Data Wrangling Project

Udacity Data Analysis NanoDegree

WeRateDogs Twitter Archive

By Lovelina Richter

This report will describe the steps taken to cleanup and analyze the WeRateDogs Twitter Archive Data. Important points to take note on this project:

1. The data WeRateDogs contains basic tweet data for all 5K plus.
2. The project only requires original ratings with images and not include retweets
3. The project requires at least 8 data issues and 2 tidiness issues.
4. Data quality involves issues related to quality of actual data values, for example null or unexpected values.
5. Tidiness involves issues related to structure of data
6. Although, data wrangling consists of three steps, Gathering of Data, Assessing the Data, and Cleaning the Data, this project requires analysis and at least 1 visualization produced in Jupyter Notebook.

I. Data Gathering

In this step, I gathered data from three (3) different sources, as required.

1. WeRateDogs Twitter archive data, the file was provided and downloaded from Udacity.
2. The tweet image predictions data is hosted in Udacity server, was downloaded through Python’s request package.
3. The third data source is WeRateDogs tweet, a JSON data gathered through Twitter API using Python's Tweepy library. The authentication for Twitter API is not included in the submitted code and tagged as “Hidden”

Each set of data was saved into CSV format and copied in another file for prior to actual data cleanup. Each CSV file is saved in UTF-8 encoding, so that the file can be opened in MS Excel. Each data set was imported into separate dataframe.

II. Data Assessment

After gathering all three required data sets, each dataframe was scanned visually and programmatically. The data was scanned visually by using panda’s .sample() function. The data was assessed programmatically by using value_counts() functions to show how many records have data quality issues. Data quality issue or dirty data refer to the actual content or value. The tidiness of data was assessed using and .info() function which displays information about a DataFrame including the index dtype and columns, non-null values.

III. Data Cleaning

The next step in this data wrangling project is the data cleanup process. The first stage of the cleaning process was to merge all three dataframe for easy data manipulation, to ensure no duplication of records, and no duplication of actual data cleanup process. Throughout the data cleanup process, the dataframe was constantly assessed visually and programmatically, using the function .sample(), .info(), .value_counts(). The following are the cleanup process completed:

Data Quality Issues:

1. Remove retweet, as the project only require original tweet.
    a. Columns related to retweet were deleted to remove unnecessary columns when displaying data
2. Remove replies, as the project doesn’t require analysis of replies
    a. Columns related to replies were deleted to remove unnecessary columns when displaying data
3. Tweets with no images were removed

4. All records with no dog_type were removed from data set.
    a. The data set was sorted by dog_type and duplicate tweet_ids were deleted
5. Name of dogs were cleaned by checking if the first character is lowercase
6. All records with rating_numerator greater than or equal to 15 to remove outliers.
7. Shorten data in source column by using regex library
8. Remove unnecessary numbers from timestamp column and change data type to datetime
9. Convert each column to its most appropriate data type
10. Finally, the data set was copied into final clean dataframe.

Tidiness Issues:

1. There were 4 columns for dog types, such as 'doggo', 'floofer', 'pupper', 'puppo'. These four types were combined into 1 column using .melt function.
2. The column "Names" was renamed to "Name"
3. Breed and predictions were combined into column Breed and Confidence, where were originally saved into three different fields respectively, 'p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog'.