Data Wrangling Report

Introduction

The target of this project is to put in practice what I learning during the section Data Wrangling. This Wrangle and Analyze Project is part of Udacity's Data Analyst Program. The dataset that is wrangles (analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. In this project I was working with the data wrangling, which consists of:

- Gathering data, download resourses from diferent references
- Assessing data, for quality and tidyness issues
- Cleaning data, that identified in previous step
- Storing, analyzing, and visualizig wrangled data
- Reporting on data analyses and visualizations

The most tools, libraries and programming language I used in this project are:

- Python
- Numpy Library
- Pandas Library
- Requests Library
- Tweepy Library
- Json Library
- Matplotlib Notebook
- Twitter's APi

Gathering

Data was gathered from 3 different sources:

1. The 'twitter-archive-enhanced' was provided by Udacity. This file includes a huge variety variables foe each tweet as tweet_id, timestamp, text, name, rating, ect.
2. The 'image_prediction file was downloaded using the Requests library.
3. Additional data, including favorite count and retweet count were gathered using the Twitter API.

Assessing Data

After gathering data were used the following methods:

- .head()
- .sample()
- .info()
- .value_counts()

Quality issue:

- Tweet_id was the incirrect data type
- The columns conteined the 'None' insted of 'NaN'
- Timestamp was the incorrect datatype

- Missing values from images datased(2075 rows insted of 2356)
- Sources are not readable

Tidiness issues:

- Dogstage was in 4 columns(doggo, floofer, pupper, puppo), there is no reason to keep them
- Merge all dataframes togethr as they all contained the proper information

Cleaning Data

This part of the data wrangling was performed in three stages: Define, Code and Test. First, copies of DataFrames were created before cleaning. First and very helpful step for me was to create a copy of three original dataframes. Then used methods for investigation: merge(), reduce(), drop(), replace(), head(), etc

Storing

The final datset was stored as 'twitter_archive_final' in a csv file. At this step, the data was successfully wrangled and therefore ready for analysis and visualization

Analysis and Visualization

In preparation for the analysis part, these three tables(datasets) needed to be put in relationship with each other. In my case the key element in each dataset is the tweet ID. Since I decided to run the analyzis usung Python Pandas library, the easiest wasy to set up the relationship, was to combine all datasets to one major table.