RWorksheet_Subosa#4c.Rmd

Gian Adree Subosa

2024-11-09

#1. Use dataset mpg

```
#a.) Show your solutions on how to import a csv file into the environment.
mpg_file <- read.csv("mpg.csv")
head(mpg_file)
```

```
##   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
```

```
str(mpg_file)
```

```
## '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?

#According to my observation, the variables from the mpg dataset that are seemly categoral are the foll
# - manufacturer: that indicates the manufacturer name
# - model: that indicates the model name
# - trans: the type of transmission system the vehicle uses
# - drv: indicating whether the vehicle is front-wheel drive (f), rear-wheel drive (r), or four-wheel d
# - fl: the type of fuel
# - class: the general category or type of the vehicle
```

```
#c.) Which are continuous variables?

#The following are the continuous variables in the mpg dataset:
# - dspl: the engine displacement in liters
# - year: the year of manufacture
# - cyl: the number of cylinders
# - cty: the city miles per hour
# - hwy: the highway miles per gallon
```

#2.1 Which manufacturer has the most models in this data set? Which model has the most variations?

```
#a.) Group the manufacturers and find the unique models
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
```

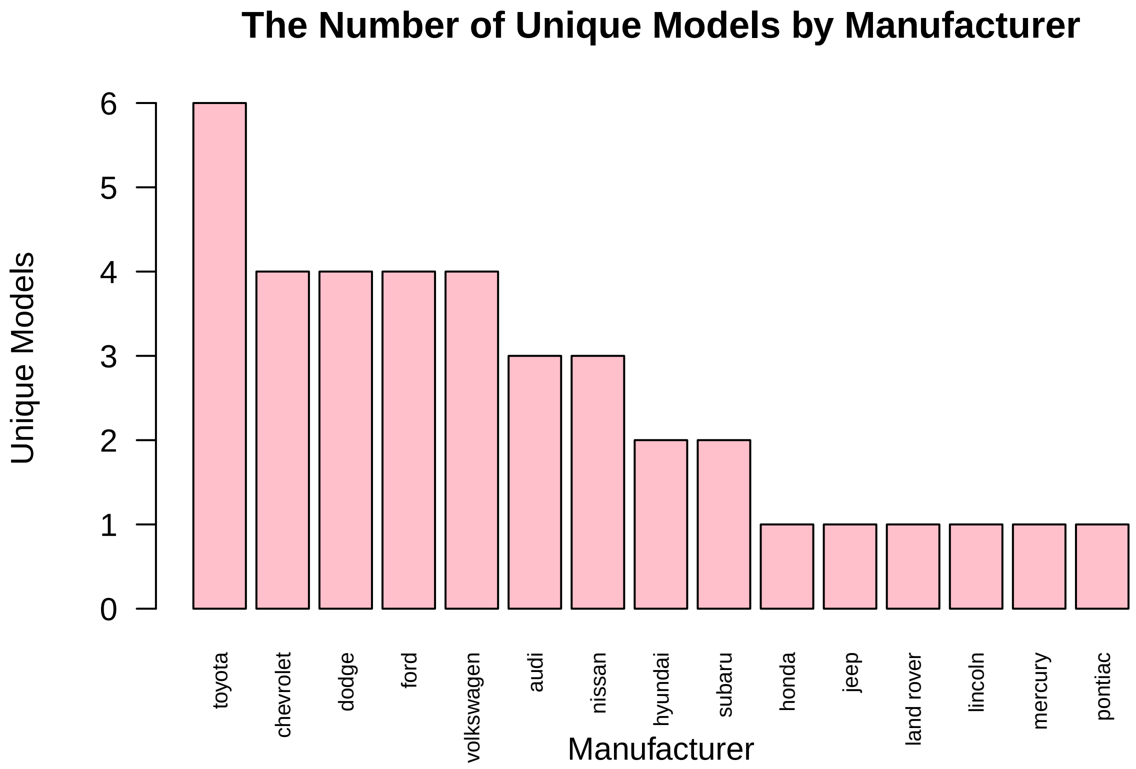
```
data(mpg, package = "ggplot2")

modelmanufacturer <- mpg %>%
  group_by(manufacturer) %>%
  summarize(unique_models = n_distinct(model)) %>%
  arrange(desc(unique_models))

print(modelmanufacturer)
```

```
## # 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
```

```
modelvariations <- mpg %>%
  group_by(model) %>%
  summarize(variations = n()) %>%
  arrange(desc(variations))

print(modelvariations)
```

```
## # 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
```

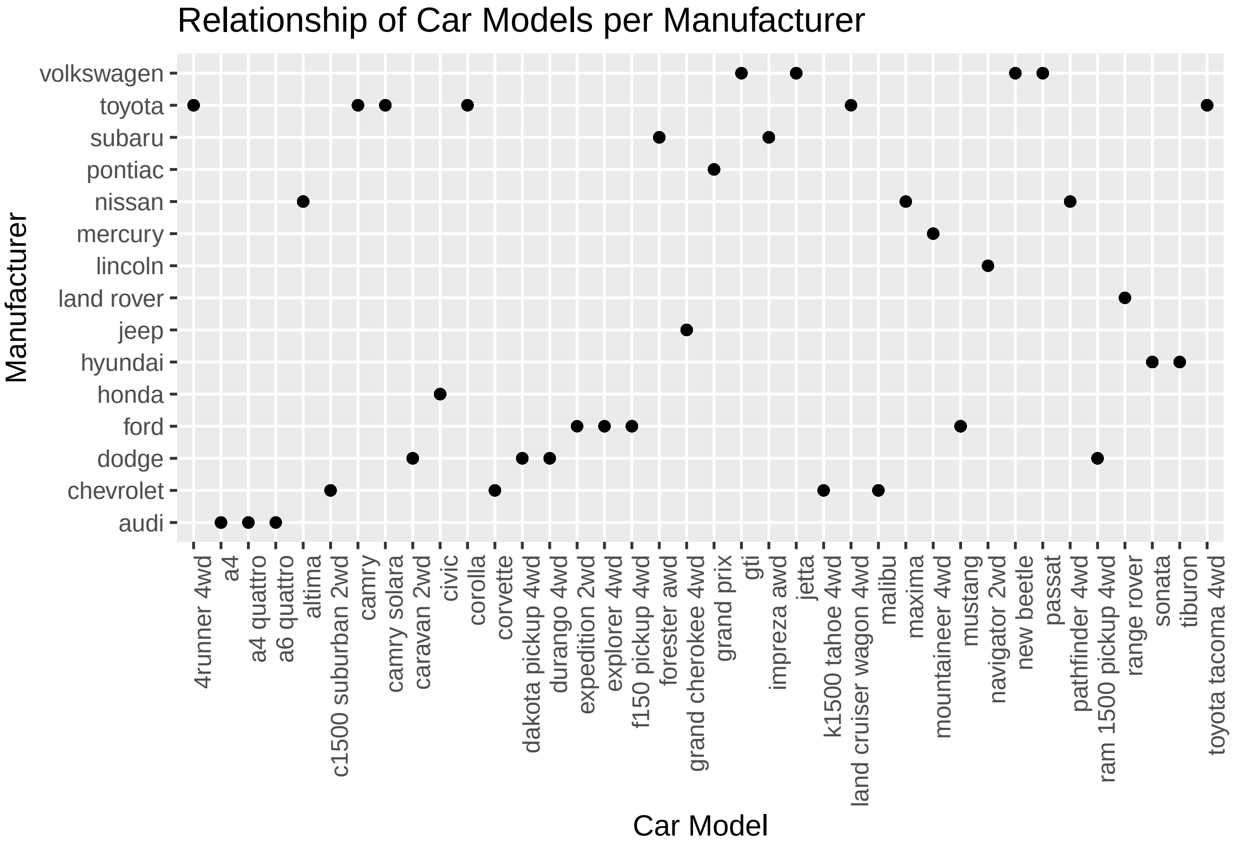
```
#b.) Graph the result by using plot() and ggplot()
modelmanufacturer$manufacturer <- factor(modelmanufacturer$manufacturer, levels = modelmanufacturer$manu

barplot(modelmanufacturer$unique_models,
        names.arg = modelmanufacturer$manufacturer,
        main = "The Number of Unique Models by Manufacturer",
        xlab = "Manufacturer",
        ylab = "Unique Models",
        col = "pink",
        las = 2,
        cex.names = 0.7)
```



#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?
library(ggplot2)
ggplot(mpg, aes(x = model, y = manufacturer)) + geom_point() + theme(axis.text.x = element_text(angle =
```

```r
  labs(title = "Relationship of Car Models per Manufacturer",
       x = "Car Model",
       y = "Manufacturer"
  )
```

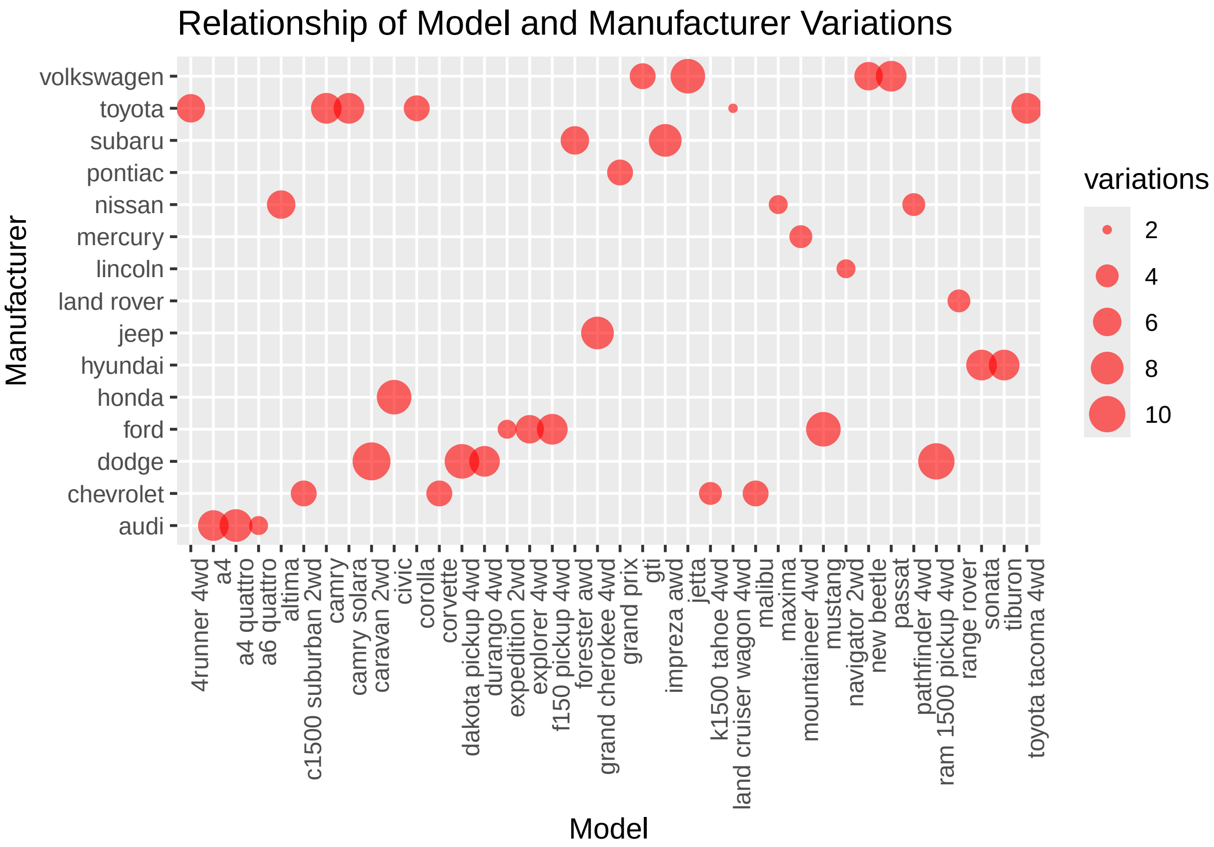


```r
#The data is presented in the form of a scatter plot, with a point at each model and manufacturer combi
```

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

#Because of the overlapping points, this visualization is currently ineffective. A better approach coul

modelmanufacturer <- mpg %>%
  group_by(model, manufacturer) %>%
  summarize(variations = n())
```

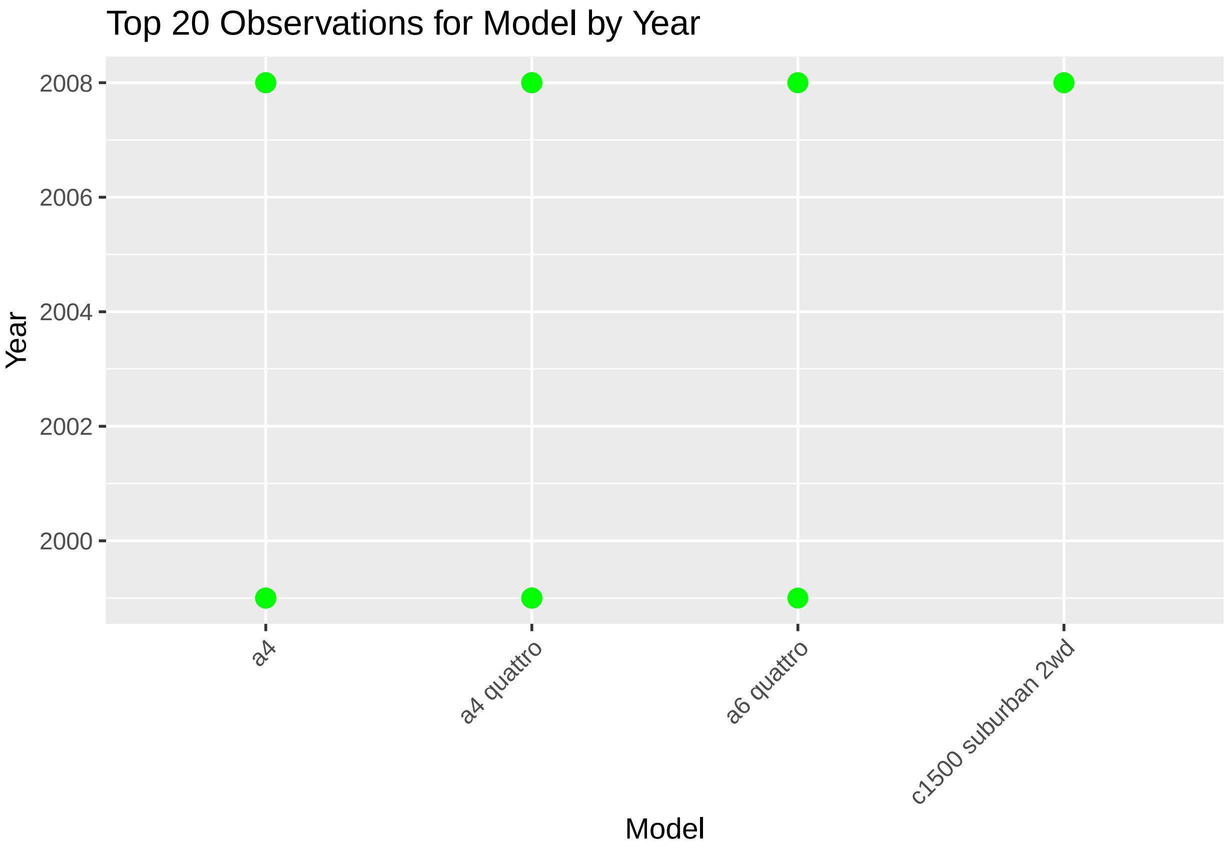
```
## `summarise()` has grouped output by 'model'. You can override using the
## `.groups` argument.
```

```r
ggplot(modelmanufacturer, aes(x = model, y = manufacturer, size = variations)) +
  geom_point(color = "red", alpha = 0.6) +
  labs(title = "Relationship of Model and Manufacturer Variations", x = "Model", y = "Manufacturer") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



#3. Plot the model and the year using ggplot(). Use only the top 20 observations.

```
library(ggplot2)
top_20_mpg <- head(mpg, 20)

ggplot(top_20_mpg, aes(x = model, y = year)) +
  geom_point(color = "green", size = 3) +
  labs(title = "Top 20 Observations for Model by Year", x = "Model", y = "Year") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Top 20 Observations for Model by Year
Year
2008
2006
2004
2002
2000
a4
a4 quattro
a6 quattro
c1500 suburban 2wd
Model