READ ME Data Version 0.4

Content

Introduction

This is the fourth version of the dataset. This improves the normalization of the traffic value

Release

Samuel Paul 13.11.2024 11:00

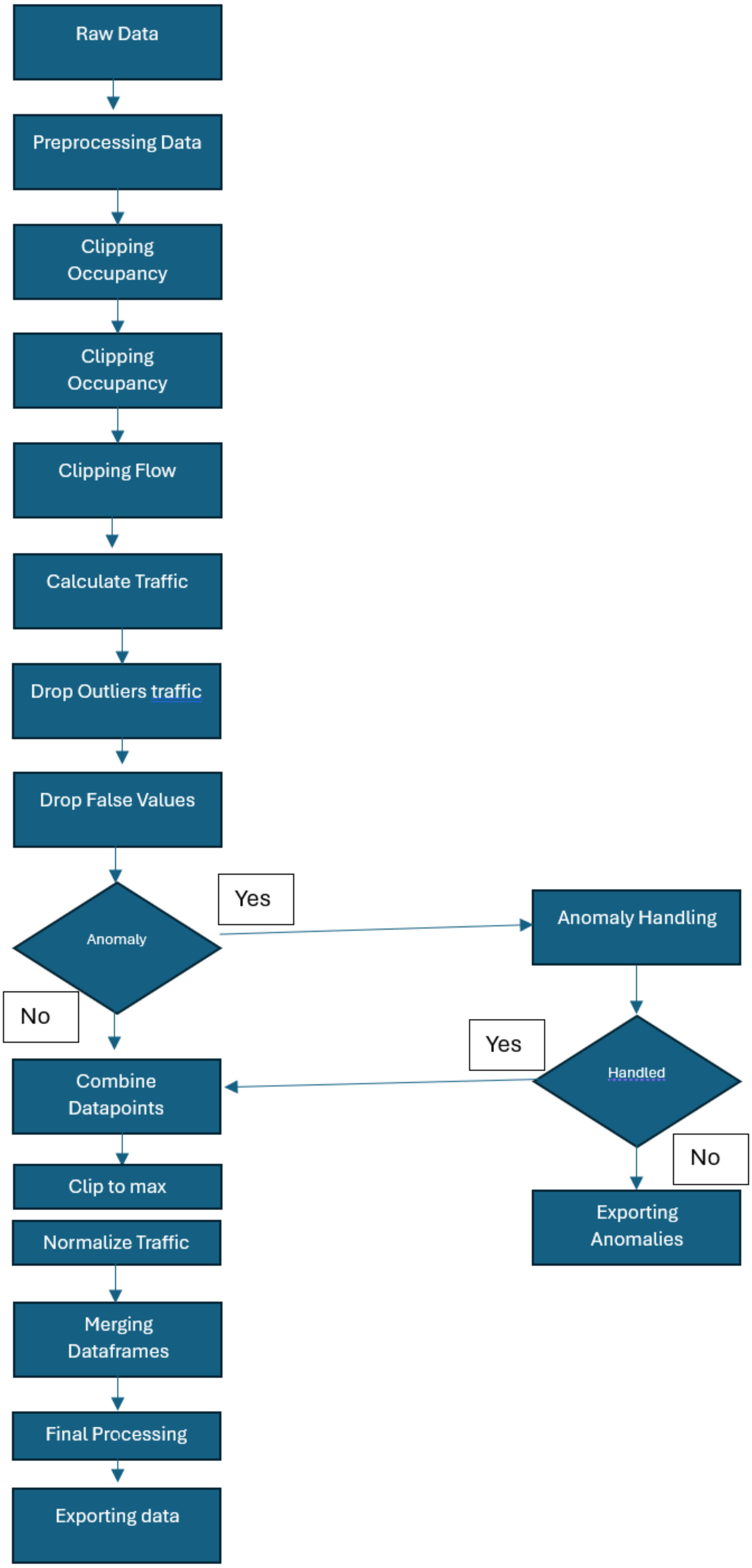
Changes

There are changes to the data data normalization.

Data Normalization

- New function that clips all values at the set max value
    - Currently set at 200

UML



Raw Data
Preprocessing Data
Clipping Occupancy
Clipping Occupancy
Clipping Flow
Calculate Traffic
Drop Outliers traffic
Drop False Values
Anomaly
Yes
No
Anomaly Handling
Handled
Yes
No
Exporting Anomalies
Combine Datapoints
Clip to max
Normalize Traffic
Merging Dataframes
Final Processing
Exporting data

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

Clipping to max value: 200

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: 10

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\London_UTD19.csv

Loading data from:  C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence
Projekt\Data\London\London_detectors.csv

Data loaded

Preprocessing data

Errors found and dropped: 12234005

Preprocessing data took 13 seconds

Drop bad days

Total outliers detected and removed: 9504

Drop bad days took 193 seconds

Clipping outliers on occ

Total outliers clipped: 81

Clipping outliers on occ took 216 seconds

Clipping outliers on flow

Total outliers clipped: 57

Clipping outliers on flow took 211 seconds

Calculating traffic

Calculating traffic took 1 seconds

Droping outliers on traffic

Total outliers dropped: 661222

Droping outliers on traffic took 32 seconds

Drop false values

Total outliers detected and removed: 0

Drop false values took 181 seconds

Detecting anomalies

Anomalies detected based on IQR: 69

Anomalies detected based on IQR too small: 139

Anomalies detected based on not enough data: 1783

Detecting anomalies took 12 seconds
```

```
Handling anomalies
Anomalies with not enough data handled: 686
Total amount of dropeed anomalies: 1163
Handling anomalies took 370 seconds
Exporting anomalies to:  C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence Projekt\Data
Exporting anomalies took 0 seconds
Combine datapoints
Combine datapoints took 8 seconds
Clipping to max traffic value
Clipping to max traffic value took 0 seconds
Normalizing traffic
traffic range was between:0.0 and 200.0
Normalizing traffic took 0 seconds
Merging dataframes
Merging dataframes took 1 seconds
Final processing
Final processing took 0 seconds
Exporting modified dataset to:  C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence Projekt\Data
Exporting modified dataset took 16 seconds
Script finished
Total script execution time: 1296 seconds
```