READ ME Data Version 0.1

Content

Introduction

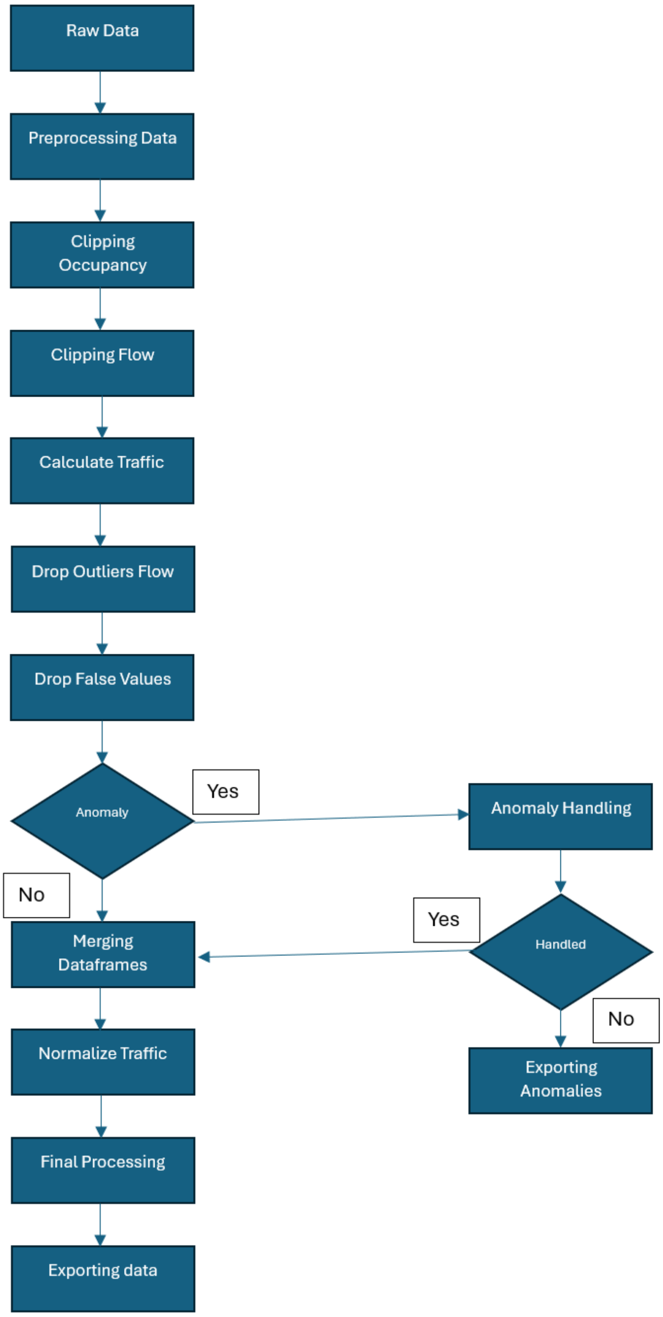
This is the first version of our dataset that will be used for training and testing. We start now with version control on the dataset. So that we can better reproduce changes in the model and know what has changed and when.

Release

Samuel Paul 02.11.2024 14:00

UML

UML Data Engineering



Parameters

In the create_dataset.py script there are a lot of parameters that can be changed. Does are the ones that got set for this version.

Data Cleaning

Clipping occupancy outlier factor: 3

Clipping flow outlier factor: 3

Dropping traffic outlier factor: 2

Dropping false values factor: 3

Anomaly Detection

Mean out of bound factor: 3

IQR to small, min IQR: 5

Not enough data, min datapoints: 4000

Anomaly Handling

The anomaly handling has now only one function. There will be more in the future.

Detectors with bad days

Minimum datapoints per day: 230

Minimum good days: 14

Minimum one good day for every weekday

Script output

```
Starting script
Loading data from:  C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence
Projekt\Data\London_UTD19.csv
Loading data from:  C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence
Projekt\Data\London_detectors.csv
Data loaded
Preprocessing data
Preprocessing data took 10 seconds
Clipping outliers on occ
Total outliers clipped: 834563
Clipping outliers on occ took 400 seconds
Clipping outliers on flow
Total outliers clipped: 168287
Clipping outliers on flow took 406 seconds
Calculating traffic
Calculating traffic took 1 seconds
Droping outliers on traffic
Total outliers dropped: 1188629
Droping outliers on traffic took 33 seconds
Drop false values
Total outliers detected and removed: 342000
Total outliers detected and removed: 1970722
Drop false values took 49 seconds
Detecting anomalies
Anomalies detected based on IQR: 78
Anomalies detected based on IQR too small: 662
Anomalies detected based on not enough data: 2867
Detecting anomalies took 97 seconds
Handling anomalies
Anomalies with not enough data handled: 35
Total amount of dropeed anomalies: 2942
```

```
Handling anomalies took 820 seconds
Exporting anomalies to:  C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence Projekt\Data
Exporting anomalies took 0 seconds
Merging dataframes
Merging dataframes took 2 seconds
Normalizing traffic
Normalizing traffic took 0 seconds
Final processing
Final processing took 1 seconds
Exporting modified dataset to:  C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence Projekt\Data
Exporting modified dataset took 73 seconds
Script finished
Total script execution time: 1913 seconds
```