# Capstone Project 1: Data Wrangling

*Youtube View Count Prediction*

Due to the requirement of utilizing the Youtube API in order to download the data for this project, several data wrangling procedures had to be utilized in order to access and clean up the data. The steps utilized in cleaning the data are outlined below:

1. Utilized Google / Youtube API online instructions to authenticate an email with Oauth2. These credentials are necessary in order to access the API with a daily quota limit.
2. Through the API, supplied 20 different random words based on categories provided by the Youtube general categories. Each search produces 50 results, therefore the dataset had 1000 rows of data that I saved as a CSV.
3. Once the data was downloaded from the API, utilized json\_normalization to flatten the data and pass this into the pandas dataframe.
4. Analyzed all of the columns, shape, head of the data to assess the current data and then removed all of the columns that were meaningless data that had no relationship to view Count and could not be used for the prediction.
5. Analysis of the view counts was then performed to assess the outliers. The rationale for this is the concern that there will be a few videos from the search that represent overly huge view counts that is not representative of an average video on youtube. From the analysis, 10 of 11 of the identified outliers were music videos. These videos would skew my model – I made the decision to remove the outliers since predicting view count is not life critical, I chose to discard these values. This should improve the overall model.
6. Missing values of the continuous data (e.g. likeCount, commentCount, dislikeCount) were filled in using the mean() function. I felt this would be the most appropriate method of addressing the continuous data since I did not want missing values to skew my model. Also, certain columns had missing lists or strings,, therefore I filled in an empty list or string for those columns using the fillna() function.
7. Instead of analyzing the NLP portion of snippet tags, I decided to just calculate the length of tags and store that in a column. NLP analysis could be considered in the future.
8. Conversion of the video publish date column to date time objects was implemented.
9. Conversion of a string to calculate the total duration of a video was required with the use of regular expression and a list comprehension plus zip function to multiply / calculate the total time in seconds.
10. Since predicting a precise view Count is rather unlikely, the dependent variable was bucketized into sizes between 0-50K, 50K-100K, 100K-150K, 150K-200K, ……500K, and then >500K. The bucket sizes were chosed after analyzing the frequency distribution, most view counts will fall under 500K and this should produce an interesting prediction that is not too open to interpretation but still can offer some accuracy.
11. Upon exploring the data, no significant outliers were detected. The removal view Counts for videos that were extremely high was considered but this would not be an appropriate action given the theory there should be a proportionally high value in other statistical values such as likeCount, etc.
12. Once the data was in a clean / useful format, continued to perform further EDA on the relationships.