Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1. **Poonam Dattu Shevkar** **Email-id :-** dattupoonam@gmail.com ❖ **Contribution -** • Checking the NAN values in the dataset. • Converting the date into date-time format. • EDA on highest watched genere on Netflix. • Understanding the type of content available in different countries. • Hypothesis findings based on release year and date feature. • Correlation matrix . • Standardized the data into scalar transformation. • Model implementation for K-means Clustering. • E-blow method & Box-plot for K-means clustering. 2. **Sanjay Ramkishan Verma** **Email-id :-** shankyverma1998@gmail.com ❖ **Contribution-** • EDA on type and ratings and to evaluate the highest rating with content preferred by the audience. • Analysis on the countries and cast in the show using word cloud visualization. • Analysis done to observe Netflix has increasingly focusing on TV shows rather than movies in recent years or not. • Data engineering on the basis of clustering text based features. • Model implementation for agglomerative clustering. • Scatter plot with type feature and with clusters after fitting K-means clustering model.
Please paste the GitHub Repo link-

GithubLink:-
https://github.com/Sara19598/Clustering-Analysis-on-Movies-TV-shows-Unsupervised-ML

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally. Netflix is a popular entertainment service used by people around the world. Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same datasetThe project's main goal is to create a model that can perform Clustering on comparable material by matching text-based attributes.

The tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.The data set contains 7787 rows and 12 columns.

The project's main goal is to create a model that can perform Clustering on comparable material by matching text-based attributes.

As the problem statement says, Understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that we have used K-means Clustering.

We perform the data wrangling on the raw data to get the useful data without NAN values and observe the summary statistics of the dataset. We prepared a dataset with feature engineering and feature scaling and also, dropped out the unnecessary columns.In the analysis ,the data was transformed into standard scalar form to implement the model.

We performed the exploratory data analysis.Reviewing all the data processing then the model was trained to form clusters.In which we observe as below- k-means clustering model gave insights of silhouette analysis consisting of 2,3,4,5,6 clusters.

For n_clusters = 2 The average silhouette_score is : 0.7049787496083262

For n_clusters = 3 The average silhouette_score is : 0.5882004012129721

For n_clusters = 4 The average silhouette_score is : 0.6505186632729437

For n_clusters = 5 The average silhouette_score is : 0.56376469026194

For n_clusters = 6 The average silhouette_score is : 0.4504666294372765

In the end, we plot boxplot to predict the hypothesis -

- After clustering, we can say that our alternative hypothesis is that the number of TV shows launched in the previous few years is not growing.
- Our second alternative hypothesis is the number of TV shows added to Netflix is high.

We evaluated that ,TV shows account for 2.8 percent of the total, while movies account for 97.2 percent.Netflix has added a lot more movies and TV episodes in the previous years, but the numbers are still low when compared to movies released in the last ten years.Movies are mostly watched in various countries rather than TV shows.

Drive Link:-

https://drive.google.com/drive/folders/11HYxRqzZ03sz5Mr78nWplwZq6fPAB-kM?usp=sharing