UDACITY

‹ Return to Classroom

Wrangle and Analyze Data

REVIEW

HISTORY

Requires Changes

1 specification requires changes

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

However, you need to work a little more on this project to meet all the specifications. Since you have already addressed most of the requirements, it is just a matter of paying attention to some finer details (please see below). I am sure you will be able to quickly get this project to meet all specifications as you have a very good python coding skills and understanding of **data wrangling process.**

To pass this project, you need to address only the comments marked as **Requires Changes**. The comments marked as **Suggestions** are optional and you do not need to address them to pass this project. But if you address these suggestions, it will improve your project.

Mainly, you need to **make the following changes** before resubmission, details of which can be found in the corresponding rubric item below.

- You need to **remove retweets** (text starting with RT @) as per project instruction. Removing retweets change your analysis results. So re-run the analysis and update your act_report.
- You need to use correct **data types** for all columns. All kinds of **IDs** should be of object type.

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 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.

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. You also did a great job **querying Twitter's API** to gather data for all the available twitter ID and importing it into a pandas dataframe, where many students struggle.

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.

Suggestions

Assessment is the most important part for data cleaning, which has huge impact on your downstream data analysis or predictive modelling. The following are some of the pandas functions you can use in the data assessment. **You have already used some of these functions.** You can try others.

- testing.assert_series_equal
- Various methods of indexing and selecting data - .loc(), .iloc()
- .duplicated()
- .isnull()
- .nunique()
- .info()
- .describe()
- .value_counts()
- .head()
- .tail()
- .sample()

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.

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** the data 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.

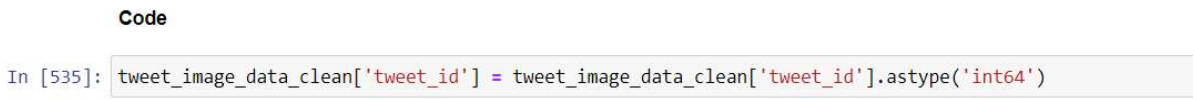
Requires Changes

As per project specification, we only want original dog ratings. So you need to **remove retweets** (text column starts with RT @) as a user can retweet their on tweet. **But you did not remove retweets. Please note that just removing the retweet-related columns like retweeted_status_id is not going to remove retweets.** You need to remove all rows that have values (not blank or non-null) in `retweeted_status_id`, `retweeted_status_user_id`, and `retweeted_status_timestamp` columns. As mentioned in the **Key Points** section of **Project. Wrangle and Analyze Data - Sublesson 2. Project Motivation**, this is an important quality

issue to be addressed, as it has a direct bearing on your analysis. You can remove retweet-related columns after removing rows with retweets as these columns become empty.
One of the ways to remove retweets is to select only rows that have null values in retweet related columns, using pandas isnull() function. You can do something like the following (the following code is just an example; you should appropriately change the name of the dataframe in the code).

```
df_1_clean = df_1_clean[df_1_clean.retweeted_status_id.isnull()]
df_1_clean = df_1_clean[df_1_clean.retweeted_status_user_id.isnull()]
df_1_clean = df_1_clean[df_1_clean.retweeted_status_timestamp.isnull()]
```

ID columns like `tweet_id` should be of `object` type (i.e. string), NOT `integers`. It is not intended to perform calculations with ID numbers. Please note that when you change the datatype of a column you should choose most suitable data type so that all the methods and functions associated with that datatype will be available for you if you want to manipulate the data. String is suitable for IDs, becauase some time these IDs may exceed limit of int. Please see this stackoverflow thread, which discusses a related question. To facilitate merging, you can convert tweet_id in all the dataframes to string.



Code

```
In [535]: tweet_image_data_clean['tweet_id'] = tweet_image_data_clean['tweet_id'].astype('int64')
```

Suggestions

You have used the string `None` to represent missing dog stage data. Instead, you should use `np.nap` to represent missing dog stage data, which is a standard way to represent Null values in pandas. This makes it easy to handle null values programmatically as there are many function available in pandas (e.g., isnull()).

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 bar and 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. For instance, for time series data (month, year etc.) **line chart** works best (e.g., favorite count over time). To depict relationship between two quantitative variables **scatterplot** works best (e.g., relationship between retweet count and favorite count). To depict distribution of variables, **histogram** is suitable (distribution of rating). Here are some resources that help you choose 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

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

Project Files

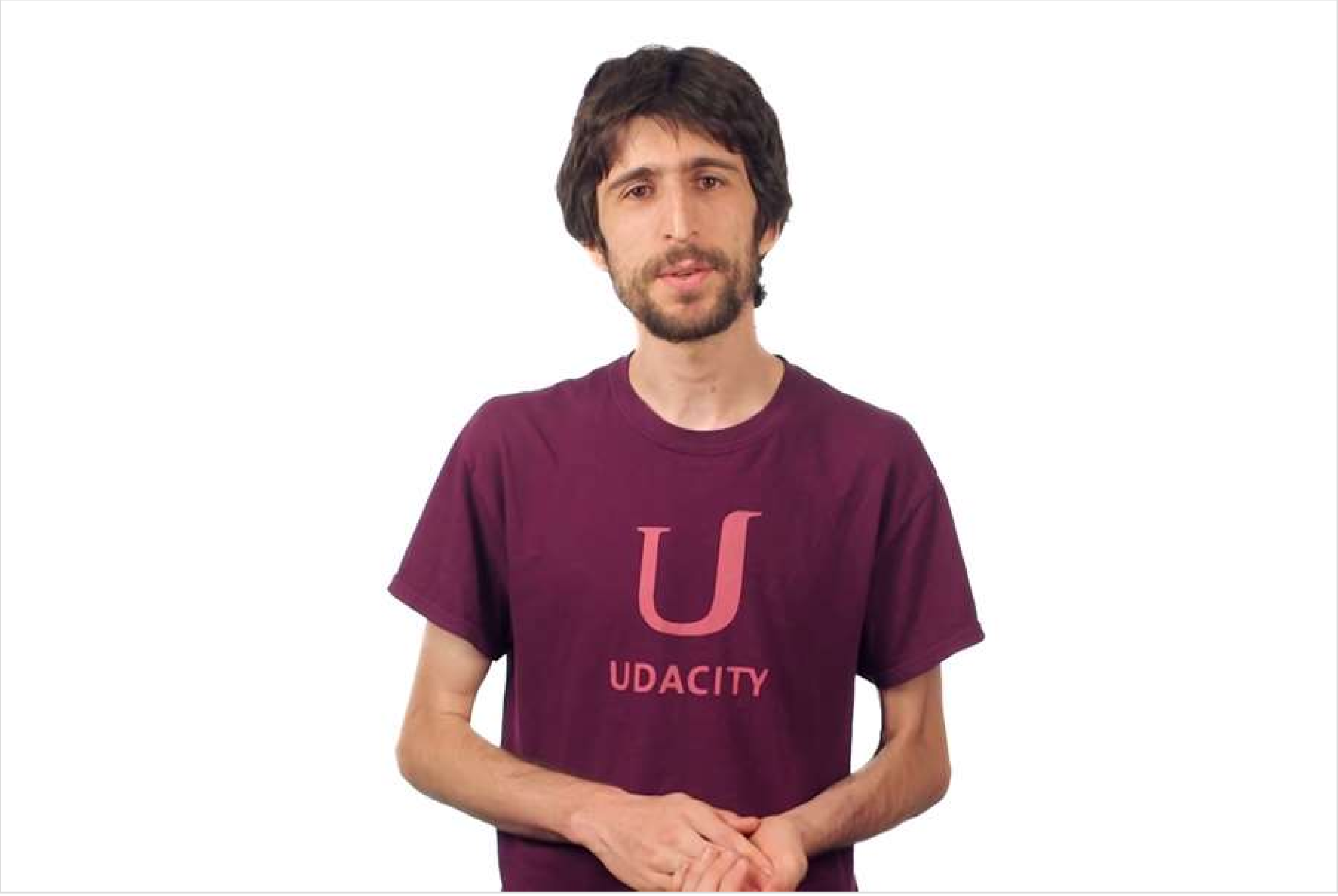
The following files (with identical filenames) are included:

- 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.

☑ RESUBMIT

DOWNLOAD PROJECT



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

RETURN TO PATH

Rate this review

START