Wrangle & Analyze WeRateDogs Data

Wrangle Report

Introduction

The dataset that will be wrangled for this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. (11/10, 12/10, 13/10).

Project details

The tasks of this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering the Data:

The data for this project is in *three* different formats:

1. **Twitter archive file:**

   WeRateDogs downloaded their Twitter archive and shared it exclusively for use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

   The **twitter-archive-enhanced.csv** was provided by Udacity, downloaded manually then loaded from the CSV file into a pandas data frame for this project.

2. **Image Predictions:**

   The tweet image predictions, i.e., what breed of dog is present in each tweet according to a neural network is stored in this file. It was hosted on Udacity's servers in tsv format and had to be downloaded programmatically using the Requests library.

   The content of **image-predictions.tsv** file is then loaded into the pandas' data frame. It consists of 2075 rows and 11 columns. The table contains the top three predictions, tweet

ID, image URL, and the image number that corresponds to the prediction with highest confidence.

3. **Twitter API File:**

Twitter API file contains tweet id, favorite count and retweet count. Data was provided by Udacity, downloaded manually then was loaded from the tweet-json.txt file into a pandas' data frame. The dataframe size is 2354 rows and 2 columns. The tweeter ID column has been used as an index.

Assessing the Data:

After gathering the data, the three tables were saved and assessed visually and programmatically for checking tidiness and quality issues.

Visual Assessment is helpful in identifying issues such as non-descriptive column headers and repetitive columns in the three Datasets.

Programmatic Assessment is useful in listing the quality issues such as incorrect data types and duplicate data that are present in the three datasets.

1. **Twitter Archive**

Initially a sample of data is assessed visually, and to get familiar with the data, we look at the description of all the columns. A summary of data types and non-null values is later displayed which helps in identifying columns with the incorrect data type and/or null values. Later, IDs are checked for duplicates. Next, the number of tweets which are replies and retweets is calculated.

Rating numerator and denominator are checked visually by printing few samples of data and later ratings with denominator greater than 10 are also printed out for further investigation. We then check programmatically the text column for any float ratings.

Expanded urls are also assessed visually and later programmatically for checking if there are 2 or more urls present in one column.

Name of dog column is assessed programmatically and checked for the number of values. And all tweets were checked for dogs if they have more than one category assigned to them.

2. **Image Predictions file**

Again, as a first step, a sample of data is assessed visually, and to get familiar with the data, we look at the description of all the columns. A summary of data types and non-null values is later displayed which helps in identifying columns with the incorrect data type and/or null values. Later we check if there are any duplicates in the jpg url column. Also, the 1st prediction is checked to see how many images have been classified as dog images.

3. Twitter API Data

To begin with, a sample of data is assessed visually, and a summary of data types and non-null values is later displayed which helps in identifying columns with the incorrect data type and/or null values. Then IDs are checked for duplicates.

Cleaning the Data:

The Quality and Tidiness issues that were listed after assessing the data are cleaned now using pandas. I followed the programmatic data cleaning process – Define, Code and Test for all the 3 datasets.

1. Twitter Archive File

As a first step, copy of the dataset is created which will be used for cleaning throughout the entire process.

Few of the gathered tweets are retweets and replies and so are removed. Along with that there are some unnecessary columns which we will not be using for further analysis and hence are dropped from the dataset.

Dog classification(stages) have 4 different columns doggo, floofer, pupper or puppo which is merged into one column.

The timestamp has an incorrect datatype (is an object) which is corrected to DateTime.

Some ratings with decimals are read incorrectly from the text of the gathered tweets which are corrected. Denominator of some ratings is not 10. Numerator of some ratings is greater than 10. So, it is confusing to interpret unstandardized ratings. To understand what % are below or above 100% a single value for rating is added as a new column to the dataset.

There were 639 expanded URLs which had more than one url address, which is corrected using the tweet id field.

2. Image Prediction

Firstly, a copy of dataset is created for use throughout the cleaning exercise. As some of the column names are confusing and do not give much information about the content, so we rename columns.
Then we clean dog breeds - we replace underscores with whitespace and capitalize the first letter to have consistent and clean formatting.
66 image_url duplicates were removed.

There was no cleaning task performed on the Twitter API dataset.

As the last step, all the 3 datasets were merged into one master dataset and converted to **twitter_archive_master.csv** file.