This problem will have you practice using the "two-lines" multiple regression model from the Math 325 Statistics Notebook.

Run the following commands in R.

```
library(mosaic)
?SaratogaHouses
```

As explained in the help file (?SaratogaHouses), the SaratogaHouses data set contains data about many houses from Saratoga County, New York in the year 2006. When it comes to buying and selling a home, one of the most important factors is determining the value (or price) of the home. Suppose a family is in search of a home that was newly constructed and that has three bedrooms. Suppose further that they are trying to decide how big of a **livingArea** they can afford and whether or not the **price** of the home is significantly impacted by adding a **fireplace** to their "dream home" wish list.

To get a group of homes that are similar to their current specifications run the following codes in R.

```
SH2 <- filter(SaratogaHouses, bedrooms == 3,  newConstruction=="Yes")
View(SH2)
```

Use the **SH2** data set and a "two-lines" multiple regression model to describe the **price** of a house according to the size of the livingArea of the house and whether or not the house has a fireplace (**fireplaces** is only 0 or 1 for this reduced SH2 data).

The two-lines regression model for this situation would be most appropriately labeled as:

$$\underbrace{Y_i}_{\text{Label A}} = \beta_0 + \beta_1 \underbrace{X_{1i}}_{\text{Label B}} + \beta_2 \underbrace{X_{2i}}_{\text{Label C}} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

where, supposing $X_{2i} = 0 \text{ or } 1$, then

[1 ▾] Label A [3 ▾] Label C [2 ▾] Label B

1. price
2. livingArea
3. fireplaces

Further, in this model the parameters would each be interpreted as

1 β_1 4 β_3 3 β_0
2 β_2

1. The change in the average price of a home without a fireplace as the living area increases by 1 additional square foot.
2. The difference in the average price of a home with a fireplace as compared to a home without a fireplace for homes with zero square feet of living area.
3. The average price of a home with no fireplace and a living area of zero square feet. Since this is unrealistic, this parameter doesn't actually carry any meaning for this particular regression model.
4. The change in the effect of 1 additional square foot in the living area on the average price of homes with a fireplace as compared to homes without a fireplace.

Perform the above regression in R. To demonstrate that you can do this, fill in the blanks in the following R code statement.

> sh2.lm <- lm (price ~ livingArea + fireplaces + livingArea:fireplaces , data=SH2)
> summary(sh2.lm)

There are four places in the R output of the above regression results that contain p-values, one p-value for each coefficient. Note that each of these p-values have a "t value" next to them implying that they came from a t test, which is cool. If you have done your work correctly, the p-value for the test of the hypothesis that livingArea effects the average price of a home is 0.0055.

What is the p-value for the test of the hypotheses that

$$H_0 : \beta_3 = 0$$
$$H_a : \beta_3 \neq 0$$

P-value = 0.06535

The p-value for the test that $\beta_3 = 0$ is not significant at the 0.05 level. However, neither is the p-value for the test of $\beta_2 = 0$. This suggests an interesting idea that either one or both of these variables is not significant. However, mutliple linear regression is a complicated world. It is best practice to "remove" only the "least significant" term from the model and then re-check all p-values to see what is now significant. This is because everything can change dramatically when just one variable is added or removed from the regression. Watch what happens to the summary output when you remove the **fireplaces** term from the lm(...) but keep the other terms, including the **fireplaces:livingArea** interaction term, in the model.

New p-value for the interaction term: 0.04819

This is now significant at the $\alpha = 0.05$ level.

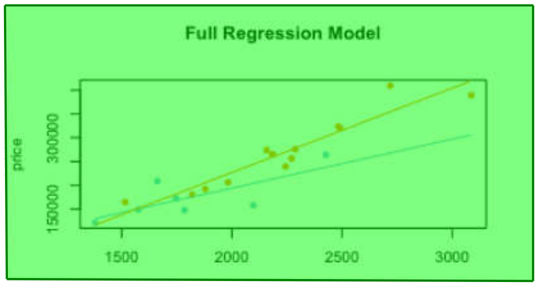
It is important to visualize a regression whenever possible so that the reader can connect with the "truth" about the situation that we are trying to show them. Reproduce the two graphics that are shown below.

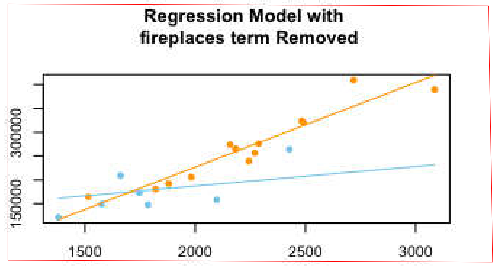
Only check these boxes if you really made the graphs:

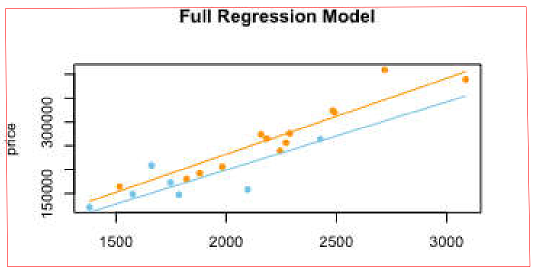
- [x] I created a scatterplot with two regression lines for the "full regression model" above.
- [x] I also created a scatterplot with two regression lines for the "reduced regression model" above.

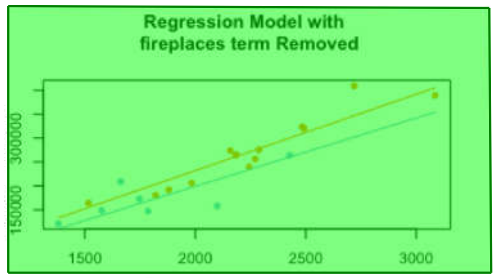
To demonstrate that you made the requested graphs, select the appropriate graph for each situation by clicking on the image that matches your graphics.

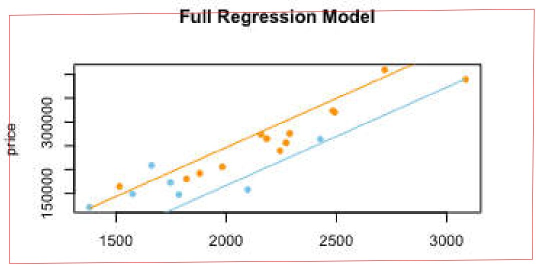
Select only 1 Graph on the "left side" and only 1 Graph on the "right side."

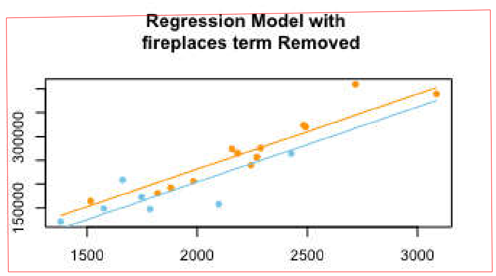


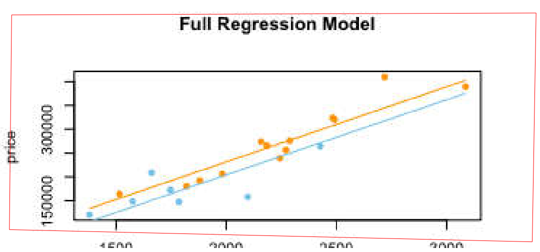


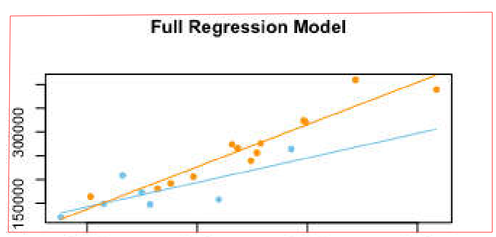












Now that we have performed the regression, found a "significant model" and drawn the regression, we are ready to interpret the results.

Suppose the family we were discussing earlier found a house that had a livingArea of 2500 square feet and a fireplace. What is the predicted cost (based on your reduced regression model) for this house?

Predicted price = $ 311915 (Round to the nearest whole dollar.)

Further, this "reduced model" claims that the average price of a homes without fireplaces increases by $ 143.1 per each additional square foot while the average price for homes with fireplaces increases by $ 159.58 per each additional square foot. **(Round both answers to the nearest cent.)**

Before we ever fully trust the results and interpretation of a regression model, it is important to diagnose the appropriateness of the model. To do this, run the commands in order to create your three diagnostic plots of regression:

> library(car)

> par(mfrow=c(1,3))

> plot(the name of your reduced lm, which=1)

> qqPlot(the name of your reduced lm$residuals, id=FALSE)

> plot(the name of your reduced lm$residuals)

Demonstrate that you have made these plots by selecting the three rows in the data set that are identified as possible "outliers"

- [ ] 1
- [x] 2
- [ ] 3
- [x] 4
- [ ] 15
- [ ] 16
- [ ] 17
- [x] 18
- [ ] 24
- [ ] 25
- [ ] 26

None of these diagnostic plots look "great" but they are "good enough" to use the regression results for interpretation and prediction.