Wrangle Report

By: Darine Battikh

Purpose

This report aims at documenting the steps that have been taken throughout the data preparation: collection, evaluation, and data cleaning.

Introduction

During this project, the WeRateDogs Twitter account has been relied on to provide advanced data preparation and cleaning processes. The collection of data has been made from different sources and using different upload methods. The data in each data frame required a substantial amount of evaluation and cleaning to be analysed and interpreted at the end.
In this report, every step in the data preparation and cleaning process will be detailed and documented.

Data Preparation

The first crucial step is to set up the environment through the import of the necessary packages. In this case, the packages required are as follows: pandas, numpy, requests, tweepy, json, seaborn, and matplotlib.
Then, gathering the data has been made from three different sources:

a. Upload of the “twitter_archive_enhanced.csv” file.
b. Upload of the “tweet image prediction” from a url using requests.
c. Gathering data from Twitter API which requires the creation and validation of a Twitter Developer Account. As the account wasn’t validated, the second method to upload this data has been used: uploading and reading the json file line by line in a pandas table. The upload has been successful.

An important step has been made by opening all the data frames in Google Sheets to be able to see the whole dataframes.

Data Evaluation

When evaluating the data, more than 8 quality issues and 2 tidiness issues have been detected through both visual and programmatic evaluation.
First, for the visual evaluation, the Google Sheets has been useful to check the totality of the data frames and detect the issues like missing values, dogs without names, wrong extraction of numinators, denominators different from 10, and one tidiness issue that concerns the four columns of dogs' stages instead of one.
Second, the programmatic evaluation has been implemented to detect different types of issues that can not be detected visually. For instance, the wrong datatype of timestamp, the duplicate data in retweets and images. Also, it was straightforward to detect the different sizes of the three dataframes, and the different order (ascending vs descending).

Data Cleaning

The data cleaning process focused on the quality and tidiness issues detected in the evaluation phase. For each issue to be solved, the process is as follows:
First, defining the issue and the steps to be taken to solve it. Second, turning the steps into code and executing it. Third, test the code and the solution through another code and compare it to the expected results.
As a first critical step of the data cleaning process, the "twitter archive enhanced" has been reversed to be in the same direction as the other two datasets. Then, the three datasets have been merged in one major dataset (named **df**).
Consequently, the cleaning process has been made on **df**. All the issues have been resolved: filling in the missing values, dropping the unnecessary columns, correcting the wrong data types, correcting the wrong extractions of the denominator and numerators, and cleaning the duplicate data.
Each cleaning code has been tested to ensure the correctness of the code.

Conclusion

This paper explained and documented the process of the data gathering, evaluation, and cleaning to make the **wrangle_act.ipynb** more understandable by the reader.