DATA CLEANING, ORGANIZING, AND VISUALIZATION USING MICROSOFT EXCEL.

THE PURPOSE OF THIS STUDY IS TO DETERMINE UNDERSERVED STREAMING DEMAND IN GLOBAL MARKETS.

GUIDING QUESTION.

1. What three movie and program genres grew the most by the number of titles added from 2016 through 2019?

DATA

Netflix data derived from Kaggle <https://www.kaggle.com/shivamb/netflix-shows>.

Import the CSV file into Excel for exploration and save it in a proper naming convention.

DATA EXPLORATION

The data has 6235 records and 12 Variables.

The variables include; show id, type, title, director, cast, country, date added, release year, rating, duration, listed in, and description.

Use the filter function to check for valid values, dirty, missing data, and potential outliers.

Findings.

1. Show id has no blanks.

2. Type has two categories; Movie and TV Show.

3. Country has many nations separated by commas.

4. Date added has two date formats. There are 12 blanks which represent an insignificant percentage of over 6000 records.

5. Rating column has blanks.

6. Duration has both number and character data types.

7. Listed in has several genres.

8. We won’t use title, director, cast, description, and release year columns in the analysis.

DATA CLEANING.

1. Check for duplicates from the data drop-down menu.

The show id column is the primary key of the data and should therefore have unique values.

Findings

No duplicates are found.

1. Check for spelling from the review drop-down menu.

Country and listed in columns should have incorrect data spellings.

Findings

The data is good to go.

1. Format the date added column for consistent dates.

Use the trim function to remove spaces, then copy and paste special the values.

Convert the date column from text to date data type.

DATA MANIPULATION.

1. Parse the data.

Separate compound data. (Data that has more than 1 data value residing in a variable).

We use text to columns to separate values in the variables.

The duration column is parsed into duration and unit. The unit column has minutes, season, and season categories.

1. Recode the data.

Text listed in a column is delimited by commas. Parse it into a new sheet.

Sort the new columns so that the blanks are placed at the bottom.

Cut and paste the text from columns B and C at the bottom of the listed in column.

The total number of records is 13,671.

Trim the data to remove spaces.

Create a pivot table to check the frequencies of the genres. This ensures we've not lost any values. Sort the values to find out the most popular genres. You can also show the values of the pivot table as a percentage of the total.

1. Transforming the data

There are 42 unique genres that can be further consolidated by similar grouping genres from movies and TV type listings. E.g., comedies and TV Comedies are one genre.

Create a lookup column next to the listed in all columns with unique values.

We are left with 30 unique values.

Use the VLOOKUP function to fill in listed in R1 to R3 columns.

=VLOOKUP(C:C,'LOOKUP TABLE'!$A:$B,2,FALSE)

DATA ANALYSIS AND VISUALIZATION.

Guiding question.

1. What three movie and program genres grew the most by the number of titles added from 2016 to 2019?

Create a new sheet and copy the useful columns; show id, listed in R1, listed in R2, listed in R3, and date added columns.

Each listed genre should have a show id, and date added. Sort the data in listed R2 and listed R3 so that the empty rows can be placed at the bottom.

On a new sheet, convert the broad data into a long format. In the long-format data, there are 13,670 observations.

Visualization.

Insert a pivot table to summarize the long data.

Filter the year to show 2016, 2017, 2018, and 2019. For the listed in data, filter the top 5 genres based on the total count of show id.

Sort the genre in ascending order based on the count of show id. The top 5 genres are international, dramas, comedies, documentaries, action & adventure in descending order.

Insert a bar graph that derives data from the pivot table.

Right-click on the pivot table and summarize values as a percentage of column totals.

To get the rate of change, use 2016 and 2019. Compare the two years. Select out the tiny genres with less than five show ids.

Show values as a percentage difference from 2016.

The top three movies that grew the most from the bar graph are; Action & Adventure, Horror, Sci-Fi & Fantasy.

1. How has the type of entertainment changed between 2016 and 2019?

We need the show id, type, and date added variables. The type has two categories which make a pie chart the best chart to use.

Findings.

TV shows have always been less than 50% of Netflix's offerings.

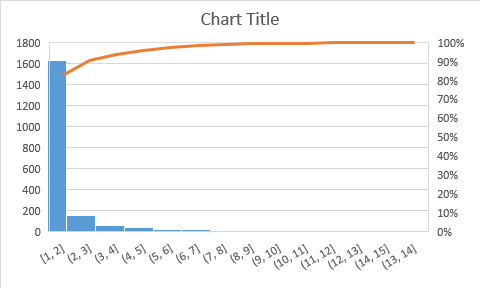
1. What can we learn from Netflix data based on duration offering?

Under duration, the TV Shows has season and seasons. Find seasons and replace the name with the season for consistency.

Insert a pivot table consisting of the duration data.

Findings.

Over 90% of TV shows have either one or two seasons.



3. Duration of movies.in minutes.

From the histogram, the data is typically distributed. Most movies last between 80 and 90 minutes.

