Real Estate Prices, Past and Future

Bob Ding

2020-10-18

Introduction

Real estate, as a major part of the economy, has been constanly and closely monitored by investors and researchers. Since 1970 [1], when most countries' statistical offices or central banks began to collect data on house prices, interest in how to predict and forecast house prices have gradually augmented, bringing out more and more sophisticated modling techniques. Due to the increased ability of modern society to collect and store more data, attempts to make prediction on real estate prices have therefore shifted to data-driven, which further improved modeling precision.

Being such a sophisticated product, real estate prices are also impacted by many factors. While most of the factors helpful in predicting the house are observable and descriptive to the house itself, such as house' size, number of bathrooms, and whether it possesses a swimming pool etc., there are also inobservable factors that also impact house prices, such as the underlying real state market economy, cyclicality of real estate prices, and so on.

Many previous researches have already proposed multiple ways of predicting house prices. From the most simple regression methods as proposed in [2], to those that account for repeated sells of houses [3], and to those that take temporal effects into consideration, such as [4] and [6]. Though these studies are drastically different and are definitly other researches proposing more sophisticated models , each study has a different but clear focus. Thus, it is important to make certain of the research question before creating model.

Therefore, we proposes our goal of this study. The only type of house that we'll be researching into is house in Durham, NC, due to our better familiarity of the terrain. The goals include 1) understand how descriptive and observable variables affect housing prices, 2) understand how temporal effect affect housing price, 3) extract past real estate market, and 3) make short term forecast of housing prices.

1. Exploratory Data Analysis

1.1 Data Discription

The Dataset is scraped from redfin official set [5]. Redfin is a real estate brokerage that was founded in 2004. It's website consist of historical purchase record of the past 3 years. We therefore scrapped these 3 years of data, ranging from 2017 April to 2020 May. This dataset contains 6962 observation. Thanks to redfin's meticulous data record, no missing value in any field was presented. Each observation is a recorded deal of house purchase. Therefore, the price is the deal price between customer and seller, which is objective enough for us to fit on.

The data set contains many covariates. Among which, there are some non-process-able string information, such as name of the community, or geographical information which is beyond the scope of our interests. Therefore, to simplify our research, we introduce the following covariates of our interest.

Name	Description	Missing
`Price`	the deal price of the house	0%
`beds`	number of beds in the house	0%
`sold.dat`	the date on which the deal is settled	0%
`baths`	the number of bathrooms the house has	0%
`square.feet`	usable area (ft^2) measured in square feet	0%
`lot.size`	total area (ft^2) of the lot	0%
`house.age`	age (years) of the house when purchased	0%
`property.type`	Townhouse, or Single family residential	0%
`latitude`	latitude of the house	0%
`longitude`	longitude of the house	0%

1.2 Exploring Data

The first impression is that the distribution of house types in category Condo, Townhouse, and Single Family Residential is very uneven. This is shown in Appendix figure 1. Besides, A complete pairwise-plot has also been attached in the Appendix in figure 2. The rest of the following section explain the major 3 concerns about data distribution, including suggesions on how to address them. Besides, an additional subsection is also added to explain engineering of additional predictors and address of interactions.

1.2.1 Multi-collinearity

Though increased number of beds in the house need not imply the increase of square feet, increasing number of baths in the house does imply the increase square feet more directly. Notice that in figure 1.2.1, a strong collinearity is shown between the number of baths and square feet of the house, achieving an correlation of 0.7843. Thus, we should be careful in the final model output for these highly correlated covariates. More pairwise distribution between variables can be found in pair plot shown at appendix at figure 2

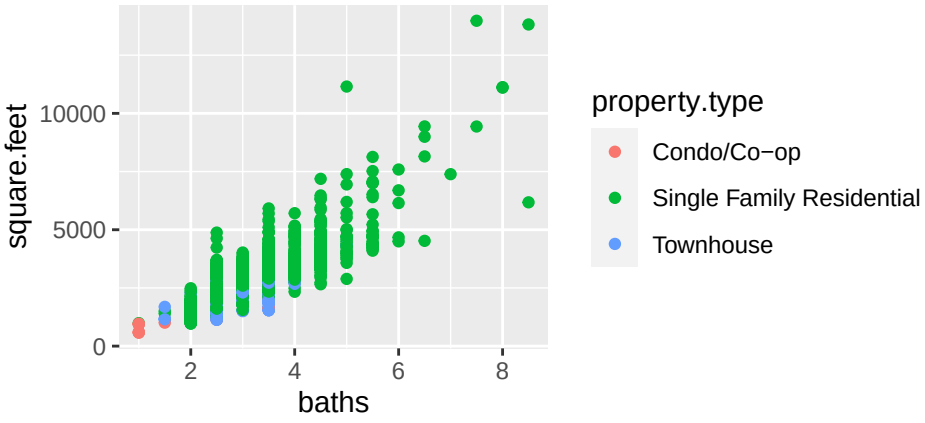


figure 1.2.1

1.2.2 Heteroscedasticity

While some strong linearity and positive correlation are evident among some predictor variables, such as `beds`, the number of beds, and `square.feet`, the usable area of the house, accompanied with the increase in these predictor values is the increase of variance. This violates the linear regression Monoscedasticity assumption, as shown in figure 1.2.2 below. To address this, we perform log-transformation on response variable and createa a final regression response variable `log.price`, which is log of house deal price. Shown in the last line of the pair-plot, Heteroscedasticity problem is resolved without harming positive correlation

between the response and predictors. Furthermore, after log transformation, the linear relationship between response and predictors become more evident.

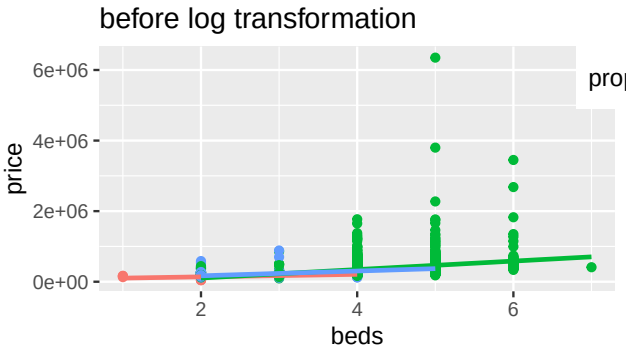


figure 1.2.2 (b)

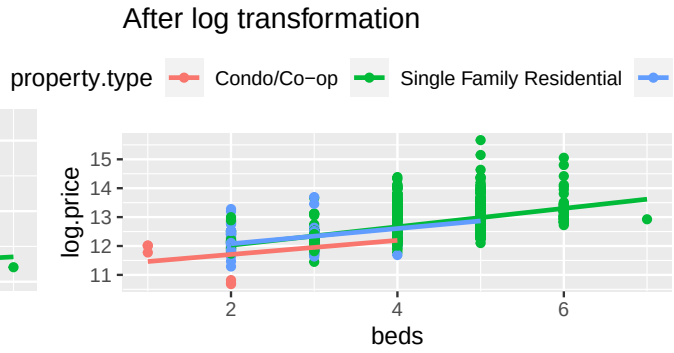
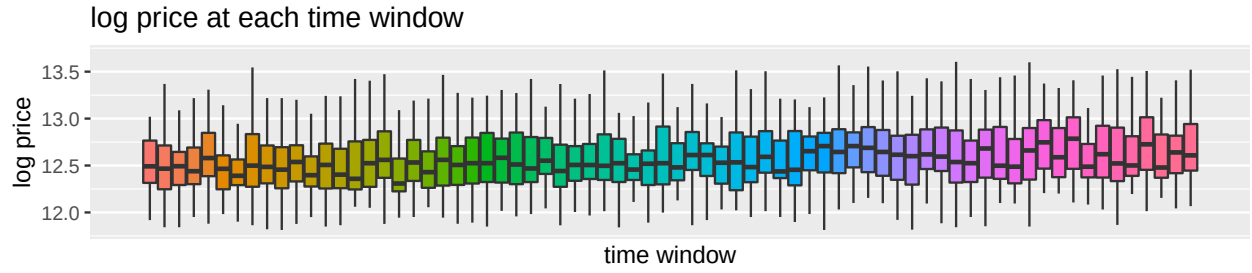


figure 1.2.2 (b)

1.2.3 Stationarity

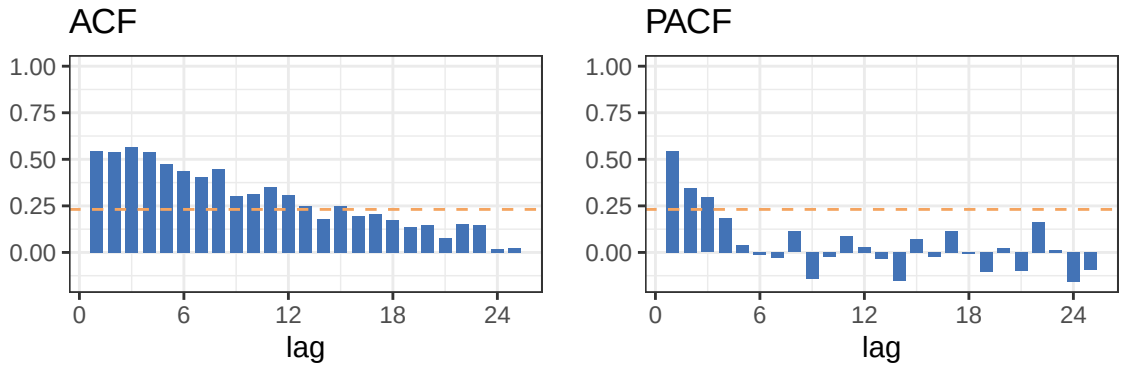
When determining the model to capture temporal effect, it is imperative to verify whether the stationarity assumption has been satisfied. In our case, due to the fact it is unlikely that there are houses sold on each day, we need to create a time series by windowed data – that is, we treat all the real estate deals that happened in a window as deals happening at the same timestep. By specifying the width of our window, we can modulate and balance between flexibility in temporal effect and validity of regression coefficients. At this point, we choose window width to be 15 days.



1.3.2 (a)

From figure 1.3.2 (a), we learn that there is a slight upward trend in the distribution of logarithm of house prices through time. Though this figure can only show distribution of log price that's marginalized out by all other exogenous variables, but it is helpful for us identify a linearly increasing time trend. In order to deal with the non-stationarity trend effect, we propose 2 solutions (might add the third later if I have time): 1) ignore, 2) adding `sold.dat`, the sold date, into the linear observation model as a predictor in order to capturing the linear trend using the linear model. We'll fit these 2 models and determine the final model by checking the moduli of eigenvalue of the AR(p) transition function.

Besides, it would also be helpful to identify the number of lags p by 1) checking ACF/PACF plots and 2) making hypothesis. Below in figure 1.3.2 (b) shows the ACF/PACF mentioned. We observe that lag 1,2,3 are statistically significantly correlated. Therefore, we should at least propose have $p \geq 3$. However, as we also hold hypothesis that house prices might have cyclical effect of very long period , we further extend $p \geq 12$ given the fact that a window is 15 days. Therefore, we will start by incorporating 12 auto-correlation terms in our model. Due to the excess amount of lags, we hope to design Bayesian ridge shrinkage prior for AR(p) coefficients.



1.2.4 Engineered Feature and Interaction

We hold hypothesis that excessive number of bed in a house with comparatively low number of bathroom will impact the house price. Therefore, we created the variable `room.Diff`, which means how much more baths does the house have than beds. We found such feature creates distinct effects across Single Family Residential, Townhouse, and Condo in affecting log of price. In figure 1.2.4 (a), we can observe a different slop of the variable for 3 type of houses. Therefore, together with `room.Diff`, interactions between the the bed-bath difference and house type should also be added to our model.

Lastly, we're believe incorporating a simple covariable can increase prediction power. Therefore, we engineered the new variable `dist.duke`, indicating the distance of the house to Duke. This is approximately represented by shortest distance of 2 points on a sphere, calculated via ellipsoid determined by latitude and longitude of the house and Duke University. Then, as in figure 1.2.4 (b) shows, aligned with our conjecture, we anticipate longer distance to Duke leading to a lower housing price.

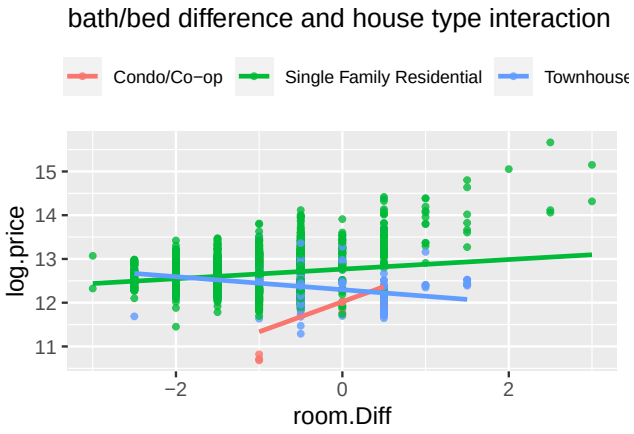


figure 1.2.4 (a)

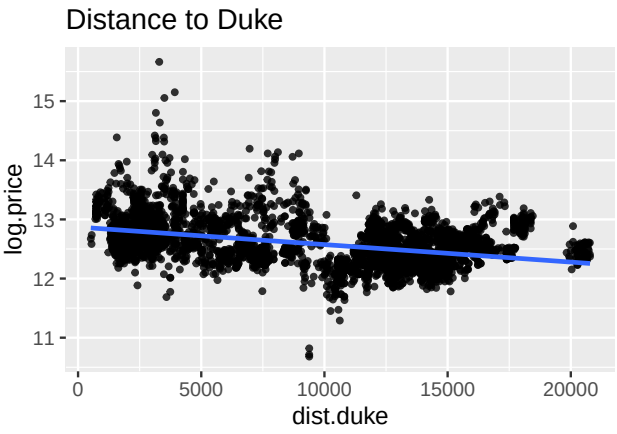


figure 1.2.4 (b)

2 Model Formulation

We define our model as a regression model on top of a AR(P) model. After EDA, we determines to use response as `log.price`: the logarithm of the house deal price. The covariate predictors are `beds`, `baths`, `square.feet`, `lot.size`, `house.age`, `property.type`, `room.Diff`, `dist.Duke`,`sold.dat`, and finally `room.Diff` interact with `property.type`. Besides, we define the intercept α_t as time-varying. The transition of α_t follows a AR(p) process.Therefore, the full model is as the following. (**Note:** This model incorporated `sold.dat` as a predictor. This correspond to the linear trend model mentioned in EDA about stationarity. To construct the ignore trend model, simply remove `sold.dat` together with its β coefficient)

$$y_t^{(i)} = \beta_{\text{beds}}\ \mathbf{beds}_t^{(i)} + \beta_{\text{sold date}}\ \mathbf{sold\ date}_t^{(i)} + \beta_{\text{beds}}\ \mathbf{baths}_t^{(i)} + \beta_{\text{square feet}}\ \mathbf{square\ feet}_t^{(i)} + \tag{1}$$

$$\beta_{\text{lot size}}\ \mathbf{lot\ size}_t^{(i)} + \beta_{\text{house age}}\ \mathbf{house\ age}_t^{(i)} + \beta_{\text{property type}}\ \mathbf{property\ type}_t^{(i)} + \tag{2}$$

$$\beta_{\text{room difference}}\ \mathbf{room\ Diff}_t^{(i)} + \beta_{\text{dist. to Duke}}\ \mathbf{dist.\ to\ Duke}_t^{(i)} + \tag{3}$$

$$\beta_{\text{interaction}}\ \mathbf{room\ Diff}_t^{(i)} \times \mathbf{property\ type}_t^{(i)} + \alpha_t + \nu_t \tag{4}$$

$$\alpha_t = \sum_{i=1}^{p} \theta_i \alpha_{t-i} + \omega_t \tag{5}$$

$$\omega_t \sim \mathcal{N}(0, w) \tag{6}$$

$$\nu_t \sim \mathcal{N}(0, v) \tag{7}$$

Where $y_t^{(i)}$ is the response of the i^{th} house sold on the t^{th} window date, which is the logarithm of deal price. (notice that suppose for each window $t \in \{1, 2, 3, \cdots, T\}$, there are n_t sold houses in the t^{th} window. Then n_t need not be equal for all t). The rests are predictive variables. α_t is an time varying intercept which will be modeled by the AR(p) model described in $(5), (6)$. ν_t is an additional observation uncertainty and ω_t is an additional evolution uncertainty. To simply our modeling process, we take ν_t, ω to have constant variance at all time.

Through reparametrization, we can simplify the model as the complete model following. A detailed explanation of why we should form the model in this compact form and how such formulation is achieved can be found in appendix B.

$$\begin{aligned}
\boldsymbol{\alpha}_t &= \boldsymbol{\Theta}\boldsymbol{\alpha}_{t-1} + \boldsymbol{W}_t \\
\boldsymbol{y}_t &= \mathbf{1}\boldsymbol{\alpha}_t + \boldsymbol{\beta}\boldsymbol{X}_t + \boldsymbol{\nu}_t \\
\boldsymbol{W}_t &\sim \mathcal{N}(\mathbf{0}, w\boldsymbol{I}) \\
\boldsymbol{\nu}_t &\sim \mathcal{N}(\mathbf{0}, v\boldsymbol{I})
\end{aligned}$$

$$\mathbf{1} := \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \boldsymbol{\alpha}_t := \begin{bmatrix} \alpha_t \\ \alpha_{t-1} \\ \alpha_{t-2} \\ \vdots \\ \alpha_{t-p} \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \dots & \theta_p \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \alpha_{t-2} \\ \alpha_{t-3} \\ \vdots \\ \alpha_{t-p} \end{bmatrix} = \boldsymbol{\Theta}\boldsymbol{\alpha}_{t-1}$$

3 Methodology

In this section, we explore methodology of using MCMC sampling with Forward Backward algorithm to make inference on parameters. Besides, we'll also have a discussion about how the methodology and the inferred parameter posterior distribution could be used to answer our pre-stated 4 research questions in introduction.

3.1 Parameter Inference

The Model requires statistical inference uppon the following parameters: $\{w, v, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\alpha}_{1:T}\}$. To obtain these, we choose to apply MCMC sampling with forward backward algorithm within the MCMC sampler. Below is the algorithm that returns a MCMC samples for these parameters. A highly detailed derivation of sampling distribution of each step can be found in appendix A. Notice that to apply MCMC, we must design a prior distribution for all the parameters. Due to the flexibility of prior design and our previous sections stated multicolinear problem, we've decided to propose Bayesian Ridge regression priors for both the AR(p)

coefficients $\boldsymbol{\theta}$ and also the linear regression model $\boldsymbol{\beta}$ coefficients. Details can also be found in Appendix A.

Algorithm 1: parameter inference algorithm

Result: A short algorithm of initializing reservoir weights with insured Echo State Property. Through empirical experiments, we recommend setting $\eta_1 = 0.97, \eta_2 = 0.85, \boldsymbol{\mu} = -2, \epsilon = 1$.

Initialize: $\mathbb{P}((\theta_1, \ldots, \theta_p)^\top), \mathbb{P}(w), \mathbb{P}(v)$ via pre-set prior distribution

while *not converged* **do**

- **Calculate**: posterior mean and covariance for $\boldsymbol{\alpha}_t|\mathcal{D}_t$: $m_t, C_t \ \forall t \in 1:T$ via forward filtering algorithm in Appendix
- **Sample**: $\boldsymbol{\alpha}_t|\mathcal{D}_T$ from $m_t^*, C_t^*, \forall t \in 1:T$ by backward smoothing
- **Sample**: $\boldsymbol{\theta}, \phi = w^{-1}|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}$ by first sampling $\phi = w^{-1}|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}$ and then $\boldsymbol{\theta}|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}, \phi$
- **Sample**: $\boldsymbol{\beta}|\boldsymbol{X}, \kappa, \boldsymbol{y}_t, \tau = v^{-1}$ and then sample $\tau = v^{-1}|\boldsymbol{X}, \kappa, \boldsymbol{y}_t, \boldsymbol{\beta}$

end

Return: samples of $w, v, \boldsymbol{\beta}, \boldsymbol{\theta}_{1:T}$

3.2 Answering Research Question

We've proposed 4 research questions at the end of the introduction section. We address each question respectively.

To answer question 1) – how non-temporal descriptive variables affect housing prices, as we've conducted MCMC sampling, we obtain a distribution of these parameters. Interpretation on such variable impact the house price is identical to the method of interpretation in a basic Bayesian linear regression. We'll provide point and interval estimate for all the covariates.

To address question 2) – how temporal effect affect house prices, we utilize eigen-decomposition of $\boldsymbol{\Theta}$ matrix to shed light on cyclicality of house market price. Notice that the MCMC sampler above returns a series of $\boldsymbol{\theta}$ samples. Thus, we only obtain the posterior distribution of $\boldsymbol{\theta}$ rather than having a point estimate for it. Thus, a point estimate for $\boldsymbol{\Theta}$ is not available. However, it is possible to make a Bayes estimator for $\boldsymbol{\Theta}$. As we're aiming to find an estimator $\hat{\boldsymbol{\theta}}$ that minimizes the posterior expected mean square error, we simply take the mean of all $\boldsymbol{\theta}$ samples to obtain this estimator $\hat{\boldsymbol{\theta}}$. Thus, a reasonable point estimate for $\boldsymbol{\Theta}$ can be constructed by transformation, and we can perform eigen decomposition on this estimator.

Once we extract eigenvalues together with its eigenvectors, we observe the complex eigenvalues, which represent the cyclicality. For example, suppose we observe a pair of eigenvalues $a \pm bi$, this correspond to the α_t, the real estate market, having a sine wave cycle of $Period = \frac{2\pi}{\arcsin(b/\sqrt{a^2+b^2})}$ measured in timestep. To re-measure the cycle in days, simply time the window size, which is 15, with $Period$. Therefore, we obtain the cycle effect. As there could be multiple cycle effects, their contribution are measured in magnitude, which can be calculated by modulus $\sqrt{a^2 + b^2}$

To address question 3) – how to extract the past real estate market α_t. In our MCMC sampler, we have also backward sampled the α_t trajectory. Thus, this task can be addressed by taking mean mean of all the α_t samples from MCMC.

To address question 4) – how to forecast short term house prices, we sample α_{T+1} in the MCMC for each MCMC iteration. Then and apply this sampled α_{T+1} to make prediction on log of prices at $\boldsymbol{y}_{T+1}$, convert the log price to real hosue prices and calculate MSE with ground truth.

4 Prelimiary Results

Currently, we've generated results using 2 models. 1) the ignore trend model and 2) the linear trend model. The results are not final but are reasonable. I'm attaching all the results in appendix in section C.

5 Model Validation and Sensitivity Analysis

Model Validation will be 2 separate parts: 1) validation on linear regression model, and 2) MCMC validation. We've finished all the model validation and relevant figures will be attached in Appendix D.1 and D.2. Formal write up will be left to the next submission.

Sensitivity Analysis will be focused on validating a series of Bayesian ridge regression shrinkage coefficient κ. Besides, variation of different time window lengths will also be tested.

Bibliography

[1] Schularick, M., & Steger, T. (2014). No Price Like Home: Global House Prices, 1870–2012. Federal Reserve Bank of Dallas, Globalization and Monetary Policy Institute Working Papers, 2014(208). doi: 10.24149/gwp208

[2], Manisaurabh. (2020, September 14). House-Prices-Advanced-Regression-Technique. Retrieved October 18, 2020, from https://www.kaggle.com/manisaurabh/house-prices-advanced-regression-technique

[3], Ghysels, E., Plazzi, A., Valkanov, R., & Torous, W. (2013). Forecasting Real Estate Prices. Handbook of Economic Forecasting, 509-580. doi:10.1016/b978-0-444-53683-9.00009-8

[4]. Aguilar, F. (2019, July 15). Time Series Analysis on US Housing Data. Retrieved October 18, 2020, from https://medium.com/@feraguilari/time-series-analysis-modfinalproyect-b9fb23c28309

[5]. Redfin. (2020). Real Estate, Homes for Sale, MLS Listings, Agents | Redfin. https://www.redfin.com/

[6]. Bhagat, N., Mohokar, A., & Mane, S. (2016). House Price Forecasting using Data Mining. International Journal of Computer Applications, 152(2), 23-26. doi:10.5120/ijca2016911775

Appendix

A Parameter Inference

A.1 Forward Filtering

$$
\begin{aligned}
\boldsymbol{\alpha}_t &= \boldsymbol{\Theta}\boldsymbol{\alpha}_{t-1} + \boldsymbol{W}_t \\
\boldsymbol{y}_t &= \mathbf{1}\boldsymbol{\alpha}_t + \beta \boldsymbol{X}_t + \boldsymbol{\nu}_t \\
\boldsymbol{W}_t &\sim \mathcal{N}(\mathbf{0}, w\boldsymbol{I}) \\
\boldsymbol{\nu}_t &\sim \mathcal{N}(\mathbf{0}, v\boldsymbol{I}) \\
\mathbf{1} &:= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

First, denote that

$$
\begin{aligned}
\boldsymbol{\alpha}_t|\mathcal{D}_{t-1}, - &\sim \mathcal{N}(\boldsymbol{\Theta}m_{t-1}, \boldsymbol{\Theta}C_{t-1}\boldsymbol{\Theta}^T + w\boldsymbol{I}) = \mathcal{N}(a_t, R_t) \\
\boldsymbol{y}_t|\boldsymbol{\alpha}_t, \mathcal{D}_{t-1}, - &\sim \mathcal{N}(\mathbf{1}\boldsymbol{\alpha}_t + \boldsymbol{X}_t\beta, v\boldsymbol{I}) \\
\mathbb{P}(\boldsymbol{\alpha}_t|\mathcal{D}_t) &\propto \mathbb{P}(\boldsymbol{\alpha}_t|\mathcal{D}_{t-1}, -)\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{\alpha}_t, \mathcal{D}_{t-1}, -) \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\alpha}_t^T(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1})\boldsymbol{\alpha} - 2\boldsymbol{\alpha}_t^T(R_t^{-1}a_t + v^{-1}\mathbf{1}^T(\boldsymbol{y}_t - \boldsymbol{X}_t\beta))\right]\right\} \\
&\sim \mathcal{N}\left(\left(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1}\right)^{-1}(R_t^{-1}a_t + v^{-1}\mathbf{1}^T(\boldsymbol{y}_t - \boldsymbol{X}_t\beta)), \left(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1}\right)^{-1}\right) \\
&= \mathcal{N}(m_t, C_t)
\end{aligned}
$$

$$
\begin{aligned}
a_t &= \boldsymbol{\Theta}m_{t-1} \\
R_t &= \boldsymbol{\Theta}C_{t-1}\boldsymbol{\Theta}^T + w\boldsymbol{I} \\
m_t &= \left(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1}\right)^{-1}(R_t^{-1}a_t + v^{-1}\mathbf{1}^T(\boldsymbol{y}_t - \boldsymbol{X}_t\beta)) \\
C_t &= \left(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1}\right)^{-1}
\end{aligned}
$$

Use this equation to update

A.2 Backward Smoothing

Sh*t, I hate this...

Suppose we already know that

$$
\mathbb{P}(\boldsymbol{\alpha}_{t+1}|\mathcal{D}_T) \sim \mathcal{N}(m_{t+1}^*, R_{t+1}^*)
$$

Let's look at log likelihood of $\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t+1}|\mathcal{D}_T$. Using conditional independence, we have

$$
\begin{aligned}
-\frac{1}{2}\ell(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t+1}; \mathcal{D}_T) &= \log \mathbb{P}\left(\boldsymbol{\alpha}_{t+1} \mid \boldsymbol{\alpha}_t\right) + \log \mathbb{P}\left(\boldsymbol{\alpha}_t \mid \mathcal{D}_t\right) - \log \mathbb{P}\left(\boldsymbol{\alpha}_{t+1} \mid \mathcal{D}_t\right) + \log \mathbb{P}\left(\boldsymbol{\alpha}_{t+1} \mid \mathcal{D}_T\right) \\
&= (w)^{-1}(\boldsymbol{\alpha}_{t+1} - \boldsymbol{\Theta}\boldsymbol{\alpha}_t)^T(\boldsymbol{\alpha}_{t+1} - \boldsymbol{\Theta}\boldsymbol{\alpha}_t) + (\boldsymbol{\alpha}_t - m_t)^T(C_t)^{-1}(\boldsymbol{\alpha}_t - m_t) - \\
&\quad (\boldsymbol{\alpha}_{t+1} - a_{t+1})^T(R_{t+1})^{-1}(\boldsymbol{\alpha}_{t+1} - a_{t+1}) + \\
&\quad (\boldsymbol{\alpha}_{t+1} - m^*_{t+1})^T(C^*_{t+1})^{-1}(\boldsymbol{\alpha}_{t+1} - m^*_{t+1}) + \text{constant} \\
&= \boldsymbol{\alpha}^T_{t+1}{C^*_{t+1}}^{-1} + w^{-1}\boldsymbol{I} + R^{-1}_{t+1})\boldsymbol{\alpha}_{t+1} + \boldsymbol{\alpha}^T_t(w^{-1}\boldsymbol{\Theta}^T\boldsymbol{\Theta} + C^{-1}_t)\boldsymbol{\alpha}_t + \\
&\quad 2\boldsymbol{\alpha}^T_{t+1}(-w^{-1}\boldsymbol{\Theta})\boldsymbol{\alpha}_t - 2\boldsymbol{\alpha}^T_t(C^{-1}_t m_t) - 2\boldsymbol{\alpha}^T_{t+1}(R^{-1}_{t+1}a_{t+1} + {C^*_{t+1}}^{-1}m^*_{t+1}) + \text{constant}
\end{aligned}
$$

One eternity of calculation later, we end up with:

$$
\begin{aligned}
J_t &= C_t\boldsymbol{\Theta}^T(\boldsymbol{\Theta}C_t\boldsymbol{\Theta}^T + w\boldsymbol{I})^{-1} \\
m^*_t &= m_t + J_t(m^*_{t+1} - \boldsymbol{\Theta}m_t) \\
C^*_t &= C_t + J_t(C^*_{t+1} - \boldsymbol{\Theta}C_t\boldsymbol{\Theta}^T - w\boldsymbol{I})J^T_t
\end{aligned}
$$

And therefore

$$
\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t+1}, \mathcal{D}_T \sim \mathcal{N}\Big(m_t + J_t(\boldsymbol{\alpha}_{t+1} - \boldsymbol{\Theta}m_t), C_t - J_tR_{t+1}J^\intercal_t\Big)
$$

A.3 Dynamic model sampling: $(\theta_1, \cdots, \theta_p)^\intercal$, $w = \phi^{-1}$

This is a simple linear regression $\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\theta} + w_t$ with design matrices as

$$
\boldsymbol{y} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \vdots \\ \alpha_T \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \alpha_{1-1} & \cdots & \alpha_{1-p} \\ \alpha_{2-1} & \cdots & \alpha_{2-p} \\ \alpha_{3-1} & \cdots & \alpha_{3-p} \\ \alpha_{4-1} & \cdots & \alpha_{4-p} \\ \alpha_{5-1} & \cdots & \alpha_{5-p} \\ \vdots & \vdots & \vdots \\ \alpha_{T-1} & \cdots & \alpha_{T-p} \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_p \end{bmatrix}
$$

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{y}; \boldsymbol{\theta}, \mathbf{X}) &\propto \phi^{\frac{T}{2}} \exp\{-\frac{1}{2}\phi(\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta})^\intercal(\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta})\} \\
\boldsymbol{\theta}|\phi, \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}/\phi\right) = \mathcal{N}((0.5, 0.5, 0.5)^T, 1/3\phi^{-1}\mathbf{I}) \\
\phi|\mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathbf{G}\left(a_0 = \frac{v_0}{2}, b_0 = \frac{v_0 s_0^2}{2}\right) = \mathbf{G}\left(\frac{1}{2}, \frac{1}{2}\right) \\
\boldsymbol{\mu}_n &= \left(\mathbf{X}^\mathrm{T}\mathbf{X} + \boldsymbol{\Lambda}_0\right)^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \mathbf{X}^\mathrm{T}\boldsymbol{y}\right) \\
\boldsymbol{\Lambda}_n &= \left(\mathbf{X}^\mathrm{T}\mathbf{X} + \boldsymbol{\Lambda}_0\right) \\
a_n &= a_0 + \frac{T}{2} \\
b_n &= b_0 + \frac{1}{2}(\boldsymbol{y}^\intercal\boldsymbol{y} + \boldsymbol{\mu}_0^\intercal\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^\intercal\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n) \\
\boldsymbol{\theta}|\phi, \boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathcal{N}\left(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n^{-1}/\phi\right) \\
\phi|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathbf{G}\left(a_n, b_n\right)
\end{aligned}
$$

A.4 Observation Model Sampling $(\beta_1 \cdots)^\intercal$, $v = \tau^{-1}$

Very similar as above, this is also a linear model. Besides, it is possible to apply Bayesian Ridge here. let's create the Bayesian ridge model

$$
\begin{aligned}
\boldsymbol{y}_t &= \boldsymbol{\alpha}_t \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \nu_t \\
\boldsymbol{z}_t = (\boldsymbol{y}_t - \boldsymbol{\alpha}_t \mathbf{1}) \mid \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \tau &\sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{l}_n/\tau\right) \\
\boldsymbol{\beta} \mid \tau, \kappa &\sim \mathrm{N}\left(\mathbf{0}, \mathbf{l}(\tau\kappa)^{-1}\right) \\
p(\tau \mid \kappa) &\propto 1/\tau
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{y}_t - \boldsymbol{\alpha}_t \mathbf{1} | \boldsymbol{X}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}) &\propto \tau^{\frac{n}{2}} \exp\{-\frac{\tau}{2}(\boldsymbol{z}_t - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{z}_t - \boldsymbol{X}\boldsymbol{\beta})\} \\
\mathbb{P}(\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{z}_t, \tau, \boldsymbol{\alpha}) &\propto \mathbb{P}(\boldsymbol{z}_t | \boldsymbol{X}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}) \mathbb{P}(\boldsymbol{\beta} | \tau, \kappa, \boldsymbol{\alpha}) \mathbb{P}(\tau | \kappa, \boldsymbol{\alpha}) \\
&\propto \tau^{\frac{n}{2}} \exp\{-\frac{\tau}{2}(\boldsymbol{z}_t - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{z}_t - \boldsymbol{X}\boldsymbol{\beta})\} (\tau\kappa)^{\frac{p}{2}} \exp\left\{-\frac{\tau\kappa}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\} \tau^{-1} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^T(\tau\boldsymbol{X}^T\boldsymbol{X} + \tau\kappa\mathbf{1})\boldsymbol{\beta} - 2\tau\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{z}_t\right]\right\} \\
&\sim N\left((\boldsymbol{X}^T\boldsymbol{X} + \kappa\mathbf{1}_p)^{-1}\boldsymbol{X}^T\boldsymbol{z}_t, \tau^{-1}(\boldsymbol{X}^T\boldsymbol{X} + \kappa\mathbf{1}_p)^{-1}\right) \\
\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{z}_t, \tau, \boldsymbol{\alpha} &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\mathbb{P}(\tau | \boldsymbol{\beta}, \kappa, \boldsymbol{z}_t, \boldsymbol{\alpha}) &\propto \tau^{\frac{n+p}{2}-1} \exp\left\{-\frac{\tau\kappa}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\} \\
\tau | \boldsymbol{\beta}, \kappa, \boldsymbol{z}_t, \boldsymbol{\alpha} &\sim \boldsymbol{G}(\frac{n+p}{2}, \frac{\kappa}{2}\boldsymbol{\beta}^T\boldsymbol{\beta})
\end{aligned}
$$

B Methodology

We define our model as a regression model on top of a AR(P) model.

B.1 Regression (Observation) Model

After EDA, we determines to use response as `log.price`: the logarithm of the house deal price. The covariate predictors are `beds`, `sold.dat`, `baths`, `square.feet`, `lot.size`, `house.age`, `property.type`, `room.Diff`, `dist.Duke`, and finally `room.Diff` interact with `property.type`. The regression model is

$$
\begin{aligned}
y_t^{(i)} = \beta_{\text{beds}}\ \mathbf{beds}_t^{(i)} + \beta_{\text{sold date}}\ \mathbf{sold\ date}_t^{(i)} + \beta_{\text{beds}}\ \mathbf{baths}_t^{(i)} + \beta_{\text{square feet}}\ \mathbf{square\ feet}_t^{(i)} + & \quad (8) \\
\beta_{\text{lot size}}\ \mathbf{lot\ size}_t^{(i)} + \beta_{\text{house age}}\ \mathbf{house\ age}_t^{(i)} + \beta_{\text{property type}}\ \mathbf{property\ type}_t^{(i)} + & \quad (9) \\
\beta_{\text{room difference}}\ \mathbf{room\ Diff}_t^{(i)} + \beta_{\text{dist. to Duke}}\ \mathbf{dist.\ to\ Duke}_t^{(i)} + & \quad (10) \\
\beta_{\text{interaction}}\ \mathbf{room\ Diff}_t^{(i)} \times \mathbf{property\ type}_t^{(i)} + & \quad (11) \\
\alpha_t + \nu_t & \quad (12) \\
\nu_t \sim \mathcal{N}(0, v) & \quad (13)
\end{aligned}
$$

Where $y_t^{(i)}$ is the response of the i^{th} house sold on the t^{th} window date, which is its logarithm of deal price. (notice that suppose for each window $t \in \{1, 2, 3, T\}$, there are n_t sold houses in the t^{th} window. Then n_t need not equal for all t). The rests are predictive variables. α_t is an time varying intercept which will be modeled by the AR(P) model described in the next session. ν_t is an additional observation uncertainty. To simply our modeling process, we take ν_t to have constant variance. Also, to simplify our notation, we write vectorized equation by merging line $(1), (2), (3), (4)$. The compact form is denoted as

$$
\begin{aligned}
\boldsymbol{y}_t &= \alpha_t \mathbf{1}_{n_t} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\nu}_t \\
\boldsymbol{\nu}_t &\sim \mathcal{N}(\mathbf{0}, v\boldsymbol{I}_{n_t})
\end{aligned}
$$

B.2 Time Series Model

We construct the AR(p) model to model the underlying intercept α_t as described above in the regression model. As we've already indicated in EDA section, we'll choose $p = 7$ for our

$$\begin{aligned}\alpha_t &= \sum_{i=1}^{p} \theta_i \alpha_{t-i} + \omega_t \\ \omega_t &\sim \mathcal{N}(0, w)\end{aligned}$$

However, the above parametrization requires us to take-in many timesteps value to predict α_t, we may simply it by vectorizing the expression into the following:

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \alpha_t \\ \alpha_{t-1} \\ \alpha_{t-2} \\ \vdots \\ \alpha_{t-p} \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \dots & \theta_p \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \alpha_{t-2} \\ \alpha_{t-3} \\ \vdots \\ \alpha_{t-p} \end{bmatrix} = \boldsymbol{\Theta} \boldsymbol{\alpha}_{t-1}$$

In this way, transition becomes easy, as dependency rely on only the past one timestep.

B.3 Combined Model

We end up with the model as the following

$$\begin{aligned}\boldsymbol{\alpha}_t &= \boldsymbol{\Theta} \boldsymbol{\alpha}_{t-1} + \boldsymbol{W}_t \\ \boldsymbol{y}_t &= \mathbf{1} \boldsymbol{\alpha}_t + \boldsymbol{\beta} \boldsymbol{X}_t + \boldsymbol{\nu}_t \\ \boldsymbol{W}_t &\sim \mathcal{N}(\mathbf{0}, w\boldsymbol{I}) \\ \boldsymbol{\nu}_t &\sim \mathcal{N}(\mathbf{0}, v\boldsymbol{I}) \\ \mathbf{1} &\coloneqq \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}\end{aligned}$$

C Current Results

C.1 Answer to question 1

Table 1: Estimate for Beta

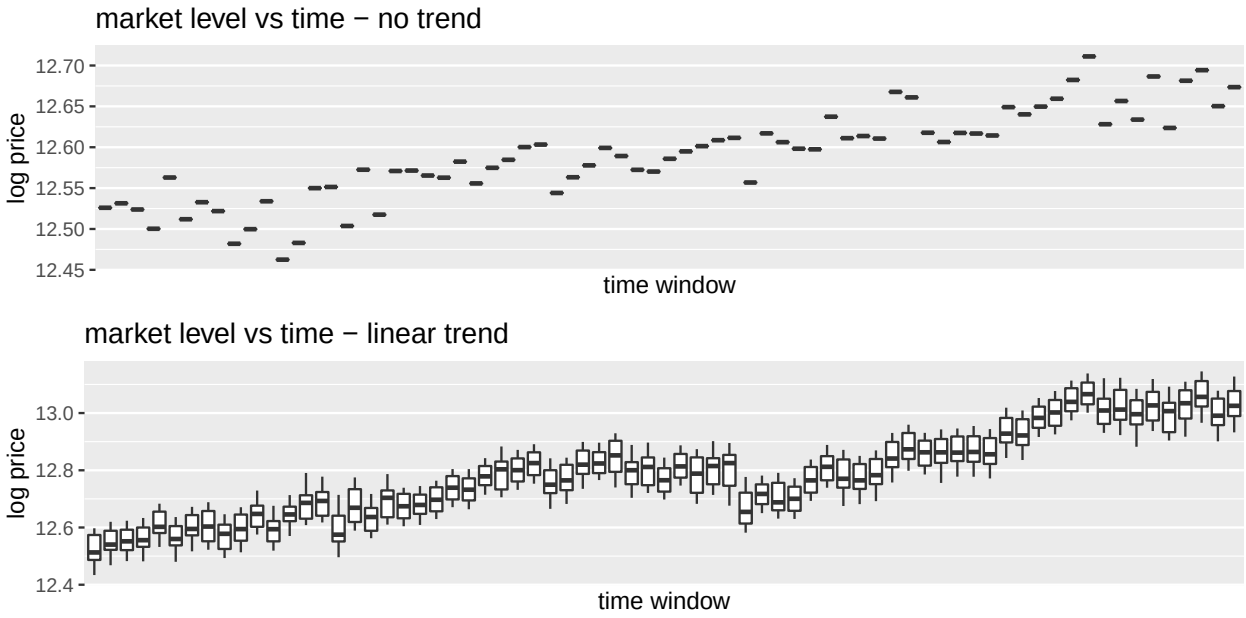
	No Trend Model			Linear Trend Model		
	mean	0.025%	0.975%	mean	0.025%	0.975%
beds	-0.1038672	-0.1107866	-0.0986698	-0.0779188	-0.1275974	-0.0179634
baths	0.1462753	0.1417107	0.1523752	-4.0867321	-4.0908893	-4.0828711
square.feet	0.2164692	0.2164179	0.2165016	0.1223538	0.0691499	0.1667891
lot.size	0.0291189	0.0290747	0.0291473	0.2157349	0.2148905	0.2167010
house.age	-0.0605298	-0.0605563	-0.0605099	0.0293895	0.0289310	0.0298071
single	0.0754878	0.0753325	0.0756679	-0.0594027	-0.0602110	-0.0589577
townhouse	0.0365382	0.0363756	0.0367080	0.0802400	0.0769373	0.0832969
room.Diff	0.2694715	0.2639533	0.2742393	0.0406870	0.0375790	0.0438255
dist.duke	-0.0775320	-0.0775541	-0.0775055	0.2787614	0.2348311	0.3259347
itr_townhouse	-0.1555703	-0.1559290	-0.1553112	-0.0782217	-0.0787153	-0.0778399
itr_single	-0.3969090	-0.3978791	-0.3961251	-0.1512971	-0.1578118	-0.1468182
sold.dat	NaN	NaN	NaN	-0.3803526	-0.3968583	-0.3683418

C.2 Answer to question 2

Table 2: Periodicity with Moduli (Magnitude)

No Trend Model			Linear Trend Model		
modulus	periods	period_in_day	modulus	periods	period_in_day
1.1916376	4.345577	65.18365	2.3559888	4.850858	72.76286
1.0519681	12.700771	190.51156	1.1476133	14.331572	214.97358
1.0024406	6.037424	90.56136	1.1112343	11.308972	169.63458
0.9041352	9.564635	143.46952	0.8691187	5.704190	85.56286
0.6746708	8.233147	123.49721	0.8281734	5.826064	87.39096

C.3 Answer to question 3



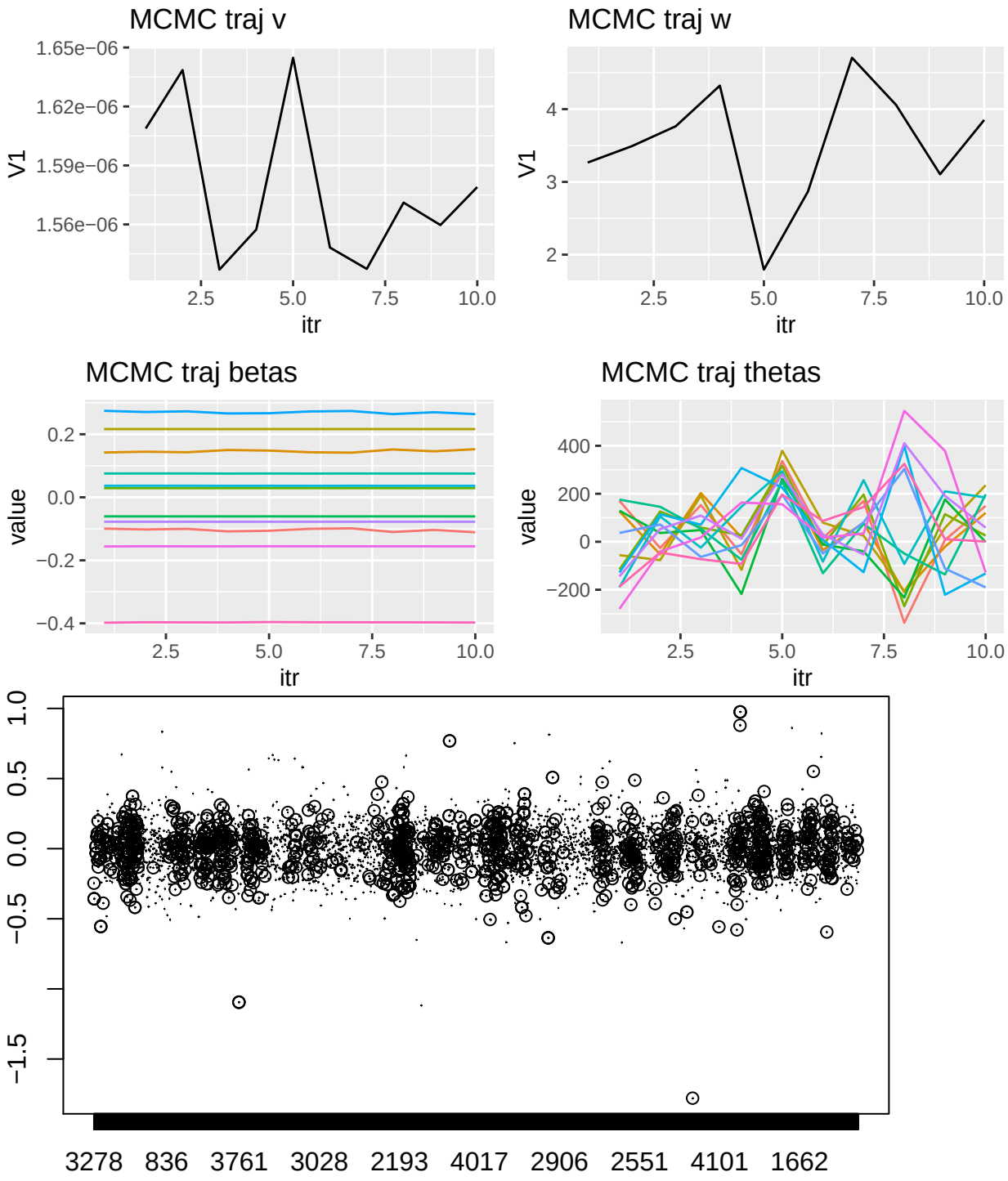
C.4 Answer to question 4

Table 3: 1 step prediction MSE for 2 models

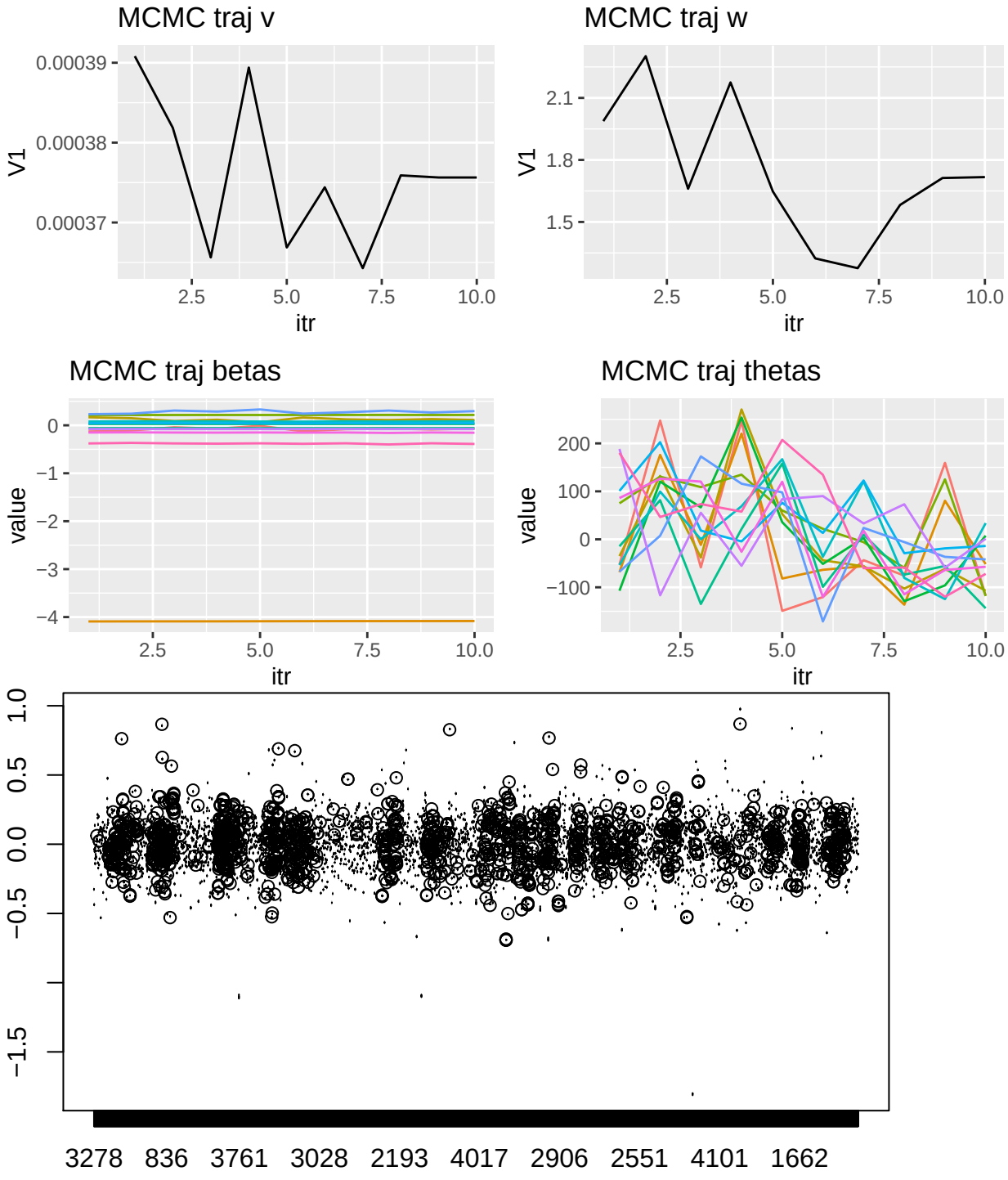
no trend	linear trend
317526.5	2433481

D Model Validation Figures

D.1 No Trend Model



D.2 Linear Trend Model



E Random Figures

{Last Revised: October 18, 2020}

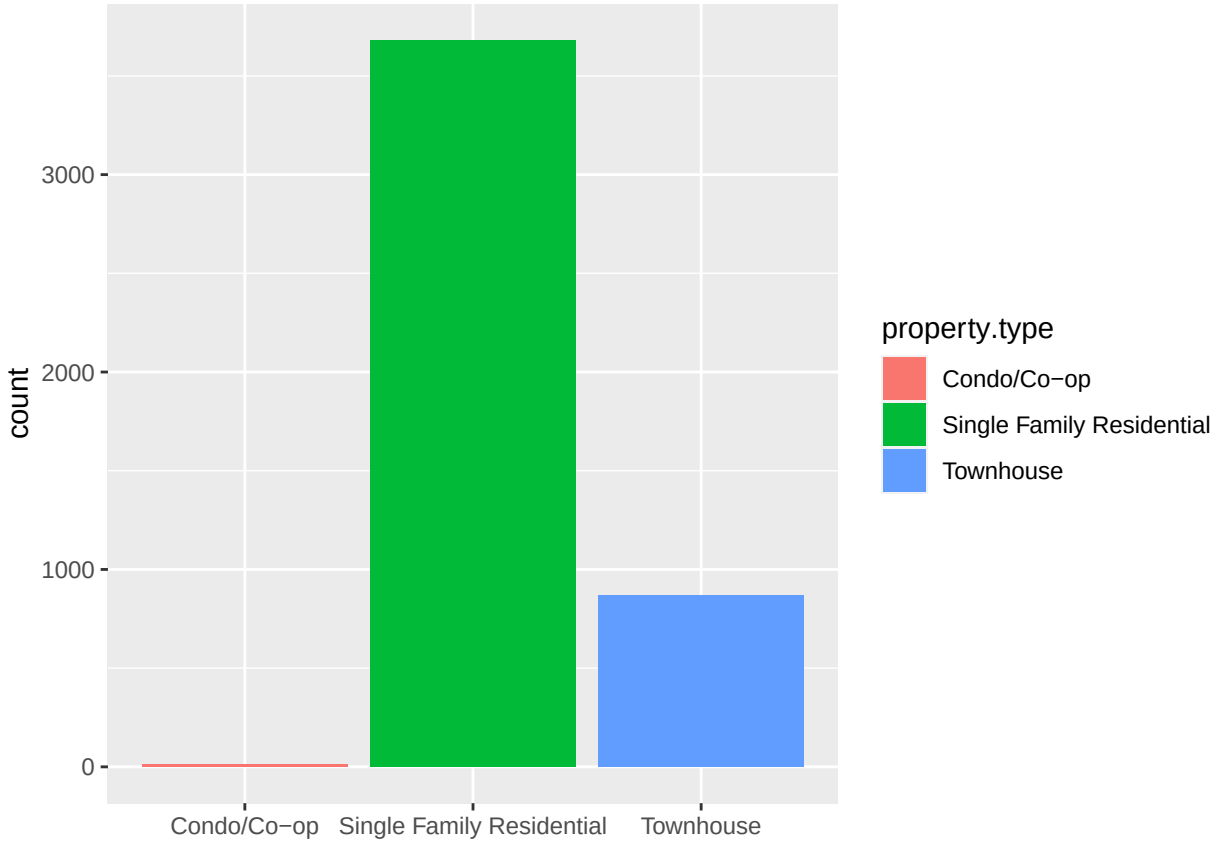


Figure 1: uneven distribution of classes

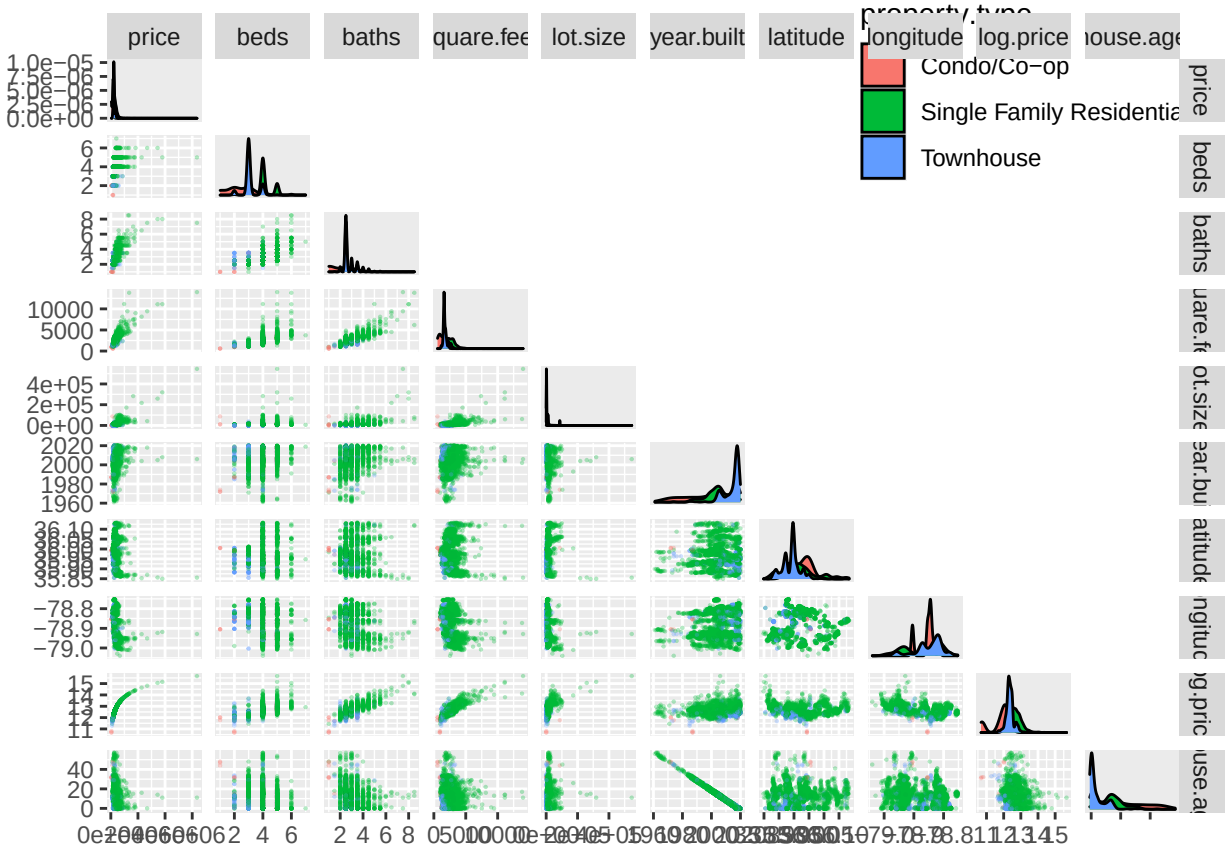


Figure 2: pair plot