Analysis on housing prices

Yena Joo

December 2020

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

Abstract

- Here you are provided a brief summary of the entire report.
- This is generally written as one long paragraph.
- what was done, what was found, and why this matters (all at a high level).

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 in order to obtain a smooth analysis. After selecting significant variables using AIC stepwise selection method[1], Multiple linear regression model is used to determine the significant factors that affect the dwelling prices. Through the analysis, we find . . . .
1. positive/negative relationships between the independent variables . . . .
2. no significant relationship
3. causal inference
Altogether, ...

Keywords

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

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

Table 1: Raw data output

Price	bedrooms	bathrooms	living area(sqft)	floors	condition	year built	year renovated
313000	3	1.50	1340	1.5	3	1955	2005
2384000	5	2.50	3650	2.0	5	1921	0
342000	3	2.00	1930	1.0	4	1966	0
420000	3	2.25	2000	1.0	4	1963	0
550000	4	2.50	1940	1.0	4	1976	1992
490000	2	1.00	880	1.0	3	1938	1994

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. It includes homes sold between May 2014 and May 2015 (Kaggle). It consists of 21613 data and 21 variables.

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.

Figure 1: before adjusting

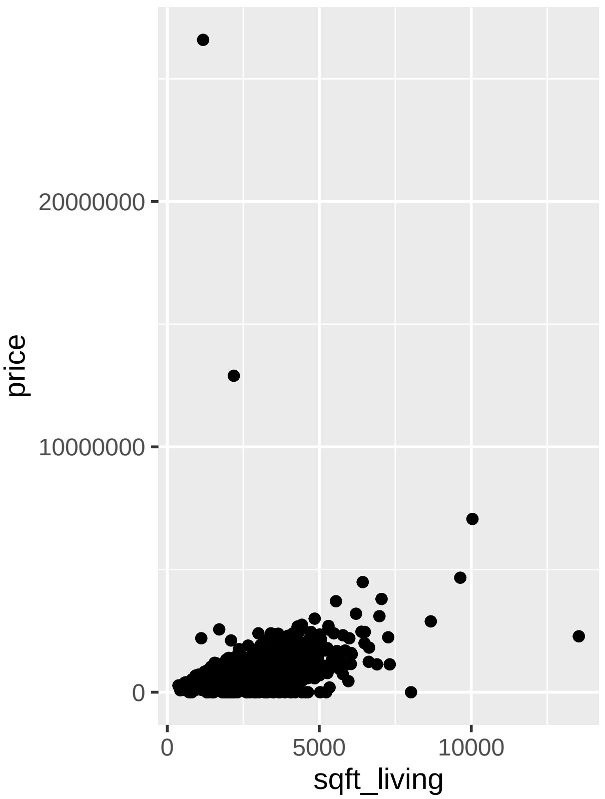
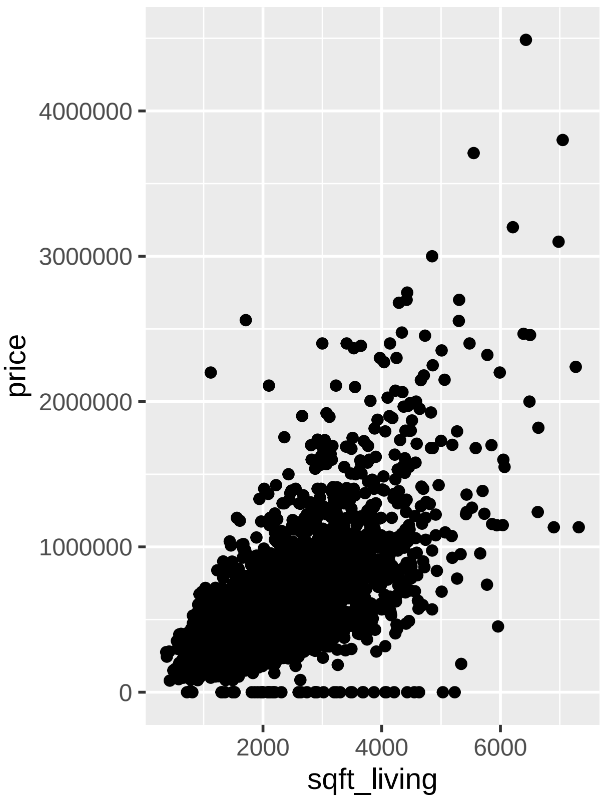


Figure 2: adjusted



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:

AIC $= -2(log - likelihood) + 2k$
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.

```
## Start:  AIC=122866.9
## price ~ -1
## 
##               Df        Sum of Sq              RSS    AIC
## + sqft_living  1 1588120582635746  316775609769700 114629
## + bathrooms    1 1477906282451794  426989909953652 116000
## + bedrooms     1 1393534534072700  511361658332746 116829
## + yr_built     1 1342423985859548  562472206545898 117266
## + floors       1 1339277096884352  565619095521094 117292
## + condition    1 1310039508227835  594856684177611 117523
## <none>                            1904896192405446 122867
## 
## Step:  AIC=114629.2
## price ~ sqft_living - 1
## 
##               Df        Sum of Sq              RSS    AIC
## + bedrooms     1    1856212286170  314919397483530 114604
## + condition    1    1314309909352  315461299860348 114612
## + floors       1     538846644232  316236763125468 114623
## <none>                             316775609769700 114629
## + yr_built     1       3483017808  316772126751892 114631
## + bathrooms    1         58802760  316775550966939 114631
## - sqft_living  1 1588120582635746 1904896192405446 122867
## 
## Step:  AIC=114604.2
## price ~ sqft_living + bedrooms - 1
## 
##               Df       Sum of Sq             RSS    AIC
## + condition    1   9003625193664 305915772289865 114473
## + yr_built     1   2858355315610 312061042167920 114564
## + floors       1   2450895518904 312468501964626 114570
## + bathrooms    1    630725517263 314288671966266 114597
## <none>                           314919397483530 114604
## - bedrooms     1   1856212286170 316775609769700 114629
## - sqft_living  1 196442260849217 511361658332746 116829
## 
## Step:  AIC=114472.9
## price ~ sqft_living + bedrooms + condition - 1
## 
##               Df       Sum of Sq             RSS    AIC
## + yr_built     1   1030158737441 304885613552424 114459
## + floors       1    668241536879 305247530752986 114465
## + bathrooms    1    134105890888 305781666398977 114473
## <none>                           305915772289865 114473
## - condition    1   9003625193664 314919397483530 114604
## - bedrooms     1   9545527570483 315461299860348 114612
## - sqft_living  1 199722419713699 505638192003564 116779
## 
## Step:  AIC=114459.4
## price ~ sqft_living + bedrooms + condition + yr_built - 1
## 
##               Df       Sum of Sq             RSS    AIC
## + floors       1   2574105258130 302311508294294 114423
## + bathrooms    1    492615695184 304392997857240 114454
```

Table 2: AIC model summary

term	estimate	std.error	statistic	p.value
sqft_living	273.87472	5.363927	51.058620	0
bedrooms	-48231.38619	5228.961695	-9.223894	0
condition	66720.36006	5672.818510	11.761413	0
yr_built	-93.99171	14.080741	-6.675196	0
floors	49595.75168	7935.000554	6.250252	0

```
## <none>                           304885613552424 114459
## - yr_built     1   1030158737441 305915772289865 114473
## - bedrooms     1   5872526865161 310758140417585 114545
## - condition    1   7175428615495 312061042167920 114564
## - sqft_living  1 200512091327678 505397704880103 116779
##
## Step:  AIC=114422.5
## price ~ sqft_living + bedrooms + condition + yr_built + floors -
##     1
##
##               Df       Sum of Sq             RSS    AIC
## <none>                           302311508294294 114423
## + bathrooms    1     24691033493 302286817260801 114424
## - floors       1   2574105258130 304885613552424 114459
## - yr_built     1   2936022458692 305247530752986 114465
## - bedrooms     1   5606087058089 307917595352383 114505
## - condition    1   9114865291940 311426373586234 114557
## - sqft_living  1 171778741717609 474090250011903 116487
```

explain the output
using this, our final equation is

The equation for the regression is:

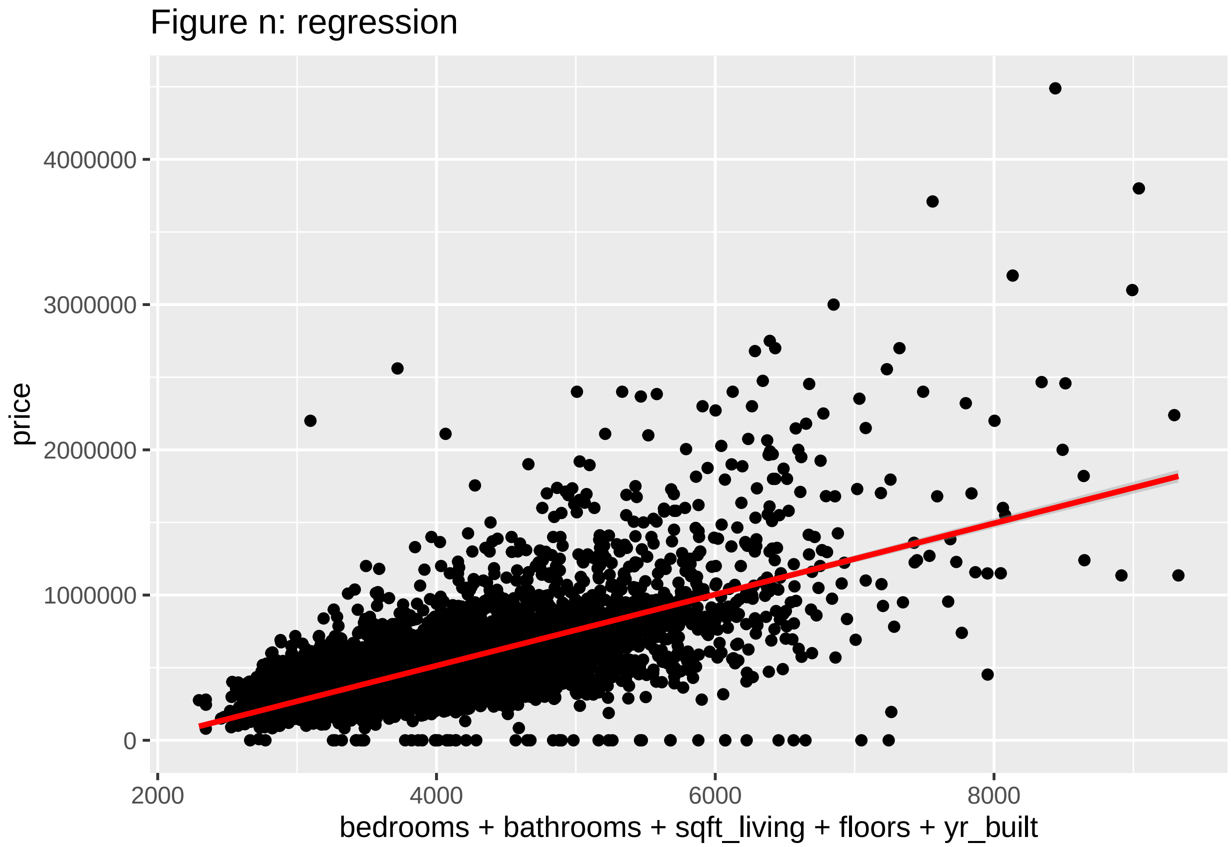
housing price $= \hat{B}_0 \; + \; \hat{B}_1 * x_{bedrooms} \; + \; \hat{B}_2 * x_{bathrooms} \; + \; \hat{B}_3 * x_{sqftliving} \; + \; \hat{B}_4 * x_{bedrooms} x_{sqftliving}$
detailed description needed

Linearity assumptions

Results

Table 3: Regression summary

term	estimate	std.error	statistic	p.value
(Intercept)	4991858.4898	298477.272256	16.724417	0.0000000
sqft_living	288.2733	5.278752	54.610127	0.0000000
bedrooms	-51816.3524	5081.551613	-10.196955	0.0000000
condition	26965.3669	5999.030480	4.494954	0.0000071
yr_built	-2596.5394	150.257639	-17.280582	0.0000000
floors	89025.2566	8057.089179	11.049308	0.0000000



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

housing price = $\hat{B}_0 \ + \ \hat{B}_1 * x_{bedrooms} \ + \ \hat{B}_2 * x_{bathrooms} \ + \ \hat{B}_3 * x_{sqftliving} \ + \ \hat{B}_4 * x_{bedrooms} x_{sqftliving}$.

As observed from the output, the price of the house increases as the number of bathrooms, living space(in sqft), number of floors increase.

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

```r
set.seed(304)
#make variables for control/treatment/time group
housing_data <- housing_data %>%
  mutate(treatment_group = case_when(
    yr_renovated != 0 ~ 1,
    yr_renovated == 0 ~ 0 ))
housing_data <- housing_data %>% mutate(time = sample(0:1, length(price), replace=T))
table(housing_data$treatment_group)
```

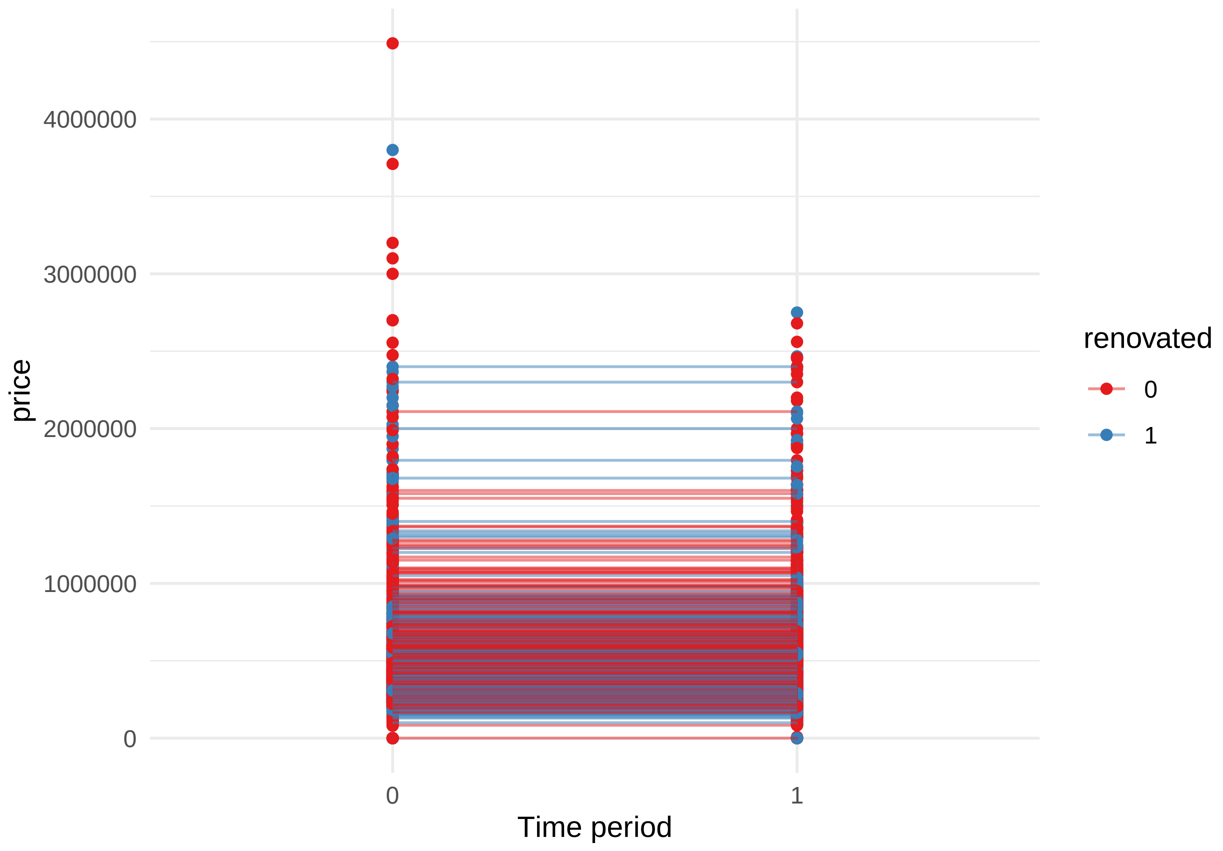
```
##
##    0    1
## 2732 1861
```

```r
#treatmenttime = 1861/2
#housing_data <-
#   for (i in 1:length(housing_data$time)) {
#     if(housing_data$yr_renovated[i] == 1){
#       mutate(housing_data, i = sample(0:1, 1, replace = T))
#     }
 #   else{
 #    mutate(housing_data, i = sample(0:1, 1, replace = T))
 #   }}
```

```r
# visualize
housing_data$treatment_group <-
  as.factor(housing_data$treatment_group)

housing_data$time <-
  as.factor(housing_data$time)

housing_data %>%
  ggplot(aes(x = time,
             y = price,
             color = treatment_group)) +
  geom_point() +
  geom_line(aes(group = price), alpha = 0.5) +
  theme_minimal() +
  labs(x = "Time period",
       y = "price",
       color = "renovated") +
  scale_color_brewer(palette = "Set1")
```



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

- This sub-section can be split into two, if needed
- Be careful here, especially if you are simulating data. You can never simulate every single scenario. So you will have some generalizability issues. - Addressing weaknesses of the analysis.

- Addressing future steps of the analysis. o Hint: a good future step might be to compare with the actual election results and do a post-hoc analysis (or at least a survey) of how to better improve estimation in future elections.

Next Steps

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/