By Nyabenge Sylvester.

Introduction:

Wrangle and Analyze Data

Data wrangling is the process of gathering your data, assessing its quality and structure, and cleaning it before you do things like analysis, visualization, or build predictive models using machine learning. The project was part of the requirement of the data wrangling section of the Udacity Data Analyst Nanodegree program.

Project Summary

Using Python and its libraries, I gathered data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This has been documented in a Jupyter Notebook, plus showcase them through analyses and visualizations.

The dataset used for wrangling, analyzing and visualizing is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

Data Gathering

Data was gathered from three different sources and all these were done programmatically to allow for reproducibility of the process:

File on hand

This file was downloaded manually read into a pandas dataframe called twitter_archive.

Web scraping

This file, image_prediction.tsv is hosted on Udacity's servers and was downloaded programmatically using the Requests library. The data was saved in a pandas dataframe called image_prediction

Additional data from the Twitter API

First you setup twitter developer account from one's twitter account, proceed to the developer's portal to acquire API keys to extract data. However, I was unfortunate in extracting the data for reasons only known by twitter thus proceeded to download the emergency data for those unable or unwilling to create a twitter account.

With the JSON text file, I read it line by line into a pandas dataframe called df_tweets.

Assessing the data

The dataframes were subjected to check for quality and tidiness issues. The issues discovered are well documented in my Jupyter notebook file.

Cleaning the data

As required and advised, copies of the original data was created to clean the data based on the issues discovered in assessment. The Define, Code and Testing format was used to document this process. The requirement for cleaning at least 7 Quality and 2 tidiness issues were met.

Analysis and Visualization

The datasets were cleaned and necessary fields were merged into a master file ready for analysis.

After analysis I was able to find the answers to the following:

1. I was able to identify the most prominent dog breeds
- The Golden Retriever is the most prominent dog breed.
- This dog breed had the most likes.
- This dog breed had the most retweets
- This dog breed was the most tweeted.

2. I was able to identify the most prominent dog slang name.
- The 'Puppo' was the most prominent dog slang name.
- A tweet with 'puppo' generated a lot of tweet interactions.

3. I was able to identify the most prominent source (app).
- The prominent source was IPhone. Many people interacting with the weratedogs used the IPhone platform to tweet.

4. Months with the highest tweet interactions
- The Month of June had the highest number of tweet interactions.

5. The record with the maximum and minimum number of retweets.
- This is clearly highlighted in the Jupyter notebooks.
- The Record with the lowest had 16 retweets.
- The Record with the Highest had 79515 retweets.

For every insight, a visualization was provided to show the insight discovered. The visualization are available in the Jupyter notebooks.