Wrangle Report

Introduction

This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree. The dataset wrangled in this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a nice and funny comments about the dog.

The Purposes of This Project Are:

- • Gathering data
- • Assessing data
- • Cleaning data

Gathering Data

Data was gathered from 3 different sources:

1- **Twitter archive file:** csv file was provided by Udacity and downloaded manually.

2- **The tweet image predictions**: tsv file is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information

3- **Twitter API & JSON:** Additional data, including favorite count and retweet count, were gathered using the Twitter API.

Assessing Data

Assessment was performed using the following methods:

- head()
- sample()
- info()
- value_counts()
- Duplicated()
- Etc

Cleaning data

The issues found during the assessment process were cleaned and checked using the following methods and techniques:

- merge()
- drop()
- isnull()
- astype()
- to_datetime()
- melt()
- Drop_duplicates()
- value_counts()
- info()
- head()
- etc.

A very helpful step was to create a copy of the originals dataframes. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original.

Conclusion

The purpose of this project is to practice of what I learned in data wrangling section from Udacity Data Analysis Nanodegree program. I have been used Python and its libraries for this project to scrape data from different sources in different formats, and clean quality and tidiness issues, before any data analysis can be performed. Data wrangling is a core skill that whoever handles data should be familiar with.