Wrangle and Analyze Data- WeRateDogs Twitter Archive

Introduction:

The dataset will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

The Data :

1. twitter-archive-enhanced.csv : Basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets.
2. Tweet image predictions (image_predictions.tsv) : what breed of dog (or other objects, animal, etc.) is present in each tweet according to a neural network.
3. json_tweet : Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting.

Data Wrangling Process :

1. Gathering data
2. Assessing data
3. Cleaning data

Step 1 : Gather Data

The first steps of wrangling data is Gathering data because simply we cant work without data on hands. In this part , we will read three files:

1. CSV File "twitter-archive-enhanced.csv"
2. TSV File "image-predictions.tsv"
3. TXT File "json-tweet.txt"

Finally we will have data the goes to the next steps "Assessing data".

Step 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:

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

There two types of assessment:

Visual assessment: scrolling through the data in your preferred software application (Google Sheets, Excel, a text editor, etc.).

Programmatic assessment: using code to view specific portions and summaries of the data (pandas' head, tail, and info methods, for example).

Step: 3 : Cleaning Data

After assessing the data and we made a list of things needs to be fixed and adjusted ,

This where the quality and tidness issues are remedied. We make sure that the data is accurate and ready for analysis. The Define, Code, and Test are the process of cleaning, were used in this sequence, with multiple definitions, cleaning operations, and tests under each header, respectively.

The issues was cleaned in the datasets:

Quality Issues

Twitter_archive Table:

1. timestamp- object
2. text are not complete.
3. There is rating denominator > 10 which is not acceptable
4. tweet id - int
5. Some dogs names has only one letter such as 'a' 'an' and 'none'is not a name. and names has different format
6. NAN value in many column, should be replaced with none.
7. Some of Expand urls has more than one link.
8. Some of the data are retweet and replies - not tweet.
9. Source written in HTML format.
10. Rating numerator > 20
11. doggo', 'floofer', 'pupper', 'puppo' - one column
12. duplicate values
13. create rating column

Image prediction Table :

1. p1,p2,p3 has both small and capital letter for the prediction.
2. The columns p1 ,p2,p3 is confusing and difficult for other people to understand.
3. The data set should includes 2356 not only 2073.
4. There is 66 jpg-url is duplicated.
5. Data format of predictions some separated with _ .

json_tweet_api Table:

1. tweet_id - int

Tidness Issues:

1. Types of dogs should be belong to one columns instead of one for each!
2. The three tables should be merge to one table since they represent the same data.