WeRateDogs Tweets

Report on Data Wrangling Steps

Introduction

Data wrangling is the process of cleaning and modeling messy and complex data sets for easy *data analytics* and *machine learning*.

Data wrangling Steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

In this project, we perform data wrangling steps on tweets of the WeRateDogs Twitter account to find insights into data. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Python libraries used for this project:

- Pandas
- NumPy
- Requests
- Tweepy
- Json

Gathering Data

Gathering Data is the first step in data wrangling. For this project, I gather three data sets from three different sources.

The first dataset is given by Udacity. It contains a huge chunk of the data about tweets of the WeRateDogs Twitter account from 2015 to 2017.

The second dataset is programmatically downloaded from the Udacity server that contains the result of the machine learning algorithm performed on the images from the WeRateDogs account.

For the third dataset, I have to do web scraping off Twitter using its Tweepy API.

Assessing Data

Assessing Data is the second step in the data wrangling process. In this step, I assess the data quality and tidiness issues visually and programmatically.

In this step, I documented 10 quality issues and 3 tidiness issues.

Quality Issues

1. Remove all the tweets that have retweets.
2. Remove all the tweets that have a reply.
3. Remove duplicates.
4. The datatype of tweet_id is int64 not str.
5. The datatype of timestamp and created_at is an object, not DateTime.
6. Removing tweets that do not have images.
7. Correcting invalid dog names.
8. Extracting tweet source from source column.
9. Fixing rating denominator.
10. Making all rows equal.

Tidiness Issues

1. Dog stages have values as columns, instead of one column with their values.
2. Prediction algorithm columns should be melted into one column.
3. All three datasets should be connected to one dataset.

Common Python methods used in assessing data:

head()

tail()

sample()

info()

describe()

value_counts()

Various methods of indexing and selecting data

Cleaning Data

Cleaning data is the final step of data wrangling. In this step, we fix the quality and tidiness issues that we identified in the assessment step.

In this step first, we make a copy of each data sets so that all the cleaning operations will be performed on the copied version of the data set so that we can view the original data sets later.

Data cleaning process:

- defining
- coding
- testing

Cleaning Sequences:

1. Addressing Missing Data
2. Cleaning for Tidiness
3. Cleaning for Quality