Dataset Description

The data NorthValleyRealtor.xlsx contain information on homes sold by the North Valley Real Estate group within the last year. Within this file you will find the following fields:

- Record - Property identification number
- Agent – Name of the real estate agent assigned to the property
- Price – Market price in US dollars
- Size – Livable square feet of the property
- Bedrooms – Number of bedrooms
- Baths – Number of baths, which takes numbers 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5
- Pool – Does this home have a pool? (1 = yes, 0 = no)
- Garage – Does the home have a detached garage? (1 = yes, 0 = no)
- Days – Number of days the property was on the market
- Township – Area where the property is located
- Mortgage type – Fixed or adjustable. The fixed mortgage is a 30 year, fixed interest rate loan. The adjustable rate loan begins with an introductory rate of 3% for the first five years, then the interest rate is based on the current interest rates plus 1% (i.e., the interest rate AND the payment is likely to change each year after the 5^{th} year.).
- Years – The number of years that the mortgage loan has been paid
- FICO – the credit score of the mortgage loan holder. The highest score is 850; an average score is 680; a low score is below 680. The score reflects a person's ability to pay their debts.
- Default – Is the mortgage loan in default? (1 = yes, 0 = no)

Correlation and Simple Linear Regression (35pts)

1. If we perform a test to investigate the correlation between Price and Size. State the null hypothesis and alternate hypothesis for this research question. (1pt*2=2pts)
   H_0: The market price is correlated with the size. (Or the correlation is zero between market price and the size)
   H_A: The market price is not correlated with the size. (Or the correlation is non-zero between market price and the size)
   Acceptable answers just need to reflect the zero/nonzero relationship.

2. What is the value of correlation and its p-value? Support your answer with an SAS output **with your first and last name in the footnote.**
   Correlation is 0.95155 and p-value is <0.0001. (1pt*2=2pts)

Pearson Correlation Coefficients, N = 105 Prob > |r| under H0: Rho=0		
	Price	Size
Price Price	1.00000	0.95155 <.0001
Size Size	0.95155 <.0001	1.00000

First and Last name

   Your output doesn't have to be this correlation table, but it must show the correlation, p-value and your name. (1pt)

3. Create a pairwise correlation matrix for variables Price, Bedrooms, and FICO. Which pair has significant linear relationship and so we can reject null hypothesis of zero relationship between them?

Pearson Correlation Coefficients, N = 105 Prob > |r| under H0: Rho=0			
	Price	FICO	Bedrooms
Price Price	1.00000	-0.00766 0.9382	0.84444 <.0001
FICO FICO	-0.00766 0.9382	1.00000	0.08928 0.3651
Bedrooms Bedrooms	0.84444 <.0001	0.08928 0.3651	1.00000

   (3pts)
   The p-value of correlation between Bedrooms and Price is <0.0001 so their correlation is significant. Other pairs are not significant. (2pts)

4. Build a simple linear relationship between Price (dependent variable) and Size. What is the R-square, parameter estimate and confidence interval? Take a screenshot of the output showing R-square, parameter estimate and confidence interval. **Make sure to add your first and last name as the footnote.**
   R-square is 0.9054. (1pt)
   Intercept estimate is -15776 and coefficient estimate of Size is 108.364. (0.5pt*2=1pt)
   The confidence interval for intercept is (-41205, 9653.45976). (1pt)

The confidence interval for size coefficient is (101.520, 115.207) (1pt)

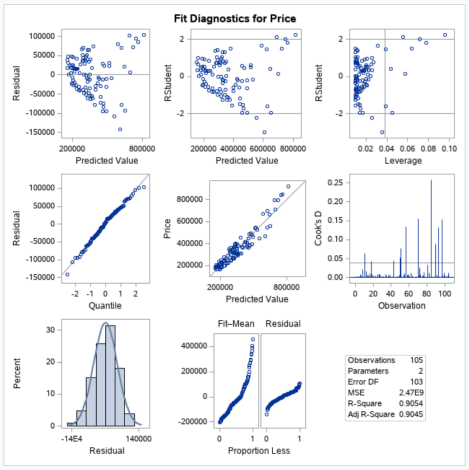
Root MSE	49656	**R-Square**	0.9054
Dependent Mean	357026	**Adj R-Sq**	0.9045
Coeff Var	13.90816		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	-15776	12822	-1.23	0.2214	-41205	9653.45976
Size	Size	1	108.36378	3.45058	31.40	<.0001	101.52037	115.20718

first and last name (1pt)

5. Take a screenshot of the diagnostic plot (the panel of plots) of the model between Price and Size. Are the assumptions satisfied?



(2pts)

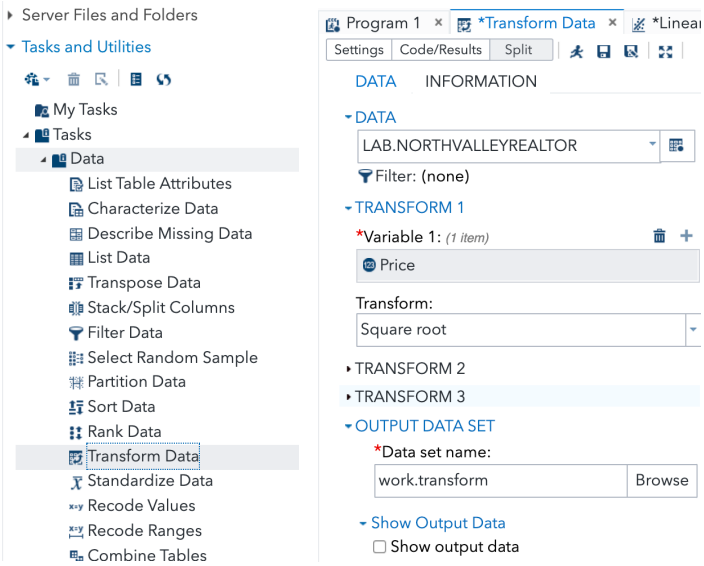
The residual or Studentized residual plots show that the linearity is satisfied because we don't see any curvature. (2pt)

Some outliers seem to be in the residual plot. We can also use Cook's D for outlier and see several values much higher than 4/n=4/105=0.038. (2pt)

Also, there is obvious non-constant variance. (2pt)

From QQ plot, the residuals follow the normal distribution. (2pt)

6. Because some assumptions are not satisfied, we decide to transform the dependent variable Price. Under Tasks, expand Data, select Transform Data. In the opened Transform Data window, under Data tab, select data, and choose 'Square root' transform on the Price variable as pictures shown below. Save transformed results to the default library and data as shown under 'Data set name'. Click run and you will see a column sqrt_Price attached to the original data which is the square root transformed Price.



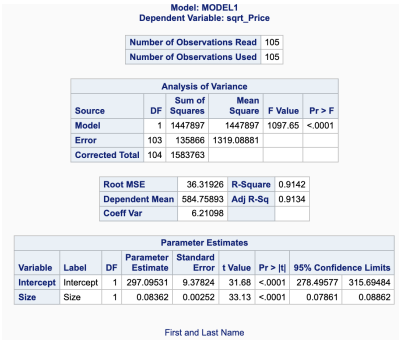
Now you can build a model between sqrt_Price (dependent variable) and size. What is the R-square, parameter estimate and confidence interval? Take a screenshot of the output showing R-square, parameter estimate and confidence interval. **Make sure to add your first and last name as the footnote.**

R-square is 0.9134. (1pt)

Intercept estimate is 297.095 and coefficient estimate of Size is 0.08362. (0.5pt*2=1pt)

The confidence interval for intercept is (278.49577, 315.69484). (1pt)

The confidence interval for size coefficient is (0.07861, 0.08862) (1pt)



Model: MODEL1
Dependent Variable: sqrt_Price

Number of Observations Read	105
Number of Observations Used	105

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1447897	1447897	1097.65	<.0001
Error	103	135866	1319.08881		
Corrected Total	104	1583763			

Root MSE	36.31926	R-Square	0.9142
Dependent Mean	584.75893	Adj R-Sq	0.9134
Coeff Var	6.21098		

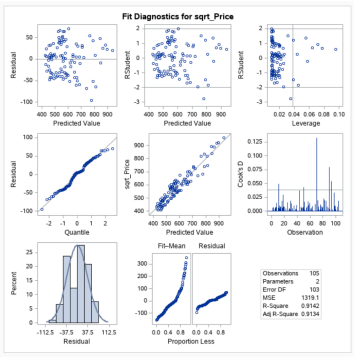
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	297.09531	9.37824	31.68	<.0001	278.49577	315.69484
Size	Size	1	0.08362	0.00252	33.13	<.0001	0.07861	0.08862

First and Last Name

(1pt)

7. Take a screenshot of the improved diagnostic plot (the panel of plots) of the model between sqrt_Price and Size. Are the assumptions better satisfied?



(2pts)

Both of nonconstant variance and outlier problem are improved. (3pts: To get the full credits, you need to at least state that the nonconstant variance is improved. You can also mention that there are less outliers and they look closer to majority of the data.)

Multiple Linear Regression (20 pts)

1. Build a model using Price as dependent variable, size, bedrooms, Baths, and mortgage type as predictors with NO selection method. Note that the categorical variable e.g., mortgage type here must be added into the Classification Variable box. Is the overall model significant? What is the degree of freedom of the model and so how many parameters are estimated in the model according to the DF? What is the sample size according to the DF of corrected total? What is the R-square and adjusted R-square? Support you answers with appropriate screenshots. What is the benefit of using adjusted R-square in multiple linear regression generally?
   The model is significant with p-value <0.0001. (1pt)
   The DF for model is 4, so there are 4+1=5 parameters including the intercept. (1pt)
   DF of total is n-1=104, so there are 105 observations. (1pt)
   R-square=0.908 and Adj R-square=0.9043. (0.5pt*2=1pt)
   Adjusted R-square penalizes addition of extraneous predictors to the MLR model. (1pt)

Least Squares Model (No Selection)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2.438732E12	6.096831E11	246.82	<.0001
Error	100	2.470191E11	2470190507		
Corrected Total	104	2.685751E12			

Root MSE	49701
Dependent Mean	357026
R-Square	0.9080
Adj R-Sq	0.9043
AIC	2382.77095
AICC	2383.62809
SBC	2289.04075

(1pt)

2. Is there any multicollinearity issue and how do you know? Support your answer with appropriate SAS output.
   If there is multicollinearity, remove the predictor with highest VIF from the model. Rerun the new model. Is there still multicollinearity?
   Using the tolerance of VIF to detect the multicollinearity. Bedrooms and Baths variables show high VIF, i.e., VIF>10. There is some multicollinearity issue. (1pt)

Model: MODEL1
Dependent Variable: Price Price

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	B	-18028	15426	-1.17	0.2453	.	0
Size	Size	1	106.83236	7.39653	14.44	<.0001	0.21803	4.58651
Bedrooms	Bedrooms	1	-14555	20812	-0.70	0.4860	0.02429	41.17280
Baths	Baths	1	25404	28238	0.90	0.3705	0.02896	34.52726
Mortgage type Adjustable	Mortgage type Adjustable	B	-11924	9887.37142	-1.21	0.2307	0.96477	1.03651
Mortgage type Fixed	Mortgage type Fixed	0	0	.	.	.	.	.

(1pt)

   After removing the Bedrooms, there seems no multicollinearity as shown in VIF or Tolerance. (1pt)

Model: MODEL1
Dependent Variable: Price Price

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	B	-15107	14812	-1.02	0.3102	.	0
Size	Size	1	104.64528	6.68586	15.65	<.0001	0.26550	3.76654
Baths	Baths	1	6753.80769	9258.46180	0.73	0.4674	0.26805	3.73060
Mortgage type Adjustable	Mortgage type Adjustable	B	-11913	9862.31756	-1.21	0.2299	0.96478	1.03651
Mortgage type Fixed	Mortgage type Fixed	0	0	.	.	.	.	.

(1pt)

3. Build a model with Price as response, size, Baths, Mortgage type and interaction of size*Baths as predictors with NO selection method. Write the linear model in an equation. Which variable(s) is/are significant predictor(s) at significance level of 0.05 and what is/are the confidence interval(s)? Support your answers with a screenshot. (You don't need to output collinearity measurements VIF or Tolerance here)

$\widehat{Price} = \beta_0 + \beta_1 Size + \beta_2 MortagageType_{adjustable} + \beta_3 Baths + \beta_4 Size * Baths$ (1pt: Mortgage type Adjustable level is in the model because the category Fixed is the reference level and takes value of 0. Any similar equation with key elements (with or without estimate values) is correct.)

Size, Baths, and their interaction are the significant predictors at significance level of 0.05. (1pt)

The confidence interval for the coefficient of size variable is (0.05189, 0.08832); CI of Baths is (-59450, -10599); CI of the interaction is (7.44, 18.48) (1pt)

Model: MODEL1
Dependent Variable: Price Price

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	B	122367	32450	3.77	0.0003	57987	186747
Size	Size	1	58.36918	11.65262	5.01	<.0001	35.25071	81.48764
Baths	Baths	1	-35024	12311	-2.84	0.0054	-59450	-10599
Mortgage type Adjustable	Mortgage type Adjustable	B	-14932	9007.85423	-1.66	0.1005	-32803	2939.72989
Mortgage type Fixed	Mortgage type Fixed	0	0	.	.	.	.	.
Size*Baths	Size*Baths	1	12.96337	2.78285	4.66	<.0001	7.44229	18.48446

(1pt)

4. Use stepwise selection method, add/remove effects with significance levels at 0.05 to automatically select a final model. Write the equation with estimated values of parameters of the final model. Support your answer with the appropriate output

$\widehat{Price} = 111463 + 57.664 Size - 31709 Baths + 12.63151 Size * Baths$ (1pt)

Model: MODEL1
Dependent Variable: Price Price

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	111463	32050	3.48	0.0007	47884	175042
Size	Size	1	57.66391	11.74517	4.91	<.0001	34.36466	80.96317
Baths	Baths	1	-31709	12253	-2.59	0.0111	-56015	-7403.29217
Size*Baths	Size*Baths	1	12.63151	2.79955	4.51	<.0001	7.07797	18.18506

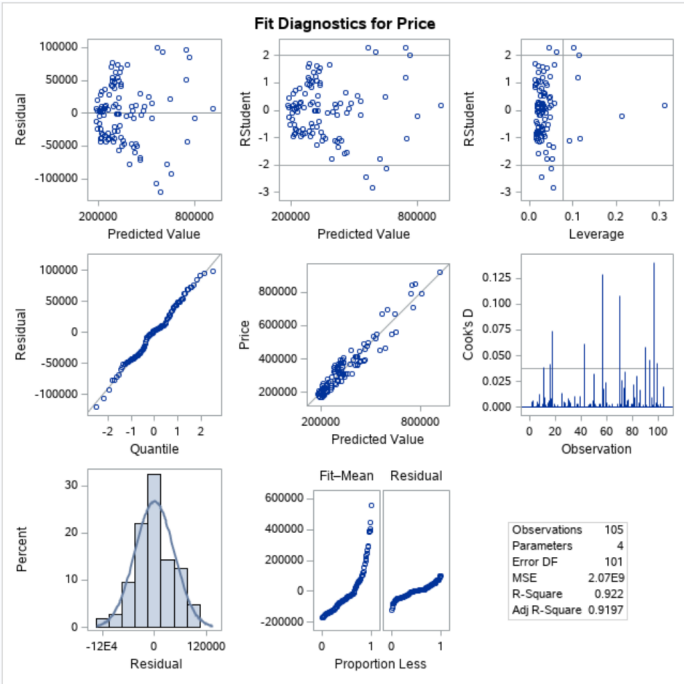
(1pts: full credit as long as the screenshot shows the estimated coefficients and intercept.)

5. For the above final selected model, is any assumption violated according to the diagnostic panel of plots? Attach the diagnostic panel of plots.

The residual or Studentized residual plot show that the linearity is satisfied but the variance seems not very constant. (1pt)

No severe outlier because no Cook's D >1 or too higher than 4/n. (1pt: full credit as long as you stated how to detect outliers even if you think there seem to be some outliers)

From QQ plot, the residuals follow the normal distribution. (1pt)


Fit Diagnostics for Price
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Price
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 105
Parameters 4
Error DF 101
MSE 2.07E9
R-Square 0.922
Adj R-Square 0.9197

(1pt)