Discovery Unit
Unlocking the power of data

Sales & Demand Analysis Technical Report

Disclaimer: To protect client confidentiality, the data presented in this report has been intentionally adjusted. All numerical values have been modified to use fictional figures, and any geographic, product-specific, or sales location information has been generalised. These changes preserve the analytical value of the work while ensuring anonymity and privacy.

Business Context

A data science consultancy partnered with a large retailer and consumer goods manufacturer to address supply and demand challenges driven by climate change. In 2022, a severe European heatwave led to a widespread shortage of temperature-sensitive goods, leading to significant missed sales opportunities and operational disruption. This project analyses sales, temperature, and event data from 2021–2022 across retail locations within a major metropolitan area in Europe to uncover patterns influencing demand.

Key questions guiding the analysis include:

Temperature & Weather

- What temperature thresholds correlate with spikes in sales?
- How do weather conditions affect purchasing behaviour?

Time & Location

- How do sales vary by day of week, month, season?
- How do holidays and events influence demand?

Demand Forecasting

- Can historical patterns predict future demand for better stock planning?

Operational Efficiency

- What is the financial impact of stockouts versus overstocking?

Key Questions Summary

- How can we balance supply with demand forecasting?
- What external factors should drive inventory decisions?
- How can profitability be optimised while ensuring product availability?
- The findings will support more resilient, data-driven supply chain strategies.

Project Development

Tools and Libraries

Python was used in Colab for analysis and modeling; Tableau supported visuals.

Core libraries included pandas, numpy, seaborn, matplotlib, with forecasting via HistGradientBoosting, and time series split using scikit-learn. Feature engineering and evaluation used scikit-learn, statsmodels, and meteostat[1] weather data.

Data Cleaning

Sales Data (2021 & 2022):

- Dates standardised (YYYY-MM-DD) with added features (e.g. weekday, season).
- No duplicates/missing values; negative sales confirmed as refunds.
- Columns renamed for consistency; locations cleaned and standardized.
- Filtered to stations within the major metropolitan area (Appendix 1).
- 2022 has fewer days (348 vs. 365 in 2021).

Temperature Data (2021 & 2022):

- Dates cleaned with derived features; columns renamed for clarity.
- Dropped empty columns; no duplicates.
- Retained missing values (e.g. avg_temp, pressure) to avoid bias.
- Outliers kept, as they reflect real weather (Appendix 2).

Final Preparation:

2021-2022 sales and temperature data were cleaned and merged to analyse climate impact on sales.

External data on public holidays and major sporting events was also added.

[1]

Assumptions:

- Products sold in 2021 and 2022 are assumed to be standard product.
- Standard products are sold at ~ €1.50 with a profit margin of ~ €0.20.
- Premium products are sold at ~ €1.90 with a margin of ~ €0.40.
- Missing daily data for certain sales locations indicates zero sales on those days.

Limitations:

- Seasonality is difficult to assess due to missing sales data from 15–31 December 2022.
- Sales limitations from stock shortages are not documented.
- No data is available on other sales points, limiting context on demand fluctuations at sales locations.

Visualisations & Insights

Exploratory Analysis

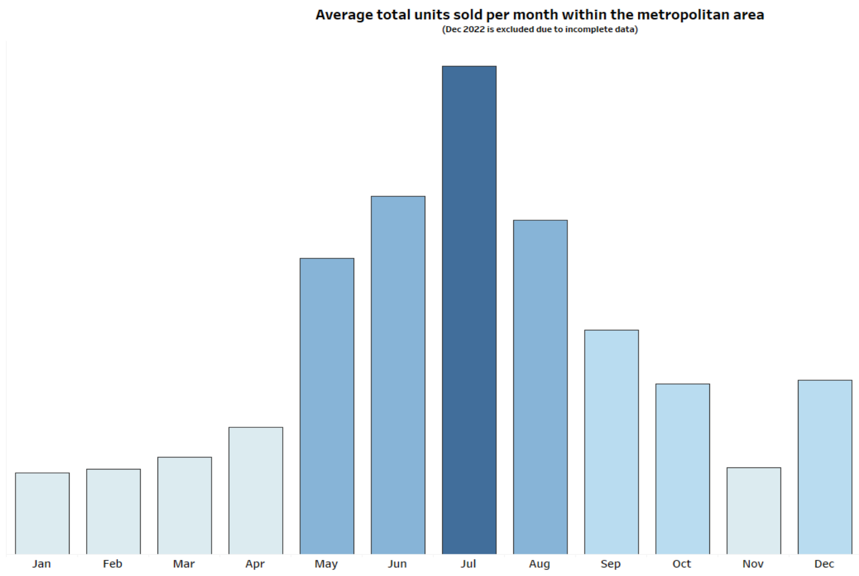
We began by analysing sales and temperature trends across country and metropolitan area, identifying top-selling locations and evaluating the impact of events and holidays. This guided deeper dives into district-level sales, North-South area differences, and the influence of local events.

Visualisations:

- Colour was used to highlight contrasts (e.g., sales vs. temperature).
- Labels aided quick insights; axes/grid lines were removed if unhelpful.
- Maps were used to show geospatial aspects, with colour and size indicating sales volume. Filters allowed detailed data exploration, and a video illustrated how locations disappeared as sales grew.

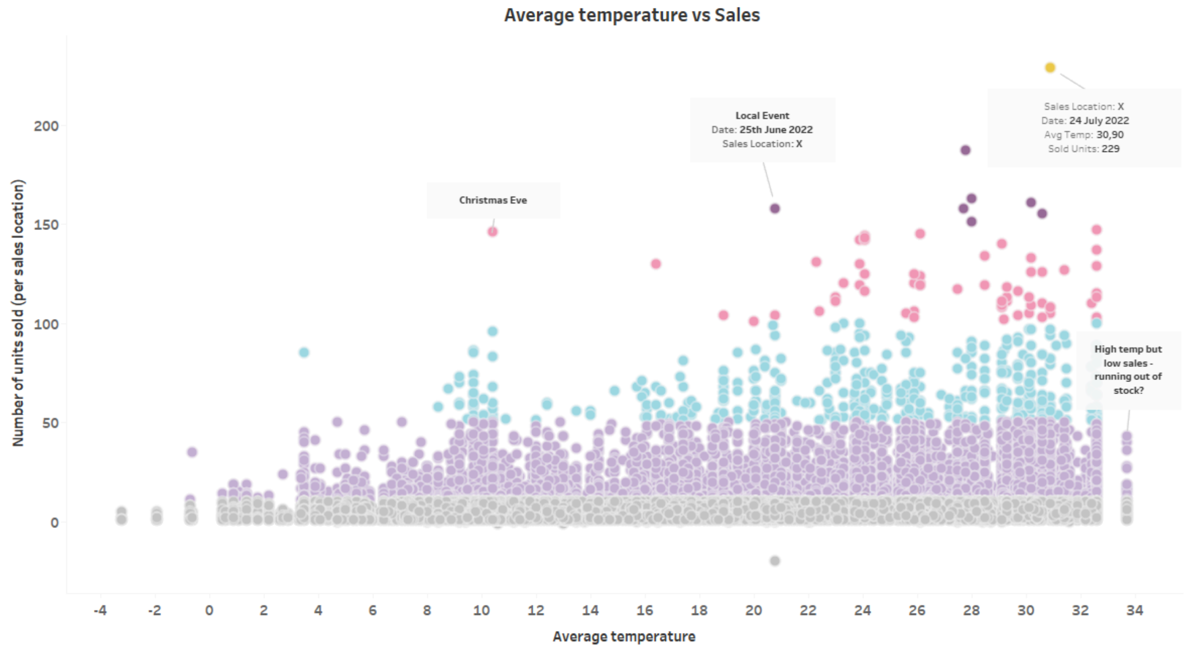
Sales Sensitivity to Temperature Trends:

1. **Seasonal temperature patterns and peak sales**:
   A monthly bar chart shows sales peaking in July, confirming summer heat drives strong seasonal demand.



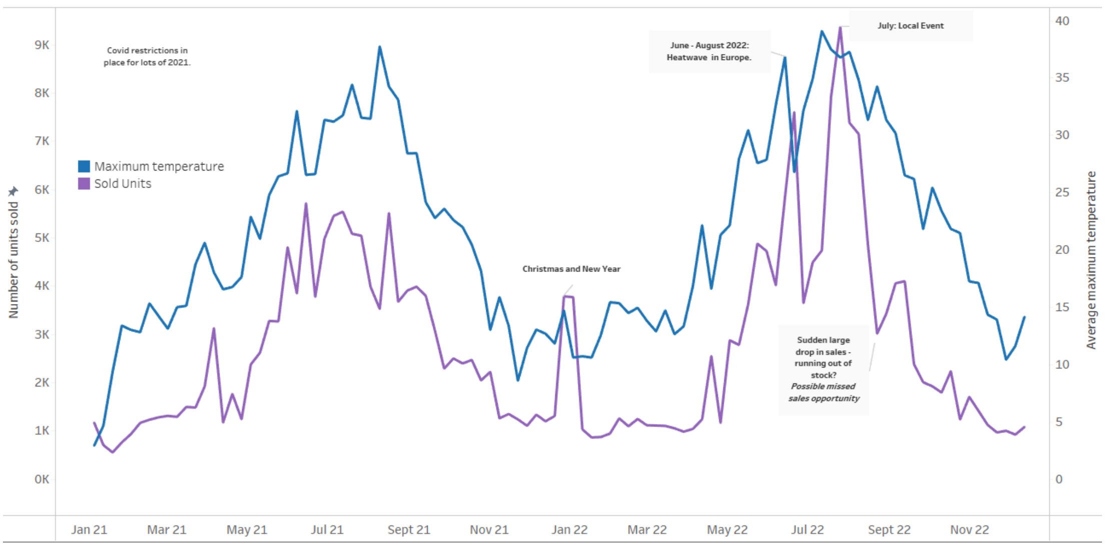
2. **Temperature thresholds and sales density**:

Sales rise above 28℃, suggesting a demand tipping point, but drop beyond 34℃ - likely due to stockouts. High heat does not guarantee high sales, as many locations still show low volumes.



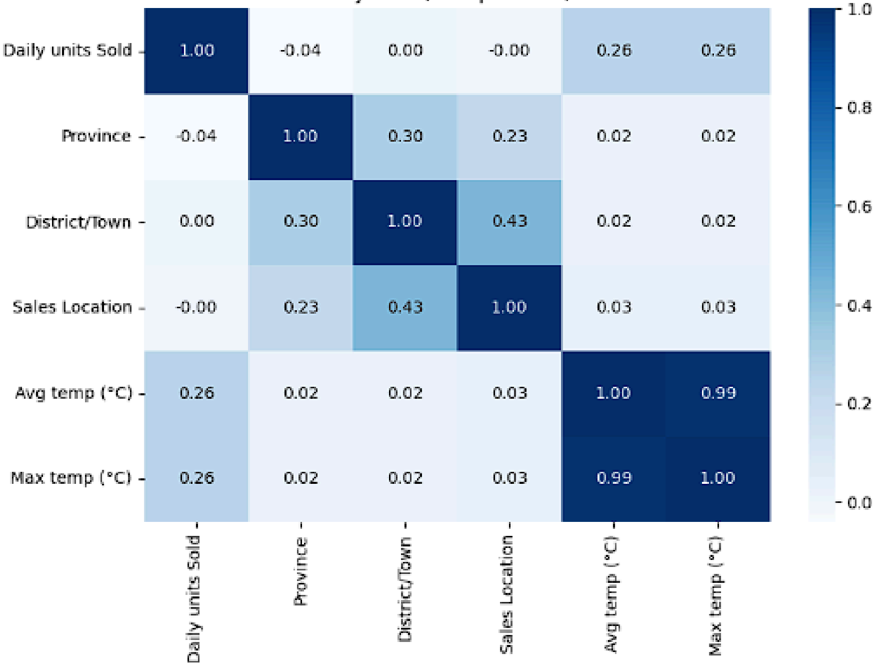
3. **Temporal interaction and potential stockouts:**
   The chart shows sales rising with temperature, but two sharp drops despite continued heat suggest missed opportunities or stockouts. A smaller cold-weather spike indicates temperature is not the only sales driver.

Weekly number of sold units and average weekly maximum temperature

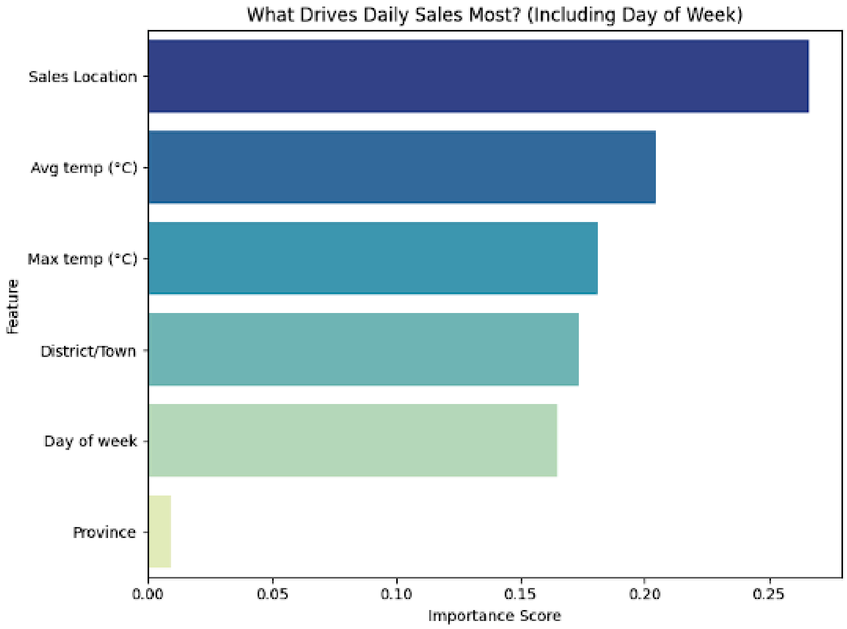


4. **Multivariate Relationships Driving Sales:**
   Correlation analysis links sales to temperature, while Random Forest's feature importance highlights timing and location as key drivers, showing sales depend on more than just weather.

Correlation Matrix - Daily Sales, Temperature, and Location Features



	Daily units Sold	Province	District/Town	Sales Location	Avg temp (°C)	Max temp (°C)
Daily units Sold	1.00	-0.04	0.00	-0.00	0.26	0.26
Province	-0.04	1.00	0.30	0.23	0.02	0.02
District/Town	0.00	0.30	1.00	0.43	0.02	0.02
Sales Location	-0.00	0.23	0.43	1.00	0.03	0.03
Avg temp (°C)	0.26	0.02	0.02	0.03	1.00	0.99
Max temp (°C)	0.26	0.02	0.02	0.03	0.99	1.00

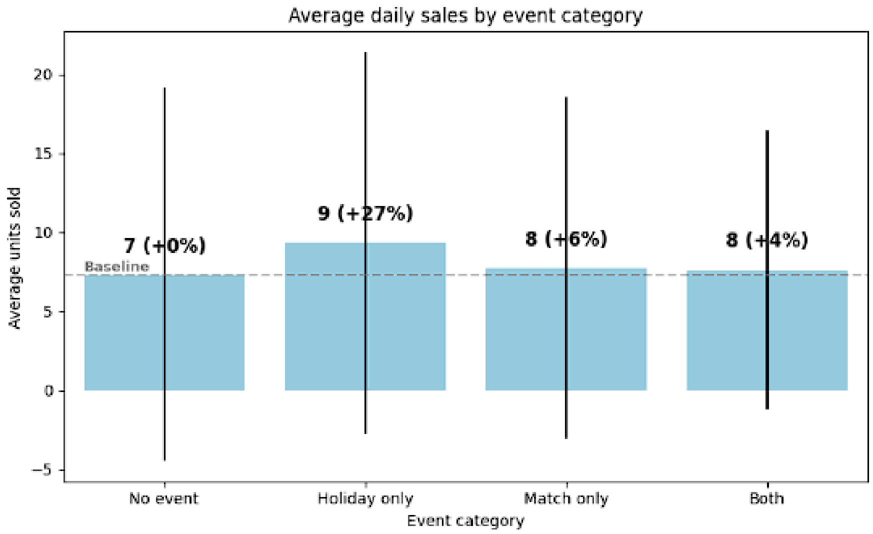


Sales, location, and event-driven demand:

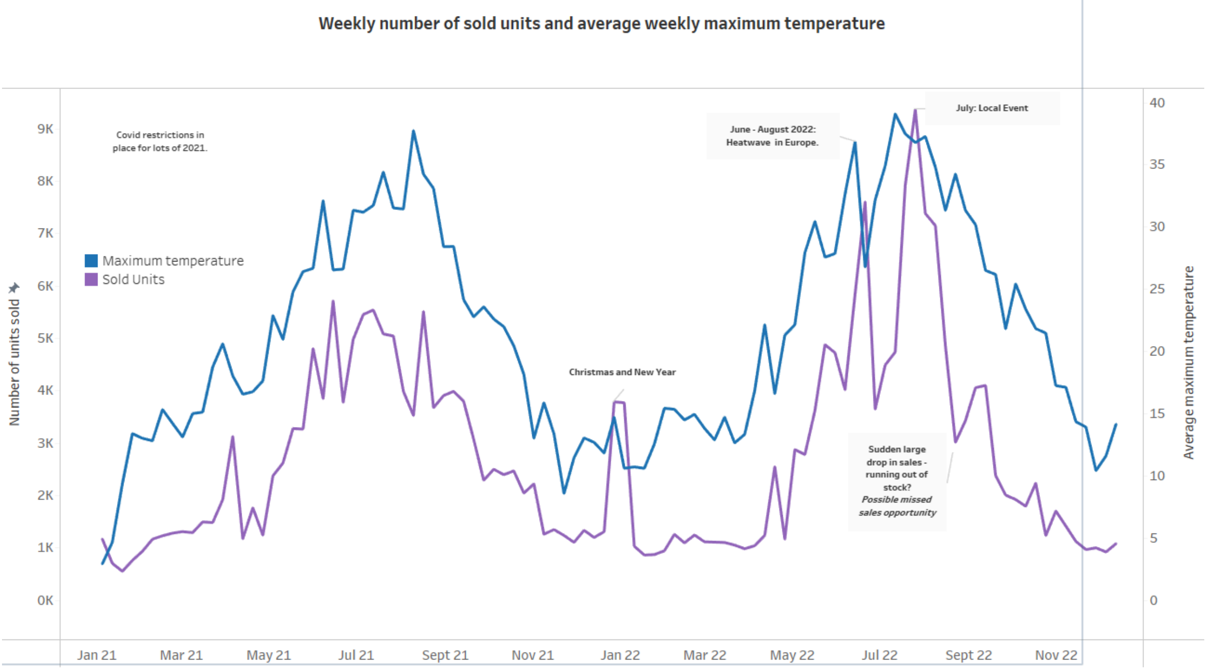
1. **Impact of location on sales**: A map was created in Tableau to display location of sales locations and volume of sales at each site (using size and colour density). There is a higher proportion of sales in the North, suggesting that it should be a higher priority.

(Maps were removed to ensure the anonymity of our client)

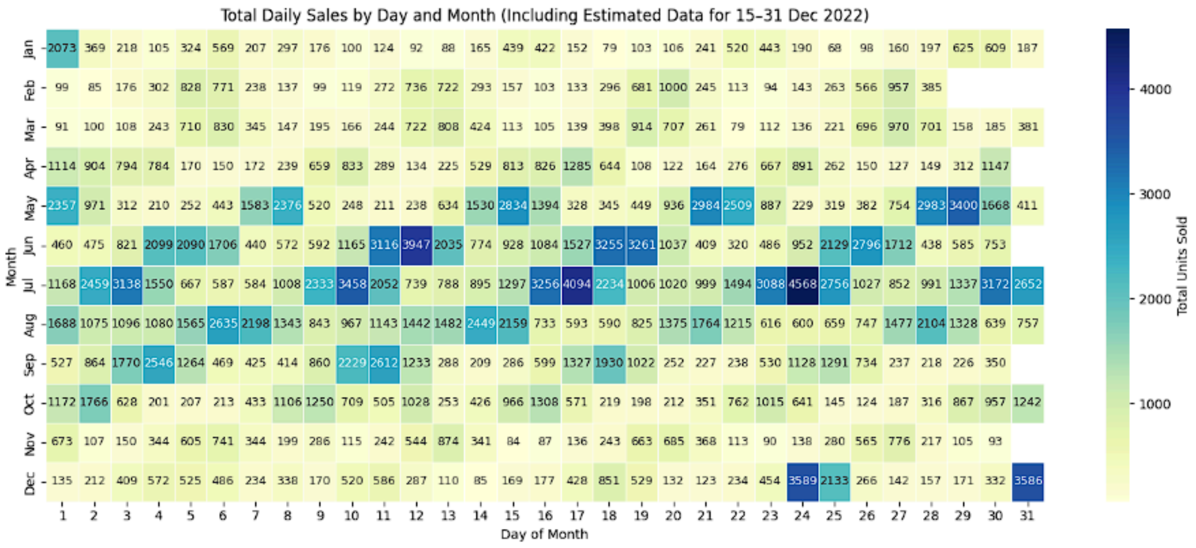
2. **Impact of events and holidays on sales**: Holiday-only days showed the highest average sales, with a 27% uplift over non-event days. A bar chart with error bars was used to clearly compare group averages while visualising variability across categories.



- Some specific events have a significant impact on sales.



3. **Calendar-based sales trends**: A day-by-month heatmap revealed peak sales during summer and major holidays, like Christmas and New Year, reflecting alignment with leisure-driven consumer behaviours. The heatmap enabled visualisation of temporal trends across the calendar year.
    - Calendar-style heatmaps for 2022 and 2021 (see Appendix 3) confirmed seasonal and weekly consistency but were excluded from the presentation as it did not highlight individual holiday impact - the focus of our heatmap analysis.
    - Day-of-week chart reveals that ~60% of sales are from Fridays to Sundays, with ⅓ from Saturdays (Appendix 3).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Jan	2073	369	218	105	324	569	207	297	176	100	124	92	88	165	439	422	152	79	103	106	241	520	443	190	68	98	160	197	625	609	187
Feb	99	85	176	302	828	771	238	137	99	119	272	736	722	293	157	103	133	296	681	1000	245	113	94	143	263	566	957	385			
Mar	91	100	108	243	710	830	345	147	195	166	244	722	808	424	113	105	139	398	914	707	261	79	112	136	221	696	970	701	158	185	381
Apr	1114	904	794	784	170	150	172	239	659	833	289	134	225	529	813	826	1285	644	108	122	164	276	667	891	262	150	127	149	312	1147	
May	2357	971	312	210	252	443	1583	2376	520	248	211	238	634	1530	2834	1394	328	345	449	936	2984	2509	887	229	319	382	754	2983	3400	1668	411
Jun	460	475	821	2099	2090	1706	440	572	592	1165	3116	3947	2035	774	928	1084	1527	3255	3261	1037	409	320	486	952	2129	2796	1712	438	585	753	
Jul	1168	2459	3138	1550	667	587	584	1008	2333	3458	2052	739	788	895	1297	3256	4094	2234	1006	1020	999	1494	3088	4568	2756	1027	852	991	1337	3172	2652
Aug	1688	1075	1096	1080	1565	2635	2198	1343	843	967	1143	1442	1482	2449	2159	733	593	590	825	1375	1764	1215	616	600	659	747	1477	2104	1328	639	757
Sep	527	864	1770	2546	1264	469	425	414	860	2229	2612	1233	288	209	286	599	1327	1930	1022	252	227	238	530	1128	1291	734	237	218	226	350	
Oct	1172	1766	628	201	207	213	433	1106	1250	709	505	1028	253	426	966	1308	571	219	198	212	351	762	1015	641	145	124	187	316	867	957	1242
Nov	673	107	150	344	605	741	344	199	286	115	242	544	874	341	84	87	136	243	663	685	368	113	90	138	280	565	776	217	105	93	
Dec	135	212	409	572	525	486	234	338	170	520	586	287	110	85	169	177	428	851	529	132	123	234	454	3589	2133	266	142	157	171	332	3586

Modeling

Can historical patterns predict future demand for better stock planning?

We tested a number of progressively more powerful and sophisticated machine learning models, each addressing the limitations of the previous. We selected the Histogram-based Gradient Boosting (HGB) paired with a Monte Carlo simulation as our final forecasting approach.

Data Preparation

To ensure optimal conditions for machine learning, preparation included:

- Insert row per date for each location where gaps, with assumed 0 sales (26,000 rows)
- Applying cyclical encoding to preserve continuing of temporal variables
- Encoding features with rolling mean to reduce noise and highlight seasonality
- Creating lag variables to capture temporal dependencies
- Retrieving temperature normals from Meteostat and filling gaps using KNNImputer

Please refer to Appendix 4 for the findings from the Decision Tree, Random Forest and the Extreme Gradient Boosting model (XGBoost).

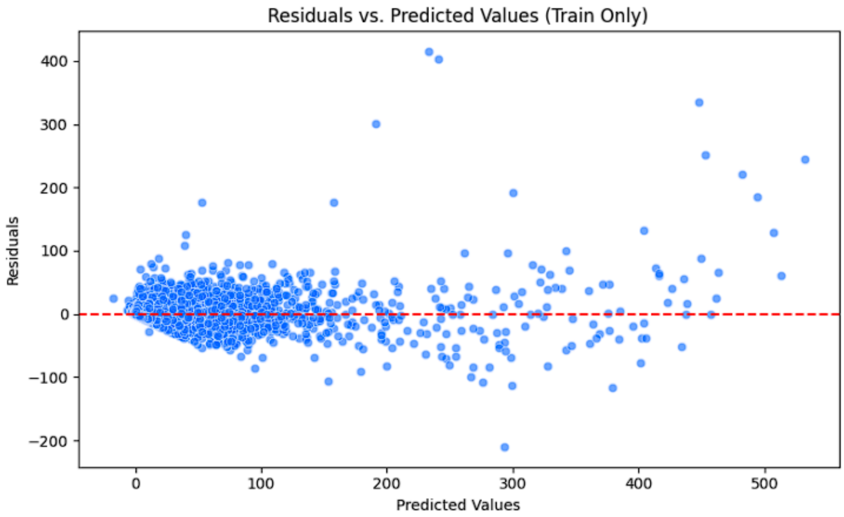
HGBoost

The final HistGradientBoosting Model showed clear improvement after aggregating sales data by town/city, reducing noise from sales locations with frequent zero-sales days. Across cross-validation folds, it achieved an average R² of 0.66, RMSE of

16.05, and MAE of 6.38, indicating stronger predictive accuracy than previous versions. Overfitting persists, as evidenced by a training R² of 0.89 versus a test R² of 0.62 in the final fold, likely stemming from the limited data range (2021–2022).

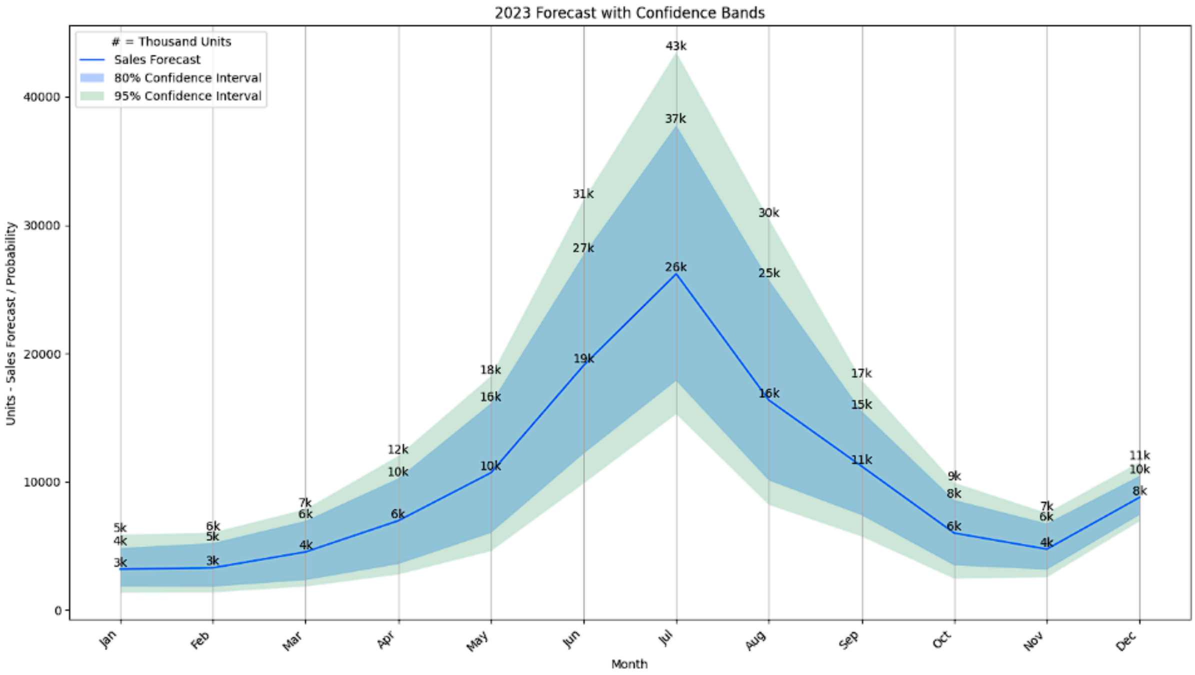
Key features included town/city and day-of-week patterns, while weather variables had minimal impact. Binary flags for holidays and recurring weekly/monthly events helped capture predictable demand shifts. Residuals were generally well distributed with a slight leftward skew (see image below), suggesting minor underprediction but overall improved generalization and reduced bias compared to prior models.

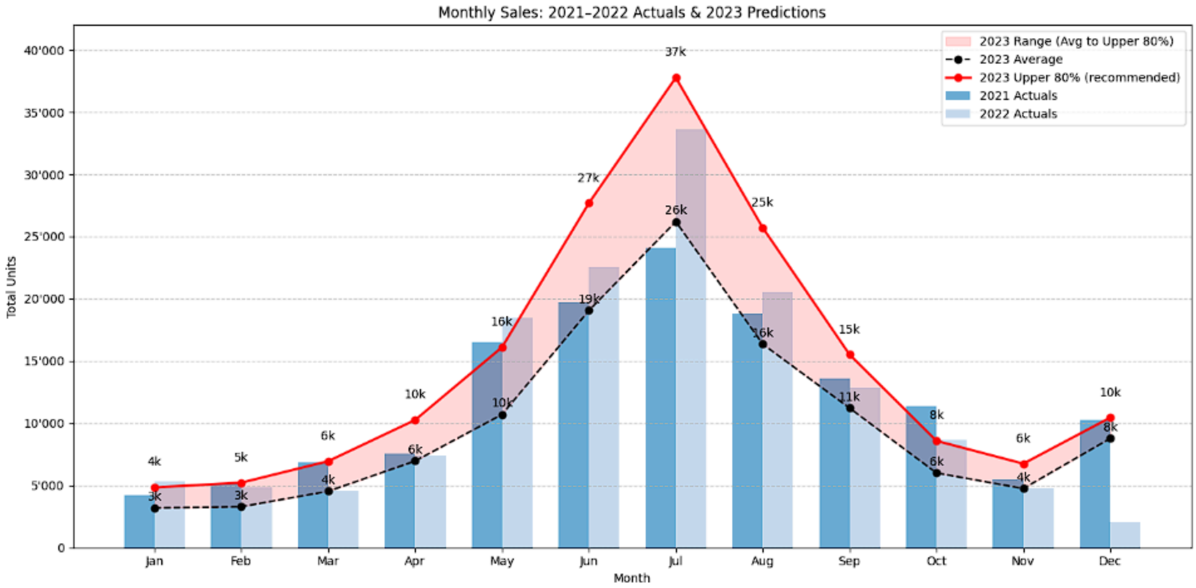
While residuals indicate a generally good fit, the model’s high Mean Absolute Percentage Error (MAPE) of 169.66% reflects considerable variability or scale issues in the target data. This limitation could be mitigated by adding more historical data to improve generalization and reduce overfitting.



Monte Carlo simulation

To project demand under future weather scenarios, we used Monte Carlo to simulate 1000 weather outcomes based on 30 years of historical Meteostat data—balancing computational efficiency with prediction reliability. Applying our Boost model to these simulations, we generated a 2023 demand forecast with 80% and 95% confidence intervals, capturing likely variability in sales:





Revenue Optimisation

Building on the model forecasts, we aimed to optimise revenue generation by applying a further layer of logic to dynamically adjust the sales strategy, accounting for event-driven demand and weather forecast data. This allows us to also compute corresponding revenue and expected profit.

We assumed a base set of product parameters based on typical industry assumptions.

The key steps of optimisation include:

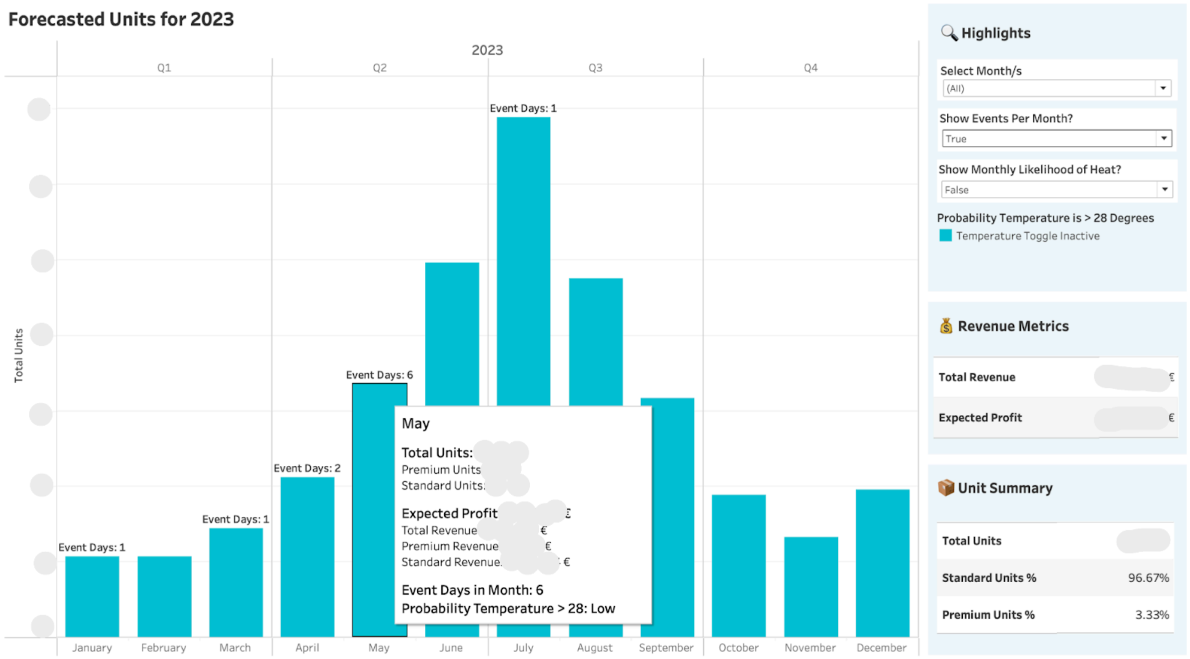
1. Demand Estimation: Uses the 95% confidence interval for peak event days otherwise the 80% confidence interval is used to estimate total units sold.

2. Heat Signal Calculation: Combines the probability of heat events with a boost based on expected temperature. Applies a dampening factor of 30% for highly uncertain temperature forecasts which deviate from the typical temperature (z_score > 3).

3. Product Ratio Adjustment: Gradually increases the premium product ratio above a defined probability and temperature threshold (capped at 10% of all units).

4. Revenue & Profit Estimation: Computes expected revenue and profit based on adjusted product mix and pricing/margin assumptions.

The final optimised forecast was visualised in a dashboard via Tableau to demonstrate the effect of events and temperature variables on the final forecast via the implementation of filters.



May

Total Units: [redacted]
Premium Units [redacted]
Standard Units. [redacted]

Expected Profit [redacted] €
Total Revenue [redacted] €
Premium Revenue [redacted] €
Standard Revenue. [redacted] €

Event Days in Month: 6
Probability Temperature > 28: Low

Highlights

Select Month/s: (All)

Show Events Per Month?: True

Show Monthly Likelihood of Heat?: False

Probability Temperature is > 28 Degrees

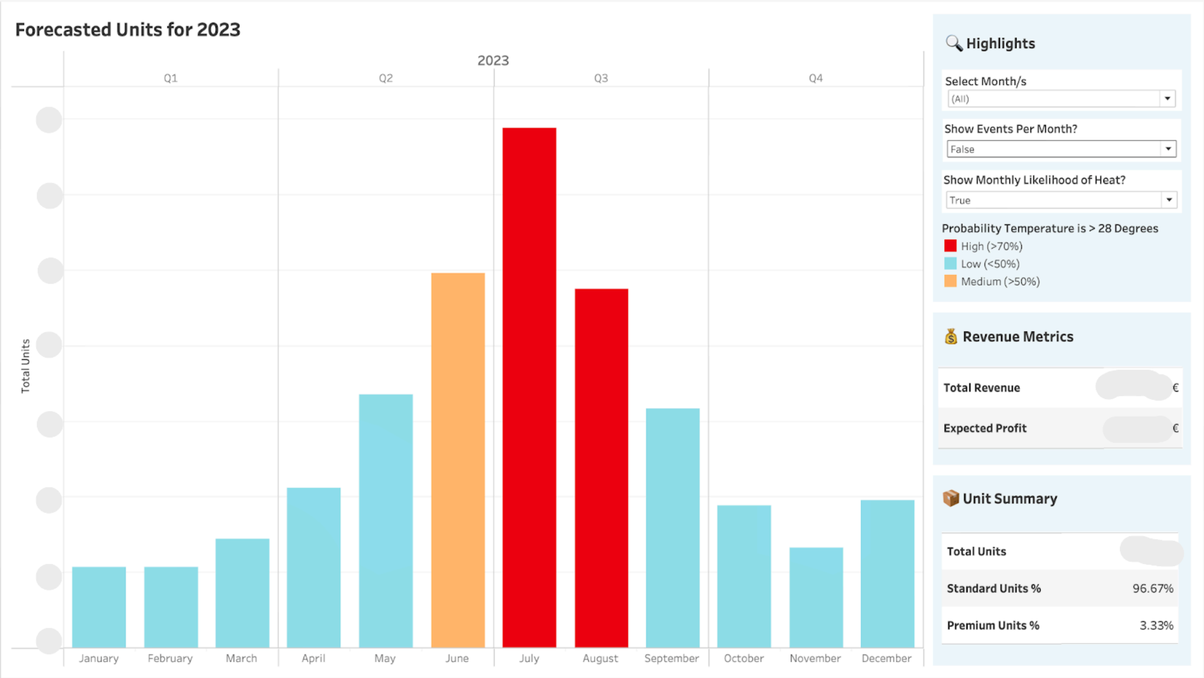
🟦 Temperature Toggle Inactive

Revenue Metrics

Total Revenue	[redacted] €
Expected Profit	[redacted] €

Unit Summary

Total Units	[redacted]
Standard Units %	96.67%
Premium Units %	3.33%



🔍 Highlights

Select Month/s: (All)

Show Events Per Month?: False

Show Monthly Likelihood of Heat?: True

Probability Temperature is > 28 Degrees
- 🟥 High (>70%)
- 🟦 Low (<50%)
- 🟧 Medium (>50%)

💰 Revenue Metrics

Total Revenue	[redacted] €
Expected Profit	[redacted] €

📦 Unit Summary

Total Units	[redacted]
Standard Units %	96.67%
Premium Units %	3.33%

Strategic Recommendations

Forecasting:

- Use forecasting model to plan production volumes for peak times (weekends, summer, Christmas, New Year and major events)
  - Use the upper 80% confidence band as standard, increase to 95% for peak events.
- Use revenue optimiser to adjust and plan distribution of premium products
  - Recommend trial at 10% cap for first year.

Geospatial:

- Align distribution schedules with weekly and seasonal cycles ensuring stock for peak periods such as weekends and events
- Prioritise stock and delivery to sales locations in the north which have nearly 75% of all sales
- Prioritise delivery to the locations with sales on over 75% of days
- Stop distributing to the locations with sales on less than 25% of days

Technical Recommendations:

- Add additional historical sales data to the model to improve the predictive power and accuracy
- Differentiate between premium and standard products in data collection
- Include stock-flow data to test assumptions on missed sales opportunities

Appendix 1 - Geographic Filtering

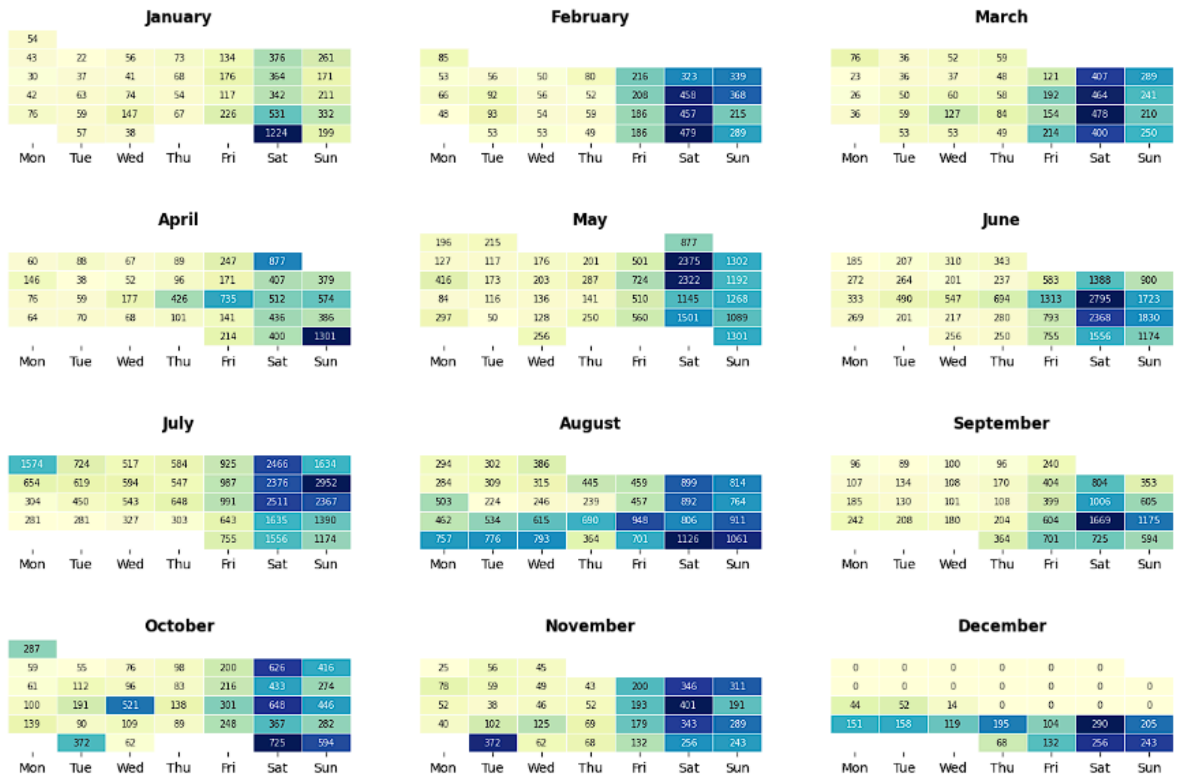
We used postcode-based radius filtering. Minor margin-of-error acknowledged due to measurement method.

Appendix 2 - Outlier Analysis

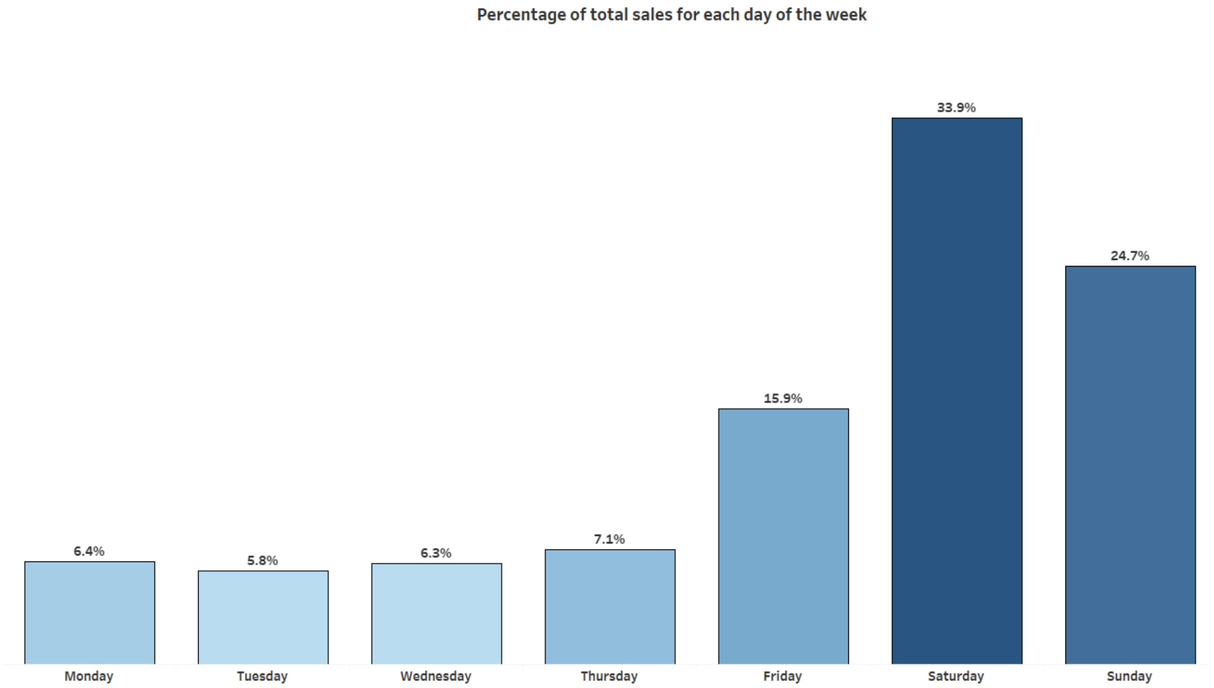
- Precipitation: Max value identified as outlier but retained as plausible during storm events.
- Wind Speed & Pressure: High-end values flagged as outliers but consistent with known weather extremes for the metropolitan area.
- Temperature: All deemed realistic.

Appendix 3 - Other Visualisations

Calendar-style heatmaps:



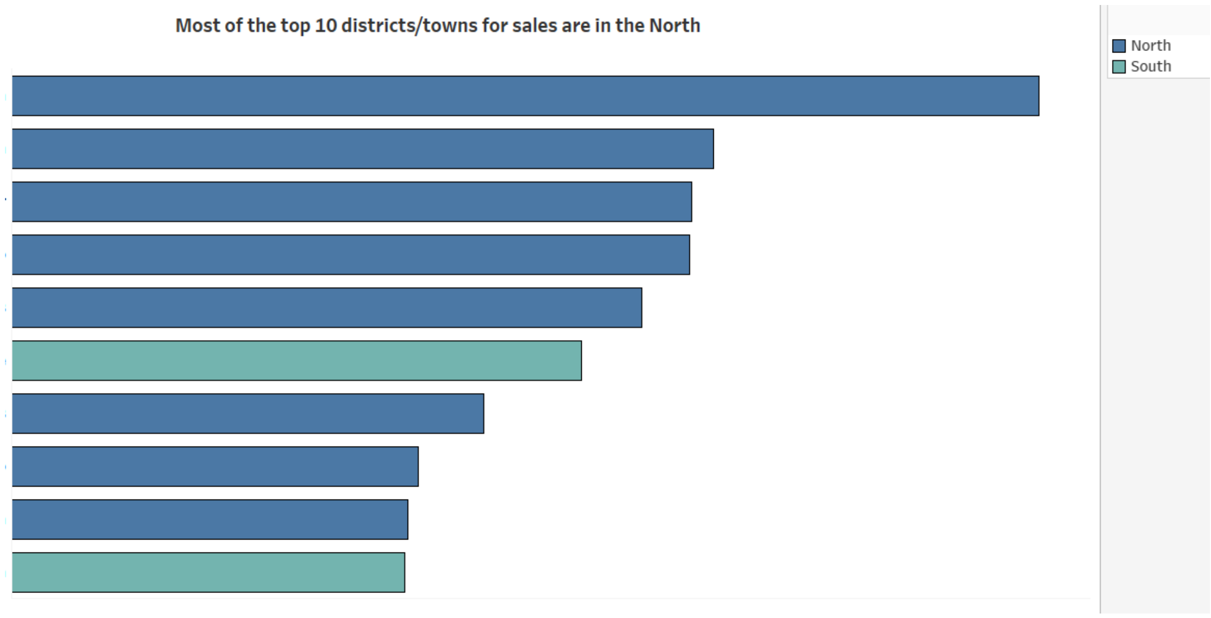
Day-of-week chart:



Percentage of days that have sales map:

(Maps were removed to ensure the anonymity of our client)

Number of total sales in the top 10 districts in the metropolitan area or the towns surrounding it:



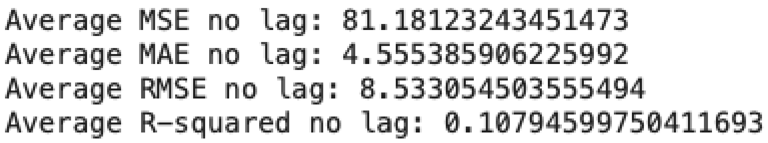
Appendix 4 - Models tested

Decision Tree

We initially tested a **Decision Tree model** to quickly identify non-linear relationships between non-linear variables, and to determine which are most predictive. The model utilised all provided weather data, location variables, dates plus external data gathered on public holidays and national and major sporting events with regional relevance.

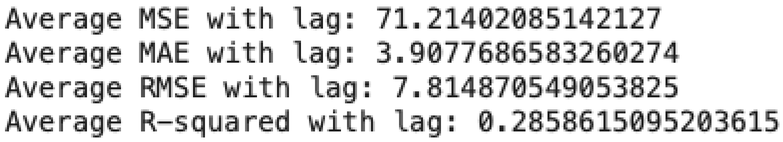
Timeseriessplit using scikit was used, with 5 splits. To attempt to improve the model various iterations were tested:

Results of first model - no adjustments:



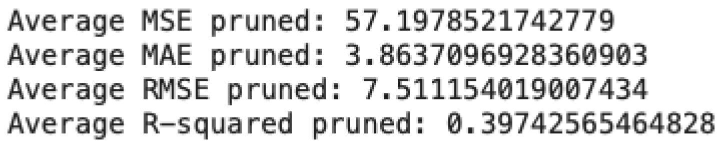
```
Average MSE no lag: 81.18123243451473
Average MAE no lag: 4.555385906225992
Average RMSE no lag: 8.533054503555494
Average R-squared no lag: 0.10794599750411693
```

Results when lag and rolling features were added for sold units over 1, 2, 7, 14 days:



```
Average MSE with lag: 71.21402085142127
Average MAE with lag: 3.9077686583260274
Average RMSE with lag: 7.814870549053825
Average R-squared with lag: 0.2858615095203615
```

Results when model pruned to dept of 5, with only 3 splits:



```
Average MSE pruned: 57.1978521742779
Average MAE pruned: 3.8637096928360903
Average RMSE pruned: 7.511154019007434
Average R-squared pruned: 0.39742565464828
```

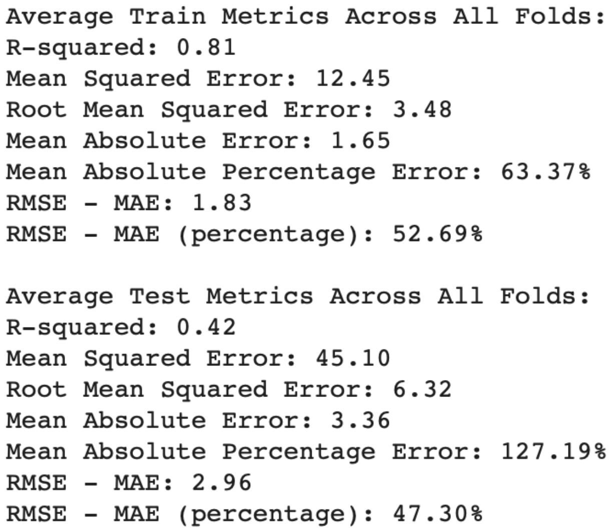
Whilst the accuracy and prediction capability improved with each adjustment, the performance is not strong enough to use for the complex forecasting needed in this project and a decision was quickly taken to test with a Random Forest model.

Random Forest

In preparing the data for the Random Forest Model various engineered features were added, including: lag-features for climate and sold_units, target encoding location date (sales location and company codes and town/city).

The performance was tested with various splits, estimators, depths and leafs and a final model run with the best hyperparameters.

Final results:



```
Average Train Metrics Across All Folds:
R-squared: 0.81
Mean Squared Error: 12.45
Root Mean Squared Error: 3.48
Mean Absolute Error: 1.65
Mean Absolute Percentage Error: 63.37%
RMSE - MAE: 1.83
RMSE - MAE (percentage): 52.69%

Average Test Metrics Across All Folds:
R-squared: 0.42
Mean Squared Error: 45.10
Root Mean Squared Error: 6.32
Mean Absolute Error: 3.36
Mean Absolute Percentage Error: 127.19%
RMSE - MAE: 2.96
RMSE - MAE (percentage): 47.30%
```

From investigations of this model we can see that the generalisation is poor. There were also additional data leakage issues, which we would later address in our boosted models. Dropping additional features and undertaking additional feature engineering could improve this model, but as a random forest overall performance is slow and time consuming. We decided that this is not an appropriate fit for a production model which will need to run in a reasonable amount of time and we progressed with further exploration using a boosted tree model.

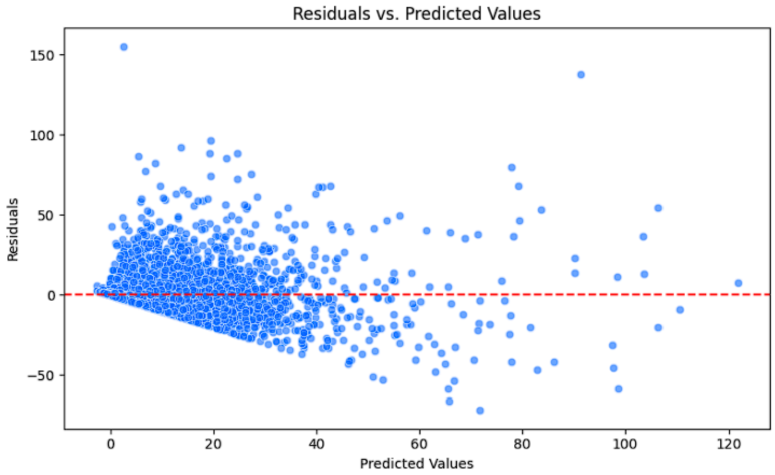
XGBoost

We explored **XGBoost** due to its strong performance on tabular data and its ability to model complex, non-linear relationships. Using a comprehensive feature set of 35 engineered variables—including temporal, location-based, weather, and event-related features—XGBoost achieved the following average performance across three time-based folds:

- **RMSE**: 7.02
- **MAE**: 3.49
- **R²**: 0.47
- **MAPE**: 134.15%

Despite extensive tuning—testing different hyperparameters, feature subsets, and feature eliminations—performance did not improve meaningfully. Analysis of the residuals (see image below) revealed clear patterns, indicating that the model was struggling to capture important dynamics in the data. In addition, signs of **overfitting** suggested limited generalization to unseen time periods.

This led to a shift in approach: we **restarted from scratch** with a smaller, more focused feature set and found that **Histogram-Based Gradient Boosting (HGB)** consistently outperformed XGBoost on this reduced input, offering better generalization and training efficiency.


Residuals vs. Predicted Values
Residuals
Predicted Values