1. Data Preparation

1.1 . Check data types

After retrieving the dataset, some of the column headers had unknown names to it. Firstly, the column names were renamed with short and meaningful names. Then, depending on the unique value counts for each column, the column data types were converted using pandas built-in function astype(). The columns which were converted into category are (Gender, Age and Household Income) whereas the columns containing ratings (rank_1 to rank_6) were converted into float.

1.2. Typos

A loop was used to iterate through all the columns checking for typos. The results were displayed for each column with their unique value counts. The typos were in 2nd , 32nd and 33rd columns with values as 'Yes' or 'No' followed by some typo errors. The typos were also in and 34th column with values as 'Female' and 'Male'. Pandas function loc[] was used to replace the values based on the given condition by assigning a value to it.

1.3. Extra-Whitespaces

By looking at the results displayed earlier, there was an extra-whitespace in 2nd column. The 2nd column had 2 values with 'Yes', this is because there was an extra-whitespace for one of the 'Yes'. A str.strip() function from pandas was used to remove the extra-whitespace.

1.4. Upper case/Lower case

To cast all data into upper case str.upper() function was used. However, some of the data represented different data types which were then converted into string data type followed by string function. To make the changes to the starwars dataset apply() function was used.

1.5. Sanity checks

After getting the unique value counts for each column, there was a sanity check error in Age column. As the Age column contained categorical data there was no need to run checks based on conditions. Loc[] was used to replace the value (i.e. 500) into 45-60 Age category. As there was only one value which was needed to be replaced, mode() function was used to calculate the highest frequency and replace the 500 value over to the highest frequency in Age category.

1.6. Missing values

Pandas' inbuilt function isnull() was used to find the null values in the dataset and then the sum() function was applied to get the total null values for each column. By looking at the result, most of data included null values. Some of the rows included null values at each column except for 2nd column where there was no null value. Assuming, as there was no demographic data for those rows, dropping those rows wouldn't make any difference. Therefore, those rows with 'No' and Null values in all of the columns were dropped. The rows were dropped by counting the total sum of null values at each row and matching them with a condition. The totals of 100 rows were dropped.

Then, based on a given condition (characters having null values and people who either watch (yes) or didn't watch (no)) the results were displayed. This helped to track how many were null values where people did watch any of the movie and where people didn't watch any of the movies. By further analyzing the behavior of the data, the missing values for some columns were replaced on a given condition and some were replaced using ffill() method (by filling the null values from previous rows). The columns which were replaced by ffill(), wouldn't had any impact on the analysis if those null values were either ignored or filled. The columns which were replaced on a condition were replaced with either one of the value from each column.

2. Data Exploration

2.1 . Explore a survey question

The dataset contains different ratings for 6 starwars films. Plot() function is used to plot the ratings for each movie. As there are multiple ratings for each movie, the sum of those values were calculated to find out which movie had highest rating and which movie had the least rating.

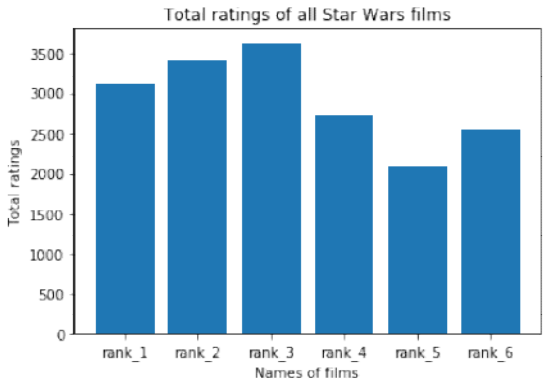

Fig 1.1

After, calculating the sum of ratings for each column, a bar graph was plotted. The fig 1.1 represents the bar graph of the total ratings for each column. The results include rank_5 (i.e. starwars movie 5) being the highest rated movie and rank_3 (i.e. starwars movie 3) with the lowest rating among all.

2.2 Relationships between columns

Gender and Household Income were the columns which were selected to find the relationship between them. These two columns were grouped and the total count for each Household value corresponding with the gender value was calculated. Then, for analysis bar graph, pie chart and scatter chart was plotted.

Fig 2.1

In Fig 2.1, we can see that most of the people selected the 3rd category 50,000 – 99,999 and very few had selected the 5th category 150,000. Moreover, most of the people who selected the 3rd category were Females and the rest were Males whereas most of the males selected the 4th category and very few Females selected the 4th category.

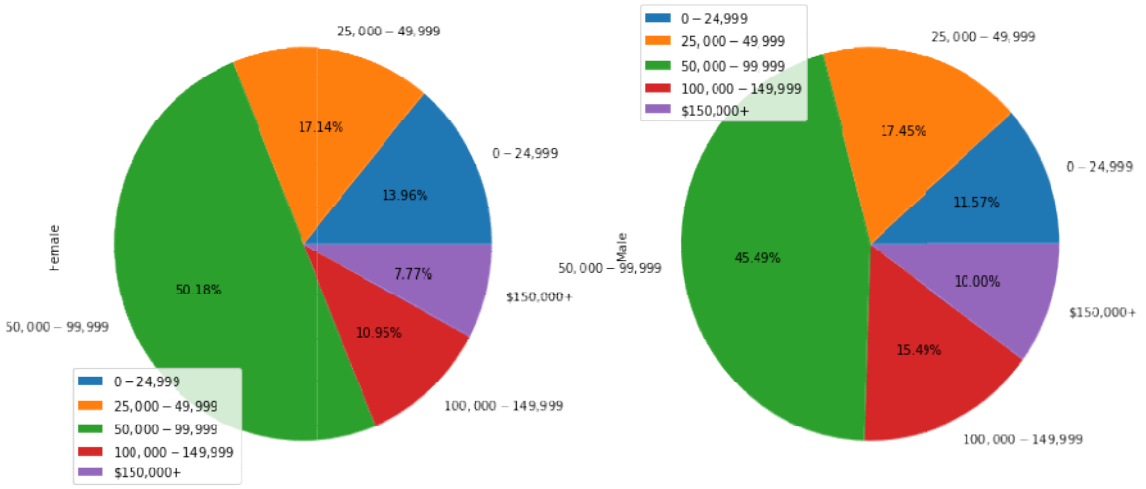

Fig 2.2

In Fig 2.2, it can be said that majority of the population selected the 3rd category 50,000 – 99,000. Half of the female population had choosen the 3rd category. Both the population had almost equally selected the 2nd category 25,000 – 49,999. Males had selected the 4th category which were more than the Females.

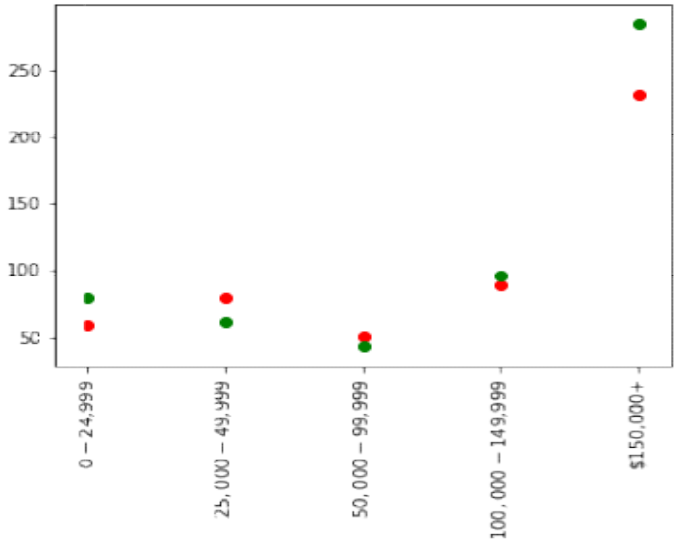

Fig 2.3

In Fig 2.3, Using scatter plot we can find the relation between 2 variables. In the 3rd category 50,000 – 99,999 the Males and Female are closed to each other which means both the population had equal numbers for the 3rd category. Whereas, for the 4th category,

both the variables are distant from each other which indicates that the female population were the most as compared to male population.

2.3 . Brief Analysis

To better understand the behavior of the data, visualization graphs were used. As graph make it easier to identify the distribution of data or how the data within data are related to each other. Exploring on the basis of data distribution or correlation will help better understand the significant factors that can create an impact and which don't. Based on the graphs, there is a relation between the demographics of the people and the characters.