Wrangle Report

Udacity
Data Analyst Nanodegree Program

Tariq Ibrahim Alhajri

For @Dograte tweets from 2017 onwards

Before start analyzing our dataset, we need to wrangle our dataset in order to get the best outcome for the analysis, since that dirty or messy data could lead to a wrong analysis conclusion that a decision might relies on.

So, What is dirty and messy data?

First let' identify Data wrangling, is the process of three steps: -

- Gathering your data form variety of sources
- Assessing its quality by observing the quality issues or structural issues
- Cleaning issue identified in the assessing step to make sure it's not messy or dirty before start analyzing your data.

Low quality data is commonly referred to as **dirty** data, whereas Untidy data is commonly referred to as **messy data**, it is about the structure of data which can be identified with three points:-

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

My wrangling process in this project: -

1- Gathering your data:-
   - I gathered three datasets in this process from three variety sources
     - Read twitter_archive CSV file via read_csv Pandas's method.
     - Image_predictions file was downloaded programmatically using the *Requests* library.
     - Read tweet_json file via read_json Panda's method.

2- Assessing the quality: -
   - Assessing the quality was done programmatically by using head(), info(), describe(), dtypes(), sample(), duplicated(), and tail() methods. Or visually by simply Open your data in Google Sheets, Excel, or any text editor, looking for quality and tidiness issues. .

   Issues I found in each dataset: -

     - For Twitter_archive:
       - Quality issues:

1- timestamp as object (string), needs to be converted to DateTime datatype.

2- tweet_id as int64, needs to be converted to String datatype.

3- Deals with records that has a denomiator higher than 10.

4- Data in source column has a href html tag, needs to be fixed.

5- Delete all retweeted tweets 'duplicate tweets'.

- Tidness issues:
  1- doggo, floofer, pupper and puppo needs to be in one column rather than 4."Each variable is a column".

  2- Remove unnecessary columns(in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp,..etc)

- For image_predictions dataset:
  - Quality issues:
    1- Remove duplicate jpg_url.

    2- Change datatype of tweet_id column.

  - Tidness Issues:
    1- Extract the breed of dog from the p,p_conf and p_dog columns.

3- Cleaning the issues:

- Cleaning was done by using methods in Pandas's library such as: -
  - astype() method is used to change the column's datatype.
  - Pd.to_datetime() method is used to convert a column from string type to datetime type.
  - drop_duplicates() method is used to remove duplicate records.
  - rename() method is used rename a column's name.
  - drop() method is used to delete unneeded columns.
  - I also create two functions one is called extract_breed and was used to store the predicted breed with most confidence.
  - The other method is called stage() which was used to convert the 4 columns (pupper, floofer, doggo, puppo) into one column to make analysis much easier.