Wrangle and Analyze Data

Introduction:

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) is the 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.
WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon

1- **Gathering Data :**

We gathering the data from 3 different resource

- **Enhanced Twitter Archive :**
  Download the file "twitter-archive-enhanced.csv" was provided by Udacity

- **Image Predictions File :**
  Download the file (image_predictions.tsv) by Udaciy, read the file programmatically using the Requests library and URL information

- **Additional Data via the Twitter API:**

Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for JSON data by using Tweepy library , store the tweet entire of JSON data in file “tweet_json.txt “ , read the file pandas dataframe

2- Accessing Data :

-This step allows us to identify quality and tidiness issues

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues.

There are two type of assessment:

- **Visual assessment:**
  Which consists in scrolling through the data in Google Sheets or Excel to name a few - I first took a look at the cvs file in Google Spreadsheet.
- **Programmatic assessment:**
  For use code such as pandas head, tail, describe, shape, sample, value_counts and info methods.

List of issue we have :

1- Tweet don’t have image
2- Missing value in dog stage
3- Some of numerator over than 10
4- Retweet therefore duplicate
5- Rating numerator less than 10
6- Dog name start with lowercase
7- Tweet_id , time stamp , source , img_num ,dog_stage are erroneous datatype
8- Breeds p1 p2 p3 have upper case

Tidiness issues :

1- Drop rating_numerator over 10

2- Drop rating_numerator less 10

3- Cleaning data

The of cleaning data of the data wrangling was divided in three parts: Define, code and test the code These three steps were on each of the issues described in the assess section.

The data are cleaning using method to clean (drop , isnull , etc) at the end of each clean section I test the dataset to make sure that the clean operations are work correctly .

Conclusion :

I use many different library in this project (Wrangling and Analysis) Panads , numpy , matplotlib , requests , time , json , tweetpy .. that allow me to gather the data , Access and clean it .

Maybe this is one of the most project that take more time and effort but finally I did it ..

wrangle_act.ipynb *: code for gathering, assessing, cleaning, analyzing, and visualizing data*
wrangle_report.pdf : *documentation for data wrangling steps: gather, assess, and clean*
act_report.pdf : *documentation of analysis and insights into final data*