AE 03: Duke Forest + data visualization

Hannah Andronyk

format: pdf

```
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.

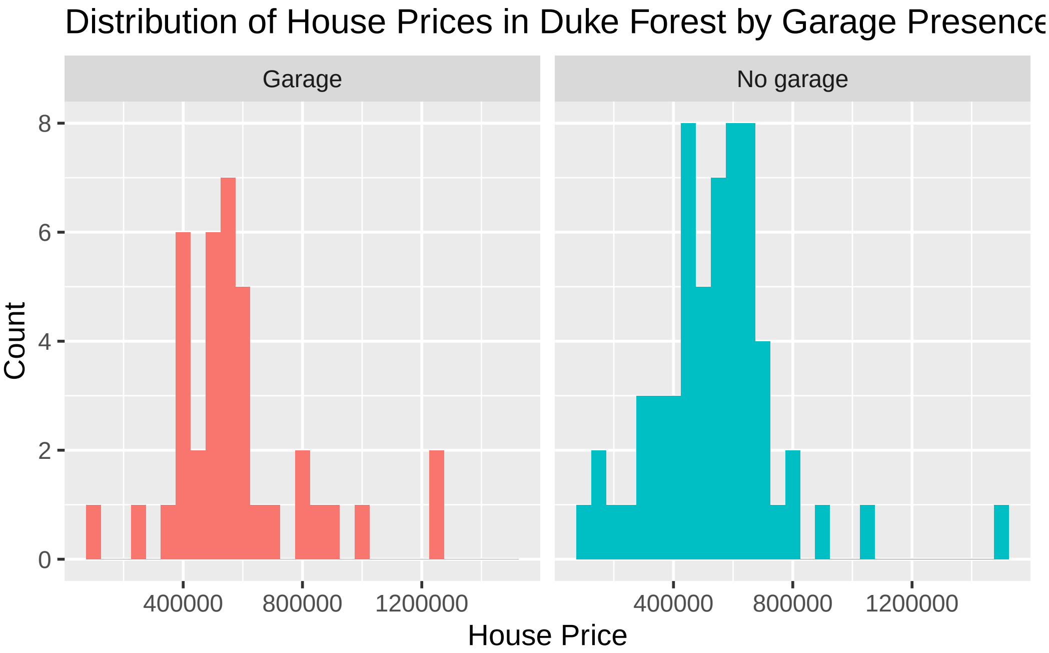
```
duke_forest |>
  mutate(garage = if_else(str_detect(parking, "Garage"), "Garage", "No garage"))
```

- 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.

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

# Histograms of house prices, facet garage
ggplot(duke_forest, aes(x = price, fill = garage)) +
  geom_histogram(binwidth = 50000) +
  labs(
    title = "Distribution of House Prices in Duke Forest by Garage Presence",
    x = "House Price",
    y = "Count"
    ) +
  facet_wrap(~ garage) +
  theme(legend.position = "none")
```



Comment: The distribution of house prices with a garage and without a garage is quite similar, both being slightly right-skewed. It would be hard to make definitive conclusions from these histograms about whether having a garage makes a difference in price because the medians look to be quite similar but there are some bins that lie farther from the cluster that could have an impact on this value.

❗ 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

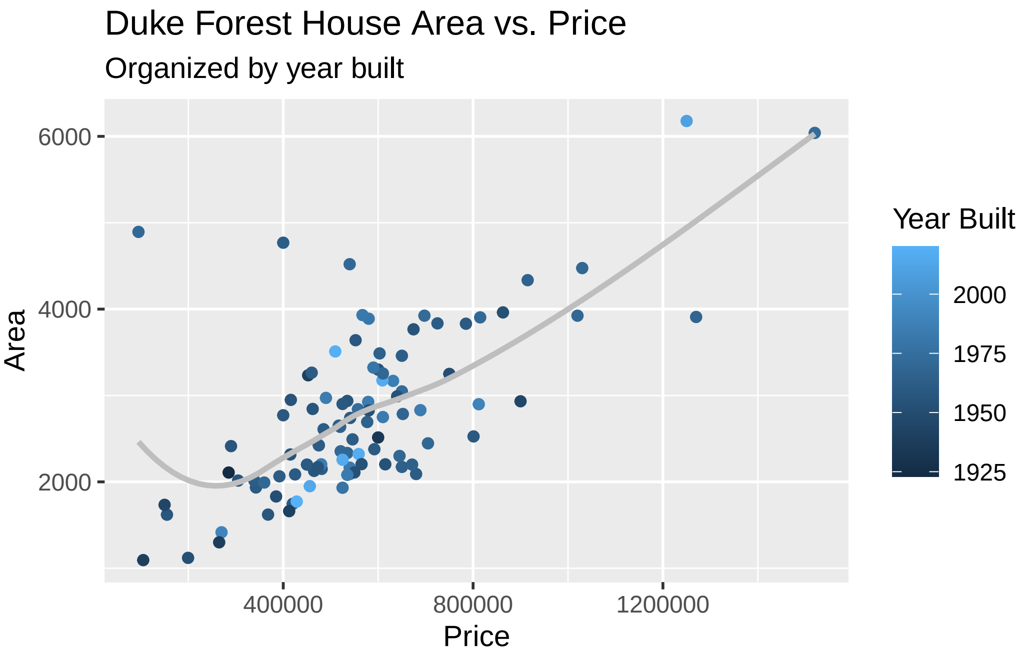
It's expected that within any given market larger houses will be priced higher. It's also expected that the age of the house will have an effect on the price. However in some markets new houses might be more expensive while in others new construction might mean "no character" and hence be less expensive. 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.

```r
# Scatterplot for area vs. price
ggplot(duke_forest, aes(x = price, y = area, color = year_built)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "grey", alpha = 0.5) +
  labs(
    title = "Duke Forest House Area vs. Price",
    subtitle = "Organized by year built",
    x = "Price",
    y = "Area",
    color = "Year Built"
  )
```

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



Claim 1: Larger houses are priced higher. This claim is supported by the general trend in the data, where the line of best fit indicates a positive relationship. However, there are a couple of outliers from this trend near the upper-left corner of the plot where these homes are relatively large and priced relatively cheap that don't support this claim.

Claim 2: Newer houses are priced higher. This claim is somewhat supported by the plot. The points located in the cluster are built in a variety of different years. However, it appears that all of the oldest homes (darkest points) are located in the bottom-left corner which supports the claim that newer houses are priced higher and thus older houses are priced lower.

Claim 3: Bigger and more expensive houses tend to be newer ones than smaller and cheaper ones. This plot does not support this claim well, as there is very few data points that lie in the top-right corner of the graph to draw conclusions from. It appears that out of the few that are there they are not consistent in colour and therefore not consistent in their year built either.

❗ 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.