Wrangle Report

Introduction

Tweet archive is the dataset that I worked on it. It’s the data 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.

Data wrangling:

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

1. Gathering data:

I gathered the data from three different resources and obtained as following:

a) **Twitter archive file:** the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

b) **The tweet image predictions:** This file is hosted on Udacity servers and was downloaded programmatically using the Requests library and URL information. URL ='https://video.udacity-data.com/topher/2018/November/5bf60c69_image-predictions-3/image-predictions-3.tsv'

c) **Tweet API:** by using the tweet IDs in the WeRateDogs Twitter archive, The Twitter API was queried for each tweet using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. (I used the json file that provided by udacity).

2.Assessing data

After the data was gathered. I started to assessing the data as following:

a) Print a sub part of The data frames and get info about columns by using (df.info, value_counts, sample, duplicated)

b) There's 4 columns (doggo, floofer, pupper, puppo) that dog stage separated into them.
c) The format of the source is bad
d) (timestamp, retweeted_status_timestamp) data type should be datetime not object.
e) The numerator and denominator have invalid values
f) There are invalid (a, an, the) and missing names
g) Missing values from images predictions dataset because there's 2075 rows instead of 2356.
h) some p1, p2 and p3 names start with uppercase characters and other with lowercase and have (underscore) '_' instead of spaces.
i) Missing data in tweet_df because there's 2354 rows instead of 2356.

3.Cleaning data

This part of the data wrangling was divided in three parts:

1. Define
2. Code
3. Test

These three steps were on each of the issues described in the assess section.

I cleaned the data as the following:

1. Rename the id in tweets_df_clean to (tweet_id) to be the same as other files.

2. Drop 181 rows that (retweeted_status_user_id) is not null.

3. drop useless columns in analysis from twitter archive ('in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp','expanded_urls', 'source') columns.

4. Melt the doggo, floofer, pupper and puppo columns to dogs and dogs stage column then I dropped The dogs' column.

5. Sort dogs stage and delete the duplicated based on tweet_id.

6. correct the denominator to be = 10

7. correct the numerator ratings and get it from the text

8. convert numerator ratings data type to float

9. replace all dogs stage that takes None to be null values.

10. get rid of invalid dog names (a, an, None).
11. convert timestamp data type to date time.
12. Separate the timestamp of the tweet archive to, day, month and year columns.
13. Delete the rows with no images (URL).
14. capitalize the first character in p1 , p2 and p3 and replace _ with space.
15. Merge the three data frames on one based on the tweet_id data frame.

4.Storing the data

Store the cleaned data into a Csv file.