We Rate Dogs Project

Data Wrangling process consists of essentially: gathering, assessing and cleaning data.

Gathering:

In this project, I gathered data from three different sources.

1. Downloaded twitter archive enhanced.csv file manually.
2. Downloaded image predictions.tsv file that is hosted on Udacity server programmatically.
3. Queried data from twitter API in json file format.

After gathering the data from these sources, I read them into three data frames using pandas read_csv() function.

Output:

1. archive_df
2. image_predictions_df
3. api_df

Assessing:

I assessed the output files from gathering process visually using Microsoft Excel and programmatically using pandas functions as info(), head(), tail(),duplicated().sum(). I found some quality and tidiness issues that should be cleaned for analysis.

Tidiness Issues:

- (doggo - floofer - pupper - puppo) columns should be represented in one column as values.

- api_df should be combained with archive_df

Quality issues:

- **archive_df** timestamp datatype is str instead of datetime.

- **archive_df** in_reply_to_status_id & retweeted_status_id non null have no ratings and should be dropped.

- **archive_df** missing values in expanded_urls.

- **archive_df** has tweets without images.

- **image_predictions_df** has retweets and replies.

- **image_predictions_df** 66 Duplicated URLs

- **archive_df** Null values are called None in (name - doggo - floofer - pupper - puppo).

- **archive_df** source contains < a > tag instead of its contents.

- **archive_df** missing values at (name - doggo - floofer - pupper - puppo) columns.

- **archive_df** inaccurate names in name column.

- **archive_df** inaccurate ratings in (rating_numerator,rating_denominator) column.

- **image_predictions_df** undescriptive column headers.

- Different number of entries in archive_df , api_df

Cleaning:

It is better to follow three steps in cleaning data: Define, Code and test. At Define step I defined specifically what issues I was going to clean and how it will do the cleaning process. In Code I converted those define points into codes, then I checked if the issued was cleaned at test step.

Define :

- archive_clean: Drop retweets , replies and Empty Urls rows.
- archive_clean: Drop Tweets without images (Not in image predictions dataframe)
- Image_predictios_clean: Drop Retweets and Replies that doesn't have ratings (Check if they are in archive_clean)
- archive_clean: extract content of a tag in source column using .str.extract() function with REGEX
- archive_clean: Convert datatype of timestamp using to_datetime() method
- archive_clean: Convert 'None' into "" using replace()
- archive_clean: make new column called dog_stage from adding stages in (doggo-puppo-pupper-floofer)columns
- Drop Unwanted Columns
- Combine archive_clean and api data into one df

Output :

Two clean data frames image_predictions_clean and archive_clean.

Storing files:

After getting the data cleaned. I stored the cleaned data into csv files them read them again to start analysis.

Insights and Visuals:

Now, We can analyze our clean data to extract some insights and visualizations.

First I asked some questions about the data sets, then I started the analysis to get the answers and end up with some insights.

Output:

- The most tweet people retweeted was for a doggo.
- The most tweet people liked was for a puppo.
- The most three frequent dog stages are: pupper, doggo and puppo.
- December is the month at which people tweeted the most in We Rate Dogs.