Wrangle Report for WeRateDogs Data

By Damla Cörüt

Introduction

In this report, we will mention the wrangling efforts for the project 'Wrangle and Analyze Data'. WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog. We used the Twitter archive of the account as our dataset.

To be able to get a healthy dataset, we performed 'Gather', 'Assess' and 'Clean' steps, which concludes our wrangling step. In this report, we will mention them in detail.

1) Gathering Data

For the project, I gathered data from three different sources and on three different formats.

- The first source was the data that contained WeRateDogs twitter archive, and it was provided by Udacity. So I directly downloaded it from the project page.
- The second source was Udacity servers. I downloaded the image prediction data programmatically using the Request library.
- Third method was to use Twitter API and pythons tweepy library to gather data. Since I couldn't manage to get a developer account, I chose to download file directly from Udacity. This data contained favorite and retweet counts for the tweets twitter archive had.

2) Assessing Data

After gathering the data, I moved on with the assessing step. I used visual and programmatic methods for assessing. Listed below are some quality and tidiness issues related to datasets. Quality issues are related to content, while tidiness is about datasets structure.

Quality issues

1. Retweets and replies should be removed from archive_df since we are interested in the ratings data only.
2. Columns related to retweets and replies should be removed later since they are not our interest.
3. On archive_df timestamp should be a datetime format.
4. On some dog names, single letters like 'a' used instead of null also some of them don't make sense.
5. We only need the most accurate prediction on image data, others are unnecessary for our analysis.
6. Dog breed predictions should be on a standart. Some of them start with lowercase and some of them contain '_'.
7. Both NaN and None values are used for the same reference on datasets.

8. On archive_df source columns should be replaced as categories.

Tidiness issues

1. On the archive_df (twitter_archive) dataset doggo, floofer, pupper, puppo should be on the same column as type.
2. All three datasets should be merged.

3) Cleaning Data

After the problems were determined, I continued with the cleaning step to get a master, tidy and quality dataset.

First I started with fixing the tidiness issues. On twitter archive data, instead of dog types being on different columns, I put them in a column called dog_type and removed the other four columns since they were no longer needed. Also I merged the three dataset I gathered on tweet_id, since they were all about tweets and it made more sense for all of them to be on the same dataset.

Then I continued with the quality issues. Archive data contained not just original tweets but replies and retweets too. Since our project motive was dog_rates I eliminated them and left original tweets only. Then removed columns related to retweets and replies too since they were mostly null.
I fixed the timestamp format so I can use it as an accurate date for my analysis.
Some dog names contain random letters like ‘a’ that don't make sense. I replaced them with None. Also I standardized the dog breeds with some replacing.
On prediction data, I only needed the true and accurate prediction for my analysis so I made a column that contains the accurate prediction only and removed the unnecessary ones.

As the last step, I stored the new clean master data in a new csv file as our new dataset.