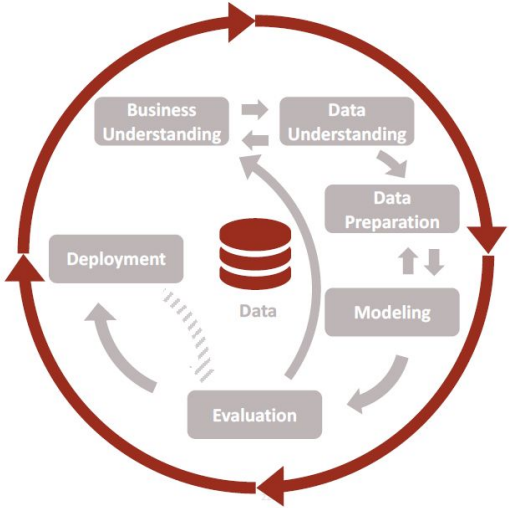
GROUP 6: WURM PROJECT

TABLE OF CONTENTS



BUSINESS UNDERSTANDING

01

PROBLEM DEFINITION

Project question: "Was it worth it to change the door seal?"

Main stakeholder: Manufacturer of a new type of door seal

Define "worth":

- Can't see if better from user experience perspective
- Regulates fridge temperature better? (However temperatures were already within acceptable bounds)
- **Less energy required to keep fridges cool? (Cost effectiveness) - best approach**

What improvements do the business stakeholders want to see?

FINAL PRODUCT

Task: Recognition of a pattern (before/after door seal change) and recommendation of whether the new door seal material is better

Model qualities: Balance complexity/accuracy with simplicity/understandability. **Our model must be explainable - not focused on automation**

Evaluating model:

- Model should show clear, significant, consistent result
- Result should relate directly to the problem, not extraneous factors
- Model should be understandable and explainable to the stakeholders

DATA EXPLORATION

02

EXPLORING DATASET

Cleaning data

- Renamed columns for understandability
- Created metadata description with units
- Set feature types - **datetime, category**

Individual fridge variables:
measuredTempFridge (the fridge's current temperature) -- temperature in **celsius**
tempAirOut, tempAirIn, (temperatures of inlet and outlet air streams) -- temperature in **celsius**

Heat pump variables (groups of fridges):
measuredTempEvap (the temperature of the evaporator) -- temperature in **celsius**
workload (utilisation of the heat pump) -- **percentage** 0-100%

Scale of data

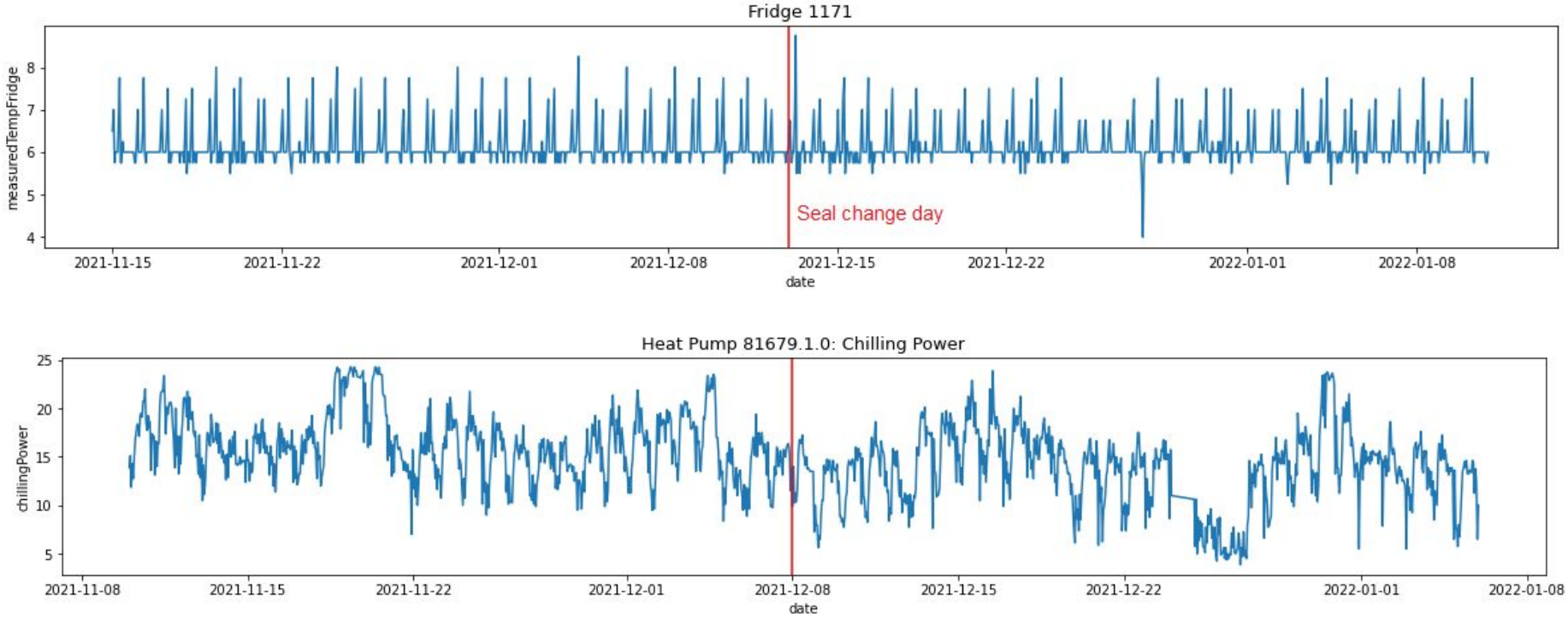
- 4 datasets: **markets, energy, heat pumps, fridges**
- 22 markets, 27 heat pumps, 270 fridges, 1-21 fridges per heat pump
- Longer than wide, few features (no high dimensionality issues)
- Mainly numeric data
- Different types of fridges included (different “set temperatures”)

```
fridges["setTempFridge"]\
.value_counts()
```

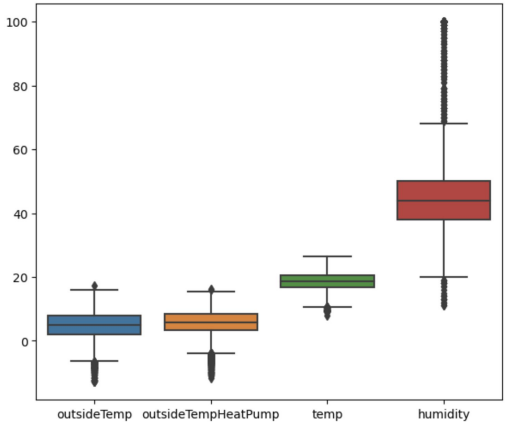
6.000	305812
1.000	34132
4.000	15048
2.000	5472
3.000	1368
5.000	1368
5.500	1368
10.000	1368
1.500	1368
7.000	1368
0.000	48

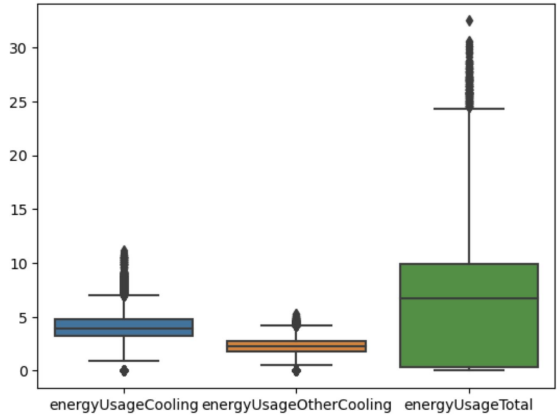
VISUALISING THE DATA

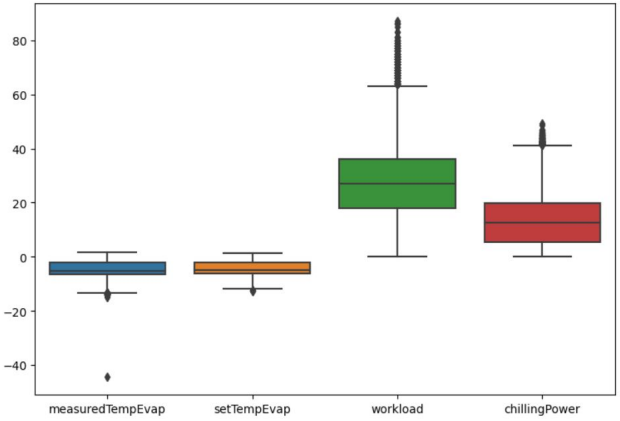
Important aspect of data: time series



VISUALISING THE DATA





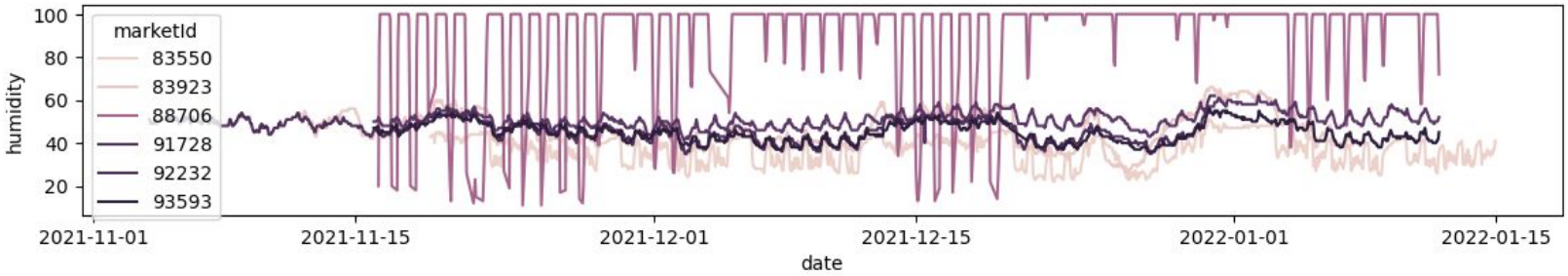


- Distribution of features tends to be normal
- Majority of data in narrow bounds
- However a few outliers visible in plots

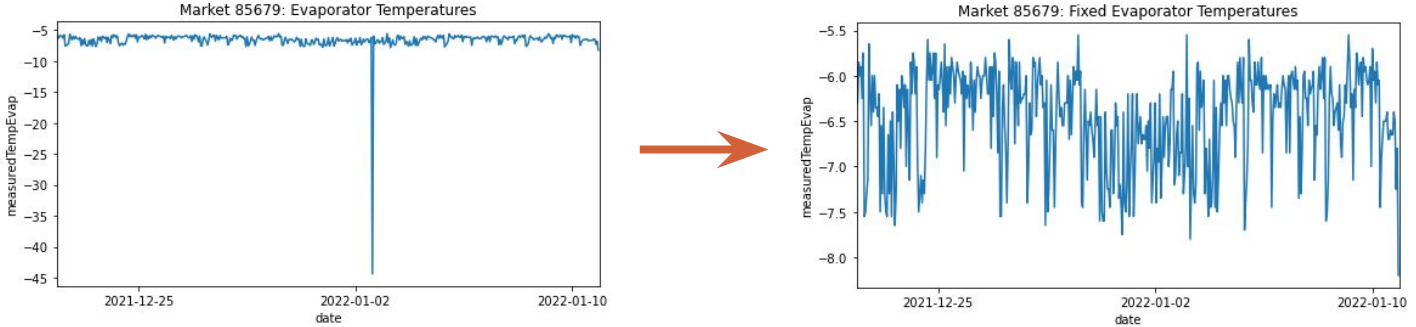
OUTLIERS

Outliers should only be removed if they are incorrect data (faulty sensors) rather than unique/unusual cases (important information)

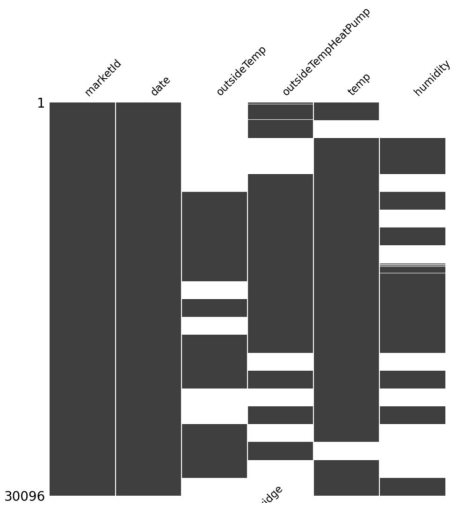
Supermarket with humidity at 100%, clearly faulty sensor. **Handle:** drop feature.

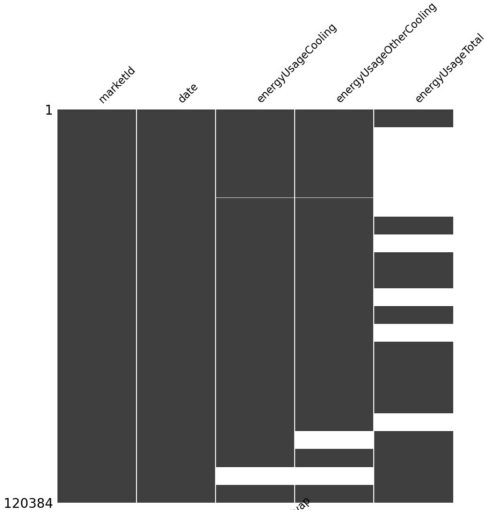


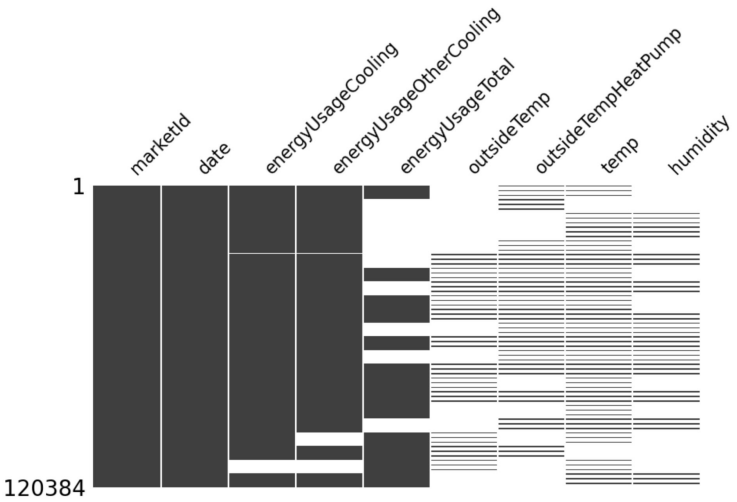
Evaporation temperature measurement -40. **Handle:** impute with previous value.



DATA





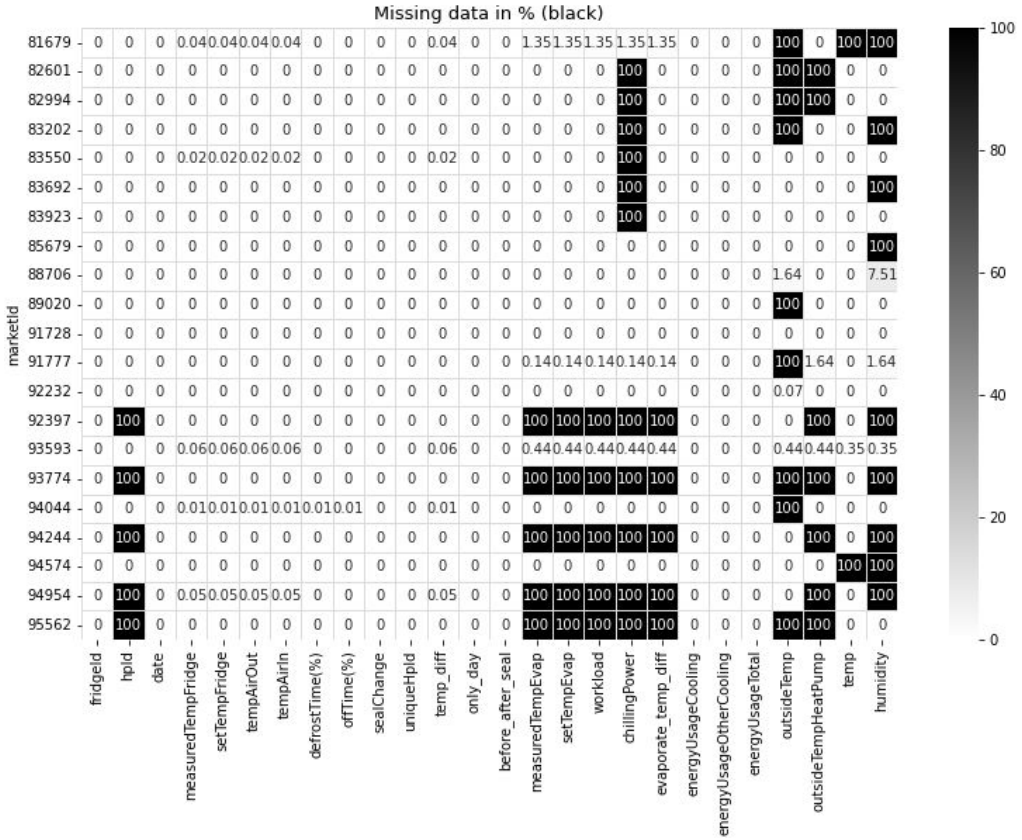


- A lot of missing data
- Some of the markets have no data regarding specific measurements.

DATA


marketId
fridgeId
hpId
date
measuredTempFridge
setTempFridge
tempAirOut
tempAirIn
defrostTime(%)
offTime(%)
sealChange
uniqueHpIdFridge
energyUsageCooling
energyUsageOtherCooling
energyUsageTotal
outsideTemp
outsideTempHeatPump
temp
humidity
measuredTempEvap
setTempEvap
workload
chillingPower
uniqueHpIdHeatpump
1
460424

DATA

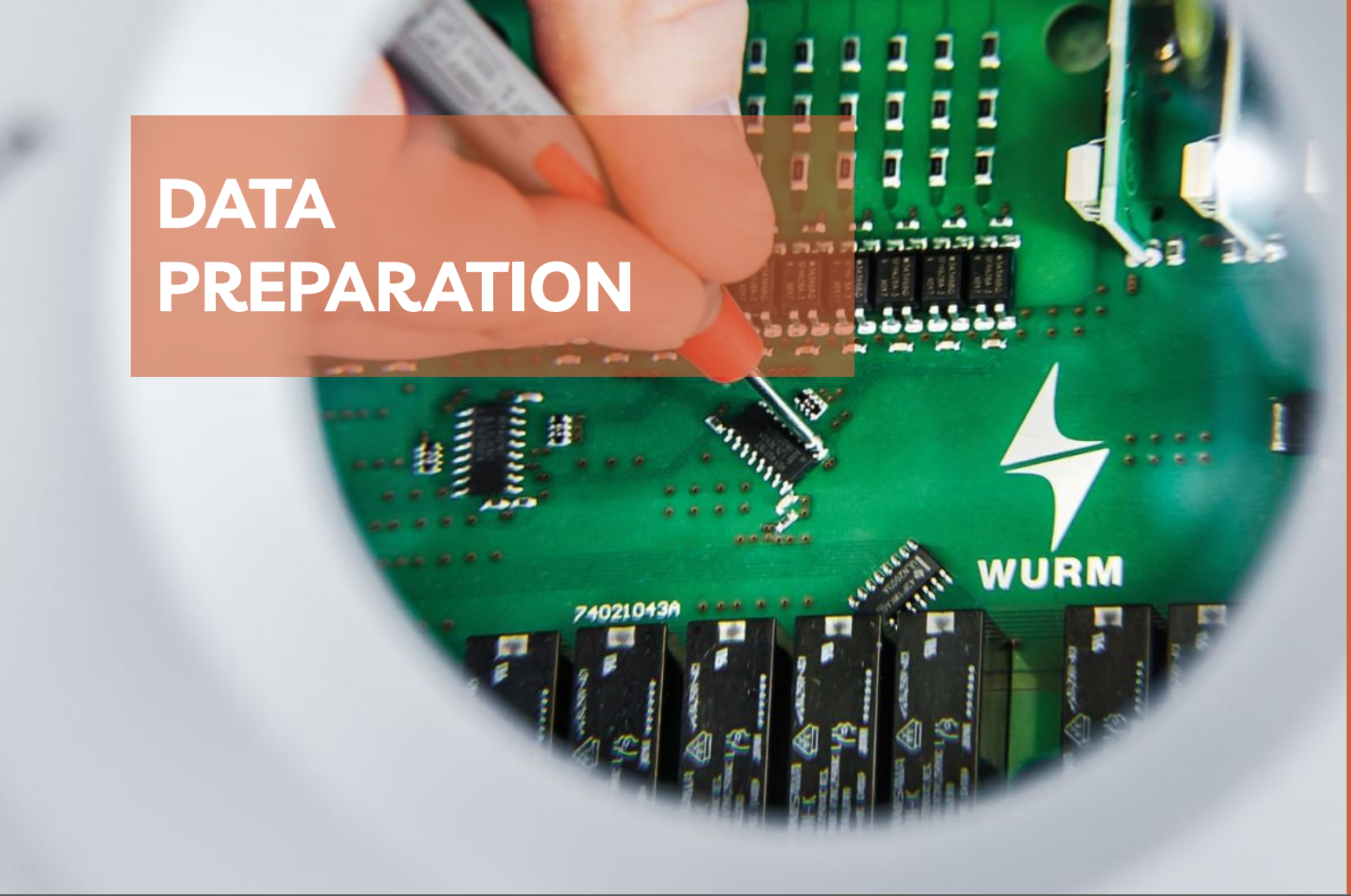


MCAR vs. MAR vs. MNAR:

"Missing values are a problem when the reason the values are missing is related to the problem of interest"

- Very high amount of missing data - missing for an entire feature/market.
- Reason for missing not associated with door seal - simply failed to record.

Handle: Cannot impute (as is entire feature) therefore have to drop. In reality would want to "**re-query**".



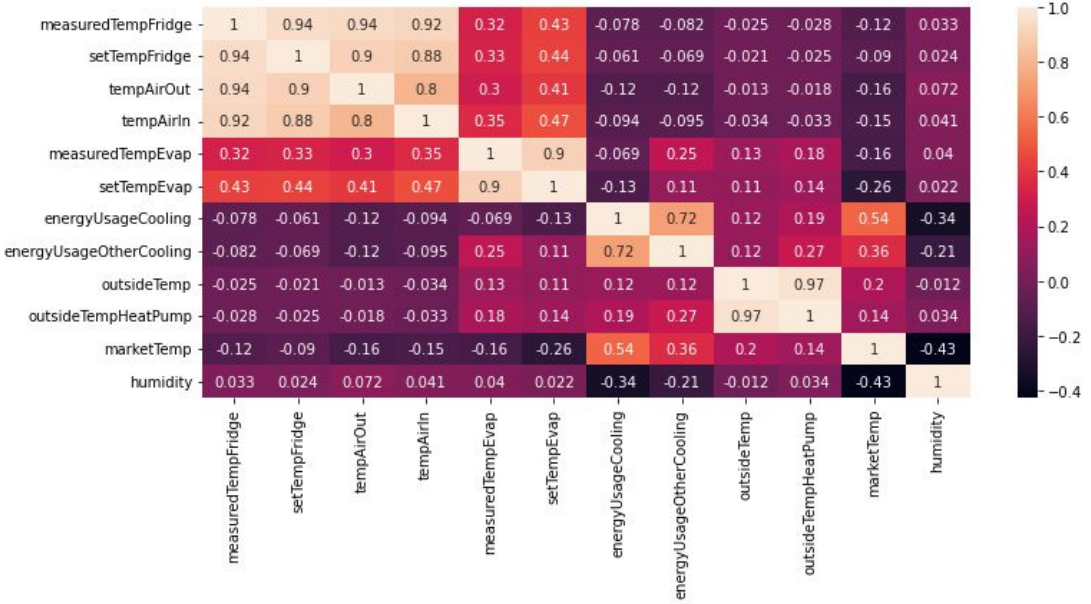
DATA PREPARATION

03

DROPPING DATA

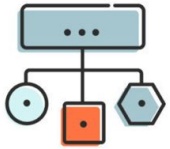
Missing data: Only markets: [89020, 91728, 92232, 91777, 93593, 94044] have data on all relevant features. Other markets must be dropped.

Correlated features:
Only include one feature from each pair of highly correlated features, e.g. [outsideTemp, outsideTempAtHeatPump]



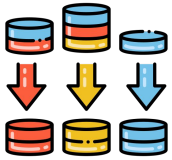
PCA: PCA unnecessary due to small number of features.
We are interested in explaining effect of features, but PCA loses understandability.

PRE-PROCESSING DATA



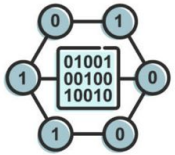
Feature engineering:

- *Describe dependencies:* create features for temperature differences
- *Add meaning to time dependent data:* mark each observation as "before" or "after" seal change (0/1). Drop all observations "on" the day of seal change



Normalising:

- Data have very different scales (% humidity versus celsius temperatures)
- *sklearn.StandardScaler* uses z-transformation -> $z = (x - u) / s$

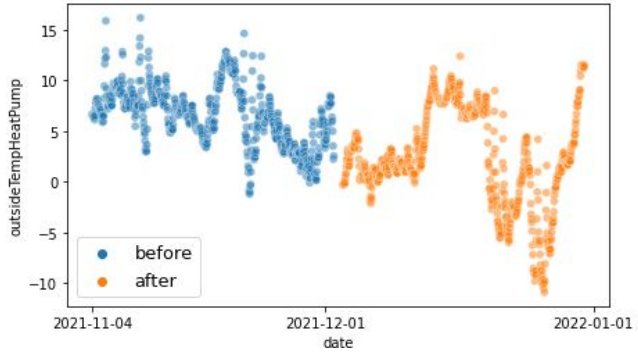


One-hot encoding:

- "One-hot-encode" categorical features (heat pump identifiers)
- "Label encoding" with ordering doesn't make sense for this data

MODELLING APPROACH

Rejected approach: modelling sealChange as dependent variable (clustering/logistic regression) fails, due to it being confounded with patterns over time



Chosen approach: model **chilling power** (kW energy required to power the heat pump) as **dependent variable**, with other features and **sealChange as independent variable.** Is sealChange significant in the model?

- Chose features based on domain knowledge. **humidity** affects heat capacity of air, **outsideTemp** affects how much heat taken off compressed gas, etc.
- **Interaction terms:** e.g. if humidity modifies the way the seal affects heat transfer
- **Polynomial terms:** represent non-linear relationships

AGGREGATING DATA

Change scale: 'energy' dataset uses 15 minute intervals, others use 1 hour. Use pandas *resample* to aggregate - take **mean** or **sum** depending on the unit.

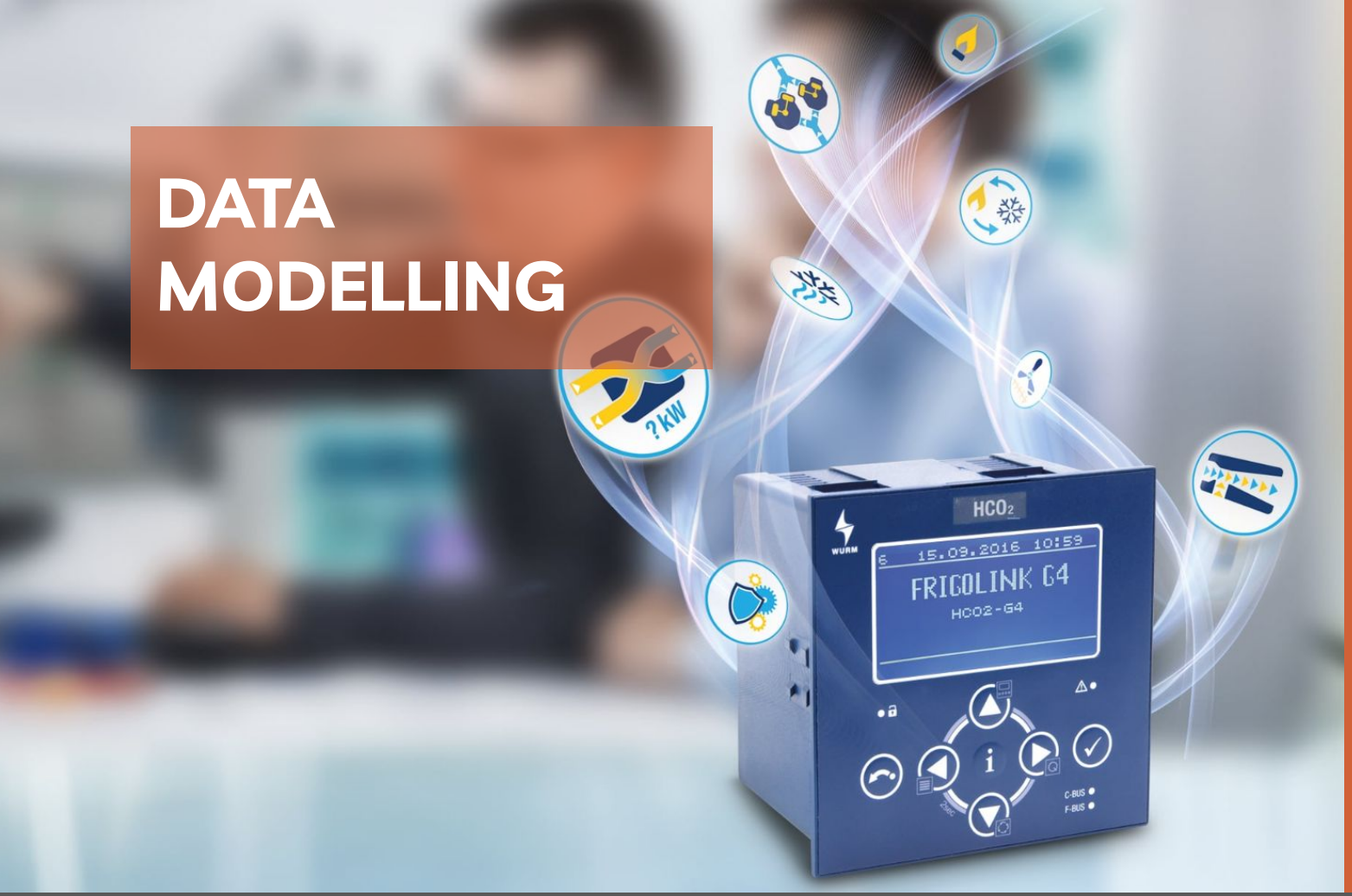
Merge datasets: 4 datasets merged to heat pump level because:

- chillingPower only given for the heat pumps
- Relationships only exist at group level. *e.g. if one fridge is hot => chillingPower increases. The other fridges in that heat pump group then have high chillingPower, even though they are still cool -* ***this confuses the algorithm***

Final dataset:

	date	uniqueHpId	chillingPower	before_after_seal	numOfFridges	...	avgMeasuredTempFridge	avgSetTempFridge	avgTempAirOut	heatPump_89020.1.0	heatPump_89020.2.0
0	2021-12-29 00:00:00+00:00	89020.1.0	-0.894	0	0.840	...	0.917	0.849	0.869	1	0
1	2021-12-29 01:00:00+00:00	89020.1.0	-0.781	0	0.840	...	0.765	0.849	0.744	1	0
2	2021-12-29 02:00:00+00:00	89020.1.0	-0.910	0	0.840	...	0.855	0.849	0.815	1	0
...	...	...	...	...	...	...	...	...	...	...	...
10687	2022-02-28 23:00:00+00:00	94044.3.0	-1.152	1	-1.280	...	-1.544	-1.293	-1.452	0	0

DATA MODELLING



04

CHOSEN MODEL

Supervised learning algorithm (for labelled data)

Models rejected:

- Deep neural network can capture complex relationships
- However black box algorithm, difficult to see impact of features. (Important for this problem). Loses understandability.

Model chosen: **Multiple linear regression.**

- Dataset relatively small - no computational performance limitations - don't need to use gradient descent, can find optimum directly
- Can derive feature importance from coefficients

MULTIPLE LINEAR REGRESSION MODEL

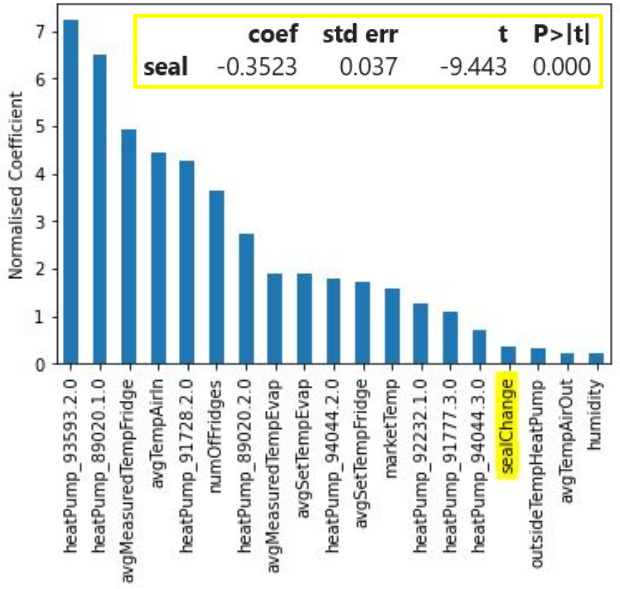
chillingPower ~ **sealChange** + numOfFridges + fridgeTemp + fridgeSetTemp + inletAirTemp + outletAirTemp + evaporatorTemp + evaporatorSetTemp + outsideTemp + marketTemp + humidity + C(heatPumpIds)

Model results:

- sealChange has **significant** p-value.
- Coefficient of sealChange is **negative**
  (less chilling power needed after seal change)

Feature importance:

- All features highly significant (except outletAirTemp)
- sealChange has relatively **small effect**



STATISTICAL EVALUATION

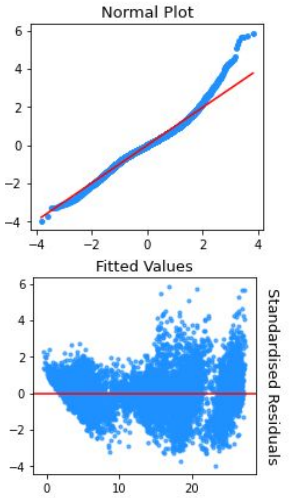
Cross validation:

- *Time series split* method to prevent "predicting past from future" fails
- Before/after seal change classes are not distributed randomly

	fit_time	score_time	test_score
0	0.005	0.001	-1.684
1	0.001	0.000	-1062220575657792765952.000
2	0.001	0.000	0.285
3	0.002	0.000	-810387041004812809273344.000
4	0.002	0.000	-1909547776291994.250

- *Shuffle split* method with 10 folds shows high R^2 scores
- Questionable normality and patterns in the variance of residuals likely reflects patterns over time in dataset

	fit_time	score_time	test_score	train_score
0	0.004	0.000	0.951	0.953
1	0.003	0.000	0.949	0.954
2	0.003	0.000	0.954	0.953
3	0.003	0.000	0.950	0.954
4	0.003	0.000	0.952	0.953
5	0.003	0.000	0.954	0.953
6	0.003	0.000	0.953	0.953
7	0.003	0.000	0.954	0.953
8	0.003	0.000	0.952	0.953
9	0.003	0.000	0.952	0.953



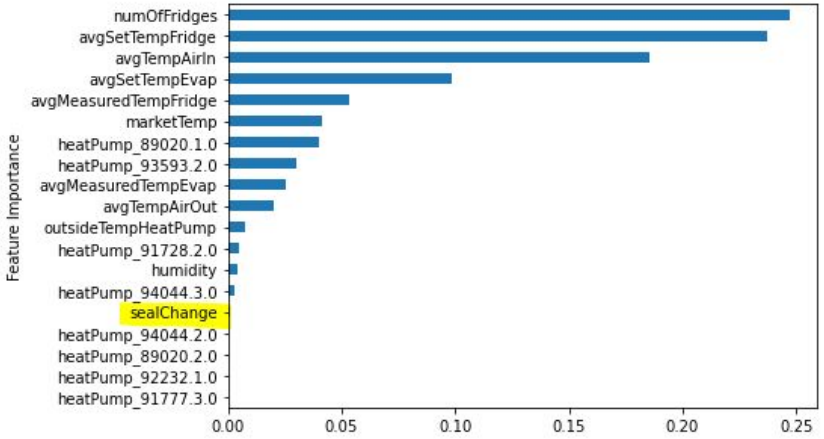
Berrak Küçük, Jannik Alexander & Natasha Randall

NON-LINEARITY

Examined non-linear relationships with **Random Forest Regressor**
(rather than adding complex polynomials to linear regression model)

```
forest_model = RandomForestRegressor(random_state=0,
                                     n_estimators=180, max_features=0.6)
forest_model.fit(data[X_vars].values, data[y_var].values)
```

scikit-learn's "Feature Importance" shows the sealChange variable is very **unimportant** for predicting chilling power



STATISTICAL EVALUATION

Cross validation with grid search:

Used to optimise hyperparameter choice.
Test and training scores very close -
validation curve not revealing.

Best parameters:
max_features: 0.6,
n_estimators: 180

params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	split5_test_score	split6_test_score	split7_test_score	split8_test_score	split9_test_score	mean_test_score	std_test_score	rank_test_score
{'max_features': 0.6, 'n_estimators': 180}	0.978	0.980	0.980	0.978	0.980	0.980	0.979	0.979	0.977	0.981	0.979	0.001	1

Nested cross validation:

Used for unbiased error estimation for model.
Model fit score is very high.

Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
0.982	0.978	0.981	0.976	0.978	0.976	0.978	0.979	0.977	0.979

mean	0.978
std	0.002
min	0.976
max	0.982

IMPACT OF HEAT PUMPS & INTERACTIONS

Linear model with heat pump categories: sealChange **p<0.001**
Linear model without heat pump categories: sealChange **p=0.143**

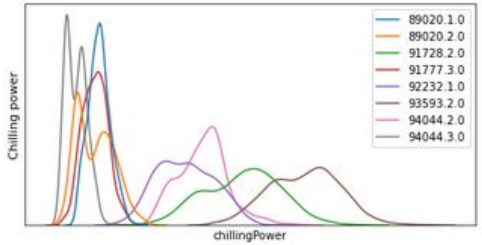
Model with interactions:
Looking at only interactions with sealChange - most important interactions are with heat pumps

terms	p-values	coefficient parameters
sealChange * heatPump_91728.2.0	0.000	1.755
sealChange * heatPump_94044.2.0	0.000	-0.724
sealChange * heatPump_89020.1.0	0.007	-0.625
sealChange * heatPump_93593.2.0	0.004	0.494

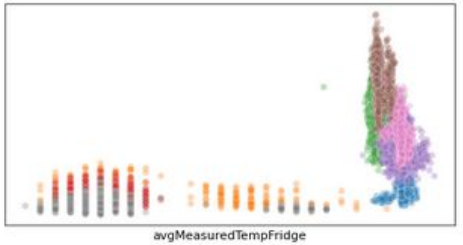
- Different heat pumps change the effect of seal change on chilling power
- Differences not captured in data (e.g. differences of humidity, market temp)
- Need to look at each heat pump individually

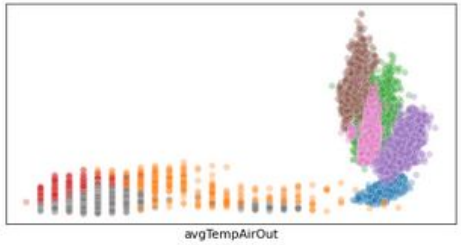
IMPACT OF HEAT PUMPS

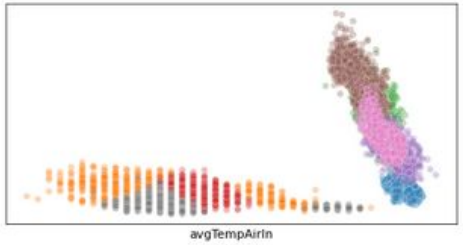
Relationships between dependent and independent variables differ by heat pump

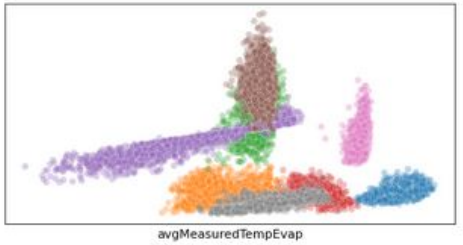


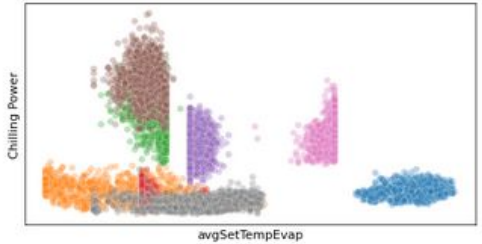
heatPumpId	RSquared	sealChange_Coef	sealChange_p-Value	numOfFridgesInHeatPump
91728.2.0	0.716	0.775	0.000	11
89020.2.0	0.660	0.293	0.000	1
93593.2.0	0.687	0.145	0.000	13
91777.3.0	0.525	-0.097	0.016	1
94044.3.0	0.662	-0.096	0.011	1

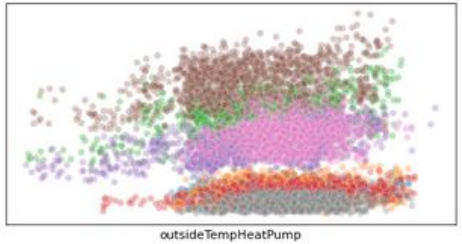


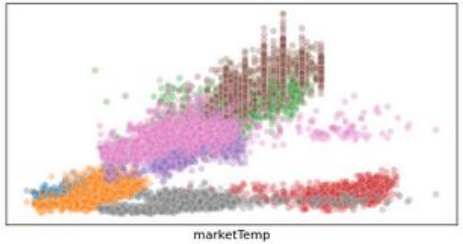


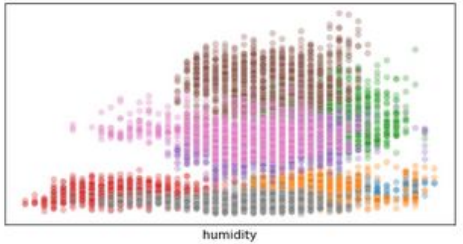












NOISE

Data recorded was single point measurement - if someone opened fridge door at that moment, measurement will be off

Modelling only data recorded between 1am-4am (no one in supermarket)

Full model at night: sealChange has **significant** p-value, coefficient is **negative,** still extremely small relative importance

coef	std err	t	P>|t|
-0.6482	0.089	-7.259	0.000

Modelling each heat pump at night:

heatPumpId	RSquared	sealChange_Coef	sealChange_p-Value	numOfFridgesInHeatPump
91728.2.0	0.803	-0.558	0.028	11
94044.2.0	0.560	-0.447	0.000	10
92232.1.0	0.711	0.262	0.045	12
89020.1.0	0.636	-0.201	0.035	12
94044.3.0	0.554	-0.128	0.007	1

With less noise, significant coefficients now tend to be negative - *"seal change reduces chilling power"*

EVALUATING RESULTS

05

SEMANTIC EVALUATION

Result: The negative coefficients in the models suggest that the door seal change has reduced the chilling power required for the fridges (**a good change**)

Problems: The impact of the change to door seal is *very small*

- Amount of air leaked by one type of door seal compared to another is tiny in scale compared to amount of air let in by someone opening the door
- The impact of other factors (e.g. fridge/evaporator temperatures) on chilling power are massively larger in scale

REQUIRED FEATURES

We are searching for a tiny effect in a huge amount of noise.
The dataset lacks many features needed to explain the process:

- what is content of the fridges (at all times)
- what is heat capacity of the content in fridges
- tracking every time they are opened
- when the doors were last replaced
- if the heat pump also power freezers
- temp directly outside fridge
- temp of fridge walls
- humidity in each fridge
- humidity directly outside each fridge
- where is the fridge in the market
- fan speed of evaporator
- if the doors were cleaned during the seal change, etc etc

CONFOUNDING WITH TIME

We cannot conclude the new seal is better because:

The before/after door seal change feature simply represents time passing

This feature captures any/all changes that occurred over time:

- if contents of fridge have changed before/after seal change
- volume of use of fridge may change over time
- procedure during seal change - fridges cleaned, doors changed

If an old seal was degraded/broken, simply **a new door seal of the old material would also show an improvement.** (No control group).

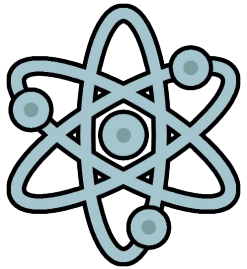
CONCLUSION

Model evaluation:

- The model is **not relevant or trustful** - we cannot be sure it is actually modelling the factor we want it to
- Effect is small and results can be inconsistent

Better approach:

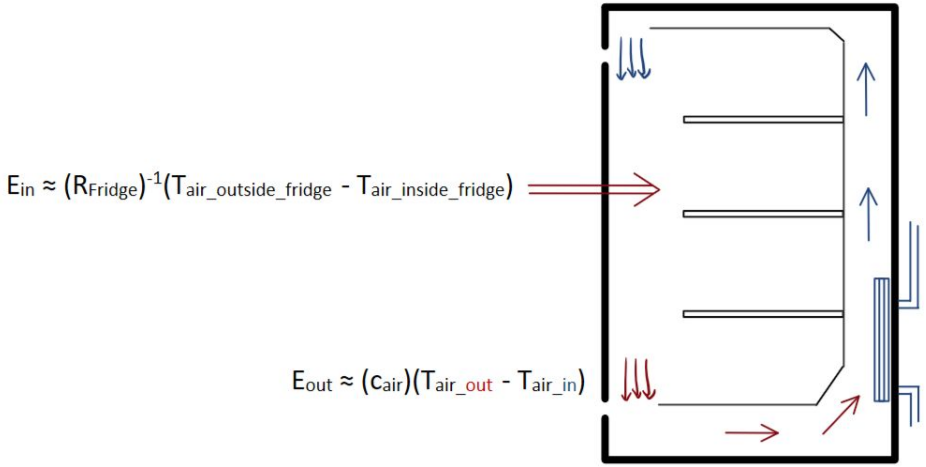
- More data, more features over a longer period of time
  - Predictions based on observations made in summer will be extrapolating from outside of the dataset.
- Different type of modelling approach - **physics based**



PHYSICAL MODEL

Laguerre, O. (2010) ***Heat transfer and air flow in a refrigerator.*** Mathematical Modeling of Food Processing. 453-482.
Feynman, R., Robert, L., Sands, M. (1965) ***The Feynman Lectures on Physics,*** Vol 1, chp 45.

By energy conservation:
thermal energy going into fridge = heat flow out via cooling



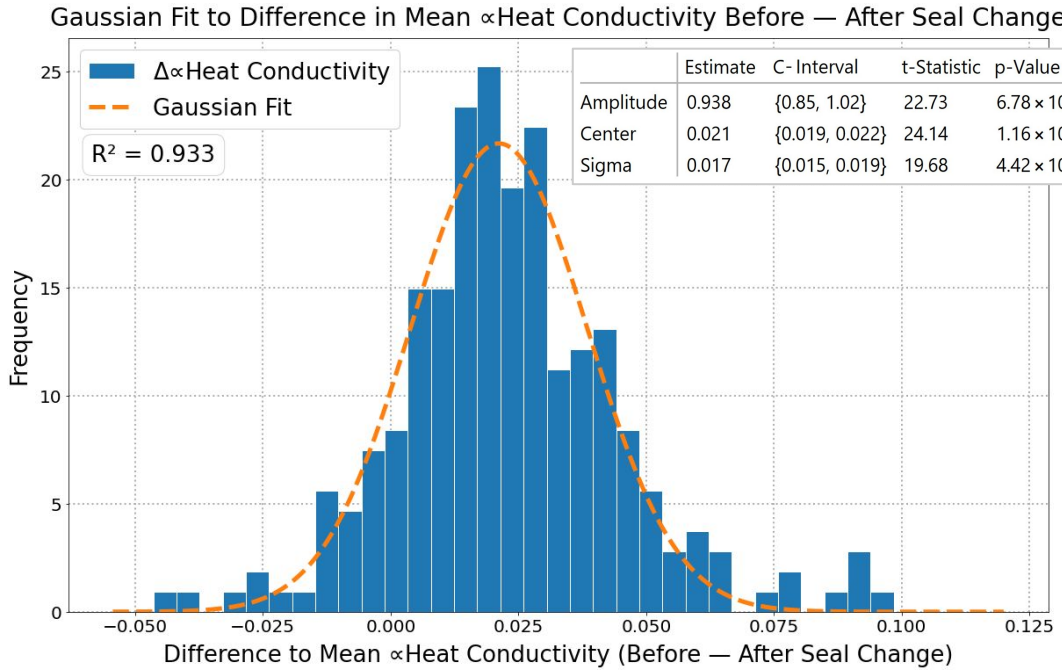
Dividing energy out by energy in, is proportional to the heat conductivity of the fridge

Does the new seal change the heat conductivity?

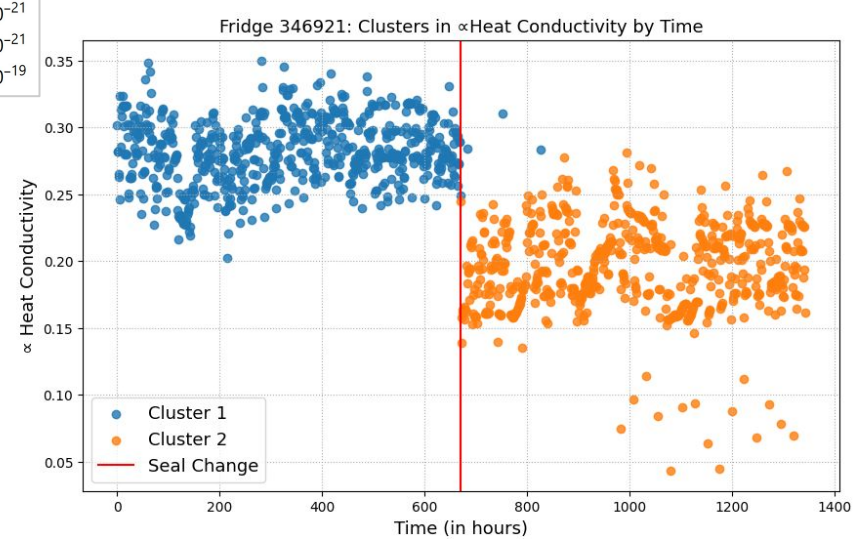
$$(R_{Fridge})^{-1} \propto (T_{air\_out} - T_{air\_in})/(T_{air\_outside\_fridge} - T_{air\_inside\_fridge})$$

PHYSICAL MODEL

For each fridge calculate: (mean heat conductivity before) – (mean heat conductivity after) the seal change



	Estimate	C- Interval	t-Statistic	p-Value
Amplitude	0.938	{0.85, 1.02}	22.73	6.78×10^{-21}
Center	0.021	{0.019, 0.022}	24.14	1.16×10^{-21}
Sigma	0.017	{0.015, 0.019}	19.68	4.42×10^{-19}



Gaussian fit mean is far above 0 => **heat resistance (1/conductivity) has improved**

Change to heat resistance corresponds to seal change!

ARE THERE ANY QUESTIONS?