Data Wrangle_Act Report

by Salomey Bezua

Introduction:

The project's primary data source is the WeRateDogs Twitter user @dog rates' tweet archive. A Twitter account called WeRateDogs awards stars to users' dogs along with amusing comments about the dogs.

The WeRateDogs Twitter project had the following objectives:

- Gathering Data
- Assessing Data
- Cleaning and storing Data
- Analysis and Visualization

Gathering Data:

Data was collected from the following sources:

- The Twitter archive CSV file was provided by Udacity
- The tweet image predictions file was programmatically downloaded from Udacity, which resulted from a machine learning algorithm performed on the images from the WeRateDogs Twitter account. This was facilitated using the requests library in Python.
- Twitter API was gathered by web scraping from Twitter using Python's Tweepy library.

Assessing Data:

After gathering the data, the tables were assessed visually and programmatically. The aim was to look at the data to solve tidiness and quality issues. Ten issues were addressed. These issues were found by looking at the structure of each data. This was done using info(), head() and describe. Below is a description of some of the issues addressed in each dataset

Twitter Archive data –

- The tweet id and timestamp data type were erroneous and needed to be changed from int and object to object and DateTime, respectively.
- Some columns had missing data and were also not too relevant to the analysis. i.e in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp, expanded_urls.
- The dog names were corrected as ***'An'***, ***'a'*** and ***'the'*** were all in the list of names.

- The source column had HTML tags which needed to be removed.
- Some of the rating numerator values were inaccurate compared to the rating given.

Image Prediction

- The tweet id was an integer
- There were no duplicates in the data frame
- The types of dogs in columns p1,p2, and p3 are inconsistent. There's a mix of uppercase and lowercase letters.
-

Twitter API

- There were missing tweets.

Cleaning Data:

After the assessment, the define-code-test framework cleaned the three datasets.
Before cleaning, copies of the original datasets were made to preserve them. The main activities were to change the data types of some variables and create new columns from merged columns. These changes are well documented in the notebook.

ACT

Analysis and Visualization:

I started the analysis by looking at the relationships between the variables. Some questions were asked in this analysis. They are stated below with the findings.

1. Correlation – The retweet counts over time positively correlated with the favourite tweets.
2. Popular Dog Name – People tend to name their dogs Charlie, Cooper, Tucker, Penny and Oliver.
3. Popular Dog Breed – The most preferred breed was the Golden Retriever.