Data Wrangling

Introduction

This is one of the project in Data Analyst Nano degree from Udacity. In this project I have wrangled the data from WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. They have a unique style of rating the dogs in a way that the numerator is always greater than the denominator on a scale of 10. The project used Jupyter notebooks for exection

Steps involved in wrangling

The wrangling of the dataset was done in three steps namely,

- Gather
- Assess
- Clean
- Store
- Analyze
- Visualise

Gather

In the 'Gather' step, all the required data set is gathered manually or programmatically. In this project, twitter_archive_enhanced.csv given by Udacity which contains twitter archive of WeRateDogs was downloaded manually. The image_prediction.tsv hosted in Udacity's server was downloaded programmatically. Then, some data are queried from the twitter using tweepy package and written into a text file.

Assess

In the 'Assess' step, the dataset is assessed for quality and tidiness issues. In this project, I have made the summary and description of the three data frames and documented 8 quality issues and 3 tidiness issues.

Clean

In the 'Clean' step, the data is cleaned for the issues documented in the 'Assess' section. Before starting the cleaning, the data is copied to another file, table or dataframe. The cleaning is done in three steps. First, the issue that is going to be cleaned is addressed in the 'Define' section. Then the codings is encapsulated in the 'Code' section. Then the result is tested in the 'Test' section.

Store

After performing the cleaning process, the cleaned data is stored in a seperate file.

Analyze

Then, analysis is performed on the cleaned data to find interesting patterns or relationships in the data.

Visualize

The visualizations are produced for the better understanding of the data and visually convey the result of the analysis.