

NOTE FROM INSTRUCTOR:

- Hello Students!
- Regarding the Module 4 Project One assignment, again there seems to be a lot of verbiage but here's what it boils down to:
 - 1. A linear regression is only appropriate if the variables are at least roughly linearly correlated. And a great way to determine whether that is the case is to plot a scatterplot of the variables and see if it shows a fairly linear pattern. Also, a correlation coefficient, R , greater than 0.5 typically shows some correlation, while values greater than 0.7 typically show strong correlation. And in determining which variable is the response variable and which is the predictor, we ask ourselves, which variable depends on the values of the other variable? That variable is the dependent, also known as the response or the Y , variable, while the other is then the independent, also known as the predictor or the X , variable. Obviously in this case, listing price should be the response or Y variable, while square footage should be the predictor or X variable
- From the Modules 2 and 3 assignments, we already know how to extract a random sample from the full National Real Estate data population. In this case our sample size should be 50 instead of the 30 we used for the previous assignments, **and** we are now supposed to extract the sample from the entire data population rather than from a region as was previously the case.
- 2. Plot the scatterplot of the Median Square Foot (which should be your X) versus Median Listing Price (which should be your Y), using the guidance provided in the tutorial titled "Scatterplots in Excel". Remember to place the column containing the X data first, and then the Y data to the right, so that Excel recognizes that the first column contains the X data and the second the Y data. As before, you can add the trend line (i.e. the regression line) and display the regression equation as well as the correlation coefficient (which is typically symbolized as R and presented in Excel as the square of R , i.e. R -squared, which simply means that we take the square root of R -squared to get R) as also described in the tutorial.
- 3. As before, obtain the descriptive statistics, i.e. the mean, median and standard deviation for the Median Listing Price and the Median Square Feet for the data sample, using the guidance provided in the tutorial titled "Descriptive statistics in Excel".
- 4. Plot the histograms for the Median Listing Price and the Median Square Feet for your data sample, using the guidance provided in the tutorial titled "Creating Histograms in Excel". If your histogram is roughly bell-shaped and symmetric, it suggests that the data are symmetrically distributed, and probably normally

distributed, with the majority of the values clustered near the middle and few if any unusually large or unusually small values to the right or left of the middle respectively. Lucky you! If your histogram has a markedly longer right tail, it means there are some unusually large values to the right of the middle, probably high outliers, suggesting a right-skewed distribution of the data. Conversely, If your histogram has a markedly longer left tail, it means there are some unusually small values to the left of the middle, probably low outliers, suggesting a left-skewed data distribution. Furthermore, if the data distribution is symmetric (and probably normal), the values of the mean and median should be close. In contrast, if the data distribution is right-skewed, the mean is typically substantially larger than the median. And if the data distribution is left-skewed, the mean is typically substantially smaller than the median. Here are some examples of histograms:

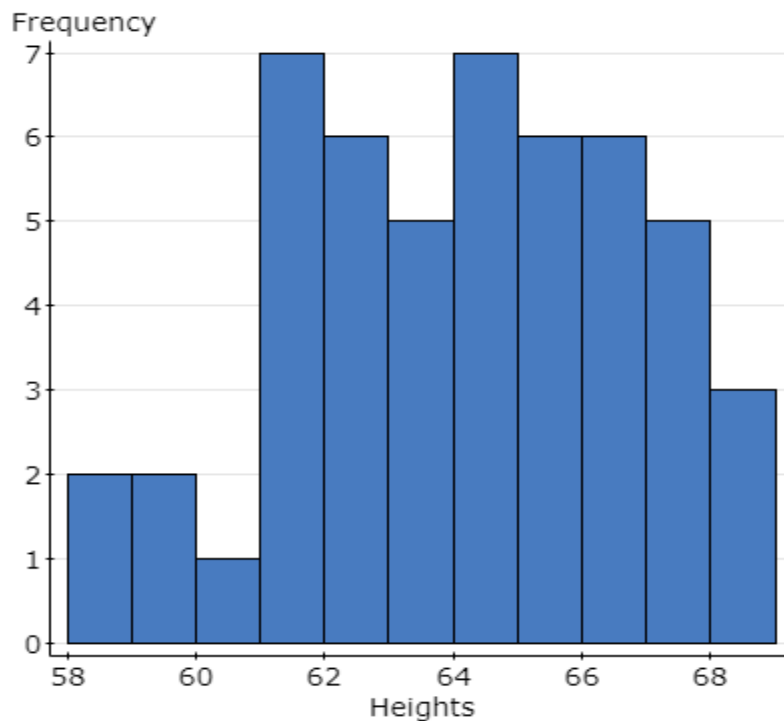


Figure 1: Histogram for Roughly Symmetric Data

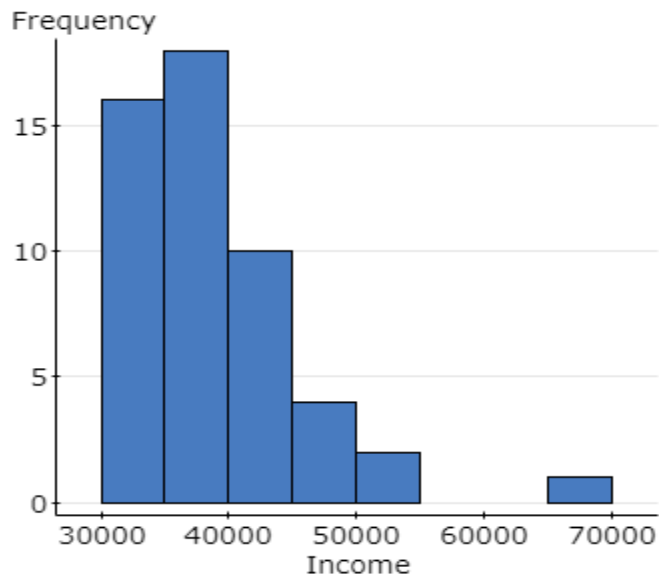


Figure 2: Histogram for Right-Skewed Data

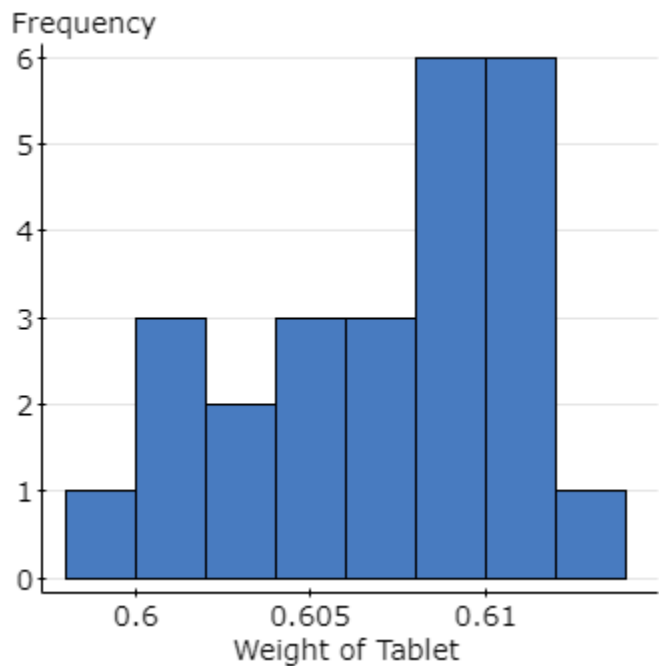


Figure 3: Histogram for Left-Skewed Data

- 5. Visually, a fairly well-defined upward or downward linear pattern in the scatterplot indicates the presence of linear correlation between the two variables. An upward pattern indicates a positive correlation, and a downward pattern indicates a negative correlation. Quantitatively, the magnitude of the correlation coefficient R provides a measure of the strength of the correlation or association between the variables.

-
- 6. As before, compute the regression equation as I described in the previous announcement titled “Guidance for Module 3 Assignment”. The regression equation gives us the slope and intercept coefficients, so that we can write the regression equation in the form:
-
- Median Listing Price = (Slope Coefficient multiplied by Square footage) + Intercept
-
- For example, if the slope is 261 and the intercept is -272175, then the equation is:
-
- $y = 261x + -272175$ which simplifies to: $y = 261x - 272175$
- where y is the median listing price and x is the median square feet.
-
- Note that this is the same equation we should get from the trend line if properly plotted, the only difference being that the trend line does not provide other important diagnostic statistics such as the P value, t-value, the confidence intervals and so on that the regression results output provides.
- Also note that the R-square provided in the regression results should be the same as the R-square provided in the trend line. And the R, which is listed as “Multiple R” in the regression results is simply the square root of the R-square shown in either the trend line or the regression results
-
- 6. As for interpretations, the slope is the mean increase (or decrease) in the value of the Y variable when the value of the X variable increases by one. In other words, the slope is the rate of change of the Y variable relative to the X. The intercept is technically the value of the Y variable when the X variable is zero. But when a zero value for the X does not make sense, then the intercept has no physical meaning or interpretation, but is simply there as a mathematical product of the least squares regression procedure. For example, a value of zero square feet for a house does not make any sense; therefore the intercept has no physical meaning in this case. The R-square, known as the coefficient of determination provides a measure of the strength of the regression equation. In other words, it indicates how well the Y value can be predicted from the X.
- 7. Finally, demonstrate an application of your regression equation by predicting a reasonable listing price for a home with a given square footage

