Wrangle Report

Introduction

For this Data Wrangling Project, I will be performing data wrangling process by gathering, assessing and testing. Specifically, I will wangle "WeRateDogs", which is from Twitter data. After the gathering process, when accessing gathered data, I will be focused on its quality and tidiness. After data is gathered and accessed, I will then move forward to cleaning process. This process includes define, code and test. At the end of this project I will have cleaned dataset to visualize my analysis.

Step 1. Gathering

Gathering is most important part of data wrangling because if you gathered wrong data, you can mess up with your whole data analysis. For this purpose of this project, I gathered 3 files from Udacity and Tweeter.

- Twitter archived file (archived)
- Image prediction (image_prediction)
- Twitter API (df)

Step 2. Assessing

After successfully gathered all necessary files, we now need to assess the problems. We need to focus on what makes my data messy and not cleaned? There is two ways to assess your data. There are visual assessment and programmatic assessment. When assessing these issues with data, we need to categorize the issues into quality and tidiness issues. For quality issue Completeness, Validity, Accuracy and Consistency are main four consideration. For tidy issues, we need to consider if each variable forms a column, each observation forms row, each observational unit forms a table.

There are issues I detected from data:

Archived

- Quality
    - Completeness
        - Couple columns are missing values in great differences. Those columns are in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
        - Many retweeted columns have missing values.Plus retweeted needs to be removed because we need only original data.
    - Validity
        - In Name column, I found invalid names for dogs such as such, a, quite, not etc.
    - Accuracy

        - Timestamp dtype is object instead of datetime.(retweeted_status_timestamp also has wrong dtype but since we don't need retweeted we don't have to worry about this yet)
        - In_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id column, the values should be in either string or integer instead of float because float is decimals.
    - Consistency
        - In rating_denominator column, I the denominator isn't consistency. Plus the demoninator should be 10
        - Sources is not easy to recognized
    - Tidiness
        - In doggo, floofer, pupper, puppo columns we need to put all those as values in one single column.

Image_prediction

    - Validity
        - There are 66 duplicated jpg_url which is the image of a dog. Duplicated pictures is not necessary. (Programmatically detected)
    - Consistency
        - Under P1, P2, and P3 columns, dog's names aren't consistently written in all lowercase, typically dog's breed name shouldn't be capitalized.

Json(df)

    - Tidiness
        - In df table, I haven't detected any issues with Quality. However, so far, I have seent tweet_id columns in three all tables. We need to join all three tables together and drop the duplicated columns if necessary. This process will minimize the tables and we can have one bigger dataset to evaluate.

Step 3. Cleaning

This process comes last step in data wrangling. However, you can always go back if you detected another issues with data. This process contains define issues, use code to fix the issues then test your code.

These are issues that I defined:

1. Drop the columns that aren't necessary to this analysis. Those columns are in_reply_to_status_id, in_reply_to_user_id.

2. Remove Retweeted columns.

3. Name column values needs to be more appropriate with dog's name.

4. In rating_denominator column, the denominator isn't consistency. Denominator should be 10.

5. Drop 66 duplicated jpg_url in image_predictions.

6. Combine doggo,floofer,pupper,puppo columns into ranking columns to minimize columns

7. Timestamp data type is object instead of datetime.

8. Under P1, P2, and P3 columns, dog's names aren't consistently written in all lowercase, typically dog's breed name shouldn't be capitalized.

9. Clean Messy HTML tags.

10. Merge all 3 tables (archived, image prediction, df) because it is better to view.

Step 4. Storing Data, analysis and Visualization.

At this point, we should have clean data to view. When this part is ready we are need to move on to analysis and visualization. When it comes to visualization we make visual effect of our data to show the readers.