Wrangle Report

WRANGLING AND ANALYZE DATA

Submitted by

Afyaa Alkhamisi

i. Introduction

The goal of this project is wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. Therefore, to achieve the main purpose we followed these steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Storing Data
5. Analyzing and Visualizing Data

Now, we will introduce all the steps and explain them with the most important details while wrangling and analyzing the data.

ii. Gathering Data

The Data: In this project, we worked on the following three datasets:

- Enhanced Twitter Archive: This is on-hand data we gathered by manually loading from twitter_archive_enhanced.csv file.
- Additional Data via the Twitter API: This is web scraping data we gathered by programmatically loading using Requests library (url = https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv).
- Image Predictions File: we gathered data via the Twitter API using Tweepy library and parsed from JSON to txt file.

iii. Assessing Data

We assessed data by using visual assessment and programmatic assessment. We discovered and documented several quality and tidiness issues in this dataset, which are:

Quality issues:

Here is a list of issues we found in the three datasets:

A. df_twitter_archive DataFrame:

1. source column has an unreadable format written in an HTML structure.
2. expanded_urls column has missing URLs.
3. Erroneous datatype

- tweet_id column should be a string, not an integer.
- in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id, and retweeted_status_user_id columns should be an object (string) not a float.
- retweeted_status_timestamp, and timestamp columns should be a datetime not an object (string).

4. The retweeted tweets are duplicates of original tweets.
5. rating_numerator and rating_denominator columns has invalid values some text columns written in number/number format, some ratings include decimal numbers, and some tweets include ratings for two dogs.
6. name column has missing dog names, and invalid names, like a, an or less than 3 characters.
7. Duplicate and unused data in rows and columns should be deleted from the dataset.

B. df_image_predictions DataFrame:
   1. Uppercase and lowercase letters for dog breed names given in columns p1, p2 and p3.
   2. Missing records (2075 rows instead of 2356).
   3. Erroneous datatype (tweet_id column should be a string, not an integer).

C. df_tweet_data DataFrame:
   1. Erroneous datatype
      - tweet_id column should be a string, not an integer
      - timestamp column should be a datetime, not object (string).

Tidiness issues:

Here is a list of issues we found in the three datasets:

1. Four columns in the df_twitter_archive dataset should be combined into dog stage column (doggo, floofer, pupper, and puppo columns).
2. Predictions and confidence for a dog breed in df_image_predictions dataset could be packed into breed_prediction and prediction_confidence columns (p1, p1_conf, and p1_dog, etc...).

3. The three datasets should be merged into a new dataset based on identifier, the tweet_id (df_twitter_archive, df_image_predictions, and df_tweet_data).

iv. Cleaning Data

In this step, we cleaned all of the issues documented while assessing in previous step. Furthermore, we used programmatic for handled all the issues, by using several functions. Firstly, we made copies of original pieces of datasets. Then, we cleaned the tidiness issues by using multiple methods such as: merge, append, and groupby. Likewise, we fixed the quality issues regarding to missing data and wrong values. Also, we fixed extraneous rows and columns, and we fixed erroneous data types. As well, we used multiple methods like drop, astype, rename, replace, lower, and regular expressions. We tested changes in the dataset by using info, head, sample, and list methods.

v. Storing Data

After we cleaned the data in above, we stored the final cleaned dataset in a CSV file with the main one named twitter_archive_master.csv by using to_csv method.

vi. Analyzing and Visualizing Data

The analyzing and visualizing data part will be in another file called act_report.pdf.

vii. Conclusion

In summary, we wrangled and analyzed the WeRateDogs Twitter data, by following several steps including gathering, assessing, and cleaning data for worthy analyses and visualizations. In general, this project is a good practice of what we learned in the course, I am able to build a well-organized data wrangling notebook that contains details of each part with an explanation and documentation of each problem and the process of solving it.