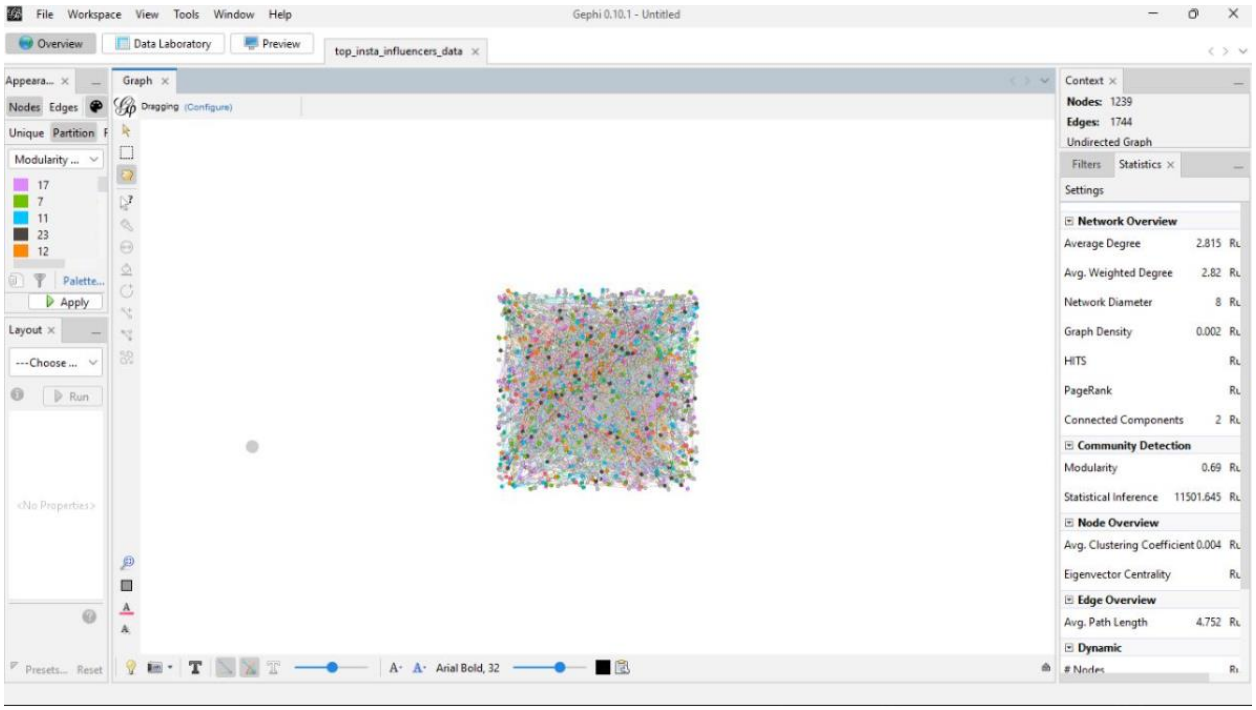
GEPHI TOOL


Import report
Source: top_insta_influencers_data.csv
Issues
Report
Nodes
Issues
Parallel edges detected, remember to choose a merge strategy
INFO
[Record #4] Missing target id at index 9
SEVERE
[Record #17] Missing target id at index 9
SEVERE
[Record #20] Missing target id at index 9
SEVERE
Graph Type:
Directed
More options...
# of Nodes:
1239
# of Edges:
1747
New workspace
Append to existing workspace
Dynamic Graph:
no
Dynamic Attributes:
no
Multi Graph:
no
OK
Cancel

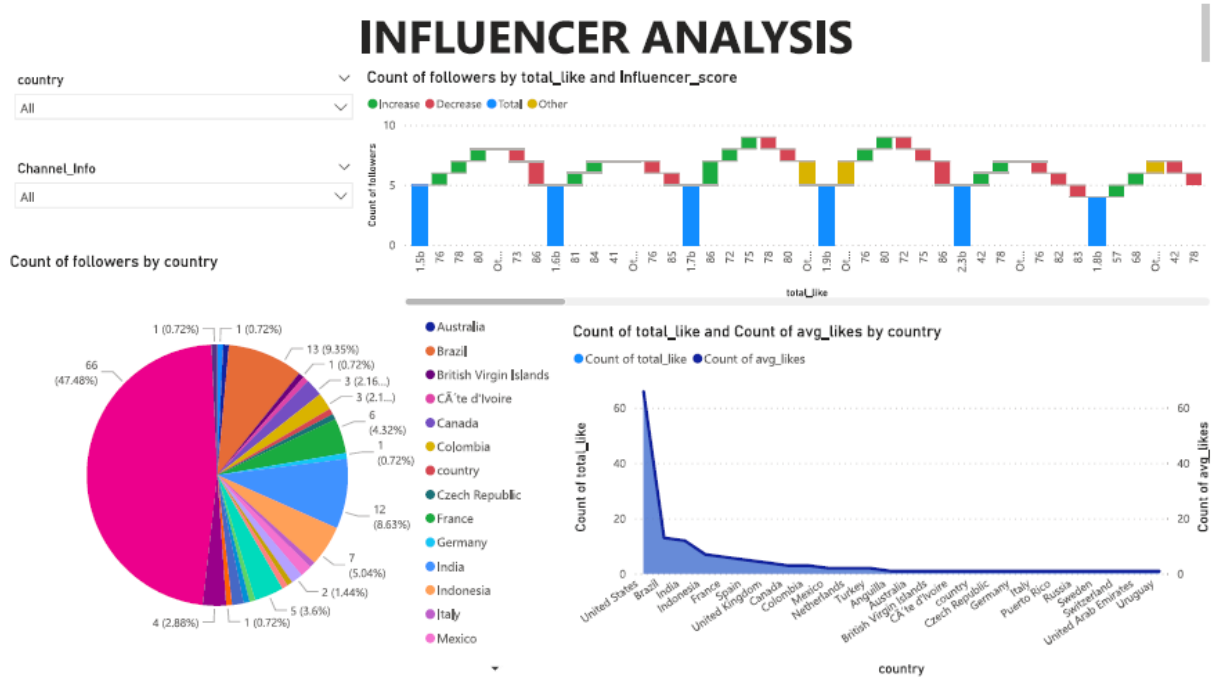



File Workspace View Tools Window Help
Gephi 0.10.1 - Untitled
Overview
Data Laboratory
Preview
top_insta_influencers_data
Graph
Dragging (Configure)
Context
Nodes: 1239
Edges: 1744
Undirected Graph
Library
Attributes
Dynamic
Edges
Operator
Topology
Saved queries
Queries
Apply
Layout
---Choose ...
Run
Presets... Reset
Arial Bold, 32
Select


File Workspace View Tools Window Help
Gephi 0.10.1 - Untitled
Overview
Data Laboratory
Preview
top_insta_influencers_data
Graph
Dragging (Configure)
Context
Nodes: 1239
Edges: 1744
Undirected Graph
Filters
Statistics
Settings
Network Overview
Average Degree 2.815 Run
Avg. Weighted Degree 2.82 Run
Network Diameter Run
Graph Density 0.002 Run
HITS Run
PageRank Run
Connected Components Run
Community Detection
Modularity Run
Statistical Inference Run
Node Overview
Avg. Clustering Coefficient Run
Eigenvector Centrality Run
Edge Overview
Avg. Path Length Run
Dynamic
Graph Distance settings
Distance
The average graph-distance between all pairs of nodes. Connected nodes have graph distance 1. The diameter is the longest graph distance between any two nodes in the network. (i.e. How far apart are the two most distant nodes).
Directed
Normalize Centralities in [0,1]
Undirected
Betweenness Centrality: Measures how often a node appears on shortest paths between nodes in the network.
Closeness Centrality: The average distance from a given starting node to all other nodes in the network.
Eccentricity: The distance from a given starting node to the farthest node from it in the network.
OK
Cancel
Apply
Layout
---Choose ...
Run
Presets... Reset
Arial Bold, 32


File Workspace View Tools Window Help
Gephi 0.10.1 - Untitled
Overview
Data Laboratory
Preview
top_insta_influencers_data
Graph
Context
Nodes: 1239
Edges: 1744
Undirected Graph
Filters
Statistics
Network Overview
Average Degree 2.815
Avg. Weighted Degree 2.82
Network Diameter 8
Graph Density 0.002
HITS
PageRank
Connected Components 2
Community Detection
Modularity 0.69
Statistical Inference 11501.645
Node Overview
Avg. Clustering Coefficient 0.004
Eigenvector Centrality
Edge Overview
Avg. Path Length 4.752
Dynamic

PowerBI


INFLUENCER ANALYSIS
country
All
Channel_Info
All
Count of followers by total_like and Influencer_score
Increase Decrease Total Other
Count of followers
total_like
Count of followers by country
Australia
Brazil
British Virgin Islands
CÃ´te d'Ivoire
Canada
Colombia
country
Czech Republic
France
Germany
India
Indonesia
Italy
Mexico
Count of total_like and Count of avg_likes by country
Count of total_like
Count of avg_likes
country

CONCLUSION

The process of arriving at a buying decision has changed dramatically in recent years. Consumers are done with brands that boast about their products and services mindlessly. They now need something more authentic to convince them to go for a product or service. This has increased the value of influencer marketing tenfold. Brand managers can no longer take social media lightly because their sales heavily depend on the same. Brands now have to compete on social media as well to keep their target audience intact. Thus analysis of influencers plays an important role in the current situation.