Wrangle Report
We Rate Dogs Project

Sarah El-Khouly

December 2020

1. Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.
This report briefly describes my wrangling efforts.

Project details:

- 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 and data analyse and visualization.

2. Gathering Data

My wrangling efforts for the WeRateDogs Twitter project included gathering data from the following sources:

- **Twitter archive file**: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
- **The tweet image predictions:** i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded manually.

- **Twitter API JSON**: by used the file provided by Udacity (tweet -json.txt)

  .

3. Assessing Data

Once the three tables were obtained I assessed the data as following:

- Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.

- Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).

Then I separated the issues encountered in quality issues and tidiness issues. Key points to keep in mind for this process was that original ratings with images were wanted.

4. Cleaning Data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described.
First and very helpful step was to create a copy of the three original dataframes.
I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original.
Whenever I made a mistake, I could create another copy of the dataframes and continue working on the cleaning part.