Analysis on the Housing Market using Multiple Linear Regression

Yena Joo

December 2020

Analysis of Key Factors that affect the Housing Price and Causal Inference of Renovation affecting the Housing Price

Code and data supporting this analysis as well as the README file are available at: https://github.com/yenajoo/304final.git

Abstract

To predict the movement of the real estate market, the Housing Price Prediction dataset from Kaggle - a data science community with various public open data - is used to study and analyze some potential factors affecting the dwelling prices. The dataset includes information about dwellings sold between May 2014 and May 2015 (10), and only useful quantitative variables are selected through a simple data cleaning, as well as eliminating outliers/influential points in order to obtain a smooth analysis. To select the significant variables that affect the dwelling prices, Akaike Information Criterion(AIC) stepwise selection method[1] is used, and Multiple linear regression model is used to determine the relationship between each predictor variables and prices, as well as the Propensity score matching to determine the causal relationship between the renovated house and the house price. Through the analysis, positive relationships are found between the independent variables - living area(sqft) and condition of the house - and dependent variable house price, but the number of bedrooms and the year the house is built has a negative effect on the house price. The propensity matching score shows that the house renovation causes an increase in housing prices.

Keywords

housing price prediction, multiple linear regression, AIC selection method, VIF, model assumptions, treatment/control group, causal inference, propensity score matching

Introduction

The world has been overturned by an unexpected virus COVID-19, and it seems that the real estate market has been under the spotlight since the Covid-19 pandemic. People started to look for houses rather than apartments and condos. The spread of the virus caused the housing price to fluctuate by almost 40% (11). The real estate market is one of the most demanding markets for people since it is one of the most important components of people living. To have a better understanding of the housing prices, it is important to know what affect house price to increase for the future home owners. On the other hand, people who already own houses have a different perspective on the real estate market. TV shows about home renovations can be

[1]*AIC stepwise selection method is explained in the Model section

easily found, and the shows are very popular. A lot of people are interested in renovating their houses, and they wonder if the renovation would make their house worth more and could be sold at a higher price.

Throughout the report, we are going to analyze some potential factors that affect housing prices to determine which characteristics of the houses have shown the higher correlation to the housing price, and analyze the causal link between the houses that have been renovated and that have not been renovated using the Propensity Score Matching[2]. between renovated houses and non-renovated houses, one being assumed as a "treatment" group[3], where they are expected higher price by renovation. The other houses would be in a "control" group[4], where the house has never been renovated.

To build a multiple linear regression model of housing prices by potential factors, Akaike Information Criterion(AIC) stepwise selection method is used, and the model is used to interpret the regression output to find relationships between the housing prices and potential factors, such as the number of bedrooms, bathrooms, living area, year built, and condition of the house. In the Data and Model section, characteristics of the dataset are explained, as well as the model that is used to analyze the relationship. The result of the Propensity score matching and the multiple linear regression are provided in the Results section and elaborated in the Discussion section, as well as the weaknesses of the analysis.

[2]*Propensity score matching is explain in "Causal inference using Propensity score matching" section.

[3]*treatment group is the group of people/observations who received a treatment based on the effect researchers wish to study, usually in an experimental study

[4]* Control group is a group of observations who do not receive a treatment

Data

The dataset contains house sale prices for King County, which includes Seattle, downloaded from a data science community Kaggle. It includes houses sold between May 2014 and May 2015 (10). It consists of 21613 observations and 21 variables. Through a simple data cleaning process[5], only 16227 observations are used, and 8 variables are included in the cleaned dataset. The target population of the analysis includes all sales of houses in the USA. The frame population is houses that can be bought or sold within King County, including Seatle, and the sampled population is the house sales record in King County in 2014-2015. There is insufficient information given regarding the population of its own dataset.

Some key features of the dataset are that every data is quantitative values, which is perfect for linear regression. Also, all the components that are planned to use in building the regression model are the factors people actually consider when they plan to buy a house, such as how many bedrooms or bathrooms the house has, when the house is built, whether the house is in a good condition, and how big the house is. Also, the reason for choosing the 2014-2015 dataset is because it shows the most recent housing prices which are before the pandemic.

However, there exists some weakness with the data. The dataset obtains the real-estate market information in King County, which might not fully reflect the housing prices in the USA. Also, since the COVID-19 pandemic made the real-estate market to fluctuate, 2014-2015 house sales data might not fully represent the prediction of the house price made in the analysis. Also, the dataset is not the best choice for the causal inference since there is no information about the prices of the house before it was renovated. In this case, propensity score matching is a better method to use rather than the difference-in-differences method(DID)[6].

Figure 1: Scatterplot of the Raw Data

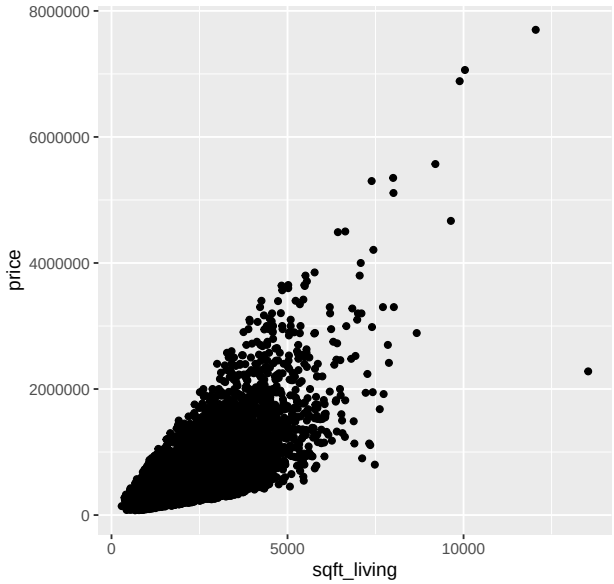
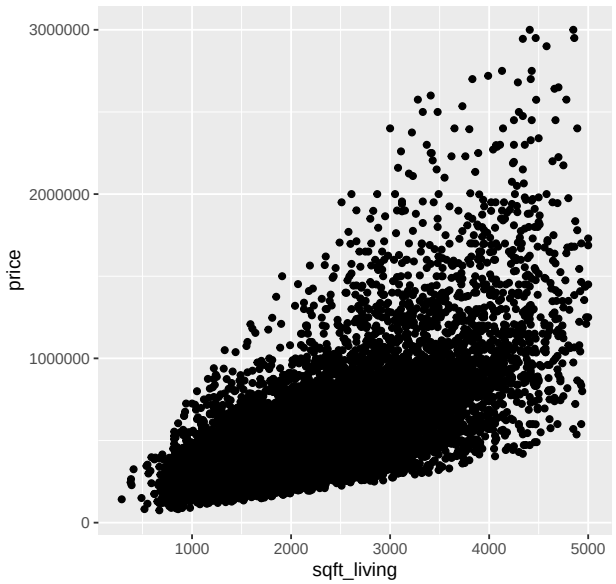


Figure 2: Scatterplot of the Cleaned Data



An important part of the data cleaning to consider is to eliminate the outliers. This process could be done by looking at the scatter plot of the data. As shown in Figure 1, The scatterplot looks densely distributed from the price range $0 to about $3,000,000. The observations that have price larger than $3,000,000 are omitted from the dataset through the data cleaning.
Also, houses with 0 bathroom are omitted since houses with 0 bathroom might be a potential estimate error, as well as houses built before 1950, since old houses are more likely to be rebuilt, and we would like to predict the future housing prices. Through this process, there are 16227 remaining observations (Figure 2) used in the future model.

[5]*The cleaning process is introduced throughout the Data section.

[6]*DID is a method that determines the causal inference by comparing the changes in outcomes over time between a population in the treatment group and the control group

Table 1: Baseline Characteristics

	Not Renovated	Renovated	p	test
n	15929	298		
price (mean (SD))	515634.01 (295080.26)	793176.22 (514652.59)	<0.001	
bedrooms (mean (SD))	3.46 (0.82)	3.61 (0.93)	0.002	
bathrooms (mean (SD))	2.24 (0.68)	2.45 (0.82)	<0.001	
sqft_living (mean (SD))	2156.10 (813.38)	2480.32 (975.24)	<0.001	
floors (mean (SD))	1.55 (0.57)	1.45 (0.52)	0.002	
condition (mean (SD))	3.34 (0.59)	3.13 (0.36)	<0.001	
yr_built (mean (SD))	1984.24 (18.86)	1964.43 (10.83)	<0.001	
yr_renovated (mean (SD))	0.00 (0.00)	2001.99 (10.13)	<0.001	

Table 1 is the baseline characteristics table and there are 9 variables included - the price of the house, number of bedrooms, bathrooms, living area(sqft), number of floors, condition, year built, year renovated, and treatment variable renovation is newly created for the future propensity score matching. The values on the left column with the title "Not Renovated" are the values of a control group, and the right side "Renovated" is a treatment group. The table shows the average value of each predictor variable, divided into two groups.

To briefly explain what each variable is, `price_house` is the price of the house. `bedrooms` and `bathrooms` are the numbers of bedrooms and bathrooms in the house. `sqft_living` is the living area of the house in the sqft unit, and `floors` variable is the number of floors in the house. the `condition` indicates how good the condition of the house is, on a scale from 0 to 5. `yr_built` is the year the house was built, and `yr_renovated` is the year the house was renovated. If the house is never renovated, the value of yr_renovated is 0. `treatment_group` is a dummy variable[7] that has a value of 0 if the house is never renovated(which is a control group), and 1 if the house was renovated at least one time(treatment).

[7]*Dummy variable is a variable that takes values of 0 and 1, where the values indicate a control variable if 0 and 1 indicating a treatment.

Model

Multiple Linear Regression model is chosen for the analysis since it contains a lot of quantitative variables that are suitable for the linear regression. The predictor variables used in the model are the number of bedrooms, bathrooms, living area in sqft, year built, and condition of the house. This section explains how the model is selected.

Model Selection

There could be various potential factors affecting house prices. Therefore, it is critical to determine which variables should be included in the multiple linear regression. There are various ways to determine, but AIC stepwise selection method is going to be used in the model.

AIC is a short form of Akaike Information criterion, which quantifies the amount of information loss due to the simplification. AIC uses a model's maximum likelihood estimation as a measure of fit. Simply, smaller AIC shows improvement in model performance. Through the process of eliminating and adding the variables, it calculates and compares AIC in each step and determines the model with the lowest AIC which is the best fit for the data (8).

Formula for AIC is the following:

$$\text{AIC} = -2(log - likelihood) + 2k \quad (1)$$

Where k is the number of predictor variables included in the model, and the Log-likelihood estimate indicates a measure of goodness of fit for any model.

There are 3 methods, forward selection, backward elimination, and the combination of both. Forward selection starts from one variable, and iteratively adds one variable at a time to compare the AIC value until the AIC is the smallest. Backward elimination starts with a full model that includes every variable in the model, and iteratively removes the least contributed predictors until it reaches the lowest AIC value. Combination of both eliminates or adds potential contributive predictor variables, and stops when you have a model where all predictors are statistically significant (15).

Table 2: AIC model summary

term	estimate	std.error	statistic	p.value
sqft_living	260.58992	3.146406	82.821457	0
bedrooms	-58352.22288	2520.791520	-23.148373	0
condition	39273.67204	2936.306710	13.375194	0
bathrooms	49460.90775	3619.622863	13.664658	0
yr_built	-41.95124	6.475582	-6.478374	0

Using the AIC selection method, the following variables in Table 2 are selected: sqft_living, bedrooms, bathrooms, year built, and condition.

The equation for the regression is:

$$\text{Housing Price} = \hat{B_0} + \hat{B_1} * x_{bedrooms} + \hat{B_2} * x_{bathrooms} + \hat{B_3} * x_{sqftliving} + \hat{B_4} * x_{yrbuilt} + \hat{B_5} * x_{condition} \quad (2)$$

The equation above[8] is going to be interpreted for the multiple linear regression in the Results section.

[8]* the description of each variable can be found in the footnotes of the Result section.

Results

Model Diagnostics

With the regression model created in the Model section, it is critical to keep in mind that the multiple linear regression analysis is well performed under the following assumptions:

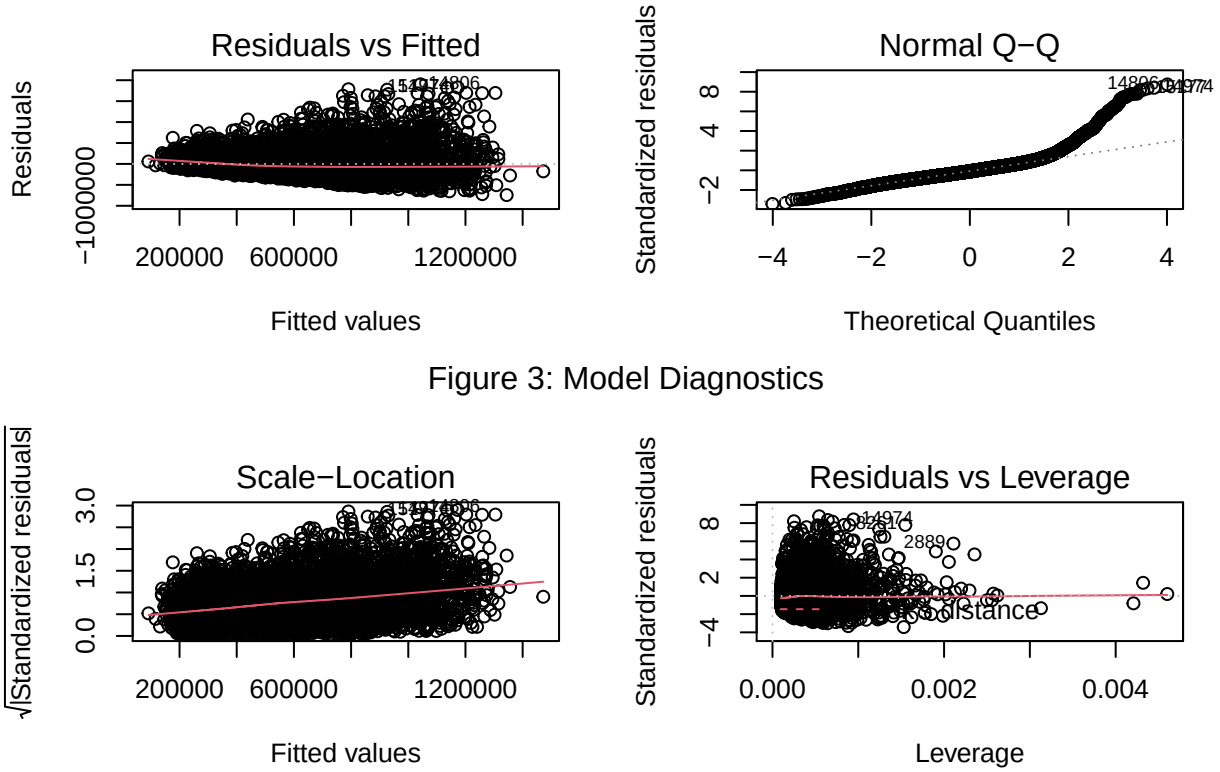


Figure 3: Model Diagnostics

1. Linearity: There should be a linear relationship between the dependent and independent variables. By showing the scatter plot above(Figure 2), it satisfies the linearity assumption. It can also be checked by the top-left graph in Figure 3, where the residual plot shows no visible pattern, and the red line is almost horizontal(17). Hence, the first assumption is satisfied.

2. Multivariate Normality: Residuals of the regression should be normally distributed. This assumption may be checked by looking at a histogram or a QQ-Plot on the top right corner of Figure 3.
Up to 2 theoretical quantiles, the data points perfectly trend the reference line and has a slight high tail at the end of the line. The data points that are off the line might indicate there exists some outliers, it should be into account that the result drawn from the regression model could be misleading or biased(16). This assumption is going to be dealt in the Weakness section. However, overall, most of the points fall approximately along the reference line, hence the second assumption is also satisfied.

3. No multicollinearity: Independent variables should not be highly correlated with each other.
This assumption could be checked by using the Variance Inflation Factor(VIF) values. VIF indicates the value of how much the variances in the regression estimates are increased due to multicollinearity. The following table shows the VIFs of the model. The left table of Table 3 is the initial model from the AIC selection method. Two variables, bathrooms and living area(sqft) have relatively high VIF which indicates that multicollinearity is a problem for the two variables. Therefore, the variable bathrooms

Table 3: VIF models

variables	VIF
bedrooms	1.595009
bathrooms	2.693725
living area(sqft)	2.293528
year built	1.790572
condition	1.245823

variables	VIF
bedrooms	1.514887
living area(sqft)	1.641526
year built	1.363597
condition	1.243540

are omitted from the model as seen in the right table, and VIF values do not exceed 2 in the adjusted model. Hence, after eliminating one variable, the third assumption is also satisfied(18).

4. Homoscedasticity: Variance of error terms should be similar across the values of the independent variables.
   This assumption can be checked through the plot of standardized residuals vs predicted values, having points equally distributed across all values of independent variables(17). To have the assumption satisfied, there should be no clear pattern in the distribution.
   The bottom left graph, which is the Scale-Location graph in Figure 3 determines whether the model satisfies the fourth assumption, by looking at the red line. A horizontal line with well-distributed points is a good indication of homoscedasticity(13). Therefore, the homoscedasticity assumption is also satisfied.

Results of the Model

The following is the summary of the multiple linear regression model using the selected predictor variables:

Table 4: Regression summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1228613.9495	213475.791438	5.755285	0
bedrooms	-56137.1984	2541.952084	-22.084287	0
living area(sqft)	290.2568	2.676011	108.466212	0
year built	-613.0295	105.363629	-5.818227	0
condition	22445.7958	3266.511068	6.871489	0

The regression equation using the estimates from Table 4, the equation is the following:

$$\text{Housing Price} = 1228613.95 - 56137.20 * x_{bedrooms} + 290.26 * x_{sqft\ living} - 613.03 * x_{yr\ built} + 22445.80 * x_{condition} \quad (3)$$

(The detailed descriptions of the x variables could be found in the footnote[9]).

In Table 4, as observed from the output, the price of the house increases as the number of a living area(in sqft), and condition increase. As 1 sqft increase of the house, the house price is expected to increase about $290, and if the house condition rating is 1 unit higher, the house has a $22445 higher expected price.

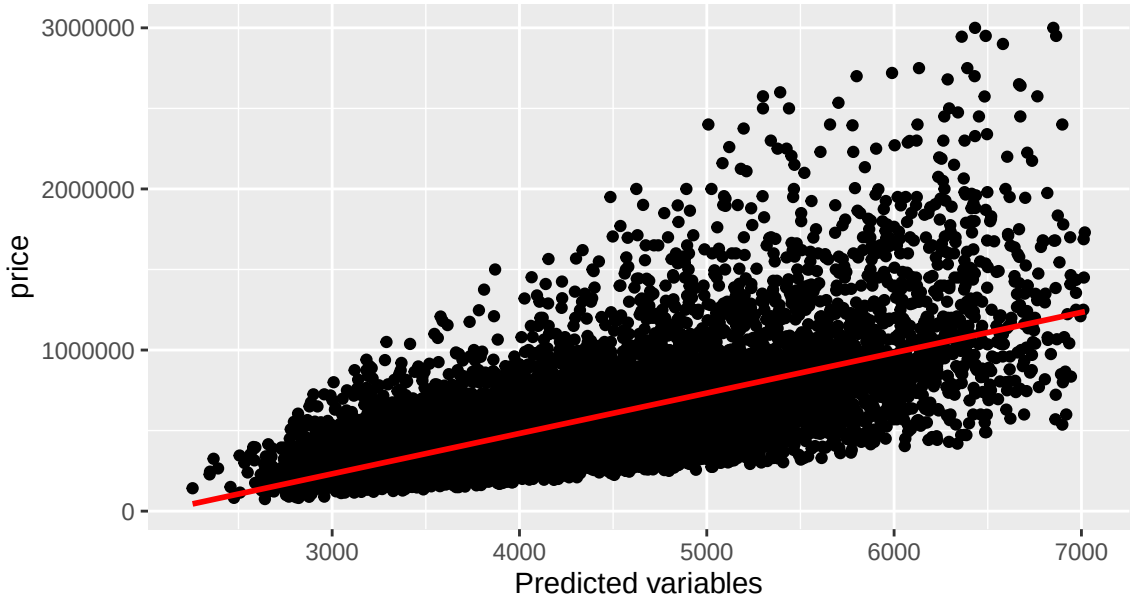
[9]* $x_{bedrooms}$ is the number of bedrooms in the house
* $x_{sqftliving}$ is a living area of the house in sqft unit measure.
* $x_{yearbuilt}$ is the year the house was built.
* $x_{condition}$ is a measure of how good condition the house is in from 0 to 5 scale
(i.e. $x_{condition} = 0$ if the house is in a terrible condition, and $x_{condition} = 5$ if the house is in perfect condition).

However, an increase in the number of `bedrooms`, and `yr built` have a negative effect on the housing prices. As the house has one more bedroom, the price of the house would likely to fall approximately $56137. Also, house A that is built 1 year earlier than house B is expected to have a higher price(approximately $613 more) than the relatively newly built house B.

Also, every variable has a very significant value at a 1% significance level, as seen in Table 3, p-value column. The p-values are rounded to 0 which indicates every p-value is significantly small and significant at a 1% significance level. Therefore, the regression model suggests that every factor has a significant linear relationship to the housing price.

Now, here is the visual image of how the model looks like:

Figure 4: Plot of the Multiple Linear Regression



As seen in Figure 4, The combined predictor variables have a positive effect on the housing price, even though number of bedrooms[10] and year built affect negatively on the house price.

[10]* Some potential bias on the variable is going to be discussed in the Weakness section.

Causal Inference Using Propensity Score Matching

The propensity score is the probability of treatment assignment conditional on observed baseline characteristics that are derived from the fitted regression model(1). The probability is constructed based on the values of independent variables before the treatment. One advantage of propensity score matching is that it allows us to easily consider many independent variables at once, and it can be constructed using logistic regression. In the analysis, the same variables used in the multiple linear regression model are used to construct the multiple logistic regression to find out the effect of the predictor variables on the renovation.

Housing price is the outcome of interest in the analysis, and year renovated is the Treatment variable. The Propensity Score matching will be for the year renovated propensity. We first construct a logistic regression model that explains whether a house is included in the treatment group(whether the house is renovated) as a function of other independent variables. After dividing them into two groups, for every house that is treated, there should be another variable that is included in the control group(not renovated) but has similar characteristics, based on the propensity score. The two variables are pair-matched one-to-one based on the characteristics from the logistic model, having similar values of the propensity score(34).

The logistic model is the following:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{bedrooms} + \beta_2 x_{sqft\ living} + \beta_3 x_{yr\ built} + \beta_4 x_{condition} \quad (4)$$

where $log(\frac{p_i}{1-p_i})$ represents log odds in each model, and p_i is the probability of an event occurring, which in this case is whether the house is renovated or not.

For the model diagnostics, logistic regressions are well performed under the following assumptions:

1. Linearity between the log odds and the predictor variables: Independent variables should be linearly related to the log odds

2. Binary response variable: Binary logistic regression requires the response variable to be binary

3. Multicollinearity among predictors is not too high. predictor variables should be independent of each other
4. Large sample size
   (Similar diagnostics is also introduced with detailed explanation in the Model Diagnostics section, so in this section, the process is not shown due to the redundancy.)

The dataset is reduced to 596 observations after matching the variables one-to-one explained above. Using this new dataset, a propensity score regression is created in the following table:

Looking at Table 5, the number of stars next to each coefficient estimates indicate how significant each value are, which is the p-value in different significance level depending on the number of stars. Every variable has a p-value smaller than 0.01 (at a 1% significance level), indicating that they are all significant predictor variables. Comparing to the original multiple linear regression summary in Table 4, every predictor variables have the same sign. Bedrooms and year built have a negative effect on the price of the house, and condition and sqft of living area have positive effects on the price. The treatment group has a value of 193578.54, which means when the house is renovated and everything else(every other predictor variables) stays constant, the renovation would cause the house price to increase by $193578.54.

Table 5: Propensity Score Regression Summary

	Model
(Intercept)	560325.532 **
	(216678.427)
bedrooms	-54993.961 ***
	(2525.826)
living area	286.160 ***
	(2.672)
year built	-286.146 **
	(106.904)
condition	28778.107 ***
	(3271.794)
treatment	193578.540 ***
	(12938.493)
N	16227
R2	0.491
logLik	-222345.766
AIC	444705.532

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Discussion

Summary

The goal of the analysis is to find the significant factors that affect the price of houses. Using the 2014-2015 housing price dataset from Kaggle, only quantitative data is used to show the accurate multiple linear regression. Through the data cleaning process in the Data Section, some data that are outliers or influential points that have a potential effect on the model are eliminated as well as unrealistic observations in the Data Section. Then, in the Model Section, to determine which variables to include in the regression, the AIC stepwise selection method is used and a linear regression model is built. After building a linear model, model diagnostics is done by showing the four assumptions of the regression - Linearity, Normality, Homogeneity of residuals variance, and Independence of the Error terms - using q-q plot, residual plot, and VIF models.

After the model selection, multiple linear regression and propensity score matching are used to determine the correlation as well as the causal effect of the significant factors on house prices. Multiple linear regression provides information about the correlation of the significant factors and the dependent variable housing price, and the Propensity score matching shows a causal relationship between the price of the house and the

observed variables.

Conclusions

The predictor variables that contribute to change in housing prices are the number of bedrooms, living area in sqft, year built, and condition of the house as shown in Table 4. Every predictor variable has a significantly small p-value under a 1% significance level which indicates that they are significant predictor variables. In the Result section, Figure 6 shows that there exists a linear relationship between the selected factors and the independent variable housing price. Among the predictor variables, the living area and condition of the house show a positive effect on the housing price, whereas the number of bedrooms and year built to show a negative effect on the price.

Condition of the house has the biggest effect on house price as we see in Table 4. To elaborate the finding of the coefficient estimates discussed in the Result section, among the four predictor variables, the condition variable has the biggest value, which means even though the house is a bit older, or the sqft of living area is smaller, as long as the house is in a good condition, it is predicted that the house could be sold at a higher price. For the negative effect on housing price, the bedroom variable has the biggest negative value, which theoretically means that as the house has more bedrooms, the price of the house would decrease. However, even though the number of the bedroom shows a negative coefficient for the regression estimate, by looking at the scatter plot in figure 5[11], it is hard to tell that the number of bedrooms has a significant negative effect on the housing price.

Moving on, the result from Propensity score matching, which could be found in Table 5, shows that renovation has a positive causal effect. This means it is more likely for the house to be priced higher in the real estate market when the house is renovated. To connect to real life, there are various TV shows where people renovate their homes and sell their houses with a jump of the house price. the original regression output determines the correlation and likelihood of the house price to increase or decrease, not the causal effect of it. Here, the condition of the house and sqft of a living area does not "cause" the house price to increase. The propensity score regression output, however, determines the causal effect of the variables. Renovation "causes" an increase in the house price. This analysis could be helpful for the people who are considering selling their house with a higher profit. According to the propensity score model which shows a causal link between renovation and the house price, renovated houses are going to be sold at a higher price.

To sum up, The Multiple Linear Regression analysis shows that houses that are in good condition are more likely to be sold at a higher price, and as the house size is bigger, the price increases. Through the causal inference using propensity score matching, we find a strong causal effect of house renovation on the housing price. That means, houses that are renovated cause the house price to increase significantly, and vice versa.

Weaknesses

Every data analysis contains some weaknesses. There are a few weaknesses the analysis includes. First, the dataset contains housing prices of a limited area which is King County, and the data is from 2014-2015. Since the data does not represent the Canadian housing price or the recent housing price, the analysis might not be the most accurate way to predict the housing prices.

Also, the AIC stepwise selection method does not consider interaction terms, there might exist some relationships between the independent variables. This might also lead to an omitted variable bias, where the omitted variables should be correlated with the dependent variable, and correlated with the explanatory variables included in the model. There might be an important variable that would affect the model, but it is hard to figure out since the variable might be missing in the dataset, or might be impossible to measure. Since the location of the house is also important, missing the location variable might have affected the model.

Another weakness is that the multiple linear regression model does not fully satisfy the model assumption introduced in the Model section. Multivariate Normality, which shows the normal distribution of residuals

[11]* Figure 5 can be found in the Appendix

of the regression is not satisfied. By looking at the QQ plot in Figure 3, it is noticeable that the data points do not trend the theoretical line, and the points at the upper tail of the data seem to jump and have higher values than the theoretical line, telling that the data might contain a gap in the values. This suggests that our model does not fully satisfy the normality assumption on the error terms. Therefore, we need to take into account that the result drawn from the regression model could be misleading or biased.

Next Steps

For the next steps, it would be a great idea to compare the housing prices after COVID-19 and do a causal inference pre & post analysis, to determine whether the COVID-19 affects the housing prices. In this analysis, the treatment variable would be houses that got affected by the COVID-19. Using the dataset that contains prices of condos, townhouses, and detached houses would lead to an interesting result.

Also, as addressed in the Weakness section, the location variable could be added with housing prices data of Toronto, or Canada, and analyze the regression model including the location. The variable might be a categorical variable, so it is important to treat it as a factor variable. It would make a big difference when the location variable is added since location affects the housing prices significantly as we can see in our real lives(urban/suburban, Toronto/other minor cities, etc).

It also would be a good idea to survey what people consider when buying a house. Which factor comes in first? We could compare the data of the survey and the regression model performed in the analysis, and determine whether the regression model actually fits into the real-life housing price prediction.

Reference

1. Propensity Score methods: Austin, Peter C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," May 2011. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.

2. Kableone function: "Bsurial/Bernr Source: R/kableone_nonnorm.R." Accessed December 22, 2020. https://rdrr.io/github/bsurial/bernr/src/R/kableone_nonnorm.R.

3. Side by Side plot: "Side-by-Side Plots with ggplot2," September 1, 1958. https://stackoverflow.com/questions/1249548/side-by-side-plots-with-ggplot2.

4. Regression Model assumption: Editor, Minitab Blog. "How to Identify the Most Important Predictor Variables in Regression Models." Accessed December 22, 2020. https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models.

5. Normal QQ plot: "How to Interpret a QQ Plot," July 1, 1963. https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot.

6. Model Diagnostics: Kassambara. "Linear Regression Assumptions and Diagnostics in R: Essentials," March 11, 2018. http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.

7. Exponential notation: Matt BannertMatt. "Force R Not to Use Exponential Notation (E.g. e+10)?," March 1, 1961. https://stackoverflow.com/questions/9397664/force-r-not-to-use-exponential-notation-e-g-e10.

8. AIC model selection: "AIC or p-Value: Which One to Choose for Model Selection?," May 1, 1960. https://stats.stackexchange.com/questions/9171/aic-or-p-value-which-one-to-choose-for-model-selection.

9. Baseline Characteristics Table: Rich, Benjamin. "Baseline Characteristics Table," November 25, 2020. https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html.

10. Data of House price prediction: Shree. "House Price Prediction," August 26, 2018. https://www.kaggle.com/shree1992/housedata.

11. Information about real estate: Steve Huebl. April 27, 2020. "COVID-19's Impact on Real Estate: Toronto Home Sales Down 69%," April 27, 2020. https://www.canadianmortgagetrends.com/2020/04/covid-19s-impact-on-real-estate-toronto-home-sales-down-69/.

12. ggplot: Susanejohnston. "A Quick and Easy Function to Plot Lm() Results with ggplot2 in R," April 23, 2015. https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-lm-results-in-r/.

13. Tableone R package: "Tableone r Package." Accessed December 22, 2020. http://rstudio-pubs-static.s3.amazonaws.com/13321_da314633db924dc78986a850813a50d5.html.

14. AIC selection: Tripathi, Ashutosh. "What Is StepAIC in R?," June 16, 2019. https://medium.com/@ashutosh.optimistic/what-is-stepaic-in-r-a65b71c9eeba.

15. AIC selection 2: "Variable Selection in Multiple Regression." Accessed December 22, 2020. https://www.jmp.com/en_in/statistics-knowledge-portal/what-is-multiple-regression/variable-selection.html.

16. Model diagnostics: Zach. "The Four Assumptions of Linear Regression," January 8, 2020. https://www.statology.org/linear-regression-assumptions/.

17. Model diagnostics: Zhang, Zhongheng. "Variable Selection with Stepwise and Best Subset Approaches," April 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4842399/.

18. Model Diagnostics: "Assumptions of Multiple Linear Regression," March 10, 2020. https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/.

19. Knitr Package: Yihui Xie. 2020. knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

20. kableExtra Package: Hao Zhu. kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1, 2020. https://CRAN.R-project.org/package=kableExtra

21. lme4 Package: Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. 2015. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

22. MASS Package: Venables, W. N. & Ripley, B. D., 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

23. tidyverse Package: Wickham et al., 2019. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

24. gridExtra Package: Baptiste Auguie, 2017. gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

25. dplyr Package: Hadley Wickham, Romain François, Lionel Henry and Kirill Müller. 2020. dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

26. caret Package: Max Kuhn, 2020. caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

27. tableone Package: Kazuki Yoshida and Alexander Bartel, 2020. tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights. R package version 0.12.0. https://CRAN.R-project.org/package=tableone

28. Matching Package: Jasjeet S. Sekhon. 2011. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. Journal of Statistical Software, 42(7), 1-52. URL http://www.jstatsoft.org/v42/i07/.

29. rbounds Pacakge: Luke J. Keele, 2014. rbounds: Perform Rosenbaum bounds sensitivity tests for matched and unmatched data.. R package version 2.1. https://CRAN.R-project.org/package=rbounds

30. broom Package: David Robinson, Alex Hayes and Simon Couch, 2020. broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.3. https://CRAN.R-project.org/package=broom

31. Omitted Variable bias: Christoph Hanck, Martin Arnold. Introduction to Econometrics with R. 15 Sept. 2020, www.econometrics-with-r.org/6-1-omitted-variable-bias.html.

32. Huxtable information: Hugh-Jones, David. Introduction to Huxtable, October 27, 2020. https://cran.r-project.org/web/packages/huxtable/vignettes/huxtable.html.

33. huxtable Package: avid Hugh-Jones, 2020. huxtable: Easily Create and Style Tables for LaTeX, HTML and Other Formats. R package version 5.1.1. https://CRAN.R-project.org/package=huxtable

34. Propensity score matching 2 & DID: "Difference in Differences." Telling Stories With Data, November 5, 2020. https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html.

Appendix

Table 6: Matched propensity score data

price	bedrooms	bathrooms	sqft_living	floors	condition	yr_built	yr_renovated	.fitted	cnts
439990	4	2.50	1540	2	3	2014	0	0.0001756	1
268000	4	3.00	1840	2	5	1988	2013	0.0001803	1
295500	3	2.50	1410	2	3	2014	0	0.0001851	1
649950	3	2.50	1500	2	3	2014	0	0.0002060	1
394950	2	1.00	1131	3	3	2011	0	0.0002201	1
239950	3	1.75	1600	1	3	2014	0	0.0002321	1

Figure 5: Scatterplot of Number of Bedrooms vs House Price