Wrangle and Analyze Data

REVIEW | CODE REVIEW | HISTORY

Meets Specifications

Congratulations

You have made it! 🎉
Wrangle and Analyze Data is a project that requires a lot of time and effort but you have been up to the task and you did everything very well! I can see that you have worked hard and that you have talent. I'm sure you'll do great things! Never stop believing in yourself and in your skills!

Student Notes:
I am very sorry that you've experienced notification problems, thank you for reporting, I will report the issue to the support team.

I wish you all the best with your Nanodegree and career!

Code Functionality and Readability

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

All the code is present in the **wrangle_act.ipynb** notebook and run without errors. Well done!

Tips

- How to use shortcuts with Jupyter Notebook
- Cheat sheet for data wrangling using python
- Use functions to avoid code repetition

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.

Every step of the wrangling process (gather, assess, and clean) is well documented with markdown text and the code is often commented. This helps anyone to follow and understand easily the wrangling workflow, that's a very good practice useful when you work in a team with other developers and also for ourselves. Especially when the code becomes very long, a clear structure helps you save a lot of time. Well done!

Tips

- 10 Tips for Writing Cleaner & Better Code

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 is successfully gathered from three different sources and in three different formats. Good job hiding your APIs credentials for your online security, to know more about it take a look at this interesting article Best practices for securely storing API keys

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.**

A very nice work here! Both visual and programmatic assessment are done properly in the notebook! Well done using info(), describe(), value_counts(), sum() and other useful functions to explore the data.

Suggestion

Please, remember that once displayed, data can additionally assess data using external applications like Microsoft Excel, Google Sheet or text editor.

Tips

- Pandas tips and tricks

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.

At least 8 data quality issues and 2 tidiness issues are correctly detected, well done!

Suggestion

Please, consider enriching the issue's descriptions in order to make them more clear and easy to understand.

Tips

A useful resource to know how to classify correctly each type of issue: Data Quality & Tidiness

Cleaning Data

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

Excellent job adding a markdown text cell in order to document clearly each **Define**, **Code** and **Test** step of the cleaning process.

Define:

Remove all the Tweets from the Archive, which Id is not in the Tweepy Dataframe.

Code:

```
In [281]: print(df_archive_clean.shape)
          print(df_tweepy_clean.shape)

          (2356, 13)
          (2339, 32)

In [282]: df_archive_clean = df_archive_clean[df_archive_clean.tweet_id.isin(df_tweepy_clean.id)]
```

Test:

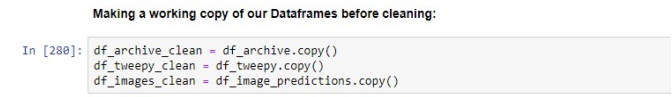
```
In [283]: df_archive_clean.shape

Out[283]: (2339, 13)
```

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.

Well done using the copy() function, making a copy of the datasets, before cleaning the data! :+1:



Making a working copy of our Dataframes before cleaning:

```
In [280]: df_archive_clean = df_archive.copy()
          df_tweepy_clean = df_tweepy.copy()
          df_images_clean = df_image_predictions.copy()
```

This is a very good practice to follow before starting the cleaning process, to know more about it you can visit this link : Why should I make a copy of a dataframe in pandas?

Cleaning issues

You did an excellent job cleaning correctly almost all the issues detected, especially those relative to the dog stages, on which you have handled properly the tweets with multiple dog stages. Many students struggle with this issue! :+1:

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.

A master dataset is correctly saved to a CSV file. Great job setting the Index Parameter to False in order to avoid an "unnamed" index column to the dataset!

```
In [356]: df_archive_clean.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.

You did an amazing job on your analyses, producing several interesting insights and visualizations. Plus, you have structured this section in a very clear manner, using question/answer pattern and describing properly each insight.

Resources

...some useful resources to improve your fantastic skills and produce always better visualizations 🚀

- How to make beautiful data visualizations in Python with Matplot
- 5 quick and easy data visualizations in Python
- The Best Python Data Visualization Libraries
- 10 Useful Python Data Visualization Libraries for Any Discipline

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.

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.