Project Proposal
Jennie

A research question. This can be presented as a hypothesis or a question statement.

Which kind of video is the most popular program in each country? Do they have something in common?

• An explanation for why that question is interesting, and what the implications would be if you could find an answer. This can take the form of a problem statement or literature review but must situate your question in a broader context.

Youtube has a strong AI algorithm that recommends great resources for each customer. Me, as a customer watching Youtube almost every day, I feel I absorb much useful knowledge that I want to know. If Youtube knows my interests so well, it must know other audiences.

I prefer to use R to do an analyze first. Find what kind of video is the most popular in each country. Using R to visualize the popular rate can be obvious. I can also do the regression to predict audiences' interest in each country. The result can be made as a comparison and may find the reason behind this result.

• The dataset you will use to address your question and the specific variables that you will analyze. This can be a dataset you find or create. It can be public or personal.

https://www.kaggle.com/datasnaek/youtube-new This data is from Youtube API.

• A description of the data – what it is about, where did it come from, what variables are available, how were they collected, and what are the limitations

My data includes many variables:

video_id / trending_date / title / channel_title / category_id / publish_time / tags / views / likes / dislikes / comment_count / thumbnail_link / comments_disabled / ratings_disabled / video_error_or_removed / description

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.

EDIT: Now includes data from RU, MX, KR, JP and IN regions (Russia, Mexico, South Korea, Japan, and India respectively) over the same time period.

Each region’s data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

Limite: The dataset doesn’t include all countries. In order to find answers to my question, I will choose countries considerably. The population size is the first step I need to consider because the data compared with a similar population will be more reliable. For instance, 1000 people like music videos in India have a different meaning with 1000 people like music videos in Japan. Therefore, according to the population all over the world, I choose Russia (1.89%_percentage of world population), Mexico (1.63%), Japan (1.63%) as a group to compare their dataset. Make UK (0.857%), Germany (1.07%) and France (0.865%) as a group to analyze.

- Your data analysis plan: will your analysis be descriptive or predictive? What methods will you use? What tests.

I would like to make my analyze process descriptive and the summary can be predictive.

The methods I use:
Data Visualization: See the majority interests / check the relationship between the columns
Confidence Intervals: Confidence intervals are essential for understanding the relationship between samples and populations.
Text Analysis: check labs
Regression: Predict the future interests of each country.

Reference：

https://zh.wikipedia.org/wiki/%E5%90%84%E5%9B%BD%E4%BA%BA%E5%8F%A3%E5%88%97%E8%A1%A8

https://www.kaggle.com/datasnaek/youtube-new