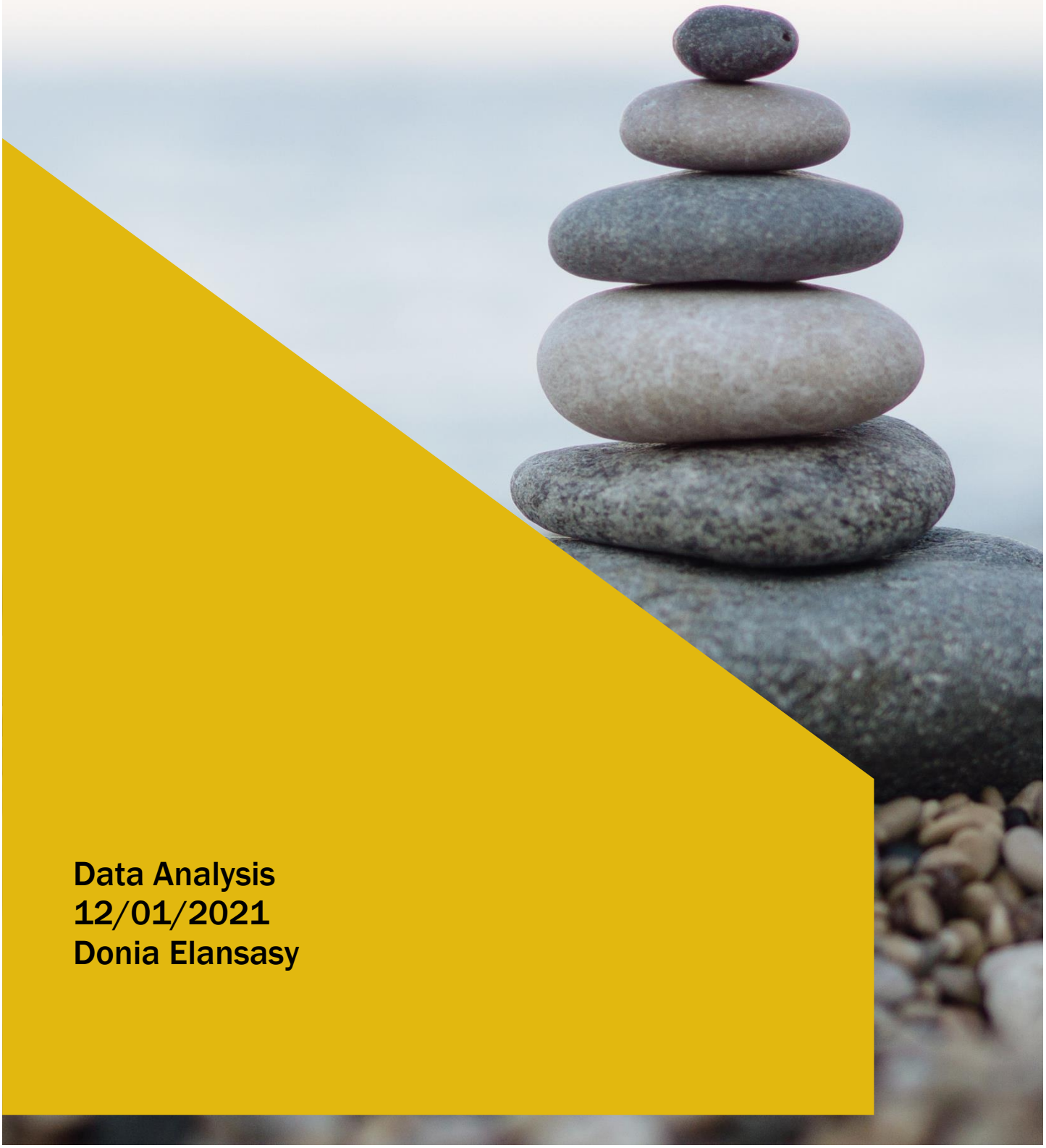
Wrangle Report



Data Analysis
12/01/2021
Donia Elansasy

Data Gathering

There were three files of data to be gathered from different sources, which are:

- archive.csv was already provided by udacity, read the file using read_csv() function in the pandas library
- image_prediction.tsv was gathered programmatically by url then extracted by read_csv function in the pandas library
- As I was denied the developer account by Twitter, I was to gather the api data from udacity directly as shown in the code comments, simply read the text file line by line

Output:

1. Archive.csv yielded df_archive
2. Image_prediction yielded df_image_prediction
3. Api.txt yielded df_api

Data Assessment:

In order to be aware of the data that has been gathered and to assess how much it might be messy or untidy I needed to check the dataframes

Visual assessment:

I opened the api.txt file and checked all of the raw data that it had and settled on retweet counts, favorite count and twitter id before extracting that data into a dataframe

Programmatic assessment:

In order to get around the data in the other 2 dataframes I used .head(), .tail() and simply calling it in jupyter notebook
I also used .info() to check the count of each dataframes rows and columns
I checked the unique values of the columns of the dataframes with .unique() and .nunique()
I checked also the count of null values, in order to get rid of them.

Output:

Quality

archive table

1. Name column with 'a' and 'none' values
2. Doggo, fluffer, pupper and puppo columns with 'None' values
3. Rimestamp column data type should be datatime (to_datetime)
4. Retweeted_status_timestamp data type should be datatime
5. Incorrect and weird values of rating_denominator
6. Expanded_url column missing values
7. in_reply_to_status_id and in_reply_to_user_id columns missing values

8. Retweeted_status_id , retweeted_status_user_id and retweeted_status_timestamp missing values
9. The row count doesn't match image_prediction table's row count, some tweets don't have images

image_predictoins table

1. Table's header is a value not a variable

api table

1. The row count doesn't match image_prediction table's row count, some tweets don't have images

Tidiness

1. Retweet count and favourit count column in api table should be in archive table
2. fix column names in image_prediction table
3. Text column got both string text of the tweet and url of the image
4. Doggo, fluffer, pupper and puppo columns should be one column

Data Cleaning:

archive table

- name column with 'a' and 'none' values
  Solution:
  Was solved by .replace() and np.nan

- doggo, fluffer, pupper and puppo columns with 'None' values
  Solution:
  Was solved by .replace() and "" empty string

- timestamp column data type should be datatime (to_datetime)
  Solution:
  Separated the date year/month/day from the rest of the timestamp for easier use in visualization process by regex and .str.extract before converting the column to datetime data type
  Was solved by .to_datetime()

- retweeted_status_timestamp data type should be datatime
  Solution:
  Was solved by .to_datetime()

- Incorrect and weird values of rating_denominator
  Solution:
  Was solved by finding out the index of the row of the value that didn't make sense and then fetching the text of the tweet itself for further investigation

Usually the correct rating is mentioned in the text column, once found it is replaced by the correct number by the .replace()

- expanded_url column missing values
  Solution:
  When the tweets with no images was removed by locating the tweet_id with the image_prediction data frame with .loc() and .isin()
  The expanded_url column had no more null values

- in_reply_to_status_id and in_reply_to_user_id columns missing values
  Solution:
  The tables were dropped after dropping the unnecessary tweets with no images as they had no value in the analysis of the data as a whole

- retweeted_status_id , retweeted_status_user_id and retweeted_status_timestamp missing values
  Solution:
  The tables were dropped after dropping the unnecessary tweets with no images as they had no value in the analysis of the data as a whole

- the row count doesn't match image_prediction table's row count, some tweets don't have images
  Solution:
  Figure that means that the archive dataframe has tweets that they are missing images and should be removed
  Got the unnecessary tweets removed by .loc() and isin() to check the matching tweet ids between archive and image prediction data frames

image_predictoins table

- table's header from value to variable
  Solution:
  The jpg_url and img_num columns were dropped from the image_predictions table and was moved to the archive datafame
  Wide_to_long() function was used to gather all the predictions under one column, the prediction confidence under on column and is_dog under one column and added trial as there are multiple trials for prediction.

api table

- the row count doesn't match image_prediction table's row count, some tweets don't have images
  Solution:
  Same as archive dataframe we needed to shed out the unnecessary rows of data of tweets that doesn't have images essential for our analysis
  With .loc() and isin() matching the tweet ids of image prediction and api dataframes

Tidiness

- retweet count and favorite count column in api table should be in archive table
  Solution:
  Merged both data frames together using merge() on tweet_id as they both have that in common, merging the the retweet and favorite columns to a master dataframe

- fix column names in image_prediction table
  Solution:
  Fixed the names of p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog,p2_dog,p3_dog for an easier use of wide_to_long() function

- text column got both string text of the tweet and url of the image
  Solution:
  Separated the url or the image form the text of the tweet itself using regex and .str.extract function

- doggo, fluffer, pupper and puppo columns should be one column
  Solution:
  Was solved by simply concatenating the 4 columns together, then replacing the empty string with np.nan using .replace()

Output:

Two files were created as .csv which are

- archive_combined.csv
- image_prediction.csv