WeRateDogs Wrangle Report

Hanna Kondrashova

Project: Wrangle and Analyze Data

This well-thought project was aimed at polishing our data wrangling skills by gathering data from different sources, assessing and then fixing various tidiness and quality issues.

1. Gather

Our data was to be gathered from three different sources. The WeRateDogs Twitter archive was already gathered for us and we only needed to import this csv file to our jupyter notebook as a pandas dataframe **twitter_archive**. The second data file, image_predictions.tsv was programmatically downloaded using the Requests library from Udacity's server using the following URL . This file was also imported as a pandas dataframe **image_predictions**. The third file was acquired from Twitter API using python library Tweepy. This data was stored in JSON format using UTF-8 encoding.

2. Assess

Having *gathered* our data, we assessed them visually and programmatically for quality and tidiness issues. The data we got to work with after gathering and accessing appeared to be not perfectly clean, but at the same time positively not dirty - **messy** only. There were both **quality** and **tidiness issues** which we had to get rid off during *cleaning* stage.

3. Clean

As for the **quality**, first thing to do with our **main data frame** was to remove *retweets data* which we didn't need for this project. *"Expanded_urls"* column also had many tweets without images - we removed them too. Column *"in_reply"* had numerous missing values and needed to be removed as this data was not helpful for our research. We also optimized *long url links and sources* for the purposes of readability, removed *extra characters after &* and went through the list of *dog names* working at the names which were not names at all. Then we also removed incorrect values from *rating_numerators and denominators* and made those with decimals showing *full floats*. Furthermore, *"timestamp" data* was incorrect and thus was fixed into datetime regime.

Image_predictions file also had some issues - there we renamed *prediction algorithm columns p1,p2* etc. for better understanding and *unified first letters* in dogs breed predictions which were both lower and upper case.

Tidiness assessment consisted of two steps. First was creating a *stage column* to contain all the '*doggo', 'floofer', 'pupper', 'puppo'* columns content from **twitter_archive**. Secondly, the *'tweet_data' and 'image_predictions'* tables were joined to the *twitter_archive table*.

Having performed all the cleaning, we exported our data to a new csv file named ***'*twitter_archive_master.csv*''*.**

Conclusion

It goes without saying that we could have done much more thorough analysis, but for the purposes of this research our decision was to stop here as full assessment of this data would require exceptional efforts. For us it appeared to be a challenging, but very exciting project of data wrangling, which led to deeper understanding of concepts, methods and good practices.