

AE 03: Duke Forest + data visualization

Jessica St Jean

```
library(tidyverse)
library(openintro)
```

Exercise 1

Suppose you're helping some family friends who are looking to buy a house in Duke Forest. As they browse Zillow listings, they realize some houses have garages and others don't, and they wonder: **Does having a garage make a difference?**

Luckily, you can help them answer this question with data visualization!

- Make histograms of the prices of houses in Duke Forest based on whether they have a garage.
 - In order to do this, you will first need to create a new variable called `garage` (with levels "Garage" and "No garage").
 - Below is the code for creating this new variable. Here, we `mutate()` the `duke_forest` data frame to add a new variable called `garage` which takes the value "Garage" if the text string "Garage" is detected in the `parking` variable and takes the test string "No garage" if not.

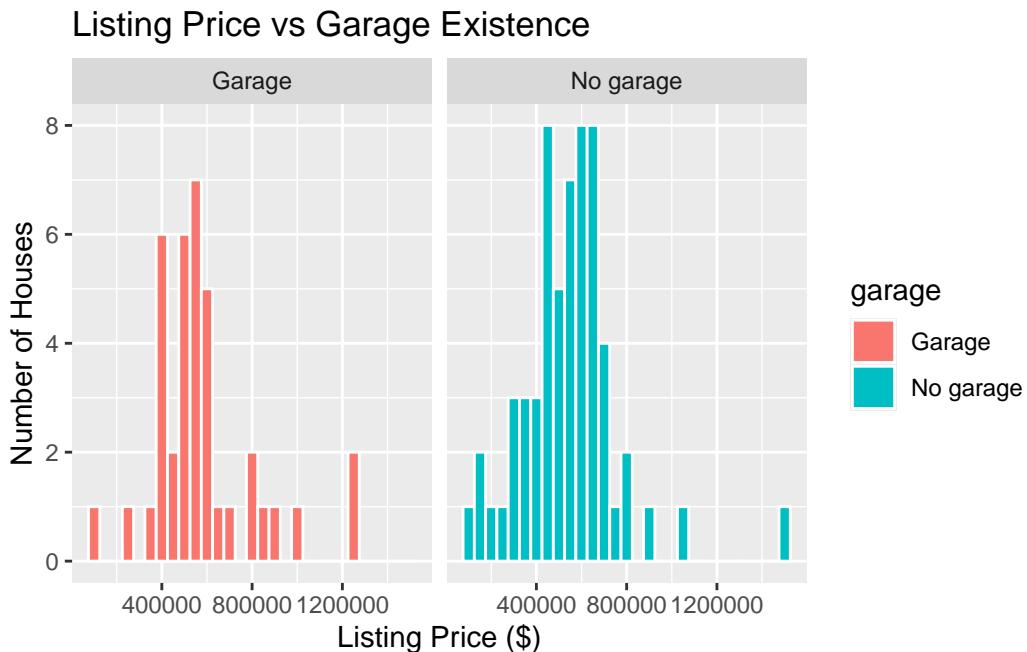
```
duke_forest <- duke_forest |>
  mutate(garage = if_else(str_detect(parking, "Garage"), "Garage", "No garage"))
```

- Then, facet by `garage` and use different colors for the two facets.
- Choose an appropriate binwidth and decide whether a legend is needed, and turn it off if not.
- Include informative title and axis labels.
- Finally, include a brief (2-3 sentence) narrative comparing the distributions of prices of Duke Forest houses that do and don't have garages. Your narrative should touch on whether having a garage "makes a difference" in terms of the price of the house.

```

ggplot(duke_forest, aes(x=price, fill= garage)) +
  geom_histogram(binwidth=50000, colour = "white")+
  facet_wrap(~ garage) +
  labs(
    x= "Listing Price ($)",
    y= "Number of Houses",
    title= "Listing Price vs Garage Existence"
  )

```



Comment: It appears that the houses without garages frequently have a higher price than those which do have a garage. While there are houses with garages that are more expensive than those without garages, the majority of the no garage houses are more expensive than the majority of the houses with a garage. This makes it appear that having a garage or not does make a difference in the price somewhat, but maybe not in the expected way.

! Important

Now is a good time to render, commit, and push. Make sure that you commit and push all changed documents and your Git pane is completely empty before proceeding.

Exercise 2

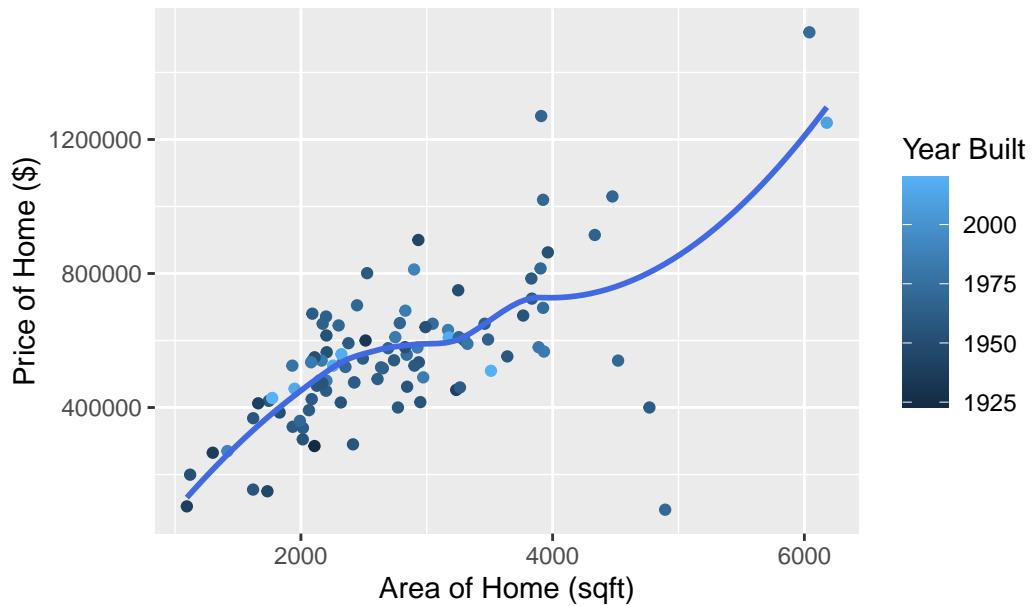
It's expected that within any given market larger houses will be priced higher. It's also expected that the age of the house will have an effect on the price. However in some markets new houses might be more expensive while in others new construction might mean "no character" and hence be less expensive. So your family friends ask: "In Duke Forest, do houses that are bigger and more expensive tend to be newer ones than those that are smaller and cheaper?"

Once again, data visualization skills to the rescue!

- Create a scatter plot to exploring the relationship between `price` and `area`, conditioning for `year_built`.
- Use `geom_smooth()` with the argument `se = FALSE` to add a smooth curve fit to the data and color the points by `year_built`.
- Include informative title, axis, and legend labels.
- Discuss each of the following claims (1-2 sentences per claim). Your discussion should touch on specific things you observe in your plot as evidence for or against the claims.
 - Claim 1: Larger houses are priced higher.
 - Claim 2: Newer houses are priced higher.
 - Claim 3: Bigger and more expensive houses tend to be newer ones than smaller and cheaper ones.

```
ggplot(duke_forest, aes(x=area, y=price, colour=year_built))+  
  geom_point() +  
  geom_smooth(se= FALSE, colour="royalblue") +  
  labs(  
    x="Area of Home (sqft)",  
    y= "Price of Home ($)",  
    title ="Home Size vs Listing Price",  
    colour= "Year Built"  
)  
  
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

Home Size vs Listing Price



Comment:

Claim 1: Larger houses are priced higher.

Response: This appears to be true as the trend line goes up when (higher price) moving towards the right side of the graph (bigger homes) this indicates a positive relationship between house size and price.

Claim 2: Newer houses are priced higher.

Response: This graph does not appear to support this claim as the newer houses are mostly concentrated in the mid range of the graph. With the slightly older houses being priced as more expensive for the most part.

Claim 3: Bigger and more expensive houses tend to be newer ones than smaller and cheaper ones.

Response: This does not appear to be the case when looking at the graph, the majority of the newer houses are in the mid range of the graph as stated previously with moderate size and price, the higher and lower ends of the size and price appear to mostly be older homes.

! Important

Now is a good time to render, commit, and push. Make sure that you commit and push all changed documents and your Git pane is completely empty before proceeding.