Shahad Nasser Asseri

Project 5:
Wrangle and Analyze Data

Data Wrangling process :

1-Gathering Data: It's the first step in data wrangling process, Before gathering,we have no data, and after it, we do.in this project,I gathered three datasets from different sources:

a) The WeRateDogs: twitter archive data: "Existing file" the WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets.

b) The tweet image predictions data: "Downable file" i.e., what breed of dog (or another object,
animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded.

c) The twitter API data: Unfortunately, twitter didn't give me the access to get the data so, instead I used the tweet-json file that Udacity provided it. "JSON API file".

2-Assessing Data: it's the second step in data wrangling. When assessing, you're like a detective at work,inspecting your dataset for two things: data quality issues (*content issues*) and lack of tidiness (*structural issues*),I did both visual assessment and programmatic assessment and I found problems related to quality and tidiness.

a) Quality: issues with content, Low quality data is also known as dirty data.

Data Quality Dimensions:

- Completeness.
- Validity.
- Accuracy.
- Consistency.

Quality issues in twitter_archive:

- Timestamp column is of type object ,I will change it to **datetime** type to be more appropriate.
- tweet_id column is of type int it should be **String (object)** because I will not need to do any calculation or manipulation
- Replace None value to **nan** to be more clear.
- name column has wrong names that start with lower case and none values replace them with **nan** value.
- Unnecessary html tags in source column,strip all html anchor tags (ex: <a..>) in source column,Then Convert the datatype from string to categorical.

Quality issues in image_predictions:

- The "p1" and "p1_conf" columns will be renamed with more meaningful titles.
- Remove unnecessary columns:

  -The column "jpg_url" will be removed since url data is already contained in the twitter archive data.

  -The "p2" and "p3" related columns will be removed as I am only using the most confident prediction ("p1").

  -After removal of "False" entries, the "p1_dog" column will be removed becouse it already did its purpose.
- tweet_id column is of type int it should be string **(object)** type because I will not do any calculation or manipulation.

Quality cleaning in Tweet_json:

- First change the id name to "tweet_id" befor marge,and convert type to **String(object).**

b) Tidiness: it's has specific structural issues, also known as **untidy** data. Untidy data has **structural issu.**

Tidy data requirements:

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

Tidiness issues in twitter_archive:

- In twitter_archive doggo, floofer, pupper, puppo columns merge them in one column called dog_stage.

Tidiness issues :

- Merge all the dataset in one table.

Assessment data steps:

1. Detect
2. Document

3-Cleaning Data: is the third step in data wrangling. It is where you fix the quality and tidiness issues that you identified in the assess step.I copyed the data in another DataFrame so I can do manipulation on it at the same time it doesn't affect the real data.and I flowed the Cleaning process :

1. Define.
2. Code.
3. Test.

I cleaned all of the problems either related to quality or tidiness programmatically. As much as I could with my previous knowledge, I tried to do it with simple code to make it more understandable and not hard coded. When I finish everything, I merged the datasets together after checking that everything is good and worked as expected, for me it was much easier doing it at end in order to easily recognize any problem.