We Rate Dogs

A Data-Wrangling project by Ramy Nouh

Wrangle Report

This report briefly describes the data wrangling efforts exerted through stages in this project. The dataset that was wrangled (and analyzed and visualized) is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. This is a Twitter account that rates people's dogs with a humorous comment about the dog." WeRateDogs has over 4 million followers and has received international media coverage.
The entirety of this project was completed on Udacity Workspace However, the reports were created and exported as PDF's using Microsoft Word.

The Wrangling process is divided into three steps:

1. ***Gathering Data***
2. ***Assessing Data***
3. ***Cleaning Data***

Each step will be further explained below:

Step1 :Gathering Data ;

In this stage, the three pieces of data were gathered and represented as pandas dataframes:

- The WeRateDogs Twitter archive (file on hand, manual download of twitter-archive-enhanced.csv') .

- The tweet image predictions ('image-predictions.tsv'). This file was be downloaded programmatically using the Requests library from a provided URL.

- Each tweet's entire set of JSON data in a file called 'tweet_json.txt' and opened by json lib to read it line by line into a pandas DataFrame and was later saved to a 'twt_data.csv' file for future use (without the index column so it wil not appear as unnamed column in the file).

Step2 : Assessing and Cleaning Data ;

While working with data, a number of observations were made. In the below table there are the observations along with actions taken in the Cleaning Step.

1: Quality ;

Dataset	Observation	Solution
*A Enhanced Twitter Archive:	_AQ1 Replace 'None' with `np.nan` for Columns (`doggo`, `floofer`, `pupper`, `puppo`).	-Replaced Non values with np.nan
	_AQ2 Extract rating scores correctly from tweet text using RegEx and convert it to float.	-Extracted the rating score correctly and converted it to float
	_AQ3 Convert `timestamp` column to datetime.	-Converted timestamp to datedime data type using pandas to_datetime function
	_AQ4 Remove retweets and replies.	-Remove retweets and replies.
	_AQ5 Remove any rows not related to dogs.	Keep rows only rate dogs and remove not related rows
	_AQ6 : - Replace 'None' with np.nan in clean_twt_archive `name` column. - Remove any rows with invalid names which starts with lower letters.	-Replaced None and invalid names with np.nan - Remove any rows with invalid names
	_AQ7 :Extract tweet source from `source` column	-Extract the source of the tweet and covert it to categorical.
**B twitter image predictions:	__BQ1 Some p names start with lowercase letters covert it to uppercase	-Covert lowercase letters to uppercase
***C Tweet data:	___CQ3 Rename `id` column in clean_twt_data to `tweet_id`	-Renamed it to match the other 2 datasets

2: Tidiness ;

Dataset	Observation	Solution
*clean_twt_archive	-T1 Create `dog_stage` column and remove the (`doggo`, `floofer`, `pupper`, `puppo`) columns.	-Created one colum dog_stage and removed the 4 columns
* clean_twt_data	-T2 Remove unnecessary columns for clean_twt_data.	-Removed other columns
* clean_img_pred	-T3 Remove `img_num` column from clean_img_pred.	-Removed it
*clean_twt_archive / clean_twt_data / clean_img_pred	-T4 All data is related but devided into 3 dataframes	- Combined all the 3 datasets into one pandas df

Result ;

A combined data set with all needed information was stored in csv file ('twitter_archive_wrangle_act.csv').