Wrangle and Analyze Data

REVIEW | HISTORY

Meets Specifications

Dear Excellent Student,
This was a great piece of work and tells a ton about the type of person you are - organized, hardworking, and quality oriented. 👏🏽 Going thoroughly through the work, I could see a lot of time and effort invested in the project, and i think this is commendable. I exhort you to keep up this good work as it will make you an outstanding Data Analyst. Keep learning with us here in Udacity! 🖖🏽

Code Functionality and Readability

✓ **All project code is contained in a Jupyter Notebook named wrangle_act.ipynb and runs without errors.**

Your code runs without errors. As a whole, you've done a fantastic job of developing coding solutions throughout the project.

Learning Notes

I am a fan of using shortcuts with Jupyter Notebook. Check out this medium post on **Jupyter Notebook Shortcuts**.

✓ **The Jupyter Notebook has an intuitive, easy-to-follow logical structure. The code uses comments effectively and is interspersed with Jupyter Notebook Markdown cells. The steps of the data wrangling process (i.e. gather, assess, and clean) are clearly identified with comments or Markdown cells, as well.**

Great job on making the work very clear and easy to follow. Markdown cells are used to clearly indicate different sections of the notebook. Apart from markdown cells, the project also makes use of inline python comments to make the work clearer. Nice!

Learning Notes

It is also good practice to use functions to avoid any code repetition.

- Why use functions in programming?

Gathering Data

✓ **Data is successfully gathered:**

- **From at least the three (3) different sources on the Project Details page.**
- **In at least the three (3) different file formats on the Project Details page.**

Each piece of data is imported into a separate pandas DataFrame at first.

Awesome, you have successfully gathered data from at least three different sources on in at least three different file formats on the project details page. Each piece of data is imported into a separate pandas dataframe at first.

Data was gathered from the following sources:

- Twitter-archive **CSV** file.
- Image prediction **TSV** file.
- **tweet_json.txt** file containing retweet counts and likes.

Assessing Data

✓ **Two types of assessment are used:**

- **Visual assessment: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes. Once displayed, data can additionally be assessed in an external application (e.g. Excel, text editor).**
- **Programmatic assessment: pandas' functions and/or methods are used to assess the data.**

Well done! Both visual and programmatic assessments are used in the notebook and the results are well documented.

✓ **At least eight (8) data quality issues and two (2) tidiness issues are detected, and include the issues to clean to satisfy the Project Motivation. Each issue is documented in one to a few sentences each.**

Good work identifying below quality and tidiness issues in the dataset:

Learning Notes

Data assessment involves examining data quality and tidiness.

Quality issues pertain to the content of data. Low quality data is also known as dirty data. There are four dimensions of quality data:

- **Completeness:** do we have all of the records that we should? Do we have missing records or not? Are there specific rows, columns, or cells missing?
- **Validity:** we have the records, but they're not valid, i.e., they don't conform to a defined schema. A schema is a defined set of rules for data. These rules can be real-world constraints (e.g. negative height is impossible) and table-specific constraints (e.g. unique key constraints in tables).
- **Accuracy:** inaccurate data is wrong data that is valid. It adheres to the defined schema, but it is still incorrect. Example: a patient's weight that is 5 lbs too heavy because the scale was faulty.
- **Consistency:** inconsistent data is both valid and accurate, but there are multiple correct ways of referring to the same thing. Consistency, i.e., a standard format, in columns that represent the same data across tables and/or within tables is desired.

Tidiness issues pertain to the structure of data. These structural problems generally prevent easy analysis. Untidy data is also known as messy data. The requirements for tidy data are:

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

Cleaning Data

✓ **The define, code, and test steps of the cleaning process are clearly documented.**

The different steps of the cleaning process are clearly documented. We have the `define`, `code` and `test` steps which are clearly stated with some explanations of what process you intend to do at each level. It is a good initiative to, first of all, begin by copying the data into separate files before manipulating with them. Excellent work!

✓ **Copies of the original pieces of data are made prior to cleaning.**

All issues identified in the assess phase are successfully cleaned (if possible) using Python and pandas, and include the cleaning tasks required to satisfy the Project Motivation.

A tidy master dataset (or datasets, if appropriate) with all pieces of gathered data is created.

Indeed, copies of the original pieces of the data are made before the cleaning process.

Storing and Acting on Wrangled Data

✓ **Students will save their gathered, assessed, and cleaned master dataset(s) to a CSV file or a SQLite database.**

The cleaned master datasets are saved to a `CSV` file. Good work!

Learning Notes

- Check out this StackOverflow Thread on **Pandas writing dataframe to CSV file**.
- Also, take a look at the pandas.DataFrame.to_sql and this Stackoverflow thread for an example of saving pandas dataframe to SQLite.
- Saving a pandas Dataframe as a CSV

✓ **The master dataset is analyzed using pandas or SQL in the Jupyter Notebook and at least three (3) separate insights are produced.**

At least one (1) labeled visualization is produced in the Jupyter Notebook using Python's plotting libraries or in Tableau.

Students must make it clear in their wrangling work that they assessed and cleaned (if necessary) the data upon which the analyses and visualizations are based.

Good work analyzing the cleaned data! The notebook contains four separate insights which are well described in the report and in addition we have visualizations produced of the data.

Learning Notes

- Check out the python visualization documentation for various ways of visualizing data.
- There are several other ways to visualize data including **Box plots**, **Line graphs**, **Pie charts**. Check out this documentation on **Visualization with Seaborn**.

Report

✓ **The student's wrangling efforts are briefly described. This document (wrangle_report.pdf or wrangle_report.html) is concise and approximately 300-600 words in length.**

The write-up (`wrangle_report.pdf`) describing the wrangling efforts made in the project with the use of 3 steps which are **Gathering Data**, **Assessing Data** and **Cleaning Data** is very detailed and within the limit required. You've produced a clear summary of your work in an 'executive report' style. It looks clean and professional. Awesome job!

Learning Notes

How To Write A Great Report: 7 Tips To Make Your Next Report Stand Out.

✓ **The three (3) or more insights the student found are communicated. At least one (1) visualization is included.**

This document (act_report.pdf or act_report.html) is at least 250 words in length.

Three interesting insights have been reported in the `act_report.pdf` and they have been analyzed in details. All information is well communicated and the write-up is more than 250 words in length. Also, good work including visualizations in the report. This made the insights very clearly communicated. Your report presents the insights in a clear and engaging way. Well done! 👍🏽

Project Files

✓ **The following files (with identical filenames) are included:**

- **wrangle_act.ipynb**
- **wrangle_report.pdf or wrangle_report.html**
- **act_report.pdf or act_report.html**

All dataset files are included, including the stored master dataset(s), with filenames and extensions as specified on the Project Submission page.

All required files are present. The submission contains the `wrangle_act.ipynb` with all the necessary code, the `wrangle_report.pdf` which contains a brief discussion of the wrangling efforts made in the project with the use of 3 steps which are Gathering Data, Assessing Data and Cleaning Data. Finally, we also have the `act_report.pdf` which is mainly a discussion on the insights discovered in the project. Brilliant!

DOWNLOAD PROJECT

RETURN TO PATH