Investigate a Dataset

REVIEW

HISTORY

Requires Changes

5 specifications require changes

Good first submission! Your project reflects your hard work and I have to congratulate you for that 😄 Your code is very solid as well, you only need some modifications in order to continue. Good luck in your next submission!

Don't hesitate to reach your help in Student Hub, we are here to help you succeed 🏆

Code Functionality

✓

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

Your code is functional and runs without error. Well done!

✓

- **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.**

As a data scientist, you'll frequently interact with NumPy arrays, pandas Series, and pandas DataFrames, and you'll leverage a variety of NumPy and pandas methods to perform your desired computations. Understanding how NumPy and pandas work together

will prove to be very useful.

- **The code makes use of at least 1 function to avoid repetitive code.**
- **The code contains good comments and meaningful variable names, making it easy to read.**

Excellent work defining the function. It makes your code so clean and efficient 😃. If you want to know more about functions and why and the importance you can check this: https://www.udacity.com/blog/2020/11/how-to-define-a-function-in-python.html#:~:text=Functions%20are%20the%20building%20blocks,how%20functions%20should%20be%20used

Quality of Analysis

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

Amazing job with the introduction and the questions of your dataset! That is very valuable information for the reader 😃

Data Wrangling Phase

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

Good work implementing and describing the Data Wrangling Phase!

Exploration Phase

- **The project investigates the stated question(s) from multiple angles.**
- **The project explores at least three variables in relation to the primary question. This can be an exploratory relationship between three variables of interest, or looking at how two independent variables relate to a single dependent variable of interest.**
- **The project performs both single-variable (1d) and multiple-variable (2d) explorations.**

They add so much valuable insights to the overall smooth flow of the project. Well done 😉

However, there are no 1d plots in your project. Please add a few 1d plots like histograms or box plots so that the audience can get an understanding about the distribution of the data.

Here are the differences between bivariate and univariate data:

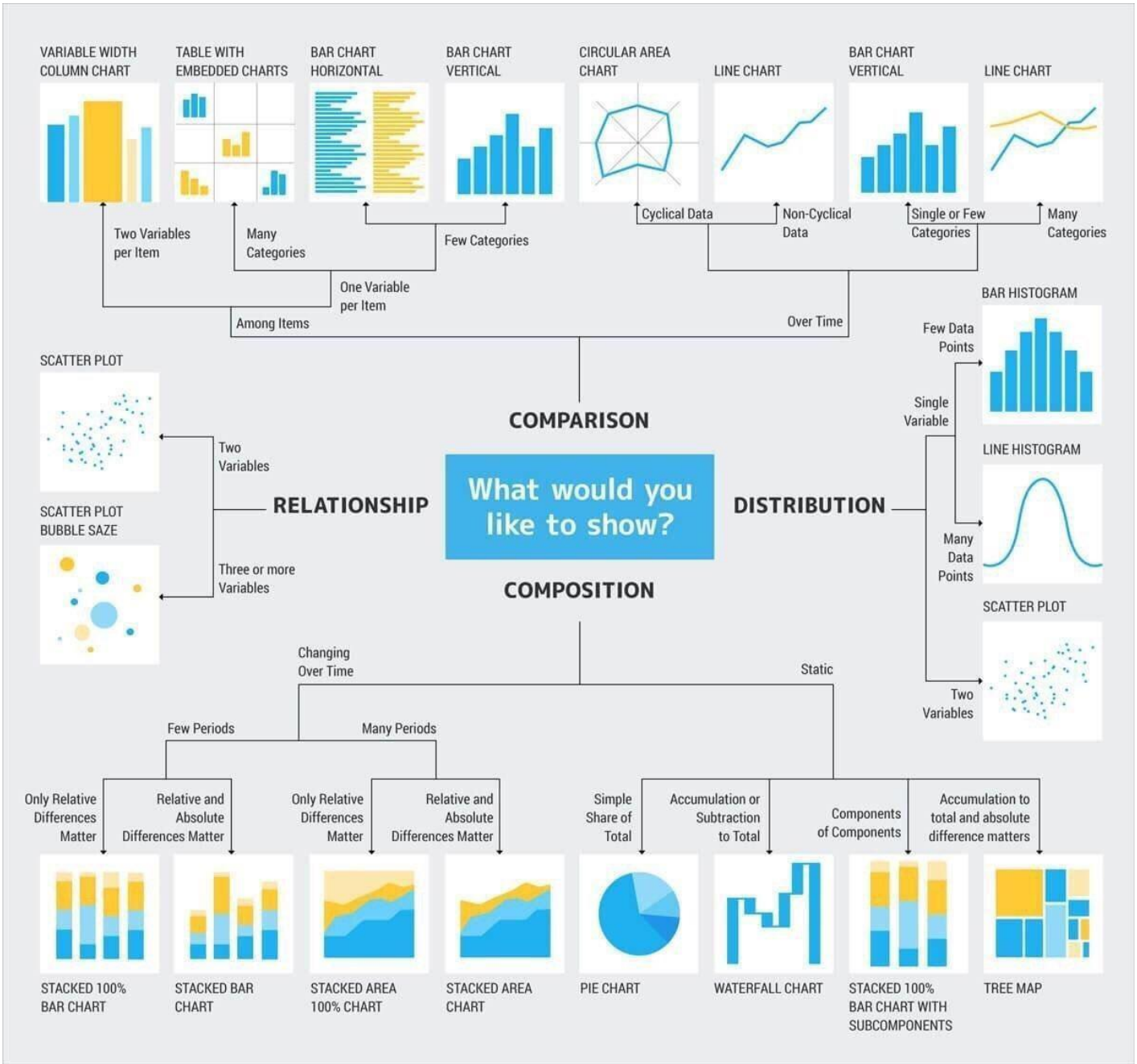
This link summarises the difference between bivariate and univariate data.

Univariate Data	Bivariate Data
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central tendency - mean, mode, median • dispersion - range, variance, max, min, quartiles, standard deviation. • frequency distributions • bar graph, histogram, pie chart, line graph, box-and-whisker plot	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations • tables where one variable is contingent on the values of the other variable. • independent and dependent variables
Sample question: How many of the students in the freshman class are female?	**Sample question:** Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

And here's some more inspirations for you: https://seaborn.pydata.org/tutorial/categorical.html

- **The project's visualizations are varied and show multiple comparisons and trends.**
- **At least two kinds of plots should be created as part of the explorations.**
- **Relevant statistics are computed throughout the analysis when an inference is made about the data.**

This is something you can improve, your report will improve as soon as you add more variable graphics, let me suggest the following tool to decide which one to use in each case 😉



please take a look at the following link and add different types of graphics, your report would look more professional with that.

Conclusions Phase

- **The Conclusions have reflected on the steps taken during the data exploration.**
- **The Conclusions have summarized the main findings in relation to the question(s) provided at the beginning of the analysis accurately.**
- **The project has pointed out where additional research can be done or where additional information could be useful.**
- **The conclusion should have at least 1 limitation explained clearly.**
- **The analysis does not state or imply that one change causes another based solely on a correlation.**

I'm going to ask to you please expand your conclusions, this is the most important part of your report and you really need to shine one here 😉 The conclusion is intended to help the reader understand why your EDA should matter to them after they have finished reading the analysis. A conclusion is not merely a summary of the main topics covered or a re-statement of your research problem but a synthesis of key points .

There should be a separate subsection inside the conclusion section called 'Limitations' where you would have to discuss the limitations of this dataset which might have adversely affected your analysis. Examples would be null or missing values, whether these samples are an effective representation of the population or not or maybe that you could dive deeper into your analysis with additional specific information.

The conclusions and limitations have 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

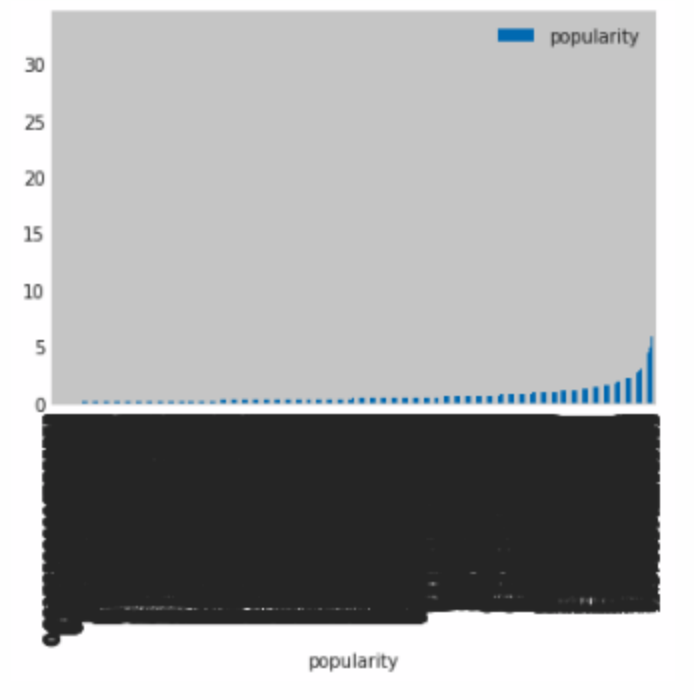
In the first section - although the literature details the phenomenon - I could not find any correlation between 'popularty' and 'release year'. It would be good to know more about what is behind the value 'popularity' and what popularity means here. Just to name a few... How was it calculated? Which 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

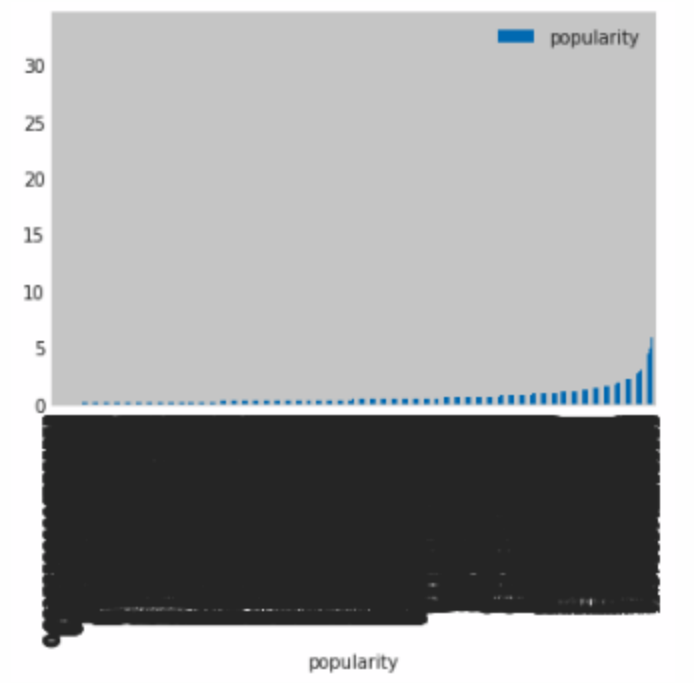
- **The code should have ideally the following sections: Introduction; Questions; Data Wrangling; Exploratory Data Analysis; Conclusions, Limitation.**
- **Reasoning is provided for each analysis decision, plot, and statistical summary.**
- **Interpretation of plots and application of statistical tests should be correct and without error.**
- **Comments are used within the code cells.**
- **Documented the flow of analysis in the mark-down cells.**

Not every analysis is followed by an explanation. For example the graphic(s):



Please include one or two sentences talking about what information you can extract from the data. The same applies to the tables you generated. What can we conclude by analyzing these values? Do not forget to use a markdown cell for your analysis (not comments) 😄

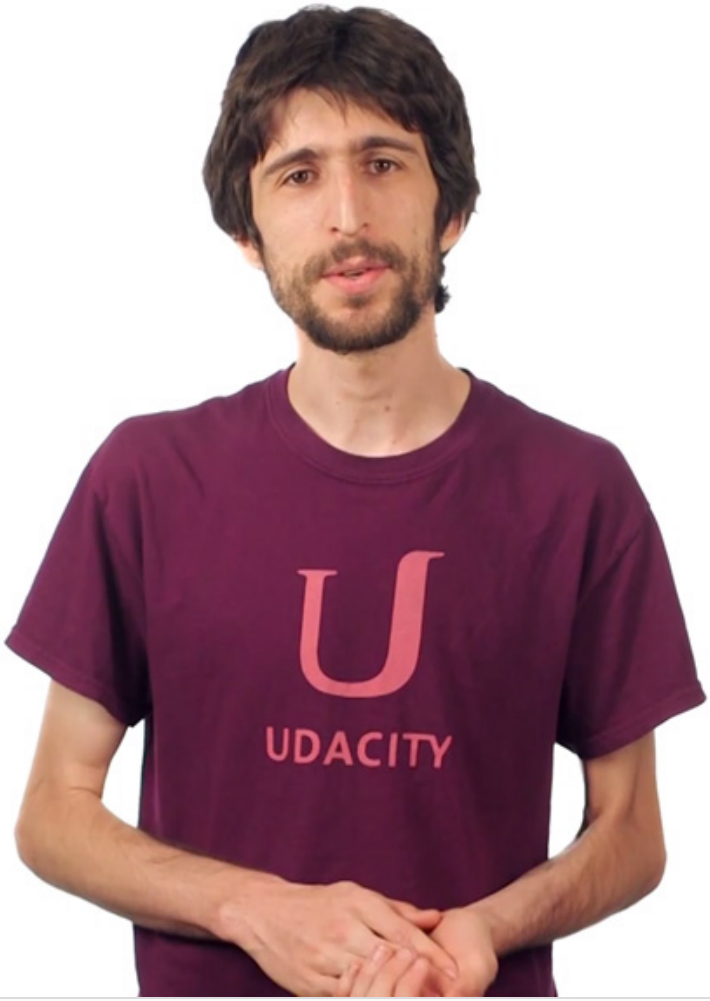
Visualizations made in the project depict the data in an appropriate manner (i.e., has appropriate labels, scale, legends, and plot type) that allows plots to be readily interpreted.



Please make sure that all the visualizations in your project are legible

RESUBMIT

DOWNLOAD PROJECT



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

RETURN TO PATH

Rate this review

START