Udacity – Wrangling Act Documentation

Introduction

Real-world data does not come so easily where you have the data neatly needed for your analysis purposes we as data analysts are required to identify the sources that consist of the data required to answer the questions we are looking for. In this project we use the data fetched from a Twitter channel called @WeRateDogs, the sources we get the data from coming from 3 main sources:

Twitter API – It is the data fetched directly from Twitter, where we get the data in a JSON format that consists of all the required details needed from the tweet.

Prediction Output tsv File – It is the data we received internally where it consists of the data of the tweet pictures passed through a model which predicts the Objects present in the model. All pictures have their unique tweeter ID where maps the tweet picture to its source

Tweeter Archive CSV File – It is our base source where we use the IDs mentioned in the dataset to limit our data for the process of analyzing and wrangling.

Wrangling

The below is the list of identified quality and tidiness issues from the data collected from the above-mentioned sources.

Quality

twitter_archive table

- **name** column consists of unusual dog names, usage of 'a' and 'None' needs to be replaced with a default value
- **in_reply_to_status_id**, **in_reply_to_user_id**, **retweeted_status_id**, **retweeted_status_user_id** a nd **retweeted_status_timestamp** are having high number of null values, hence ignoring these columns as it will have small effect on our analysis.
- **source** column needs to be re-defined into a categorical column distinguishing the actual source in simple terms instead of the usage of HTML Tags.
- lower case all names in the **name** column
- To remove all rows which has null values for the **extended_url** value, as these all look to be replies or retweets to the original tweets.
- To convert the object to timestamp datatype for the **timestamp** column
- Convert **tweet_id** to string from int

image_prediction table

- make the **p1, p2, p3** column values in lowercase to maintain consistency

archive_json table

Tidiness

- **doggo, floofer, pupper, puppo** - needs to be converted to a categorical column of whether the tweet includes these dog terms
- **rating_numerator** and **rating_denominator** have varied rating metric systems, it is best to evaluate this in terms of categorical types by having a standard of all dogs rated as a base of 10 and for anything above 10 will be termed as a good rating and anything below 10 is an average/bad rating
- Join all 3 data frames into a master dataframe
- Separate the Tweet content with the link URL in the **text** column

Tackling the Identified Issues:

In short to tackle some of the common issues listed in the list are resolved by the below actions performed.

1. **Null Values –** Dropping the rows/columns, depending on the effect it can have on our analysis.
2. **Incorrect Names/Consistency issues –** Modifying the data giving a default name for the supposed incorrect names and regarding consistency issues with regards to lowercasing the records or modifying the column names so it is easier to call and analyze during the analysis phase
3. **DataType Mismatch –** Casting the columns to match their respective data type.
4. **To Tidy the Data –** This process or step varies depending on how the analyst requires its data to be in a format where it can effectively use for analysis.
    a. **Converting it into Categorical columns –** This makes it easier to identify and answer questions regarding the field
    b. **Joining tables into 1 Master Table –** By doing so we are only taking in the columns which are useful and also identifying if any records do not have the records present for the particular ID.
    c. **Text level wrangling process –** when trying to get information such as common wordings and spaces used we need to make sure the column content does not contain any consistencies such as the inclusion of a short link to the tweet content at the end of the text content. Removal of such consistent text/link makes the column/field more valuable and helps in the process of analysis.
    d. **Rating System –** In the mentioned dataset the rating system is not well defined hence a logical operation or system must be introduced where it will be made easier for analysis. Where if the numerator value is greater than the denominator value then it would be classified as a "*Good Rating*" else it will be classified as *"Bad Rating"*.

Doing the above activity helps in making our data ready for analysis purposes and also ready to tackle a majority of the questions with more efficiency.