WRANGLE REPORT

Data Wrangling Steps: Gather, Assess, and Clean

By

Ebunoluwa Alabi

29-11-22

Introduction

Real-world data rarely come clean. Using Python and its libraries, you will gather data from various sources and formats, assess its quality and tidiness, and then clean it. This is called data wrangling.

This project walks through cleaning data from different sources and combining them to draw critical insights.

Some of the project goals include;

1. The wrangling of the data collected
2. Storing, analyzing, and visualizing wrangled data
3. Reporting on the data-wrangling efforts

Data Gathering

I collected data from the following sources;

1. The WeRateDogs Twitter archive provided in the project overview by Udacity, The “twitter-archive-enhanced.csv” contains basic information like tweet-id, text, timestamp, etc.
2. The image prediction file was also supplied by Udacity and includes the dog breeds and the level of prediction accuracy.
3. Twitter API (Tweepy) was used to collect data like the latest tweets from the @dog_rates handle and other interesting information.

Assessing Data

With the data collected, I started to assess the quality of the data collected;

1. **Quality dimensions - ‘twitter-archive-enhanced.csv’**
    i. **Completeness**; missing data in the following columns: in_reply_to_status_id, retweeted_status_id, expanded_urls

-The tweet_id is recorded as an integer across all tables, I needed to change to string format eventually.

ii. **Validity**; a lot of the cells under dog names have ‘none’ in them.

- This data set contains retweets suggesting there is duplicate data

iii. **Accuracy**; The timestamp data is presented as objects

iv. **Consistency**; All the ratings data should be capped at 10

b. Tidiness requirements

i. There are four columns all related to the same issue (dogoo, floofer, pupper, puppo)

2. Quality dimensions - ‘image_predictions.tsv’

i. **Completeness**; All the cells actually seem to have relevant data to the overall dataset

ii. **Validity**; p1, p2, and p3 have invalid data some dogs have been labeled as bagel, banana, and spatula, suggesting that the predictions can not be trusted.

iii. **Accuracy**; the p1_conf column has high confidence levels for a Labrador Retriever and still labels t=it false in the p1_dog column

iv. **Consistency**; The sentence cases for the breed name are inconsistent, with mixtures of lowercase and uppercase.

b. Tidiness requirements

i. This data set seems to be observationally organized for the single table it belongs to.

3. Quality dimensions - ‘tweet_json’

i. **Completeness**; There are some missing data sets

b. Tidiness requirements

i. This data set seems to be observationally organized for the single table it belongs to.

Cleaning Data

In cleaning the data, the following steps were followed;

1. Define; I decided exactly what needed to be cleaned
2. Code; Wrote the code to clean it
3. Test; Then tested to make sure it was clean

Cleanings tasks completed;

1. I merged all the gathered file
2. I created a single column for all the dog types
3. I deleted unused data
4. I deleted unused columns
5. I removed duplicate data
6. I converted tweet_id from integer to string
7. I changed the timezone format
8. I corrected existing naming issues
9. I made the rating columns uniform
10. I analyzed the image data frame
11. I created a dogbreed column from the image_prediction portion of the data frame