WeRateDog Wrangle Report

Introduction

The goal of this project is to wrangle WeRateDogs Twitter data to to create interesting and trustworthy analyses and visualizations.WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.This project is part of Udacity Data Analyst Nanodegree programme and the main purpose of this project is focussed on data wrangling from the Twitter account of WeRateDogs using python and its libraries.

Project Details

Real-world data rarely comes clean. Using Python and its libraries, I gathered data from a variety of sources and in a variety of formats, assessed its quality and tidiness, then cleaned it. Assessing and cleaning the entire dataset completely would require a lot of time and effort so only 8 quality issues and 2 tidiness issues(minimum) needed to be cleaned in this dataset.

The wrangling process of the project involved these steps-

1.Dtaa Wrangling which consisted of-

- Gathering Data
- Assessing Data
- Cleaning Data

2.Storing,Analyzing and visualizing the wrangled data

3.Reporting the analysis and visualization

Gathering

The data for this project was stored in three different format and obtained as mentione below:-

1. twitter-archive-enhanced.csv file-This was extrcted programmatically by udacity and provided us to use.

2.Image_prediction file-The tweet image predictions, i.e., what breed of dog (or other 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 following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

3.Twitter API and Tweet_json file-By using the tweet IDs in the WeRateDogs Twitter archive,I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assessing

After gathering each of the above pieces of data,I assessed them visually and programmatically for quality and tidiness issues. Then Detected and documented (8) quality issues and two (2) tidiness issues in wrangle_act.ipynb Jupyter Notebook.

By Visual assessement I got acquinted with data and found some tidiness issues such as there should be only one column 'stage' instead of for columns 'doggo','floofer','pupper','puppo'in twitter_archive dataset and retweet_count_and_favorite_count should be part of twitter_archive dataset.

Programmatically Assessement gave me the most of the quality issues that are present in the three datasets.I seperated the issue in two group quality and tidiness and after that i divided the quality issues according to datasets and checked for completeness,validity,accuracy and consistancy.

Cleaning

This part of the data wrangling is divided into three steps-Define,Code and Test.

First i have created the copies of the original datasets so that i can do trial and error with copy dataset rather than the original. Further i have divided it in three steps.In step 1, I addressed completeness issues and in step 2, i tackled the tidiness issues and in step 3 ,i fixed the quality issues.

After fixing the tidiness issues and quality issues i joined all 3 tables in one.

Storing the data

After cleaning the data,i stored the wrangled data into twitter_archive_master.csv

In [ ]: