Name: Nikhil H

DATA PREPARATION:

Task 1.1 : Data Retrieving

Load the csv data using the pandas read_csv() function with specific parameters.

Use comma for separator parameter sep(,)

Header is set to none as we are defining the column names.

Skip the first two rows using the skip parameter(skiprows=2) of the dataset as it contains a question and character names, rather pass a number of user defined names to each of the attributes in the names array.

Task 1.2: Check data type

Using the dtypes function to check the data type of each attribute.

The are three different datatypes for the attributes int, float and object.

The data types of Rank_E1 to Rank_E6 attributes is displayed as float as it contains NaN values.

All the other attribute data types are correct.

Task 1.3: Typos

The common typos are (Yess) for Yes, (Noo) for No, (F) for Female and (M) for Male

Correcting these typos by using mask functions as shown below:

starwars.loc[:] = starwars.mask(starwars=='Yess','Yes') starwars.loc[:]

= starwars.mask(starwars=='Noo','No')

There are other common typos which are in lowercase such as (no, yes, female,male), which will be handled by Task 1.5 which casts the data to Upper case.

Task 1.4: Extra-whitespaces

A lambda function is used to strip the white spaces, based on the condition that the strip function will be applied if the data type of the attribute is 'object'(as it contains strings), else the strip function won't be applied.

There is a white space in one of the Fan_Startrek attribute which is displayed as 'no '. After the strip function is applied it gets converted to 'no' .

TASK 1.5: UPPER CASE

A lambda function is used to cast the text data to upper case, based on the condition that the upper() function will be applied if the data type of the attribute is 'object'(as it contains strings), else the upper() function won't be applied.

Task 1.6: Sanity checks

To find the unexpected values check for the unique value counts of each attribute. All the attributes have correct unique value counts except for Age attribute which displays 5 unique values whereas it only has 4. Check the unique values of age attribute. It displays an unexpected value 500.

Task 1.7: Missing values

Check the number of missing values in each of the attributes by using the function (starwars.isnull().sum()).

For a person who has watched the star wars movies the values in the dataset are displayed with the movie name, need to replace the movie name with 'YES'(Assuming person has watched the movie).

For a person who has not watched star movies it is displayed as blank, need to replace the blank values with 'No'(Assuming the person has not watched the movie).

The other blank values should be filled with Nan.

Verify using the isnull() function that the number of missing values for the movie attributes should be set to zero as they replaced with 'NO'.

TASK 2: DATA EXPLORATION

Task 2.1: Explore a survey question

Plot a pie chart for each of the six starwars movies.

Each of the pie chart depicts the percentage of each ranking of the movies from 1 to 6.

From the analysis of the pie chart of all the movies we can conclude that the Star Wars: Episode V The Empire Strikes Back is the favorite film as it is Ranked 1 with 34.57%(Figure 1) and Star Wars: Episode III Revenge of the Sith is the least favorite film as it is ranked 6 with 25.99%(Figure 2) by majority of respondents.

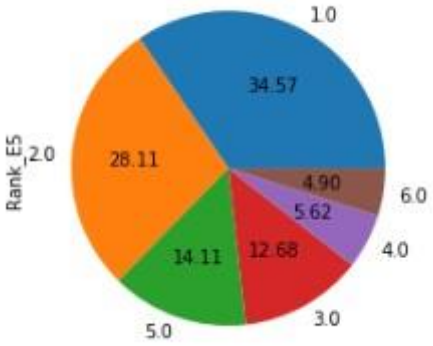


Figure 1

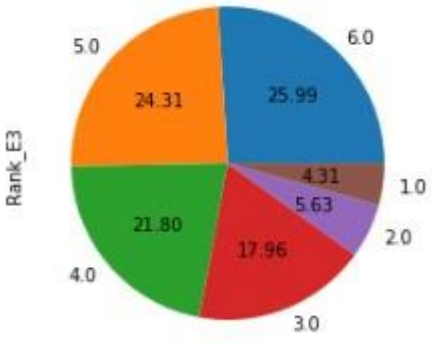
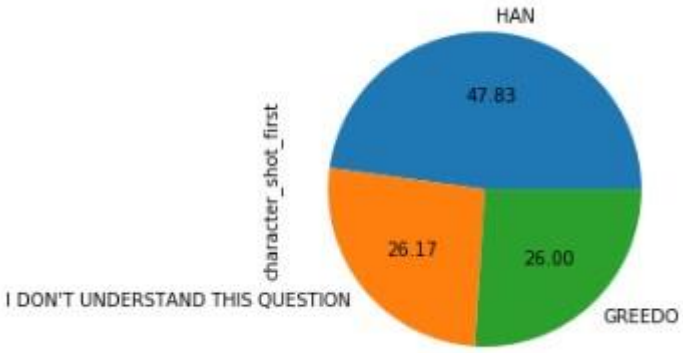


Figure 2

TASK 2.2: Explore the relationships between columns; at least 3 visualisations with plausible hypothesis

Task 2.2.1) Relationship between Star Wars: Episode IV A New Hope and Which character shot first attribute:

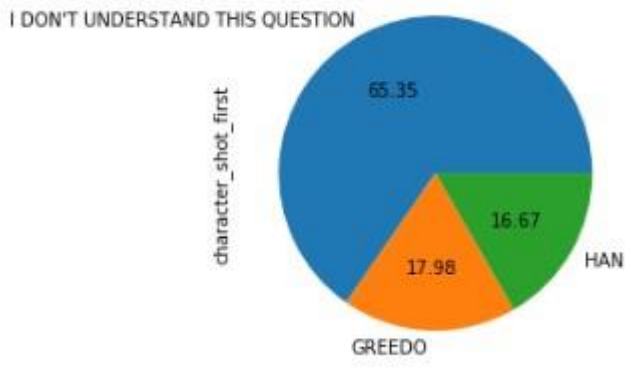
Hypothesis: The 'Who shot first' attribute refers to a controversial change made to a scene in Star Wars: Episode IV A New Hope. In the older version of Episode IV HanSolo shot first whereas in the new version of Episode IV Greedo shot first. So majority respondents who have watched starwars Episode IV will be able to answer the question who shot first. Majority of the respondents who have not watched Episode IV won't be able to understand the question.



Watched EP IV A New Hope

Figure 3

Analysis of Figure 3: The above pie chart depicts the count for each of the values in the character shot first attribute for the respondents who have watched the EP IV. As displayed in the Figure 3 majority of the respondents were able to answer the question and only 26.17% said 'I DON'T UNDERSTAND THE QUESTION'.



Not Watched EP IV A New Hope

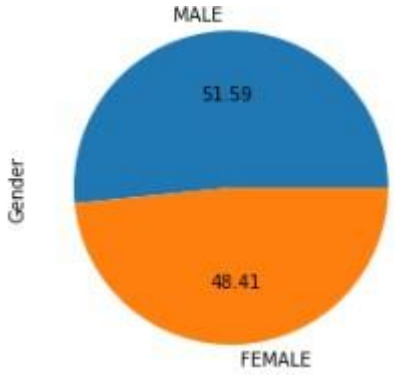
Figure 4

Analysis of Figure 4: The above pie chart depicts the count for each of the values in the character shot first attribute for the respondents who have not watched the EP IV. As displayed in the (Figure 4) 65.35% respondents said 'I DON'T UNDERSTAND THE QUESTION'.

Conclusion: According to the above hypothesis and the data summarised from the plots we can conclude that majority of respondents who have not watched Episode IV said 'I DON'T UNDERSTAND THE QUESTION' , as they don't know about the controversial change in the Episode which relates to the question (Who shot first?).

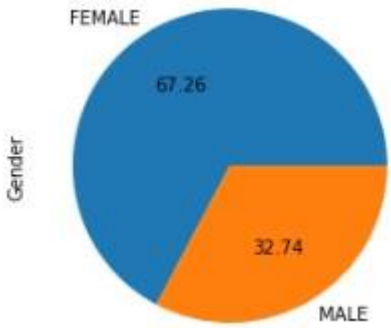
Task 2.2.2) Relationship between the attributes Have you seen any of the 6 films in the Star Wars franchise? With Gender

Hypothesis: To check the percentage of male and female respondents with respect to whether they have seen or not seen any of the star wars movies.



Watched_starwars

Figure 5



Not_Watched_starwars

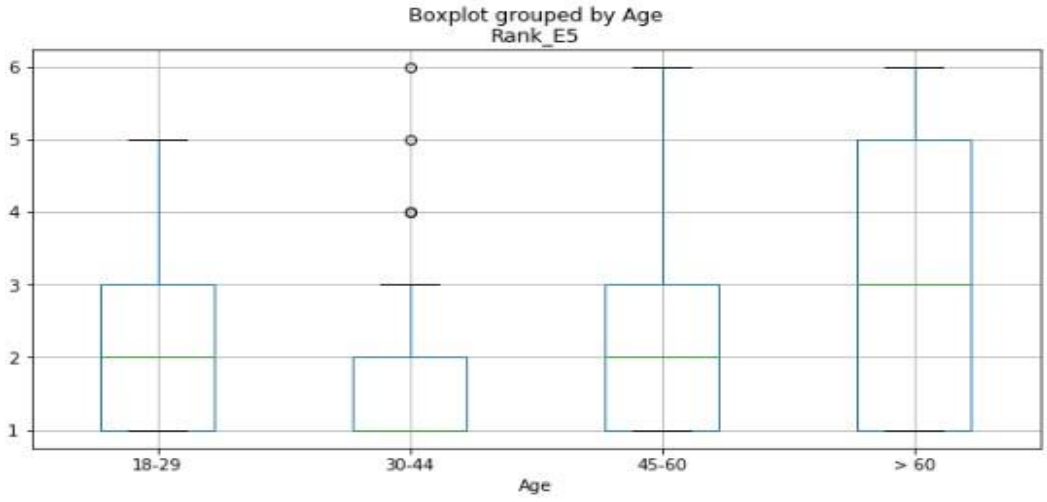
Figure 6

Analysis of the Figure 5 and Figure 6: *The Figure 5 depicts the number of respondents who have watched any of the star wars movies and Figure 6 depicts the number of respondents who have not watched the star wars movies.*

Conclusion: According to the data summarised from the above plots we can conclude that majority of Females (67.26%) have not watched any star wars movies as compared to the male respondents.

Task2.2.3) Relationship between the best rated favourite film(Star Wars: Episode V The Empire Strikes Back) and Age:

Hypothesis: To recognise the pattern as to which Age group has rated the Star Wars: Episode V The Empire Strikes Back as favorite on a scale of 1(favorite) to 6(least favorite).



Analysis of Figure 7: The Figure 7 displays the Box plot grouped by age group(X axis) and the y axis displays the Rank of the movie (Star Wars: Episode V The Empire Strikes Back)) from 1 to 6. The respondents in the age group > 60 have ranked it from 1 to 6, with majority of the repsondents rating it from 1 to 5 as displayed in the IQR region of boxplot. The Respondents in the age group 3044 have ranked it from 1 to 3 with majority of them ranking the movie from 1 to 2 as displayed in the IQR region of boxplot. Very less number of respondents in the age group of 30-44 have ranked it as 4,5,6 and hence it is displayed as outliers.

Conclusion: From the information acquired from the box plot we can conclude that the respondents in the Age Group 30-44 have ranked the movie more favorite(scale of 1 to 2) compared to the respondents in the other age groups.

Task 2.3: Explore whether there are relationship between people's demographics (Gender, Age, Household Income, Education, Location) and their attitude to Start War characters.

Task 2.3.1) Relationship between Gender and star wars character Emperor Palpatine:

Emperor Palpatine is the ranked as Unfamiliar character by majority of the respondents. As we already know from previous data Exploration results that compared to male's most of the Female's are not a Fan of starwars Franchise. So based on this majority of females should be unfamiliar with Emperor Palpatine .

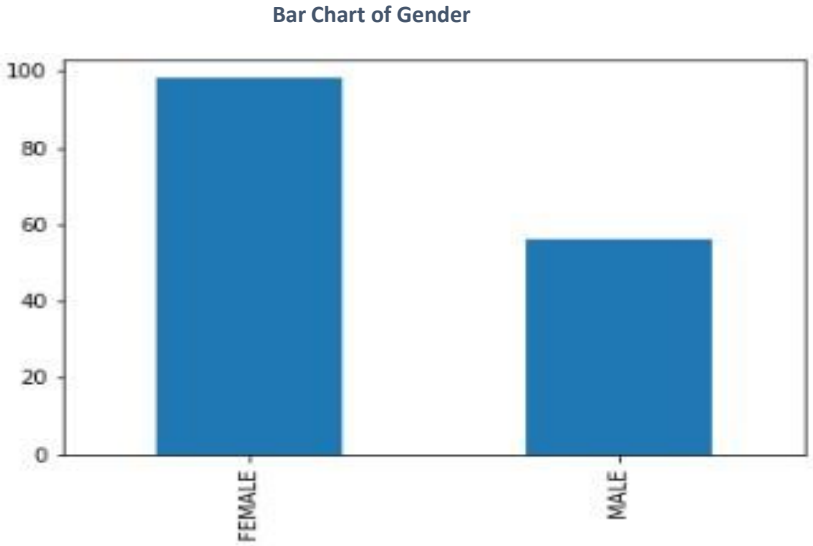


Figure 8

Analysis of Figure 8: The Figure 8 displays the Bar chart for the male and female respondents with respect to the frequency of Emperor Palpatine Rating (== 'UNFAMILIAR N/A'). The number of female respondents who have ranked Emperor Palpatine as Unfamiliar is approximately equal to 98 which.

Conclusion : From the information gathered from the Bar chart we can conclude that majority of Female respondents are Unfamiliar with Emperor Palpatine as the most of the Female respondents are not a Fan of Starwars.

Task 2.3.2) Relationship between Age group and starwars character yoda:

Yoda is rated as the second most Familiar character. So need to analyse the relationship by checking which age group has ranked yoda as the most Familiar.

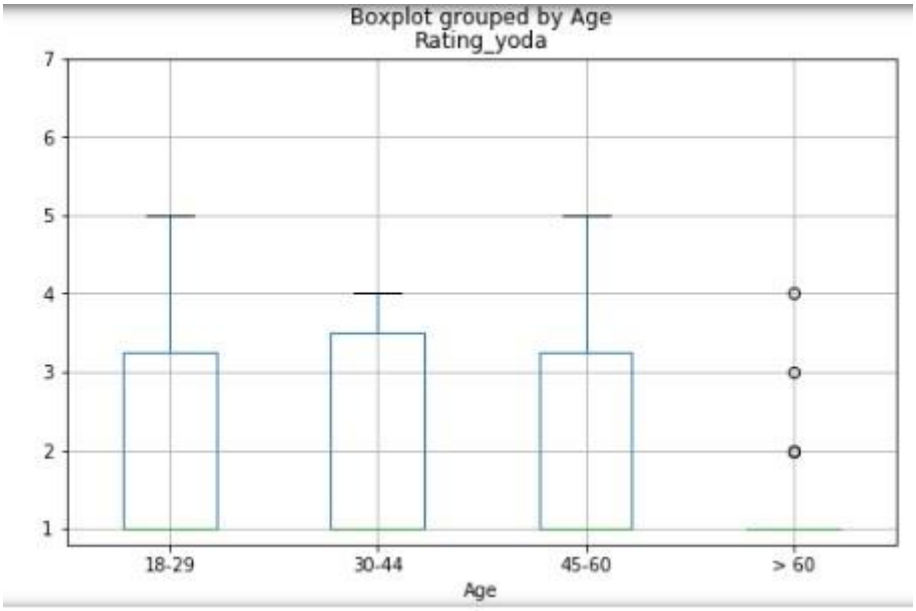


Figure 9

Analysis of Figure 9: The above Figure 9 displays a Bar plot grouped by age with respect to rating character yoda. The Y axis displays the Rank for Familiarity of character yoda from 1(Very favorably) to 6(Unfamiliar N/A) and the x-axis displays the age Groups. The majority of respondents in the age group (> 60) have ranked yoda as Very Favorably(1) and very less respondents in the age group have ranked yoda as 2,3,4 which is displayed as outliers in the Box plot.

Conclusion: From the summarised data from the boxplot we can conclude that respondents in the age group > 60 consider Yoda to be Very Familiar compared to the respondents in other age groups

REFERENCES: pandas documentation — pandas 1.0.3 documentation

pandas documentation — pandas 1.0.3 documentation (2020). Available at: https://pandas.pydata.org/docs/

Overview — Matplotlib 3.2.1 documentation

Overview — Matplotlib 3.2.1 documentation (2020). Available at: https://matplotlib.org/3.2.1/contents.html

7.1. string — Common string operations — Python 2.7.18rc1 documentation

7.1. string — Common string operations — Python 2.7.18rc1 documentation (2020)

Pandas : Get unique values in columns of a Dataframe in Python – thispointer.com

Pandas : Get unique values in columns of a Dataframe in Python – thispointer.com