# AE 03: Duke Forest + data visualization

## Ayden Frost

```
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, color = "white") +
  facet_wrap(~ garage) +
  labs(
    x = "House Price ($)",
    y = "Number of Houses",
    title = "House Prices vs Garage Presence"
  ) +
  theme(
    legend.position = "none"
  )
```

## House Prices vs Garage Presence
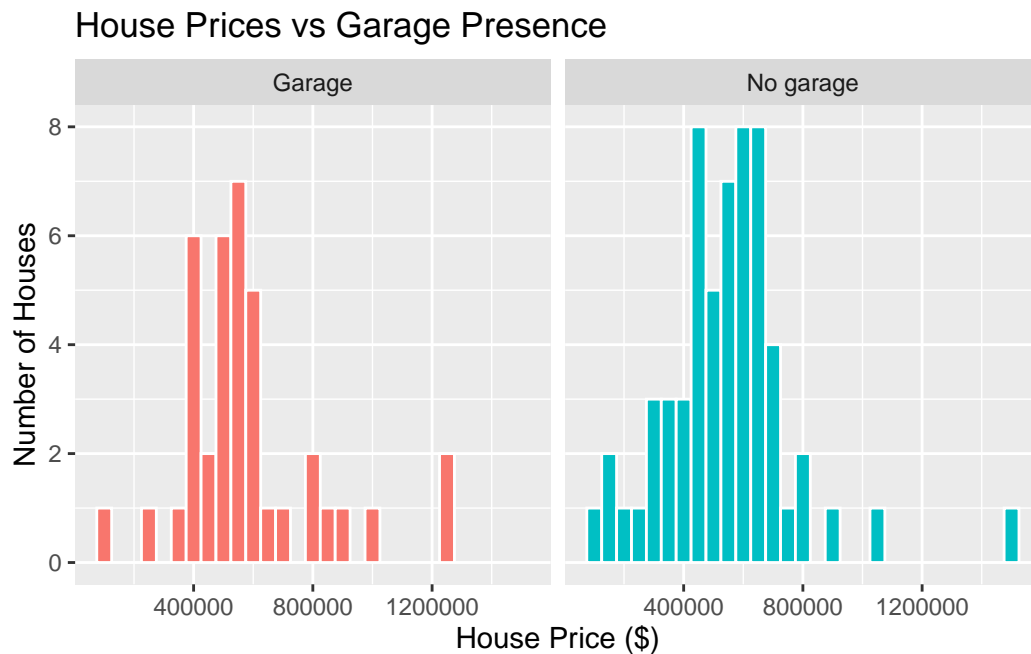


The majority/mode of houses without a garage seem to be higher in value when compared to the majority/mode of houses with a garage. It seems as though having a house with a garage somehow lowers the overall price of the home than if the home were to not have a garage. There also seems to be far more homes without garages relative to houses with them.

> **!** 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 marker 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, color = year_built)) +
geom_point() +
geom_smooth(se = FALSE, color = "black") +
  labs(
    y = "Price of Home ($)",
    x = "Size of Home (Square Feet)",
    title = "Home Size vs Price",
    color = "Year Built"
  )
```

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

## Home Size vs Price
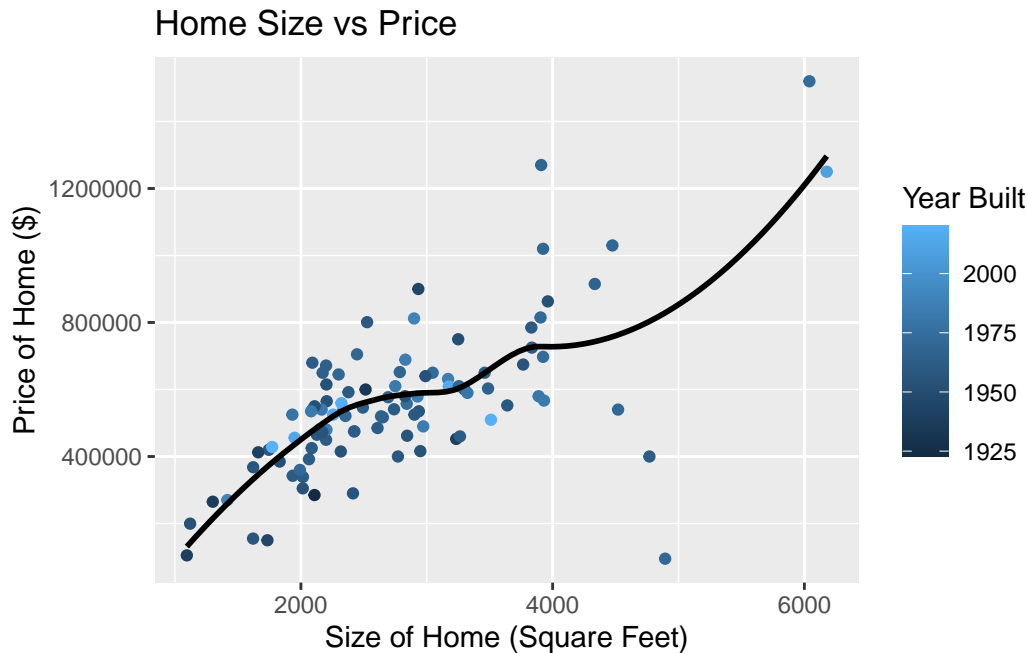


- Claim 1: Larger houses are priced higher.

  Answer: The graph shows that as home sizes increase, homes tend to get more expensive. The trend line shows an upward slope implying that there is a positive relationship between the two variables.

- Claim 2: Newer houses are priced higher.

  Answer: From the graph we can see very few lightly colored points sitting too high or too low in terms of price. The new homes seem to sit between or in the middle of an expensive and lowerend subset of older homes, thus i would not conclude that newer homes are in fact more expensive.

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

  Answer: From the graph I wouldn't come to this conclusion as most homes that are large and expensive seem to be old as shown by the points on the graph. The graph also displays that homes which are small and cheap tend to be older. The newer houses tend to moderately sized and priced when compared to the wide range of prices and sizes for old 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.