BSTA 450 Section A

Final Project – King County Housing Project

Mather Rahhal - 40060748

Kevin Heng - 40055424

Professor: Mohsen Farhadloo, Ph.D.
Teacher Assistant : Parisa Foroutan

Concordia University

December 7th, 2020

Table of Contents

Executive Summary 3
Introduction 4
  Data Collection/Cleaning 4
  Random sample and prediction sample creation through SAS EG random sample function: 5
  Backwards, Forward & Stepwise (0.05, 0.10 & 0.15 significance): 5
  Dummy Variable 5
    Condition: 5
    Grade 5
Statistical Analyses 6
  Data transformation: 6
  Summary Statistics 6
  Why did we transform certain variables? 6
  Residual Plot Analysis 7
  ANOVA Table - A: Variable selected dummies regression Table 8
  SAS – Partial Regression ANOVA Table 8
  Hypothesis Testing 8
  Discussion of result, interpretation, and conclusions 10
References 11
Appendix A – Variable Grade Description 12
Appendix B – Variable Condition Description 13
Appendix C – Linear Regression Model 13
  Linear Regression Equation Model 13
  Linear Regression Equation – Revert back with $e^{logprice}$ 13
  Figure A – ANOVA table without grade and condition 14
  Figure B – ANOVA table with Grade polynomial and Condition 15
  Figure C – ANOVA table with Grade and Condition dummies 16
  Figure D – Backward Selection ($\alpha = 0.05; 0.10; 0.15$) 17
  Figure E – Forward($\alpha = 0.05; 0.10; 0.15$) 18
  Figure F – Stepwise Selection($\alpha = 0.05; 0.10; 0.15$) 18
Appendix D - Program 19
  Problem encountered 20
Appendix E - Variables Kept 20
Appendix F – Kaggle Dataset 21

Executive Summary

This study was based on a dataset of house pricing in King County with a dataset size of 21,597 observations. The dataset was cleaned from variables that added no value to the study or were too difficult to transform to bring statistical meaning to the study (all cleansing was validated through approval from the Teacher/TA). The large dataset was randomly split to create a random sample of 10,799 and a prediction sample of 10,798 observations. The data study conclusions are valid and can be accepted since none of the four regression analysis assumptions were violated. The study has shown that the model with dummy variables is advantageous to use. The variable selection was also performed and has eliminated some of the dummy variables and variables. The final model includes optimal model contained 21 variables (including dummy variables): bathrooms, floors, waterfront, view, condition_3, condition_4, condition_5, grade_4, grade_5, grade_6, grade_7, grade_8, grade_9, grade_10, grade_11, grade_12, basement, renovated, yr_built, sqrt_living15, sqrt_lot15. This model gave us an adjusted R^2 0.6548, and this means that our model explains 65.48% of the variations in price in King County. To our surprise, floors, bedrooms, and sqrt_living15 only contributed by 5.32%. Finally, we predicted prices using our model mentioned above. We received a Mean Absolute Error of approximately 1.87%, which shows that the model and the variables included within can reasonably predict houses' prices in King County.

Introduction

We chose the study of house pricing in King County because of our curiosity about real estate. As students, we will eventually be buyers in the housing market, and we thought it would be a great idea to analyze the variables that come into play in explaining house pricing.

We hypothesize that removing floors, bathrooms, and sqrt_living15 will affect the variation in price.

Data Collection/Cleaning

The Kaggle dataset, *kc_house_data.csv,* covers King County's house price in Washington, USA. (Achath, 2018)

Variables kept: Price, Bedroom, Bathroom, Waterfront, Views, Sqft_living15, Sqft_lot15, Grade, Condition. (see Appendix E)

Removed Variables	**Description of Variables**	**Reason for Removal**
ID	ID of buyer	Irrelevant to the study
Date	Date Purchased	Complicated. Suggested to remove
Sqft_above	Measurements before renovation	Irrelevant (not the current size)
Sqft_living	Size inside the house	[…] Same as above
Sqft_Lot	Size of the Lot	[…] Same as above
Zipcode	Zipcode (Similar to Postal Code in Canada)	Very complicated to turn these numbers into meaningful data
Lat	Geographical coordinates	[…] Same as above
Long	Geographical coordinates	[…] Same as above

Modified Variable	**Transformation**
Yr_renovated Turned into 1 or 0 variables	Transformed it into Yes/No (Query Builder)
Sqft_basement Turned into 1 or 0 variables	Transformed it into Yes/No (Query Builder)
Grade (1-13) turned into 1 or 0 dummy variables	13 Grades are categorical data; therefore, we separate into Dummy Variables (Query Builder)
Condition (1-5) turned into 1 or 0 dummy variables	5 Conditions are categorical data; therefore, we separate into Dummy Variables (Query Builder)

Temporary observation column variable was added on excel and used to create a random sample and random predicted sample through SAS.

Random sample and prediction sample creation through SAS EG random sample function:

We randomly select 10,799 from the original sample N size: 21,597

- To create a random sample
- With the remaining 10,798 observations, we created the prediction sample.

The prediction sample was created using SAS programming. (see Appendix D).

Variable Selection Model & Selected Predictors Variables:

We ran variable selection techniques like Mallow's CP, Backward Selection, Forward Selection and Stepwise Selection. It has shown that all our current variables are significant.

Malow's CP. The first option was chosen with the Lower Cp and higher R-Squared.

Model Index	Number in Model	C(p)	R-Square	Variables in Model
1	12	13.0000	0.6148	bedrooms bathrooms floors waterfront view condition grade basement renovated yr_built sqft_living15 sqft_lot15
2	11	16.6734	0.6146	bedrooms bathrooms floors waterfront view condition grade basement renovated yr_built sqft_living15

Backwards, Forward & Stepwise (0.05, 0.10 & 0.15 significance):

Gave the same results: Bedrooms, Bathrooms, Floors, Waterfront, View, Condition, Grade, Basement, Renovated, Yr_built, Sqft_living15, Sqft_lot15 (View SAS for all 9 SAS results if desired).

Dummy Variable

After further investigation, we provided meaning into Grade and Condition by creating dummies variables

Condition: Condition_1, Condition_2, Condition_3, Condition_4, Condition_5 (see condition)

Grade: Grade_3, Grade_4, Grade_5, Grade_6, Grade_7, Grade_8, Grade_9, Grade_10, Grade_11, Grade_12, Grade_13(see grade description)

Use Reference for the indicator variable selection

- Dummy variable condition: Condition_1
- Dummy variable grade: Grade_13
- Dummy variable Grade_3 is dropped because it has no data points in the random sample

Statistical Analyses

Summary Statistics

Variable	Mean	Std Dev	Minimum	Maximum	N
price	542117.98	382753.58	78000.00	7700000.00	10799
bedrooms	3.3842022	0.9485244	1.0000000	33.0000000	10799
bathrooms	2.1251273	0.7789221	0.5000000	8.0000000	10799
floors	1.4861561	0.5339190	1.0000000	3.5000000	10799
waterfront	0.0072229	0.0846840	0	1.0000000	10799
view	0.2339105	0.7634175	0	4.0000000	10799
condition	3.4113344	0.6531764	1.0000000	5.0000000	10799
grade	7.6573757	1.1675003	4.0000000	13.0000000	10799
basement	0.3962404	0.4891381	0	1.0000000	10799
renovated	0.0402815	0.1966278	0	1.0000000	10799
yr_built	1971.15	29.2762727	1900.00	2015.00	10799
sqft_living15	1994.18	686.8423894	399.0000000	6110.00	10799
sqft_lot15	12840.39	28644.08	750.0000000	871200.00	10799

Data transformation:

1. We use log transformation on price, as well as Bedroom and Sqft_lot15
2. We added a polynomial variable for gradeSquare (exponent = 2) and gradeCube (exponent = 3)

Why did we transform certain variables?

- Log was performed on variables (Price (dependent) and Bedroom, Sqft_lot15 (independent)) due to skewness in the scatterplots where a few points were much larger than most of the dataset.
- Polynomial variables were added to grade due its scatterplots had a quadratic shape.

Collinearity Test

We ran a co-linearity test on the 13 selected transformed variables: (LogPrice, LogBedroom, bathrooms, floors, waterfront, view, condition, grade, basement, renovated, yr_built, sqft_living15 and logSqftLot15). The collinearity is not a problem because $\forall\ predictors\ variable\ \rho < 0.8 \therefore$ collinearity between the selected variable is not a concern.

Pearson Correlation Coefficients, N = 10799
Prob > |r| under H0: Rho=0

	logPrice	logBedroom	bathrooms	floors	waterfront	view	condition	grade	basement	renovated	yr_built	sqft_living15	logSqftLot15
logPrice	1.00000	0.34189	0.55092	0.31303	0.17376	0.34899	0.05087	0.70055	0.21997	0.11323	0.07524	0.61648	0.12362
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
logBedroom	0.34189	1.00000	0.52183	0.19810	-0.02634	0.06250	0.03777	0.37857	0.16435	0.01565	0.19253	0.39723	0.16582
	<.0001		<.0001	<.0001	0.0062	<.0001	<.0001	<.0001	<.0001	0.1038	<.0001	<.0001	<.0001
bathrooms	0.55092	0.52183	1.00000	0.50548	0.06211	0.18637	-0.11523	0.66285	0.17618	0.04418	0.50243	0.56275	0.08562
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
floors	0.31303	0.19810	0.50548	1.00000	0.01860	0.02862	-0.25613	0.45763	-0.24247	0.00708	0.48322	0.28708	-0.21726
	<.0001	<.0001	<.0001		0.0533	0.0029	<.0001	<.0001	<.0001	0.4622	<.0001	<.0001	<.0001
waterfront	0.17376	-0.02634	0.06211	0.01860	1.00000	0.40218	0.00656	0.07749	0.03374	0.10488	-0.02356	0.08067	0.07122
	<.0001	0.0062	<.0001	0.0533		<.0001	0.4957	<.0001	0.0005	<.0001	0.0143	<.0001	<.0001
view	0.34899	0.06250	0.18637	0.02862	0.40218	1.00000	0.06370	0.24984	0.18677	0.10257	-0.06436	0.27340	0.11051
	<.0001	<.0001	<.0001	0.0029	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
condition	0.05087	0.03777	-0.11523	-0.25613	0.00656	0.06370	1.00000	-0.13857	0.13708	-0.04538	-0.36306	-0.07902	0.08631
	<.0001	<.0001	<.0001	<.0001	0.4957	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
grade	0.70055	0.37857	0.66285	0.45763	0.07749	0.24984	-0.13857	1.00000	0.06229	0.01293	0.44701	0.71460	0.20002
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.1792	<.0001	<.0001	<.0001
basement	0.21997	0.16435	0.17618	-0.24247	0.03374	0.18677	0.13708	0.06229	1.00000	0.03913	-0.15888	0.04249	-0.06114
	<.0001	<.0001	<.0001	<.0001	0.0005	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
renovated	0.11323	0.01565	0.04418	0.00708	0.10488	0.10257	-0.04538	0.01293	0.03913	1.00000	-0.21609	0.00121	0.02642
	<.0001	0.1038	<.0001	0.4622	<.0001	<.0001	<.0001	0.1792	<.0001		<.0001	0.8997	0.0060
yr_built	0.07524	0.19253	0.50243	0.48322	-0.02356	-0.06436	-0.36306	0.44701	-0.15888	-0.21609	1.00000	0.32793	0.03075
	<.0001	<.0001	<.0001	<.0001	0.0143	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.0014
sqft_living15	0.61648	0.39723	0.56275	0.28708	0.08067	0.27340	-0.07902	0.71460	0.04249	0.00121	0.32793	1.00000	0.38040
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.8997	<.0001		<.0001
logSqftLot15	0.12362	0.16582	0.08562	-0.21726	0.07122	0.11051	0.08631	0.20002	-0.06114	0.02642	0.03075	0.38040	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0060	0.0014	<.0001	

We tried four models:

1. M1: Without Grade and Condition (see Appendix – Figure A)
2. M2: With Grade polynomial and Condition (See Appendix – Figure B)
3. M3: With Grade dummies and Condition dummies (See Appendix – Figure C)
4. M4: We ran variable selection on Dummies regression (See ANOVA Table – A)

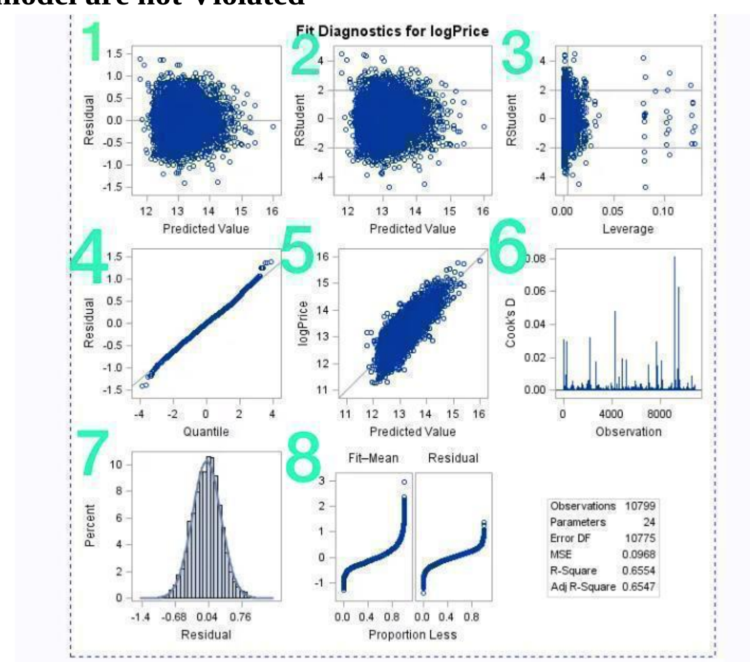
For model 4, we ran stepwise selection at $\alpha = 0.05$ and Malow CP. We choose Malow Cp selected variable as our predictors for model 4. We compare R^2_{adj} among the model to select the best fit linear regression model. We choose Model 4 because its adjusted R square is higher than the other models:

$$M4's\ R^2_{adj} = 0.6548\ < M3's\ R^2_{adj} = 0.6547 < M2's\ \ R^2_{adj} = 0.6538 < M1's\ R^2_{adj} = 0.5566$$

Residual Plot Analysis

Ensuring the assumptions of the regression model are not Violated

a) Residual has a mean = 0 (Graphs 1 & 2) (No Pattern)
b) Residual variances are constant (Graphs 1 & 2) (No pattern)
c) Residuals are normally distributed (Graphs 4 & 7). (4) The Q-Q Plot follows the reference normal line. (7) It follows the bell shape curve.
d) Residuals assume a linear curve (Graph 5) It is linear.



The assumptions are not violated, and thus we are fine to conclude that the results that come out of the regression can be accepted.

ANOVA Table - A: Variable selected dummies regression Table

Number of Observations Read	10799
Number of Observations Used	10799

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	1984.02632	94.47744	976.14	<.0001
Error	10777	1043.06975	0.09679		
Corrected Total	10798	3027.09607			

Root MSE	0.31111	R-Square	0.6554
Dependent Mean	13.04838	Adj R-Sq	0.6548
Coeff Var	2.38425		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	25.20319	0.30460	82.74	<.0001
bathrooms	1	0.11372	0.00616	18.46	<.0001
floors	1	0.09755	0.00812	12.01	<.0001
waterfront	1	0.45675	0.03892	11.74	<.0001
view	1	0.04060	0.00461	8.80	<.0001
condition_3	1	0.17158	0.03188	5.38	<.0001
condition_4	1	0.20340	0.03199	6.36	<.0001
condition_5	1	0.26575	0.03324	7.99	<.0001
grade_4	1	-2.05115	0.15070	-13.61	<.0001
grade_5	1	-1.86205	0.11708	-15.90	<.0001
grade_6	1	-1.65666	0.11373	-14.57	<.0001
grade_7	1	-1.40550	0.11275	-12.47	<.0001
grade_8	1	-1.17694	0.11217	-10.49	<.0001
grade_9	1	-0.90474	0.11188	-8.09	<.0001
grade_10	1	-0.72577	0.11195	-6.48	<.0001
grade_11	1	-0.54525	0.11303	-4.82	<.0001
grade_12	1	-0.34163	0.11857	-2.88	0.0040
basement	1	0.10456	0.00721	14.51	<.0001
renovated	1	0.03811	0.01626	2.34	0.0191
yr_built	1	-0.00586	0.00014720	-39.80	<.0001
sqft_living15	1	0.00017975	0.00000693	25.95	<.0001
logSqftLot15	1	-0.03790	0.00448	-8.46	<.0001

SAS – Partial Regression ANOVA Table

Number of Observations Read	10799
Number of Observations Used	10799

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	1823.18967	101.28832	906.95	<.0001
Error	10780	1203.90639	0.11168		
Corrected Total	10798	3027.09607			

Root MSE	0.33419	R-Square	0.6023
Dependent Mean	13.04838	Adj R-Sq	0.6016
Coeff Var	2.56112		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	22.97380	0.31532	72.86	<.0001
waterfront	1	0.42507	0.04177	10.18	<.0001
view	1	0.06123	0.00491	12.46	<.0001
condition_3	1	0.18558	0.03424	5.42	<.0001
condition_4	1	0.20794	0.03435	6.05	<.0001
condition_5	1	0.31022	0.03565	8.70	<.0001
grade_4	1	-2.99426	0.15974	-18.75	<.0001
grade_5	1	-2.81296	0.12298	-22.87	<.0001
grade_6	1	-2.60137	0.11935	-21.80	<.0001
grade_7	1	-2.25225	0.11881	-18.96	<.0001
grade_8	1	-1.89129	0.11874	-15.93	<.0001
grade_9	1	-1.49076	0.11885	-12.54	<.0001
grade_10	1	-1.20481	0.11927	-10.10	<.0001
grade_11	1	-0.87950	0.12081	-7.28	<.0001
grade_12	1	-0.55407	0.12709	-4.36	<.0001
basement	1	0.11728	0.00696	16.85	<.0001
renovated	1	0.09961	0.01724	5.78	<.0001
yr_built	1	-0.00404	0.00014494	-27.88	<.0001
logSqftLot15	1	-0.02409	0.00421	-5.72	<.0001

Hypothesis Testing

1. ***Can any of the predictors explain price?***

$H0: \beta 1 = \beta 2 = \beta 3 = \cdots = \beta 24 = 0$

Ha: At least one of β_i is different

Decision Rule: Reject H0 if:

$F > F(\alpha, k, n - k - 1) =$

$F(0.05, 21, 10798 - 21 - 1) = \frac{1.58+1.52}{2} =$

$1.55 \therefore F > 1.55$

1. (cont)

Do not reject H0 if:

$F \leq F(\alpha, k, n - k - 1) = 1.55 \therefore F \leq 1.55$

F – test:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE} = 976.14$$

Decision: *Reject H0:* 976.14 > *1.55.*

Conclusion: At least one of the βi is different. Then, at least one of x_i can the variation in price with the model.

2. **How much does inclusion #floors, bathrooms, and sqrt_living15 improve or worsen the linear regression model?**

(Full table: See ANOVA Table A);

(Partial Table: See Partial)

$H0 = \beta_{bathroom} = \beta_{floors} = \beta_{sqrt_{living15}} = 0$

Ha: At least one of β_i is different

Decision rule:

Reject H0 if:

F > F (a,k – L, n – k – 1)

= F (0.05,21-18,10798-21-1)=2.61

2. (Cont)

Do not reject H0 if:

$F \le F(\alpha, k - L, n - k - 1) = 2.61$

F Test:

$$F = \frac{\frac{SSE_r - SSE_f}{k-l}}{\frac{SSE_f}{n-k-1}} = \frac{\frac{1203.906-1043.07}{21-18}}{\frac{1043.07}{10798-21-1}} = 90.4099$$

Decision: Reject H0 ∴ 90.40995 > 2.61

Conclusion: At least one of the chosen predictors ($\beta_{bathroom}$, β_{floors}, $\beta_{sqrt_{living15}}$) is different, and it is significant in explaining the variation in price in the full model.

3. ***Does Logsqft_lot15 influence price?*** *(see ANOVA Table – A)*

H0: $\beta_{logsqft_{lot15}} = 0$

Ha: $\beta_{logsqft_{lot15}} \neq 0$

(Question 3 – Continued) Decision rule:

Reject H0 if:

$t > t\left(\frac{\alpha}{2}, N - k - 1\right) = t\left(\frac{0.05}{2}, 10798 - 21 - 1\right) = 1.96$

$t < -t\left(\frac{\alpha}{2}, N - k - 1\right) = -t\left(\frac{0.05}{2}, 10798 - 21 - 1\right) = -1.96$

∴ t > 1.96 or t < -1.96

Do not reject H0 if:

$t \le t\left(\frac{\alpha}{2}, N - k - 1\right) = 1.96$

T – test:

$t = \frac{b_j}{S_j} = -8.46$

Decision: Reject H0: $-8.46 < -1.96$

Conclusion: LogSqrt_15 is not equal zero, then LogSqrt_15 predictor is significant in explaining the variation in price.

4. **We calculated the prediction value**

(Please refer to *prediction_using_model.xlsx* **for the calculation of the MAE using the price predicted and the actual price for more details)**

LOG MAE	MAE	MAE %
0.243466518	$1.2757	1.8659%

When comparing the predicted price and the actual price, it indicates a small MAE of roughly 1.28$, which is 1.8659 % over the actual price mean. It shows that our model can adequately predict King County's house prices with the lowest error possible.

Discussion of result, interpretation, and conclusions

Reverting the price from the linear logPrice model (see linear regression model). Excel was causing issues, and thus we used SAS. Interpretation has been made throughout the study, and many variables were removed from the study in the data cleaning phase due to their irrelevance. Interestingly, the number of bedrooms was found irrelevant in determining the price. The optimal model was the one with dummy variables. After variable selection, the optimal model contained 21 variables (including dummy variables): bathrooms, floors, waterfront, view, condition_3, condition_4, condition_5, grade_4, grade_5, grade_6, grade_7, grade_8, grade_9, grade_10, grade_11, grade_12, basement, renovated, yr_built, sqrt_living15, logSqrtLot15. This model gave us an adjusted $R^2 = 0.6548$, which means that our model explains 65.48% of the variations in price. Log was performed on certain variables as mentioned above, and none of the regression assumptions were violated. Interestingly, the removal of floors, bathrooms, and sqrt_living15 reduced the adjusted R^2 by 5.32%. Finally, we tested the model and predicted prices from the prediction sample. The predictions gave a Mean Absolute Error of approximately 1.87%. This shows that our model and the variables within it can be used to reasonably predict the prices of houses in King County.

References

King County Government. (2017, August 16). https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r

Achath, S. (2018). *KC_Housesales_Data*. Kaggle. https://www.kaggle.com/swathiachath/kc-housesales-data

Appendix A – Variable Grade Description

Represents the construction quality of improvements. Grades run from grade 1 to 13. (King County Government, 2017). Generally defined as:

1-3. Falls short of minimum building standards. Normally cabin or inferior structure.

4. Generally older, low quality construction. Does not meet code.

5. Low construction costs and workmanship. Small, simple design.

6. Lowest grade currently meeting building code. Low quality materials and simple designs.

7. Average grade of construction and design. Commonly seen in plats and older sub-divisions.

8. Just above average in construction and design. Usually, better materials in both the exterior and interior finish work.

9. Better architectural design with extra interior and exterior design and quality.

10. Homes of this quality generally have high quality features. Finish work is better, and more design quality is seen in the floor plans. Generally, have a larger square footage.

11. Custom design and higher quality finish work with added amenities of solid woods, bathroom fixtures and more luxurious options.

12. Custom design and excellent builders. All materials are of the highest quality and all conveniences are present.

13. Generally custom designed and built. Mansion level. Large amount of highest quality cabinetwork, wood trim, marble, entryways etc." (King County Gouv, 2017)

Appendix B – Variable Condition Description

Condition	Description
1	Inferior
2	Below average
3	Average
4	Above average
5	Excellent

Appendix C – Linear Regression Model

Linear Regression Equation Model

(see SAS EG file: Code for Selected Linear Regression)

LogPrice = 25.20319 + 0.11372 bathrooms + 0.9755 floors + 0.45675 waterfront + 0.0406 view + 0.17158 condition_3 + 0.2034 condition_4 + 0.26575 condition_5 – (2.05115 grade_4 + 1.86205 grade_5 + 1.65666 grade_6 + 1.4055 grade_7 + 1.17694 grade_8 + 0.90474 grade_9 + 0.72577 grade_10 + 0.54525 grade_11 + 0.34163 grade_12) + 0.10456 basement + 0.03811 renovated -0.00586 yr_built + 0.00018 sqft_living15 - 0.0379 logSqftLot15

Linear Regression Equation – Revert back with $e^{\log(price)}$

(see SAS files: Prediction & Linear equation - linear_regression_parameter)

$Price = e^{LogPrice}$ = e^(25.20319 + 0.11372 bathrooms + 0.9755 floors + 0.45675 waterfront + 0.0406 view + 0.17158 condition_3 + 0.2034 condition_4 + 0.26575 condition_5 – (2.05115 grade_4 + 1.86205 grade_5 + 1.65666 grade_6 + 1.4055 grade_7 + 1.17694 grade_8 + 0.90474 grade_9 + 0.72577 grade_10 + 0.54525 grade_11 + 0.34163 grade_12) + 0.10456 basement + 0.03811 renovated -0.00586 yr_built + 0.00018 sqft_living15 - 0.0379 logSqftLot15)

Figure A – ANOVA table without grade and condition

Number of Observations Read	10799
Number of Observations Used	10799

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1686.14052	168.61405	1356.50	<.0001
Error	10788	1340.95554	0.12430		
Corrected Total	10798	3027.09607			

Root MSE	0.35256	R-Square	0.5570
Dependent Mean	13.04838	Adj R-Sq	0.5566
Coeff Var	2.70197		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21.88743	0.29493	74.21	<.0001
logBedroom	1	0.00342	0.01456	0.23	0.8142
bathrooms	1	0.20044	0.00714	28.09	<.0001
floors	1	0.18193	0.00891	20.43	<.0001
waterfront	1	0.41398	0.04405	9.40	<.0001
view	1	0.06604	0.00520	12.70	<.0001
basement	1	0.13443	0.00809	16.62	<.0001
renovated	1	0.04460	0.01813	2.46	0.0139
yr_built	1	-0.00512	0.00015231	-33.64	<.0001
sqft_living15	1	0.00036113	0.00000681	53.01	<.0001
logSqftLot15	1	-0.02597	0.00503	-5.16	<.0001

Figure B – ANOVA table with Grade polynomial and Condition

Number of Observations Read	10799
Number of Observations Used	10799

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1980.34634	141.45331	1457.30	<.0001
Error	10784	1046.74973	0.09707		
Corrected Total	10798	3027.09607			

Root MSE	0.31155	R-Square	0.6542
Dependent Mean	13.04838	Adj R-Sq	0.6538
Coeff Var	2.38767		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	22.39947	0.43061	52.02	<.0001
logBedroom	1	-0.00874	0.01314	-0.67	0.5059
bathrooms	1	0.11565	0.00656	17.63	<.0001
floors	1	0.09959	0.00810	12.29	<.0001
waterfront	1	0.45034	0.03895	11.56	<.0001
view	1	0.04049	0.00463	8.75	<.0001
condition	1	0.04754	0.00509	9.34	<.0001
grade	1	0.07110	0.12607	0.56	0.5728
gradeSquare	1	0.02500	0.01534	1.63	0.1031
gradeCube	1	-0.00120	0.00061076	-1.97	0.0486
basement	1	0.10406	0.00722	14.42	<.0001
renovated	1	0.04297	0.01625	2.64	0.0082
yr_built	1	-0.00580	0.00014668	-39.54	<.0001
sqft_living15	1	0.00018109	0.00000700	25.87	<.0001
logSqftLot15	1	-0.03844	0.00449	-8.57	<.0001

Figure C – ANOVA table with Grade and Condition dummies

Number of Observations Read	10799
Number of Observations Used	10799

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	1984.08944	86.26476	891.18	<.0001
Error	10775	1043.00663	0.09680		
Corrected Total	10798	3027.09607			

Root MSE	0.31112	R-Square	0.6554
Dependent Mean	13.04838	Adj R-Sq	0.6547
Coeff Var	2.38440		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	23.16241	0.31015	74.68	<.0001
logBedroom	1	-0.01046	0.01316	-0.79	0.4269
bathrooms	1	0.11556	0.00659	17.55	<.0001
floors	1	0.09762	0.00812	12.02	<.0001
waterfront	1	0.45550	0.03897	11.69	<.0001
view	1	0.04035	0.00463	8.72	<.0001
condition_2	1	0.01291	0.09341	0.14	0.8901
condition_3	1	0.18330	0.08761	2.09	0.0364
condition_4	1	0.21542	0.08760	2.46	0.0139
condition_5	1	0.27781	0.08804	3.16	0.0016
grade_5	1	0.19265	0.10339	1.86	0.0624
grade_6	1	0.39903	0.09983	4.00	<.0001
grade_7	1	0.65135	0.09978	6.53	<.0001
grade_8	1	0.87969	0.10009	8.79	<.0001
grade_9	1	1.15180	0.10058	11.45	<.0001
grade_10	1	1.33013	0.10136	13.12	<.0001
grade_11	1	1.50991	0.10356	14.58	<.0001
grade_12	1	1.71287	0.11045	15.51	<.0001
grade_13	1	2.05301	0.15078	13.62	<.0001
basement	1	0.10486	0.00722	14.52	<.0001
renovated	1	0.03775	0.01627	2.32	0.0203
yr_built	1	-0.00587	0.00014798	-39.67	<.0001
sqft_living15	1	0.00018054	0.00000700	25.80	<.0001
logSqftLot15	1	-0.03755	0.00450	-8.34	<.0001

Figure D – Backward Selection ($\alpha = 0.05; 0.10; 0.15$)

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.6148 and C(p) = 13.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	9.725695E14	8.104746E13	1434.63	<.0001
Error	10786	6.093407E14	56493670397		
Corrected Total	10798	1.58191E15			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	7047043	214466	6.099552E13	1079.69	<.0001
bedrooms	-9017.92692	2874.84229	5.558853E11	9.84	0.0017
bathrooms	112867	4983.71191	2.897506E13	512.89	<.0001
floors	26177	5714.16063	1.185548E12	20.99	<.0001
waterfront	695468	29659	3.106379E13	549.86	<.0001
view	47956	3524.63189	1.045841E13	185.13	<.0001
condition	18257	3857.99446	1.265073E12	22.39	<.0001
grade	162364	3269.04159	1.393601E14	2466.83	<.0001
basement	31739	5356.14940	1.98371E12	35.11	<.0001
renovated	42591	12346	6.723748E11	11.90	0.0006
yr_built	-4191.97429	109.34278	8.303415E13	1469.80	<.0001
sqft_living15	88.38614	5.02392	1.748566E13	309.52	<.0001
sqft_lot15	-0.19487	0.08181	3.205085E11	5.67	0.0172

Figure E – Forward($\alpha = 0.05; 0.10; 0.15$)

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	grade	1	0.4341	0.4341	5051.97	8281.26	<.0001
2	yr_built	2	0.0753	0.5093	2946.30	1656.15	<.0001
3	bathrooms	3	0.0446	0.5540	1698.82	1079.93	<.0001
4	waterfront	4	0.0379	0.5918	639.870	1002.01	<.0001
5	sqft_living15	5	0.0121	0.6040	302.943	329.85	<.0001
6	view	6	0.0078	0.6117	86.9717	216.37	<.0001
7	basement	7	0.0008	0.6125	67.5398	21.31	<.0001
8	floors	8	0.0007	0.6133	48.5622	20.90	<.0001
9	condition	9	0.0006	0.6138	34.2018	16.32	<.0001
10	renovated	10	0.0004	0.6143	24.0232	12.16	0.0005
11	bedrooms	11	0.0003	0.6146	16.6734	9.35	0.0022
12	sqft_lot15	12	0.0002	0.6148	13.0000	5.67	0.0172

Figure F – Stepwise Selection($\alpha = 0.05; 0.10; 0.15$)

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	grade		1	0.4341	0.4341	5051.97	8281.26	<.0001
2	yr_built		2	0.0753	0.5093	2946.30	1656.15	<.0001
3	bathrooms		3	0.0446	0.5540	1698.82	1079.93	<.0001
4	waterfront		4	0.0379	0.5918	639.870	1002.01	<.0001
5	sqft_living15		5	0.0121	0.6040	302.943	329.85	<.0001
6	view		6	0.0078	0.6117	86.9717	216.37	<.0001
7	basement		7	0.0008	0.6125	67.5398	21.31	<.0001
8	floors		8	0.0007	0.6133	48.5622	20.90	<.0001
9	condition		9	0.0006	0.6138	34.2018	16.32	<.0001
10	renovated		10	0.0004	0.6143	24.0232	12.16	0.0005
11	bedrooms		11	0.0003	0.6146	16.6734	9.35	0.0022
12	sqft_lot15		12	0.0002	0.6148	13.0000	5.67	0.0172

Appendix D - Program

Program to create: To create a separate sample file from the random sample

```
/* proc sort by id to prepare sample merge */

    proc sort data= work.random_sample_obs;

        by observation;

    run;

    proc sort data= work.observed_house_v6_0000;

        by observation;

    run;

/*Seperate random sample from the main sample, and keep the remaining */

    data work.PREDICTION_SAMPLE_OBS(keep= observation    price   bedrooms

        bathrooms    floors  waterfront    view   condition    condition_1

        condition_2  condition_3  condition_4  condition_5  grade  grade_3

        grade_4      grade_5      grade_6      grade_7      grade_8

        grade_9      grade_10     grade_11     grade_12     grade_13

        basement     renovated    yr_built     sqft_living15 sqft_lot15

    );

     merge  work.random_sample_obs (in= Randsample_obs)

            work.observed_house_v6_0000 (in= KcHouse_obs);

        by observation;

    if KcHouse_obs and not Randsample_obs;

     run;
```

Problem encountered

We encountered a problem running the SAS, EG linear regression task on our linear regression. ODS Graphic suppresses 5000 points, which causes the regression task not to show residual plots. As such, we had to change our dummy variable linear regression source code:

- In Proc Reg, we change "Plot (ONLY)=ALL" to "Plot (MAXPOINTS=NONE)," which fixes the error. Now, we can see the residual plots.

Appendix E - Variables Kept

Variables Kept	**Description of Variables**	**Variables Kept**	**Description of Variables**
Price	Price of the house	Yr_built	Year the house was built
Bedrooms	Number of bedrooms	Sqft_living15	Current size of the house
Bathrooms	Number of bathrooms	Sqft_lot15	Current size of Lot
Floors	House's number of floors	Grade	King County's real estate grading scale (1-13) from inferior to excellent.
Waterfront	Waterfront view? Yes/No	Condition	House's condition 1-5s
View	Number of people that viewed the house		

Appendix F – Kaggle Dataset

Variable	Description of Variable	Variable	Description of Variable
Price	Price of the house	**Waterfront**	House's waterfront view
ID	House's ID	**View**	Number of people view the house
Date	House' sale date	**Condition**	House's condition inferior to excellent (1-5)
Lat	Latitude coordinate	**Long**	Longitude coordinate
Bedrooms	Number of bedrooms within the house	**Sqft_lot15**	New lot size after renovation in 2015
Bathrooms	Number of bathrooms within the house	**Sqft_above**	House's square footage excluding the basement
Floors	House's total square footage	**Sqft_basement**	House's basement square footage
Sqft_lot	House's lot square footage	**Yr_built**	The year when the house was built
Sqft_living	House's total floors	**Yr_renovated**	The year when the house was renovated

Sqft_living15	New living room square footage after renovation in 2015	**Grade**	King County's real estate grading scale (1-13) from inferior to excellent.