Data Analysis Report

- **Hikansh Kapoor (S3803669)**

- ## Data Preparation

This was the very first step in this project that I conducted for data cleaning and organization. The main focus was to organize the data in such a way that we can use in the next step that is data exploration to visualize the data using bar graphs and charts. I followed the steps below to ensure that the dataset is clean and organized in an appropriate manner :

1. **Loading the dataset and renaming columns :** I used pandas as the core library for this project to load the dataset of Star Wars into a variable called 'data' and then renamed all the columns to some short names by using appropriate string functions like 'lower()'.
2. **Checking data types :** I used the 'dtypes' attribute to check and confirm the data types of the loaded data is equivalent to the original data in csv file.
3. **Typos :** Typos are constantly checked for each and every column by using value_counts() method and then fixing the typos right there. For example, replacing 'Yess' with 'Yes' etc.
4. **Extra-Whitespaces :** Again by using value_counts(), I have checked my data and every column so as to make sure that there are no extra whitespaces. If there are some, I used strip() method to strip off the whitespaces.
5. **Lower-Case :** I used str.lower() method to lowercase all the column names so as to make it easy to select a column.
6. **Sanity checks :** Sanity checks like info, value_counts(),isnull() and dtypes are used to check if the data contains any impossible or invalid values. value_counts() is used for every column so as to confirm this.
7. **Missing Values:**
    a. **Removing Null values from ID column :** As RespondentID has to be unique and not null, I simply used the pandas 'notnull()' function to remove any null value from that column and by this the very first row was removed that had NaN as the ID. This was essential for us to explore the data in the further stages of this project.
    b. **Handling Null values in the rating columns :** I simply filled the null values in these columns with rating 0 as it would not be considered while exploration of the data. I simply used 'fillna()' method for this task.
    c. **Handling Null values in character columns and further columns :** Again, I used 'fillna()' but this time, I filled all the null values to 'Not Voted' instead of any mean or median value or 0. In case of Gender I filled null values with 'Not interested in disclosing'.
8. **Dropped a column with more than half null values :** The column which was named 'Fan of expanded universe?' was dropped as it had almost 1000 value as null which is not good for data analysis.

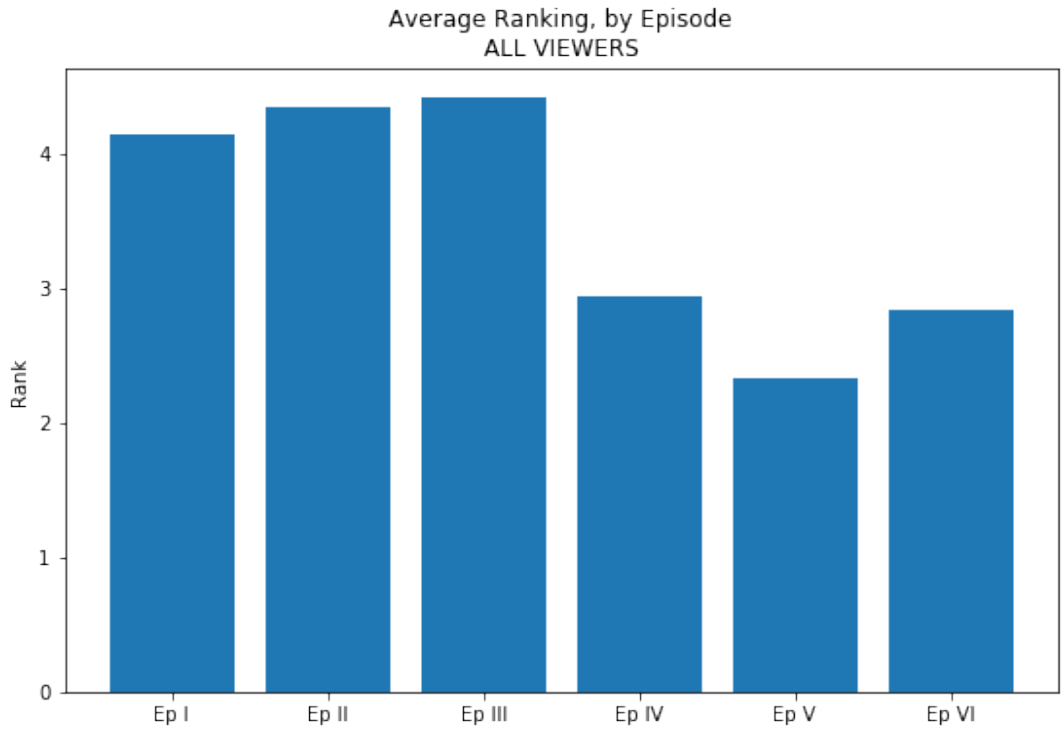
Note : I have also changed the column names as well and mapped the values where it was needed to such as in columns 3 to 8 and 15 to 29 as well.

- ## Data Exploration

In this step, I tried to find relations between several columns in our dataset. Different visualization techniques were also used and data was manipulated as per the requirements of the technique as well in the same step. I tried to compare as much columns as I could and present them in a beautiful pictorial representation at the same time. In all of these representations I used matplotlib library for the beautiful plots.

1. **Exploring the survey question : Please rank the Star Wars films in order of preference with 1 being your favourite film and 6 being your least favourite film. Then analyse how people rate these films.**

I started off by taking the mean in every rating column and then plotting it with the average rating being on the Y-axis and Episodes on X-axis.

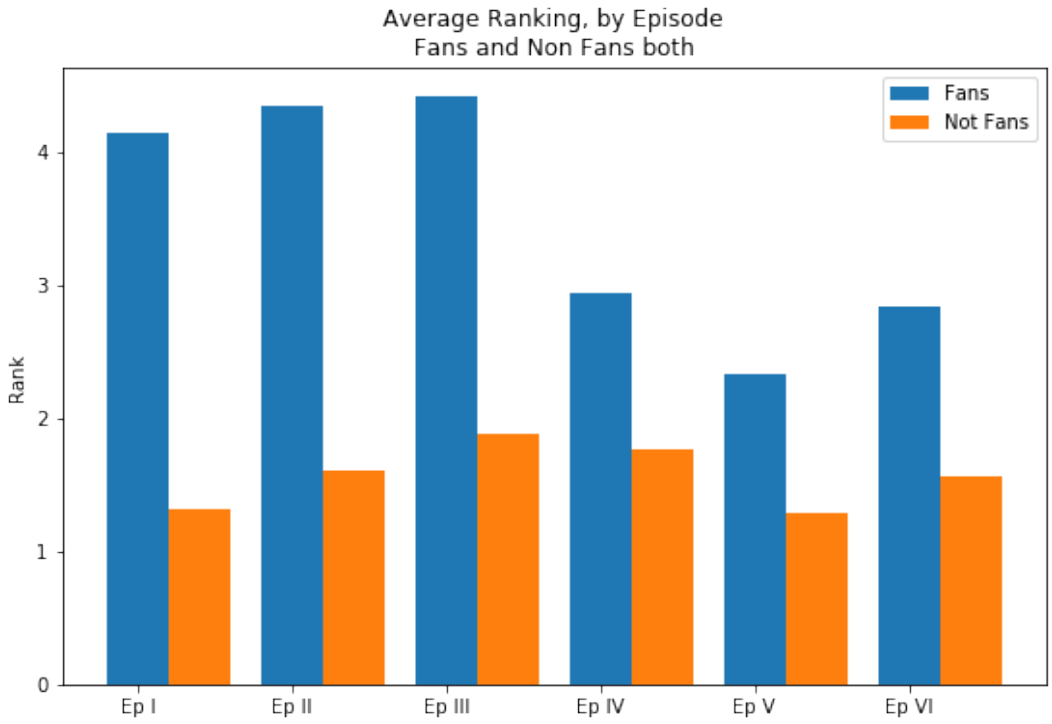


Looking at this bar graph one can easily see that Episode 5 was the favourite for most of the people as its average rating is low and remember 1 is the most favourite and 6 is the least favourite movie. Similarly, episode 3 was the least favourite on an average.

2. **Exploring relationships between columns :**
    a. **Average ranking of episodes by Fans and Non-Fans :**

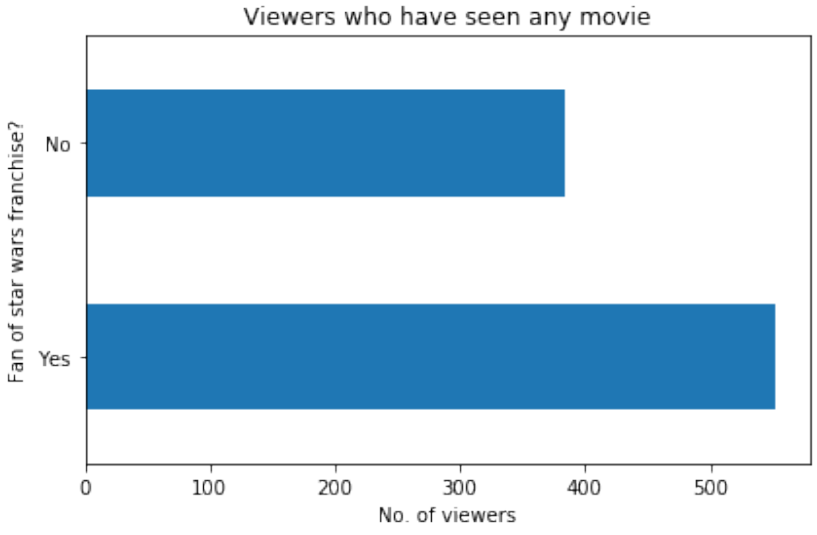
    I segregated the data frame into 2, one with all the fans of the Star Wars franchise and the other with non-fans and then tried to visualize how the two categories rate each and every episode.

Hereby showing the comparison between the average rating of fans(Blue) and that of non-fans(Orange) as well. Its interesting to see that both of them has rated episode 5 as their favourite episode.



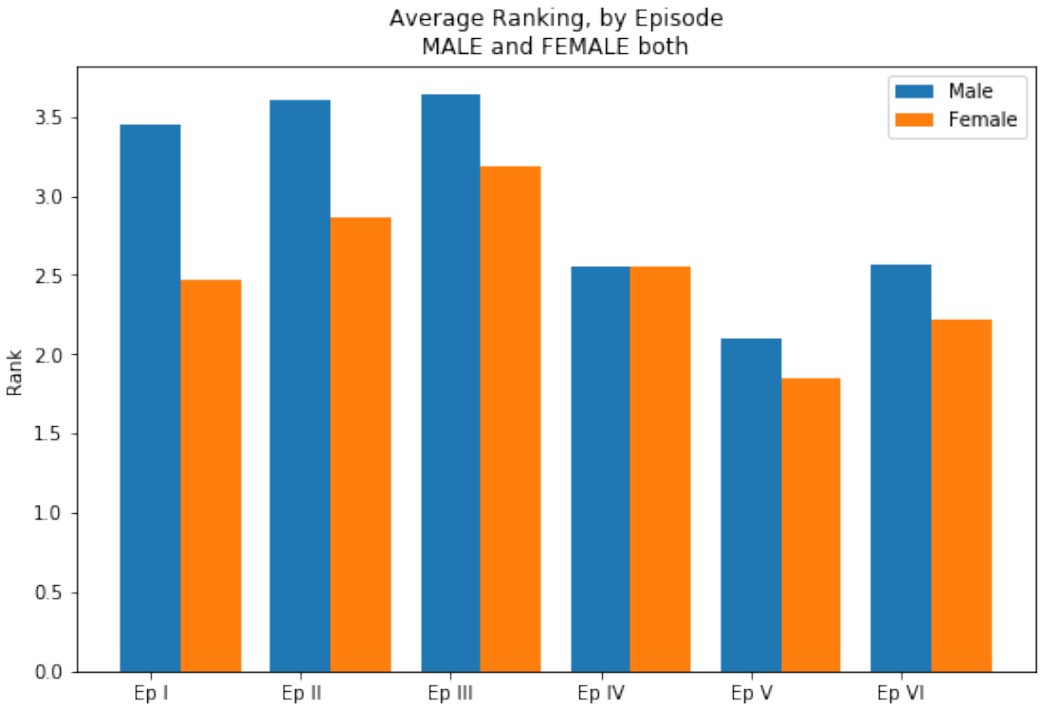
b. Relationship between viewers and Fans :

In this relationship, I tried checking how many viewers are actually fans of the Star Wars franchise. I simply segregated all the viewers(seen any movie) into a data frame and then plotted a horizontal bar graph of whether they are a fan of franchise or not. We can easily see that most of the people came out to be fans of the franchise.



c. **Relation between gender and movie rankings :**

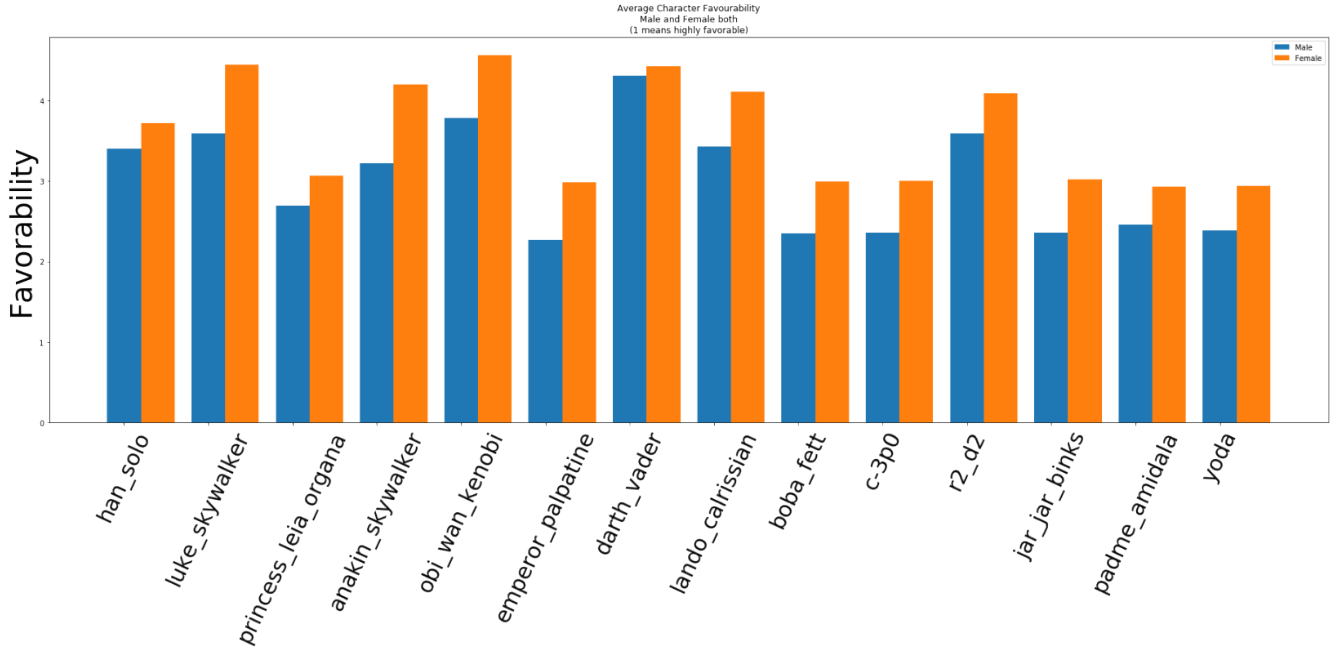
I segregated the data frame into 2, one having all the males and other with all the female gender. Then by using the same technique, I tried to plot the comparison graph as how the two genders rate the movies and the corresponding graph is as below.



We can observe that irrespective of the gender, episode 5 was the one with lowest mean and hence was the most liked episode by both males and females.

3. **Exploring the relationships between people's demographics and their attitude to Star Wars characters :**
   a. **Relationship between gender and character preferences :**

In the same manner as above, I segregated the dataframe on the basis of gender but this time I started comparing the character preferences for males(Blue) and females(Orange). The corresponding graph is as below. Its quite difficult to see the preferences as some of the rankings by females are almost same for most of the characters but for males, one can see 'emperor_palpatine' was the most favourable character and for females, it were 'yoda' and 'padme_amidala' with the same amount of ratings for highly favourable.



b. Relationship between Age group and character preference :

For this, I first had to drop the outliers in the age column that was age=500 as it is impossible value for a human being. Next, I divided the dataframe on the basis of age groups in the age column :

18-29 : Blue

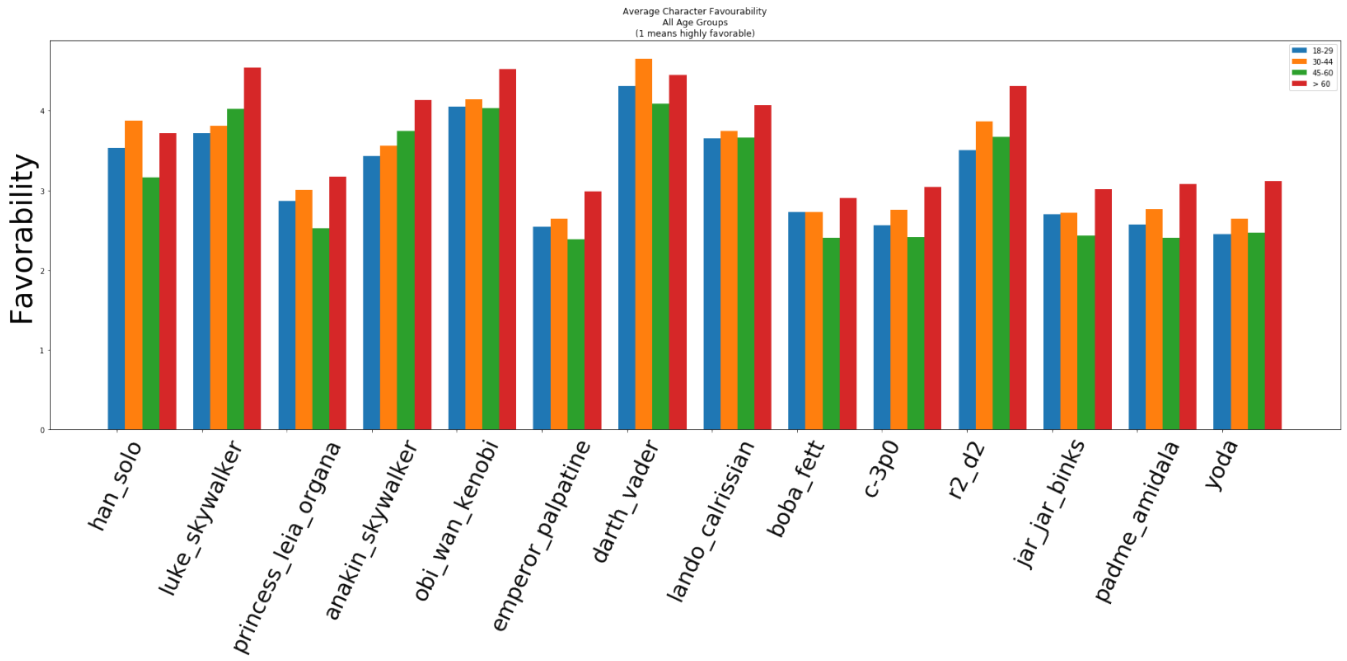
30-44 : Orange

45-60 : Green

Above 60 : Red

The plotted graph is as below. One can easily observe that ‘darth_vader’ was the least favourite amongst all the age groups while for other characters, different age groups had different preferences as one could see.

For instance,‘emperor_paltapine’ was liked by age groups 18-29 and also 30-44. It can be seen that the other 2 age groups had several character coinciding with almost same amount of preferences like 'Jar_Jar_Binks','Padme_Amidala' and 'Yoda'.



- References

1. Visualization techniques : https://www.kaggle.com/alexisbcook/bar-charts-and-heatmaps
2. Data manipulation : https://www.kaggle.com/residentmario/data-types-and-missing-values