* Logout

  Return to Classroom

Investigate a Dataset

* **REVIEW**
* **HISTORY**

**Requires Changes**

2 specifications require changes

You've made a very nice first submission. Good job! The project shows how you apply what you've learned into a real-world problem. I am proud of you! There are only a few suggestions that may help you improve your analysis to a higher level. You could refer to the below. Good luck and look forward to your next submission!

**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](https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet) 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](https://www.reddit.com/r/Python/comments/3aeled/dataframe_vs_nested_list_vs_dictionary_for/), 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](https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html)
* [Pandas Cheat Sheet](https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf)

**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 (not in the code cell but clear enough for your readers to understand your decisions). It makes it easy for your readers to follow. Nice work! However, the rubric requires to have at least one user-defined function to avoid repetitive code. Here are some suggestions that might be helpful:

* Take a look if there are some similar plots that share common codes. You might be able to create a function for these plots
* Take a look if there are similar workflows when you wrangle your data. A function will also help to avoid 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**

**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](https://en.wikipedia.org/wiki/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 (histograms and bars) 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.**

A section of the conclusion is included at the end of the report. Nice work! I would also recommend you discuss a little bit about the limitations of our current analysis. There are a few ideas you may think about:

* Is the data we have sufficient to draw the conclusions?
* Is there some possibility that we are missing some important hidden variables?
* Does correlation mean causation?

**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!

**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!

 RESUBMIT PROJECT

[**DOWNLOAD PROJECT**](https://review-api.udacity.com/api/v1/submissions/2563719/archive)



**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

( )( )( )( )( )