2021

Human Besties

Salma Gouda

Udacity Projects

1/6/2021

Data Wrangling Report

By Salma Ahmed Gouda

1/6/21

This report is presented as an assignment for the Udacity Data Analyst Nanodegree. In this report, the main steps involved in the data wrangling of the Twitter account "WeRateDogs'" are illustrated.

Data Gathering

In this first step, we collect data from different sources through various methods. For this project, there were three main sources for the data required:

1. Twitter_archive_enhanced.csv file, the first file was provided in the Udacity project classroom and downloaded manually to our working directory and then imported into our working environment using Pandas function "pd.read_csv".

2. Image_predictions.tsv file, the second file that has been hosted on a webpage and downloaded from its relevant URL using the Requests library "get" function and Pandas function "pd.read_csv". This file contained image predictions for the dog breeds obtained through a neural network on most of the tweets in the twitter archive file.

3. The third and final dataset was gathered from Twitter REST API via the Tweepy library by querying the API to obtain extra information pertinent to the tweets ids in the first file, e.g. retweet count and favourite count aspects. In case of dtudents having troubles establishing their Twitter developer account, the code and file containing information was provided for them in the Udacity project classroom.

Data Assessing

In this second step, we investigate our imported datasets both visually and programmatically for quality and tidiness issues.

Visual Assessment:

The visual assessment is done on a spreadsheet application or in a Jupiter notebook to find some issued than can be easily detected with the eye.

Programmatic Assessment:

The programmatic assessment is done in a Jupiter notebook using Pandas functions:

- **head()**
- **info()**
- **columns()**
- **value_counts()**
- **sample()**
- **describe()**

While assessing the data, I found the following issues:

Tidiness Issues:

- All tables should be part of one master dataset then duplicate columns should be removed

a. archive_df

- values are column names (doggo, floofer, pupper, puppo)
- timestamp column should be date and time columns

Quality Issues:

a. archive_df

- retweets (3 tables)
- tweet_id dtype should be a string
- inaccurate name data
- inaccurate dog ratings (numerator and denominator data)
- some columns aren't useful for analysis

b. image_predictions

- 66 duplicated urls
- Various columns for dog breed predictions and their confidence levels
- tweets with no images
- some columns aren't useful for analysis

Note: There're more issues with the datasets, but those are the ones I decided to fix according to the project rubric

Data Cleaning

In this third and last step in data wrangling, we define the issues we found, build a code to fix it, and then test if our code worked. We should end up with a clean dataset for further analysis and visualization.

Before cleaning, we should first make copied of the datasets to work with.

The issues addressed in the assessment process were fixed as follows:

Tidiness Issue	Solution
• All tables should be part of one master dataset then duplicate columns should be removed	• Merge the 3 datasets into one using "concat" function then remove duplicate columns
• Values are column names (doggo, floofer, pupper, puppo)	• Create one column to include the values (doggo, floofer, pupper, puppo) using "extract" function, then drop the values columns using "drop" function
• "timestamp" column should be date and time colu mns	• Convert timestamp column into datetime format using "to_datetime" functio n then convert timestamp column into date and time columns

Quality Issue	Solution
• Retweets (3 tables)	• Keep only original tweets (delete all retweets) by filtering NaN values in the retweeted_status_user_id column
• "Tweet_id" dtype should be a string	• Change tweet_id dtype from *int* to *str* using "astype" function
• Inaccurate name data	• Change inaccurate name data to None using "replace" function
• Inaccurate dog ratings (numerator and denominator data)	• Fix dog ratings (numerator and denominator issues) both manually and programmatically then merge both columns into one "rating" column
• 66 duplicated *"jpg_urls"*	• Drop 66 duplicated "jpg_url" using "drop" function
• Various columns for dog breed predictions and their confidence levels	• Create 1 column for "dog_breed" and 1 column for "conf_lvl"
• Tweets with no images	• Remove tweets with no images
• Some columns are not useful for analysis	• Drop all columns not useful for analysis using "drop" function