wrangle_report

October 29, 2019

1 Wrangle Report

This report aim is about the data wrangling that was conducted in Data Analyst Nanodegree Project "Wrangle and Analyze Data".

1.1 Project Report

The task is as follows: 1. Gathering Data 2. Assessing Data 3. Cleaning Data

1.2 1.Gathering Data

The Data we'll be wrangling is tweet archive of Twitter user @dog_rates (https://twitter.com/dog_rates).

1.2.1 this project uses three different daset obtained as following:

- **Twitter Archive file**: this file was provided by udacity and can be downloaded manually as. ()
- **Tweet Image Predictions**: this file is downloaded programatically using pythons request library adn saved locally to image_prediction.tsv file
- **Twitter API or JSON**: this file is also provided by udacity. this consists of JSON objects in file name called tweet_json.txt file.

1.3 2. Assessing Data

I used pandas info method to spot datatypes and other quality issue. i used code such as head(),tail(), describe(), shape() and info() methods.

here is a list of issue that have been discovered:

Quality Issue - Data contained retweets - Tweet_id,Timestamp was of incorrect data types - Undesired columns were present - Name contained inaccuracies

Tidiness issues - combining all dataframes together as they contained info about same tweets

1.4 3.Cleaning Data

The data have been cleaned by programmatic method. following methods and techniques were used: - merge() - drop() - islower() - regular expressions - replace - extract()

1.5 4. Storing Data

After the completion of cleaning Data. I stored the dataframe in twitter_archive_master.csv file

1.6 Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with.

this project concludes that we will have to use pyhton and its various libraries to scrape data from various sources in various formats and clean various quality and tidiness issues, before any data analysis can be performed.

```
In [ ]:
```