Slide 2:

Being a fan of starWars, the data set which was given was very interesting to me. This also refreshed different characters to me and was having extra fun Pre- Processing the data. The steps taken for doing this was, Data Retrievin, Checking Data Types, Upper/ Lower Cases, Extra White spaces, Typos, Sanity Checks and Missing Values.

Slide 3:

Data Retrieving:

After importing the dataset, we can notice that the column headings are multi-indexed and the best way to fix this is to remove unwanted columns. Many columns were unnamed at index zero but had a name in index one. This had to be removed and had to be made uniform. To do this rename function is used to rename all the unwanted columns and the extra column heading were dropped. This made the data more easily readable.

Check Data Types:

Using the dypes function, we can find the data types for all the columns. It is noted that all the columns are in object data type. Only the respondent ID was in float type.

Upper/Lower case

The upper case function is used to make all the characters into upper case so that there is uniformity with the data to be manipulated. This becomes easy to rectify typos and having to replace lesser values.

Extra-whitespaces

These are hard to detect in the data set and need to be dealt with initially. This is done using the . strip function which instantaneously removes all the white spaces in the dataset.

Typos

Typos were easily found by running a for loop in the dataset and retrieving the value counts of each of the columns. This way I could find and tackle all the missing values using the replace function.

Sanity Checks

To make Sanity Checks, value\_counts was used under each column. This way I could find some anomalies in the data set. It was found that one value in the age column had an obnoxious value of age 500. This was considered as a typo and was fixed using the replace function.

Missing Values

Missing Values were dealt a little differently than the traditional way. Columns which had whether if people had watched the movie was indicated by ‘the movie name’ and not watched was denoted as a missing value. The missing value however is vital information because it really is not missing value. So, movie watched was replaced with True while missing value was replaced with False. This way we recovered a lot of missing values. Moving on to the columns which had ratings, people who have not watched the movie also had given a rating. This gives an biased result in the end. To make it fair, only such information was dropped. Rest of the missing values were dealt with for each analysis / visualization.

Slide 4

Read what is there in the ppt

Slide 5

Here we have found which movies have ranked how much. Firstly all the ranks had to be converted int values. After which mean of each of the columns were taken we can see which movie had been rated the best. A bar graph is then plotted and we can see that movie 5 is the best rated followed by Movie 4. The worst rated movie is Movie 3

Slide 6

Here we are comparing Fans of starwars and their age. Firstly creating a copy of the dataset and sub setting it into only essential columns. Taking value counts of these 2 columns and grouping by with each other, this is ready to be plotted against each other. But before that normalization of the data had to be done. After all this, putting the group-by data frame based age into a stacked bar graph we can see We can see that the age group 18-29 had the highest fans. Followed by 45-60, 30-44 and then finally >60 age group.