CAF Energy Efficiency Analysis

Master in Big Data and Business Analytics

Corporate Project

Full Report

Group 07

July 2024

Thomas Russell

Mario Garcia

Rodrigo Reyes

Carlos Eid

Yazeed Yabroudi

"We hereby certify that this report is our own original work in its entirety, unless where indicated and referenced"

Table of Contents:

1. Background and Context

This report details an analysis of the CAF train driver journey dataset to assess the energy efficiency of train drivers and anomalous energy trends over the time frame of 2022-2024. The objective is to identify key drivers who optimize energy consumption focused on metrics such as energy per kilometre and total variance. Data preprocessing, optimization, evaluation, and data visualization are some of the techniques used to find consensus for the two assigned questions.

2. Introduction

CAF (Construcciones y Auxiliar de Ferrocarriles) founded in 1917, is a Spanish company that specializes in the design, manufacture, and maintenance of railway rolling stock, signaling systems, and infrastructure. According to the Spanish Ministry of Development, both passenger and freight trains have over 1.3 billion journeys per year. There are over 1,200 CAF-operated trains daily, with over 400 million passengers per year. CAF operates over 1,500 freight trains moving over 100 million tonnes of goods per year. CAF has a fleet of over 2,000 trains including diesel, electric, and hybrid.

Energy-efficient operations can provide many financial incentives, improve overall operations, and contribute to sustainability. Our mandate from CAF is to advance and deepen two questions related to energy efficiency. First question is about driving improvement, **which train drivers carry out the most optimal driving from the point of view of energy without altering schedules**. Second question is about Reducing out of service energy, **which equipment/trains have an anomalous consumption**.

Our approach to answering the questions follows the scientific method of data collection, cleaning, data selection, data normalisation and transformation, use of analytical techniques, categorization, and further analysis. We aimed to understand the various challenges of working with a dynamic dataset and using multiple machine-learning models to derive meaningful insights.

Question 1 Report: Train drivers with optimal driving energy consumption in CAF trains

3. Data Preparation and Methodology

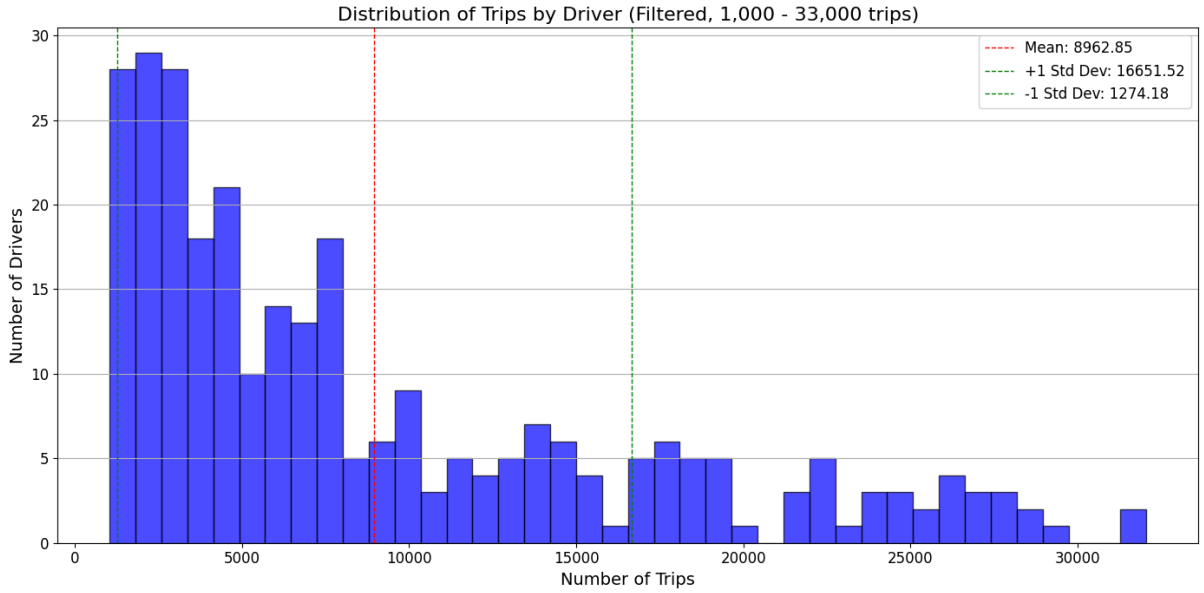
3.1 Data Collection, Exploration

The dataset comprises trip records from April 2022 to April 2024, capturing details like driver ID, trip duration, distance covered, and various energy consumption metrics (net, auxiliary,

and traction energy). It has temporal data that specify trip duration, station data that specifies journey points, operational data which provides operational statuses, driver information such as “si_siv_ndriver_def” Energy consumption metrics include “net_energy consumed by UT” and weighted versions such as “net_energy consumed by UT weight” Data types include numerical and categorical which are suitable for quantitative analysis and categorization.

3.2 Data Cleansing and Selection

This data was loaded from a CSV file into a pandas data frame, which allows easy manipulation and analysis. The drivers were filtered for those with 1,000 trips and no more than 33,000 trips in order to have a balanced dataset. The data was further refined by calculating the number of trips each driver made. A filtered data frame containing only the drivers who met this criteria was then created. This filtered data frame relevant to our case study was used for further analysis ensuring only consistent and relevant data was used. In addition, Initial data cleansing involved removing entries with missing driver IDs or null values in critical fields. We also excluded outlier data that could skew analysis results, such as trips with extremely high or low energy consumption figures which stemmed from drivers who fell outside our minimum or maximum trip parameters. In addition, the column si_siv_ndriver_def had 412,569 instances of “No Disponible” which indicates there was no driver ID that was not recorded for the trip and all these values were excluded from the analysis.



3.3 Analytical Techniques Employed

Using Python software on VS Code platform with required libraries such as Matplotlib and Seaborn, we used various techniques to build our investigation into the dataset. The process was divided into the following steps

1) Descriptive statistics – To check the total number of drivers, total trips per driver, total number of unique trips, unique trips per driver

2) Data filtering – To filter drivers based on number of trips

3) Histogram visualizations – To visualize trips made by drivers

4) Energy calculations – To check the metrics of the trips including energy per km and per trip

5) Ranking and sorting – To rank efficient drivers over the less efficient drivers based on criteria

6) Variance calculations – Computing variances across trips for different drivers

7) Data aggregations – To group data per driver to gather insights

8) Visualizations – To identify trends and outliers and explore impact of different variables

9) Conditional Logic – To refine data further to identify trends within certain thresholds which includes normalization of data which is used to reduce bias

The above techniques focused on the driver behaviors which allowed us to identify efficiency patterns, and guided decision-making towards answering the questions of driver efficiency and optimal journey energies.

4. Detailed Analysis of Energy Consumption

4.1 Energy Consumption per Kilometre

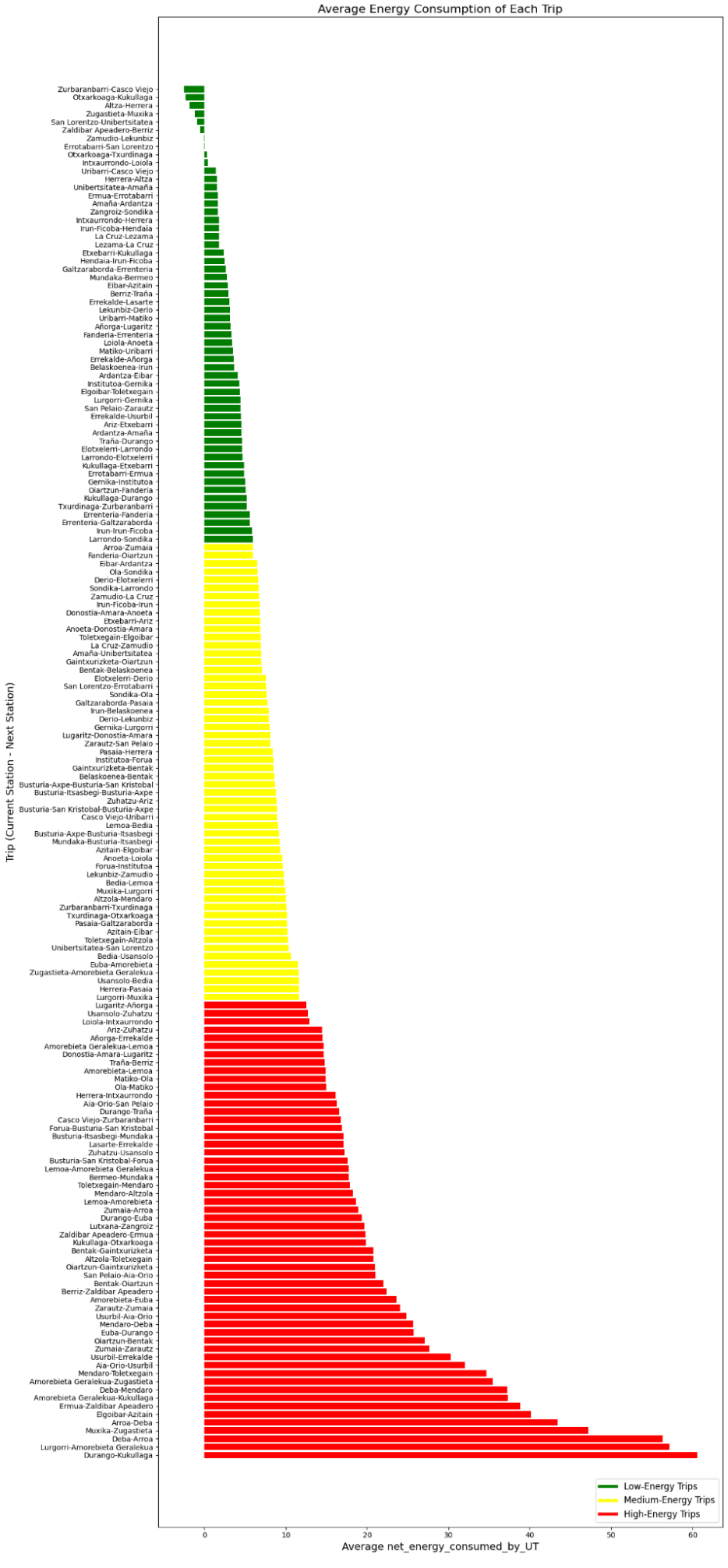
This metric under column name “net_energy_consumed_by_UT” was divided by kilometres per trip in order to provide insight for the energy each driver uses per kilometre traveled. A comparison of these values highlights the drivers who manage to use less energy over distance which is a direct indicator of driver efficiency. Drivers with a lower energy consumption are classified as high efficiency. This metric can provide practical improvement in the operations.

4.2 Variance in Energy Consumption

Variance was calculated for each driver using “net_energy_consumed_by_UT” and the number of trips, and insights on the performance of a driver regardless of the conditions such as distance, route, traffic, or skill level. Variance per driver can provide a stable performance or low variance compared with less adequate performance or high variance. This affects the energy consumption during journeys and further highlights driver performance.

4.3 Categorization of Trips

A key step to our analysis was to split the energy consumption data based on net_energy_consumed_by_UT into three categories of low, medium, and high energy trips. Using a fixed quantile of 33% provides a relative comparison rather than an arbitrary figure. This was due to finding a significant impact on overall performance when a driver frequently logged more demanding trips. It also provides a more natural distribution to avoid bias or any mischaracterization of the dataset and further analysis. Below is a chart to display the energy per trip for a visualization of the journeys contained in this CAF dataset.


Average Energy Consumption of Each Trip
Trip (Current Station - Next Station)
Zurbaranbarri-Casco Viejo
Otxarkoaga-Kukullaga
Altza-Herrera
Zugastieta-Muxika
San Lorentzo-Unibertsitatea
Zaldibar Apeadero-Berriz
Zamudio-Lekunbiz
Errotabarri-San Lorentzo
Otxarkoaga-Txurdinaga
Intxaurrondo-Loiola
Uribarri-Casco Viejo
Herrera-Altza
Unibertsitatea-Amaña
Ermua-Errotabarri
Amaña-Ardantza
Zangroiz-Sondika
Intxaurrondo-Herrera
Irun-Ficoba-Hendaia
La Cruz-Lezama
Lezama-La Cruz
Etxebarri-Kukullaga
Hendaia-Irun-Ficoba
Galtzaraborda-Errenteria
Mundaka-Bermeo
Eibar-Azitain
Berriz-Traña
Errekalde-Lasarte
Lekunbiz-Derio
Uribarri-Matiko
Añorga-Lugaritz
Fanderia-Errenteria
Loiola-Anoeta
Matiko-Uribarri
Errekalde-Añorga
Belaskoenea-Irun
Ardantza-Eibar
Institutoa-Gernika
Elgoibar-Toletxegain
Lurgorri-Gernika
San Pelaio-Zarautz
Errekalde-Usurbil
Ariz-Etxebarri
Ardantza-Amaña
Traña-Durango
Elotxelerri-Larrondo
Larrondo-Elotxelerri
Kukullaga-Etxebarri
Errotabarri-Ermua
Gernika-Institutoa
Oiartzun-Fanderia
Kukullaga-Durango
Txurdinaga-Zurbaranbarri
Errenteria-Fanderia
Errenteria-Galtzaraborda
Irun-Irun-Ficoba
Larrondo-Sondika
Arroa-Zumaia
Fanderia-Oiartzun
Eibar-Ardantza
Ola-Sondika
Derio-Elotxelerri
Sondika-Larrondo
Zamudio-La Cruz
Irun-Ficoba-Irun
Donostia-Amara-Anoeta
Etxebarri-Ariz
Anoeta-Donostia-Amara
Toletxegain-Elgoibar
La Cruz-Zamudio
Amaña-Unibertsitatea
Gaintxurizketa-Oiartzun
Bentak-Belaskoenea
Elotxelerri-Derio
San Lorentzo-Errotabarri
Sondika-Ola
Galtzaraborda-Pasaia
Irun-Belaskoenea
Derio-Lekunbiz
Gernika-Lurgorri
Lugaritz-Donostia-Amara
Zarautz-San Pelaio
Pasaia-Herrera
Institutoa-Forua
Gaintxurizketa-Bentak
Belaskoenea-Bentak
Busturia-Axpe-Busturia-San Kristobal
Busturia-Itsasbegi-Busturia-Axpe
Zuhatzu-Ariz
Busturia-San Kristobal-Busturia-Axpe
Casco Viejo-Uribarri
Lemoa-Bedia
Busturia-Axpe-Busturia-Itsasbegi
Mundaka-Busturia-Itsasbegi
Azitain-Elgoibar
Anoeta-Loiola
Forua-Institutoa
Lekunbiz-Zamudio
Bedia-Lemoa
Muxika-Lurgorri
Altzola-Mendaro
Zurbaranbarri-Txurdinaga
Txurdinaga-Otxarkoaga
Pasaia-Galtzaraborda
Azitain-Eibar
Toletxegain-Altzola
Unibertsitatea-San Lorentzo
Bedia-Usansolo
Euba-Amorebieta
Zugastieta-Amorebieta Geralekua
Usansolo-Bedia
Herrera-Pasaia
Lurgorri-Muxika
Lugaritz-Añorga
Usansolo-Zuhatzu
Loiola-Intxaurrondo
Ariz-Zuhatzu
Añorga-Errekalde
Amorebieta Geralekua-Lemoa
Donostia-Amara-Lugaritz
Traña-Berriz
Amorebieta-Lemoa
Matiko-Ola
Ola-Matiko
Herrera-Intxaurrondo
Aia-Orio-San Pelaio
Durango-Traña
Casco Viejo-Zurbaranbarri
Forua-Busturia-San Kristobal
Busturia-Itsasbegi-Mundaka
Lasarte-Errekalde
Zuhatzu-Usansolo
Busturia-San Kristobal-Forua
Lemoa-Amorebieta Geralekua
Bermeo-Mundaka
Toletxegain-Mendaro
Mendaro-Altzola
Lemoa-Amorebieta
Zumaia-Arroa
Durango-Euba
Lutxana-Zangroiz
Zaldibar Apeadero-Ermua
Kukullaga-Otxarkoaga
Bentak-Gaintxurizketa
Altzola-Toletxegain
Oiartzun-Gaintxurizketa
San Pelaio-Aia-Orio
Bentak-Oiartzun
Berriz-Zaldibar Apeadero
Amorebieta-Euba
Zarautz-Zumaia
Usurbil-Aia-Orio
Mendaro-Deba
Euba-Durango
Oiartzun-Bentak
Zumaia-Zarautz
Usurbil-Errekalde
Aia-Orio-Usurbil
Mendaro-Toletxegain
Amorebieta Geralekua-Zugastieta
Deba-Mendaro
Amorebieta Geralekua-Kukullaga
Ermua-Zaldibar Apeadero
Elgoibar-Azitain
Arroa-Deba
Muxika-Zugastieta
Deba-Arroa
Lurgorri-Amorebieta Geralekua
Durango-Kukullaga
0
10
20
30
40
50
60
Average net_energy_consumed_by_UT
Low-Energy Trips
Medium-Energy Trips
High-Energy Trips

5. Comparative Analysis and Findings

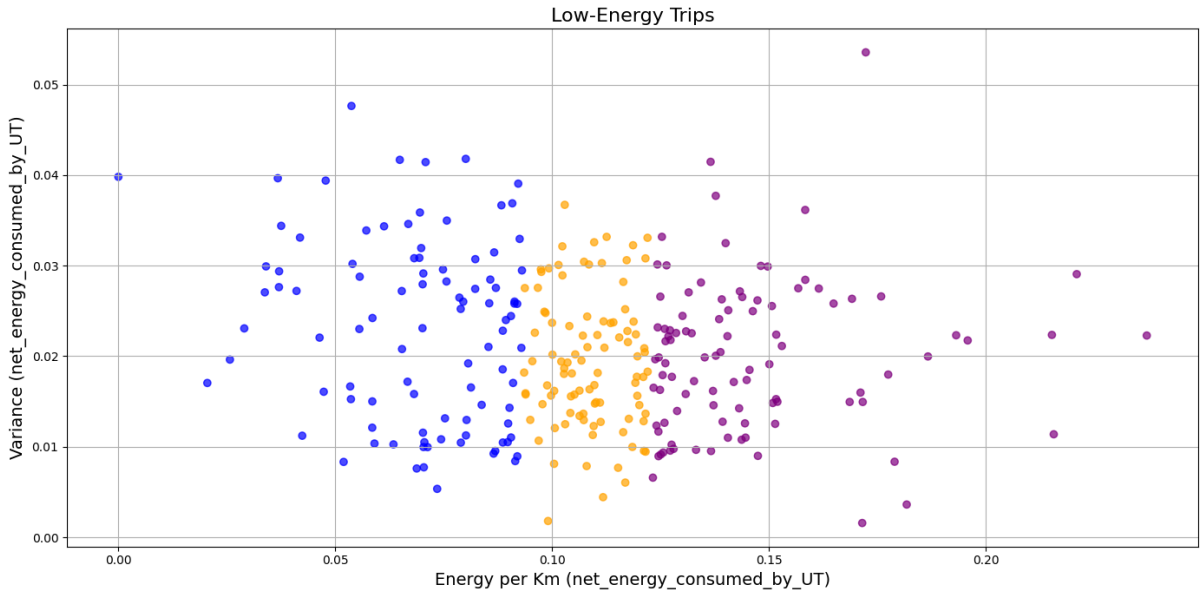
Our comparative study allowed us to visually inspect the three graphs, and find the top performers across all types of trips. For each type of trip, we used unique values to identify the top drivers based on the clustering who were most efficient in variance and energy per km. The values for each were as follows:

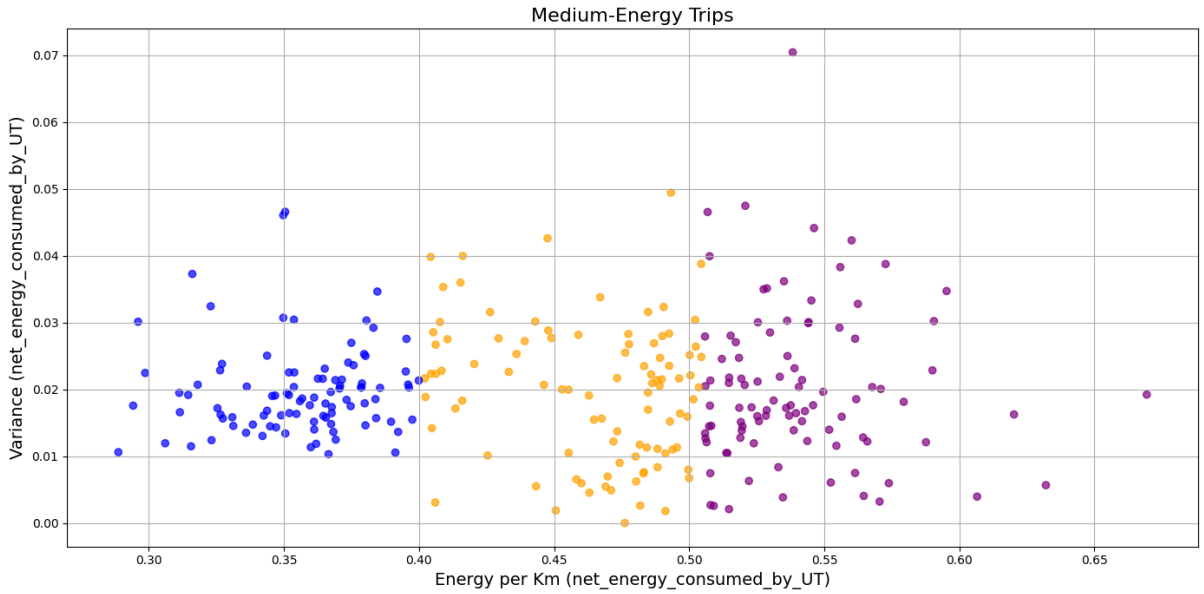
Low Energy Trips: Variance: 0 to 0.05 and Energy per KM: 0 to 0.1

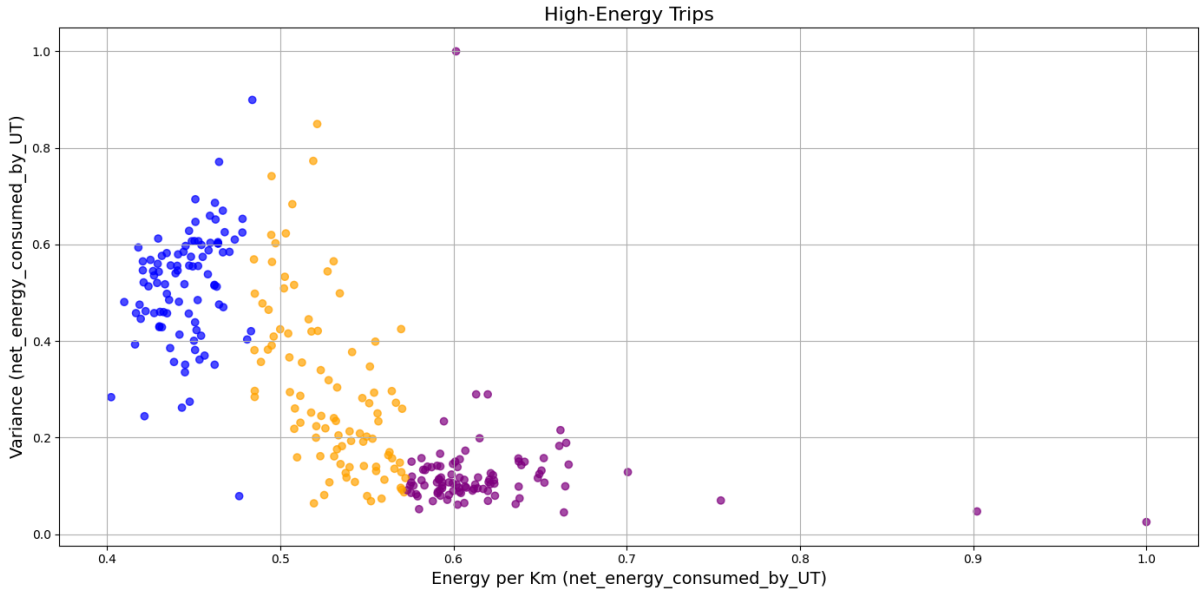
Medium Energy Trips: Variance 0 to 0.05 and Energy per KM: 0 to 0.4

High Energy Trips: Variance 0 to 0.4 and Energy per KM: 0 to 0.5

Using this method we identified 11 drivers who performed within our parameters. These drivers demonstrated consistent superior energy efficiency per KM across all kinds of trips. Below are the charts we developed to observe the driver efficiencies.







6. Further findings

As per the below chart, we were able to identify the top drivers against all the drivers based on comparing average energy per km and average energy per km (UT)

Comparison of average energy per km by year:

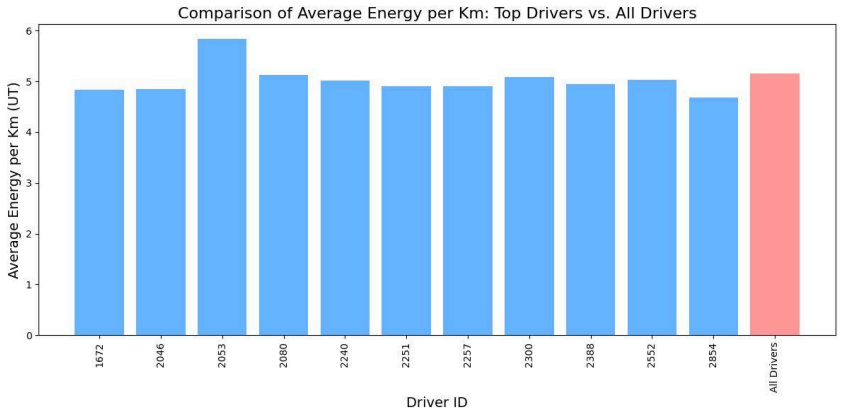
year	energy_per_km(All Drivers)	energy_per_km(Top 11)	energy_saved_per_km
2022	5.105	5.118	-0.0130
2023	5.210	4.984	0.2260

2024	5.150	4.956	0.1931

years	Total_energy_saved_per_year (kWh) top 11
2022	-27,271
2023	624,250
2024	186,767

Total energy saved per km over 3 years	0.41 kWh
Average energy saved per km per year	0.14 kWh
Total energy saved over the years	783,745 kWh
Total energy used for all drivers	30,074,719 kWh
Average energy per km for top 11 drivers	5.02 kWh
Average energy per km for all of the drivers	5.18 kWh

We saw an improvement with the top 11 drivers, however, out of the top 11 three drivers showed exceptional performance which could lead to even greater savings. The column details are si_siv_ndriver_def, and the exceptional drivers we identified have the following IDs: 1672, 2046, and 2854.



year	Top 3 average energy per km	All driver's average energy per km	Energy saved per km	Percentage energy saved

2022	4.91	5.11	259561	3.92%
2023	4.88	5.21	858548	6.43%
2024	4.56	5.16	533388	11.64%

7. Conclusions

The study conclusively demonstrates that focused driver training and performance monitoring can lead to substantial improvements in energy efficiency. The total energy savings over the three years is 1,651, 498 kWh and the percentage of energy savings over the three years will be 6.73%. Using the cost of electricity in Spain of €0.13 per kWh as per https://www.globalpetrolprices.com/Spain/electricity_prices/, a saving of approximately 200k euros can be achieved. If every driver drove similarly to the top 3 drivers based on the total kilometres driven from 2022 to 2024 the amount of savings would have been much higher.

Question 2 Report: Detection of Anomalous Energy Consumption in CAF Trains

8. Introduction to Detection of Anomalous Energy

Background and Context

Energy consumption in railway operations is a critical aspect of sustainable transport management. Efficient energy use not only reduces operational costs but also minimizes environmental impact. For CAF trains, understanding and managing energy consumption is vital for maintaining competitiveness and meeting regulatory standards. Identifying and addressing anomalous energy consumption can lead to significant improvements in efficiency, reliability, and cost-effectiveness.

9. Objectives

The primary objective of this analysis is to detect trains or equipment within the CAF fleet that exhibit anomalous energy consumption patterns. By pinpointing these anomalies, we aim to identify potential issues that may require maintenance or operational adjustments, thereby enhancing overall energy efficiency.

10. Scope

This report covers the analysis of energy consumption data for CAF trains from April 2022 to 2024. It includes data preprocessing, exploratory data analysis (EDA), anomaly detection using machine learning techniques, and visualization of the results. The focus is on detecting anomalies in various energy consumption metrics and providing actionable insights based on the findings. Limitations include the data's temporal scope and potential inaccuracies inherent in the dataset.

11. Methodology

11.1 Data Collection

The dataset used for this analysis was provided by CAF, spanning from April 2022 to 2024. It includes metrics such as net energy consumed, auxiliary energy consumed, traction energy, rheostatic energy, and regenerated traction energy. The data was pre-processed to ensure accuracy, including converting date columns to appropriate formats and handling missing values by imputing mean values for specific metrics like kilometres and avg_voltage.

11.2 Analytical Methods

1. **Descriptive Statistics**: Calculating basic statistical measures to understand the distribution and central tendencies of the data.
2. **Visualization**: Creating plots to visualize energy consumption patterns and identify potential outliers visually.
3. **Anomaly Detection**: Employing the Isolation Forest algorithm, a machine learning technique particularly suited for identifying outliers in high-dimensional data. This method isolates anomalies by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

11.3 Tools and Software

- **Programming Language**: Python
- **Libraries**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- **Environment**: Jupyter Notebook

12. Data Analysis

12.1 Descriptive Statistics

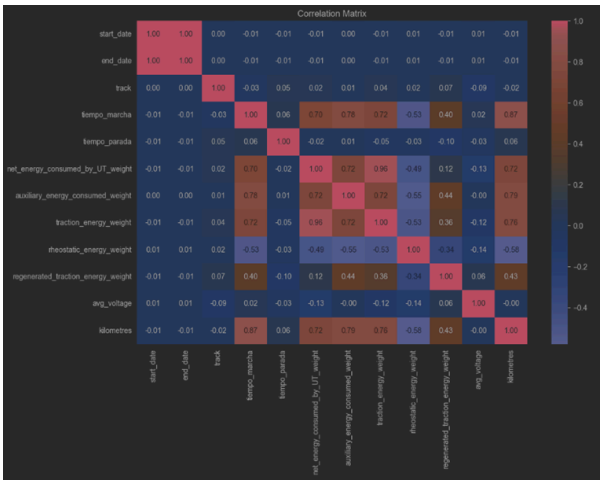
The dataset includes various metrics related to energy consumption and train operation. Summary statistics provided insights into the data distribution:

- **Net Energy Consumed**: Mean and standard deviation to understand typical energy usage and variability.
- **Auxiliary Energy Consumed**: Measures central tendency and dispersion.
- **Traction Energy**: Analysing energy used for train movement.
- **Rheostatic Energy**: Energy dissipated as heat during braking.
- **Regenerated Traction Energy**: Energy recovered during braking.

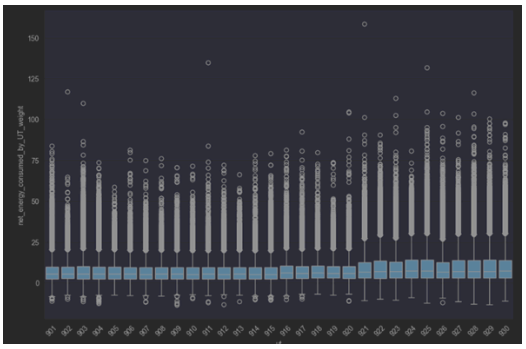
12.2 Visualization

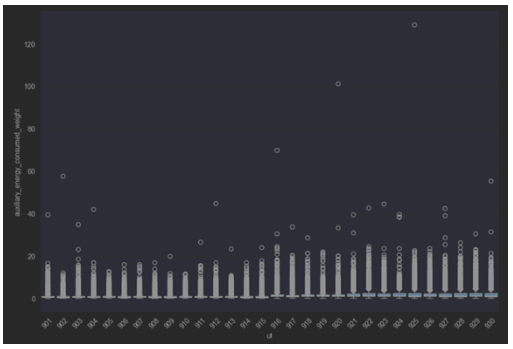
Several plots were created to illustrate energy consumption patterns:

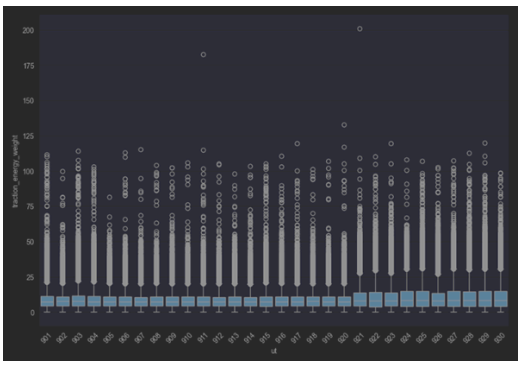
- **Correlation Heatmap**: Visualized relationships between different energy metrics, aiding in understanding potential interactions and dependencies.

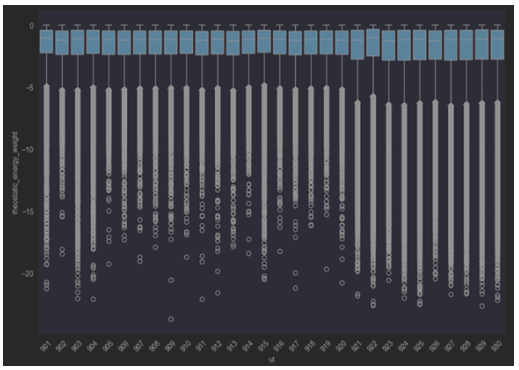


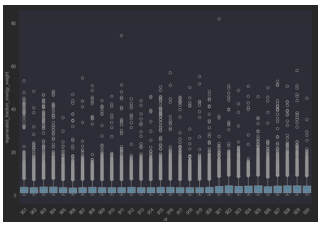
- **Box Plots**: Showed energy consumption metrics for each train, highlighting variations and potential outliers











12.3 Anomaly Detection

The Isolation Forest model was used to detect anomalies. Is a machine learning algorithm used for anomaly detection, particularly in detecting outliers in data. It works by creating an ensemble of isolation trees during the training phase, where each tree is trained on a random subset of the data and recursively splits the data based on randomly selected features.

- **Model Training**: The model was trained on key energy consumption metrics to learn normal consumption patterns.
- **Anomaly Scoring**: Each observation was assigned an anomaly score, with higher scores indicating more significant deviations from the norm.
- **Thresholding**: Anomalies were identified based on a predefined contamination level, indicating the proportion of outliers in the data.

```
param_grid = {
    'n_estimators': [100],
    'max_samples': ['auto'],
    'contamination': [0.05],
    'random_state': [42]
}
```

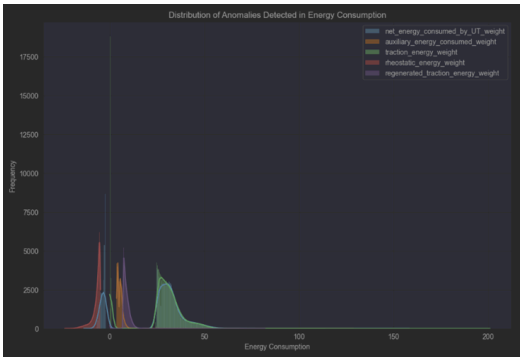
The Isolation Forest model was applied in different categories to detect anomali
The purpose of this was to ensure that only the trains that consistently showed
consumption of energy in multiple categories were detected as true anomalies.

- **Detecting outliers in energy consumption**: In this category the machine learning model identified anomalies in the data related to energy consumption features.

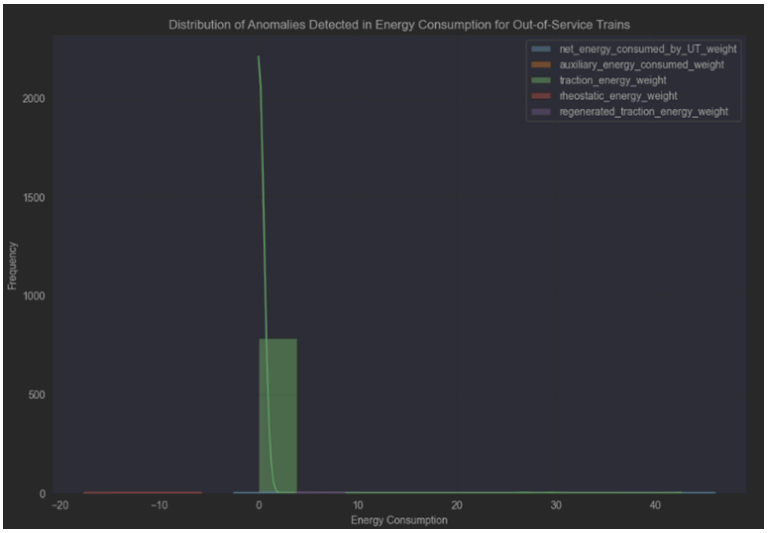
```
outlier_columns = ['net_energy_consumed_by_UT_weight', 'auxiliary_energy_consumed_weight', 'traction_energy_weight',
'rheostatic_energy_weight', 'regenerated_traction_energy_weight']
```

After training the model, it outputted the detected anomalies for each type of energy consumption

```
net_energy_consumed_by_UT_weight: 154470 anomalies detected out of 3162164 non-null values (4.88%)
auxiliary_energy_consumed_weight: 159951 anomalies detected out of 3162164 non-null values (5.06%)
traction_energy_weight: 158121 anomalies detected out of 3162164 non-null values (5.00%)
rheostatic_energy_weight: 158951 anomalies detected out of 3162001 non-null values (5.03%)
regenerated_traction_energy_weight: 151488 anomalies detected out of 3162164 non-null values (4.79%)
```



- **Detecting outliers in out-of-service trains**: In this category the machine learning model identified anomalies in the data related to out of service trains, labelling trains as a -1 if it was detected as an anomaly and labelled as a 1 if not.



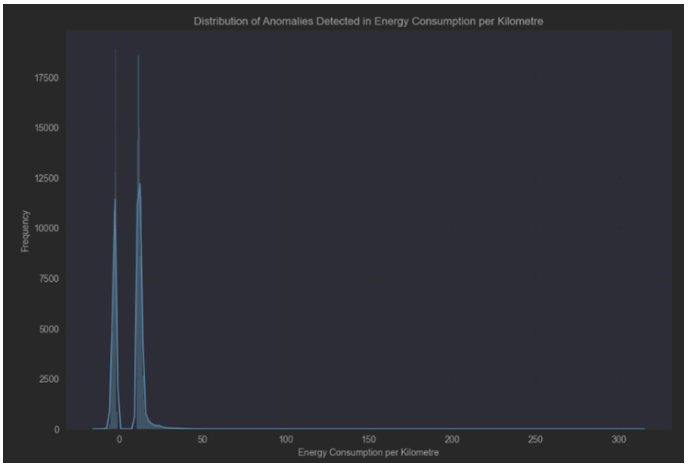
- **Detecting outliers in energy consumption per kilometre**: In this category the machine learning model identified anomalies in the data related to energy consumption per kilometre. After training the model, it outputted the detected anomalies.

```
Number of energy per km outliers: 157410
```

```
Average energy per km for each train (top 10):
    ut  avg_energy_per_km
0  924           4.758186
1  929           4.746744
2  930           4.737158
3  918           4.735066
4  925           4.718116
5  926           4.705356
6  928           4.698838
7  927           4.676596
8  919           4.652739
9  923           4.640602

Overall average energy per km: 4.46

Percentage of anomalous trips: 4.98%
```



- **Detecting outliers in energy consumption by route**: In this category the machine learning model identified anomalies in the data related to energy consumption by routes. After training the model, it outputted the detected anomalies.

```
Number of energy consumption outliers: 157795
```

```
Average energy consumption by route (top 20):
          current_station          next_station  avg_route_energy
49                Durango             Kukullaga         50.356766
110              Lurgorri  Amorebieta Geralekua         47.467674
42                  Deba                Arroa         46.016695
122               Muxika            Zugastieta         38.711485
18                Arroa                 Deba         35.610177
53              Elgoibar               Azitain         32.632523
58                 Ermua     Zaldibar Apeadero         31.726586
9    Amorebieta Geralekua             Kukullaga         30.601509
43                  Deba               Mendaro         30.379208
11   Amorebieta Geralekua            Zugastieta         29.024458
118              Mendaro           Toletxegain         28.145839
1              Aia-Orio              Usurbil         26.134400
152             Usurbil             Errekalde         24.673842
165              Zumaia               Zarautz         22.688095
123             Oiartzun               Bentak         22.280831
69                 Euba               Durango         20.878094
117              Mendaro                 Deba         20.799460
151             Usurbil              Aia-Orio         20.207184
159             Zarautz                Zumaia         19.515693
7             Amorebieta                 Euba         19.151778

Overall average energy consumption: 7.93

Percentage of anomalous routes: 4.99%
```

```
Top 10 routes with highest number of anomalies:
current_station  next_station
Aia-Orio         Usurbil             16488
Usurbil          Errekalde           15059
Ermua            Zaldibar Apeadero   13877
Elgoibar         Azitain             12435
Deba             Mendaro             11441
                 Arroa               10993
Arroa            Deba                10723
Usurbil          Aia-Orio             8797
Mendaro          Toletxegain          7399
Zarautz          Zumaia               6982
```

- **Detecting outliers n energy consumption by season**: In this category the machine learning model identified anomalies in the data related to energy consumption by season of the year. After training the model, it outputted the detected anomalies.

```
Number of seasonal energy outliers: 1

Seasonal energy consumption:
   season  season_energy  anomaly_season_energy
0  Autumn       7.869306                      1
1  Spring       7.975565                      1
2  Summer       7.805433                      1
3  Winter       8.068962                     -1

Details of seasonal energy outliers:
   season  season_energy
3  Winter       8.068962

Overall average energy consumption: 7.93
Season with highest energy consumption: Winter
Season with lowest energy consumption: Summer
```

13. Results and Findings

13.1 Summary of Findings

The analysis revealed several trains exhibiting anomalous energy consumption patterns. Key findings include:

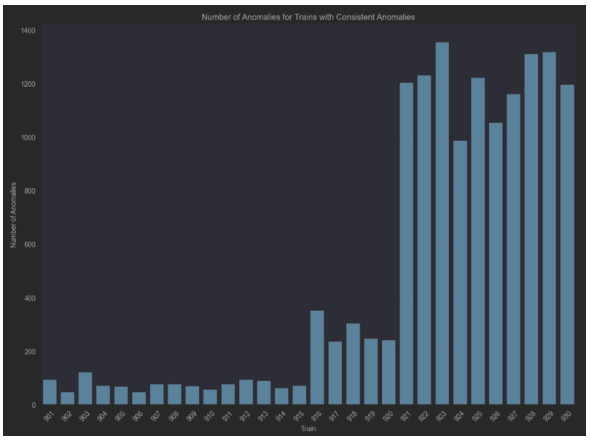
- **Outliers in Net Energy Consumption**: Some trains showed significantly higher or lower net energy consumption compared to the fleet average.
- **Auxiliary Energy Anomalies**: Certain trains had unusual auxiliary energy consumption, potentially indicating operational inefficiencies.
- **Traction Energy Variability**: A few instances of excessive traction energy usage were identified, suggesting possible mechanical issues or inefficient driving practices.

Statistical Significance

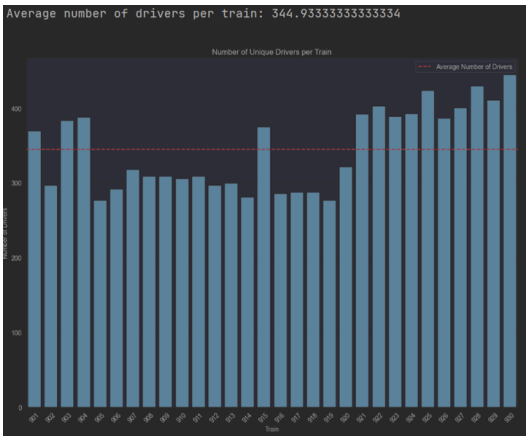
The detected anomalies were statistically significant, as they deviated substantially from the fleet's average energy consumption patterns. The Isolation Forest model effectively isolated these outliers, providing confidence in the robustness of the findings

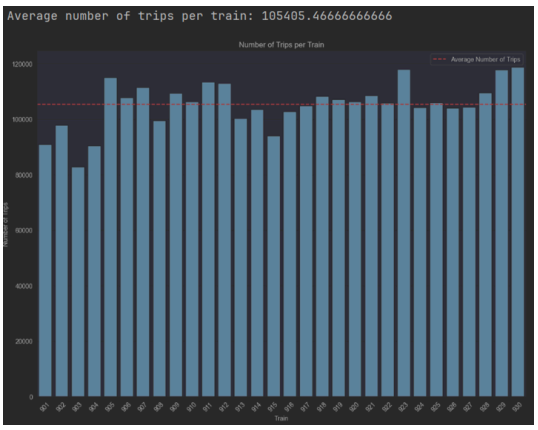
Top 10 trains with consistent anomalies:

	ut	num_anomalies
0	923	1356
1	929	1319
2	928	1311
3	922	1233
4	925	1223
5	921	1204
6	930	1197
7	927	1162
8	926	1054
9	924	988



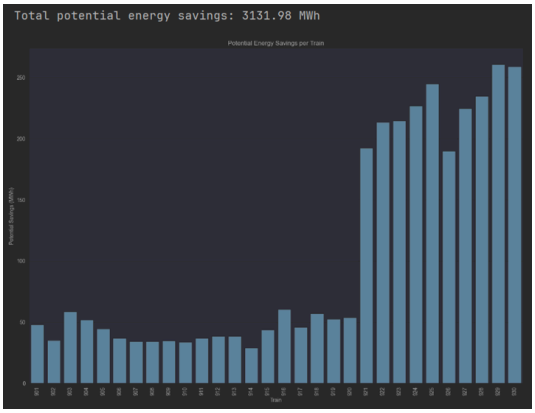
In order to compare trains fairly, an analysis on number of drivers and number of trips has been performed. This way, it can be ensured that the consumption of energy for the anomalous trains is not due to the skills of the driver or the number of trips it has performed.





Potential Savings

After detecting the anomalous trains, our group performed an analysis to calculate the potential savings the company could had have if these anomalies had been detected and maintenance had been performed



Top 10 trains with highest potential savings:

	ut	net_energy_consumed_by_UT_weight	normal_energy	potential_savings
28	929	1.189468e+06	928706.420363	260.761631
29	930	1.195801e+06	936814.896967	258.985753
24	925	1.078683e+06	833844.312972	244.838718
27	928	1.086858e+06	851859.827188	234.998467
23	924	1.049336e+06	822509.847852	226.825798
26	927	1.040051e+06	815167.059519	224.883483
22	923	1.106772e+06	892203.492861	214.568029
21	922	1.020892e+06	807347.933294	213.544457
20	921	1.010393e+06	818101.633626	192.290931
25	926	9.740088e+05	784124.126733	189.884685

Percentage of total energy that could be saved: 12.49%

Total potential savings in 2022: €639863.54

Total potential savings in 2023: €311945.22

Total potential savings for both years: €951808.76

The potential monetary savings were calculated using the average price per megawatt/hour according to Statista. https://es.statista.com/estadisticas/993787/precio-medio-final-de-la-electricidad-en-espana/

14. Discussion

14.1 Interpretation of Results

The anomalies detected in the analysis indicate potential issues with specific trains or operational practices. High net energy consumption could result from mechanical inefficiencies, while irregular auxiliary energy usage might point to faulty equipment. Excessive traction energy consumption could be due to suboptimal driving practices or technical problems.

14.2 Implications

Detecting these anomalies has several implications:

- **Operational Efficiency**: Addressing anomalies can lead to more efficient energy use, reducing operational costs.
- **Maintenance**: Identifying faulty equipment allows for timely maintenance, preventing more severe issues and ensuring reliability.
- **Energy Management**: Insights from the analysis can inform strategies for better energy management and sustainability.

14.3 Limitations

- **Data Quality**: Inherent inaccuracies in the dataset could affect the results.
- **Temporal Scope**: The analysis covers a specific period, and patterns may change over time.
- **Model Limitations**: The Isolation Forest model, while effective, may not capture all types of anomalies.

15. Recommendations

15.1 Actionable Steps

1. **Investigate Detected Anomalies**: Conduct detailed inspections of the trains identified as anomalies to determine the root causes.
2. **Enhance Monitoring Systems**: Implement continuous monitoring using the anomaly detection model to identify issues in real-time.
3. **Optimize Operations**: Use the insights to adjust operational practices, such as driving techniques, to improve energy efficiency.

15.2 Future Work

1. **Expand Data Collection**: Include additional data points and extend the analysis period for more comprehensive insights.

2. **Refine Models**: Explore other machine learning models and techniques to improve anomaly detection accuracy.
3. **Energy Efficiency Programs**: Develop targeted programs based on the findings to enhance energy efficiency across the fleet.

16. Conclusion

16.1 Recap of Key Points

This analysis successfully identified trains with anomalous energy consumption patterns within the CAF fleet. By utilizing data preprocessing, EDA, and an Isolation Forest model, we pinpointed specific instances of abnormal energy usage and provided actionable insights.

16.2 Final Thoughts

The findings from this project underscore the importance of continuous monitoring and proactive management of energy consumption in railway operations. Addressing detected anomalies not only improves operational efficiency but also contributes to sustainability and cost savings. Future efforts should focus on expanding data collection and refining analytical techniques to build on these insights.