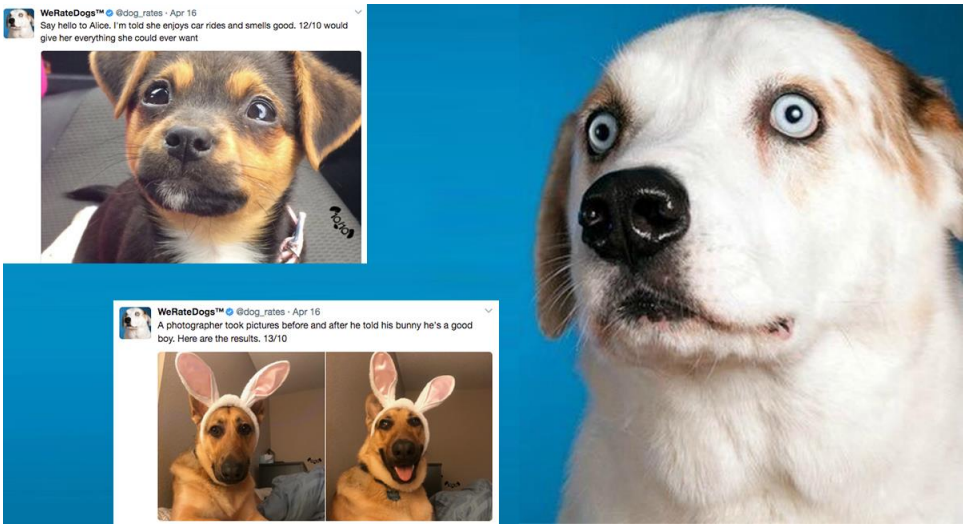
WRANGLE AND ANALYZE DATA PROJECT

Adedamola Babayato | Udacity Data Analysis Nanodegree | June 22, 2022

Table of Contents

Introduction

The dataset WeRateDogs, is the tweet archive of Twitter user @dog_rates. WeRateDogs is a Twitter account that rates dogs with humorous comment about the dog. This archive contains basic tweet data (tweet ID, timestamp, text, etc.).



What Software Is Used?

I worked outside the Udacity classroom and worked on my local machine using Jupyter notebook. I had pre-installed Pandas, NumPy, Request and Json libraries so I only installed the Tweepy library. I made use of text editor and visual studio code to access the twitter-json.txt and twitter_api.py files respectively. Also made use of excel to visually assess the three datasets.

Project Goal

The goal of this project is to wrangle the WeRateDogs Twitter data to gather data from different source, to be assessed and to perform cleaning techniques to raise the quality and tidiness of the data. Hence it can be used to perform exploratory data analysis.

Data Wrangling

Step 1: Gathering Data

In this project, I worked on three datasets which are:

- Enhanced twitter archive (twitter_archived_enhanced.csv)- I gathered this data by manually downloading it from the Udacity Project workspace then I uploaded it into the Jupyter notebook workspace on my local machine and read the data into a pandas Dataframe.
- Image Prediction File (image_prediction_tsv) – The file was hosted on Udacity's servers. I gathered this by programmatically downloading it the using the Request library following the provided URL.
- Additional data via Twitter API – Unable to gain elevated access from Twitter, I manually downloaded the twitter_api.py and tweet_json.txt file provided. I read the Json file with the (with open as file:) function to read through the line in file then I appended the line to a list of dictionaries then I converted it to a dataframe to begin assessment on the data.

Step 2: Assessing Data

After gathering the three datasets, I assessed them visually and programmatically for quality and tidiness issues.

Visual Assessment – I read the three datasets into jupyter notebook to visually assess and additional made use of excel to assess the image prediction file and enhanced twitter archive file while I made use of text editor and visual studio code to visually assess the json file and python file respectively.

Programmatical assessment – I used Pandas' info(), describe(),tail(), sample() , duplicated(), isnull() and value_counts() methods to assess the data.

After assessing the three datasets I was able to find 12 quality issues across the three datasets and 2 tidiness issues.

Step 3: Cleaning Data

Before I cleaned these issues, I made copies of each dataset using Pandas' copy() method. During the cleaning process, I used the Define-Code-Test framework and also documented my approach making use of comments. In my cleaning process I merged the three datasets as they all have a relationship. Finally, I successfully cleaned all the issues addressed in the Assessing phase and created a master dataset containing all gathered and cleaned data.

Step 4: Storing

I saved the gathered, assessed and cleaned master Dataframe to a CSV file name twitter_archive_master.csv

Step 5: Analyzing and Visualizing

I performed analysis on cleaned master dataset and was able to come up with three insights and visualizations about:

1. Popular Dog Stage by percentage – I used the different dog stages to carry out this analysis
2. Most popular Dog breeds – I extracted dog breeds from the prediction data then created a new column for Dog breeds.
3. Most dog tweet favorited and retweeted – I used the favorite count and retweet count to derive this insight.
4. Relationship between the favorite count and retweet count – I created a scatterplot to depict this relationship between the two variables.

Step 6: Reporting

I created two written reports namely wrangle report and act report respectively using Microsoft Word. This wrangle report document describes my wrangling efforts and my approach while the act report document communicates insights and displays visualizations made from my wrangled data.