INSTAGRAM INFLUENCER ANALYTICS FOR BRAND PROMOTIONS

Project submitted to the

APSSDC

Bachelor of Technology

In

Information Technology

Submitted By

REDDY SURESH - 24P35A1212

Table of Contents

ABSTRACT:

Influencer marketing has become a powerful tool for brands to connect with their target audience, especially on platforms like Instagram. However, selecting the right influencers can be difficult due to the presence of fake engagement, inconsistent content, and a lack of reliable performance data. This project aims to solve that problem by creating a clean, structured system to identify and rank top-performing Instagram influencers across various niches.

Using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn, a synthetic dataset of 500 influencers was generated. Important metrics like engagement rate, authenticity score, influencer score, and promotions rate were calculated. Data cleaning and preprocessing steps were applied to remove incomplete or misleading entries, and visualization techniques were used to understand patterns in the data. A key feature of the project is the calculation of “Expected Money,” which helps estimate the value each influencer might bring to a brand campaign.

The final output is a ranked list of the top 5 influencers in each niche, based on their performance and promotional potential. This data-driven approach gives businesses a smarter way to plan promotions, reduce marketing risks, and improve return on investment. It simplifies influencer selection and provides a reliable starting point for brand collaborations.

INTRODUCTION:

In today's digital world, **influencer marketing** has become a powerful strategy for brands to promote their products and build customer trust. Among various platforms, **Instagram** stands out as one of the most influential spaces, where content creators reach thousands or even millions of followers through visual and engaging content. Companies are increasingly investing in collaborations with influencers to enhance brand visibility and drive sales.

- However, choosing the right influencer is not as simple as it may seem. Many influencers have large follower counts but poor engagement or even fake interactions. This can lead to **low campaign performance**, wasted marketing budgets, and missed opportunities. Therefore, businesses need a **reliable, data-driven method** to identify the most suitable influencers for their promotions.

- This project provides a solution by analyzing Instagram influencers using Python and key data science libraries such as **Pandas, NumPy, Seaborn, and Matplotlib**. It focuses on generating a synthetic dataset of influencers from various niches like Fitness, Fashion, Health, and Technology. Important metrics such as **Engagement Rate**, **Influencer Score**, and **Promotions Rate** are calculated to measure the value each influencer offers.

- The ultimate goal is to **rank influencers within their niches** and estimate their **Expected Money**, which represents potential earnings in a promotional campaign. This helps companies make better marketing decisions, **reduce risk**, and improve campaign

performance by partnering with influencers who truly drive results.

- 📱 **Instagram is a leading platform** for influencer marketing, used by brands to reach targeted audiences.
- ⚠️ **Selecting the right influencer is difficult** due to fake followers, low engagement, and misleading metrics.
- 🎯 **Brands risk wasting time and money** on ineffective promotions without data-backed decisions.
- 📊 This project provides a **data-driven solution** to rank influencers based on real performance metrics.
- 🧠 **Metrics used include:** Engagement Rate, Influencer Score, Promotions Rate, and Expected Money.
- The system helps companies make **smarter, faster, and more effective** influencer marketing choices.

SYSTEM REQUIREMENTS:

This project was developed using basic tools that are freely available and easy to set up. The following hardware and software components are needed to run the analysis smoothly:

Hardware Requirements

- A laptop or desktop with at least:
    - **4 GB RAM**
    - **Dual-core processor**
    - **500 MB of available disk space**

Software Requirements

- **Operating System:**
    - Windows 10 / 11, macOS, or any Linux distribution
- **Python Environment:**
    - **Python 3.8 or higher**
    - IDEs such as **Jupyter Notebook**, **VS Code**, or **Google Colab**

Python Libraries Used

- `pandas` – For data loading and manipulation
- `numpy` – For numerical operations
- `matplotlib` – For creating basic plots
- `seaborn` – For advanced visualizations
- `scikit-learn` – For normalization and data scaling

To solve this problem, we created a synthetic dataset of 500 Instagram influencers across different niches such as Fitness, Fashion, Health, and Technology. We used Python libraries like **Pandas**, **NumPy**, **Matplotlib**, and **Seaborn** to process the data, clean it, and visualize it. Important metrics like **Engagement Rate**, **Influencer Score**, **Promotions Rate**, and **Expected Money** were calculated to rank the top influencers in each niche. The final output is a CSV file with the top influencers per category, helping brands make smarter and more confident marketing decisions.

4. Project Architecture

The architecture of this project is designed as a step-by-step workflow, starting from generating and importing influencer data, cleaning and exploring it, then performing calculations and visualizations. Each phase plays an important role in ensuring that the final influencer ranking is accurate, reliable, and useful for brands. The architecture is implemented using Python and popular data libraries like Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn.

This section explains the core steps that make up the project's logical flow:

4.1 Importing Dataset (Data Acquisition)

The project begins by creating a synthetic dataset of Instagram influencers. This data is generated using Python's NumPy library to simulate realistic values for followers, likes, comments, and authenticity scores. Once generated, the dataset is structured using Pandas and saved as a CSV file.

The dataset includes key fields such as:

- **Influencer ID**
- **Niche (category)**
- **Followers**
- **Average Likes**
- **Average Comments**
- **Authenticity Score**
- **Engagement Rate**

- **Influencer Score**
- **Promotions Rate**

The goal of this step is to create a foundational dataset that mirrors real-world influencer statistics across various niches like Fashion, Fitness, Technology, Health, and Education.

4.2 Data Exploration

Once the dataset is ready, the next step is to explore the data and understand its structure. Using Pandas functions such as `.head()`, `.info()`, and `.describe()`, we examine:

- **Number of records and columns**
- **Data types (e.g., int, float, object)**
- **Summary statistics (mean, max, min, std)**
- **Any missing values (NaNs)**

Data exploration helps reveal patterns and outliers. For instance, we may notice that some influencers have extremely high followers but low engagement, or that certain niches are more active than others. This insight is important before proceeding to cleaning or scoring.

4.3 Data Cleaning

Data cleaning is essential for ensuring the quality and consistency of the analysis. In this phase, we handle:

- **Missing values (using `.dropna()` or `.fillna()`)**
- **Recalculation of metrics like Engagement Rate**

- Checking for inconsistent or unrealistic values (e.g., zero followers)

Engagement Rate is recalculated using:

```
Engagement Rate = (Avg Likes + Avg Comments) / Followers
```

After cleaning, only complete and reliable influencer records are retained. This ensures that further analysis, especially ranking, is not affected by gaps in the data.

4.4 Feature Engineering

Feature engineering involves creating new calculated metrics from the existing data that better represent influencer performance. The key features developed in this project are:

- Influencer Score
  A combination of engagement rate and authenticity score:
  ```
  Influencer Score = 0.6 * Engagement Rate + 0.4 * Authenticity Score
  ```

- Promotions Rate
  A normalized combination of Followers, Engagement Rate, and Influencer Score using MinMaxScaler.

- Expected Money
  A calculated estimation of how much a company might pay the influencer, based on their total performance impact.

This phase turns raw data into meaningful indicators that can be compared across all influencers.

4.5 Influencer Ranking Logic

To help brands identify top-performing influencers, we apply a ranking logic that filters the best candidates from each niche. Here's how it works:

- **Sort influencers by:**
    - **Niche**
    - **Influencer Score (descending)**
    - **Promotions Rate (descending)**
- **Select the top 5 influencers in each niche using `groupby()` and `head()`.**

This approach ensures that companies see only the highest-impact influencers in their target niche, saving time and improving decision-making.

4.6 Expected Money Calculation

This step estimates the earning potential of each influencer. The Raw Earning Score is calculated as:

java

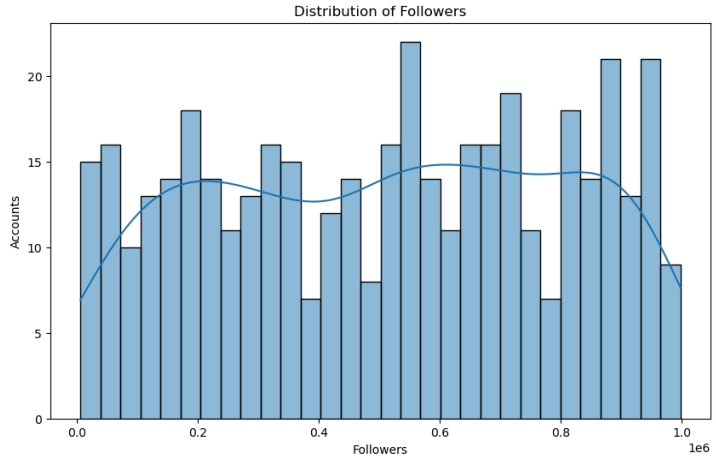
CopyEdit

```
Raw Score = Followers × Engagement Rate × Influencer Score × Promotions Rate
```

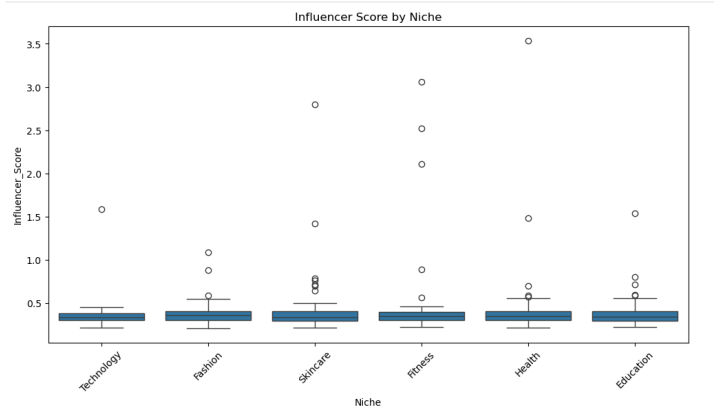
Then we scale the maximum Raw Score to ₹100,000, and calculate other influencers' `Expected Money` proportionally. This helps brands get a realistic idea of what they might pay an influencer based on performance—not just follower count.

This final value adds commercial value to the analysis and prepares the dataset for business use.

5.Data Visualization



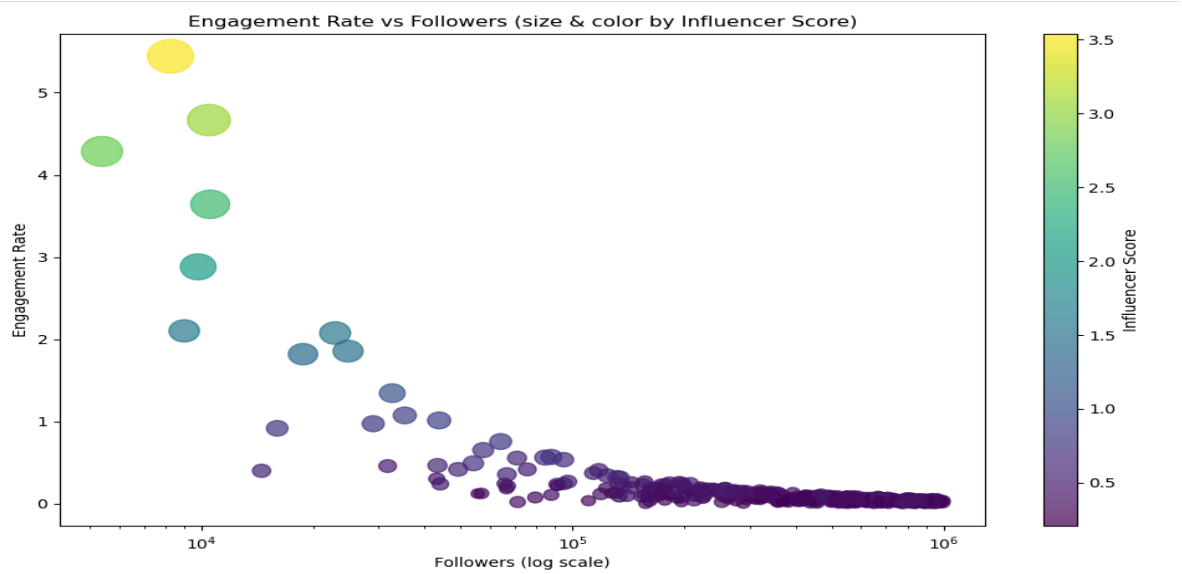
The histogram above illustrates the distribution of followers among influencers in the dataset. By using seaborn's histplot function with a kernel density estimate (KDE), we can observe how follower counts are spread across the sample. Most influencers fall within a lower to mid-range follower count, while fewer have very high follower numbers, indicating a right-skewed distribution. This insight helps identify the general reach of influencers and highlights the presence of both micro and macro influencers in the dataset.



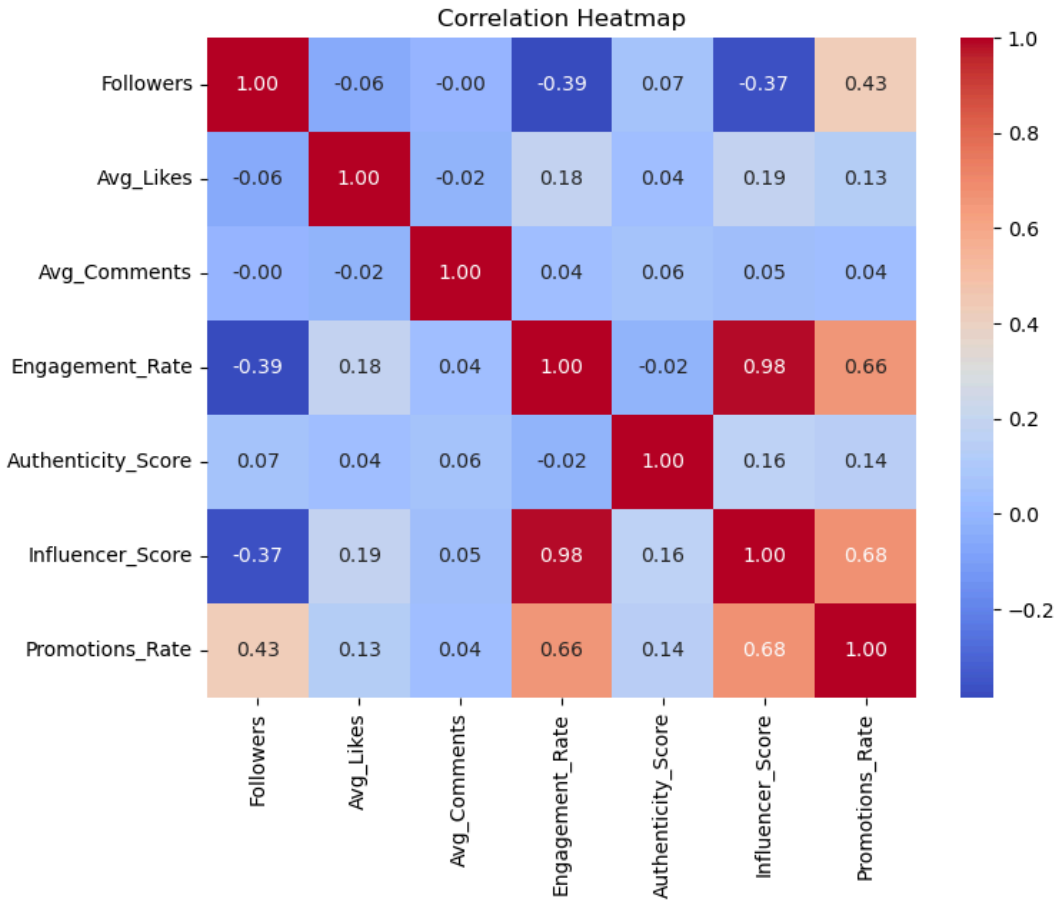
Influencer Score by Niche Box Plot:

This box plot compares the **Influencer Score** distribution across different content niches like Technology, Fashion, Fitness, etc. Most niches show similar median influencer scores, but the **Fitness**, **Health**, and **Skincare** categories have a larger number of high outliers, indicating the presence of some exceptionally impactful influencers in those fields. The spread is relatively narrow in all categories, implying that while outliers exist, most influencers in each niche have modest scores, reinforcing the need for careful individual evaluation rather than relying solely on niche averages.

Engagement Rate vs Followers Scatter Plot:



This bubble plot illustrates the relationship between **Engagement Rate** and **Follower Count** (on a log scale), with bubble size and color representing the **Influencer Score**. It clearly shows a negative trend: influencers with fewer followers tend to have significantly higher engagement rates and influencer scores. As follower count increases, both engagement and influencer score generally decrease, emphasizing that micro-influencers (with smaller audiences) may be more effective at driving interactions than large influencers with massive but less engaged audiences.



Correlation Heatmap:

This heatmap visualizes the pairwise correlation between different influencer-related metrics such as Followers, Engagement Rate, Authenticity Score, and Promotions Rate. Notably, there is a strong positive correlation between **Engagement Rate** and **Influencer Score** (0.98), as well as between **Influencer Score** and **Promotions Rate** (0.68), indicating that engagement plays a major role in determining influencer credibility and promotion frequency. Interestingly, **Followers** show a weak or negative correlation with key metrics like Engagement Rate (-0.39) and Influencer Score (-0.37), suggesting that follower count alone is not a good indicator of influence.

6.Uses of Python Libraries

6.1 Pandas

Pandas is a powerful Python library used for data manipulation and analysis. In this project, it plays a central role in reading the dataset, storing it in a structured format (DataFrame), and performing basic operations like viewing, sorting, and filtering data. Functions like `read_csv()`, `head()`, `describe()`, and `dropna()` were used to explore and clean the dataset efficiently.

Additionally, Pandas made it easy to create new columns such as **Engagement Rate** and **Influencer Score** by performing mathematical operations across columns. Its flexible syntax and built-in methods allowed for fast and clean handling of the dataset, making it an essential part of the analysis process.

6.2 NumPy

NumPy is a numerical computing library in Python that is especially useful for working with arrays and performing mathematical operations. In this project, NumPy was used to generate synthetic data such as random values for followers, likes, comments, and authenticity scores using functions like `np.random.randint()` and `np.random.uniform()`.

It also helped in performing fast element-wise operations when calculating metrics like **Engagement Rate** and **Expected Money**. NumPy's efficiency with large numerical datasets made the data generation and manipulation process smooth and accurate.

6.3 Matplotlib

Matplotlib is a basic but powerful plotting library that was used to create visual representations of the data. It allowed us to build line charts, bar graphs, histograms, and other plot types to help understand the data distribution and relationships between variables.

In this project, Matplotlib was mainly used alongside Seaborn to set figure sizes and customize titles, labels, and plot layouts. It served as the foundation for visual exploration, helping turn raw data into easy-to-understand visuals that highlight important patterns.

6.4 Seaborn

Seaborn is built on top of Matplotlib and is specifically designed for creating more visually appealing and informative statistical plots. It simplifies the process of making complex charts like boxplots, scatterplots, and heatmaps with minimal code.

In this project, Seaborn was used to explore how different features such as **Followers**, **Engagement Rate**, and **Influencer Score** are related. It also helped show differences across niches and spot outliers

6.5 Scikit-learn

Scikit-learn is a machine learning and data preprocessing library in Python. Although no machine learning models were built in this project, scikit-learn's `MinMaxScaler` was used to normalize numerical features like Followers and Influencer Scores to a common scale before combining them.

7.Sample Output

7.1 Top Influencers Table

- Table for top5 influencers with expected money

	Influencer_ID	Niche	Followers	Engagement_Rate	Influencer_Score	Promotions_Rate	Expected_Money
1	influencer_41	Education	9000.0	2.101777777777778	1.5371	0.3349	7630.65
2	influencer_485	Education	29071.0	0.9721715799250112	0.7993	0.1724	3051.86
3	influencer_317	Education	87844.0	0.5678133964755703	0.7127	0.1496	4167.46
4	influencer_49	Education	43304.0	0.4646914834657306	0.5948	0.1184	1110.53
5	influencer_304	Education	132016.0	0.3188931644649133	0.5873	0.127	2460.65
6	influencer_339	Fashion	32712.0	1.3436047933480069	1.0862	0.2341	8758.0
7	influencer_228	Fashion	35355.0	1.07325696506859	0.88	0.1906	4987.41
8	influencer_420	Fashion	49238.0	0.416751289654332	0.5861	0.1157	1090.43
9	influencer_21	Fashion	14540.0	0.3991746905089408	0.5475	0.1023	254.74
10	influencer_100	Fashion	201582.0	0.160073816114534	0.492	0.1188	1477.98
11	influencer_142	Fitness	10486.0	4.666984550829677	3.0562	0.6913	81023.25
12	influencer_468	Fitness	10569.0	3.642066420664207	2.5172	0.5586	42414.62
13	influencer_375	Fitness	9809.0	2.8810276276888573	2.1046	0.4582	21355.61
14	influencer_382	Fitness	43765.0	1.0125442705358163	0.8835	0.1895	5813.96
15	influencer_239	Fitness	66629.0	0.354515301145147	0.5647	0.1127	1178.03
16	influencer_75	Health	8267.0	5.442724083706303	3.5376	0.8017	100000.0
17	influencer_209	Health	24870.0	1.8556091676718935	1.4854	0.3172	17039.39
18	influencer_127	Health	84459.0	0.5602363276856226	0.6961	0.1462	3773.56
19	influencer_71	Health	70953.0	0.5544092568319874	0.5846	0.1274	2295.86
20	influencer_123	Health	124180.0	0.3408680947012401	0.5725	0.1246	2366.17
21	influencer_359	Skincare	5404.0	4.285529237601777	2.7993	0.6329	32152.85
22	influencer_196	Skincare	18807.0	1.8183123305152336	1.415	0.304	11527.45
23	influencer_3	Skincare	64101.0	0.7554640333224131	0.7853	0.1655	4932.05
24	influencer_287	Skincare	16023.0	0.9163702178118954	0.7618	0.1614	1414.73
25	influencer_271	Skincare	57528.0	0.6513871506049228	0.7108	0.1479	3087.09
26	influencer_97	Technology	22955.0	2.0754955347418864	1.5813	0.3425	20220.37
27	influencer_149	Technology	354652.0	0.130981920304975	0.4546	0.1425	2358.16
28	influencer_169	Technology	94930.0	0.2431897187401243	0.4459	0.0954	769.57
29	influencer_96	Technology	90999.0	0.2347608215474895	0.4369	0.0929	679.48
30	influencer_198	Technology	162504.0	0.1520885639738098	0.4273	0.1013	838.34

8.Advantages of the Project

- **Identifies Genuine Influencers**
  The project filters out inauthentic influencers by analyzing metrics like engagement and authenticity scores.
  This helps brands and users avoid fake followers or bots.
  Ensures collaboration with credible, trustworthy content creators.

- **Supports Smart Marketing Decisions**
  By scoring and ranking influencers, it guides brands in choosing high-impact partnerships.
  Promotes data-driven decision-making in influencer campaigns.
  Saves money by avoiding low-performing influencers.

- **Highlights Niche-Based Performance**
  The niche-wise analysis shows which categories have the most effective influencers.
  Enables targeted marketing based on niche trends.
  Helps businesses align with relevant content creators.

- **Reveals Engagement vs. Follower Trends**
  Shows that micro-influencers often have better engagement than celebrities.
  Encourages focusing on quality of interaction over quantity of followers.
  Helps improve ROI on influencer marketing.

- **Promotes Authenticity and Transparency**
  Authenticity scores reduce reliance on inflated numbers.
  Encourages influencers to build real relationships with audiences.
  Builds trust between brands, influencers, and followers.

- **User-Friendly Visual Insights**
  The project uses clear visualizations like heatmaps, box plots, and bubble charts.

Makes it easier for non-technical users to understand influencer impact.
Enhances presentation and communication of insights.

- **Scalable for Large Datasets**
  Can be applied to thousands of influencer profiles across regions and niches.
  Suitable for influencer agencies and brand teams.
  Easily extendable with more metrics or machine learning.

9.Conclusion

This project helps companies choose the right Instagram influencers using data. We analyzed follower count, engagement, and other important scores to rank influencers. It also shows how much money brands might spend on promotions. This method makes marketing easier, smarter, and more cost-effective.

we can rank top influencers in each categories and help businesses:

1. Choose the right influencers faster
2. Save time and money
3. Make smarter marketing decisions
4. Get better results from promotions

10.References

- **https://scikit-learn.org**
- **https://www.kaggle.com**
- **https://streamlit.io**
- **https://realpython.com**