Wrangle Report

February 12, 2020

The purpose of this report is to introduce this project and walkthrough the efforts invested to gather, clean, assess and analyze data.

0.1 Introduction

The purpose of this project is to practice learnings gathered in data wrangling section of Udacity's Data Analytics Nanodegree program. The dataset of interest pertains to Twitter archives for @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

0.2 Gather

The dataset that was gathered from twitter archives was provided as a downloadable csv file. (twitter_archive_enhanced.csv) The were two other datasets that were collected and merged with the twitter archives to enhance the analysis. One of those datasets was image predictions file, which was processed by Udacity using neural networks to predict dog breed based on image URLs provided in archives. This dataset was gathered by using the requests library in python. (image_predictions.tsv) The other data set was collected using the tweepy api to gather favorite and retweet count for each twitter post in the archives. (tweet_fav_retweet.txt)

0.3 Assess

To assess the quality, these datasets were observed manually (by extracting data on local machine in csv/tsv formats), and programatically (using basic python pandas functions like - head(), duplicated(), info(), sample()).

0.4 Clean

This was the section which tested technical and analytical skills on data wrangling efforts. While messy data was easier to clean, the main challenges that were faced are - 1. Constructing regex to extract the correct ratings from plain text in twitter archives and further clean the dataset by working around the extracted ratings. 2. Getting all image predictions merged into one column by using multiple if-else statements.

0.5 Analyze

After data was assessed to be clean in all form, analysis was done around getting insights and building visuals on dog breed, ratings, retweets and favorite count. Please see 'act_report.pdf' file for detailed analysis.