Wrangle and Data Project

WRANGLE REPORT

DATA WRANGLING STEPS: GATHER, ASSESS, AND CLEAN

Oseremen Sandra Osara | Udacity’s Junior Data Analysis Nanodegree | May 20, 2022

WRANGLING MY DATA

The dataset wrangle in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a twitter account with comments about dogs.

In this project, my goals were:

Wrangling the twitter data through the following processes:

- ▪ Gathering Data
- ▪ Assessing Data
- ▪ Cleaning Data

- Storing, analyzing and visualizing your wrangled data
- Reporting on the data wrangling efforts, data analysis and visualization

GATHER:

This project gathered data from the following sources:

- The WeRateDogs Twitter archive. The 'twitter-archive-enhanced-2.csv' file was provided to Udacity Students (Like me). This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

- The Tweet image prediction, i.e., what breed dogs (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students (Like me).

- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favourite ("like") count at minimum and any additional data I find interesting.

ASSESSING DATA:

After gathering all three pieces of data, assess them visually and programmatically for quality and tidiness issues.

I used two types of assessment:

Visual assessment: Each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes. Once displayed, data can additionally be assessed in an external application (e.g., Excel, text editor).

Programmatic assessment: Here, I used different pandas' functions and/or methods to assess the data.

At the end of these, I came out with some quality and Tidiness issues which were taken care of in the cleaning stage.

Tidiness

1. Dog stage data is separated into 4 columns
2. All data are related but are in 3 deferent data frames

Quality

Enhanced Twitter Archive

1. There are 181 retweets as under the retweeted_status_id
2. There are invalid dog names under the name column
3. The timestamp has strings has its data types instead of datetime.
4. Invalid tweet_id datatypes i.e having integers instead of strings
5. Row 313 has a 0 as its denominator rating
6. 23 rating denominators are not equal to 10
7. 440 rating numerators are less tha 10

Image Predictions

1. Some P names begin with uppercase while others begin with lowercase

CLEAN

Cleaning data is tedious and often iterative. Just when data analyst believe they found all quality and tidiness issue, they often found additional issue arises. The cleaning process involves three steps:

In this stage, All the issues that were identified when accessing were cleaned.

Before cleaning, copy of the original data was created. The Define-Code-Test Framework was used for the session:

Define: Determine exactly what needs to be clean and how and below are the definitions I worked on in this project.

- Merge the clean versions of df_twitter_archive, df_image_predictions, and tweet_json dataframes Correct the dog types
- Create one column for the various dog types: doggo, floofer, pupper, puppo
- Delete retweets
- Remove columns relating to retweets: retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
- Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id
- Change tweet_id from an integer to a string
- Change the timestamp to correct datetime format
- Convert invalid dog names:
- Convert lowercase Letters to Uppercase
- Creating a new dog breed column using the image prediction data

Code: Programmatically clean the code and this is contained in my wrangle_act.ipnb file

Test: Evaluate the code to ensure the data set was cleaned properly and this is contained in my wrangle_act.ipnb file

–