AE 03: Practicing a bunch of geoms

Suggested answers

APPLICATION EXERCISE ANSWERS

MODIFIED
January 30, 2025

Important

These are suggested answers. This document should be used as reference only, it's not designed to be an exhaustive key.

```
library(tidyverse)

options(scipen = 999)
```

For the following exercises we will work with data on houses that were sold in Tompkins County, NY in 2022-24.[1]

The variables include:

- `sold_date` - date of last recorded sale
- `price` - sale price (in dollars)
- `beds` - number of bedrooms
- `baths` - number of bathrooms. Full bathrooms with shower/toilet count as 1, bathrooms with just a toilet count as 0.5.
- `area` - living area of the home (in square feet)
- `lot_size` - size of property's lot (in acres)
- `year_built` - year home was built
- `hoa_month` - monthly HOA dues. If the property is not part of an HOA, then the value is `NA`
- `town` - Census-defined town in which the house is located.
- `municipality` - Census-defined municipality in which the house is located. If the house is located outside of city or village limits, it is classified as "Unincorporated"
- `long` and `lat` - geographic coordinates of house

The dataset can be found in the `data` folder of your repo. It is called `tompkins-home-sales.csv`. We will import the data and create a new variable, `decade_built_cat`, which identifies the decade in which the home was built. It will include catch-all categories for any homes pre-1940 and post-1990.

```
tompkins <- read_csv("data/tompkins-home-sales.csv") |>
  mutate(decade_built = (year_built %/% 10) * 10) |>
  mutate(
```

```
    decade_built_cat = case_when(
      decade_built <= 1940 ~ "1940 or before",
      decade_built >= 1990 ~ "1990 or after",
      .default = as.character(decade_built)
    )
  )
```

Part 1

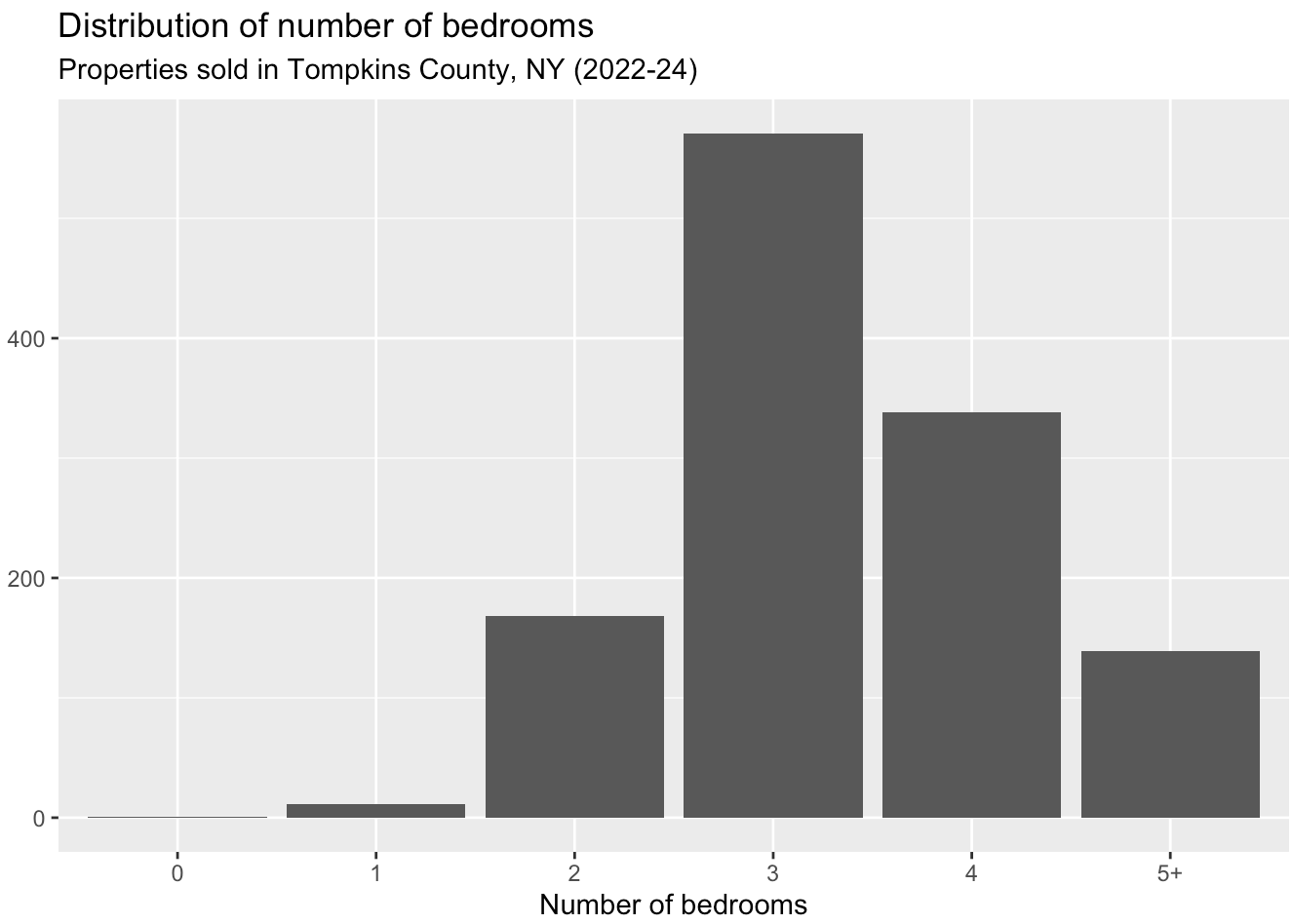
Let's start by visualizing the distribution of the number of bedrooms in the properties sold in Tompkins County, NY in 2022-24. To simplify the task, let's collapse the variable `beds` into a smaller number of categories and drop rows with missing values for this variable.

```
tompkins_beds <- tompkins |>
  mutate(beds = factor(beds) |>
    fct_collapse(
      "5+" = c("5", "6", "7", "9", "11")
    )) |>
  drop_na(beds)
```

Since the number of bedrooms is effectively a categorical variable, we should select a geom appropriate for a single categorical variable.

Your turn: Create a bar chart visualizing the distribution of the number of bedrooms in the properties sold in Tompkins County, NY in 2022-24.

```
ggplot(data = tompkins_beds, mapping = aes(x = beds)) +
  geom_bar() +
  labs(
    title = "Distribution of number of bedrooms",
    subtitle = "Properties sold in Tompkins County, NY (2022-24)",
    x = "Number of bedrooms",
    y = NULL
  )
```



Now let's visualize the distribution of the number of bedrooms by the decade in which the property was built. We will still use a bar chart but also color-code the bar segments for each decade. Now we have a few variations to consider.

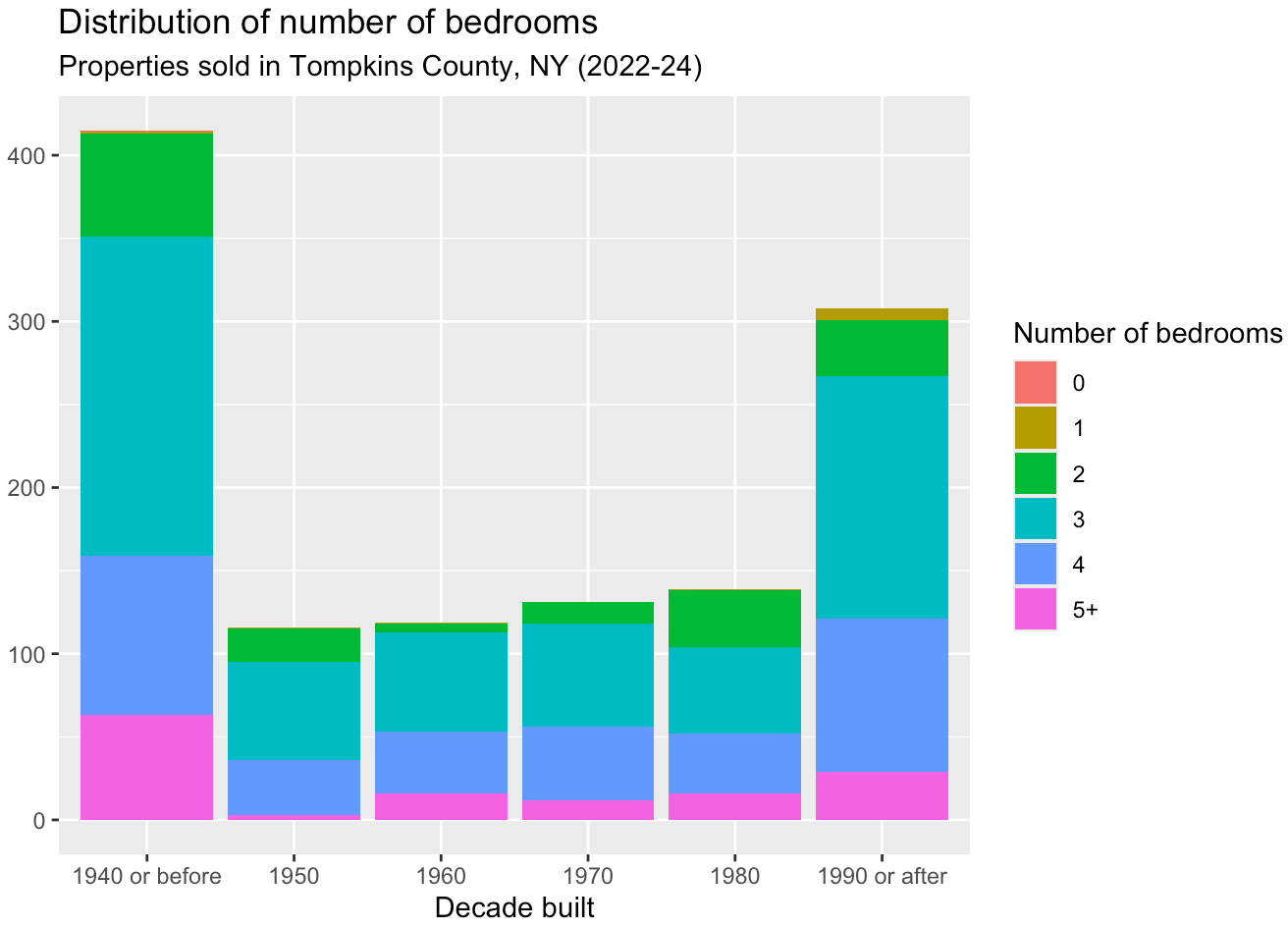
- **Stacked bar chart** - each bar segment represents the frequency count and are stacked vertically on top of each other.[2]
- **Dodged bar chart** - each bar segment represents the frequency count and are placed side by side for each decade. This leaves each segment with a common **origin**, or baseline value of 0.
- **Relative frequency bar chart** - each bar segment represents the relative frequency (proportion) of each category within each decade.

Your turn: Generate each form of the bar chart and compare the differences. Which one do you think is the most informative?

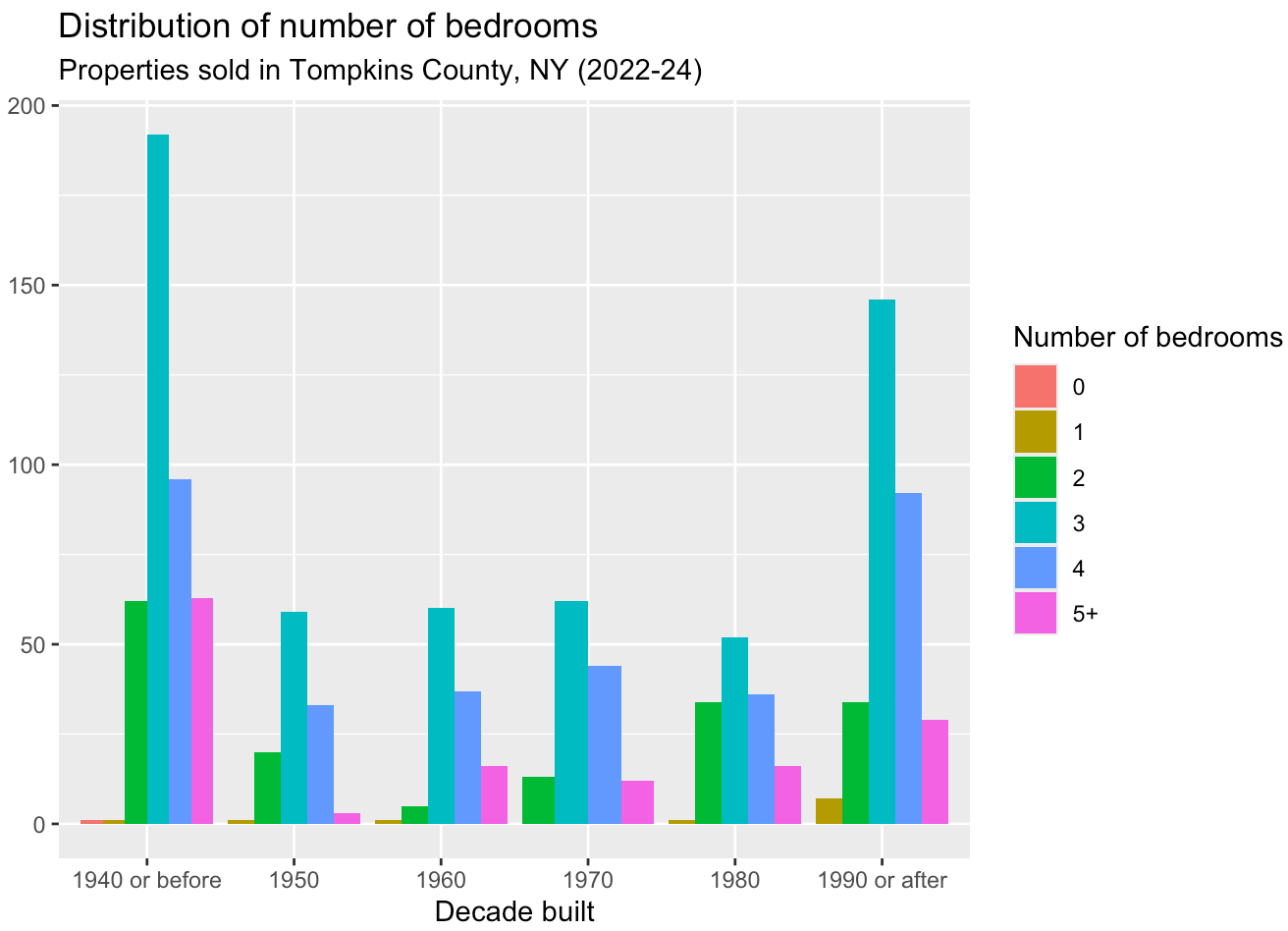
> **Tip**
>
> Read the documentation for `geom_bar()` to identify an appropriate argument for specifying each type of bar chart.

```
ggplot(data = tompkins_beds, mapping = aes(x = decade_built_cat, fill = beds)) +
  geom_bar() +
  labs(
    title = "Distribution of number of bedrooms",
```

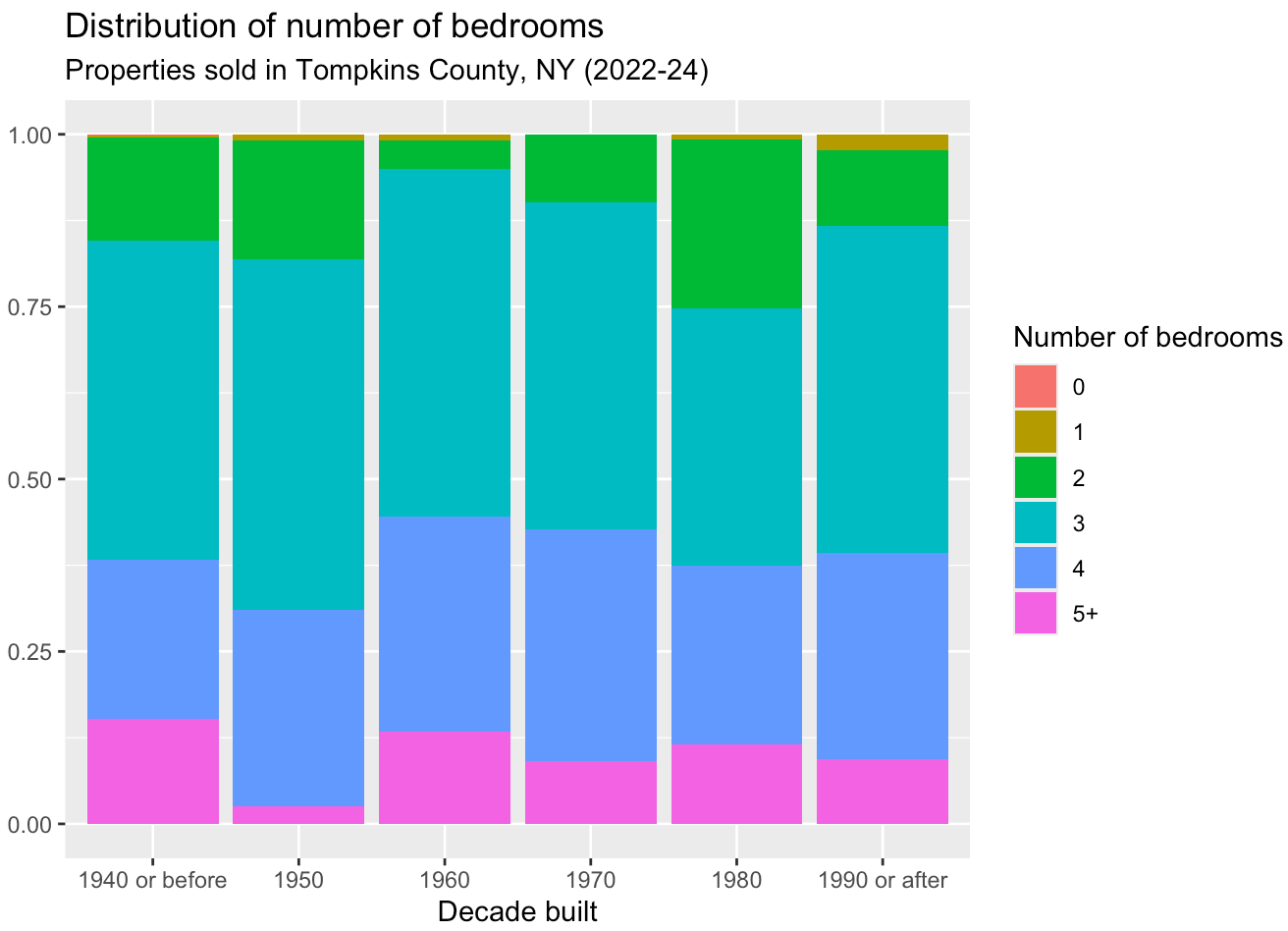
```
    subtitle = "Properties sold in Tompkins County, NY (2022-24)",
    x = "Decade built",
    y = NULL,
    fill = "Number of bedrooms"
  )
```




```
ggplot(data = tompkins_beds, mapping = aes(x = decade_built_cat, fill = beds)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Distribution of number of bedrooms",
    subtitle = "Properties sold in Tompkins County, NY (2022-24)",
    x = "Decade built",
    y = NULL,
    fill = "Number of bedrooms"
  )
```




```
ggplot(data = tompkins_beds, mapping = aes(x = decade_built_cat, fill = beds)) +
  geom_bar(position = "fill") +
  labs(
    title = "Distribution of number of bedrooms",
    subtitle = "Properties sold in Tompkins County, NY (2022-24)",
    x = "Decade built",
    y = NULL,
    fill = "Number of bedrooms"
  )
```



Part 2

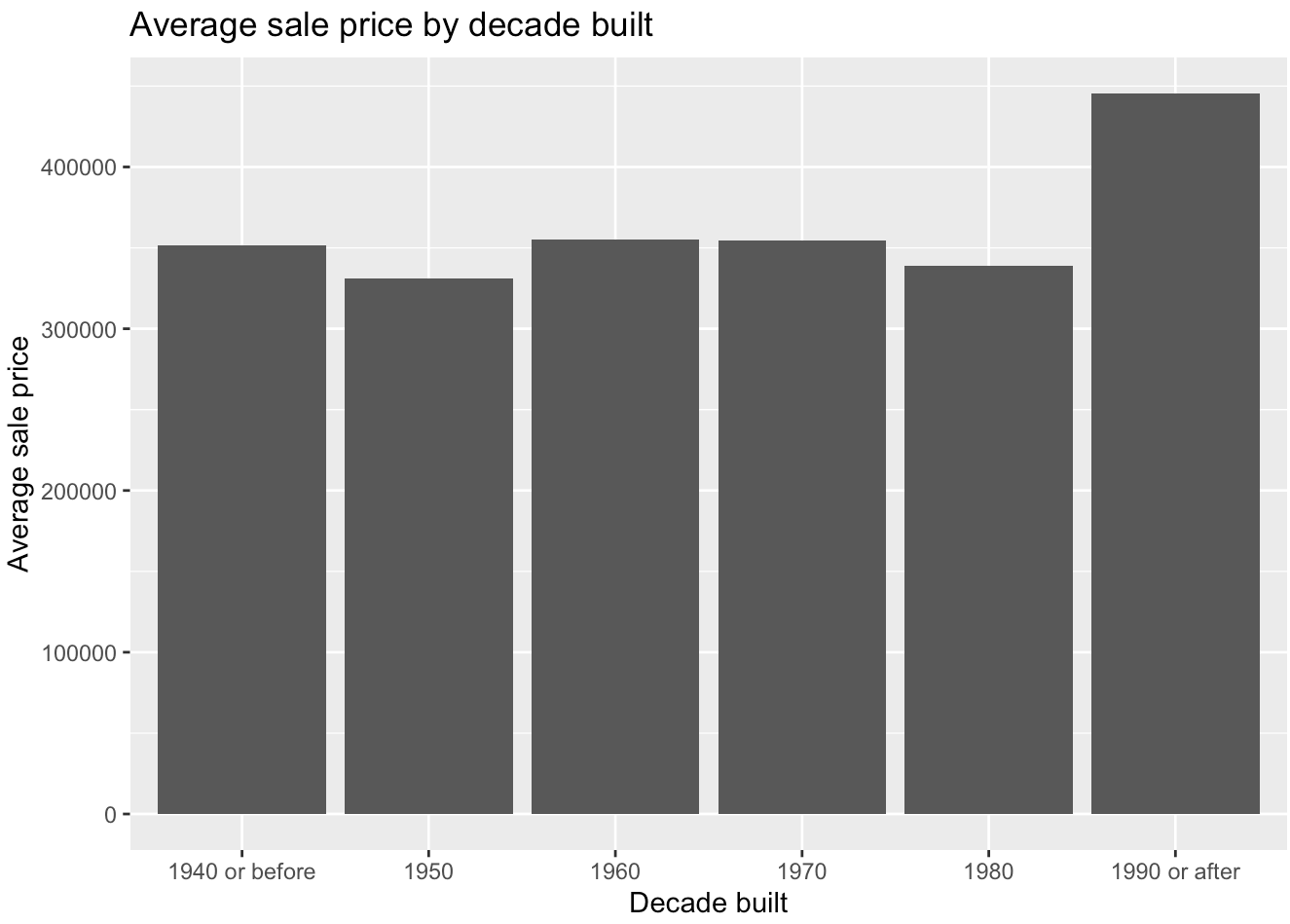
Now let's evaluate the typical sales price (`price`) by the decade in which the property was built. We will start by summarizing the data and then visualize the results using a bar chart and a boxplot.

```
mean_price_decade <- tompkins |>
  group_by(decade_built_cat) |>
  summarize(mean_price = mean(price))
```

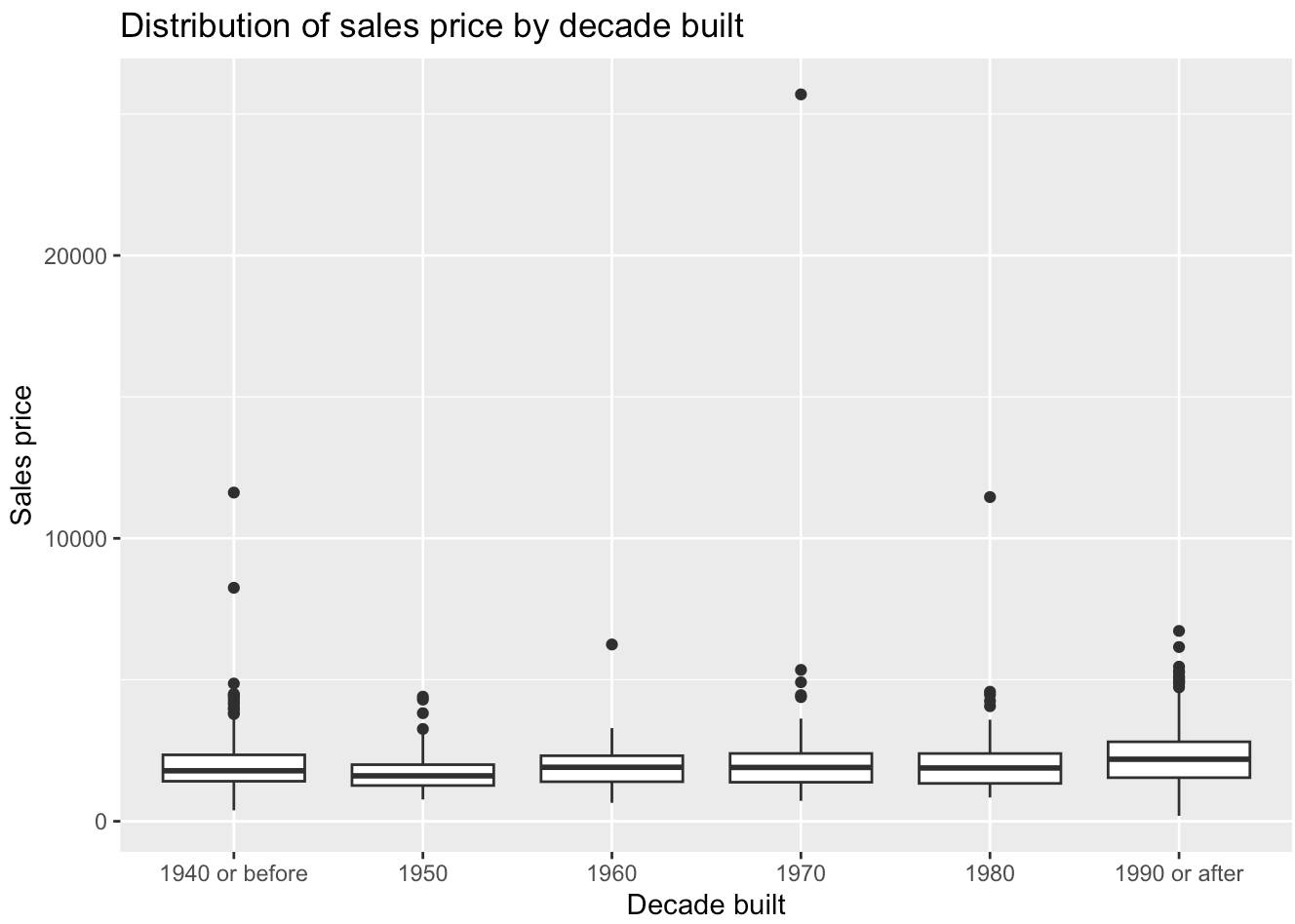
Your turn: Visualize the sales price by the decade in which the property was built. Construct a bar chart reporting the average sales price, as well as a boxplot, violin plot, and strip chart (e.g. jittered scatterplot). What does each graph tell you about the distribution of sales price by decade built? Which ones do you find to be more or less effective?

```
ggplot(data = mean_price_decade, mapping = aes(x = decade_built_cat, y = mean_price)) +
  geom_col() +
  labs(
    title = "Average sale price by decade built",
    x = "Decade built",
```

```
    y = "Average sale price"
  )
```

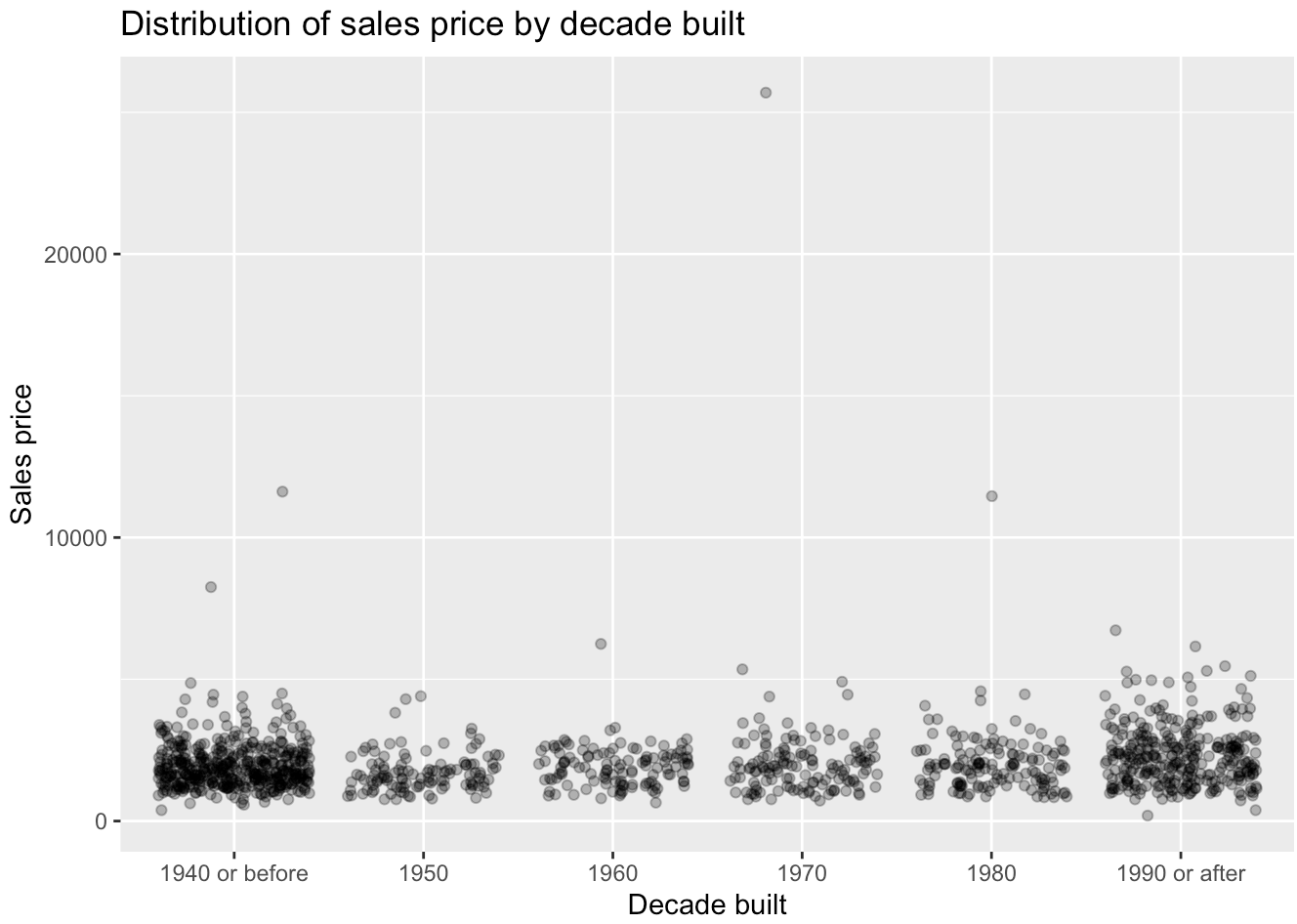



```
ggplot(data = tompkins, mapping = aes(x = decade_built_cat, y = area)) +
  geom_boxplot() +
  labs(
    title = "Distribution of sales price by decade built",
    x = "Decade built",
    y = "Sales price"
  )
```




```
ggplot(data = tompkins, mapping = aes(x = decade_built_cat, y = area)) +
  geom_violin() +
  labs(
    title = "Distribution of sales price by decade built",
    x = "Decade built",
    y = "Sales price"
  )
```


```
set.seed(123) # for reproducibility
ggplot(data = tompkins, mapping = aes(x = decade_built_cat, y = area)) +
  geom_jitter(alpha = 0.3) +
  labs(
    title = "Distribution of sales price by decade built",
    x = "Decade built",
    y = "Sales price"
  )
```



Session information

Footnotes

1. Data source: Redfin. ↩
2. Or horizontally for a horizontal bar chart. ↩