Wrangle Report

Introduction

This is a report fle for the wrangle process of WeRateDogs.The dataset that is used in this project is that of twitter users at dog rates. The purpose of the project was to practise data wrangling skills which is part of te udacity data analysis program. This report works through the data wrangling process, focusing on the gathering, assessing and cleaning of data

The WeRateDogs Twitter project goals included:

1. Wrangling the twitter data through the following processes:
   - Gathering Data
   - Assessing Data
   - Cleaning Data
2. Storing, analyzing and visualizing your wrangled data
3. Reporting on the data wrangling efforts and data analyse and visualization

Process

Gathering data

Data was gathered from three source

1. Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually
2. 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 programmatically using the Requests library and URL information
3. Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

Assessing Data

The data was assessed visualy and programitial for both quality issues an tidines some of the tidiness and quality issues found are

1. missing records
2. timestamp should be in datetime format and not object
3. source columns contain html link tags
4. the denominator has a max value of 170
5. p1, p2, and p3 ave lowercase given names
6. img_num are not use full

7. datetime should be of datatime datatype and not object
8. tweet_id should be string instead of int
9. Combine the 4 dog stage columns into a single column in the archive table
10. merge WeRateDogs_Achives, imagePredictions and apiTweets tables

Cleaning Data

Issues found wile assessing data were cleaned following the define code test method

1. Define: defines the problem
2. code : show the working code to solving the define problem
3. Test : show the cleaned output

Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with. I have used Python programming language and some of its packages. There are several advantages of this tool (as compared to e.g. Excel) that is used by many data scientists (including the guys at Facebook). Overall, this project was completed successfully and I'm extremely pleased with the new skills I acquired