Wrangle and Analyze Data

REVIEW | HISTORY

Requires Changes

1 SPECIFICATION REQUIRES CHANGES

You have done really well in your initial submission for this project, well done. There is one more update that needs to be incorporated before we can mark this project as passing. Good luck with your next submission!

Code Functionality and Readability

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

✓ **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.**

Code blocks in the notebook are properly documented, good work here.

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.

Data were successfully gathered from three different sources and each piece of data was imported into a separate object at first. Good work.

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.**

Great job doing both visual and programmatic assessments properly and documenting the process in the Jupyter notebook.

🔄 **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.**

Regarding these items:

> - column names p1, p1_conf, p1_dog... should be more descriptive
> - id column should be named twitter_id
> - not needed columns: in*reply_to*.... , retweeted_...

Following the Hadley Wickham's Tidy Data description, tidy data is a standard way of mapping the meaning of a dataset to its structure. Issues that only requires you to remove rows or columns or to update values or data types to fix them and does not require you to change the structure of the columns are hence should be considered Quality issues.

Please move them to the Quality issues section.

Check out these rules of tidy data. There are two interesting points to take here:

Firstly, tidy data should have the following characteristics:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

And here are five most common problems with messy datasets (read up on the article for ways to correct them):

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

Hence, there are two prominent tidiness issues in this project:

1. Information about one type of observational unit (tweets) is spread across three different dataframes. Therefore, these three dataframes should be merged as they are part of the same observational unit.
2. Dog stages should be a single column rather than four; one of the requirements for tidy data is that each variable forms a column.

Cleaning Data

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

✓ **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.

DataFrame objects were copied before cleaning, and a final cleaned dataset was created and filled with the cleaned data. All the important issues have also been cleaned, excellent work here.

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 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.

Master dataset has been properly analyzed and several insights have been produced. Visualizations are included in the report. Excellent work, here.

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 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.

The act report document is well-written and it is such a pleasure to read, 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.

RESUBMIT PROJECT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH