Wrangle and Analyze data

Wrangle report

This project is about wrangling and analyzing the tweet archive of Twitter account “@dog_rates”, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." This was a quote from @dog_rates account during a celebrated exchange in which the account shut down a person taking issue with their rating system. WeRateDogs has over 4 million followers and has received international media coverage.

In this project, I was challenged to wrangle and analyze a very messy data frame. Starting from gathering data using three different ways, one where the file was a given and easily downloaded, second was downloading the file programmatically from Udacity servers using Requests library, and finally the third by using Twitter API by Tweepy library (Although in my project, I couldn’t get validation from Twitter to use the API, so I just downloaded the file from Udacity). After gathering the data, I started to assess it, where I should document any issues I can find visually and programmatically. After documenting the issues, I started the cleaning process, where I will be fixing all (or most) of the issues that I documented it in the assessing step.

Now I’ll briefly describe the process of wrangling:

1- Gathering Data

Depending on the source of your data, and what format it's in, the steps in gathering data vary.

The high-level gathering process:

• **Obtaining data** (downloading a file from the internet, scraping a web page, querying an API, etc.)

• **importing that data into your programming environment** (e.g. Jupyter Notebook)

2- Assessing Data

There are two types of issues you are looking for:

• **Quality**: Issues with content.

• **Tidiness**: Issues with structure that prevent easy analysis.

and you can assess by:

• **Visual assessment**: Scrolling through the data.

• **Programmatic assessment**: Using code to view specific portions and summaries of the data.

3- Cleaning Data

You can clean:

• **Manually**

• **Programmatically**

o **Define**: Convert our assessments into defined cleaning tasks.

o **Code**: Convert those definitions to code.

o **Test**: Test the dataset.

Wrangle Report

Introduction:

Gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The dataset I worked on is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

The wrangling Process go through some steps:

- Gathering data
- Assessing data
- Cleaning data

And ends with Storing, analyzing, and visualizing the wrangled data.

Step 1: Gathering.

There are 3 resource of data

• WeRateDogs Twitter archive (CSV file)

• tweet image predictions (Using the Requests library)

• tweets retweet count and favorite using twitter APIs (Text file)

In my case unfortunately twitter reject my request, so I used twitter Json file provided to me by Udacity.

Step 2: Assessing.

After gathering all necessary data, I start assessing data for Quality & Tidiness Issues.

Quality Issues:

• timestamp & retweeted_status_timestamp type in twitter_archive_enhanced should be datetype.

• tweet_id type in twitter_archive_enhanced should be string

• p1,p2,p3 in image_predictions table should be renamed to clear meaning

• wrong names in twitter_archive_enhanced (like : a , an , the)

• source column content in twitter_archive_enhanced contain HTML link tags surrounding the text.

• Some tweets are not original tweets "retweets"

• rename id in tweet_rt_fav table to tweet_id and covert it to string

• doggo, floofer, pupper, and puppo columns have values with None instead of NaN

• missin rows in tweet_rt_fav and image_predictions.(2 missing row in tweet_rt_fav and 281 missing row in image_predictions)

• missing expanded_urls in twitter_archive_enhanced

• extraction of ratings of some rows are not correct

• The rating_numerator column should of type float.

Tidiness Issues:

• doggo, floofer, pupper, and puppo column in twitter_archive_enhanced better to be one column with this value (doggo, floofer, pupper, and puppo).

• combine 3 data resources to be one dataset.

Step 3: Cleaning.

• This step is last process of wrangling data after assessing data. I follow this process of cleaning (Define, Code, Test).

• change timestamp & retweeted_status_timestamp type to datetime.

• Separate timestamp column into 2 columns date and time.

• change tweet_id type to string and rating numerator to float.

• Correct wrong extracting of rating.

• Correct wrong names by replacing wrong names by NaN.

• renaming p1, p2, p3 column names to clean names to become (1st_prediction, 2nd_prediction, 3rd_prediction).

• Correct source column content in twitter_archive_enhanced table to be without <a> tag.

• Dropping retweets tweets.

• Rename id column to tweet_id.

• Change doggo, floofer, pupper, and puppo have values with "None" to NaN.

• Combine all 3 dataframes together.

• Store datafram to Csv file.