Data Wrangling Report

Introduction

This Project was made because it would help me in the practical application of the skill and knowledge I learnt in the Data Wangling section of the Udacity Data Analysis nanodegree. To carry out this project I made use of a data set which I extracted from twitter. It was a data set of WeRateDogs which is a twitter account which gives ratings to dogs on twitter.

Project Flow

Following are the steps:

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

GATHERING DATA

Enhanced Twitter Archive: This File was provided by Udacity named twitter_archive_enhanced.

Image Prediction File: This File gave the information about the dogs breed and was hosted on the udacity servers which I downloaded programmatically.

Twitter Jason File: By using the twitter API I read the WeRateDogs data into a python data frame using the tweet ID.

ASSESSING DATA

For the assessment part of the project I made use of functions like info() which are part of pandas library of python to display the various information like number of rows and columns and the data type of each column. Once this Initial information was collected to get an in-depth knowledge of the data sets I used the value_counts() function to see the counts of each attribute value.

After this I gave a point by point description of the tidiness and quality issues in the data set.

CLEANING DATA

In this step each and every data set was cleaned separately starting with the twitter_archive data set.

The first this I did to clean this data set was to combine the various dog names used like pupper, puppo etc which were given as separate columns so I combined them to create one single column called types_of_dogs. After this I removed all the tweets and retweets without image from the data set. I had to also correct the numerators in decimal form and in fractions also the denominators which were greater than 10.

For the Image prediction file I dropped the unnecessary columns once they were combine into one. After all the data was cleaned I combined all the data sets in to a single data frame. For easy analysis.

Conclusion

This project was a great way to practically apply the knowledge and skill learnt inthis section.

This gave me an insight as to how to use the python libraries and twitter api to extract data from the web and then how to clean and analyze the data to give interesting insights into the data .