UDACITY

PROJECT

Wrangle and Analyze Data

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Meets Specifications

Excellent work passing this project on your first submission. This was a challenging project, but you managed to push through. Congratulations, and good luck with your next Nanodegree 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.

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

- 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 the visual assessment 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.

When extracting rating numerators, a more robust way to do this is by using regex. You may try the following (where `df` here is the twitter archive dataset):

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

`ratings` dataframe will then contain all rating numerators with decimals and rating denominators (without decimals). You may then update your dataset's fields with extracted rating numerators and denominators. To improve it even further, you may also want to try adjusting the code so rating denumerators would also capture decimal values.

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.

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. All the important issues have also been cleaned, excellent work here.

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 cleaned dataset has been saved to a csv file, good work. When saving to csv, we recommend removing the indexes to avoid adding an "Unnamed" index column to the dataset. Setting `index` parameter to `False` i.e. `to_csv(filename, index=False)` would help here.

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.

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.

The act report document is really well-written and it was such a pleasure to read, well done.

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. Frame this report as a blog post or magazine article; we want people to be engaged and have fun while reading.

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.

DOWNLOAD PROJECT

RETURN TO PATH

Student FAQ