Data Preparation

1. Check data types

Begin with the reading of CSV file that is encoded and observing the couple of rows of dataset that look alike dirtier and need some cleaning for further analysis. This type of data is difficult to represent in columnar format. Most of the columns are unnamed and rows are without respondent id.

2. Extra-whitespaces

The data with extra whitespaces has large damage for the analysis so I checked them every column, the first extra-whitespace found in question “Have you seen any of the 6 films in the Star Wars franchise?” There is an extra whitespace behind “Yes ”, so I use the function“str.strip()” to delete the whitespace.

3. Upper/Lower-case

There were some data changed the upper/lower case, in the question”Gender”, the data “Female” “FEMALE” are the same meaning, so in order to get the right data,I used the “str.upper()”, changed them to “FEMALE” and “MALE”.

4. Typos

There are couple of columns has the wrong data, i.e in the question “Do you consider yourself to be a fan of the Star Wars film franchise?” there had answer “Noo” and “Yess”, so I use the function “replace(“A”,”B”)” to get the right answer “Yes” and “No”.

5. Missing value

For the missing value I have different processing method for different situation: normally missing value will use an appropriate function to replace them with one of the following values: a fixed value,the column,wise median value,the column-wise mean value or ignoring all observations containing missing values.In the dataset there are many column is string and named “unnamed”, so for the string column I decide to delete all missing value because it’s hard to replace the value, ignoring them are the best choose.

6. Sanity checks

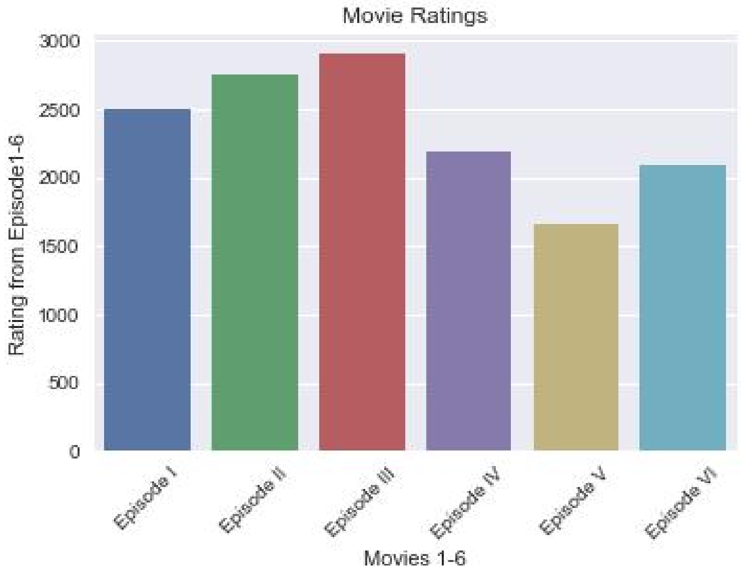
There are a wrong answer in the column “Age”, there is a “500” in the age, nobody’s age can be 500, however it’s hard to say what is the correct age, so I ignored it because it only 1 data in 1186 data.

Also for some columns have string types, because the main values they contain are Yes and No. We can make the data a bit easier to analyze down the road by converting each column to a Boolean having only the values True, False, and NaN. Booleans are easier to work with because we can select the rows that are True or False without having to do a string comparison and this process is known as Mapping.

Data Exploration

1. Explore a survey question

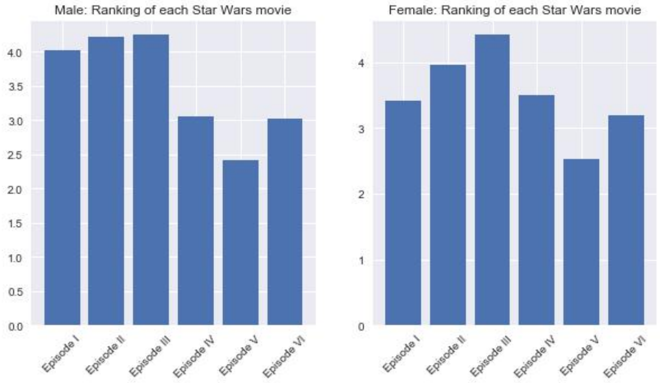
Highest Ranked Movie:



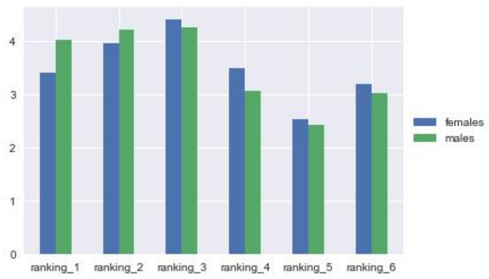
From the above Bar chart we can see that Star Wars Episode 5 has the highest rating (Lower Rating is better as mentioned in Dataset.Column 10 in dataset contains the following string: 'Please rank the Star Wars films in order of preference with 1 being your favourite film in the franchise and 6 being your least favourite film. So columns with lower ranking values are considered better by the survey respondents.) From the chart, it looks like the movies(4-5-6) have higher rankings than the others.

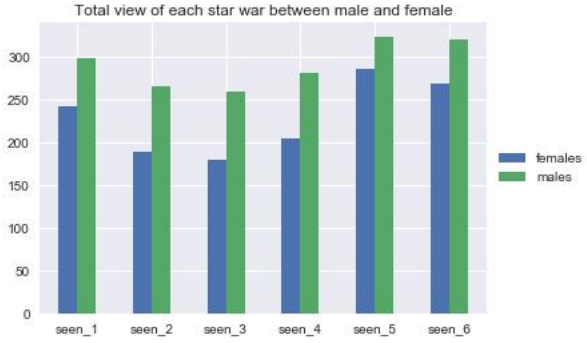
I mapped the String to Boolean and rename the column 4-9 as “seen_ ” and column9-15 as “ranking_ ” and calculate the sum of column 9-15.

2. Relationships between column

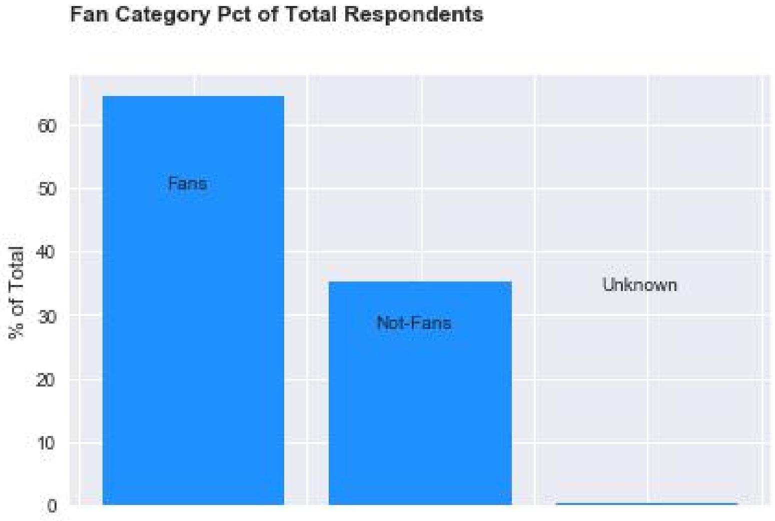


As the graph shows gender has different attitude on different Episode, in the survey Episode 1 is the second favorite movie and it has 0.5 higher than male’s. Also for episode 4 male prefer female.

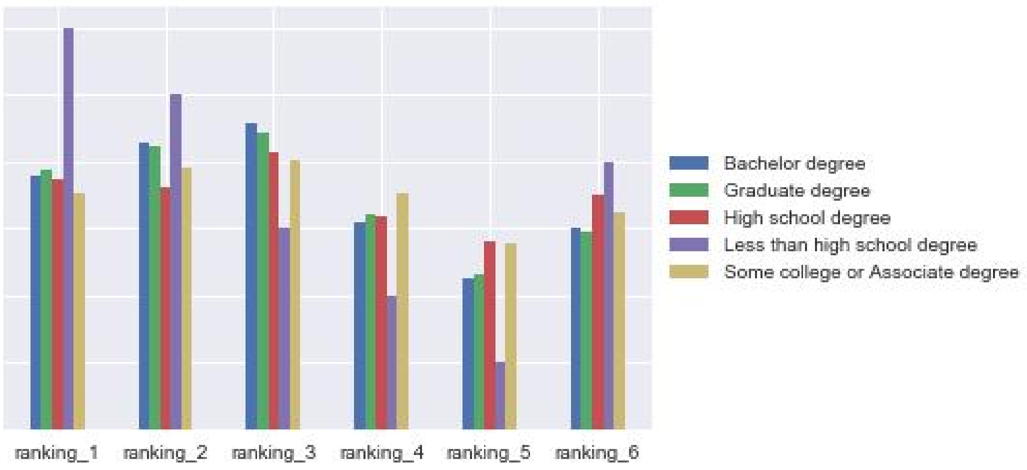


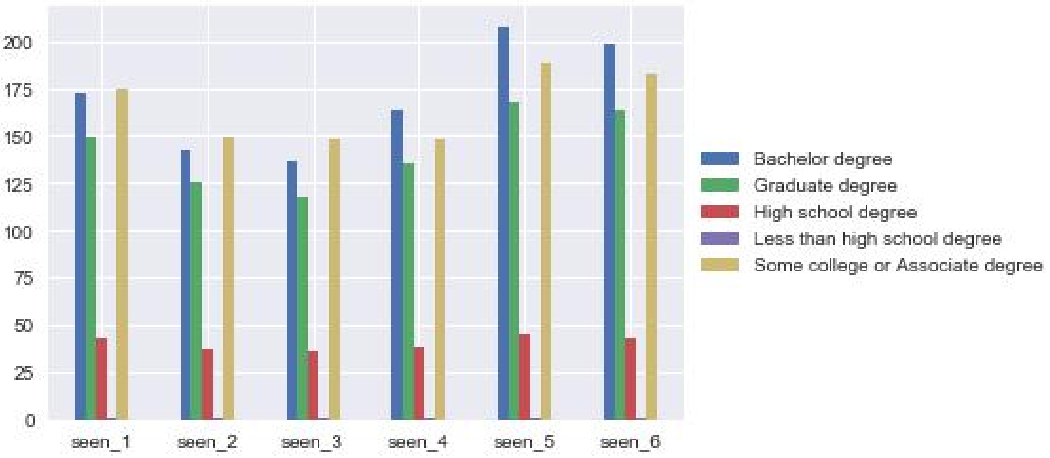


As compared to females more number of males have watched the movies, and in case of reviews given by females are more as compared to male by this we can say that most of the females who have watched the movie have given the review

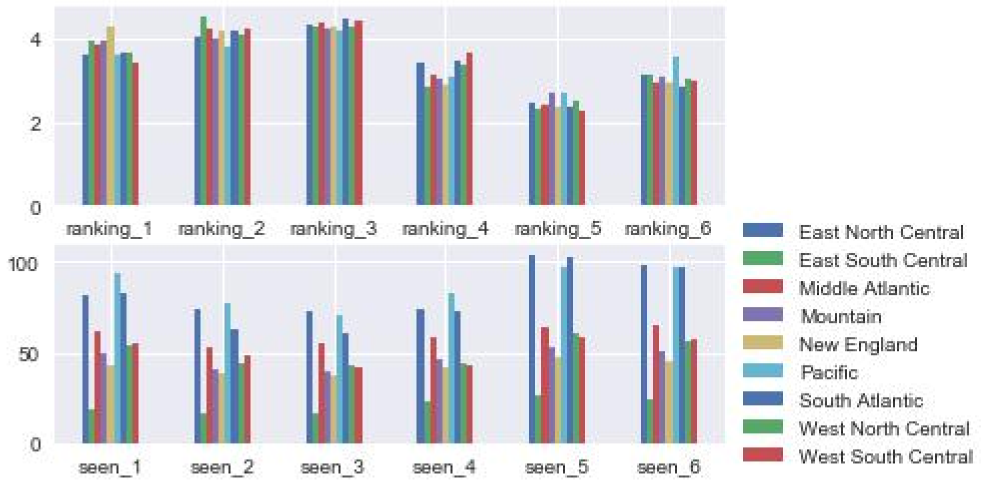


We can see that more than 50% respondents are fans and About 30% of the respondents who did not specify whether they were fans or not





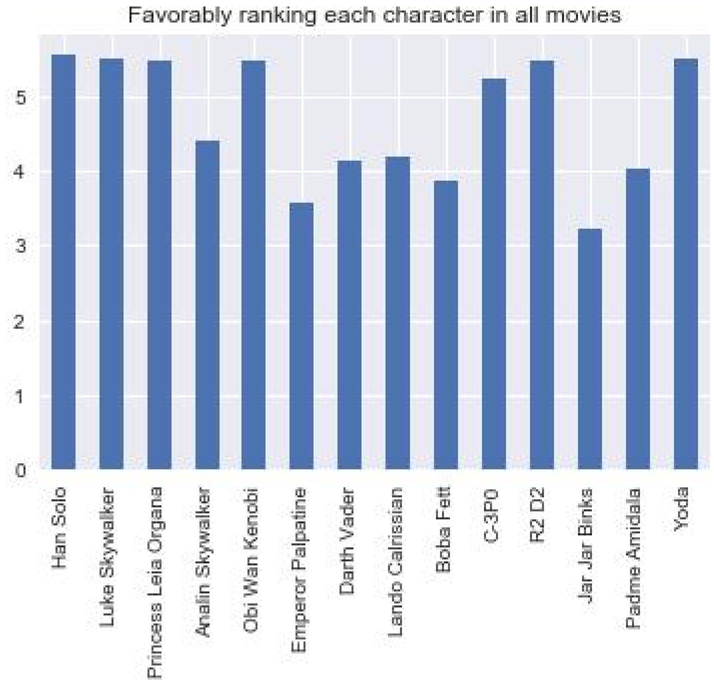
From this categorical Bar Chart we can see People watching star wars movies have educational background related to Bachelor Degree,Graduate degree,Some college or Associate degree. Most people lie in the category of some college and associate degree.



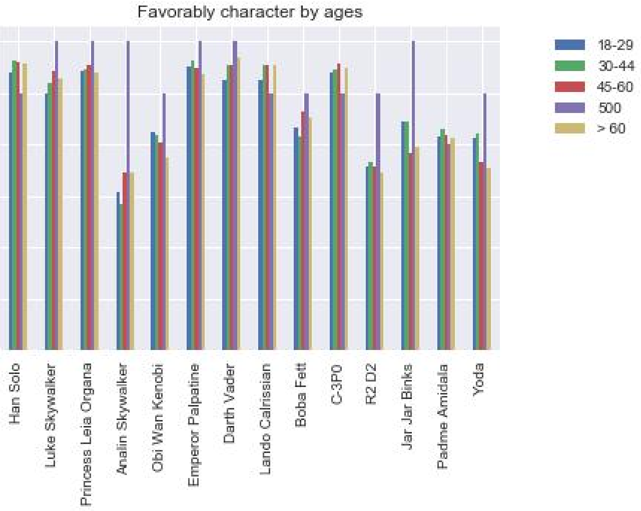
Highest star wars movie watchers belong are from East North Central, Pacific, South Atlantic. The least star movie watchers are from East South Central.

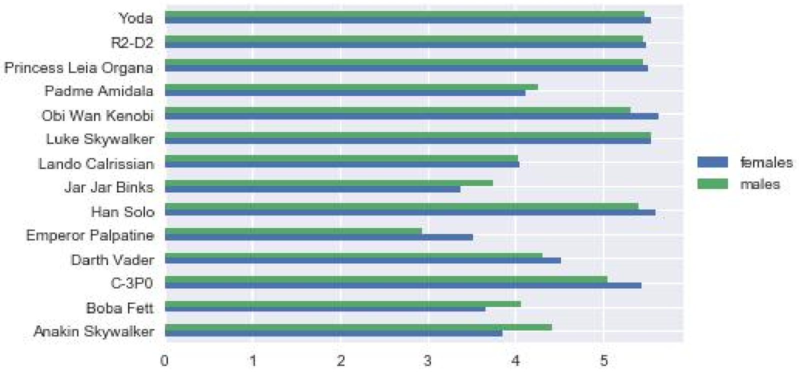
2.3 Explore a specific relationship

As we can see form the data for the related columns are strings. I choose to use stacked chart to show the attitude of people. It can show different categorical groups to us and also showing some statistic information at same time.



The data is grouped by using different filter. Chart shows the percentage of designed groups in each attitude. It is easy to see the composition ratio of different data.





Charts were shown in total number for each attitude.And I found that is meaningless to show the total number even it is divided into groups. Because the number for each attitude group varies. I showed the different favorably character by ages. We can find that most stacked bars in each chart are "colourful", which shows that most attitude type consist of all groups in one demographic feature. In most case, people's gender, age, income, education and location do not affected their attitude to Star War characters and movies deeply.