DAND: Data Wrangle and Analyze

Wrangle Report

Introduction

Data Analysis nanodegree illustrates the concept of Data analysis and how to move over all processes to get clean and understandable data. For over years, data has become more valuable to solve problems in real life and male accurate decisions. Data analysis programs show the best to take the data from early stage and agater to clean and get some available information. The process start with data wrangling which is when you gather the data from several resources and assess then clean. This report will show how we performed the data gathering over Udacity project 5 or what we call it (we rate the dog project).

Gathering the Data

The Data for the project needs to be collected from three resources, the first resource which is twitter archived that has been provided by Udacity. I have imported the pandas library to read the data from a csv file then saved in the dataframe. The second resource which is image prediction that has been provided to be collected programmatically and saved in directory then read it and saved in dataframe. The third resource is twitter API, I have contact twitter support and request API access with Twitter developer account, Twitter support request some information about the project and how it will be implemented and how as developer use the data I will collect. I have provided all necessary information and I have earned it. I used the API Token to collect information Data from Twitter account (WeRateDogs) with compared wit hTwitter archived that has been provided from Udacity and stored in a JSON file to be read. The JSON file has plenty of Data so I had to read it in a Dictionary then save it in a python list. From the Python list I have created the DataFrame by specifying the necessary information.

Assessing the Data

For assessing Data and finding issues I have specified nine quality issues and two tidiness. Quality issue that means something either missing with the data or anything that can affect the process of analyzing the data. By assessing the nine issues and writing them down I would be able to fix these issues. Tidiness is the issue with table structure and by fixing the issue it can be easy to group the data and analyze it.

Cleaning the Data

Cleaning data is independent to assessing, once assessing all issues that may affect the analysis phase you can start cleaning the data. I have fixed the quality and tidiness issues, then I have merged all data frames in one dataframe and exported the clean data frame to csv file.