**BACKGROUND**

This Guided Project was initiated as part of Dataquest’s course entitled “Data Cleaning”. The data used for the project was taken from Andrew Flowers’ [fivethirtyeight github repository](https://github.com/fivethirtyeight/data/tree/master/star-wars-survey) for an online survey conducted in 2014 (before Star Wars Episode VII: The Force Awakens came out). For this project, emphasis was placed on cleaning the data in preparation for analysis.

**DESCRIPTION**

The following strategies were used to clean the data:

1. Survey responses with “Yes” or “No” answers were replaced with Boolean “True” or “False” values using pandas.Series.map(). Boolean values make it easier to sum up a column than creating loops with counters for counting “Yes” answers.
2. Column names were renamed using pandas.DataFrame.rename(). Column names were renamed for easier referencing.
3. Numerical values stored as character values were converted to floating point numbers using float(), so that mean values can easily be calculated.
4. Descriptive ratings such as “very favorably”, “unfamiliar”, etc. were converted to numerical floating point equivalents, so that mean values can be calculated.

The analysis was mostly focused on finding the most popular Star Wars franchise and the number of respondents watching each franchise. For the most popular Star Wars franchise, the mean for the respondent rankings per franchise was calculated using pandas.DataFrame.mean(). For the number of respondents watching each franchise, the total number per franchise was calculated using pandas.Dataframe.sum(). Bar plots were then generated using pandas.DataFrame.plot().

A variety of approaches was used to analyze the overall ranking and total number of viewers for each franchise. Throughout the project, the following strategies were used for the analysis:

1. Segmenting the data into Star Wars fans and those who are not Star Wars fans;
2. Segmenting the data into Star Trek fans and those who are not Star Trek fans;
3. Segmenting the data into Male and Female respondents;
4. Segmenting the data according to educational attainment;
5. Segmenting the data according to location;
6. Segmenting the data according to respondents who believed that Han Solo shot first and those who believed that Greedo shot first (from a controversial scene in Star Wars Episode I: A New Hope);

The final analysis was devoted to finding which characters are popular, which characters are least liked, and which characters were controversial (in both the positive and negative sense). Once more, the analysis used pandas.DataFrame.mean(), pandas.DataFrame.sum() and pandas.DataFrame.plot().

**FILES**

The following files are included in this project:

|  |  |
| --- | --- |
| FILE NAME | DESCRIPTION |
| GP\_Star\_Wars.ipynb | The Jupyter Notebook version of the project. |
| star\_wars.csv | The data set in \*.csv format |
| GP\_Star\_Wars.html | The project in html format for easy viewing |
| GP\_STARWARS\_README.docx | Short documentation for the project. |