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 have done an excellent job wrangling the given data and producing some interesting insights like **iphone is the most frequent platform for tweeting on this account**

You also did a fantastic job of incorporating the **previous reviewer** suggestions. **As a different reviewer,** I have left some additional comments. I made these comments marked as Suggestions to help you improve the project. It does not require you to resubmit the project. You have already passed the project. **Congratulations!** 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!

Here are a few resources that may help your continued learning:

- A critical skill for data scientist/analyst is data visualization. With python seaborn library, you can plot many different kind of charts apart from the ones you plotted. You can take a look at it.
- Having developed a wide array of data wrangling skills, the next thing you may want to learn more about is predictive analytics. Luckily, there is a free machine learning course available with this nanodegree. You can see a lesson **Intro to Machine Learning** in EXTRACURRICULA section. You can take a look at it.
- PEP8 is the style guide for python. This style guide provides guidelines and best practices on how to write Python code to improve the readability of code and make it consistent across the wide spectrum of Python code. You can take a look at this guide here; https://www.python.org/dev/peps/pep-0008/ and should strive to adhere to these guidelines.

Code Functionality and Readability

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

Good job adding a **hyper-linked Table of Contents** so that it is very easy to navigate through your notebook. This shows your attention to details.

Excellent job writing functional code, executing the code and displaying the output without any 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. A good notebook structure also makes code **maintenance** easier.

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.csv` and from a URL (`image_predictions.tsv`) and imported them into separate pandas dataframes.

Suggestions

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.

✓ 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.

Suggestions

You identified **the presence of more than one dog stage** within a row/tweet as a tidiness issue. This is not an issue, because some tweets may have more than one dog with different stages. That is the reason why you captured multiple stages as a list of different stages delimited by comma (e.g., 'doggo, pupper') right?

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 this stack overflow thread. Copying is also important if at some point you need to trace back on your steps.

Nice job **capturing all stages of dogs** when an image has dogs with different stages. Many students miss this critical issue. Nice job digging deep into data and identify this issue.

Good job correctly extracting **decimal numerators**.

Good job **removing retweets** and retaining only original tweets as per project instruction as we are only interested in original tweets in our analysis.

Suggestions

> Data type of tweet_id column is int64 instead of category

ID columns like `tweet_id` **should be of** `object` **type (i.e. string), NOT** `category` . As each `tweet_id` is unique there is no real advantage of using `category` datatype.

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.

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.

✓ 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.

Suggestions

You used **only pie chart** in this project, which is fine as far as completing this project **as project requires you to create just one visualization**. But it is a good idea to know about different kind of charts one can use to represent different kinds of insights. Here are some resources that help you choosing right kind of visualizations to represent various types of data/insights.
Choosing the right chart type for your insight.
How to pick the right chart type.
When to use line chart vs area chart.
The difference between a bar chart and a histogram.
Why pie charts are not an ideal choice in most cases.
Quickly plot correlation between multiple variables is pandas using scatter matrix.

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.

You have done an excellent job producing this very interesting report explaining the insights you gained from your analysis.

Suggestions

This report should be like a **blog post** or **magazine article**. Hence, you can also include a screenshot of a specific tweet, a specific breed of dog, etc. (anything to get the reader engaged).

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