Wrangle Report

Wrangle and Analyze Data Project

Introduction

This project is part of Udacity's Data Analyst Nanodegree program, and it's about wrangling and analyzing data. The dataset that I will be wrangling is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

The method of data wrangling is split into three important sections:

- Gathering data.
- Assessing data.
- Cleaning data.

Gathering Data

I gathered the data from three different sources:

1. **The WeRateDogs Twitter archive:** I downloaded this file (twitter_archive_enhanced.csv) manually from Udacity's website.
2. **The tweet image predictions:** this file (image_predictions.tsv) is hosted on Udacity's servers and I downloaded programmatically using the Requests library.
3. **Twitter API:** by using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assessing Data

After gathering the data, I assessed them visually and programmatically for quality and tidiness issues. In the end of assessing process, I detected and documented many points about quality and tidiness issues as following:

Quality Issues

twitter_archive

- Delete retweets to keep the original tweets because we want only original ratings that have images.
- The data type of (tweet_id) column should be converted to string.
- The data type of (timestamp) column should be converted to date.
- Remove the rows and columns that we don't need them.
- The numerator and denominator columns have invalid values.

image_predictions

- Remove duplicate in (jpg_url).
- The data type of (tweet_id) column should be converted to string.
- Capitalize the first letter of (p1, p2, p3) columns values.
- p1,p2 and p3 have (_ and -) instead of space.
- Remove columns that we don't need them.

df_tweet

- The data type of (tweet_id) column should be converted to string.

Tidiness Issues

- Merge the three dataframes into one dataframe.
- Melt (doggo, floofer, pupper, puppo) columns into one column.

Cleaning Data

The third process of wrangling the data is cleaning. First, I made copies of the DataFrames. Second, I started to clean quality issues, then tidiness issues. In every issue, there were three key steps: define, code and test.

Conclusion

I'm still learning about data processing, but in this moment, after completing this project, I learned and discovered a lot of stuff that would help to wrangle and analyze the data. I used codes and some of Python 's libraries in Jpyter Notebook for this project.