Wrangle Report

Introduction

The purpose of this project is to put our knowledge into what we have learnt in Data Wrangling lesson in Udacity Data Analyst Nanodegree course.
Here we wrangle the data @WeRateDogs given by Twitter API.
WeRateDogs is a Twitter account consisting all lots of peoples dogs and rate them through their comments and other attributes

This report tells us the wrangling process of this dataset

Project Details

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

Gathering Data

The project consist of 3 datasets that 2 of them are given by Udacity and remaining one we need to access it from Twitter Dev account

Twiitter_archive_data : This is the first data that is give by Udacity and downloaded manually.In this data we are given the Twitter post link and caption for the given post

Image_predictions_data : This is the second data given by Udacity and downloaded manually. In this data we get the prediction output of a 3 different neural network that which breed of dog is in each twitter post and the confidence level of the prediction

Twitter_additional_data : This is the additional data we need to download it from the twitter dev console.
We need to create a twitter account if you don't have else we log into the twitter dev console using our twitter account and create an dev account.After that you are given a confirmation link when we sign up,this link will redirect to the console page where we can create an app or website for twitter tweets analysis.We can get the API key for downloading this data.We need to download this data using the twitter "Tweepy" library.It will take upto 30-40 minutes to download this json file and after this we need to convert it into Pandas DataFrame

Assessing Data

Here we will analyse the data visually and programmatically

Visually assessing means we will look into the 3 data in an Excel and checking if there is any quality issues in the data

Programmatically assessing means we will use python libraries to analyse the data of having any quality and tidiness issues

We will make notes for each dataset on where we have any issues in the data and we will use this notes in cleaning phase

Cleaning the data

Here we will make use of the issue notes in the assessing phase and try to solve the issues
This is the main part of wrangling process.Only after this process we can proceed into analysis and making deep learning model otherwise our analysis and prediction will be wrong.
First we need to make a copy of every data using pandas libraries and we try to solve issues one by one. The reason we are making a copy is because when we encounter any errors we can recreate the copy of original data and we can continue our cleaning process.

There was a couple of cleaning steps and most challenging for me is to extract the algorithm prediction and its confidence levels using an if-else statement

Conclusion:

Data wrangling is a core skill whoever handles the data This is not a one time step but its an iterative step where we have to go back steps in order to get a full clean data

I have used python programming language and jupyter-notebook and Excel to gather,assess,clean the data

- For gathering data we have vast python libraries that makes gathering easy.I used requests,tweepy libraries in this stage
- For Assessing in visual assessment I used Excel, and for programmatic assessment I used numpy and pandas libraries
- For Cleaning I also used numpy and pandas as my core cleaning libraries

After all this I can make insights and correct visualization using my matplotlib and seaborn library