Data Wrangling Report

Introduction

Real-world data rarely comes clean, they always come with quality and tidiness issues. Data used for analysis also rarely come from one source. They are aggregated from various sources and cleaned before analysis. This is known as **Data Wrangling**.

This data wrangling project involves tweet archive of Twitter user, @dog_rates, also known as **WeRateDogs**. **WeRateDogs** is a Twitter account that rates people's dogs with a humorous comment about the dog.

The three steps in data wrangling - Gathering, Accessing and Cleaning was implemented in this project, using python and its libraries. **Twelve quality issues** and **three tidiness issues** were addressed.

1. Gathering Data

I sourced the data for the project from three different sources as required. The twitter archive csv file shared with Udacity by twitter was downloaded manually from the website, which contained basic information on tweets as stored in twitter objects.

The second dataset was the image predictions on each tweet using neural network algorithm, which was programmatically downloaded from Udacity server using the weblink, and saved it locally as a tsv file.

The final dataset was sourced from twitter using twitter API (tweepy). This involved creating a developer account and app with twitter and generating access keys.

Using the tweet IDs in the Twitter archive, I accessed the entire data for every tweet from Twitter API, returning just the tweet_id, retweet_count and favorite_count for the tweets that could be accessed, which I converted later into a pandas dataframe.

The above datasets were then read into jupyter notebook, using the correct separator, for the data wrangling project

2. **Assessing Data**

Observing the twitter archive dataset visually in excel, I spotted some quality issues like multiple and duplicate entries in the expended_urls column, urls not linked to twitter, source column information recorded as html tags, rather than utility names.

Other quality and tidiness issues were spotted programmatically using pandas functions. They include:

- Incorrect datatypes
- The dog stage variable recorded in four columns
- Null values

- Incomplete rows of data for image prediction table and tweet_json table
- Columns that should be dropped, which represented duplicates (retweets)
- Some errors already captured in the motivation page, such as incorrect dog names, wrong ratings on both the numerator and denominator, etc

3. Cleaning Data

During the accessing operation, I documented the quality and tidiness issues I spotted in the datasets and they were all addressed in this stage, using the three steps of the data cleaning process – **Define, Code, Test.**

I made a copy of the tweet archive and image predictions datasets, on which the most of the cleaning process occurred, ranging from quality to tidiness issues. I completed the major tidiness issues first which also enhanced cleaning of the quality issues.

4. Storing Data

After the completion of the cleaning process, I stored the clean datasets in twitter_archive_master.csv and image_predictions_master.csv files as required.