

Springboard Capstone Project 1 – Project Proposal

What is the problem you want to solve?

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

Project Goals:

This capstone project is based on dataset for YouTube's [top trending videos](#). The problem many potential YouTube advertisers face today is how to define what makes a YouTube content 'great' and has a potential to become one of the top trending contents on YT. Secondly, advertisers also are looking for new and efficient ways to define YouTube's audience/users in terms of their profile, viewing habits and preferences. User profiling and segmentation is the key input in targeted advertising.

This project aims to solve following problems:

1. Identify and characterize a 'great' content which is able to make it to the 'Top Ten Trending' list of YouTube
2. Identify and define potential user segments based on the viewing and 'likes' data provided in the dataset.

What data are you using? How will you acquire the data?

A little more information about the dataset:

- YouTube (the world-famous video sharing website) maintains a list of the [top trending videos](#) on the platform. [According to Variety magazine](#), "To determine the year's top-trending videos, YouTube uses a combination of factors including measuring users' interactions (number of views, shares, comments and likes). Note that they're not the most-viewed videos overall for the calendar year". Top performers on the YouTube trending list are music videos (such as the famously virile "Gangam Style"), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for.
- This dataset is a daily record of the top trending YouTube videos.

Content

- This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, FR, RU, MX, KR, JP and IN regions (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India respectively), with up to 200 listed trending videos per day.
- 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.

Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

Project Plan / Approach to achieve the goals outlines above:

- a. As in any good data analysis project, I first plan to do a good deal of EDA (Exploratory Data Analysis) of the dataset. The objective is to thoroughly understand the data and then based on that prepare the data analysis plan/framework.
- b. Based on the EDA, I will identify the key hypothesis which should answer following questions
 - What makes a video worthy of 'top trending videos list'?
 - How can we segment these videos (based on the data attributes given in the data)
- c. Use regression methods to determine which attributes have higher impact/influence
- d. Plan how to use the visualization techniques learned to present the findings
- e. Use Supervised learning methods to identify patterns, segments and attributes which drive the results
- f. Use Unsupervised learning methods to identify patterns, segments and attributes which drive the results

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

Key Deliverables, Collaterals

I envision following documents to document and present the result of this effort

- iPython Notebooks containing all the codes, analysis and results
- A summary/executive finding report
- Presentation slides