wrangle_report

September 30, 2022

1 Wrangle Report

The project focused on wrangling data from the 'WeRateDogs' account on twitter. The efforts put into the wrangling process are highlighted below:

- Data Gathering
- Assessing the Data
- Cleaning the Data
- Analysing the Data
- Visualising the Data

1.1 Data Gathering

The data was gathered from three different sources: 1. A twitter-archived csv file containing information about the kind of dog per dogsize, wheter the tweet was a retweet and the day and time of the tweet. it was then loaded into a pandas dataframe

2. An Image predictions tsv file that was downloaded programmatically using the requests library of which a folder was created for and the file was loaded into a pandas dataframe
3. Additional twitter API data, an alternative approach was used as made available to me via the tweet-json text file and the code provided to load the data was used.

1.2 Assessing the Data

All data were assessed via two main approaches, the virtual assessment and the programatic assessment which required the use of some pandas functions to describe, summarise the data. The quality issues found were documented as well as the tidyness issues.

The issues are highlighted as below:

1.2.1 Quality issues

1. Twitter Archived - rating_denominator has values other than the normal value of 10 in some cases
2. Twitter Archived - Some records have entries for retweeted_status_id
3. Twitter Archived - Missing data in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
4. Twitter Archived - timestamp and retweeted_status_timestamp are object type instead of datatime type
5. Twitter Archived - text in source column contains href tag and un-needed twitter url
6. tweet_id column are off integer data type
7. Twitter Archived - inconsistent records found in expanded_url column like 'https://www.gofundme.com/mingusneedsus,https:/..'
8. Twitter Archieved - records like 'None','a','an' found in the name column which is unlikely in reality, inconsitent alphabetic case

1.2.2 Tidiness issues

1. Twitter Archived - doggo,floofer,pupper,puppo columns representing dog sizes are captured in four different columns
2. Twitter Archived - timestamp column break the tidiness rule 'Each variable forms a column', it has two variables date and time
3. Separate dataframes 'twitter archived, image predictions, twitter API data' containing the desired information.

1.3 Cleaning the Data

Copy's of the original dataframes were created and the copied dataframe was then cleaned after assessment to resolve both quality and tidyness issues found. Aftwards the three dataframes were combined and saved into one master csv file. This is an important step as to ensure we are able to find the right insights while analysing the data.

1.4 Analysing the Data

The data was further analysed using a number of pandas methods as well as other python functions to find notable insights within the data.

1.5 Visualising the Data

After analysis the insights found were then translated into visuals using charts and graphs by pandas plot function. Visualising this insights is an important step in communicating them to the various stakeholders.