House Price Analysis (Team 2)

(Homework Assignment 1)

Douglas Reed, Jason Tompkins,

Joe Shaxted, Michael Farrell, Timothy Hulak

SCM 651 (Business Analytics)

02/04/2021

House Price Analysis

Initial Analysis of Raw Data

Normally, analysts will look at data to answer questions regarding a business need. this report attempts to categorize, visualize, correlate, and analyze the data to determine best purchase price for homes in each of three neighborhoods.

The house price data is made up of 128 observations of eight variables effecting the price of a house. In addition to a serialized, unique identifier for each listing, there are five quantitative (numeric) values and two categorical or qualitative values. Our team made the following clarifications and assumptions regarding the data:

Quantitative Values:
Price – the price of home in US dollars. As the data also contained offers received, the team assumed that this was the final sale price for the home.
SqFt – the square footage of the interior of the home.
Bedrooms – the total number of bedrooms within the home.
Bathrooms – the total number of bathrooms within the home.
Offers – the number of offers received from prospective buyers before the sale of the home.

Qualitative values:
Brick – the data provides a YES/NO value referring to whether the home exterior was brick.
Neighborhood – the location of the home within three designated quadrants of the city (East, West, and North)

Categorizing the Data:

The data was initially broken out into counts by neighborhood to determine if there were any patterns. The table below aggregates the data to begin the comparison.

Counts:	128	49 / 38%	79 / 62%	387	313	330
	Homes	**Brick**	**No Brick**	**Beds**	**Baths**	**Offers**
East	45 / 35.2%	19 / 42%	26 / 58%	132	110	115
North	44 / 34.3%	7 / 16%	37 / 84%	117	97	135
West	39 / 30.5%	23 / 59%	16 / 41%	138	101	80

Though the comparison of counts does not provide much detail, a pattern does begin to appear. As shown above, approximately one third of the homes are in each neighborhood. Additionally, while it is overall more likely that homes do not have a brick exterior (only 38% do), there is a large disparity among the neighborhoods with only 16% of the homes in the North of the city were made of brick compared to 58% in the East. The other data points provided little initial context, though the counts proved useful later during the analysis.

Next, the data was refined based on the two categorical data points in our data set. The qualitative variables of neighborhood and brick provided additional context to determine if there are correlations as seen in the tables below.

The Average Price and Average Square Footage Pivot Tables

Average Price	Brick	
Neighborhood	No	Yes
East	$117,750	$135,468
North	$108,584	$118,457
West	$148,230	$175,200

Average of SqFt	Brick	
Neighborhood	No	Yes
East	2001.5	2031.1
North	1928.1	1857.1
West	2073.5	2091.3

The tables above denote the average sales price and square footage of the homes in each neighborhood, grouped by building exterior. In all three neighborhoods, homes made of brick tended to be more expensive. Homes in the North neighborhood tended to be least expensive, followed by the East neighborhood. The West neighborhood is the most expensive neighborhood, with a home prices being $30-40,000 more expensive than a home in the East.

Regarding square footage, the data follow the same pattern as price. The North (least expensive homes) also have the smallest average square footage, followed by the East neighborhood. Further, except for homes in the North, brick homes average more square footage than homes without brick. Only in the North neighborhood do homes made of brick average about 70 SqFt less in total area.

When looking at home prices, there are myriad qualitative variables not available at this time that affect housing prices. These include specific neighborhoods or quadrants of a city, lot size, amenities available (pools, parks), distance to workplaces, housing density, age of homes or neighborhoods, etc.

Looking strictly at the data provided, the West neighborhood has homes that trend to be larger and more expensive, and likely a "better" neighborhood. In the North neighborhood, the homes are smaller and $40-60,000 less expensive than homes in the West. The East neighborhood appears to be a more median neighborhood, with the prices of homes without brick very similar to the price of homes in the North with brick.

While there is an approximate difference of $17,000 between the North/brick home and the East/brick home, a closer examination shows that the East/brick home may not be as expensive as it appears. The table below shows the price per square foot.

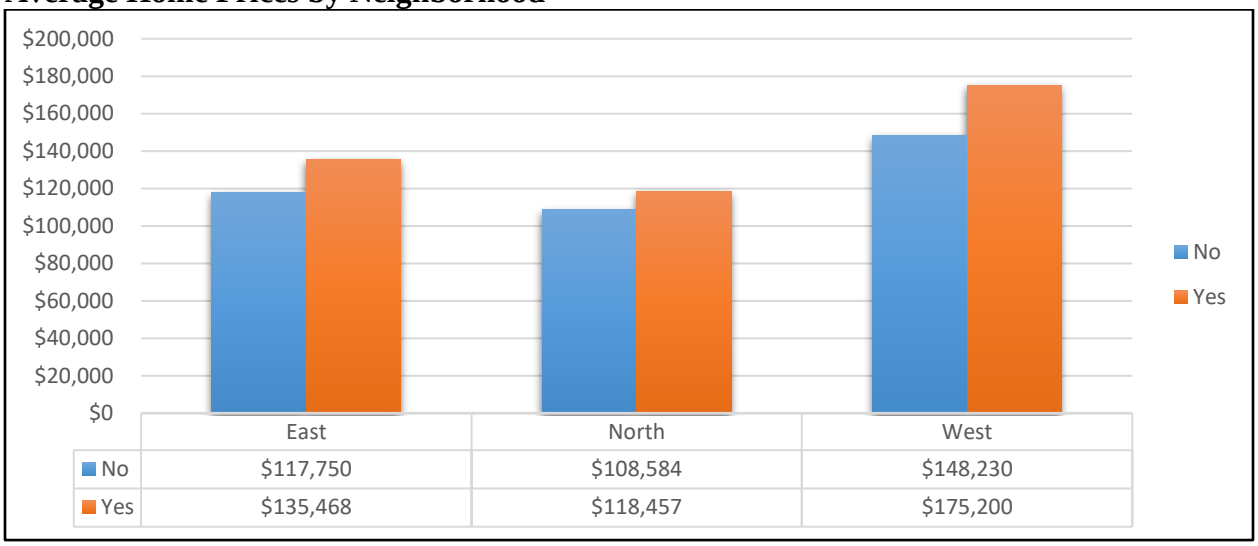
Average Price per Square Foot

AVG Price/SQFT	Brick	
Neighborhood	No	Yes
East	$59.01	$66.81
North	$56.42	$64.28
West	$71.89	$83.85

Looking at the above data, the team saw that there is only a $2.53/SqFt difference between the average brick home in the North and East neighborhoods. This data combination shows that the differences in price between the North and East neighborhoods are more a factor of square footage than neighborhood thereby affecting home values. Looking at this data visually provide further insights.

Visualizing the Data

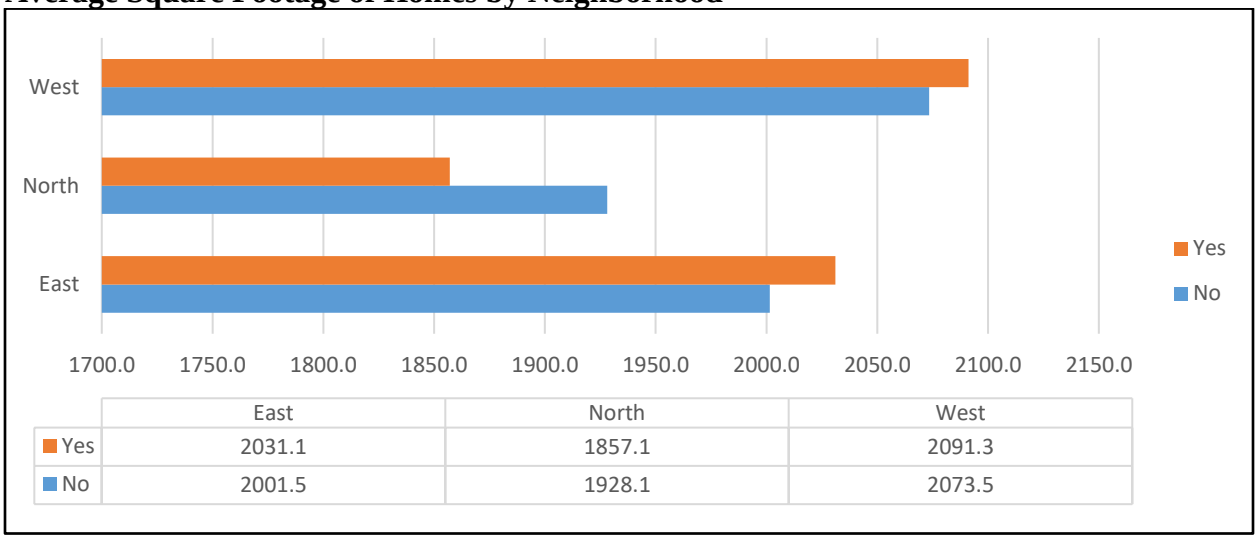
Average Home Prices by Neighborhood



	East	North	West
No	$117,750	$108,584	$148,230
Yes	$135,468	$118,457	$175,200

The above chart shows a graphical representation of the home prices. While there is little difference between the average prices in the East and North, there is a more significant increase in the sales price for homes in the West neighborhood. The prices for homes in the East without brick average about the same as homes made of brick in the North. In the chart below depicting the average square footage, some of the reason behind the differences between the North and East begin to emerge.

Average Square Footage of Homes by Neighborhood



	East	North	West
Yes	2031.1	1857.1	2091.3
No	2001.5	1928.1	2073.5

On average, the homes in the North neighborhood are the smallest compared to the other neighborhoods. Looking at the counts, only 16% of the homes in the North are made of brick, and these homes are the smallest in total area within the data provided. From data categorization and visualization, the data trends show that on average, size of home played a big part in the overall sales price. It should be noted that in real estate, a significant factor in price is location,

therefore neighborhood location is also affecting home prices. Correlation analysis can show us how the differing data points provided affect each other.

Correlation Analysis:

In the table below, is the correlation analysis between the five quantitative variables provided in the original data.

Correlation Diagram

	Price	SqFt	Bedrooms	Bathrooms	Offers
Price	**1**				
SqFt	**0.552982243**	**1**			
Bedrooms	**0.525926058**	**0.483807112**	**1**		
Bathrooms	**0.523257758**	**0.522745301**	**0.414555956**	**1**	
Offers	**-0.313635883**	**0.336923352**	**0.11427061**	**0.143793404**	**1**

As the diagram shows, the home price and square footage show the greatest correlation, meaning square footage has the greatest impact on the sales price. On the other hand, the number of offers has the least correlation with bedrooms. The reason is non-intuitive, but the meaning is clear: the number of bedrooms has the least impact on whether an offer was made on the home.

Of particular interest, offers has a negative correlation with the final sales price of the home. During our initial team discussions, the team discussed how the number of offers would reduce the overall price. Finally, deciding that offers normally do not reduce the sale price; rather it increased the price as multiple offers may have started bidding wars. What can be determined is that there is a correlation between the number of offers and each of the other variables. Regression can provide a better look at how each of the variables affect the final sales price .

Deeper Analysis

Running multivariate regression on the data using Sales Price as the “Y” axis, the data was correlated to provide a more accurate equation regarding how each variable affects price.

Regression Analysis with Sales Price as "Y"

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.835573066
R Square	0.698182349
Adjusted R Squar	0.688367141
Standard Error	14999.24552
Observations	128

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	64012998276	16003249569	71.13270927	4.4375E-31
Residual	123	27672216021	224977366		
Total	127	91685214297			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-17347.37695	12724.89631	-1.36326274	**0.175289936**	-42535.529	7840.77506	-42535.529	7840.77506
SqFt	61.8399461	8.263773843	7.48325732	**1.20211E-11**	45.4823125	78.1975797	45.4823125	78.1975797
Bedrooms	9319.752602	2148.75444	4.33728137	**2.97311E-05**	5066.42494	13573.0803	5066.42494	13573.0803
Bathrooms	12646.34749	3109.662029	4.066791622	**8.44849E-05**	6490.96217	18801.7328	6490.96217	18801.7328
Offers	-13601.01141	1324.818659	-10.2663193	**3.08843E-18**	-16223.4087	-10978.6141	-16223.4087	-10978.6141

After running a multivariate regression on the data, there were some interesting results. First, from the R-Squared output indicates that approximately 70% of the sales price is explained by the five quantitative variables. Further, the Significance F in the chart approaches zero, which indicates that the data provides a very high correlation to the sales price. Also of note, the R-Squared value also indicates that 30% of the price cannot be explained by these variables. In other words, an unknown number of unexplained variables contribute to 30% of the final sales price.

The P-value in the chart indicates that all five quantitative variables are significantly contribute to the price. Though the square footage showed the greatest correlation during our correlation analysis, the Offers variable appears to be slightly more significant. Though the team rationalized that Offers does not reduce the price, it is still a significant factor that will remain in the final analysis.

Using the coefficients provided in the regression, a specific formula was developed to determine the sales price for a home. Using the standard formula "Y=F(x)" with Sales Price as the "Y" value and the results of the regression analysis, the formula to calculate sales price in US$ is:

*Sales Price (Y) = the Intercept (-$17,347.38) + ($61.84 * Square Feet) + ($9319.75 * Bedrooms) + ($12,646.35 * Bathrooms) + (-$13,601.01 * Offers)*

The large coefficients for many of the variables are attributed to the small range of values for each. On the other hand, the small coefficient for square feet (which has the greatest correlation to price) is due to the larger data point regarding number of square feet.

Another consideration regarding the equation is that the results will have a range of +/- 30%, as indicated by R-Squared. By plugging in some known parameters into the equation, differences emerge. For example, using the first entry on the table (ID 1), produces an output of $110,076, a 4% difference from our provided data.

Predictive Analysis

Using the coefficients generated and further looking at the sensitivity of the data, predicting the sales price of homes as the team changed the data entries of two variables. First, the coefficients were placed into a table, multiplied by arbitrary values, the below table of results was developed.

Coefficients and Sample Values

Variables	Coefficients	value (Median)	coef*value
Intercept	-17347.37695	1	-$17,347.38
SqFt	61.8399461	2000	$123,679.89
Bedrooms	9319.752602	3	$27,959.26
Bathrooms	12646.34749	2	$25,292.69
Offers	-13601.01141	3	-$40,803.03
		Total Price:	$ 118,781.43

To determine a base sales price, the median value for each of the five variables was used: 2000 square feet, 3 bedrooms, 2 bathrooms, and 2 offers. Using these variables, a home with these characteristics is predicted to have a sales price of $118,781.

The next step is to determine which two variables to vary. The first variable used was square footage. This variable has the greatest correlation with the sales price, and just as important, square footage has the greatest range to allow us to quantify differences in output. For the second variable, the team debated on which to use. The offers variable has the greatest range (6 values), but the number of bathrooms was chosen instead. Bathrooms not only have similar correlation to bedrooms in determining sales price, but bathrooms also have the greatest correlation with square footage. Using a range of 1400-2600 square feet and the entire range of 2-4 bathrooms, the sensitivity analysis returned the following .

Two Way Sensitivity Analysis (SqFt and Bedrooms)

		Square Feet											
	$ 118,781.43	1400	1500	1600	1700	1800	1900	2000	2100	2200	2300	2400	2500	2600
	2	$ 81,677.47	$ 87,861.46	$ 94,045.46	$100,229.45	$106,413.44	$112,597.44	$118,781.43	$124,965.43	$131,149.42	$137,333.42	$143,517.41	$149,701.41	$155,885.40
Bathrooms	3	$ 94,323.81	$ 100,507.81	$106,691.80	$112,875.80	$119,059.79	$125,243.79	$131,427.78	$137,611.78	$143,795.77	$149,979.77	$156,163.76	$162,347.75	$168,531.75
	4	$ 106,970.16	$ 113,154.16	$119,338.15	$125,522.14	$131,706.14	$137,890.13	$144,074.13	$150,258.12	$156,442.12	$162,626.11	$168,810.11	$174,994.10	$181,178.10

The overall results in this table are straight forward: the lowest sales price belonged to the home with the least square footage and least number of bathrooms. The inverse was true for the highest sales price. Though the results are predictable, this table becomes important to a homebuyer with a specific budget. Using this table, anyone can quickly see what factors they can afford based on these two variables. For instance, if the buyer had a large family and needed four bathrooms, what is the maximum square footage I can afford and stay under $140,000? Not factoring in other variables, the buyer can use this table to quickly see that they need to look for homes within a range of 1900 to 2000 square feet.

Final Analysis and Limitations of Data

The quantitative variables all proved to have a significance in the price value of a home. With that said, most of these variables effect on price were quite intuitive, as it can be assumed that a home with more square feet, more bathrooms, more bedrooms, and more offers would, in turn, result in a higher value. the five quantitative variables accounted for almost 70% of the final sales price for a home. Each significant data point was well correlated with each other, either positively (i.e., square footage and sales price) or negatively (number of offers and sales price). The only factors that were not intuitive was the lack of correlation between the number of offers and both bedrooms and bathrooms. The data that we gathered throughout this analysis supported our initial presumption from the start.

At first, the supposition was that this may be due to the small number possible results (2-5 bedrooms, 2-4 bathrooms, and 1-6 offers). Looking closer, the data showed a stronger correlation between the two smallest ranges (bedrooms and bathrooms). This report looked at the data a different way.

Next, the data was correlated using a simple 50% rule, showing which elements trended above (in green) or below (in red) the median for that specific variable. For Offers, the highest (green) and lowest (red) quartiles were used to better show contrast in trends. The data was also sorted by neighborhood to see if this affected outcomes (see table on the next page).

ID	Price	SqFt	Bedrooms	Bathrooms	Offers	Brick	Neighborh
105	82300	1910	3	2	4	No	East
4	94700	1980	3	2	3	No	East
28	99300	1700	3	2	2	No	East
46	103200	1810	3	2	3	No	East
10	104000	1730	3	3	3	No	East
43	105600	1990	2	2	3	No	East
122	105600	1930	3	3	3	No	East
41	106600	1560	2	2	1	No	East
110	108700	2110	3	2	3	No	East
19	111400	1700	2	2	1	Yes	East
2	114200	2030	4	2	3	No	East
1	114300	1790	2	2	2	No	East
3	114800	1740	3	2	1	No	East
49	115900	1980	2	2	2	No	East
21	116200	1790	3	2	3	No	East
92	116500	2150	3	2	2	No	East
109	117000	1990	3	3	3	Yes	East
9	119200	2110	4	2	3	No	East
124	119700	1900	3	3	3	Yes	East
5	119800	2130	3	3	3	No	East
64	120500	1910	2	3	2	No	East
12	123000	1870	2	2	2	Yes	East
102	123100	2260	3	3	5	No	East
113	123600	1940	2	2	2	Yes	East
115	124500	2010	4	3	2	No	East
74	125700	2040	3	3	2	No	East
56	125700	1720	2	2	2	Yes	East
98	126800	2000	2	2	1	Yes	East
11	132500	2030	3	2	3	Yes	East
97	133300	2440	3	3	3	No	East
108	134000	1890	3	2	1	Yes	East
33	135000	2250	3	3	3	Yes	East
103	136800	2410	3	3	4	No	East
34	139600	2280	5	3	4	Yes	East
57	140900	2190	3	2	3	Yes	East
81	143400	2190	3	3	4	Yes	East
123	144800	2060	2	2	1	Yes	East
17	147100	2190	3	3	4	Yes	East
84	147700	2410	3	3	2	No	East
125	147900	2160	4	3	3	Yes	East
51	151100	2100	3	2	3	Yes	East
68	151900	2040	4	3	3	No	East
44	154000	1920	3	2	1	Yes	East
25	156400	2210	4	3	2	Yes	East
94	157100	2080	3	3	2	No	East
29	69100	1600	2	2	3	No	North
55	81300	1650	3	2	3	No	North
18	83600	1990	3	3	4	No	North
48	90300	2050	3	2	6	No	North
85	90500	1520	2	2	3	No	North
52	91100	1860	2	2	3	No	North
23	91700	1690	3	2	3	No	North
69	93600	2140	3	2	4	No	North
90	97800	2010	2	2	4	No	North
62	100900	1610	2	2	2	No	North
116	102500	1900	3	3	3	No	North
13	102600	1910	3	2	4	No	North
87	102700	1900	4	2	4	No	North
101	103200	2010	3	2	5	No	North
24	106100	1820	3	2	3	Yes	North
76	106900	1900	2	2	2	No	North
73	107300	1650	3	2	3	No	North
50	107500	1700	3	2	3	Yes	North
40	108200	1740	3	2	2	No	North
107	108500	2130	3	2	4	No	North
120	109700	1920	2	2	4	No	North
121	110400	1930	2	3	3	No	North
66	111100	1450	2	2	1	Yes	North
111	111600	1710	2	2	1	No	North
32	112300	1930	2	2	2	Yes	North
126	113500	2070	2	2	2	No	North
22	113800	2000	3	2	4	No	North
6	114600	1780	3	2	2	No	North
112	114900	1740	2	2	2	No	North
114	115700	2000	3	2	3	Yes	North
36	117100	2080	3	3	3	No	North
53	117400	2150	2	3	4	No	North
37	117500	1880	2	2	2	No	North
35	117800	2000	2	2	3	No	North
118	117800	1920	3	2	2	No	North
79	121300	2130	3	2	3	No	North
128	124600	2250	3	3	4	No	North
67	126200	2210	3	3	4	No	North
14	126300	2150	3	3	5	Yes	North
89	127700	1930	3	3	2	No	North
47	129800	1990	2	3	2	No	North
54	130800	2100	3	2	3	No	North
26	149300	2290	4	3	3	No	North
119	150200	1950	3	2	3	Yes	North
77	129800	1930	3	2	2	No	West
65	130300	1860	3	2	2	No	West
39	131300	1720	3	2	1	No	West
42	133600	1840	4	3	2	No	West
27	137000	2000	4	2	3	No	West
59	138100	1840	3	3	1	No	West
93	142600	2110	3	2	2	No	West
91	143100	1920	4	2	2	No	West
80	143600	1780	4	2	1	No	West
75	144200	2140	3	3	3	No	West
99	145500	2060	3	2	1	No	West
16	145800	1780	4	2	1	No	West
106	146900	2530	4	3	4	No	West
38	147000	2420	4	3	4	No	West
127	149900	2020	3	3	1	No	West
8	150700	2160	4	2	2	No	West
7	151600	1830	3	3	3	Yes	West
58	152300	2240	4	3	3	No	West
96	152500	1970	2	2	1	Yes	West
60	155400	2090	4	2	1	No	West
72	157600	2160	4	2	1	No	West
95	160600	2150	4	3	3	Yes	West
63	161300	2220	4	3	2	No	West
83	164800	2050	2	2	1	Yes	West
70	165600	2080	4	3	3	No	West
45	166500	1940	3	3	2	Yes	West
71	166700	1950	3	3	3	Yes	West
20	167200	1920	3	3	2	Yes	West
100	171000	2080	3	3	2	Yes	West
88	172500	1880	3	3	1	Yes	West
78	176500	2280	4	3	3	Yes	West
15	176800	2590	4	3	4	No	West
61	180900	2200	3	3	1	No	West
31	182000	2250	4	3	3	Yes	West
82	184300	2140	4	3	2	Yes	West
30	188000	2040	4	3	1	Yes	West
86	188300	2250	4	3	2	Yes	West
117	199500	2290	5	4	1	Yes	West
104	211200	2440	4	3	3	Yes	West

As shown on the above table, data patterns emerge, not just between the quantitative variables, but also between the three different neighborhoods. Most of the homes with larger square footage belong to the West and East neighborhoods. Further, most of the homes with four or more bedrooms are in the West neighborhood, with the larger square footage. Lastly, homes with the largest number of bathrooms are aligned with the homes with largest square footage in the East and West neighborhoods.

Moving to the Offers column, the lowest quartile of offers (containing only one offer) mostly appear in the West region, followed by the East region. These homes are the larger homes with greater number of bedrooms. On the other hand, the highest quartile of offers belong to homes in the North. Regarding Offers with both Beds and Bathrooms, there is a lack of correlation. There are equal amounts of homes with above average and below average bedrooms and bathrooms in each quartile for Offers.

This project provided a good context for categorizing, visualizing, correlating, and analyzing the regression of data to find answers to business needs. As the reading assignments have stated, there is always more data out there to refine your solutions. For this project, the initial list price would greatly benefit in helping refine any correlations with the offers received.

In conclusion, there were several limitations to the data that was offered in the analysis. To begin, it did not provide information as to buyer family size. Family size drives variables such as bathrooms and bedrooms, school districts, lot size, garage details, amenities available (pools, parks), distance to workplaces, housing density, age of homes or neighborhoods, availability of public transportation, commute times and other factors that contribute to the 30% of the unknown factors. Having this data would have reduced that 30%. In addition, the data did not specify any time range in which the offers were made. This can significantly skew the data as we have no visibility to the increase, or decrease, in value over time and the effect they have on price. Finally, based on the clear differences in price in relation to neighborhood, it can be argued that each neighborhood should be analyzed separately being aware that there may not be a representative number of samples in each neighborhood. To dive further into these relationships, the analysis would need to be run separately by neighborhood with enough additional samples to eliminate any questions about the team’s ability to generalize the data.