wrangle_report

June 19, 2022

0.1 Reporting: wrangle_report

0.2 stated date for the report: 10th of June 2022.

This wrangle_report as it name entails is to document the wrangling procedure which was stored as twitter_achive_master.csv.

Definition of Data Wrangling: it is a process that includes three aspect, which are Gathering, Accessing and Cleaning data.

Details of what is done on the three pocess of data wrangling

1. GATHERING DATA: three method was used in gathering the data;

a. The first dataset was read as CSV, that is 'twitter_enhanced.csv', using pd.read_csv.

b. The second dataset is image_predictions.tsv, it was firstly programmatically extracted using 'requests' and 'os'and read as CSV.

c. The 'tweet_json.txt' was gathered without twitter acccout,it was gathereed through the suppoting materials provide in udacity workspace.

2. ASSESSING DATA: the three data was assessed visually and programmatically, which was to detect the quality and tidiness issue.

Quality issues

a. The archive_enhanced_df has 181 retweeted value that are non-null.

b. columns that are not needed, like retweeted column, etc.

c. The timestamp column is a string instead of datetime.

d. Text for canela is duplicated.

e. Dog name like O'Malley was 'O', Quizno was his.

f. The column for 'name' have some invalid names such as; a, none, etc.

g. There are some typographical error in the names of the dog.

h. 'rating_numerator' less than 10, which is wrongly extracted.

Tidiness issues a. the three datasets which are archive_enhanced, image_predictions and tweet_df have to be one dataset

b. archive_enhanced_df have four columns which are, doggo, floofer, pupper and puppo, the four columns is all about the stages of dog.

3. CLEANING DATA: this part is done programmatically to make amendment to quality and tidiness issue. The following are steps taken to successfully clean the data.

a. The non-null for retweeted value was drop

b.Timetamp was change from string to datetime

c. Retweeted value was remove as it was not needed

d. The duplicated canela name of dog was dropped

e. Correction was made on incomplete name typographical error made

f. The rating_denominator was dropped and the rating_numerator was renamed

g. Four stages of dog was join together to a single column

h. The three was concatinated into one dataframe.