

# Week 2

## Application Exercises

Include tidyverse:

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## — Attaching core tidyverse packages — tidyverse 2.0.0 —
## ✓ dplyr      1.1.4      ✓ readr      2.1.5
## ✓ forcats    1.0.0      ✓ stringr    1.5.1
## ✓ ggplot2    3.4.4      ✓ tibble     3.2.1
## ✓ lubridate  1.9.3      ✓ tidyr      1.3.1
## ✓ purrr      1.0.2
## — Conflicts — tidyverse_conflicts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
theme_set(theme_minimal())
```

Read the data:

```
df <- read_csv("homesales.csv")
```

```
## Rows: 1897 Columns: 12
## — Column specification —
## Delimiter: ","
## chr (4): property_type, address, city, state
## dbl (8): zip_code, price, beds, baths, area, lot_size, year_built, hoa_month
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Average home size by decade:

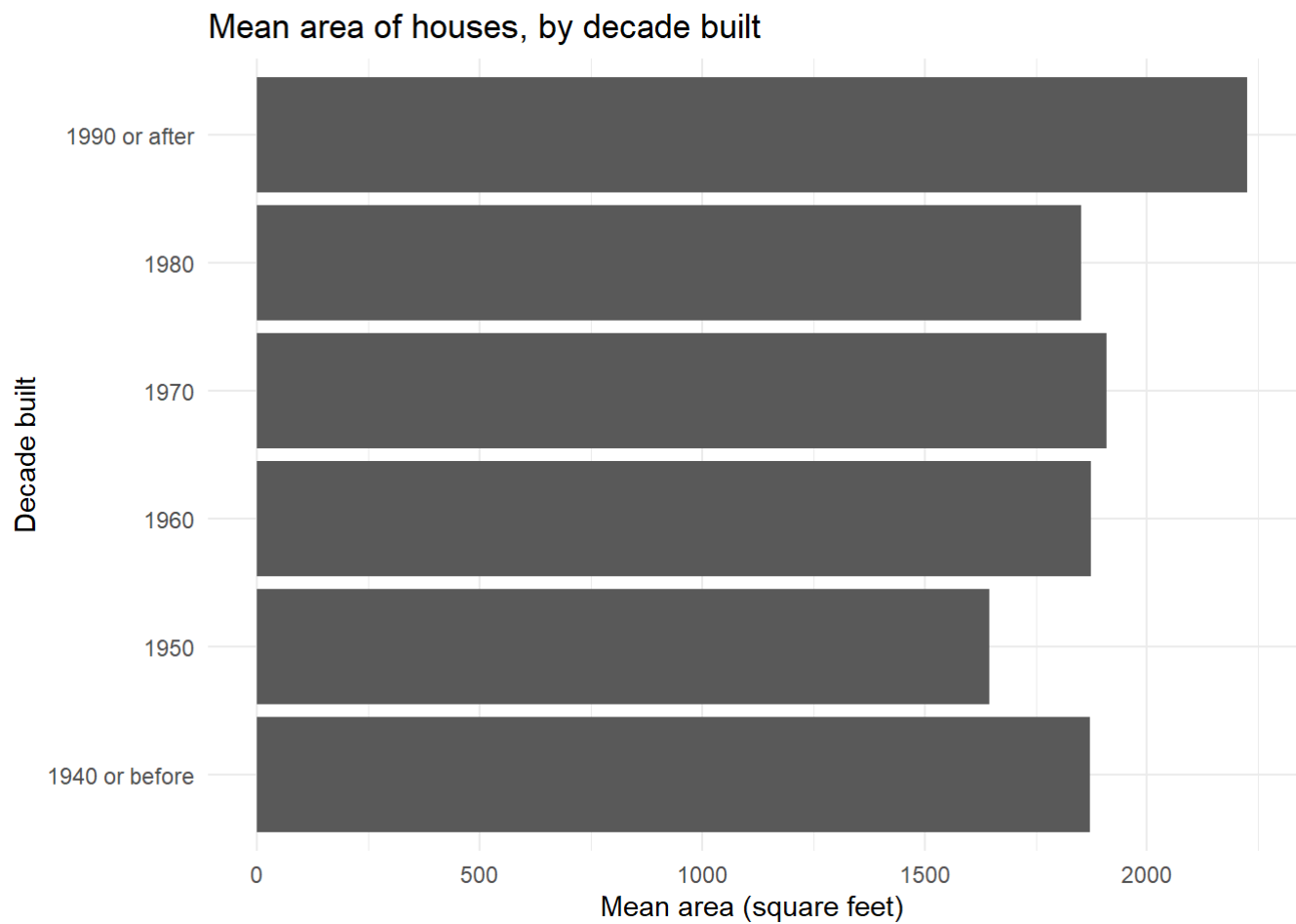
```
# create decade variable
df <- df |>
  mutate(
    decade_built = (year_built %% 10) * 10,
    decade_built_cat = case_when(
      decade_built <= 1940 ~ "1940 or before",
      decade_built >= 1990 ~ "1990 or after",
      .default = as.character(decade_built)
    )
  )

# calculate mean area by decade
mean_area_decade <- df |>
  group_by(decade_built_cat) |>
  summarize(mean_area = mean(area))
mean_area_decade
```

```
## # A tibble: 6 × 2
##   decade_built_cat mean_area
##   <chr>             <dbl>
## 1 1940 or before    1872.
## 2 1950             1645.
## 3 1960             1874.
## 4 1970             1908.
## 5 1980             1852.
## 6 1990 or after    2226.
```

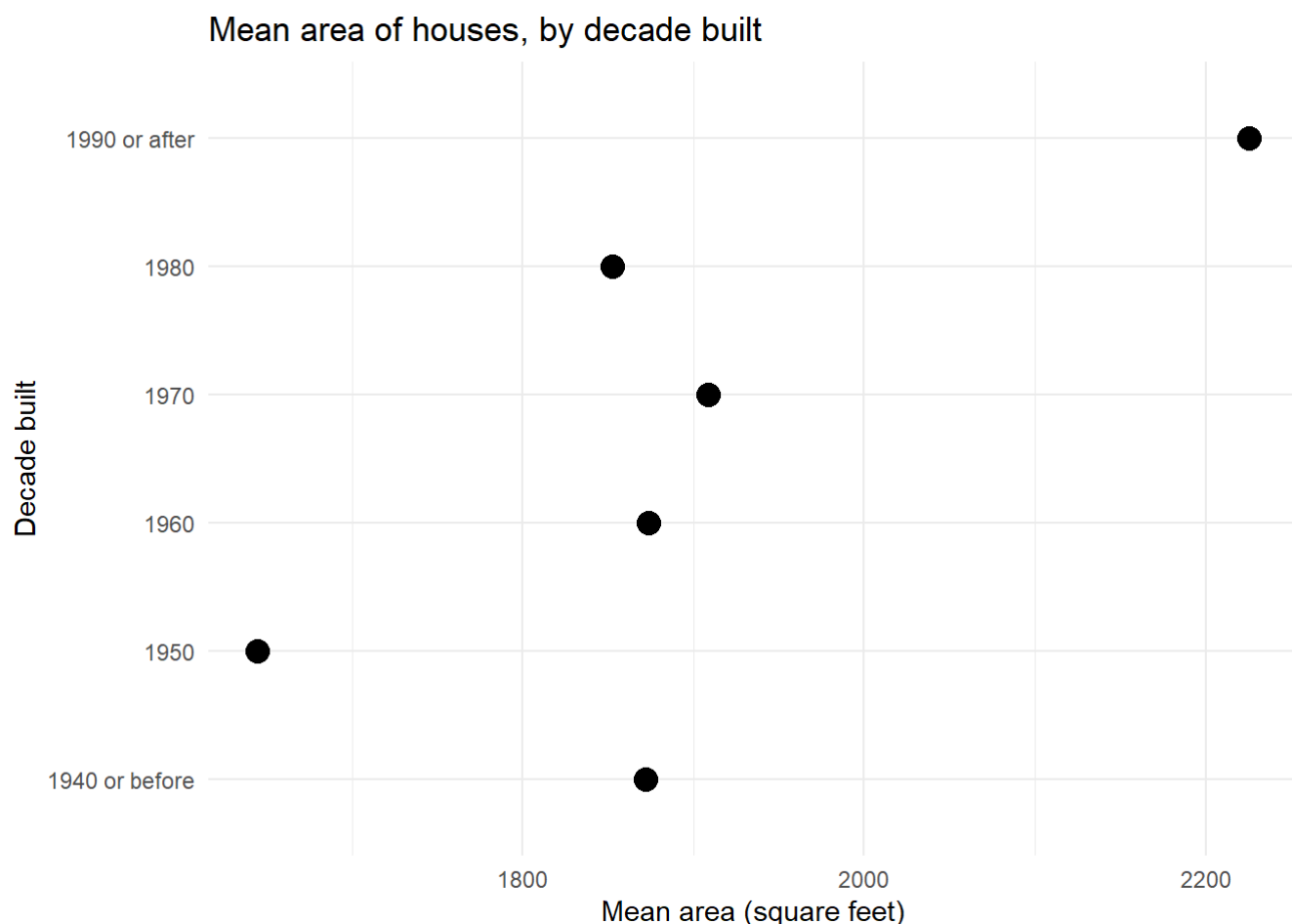
Visualizing the data as a bar chart:

```
ggplot(
  data = mean_area_decade,
  mapping = aes(x = mean_area, y = decade_built_cat)
) +
  geom_col() +
  labs(
    x = "Mean area (square feet)", y = "Decade built",
    title = "Mean area of houses, by decade built"
  )
```



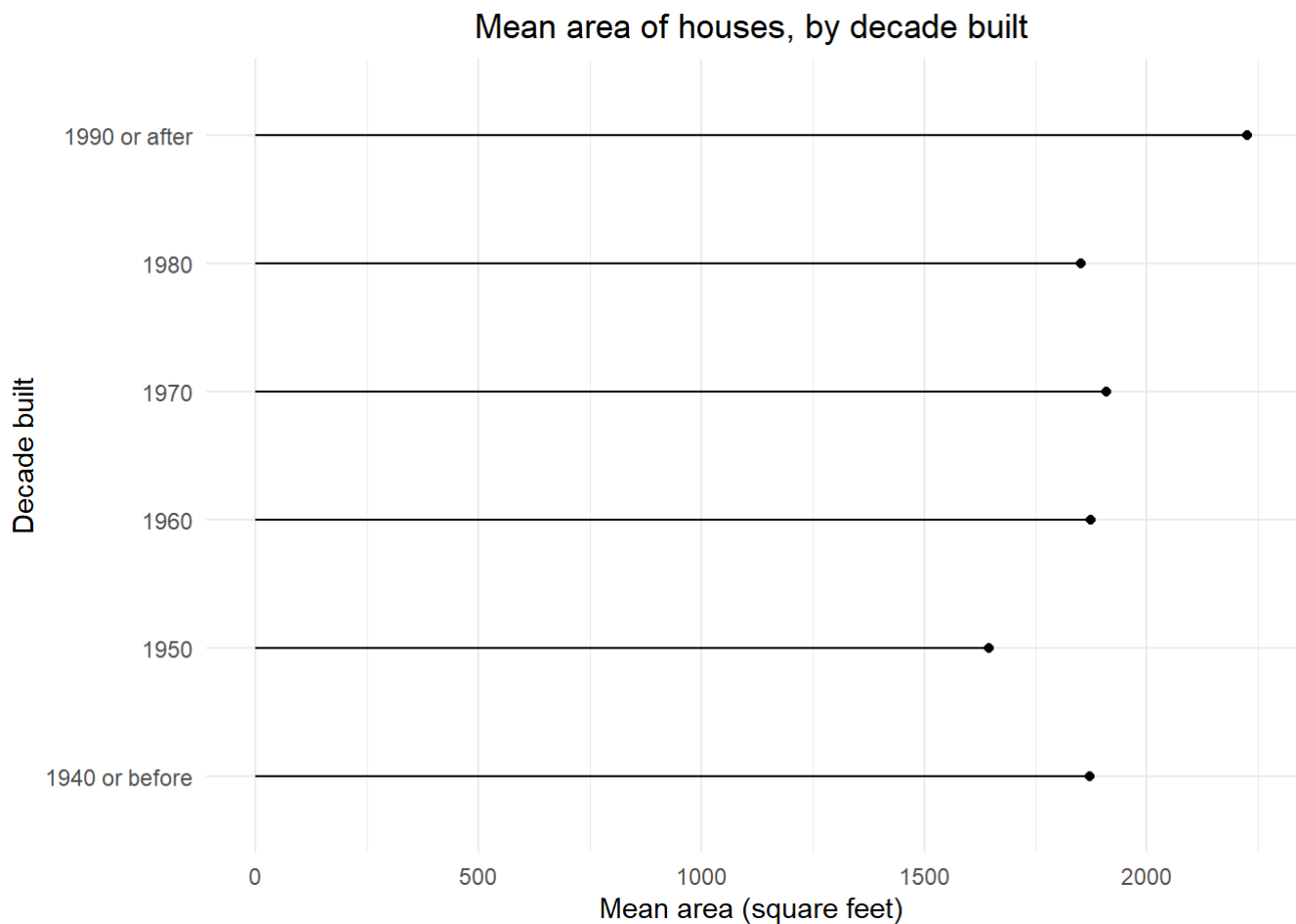
Visualizing the data as a dot plot:

```
ggplot(  
  data = mean_area_decade,  
  mapping = aes(x = mean_area, y = decade_built_cat)  
) +  
  geom_point(size = 4) +  
  labs(  
    x = "Mean area (square feet)", y = "Decade built",  
    title = "Mean area of houses, by decade built"  
  )  
)
```



## TASK 1. Visualizing the data as a lollipop chart

```
# YOUR CODE HERE
ggplot(
  data = mean_area_decade,
  mapping = aes(x = mean_area, y = decade_built_cat)
) +
  geom_segment(aes(x = 0, xend = mean_area, y = decade_built_cat, yend = decade_built_cat)) +
  geom_point() +
  labs(
    x = "Mean area (square feet)", y = "Decade built",
    title = "Mean area of houses, by decade built"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```

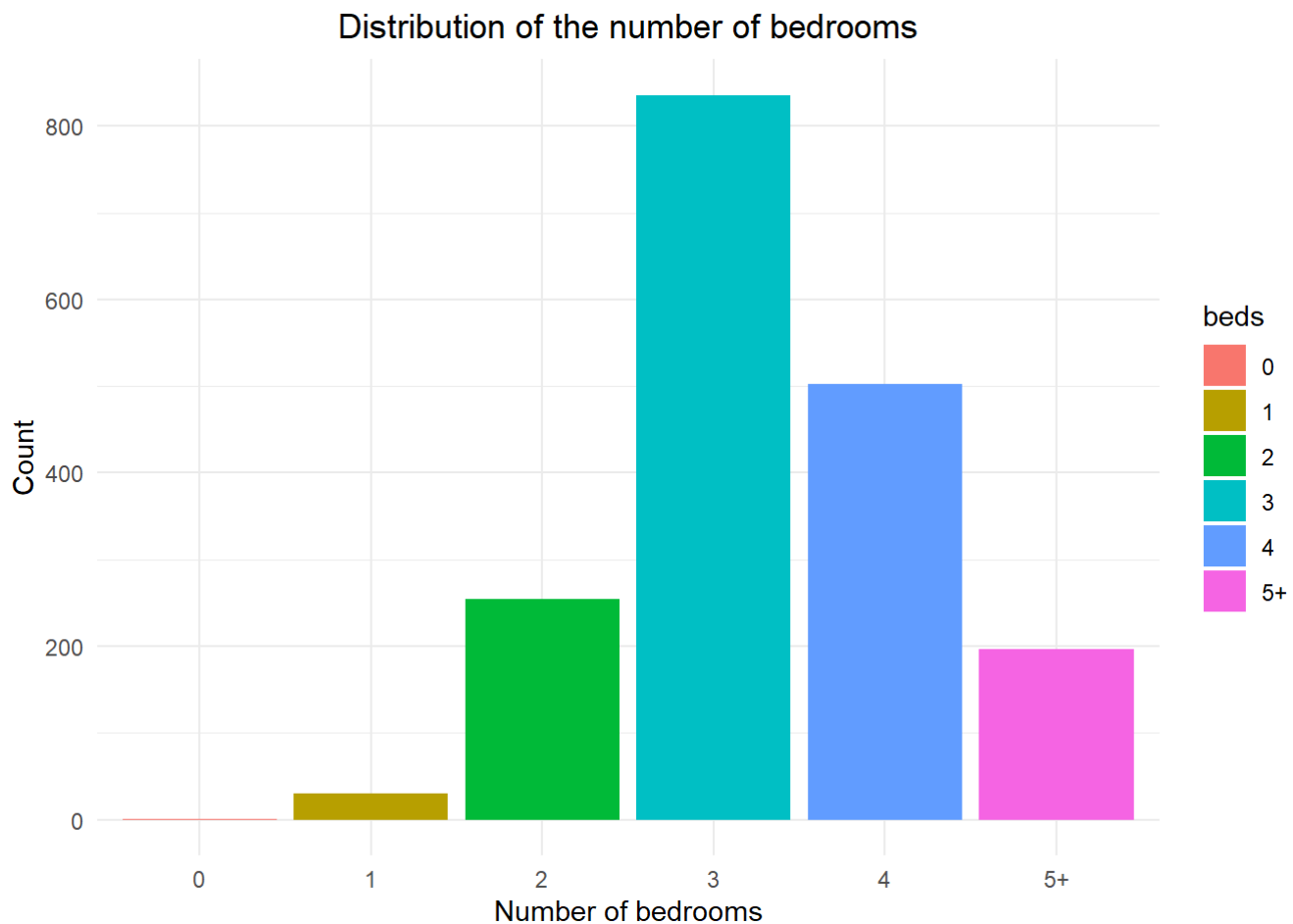


## TASK 2. Visualizing the distribution of the number of bedrooms

Collapse the variable `beds` into a smaller number of categories and drop rows with missing values for this variable:

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

```
# YOUR CODE HERE
ggplot(data = df_bed, aes(x = beds, fill = beds)) + geom_bar() +
  labs(x = "Number of bedrooms", y = "Count",
    title = "Distribution of the number of bedrooms") +
  theme(plot.title = element_text(hjust = 0.5))
```



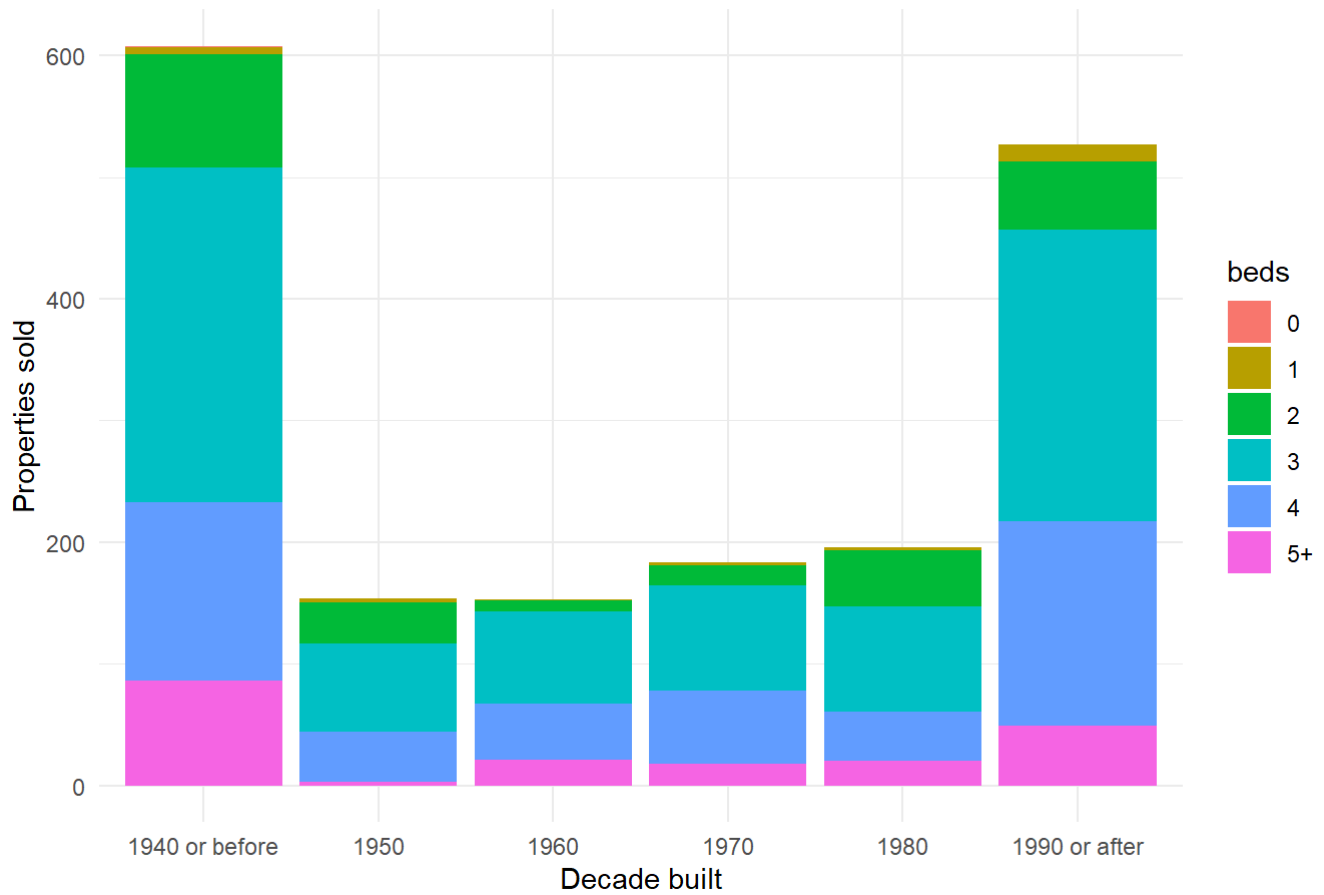
## TASK 3. Visualizing the distribution of the number of bedrooms by the decade in which the property was built

Stacked bar chart (number of bedrooms by the decade built):

# YOUR CODE HERE

```
ggplot(data = df_bed, aes(x = decade_built_cat, fill = beds)) + geom_bar() +  
labs(x = "Decade built", y = "Properties sold",  
title = "Distribution of the number of bedrooms by decade") +  
theme(plot.title = element_text(hjust = 0.5))
```

Distribution of the number of bedrooms by decade

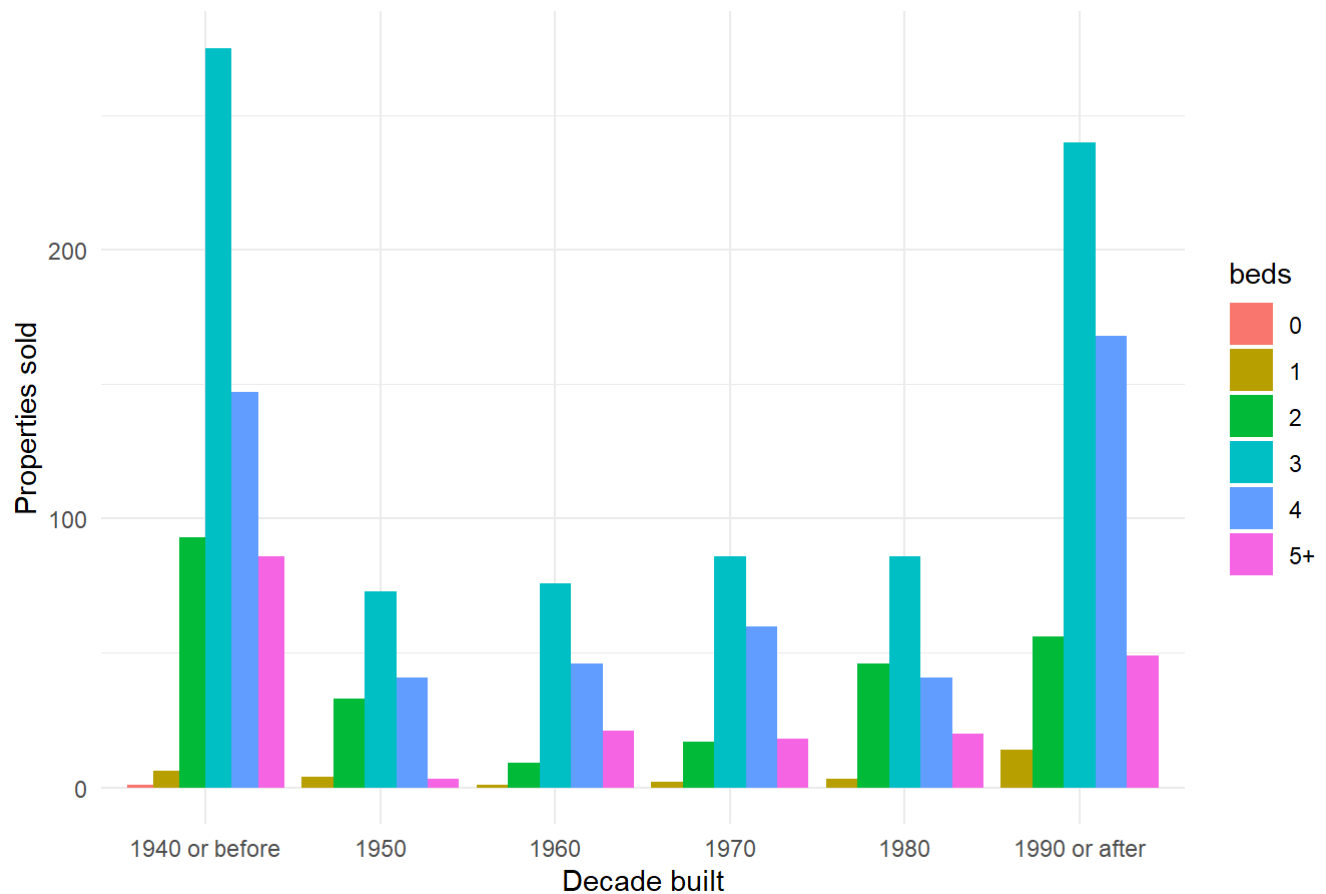


Dodged bar chart (number of bedrooms by the decade built):

# YOUR CODE HERE

```
ggplot(data = df_bed, aes(x = decade_built_cat, fill = beds)) + geom_bar(position = "dodge")
+
labs(x = "Decade built", y = "Properties sold",
title = "Distribution of the number of bedrooms by decade") +
theme(plot.title = element_text(hjust = 0.5))
```

Distribution of the number of bedrooms by decade

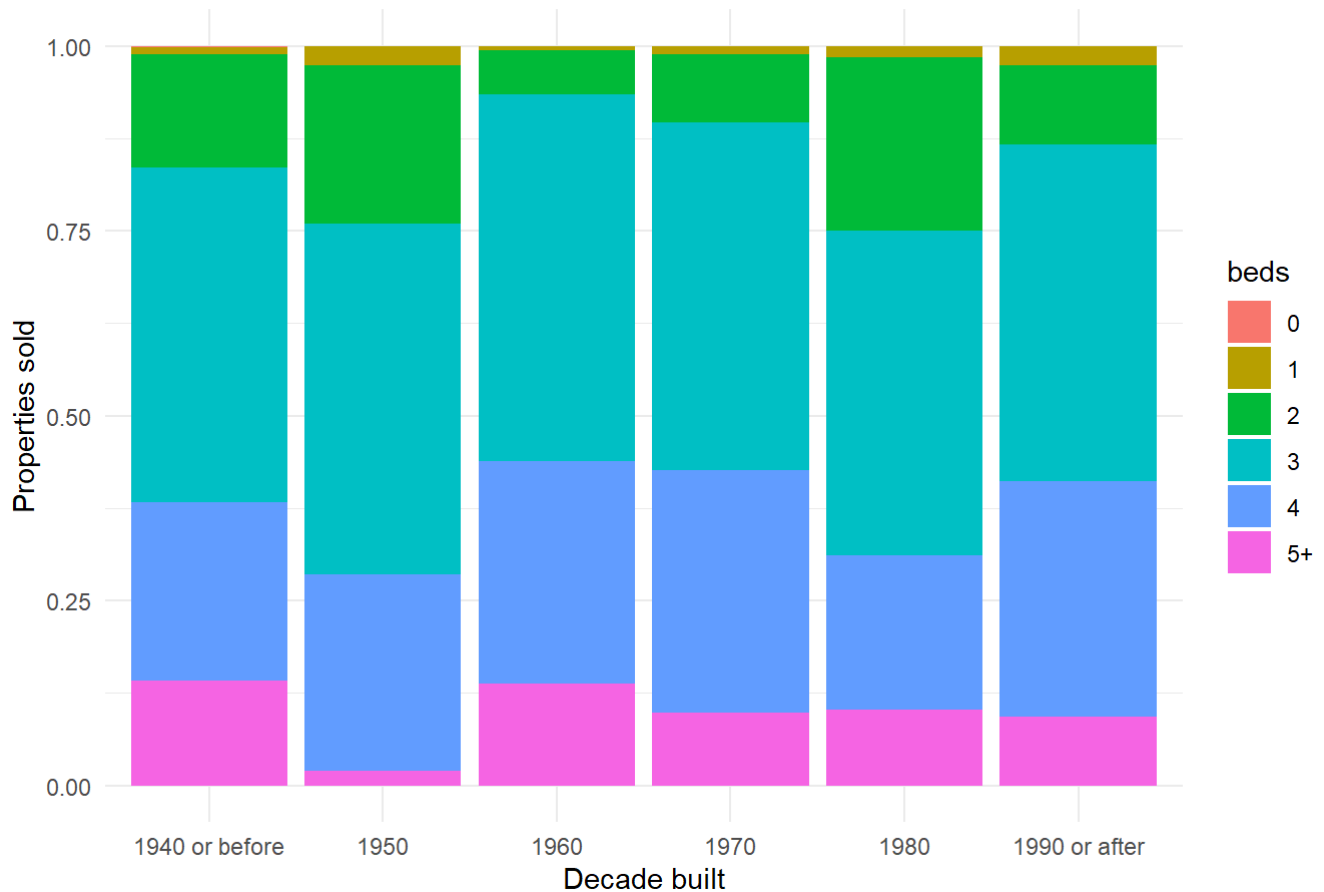


Relative frequency bar chart (number of bedrooms by the decade built):

```
# YOUR CODE HERE
ggplot(data = df_bed, aes(x = decade_built_cat, fill = beds)) + geom_bar(position = "fill") +
  labs(x = "Decade built", y = "Properties sold",
  title = "Distribution of the number of bedrooms by decade") +
  theme(plot.title = element_text(hjust = 0.5))
```



Distribution of the number of bedrooms by decade



## Task 4. Visualizing the distribution of property size by decades

Getting mean of area of each decade category:

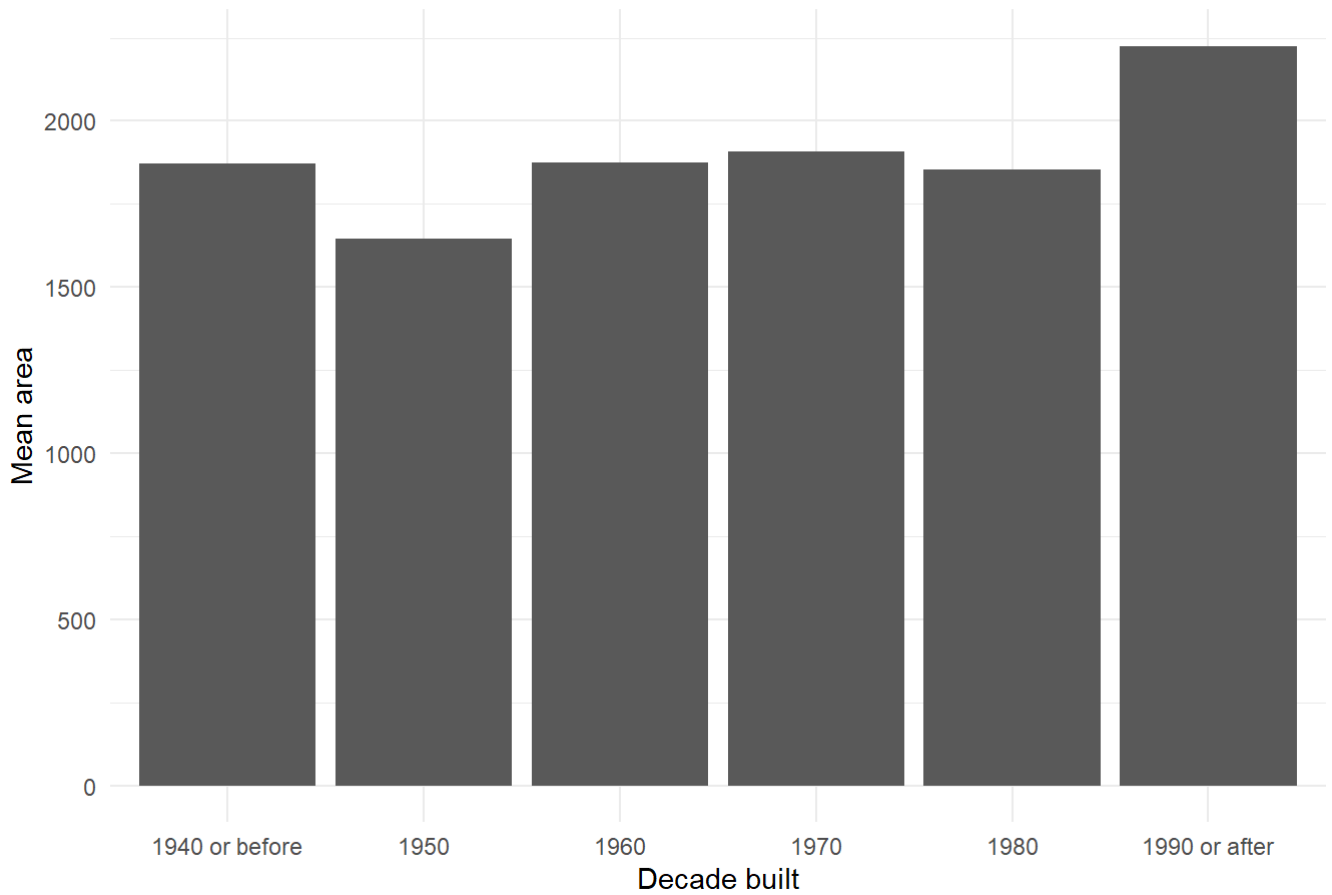
```
mean_area_decade <- df |>
  group_by(decade_built_cat) |>
  summarize(mean_area = mean(area))
```

Bar chart (mean area by decade built):

```
# YOUR CODE HERE

ggplot(data = mean_area_decade, aes(x = decade_built_cat, y = mean_area)) + geom_bar(stat =
"identity") +
labs(x = "Decade built", y = "Mean area",
title = "Mean area by decade") +
theme(plot.title = element_text(hjust = 0.5))
```

Mean area by decade

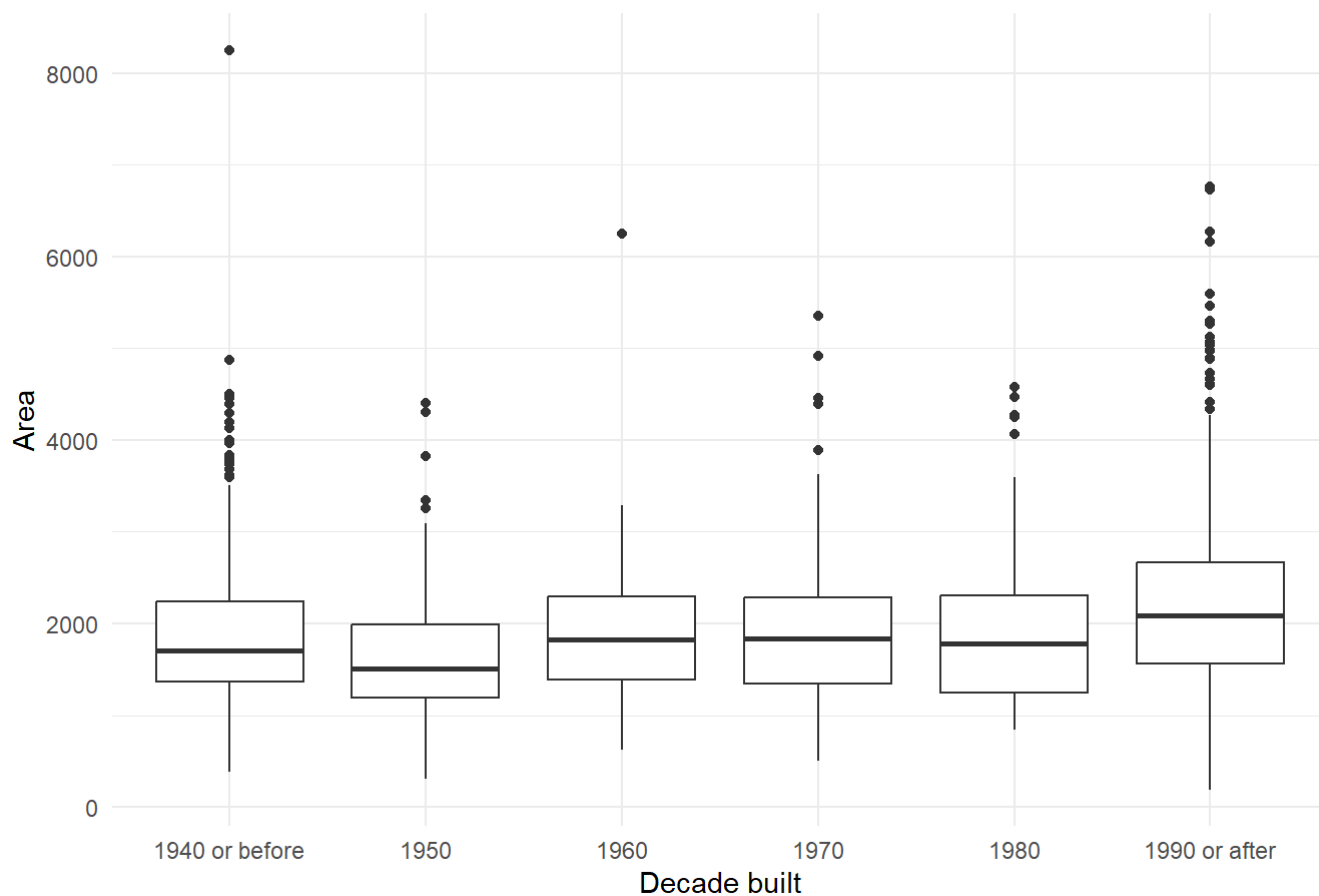


Box plot (area by decade built):

*# YOUR CODE HERE*

```
ggplot(data = df, aes(x = decade_built_cat, y = area)) + geom_boxplot() +  
labs(x = "Decade built", y = "Area",  
      title = "Distribution of area by decade") +  
theme(plot.title = element_text(hjust = 0.5))
```

Distribution of area by decade

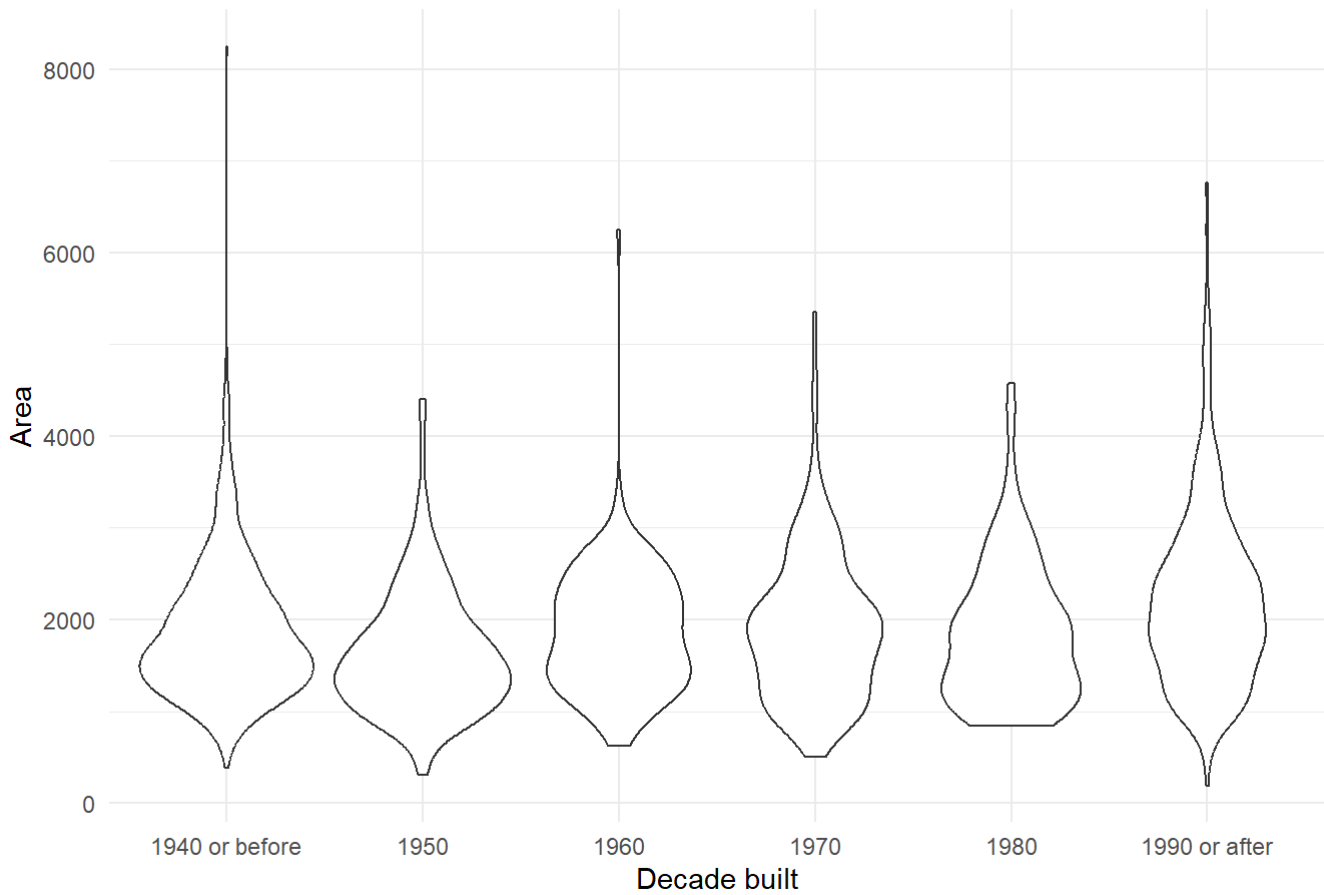


Violin plot (area by decade built):

# YOUR CODE HERE

```
ggplot(data = df, aes(x = decade_built_cat, y = area)) + geom_violin() +  
labs(x = "Decade built", y = "Area",  
      title = "Distribution of area by decade") +  
theme(plot.title = element_text(hjust = 0.5))
```

Distribution of area by decade



Strip chart (area by decade built):

```
set.seed(4010)

# YOUR CODE HERE

ggplot(data = df, aes(x = decade_built_cat, y = area)) + geom_jitter(alpha = 0.3) +
  labs(x = "Decade built", y = "Area",
       title = "Distribution of area by decade") +
  theme(plot.title = element_text(hjust = 0.5))
```

Distribution of area by decade

