# Wrangle_Report

September 7, 2020

## 0.1 WeRateDogs Twitter Archive - Wrangle Report

### 0.1.1 Project Wrangle and Analyze Data on the WeRateDogs Twitter Archive

**Sherif Sakr, Egypt, Cairo, 7 September 2020**

**The Project wrangle and analyze data as a part of Data Analysis Nanodegree program, Udacity is focus on the** wrangling efforts in assembling , gathering, data required for analysis of the WeRateDogs Twitter Archive. However,these project consists of five steps (data gathering, assessing data,cleaning data, store data, and finally anlyze, visualis and report data)

Dataset from Twitter archive of user @dog_rates, also known as WeRateDogs, which is a Twitter account that rates people's dogs with a humorous comment about the dog. Only the original ratings not retweets that have images. There are more than five thousands tweets in the data set, not all are dog ratings and some are retweets. I do not need to gather the tweets beyond August 1, 2017, because i can not gather the image predictions for these tweets. since I can not have access to the algorithm used. Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate my skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset. Cleaning data including merging individual pieces of data according to the rules of tidy data. The fact that the rating numerators are greater than the denominators does not need to be clean. This unique rating system is a big part of the popularity of WeRateDogs.

**The project Consists of five Stages**

- First: Gathering data from three different sources, and three different format of files
- Second: Assess Data (Quality & Tidiness)
- Third: Clean data (Define, Code, Test)
- Fourth: Store data (Flat files and Database)
- Fifth: Analyze, Visualize, and Report

**The Programms Used**

- Using Command line rather than graphical user interface (GUI) for install programmes and execute various command line codes, such as Git Bash
- Using Atom or Sublime as a text editor

- Using Python programmatic language
- Using Python packages(libraries) such as NumPy, Pandas, Matplotlib, tweepy, seaborn etc.
- Using Jupyter Notebook, Anaconda, as a programming environment

- Using Microsoft Excel or Google Sheets for visual assessment on datasets.

**Jupyter Notebook environment and Python libraries:**

- Using Numpy for accessing, deleting, inserting elements into ndarrays, however, Numpy is the fundamental package for scientific computing in Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Using Pandas as a package of Python for data manipulation and analysis, however, Pandas incorporates two adational data structure Python, namely Pandas Series and Pandas DataFram.
- Using Matplotlib for data visualization. However, Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

**Using various functions and methods, for instance:**

**Pandas functions and methods**

- Use the Pandas.head() function to view the first few rows of a Series or DataFrame object.
- Use the Pandas.tail() function to view the last few rows of a Series or DataFram object.
- Use the Pandas.sample() function to view a random rows of a Series or DataFrame object.
- Use the Pandas.DataFrame.info() to print information about a DataFrame including the index dtype and columns, non-null values and memory usage.
- Use the Pandas.DataFrame.shape to return a tuple representing the dimensionality of the DataFrame.
- Use the Pandas.DataFrame.size to Return an int representing the number of elements in this object.
- Use the Pandas.DataFrame.ndim to Return an int representing the number of axes / array dimensions.

**NumPy functions and methods**

- Use the numpy.random.rand() function to generate random values in a given shape.
- Use the numpy array slicing [start:end].
- Use the numpy.indices() function to return an array representing the indices of a grid.
- Use the numpy for importing and exporting files (np.loadtxt('file.txt')), ('np.read_csv(file.csv') and np.savetxt('file.txt',arr,delimiter=" ").

### 0.1.2 The project Consists of five Stages

**First: Gathering data from three different sources, and three different format of files**

- Twitter archive file (twitter-archive-enhanced.csv)
- the Twitter image predictions file (image-preditions.tsv)
- Twitter API & JSON file

**Second: Assess Data (Quality & Tidiness)** There are two types of Data Assessment:

- Visual assessment: the three datasets (gathered data) is displayed in the Jupyter Notebook for visual assessment purposes, once displayed, data can additionally be assessed in an extra applications such as Microsoft Excel, Google Sheets, and Text editor.
- Programmatic assessment: using Pandas' functions and methods.

### Third: Clean data (Define, Code, Test)

- This part of data wrangling was divided into three parts: Define, Code, and Test the Code.

### Fourth: Store data (Flat files and Database)

- Save the Pandas DataFrame into CSV file format (flat file).
- Store the Pandas DataFrame in SQLite DB.

### Fifth: Analyze, Visualize, and Report

- Apply Descriptive Statistics.
- Apply Using Matplotlib as a plotting library for data visualization.