Data Analyst Nanodegree Program

Wrangle and Analyze Data Project

(WeRateDogs Twitter account)

Name: Abdulrahman Alghaligah

September 2021

Introduction:

In this report, I will explain all the steps that I went through wrangling the dataset from "WeRateDogs" Twitter account.

Data Gathering:

The first step that data analysts usually face is to gather the data, in this project I gathered the data from three different resources.

1. Twitter archive from "WeRateDogs" Twitter account, the account asked for their archive, and they sent it to Udacity, it has more than 5000 tweets. I downloaded the file in CSV format then I read by pandas read function.
2. Tweets image predictions, this file has three predictions from neural network algorithms, this file hosted in Udacity's servers, and I downloaded programmatically using request library, then I wrote in the local machine as TSV file, finally, I read it from pandas read function.
3. The third dataset I gathered is from Twitter API through tweepy library, this data contains helpful information about the retweet and favourite data in each tweet. I wrote the data in a JSON file after that I read it using the pandas read function.

Data Assessing:

After gathering the needed data, it is time to play the role of detective and assess the data, this step is important to look at the data and make notes of the issues. There are two types of issues quality and tidiness issues, quality issues are for the problems in the content of data, where tidiness issues focus on the structural issues. I assessed the data visually by scrolling in jupyter notebook and Excel, then I found most of the issues in the programming way. I found 9 quality issues and 3 tidiness issues I will provide them blow:

Quality issues:

1. Incorrect missing values recorded as 'a', 'the', 'an' and 'None'.
2. Removing retweeted tweets from the dataset.
3. Removing unneeded columns.
4. In the source column it is better to get rid of the URL.
5. Changing the type of source column to category datatype.
6. Changing the type of timestamp column to datatime type.
7. Incorrect ratings need to be changed.
8. Changing the type of img_num column to category datatype.
9. Describe the columns in image prediction dataset to more descriptive names.

Tidiness issues:

1. Convert each breed of the dogs to one column.
2. Separate the date and the time to two columns.
3. Create a tidy master dataset that combine all the datasets.

Data Cleaning:

The final step for wrangling the data is to clean the data, in this step, I cleaned all the issues that I found in the assessing step, I started cleaning by finding the missing values to prevent any problem during the cleaning because maybe some cleaning issues will affect the others. Also, it's important to say that I made too many iterations in the assessment and the cleaning process, sometimes during the cleaning process I found more issues then I iterate by coming back to the assessing step and making notes. I cleaned each issue by defining the problem and how I'm going to fix it then I wrote the code after that I test the issue to see if it was fixed or not.