Question_No_6

Utsab Bhattarai

2024-05-31

Do the followings in R studio using ggplot2 package with R script to knit PDF output:

a. Create a dataset with following variables: age(10-99 years), sex(male/female), educational levels(No education/Primary/Secondary/Beyond secondary), socio-economic status(Low, Middle, High) and body mass index(14-38) with random 200 cases of each variable. Your roll number must be used to set the random seed.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(dplyr)
```

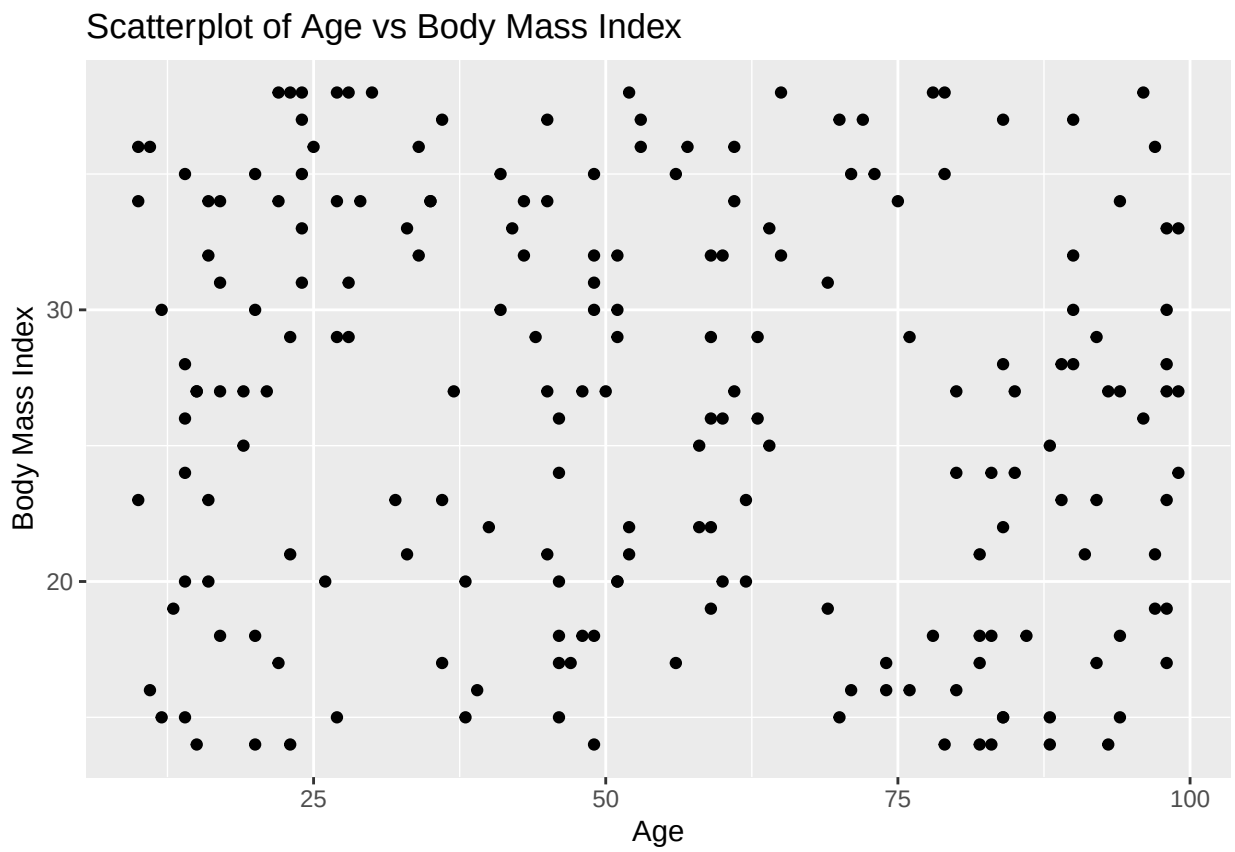
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
set.seed(35)
data <- tibble(
  age = sample(10:99,
               size = 200,
               replace = TRUE),
  sex = sample(c("Male", "Female"),
               size = 200,
               replace = TRUE),
  educational_level = sample(c("No education", "Primary", "Secondary", "Beyond Secondary"),
```

```
                              size = 200,
                              replace = TRUE),
  socio_economic_status = sample(c("Low", "Middle", "High"),
                                 size = 200,
                                 replace = TRUE),
  body_mass_index = sample(14:38,
                           size = 200,
                           replace = TRUE)
)
```

b. Create a scatterplot of age and body mass index variables using ggplot2 package and interpret the result carefully.

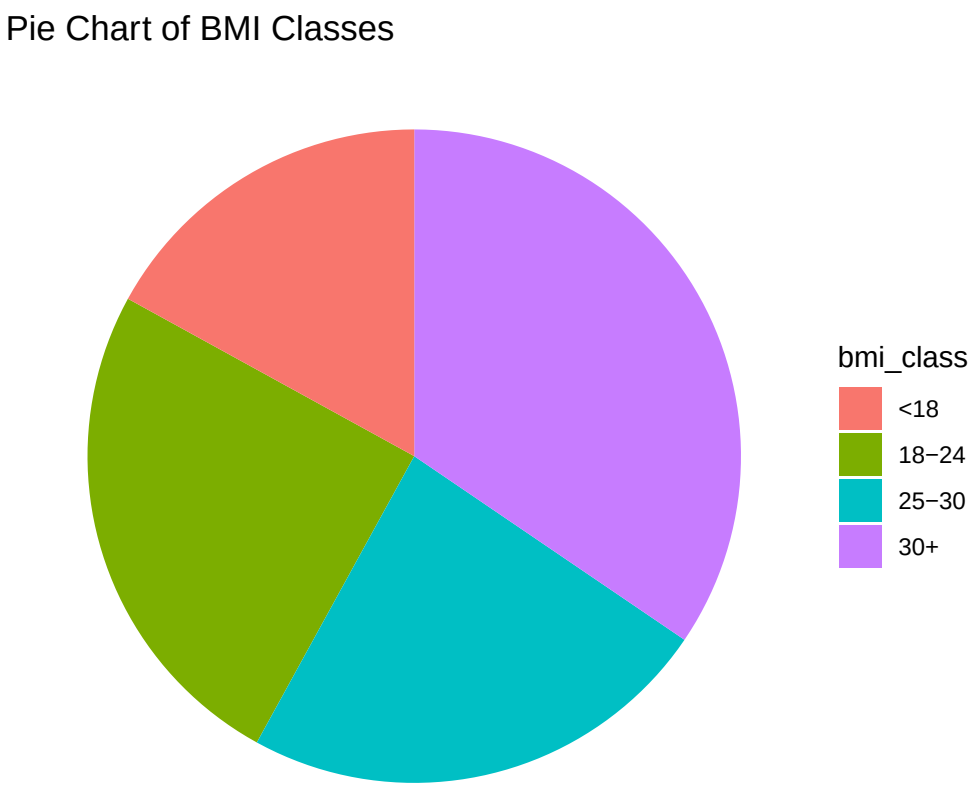
```
ggplot(data, aes(x = age, y = body_mass_index)) +
  geom_point() +
  labs(title = "Scatterplot of Age vs Body Mass Index",
       x = "Age",
       y = "Body Mass Index")
```



Interpretation: The scatterplot shows the distribution of body mass index across different ages. There is no clear pattern or correlation observed between age and body mass index in this "data" dataset.

c. Create classes of body mass index variable as : <18, 18-24, 25-30, 30+ and show it as pie chart using ggplot2 package and interpret it carefully.

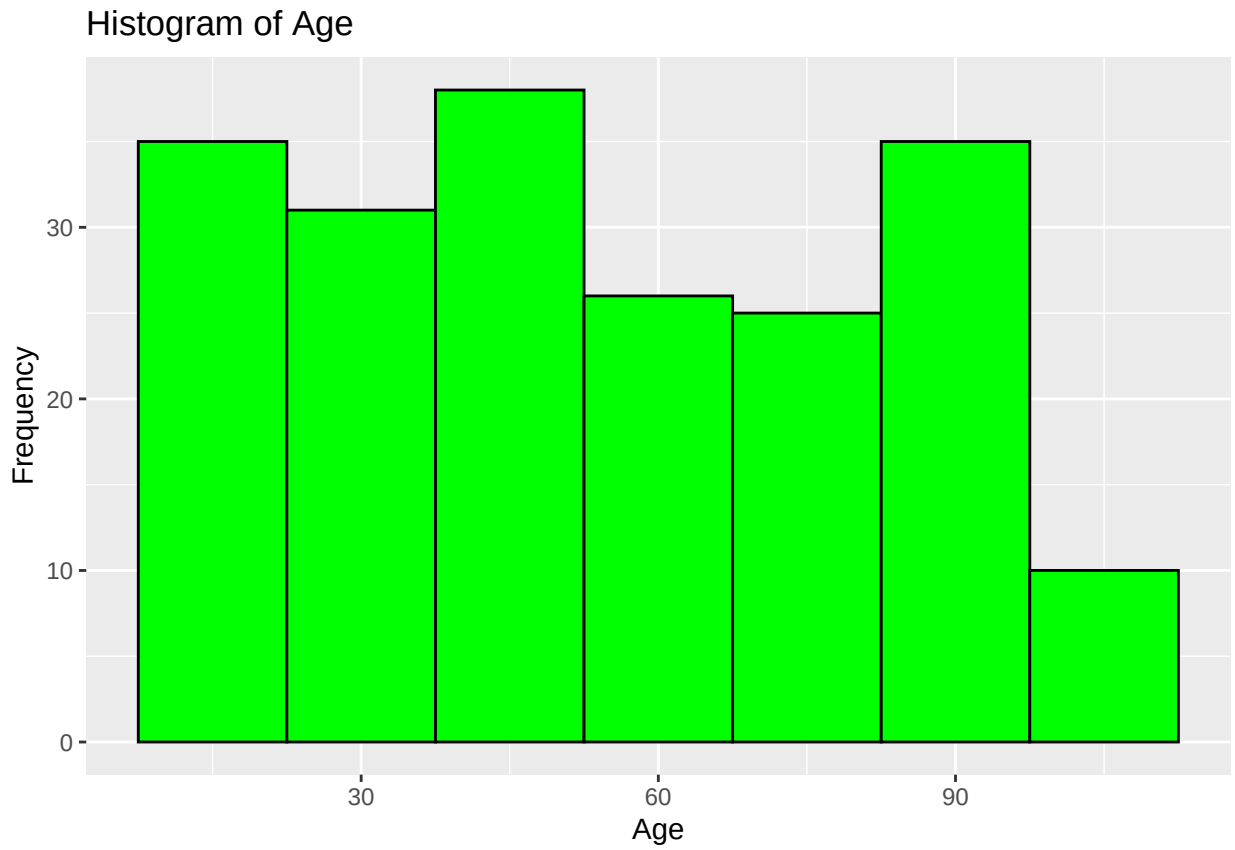
```
data <- data %>%
  mutate(bmi_class = case_when(
    body_mass_index < 18 ~ "<18",
    body_mass_index >= 18 & body_mass_index <= 24 ~ "18-24",
    body_mass_index >= 25 & body_mass_index <= 30 ~ "25-30",
    body_mass_index > 30 ~ "30+"
  ))
bmi_class_counts <- data %>%
  count(bmi_class)
ggplot(bmi_class_counts,
       aes(x = "",
           y = n,
           fill = bmi_class)) +
  geom_bar(stat = "identity",
           width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Pie Chart of BMI Classes") +
  theme_void()
```



Interpretation: The pie chart displays the distribution of body mass index classes among the dataset. Each segment represents the proportion of cases falling into each BMI class. As it seems that the BMI index of <18 is greater than other BMI index classes.

d. Create histogram of age variable with bin size of 15 using the ggplot2 package and interpret it carefully.

```
ggplot(data, aes(x = age)) +
  geom_histogram(binwidth = 15, fill = "green", color = "black") +
  labs(title = "Histogram of Age",
       x = "Age",
       y = "Frequency")
```



Interpretation: The histogram shows the distribution of ages in the dataset with a bin size of 15. It helps to identify the age groups that are most and least represented in the dataset. It shows that the age class of 30-37.5 is greater than other age classes.