Data Wrangling Report

Introduction

In this project, I will describe the data wrangling process that I worked on.

Data Gathering:

In the data Gathering step, I have gathered three pieces of data

1- **Twitter archive**

The first data frame is “twitter archive enhanced “, I have downloaded this file manually by clicking the link then upload it to the project workspace**.**

2- **Image predictions**

The second data frame is “Image predictions“, I have downloaded this file programmatically using the Requests library.

3- **Tweet json**

The Third data frame is “Tweet json”, since I can't set up a Twitter developer account for some reason. So, I have downloaded the file ‘tweet_json.txt’ then read it line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

Data Assessing:

After gathering the data, I have assessed them visually and programmatically for quality and tidiness issues.

1- Quality Issues:

Low quality data has content issues ex: missing, duplicates, incorrect data:

First data frame

1- Missing values(NaN); in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
2- tweet_id, in_reply_to_status_id, in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id should be object.
3- rating_denominator column has values other than 10
4- None in columns names,doggo,floofer,pupper,puppo should change to NaN.
5- timestams and retweeted_status_timestamp shoud be datetime
6- incorrect dog names starting with lowercase characters, ex(a, an, the)
7- Clean the tweet source.

Second data frame

8- tweet_id should be an object rether than integer.
9- Column headers are not discriptive

Third data frame

10- tweet_id should be an object rether than integer.

2- Tidiness Issues:

Untidy data has structural issues:

11- The Four columns doggo,floofer,pupper and puppo should be in one column.
12- The three dataframes 'df_1' table, 'df_2' table, 'df_3' table need to be combined in one dataframe because the rows in each are all for the same observations.

Data Cleaning:

In data cleaning, I will clean the issues documented in assessing step.

- Drop the columns that have many null values and doesn't necessary for analysis.
- Change the datatype in df_1 table of column (tweet_id) to object.
- Set 10 for all cells in rating_denominator column.
- Replace None in columns names,doggo,floofer,pupper,puppo should to NaN.
- Change datatype of timestamp to datetime.
- Replace names with lowercase values to NaN.
- Change the datatype in df_2 of tweet_id column to object.
- Change column headers to be more descriptive.
- Change the datatype in df_3 of tweet_id column to object
- Merge the doggo,floofer,pupper and puppo columns to one column named stages.
- Drop doggo,floofer,pupper and puppo columns.
- Combine the three dataframes 'df_1' table, 'df_2' table, 'df_3' table to one dataframe.
- Strip all html tags and retain the text in between the tags. Convert the datatype from string to categorical.