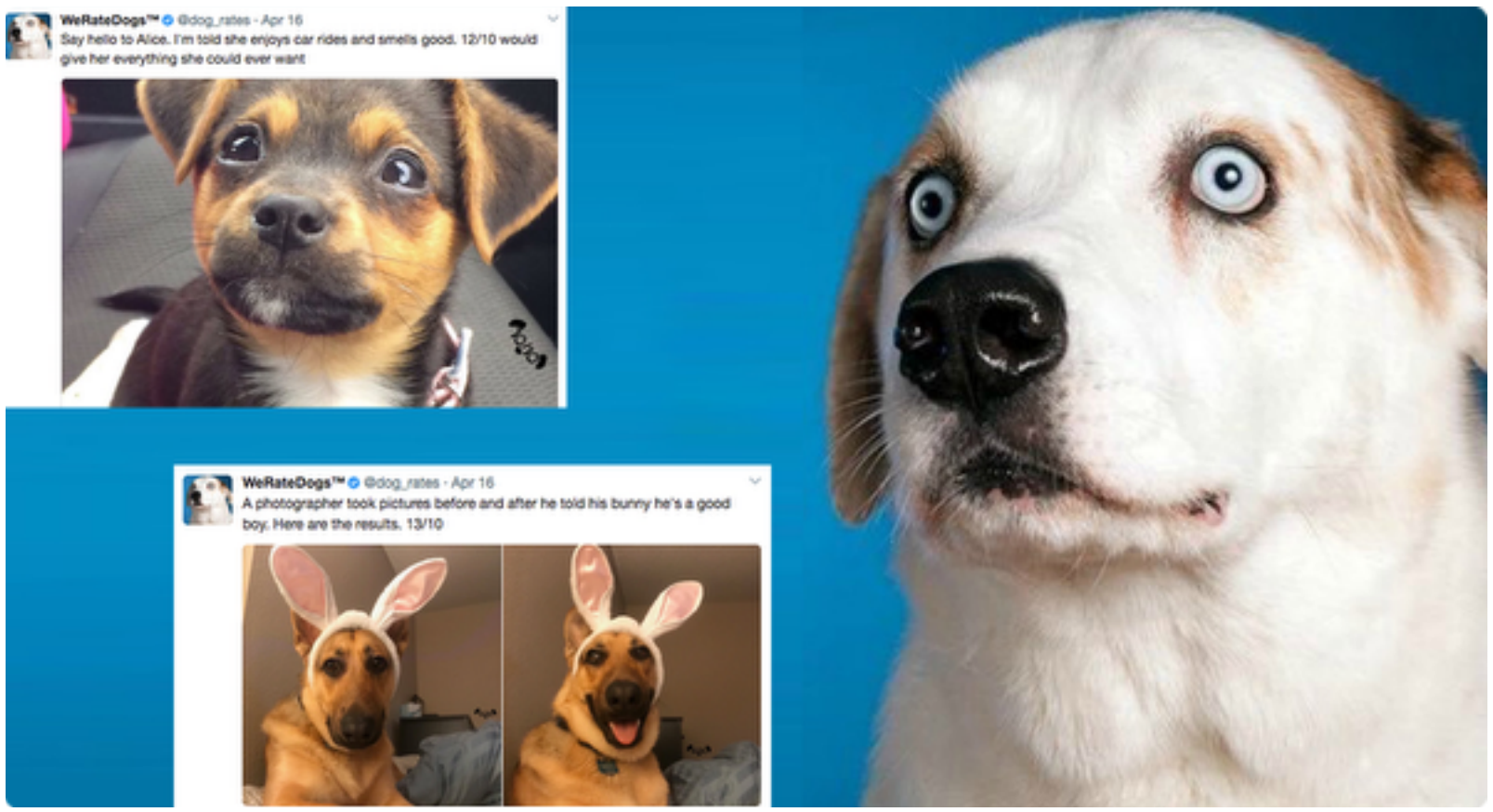
Udacity

Data Wrangling Project



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.

Udacity

Wrangling Actions Taken

The project data files were downloaded and acted upon in jupyter notebook. Few sample data problems that had to be solved in the wrangling stage are illustrated below:

For df_twitter_archive - Wrong data existed in the following columns: 1.in_reply_to_status_id 2.in_reply_to_user_id 3.retweeted_status_id 4.retweeted_status_user_id 5.timestamp 6.retweeted_status_timestamp. Missing data in the following: incorrect rating_numerator and rating_denominator for row numbers : 1069, 1166, 2336. For df_image_predictions - Missing Data (2075 records instead of 2356)

For new_df - All dataframes exist as individual dataframes but they contain data for same tweet ids so it should be a single dataframe.

Wrangling Actions Summary - A copy of the dat was made. Datatypes were corrected. Twitter data was checked for accuracy. URL was expanded in case it wasn't null. Incorrect ratings were rectified. Redundant entry was removed. Missing values were detected and corrected. Retweets and Ratings were made consistent. Cleaned data was merged to give one master file. The cleaned data was stored as master file for analytical visualisations to be done.