Murugan_HW2

2024-09-03

Problem 1a:

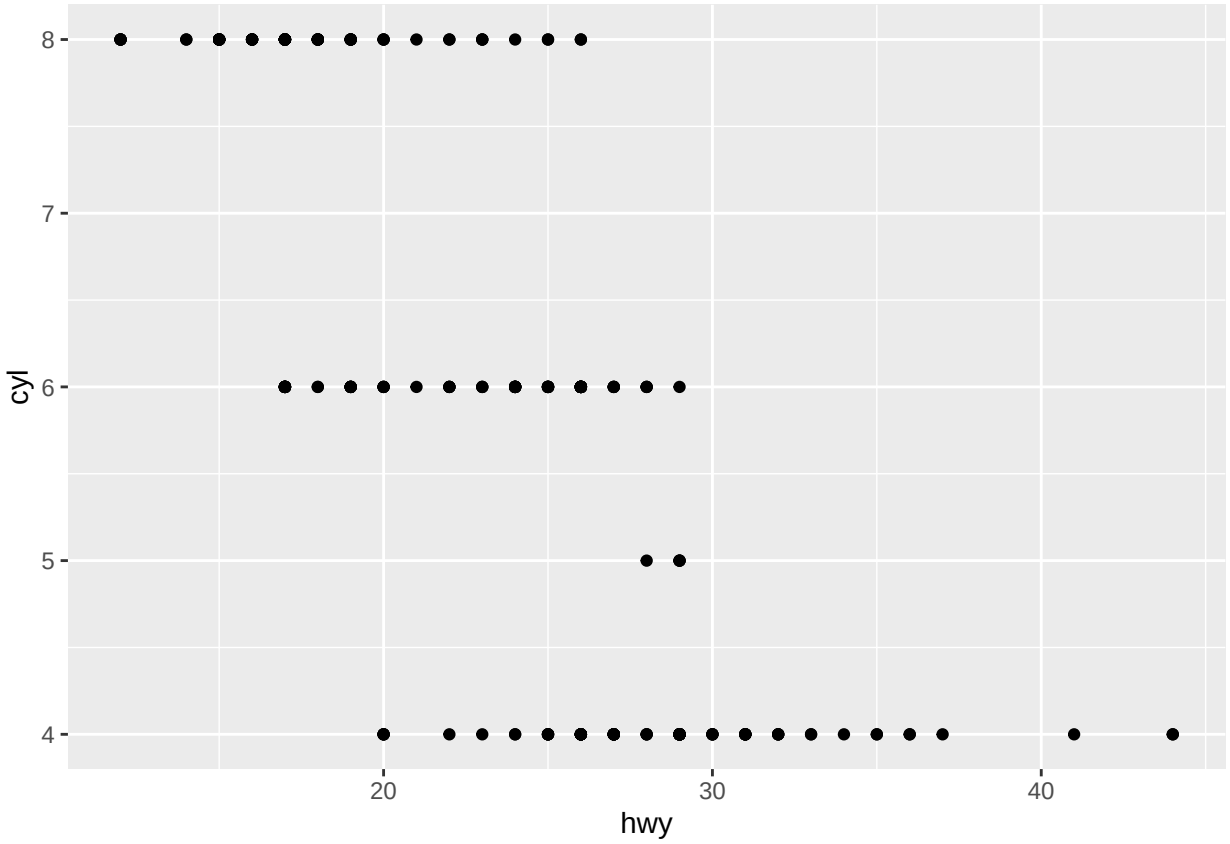
Exercise 3.2.4

Question 4 : Make a scatterplot of hwy vs cyl.

Scatter Plot of Highway Mileage vs. Cylinder Count

This scatter plot helps in understanding the relationship between highway mileage (hwy) and the number of cylinders (cyl).

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x=hwy, y=cyl))
```



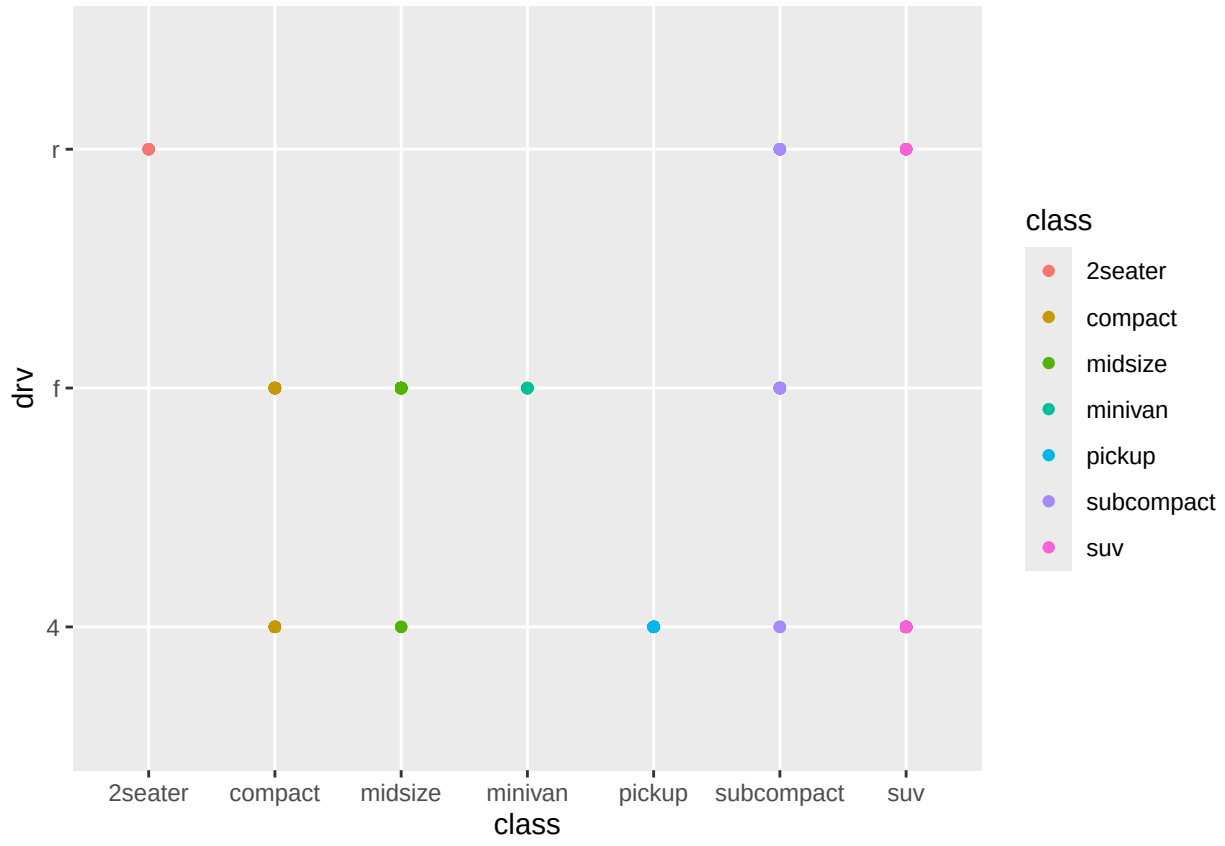
Exercise 3.2.4

Question 5 : What happens if you make a scatterplot of class vs drv? Why is the plot not useful?

Scatter Plot of Car Class vs. Drive Type

When you create a scatterplot of class versus drv in the mpg dataset, the plot is not useful because both class and drv are categorical variables, not continuous. In a scatterplot, categorical variables result in points that stack on top of each other along the axes, leading to overplotting where multiple data points occupy the same positions. This clustering at discrete points fails to convey meaningful relationships or patterns between the variables, making it difficult to interpret or gain insights from the plot.

```
ggplot(mpg)+
  geom_point(mapping = aes(x=class, y=drv, colour = class))
```

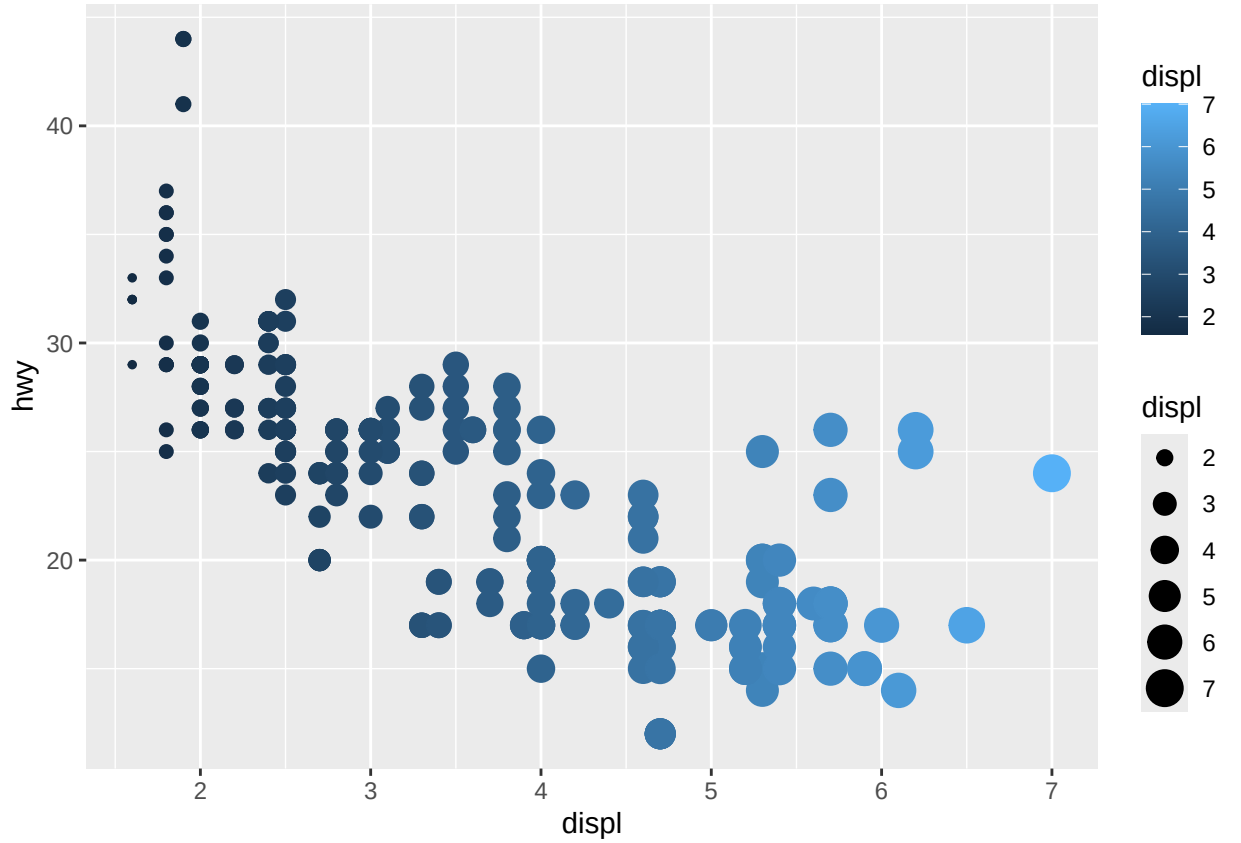


Exercise 3.3.1

Question 3 : Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical vs. continuous variables?

When mapping a continuous variable to color or size, color forms a gradient and size varies proportionally, effectively showing the range of values. Continuous variables can't be mapped to shape because shapes are finite and categorical. For categorical variables, distinct colors and shapes are used, with each category getting a different one, allowing clear differentiation without implying magnitude or order. This makes color and size ideal for continuous data, while shapes and distinct colors work best for categorical data.

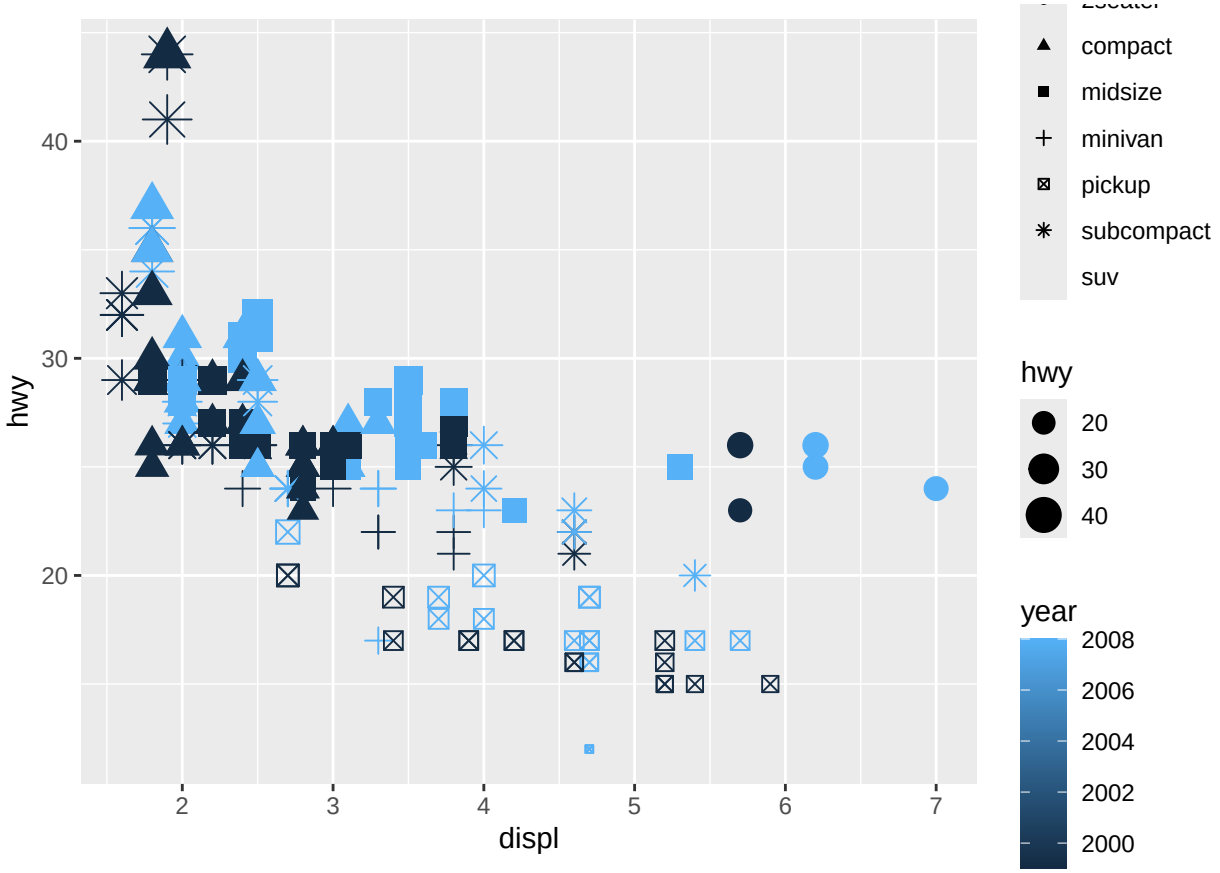
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = displ, size = displ))
```




```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = year,
                           shape = class, size = hwy))
```

```
## Warning: The shape palette can deal with a maximum of 6 discrete values because more
## than 6 becomes difficult to discriminate
## i you have requested 7 values. Consider specifying shapes manually if you need
##   that many have them.
```

```
## Warning: Removed 62 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Exercise 3.3.1

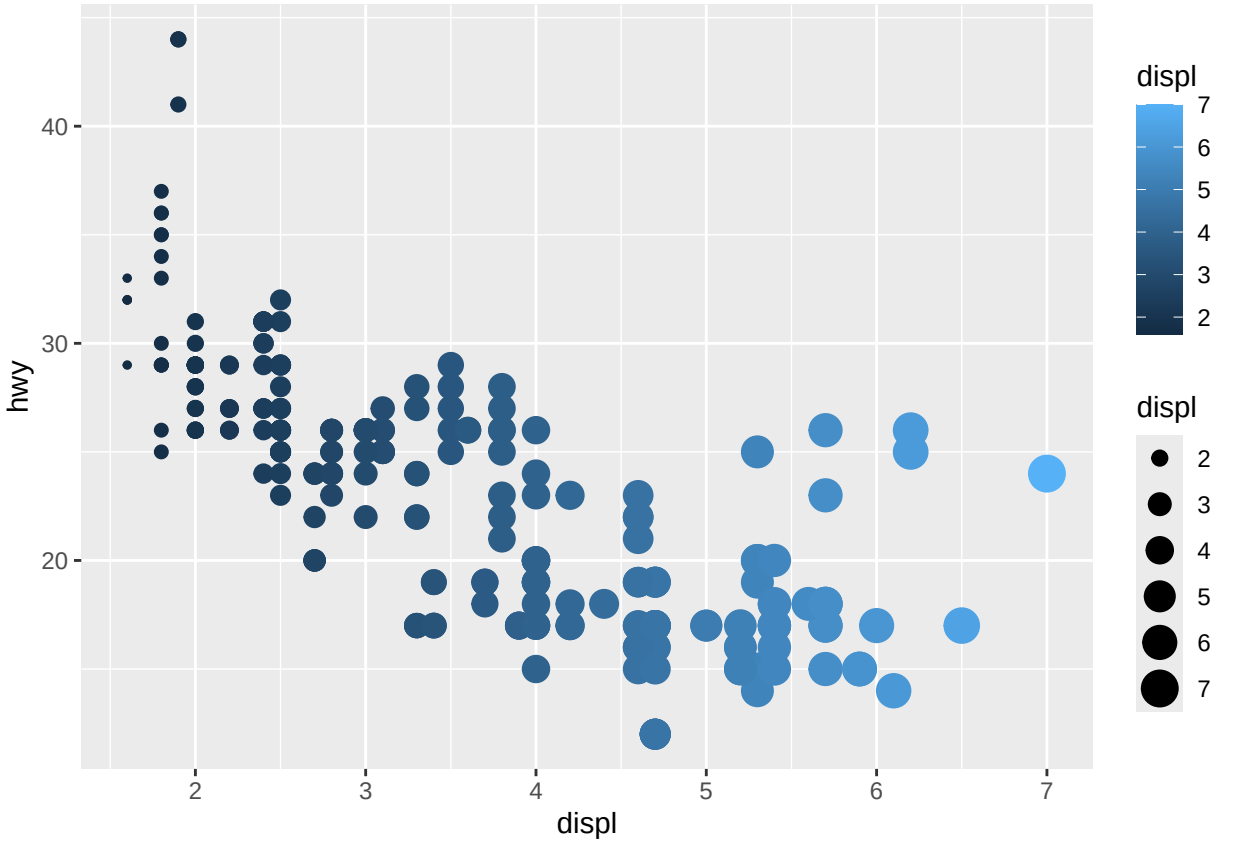
Question 4 : What happens if you map the same variable to multiple aesthetics?

Mapping the same variable ('displ') to both color and size.

This can increase emphasis on engine displacement, making its impact more visible in the plot.

However, care must be taken to avoid visual clutter and ensure the plot remains interpretable.

```
ggplot(mpg, aes(x=displ, y=hwy)) +
  geom_point(aes(color=displ, size=displ))
```

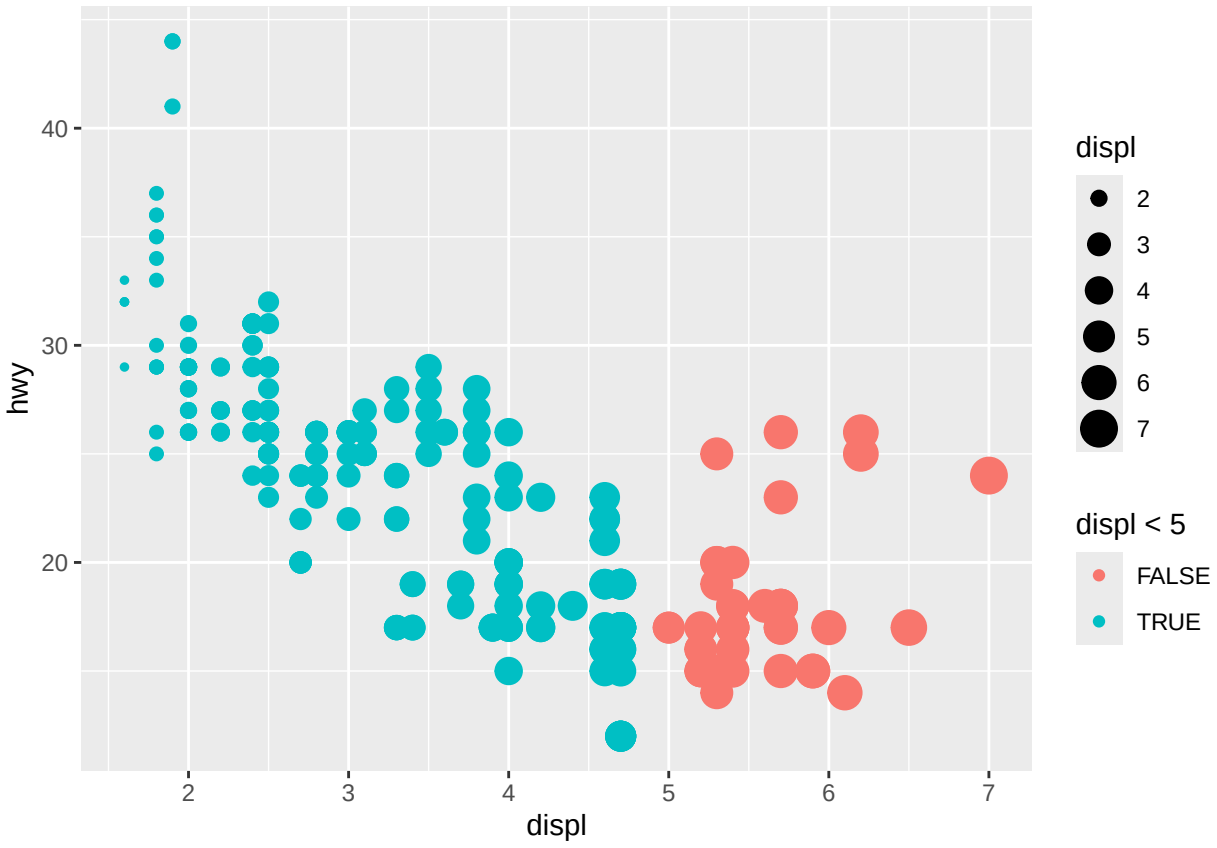


Exercise 3.3.1

Question 6 : What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5)? Note, you'll also need to specify x and y.

If you were to map disp < 5 , then the color coding will be true or False.

```
ggplot(mpg, aes(x=displ, y=hwy)) +
  geom_point(aes(color=displ<5, size=displ))
```



Exercise 3.5.1

Question 4 :

Take the first faceted plot in this section:

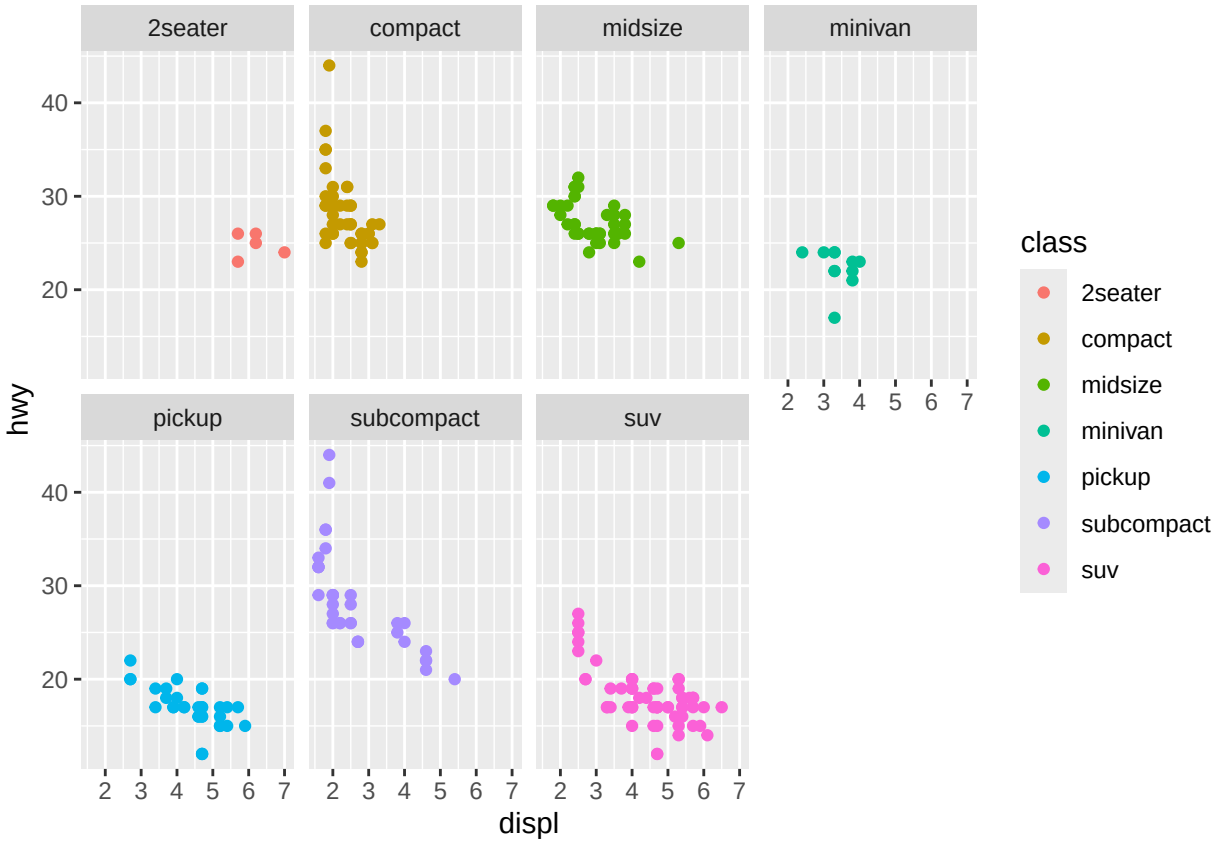
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy)) + facet_wrap(~ class, nrow = 2)

What are the advantages to using faceting instead of the colour aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset?

Ans :

Faceting provides clear separation of categories by creating individual plots for each category, which reduces overplotting and makes patterns easier to identify. This approach is particularly beneficial in larger datasets, as it prevents clutter that could occur if all points were differentiated only by color on a single plot. However, faceting can be space-intensive, especially with many categories, leading to smaller, harder-to-read plots. It also makes direct comparison across categories less straightforward compared to using color within a single plot.

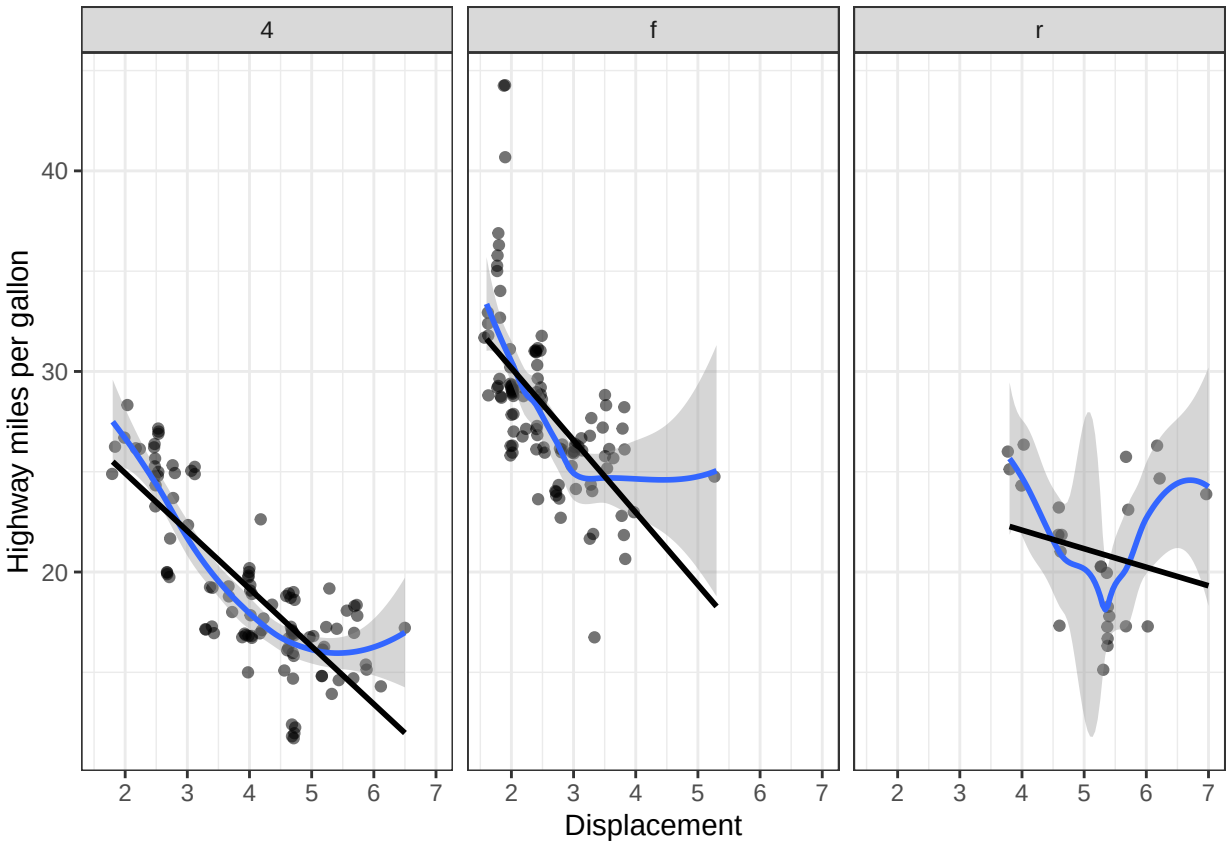
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, colour = class)) +
  facet_wrap(~ class, nrow = 2)
```



Problem 1b:

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_jitter(mapping = aes(alpha = 0.1)) +
  facet_wrap(~ drv) +
  geom_smooth() +
  geom_smooth(method = "lm", color = "black", se = FALSE) +
  theme_bw() +
  labs(x = "Displacement", y = "Highway miles per gallon") +
  theme(legend.position = "none")
```

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



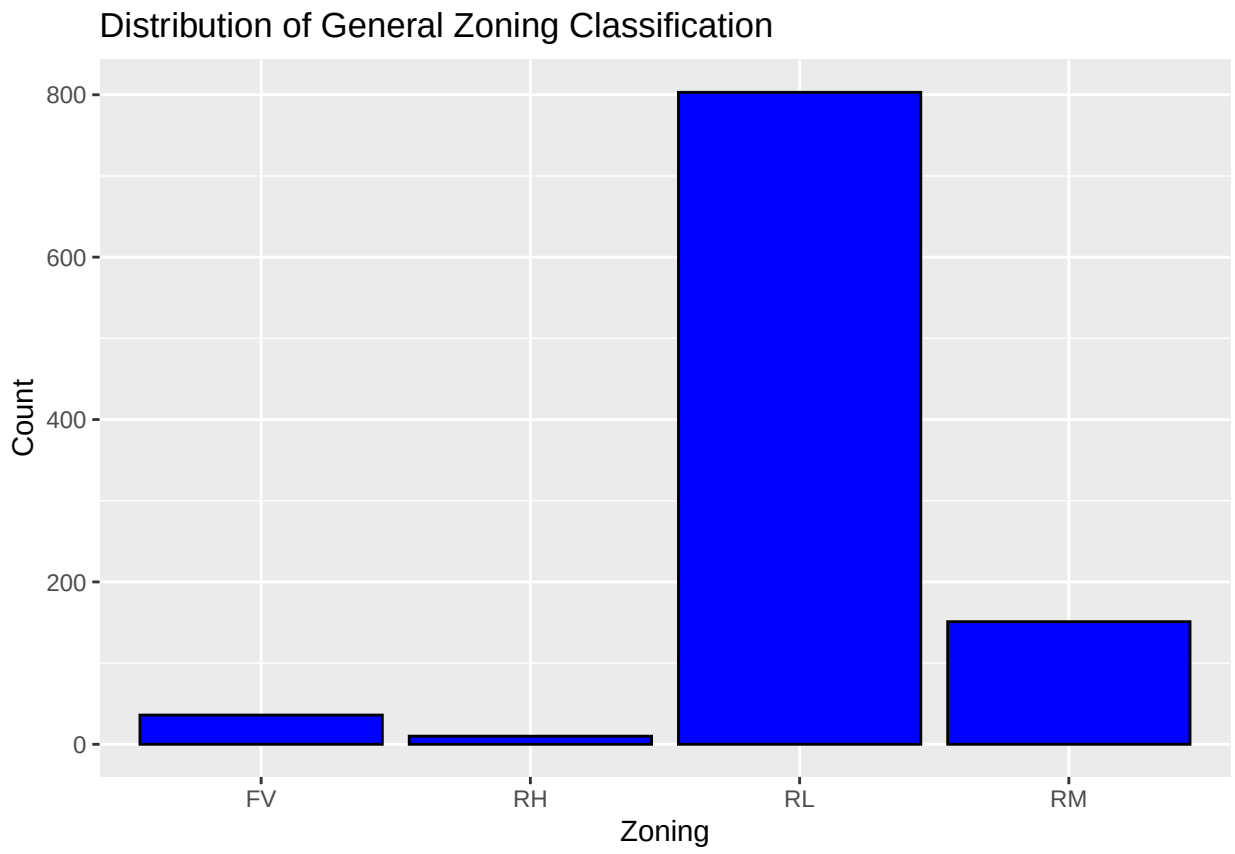
Problem 2: Analyzing Housing Data

Visualizing Housing Data with Various Plots

Zoning Classification Distribution

This bar plot shows the frequency of different zoning classifications in the dataset, helping to understand the prevalence of each zoning type, which can be useful for zoning-related analyses and understanding market segmentation.

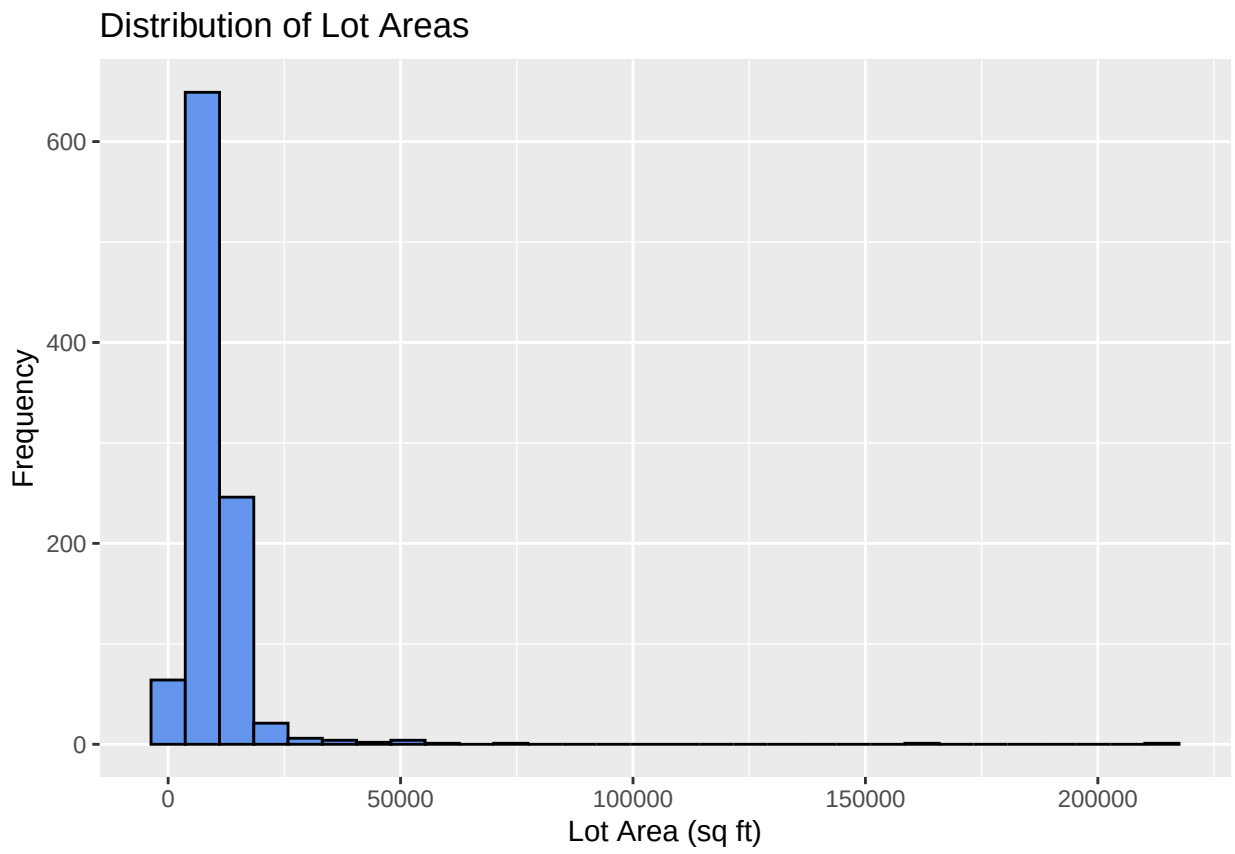
```
ggplot(housingData, aes(x = MSZoning)) +
  geom_bar(fill = "blue", color = "black") +
  labs(title = "Distribution of General Zoning Classification",
       x = "Zoning", y = "Count")
```



Lot Area Distribution

The histogram of lot areas provides a visual representation of the spread and frequency of lot sizes, highlighting any skewness or common lot sizes, which can inform property valuation and development considerations.

```
ggplot(housingData, aes(x = LotArea)) +
  geom_histogram(bins = 30, fill = "cornflowerblue", color = "black") +
  labs(title = "Distribution of Lot Areas",
       x = "Lot Area (sq ft)", y = "Frequency")
```

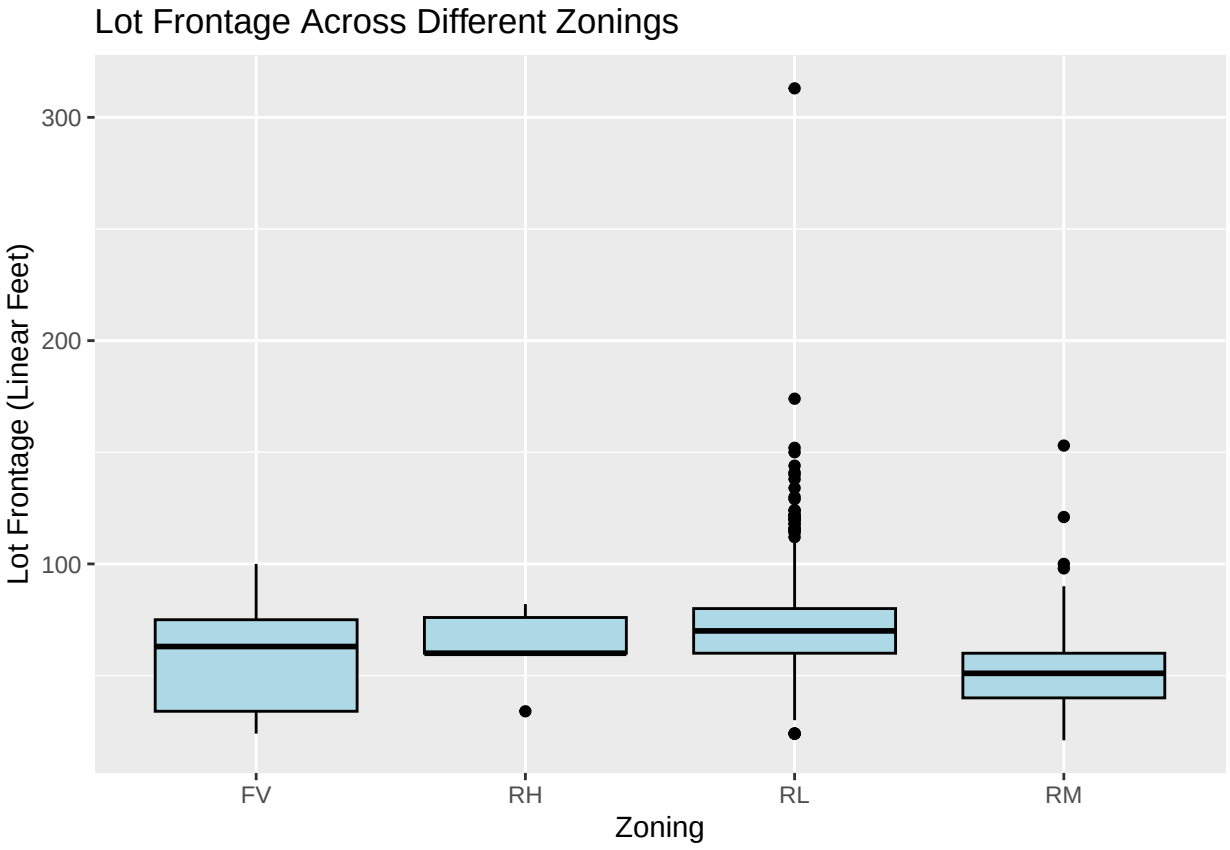


Lot Frontage by Zoning

This boxplot illustrates the distribution of lot frontage across various zoning classifications, revealing how lot frontage varies by zoning type, which is important for assessing zoning regulations and property accessibility.

```
ggplot(housingData, aes(x = MSZoning, y = LotFrontage)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Lot Frontage Across Different Zonings",
       x = "Zoning", y = "Lot Frontage (Linear Feet)")
```

```
## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Relationship Between Living Area and Sale Price

The scatter plot with a linear trend line shows the relationship between above-grade living area and sale price, highlighting trends and correlations which are crucial for predicting property values and identifying outliers.

The first scatter plot with points (geom_point) visualizes the overall relationship between living area and sale price across different zoning types, highlighting trends and outliers. The second plot with jitter (geom_jitter) addresses overplotting by slightly dispersing overlapping points, making it easier to observe the density and distribution of data within each zoning category. The faceted plot (facet_wrap(~MSZoning)) breaks down the data into separate plots for each zoning classification, allowing for a clearer comparison of trends and patterns within each zone independently.

```
ggplot(housingData, aes(x = GrLivArea, y = SalePrice)) +
  geom_point(aes(color = MSZoning), alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Living Area vs Sale Price by Zoning",
       x = "Above Grade Living Area (sq ft)", y = "Sale Price ($)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


```
ggplot(housingData, aes(x = GrLivArea, y = SalePrice)) +
  geom_jitter(aes(color = MSZoning), alpha = 0.6, width = 0.3, height = 0) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Living Area vs Sale Price by Zoning (Jitter)",
       x = "Above Grade Living Area (sq ft)", y = "Sale Price ($)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

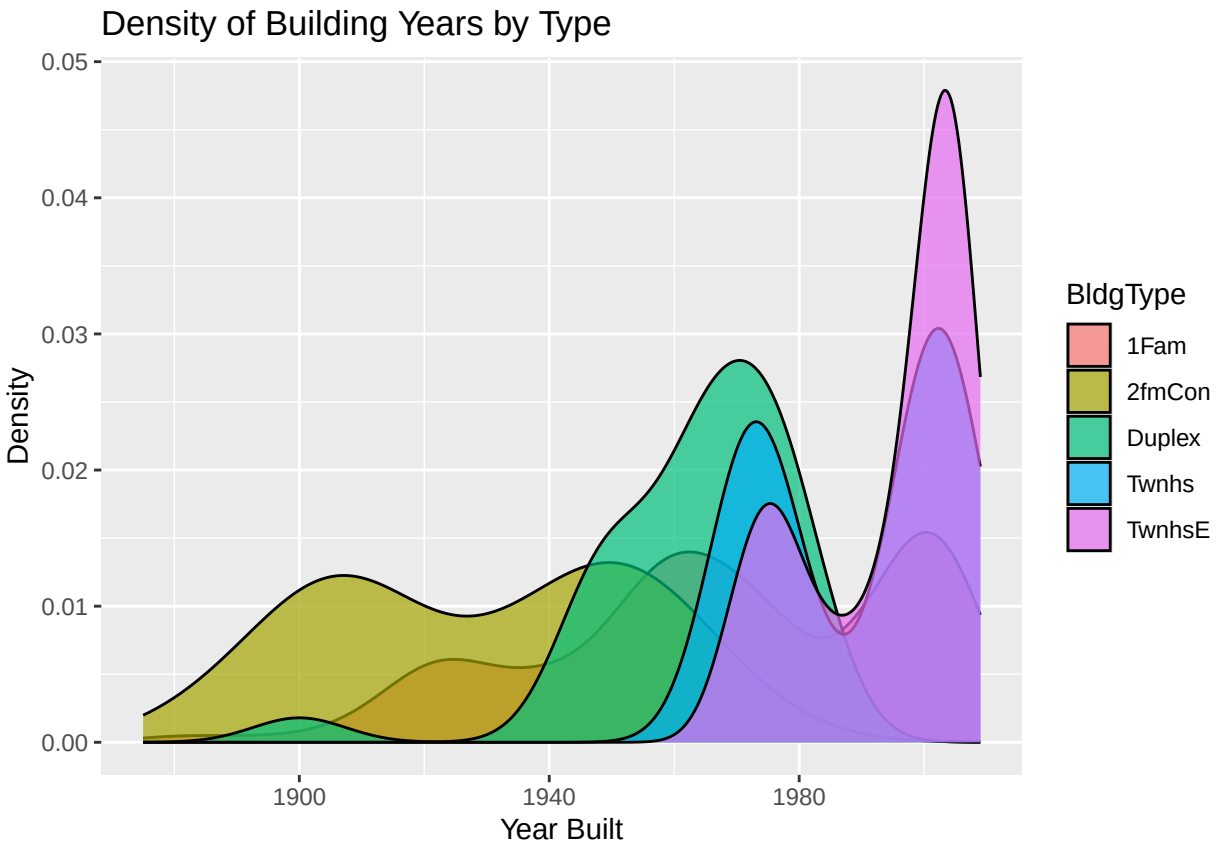

```
ggplot(housingData, aes(x = GrLivArea, y = SalePrice)) +
  geom_point(aes(color = MSZoning), alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  facet_wrap(~MSZoning) +
  labs(title = "Living Area vs Sale Price by Zoning (facet)",
       x = "Above Grade Living Area (sq ft)", y = "Sale Price ($)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Building Year Density by Building Type

The density plot visualizes the distribution of construction years across different building types, providing insights into the age distribution of homes which can affect market value and renovation needs.

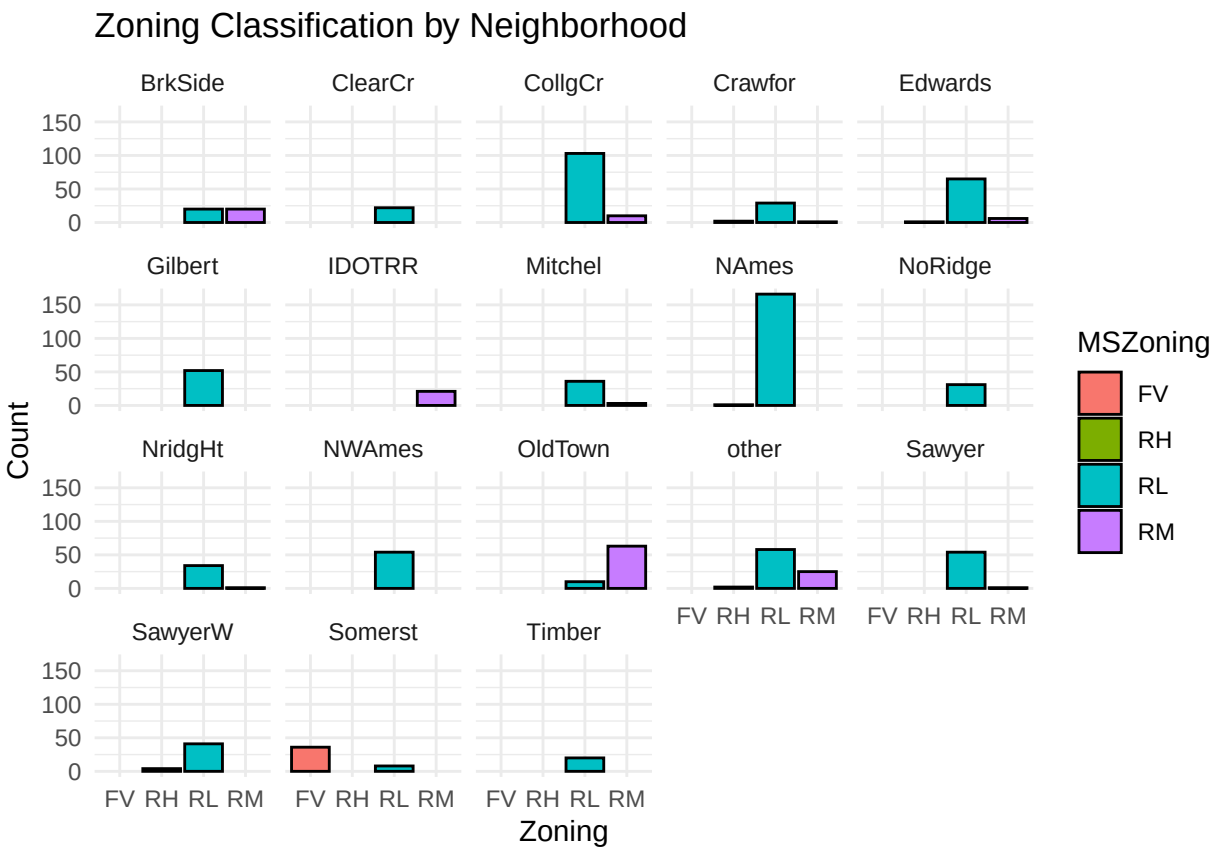
```
ggplot(housingData, aes(x = YearBuilt, fill = BldgType)) +
  geom_density(alpha = 0.7) +
  labs(title = "Density of Building Years by Type",
       x = "Year Built", y = "Density")
```



Zoning Classification by Neighborhood

The faceted bar plot shows zoning classification distributions across neighborhoods, allowing for a comparative analysis of zoning preferences and real estate development patterns in different areas.

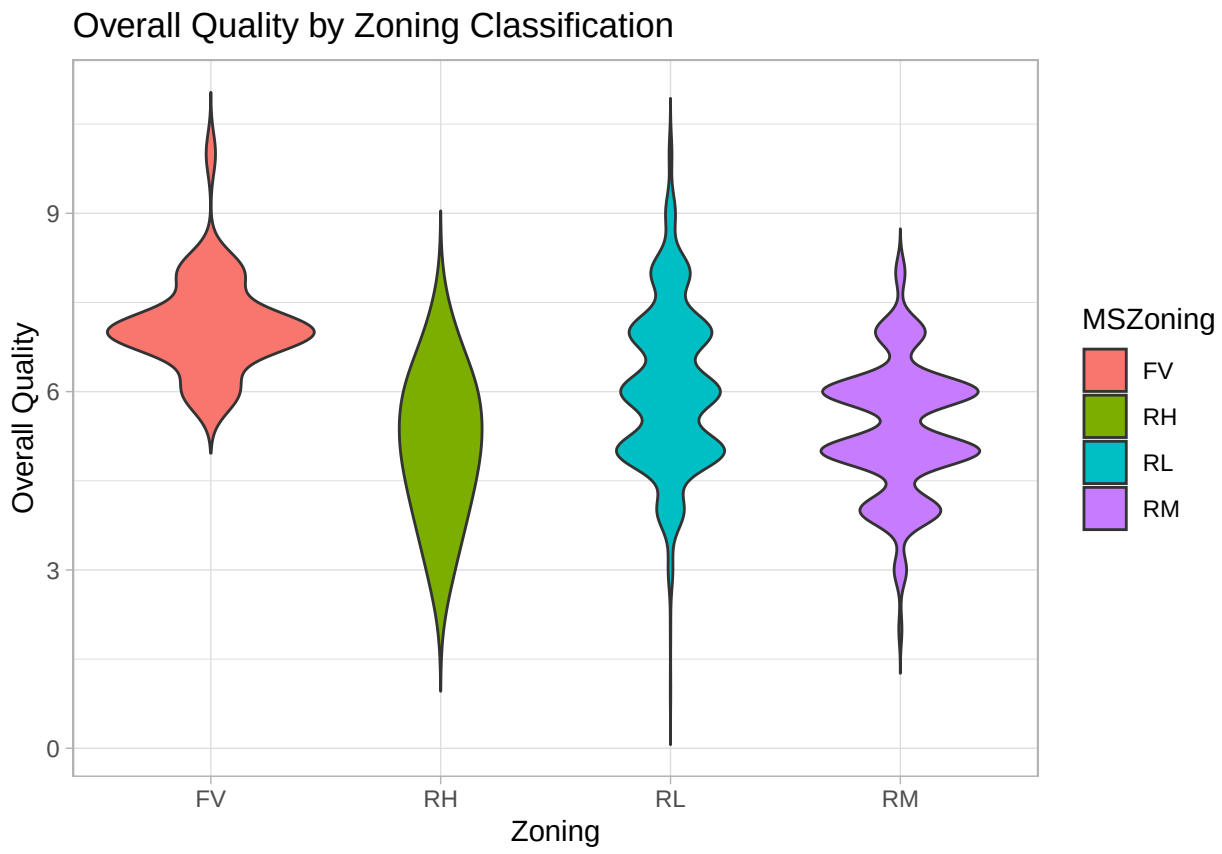
```
ggplot(housingData, aes(x = MSZoning)) +
  geom_bar(aes(fill = MSZoning), color = "black") +
  facet_wrap(~ Neighborhood) +
  labs(title = "Zoning Classification by Neighborhood",
       x = "Zoning", y = "Count") +
  theme_minimal()
```



Overall Quality by Zoning Classification

This violin plot displays the distribution of overall quality ratings across different zoning classifications, offering insights into the quality standards associated with various zoning types, which can influence property prices and buyer preferences.

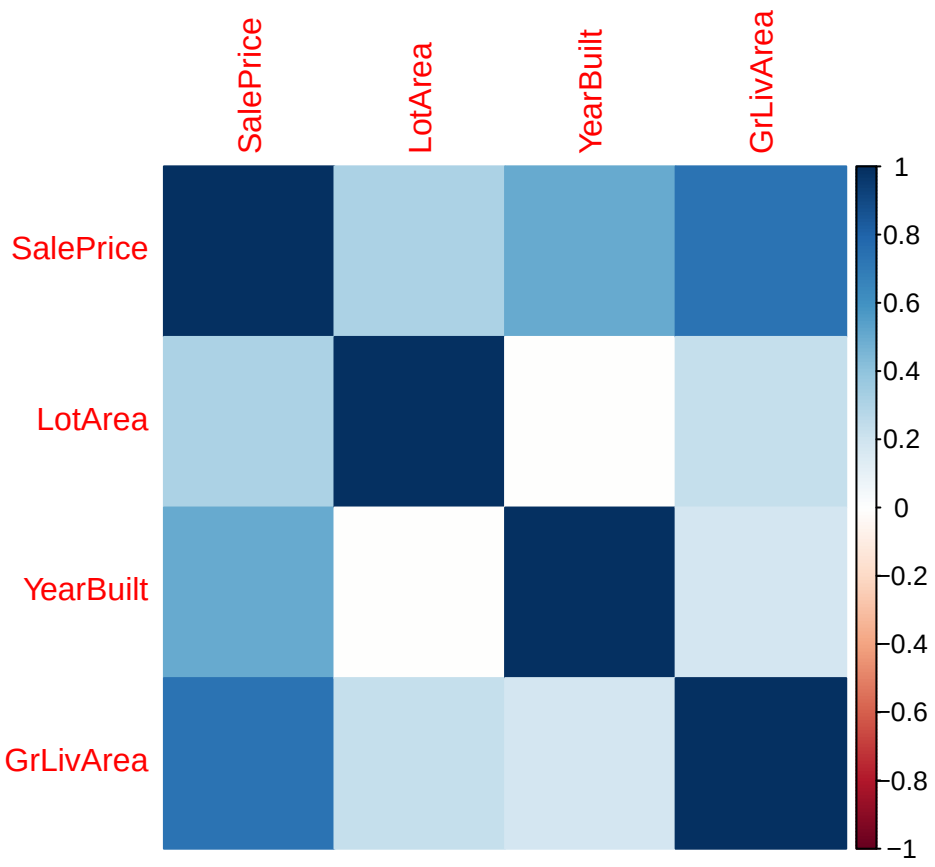
```
ggplot(housingData, aes(x = MSZoning, y = OverallQual, fill = MSZoning)) +
  geom_violin(trim = FALSE) +
  labs(title = "Overall Quality by Zoning Classification",
       x = "Zoning", y = "Overall Quality") +
  theme_light()
```



Correlation Matrix

The correlation plot visualizes the relationships between features like - SalePrice, LotArea, YearBuilt, GrLivArea, helping to identify strong correlations that can be used for predictive modeling and multivariate analysis, particularly in understanding factors that drive property prices.

```
numericalData <- housingData %>% select(SalePrice, LotArea, YearBuilt, GrLivArea)
corrMatrix <- cor(numericalData, use = "complete.obs")
corrplot(corrMatrix, method = "color")
```


SalePrice
LotArea
YearBuilt
GrLivArea
SalePrice
LotArea
YearBuilt
GrLivArea
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1