Data Wrangle Report

Introduction:

This project focused on wrangling data from the **WeRateDogs** Twitter account using Python. In this project I learned how to do Data Wrangling in real world. The main purpose of this Project is to Practice the things which I learned in Data Analyst nanodegree program Data Wrangling Section. **WeRateDogs** is Twitter account that rates people dog's with a comment about the dog. And give ratings to the Dog.

Wrangling Efforts

Project Stages :

The task given in the project are in different stages are given below..

1. *Gathering Data*
2. *Assessing Data*
3. *Cleaning Data*

Let's understand what I have done in these stages one by one ….

Gathering Data:

It is the first stage of Data wrangling in which we Assess Data from different resources. So, in our Project there are three data sets .

1. **Twitter archive enhanced Csv file:** This twitter archive enhanced csv file is provided by the Udacity and I easily downloaded it manually and further used this file in Data Wrangling

2. **Image Prediction Tsv file:** This file contain breed of dog, images Url but there is two ways to download this file manually or programmatically so I choose the programmatical way because it enhanced our skill. The file image_predictions.tsv is hosted on Udacity's Server we have to download it by using **Request library in python** and url information which are provided in project details section.

3. **Twitter API & Jason :** This is the most complicated part of our Data wrangling where we have to use **twitter Api keys** and tokens to archive the data of **WeRateDogs**. I query the twitter Api for each tweets which is in Jason format by using Python **Tweepy's Library** and I make two empty list one list contain number of error we got and another list contain the data we successfully archived, and stored in file '**tweet_jason.txt'** with encoding of **utf-8** then I create new dataframe of Pandas named as **tweets_data**. Then I further used this Data .

Assessing Data :

After Gathering Data , now I have three files. Assessing is also main part of Data wrangling . in this part I assess the document and wrote the quality and tidiness issues, which I have to solve this issue in cleaning Part.

Two Ways of Assessing :

1. *Visually*
2. *Programmatically*

Let's understand these ways one by one

Visually: for visual assessment I open 2 files in Excel and other ways for visual assessment using google spreadsheet. And I also visualize by printing all three entire dataframes in jupyter notebook.

Programmatically : I also visualize all three dataset by using different methods in python like **info(), describe(), value_counts(), len(), duplicated(), sample() etc..**

The best way to visualize is Programmatically because visually you can't find all issues. Then after assessing I wrote all quality issues and tidiness issues and try to fix them all in cleaning Section

Cleaning Data :

This Part of Data Wrangling is divided in three parts .

Define : In this Part I define the issue which I'm going to fix.

Code : In this part I fix the defined issue

Test : After fixing I checked that issue is fix or not.

- Firstly I make copies of original data so that after cleaning the Data I can see the changes.
- Then I join all data frames in single dataframe named as twitter archive copy so that I can perform all actions in one dataframe rather than all three dataframe .
- Then I melt 4 columns(,doggo,flooper,pupper,puppo) and make one column 'dog stage'.
- Then I delete column related to retweets because we didn't want these column, then I remove duplicate id's.
- Then I make two columns Prediction algorithm and confidence interval from several columns(p1, p1_conf , p2, p2_conf , p3, p3_conf) and then I drop these columns.
- Then I change the data types of columns.
- Then I convert the Null values to None types.

- Then I make dog gender column and I found that dog is male or female by extracting text.
- Then I fix numerator with decimals.

Conclusion:

Data wrangling is a core skill of an individual who handle the data.

In the whole project I use python programming language and some of it's important packages . but for visual assessment I use excel . This project is really interesting I learned a lot from this project. I learn how to use twitter api keys to archive data from twitter.

Pandas have special features and lots of libraries which are very useful in data wrangling. It can deal with variety of data .