wrangle_report

July 13, 2021

1 Introduction

The main goal of the project were to apply:

1 - Data wrangling, which consists of the main three parts: - Gathering data - Assessing data - Cleaning data

2 - Storing, analyzing, and visualizing our wrangled data

3 - Reporting whcih has: - Data wrangling efforts - Data analyses and visualizations

2 Gather

The data was gathered from three main sources. Twitter Archive Enhanced, is the first source that has been provided by us through CSV. We have loaded the source by using **pd.read_csv**. The second source is Image Predictions. We have downloaded the file programmatically using the Requests library in python. Twitter API was the third source to gather from. We created a new development twitter app to generate API keys, which will let us interact with their API. The file *tweet_json.txt* contains each tweet's JSON data written into its own line. Then, loading each line and extract the most important columns to a dataframe.

We have applied all these three methods in order to gather the data.

3 Assess

In the assessing step, we looked into all the dataframes for any quality or tidiness issues. In addition, we documented all the issues and list them in a notebook. However, below is the list of all the issues we found in the assessing step:

Quality

For The `Twitter Archive` table

- Delete Retweeted tweets
- Delete unnecessary columns (retweeted_status_user_id,retweeted_status_id,retweeted_status_timestamp,e
- Replace faulty names or uncorrect names
- Separate timestamp to (month_arc, year_arc)
- Convert tweet_id type to str

For The `Image Predictions` table

- Capitalize the first letter of each word in (p1, p2, p3) columns
- Remove the underscore between the words in (p1, p2, p3) columns
- Convert tweet_id type to str

For The `Tweets Description` table

- Delete Retweeted tweets
- Separate timestamp to (month, year)

Tidness

- Convert (doggo,floofer, pupper, puppo) to one column (Kind)
- Merge (Inner Join) all three Dataframes into one dataframe, doing inner join will automatically exclude deleted tweets.

After doing the inner join the volume of the data got decreased to 654, this merge will automatically exclude the deleted tweets.

4 Clean

In the Cleaning process, all the above issues were resolved in this section. We have mainly used pandas and numpy library and we will list all the fucntions used.

- copy()
- reset_index()
- query()
- drop()
- lower()
- replace()
- to_datetime()
- rename()
- value_counts()
- join()
- merge()