James Taylor

Assignment 2

University of Maryland University College

DATA610-9040 Summer 2019

Dr. Saleeb

james.taylor.ps3@gmail.com

Introduction

The data set for this assignment is a sample of home sales around Seattle, Washington from May until July in 2014. There are 4,600 observations of a sale. Included in each observation are basic attributes of the home like the number of bathrooms and bedrooms, number of floors, square footage, price, date of sale, year built, year renovated and location information. Each observation also has arbitrary rating on the home like the quality view and the condition of the home. The table below displays a sample of the data set, not all variables are displayed.

price	bedrooms	bathroom	sqft_living	sqft_lot	floors	waterfron	view	condition	sqft_abov
2,384,000	5	2.5	3650	9050	2	0	4	5	3370
1,400,000	4	2.5	2920	4000	1.5	0	0	5	1910
1,200,000	5	2.75	2910	9480	1.5	0	0	3	2910
838,000	4	2.5	3310	42998	2	0	0	3	3310
805,000	3	2	2710	4500	1.5	0	0	4	1880
790,000	3	2.5	2600	4750	1	0	0	4	1700
785,000	5	3.25	3660	11995	2	0	2	3	3660
750,000	3	1.75	2240	10578	2	0	0	5	1550
750,000	3	2.5	2390	6550	1	0	2	4	1440

Data Exploration

The first necessary step to looking at data is to determine what is there, what is missing and what is inaccurate. What the dataset contains is mentioned above. There are 49 observations where a price is missing. A significant proportion of observations have their basement square footage listed as zero, but it is reasonable to assume it is because it has no basement. The same is true for the year of renovation, if there was no renovation it wouldn't be listed. There is concern for missing view discretionary variable. The view variable is ordinal, listing a number from zero up to four. If a home is rated zero, either the variable was never evaluated or the home has the lowest level for that variable. The worry is if the home was not evaluated for its view, it becomes a zero because it is the default value if no other was

collected. To avoid this issue the view and condition variables won't be in calculations. Another issue I saw was there are observation where the house was renovated before it was built. As an example, a home in Snoqualmie was renovated was built in 2012 but was renovated in 1912. A house can't be renovated before it was even built.

Price appears to have inaccurate values. The larges price value in the dataset is $26 million dollars with three bedrooms, two bathrooms on less than 8000 square-feet of land (See *Figure 1*. Outlier House). Another house price seems too high, $13 million for three bedrooms and 2.5 baths. I discovered the sale price of this home on a real estate sight for $1,899,000 but the price in the data set is $12,899,000 (Zillow, 2019). It appears the is an unintentional '2' in the price possibly from a typing mistake. I looked at the outliers for the other variables and there was not strong evidence to exclude them those observations. One observation was the purchase of multiple homes at once, which I would consider unfit for this dataset.

I was immediately curious about how a renovation could affect the condition evaluation of the house. I'm also interested in the value of a house given the condition. This would be useful information for professionals who flip houses, or those who plan to invest in their home to get a better return when they sell. I'm interested in how people evaluate a house. The only variables that do that is price and condition of the house.

The condition is rated as '3' in 62.5% of the observation. The average price of this condition is $550,111. Comparing the average price of a house with a '1' condition to that of a '5', the price more than doubles from $306,633 to $637,041. The average home with a '4' condition is about $13,000 less than the average for '3'. So, there is a relationship between condition and price but not very strong.

Data Visualizations

I want to explore the relationship between when a home was remodeled relative to when it was built, and how that effects the condition and price. The first step is to see if condition is a good predictor of house price, which *Figure 2*-Average Price per Condition, shows the average price of a house for each condition. You can see there is a positive correlation but not very strong.

How does the home's age relate its price? This is displayed in *Figure 3*-Average Price per Year Built, showing old homes and new homes are the most expensive. It is the homes aged in the middle-third of this dataset that are least expensive, and by deduction least desired. Basic economics would say the supply for these old-style homes is low compared to the demand. Old homes may have desirable locations. Old homes would be closer to the metropolitan centers because they were built before all the recent growth and urban sprawl. The land, not the actual home, could explain the value of the purchase. There could be another explanation.

A renovation to an old home could make it as good as a modern one. This could explain why old and new home prices are comparable. *Figure 4*-Year Built and Year Renovated by Condition is displaying the relationship between year built, year renovated and the condition of that home. The first thing to note is that the four points in the bottom right quadrant of the graph is the inaccurate data where a house was renovated before it was built. It appears that homes renovated before 1992 have a significantly higher condition than after it. The homes renovated in 1992 or earlier average a 3.7 condition rating, where houses after have 3.15 average. Data points in the top-left corner are devoid of any homes above a condition 3.

Homes built early and renovated recently are not in exceptional condition. This may be because very old homes are too worn-down that even a new remodel can't improve the state of the house. The section with the highest average are homes built before 1956 and renovated before 1990.

This is an unexpected result. Homes that are older and renovated more than 20 years ago are rated better than new homes. Wouldn't new homes and recently renovated homes be in better condition? *Figure 5*-Average Price and Average Condition per Year Built, displays the average condition and average price by year. The conditions of homes built in 1970 and later show a decreasing trend. This decrease coincides with raising average home prices, meaning people are paying more for new home with lower conditions.

I could see expectation bias being a factor here. When an old home is rated for it condition, the person rating it may be surprised at how good it looks given its age. This would cause them to rate it higher. Contrarily, someone rating a new home would expect it to look new, making minor issues standout worse than they should. In other words, the rater's expectations bias their evaluation because they expect old home to look old and vice versa for new ones.

There may not be this bias however. The cities vary in their average prices and condition. *Table 6*-Average Price and Average Condition per City, is a sample of the cities in the dataset, showing the variety in price and condition. House sales for this data set may not be evenly distributed across all the cities. For example, a city may have had a building project where lots of new houses were built. All these new homes would flood the housing market, bringing housing prices down because of the increased supply.

Calculation

If I were moving to this are, I would want to know what areas I would get the best condition home for the price. I calculated the dollars per condition average for each city in the data set. In other words, how much money do I have to spend in each city for each condition I get. In *Figure 7*-Price per Condition by City, which displays the calculation, shows that for Woodinville, I would have to spend $185K for each level of condition for a house. This would mean I would have to spend $370K on average to get a house with a condition of '2'. For this calculation, I left the discovered outliers in because it proves that they are there. For instance, in Medina I would spend on average $1.2 million to get a home with a 2 condition. This should call into question what houses sold in Medina that would cause such a difference.

Some cities may have stricter building codes. This would prevent the building of homes, keeping supply down thus inflating prices in this area. There has been a significant amount of academic work that supports the idea that the Housing Bubble was caused by these building codes (Sowell, 2009). There wasn't enough housing to support the population increases in cities because large apartment complexes wouldn't be approved. Many of these laws are still on the books today even after the bubble burst.

References

Zillow, Inc. (2019). 5426 40th Ave W Seattle, WA 98199. Retrieved from https://www.zillow.com/homedetails/5426-40th-Ave-W-Seattle-WA-98199/48677285_zpid/

Sowell, T. (2009). Thomas Sowell: regulators started housing crisis. Newsmax. Retrieved from https://www.newsmax.com/newsfront/sowell-housing-crisis/2009/05/17/id/330167/

Google Maps. (2019). 12005 SE 219th Ct Kent, WA. Retrieved from https://www.google.com/maps/place/12005+SE+219th+Ct,+Kent,+WA+98031/@47.4054595,-122.1809714,3a,90y,229.39h,83.08t/data=!3m6!1e1!3m4!1sVZlEQn-ajbkEvDEuDH5lIg!2e0!7i16384!8i8192!4m5!3m4!1s0x54905c2cb21ecd45:0xf24bf45e3c587184!8m2!3d47.4052813!4d-122.1810786

Appendix A

Figure 1. Outlier House (Google Maps, 2019)

This how was listed as $26 million with 3 bedrooms and 2 bathrooms, clearly a mistake.

Figure 2-Average Price per Condition

This displays the average price for each condition of the homes.

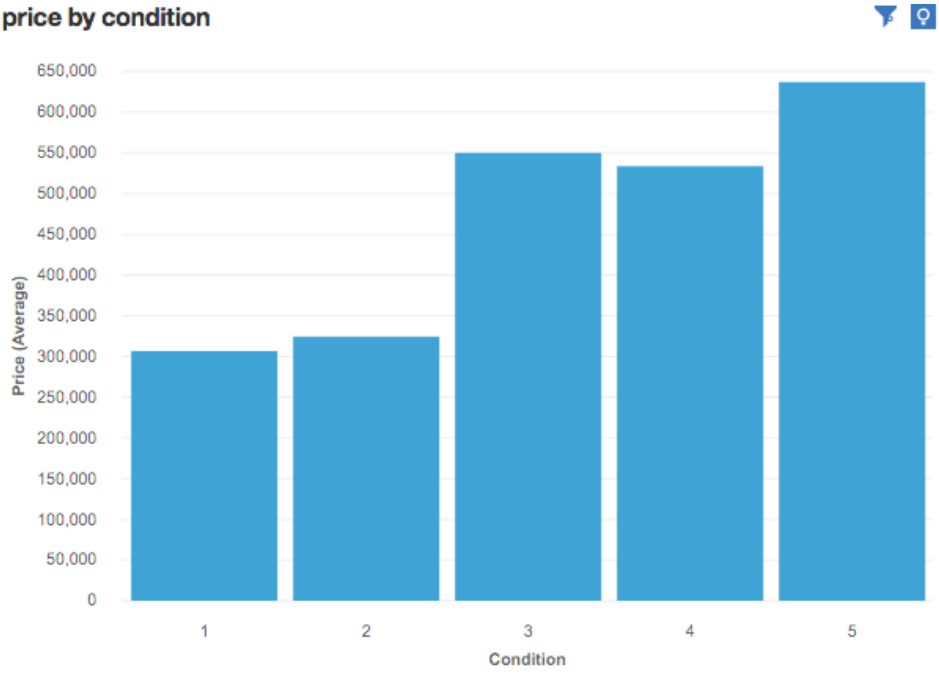

Figure 3-Average Price per Year Built

The years are grouped to reduce some variability. There is no strong reason to believe a home built in 1952 is very different than one built in 1951.

Figure 4-Year Built and Year Renovated by Condition

The darker the point, the higher the condition.

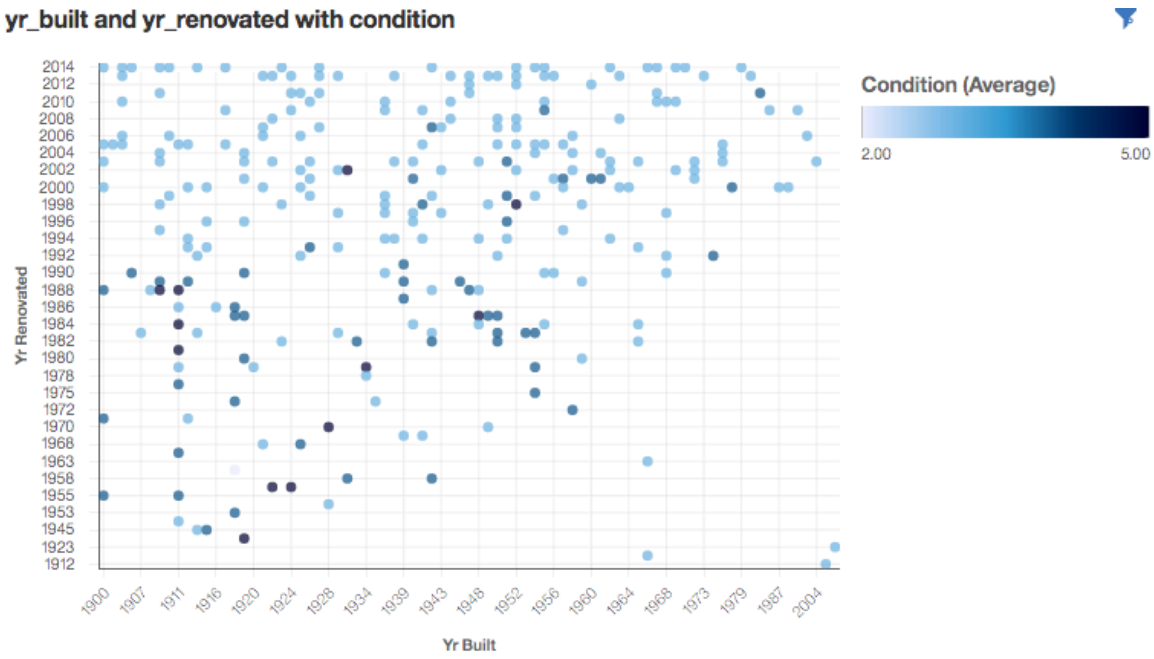

Figure 5-Average Price and Average Condition per Year Built

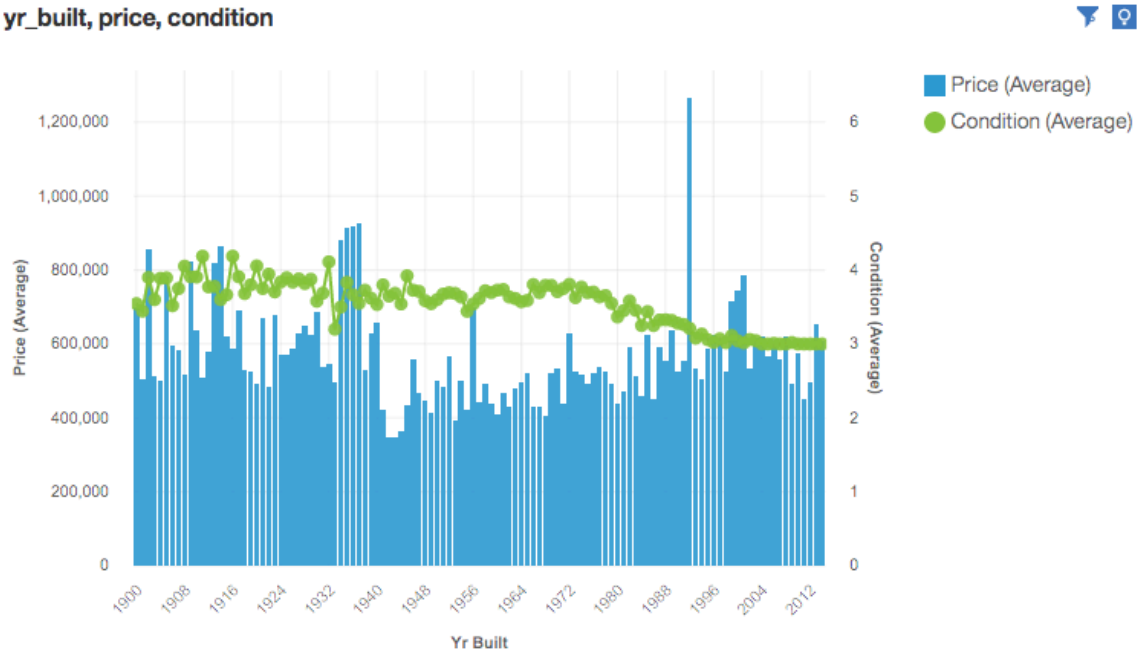

Table 6-Average Price and Average Condition per City

city, condition, price

City	Condition	Price ▾
Clyde Hill	3.55	1,321,945
Mercer Island	3.79	1,123,818
Beaux Arts Village	4	745,000
Kirkland	3.49	651,584
Woodinville	3.38	609,565
Issaquah	3.31	596,164
Preston	4.5	562,450
Bothell	3.15	481,442
Shoreline	3.64	420,392
Renton	3.43	377,041
Des Moines	3.48	304,993
Federal Way	3.43	289,888
Algona	3.2	207,288

Figure 7-Price per Condition by City