Carmen Requena, Data Analyst Nanodegree

WRANGLE REPORT

In this project I did all the wrangling process for data related to a Twitter user called WeRateDogs. Our goal was to create insights and visualization of the data.

The 3 steps carried out in this project to wrangle the data were:

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

With this information, I extracted 3 insights and 3 visualizations.

1. Gathering Data

I gather along 3 pieces of data:

1. “Twitter_archive_enhanced.csv” which was downloaded manually, and called “df”.
2. The “tweet image predictions”, a file that was hosted in the Udacity server, and I downloaded programmatically using the request library. It contains 3 predictions for the type of breed of the dogs in the picture. Number 1 prediction was the most accurate one.
3. Each tweet's retweet count and favorite ("like") count were extracted using the tweet IDs in the WeRateDogs Twitter archive, querying the Twitter API for each tweet's JSON data using Python's Tweepy library. This was archived in a file called “Tweets_df”, that I later converted in a file called “Tweets_short”.

2. Assessing Data

For this phase, I used both visual and programmatic techniques to assess the data.

I first made a copy of the 3 data sets in order to conserve the original ones. Then I did a selection of columns from the tweets_df, since it was very large, and I only was going to analyze 2 of their columns.

For the Visual Assessment I downloaded the 3 files and explored them in both Excel and the tables of the Jupiter notebook. I used scroll, filter and review of the columns names to do my analysis.

For the Programmatic Assessment, I used a range of methods to assess the data of the 3 data frames. Mainly .info, .describe, .unique, .count_values, etc.

The main results obtained were:

- **Quality issues found:**

1. There are 746 names missed, but cataloged at none, that could generate a possible pandas misleading result for the None.

2. In the prediction table, there are predictions of not dog breed. Not accurate data.

3. Timestamp and retweeted_status_timestamp had incorrect data type (object instead of date time)

4. Rating denominator has extreme values: 110(1),120(1),130(1),150(1),170(1). I decided to consider extreme values over 100. They are possible outliers for visualizations and medias

5. Rating numerator also has extreme values: 420(2),666(1),960(1),1776(1). I decided to consider as extreme those over 400.

6. There are 181 retweeted tweets, but we are only interested in analyzing the originals.

7. There are 78 replies to the tweets, and again, we are only interested in the original tweets.

8. There are some tweets that do not have an image, when this is the core of our scope.

9. The Source column in the Archive table is confusing, it is difficult to clearly see where the tweet comes from.

- **Tidiness Issues:**

1. Twitter archive, Image prediction and twitter_df tables can be merged in one unique table, since they have one common column (tweet id). This will make it easier and more clean for our Exploratory Data Analysis.

2. Dog stages are 4 possible (doggo,pupper,floofer,puppo) and they are a variable, more than separate columns, so we decided to merge them in one column.

3. Since the more accurate prediction in the Image Prediction Table is the P1, we can drop the other to facilitate analysis.

3. Cleaning Data

Finally, for the wrangling process, I used programmatic methods to clean the quality and tidiness issues detected while assessing the data.

I followed a scheme for the cleaning: First, I cleaned the completeness issues, mainly missing data. Then, the tidiness issues. And at the end, the rest of the quality issues.

The final file after merging all the data frames was called master_df. Once all the cleaning was completed, I uploaded this file as a “csv” file called “twitter_archive_master”, which is the final document I then analyzed and made visualizations with.