wrangle_report

June 1, 2024

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

0.1.1 Twitter archive data

Data Gathering and Assessing

1. the twitter archives enhanced is loaded in to a dataframe df_1.
2. the .info method indicates data types issues with tweet_id and timestamp columns and some missing values with other columns.
3. the .sort_values method when applied on rating_numerator and rating_denominator indicates that there are numerators and denominators that are less than 10.
4. the name, doggo, floofer, pupper, puppo columns have values None which are actually NaN.
5. the name column have values other than name like an, quite and etc..
6. there is retweeted_status_id column with values and NaN.
7. the .describe method revealed that there are tweets that have higher ratings which might be outliers are wrongly interpreted.
8. the data is assessed for duplicates and no duplicates are found.

Data Cleaning

1. the tweets that are retweets and tweets with no image urls are filtered from the dataframe.
2. then the columns with zero non null values are dropped from the dataframe.
3. the tweet_id and the timestamp are changed to respective data types.
4. the reply fields are also dropped from the dataframe.
5. the name field is renamed to dog_name to maintain naming convention.
6. the names of the dog values checked and is corrected.
7. the dnominator values less than 10 are selected and is correctly replaced as per the tweet.
8. the source columns are extracted for correct field values.
9. the dog_stage is created by melting the ′doggo′, ′floofer′, ′pupper′, ′puppo′ columns.
10. the dog_stage is changed to category data type.

0.1.2 Tweet image prediction

Data Gathering and Assessing

1. the url is downloaded programmatically.

2. the .info method indicates no missing values and data type issue with tweet_id.
3. the dataset has column names that are confusing with the value it holds.
4. there are some images that are predicted as dogs and some predictions are not dogs.
5. the predicted values are inconsistent data types.
6. the .describe method indicates that the p1_conf has value 1 which is to be inspected for what the predictions are for.
7. the data is assessed for duplicates and no duplicates are found

Data Cleaning

1. the tweet_id is changed to object data type.
2. the columns are renamed and dropped as required for analysis.
3. the ′prediction_1′, ′prediction_2′, ′prediction_3′ columns are curated by replacing the _, - with spaces and made as title case.

0.1.3 Twitter API

Data Gathering and Assessing

1. the twitter developer account is created to get the key and token and the json file is downloaded as .txt file
2. the tweet_id, retweet_count and favorite_count has been collected as a list and converted to dataframe.
3. the .info method reveals that there are no missing values in the dataframe.
4. the .sort_values mehtod indicates there are many tweets with zer0 favorite count which might be an error in the the data.
5. the .describe method reveals there are tweets with zero retweets.
6. the data is assessed for duplicates and no duplicates are found

Data Cleaning

1. the zero values are from the data source so we are not accepting as it is