Wrangle Report

David Chi Nam

Udacity We rate dog project

Data wrangling is the most important part of the data analyzing the process. Data always have to be cleaned to convey the right insights to readers and data analyst itself. Through 'We rate dog' project from Udacity, I have masters skills in the very first and important part of data preprocessing steps. I have focused on data wrangling process, which includes data gathering, assessing, and cleaning data.

In the data gathering process, two data files were given, and one data file was created; 1 CSV file, 1 TSV file, and 1 created CSV file. I had to extract information from tweet URL to create a tweeter data frame which will be needed for the analyzing process. For the gathering process "tweepy" and "pandas" library were used to extract and import data on python.

For assessing data process, data quality and tidiness issue were recorded. 3 files were imported in the data gathering process; arch_df, ima_df, tweet_df. In these three data frames, a lot of different issues were found, such as wrong data format, missing values, duplicated values, and outliers. However, since tweet_df were cleaned while extracting data from a given URL, I could not find any problems.

Data cleaning process were the step that had to put most of the effort into the project. In the data cleaning process, I have defined the problem, cleaned with python code, and tested them. At the beginning of the process instead of cleaning original dataset, the copy of data frames was imported, in the way of preventing transforming the whole dataset. The issues that were found on the assessing

process were cleaned one by one and tested them until I find no problems in data. After cleaning all the data, the three data frames were merged into one data frame.

Thus all three datasets are cleaned; the data will be analyzed and visualized for insights. The data will be re-cleaned if any problems are found during analyzing.