UDACITY WE RATE DOGS DATA WARNGLING PROJECT

FIRST STEP: DATA GATHERING

Gathered, data from different sources both programmatically with request library and through web browser, these datasets include:

include:

a.Twitter_archive.csv:- The data was downloaded and passed into a data frame.

b.image_prediction.tsv:- The data was downloaded programmatically using the request library and loaded into a data frame.

c.tweets.json:- The txt file format was loaded with an open function into columns and loaded into a dataframe for analysis.

All gathered data was arranged into a pandas data frame

SECOND STEP: ASSESSING DATA

This was carried out on various datasets by making use of the functions available for assessing data: This function includes:

The info ():-gets the information of each column

Describe function: -Gets the numerical columns for statistical analysis

Sample function: -Used in analysing the top rows in the dataset.

Duplicate function: - I looked for any duplicated rows to make sure there were no duplicated values in the dataset.

THIRD STEP: DATA WRANGLING

ISSUES:

1.Twitter_archive file contains name column contains none values.

Solution

This was resolved by replacing all the value with null. values.

2.Some names in twitter archive name column aren’t actual

names of individuals.

Solution

This was resolved by replacing all the non-name values.

3.Twitter_archivecopy column has a different filetype i.e. it is

seen as an object type.

Solution

Columns with timestamp were changed to their appropriate datetime format.

4.No column for month or year for easy analysis.

Solution

new columns for year and month were created from the timestamp column for easy analysis.

5.Too many redundant columns that can be combined.

Solution

Rredundant columns for duggo,puppo,floofer,pupper were merged into one dog stage column.

6.Not all values in the image_predictions dataset has .jpg

extension.

Solution

Not all values in jpg column had the .jpg extension and it was fixed.

7.Delete unusual column.

Solution

Columns that won't be used for analysis were deleted.

8.Column type were changed to make merging easier.

Solution

Tweet_id column type was changed to string to allow for easy merging of all the tables

9.Different tables merged into one master dataframe.

Solution

The three tables were merged into one to form a master dataframe.