Amrutha valli

house\_sales

DSSTCS (2ND YEAR)

AWDCRJY

**➢ step 1: Data Loading and Overview**

The data is loaded from a CSV file, and a quick overview is provided using functions such as head(), sample(), and str() to understand the structure and content of the dataset.

**Code: -**

data <- read.csv(file.choose(), header=TRUE)

▪ The read.csv() function is used to load a CSV file. file.choose() opens a file dialog for the user to select the file interactively. The resulting data is stored in a variable called data.

**#LOAD PACKAGE INTO LIBRARY FOR USE:**

**# Install and load necessary libraries**

library(dplyr)

library(ggplot2)

#-------------------------------------------------------------------------------------

**## Importing the dataset**

house\_sales <- read.csv(file.choose(),header = T)

#gives entire data set.

View(house\_sales)

#gives first few rows

head(house\_sales)

# list of last five rows

tail(house\_sales)

#gives the columns names

names(house\_sales)

#dimensions of dataset

#gives no.of rows and columns

dim(house\_sales)

# displays the internal structure of an object in a compact way.

str(house\_sales)

# We can also see the number of unique values in the dataset.

library(dplyr)

n\_distinct(house\_sales$price)

n\_distinct(house\_sales$sqft)

n\_distinct(house\_sales$bedrooms)

n\_distinct(house\_sales$bathrooms)

n\_distinct(house\_sales$zip\_code)

n\_distinct(house\_sales$house\_style)

n\_distinct(house\_sales$year\_built)

n\_distinct(house\_sales$lot\_size)

n\_distinct(house\_sales$garage\_size)

n\_distinct(house\_sales$fire\_place)

n\_distinct(house\_sales$air\_condition)

n\_distinct(house\_sales$gender)

# Step 1: Calculate the number of null values to add

total\_rows <- nrow(house\_sales)

total\_rows

null\_count <- round(0.05 \* total\_rows)

# Step 2: Randomly assign null values to the column

null\_indices <- sample(1:total\_rows, null\_count)

null\_indices

house\_sales$garage\_size[null\_indices] <- NA

house\_sales$garage\_size

# check for duplicates values

sum(duplicated(house\_sales))

#checking for null values

sum(is.null(house\_sales))

## check for na values

sum(is.na(house\_sales))

#checking summary of garage\_size

summary(c(house\_sales$garage\_size))

#replace na values with with 0

house\_sales$garage\_size[is.na(house\_sales$garage\_size)] <- 0

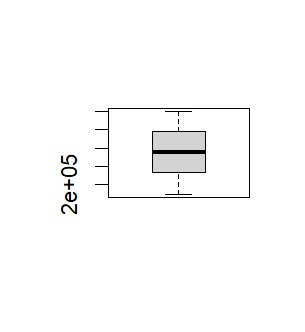
house\_sales$garage\_size

# to get a better understanding of the data set,

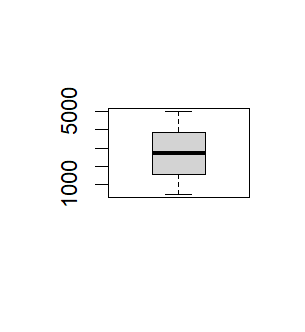
# we can also see the statistical summary of the data set.

summary(house\_sales)

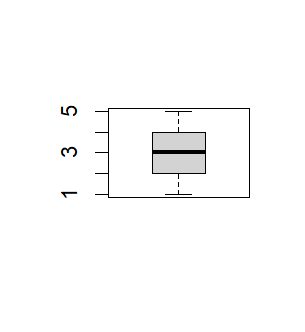
boxplot(house\_sales$price)



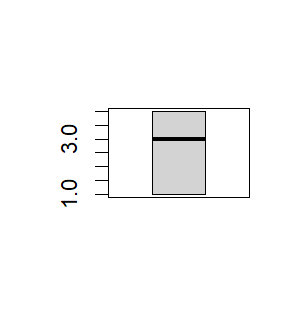
boxplot(house\_sales$sqft)



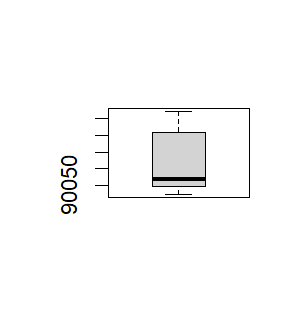
boxplot(house\_sales$bedrooms)



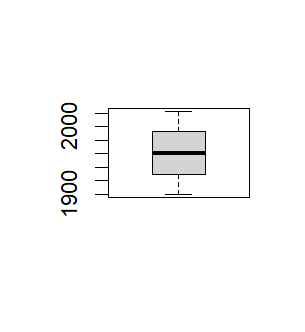
boxplot(house\_sales$bathrooms)



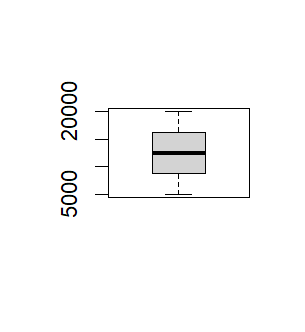
boxplot(house\_sales$zip\_code)



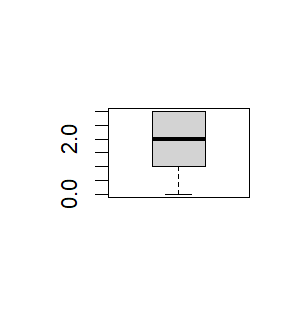
boxplot(house\_sales$year\_built)



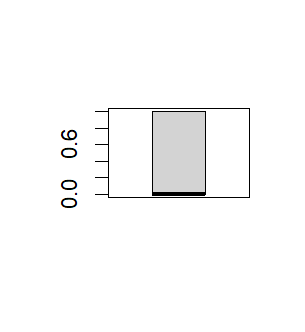
boxplot(house\_sales$lot\_size)



boxplot(house\_sales$garage\_size)



boxplot(house\_sales$fire\_place)



# Find the year in which the house built first

first\_house <- house\_sales$year\_built[which.min(house\_sales$year\_built)]

first\_house

# Find the year in which the house built recently

last\_house <- house\_sales$year\_built[which.max(house\_sales$year\_built)]

last\_house

# Calculate the number of years since the house was built

current\_year <- as.integer(format(Sys.Date(), "%Y"))

house\_sales$years\_since\_built <- 2023 - house\_sales$year\_built

# View the updated dataset

View(house\_sales)

# Rename the "gender" column to "house\_owner"

names(house\_sales)[names(house\_sales) == "gender"] <- "house\_owner"

house\_sales$house\_owner

# Reorder the columns

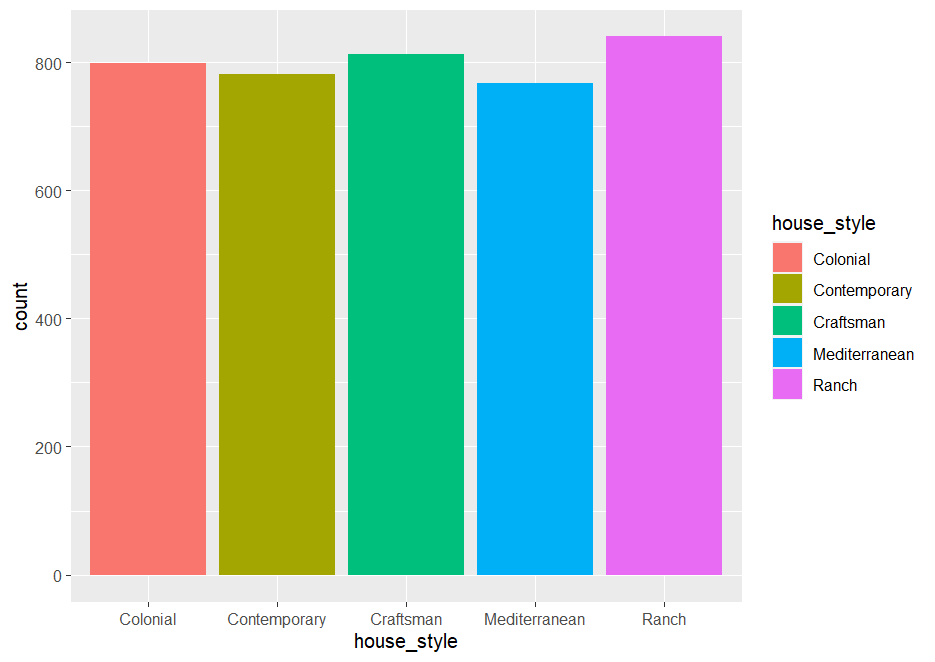
house\_sales <- house\_sales[,c("house\_style","year\_built","years\_since\_built","sqft","price","bedrooms","bathrooms","air\_condition","lot\_size","garage\_size","fire\_place","house\_owner","zip\_code")]

library(ggplot2)

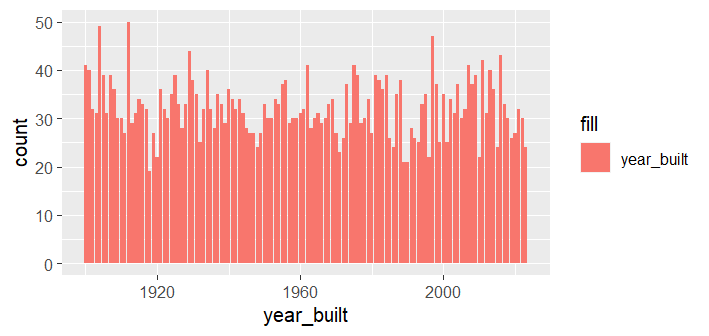
library(dplyr)

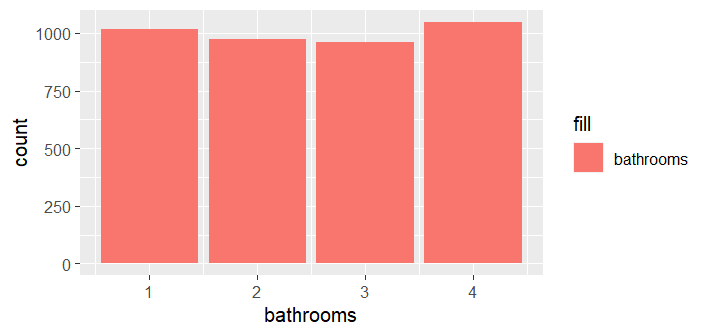
#bar plot

ggplot(house\_sales,aes(x=house\_style))+geom\_bar()



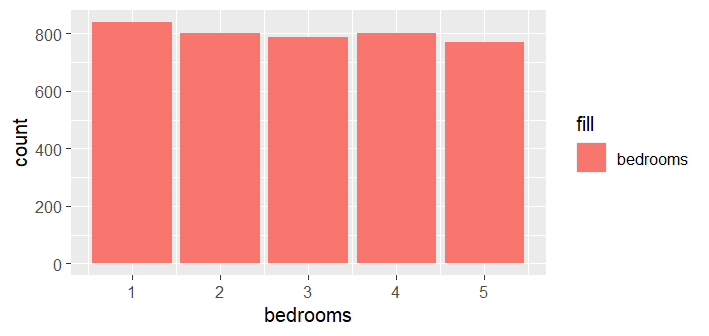
#we can observe in this graph is that the Ranch style is the most common, followed very closely by the Craftsman style and the Colonial style. The Mediterranean style is less common than these, and the Contemporary style is the least frequent of all in this dataset.

ggplot(house\_sales,aes(x=year\_built,fill ="year\_built"))+geom\_bar()

ggplot(house\_sales,aes(x=bathrooms,fill = "bathrooms"))+geom\_bar()

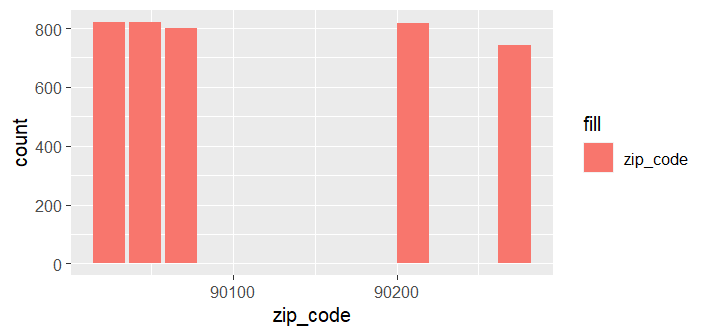
#we can observe that 4 bathrooms are the most among them followed by 1 and 2 bathrooms.

ggplot(house\_sales,aes(x=bedrooms,fill ="bedrooms"))+geom\_bar()



##we can observe that 1 bedroom are the most among them followed by 2 and 3 bathrooms.

ggplot(house\_sales,aes(x=zip\_code,fill = "zip\_code"))+geom\_bar()



# Filter the data set for houses with the "Colonial" style

Colonial\_houses <- house\_sales[house\_sales$house\_style == "Colonial", ]

Colonial\_houses

# Count the number of Colonial houses sold in each year

Colonial\_houses\_counts <- table(Colonial\_houses$year\_built)

Colonial\_houses\_counts

# Convert year\_counts to a data frame

Colonial\_houses\_counts\_df<- data.frame(year = as.numeric(names(Colonial\_houses\_counts)),

count = as.numeric(Colonial\_houses\_counts))

Colonial\_houses\_counts\_df

# Create a bar graph of the number of Colonial style houses built in each year

ggplot(Colonial\_houses\_counts\_df, aes(x = factor(year), y = count)) +

geom\_bar(stat = "identity", fill = "purple3") +

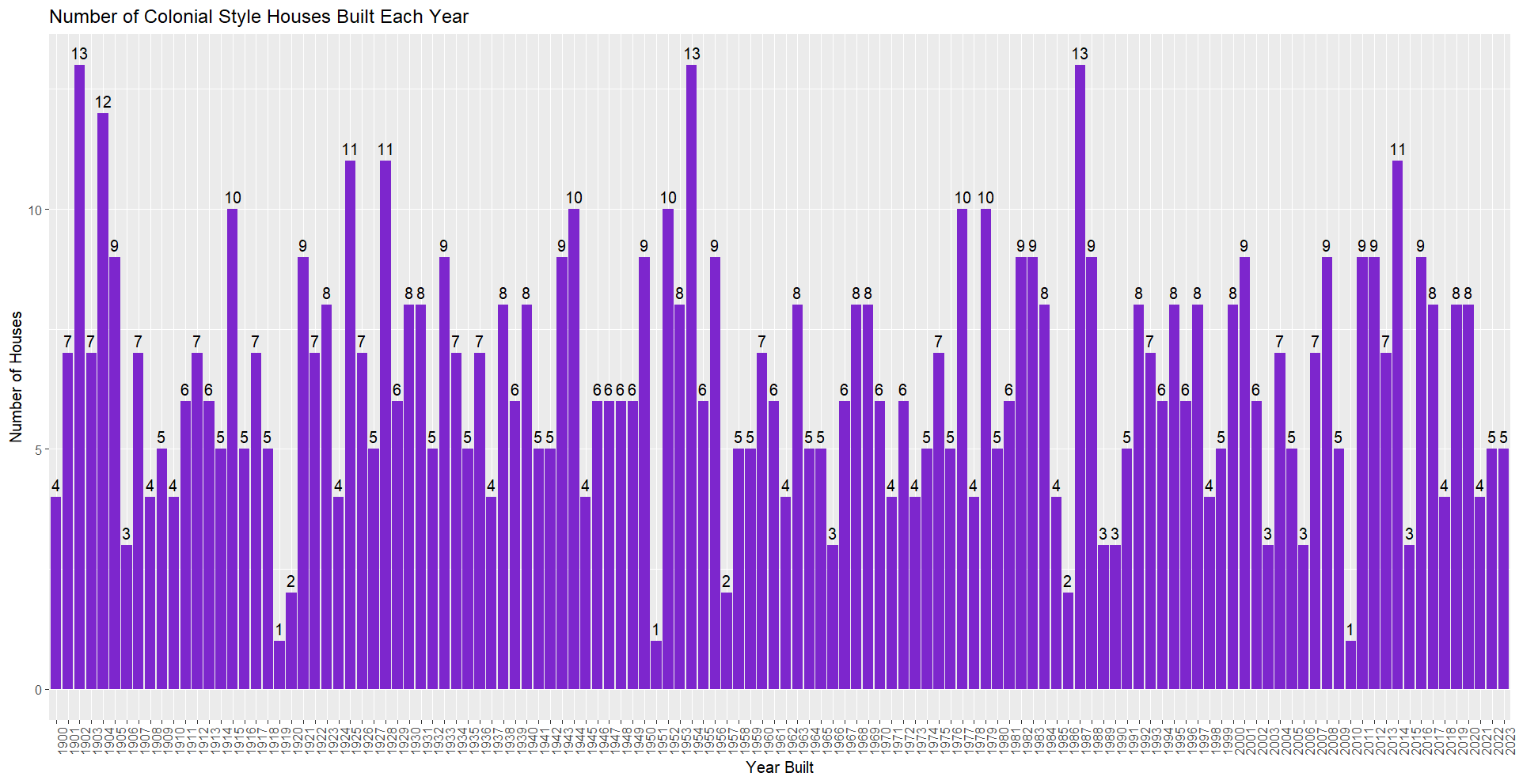
geom\_text(aes(label = count), vjust = -0.5, color = "black") +

labs(title = "Number of Colonial Style Houses Built Each Year",

x = "Year Built",

y = "Number of Houses") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))



#In the year 1902,1953 and 1987 colonial style houses are build most.(13 houses)

# Filter the data set for houses with the "Contemporary" style

Contemporary\_houses <- house\_sales[house\_sales$house\_style == "Contemporary", ]

Contemporary\_houses

# Count the number of Contemporary houses sold in each year

Contemporary\_houses\_counts <- table(Contemporary\_houses$year\_built)

Contemporary\_houses\_counts

# Convert year\_counts to a data frame

Contemporary\_houses\_counts\_df<- data.frame(year = as.numeric(names(Contemporary\_houses\_counts)),

count = as.numeric(Contemporary\_houses\_counts))

Contemporary\_houses

# Create a bar graph of the number of Contemporary style houses built in each year

ggplot(Contemporary\_houses\_counts\_df, aes(x = factor(year), y = count)) +

geom\_bar(stat = "identity", fill = "seagreen") +

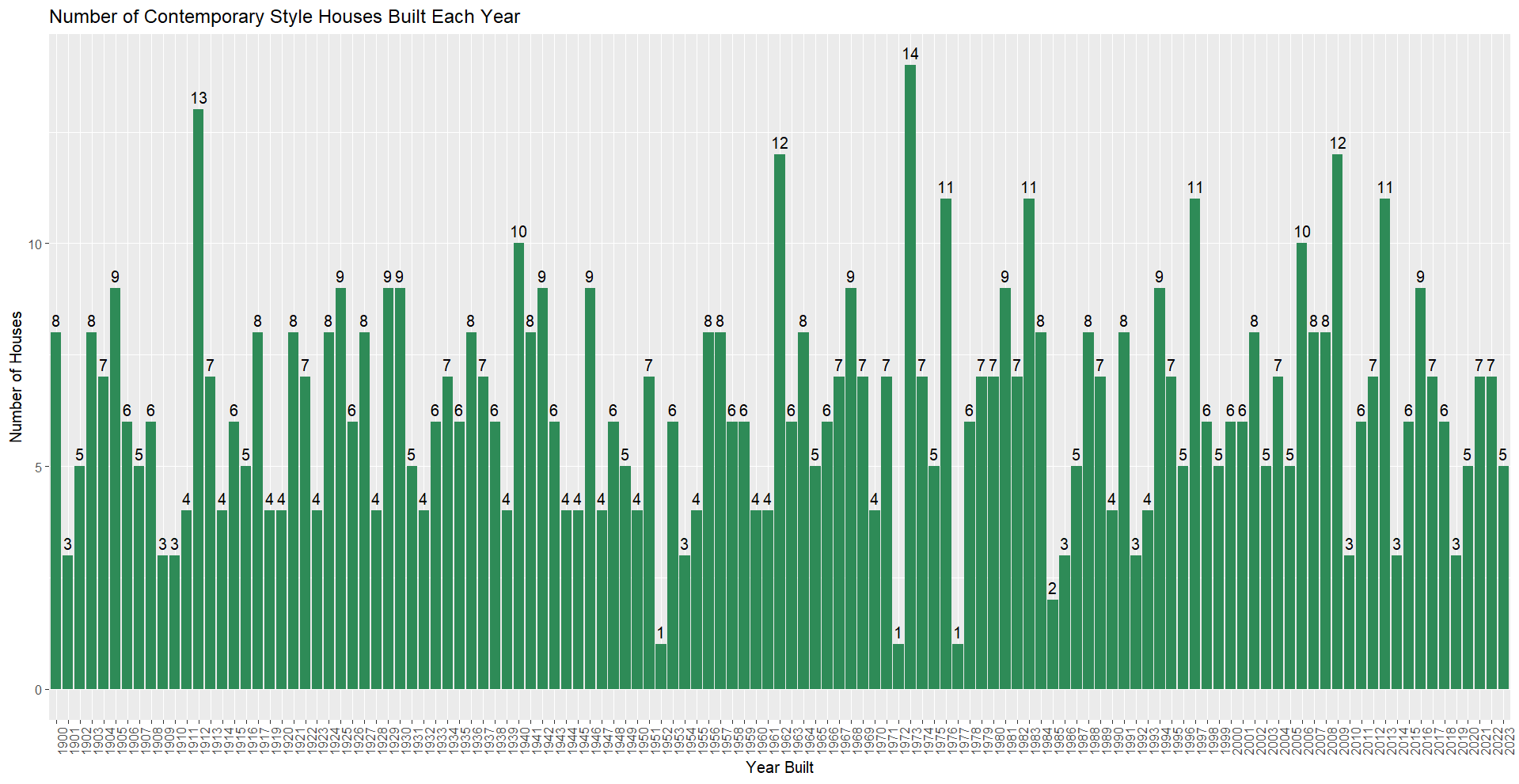
geom\_text(aes(label = count), vjust = -0.5, color = "black") +

labs(title = "Number of Contemporary Style Houses Built Each Year",

x = "Year Built",

y = "Number of Houses") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))



##In the year 1972 contemporary style houses are build most.(14 houses)

# Filter the data set for houses with the "Craftsman" style

Craftsman\_houses <- house\_sales[house\_sales$house\_style == "Craftsman", ]

Craftsman\_houses

# Count the number of Craftsman houses sold in each year

Craftsman\_houses\_counts <- table(Craftsman\_houses$year\_built)

Craftsman\_houses\_counts

# Convert year\_counts to a data frame

Craftsman\_houses\_counts\_df<- data.frame(year = as.numeric(names(Craftsman\_houses\_counts)),

count = as.numeric(Craftsman\_houses\_counts))

Craftsman\_houses\_counts\_df

# Create a bar graph of the number of Craftsman style houses built in each year

ggplot(Craftsman\_houses\_counts\_df, aes(x = factor(year), y = count)) +

geom\_bar(stat = "identity", fill = "hotpink3") +

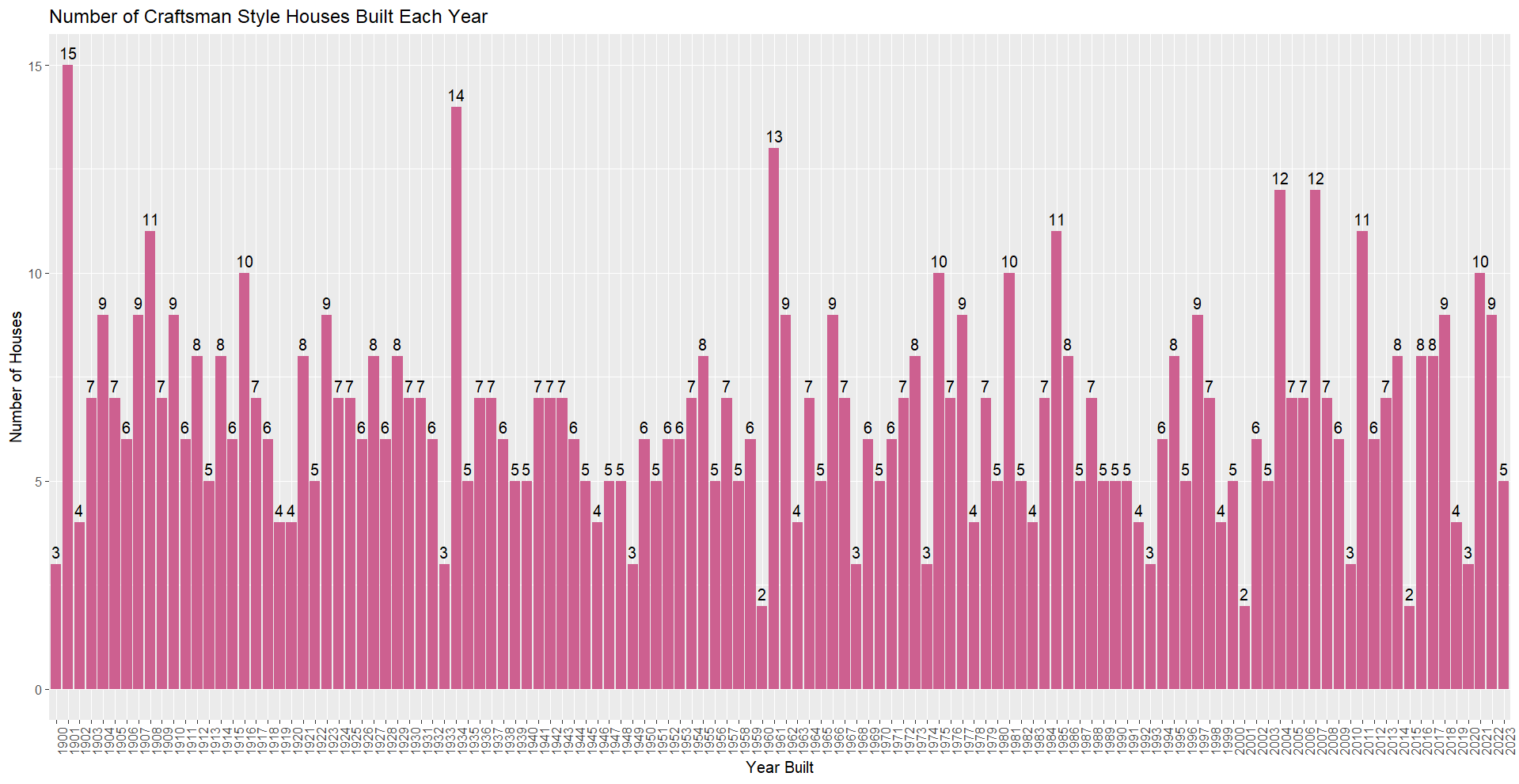
geom\_text(aes(label = count), vjust = -0.5, color = "black") +

labs(title = "Number of Craftsman Style Houses Built Each Year",

x = "Year Built",

y = "Number of Houses") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))



#In the year 1901 craftsman style houses are build most.(15 houses).

#it is the hightest .

# Filter the dataset for houses with the "Mediterranean" style

Mediterranean\_houses <- house\_sales[house\_sales$house\_style == "Mediterranean", ]

Mediterranean\_houses

# Count the number of houses sold in each year

Mediterranean\_house\_counts <- table(Mediterranean\_houses$year\_built)

Mediterranean\_house\_counts

# Convert year\_counts to a data frame

Mediterranean\_houses\_counts\_df <- data.frame(year = as.numeric(names(Mediterranean\_house\_counts)),

count = as.numeric(Mediterranean\_house\_counts))

Mediterranean\_houses\_counts\_df

# Create a bar graph of the number of Mediterranean style houses built in each year

ggplot(Mediterranean\_houses\_counts\_df, aes(x = factor(year), y = count)) +

geom\_bar(stat = "identity", fill = "yellow3") +

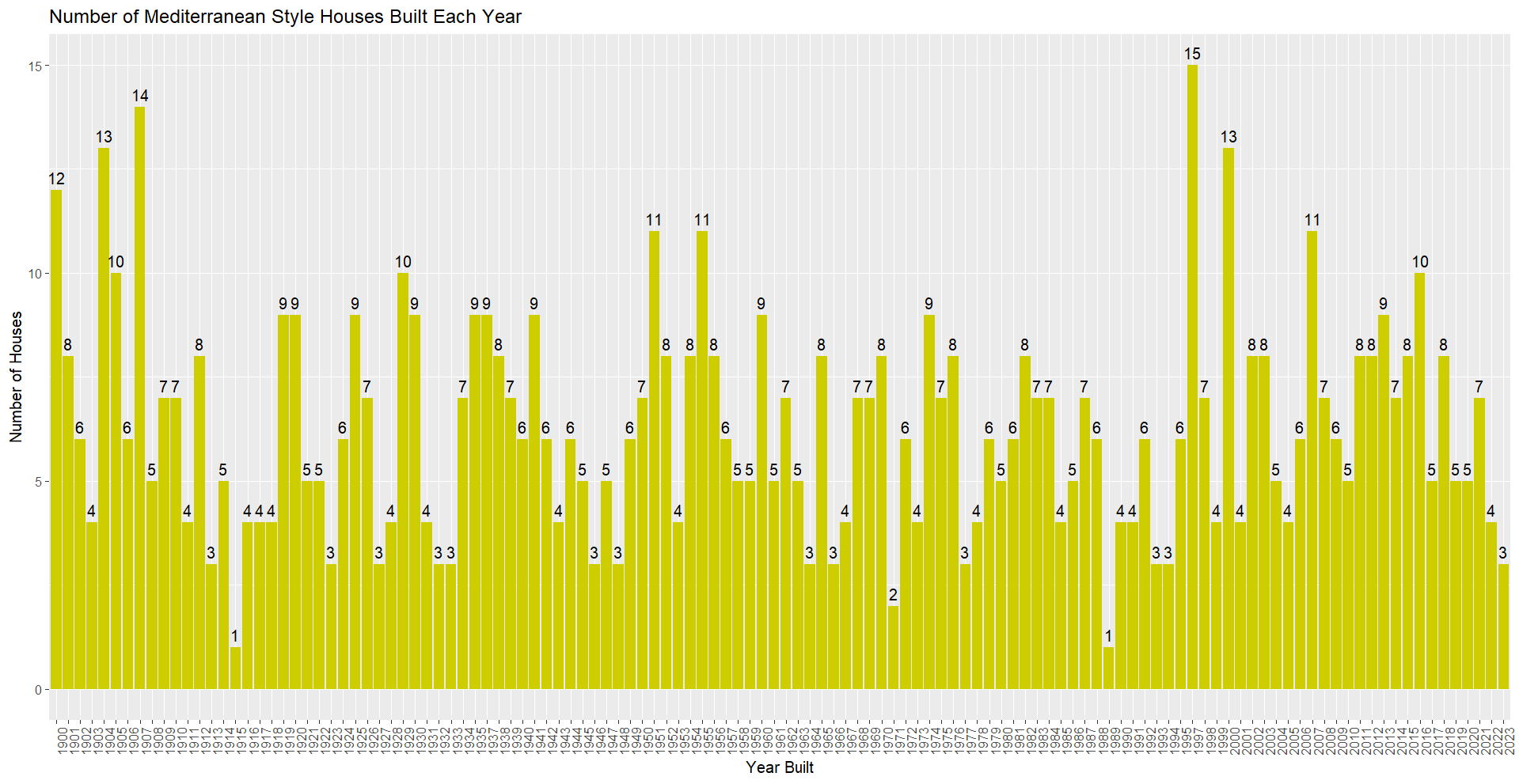
geom\_text(aes(label = count), vjust = -0.5, color = "black") +

labs(title = "Number of Mediterranean Style Houses Built Each Year",

x = "Year Built",

y = "Number of Houses") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))



#In the year 1997 Mediterranean style houses are build most.

#histogram

# Filter the dataset for houses with the "Ranch" style

ranch\_houses <- house\_sales[house\_sales$house\_style == "Ranch", ]

ranch\_houses

# Count the number of houses sold in each year

Ranch\_year\_counts <- table(ranch\_houses$year\_built)

Ranch\_year\_counts

# Convert year\_counts to a data frame

Ranch\_houses\_counts\_df <- data.frame(year = as.numeric(names(Ranch\_year\_counts)),

count = as.numeric(Ranch\_year\_counts))

Ranch\_houses\_counts\_df

# Create a bar graph of the number of Ranch style houses built in each year

ggplot(Ranch\_houses\_counts\_df, aes(x = factor(year), y = count)) +

geom\_bar(stat = "identity", fill = "steelblue") +

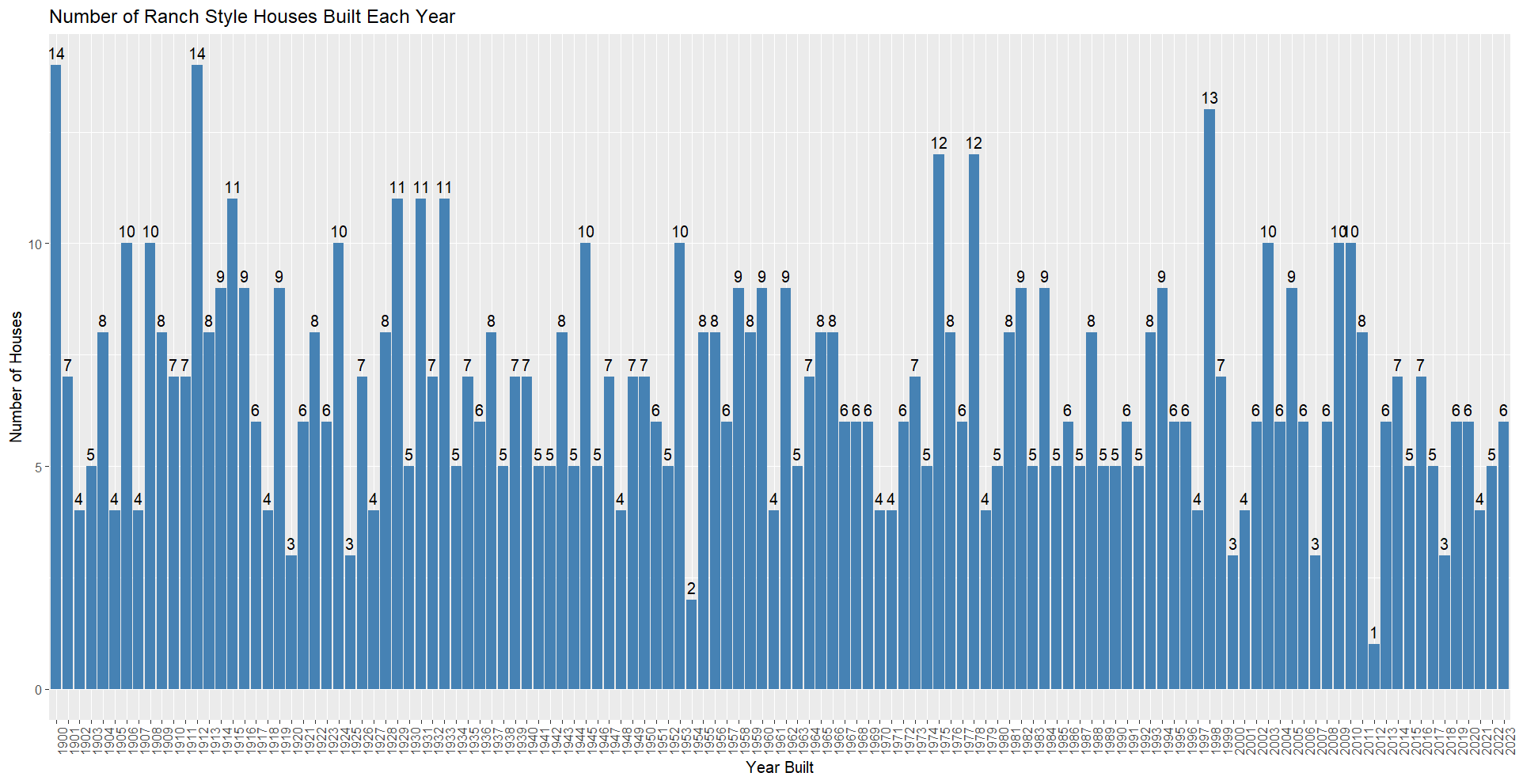
geom\_text(aes(label = count), vjust = -0.5, color = "black") +

labs(title = "Number of Ranch Style Houses Built Each Year",

x = "Year Built",

y = "Number of Houses") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))



#In the year 1900,1912 Ranch style houses are build most.

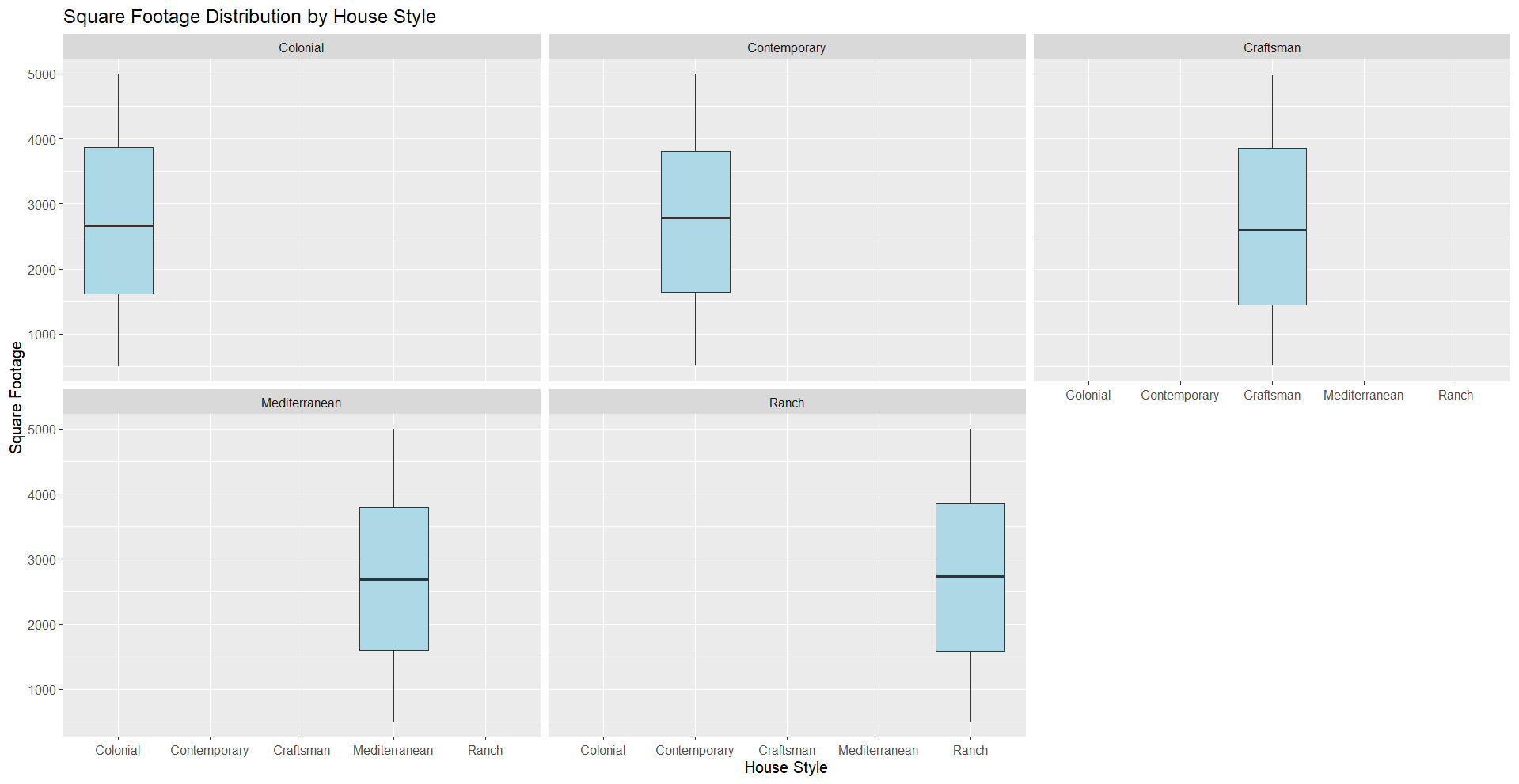
# Create a boxplot to show the square footage distribution for each house style

ggplot(house\_sales, aes(x = house\_style, y = sqft)) +

geom\_boxplot(fill = "lightblue") +

labs(title = "Square Footage Distribution by House Style", x = "House Style", y = "Square Footage") +

facet\_wrap(~house\_style)

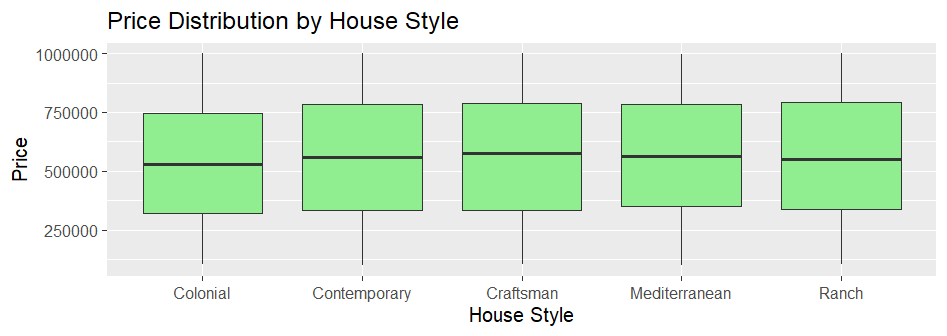


# Box plot for price by house style

ggplot(house\_sales, aes(x = house\_style, y = price)) +

geom\_boxplot(fill = "lightgreen") +

labs(title = "Price Distribution by House Style", x = "House Style", y = "Price")



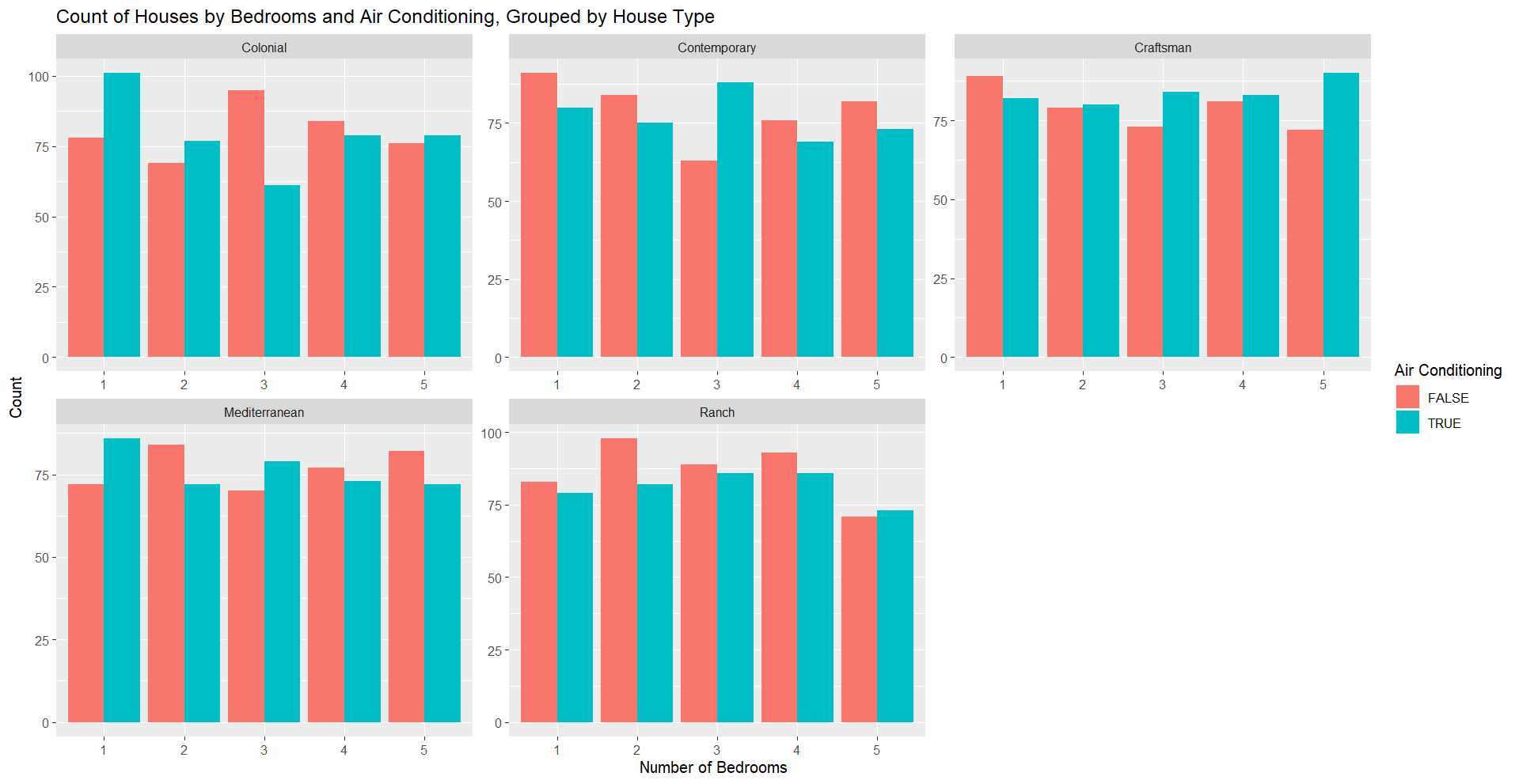
# Create a bar plot for count of houses by the number of bedrooms and presence of air conditioning, grouped by house type

ggplot(house\_sales, aes(x = factor(bedrooms), fill = factor(air\_condition))) +

geom\_bar(position = "dodge") +

facet\_wrap(~ house\_style, scales = "free") +

labs(title = "Count of Houses by Bedrooms and Air Conditioning, Grouped by House Type", x = "Number of Bedrooms", y = "Count", fill = "Air Conditioning")



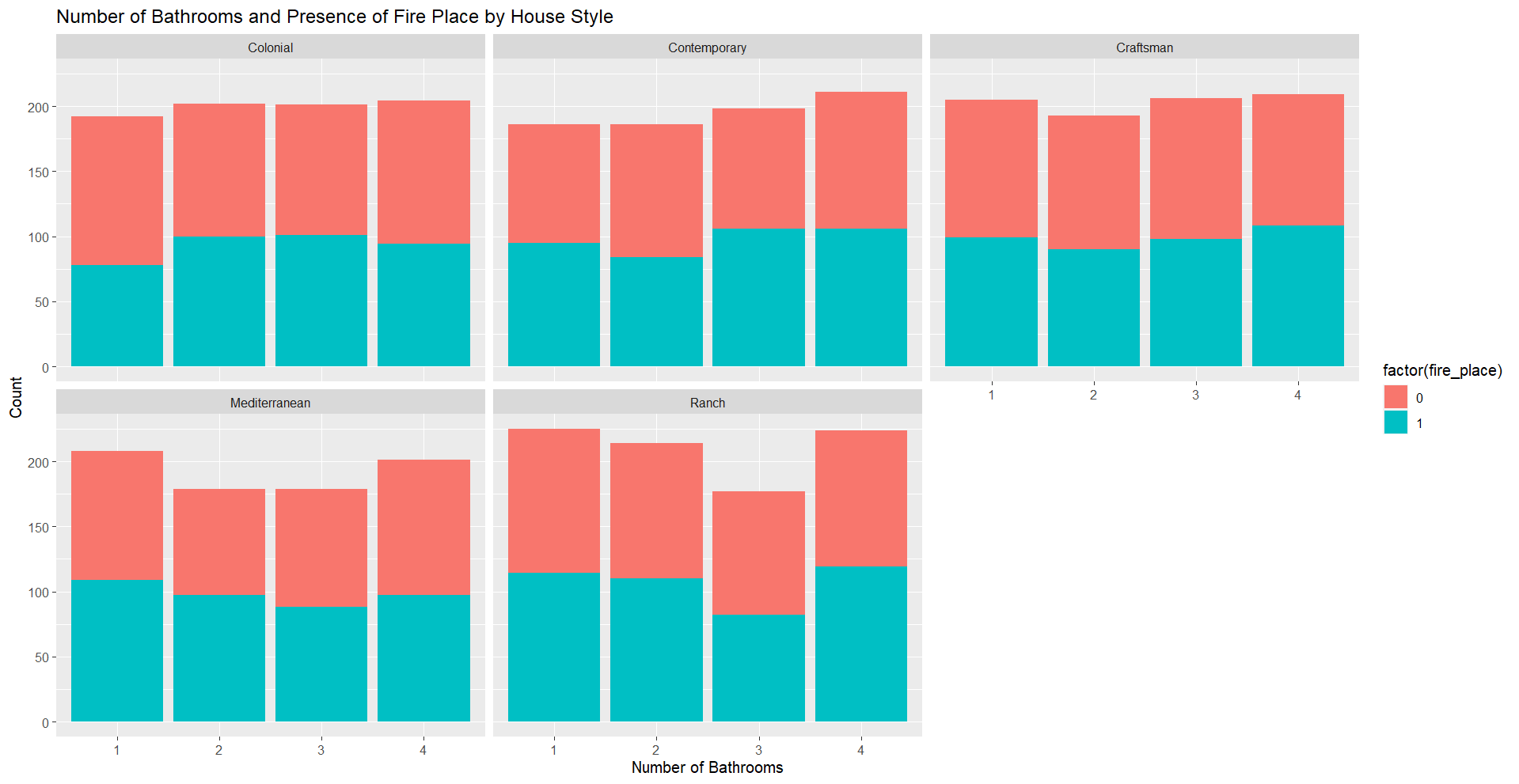
# Stacked bar plot for bathrooms and presence of fire place by house style

ggplot(house\_sales, aes(x = factor(bathrooms), fill = factor(fire\_place))) +

geom\_bar(position = "stack") +

facet\_wrap(~house\_style) +

labs(title = "Number of Bathrooms and Presence of Fire Place by House Style", x = "Number of Bathrooms", y = "Count")

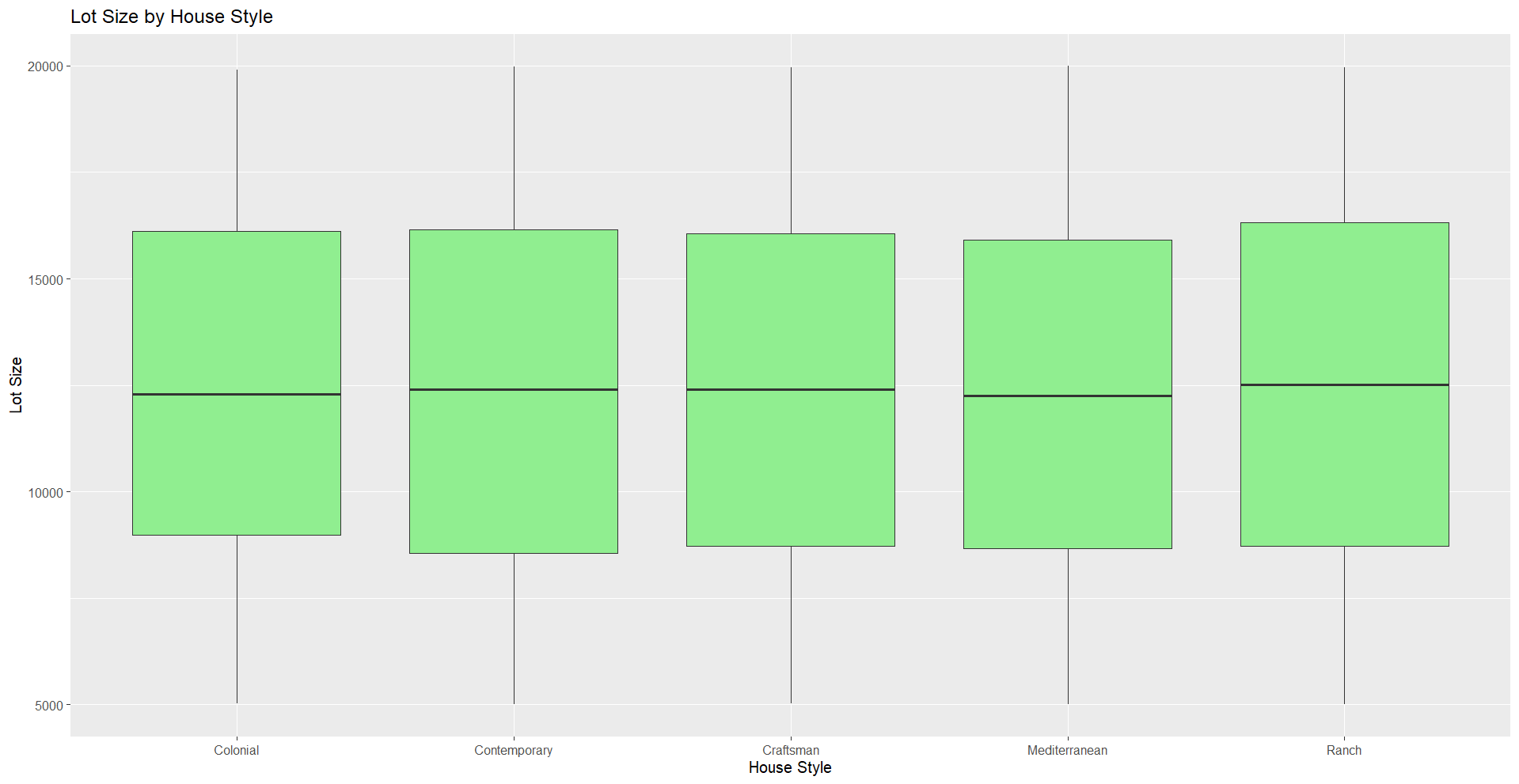


# Box plot for lot size by house style

ggplot(house\_sales, aes(x = house\_style, y = lot\_size)) +

geom\_boxplot(fill = "lightgreen") +

labs(title = "Lot Size by House Style", x = "House Style", y = "Lot Size")



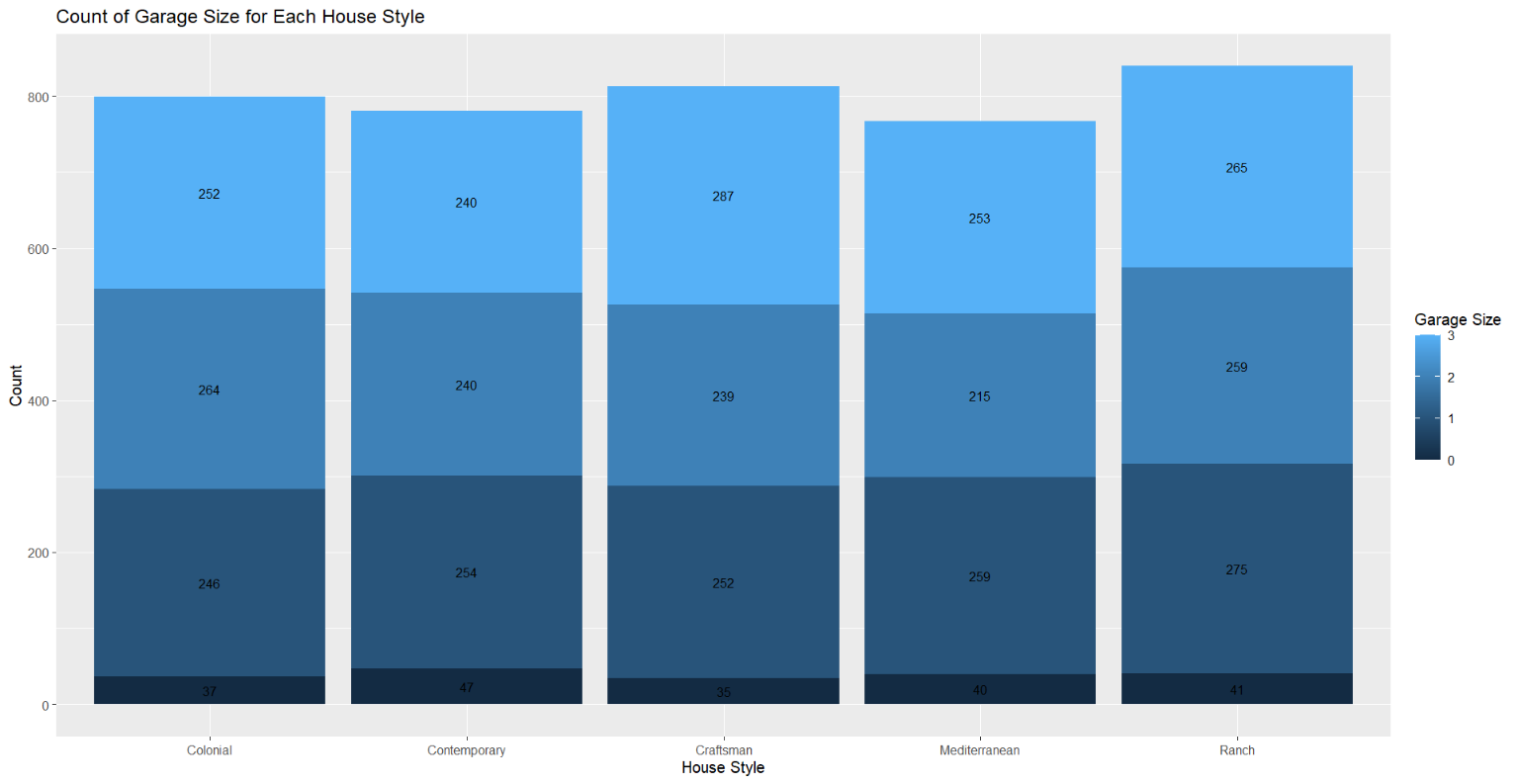
# Visualize the count of each garage size for each house style with count labels

ggplot(garage\_summary, aes(x = house\_style, y = count, fill = garage\_size, label = count)) +

geom\_bar(stat = "identity", position = "stack") +

geom\_text(position = position\_stack(vjust = 0.5), size = 3) +

labs(title = "Count of Garage Size for Each House Style", x = "House Style", y = "Count", fill = "Garage Size")



#from above graph we know that the houses which contain 0 garage sizes are contemporary house styles.47 houses doesn’t contain in garages in contemporary house\_style.

# the houses which contain maximum no.of 1 garage sizes are ranch house styles.275 houses contain 1 garages in ranch house style houses.

# the houses which contain maximum no.of 2 garage sizes are ranch house styles.275 houses contain 2 garages in colonial house style houses.

#craftsman stands in first place that contain maximum no.of 3 garages with count of 287.

**# Calculate the count of houses built in each zip code area**

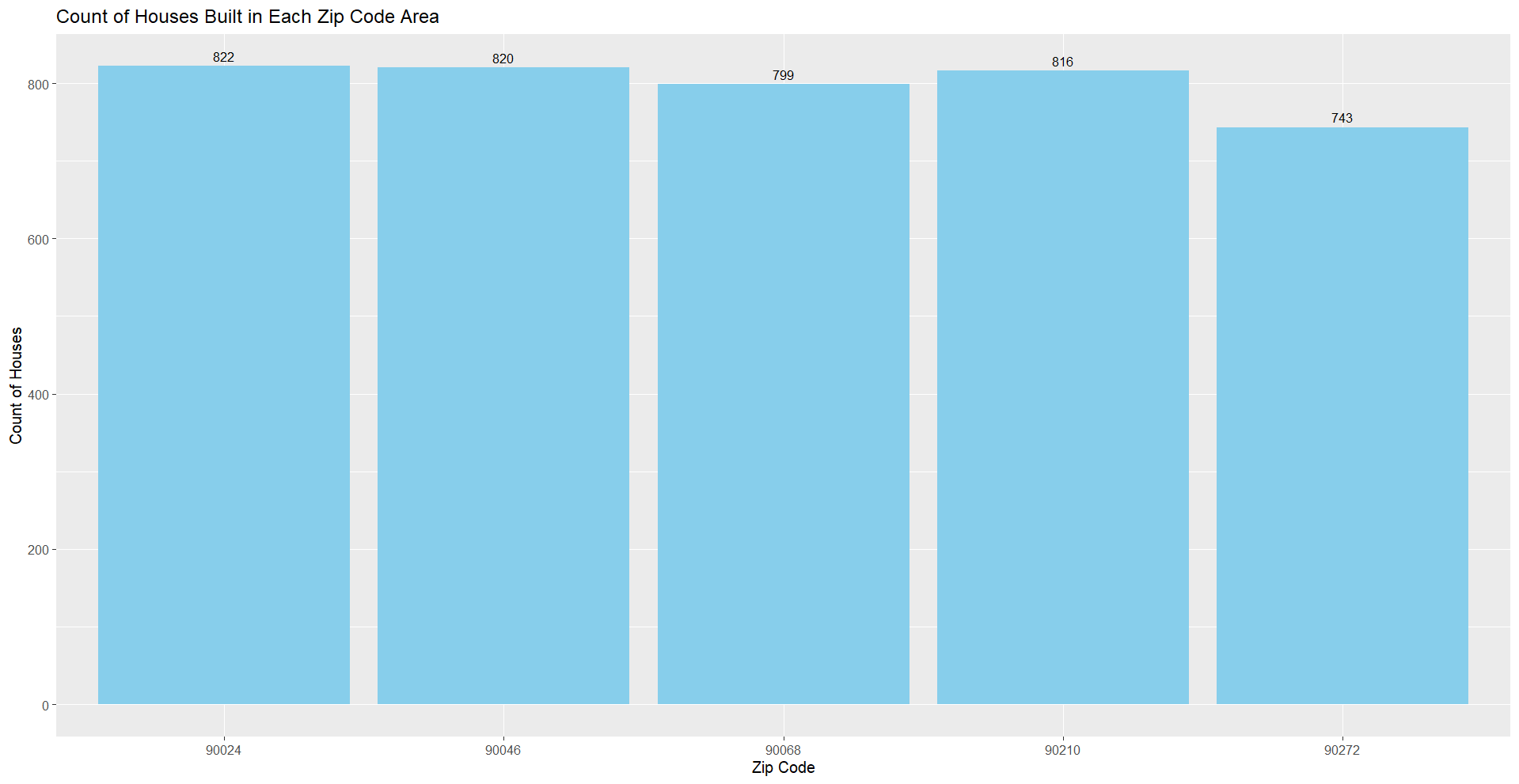
house\_count\_by\_zip <- house\_sales %>%

group\_by(zip\_code) %>%

summarise(count = n())

**# Bar plot for count of houses built in each zip code area with count labels**

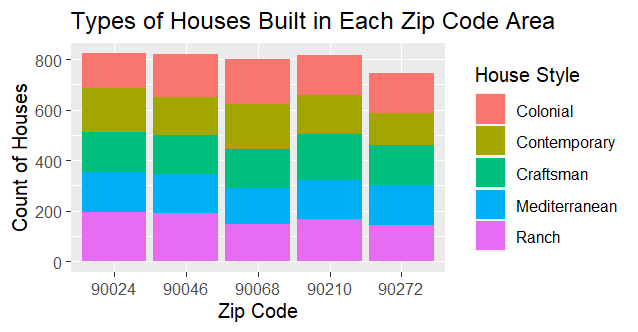
ggplot(house\_count\_by\_zip, aes(x = factor(zip\_code), y = count, label = count)) + geom\_bar(stat = "identity", fill = "skyblue") + geom\_text(vjust = -0.5, size = 3) + labs(title = "Count of Houses Built in Each Zip Code Area", x = "Zip Code", y = "Count of Houses")



#Maximum number of houses are build in the area with zip code 90024 with count of 822.

**# Create a bar plot to show different types of houses built in each zip code area**

ggplot(house\_sales, aes(x = factor(zip\_code), fill = house\_style)) + geom\_bar(position = "stack") + labs(title = "Types of Houses Built in Each Zip Code Area", x = "Zip Code", y = "Count of Houses", fill = "House Style")



# Calculate the count of each type of house built in each zip code area

house\_count\_by\_type <- house\_sales %>%

group\_by(zip\_code, house\_style) %>%

summarise(count = n())

# View the resulting table

house\_count\_by\_type

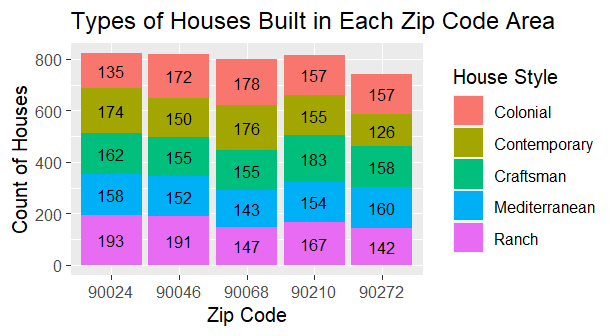
# Create a bar plot to show different types of houses built in each zip code area with count labels

ggplot(house\_count\_by\_type, aes(x = factor(zip\_code), y = count, fill = house\_style, label = count)) +

geom\_bar(stat = "identity", position = "stack") +

geom\_text(position = position\_stack(vjust = 0.5), size = 3) +

labs(title = "Types of Houses Built in Each Zip Code Area", x = "Zip Code", y = "Count of Houses", fill = "House Style")

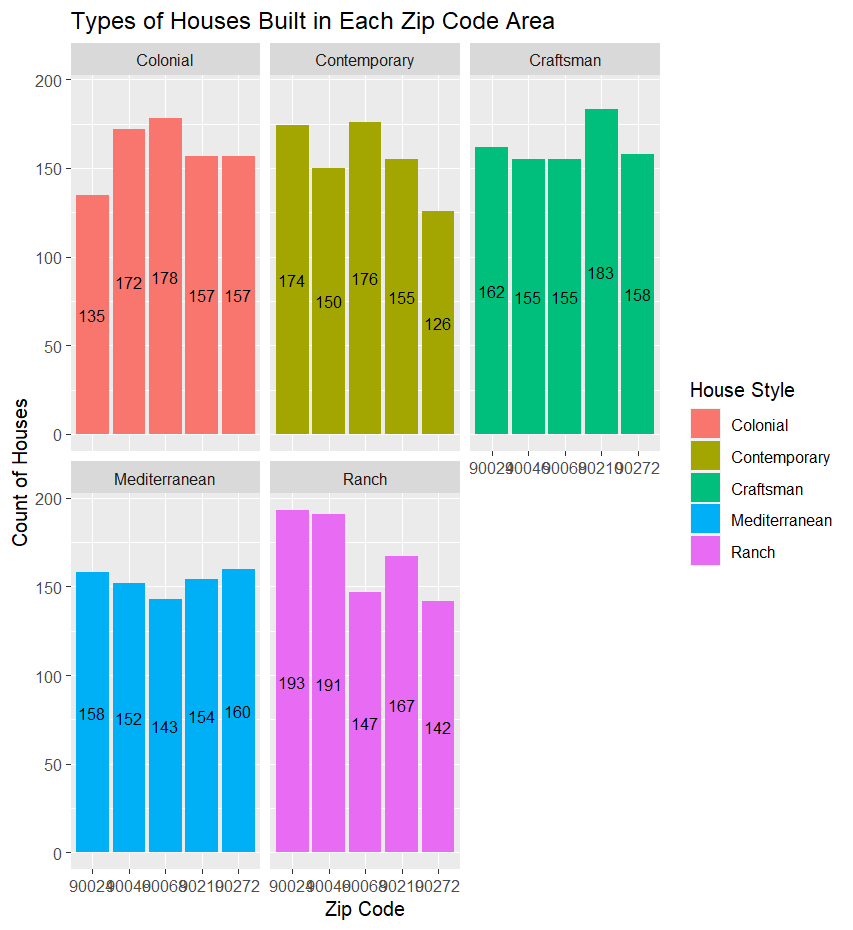


# in the area 90024, 193 Ranch houses are build with highest no.of houses than any other house styles (193 houses)

# Create a bar plot to show different types of houses built in each zip code area with count labels and price range represented by color, separated by house type

ggplot(house\_sales, aes(x = factor(zip\_code), fill = house\_style, color = price)) +geom\_bar(position = "stack") + geom\_text(aes(label = ..count..), stat = "count", position = position\_stack(vjust = 0.5), size = 3) +scale\_color\_gradient(low = "blue", high = "red") +labs(title = "Types of Houses Built in Each Zip Code Area", x = "Zip Code", y = "Count of Houses", fill = "House Style", color = "Price Range") +

facet\_wrap(~house\_style)



# from the above graph we know that the colonial house style mostly prefers area with zip\_code 90068.(178 houses)

# from the above graph we know that the contemporary house style mostly prefers area with zip\_code 90068. .(176 houses)

# from the above graph we know that the craftsman house style mostly prefers area with zip\_code 90210. .(173 houses)

# from the above graph we know that the contemporary house style mostly prefers area with zip\_code 90270. .(160 houses)

# from the above graph we know that the contemporary house style mostly prefers area with zip\_code 90024. .(193 houses)

# Calculate average price by house style

avg\_price\_by\_style <- aggregate(price ~ house\_style, data = house\_sales, FUN = mean)

print("Average price by house style:")

print(avg\_price\_by\_style)

# Filter for houses built within the last 10 years

recent\_houses <- house\_sales %>%

filter(year\_built >= (2023 - 10))

# Calculate the count, average bedrooms, bathrooms, and price by house style

house\_summary <- recent\_houses %>%

group\_by(house\_style) %>%

summarise(count = n(),

avg\_bedrooms = mean(bedrooms),

avg\_bathrooms = mean(bathrooms),

avg\_price = mean(price))

# View the summary

house\_summary

# Filter for houses built recently with 3 bedrooms, 3 bathrooms, and good price

selected\_houses <- house\_sales %>%

filter(year\_built >= (2023 - 10), # Assuming "recently" means within the last 10 years

bedrooms == 3,

bathrooms == 3,

price <= 500000) # Adjust the price criteria as needed

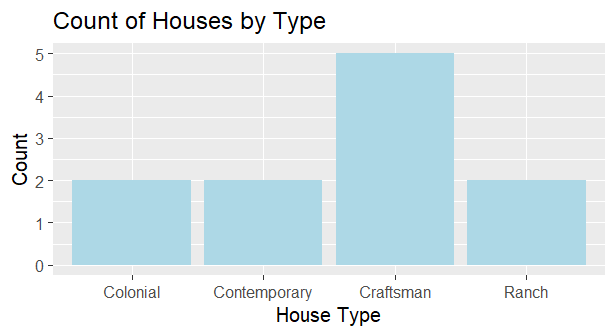
selected\_houses

# Create a bar plot for the count of each house type

ggplot(selected\_houses, aes(x = house\_style)) +

geom\_bar(fill = "lightblue") +

labs(title = "Count of Houses by Type", x = "House Type", y = "Count")



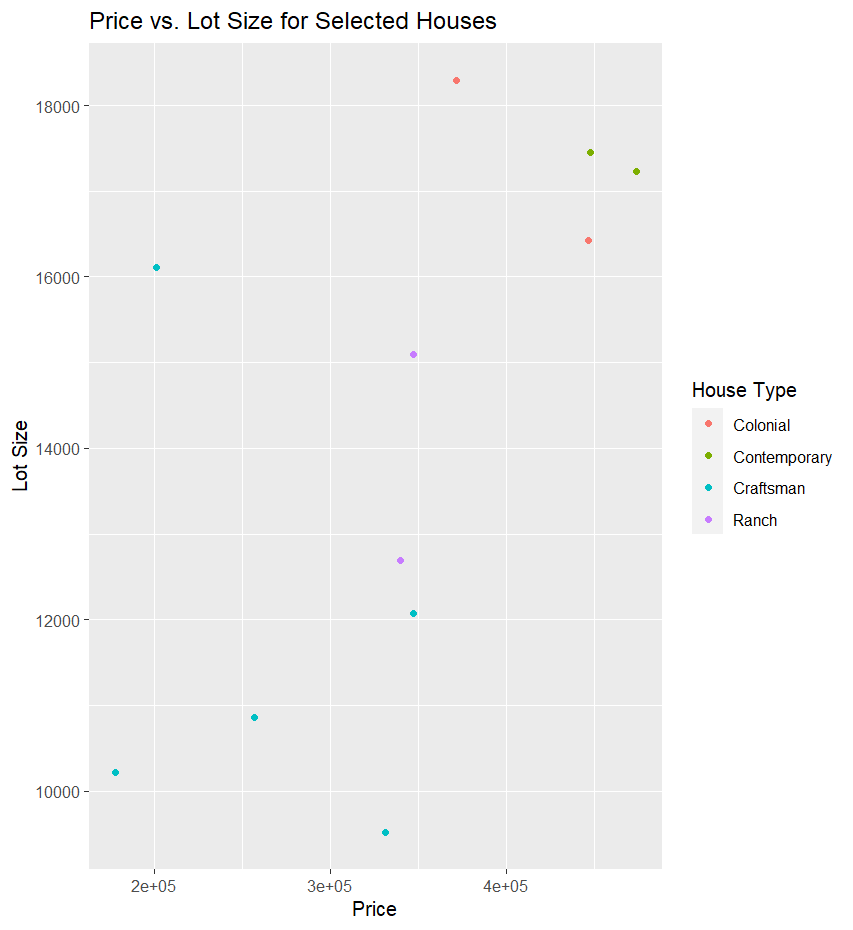
# for houses built recently with 3 bedrooms, 3 bathrooms, and good price is craftsman.

# Scatter plot for the relationship between price and selected houses

ggplot(selected\_houses, aes(x = price, y = lot\_size, color = house\_style)) +

geom\_point() +

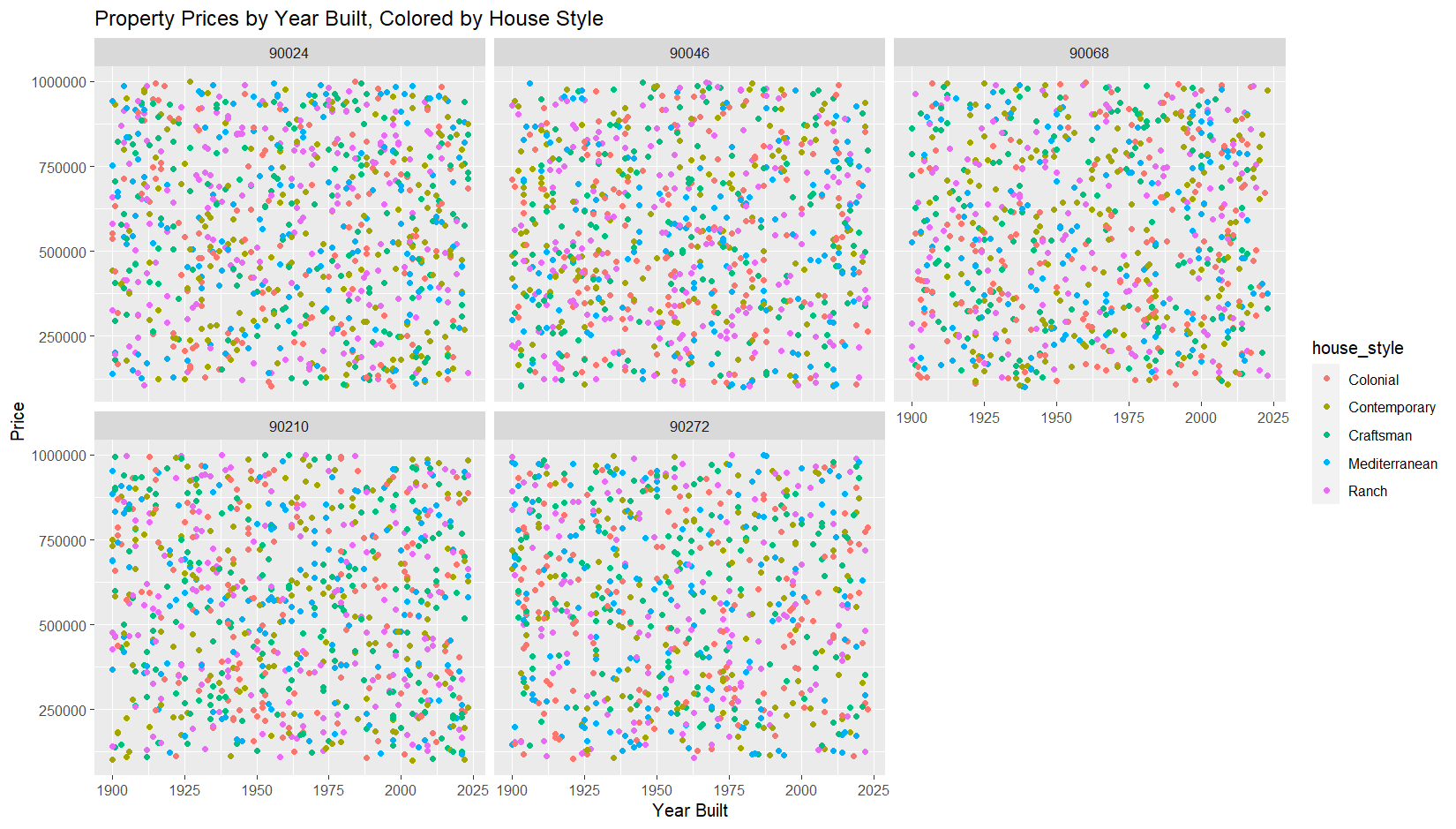
labs(title = "Price vs. Lot Size for Selected Houses", x = "Price", y = "Lot Size", color = "House Type")



#if priceincreases then lot size also increases.

# Scatter plot for property prices by year built, colored by house style and faceted by zip code

ggplot(house\_sales, aes(x = year\_built, y = price, color = house\_style)) + geom\_point() + facet\_wrap(~zip\_code) + labs(title = "Property Prices by Year Built, Colored by House Style", x = "Year Built", y = "Price")



# Create separate bar plots for count of properties by house style and year built for each zip code with correct year order

for (zip\_code in unique(house\_sales$zip\_code)) {

data\_subset <- house\_sales %>% filter(zip\_code == zip\_code)

data\_subset$year\_built <- factor(data\_subset$year\_built, levels = unique(data\_subset$year\_built))

plot\_title <- paste("Count of Properties in Zip Code", zip\_code)

p <- ggplot(data\_subset, aes(x = year\_built, fill = house\_style)) +

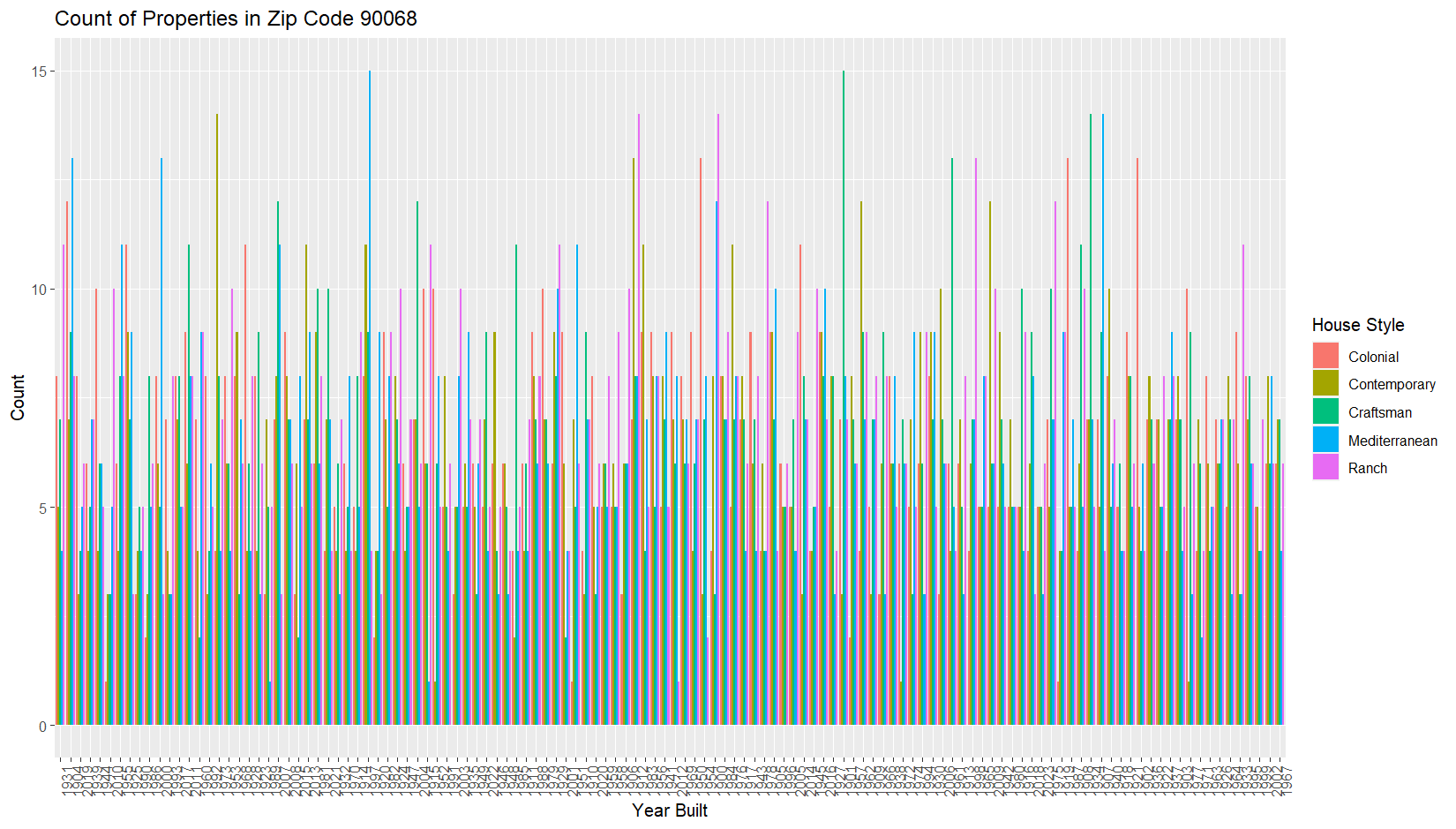
geom\_bar(position = "dodge") +

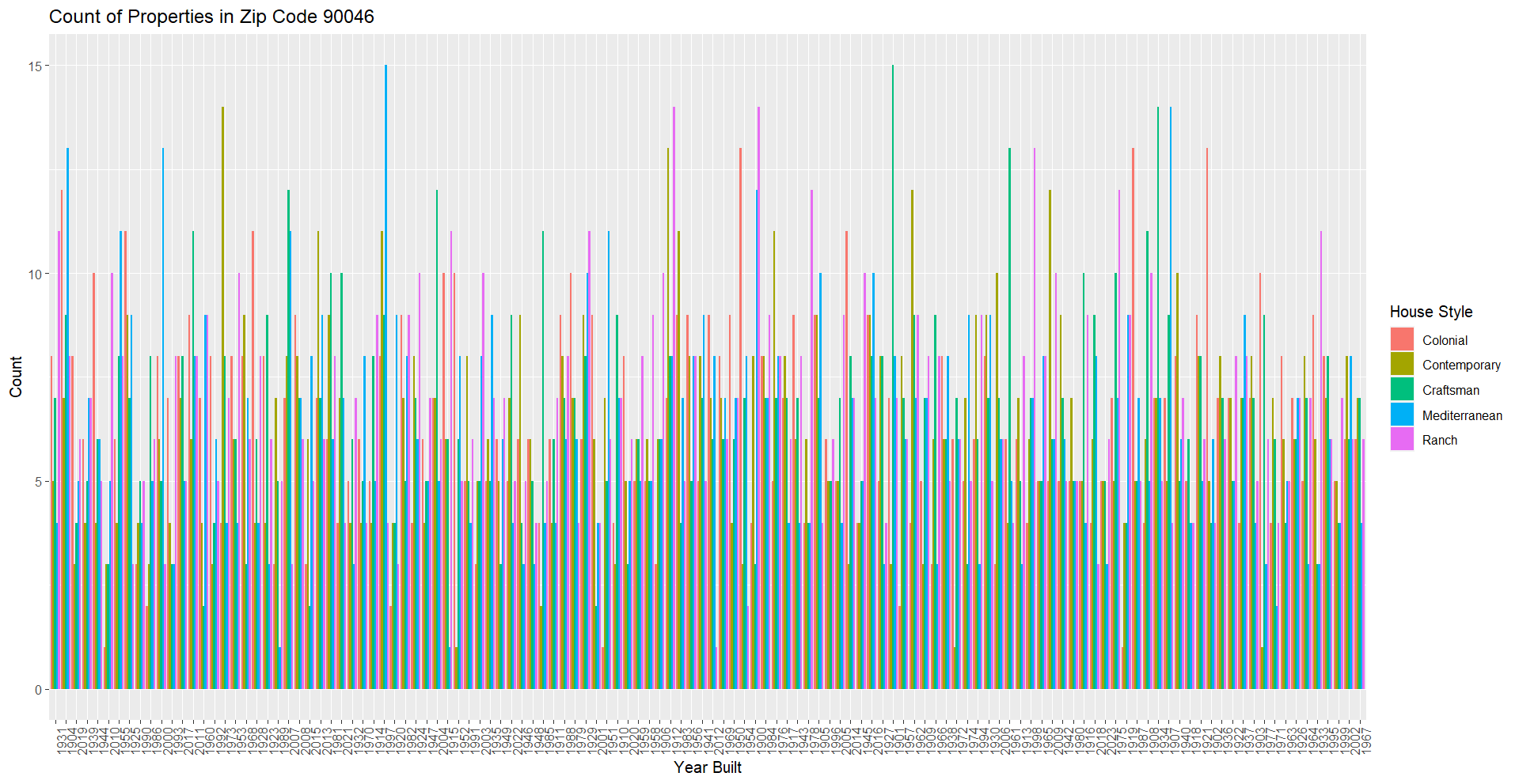
labs(title = plot\_title, x = "Year Built", y = "Count", fill = "House Style")

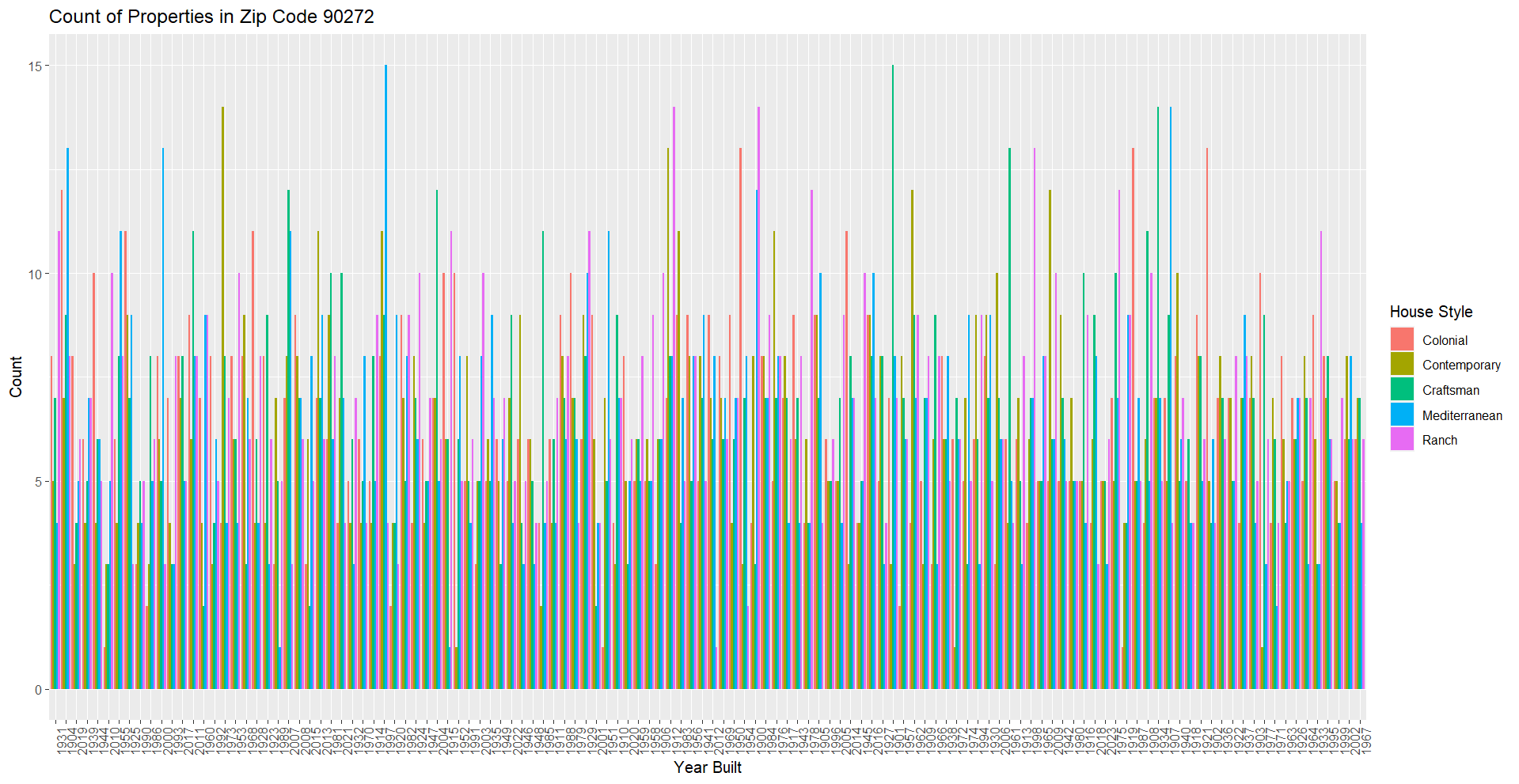
print(p) +

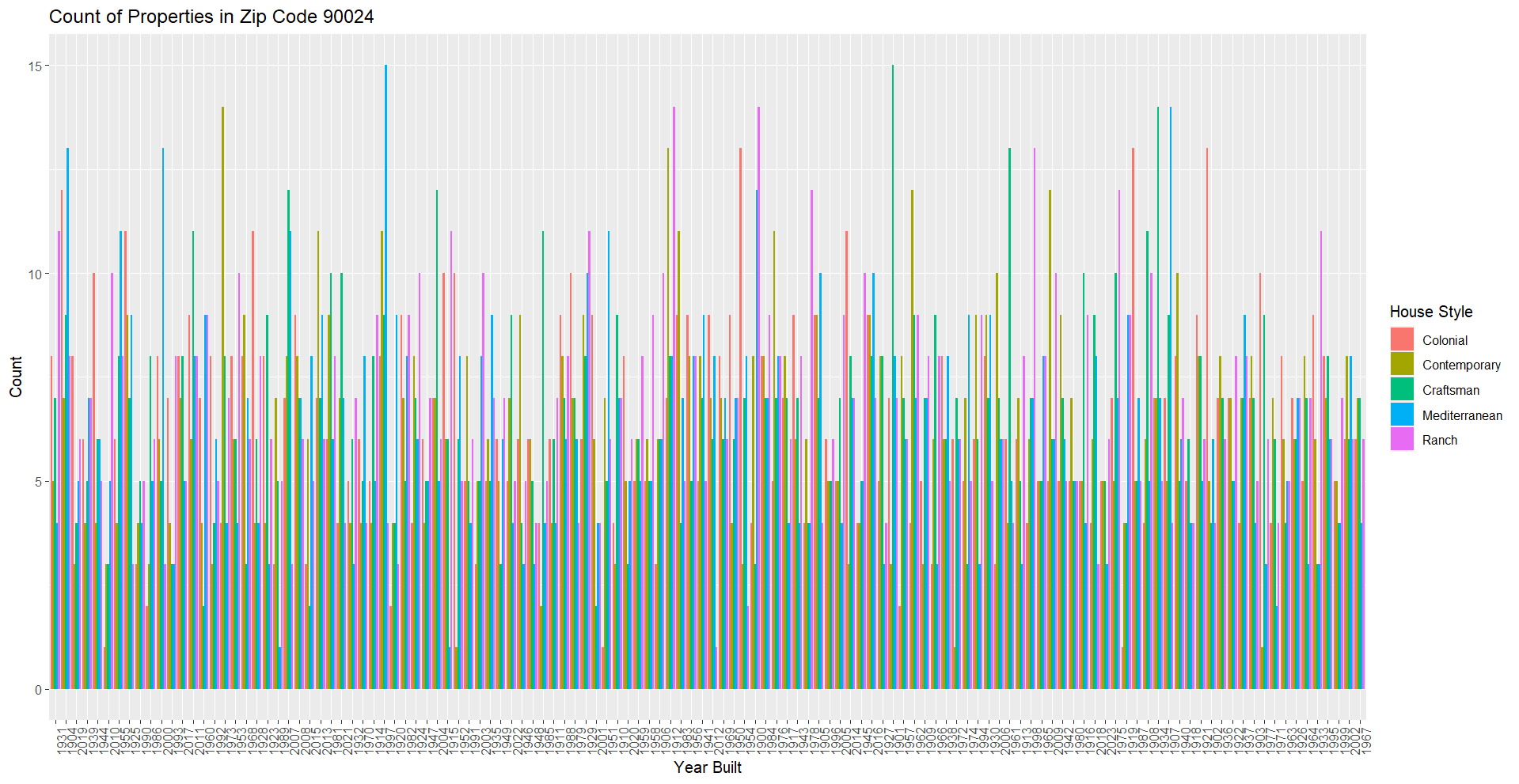
theme(axis.text.x = element\_text(angle = 90, hjust = 1))

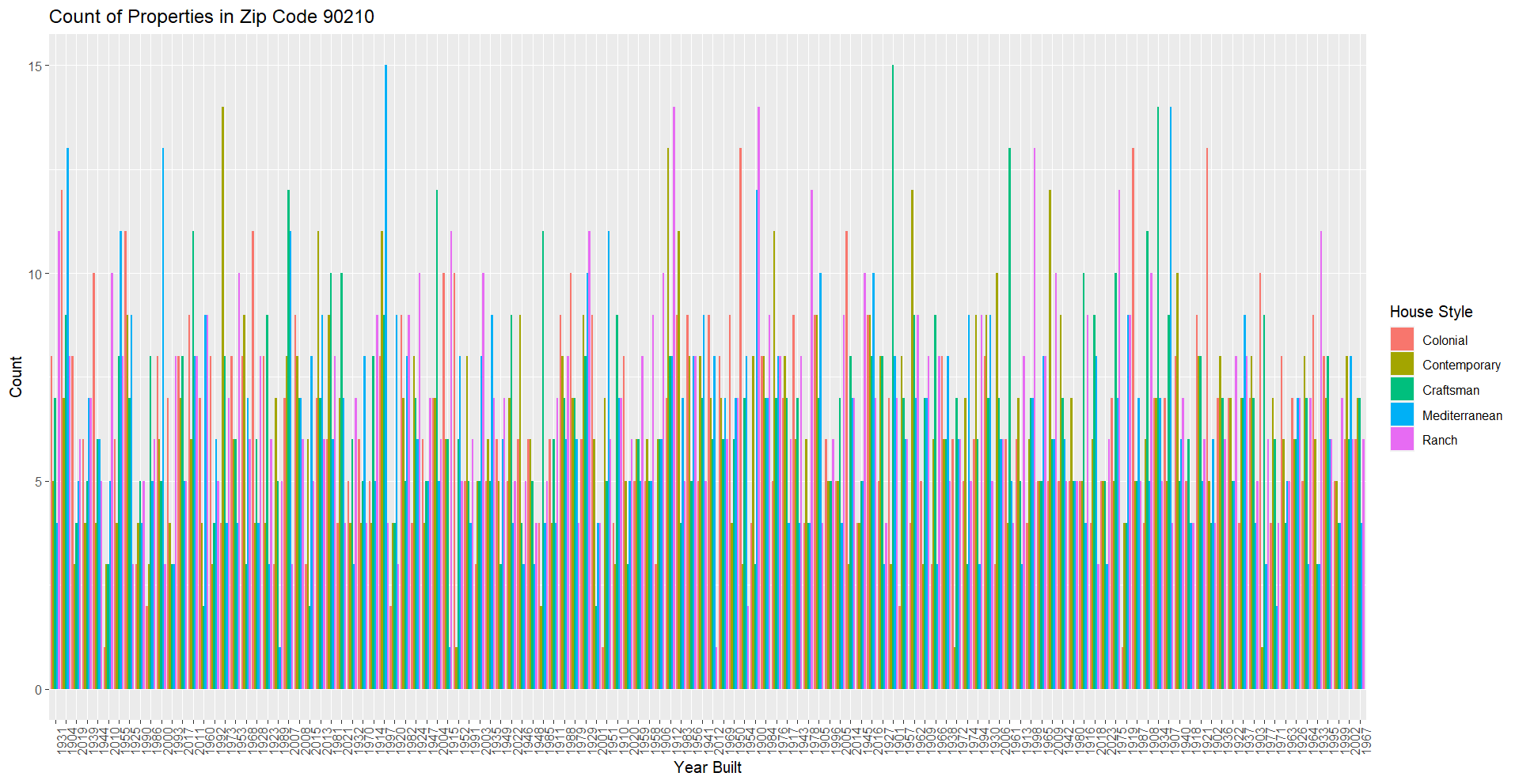
}











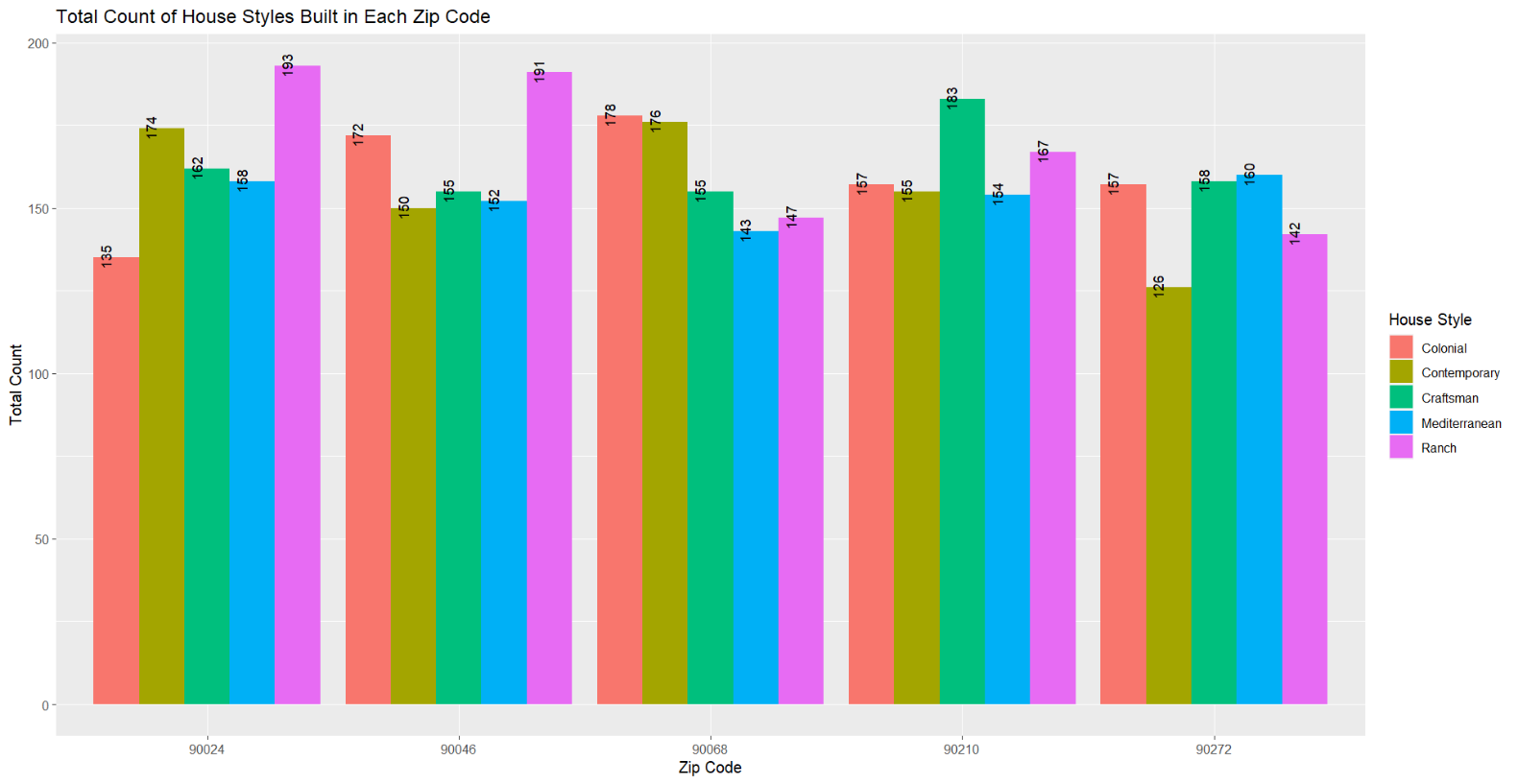
# Create a bar plot to display the total count of house styles built in each zip code with clear numeric values

ggplot(house\_styles\_count, aes(x = as.factor(zip\_code), y = total\_count, fill = house\_style, label = total\_count)) +

geom\_bar(stat = "identity", position = "dodge") +

geom\_text(position = position\_dodge(width = 0.9), vjust = -0.5, size = 3.5, angle = 90, hjust = 0.5) +

labs(title = "Total Count of House Styles Built in Each Zip Code", x = "Zip Code", y = "Total Count", fill = "House Style")



#conclusion :

As from my analysis I think craftsman is the most likely to be sold as these houses are built recently (time span of 10 years ) with 3 bedrooms, 3 bathrooms, and good price.

.