Assignment 1: Data Cleaning and Summarising

Introduction :

Aim of this assignment is to load and analyze the StarWars dataset and perform operations for data cleaning as well as summarising.

Discussion :

Data preparation -

1.1) We have loaded 'StarWars.csv' in the **StarWarsDF** variable and modified it with the appropriate column names. We read the data by **read.csv()** function where we assigned the file name and modified column names vector to names attributes. Encoding was set to **iso-8859-1** just to ensure that no latin character caused the error while reading. There exists a latin character in the first row of the dataset but we have ignored it by making **header = 1.** Finally we have run **head()** function to check whether the data is loaded properly.

1.2) We did check the datatype of all the columns by **StarWarsDF.dtypes**. Datatypes of all the columns were exactly the same as I wanted it to be. For example " Have you seen Star Wars : Episode 1 The Phantom Menace? " is a float.

1.3) By **value_counts()** we can see what are the unique values and their respective frequency, based on which we can spot whether there are typos or not and if yes how many. If typos exist we mask the typos and then modify them with the correct value. After masking we can see that typos have been eliminated.

1.4) Sometimes we can see that the spelling and case of the unique columns are the same but still they have been treated as different due to a whitespace which is invisible to our naked eye. 'Have you seen any of the film?' column also has such value which we did check by value_counts(). We removed the whitespace by **strip()** function. strip() removes the leading and trailing whitespaces.

1.5) We have changed the case of all the values in the dataset to upper case with the help of **upper()** function. To check whether the operation was performed successfully we executed head() function.

1.6) We plotted the barplot of the 'Age' column to check the various categories in age where we could observe that age = 500 was an impossible value. This is the sanity check for our dataset. We eliminated the observation with age=500 by subsetting the dataset.

1.7) There are a lot of columns with missing values in our dataset. Initially I thought of dropping the observations with missing values from the dataset but after executing **StarWarsDF.isnull().sum()** it was observed that 972 was the highest missing value from a single column. Dropping 972 observations from 1185 observations was not feasible according to me hence I just replaced the missing values of the ratings columns with their respective means. Rest all columns had observations with missing values somewhere around 350 hence while dealing with those columns in the analysis I preferred to ignore

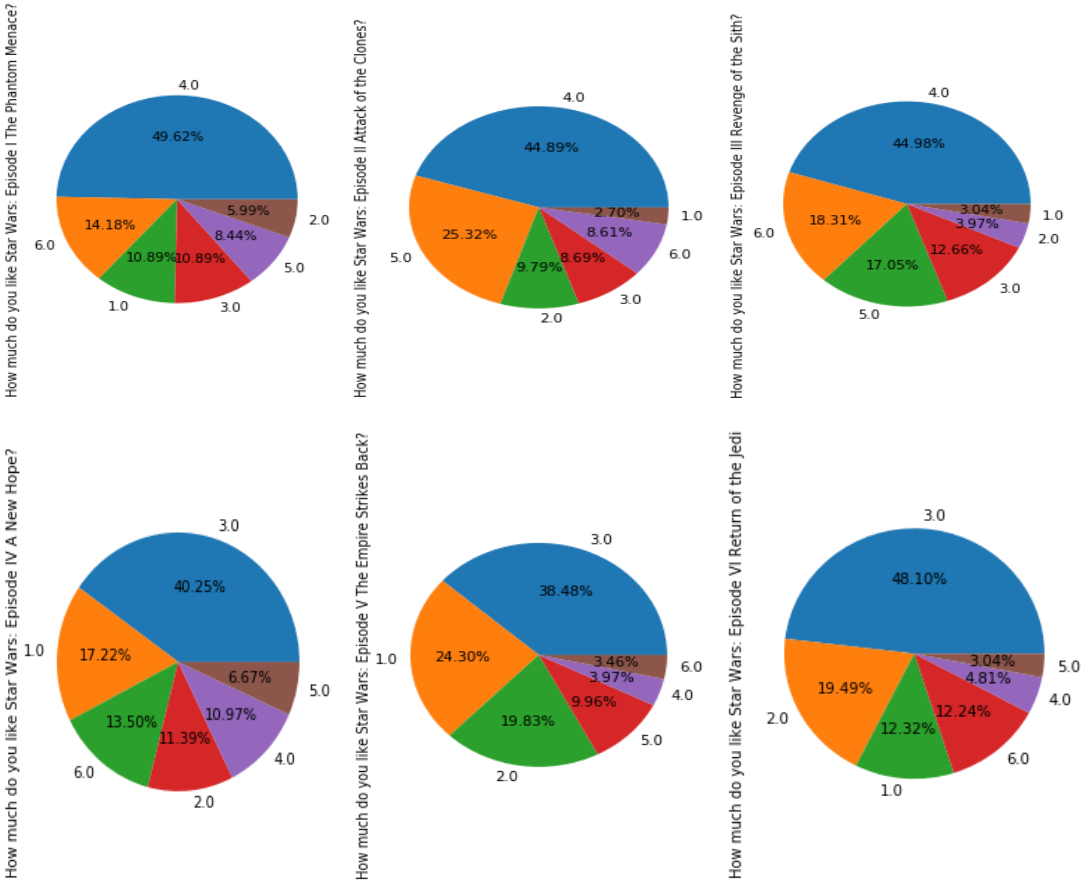
them simply. I have used the **mean()** function to calculate the mean value of the entire column. Calculated mean was then filled in place of missing value by the function **fillna().**

Data Exploration -

2.1)

Subsection 1 = Inorder to analyze how people rated for Star Wars movies we have plotted pie charts of the individual movies. Pie chart shows the percentage of the people rating(rating 1 to rating 6) for respective movies. In addition we have also calculated the count of people giving rating 6 for those movies for our analysis(refer section 2.1.b in assignment).

Subsection 2 =



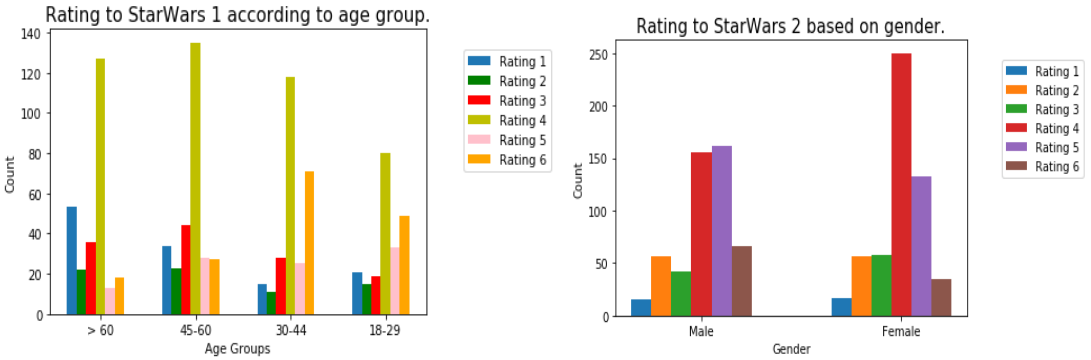
With the help of the above charts we can observe how people like the Star War movies. Let us take the pie chart of movie 1. We can see that 49.62% of people have voted rating 4 for the movie which is the maximum. Percentage of people who gave rating 6 is 14.18%. 10.89% people have given the least rating which is rating 1. In the same way we can gain insights of other movies with the help of pie charts.

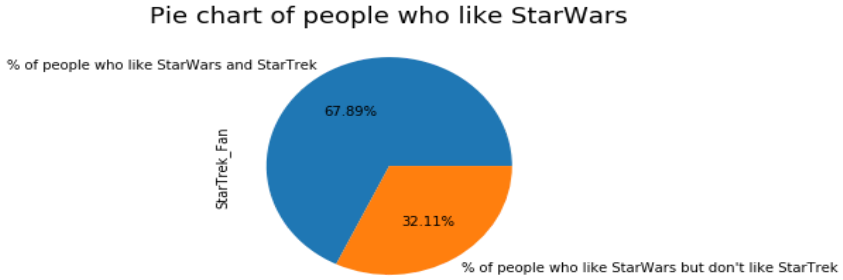
Subsection 3 = Comparing all the above pie charts we can say that Star Wars Episode 3 was the most liked movie with rating 6. Star Wars Episode 5 was rated 1 by 24.03% of people. Hence we can conclude by studying the given charts that Star War Episode 1 is least favourite among all.

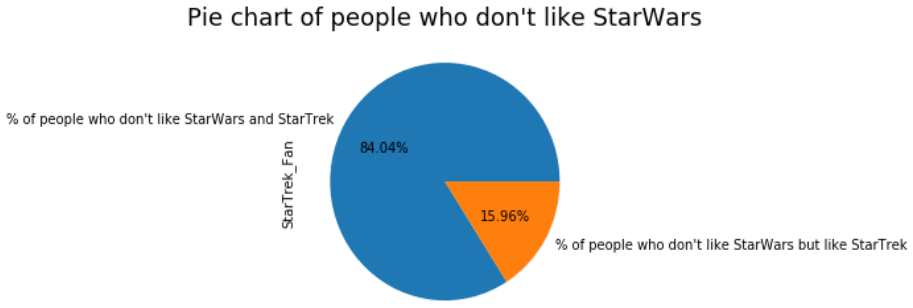
2.2)

Subsection 1 = Inorder to explore the relationship between the columns we have chosen 3 pairs of attributes. ‘How much do you like Star Wars: Episode I The Phantom Menace?’ and ‘Age’ is the first pair, we have done the analysis of this pair by a grouped bar chart. ‘How much do you like Star Wars: Episode II Attack of the Clones?’ and ‘Gender’ is the second group, to obtain relationships we have plotted their grouped bar chart. 'Are you a fan of StarWars?' and 'Are you fan of Star Trek franchise?' is the third pair we have plotted pie charts to see their relationship.

Subsection 2 =







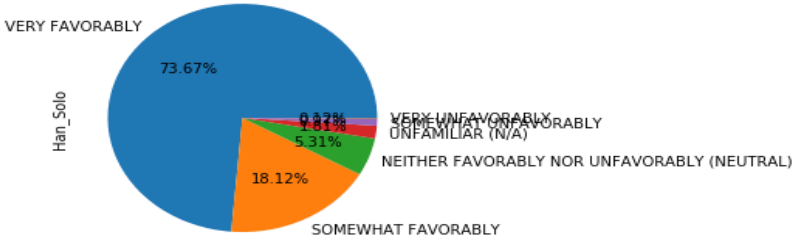
In the first chart we have plotted ratings given to the Star War episode 1 based on the age. Grouped bar chart has categories of age on the x-axis and its frequency on y-axis hence it can be used to see how people vote for the movie in particular age group. In the second chart we have plotted ratings given to Star War episode 2 based on the gender. We can gain insight on how the movie was rated differently gender wise. In the third pair we have plotted two pie charts. First pie chart is of the people who like Star Wars movies and the wedges are divided based on whether people like Star Trek or not. Second pie chart is of people who don't like Star Wars movies and the wedges of the pie chart is divided based on whether they like Star Trek or not.

Subsection 3 = From the first diagram we can observe that all the age groups rated mostly 4 to the movie. Rating 6 was given mostly by age group 30-44 amongst all. From the second chart we can conclude that male have given rating 5 for most of the time and rating 4 being next most voted. Females have voted rating 4 for most of the time. From the third and fourth chart we can say that 67.89% of people like Star Wars and Star Trek both. It can be seen that 84.04% of people who like Star Wars don't like Star Trek as well.

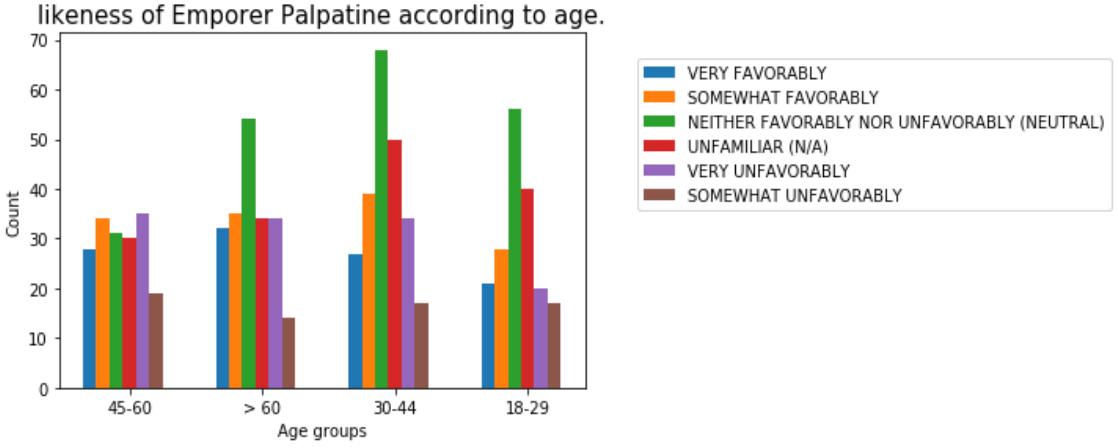
2.3)

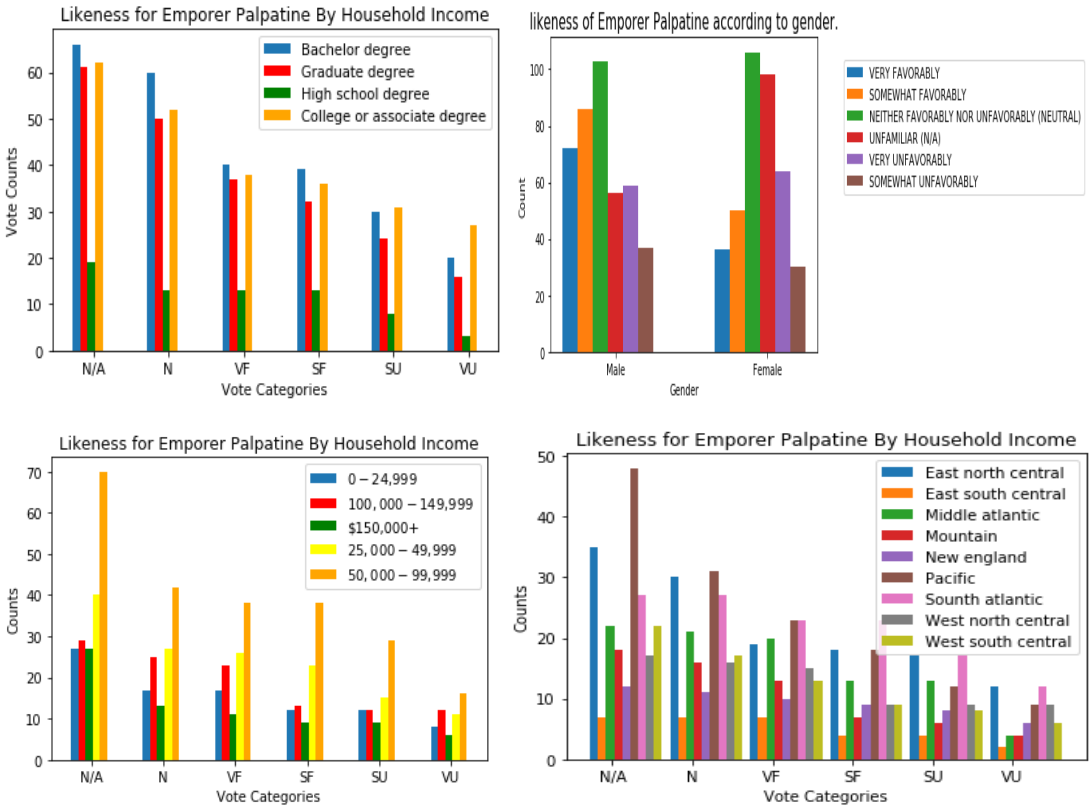
Subsection 1 = Inorder to explore whether there are relationships between people's demographics and their attitude towards Star Wars characters we have plotted pie charts of each character. Comparing these pie charts we can conclude which character is mostly liked by the people.

Subsection 2 =



Gender	Age	Income	Education	Location	Han_Solo
FEMALE	18-29	$0 - $24,999	BACHELOR DEGREE	MIDDLE ATLANTIC	SOMEWHAT FAVORABLY
1					
				MOUNTAIN	VERY FAVORABLY
2					
				NEW ENGLAND	VERY FAVORABLY
3					
					SOMEWHAT FAVORABLY
1					
				SOUTH ATLANTIC	VERY FAVORABLY
1					
				WEST NORTH CENTRAL	UNFAMILIAR (N/A)
1					
			GRADUATE DEGREE	EAST NORTH CENTRAL	SOMEWHAT FAVORABLY
1					
					UNFAMILIAR (N/A)
1					
				PACIFIC	VERY FAVORABLY
1					





From the pie chart we can understand how much people like the character. I Have plotted pie charts for all the characters but include only Han Solo in the report. We can say that Han Solo was rated as 'Very Favourable' by 73.6% of the voters. Second snippet is of the output produced by the code in section 2.3.b(refer assignment), in this code we have grouped all the demographics and produced the count of people in that group. Rest all the grouped bar plots are produced for character Emperor Palpatine with respect to single demographics.

Subsection 3 = Comparing all the pie charts for the characters we can conclude that Han Solo is the most favourable character and Jar Jar Binks is most unfavourable from all. From the snippet we can say that there are 3 females in Age group 18-29 having income $0-$24,999 studying bachelor living in New England who voted Han Solo as very favourably. From chart 3 we can say that Emporer Palpatine was voted very favourably mostly by age group >60. From chart 4 we can say that Emporer Palpatine was voted very favourably mostly by people studying bachelors degree. From chart 5 we can say that male were the most to vote 'very favourably' for Emporer Palpatine based on gender. From chart 6 we can conclude that people with income 50000-99999 have voted 'very favourably' for most times. From chart 7 we can say that people living in Pacific and South Atlantic have voted for 'very favourably' most times.

References

(pandas - Python Data Analysis Library, 2020)
(Matplotlib: Python plotting — Matplotlib 3.2.1 documentation, 2020)
Module Notes (Lecture