Ebtisam Alshehri

Data Wrangle Act Report

Udacity Data Wrangle and Analyze Project

Introduction

The data wrangle and analyze project focuses on applying the 4-step data wrangling process, which is composed of data gathering, data assessing, data cleaning, and finally data visualization and analysis. The data set provided for this project is based on a Twitter account called @dog_rates, aka WeRateDogs, which rates dogs provided by their audience.

Data gathering

- I gathered data from three different files. the Twitter-archive-enhanced.csv file
- We gathered the image prediction tsv file using the request library.
- Using the Twitter API provided by Udacity and converting the json file into a data frame, we additionally copied the most important columns into a new data frame.

Data assessing

When viewing the data frames, it had a lot of issues, but the quality and tidiness issues I spotted while assessing the data frames were:

Quality issues

I. The rating numerator should be greater than or equal to 10.
II. The denominator of the rating should be ten.
III. In df1 we delete (tweets) that have images.
IV. wrong data type for timestamp (it should be a datatime64 instead of an object)
V. +0000 appears at the end of the timestamp content.
VI. Missing values in df1
VII. Drop unnecessary columns (that we will not use in our analysis)
VIII. Invalid dog names(none,a)
IX. In the invalid prediction in df2, it is predicting dogs as (orange and paper towel).

Tidiness issues

I. Columns (doggo, floofer, pupper, puppo) must be in one column, not divided into 4 columns.
II. 2.columns (p1, p2, p3) should be in one column and have a clear column name.

Data cleaning

This is the longest process in our project, and since there are many issues with the data frames, we used different methods in Pandas such as drop, mask, replace, to_datatime for conversion, value_counts, and head for testing, and the string function rstrip to delete unwanted string values in the timestamp column.

Data storing

We merged the data frames by using join and the intersection of them, and then we stored it into a new csv file.

Data visualization and analysis

During the exploration of our wrangled dataset, we have discovered the dog with the lowest rate, the most common dog names and breeds, and the most favorited dog.