Nawaf Alhajri-Wrangle Report

Wrangle and analyze data project

Project details:

We have the following tasks in this project:

•Gathering data

•Assessing data

•Cleaning data

Gathering data:

In this project we can obtain the data from three datasets, so we have different dataset as following:

•Twitter archive file: the twitter archive enhanced.csv was provided by Udacity that can be downloaded.

•The tweet image predictions: what kind of the image which will be presented in each tweet according to a neural network. This file (image predictions.tsv) is hosted by Udacity's servers and can be downloaded programmatically using the Requests library and URL information.

.•Twitter API: Read (tweet-json.txt) file line by line to create date frame.

Assessing data:

After looking to the collecting data and applying programming methods such as:

1. sample ()
2. info()
3. Value_counts()
4. isnull().sum()
5. List()

Quality issues have been discovered:

I have founded the following issues on quality and tidiness:

- Wrong Datatype img_num Column should be in string

- Change tweet_id from an integer to a string .
- Timestamp is not of datetime format .

remove columns with too many missing values.

- retweeted_status_user_id
- retweeted_status_id
- retweeted_status_timestamp
- in_reply_to_user_id
- in_reply_to_status_id

Tidness issues I have discovered:

- doggo, floofer, pupper, puppo these 4 variables shoule be combined into one categorical variable Dog Type.
- merge the dataframe twitter_archive, dataframe image_predictions, and tweet_json dataframes .

Cleaning data:

Cleaning data steps:

1- define 2- code 3- Test

First, we should copy the data frame before cleaning and then start cleaning process, so In the "define" type: the issue has been defined, then I identified the dataset that the issues come from.

In the next step: I have put some special "code" (astype (),drop(),to_datetime ()), which will help us in the issues cleaning.

In the last step: the cleaning ability of the code has been tested, to check the output whether if it is corrected or not.