UDACITY

< Return to "Data Analyst Nanodegree" in the classroom

DISCUSS ON STUDENT HUB

Wrangle and Analyze Data

REVIEW

HISTORY

Requires Changes

3 SPECIFICATIONS REQUIRE CHANGES

Dear Student,
You have done an excellent job wrangling the given data for the most part and producing some interesting insights. However, you need to work a little more on this project to meet all the specifications. Since you have already addressed most of the requirements, it is just a matter of paying attention to some finer details (see my comments below). I am sure you will be able to quickly get this project to meet all specifications as you have a very good python coding skills and understanding of data wrangling process. After you make changes, please use the project rubric below to review your project before resubmission.

Good luck with your resubmission. Looking forward to seeing your resubmission!

Note: to pass this project, you need to only address issues marked as *Required*. The issues marked as *Suggested* are optional and you do not need to address them to pass this project. But if you address these suggested issues, it will improve your project.

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.

Good job clearly identifying the steps of the data wrangling process in markdown cells. The notebook is structured well. This helps to easily follow your 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.

Excellent job successfully gathering data from local file 'twitter_archive_enhanced' and from a URL ('image_predictions.tsv') and imported them into separate pandas dataframes.

Suggested:

You have used `tweet_json.txt` provided in the supporting material in the project instruction. It is fine as far as completing the project for this nanodegree is concerned. However, I strongly encourage you to query twitter API and gather data by yourself if possible as it is an invaluable skill.

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.

Suggested:

You can also use describe() function, which gives you descriptive statistics, to assess quantitative variables.

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.

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.

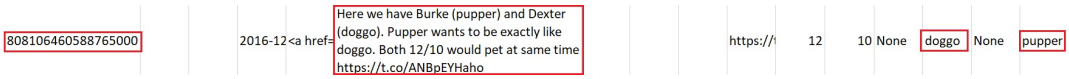
Good job copying all the dataframes prior to cleaning. If you want to know more about why it is important to copy the dataframes please see the following link; https://stackoverflow.com/questions/27673231/why-should-i-make-a-copy-of-a-data-frame-in-pandas. Copying is also important if at some point you need to trace back on your steps.

Required:

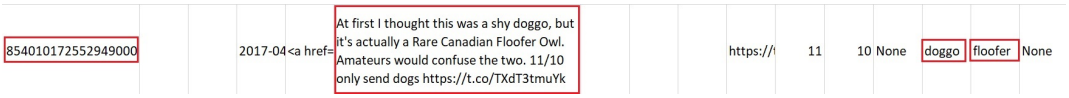
Quality:
You only want original dog ratings (no retweets) that have images (a user can retweet their on tweet), so you need to remove all rows that have values (not blank or non-null) in retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp columns. As mentioned in the **Key Points** section of **Project. Wrangle and Analyze Data - Sublesson 2. Project Motivation** (link), this is an important quality issue to be addressed, as it has a direct bearing on your analysis.
Just removing the columns like retweeted_status_id is not going to address the retweet issue. You can remove these columns only after removing rows with retweets.

Suggested:

Did you notice that certain tweets have more than one stage. For instance take a look at the following tweet, where both doggo and pupper is present.



This is because some tweets may have more than one dog with different stages (https://twitter.com/dog_rates/status/808106460588765185/photo/1). When there are multiple stages for a tweet (e.g., doggo and pupper) like this, your code capture only one stage. Instead you should capture all the stages as a list delimited by comma (e.g., 'doggo, pupper'), or you can capture them something like 'multiple_stages'. However, there is one more issue. In certain cases, although there are more than one stage, if you look at the text, there is supposed to be only one stage. For example, take a careful look at the following tweet;



This is supposed to be floofer, but it is captured as doggo and floofer, which is wrong. So if you want to clean this column perfectly, you may have to do some manual cleaning. This shows how data wrangling can get really complicated on occasions.

Removing rows with denominator != 10 without further inspection is not a good idea. For instance, take a look at this tweet https://twitter.com/dog_rates/status/704054845121142784; in this case the denominator is intended to be 50, because there are 5 puppers. Project instruction says denominator is **almost always** 10, it does not say it is **always** 10.

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.

Suggested:

Please note that in your master dataset saved to csv, an unwanted index column is added. To avoid this you need to set index argument in to_csv function to False as `pd. 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.

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.

Required:

Although you have produced some interesting insights, your analysis and visualizations are based on partially cleaned data (retweets not removed) as mentioned above. So please repeat the analysis and visualization section after you remove retweets.

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

visualization is included.

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

Required:

Although your report is interesting, your analysis and visualizations are based on partially cleaned data (retweets not removed) as mentioned above. So please rerun the analysis and visualization section after you remove retweets and **then update this report**.

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

DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

RETURN TO PATH

Rate this review