Instagram Influencer Data Analysis — Detailed Summary Report

K. Kiran Sai Karthik

Project by K. Kiran Sai Karthik

Contents

Step 1: Import Libraries

Imported essential Python libraries:

- pandas, numpy for data handling
- matplotlib, seaborn for visualizations
- scipy for statistical tests
- sklearn for model building and evaluation
- Suppressed warnings for cleaner outputs.

Step 2: Data Cleaning

- Loaded the influencer dataset (top_insta_influencers_data.csv).
- Converted string values like '1.2M' and '14.5K' into numeric floats.
- Cleaned percentage signs and standardized column formats.
- Filled missing values using:
  - Mode for categorical (country)
  - Dropped remaining rows with NaNs.
- Removed duplicates and verified final dataset shape.

Step 3: Basic Information

- Displayed the first few rows and dataset info.
- Confirmed all columns were correctly typed (float/int/string).
- Used .describe() and .info() for initial understanding.

Step 4: Visualization 1 (Distribution of Influence Scores)

- Plotted a histogram of influence_score.
- Most influencers had scores in the 60–80 range.
- Helped visualize overall influence level distribution.

Step 5: Visualization 2 (Top 10 Countries by Influencer Count)

- Created a horizontal bar plot.
- Highlighted countries like USA, India, and Brazil with the highest influencer counts.

Step 6: Visualization 3 (Correlation Heatmap)

- Cleaned shorthand like '475.8M', '1.2K' before applying .corr().
- Heatmap showed strong correlation between:
  - followers and avg_likes
- Weak correlation between engagement_rate_60d and follower count.

Step 7: Visualization 4 (Engagement Rate vs Followers)

- Created a scatterplot to observe patterns.
- Identified that high follower counts generally correspond to lower engagement rates.

Step 8: Visualization 5 (Boxplot of Engagement Rate by Top Countries)

- Used seaborn.boxplot() for engagement rate grouped by country.
- Some countries showed outliers and high variability.

Step 9: Group Analysis (By Country)

- Grouped by country to calculate:
  - Mean, median of engagement rate
  - Count of influencers
- Created new insights on geographic trends.

Step 10: Visualization 6 (Average Engagement Rate by Country)

- Bar plot showing countries with highest and lowest average engagement rates.
- Countries with fewer influencers showed higher engagement.

Step 11: Outlier Detection (Engagement Rate)

- Used IQR method to detect outliers in engagement_rate_60d.
- Counted number of influencers outside upper/lower bounds.

Step 12: Visualization 7 (Outliers in Engagement Rate)

- Plotted boxplot with outliers highlighted.

- Showed few extremely high engagement influencers skewing the data.

Step 13: T-Test (High vs Low Followers)

- Split data by median followers into two groups.
- Performed T-Test on engagement_rate_60d.
- Result: Statistically significant difference ($p < 0.05$).

Step 14: Z-Test (Sample Mean Engagement Rate vs 100)

- Performed Z-Test to compare mean engagement to benchmark value (e.g., 100).
- Result: Significant difference, confirming deviation from benchmark.

Step 15: Regression Model (Base Features)

- Built Linear Regression model using:
  - followers, avg_likes, engagement_rate_60d as predictors
- Evaluated using:
  - R^2 score
  - Mean Squared Error (MSE)
  - Residual plot

Step 16: Feature Engineering

- Created new features:
  - Log-transformed values: Log_Followers, Log_Posts, Log_Avg_Likes
  - Binned followers into: Small, Medium, Large, Very Large
  - Interaction term: posts × engagement_rate_60d
  - Boolean: Is_High_Avg_Likes
- Applied one-hot encoding to follower bins.

Step 17: Regression with Feature Engineering and Country

- Retrained regression using enhanced features.
- Included country dummy variables.
- Final model achieved better performance (higher R^2).
- Captured more nuanced influencer behavior across regions.

Step 18: Conclusion

This project provided a deep-dive analysis of Instagram influencers by performing:

- Data cleaning
- Statistical tests
- Visualization
- Predictive modeling

The feature-engineered regression model significantly improved performance and offered practical insights into influencer metrics by geography and engagement.