wrangle_report

June 14, 2018

1 Wrangling Report of WeRateDogs

1.1 Introduction

This project is focused on wrangling data from the WeRateDogs Twitter account using Python.This Twitter account rates dogs with humorous commentary.Here Data wrangling is performed in three stages. 1.Gather 2.Assess 3.Clean

1.1.1 Gather

We gathered three pieces of data for this project and saved them as three data frames.

- We manually downloaded the file 'twitter_archive_enhanced.csv' and saved it to the dataframe df_archive.

- Programatically downloaded 'image_predictions.tsv' file using python Requests library and saved it to the dataframe df_images.

- Using the tweet IDs in the WeRateDogs Twitter archive, we could query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data stored in a line.Saved the file as a text file tweets_json.txt and read the file and saved the file in dataframe df_tweets

1.1.2 Assess

After gathering each of the above pieces of data, we assessed them visually and programmatically for quality and tidiness issues was our next step. We could detect and document the following quality issues and tidiness issues.

Quality

- There are several columns with empty values 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'date_time','in_reply_to_status_id', 'in_reply_to_user_id' and 'user_favourites'.
- Several duplicated tweet ids and tweets with no images.
- Values entered in the rating_numerator column is different from real numerator values that are provided in the ratings in text column along with description.

- Ratings not given as one column.Instead two different columns are given rating_numerator rating_dinominator.
- Columns with dog breeds and predictions confidence can be condensed.

- Some columns are not in appropriate data types.
- Change datatypes of timestamp to datetime, dog_stage to categorical, and tweet_id, in_reply_to_status_id, and in_reply_to_user_id to strings.
- Categories in Sources colmns are not easily readable.

Tidiness

- The four dog Stages(doggo, floofer, pupper, puppo) have four different columns, instead of one column for stages filled with the values.
- All tables should be part of one dataset.

1.1.3 Clean

Cleaning our data is the third step in data wrangling. It is where we fixed the quality and tidiness issues that we identified in the assess step.We did both the manual and programmatic cleaning.First we made a copy of all the three dataframes.We repeated 'Define','Test' and 'Code' method for all the issues. Finally saved the cleaned file to a CSV file named 'twitter_archive_master.csv'.

1.2 Conclusion

Data Wrangling is a core skillset every data analyst should be familiar with.Analyzing and Visualizing data without cleaning the data will be mere waste of time,as we may end up at totally incorrect conclusions. My biggest challenge while performing the wrangling was using the twitter API to gather the JSON data.