5th Sept 2022

DATA WRANGLING REPORT

WeRateDogs TWITTER ARCHIVE USING API

BY: IKENGWA MAXIMUS CHINAZO

INTRODUCTION

This is project on data Wrangling , which involves the transformation of raw data to a very clean data for further analysis. Originally most data are gotten in their original state, and in that state , they are dirty, messy and untidy, hence the need for data wrangling. There are three stages in the data wrangling process, this includes the following

1. **Gathering Data**: This involves gathering data from the source. In this project , I used the **WeRateDogs** tweets from twitter, but the data were gathered in three different ways :
    i. **Twitter Archive data**: This I downloaded straight from Udacity repository. It was given to them by the owners of the WeRateDogs twitter handle.
    ii. **Image Prediction File**: This I downloaded pragmatically using the Request library from url(https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv) to twitter web page
    iii. **Twitter API Data**: this I got through the use of twitter Api. In this project I got the Json file by querying the twitter Api using the tweepy library to obtain more information as regards to the tweets' id in the Archive file such as the retweet counts and favorite counts.

2. **Assessing Data**: Here we look at the data to check and remove unwanted parameters, hence assessing them for Quality and Tidiness issues; this could be inconsistencies, completeness or validity issues. we watch out for them and

document them so as to aid our cleaning process. There are two as to assess data; either visually looking at the data or using built in functions and codes(programmatically)

3. **Data Cleaning**; This is the final stage of the Data wrangling process, it involves removal of unwanted data, it follows three definite stages which includes Defining the issue and what to do about it, secondly, writing codes for the cleaning process and finally testing the code to make sure it works. This can be done by defining some functions, such as **try, exceptions, for loop** etc Merging data together using some pandas built-in functions such as the **melt, merge, drop**, etc, which are some of the functions I used in this project.

CONCLUSIONS

Data Wrangling process helps in delivering a very clean and tidy data for further analysis and visualization.

After cleaning the data I save my cleaned data to a CSV file name twitter_master_archive data. It can be shared to anybody that wants to explore the data without any further cleaning