WRANGLE REPORT

Introduction

This project uses data wrangling steps learned in Data Analysis Nano degree from Udacity.

He the tweet archive of user @dog_rates is used which accounts the rates given by people to dogs.

This report describes the steps taken by me to wrangle the data

Details

Steps are :

Gathering Data

Analysing Data

Cleaning Data

Gathering Data

Data was gathered from below sources:

The enhanced tweeter archive file (twitter-archive-enhanced.csv).

Data gathered from tweeter archive using tweeter api.

Tweet Image predictions file downloaded from Udacity server using request api.

Analysing Data

The data gathered was viewed using the .head(), .sample(), .info() methods provided by the panda library.

The data had issues like :

- Data contained retweets.
- Name contained "None" instead of NaN
- Name contained inappropriate data such as adjectives and verbs.

- Ratings were non standardized
- Columns present that were not needed.
- Had to create few columns to create a better view for data visualization.
- Combining data from different sources into one to have one true source

Cleaning Data

The issues found during the analysis phase were cleaned using the methods found in pandas library such as:

- Data was merged using the .merge() command to merge all the three data sources.
- .drop() was used to drop any unnecessary columns from the data sources.
- Data for columns were created and appended with command such as .extract()
- .replace() was used to replace data that was not so standardized.
- Many other methods were also used to make the data cleaner

Conclusion

Data can be obtained from many sources, but we can draw certain conclusions only when the data is not ambiguous and has all the possible values. Not all but some of these values and properties can be achieved using proper data wrangling techniques.