Exploring Data with R

Section 1: Identifying Variables in a Dataset

Question 1: Identify the elements in the dataset and determine what variables are measured by identifying each as categorical or quantitative.

In [2]:
```
# Load the data
data <- data.frame(
  State = c("Florida", "Alabama", "California"),
  Zip_Code = c(32116, 35236, 94565),
  Family_Size = c(6, 5, 1),
  Annual_Income = c(13500, 800, 23000)
)

# Print the data
data
```

A data.frame: 3 × 4

State	Zip_Code	Family_Size	Annual_Income
<chr>	<dbl>	<dbl>	<dbl>
Florida	32116	6	13500
Alabama	35236	5	800
California	94565	1	23000

Explanation:

- **State**: Categorical
- **Zip_Code**: Categorical
- **Family_Size**: Quantitative
- **Annual_Income**: Quantitative

Section 2: Visualizing Study Time Data

Question 2: Create a dot plot, back-to-back stem-and-leaf plot, and histogram for the study times of men and women. Determine the approximate center and shape of each.

In [4]:

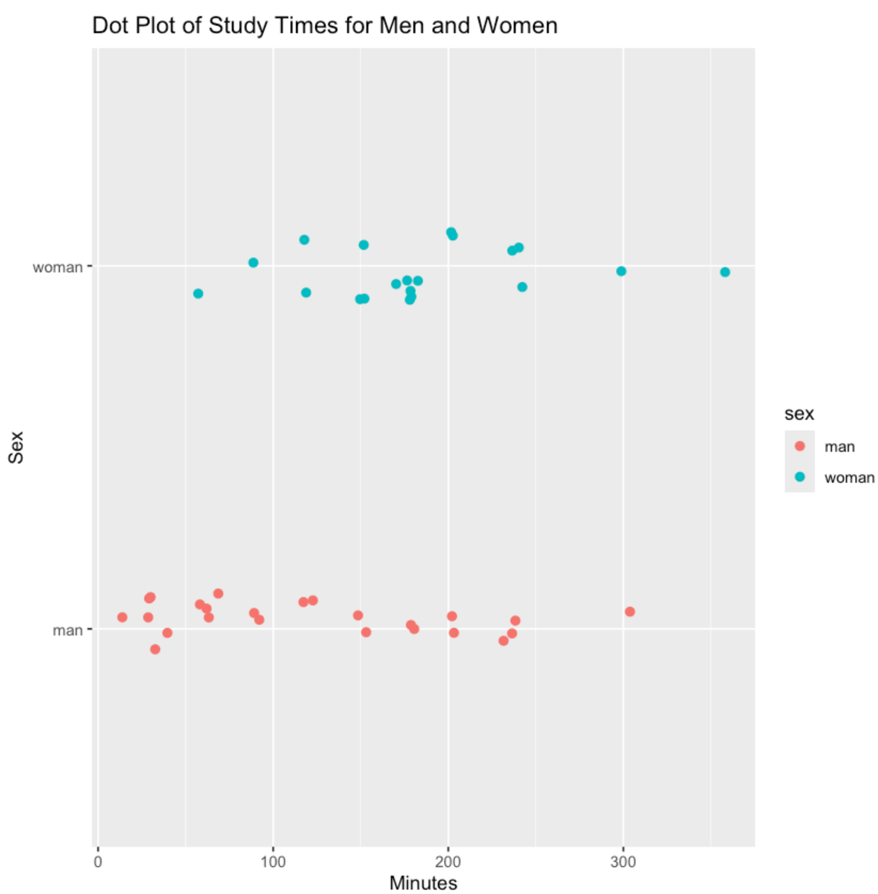
```r
# Study times for men and women

men <- c(10, 30, 30, 30, 30, 40, 60, 60, 60, 70,
         90, 90, 120, 120, 150, 150, 180, 180, 200,
         200, 230, 240, 240, 300)

women <- c(60, 90, 120, 120, 150, 150, 150, 170, 180, 180,
           180, 180, 180, 200, 200, 240, 240, 240, 300, 360)
```

In [12]:

```r
# Dot Plot
library(ggplot2)

df_men <- data.frame(study.time = men,
                     sex = "man")

df_women <- data.frame(study.time = women,
                       sex = "woman")

data_combined <- rbind(df_men, df_women)
```

In [13]:

```r
# Dot plot visualization
ggplot(data_combined, aes(x = study.time, y = sex, color = sex)) +
  geom_point(position = position_jitter(height = 0.1), size = 2) +
  labs(title = "Dot Plot of Study Times for Men and Women", x = "Minutes", y
```



In [110…

```r
# 5 Stat Summary
library(dplyr)
```

```r
data_combined %>%
    group_by(sex) %>%
    summarize(n = n(),
              mean = mean(study.time),
              min = min(study.time),
              q1 = quantile(study.time,.25),
              q2 = quantile(study.time,.50),
              q3 = quantile(study.time,.75),
              max = max(study.time))
```

A tibble: 2 × 8

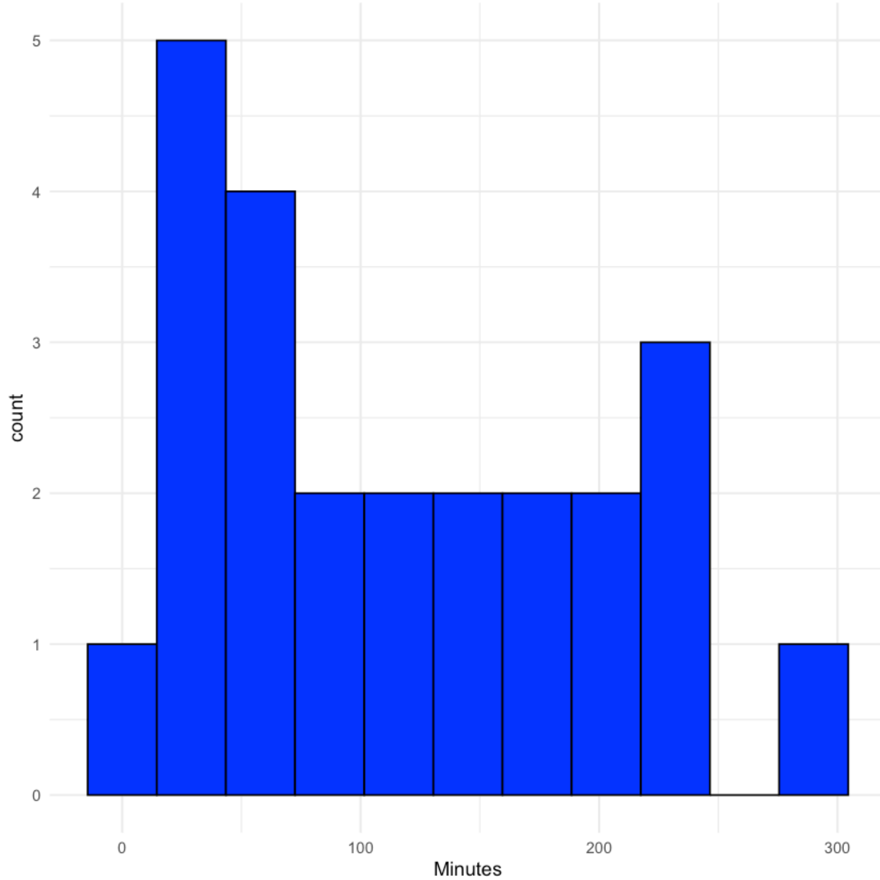
sex	n	mean	min	q1	q2	q3	max
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
man	24	121.25	10	55	105	185	300
woman	20	184.50	60	150	180	210	360

In [57]:

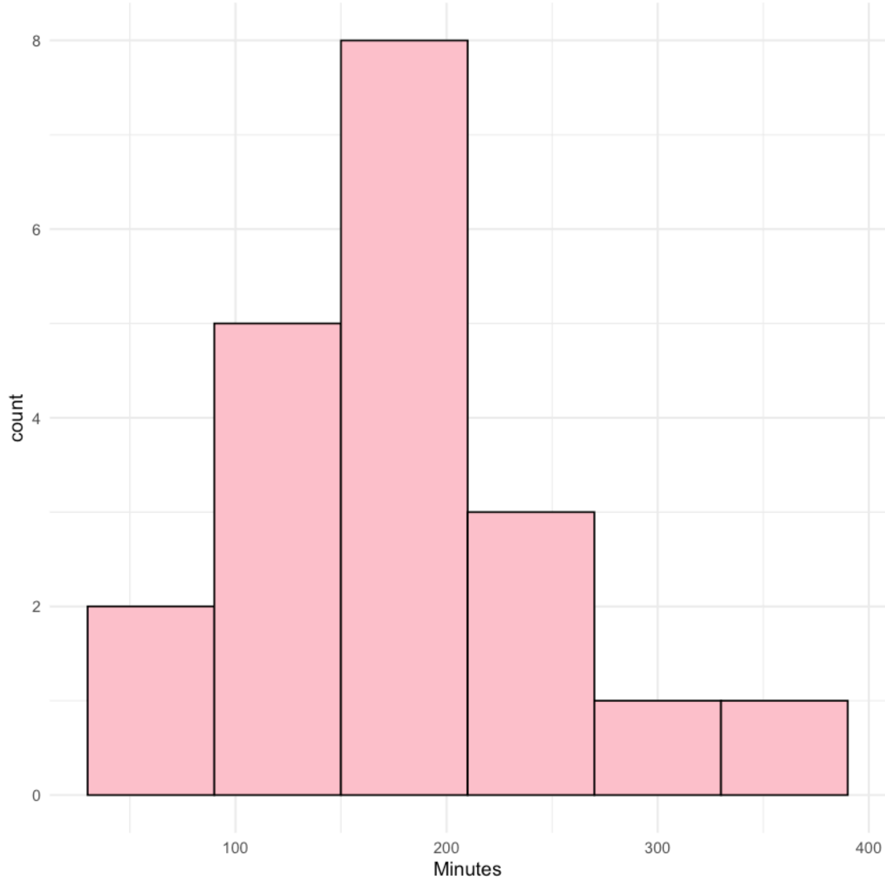
```r
# Histogram Width
men.iqr <- 185-55
women.iqr <- 210-150

men.width <- ceiling( (2*men.iqr)/24 )
women.width <- ceiling( (2*women.iqr)/20 )
```

In [88]:

```r
# Histogram Men
data_combined %>%
    filter(sex == "man") %>%
    ggplot(aes(study.time)) +
    geom_histogram(bins = men.width,color="black",fill='blue') +
    labs(x="Minutes","Histogram of Men's Study Time") +
    theme_minimal()
```



In [89]:
```
# Histogram Men
data_combined %>%
    filter(sex == "woman") %>%
    ggplot(aes(study.time)) +
    geom_histogram(bins = women.width,color="black",fill='pink') +
    labs(x="Minutes","Histogram of Women's Study Time") +
    theme_minimal()
```



Summary:

- **Men**: Center ~120, slightly right-skewed.
- **Women**: Center ~150, slightly right-skewed.

Section 3: Summary Statistics and Boxplot

Question 3: Compute summary statistics for the given data and create a boxplot.

In [114…

```
# Data
data <- c(32, 31, 30, 33, 34, 32, 35, 40, 32, 30, 32, 31)
mean(data)
```

32.6666666666667

In [111…

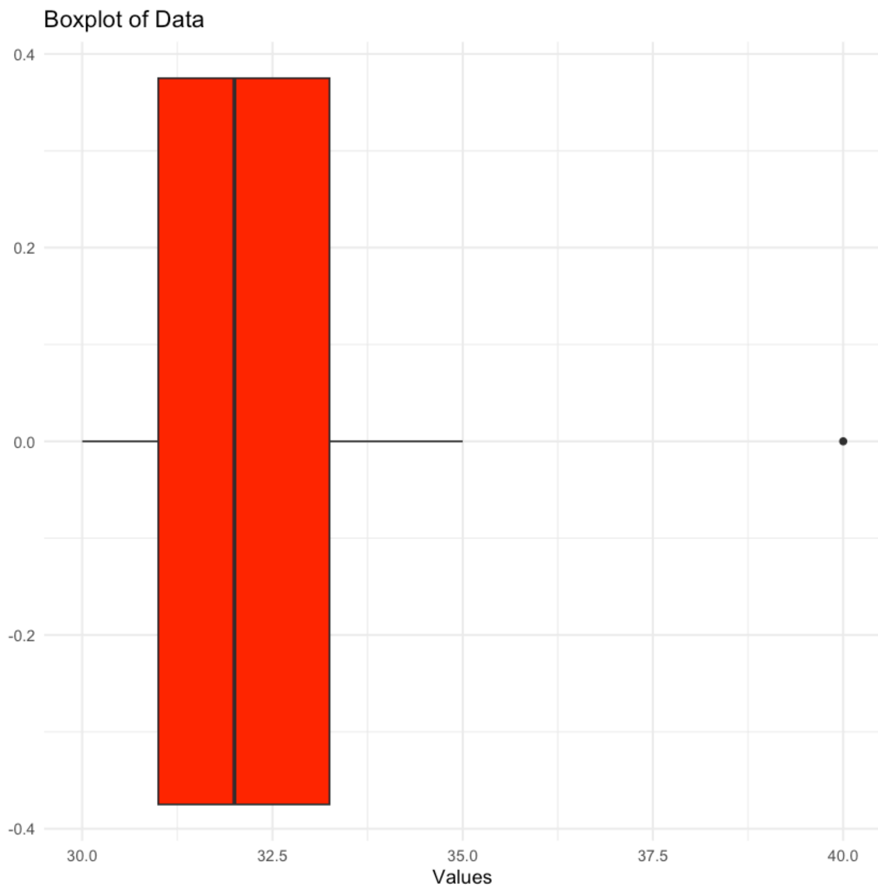
```
# Summary statistics
summary(data)
iqr <- IQR(data)
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
30.00   31.00   32.00   32.67   33.25   40.00 
```

In [112…

```
# Boxplot
df <- data.frame(values=data)
```

```r
ggplot(df,aes(x=values)) +
geom_boxplot(fill='red') +
labs(title ="Boxplot of Data",x="Values") +
theme_minimal()
```



In [113…

```r
# Print IQR
cat("Interquartile Range (IQR):", iqr)
```

```
Interquartile Range (IQR): 2.25
```

Results:

- **Mean**: 32.7
- **Median**: 32
- **IQR**: 2.25
- Boxplot highlights outliers and distribution symmetry.

Section 4: Two-Way Frequency Table

Question 4: Analyze the conditional distributions and marginal values for hair color and gender.

In [115…

```r
# Create table
hair_data <- matrix(c(26, 24, 10, 3, 20, 35, 12, 9), ncol = 4, byrow = TRUE)
colnames(hair_data) <- c("Brown", "Blond", "Black", "Ginger")
```

```r
rownames(hair_data) <- c("Male", "Female")

hair_table <- as.table(hair_data)
print(hair_table)
```

```
       Brown Blond Black Ginger
Male      26    24    10      3
Female    20    35    12      9
```

In [116…:

```r
# Marginal distributions
margin.table(hair_table, 1) # Gender
margin.table(hair_table, 2) # Hair color
```

```
  Male Female
    63     76
 Brown  Blond  Black Ginger
    46     59     22     12
```

In [117…:

```r
# Conditional distribution for gender
prop.table(hair_table, 1)

# Conditional distribution for hair color
prop.table(hair_table, 2)
```

```
            Brown      Blond      Black     Ginger
Male   0.41269841 0.38095238 0.15873016 0.04761905
Female 0.26315789 0.46052632 0.15789474 0.11842105
           Brown     Blond     Black    Ginger
Male   0.5652174 0.4067797 0.4545455 0.2500000
Female 0.4347826 0.5932203 0.5454545 0.7500000
```

Section 5: Normal Distribution and Probability

Question 5: Analyze gestation periods using a normal distribution.

In [ ]:

In [ ]:

Studies have shown that the average gestation period (conception to birth) for humans is 26 days (38 weeks) with astandard deviation of 16 days. The length of gestation is considered to follow a normal distribution.

In [119…:

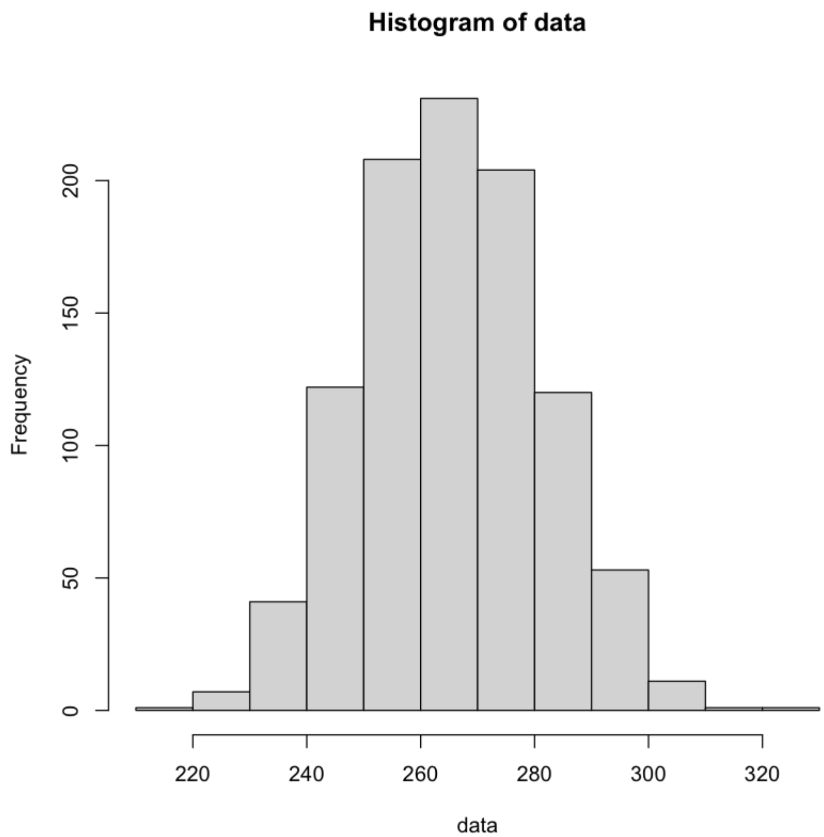
```r
# Given data
mean <- 266
sd <- 16
```

In [ ]:

In [ ]:

In [129...

```
data <- rnorm(1000,mean,sd)
hist(data)
```



In [127...

```
# (a) Sketch omitted (use theory)
# (b) P(gestation < 36 weeks)
p_36 <- pnorm(36 * 7, mean, sd)
```

In [126...

```
# (c) P(gestation > 39 weeks)
p_39 <- 1 - pnorm(39 * 7, mean, sd)
```

In [125...

```
# (d) P(245 <= gestation <= 270)
p_range <- pnorm(270, mean, sd) - pnorm(245, mean, sd)
```

In [124...

```
# (e) Top 5% of gestations
top_5 <- qnorm(0.95, mean, sd)
```

In [128...

```
cat("P(<36 weeks):", p_36, "\n")
cat("P(>39 weeks):", p_39, "\n")
cat("P(245-270):", p_range, "\n")
cat("Top 5% gestation threshold:", top_5)
```

```
P(<36 weeks): 0.190787
P(>39 weeks): 0.3308744
P(245–270): 0.5040306
Top 5% gestation threshold: 292.3177
```

Results:

- Probability calculations based on Z-scores and the normal curve.

In [ ]: