Data Wrangling Report

Introduction

The dataset that I will be wrangling ,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.

The takes of this project are as follow:

- Data Gathering
- Data Assessing
- Data Cleaning

Data Gathering

Gather each of the three pieces of data as described below in a Jupyter Notebook titled Wrangle_act.ipynb.

1. The WeRateDogs Twitter archive:
   This file called twitter_archive_enhanced.csv.

   The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file called image_predictions.tsv.

2. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Data Assessing

In Data Assessing step, we will analyze

Quality of the data:

Quality of the data include the issues such as if there any missing , duplicated , incorrect entries in the data set.

Tidiness of the data:

Checking the structure of the data set .

In the dataset that I use , I found these issues

Quality Issues:

1-Tweet_id is an integer.It has to be a string.

2-Erroneous dog names starting with lowercase characters (e.g. a, an, actually, by).

3-Data type should be datetime for retweeted_status_timestamp,timestamp insted of string.

4-Several columns have empty values, like in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

5- Incorrect values in rating numerators .

6-Found an instance of a name being "O" instead of "O'Malley".

7-Name has values that are the string "None" instead of NaN.

8-343rd entry is not a dog rating.

9-A lot of columns that won't be used for analysis.

Tidiness Issues:

1- Doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into one column named "stage".

2- Three data frames twitter_archive, image_predictions, and df_tweet_info should be one (combined table)

Data Cleaning

Quality Issues:

1-Change 'tweet_id' to a string.

2- Replace all names that start with a lowercase letter with a NaN.

3- Convert timestamp to datetime object.

4- Remove 'retweeted_status_id', 'retweeted_status_user_id' and

'retweeted_status_timestamp' columns

5- Calulate the value of 'rating'.

6- Replace the name 'O' with "O'Malley".

7- Replace all 'None's with a NaN.

8- Drop 343rd entry that is not a dog rating.

9- Delete columns that won't be used for analysis.

Tidiness Issues:

1- Merger doggo, floofer, pupper and puppo columns in twitter_archive data set into one column named "stage"

2- Three data frames twitter_archive, image_predictions, and df_tweet_info should be one (combined table)