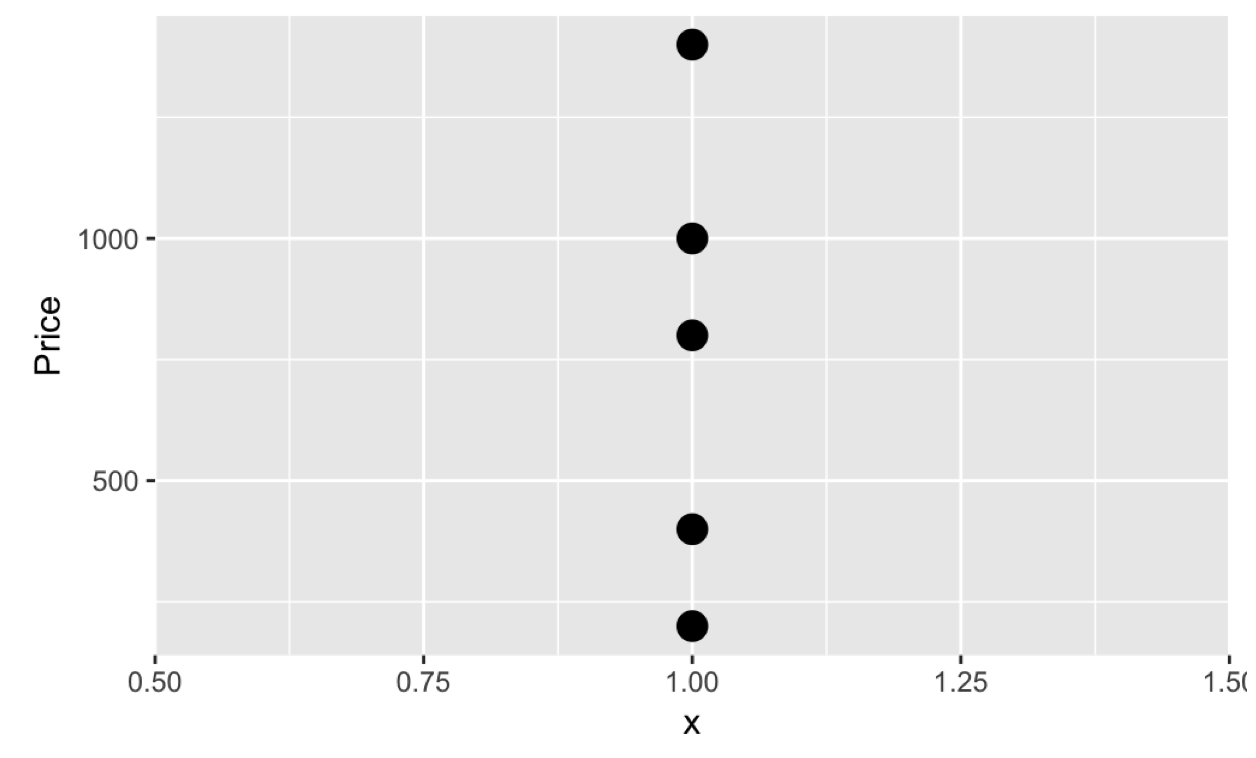
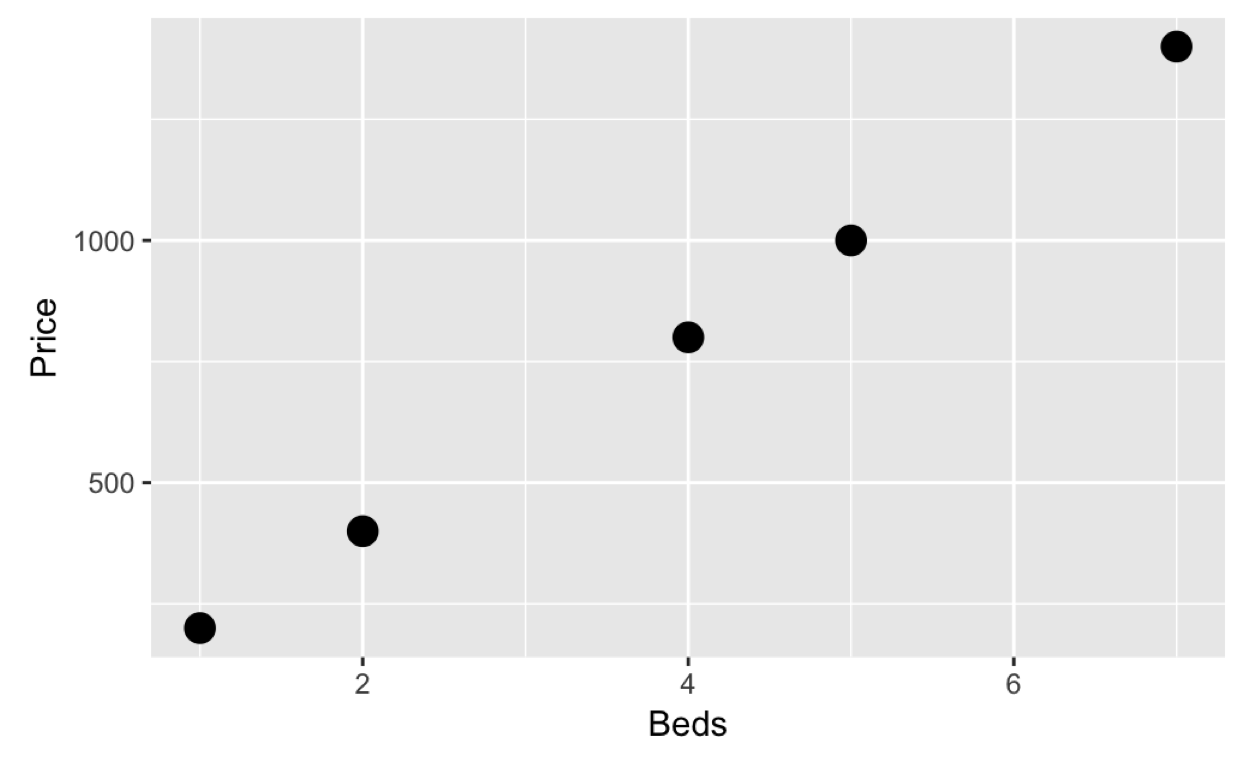
**Name: Classwork 9.1**

1. Let’s look at some houses. How much do houses in this area cost in general? Is there variation in how much houses are worth?
2. What do you think makes a house cost more versus less? Name a few features.
3. Before we get to look at real data on home costs, let’s take a look at a tiny data set of **FakeHomes**. Where do you see variation in price? (draw)



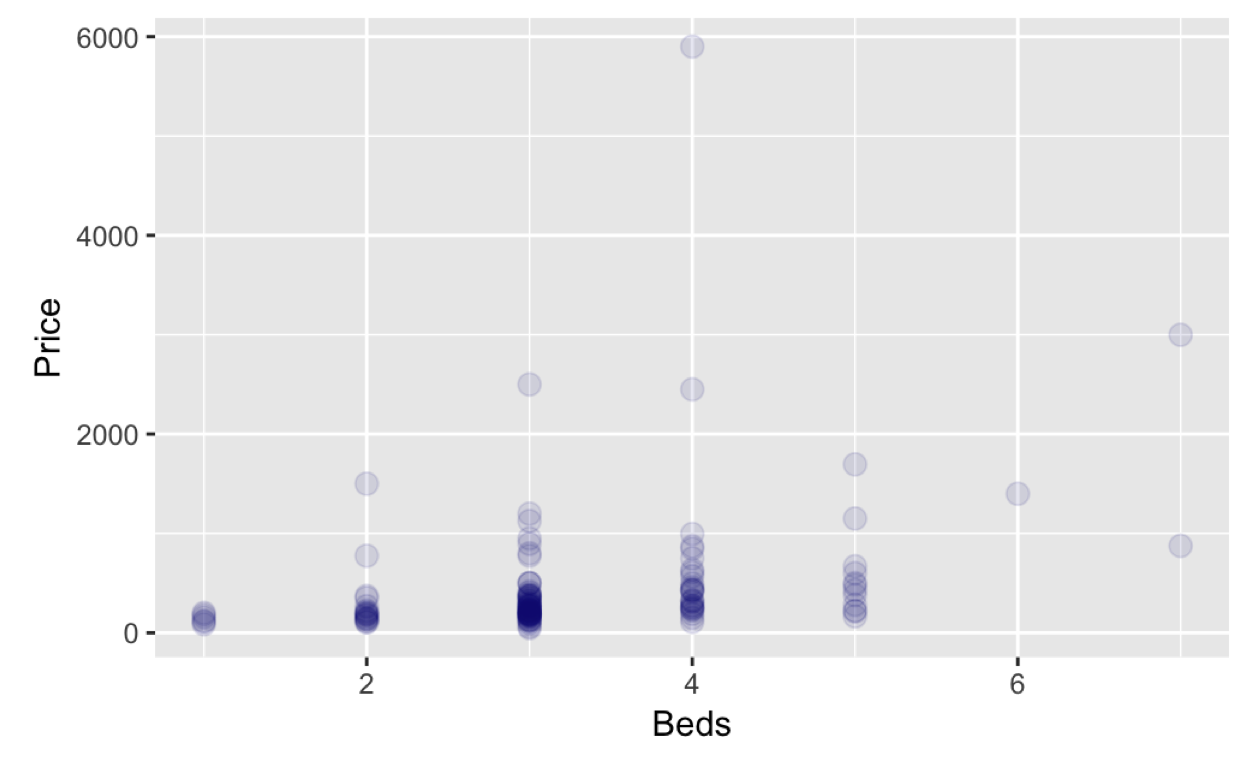
1. Let’s create a story of a DGP that might be responsible for the variation we see in price. How might number of bedrooms explain some of the variation we see in price? Write this as a story and then as a word equation.
2. What do we mean by “explain variation” again?
3. Using our tiny little bit of fake data (**FakeHomes**), I’ve created a visualization that could help you see if bedrooms might be a good explanatory variable for price. Would knowing Beds help you make a better prediction about Price?



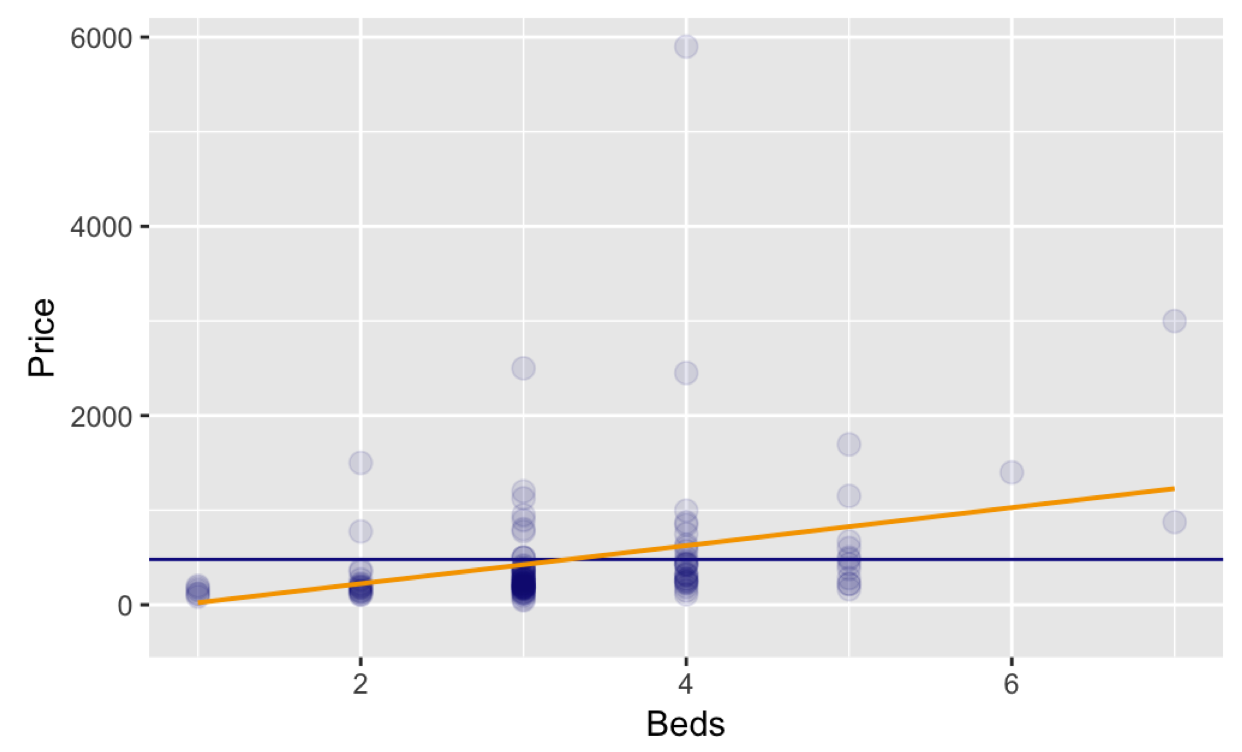
1. Sketch in the empty model and sketch the bedroom model. Use different colors.
2. What code would you use to find the best fitting estimates for the General Linear Model (GLM) that uses **Beds** to predict **Price**? After Dr. Ji runs the code, write down the GLM equation and label it so that you could understand it.
3. What would the Beds model predict is the price of a home with 8 bedrooms?
4. Just by looking, is there any leftover error after using the Beds model? What do you think the SS Error would be? What would be the PRE?
5. What do we mean when we say, “This model explains 100% of the variation.”?
6. Now we had to do all that with a fake data set because real variation isn’t usually so easy to explain. Now let’s look at real data on homes: **HomesForSale** data frame.Here is what a scatterplot of just prices will look like. Sketch what the histogram might look like.

|  |  |
| --- | --- |
|  |  |

1. Add the empty model into the scatterplot above. Does it seem like it balances the positive and negative deviations? Why or why not? If it helps you might want to sketch a few of the deviations on scatterplot above.
2. Here is a scatterplot with Beds predicting Price. See that very expensive house with 4 bedrooms? Find that data point on both this scatterplot and the one in 11.



1. Draw the empty model in 14 as well. Is it the same empty model from 12 and 13? Draw a few residuals from the empty model and a few of the squares that make up SS Total on this visualization as well.
2. Write the simple model in GLM notation. Label each part so that you can understand it.
3. Does the simple model “explain” any variation in Price?
4. Write the R code to add the best fitting regression line (using Beds to predict Price) to the visualization. (You want to be able to create visualization below.) Then draw in three different types of deviations using three different colors:
   1. Residuals from the simple model (variation *\_\_\_\_\_\_\_\_\_\_\_\_\_\_*)
   2. Residuals from the complex model (variation \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)
   3. Variation \_\_\_\_\_\_\_\_\_\_\_\_\_ by the complex model
   4. If we square and add up these deviations, which one would we give us SS Total, SS Model or SS Error?



1. Write down the GLM equations for the empty and complex models. Label them so that you could understand them.
2. Some house prices are better predicted by the empty model than the complex model. But if we run supernova, the complex model has a smaller leftover error than the empty model. Why?
3. If we coded **Beds** to be a factor, how would our variable be different?
4. How would the resulting model be similar? How would it be different?

|  |  |  |
| --- | --- | --- |
|  | **Beds as a number** | **Beds as a factor (Beds.factor)** |
| GLM for model |  |  |
| Best fitting estimates |  |  |
| Degrees of Freedom (and why) |  |  |
| What is the PRE and what does it mean? |  |  |
| What is the F and what does it mean? |  |  |
| Similarities:  Why do they have same SS Total?  Why are they both considered “complex models”? |  | |

1. Ultimately, which model do you think explains more of the variation seen in Price? Why?