Wrangle Report

Introduction

This Wrangling data project is from Udacity academy to teach about the techniques of wrangling data which includes gathering, assessing, and cleaning data for a Twitter account called WeRateDogs where this account is famous for rating dogs with a denominator of 10. We have been asked to wrangle the data of this account by various sources.

Gathering Data

The data was gathered from 3 sources, as follows:

- **The twitter archive** file which was given by Udacity and downloaded by me, and it contains detailed data about 2000+ tweet from there WeRateDogs account.
- **Additional Data via the Twitter API** this additional data contains favorite and retweets count.
- **Image Predictions File** which was given by an URL link then I requested it in Jupyter notebook.

Assessing Data

After gathering all data frames, I started to use the following tools to assess the data:

- .info()
- .isnull()
- .sum()
- .value_counts()

and after assessing the data I separated the issues that were encountered in the data frames in two: tidiness issues and quality issues.

Where tidiness issues are about the order and arranging of the data frame.

And quality issues are about wrong and unnecessary data.

Cleaning Data

This process has three steps for each cleaning process: Define the process, The code, and Test.

Before I started cleaning, I copied every data frame to new data frames to clean it separately, And for the cleaning process, I used basic pandas tools such as:

- .replace()
- .drop()
- .merge()
- to_datetime()
- .info()
- .head()
- .count()
- .sum()
- etc.

Conclusion

In conclusion, wrangling data is a very important step in analyzing data it allows the researcher to see the data more thoroughly and it can be analyzed in a much faster time than a messy data frame, it also makes the data clear in a way a person could raise questions about the data frame to allow him to do a better analysis.