Data Wrabling 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- WeRateDogs Twitter archive given by Udacity in csv format

I used pandas .read function to be able to import the archive file in my workspace, that file was named as twitter-archive-enhanced-2.csv, i loaded that file as image_df that file had many issues that will be cleaned later

2- Image prediction file

I donwload this file manually from udacity workspace and also imported it using the function .read_csv and the file was named as image-predictions-3.tsv , i loaded this file as archive_df

Data retrieved by Twitter's APIs

I donwload this file manually from my classroom workspace cuz i had problems with twitter developer account that file was named as tweet-json.txt and i loaded it as json_df

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

Issues with content. Low quality data is also known as dirty data. Identified quality issues are

- We need to remove the retweeted tweets
- Changeing the typ of those columns types to strings (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and tweet_id) to strings as we don't need to do any operatins on them
- The numerator and denominator columns have non true values.
- There is invalid names (less than 3 chars).
- reply_to_status_id, reply_to_user_id, retweeted_status_id, retweeted_status_user_id, should be integers or strings instead of float.
- Some tweet_ids have same jpg_url.
- Some tweets have 2 different tweet_id.
- retweeted_status_timestamp, timestamp should be datetime instead of strings.
- Missing some images from dataset 2075 rows instead of 2356.

Tidiness

Issues with structure that prevent easy analysis. Untidy data is also known as messy data.

- There is 4 stage columns (doggo, floofer, pupper, puppo) We should delete them.
- We need to merge all the dataframes(image_df, json_df) into archive_df

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 'archive_clean' contains with the correct data types. The dataset is then stored in a csv file called '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

In [ ]: