RWorksheet_Elizalde#4c

Christian Alimace Elizalde

2024-11-04

#1. #a.

```r
mpg <- read.csv("/cloud/project/Rworksheet4c/mpg.csv")
```

#b. #The categorical variables are manufacturers, model, trans, drv, fl, and class

#c. #The continuous variables are displ, year, cyl, cty, and hwy

#2.

```r
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
```

```r
manufacturers <- mpg %>%
  group_by(manufacturer) %>%
  summarise(unique_models = n_distinct(model)) %>%
  arrange(desc(unique_models))
print(manufacturers)
```

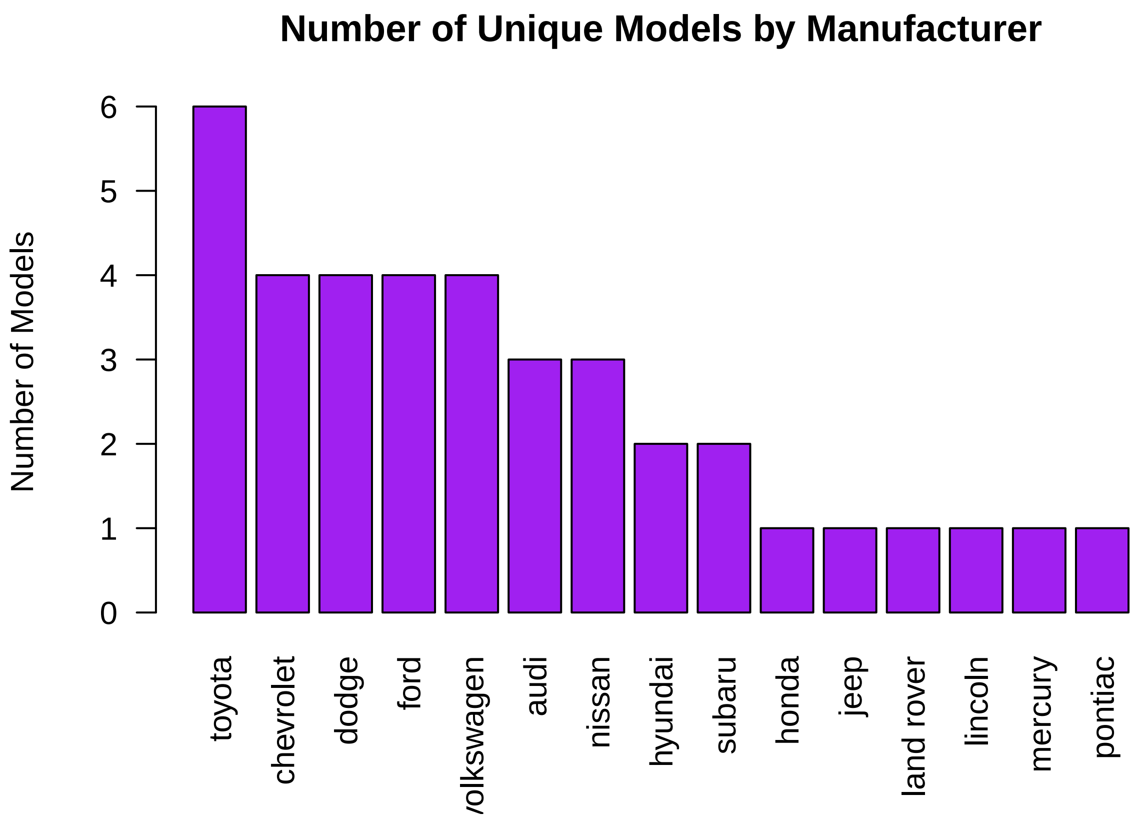
```
## # A tibble: 15 x 2
##    manufacturer unique_models
##    <chr>                <int>
##  1 toyota                   6
##  2 chevrolet                4
##  3 dodge                    4
##  4 ford                     4
##  5 volkswagen               4
##  6 audi                     3
##  7 nissan                   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
```

```
models <- mpg %>%
  group_by(model) %>%
  summarise(variations = n()) %>%
  arrange(desc(variations))


print(models)
```

```
## # 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
```

#b.

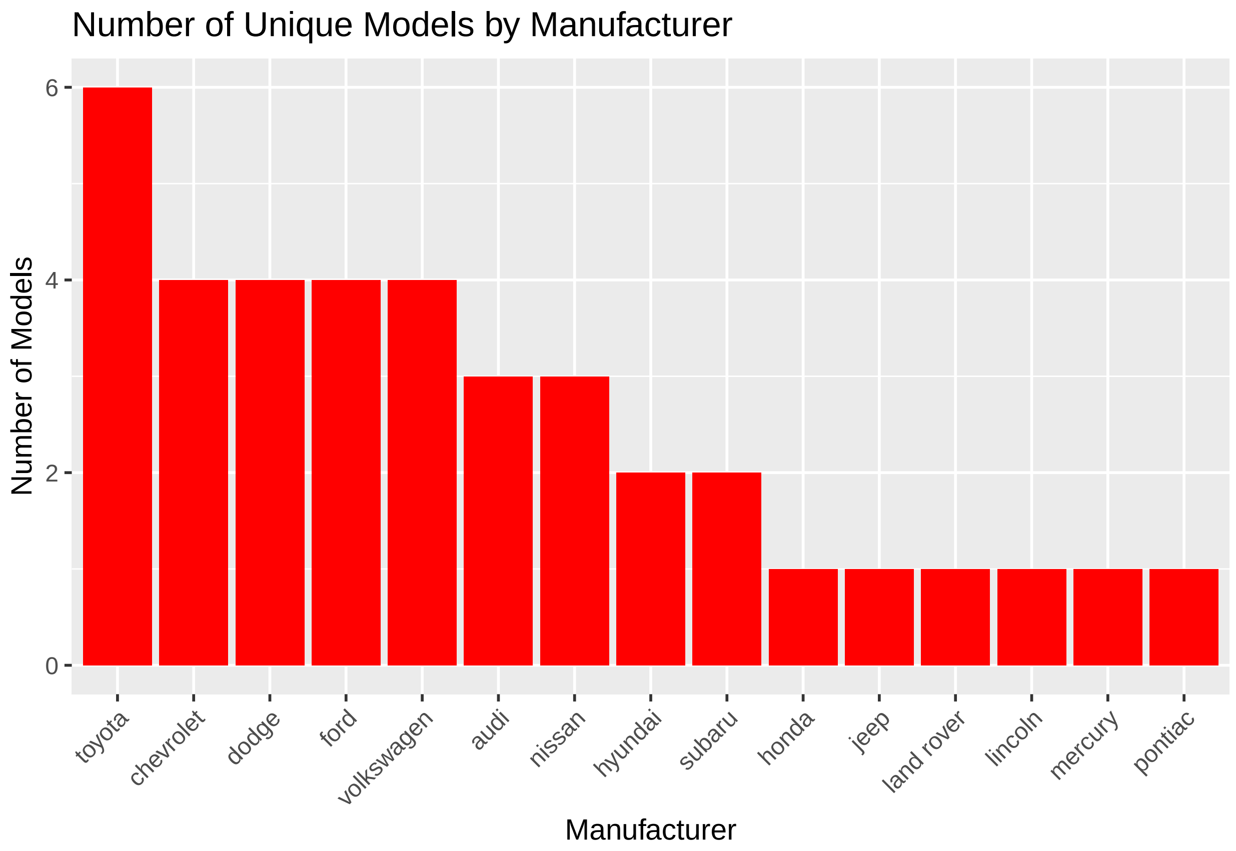
```
barplot(
  manufacturers$unique_models,
  names.arg = manufacturers$manufacturer,
  las = 2,
  col = "purple",
  main = "Number of Unique Models by Manufacturer",
  ylab = "Number of Models"
)
```



```
library(ggplot2)
```

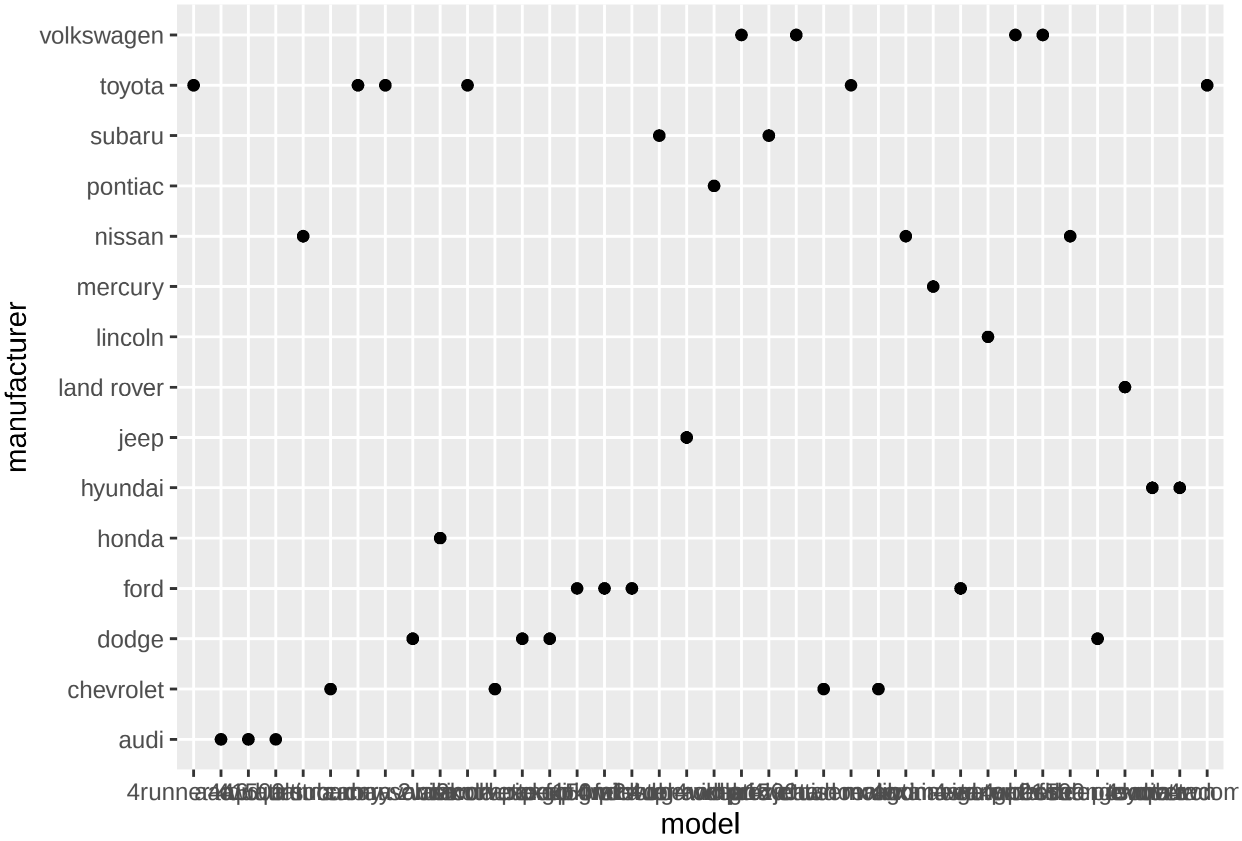
```
##
## Attaching package: 'ggplot2'

## The following object is masked _by_ '.GlobalEnv':
##
##     mpg
```

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



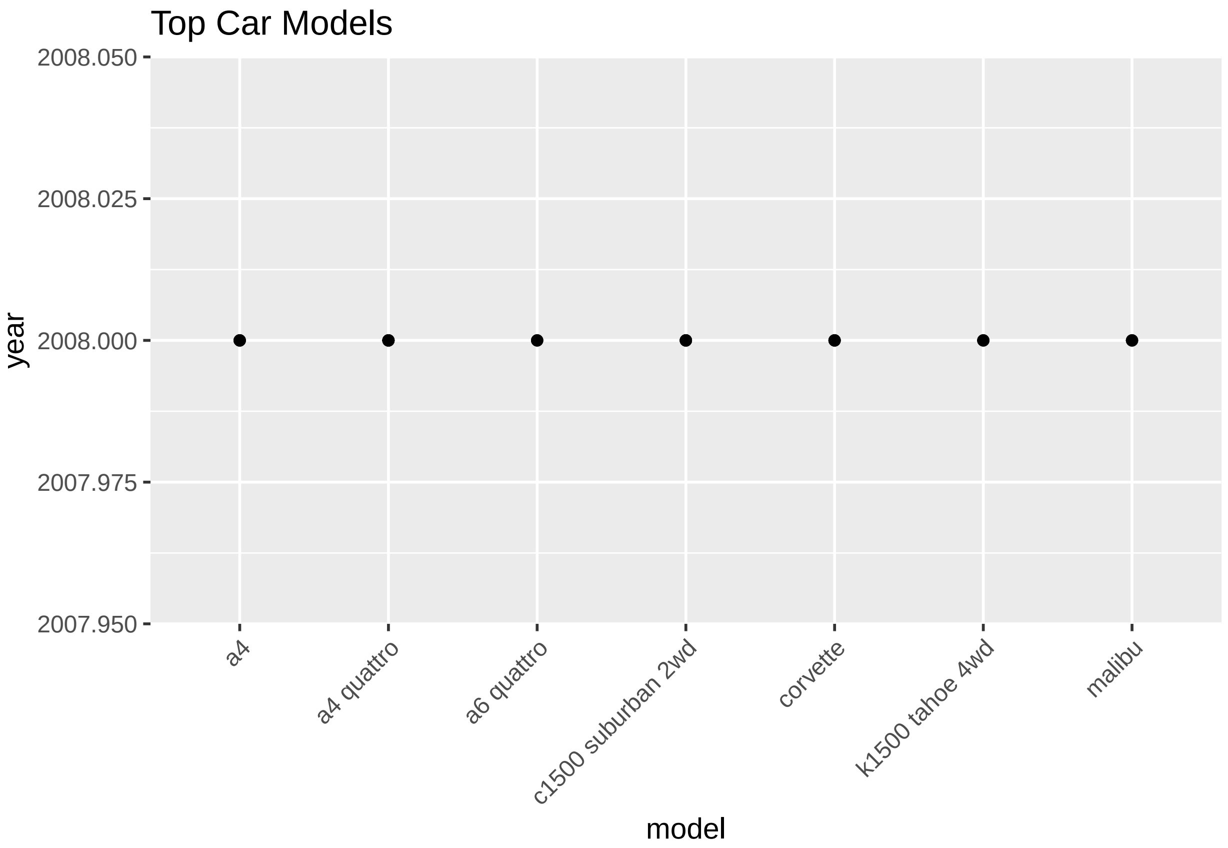
#2. #a.

```
ggplot(mpg, aes(model, manufacturer)) + geom_point()
```



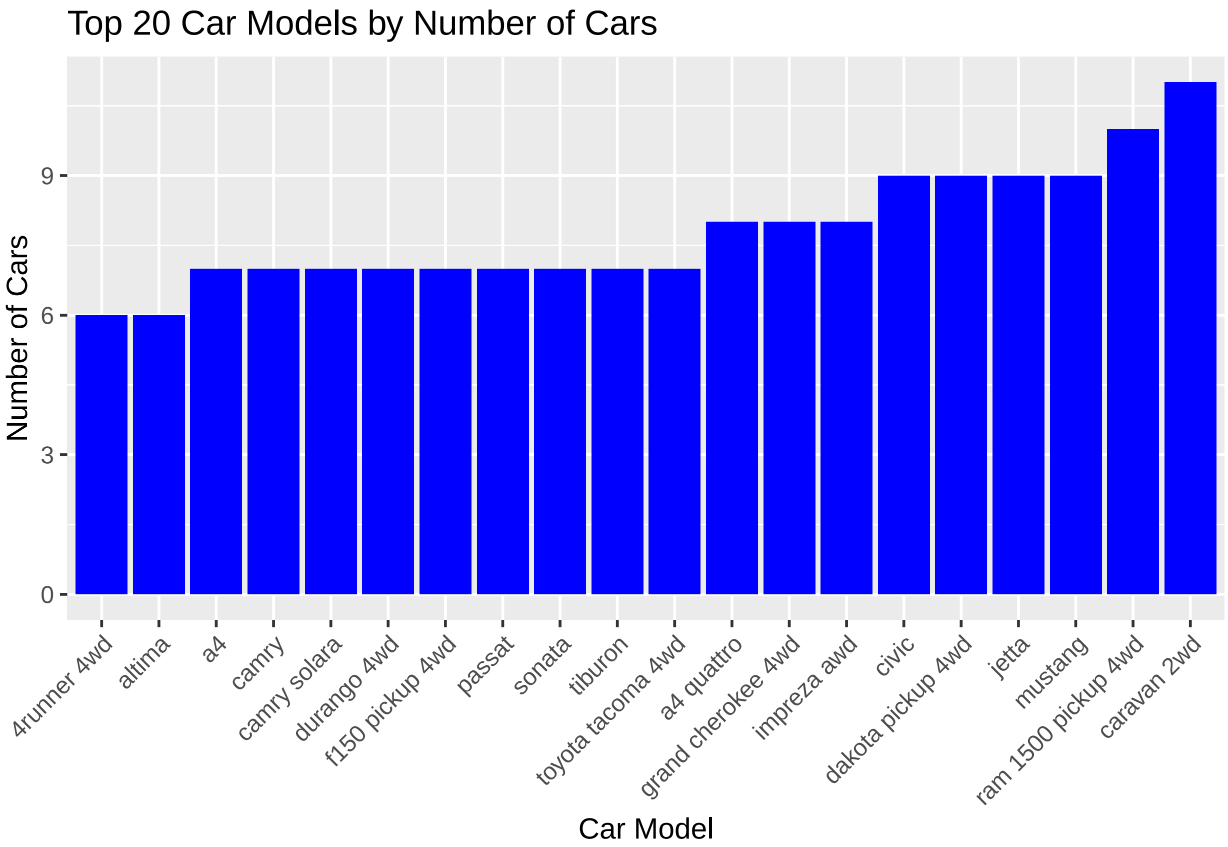
#b. #For me it is not useful because it lacks context and the text in x-axis is not readable. To make it more readable, I will use barplot to show the count of models for each manufacturer.

```
library(dplyr)
library(ggplot2)

top20 <- mpg %>%
  arrange(desc(year)) %>%
  head(20)

ggplot(top20, aes(x= model, y= year,)) + geom_point() + labs(title = "Top Car Models", xlab="Car model"
```



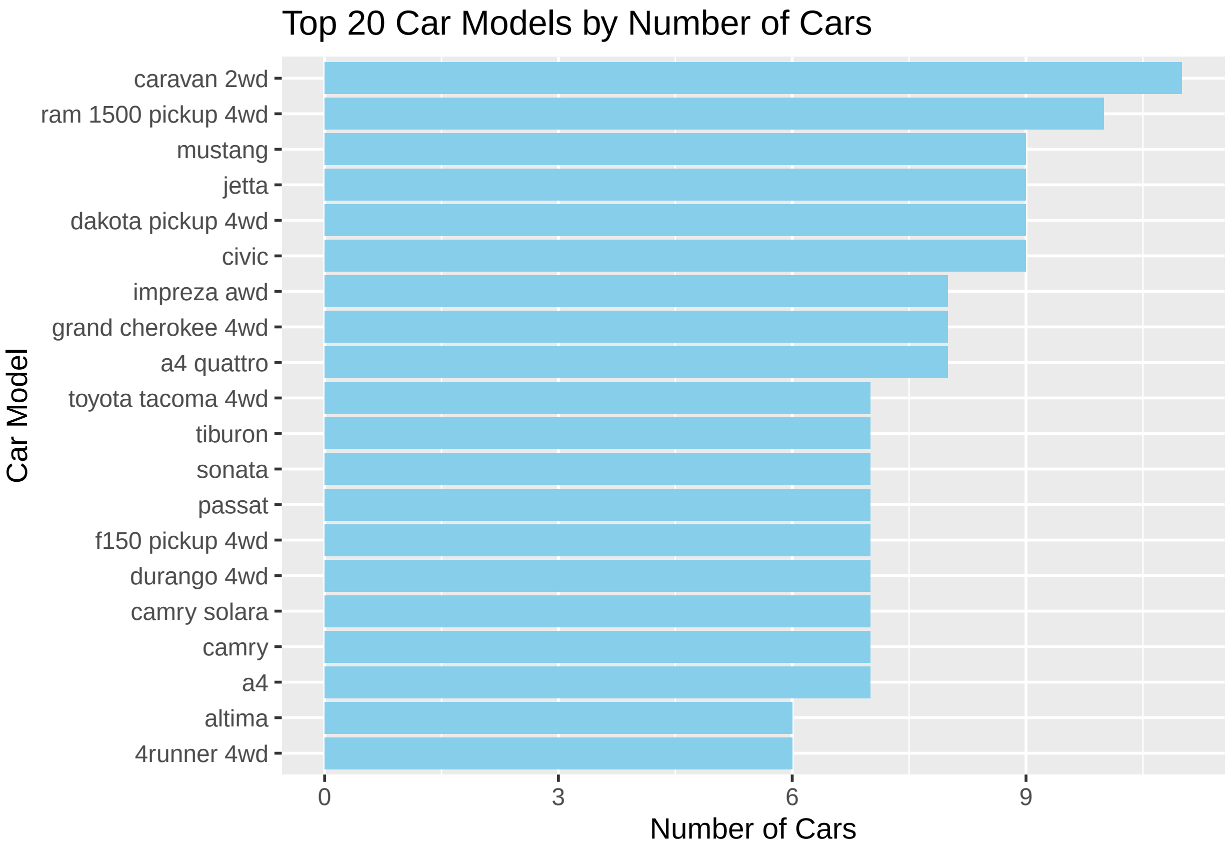
#4. #a.

```
library(dplyr)
library(ggplot2)

car_counts <- mpg %>%
  group_by(model) %>%
  summarise(number_of_cars = n()) %>%
  arrange(desc(number_of_cars))



top_20_models <- head(car_counts, 20)


ggplot(top_20_models, aes(x = reorder(model, number_of_cars), y = number_of_cars)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Top 20 Car Models by Number of Cars",
       x = "Car Model",
       y = "Number of Cars") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



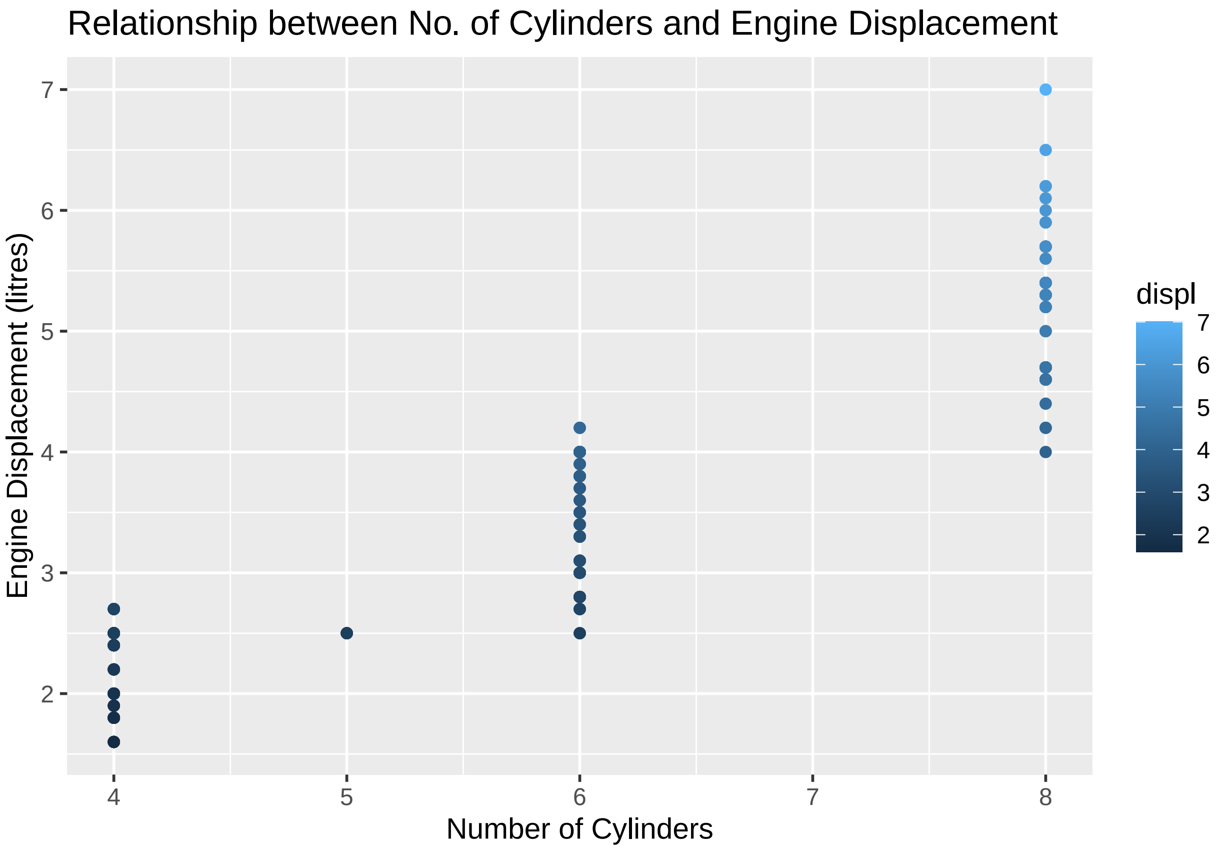
#b.

```
ggplot(top_20_models, aes(x = reorder(model, number_of_cars), y = number_of_cars)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Top 20 Car Models by Number of Cars",
       x = "Car Model",
       y = "Number of Cars") +
  coord_flip()
```



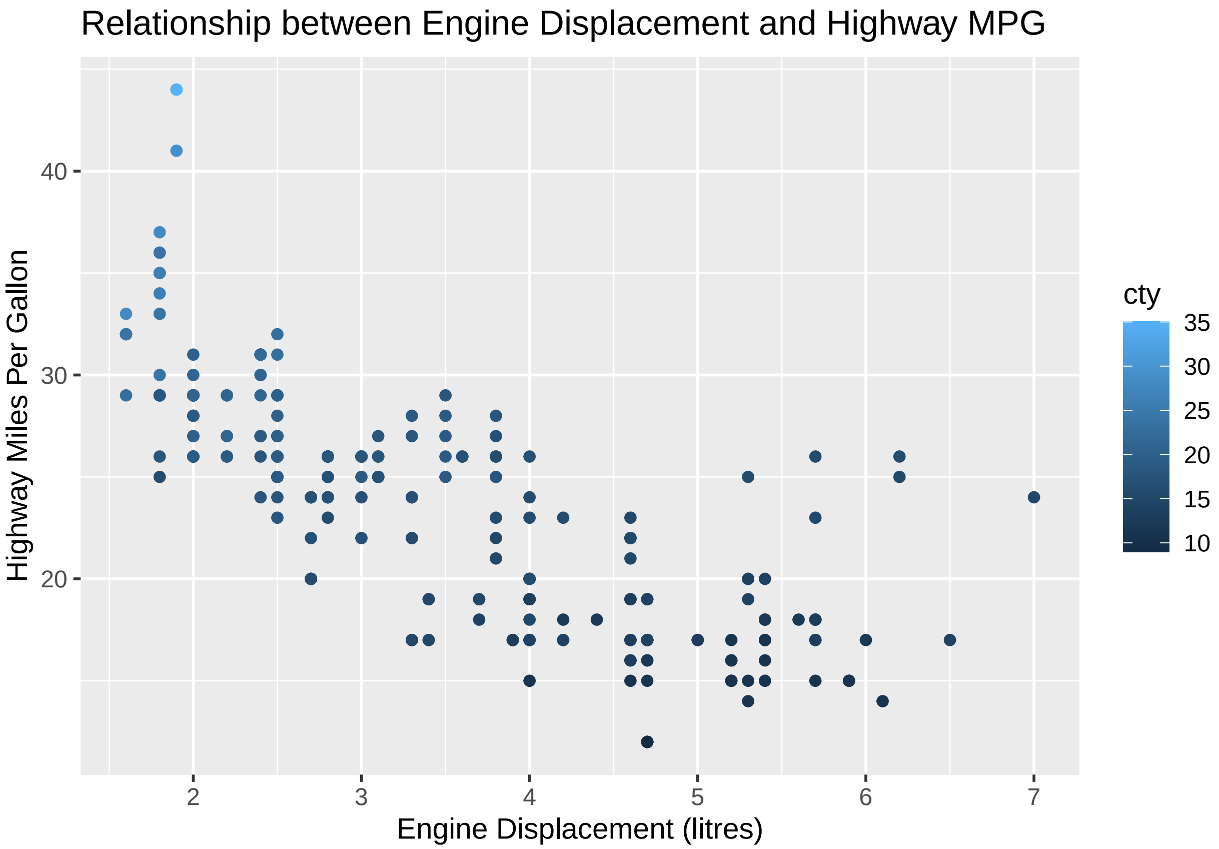
#5 #a.Vehicles designed with more cylinders are likely to have larger engine capacities

```r
ggplot(mpg, 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 (litres)")
```



#6.The result displays the relationship between engine displacement and highway miles per gallon, with points colored based on city mpg

```
library(ggplot2)


ggplot(mpg, aes(x = displ, y = hwy, color = cty)) +
  geom_point() +
  labs(title = "Relationship between Engine Displacement and Highway MPG",
       x = "Engine Displacement (litres)",
       y = "Highway Miles Per Gallon")
```



#a.

```
traffic <- read.csv("/cloud/project/Rworksheet4c/traffic.csv")
obs <- nrow(traffic)
vrbls <- names(traffic)
obs
```

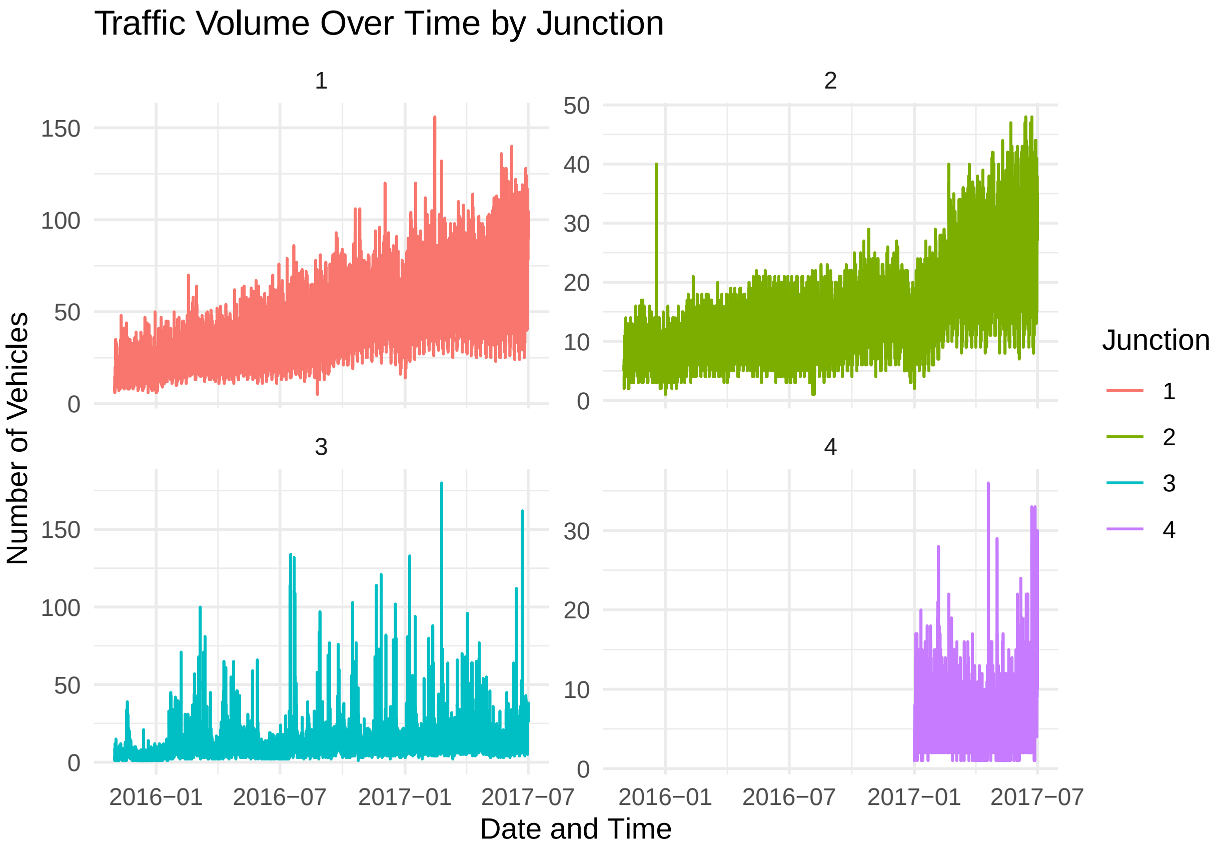
```
## [1] 48120
```

```
vrbls
```

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

#b.

```
junctions_data <- split(traffic, traffic$Junction)


traffic$DateTime <- as.POSIXct(traffic$DateTime, format="%Y-%m-%d %H:%M:%S")

ggplot(traffic, aes(x = DateTime, y = Vehicles, color = factor(Junction))) +
  geom_line() +
  labs(title = "Traffic Volume Over Time by Junction",
       x = "Date and Time",
       y = "Number of Vehicles",
       color = "Junction") +
  theme_minimal() +
  facet_wrap(~ Junction, scales = "free_y")
```



#7. #a.

```
library(openxlsx)
library(readxl)

alexa <- read.xlsx("/cloud/project/Rworksheet4c/alexa_file.xlsx")
observation <- nrow(alexa)
columns <- ncol(alexa)
observation
```

```
## [1] 3150
```

```
columns
```

```
## [1] 5
```

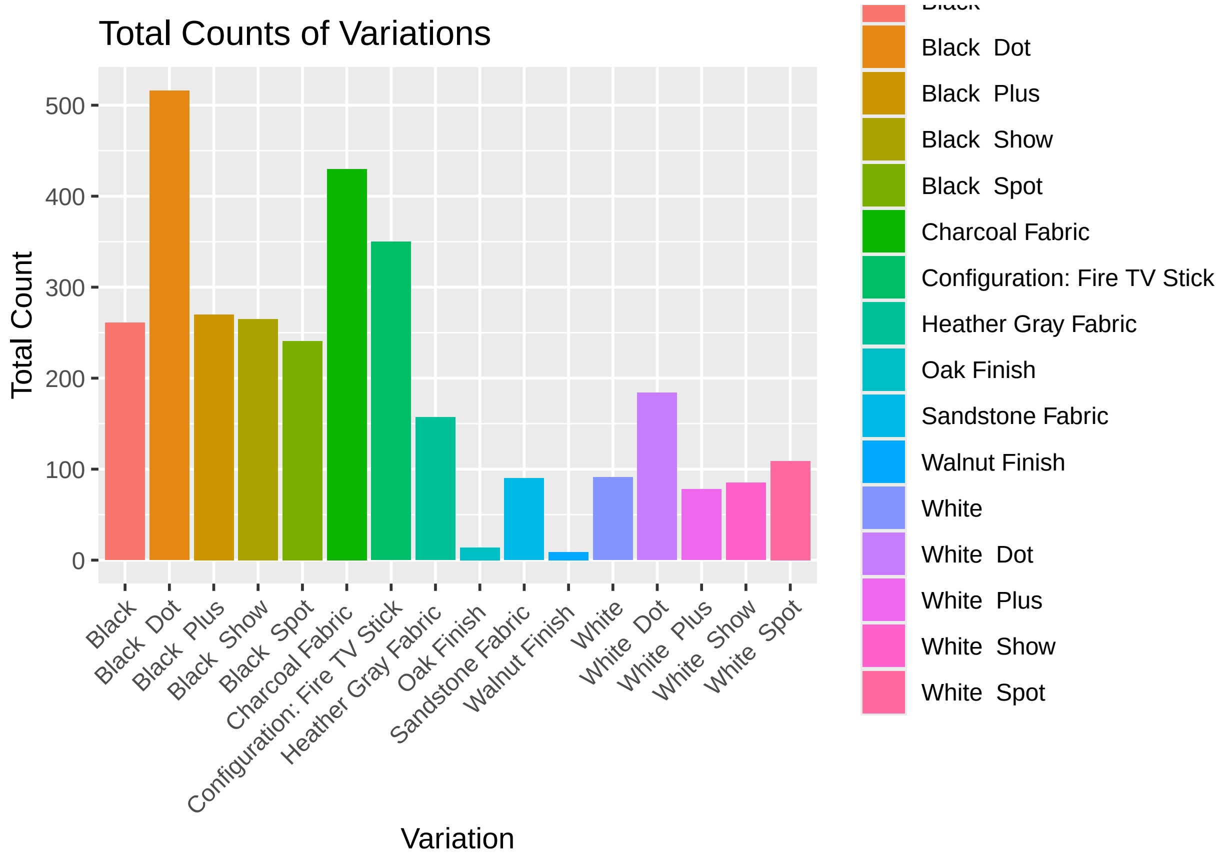
#b.

```
variation_counts <- alexa %>%
  group_by(variation) %>%
  summarise(total = n())
variation_counts
```

```
## # A tibble: 16 x 2
##    variation                      total
##    <chr>                          <int>
##  1 "Black"                          261
##  2 "Black  Dot"                     516
##  3 "Black  Plus"                    270
##  4 "Black  Show"                    265
##  5 "Black  Spot"                    241
##  6 "Charcoal Fabric "               430
##  7 "Configuration: Fire TV Stick"   350
##  8 "Heather Gray Fabric "           157
##  9 "Oak Finish "                     14
## 10 "Sandstone Fabric "               90
## 11 "Walnut Finish "                   9
## 12 "White"                           91
## 13 "White  Dot"                     184
## 14 "White  Plus"                     78
## 15 "White  Show"                     85
## 16 "White  Spot"                    109
```

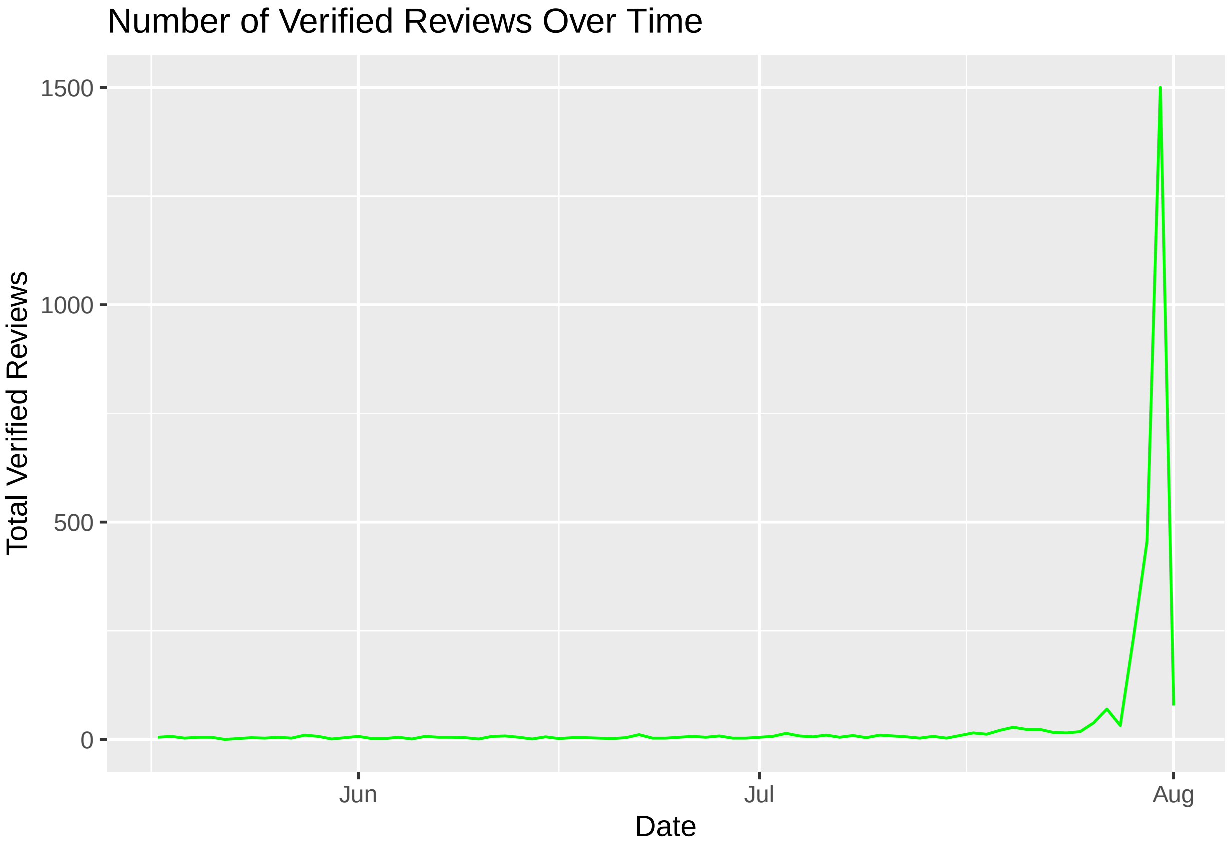
#c. There are only 3 variations that has more counts and 2 lowest counts than the others.

```
ggplot(variation_counts, aes(x = variation, y = total, fill = variation)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Counts of Variations",
       x = "Variation",
       y = "Total Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



#4d.

```
library(dplyr)
library(ggplot2)

alexa$date <- as.Date(alexa$date)

daily_reviews <- alexa %>%
  group_by(date) %>%
  summarise(total_verified_reviews = sum(feedback))

ggplot(daily_reviews, aes(x = date, y = total_verified_reviews)) +
  geom_line(color = "green") +
  labs(title = "Number of Verified Reviews Over Time",
       x = "Date",
       y = "Total Verified Reviews")
```

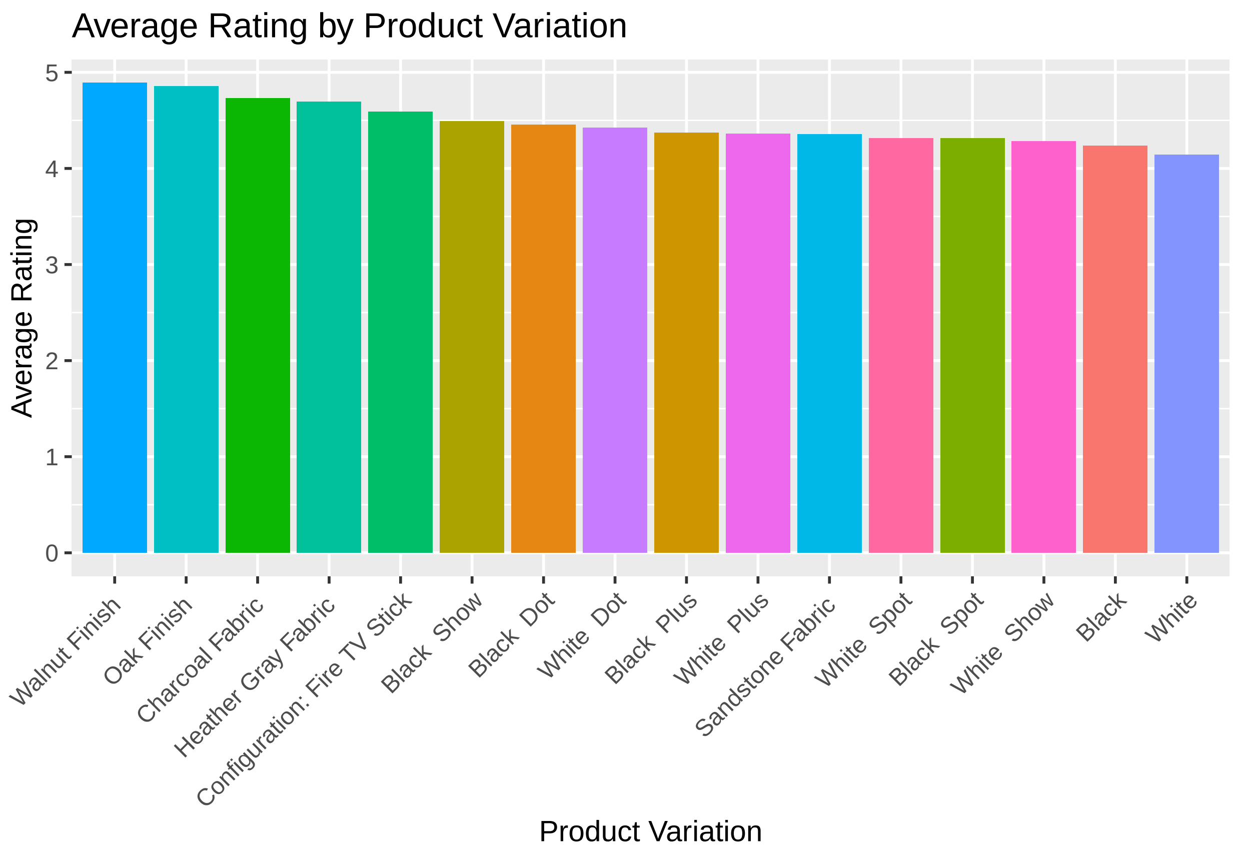


#e.

```
library(dplyr)
library(ggplot2)

variation_ratings <- alexa %>%
  group_by(variation) %>%
  summarise(average_rating = mean(rating, na.rm = TRUE)) %>%

arrange(desc(average_rating))

variation_ratings
```

```
## # A tibble: 16 x 2
##    variation                          average_rating
##    <chr>                                       <dbl>
##  1 "Walnut Finish "                             4.89
##  2 "Oak Finish "                                4.86
##  3 "Charcoal Fabric "                           4.73
##  4 "Heather Gray Fabric "                     4.69
##  5 "Configuration: Fire TV Stick"             4.59
##  6 "Black  Show"                              4.49
##  7 "Black  Dot"                               4.45
##  8 "White  Dot"                               4.42
##  9 "Black  Plus"                              4.37
## 10 "White  Plus"                              4.36
## 11 "Sandstone Fabric "                        4.36
## 12 "White  Spot"                              4.31
## 13 "Black  Spot"                              4.31
## 14 "White  Show"                              4.28
## 15 "Black"                                    4.23
## 16 "White"                                    4.14
```

```r
ggplot(variation_ratings, aes(x = reorder(variation, -average_rating), y = average_rating, fill = varia
  geom_bar(stat = "identity") +
  labs(title = "Average Rating by Product Variation",
       x = "Product Variation",
       y = "Average Rating") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  guides(fill = "none")
```


Average Rating by Product Variation
Average Rating
Walnut Finish
Oak Finish
Charcoal Fabric
Heather Gray Fabric
Configuration: Fire TV Stick
Black Show
Black Dot
White Dot
Black Plus
White Plus
Sandstone Fabric
White Spot
Black Spot
White Show
Black
White
Product Variation