Final_Project

Soma Shekar Vayuvegula

2022-07-24

Introduction:

Social media is important arm of the current internet world. Social media not just provides content and entertainment but also has become a good socurce of income and popularity.

With the datasets available from Kaggle, I will process the data to show which category has is trending per country.

Problem statement addressed:

To find a top successful categories of the social media handle per country.

Approach:

1. Data Collection
2. Merging different datasets
3. Data Cleansing
4. Data plotting using following plots:
    - Scatter plot
    - Histogram

How your approach addresses (fully or partially) the problem:

I will be providing a prediction on which category should a person be starting a social media channel in a country of his liking depending on the data provided.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows):

Below are the data sets used in the analysis (with column details and descriptions):

1. iso-country-codes.xls

- Alpha-2 code: 2 bytes ISO alpha country code
- Alpha-3 code: 3 bytes ISO alpha country code
- English short name lower case: Country name
- Numeric code ISO 3166-2: ISO numeric country code

2. top_200_instagrammers_categories.csv

- Username: Name of the influencer's account
- Main Category: Main topic of the page
- Main Video Category: Category of the reels and video

3. top_200_instagrammers_details.csv

- Username: Name of the influencer's account
- Channel name: Name of the Channel
- Country: Influencer's country
- Url: Instagram Url

4. top_200_instagrammers_likes_followers_Jul2022.csv

- Username: Name of the influencer's account
- Likes: Total Likes count
- Likes Avg. : Average likes
- Posts: Total Posts
- Followers: Total number of the followers
- Boost Index: Boost index value
- Comments Avg.: Average comments number.
- Views Avg: Average Views.
- Avg. 1 Day: Average views perday
- Avg. 3 Day: Average views for 3 days
- Avg. 7 Day: Average views for 7 days
- Avg. 14 Day: Average views for 14 days
- Avg. 30 Day: Average views for 30 days
- Engagement Rate: Percentage of Engagement with users.
- Engagement Rate (60 Days): Percentage of Engagement with users for 60days

Required Packages:

We require the below packages but not limited to:

1. ggplot2
2. readr
3. tidyr
4. dplyr

Plots and Table Needs:

We need below plots:

```
- Scatter plot: To show the followers of each country handles
- Histogram: To show the category wise followers and likes and sub-category wise followers and likes
```

We need the below tables:

```
- Output by category
- Output by sub category
- Output by country
- Output by country and show top 2 categories
- Output by country and show top 4 sub-categories
```

Questions for future steps:

- How to show the data in the required output tables?
- Are the listed plots are enough or should we use any other plot?