Udacity Data Analyst Nanodegree -Project "Data Wrangling and Analyzing

21-03-2021

Omar Ahmed

Overview

This project is about wrangling and analyzing data collected from WeRateDogs twitter account to create interesting and trustworthy analyses and visualizations.

Goals

1. Gather Data.
2. Assessing Data.
3. Cleaning Data.
4. Storing Data.

Gathering Data

The WeRateDogs Twitter archive. Is downloaded manually from udacity .

The tweet image predictions hosted on Udacity's servers are downloaded programmatically using the Requests library .

The JSON data is supposed to be downloaded by the twitter API but because i couldn't create a twitter developer account i downloaded `tweet_json.txt from Udacity` .

Assessing Data

Twitter archive.

Visual Assessment

By Assessing the Data Visually you get many Nan values in (in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id

,retweeted_status_user_id,retweeted_status_timestamp)

Programmatic Assessment

By programmatic assessment you can easily catch Quality Issues like.

- Tweet_id is int not string .
- Wrong parsed names like 745 None , 55 a .
- Time_stamp is object not datetime.
- Nan values at expanded_urls columns .
- Dog stage (doggo,floofer,pupper,puppo) each as a column having many non-values.
- Tweets with decimal numerator
- Outliers rating_denominator
- Zero as rating_denominator

Image Prediction

Programmatic Assessment

- P1,P2,P3 columns have invalid data like (orange,paper_towel,starfish, boathouse, mailbox.
- create a new dog_breed column using the image prediction data.

Tidiness Issues

- doggo,floofer,pupper,puppo as multipe columns
- merge the 3 dataframes in 1 dataframe
- drop unnecessary columns

Cleaning Data

After assessing the data ,its cleaning time .

Firstly make copies of the 3 dataframes .

Change the tweet_id from int to string.

Change timestamp to DateTime.

Delete the retweets.

Drop tweets with no image .

Extract the dog stage from the text column.

Delete tweets with decimal numerator in text.

Delete tweets with a denominator less than 10 or bigger than 10.

create a new dog_type column using the image prediction data.

Drop(p1,p1_conf,p1_dog,p2,p2_conf,p2_dog,p3,p3_conf,p3_dog,img_num)columns.

Drop unnecessary columns (in_reply_to_status_id,in_reply_to_user_id, doggo,floofer,pupper,puppo) in tweet_archive_clean.

merge the 3 dataframes in 1 dataframe.

Saving the DataFrame to a csv file.