UDACITY

Logc

‹ Back to Data Analyst Nanodegree

Investigate a Dataset

REVIEW | HISTORY

Requires Changes

2 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.

Code is functional and working fine.

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 sui
ted for data analysis via its DataFrame and Series data structures in a c
oncise manner.

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

Concise and clear code: Its concise and clear API allows the user to focu
s more on the core goal at hand, rather than have to write a lot of scaff
olding code in order to 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.

Awesome work, you have made a lot of functions in your project and all of them had some complex calculations in them, nice work. Functions were made for plotting charts also. Keep it up!!!

Quality of Analysis

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

Good job writing a brief introduction and stating your questions, Just a suggestion it would be nice if you highlighted your questions in bullet point.

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, you have used many pandas functions like merge, group by, sort_values, reset_index() etc. Searched for NaN values in the data set, dropped irrelevant columns. You have even used lambda function also. You could have searched for duplicate values also. The data wrangling process clearly shows that you had put a lot of time and effort into it. Keep it up!!!.

Some general suggestions.

The most important aspect of Data Wrangling is to clean or transform the data preparing it for analysis.

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.

You have only plotted 1d charts. For meeting requirements you need to plot 2d charts also. Both boxplots and pie chart are 1d charts. Try to plot bar charts, line chart, scatter chart, heat map etc.

And , I would like to note that in one of the functions you generated only cells or tables. Well cells are a very bad way to show data. Always try to use charts or visualization to show data. Its always easy for the end user to see and understand data from charts rather than understanding it from a table.

But I must say, I am very much impressed by the way you made the charts from a function. Its very complex though, I must admit. That's some great work out there.

Since you are practicing and learning here's a link to get knowledge about other charts too.

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

So make the changes and submit the project again.

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.

Good work, but I suggest that you study other chart types too.

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 was written addressing each question, nice job.

However for meeting specification you need to write about the limitations also.

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.

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

All charts were labelled properly with appropriate chart titles.

RESUBMIT PROJECT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH