Project: Housing Prices Analysis

Albert Joseph, abj37
Marc Yarkony, mdy7

1.
    a.

Quantitative	Qualitative
price ($1000) bedrooms bathrooms sqft_living floors sqft_above yr_built yr_renovated latitude longitude sqft_living15	Id view condition grade

    b. Price: The distribution is right-skewed. As shown in Figure 1.
    c. As shown in Figure 2.
    bedrooms: The distribution is right-skewed.
    bathrooms: The distribution is left-skewed.
    sqft_living: The distribution is right-skewed.
    floors: The distribution is right-skewed.
    sqft_above: The distribution is right-skewed.
    yr_built: The distribution is left-skewed.
    yr_renovated: The distribution is right-skewed.
    latitude: The distribution is normal.
    longitude: The distribution is normal.
    sqft_living15: The distribution is right-skewed.
    view: The distribution is right-skewed (zero has the highest frequency).
    condition: The distribution is somewhat normal (three has the highest frequency).
    grade: The distribution is normal.
    d. There seems to be a positive correlation (>.5) between price and sqft_living, grade, sqft living15. There seems to be a positive correlation between bedroom and sqft_living, There seems to be a positive correlation between bathrooms and sqft_living, floors, grade, sqft_above, yr_built, sqft_living15. Overall, as sqft_living increased, other variables associated with area increased as well (price, bathroom, etc.). As shown in Figure 3, Figure 4, and Figure 5.

e. To predict price, we should use sqft_living, grade, and sqft living15 because there is a positive correlation (>.5) between these variables. Correlation between variables is shown in Figure 3.

2.

a. At 95% confidence, the true average selling price of King County houses lies between [462.1161,529.5443]. At 95% confidence, the true average square footage of King County houses lies between [1933.71,2158.041]. As shown in Figure 6.

b. The P-value for this test is .0054. Since P<alpha, we reject the null hypothesis and conclude that houses with basements have a higher average selling price than houses without basements. As shown in Figure 7.

c. CI: [21.419,161.857] Since 0 is outside of CI, we reject the null hypothesis and conclude that houses with basements have a higher average selling price than houses without basements. As shown in Figure 7.

d. Since there is overlap between the box plots, we cannot make a conclusion about the difference in means. As shown in Figure 8.

3.

a.

i) Yes, there is a relationship between the Predictor and the Response. (See Figure 9)

ii) The strength of relationship between the predictor and the response can be measured by the correlation between the two. The correlation between the predictor and the response is 0.674729. (See Figure 9)

iii) The relationship between the predictor and the response is positive, evident from its positive correlation and its positive slope. (See Figure 9)

iv) Based on the linear fit line from our graph, b1 returned a value of $0.2028616 (In $1000). Therefore, $0.2028616 (In $1000) is the cost for each additional square foot. (See Figure 9)

v) The predicted selling price for a 2000 square foot home is $486.517691 (In $1000). The 95% Confidence Interval for a selling price for a 2000 square foot house is [461.53, 511.5]. The 95% prediction Interval for selling price for a 2000 square foot home is [122.48, 859.55]. (See Figure 19)

b. There are a few problems we noticed while analyzing the diagnostic plots of the least squares regression fit. The residual by predicted plot looked like a funnel plot, inferring that the variability of variables is different across points of the graph. For the actual by predicted plot, the plotted points were slightly lower than the preferred angle of 45 degrees. Lastly, the residual Normal Quantile plot had a handful of plots outside of the normal distribution lines.

c. **Quadratic fit - R^2, R^2_adj = (0.457387, 0.452194)**
Price ($1000) = 89.93 + 0.194*sqft_living + 1.28e^-5*(sqft_living)^2. (See Figure 10)
Linear fit with logarithm transformation - R^2, R^2_adj = (0.4361, 0.4334)
Log(price ($1000)) = 5.3135436 + 0.0003804*sqft_living. (See Figure 11)
Linear Fit - R^2, R^2_adj = (0.45526, 0.452666)
price ($1000) = 80.794491 + 0.2028616*sqft_living. (See Figure 12)
Although the Quadratic fit has a slightly better R^2 and R^2_adj value, it is not the only important criterion when picking models. Of the 3 models, the linear fit with logarithmic transformation had a significantly better RMSE. When looking over the linear models on the graph, it becomes even more evident that this model clearly plots the points better than the other two.

d. **Linear regression models to predict price for each predictor**
Bedrooms - price ($1000) = 168.87516 + 95.474471*bedrooms
Bathrooms - price ($1000) = 119.01325 + 180.63356*bathrooms
Sqft_living - price ($1000) = 80.794491 + 0.2028616*sqft_living
Floors - price ($1000) = 296.20924 + 129.22028*floors
View - price ($1000) = 470.8305 + 120.45303*view
Condition - price ($1000) = 441.6712 + 16.263039*condition
Grade - price ($1000) = -682.9709 + 154.16769*grade
Sq_ft above - price ($1000) = 169.16239 + 0.1851492*sqft_above
Yr built - price ($1000) = -1162.812 + 0.8414649*yr_built
renovated - price ($1000) = 489.15857 + 0.0888655*yr_renovated
Latitude - price ($1000) = -38511.59 + 820.36334*latitude
Longitude - price ($1000) = 15389.831 + 121.86585*longitude
Sqft_living15 - price ($1000) = 113.79225 + 0.1934628*sqft_living15

With our models, the model with the statistically significant association between the response and the predictor was grade. Grade resulted in the highest correlation, R^2 value, R^2_adj value, and the lowest RMSE value.
(See Figure 13)

4.

a. i) There is definitely a relationship between the predictors and the response with the R^2 value being higher than all other single predictor options and the RMSE lower than all other single predictor options. (See Figure 14)
ii) The top predictors that seem to have a statistically significant relationship with the response are Latitude, Grade, view, year built, bathrooms and conditions (See Figure 14)
iii) Because of the large coefficient value for "Latitude" the variable suggests that the response is significantly affected by each individual change in the predictor value. It is also a positive line based on the coefficient. (See Figure 14)

iv) The coefficient value for “grade” suggests a positive line as well at a significant slope value, causing a significant difference for each change in predictor value. (See Figure 14)

b. Based on our plotted graph there are many more residual outliers plotted compared to when we plot the response to only one predictor. (See Figure 14)

c. Our results from 4A resulted in a higher R^2 value and a lower RMSE than any plotted graph from 3D. When all predictors are taken into account for one plot it clearly makes a difference.
(Compare Rsquare of Figure 14 to every Rsquare in Figure 13)

d. A few interactions that appear to be statistically significant are sqft_living*Grade, Grade*Latitude, Grade*View, and Sqft_Living*bedroom. The interactions are clear based on the non parallel lines. (See Figure 20)

e. The transformations done were able to improve RMSE scores and R^2 values by a small margin. (See Figures 15-17)

f. We narrowed down our Final Model predictors to the significant variables which were Latitude, grade, view, yr built, bathrooms and conditions. The log(x) transformation gave the best results compared to other transformations. Putting all of this together, our simplified model consisting of 6 predictors resulted in a R^2 value of .79, and a RMSE value of 0.2242. (See Figure 18)

5.

a. Validation Column created in JMP file. (See JMP File, Last column)

b. A few variables that help understand the association between houses built before 1980 and the variables are Floors, Grade, and Bathroom. Floors being the best indicator of the three. (See Figure 21)

c. Completed, See JMP file. Split up the proportion of Training and Test as .8 to .2 (80% to 20%).

d. The test error (Misclassification) of the model obtained is 0.1887 (See Figure 23)

e. The value of K that seemed to perform the best data was when there was a range of k from 1-75 and the K value was k=75. This resulted in a misclassification of .16327 (See Figure 22)

f. Our models returned slightly better classification results for the Logistic Regression Method instead of the K-Nearest neighbor method. Therefore, we would recommend using the Logistic Regression classification model to predict whether or not a house was built before 1980.

Bonus:

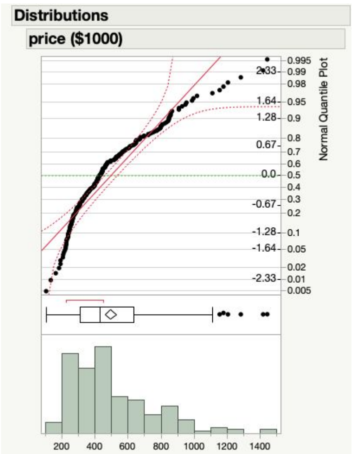
For the extra credit, we decided to put ourselves in the shoes of perspective home buyer in the King County area, Washington State, between May 2014 and May 2015. As a home buyer, there are a few factors that are important: bedrooms, bathrooms, sqft_living, floors, sqft_above, yr_built, yr_renovated , and sqft_living15. Since there are hundreds of houses to choose from, we performed cluster analysis to identify which houses are similar to each other, so that home buyers could fit themselves into a cluster, and then decide which houses to buy from there.

We decided to use Kmeans clustering because that was the easiest method for us to understand and apply. In order to determine how many clusters to use, we ran a factor analysis on the previously mentioned variables in JMP. From the elbow plot, we learned that two clusters would be sufficient. The factor analysis is shown in Bonus Figure 1.

Next, we ran the Kmeans cluster JMP, and found that in cluster one there would be 75 homes, and a cluster two there would be 137 homes. The average price for cluster one is $689.893 and the average price for cluster two is $389.591. Cluster one would be best for a home buyer with a higher budget who wants a newer home in larger living space. Cluster two would be best for a home buyer with a lower budget who wants a smaller living space, and older home. The Kmeans cluster results are shown in Bonus Figure 2.

Appendix:

Figure 1:



Quantiles

100.0%	maximum	1445
99.5%		1443.375
97.5%		1172.525
90.0%		850
75.0%	quartile	632.5
50.0%	median	432.5
25.0%	quartile	309.25
10.0%		244.3
2.5%		194
0.5%		108.82
0.0%	minimum	107

Summary Statistics

Mean	495.83019
Std Dev	249.01944
Std Err Mean	17.102725
Upper 95% Mean	529.54429
Lower 95% Mean	462.11609
N	212

Figure 2:

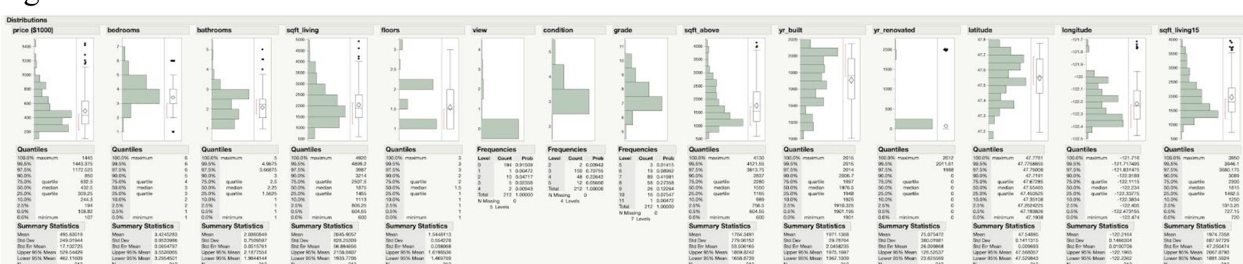


Figure 3:

Multivariate

Correlations

	price ($1000)	bedrooms	bathrooms	sqft_living	floors	view	condition	grade	sqft_above	yr_built	yr_renovated	latitude	longitude	sqft_living15
price ($1000)	1.0000	0.3655	0.5447	0.6747	0.2876	0.3475	0.0389	0.6931	0.5792	0.1007	0.1356	0.4649	0.0718	0.5345
bedrooms	0.3655	1.0000	0.5230	0.6235	0.1611	0.0714	-0.0811	0.3456	0.5263	0.2080	-0.0100	-0.0296	0.1740	0.3534
bathrooms	0.5447	0.5230	1.0000	0.7609	0.5187	0.1117	-0.2042	0.6522	0.7037	0.5179	0.0116	-0.0047	0.2689	0.5041
sqft_living	0.6747	0.6235	0.7609	1.0000	0.2958	0.3137	-0.1478	0.7595	0.8611	0.3865	-0.0246	0.0898	0.3349	0.7544
floors	0.2876	0.1611	0.5187	0.2958	1.0000	-0.0354	-0.2675	0.4343	0.5129	0.5170	-0.0155	0.0551	0.1174	0.1844
view	0.3475	0.0714	0.1117	0.3137	-0.0354	1.0000	0.0938	0.2921	0.1355	0.0228	0.1154	-0.0600	-0.0276	0.3206
condition	0.0389	-0.0811	-0.2042	-0.1478	-0.2675	0.0938	1.0000	-0.2006	-0.2268	-0.4245	0.0557	0.0076	-0.1727	-0.0770
grade	0.6931	0.3456	0.6522	0.7595	0.4343	0.2921	-0.2006	1.0000	0.7499	0.5244	-0.0476	0.1333	0.2980	0.6879
sqft_above	0.5792	0.5263	0.7037	0.8611	0.5129	0.1355	-0.2268	0.7499	1.0000	0.4986	-0.0526	0.0310	0.4253	0.6761
yr_built	0.1007	0.2080	0.5179	0.3865	0.5170	0.0228	-0.4245	0.5244	0.4986	1.0000	-0.2724	-0.2059	0.4560	0.3507
yr_renovated	0.1356	-0.0100	0.0116	-0.0246	-0.0155	0.1154	0.0557	-0.0476	-0.0526	-0.2724	1.0000	0.0847	-0.0753	-0.0485
latitude	0.4649	-0.0296	-0.0047	0.0898	0.0551	-0.0600	0.0076	0.1333	0.0310	-0.2059	0.0847	1.0000	-0.2177	0.1260
longitude	0.0718	0.1740	0.2689	0.3349	0.1174	-0.0276	-0.1727	0.2980	0.4253	0.4560	-0.0753	-0.2177	1.0000	0.4598
sqft_living15	0.5345	0.3534	0.5041	0.7544	0.1844	0.3206	-0.0770	0.6879	0.6761	0.3507	-0.0485	0.1260	0.4598	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix

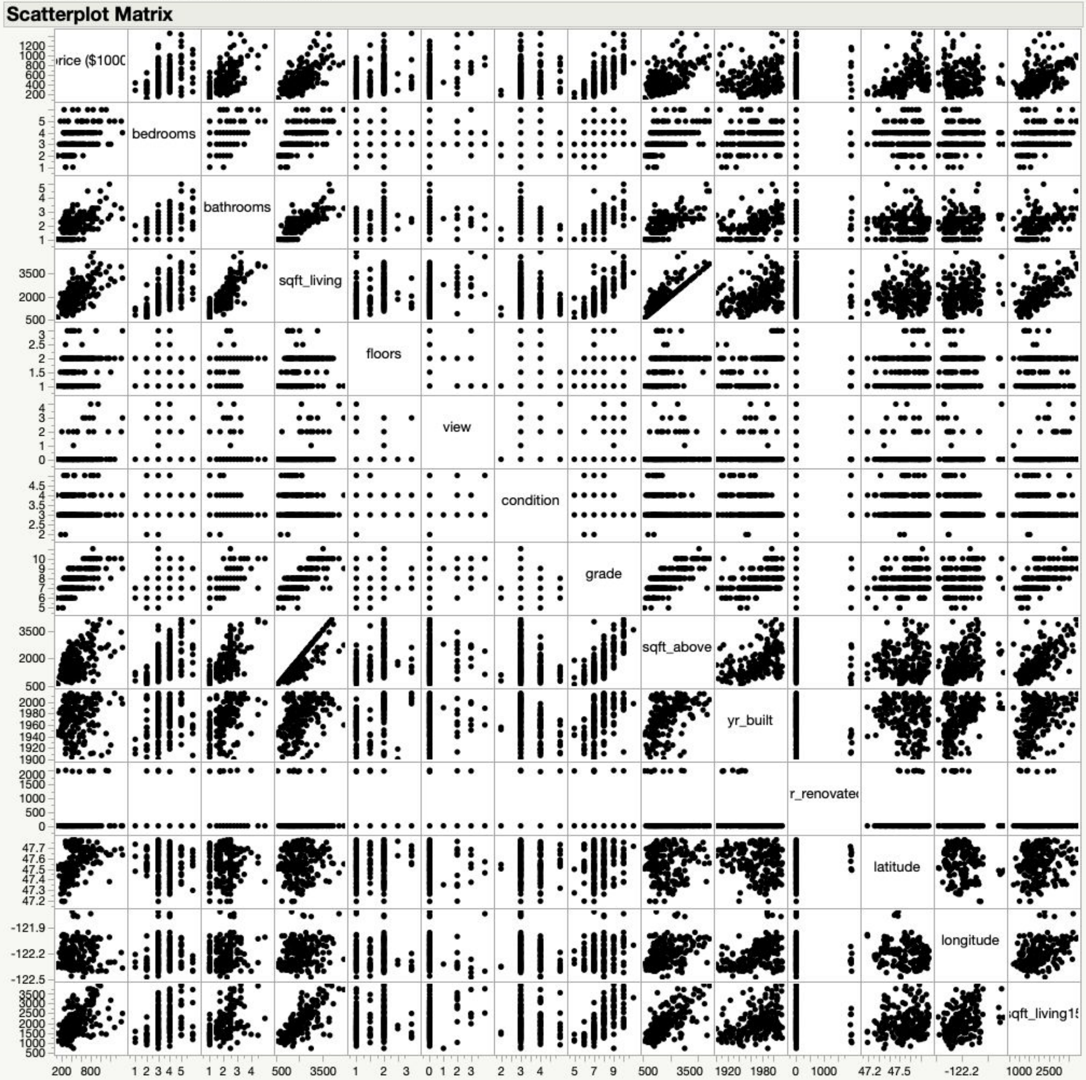


Figure 4:

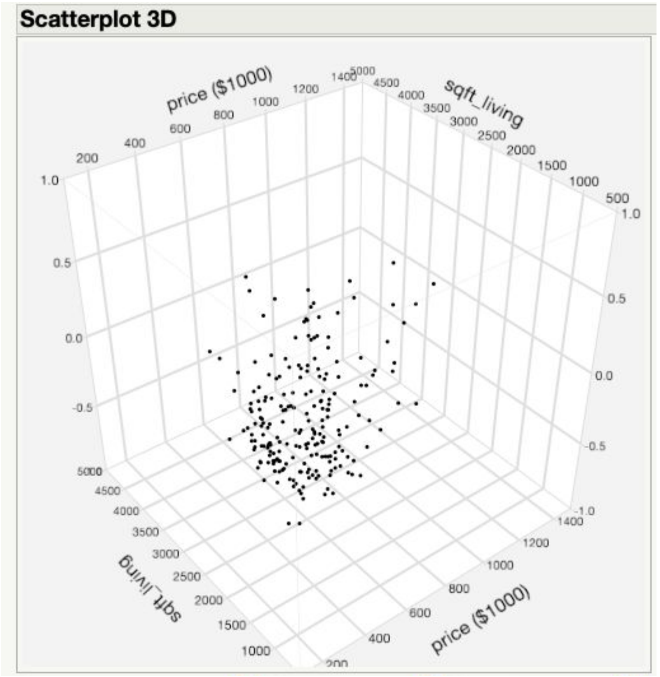


Figure 5:

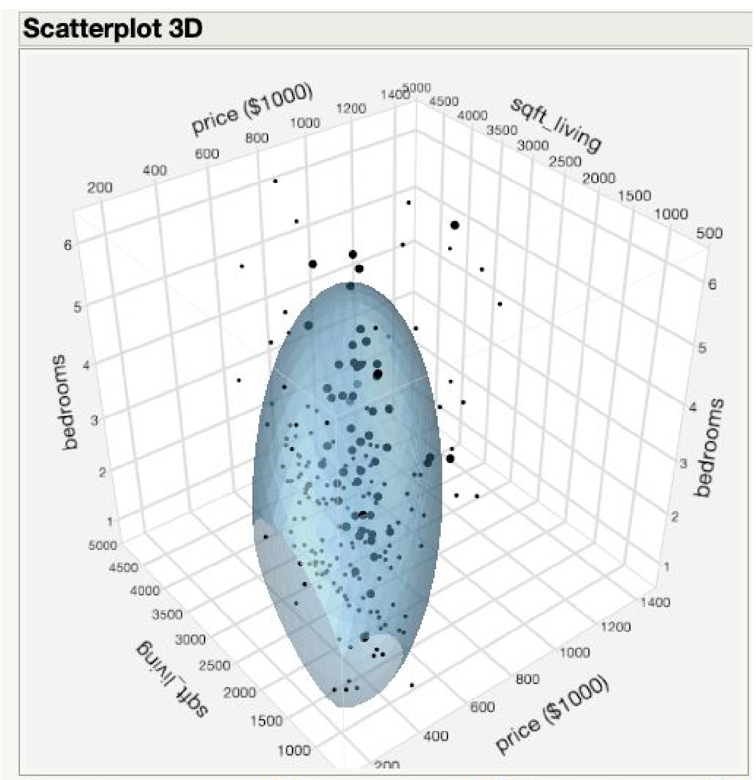


Figure 6:

Distributions

price ($1000)

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	495.8302	462.1161	529.5443	0.950
Std Dev	249.0194	227.3574	275.2796	0.950

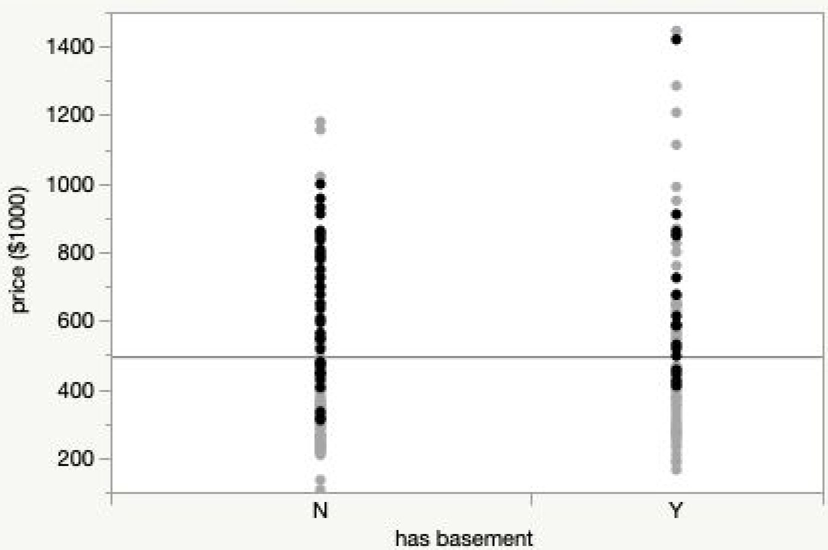
sqft_living

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	2045.906	1933.771	2158.041	0.950
Std Dev	828.2531	756.204	915.596	0.950

Figure 7:

Oneway Analysis of price ($1000) By has basement



t Test

Y-N

Assuming unequal variances

Difference	91.638	t Ratio	2.57732
Std Err Dif	35.556	DF	159.8881
Upper CL Dif	161.857	Prob > |t|	0.0109*
Lower CL Dif	21.419	Prob > t	0.0054*
Confidence	0.95	Prob < t	0.9946

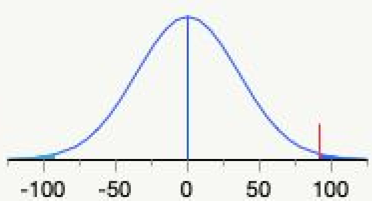
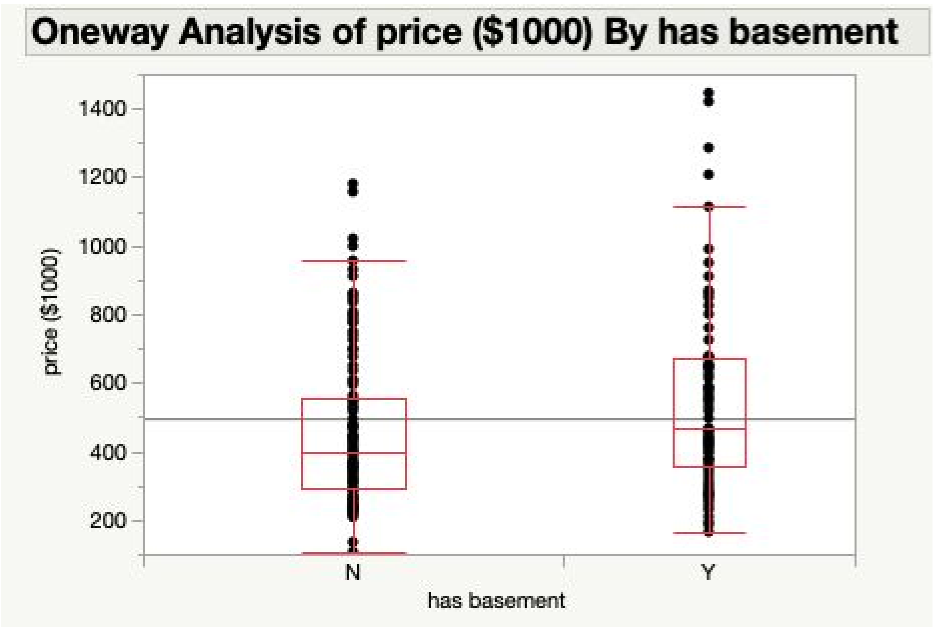


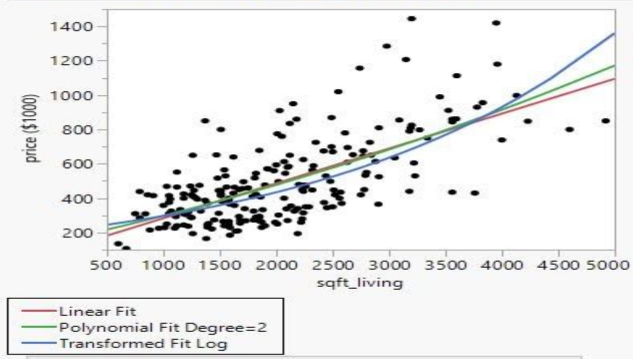
Figure 8:



Oneway Analysis of price ($1000) By has basement
price ($1000)
1400
1200
1000
800
600
400
200
N
Y
has basement

Figure 9

Bivariate Fit of price ($1000) By sqft_living



Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.674729	0.593985	0.742015	<.0001*
Covariance	139163.7			
Count	212			

Variable	Mean	Std Dev
sqft_living	2045.906	828.2531
price ($1000)	495.8302	249.0194

Polynomial Fit Degree=2

price ($1000) = 89.931119 + 0.1941187*sqft_living + 1.2816e-5*(sqft_living-2045.91)^2

Summary of Fit

RSquare	0.457387
RSquare Adj	0.452194
Root Mean Square Error	184.309
Mean of Response	495.8302
Observations (or Sum Wgts)	212

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	5984567	2992284	88.0866
Error	209	7099687	33970	Prob > F
C. Total	211	13084254		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	89.931119	35.27673	2.55	0.0115*
sqft_living	0.1941187	0.01811	10.72	<.0001*
(sqft_living-2045.91)^2	1.2816e-5	1.416e-5	0.91	0.3664

Transformed Fit Log

Log(price ($1000)) = 5.3135436 + 0.0003804*sqft_living

Summary of Fit

RSquare	0.436105
RSquare Adj	0.43342
Root Mean Square Error	0.359161
Mean of Response	6.091898
Observations (or Sum Wgts)	212

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	20.950389	20.9504	162.4100
Error	210	27.089360	0.1290	Prob > F
C. Total	211	48.039749		<.0001*

Parameter Estimates

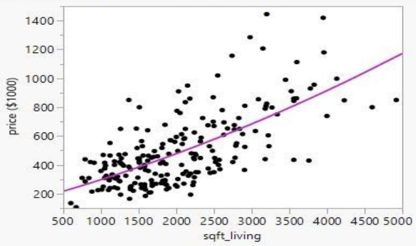
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.3135436	0.065869	80.67	<.0001*
sqft_living	0.0003804	2.985e-5	12.74	<.0001*

Fit Measured on Original Scale

Sum of Squared Error	7502167.3
Root Mean Square Error	189.00954
RSquare	0.4266263
Sum of Residuals	6174.4857

Figure 10

Bivariate Fit of price ($1000) By sqft_living



Polynomial Fit Degree=2

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.674729	0.593985	0.742015	<.0001*
Covariance	139163.7			
Count	212			

Variable	Mean	Std Dev
sqft_living	2045.906	828.2531
price ($1000)	495.8302	249.0194

Polynomial Fit Degree=2

price ($1000) = 89.931119 + 0.1941187*sqft_living + 1.2816e-5*(sqft_living-2045.91)^2

Summary of Fit

RSquare	0.457387
RSquare Adj	0.452194
Root Mean Square Error	184.309
Mean of Response	495.8302
Observations (or Sum Wgts)	212

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	5984567	2992284	88.0866
Error	209	7099687	33970	Prob > F
C. Total	211	13084254		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	89.931119	35.27673	2.55	0.0115*
sqft_living	0.1941187	0.01811	10.72	<.0001*
(sqft_living-2045.91)^2	1.2816e-5	1.416e-5	0.91	0.3664

Figure 11

Bivariate Fit of price ($1000) By sqft_living

Transformed Fit Log

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.674729	0.593985	0.742015	<.0001*
Covariance	139163.7			
Count	212			

Variable	Mean	Std Dev
sqft_living	2045.906	828.2531
price ($1000)	495.8302	249.0194

Transformed Fit Log

Log(price ($1000)) = 5.3135436 + 0.0003804*sqft_living

Summary of Fit

RSquare	0.436105
RSquare Adj	0.43342
Root Mean Square Error	0.359161
Mean of Response	6.091898
Observations (or Sum Wgts)	212

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	20.950389	20.9504	162.4100
Error	210	27.089360	0.1290	Prob > F
C. Total	211	48.039749		<.0001*

Parameter Estimates

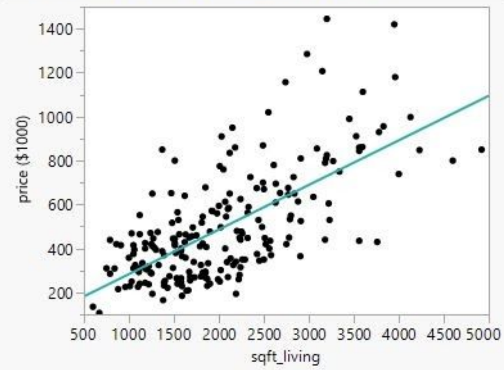
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.3135436	0.065869	80.67	<.0001*
sqft_living	0.0003804	2.985e-5	12.74	<.0001*

Fit Measured on Original Scale

Sum of Squared Error	7502167.3
Root Mean Square Error	189.00954
RSquare	0.4266263
Sum of Residuals	6174.4857

Figure 12

Bivariate Fit of price ($1000) By sqft_living



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.674729	0.593985	0.742015	<.0001*
Covariance	139163.7			
Count	212			

Variable	Mean	Std Dev
sqft_living	2045.906	828.2531
price ($1000)	495.8302	249.0194

Linear Fit

price ($1000) = 80.794491 + 0.2028616*sqft_living

Summary of Fit

RSquare	0.45526
RSquare Adj	0.452666
Root Mean Square Error	184.2297
Mean of Response	495.8302
Observations (or Sum Wgts)	212

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5956735	5956735	175.5049
Error	210	7127519	33941	Prob > F
C. Total	211	13084254		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	80.794491	33.78724	2.39	0.0177*
sqft_living	0.2028616	0.015313	13.25	<.0001*

Figure 13

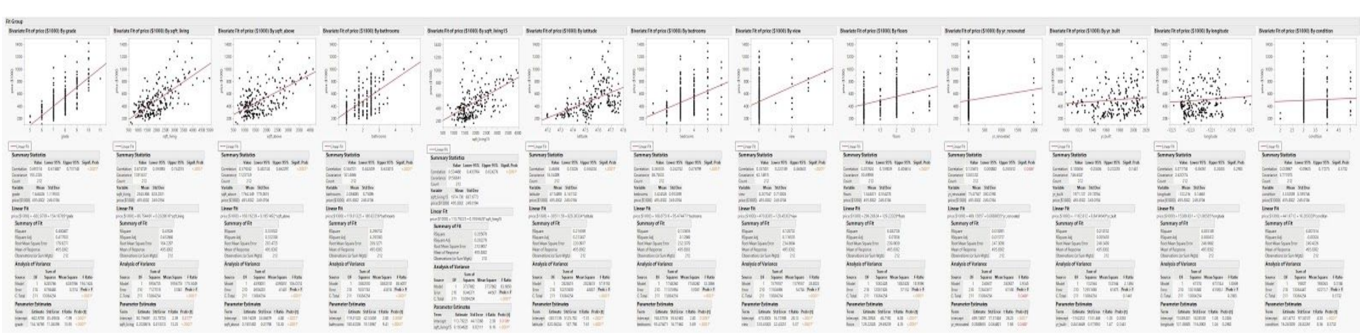
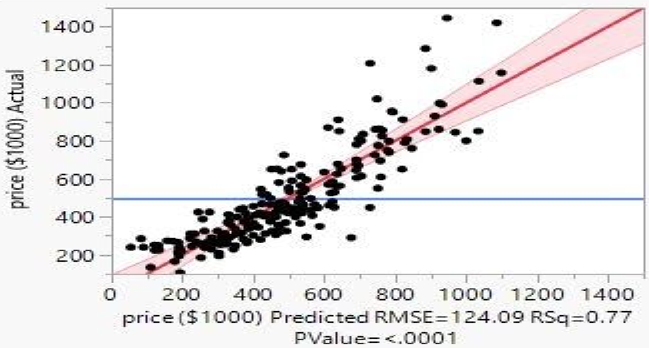


Figure 14)

Response price ($1000)

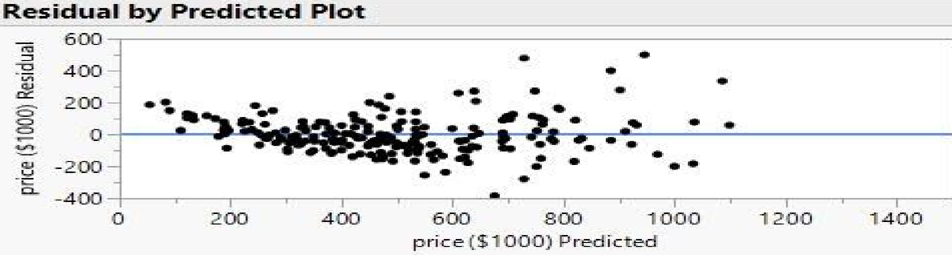
Actual by Predicted Plot



Effect Summary

Source	LogWorth	PValue
latitude	16.866	0.00000
grade	10.558	0.00000
view	4.605	0.00002
yr_built	3.760	0.00017
bathrooms	2.284	0.00520
condition	2.270	0.00537
sqft_living15	1.002	0.09960
sqft_living	0.662	0.21763
yr_renovated	0.644	0.22720
bedrooms	0.562	0.27435
sqft_above	0.470	0.33863
longitude	0.437	0.36580
floors	0.344	0.45303

Residual by Predicted Plot



Summary of Fit

RSquare	0.766997
RSquare Adj	0.751699
Root Mean Square Error	124.0859
Mean of Response	495.8302
Observations (or Sum Wgts)	212

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-19742.19	9827.396	-2.01	0.0459*
bedrooms	13.409008	12.23306	1.10	0.2744
bathrooms	61.87112	21.89329	2.83	0.0052*
sqft_living	0.0423391	0.034233	1.24	0.2176
floors	17.739773	23.59449	0.75	0.4530
view	62.513922	14.47616	4.32	<.0001*
condition	45.557019	16.18277	2.82	0.0054*
grade	101.83778	14.42699	7.06	<.0001*
sqft_above	0.0276489	0.028825	0.96	0.3386
yr_built	-1.813369	0.473814	-3.83	0.0002*
yr_renovated	0.0295795	0.024418	1.21	0.2272
latitude	656.64873	69.8416	9.40	<.0001*
longitude	70.540058	77.81957	0.91	0.3658
sqft_living15	-0.036851	0.022273	-1.65	0.0996

Figure 15

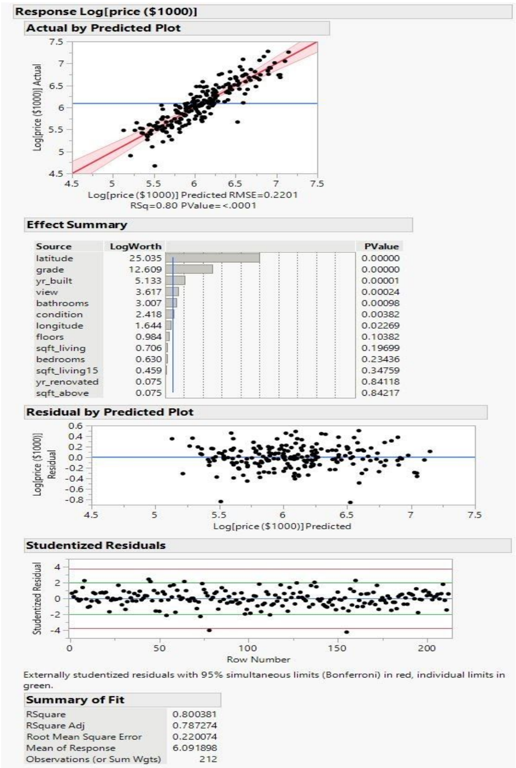


Figure 16

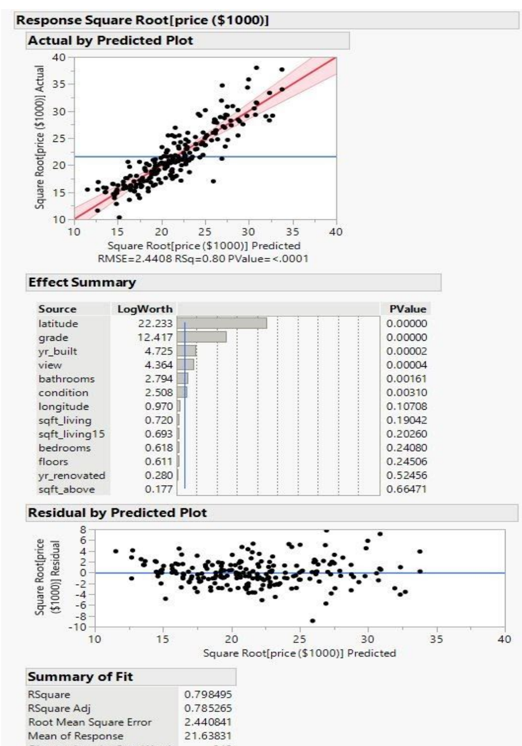
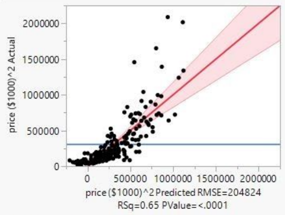


Figure 17

Response price ($1000)^2

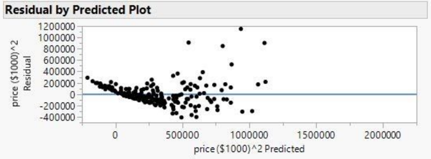
Actual by Predicted Plot



Effect Summary

Source	LogWorth	PValue
latitude	7.754	0.00000
grade	6.215	0.00000
view	4.350	0.00004
yr_built	1.969	0.01074
sqft_living15	1.652	0.02229
condition	1.604	0.02491
bathrooms	1.323	0.04752
yr_renovated	0.915	0.12153
sqft_above	0.819	0.15154
sqft_living	0.526	0.29808
bedrooms	0.433	0.36898
floors	0.157	0.69612
longitude	0.015	0.96626

Residual by Predicted Plot



Summary of Fit

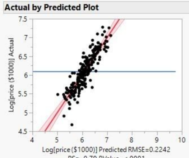
RSquare	0.648864
RSquare Adj	0.625809
Root Mean Square Error	204824
Mean of Response	307565.8
Observations (or Sum Wgts)	212

Figure 18

Response Log[price ($1000)]

Whole Model | bathrooms | view | condition | grade | yr_built | latitude

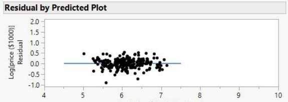
Actual by Predicted Plot



Effect Summary

Source	LogWorth	PValue
latitude	25.663	0.00000
grade	20.286	0.00000
bathrooms	10.490	0.00000
yr_built	4.954	0.00001
view	3.960	0.00011
condition	2.151	0.00706

Residual by Predicted Plot



Summary of Fit

RSquare	0.785475
RSquare Adj	0.779196
Root Mean Square Error	0.224214
Mean of Response	6.091898
Observations (or Sum Wgts)	212

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	37.734018	6.28900	125.0999
Error	205	10.305731	0.05027	Prob > F
C. Total	211	48.039749		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-59.43272	6.24748	-9.51	<.0001*
bathrooms	0.1984771	0.028283	7.02	<.0001*
view	0.0917753	0.023265	3.94	0.0001*
condition	0.0786871	0.028914	2.72	0.0071*
grade	0.2215552	0.021035	10.53	<.0001*
yr_built	-0.003272	0.000726	-4.51	<.0001*
latitude	1.4634342	0.119021	12.30	<.0001*

Prediction Expression

-59.43272001

+ 0.1984770878 • bathrooms

+ 0.0917752513 • view

+ 0.0786871108 • condition

+ 0.2215552215 • grade

+ -0.003271885 • yr_built

+ 1.4634341792 • latitude

Figure 19

Lower 95% Indiv price ($1000)	Upper 95% Indiv price ($1000)	Lower 95% Mean price ($1000)	Upper 95% Mean price ($1000)
71.660693359	799.94389215	409.30807354	462.29651197
126.54205733	854.60779256	465.61960595	515.53024395
-50.94988683	679.12055147	277.29470221	350.87596244
-75.57951088	655.06339135	250.21218068	329.27169979
158.97824912	887.087313	497.763403	548.30215913
161.00338362	889.11941052	499.74191051	550.38088363
81.837670676	810.0530749	419.9211921	471.96955348
245.83320475	974.693334	580.06208983	640.46444891
-75.57951088	655.06339135	250.21218068	329.27169979
-110.5325065	621.0434427	211.6242683	298.88666792
81.837670676	810.0530749	419.9211921	471.96955348
22.724211935	751.50680522	357.38578754	416.84522962
-42.74795233	687.14754503	286.28742366	358.11216904
368.25559273	1099.7620989	690.669702	777.34798962
306.1564177	1036.0870815	635.03113138	707.2123678
414.17583404	1147.1581939	731.48750386	829.84652409
223.65847994	952.23850664	559.48706205	616.40992453
85.906710583	814.09849902	424.14415491	475.8610547
370.25483428	1101.8200894	692.45111244	779.62381119
87.940855267	816.12158635	426.25073353	477.81170809
-124.9456575	607.05596967	195.67415998	286.43615215
436.09176605	1169.8718141	750.9132875	855.05029261
49.249338429	777.72569492	385.69741531	441.27761804
24.766104459	753.52214472	359.577576	418.71067318
63.514609822	791.86104763	400.76251335	454.6131441
6.3800894275	735.39307162	339.77675421	401.99640684
122.48302778	850.55235809	461.53616625	511.49921962
280.04386999	1009.455613	611.38491041	678.11457258
26.80774738	755.53773381	361.767171	420.57831019
34.971822506	763.60258674	370.50288789	428.07152136
-1.797956925	727.34218991	330.92596012	394.61827287
-57.10394456	673.10291316	270.5381308	345.4608378

Figure 20

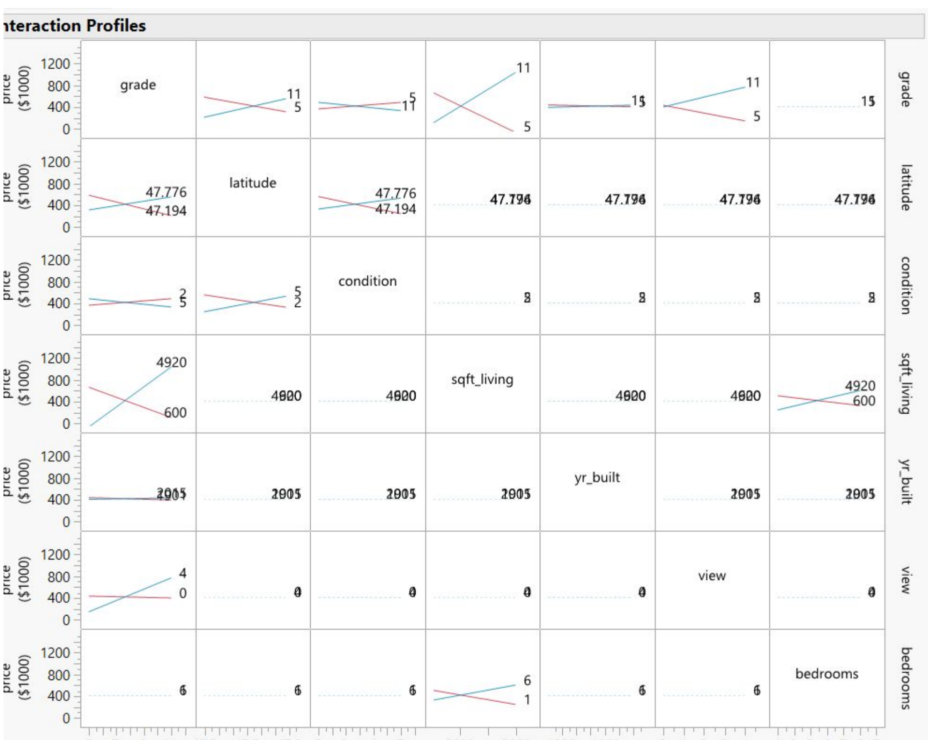


Figure 21

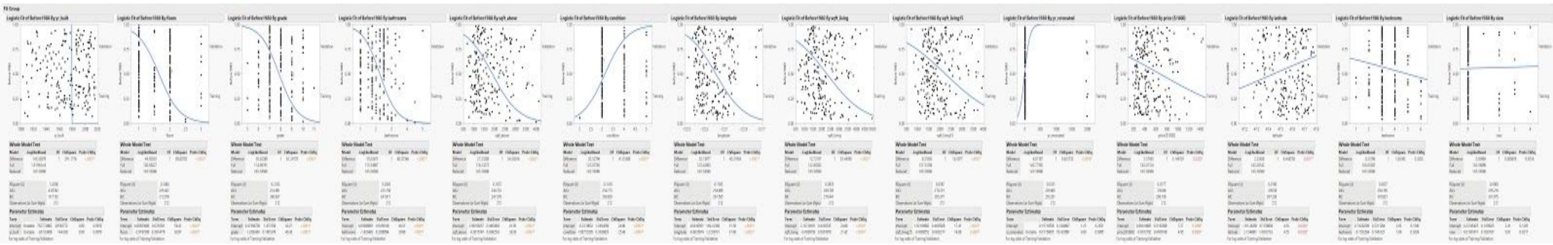
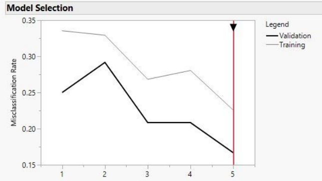


Figure 22

K Nearest Neighbors

Validation

Model Selection



Training				Validation			
K	Count	Misclassification Rate	Misclassifications	K	Count	Misclassification Rate	Misclassifications
1	164	0.33537	55	1	48	0.25000	12
2	164	0.32927	54	2	48	0.29167	14
3	164	0.26829	44	3	48	0.20833	10
4	164	0.28049	46	4	48	0.20833	10
5	164	0.22561	37 *	5	48	0.16667	8 *

Confusion Matrix for Best K=5

Training

Actual	Predicted Count	
Validation	Training	Validation
Training	122	6
Validation	31	5

Validation

Actual	Predicted Count	
Validation	Training	Validation
Training	40	2
Validation	6	0

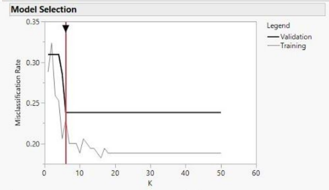
Mosaic Plot

Mosaic Plot for K=5

K Nearest Neighbors

Validation

Model Selection



Training				Validation			
K	Count	Misclassification Rate	Misclassifications	K	Count	Misclassification Rate	Misclassifications
1	170	0.28824	49	1	42	0.30952	13
2	170	0.32353	55	2	42	0.30952	13
3	170	0.25882	44	3	42	0.30952	13
4	170	0.25294	43	4	42	0.30952	13
5	170	0.20588	35	5	42	0.28571	12
6	170	0.22941	39	6	42	0.23810	10 *
7	170	0.20000	34	7	42	0.23810	10
8	170	0.20000	34	8	42	0.23810	10
9	170	0.20000	34	9	42	0.23810	10
10	170	0.18824	32	10	42	0.23810	10
11	170	0.20588	35	11	42	0.23810	10
12	170	0.20000	34	12	42	0.23810	10
13	170	0.19412	33	13	42	0.23810	10
14	170	0.19412	33	14	42	0.23810	10
15	170	0.18824	32	15	42	0.23810	10
16	170	0.18235	31 *	16	42	0.23810	10
17	170	0.19412	33	17	42	0.23810	10
18	170	0.18824	32	18	42	0.23810	10
19	170	0.18824	32	19	42	0.23810	10
20	170	0.18824	32	20	42	0.23810	10
21	170	0.18824	32	21	42	0.23810	10
22	170	0.18824	32	22	42	0.23810	10
23	170	0.18824	32	23	42	0.23810	10
24	170	0.18824	32	24	42	0.23810	10
25	170	0.18824	32	25	42	0.23810	10
26	170	0.18824	32	26	42	0.23810	10
27	170	0.18824	32	27	42	0.23810	10
28	170	0.18824	32	28	42	0.23810	10
29	170	0.18824	32	29	42	0.23810	10
30	170	0.18824	32	30	42	0.23810	10
31	170	0.18824	32	31	42	0.23810	10
32	170	0.18824	32	32	42	0.23810	10
33	170	0.18824	32	33	42	0.23810	10
34	170	0.18824	32	34	42	0.23810	10
35	170	0.18824	32	35	42	0.23810	10
36	170	0.18824	32	36	42	0.23810	10
37	170	0.18824	32	37	42	0.23810	10
38	170	0.18824	32	38	42	0.23810	10
39	170	0.18824	32	39	42	0.23810	10
40	170	0.18824	32	40	42	0.23810	10
41	170	0.18824	32	41	42	0.23810	10
42	170	0.18824	32	42	42	0.23810	10
43	170	0.18824	32	43	42	0.23810	10
44	170	0.18824	32	44	42	0.23810	10
45	170	0.18824	32	45	42	0.23810	10
46	170	0.18824	32	46	42	0.23810	10
47	170	0.18824	32	47	42	0.23810	10
48	170	0.18824	32	48	42	0.23810	10
49	170	0.18824	32	49	42	0.23810	10
50	170	0.18824	32	50	42	0.23810	10

Confusion Matrix for Best K=6

Training

Actual	Predicted Count	
Validation	Training	Validation
Training	129	9
Validation	30	2

Validation

Actual	Predicted Count	
Validation	Training	Validation
Training	32	0
Validation	10	0

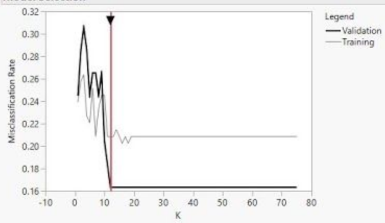
Mosaic Plot

Mosaic Plot for K=6

K Nearest Neighbors

Validation

Model Selection



Training

K	Count	Misclassification Rate	Misclassifications
1	163	0.23926	39
2	163	0.25767	42
3	163	0.26380	43
4	163	0.22699	37
5	163	0.22086	36
6	163	0.25153	41
7	163	0.20859	34
8	163	0.23313	38
9	163	0.24540	40
10	163	0.24540	40
11	163	0.20859	34
12	163	0.20859	34
13	163	0.20859	34
14	163	0.21472	35
15	163	0.20859	34
16	163	0.20245	33 *
17	163	0.20859	34
18	163	0.20245	33
19	163	0.20859	34
20	163	0.20859	34
21	163	0.20859	34
22	163	0.20859	34
23	163	0.20859	34
24	163	0.20859	34
25	163	0.20859	34
26	163	0.20859	34
27	163	0.20859	34
28	163	0.20859	34
29	163	0.20859	34
30	163	0.20859	34
31	163	0.20859	34
32	163	0.20859	34
33	163	0.20859	34
34	163	0.20859	34
35	163	0.20859	34
36	163	0.20859	34
37	163	0.20859	34
38	163	0.20859	34
39	163	0.20859	34
40	163	0.20859	34
41	163	0.20859	34
42	163	0.20859	34
43	163	0.20859	34
44	163	0.20859	34
45	163	0.20859	34
46	163	0.20859	34
47	163	0.20859	34
48	163	0.20859	34
49	163	0.20859	34
50	163	0.20859	34
51	163	0.20859	34
52	163	0.20859	34
53	163	0.20859	34
54	163	0.20859	34
55	163	0.20859	34
56	163	0.20859	34
57	163	0.20859	34
58	163	0.20859	34
59	163	0.20859	34
60	163	0.20859	34
61	163	0.20859	34
62	163	0.20859	34
63	163	0.20859	34
64	163	0.20859	34
65	163	0.20859	34
66	163	0.20859	34
67	163	0.20859	34
68	163	0.20859	34
69	163	0.20859	34
70	163	0.20859	34
71	163	0.20859	34
72	163	0.20859	34
73	163	0.20859	34
74	163	0.20859	34
75	163	0.20859	34

Validation

K	Count	Misclassification Rate	Misclassifications
1	49	0.24490	12
2	49	0.28571	14
3	49	0.30612	15
4	49	0.28571	14
5	49	0.24490	12
6	49	0.26531	13
7	49	0.26531	13
8	49	0.24490	12
9	49	0.26531	13
10	49	0.20408	10
11	49	0.18367	9
12	49	0.16327	8 *
13	49	0.16327	8
14	49	0.16327	8
15	49	0.16327	8
16	49	0.16327	8
17	49	0.16327	8
18	49	0.16327	8
19	49	0.16327	8
20	49	0.16327	8
21	49	0.16327	8
22	49	0.16327	8
23	49	0.16327	8
24	49	0.16327	8
25	49	0.16327	8
26	49	0.16327	8
27	49	0.16327	8
28	49	0.16327	8
29	49	0.16327	8
30	49	0.16327	8
31	49	0.16327	8
32	49	0.16327	8
33	49	0.16327	8
34	49	0.16327	8
35	49	0.16327	8
36	49	0.16327	8
37	49	0.16327	8
38	49	0.16327	8
39	49	0.16327	8
40	49	0.16327	8
41	49	0.16327	8
42	49	0.16327	8
43	49	0.16327	8
44	49	0.16327	8
45	49	0.16327	8
46	49	0.16327	8
47	49	0.16327	8
48	49	0.16327	8
49	49	0.16327	8
50	49	0.16327	8
51	49	0.16327	8
52	49	0.16327	8
53	49	0.16327	8
54	49	0.16327	8
55	49	0.16327	8
56	49	0.16327	8
57	49	0.16327	8
58	49	0.16327	8
59	49	0.16327	8
60	49	0.16327	8
61	49	0.16327	8
62	49	0.16327	8
63	49	0.16327	8
64	49	0.16327	8
65	49	0.16327	8
66	49	0.16327	8
67	49	0.16327	8
68	49	0.16327	8
69	49	0.16327	8
70	49	0.16327	8
71	49	0.16327	8
72	49	0.16327	8
73	49	0.16327	8
74	49	0.16327	8
75	49	0.16327	8

Confusion Matrix for Best K=12

Training

Actual	Predicted Count	
Validation	Training	Validation
Training	128	1
Validation	33	1

Validation

Actual	Predicted Count	
Validation	Training	Validation
Training	41	0
Validation	8	0

Mosaic Plot

Mosaic Plot for K=12

Training

Validation

Figure 23

Nominal Logistic Fit for Validation

Effect Summary

Source	LogWorth	PValue
floors	1.643	0.02274
grade	1.183	0.06559
bathrooms	0.699	0.20000

Converged in Gradient, 4 iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	4.86199	3	9.723985	0.0211*
Full	100.66644			
Reduced	105.52843			

RSquare (U)	0.0461
AICc	209.526
BIC	222.759
Observations (or Sum Wgts)	212

Fit Details

Measure	Training	Definition
Entropy RSquare	0.0461	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.0711	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4748	$\sum -\mathrm{Log}(\rho[j])/n$
RASE	0.3861	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3002	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.1887	$\sum (\rho[j]\neq\rho\mathrm{Max})/n$
N	212	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	66	32.30285	64.6057
Saturated	69	68.36359	**Prob>ChiSq**
Fitted	3	100.66644	0.5255

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	0.65986349	1.3029309	0.26	0.6125
grade	0.38337537	0.2133067	3.23	0.0723
floors	-0.8128369	0.3574347	5.17	0.0230*
bathrooms	-0.4095722	0.3175913	1.66	0.1972

For log odds of Training/Validation

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
grade	1	1	3.39018117	0.0656
floors	1	1	5.18850528	0.0227*
bathrooms	1	1	1.64236263	0.2000

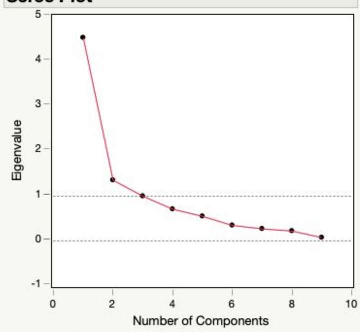
Bonus Figure 1:

Factor Analysis

Eigenvalues

Number	Eigenvalue	Percent	Cum Percent
1	4.5219	50.243	50.243
2	1.3480	14.978	65.222
3	0.9919	11.021	76.243
4	0.7026	7.807	84.050
5	0.5437	6.041	90.091
6	0.3413	3.792	93.883
7	0.2647	2.941	96.823
8	0.2163	2.404	99.227
9	0.0696	0.773	100.000

Scree Plot



Factor Analysis on Correlations with 2 Factors: Maximum Likelihood / Varimax

Final Communality Estimates

price ($1000)	0.46116
bedrooms	0.38925
sqft_living15	0.57037
bathrooms	0.68370
sqft_living	1.00000
sqft_above	0.82162
yr_built	0.35956
yr_renovated	0.00134
floors	0.90380

Variance Explained by Each Factor

Factor	Variance	Percent	Cum Percent
Factor 1	3.3478	37.198	37.198
Factor 2	1.8430	20.477	57.675

Significance Test

Test	DF	ChiSquare	Prob>ChiSq
H0: no common factors. HA: at least one common factor.	36	1193.977	<.0001*

Test	DF	Criterion	ChiSquare	Prob>ChiSq
H0: 2 factors are sufficient. HA: more factors are needed.	19	0.506	104.094	<.0001*

Measures of Fit

Measures of Fit	Fit Index
Chi-Square without Bartlett's Correction	106.707
AIC	68.707
BIC	4.932
Tucker and Lewis's Index	0.861
Root Mean Square Error of Approximation	0.148

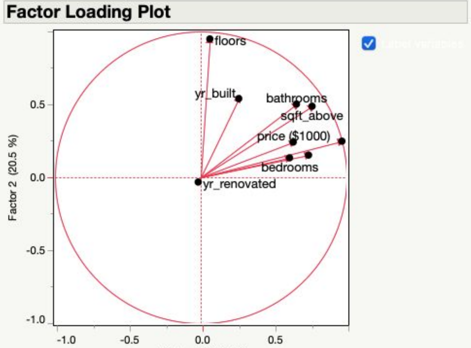
Rotated Factor Loading

	Factor 1	Factor 2
sqft_living	0.968822	0.247757
sqft_above	0.764158	0.487523
sqft_living15	0.739700	0.152368
bathrooms	0.656978	0.502076
price ($1000)	0.634660	0.241596
bedrooms	0.609437	0.133540
yr_renovated	-0.017157	-0.032365
floors	0.062697	0.948613
yr_built	0.260899	0.539896

Suppress Absolute Loading Value Less Than 0.3

Dim Text 0.4

Factor Loading Plot



Bonus Figure 2:

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	2	-2.267	Optimal CCC

Columns Scaled Individually

K Means NCluster=2

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	75
2	137

Step	Criterion
9	0

Cluster Means

Cluster	price ($1000)	bedrooms	bathrooms	sqft_living15	sqft_living	sqft_above	yr_built	yr_renovated	floors
1	689.893333	3.96	2.73	2602.50667	2912.82667	2621.76	1988.61333	79.8133333	1.87333333
2	389.591241	3.13138686	1.73357664	1631.06569	1571.31387	1294.9635	1961.56934	72.4817518	1.3649635