RWorksheet_calvario#4c.Rmd

Jolien

2024-11-19

1. Use the dataset mpg

a. Solutions on how to import a csv file into the environment.

```
library(ggplot2)

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

```
## 'data.frame':    234 obs. of  12 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 categorical variables in the mpg dataset are manufacturer, model, trans, drv, fl, and class. These represent distinct groups or categories, such as car brand, transmission type, drivetrain, fuel type, and car class.

c. Which are continuous variables?

The continuous variables in the mpg dataset are displ, cty, hwy, and cyl, as they represent measurable numerical values.

2.1. 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
```

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

manufacturer_model
```

```
## # A tibble: 15 x 2
##    manufacturer model_num
##    <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
```

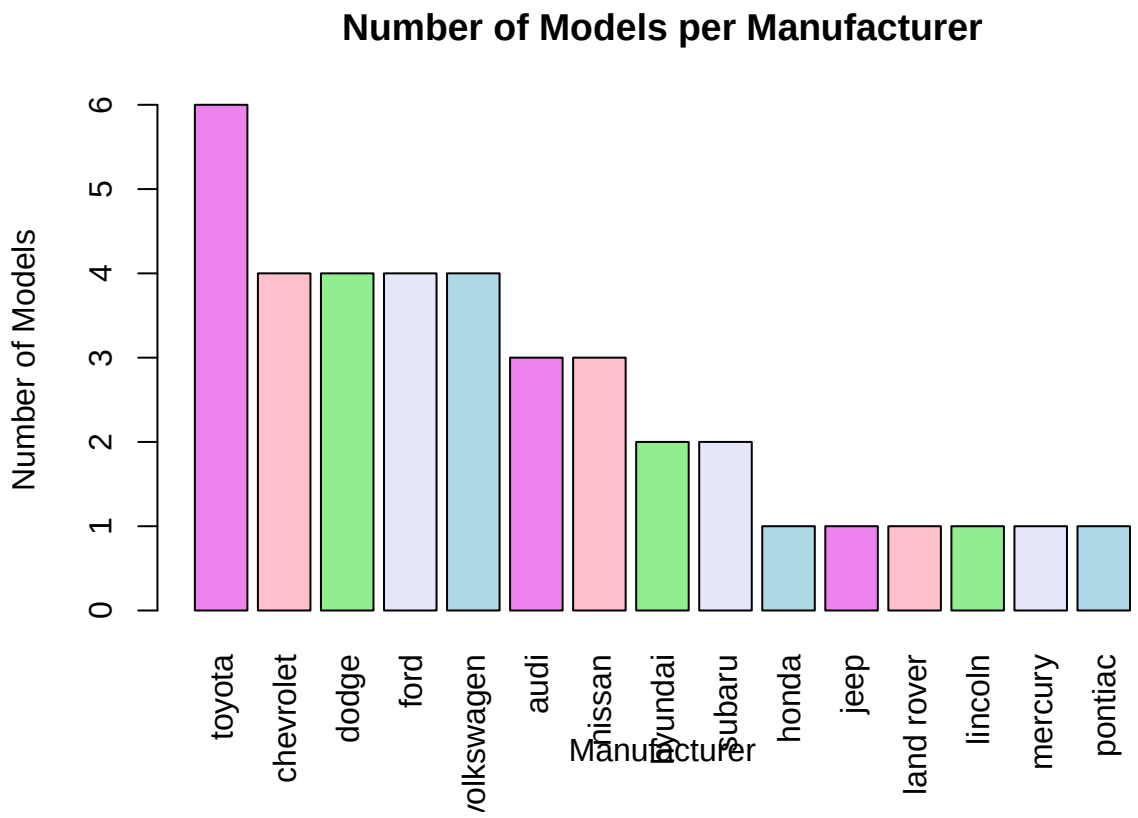
```
variations_num <- table(mpg$model)
variations_num [variations_num == max(variations_num)]
```

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

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

```
manufacturer_data <- setNames(
  manufacturer_model$model_num,
  manufacturer_model$manufacturer
  )

barplot(manufacturer_data,
        main = "Number of Models per Manufacturer",
        xlab = "Manufacturer",
        ylab = "Number of Models",
        col = c("violet", "pink", "lightgreen", "lavender", "lightblue"),
        las = 3)
```

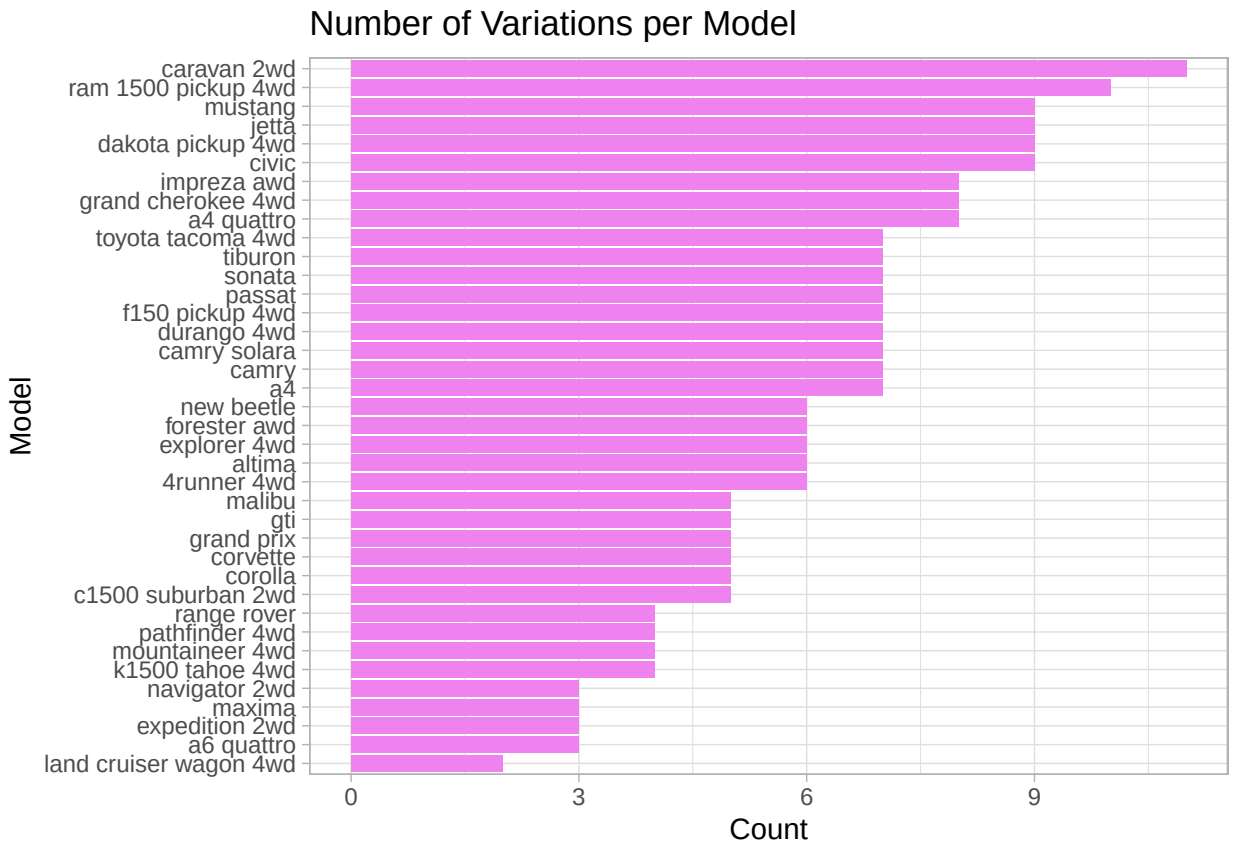


```r
variations_num <- mpg %>%
  group_by(model) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

variations_num
```

```
## # 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
```

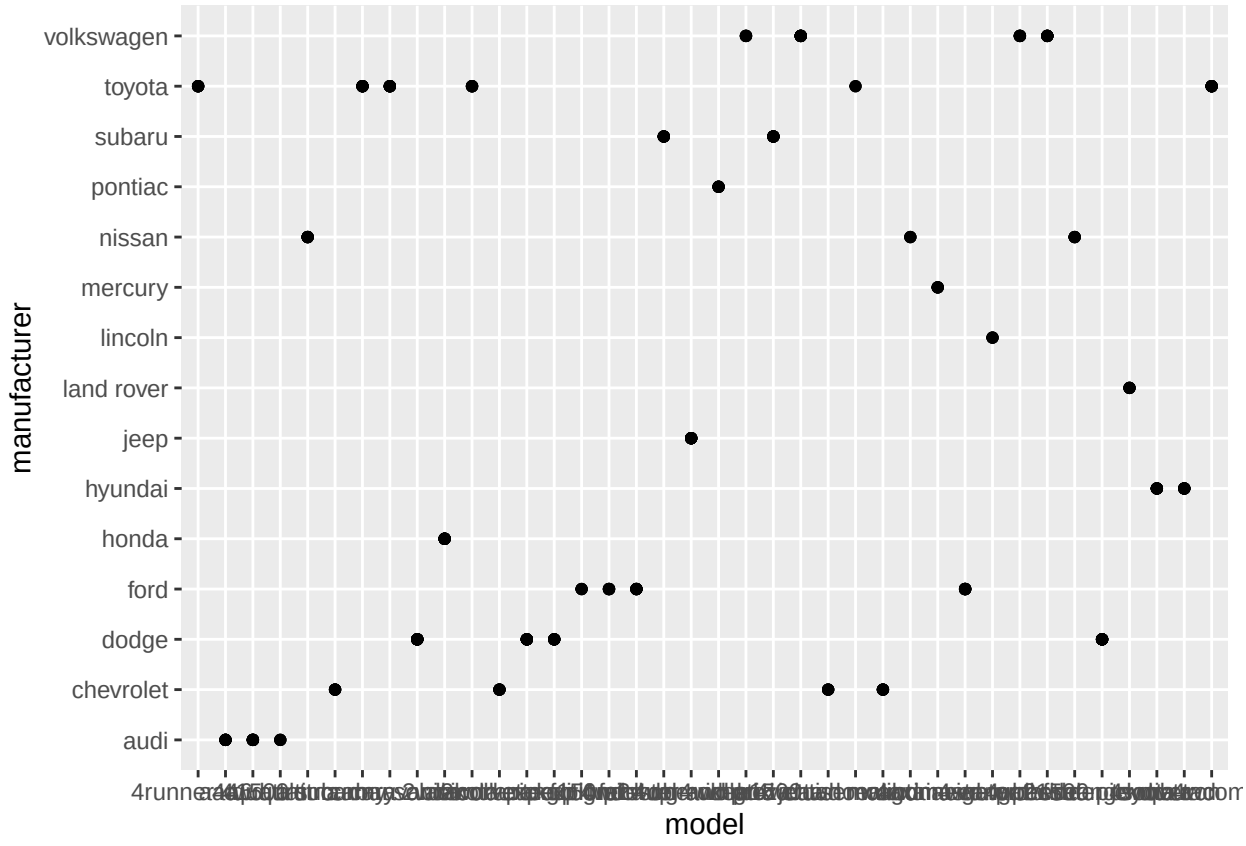
```r
ggplot(variations_num, aes(x = reorder(model, count), y = count)) +
  geom_bar(stat = "identity", fill = "violet") +
  coord_flip() +
  labs(title = "Number of Variations per Model", x = "Model", y = "Count") +
  theme_light()
```



2.2. 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()
```



ggplot(mpg, aes(model, manufacturer)) + geom_point() creates a scatter plot showing the relationship between car models (model) on the x-axis and manufacturers (manufacturer) on the y-axis, with each point representing a car observation.

b. For you, is it useful? If not, how could you modify the data to make it more informative?

In its current form, this plot isn't very useful as it doesn't effectively visualize the relationship between car models and manufacturers. Since both variables are categorical, a scatter plot isn't the most appropriate way to represent this relationship.

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

```
ggplot(head(mpg_data, 20), aes(x = model, y = year)) +
  geom_point() +
  labs(title = "Model vs Year (Top 20 Observations)", x = "Model", y = "Year") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```