NETFLIX DATASET POWER BI PROJECT

BY: DONIA SALHEEN

HOSSAM EL-DEEN HAMED

MAHMOUD EL-SAYED

BASSANT HANY

A. Data Preprocessing:

- Acquiring the needed datasets which were 3 (Netflix's dataset, Disney + dataset, Netlfix Subscription fees dataset)
- Loading the datasets into powerbi
- Cleaning the netflix dataset which had a few corrupted rows by removing them
- The other two datasets required no cleaning
- Change the data types of the columns to its appropriate types (text to date, text to numeric .. etc.)
- Adding a new column called platform to both disney and netflix tables containing the platform's name to distinguish them upon appending
- Appended the two tables together as they have the same structure
- After appending it appeared that there are duplicates in show_id column so we created an index column
- The column "listed_in" has the genres separated by commas, so we took this column along with other relevant columns in an other table to expand the "listed_in" column by rows so we each row represents a movie/tv show and one of its genres to help in analysis and visuals related to genres
- Same procedure were done to the cast column
- The column ratings was used to generate the column "rating_ages" as each group of ratings that has the same age group audience was classified into (Kids, older kids, teens, adults) for simplifing this info.
- The table subscription_fees required no cleaning nor transforming

B. Challenges:

- The netflix data had a lot of missing data in the director column
- There were no data about disney + subscription fees as the platform is fairly new and isn't available world wide yet
- The netflix subscription data doesn't include every country netflix is available in
- We wanted to include amazon prime data along disney and netflix but it was full of null and wrong values and it was not very clean
- The duration column was tricky to deal with as it had two different units of measurements, minutes for movies and seasons for tv shows