AE 03: Duke Forest + data visualization

David Bell

```
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!

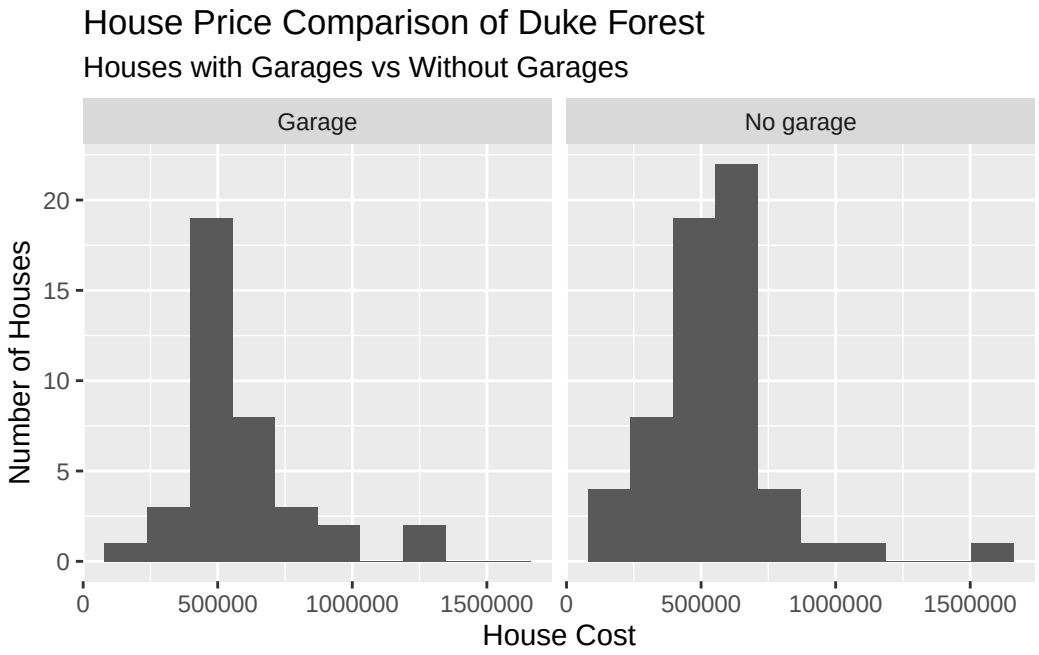
- Make histograms of the prices of houses in Duke Forest based on whether they have a garage.
  - In order to do this, you will first need to create a new variable called `garage` (with levels `"Garage"` and `"No garage"`).
  - Below is the code for creating this new variable. Here, we `mutate()` the `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.

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

ggplot(duke_forest, aes(x = price)) + geom_histogram(show.legend = FALSE,
  bins = 10) +  labs(title = "House Price Comparison of Duke Forest",
 subtitle = "Houses with Garages vs Without Garages",
 x = "House Cost", y = "Number of Houses") + facet_wrap(~garage)
```



From observation of the two histograms, it can be determined that there is a higher amount of houses in duke forest without a garage. Alltough more houses with no garages, it appears that the house have a similar median around the $500K mark.

> **!** 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

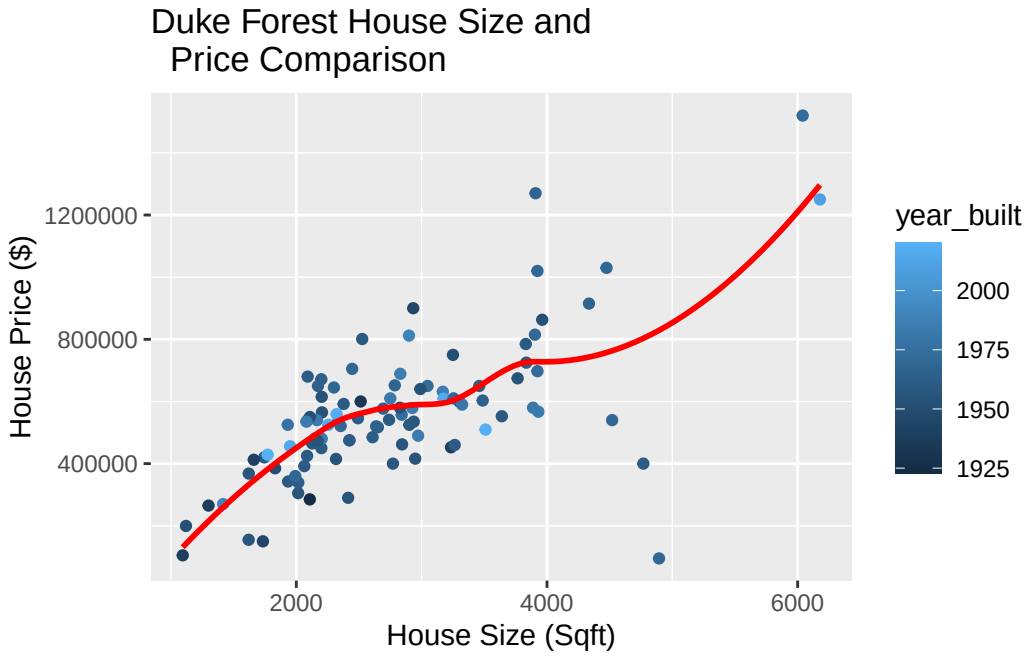
It's expected that within any given market larger houses will be priced higher. It's also expected that the age of the house will have an effect on the price. However in some markets new houses might be more expensive while in others new construction might mean "no character" and hence be less expensive. 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)) + geom_point(aes(color = year_built)) +
  geom_smooth(se = FALSE, color = "red") + labs(title = "Duke Forest House Size and
  Price Comparison", x = "House Size (Sqft)", y = "House Price ($)")
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



1) As shown in the scatter plot, it appears that the majority of larger houses are priced higher then small houses. There are some outliers, but the majority of houses greater then 4,000 square feet are larger in size.

2) All newer houses (>~2020) are not priced higher then older homes. Other then the lone outlier, newer homes are priced in similar regions of older home of between 400k and 800k dollars

3) Claim 3 is false, it appears the larger and more expensive homes appear to be older houses.

> **!** 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.