Udacity's Data Wrangling Project Report

Prepared by: Suhaila Ehab Alkholy

Project Overview

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. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage. In this report I report on my efforts to gather, assess, and clean the twitter data to gain some useful insights. Which do you think is the most loveable dog?

Methodology

Data Gathering

This is the part where I start to gather the data.

- Firstly, I import the Twitter archive as a DataFrame into my Jupyter Notebook.
- Secondly, I decided to download the image prediction file programmatically and then create a DataFrame with its data using Python's Pandas library.
- Thirdly, I used my Twitter's Developer account and Python's Tweepy to create an API object. Each Tweet is a JSON file and I wanted to import all of its data into a text file where each new tweet forms a new line. I did this by looping over each file and dumping its contents into the tweet_json.txt. I then created a Panda's DataFrame with this data.

Data Assessment

This is the part where I assess the gathered data.

- I assess the data visually and programmatically.
- I find 8 quality issues and 2 Tidiness issues.

Data Cleaning

This is the part where I clean the data, and store the data to CSV files.

- I use the method of **define**, **code** and **test**.
- For each issue I define how I would tackle it, then I write the code and then test if the quality has been resolved.
- I then store the data to CSV files to start my analysis process.