Wrangle Report

Udacity DAND - Wrangle and Analyze Data Project

By Jiwoon Kim

■ Introduction

In this project, I wrangled the dataset of the Twitter, WeRateDogs.

This twitter rates dogs with a comment in humorous manner.

This report briefly describes my wrangling efforts from the course learning including :

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

Let's me describe the wrangling process and result step by step.

■ Gathering Data

In this project, three of dataset is provided to me.

1. Twitter archive

I could download twitter_archive_enhanced.csv file from the link provided to me.

Using pandas of python, I imported this csv file into dataframe.

This file has a variety of basic information tweets such as tweet id, comment text, rating value, timestamp and so on.

2. Tweet image predictions

This file (image_predictions.tsv) was hosted on the server and I could download programmatically on python, using 'Request' library. I imported into dataframe.

It has the information that what is the breed of the animals in the picture.

3. Twitter API and JSON

I accessed the each tweet's JSON data using tweet ID in archive data with 'Tweepy' library.

I stored entire tweet's JOSN data into tweet_json.txt file and imported into pandas dataframe.

There are some additional data such as retweet count, favorite count.

■ Assessing data

First, I browsed datasets using some pandas method such as head, info and I saw entire dataset for visual approach.

I found the both quality and tidiness issues from each dataset

1. Twitter Archive

- **Quality Issues**
  - ✓ There are some typo or mis-spelled values in name
  - ✓ Retweet and replies tweet exist in this Dataframe
  - ✓ Datatype of tweet_id is wrong

- **Tidiness Issues**
  - ✓ Source column have html code , not normal string object
  - ✓ Some of extended_URLs columns have multiple URL
  - ✓ Timestamp should be split into data and time
  - ✓ Text column contains url information
  - ✓ Stage column should newly created after combining split columns (doggo,pupper...)

2. Image Prediction

- **Quality Issues**
  - ✓ Datatype of tweet_id is wrong
  - ✓ p1,p2,p3 should be capitalized for consistency
  - ✓ Some tweets have same URL of JPG FILE

3. Twitter JSON

- **Quality Issues**
  - ✓ Retweets should be deleted

■ Cleaning data

Those issues I founded and described in the assess section was be cleared on this cleaning process.

After solving each issues, I merged 3 dataset into one with the columns I really need to use.

■ Conclusion

This project gave me a wide of skill and knowledge about data analysis.

I've learned how to use API, programmatic wrangling skills and various python library and pandas methods.

It is very essential and critical procedure for data analysis, but very hard and long-time process.

I will do my best because It should be perfect for accurate analysis when I do every analysis.