

# House Property Sales Time Series and Linear Regression



**MAS 766 Fall 2020 | Final Project Report | Group 5**

**Jieer Chen**, School of Management, jchen257@syr.edu

**Ruiwei Zhan**, School of Information Study, rzhang72@syr.edu

**Liuqi Qian**, School of Information Study, lqian100@syr.edu

**Renjie Zhu**, School of Information Study, rzhu20@syr.edu

## **Abstract:**

Property selling prices vary from time to time. In this project, we focus on the discussion of two questions. The first question is how property selling prices change for the 2007 - 2019 time period in a region contained in our raw data. We answered this question through descriptive statistics and exploratory data analysis. The second question is how to make the best prediction of the property selling price in a certain number of following time periods in this region. We chose two types of models, Time Series Model and Linear Regression Models, for model training. As for the time series model, we ran and compared a total of nine models and reached a conclusion that the Seasonal AutoRegressive and Trend (SAR and Trend) model is the most appropriate model given its high R-squared value 0.99 and its good interpretability. As for the linear Regression models, the R-square of the best linear model is only around 0.24. Therefore, we have chosen the SAR+Trend model as our final model to make predictions.

# 1. Introduction

## 1.1 Project Purpose

The purpose of this study is to understand how the house property sales changed for the 2007-2019 period for one specific region. On the one hand, we are interested in knowing if there is any pattern in house property selling prices. For instance, whether there is a seasonal effect in our time series data and how different factors might influence property selling prices. On the other hand, we would like to study how to predict the house property selling price in this region which gives us an opportunity to employ what we have learnt from the classes and to determine a best prediction model for this dataset.

Predicting the price of house property sales in one certain area is beneficial for both individuals and companies. For individuals, we could use the model to predict the following trend of house prices so that we could figure out whether it is a good time to buy houses or sell houses. For companies, the trend of sale prices could exert an important influence on its sale strategy. Also, the analysis could help us have a better understanding of customer preferences in property types (house or apartment) and number of bedrooms. What is more, the house sale price trend could reflect the economic situation of this certain region to some extent.

## 1.2 Key findings of the project

The author who uploaded this Kaggle dataset used variance autoregressive (VAR) time series model to predict the following eight future quarters for each property type and # of bedrooms. The author found that the VAR model did not work well in this data set using MAPE as evaluation metrics and thought the result could be significantly improved.

In this project, we employed both Time Series Models and Linear Regression models. In the first part, we first used the monthly average property selling price as y variable. We found that ARIMA and SAR+Trend Model have relatively good performance compared to the rest in terms of R-squared. They both correctly show an overall upward trend. However, they failed to provide a good description of the seasonality shown by our data. In the second part, we tried to use quarterly average property selling price as the y variable. Both the ARIMA (R-squared = 0.99) model and SAR+Trend model (R-squared = 0.95) perform better in describing the trends and in making predictions for the following three years. In the third part we decided to use the SAR+Trend model instead of ARIMA model. Although the ARIMA model has a higher R square than SAR+Trend model, it is challenging to interpret the intercept and slope of the ARIMA model considering the complexity of this model. Therefore, we prefer to use the SAR+Trend model. While running the SAR+Trend model, we improved the R-squared of the model from 0.95 to 0.98 by performing a residual analysis and adding a new variable to solve the parabola trend problem. The SAR+Trend model performed much better than the best linear regression model (R-squared = 0.24) and thus is chosen as our final model.

### 1.3. Identify dependent variable and a set of independent variables

The dependent variable is price while the independent variables are Date of Sale, Number of bedrooms, Property Types. The postcode variable is discarded because we aren't able to get more information about the region and because this variable would not help improve model performance.

As for the techniques, since the data set contains not only the price of sale and date of sale, but also the property type, post codes and number of bedrooms, we decided to do time series analysis to analyze and predict the price of sale. We would also try to use other features to build linear models.

## 2. Data Description

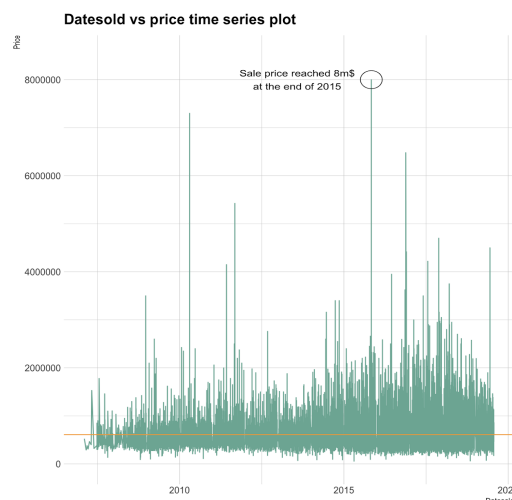
### 2.1 Data Source, Time Span, Number of Observations

We obtained our dataset from [Kaggle](https://www.kaggle.com/datasets/robertmiller/australian-property-sales). The raw data contains a total of 29580 rows and 5 columns. There is no missing value in this dataset. The five variables are Date of Sale, Price, Property Type, Number of Bedrooms, and 4-digit Postcode for locations within the region contained in this database. The variables can be summarized as:

- Date of sale: 2007 - 2019 time period
- Price: in dollars
- Property type: unit or house
- Number of bedrooms: 1,2,3,4,5
- 4-digit postcode of this specific region

### 2.2. Time series plot (Date sold vs price) and Scatter Plots

- (1) **Time series plot (Date sold vs price):** There seems to be a periodic phenomenon in the time slice, indicating that the Seasonal Autoregressive models + Trend model might be a good candidate for model selection.



- (2) **Scatter plot (Number of bedrooms vs sale price; Number of Bedrooms vs log transformation of sale price):** After transforming the sale price, we observe a clear positive correlation between the number of bedrooms and the property selling price. . The larger number of bedrooms, the higher price the property is, for both property types. Red represents houses and green represents units.

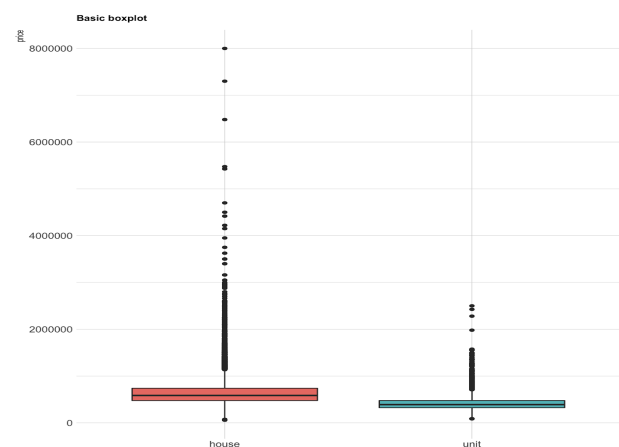
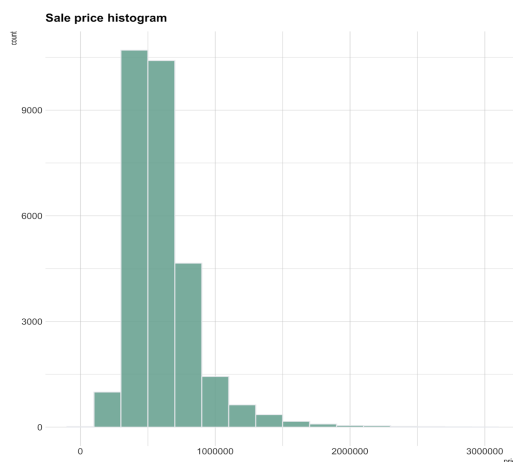


## 2.3 Summary statistics & Exploratory Data Analysis

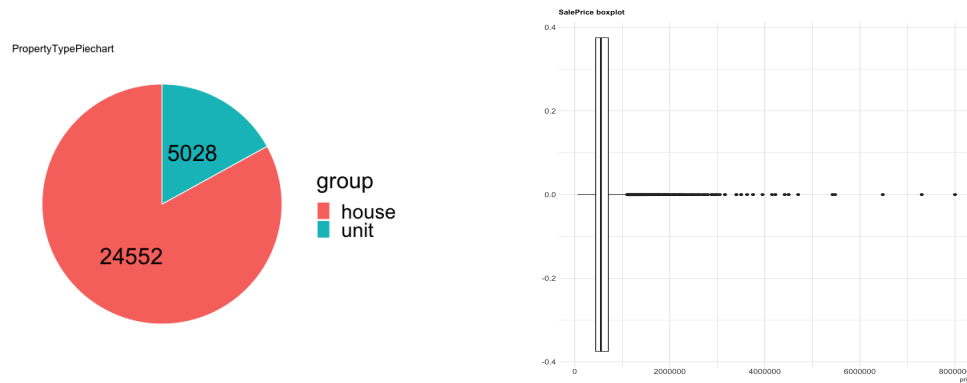
- **Descriptive statistics of target variable sale price:** The minimum selling price is as high as \$ 8,000,000 and as low as 56,500. The mean 609, 736 is not less than the median 550,000, indicating that the distribution of selling price is right skewed.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
56500	440000	550000	609736	705000	8000000

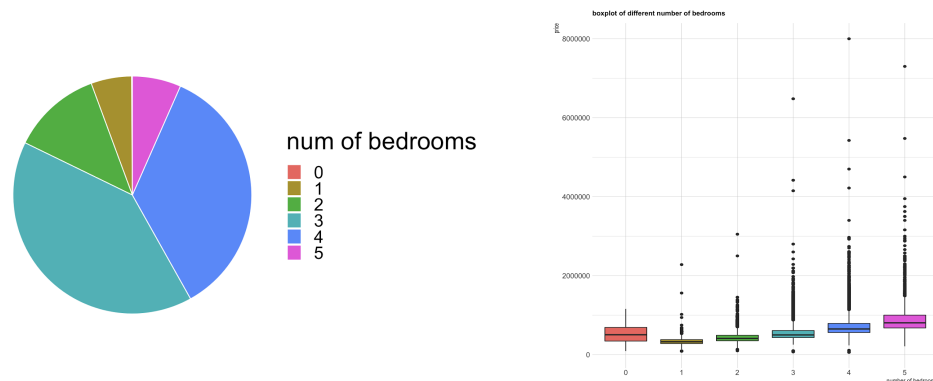
- **Distribution of target variable sale price (Histogram and Boxplot):** The distribution of prices is right skewed and most sale prices are lower than one million dollars. Since houses or units sold at high prices are common in the real estate market, we consider it normal to see outliers.



- **Dependent Variable-Property type:** There are 24552 house type observations and 5028 unit type observations. In general, unit types of properties are relatively cheaper than houses.

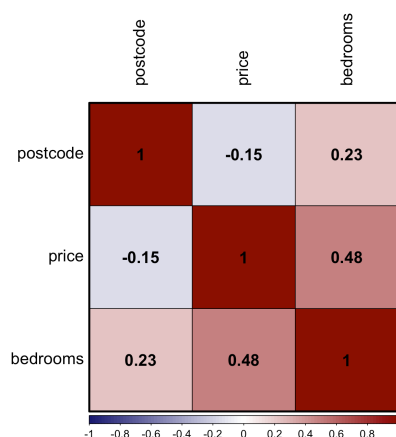


- **Dependent Variable-Number of bedrooms:** Most houses or units have 3 or 4 bedrooms. Only 30 observations have 0 bedrooms. The property with 0 bedrooms are We also draw the boxplot for each number of bedrooms and corresponding sale prices. We could see a trend that houses with more bedrooms have relatively higher average prices but not very significant.



- **Correlations:** The Pearson correlation between *the number of bedrooms* and *the price* is the highest, 0.4832117. The lowest is between *postcode* and *price*, -0.15.

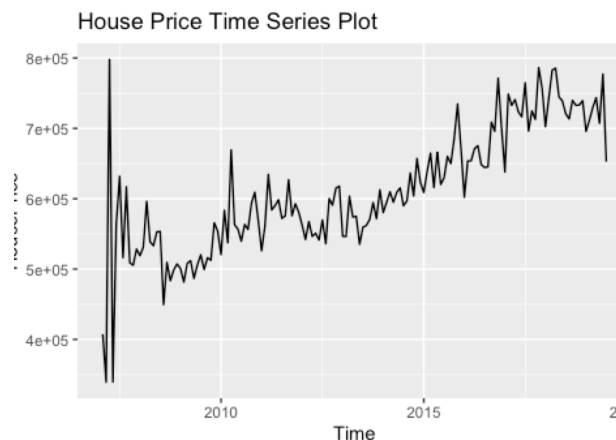
Correlation Plot



### 3. Model Selection

#### 3.1 Time series analysis with Monthly average house price as y variable

As previously mentioned, the majority of the property is house, therefore we would like to focus on only the house property type for all the time series models. We extracted the house type data, grouped them by months, and then calculated the average sale price of each month. Then, we drew a time series plot to have an understanding of the trend. The x axis is time and the y axis is the house selling price. We could see from the visualization that there are violent swings in the beginning which might be caused by the 2008 financial crisis. After that there is an increasing trend of monthly average sale price. And we could not identify the seasonality from the plot.



After that we built a series of time series models ranging from basic trend models to complicated ARIMA models. We use R-squared / Adjusted R-squared values to evaluate those models. As for the trend model, autoregressive model, and the seasonal autoregressive model, we used the basic *lm* function in R language. As for the ARIMA model, we used the *auto.arima* function from *tseries* package to automatically find the optimal model parameters.

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a widely used statistical method that captures a suite of different standard temporal structures in time series data. We firstly made an Augmented Dickey-Fuller Test to test whether the time series data is stationary. The result on the left shows that the data needs to be differencing so that it will be stationary and that we could build ARIMA models. Then, we used the *auto.arima* function in *tseries* package to automate the differencing and parameters tuning. Using that convenient function, we got our optimal ARIMA model which had 0.77 R-squared value.

```
> adf.test(tsData)
```

Augmented Dickey-Fuller Test

```
data: tsData
Dickey-Fuller = -2.2792, Lag order = 4, p-value = 0.4605
alternative hypothesis: stationary
```

```
Series: tsData
ARIMA(2,1,3)(1,0,0)[12]
```

Coefficients:

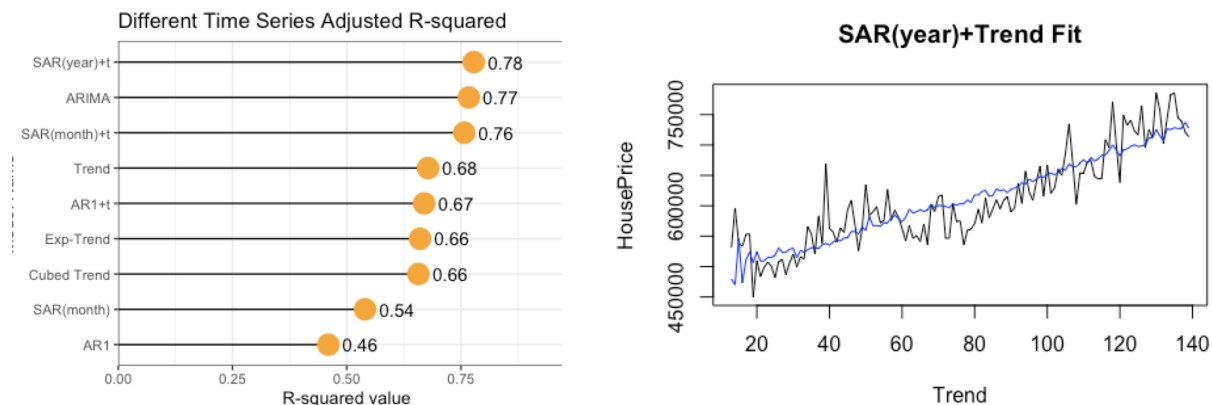
	ar1	ar2	ma1	ma2	ma3	sar1
	-1.3081	-0.905	0.2180	-0.1678	-0.2923	0.2537
s.e.	0.1180	0.093	0.1468	0.1664	0.1315	0.0988

```
sigma^2 estimated as 2.037e+09: log likelihood=-1806.83
AIC=3627.66 AICc=3628.46 BIC=3648.69
> r2<- cor(fitted(fitARIMA),tsData)^2
> r2
[1] 0.7670909
```

Below we summarized the results of 9 models in a table for a better and clear comparison. For each model, we specified its X variable, y variable, R-squared, Adjusted R-squared, and its F-test result.

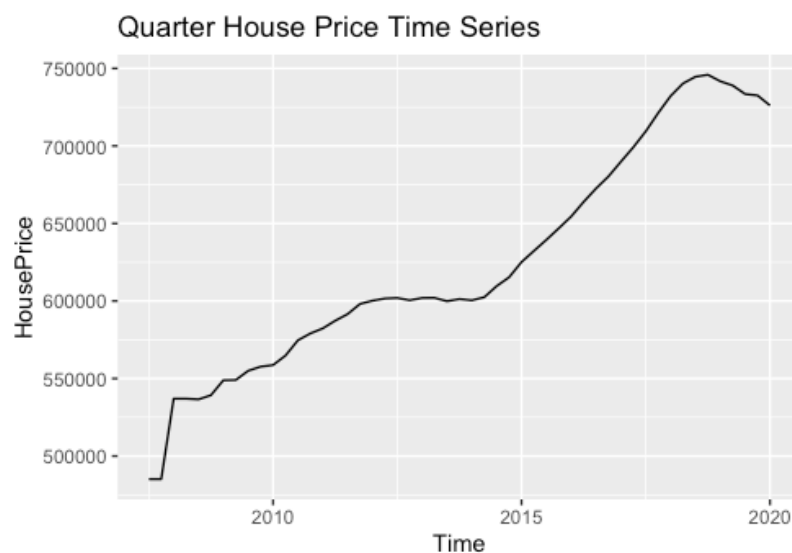
	Models	X variable	Y variable	R-squared	Adjusted R-squared	Model Significance
1.	Trend Model	Trend	Monthly Average House Price	0.6802	0.678	Significant
2.	Cubic Trend Model	Trend <sup>3</sup>	Monthly Average House Price	0.659	0.6567	Significant
3.	Exponential Trend Model	Trend	Ln(Monthly Average House Price)	0.6627	0.6604	Significant
4.	Autoregressive(AR1) Model	Price lag 1	Monthly Average House Price	0.4634	0.4597	Significant
5.	Trend + Autoregressive (AR1) Model I	Trend Price lag 1	Monthly Average House Price	0.6739	0.6694	Significant
6.	Seasonal Autoregressive (Quarter) Model	Price lag 4	Monthly Average House Price	0.5434	0.5402	Significant
7.	Trend + Seasonal Autoregressive (Quarter) Model	Trend Price lag 4	Monthly Average House Price	0.7606	0.7572	Significant
8.	Trend + Seasonal Autoregressive (Year) Model	Trend Price lag 12	Monthly Average House Price	0.7815	0.778	Significant
9.	ARIMA Model	Trend Price lag 12	Monthly Average House Price	0.7671	\	\

We also used ggplot2 to visualize the different time series model performance as follows. From the bar chart on the left we could see that the best two models are SAR+Trend model and ARIMA model. For better interpretation, we used the SAR+Trend model to fit our data to see how the model performed compared to original data. From the time series plot on the right, we see that the SAR+Trend model successfully catches the linear increasing trend generally, but it could not explain the fluctuations especially the spikes.



### 3.2 Time series analysis with quarterly average house price as y variable

The time series models we built in the previous section did a good job on identifying the increasing trend, but they aren't good enough to explain the details of the trend. Therefore, in this section, we would like to improve the performance of the best 2 models we identified from the previous section, which are the SAR+Trend model and ARIMA model. Then, we will try to forecast the sale price in the next three years (year 2020, 2021, 2022). Instead of using monthly average house selling price, we will be using quarterly average house price as the y variable. Then we built the SAR+Trend model and ARIMA model to see if the performance could be improved.





- **SAR +Trend:** Here we chose three as the cut number to neglect the influence of 2008 finance crisis on the sale price. With the new y variable, the model has a 0.9536 adjusted R-squared value. The graph on the right shows the fitted line along with original data

```
Call:
lm(formula = y ~ t + ylag4)
```

Residuals:

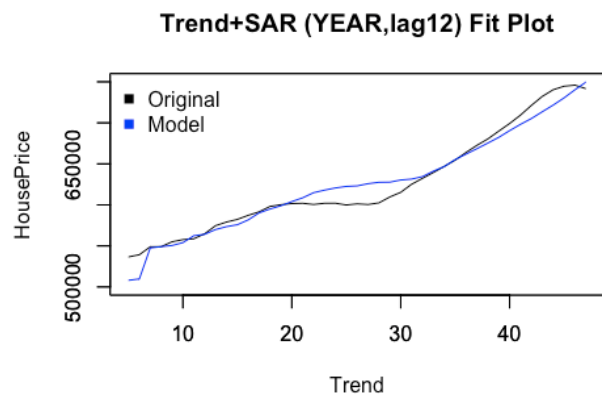
	Min	1Q	Median	3Q	Max
	-25351	-6497	1316	5613	29496

Coefficients:

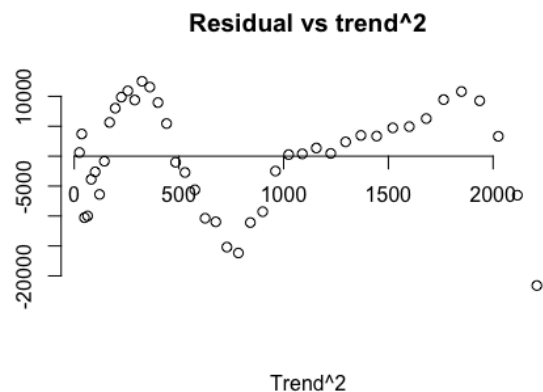
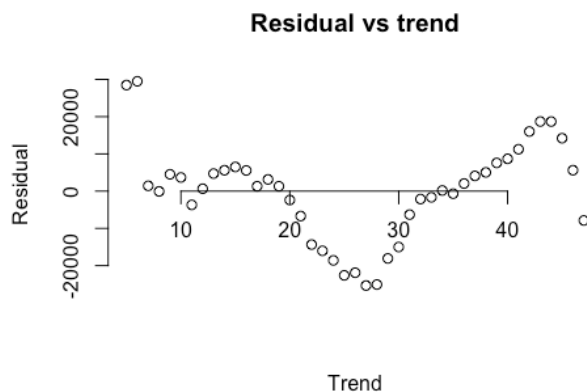
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.625e+05	6.341e+04	2.563	0.0142 *
t	1.660e+03	5.987e+02	2.773	0.0084 **
ylag4	6.952e-01	1.295e-01	5.369	3.63e-06 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13370 on 40 degrees of freedom  
Multiple R-squared: 0.9558, Adjusted R-squared: 0.9536  
F-statistic: 432.6 on 2 and 40 DF, p-value: < 2.2e-16



We continued to perform a residual analysis. The Residual vs Trend plot shows an obvious parabola trend. Thus, we added the Squared Trend to the model. The Residual vs Trend<sup>2</sup> graph on the right shows no parabola trend so this problem has been taken care of.



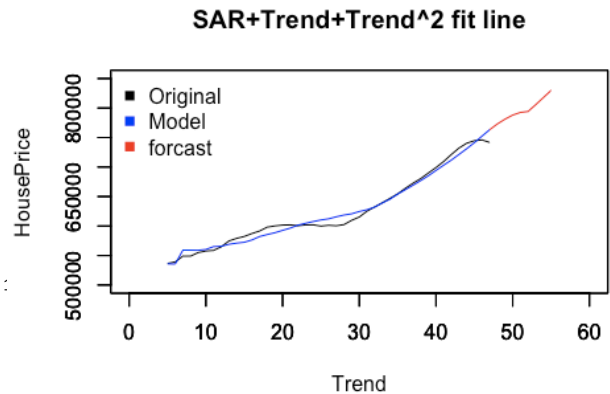
By adding the Trend Squared variable, the model's adjusted R-squared increased from 0.9536 to 0.9801. The output and the corresponding fitted plot are shown below.

```
Call:
lm(formula = y ~ t + t2 + ylag4)

Residuals:
    Min       1Q   Median       3Q      Max
-21616.1  -6003.0   608.1   5956.1  12509.1

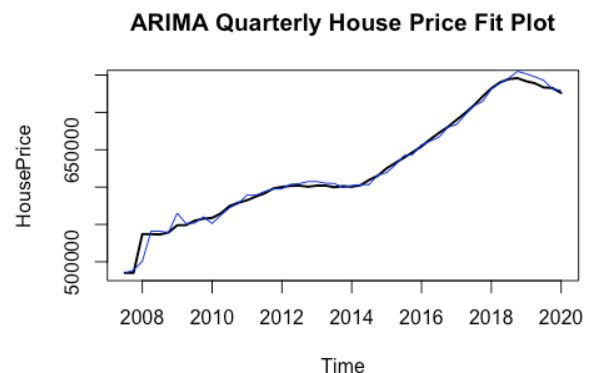
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.168e+05  4.650e+04   6.812 3.89e-08 ***
t           -1.267e+03  5.582e+02  -2.271  0.0288 *
t2            7.632e+01  1.036e+01   7.368 6.74e-09 ***
ylag4         4.610e-01  9.055e-02   5.091 9.40e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8754 on 39 degrees of freedom
Multiple R-squared:  0.9815,    Adjusted R-squared:  0.9801
F-statistic: 690.7 on 3 and 39 DF,  p-value: < 2.2e-16
```

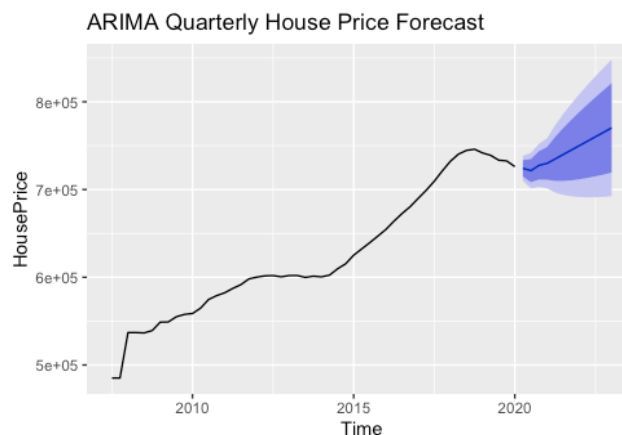


- **ARIMA Model:** Using the quarterly aggregated house selling price, the r-squared of the ARIMA model reaches 0.99092. In the ARIMA quarterly house price fit plot, the blue line shows the fitted ARIMA value while the black line was the original data. We can clearly see that the predicted values and the actual values are very close.

```
> r2<- cor(fitted(fitARIMA),tsData)^2
> r2
[1] 0.99092
```



When we applied the ARIMA model to make predictions for the following three years. In below prediction graph, the blue line was the prediction, and the purple area was the standard deviation of model prediction. We observed an upward trend.



### 3.3 Make predictions with the improved SAR+Trend Model and ARIMA Model

In this section, we applied the SAR+Trend model and ARIMA model to predict the house selling price for the next three quarters. The SAR+Trend prediction shows that the house sale price is predicted to increase steadily in the coming three years. The ARIMA prediction shows that the house selling price increases for two consecutive quarters and then decreases in the third quarters. When looking at the Mean Squared Error (MSE) and Mean Absolute Error (MAE), the ARIMA model has lower values for both thus performs better in making predictions.

- **Table:** Out of sample forecast evaluation (Using SAR+Trend +Trend2 model)

Date	PredictValue	PredictResidual
03/2019	741299.5	3318.258
06/2019	752429.2	-6521.946
09/2019	763306.4	-21616.105

MSE	MAE
13175.77	10485.44

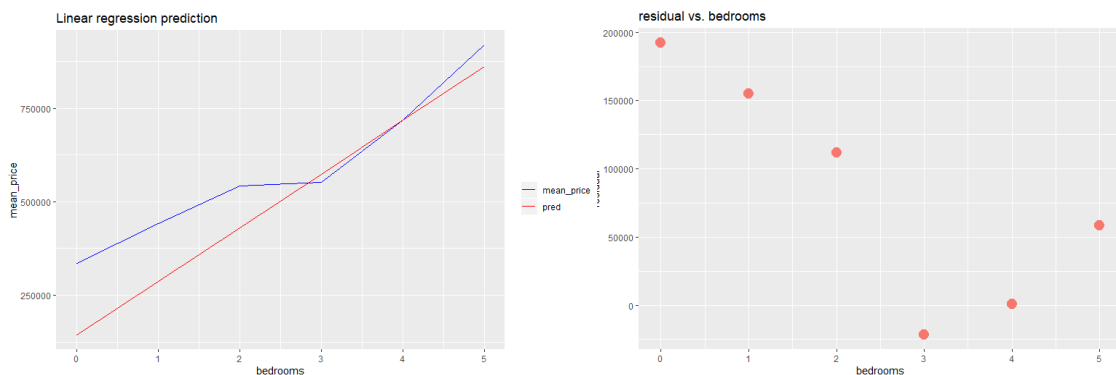
- **Table:** Out of sample forecast evaluation (Using ARIMA)

Date	PredictValue	PredictResidual
03/2019	747652.5	-14199.3
06/2019	743026.4	-10461.4
09/2019	731305.0	-5229.8

MSE	MAE
10620.9	9963.5

### 3.1 Linear Regression Analysis

In the linear regression model analysis, we first built a full model without categorical variables. That is a simple linear regression model between the number of bedrooms and the sale price in our case. The R-squared of this regression model is 0.2344 which is unexpectedly low. Secondly, we performed residual analysis. According to the plot on the left, the average selling prices increase as the bedroom number increases. The Residual vs bedrooms plot shows that when the bedroom number is low, our prediction is quite far from the average price, but in the higher number period, the prediction accuracy becomes better.



Thirdly, we added in categorical variables. We created dummy variables for the number of bedrooms and units. The R-squared increased from around 0.23 to around 0.24. Given that the R-squared is low and below 0.5. We consider the linear regression models inapplicable and will not further discuss the linear regression models.

```
Call:
lm(formula = df_1$price ~ df_1$bedrooms)

Residuals:
    Min       1Q   Median       3Q      Max
-660757 -136864  -52078   64529  7282743

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  143684      5102   28.16  <2e-16 ***
df_1$bedrooms  143393      1507   95.18  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 246500 on 29578 degrees of freedom
Multiple R-squared:  0.2345,    Adjusted R-squared:  0.2344 
F-statistic: 9059 on 1 and 29578 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = df_1$price ~ df_1$bedrooms + df_1$house)

Residuals:
    Min       1Q   Median       3Q      Max
-661772 -135124  -50956   63412  7281728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  142548      5101   27.95  < 2e-16 ***
df_1$bedrooms  152684      2033   75.11  < 2e-16 ***
df_1$house    -35012      5148   -6.80  1.06e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 246300 on 29577 degrees of freedom
Multiple R-squared:  0.2357,    Adjusted R-squared:  0.2356 
F-statistic: 4559 on 2 and 29577 DF,  p-value: < 2.2e-16
```

## 4. Final Model Interpretation: Time Series SAR+ Trend Model

We have chosen the Time Series SAR+Trend Model over the ARIMA model in the above discussion. Although the ARIMA model performed nearly perfectly, we prefer to use the SAR + Trend model because the intercepts and slopes in the ARIMA are not constant in this case. It would be relatively difficult to interpret the model in spite of its great forecast capacity.

### (1) Model assumption

- Linear
- $\varepsilon$  are serially random variables
- $\varepsilon$  are normal distributed
- $\varepsilon$  have equal variance
- Prices are influenced by *Trend* and *Seasonality*.

### (2) Estimated regression line

$$\text{SalesPrice} = 316800 - 1267 \times \text{Trend} + 76.32 \times \text{Trend}^2 + 0.4610 \times \text{ylag4}$$

### (3) Interpretation of the slope, intercept and R2

- Slope of *Trend*:

All else equal (hold the previous quarter's sale price and trend times trend fixed), as each quarter passes, the sale price is predicted to decrease by 1267 units on average.

- Slope of  $\text{Trend}^2$ :

All else equal (hold the previous quarter's sale price and trend fixed), as each quarter times quarter increases 1 unit, the sale price is predicted to increase by 76.32 units on average.

- Slope of *ylag4*:

All else equal (hold the effect of time fixed), when the sale price in the same quarter of the previous year increases by 1 unit, the sale price is predicted to increase by 0.46 units on average.

- Intercept:

When  $t$  is zero and the sale price in the same quarter of the previous year is zero, the sale price is predicted to be 316800 units on average.

- R2:

98.15% of the variation in the sale price is accounted for by linear trend and the same quarter of the previous year's sale price. Thus, about 1.85% of the variation in sale price remains unexplained.

## 5. Hypothesis test (one sided or two sided)

**Question:** Is the model as a whole a reliable model (use 5% level of significance)

Step1:  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_a$ : At least one of the  $\beta_s$  not equal to 0

Step2: Compute the test statistics

From the table we could see that the F is 690.7

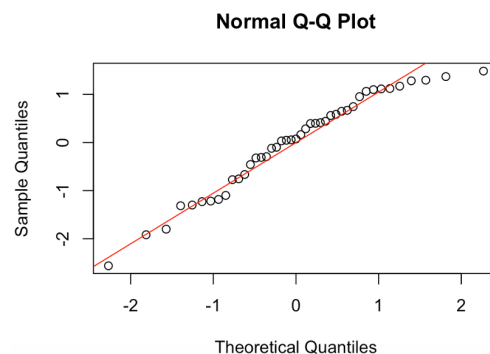
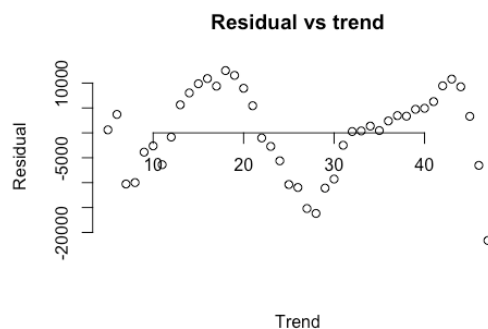
Step3: Decision Rule: Reject  $H_0$  if  $F > F_{(5\%, 3, 39)}$

Step4: Conclusion: Since  $F = 1529 > F_{(5\%, 3, 39)} = 0.12$ , reject  $H_0$ . The model is reliable.

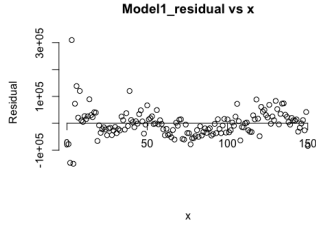
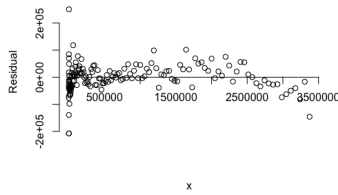
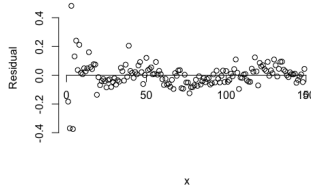
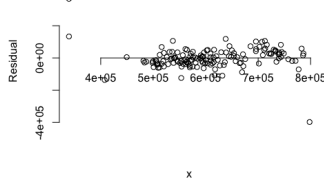
## 6. Residual analysis

We have previously performed a residual analysis to improve our time series SAR+Trend Model (with quarterly average house selling prices as the y variable). Below are the details of this residual analysis.

- **Non-linear:** A parabola problem was observed so we added a variable ( $Trend^2$ ) to fix this problem.
- **Equal variance:** From the scatter plot, we could not identify the unequal variance shapes.
- **Seasonal Pattern:** From the residual plot, there seems to be a seasonal pattern (every three year). We tried to add this feature to the model, but the model performance did not improve. So we could treat the swings as random patterns in this case.
- **Normality:** Since all the variables' p values were less than 0.05, we did not need to repeat the last three steps. Also the outliers (first several quarters) were cut due to the lag variable. Then we draw the Q-Q plot to check the normality. We could see that the majority of the data points were on the red line excluding the last three dots. In brief, through residual analysis, we found that the final model could meet the assumptions and could explain the variation of the sale price quite well.



We also drew the residual plots for the first 9 time series model we trained.

Models	Output	Residual Analysis	Observations																											
1.Trend Model	<p>Call: lm(formula = Price ~ Trend)</p> <p>Residuals:</p> <table><tr><td></td><td>Min</td><td>1Q</td><td>Median</td><td>3Q</td><td>Max</td></tr><tr><td></td><td>-150506</td><td>-29929</td><td>-2389</td><td>21657</td><td>309685</td></tr></table> <p>Coefficients:</p> <table><tr><td></td><td>Estimate</td><td>Std. Error</td><td>t value</td><td>Pr(&gt; t )</td></tr><tr><td>(Intercept)</td><td>483242.03</td><td>8295.43</td><td>58.25</td><td>&lt;2e-16 ***</td></tr><tr><td>Trend</td><td>1690.88</td><td>95.31</td><td>17.74</td><td>&lt;2e-16 ***</td></tr></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 50550 on 148 degrees of freedom Multiple R-squared: 0.6802, Adjusted R-squared: 0.678 F-statistic: 314.7 on 1 and 148 DF, p-value: &lt; 2.2e-16</p>		Min	1Q	Median	3Q	Max		-150506	-29929	-2389	21657	309685		Estimate	Std. Error	t value	Pr(> t )	(Intercept)	483242.03	8295.43	58.25	<2e-16 ***	Trend	1690.88	95.31	17.74	<2e-16 ***		<b>Problem:</b> None
	Min	1Q	Median	3Q	Max																									
	-150506	-29929	-2389	21657	309685																									
	Estimate	Std. Error	t value	Pr(> t )																										
(Intercept)	483242.03	8295.43	58.25	<2e-16 ***																										
Trend	1690.88	95.31	17.74	<2e-16 ***																										
2.Cubic Trend Model	<p>Call: lm(formula = Price ~ Cubed_Trend)</p> <p>Residuals:</p> <table><tr><td></td><td>Min</td><td>1Q</td><td>Median</td><td>3Q</td><td>Max</td></tr><tr><td></td><td>-207822</td><td>-27164</td><td>1635</td><td>26824</td><td>250843</td></tr></table> <p>Coefficients:</p> <table><tr><td></td><td>Estimate</td><td>Std. Error</td><td>t value</td><td>Pr(&gt; t )</td></tr><tr><td>(Intercept)</td><td>5.472e+05</td><td>5.689e+03</td><td>96.17</td><td>&lt;2e-16 ***</td></tr><tr><td>Cubed_Trend</td><td>7.456e-02</td><td>4.408e-03</td><td>16.91</td><td>&lt;2e-16 ***</td></tr></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 52190 on 148 degrees of freedom Multiple R-squared: 0.659, Adjusted R-squared: 0.6567 F-statistic: 286 on 1 and 148 DF, p-value: &lt; 2.2e-16</p>		Min	1Q	Median	3Q	Max		-207822	-27164	1635	26824	250843		Estimate	Std. Error	t value	Pr(> t )	(Intercept)	5.472e+05	5.689e+03	96.17	<2e-16 ***	Cubed_Trend	7.456e-02	4.408e-03	16.91	<2e-16 ***		<b>Problem:</b> Non-linearity
	Min	1Q	Median	3Q	Max																									
	-207822	-27164	1635	26824	250843																									
	Estimate	Std. Error	t value	Pr(> t )																										
(Intercept)	5.472e+05	5.689e+03	96.17	<2e-16 ***																										
Cubed_Trend	7.456e-02	4.408e-03	16.91	<2e-16 ***																										
3. Exponential Trend Model	<p>Call: lm(formula = LnPrice ~ Trend)</p> <p>Residuals:</p> <table><tr><td></td><td>Min</td><td>1Q</td><td>Median</td><td>3Q</td><td>Max</td></tr><tr><td></td><td>-0.37445</td><td>-0.04552</td><td>0.00229</td><td>0.03521</td><td>0.48301</td></tr></table> <p>Coefficients:</p> <table><tr><td></td><td>Estimate</td><td>Std. Error</td><td>t value</td><td>Pr(&gt; t )</td></tr><tr><td>(Intercept)</td><td>1.310e+01</td><td>1.442e-02</td><td>908.17</td><td>&lt;2e-16 ***</td></tr><tr><td>Trend</td><td>2.826e-03</td><td>1.657e-04</td><td>17.05</td><td>&lt;2e-16 ***</td></tr></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.08788 on 148 degrees of freedom Multiple R-squared: 0.6627, Adjusted R-squared: 0.6604 F-statistic: 290.8 on 1 and 148 DF, p-value: &lt; 2.2e-16</p>		Min	1Q	Median	3Q	Max		-0.37445	-0.04552	0.00229	0.03521	0.48301		Estimate	Std. Error	t value	Pr(> t )	(Intercept)	1.310e+01	1.442e-02	908.17	<2e-16 ***	Trend	2.826e-03	1.657e-04	17.05	<2e-16 ***		<b>Problem:</b> Non-linearity <b>Solution:</b> Use Ln(Price) instead of Price as y variable
	Min	1Q	Median	3Q	Max																									
	-0.37445	-0.04552	0.00229	0.03521	0.48301																									
	Estimate	Std. Error	t value	Pr(> t )																										
(Intercept)	1.310e+01	1.442e-02	908.17	<2e-16 ***																										
Trend	2.826e-03	1.657e-04	17.05	<2e-16 ***																										
4. Autoregressive(AR 1) Model	<p>Call: lm(formula = y ~ ylag1)</p> <p>Residuals:</p> <table><tr><td></td><td>Min</td><td>1Q</td><td>Median</td><td>3Q</td><td>Max</td></tr><tr><td></td><td>-398148</td><td>-30139</td><td>-1728</td><td>31515</td><td>367260</td></tr></table> <p>Coefficients:</p> <table><tr><td></td><td>Estimate</td><td>Std. Error</td><td>t value</td><td>Pr(&gt; t )</td></tr><tr><td>(Intercept)</td><td>2.037e+05</td><td>3.665e+04</td><td>5.558</td><td>1.25e-07 ***</td></tr><tr><td>ylag1</td><td>6.691e-01</td><td>5.939e-02</td><td>11.267</td><td>&lt; 2e-16 ***</td></tr></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 64520 on 147 degrees of freedom Multiple R-squared: 0.4634, Adjusted R-squared: 0.4597 F-statistic: 126.9 on 1 and 147 DF, p-value: &lt; 2.2e-16</p>		Min	1Q	Median	3Q	Max		-398148	-30139	-1728	31515	367260		Estimate	Std. Error	t value	Pr(> t )	(Intercept)	2.037e+05	3.665e+04	5.558	1.25e-07 ***	ylag1	6.691e-01	5.939e-02	11.267	< 2e-16 ***		<b>Problem:</b> None
	Min	1Q	Median	3Q	Max																									
	-398148	-30139	-1728	31515	367260																									
	Estimate	Std. Error	t value	Pr(> t )																										
(Intercept)	2.037e+05	3.665e+04	5.558	1.25e-07 ***																										
ylag1	6.691e-01	5.939e-02	11.267	< 2e-16 ***																										

5. Trend + Autoregressive (AR1) Model I	<pre>Call: lm(formula = y ~ ylag1 + t)  Residuals:     Min       1Q   Median       3Q      Max -153405  -27935  -3172   20973  308033  Coefficients:             Estimate Std. Error t value Pr(&gt; t ) (Intercept)  4.840e+05  4.069e+04  11.895  &lt;2e-16 *** ylag1       2.776e-03  8.289e-02  0.033  0.973 t           1.665e+03  1.715e+02  9.707  &lt;2e-16 *** --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  Residual standard error: 50470 on 146 degrees of freedom Multiple R-squared:  0.6739,    Adjusted R-squared:  0.6694 F-statistic: 150.8 on 2 and 146 DF,  p-value: &lt; 2.2e-16</pre>		<b>Problem:</b> None
6. Seasonal Autoregressive (Quarter) Model	<pre>Call: lm(formula = y ~ ylag4)  Residuals:     Min       1Q   Median       3Q      Max -226675  -32173  -8038   30291  199443  Coefficients:             Estimate Std. Error t value Pr(&gt; t ) (Intercept)  2.027e+05  3.197e+04  6.341  2.89e-09 *** ylag4       6.768e-01  5.224e-02  12.954  &lt; 2e-16 *** --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  Residual standard error: 54560 on 141 degrees of freedom Multiple R-squared:  0.5434,    Adjusted R-squared:  0.5402 F-statistic: 167.8 on 1 and 141 DF,  p-value: &lt; 2.2e-16</pre>		<b>Problem:</b> None
7. Trend + Seasonal Autoregressive (Quarter) Model	<pre>Call: lm(formula = y ~ t + ylag4)  Residuals:     Min       1Q   Median       3Q      Max  -74463  -26348  -6124   19606  148600  Coefficients:             Estimate Std. Error t value Pr(&gt; t ) (Intercept)  4.482e+05  3.184e+04  14.074  &lt;2e-16 *** t           1.561e+03  1.385e+02  11.270  &lt;2e-16 *** ylag4       7.578e-02  6.546e-02  1.158  0.249 --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  Residual standard error: 39650 on 140 degrees of freedom Multiple R-squared:  0.7606,    Adjusted R-squared:  0.7572 F-statistic: 222.4 on 2 and 140 DF,  p-value: &lt; 2.2e-16</pre>		<b>Problem:</b> None
8. Trend + Seasonal Autoregressive (Year) Model	<pre>Call: lm(formula = y ~ t + ylag12)  Residuals:     Min       1Q   Median       3Q      Max -68104  -25200  -4515   18378  131519  Coefficients:             Estimate Std. Error t value Pr(&gt; t ) (Intercept)  3.930e+05  3.053e+04  12.87  &lt;2e-16 *** t           1.600e+03  1.319e+02  12.13  &lt;2e-16 *** ylag12      1.611e-01  6.291e-02  2.56  0.0117 * --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  Residual standard error: 36570 on 124 degrees of freedom Multiple R-squared:  0.7815,    Adjusted R-squared:  0.778 F-statistic: 221.8 on 2 and 124 DF,  p-value: &lt; 2.2e-16</pre>		<b>Problem:</b> Parabola
9. ARIMA Model	<pre>Series: tsData ARIMA(2,1,3)(1,0,0)[12]  Coefficients:             ar1      ar2      ma1      ma2      ma3      sar1       -1.3081  -0.905  0.2180  -0.1678  -0.2923  0.2537 s.e.   0.1180  0.093  0.1468  0.1664  0.1315  0.0988  sigma^2 estimated as 2.037e+09: log likelihood=-1806.83 AIC=3627.66  AICc=3628.46  BIC=3648.69 &gt; r2&lt;- cor(fitted(FitARIMA),tsData)^2 &gt; r2 [1] 0.7670909</pre>		<b>Problem:</b> None



## 7. Conclusion

The above analysis helped us answer the two questions we were interested in. The first question is how property selling prices change for the 2007 - 2019 time period in the region contained in the dataset. Through descriptive statistics and exploratory data analysis, we know that the property selling prices is right-skewed and has an upward trend. Beside the increasing trend over the last decade, our model indicated that house prices would continue to rise in the coming years.

The second question is how to make the best prediction of the property selling price in a certain number of following time periods in the region contained in this dataset. We have built both linear regression models and time series models and found that the time series models have distinct advantages over linear regression models. We then focused on the best two time series models we found, the ARIMA model and the SAR+Trend model. At this stage, we found it challenging to capture the monthly price fluctuations even for powerful models like ARIMA. Therefore, in the next stage, we solved this problem by using the quarterly average price data instead of the monthly average price data. Then we proceeded to do residual analysis and applied the improved models to make predictions. At the end, we chose the SAR + Trend + Trend<sup>2</sup> time series model over the ARIMA model as our final model. The final model has a 0.9815 R-squared value and a 0.9801 adjusted R-squared value.

## **8. Appendix**

1. ARIMA material:

<https://otexts.com/fpp2/arma-r.html>

2. ggplot2 visualization examples:

<https://www.r-graph-gallery.com/histogram.html>

3. Combining Plots in R

<https://www.statmethods.net/advgraphs/layout.html>