DATA WRANGLING

ABSTRACT

This report was made to illustrate out the approach of gathering , assessing and cleaning data in the WeRateDogs project.

Abdelrahman Emara

Tables of Content

A-) Gathering

In our first step in this project (Gathering) which is based on gathering the data required to do the analysis later on; we gathered our data from three sources:

1- Downloaded the twitter-archive-enhanced.csv file manually from Udacity Resources.

2- Downloaded programmatically image-predictions.tsv from the URL given

3- Downloaded the tweet-json.txt file manually since I didn’t have access on the twitter-api yet.

Then we read each file using the pd.read_csv() function to read our files whether they are .csv or .tsv or .txt . and they were saved as data frames through the following variables:

1- df_archive

2- df_pred

3- df_json

B-) Assessing

After reading the files in the gathering step, we assessed our data through two ways:

1- Visually by reading our data in Jupyter Notebook and through Excel

2- Programmatically using pandas’ functions (e.g: info (), describe (), unique () …etc.)

Then divided the issues we faced for each data frame into two categories:

- Quality Issues
- Tidiness Issues

B.1-) Archived enhanced data:

B-1.1-) Quality Issues

- Some dog names were extracted incorrectly from text (e.g: None, a, an, such, quite …. etc.)
- Missing values in doggo, pupper, floofer and puppo were entered as “none”.
- Timestamp column data type should be converted to datetime.
- Source column needs a regular expression to extract the real sources (e.g: Twitter for iPhone, Twitter Web Client, Vine - Make a Scene, TweetDeck)
- Tweet links need to be extracted from text using regular expressions and saved in another column.
- Removing rows that includes retweets, we only care about WeRateDogs Tweets.
- Snoop Dogg name was missing and also had a duplicate.
- Rating Numerator needed to be extracted again using regex and also some ratings were the multiplication of the rating by the number of dogs.
- Rating Denominators was bigger than 10 because it was multiplied by the number of dogs.
- After doggo, floofer, pupper and puppo joined into one column they should be converted into categorical data type.
- Rating Numerator and Denominator columns should be converted to float datatype for decimals

B-1.2-) Tidiness Issues

- Doggo, floofer, pupper and puppo should be joined into one column called dog_stage
- Retweet and Reply related columns should be dropped.

B-2-) Image Predictions Data:

B-2.1-) Quality Issues

- Dog predictions and confidence columns should be converted into categorical data after they are combined into one column called “confidence_per“ and “breed” respectively.

B-2.2-) Tidiness Issues

- The p1, p2 and p3 columns should all be joined in one column according to the prediction truth and confidence percentages.
- tweet_id , jpg_url , img_num , dog_stage and confidence_per should be the only columns we need.

B-3-) Json Twitter API Data:

B-3.1-) Quality Issues

B-2.2-) Tidiness Issues

- We only need the retweet_count, favorite_count columns.
- Join the resulted data frame with the other two data frames to form the master data frame (twitter_archive_master.csv)

🧹 C-) Cleaning

After assessing our missing, in-valid, in-accurate and in consistent data set we used the Define, Code and Test method to clean and organize the data in each point the assessing phase until we finish them all.

All the cleaning done was done programmatically.

💾 D-) Storing

We stored the data as a csv file called twitter_archive_master.csv