Mini Project 1: Structured Data

IST652 - Scripting for Data Analysis

A. The data and its source

The dataset utilized for this analysis was sourced from Kaggle, This specific dataset comprises information about shows and movies available on Netflix as of 2021.

Source Details:

Title: Netflix Shows and Movies
URL: Kaggle Dataset
Dataset Overview:

The dataset provides a comprehensive list of TV shows and movies available on Netflix. Each entry in the dataset includes details such as the ‘title’, ‘director’, ‘cast’, ‘country’, ‘date added’, ‘release year’, ‘rating, duration’, and ‘listed genre’ for the content, making it an rich dataset for understanding trends and patterns in Netflix's content..

Utilizing this dataset has facilitated a deep dive into the evolution of content themes and genres on Netflix, enabling insights into the platform's content strategy, preferences, and shifts over the years.

B. A description of your data exploration and data cleaning steps.

Data Exploration:

I started by exploring the dataset to understand its structure and identify any problems. We used functions like data.info() and data.head() to get an overview of the dataset, and data.describe() to understand the distribution of the data. We also used data.isnull().sum() to identify missing values.

Data Cleaning

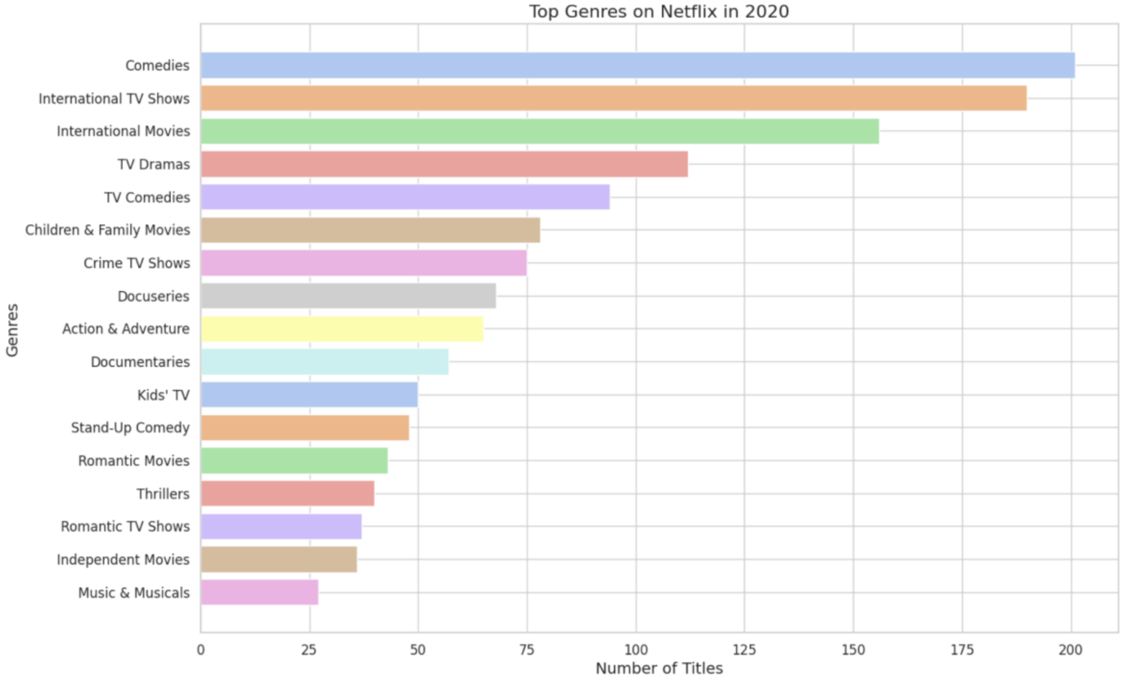
Once I had a good understanding of the dataset, I started the cleaning process. This involved handling missing values, standardizing the date information, restructuring the genre information, and ensuring that the release year field was consistently formatted.

I replace missing values with the placeholder "Unknown" and converted the date_added field from a string to a datetime object to make it easier to manipulate and extract specific time units. And then removed any entries with null date_added fields to prevent inconsistencies in later time-based comparisons and aggregations. We restructured the listed_in field by translating the genre information into separate binary columns for a more refined genre-based analysis. And finally made sure that the release_year field was consistently formatted as an integer across all entries.

C. Two clearly stated comparison questions with the unit of analysis, the comparison values and how they are computed.

1. What are the common characteristics of content added to Netflix in 2020?

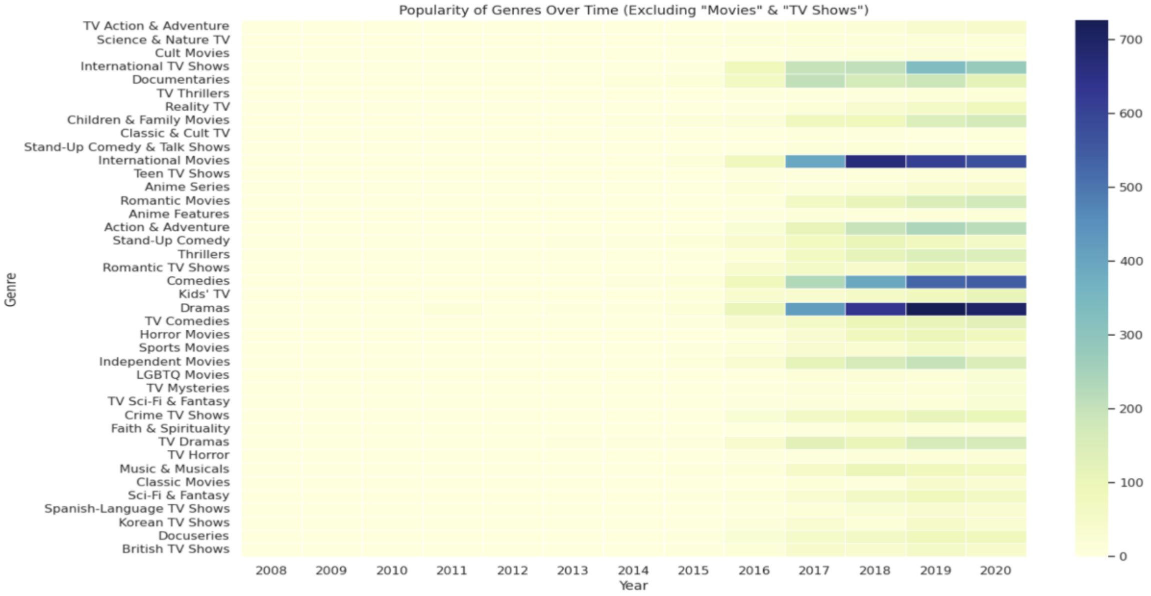
Unit of Analysis: Individual content titles added in 2020.
This analysis involves looking into several aspects of the content, such as the type of content (movie, TV show), the genres into which it falls, and the regions that create the most content. The calculation is performed by aggregating the data for the content added specifically in 2020, and using techniques such as 'value_counts()' to understand the frequency of different categories within each attribute. For example, to find the most prevalent qualities, analyzing the frequency of each genre and content type. Also examining the 'country' field to find the top content-producing countries. The findings show the most prominent genres, categories, and countries producing content in 2020..



2. How have the themes and genres of Netflix content evolved since the platform's inception?

Unit of Analysis:* Content titles across all available years on Netflix.
This section of the analysis focuses on genre historical trends over time. The counts of different genres across years are the major comparative values here, showing shifts and increasing preferences in content themes. The computation entails yearly data organizing and estimating

the frequency of each genre. the plan is to build a year-on-year snapshot of genre popularity by using methods like 'groupby' on the 'year_added' column and aggregating genre data. The evolution of the analysis is extended by recognizing new genres that have acquired traction in recent years and observing notable shifts in genre dominance, which provide insights into Netflix's shifting content strategy.



D. A description of the program

1. First I have started off by by filtering the data to year 2020 as the first question involves with 2020 data with the code “data_2020 = data[data['release_year'] == 2020]”. Then to find out the genre with highest number of titles used the following code and also sorted them in descending order “genre_counts = data_2020[genre_columns].sum().sort_values(ascending=False)”.
   In the same way found out the countries who have more titles on Netflix in 2020 by using the code “top_countries_2020 = data_2020['country'].value_counts()”
   Finally summarized the findings using the following code “summary_2020 = {
       'content_type_distribution': content_type_distribution,
       'top_genres': genre_counts,
       'top_countries': top_countries_2020
   }
   summary_2020”
2. For the second Question found out the top genre by year by using the groupby function with the following code “genre_counts_per_year = data.groupby('year_added')[genres_excluding_movies].sum()”. then found out the CAGR for genres over the years . Also used. sort_values() function to sort them. And finally used matplotlib to create some visualizations.

E. A description of the output files
   Output file shows the analysis the first sheet shows the summary of Netflix on the year 2020 and the next 2 sheets shows the trend of Netflix over the years

F. The source data file.

Kaggle Dataset