UDACITY

‹ Return to Classroom

Investigate a Dataset

REVIEW

HISTORY

Requires Changes

2 specifications require changes

Dear Student,
This project qualifies for a great first submission! I am impressed with your communication and organisational skills in coming up with a structured report. Your awesome storytelling skills will make you a remarkable data scientist. Keep shining, you star! ⭐

Here is an overview of the required changes:

- the limitations section does not talk about the data.
- reasoning is missing under some visualisations.

Read the detailed suggestions below on how to improve the project and I'm sure you'll pass in the next submission soon. 😛
I wish you all the best for the next submission. 👍🏻
Stay Safe. Keep Learning. Stay U!

Code Functionality

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

Comments

Well done! The submitted code works well as it doesn't produce errors when run. Also, it's 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.

Comments

Excellent work using NumPy and Pandas functions to facilitate the work for this submission. 👏🏼

- 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.

Comments

You have given comments and variable names. You've also implemented a custom Python function eloquently in the project.

Quality of Analysis

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

Comments

Nice job stating questions and addressing the questions throughout 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.

Comments

You've used Markdown and code comments efficiently in documenting the project step-by-step. It is easy to understand and follow through. 😊

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.

Comments

You've thoroughly investigated the questions from various angles, and also use both 1d and 2d explorations for several variables in consideration. 🤝

I like the way you have represented the number of movies every year using treemap. It is a unique visualisation that is beautiful and easy to understand. Keep it up!

- 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.

Comments

Visualising data requires a lot of patience and determination because it's not easy selecting the best visualisation to match with a given data type. Well enough, the project rightly builds descriptive visualisations using a variety of plots. 👌🏻

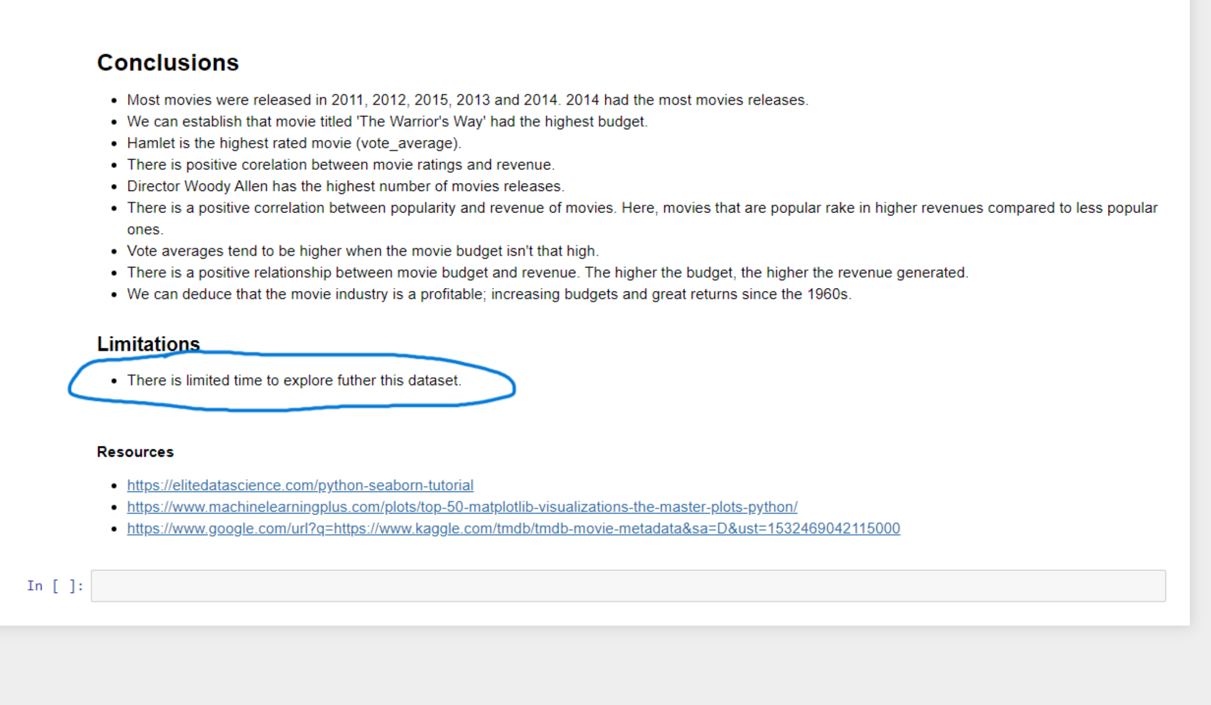
Conclusions Phase

- 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 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.

Requires Changes

Do you consider your explanation a limitation of the data?



Conclusions

- Most movies were released in 2011, 2012, 2015, 2013 and 2014. 2014 had the most movies releases.
- We can establish that movie titled 'The Warrior's Way' had the highest budget.
- Hamlet is the highest rated movie (vote_average).
- There is positive corelation between movie ratings and revenue.
- Director Woody Allen has the highest number of movies releases.
- There is a positive correlation between popularity and revenue of movies. Here, movies that are popular rake in higher revenues compared to less popular ones.
- Vote averages tend to be higher when the movie budget isn't that high.
- There is a positive relationship between movie budget and revenue. The higher the budget, the higher the revenue generated.
- We can deduce that the movie industry is a profitable; increasing budgets and great returns since the 1960s.

Limitations

- There is limited time to explore futher this dataset.

Resources

- https://elitedatascience.com/python-seaborn-tutorial
- https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/
- https://www.google.com/url?q=https://www.kaggle.com/tmdb/tmdb-movie-metadata&sa=D&ust=1532469042115000

```
In [ ]:
```

No, **limitations are something you come across while doing the analysis part in the dataset itself which may or may not affect the final predictions.** What hinders your analysis of the current data? And be elaborate in your analysis why you say there are hindrances?

Say, for example, there are more than 5 to 10% of data having null values or highly correlated having erroneous or missing values or imbalanced data. The sample doesn't represent the population. All these will lead either to wrong analysis which will lead to wrong predictions or biased analysis. Such ones only should be mentioned as your limitations.

In such cases, it's always good to list down and give an elaborate explanation about the limitations and what way they may affect the output. Be very specific while mentioning limitations.

After completing the analysis if you feel there are no limitations in the input dataset, you can as well mention that there are no limitations in this project.

The sample must be good enough to represent the population. So that our analysis will have enough data to generalise to the entire population.

A limitation is a roadblock you have come across during your analysis. For example, a higher % of missing or wrong data or outliers will reduce the accuracy of the analysis outcome. The lesser or imbalanced sample may not represent the population. Any such cases will produce biased or inaccurate analysis and we can't generalise such output for the entire population. Those need to be given under limitations to give a caution to the user while using this analysis for further prediction or to take any important business call.

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.

Requires Changes

Please provide your reasoning under all visualisations. There are some ideas that might be helpful:

- Explanation about the choice of the graph?
- Why do you want to explore these variables?
- Any findings from the visualizations?

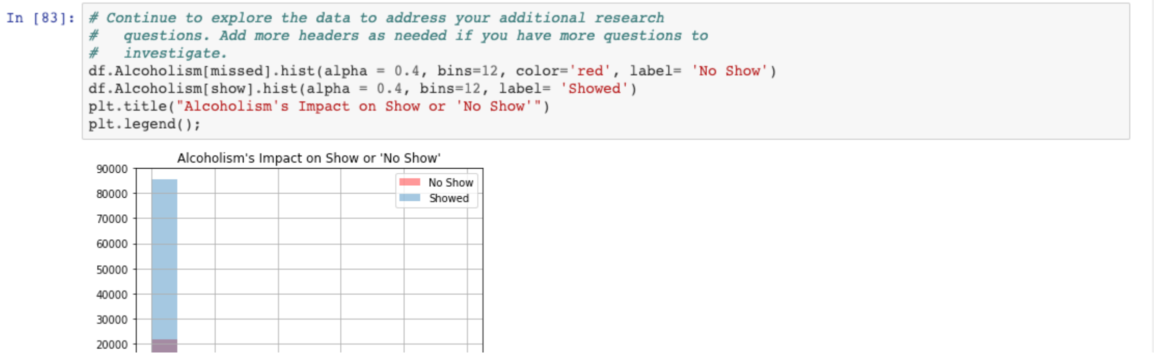
It seems that some of your discussions in the data wrangling phase are placed in the commented lines of code cells. These are also good reasoning that could help your readers understand why you want to explore a specific relationship between variables or why you do some kind of data wrangling. In your future work, some of your readers are not able to code, so they might just skip the content in the code cells, including your reasoning put in the code cells. As such, I suggest you could also state this kind of reasoning in the markdown cells.

The reasoning is another important factor in data analysis. Though we give an overall conclusion at the end of the project, reasoning on **each and every statistical summary** and data visualisation of every question is required. That way we can convey the intention of each step of the analysis.

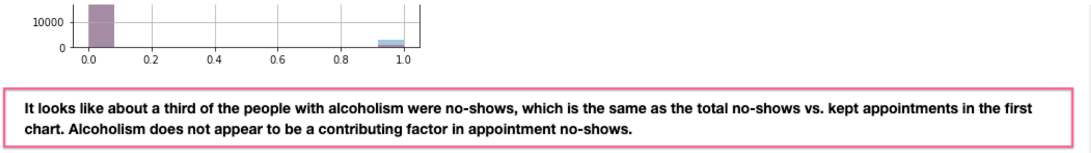
For any Data Science Project, you need to give an elaborate explanation for every part of your analysis. The reason because you need to convince the client about your analysis. This analysis doesn't stop here. The client will be using this analysis outcome for further analysis or some predictions to take any important business decisions. That's the value this analysis carries. Have more write up will ease the part of conveying your analysis inference.

Add a few lines after each visual about what do you infer from each visual so as to answer the respective question.

Here is an example:



```
In [83]: # Continue to explore the data to address your additional research
         #   questions. Add more headers as needed if you have more questions to
         #   investigate.
         df.Alcoholism[missed].hist(alpha = 0.4, bins=12, color='red', label= 'No Show')
         df.Alcoholism[show].hist(alpha = 0.4, bins=12, label= 'Showed')
         plt.title("Alcoholism's Impact on Show or 'No Show'")
         plt.legend();
```



It looks like about a third of the people with alcoholism were no-shows, which is the same as the total no-shows vs. kept appointments in the first chart. Alcoholism does not appear to be a contributing factor in appointment no-shows.

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.

Comments

Awesome! The plots are well labelled and easy to interpret. 👍

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Rate this review

START