Student Survey Solution

Statistics Assignment with R Solutions

Instructions

The following code assumes the dataset is loaded as `students_data` in R. Replace `path_to_file` with the actual path to your dataset to reproduce the results.

Part 1: Descriptive Statistics

1. **Basic Summary**

```r
# Load dataset
students_data <- read.csv("path_to_file")

# Summary for Age, Mid-exam score, and GPA
summary_stats <- students_data %>%
  select(Age, What.was.your.score.in.MID.exam., What.is.your.current.GPA.) %>%
  summarise_all(list(
    Mean = mean,
    Median = median,
    StdDev = sd,
    Min = min,
    Max = max,
    Mode = ~ names(sort(table(.), decreasing = TRUE))[1]
  ))
print(summary_stats)
```

2. **Frequency Distributions**

```r
# Frequency tables
gender_dist <- table(students_data$Gender)
year_dist <- table(students_data$Year.of.Study)
distance_dist <- table(students_data$How.far.is.your.home.from.the.university.)
print(gender_dist)
print(year_dist)
print(distance_dist)
```

3. **Percentile Analysis**

```r
# Percentiles for GPA
gpa_percentiles <- quantile(students_data$What.is.your.current.GPA., probs = c(0.25, 0.5, 0.75))
print(gpa_percentiles)
```

Part 2: Exploratory Data Analysis (EDA)

4. **Age Distribution**

```
# Histogram and density plot
hist(students_data$Age, main = "Age Distribution", xlab = "Age", col = "lightblue", breaks = 10)
print(summary(students_data$Age))  # Outlier detection
```

5. **Satisfaction Levels**

```
# Boxplots for satisfaction levels across genders
boxplot(students_data$How.satisfied.are.you.about.your.experience.at.the.University.of.Petra. ~ s
        main = "University Satisfaction by Gender", xlab = "Gender", ylab = "Satisfaction Level")
```

6. **Home Distance vs. Recommendation**

```
# Boxplot for recommendation likelihood
boxplot(students_data$How.likely.are.you.to.recommend.the.University.of.Petra.to.a.friend.or.colle
          students_data$How.far.is.your.home.from.the.university.,
        main = "Recommendation Likelihood by Home Distance", xlab = "Home Distance", ylab = "Recor
```

7. **Correlation Analysis**

```
# Correlation matrix
correlation_matrix <- cor(students_data %>% select(What.was.your.score.in.MID.exam., What.is.your
                                                   How.satisfied.are.you.about.your.experience.at
                                                   How.likely.are.you.to.recommend.the.University
print(correlation_matrix)
```

Part 3: Inferential Statistics

8. **Hypothesis Testing (T-Test)**

```
# GPA comparison between males and females
t_test_result <- t.test(What.is.your.current.GPA. ~ Gender, data = students_data)
print(t_test_result)
```

9. **ANOVA**

```
# Effect of year of study on GPA
anova_result <- aov(What.is.your.current.GPA. ~ Year.of.Study, data = students_data)
```

```
summary(anova_result)
```

10. **Chi-Square Test**

```
# Association between work status and satisfaction
work_satisfaction_table <- table(students_data$Do.you.work.while.studying., students_data$How.sat
chi_square_result <- chisq.test(work_satisfaction_table)
print(chi_square_result)
```

Part 4: Regression Analysis

11. **Simple Linear Regression**

```
# Predict GPA using mid-exam score
lm_model <- lm(What.is.your.current.GPA. ~ What.was.your.score.in.MID.exam., data = students_data
summary(lm_model)
```

12. **Multiple Linear Regression**

```
# Predict GPA using multiple variables
multi_lm_model <- lm(What.is.your.current.GPA. ~ What.was.your.score.in.MID.exam. +
                     How.satisfied.are.you.about.your.experience.at.the.University.of.Petra. +
                     How.satisfied.are.you.about.the.Business.Intelligence.major., data = student
summary(multi_lm_model)
```

13. **Logistic Regression**

```
# Logistic regression for recommendation likelihood
students_data$Recommend <- ifelse(students_data$How.likely.are.you.to.recommend.the.University.of
logistic_model <- glm(Recommend ~ How.satisfied.are.you.about.your.experience.at.the.University.o
                      How.satisfied.are.you.about.the.Business.Intelligence.major. + Year.of.Study
                      family = binomial, data = students_data)
summary(logistic_model)
```

Part 5: Specific Questions

14. **Key Insights**

```
# Average GPA by year
avg_gpa_by_year <- aggregate(What.is.your.current.GPA. ~ Year.of.Study, data = students_data, FUN
```

```
print(avg_gpa_by_year)

# Satisfaction by group
satisfaction_by_gender <- aggregate(How.satisfied.are.you.about.your.experience.at.the.University
                                    data = students_data, FUN = mean)
print(satisfaction_by_gender)
```

15. **Outliers**

```
# Identify outliers for GPA and mid-exam scores
boxplot(students_data$What.is.your.current.GPA., main = "GPA Outliers", horizontal = TRUE)
boxplot(students_data$What.was.your.score.in.MID.exam., main = "Mid-Exam Score Outliers", horizon
```

Part 6: Bonus Tasks

16. **Clustering**

```
# K-means clustering
set.seed(123)
clustering_data <- students_data %>% select(What.is.your.current.GPA., What.was.your.score.in.MID
kmeans_result <- kmeans(clustering_data, centers = 3)
students_data$Cluster <- kmeans_result$cluster
print(table(students_data$Cluster))
```