Data Wrangling (DAND) Report

Prepared By: Faris AlShammari

Table of Contents

Introduction.

First of all, in this project I am going to wrangling some datasets which I gather it from different source, and the datasets are about twitter account that interested in Dogs. And it includes pictures, rating, and dog's breed.

Second, the datasets contain a lot of information such number of people who rewets and likes a tweet, the date published of a tweet, the tweet's URL, number of images included the tweet, the source of tweet and also has an algorithm which predict the dog' breed.

Third, the datasets are dirty, so I am going to wrangle it to make it clearly by Three steps which is Gathering, Assessing and Clean.

Wrangling

Gathering

In the first step, I collect three datasets which called (tweet_json, image-Prediction and Twitter_Archive_Enhanced) each of them has unique information so I need them all.

Assessing

In the second step, I found too many issues in the three datasets either Quality or Tidiness issues.

Quality Issues

Twitter Archive:

- Null URL.
- Incorrect dog's names.

- Duplicated tweet ID.
- invalid URL which contains the HTML tags.
- Incorrect rate.
- Converting timestamp datatype into datetime.

Image Prediction:

- Duplicates Tweet ID
- Unclear prediction percentage.

Tidiness Issues

Twitter Archive:

- Combined Four columns together into one column [doggo, floofer, pupper, puppo] into [Category].
- Merging Rating numerator and Rating denominator into one column called [Rating].

- Merging all the three dataframes into one dataframe called (Twitter Archive Master).

Cleaning

In the last step I just took the quality and tidiness issue and fixed it by using jupytr Notebook tool using pandas and python libraries. First I fixed the twitter archive dataframe quality and tidiness issues then the Image prediction. And I see that Twitter API dataframe does not need a cleaning. At the end of cleaning I merged all dataframes into one dataframe saved as CSV.

Conclusion

In conclusion this report was for Data Wrangling only and how I wrangle the data by using specific tools and programming language.