PROJECT TITLE : PROJECT 2

TITLE :DATA WRANGLING AND ANALYSIS

STUDENT NAME : DENZEL AKWANY

What is data wrangling?

This is the process of cleaning and unifying messy and complex data sets for easy access and analysis. The munging process being the backbone of all data decisions largely includes statistical analysis,aggregation processes,business intelligence and visualisation and advanced analytics such as forecasting[1].

Why?

The data wrangling process is key and mandatory in all data engagement processes.It is agreed among scientists that a huge chunk of time, 80% being spent on data munging process[2]. Data engineers,data analysts as well as data scientists must partake in this process for the following reasons.

- Data wrangling enables data enrichment. This is by joining data from different sources facilitating rich and deep insights.
- The data reliability is improved making it easy for the consumers such as analysts to have actionable and accurate data.
- The process reduces time spent by analysts and scientists in the data cleaning process and therefore has time to focus on insights and decisions.
- Exceptionally wrangled data serves the purpose of quick and easy data driven decision making by key stakeholders in a company or institution.

The wrangling process is a necessity that can be summed into data enrichment, data structuring,data cleaning, validation and discovery .[3]

How?

The data wrangling process is procedural and include the following process:

- Data centralization and acquisition.
- Data combination or merging
- Data cleaning.

Data centralization

The data acquisition was done by sourcing data from different sources. The process involved:

- Loading data , the twitter_archive_enhanced.csv through the pandas module.
- The second data source was downloaded through the use of the requests library. The data,image_predictions.tsv, which was tab separated was transformed into csv.
- Through the use of tweepy library, the query fetched retweets and favourite count. The data was then saved into a data frame.

Data combination

The data from the different sources and formats as described above were harmonised into common file type,csv and merged to form a wide data mart. The wide table is formed by joining the twitter archived data and the tweet_json, with fields of retweet and likes into one joined dataframe.

Data cleaning

The data cleaning process was tedious and required expansive research. The research was conceived through the need to fetch the tweet_json.txt data. The tweepy module which is intuitive to use made it easier to fetch the retweets and likes by using the tweet id. There were missing ids which prompted the use of exceptions.

These exceptions failed at some stage with output errors such as NotFound and forbidden error. Through the expansive research I was able to exclude the errors and fetched the data successfully.

The cleaning process included identification of nulls. The nulls were dropped where the majority was above or equal to 60% relative to the total.

The data types for column were also transformed with timestamp cast to datetime,tweet id casted to object instead of int as this often affect summary statistics and statistical processes that focus on numerical columns.

The cleaning process also focused on the outliers and their possible effect in the data. The data column rating denominator as well as the numerator had outliers. The outliers were further analysed by creation of interquartile ranges, stating of upper cutoff and lower cutoff.

The data happened not to have duplicates

Conclusion

The data wrangling process is longer and not made easier by the data not having a descriptive schema of a number of columns. The lack of data fields knowledge made it difficult in formulating insight oriented questions and in depth understanding.However, the process was successful with the data sourcing and was an excellent learning process.

References

[1]https://www.tweepy.org/

[2]https://www.simplilearn.com/data-wrangling-article

[3]https://www.analyticsvidhya.com/blog/2021/08/lets-understand-all-about-data-wrangling/