**COIMBATORE INSTITUTE OF TECHNOLOGY**

**COIMBATORE – 014**

**DEPARTMENT OF COMPUTING**

**15MSSL10 – BUSINESS INTELLIGENCE LABORATORY**

**CONTINUOUS ASSESSMENT TEST – 03**

**Analytics on Movies & TV Shows in streaming platforms**

**DATE:** 03 NOVEMBER 2022

**DEGREE:** MSC

**BRANCH:** SOFTWARE SYSTEMS

**SEMESTER:** 9

**TEAM MEMBERS:**

1831009 – BOOBALARAGAVAN. P

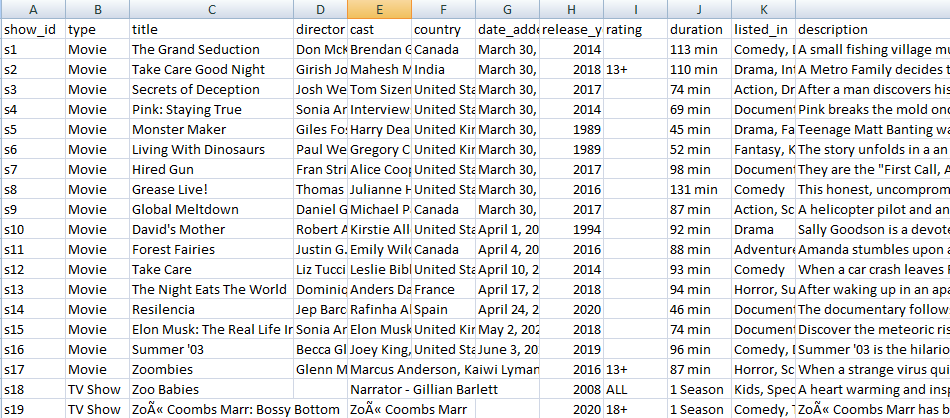
1831013 – DHARANI PRASAD. S

**ABSTRACT:-**

Business Intelligence systems combine the operational data with Analytical tools to present complex and competitive information to planners and decision makers. Analytics, a procedure that explores and analyses data to find hidden patterns to analyze patterns, to find deviations and data inconsistencies. Transforming the data into useful insights can be used to improve the speed and efficiency of the decisions to be taken. The objective is to improve the timeliness and quality of inputs to the decision processes. OTT platforms are becoming increasingly popular as more and more people shift away from traditional cable TV. Analyzing the details of Movies and TV shows streamed in Netflix, Amazon prime, Disney + to provide various reports and build a Data Dashboard. Fields in the datasets include Show ID, Type, Title, Director, Cast, Country, Date Added, Release Year, Rating, Duration etc.

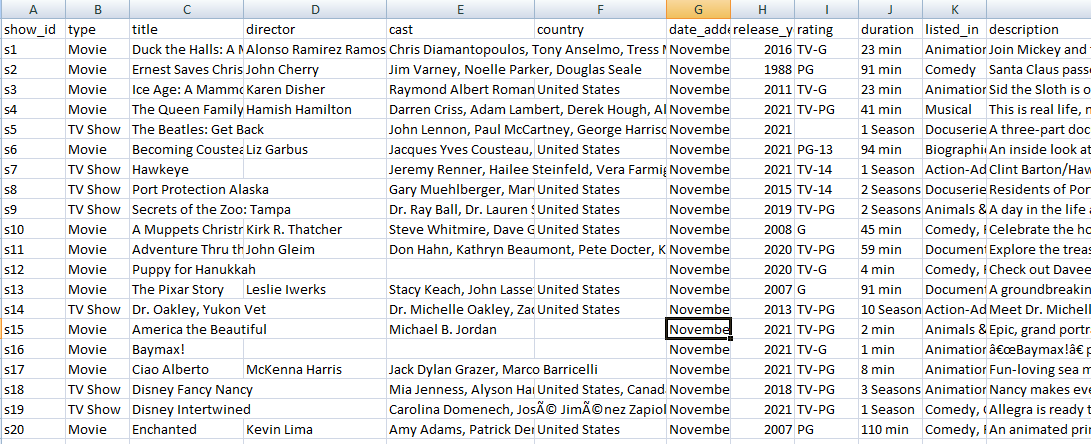
**DATA INTERPRETATION:-**

* ***Data Elucidation:***
  + OTT platforms which we took for analysis are Amazon prime, Disney + and Netflix. So, we have collected the dataset about these platforms from the website [Kaggle](https://www.kaggle.com/). So the datasets are
  + ***Amazon Prime :*** [amazon\_prime\_dataset](https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows)

****

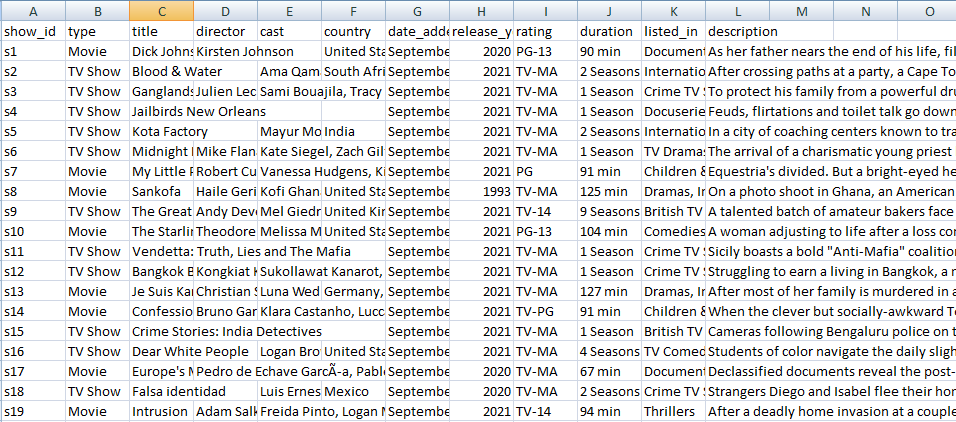
**Figure 1: Sample amazon prime data**

* + ***Disney + :*** [Disney\_+\_hotstar\_dataset](https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows)

****

**Figure 2: Sample Disney + hotstar data**

* + ***Netflix :*** [Netflix\_dataset](https://www.kaggle.com/datasets/shivamb/netflix-shows)

****

**Figure 3: Sample netflix data**

* ***Data Wrangling:***

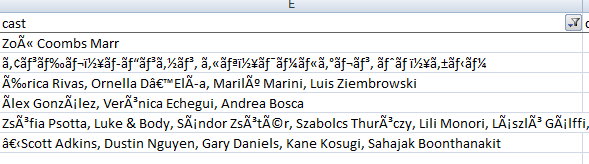
Data cleaning or wrangling prevents us from wasting time on wobbly or even faulty analysis and also prevents us from making the wrong conclusions, which would make you look bad. An improper data arises in the following scenarios

* + - ***Missing data:***



**Figure 4: Sample missing data from different columns**

* + - ***Invalid data:***

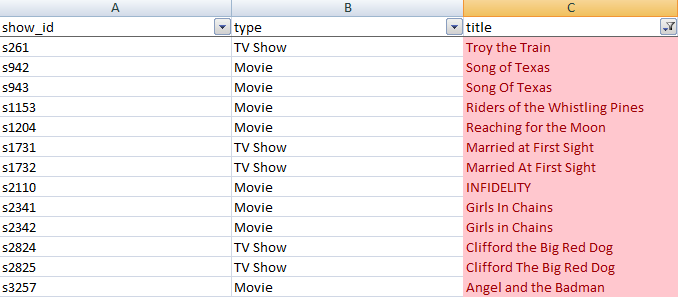


**Figure 5: List of invalid text or characters present in the dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| â | Ÿ | « | ‹ | ¼ |
| © | § | €™ | § | › |
| Ã | Ä | Ã³ | ‡ | € |
| ¶ | º | ‰ | ¶ | - |
| ± | È | œ | ¨ | # |
| Å | ¤ | ¦ | $ |  |

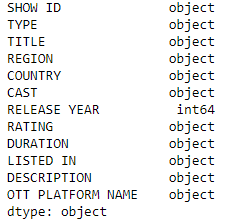
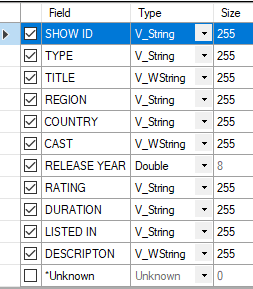
**Table 1: List of invalid characters**

* + - ***Duplicate data:***



**Figure 6: Duplicated rows with different ids**

* + - ***Data type issues:***

**Figure 7: Data attributes with improper Figure 8: Changing to the corresponding data types data types using Alteryx Tool**

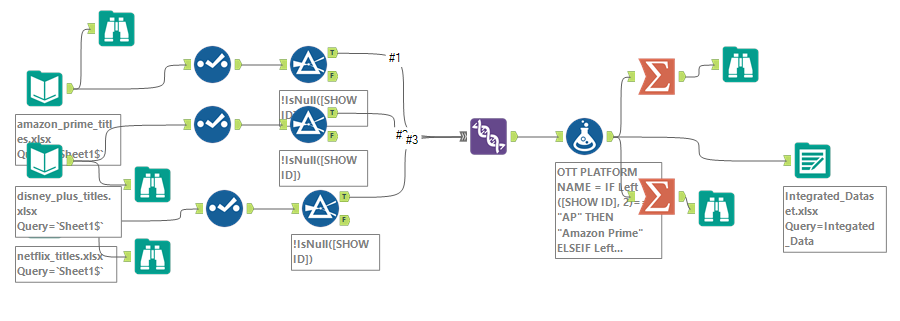
**ANALYSIS TOOLS AND IMPLEMENTATION:-**

Analysis tools and implementation briefly describes about the tools and features in the tools used for the analysis. It also describes about the programming languages used and how it is used and for what it is used. Also, it explains about the implementation of exploratory analysis in jupyter notebook using the python code and general ledger workflows using alteryx.

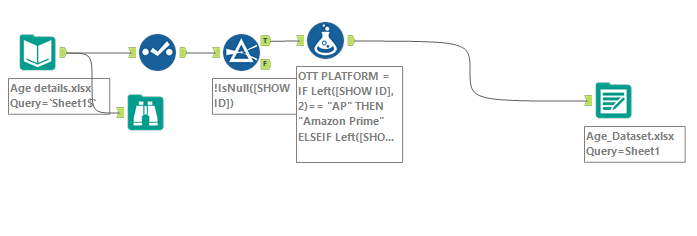
|  |  |
| --- | --- |
| **TOOLS USED FOR ANALYSIS** | Microsoft Excel |
| PowerBI |
| Alteryx |
| Jupyter Notebook |

**Table 2: List of tools used for analysis in project**

* ***Alteryx:***

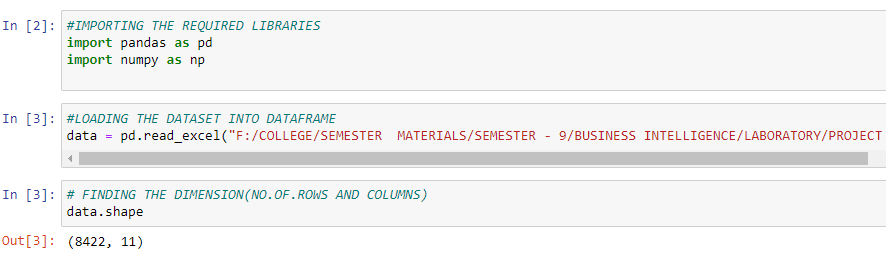
****

**Figure 9: A workflow in an alteryx tool which depicts the integration of multiple datasets using different nodes present in it.**

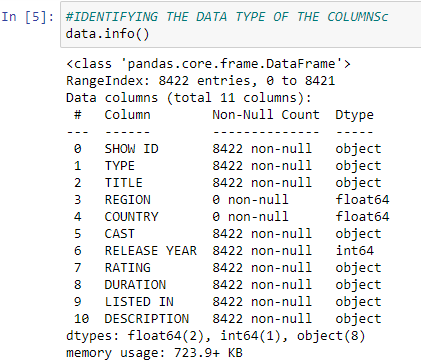
****

**Figure 10: A workflow in an alteryx tool which depicts the integration and pre processing of age dataset using different nodes present in it.**

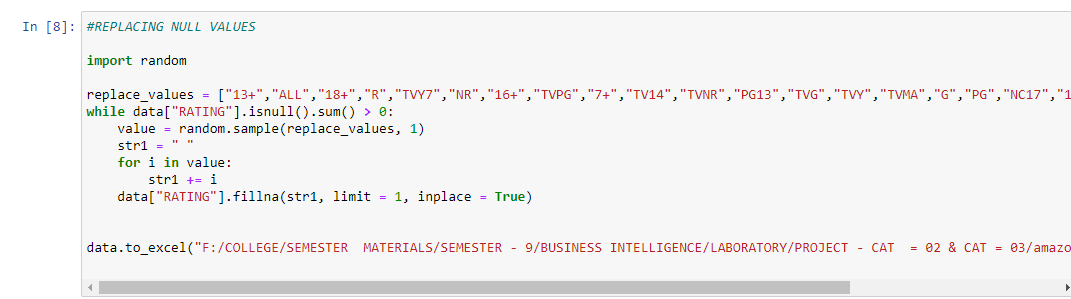
* ***Jupyter Notebook:***



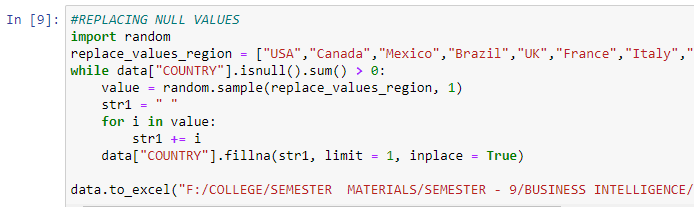
**Figure 11: Python implementation of EDA. Loading the dataset and finding the dimension.**

****

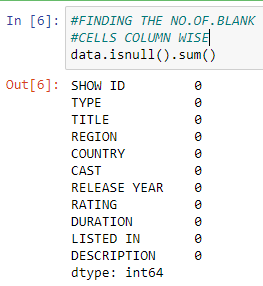
**Figure 12: Describing the dataset using each attributes with number of line items, data type and NULL & NON-NULL type.**

****

**Figure 13: Replacing the blank cells [NaN] in RATING attribute using random generator with fillna method and with list concept containing the summarized RATING.**

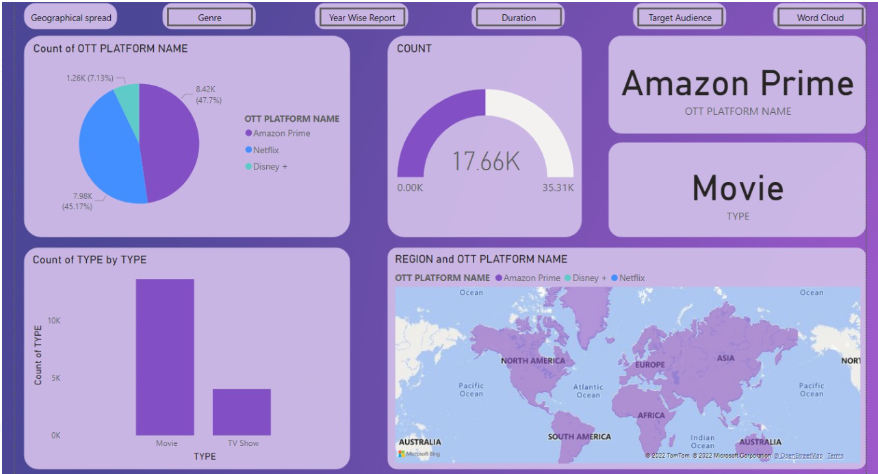
****

**Figure 14: Replacing the blank cells [NaN] in COUNTRY attribute using random generator with fillna method and with list concept containing the summarized COUNTRY.**

****

**Figure 15: Checking the availability of blank cells [NaN] for each column attribute.**

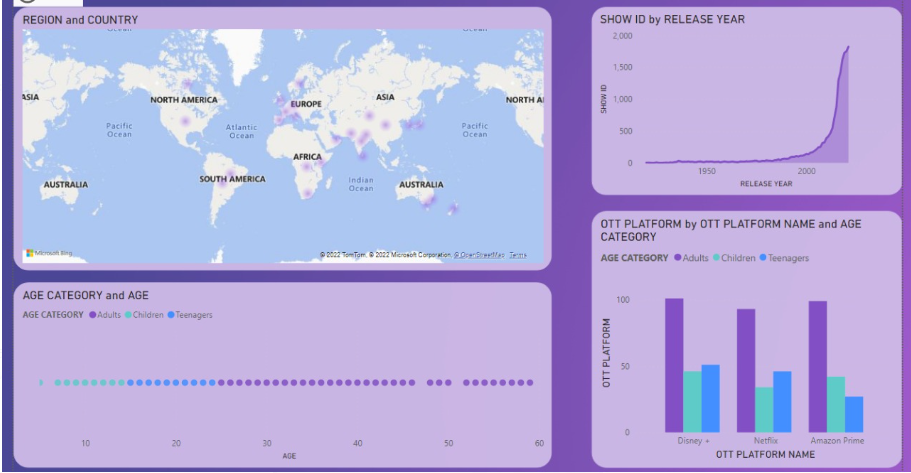
**REPORTS AND VISUALIZATIONS:-**

****

**Figure 16**

**Report 1:**

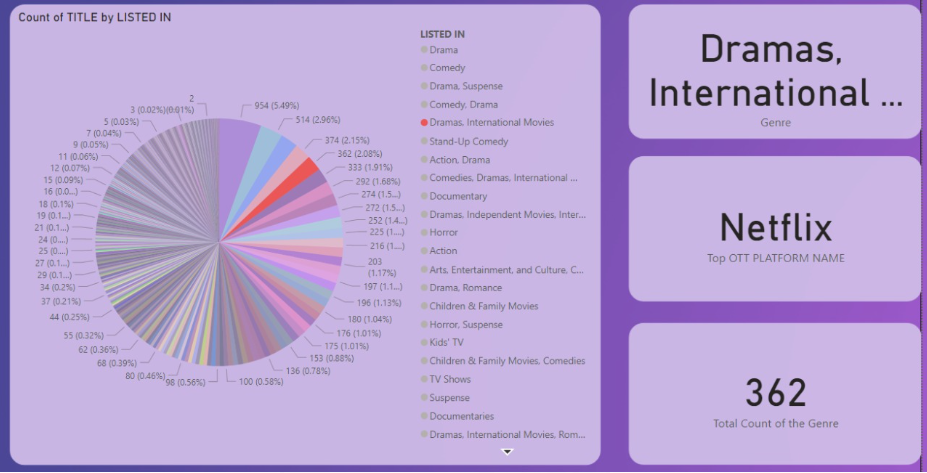
An overview dashboard which provides insights of OTT Platforms with Show types and count across different global regions. The buttons in the top are used for navigating to the further reports. The reports are dynamically updated for the user’s constraints available in the dashboard. The pie, bar charts and the gauge meter gives us the values of the contents for the respective selection.

****

**Figure 17**

**Report 2:**

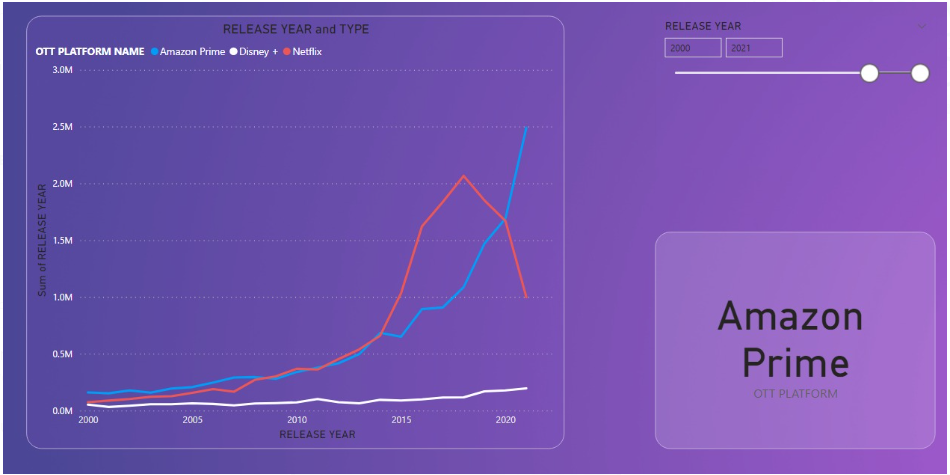
A geographical spread dashboard along with subscribers of varying age groups provides us the insights for most viewed content in each region. The scatter plot allows us to infer the age details cross to the age category. The reports in this dashboard are dynamically updated for the user’s selection. Release year with respect to contents for the specified region is depicted in the stacked area chart. The world heat map allows us to engage with the drill up and down process to view specific region or country release details.

****

**Figure 18**

**REPORT 3:**

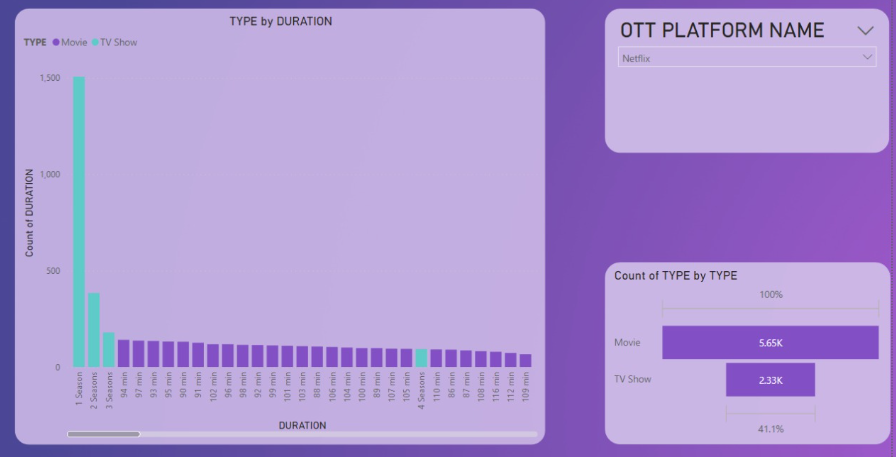
Genre dashboard provides the insights of genre across the different OTT platforms with total count. Dashboard is capable of updating dynamically for the user’s selection. This also provides the top streaming platform for the selected genre with the total count available in the information card. From the pie chart it is evident that there are wide spread of genres in the streaming platforms.

****

**Figure 19**

**Report 4:**

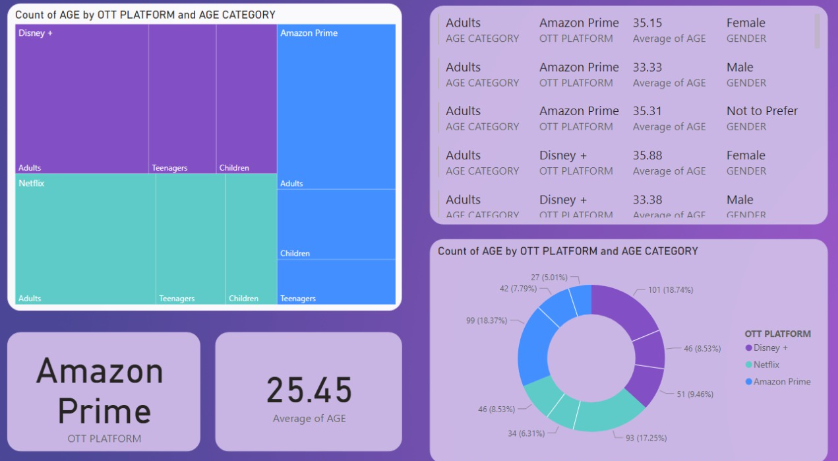
A year wise report dashboard helps us identifying the release year of the contents. Also depicts the top OTT platform trend over the years. It dynamically updates the line chart for the release year of the contents with the constraints having the range of year. It also depicts the lead streaming platform which infers more number of releases for the corresponding constraint.

****

**Figure 20**

**Report 5:**

The duration dashboard gives the streaming time of the content for all three platforms. A line and stacked column chart differentiates type of content with two column series that is movie and TV shows. The line gives the value of duration, total count for the selected platform dynamically.

****

**Figure 21**

**Report 6:**

Target audience dashboard consisting of Tree Map with age categories for the details, OTT platforms for the group data and age for the values. The Multicard is used to depict the age category with corresponding OTT platforms along with gender and average of age. The Donut chart and individual info card depicts the top streaming platform for the average age of subscribers. All the visualizations are updated dynamically for the constraints set.

****

**Figure 22**

**Report 7:**

A word cloud dashboard consists of three different word clouds for the corresponding attributes such as Title of the contents, Cast and crew details and Description of the contents.

**CONCLUSION:-**

The project provides analysis of all major streaming OTT platforms, performance reports and detects the risk factors in the OTT platforms.This provides visualizations for the multiple OTT platform data dashboard. It helps in identifying the performance based on geographical spread, top genre, target audience, year wise, duration reports, word cloud which enables the user to get insights from the huge volume of datasets which can never be possible in the short span of time.

This report ensures to result us the top OTT platform based on the filters we set. Across many filters, Amazon prime tops the list, followed by Netflix and Disney + OTT platforms. Amazon prime shall continue to enhance the richness of content they produce to the users across the globe to maintain the top track. Netflix must focus and revise their business plan and subscription plan to retrieve the top place leading Amazon prime. Disney+ as the emerging OTT platform shall focus on enriching and widening the contents to provide the subscribers the eccentric streaming experience.

**REFERENCES:-**

1. [Data Visualization Cheat sheets and Resources | Kaggle](https://www.kaggle.com/datasets/kaushiksuresh147/data-visualization-cheat-cheats-and-resources)
2. [🎥📀For Beginners:Netflix Visualization EDA🍿 | Kaggle](https://www.kaggle.com/code/melikedilekci/for-beginners-netflix-visualization-eda/notebook)
3. [EDA of Netflix success by Covid 19 | Kaggle](https://www.kaggle.com/code/abdallhwerby/eda-of-netflix-success-by-covid-19/notebook)
4. [Power BI Interactive Movie Dashboard Example](http://www.cmbi.com.au/Power_BI_Movie_Dashboard_Sample.html)
5. [Exploratory Visual System for IMDB Movies.pdf | Powered by Box](https://app.box.com/s/9sogilrr2i3iwn100b1vb6en2sxgm053)
6. [Word Cloud in Python | How to Build Word Cloud in Python? (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2021/05/how-to-build-word-cloud-in-python/#:~:text=A%20Word%20Cloud%20in%20Python%20can%20be%20created,task.%20...%204%204.%20Check%20for%20NULL%20values)
7. [Disney+ Movies and TV Shows | Kaggle](https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows)