RWorksheet_Alpanghe#4c

Jersey Gabriel E. Alpanghe

2024-11-11

1. Use the dataset mpg A data frame with 234 rows and 11 variables:

• Download and open the mpg file. Upload it to your OWN environment a. Show your solutions on how to import a csv file into the environment.

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

b. Which variables from mpg dataset are categorical?

Categorical variables in mpg: manufacturer, model, trans, drv, fl, class.

c. Which are continuous variables?

Continuous variables in mpg: displ, year, cyl, cty, 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
```

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

model_variations <- mpg %>%
  group_by(model) %>%
  summarise(num_variations = n()) %>%
  arrange(desc(num_variations))

print(manufacturer_count)
```

```
## # 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
```

```
print(model_variations)
```

```
## # A tibble: 38 x 2
##    model               num_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(). Write the codes and its result.

```
library(ggplot2)
```

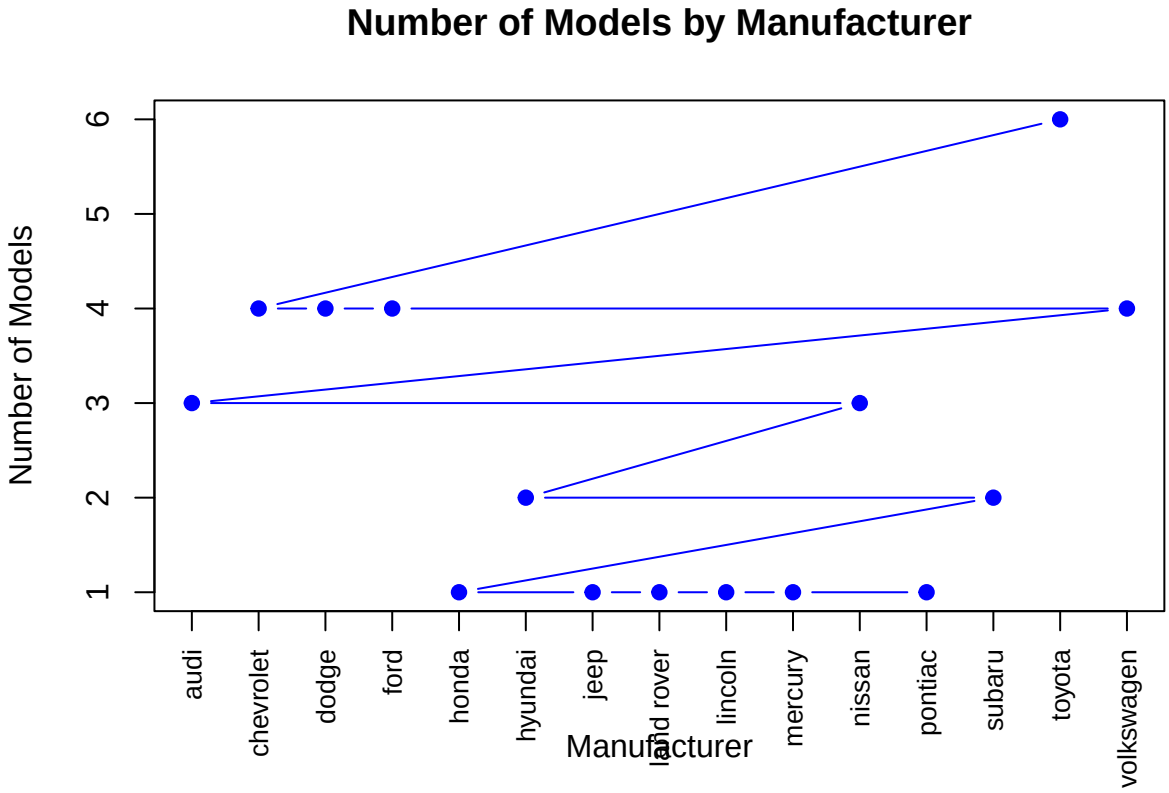
```
##
## Attaching package: 'ggplot2'
```

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

```
manufacturer_count$manufacturer <- factor(manufacturer_count$manufacturer)

plot(as.numeric(manufacturer_count$manufacturer), manufacturer_count$num_models,
     type = "b", pch = 19, col = "blue",
     main = "Number of Models by Manufacturer",
     xlab = "Manufacturer", ylab = "Number of Models",
     xaxt = "n")

axis(1, at = 1:length(manufacturer_count$manufacturer),
     labels = levels(manufacturer_count$manufacturer), las = 2, cex.axis = 0.8)
```



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

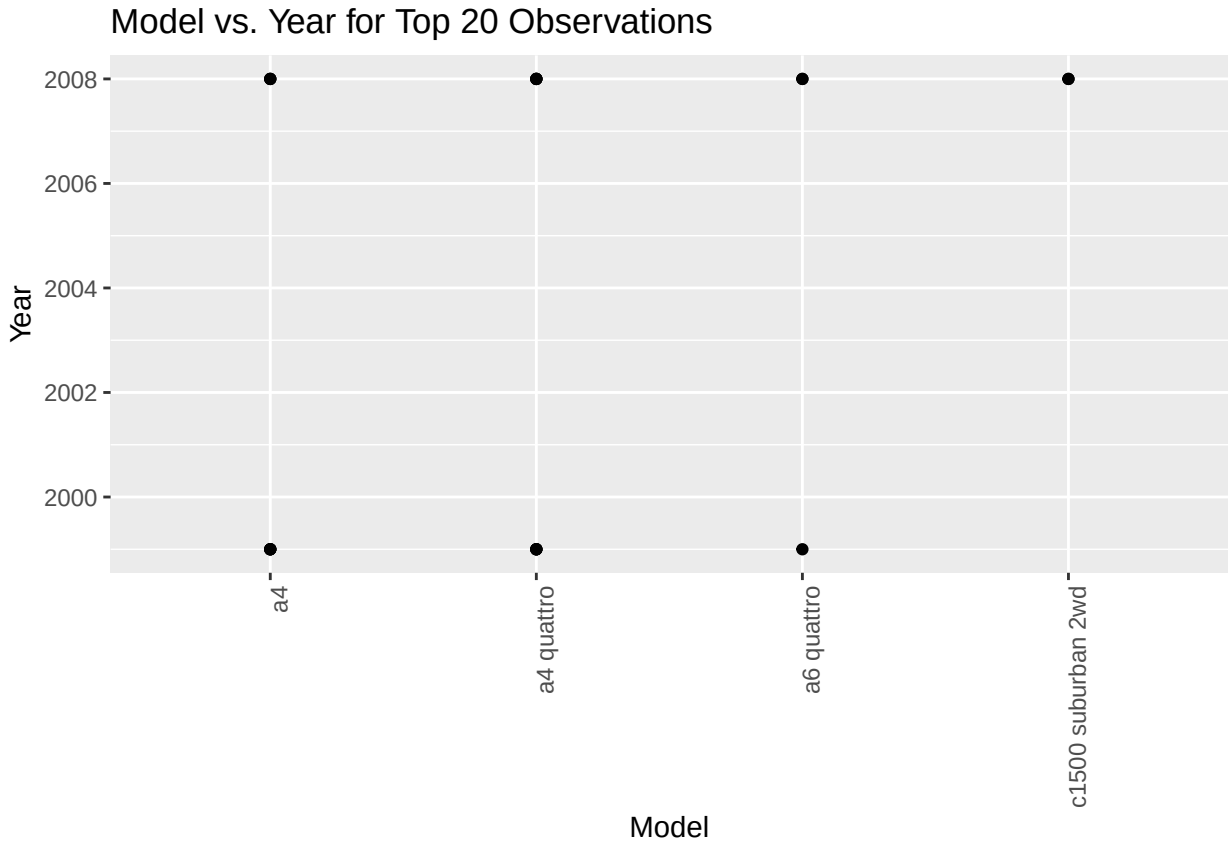
a. What does ggplot(mpg, aes(model, manufacturer)) + geom_point() show?

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

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

```
# Top 20 observations by model and year
top_20 <- mpg %>% slice(1:20)

ggplot(top_20, aes(x = model, y = year)) +
  geom_point() +
  labs(title = "Model vs. Year for 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

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

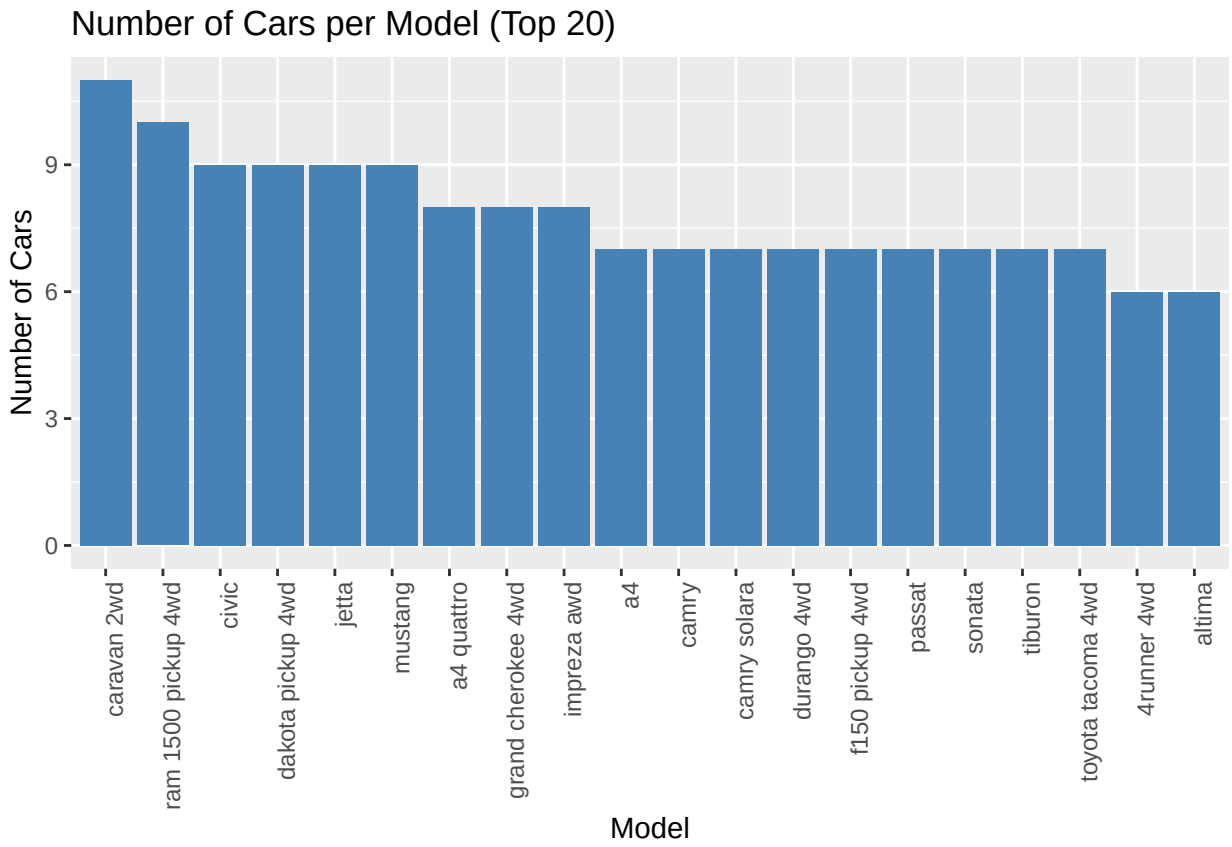
```
car_count <- mpg %>%
  group_by(model) %>%
  summarise(num_cars = n()) %>%
  arrange(desc(num_cars))

print(car_count)
```

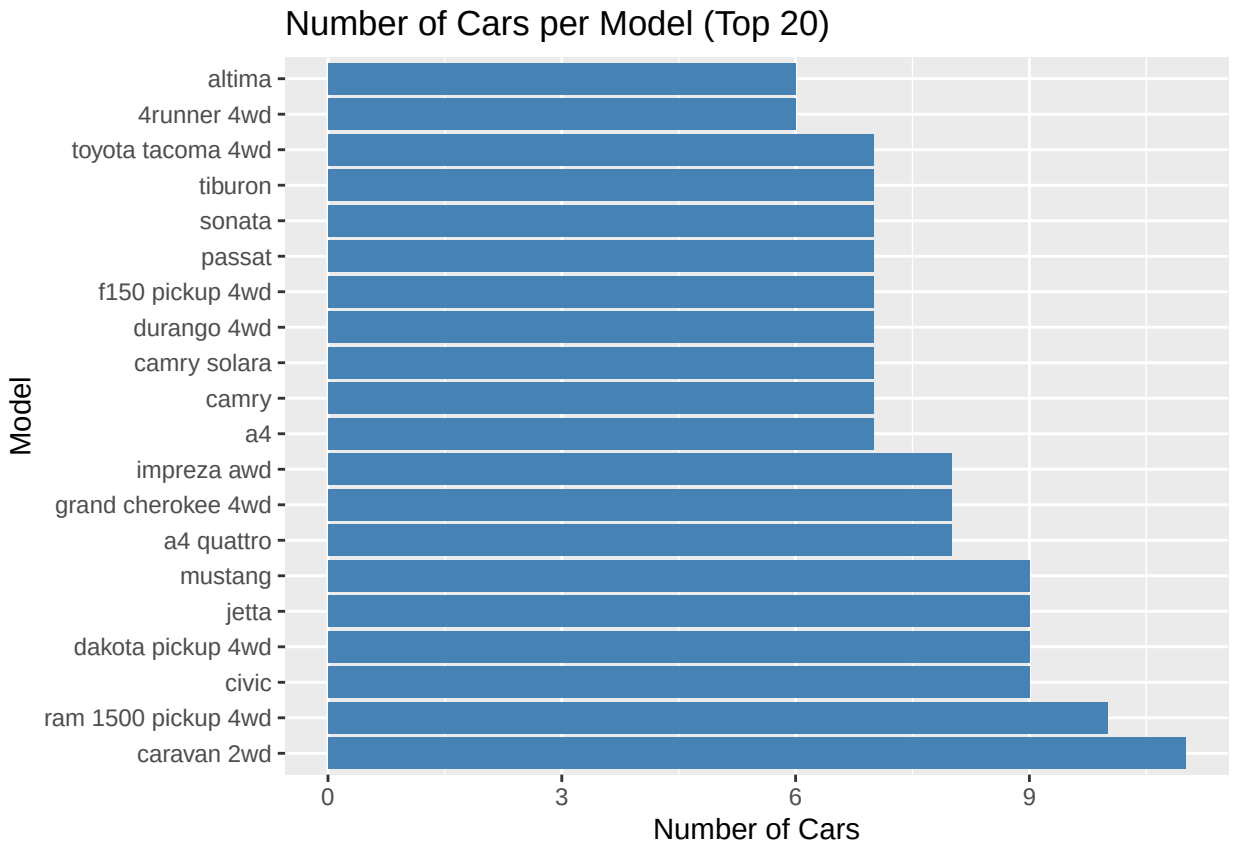
```
## # A tibble: 38 x 2
##    model               num_cars
##    <chr>                  <int>
##  1 caravan 2wd               11
##  2 ram 1500 pickup 4wd       10
##  3 civic                      9
##  4 dakota pickup 4wd          9
##  5 jetta                      9
##  6 mustang                    9
##  7 a4 quattro                 8
##  8 grand cherokee 4wd         8
##  9 impreza awd                8
## 10 a4                         7
## # i 28 more rows
```

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

```r
# Basic bar plot
ggplot(car_count %>% slice(1:20), aes(x = reorder(model, -num_cars), y = num_cars)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Number of Cars per Model (Top 20)", x = "Model", y = "Number of Cars") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

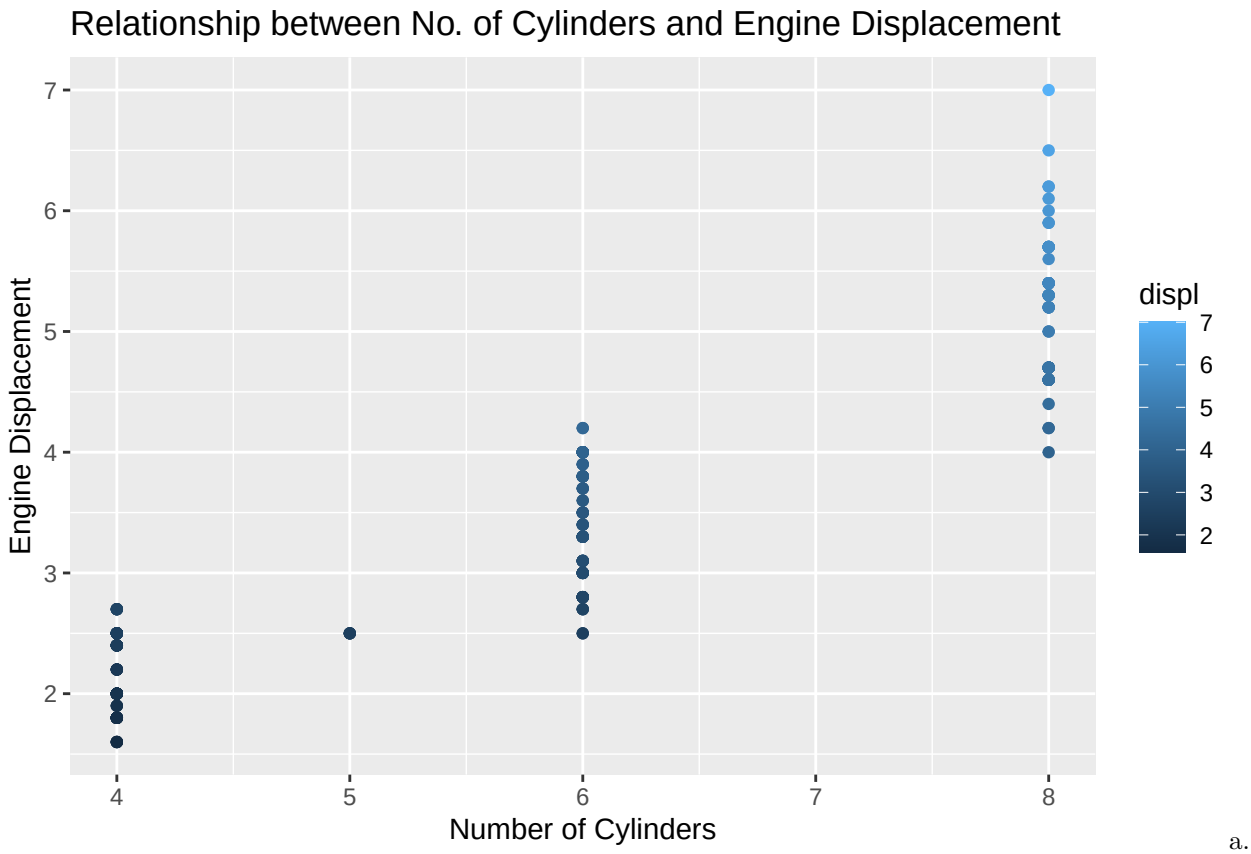



```r
# Bar plot with coord_flip()
ggplot(car_count %>% slice(1:20), aes(x = reorder(model, -num_cars), y = num_cars)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Number of Cars per Model (Top 20)", x = "Model", y = "Number of Cars") +
  coord_flip()
```



5. Plot the relationship between cyl - number of cylinders and displ - engine displacement using geom_point with aesthetic color = engine displacement. Title should be "Relationship between No. of Cylinders and Engine Displacement".

```
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")
```

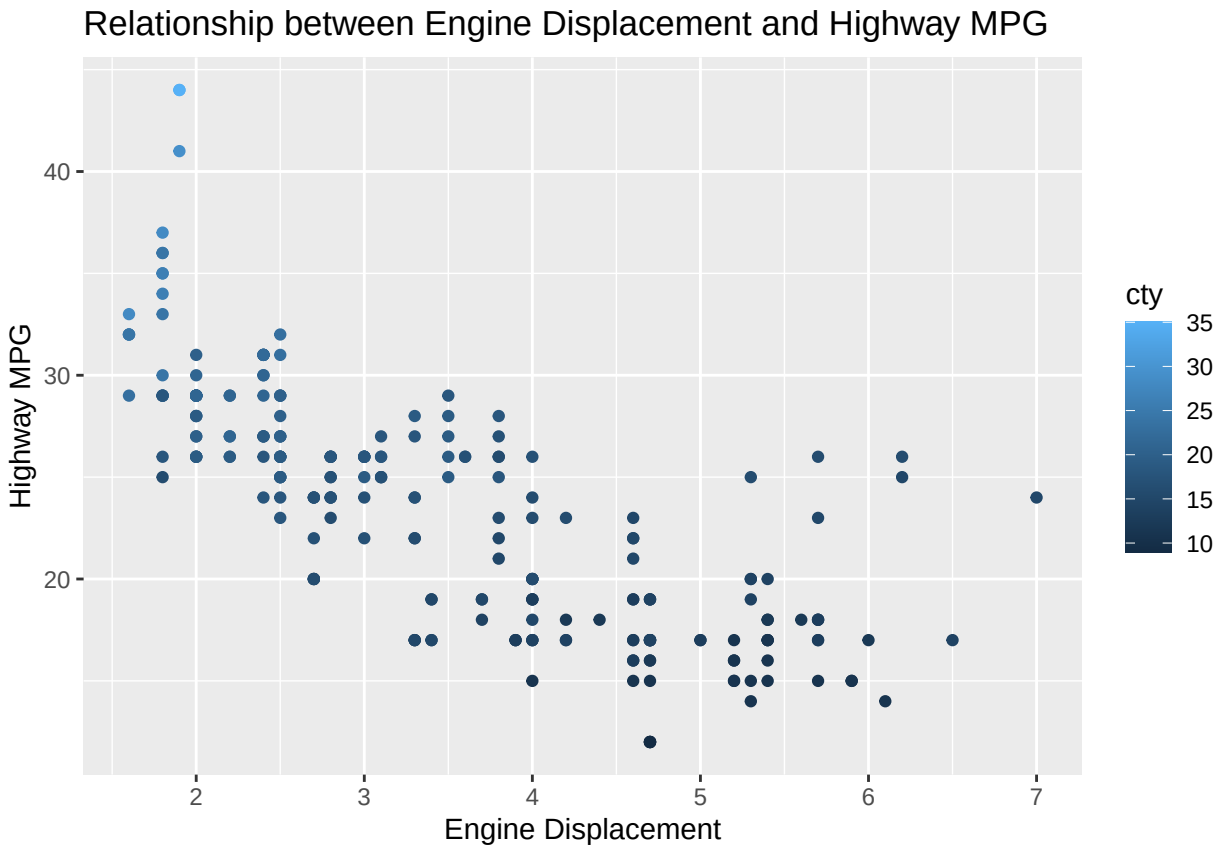


a.

How would you describe its relationship? Show the codes and its result.

6. Plot the relationship between displ (engine displacement) and hwy(highway miles per gallon). Mapped it with a continuous variable you have identified in #1-c. What is its result? Why it produced such output?

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



6. Import the traffic.csv onto your R environment.

```
traffic <- read.csv("traffic.csv")
```

a. How many numbers of observation does it have? What are the variables of the traffic dataset the Show your answer.

```
print(dim(traffic))
```

```
## [1] 48120     4
```

```
print(names(traffic))
```

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

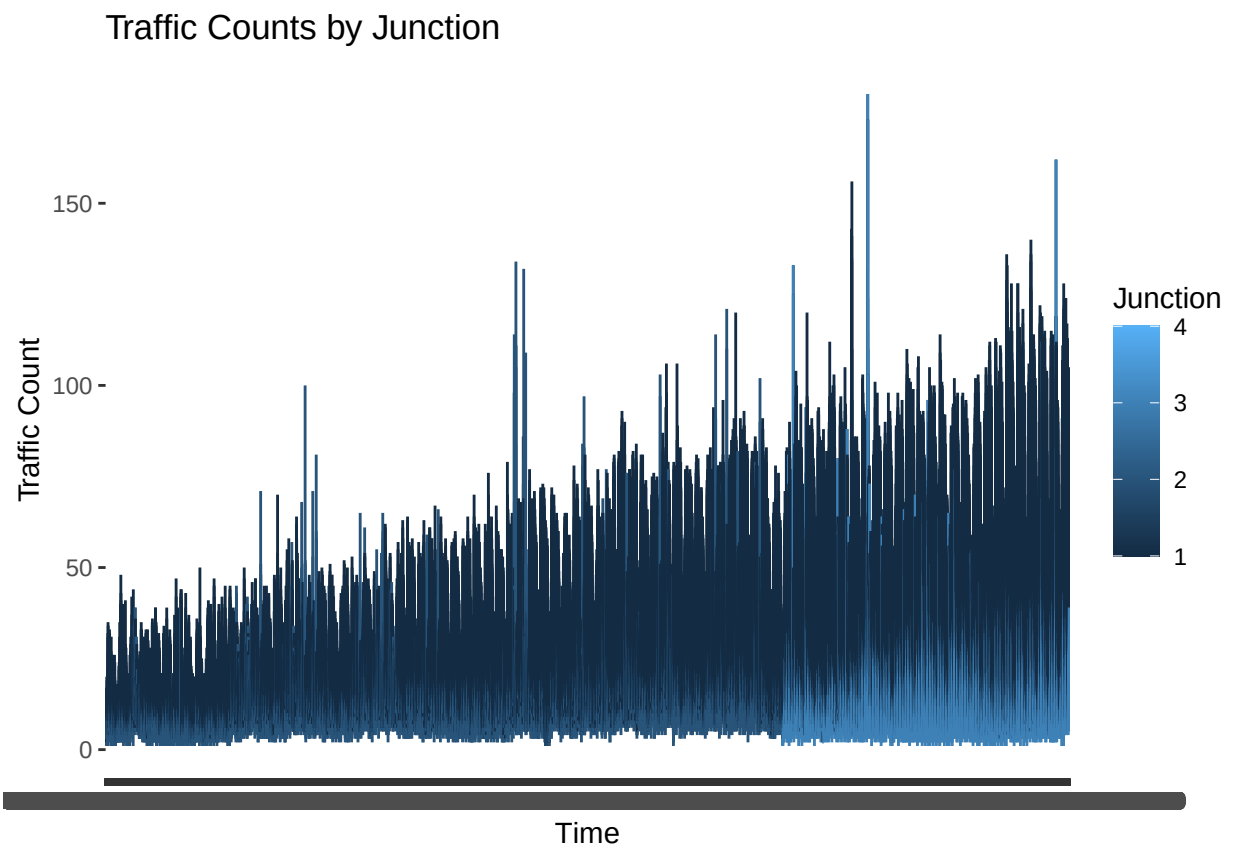
b. subset the traffic dataset into junctions. What is the R codes and its output?

```
junction_traffic <- traffic %>% filter(Junction == "junction")
print(junction_traffic)
```

```
## [1] DateTime Junction Vehicles ID
## <0 rows> (or 0-length row.names)
```

c. Plot each junction in a using geom_line(). Show your solution and output.

```
ggplot(traffic, aes(x = DateTime, y = Vehicles, color = Junction)) +
  geom_line() +
  labs(title = "Traffic Counts by Junction", x = "Time", y = "Traffic Count")
```



7. From alexa_file.xlsx, import it to your environment

a. How many observations does alexa_file has? What about the number of columns? Show your solution and answer.

```
library(readxl)
alexa_data <- read_excel("alexa_file.xlsx")
print(dim(alexa_data))
```

```
## [1] 3150    5
```

b. group the variations and get the total of each variations. Use dplyr package. Show solution and answer.

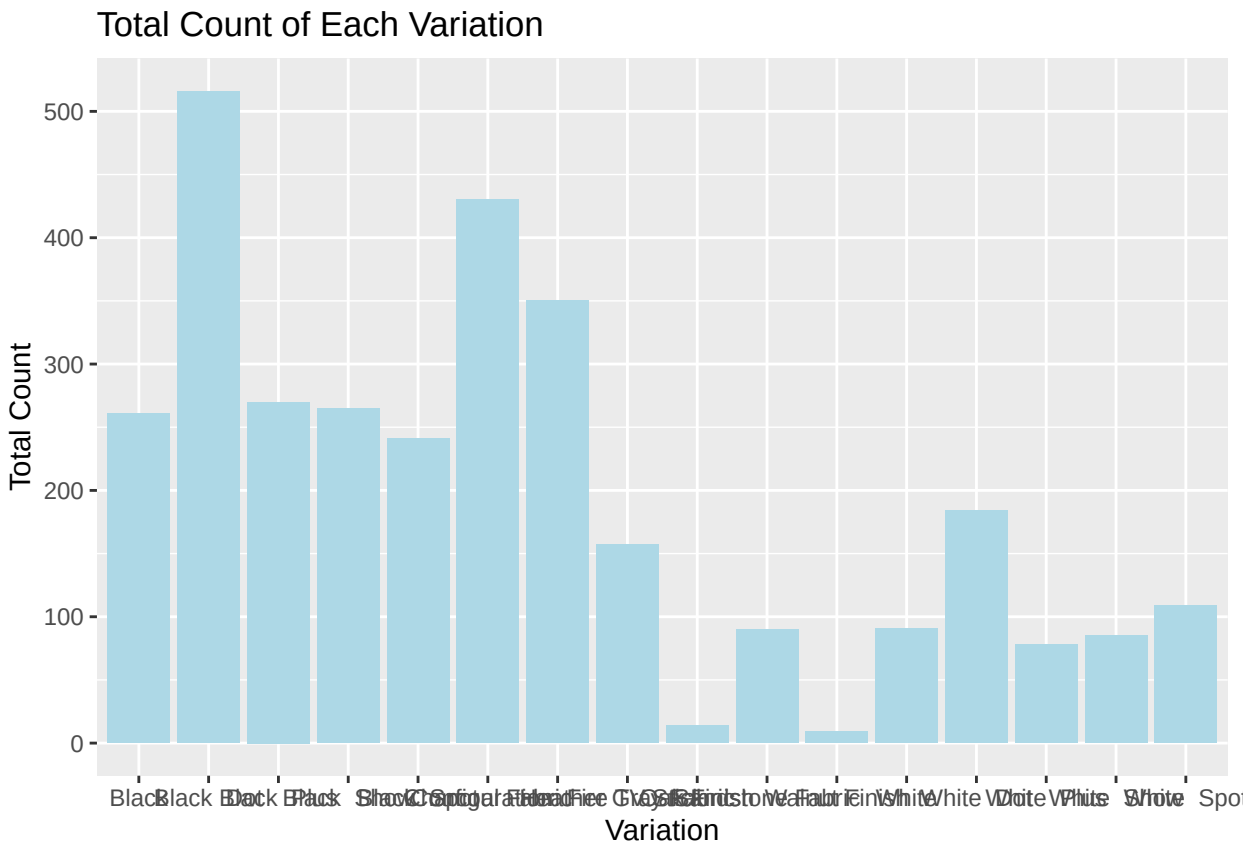
```
alexa_variations <- alexa_data %>%
  group_by(variation) %>%
  summarise(total = n())
print(alexa_variations)
```

```
## # 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. Plot the variations using the ggplot() function. What did you observe? Complete the details of the graph. Show solution and answer.

```
ggplot(alexa_variations, aes(x = variation, y = total)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  labs(title = "Total Count of Each Variation", x = "Variation", y = "Total Count")
```



d. Plot a geom_line() with the date and the number of verified reviews. Complete the details of the graphs. Show your answer and solution.

```
ggplot(alexa_data, aes(x = date, y = verified_reviews)) +
  geom_line() +
  labs(title = "Verified Reviews Over Time", x = "Date", y = "Verified Reviews")
```

are some serious flaws, particularly if you are the last one to bed or the first to wake. It doesn't seem like the engineer

expensive alternative option to fill the gap. Ordered the Amazon Fire Stick from Best Buy. Instructions were short and

one of the lights by saying "Alexa, turn off the second light". In the Alexa app, I created a 'Group' with &#34
, but lately I've been getting terrible support. The guy that took my call just rambled off a (completely unhelpful) script a

nting to add this both to my Alexa Echo Plus. Everything I tried ended in a Discovery Failed message. I tried to set
gh multiple pages. The one thing that I am not a fan of is the home screen cards do not really rotate that much - they a

e. Get the relationship of variations and ratings. Which variations got the most highest in rating? Plot a graph to show its relationship. Show your solution and answer.

```
variation_ratings <- alexa_data %>%
  group_by(variation) %>%
  summarise(avg_rating = mean(rating))

ggplot(variation_ratings, aes(x = reorder(variation, -avg_rating), y = avg_rating)) +
  geom_bar(stat = "identity", fill = "coral") +
  labs(title = "Average Rating by Variation", x = "Variation", y = "Average Rating") +
  coord_flip()
```