UDACITY

PROJECT

Investigate a Dataset

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

4 SPECIFICATIONS REQUIRE CHANGES

Hi. Thanks for submitting your report. It is coming up great! 😃

I made some suggestions that will help you improve it.

On your submission, you sent the following comment:

> I know we're not supposed to be focused on statistical inference for this project, but I am not certain the kind of analysis I was doing is even coherent (much less correct). What kinds of assessment should I expect myself to know at the completion of this project?

It is true that statistical tests are not required for this project, but you are free to do it, and we will review them for you.

The primary purpose of this project is to evaluate if you can make a question, answer it using data, and communicating it to another person through a report. This routine is very common in the daily work of a Data Scientist.

At this point, the report is missing a specific question, and it seems to be investigating the dataset. I recommend that you come up with a hypothesis about movies where you can investigate using this dataset.

Keep up the good work.

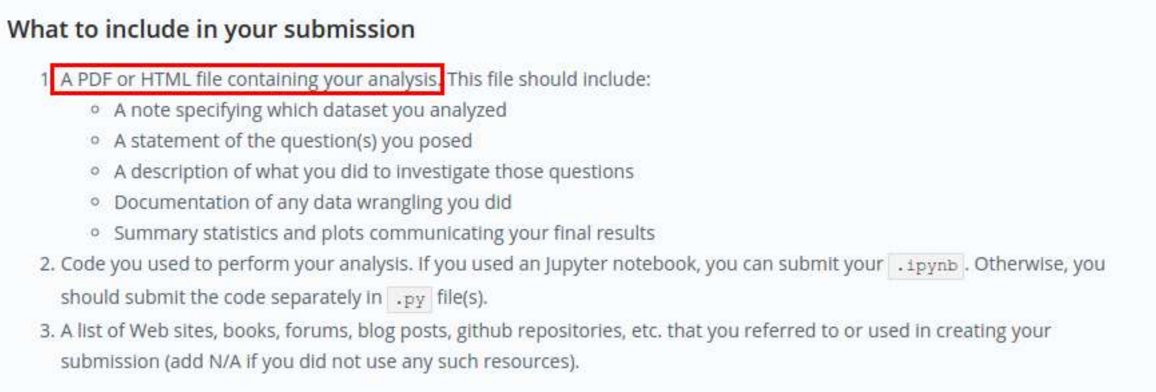
Gustavo ✌

Code Functionality

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

Required

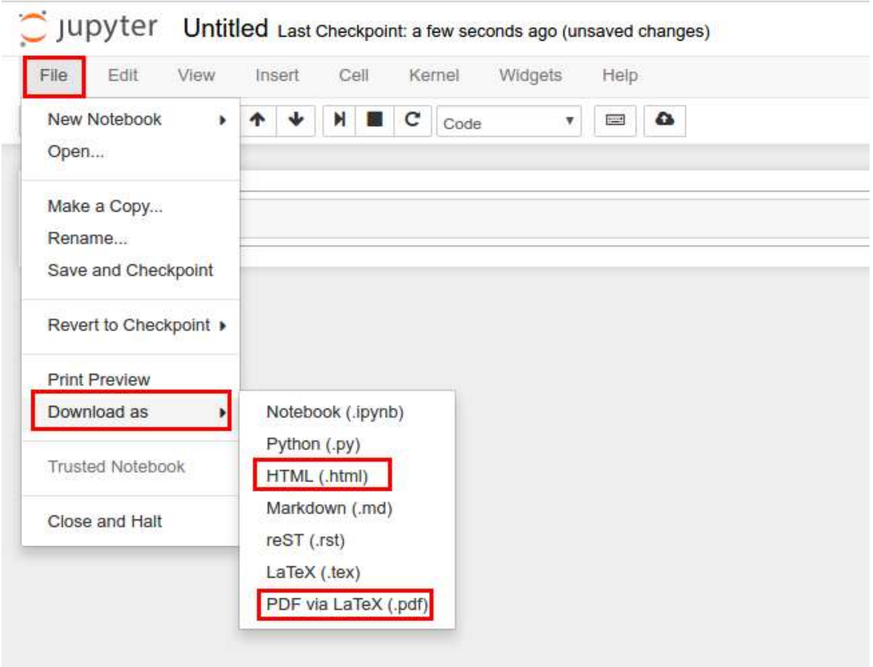
The project description asks that the report should be sent in HTML or PDF:



What to include in your submission

1. A PDF or HTML file containing your analysis. This file should include:
   - A note specifying which dataset you analyzed
   - A statement of the question(s) you posed
   - A description of what you did to investigate those questions
   - Documentation of any data wrangling you did
   - Summary statistics and plots communicating your final results
2. Code you used to perform your analysis. If you used an Jupyter notebook, you can submit your `.ipynb`. Otherwise, you should submit the code separately in `.py` file(s).
3. A list of Web sites, books, forums, blog posts, github repositories, etc. that you referred to or used in creating your submission (add N/A if you did not use any such resources).

The image below shows how to export the Notebook to one of those formats.



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.

Suggestions

The following resources can help you take your Pandas skills to the next level:

- Essential Basic Functionality
- Indexing and Selecting Data
- Working with missing data
- Working with Text Data
- Group By: split-apply-combine
- Merge, join, and concatenate
- Visualization

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

The code is clean and the Markdown comments help us understanding them.

Quality of Analysis

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

Required

The introductory section is well written. Well done! 👍

As I note in my initial comments, the report is missing a specific question/hypothesis which will guide the analysis. Here are some suggestions:

- Is it true that more movies are released before the Oscars Awards?
- When Horror movies are released? Near the Halloween?
- Are hero movies profitable?
- If a friend like Action that are also Adventure movies, which one should I recommend?

These are just some ideas, but you can come up with others.

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.

Suggestions

A nice set of plots was chosen.

The following resources can be useful to know more about the various types of plots available.

- The Data Visualization Catalogue
- And this gallery "displays hundreds of charts, always providing the reproducible python code!"

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.

Required

Please, write a conclusion section summarizing your main results and limitations of the analysis and/or dataset.

Communication

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

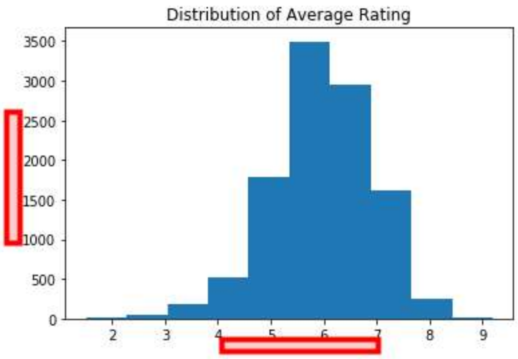
The report is well detailed and results are being explained.

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

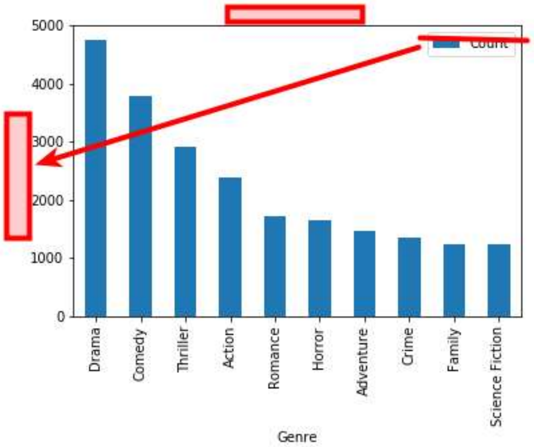
Required

You are doing an excellent job with the plots. We can change/add some features to improve their understanding:

- We must make sure that both axes are labels. For example, the plot below is missing labels on both:



- Legends are used when we have two or more different categories in the plot. In the plot below, we have only one category, so we can remove the legend and add a label to the y-axis that explains the same thing.



The plot is also missing a title.

Suggestion

This dataset has a lot of entries, so when we make a scatter plot, the dots get on top of each other, which can hide some features like clumps. I have two suggestions to improve them, and you can pick one that you like:

1. Decrease the opacity of the dots using the `alpha` argument. This parameter goes from 0 (transparent) to 1 (opaque).
2. Use one of the strategies described in this tutorial.

The following resources have great tips on how to make better plots:

- Ten Simple Rules for Better Figure
- A guide on how to make better Matplotlib figures.
- Effectively Using Matplotlib

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

Student FAQ