# Name:

# Lab 3: **HomesForSale**

This data frame is already available in RStudio. The data frame is called **HomesForSale**.

This contains a sample of homes for sale in each state, selected from zillow.com. There are 120 observations on the following 5 variables.

|  |  |
| --- | --- |
| State | Location of the home: CA NJ NY PA |
| Price | Asking price (in $1,000's) |
| Size | Area of all rooms (in 1,000's sq. ft.) |
| Beds | Number of bedrooms |
| Baths | Number of bathrooms |

1. Let’s look at the prices of houses. Make a visualization to explore the variation in house cost. How much do these houses cost? (Describe the distribution: shape, center, spread.)
2. In your opinion (before you look at the data), which of these features is most useful for predicting house **Price**: State, Size, Beds, or Baths?

Why do you think that feature is most predictive of price?

Write a word equation for your idea. We will call this your “best model.”

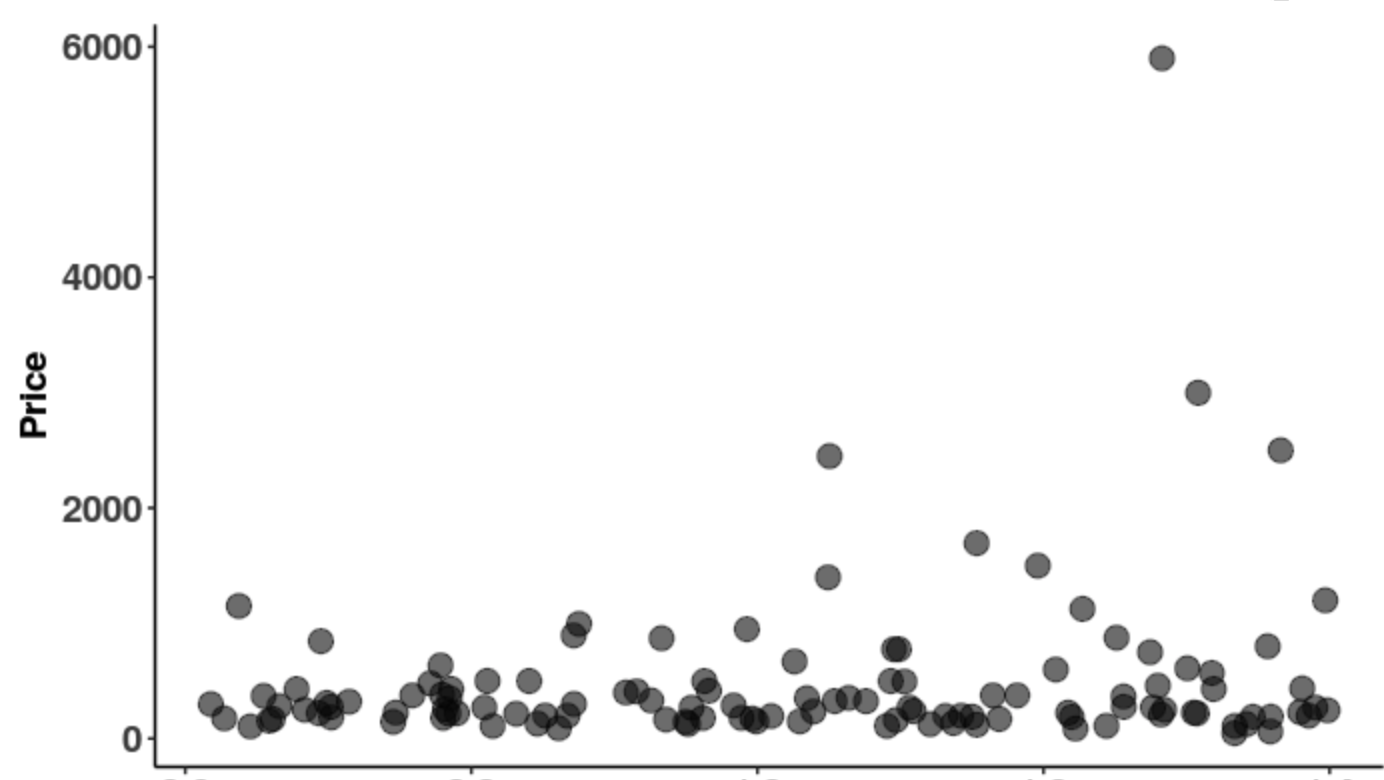
1. In your opinion (before you look at the data), which of these features is most least for predicting house **Price**: State, Size, Beds, or Baths?

Why do you think that feature is least predictive of price?

Write a word equation for your idea. We will call this your “worst model.”

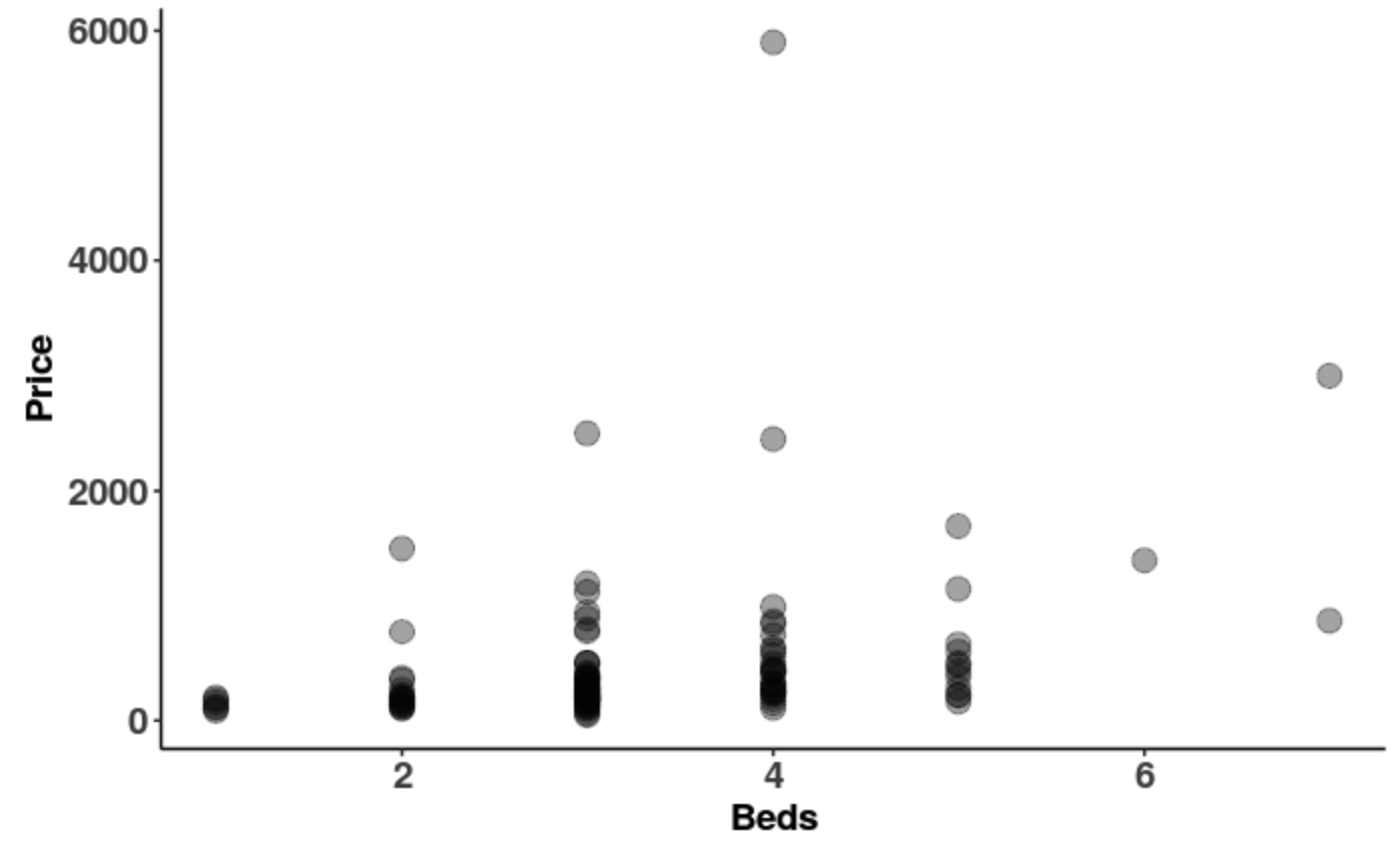
1. What do we mean by “explain variation” again?
2. Do you think your best model will explain more variation than the empty model? Why or why not?
3. Do you think your worst model will explain more variation than the empty model? Why or why not?
4. Here is a jitter plot made with this R code (below). Draw the empty model and a few residuals here.

gf\_jitter(Price ~ 1, data = HomesForSale, size = 4, height = 0, alpha = .5)

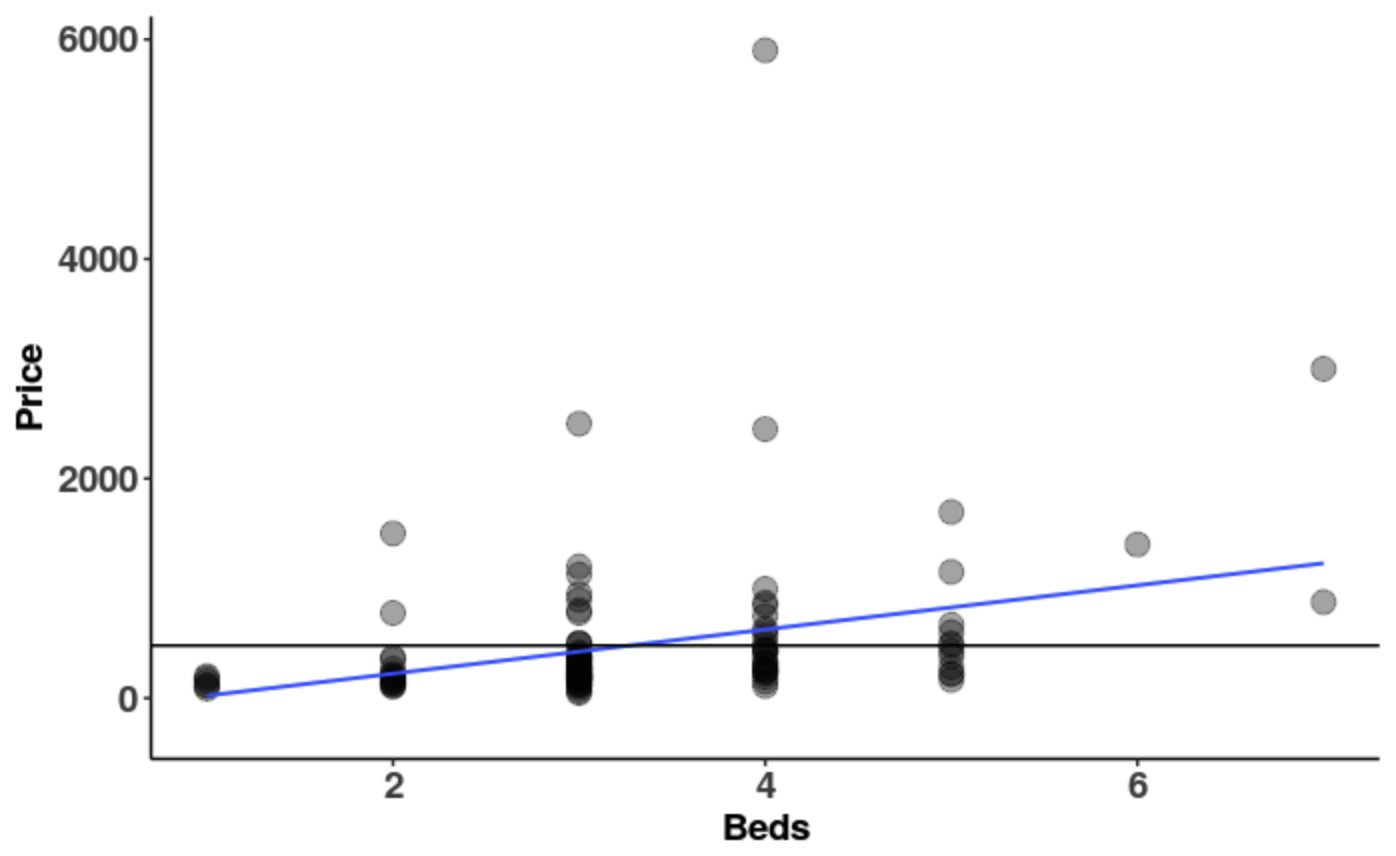


1. Find the best fitting empty model. Write the GLM for the empty model. What does the number mean?

1. Now find the best fitting estimates for your “best model”. For example, if you thought **State** explains variation in **Price** the best, then you would create a **State model**. Write the GLM for your best model. What do the numbers mean?
2. Now find the best fitting estimates for your “worst model”. For example, if you thought **State** explains variation in **Price** the least, then you would create a **State model**. Write the GLM for your least model. What do the numbers mean?
3. Which model (your best versus worst model) do you think explains more of the variation seen in Price? Why?
4. What code would you use to find the best fitting estimates for the General Linear Model (GLM) that uses **Beds** to predict **Price**?
5. What would the **Beds** model predict as the price of a home with 8 bedrooms?
6. What would the empty model predict as the price of a home with 8 bedrooms?
7. Here is a scatterplot with **Beds** predicting **Price**. See that very expensive house with 4 bedrooms? Which **State** is that house in? How much does it cost? Does it cost only $6000?



1. Here we have depicted the best fitting regression line (using **Beds** to predict **Price**) to the scatterplot as well as the empty model. Label each below. Then draw in three different types of deviations using three different colors:
   1. Residuals from the simple model: is this variation explained or unexplained?
   2. Residuals from the complex model: is this variation explained or unexplained?
   3. Variation explained by the complex model



1. If we square and add up these deviations, which one would we give us SS Total? SS Model? SS Error?
2. Some house prices are better predicted by the empty model than the **Beds** 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. Try using **Beds** as a number and **Beds** as a factor. Fill out the following table to show how the resulting models would be different and similar.

|  |  |  |
| --- | --- | --- |
|  | **Beds as a number** | **Beds as a factor (Beds.factor)** |
| GLM notation for model |  |  |
| Best fitting estimates (the b’s) |  |  |
| 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. Which way should we create a **Beds** model: using **Beds** as a number or as a factor? Why?