Wrangle Report

Udacity Nanodegree: Wrangle and Analyze Data Project

By: Mahmoud Medhat

Introduction

This is Wrangle and Analyze Data for Udacity Nanodegree. In this project we practice wrangling data from several sources, like tweets from the Twitter user @dog_rates, also known as WeRateDogs. Below I will illustrate how exactly I have done my project

Project Details

Data Wrangling steps:

- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data

Data with gathered from 3 different sources making 3 different data frames

1. The Twitter archive file was provided by Udacity and downloaded manually, this file included various variables for each tweet like: tweet id, timestamp, name, etc.
2. I used Twitter API to gather more data including favorite counts and retweet counts
3. Using the requests library, I was able to download the tweet image prediction file programmatically

Assessing Data

After the data was gathered, I performed data assessing manually and programmatically using excel sheets and some methods as:

- .head()
- .sample()
- .info()
- .value_counts ()

Quality Issues found

1. tweet id is integer
2. remove retweets
3. timestamp is object
4. some names are inaccurate
5. none instead of NaN
6. unnecessary columns
7. Rename some columns
8. missing urls
9. rating column for data analysis
10. tweet Id is integer in both APIs df and images df
11. dataframe doesn't contain breed column and many columns for same thing and columns can be dropped

Tidiness Issues found

1. The 3 datasets have the same observational unit so they should be merged in one dataset
2. 4 columns (doggo, floofer, pupper, and puppo) in archive dataframe stands for the same variable which is ‘Dog Stage’

Cleaning Data

All the Issues found have been cleaned and tested using several methods such as:

- .merge()
- .reduce()
- .drop()
- .isnan
- .astype()
- .to_datetime()
- .islower()
- .replace()
- .rename
- .value_counts()
- .info()
- .head()
- Loops
- Regular expressions

Conclusion

As a detective I started to discover the data and understand it using several methods and libraries. Time by time I felt like I’m more familiar with the data I’m pretty sure that there is still more and more dirts in my dataset to clean. It’s so rare to find data already clean and tidy