Real Estate Sales

Hyunjoon Rhee (hrhee8)
Jinho Lee (jinhohl2)

Introduction

In this case study, we are given a dataset regarding the sales prices of real estates and eleven other factors of real estates that affects the sales prices. An objective of this study is to perform a multivariate statistical analysis using ANOVA method to observe correlations between sales prices and other deciding factors and see how the factors affect the sales prices.

1. Descriptive Statistics

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
price	522	277894.15	137923.40	84000.00	920000.00
sqfeet	522	2260.63	711.0659325	980.0000000	5032.00
bed	522	3.4712644	1.0143585	0	7.0000000
bath	522	2.6417625	1.0641692	0	7.0000000
aircon	522	0.8314176	0.3747418	0	1.0000000
garage	522	2.0996169	0.6539705	0	7.0000000
pool	522	0.0689655	0.2536386	0	1.0000000
year	522	1966.90	17.6379243	1885.00	1998.00
quality	522	2.1839080	0.6414128	1.0000000	3.0000000
style	522	3.3448276	2.5628121	1.0000000	11.0000000
lot	522	24369.70	11684.08	4560.00	86830.00
highway	522	0.0210728	0.1437648	0	1.0000000

Figure 1

We started off by investigating some statistical information of all twelve factors. First, we obtained mean, standard deviation, minimum and maximum values of each factor by using the mean procedure of SAS[(1)]

Statistics/Factor	Median	Mode	IQR
Sales price	229900	175000	155000
Finished square feet	2061	1592	937
Number of bedrooms	3	3	1
Number of bathrooms	3	3	1
Air conditioning	1	1	0
Garage size	2	2	0
Pool	0	0	0
Year built	1966	1956	25
Quality	2	2	1
Style	2	1	6
Lot size	22200	15001	9631
Adjacent to highway	0	0	0

Figure 2

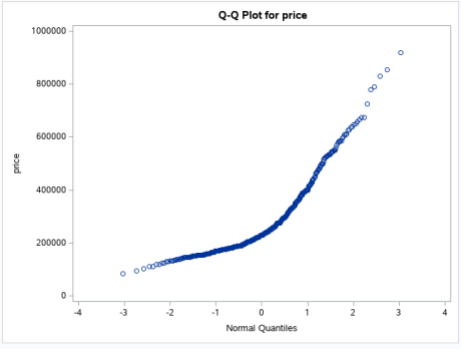


Figure 3

Afterwards, through the UNIVARIATE procedure in SAS, we also obtained the median, mode, iqr (interquartile range) of all factors[2].

Finally, as shown in figure 3, we created a histogram for the response variable (price) using UNIVARIATE procedure again[3] . By looking at the plot, we can observe that the sales price of real estates are not normally distributed, which implies that we have to transform data in order to use ANOVA model for the analysis.

2. Two way ANOVA

In this part of the study, we have to perform 2-way ANOVA analysis with two crossed factors, quality and style. There are 522 data about the style of the real estate. Note that the dominant value (mode) of this column is 1 while others vary from 1 to 11. We do not have any further detailed information about this indicator. Therefore, we re-coded the column so that the value is equal to 1 if the original value is 1, or 0 otherwise[4].

We used a significance level α=0.1 for this analysis. Figure 9 shows that ANOVA table shows high mean squared error(MSE) with a value approximately 6*10^9. On the other hand, we could find an interaction between two variables[15]. Figure 10 shows the interaction plot, which indicates that there is a point where two of the lines cross each other, implying the interaction. Also interaction could be measured by looking at p value in Figure 9. The p value for the interaction between quality and style was below 0.1 which indicated that there is an interaction between those two.

After the analysis with ANOVA table, analysis of the assumption of performing ANOVA was made. Three different assumptions - constant variance, normality, and independence - were analyzed individually through fitted versus residual plot, qqplot and sequence plot[13]. Figure 11 shows the fitted vs residual plot which could be interpreted as non constant variance. The trend of the dots are not showing the constant variance, meaning that the dots are scattered with a specific pattern. Figure 12 shows the histogram and qqplot. Looking at qqplot, it shows a trend of non linear graph, meaning that the normality is violated. Sequence plot in Figure 13 indicates a non independent trend of the data points.

Considering these factors mentioned above, transformation of the data was needed, and Box Cox transformation was applied on these data[14]. Figure 14 shows that the optimal λ value is 0.5. Therefore, considering this λ value, the transformation of the data was made. After that, Figure 4 and 5 shows the ANOVA table, and another assumption check was made in Figures 17, 18, and 19. The normality and constant variance was much improved than the one before transformation.

The GLM Procedure

Dependent Variable: newprice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.00005655	0.00001131	177.24	<.0001
Error	516	0.00003293	0.00000006		
Corrected Total	521	0.00008948			

R-Square	Coeff Var	Root MSE	newprice Mean
0.632007	12.38661	0.000253	0.002039

Source	DF	Type I SS	Mean Square	F Value	Pr > F
quality	2	0.00005539	0.00002770	434.02	<.0001
newstyle	1	0.00000070	0.00000070	11.03	0.0010
quality*newstyle	2	0.00000045	0.00000023	3.56	0.0290

Source	DF	Type III SS	Mean Square	F Value	Pr > F
quality	2	0.00004750	0.00002375	372.20	<.0001
newstyle	1	0.00000021	0.00000021	3.22	0.0734
quality*newstyle	2	0.00000045	0.00000023	3.56	0.0290

Level of quality	Level of newstyle	N	newprice	
			Mean	Std Dev
1	0	43	0.00140851	0.00016045
1	1	25	0.00133672	0.00014215
2	0	213	0.00194877	0.00027219
2	1	77	0.00204303	0.00028872
3	0	52	0.00234115	0.00020452
3	1	112	0.00246836	0.00025445

Figure 4 & 5

Mean Analysis

When we look at the ANOVA table(Figure 4 & 5), the P-value for the interaction effect is 0.0290. This value is smaller than our significance level, which indicates that the interaction effect of two different factors exists, and is statistically significant. Therefore, we have to perform a treatment mean analysis.

First, we used TUKEY analysis with factor level means to have a brief idea of which treatment means we have to observe[5]. By looking at the confidence intervals from TUKEY analysis, we can observe a clear main factor effect of the quality of the real estates on sales prices.

Comparisons significant at the 0.1 level are indicated by ***.

quality Comparison	Difference Between Means	Simultaneous 90% Confidence Limits		
3 - 2	0.00045423	0.00040346	0.00050499	***
3 - 1	0.00104590	0.00097096	0.00112085	***
2 - 3	-.00045423	-.00050499	-.00040346	***
2 - 1	0.00059168	0.00052167	0.00066169	***
1 - 3	-.00104590	-.00112085	-.00097096	***
1 - 2	-.00059168	-.00066169	-.00052167	***

Figure 6

Therefore, we have conducted a set of pairwise analysis of treatment means so we can observe the interacting effect of quality and style for each quality level of real estates.
Our MSE value is 6*10^(-8). We used n = (522/6) = 89 since there are 6 different treatments and 522 data. Also, we used the Bonferroni confidence interval here because we are making only a few pairwise comparisons.

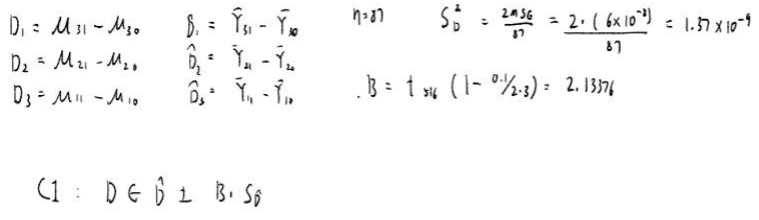


Figure 7

Y31	0.002468	Y30	0.00234115	
Y21	0.002043	Y20	0.00194877	
Y11	0.001337	Y10	0.00140851	
Bon	2.13376		Std.Dev	3.70135E-05
	Estimator	Lower	Uper	
CI1	0.000127	4.8232E-05	0.00020619	
CI2	9.43E-05	1.5282E-05	0.00017324	
CI3	-7.2E-05	-0.0001508	7.188E-06	

Figure 8

By looking at the confidence interval, we can see that when the quality level is 2 or 3, the confidence interval only includes positive range. Therefore, we can conclude that the real estates with style 1 have higher sales price then when they have quality level 2 or 3. However, when the quality level is 1, the confidence interval includes zero. Therefore, we can conclude that the style is statistically not significant when the quality level is zero.

The GLM Procedure

Dependent Variable: price

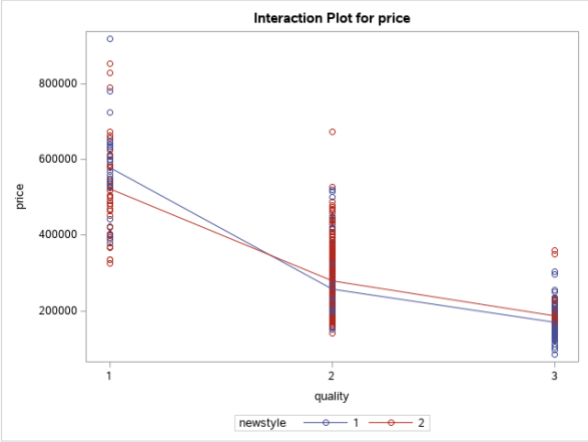
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6.6313075E12	1.3262615E12	208.67	<.0001
Error	516	3.2796044E12	6355822530.8		
Corrected Total	521	9.9109119E12			

R-Square	Coeff Var	Root MSE	price Mean
0.669092	28.68841	79723.41	277894.1

Source	DF	Type I SS	Mean Square	F Value	Pr > F
quality	2	6.5417834E12	3.2708917E12	514.63	<.0001
newstyle	1	11695175157	11695175157	1.84	0.1755
quality*newstyle	2	77828934050	38914467025	6.12	0.0024

Source	DF	Type III SS	Mean Square	F Value	Pr > F
quality	2	6.1355251E12	3.0677626E12	482.67	<.0001
newstyle	1	1404344929.7	1404344929.7	0.22	0.6385
quality*newstyle	2	77828934050	38914467025	6.12	0.0024

Figure 9: Two Way ANOVA table (Before Box Cox Transformation)



Level of quality	Level of newstyle	N	price	
			Mean	Std Dev
1	1	25	577850.000	122372.425
1	2	43	523704.047	121985.740
2	1	77	256456.429	85835.971
2	2	213	280023.831	84470.887
3	1	112	169247.321	34884.338
3	2	52	187448.077	41420.968

Figure 10: Interaction plot

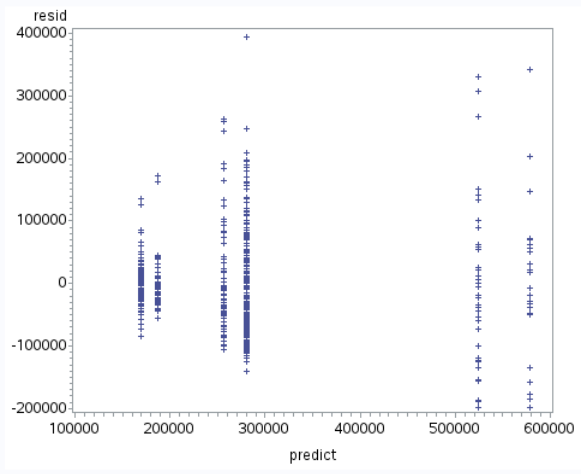
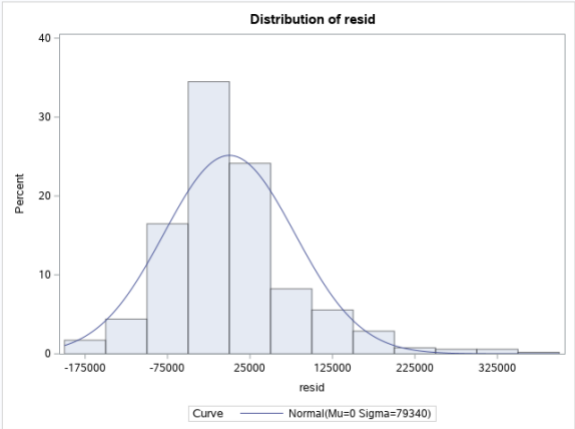


Figure 11: Fitted vs Residual (Before Box Cox Transformation)



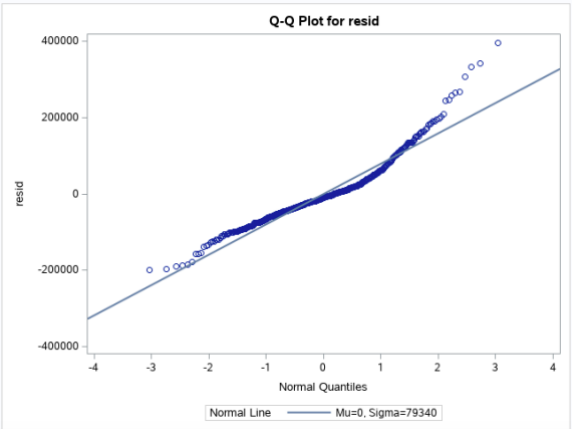


Figure 12: Histogram and QQplot before Box Cox transformation

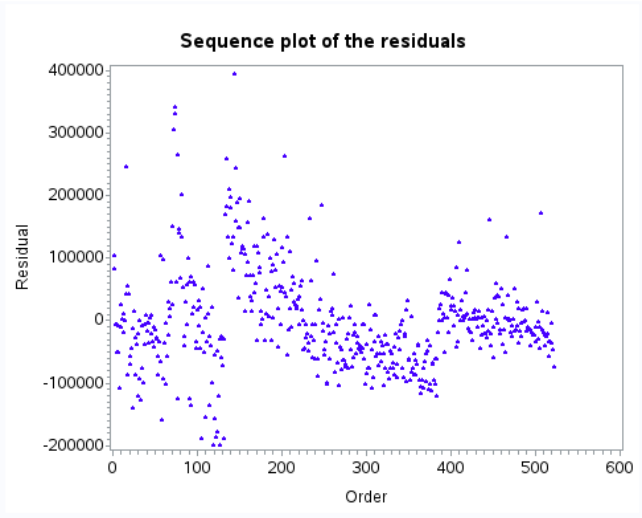


Figure 13: Sequence plot before transformation

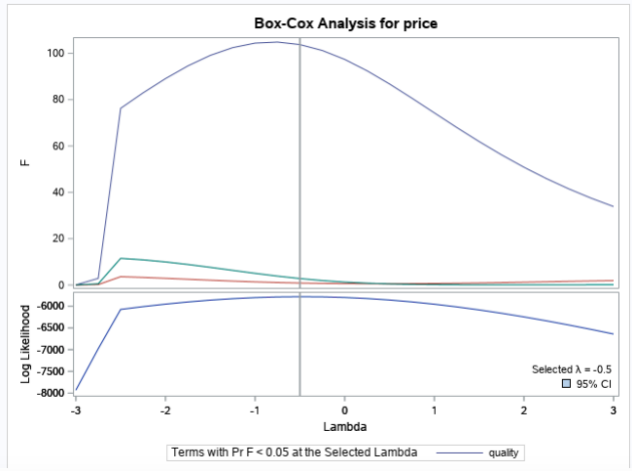


Figure 14 Box-Cox Transformation graph and optimal λ value

The GLM Procedure

Dependent Variable: newprice

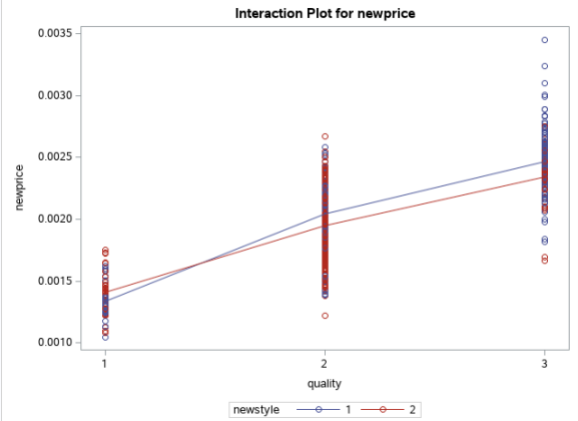
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.00005655	0.00001131	177.24	<.0001
Error	516	0.00003293	0.00000006		
Corrected Total	521	0.00008948			

R-Square	Coeff Var	Root MSE	newprice Mean
0.632007	12.38661	0.000253	0.002039

Source	DF	Type I SS	Mean Square	F Value	Pr > F
quality	2	0.00005539	0.00002770	434.02	<.0001
newstyle	1	0.00000070	0.00000070	11.03	0.0010
quality*newstyle	2	0.00000045	0.00000023	3.56	0.0290

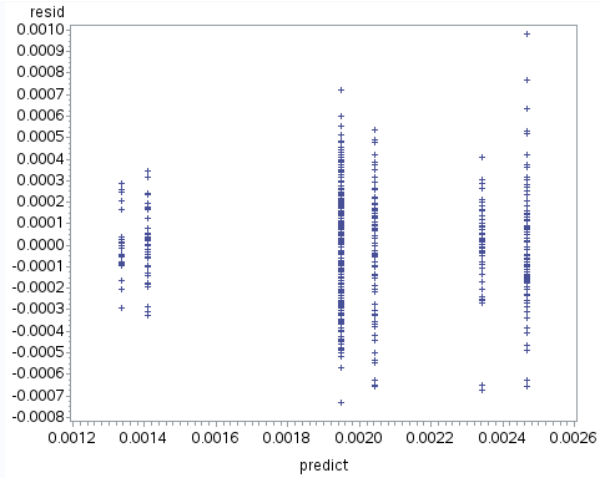
Source	DF	Type III SS	Mean Square	F Value	Pr > F
quality	2	0.00004750	0.00002375	372.20	<.0001
newstyle	1	0.00000021	0.00000021	3.22	0.0734
quality*newstyle	2	0.00000045	0.00000023	3.56	0.0290

Figure 15: ANOVA table after transformation



Level of quality	Level of newstyle	N	newprice	
			Mean	Std Dev
1	1	25	0.00133672	0.00014215
1	2	43	0.00140851	0.00016045
2	1	77	0.00204303	0.00028872
2	2	213	0.00194877	0.00027219
3	1	112	0.00246836	0.00025445
3	2	52	0.00234115	0.00020452

Figure 16: Interaction plot after transformation



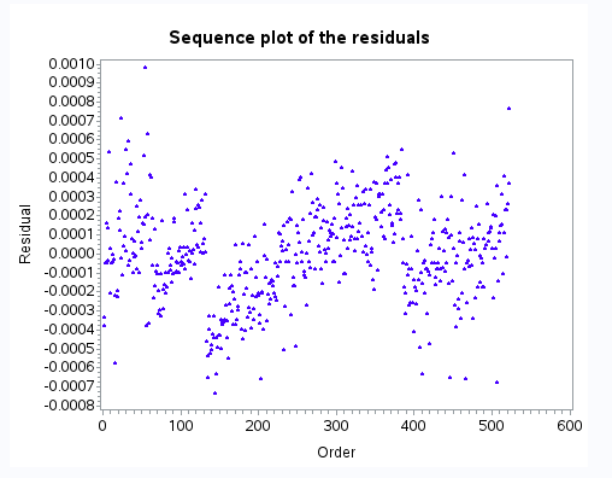
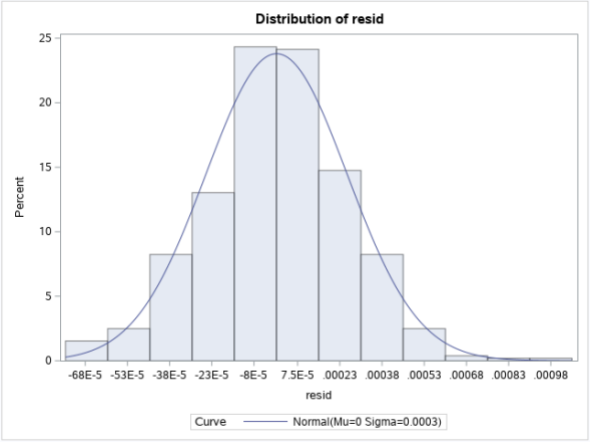


Figure 17 & 18: Fitted vs Residual and Sequence Plot after transformation



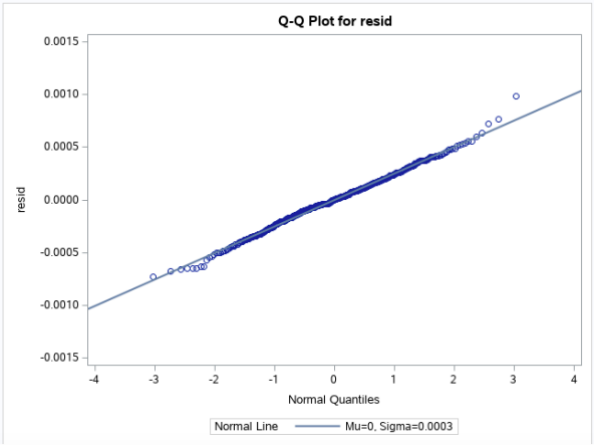


Figure 19 Histogram and QQplot after transformation

Analysis:

3. Four way ANOVA

Last approach of analysis was done with 4-way ANOVA analysis with square feet, number of bedrooms, number of bathrooms, and year built. There are 522 data about the style of the real estate. Before beginning the ANOVA analysis, because the values of number of bedrooms, bathrooms, square feet, and year all have different ranges, we had to categorize those into values from 1 to 4[(6)].

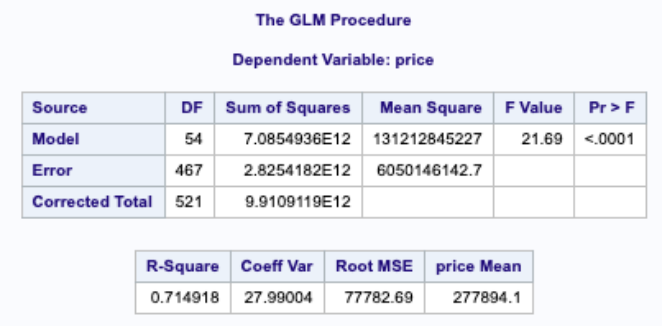
We used a significance level $\alpha=0.1$ for this analysis. Before making an analysis with the results of the model, first we went through the process of figuring out the apt model to analyze. The method was to first sort out the highest order of interaction and watch the p value of the interactions. If the p values were below α, the null hypothesis of interaction equal 0 could be rejected meaning that there is an interaction

for the specific term.
Four Way ANOVA (Before Box Cox Transformation)

The model we have started off was interaction of all four[7]. Looking at the Type III table of Figure 20, we can derive that the degree of freedom for 4 way interaction does not exist and the SS value is 0, which means that there cannot be any interaction among 4 variables. So the model eliminated the 4 way interaction. Then, we have moved on to the model with all three interactions in Figure 21[8]. Based on the SS value, the highest order of interaction is 3 way interaction. Looking at the p values, when $\alpha = 0.1$, three way interaction of square feet, bathroom, and year shows an interaction, but any other three way interaction is above 0.1, meaning that there is no interaction among them. Therefore, we eliminated the three way interaction except one. We moved on to Figure 22[9]. Table III of Figure 22 shows that based on the SS value, the highest order of interaction is 3 way interaction. Looking at the p values, when $\alpha = 0.1$, three way interaction of squarefeet, bathroom, and year shows an interaction, but any other three way interaction is above 0.1, which means that null hypothesis cannot be rejected, meaning that there is no interaction among them[10]. By this process until Figure 22, it could be concluded that there are no three way interactions for the model. Next with Figure 23, the model includes two-way interactions. Type III table shows that the highest order of interaction is two way interaction, and interactions of squarefeet and bathroom, bedroom and year, and bathroom and year have p value of greater than 0.1. This means that these three of two-way interactions fail to reject the null hypothesis, meaning that there are no interactions among them[11]. Lastly, Figure 24 shows that the model is set as final for the analysis. Table III of above shows that the highest order interaction is two way interaction, which none of them have p value greater than 0.1. This means that this model could be the finalized model[11].

Figure 24 shows that ANOVA table shows high mean squared error(MSE) with a value approximately 6*10^9. Two way interactions of square feet and number of bedrooms, square feet and year, and number of bedrooms and year shows interaction in the model. Figure 29 to 31 shows the analysis of assumption of the model. Normality is hard to say that it is violated but there are some deviance of the qqplot, constant variance is violated and independence is not shown, due to the patterned dots of the sequence plot.

Four way ANOVA (Before Box Cox Transformation)



The GLM Procedure

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	54	7.0854936E12	131212845227	21.69	<.0001
Error	467	2.8254182E12	6050146142.7		
Corrected Total	521	9.9109119E12			

R-Square	Coeff Var	Root MSE	price Mean
0.714918	27.99004	77782.69	277894.1

Source	DF	Type III SS	Mean Square	F Value	Pr > F
newsqfeet	3	128585288941	42861762980	7.08	0.0001
newbed	2	14830399687	7415199843.7	1.23	0.2945
newsqfeet*newbed	4	76314431373	19078607843	3.15	0.0142
newbath	2	19088293617	9544146808.3	1.58	0.2076
newsqfeet*newbath	3	26978364568	8992788189.3	1.49	0.2175
newbed*newbath	4	73317848541	18329462135	3.03	0.0174
newsqf*newbed*newbat	2	15645259196	7822629598.1	1.29	0.2754
newyear	3	375255067536	125085022512	20.67	<.0001
newsqfeet*newyear	6	149797047863	24966174644	4.13	0.0005
newbed*newyear	6	35787476085	5964579347.4	0.99	0.4341
newsqf*newbed*newyea	5	11773819413	2354763882.6	0.39	0.8563
newbath*newyear	4	32389262396	8097315599	1.34	0.2547
newsqf*newbat*newyea	2	38640253480	19320126740	3.19	0.0419
newbed*newbat*newyea	2	16131408000	8065703999.8	1.33	0.2646
news*newb*newb*newye	0	0	.	.	.

Figure 20[7]

The GLM Procedure

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	54	7.0854936E12	131212845227	21.69	<.0001
Error	467	2.8254182E12	6050146142.7		
Corrected Total	521	9.9109119E12			

R-Square	Coeff Var	Root MSE	price Mean
0.714918	27.99004	77782.69	277894.1

Source	DF	Type III SS	Mean Square	F Value	Pr > F
newsqfeet	3	59028494009	19676164670	3.25	0.0216
newbed	2	12648433538	6324216768.9	1.05	0.3524
newbath	2	16828127921	8414063960.7	1.39	0.2499
newyear	3	359445785699	119815261900	19.80	<.0001
newsqfeet*newbed	4	68015838146	17003959536	2.81	0.0251
newsqfeet*newbath	3	33112490584	11037496861	1.82	0.1418
newsqfeet*newyear	6	149826315242	24971052540	4.13	0.0005
newbed*newbath	4	69554972492	17388743123	2.87	0.0226
newbed*newyear	6	34443216005	5740536000.9	0.95	0.4597
newbath*newyear	4	39333205579	9833301394.8	1.63	0.1667
newsqf*newbed*newbat	2	15645259196	7822629598.1	1.29	0.2754
newsqf*newbed*newyea	5	11773819413	2354763882.6	0.39	0.8563
newsqf*newbat*newyea	2	38640253480	19320126740	3.19	0.0419
newbed*newbat*newyea	2	16131408000	8065703999.8	1.33	0.2646

Figure 21[8]

The GLM Procedure

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	43	7.0386497E12	163689526850	27.24	<.0001
Error	478	2.8722622E12	6008916812.6		
Corrected Total	521	9.9109119E12			

R-Square	Coeff Var	Root MSE	price Mean
0.710192	27.89451	77517.20	277894.1

Source	DF	Type III SS	Mean Square	F Value	Pr > F
newsqfeet	3	58433238917	19477746306	3.24	0.0219
newbed	2	11311549042	5655774520.9	0.94	0.3909
newbath	2	19824101597	9912050798.6	1.65	0.1932
newyear	3	209484045542	69828015181	11.62	<.0001
newsqfeet*newbed	4	95475576462	23868894116	3.97	0.0035
newsqfeet*newbath	3	16575874247	5525291415.8	0.92	0.4312
newsqfeet*newyear	6	108743875634	18123979272	3.02	0.0067
newbed*newbath	4	89935945742	22483986436	3.74	0.0052
newbed*newyear	6	38310098399	6385016399.9	1.06	0.3841
newbath*newyear	6	60823206838	10137201140	1.69	0.1222
newsqf*newbat*newyea	4	36330709473	9082677368.2	1.51	0.1976

Figure 22[9]

The GLM Procedure

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	39	7.0023189E12	179546639617	29.75	<.0001
Error	482	2.9085929E12	6034425198.9		
Corrected Total	521	9.9109119E12			

R-Square	Coeff Var	Root MSE	price Mean
0.706526	27.95365	77681.56	277894.1

.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
newsqfeet	3	45480220517	15160073506	2.51	0.0579
newbed	2	11195566115	5597783057.3	0.93	0.3962
newbath	2	17990824857	8995412428.6	1.49	0.2263
newyear	3	195389166097	65129722032	10.79	<.0001
newsqfeet*newbed	4	91440613307	22860153327	3.79	0.0048
newsqfeet*newbath	3	8361467115.3	2787155705.1	0.46	0.7090
newsqfeet*newyear	6	138654610194	23109101699	3.83	0.0010
newbed*newbath	4	83812755807	20953188952	3.47	0.0082
newbed*newyear	6	36527526944	6087921157.3	1.01	0.4187
newbath*newyear	6	49679654090	8279942348.3	1.37	0.2240

Figure 23[10]

The GLM Procedure

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	24	6.8924835E12	287186811776	47.29	<.0001
Error	497	3.0184284E12	6073296596.2		
Corrected Total	521	9.9109119E12			

R-Square	Coeff Var	Root MSE	price Mean
0.695444	28.04354	77931.36	277894.1

Source	DF	Type III SS	Mean Square	F Value	Pr > F
newsqfeet	3	308713730945	102904576982	16.94	<.0001
newbed	2	860789515.01	430394757.51	0.07	0.9316
newbath	2	89093911482	44546955741	7.33	0.0007
newyear	3	473189152861	157729717620	25.97	<.0001
newsqfeet*newbed	4	79344054996	19836013749	3.27	0.0117
newsqfeet*newyear	6	221347974789	36891329132	6.07	<.0001
newbed*newbath	4	86966730160	21741682540	3.58	0.0068

Figure 24(11)

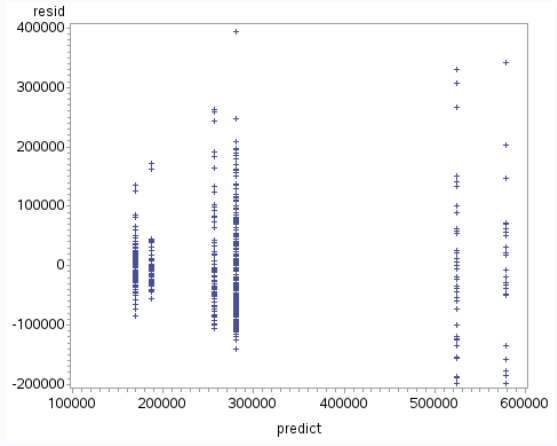


Figure 29: Fitted versus Residual before Box Cox transformation

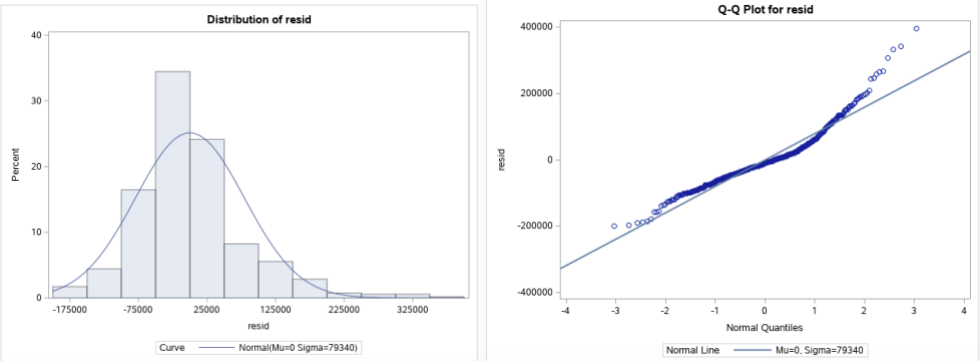


Figure 30: Histogram and QQplot before Box Cox transformation

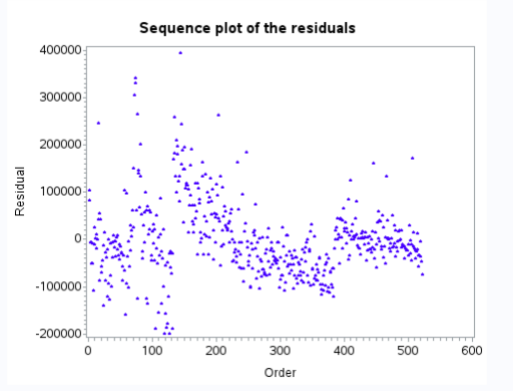


Figure 31: Sequence plot before Box Cox transformation

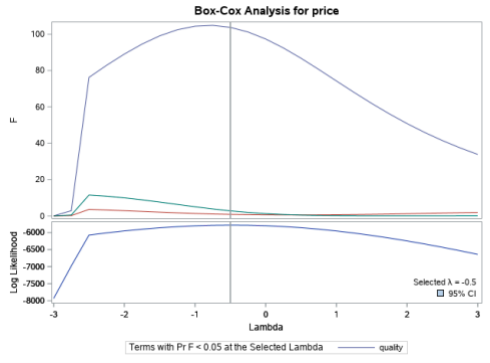


Figure 32: Box Cox transformation value

The Box Cox transformation value was calculated as -0.5 in Figure 32. Transformation was made on the values of the data. After the transformation, the ANOVA table and the analysis of assumption was made in Figure 25 to 28. Mean Square error was reduced from 6*10^9 to near zero. All the p values of the interaction remains the same. The assumption check with fitted versus residual plot was much improved with better deviance of the dots. The normality was much improved with better histogram and qqplot almost fitting the line. Also, the independence was not violated. The dots of the sequence plot was randomly scattered, which shows that considering all the assumptions analysis and the ANOVA table, the transformation was useful in terms of performing the ANOVA analysis[16].

The GLM Procedure

Dependent Variable: newprice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	24	0.00006665	0.00000278	60.46	<.0001
Error	497	0.00002283	0.00000005		
Corrected Total	521	0.00008948			

R-Square	Coeff Var	Root MSE	newprice Mean
0.744859	10.50919	0.000214	0.002039

Source	DF	Type III SS	Mean Square	F Value	Pr > F
newsqfeet	3	4.6785836E-6	1.5595279E-6	33.95	<.0001
newbed	2	1.8526538E-8	9.2632692E-9	0.20	0.8174
newbath	2	1.1687307E-6	5.8436534E-7	12.72	<.0001
newyear	3	2.4922404E-6	8.307468E-7	18.08	<.0001
newsqfeet*newbed	4	4.8369261E-7	1.2092315E-7	2.63	0.0336
newbed*newbath	4	4.1372077E-7	1.0343019E-7	2.25	0.0625
newsqfeet*newyear	6	6.1330135E-7	1.0221689E-7	2.23	0.0396

Figure 25: ANOVA table after transformation

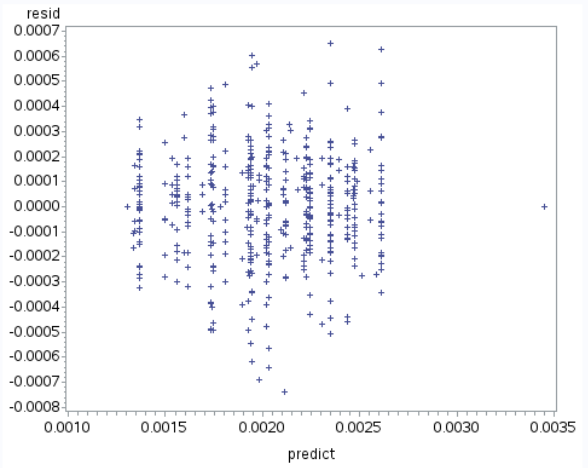
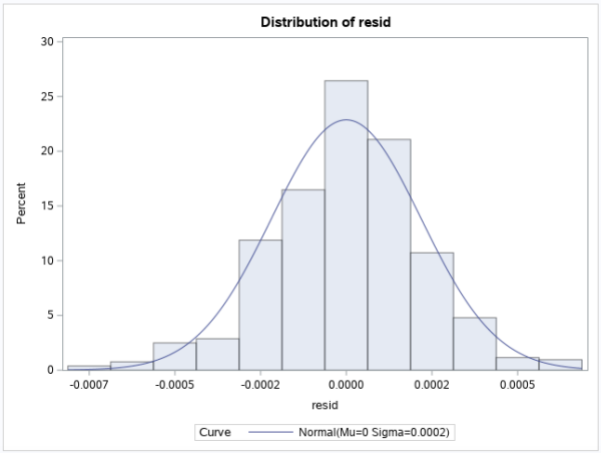


Figure 26: Fitted versus Residual after Box Cox transformation



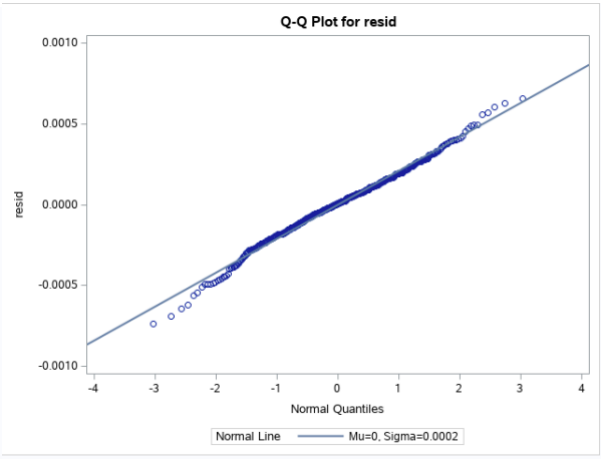


Figure 27 Histogram and QQplot after transformation

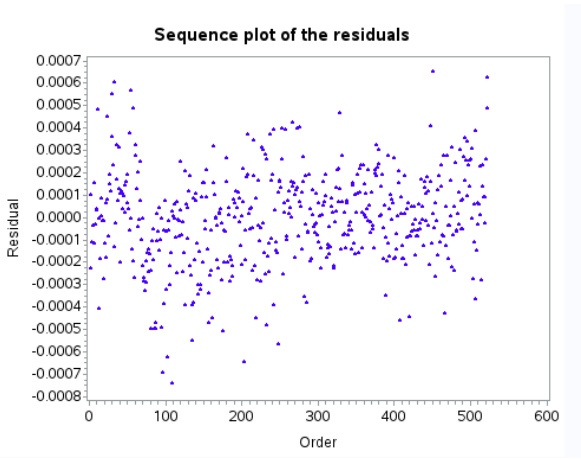
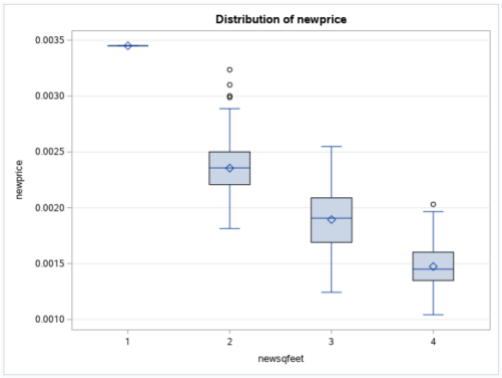
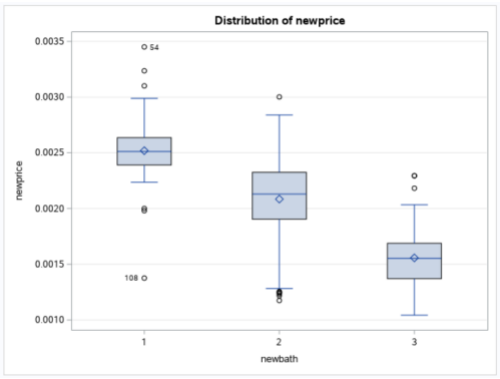
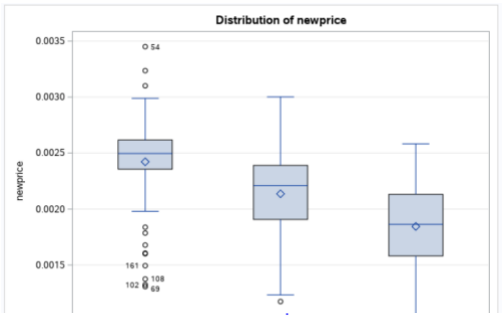


Figure 28 Sequence plot after transformation


Distribution of newprice
newprice
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
1
2
3
4
newsqfeet


Distribution of newprice
newprice
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
54
108
1
2
3
newbath


Distribution of newprice
newprice
0.0035
0.0030
0.0025
0.0020
0.0015
54
161
102
108
89
1
2
3
4

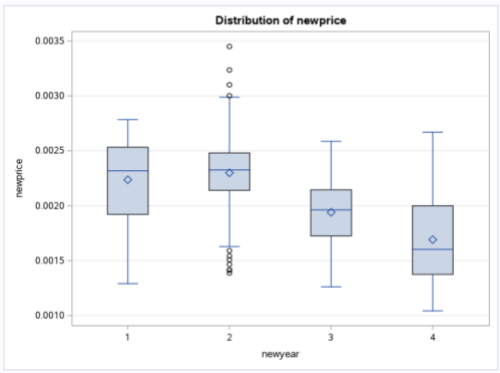
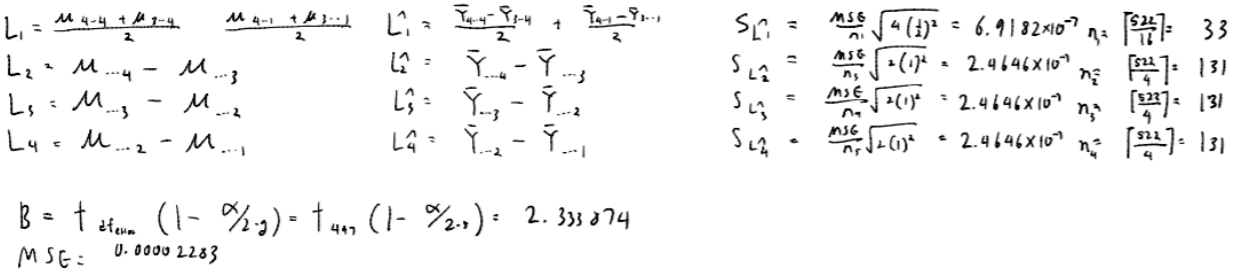

Distribution of newprice
newprice
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
1
2
3
4
newyear

Figure 33-36

Mean Analysis

Our model concludes that at most 2 factors have interacting effects. From boxplots, we have a brief idea of the factor effect of finished square feet, number of bedrooms and bathrooms on the sales price. However, the factor effect of the year looks vague. We have conducted treatment mean analysis in order to compare the treatment means and factor effect of the year. We obtained n by dividing the total sample number by the number of treatments on a level that we are interested in. Also, we have used significance level =0.1 and Bonferonni analysis since we are comparing only a few comparisons.



Y4..4	0.0013879	Y3..4	0.001741	Y4..1	0.001936	Y3..1	0.002
Y...4	0.00169139	Y...3	0.00194	Y...2	0.002299	Y...1	0.002236
Std.Dev1	6.92E-07		Bon	2.333874			
Std.Dev2	2.46E-07						
Std.Dev3	2.46E-07						
Std.Dev4	2.46E-07						
	Estimator	Lower	Uper				
CI1	-0.0004034	-4.05E-04	-4.02E-04				
CI2	-0.0002488	-2.49E-04	-2.48E-04				
CI3	-0.0003591	-3.60E-04	-3.59E-04				
CI4	0.00006294	6.24E-05	6.35E-05				

Figure 37-38

From the first confidence intervals we have, we can observe that when the finished square feet factors are equal, years built have a negative interaction effect on price. i.e) if the real estate has the same level of finished square feet, real estates built earlier have higher prices. Also, from confidence interval 2,3, since they are negative, we can conclude that the factor mean of year 3 has higher price than 4, and 2 has higher price than 3. From confidence interval 3, we can conclude that real estates built in year2 is more expensive than the real estates built in year 1.

Conclusion

We have conducted multiple ANOVA analysis in order to study the effects of quality, style, finished square feet, number of bedrooms, number of bathrooms, and years built. Through ANOVA analysis, we could figure out which interactions are influential by using the ANOVA table. We checked normality, constant variance, and independence of the dataset before analyzing data. Afterwards, we used Box-Cox transformation to make a dataset suitable for ANOVA models. In this process, we applied negative exponents to the sales price data, which invert the correlation between response variable and other factors. Therefore, if we were to use this analysis in the real world, we would have to invert such relationships. From ANOVA analysis, we have concluded that quality and style has interaction. Also, finished square

feet and number of bedrooms, number of bedrooms and bathrooms, finished square feet and years built have interaction effects.

CODE APPENDIX

```
/* Bring the data and name the variables of the data */
data sales;
   infile '/home/u49619011/my_shared_file_links/achronop/c_8944/IE 400/sales.txt';
   input ID price sqfeet bed bath aircon garage pool year quality style lot highway;

/* Calculate the statistics of the data using proc means */   -(1)
proc means data=sales;
   var price;
   var sqfeet;
   var bed;
   var bath;
   var aircon;
   var garage;
   var pool;
   var year;
   var quality;
   var style;
   var lot;
   var highway;
run;

/* Calculate the statistics of the data using proc univariate */    -(2)
proc univariate data=sales mode;
   var price;
   var sqfeet;
   var bed;
   var bath;
   var aircon;
   var garage;
   var pool;
   var year;
   var quality;
   var style;
   var lot;
   var highway;
run;

/* Plot histogram to check the distribution of response variable */  -(3)

proc univariate data=sales;
   histogram price;
run;


/* TWO Way ANOVA */
```

```
/* Modify the data so that newstyle variable could be categorized in binary options */  -(4)
data sales;
       set sales;
       newstyle = .;
run;

data sales;
       set sales;
       if style ^= 1 then newstyle = 2;
       else if style = 1 then newstyle = 1;
run;


/* Plot two way ANOVA */
proc glm data=sales;                                                    -(15)
    class quality newstyle;
    model price = quality|newstyle;
    means quality*newstyle / hovtest=bartlett hovtest=levene;
    output out=temp r=resid p=predict;
run;

/* Dataset Containing Residuals */
data residsF;
    set temp;
run;

/* Fitted vs Residual to check constant variance */                        -(13)
goptions hsize=5;
goptions vsize=4;
proc gplot data=residsF;
plot resid*predict;
run;

/* QQ plot to check normality */                                        -(13)
goptions hsize=5;
goptions vsize=4;
proc univariate data=residsF noprint;
       var resid;
       histogram resid/ normal;
       qqplot resid / normal (L=1 mu=est sigma=est);
run;

/* Sequence Plot to check Independence */                                   -(13)
data resids2;
  set residsF;
  order = _n_;
run;

goptions reset=all;
goptions hsize=5;
goptions vsize=4;
proc gplot data=resids2;
  plot resid*order / vaxis=axis1 haxis=axis2 ;
  title2 "Sequence plot of the residuals";
  axis1 label = (a=90 'Residual');
  axis2 label=('Order');
  symbol1 v=dot c=blue h=.8;
run;
```

```
/* Box Cox transformation */                                          -(14)
proc transreg data=sales maxiter=0 nozeroconstant;
    model BoxCox(price) = identity(quality|newstyle);
    output;
run;

/* Estimated Lambda value turned out to be -0.5*/
data newsales;
set sales;
newprice = 1/sqrt(price);
run;

/* TWO Way ANOVA again with changed value due to Box Cox transformation */
proc glm data=newsales;
    class quality newstyle;
    model newprice = quality|newstyle;
    means quality*newstyle;
    output out=temp r=resid p=predict;
run;

data residsF;
    set temp;
run;

goptions hsize=5;
goptions vsize=4;
proc gplot data=residsF;
plot resid*predict;
run;

goptions hsize=5;
goptions vsize=4;
proc univariate data=residsF noprint;
       var resid;
       histogram resid/ normal;
       qqplot resid / normal (L=1 mu=est sigma=est);
run;

/* Sequence Plot to check Independence */
data resids2;
  set residsF;
  order = _n_;
run;

goptions hsize=5;
goptions vsize=4;
proc gplot data=resids2;
  plot resid*order / vaxis=axis1 haxis=axis2 ;
  title2 "Sequence plot of the residuals";
  axis1 label = (a=90 'Residual');
  axis2 label=('Order');
  symbol1 v=dot c=blue h=.8;
run;

proc glm data=newsales;                                              -(5)
       class quality newstyle;
```

```
        model newprice = quality newstyle quality*newstyle;
        means quality newstyle quality*newstyle / TUKEY alpha=0.1 cldiff;
run;


/* FOUR Way ANOVA*/

/* rearrange data so that numbers within certain categories could be categorized as 1, 2, 3
or 4. */                                        -(6)
data sales;
        set sales;
        if bed<=2 then newbed = 1;
        else if bed=3 then newbed = 2;
        else if bed>=4 then newbed = 3;
        if sqfeet <1000 then newsqfeet = 1;
        else if 1000<=sqfeet<2000 then newsqfeet = 2;
        else if 2000<=sqfeet<3000 then newsqfeet = 3;
        else if 3000<=sqfeet then newsqfeet = 4;
        if year<1930 then newyear = 1;
        else if 1930<=year<1965 then newyear = 2;
        else if 1965<=year<1980 then newyear = 3;
        else if year>=1980 then newyear = 4;
        if bath<=1 then newbath = 1;
        else if 2<=bath<=3 then newbath = 2;
        else if 4<=bath then newbath = 3;
run;




/* Four way ANOVA */                                                    -(7)
proc glm data=sales;
    class newsqfeet newbed newbath newyear;
    model price = newsqfeet|newbed|newbath|newyear;
    means newsqfeet*newbed*newbath*newyear;
    /*output out=temp r=resid p=predict;*/
run;

/* eliminated 4 way interaction */

proc glm data=sales;                                                    -(8)
    class newsqfeet newbed newbath newyear;
    model price = newsqfeet newbed newbath newyear
    newsqfeet*newbed newsqfeet*newbath newsqfeet*newyear
    newbed*newbath newbed*newyear newbath*newyear
    newsqfeet*newbed*newbath newsqfeet*newbed*newyear
    newsqfeet*newbath*newyear newbed*newbath*newyear;
run;

/* eliminated three way interaction except newsqfeet*newbath*newyear */
proc glm data=sales;                                                    -(9)
    class newsqfeet newbed newbath newyear;
    model price = newsqfeet newbed newbath newyear newsqfeet*newbed newsqfeet*newbath
newsqfeet*newyear
    newbed*newbath newbed*newyear newbath*newyear newsqfeet*newbath*newyear;
run;

/* eliminated three way interaction */
```

```
proc glm data=sales;                                                    -(10)
    class newsqfeet newbed newbath newyear;
    model price = newsqfeet newbed newbath newyear newsqfeet*newbed newsqfeet*newbath
newsqfeet*newyear
    newbed*newbath newbed*newyear newbath*newyear;
    output out=temp r=resid p=predict;
run;

/* eliminated 3 two-way interactions */
proc glm data=sales;                                                        -(11)
    class newsqfeet newbed newbath newyear;
    model price = newsqfeet newbed newbath newyear newsqfeet*newbed newsqfeet*newyear
    newbed*newbath;
    output out=temp r=resid p=predict;
run;


/* Again check normality, variance, and independence for ANOVA assumptions*/    -(16)
data residsF;
    set temp;
run;

goptions hsize=5;
goptions vsize=4;
proc gplot data=residsF;
plot resid*predict;
run;

goptions hsize=5;
goptions vsize=4;
proc univariate data=residsF noprint;
       var resid;
       histogram resid/ normal;
       qqplot resid / normal (L=1 mu=est sigma=est);
run;

goptions hsize=5;
goptions vsize=4;
proc gplot data=resids2;
  plot resid*order / vaxis=axis1 haxis=axis2 ;
  title2 "Sequence plot of the residuals";
  axis1 label = (a=90 'Residual');
  axis2 label=('Order');
  symbol1 v=dot c=blue h=.8;
run;


/* Box Cox transformation to find the optimal lambda */
proc transreg data=sales maxiter=0 nozeroconstant;
    model BoxCox(price) = identity(newsqfeet newbed newbath newyear newsqfeet*newbed
newsqfeet*newyear
    newbed*newbath);
    output;
run;

/* lambda turned out to be -0.5 again, so modify the data again with -0.5 lambda*/
data newsales2;
set sales;
newprice = price**(-1/2);
run;
```

```
/* Run ANOVA with the optimal model found by Box Cox transformation */
proc glm data=newsales2;
    class newsqfeet newbed newbath newyear;
    model newprice = newsqfeet newbed newbath newyear newsqfeet*newbed newbed*newbath
newsqfeet*newyear;
    output out=temp r=resid p=predict;
run;

data residsF;
    set temp;
run;

/* Find fitted vs residual, histogram, qqplot, and sequence plot to check ANOVA
assumptions*/

goptions hsize=5;
goptions vsize=4;
proc gplot data=residsF;
plot resid*predict;
run;

goptions hsize=5;
goptions vsize=4;
proc univariate data=residsF noprint;
       var resid;
       histogram resid/ normal;
       qqplot resid / normal (L=1 mu=est sigma=est);
run;

data resids2;
  set residsF;
  order = _n_;
run;

goptions hsize=5;
goptions vsize=4;
proc gplot data=resids2;
  plot resid*order / vaxis=axis1 haxis=axis2 ;
  title2 "Sequence plot of the residuals";
  axis1 label = (a=90 'Residual');
  axis2 label=('Order');
  symbol1 v=dot c=blue h=.8;
run;
```