

Analysis of Youtube Trending Statistics

- The aim of this project is to perform exploratory data analysis on the Youtube trending statistics dataset and deduce insights and conclusions about the relationships between various attributes.
- The Youtube trending statistics dataset is a publicly available dataset on Kaggle which gives details of the stats for various nations about atleast 200 trending videos per day. We have chosen to work with the US version of the dataset.
- The dataset consists of 16 attributes which describe different aspects of each trending video. The attributes are: video_id, trending_date (The date the video was collected on), title, channel_title, category_id, publish_time (The time at which the video was published on YouTube), tags (Separated by | character, [none] is displayed if there are no tags), views, likes, dislikes, comment_count, thumbnail_link, comments_disabled (if comments are disabled for the video), ratings_disabled (if ratings are disabled for the video), video_removed (if the video has been removed), description.
- This dataset involves the cleaning of missing values of string type attributes more than the integer types, which makes data cleaning an important aspect of the analysis of this dataset. The dataset contains 40950 rows entirely, which makes it a viable dataset to be used for other practical applications.
- The inferences drawn from this dataset can be used for Machine Learning based applications, such as sentiment analysis and opinion mining, recommender systems and classification of videos. Since there are a relatively large number of variables in the dataset, it is also very beneficial to draw relationships between the various variables which can be used for the above mentioned applications.