✅Mixed-Effects Models (Hierarchical Linear Models) – These models are particularly well-suited for your research question because they explicitly model the nested structure of houses within districts. By specifying random slopes for each characteristic across districts, you can directly measure the variance in effect sizes. The model allows you to decompose the variation in house prices into within-district and between-district components, while simultaneously estimating both district-specific effects and overall population effects. This approach naturally handles the multilevel structure of your data and provides direct estimates of the variance in effects across districts.

<https://www.pythonfordatascience.org/mixed-effects-regression-python/>

<https://en.wikipedia.org/wiki/Mixed_model>

Geographically Weighted Regression (GWR) – This model extends traditional regression by allowing coefficients to vary continuously across space rather than discreetly by district. GWR fits local regressions for each location using nearby observations, weighted by their distance. This approach could reveal more nuanced spatial patterns in the effect sizes of house characteristics, potentially identifying gradual changes or transitions between areas that might be missed when using district boundaries. The model produces coefficient surfaces that can be mapped to visualize how effects vary across space.

Gaussian Graphical Models.

❌Spatial Durbin Model – This model incorporates both spatial lag and spatial error components while also accounting for spillover effects from neighboring areas. It could capture how the impact of house characteristics in one district might be influenced by characteristics of houses in adjacent districts. This is particularly relevant for your research because the effect of certain characteristics might not be confined to district boundaries - for example, the value of a parking space might be influenced by parking availability in neighboring districts as well.

Not relevant for our research.

✅Bayesian Hierarchical Models – These models offer a flexible framework for modeling district-level variations while incorporating prior knowledge about the housing market. The Bayesian approach allows you to quantify uncertainty in your estimates of effect size variations and can handle complex dependency structures. It's particularly useful when you have some districts with sparse data, as it can borrow strength from other districts through partial pooling. The model can also naturally incorporate your missing value imputation strategy since it's already in a Bayesian framework.

Questions:

CouncilArea 1369 → these are the missing values. How do we want to proceed?

Key notes:

* missing values for CouncilArea occurs at rows:
  + 7584
  + 10797
  + From 12213 → to 13579 → WHY?

From 12213 to 13579 other columns have missing values too:

* Car 4.54%
* BuildingArea 49.96%
* YearBuilt 44.62%

- We can address it through imputation using latitude and longitude

- Clustering → elbow method for choosing k

After Clustering Imputation, in housing.loc[12213:13579]:

Boroondara 467

Moreland 328

Glen Eira 290

Moonee Valley 232

Melton 50

- External sources

- Rename Lattitude in Latitude

Mixed effects model is made of two components:

* Random effects: house features of which effect varies across council area
* Fixed effects: house features of which effect is constant across council area

Nuance: all effects are random; however, some have a high variance, while others have a small variance.

We disregard covariance between random effects across council areas. Let's say we're looking at two random effects: Rooms and Bathrooms. For a given council area j:

* The effect of Rooms (s\_rooms[j]) depends on its neighboring councils' Room effects
* The effect of Bathrooms (s\_baths[j]) depends on its neighboring councils' Bathroom effects
* BUT there's no direct relationship between the Room effect and Bathroom effect, either within a council or across councils

Keeping the effects independent, has several advantages:

1. Computational Efficiency: The model is much easier to fit because each random effect can be sampled somewhat independently
2. Interpretability: The spatial correlation parameter (rho) for each effect has a clear meaning
3. Diagnostic Simplicity: It's easier to identify which effects show strong spatial patterns
4. Robustness: Less chance of the model failing to converge due to complex interactions

Computationally expensive because:

* N observations, data
* K average neighbours per observation, spatial dependencies
* F features per observation
* S sample, MCMC process
* C chains, MCMC process

= O(N^2KFSC)

**Non-centered parameterization**: This is a hierarchical model - each child's height depends on their family's typical height, which itself varies across families. This is called a "centered" parameterization because the child's height is centered around their family's height.

Now, imagine trying to figure out both the family's typical height AND the spread of heights within the family at the same time. It's tricky because these parameters are tangled together - if we see a tall child, is it because:

1. Their family tends to be tall? (high family\_height)
2. Or because they're particularly tall for their family? (large within\_family\_spread)

This "tangling" makes it hard for our statistical algorithms to find good estimates.

Here's where non-centered parameterization comes in. The key insight is that child\_offset is now independent of both family\_height and within\_family\_spread.

Resources for CAR

* <https://www2.stat.duke.edu/~cr173/Sta444_Sp17/slides/Lec18.pdf>
* <https://web.stanford.edu/class/stats253/lectures/lect7.pdf>
* <https://www.youtube.com/watch?v=4zsh_o-qli4>
* <https://stats.stackexchange.com/questions/277/spatial-statistics-models-car-vs-sar>
* <https://arxiv.org/pdf/1710.07000>
* <https://mc-stan.org/users/documentation/case-studies/mbjoseph-CARStan.html>
* <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>

Resources for non centered parametrisation

* <https://arxiv.org/pdf/1312.0906>
* <https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/>
* <https://www.econ.upf.edu/~omiros/papers/val7.pdf>
* <https://www.youtube.com/watch?v=DPnLb5EaCkA>

Metrics in the result:

* Variance of effect for each house characteristic.

Da fare:

* Lettura e capire se alcune parti richiedono più spiegazione
* Run di prova del codice
* Introdurre una metrica tipo R^2

NAIVE ESTIMATOR  
  
 *We split our real estate dataset by council area and run separate linear regressions. Formally, for area* ***i****:*

*Price(i,j) = b0(i)*

*+ b1(i) \* Distance(i,j)*

*+ b2(i) \* BuildingArea(i,j)*

*+ b3(i) \* Rooms(i,j)*

*+ b4(i) \* Car(i,j)*

*+ b5(i) \* YearBuilt(i,j)*

*+ e(i,j).*

*This approach allows each area to have its* ***own*** *intercept and slopes, letting us capture differences in how property characteristics affect price across different locations.*