Data Wrangling Report

By : Mohamed mohamed abdraboh abd elsayed kansouh

November/2020

This is an assignment is written for Udacity Nanodegree in Data Analysis Professional

Project 2 'we rate dogs'

Its about the steps I've made to wrangle the data which have been collected from we rate dogs account on twitter .

First of all what is wrangling data its separated into 3 sections :

Gathering

1. **Data from archive data frame which have been saved in .csv format.**
2. **Downloading programmatically image_prediction data frame from udacity server.**
3. **Collect data from Twitter API (which unfortunately I couldn't get them as I didn't have dev access to twitter api)**

Assessing

- **Which is separated into 2 sections**
  - ### Visually

    Which I have done using excel and I found bunch of issues

    1. **source column has html tags which i think they aren't important and device column should show me the device and application used if it web , vine or phone app (html tags meant to download the app for each device type) .**
    2. **timestamp has 00:00 which is duplicated and not necessary should be deleted .**
    3. **denominator and numerators has outliers (this one need to be checked with programmatic assessing to make sure) .**
    4. **There's none values as string not NAN so validation issue .**

  - ### Programmatically

    1. **Columns that is not needed for replies and retweets and the data inside them.**
    2. **Tweets doesn't have image.**
    3. **Variables that should be values in columns not the name of it.**
    4. **types that is not supposed to be**
    5. **strange names in name column for dogs .**

6. **un useful rating for analysis some are too big in dominator and bunch of outliers .**
7. **When merging I found out that there's un useful columns .**
8. **Html tags in one column (source) .**
9. **Null values in column (expanded_urls).**
10. **Restructure the data after cleaning to make it easier in analysis**

Cleaning

For cleaning I made it 20 stages to make the data fully cleaned

Starting with retweets and replies I deleted their tweets then dropped their columns , after that I changed the None values in (doggo) (floofer) (pupper) (puppo) into ('') empty string .

Melting these 4 columns into one called dog_types then drop these 4 columns.

Removing null values from expanded_urls.

Change the type of most of columns timestamp to date time , source to category

Take valid values from (p,p_conf,p_dog) columns to make 2 columns breed and machine probability of choosing this breed (confidence)

While assessing I found that the strange names for dogs was taken randomly from text and most of them are lower case so I used regex to specify them then change them to empty string as well as when I tried to delete 00:00 UTC from date column .

Merging all dfs into big one .

change outliers in dominator and numerator and get the valid data from text or calculate it if possible .