Analysis on Housing Market using Multiple Linear Regression

Yena Joo

December 2020

Code and data supporting this analysis is available at: https://github.com/yenajoo/304final.git

Abstract

To predict the movement of the real-estate market, 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(cite). 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, AIC stepwise selection method[1] is used, and Multiple linear regression model is used to determine the relationship between each predictor variables and prices. Through the analysis, we find ....
1. positive relationships between the independent variables ..... .
2. negative relationships between .....
2. no significant relationship
3. causal inference
Altogether, ...

Keywords

housing price, linear regression, treatment/control group, causal inference, housing price prediction, AIC selection method

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%. House is one of the most demanding ??? to people, since it is one of the most important components of people living. To have a better understanding of the real-estate market is the key of predicting the dwelling price and purchase an ideal house.

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 most correlation to the housing price, and analyze the causal effect using the difference in differences between renovated houses and non-renovated houses, one being assumed as a "treatment" group, where they are expected higher price by renovation. The other houses

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

would be in a "control" group, where the house have never been renovated. The difference is differences are measured by two different variables, control and treatment, and by first calculating the difference in first and second time periods, and then subtracting the average change in the control group from the average change in the treatment group(3).

we will use statistical methods to build a regression model of housing prices by potential factors, and interpret the regression output to find relationships between the houseing prices and potential factors.

Three housing price data will be used to investigate the relationship between housing prices and potential factors such as number of bedrooms, bathrooms, sqft, etc. In the Methodology section, I describe the data and the model that is used to analyze the relationship. Results of the Difference in differences are provided in the Results section.

Methodology

Data

- describe dataset

1. source of the data: Kaggle

- The methodology and approach that is used to collect and process the data.
- The population, the frame, and the sample.

The contents of 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 (cite). It consists of 21613 observations and 21 variables. However, through a simple data cleaning process, only 21604 observations are used, and 8 variables are included in the cleaned dataset.

should the target population be all housing in USA? The target population of the dataset includes all houses in the King County, USA. The frame population and sample population are whichever bought or sold within the county. There is not enough information regarding the dataset.

The reason for choosing the 2014-2015 data is because it shows the most stable housing prices which is before the pandemic.

These are the first 6 observations in the cleaned dataset:

Table 1: Raw data output

Price	bedrooms	bathrooms	living area(sqft)	floors	condition	year built	year renovated
221900	3	1.00	1180	1	3	1955	0
538000	3	2.25	2570	2	3	1951	1991
180000	2	1.00	770	1	3	1933	0
604000	4	3.00	1960	1	5	1965	0
510000	3	2.00	1680	1	3	1987	0
1225000	4	4.50	5420	1	3	2001	0

The dataset contains 8 variables. The included variables are price of the house, bedrooms, bathrooms, living area(sqft), floors, condition, year built, and year renovated.

After reducing the number of variables, another thing to consider is to eliminate the outliers. This process could be done though looking at the scatter plot of the data.

Figure 1: before adjusting

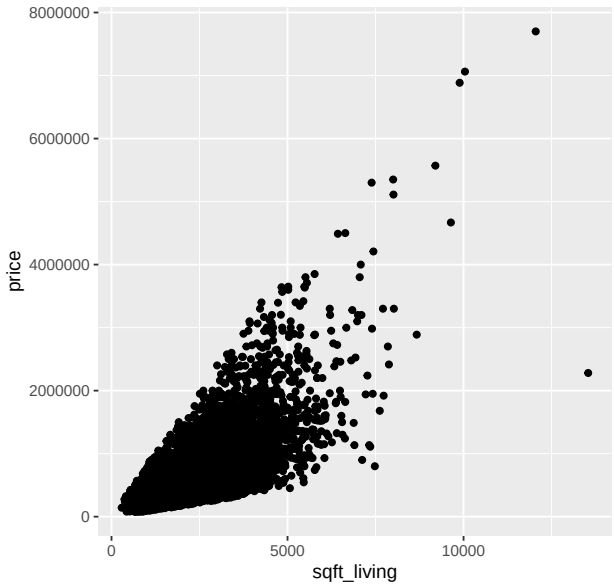
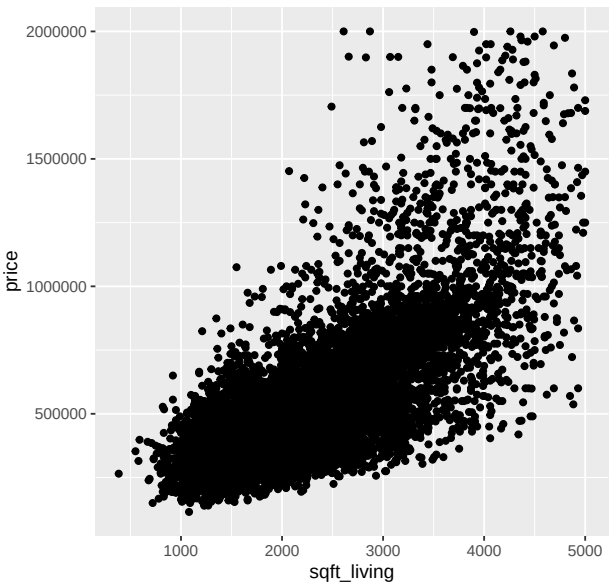


Figure 2: adjusted



As shown in Figure 1, The scatterplot looks too densely distributed from price range 0 to about 4,000,000. The observations that have price larger than $7,000,000 are omitted from the dataset through the data cleaning.
Also, houses with 0 bedrooms are ommitted since houses with 0 bathrooms might be an potential estimate error, as well as houses built before 1980, since old houses are more likely to be rebuilt, and we would like to predict the future housing prices.

- *What are its key features, strengths, and weaknesses about the study generally.* Some key features of the dataset is that every data is quantitative values, which is perfect for linear regression. If there are

add data addition for causal inference

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 number of bedrooms, number of bathrooms, living space in sqft, and an interaction term of bedrooms and living space in sqft. Interaction terms are used *interaction terms explained*

Model Selection

There could be a various potential factors affecting the 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 to the model.

AIC is a short form of Akaike Information criterion, which quantifies the amount of information loss due to the simplication. 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 compares AIC in each step and determines the model with the lowest AIC.

Formula for AIC is the following:

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

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

which starts with all predictors in the model (full model) along with the model without any predictor, iteratively removes the least contributive predictors or add a potential contributive predictors, and stops when you have a model where all predictors are statistically significant.

Table 2: AIC model summary

term	estimate	std.error	statistic	p.value
sqft_living	249.33892	3.473835	71.776276	0.0000000
bedrooms	-74268.77379	3398.040914	-21.856351	0.0000000
bathrooms	69143.39143	5203.575607	13.287669	0.0000000
floors	46078.84489	4507.377500	10.222983	0.0000000
yr_built	-65.76804	12.238782	-5.373741	0.0000001
condition	30318.20452	6082.663490	4.984363	0.0000006

explain the output
using this, our final equation is

The equation for the regression is:

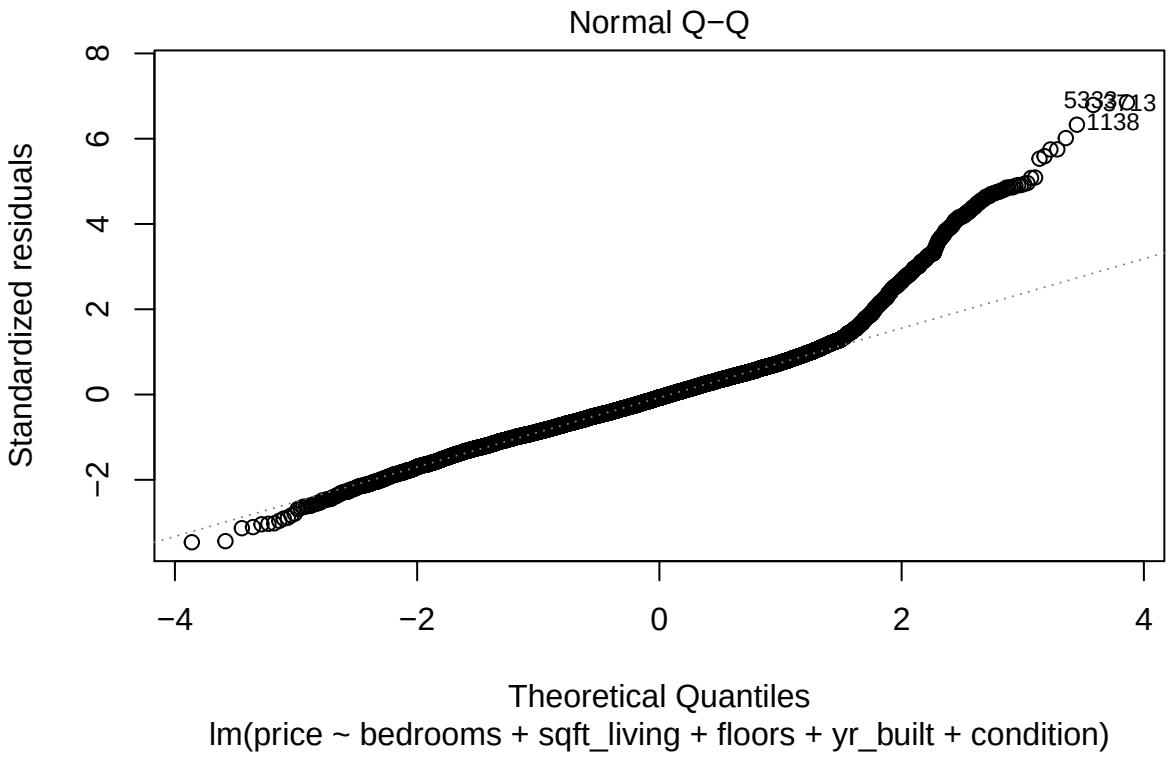
$$\text{Housing Price} = \hat{B_0} \ + \ \hat{B_1} * x_{bedrooms} \ + \ \hat{B_2} * x_{bathrooms} \ + \ \hat{B_3} * x_{sqftliving} \ + \ \hat{B_4} * x_{bedrooms} x_{sqftliving} \quad (2)$$

detailed description needed

Linearity assumptions

Multiple linear regression analysis are well performed under the following assumptions: 1. Linear relationship between the dependent and independent variables.
By showing the scatter plot above(Figure n), it satisfies the linearity assumption.

2. multivariate Normality: residuals of the regression should be normally distributed. This assumption may be checked by looking at a histogram or a Q-Q-Plot.



As we can see in the plot, the data points do not trend the theoretical line, and the points at each tail of the data seem to fall off the line, revealing that the distribution of residuals may have long tails.Normality is one of the assumptions in the linear regression model. However, the normal QQ plot above suggests that our model does not 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.

3. No multicollinearity: independent variables are not highly correlated with each otheer. This assumptioin could be checked by using Varince 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.

Table 3: VIF models

variables	VIF	variables	VIF
bedrooms	1.713619	bedrooms	1.598278
bathrooms	1.983134	living space(sqft)	1.603082
living space(sqft)	2.048710	floors	1.215371
floors	1.246917	year built	1.358351
year built	1.395044	condition	1.160340
condition	1.160381		

```
##  bedrooms bathrooms    floors  yr_built condition
##  1.440226  1.551770  1.245467  1.390289  1.160157
```

Left table in Table 3 is the initial model from the AIC selection method. Two variables, bathrooms and living space(sqft) have relatively high VIF which indicates that multicollinearity is a problem for the two variables. Therefore, variable bathroom is omitted from the model in the right table, and VIF values do not exceed 2 in the adjusted model. Hence, the second assumption is also satisfied.

4. Homoscedasticity: variance of error terms are similar accross 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. To have the assumption satisfied, there should be no clear pattern in the distribution.

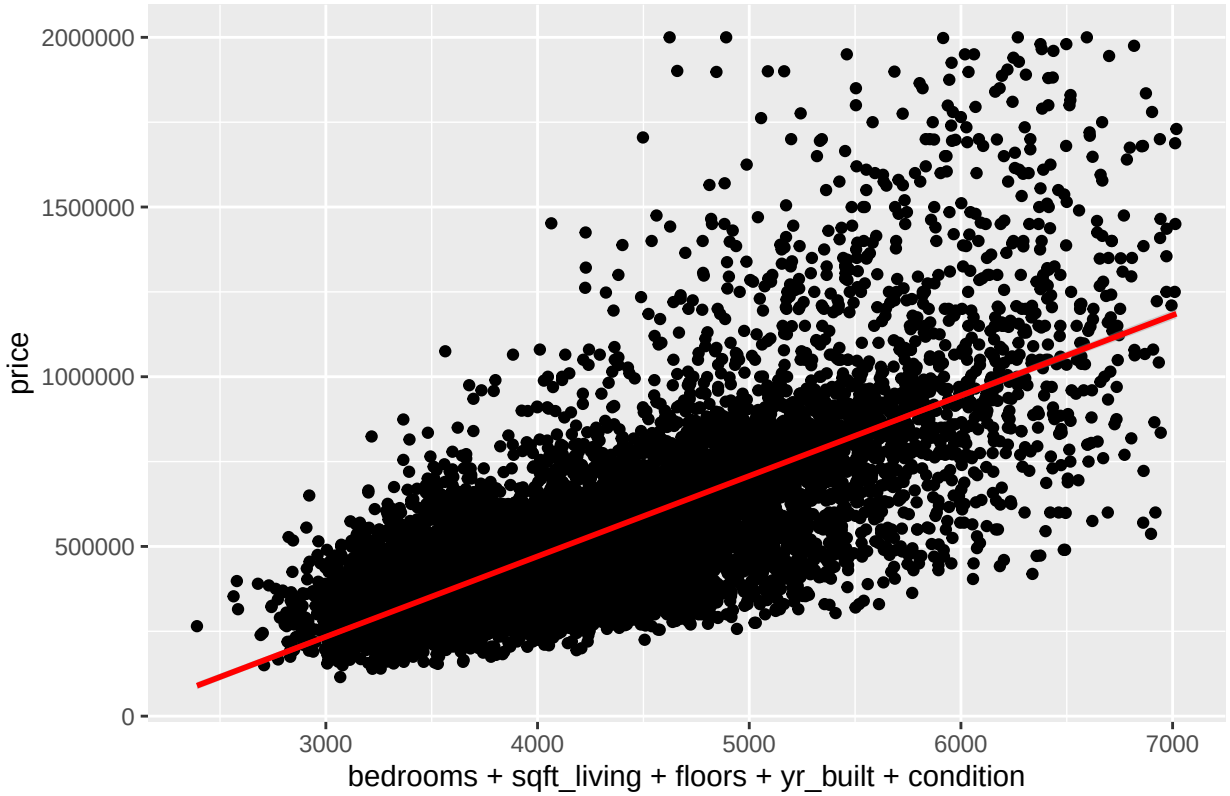
Results

Table 4: Regression summary

term	estimate	std.error	statistic	p.value
(Intercept)	-1507286.8082	506081.922990	-2.978346	0.0029060
bedrooms	-62769.1074	3313.117094	-18.945635	0.0000000
sqft_living	270.5899	3.106438	87.106179	0.0000000
floors	54370.9737	4745.746949	11.456779	0.0000000
yr_built	718.9319	251.560817	2.857885	0.0042747
condition	33173.0246	6580.261484	5.041293	0.0000005

```
##
## Call:
## lm(formula = price ~ bedrooms + sqft_living + floors + yr_built +
##     condition, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -676366 -121133  -13858   93415 1338916
##
## Coefficients:
##                 Estimate   Std. Error t value             Pr(>|t|)
## (Intercept) -1507286.808   506081.923  -2.978              0.00291 **
## bedrooms      -62769.107     3313.117 -18.946 < 0.0000000000000002 ***
## sqft_living      270.590        3.106  87.106 < 0.0000000000000002 ***
## floors         54370.974     4745.747  11.457 < 0.0000000000000002 ***
## yr_built         718.932      251.561   2.858              0.00427 **
## condition      33173.025     6580.261   5.041          0.000000472 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195400 on 8821 degrees of freedom
## Multiple R-squared:  0.5286, Adjusted R-squared:  0.5284
## F-statistic:  1979 on 5 and 8821 DF,  p-value: < 0.00000000000000022
```

Figure n: regression



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

```
model_practice <- lm(price ~ sqft_living + condition + bathrooms + yr_built + bedrooms + yr_built*bedroo
summary(lm(price ~ bedrooms + bathrooms + floors + yr_built + condition, data = housing_data))
```

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + floors + yr_built +
##     condition, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1449123  -171651   -35370   120381  1494482
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 1450338.6   637255.6   2.276              0.02288 *
## bedrooms      23148.4     3920.1   5.905        0.00000000366 ***
## bathrooms    244450.2     5859.3  41.720 < 0.0000000000000002 ***
## floors        32893.0     5988.1   5.493        0.00000004060 ***
## yr_built       -870.4      317.2  -2.744              0.00609 **
## condition     26169.5     8201.2   3.191              0.00142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 243600 on 8821 degrees of freedom
```

```
## Multiple R-squared:  0.2677, Adjusted R-squared:  0.2673
## F-statistic: 644.9 on 5 and 8821 DF,  p-value: < 0.00000000000000022
```

```
summary(model_practice)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + condition + bathrooms + yr_built +
##     bedrooms + yr_built * bedrooms + sqft_living * bedrooms +
##     condition * yr_built, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -806283 -120188  -16853   94470 1305126
##
## Coefficients:
##                            Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)          -34652065.692   7171191.871  -4.832          0.000001374
## sqft_living                129.877        10.614  12.237 < 0.0000000000000002
## condition              7968719.680   2262738.746   3.522             0.000431
## bathrooms                74071.836      5182.901  14.292 < 0.0000000000000002
## yr_built                 17508.752      3604.552   4.857          0.000001210
## bedrooms               2632834.617    546467.412   4.818          0.000001475
## yr_built:bedrooms        -1394.500       273.197  -5.104          0.000000339
## sqft_living:bedrooms        34.121         2.848  11.979 < 0.0000000000000002
## condition:yr_built       -3993.861      1138.197  -3.509             0.000452
##
## (Intercept)          ***
## sqft_living          ***
## condition            ***
## bathrooms            ***
## yr_built             ***
## bedrooms             ***
## yr_built:bedrooms    ***
## sqft_living:bedrooms ***
## condition:yr_built   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 192600 on 8818 degrees of freedom
## Multiple R-squared:  0.5424, Adjusted R-squared:  0.542
## F-statistic:  1306 on 8 and 8818 DF,  p-value: < 0.00000000000000022
```

```
summary(lm(price ~ bedrooms, data = housing_data))
```

```
##
## Call:
## lm(formula = price ~ bedrooms, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -744379 -192144  -49017  116669 1611228
##
```

```
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   156027      12766   12.22 <0.0000000000000002 ***
## bedrooms      115372       3613   31.93 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269400 on 8825 degrees of freedom
## Multiple R-squared:  0.1036, Adjusted R-squared:  0.1035
## F-statistic:  1020 on 1 and 8825 DF,  p-value: < 0.00000000000000022
```

$$\text{Housing Price} = \hat{B_0} + \hat{B_1} * x_{bedrooms} + \hat{B_2} * x_{sqftliving} + \hat{B_3} * x_{floors} + \hat{B_4} * x_{yearbuilt} + \hat{B_4} * x_{condition} \quad (3)$$

.

In Table 4, as observed from the output, the price of the house increases as the number of living space(in sqft), number of floors, and condition increase.
However, increase in number of bedrooms, and year built have negative effect to the housing prices.

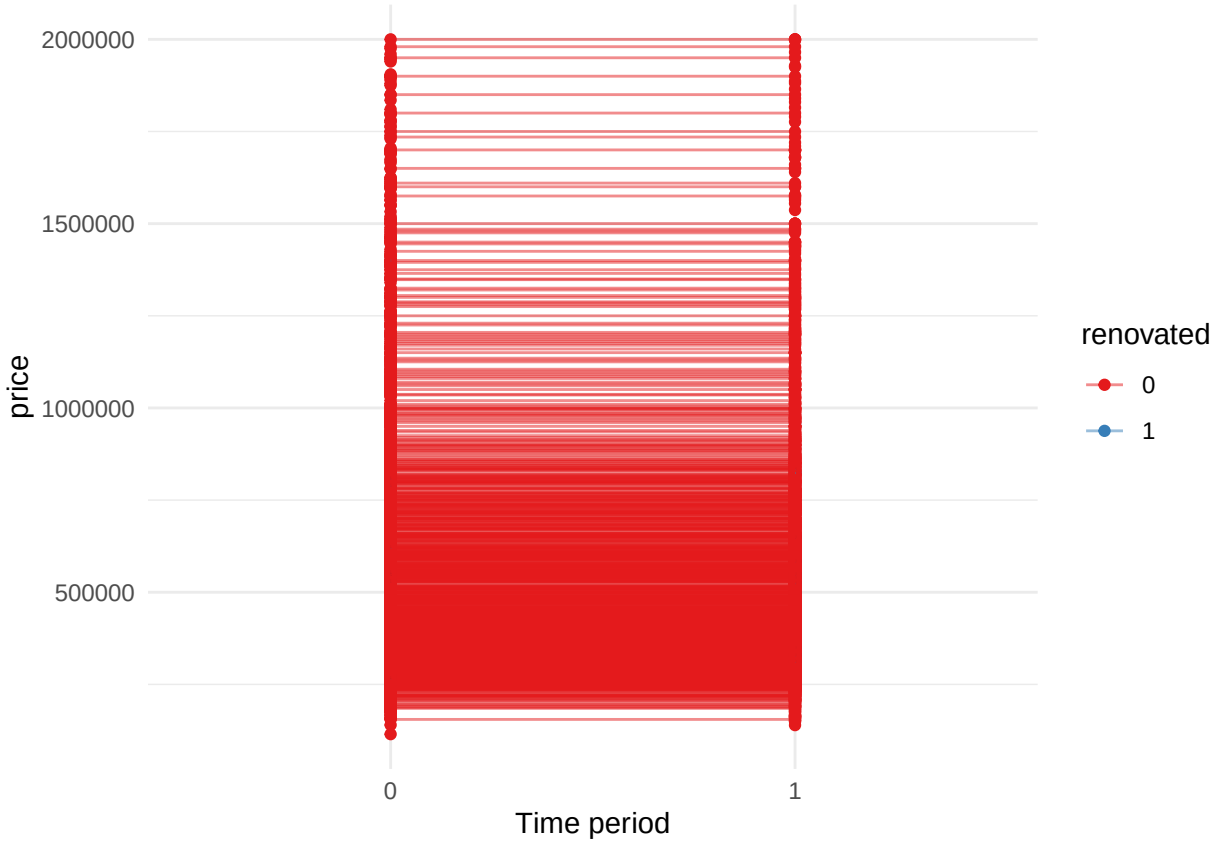
As seen in Figure n, The combined predictor variables have positive effect to the housing price.

what increases
pvalues
Every variable has a very significant value at 1% significance level, as seen in Table 3, p-value column. Therefore, the regression model suggests that every factor has a linear relationship to the housing price.

Causal inference

if yr_renovated == 0, treatment = 0

```
##
##    0    1
## 8805   22
```



Discussion

Summary

The goal of the analysis is to find the significant factors that affect price of houses. Using the housing price dataset from Kaggle, only quantitative data is used to show the accurate multiple linear regression. To determine which variables to include in the regression, AIC stepwise selection method is used. The predictor variables that contribute to change in housing prices are number of bedrooms, bathrooms, ...

multiple linear regression & analyzed as well as causal inference using difference-in-differences.
For the results, we got **result**

Conclusions

- Here is where you explain what the results really mean, and identify any relevant findings.
- Make sure to touch on global impacts. For example,

Weaknesses

Every data analysis sontains 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 selction 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 our data set, 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 of the regression is not satisfied. By looking at the QQ plot, it is noticable 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 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.
r-squared around 50%. is it ok?

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. Also, as addressed in the Weakness section, location variable could be added with housing prices data of Toronto, or Canada, and analyze the regression model including the location. It would make a big difference, since location affects the housing prices significantly.

Reference

1. linear regression assumptions: https://www.statology.org/linear-regression-assumptions/
2. ggplot regression: https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-lm-results-in-r/
3. exponential notation: https://stackoverflow.com/questions/9397664/force-r-not-to-use-exponential-notation-e-g-e10
4. ggplot side by side: https://stackoverflow.com/questions/1249548/side-by-side-plots-with-ggplot2
5. Kaggle https://www.kaggle.com/shree1992/housedata
6. Toronto home Sales: https://www.canadianmortgagetrends.com/2020/04/covid-19s-impact-on-real-estate-toronto-home-sales-down-69/
7. r-squared?: https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models
8. Forward selection: https://www.jmp.com/en_in/statistics-knowledge-portal/what-is-multiple-regression/variable-selection.html https://medium.com/@ashutosh.optimistic/what-is-stepaic-in-r-a65b71c9eeba
9. AIC stepwise selection: https://stats.stackexchange.com/questions/9171/aic-or-p-value-which-one-to-choose-for-model-selection
10. AIC anova: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4842399/
11. Assumption for multiple linear regression: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1111&context=pare https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/

12. normal qq plot interpretation: https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot

don't forget to cite packages