**Data Source:**

The data set used for our project consists of data extracted from the YouTube Data Analytics Application Programming Interface (API) that allows users and developers to access YouTube video statistics and various channel data available on YouTube through two different types of data calls. These API resources have been made available by Google for analysis of the YouTube data and implementation of the YouTube experience over various platforms such as websites, applications or devices. The YouTube Data Analytics API was accessed and extracted using Python’s Beautiful soup methodology. The monthly and yearly earning for each YouTube channel were extracted from the Influencer marketing hub - web application that provides an estimate of the earnings based on the channel URL. The application evaluates these values on the basis of view count, subscriber count, likes, dislikes, number of videos uploaded and the average YouTube revenue model.

The various data fields extracted from the API and the web application includes Channel Title, Region, Subscriber Count, Number of Views over the videos, Number of videos uploaded, Genre, Monthly and Yearly Channel Earnings, Rating, Number of likes, Number of dislikes, channel description, Channel ID and Number of Comments for each YouTube channel.

YouTube: Analytics and Reporting APIs (2018). *Retrieve viewing statistics, popularity metrics, and more for YouTube videos and channels.* Retrieved from <https://developers.google.com/youtube/analytics> . Accessed on November 1, 2018.

Influencer Marketing Hub, 2016. *YouTube Money Calculator*. Retrieved from <https://influencermarketinghub.com/youtube-money-calculator/> . Accessed on November 5,2018.

**Licensing and Authorization:**

To access the YouTube Analytics Data API, developers need to complete the API authorization process as per the terms and services of usage. This includes generating the OAuth2.0 authorization credential and API keys for our application to access the API. The steps followed to create the authorization credentials included:

* Signing up on the Google Cloud platform to raise access for OAuth2.0 credential keys and creating a project on Google’s developer console so that the application can submit API requests.
* The application (Python script) must send this OAuth2.0 token using authorized redirect URIs with the request to access the private data.
* The application request is redirected to Google’s authorization server that specifies the scope of access such as the resources that the application can retrieve, update, insert or delete.

This authorization process allowed us to access the API and store the data retrieved in the application dictionary.

Terms of Usage:

Google APIs Terms of Service (2014). *Section 2: Using Our APIs*. Retrieved from <https://developers.google.com/terms/#section_2_using_our_apis> . Accessed on November 5, 2018.

**Metadata:**

The following table gives a brief description of the various resources used in our data set.

|  |  |
| --- | --- |
| Data Field | Description |
| Channel Description | Contains information about a single YouTube channel |
| Channel Title | Specifies the YouTube channel name. |
| Genre | Contains information about a set of videos that a channel has chosen to feature. It also identifies a category that YouTube associates with channels based on their content. |
| Number of likes | Specifies the total number of users that have liked the videos uploaded in the channel |
| View count | Specifies the total number of views over all the videos uploaded by channel. |
| Region | Identifies a geographic area that a YouTube user can select as the preferred content region. The content region can also be referred to as a content locale. |
| Estimated Monthly earnings | Represents an estimate of channel’s monthly earnings. |
| Estimated Yearly earnings | Represents an estimate of channel’s yearly earnings. |
| Number of dislikes | Specifies the total number of users that have disliked the video content uploaded in the channel |
| Number of Comments | Specifies the total number of comments made by users over all the videos uploaded in the channel |
| Subscriber count | Contains information about a YouTube user subscription. This defines the number of users subscribed to the particular YouTube channel who receive daily notifications on any new video uploaded by the channel. |
| Videos uploaded | Represents the total number of videos uploaded by the YouTube Channel. |
| Rank | Represents the channels performance over other channels in YouTube. |

**Data Limitations:**

The data set consisted of various anomalies, missing values and unstandardized data format that needed to be cleaned and rectified. Some of the limitations identified in our data set included:

* Missing values in number of likes, dislikes and View Count data fields.
* Inconsistencies in channel earnings with respect to subscriber and view count which can be considered as anomalies in our dataset.
* Earnings were represented in string format and included range of values specified by multiple special characters such as dollar sign, hyphen and denomination string characters such ‘K’ for thousand and ‘M’ for Millions..

**Data Cleaning Steps:**

The following were the steps taken to identify and clean the inconsistencies in the data format.

* Some of the videos didn’t have records for likes, dislikes and comment section and this needed to be treated as exception by making the count as zero since there are many YouTube channels that prefer disabling this section for their videos.
* The salary data was in the form of a string with multiple non-numeric and special character values and were represented in a range of values for each YouTube channel. These inconsistencies in data format were removed and the range of values were replaced with an average earning value which can be used for our analysis of the channel’s performance over a specific period of time.
* Ranks were represented in ordinal values and string format and were converted into cardinal values so that the numeric values can be compared to identify the top performing and popular channels.
* There were various outlier values based on the subscriber to earnings field where some channels had low subscriber count but unusually high earning value. These outliers were identified and replaced based on regression model and Z test in R.

**Data Cleaning Process**

The data cleaning process has been implemented using R script with a brief description on the various steps performed. Please find the R script attached in the below mentioned Git Repository.

Git URL: <https://github.com/krishnanambi/INFM600-FinalProject.git>

Word Count:985