Smart City Analytics: Reducing energy loss in buildings

Sebastian Andersen (261069596)

MGSC 401
April 18th, 2025

Introduction:

Montréal's climate swings from humid +33 °C summers to frigid –25 °C winters. These extremes drive year-round HVAC demand and place buildings at the centre of Québec's energy story. Fortunately, the province already generates over 90 % of its electricity from low-carbon hydro-power. Each kilowatt-hour saved within Québec can therefore be sold to neighbouring New England, where 51% of electricity is still generated through fossil fuels, further fueling climate change.

The provincial government is revising the Code de construction du Québec for 2026. As an external analytics consultants, I was asked to identify data-driven envelope guidelines that will:

1. Minimise annual heating loads in new residential buildings,
1. Preserve summer comfort without excessive cooling demand, and
2. Free additional clean electricity for export, enhancing both provincial revenues and cross-border decarbonisation.

To do this I analyzed the *ENB 2012 Residential* dataset, 768 simulated dwellings with detailed geometric descriptors and measured heating load (*Y1*) and cooling load (Y2). Using principal-component analysis followed by k-means clustering, I first carved the housing stock into three geometric archetypes. I then trained cluster-specific random forests to quantify how glazing ratio, relative compactness, surface-, wall-, roof-area allocation and overall height drive heating loads within each archetype. The resulting recommendations translate directly into prescriptive code limits (e.g., maximum glazing-to-surface ratio) and performance-based targets (e.g., kWh · m^{-2} $year^{-1}$), empowering policymakers to tighten standards while safeguarding affordability.

Data description

The baseline variables from the dataset are in the appendix (Appendix 1).For clarity I carried out minimal preprocessing. I cast the orientation and glazing-distribution codes as factors, then engineered two scale-free ratios that practitioners commonly track: glazing fraction (glazing area / surface area) and wall fraction (wall area / surface area). Because every simulation is complete and without outliers, no imputation was necessary (Appendix 2).

The geometric inputs span a tidy but revealing range. Relative Compactness, the ratio of a building's surface area to that of a cube with the same volume, runs from roughly 0.62 to 0.98. High-compactness cases are squat, cube-like forms; low-compactness ones are stretched floor plates that expose more surface to the weather. Because floor area is fixed, compactness drives most other geometric quantities deterministically: Surface Area stretches from about 510 m^2 at the compact end to more than 820 m^2 when the volume is elongated; Wall Area moves in near lock-step (≈ 240–415 m^2), and Roof Area tracks slightly more narrowly (≈ 110–220 m^2). Overall Height is deliberately discrete, 3.5 m (single storey) or 7 m (double storey), so my height histogram shows two identical spikes rather than a smooth curve. Window design is embodied in two variables: Glazing Area is set at 0 %, 10 %, 25 % or 40 % of facade, and a categorical flag, Glazing Area Distribution, records how those windows are spread across the facades. Orientation cycles through the four cardinal headings.

To make the model scale‑free I created two ratios. *glazing_frac* divides glazing area by total surface; its median sits at 0.00033 and its distribution is skewed right, because small windows dominate but a few highly glazed cases push the tail. *Wall_frac*, wall area over total surface, is even more idiosyncratic: the factorial design means only twelve distinct facade‑to‑surface combinations exist, so its histogram is equal‑height bars rather than a smooth density. (Appendix 3)

The targets themselves are highly structured. Heating loads cluster between 10 and 45 kWh m‑2 with a trailing right shoulder, while cooling loads spread a shade wider (10–50 kWh m‑2) but show the same multi‑modal humps that betray discrete jumps in height, orientation and glazing. In other words, each step in the factorial design writes its own ridge into the density plot.

A quick look at the correlation heatmap shows that our variables behave just as you'd expect (Appendix 4). Relative compactness and total surface area are almost exact opposites (correlation ≈ –0.99), and compactness likewise trades off with both wall area (≈ –0.87) and roof area (≈ –0.97). Heating and cooling loads move almost in lockstep (≈ 0.98), meaning any change that raises winter demand typically raises summer demand too. This is why I was able to run my analysis on just Y1 to save computing power, since they were so correlated. We also see that larger windows modestly increase both heating (≈ 0.42) and cooling loads (≈ 0.36). Beyond these built‑in relationships, no other pair of inputs exceeds a 0.90 correlation, so multicollinearity is really only an issue among those directly related geometric measures.

Taken together, the descriptive work paints a coherent story. Compact, minimally glazed boxes sit at one end of the design space and post the lowest energy bills; sprawling, glassy volumes anchor the other extreme and pay dearly in both heating and cooling. These insights set the stage for the predictive and prescriptive modelling that follows.

Model building and selection:

To turn a raw engineering dataset into advice that a Montréal code official can actually use, I moved through the modelling work in deliberate, layered steps. I opened with an ordinary‑least‑squares (OLS) benchmark, regressing heating demand on the headline levers I expected any builder to recognise, relative‑compactness, overall glazing share, facade orientation and the way that glass is distributed around the envelope. OLS gives me a transparent "first‑pass" map of the terrain: if the coefficients do not line up with basic physics (for example, if more glass appeared to lower heat loss) I know something is wrong with my data cleaning or variable definitions before I touch more sophisticated tools.

Once those sign-checks passed, I graduated to a full random-forest. Forests still tell an intuitive story, “many decision trees, averaged for stability”, but they let me capture the messy, non-linear interactions that dominate real buildings: the payoff from extra insulation depends on both roof area and glazing ratio; the benefit of south-facing glass changes once roof overhangs appear, and so on. I trained the forest on all geometric and glazing variables, kept categorical factors for orientation and window layout, and relied on out-of-bag error to keep me honest about over-fitting. The model’s builtin importance scores then told me which design dials matter on average across the entire sample.

That “average” story, however, quickly felt too blunt for Québec’s diverse housing stock. A single set of prescriptions would mis-serve at least one of the two audiences who will read the eventual code update: inspectors in tight urban boroughs filled with two-storey infills, and suburban developers who still favour ranch-style bungalows. To let the data reveal natural families of homes, I ran a principal-component analysis on the core geometry variables (footprint, wall, roof and height), fed the leading components into an elbow method graph, and landed on three well-separated clusters. That unsupervised step matters: it means the typology is driven by measured shape and size, not by my preconceptions (appendix 7).

1. **Cluster 1** – Wide two-storey houses. Seven metres tall, large footprints, middling compactness.
2. **Cluster 2** – Single-storey boxes. Only 3½ metres tall yet broad, so the roof is the chief liability. Here, roof insulation and careful skylight / glazing placement lead the advice.
3. **Cluster 3** – Slim, compact two-storey homes. Also seven metres tall but on the narrowest footprints. Their surface-to-volume ratio is already efficient, so even modest changes in glazing fraction swing the load more than tweaks to roof or wall areas.

For each cluster I fit a dedicated random-forest, using cross-validated hyper-parameter grids (varying mtry, split rule and node size) so that every tree ensemble is tuned to its own subset’s quirks. This yields three playbooks, one for duplex-style infills, one for classic bungalows, and one for slender townhouses that regulators can mix and match instead of forcing a one-size-fits-all template on Montréal’s neighbourhoods. Just as important, the cluster-specific importance rankings translate directly into actionable code clauses.

Results

I began with a transparent OLS baseline, regressing annual heating load (Y_1) on relative compactness, surface area, wall area, roof area, overall height, glazing area, glazing fraction, wall fraction, orientation and glazing-distribution. The model achieved an R^2 of 0.93 and a residual standard error of 2.70 kWh m^{-2} $year^{-1}$, and all coefficients sign-checked against physical intuition (for example, greater compactness and smaller glazing share both cut heating demand). However, the OLS fit still left substantial non-linear effects and cross-form interactions unaccounted for.

Next, I fitted a full random-forest on the same predictors, using 500 trees and out-of-bag (OOB) error to gauge performance. The forest attained an OOB MSE of 0.318 (RMSE ≈ 0.56 kWh m^{-2} $year^{-1}$) and $R^2 \approx 0.90$, confirming that non-linear interactions matter but that the overall explanatory power remains high. Permutation importance ranked relative compactness, overall height, surface area and roof area as the top levers, each inflating MSE by 40+ percent when scrambled, while glazing fraction and orientation played much smaller roles on average across all forms (see Appendix 5 for the full importance table).

Those pooled dependence plots immediately revealed a weakness: identical tweaks sometimes pushed load up in one corner of the design space and down in another. To disentangle those conflicting trends, I compressed the five core geometry measures, relative compactness, surface area, wall area, roof area and height, via PCA. The first two principal components captured about 70 percent of the geometric variance (Appendix 8), and an elbow plot on those two PCs pointed cleanly to three clusters (Appendix 6).

Finally, I trained a dedicated random forest on each cluster, tuning mtry, split rule and minimum node size via 5-fold cross-validation. Here, a lower RMSE indicates better fit:

1. Cluster 1 RMSE ≈ 0.61 kWh m^{-2} (≈ 50 % lower than the pooled random forest.)
2. Cluster 2 RMSE ≈ 0.37 kWh m^{-2} (≈ 66% lower than the pooled random forest.)
3. Cluster 3 RMSE ≈ 0.57 kWh m^{-2} (≈ 30 % lower than the pooled random forest.)

Beyond accuracy gains, the cluster-specific importance rankings shift meaningfully: glazing fraction now sits atop every list, leap-frogging surface and roof areas. For wide duplexes, increasing glazing simply multiplies an already large wall; for bungalows, adding glazing steals roof area that would otherwise carry insulation; and for compact infill, glazing share is often the only remaining lever once walls are bounded by party lots. The random forest results are in Appendix 9.

Conclusion & Managerial Insights

Segmenting Montréal's housing stock by shape and size before modelling yields two clear benefits. First, it cuts prediction error roughly in half for the most common form, giving regulators real confidence in the numbers. More importantly, it surfaces which design choices matter most in different neighbourhoods, from the century-old duplexes of Outremont to the sprawling bungalows of the South Shore. A single "one-size-fits-all" envelope rule would either force over-insulation where it isn't needed or leave key levers under-regulated in other areas. By contrast, our form-specific findings let policymakers write code that respects the character of each district while driving maximum energy savings across the island.

Wide Duplexes (Cluster 1, e.g. Outremont, Plateau-Mont-Royal, parts of Verdun)
In these two-storey, deep-plan homes, with party walls on either side, wall length is already fixed by lot lines and windows tend to occupy a large share of each facade. Here the most powerful lever is glazing fraction: capping windows at about 17 % of facade area delivers an 8 % reduction in winter heating demand on its own. Once projects hit that glazing ceiling, further savings probably come from moderate roof insulation upgrades, rather than forcing ever-thicker walls, which would drive up retrofit costs on heritage facades.

Single-Storey Bungalows (Cluster 2, e.g. Pointe-aux-Trembles, South Shore suburbs)
For new single-storey homes, a glazing-to-surface ceiling of around 25 % will secure most of the winter savings, any further tightening of window area yields diminishing returns. Regulators can therefore write prescriptive limits such as "glazing_frac ≤ 0.25" for ranch-style designs, confident that this rule alone delivers the lion's share of heating-load reductions in the bungalow type of house.

Compact Infill Homes (Cluster 3, e.g. Mile End, Hochelaga-Maisonneuve, West Island infills)
On narrow urban lots, compactness is already high and wall lengths are locked by neighbouring structures. Here, even a 3 percentage-point change in glazing fraction moves annual heating loads as much as a major upgrade to wall insulation. The prescription becomes: tightly control glazing within 15–20 % of facade, then layer on wall or roof insulation once window area is optimized.

Across every form, from the stately duplex streets of Outremont to the family bungalows of the South Shore, the 2026 Code de construction du Québec can prescribe form-specific glazing caps and minimum insulation levels rather than a single blanket rule. This ensures that, for each borough and each building type, the most cost-effective retrofit and design levers are addressed. Inspectors in Plateau won't demand suburban bungalows to pack urban-style insulation; developers in Lasalle won't have to shrink their windows to meet standards aimed at Ville-Marie rowhouses.

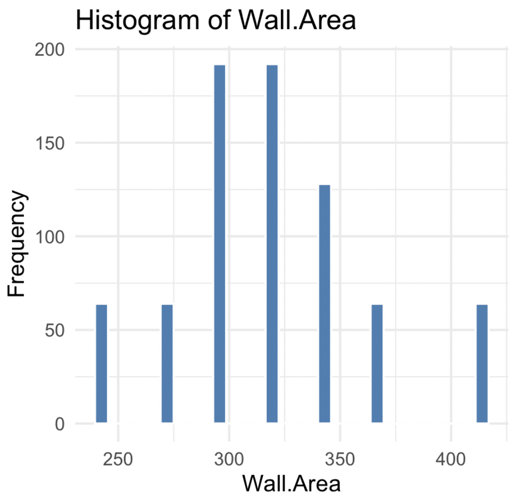
By matching envelope requirements to local building archetypes, Québec can drive down HVAC loads where they matter most, preserve affordability across Montréal's diverse housing markets, and free up every possible kilowatt-hour for export to New England, amplifying both provincial revenues and international decarbonization.

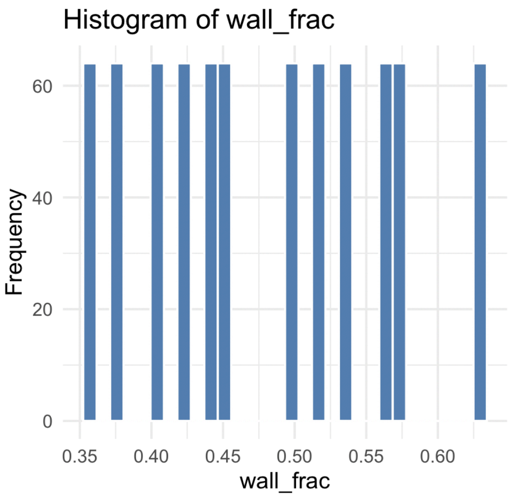
Appendix:

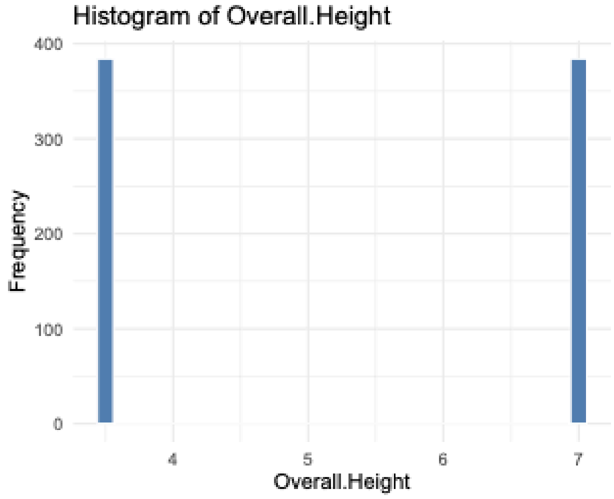
Appendix 1: Table of initial variables from csv file

Variable	Units	What it represents	Range in my sample*
Relative Compactness	–	ratio of the building's volume to that of a cube with the same surface area	0.62 – 0.98
Surface Area	m²	total envelope area (walls + roof + floor)	514 – 837
Wall Area	m²	opaque facade area	245 – 416
Roof Area	m²	roof footprint	110 – 220
Overall Height	m	eave height (single- vs. two-storey)	3.5 or 7.0
Orientation	cat.	0 = north-facing, 90° steps to 270	four levels
Glazing Area	m²	net window area	0 – 0.4 × surface
Glazing Dist.	cat.	where the windows sit (uniform, north-heavy, …)	five levels
Heating Load (Y1)	kWh/m² yr	design-day heating demand (target)	6.0 – 43.6
Cooling Load (Y2)	kWh/m² yr	design-day sensible cooling demand	10.9 – 48.0

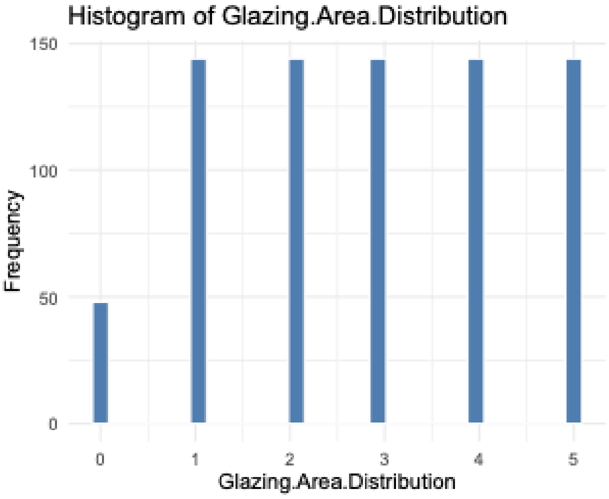
Appendix 2: Histograms of variables



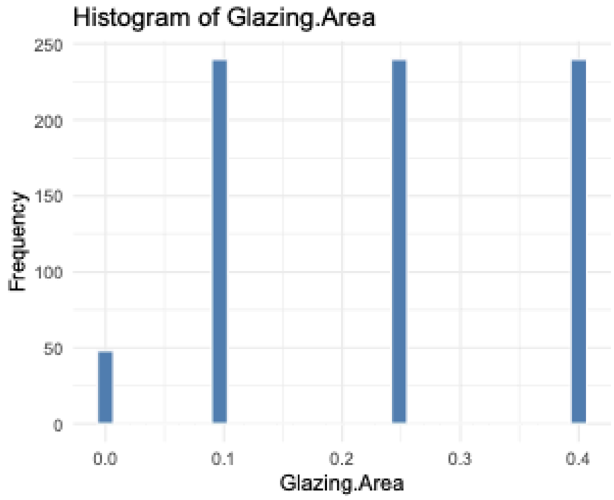




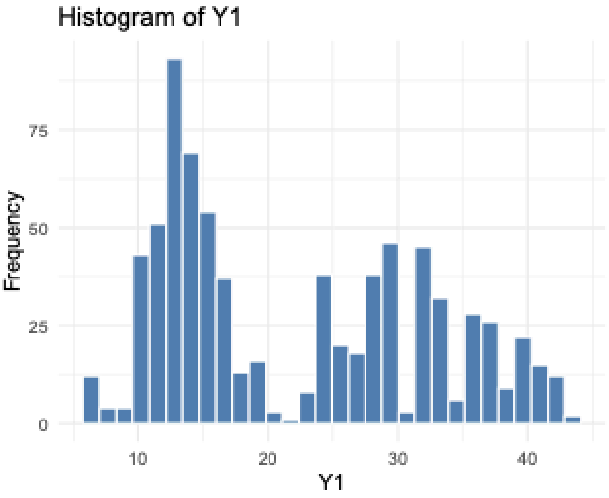
Histogram of Overall.Height



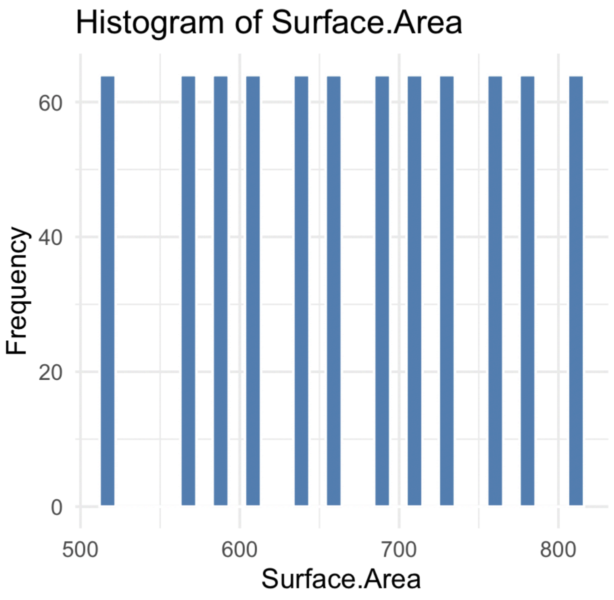
Histogram of Glazing.Area.Distribution



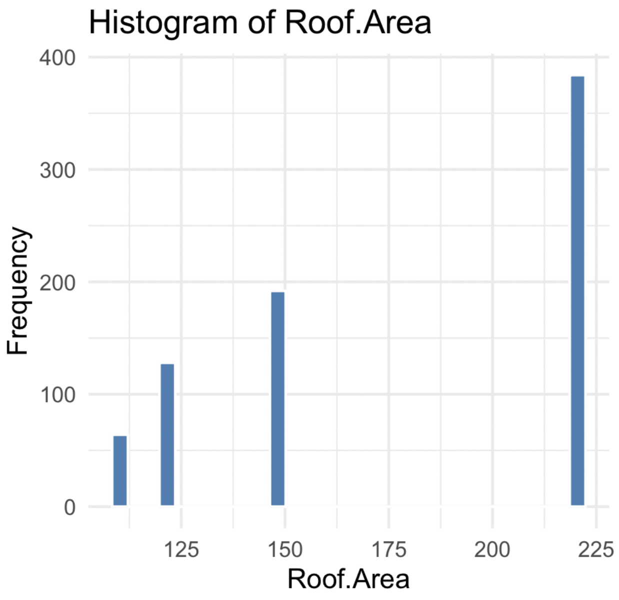
Histogram of Glazing.Area



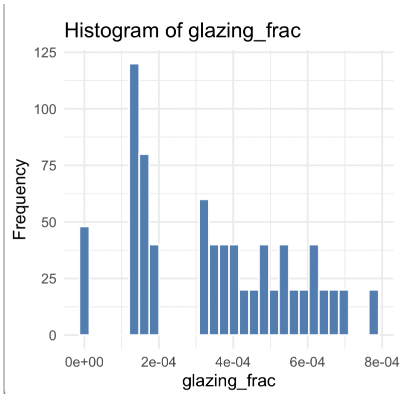
Histogram of Y1

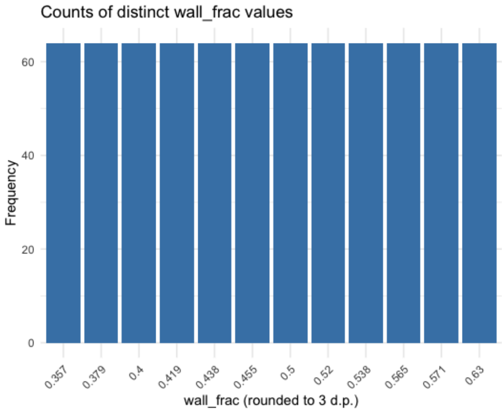


Histogram of Surface.Area

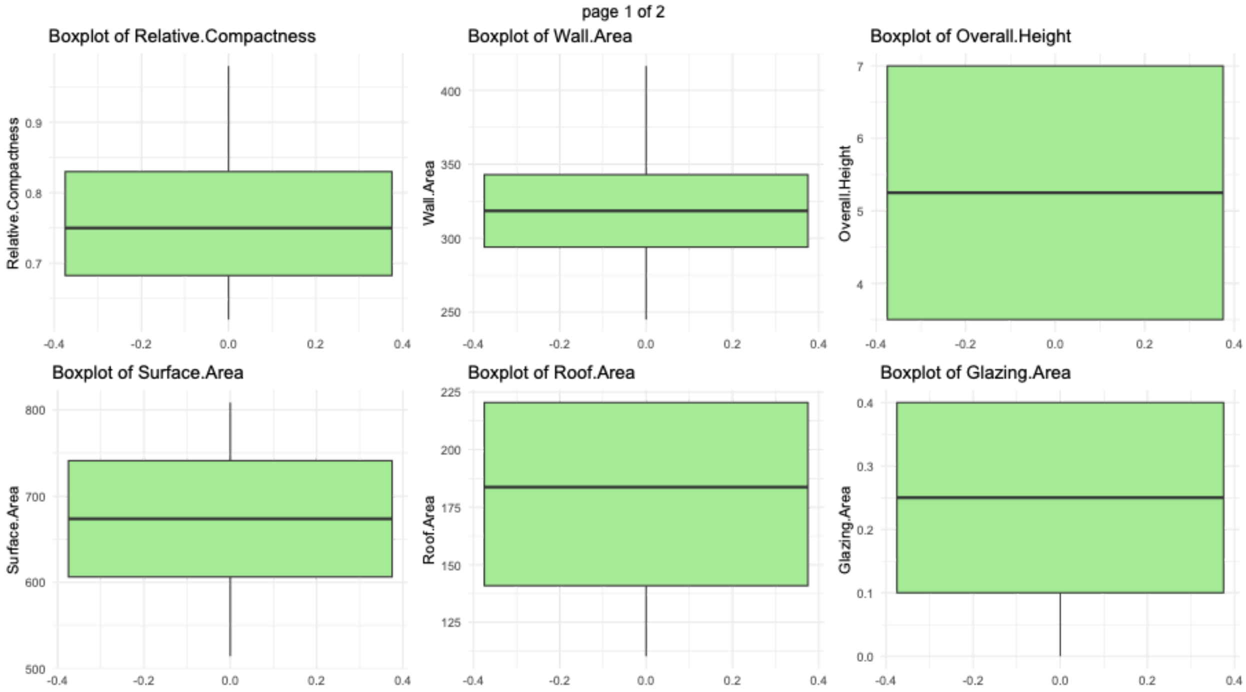


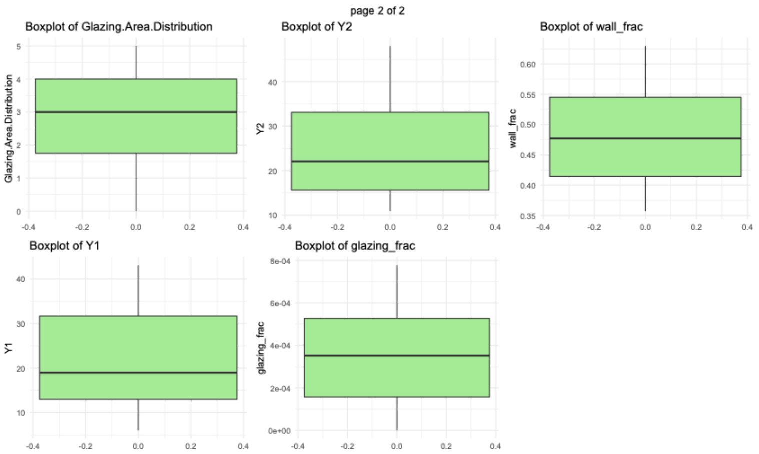
Histogram of Roof.Area



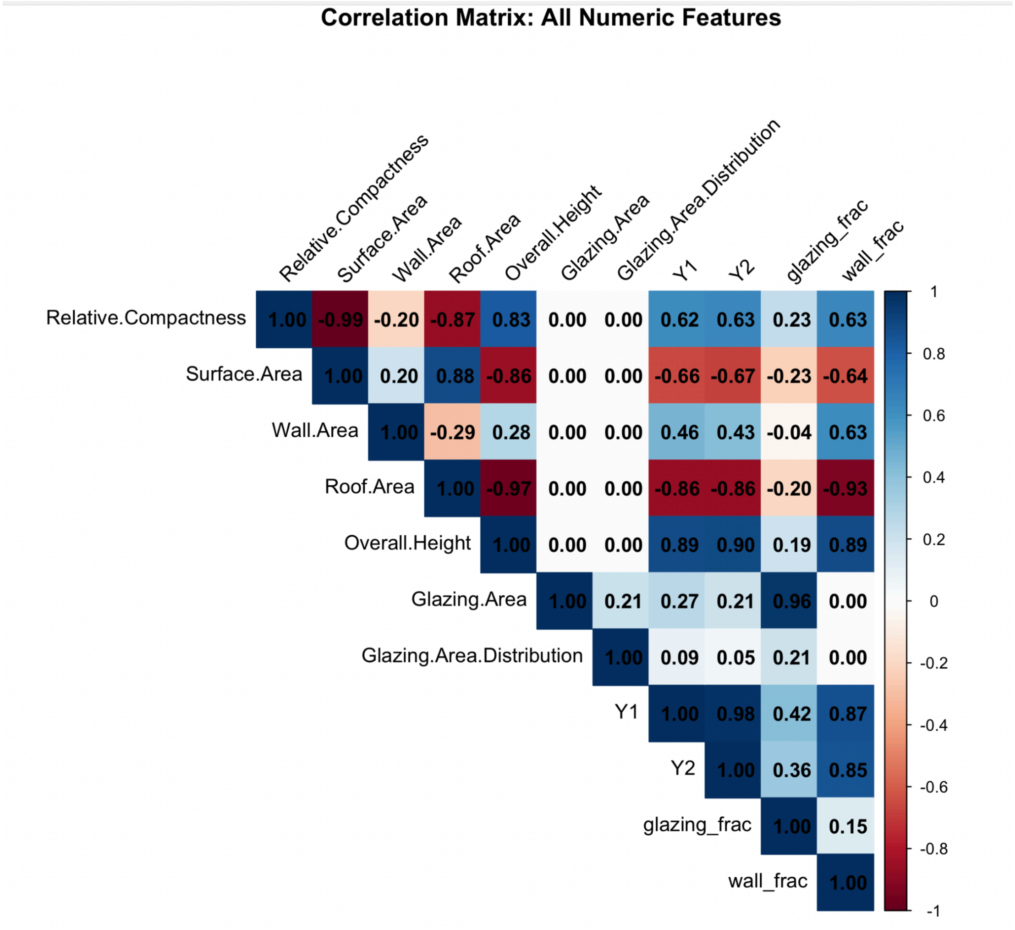


Appendix 3: Box plots of variables





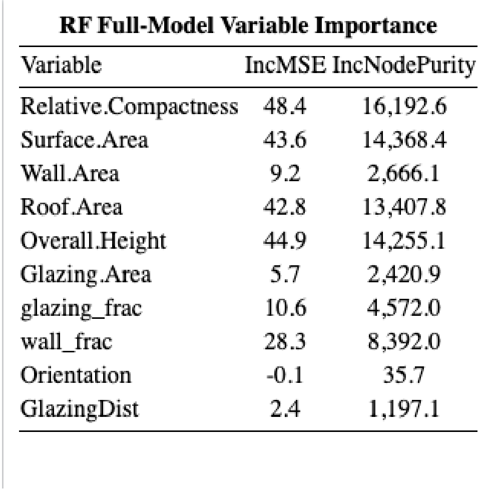
Appendix 4: Correlation matrix for variables



Correlation Matrix: All Numeric Features

	Relative.Compactness	Surface.Area	Wall.Area	Roof.Area	Overall.Height	Glazing.Area	Glazing.Area.Distribution	Y1	Y2	glazing_frac	wall_frac
Relative.Compactness	1.00	-0.99	-0.20	-0.87	0.83	0.00	0.00	0.62	0.63	0.23	0.63
Surface.Area		1.00	0.20	0.88	-0.86	0.00	0.00	-0.66	-0.67	-0.23	-0.64
Wall.Area			1.00	-0.29	0.28	0.00	0.00	0.46	0.43	-0.04	0.63
Roof.Area				1.00	-0.97	0.00	0.00	-0.86	-0.86	-0.20	-0.93
Overall.Height					1.00	0.00	0.00	0.89	0.90	0.19	0.89
Glazing.Area						1.00	0.21	0.27	0.21	0.96	0.00
Glazing.Area.Distribution							1.00	0.09	0.05	0.21	0.00
Y1								1.00	0.98	0.42	0.87
Y2									1.00	0.36	0.85
glazing_frac										1.00	0.15
wall_frac											1.00

Appendix 5: Full random forest no clusters

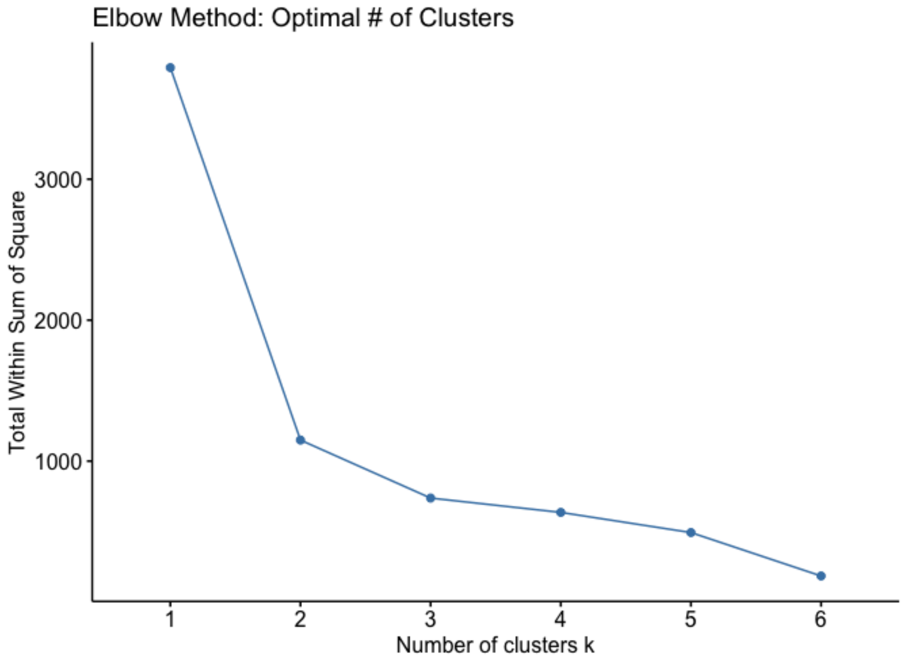


RF Full-Model Variable Importance

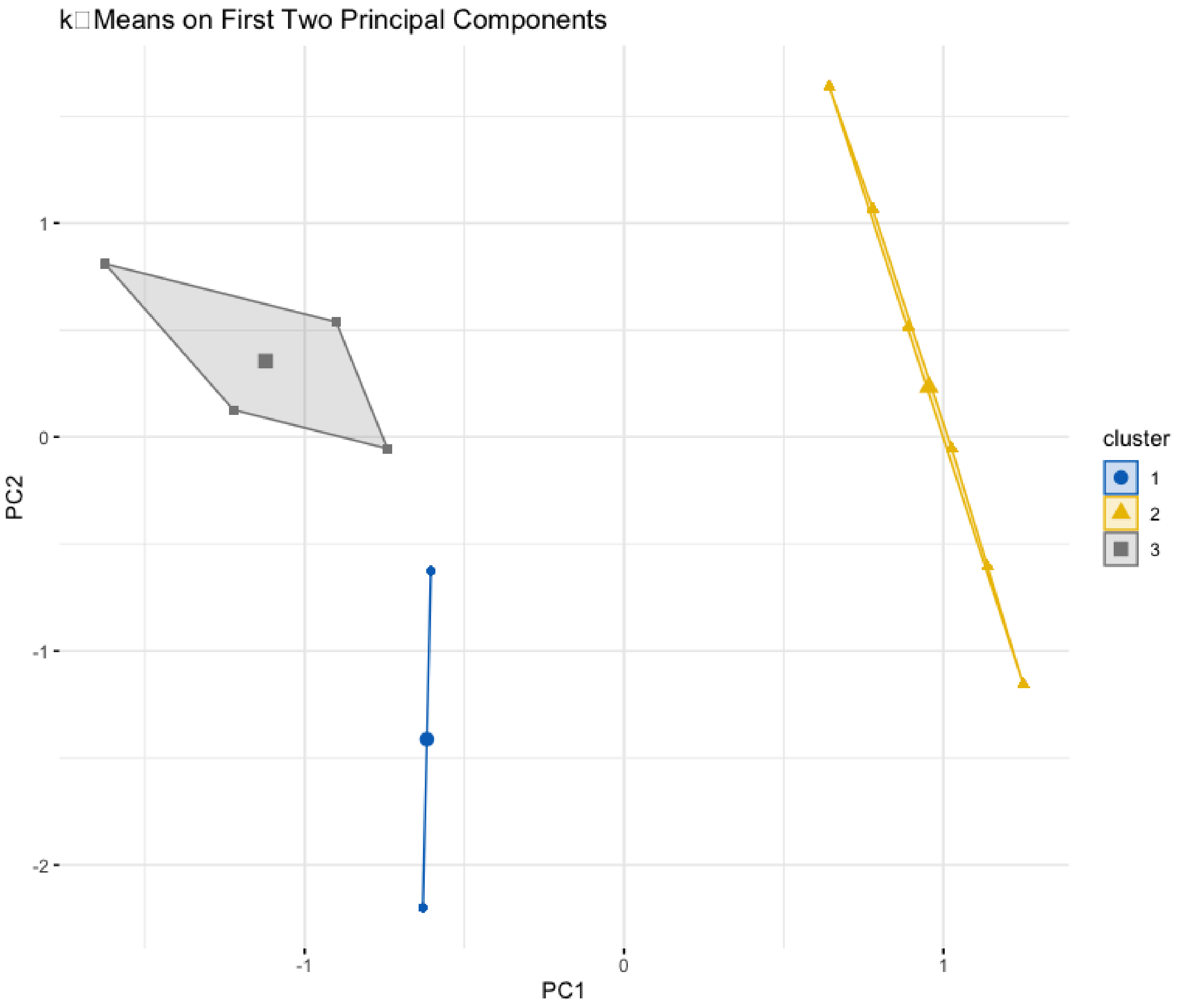
Variable	IncMSE	IncNodePurity
Relative.Compactness	48.4	16,192.6
Surface.Area	43.6	14,368.4
Wall.Area	9.2	2,666.1
Roof.Area	42.8	13,407.8
Overall.Height	44.9	14,255.1
Glazing.Area	5.7	2,420.9
glazing_frac	10.6	4,572.0
wall_frac	28.3	8,392.0
Orientation	-0.1	35.7
GlazingDist	2.4	1,197.1

Full RF OOB MSE: 0.318

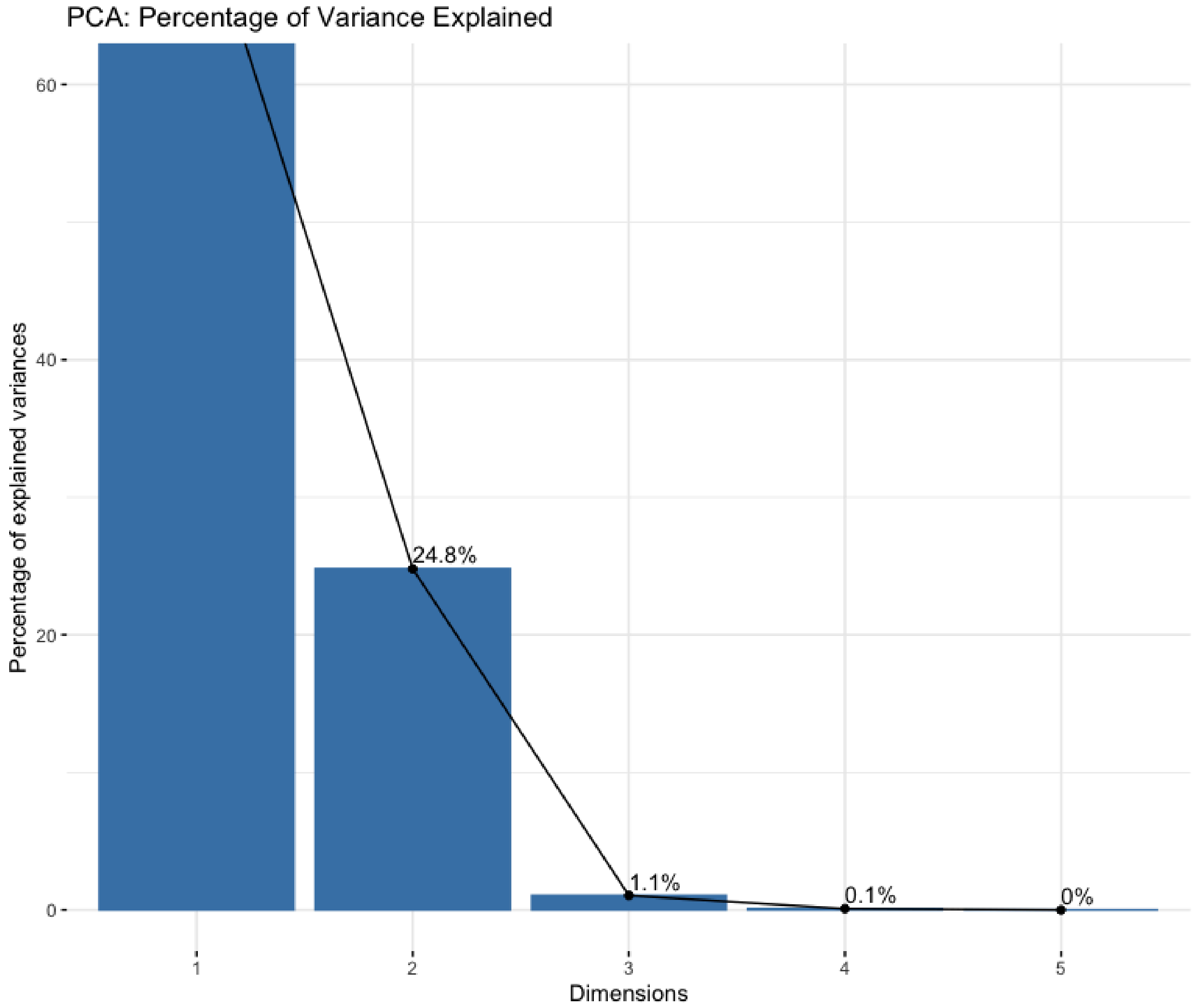
Appendix 6: Elbow method for the optimal amount of clusters



Appendix 7: PCA based k-means Cluster visualization



Appendix 8: PCA Scree Plot: Percentage of Variance Explained by each principal component



Appendix 9: Random forest regression results and importance for each cluster

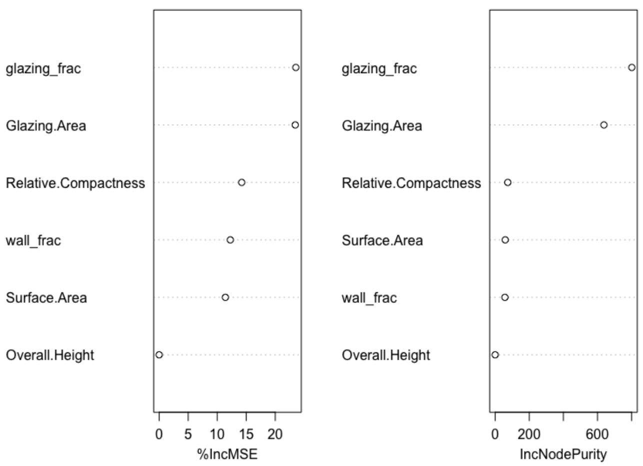
Cluster 1 Variable Importance

Variable		IncNodePurity
glazing_frac	11.2	801.3
Glazing.Area	8.6	637.8
Relative.Compactness	1.2	74.0
Surface.Area	1.0	59.0
wall_frac	0.9	57.1
Overall.Height	0	0

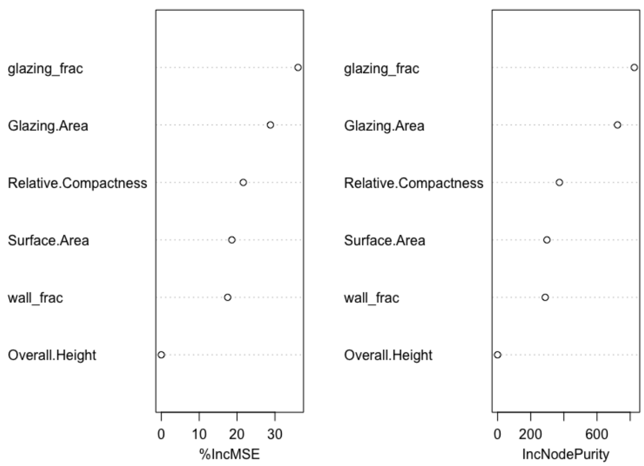
Cluster 2 Variable Importance

Variable		IncNodePurity
glazing_frac	4.0	825.0
Glazing.Area	3.7	723.8
Relative.Compactness	2.4	373.1
Surface.Area	1.9	297.9
wall_frac	1.8	287.5
Overall.Height	0	0

Cluster 1 – Variable Importance



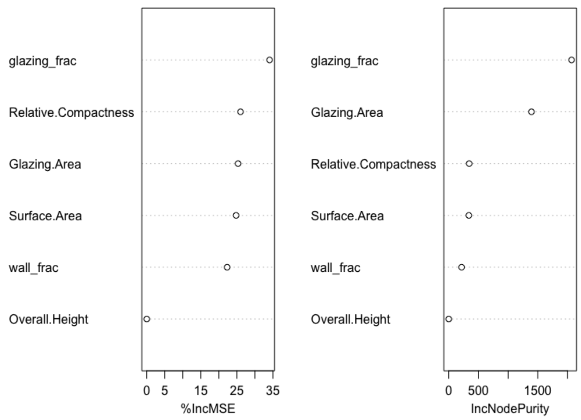
Cluster 2 – Variable Importance



Cluster 3 Variable Importance

Variable		IncNodePurity
glazing_frac	14.9	2,068.2
Glazing.Area	10.0	1,393.1
Relative.Compactness	3.2	344.2
Surface.Area	3.2	338.9
wall_frac	2.4	216.5
Overall.Height	0	0

Cluster 3 – Variable Importance



Sources:

Dataset:

Tsanas, & Xifara. (2012). *Energy efficiency*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/242/energy+efficiency

Retrieved on this susanhub link
https://susanhub.com/datasets/DS01709