Wrangle Report

1. Introduction

This paper will describe the efforts which made for the Udacity Data Analyst Nanodegree Program of project “WeRateDogs”

The report will have to following structure:

- Gathering Data
- Assessing Data
- Cleaning Data

2. Gathering Data

The data for this project came from three different sources:

- Original twitter archive data: downloaded from the udacity project details site and uploaded into the udacity project workspace
- Predictions data: programmatically downloaded from the udacity server
- Twitter data: obtained from the Twitter API using Tweepy

3. Assessing

After gathering the data from the individual sources, the next step was to assess this data visually and programmatically. Following quality and tidiness issues were detected:

Archive file

The archive column “expanded_urls” had empty rows without links. There are also a lot of duplicated twitter links and links of other sources.

The “name” column of the archive table isn’t always filled. A lot of those names are stored in the “text” column.

The “doggo”, “floofer”, “pupper” and “puppo” column stores the same data. These columns should be merged into one single column.

The “text” column includes more than just the photo description. It also includes the rating, dog names, IG names and short URLS.

Predictions:

Some of the names are lower case.

4. Cleaning

Cleaning the data is the third step in data wrangling. Following quality and tidiness issues were cleaned:

Quality

1. Remove short URL from archive table "text" column and move it to a new column called "url_short"

2. Missing names in archive table name column

3. Values for rating_numerator and rating_denominator are sometimes incorrect

4. Replace the value 'None' with NaN (missing value)

5. Remove replies and retweets from in the archive table and non-matching ID´s from prediction table.

6. Drop unnecessary columns from archive table

7. Expanded URL column got rows with data duplicates or other website sources

8. Drop columns without pictures

9. Change timestamp to datetime

Tidiness

1. Merge the four dog type columns "puppo", "pupper", "floofer" and "doggo" into one column and change the data type to categorical

2. Merge archive table with api_data table

3. Multiple columns store the same type of data in the predictions table.

4. Merge tables into archive_clean data frame

5. Conclusion

Since there is a lot of unclean data around the world. Data wrangling is a skill every data analyst should be familiar with. After the gathering, assessing and cleaning part of the data, there is one last step to come. The results need to be analysed and visualized to create better insights about the data.