DAND Wrangle And Analyse Data Project
Wrangling Report

By Hamad Sami Al-Assafi

Table of Contents

Introduction:

In this project I will take three steps to wrangle dataset from weRateDog account in twitter the three steps are:

1-Gathering Data: In this step I will gather data from multiple resources that i must have to complete this project

2-Assessing Data: In this step I will look through the data that i have gathered and see if there is any quality or tidiness issues in the dataset that i gathered if there are any issues, I will document it

3-Cleaning Data: The last step in wrangling process, in this step i will transfer the documentation that I have in the assessing step to code to clean our dataset, in this step I will use (Define, Code, Test) methodology

Then our dataset will become somehow clean! afterward I will go through the analysis process

Gathering Data:

In this step I have gathered datasets from three different sources:

1-From (CSV) file that contains archive tweets for weRateDog account, which was given from Udacity.

2-From (TSV) file that contains detailed data about the images in each tweet, I have extracted the file from the internet using Python and (Request) library, then I have installed the file in my local machine.

3-From (JSON) file that contains detailed data about the count of the favourite and retweets, I have extracted the file using Twitter API (tweepy), then I have installed the file in my local machine

Assessing Data:

In this step, I have through the datasets to discover any quality or tidiness issues, either visually or programmatically. And I have discovered many quality or tidiness issues:

Quality Issues:

1-Missing values in (expanded_url) column

2-Incorrect dogs name

3-dupliacted tweets

4-Unwanted HTML tag should be removed in (source) column

5-Data types problem (timestamp) should be converted to datatime, (img_num) and (source) should be converted to category data type

6-Fill the missing value with null values instead of (None) or other words

Tidiness Issues:

1-Convert the nominator rating and denominator rating to actual rating, and delete them from our dataframe

2-Create single column that express the type of the dog instead of four columns for each type

3-Merge the dataframes together, and create master dataframe

Cleaning Data:

In the cleaning process I have cleaned the missing values firstly, because maybe some of the other issues will disappear when I deal with the missing values, then I have cleaned the other quality issues that I have mentioned them above. In this process I have used the (Define, Code, Test) methodology for each issues to make the cleaning process more organized.

In the other report I will go through the analysis process. Thanks !