UDACITY

‹ Return to "Data Analyst Nanodegree" in the classroom

Investigate a Dataset

REVIEW

HISTORY

Meets Specifications

Dear student:

Congratulations! You approved this project!

I want to say thank you for the good job you have done. You have built a very good report. It is easy to understand. It is organized and shows very interesting findings.

Always keep in mind your audience. Ask yourself what your audience need, and work to answer that question. It is very important to create easy to read and understandable documents. Describe each step, analysis, or plot because it will help your audience to understand what and why you are doing that. Finally, as a Data Scientist, you must take care of all the details.

Continue working with the same enthusiasm and dedication in the next projects I am sure that you will be an excellent Data Scientist.

Thank you again for your commitment to excellence,

Keep learning, keep working!

Code Functionality

✓

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

✓

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 Data Frames, 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 data analysis via its DataFrame and Series data structures in a concise 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 focus more on the core goal at hand, rather than have to write a lot of scaffolding 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.

Excellent job!

You constructed reusable code blocks using functions.

COMMENTS:

"Reusing code is key to building a maintainable system, and when it comes to reusing code in Python, it all starts and ends with the humble function." (https://www.safaribooksonline.com/library/view/head-first-python/9781491919521/ch04.html)

Quality of Analysis

✓

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

Good job with your questions!

COMMENTS:

Either you're given data and ask questions based on it, or you ask questions first and gather data based on that later, great questions help you focus on relevant parts of your data and direct your analysis towards meaningful insights. Questions should be measurable, clear and concise. They should be designed to either qualify or disqualify potential solutions to your specific problem or opportunity.

Data Wrangling Phase

✓

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

Good job!

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. It is about targeting a field, row or column in a data set, and applying an action such as joining, parsing, cleansing, consolidating, or filtering to create the desired output, which will then be used down the road. Data wrangling involves activities like:

- Remove unused columns.
- Remove duplicate rows.
- Change data formats (date columns)
- Discard missing values.

Benefits:

- it makes your data useful
- it can be organized into a standardized and repeatable process that moves and transforms data sources into a common format, which can be reused multiple times.

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.

Well done!

You did an extraordinary analysis joining both single-variable and multiple-variable explorations in your work.

COMMENTS:

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set. The graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1) Plotting the raw data such as histograms, bihistograms, probability plots, lag plots, block plots, scatter plots.
2) Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.

Remember:

What is very important when you analyze data is to stay focused on your questions. Build plots or statistical summaries which answer your questions, and not just because they are nice.

✓

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 job!

COMMENTS

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.

Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments.

Data visualization can also:
• Identify areas that need attention or improvement.
• Clarify which factors influence customer behavior.
• Help you understand which products to place where.
• Predict sales volumes.

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.

Awesome job!

The results of the analysis are presented such that any limitations are clear.

COMMENTS:

In the data analysis process, a well-crafted conclusion can provide the final word on the value of your analysis. It shows readers the value of your completely developed argument or thoroughly answered question. Some aspects you must consider about conclusions are:

- Build the conclusion from the reader's perspective.
- Summarize the main points you made in your introduction (questions).
- Review (very briefly) the research methods and/or design you employed.
- Repeat (in abbreviated form) your findings.
- Discuss the broader implications of those findings.
- Mention the limitations of your research (due to its scope or its weaknesses)
- Offer suggestions for future research related to yours.

Communication

✓

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

Well done!

It is very important to communicate the results adequately; however, it is also very important to describe each activity, analysis, or graph. This will allow your audience to understand what you are doing and how you are doing it. Moreover, the reasoning makes your work organized, formal, and sophisticated.

✓

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

Good job!

COMMENTS:

All visualizations are properly labeled, titled and have legends where necessary that depict the data correctly.

One of the most important steps in creating an impactful visualization is making sure all of its elements are labeled appropriately. The text components of a graph give your reader visual clues that help your data tell a story and should allow your graph to stand alone, outside of any supporting narrative.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

☆☆☆☆☆