DATA WRANGLING REPORT

Data wrangling is the process of gathering data, assessing its quality and structure and cleaning it before performing any analysis, visualization or build predictive models. In this analysis, I explored all the three stages of data wrangling; Data Gathering, Assessing Data and Cleaning data.

Data Gathering

This analysis made use of three datasets. An already sitting data (tweet archive data) provided in the Udacity classroom, a test delimiter (tsv) file (tweet images) was provided which was extracted from the website with the help of the requests library in python. The third dataset, data that was scrapped from twitter which contains the favorite likes and retweets of dogs. This data was in a json file which was read into Jupyter notebook.

Assessing Data and Cleaning Data

The two techniques of data assessment were used to evaluate the data: the data quality issue, which deals with the contents of the data, such as missing, inaccurate, inconsistent, duplicate, etc., and the tidiness issue, which deals with the data's structure. I employed both programmatic and visual techniques to evaluate the quality and tidiness issues.

After getting a concise summary of the data, it was detected that some of the columns were not in the correct datatype format. The timestamp column was in a string (object) format meanwhile it was supposed to be a datetime. Also, the ratings column was integers which was supposed to be floats. In cleaning the data, I changed these datatypes from object and integers to datetime and floats datatype respectively.

There was also membership constraints issue in the name column in the twitter archive data. Some of the dog’s name was entered wrongly. This was dealt with by using the mask function. In the end, I only used the correct dog names (dog names that are capitalized) in the analysis.

There were also missing data in some of the tweet archive data. Columns that had over 80% of missing values was dropped from the data since the percentage of missing values was too high to be replaced.

The tweet source column in the tweet archive data also contained html tags and text we wouldn’t need in the analysis. These tags were removed so that we keep only the source of tweets only in that column.

The twitter images data had some duplicated URL. All duplicates were taken off.

Also, the text column in the twitter archive data contained URL links, we would not need per the analysis. Here I used regular expressions (regex) to take off all the links out of the text to keep only the text.

There were also irrelevant columns (columns that are not of importance to us in the analysis). These columns were as well dropped from the tweet archive data to keep only relevant ones.

The date column had time zones attached we wouldn’t need for our analysis; this was taken off from the date column so we could pick only the dates from the timestamp column. To make sure the structure of the data is correct (tidy data), I combined all the three datasets to form one data using the merge method in pandas. I used the inner join to combine these datasets so as to get only the intersecting rows from these three datasets.

Also, for the tweet archive data, the dog stage, which was supposed to be in one column was stored in 3 different columns each for a particular stage. These 3 columns were brought together as one comprising the different dog stages. Here I used the melt function in pandas.