|  |
| --- |
| **Wrangle Report** |
| Yousef Majeed |
| This report is for the wrangling data project (fourth project) in the Data Analyst Nanodegree Program that running by Udacity corporate. |
|  |

Data Analyst Nanodegree

May-13, 2019

## Overview

In this report I will describe the data wrangling process performed in the project named #Data Wrangling - Enhanced Twitter Archive.

Data wrangling process consists of 3 steps:

* Gathering data
* Assessing data
* Cleaning data

## Gathering Data

Gathering data is the first step of data wrangling. For this project the data needed to be gathered was the following as described below in a Jupyter Notebook titled wrangle\_act.ipynb:

1. The WeRateDogs Twitter archive stored in a csv file: twitter\_archive\_enhanced.csv. The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I was used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced."
2. The tweet image predictions file, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (image\_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the followingURL:<https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv>
3. Additional Data via the Twitter API: each tweet's retweet count and favorite ("like") count. Using the tweet IDs in the WeRateDogs Twitter archive, Twitter API was queried for each tweet's JSON data using Python's Tweepy library and stored in a file called tweet\_json.txt file.

Gathering data for this project was done in Python Jupyter Notebook using pandas, requests and tweepy libraries. What I found challenging is using tweepy library to query Tweepy API, especially because it’s the first time using such library, it’s the first-time accessing Tweeter via secret keys and access. It took a while to download the data and many attempts and many failures, but in the end, I managed to extract it successfully.

## Assessing Data

* For extraction of data from the database **I used SQL queries.**
* For making a line chart visualization and analysis **I used Microsoft Excel.**

## Cleaning Data

* To observe the trends in temperature I calculated moving average (MA).
* I used built-in tool pack that called ‘Data Analysis’ to calculate the moving average