UDACITY

‹ Return to Classroom

Wrangle and Analyze Data

REVIEW

HISTORY

Meets Specifications

Dear Student,

You have put dedicated effort into this project and it paid off. Congratulations on meeting all the specifications of the project! You have demonstrated a very good python coding skills and understanding of data wrangling process. You also did a fantastic job incorporating suggestions from **the previous reviewer. As a different reviewer, I have left some additional comments.** If you are uploading this project to your portfolio or sharing it with your potential employer, it is a good idea to address these comments. It also gives you an opportunity to appreciate the complete essence of this project. Keep up all the great work you are doing.

Good luck with your future projects!

Note: I made the comments marked as **Suggested** to help you improve the project. It does not require you to resubmit the project. You have already passed the project. Congratulations!

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.

Suggested:

Since it is a long notebook, a Table of Contents (TOC) would really help. Here is how to add a table of contents; The following markdown cell creates a TOC.

```
## <font color='blue'>Table of Contents</font>

* [Gather](#step0)
* [Assess](#step1)
* [Clean](#step3)
## Add more contents as necessary
```

To link the headers to TOC, in each header you need to add a tag as follows;

```
<a id='step0'></a>
## <font color='blue'>Gather</font>
```

Please also see this link; https://medium.com/@sambozek/ipython-er-jupyter-table-of-contents-69bb72cf39d3.

Gathering Data

Data is successfully gathered:

- **From at least the three (3) different sources on the Step 1: Gathering Data page.**
- **In at least the three (3) different file formats on the Step 1: Gathering Data 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. You also did a great job querying Twitter's API to gather data for all the available twitter ID and importing it into a pandas dataframe, where many students struggle.

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.

Suggested:

One of the quality issues you cleaned is,

- Change row whose denominator rating is 0 to 1000

What is the logic behind changing the denominator to 1000? Do you mean 10? Most of other denominators are 10.

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.

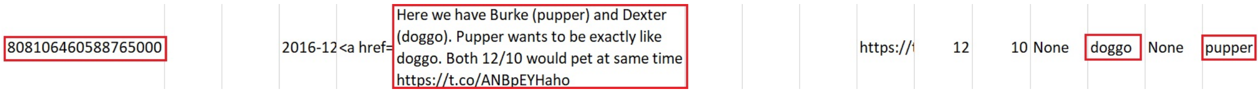
All issues identified in the assess phase are successfully cleaned (if possible) using Python and pandas, and include the cleaning tasks required to satisfy the Project Motivation.

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.

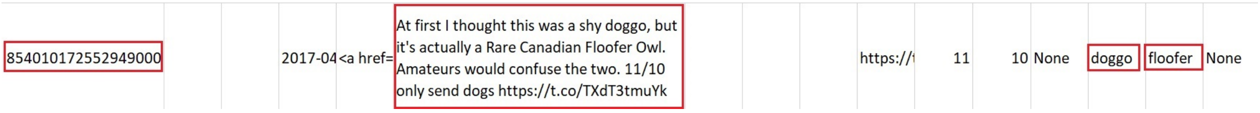
Suggested:

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.



808106460588765000			2016-12	<a href=	Here we have Burke (pupper) and Dexter (doggo). Pupper wants to be exactly like doggo. Both 12/10 would pet at same time https://t.co/ANBpEYHaho				https://t	12	10	None	doggo	None	pupper

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;



854010172552949000			2017-04	<a href=	At first I thought this was a shy doggo, but it's actually a Rare Canadian Floofer Owl. Amateurs would confuse the two. 11/10 only send dogs https://t.co/TXdT3tmuYk				https://t	11	10	None	doggo	floofer	None

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.

You removed the text column. It is not a good idea to remove text column as it contains lot of useful information and it is kind of heart of this dataset.

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.

In the following line of code, you have done a good job using index argument in to_csv function and setting it to False to avoid adding a unwanted index column in the saved file.

```
combined_clean_data.to_csv('twitter_archive_master.csv', 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.

Suggested:

Your code for analysis and visualizations should be in wrangle_act.ipynb, not in a separate file. Please see the section **Storing, Analyzing, and Visualizing Data for this Project** in the lesson **Project. Wrangle and Analyze Data - Sublesson 3. Project Details**.

You have only used bar chart in this project. That is fine as far as completing this project. But it is a good idea to know about different kind of charts one can use to represent different kinds of insights.
The following link gives a good guideline as to when to use what chart; http://www.infographicsblog.com/chart-suggestions-a-thought-starter-andrew-abela/.

It is very easy to plot different kinds of charts with python seaborn libray. You can take a look at it in this link; https://seaborn.pydata.org/tutorial.html. Data visualization is one of most valueble skills for data scientists/analysts.

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.

Suggested:

In this report, you identified the following issue as a tidiness issue. This is not a tidiness issue.

- The number of tweets_id for each table is different

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

specified on the Project Submission page.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START