Attribution Analysis of the House Sale Prices Based on the Multiple Linear Regression

Li Si Xuan

21/12/2020

Link to a Github repository

Code and data supporting this analysis is available at: https://github.com/LLINDAx/STA304-Final-Project

Abstract

The visual analysis of the cleaned data and further imputation analysis of the house sale prices were performed. Since the relationship between other variables and the house sale prices was of concern, a multiple linear regression model was used and the most significant house's characteristics affecting the house sale prices were found to be the number of bedrooms, the square footage of the land space and the number of floors.

Keywords

Multiple Linear Regression, Real Estate, Prediction, House Sale Prices

Introduction

Statistics play an important role in many fields, such as real estate. For most people, buying or selling a house is a relatively large financial transaction in their lifetime. First-time buyers might meet overpaying. They may have met a seller with a high bid, but the house is not worth that much money. On the other hand, current homeowners might meet problems with losing money in a sale because they don't know how much their house should sell. Hence, it can be seen that both buyers and sellers may encounter complex problems in real estate transactions.(S., 2018)

Furthermore, real estate investment has always been a hot topic. Many people choose to invest in real estate. Because compared with other investment projects, such as stocks and bonds, there tends to be less volatility in real estate. Also, real estate investment is relatively safe because it has a high tangible asset value and a large appreciation space(R., 2020).

Whether you want to buy a house for your own living or for investment, you need to be able to understand the housing prices in recent years. And what will affect the price of a house, such as the location, size, etc. Therefore, we analyzed the regression of these houses' characteristics on the house sale prices by building a multiple linear regression model. Further, the evaluation of the multiple linear regression model led to some conclusions on the relationship between the house's characteristics and the house sale prices, and to achieve our goal of predicting housing prices. Finally, in the discussion of the experimental procedure, it is suggested that the model may have some stability problems caused by endogeneity and some directions for improvement are proposed.

Methodology

Data

The dataset includes houses sold between May 2014 and May 2015 for King Country, USA. And this dataset comes from Kaggle, which is a website that allowed users to find freely available dataset (Harlfoxem, 2016). There are 21613 observations and 21 variables in the dataset.

The variables in the dataset are:
id: Unique ID for each home sold
date: Date of the home sale
price: Price of each home sold
bedrooms: Number of bedrooms
bathrooms: Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living: Square footage of the apartments interior living space
sqft_lot: Square footage of the land space
floors: Number of floors
waterfront: - A dummy variable for whether the apartment was overlooking the waterfront or not
view: An index from 0 to 4 of how good the view of the property was
condition: - An index from 1 to 5 on the condition of the apartment,
grade: An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
sqft_above: The square footage of the interior housing space that is above ground level
sqft_basement: The square footage of the interior housing space that is below ground level
yr_built: The year the house was initially built
yr_renovated: The year of the house's last renovation
zipcode: What zipcode area the house is in
lat: Lattitude
long: Longitude
sqft_living15: The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15: The square footage of the land lots of the nearest 15 neighbors
(Mantero, 2020)

Since our goal is to predict housing prices, we need to remove some unrelated variables and some unuseful variables, such as **id**, **lat**, **date**, etc. Also, we need to remove overly complex variables and some variables are almost the same, such as **zipcode**, **sqft_lot15**, **sqft_living**, etc. Then, we use the remaining variable to form a new dataset. There are 7 variables in it. The data is cleaned and analyzed by RStudio. The following table is showing summary statistics of this dataset:

Table 1: Summary Statistics of Data

	Overall
n	21613
price (mean (SD))	540088.14 (367127.20)
bedrooms (mean (SD))	3.37 (0.93)
bathrooms (mean (SD))	2.11 (0.77)
sqft_lot (mean (SD))	15106.97 (41420.51)
floors (mean (SD))	1.49 (0.54)
sqft_above (mean (SD))	1788.39 (828.09)
sqft_basement (mean (SD))	291.51 (442.58)

From the table 1, we can see the means and the SDs of each variables in the dataset.