Wrangle Report

By Tanmay Jain

Introduction

This report is a part of Udacity Data Analyst Nanodegree program. This project involves wrangling of data from various sources associated with tweets from the user WeRateDogs(@dog_rates). WeRateDogs rates pictures of people's dogs. Instead of using the typical 10/10 format, it rates in a fun way and gives a rating more than 10 because they say Dogs always deserve a 10 and sometimes more.

This wrangling report describes the Gathering, Assessing and Cleaning of the data.

Gathering Data

The data used in the project was gathered from three different sources.

- **Twitter Archive File**: This dataset was provided by Udacity and had to be downloaded manually.
- **Image Predictions File**: This file was to be downloaded programmatically using the requests library.
- **Twitter API**: We had to use the "Tweepy" library in Python to gather tweets data by the tweet_id provided in other datasets.

Assessing the Data

After gathering the data, it was assessed using pandas functions like

- .head()
- .info()
- .duplicated()
- .value_counts()
- .contains()
- RegEx

After visualizing the data, it was easy to find the issues in them. These issues could be divided into 2 categories:

1. **Quality Issues**
    - Data contains retweets.
    - Tweet_id had incorrect data types.
    - Name column had “None” instead of NaN.
    - Name column had values like “a”, “such”, “all”.
    - Dataset consists of unwanted columns.
    - Standardizing the columns rating_numerator and rating_denominator into a single column rating = numerator / denominator.
    - Source column is in HTML format with action tags.
    - Decimal ratings not extracted properly.
    - There are 2075 rows in the dataset which is not equal to the number of tweet_ids(2356) meaning incomplete data or tweets without images.
2. **Tidiness Issues**
    - Columns "floofer", "doggo", "pupper" should be single column.
    - The three datasets should be one since they are related.

Cleaning Data

To avoid the tampering of the original data while cleaning, we create a copy of the original datasets using the pandas “.copy()” feature.

The cleaning process consists of three parts:

- **Define**: Define the problem which is to be cleaned.
- **Code:** Remove the problem by coding.
- **Test:** Check for assuring the removal of the problem.

In this process, we fix the issues we encountered in the assessing step. This is done using functions like:

- .str.cat()
- .astype()
- .replace()
- .drop()
- .apply()
- RegEx
- .islower()
- .str.contains()