AE 03: Duke Forest + data visualization

Gracie Carlaw

```
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!

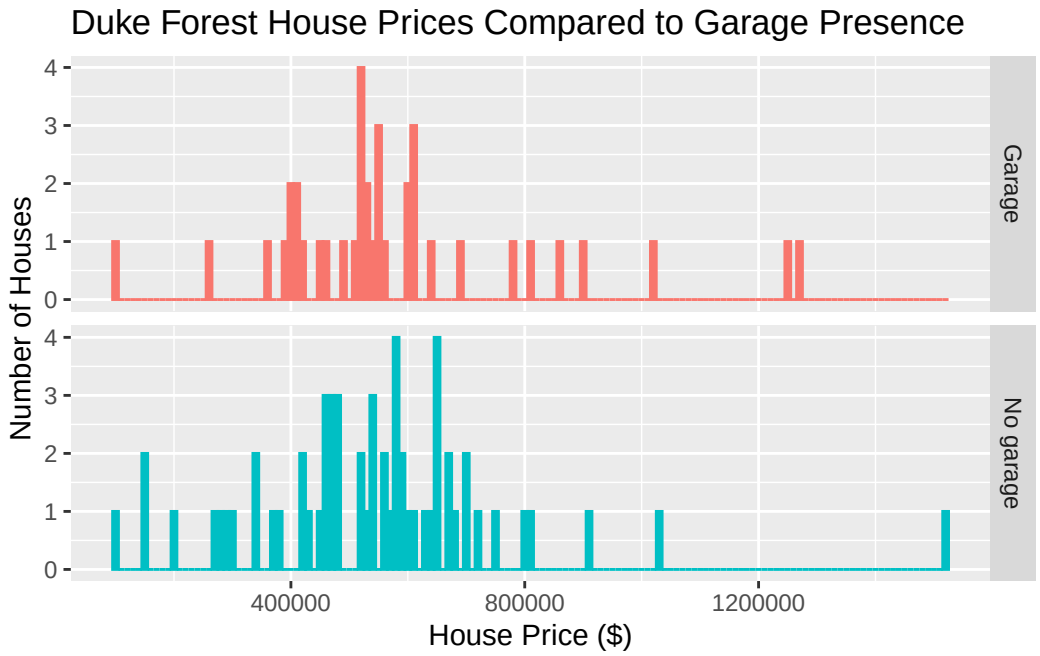
- Make histograms of the prices of houses in Duke Forest based on whether they have a garage.
  - In order to do this, you will first need to create a new variable called `garage` (with levels `"Garage"` and `"No garage"`).
  - Below is the code for creating this new variable. Here, we `mutate()` the `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 %>%
  dplyr::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 %>%
  ggplot(aes(x = price, color = garage, fill = garage)) +
  guides(color = FALSE, fill = FALSE) +
  geom_histogram(binwidth = 10000) +
  facet_grid("garage") +
  labs( x = "House Price ($)",
        y = "Number of Houses",
        title = "Duke Forest House Prices Compared to Garage Presence")
```

```
Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
of ggplot2 3.3.4.
```



Presence of a garage does not appear to affect the price of a house. If anything, lack of a garage may lead to a slightly higher house value, and the most expensive house does not have a garage, but it is negligible as it is an outlier.

> **!** 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?"

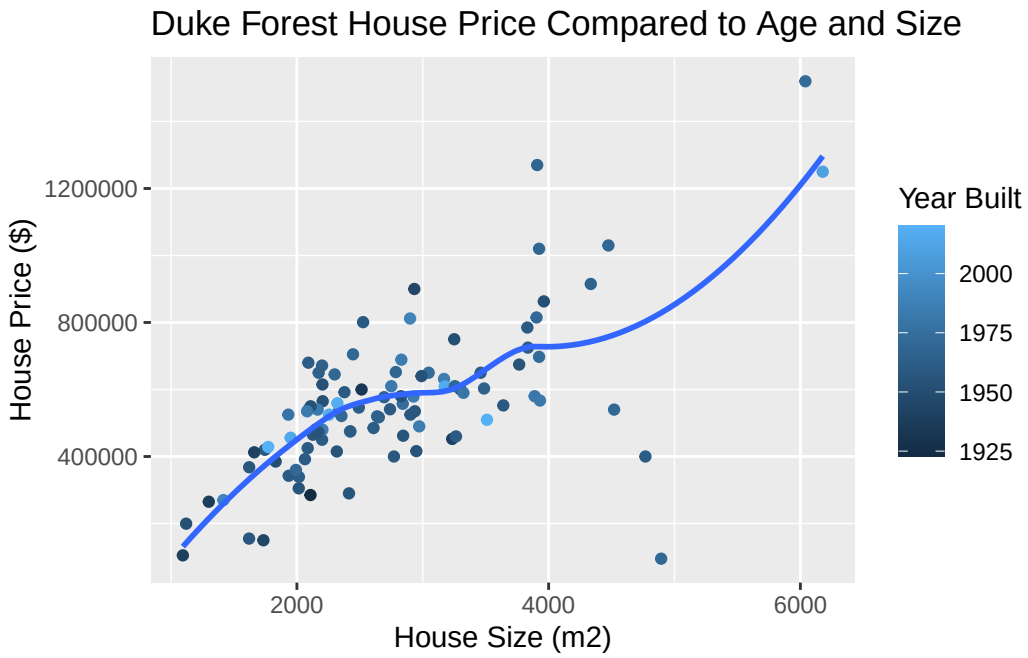
Once again, data visualization skills to the rescue!

- 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) +
  labs( x = "House Size (m2)",
        y = "House Price ($)",
        title = "Duke Forest House Price Compared to Age and Size",
        color = "Year Built")
```

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

```
Warning: The following aesthetics were dropped during statistical transformation:
colour.
i This can happen when ggplot fails to infer the correct grouping structure in
  the data.
i Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```



Claim 1: Generally, houses that are bigger tend to be more expensive, as seen by the line of best fit, which trends positively up and to the right. Claim 2: Overall, there does not appear to be a strong trend that associates house age and price. Older houses and younger houses appear in the same places on thr graph. However, extremely old houses tend to be concentrated around the cheaper end. Claim 3: Big and expensive houses actually tend to be older than younger. There is one outlier, but aside from that, the biggest and most expensive houses are older, around 1975 in age.

> **!** 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.