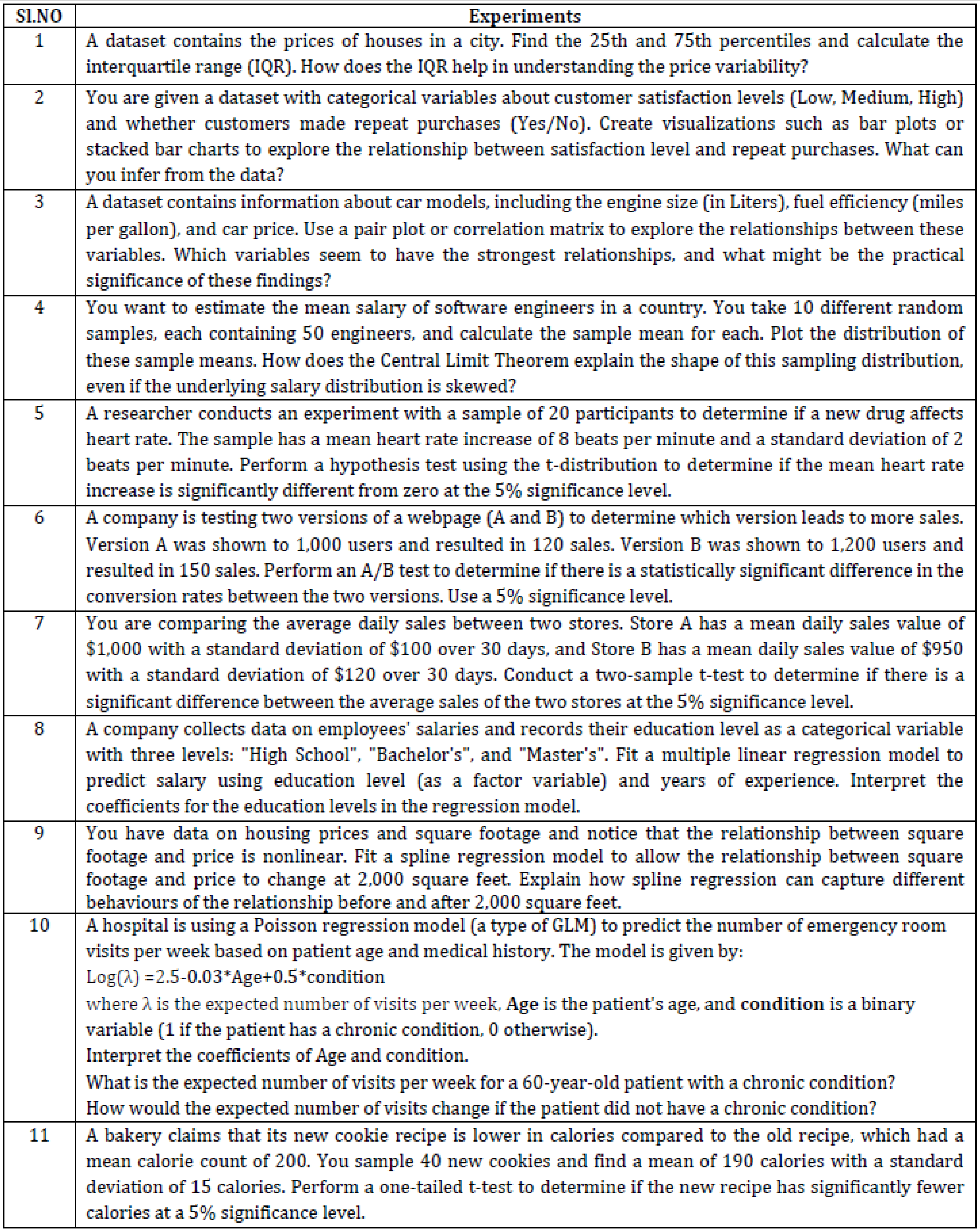


Sl.NO	Experiments
1	A dataset contains the prices of houses in a city. Find the 25th and 75th percentiles and calculate the interquartile range (IQR). How does the IQR help in understanding the price variability?
2	You are given a dataset with categorical variables about customer satisfaction levels (Low, Medium, High) and whether customers made repeat purchases (Yes/No). Create visualizations such as bar plots or stacked bar charts to explore the relationship between satisfaction level and repeat purchases. What can you infer from the data?
3	A dataset contains information about car models, including the engine size (in Liters), fuel efficiency (miles per gallon), and car price. Use a pair plot or correlation matrix to explore the relationships between these variables. Which variables seem to have the strongest relationships, and what might be the practical significance of these findings?
4	You want to estimate the mean salary of software engineers in a country. You take 10 different random samples, each containing 50 engineers, and calculate the sample mean for each. Plot the distribution of these sample means. How does the Central Limit Theorem explain the shape of this sampling distribution, even if the underlying salary distribution is skewed?
5	A researcher conducts an experiment with a sample of 20 participants to determine if a new drug affects heart rate. The sample has a mean heart rate increase of 8 beats per minute and a standard deviation of 2 beats per minute. Perform a hypothesis test using the t-distribution to determine if the mean heart rate increase is significantly different from zero at the 5% significance level.
6	A company is testing two versions of a webpage (A and B) to determine which version leads to more sales. Version A was shown to 1,000 users and resulted in 120 sales. Version B was shown to 1,200 users and resulted in 150 sales. Perform an A/B test to determine if there is a statistically significant difference in the conversion rates between the two versions. Use a 5% significance level.
7	You are comparing the average daily sales between two stores. Store A has a mean daily sales value of $1,000 with a standard deviation of $100 over 30 days, and Store B has a mean daily sales value of $950 with a standard deviation of $120 over 30 days. Conduct a two-sample t-test to determine if there is a significant difference between the average sales of the two stores at the 5% significance level.
8	A company collects data on employees' salaries and records their education level as a categorical variable with three levels: "High School", "Bachelor's", and "Master's". Fit a multiple linear regression model to predict salary using education level (as a factor variable) and years of experience. Interpret the coefficients for the education levels in the regression model.
9	You have data on housing prices and square footage and notice that the relationship between square footage and price is nonlinear. Fit a spline regression model to allow the relationship between square footage and price to change at 2,000 square feet. Explain how spline regression can capture different behaviours of the relationship before and after 2,000 square feet.
10	A hospital is using a Poisson regression model (a type of GLM) to predict the number of emergency room visits per week based on patient age and medical history. The model is given by: Log(λ) =2.5-0.03*Age+0.5*condition where λ is the expected number of visits per week, **Age** is the patient's age, and **condition** is a binary variable (1 if the patient has a chronic condition, 0 otherwise). Interpret the coefficients of Age and condition. What is the expected number of visits per week for a 60-year-old patient with a chronic condition? How would the expected number of visits change if the patient did not have a chronic condition?
11	A bakery claims that its new cookie recipe is lower in calories compared to the old recipe, which had a mean calorie count of 200. You sample 40 new cookies and find a mean of 190 calories with a standard deviation of 15 calories. Perform a one-tailed t-test to determine if the new recipe has significantly fewer calories at a 5% significance level.

1.A dataset contains the prices of houses in a city. Find the 25th and 75th percentiles and calculate the interquartile range (IQR). How does the IQR help in understanding the price variability?

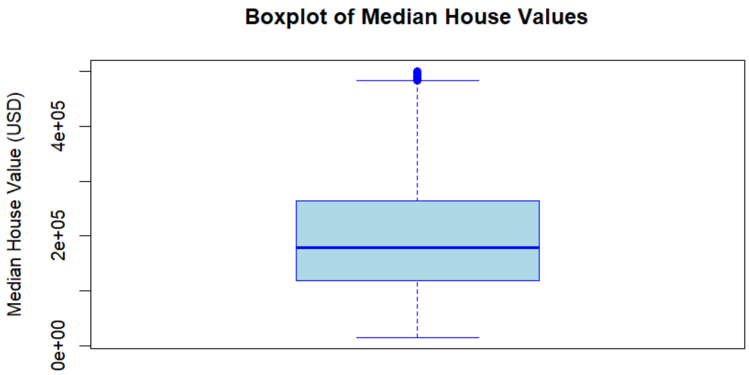
Dataset:

https://www.kaggle.com/datasets/camnugent/california-housing-prices

Program:

```
#Load the data (update path as needed)
housing <- read.csv("C:/Users/KUMARESH/Downloads/archive/housing.csv")
#Check column names (optional)
names(housing)
#Calculate Q1 and Q3 for median house value
percentiles <- quantile(housing$median_house_value, probs = c(0.25, 0.75), na.rm = TRUE) Q1 <-
percentiles[1] Q3 <- percentiles[2]
#Calculate IQR
IQR_value <- Q3 - Q1
#Output
cat("25th Percentile (Q1):", Q1, "\n")
cat("75th Percentile (Q3):", Q3, "\n")
cat("Interquartile Range (IQR):", IQR_value, "\n")
#Optional: Visualize with a boxplot
boxplot(housing$median_house_value, main = "Boxplot of Median House Values", ylab = "Median
House Value (USD)", col = "lightblue", border = "blue")
```

Output:



2. You are given a dataset with categorical variables about customer satisfaction levels (Low, Medium, High) and whether customers made repeat purchases (Yes/No). Create visualizations such as bar plots or stacked bar charts to explore the relationship between satisfaction level and repeat purchases. What can you infer from the data?

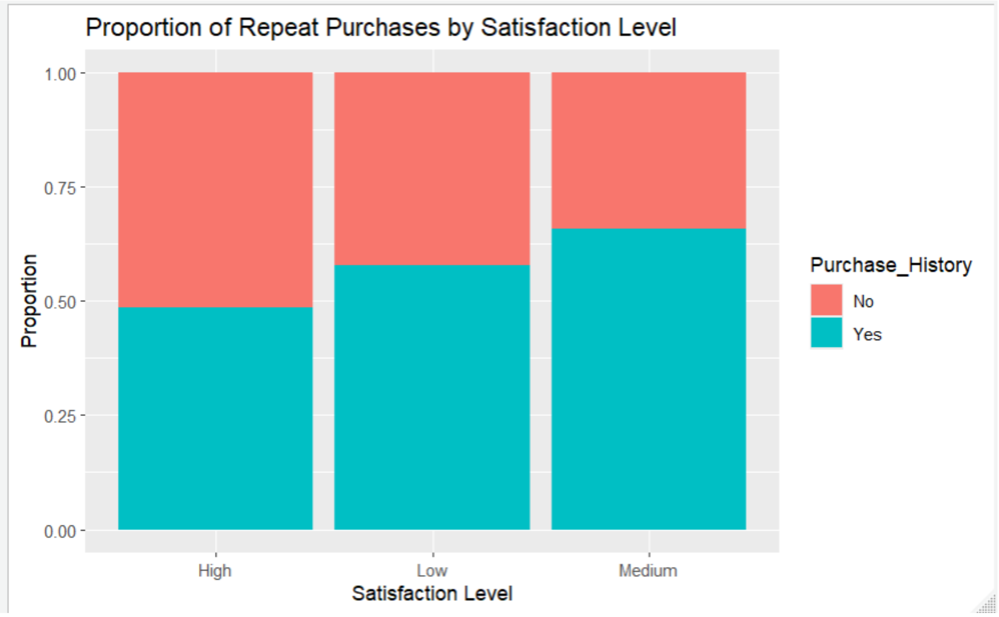
Dataset:

https://www.kaggle.com/datasets/salahuddinahmedshuvo/customer-satisfaction-scores-and-behavior-data

Program:

```
#Load libraries

library(ggplot2)

library(dplyr)

#Read dataset

data <- read.csv("C:/Users/KUMARESH/Downloads/archive/Customer Satisfaction Scores and Behavior Data.csv")

#Bar plot (count plot with hue)

ggplot(data, aes(x = Loyalty_Level, fill = Purchase_History)) + geom_bar(position = "dodge") + labs(title = "Customer Satisfaction (Loyalty Level) vs Repeat Purchase", x = "Satisfaction Level (Loyalty Level)", y = "Count", fill = "Repeat Purchase")

#Stacked bar chart

ggplot(data, aes(x = Loyalty_Level, fill = Purchase_History)) + geom_bar(position = "stack") + labs(title = "Stacked Bar: Satisfaction vs Repeat Purchase", x = "Satisfaction Level", y = "Number of Customers")

#Proportion stacked bar chart

ggplot(data, aes(x = Loyalty_Level, fill = Purchase_History)) + geom_bar(position = "fill") + labs(title = "Proportion of Repeat Purchases by Satisfaction Level", x = "Satisfaction Level", y = "Proportion")

#Cross-tabulation

table_summary <- table(data$Loyalty_Level, data$Purchase_History) prop_summary <- prop.table(table_summary, 1)

print(table_summary) print(round(prop_summary, 2))
```

Output:


Proportion of Repeat Purchases by Satisfaction Level
Proportion
1.00
0.75
0.50
0.25
0.00
High
Low
Medium
Satisfaction Level
Purchase_History
No
Yes

3. A dataset contains information about car models, including the engine size (in Liters), fuel efficiency (miles per gallon), and car price. Use a pair plot or correlation matrix to explore the relationships between these variables. Which variables seem to have the strongest relationships, and what might be the practical significance of these findings?

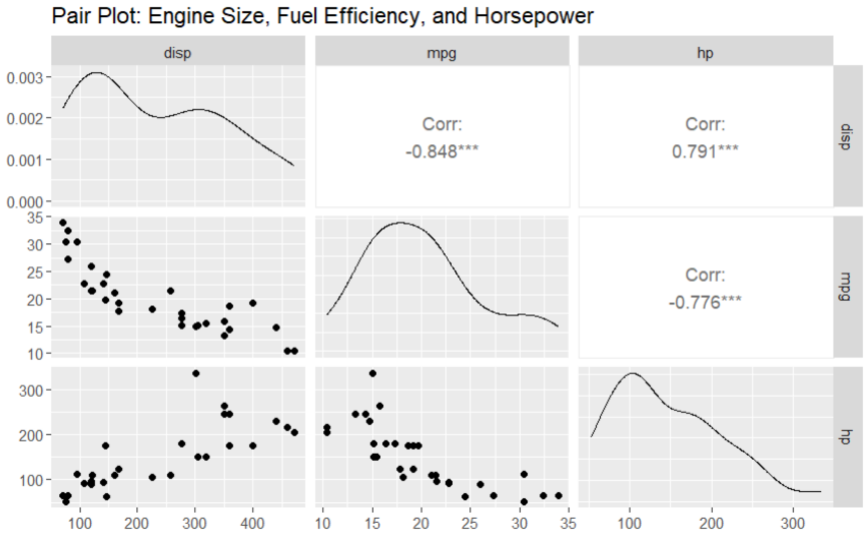
Dataset:

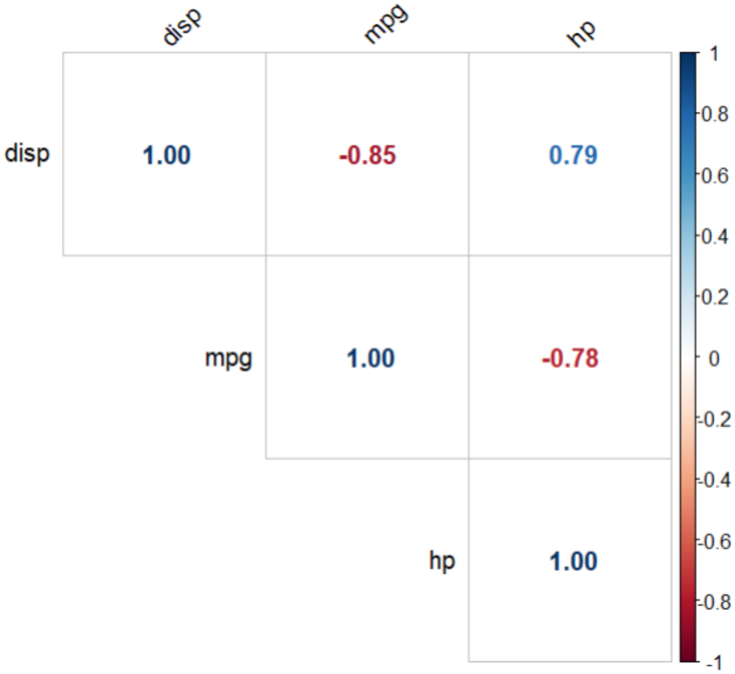
mtcars-R dataset

Program:

```
#Load libraries

library(GGally) # for ggpairs

library(corrplot) # for correlation plot

#Select relevant variables

data <- mtcars[, c("disp", "mpg", "hp")]

#---- Pair Plot ----

ggpairs(data, title = "Pair Plot: Engine Size, Fuel Efficiency, and Horsepower")

#---- Correlation Matrix ----

cor_matrix <- cor(data) print(cor_matrix)

#Visualize correlation matrix

corrplot(cor_matrix, method = "number", type = "upper", tl.col = "black", tl.srt = 45)
```

Output:





Conclusion:

The strongest relationships are Engine Size vs Fuel Efficiency (negative) and Engine Size vs Horsepower (positive).

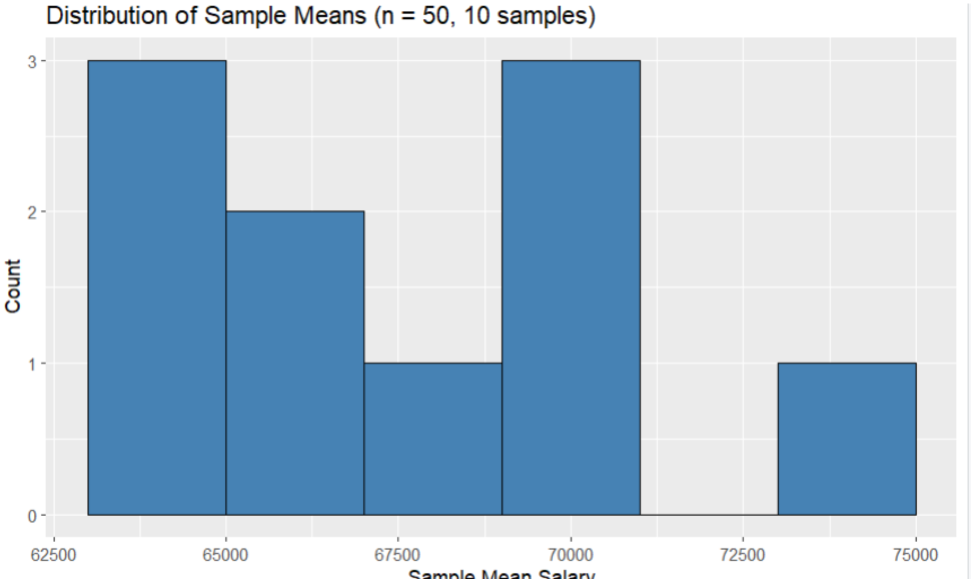
This reflects real-world trade-offs: fuel-efficient cars tend to be smaller and cheaper, while powerful cars tend to be larger and more expensive.

4. You want to estimate the mean salary of software engineers in a country. You take 10 different random samples, each containing 50 engineers, and calculate the sample mean for each. Plot the distribution of these sample means. How does the Central Limit Theorem explain the shape of this sampling distribution, even if the underlying salary distribution is skewed?

Program:

```
#Load library

library(ggplot2)

set.seed(123) # for reproducibility

#Step 1: Simulate population (skewed distribution of salaries)

#Assume salaries follow a right-skewed log-normal distribution

population <- rlnorm(100000, meanlog = 11, sdlog = 0.5)

#Step 2: Take 10 samples, each of size 50, and compute sample means

sample_means <- replicate(10, mean(sample(population, 50, replace = TRUE)))

#Step 3: Plot distribution of sample means

sample_means_df <- data.frame(sample_mean = sample_means)

ggplot(sample_means_df, aes(x = sample_mean)) + geom_histogram(binwidth = 2000, fill =
"steelblue", color = "black") + ggtitle("Distribution of Sample Means (n = 50, 10 samples)")
+ xlab("Sample Mean Salary") + ylab("Count")
```

Output:



5. A researcher conducts an experiment with a sample of 20 participants to determine if a new drug affects heart rate. The sample has a mean heart rate increase of 8 beats per minute and a standard deviation of 2 beats per minute. Perform a hypothesis test using the t-distribution to determine if the mean heart rate increase is significantly different from zero at the 5% significance level.

Program:

```
#Summary statistics
n <- 20
xbar <- 8
s <- 2
alpha <- 0.05
#Calculations
se <- s / sqrt(n)
 t_stat <- xbar / se
df <- n - 1
p_value <- 2 * (1 - pt(abs(t_stat), df))
t_crit <- qt(1 - alpha/2, df)
```

```
ci_lower <- xbar - t_crit * se

ci_upper <- xbar + t_crit * se

#Print results

cat("n =", n, "\n") cat("Mean =", xbar, " SD =", s, "\n")

cat("SE =", round(se, 6), "\n")

 cat("t =", round(t_stat, 4), " df =", df, "\n")

cat("p-value =", format.pval(p_value, digits = 6), "\n")

cat("95% CI = [", round(ci_lower, 3), ",", round(ci_upper, 3), "]\n")
```

Output:

```
> # Print results
> cat("n =", n, "\n")
n = 20
> cat("Mean =", xbar, " SD =", s, "\n")
Mean = 8  SD = 2
> cat("SE =", round(se, 6), "\n")
SE = 0.447214
> cat("t =", round(t_stat, 4), " df =", df, "\n")
t = 17.8885  df = 19
> cat("p-value =", format.pval(p_value, digits = 6), "\n")
p-value = 2.39586e-13
> cat("95% CI = [", round(ci_lower, 3), ",", round(ci_upper, 3), "]\n")
95% CI = [ 7.064 , 8.936 ]
```

6. A company is testing two versions of a webpage (A and B) to determine which version leads to more sales. Version A was shown to 1,000 users and resulted in 120 sales. Version B was shown to 1,200 users and resulted in 150 sales. Perform an A/B test to determine if there is a statistically significant difference in the conversion rates between the two versions. Use a 5% significance level.

Program:

```
Given data

successes <- c(120, 150) # number of sales for A and B

n <- c(1000, 1200) # number of users for A and B

 alpha <- 0.05 # significance level

test_result <- prop.test(successes, n, correct = FALSE)

#Extract values

z_stat <- sqrt(test_result$statistic)

if (successes[1]/n[1] - successes[2]/n[2] < 0)

 { z_stat <- -z_stat}

p_value <- test_result$p.value

#Print results

cat("Z-statistic:", round(z_stat, 3), "\n") cat("P-value:", round(p_value, 5), "\n")

#Conclusion

if (p_value < alpha)

 { cat("Reject the null hypothesis: Significant difference in conversion rates between A and B.\n") }

else { cat("Fail to reject the null hypothesis: No significant difference in conversion rates between A and B.\n") }
```

Output:

```
Fail to reject the null hypothesis: No significant difference in conversion rates between A and
B.
```

>

7. You are comparing the average daily sales between two stores. Store A has a mean daily sales value of $1,000 with a standard deviation of $100 over 30 days, and Store B has a mean daily sales value of $950 with a standard deviation of $120 over 30 days. Conduct a two-sample t-test to determine if there is a significant difference between the average sales of the two stores at the 5% significance level.

Program:

```
mean_A <- 1000
std_A <- 100
n_A <- 30
mean_B <- 950
std_B <- 120
n_B <- 30
alpha <- 0.05

t_stat <- (mean_A - mean_B) / sqrt((std_A^2 / n_A) + (std_B^2 / n_B))

df <- ((std_A^2 / n_A + std_B^2 / n_B)^2) / ( ((std_A^2 / n_A)^2 / (n_A - 1)) + ((std_B^2 / n_B)^2 / (n_B - 1)) )

p_value <- 2 * (1 - pt(abs(t_stat), df))
cat("T-statistic:", round(t_stat, 3), "\n")
 cat("Degrees of freedom:", round(df, 2), "\n")
cat("P-value:", round(p_value, 5), "\n")
if (p_value < alpha) { cat("There is a significant difference between the average sales of the two stores.\n") }
else { cat("No significant difference between the average sales of the two stores.\n") }
```

Output:

```
No significant difference between the average sales of the two stores.
```

8. **A company collects data on employees' salaries and records their education level as a categorical variable with three levels: "High School", "Bachelor's", and "Master's". Fit a multiple linear regression model to predict salary using education level (as a factor variable) and years of experience. Interpret the coefficients for the education levels in the regression model.**

Program:

```
library(dplyr)

# dataset
df <- data.frame(
  Salary = c(40000, 50000, 60000, 45000, 55000, 65000, 48000, 58000, 70000, 62000),
  Education = c("High School", "Bachelor's", "Master's", "High School", "Bachelor's",
               "Master's", "High School", "Bachelor's", "Master's", "Bachelor's"),
  Experience = c(2, 5, 7, 3, 6, 8, 4, 5, 9, 6)
)
 # Set "High School" as the reference category
df$Education <- factor(df$Education, levels = c("High School", "Bachelor's", "Master's"))
 # Fit multiple linear regression model
model <- lm(Salary ~ Experience + Education, data = df)
 # Print model summary
summary(model)
```

Output:

```
Residuals:
   Min      1Q  Median      3Q     Max
-4000.0  -750.0    83.3   625.0  4000.0
 Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         30833.3     4547.7   6.780 0.000503 ***
Experience           4500.0     1392.4   3.232 0.017871 *
EducationBachelor's    666.7     4215.8   0.158 0.879539
EducationMaster's    -1833.3     7411.8  -0.247 0.812882
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

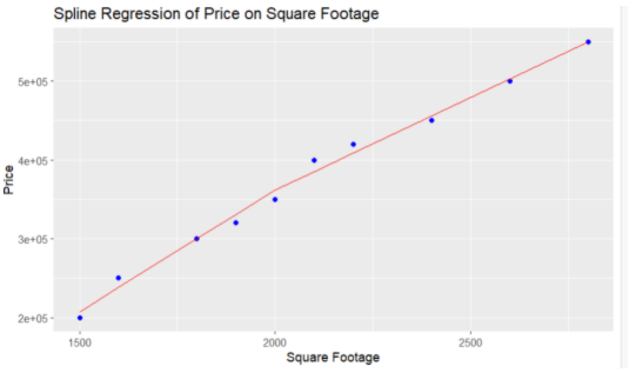
```
Residual standard error: 3114 on 6 degrees of freedom
Multiple R-squared:  0.9278,  Adjusted R-squared:  0.8918
F-statistic: 25.72 on 3 and 6 DF,  p-value: 0.0007993
```

9.You have data on housing prices and square footage and notice that the relationship between square footage and price is nonlinear. Fit a spline regression model to allow the relationship between square footage and price to change at 2,000 square feet. Explain how spline regression can capture different behaviours of the relationship before and after 2,000 square feet.

Program:

```
library(ggplot2)
# dataset
df <- data.frame(
  Price = c(200000, 250000, 300000, 320000, 350000, 400000, 420000, 450000, 500000,
550000),
  SqFt  = c(1500, 1600, 1800, 1900, 2000, 2100, 2200, 2400, 2600, 2800)
)

# Define spline term with knot at 2000
df$sqft_knot <- pmax(0, df$SqFt - 2000)
# Fit linear regression with spline
model <- lm(Price ~ SqFt + sqft_knot, data = df)
# Print summary
summary(model)
# Plot spline regression
ggplot(df, aes(x = SqFt, y = Price)) +
  geom_point(color = "blue") +
  geom_line(aes(y = predict(model)), color = "red") +
  labs(
    title = "Spline Regression of Price on Square Footage",
    x = "Square Footage",
    y = "Price"
  )
```

Output:



10 A hospital is using a Poisson regression model (a type of GLM) to predict the number of emergency room visits per week based on patient age and medical history. The model is given by: Log(λ) =2.5-0.03*Age+0.5*condition where λ is the expected number of visits per week, Age is the patient's age, and condition is a binary variable (1 if the patient has a chronic condition, 0 otherwise). Interpret the coefficients of Age and condition. What is the expected number of visits per week for a 60-year-old patient with a chronic condition? How would the expected number of visits change if the patient did not have a chronic condition?

Program:

```
# Given Poisson regression coefficients
intercept <- 2.5
coef_age <- -0.03
coef_condition <- 0.5

# Patient information
age <- 60
condition <- 1   # 1 = chronic condition, 0 = no condition

# Calculate log(lambda)
log_lambda <- intercept + coef_age * age + coef_condition * condition

# Convert to expected number of visits
expected_visits <- exp(log_lambda)
cat("Expected number of visits (with chronic condition):", round(expected_visits, 2), "\n")

# If patient does NOT have chronic condition
condition <- 0
log_lambda_no_condition <- intercept + coef_age * age + coef_condition * condition
expected_visits_no_condition <- exp(log_lambda_no_condition)
cat("Expected number of visits (without chronic condition):",
round(expected_visits_no_condition, 2), "\n")
```

Output:

```
Expected number of visits (with chronic condition): 3.32
Expected number of visits (without chronic condition): 2.01
```

11.A bakery claims that its new cookie recipe is lower in calories compared to the old recipe, which had a mean calorie count of 200. You sample 40 new cookies and find a mean of 190 calories with a standard deviation of 15 calories. Perform a one-tailed t-test to determine if the new recipe has significantly fewer calories at a 5% significance level.

Program:

```
# Given data
old_mean <- 200        # mean calories of old recipe
sample_mean <- 190     # mean calories of new recipe
sample_sd <- 15        # standard deviation of new recipe
n <- 40                # sample size
alpha <- 0.05          # significance level

# Calculate t-statistic
t_stat <- (sample_mean - old_mean) / (sample_sd / sqrt(n))

# Degrees of freedom
df <- n - 1

# One-tailed p-value (testing if new mean < old mean)
p_value <- pt(t_stat, df = df)   # pt() gives CDF of t-distribution

# Print results
cat("T-statistic:", round(t_stat, 3), "\n")
cat("P-value:", round(p_value, 5), "\n")

# Conclusion
if (p_value < alpha) {
  cat("Reject the null hypothesis: The new recipe has significantly fewer calories.\n")
} else {
  cat("Fail to reject the null hypothesis: No significant difference in calories.\n")
}
```

Output:

Reject the null hypothesis: The new recipe has significantly fewer calories.