UDACITY

Return to Classroom

Wrangle and Analyze Data

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Congratulations on completing the project! It's true that data wrangling is seen as a somewhat tedious task, but I hope you have found this project engaging, interesting, and fun. You've displayed good understanding of concepts.. excellent!

Distinction between data quality and data tidyness

```
QUALITY
Completeness. i.e: Do we have all of the records that we should? Do we have missing rec
ords or not? Are there specific rows, columns, or cells missing?
Validity. i.e: Do they conform with real-world constraints? (e.g. negative retweets, fa
vorites, etc.)
Accuracy issues. i.e: Incorrect data like denominator ratings greater than 10 for examp
le
Consistency. i.e: Make sure the columns all have a standard format. in your case detail
s about dog tweets and metrics
TIDINESS
Making sure every column is a variable
Making sure every row is an observation
```

```
Making sure every cell is a single value
```

Code Functionality and Readability

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

All code is running fine

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.

What went well:

- You're thinking in the right direction by exploring the datasets and then trying to identify/address the issues
- notebook is organised properly...

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.

You successfully gathered data from three dierent sources: local file (xls), URL, and API. You have also stored correctly the gathered data in a format according the project instructions.

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.

You have correctly used functions like info, sample, tail, isunique, duplicated, describe, value_counts.... to perform the assessment of data. In general, it's always a good idea to use another application (like Sheets, Excel, etc) to perform a quick visual assessment.

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.

Good work identifying data quality and data tidyness issues.. IMO creating a new column rating belongs to more of data tidyness issue

---

Suggestions:

- One way for handling/cleaning rating_numerator and rating_denominator is

```
import re

regex = r'''([+-]?([0-9]+[.])?[0-9]+\/[+-]?([0-9]+[.])?[0-9]+)'''

          #[+-]?([0-9]*[.])?[0-9]+\/[+-]?([0-9]*[.])?[0-9]+
def get_pattern(pat):
  try:
      return re.findall(regex, pat)[0][0]
  except Exception as e:
      return ''


df_archive_data['pattern'] = df_archive_data['text'].apply(get_pattern)
df_archive_data['fraction'] = df_archive_data['rating_numerator'].astype(str) + '/'
+ df_archive_data['rating_denominator'].astype(str)
df_archive_data[df_archive_data['pattern'] != df_archive_data['fraction']][['patter
n', 'fraction']]
```

Cleaning Data

The define, code, and test steps of the cleaning process are clearly documented.

Good work following define -> code -> test steps...

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.

You copied the original data before cleaning (important if at some point you need to trace back on your steps),

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.

Good work storing the results into a csv 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.

Good work sharing insights and corresponding visualisations

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.

Fine work... here you mainly discuss/explain your thought process in sufficient detail... This document is mostly for reflective learning purposes and to practice your communication skills

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.

Excellent work sharing insights and visualisations

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.

All files are shared

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START