Requires Changes

6 SPECIFICATIONS REQUIRE CHANGES

Good start here! There are a couple of updates that are still required before we can pass this project, but they are all quite straightforward to implement. This is a challenging project, so please do not be discouraged for not passing it on your first try.

You may also go to the Student Hub to get some assistance from your fellow students and mentors.

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.

Please ensure to follow the recommended steps in your wrangling process:

Gather -> Assess -> Clean -> Store -> Analyze

Currently, you cleaned the dataset right after you have gathered data from each data source. It makes the analysis harder to follow.

Relevant lesson:
Part 8. Data Wrangling, Lesson 1. Introduction to Data Wrangling, Sublesson 23. Quiz: Clean (Test)

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.

Joining the tables should be a part of Tidiness issues

I noticed you have joined the tables. This is a solution to one of the tidiness issues as described in the Hadley Wickham's Tidy Data description: 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.

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. Excellent work on this part. Please correct these issues to pass this specification:

Incorrect values in rating numerators - Quality issue

Rating numerators have not been properly cleaned. The current pipeline captures incorrect values when rating numerators contain decimals. For example, here is a value from one observation with tweet id 786709082849828864:

"This is Logan, the Chow who lived. He solemnly swears he's up to lots of good. H*ckin magical af
9.75/10 https://t.co/yBO5wuqaPS"

Currently, the value 75 would be captured as the rating numerator. Try to capture the entire value from the text instead. Here is a code snippet as an example, where df here is the twitter archive dataset:

```
ratings = df.text.str.extract('((?:\d+\.)?\d+)\/(\d+)', expand=True)
```

ratings series object will then contain all rating numerators with decimals and rating denominators (without decimals). The next step is to extract only the numerators and denumerators from ratings dataframe, and then update your dataset's fields with extracted rating numerators and denominators (**NOTE: Do not forget to convert the field datatype into Float, astype function may be used here**):

```
df.rating_numerator = ratings
```

To improve it even further, you may also want to try adjusting the code so rating denumerators would also capture decimal values.

I found tools such as this one to be helpful in finding the correct regex.

Retweets need to be removed - Quality issue

Retweets need to be removed to avoid duplication in our analysis. This may be done by removing rows that have non-empty retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp. When this step is correct, there should be a fewer number of non-empty tweet ids.

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.

Once you have properly cleaned the issues as recommended above, please redraw the visualizations and update your insights to reflect the new data.

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.

Please update your wrangle report document to reflect the changes from my comments in the above specifications.

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.

Please redraw the visualizations to reflect the new data due to the above corrections and update your report.

(Optional) We suggest including pictures for aesthetic and additional context purposes on top of the required visualizations. Example: include a screenshot of a specific tweet, a specific breed of dog, etc. Anything to get the reader engaged. Picture this report like a blog post or magazine article; we want people to be engaged and have fun while reading.

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.