Wrangle and Analyze Data

REVIEW

HISTORY

Meets Specifications

Dear Student:

You have done an excellent job with this project. it's clear that you have put in a great amount of thought and effort into its completion. Well done.

I thoroughly enjoyed reviewing your work and I wish you the best of luck with the remainder of the course, as well as upcoming projects :)

Code Functionality and Readability

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

All code cells run without execution errors.

Your coding practices and structure thereof is spot on. It's clear that you have a good understanding of code structure, syntax as well as how to correctly apply appropriate code for an intended purpose and to reproduce the results described.

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.

You have done an exceptional job ensuring that the notebook is very well structured, whereby you have included clear headings, descriptions, as well as embedded links which makes the notebook easy to navigate. This attention to details really goes a long way, very well done here.

Moreover, the notebook has an intuitive, easy-to-follow logical structure which clearly documents the gather, assess and clean steps. Great job.

You have ensured that your work is thoroughly commented, whereby you have incorporated many in-code comments, as well as markdown cells. Comments are very important as this allows the reader to follow along with the intentions of the author.

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.

Data has been successfully gathered from three different sources and you have also correctly stored the gathered data in a format according the project instructions. Well done.

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.**

It is clear that you have thoroughly and thoughtfully assessed the gathered data for data quality and tidiness issues. This takes a considerable amount of time and effort. Well done.

It's evident in your work that you have a good understanding of built-in functions and methods to assess data both visually and programmatically.

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.

You have done an excellent job identifying quality and tidiness issues. You have gone above and beyond in this area, where you have identified and document more quality and tidiness issues than what is required to meet specifications. Well done.

Furthermore, each issue is correctly classified as either a quality or tidiness issue.

Cleaning Data

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

The define, code, and test steps of the cleaning process have been followed throughout the cleaning phase. This ensures that the cleaning phase is structured well and your audience can easily follow along with each issue. Well done.

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.

Copies of the original datasets are made prior to cleaning.

At least eight data quality issues and two tidiness issue have been appropriately cleaned according to their definitions.

You've done an excellent job successfully combining all four dog stage values('doggo , floofer , pupper , and puppo') into a single column 'stage'. Additionally, very well done appropriately dealing with the duplicate records that resulted from the `pd.melt()`.

Thoroughly cleaning data can take an exceptional amount of time and effort and it's evident that you have put in a great deal of work and effort to thoroughly clean data for analysis and visualization. You have done an excellent job successfully identifying and cleaning important issues. Very well done.

Additional Comments and Suggestions:

Great job identifying that there are instances where dog names have been incorrectly extracted and stored in the name column. You had replaced names such as 'a', 'an', 'my', 'by' with np.nan, well done. However, there are other instances, in addition to 'a', 'an', 'my', 'by', where dog names have been incorrectly recorded. For example:

```
#assess naming issue
enhanced_archive_clean.name.unique()
```

```
'Atticus', 'Blu', 'Dietrich', 'Divine', 'Tripp', 'his', 'Cora',
'Huxley', 'Keurig', 'Bookstore', 'Linus', 'Abby', 'Shaggy',
'Shiloh', 'Gustav', 'Arlen', 'Percy', 'Lenox', 'Sugar', 'Harvey',
'Blanket', 'actually', 'Geno', 'Stark', 'Beya', 'Kilo', 'Kayla',
'Maxaroni', 'Doug', 'Edmund', 'Aqua', 'Theodore', 'Chase',
'getting', 'Rorie', 'Simba', 'Charles', 'Bayley', 'Axel',
'Storkson', 'Remy', 'Chadrick', 'Kellogg', 'Buckley', 'Livvie',
'Terry', 'Hermione', 'Ralpher', 'Aldrick', 'this', 'unacceptable',
'Rooney', 'Crystal', 'Ziva', 'Stefan', 'Pupcasso', 'Puff',
'Flurpson', 'Coleman', 'Enchilada', 'Raymond', 'all', 'Rueben',
'Cilantro', 'Karll', 'Sprout', 'Blitz', 'Bloop', 'Lillie',
'Ashleigh', 'Kreggory', 'Sarge', 'Luther', 'Ivar', 'Jangle',
'Schnitzel', 'Panda', 'Berkeley', 'Ralphé', 'Charleson', 'Clyde',
'Harnold', 'Sid', 'Pippa', 'Otis', 'Carper', 'Bowie',
'Alexanderson', 'Suki', 'Barclay', 'Skittle', 'Ebby', 'Flávio',
'Smokey', 'Link', 'Jennifur', 'Ozzy', 'Bluebert', 'Stephanus',
'Bubbles', 'old', 'Zeus', 'Bertson', 'Nico', 'Michelangelope',
'Siba', 'Calbert', 'Curtis', 'Travis', 'Thumas', 'Kanu', 'Lance',
'Opie', 'Kane', 'Olive', 'Chuckles', 'Staniel', 'Sora', 'Beemo',
'Gunner', 'infuriating', 'Lacy', 'Tater', 'Olaf', 'Cecil', 'Vince',
```

- Most incorrectly recorded names are lower case rather than correctly recorded names which are capitalized, therefore, we can use `str.islower()` to find these incorrect dog names and then use replace method to clean this issue. For example:

```
#all lowercase names
wrong_names = enhanced_archive_clean[enhanced_archive_clean.name.str.islower()]
wrong_names = wrong_names['name'].unique()
wrong_names
```

```
array(['such', 'a', 'quite', 'not', 'one', 'incredibly', 'mad', 'an',
       'very', 'just', 'my', 'his', 'actually', 'getting', 'this',
       'unacceptable', 'all', 'old', 'infuriating', 'the', 'by',
       'officially', 'life', 'light', 'space'], dtype=object)
```

```
# replace values equals to invalid names with none or nan
enhanced_archive_clean['name'].replace(wrong_names, np.nan, inplace = True)
```

```
#test

enhanced_archive_clean.loc[enhanced_archive_clean.name.str.islower()==True,'name']
```

```
Series([], Name: name, dtype: object)
```

Moreover, you had noted that:

> "the tweet has no dog name to fetch"

I'm unsure whether or not you had been referring to a specific tweet, but there are some instances where the dog name is in the text but has been incorrectly recorded. For example:

```
pd.set_option('display.max_colwidth', None) # run to see full width of columns in df
enhanced_archive_clean[(enhanced_archive_clean.name == 'a') &
                       (enhanced_archive_clean['text'].str.contains('named'))][['text', 'name']]
#there are other variations of phrases that preceed the dog name in the text column, other than 'named',
#such as 'T(h)is is', 'H(h)ere is', 'name is', ect
```

	text	name
1853	This is a Sizzlin Menorah spaniel from Brooklyn named Wylie. Lovable eyes. Chiller as hell. 10/10 and I'm out.. poof https://t.co/7E0AiJXPml	a
1955	This is a Lofted Aphrodisiac Terrier named Kip. Big fan of bed n breakfasts. Fits perfectly. 10/10 would pet firmly https://t.co/gKlLpNzII3	a
2034	This is a Tuscaloosa Alcatraz named Jacob (Yacōb). Loves to sit in swing. Stellar tongue. 11/10 look at his feet https://t.co/2IsIQ8ZSc7	a
2066	This is a Helvetica Listerine named Rufus. This time Rufus will be ready for the UPS guy. He'll never expect it 9/10 https://t.co/34OhVhMkVr	a
2116	This is a Deciduous Trimester mix named Spork. Only 1 ear works. No seat belt. Incredibly reckless. 9/10 still cute https://t.co/CtuJoLHiDo	a
2125	This is a Rich Mahogany Seltzer named Cherokee. Just got destroyed by a snowball. Isn't very happy about it. 9/10 https://t.co/98ZBi6o4dj	a
2128	This is a Speckled Cauliflower Yosemite named Hemry. He's terrified of intruder dog. Not one bit comfortable. 9/10 https://t.co/yV3Qgjh8iN	a
2146	This is a spotted Lipitor Rumpelstiltskin named Alphred. He can't wait for the Turkey. 10/10 would pet really well https://t.co/6GUGO7azNX	a
2161	This is a Coriander Baton Rouge named Alfredo. Loves to cuddle with smaller well-dressed dog. 10/10 would hug lots https://t.co/eCRdwouKCl	a
2191	This is a Slovakian Helter Skelter Feta named Leroi. Likes to skip on roofs. Good traction. Much balance. 10/10 wow! https://t.co/Dmy2mY2Qj5	a
2218	This is a Birmingham Quagmire named Chuk. Loves to relax and watch the game while sippin on that iced mocha. 10/10 https://t.co/HvNg9JWxFt	a
2235	This is a Trans Siberian Kellogg named Alfonso. Huge ass eyeballs. Actually Dobby from Harry Potter. 7/10 https://t.co/XpseHBlAAb	a
2249	This is a Shotokon Macadamia mix named Cheryl. Sophisticated af. Looks like a disappointed librarian. Shh (lol) 9/10 https://t.co/J4GnJ5Swba	a
2255	This is a rare Hungarian Pinot named Jessiga. She is either mid-stroke or got stuck in the washing machine. 8/10 https://t.co/ZU0i0KJyqD	a
2264	This is a southwest Coriander named Klint. Hat looks expensive. Still on house arrest :(\n9/10 https://t.co/lQTOMqDUle	a
2273	This is a northern Wahoo named Kohl. He runs this town. Chases tumbleweeds. Draws gun wicked fast. 11/10 legendary https://t.co/J4vn2rOYFk	a
2304	This is a curly Ticonderoga named Pepe. No feet. Loves to jet ski. 11/10 would hug until forever https://t.co/cyDfaK8NBc	a
2311	This is a purebred Bacardi named Octaviath. Can shoot spaghetti out of mouth. 10/10 https://t.co/uEvsGLOFHa	a
2314	This is a golden Buckminsterfullerene named Johm. Drives trucks. Lumberjack (?). Enjoys wall. 8/10 would hug softly https://t.co/uQbZJM2DQB	a

- There aren't too many records, and therefore, shouldn't affect the results if analyzed. This is to show how extensive data wrangling and cleaning can be, and also, it's an interesting quality issue to explore.

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 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.

You have done a fantastic job analyzing and visualizing the data. You have gone above and beyond here whereby you have produced more insights, as well as visualizations than what is required meet project specifications. Well done.

Moreover, your plots and insights clearly and appropriately represent the gathered and clean datasets. Additionally your visualizations are structured well and easy to read and interpret.

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.

This report has been compiled well.

Wrangling efforts are documented in a clear and concise manner. This report is structured well and efforts described are consistent with work done in the notebook.

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.

This is a great external report.

This report is compiled as such that makes it easy for the audience to know what was investigated, as well as the purpose of the report. Furthermore, findings are communicated in a concise, easy to understand manner. Great job.

Additionally, you have ensured that the plots presented can be readily interpreted by your audience with clearly represented titles, good work.

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

/