UDACITY

Investigate a Dataset

REVIEW

HISTORY

Meets Specifications

Very impressive submission! I can see your hard work reflected in your project 🏆 Congratulations on achieving this and good luck on your way to master data analysis 😃

Code Functionality

All code is functional and produces no errors when run. The code given is sufficient to reproduce the results described.

The project uses NumPy arrays and Pandas Series and DataFrames where appropriate rather than Python lists and dictionaries. Where possible, vectorized operations and built-in functions are used instead of loops.

The code makes use of functions to avoid repetitive code. The code contains good comments and variable names, making it easy to read.

Quality of Analysis

The project clearly states one or more questions, then addresses those questions in the rest of the analysis.

Data Wrangling Phase

The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.

Exploration Phase

The project investigates the stated question(s) from multiple angles. At least three variables are investigated using both single-variable (1d) and multiple-variable (2d) explorations.

The project's visualizations are varied and show multiple comparisons and trends. Relevant statistics are computed throughout the analysis when an inference is made about the data.

At least two kinds of plots should be created as part of the explorations.

Conclusions Phase

The results of the analysis are presented such that any limitations are clear. The analysis does not state or imply that one change causes another based solely on a correlation.

Brilliant job here 👏 just give to the conclusions and limitations the following structure:

Conclusions

In the first section I examined the popularity of Western movies over the decades. I made my analyzation based on the values of 'released_year' and 'popularity'. I could not find any correlations between the numbers and the assumptions but I found it by taking into account the numbers of released movies.

After that I analyzed the ratings of the most and least expensive movies and I found out that the more expensive movies got higher votes than the cheaper ones.

Limitations

In the first section - although the literature details the phenomenon - I could not find any correlation between 'popularty' and 'release year'. It would be good to know more about what is behind the value 'popularity' and what popularity means here. Just to name a few... How was it calculated? Which criterias and values were measured exactly to get these numbers? It could be caculated based on ticket sales? Or based on audience appraisal? However, I found correlation between my assumptions and the number of released western movies but I would not name it causation without a much more detailed further analysis.

In the second section, I made my calculations based on the values of budget adjustment to take the fluctuations into account, I found this really useful. But there were more missing values in the 'budget_adj' column. During the cleaning process I replaced the missing values with the average, but it still can distort the result (for instance, there would be other movies among the most expensive 200 movies).

Communication

Reasoning is provided for each analysis decision, plot, and statistical summary.

Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted.

⬇ DOWNLOAD PROJECT

Rate this review