# Practical Data Science with Python - Assignment 1

**Yashwin Sudarshan**

RMIT

# Data Preparation

The beginning of the data preparation stage included loading the StarWars.csv file into the Jupiter notebook and converting it to a Pandas data frame. Then the names of the columns were renamed, as many respondents' answers spanned several columns after the name of the column containing the question they were asked. For example, in the original StarWars.csv file was a column asking respondents to select all the Star Wars movies they have seen, and the respondents' answers to this question could be found in columns immediately after this column, thus, this column and the five columns after were renamed to 'Seen SWE1' (seen Star Wars Episode 1), 'SWE2' etc. The Pandas dataset was then checked to see if it was equivalent in shape and contained the same types as the StarWars.csv dataset. Data cleaning was then performed on for each feature in the dataset. Issues such as extra whitespaces, typos, and missing values were sought after so that they might be dealt with to aid data exploration. Checks were performed for all of these issues, as well as impossible values, at the same time for each feature of the data set (instead of checking for each issue separately), which means that sanity checks for impossible values had been conducted automatically as well.

## Typos

The method that was utilised to search for typos in each attribute in the dataset was the value_counts() method, which returned the observations in the column and their quantity. As the names of the observations were in the output, it can be seen if there is an extra observation that should not be present in the output, that is present only because it is spelt differently to the actual way it is supposed to be spelt. For example, when checking for typos for the 'Do you consider yourself to be a fan of the Star Wars film franchise?' attribute, the output should have only contained 'Yes' and 'No', but it contained 'Noo' and 'Yess' in addition to the proper values. In order to fix the typos, the str.replace() method was used on the dataset for each particular column which had typos, to replace the typo with the correct spelling.

## Extra Whitespaces

The value_counts() method was also utilised to search for extra whitespaces in each column of the dataset. Extra whitespace in an observation, even if it had to correct spelling, caused that particular value to be in the output as a unique observation. For example, when checking for extra whitespaces in the 'Have you seen any of the 6 films in the Star Wars franchise?' column, Two 'Yes' values were in the output, which meant that one of them had extra whitespace at the end of the value. In order to fix extra whitespaces, the function str.stip() was called on the dataset column that had extra whitespaces, which removed all the extra whitespace from the values with it.

## Missing Values

The value_counts() method was again used to check for the presence of missing values for each attribute of the dataset. If the total quantity of each observation in the output was not equal to 1186 (this is the number of rows that the dataset contains), then this would confirm the existence of missing values. In order to fix this issue, in most of the columns with missing values, the string 'UNANSWERED' was assigned using the fillna() function, to each instance of a missing value. In the columns labelled 'Seen SWE1', 'Seen

SWE2' etc., the string 'NOT SEEN' was assigned to the missing observations, and in the columns for preference ranking of the Star Wars movies, the missing values were not dealt with, as a large number of rows contained missing numeric values, and it was thought that filling a large number of missing numeric values with a mean or median value would lead to a more inaccurate modelling of preference ranking data; if there were only a few missing numeric values, then the impact of using a mean or median value would not be as large.

### Sanity Checks for Impossible Values

Again, the value_counts() function was used to check for impossible values for each feature in the Star Wars Pandas dataset. Since the output was all the unique different observation values, it was easy to see if there are any observations that simply should not exist. For example, when calling the value_counts() method on the 'Age' column in the dataset, there was one observation of age '500', which is clearly an impossible value. To fix this, the function str.replace() was used to replace '500' with the string 'UNANSWERED', as it was not desired that the entire row be dropped because of this impossible value, as the row would contain other important information contained in other columns of the dataset. The only impossible value dealt with was the impossible 'Age' value previously mentioned; the use of the value_counts() method to check for other errors in the dataset showed that there were no other impossible values in other columns of the dataset, besides extra whitespace and typos.

## Data Exploration

### Task 2.1

Exploration of the preference rating allocated to each Star Wars movie by the respondents was conducted. Firstly, it was desired to find out which Star Wars movie was favoured by the respondents, and where the rest of the other Star Wars movies rank in terms of favouritism, as the attitudes towards the movies was of great interest. In order to perform this exploration, a subset of the dataset was obtained, showing only the preference ranking allocations for each Star Wars movie. The large number of missing values, as mentioned earlier, were then removed from this new subset dataset. A new dataset with one column was then created manually, which contained the number of 1st rankings by the respondents, for each Star Wars movie. For instance, Star Wars episode 1 had 128 people ranking it as their favourite Star Wars movie. Then a pie chart was constructed using this new dataset . As can be seen from Figure 1, Star Wars Episode 5 was the favourite movie, with 34.5% of respondents who provided preference allocations, voting it as their favourite. Star Wars Episode 4 movie was also quite popular, and Episode 2 and Episode 3 were the least favourite. Secondly, it was desired that exploration into how respondents actually rated Star Wars movies be conducted. In order to do this, pie charts were constructed containing all the preference ranking allocations for each Star Wars Movie.

Figure 1: Pie Chart of Favourite Star Wars Movies

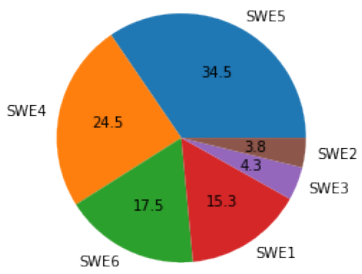


Figure 2: Pie Chart of Proportions of Ranking Allocations for Star Wars Episode 1

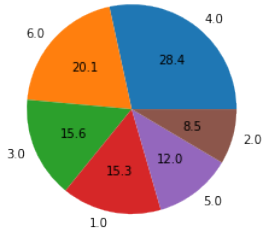



Figure 3: Pie Chart of Proportions of Ranking Allocations for Star Wars Episode 2

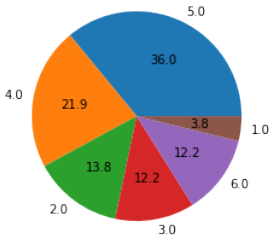



Figure 4: Pie Chart of Proportions of Ranking Allocations for Star Wars Episode 3

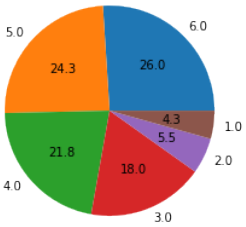



Figure 5: Pie Chart of Proportions of Ranking Allocations for Star Wars Episode 4

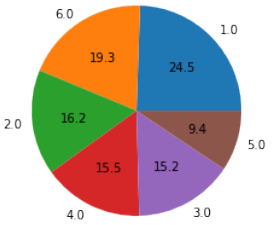



Figure 6: Pie Chart of Proportions of Ranking Allocations for Star Wars Episode 5

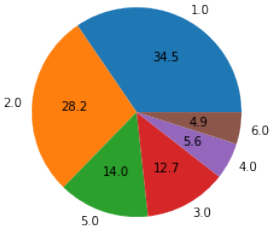
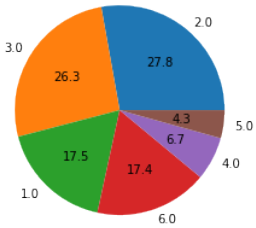


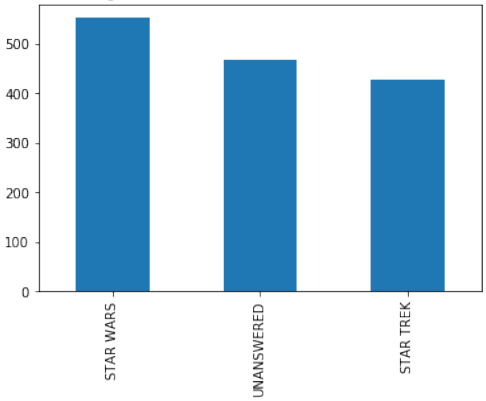

Figure 7: Pie Chart of Proportions of Ranking Allocations for Star Wars Episode 6




As can be seen in the figures above, it is more apparent that the more lower ratings are given to the first three episodes of the Star Wars movies, and more higher ratings are given to the last three Episodes of the Star Wars movies. This suggests that Episodes 1 to 3 are held in lower regard by the respondents as a whole, and Episodes 4 to 6 are the favourites of the Star Wars Franchise, especially the fifth Star Wars movie. Therefore, it is apparent that the respondents do not prefer the prequel movies of the Star Wars Franchise, and thus rate them lower.

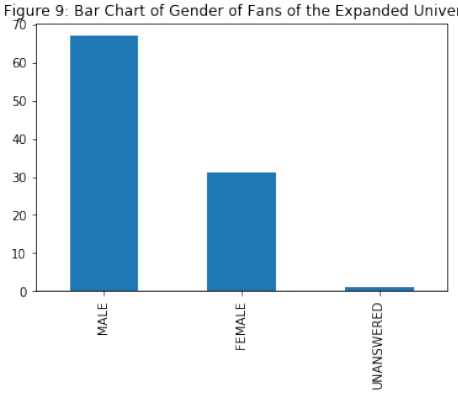
## Task 2.2

Figure 8: Bar Chart of Favourite Franchise



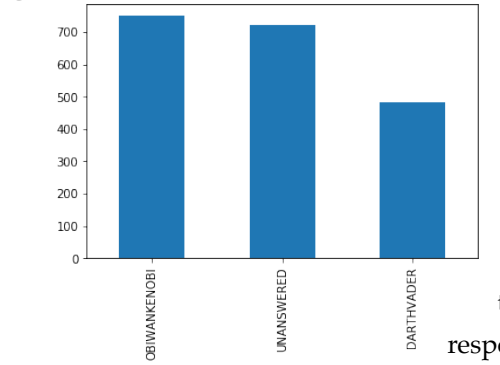

The first hypothesis that was intended to be investigated was whether more fans liked the Star Wars Franchise, or the Star Trek Franchise. According to Figure 8, it can be seen that more respondents are a fan of the Star Wars Franchise as opposed to the Star Trek Franchise. However, the difference is not that large, which was not expected, as it was thought that Star Wars fans would not have much affinity to the Star Trek Franchise. There were a lot of unanswered values, which definitely influences the gap between the number of Star Wars fans and Star Trek fans.

Figure 9: Bar Chart of Gender of Fans of the Expanded Universe




The second hypothesis that was desired to be investigated was whether male respondents were more likely to be fans of the Star Wars Expanded Universe than female respondents. According to Figure 9, it is clear that much more male respondents are fans of the Star Wars Expanded Universe. These results were as expected, however the data might not be the most accurate representation of the relationship between the 'Gender' and 'Do you consider yourself to be a fan of the Expanded Universe?' columns, as there were not many respondents who answered that they were familiar with the Expanded Universe, and an even larger amount of respondents did not indicate whether or not they were a fan of the Expanded Universe.

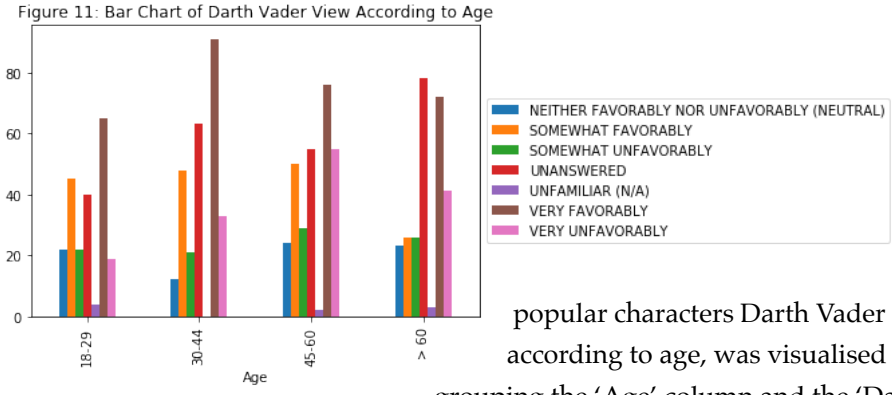
Figure 10: Bar Chart of View of fan favourites - Darth Vader vs Obi Wan Kenobi




The last hypothesis that was investigated was whether Darth Vader was viewed more favourably than Obi Wan Kenobi. Darth Vader and Obi Wan Kenobi are some of the most popular Star Wars characters (TheTopTens, 2020), so this hypothesis was particularly desired to be investigated. According to Figure 10, Obi Wan Kenobi was actually the character that was viewed more favourably. This data may not be the best representation of this fact as there were also a very large number of unanswered responses.

## Task 2.3

Figure 11: Bar Chart of Darth Vader View According to Age




Any relationship between the demographics of the respondents, and their attitudes towards Star Wars characters, was attempted to be ascertained. Firstly, the way attitudes towards the popular characters Darth Vader and Obi Wan Kenobi change according to age, was visualised using two bar charts, after grouping the 'Age' column and the 'Darth Vader View', and the 'Age' column again with the 'Obi Wan Kenobi View' column. According to Figures 11 and 12, it seems that the views towards both characters are similar in that they are both viewed very favourably by all age groups. However, it seems that a larger amount of people aged between 30-44 view Darth Vader more favourably as opposed to Obi Wan Kenobi. However, it seems more respondents aged 45-60 view Obi Wan Kenobi most favourably as opposed to a fewer amount of people for Darth Vader.

Figure 12: Bar Chart of Obi Wan Kenobi View According to Age

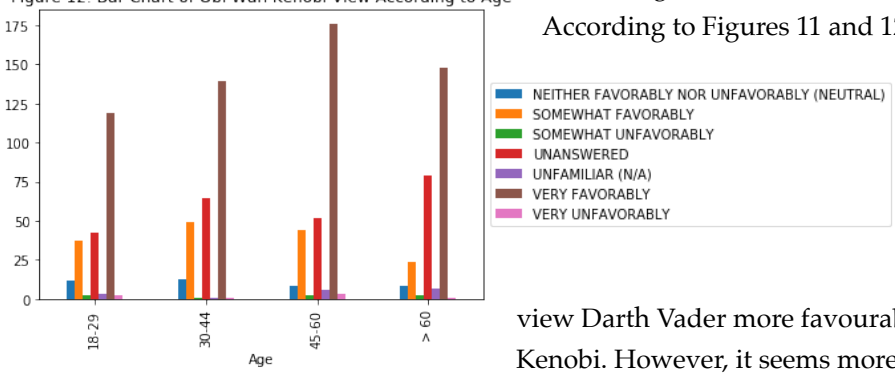


Figure 13: Bar Chart of Han Solo View According to Location

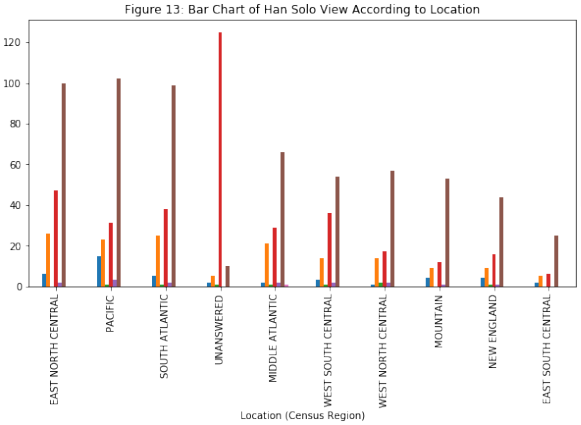
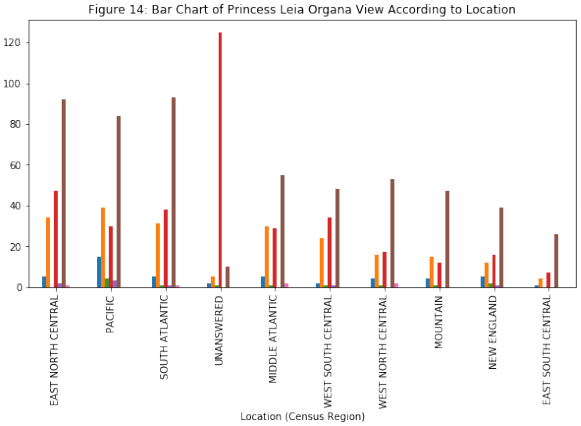
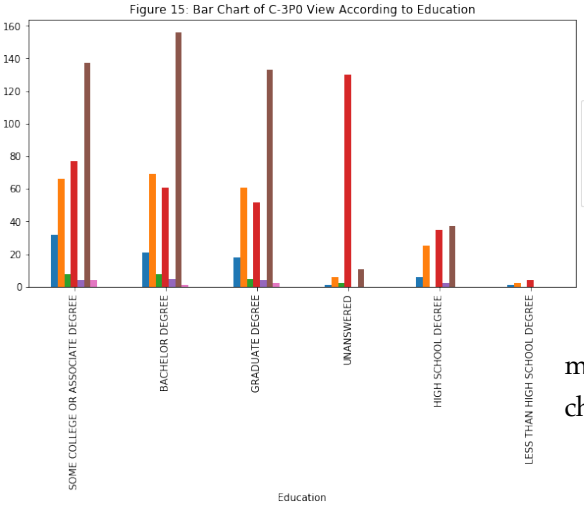



Figure 14: Bar Chart of Princess Leia Organa View According to Location



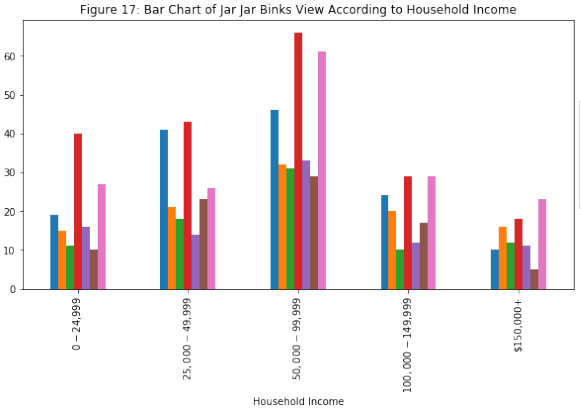

Secondly, the way attitudes towards two other popular characters Han Solo and Princess Leia Organa change according to respondents' location, was visualised using two bar charts after grouping the 'Location' column and the 'Han Solo View', and the 'Location column again with the 'Princess Leia Organa View'. According to Figures 13 and 14, it seems that the relative views for both characters are very similar. It is interesting that for both characters, the respondents' views who are from East North Central are much higher than those who are from East South Central. A similar contrast can be found when comparing the respondents' from the South Atlantic and Middle Atlantic.

Figure 15: Bar Chart of C-3P0 View According to Education



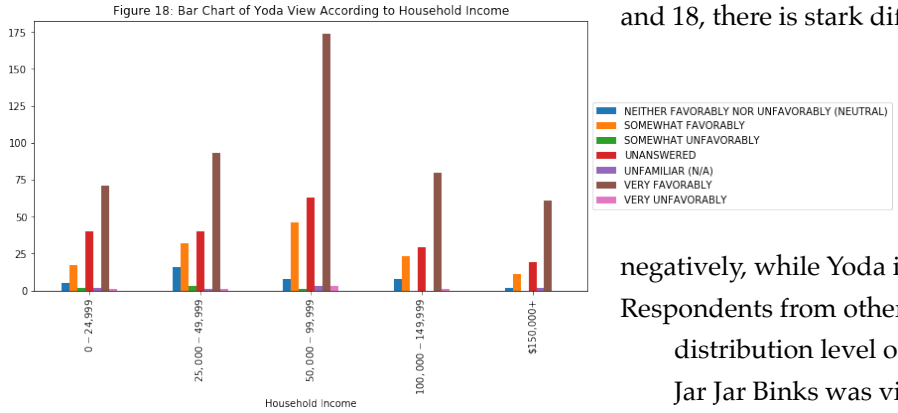

Thirdly, the way attitudes towards C-3P0 according to respondents' education , was visualised using a bar chart after grouping the 'Education' column and the 'C-3P0 View'. According to Figure 15, is seems the highly favourable view towards the popular character does not change, with more respondents who have bachelor degrees finding the character the most favourable.

Figure 17: Bar Chart of Jar Jar Binks View According to Household Income



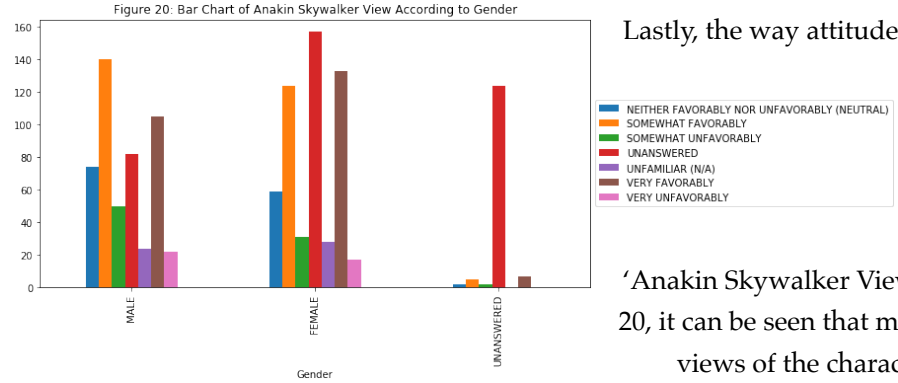

Fourthly, the way attitudes towards two well known characters Jar Jar Binks and Yoda according to respondents' household income, was visualised using two bar charts after grouping the 'Household Income' column with the 'Jar Jar Binks View' column, and again grouping the income column with the 'Yoda View' column. According to Figures 17

Figure 18: Bar Chart of Yoda View According to Household Income




and 18, there is stark differences in the view towards both characters, with a lot of respondents whose income is $50k-99k letting their contrasting views known. Jar Jar Binks is viewed very negatively, while Yoda is viewed very favourably. Respondents from other income categories had a similar distribution level of views for each character, but Jar Jar Binks was viewed negatively on the whole.

Figure 20: Bar Chart of Anakin Skywalker View According to Gender




Lastly, the way attitudes towards the character Anakin Skywalker according to respondents' gender was visualised again using a bar chart after grouping the 'Gender' column and the 'Anakin Skywalker View' column. According to Figure 20, it can be seen that males and females have similar views of the character, with females having more of a very favourable outlook of the character as compared to males. With all of the bar charts above and on the previous pages, the large number of unanswered responses may affect the actual results shown in the bar charts greatly.

# References

TheTopTens 2020, *Top Ten Star Wars Characters*, TheTopTens, viewed 18 April 2020,< https://www.thetoptens.com/star-wars-characters/ >


# Bibliography

*How to put the legend out of plot* 2020, stack overflow, viewed 18 April 2020, < https://stackoverflow.com/questions/4700614/how-to-put-the-legend-out-of-the-plot >

Ma, C 2020, n.t., Canvas Discussions, COSC2670, RMIT University, viewed 18 April 2020, <https://rmit.instructure.com/courses/67430/discussion_topics/669912/page-1 >

TheTopTens 2020, *Top Ten Star Wars Characters*, TheTopTens, viewed 18 April 2020,< https://www.thetoptens.com/star-wars-characters/>

*What's the point of "plt.figure"?* 2020, stack overflow, viewed 18 April 2020, < https://stackoverflow.com/questions/46091681/whats-the-point-of-plt-figure >