TWITTER DATA WRANGLING

INTRODUCTION

In this project we used the weRateDogs data from twitter for our analysis.

Details:

First I extracted the data from twitter using api, I had to apply for the twitter api keys which I was given and I felt really nice. I extracted the data and saved it as tweet_json.txt which I later converted to a csv for easier analysis.

I downloaded the archive data and image data from udacity's server programmatically using python libraries.

ASSESSING.

I assessed the data both visually and programmatically and this are the issues that I found out:

Issues

archive data

Timestamp columns should be datetime
The ratings in the numerator should exceed 10 and also remove other extreme values
The denominator ratings should be 10
remove html tags from source.
Remove retweets since they are duplicates
removing a, an, the from names
drop: retweeted_status_user_id,retweeted_status_user_id,retweeted_status_timestamp columns

exctracted tweets

Remove columns we don't need.

tidiness
Dog names are in different columns

merge the archive dataset with the extracted one.

WRANGLING

I created a copy of the three datasets

archive_clean = archive_df.copy()

image_clean = image_df.copy()

twitter_clean = df_exctracted_tweets.copy()

and then I defined coded and tested the issues.

Saving the datasets.

I combined the datasets programmatically and saved them as csv files.

The first one was between the archive data that I had downloaded from udacity's server together with the one I exctracted from twitter using api.

twitter_archive_master.to_csv('twitter_archive_master.csv',index=False)

The second one which was the master dataset I combined the above the dataset with the image dataset to have a complete combined dataset.

twitter_archive_master_clean.to_csv('twitter_archive_master_clean.csv',index = False)

What followed was Analysis.