RWorksheet_Aguire#4C

Ryza Faith Aguirre

2024-11-04

1.

```
mpg_data <- read.csv("mpg.csv")
head(mpg_data)
```

```
##   X manufacturer model displ year cyl      trans drv cty hwy fl   class
## 1 1         audi    a4   1.8 1999   4   auto(l5)   f  18  29  p compact
## 2 2         audi    a4   1.8 1999   4 manual(m5)   f  21  29  p compact
## 3 3         audi    a4   2.0 2008   4 manual(m6)   f  20  31  p compact
## 4 4         audi    a4   2.0 2008   4   auto(av)   f  21  30  p compact
## 5 5         audi    a4   2.8 1999   6   auto(l5)   f  16  26  p compact
## 6 6         audi    a4   2.8 1999   6 manual(m5)   f  18  26  p compact
```

1b. Categorical Variables in the mpg Dataset manufacturer, model, trans, drv, fl, class

1c. Continuous Variables in the mpg Dataset displ, cty, hwy, year

2a.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
model_count_by_manufacturer <- mpg_data %>%
  group_by(manufacturer) %>%
  summarise(unique_models = n_distinct(model)) %>%
  arrange(desc(unique_models))

print(model_count_by_manufacturer)
```

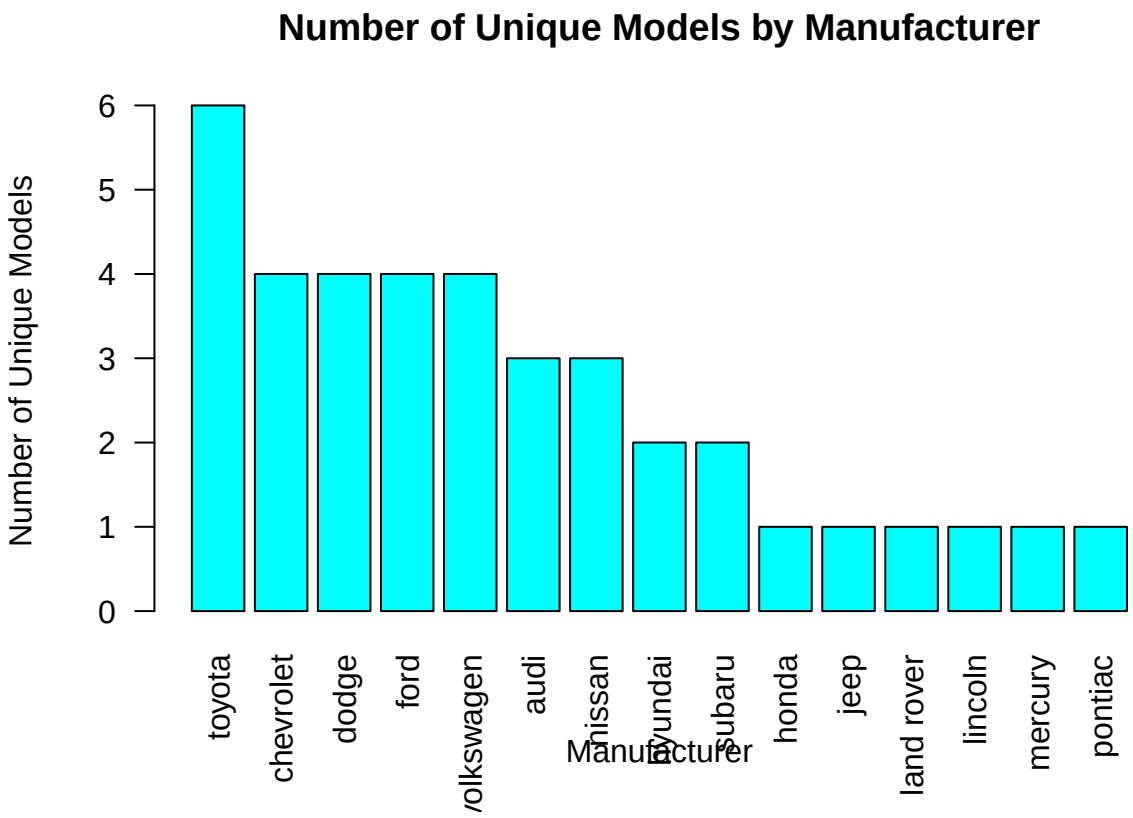
```
## # A tibble: 15 x 2
##    manufacturer unique_models
##    <chr>                <int>
##  1 toyota                   6
##  2 chevrolet                4
##  3 dodge                    4
##  4 ford                     4
##  5 volkswagen               4
##  6 audi                     3
##  7 nissan                   3
##  8 hyundai                  2
##  9 subaru                   2
## 10 honda                    1
## 11 jeep                     1
## 12 land rover               1
## 13 lincoln                  1
## 14 mercury                  1
## 15 pontiac                  1
```

```
variation_count_by_model <- mpg_data %>%
  group_by(model) %>%
  summarise(variations = n()) %>%
  arrange(desc(variations))

print(variation_count_by_model)
```

```
## # A tibble: 38 x 2
##    model               variations
##    <chr>                    <int>
##  1 caravan 2wd                 11
##  2 ram 1500 pickup 4wd         10
##  3 civic                        9
##  4 dakota pickup 4wd            9
##  5 jetta                        9
##  6 mustang                      9
##  7 a4 quattro                   8
##  8 grand cherokee 4wd           8
##  9 impreza awd                  8
## 10 a4                           7
## # i 28 more rows
```

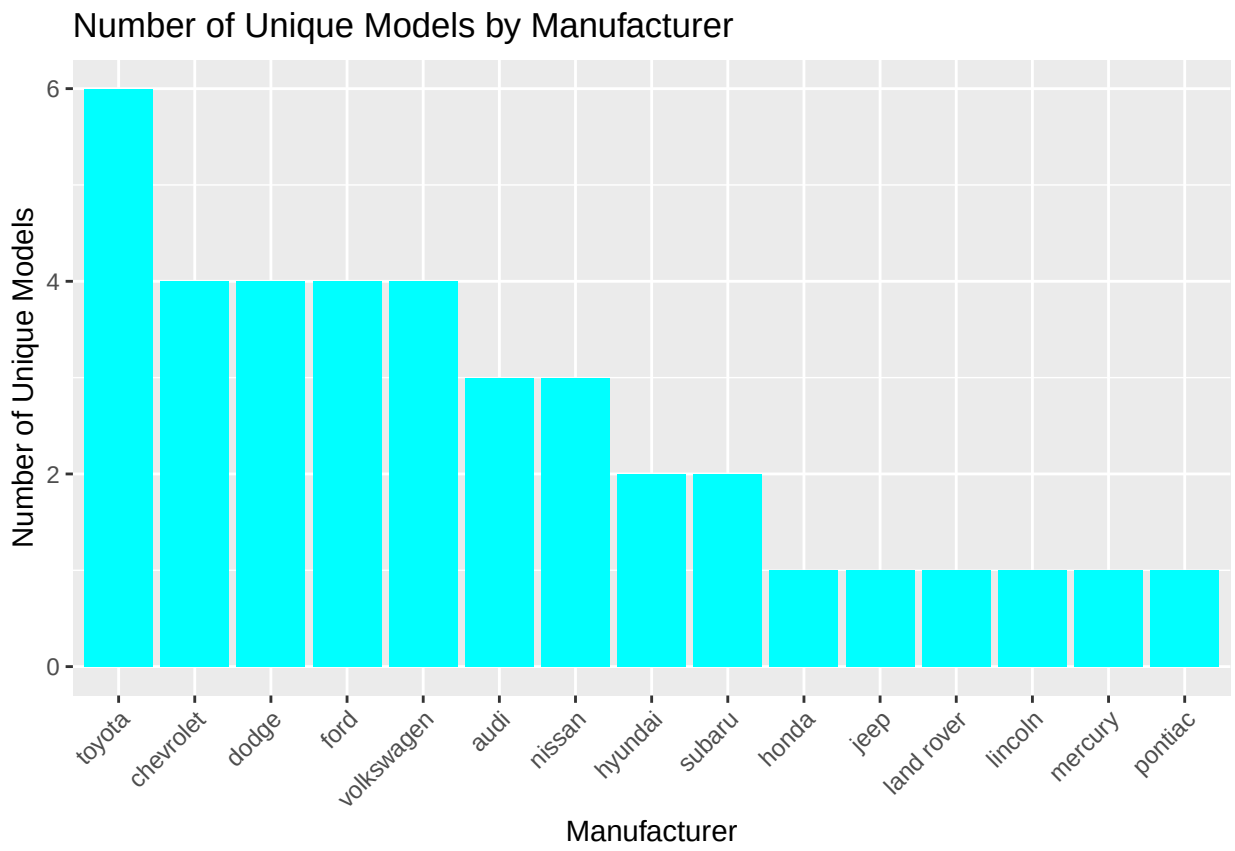
2b. Using plot()

```
sorted_data <- model_count_by_manufacturer[order(model_count_by_manufacturer$unique_models, decreasing =
barplot(sorted_data$unique_models, names.arg = sorted_data$manufacturer, las = 2, col = "cyan",
        main = "Number of Unique Models by Manufacturer",
        xlab = "Manufacturer", ylab = "Number of Unique Models")
```



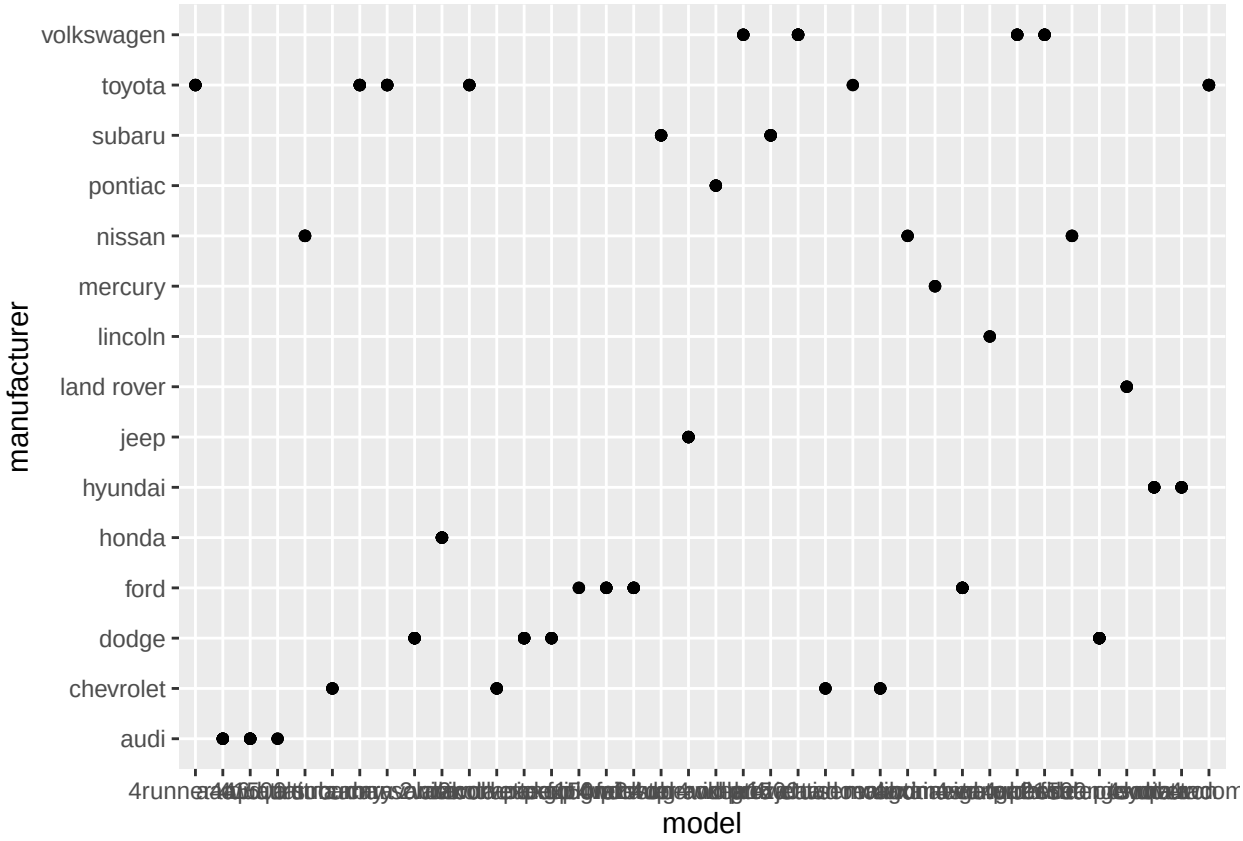
2b. ggplot()

```
library(ggplot2)
ggplot(model_count_by_manufacturer, aes(x = reorder(manufacturer, -unique_models), y = unique_models)) +
  geom_bar(stat = "identity", fill = "cyan") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Number of Unique Models by Manufacturer",
       x = "Manufacturer", y = "Number of Unique Models")
```



2a.

```
ggplot(mpg_data, aes(x = model, y = manufacturer)) + geom_point()
```

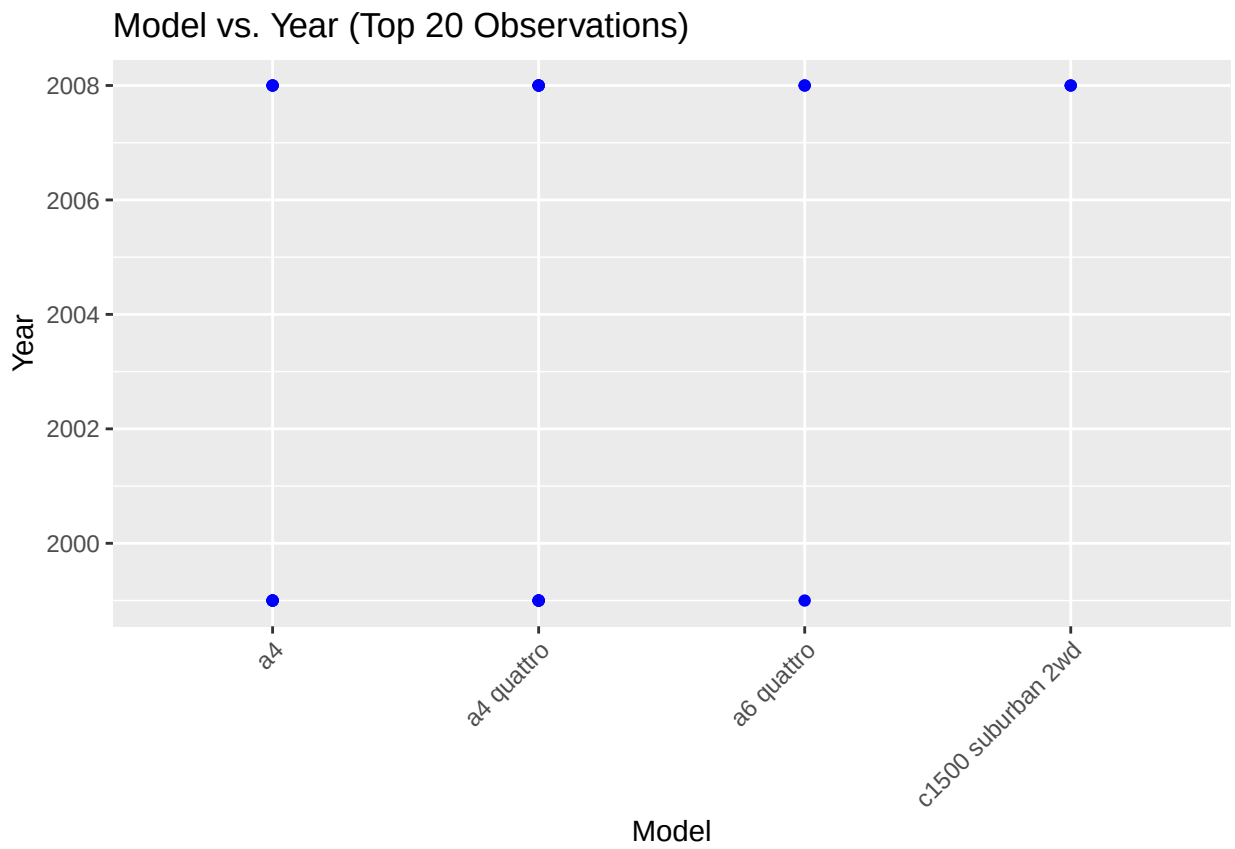


2a. This plot shows the relationship between model and manufacturer but may not be particularly informative. Since there are many model names, they may overlap, making the plot cluttered and difficult to interpret. It essentially shows which manufacturers have which models but doesn't provide much additional insight.

2b. Is This Plot Useful? Suggested Improvements The scatter plot isn't particularly useful in this format due to overlap and lack of quantitative insight.

3.

```
top_20_mpg <- mpg_data[1:20, ]

ggplot(top_20_mpg, aes(x = model, y = year)) +
  geom_point(color = "blue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Model vs. Year (Top 20 Observations)",
       x = "Model", y = "Year")
```



4.

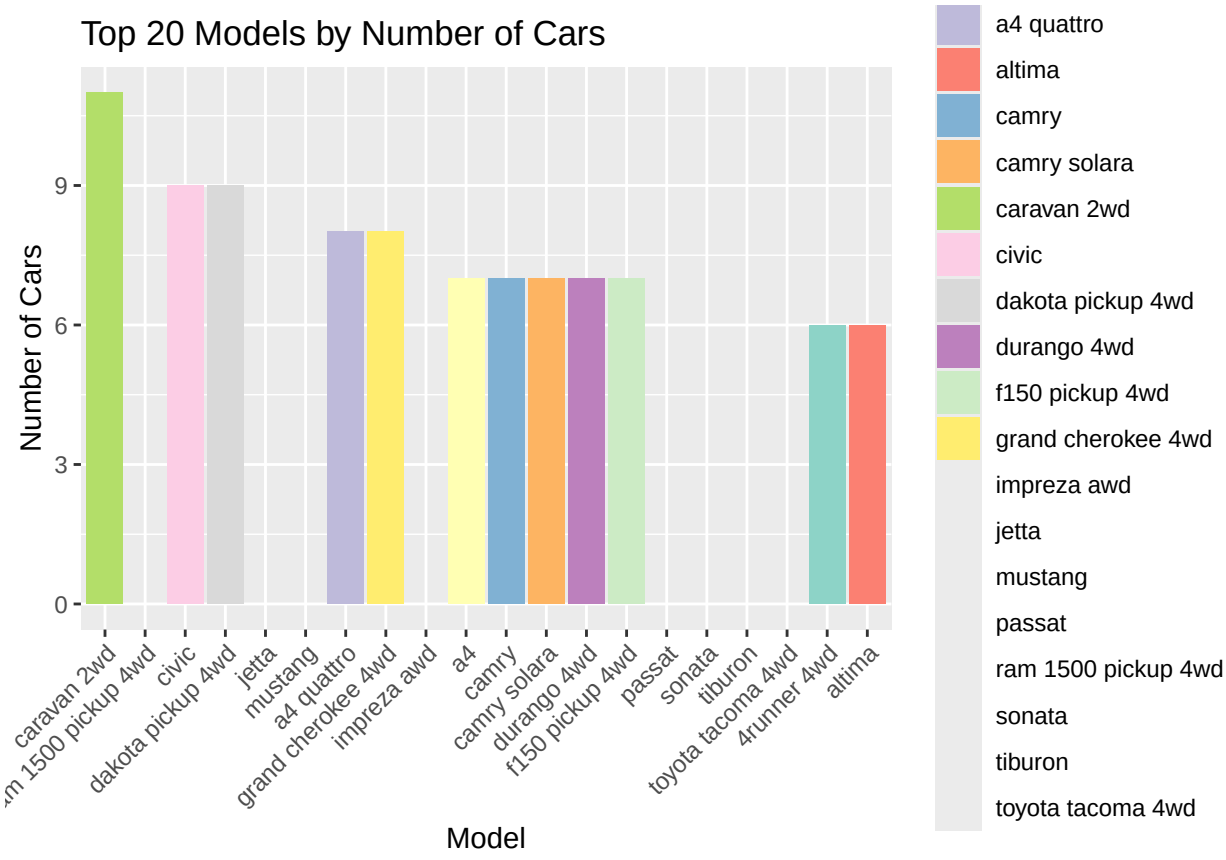
```
library(dplyr)

cars_per_model <- mpg_data %>%
  group_by(model) %>%
  summarise(car_count = n()) %>%
  arrange(desc(car_count))

print(cars_per_model)
```

```
## # A tibble: 38 x 2
##    model               car_count
##    <chr>                   <int>
##  1 caravan 2wd                11
##  2 ram 1500 pickup 4wd        10
##  3 civic                       9
##  4 dakota pickup 4wd           9
##  5 jetta                       9
##  6 mustang                     9
##  7 a4 quattro                  8
##  8 grand cherokee 4wd          8
##  9 impreza awd                 8
## 10 a4                          7
## # i 28 more rows
```

4a.

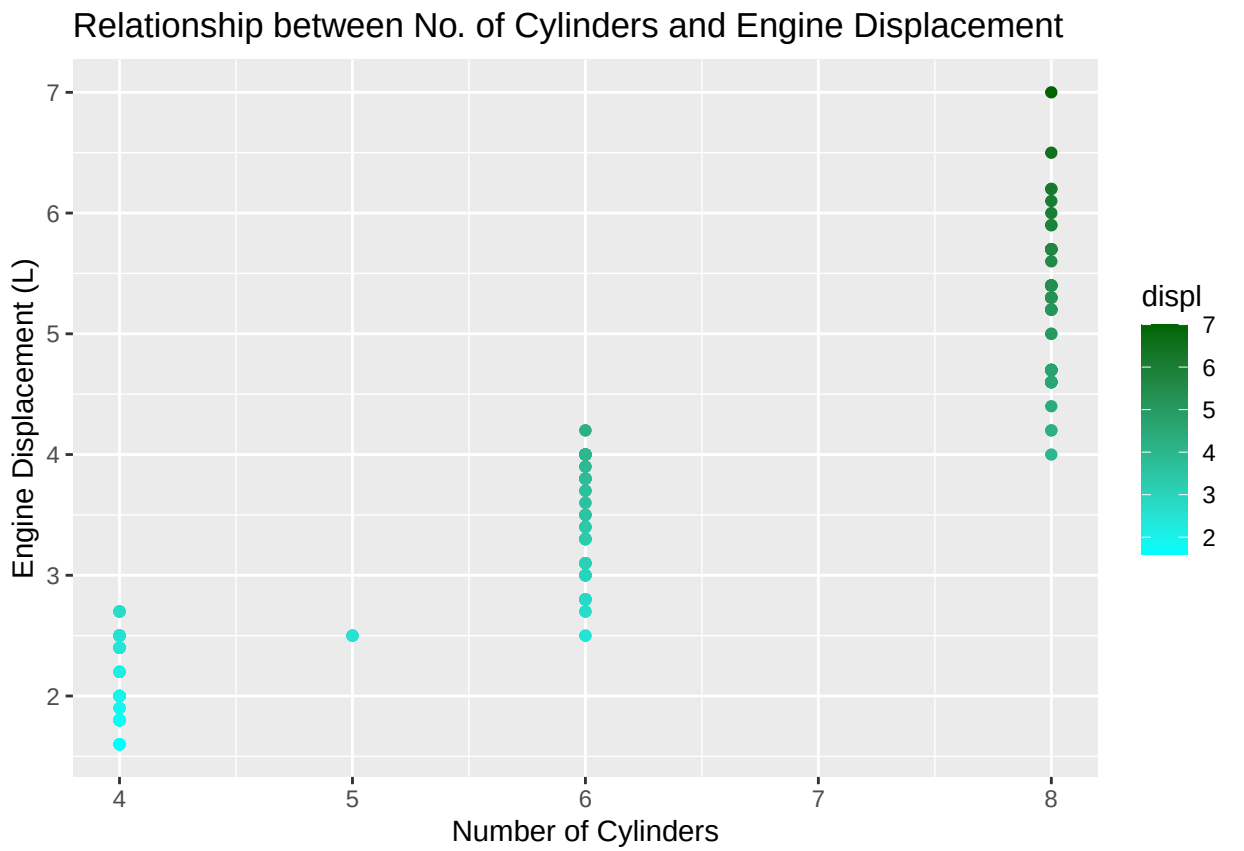
```r
top_20_models <- cars_per_model[1:20, ]

library(ggplot2)
ggplot(top_20_models, aes(x = reorder(model, -car_count), y = car_count, fill = model)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 20 Models by Number of Cars",
       x = "Model",
       y = "Number of Cars") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")
```

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Set3 is 12
## Returning the palette you asked for with that many colors
```



5.

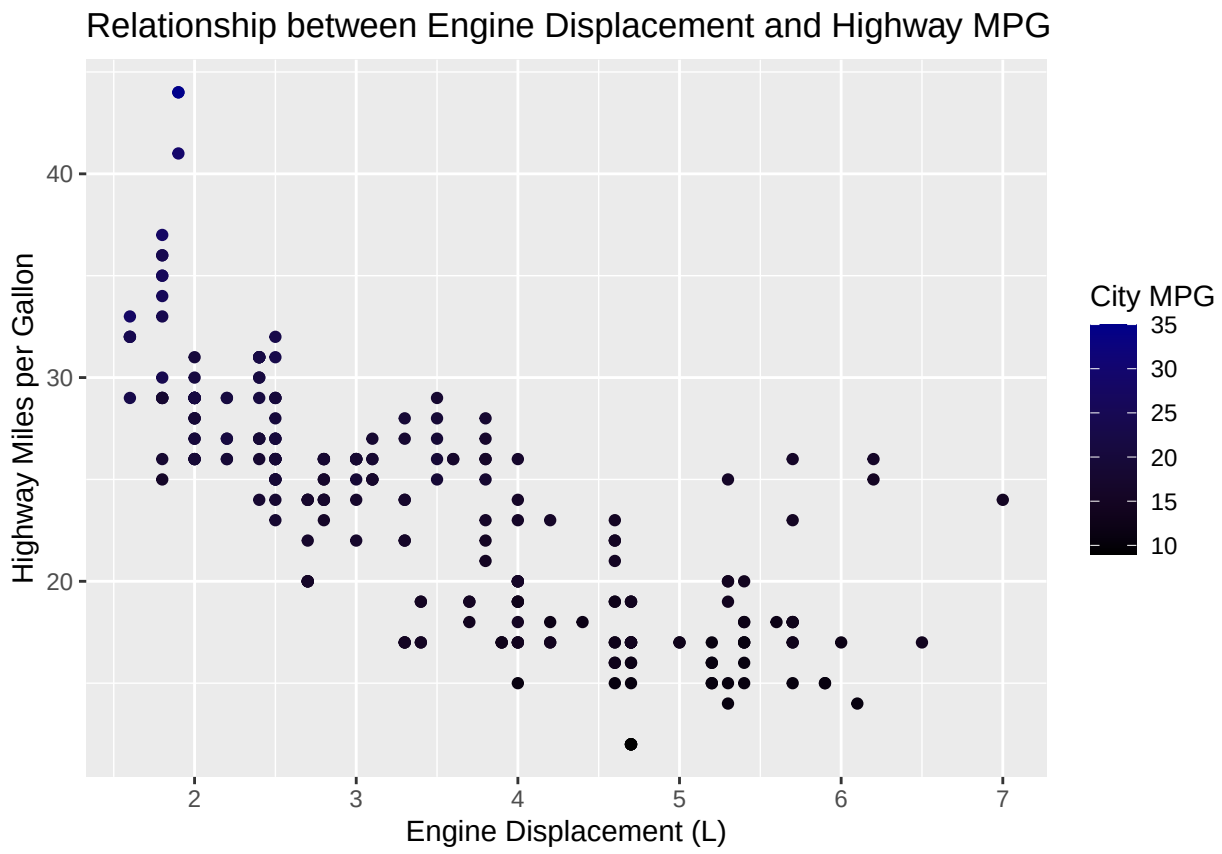
```r
ggplot(mpg_data, aes(x = cyl, y = displ, color = displ)) +
  geom_point() +
  labs(title = "Relationship between No. of Cylinders and Engine Displacement",
       x = "Number of Cylinders",
       y = "Engine Displacement (L)") +
  scale_color_gradient(low = "cyan", high = "darkgreen")
```



The plot shows that as the number of cylinders (cyl) increases, the engine displacement (displ) also generally increases. The positive trend suggests that cars with more cylinders tend to have larger engines, which makes sense because more cylinders typically mean a larger engine size. The color gradient further emphasizes the increase in displacement as the cylinder count goes up.

6.

```
ggplot(mpg_data, aes(x = displ, y = hwy, color = cty)) +
  geom_point() +
  labs(title = "Relationship between Engine Displacement and Highway MPG",
       x = "Engine Displacement (L)",
       y = "Highway Miles per Gallon",
       color = "City MPG") +
  scale_color_gradient(low = "black", high = "darkblue")
```



6.

```
traffic_data <- read.csv("~/DataScience/CS101/worksheet4c/traffic.csv")
head(traffic_data)
```

```
##              DateTime Junction Vehicles          ID
## 1 2015-11-01 00:00:00        1       15 20151101001
## 2 2015-11-01 01:00:00        1       13 20151101011
## 3 2015-11-01 02:00:00        1       10 20151101021
## 4 2015-11-01 03:00:00        1        7 20151101031
## 5 2015-11-01 04:00:00        1        9 20151101041
## 6 2015-11-01 05:00:00        1        6 20151101051
```

```
str(traffic_data)
```

```
## 'data.frame':    48120 obs. of  4 variables:
##  $ DateTime: chr  "2015-11-01 00:00:00" "2015-11-01 01:00:00" "2015-11-01 02:00:00" "2015-11-01 03:0
##  $ Junction: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Vehicles: int  15 13 10 7 9 6 9 8 11 12 ...
##  $ ID      : num  2.02e+10 2.02e+10 2.02e+10 2.02e+10 2.02e+10 ...
```

```
nrow(traffic_data)
```

```
## [1] 48120
```

```r
colnames(traffic_data)
```

```
## [1] "DateTime" "Junction" "Vehicles" "ID"
```