Report: wragle_report

This report of project Data Wraling that communicates the insights and displays the visualization(s) produced from my wrangled data. This is to be framed as an external document, like a blog post or magazine article, for example.

This project have 5 steps as follows:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analyzing, and visualizing data

1. Gathering data

In this step, I will gather all three pieces of data as described below in the "Data Gathering" section in the wrangle_act.ipynb notebook.

a. The WeRateDogs Twitter archive

Download this file manually by clicking the following link: twitter_archive_enhanced.csv.

Once it is downloaded, I upload it and read it into a pandas DataFrame using Python.

b. The tweet image predictions

This file (image_predictions.tsv) is present in each tweet according to a neural network.

It is hosted on Udacity's servers and be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

c. Additional data from the Twitter API

Gather **each tweet's retweet count** and **favorite ("like") count** at the minimum and any additional data you find interesting.

Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called `tweet_json.txt` file.

2. Assessing data

After gathering all three pieces of data, assess them visually and programmatically for quality and tidiness issues. Detect and document **eight (8) quality issues** and **two (2) tidiness issues**

a. Dataset twitter_archive_enhanced.csv

Quality issues

1. Data Types

   Data Format for timestamp column in df_twitter table ("twitter-archive-enhanced.csv") must be timestamp format instead of string format

2. Data Types

   Data Format for tweet_id column in df_twitter table ("twitter-archive-enhanced.csv") must be object format instead of int64 format

3. Missing or inaccurate data

   Value data of in_reply_to_status_id, in_reply_to_user_id column in df_twitter table ("twitter-archive-enhanced.csv") have 96,67% null

4. Inaccurate data

   Value data of "name" column in df_twitter table ("twitter-archive-enhanced.csv") have "a", "quite","an", this are not name

5. Inaccurate data

   Value data of "source" column in df_twitter table ("twitter-archive-enhanced.csv") have ".. href = "http://"". It must be clean

6. Inaccurate data

   Name of "floofer" column in df_twitter table ("twitter-archive-enhanced.csv") should be "floof"

7. Inaccurate data

   Value of doggo col is none and doggo, it must be true and false
   The same with floofer ,pupper ,puppo column in df_twitter table ("twitter-archive-enhanced.csv")

8. Inaccurate data

   Value data of "expanded_urls" column in df_twitter table ("twitter-archive-enhanced.csv") have repeated value "https://" (ex tweet_id = 863062471531167744).

Tidiness issues

1. Each observation forms a row

   4 columns doggo, floofer ,pupper ,puppo in df_twitter table ("twitter-archive-enhanced.csv") need to be merged into a single column

b. Dataset image_predictions.tsv

Quality issues

1. Data Types

   Data Format for id column in df_web_tweet table ("tweet-json.txt") must be object format instead of int64 format

2. Missing or inaccurate data

   Value data of contributors, coordinates, geo,place column have 99-100% null in df_web_tweet table ("tweet-json.txt")

3. Inaccurate data

   Value data of "source" column in df_web_tweet table ("tweet-json.txt") have ".. href = "http://"". It must be clean

c. Dataset tweet_json.txt

Quality issues

1. Duplicated data

   Value data of jpg_url column in df_image table ("image-predictions.tsv") have 66 rows duplicated

2. Human error - typo

   Value data of p1, p2, p3 column in df_image table ("image-predictions.tsv") have the first character is capital, but value: ice_bear,laptop.. are not capital

3. Data type

   Data Format for id column in df_clean_image table ("image-predictions.tsv") must be object format instead of int64 format

3. Cleaning data

This step Clean all of the issues you documented while assessing.

Quality issues

1. Issue 1: Data Types

   Data Format for timestamp column in df_twitter table ("twitter-archive-enhanced.csv") must be timestamp format instead of string format

2. Issue 2: Data Types

Data Format for tweet_id column in df_twitter table ("twitter-archive-enhanced.csv") must be object format instead of int64 format
Data Format for id column in df_web_tweet table ("tweet-json.txt") must be object format instead of int64 format
Data Format for id column in df_clean_image table ("image-predictions.tsv") must be object format instead of int64 format

3. Issue 3: Missing or inaccurate data

   Value data of in_reply_to_status_id, in_reply_to_user_id column in df_twitter table ("twitter-archive-enhanced.csv") have 96,67% null
   Value data of contributors, coordinates, geo,place column have 99-100% null in df_web_tweet table ("tweet-json.txt")

4. Issue 4: Inaccurate data

   Value data of "name" column in df_twitter table ("twitter-archive-enhanced.csv") have "a", "quite","an", this are not name

5. Issue 5: Inaccurate data

   Value data of "source" column in df_twitter table ("twitter-archive-enhanced.csv") have ".. href = "http://"". It must just be clean
   Value data of "source" column in df_web_tweet table ("tweet-json.txt") have ".. href = "http://"". It must just "http://"

6. Issue 6: Human error - typo

   Value data of p1, p2, p3 column in df_image table ("image-predictions.tsv") have the first character is capital, but value: ice_bear,laptop.. are not capital

7. Issue 7: Inaccurate data

   Name of "floofer" column in df_twitter table ("twitter-archive-enhanced.csv") should be "floof"

8. Issue 8: Inaccurate data

   Value data of "expanded_urls" column in df_twitter table ("twitter-archive-enhanced.csv") have repeated value "https://" (ex tweet_id = 863062471531167744).

Tidiness issues

1. Issue 9: Each observation forms a row

   4 columns doggo, floofer ,pupper ,puppo in df_twitter table ("twitter-archive-enhanced.csv") need to be merged into a single column

2. Issue 10: Merge 3 dataset

   Need to merge 3 dataset: df_twitter table ("twitter-archive-enhanced.csv"), df_web_tweet table ("tweet-json.txt") and df_image table ("image-predictions.tsv") together

4. Storing data

This step is saving gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv".

5. Analyzing, and visualizing data

These steps are:

- Describe table twitter_archive_master.csv and conclusion
- Create Correlection data and conclusion
- Create bar chart to compare the total number of 4 type of dog and conclusion
- Create bar chart to compare the average rating of 4 type of dogs and conclusion

Thi My Hao Pham