Data Analysis Professional Nanodegree

Wrangle and Analyze Data Project

Project Overview :

Its real-world data from Twitter user @dog_ratest, also known as WeRateDogs

The user rates the people's dogs with a humorous comment, the task is gathering the data from different sources and clean and analyze it

1 - Gathering Data

I gathered from three different sources :

1 - Enhanced Twitter Archive :

This file was given by udacity and it contains 2356 tweets

2- Additional Data via the Twitter API :

There are data not including in the Enhanced Twitter Archive such as retweet count and favorite count so I used Twitter API to get this information

3- Image Predictions File :

Three algorithms predicted the dog breeds and saved them in the Image Prediction file

2 - Assessing The Data

After examining the three files manually and programmatically I detect some Issues some of them are quality and others are tidiness here the summary :

Tidiness Issues

1- prediction data should be in the same file in the archived Twitter data file

2- columns (tweet_id, retweet_count,favorite_count) in the JSON file should be in the same file with twitter data and prediction data

Quality Issues in twitter archived data

1- there are retweets in the data

2- timestamp not in a datetime format

3- the most name in the "name" column is 'a'

4- there some outliers in (rating_numerator, rating_denominator)

Quality Issues in Twitter prediction data

5- the prediction for the dog breed it's not actually a dog all the time

Quality Issues in JSON file

6- there are some uninformative columns

7- some tweets in the original file doesn't have a match in the JSON file

8- there a min values that don't make sense in (retweet_count,favorite_count)

3- Cleaning The Data:

the tidiness issues in point 1 and 2 solved by :

- Merge all dataframes in one place

the quality issues solved by :

1- drop the uninformative columns

2- remove retweets

3- convert object type to datetime in the timestamp column

4-replace the name 'a' with None value of the dog's name

5-drop the outliers in (rating_numerator, rating_denominator)

6-merge all dataframes in one place

Resources

Project overview in the nano degree