Problem 1: Data Input

Obs	address	info	z_address	bathrooms	bedrooms	finishedsqft	lastsolddate	lastsoldprice	latitude	longitude	neighborhood	totalrooms	usecode	yearbuilt	zestimate	zipcode	zpid
1	Address: 1160 Mission Street #2007	San FranciscoSales price: 1300000Sales date: 02/17/2016Bedrooms: 2Square feet: 1043Lot size: 43557	1160 Mission St UNIT 2007	2	2	1043	20501	1300000	37.7787	-122.413	South of Market	4	Condominium	2007	1167508	94103	83152781
2	Address: 260 King Street #475	San FranciscoSales price: 750000Sales date: 02/17/2016Bedrooms: 1Square feet: 908Lot size: 213232	260 King St UNIT 475	1	1	903	20501	750000	37.7776	-122.393	South of Market	3	Condominium	2004	823719	94107	69819817
3	Address: 560 Missouri Street #B	San FranciscoSales price: 1495000Sales date: 02/17/2016Bedrooms: 3Square feet: 1425Lot size: 2496	560 Missouri St # B	4	3	1425	20501	1495000	37.7592	-122.397	Potrero Hill	6	Condominium	2003	1708594	94107	64972847
4	Address: 350 Missouri Street	San FranciscoSales price: 2700000Sales date: 02/17/2016Bedrooms: 3Square feet: 1984Lot size: 2469	350 Missouri St	3	3	2231	20501	2700000	37.7619	-122.397	Potrero Hill	10	SingleFamily	1927	2411236	94107	15149489
5	Address: 3658 Folsom Street	San FranciscoSales price: 1530000Sales date: 02/17/2016Bedrooms: 0Square feet: 924Lot size: 1750	3658 Folsom St	3	3	1300	20501	1530000	37.7408	-122.413	Bernal Heights	4	SingleFamily	1900	1918539	94110	15161978

Problem 2: Drop Tables (Data Statement)

Obs	bathrooms	bedrooms	finishedsqft	lastsolddate	lastsoldprice	totalrooms	usecode	yearbuilt	zestimate	zipcode
1	2	2	1043	20501	1300000	4	Condominium	2007	1167508	94103
2	1	1	903	20501	750000	3	Condominium	2004	823719	94107
3	4	3	1425	20501	1495000	6	Condominium	2003	1708594	94107
4	3	3	2231	20501	2700000	10	SingleFamily	1927	2411236	94107
5	3	3	1300	20501	1530000	4	SingleFamily	1900	1918539	94110

Problem 2: Drop Tables (SQL)

Obs	bathrooms	bedrooms	finishedsqft	lastsolddate	lastsoldprice	totalrooms	usecode	yearbuilt	zestimate	zipcode
1	2	2	1043	20501	1300000	4	Condominium	2007	1167508	94103
2	1	1	903	20501	750000	3	Condominium	2004	823719	94107
3	4	3	1425	20501	1495000	6	Condominium	2003	1708594	94107
4	3	3	2231	20501	2700000	10	SingleFamily	1927	2411236	94107
5	3	3	1300	20501	1530000	4	SingleFamily	1900	1918539	94110

Problem 3: Add Price Per Square Foot (Data Statement)

Obs	bathrooms	bedrooms	finishedsqft	lastsolddate	lastsoldprice	totalrooms	usecode	yearbuilt	zestimate	zipcode	price_per_square_foot
1	2	2	1043	20501	1300000	4	Condominium	2007	1167508	94103	1246.40
2	1	1	903	20501	750000	3	Condominium	2004	823719	94107	830.56
3	4	3	1425	20501	1495000	6	Condominium	2003	1708594	94107	1049.12
4	3	3	2231	20501	2700000	10	SingleFamily	1927	2411236	94107	1210.22
5	3	3	1300	20501	1530000	4	SingleFamily	1900	1918539	94110	1176.92

Problem 3: Add Price Per Square Foot (SQL)

Obs	bathrooms	bedrooms	finishedsqft	lastsolddate	lastsoldprice	totalrooms	usecode	yearbuilt	zestimate	zipcode	price_per_square_foot
1	2	2	1043	20501	1300000	4	Condominium	2007	1167508	94103	1246.40
2	1	1	903	20501	750000	3	Condominium	2004	823719	94107	830.56
3	4	3	1425	20501	1495000	6	Condominium	2003	1708594	94107	1049.12
4	3	3	2231	20501	2700000	10	SingleFamily	1927	2411236	94107	1210.22
5	3	3	1300	20501	1530000	4	SingleFamily	1900	1918539	94110	1176.92

Problem 4: Average of Last Sold Price by Zipcode (Data Statement)

The MEANS Procedure

Analysis Variable : lastsoldprice		
zipcode	N Obs	Mean
94102	318	835957.49
94103	360	1057806.08
94104	1	1300000.00
94105	340	1202028.24
94107	857	1115271.06
94108	114	1479360.53
94109	653	1274315.54
94110	935	1235720.02
94111	38	1078907.89

Analysis Variable : lastsoldprice		
zipcode	N Obs	Mean
94112	877	763993.29
94114	651	1685736.43
94115	534	1860896.42
94116	655	1019365.35
94117	473	1454416.58
94118	466	1835655.89
94121	527	1474686.53
94122	601	1070692.68
94123	430	2449347.67
94124	393	573805.64
94127	435	1365385.56
94131	687	1382432.57
94132	292	774159.25
94133	183	1718297.81
94134	439	650733.71
94158	71	976035.21

Problem 4: Average of Last Sold Price by Zipcode (SQL)

AveragePrice	zipcode
835957.5	94102
1057806	94103
1300000	94104
1202028	94105
1115271	94107
1479361	94108
1274316	94109
1235720	94110
1078908	94111
763993.3	94112
1685736	94114
1860896	94115
1019365	94116
1454417	94117
1835656	94118
1474687	94121
1070693	94122
2449348	94123
573805.6	94124
1365386	94127
1382433	94131
774159.2	94132
1718298	94133
650733.7	94134
976035.2	94158

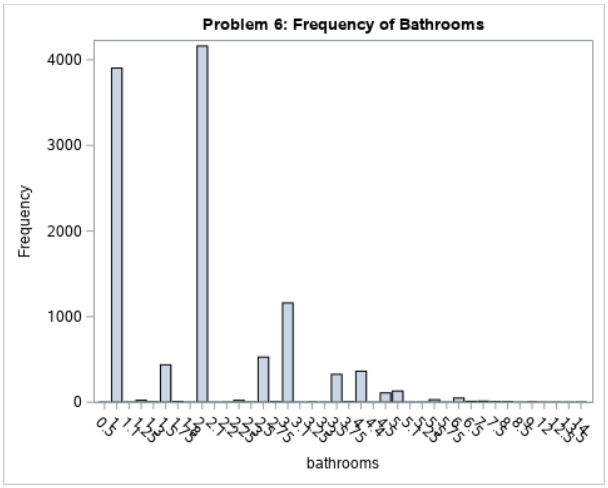
Problem 5: Average of Last Sold Price, Total Rooms, and Bedrooms by Usecode (Data Statement)

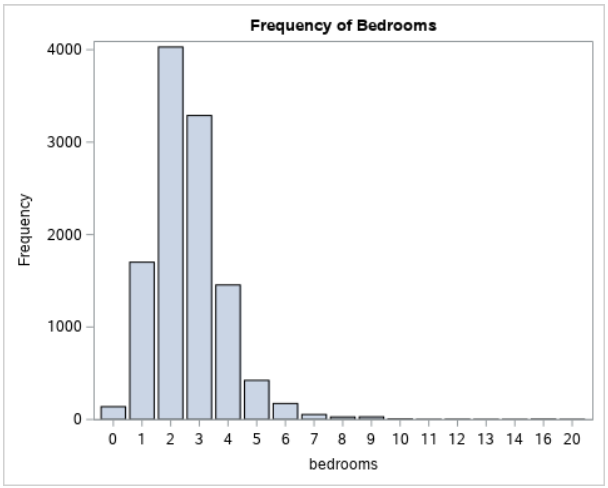
The MEANS Procedure

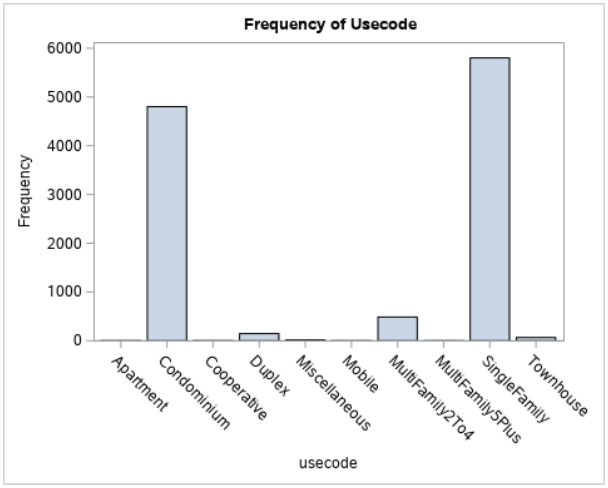
usecode	N Obs	Variable	Mean
Apartment	3	lastsoldprice totalrooms bedrooms	1341666.67 5.6666667 1.6666667
Condominium	4802	lastsoldprice totalrooms bedrooms	1083701.40 4.9952103 1.9137859
Cooperative	3	lastsoldprice totalrooms bedrooms	906666.67 11.0000000 6.3333333
Duplex	146	lastsoldprice totalrooms bedrooms	1416006.85 11.8904110 4.8082192
Miscellaneous	17	lastsoldprice totalrooms bedrooms	1579941.18 8.4705882 3.3529412
Mobile	2	lastsoldprice totalrooms bedrooms	1697000.00 5.0000000 2.5000000
MultiFamily2To4	486	lastsoldprice totalrooms bedrooms	1322210.88 8.9053498 3.7119342
MultiFamily5Plus	2	lastsoldprice totalrooms bedrooms	678000.00 5.0000000 2.0000000
SingleFamily	5803	lastsoldprice totalrooms bedrooms	1408017.87 6.6539721 3.0470446
Townhouse	66	lastsoldprice totalrooms bedrooms	878167.20 5.5454545 2.3333333

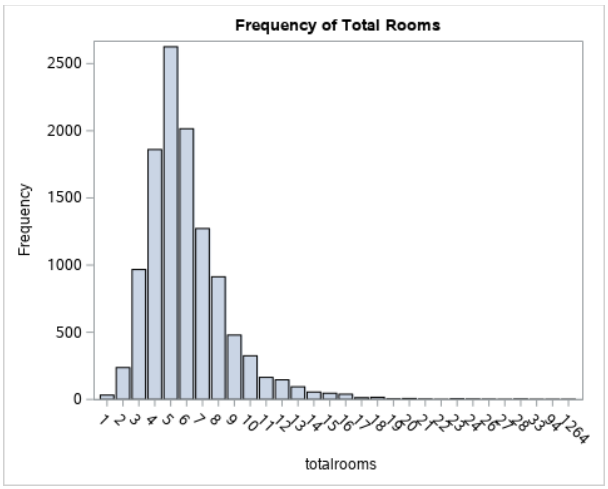
Problem 5: Average of Last Sold Price, Total Rooms, and Bedrooms by Usecode (SQL)

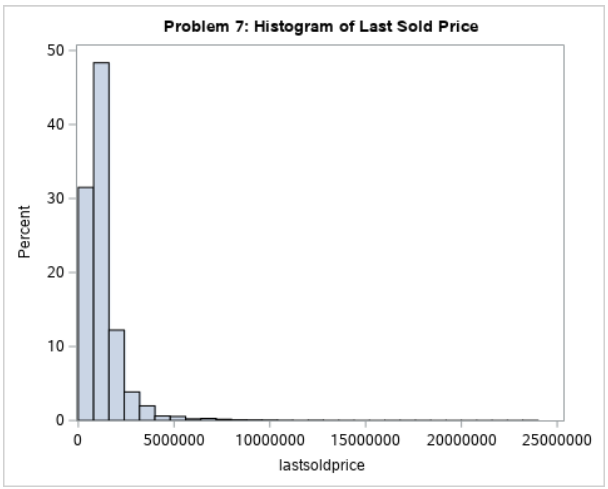
AveragePrice	AverageRooms	AverageBedrooms	usecode
1341667	5.666667	1.666667	Apartment
1083701	4.99521	1.913786	Condominium
906666.7	11	6.333333	Cooperative
1416007	11.89041	4.808219	Duplex
1579941	8.470588	3.352941	Miscellaneous
1697000	5	2.5	Mobile
1322211	8.90535	3.711934	MultiFamily2To4
678000	5	2	MultiFamily5Plus
1408018	6.653972	3.047045	SingleFamily
878167.2	5.545455	2.333333	Townhouse

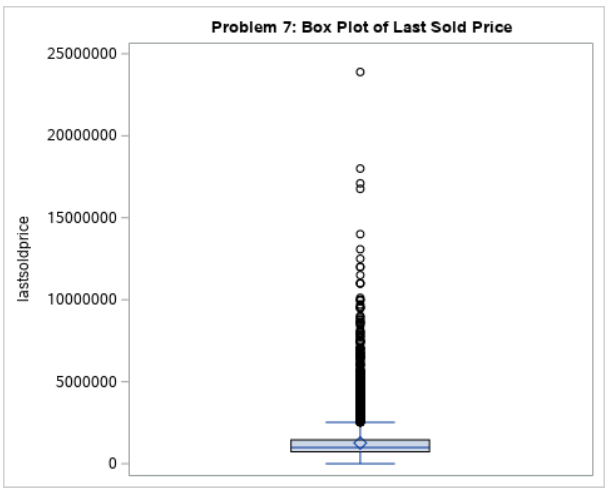


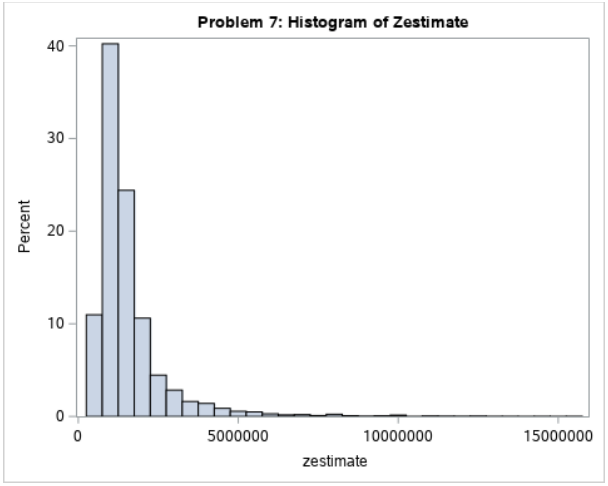


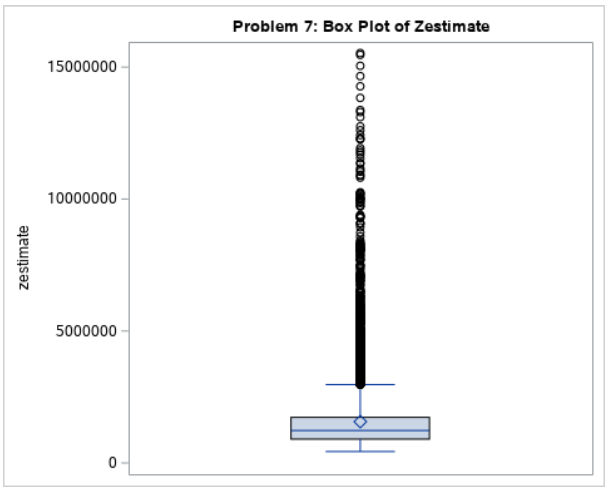



Frequency of Total Rooms
Frequency
2500
2000
1500
1000
500
0
totalrooms


Problem 7: Histogram of Last Sold Price
Percent
50
40
30
20
10
0
0
5000000
10000000
15000000
20000000
25000000
lastsoldprice


Problem 7: Box Plot of Last Sold Price
lastsoldprice
25000000
20000000
15000000
10000000
5000000
0





Problem 7: Median of Last Sold Price

The MEANS Procedure

Analysis Variable : lastsoldprice
Median
990000.00

Problem 7: Median of Zestimate

The MEANS Procedure

Analysis Variable : zestimate
Median
1230758.00

Problem 8: T-Test Between Zipcodes 94103 and 94107

The TTEST Procedure

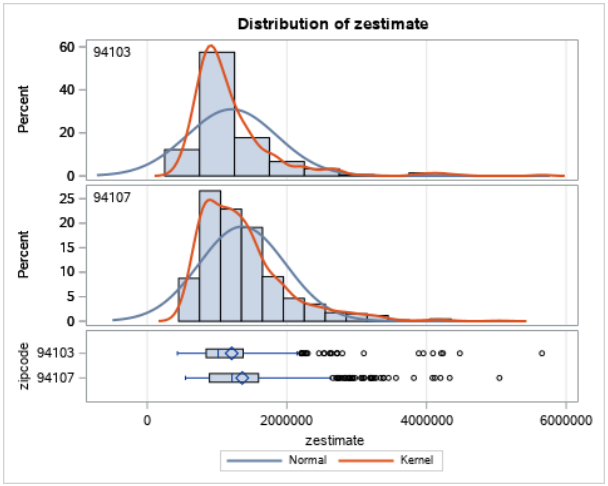
Variable: zestimate

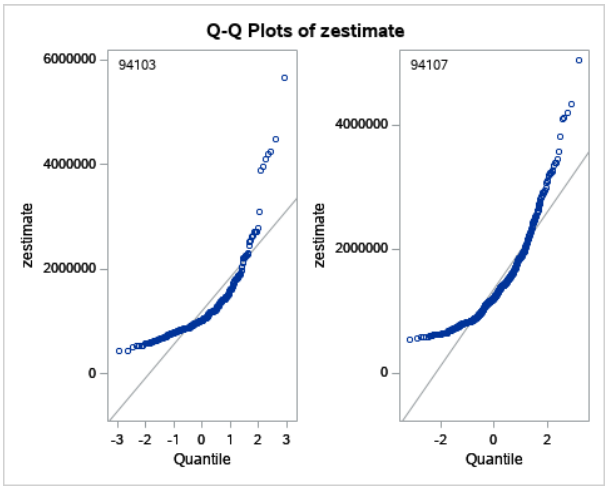
zipcode	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
94103		360	1210217	643797	33931.1	432385	5653999
94107		857	1361379	619847	21173.6	548210	5044710
Diff (1-2)	Pooled		-151163	627019	39380.8		
Diff (1-2)	Satterthwaite		-151163		39995.5		

zipcode	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
94103		1210217	1143488	1276945	643797	599954	694605
94107		1361379	1319821	1402938	619847	591827	650673
Diff (1-2)	Pooled	-151163	-228425	-73900.8	627019	603051	652985
Diff (1-2)	Satterthwaite	-151163	-229698	-72627.1			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1215	-3.84	0.0001
Satterthwaite	Unequal	651.59	-3.78	0.0002

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	359	856	1.08	0.3850





Problem 9: T-Test Between Zestimate and Last Sold Price

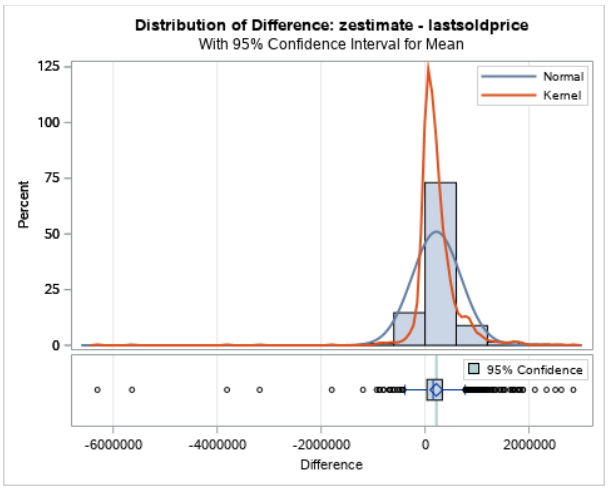
The TTEST Procedure

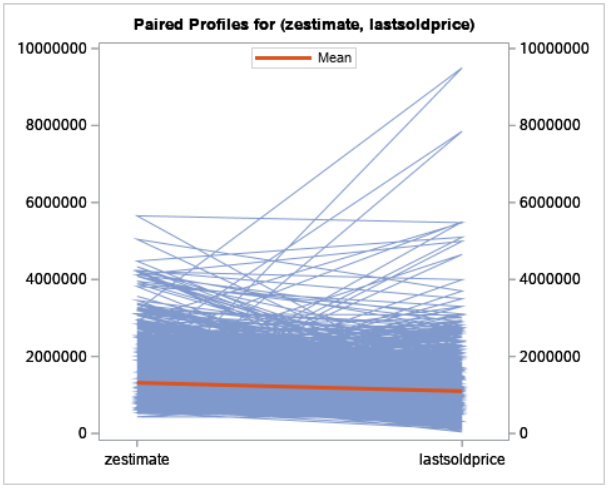
Difference: zestimate - lastsoldprice

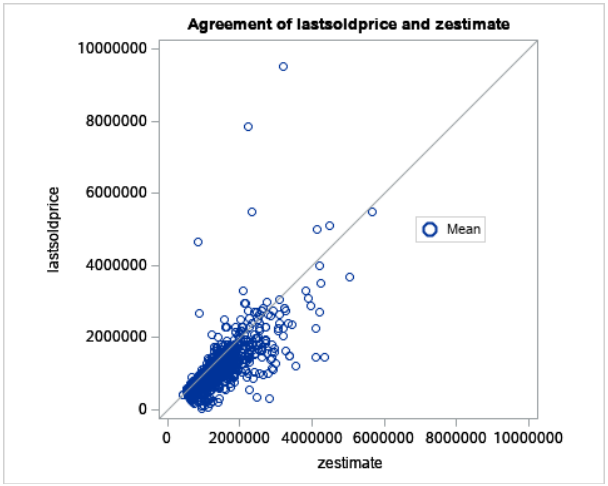
N	Mean	Std Dev	Std Err	Minimum	Maximum
1217	218392	469186	13449.3	-6303245	2853593

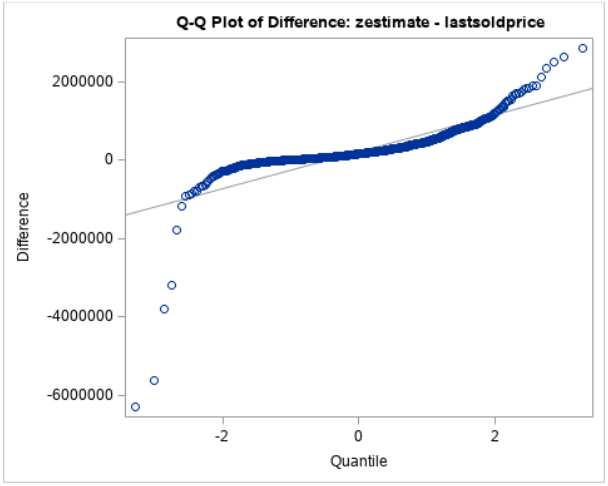
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
218392	192005	244778	469186	451259	488608

DF	t Value	Pr > |t|
1216	16.24	<.0001


Distribution of Difference: zestimate - lastsoldprice
With 95% Confidence Interval for Mean
Normal
Kernel
Percent
125
100
75
50
25
0
95% Confidence
-6000000
-4000000
-2000000
0
2000000
Difference


Paired Profiles for (zestimate, lastsoldprice)
Mean
10000000
8000000
6000000
4000000
2000000
0
zestimate
lastsoldprice


Agreement of lastsoldprice and zestimate
lastsoldprice
zestimate
Mean
10000000
8000000
6000000
4000000
2000000
0



Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=Apartment

1 Variables:	bedrooms

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	3	1.66667	0.57735	5.00000	1.00000	2.00000

Pearson Correlation Coefficients, N = 3 Prob > |r| under H0: Rho=0	
	bedrooms
bedrooms	1.00000

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=Condominium

1 Variables:	bedrooms

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	4802	1.91379	0.89040	9190	0	20.00000

Pearson Correlation Coefficients, N = 4802 Prob > |r| under H0: Rho=0	
	bedrooms
bedrooms	1.00000

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=Cooperative

1 Variables:	bedrooms

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	3	6.33333	6.65833	19.00000	2.00000	14.00000

Pearson Correlation Coefficients, N = 3 Prob > |r| under H0: Rho=0	
	bedrooms
bedrooms	1.00000

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=Duplex

1 Variables:	bedrooms

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	146	4.80822	1.76289	702.00000	1.00000	9.00000

Pearson Correlation Coefficients, N = 146 Prob > |r| under H0: Rho=0	
	bedrooms
bedrooms	1.00000

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=Miscellaneous

1 Variables:	bedrooms

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	17	3.35294	2.23442	57.00000	1.00000	10.00000

Pearson Correlation Coefficients, N = 17 Prob > |r| under H0: Rho=0	
	bedrooms
bedrooms	1.00000

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=Mobile

1 Variables:	bedrooms

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	2	2.50000	0.70711	5.00000	2.00000	3.00000

Pearson Correlation Coefficients, N = 2 Prob > |r| under H0: Rho=0	
	bedrooms
bedrooms	1.00000 .

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=MultiFamily2To4

1 Variables:	bedrooms

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	486	3.71193	2.57172	1804	0	16.00000

Pearson Correlation Coefficients, N = 486 Prob > |r| under H0: Rho=0	
	bedrooms
bedrooms	1.00000

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=MultiFamily5Plus

1 Variables:	bedrooms

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	2	2.00000	0	4.00000	2.00000	2.00000

Pearson Correlation Coefficients, N = 2 Prob > |r| under H0: Rho=0	
	bedrooms
bedrooms	. .

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=SingleFamily

1 Variables:	bedrooms

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	5803	3.04704	1.07012	17682	0	11.00000

Pearson Correlation Coefficients, N = 5803
Prob > |r| under H0: Rho=0

	bedrooms
bedrooms	1.00000

Problem 10: Correlation of Bedrooms with Usecode

The CORR Procedure

usecode=Townhouse

1 Variables:	bedrooms

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bedrooms	66	2.33333	0.96609	154.00000	1.00000	6.00000

Pearson Correlation Coefficients, N = 66
Prob > |r| under H0: Rho=0

	bedrooms
bedrooms	1.00000

Problem 11: Correlation of Bedrooms with Bathrooms

The CORR Procedure

1 With Variables:	bathrooms
1 Variables:	bedrooms

Simple Statistics

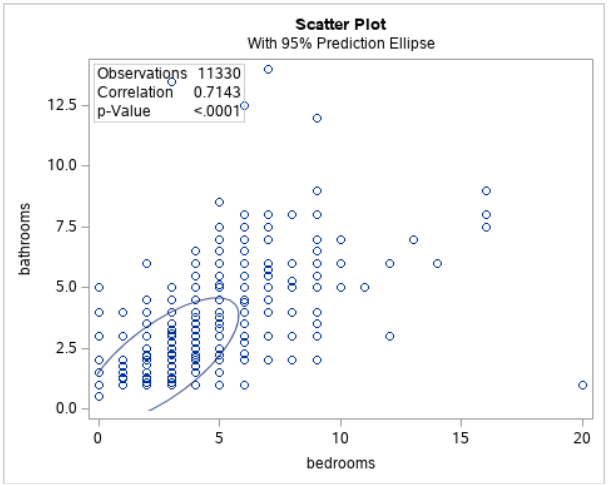
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
bathrooms	11330	1.98023	1.04736	22436	0.50000	14.00000
bedrooms	11330	2.61447	1.29946	29622	0	20.00000

Pearson Correlation Coefficients, N = 11330
Prob > |r| under H0: Rho=0

	bedrooms
bathrooms	0.71427 <.0001

Problem 11: Correlation of Bedrooms with Bathrooms

The CORR Procedure



Problem 12: Correlation Coefficients of All Numerical Variables with Zestimate

The CORR Procedure

1 With Variables:	zestimate
7 Variables:	bathrooms bedrooms finishedsqft lastsolddate lastsoldprice totalrooms price_per_square_foot

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum

Simple Statistics

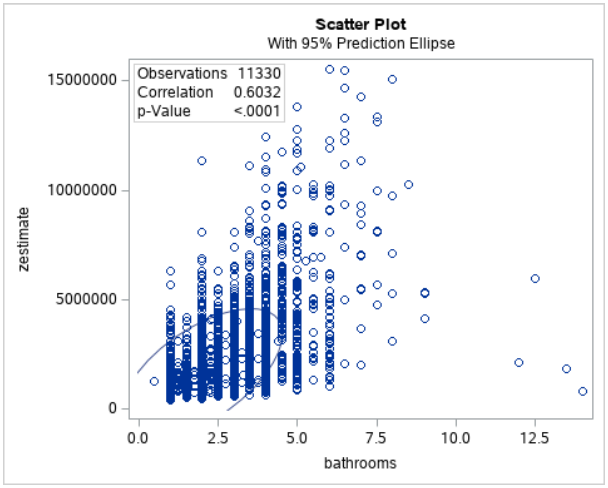
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
zestimate	11330	1565695	1229417	1.77393E10	432385	15533253
bathrooms	11330	1.98023	1.04736	22436	0.50000	14.00000
bedrooms	11330	2.61447	1.29946	29622	0	20.00000
finishedsqft	11330	1585	921.97825	17962819	1.00000	27275
lastsolddate	11330	19851	527.69286	224917260	3927	20541
lastsoldprice	11330	1263928	1042079	1.43203E10	535.00000	23889000
totalrooms	11330	6.11156	12.12582	69244	1.00000	1264
price_per_square_foot	11330	1761	33861	19949622	0.46725	1930000

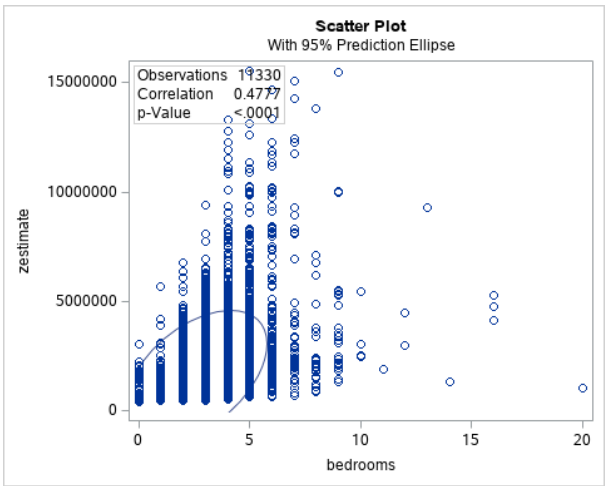
Pearson Correlation Coefficients, N = 11330
Prob > |r| under H0: Rho=0

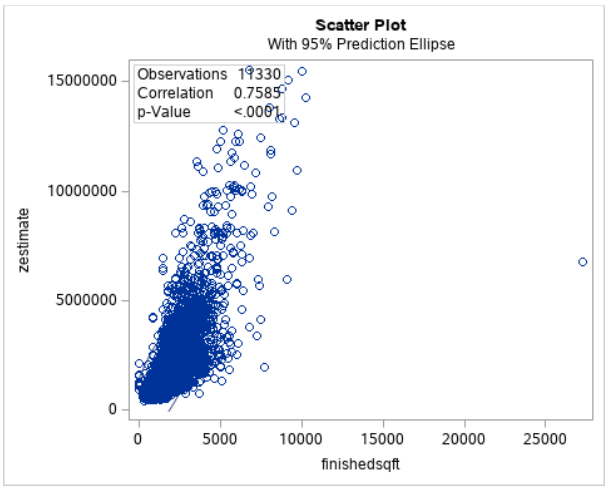
	bathrooms	bedrooms	finishedsqft	lastsolddate	lastsoldprice	totalrooms	price_per_square_foot
zestimate	0.60316 <.0001	0.47768 <.0001	0.75849 <.0001	-0.02256 0.0163	0.87694 <.0001	0.11199 <.0001	0.00093 0.9210

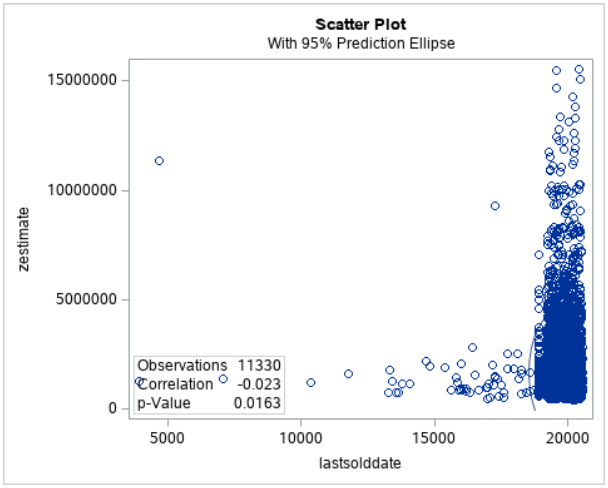
Problem 12: Correlation Coefficients of All Numerical Variables with Zestimate

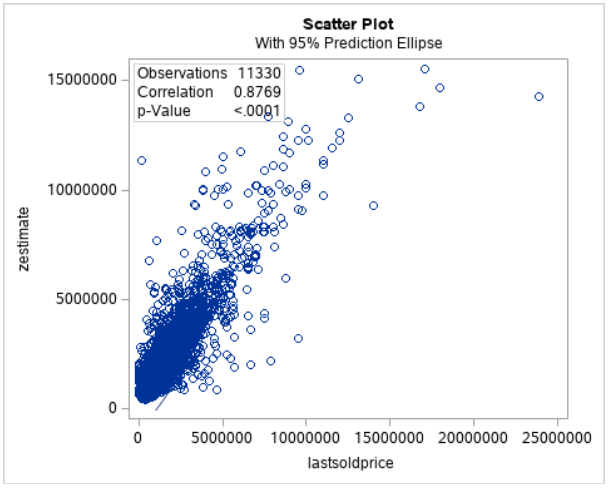
The CORR Procedure

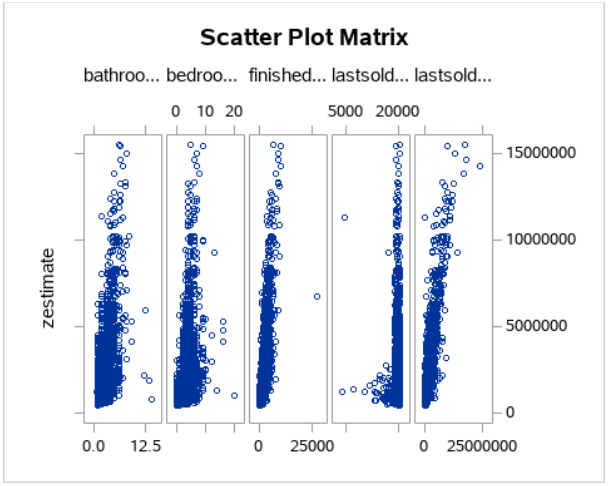





Scatter Plot
With 95% Prediction Ellipse
Observations 11330
Correlation 0.7585
p-Value <.0001
zestimate
15000000
10000000
5000000
0
0
5000
10000
15000
20000
25000
finishedsqft


Scatter Plot
With 95% Prediction Ellipse
Observations 11330
Correlation -0.023
p-Value 0.0163
zestimate
15000000
10000000
5000000
0
5000
10000
15000
20000
lastsolddate


Scatter Plot
With 95% Prediction Ellipse
Observations 11330
Correlation 0.8769
p-Value <.0001
zestimate
15000000
10000000
5000000
0
0
5000000
10000000
15000000
20000000
25000000
lastsoldprice



Problem 13: Multiple Linear Regression with Top 3 Correlated Variables

The REG Procedure
Model: MODEL1
Dependent Variable: zestimate

Number of Observations Read	11330
Number of Observations Used	11330

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.424625E16	4.748749E15	18693.6	<.0001
Error	11326	2.877154E15	2.540309E11		
Corrected Total	11329	1.71234E16			

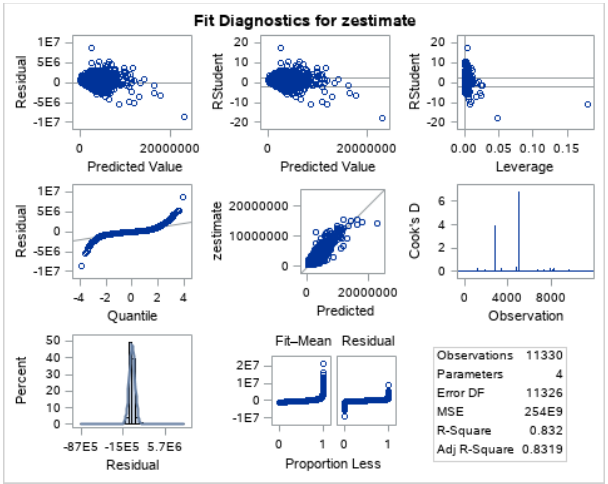
Root MSE	504015	R-Square	0.8320
Dependent Mean	1565695	Adj R-Sq	0.8319
Coeff Var	32.19112		

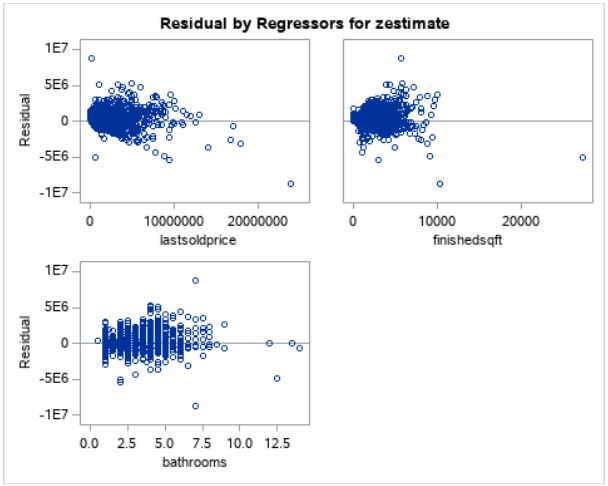
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-136521	10421	-13.10	<.0001
lastsoldprice	1	0.78058	0.00602	129.75	<.0001
finishedsqft	1	419.68249	8.25507	50.84	<.0001
bathrooms	1	25373	6566.18158	3.86	0.0001

Problem 13: Multiple Linear Regression with Top 3 Correlated Variables

The REG Procedure
Model: MODEL1
Dependent Variable: zestimate





Problem 14: Multiple Linear Regression with Top 3 Correlated Variables

The REG Procedure
Model: MODEL1
Dependent Variable: zestimate

Number of Observations Read	11330
Number of Observations Used	11330

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.426137E16	2.852273E15	11285.4	<.0001
Error	11324	2.862035E15	2.527406E11		
Corrected Total	11329	1.71234E16			

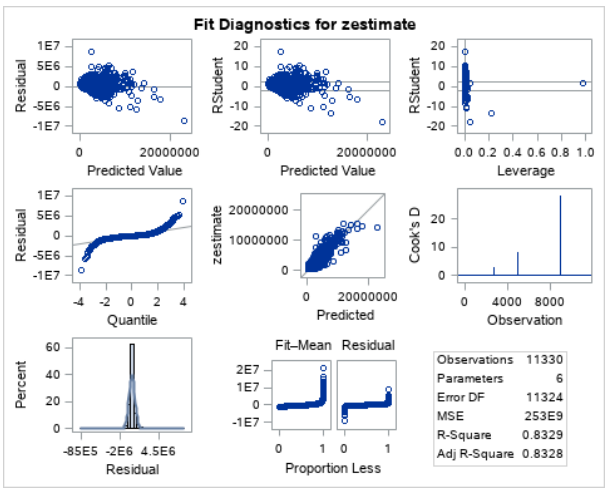
Root MSE	502733	R-Square	0.8329
Dependent Mean	1565695	Adj R-Sq	0.8328
Coeff Var	32.10926		

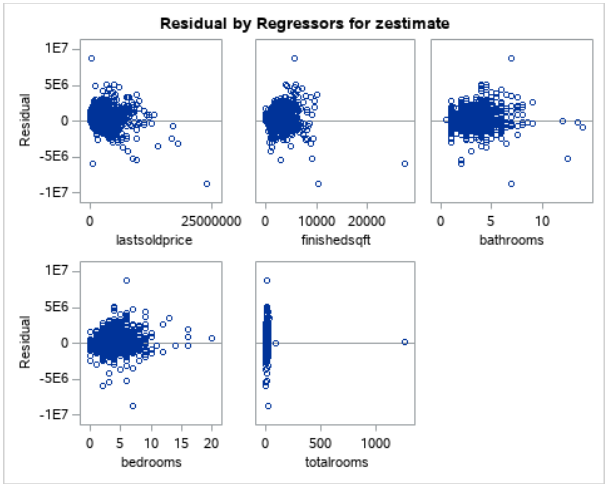
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-106325	11125	-9.56	<.0001
lastsoldprice	1	0.77215	0.00610	126.61	<.0001
finishedsqft	1	448.56394	9.05092	49.56	<.0001
bathrooms	1	50717	7323.48884	6.93	<.0001
bedrooms	1	-44038	5709.85853	-7.71	<.0001
totalrooms	1	-62.61864	395.24897	-0.16	0.8741

Problem 14: Multiple Linear Regression with Top 3 Correlated Variables

The REG Procedure
Model: MODEL1
Dependent Variable: zestimate





Problem 16: Predicted Zestimate for House

Obs	lastsoldprice	finishedsqft	bathrooms	bedrooms	totalrooms	MODEL1
1	2550000	2050	3.0	3	9	2801694.36
2	1300000	1325	1.0	2	5	1454148.65
3	5640000	5100	5.5	7	13	6506157.21
4	990000	1120	2.0	2	4	1173604.96

Problems 19/20: Average of price_per_square_foot by zipcode

Obs	zipcode	price_per_square_foot
1		1760.78
2	94102	928.85
3	94103	3018.24
4	94104	866.67
5	94105	2856.13
6	94107	1659.79
7	94108	1001.95
8	94109	919.28
9	94110	3474.67
10	94111	823.54
11	94112	573.12
12	94114	1012.22
13	94115	940.64
14	94116	650.60
15	94117	2691.56
16	94118	834.23
17	94121	695.39
18	94122	642.40
19	94123	1188.12
20	94124	431.21
21	94127	710.00
22	94131	3703.84
23	94132	3190.92
24	94133	7832.94
25	94134	499.77
26	94158	21914.18

Problems 19/20: Average of zestimate by totalrooms

Obs	totalrooms	zestimate
1	.	1565695.26
2	1	864987.34
3	2	790535.91
4	3	872228.57
5	4	1096498.89
6	5	1269967.61
7	6	1530213.50
8	7	1738700.67
9	8	1927580.83
10	9	2472959.43
11	10	2840516.66
12	11	3267640.75
13	12	3047366.25
14	13	3692985.06

Obs	totalrooms	zestimate
15	14	3530035.11
16	15	4273266.89
17	16	4120780.28
18	17	3178413.92
19	18	4178196.19
20	19	3739224.25
21	20	6404786.86
22	21	2012261.00
23	22	5684941.00
24	23	3106397.40
25	24	3662498.33
26	26	8765272.00
27	27	5430463.00
28	28	8381745.00
29	33	9304469.00
30	94	502285.00
31	1264	1591608.00