We Rate Dogs
Wrangling Data Udacity Project

By Mohamed Hassona

Introduction

This is a Udacity project “Wrangle and Analyze Data”, the goal of the project was to get data from multiple sources from the twitter group WeRateDogs then assess and clean the data and make a analysis of the data.

Gathering

To start with there there was 3 sources of data a csv file, a link to a csv file and the data on Twitter’s api using tweepy. Getting the data from the csv files was easy enough as the data was there and I just needed to read the csv files, for the Twitter data I had to gather the JSON data save it as a txt file. After I had finished I had 3 pandas dataframes.

Assessment

Once I had the data I was able to look for ways to make my data cleaner and tidier I did this visually and programmatically. As I found things that needed changing I would document the changes that need to be made.

Cleaning

I had to change some of the data types delete some of the columns that had a lack of data and delete retweets as they were not needed and would have made results of my findings skewed, I found the rating_denominator values to be all over the place but decided to leave the ratings as I think it’s intentional to give dogs over the top ratings for humour. Once I was finished cleaning the data I combined the 3 dataframes and saved it as a CSV file.

Presenting the data

I decided to focus of the dog breeds as I thought it was the most interesting way to present the data and gives me an excuse to show the cute dogs. I used horizontal bar graphs to make it easier to read the dog names