Wrangle Report

Udacity Project: Wrangle and Analyze Data

Wrangle and Analyze Data

In this project, we wrangled WeRateDogs Twitter data. The data was then used to create visualizations and related analysis to get more insight into data.

We followed the data wrangling steps:

1. Gather
2. Assess
3. Clean

Data had to be gathered from multiple sources:

1. Udacity servers to get image prediction data in a TSV file
2. Provided 'twitter-archive-enhanced.csv' file
3. Twitter servers to get 'tweet_json.txt' file

Data was gather using multiple methods:

1. Used request API to download data from Udacity servers
2. Used tweepy library to get data from twitter.com
3. Used provided CSV data

After gathering the data, it was stored in 3 pandas dataframes where assessment and cleaning process was done. I used following functions to assess the data quality issues:

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

Data quality issues that were cleaned are as follows:

- In archive data expanded_url column
  1. Urls which are not hosted on twitter.com (vine.co, gofundme etc)
  2. Urls not related to dog_rates/status

3. Urls which are video instead of photo
4. Same Url inserted multiple times in a cell

- In archive data rating numerator and denominator columns
  1. Denominators are found to be other than 10, including 0
  2. Some numerator values are very high
  3. No standard rating available due to changing denominators

- In archive data timestamp and name columns
  1. Changing timestamp column data type and filter records
  2. Remove data after 08/01/2017
  3. Wrong names, some random words entered as names

- In archive data source column
  1. Remove unnecessary data and keep only the name of the source as a single word

Apart from quality issues, the data tidiness issues which were fixed are:

1. In archived data, dog stage should ideally be a column with doggo, puppo, pupper and floofer as values for the stage columns. Instead, we see these values in their own columns.
2. Merge 3 different data sets based on tweet id
3. Remove unnecessary columns which are not required, especially the ones which are related to retweets

Total 11 quality and 3 tidiness issues are targeted to be fixed.

Functions used for quality and tidiness fixes are:

1. Function merge()
2. Function melt()
3. Function apply()
4. Function unique()
5. Function notnull
6. Function sorted()
7. Function append()
8. Function drop()
9. Function astype()

10. Function drop_duplicates()
11. Function reset_index()
12. Function remove()

There could be more functions used but provided above a list of functions most used to clean the data with tidiness and quality issues.

The data available to us from the various sources was good but still not in a state where we can start using it to gain insight into the topic. It had to wrangled before any visualization could be made. Assessing the data to find all the issues a challenging and time-consuming task. In the end, I was able to clean the data and get it ready for analysis using various charts. Library pandas provide great power to data analysts to wrangle the data to fix all different types of issues before it becomes ready for further analysis.