UDACITY

Back to Data Analyst Nanodegree

Investigate a Dataset

REVIEW

HISTORY

Requires Changes

7 SPECIFICATIONS REQUIRE CHANGES

Code Functionality

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

Please submit the notebook so that I can run and check the functionality of the code.

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.

Wonderful work!

As a data scientist, you'll frequently interact with NumPy arrays, pandas Series, and pandas DataFrames, and you'll leverage a variety of NumPy and pandas methods to perform your desired computations. Understanding how NumPy and pandas work together will prove to be very useful.

COMMENTS:

NumPy is a Python extension module that provides efficient operation on arrays of homogeneous data. It allows python to serve as a high-level language for manipulating numerical data, much like IDL, MATLAB, or Yorick. (https://www.scipy.org/scipylib/faq.html)

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Benefits of Pandas are:

```
Data representation: It can easily represent data in a form naturally suited for dat
a analysis via its DataFrame and Series data structures in a concise manner.

Data subsetting and filtering: It provides for easy subsetting and filtering of dat
a, procedures that are a staple of doing data analysis.

Concise and clear code: Its concise and clear API allows the user to focus more on t
he core goal at hand, rather than have to write a lot of scaffolding code in order t
o perform routine tasks. (https://goo.gl/BvBkL2)
```

Some of the few Pandas built-in methods that are very useful for exploring variables in this project:
• Boolean-Indexing: http://pandas.pydata.org/pandas-docs/stable/indexing.html#boolean-indexing
• Group-by: http://pandas.pydata.org/pandas-docs/stable/groupby.html
• Value-Counts: https://chrisalbon.com/python/data_wrangling/pandas_dataframe_count_values/
• Series.map: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.map.html
• Working-with-text-data: https://pandas.pydata.org/pandas-docs/stable/text.html

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

Good work, there are no repetitive code present.

But for meeting specifications you need to provide comments.

As a developer/ programmer, it is good practice to add comments and/or doc strings to your code. Comments and/or doc strings are important because they briefly tell the reader what type of variable, function or execution is being done within the code block. Comments and/ or doc strings do not need to be paragraphs, but a good sentence or two will do.
No matter how many lines of code it may be from one line of code to multiple lines of code, it is necessary.

This article is a very good read as it pertains to this section https://www.python.org/dev/peps/pep-0008/#comments

To meet specifications

```
Provide comments and/or doc strings for ALL of the code blocks where necessary withi
n the Python notebook
```

Quality of Analysis

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

This is one the important part of the project. In this section we have to mention all the questions in the Introduction section. This is necessary as this sets the tone for the whole project. By reading the questions in the beginning the reviewer/reader will have a better idea about the coming analysis and this is a good practice to have Questions which are you going to analyse in the beginning only.

So please give a brief intro and state your questions that you are going to analyze.

Data Wrangling Phase

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

Awesome work, I can see that you have found outliers in age variable and used various pandas functions like groupby, join. Keep it up.

However if you would have noticed the Handcap variable it contains 3 extra values, since it should have binary data those extra values are incorrect data. So always use describe and value count for checking outliers and number of rows for a variable.

Now some general suggestion for future work.

One main issue is having missing data while conducting analysis, which can provide skew/bias results. There are some strategies to deal with these issues:

• Identify the missing values within the dataset. (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.isnull.html)
• Drop the missing rows (http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.dropna.html)
• Replace missing values (http://pandas.pydata.org/pandas-docs/version/0.17.0/generated/pandas.DataFrame.fillna.html)
• If there are way too many missing values within a column it is best to drop the column completely. (http://pandas.pydata.org/pandas-docs/version/0.17.0/generated/pandas.DataFrame.drop.html)

Also you should consider to:

• Detect and exclude outliers (https://stackoverflow.com/questions/23199796/detect-and-exclude-outliers-in-pandas-dataframe, https://ocefpaf.github.io/python4oceanographers/blog/2015/03/16/outlier_detection/, https://www.kdnuggets.com/2017/02/removing-outliers-standard-deviation-python.html
)

• Groups continuous or numerical values into smaller groups or 'bins' (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.cut.html)
• Transforms categorical data into dummy/indicator variables (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html)

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.

To meet specification you need to explore at least 3 variables using both 1d and 2d charts.

This link contains good material on Differences between Univariate(1d) and Bivariate(2d) data.
http://www.math.kent.edu/~reed/Instructors/MATH%2010041/Ch4/Univariate%20vs%20bivariate%20data.pdf

So please make some charts in your project. You can refer the below link for chart suggestions and examples.

https://udacity-reviews-uploads.s3.us-west-2.amazonaws.com/_attachments/144663/1524172507/suggestion_plot.png

https://python-graph-gallery.com/

SUggestions: For 1st question you can plot bar chart with top 10 and bottom 10 neighbourhood.

For 2nd question you can make stacked bar chart for the 4 variables with Show/No-Show in X axis and bars containing the patients illness. Some random chart you can plot a bar chart of no of patients admitted w.r.t month.

And for 1d charts you can make histogram of age variable.

So please update this section.

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.

Same as above.

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.

Conclusion is not present.

This is the section where you will reflect on the question(s) asked at the beginning and state whether you have addressed the findings or not.

To give a thorough idea of how exactly the conclusion should be written:

```
Determine whether the questions asked at the beginning before the analysis can be an
swered after finishing the analysis. (Address your questions which you posted by sum
marizing the findings of each question!)

Provide the results of the exploration accurately, and note where additional researc
h can be done or where additional information could be useful to improve analysis.(R
eviewing statistical information from the visualizations can be very useful!)

Explain some of the limitations of this analysis.
Lastly, provide a reference
```

RECOMMENDATIONS (PART TWO)

To give you an idea to address the limitations or challenges you personally faced while doing this project, here are some factors to consider and can be addressed:

Common limitations would be : more missing values, imbalanced data, highly correlated having erroneous or missing data, sample not representing the population correctly. 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.

Communication

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

Good job, when you make the plots please explain the plots and why you made such plots.

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

Same as told in Exploration phase. Jut try to label each axes and give chart title for every chart.

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

Rate this review

Student FAQ