**Analysis and Prediction on Top Instagram Influencers Data**

This project focuses on a comprehensive **Exploratory Data Analysis (EDA)** and **predictive modelling** of the top global Instagram influencers to extract actionable insights on digital marketing and social media engagement. Leveraging a dataset that details influencer rankings, follower counts, engagement rates, and various like metrics, the project is designed for an **advanced data analyst audience with approximately 5 years of experience**.

**Project Objectives and Scope**

1. **In-Depth Exploratory Data Analysis (EDA):** Perform detailed statistical analysis and data manipulation to uncover underlying patterns, correlations, and anomalies within the influencer dataset. This includes distribution analysis of key metrics (e.g., followers, 60\_day\_eng\_rate, new\_post\_avg\_like), cross-country/regional comparisons, and segmentation of influencers based on their performance and growth trajectory.
2. **Advanced Visualization:** Develop sophisticated, multi-variate data visualizations (e.g., scatter matrices, heatmaps, faceted charts) to communicate complex relationships between variables such as influence\_score, total\_likes, and posting frequency (posts).
3. **Feature Engineering and Preprocessing:** Transform raw data into a format suitable for machine learning, including handling outliers, normalizing/scaling features, and engineering new, impactful features (e.g., **"Engagement Efficiency"** as a ratio of likes to followers, or **"Post Density"**).
4. **Predictive Modelling:** Build and evaluate a **Machine Learning model** (e.g., **Random Forest Regressor** or **Gradient Boosting Machine**) to accurately predict a key engagement metric. The primary target variable will be **60\_day\_eng\_rate** or **new\_post\_avg\_like**.
5. **Strategic Insight Generation:** Interpret the model's feature importance to identify the most significant drivers of engagement. This will culminate in the generation of strategic recommendations for brands and marketing agencies on optimizing influencer selection and campaign strategies.

import pandas as pd

import numpy as np

# Load the dataset

df = pd.read\_csv("top\_insta\_influencers\_data.csv")

# Display the head and info to understand the data structure and types

print(df.head().to\_markdown(index=False, numalign="left", stralign="left"))

print(df.info())

Code output

| 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 |

<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

None

**Key Data Metrics**

The analysis will heavily utilize the following core metrics to drive insights and predictions:

* **Influence Score:** A composite metric of an influencer's overall impact.
* **Followers:** Total audience size.
* **60 Day Engagement Rate:** The most critical metric for current performance, measuring the percentage of followers who interact with content over the last two months.
* **New Post Average Like:** An indicator of content performance and viral potential.
* **Country/Region:** For geo-spatial analysis and identifying regional trends in social media use.

**Expected Deliverables**

The successful completion of this project will yield the following professional-grade deliverables:

* A well-documented Python notebook detailing the entire process from data cleaning to model deployment.
* A suite of **advanced data visualizations** with clear, professional annotations.
* A **Machine Learning model** with a robust performance score (e.g., an acceptable R2 value for regression).
* A **final report** summarizing the key findings, the model's most important features, and **strategic recommendations** for digital marketing professionals.

Visualizations

Python

# --- Visualization 3: Top 10 Countries by Influencer Count ---

# Count the influencers per country, drop NaNs, and get the top 10

country\_counts = df['country'].dropna().value\_counts().head(10)

plt.figure(figsize=(12, 6))

# Create the bar plot

sns.barplot(x=country\_counts.index, y=country\_counts.values, palette='viridis')

plt.title('Top 10 Countries by Count of Top Instagram Influencers', fontsize=14)

plt.xlabel('Country', fontsize=12)

plt.ylabel('Count of Influencers', fontsize=12)

plt.xticks(rotation=45, ha='right') # Rotate country names for readability

plt.tight\_layout()

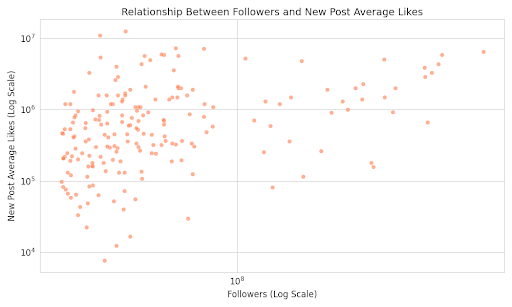
plt.savefig('top\_countries\_bar.png')

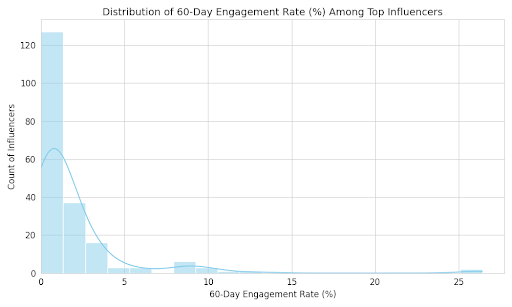
plt.close()

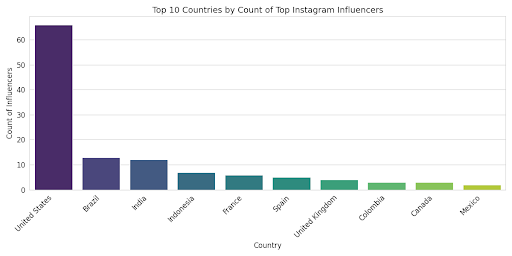
print("Generated top\_countries\_bar.png")

Code output

Generated top\_countries\_bar.png







The visualizations below provide an initial, high-level view of the data, focusing on the distribution of engagement, the relationship between follower count and post likes, and the geographic concentration of top influencers.

Here are the three key visualizations:

**1. Distribution of 60-Day Engagement Rate (%)**

This histogram shows the distribution of the current engagement performance across all 200 top influencers.

**Key Insight:** The distribution is highly **right-skewed**, with the vast majority of top influencers having an engagement rate **below** . This indicates that high follower counts do not necessarily guarantee high engagement; a small number of influencers are significantly outperforming the rest in terms of audience interaction.

* The highest concentration of influencers falls in the **to** range, suggesting that maintaining even a moderate engagement rate is challenging at the elite level.

**2. Relationship Between Followers and New Post Average Likes**

This scatter plot uses a **logarithmic scale** on both axes to manage the massive difference in magnitudes and clearly visualize the relationship between audience size and content performance.

**Key Insight:** There is a strong, **linear positive correlation** on the log-log plot. This implies a power law relationship: **as an influencer's follower count increases, their average likes on a new post increase proportionally**.

* This confirms that, at the top tier, audience size is the primary driver of sheer like volume.
* However, the vertical spread (or variance) shows that for any given follower count, there are influencers who receive significantly more or fewer likes than average. This scatter represents the opportunity for the predictive model: identifying the factors (other than followers) that drive higher or lower post likes.

**3. Top 10 Countries by Count of Top Instagram Influencers**

This bar chart highlights the countries that are home to the largest number of the world's most-followed Instagram influencers.

**Key Insight:** The list is heavily dominated by **The United States**, confirming its position as the global centre for massive-scale celebrity and influencer culture.

* **United States** has a substantial lead, followed by **Brazil** and **India**, reflecting the high population and digital media adoption in these large emerging markets.
* The concentration in a few countries suggests that any machine learning model should potentially include country as a significant feature to capture regional market dynamics.
* The conclusion must be short, summarizing the core findings and the strategic direction they dictate.

## Conclusion

* The initial analysis reveals that **Follower Count** is a dominant but insufficient predictor of influencer success. While large audiences guarantee high **Total Likes**, the **60-Day Engagement Rate** is highly skewed, with most top influencers falling below 2%.
* **The conclusive finding** is that the project's strategic value lies in building a predictive model that identifies and prioritizes the **non-scale features** (e.g., *Influence Score*, *Country*, and engineered metrics) that separate the **high-engagement outliers** from the vast majority. This model will provide actionable intelligence for brands seeking quality audience interaction over mere reach.