**Group Assignment 1**

The following questions refer to the dataset “Grp.Assgn1.data.csv” posted on eCampus; download the data and save it to your hard drive. The data contains housing information of 2901 houses in Ames, Iowa. Our goal is to develop a model to predict house sales price. The variables in the data set are as follow:

|  |
| --- |
| 1. Sales\_Price 2. Zoning 3. Lot\_Area 4. Alley 5. Lot\_Config 6. Overall\_Qual 7. Foundation 8. Bsmt\_SF 9. Central\_Air 10. First\_Flr\_SF 11. Second\_Flr\_SF 12. Full\_Bath 13. Half\_Bath 14. Bedrooms 15. Kitchen 16. TotalRooms 17. Fireplaces 18. Garage\_Cars |

1. Let’s investigate the data by start with **single variable** at a time.
   1. Develop **Histograms** for 3 variables. “*Lot\_Area*”, “*Sales\_Price*”, “Second\_Flr\_SF”. Submit your visual charts. Describe each variable by stating the values of:
      1. Min, Max
      2. Skewness
      3. Mean, Median, Mode (5 points)
   2. Develop **Histogram Distributions**, using Rattle, for 3 variables: “*Lot\_Area*”, “*Sales\_Price*”, “Second\_Flr\_SF”, grouped by “***Zoning”.*** Submit your visual charts. Describe **1 key observation for each** variable distribution based on the “Zoning” variable. (10 points)

1. Further explore the data for **pairs of variables**.
   1. Create Scatterplot Matrix, using Rcmdr, to explore relationships between pairs of variables. Use the following **5 variables:** *“Bsmt\_SF”, “First\_Flr\_SF”,“Second\_Flr\_SF” “Lot\_Area”, “Sales\_Price”* to explore more deeply into relationships among those variables. Submit your scatterplot matrix and make **3 observations** on relationships among the variables by studying your scatterplot matrix. (10 points)
   2. Develop a **Distribution Pairs+ Histogram**, using Rattle, for 2 variables: *“First\_Flr\_SF”, “Lot\_Area”.* Submit your distribution visual chart. State your observation about “First\_Flr\_SF” for **very large** lots (“Lot\_Area”)? (10 points)

1. Now, let’s investigate the impact of all variables: To that end, run a regression model with “Sales\_Price” as the response variable and all other variables as predictors; Submit regression output R screenshot and answer the following questions:
   1. Are all variables in the model important predictors for Sales\_Price? State your reasons for your answers (be concise and clear). ( 10 points)
   2. Is this a good model? Why or why not? (5 points)
   3. Make one suggestion to improve the model. (10 points) \*Keep your answers **short, clear and precise**.

1. Now, transform some variables using **log transformation**.

*(\* “some variables” means any of the variables in the data set, use your logic and intuition for which variables need log-transformation)*

* 1. Use Scatterplot Matrix to explore relationships among the log-transformed variables. Explore more deeply into relationships between “*Sales\_Price*”, “First\_Flr\_SF”. Submit your scatterplot matrix and make **3 observations** about your scatterplot matrix. (10 points)

1. Run a few regression models with log-transformed variables. **Pick your best model**, submit regression output R screenshot, and answer the following question:  
   1. Is this a better model as compared to the previous model for question 3? State your reason. (10 points)
2. Use your **BEST regression model** obtained in question 5 to perform “Sales\_Price” prediction for the following house.

A one-story CulDSac home on a 20,000 SqFt Lot Area, with 4,500 SqFt on the first floor, slab foundation, central air, 3-car garage, 4 bedrooms, 2 full- bathrooms, 1 half-bathroom, and a fireplace. The house is in an excellent overall condition with a paved alley, and is located in medium density zoning.

1. What’s your point forecast for Sales\_Price (in dollars)? (10 points)
2. What are your 95% confidence interval upper and lower bounds for this house (in dollars)? (10 points)