WeRateDogs Project Report

Introduction:

The below report which made for the Udacity Data Analyst Nanodegree Program of project "WeRateDogs", I'll try top explain the process in which my report has gone through. The goal of this project is to practice the process of wrangling and cleaning data, which was made through this twitter account tweet data. Tweets went through a process in which I performed the following activites:

- Gathering Data
- Assessing Data
- Cleaning Data

- Gathering Data

I this process data is being obtained from csv files and loaded to tables in which it will go through the wrangling process.

- Twitter archive data was loaded to `twitter_archive` table whcih contains WeRateDogs Twitter archive, which was provided by the Course and data was imported into the dataframe.

- Image prediction data was imported from Image prediction file provided by the course and hosted in Udacity's servers and added data to `predictions` table. The tweet image predictions, basically predicts whether the object in a said image is a dog or other object.

- API data was provided through a file in the course material as my twitter developer account wasn't created when I've started to work on the project, in this file I was able to query API data in JOSN file to read twitter data to api_df_now table.

- Assessing Data

In this step data is being assessed visually and programmatically to detect quality and tidiness issues in the gathered data.

- 'twitter_archive' has missing data in multiple tables example, "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id". Lower case dog names was an issue too.

- Another issue is the dog names that can make confusion like doggo, pupper, floofer and puppo.

- Timestamp is another issue that needs an attention. Source of content needs to be organized.

- Rating values needed some changes.

- image predictions columns was making a confsuion.

- ‘api_df_now’ file is separate from Twitter archive data.

Cleaning Data

In this step data is being cleaned and added to new tables twitter_archive_clean, prediction_clean and api_df_now_clean according to the issues observed in after assessing the data.

1- Fixing Quality issues

1- Dropped unnecessary columns containing missing data “in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id".

2- Replaced missing “None” values with “NaN”.

3- Joined “api_df_now” table with “twitter_archive” table and renaming ‘tweet_id’ column.

4- Combined all dog names; doggo, pupper, floofer and puppo under one column name ‘dog’

5- Changed timestamp to datetime.

6- Optimized source of content: Twitter for iphone, Vine - Make a Scene, Twitter Web Client and TweetDeck.

7- Made a default value for numerator and denominator values.

8- Capitalized first letters of dogs names.

2- Tidiness

1- Changed Image predictions p1, p2 and p3 names to potential_dog1, potential_dog2 and potential_dog3.

2- Merged the cleaned data into the clean tables.

Conclusion

Through this project I've learned to express the data analysis process through code and different tools offered through the Jupyter lab application.

Data wrangling is crucial in the data analysis process as it's the only way top obtain a reliable data to take the proper decisions in any organization. And using python in the process made it much easier and more efficient, also, the different libraries used in the process allowed the data to be read and manipulated in a relatively easier way, which will facilitates the process if dealt with much larger data amounts like Big Data.

This proves that using code is a way to manipulate data and alter it efficiently. I believe that through my learning process I'll be able to dig deeper into more processes and tools which will make the process more fruitful and efficient.