Investigate a Dataset

REVIEW | HISTORY

Requires Changes

1 SPECIFICATION REQUIRES CHANGES

Dear Student, Awesome first submission !!! Good work !!! I could see lots of learning and implementation of the same in the project !!! Especially the way plots are created using Python packages !!!

You are just one step closer to the project completion !!! Only one specification is missing in the project. Include that to complete this project !!!

Keep up your good work !!!

This is starting point of your Data Science Career. And I would say you landed in the right place. If interested, download more real time dataset from kaggle.com and do analysis on those datasets.

Here are few references you could use :

https://chrisalbon.com/
https://machinelearningmastery.com/
https://machinelearningmastery.com/data-cleaning-turn-messy-data-into-tidy-data/
www.kaggle.com
www.analyticsvidhya.com

As you know math algebra is the basis of all our data analysis. Again to refresh those topics there are many links and you might have come across them in the course itself.

khanacademy
mathisfun
https://www.youtube.com/watch?v=Dft1cqjwIXE

Code Functionality

✓	**All code is functional and produces no errors when run. The code given is sufficient to reproduce the results described.**
	Code ran without any error. Meet the specification.

✓	**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.**
	Numpy and pandas series and dataframes are implemented very well in the project. Meet the specification. The main objective of this project is to make use of pandas methods in the place of lists/dictionaries and loops. Using loops will take more run time as well as more lines of code which can be replaced with just one or two pandas methods that has lesser run time and lesser lines of code.

✓	**The code makes use of functions to avoid repetitive code. The code contains good comments and variable names, making it easy to read.**
	You have given comments and variable names. Function implementation is also done very well in the project. Meet the specification.

Quality of Analysis

✓	**The project clearly states one or more questions, then addresses those questions in the rest of the analysis.**
	You have stated questions to be analysed in the beginning of the project. Meet the specification. That's the right way to start any Data Science Project. First thing to do is to state the questions to be analysed. All the other phases will be done based on those initially stated questions.

Data Wrangling Phase

✓	**The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.**
	You have done some cleaning required for the analysis. Meet the specification. Like you have considered age lesser than 0 as outlier, you must consider age greater than 100 especially 115 also to be considered as outlier. Though those are possible values in the real time scenario, still those are outliers. Any outliers must be cleaned before analysis so that analysis will be accurate.

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.**

You have done both single and multi variate exploration in the project. Meet the specification. Explore more on matplotlib, seaborn, bokeh python libraries. As you work on more these libraries, you will be able to show third dimension also in the visualisation.

✓ **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.

You have used histogram and bar plots in your project. Though both the plots belong to the same kind, the way of creating the plots are different. That way i considering these two as different types.
To find the impact of a parameter or ailment on patient turn out, you have used both no-show and show data. Thats very good way and right way to find the impact. Otherwise usual tendency is to use only no-show and that won't give the full picture of impact. Good ones !!!

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.**

Overall conclusion is given in the end of the project. Good ones !!! Conclusion is very important phase of a project especially analysis kind of project wherein you will give your overall intention and interpretation about your analysis. You got it right !!! Limitation is missing in the conclusion phase. Add the limitations you have come across with the input dataset during your analysis. Add limitation as a separate heading under conclusion phase itself.

Here are few suggestions about limitations :

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 with 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 is having null values or highly correlated having erroneous or missing values or imbalanced data. 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, its always good to list down and give elaborate explanation about the limitations and what way it may affect the output. Be very specific while mention about 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.

Sample must be good enough to represent the population. So that our analysis will have enough data to generalize to entire population.

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

Communication

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

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

Every visual must have axes and data labels, title and legend (if any). All the visuals in the project are neat and easy to interpret.

RESUBMIT PROJECT

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