Reporting: wrangle_report

Introduction

This report describes the wrangling efforts performed in the **Wrangle and Analyze Project.** The wrangling stage involves gathering, assessing and cleaning datasets (i.e., twitter archive, image predictions and twitter api data). After this stage, the datasets are combined and stored as one master dataset before data analysis and visualisation.

The following sections outline the process undertaken in wrangling the three datasets provided for the project.

Gathering Data

The datasets were gathered (via manual download, programmatic download and Twitter API extraction) and loaded as pandas dataframes.

Note: The twitter api data was downloaded manually due to the challenges encountered during the API extraction using “tweepy” library.

Assessing Data

The datasets were visually and programmatically assessed to identify both data quality issues (i.e. dirty data) and tidiness issues (i.e. messy data). This involves checking for issues such as duplicate entries, missing values, invalid and inaccurate values and irrelevant columns. The data quality and tidiness issues identified are:

1. The **source** column contains redundant values ('anchor' html tags) attached.
2. The **timestamp** column has a wrong datatype (object).
3. Column names are non-descriptive hence do not provide context for the image prediction data.
4. There are duplicate images in the image predictions.
5. Wrong data type (i.e. integer) for **tweet_id** columns.
6. There are some unnecessary columns in the twitter archive dataset.

7. Dog breed names in the image predictions data are in lower cases and have underscores.
8. Some tweet entries have more than one dog stage (i.e. doggo, puppo, pupper and floofer).
9. The retweets are on the same table as the original tweets.
10. The dog stages (doggo, puppo, pupper and floofer) are on separate columns

Note: Only tweet entries that are original ratings and have images were assessed in this stage. A total of 1971 entries met this condition.

Cleaning Data

The Define-Code-test Framework was used to clean the datasets to resolve the aforementioned issues. The framework involves defining the steps to clean the data based on identified issues, writing codes to clean the data and testing to check if these issues have been resolved. The following are data cleaning steps undertaken to fix the issues:

1. Extract the twitter sources from the anchor html tags.
2. Convert **timestamp** column to datetime datatype
3. Change the names of all columns in the image prediction data for easy interpretation.
4. Remove duplicate image urls in the data
5. Convert the datatype of the **tweet_id** columns to string
6. Drop unnecessary columns in the data
7. Change the dog breed names to proper case and replace underscores with whitespaces.
8. Re-extract the accurate dog stages from the **text** column in a new column.
9. Separate retweets data on a new table
10. Drop the dog stage columns (doggo, puppo, pupper and floofer)

Storing Data

After the data cleaning process, the three datasets are combined and stored as one master dataset to be used during the analysis and visualisation stage.

Note: Only the tweet_id entries in the cleaned twitter archive data were combined with the cleaned image prediction data and retweet counts data.