Reporting: wragle_report

By: Saeed Ashri

Introduction

This paper illustrates the steps of Data Wrangling of Twitter account (@WeRateDogs). This account focuses on rating people's dogs. The steps of wrangling are:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

In this part, the steps of gathering data for this project are dealing with three pieces of data.

Firstly, the Twitter archive is downloaded manually by clicking this link: twitter_archive_enhanced.csv

Secondly, the tweet image predictions data has the predictions of breeds of dogs. This file is downloaded by using the library of requests in python language. It requested this link:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

The last file has got by getting access to the API of Twitter. Using the tweepy library helps to get tweet and save it in t(weet_json.txt). Then, the JSON library is used to get specific data from JSON objects and put it in the data frame.

Assessing data

This section demonstrates three dataset's assessments using visual and programmatic assessment. The assessing detects some of quality and tidiness issue.

Quality issue

Twitter Archive Data

1. Only using original tweets.

2. Erroneous datatypes of timestamp column and separate date and time.

3. Getting the type of source that has been created a tweet from the source column.

4. rating_denominator column has wrong values.

5. Outliers in rating_numerator.

6. names of dogs are not correct such as (a, an). Also, the None values transform to NaN.

Image Prediction Data

7. jpg_url column has 66 duplications.

8. dog breeds in the columns (p1, p2, p3) are not lower or uppercase letters and use underscore rather than space.

9. type of dog and confidence rate in from columns (p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog)

Tweet JSON Data

- We don't have any issue with data quality because we get specific columns that are used in the analysis

Tidiness issue

10. (doggo, floofer, pupper and puppo) columns are (transform to one column).

11. Delete columns that are not used in the analysis.

12. We've three datasets that need to merge into one dataset by using tweet id.

Cleaning data

The third step in data wrangling is cleaning. All issues in the previse part are repaired as clean data. First, we are getting a copy of each dataset. Then, define each issue and write code then test the data. Then, we merge all data into one dataset to be ready for the next stage which is analysis.

Conclusion

Data Wrangling is an essential skill for people who deal with data. We did all stages of data wrangling gathering, assessing, and cleaning for all datasets that we have. After that, the data is ready for analysis.