**Data Preparation**

3.1.2: Upon loading the corresponding dataset successfully we will have to ensure that the loaded data is equivalent to the data in the source (CSV) file. This can be done using the .dtypes function, which lists the data type of each variable.

3.1.3: Here we should ensure that the data does not have any typos. Using the value\_counts() function helps us to get the values of each value present in a particular column. This in turn helps us to find out any typos if present. We can correct these typos present by using masks or replace function. Mask function fixes the typos if any and helps in making the data set clean.

3.1.4: Another important step in data cleaning is to examine if there are any instances of extra whitespaces in the data and eventually removing them. This task can be achieved with the help of a function called .strip() function. This function removes extra white spaces if any from both ends of a string, that is from both trailing and leading ends.

3.1.5: This tasks is used to cast all text data to upper-case by using an appropriate function. With the help of .upper() function this can easily be achieved. The final dataset upon checking will have all test data converted in upper case.

3.1.6: Data also consists of some impossible values sometimes making the data unsuitable for analysis. In order to get rid of these values we can perform a sanity check on the data set provided. This step helps to get rid of these impossible values, for instance the age in this data set is given as 500 which is not possible. We can fix this with the help of sanity checks.

3.1.7 The data also needs to be checked for any missing values if any. We have to ensure that the missing values are replaces with some other value. This can be done in many ways, however, in this case I have used the .fillna() function to fill up the missing values with the value of my choice.

**Data Exploration**

3.2.1 The given task expects us to survey about how people rank the Star Wars films in order of preference with 1 being their favorite film in the franchise and 6 being their least favorite film. This can be easily obtained with the help of a graph. In this case I have taken a histogram to understand the ranking given by people for different episodes.

Rank for episode 1

Rank for episode 2

Rank for episode 3

Rank for episode 4

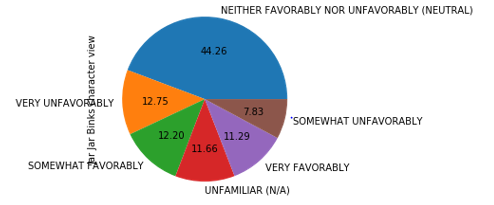
Rank for episode 5

Rank for episode 6

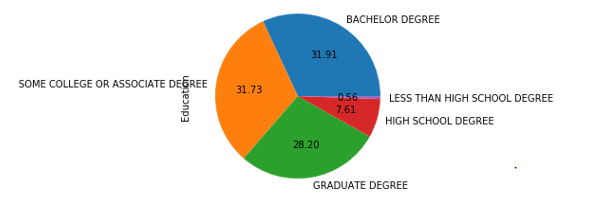
For instance, StarWars['Rank for episode 1'].plot(kind='hist',bins=20) is used to determine the ranks given by people in the range 1-6 for Episode 1, which is Star Wars: Episode I The Phantom Menace. Likewise there are 5 more histograms for 5 other movies where people have ranked the movies in the range 1-6. From the Histograms we see that Episode 5 was ranked number 1 by majority of the people and on the other hand Episode 2 was not enjoyed by many.

3.2.2

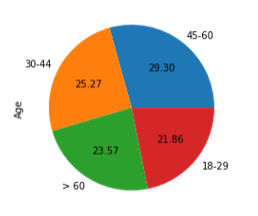
A] From the above Pie-Chart I have tried to analyze the liking of females towards the character Jar Jar Binks. We can see that 44.26% of the females have a neutral view about this particular character, that is majority of them have neither favorable nor unfavorable view about this character. However, we can see that this character was not very popular amongst the females as only 11.29% of them liked this character Jar Jar Binks.



B] This pie chart gives us the information that most of people who have seen this series are educated or they hold any type of degree. Only 0.56% of the people who are fans of Star Wars movie are those who have less than high school degree.

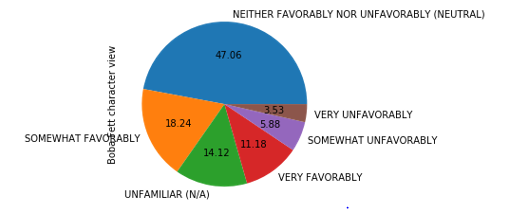


C] The above pie chart is used to examine the percentages of diverse age groups of people who have seen any of the 6 films from the Star Wars series. From the chart we see that this movie was widely seen by all age groups of people, as the pie chart is almost equally distributed. However, we can see that the Star Wars franchise was most popular amongst the age group of 45-60 years.



3.2.3

From this pie chart we can interpret Boba Fett character’s view by the people of South Atlantic. We can clearly see that 47.06% of the people have a neutral view about this particular character. Whereas 18.24% of the people from South Atlantic liked Boba Fett character somewhat favorably, and 14.12% of them were unfamiliar with this character, the reason for this could be, either they are not aware of the character or they have not seen the Star Wars franchise. About 11.18% of them like this character very favorably. With 3.53% of them choosing very unfavorably we can say that only a small percentage of the people from this region did not like this character at all.



Assumption: Here I have replaced the missing values with neither favorably nor unfavorably (neutral), as I have made an assumption that the people who have not given any response have a neutral opinion about this character.