APPLIED STATISTICAL ANALYSIS LAB

NAME: ADITI KULKARNI

ROLL NO: 55

YEAR: SY

DIVISION: E (E3)

SRN NO: 202201893

ASSIGNMENT 6

STATEMENT: *To run the Independent-Samples T Test, to interpret the output and visualize the results with an error bar chart. Using the preexisting data file*

THEORY:

1. Data Import and Exploration:

- The code starts by importing a dataset from a CSV file using a file dialog.

- It then opens a data viewer to explore the imported dataset.

- The dimensions (rows and columns) of the dataset are calculated and displayed.

- The structure of the dataset, including data types of columns, is examined.

2. Creating a Contingency Table:

- The code constructs a contingency table (`tab1`) to analyze the relationship between two variables.

- Marginal sums of the table are computed and displayed.

3. Independent-Samples T-Test :

- The dataset is divided into two groups: 'males' and 'females,' based on the 'sex' column.

- An independent-samples t-test is conducted to compare the means of the 'G3' variable between these gender-based groups.

- The results of the t-test, including statistics like the t-statistic, degrees of freedom, and p-value, are stored in `t_test_result`.

4. Interpreting T-Test Results:

- The code prints out the results of the independent-samples t-test to determine if there's a statistically significant difference in 'G3' scores between males and females.

5. Creating an Error Bar Chart :

- Mean and standard deviation values for 'G3' scores are calculated separately for males and females.

- A data frame (`error_data`) is created to hold the mean and standard deviation values for both gender groups.

- An error bar chart is generated using the `ggplot2` library, showing the means of 'G3' scores for males and females with error bars representing standard deviations.

SOURCE CODE:

```
data <- read.csv(file.choose())

View(data)

dim(data)

str(data)
```

```
tab1 =table(data$ selling_price..in.thousands.
,data$km_driven..in.thousands.)
margin.table(tab1)
#Independent-Samples T-Test for 'G3'
males <- subset(data, sex == "M")
females <- subset(data, sex == "F")

t_test_result <- t.test(males$G3, females$G3)

# Interpret the T-Test results
print("Independent-Samples T-Test:")
print(t_test_result)

# Calculate means and standard deviations
mean_males <- mean(males$G3)
mean_females <- mean(females$G3)
sd_males <- sd(males$G3)
sd_females <- sd(females$G3)

# Create a data frame for the error bar chart
error_data <- data.frame(
  sex = c("Male", "Female"),
  mean = c(mean_males, mean_females),
  sd = c(sd_males, sd_females)
```

```
)

# Create the error bar chart

library(ggplot2)
library(dplyr)

p <- ggplot(error_data, aes(x = sex, y = mean)) +
  geom_bar(stat = "identity", fill = "lightsteelblue1") +
  geom_errorbar(aes(ymin = mean - sd, ymax = mean + sd), width = 0.4, colour = "orange", alpha = 0.9, size = 1.3) +
  labs(
    title = "Error Bar Chart for G3 by Sex",
    x = "Sex",
    y = "Mean G3"
  )
# Display the error bar chart
print(p)
```

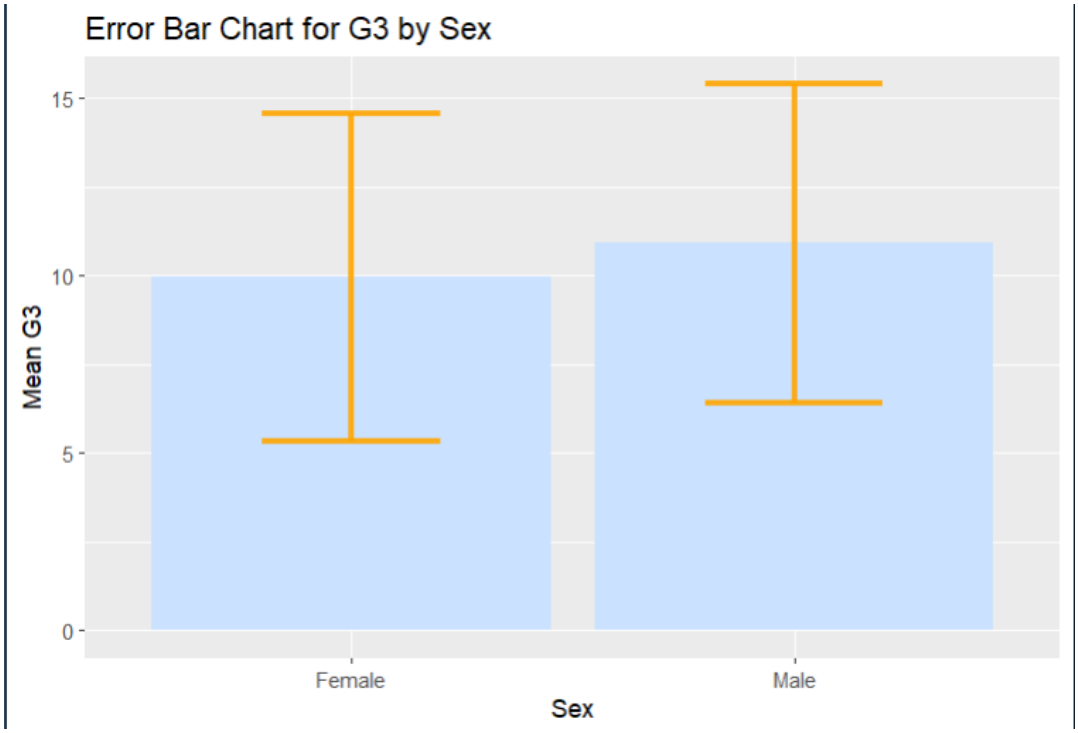
OUTPUT:

```
> dim(data)
[1] 395  33
> str(data)
'data.frame':   395 obs. of  33 variables:
 $ school    : chr  "GP" "GP" "GP" "GP" ...
 $ sex       : chr  "F" "F" "F" "F" ...
 $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address   : chr  "U" "U" "U" "U" ...
 $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
 $ Pstatus   : chr  "A" "T" "T" "T" ...
 $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
 $ Fjob      : chr  "teacher" "other" "other" "services" ...
 $ reason    : chr  "course" "course" "other" "home" ...
 $ guardian  : chr  "mother" "father" "mother" "mother" ...
 $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup : chr  "yes" "no" "yes" "no" ...
 $ famsup    : chr  "no" "yes" "no" "yes" ...
 $ paid      : chr  "no" "no" "yes" "yes" ...
 $ activities: chr  "no" "no" "no" "yes" ...
 $ nursery   : chr  "yes" "no" "yes" "yes" ...
 $ higher    : chr  "yes" "yes" "yes" "yes" ...
 $ internet  : chr  "no" "yes" "yes" "yes" ...
 $ romantic  : chr  "no" "no" "no" "yes" ...
 $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
 $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
 $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
 $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
 $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
> tab1 =table(data$ selling_price..in.thousands. ,data$km_driven..in.thousands.)
> margin.table(tab1)
[1] 0
```

```
> # Interpret the T-Test results
> print("Independent-Samples T-Test:")
[1] "Independent-Samples T-Test:"
> print(t_test_result)

        Welch Two Sample t-test

data:  males$G3 and females$G3
t = 2.0651, df = 390.57, p-value = 0.03958
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.04545244 1.85073226
sample estimates:
mean of x mean of y
10.914439  9.966346
```



CONCLUSION:

In summary, the code analyzes a dataset to assess whether there is a statistically significant difference in academic performance (variable 'G3') between males and females. It does so by conducting a t-test and visualizing the results with an error bar chart. The code's output provides insights into potential gender-based disparities in academic outcomes.