Report

On

Data Wrangling Steps: Gather, Assess, and Clean

By:Potros Shenoda

Wrangle Report:

WeRate Dogs is a Twitter account that categorizes people's dogs with dog comic comments.

The Purposes of Wrangling twitter data:-

- Gathering Data
- Assessing Data
- Cleaning Data
- Storing and visualizing your wrangled data
- Reporting on the data wrangling efforts

Gathering Data:

- From twitter_archive_enhanced.csv
- From image predictions
- From Twitter API

Assessing Data:

- This data-set includes retweets, which means there is duplicated data (as a result, these columns will be empty

- Some column must be delete

- The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo).

- There is {+0000] should be remove

- [rating_denominator,rating_numerator] in integer type and should transform to float type.

- dog names: some dogs have 'None' as a name, or 'a', or 'an.'

- rating_denominator should be a standard 10, but there are a multitude of other values

- there are duplicated in [jpg_url] must be delete

- Change tweet_id to type int64 to meet other dataset

- All tables should be in one table

Cleaning Data:

Quality:-

1-Have only tweets

2-Delete columns that won't use

3-Dog types issue

4-Remove [+0000] from Timestamps

5-Transform [rating_denominator,rating_numerator] from integer to float

6-Column [name] have [None,a,an] and convert them to [Nan]

7-Ceate column called 'rating'

8- Drop [jpg_url] duplicated

Tidiness:-

1-Change tweet_id to type int64

2-All tables should be in one table

Store:-

Store data to =twitter_archive_master.csv

Visualizing Data:-

1-Most Rated Dog Breed

2-Most Common Names

3-Tweet Count vs. Retweet Count

On

Data Wrangling Steps: Gather, Assess, and Clean