Final Written Report

NMA Group (Nihel Charfi, Mi Li, Atlas Li)

1. Brief descriptions of our dataset.

The public dataset we choose consists of historical house prices from King County, an area in the US State of Washington, between May 2014 to May 2015. The dataset was obtained from Kaggle:

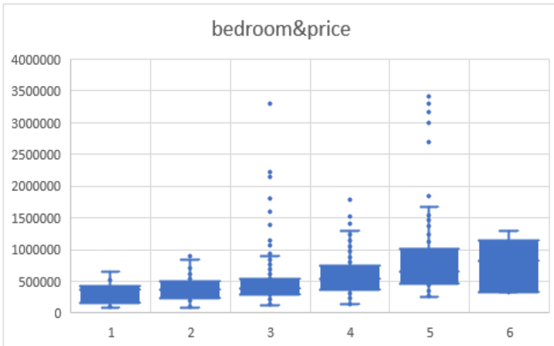
https://www.kaggle.com/swathiachath/kc-housesales-data

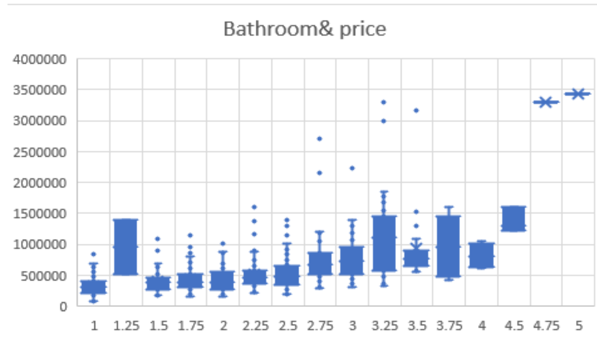
The dataset consisted of 21 variables and over 20000 observations. But we only use 'price', 'bedrooms', 'bathrooms', 'SQFT_Living', and 'SQFT_Lot' as the variables in the project to build our model. Since in common sense, these are the basic concerns when people plan to purchase house. The 'price' is the prediction target. 'Bedrooms' and 'bathrooms' are the number of bedrooms and bathroom of a single house, 'SQFT_Living' and 'SQFT_Lot' are the square footage of the living room and the lot of a house.

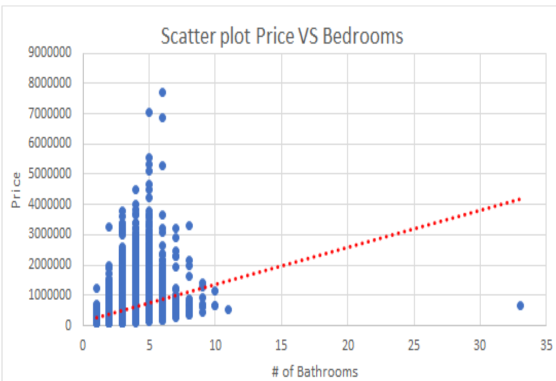
We choose this dataset because the most of data is the numerical values, it is very straightforward to show the plot and the analytic results. And it is a very interesting dataset as well, we hope we can make something valuable out.

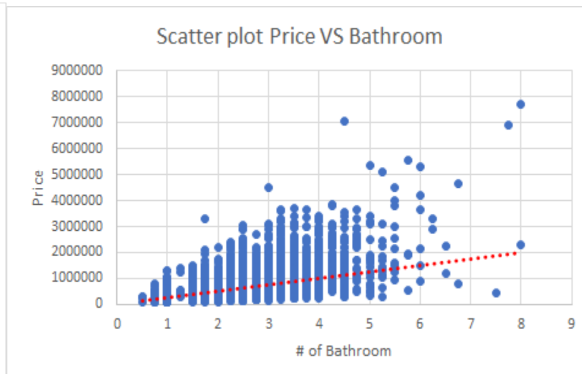
2. Pertinent plots and tables to support your work.

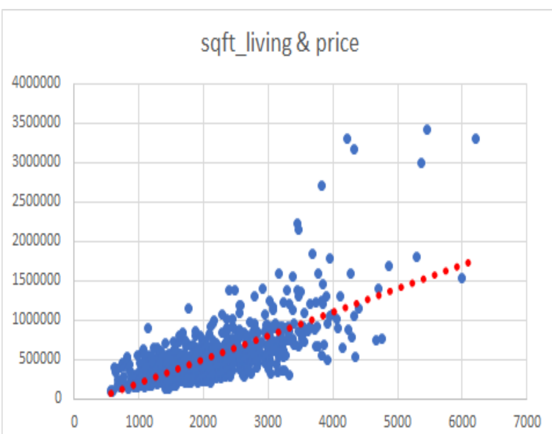
We created the line charts and box plot charts for each of the attribute's relationship with price to tell how they are correlated and how significant the features could be in the prediction in this preprocessing data phase.

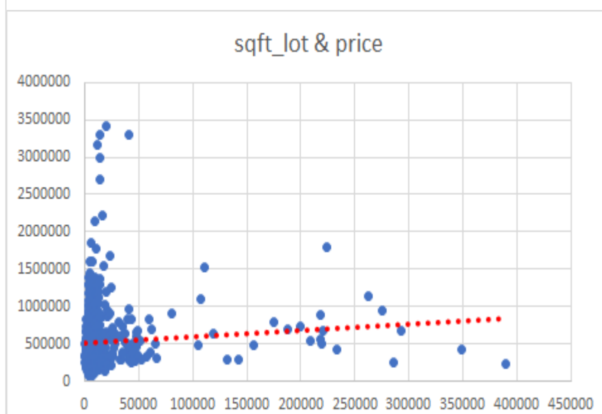












After observation, as we can see from the Scatter plots above the relation between our features sqft_living and sqft_lot and the price support the concept of linear relationship. For example, Sqft_living VS Price and Sqft_lot VS Price demonstrate a strong linear relationship or correlation between them and as the number of Sqft_lot or Sqft_living increase as the price increase as well. The boxplots indicate the price will increase as the number of bedrooms increased showed by the Interquartile range IQR of this different box plot especially for the number of bedroom 6 which has no outliers. For the bathroom and price, each of the box range are vary from one to another means the number of bathrooms has significant effect to the price. Therefore, we insisted on our original plan to build the first model based on those four features.

3. The sample in your analysis

As previously described, we have a total of 20000 data sets that means we have a big number of data rows. Thus, by Xlminer we randomly selected 800 groups of data sets(in the ribbon of data mining, we selected get data and worksheet to make it) to train and build our model. Then, we partition our 800-sample data in 60% for training and 40% for validation, then build the multi linear regression model to predict the house price. This is how it looks our new dataset after random selection:

Data Mining: Sampling Da

Output Navigator	
Inputs	Data Sample

Elapsed Times in Milliseconds			
Data Reading Time	gorithm Tin	Report Time	Total
470	104	22	596

Inputs

Data	
Workbook	kc_house_data.xlsx
Worksheet	new_Kc_house_data
Range	$A$1:$E$21598
# Records in the input data	21597

Variables					
# Selected Variables	5				
Selected Variables	bedrooms	bathrooms	sqft_livin	sqft_lot	price

Sampling Parameters	
Sampling type	Random
Sample size	800
With replacement?	FALSE
Random seed	12345
Sort indices?	TRUE

Data Sample

Record ID	bedrooms	bathrooms	sqft_living	sqft_lot	price
Record 7	3	2.25	1715	6819	257500
Record 26	3	2	1710	4697	233000
Record 44	3	1	1570	2280	685000
Record 67	4	2.5	2720	11049	975000
Record 105	3	1	1210	33919	290000
Record 166	3	1.75	1580	7000	370000

And this is how it looks our standard partition:

Data Mining: Standard Partitioner

Output Navigator		
Inputs	Partition Summary	Partitioned Data

Elapsed Times in Milliseconds			
Data Reading Time	gorithm Tin	Report Time	Total
16	135	33	184

Inputs

Data	
Workbook	kc_house_data.xlsx
Worksheet	Sampling1
Range	$C$31:$H$831
# Records in the input data	800

Variables					
# Selected Variables	5				
Selected Variables	bedrooms	bathrooms	sqft_living	sqft_lot	price

Partitioning Parameters	
Partitioning type	RANDOM
Random seed	12345
Ratio - Training	0.6
Ratio - Validation	0.4

Partition Summary

Partition	# Records
Training	480
Validation	320

Partitioned Data

Record ID	bedrooms	bathrooms	sqft_living	sqft_lot	price
Record 1	3	2.25	1715	6819	257500
Record 5	3	1	1210	33919	290000
Record 8	4	1	1330	2400	615000
Record 11	3	1	1300	8316	225000
Record 12	2	1.5	720	5120	181000
Record 15	3	2	1650	7352	252350

4. Describe the results of your feature engineering process describing which features were considered and ultimately used. Include plots of your real data and predicted data, demonstrating the improvement from specific features (plot of actual data and prediction for each feature).

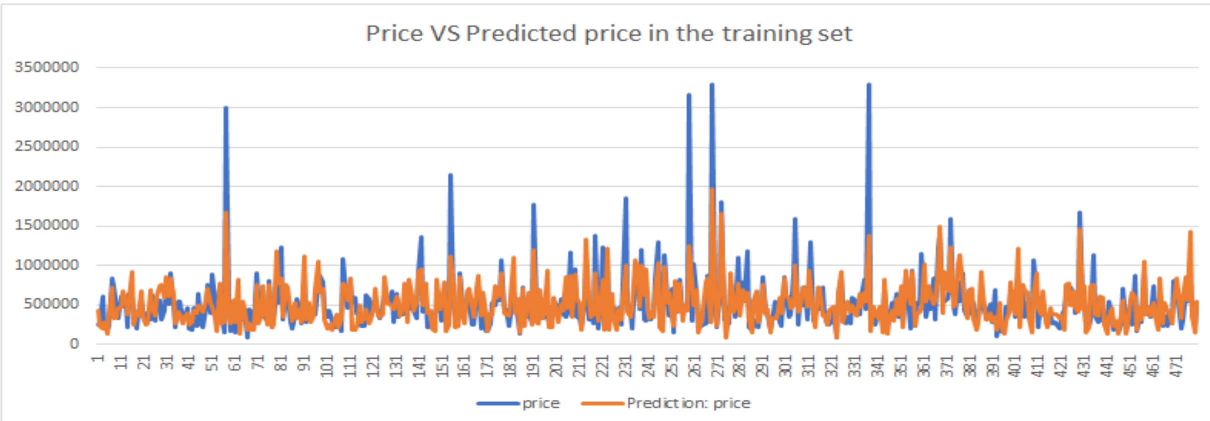
We initially planned to use the features as 'bedrooms', 'bathrooms', 'sqft_living', and 'sqft_lot' to predict the 'price' and to build our first model.

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	74357.88953	-22871.14582	171586.9249	49481.16108	1.50275151	0.1335676
bedrooms	-92795.43336	-128398.6164	-57192.25029	18118.93773	-5.121461024	4.415E-07
bathrooms	-18992.6496	-68962.23263	30976.93344	25430.19149	-0.746854368	0.455521
sqft_living	397.5487415	348.5818552	446.5156278	24.91990565	15.95305966	3.44E-46
sqft_lot	-0.808261438	-1.647325481	0.030802604	0.427010953	-1.892835377	0.0589874

Based on the first model, we can conclude that a house with biggest square footage of living could be sold at a higher price due to the positive coefficient it has. The variables bedrooms, bathrooms and SQFT_Lot have relatively lower contribution on the price, because they have a negative estimate values in our model, which means those elements have negative impact on the total price of the houses. Our first predict equation is as below:

PredictedPrice_1*=74357.88953-92795.43336****(bedrooms)****-18992.6496****(bathrooms)***
*+397.5487415****(sqft_living)****-0.808261438****(sqft_lot)***



After the observed of the actual price and predicted price, we found our predict price is generally matched good with the actual price with our first model. But for some specific price, for example in the ID 61, 151, 191,231, 261 ,271 and 341 are dramatically not accurate enough. Most of the prices we mentioned here are two times higher than the original actual price. We want to improve our model to eliminate this situation in the next model.

Validation: Prediction Summary

Metric	Value
SSE	2.21E+13
MSE	6.92E+10
RMSE	262989.2
MAD	179178.9
R2	0.44308

Since RMSE is a good measure of how accurately the model predicts the response, by defining Lower values of RMSE indicate better fit [1]. In our case, ***RMSE*** is equal to ***262989.2*** which is pretty a high value. Thus, we can say that our model is not the best model to fit our data.

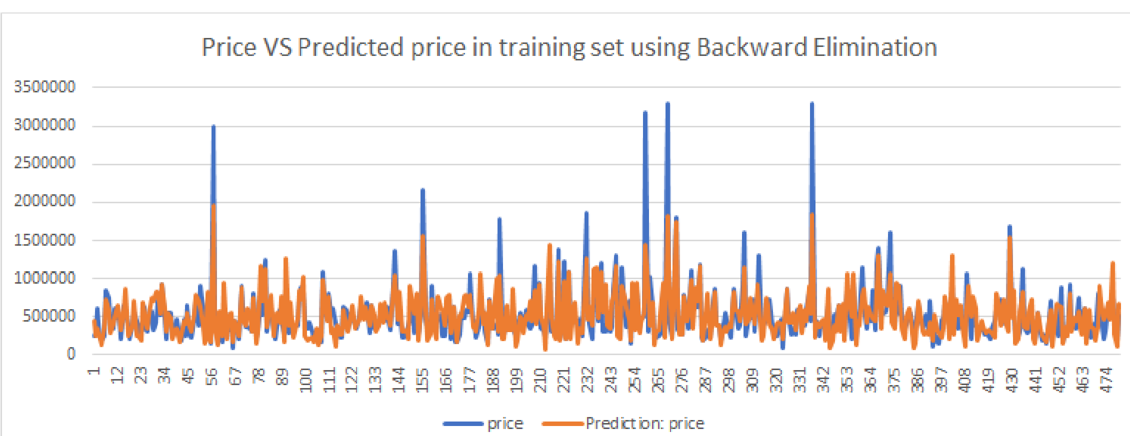
To improve our model, we can deal with other variables in the data set to include them to our current features in the random sampling data of 800 rows like "**floors, sqft_basement, yr_built, yr_renovated, zipcode, sqft_living15, sqft_lot15**" and we create other partition with 60% for the training and 40% for the validation set , then we can build a new model using the Backward elimination selection to eliminate these not significant variables which has a ***highest p-value and larger than the level of significance of 0.05*** [2].

After running multiple models by eliminating the non-significant features, we found this result:

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	5731532.099	3897926.906	7565137.293	933131.1429	6.142257863	1.72922E-09
bedrooms	-84904.65625	-116850.2898	-52959.02268	16257.29774	-5.222556516	2.65154E-07
sqft_living	269.1328453	214.9971659	323.2685246	27.5499265	9.768913368	1.18615E-20
floors	95612.92336	43455.24303	147770.6037	26543.31262	3.602147356	0.000349074
sqft_basem	127.1598846	61.63049368	192.6892755	33.34824511	3.813090738	0.000155444
yr_built	-3019.741511	-3973.888114	-2065.594909	485.5701286	-6.218960627	1.10317E-09
yr_renovate	128.4288379	74.17094668	182.686729	27.61212075	4.651176163	4.29156E-06
sqft_living1	149.622281	95.22707913	204.0174829	27.68199889	5.405038908	1.03034E-07

PredictedPrice_2*=5731532.099-84904.65625****(bedrooms)****+269.1328453****(sqft_living)***
*+95612.92336****(floors)****+127.1598846****(sqft_basement)****-3019.741511****(yr_built)****+***
*128.4288379****(yr_renovated)+149.622281*(sqft_living15)***



This modified model has largely improved our model's accuration. As we can see, the actual price is twice than the predicted price has reduced from the 7 cases into 4 cases(only ID 56, 254,265 and 342). Besides, the most of other ID cases match better than previous model.

Validation: Prediction Summary

Metric	Value
SSE	1.96E+13
MSE	6.12E+10
RMSE	247301.7
MAD	166493.8
R2	0.50754

Compared to the first model, the ***RMSE*** is equal now to ***247301.7*** instead of ***262989.2*** so, we can notice a decrease of the root-mean-squared error . Thus, we can say that we improved a little our model and we can say that it fits better than the first model.

5. Insights obtained from the model

Coefficients

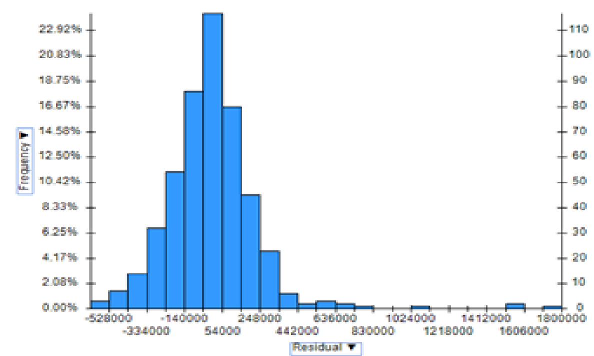
Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	5731532.099	3897926.906	7565137.293	933131.1429	6.142257863	1.72922E-09
bedrooms	-84904.65625	-116850.2898	-52959.02268	16257.29774	-5.222556516	2.65154E-07
sqft_living	269.1328453	214.9971659	323.2685246	27.5499265	9.768913368	1.18615E-20
floors	95612.92336	43455.24303	147770.6037	26543.31262	3.602147356	0.000349074
sqft_basem	127.1598846	61.63049368	192.6892755	33.34824511	3.813090738	0.000155444
yr_built	-3019.741511	-3973.888114	-2065.594909	485.5701286	-6.218960627	1.10317E-09
yr_renovate	128.4288379	74.17094668	182.686729	27.61212075	4.651176163	4.29156E-06
sqft_living1	149.622281	95.22707913	204.0174829	27.68199889	5.405038908	1.03034E-07

Based on the improved regression model, we can conclude the features that increase the house price is the square footage of it, total floors that it has, the square footage of the basement and the year when the house was renovated, due to all the related features we got from this model have relatively low P-value. We believe the most significant features we got from the model matches our general knowledge of the factors decide the market house price. Therefore, we believe our model did a great job for helping predict the house price for the people who living in this zip code places.

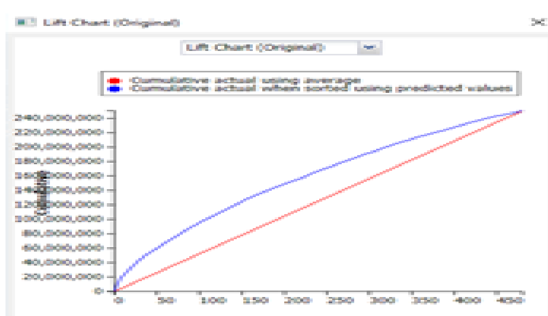
6. Reflection

6.1 Evaluate the predictive model

To evaluate the predictive model, we already explain how the ***RMSE*** decreased from the first model and based on it we concluded that our model improved. Furthermore, to evaluate our predictive model we can create a histogram for the residuals on the training set to check how the distribution looks like and how it affects to the performance or our model.



This histogram does support a bell-shape for normal distribution with some outliers. That means the model residuals do appear to follow a normal distribution. Thus, the predictive performance of the model demonstrates that our prediction is better than the baseline model.



6.2 Describe your challenges.(M)

It was hard for us to decide the best features selected to build our model even we improved our model using the backward elimination. Therefore, the feature selection “best subset method” is especially important for us in this project since it choose the best features from the whole data set automatically to figure out the best result and to evaluate our prediction.

6.3 Are you aware of other datasets could you incorporate to improve your model?

As we mentioned before the best subset selection is the method to select the best features of our model. Thus, we did another one a random data sample (800 rows) of the whole data set without exemption of any variables and then from this new data we can run the Standard partition and the linear regression prediction by selecting the option ***best subset method***. From, the first model of the best subset method we select the subset feature with ***higher R squared adjusted*** because how much it is high how much our model fits very well the dataset[3]. This illustrated in the following figure:

Best Subsets Details

Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability
Subset 1	1	6.81099E+13	1641.546703	-2.22045E-16	-2.22045E-16	1.029E-140
Subset 2	2	3.3441E+13	564.6680221	0.509013875	0.507986708	4.51318E-71
Subset 3	3	2.56738E+13	324.9567326	0.623053018	0.621472528	5.83275E-46
Subset 4	4	2.18277E+13	207.268016	0.679522034	0.677502215	5.60132E-31
Subset 5	5	1.89952E+13	121.1213813	0.721109528	0.718760976	1.79081E-18
Subset 6	6	1.71317E+13	65.12953603	0.748470022	0.745816752	1.23083E-09
Subset 7	7	1.65342E+13	48.53493043	0.757242938	0.754163567	5.6768E-07
Subset 8	8	1.60024E+13	33.98818547	0.765049676	0.761565243	0.000127201
Subset 9	9	1.57486E+13	28.08734811	0.768777283	0.764849933	0.001168192
Subset 10	10	1.55073E+13	22.57964238	0.772319411	0.76795957	0.009380391
Subset 11	11	1.52762E+13	17.3891279	0.775711888	0.770929626	0.065798642
Subset 12	12	1.51607E+13	15.79417196	0.777407985	0.772176121	0.136327296
Subset 13	13	1.50388E+13	14.00117241	0.779197518	0.773523793	0.307989579
Subset 14	14	1.49432E+13	13.02414448	0.780602077	0.774481534	0.554350018
Subset 15	15	1.48981E+13	13.62042103	0.781264352	0.774678763	0.655013849
Subset 16	16	1.48888E+13	15.33227895	0.781400297	0.774333497	0.514180535
Subset 17	17	1.48479E+13	16.06046676	0.782000337	0.774466871	0.805868923
Subset 18	18	1.4846E+13	18.00000004	0.782028865	0.774008282	0

Then we can click on this subset and we find only the selected best features which are “***bedrooms, Sqft_living, floors, waterfront, view, condition, grade, sqft basement, yr_built, yr_renovated, zipcode, lat, long*** and ***sqft_living 15*** ”and automatically it eliminates the other once. After that, we can rerun other model with only these selected features [3].The results are the following:

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	49196458.94	14800490.12	83592427.76	17503608.6	2.810646654	0.005152915
bedrooms	-37193.25081	-62648.16466	-11738.33695	12953.63568	-2.8712596	0.004275045
sqft_living	176.4361482	130.2535897	222.6187067	23.50163279	7.507399584	3.10306E-13
floors	43572.30566	627.0358422	86517.57547	21854.22362	1.993770468	0.046760785
waterfront	931308.7591	722446.965	1140170.553	106286.7313	8.76222975	3.60086E-17
view	54036.73858	28707.76093	79365.71623	12889.54858	4.192291006	3.30674E-05
condition	45260.9757	17158.43786	73363.51354	14300.97305	3.16488784	0.001653214
grade	105090.1627	79424.76097	130755.5645	13060.74991	8.046257941	7.17403E-15
sqft_basem	69.64829893	16.78727522	122.5093226	26.90020665	2.589136204	0.009923264
yr_built	-2791.726958	-3661.256887	-1922.197029	442.4911426	-6.309113764	6.54478E-10
yr_renovate	42.05841531	-2.877409576	86.99424019	22.86718816	1.839247354	0.066516181
zipcode	-913.7752341	-1310.815621	-516.7348474	202.048082	-4.52256327	7.76418E-06
lat	512713.6652	389738.0138	635689.3166	62580.5216	8.192863403	2.49082E-15
long	-169497.3424	-323914.4269	-15080.25787	78580.60992	-2.156986852	0.031518459
sqft_living1	27.80364384	-18.24275741	73.8500451	23.43234433	1.186549815	0.23601111

Validation: Prediction Summary

Metric	Value
SSE	1.31E+13
MSE	4.09E+10
RMSE	202150.7
MAD	128981.3
R2	0.670946

Compared to the first two models , the ***RMSE*** is equal now to ***202150.7*** instead of ***247301.7*** from the backward elimination model and we can notice another decrease of the root-mean-squared error. Thus, we can say that those features lead us to the best fit model of our dataset.

6.4 In this project, what skills did you employ? What skills do you think you can improve upon in the future? How might you go about improving those skills?

In this project, we used line charts and box plot charts to see the relationship between the variables and the price. Then we used XLMiner to randomly select a part of data to separate them into training and validation. After this we built the multiple linear regression model to predict the house price and compared the predicted price with the previous actual price to check the accuracy of our model. What is more, we also used the Backward Elimination Method to determine what is the unimportant data, after switched these data, our model become much better. In the future, we think we need to improve the ability of how to select the best features for our analytic model at the first time. Also, we need to be more familiar with the tools we used in this project. Those are not easy, but we can make it by practicing and having a good plan in mind before acting. We think these will be helpful.

References:

[1] Assessing the Fit of Regression Models Available online at:
https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/

[2] How to use Excel to do the backward Elimination to find the best model: https://www.youtube.com/watch?v=8xpSfhrdlEs

[3] Business Intelligence -- Regression model in XLMiner
https://www.youtube.com/watch?v=qZpzZHIZ2eU&t=104s