wrangle_report

August 26, 2022

0.1 WRANGLING REPORTING:

GATHERING STAGE OF THE DATA WRANGLING PROCESS For this Project, three Datasets were gathered. The Datasets include WeRateDogs Twitter archive data, the image predictions Dataset and the Retweets and likes Dataset fetched from twitter Database. The Twitter archive Data was directly downloaded in a csv format. The image prediction Dataset was downloaded programmatically. The retweets and likes Dataset was gathered using the tweepy library via the twitter API. All three Datasets were loaded into a pandas data frame.

ACCESSING STAGE OF THE DATA WRANGLING PROCESS The Datasets were accessed for Data quality and tidiness issues. The assessment was done visually and programmatically. The visual assessment was done by just skimming through the Data while the programmatic Dataset involved the use of pandas functions. Some of the functions used in accessing the Dataset includes: - df.head()- This is used to just display the first few rows of the Data. - df.info()- This is used to display information about the Datasets such as number of columns, column labels, column data types, memory usage, range index. - df.dtypes- This is used to display the different data types of the columns - df[df.duplicated()]: This is used to check for duplicates in the whole Dataset - df[df.duplicated('colname')]: This is used to check for duplicates in a particular column that should hold unique values. - df.value_counts(): This is used to return the count of unique values in a column

After the datasets were accessed programmatically and visually, the quality and tidiness issues were documented in few sentences for cleaning. The Dataset cleaned for different tidiness and quality issues which included:

QUALITY ISSUES

1. Incorrect ratings in the twitter_df dataframe
2. Some tweets in the twitter_df are retweets not dog ratings
3. Some tweets in the twitter-df are replies not dog ratings
4. Incorrect names for some of the dogs.
5. Incorrect Datatype for the timestamp column in Twitter_df table
6. Twitter_Id is stored as integers in all three Datasets.
7. create_date column in twitter_data is stored as object instead of DateTime
8. Some of the null values are stored as none instead of NaN

TIDINESS ISSUES 1. The create_date duplicate column should be dropped 2. Dog stages (doggo, floofer, pupper, puppo) are spread in different columns. 3. The dog breed predictions are spread out in different columns. 4. All three datasets should be merged into one.

CLEANING STAGE OF THE DATA WRANGLING PROCESS The Data was cleaned for the different quality and tidiness issues by using different pandas functions like the replace, merge, drop functions. After the individual datasets were cleaned, they were merged into a single file.

In [ ]: