1.

a)

Sample mean and sample variance of both stores can be calculated as follows:

	Miller's	Albert's
	119.25	111.99
	121.32	114.88
	122.34	115.11
	120.14	117.02
	122.19	116.89
	123.71	116.62
	121.72	115.38
	122.42	114.4
	123.63	113.91
	122.44	111.87
Mean	121.916	114.807
Variance	1.955004	3.386712

Let $\overline{x}_1$ = Mean weekly expense of Miller's = 121.916 and $\overline{x}_2$ = Mean weekly expense of Albert's = 114.807

Two samples are assumed to be independent and have equal variances. So, we can calculate the Pooled estimate of Standard Error:

$$se_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{1.955004}{10} + \frac{3.386712}{10}} = 0.731$$

This is a two-tailed hypothesis testing with the mean different between $\overline{x}_1$ and $\overline{x}_2$ = 0

Test statistic can be calculated as:

$$t = \frac{(121.916 - 114.807) - 0}{0.731} = 9.73$$

At 0.05 level of significance with degree of freedom = 18, the cutoff t-statistic = 2.101

Since the Test statistic value exceeds the cutoff t-statistic, we reject the null Hypothesis and conclude that the mean weekly expense between Miller's and Albert's supermarket chains are different <u>ANS</u>

b)

A 95% confidence interval band can be calculated as follows:

$$(121.916 - 114.807) \underline{+}\, 2.101 * 0.731 = [8.645\,, 5.573]$$

So, at 95% confidence interval, the mean weekly expense different between Miller's and Albert's is between \$8.645 and \$5.573 <u>ANS</u>

c)

Null Hypothesis: $H_0: \mu_M - \mu_A \leq 5$

Alternative Hypothesis: $H_a: \mu_M - \mu_A > 5$

From a), the Test statistic can be calculated as follows:

$$t = \frac{(121.916 - 114.807) - 5}{0.731} = 2.885$$

This is a one-tailed hypothesis testing. At 0.05 level of significance with degree of freedom = 18, the cutoff t-statistic = 1.734

Since the Test statistic value exceeds the cutoff t-statistic, we reject the null Hypothesis and conclude that the mean weekly expense of Miller's supermarket chain is more than \$5 higher than Albert's.<u>ANS</u>

2.

a)

Means and Std Deviations

Level	Number	Mean
Condo/Coop	73	1497997
Single Family Residential	144	6871277

For Home Type, as Condo/Coop estates generally have a lower price, it is set as a baseline. A dummy variable 'is_single_family_resident' is created and has a value of 1 if the Home Type is 'Single Family Residential', 0 otherwise.

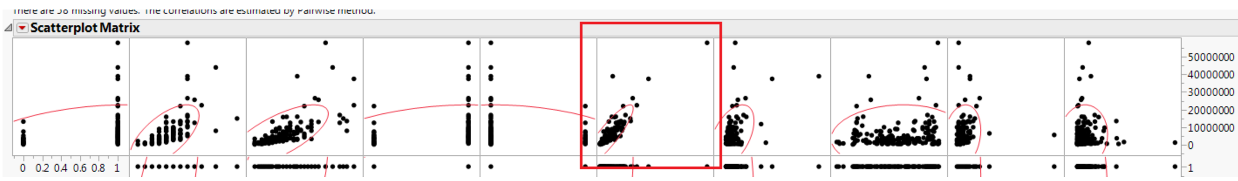
Means and Std Deviations

Level	Number	Mean
Beverly Hills	132	6996606
Downtown Los Angeles	11	467173
Santa Monica	74	2299011

For Location, Downtown LA estates generally have a lower price, 2 dummy variable are crated:

'is_Beverly_Hills' – equals 1 if the location is Beverly Hills, 0 otherwise.

'is_Santa_Monica' – equals 1 if the location is Santa Monica, 0 otherwise.



From the correlation matrix, SQFT(Square footage of the living space) seems to be the best predictor. ANS

b)

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-765663.5	302863.9	-2.53	0.0122*
SQFT	1305.8265	48.68377	26.82	<.0001*

Fit price by SQFT, the regression function is:

$$Price = -765663.5 + 1305.83 * SQFT$$

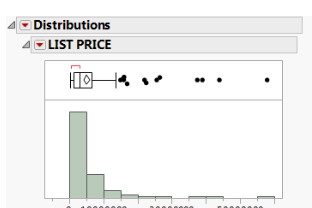
ANS

c)

Summary of Fit

RSquare	0.780781
RSquare Adj	0.779696
Root Mean Square Error	3167283
Mean of Response	4767322
Observations (or Sum Wgts)	204

Standard error of the estimate is $3,167,283. This is lower than the standard deviation of Y variable(Price), which is $7,230,747. That is, using SQFT is helpful in predicting the real estate price, but the error is still high because the Price distribution is heavily right skewed:



ANS

d)

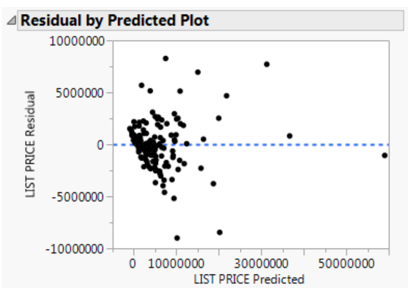
C. Total	158	8.4716e+15	<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	29523154	15786815	1.87	0.0634	.
is_single_family_resident	905992.75	652787.9	1.39	0.1673	2.170404
BEDS	-46627.99	230373.9	-0.20	0.8399	5.2217297
BATHS	-87015.75	167918.5	-0.52	0.6051	6.82847
is_Beverly_Hills	917722.71	1078009	0.85	0.3960	7.5095682
is_Santa_Monica	1828304	1020278	1.79	0.0752	6.2238909
SQFT	1255.1488	60.39279	20.78	<.0001*	2.5719178
LOT SIZE	55.307427	4.456459	12.41	<.0001*	1.1915946
YEAR BUILT	-16950.61	7894.134	-2.15	0.0334*	1.4201257
PARKING SPOTS	61365.937	37801.16	1.62	0.1066	1.09739
DAYS ON MARKET	814.07215	1167.375	0.70	0.4867	1.0577051

By fitting the model to all variables, some variables have Variance Inflation Factor that exceed 5, but none has VIF more than 10. According to the guideline, this shows that multicollinearity exists, but not severe. If we were serious about the multicollinearity, I'd try to drop the variable that has highest VIF first and iteratively repeat the step if VIF of the remained variables are not within the safe area. I will assume it's a benign problem here. ANS

e)



According to the Residual plot, there is a heteroscedasticity issue in the error term. This is likely because of the right skewed distribution of the dependent variable. I'd try to apply a log transformation to the Y(Price) variable. ANS

f)

After applying Ln function to the price variable (to adjust for fanned out error term) and iteratively remove variables that has coefficient with P-value exceed 5% significance level (remove variables that may not play a significant role in the prediction), I get the following model:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	12.375266	0.187541	65.99	<.0001*	.
is_single_family_resident	0.9051683	0.117247	7.72	<.0001*	1.7232527
BATHS	0.1494584	0.023225	6.44	<.0001*	3.2663137
is_Beverly_Hills	0.835294	0.216136	3.86	0.0002*	7.5583244
is_Santa_Monica	0.7284416	0.203187	3.59	0.0004*	6.1805289
SQFT	0.0000535	1.218e-5	4.39	<.0001*	2.5829222
LOT SIZE	3.4316e-6	8.863e-7	3.87	0.0002*	1.1778144

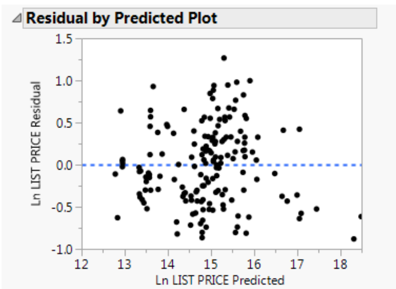
Summary of Fit

RSquare	0.826102
RSquare Adj	0.819498
Root Mean Square Error	0.473839
Mean of Response	14.90058
Observations (or Sum Wgts)	165

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	168.52280	28.0871	125.0965
Error	158	35.47476	0.2245	Prob > F
C. Total	164	203.99756		<.0001*

The residual plot looks much better:

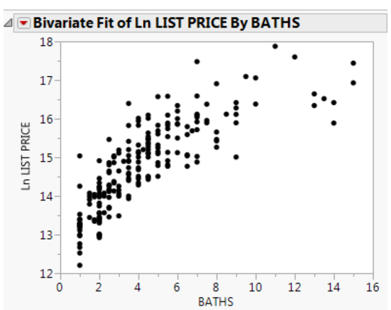


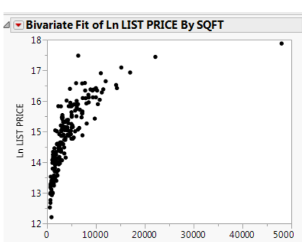
The equation of the final model is:

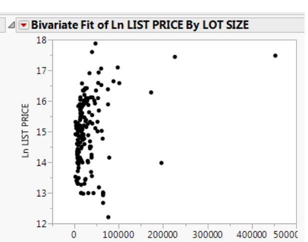
$$\ln(Price) = 12.38 + 0.91 * IS\ SINGLE\ FAMILY\ RESIDENT + 0.15 * BATHS + 0.84 * IS\ BEVERLY\ HILLS + 0.73 * IS\ SANTA\ MONICA + 0.00005 * SQFT + 0.000003 * LOT\ SIZE$$

ANS

g)

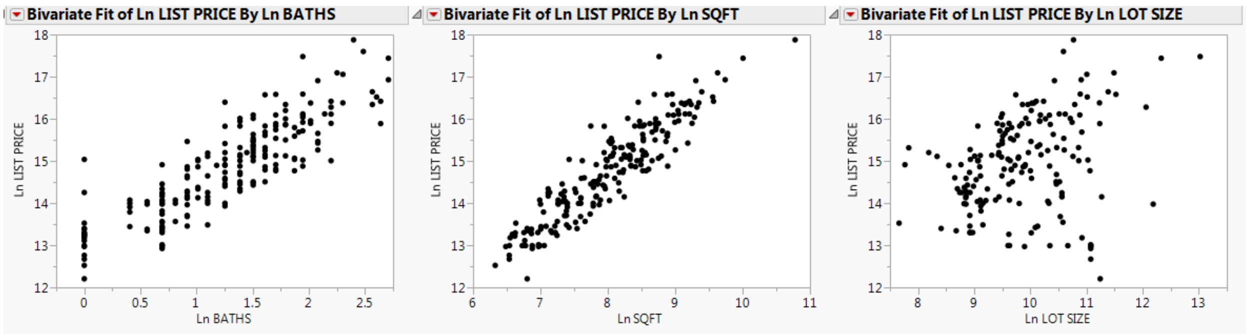






Inspecting the scatter plots, BATHS, SQFT, and LOT SIZE may have a non-linear relationship with Ln(Price). My recommendation is to try applying Log transformation to all 3 variables. Reciprocal transformation in X is not recommended because I don't think we have a theoretical reason to believe that Ln(Price), a % increase in Price, will be bounded by those three variables.

After applying the transformation, I get the following scatterplots on those 3 variables:



ANS

h)

Fitting the model with the new transformed variable, I get the following data:

Summary of Fit

RSquare	0.886898
RSquare Adj	0.882603
Root Mean Square Error	0.382137
Mean of Response	14.90058
Observations (or Sum Wgts)	165

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	180.92499	30.1542	206.4945
Error	158	23.07257	0.1460	Prob > F
C. Total	164	203.99756		<.0001*

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	3.7203644	0.797445	4.67	<.0001*	.
is_single_family_resident	0.5215365	0.11322	4.61	<.0001*	2.4706601
is_Beverly_Hills	0.5220293	0.180107	2.90	0.0043*	8.0697036
is_Santa_Monica	0.7618128	0.170319	4.47	<.0001*	6.6769969
Ln BATHS	0.0840979	0.13419	0.63	0.5318	8.6276345
Ln SQFT	0.9932456	0.11578	8.58	<.0001*	10.151634
Ln LOT SIZE	0.2041793	0.042455	4.81	<.0001*	1.5554981

As the P-Value of Ln(BATHS) is quite high, I tried dropping it and get the following result:

Summary of Fit

RSquare	0.886617
RSquare Adj	0.883051
Root Mean Square Error	0.381407
Mean of Response	14.90058
Observations (or Sum Wgts)	165

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	180.86764	36.1735	248.6645
Error	159	23.12992	0.1455	Prob > F
C. Total	164	203.99756		<.0001*

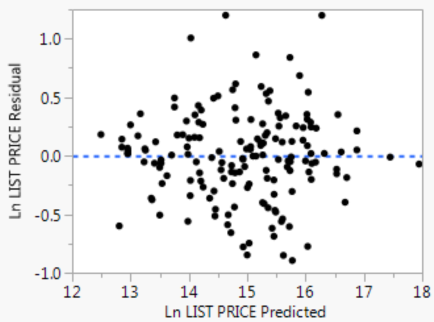
Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	3.3240181	0.484824	6.86	<.0001*	.
is_single_family_resident	0.5162744	0.112692	4.58	<.0001*	2.4570721
is_Beverly_Hills	0.5260608	0.179649	2.93	0.0039*	8.0594092
is_Santa_Monica	0.7584207	0.169907	4.46	<.0001*	6.6702535
Ln SQFT	1.0551262	0.060346	17.48	<.0001*	2.7683815
Ln LOT SIZE	0.2049703	0.042355	4.84	<.0001*	1.5541233

Effect Tests

Residual by Predicted Plot



As the doing this improved Adjusted R-Square from 0.8826 to 0.8830, Residual plot looks nicer, and VIF of all independent variables looks fine, I decided this to be the final model.

The final model is:

$$\ln(Price) = 3.32 + 0.52 * IS\ SINGLE\ FAMILY\ RESIDENT + 0.53 * IS\ BEVERLY\ HILLS + 0.76 * IS\ SANTA\ MONICA + 1.06 * \ln(SQFT) + 0.20 * \ln(LOT\ SIZE)$$

ANS

i)

On average, the list price of Single Family Resident real estate is 52% higher than Condo/ Coop, holding other variables constant.

On average, the list price of a real estate in Beverly Hills is 53% higher than in Downtown LA, holding other variables constant.

On average, the list price of a real estate in Santa Monica is 76% higher than in Downtown LA, holding other variables constant.

On average, increase a square footage of the living space by 1% resulted in 1.06% increase in list estate price, holding other variables constant.

On average, increase a square footage of the lot property by 1% resulted in 0.20% increase in list price, holding other variables constant.

ANS

3.

a) The result of fitting model is as follow:

Summary of Fit

RSquare	0.961683
RSquare Adj	0.959494
Root Mean Square Error	1113.093
Mean of Response	4499.184
Observations (or Sum Wgts)	38

Analysis of Variance

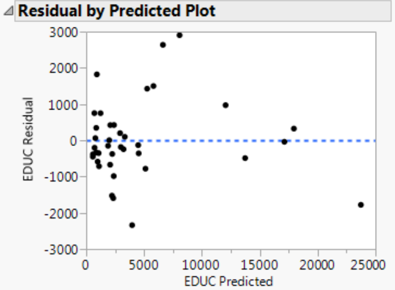
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1088365570	544182785	439.2200
Error	35	43364140	1238975.4	Prob > F
C. Total	37	1131729710		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	414.45834	266.7247	1.55	0.1292	.
GDP	0.0523186	0.00185	28.27	<.0001*	1.2644503
POP	-50.04769	9.958169	-5.03	<.0001*	1.2644503

Effect Tests

Residual by Predicted Plot



The equation is:

$$EDUC = 414.46 + 0.05 * GDP - 50.05 * POP$$

The model has relevant values:

- F-Statistic = 439.22, with P-Value < 0.0001
- t-Statistics of the intercept term = 1.55, with P-Value = 0.13
- t-Statistics of GDP = 28.27, with P-Value < 0.0001
- t-Statistics of POP = -5.03, with P-Value < 0.0001
- R-Square and Adjusted R-Square are around 0.96
- Standard Error of the estimate = 1113

ANS

b) The result of fitting model is as follow:

Summary of Fit

RSquare	0.885311
RSquare Adj	0.878757
Root Mean Square Error	0.441905
Mean of Response	7.712535
Observations (or Sum Wgts)	38

Analysis of Variance

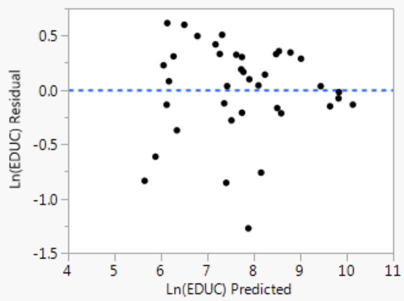
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	52.759224	26.3796	135.0863
Error	35	6.834792	0.1953	Prob > F
C. Total	37	59.594016		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-5.431023	0.810184	-6.70	<.0001*	.
Ln(GDP)	1.2328757	0.079918	15.43	<.0001*	1.2623859
Ln(POP)	-0.15594	0.078607	-1.98	0.0552	1.2623859

Effect Tests

Residual by Predicted Plot



The equation is:

$$\ln(EDUC) = -5.43 + 1.23 * \ln(GDP) - 0.16 * \ln(POP)$$

The model has relevant values:

- F-Statistic = 135.09, with P-Value < 0.0001
- t-Statistics of the intercept term = -6.70, with P-Value < 0.001
- t-Statistics of ln(GDP) = 15.43, with P-Value < 0.0001
- t-Statistics of ln(POP) = -1.98, with P-Value = 0.0552
- R-Square and Adjusted R-Square are around 0.96
- Standard Error of the estimate = 0.44

<u>ANS</u>

c)

- On average, 1% increase in GDP resulted in 1.23% increase in education expenditure
- On average, 1% increase in POP(population) resulted in 0.16% decrease in education expenditure

<u>ANS</u>

d) Even though the R-Square of the log-linear model is lower and the P-value of ln(POP) is a bit on a high side, I think the log-linear is the appropriate model because:

- Residual plot of the log-linear model is more satisfactory. Heteroscedasticity does not look noticeably bad.

- The model’s interpretation sounds more plausible to me; increasing in a number of population has a diminishing effect on a reduction in an education expense. It should not go straight down to zero.

ANS