Wrangle report

Introduction

Data wrangling is an important part of any big project these days. As fields like neural networks become more popular collecting datasets for them is becoming more popular as well. The process isn't that easy for anyone starting it for the first time, but I think you get the hang of it quickly.

WeRateDogs is a twitter account that rates dog photos & videos however, they seem to ignore the traditional rating systems and made up their own.

Data wrangling

This consisted of 3 parts:

- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data

There were 3 different datasets and each of them was gathered in a different way

- **twitter-archive-enhanced.csv:** This one was the easiest as it was available to download directly. It basically was an archive for old tweets from the user @dog_rates, but it was missing some key data like favorite and retweet count (I guess fetching that data later made more sense as it would be up to date).
- **image-predictions.tsv:** This file was provided as a URL to download the file from. There were two methods to do this either using the io library or using pd.read_csv()'s ability to load from URL. The first method made a problem later as I had to use skimage's io library which overwrote python's io library. This file contained predictions for a neural network on the images that @dog_rates tweeted to detect an object in the photo.
- **tweet-json:** This was the trickiest file to get but it was quite simple as well. I just collected all the tweet IDs from **twitter-archive-enhanced.csv** and used the twitter API to get a json file for each tweet, Stored that json file locally so that I won't have to waste 20-30 mins every time I needed to run the program.

Assessing Data

After collecting the three tables I started looking at them to find quality and tidiness issues this was done

1) Visually: By viewing the data using excel.

2) programmatically: By using python functions to find certain data easier.

Cleaning Data

In this step I started by making copies of the original datasets and then started fixing the problems observed in the previous step one by one. This was done mostly using Dataframe's functions. And finally, I merged all datasets into one.