Data Wrangling Report

Introduction

Data wrangling is a core skill that everyone who works with data should be familiar with since so much of the world's data is not clean. It is a process divided into 3 main steps:

- Gathering.
- Assessing.
- Cleaning.

Gathering

Data was gathered from 3 different sources:

1. From WeRateDogs Twitter archive in csv format:

Using panda's method 'read_csv', I managed to read the data stored in the file 'twitter-archive-enhanced.csv'. I stored it in a DataFrame called 'twitter_archive'. The data has many issues that will be cleaned and resolved later.

2. Image prediction file downloaded programmatically using Requests library and the URL in tsv format:

Using Requests library and 'get' method, data was downloaded in a file 'image_predictions.tsv'. Then, the content was stored in a DataFrame called 'image_predictions' using pandas' method 'read_csv'.

3. Data retrieved by querying Twitter's APIs and using Tweepy library.

I did not create the account, but used ‘tweet_json.txt’.

Assessing

After gathering the data and storing them in DataFrames, the following step was assessing the data for quality and tidiness. Data were assessed programmatically and visually.

Quality and Tidiness

- No need to all the information in images dataset, (tweet_id and jpg_url what matters).

- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo.
- Join 'tweet_info' and 'image_predictions' to 'twitter_archive'.

twitter_archive dataset

- Columns like in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers/strings instead of float.
- Name column have invalid names i.e 'None', 'a', 'an' and less than 3 characters.
- Cloumns like retweeted_status_timestamp, timestamp should be datetime instead of object (string).
- Sources are unreadable.
- The numerator and denominator columns have invalid values.
- In several columns null objects are non-null (None to NaN).

image_predictions dataset

- Some tweet_ids have the same jpg_url

tweet_data dataset

- This tweet_id (666020888022790149) duplicated 8 times

Cleaning

It is the process of fixing and resolving issues identified in the Cleaning process. The (define, code, and test) steps were used in the cleaning process. First, copies of
the DataFrames were created before cleaning. Then, the steps of cleaning were applied iteratively on all issues.
Storing
The final DataFrame called 'twitter_archive_clean' contains 1981 rows and 13 columns with the correct data types. The dataset is then stored in a csv file called 'twitter_archive_master.csv'. At this point, the data was successfully wrangled and
therefore ready for analysis and visualization.

Analysis & Visualization

These steps are not part of data wrangling process. However, it cannot reflect correct and accurate insights without performing data wrangling first. Visualizations and insights are provided in ‘act_report.pdf