# Data Exploration

**EXPLORING THE GARAGE AREA**

1. Open the small\_ames dataset. How many Houses are in your dataset?
2. Create a histogram of the Garage Area. Make the BinWidth of 50. What is the range of the most common Garage Area?
3. Calculate the mean, median and mode of the Garage Area.
4. Calculate the max, minimum and range of the Garage Area.
5. Is the Mode the same as the most common Garage Area range? Why or Why not?
6. Create a new column that indicates if the Garage Area is greater than 1 standard deviation from the mean. How many houses are more than 1 standard deviation? What percentage of the houses are more than 1 standard deviation?
7. Repeat the previous with 2 and 3 standard deviations.
8. How do the previous results compare with the Gaussian distribution?
9. Where would you set the threshold for outliers?
10. What would you do with the outliers?

**PREDICTING HOME PRICES**

1. Plot the relationship between the “SalePrice” and the “GrLivArea”?
2. Fit the plot to a straight line. What is the equation? What is the interpretation of the equation. How would you determine if the line is a good prediction of the SalePrice?
3. Calculate the root mean squared error (RMSE) between each data point and the line.

Where yi is the SalePrice of each house, yline is the price predicted by the line, m is the number of houses.

* Data > Data Analysis
* Descriptive Statistics – Select “Label on first Row”, “Summary Statistics,”ConfidenceLevel”- 95%

A:

1. 1460
2. 450-500
3. Average = 472.9, Median = 480, Mode = 0
4. No, because the most common is a garage of 0.
5. Range = 1418, Max = 1418, Min = 0
6. =ABS($J2-AVERAGE($J:$J))>STDEV.P($J:$J), =COUNTIF(M:M,TRUE)
   1. 376/1460 = 25.75%
7. 113/1460 = 7.7%, 7/1460 = 0.47%
8. 68-95-99.7, 32%, 5%, 0.3%.
9. .
10. .
11. .
12. Y = 107.13x+18569 (Note the X needs to be GrLivArea)
13. =(E2-D2)^2, =SQRT(SUM(F:F)/1460), RMSE = $56034