RWorksheet_Cautivar#4c.Rmd

James Clark Cautivar

2024-11-1

1. Use the dataset mpg

a. Show your solutions on how to import a csv file into the environment.

```
library(ggplot2)

write.csv(mpg, "mpg.csv", row.names = FALSE)
mpg_data <- read.csv("mpg.csv")
str(mpg_data)
```

```
## 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

b. Which variables from mpg dataset are categorical? The variables that are categorical in the mpg dataset are manufacturer, model, year, cyl, trans, drv, fl, and class.

c. Which are continuous variables? The continuous variables are displ, cty, and hwy.

2. Which manufacturer has the most models in this data set? Which model has the most variations? Show your answer.

a. Group the manufacturers and find the unique models. Show your codes and result.

```
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
```

```
manufacturerModelCount <- mpg %>%
  group_by(manufacturer) %>%
  summarize(num_models = n_distinct(model)) %>%
  arrange(desc(num_models))

manufacturerModelCount
```

```
## # A tibble: 15 x 2
##    manufacturer num_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
```

```
modelVariationCount <- table(mpg$model)
modelVariationCount [modelVariationCount  == max(modelVariationCount )]
```

```
## caravan 2wd
##          11
```

The manufacturer that has the most models in this data set is toyota which has 6 models.

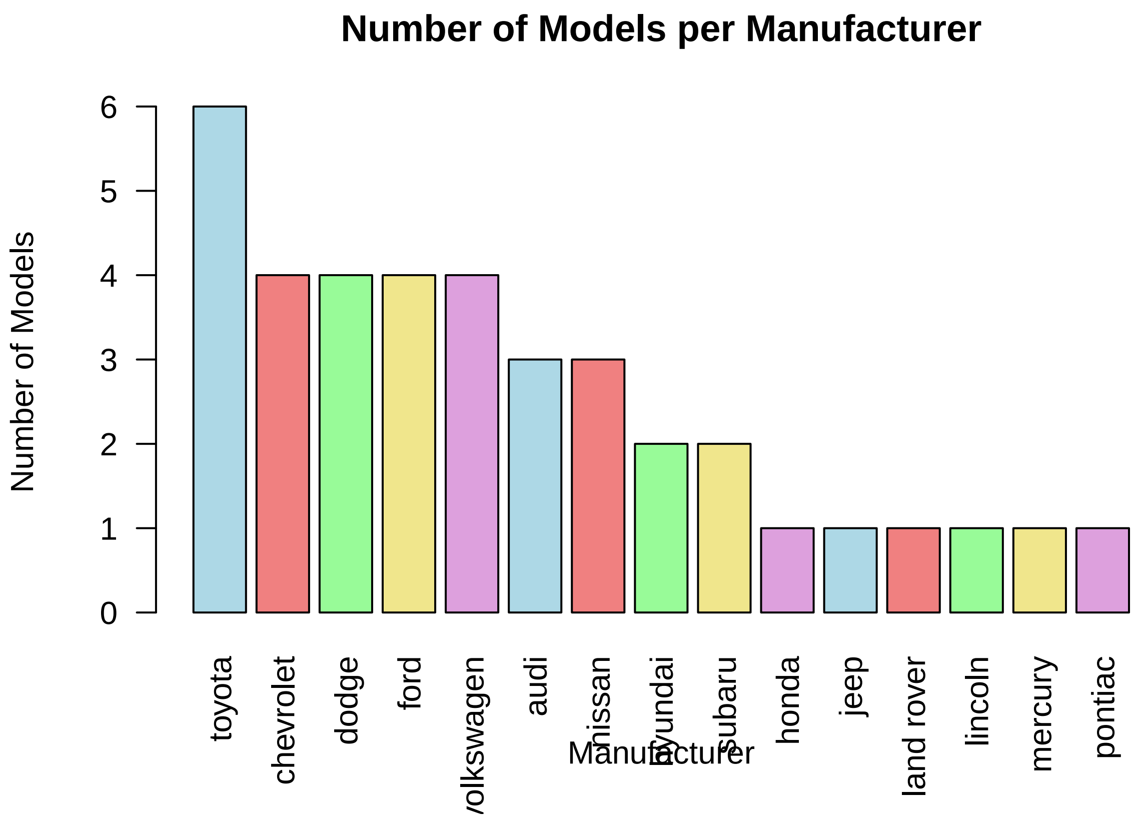
The model that has the most variations is the caravan 2wd which has 11C variarions.

b. Graph the result by using plot() and ggplot(). Write the codes and its result.

```
library(dplyr)
library(ggplot2)

manufacturer_counts <- setNames(manufacturerModelCount$num_models, manufacturerModelCount$manufacturer)

barplot(manufacturer_counts,
        main = "Number of Models per Manufacturer",
        xlab = "Manufacturer",
        ylab = "Number of Models",
        col = c("lightblue", "lightcoral", "palegreen", "khaki", "plum"),
        las = 2)
```

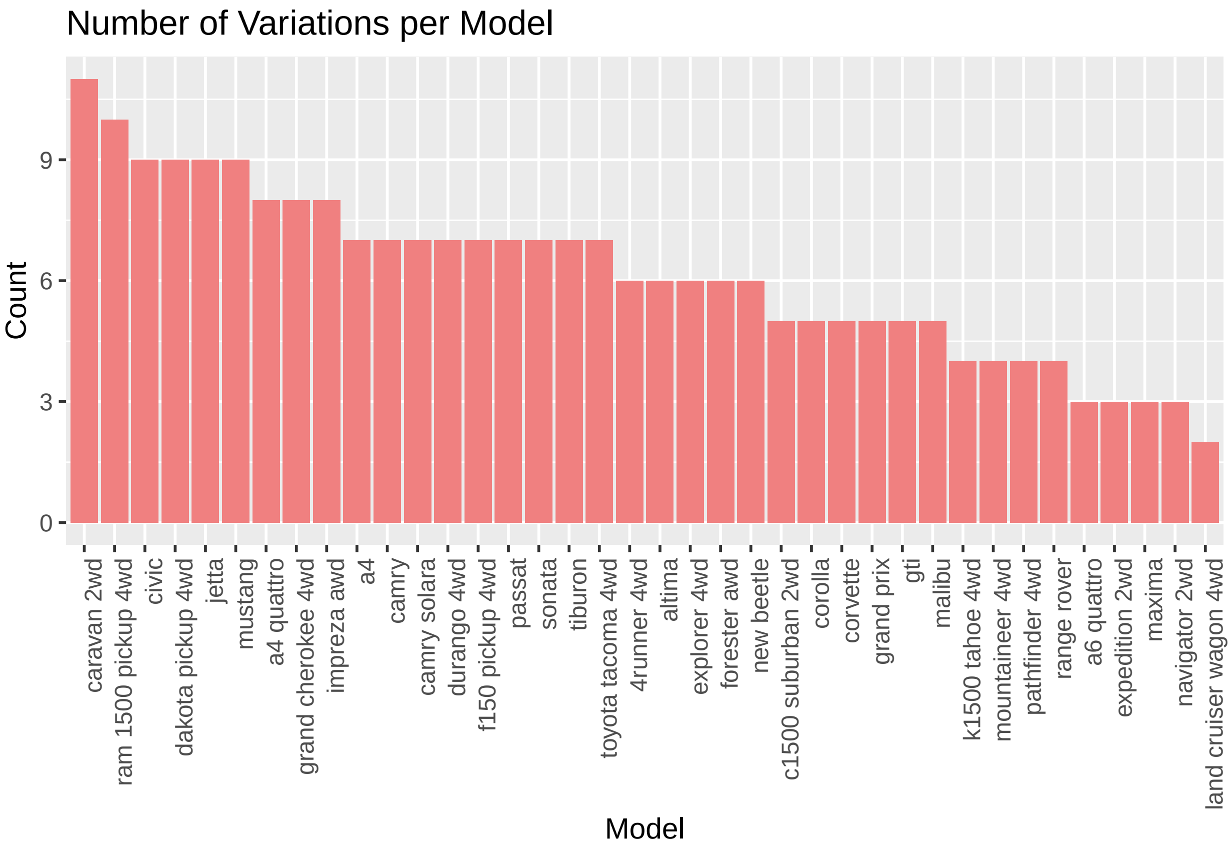


```
modelVariationCount <- mpg %>%
  group_by(model) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

print(modelVariationCount)
```

```
## # A tibble: 38 x 2
##    model                 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
```

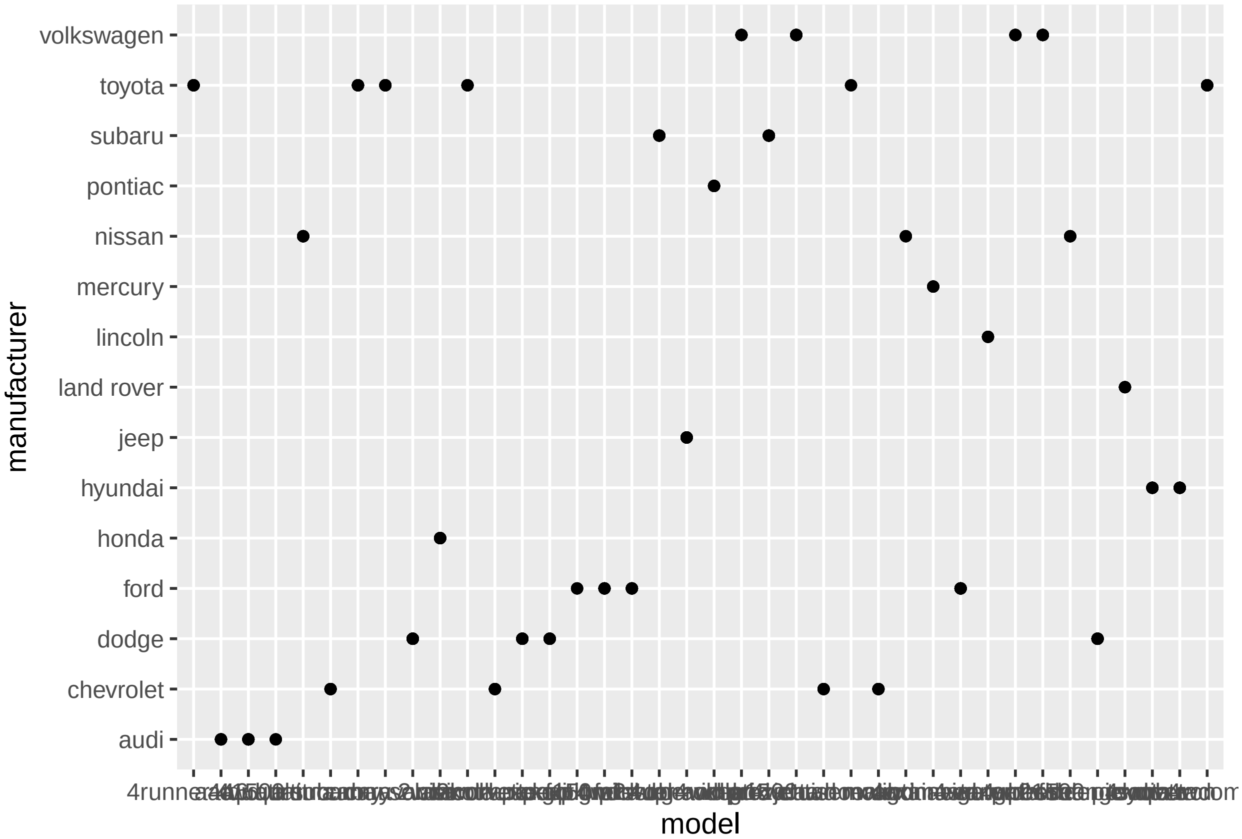
```
ggplot(modelVariationCount, aes(x = reorder(model, -count), y = count)) +
  geom_bar(stat = "identity", fill = "lightcoral") +
  labs(title = "Number of Variations per Model", x = "Model", y = "Count") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



2. Same dataset will be used. You are going to show the relationship of the modeland the manufacturer.

a. What does ggplot(mpg, aes(model, manufacturer)) + geom_point() show?

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



It shows a scatter plot of the mpg models and manufacturers.

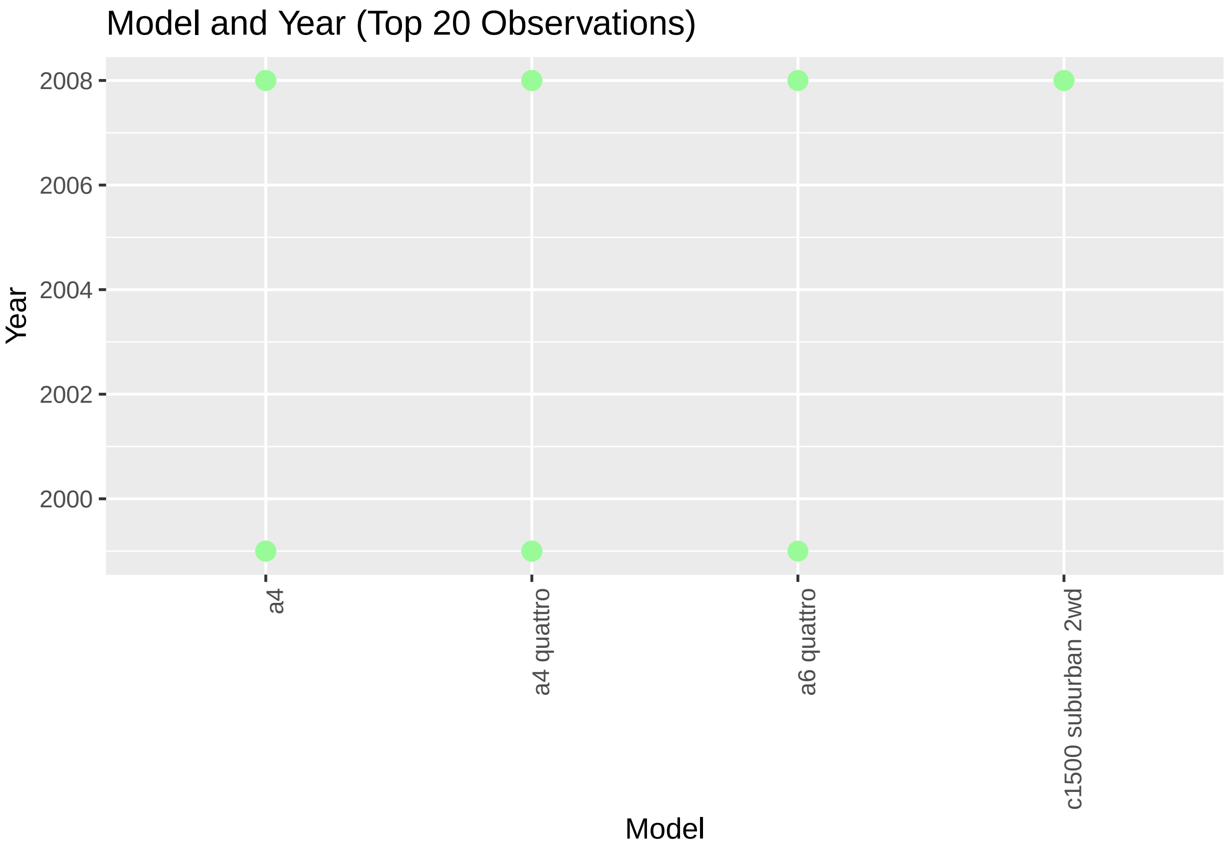
b. For you, is it useful? If not, how could you modify the data to make it more informative? For me, it's not that useful due to the visualization of it. It's not clear and the labels are covering each other which is confusing. To make it more informative, i'll change it into a bar graph and fix the labels to make it organized.

3. Plot the model and the year using ggplot(). Use only the top 20 observations. Write the codes and its results.

```r
library(ggplot2)

top20Obs <- mpg[1:20, ]

ggplot(top20Obs, aes(x = model, y = year)) +
  geom_point(color = "palegreen", size = 3) +
  labs(title = "Model and Year (Top 20 Observations)",
       x = "Model",
       y = "Year") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



4. Using the pipe (%>%), group the model and get the number of cars per model. Show codes and its result

```
library(dplyr)

carCounts <- mpg %>%
  group_by(model) %>%
  summarise(count = n())

carCounts
```

```
## # A tibble: 38 x 2
##    model              count
##    <chr>              <int>
##  1 4runner 4wd            6
##  2 a4                     7
##  3 a4 quattro             8
##  4 a6 quattro             3
##  5 altima                 6
##  6 c1500 suburban 2wd     5
##  7 camry                  7
##  8 camry solara           7
##  9 caravan 2wd           11
## 10 civic                  9
## # i 28 more rows
```

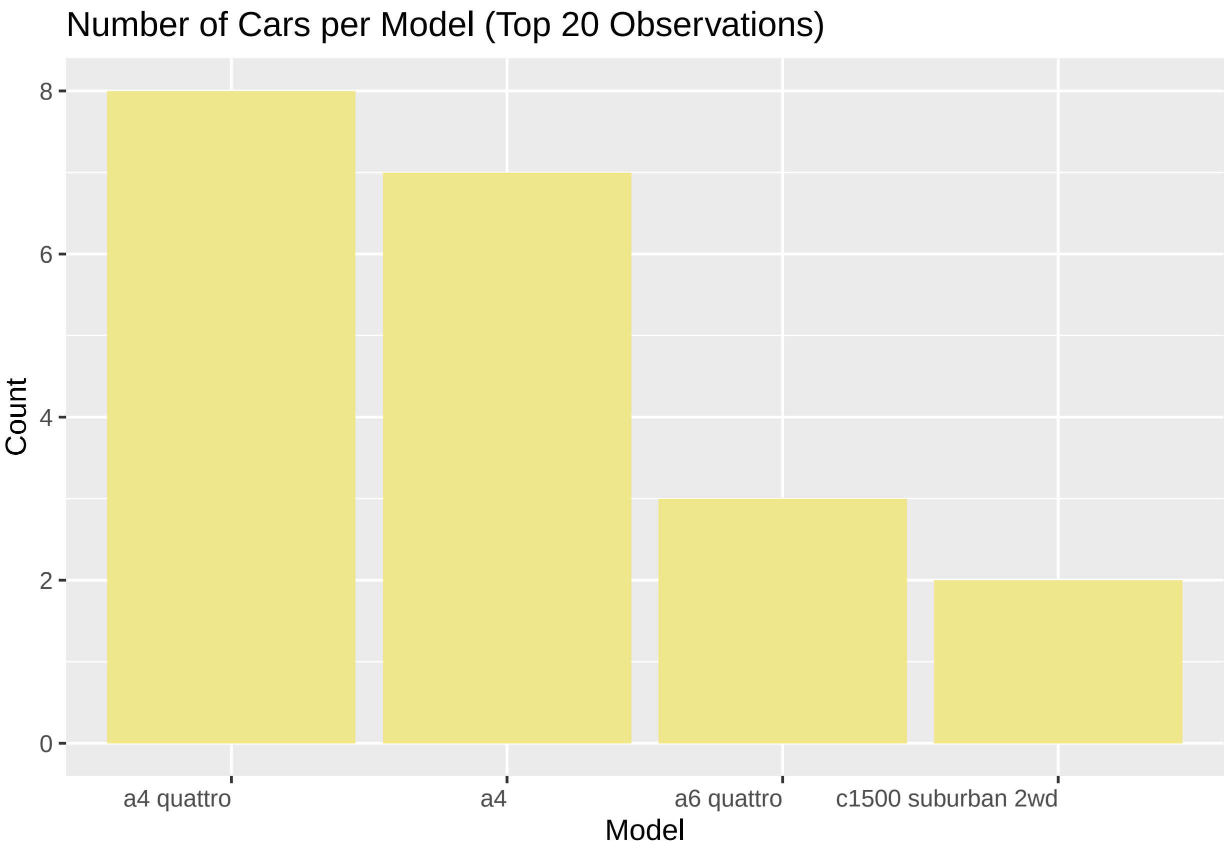
a. Plot using geom_bar() using the top 20 observations only. The graphs shoudl have a title, labels and colors. Show code and results.

```r
library(ggplot2)
library(dplyr)

top20Obs <- mpg[1:20, ]

carCounts20 <- top20Obs %>%
  group_by(model) %>%
  summarise(count = n())

ggplot(carCounts20, aes(x = reorder(model, -count), y = count)) +
  geom_bar(stat = "identity", fill = "khaki") +
  labs(title = "Number of Cars per Model (Top 20 Observations)",
       x = "Model",
       y = "Count") +
  theme(axis.text.x = element_text(hjust = 1))
```



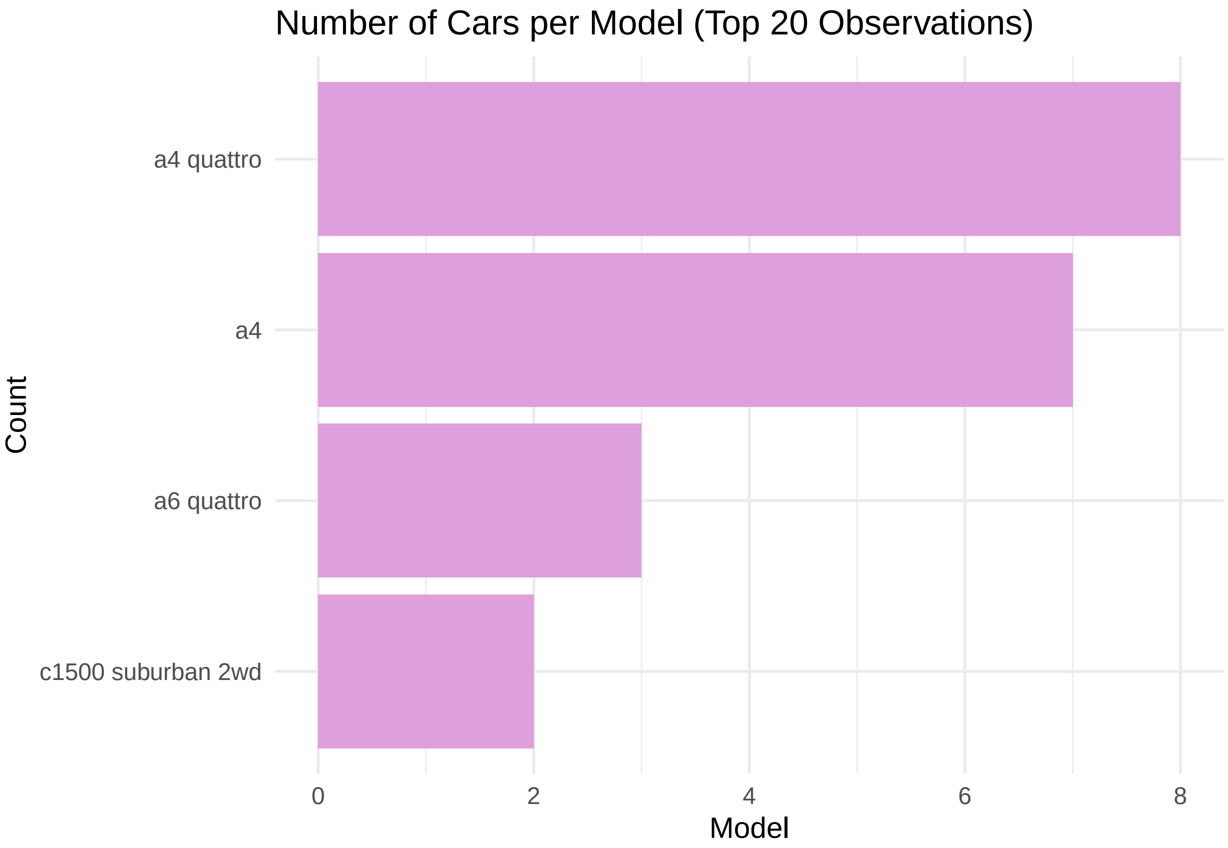
b. Plot using the geom_bar() + coord_flip() just like what is shown below. Show codes and its result.

```r
library(ggplot2)
library(dplyr)

top20Obs <- mpg[1:20, ]

carCounts20 <- top20Obs %>%
  group_by(model) %>%
  summarise(count = n())

ggplot(carCounts20, aes(x = reorder(model, count), y = count)) +
  geom_bar(stat = "identity", fill = "plum") +
  coord_flip() +
  labs(title = "Number of Cars per Model (Top 20 Observations)",
       x = "Count",
       y = "Model") +
  theme_minimal()
```


Number of Cars per Model (Top 20 Observations)
Count
a4 quattro
a4
a6 quattro
c1500 suburban 2wd
Model