Wrangling and Analysis of WeRateDogs Dataset

Introduction:

WeRateDogs is a twitter account that focuses on posting humorous tweets on dogs and provides ratings. WeRateDogs achieved universal popularity when a twitter user complained about their rating system. Their reply to user stating that ratings over 10 are provided because “they're Good Dogs Brent**”** became a celebrated exchange**.** Currently they over 7 million followers and own a clothing line.

WeRateDogs twitter data was provided by Udacity for analysis. The original dataset has several quality issues that we will clean. We will follow that with visualizations identifying key trends and useful insights.

Project details

- Data wrangling has three steps shown below:
  - Gathering
  - Assessing
  - Cleaning

- Storing, analysis and visualization

- Reporting on a) data wrangling efforts b) data analysis and visualizations

Data wrangling

Gathering

We gathered three different datasets for our analysis.

1. WeRateDogs twitter archive was provided by Udacity. This dataset was then read to a python dataframe called twitter_excel.

2. Each tweet's retweet count, favorite count, user favorites count and user followers count was obtained using the twitter API. The orginal information for authorization was obtained from the twitter account. This data was then parsed into a json object using tweepy. Data for each tweet id was obtained and stored in a dictionary. This json data was then stored to a tweet_json.txt file.

3. Tweet image predictions data was obtained using the url provided by Udacity. The data was pulled using the requests library and the content was written to image-predictions.tsv file. This dataset consists of dog breed predictions and the confidence level for each predicton.

Assessing

Each dataset was assessed using various python functions. The variables in the dataset where described for better understanding of the reader. The goal of the project was to identify 8 quality issues and 2 tidiness issues. I identified 12 quality issues and 2 tidiness issues. Finally 8 quality issues and 2 tidiness issues were fixed in the dataset.

Cleaning

I cleaned the dataset by focusing on individual issues in quality and tidiness. Each issue was defined with the process for correction in the define section, the code for correction was provided in the code section and the new dataset was tested in the test section.

- The original twitter archive dataset consists of the retweet and reply data. This retweet and reply data was removed from the dataset in the quality issues section.
- The rating numerator, rating denominator and dog names were incorrect. This was fixed in the quality issues section
- The dog stages were divided into four columns. This was changed to a single section called dog type in the tidiness section.

- All 3 datasets have the tweet id to be column. This column was used to merge the datasets in the tidiness section.

Storing, analysis and visualization

The cleaned data is stored in the twitter_archive_master.csv file. This data is then read to a dataframe for analysis and visualization. I was required to produce atleast 3 insights and 1 visualization. I created 9 visualizations and several insights from the plots.

Reporting:

Two reports were created from the data gathered, cleaned and visualized. First is the current report, detailing the steps in gathering, assessing and cleaning. The second report called act_report details the insights and trends observed from the visualizations.