Wrangle and Analyze Data Project

Wrangle Report

Introduction

In this project we will wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. We will gather the data from three places twitter archive, twitter API's, and Image Predictions File. This document will discuss the process of wrangling the data.

Gathering

We gather the data from three places each in different way.

Twitter archive

- We read the twitter archive data from a csv file that provided by Udacity. First I download the file and address it the project file. Then, I read it by using pandas library using read_csv method.

Twitter API's

- This data we read it by using Twitter API'S, I need first to have twitter developer account to use the API'S, I have an account before. I use my keys to fetch the tweet for WeRateDogs user. Then I save the tweet in .txt file as required. After that I wrote the same data in .json file to read it easier. I used json library to read from .json file and save it in a dataframe. Before save it in dataframe we need to normalize it, since there are nested object in json.

Image Predictions File

- This data we read it from url link using requests library then save it in .txt file and read it using pandas and tap as delimiter.

Assessing

For assessing data, I used different method to assess the data. I used some method in pandas like .info() , .shape , .duplicated. Also, some of the issues I found it by visualizing the dataframe and other by understanding the data and look how is the structure of it.

List of Quality Issues

1. expanded_urls column in df_twitter has missing value.
2. the link tag in df_twitter_archive column source extract the tag.
3. the link tag in df_twitter_json column source extract the tag.
4. the data type of timestamp column in df_twitter.
5. column names in df_image_predictions data frame p1,p2,p3.
6. column names in df_image_predictions data frame p1_conf,p2_conf,p3_conf.
7. column names in df_image_predictions data frame p1_dog, p2_dog, p3_dog.
8. duplicated expanded_urls in df_twiiter.
9. a lot of Nan value column in df_twitter_archive.
10. evaluate the tweet text.
11. extract the tweet hour in new column.

List of Tidiness Issues

12. dogs stages column in df_twitter.
13. linking the dataframes df_twitter_archive & df_twitter_json by tweets id.
14. linking the dataframes df_image_predictions & with the merged from 11 by tweets id.

Cleaning

In the phase I wrote a definition of what is the wrong with the data and how can I fix it. Then, I wrote the code of cleaning. A lot of the cleaning code I search it in the internet and look to python library documentation to get the idea of how we can fix the issue. After that I wrote I test code to see if we fix the issues correctly.