wrangle_report

December 12, 2020

1 Wrangle Reporting for this Project

Create a 300-600 word written report called wrangle_report.pdf or wrangle_report.html that briefly describes your wrangling efforts. This is to be framed as an internal document.

For this poject I was tasked to look into three data set that needed to be gathered, merged and cleaned. Two of these report were provided for us while the other were had to attain using an API. This was a Twitter API and I used the Twitter IDs from the twitter_archive-enhanced.csv to looks for the tweets on the APi. There was lot of missing data and even some instances where the API was not able to locate some of the twitter_ids. There were some that had 0 favorit counts which I thought was very odd considering how well liked the WeRateDogs page on twitter is.

During the assessing process i noticed there was a lot to be fixed:

2 Quality issues:

1.Quaity: Drop Column we dont need:

-There were a few columns that were not relevent to my assesmnet such as retweeted_status_id that did not give any valuable insight.

2.Quality : Fix denominator

All of the denominators are suppose to be 10 yet some were variously different numbers.

3.Quality: Viewing and drop outliers

For the rating numerator there were plenty of ratings that went beyond 10 but that was part of the fun on the twitter page so i only tried to remove outliers above 100 because they seemed far too exagerated.

4.Quality Convert P1-P3 to string

felt is was eaier to hand the value if they were string values.

5.Quality: if there is a missing confidence , replace with 0 to make easier to compare to the other two.

There were some confidence numbers that were missing and so to make it easier to read i just turned them into 0

6.QUALITY : Clear all false names and convert to Nan

There was some data that rated possible predictions(3) and choose the best prediction, therefore i deleted the predictions that were likly since i can not speculate and just need best guess on the breed of dog.

7.Quality issue: Some of the name are capatolized and some are not

This could cause fuuture issue if one wanted to catergorize the breeds in a graph

8.Quality issue: Change Timestape to to_datetime

Allows one to analyse the dates more effectivly

9.Quality: Drop nan values in rows since we wont know wha dog they are rating or talking about.

There some rows that did not have any dog predictions therefor we woud not be able to tell which dog was even liked.

3 Tidiness issues:

1.Tidy Issue : Meerge alld dataframes to a master table

-I had to merge all the data based on the twitter_id

2.Tidy Issue , Clear all false name and combine 3 columns to give the bread and confidence only

-has to combine the prediction do we only have the best guess prediction to work with since the other would be irrelevent.

3.Tidy issue: Combine the stage column into one dog stage column

-There were three seprate columns for the stages and I had to combine the to have one column to work with

4.Tidy issue: Drop unneeded columns

more columns needed to be dropped as the table became simpler.

Conclusion: It becomes a huge skill to wrangle and asses data because often times the data that you wrangle will more often then not be very dirty data. It will have the most Tidiness issue along with the most Quality issues.