Investigate a Dataset

REVIEW

HISTORY

Meets Specifications

Congratulations! You've made a strong submission. In this project, you've addressed questions first and then use your data skills learned in the courses to clean and explore the data. There are a few decisions you've made along with your analysis. It is terrific that you've documented them well so that it's much easier for your readers to follow. Different variables are also explored and visualized. Good job, I am super proud of you!

Code Functionality

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

The code is provided along with the project. All code is functional and sufficient to reproduce the results described. Nice work!

Jupyter Notebook is a powerful tool for data scientists as it provides a super convenient way to experiment when you are exploring your dataset. You could reuse the data produced by your previous cells, add a filter and test new ideas. You could also document your reasoning and decisions you've made along with your code. That is a more friendly way for your readers to follow.

Sometimes organize your Jupyter Notebook is a non-trivial task. You will have to document your analysis in the Markdown cells (instead of Code cells, which is not easy to read for non-technical people) with Markdown

syntax. Here is a Markdown Cheat Sheet that I often refer to. Hope that might be helpful when you need to consult for some tricks.

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.

Pandas Series and DataFrames are appropriately applied when wrangling and exploring the dataset. Good job!

Where possible, stick to Pandas or Numpy rather than handling data with Python lists or dictionaries, as Pandas and Numpy has much faster speed when operating your data (filtering, transforming, etc..). Pandas is highly optimized for data analysis. Here is also a brief discussion about Pandas vs. Python List, which inspired me when I started my journey in the data analysis field. In general, Pandas is a good start and will solve many problems in your daily life.

Sometimes we may need a quick brush up on Pandas. I would suggest:

- Official Quick Start
- Pandas Cheat Sheet

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

Comments are provided along with the code, which makes it easy for your readers to understand your code and follow your analysis. Nice work!

Suggestion:

If possible, I encourage you to include at least one user-defined function, which is a way to avoid potential repetitive code. One of the beauties of the programming is the ability of abstraction. It could abstract some common ideas into a tool that you could reuse for multiple times. This idea avoids repeating yourself again and again. This is also a guideline called DRY (Don't repeat yourself).

Quality of Analysis

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

Questions are clearly stated at the beginning of the project. In the rest of the analysis, the dataset is wrangled and explored to answer the stated questions. It is well structured and easy to follow. Good job!

Data Wrangling Phase

Data Wrangling Phase

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

When wrangling data, changes are well documented. This makes it easy for your readers to understand what have been performed over the dataset. Great job!

In most cases, the dataset you've got is not ready for data exploration, as datasets in the real world are often in a messy. You want to try different approaches to clean and transform into tidy data. It might not be obvious when you are already given tidy data sources. It is more important if you are dealing with combinations of different sources, or unstructured/semi-structured data sources.

Here are a few helpful guidelines that will help during the wrangling phase:

- Is there any missing data under each column? If yes, how to deal with them? Is it safe to just drop lines with missing values? Is it safe to replace with a statistical average (mean, median, mode)?
- What's the distribution of the variable? Are the distribution and the data range sound reasonable to you? For example, if you see a negative age in the distribution, you might want to take a deep look, as it is impossible in the real world.
- Do you want to create some new variables to support your data exploration? As we are exploring our dataset, so it is totally acceptable to create new dimensions based on our knowledge, even if we find it not that useful in the end. For example, if we have `investment` and `revenue`, we might want to create a new variable `roi (revenue/investment)` to see the efficiency of capital use.

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.

Many variables are investigated using both single-variable and multiple-variable exploration. Good job!

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.

Different types of visualization are included in the project. Nice work!

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.

The results of the analysis are stated at the end of the project. Limitations are also touched to inform the readers required information. Nice work!

Conclusion is not only for you but also for your readers to wrap up the key findings or interesting trends from your overall analysis. If your readers (sometimes it's your boss) don't have enough time to go through your report, this is the place that they can consult and quickly get the idea you want them to know.

Communication

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

Reasoning is provided under each graph. It is easy for your readers to understand what's going on and insights we've learned from the graphs. Well done!

Suggestion:

It seems that some of your discussions are placed in the commented lines of code cells (e.g.: cell 59). 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.

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

Visualizations are all well done with clear axis labels and graph titles. Good job!

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START