Wrangle Report of Twitter WeRateDogs Data

Harry Xiong

Introduction

This data wrangling project utilizes data from twitter account WeRataDogs. Through gathering, assessing and analyzing, I tried to wrangle with the data and provide some meaningful insights.

Gather

First, three data frames were collected and generated from three distinct sources. The archive was given by Udacity. The enhanced archive was extracted from the tweet content and needed to be assessed and cleaned. The image prediction was given as a url and needed to be programmatically gathered. The tweet-json.txt was given as a json text, so I first read the file line by line into a list and then converted the list to a data frame.

Assess & Clean

For the assess and clean part, I observed and acted on 8 quality issues and 2 tidiness issues. Most of the quality issues have been assessed and cleaned in the enhanced archive dataset. Notably, as instructed by the project details, I remained only the rows without retweet or reply status ids. Keeping only the original tweets would ensure that the analysis could treat all the tweets with same weight. Also, a tidiness issue was pointed out in the project details as well. Since the data was extracted programmatically, the dog stages were split into 4 columns instead of 1. I have cleaned this issue by aggregating them into one column. Other assessment process including inspecting the column names and making sure they were readable, observing if the columns were in the correct data types and formats, and maintaining a level of consistency across data frames for merging purpose. Finally, three data frames were merged into one with tweet ids being the index.

Summary

The wrangling project focused on a multi-source dataset in the context of a twitter account named WeRateDogs. Data comprised of tweet content, neural network image prediction and additional tweet information has been gathered, assessed and cleaned for further analysis.