Min Yang

STAT2112

Project 2: Multiple Regression Analysis on Sale Price (in $1000)

Dr. Vadakkeveetil

I. Introduction.

Residence	Sales Price (in $1000)	Square feet	Rooms	Bedrooms	Age	View	Area x Age
2	49.0	1290	6	3	36	0	46440
4	49.9	912	5	3	41	0	37392
6	55.0	1204	5	3	10	0	12040
7	80.5	1764	8	4	64	0	112896
9	69.0	1255	5	3	16	0	20080
10	149.0	3600	10	5	17	1	61200
12	38.0	720	4	2	41	0	29520
13	49.5	1008	6	3	35	0	35280
16	85.0	2011	9	4	76	1	152836
18	58.5	1232	5	2	69	0	85008
19	101.0	1736	7	3	67	1	116312
21	125.0	1996	7	3	9	1	17964
23	80.0	1580	5	3	11	0	17380
25	74.0	1430	9	3	16	0	22880
26	69.0	1486	6	3	27	0	40122
28	67.5	1282	5	3	20	0	25640
31	92.2	1701	5	3	15	1	25515
32	56.0	1020	6	3	16	0	16320
34	60.0	1728	6	3	26	0	44928
37	75.0	1496	6	3	30	0	44880
39	60.0	1904	7	4	32	0	60928
41	71.0	1768	8	4	74	0	130832
43	90.0	1736	7	3	16	1	27776
45	115.0	2186	8	4	12	0	26232
48	61.0	1400	5	3	33	0	46200
49	147.0	2165	7	3	2	1	4330
50	210.0	2353	8	4	15	1	35295
52	100.0	1972	8	3	37	1	72964
54	55.0	1664	7	3	79	0	131456
55	53.4	925	5	3	20	0	18500
57	73.0	1400	5	3	2	0	2800
58	40.0	1376	6	3	103	0	141728
60	68.0	1572	6	3	29	0	45588
61	139.0	1545	6	3	9	1	13905
63	55.0	1130	5	2	21	0	23730

My data is based on above 35 single-family residences data. The propose for this multiply regression analysis is to test whether the residence's area size, number of rooms, number of bedrooms, the owner's age, and whether the residence has a view influence the sale price of the residence. So the hypothesis test would be:
Null hypothesis: there is no correlation between sales price and age, area, number of bedroom, number of rooms, age, and area times age.
Alternative hypothesis: there is at least one independent variable has correlation with the dependent variable. We use value of $\alpha = 0.05$

II. Methodology

a) multiple linear regression with all quantitative variables

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Age, Number of Bedroom, Number of Rooms, Area[b]	.	Enter

a. Dependent Variable: Sales Price

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.809[a]	.654	.608	22.951

a. Predictors: (Constant), Age, Number of Bedroom, Number of Rooms, Area

This regression model tests the correlation between sales price and age, number of bedroom, number of room, and area. It contains four independent variables (age, number of bedrooms, number of rooms, area) and one dependent variable (sale price). Age, number of bedrooms, number of rooms, areas, and sale price are all quantitative variables. The R square of 0.654 indicate that 65.4 presents of the variation in the dependent variable can be explained by the variation in independent variables. We assume the relationship between independent variables and dependent variables are linear, all variables are normal, and there is little or no multicollinearity in the data.

b) multiple linear regression with all the quantitative variables and dummy variable

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	River View, Age, Number of Bedroom, Number of Rooms, Area[b]	.	Enter

a. Dependent Variable: Sales Price

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.876[a]	.768	.727	19.130

a. Predictors: (Constant), River View, Age, Number of Bedroom, Number of Rooms, Area

This regression model tests the correlation between sales price and river view, age, number of bedrooms, number of rooms, and area. It contains five independent variables (age, number of bedrooms, number of rooms, area, river view) and one dependent variable (sale price). Age, number of bedrooms, number of rooms, areas, and sale price are all quantitative variables, river view is a nominal variable. The R square of 0.768 indicate that 76.8 presents of the variation in

the dependent variable can be explained by the variation in independent variables. We assume the relationship between independent variables and dependent variables are linear, all variables are normal, and there is little or no multicollinearity in the data.

c) multiple linear regression with the quantitative variables and interaction effect

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Area times Age, Area, Number of Bedroom, Number of Rooms, Age[b]	.	Enter

a. Dependent Variable: Sales Price

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.811[a]	.658	.599	23.194

a. Predictors: (Constant), Area times Age, Area, Number of Bedroom, Number of Rooms, Age

This regression model tests the correlation between sales price and are times age, area, number of bedrooms, number of rooms, and age. It contains five independent variables (age, number of bedrooms, number of rooms, area, area times age) and one dependent variable (sale price). Age, number of bedrooms, number of rooms, areas, area times age, and sale price are all quantitative variables. The R square of 0.658 indicate that 65.8 presents of the variation in the dependent variable can be explained by the variation in independent variables. We assume the relationship between independent variables and dependent variables are linear, all variables are normal, and there is little or no multicollinearity in the data.

III. Descriptive Statistics and Graph

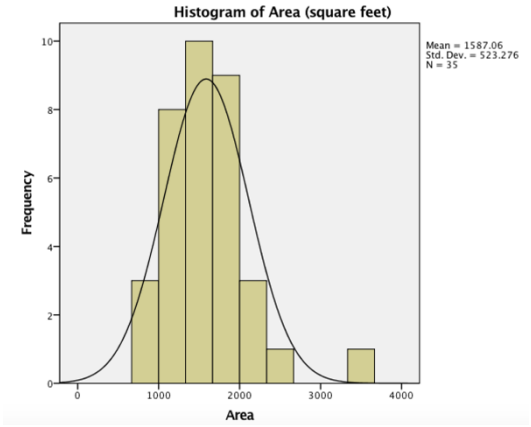
a) Area

Statistics

Area

N	Valid	35
	Missing	0
Mean		1587.06
Std. Error of Mean		88.450
Median		1545.00
Mode		1400[a]
Std. Deviation		523.276
Variance		273818.114
Skewness		1.595
Std. Error of Skewness		.398
Range		2880
Minimum		720
Maximum		3600
Sum		55547
Percentiles	25	1255.00
	50	1545.00
	75	1768.00

a. Multiple modes exist. The smallest value is shown

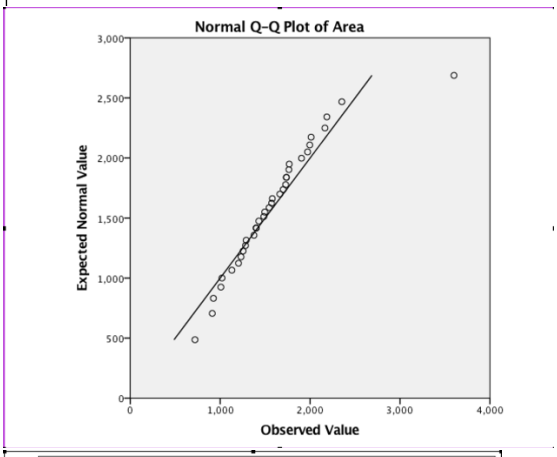


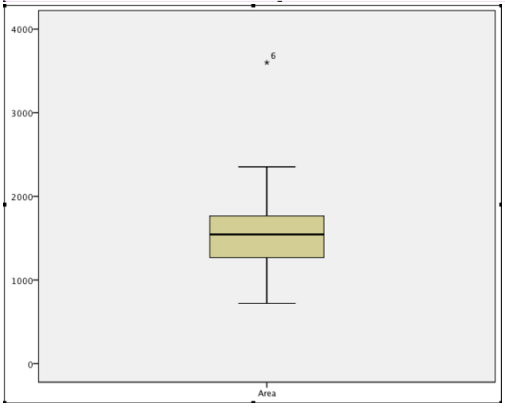

```
Area Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3.00        0 .  799
    14.00        1 .  00122222344444
    13.00        1 .  5556777777999
     4.00        2 .  0113
     1.00 Extremes    (>=3600)

 Stem width:      1000
 Each leaf:        1 case(s)
```





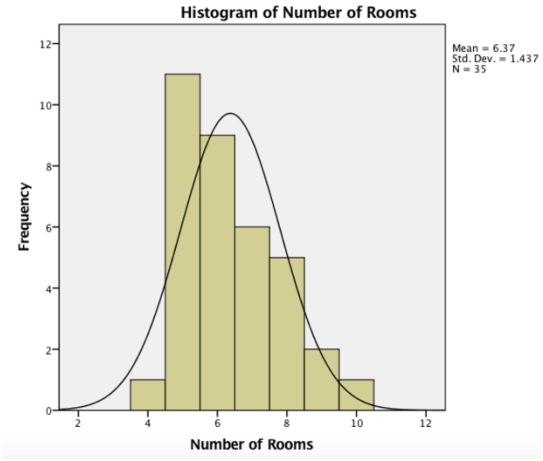
b) Number of rooms

Statistics

Number of Rooms

N	Valid	35
	Missing	0
Mean		6.37
Std. Error of Mean		.243
Median		6.00
Mode		5
Std. Deviation		1.437
Variance		2.064
Skewness		.684
Std. Error of Skewness		.398
Range		6
Minimum		4
Maximum		10
Sum		223
Percentiles	25	5.00
	50	6.00
	75	7.00

Histogram of Number of Rooms

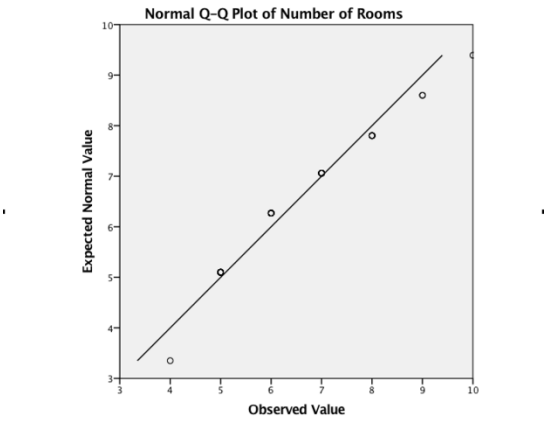


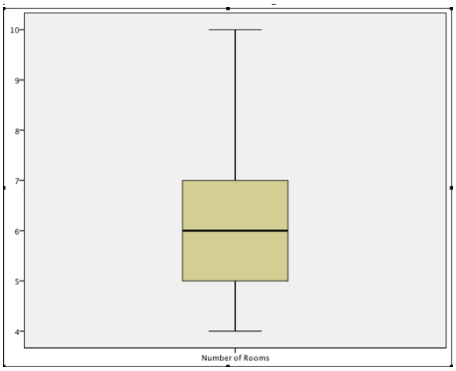
Number of Rooms

Number of Rooms Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	4 .	0
11.00	5 .	00000000000
9.00	6 .	000000000
6.00	7 .	000000
5.00	8 .	00000
2.00	9 .	00
1.00	10 .	0

Stem width: 1
Each leaf: 1 case(s)



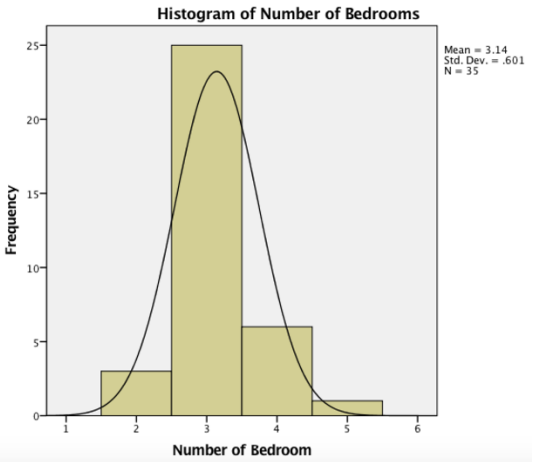


c) Number of bedrooms

Statistics

Number of Bedroom

N	Valid	35
	Missing	0
Mean		3.14
Std. Error of Mean		.102
Median		3.00
Mode		3
Std. Deviation		.601
Variance		.361
Skewness		.809
Std. Error of Skewness		.398
Range		3
Minimum		2
Maximum		5
Sum		110
Percentiles	25	3.00
	50	3.00
	75	3.00

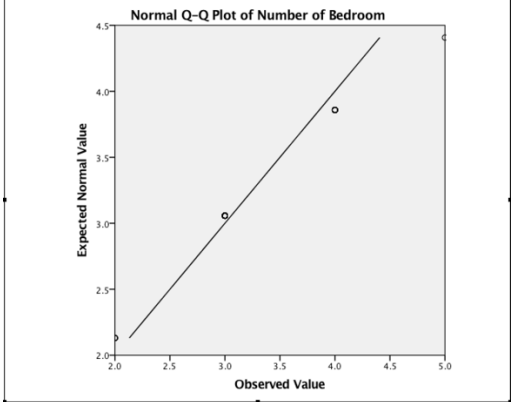



```
Number of Bedroom Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3.00 Extremes    (=<2)
    25.00        0 .  3333333333333333333333333
     7.00 Extremes    (>=4)

 Stem width:        10
 Each leaf:       1 case(s)
```

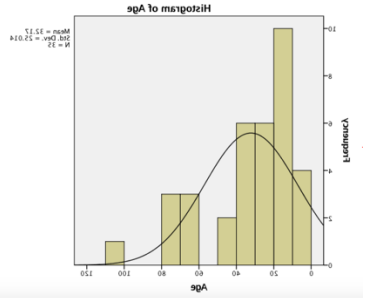


d) Age

Statistics

Age

N	Valid	35
	Missing	0
Mean		32.17
Std. Error of Mean		4.228
Median		26.00
Mode		16
Std. Deviation		25.014
Variance		625.676
Skewness		1.190
Std. Error of Skewness		.398
Range		101
Minimum		2
Maximum		103
Sum		1126
Percentiles	25	15.00
	50	26.00
	75	41.00

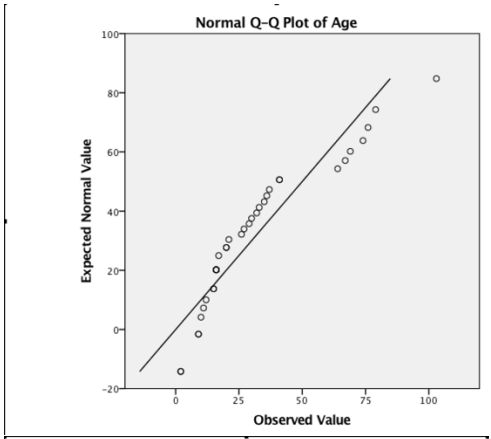


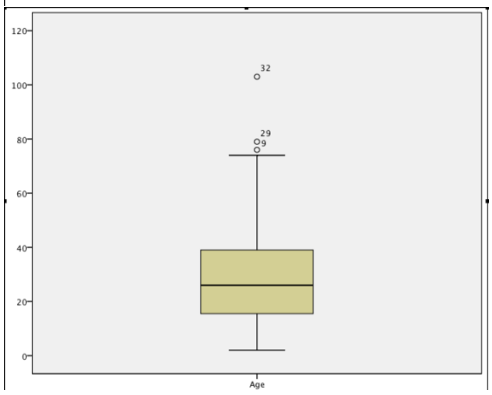
Age Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

     4.00        0 .  2299
    10.00        1 .  0125566667
     6.00        2 .  001679
     6.00        3 .  023567
     2.00        4 .  11
      .00        5 .
     3.00        6 .  479
     1.00        7 .  4
     3.00 Extremes    (>=76)

 Stem width:        10
 Each leaf:        1 case(s)
```





e) Area times age

Statistics

Area times Age

N	Valid	35
	Missing	0
Mean		49911.34
Std. Error of Mean		7074.820
Median		35295.00
Mode		2800[a]
Std. Deviation		41855.201
Variance		1.752E+9
Skewness		1.262
Std. Error of Skewness		.398
Range		150036
Minimum		2800
Maximum		152836
Sum		1746897
Percentiles	25	20080.00
	50	35295.00
	75	61200.00

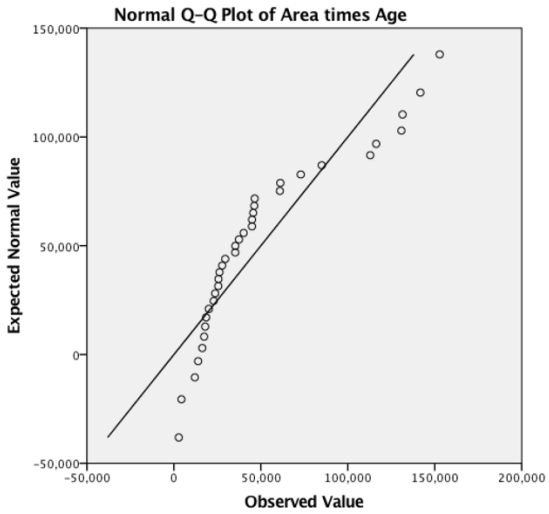
a. Multiple modes exist. The smallest value is shown

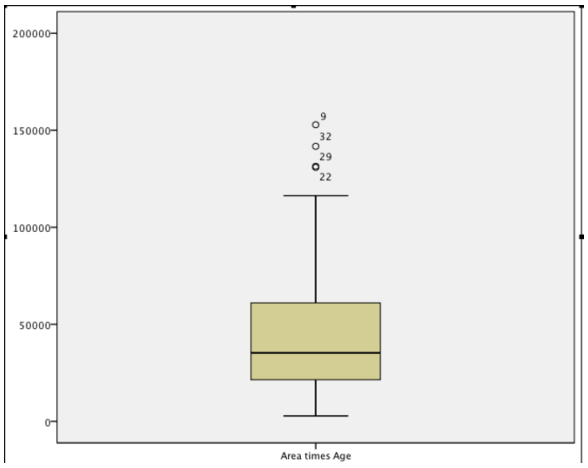
Area times Age Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

     8.00        0 .  00111111
    11.00        0 .  22222222333
     6.00        0 .  444444
     3.00        0 .  667
     1.00        0 .  8
     2.00        1 .  11
     4.00 Extremes    (>=130832)

 Stem width:    100000
 Each leaf:        1 case(s)
```



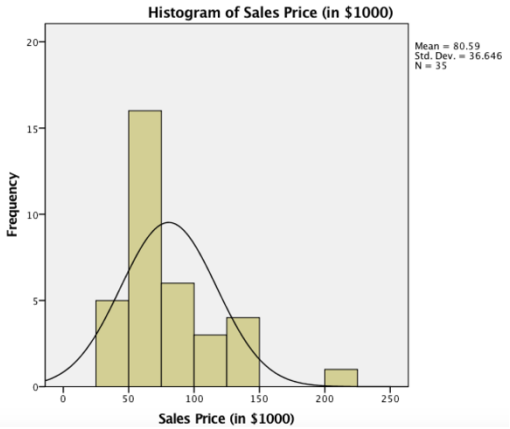


f) Sales Price

Statistics

Sales Price (in $1000)

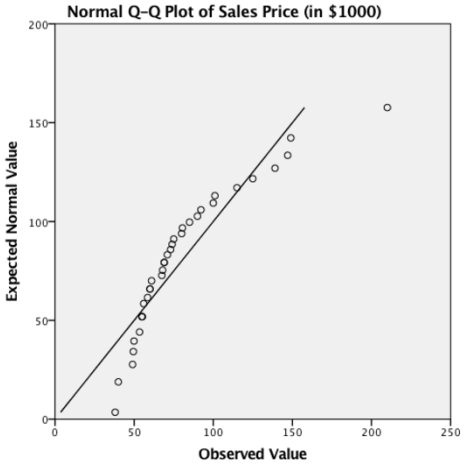
N	Valid	35
	Missing	0
Mean		80.59
Std. Error of Mean		6.194
Median		69.00
Mode		55
Std. Deviation		36.646
Variance		1342.953
Skewness		1.751
Std. Error of Skewness		.398
Range		172
Minimum		38
Maximum		210
Sum		2821
Percentiles	25	55.00
	50	69.00
	75	92.20

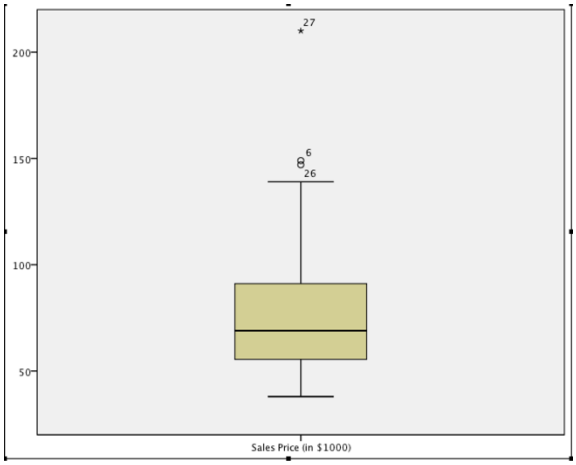


Sales Price (in $1000) Stem-and-Leaf Plot

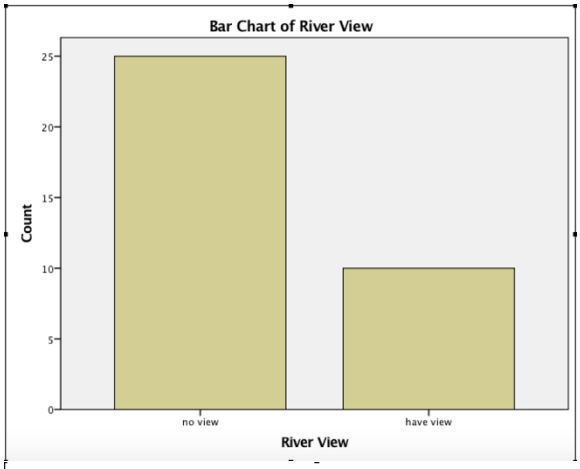
Frequency	Stem &	Leaf
1.00	0 .	3
10.00	0 .	4444555555
11.00	0 .	66666667777
5.00	0 .	88899
3.00	1 .	001
2.00	1 .	23
3.00	Extremes	(>=147)

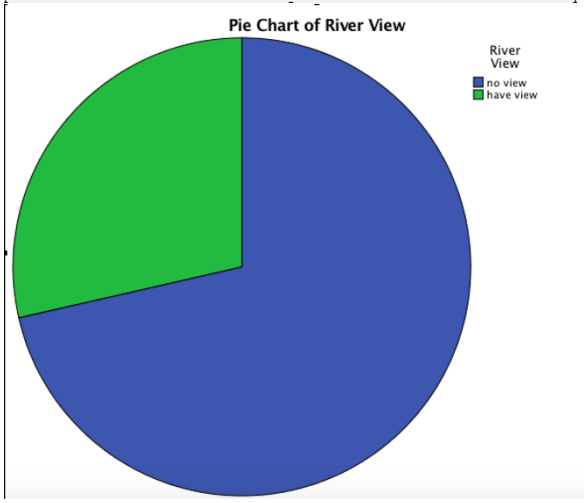
Stem width: 100
Each leaf: 1 case(s)





g) River view





Major findings:

From the descriptive analysis, for area, we find the minimum is 720 square feet, the maximum is 3600 square feet, the q25 is 1255 square feet, median is 1545 square feet, and q75 is 1768 square feet. For number of rooms, we find the minimum is 4, the maximum is 10, the q25 is 5, median is 6, and q75 is 7. For number of bedrooms, we find the minimum is 2, the

maximum is 5, the q25 is 3, median is 3, and q75 is 3. For age, we find the minimum is 2 years, the maximum is 103 years, the q25 is 15 years, median is 26 years, and q75 is 41 years. For area times age, we find the minimum is 2800, the maximum is 152836, the q25 is 20080, median is 35295, and q75 is 61200. For sale price, we find the minimum is $38000, the maximum is $210000, the q25 is $55000, median is $69000, and q75 is $92000. From histogram, we found the skewness for area, number of rooms, number of bedrooms, and sale price are very small, so we can conclude that these data are roughly normal distributed. Both histograms for age and area times age are slightly skew to the lest, but the skewness of age is due to one extreme outlier, since the data number is greater than 30, and the skewness are under reasonable range -2 to2, the descriptive statistics shows that we did not violate ANOVA's assumption and are allowed to use it.

From the steam and leaf, except area times age are concentrated more to the left, all the other variables are concentrated to the middle. From all the Q-Q plots except the one for area times age, most of the point are locating on the line, so we can tell the variance within each of the populations is roughly equal. From the boxplot, although there are outliers, it does not significantly skew the pattern, so the data is normal distributed. Thus, we can conclude that none of the ANOVA assumption is violate. From the bar chart and pie chart, we find that there are more residences with no view than residences with view. Except the data analysis for age and areas times age, we can conclude from the analysis that area, number of room, number of bedrooms, and sale price are normally distributed, and age and area times ages are slightly skewed. So we still need to take furtherer analysis to test whether the skewness is caused by interaction, and these variable's single correlation with dependent variable sale price.

IV. Correlation Analysis

a. Correlation analysis

Correlations

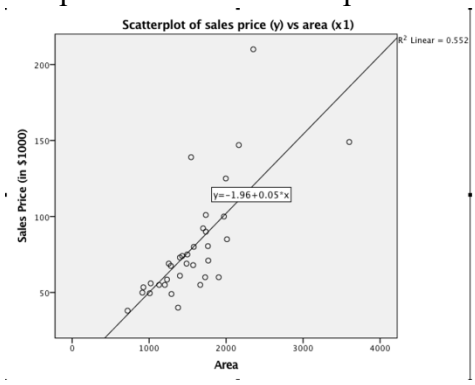
		Area	Number of Rooms	Number of Bedroom	Age	River View	Area times Age
Area	Pearson Correlation	1	.779**	.764**	-.090	.606**	.206
	Sig. (2-tailed)		.000	.000	.606	.000	.235
	N	35	35	35	35	35	35
Number of Rooms	Pearson Correlation	.779**	1	.754**	.159	.459**	.423*
	Sig. (2-tailed)	.000		.000	.360	.005	.011
	N	35	35	35	35	35	35
Number of Bedroom	Pearson Correlation	.764**	.754**	1	.030	.274	.296
	Sig. (2-tailed)	.000	.000		.866	.111	.084
	N	35	35	35	35	35	35
Age	Pearson Correlation	-.090	.159	.030	1	-.151	.937**
	Sig. (2-tailed)	.606	.360	.866		.388	.000
	N	35	35	35	35	35	35
River View	Pearson Correlation	.606**	.459**	.274	-.151	1	.044
	Sig. (2-tailed)	.000	.005	.111	.388		.800
	N	35	35	35	35	35	35
Area times Age	Pearson Correlation	.206	.423*	.296	.937**	.044	1
	Sig. (2-tailed)	.235	.011	.084	.000	.800	
	N	35	35	35	35	35	35

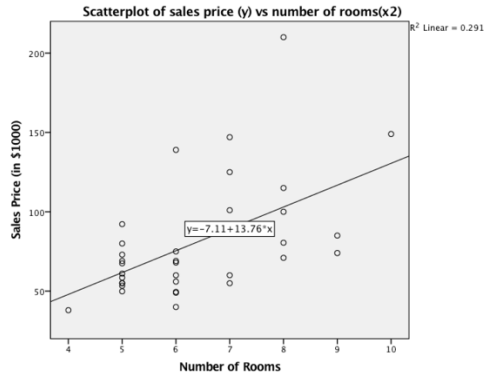
**. Correlation is significant at the 0.01 level (2-tailed).

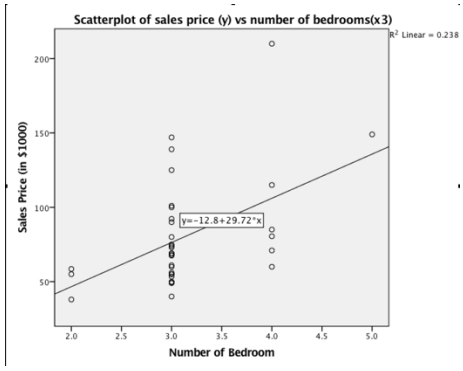
*. Correlation is significant at the 0.05 level (2-tailed).

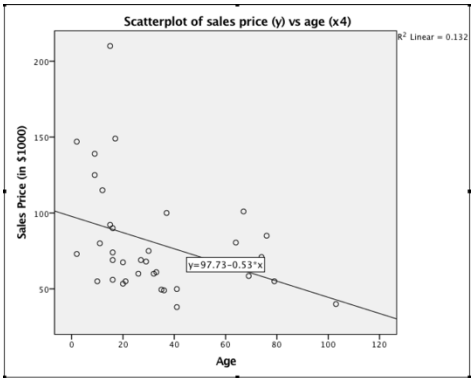
From the correlation analysis, we find that area has a strong positive correlation with number of rooms, number of bedroom, and river view; number of rooms has a strong positive correlation with area, number of bedrooms, and river view; number of bedroom has a strong positive correlation with area and number of rooms; age has a strong positive correlation with ages times area; river view has a strong positive correlation with area and number of rooms; area times age has a strong positive correlation with river view.

b. Scatter plots with the quantitative variables
   a. Independent variable to dependent variable scatter plots



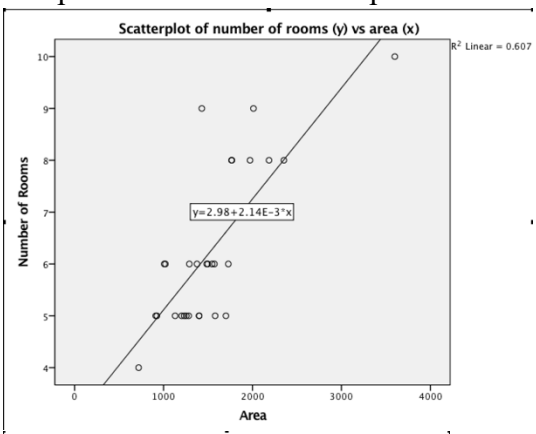


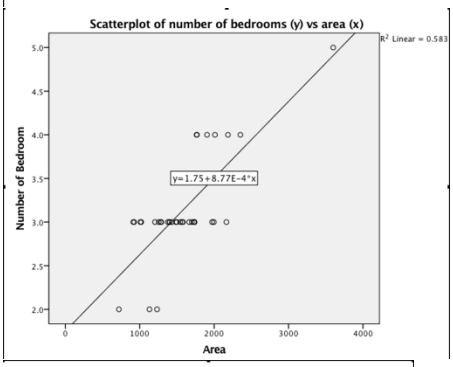


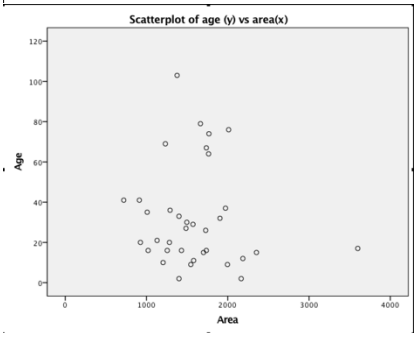


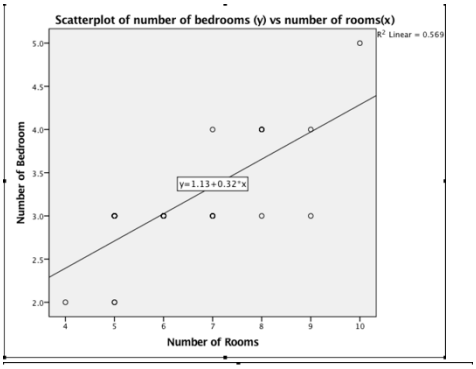
Sale price (y) appears highly correlated with area (x1)
Sale price (y) appears somewhat correlated with number of rooms (x2), number of bedrooms (x3), and age (x4)

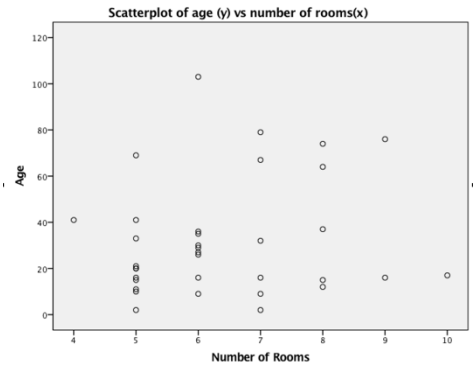
b. Independent variable vs independent variable scatterplots

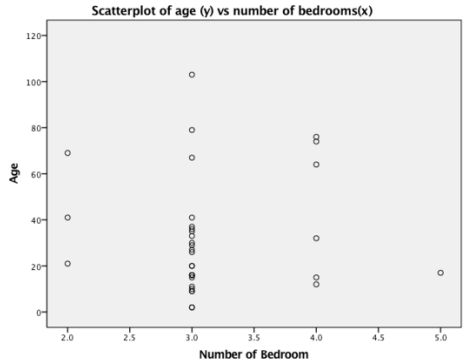












area (x1) appears highly correlated with number of rooms (x2) and number of bedrooms (x3), number of bedrooms (x3) appears highly correlated with number of rooms (x2), these might be multicollinearity. Number of bedrooms (x3) and number of rooms (x2) are both highly correlated with each other and therefore are redundant.

V. Regression Analysis

a) Linear regression analysis with the quantitative variables

i. Explain the findings of your regression analysis

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	29858.413	4	7464.603	14.172	.000[b]
	Residual	15801.990	30	526.733		
	Total	45660.403	34			

a. Dependent Variable: Sales Price (in $1000)

b. Predictors: (Constant), Age, Number of Bedroom, Number of Rooms, Area

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	23.086	22.753		1.015	.318					
	Area	.051	.014	.726	3.625	.001	.743	.552	.389	.288	3.474
	Number of Rooms	4.032	5.053	.158	.798	.431	.540	.144	.086	.294	3.401
	Number of Bedroom	-10.768	11.032	-.177	-.976	.337	.487	-.175	-.105	.352	2.839
	Age	-.466	.170	-.318	-2.744	.010	-.364	-.448	-.295	.857	1.166

a. Dependent Variable: Sales Price (in $1000)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Area	Number of Rooms	Number of Bedroom	Age
1	1	4.601	1.000	.00	.00	.00	.00	.01
	2	.334	3.714	.00	.01	.00	.00	.79
	3	.045	10.083	.36	.25	.00	.00	.08
	4	.011	20.533	.33	.67	.74	.02	.11
	5	.009	22.698	.31	.07	.25	.97	.01

a. Dependent Variable: Sales Price (in $1000)

]

Since the p value of 0 is smaller than α, we can reject the null hypothesis and conclude that there is at least one independent variable has correlation with sale price. From the coefficients chart, we find that area and age have significant p value that is smaller than α, so that they have correlation with sale price, while number of rooms and number of bedroom do not show correlation. Among these four independent variables, since all the VIF are smaller than 5, the multicollinearity is not significant to remove.

ii. Conduct one forecast for a set of independent variables and find the residual.
Independent variable: area, the regression line is calculated from 35 data, my forecast will show the first five values.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.742545
R Square	0.55137307
Adjusted R S	0.53777831
Standard Er	24.8998945
Observation	35

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	25146.0147	25146.0147	40.5577777	3.2748E-07
Residual	33	20460.1567	620.004748		
Total	34	45606.1714			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.8529381	13.6181993	-0.1360634	0.89259783	-29.559373	25.8534967	-29.559373	25.8534967
X Variable 1	0.05197135	0.00816069	6.36849886	3.2748E-07	0.0353683	0.06857441	0.0353683	0.06857441

according from the output, the regression line function is y^=-1.85+0.0519x

observation	X (area/square feet)	Predicted y(in $1000)	Residual (observed-predicted)(in $1000)
1	1290	65.19	-16.19
2	912	45.54	4.45
3	1204	60.72	-5.72
4	1764	89.82	-8.82
5	1255	63.37	5.63

iii. Conduct hypothesis testing for the slope of an independent variable
Independent variable: area
Null hypothesis: slope=0
Alternative hypothesis: slope does not equal 0
Value of $\alpha = 0.05$

P-value=3.27x10^(-7)<0.05, so we can reject the null hypothesis and conclude that the slope is not 0, so there is correlation between area and sale price.

iv. Conduct interval estimate for the slope of an independent variable
Independent variable: area, Value of $\alpha = 0.05$
Interval estimate= $slope \pm t * standard\ error$
=0.0519±2.042*0.00816069=(0.035,0.069)
So the interval of the slope with a 95% significance is between 0.035 and 0.069.

b) Linear regression analysis with all the six variables which includes the dummy variable
i. Explain the findings of your regression analysis with the dummy variable

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35047.542	5	7009.508	19.154	.000[b]
	Residual	10612.861	29	365.961		
	Total	45660.403	34			

a. Dependent Variable: Sales Price (in $1000)

b. Predictors: (Constant), River View, Age, Number of Bedroom, Number of Rooms, Area

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	23.701	18.966		1.250	.221					
	Area	.024	.014	.350	1.798	.083	.743	.317	.161	.212	4.718
	Number of Rooms	1.251	4.276	.049	.292	.772	.540	.054	.026	.285	3.506
	Number of Bedroom	3.907	9.988	.064	.391	.699	.487	.072	.035	.299	3.349
	Age	-.399	.143	-.272	-2.792	.009	-.364	-.460	-.250	.844	1.185
	River View	37.093	9.851	.464	3.766	.001	.757	.573	.337	.528	1.894

a. Dependent Variable: Sales Price (in $1000)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	Area	Number of Rooms	Number of Bedroom	Age	River View
1	1	4.981	1.000	.00	.00	.00	.00	.01	.01
	2	.698	2.671	.00	.00	.00	.00	.09	.41
	3	.269	4.303	.01	.00	.00	.00	.76	.18
	4	.034	12.113	.49	.23	.01	.00	.02	.21
	5	.011	21.462	.22	.48	.82	.00	.11	.00
	6	.007	25.874	.28	.28	.18	1.00	.01	.19

a. Dependent Variable: Sales Price (in $1000)

Since the p value of 0 is smaller than α, we can reject the null hypothesis and conclude that there is at least one independent variable has correlation with sale price. From the coefficients chart, we find that area, age and river have significant p value that is smaller than α, so that they have correlation with sale price, while number of rooms and number of bedroom do not show correlation. Among these five independent variables, since all the VIF are smaller than 5, the multicollinearity is not significant to remove. River view is a nominal variable here, and when other independent variable maintains constant, the residence with a river view significantly increase the sale price.

c) Linear regression analysis all the quantitative variables and an interaction of two independent variables.
i. Explain the findings of your regression analysis with the interaction effect.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	30058.929	5	6011.786	11.175	.000[b]
	Residual	15601.474	29	537.982		
	Total	45660.403	34			

a. Dependent Variable: Sales Price (in $1000)

b. Predictors: (Constant), Area times Age, Area, Number of Bedroom, Number of Rooms, Age

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	4.564	38.068		.120	.905					
	Area	.055	.016	.788	3.478	.002	.743	.543	.377	.229	4.358
	Number of Rooms	4.917	5.308	.193	.926	.362	.540	.170	.101	.272	3.675
	Number of Bedroom	-8.919	11.554	-.146	-.772	.446	.487	-.142	-.084	.328	3.048
	Age	.064	.886	.044	.073	.943	-.364	.013	.008	.032	31.036
	Area times Age	.000	.001	-.387	-.611	.546	-.146	-.113	-.066	.029	34.157

a. Dependent Variable: Sales Price (in $1000)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	Area	Number of Rooms	Number of Bedroom	Age	Area times Age
1	1	5.371	1.000	.00	.00	.00	.00	.00	.00
	2	.536	3.167	.00	.00	.00	.00	.01	.01
	3	.068	8.862	.06	.09	.00	.00	.02	.03
	4	.012	21.565	.05	.77	.38	.11	.06	.03
	5	.009	23.923	.01	.02	.53	.68	.02	.02
	6	.004	37.733	.88	.12	.08	.21	.89	.91

a. Dependent Variable: Sales Price (in $1000)

Since the p value of 0 is smaller than α, we can reject the null hypothesis and conclude that there is at least one independent variable has correlation with sale price. From the coefficients chart, we find that area has significant p value that is smaller than α, so that they have correlation with sale price, while number of rooms, number of bedroom, age and area times age do not show correlation. Among these five independent variables, there are two VIF (age and area times age are greater than 5), so there are multicollinearity in this test.

VI. Conclusions (including Step 6 tasks)

In testing the correlation between sale price and independent variables (area, number of room, number of bedrooms, age, river view, and area times age). With an α of 0.05. We reject the null hypothesis in all three multiple linear regression analyses and conclude that there is at least one independent variable has correlation with sale price. We find out that area has a strong correlation with sale price as it has correlation with sale price in all three analyses. We also find number of rooms and number of bedrooms are redundant, since they are highly correlated with each other. A residence with river view would significantly increase the sale price when other independent variables maintains constant. The interaction of area times age would significantly influence the output of age by making age highly correlated with the interaction. Even though area has strong correlation with sale price, it also has strong correlation with number of rooms and number of bedroom, which make sense in real life as more more rooms requires more space. Because the correlation between area and number of rooms is not as strong as the correlation between area and sale price, thus the high correlation between area and number of rooms and bedrooms does not cause multicollinearity.