

**School of Computer Science and Engineering**

# J Component report

**Programme : B.Tech (CSE)**

**Course Title : Data Visualization**

**Course Code : CSE3020 Slot : D2**

**Title: Streaming Content Dashboard**

**Team Members: SAURISH SHARMA (20BCE1370)**

**RITIK SINGH (20BCE1384)**

**AKASH RAJ BEHERA (20BCE1829)**

# Faculty: Dr. PARVATHI R. Sign:

**Date:**

## Streaming Content Dashboard

**20BCE1370, SAURISH SHARMA, Vellore Institute of Technology, Chennai, India 20BCE1384, RITIK SINGH, Vellore Institute of Technology, Chennai, India 20BCE1829, AKASH RAJ BEHERA, Vellore Institute of Technology, Chennai, India**

## ABSTRACT

As we all know in today’s world data analysis and visualization is becoming important thing because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments. Now a days people do not want to waste any time on viewing bad shows and they first look at the ratings and later they decide what to see. According to this situation we designed our project to make streaming content dashboard which will enable us to visualize all the famous shows in every aspect we can understand in a clear way. We also clustered the combined data from Netflix, Hulu, Disney Plus and Amazon Prime using K-Means and created a recommendation system to find similar movies to what the viewer has watched.

## KEYWORD

Netflix, Hulu, Disney Plus, Amazon Prime, recommendation system, text clustering, data visualization and analytics, OTT Content and k-means algorithms

## INTRODUCTION

Recommender Systems (RSs) are characterized by the capability of filtering large information spaces and selecting the items that are likely to be more interesting and attractive to a user.

OTT Platforms are the biggest users of recommendation systems. So, in this Project we aim to visualize the content library of top OTT Platforms like Netflix, Disney Plus, Hulu and Amazon Prime. While doing this we will also discover correlations and recurring patterns in the dataset with interesting inferences.

Finally, we will see how the recommendation engine works to deliver similar content as quickly as possible.

## About The Dataset

For this project we will use 4 datasets containing of listings of all the movies and tv shows

available on Netflix, Hulu, Disney Plus and Amazon Prime, along with details such as - cast, directors, ratings, release year, duration, etc. In total there are approximately 22k observations. It is obtained from Kaggle Open-Source Dataset Library ([Source](https://www.kaggle.com/datasets/shivamb/netflix-shows)).

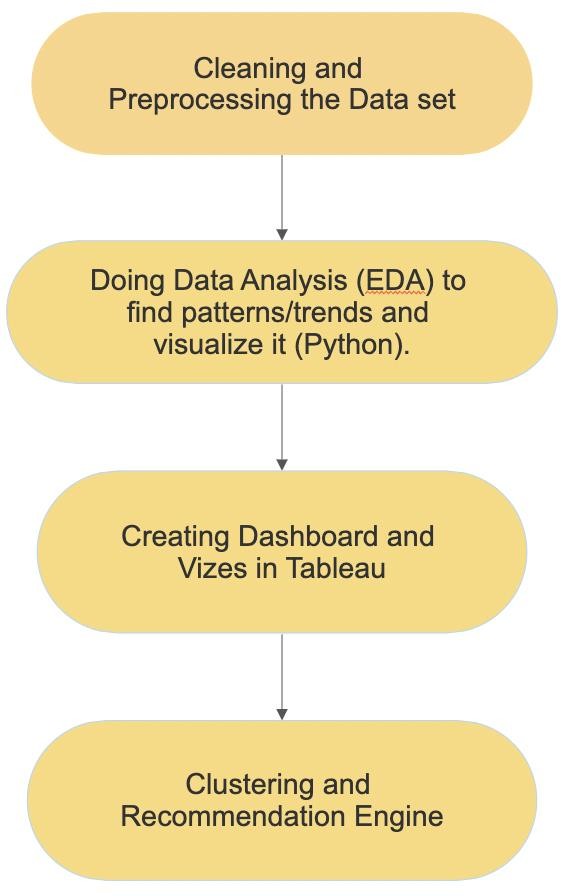
## Feature components for analysis & visualization

For this visualization and analysis, we use feature attributes from the dataset, namely,

* + - Type
    - Title
    - Director
    - Cast
    - Country
    - Date Added
    - Release Year
    - Rating
    - Duration
    - Listed In
    - Description

Each individual dataset contains all the following attributes. During the project we will combine all 4 datasets into one and then we will append a column denoting the OTT platform.

## Design and flow of models



**Fig.1 design and flow of model**

For the Visualization we have used the following modules and analysis parameters:

## Module 1: data cleaning and dataset analysis

After Importing the data set, we need to clean it and analyze what data we were able to collect. After this we can easily plan which parameters to visualize.

## Module 2: Doing Data Analysis (EDA) to find patterns/trends and visualize it (Python).

The attributes from the obtained data set are compared with each other to find correlations and dependencies and then these are visualized using different types of graphs. We can use these graphs to visualize common trends in the dataset.

## Module 3: Creating Dashboard and Vizes in Tableau

We then use Tableau to further Visualize the Dataset and create interactive Dashboards. We found Tableau to be an incredibly versatile and powerful tool for this purpose.

## Module 4: Clustering and Recommendation Engine

We will use K-Means clustering to cluster similar data. We then append the cluster id generated to the combined dataset to facilitate the recommendation engine

**K-means**

K-means algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

The way kmeans algorithm works is as follows:

* + 1. Specify number of clusters *K*.
    2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
    3. Keep iterating until there is no change to the centroids.
    4. Compute the sum of the squared distance between data points and all centroids.
    5. Assign each data point to the closest cluster (centroid).

Compute the centroids for the clusters by taking the average of the all- data points that belong to each cluster.

The objective function is:



**Recommendation Engine**

The recommendation Engine takes a Movie or Show Title as an input. It then finds the cluster id of that entry. It uses the cluster id to reduce the search space.

Now it runs a text similarity check between the description of entered show or movie to find similar content from that cluster.

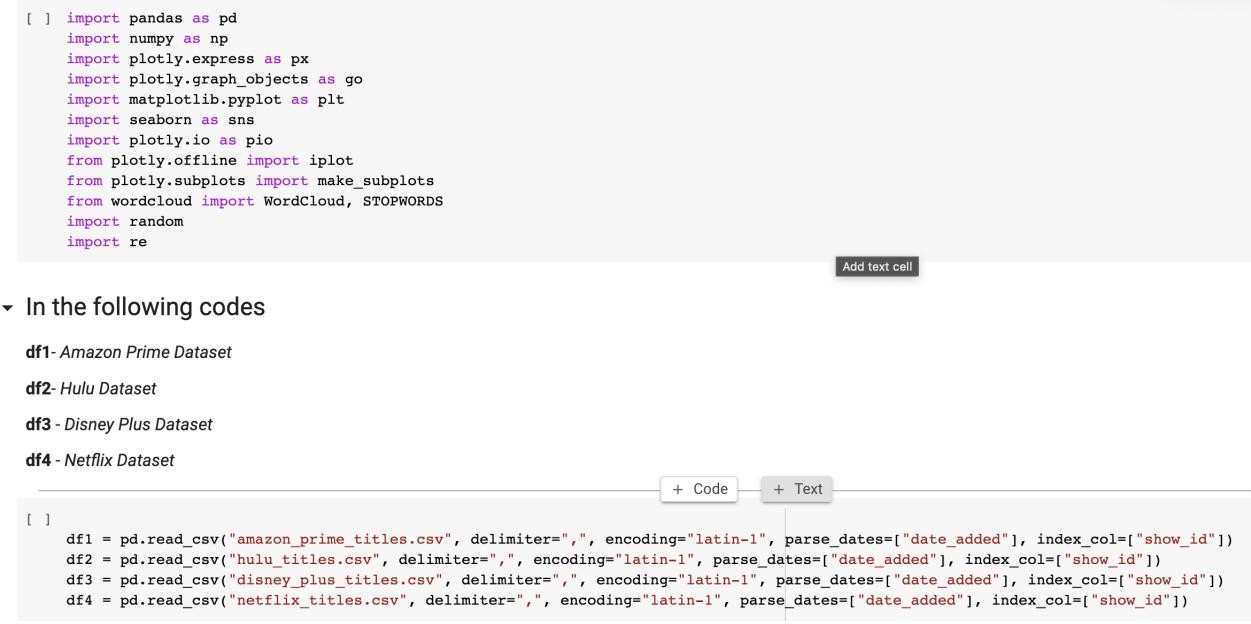
Thus, using K-Means and Text Similarity, it achieves fast and accurate results.

## IMPLEMENTATION

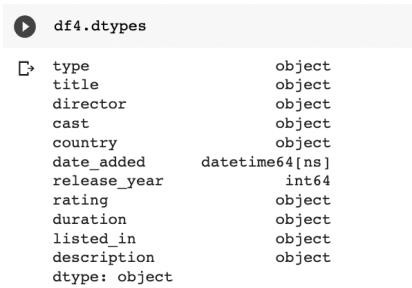
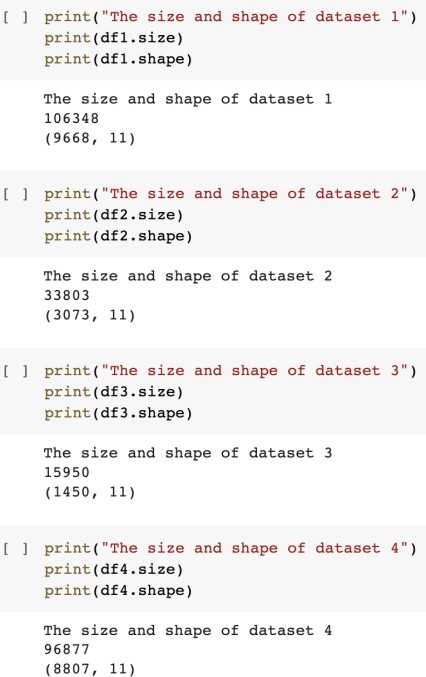
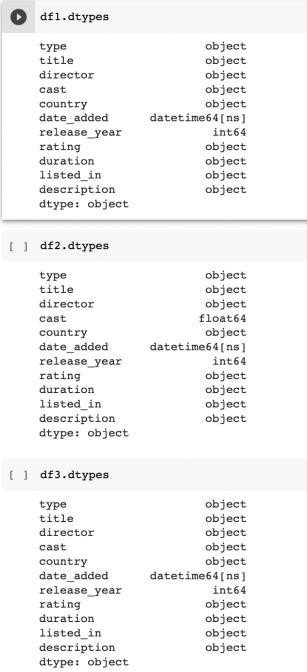
* 1. **First we import modules and datasets.**

## Importing all the libraries and modules

First import the libraries to better analyze the data set. Here matplotlib and plotly are used for visualization and word cloud.

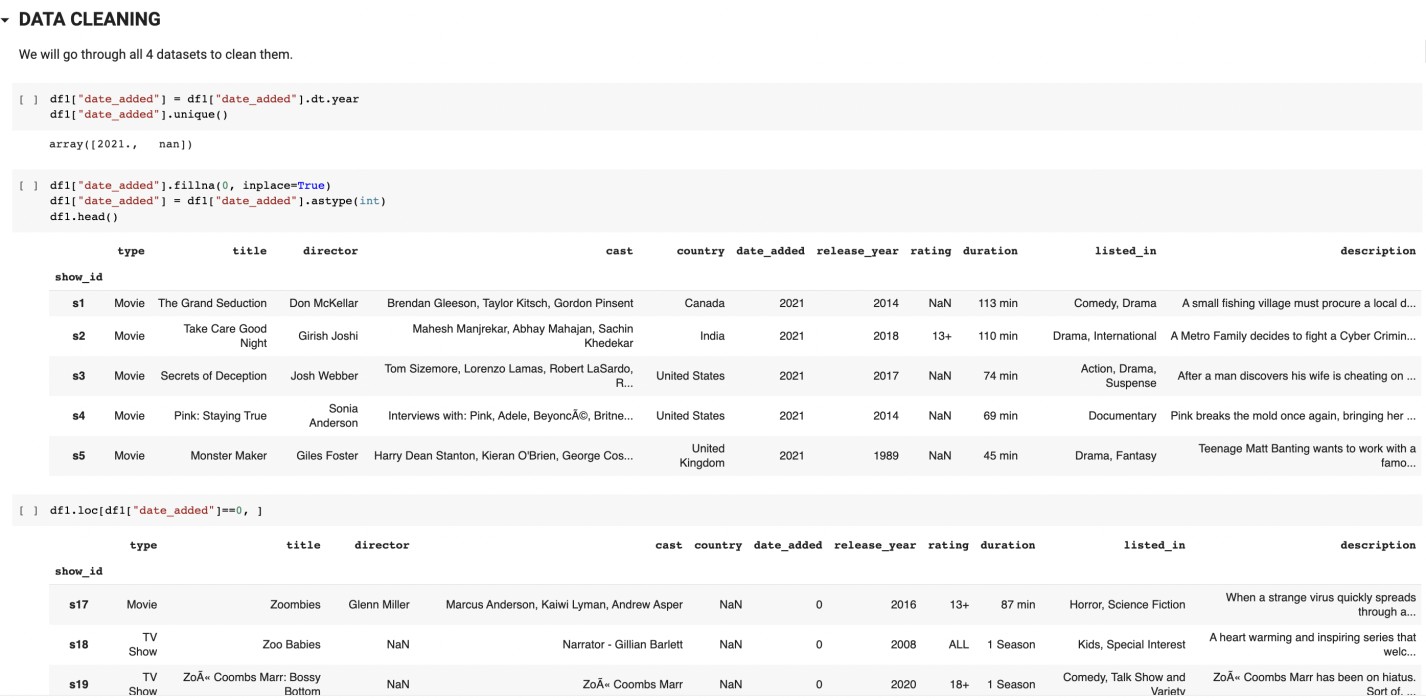


* 1. **Dataset Analysis**

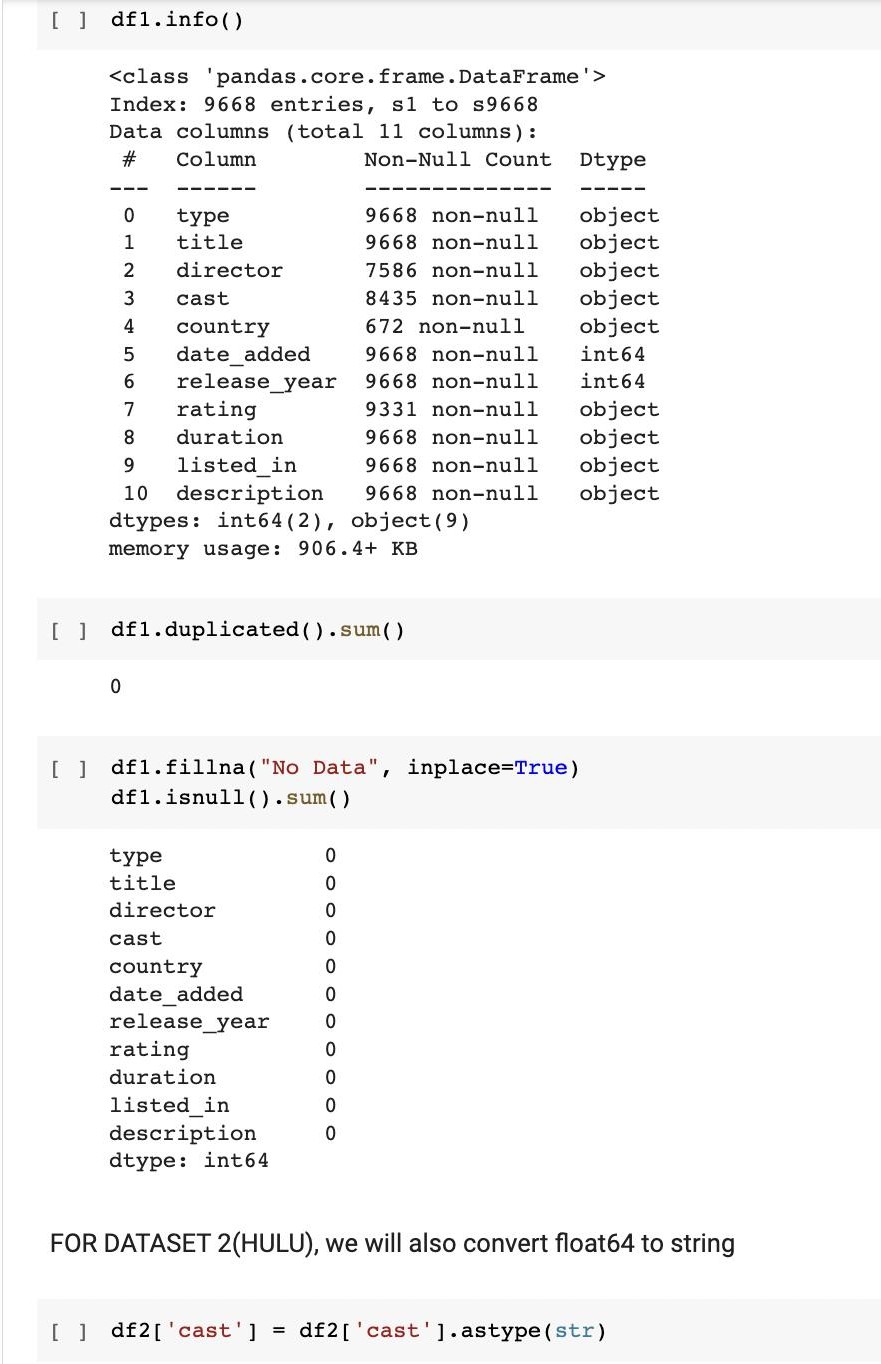


Here we can see the size and attributes of the dataset. All data is in correct form except cast in df2 (Hulu) which is in float64 format. We will resolve that in pre processing

* 1. **Data Cleaning and Preprocessing**

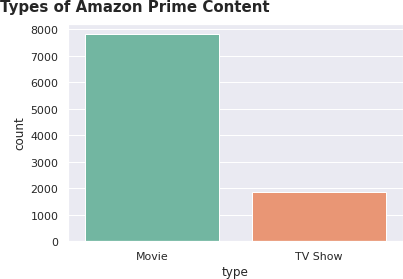


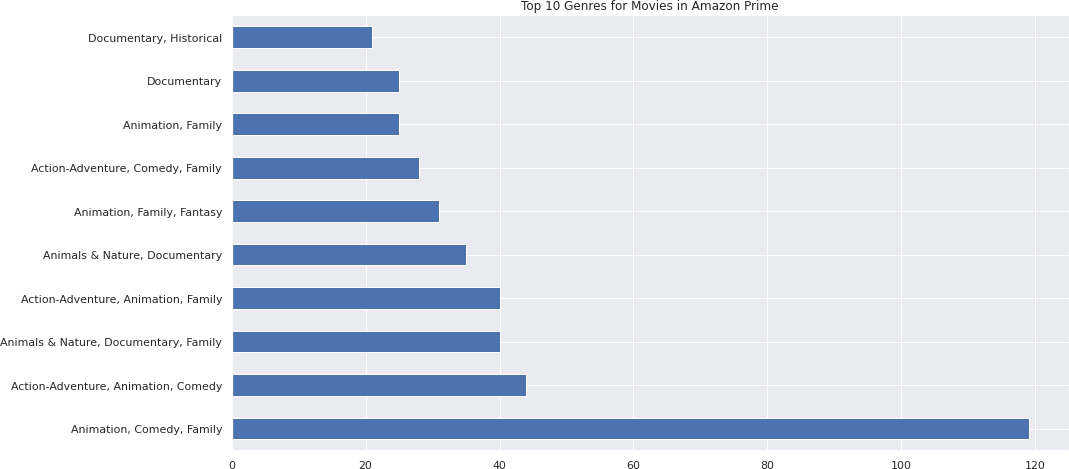
Here we transform the date\_added and date\_released fields to extract years from it. We also check for and remove Null values. We do the same for all dataset.

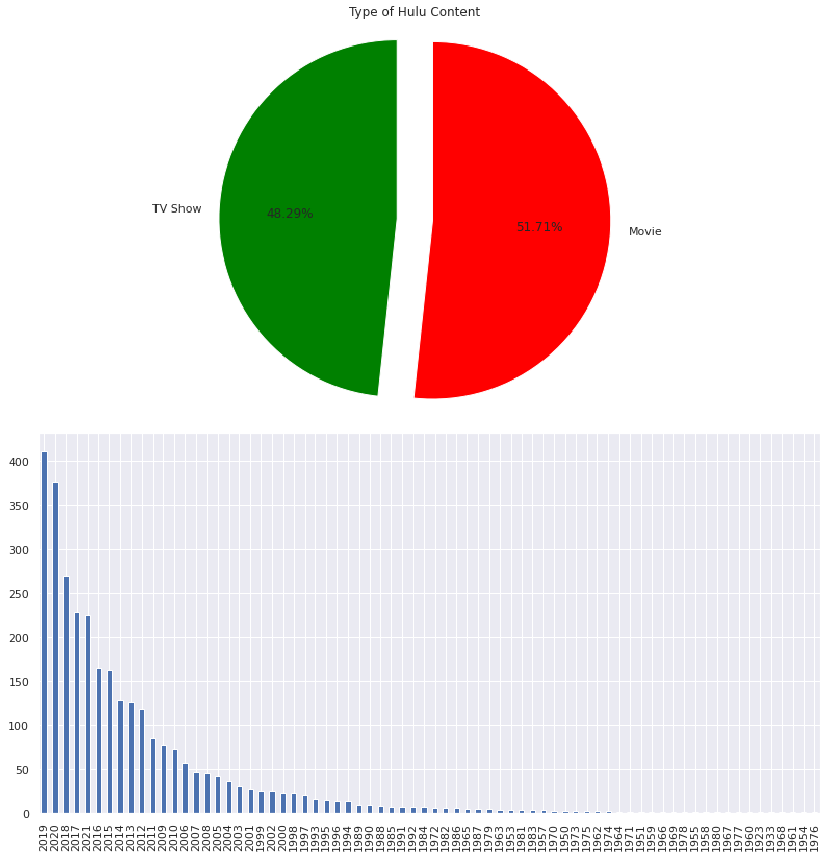


Here, we check the datasets for null and duplicated values as well as missing data. We also convert the float64 column from Hulu dataset to String format.

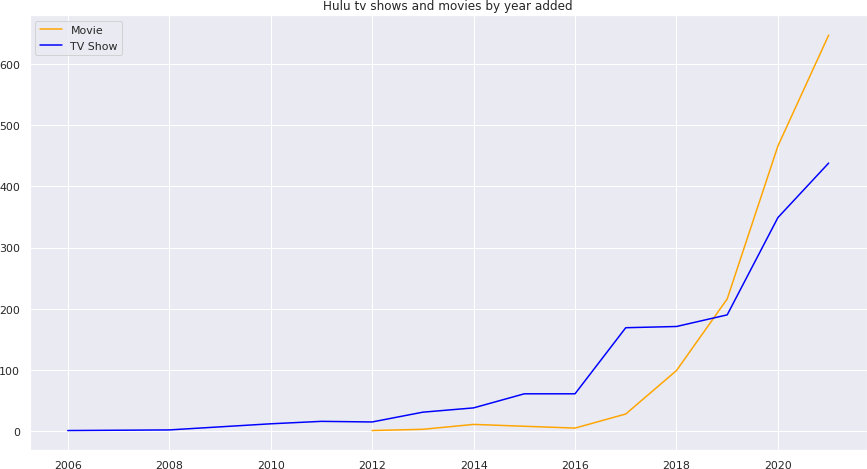
* 1. **Visualization of Datasets**

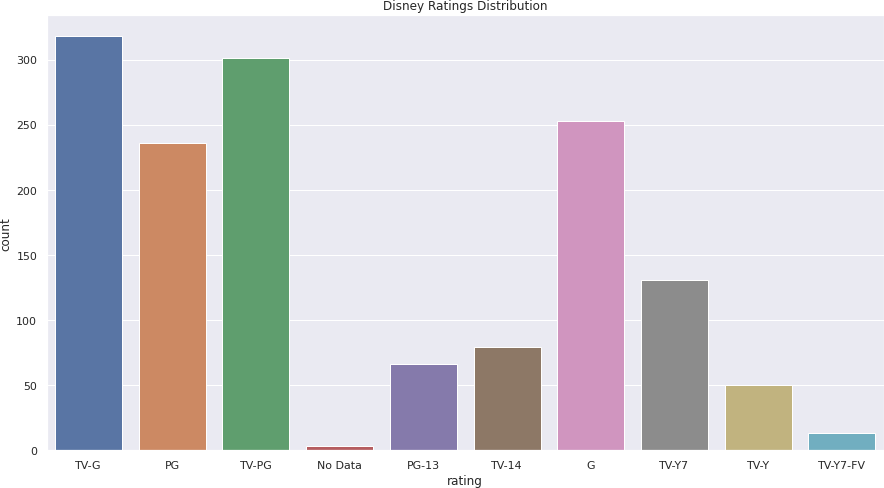
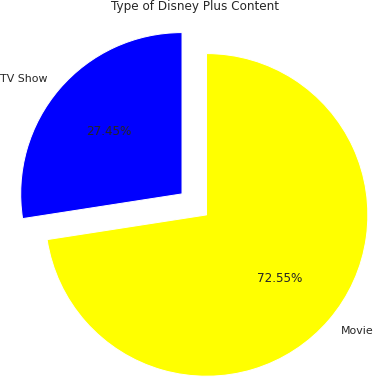


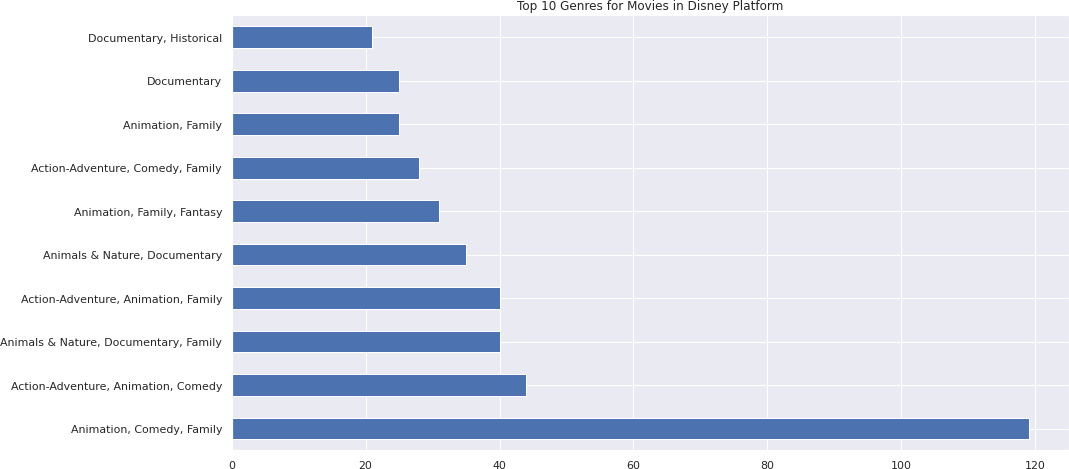


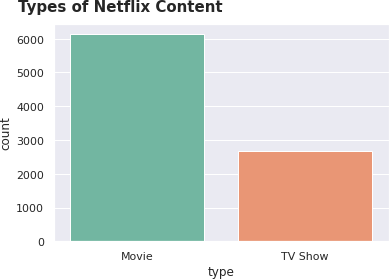
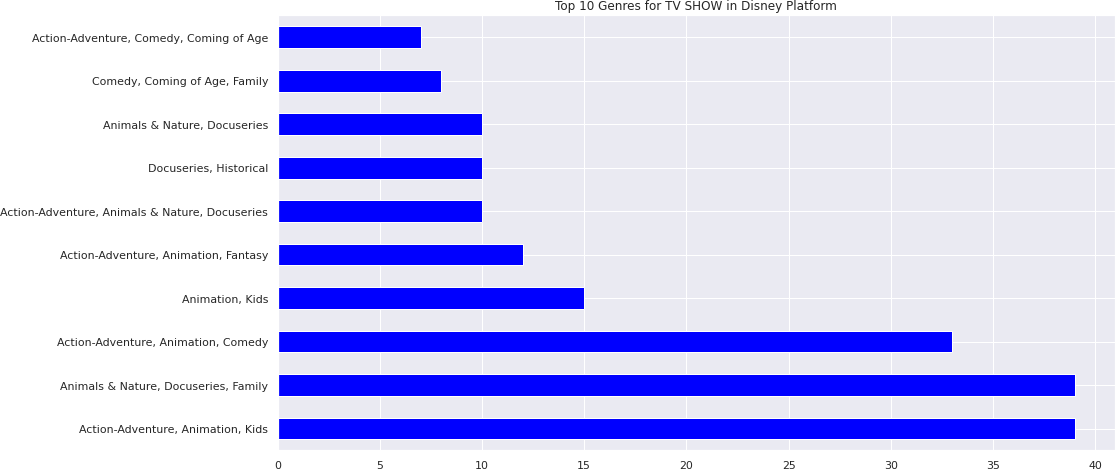


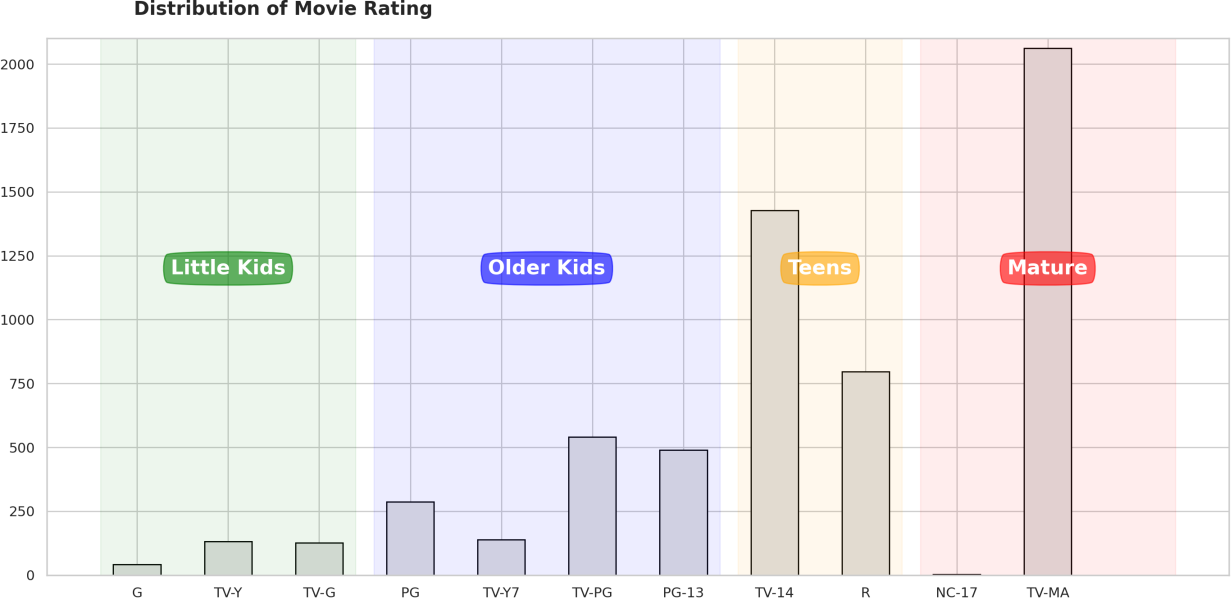
Content by their Release year on Hulu

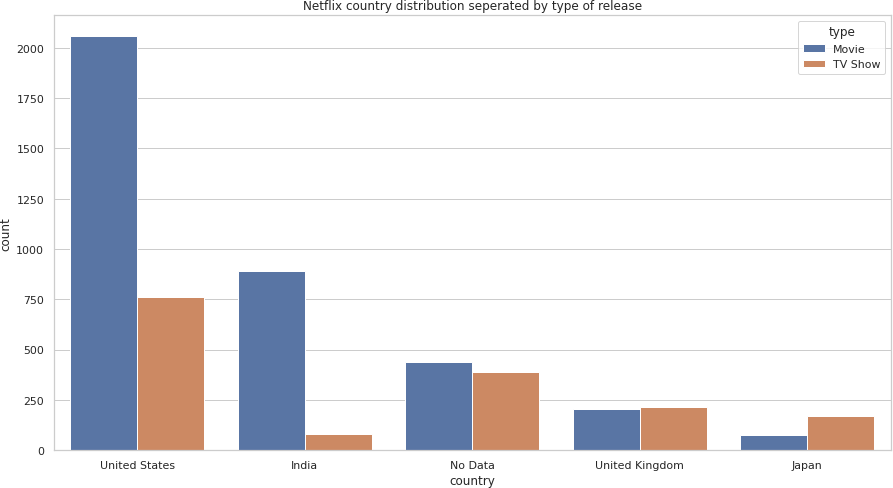


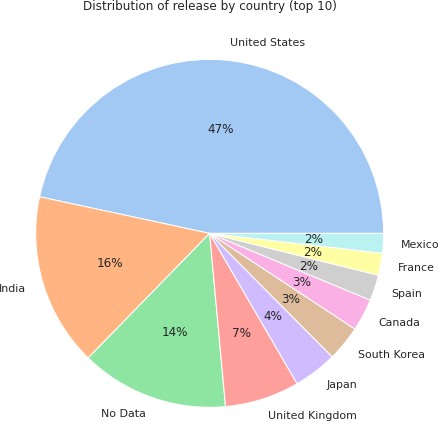




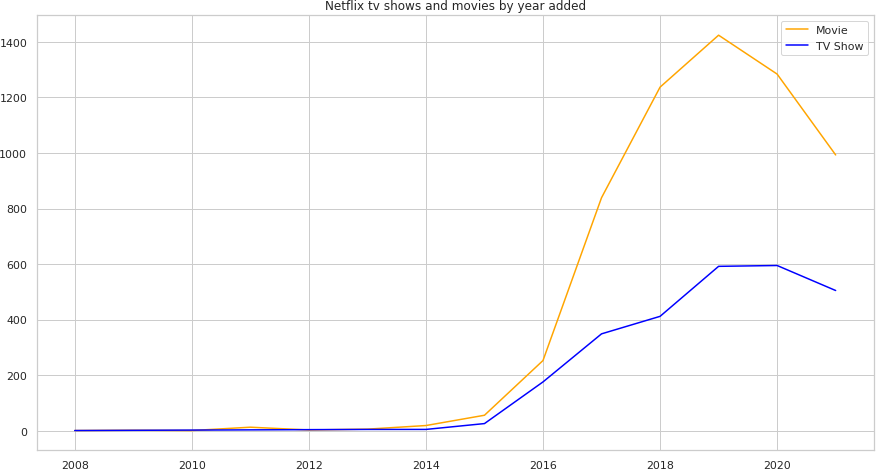




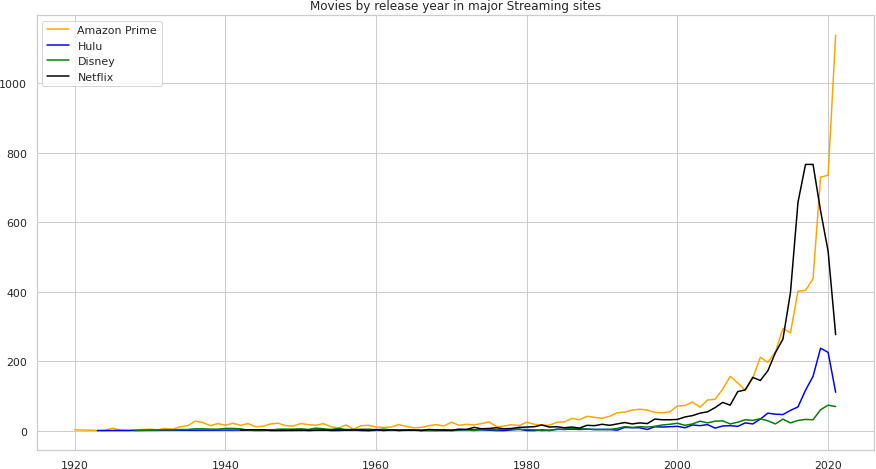


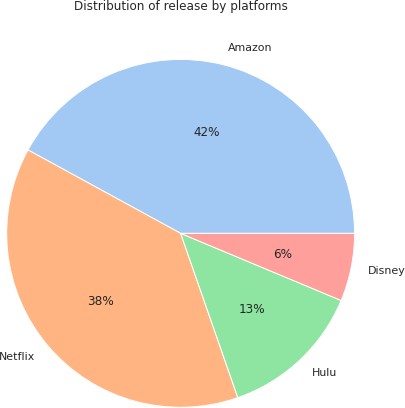


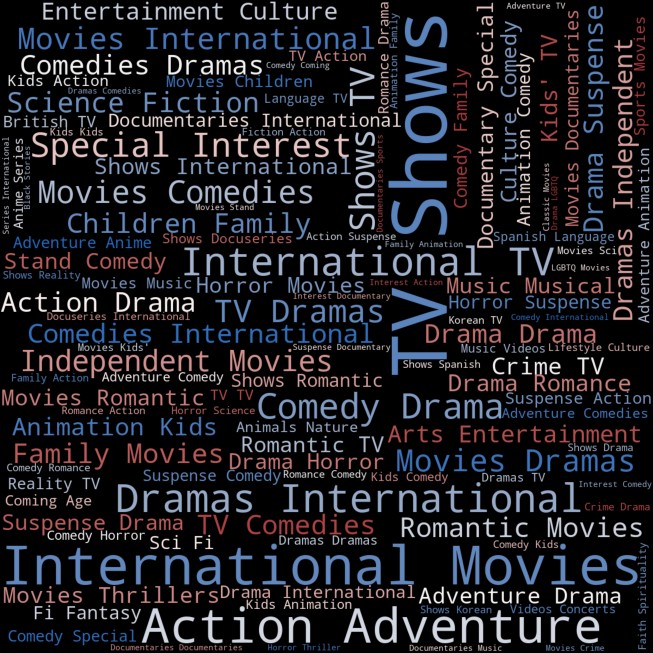
Top 10 Country by content type in country

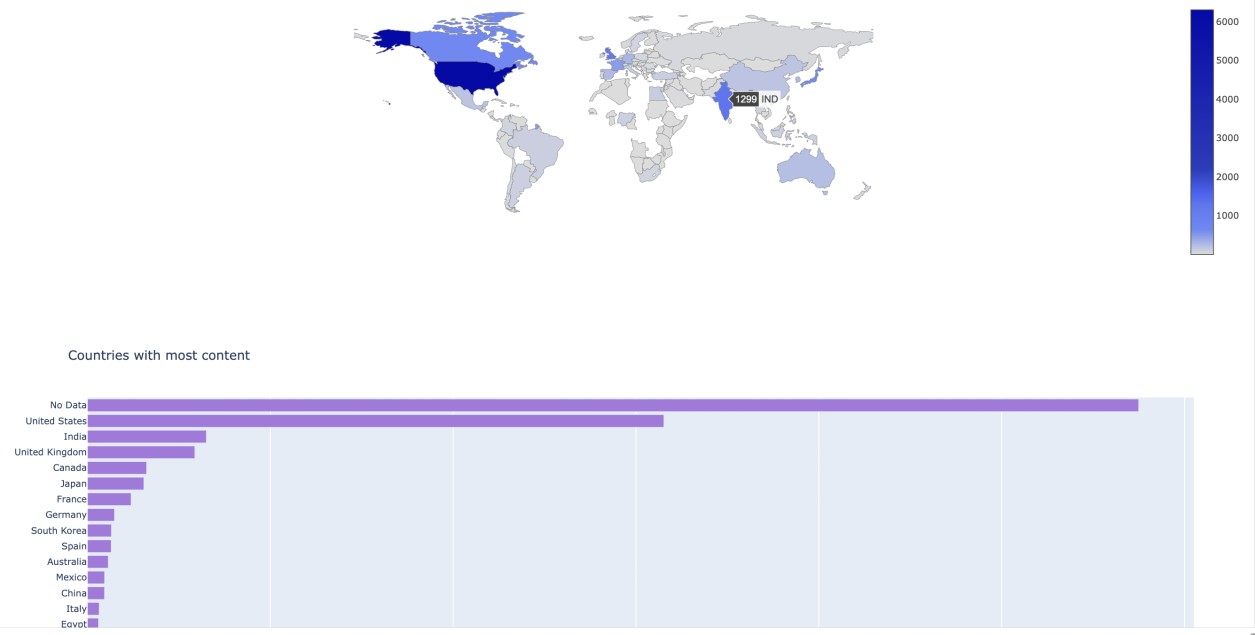


# Creating One unified Dataset and Visualizing it



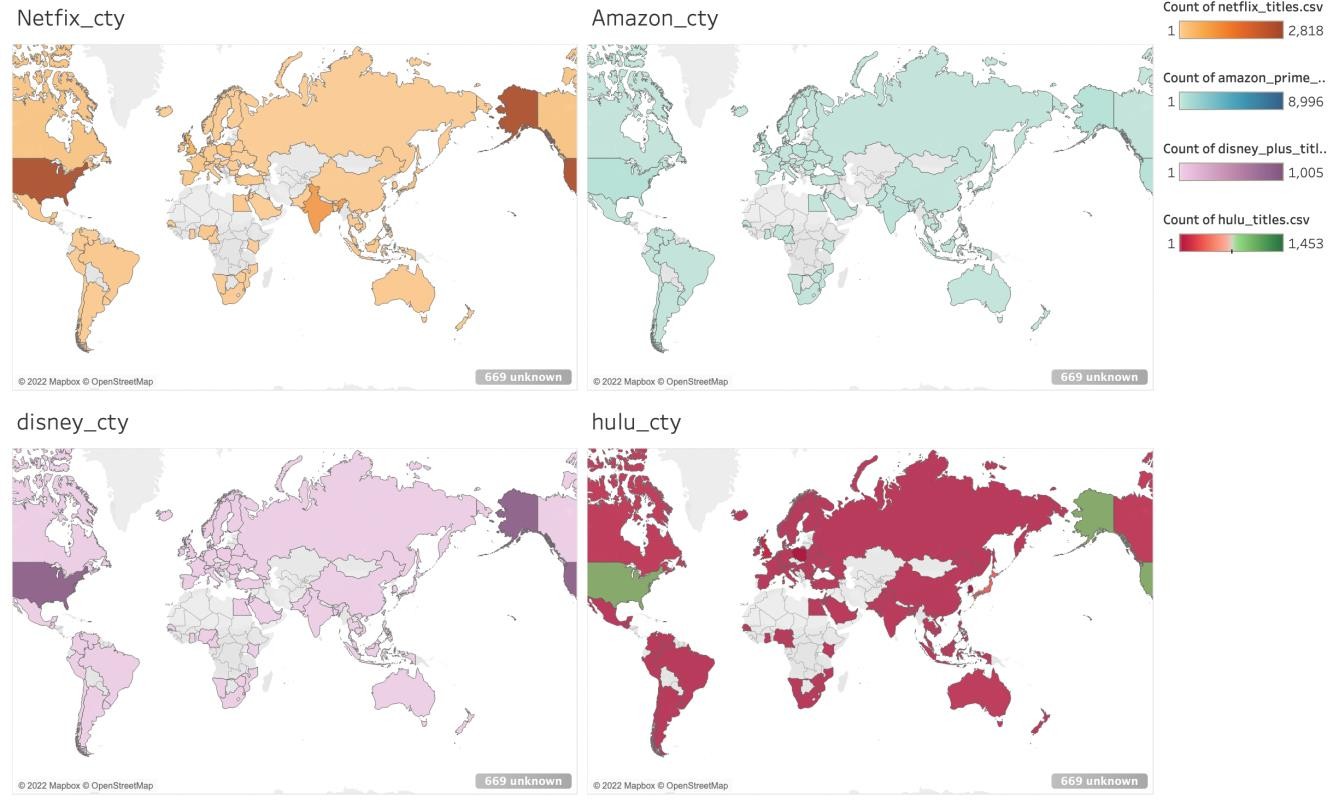




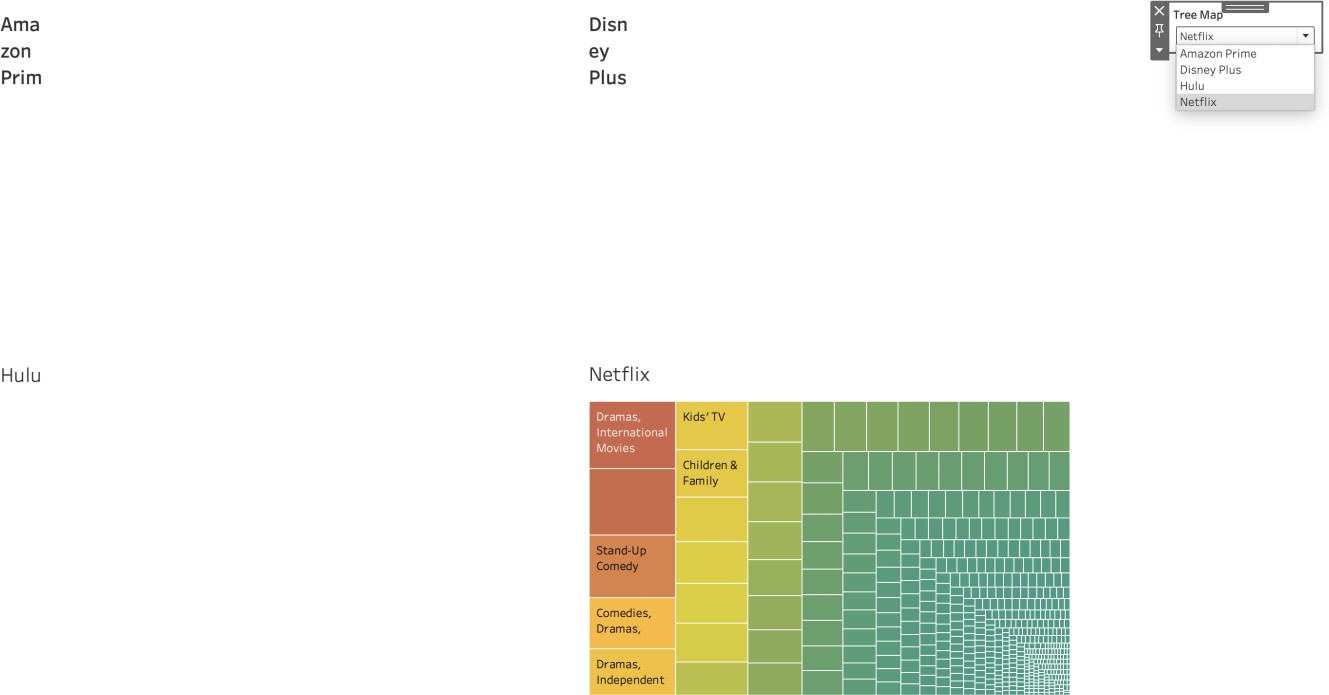


Interactive Plotly Graph in Python

# Creating Tableau Dashboard

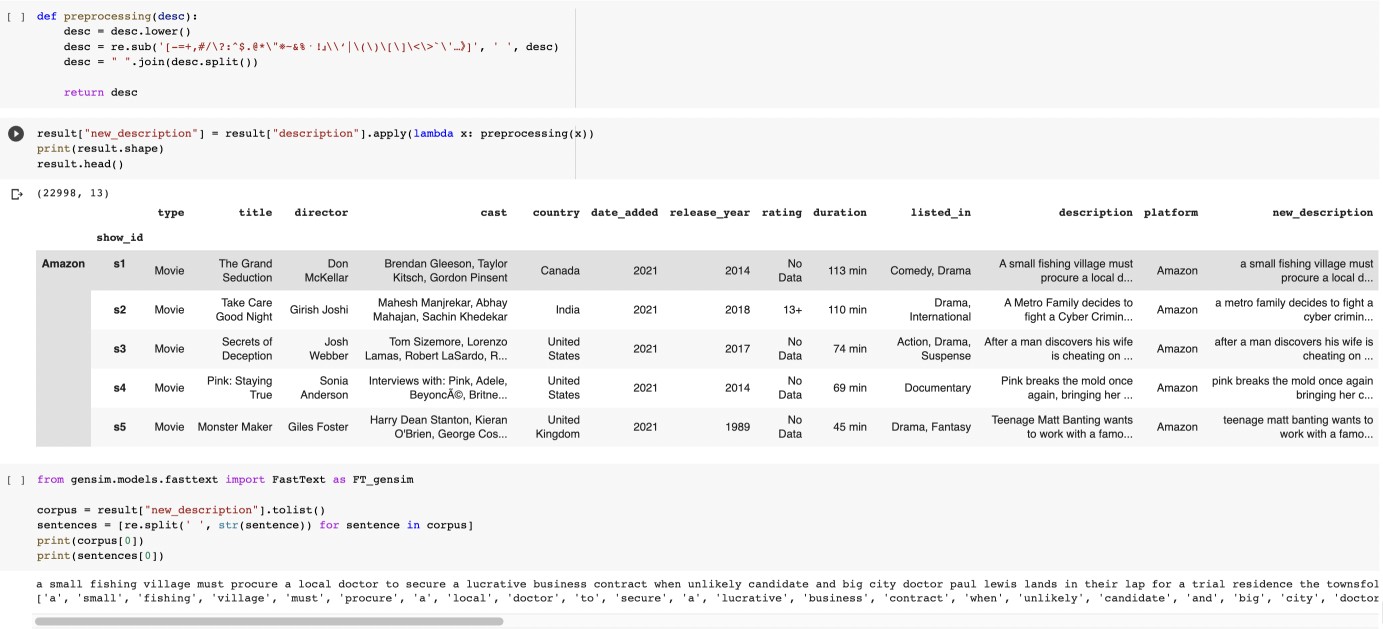


Dashboard displaying content by country in all OTT Platforms



Using Calculated Field to create dynamic Dashboard with a drop-down menu selector

# Creating Clustering and Regression Model



We first preprocess the description field to make it compatible with similarity checks

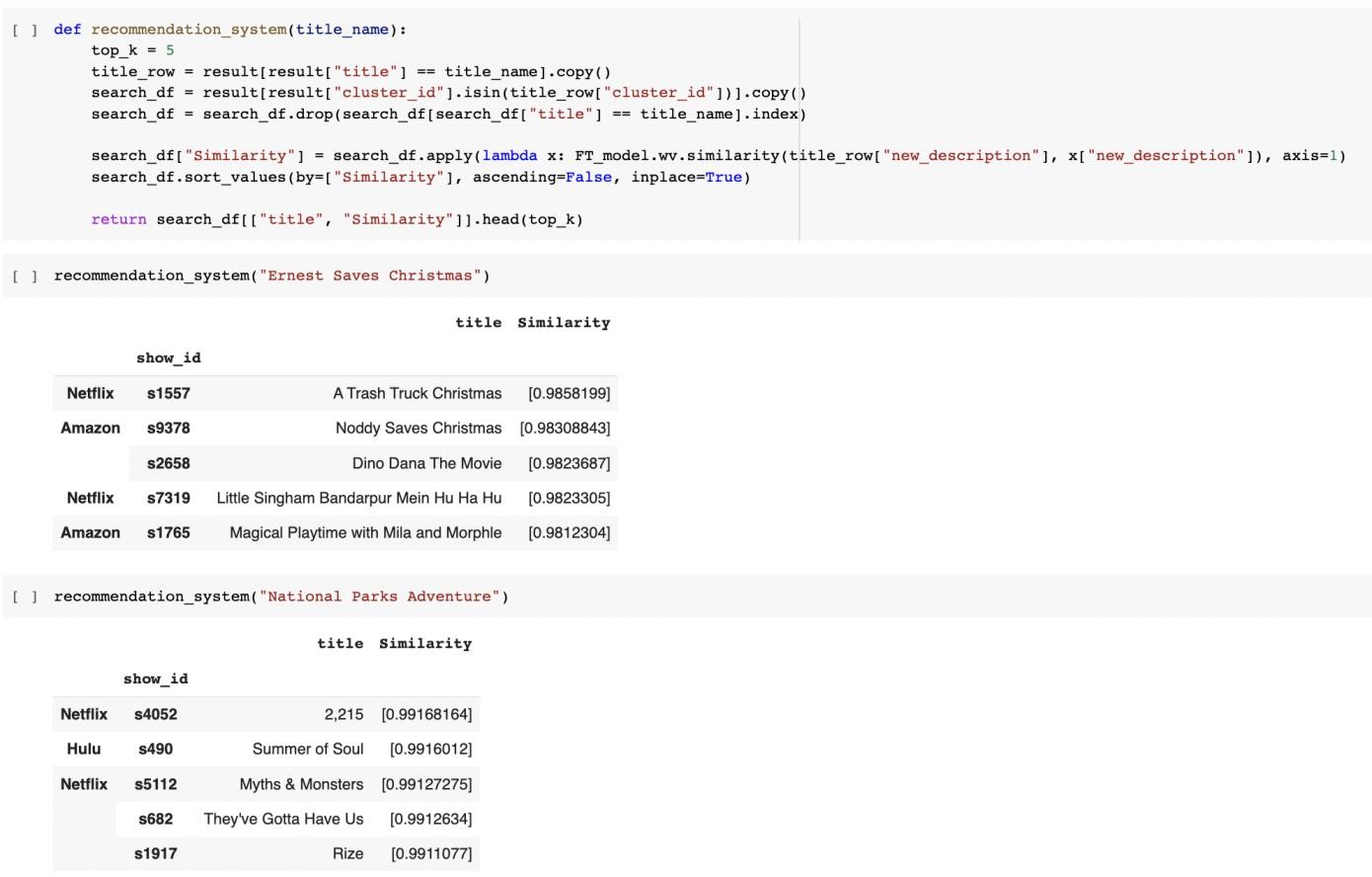


Then we make clusters using the K-Means algorithm and appending the cluster id with the dataset.

The Data is divided into a total of 49 clusters.



Now, we create the recommendation system



Our Recommendation System takes a movie or show name as input and then narrows its search space to the cluster that they belong to. Then it runs a similarity check on the description of the entered title with every entry on the cluster.

It then returns a list of similar movies and which OTT platform you can watch that content.

## CONCLUSION

From the Visualization we gained a lot of Inferences. Like how each platform values movies more than tv shows. We also found that Amzon and Netflix has the biggest content library with Disney & Hulu slowly building their catalogues. We also saw how US is the biggest producer of OTT Content with India coming at a close Second. We also inferred how the growth of OTT Content libraries has been meteoric in recent years, almost growing exponentially. We also saw the rating distribution between the OTTs and how they favor older teens/Adult markets as their main customer segment.

Finally, we created and tested the recommendation engine. We can see how such engines use clustering to reduce runtime dramatically while producing high quality results. This also highlights the importance of clustering data in large corporate environments like multinational OTT providers.

This proves that clustering isn’t just a mere visualization tool but also a very important machine learning implementation that reduces runtimes in such demanding worloads drastically.