WeRateDogs

Wrangle Report

---

* Introduction

Wrangling data is very important for every data analysis projects because it helps you to be sure that your data is clean.

Our purpose of this project (WeRateDogs) is making all data clean and suitable for analysis.

* Project details

- Gathering Data
- Assessing Data
- Cleaning Data

For wrangling data I followed the following steps:

1- Gathering Data
In this project it was three sources of data to gather from.
   - `twitter_archive_enhanced.csv` From Udacity classroom , I downloaded manually this file and upload it into jupyter notebook .
   - `image_predictions.tsv` this file is hosted in Udacity server and was downloaded programmatically using Request library and URL
   - `api_df` it's alternative file to (wrangle_act) because I don't have twitter API developer account yet but I will .
     this file was downloaded manually and starting assessing.

2- Assessing data
There are two ways for assessing data –Quality and tidiness-
   - Quality
     Completeness, validity, accuracy, consistency (content issues)
     - twitter_archive
       - ✓ Keep original ratings (no retweets) that have images
       - ✓ Delete columns that won't be used for analysis
       - ✓ Erroneous datatypes (doggo, floofer, pupper and puppo columns)
       - ✓ Correct numerators with decimals
       - ✓ Correc denominators other than 10
     - image_prediction
       - ✓ Drop 66 jpg_url duplicated

- ✓ Create 1 column for image prediction and 1 column for confidence level
- ✓ Delete columns that won't be used for analysis

- Tidiness
  - (twitter_archive)
    Separate timestamp into day - month - year (3 columns)
  - use tweet_id as type int24 to merge all tables in one data set

3- cleaning data
This part was divided in three parts define, code and test
Each part took its time and effort but it were very challenging.
Firstly I made a copy from all data sets to make data sources in safe
After that I started a walkthrough in my assessing notes to solve all quality and tidiness issues. It was very enjoyable.
One very challenging cleaning was when I had to correct some numerator there were decimals and wrong values. in addition to, denominators were have values upper than 10 . I solved these problems by programming and Excel too.