INFO-H423 Data Mining course

Cool train project from SNCB

- ASSONFACK DEUMANI Patrice Miguelle
- KASHAEVA Gulnur
- LECLERCQ Florian
- WORRALL Stephen

Agenda

- Business context
- Dataset overview
- Preprocessing steps
- Anomalies report
- Models
- Dashboard

Business context

The National Railway Company of Belgium SNCB is responsible for organizing and operating the rail service in Belgium.

The purpose of the excercise is to develop a robust method to report anomalies for the cooling system of the train.

Dataset overview

Dataset

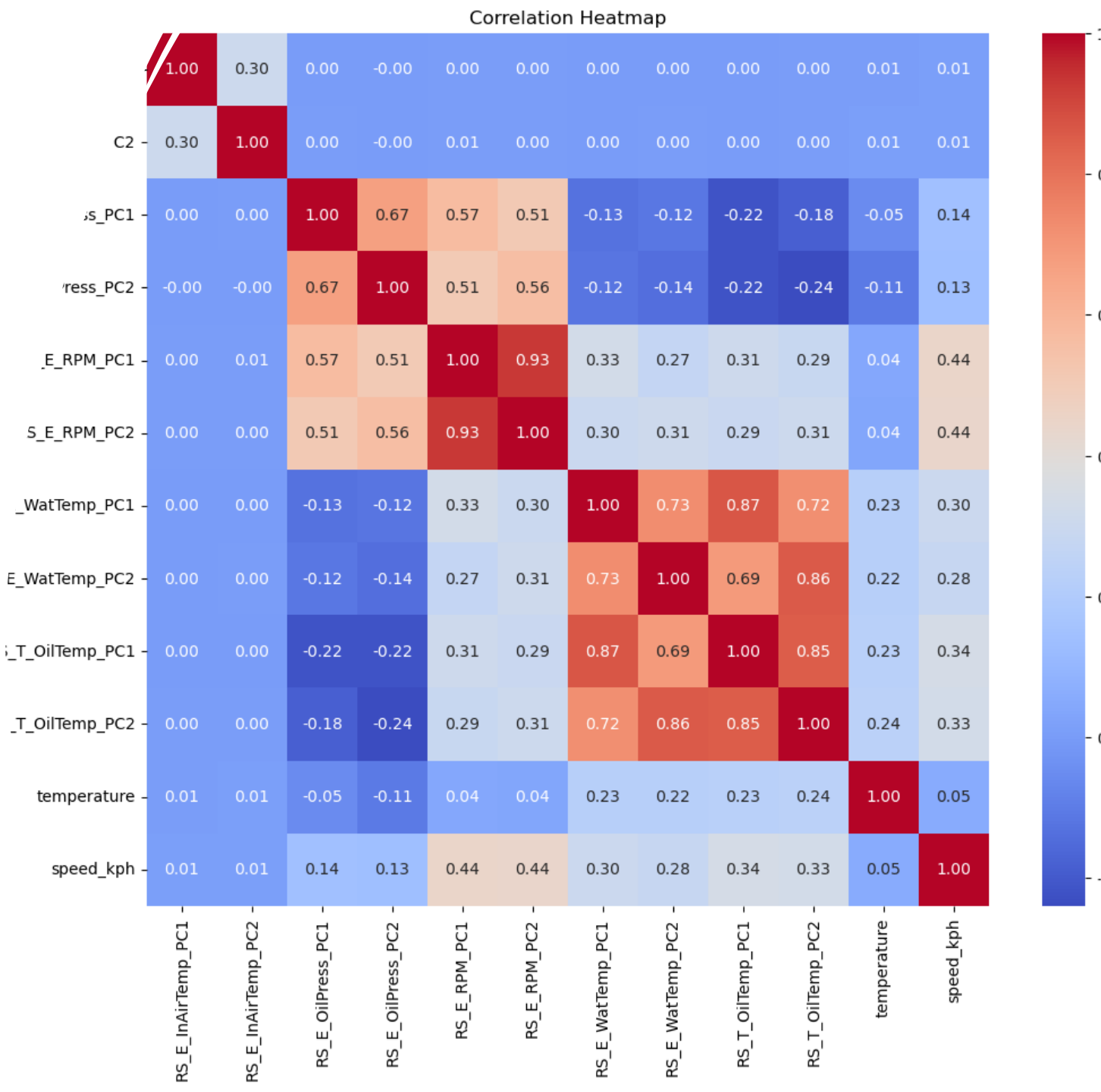
- 14 columns in the orignal dataset.
- 13 numerical and 1 datastemp column.
- One unique ID per train.

	mapped_veh_id	timestamps_UTC	lat	lon	RS_E_InAirTemp_PC1	RS_E_InAirTemp_PC2	RS_E_OilPress_PC1	RS_E_OilPress_PC2	RS_E_RPM_PC1	RS_E_RPM_PC2	RS_E_WatTemp_PC1	RS_E_WatTemp_PC2	RS_T_OilTemp_PC1	RS_T_OilTemp_PC2
0	181	2023-08-01 3:44:12	50.7698183	3.8721144	27	23	255	238	794	801	83	81	76	77
1	143	2023-08-01 6:36:29	51.0399934	3.6934285	33	32	272	324	802	804	78	78	73	74
2	183	2023-08-24 6:53:54	50.7422026	3.6020347	31	33	234	182	799	802	82	82	85	87
3	177	2023-08-01 13:53:38	50.9309143	5.3271318	35	38	220	244	794	801	77	81	78	82
4	143	2023-08-24 7:02:30	51.1807725	3.5752586	41	34	227	282	806	800	85	78	82	79

Variables	N	Mean (SD)	Missing Values
V1	17679273	8.84*10^6 (5.10*10^6)	0
Latitude	17679273	5.09*10^1(3.13*10^1)	0
Longitude	17679273	4.23*10^0 (5.98*10^1)	0
Air temperature in PC1	17679273	3.20*10^1 (3.28*10^2)	0
Air temperature in PC2	17666547	3.23*10^1 (3.48*10^2)	12726
Oil pressure in PC1	17679273	2.64*10^2 (1.15*10^2)	0
Oil pressure in PC2	17666547	2.71*10^2 (1.16*10^2)	12726
RPM PC1	17679273	9.12*10^2 (3.83*10^2)	0
RPM PC2	17666547	9.08*10^2 (3.88*10^2)	12726
Water temperature in PC1	17679273	7.69*10^1 (1.37*10^1)	0
Water temperature in PC2	17666547	7.61*10^1 (1.45*10^1)	12726
Oil temperature in PC1	17679273	7.65*10^1 (1.45*10^1)	0
Oil temperature in PC2	17666547	7.62*10^1 (1.54*10^1)	12726

- Correlation heat map

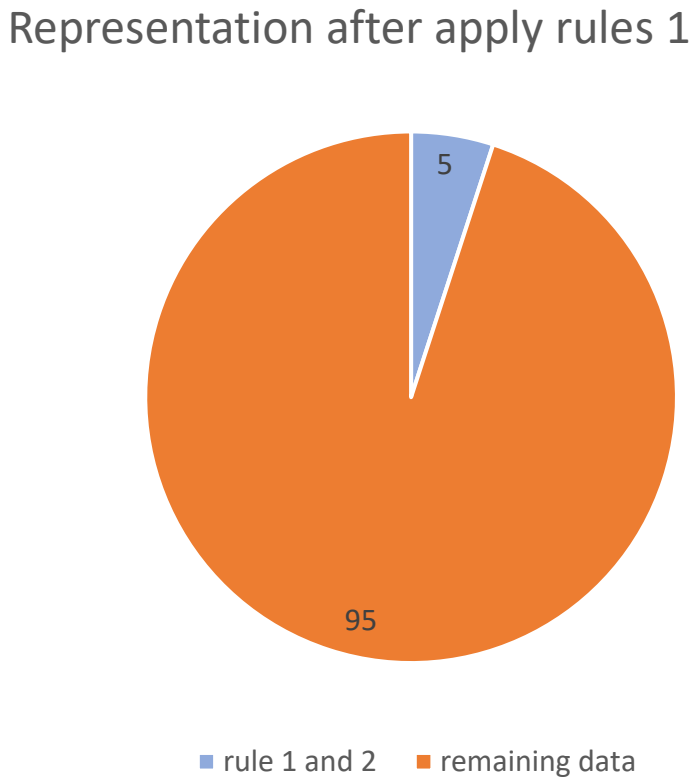
- #RS_E_OilPress_* has an inverse correlation with RS_T_OilTemp_* due to higher viscosity of the oil.
- #RS_E_RPM_PC1 and RS_E_RPM_PC2 show a strong positive correlation, likely because the engines operate in tandem in normal operating conditions.
- Oil and water Temperatures also show a strong positive correlation as they are thermally coupled physical systems.



Preprocessing steps

1. Working with noisy data
2. Enrichment data
3. Derivative columns

1. Working with noisy data

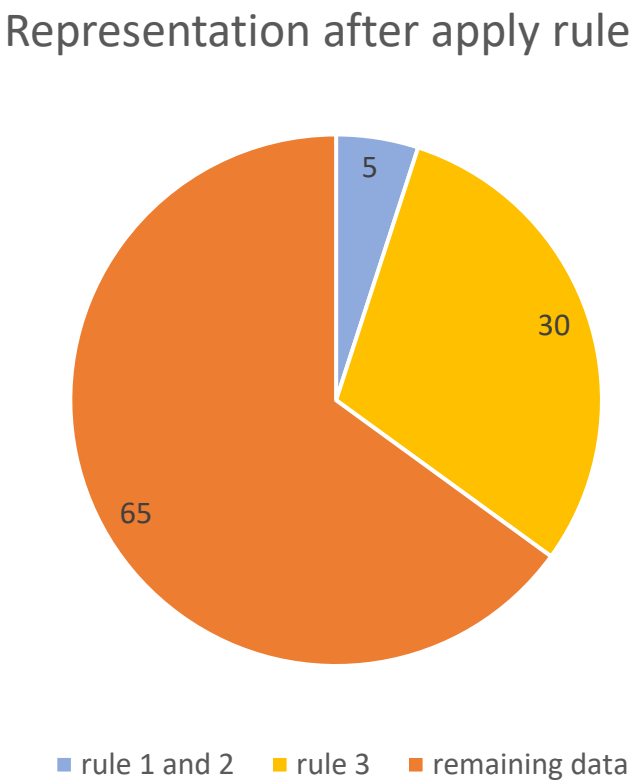


The problem of defining « good » business rules .

1. The 0 value rule.
2. The NaN rule.
3. The time elapsed rule.

Impact visualization - after running Rules 1 and 2 dataset reduced from 17.679.273 to 16.805.603.

1. Working with noisy data



The problem of defining « good » business rules .

1. The 0 value rule.
2. The NaN rule.
3. The time elapsed rule.

Impact visualization - after running Rule 3
dataset reduced from 16.805.603 to 10.369.748

1. Working with noisy data

Processing the Rules: Performance.

1.	The 0 value rule.	Python Pandas, took approx. 1 hour to run.
2.	The NaN rule.	Python Pandas, took approx. 15 mins to run
3.	The time elapsed rule.	RapidMiner, took approx. 4 mins to run

Rapidminer has advantage of using high system memory and native multiprocessing

Python implementation's performance can be increased further by using DASK multi-processing library

2. Enriching the data

- Adding weather data to the dataframe.
- Adding a workstation list to the dataframe.
- Adding elevation data from GPS to the dataframe.

	Row_number	station	round_time	temperature	rel_hum	precipitation	pressure
0	0	06428	2023-08-01 04:00:00	15.3	91.0	0.0	1001.0
1	1	06434	2023-08-01 07:00:00	16.6	87.0	0.0	1001.9
2	2	06428	2023-08-24 07:00:00	19.4	85.0	0.0	1014.3
3	3	06477	2023-08-01 14:00:00	19.4	64.0	0.4	1003.2

Weather dataset.

	ID	Type	City	Place_Name	Company	Speciality	Coordinates	Latitude	Longitude
0	1	Traction	Charleroi	Atelier de traction de Charleroi	SNCB	locomotives & railcars	50.400, 4.458	50.400	4.458
1	2	Traction	Kinkempois	Atelier de traction de Kinkempois	SNCB	locomotives & railcars	50.609, 5.583	50.609	5.583
2	3	Traction	Hasselt	Atelier de traction d'Hasselt	SNCB	locomotives & railcars	50.937, 5.305	50.937	5.305
3	4	Traction	Arlon	Atelier de traction d'Arlon	SNCB	locomotives & railcars	49.678, 5.817	49.678	5.817
4	5	Traction	Anvers	Atelier de traction d'Anvers-Nord	SNCB	locomotives & railcars	51.294, 4.383	51.294	4.383
5	6	Traction	Schaerbeek	Atelier de Traction de Schaerbeek	SNCB	locomotives & railcars	50.878, 4.379	50.878	4.379
6	7	Traction	Ostende	Atelier de traction d'Ostende	SNCB	locomotives & railcars	51.216, 2.946	51.216	2.946
7	8	Central	Cuesmes	Atelier central de Cuesmes	SNCB	passenger equipment	50.446, 3.933	50.446	3.933
8	9	Central	Malines	Atelier central de Malines	SNCB	passenger equipment	51.057, 4.292	51.057	4.292
9	10	Central	Salzinnes	Atelier central de Salzinnes	SNCB	passenger equipment	50.468, 4.844	50.468	4.844
10	11	Polyvalent	Melle	Atelier polyvalent de Melle	SNCB	passenger equipment & wagons	51.013, 3.779	51.013	3.779
11	12	TGV	Forest	Atelier TGV de Forest	SNCB	TGV	50.822, 4.318	50.822	4.318

[Ateliers et postes d'entretien | SNCB (belgiantrain.be)]

Workstation list.

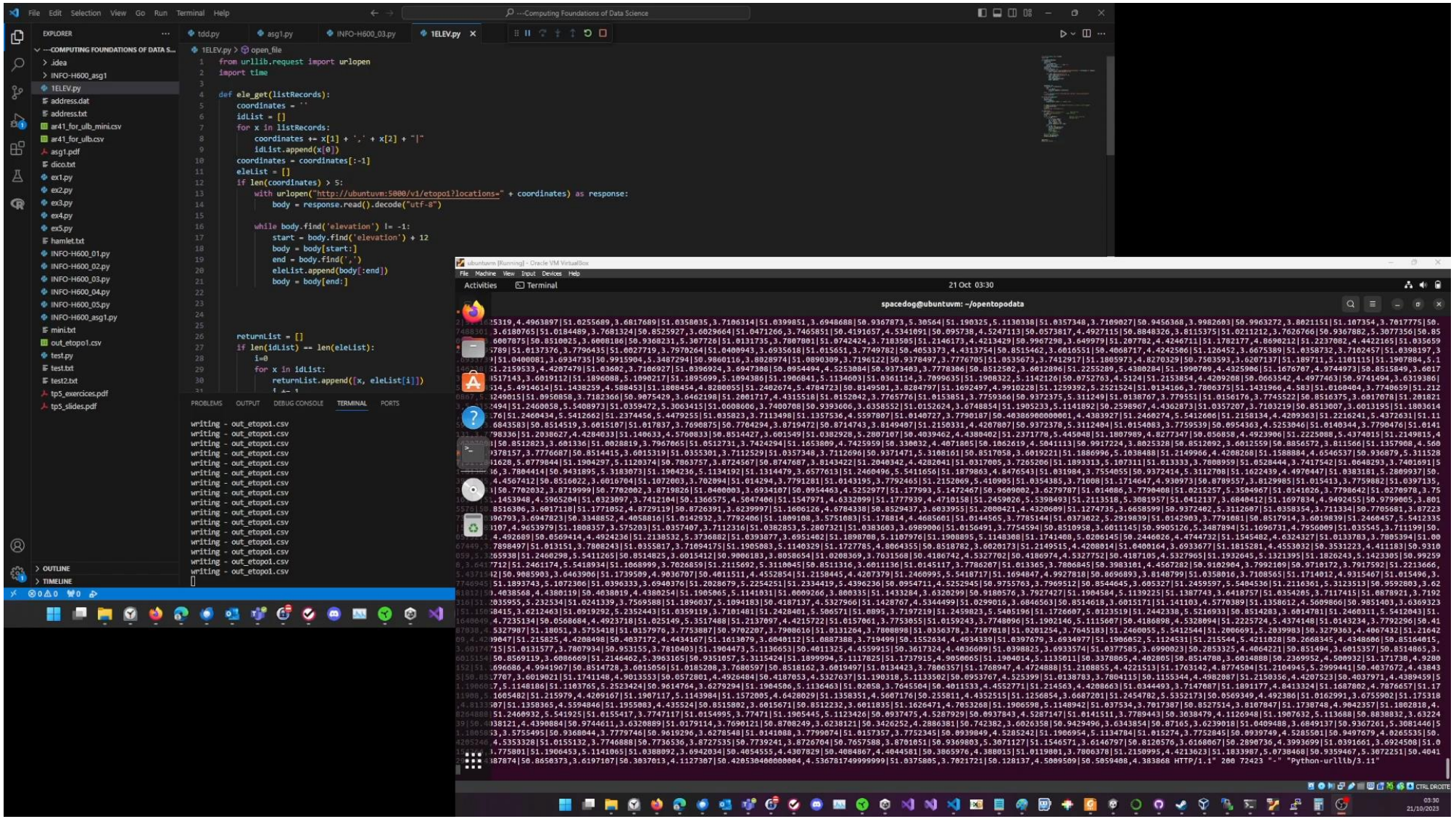
Elevation data.

Local python script →

Calling API over https and parsing json output, populating new CSV.

VirtualBox Ubuntu LTS

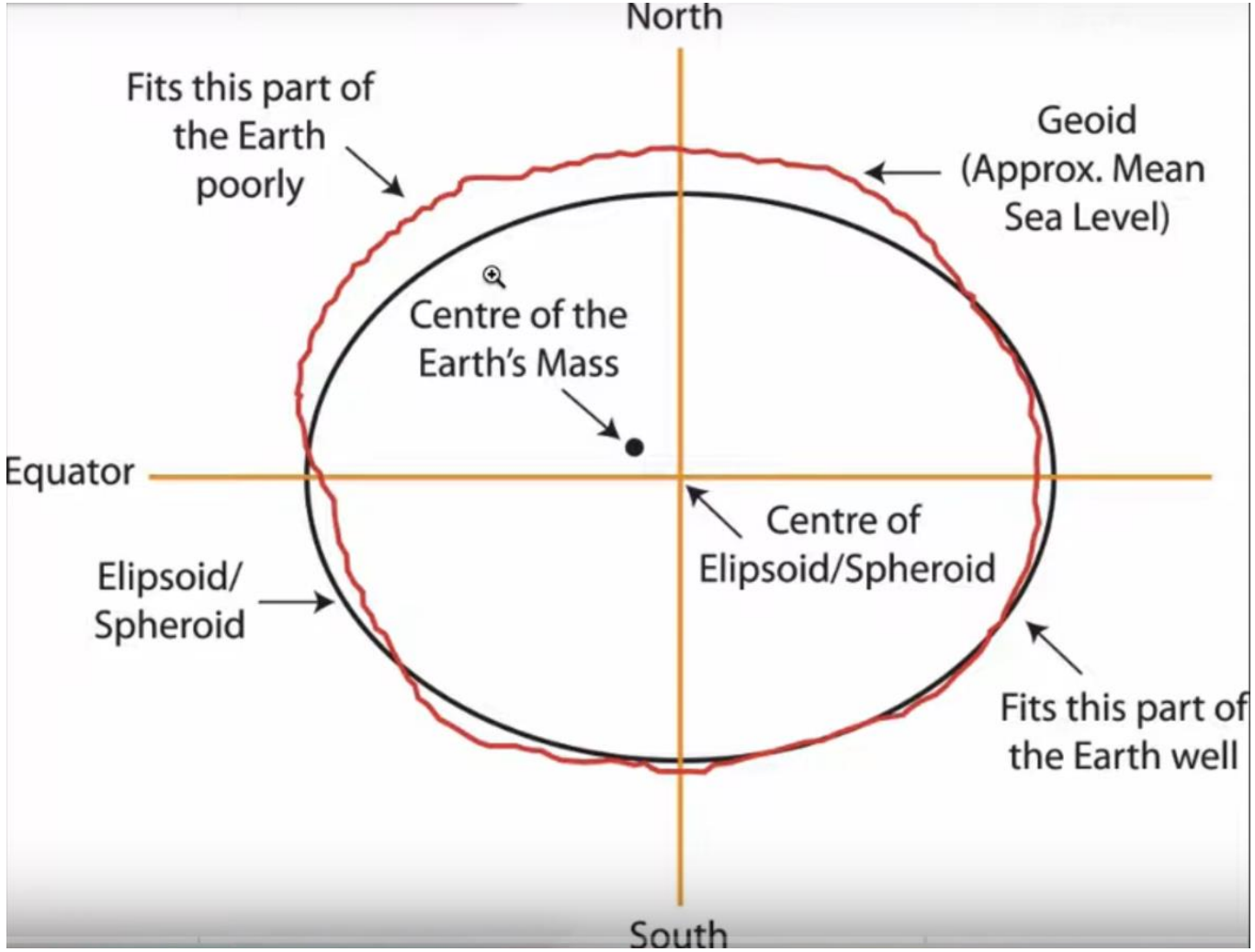
Running docker container of opentopodata API server configured to serve >1000 requests per second →



opentopodata.org server running on a local VM, being called by python to populate a pandas dataframe.

Conversion of Lat/Lon GPS to the (x, y) Cartesian coordinate system

- Not a simple scale conversion
- Approximation shape for earth as an ellipsoid
- CRS (Coordinate Reference System) definitions/standards to localized terrains
- We used 2 step CRS conversion process:
- Lat/lon → EPSG:4326 → EPSG:31370 → (x, y)



A Coordinate Reference Ellipsoid approximation of the earth does not suit all regions equally.

https://epsg.io/31370

3. Derivate columns

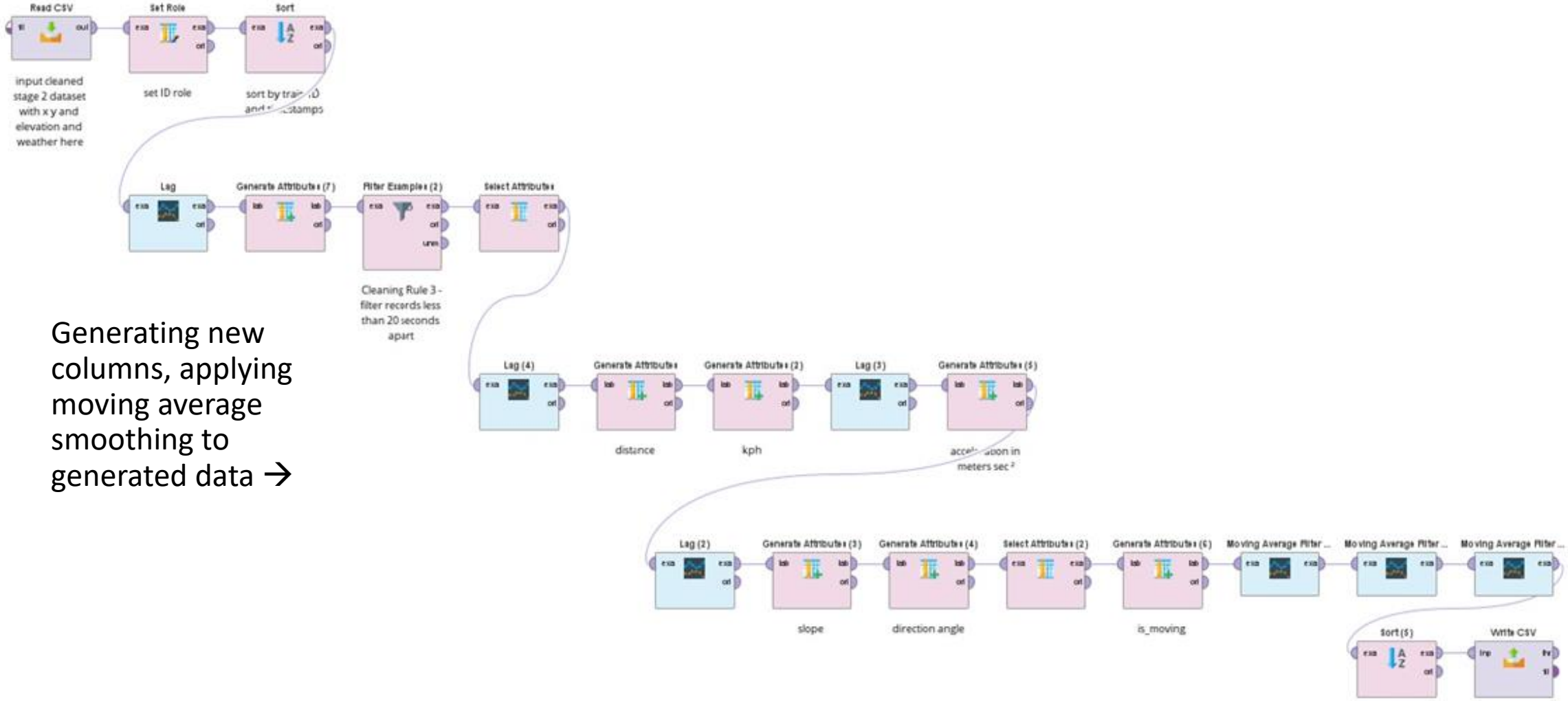
Goal is to extend the columns in the dataframe to add valuable additional information.

For this, we had three objectives :
1. Work with GPS data.

2. Work with exécution time.

3. Gather all the columns into a new, expanded dataset.

3. Derivate columns



Generating new columns, applying moving average smoothing to generated data →

RapidMiner flow

3. Derivate columns

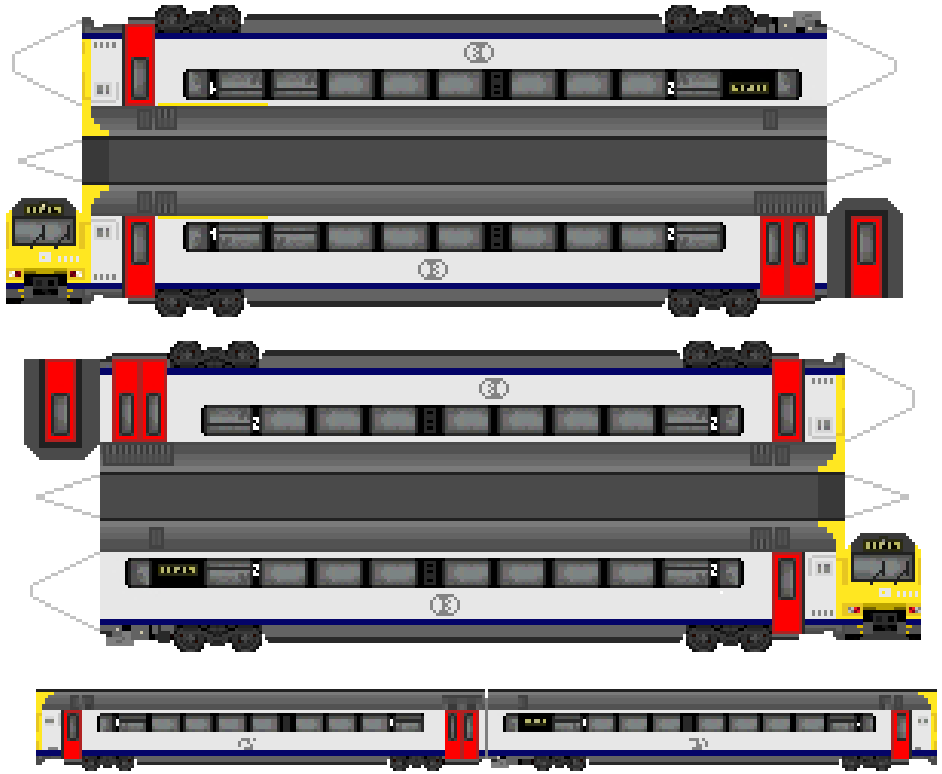
etopo1_elevation	elevation from GPS rounded to nearest 1 meter.
x y	Cartesian coordinates calculated from GPS in meters.
station	nearest weather station.
round_time	time rounded to the hour.
temperature	External temperature.
rel_hum	Humidity.
precipitation	Precipitation.
pressure	Air pressure.
distance	Distance calculated from the previous record for this vehicle.
speed_kph	Speed.
acceleration	Acceleration.
slope	Estimated slope (typically a small number as train tracks are generally designed to be on level terrain).
angle	Direction the train is moving 0-360 degrees where 0:West, 90:South, 180:East, 270:North.
is_moving	It's a Boolean measure to know if the train moving. It's estimated if oil pressure of either engine exceeds 900.
etopo1_elevation_filtered_ma_2_2	Elevation smoothed using moving average 2 before, 2 after.
acceleration_filtered_ma_2_2	Acceleration smoothed using moving average 2 before, 2 after.
slope_filtered_ma_5_5	Slope smoothed using moving average 2 before, 2 after.

Features selection

NMBS AR41 Papercraft

By Tim1995

http://tim1995.deviantart.com/



Features Selection (1)

- **Train Engine Status:**
- ***Speed*** : This feature indicates the current speed of the train.
- ***Prev_speed*** : This feature indicates the previous speed of the train, allowing for the calculation of acceleration.
- ***Acceleration*** : This feature represents the rate of change of the train's speed, which can provide insights into potential mechanical issues.
- ***RS_E_RPM_PC1*** : This feature represents the RPM (revolutions per minute) of engine PC1.
- ***RS_E_RPM_PC2*** : This feature represents the RPM (revolutions per minute) of engine PC2.

- **Train Material Temperature:**
- ***RS_E_InAirTemp_PC1*** : This feature represents the intake air temperature for engine PC1.
- ***RS_E_InAirTemp_PC2*** : This feature represents the intake air temperature for engine PC2.
- ***RS_E_OilPress_PC1*** : This feature represents the oil pressure for engine PC1.
- ***RS_E_OilPress_PC2*** : This feature represents the oil pressure for engine PC2.
- ***RS_E_WatTemp_PC1*** : This feature represents the water temperature for engine PC1.
- ***RS_E_WatTemp_PC2*** : This feature represents the water temperature for engine PC2.
- ***RS_T_OilTemp_PC1*** : This feature represents the oil temperature for engine PC1.
- ***RS_T_OilTemp_PC2*** : This feature represents the oil temperature for engine PC2.

Features Selection (2)

- **Train Material Status :**
- ***Mean_RS_E_RPM_PC1*** : This feature represents the average RPM of engine PC1 over a period of time.
- ***Mean_RS_E_RPM_PC2*** : This feature represents the average RPM of engine PC2 over a period of time.
- ***Combined_Mean_RS_E_RPM*** : This feature represents the combined average RPM of both engines PC1 and PC2 over a period of time.
- ***Mean_RS_E_OilPress_PC1*** : This feature represents the average oil pressure of engine PC1 over a period of time.
- ***Mean_RS_E_OilPress_PC2*** : This feature represents the average oil pressure of engine PC2 over a period of time.
- ***Combined_Mean_RS_E_OilPress*** : This feature represents the combined average oil pressure of both engines PC1 and PC2 over a period of time.
- ***Mean_RS_E_InAirTemp_PC1*** : This feature represents the average intake air temperature for engine PC1 over a period of time.
- ***Mean_RS_E_InAirTemp_PC2*** : This feature represents the average intake air temperature for engine PC2 over a period of time.
- ***Combined_Mean_RS_E_InAirTemp*** : This feature represents the combined average intake air temperature of both engines PC1 and PC2 over a period of time.
- ***Mean_RS_E_WatTemp_PC1*** : This feature represents the average water temperature of engine PC1 over a period of time.
- ***Mean_RS_E_WatTemp_PC2*** : This feature represents the average water temperature of engine PC2 over a period of time.
- ***Combined_Mean_RS_E_WatTemp*** : This feature represents the combined average water temperature of both engines PC1 and PC2 over a period of time.

Features Selection (3)

- **Status Columns:**
- ***Air_Temp_Status_PC1*** : This feature indicates the status of the intake air temperature for engine PC1.
- ***Air_Temp_Status_PC2*** : This feature indicates the status of the intake air temperature for engine PC2.
- ***Water_Temp_Status_PC1*** : This feature indicates the status of the water temperature for engine PC1.
- ***Water_Temp_Status_PC2*** : This feature indicates the status of the water temperature for engine PC2.
- ***Oil_Temp_Status_PC1*** : This feature indicates the status of the oil temperature for engine PC1.
- ***Oil_Temp_Status_PC2*** : This feature indicates the status of the oil temperature for engine PC2.

- **Train Status:**
- ***Repair*** : This feature indicates whether the train is currently undergoing repair.
- ***Breakdown*** : This feature indicates whether the train has experienced a breakdown.

- These features provide a comprehensive overview of the train's mechanical performance and potential maintenance needs. By analyzing these features, you can identify patterns and anomalies that may signal the need for preventative maintenance or further investigation

Anomalies report & model

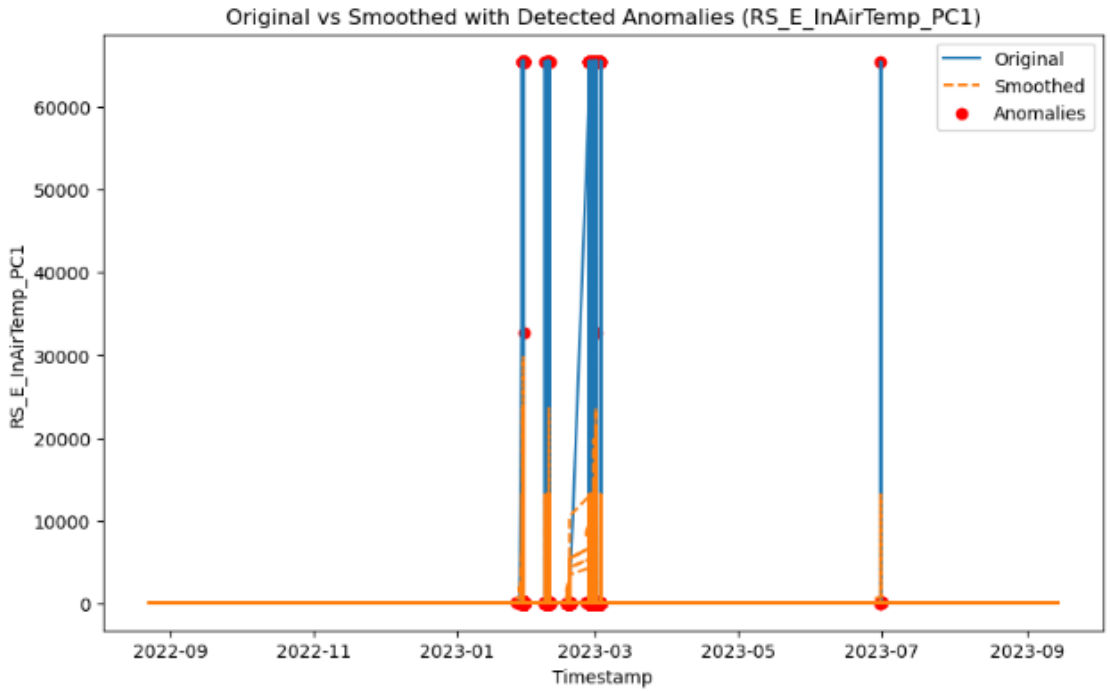
Exponential Smoothing Approach:

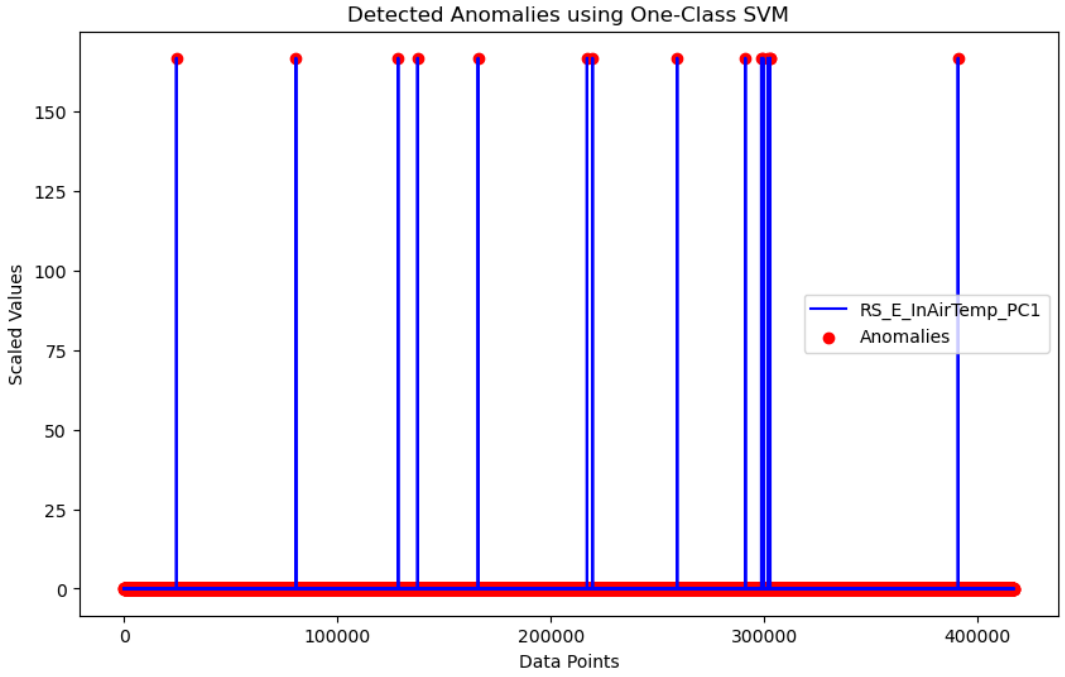
- **Variable Selection:** Choose the target variable (e.g., 'RS_E_InAirTemp_PC1') for anomaly detection.
- **Exponential Smoothing:** Smooth the variable by assigning decreasing weights to past observations.
- **Residual Calculation:** Find the difference between original and smoothed values to get residuals.
- **Statistical Analysis:** Compute mean and standard deviation of residuals for establishing anomaly thresholds.
- **Threshold Setting:** Define anomalies when residual values exceed a set threshold.
- **Anomaly Detection:** Identify potential anomalies by comparing residual values against the threshold.
- **Visualization:** Generate visual representations to observe deviations and anomalies.

One-Class SVM Technic

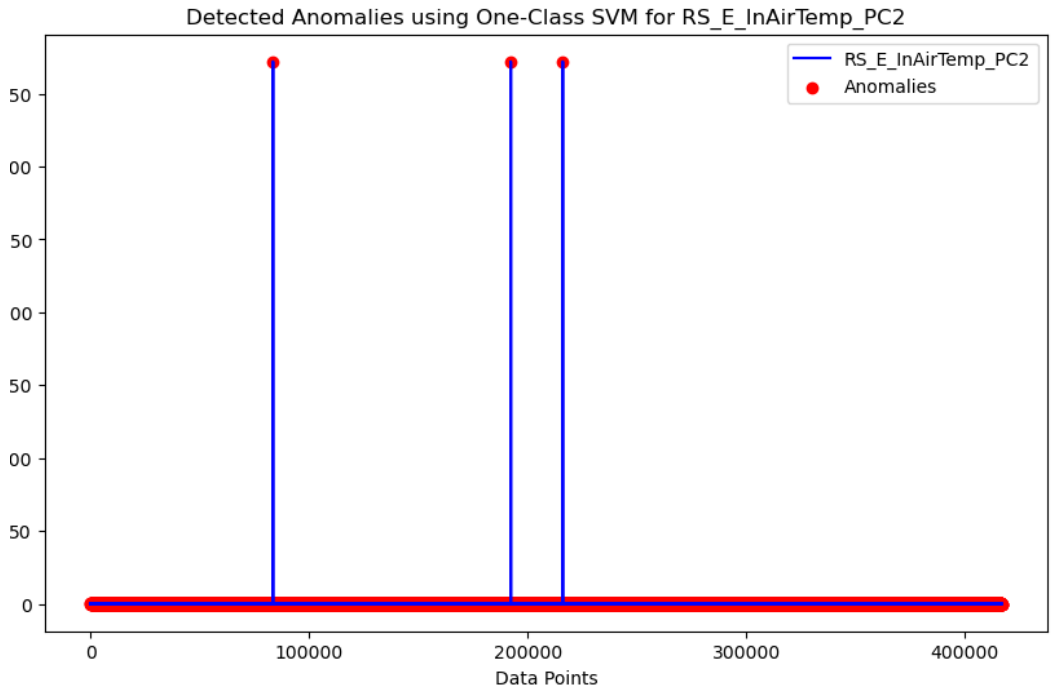
Choose	Column Selection: Choose relevant columns (e.g., temperature, pressure) for anomaly detection.
Extract	Data Extraction: Extract essential columns to form a focused subset for anomaly detection.
Handling	Handling Missing Values: Address missing data using common methods like imputation.
Scale	Standardization (Scaling): Scale the dataset for better model performance.
Train	One-Class SVM Model Creation: Train a model solely on 'normal' data to learn boundaries of normal behavior.
Predict	Anomaly Prediction: Predict anomalies using the trained model, labeling significant deviations as anomalies.
Counting	Anomaly Counting: Count instances labeled as anomalies, estimating abnormal occurrences.

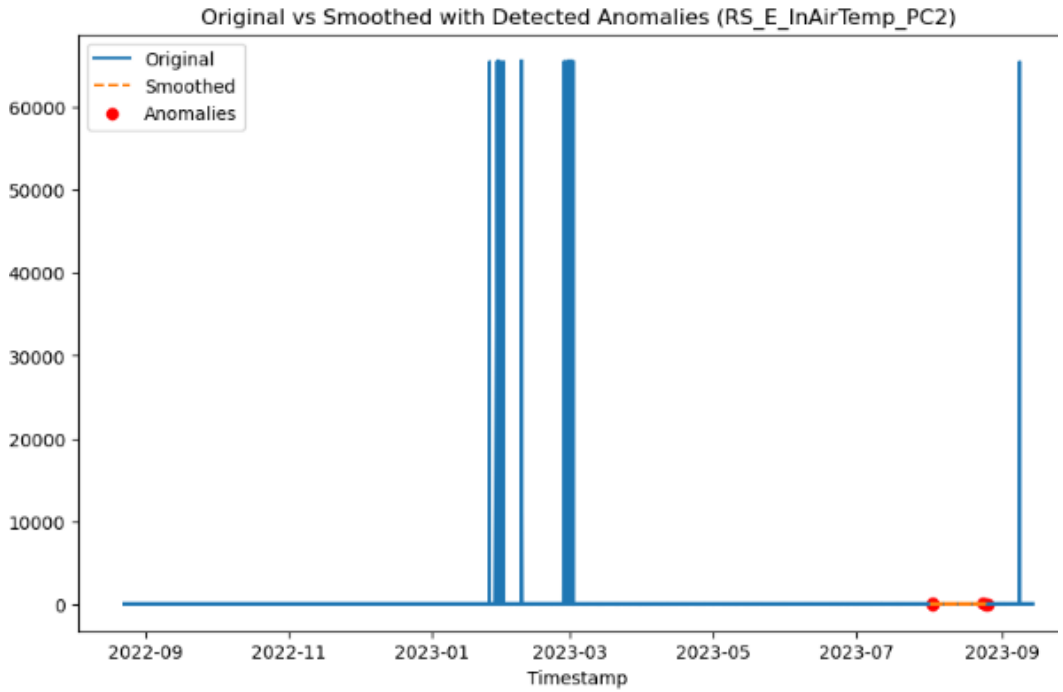
Anomalies Detection model variable 1



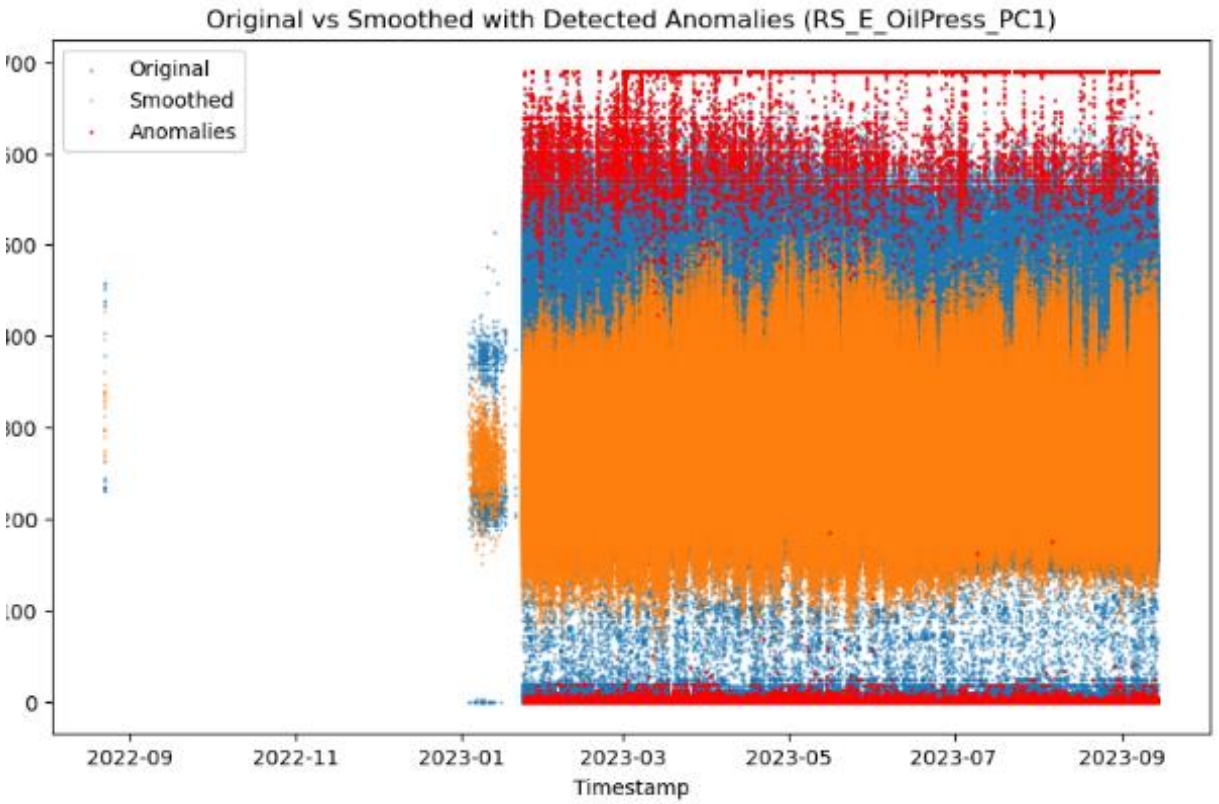


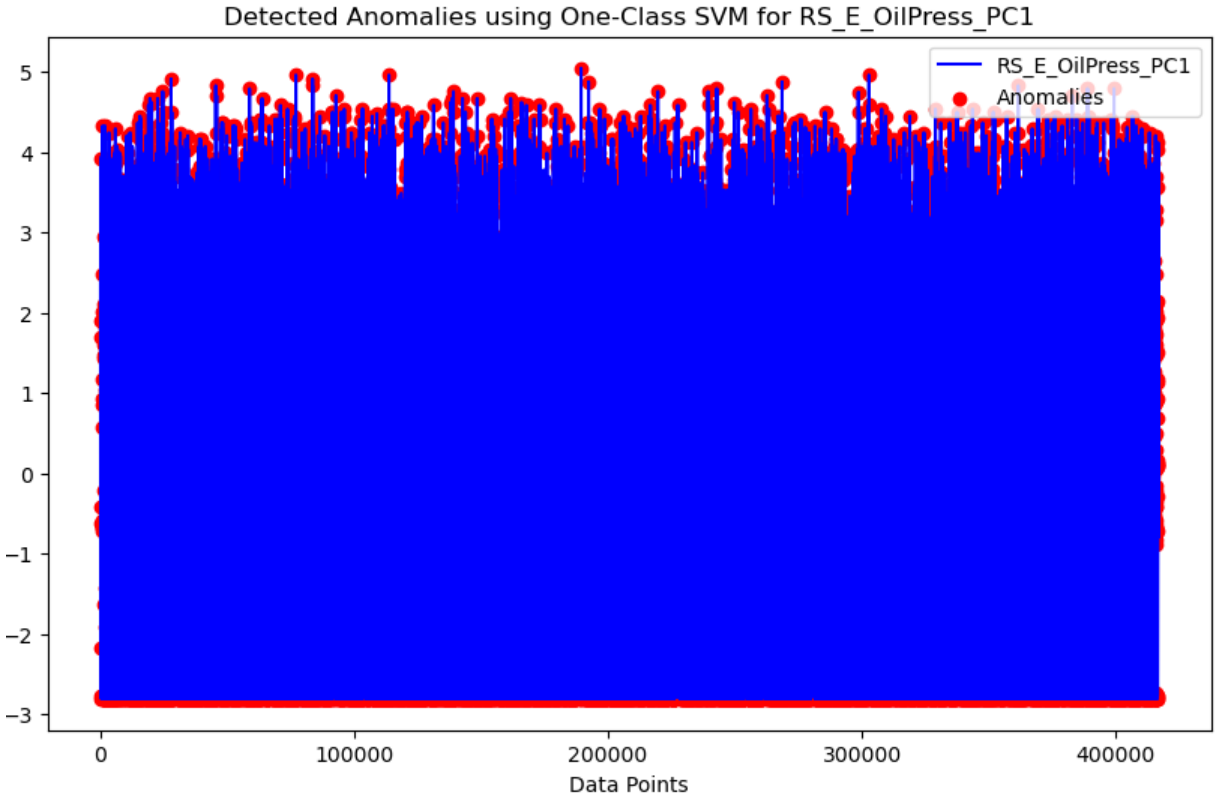
Anomalies Detection model variable 2



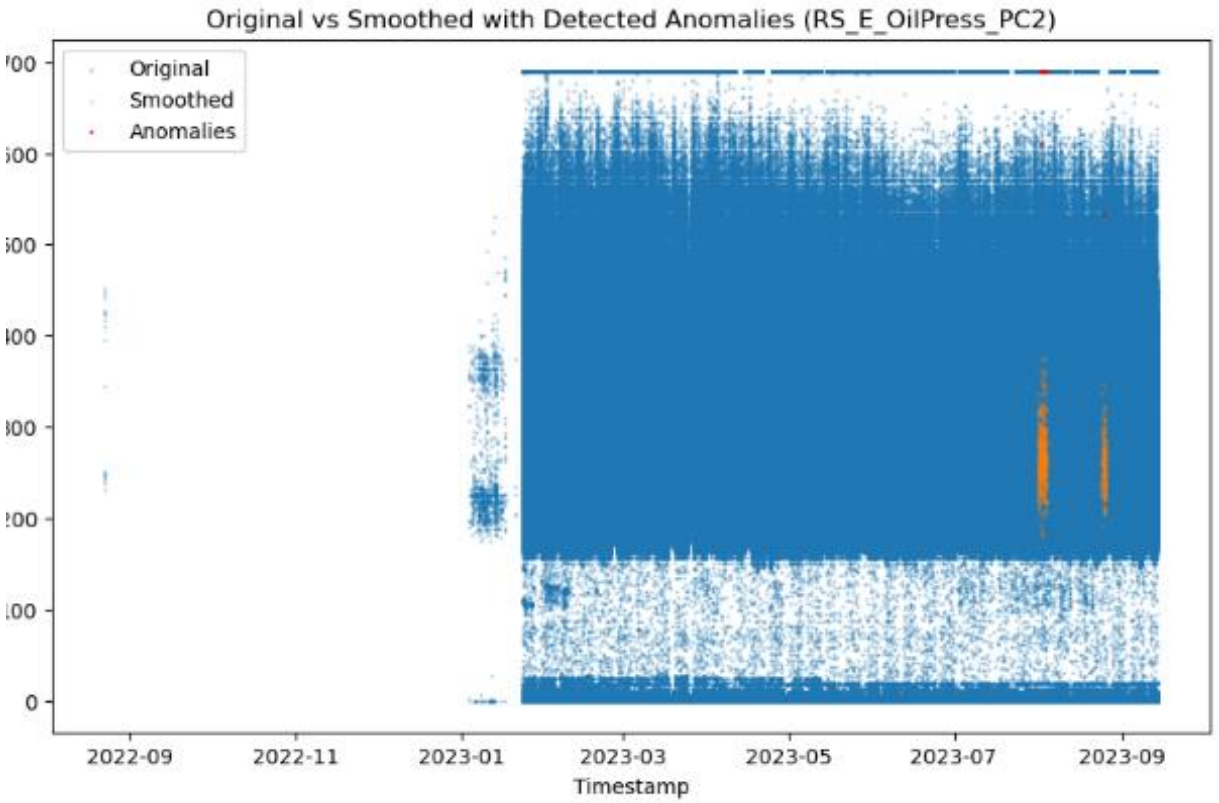


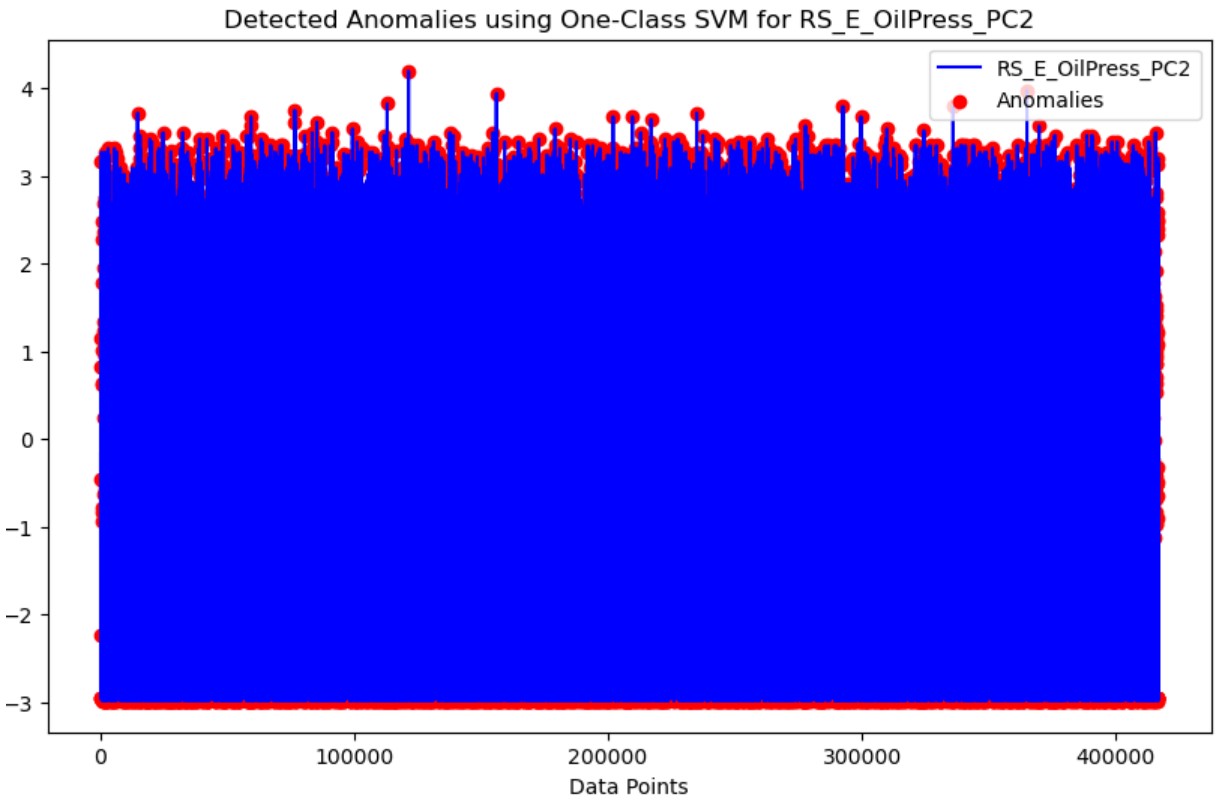
Anomalies Detection model variable 3



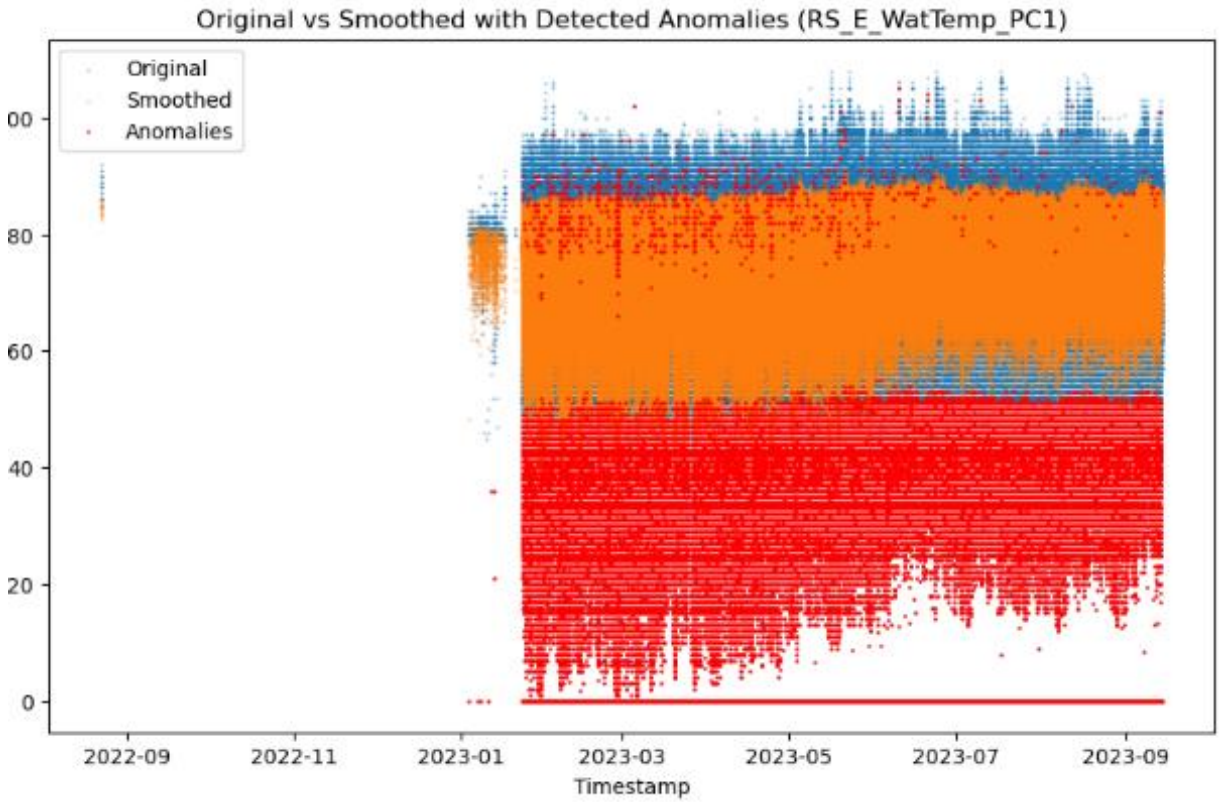


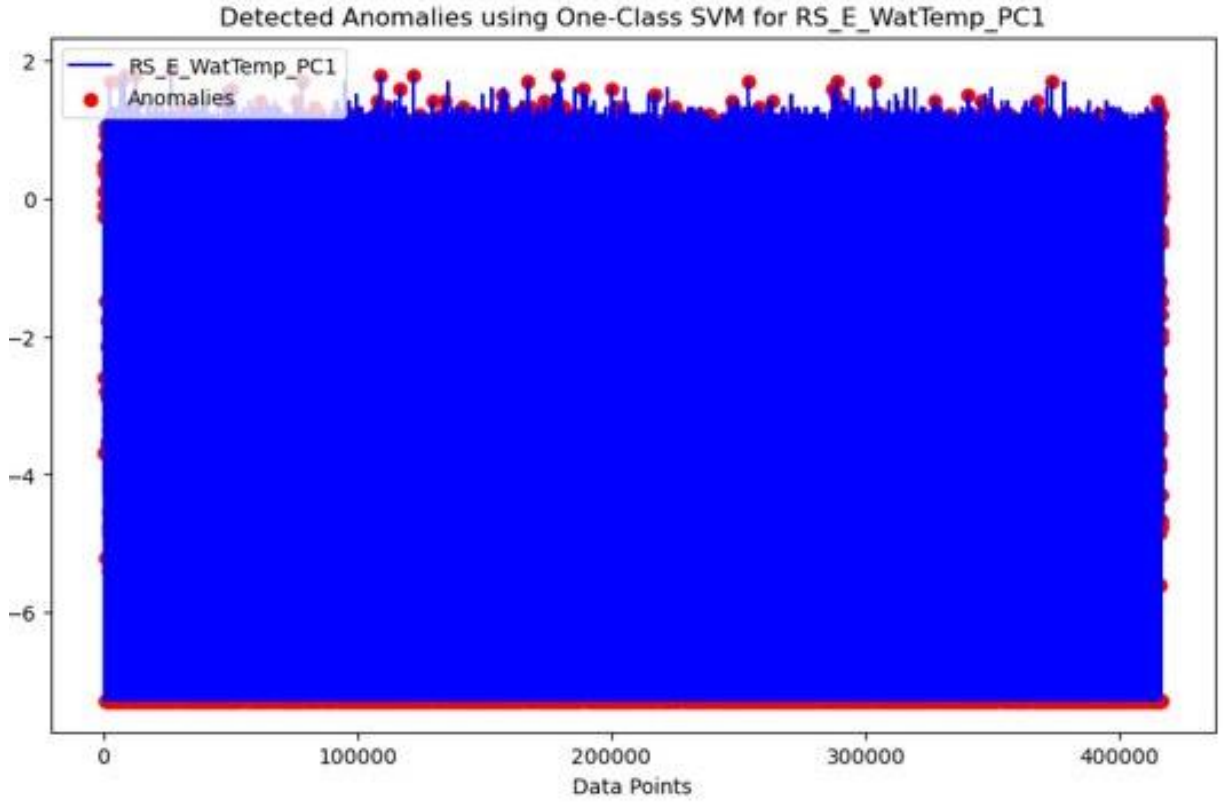
Anomalies Detection model variable 4



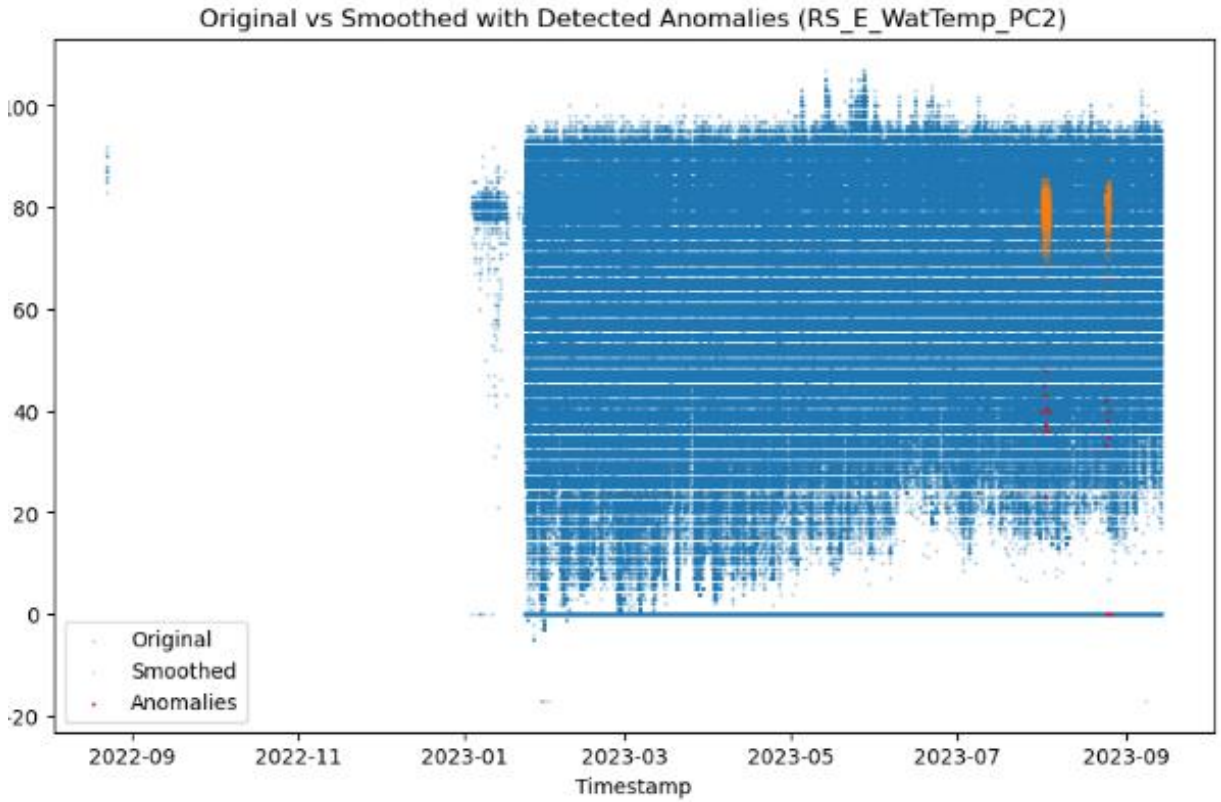


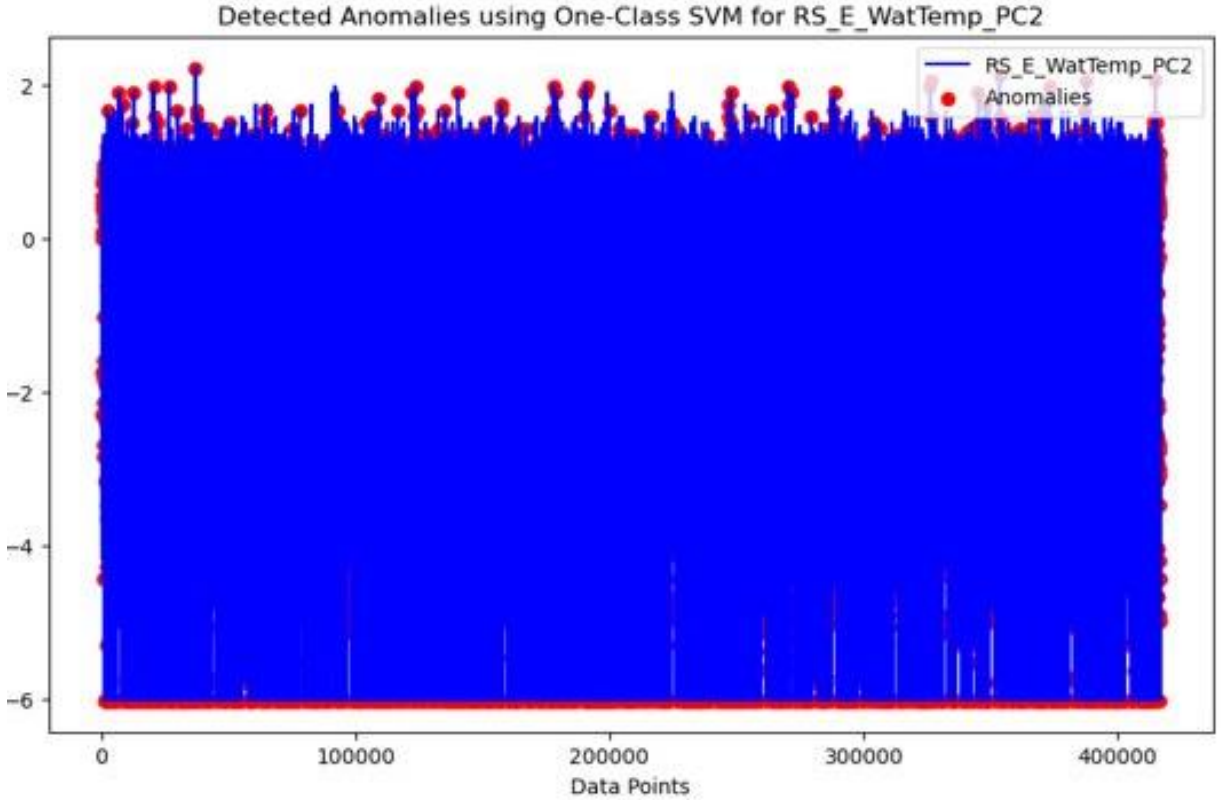
Anomalies Detection model variable 5



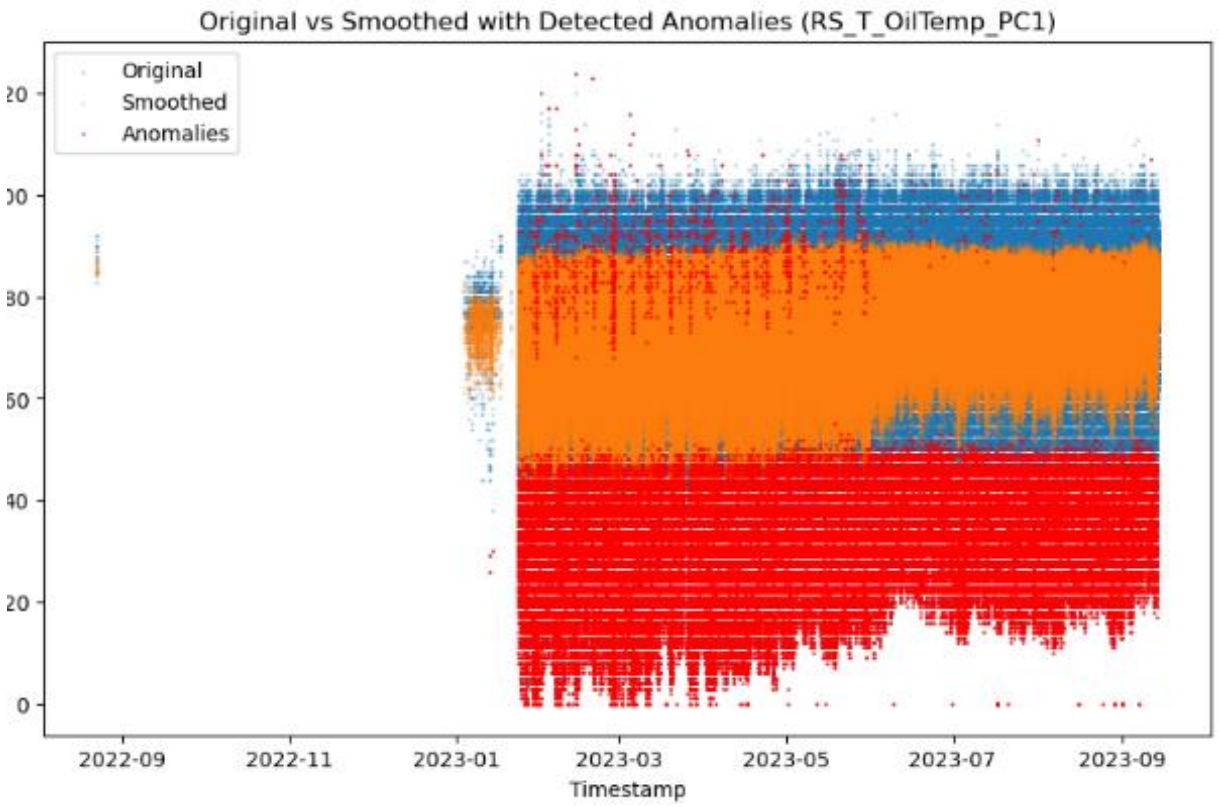


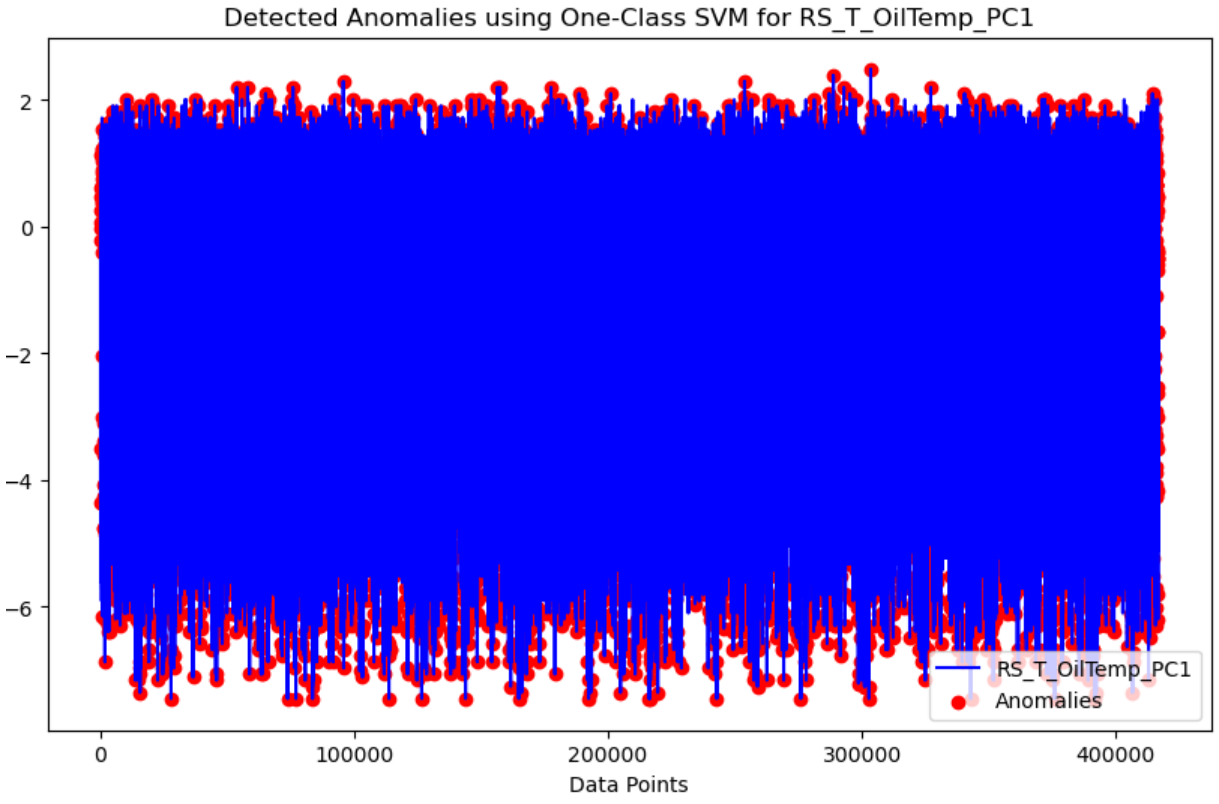
Anomalies Detection model variable 6



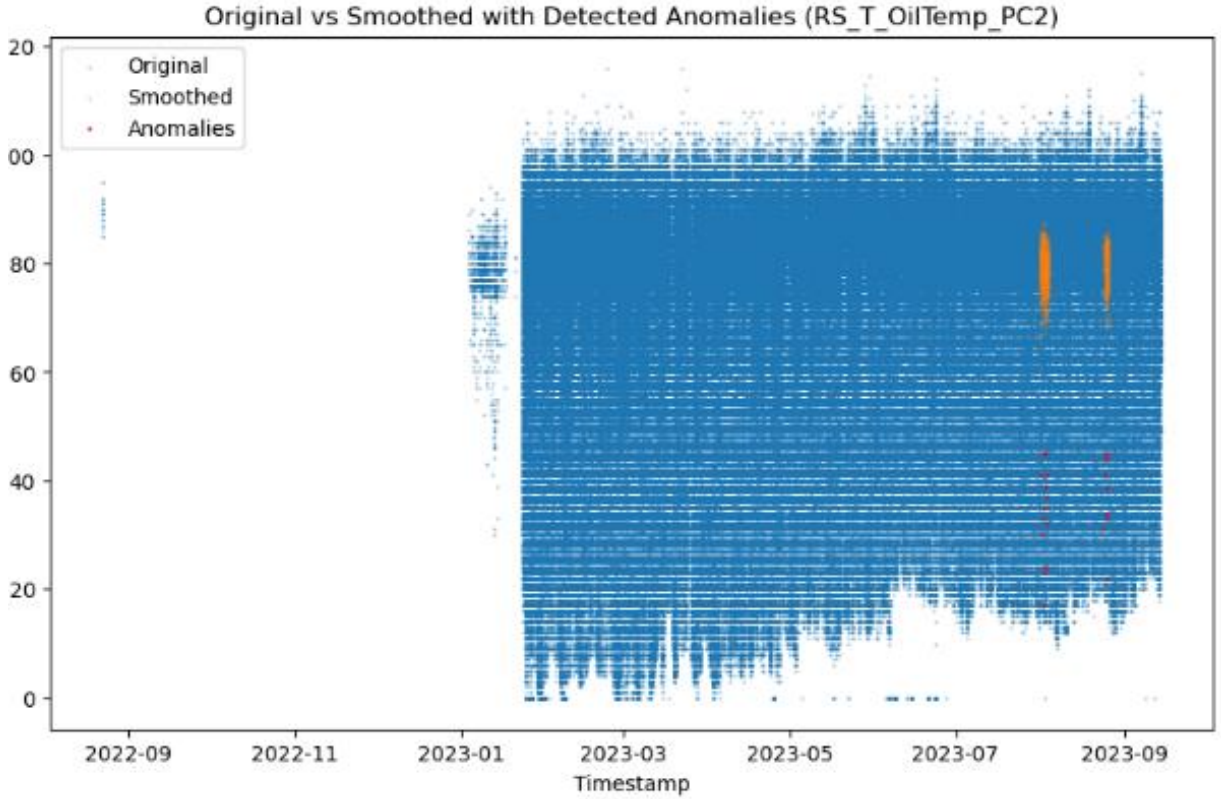


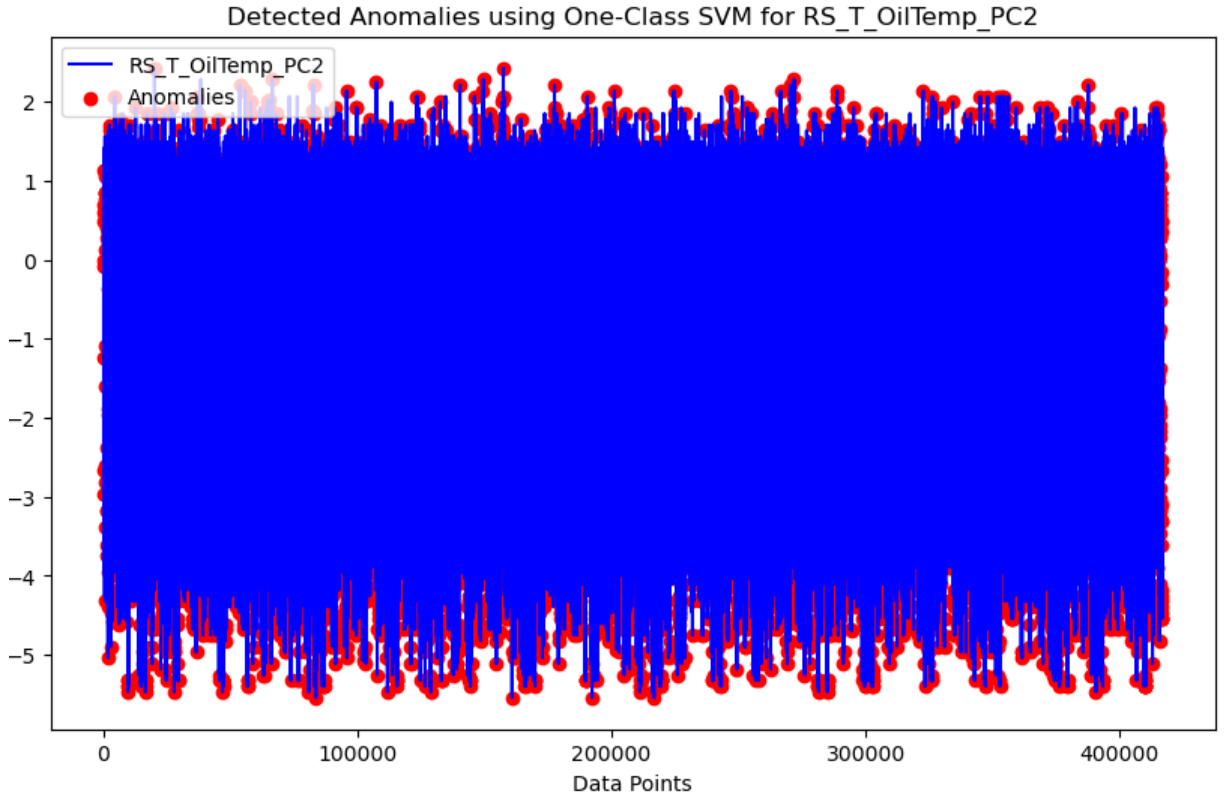
Anomalies Detection model variable 7



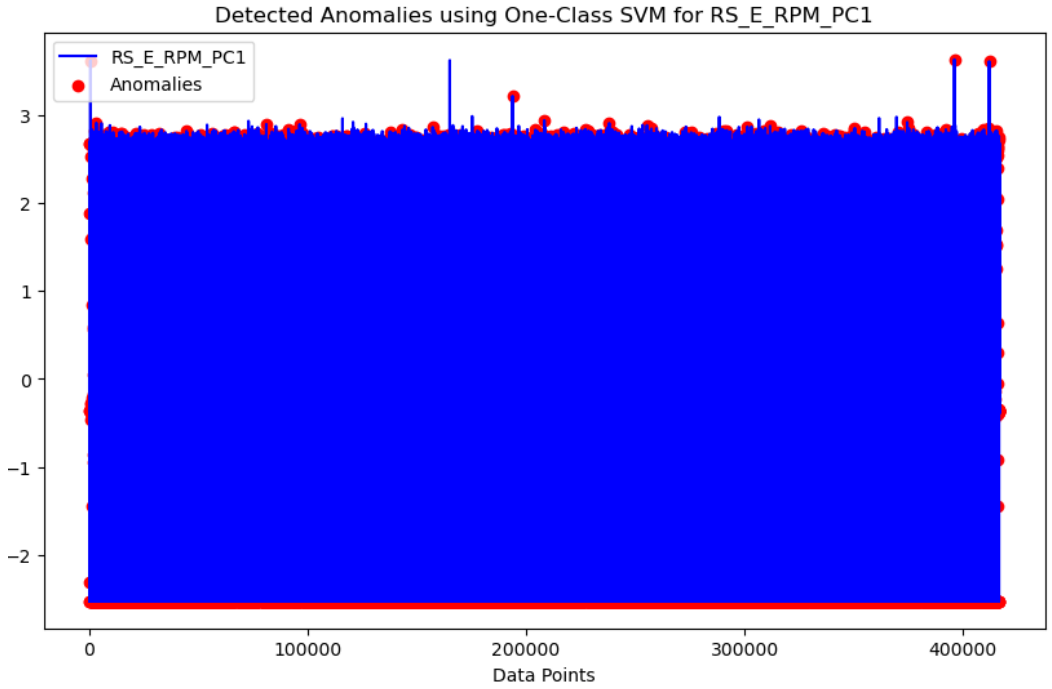


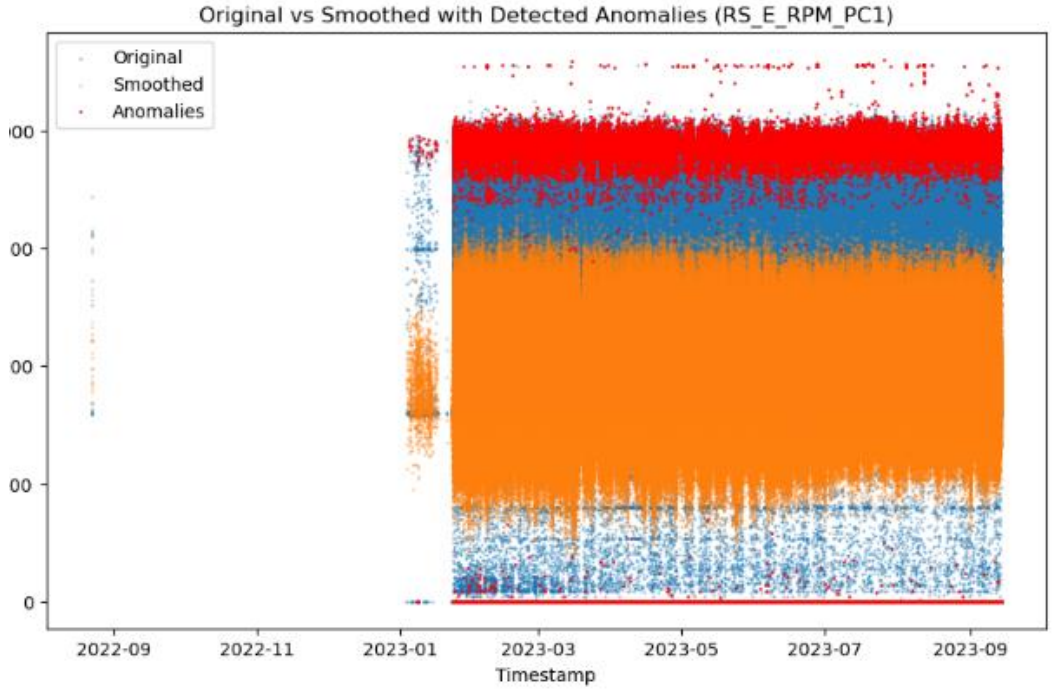
Anomalies Detection model variable 8



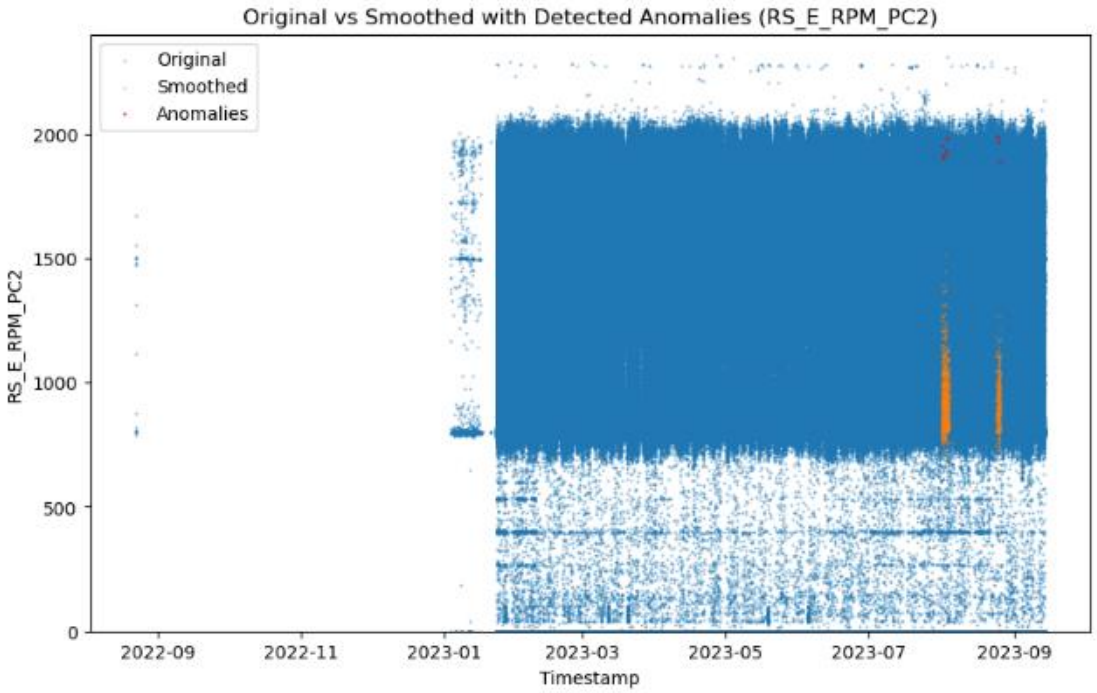


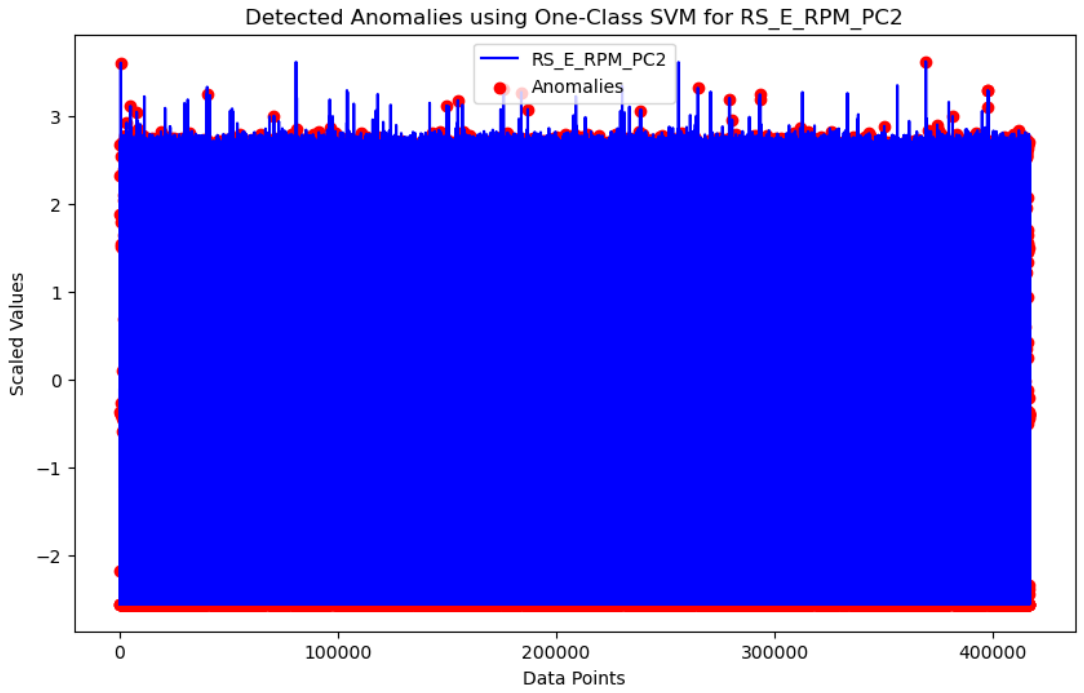
Anomalies Detection model variable 9





Anomalies Detection model variable 10

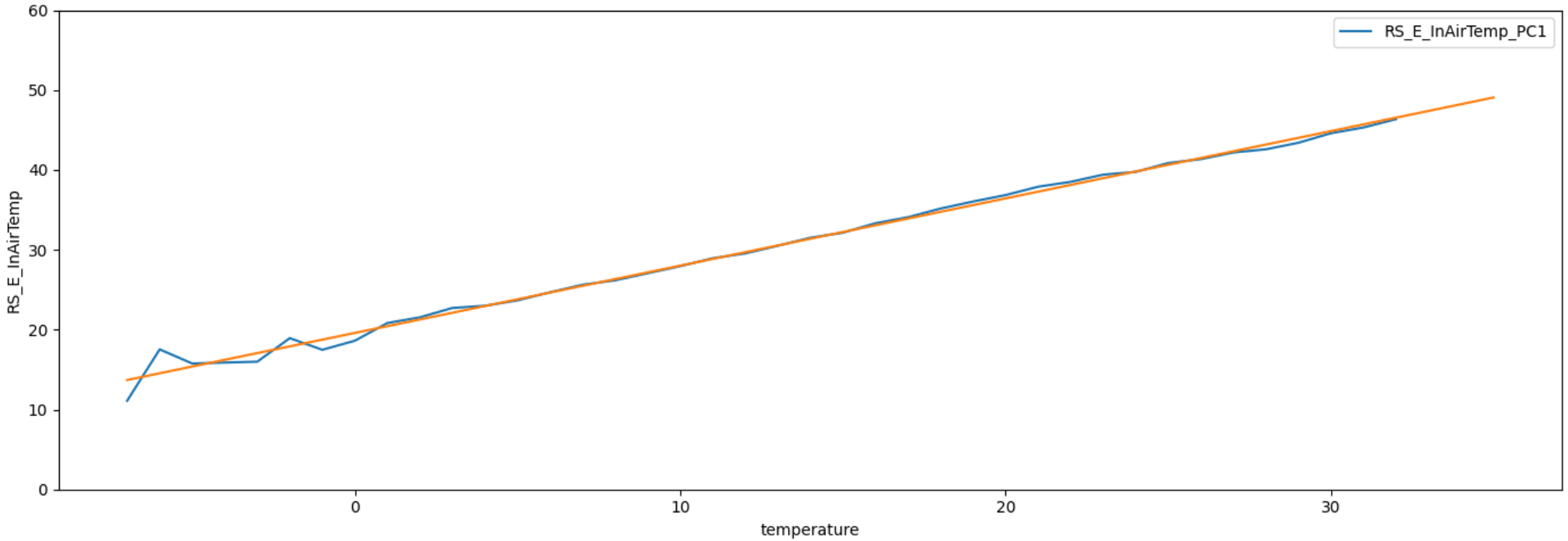




Model

A correlation between column values and temperature.

- Specifically, the correlation between outside temperature and the mean value of 'RS_E_InAirTemp_PC1,' calculated for each outside temperature value rounded to integer, demonstrated linearity with a slope of 19.6 and an intercept of 0.84



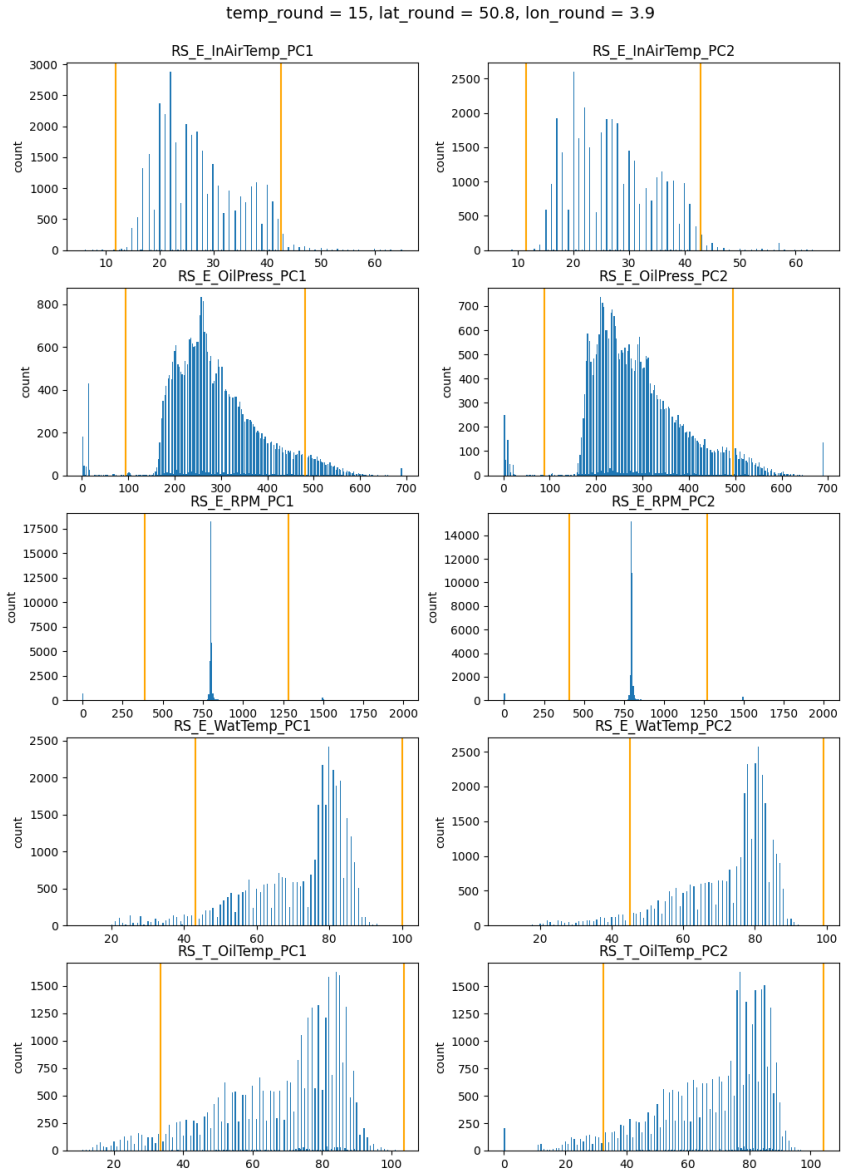
Clusterization

We removed all lines containing at least one zero or inappropriate value, except for two RPM values (since these can be zero)

We grouped all rows by latitude and longitude rounded to the first decimal place, and outside temperature rounded to the nearest integer

We calculated mean value and standard deviation of each parameter for all clusters

Each line in the original dataset was assigned to a specific cluster.



Flag generation

- We generated a code for each line in the original dataset, indicating potential outliers. If a parameter deviated by more than 2 times the standard deviation, it was marked as a potential anomaly.

- The code could contain the following numbers:
- -2: no potential outliers,
- 0: outlier is 'RS_E_InAirTemp_PC1'
- 1: outlier is 'RS_E_InAirTemp_PC2'
- 2: outlier is 'RS_E_OilPress_PC1'
- 3: outlier is 'RS_E_OilPress_PC2'
- 4: outlier is 'RS_E_RPM_PC1'
- 5: outlier is 'RS_E_RPM_PC2'
- 6: outlier is 'RS_E_WatTemp_PC1'
- 7: outlier is 'RS_E_WatTemp_PC2'
- 8: outlier is 'RS_T_OilTemp_PC1'
- 9: outlier is 'RS_T_OilTemp_PC2'

- For example, a code of '-2' indicates that our algorithm did not identify any anomalies in that line. Alternatively, a code like '2468' signifies that the values of 'RS_E_OilPress_PC1', 'RS_E_RPM_PC1', 'RS_E_WatTemp_PC1', and 'RS_T_OilTemp_PC1' are considered anomalies.

Final dataset

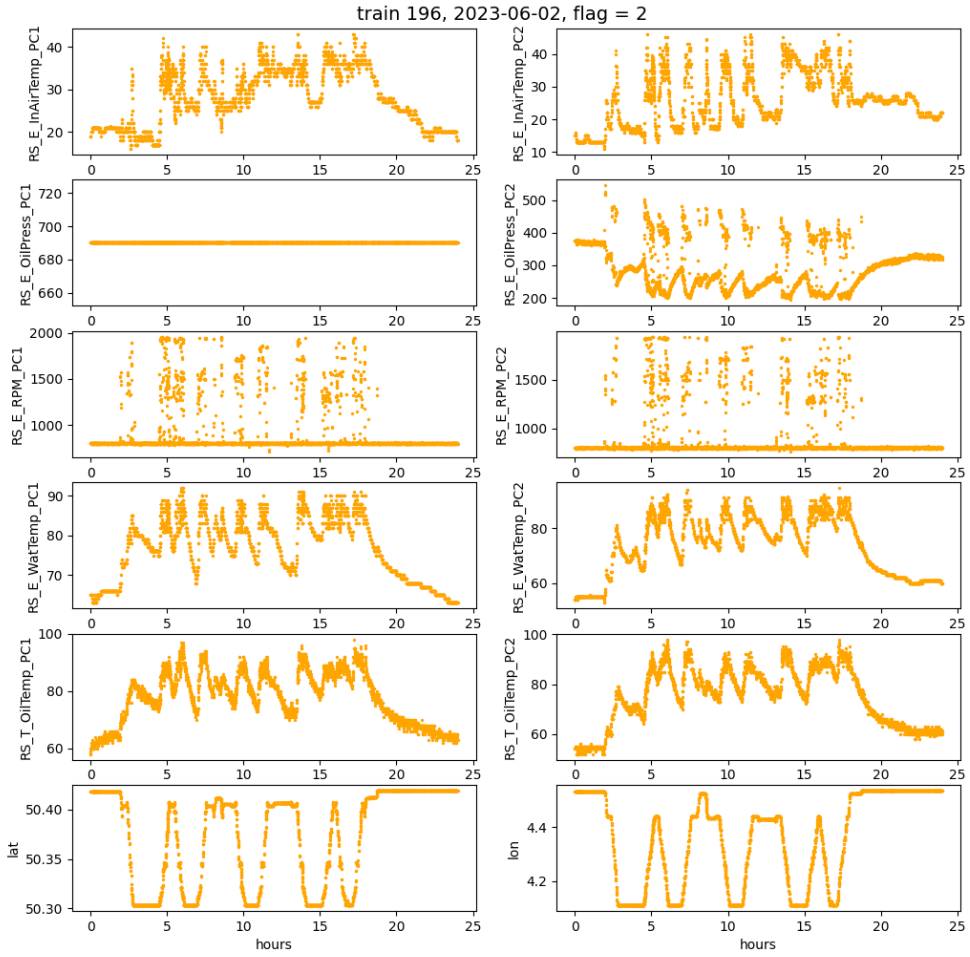
We grouped all rows by train_id, flag and date and calculated the count of each code and its percentage relative to the total number of rows.

We stated that a train has a real problem if the percentage is more than 40% and number of entries is more than 100(if we have a small number of rows we do not have enough statistics).

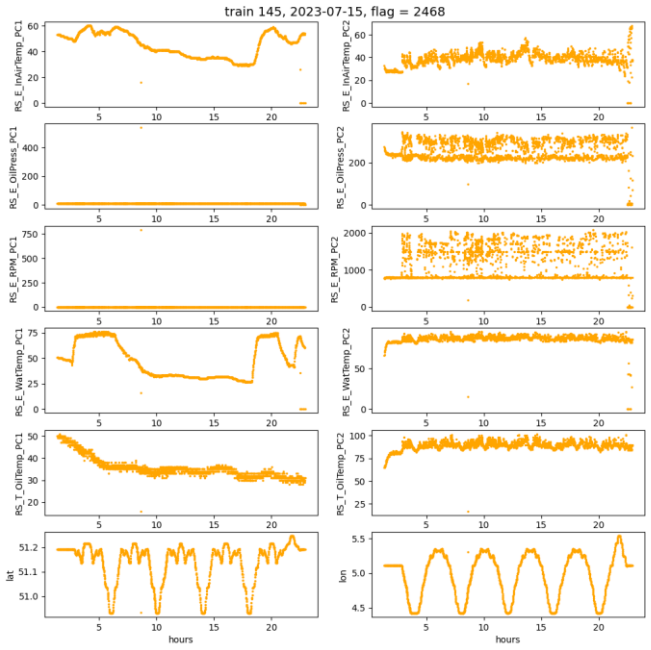
Applying these conditions we got a data frame with 1138 rows:

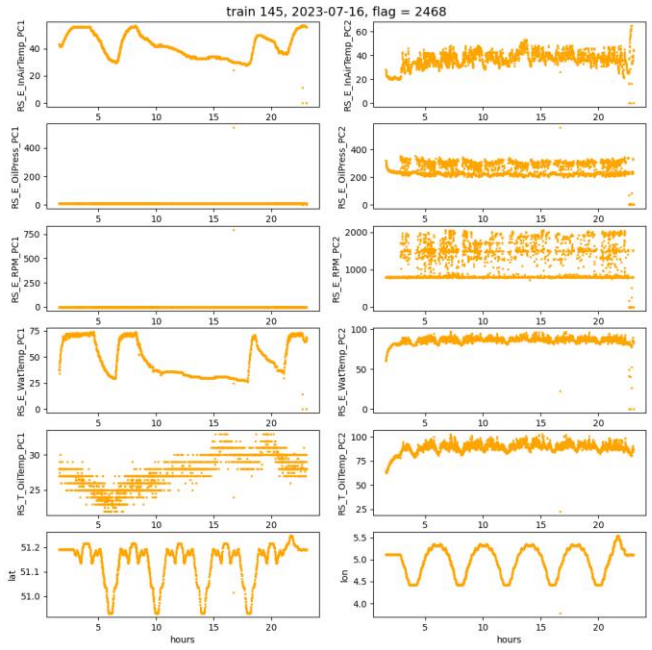
	mapped_veh_id	date	flag	count	percent
309242	196.0	2023-06-02	2	3148	76.705653
310461	196.0	2023-08-09	2	3079	86.854725
310656	196.0	2023-08-30	2	2996	84.632768
310597	196.0	2023-08-28	2	2919	82.155925
308501	196.0	2023-05-03	2	2842	81.292906
137450	148.0	2023-01-27	5	2767	88.799743
309260	196.0	2023-06-03	2	2762	99.495677
137550	148.0	2023-02-02	5	2726	90.384615
95359	136.0	2023-01-24	3	2662	93.142057
309172	196.0	2023-05-31	2	2655	74.411435

Illustration of a problem with one sensor

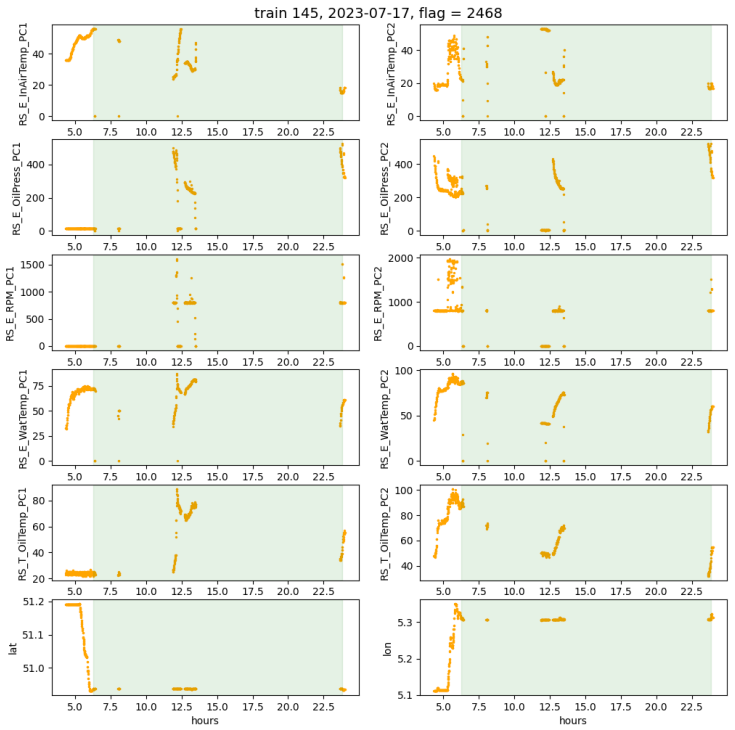


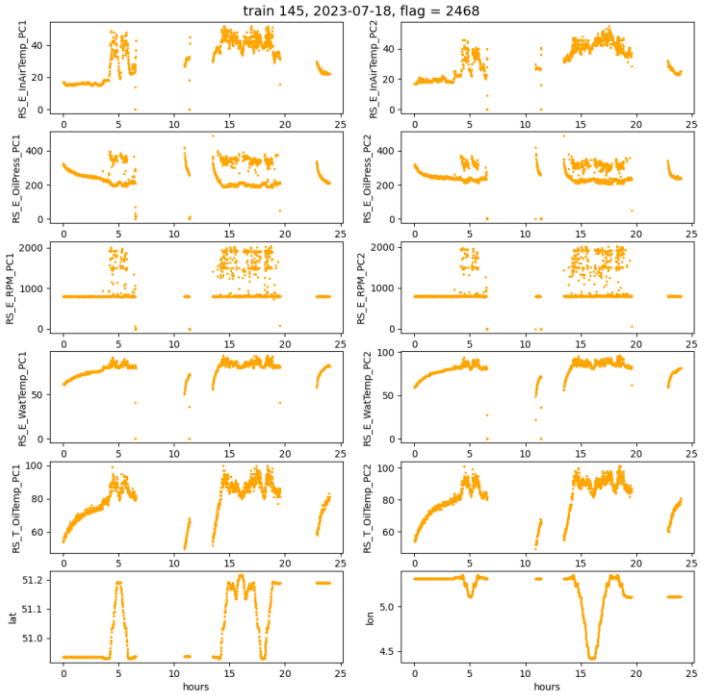
- An important problem with first engine or cooling system





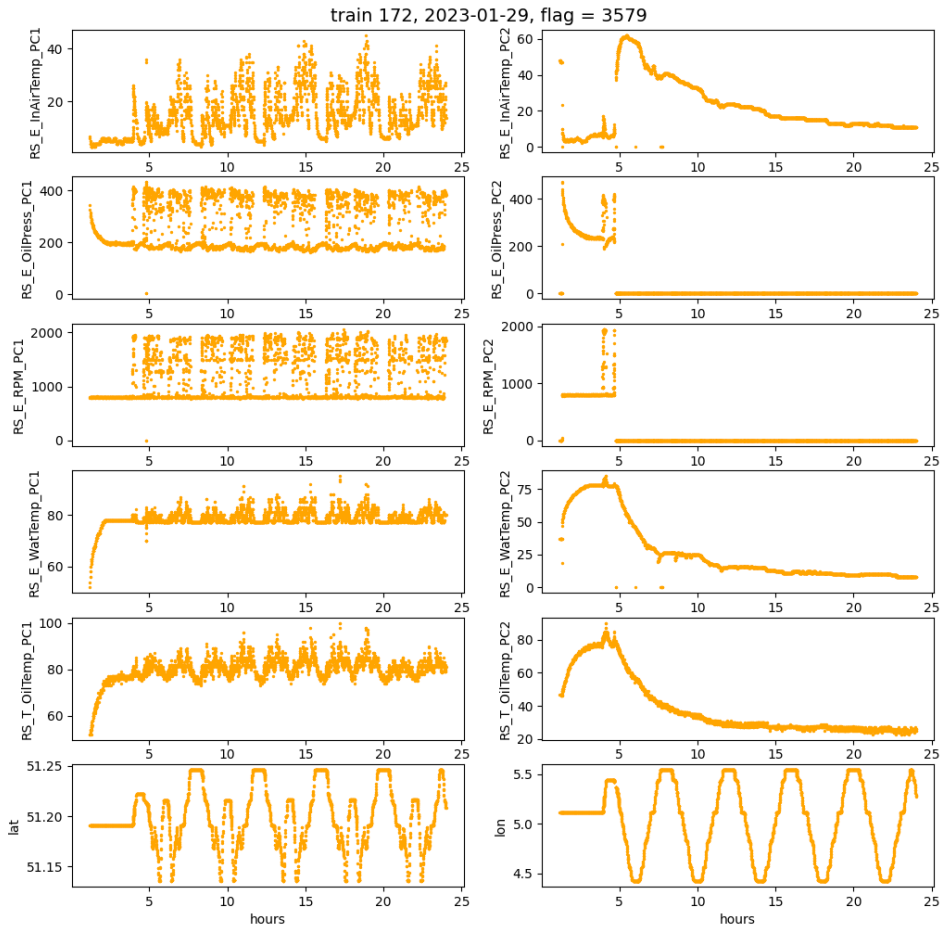
Train in workshop after breakdown



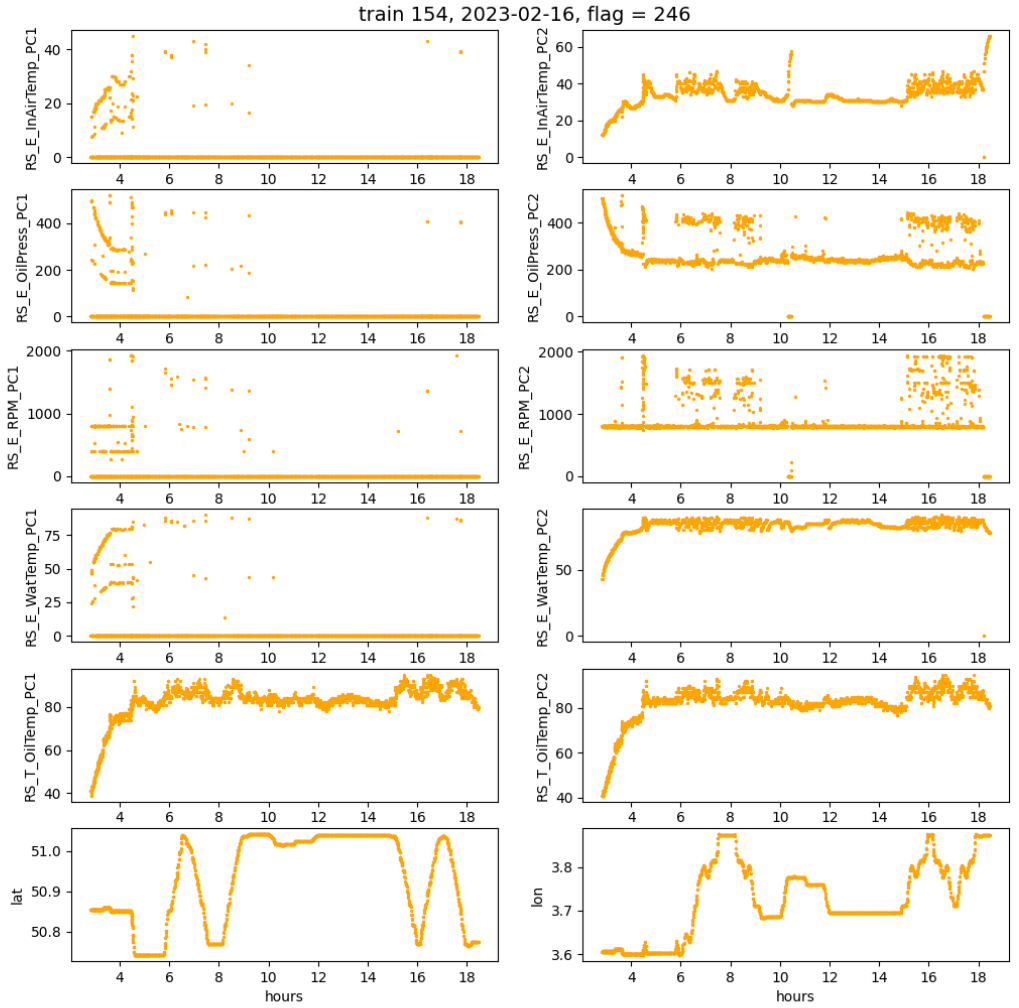


3	Traction	Hasselt	Atelier de traction d'Hasselt	SNCB	locomotives & railcars	50.937, 5.305	50.937	5.305

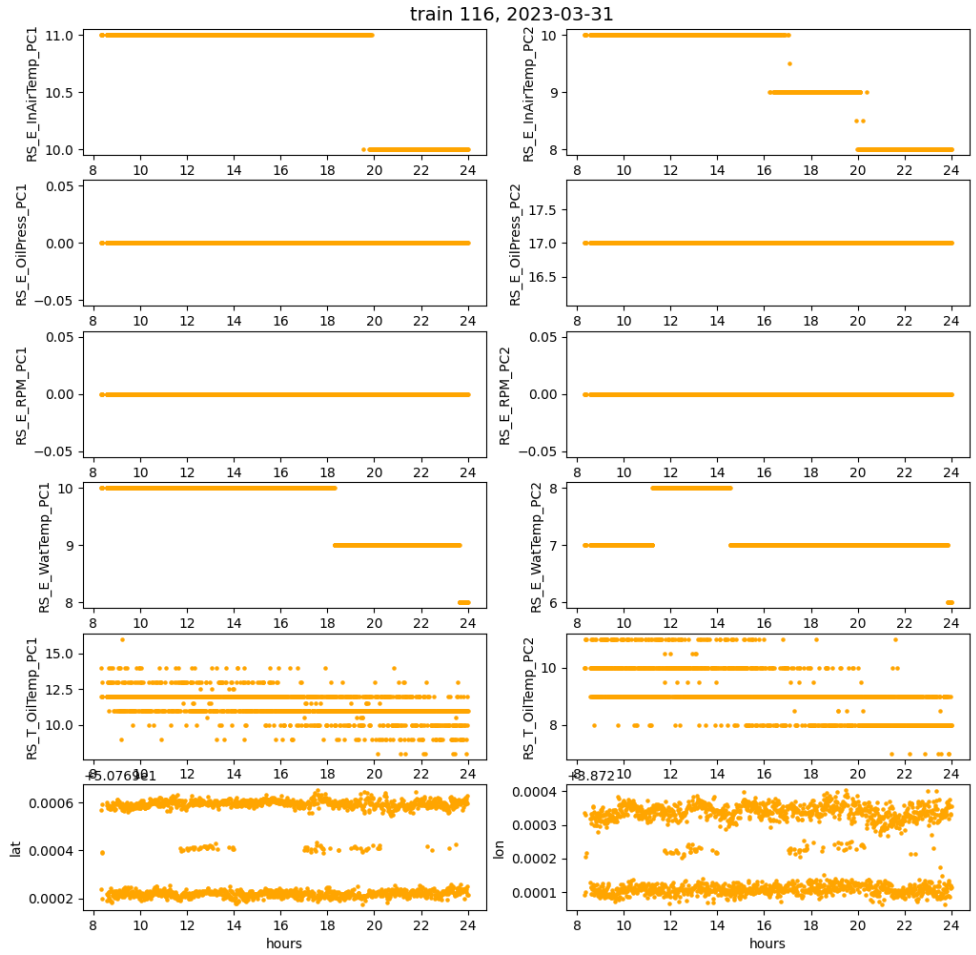
- An important problem
- with second engine or cooling system



- Problem with several sensors



- Problem with all sensors and GPS



- Our model is quite simple and easy to understand.
- However, it is flexible and has a number of parameters that can be tuned to increase performance. For example, we can change bin sizes of temperature and coordinats. Also, we can develop our model taking into account heating regime during winter, also when outside temprerature is low, model sometimes does not find an anomaly when InAirTemperature is low. Moreover, when temperatures are low standart deviation is large and some anomalies can not be detected.
- Our model can be directly used in streaming mode. Using the calculated lookup table our model can on fly determine if a given line of parameters is anomalous or not. If the anomaly persists, for example if 50 of 100 last parameter lines are detected as anomalous with the same flags, this indicates problems with sensors or breakdown of one of the systems of the train.

Dashboard

Use case

- To develop how the dashboard is working, we will define a use case.

- In our use case, we will check for train 172 on the 28.08.2023. This train is chosen because it's the topest anomalies counted in our prototype dashboard.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

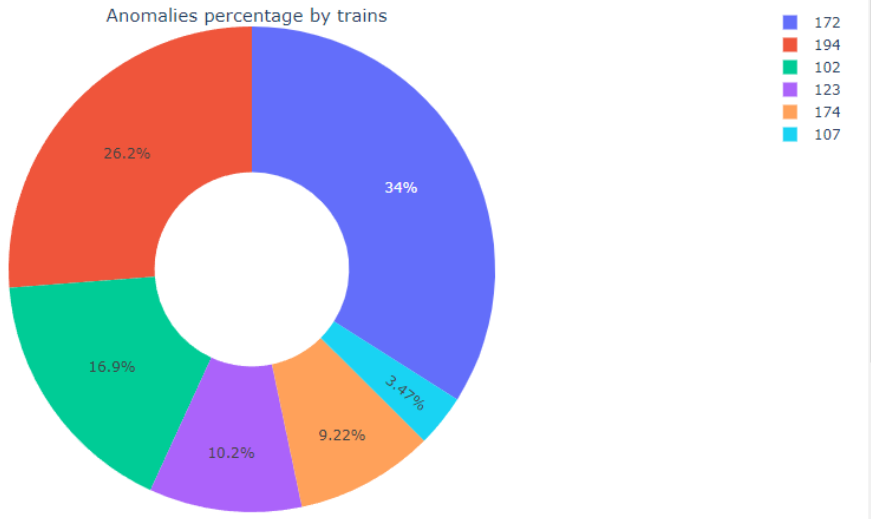
INFO-H423 - Data Mining - 202324

SNCB Train Anomalies Analysis

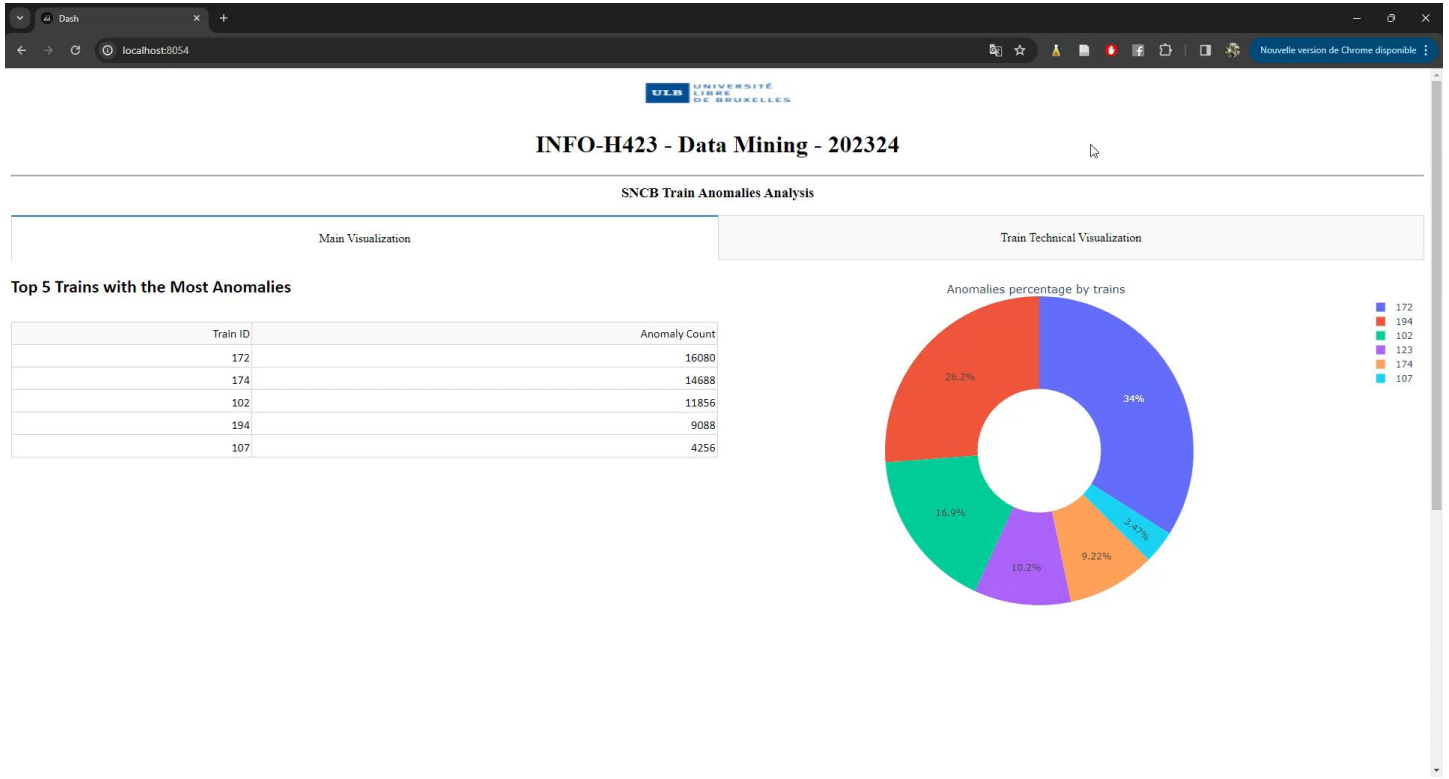
Main Visualization | Train Technical Visualization

Top 5 Trains with the Most Anomalies

Train ID	Anomaly Count
172	16080
174	14688
102	11856
194	9088
107	4256



Demo


Dash
localhost:8054
Nouvelle version de Chrome disponible
INFO-H423 - Data Mining - 202324
SNCB Train Anomalies Analysis
Main Visualization
Train Technical Visualization
Top 5 Trains with the Most Anomalies
Train ID
Anomaly Count
172
16080
174
14688
102
11856
194
9088
107
4256
Anomalies percentage by trains
34%
26.2%
16.9%
10.2%
9.22%
3.47%
172
194
102
123
174
107

Source code and process files.

- All code built for this project is publicly available at : https://github.com/steveworrall-ulb/INFO-H423_SNCB

- Includes the components necessary for real time implementation if the filtering and evaluation process of new data

Thank you !