Wrangle Report

Udacity

By: Sarah Alamri

1-1-2021

Wrangle Report:

In this project, I have practiced what I learned in the data wrangling section and apply all data wrangling processes on real data set.

The dataset that I wrangling is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Data wrangling processes:

- Gathering data
- Assessing data
- Cleaning data

Gathering data:

In the gathering data step, I gathered data from different sources of data by different methods:

1- **(** twitter_archive_enhanced.csv) this data set provided from Udacity, I was download and read it.

2- (image_predictions.tsv) I have downloaded this dataset programmatically using the Requests library.

3- (tweet_json.txt) I was download and read this dataset.

Assessing data:

In Assessing data step , I used the two types of assessment **Programmatic** assessment and **Visual** assessment for Detect **quality** and **tidiness issues**

Quality issues

- (tweet_id) Change it string
- Delete (retweets)
- Missing values in (in_reply_to_status_id ,in_reply_to_user_id ,retweeted_status_id , retweeted_status_user_id,retweeted_status_timestamp)

- The data in columns (p1, p2, and p3) had uppercase
- Wrong datatype for(timestamp)change it to datetime
- Wrong names of dogs
- Standardize ratings
- Non-descriptive names of columns.

Tidiness issues

- Create one column for dog stags (doggo, floofer, pupper, puppo)
- All table must be in same data set

Cleaning data

In Cleaning data step, I Cleaned all the issues (quality and Tidiness) detected while assessing. First, make copy of all dataset and then I cleaned the issues using Defined-Code-Test framework.

- Incorrected datatype on tweet_id it int, and I have converted it to string

- remove retweets because we don't need it and it is not original rate

- Missing values in (in_reply_to_status_id ,in_reply_to_user_id ,retweeted_status_id , retweeted_status_user_id,retweeted_status_timestamp), I removed this columns

- Create one column for dog stages(doggo, floofer, pupper, puppo)

- The data in columns (p1, p2, and p3) had uppercase, that must be standardized into lowercase letters

- Wrong datatype for(timestamp)change it to datetime

- incorrect some of dogs name like (''one'', "a", "an", "by","very"), replaced with no name

- Standardize ratings to float data type

- change the Non-descriptive names of columns to descriptive names

- All table must be in same data set by using MERGE