Wrangle Report

Introduction

The purpose of this project is to wrangle data in order to create interesting and trustworthy analyses and visualizations. It is also to put to test the lessons on data wrangling that I've been taking in the last 3 weeks from Udacity Data Analysis Nanodegree program.

Project Steps Overview

The tasks in this project are carried out as follows:

1. Gathering data

The data for this project were gathered from three different format and sources. They were obtained as follows;

a. Twitter archive file: The twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
b. The tweet image predictions: This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
c. Twitter API & JSON: By using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

2. Assessing data

Once the three datasets were gathered, I assessed the data using two methods:

a. Visually: I used the .shape, .head(), .tail() and .sample() methods of the pandas library to scan through the datasets. I also opened the two flat files in Microsift Excel to have a better look.
b. Programmatically: This is done by using different methods of the pandas library (e.g. info, value_counts, duplicated, describe, unique, etc).

The issues I discovered about the datasets were documented under quality and tidiness issues subheading.

3. Cleaning data

This part of the data wrangling is where the problems identified and documented in the assess stage gets fixed. The very first thing I did in this stage was to create a copy of each of the dataset. All of the cleaning was done on the copied datasets. The cleaning was done under three sub-headings: Define, code and test, and it is just as straightforward. In define, I stated the cleaning plan, in code, I wrote codes to execute the plan and, in the test, I confirm the fix using another code.

I wrote the codes to manipulate the copies until I arrived at a dataset clean enough for the kind of analysis I wanted to carry out. In the cleaning stage, I deleted irrelevant columns, removed duplicates, removed retweets and replies, changed datatypes and melt four columns into one, before combining the datasets into one using the merge method of pandas.

4. Storing Data

The wrangled and combined dataset was then stored as a csv file in my directory. I did this to be able to load the dataset whenever I want without needing to go over the wrangling again.

5. Analysis and Visualizing Data

The wrangled data was analysed and visualized using seaborn, numpy, matplotlib, Image and Pandas libraries. Insights were drawn and documented. More on this is recorded in the jupyter notebook and the act report.

Conclusion

Data wrangling is a core skill required by data analysts. It is a serious business and the consequences of not using it, can have a major impact. It ensures that an analyst is building on a solid foundation.

This project has been instrumental in reinforcing the data wrangling skills that I have been taught and I am confident that I can now take up any dataset, no matter how untidy and messy it may be, and make something meaningful from it.