Wrangle and Analyze Data

REVIEW | HISTORY

Requires Changes

3 SPECIFICATIONS REQUIRE CHANGES

Hi Sebastian,

Nice work on the project and it's almost complete. There are a couple of minor changes which should not take much time to complete. Detailed feedback is provided for them in the cleaning and assessment section.

I am sure the next submission will pass.

All the best and keep learning.

Code Functionality and Readability

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

All the code is present in the `wrangle_act.ipynb` notebook and run without errors. Good work on checking that every cell works correctly.

✓ **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.**

It's great to see that you have organized the notebook in the 4 distinct sections of GATHER / ASSESS / CLEAN and ANALYZE. The notebook is interspersed with code and markdown text. This helps anyone in following along the work and can also understand the process flow that you have taken.
Nice job.

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 is successfully gathered from all the 3 sources and is saved to file locally.

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 done in the notebook.

Nice job on using the functions like `info(), describe(), value_counts()` to explore more about the data.

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.

Issue list has been identified in the notebook. Each issue is classified into Quality and Tidiness.
The issues are described in 1-2 sentences. Awesome. Most of the issues have been identified.

For the tidiness issue of merging the API data to twitter archive, the image predictions dataset should also be merged as that is also part of the same observational unit (dog info tweets).
So the tidiness issues should be to merge all 3 datasets into one.

Cleaning Data

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

The DEFINE / CODE / TEST steps are clearly documented in the cleaning section. This helps us a lot in identifying each issue and how it's cleaned and tested.

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.

Copies of the dataset are made prior to cleaning. That's an important step as we may need to refer to original dataset later on.

Most of the issues are cleaned properly. Awesome work for ratings. Many students miss this issue in first submission and you nailed it.

There are just 2 required changes

Required Changes:
For the issue

For the issue

> - Convert doggo, floofer,pupper, puppo to stages

You have correctly created one column and put the stages in that but some tweets have multiple stages present in them. These multiple stages should be saved separated by comma or using the word `multiple`

Eg. in tweet id `855851453814013952`

```
In [102]: 1 twitter_archive[twitter_archive.tweet_id == 855851453814013952]
```

Out[102]:

d_status_timestamp	expanded_urls	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
NaN	https://twitter.com/dog_rates/status/855851453814013952/photo/1	13	10	None	doggo	None	None	puppo

we see that `doggo` and `puppo` column both have values. So after cleaning the `stage` column should have values delimited by comma like `doggo, puppo` or `mulitple` . Any of these are acceptable.

Merge all 3 datasets into 1 as a tidiness issue.

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 gathered/cleaned data is saved to a CSV file.

To remove writing the row number to CSV File it is recommended to use `index=False` argument in `to_csv()` function.

Eg. `df.to_csv(filename, index=False)`

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

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.

The master dataset is analyzed using the pandas and insights and visualizations are given.
Good job.

Please redo the same analysis after completing the Assessment and Cleaning.

This section is marked as Require Changes because based on new Assessment and cleaning the insights may change. So once both assessment and Cleaning sections pass, this will be evaluated.

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.**

Good work on creating the report for the wrangling efforts. It's clear and concise and reflects the wrangling process taken for the data set.

Since the report is based on the Assessment and Cleaning process so please review the report again after completing those sections and make the necessary changes.

✓ **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 Analysis report is present and the insights, visualizations are communicated. Multiple visualizations are present.

Since the report is based on the Assessment and Cleaning process so please review the report again after completing those sections and make the necessary changes.

Beautiful to add dog images. It makes the report fun to read.

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

Wrangle and Analyze Data

REVIEW | HISTORY

Meets Specifications

Good work!
You have provided a wonderful project :)
However, to improve your skill and knowledge for your journey of being an aUdacious data analyst, here I provide you with some comments

Code Functionality and Readability

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

No error found, well done!

- If you want a very succinct cheat sheet for data wrangling using python, I think this will be very helpful for you :)

✓ **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.

I love how you have structured your project and commented all complex code structures. This is a very good practice. As a reviewer myself, I found it very helpful to understand the code and how the code produces a correct/wrong result. In a workplace, such clear structure and a well-documented code will be very helpful for colleagues that might be continuing your work or learning from your work.

Keep doing the good practice!

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.

You have included the three data sources correctly with the correct methods :)

However, regarding the API dataset, as you have chosen to use the provided tweet-json.txt, it is required to copy and paste the provided code in the project brief. There you have to modify a variable to match your code.

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.**

✓ **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.

All issues are correctly mentioned :)

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.

You have copied the original dfs before you clean it :)
This is a good practice.
For further information why is it so important, please read this.

You also have cleaned all mentioned issues and merge the tables into a master table :)

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

which the analyses and visualizations are based.

Your analyses are perfect :) You also have provided some visualisations

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.

DOWNLOAD PROJECT

RETURN TO PATH