440Independent

Ziyang Ding

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]. 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 cyclical, trend, and mean-shift effect of past real estate, and 3) make short term forecast of housing prices.

EDA

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

As we've already indicated above, we're interested only in houses and apartments in Durham. Therefore, after we filter out the data, the `city` variable no longer exist.

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%
`year.built`	the date the house was built.	0%
`property.type`	Townhouse, or Single family residential	0%
`city`	Durham, Chapel Hill, or Morresville	0%

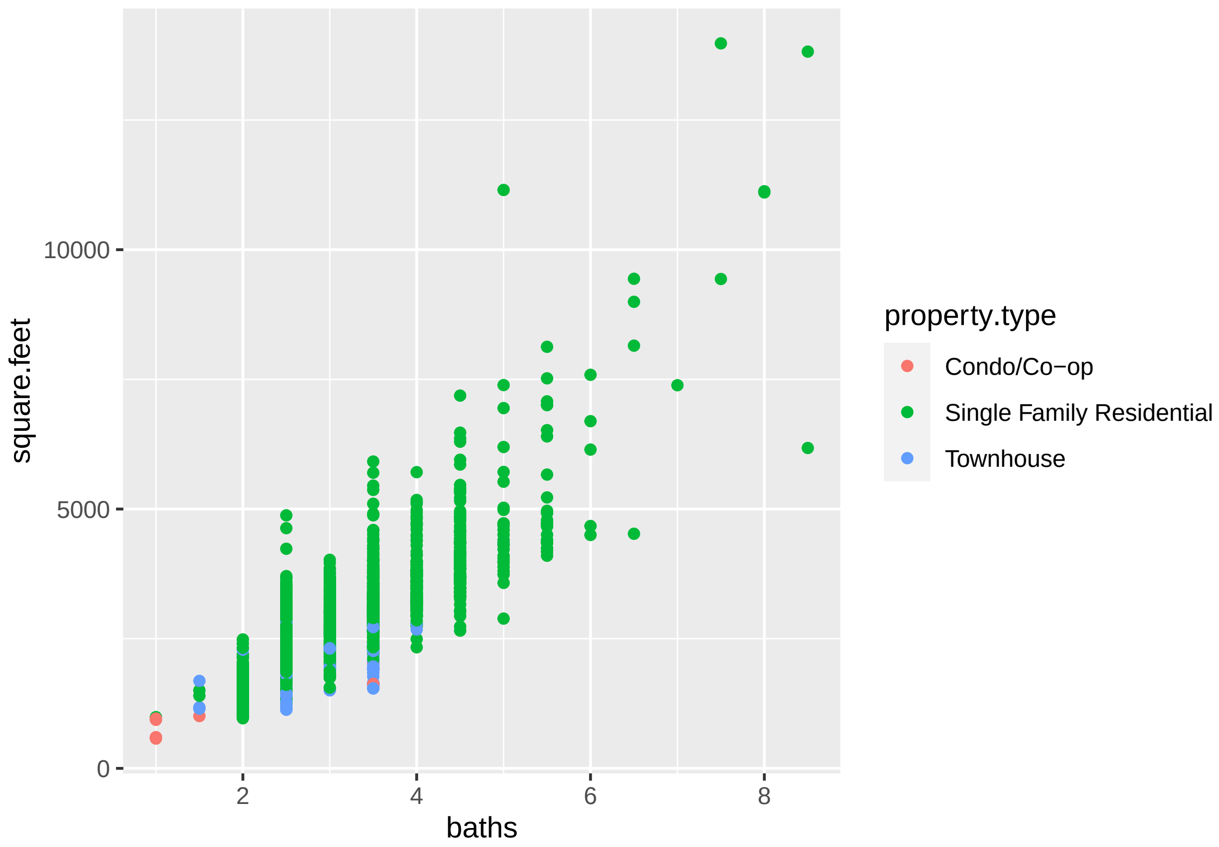
The real EDA

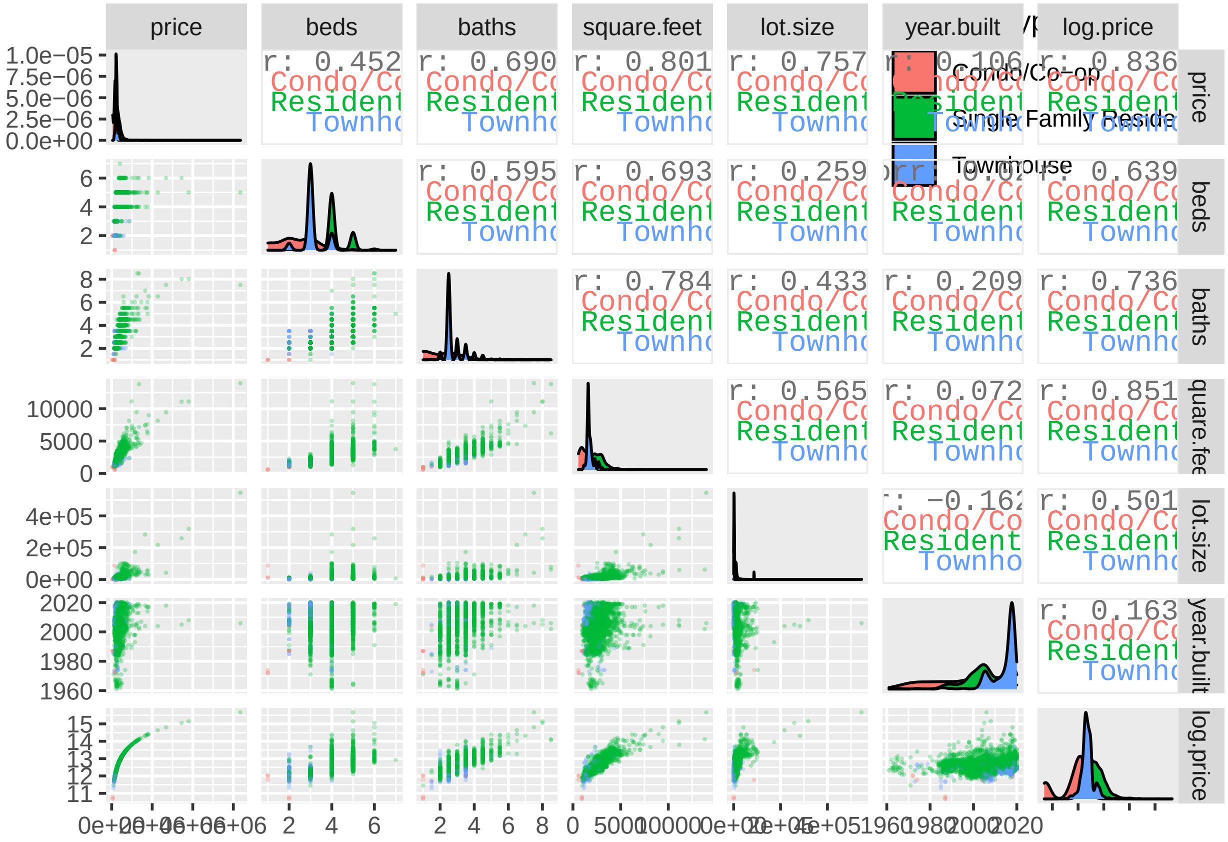
A complete pairwise-plot has been attached in the Appendix. Below are 3 main observations (concerns) and their solutions

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(***), 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 (***___)

```
## [1] 0.7843887
```


square.feet
10000
5000
0
2
4
6
8
baths
property.type
Condo/Co−op
Single Family Residential
Townhouse



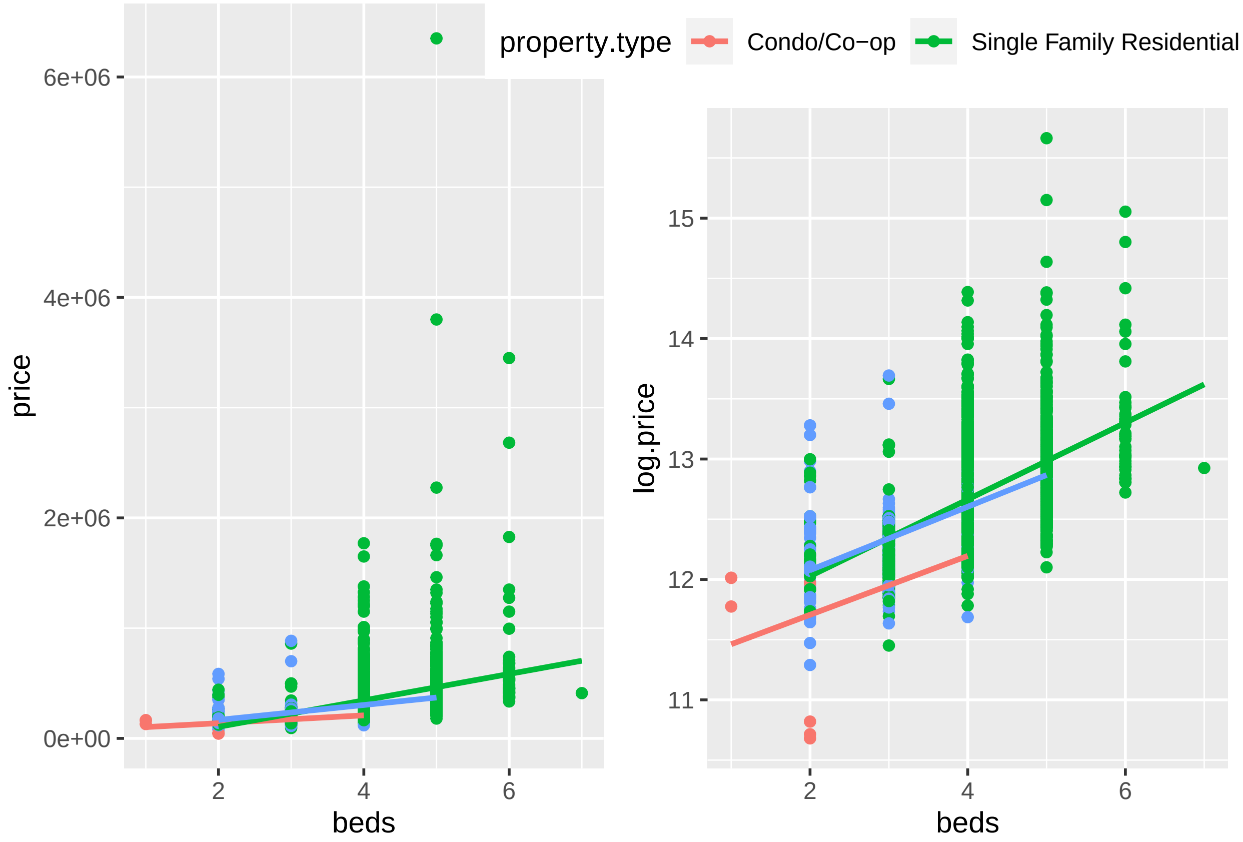
Heteroskedasticity

While some strong linearity and positive correlation is evident between 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 variables is the increase of variance. This violates the linear regression monoskedasticity assumption. To address this, we perform log-transformation on response variable and create response variable `log.price`. Shown in the last line of the pair-plot, heteroskdasticity problem is sufficiently solved without harming positive correlation between the original response and predictors. Furthermore, distribution supports a stronger linearity becomes transformed `prices`, which is `log.price`, and its predictors.

```
## Loading required package: gridExtra

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```



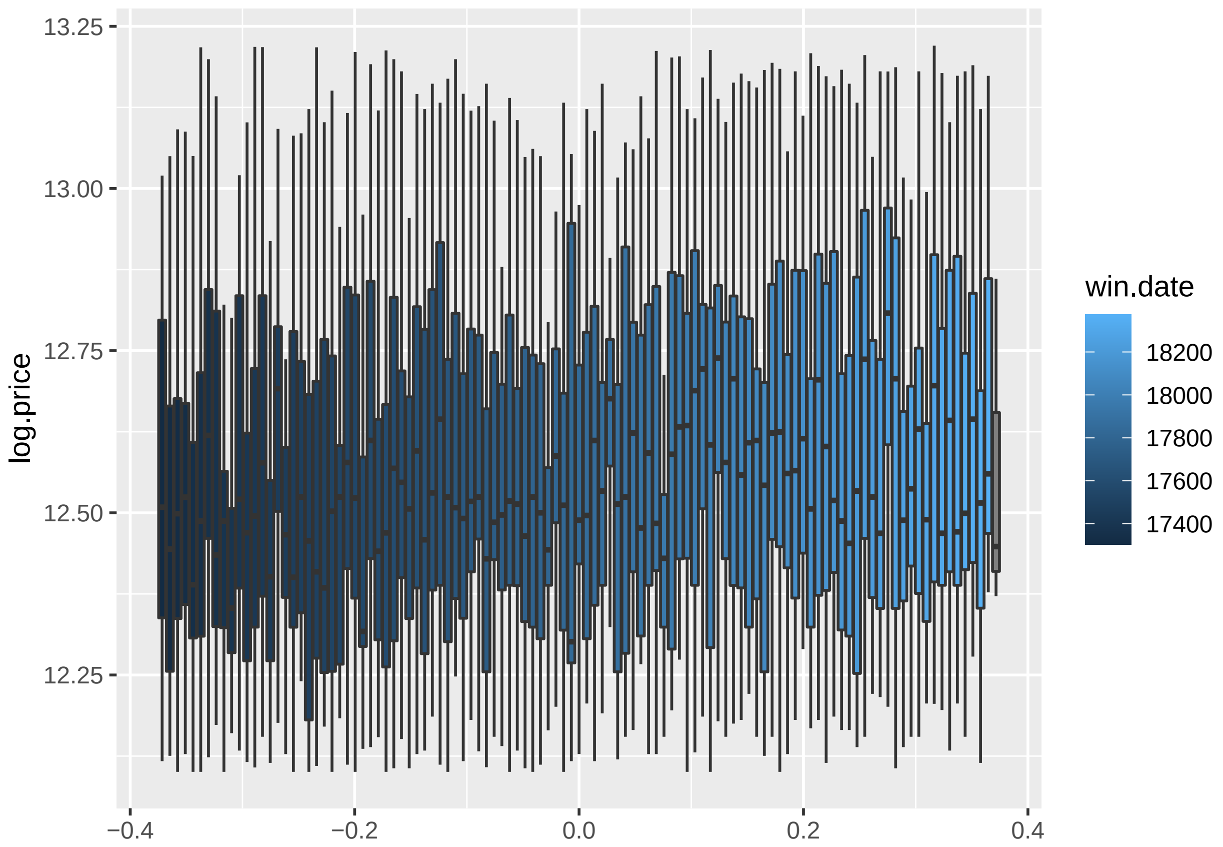
Stationarity

When determining the model for temporal effect, it is imperative to verify whether the stationarity assumption has been satisfied. In our case, due to the fact it is very unlikely that there are houses are sold everyday, we're create out time series by window period – that is, we treat all the real estate deals that lie in a prespesified window as deals happening at the same timestep. By specifying the width of our window, we can modulate and balance the amount of information to make better inference of exogenou variables (regression coefficients) versus the flexibility and variability of temporal effect. At this point, we choose window width to be 5 days.

```
onlyDurham$sold.dat <- mdy(onlyDurham$sold.dat)
begin <- min(onlyDurham$sold.dat)
end <- max(onlyDurham$sold.dat)
windowGrid <- seq(begin, end, 10)
```

```
ggplot(data=onlyDurham, aes(group=win.date, y=log.price)) +
        geom_boxplot( aes(fill=win.date), outlier.shape = NA) +
        scale_y_continuous(limits = quantile(onlyDurham$log.price, c(0.05, 0.95)))
```

```
## Warning: Removed 455 rows containing non-finite values (stat_boxplot).
```



When treating AR(p) process as the dynamic generating model for Dynamic Linear Model, one verification is stationarity of the time series. We understand that the series is the intercept of the regression model. Thus, it is infeasible to verify stationarity without knowing the coefficient series. However, at this stage, we've observed that *beds* help to explain most of variances, it suffices to use marginal distribution of prices given *sold.dat* to verify stationarity. The results for 3 beds group, the biggest group, is shown here. Other group's stationarity EDA has been attached in Appendix. The impression from the plot shows that there is a very slight upward trend in the past 3 years. Therefore, *sold.dat* should also be considered as a potential predictor variable in the regression model.

Methodology

From the introduction above, we necessarily define our model here to help you understanding the modeling process and our notation.

Regression (Observation) Model

A house's individual regression features are stated above at ※. The regression model is defined as

$$y_{ti} = \alpha_t + \beta \boldsymbol{x}_{ti} + \nu$$
$$\nu \sim \mathcal{N}(0, v)$$

Where y_{ti} is the response, the log transformed *price* (reason for transformation be explained in EDA section). α_t is the time varying intercept which will be discussed later. $\boldsymbol{\beta}$ is the regression coefficient, and $\boldsymbol{x_{ti}}$ is the ith observation at time t, which is a house vector.

Time Series (Dynamic) Model

The questions then proceed into the transition of α_t: what is the dependencies of α_t? How does α_t evolve over time?

Here, we pick a simple AR(P) process for α_t:

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

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:

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

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

Combined Model

We end up with the model as the following

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

Appendix

{Last Revised: September 13, 2020}

Some reference

[1] https://www.dallasfed.org/-/media/documents/institute/wpapers/2014/0208.pdf

[2], https://www.kaggle.com/manisaurabh/house-prices-advanced-regression-technique

[3], https://rady.ucsd.edu/faculty/directory/valkanov/pub/docs/HandRE_GPTV.pdf

[4]. https://medium.com/@feraguilari/time-series-analysis-modfinalproyect-b9fb23c28309

[5]. https://www.redfin.com/