Trending Videos

Data Wrangling

Data wrangling is one of the most essential parts of a data science project. Data wrangling is the process of mapping data from raw to a format appropriate for the data to be analyzed. Data wrangling consists of selecting, indexing, and mutating the tables to form them for your need. This process helps with better decision making, but it takes al long time to complete. If it is done correctly, then a data scientist life is made easier. These steps will help me have more accurate results.

The data I used contained 14 variables and 40,881 observations. They are multiple techniques used to clean and conform the data before the analysis. The first packages loaded was dplyr and tidyr. These packages allow syntax that is simple to use when cleaning the data. By removing insignificant variables from the data frame, the new data set contains the variables trending date, title likes, view, dislikes, and comment count. Theses variable provides information pertinent information regarding the trending videos. This method reduced the number of variables from 14 to 6.

Since the data set is large, I eliminated the observations that included NA.  To this NA was changed to 0 and then omitted them. The omit function was also used to remove the variable with missing data values. The new data frame now has 39924 observations. Another observation from the data set was that row with the lowest number of views was 1023; this didn’t seem logical for a trending video. Thus I created a filter to include only observations at a certain threshold. A popularity index was created to distinguish between highly trending and low trending videos based on the number of views. The clean data set includes 29943 observations and six variables. It is labeled df\_1clean.csv.