WRANGLE REPORT

Udacity : Wrangle and Analyze Data Project

INTRODUCTION

This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree . The project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. We first gather the data from three sources, then assess them to pick out quality and tidiness issues and finally clean the issues to analyse and visualise them to infer some facts.

GATHERING DATA

Data was gathered from 3 different sources:
1) The enhanced twitter archive file was provided and downloaded manually. This file includes various variables for each tweet including tweet id, timestamp, rating numerator and denominator, name, etc.

2) Additional data, including favorite count and retweet count, were gathered using the Twitter API, through a json file.

3) The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers.

ASSESSING DATA

After the data was gathered, assessment was performed using the following methods:

1. .head()
2. .tail()
3. .sample()
4. .info()
5. .value_counts()

CLEANING DATA

Tidiness issues:

1. Combining all dataframes together to make one master dataset as they all contained information about the same tweets.

2.We combine all the doggo, floofer, pupper, puppo columns into a single dog stage column.

Quality issues :

1.Dropping unwanted columns not needed for our analysis
2.Fix the timestamp column by changing it into datetime dataype
3.Replace the none names with NaN
4.Fix the rating by explicitly calculating it
5.Making a single column for all dog breeds for easier visualisation
6.Removing tweets without images
7.Removing retweets
8.Chaning tweet_id to string datatype

CONCLUSION

We usually never find all the data from 1 source.We need to gather the data from various sources and then make it clean and tidy for analysing purposes. This project emphasized that you will need Python and its various libraries to gather data from various sources, and then clean various quality and tidiness issues, before any data analysis can be performed on it.