Data wrangling

@WeRateDogs

1)Gathering data:

Data were collected from three different sources and in different ways:

1- The WeRateDogs Twitter archive: Manually downloaded from Udacity.

2-The tweet image predictions: Uploaded programmatically using the Requests library.

3- Each tweet's retweet count and favorite: Uploaded by Twitter API using Python's Tweepy library.

2)Assessing data:

•Quality: issues with content. Low quality data is also known as dirty data.
•Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data.
Tidy data requirements:

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

Using two types of assessment:

- Visual assessment: scrolling through the data software application Excel.
-Programmatic assessment: using code to view specific portions and summaries of the data (pandas' head, tail, and info methods)

Quality:

Twitter archive table

- Contain some text on the replay or retweet Must be removed.
- The text id (666287406224695296, 778027034220126208, 786709082849828864) it has an error rating.
- Column rating_numerator contains a rating greater than 14 Must be removed.
- Column rating_denominator contains a rating greater than 10 and Smaller then 10 Must be removed.
- Convert the timestamp variable from object to dataframe.
- There are missing values in column expanded_urls but most of them contain retweet or replies and we will fix the missing values after deleting the retweet and replies.
- Change the name of the rating_numerator column to dog_ratin_of_10 and delete the rating_denominator column.
- Delete the columns (in_reply_to_status_id,in_reply_to_user_id,source,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,rating_denominator,,)

image predictions table

- Delete the rows that contain the false column p1_dog.

- Change the name p1 column to Dog_Type.
- Change the name of the jpg_url column to Image_Link.
Tidiness
- In the image classification table, we are interested in the highest predictive ratio and we will only take the p1 column

Tidiness
- In the image classification table, we are interested in the highest predictive ratio and we will only take the p1 column
- Four columns constitute one variable in `Twitter archive` (doggo,floofer,pupper,puppo)
-All three tables must be merged.

3)Cleaning

Before the cleanup operation, a copy of the data was copied

The data cleaning process is divided into three stages

1. Define: convert our judgments into defined cleaning tasks.
2. Code: convert those definitions to code and run that code.
3. Test: test dataset, visually or with code, to make sure your cleaning operations worked.