Modelling Spatial Heterogeneity in Real Estate Markets

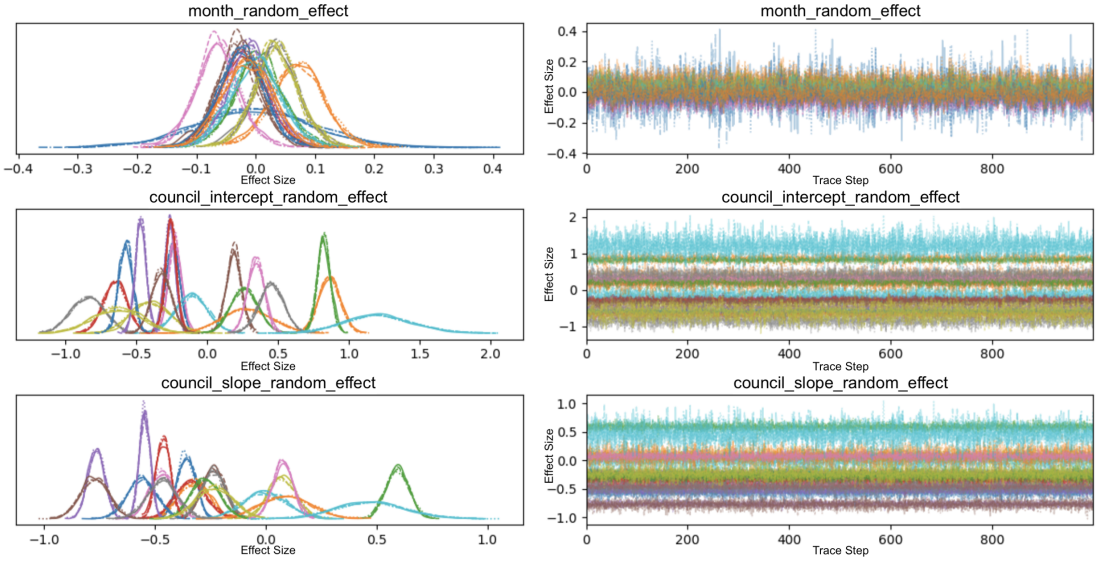
Bocconi Students for Machine Learning

Marco Lomele
marco.lomele@studbocconi.it

Giovanni Mantovani
giovanni.mantovani@studbocconi.it

Letizia Dimonopoli
letizia.dimonopoli@studbocconi.it

Sofia Villa
sofia.villa@studbocconi.it



1 Introduction and Research Goal

Real estate is one of the few markets that can truly impact the economy of a country. Therefore, policy makers, businesses, and consumers look to understand its dynamics and predict its movements. A key metric that stakeholders track is sale prices. Houses have certain features, all of which have an impact on the price. Morevoer, as a city evolves, different areas emerge, with some being more premium and expensive than others.

The aim of this research is to identify which house features have the most variable effect on the price across Melbourne's council areas. We hypothesize that the importance of house features is heterogeneous across space, meaning that features have varying value across Melbourne's geography. For example, parking spaces are more valuable in the city center, due to space constraints, compared to the suburbs, where space is abundant. Understanding this spatial heterogeneity would benefit all types of stakeholders in real estate.

2 Data

2.1 Exploration and Imputation

The provided dataset contains 13580 rows and 21 columns, with missing values in `Car`, `Building Area`, `Year Built` and `Council Area`. We decide to impute the values that are missing using Bayesian Imputation, in order not to lose anything that could be valuable in the analysis.

In fact, removing all rows with missing values would lead to a 54.37% loss of observations in the dataset. We use Bayesian imputation due to its robustness to outliers, and we proceed with outlier detection and treatment in a subsequent section.

Initially, we planned to impute the missing `Car` values using the average for each `Council Area`. However, after analyzing the data, we found that the proportion of missing `Car` values in all Council Areas was zero. Therefore, we impute using the average `Car` value within a 10km radius. For `Building Area`, we use the correlated features `Landsize`, `Rooms`, `Price` and `Type` for imputation, applying a truncated Normal distribution to ensure non-negative values. Missing `Council Area` values are imputed by assigning the council of the suburb. For the 29 suburbs with multiple councils, we compute the mean geographic location (using `Longitude` and `Latitude`) and impute based on proximity. For `Year Built`, we fill missing values using the mean `Year Built` value per `Council Area` and `Type`, capturing differences across administrative zones and property types. In cases where data is missing for a specific combination, we use the average year per `Council Area` or, if needed, per `Type`. Finally, we round all the discrete variables and set the appropriate variable types.

2.2 Analysis

Next, we examine correlations among the variables in the dataset. The heatmap (see Figure 1 in the Appendix) highlights notable relationships: `Rooms`, `Bedroom2`, `Bathroom`, and `Car` exhibit positive correlations, suggesting that properties with more rooms tend to have additional bedrooms, bathrooms, and car spaces, which aligns with expectations for larger properties. `Landsize` and `Building Area` also display a moderate positive correlation, indicating that larger lands typically have bigger buildings. `Price` correlates moderately with `Building Area` and `Rooms`, suggesting these factors significantly influence property prices. In contrast, `Year Built` shows weak correlations, likely due to variations in design, features, and external factors like location and renovations.

To ensure our analysis remains robust and representative of general trends, we remove outliers using the Interquartile Range (IQR) method. This approach allows us to filter out extreme values for each numerical variable, focusing on observations within 1.5 times the IQR from the lower and upper quartiles. By excluding these rare occurrences, we reduce the skewness in our data, particularly evident in variables like `Price`, `Landsize`, and `Building Area`. This step is crucial in refining our dataset to better capture the central tendencies and relationships of interest. Moreover, managing outliers enhances model accuracy by ensuring that extreme values do not skew coefficients.

We proceed with producing a series scatter plots to infer potential relationships. We observe that as `Landsize` increased, `Price` generally goes up, though variability suggested other factors influence property values. A similar trend appears between `Price` and `Building Area`, with diminishing returns at larger sizes. For a detailed visual comparison of the data before and after outlier treatment, as well as to see the scatter-plots, please refer to figures 3 to 7 in the Appendix.

We focus our study on the following variables: `Price`, `Building Area`, `Council Area`, `Distance`, `Car`, `Bathroom`, `Bedroom2`, `Month`. These are chosen because: (1) they are all numerical, thus easier to manipulate; (2) they represent house characteristics and exclude external factors such selling skills of the real estate agent, except for time; (3) they are all possibly correlated with `Price`. Moreover, we restrict our attention to the active council areas, i.e.where the total number of properties sold exceeds 200. This threshold corresponds to an average of approximately 8 properties sold per month.

Finally, we test our spatial heterogeneity hypothesis by we creating a correlation barplot to analyze the relationship between our selected house feature and house prices, segmented by council areas. 1 The plot shows that the influence of covariates on house prices varies across council areas, thus supporting our hypothesis and emphasizing the importance of our research.

2.3 Preparation

Before modelling, we prepare the data via a couple of processing steps. First, we retain the IQR filter, thus excluding rare occurrences and focusing our research on typical patterns, which in turn improves model stability. Then, we apply standardization (or z-score transformation) on the data by subtracting the mean and dividing by the standard deviation for each variable. This removes differences in units and magnitudes between the variables, enabling an unbiased interpretation and fair comparison of the estimated coefficients.

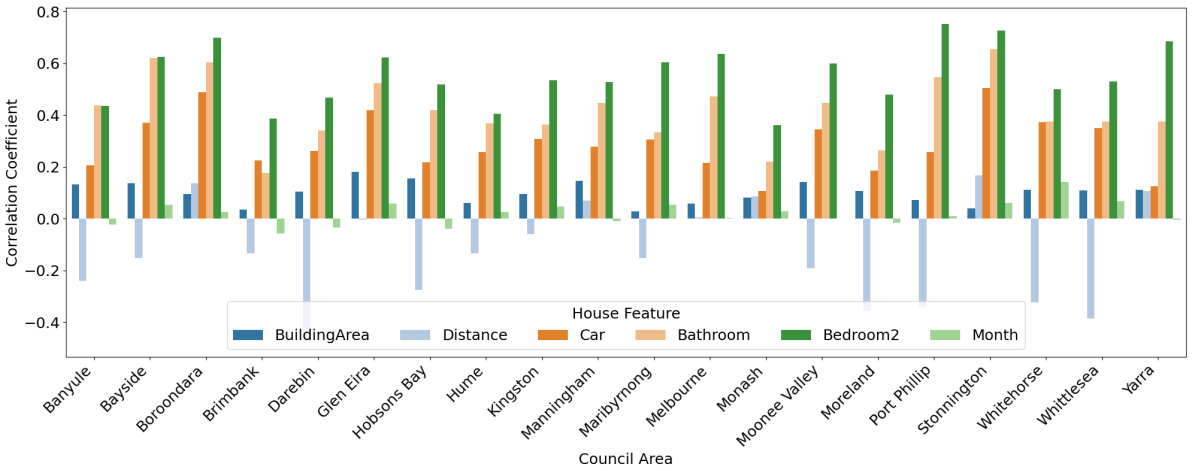


Figure 1: Correlation of covariates with Price segmented by the selected council areas

3 Modelling

3.1 Hierarchical – Mixed Effects Model

For our research we employ a hierarchical-mixed effects model. The hierarchical structure offers a principled way to accommodate council-level heterogeneity. [3] The mixed effects framework allows us to distinguish between house features with a fixed effect on the price, i.e. constant across council areas, and the features with a random effect on the price, i.e. varying across across council areas. [4]

Let each house sale i be characterized by a sale price $y_i \in \mathbb{R}$ (our target variable), a set of features $\mathbf{X}_i \in \mathbb{R}^F$ having a fixed effect on y_i, and a feature $Z_i \in \mathbb{R}$ having random effect on y_i across council areas. Furthermore, we define two grouping factors to account for spatial and temporal effects. Let index $j(i)$ indicate the council area in which i is located in, for $j(i) \in \{1, ..., J\}$ levels, where J is the number of council areas, and let index $m(i)$ be the month when the sale of house i is recorded, for $m(i) \in \{0, ..., 11\}$ levels. Then, we model the sale price y_i of house i as

$$y_i \sim N(t_{m(i)} + u_{j(i)} + X_i^T \beta + Z_i v_{j(i)}, \quad \sigma_\epsilon^2)$$

The model's parameters of interest are:

- $t_{m(i)} \in \mathbb{R}$, the intercept random effect specific to month $m(i)$;
- $u_{j(i)} \in \mathbb{R}$, the intercept random effect specific to council $j(i)$;
- $\beta \in \mathbb{R}^{\mathbb{F}}$, the slope fixed effects of the features X_i;
- $v_{j(i)} \in \mathbb{R}$, the slope random effect of the feature Z_i specific to council $j(i)$;
- $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, the error term accounting for the variance in y_i unexplained by the model.

3.2 Bayesian Priors and Non-Centered Parametrization

Following the Bayesian paradigm, we define prior distributions over the parameters of interest of the model to incorporate information and quantify uncertainty. We begin with classical priors on the global standard deviations for t, u, v, y_i, which indicate how much a variable can vary at the population level, and β:

$$\sigma_t, \sigma_u, \sigma_v, \sigma_\epsilon \sim InverseGamma(2, 1)$$
$$\beta_f \sim N(0, 1) \qquad \text{for fixed effect feature } f = 1, ..., F$$

Next, we apply non-centered parametrization. First, it supports inference, flattening out the space of parameter values to be explored. [8] Second, it allows us to characterize the prior distribution of a parameter as a deterministic sum of its global mean and an offset component for each level of the random effect. [2] For simplicity, we set all global means to 0 and assume full independence between group levels.

$$t_{m(i)} = 0 + \sigma_t \times t_{offset,m(i)} \qquad t_{offset,m(i)} \sim N(0, 1)$$

$$v_{j(i)} = 0 + \sigma_v \times v_{offset,j(i)} \qquad v_{offset,j(i)} \sim N(0,1)$$

For the prior distribution of $u_{j(i)}$, we implement a Conditional Autoregressive (CAR) component. This captures the spatial dependencies and accounts for the spill-over effects among the council areas. Intuitively, if house i neighbors a premium council area, then y_i will reflect part of this premium because i is physically close to the factors that make the neighbor council area exceptionally valuable.

We represent the spatial relationships between council areas with an adjacency matrix $A \in \mathbb{R}^{J\times J}$, where $A[k,l] = 1$ if councils k and l are adjacent, and 0 otherwise, with a forced 0-diagonal. Then, we encode the variance-covariance structure of the prior distribution using precision matrix Q and define the joint probability distribution of $u_{j(i)}$ for all $j(i)$ as

$$\mathbf{u} \sim MultivariateNormal(\mathbf{0}, Q^{-1}) \qquad Q = \alpha^2[D - \rho A]^{-1}$$

Here, $D \in \mathbb{R}^{J\times J}$ is a diagonal matrix holding the count of neighbors per council, while α^2 and ρ are a spatial variance and a smoothing parameter respectively. Using Brook's theorem, Besag (1974) showed that the joint probability distribution defined above is equivalent to the more common definition of CAR that uses the conditional distribution $u_{j(i)}|u_{n_j} \ \ \forall n_j \in \mathcal{N}(j(i))$, where $\mathcal{N}(j(i))$ is the set of adjacent council areas to $j(i)$. [1] This is true only when Q is symmetric and positive definite, which in turn bounds ρ between the reciprocal of the smallest eigenvalue and the reciprocal of largest eigenvalue of A. [5] However, for simplicity, we fix $\rho = 0.1$ and $\alpha = 0.09$.

3.3 Inference

The purpose of having one slope random effect per council area is to simplify inference. A fully specified model with random slopes for all the five selected regressor would require estimating $J \times F$ random slope parameters, which risks overfitting and computational instability, as well produces an unidentifiable covariance matrix.

To address this limitation and satisfy our research goal, we employ the following approach: fit $F + 1$ separate models, each using a different house feature Z_i for the slope random effect; then, collect the inferred coefficients $v \in \mathbb{R}^J$ for the slope random effect across all $F + 1$ models and compare their variances. Note that $F + 1$ is the number of house features at our disposal: `Building Area`, `Distance`, `Bedroom2`, `Bathroom`, and `Car`.

This approach requires some assumptions. Namely, (i) that slope random effects across different features are uncorrelated, (ii) that month-specific intercept random effects don't interact with slope random effects, (iii) that fixed effects estimates β remain stable regardless of which feature has random slopes, (iv) that CAR prior spatial relationships between councils are feature-independent, and (v) that errors are homoscedastic between the models.

We implement and fit the models with the PyMC library in Python.[7] In particular, we use the No-U-Turn Sampler (NUTS), a Hamiltonian Monte Carlo algorithm that dynamically adjusts step sizes and trajectory lengths to robustly explore complex, high-dimensional posterior distributions. [6] We configure the sampler with 3,000 posterior draws, 3,000 tuning steps, a target acceptance rate of 0.95 to avoid divergent transitions, and a maximum tree depth of 15 to balance precision and computational cost.

Then, we perform the inference across four independent chains to enable consistency checks through convergence diagnostics, including the Gelman-Rubin $\hat{R}$ statistic, the effective sample size, and the Highest Density Interval (HDI). Finally, we extract the posterior distributions from the sampler's trace and analyze the results.

4 Results and Conclusion

4.1 Interpretation of Results

The standard deviations of the random slopes across council areas, which quantify spatial heterogeneity, along with other metrics, are summarized in Table 1.

The largest standard deviation is observed for `Distance` (0.3484), indicating significant spatial variability in its impact on house prices, followed by `Bedroom2` (0.2532), reflecting differing preferences for additional bedrooms across council areas. In contrast, `Building Area` has the lowest standard deviation (0.0489), suggesting a uniform influence on property values. Moderate variability is found for `Car` (0.1508) and `Bathroom` (0.1651), highlighting some spatial differences in their effects. The mean, minimum, and maximum values of the slope random effect per house feature give a further insight into how the same characteristic can have effect with opposite sign in different council areas, further supporting our spatial heterogeneity hypothesis.

Feature	Standard Deviation	Mean	Minimum Value	Maximum Value
Distance	0.3484	-0.2412	-0.7800	0.5370
Bedroom2	0.2532	0.5139	0.1480	1.0180
Bathroom	0.1651	0.0689	-0.1080	0.4660
Car	0.1508	0.0627	-0.0920	0.5170
Building Area	0.0489	0.0545	-0.0240	0.1760

Table 1: Standard deviations and other metrics of slope random effects for selected house characteristics.

4.2 Robustness of Estimates

Convergence diagnostics confirm the reliability of the results, with $\hat{R} \approx 1$ for all parameters and effective sample sizes exceeding 5,000. We proceed with validating the model assumptions to gain a deeper insight into the validity of the results. First, we compare the means and variances of the posterior distributions of the β coefficients and find that the differences across models are statistically insignificant, thus validating assumption (iii) (see tables 2 and 3). Then, we investigate the intercept random effects for the month ($t_{m(i)}$) and for the council areas ($u_{j(i)}$) and find similar robustness across models, supporting assumptions (i) and (iv) (see tables 7 and 6). Finally, we compare global variance across models to check for homoscedastic errors, and again find supporting evidence (see tables 4 and 5). These findings give us confidence in the results and demonstrate the existence of spatial heterogeneity in the effects of house characteristics on prices.

Before concluding, we would like to identify future research opportunities with regards to spatial heterogeneity. On the one hand, more nuanced models could be explored, for instance with multiple slope random effects or a learnable precision matrix in the CAR component. On the other hand, the dimensionality of the dataset could be expanded, including additional layers of hierarchy and additional factors affecting house prices.

5 References

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B, 36, 192–236*, 1974.

[2] Michael Betancour. Hamiltonian monte carlo for hierarchical models. *Department of Statistical Science, University College London*, 2013.

[3] Maria M. Ciarleglio and Robert W. Makuch. Hierarchical linear modeling: An overview. *Child Abuse Neglect, Volume 31, Issue 2*, 2007.

[4] Silveira L. T. Y. D., Ferreira J. C., and C. M. Patino. Mixed-effects model: a useful statistical tool for longitudinal and cluster studies. *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia, 49(2), e20230137.*, 2023.

[5] Ephraim M. Hanks et al. On the relationship between conditional (car) and simultaneous (sar) autoregressive models. *Department of Statistics, The Pennsylvania State University*, 2017.

[6] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Department of Statistics – Columbia University*, 2011.

[7] PyMC Team. Pymc – probabilistic programming library. *link: https://www.pymc.io/welcome.html*, 2024.

[8] Thomas Wiecki. Why hierarchical models are awesome, tricky, and bayesian. *blog: https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/*, 2017.

6 Appendix

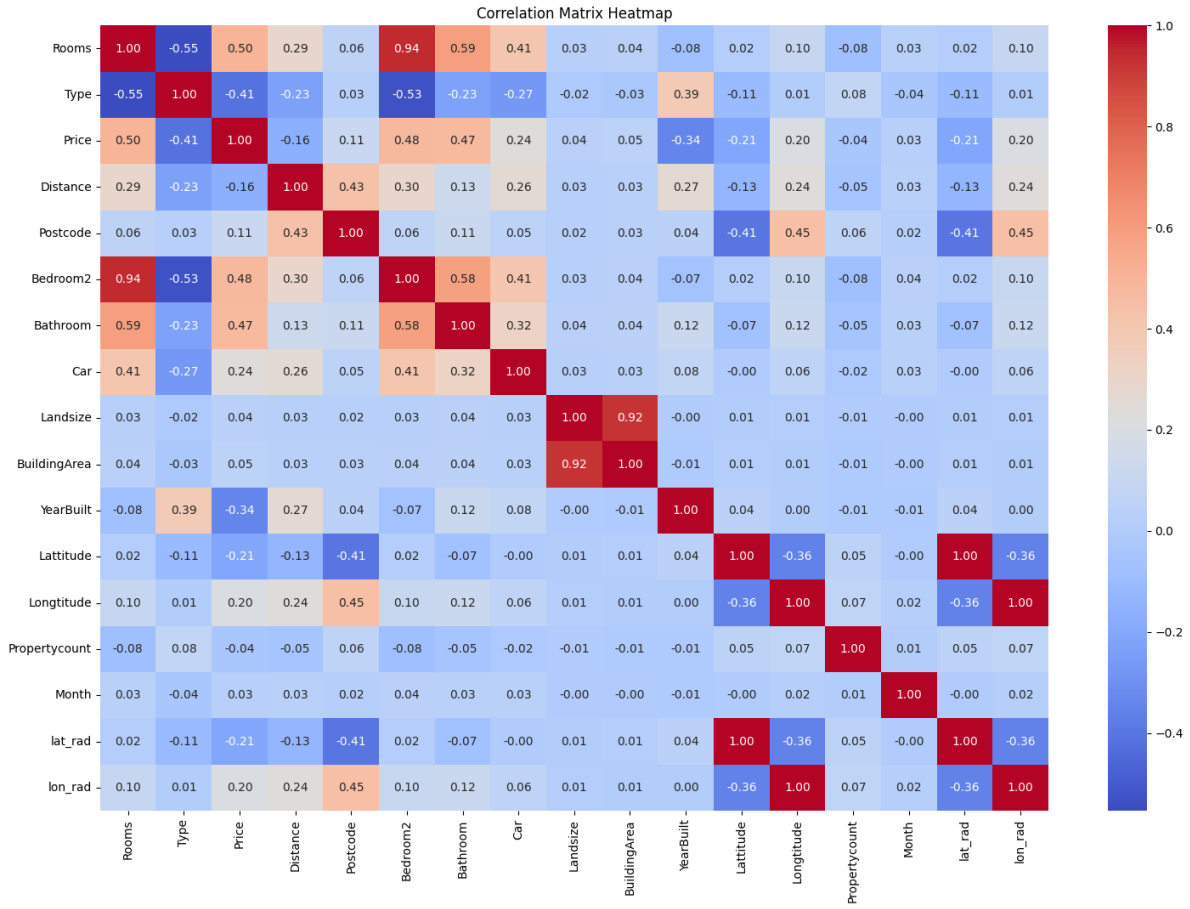


Figure 2: Correlation Heatmap

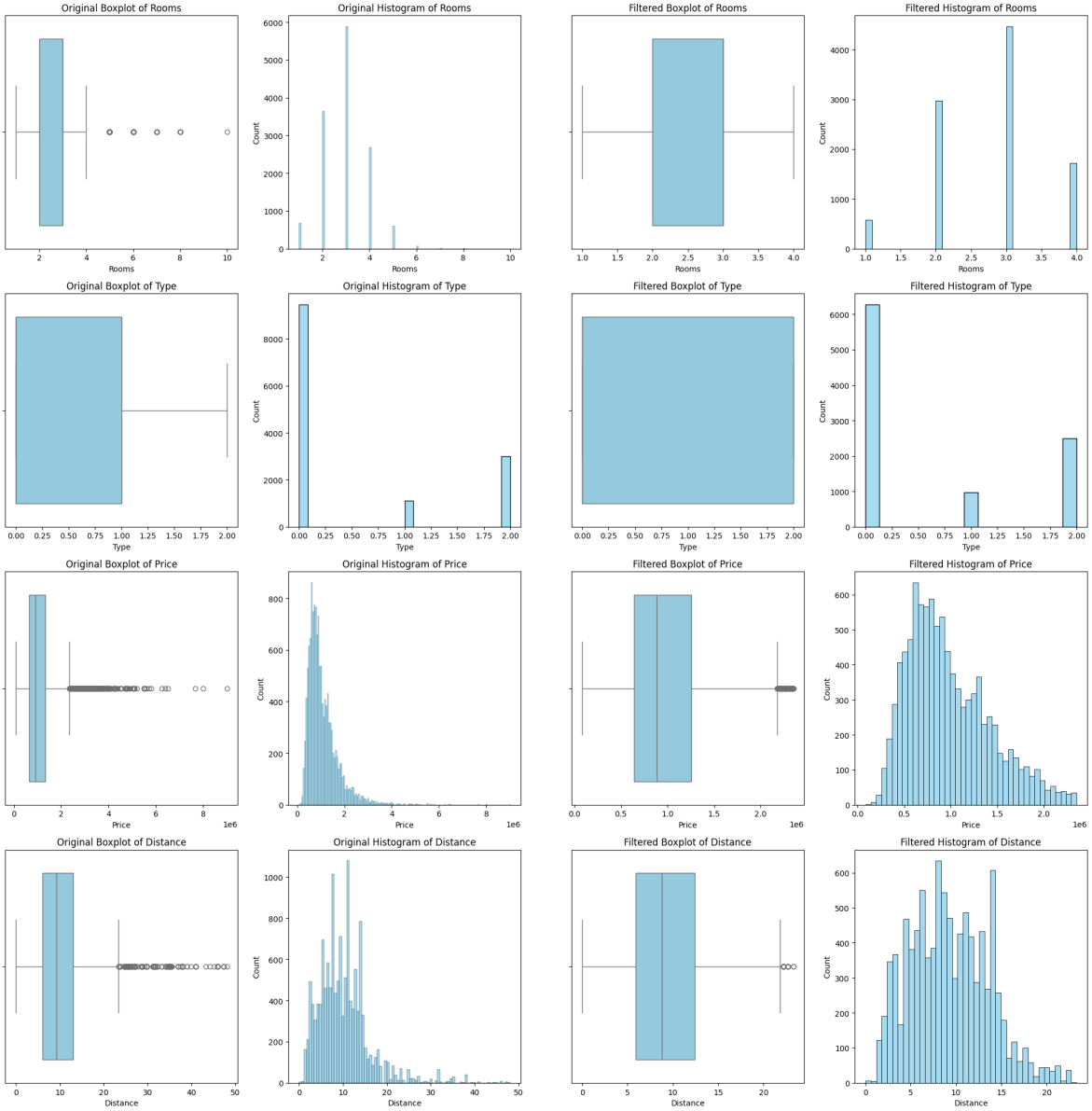


Figure 3: Boxplot and Histogram for the all variables before and after Inter Quartile Range filtering. (1/4)

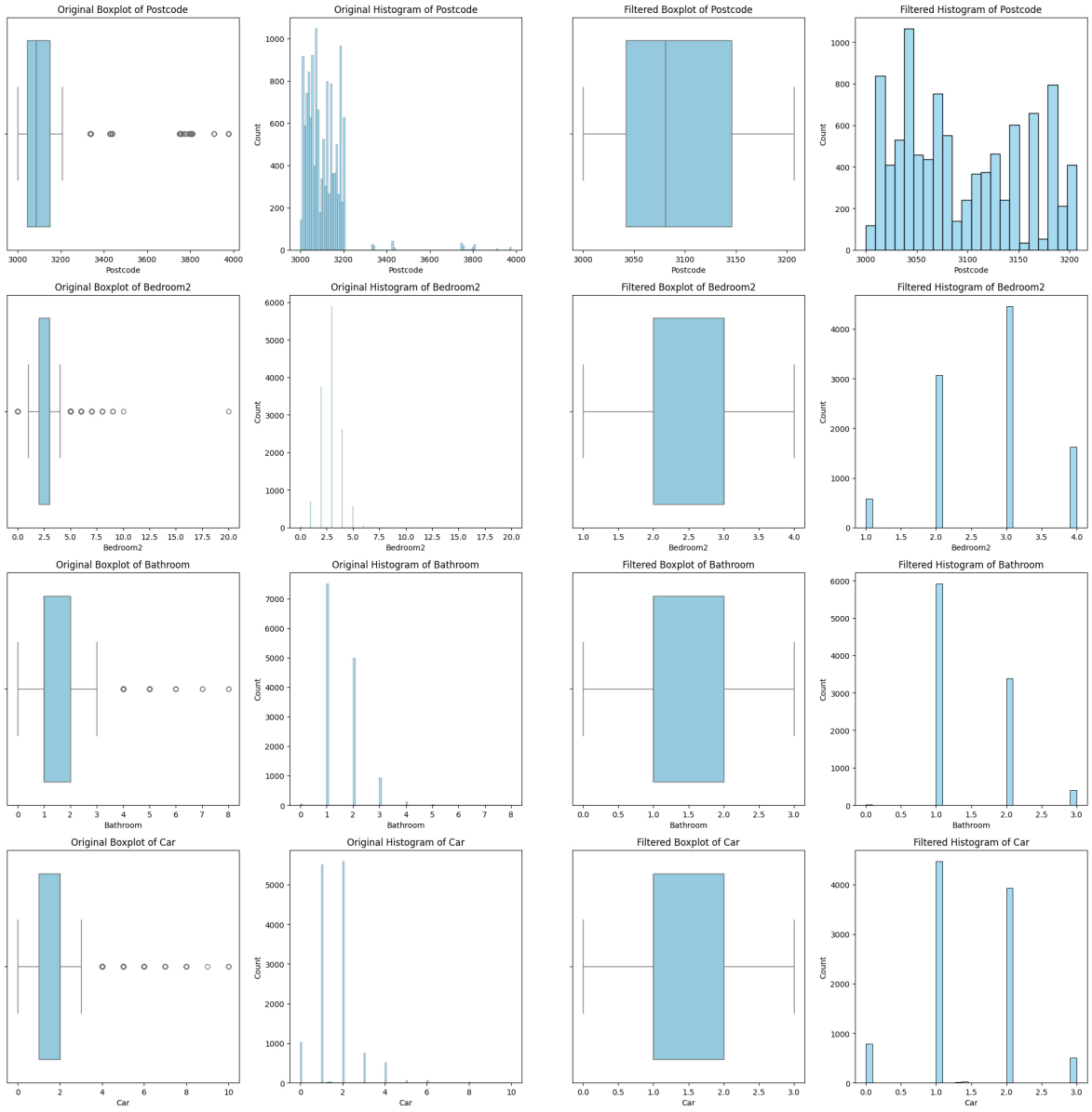


Figure 4: Boxplot and Histogram for the all variables before and after Inter Quartile Range filtering. (2/4)

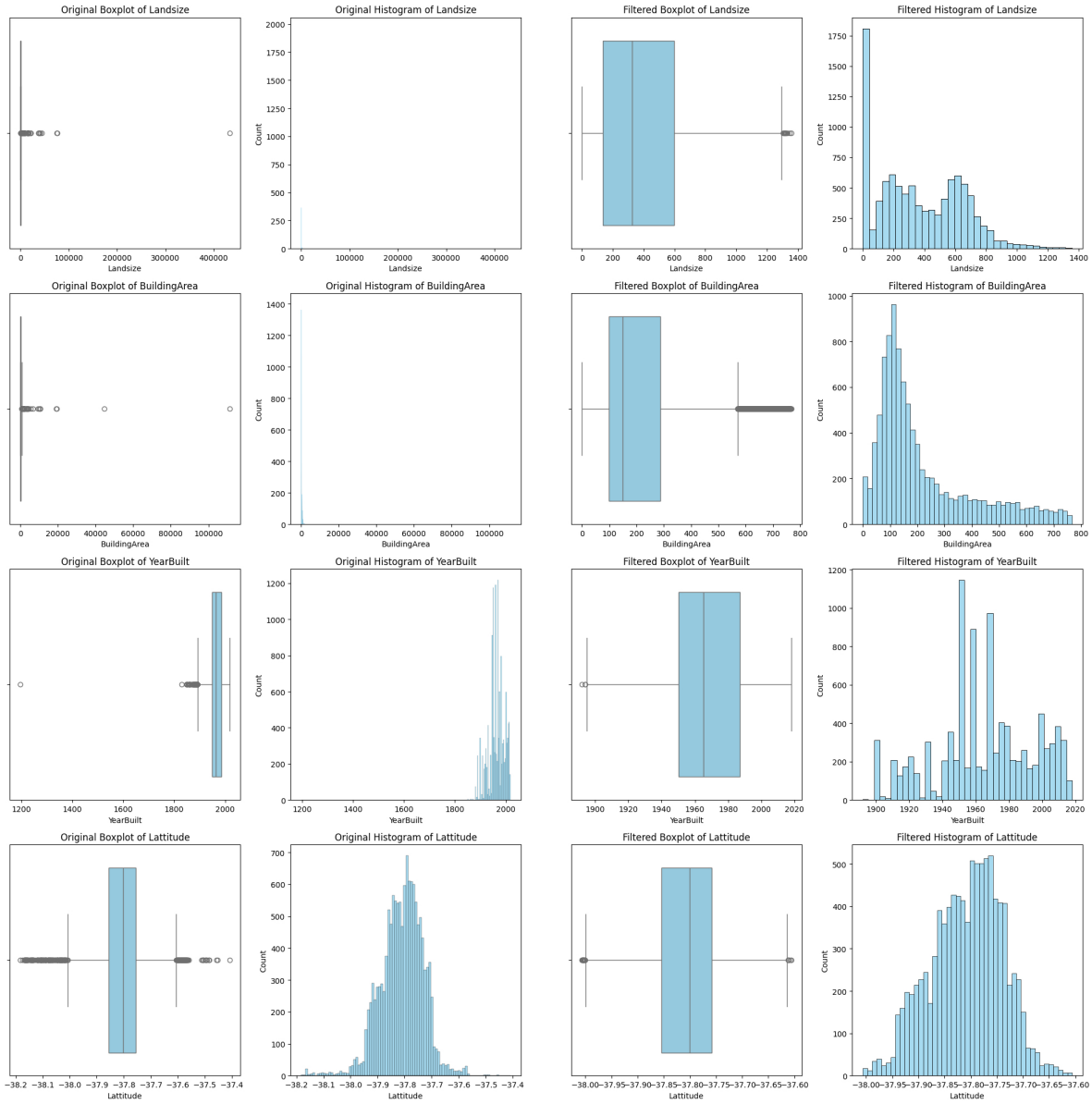


Figure 5: Boxplot and Histogram for the all variables before and after Inter Quartile Range filtering. (3/4)

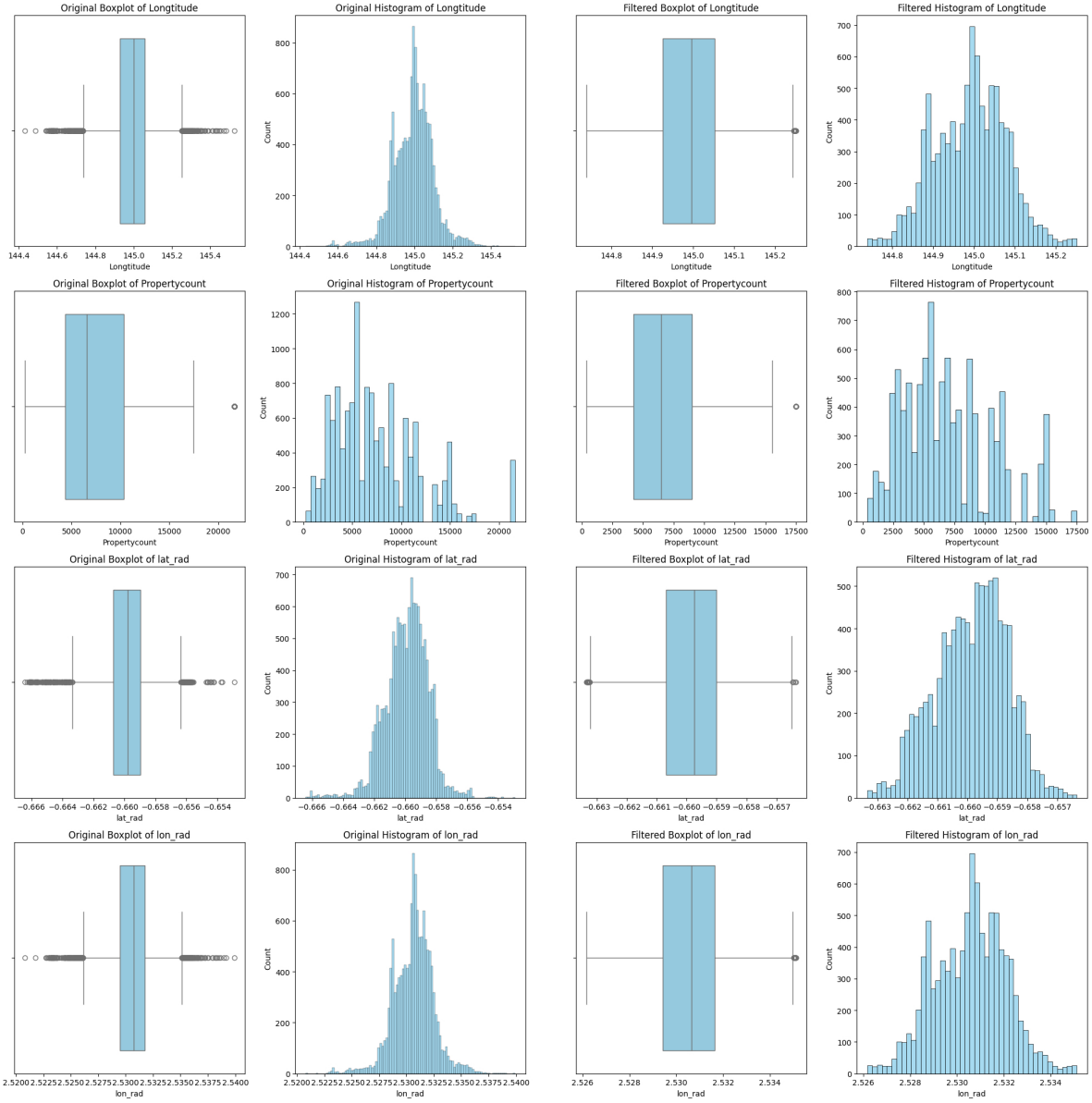


Figure 6: Boxplot and Histogram for the all variables before and after Inter Quartile Range filtering. (4/4)

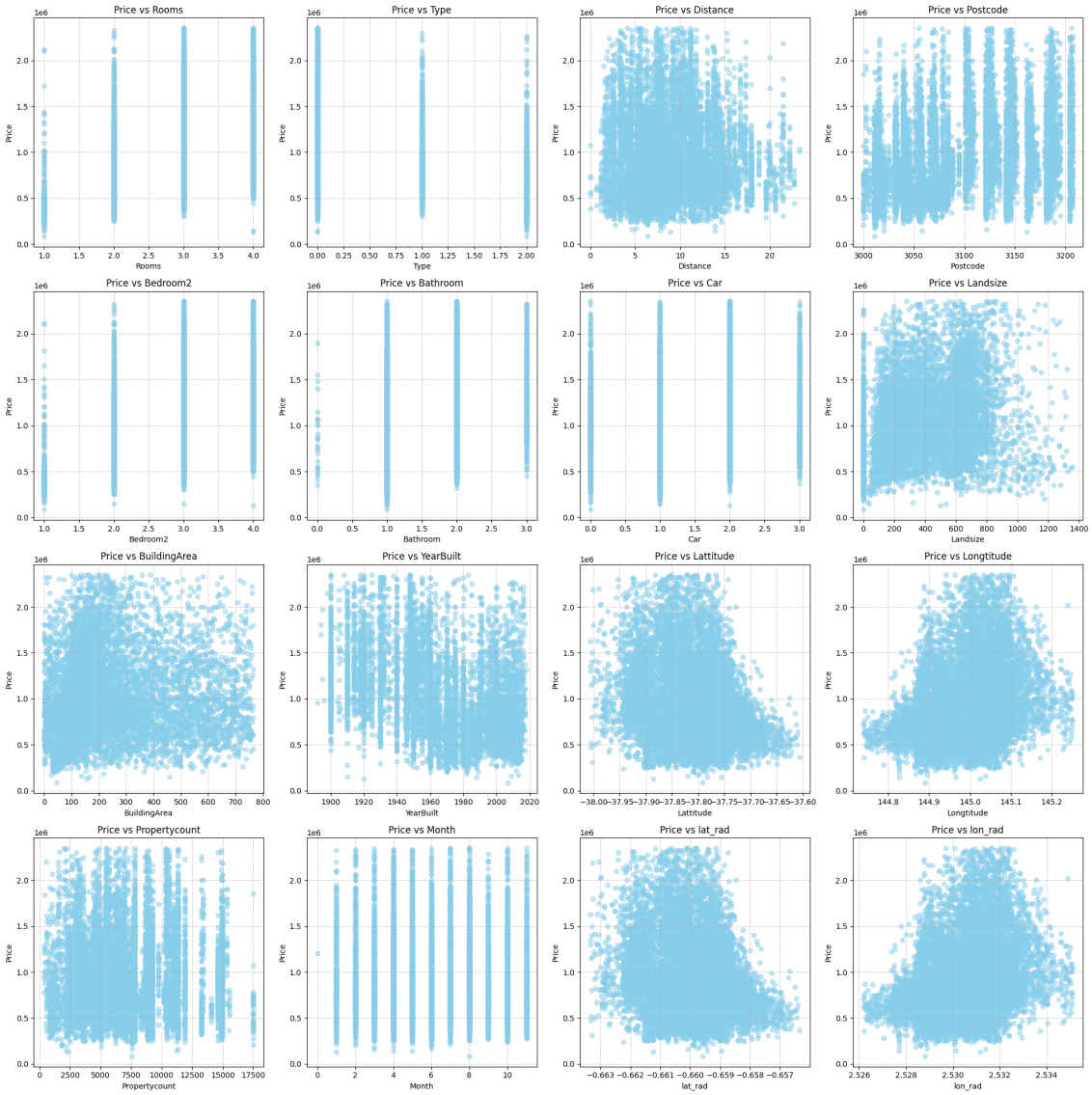


Figure 7: Scatter plots between price and all other house features.

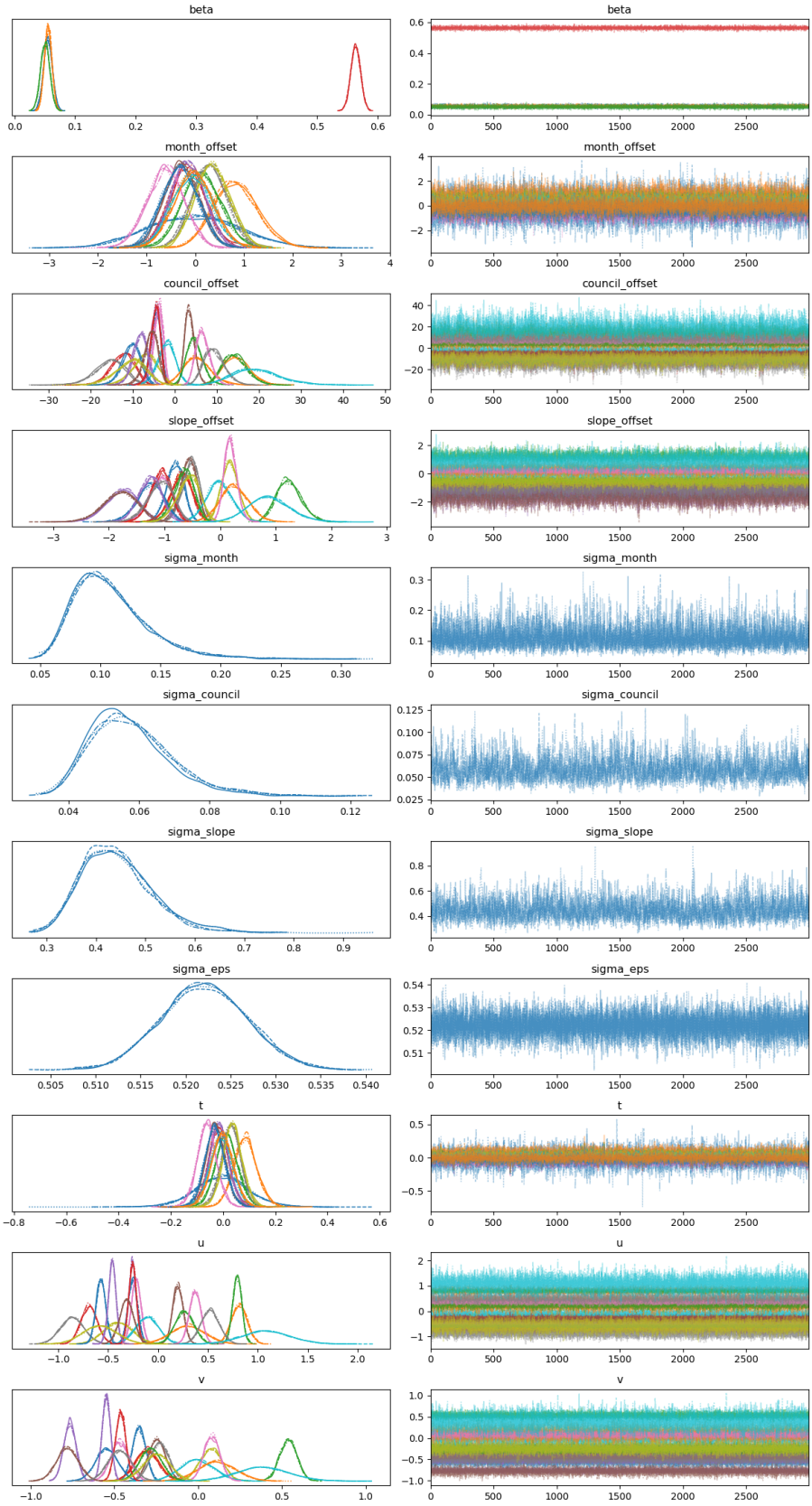


Figure 8: Trace plots and posterior distributions for model parameters with distance having a random slope.

Note: for tables 2, 3, 4, 5, 6, 7, 8, when we write "{House Feature} Model" in the header we refer to the instance of the model where {House Feature} is modeled as the slope random effect. For instance, the "Car Model" column refers to a model where the house feature Car is modeled with a random slope v, while all the other house features are modeled as fixed effects in β.

Mean values	**Building Area Model**	**Bathroom Model**	**Bedroom Model**	**Distance Model**	**Car Model**
Beta_CAR	0.047	0.039	0.051	0.055	-
Beta_DISTANCE	-0.375	-0.364	-0.360	-	-0.377
Beta_BATHROOM	0.062	-	0.056	0.051	0.055
Beta_BUILDINGAREA	-	0.056	0.058	0.055	0.053
Beta_BEDROOM	0.577	0.560	-	0.564	0.559

Table 2: Mean values for fixed effects across models.

Standard Deviation values	**Building Area Model**	**Bathroom Model**	**Bedroom Model**	**Distance Model**	**Car Model**
Beta_CAR	0.007	0.007	0.007	0.007	-
Beta_DISTANCE	0.012	0.012	0.011	-	0.012
Beta_BATHROOM	0.008	-	0.007	0.008	0.008
Beta_BUILDINGAREA	-	0.006	0.006	0.006	0.006
Beta_BEDROOM	0.008	0.008	-	0.008	0.008

Table 3: Standard deviations for fixed effects across models.

Sigma Eps	**Building Area Model**	**Bathroom Model**	**Bedroom Model**	**Distance Model**	**Car Model**
Mean Sigma Eps	0.54	0.523	0.501	0.522	0.525
Std. Dev Sigma Eps	0.005	0.005	0.005	0.005	0.005

Table 4: Sigma Eps values across models.

Sigma Council	Building Area Model	Bathroom Model	Bedroom Model	Distance Model	Car Model
Mean Sigma Council	0.055	0.055	0.054	0.058	0.055
Std. Dev Sigma Council	0.012	0.011	0.011	0.012	0.012

Table 5: Sigma Council values across models.

Council Parameters	Building Area Model	Bathroom Model	Bedroom Model	Distance Model	Car Model
u[0]	-0.237	-0.248	-0.233	-0.253	-0.240
u[1]	0.866	0.823	0.828	0.814	0.818
u[2]	0.511	0.416	0.525	0.794	0.419
u[3]	-0.623	-0.657	-0.452	-0.691	-0.608
u[4]	-0.240	-0.252	-0.268	-0.256	-0.244
u[5]	0.287	0.276	0.287	0.199	0.250
u[6]	-0.229	-0.232	-0.224	-0.219	-0.222
u[7]	-0.611	-0.613	-0.428	-0.861	-0.584
u[8]	0.504	0.505	0.540	-0.399	0.505
u[9]	0.253	0.317	0.503	-0.108	0.290
u[10]	-0.512	-0.522	-0.523	-0.574	-0.520
u[11]	-0.407	-0.394	-0.336	0.297	-0.512
u[12]	0.374	0.418	0.427	0.264	0.400
u[13]	-0.253	-0.252	-0.252	-0.256	-0.254
u[14]	-0.445	-0.483	-0.447	-0.459	-0.450
u[15]	0.049	0.119	0.314	-0.314	0.016
u[16]	0.032	0.125	0.342	0.373	0.126
u[17]	0.412	0.417	0.404	0.521	0.424
u[18]	-0.336	-0.341	-0.170	-0.583	-0.285
u[19]	-0.057	-0.067	0.039	1.086	-0.179

Table 6: Council parameters $u[.]$ for every council area, across models.

Month parameters	Building Area Model	Bathroom Model	Bedroom Model	Distance Model	Car Model
t[0]	0.001	-0.002	-0.001	-0.004	0.001
t[1]	0.076	0.078	0.070	0.085	0.061
t[2]	0.013	0.002	0.008	0.013	0.006
t[3]	-0.012	-0.002	0.005	-0.017	-0.007
t[4]	-0.009	-0.017	-0.025	-0.016	-0.012
t[5]	-0.024	-0.023	-0.020	-0.030	-0.023
t[6]	-0.063	-0.051	-0.053	-0.063	-0.054
t[7]	0.028	0.023	0.025	0.030	0.032
t[8]	0.054	0.052	0.051	0.036	0.056
t[9]	-0.008	-0.020	-0.014	-0.008	-0.016
t[10]	-0.039	-0.040	-0.039	-0.033	-0.033
t[11]	-0.009	-0.006	0.001	-0.004	-0.013

Table 7: Month parameters $t[.]$ for every month, across models.

Random slopes	**Building Area Model**	**Bathroom Model**	**Bedroom Model**	**Distance Model**	**Car Model**
v[0]	0.008	0.042	0.402	-0.353	0.038
v[1]	0.113	0.210	0.745	-0.303	0.207
v[2]	0.176	0.370	1.018	0.537	0.517
v[3]	-0.003	-0.100	0.148	-0.307	-0.022
v[4]	0.015	0.032	0.443	-0.765	-0.020
v[5]	0.076	0.135	0.688	-0.247	0.242
v[6]	0.040	-0.008	0.421	-0.466	-0.009
v[7]	-0.024	0.030	0.231	-0.230	0.010
v[8]	-0.004	-0.077	0.313	0.077	0.043
v[9]	0.065	-0.029	0.259	-0.007	0.016
v[10]	0.026	-0.044	0.402	-0.545	-0.052
v[11]	0.108	0.159	0.648	0.110	-0.057
v[12]	0.046	-0.108	0.425	-0.283	0.010
v[13]	0.047	0.055	0.523	-0.463	0.071
v[14]	0.045	-0.098	0.347	-0.548	-0.025
v[15]	0.100	0.350	0.874	-0.780	0.056
v[16]	0.036	0.466	0.941	0.081	0.333
v[17]	0.105	0.016	0.542	-0.468	0.032
v[18]	0.036	-0.036	0.182	-0.248	-0.044
v[19]	0.078	0.012	0.726	0.385	-0.092

Table 8: Random slope $v[.]$ for every council area, across models.