Analyzing and Visualizing WeRateDogs

Udacity: Wrangle and Analyze Project

By: Dwiti Shah

Introduction

As a Udacity Data Analyst Nanodegree candidate I need to wrangle and analyze the WeRatedog data. It is a twitter account that rates people's dog. It has over 5000+ records, however after filtering Udacity has provided us with 2356 records from Nov 2015 to Aug 2017.

Gather Data

For this project, 3 data sources were used:

- Enhanced Twitter Archive file was provided by Udacity and I have manually downloaded it. It provides various attributes like timestamp, name, rating, text and more.
- Image prediction file is hosted on Udacity's servers and was downloaded programmatically. This file predicts a dog breed using neural network.
- Additional Data, query Twitter API using tweet id to gather retweet and favorite count.

All the 3 data sources are helpful since, each provide different features about the tweet which will lead to more in depth analysis.

Accessing Data

Two types of assessment are used:

- Visual assessment: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes.
- Programmatic assessment: pandas' functions and/or methods are used to assess the data.

I have performed data assessment using the following commands:

- .info()
- .values_count()
- .head()

After data assessment, I have identified the following quality and tidiness issues.

Quality Issues

- Twitter DataFrame
    - As we can see 181 values in the retweet column, therefore need to remove from the data
    - Incorrect datatype as retweeted_status_timestamp and timestamp are defined as object
    - Tweet_id should be string and integer
    - Missing data in expanded_url column
    - Incorrect dog names like: a, an
    - Nulls represnted as None in doggo', 'floofer', 'pupper','puppo'
    - Remove unnecessary tags in source and we can easily identify the source
    - On assessing the twitter_enhanced.csv file you will see on row 46 that the correct rating in the tweet is 13.5 but it's extracted as 5. Hence, changing the data type as float. Extracting decimal value from text column

- Image Prediction DataFrame
    - First letter for few names in columns P1,P2 and P3 are capital
    - Combining P1, P2, P3 values into columns called as Breed and Confidence
    - Less tweet id count compared tweet id count in Twitter dataframe
    -
- Twitter1 DataFrame (Json text file)
    - Less tweet id count compared tweet id count in Twitter dataframe

Tidiness Issues

- Drop columns are not needed for analysis
- doggo, floofer, pupper and puppo columns in twitter table can be merged into one column “Stage”
- join all tables together since they are talking about the same tweet

Cleaning Data

After assessment, next step is to clean and test the dataset. First, I have created a copy of all the 3 files and then programmatically cleaned dividing them into 3 stages: Define, Code and Test.

I have applied the following commands and techniques

- .drop()
- .info()
- .value_counts()
- . isnull()
- .to_datetime()
- . notnull()
- .astype()
- .apply()
- .str.match()
- .replace()
- str.lower()
- .merge()
- Loops

Store Data

After completing the cleaning process, I have stored the final file into Csv format.

Conclusion

Today each company generates millions of data points and none of them are in correct format. Through this project I understood how important is to first assess and clean the data, before performing any analysis on it.