Udacity - Data Analyst Nanodegree - Wrangle and Analyze Data

27th Mar. 2020

Wrangle Report

1. Introduction

Real-world data rarely comes clean. Using Python and its libraries, we gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called "Data Wrangling". We will describe our wrangling efforts in this paper.

The dataset that we wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs has over 4 million followers and has received international media coverage.

2. Gathering

"Data Wrangling" consists of following three steps, "Gathering", "Assessing" and "Cleaning". And all these step are iterative. Fist of all, we gathered data for this project.

- Enhanced Twitter Archive : WeRateDogs Twitter archive contains basic tweet
- Image Predictions File : Results neural network classified WeRateDogs Tweet images (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)
- Additional Data via the Twitter API : Raw data of WeRateDogs(tweet_json.txt)

3. Assessing

We identified data issues of "Quality" and "Tidiness".

3.1. Quality

Quality issues (completeness, validity, accuracy and consistency) were following.

3.1.1. Twitter archive

(1) Datatype of timestamp was incorrect.

(2) Some URLs were duplicated.

(3) Rating Denominator should be always 10.

(4) Retweets were included but not needed.

(5) Replies were included but not needed.

(6) Source should not include some unnecessary characters.

3.1.2.Image Predictions

(1) Some URLs were duplicated.

3.1.3.Twitter additional data

(1) Datatype of timestamp was incorrect.

3.2.Tidiness

Tidiness issues (structural issues) were following.

3.2.1.Twitter archive

(1) Dog type (doggo, floofer, pupper and puppo) should be in one column.

3.2.2.Image Predictions

(1) Predicted result should be in one column.

3.2.3.Twitter additional data

(1) There were unnecessary columns.

4. Cleaning

We fixed all of the data issues. First of all, we copied data for the purpose to test whether the issues were fixed after cleaning. This process consists of Define, Code and Test. "Define" means to put solutions for the issues into words. And “Code” means to write code to implement the solutions. “Test“ means to test whether the issues were fixed.

5. Conclusion

Finally, we made all of the issues fixed and got cleaned data. “Data Wrangling” is one of most important step in data analyze. Because low quality and untidy data never give us excellent insights.