Data Wrangling Report

Ahmed Tariq Balkhair

Contents

Introduction

Real-world data rarely come clean. Using Python and its libraries, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. We will document our wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that we will be wrangling (and analyzing and visualizing) 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.

The purpose and goal of this project are to create a trustworthy and interesting analysis and visualization based on the wrangled data.

Project Details

This part of the project is divided into three steps, which are as follows:

- Data Gathering
- Data Assessment
- Data Cleaning

Data Gathering

This project requires working on three different datasets, which are acquired as follows:

- **Twitter archive:** The twitter_archive_enhanced dataset is a csv file that was given by Udacity to be downloaded manually from their side, uploaded into the Jupyter notebook, and read into a panda's DataFrame.
- **Tweet image predictions:** This file (image_predictions.tsv) is present in each tweet according to a neural network. It is hosted on Udacity's servers and should be downloaded programmatically using the Requests library.
- **Twitter API and JSON:** We used Udacity's ready-scrapped Twitter data, which was stored in txt file, we have extracted the important data (tweet-id, retweet-count, and favorite-count) and stored it in a separate DataFrame.

Data Assessment

After gathering all three pieces of data, assess them visually and programmatically for quality and tidiness issues. Visually by displaying data in the Jupyter Notebook and in Excel sheets. And programmatically by using pandas' functions and/or methods used to assess the data.

Quality issues

The four main data quality dimensions are:

- **Completeness:** missing data?
- **Validity:** does the data make sense?
- **Accuracy:** inaccurate data? (wrong data can still show up as valid)
- **Consistency:** standardization?

df_archive table

- **Completeness:**
  - Missing data in columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, because original tweets are not isolated from retweets.
  - Remove columns not needed for analysis: in_reply_to_status_id, in_reply_to_user_id, source, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls.
  - Create a rating column, which contains the results of the rating_numerator divided by rating_denominator.
  - Split timestamp column into date and time columns and remove the timestamp column.

- **Validity:**
  - 'None' values in the name, should be 'NaN'.
  - 'None' values in dog stage columns, should be 'NaN'.
  - There are invalid name data: 'a' or 'an'.
  - Some records have more than one dog stage.

- **Accuracy:**
  - Erroneous datatypes:
    - tweet_id datatype is int, should be str.
    - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id data types are float, should all be str.
    - retweeted_status_timestamp, timestamp should be DateTime instead of str.
  - Abnormal values in both rating_numerator and rating_denominator.

- **Consistency:**
  - rating_denominator should be a standard 10.
  - The source column still has the HTML tags.

df_image table

- **Completeness:**
  - There are jpg_url duplications.
  - Unnecessary columns for dog breed prediction (p1, p2 , and p3) i.e. could be packed into one column and get rid of the other three and any unneeded columns.
  - Unnecessary columns for prediction confidence (p1_conf, p2_conf, and p3_conf) i.e. could be packed into one column and get rid of the other three and any unneeded columns.
- **Accuracy:**
  - Some image predictions are not dog breeds.
  - Erroneous datatype for column tweet_id, should be str.
- **Consistency:**
  - The dog breeds in the columns (p1, p2, and p3) are not consistently lower or uppercase, and an underscore is used instead of space.

df_tweet table

- **Completeness:**
  - Few missing data.
- **Accuracy:**
  - Erroneous datatype for column tweet_id, should be str.

Tidiness issues

Three requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

1. Dog stage is in 4 columns (doggo, floofer, pupper, puppo), no need for that.
2. Merge df_tweet and df_image into df_archive.

Data Cleaning

After we assessed the gathered data, we made copies of the datasets, merged them into one, and fixed the quality and tidiness issues, and this is the third and final step in the data-wrangling process. This step consists of three stages that would be applied to each assessment point in order for it to be cleaned. The stages are:

- **Define:** define how you will clean the issue in words.
- **Code:** convert your definitions into executable code.
- **Test:** test your data to ensure your code was implemented correctly.

Data Storing

Finally, we stored the cleaned data as a csv file to be ready for analysis and visualization.