Wrangle Report

Introduction

The goal of this project is to put into reality what I learnt in the data wrangling data component of the Udacity Data Analysis Nanodegree program.The dataset in question is the tweet collection of Twitter account @dog ratings, commonly known as WeRateDogs

WeRateDogs is a Twitter account that assesses people's dogs with a funny remark about the dog. These ratings almost always have a denominator of ten.

Project details

The tasks of this project are as follows:

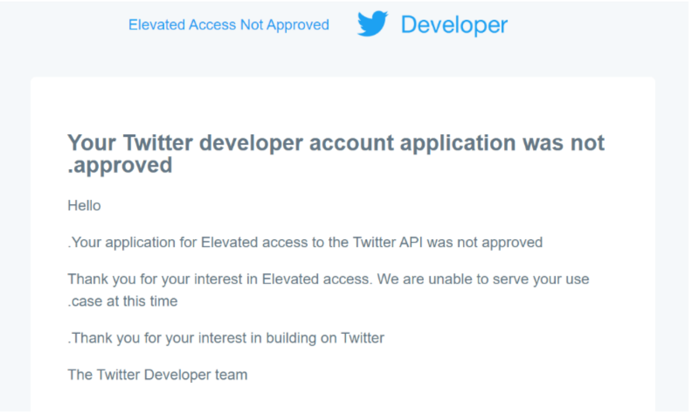
- Gathering data
- Assessing data
- Cleaning data

Gathering data

The data for this project consists of three different datasets but unfortunately, Twitter will not allow me to do so:

Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually

The tweet image predictions, i.e., what breed is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information.



Elevated Access Not Approved

Your Twitter developer account application was not .approved

Hello

.Your application for Elevated access to the Twitter API was not approved

Thank you for your interest in Elevated access. We are unable to serve your use .case at this time

.Thank you for your interest in building on Twitter

The Twitter Developer team

Assessing data

After obtaining the tables, I evaluated the data as follows

I utilized two visual tools. One method was to print the complete data frames separately in Jupyter .Notebook, and the other was to examine the CSV files in Excel

Programmatically, by employing various methods (for example, info, value counts, sample, duplicated, groupby, and so on)

Quality Issue:

1- missing data in the following columns

[,in_reply_to_status_id, in_reply_to_user_id

,retweeted_status_id, retweeted_status_user_id

retweeted_status_timestamp, expanded_urls]

2- rating_denominator should be a standard 10, but there are a multitude of other values

3- retweeted_status_timestamp is also an object (the other retweeted statuses are floats)

Tidiness Issue:

1- The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo)

2- This data set is part of the same observational unit as the data in the 'twitter-archive-enhanced-2.csv' - one table with all basic information about the dog ratings.

Cleaning Data

After the assessment, I cleaned the data through the following means:

Merge the clean versions of the data frames df Twitter archive, df image predictions. Correct the dog .breeds

.Retweets should be removed

Make one column for each type of dog: floofer, pupper, and puppo Remove any columns that are no longer required: in reply to status id, in reply to user id, retweeted status id, retweeted status user .id, and retweeted status timestamp are all possible values

Make tweet_id from an integer to a string

.Remove any columns that are no longer required

.Change the timestamp to the proper date/time format

Issues with proper naming

Standardize dog ratings