RWorksheet_Malayas#4c

Andrew Miguel M. Malayas BSIT2A

2024-11-01

1. Use the dataset mpg

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

```
library(readr)
mpg <- read_csv("mpg.csv")
```

```
## New names:
## Rows: 234 Columns: 12
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (6): manufacturer, model, trans, drv, fl, class dbl (6): ...1, displ, year,
## cyl, cty, hwy
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
mpg
```

```
## # A tibble: 234 x 12
##     ...1 manufacturer model      displ  year   cyl trans drv     cty   hwy fl
##    <dbl> <chr>        <chr>      <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <chr>
##  1     1 audi         a4           1.8  1999     4 auto~ f        18    29 p
##  2     2 audi         a4           1.8  1999     4 manu~ f        21    29 p
##  3     3 audi         a4           2    2008     4 manu~ f        20    31 p
##  4     4 audi         a4           2    2008     4 auto~ f        21    30 p
##  5     5 audi         a4           2.8  1999     6 auto~ f        16    26 p
##  6     6 audi         a4           2.8  1999     6 manu~ f        18    26 p
##  7     7 audi         a4           3.1  2008     6 auto~ f        18    27 p
##  8     8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p
##  9     9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p
## 10    10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p
## # i 224 more rows
## # i 1 more variable: class <chr>
```

b. Which variables from mpg dataset are categorical?

manufacturer - The car manufacturer (e.g., Audi, Chevrolet). - model - trans - drv - cyl - fl - class - manufacturer

c. Which are continuous variables?

- displ
- year
- cyt
- hwy

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

```
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_count <- mpg %>%
  group_by(manufacturer) %>%
  summarise(model_count = n_distinct(model)) %>%
  arrange(desc(model_count))

most_models_manufacturer <- manufacturer_model_count[1, ]

model_variation_count <- mpg %>%
  group_by(model) %>%
  summarise(variation_count = n()) %>%
  arrange(desc(variation_count))

most_variations_model <- model_variation_count[1, ]

most_models_manufacturer
```

```
## # A tibble: 1 x 2
##   manufacturer model_count
##   <chr>              <int>
## 1 toyota                 6
```

```
most_variations_model
```

```
## # A tibble: 1 x 2
##   model       variation_count
##   <chr>                 <int>
## 1 caravan 2wd              11
```

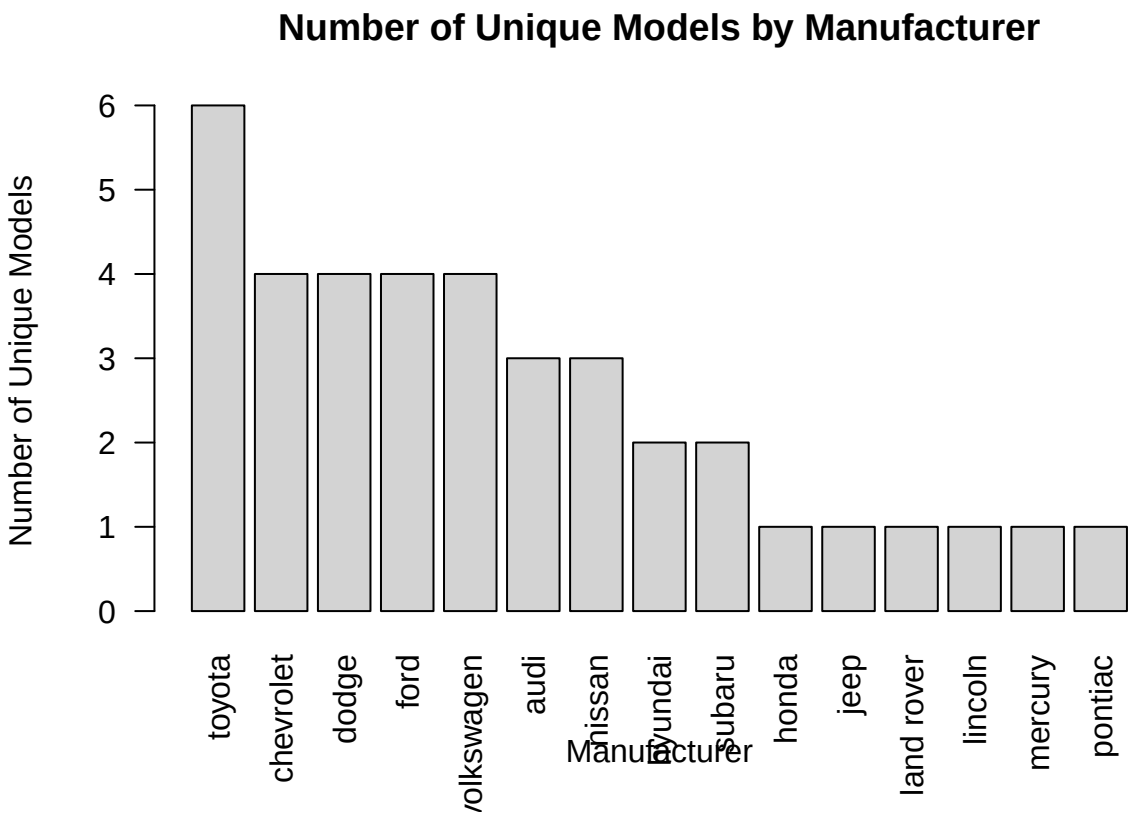
a. Group the manufacturers and find the unique models. Show your codes and result.

```r
unique_models <- mpg %>%
  group_by(manufacturer) %>%
  summarise(unique_models = list(unique(model))) %>%
  arrange(manufacturer)

print(unique_models)
```

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

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

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


barplot(model$unique_count,
        names.arg = model$manufacturer,
        las = 2,
        col = "lightgrey",
        main = "Number of Unique Models by Manufacturer",
        xlab = "Manufacturer",
        ylab = "Number of Unique Models")
```

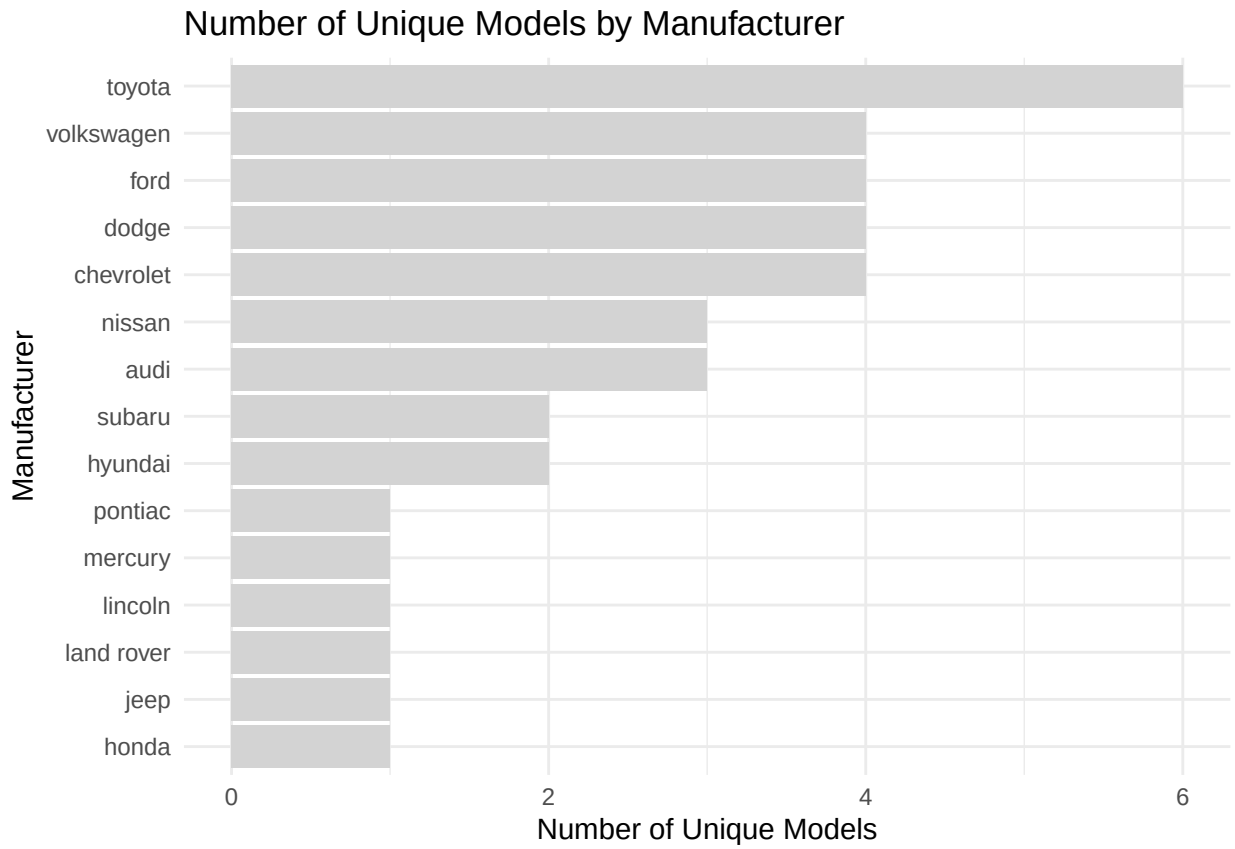



```
library(ggplot2)
```

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

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

```
ggplot(model, aes(x = reorder(manufacturer, unique_count), y = unique_count)) +
  geom_bar(stat = "identity", fill = "lightgrey") +
  coord_flip() +
  labs(title = "Number of Unique Models by Manufacturer",
       x = "Manufacturer",
       y = "Number of Unique Models") +
  theme_minimal()
```



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

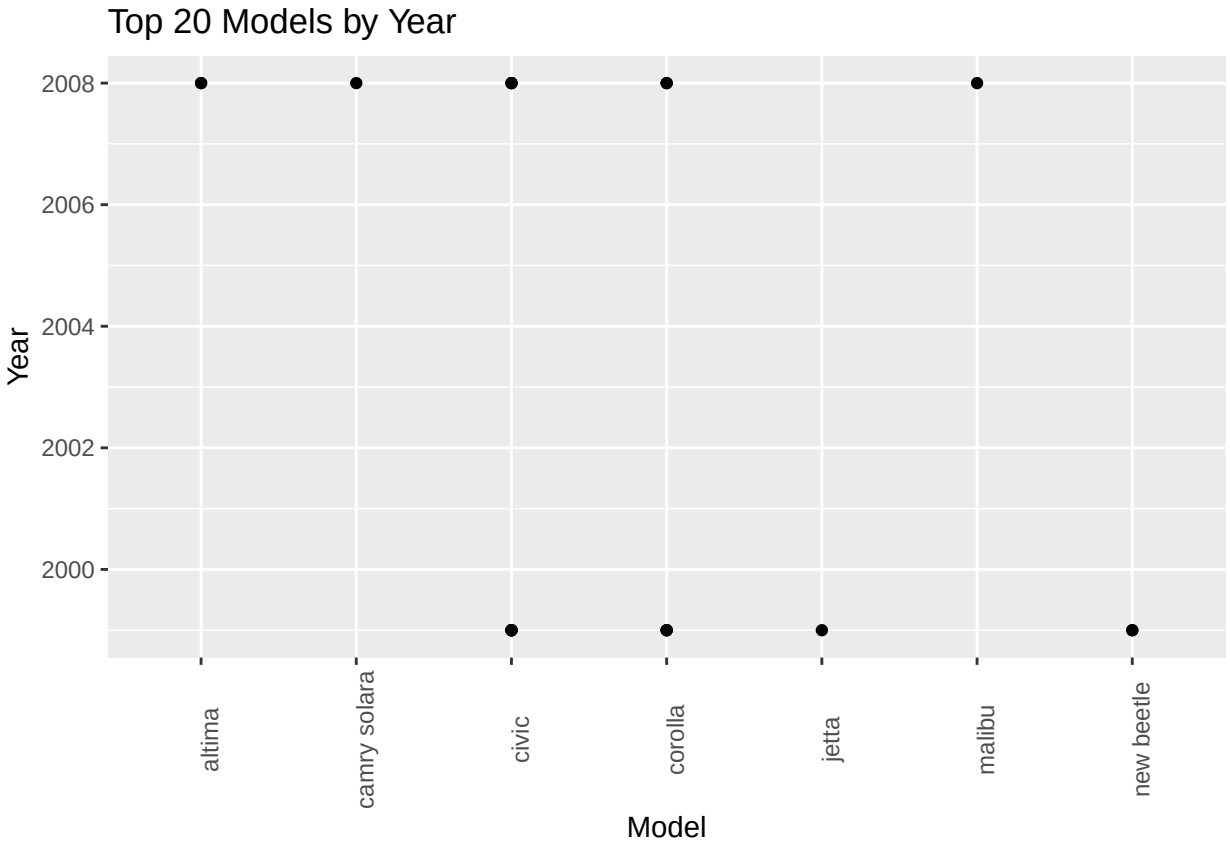
   a. What does ggplot(mpg, aes(model, manufacturer)) + geom_point() show? x-axis: represents the model. y-axis: represents the manufacturer. This code generates a scatter plot where each point represents a unique pairing of model and manufacturer. It illustrates which models are produced by each manufacturer, showing the distribution of models across different manufacturers.

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

As it stands, the plot isn't particularly helpful for analyzing the relationship between model and manufacturer because of the overplotting of points and the absence of numerical or continuous variables that could offer deeper insights.

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

```
Top_20 <- mpg %>%
  arrange(desc(cty)) %>%
  head(20)
ggplot(Top_20, aes(x = model, y = year)) +
  geom_point() +
  labs(title = "Top 20 Models by Year", x = "Model", y = "Year") +
  theme(axis.text.x = element_text(angle = 90, hjust = 0.5))
```


Top 20 Models by Year
Year
2008
2006
2004
2002
2000
altima
camry solara
civic
corolla
jetta
malibu
new beetle
Model