

[< Back to Data Analyst Nanodegree](#)

Wrangle and Analyze Data


REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Greetings Student,

This is 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 and remain dacious.

Please, do not forget to rate this review and give a comment about the challenges encountered in this project to help us improve our work. Thank you and good learning at Udacity. 😊

Code Functionality and Readability

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

Good work! All cells of the notebook run on my end without errors.

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.

A remarkable job was done to organize the notebook in an easy-to-follow format. The steps of data wrangling (gather, assess, and clean) were well identified with markdown cells. Good work!

Suggestions and Comments

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

- [Why use functions in programming?](#)

Inline comments could also help with code follow up and ease the work of other programmers working on the same project.

- [Why should we comment code?](#)

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 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.

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 quality issues and tidiness issues in the dataset.

Quality Issues:

tweet_archive_df

- tweets archive includes retweets and replies, which are less curated data
- datetime column timestamp is a object type (string)
- source contains HTML fragments, but the column isn't needed
- DataFrame contains extraneous variables to analysis
- numerator extraction of decimals extracted the number behind the '.' as the whole number rating
- numerators/denominator pairs with denominator above 10 are for multiple dogs
- denominators that aren't divisible by 10 are invalid
- 'name' column contains non-dog names
- manual fixes:
 - tweet_id = 855862651834028034 Snoop Dogg is not a real dog

image_predictions_df

- tweeted pictures aren't always of dogs

tweet_json_df

- missing entries from the api versus the id's present in the tweet archive
- too much information; a lot of trash fields

all files

- tweet_id has an inconsistent data type

Tidy Issues:

tweet_archive_df

- doggo, floofer, pupper, and puppo columns should be one variable column

image_predictions_df

- many columns leading to most probable dog breed

tweet_json_df

- multiple data in several columns
 - JSON tree includes compound fields with another dictionary tree
- contains duplicate data fields as 'tweet_archive_df'

all files

- data should be consolidated into one DataFrame

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.

Suggestions and Comments

- Please make sure to checkout [StackOverflow-why should I make a copy of a data frame in pandas](#)

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 dataset is saved to a `csv` file. Good work!

Suggestions and Comments

- 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.

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 and plotting some visualizations of the data.

Suggestions and Comments

- Check out the [python visualization documentation](#) for various ways of visualizing data.
- 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 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.

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.

The required files are included. Excellent!!!

 [DOWNLOAD PROJECT](#)

[RETURN TO PATH](#)

[Rate this review](#)

[Student FAQ](#)