

Housing data visualization and analysis

Yuxuan Tong

School of Statistics and Mathematics, Zhongnan University of Economics and Law,
Wuhan, 430073, China

Adrian_tyxuan@163.com

Abstract. Data visualization is a powerful tool that can assist individuals and organisations in comprehending vast amounts of data and extracting valuable insights from it. The most significant function of data visualization is to make recommendations by figuring out the essence of the occurrence of the data. This paper will take housing data as an example, raise relevant questions, and reveal the logic behind the data and the relationship between variables through data visualization, linear regression, and other statistical methods.

Keywords: Housing price, Visualization, ARMA model, Regression

1. Introduction

The Housing Data totally contains 999 samples and 21 variables. It includes id, date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, re_renovated, zipcode, lat, long, sqft_living15 and sqft_lot15. Among them, it Data visualization involves presenting data in a visually appealing and easy-to-understand manner using graphical elements like charts, graphs, and maps, aiming to help individuals grasp and interpret information more effectively. [1]The main function of data visualization is to simplify complex data sets and make it easier for users to identify patterns, trends, and relationships within the data. By presenting data visually, it allows for quick and effective analysis, decision-making, and communication of insights.

The objectives of data visualization are:

- Identify trends and patterns within the data

- Compare and contrast different data sets

- Discover insights and relationships within the data

- Make data-driven decisions based on visual representations of data

Linear regression is a statistical technique employed to establish the correlation between a dependent variable and one or multiple independent variables, by creating a linear equation that fits the observed data. The primary objective of linear regression is to identify the most optimal line that reduces the sum of squared variances between the actual data points and the model-predicted values. [2].

The main function of linear regression is to predict the value of the dependent variable based on the values of the independent variables. It helps in understanding the relationship between variables, making predictions, and identifying trends in the data.

The objectives of linear regression are:

- Predict future values of the dependent variable based on the values of the independent variables.

- Determine the strength and direction of the relationship between variables.

Evaluate the significance of the independent variables in predicting the dependent variable.

Make data-driven decisions based on the insights derived from the regression analysis.[3]

ARIMA (Auto-Regressive Integrated Moving Average Model)

The auto-regressive integrated moving average (ARIMA) model, a traditional method for time series forecasting, was introduced by Box and Jenkins in the early 1970s and has been widely utilized in market prediction. ARIMA is a mathematical model that leverages past data to predict future values of a variable. The fundamental equation for ARMA is as follows:

$$\Theta_p(B^s)\theta_p(B)(1-B^s)^D(1-B)^d y_t = \Phi_Q(B^s)\varphi_q(B)\varepsilon_t \quad (1)$$

In this equation, y_t denotes the predicted value, B stands for the lag operator, ε_t is the residuals from time series, Θ_p and θ_p , Φ_Q , and φ_q correspond to the four parameters of the ARMA model denoted as p , q , P , and Q , respectively. [5] Here, d and D represent the degrees of the seasonal and trend differences, respectively. model denote the auto-regression order, seasonal auto-regression lag, moving average order, seasonal moving average, and seasonal periodicity, respectively [6]

Typically, the ARIMA model is defined as $ARIMA(p, d, q)(P, D, Q)_s$. Nevertheless, in this paper, the ARIMA model was simplified as $ARMA(p, q)$ due to the non-seasonal nature of the monthly housing price data in the time series. The equation for this scenario can be represented as follows:

$$\theta_p(B)y_t = \varphi_q(B)\varepsilon_t \quad (2)$$

The construction process of the ARIMA model involves multiple stages. Initially, the sequence of monthly housing price cases is graphed to ascertain the stationarity of the time series. Non-stationary sequences transform such as differencing and logarithmic transformations to achieve stationarity. Subsequently, the ARIMA model parameters are estimated by examining the auto-correlation function (ACF) and partial auto-correlation function (PACF) plots post-transformations. An initial candidate ARIMA model is identified. Next, the diagnosis and assessment of the ARIMA model are conducted using the Ljung-Box (Q) test and t-test, respectively. The Ljung-Box (Q) test necessitates that the residuals of the monthly housing price time series exhibit white noise characteristics (with a significance level of $p > 0.05$). The t-test is employed to ascertain the significance of the parameters in each candidate ARIMA model. The optimal model selection is based on achieving the highest R-square value, the lowest normalized AIC, and RMSE values, along with ensuring that the residuals form white noise sequences. The Akaike information criterion is a widely utilized metric for model selection in time series forecasting, developed by Akaike, and is defined as:

$$AIC = -2 \ln(L) + 2k \quad (3)$$

Where L is the maximized value of likelihood function of the model, k is the number of parameters estimated by the model. The normalized Akaike information criterion (AIC) was used to confirm the adequacy of the model. The smaller the value of the normalized AIC, the more adequate the model fits.

To understand the meaning of data better, this paper intends to raise some questions.

1)What is the price trend?

2)Does the price have anything to do with whether it is nearby the sea? How much does the correlation have?

3)What coefficients of the housing space that is above or below the ground level? Which type do people prefer?

2. Analysis Model

Based on the housing data, with visualization tools like R and excel, this paper is able to plot the time series of the sum price, sales number, and average house price of each month.



Figure 1. month sum price trend

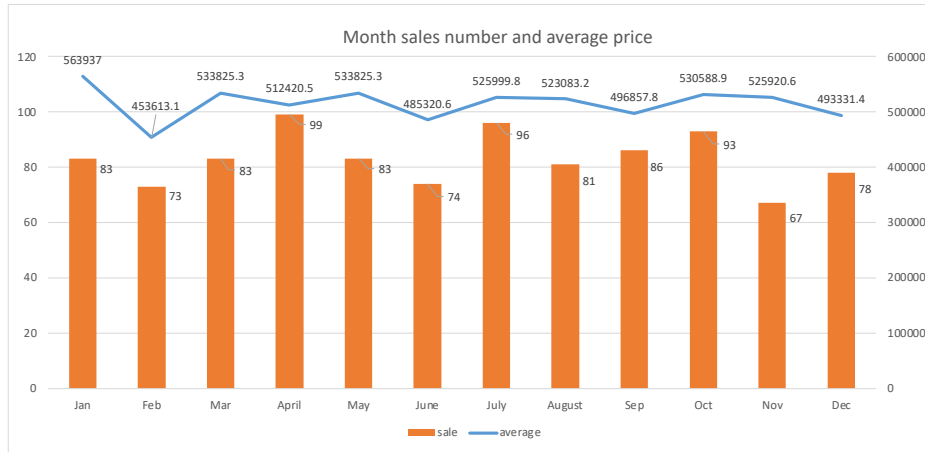


Figure 2. month sales number and average price trend

From the figure 1, it indicates that the highest sum price point is happened in April, and the lowest sales point is happened in February. From the figure 2, it suggests that the April is the month when houses are sold the most and the November is the month when houses are sold the least. From the “month average price” line chart, it shows that there is no big fluctuation in month average price so that the sum price is affected by sales number primarily.

So, by doing data visualization, the paper could suppose that the house market is become popular in April. With this conclusion, if the house agency wants to achieve a better sales goal, selecting the April as the house introduced time is a great choice.

To get more information about price trend, it is necessary to model the ARMA model due to the price is the time series.

At first, we set some models to fit the series.

Model 1:

$$\theta_p(B)y_t = \varepsilon_t, \theta_p(B) = 1 - \theta_1 B \quad (4)$$

Model 2:

$$y_t = \varphi_q(B)\varepsilon_t, \varphi_q(B) = 1 - \varphi_1 B \quad (5)$$

Model 3:

$$\theta_p(B)y_t = \varphi_q(B)\varepsilon_t, \theta_p(B) = 1 - \theta_1 B, \varphi_q(B) = 1 - \varphi_1 B \quad (6)$$

Then, by R, obtaining those parameters of three models.

For Model 1

the $\theta_1 = 0.2826$. Model 2, the $\varphi_1 = 1$. Model 3, the $\theta_1 = -0.0485$, $\varphi_1 = 1$.

From the summary of the model fit, the AIC is the value that to choose the best model of these.

Table 1. fitness comparison

	Likelihood	AIC
Model 1	-203.4	412.79
Model 2	-201.38	408.76
Model 3	-201.37	410.73

As a result, the performance of model 2 is the best. The model of price trend is

$$y_t = \varepsilon_t - \varepsilon_{t-1} \quad (7)$$

3. Waterfront

First, divide the house price into two groups. Group 1 is the house price when its waterfront = 0, and group 2 is the house price when its waterfront = 1. Then plot the two groups with R.

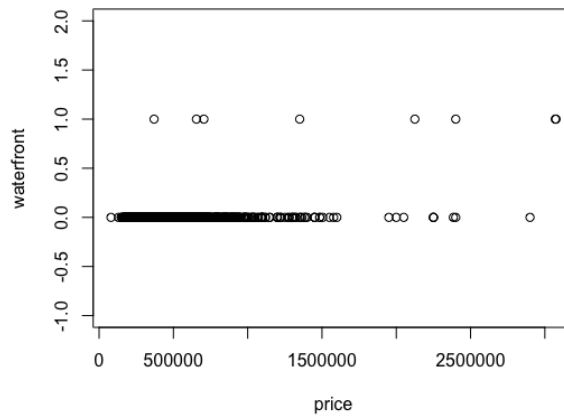


Figure 3. price grouped by waterfront

From the Figure 3, the paper can know that when waterfront = 1, the house price is distributed uniformly with no central tendency. But when waterfront = 0, the values of house price concentrate the value about 500,000, like a right-skew distribution.

So, by doing scatter plot of price with waterfront, it is preliminary reasonable to consider that waterfront have some influence to the house price.

In order to get more persuasive evidence, the paper can do a T-test to figure out the difference between two groups.

H_0 : the mean price of group 1 = the mean price of group 2

H_1 : the mean price of group 1 \neq the mean price of group 2

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (8)$$

By R, Calculating the t and get the corresponding p-value.

```
> t.test(price0,price1)

Welch Two Sample t-test

data: price0 and price1
t = -3.1157, df = 7.009, p-value = 0.01692
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
-2124127.5 -291398.1
sample estimates:
mean of x mean of y
510974.7 1718737.5
```

Figure 4. T-test

From figure 4, it is shown in the statistical results of the t-test that the p-value > 0.01. So the paper couldn't reject the null hypothesis. It should be supposed that the price doesn't have a significant relation with whether the house is near the sea.

In addition, because the distribution of price is unknown, the nonparametric test could be recommended.

The Permutation test, also known as the randomization test, is a nonparametric test proposed by Fisher and Pitman in the 20th century. The purpose of the test is whether there is a significant difference in the distribution of the two samples.[7] The idea of the test is that the empirical distribution of the test can be constructed by randomly substituting the two groups of samples in order and recalculating the test quantity. [8] Usually due to tests such as whether a drug can improve the effectiveness of treatment.

Suppose that $X_1, X_2, \dots, X_{n1} \sim F_X$ and $Y_1, Y_2, \dots, Y_{n2} \sim F_Y$ are the independent samples. The null hypothesis is the two samples come from the same distribution, $H_0: F_X = F_Y \leftrightarrow H_1: F_X \neq F_Y$.

Let $T(x_1, x_2, \dots, x_{n1}, y_1, y_2, \dots, y_{n2})$ is a test statistic, often expressed as the position difference between two sets of data:

$$T(x_1, x_2, \dots, x_{n1}, y_1, y_2, \dots, y_{n2}) = |\bar{X}_{n1} - \bar{Y}_{n2}| \quad (9)$$

The probability of a new T statistic formed by each permutation is $1/N$, where $N = C_{n1+n2}^{n1}$, then the p value of the permutation test is [9]:

$$p = \frac{\sum_{i=1}^N I(T_i > t_o)}{N} \quad (10)$$

So, in this paper, the price of waterfront equal 0 is X, and the price of waterfront equal 1 is Y. R code:

```
> permutation.test=function(x,y)
{mixsample=c(x,y)
avg1=mean(x)
avg2=mean(y)
n1=length(x)
n2=length(y)
n=avg1*n1+avg2*n2
t0=avg2-avg1
B=1000
t=0
for (i in 1:B) {
sample1=sample(mixsample,n1)
avg.sample1=mean(sample1)
avg.sample2=(n-n1*avg.sample1)/n2
t1=avg.sample2-avg.sample1
```

```

    if(t1>=t0){t=t+1}else{t=t}
  }
  p.value=t/B
  p.value
}

```

By R, it can get the value of permutation is 0, which indicates that there is no significant difference between these two price groups.

4. Regression

To establish the relation with sqft_above, sqft_basement and price, the paper can model them by linear regression.

The multiple linear regression model (MLR), an extension of simple linear regression, is employed to depict the linear correlation between numerous independent variables and a sole dependent variable. The formula for the MLR model is presented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (11)$$

Where Y is the dependent variable; X_1, X_2, \dots are the independent variables; β is the Y-intercept; $\beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients; and ε is the random error term.[10]

Multiple Linear Regression Model:

$$P = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + \alpha_0 \quad (12)$$

P : the house price, S_1 : sqft of above,
 S_2 : sqft of basement, α_0 : erro

In order to estimate the coefficient, transforming the model to the matrix.

$$P = \beta^T \times S + \varepsilon \quad (13)$$

$$\beta = (\beta_0, \beta_1, \beta_2)', \quad S = (1, S_1, S_2)'$$

Using the Ordinary Least Squares(OLS) method, the form of coefficient is:

$$\hat{\beta} = (S^T S)^{-1} S^T P \quad (14)$$

So, from the figure 5, the summary of model is:

$$P = -24677 + 259S_1 + 301S_2 + \alpha_0 \quad (15)$$

```
> summary(model3)
```

Call:

```
lm(formula = price ~ sqft_above + sqft_basement, data = house_price)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-734411	-134959	-20044	95337	2007396

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-24677.376	19453.211	-1.269	0.205
sqft_above	259.722	9.648	26.920	<2e-16 ***
sqft_basement	301.331	16.909	17.821	<2e-16 ***

Figure 5. linear regression

Further, to measure which variable has more influence on the house price, the paper can calculate the correlation between dependent variable and independent variable respectively.

$$r(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \quad (16)$$

Then, paint the correlation coefficient diagram with R



Figure 6. correlation diagram

From the figure 6, it suggests that the correlation between price and sqft_above is bigger than the correlation between price and sqft_basement. So that, the paper could reckon that sqft_above has a greater impact on the house price.

5. Conclusion

In the study, the paper collected data on American house prices in 1997 from American government data. ARIMA, MLR and Permutation tests were constructed and compared. The house market is become popular in April and the house may be sold well if it is introduced in April. the price doesn't have significant relation with whether the house is near the sea. the linear regression model is:

$$P = -24677 + 259S_1 + 301S_2 + \alpha_0 \quad (17)$$

And sqft_above has a greater impact on the house price than sqft_basement.

The study is subject to several limitations. Firstly, the ARIMA model is adept at handling linear issues, yet struggles with the nonlinear aspects of a time series. Secondly, there is a potential for under-reporting of cases and house prices in notifications, potentially introducing bias into the results. Thirdly, house prices are predominantly influenced by policy decisions. While the predicted house price values all fall within the 95% confidence interval, the absence of consideration for policy effects renders the findings less compelling.

Ignoring the distribution of price is unknown, so that the result of linear regression may be default. Besides, the factors included are too few so that the prediction effect of model is too poor. What's more, the distribution of housing price is not the normal distribution, so that linear regression shouldn't be take to fit the data.

Therefore, in future studies, it is needed to consider the influential factors that affect the fluctuation of house price in the model and update the data continuously to obtain more accurate predictions. More non-parametric statistical methods should be used to estimate and analyze the statistical housing prices. Take the General Linear Regression Model to fit the data.

References

- [1] Kolyan Ray, Botond Szabo, Variational bayes for high-dimensional linear regression with sparse priors, Journal of the American Statistical Association, 2021:142-144.
- [2] Alim, Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: A time-series study, BMJ open.10,2022:3-5.
- [3] Alabdulrazzaq, H.et al. On the accuracy of ARIMA based prediction of COVID-19 spread, ResultsPhys.27, 2021:4.

- [4] Zhai, M.et al. Research on the predictive effect of a combined model of ARIMA and neural networks on human brucellosis in Shanxi Province, China: A time series predictive analysis, *BMC Infect. Dis.* 21, 2021: 280.
- [5] SchwarzG, E. Estimating the dimension of a model, *Ann. Stat.* 2, 1978:461-464.
- [6] Damette, O., Mathonnat, C. & Goute, S. Meteorological factors against COVID-19 and the role of human mobility. *Plos one* 16, 2021:4-12.
- [7] Kai Xu, Hao Yin, Mengxiao Chen, Jialong Zhang. A robust permutation test for Kendall's tau. *Journal of Statistical Computation and Simulation*, 2022:3-8.
- [8] Thomas B. Berrett. The conditional permutation test for independence while controlling for confounders. *Journal of the royal statistical society series B-statistical methodology*. 2019:66-70.
- [9] Gary Doran. A permutation-based kernel conditional independence test. *Uncertainty in Artificial intelligence*, 2014:42-45.
- [10] Boxian Zhou. Network traffic prediction based on ARMA model. *Journal of computer research and development*, 2002:19-23.