**Module7\_** **Critical thinking\_Option#1**

PinJu Chen

Colorado State University Global

MIS581: Capstone: Business Intelligence and Data Analytics

Orenthio Goodwin

01/08/2023

**ABSTRACT**

The industry analyzed this time is the real estate industry. Due to the dual nature of real estate, on the one hand, real estate is a necessity of life and one of the basic needs of people. On the other hand, real estate is often used as an investment target. In the United States, real estate has experienced rapid growth and recession in recent years, from explosive growth during the COVID-19 period to the continuous decline in mid-2022 due to the increase in interest rates. After this period, we believe that a reasonable valuation of the house is a very important thing. We downloaded a dataset of 82 explanatory variables from the residential real estate website in Ames, Iowa, and tried to find the variables that had the greatest impact on real estate prices, and built a multiple linear regression model to predict the housing price. In addition, we also find out the season when the property is in great demand, and the number of rooms, the number of bathrooms that are popular. Through the multiple linear regression model, we found that the biggest factors affecting housing prices are size in square feet, masonry veneer area in square feet, above-grade living area in square feet, size of a garage in square feet, and remodel date, especially the influence of remodel date is huge, and the price has nothing to do with the original construction date. In addition, we also know that in the Ames, Iowa area, the most popular number of rooms and the number of bathrooms are 6 rooms and 3 bathrooms, respectively. The hottest season is summer, especially June. Based on the models we construct, architects, planners, and real estate managers can use this research to obtain pricing and house construction advice. While no single study can cover all possible, our study provides a direction for the study of real estate prices in Ames, Iowa, but the variables that affect real estate prices in different regions are also different.

**INTRODUCTION**

The target industry we want to analyze this time is the real estate industry. In law, the word real, as it relates to property, means land as distinguished from personal property; and the estate is defined as the interest one has in the property (Funk & Wagnalls New World Encyclopedia, 2018). Due to the duality of real estate, on the one hand, real estate is a necessity that is one of the basic needs of people. On the other hand, real estate is also a good investment target for investors, although involves many risks, as well as various forms of financial support and investment, the rewards are huge. In most societies, rich or poor, a significant fraction of the total wealth is in the form of land and buildings, dealing in real estate can be very profitable, due to factors such as the rising demand for property (Tetteh, V. A., 2021). The rapid growth and decline of the real estate industry in the United States in recent years have deepened the uncertainty of this industry. For the 11th straight month in November and plunged to its lowest level since June 2012, excluding the COVID-19 pandemic, according to the National Association of Home Builders’ monthly survey(Thomas, 2022). so a reasonable estimate of housing has become more important.

We downloaded the dataset from the website which is the residential real estate in Ames, Iowa with 82 explanatory variables. Try to find the variables that most affect real estate prices among these variables, and build a multiple linear regression model to predict each house price in Ames, Iowa.

**OBJECTIVES**

There are a total of 82 variables in the data set, and we must use statistical analysis to find important factors that have a potential impact on prices, and build a model to predict housing prices in the area. We think the multiple linear regression model is the best method to help us predict housing prices. Use multiple regression to find the value that best calculates the specific data for which they are searching. This data can help predict the factors that result in an outcome or forecast an effect or trend (Sheposh, 2020). By predicting the housing price in Ames, Iowa, we will use the multiple linear regression model to find the relationship between variables and housing price, if there has a significant relationship between them, this means that these variables have a large impact on house prices.

**RESEARCH QUESTIONS AND HYPOTHESES**

**Q1: Housing prices are affected by lot size, type of dwelling, remodel date, type of foundation, number of kitchens.**

**Null hypothesis (H0):** The top five variables that affect house prices the most are not lot size in square feet (LotArea), type of dwelling (BldgType), remodel date (YearRemodAdd), type of foundation (Foundation), number of kitchens (Kitchen).

**Alternate hypothesis (Ha):** The top five variables that affect house prices the most are lot size in square feet (LotArea), type of dwelling (BldgType), remodel date (YearRemodAdd), type of foundation (Foundation), number of kitchens (Kitchen).

**Q2: If the bid is less than $170,000, the buyer can buy a house built before 2000.**

**Null hypothesis (H0):** If the bid is less than $170,000, the buyer can not buy a house built before 2000.

**Alternate hypothesis (Ha):** If the bid is less than $170,000, the buyer can buy a house built before 2000.

**Q3: The most popular total rooms and the number of bedrooms in this area are 5 and 2 respectively.**

**Null hypothesis (H0):** The most popular total rooms (TotRmsAbvGrd) and Number of bedrooms (Bedroom) in this area are not 5 and 2.

**Alternate hypothesis (Ha):** The most popular total rooms (TotRmsAbvGrd) and Number of bedrooms (Bedroom) in this area are 5 and 2 respectively.

**Q4: June is the hottest time of year for house sales in Ames, Iowa.**

**Null hypothesis (H0):** June is not the hottest time of year for house sales in Ames, Iowa. **Alternate hypothesis (Ha):** June is the hottest time of year for house sales in Ames, Iowa.

**LITERATURE REVIEW**

In Phipps, A. G., & Li, D’s study, they compared the hybrid price model from Quigley with the repeat sales model and the single sales hedonic model. To evaluate the effect of these three models, the authors use observations' 70 variables that were analyzed within the executable limits of a linear regression procedure in Excel, to predict the prices of 1,275 sold and 1,284 resold homes observed in two inner-city neighborhoods in Windsor, Ontario, during a 30-year period. Through this paper comparing the correctness of various models for price prediction, we know that the goodness of fits obtained under different conditions is different. And multiple linear regression is a good method for price prediction and it statistically explains more of the variation in the house price dependent variable.

The other paper using a linear regression model to build a real estate price prediction model is “Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model”. In this paper, The authors think the multiple regression model only considers the variables that affect housing prices. But there have some factors that are nonparametric but also affect the housing price, failure to account for these variables can lead to a decrease in the accuracy of the model. So the author uses the generalized linear regression model to improve the reliability of housing price prediction. As the result, the housing price prediction system based on the generalized regression model proposed in this article has a high housing price prediction accuracy(Li X, 2022).

The hedonic price model is one of the popular models to predict the housing markets, the study” The Choice of Functional Form and Variables in the Hedonic Price Model in Seoul” tries to prove the applicability of the hedonic price model in different countries and housing markets. According to this article, we know that the characteristic price structure of the real estate market in each region is different. Like in my research object Ames, Iowa, there may be some factors that affect the proportion of housing prices higher than others, so we must consider the region and cultural background, and then adjust.

The last paper we review this time is “Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model.” which analyzes the factors of the house price, and establishes a multiple linear regression model. Then uses the least square method to solve the unknown parameters of the model. Through such a research method, we can find the factors that affect housing prices and show a high degree of positive correlation and build a model to predict future prices and analysis.

**RESEARCH DESIGN**

**Methodology**

There are 82 variables in the dataset we use this time, 38 of those variables are numeric data type and 44 variables are varchar data type. The methodology we choice is quantitative research, because our reaserch aim and objectives are not exploratory in nature, we focus on the relathipnship between variables hosuing price and other varilables and test a set of hypotheses.



**Methods**

From the 82 variables in the dataset, find the five variables that affect the house price the most, which allows us to build a model to predict the appropriate price of each house. In the future, in Ames, Iowa, the analysis of the process of house sales or purchases can focus on these five variables variable. The method we use this time is quantitative research which uses a large amount of data to support an argument. The source is usually in the form of data and collect by statistical methods, then the data is analyzed to obtain results. To a positivist researcher, the reality(quantitative research) is objective and independent of the researcher. Research is formal, value-free, and unbiased(Smith, T, 2021)

The statistical test that can be used to check the relationship between variables is the linear regression model. The multiple linear regression model is the best method to help us predict housing prices. Use multiple regression to find the value that best calculates the specific data for which they are searching. This data can help predict the factors that result in an outcome or forecast an effect or trend (Sheposh, 2020). The multiple linear regression model formula in our model is Y= 𝛽0+ 𝛽1X+ 𝛽2X+𝛽3X+...+𝛽82X+ε. Then R-squared can be used to check the linear regression performance. If R-squared=0, then it is a very poor fit for the data. If R-squared =1 indicates the line is a perfect fit.

The statistical test that can be used to predict the probability corresponding to an event that has two outcomes is the logistic regression model. Logistic regression also referred to as the logit model, is used to model dichotomous outcome variables. In a logit model, the log odds of the outcome are modeled as a linear combination of the predictor variables. A BLR model is a regression model established for binary response variables, and its independent variables can represent qualitative or quantitative data(Cui, H. & Wang, 2021).

To validate the model, we have two hypotheses which are H0 and H1 and use P-value to check the results. Usually, statistical significance in this context is defined as a pre-set P-value <0.05. Ninety-six percent of articles reported in PubMed Central with one or more P-values claimed a statistically significant result with more than one P-value<0.05.( Gale. & Zhang, 2016)

After Finding the important predictors, we can use them to verify the correctness of the model and find the individual odds of the two outcomes. The probability of success is related to the independent variable X, which is the year the house was built. In the logistic regression model, Y is a binary variable (yes or no, exist or not exist), and Y in our model is the house sales amount 170,000. If the result is greater than or equal to 0, then it is one type; otherwise, it is another type. If the Alternate hypothesis is true, we can predict whether houses built after 2000(original construction date) can be purchased for 170,000 dollars.

Analyzing the combination of the number of bedrooms(variable Bedroom) and total rooms (TotRmsAbvGrd) in this area can be used as a reference for future house construction, and it can also let sellers who want to sell their houses know whether their houses are popular in the market. This also uses a linear regression model to calculate the influence of variables and to visualize the results.

The hot time of sale is a very important factor in real estate sales. It is generally believed that July and August are the hottest months for sales. Through data analysis, we can understand which month is the hottest month of the year for sales in Ames, Iowa. Then which year is the time when most homes are sold, those can give sellers a reference for when is the best time to sell a home.

Regarding ethical considerations for the type of data, there are very high standards in the medical industry and the banking industry. For example, Tabary, Mohammadreza, author of the article "Ethical considerations in neurology during the COVID-19 pandemic" say, what is widely acknowledged is that each ethical framework must engage and reflect the views of the public and at the same time address the priorities of the society (Tabary 2021).

**Limitations**

In the real estate industry, every buyer wants to buy their dream house at the lowest price, and every seller wants to sell their house at the highest price as quickly as possible. But no one can know the right price for that house. We try to use the housing prices dataset in Ames, Iowa to find the best model to predict the housing price. Based on the model we build, no matter whether buyers, sellers, architects, planners, or real estate managers, can get price-setting advice in this area can be found through this research.

No single study can cover all possible angles. Our research provides a direction for real estate price research, but the variables affected by real estate prices in different regions must be different. This model is suitable for Ames, Iowa, but not necessarily applicable to all of the world, and also maybe not necessarily applicable to all of America. This time there are 82 sample variables and 1,459 samples, but I think a larger sample size is needed to ensure fair representation.

**Ethical Considerations**

Unlike the medical industry or the banking industry, which require confidentiality of customer information, real estate information is publicly available on the Internet for everyone to check, such as transaction date, construction date, transaction price, location, interior decoration, garage type, etc. Information about the house itself can basically be found on various real estate websites. But we still need to avoid violations of IT ethics such as illegal use of personal data, license infringement, and exposure of individual house addresses and private information on the Internet during the research process.

**Finding**

**Q1: Housing prices are affected by lot size, building date, remodel date.**

**Null hypothesis (H0):** The top seven variables that affect house prices the most are not lot size in square feet (LotArea), Original construction date (YearBuilt), Remodel date(YearRemodAdd), Masonry veneer area in square feet (MasVnrArea), Above grade iving area square feet (GrLivArea), Size of garage in square feet(GarageArea), Pool area in square feet(PoolArea).

**Alternate hypothesis (Ha):** The top seven variables that affect house prices the most are size in square feet (LotArea), Original construction date (YearBuilt), Remodel date(YearRemodAdd), Masonry veneer area in square feet (MasVnrArea), Above grade iving area square feet (GrLivArea), Size of garage in square feet(GarageArea), Pool area in square feet(PoolArea).

In the multi-linear regression model we created by SAS, the formula is sale price = 323,278+1.99634X1-81.24917X2-11.47399x3+16.43253x4-7.05923X5-9.45242X6-29.65547X7+ε. The P-values in are LotArea, YearRemodAdd, GrLivArea, MasVnrArea, and GarageArea are all <0.001 which is significant. The P-values in YearBuilt and PoolArea are not <0.001 which is not significantly affect the Y(sale price). R-squared in our model = 0.6584 which means there has 65% of the data can be explained by our model. It can be seen that there is a positive correlation between the variable LotArea and Remodel date with the Y(sale price). The price rises when LotArea increases and Remodel date is closer to now.

Due to the results, we can not accept the Ha, the Housing prices are affected by size in square feet (LotArea), Remodel date(YearRemodAdd), Masonry veneer area in square feet (MasVnrArea), Above grade living area square feet (GrLivArea), Size of garage in square feet(GarageArea), but does not have a significant relationship with Original construction date (YearBuilt) and Pool area in square feet(PoolArea).

**Q2: If the bid is less than $170,000, the buyer can buy a house built before 2000.**

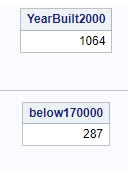
**Null hypothesis (H0):** If the bid is less than $170,000, the buyer can not buy a house built before 2000.

**Alternate hypothesis (Ha):** If the bid is less than $170,000, the buyer can buy a house built before 2000.

In order to analyze this problem, we wrote code in SAS.



There have 1,459 houses in the dataset, 1,064 houses built before 2000, and 287 of them are under 170,000. So accept the Alternate hypothesis (Ha) that the bid is less than $170,000, the buyer still can buy a house built before 2000.



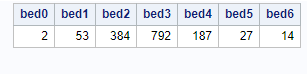
**Q3: The most popular total rooms and the number of bedrooms in this area are 5 and 2 respectively.**

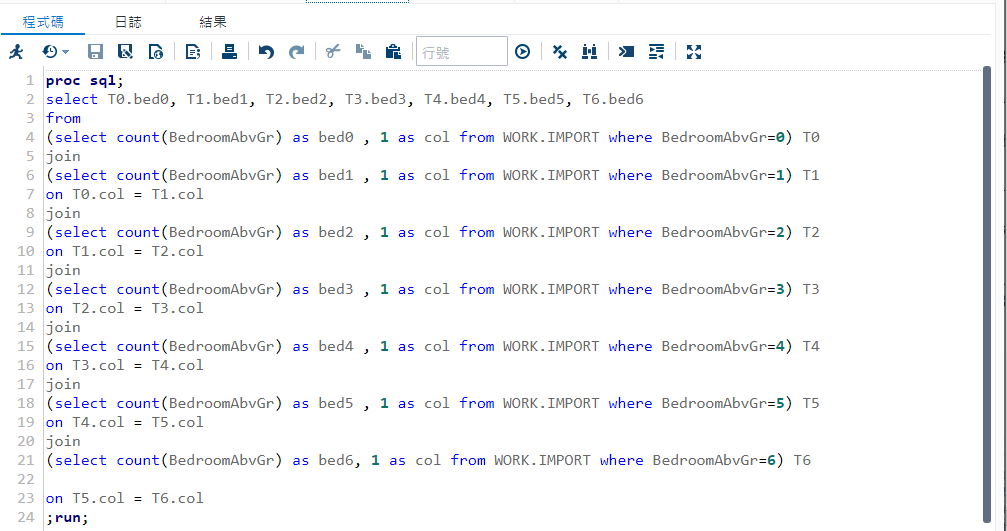
**Null hypothesis (H0):** The most popular total rooms (TotRmsAbvGrd) and Number of bedrooms (BedroomAbvGr) in this area are not 5 and 2.

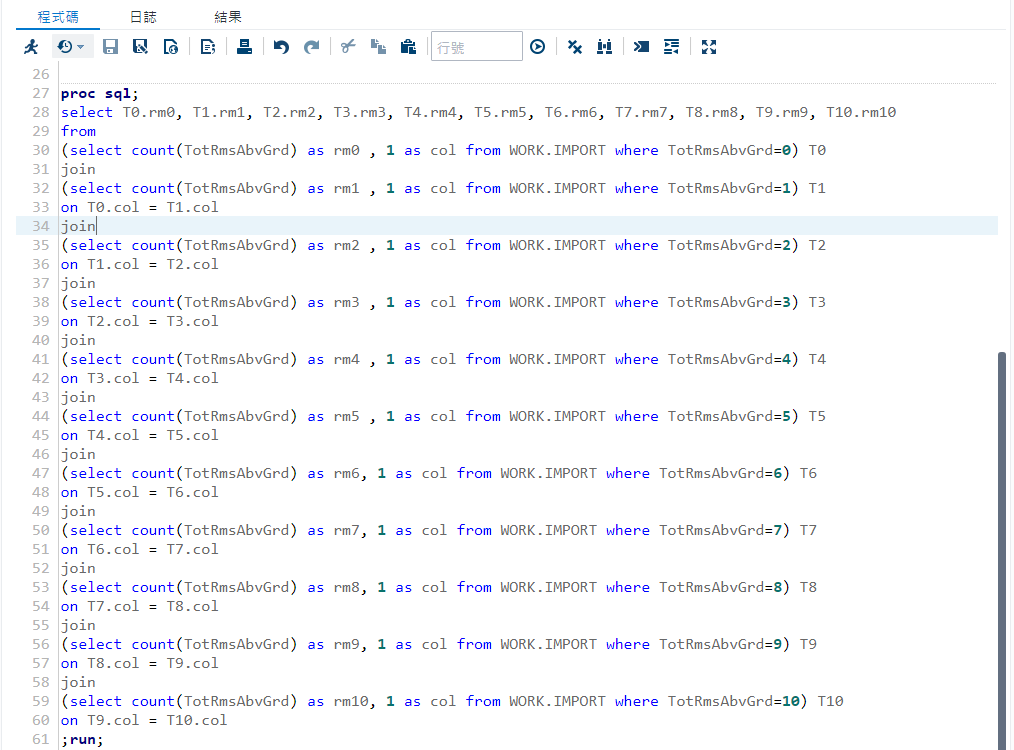
**Alternate hypothesis (Ha):** The most popular total rooms (TotRmsAbvGrd) and Number of bedrooms (BedroomAbvGr) in this area are 5 and 2 respectively.

By listing all category’s quantities, the most popular total room number is 6 rooms not including bedrooms, the most popular number of bedrooms is 3 bedrooms. So we reject the Alternate hypothesis (Ha).





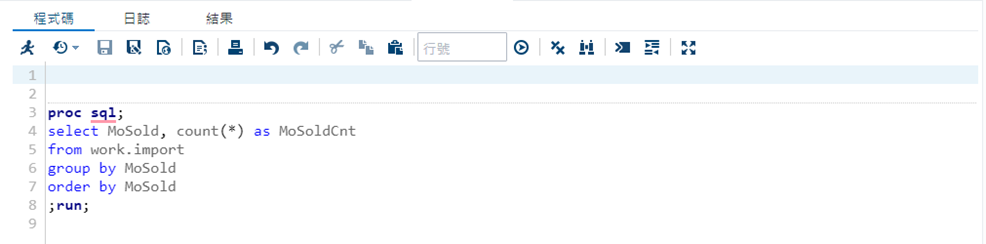


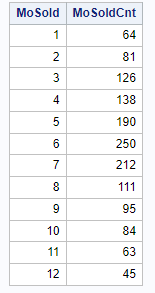


**Q4: June is the hottest time of year for house sales in Ames, Iowa.**

**Null hypothesis (H0):** June is not the hottest time of year for house sales in Ames, Iowa. **Alternate hypothesis (Ha):** June is the hottest time of year for house sales in Ames, Iowa.

To analyze this problem, we wrote code in SAS. The hottest time of year for house sales in Ames, Iowa is June. May and July are also very hot for sales. We can accept the Ha and suggest to any customer who wants to sell their house in Ames, Iowa, summer is the best time.





**REFERENCES**

Phipps, A. G., & Li, D. (2019). Calibration and evaluation of Quigley’s hybrid housing price model in Microsoft Excel. *PloS One, 14*(4), e0215954. https://doi.org/10.1371/journal.pone.0215954

Li, X. (2022). Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model. *Computational Intelligence and Neuroscience, 2022, 3590224.* https://doi.org/10.1155/2022/3590224

Huh, S., & Kwak, S.-J. (1997). The Choice of Functional Form and Variables in the Hedonic Price Model in Seoul. *Urban Studies*, 34(7), 989–998.

Liu, G. (2022). Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. *Scientific Programming*, 1–8. https://doi.org/10.1155/2022/5750354

Tetteh, V. A. (2021). Real Estate Entrepreneurship. *Salem Press Encyclopedia.*

Real Estate. (2018). *Funk & Wagnalls New World Encyclopedia,* 1;

Sheposh, R. (2020). Multiple regression. *Salem Press Encyclopedia.*

Gale, R., Hochhaus, A. & Zhang, MJ. What is the (p-) value of the P-value?. *Leukemia 30*, 1965–1967 (2016). https://doi.org/10.1038/leu.2016.193

Cui, H., Wu, L., Hu, S., Lu, R., & Wang, S. (2021). Research on the driving forces of urban hot spots based on exploratory analysis and binary logistic regression model. Transactions in GIS, 25(3), 1522–1541. <https://doi.org/10.1111/tgis.12739>

Smith, T. (2021). Qualitative and quantitative research. *Salem Press Encyclopedia.*