wrangle_report

September 24, 2021

1 Wrangle Report

The dataset wrangled (and analyzed and visualized) 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.

2 Project Steps

The tasks in this project are as follows:
Step 1: Gathering data
Step 2: Assessing data
Step 3: Cleaning data
Step 4: Storing data
Step 5: Analyzing, and visualizing data
Step 6: Reporting

2.0.1 Gathering Data:

In this step, I gathered all three pieces of data in the wrangle_act.ipynb notebook.

2.0.2 Assessing Data:

After gathering all three pieces of data, assessed them visually and programmatically for quality and tidiness issues. ###### Pandas functions used for programmatically assessing: .head() .info() .describe() .sample() .isnull() .value_counts() .columns()

Quality issues 1.tweet_id in twitter_archive & image_predictions, and id column in tweets_json, are an intgers.
twitter_archive
2.timestamp is an object.
3.name mistake (such, an, a, the) which is not a name.
4.the dataset has retweets.
5.alot of missing data, in 6 columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls).
image_predictions
6.there are 66 duplicated urls.

tweets_json

7.These columns (in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str) are floats. And these two columns (retweet_count and favorite_count) are integers.

8.there are null values in some columns, and other columns with NaN in all rows. 16 columns.

Tidiness issues 1.the dog "stage" in twitter_archive (doggo, floofer, pupper, and puppo) are in 4 columns.

2.rename id column to tweet_id, like the other two DataFrames, so we can merge later.

3.merge all 3 DataFrames into 1 DataFrame.

2.0.3 Cleaning Data:

After data assessing I cleaned each issue in three steps (Define, Code, Test).

1. Converted tweet_id in twitter_archive & image_predictions, and id column in tweets_json, to a string.
2. Converted timestamp to datetime in twitter_archive_clean.
3. Changed all name mistakes in twitter_archive_clean "such, an, a, the" (which are lowercase) to None.
4. This dataset has retweets so I had to remove all rows with retweets and keep all rows with NaN, in retweeted_status_id.
5. Drop these columns in twitter_archive_clean (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id, expanded_urls) as it has alot of missing data.
6. Drop the 66 duplicated urls in image_predictions_clean.
7. Changed these columns (in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str) to objects. Also these two columns (retweet_count and favorite_count) changed to floats.
8. Drop null values in tweets_json_clean.
9. Created the dog_stage in twitter_archive_clean, and then put (doggo, floofer, pupper, and puppo) as values in it.
10. Renamed id in 'tweets_json_clean' column to tweet_id, like the other two DataFrames, so we can merge later.
11. Finally, merged all 3 DataFrames into 1 DataFrame.

2.0.4 Storing Data:

And after gathering, assessing & cleaning the data I stored it to a CSV file to do the last step which is Analyzing, and visualizing data.