MANICALAND STATE UNIVERSITY

OF

APPLIED SCIENCES

FACULTY OF APPLIED SCIENCES & TECHNOLOGY

DEPARTMENT OF APPLIED STATISTICS

MODULE: STATISTICAL COMPUTING II

CODE: HAST215

SESSIONAL EXAMINATIONS

JUNE 2023

EXAMINER: MRS S MANDIZVIDZA

INSTRUCTIONS

1. Answer **All** in Section A.
2. Answer **three** questions in Section B.
3. Start a new question on a fresh page.
4. Total marks: 100.

Additional material(s)

- Statistical tables, Non-programmable electronic scientific calculator, List of formulae.

SECTION A [40 MARKS]

Answer **ALL** questions in this section

A 1

Write the R codes for finding the following

(a) Number of variables in a dataset. (2)

(b) Variance. (2)

(c) Mean. (2)

(d) Minimum. (2)

A 2

In detail explain

(a) the data import procedures in R Statistical package from Excel. (4)

(b) the advantages and disadvantages of using R over Stata or SPSS. (4)

(c) why are the data frames exported from R and how do your export from R to any Statistical package . (4)

A 3

Figure1 shows the summarises of the clay contents at the three depths.

```
> attach(obs)
> summary(Clay1); summary(Clay2); summary(Clay5)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   10.0    21.0    30.0    31.3    39.0    72.0

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    8.0    27.0    36.0    36.7    47.0    75.0

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   16.0    36.5    44.0    44.7    54.0    80.0
```

Figure 1: Summarises of the clay contents

(a) What does the summary say about the trend of clay content with depth?. (4)

(b) What evidence does the summary give that the distribution is somewhat symmetric?. (3)

A 4

Figure 2 shows the computed best estimate of the population mean of topsoil clay content from a sample, its 99% conffidence interval, and the probability that it is not equal to 30% clay as shown.

```
> t.test(Clay1, mu=30, conf.level=.99)

        One Sample t-test

data:  Clay1
t = 1.11, df = 146, p-value = 0.27
alternative hypothesis: true mean is not equal to 30
99 percent confidence interval:
 28.272 34.272
sample estimates:
mean of x
   31.272
```

Figure 2: Topsoil clay

(a) What is the estimated population mean and its 99% conffidence interval? Express this in plain language. (3)

(b) What is the probability that we would commit a Type I error if we reject the null hypothesis that the population mean is 30% clay?. (3)

A 5

The histogram function has a number of parameters which can be changed to make our plot look much nicer as shown in Figure 3 and 4.

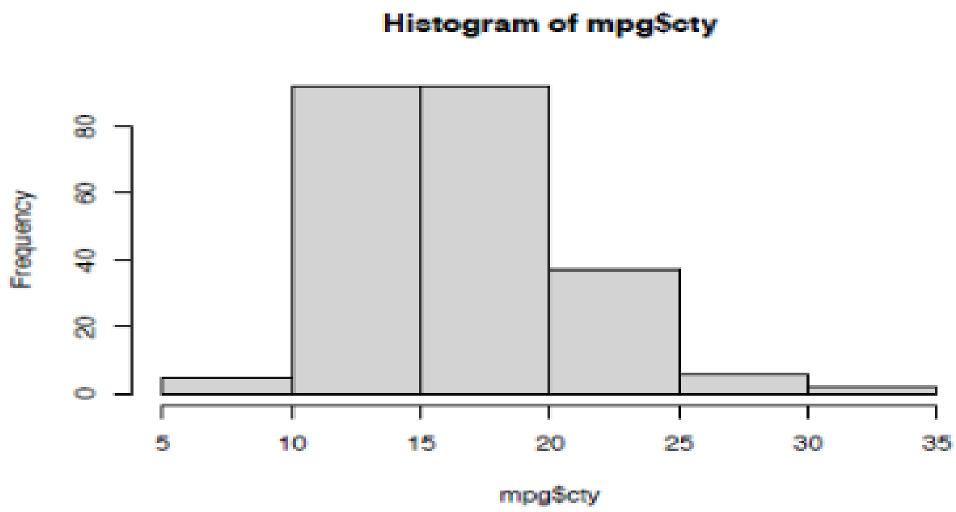

Figure 3: Histogram1

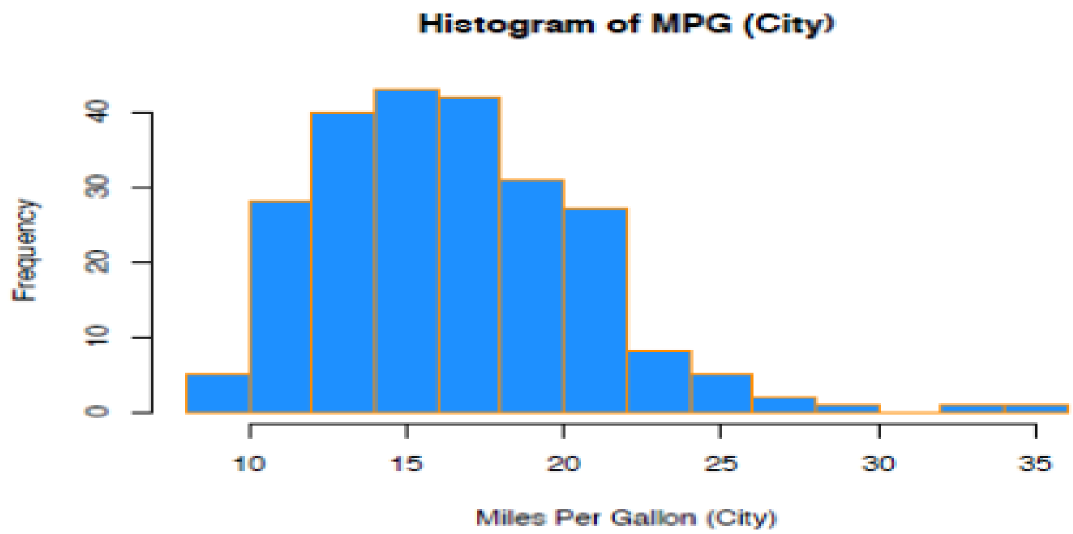

Figure 4: Histogram2

(a) Using Figure 3 and 4 can you state parameters which have been changed to make the plot look much nicer (4)

(b) Does the distribution look symmetric?, skewed?, or peaked? and support your answer with a valid reason. (3)

SECTION B [60 MARKS]

Answer any **THREE** questions in this section

A 6

(a) Carry out a hypothesis test R output in Figure 5, that is whether or not the mean weight of a certain species of some Turtle is equal to 310 kgs. Below is a simple random sample of Turtles with the following weights: (6)

`300, 315, 320, 311, 314, 309, 300, 308, 305, 303, 305, 301, 303`

```
> #define vector of turtle weights
> turtle_weights <- c(300, 315, 320, 311, 314, 309, 300, 308, 305, 303, 305, 301, 303)
>
> #perform one sample t-test
> t.test(x = turtle_weights, mu = 310)

        One Sample t-test

data:  turtle_weights
t = -1.5848, df = 12, p-value = 0.139
alternative hypothesis: true mean is not equal to 310
95 percent confidence interval:
 303.4236 311.0379
sample estimates:
mean of x
 307.2308
```

Figure 5: Species of Turtles

(b) Carry out a hypothesis test using R output in Figure 6, that is whether or not the mean weight between two different species of Turtles is equal. (7)

Sample 1: `300, 315, 320, 311, 314, 309, 300, 308, 305, 303, 305, 301, 303`

Sample 2: `335, 329, 322, 321, 324, 319, 304, 308, 305, 311, 307, 300, 305`

```
> #define vector of turtle weights for each sample
> sample1 <- c(300, 315, 320, 311, 314, 309, 300, 308, 305, 303, 305, 301, 303)
> sample2 <- c(335, 329, 322, 321, 324, 319, 304, 308, 305, 311, 307, 300, 305)
>
> #perform two sample t-test
> t.test(x = sample1, y = sample2)

        Welch Two Sample t-test

data:  sample1 and sample2
t = -2.1009, df = 19.112, p-value = 0.04914
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.73862953  -0.03060124
sample estimates:
mean of x mean of y
 307.2308  314.6154
```

Figure 6: Species of Turtles

Suppose we want to know whether or not a certain training program is able to increase the maximum vertical jump (cm) of basketball players. A simple random sample of 12 college basketball players was recruited and measured each of their maximum vertical jumps was recorded. Each player used the training program for one month and then was measured their maximum vertical jump again at the end of the month.Given the following data which shows the maximum jump height(cm) before and after using the training program for each player:

Before: `22, 24, 20, 19, 19, 20, 22, 25, 24, 23, 22, 21`

After: `23, 25, 20, 24, 18, 22, 23, 28, 24, 25, 24, 20`

```
> #define before and after maximum jump heights
> before <- c(22, 24, 20, 19, 19, 20, 22, 25, 24, 23, 22, 21)
> after <- c(23, 25, 20, 24, 18, 22, 23, 28, 24, 25, 24, 20)
>
> #perform paired samples t-test
> t.test(x = before, y = after, paired = TRUE)

        Paired t-test

data:  before and after
t = -2.5289, df = 11, p-value = 0.02803
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.3379151 -0.1620849
sample estimates:
mean of the differences
                  -1.25
```

Figure 7: Maximum jump height

(c) Carry out a hypothesis test using the R output in Figure 7, to compare the means of two samples. (7)

A 7

The speed of a car affects its stopping distance, that is, how far it travels before it comes to a stop. The simple linear regression model is defined by, $Y_i = \beta_0 + \beta_1 X_i + e_0$ where $e_0 \sim N(0, \sigma^2)$ for the car speet and its stopping distance.

We can use R to check that our data meet the four main assumptions for linear regression as shown in Figure 8,9 and 10

(a) Use the output to test whether the assumptions have been met indicating clearly the assumption, its corresponding test and the conclusion on the test. (5)

- Normality.
- Independence of observations(no autocorrelation).
- Linearity.
- Homoscedasticity(homogeneity of variance).

```
> cor(cars$speed,cars$dist )
[1] 0.8068949
```

Figure 8: Car correlation

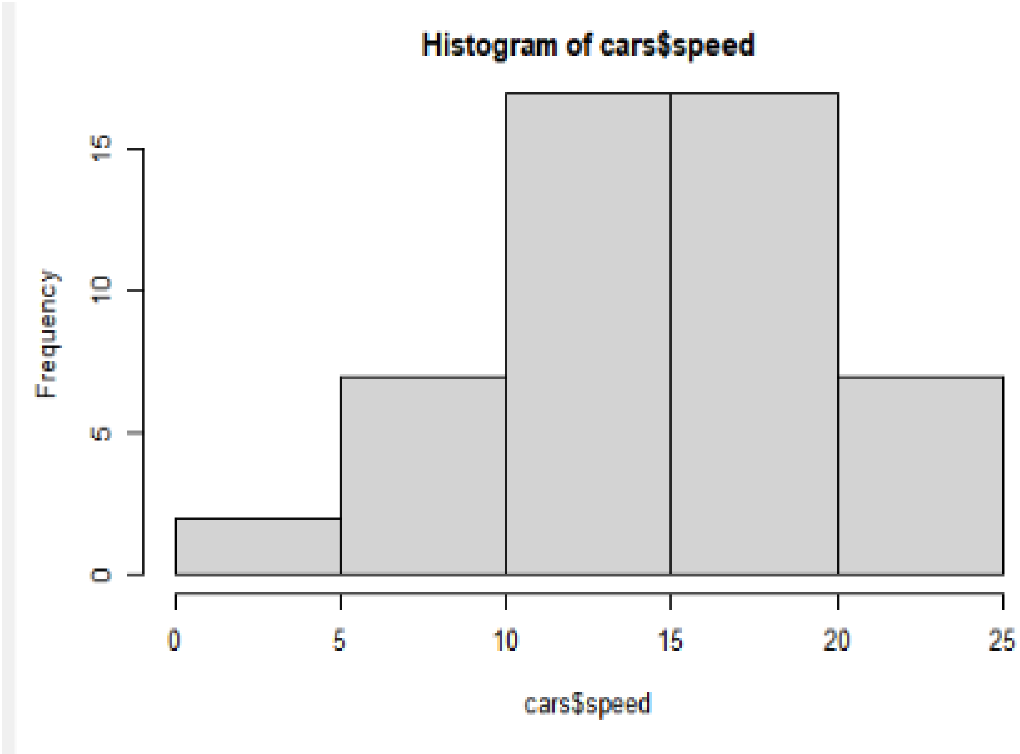

Figure 9: Car Histogram

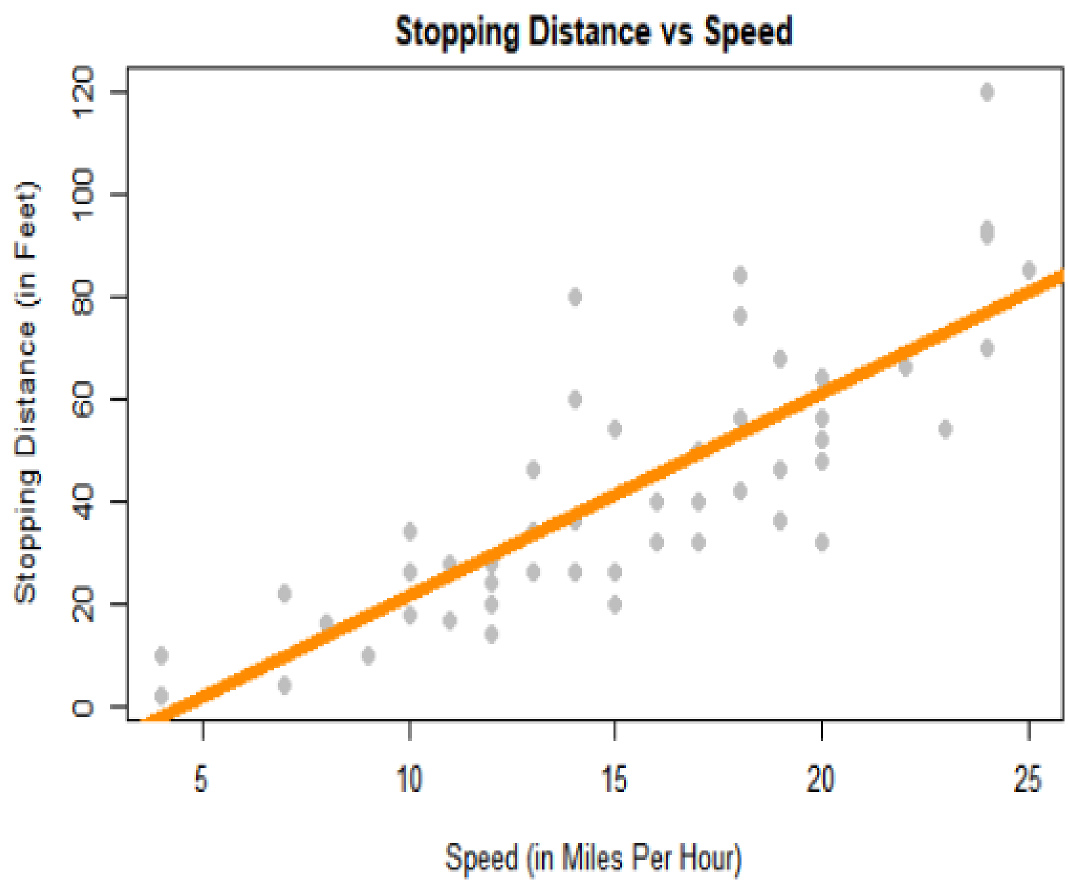

Figure 10: Car scatter plot

Using Figure 11 answer the following questions

```
> stop_dist_model = lm(dist ~ speed, data = cars)
> summary(stop_dist_model)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Figure 11: Car stopping distance

(b) Fit the linear regression model. (2)

(c) Comment on the significance of of the model coefficients. (5)

(d) Predict the stop-dist-model, when speed = 8. (1)

(e) What does the residual standard error tells you. (2)

(f) Carry out a significance test on the linear relationship between speed and stopping distance. (5)

A 8

Crop yield data was modelled below as a function of the type of fertilizer used and planting density.

(a) Determine whether there is significant variation among the crop yield formed by the type of fertilizer using R output in Figure 12. (4)

```
> one.way <- aov(yield ~ fertilizer, data = crop.data)
>
> summary(one.way)
            Df Sum Sq Mean Sq F value Pr(>F)
fertilizer   2   6.07  3.0340   7.863  7e-04 ***
Residuals   93  35.89  0.3859
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Figure 12: Crop yield one-way ANOVA

(b) Determine whether there is significant variation in crop yield as a function of type of fertilizer and planting density using R output in Figure 13. (4)

```
> two.way <- aov(yield ~ fertilizer + density, data = crop.data)
>
> summary(two.way)
            Df Sum Sq Mean Sq F value   Pr(>F)
fertilizer   2  6.068   3.034   9.073 0.000253 ***
density      1  5.122   5.122  15.316 0.000174 ***
Residuals   92 30.765   0.334
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 13: Crop yield two-way ANOVA

Sometimes you have reason to think that two of your independent variables have an interaction effect rather than an additive effect, that it is possible that planting density affects the plant's ability to take up fertilizer. This might influence the effect of fertilizer type in a way that isn't accounted for in the two-way model.

(c) Test whether two variables have an interaction effect using R out put in Figure 14. (4)

```
> interaction <- aov(yield ~ fertilizer*density, data = crop.data)
>
>
> summary(interaction)
                   Df Sum Sq Mean Sq F value   Pr(>F)    
fertilizer          2  6.068   3.034   9.001 0.000273 ***
density             1  5.122   5.122  15.195 0.000186 ***
fertilizer:density  2  0.428   0.214   0.635 0.532500    
Residuals          90 30.337   0.337                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 14: Crop yield interaction effect

(d) Find the best-fit model using the Akaike Information Criterion (AIC) using R output in Figure 15. (4)

```
> model.set <- list(one.way, two.way, interaction)
> model.names <- c("one.way", "two.way", "interaction")
>
> aictab(model.set, modnames = model.names)

Model selection based on AICc:

            K   AICc Delta_AICc AICcWt Cum.Wt     LL
two.way     5 173.86       0.00   0.83   0.83 -81.59
interaction 7 177.12       3.26   0.16   1.00 -80.92
one.way     4 186.41      12.56   0.00   1.00 -88.99
```

Figure 15: Crop yield AIC

To check whether the model fits the assumption of homoscedasticity, below is the model diagnostic plots.

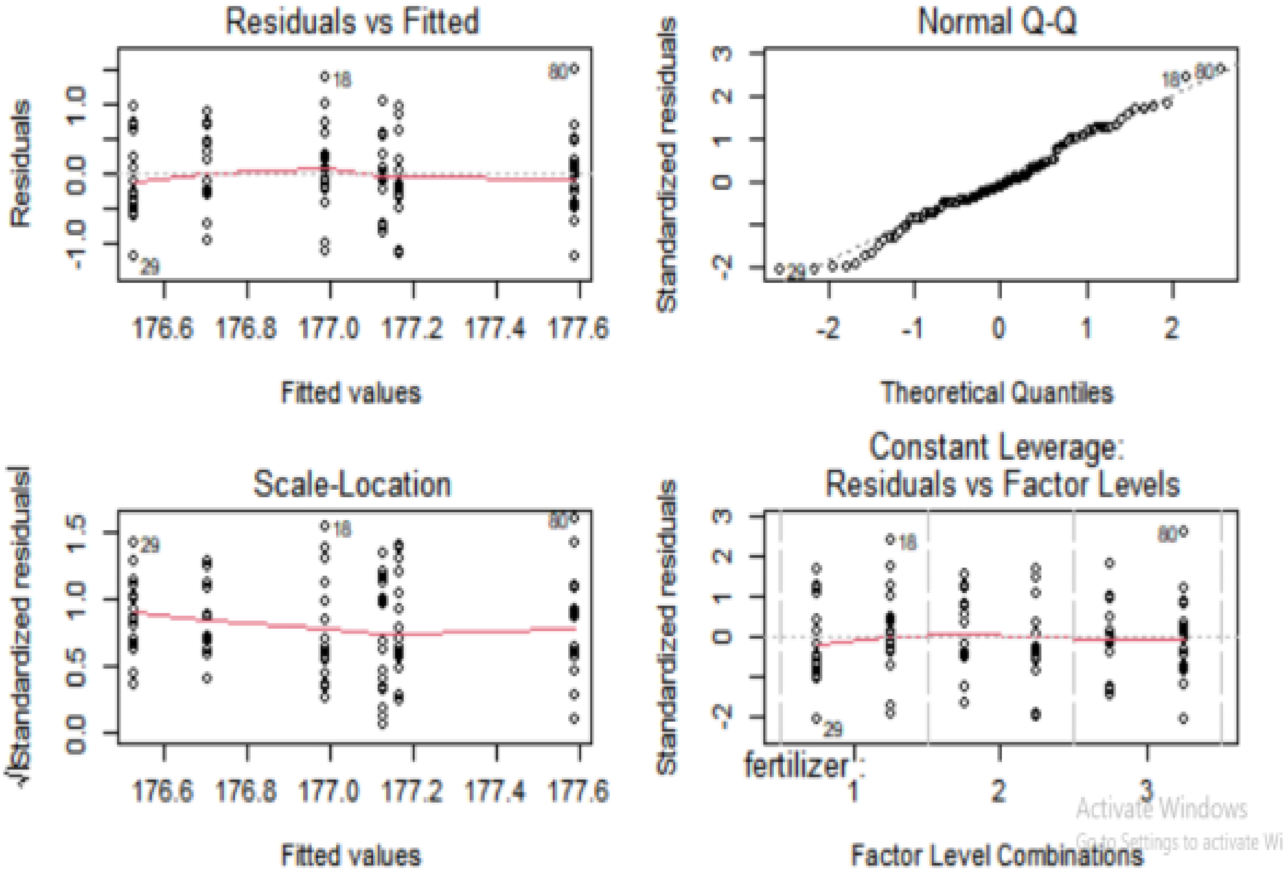

Figure 16: Model diagnostic plots

(e) Using R out put in Figure 16, Comment on the output. (4)

A 9

Using the survey data in the MASS library which represents the data from a survey conducted on student, Figure 17 shows the output in R.

```
> library(MASS)
> print(str(survey))
'data.frame':   237 obs. of  12 variables:
 $ Sex   : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
 $ Wr.Hnd: num  18.5 19.5 18 18.8 20 18 17.7 17 20 18.5 ...
 $ NW.Hnd: num  18 20.5 13.3 18.9 20 17.7 17.7 17.3 19.5 18.5 ...
 $ W.Hnd : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 2 2 2 ...
 $ Fold  : Factor w/ 3 levels "L on R","Neither",..: 3 3 1 3 2 1 1 3 3 3 ...
 $ Pulse : int  92 104 87 NA 35 64 83 74 72 90 ...
 $ Clap  : Factor w/ 3 levels "Left","Neither",..: 1 1 2 2 3 3 3 3 3 3 ...
 $ Exer  : Factor w/ 3 levels "Freq","None",..: 3 2 2 2 3 3 1 1 3 3 ...
 $ Smoke : Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2 2 2 2 2 2 2 ...
 $ Height: num  173 178 NA 160 165 ...
 $ M.I   : Factor w/ 2 levels "Imperial","Metric": 2 1 NA 2 2 1 1 2 2 2 ...
 $ Age   : num  18.2 17.6 16.9 20.3 23.7 ...
....
```

Figure 17: Survey data

(a) Desribe the type of variables that we have in the survey dataset. (3)

Figure 18 shows the dataset has many Factor variables which can be considered as categorical variables. For our model, have considered the variables "Exer" and "Smoke".The Smoke column records the students smoking habits while the Exer column records their exercise level.

```
> # Create a data frame from the main data set.
> stu_data = data.frame(survey$Smoke,survey$Exer)
>
> # Create a contingency table with the needed variables.
> stu_data = table(survey$Smoke,survey$Exer)
>
> print(stu_data)

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
> # applying chisq.test() function
> print(chisq.test(stu_data))
```

Figure 18: "Exer" and "Smoke"

(b) Test the hypothesis whether the students smoking habit is independent of their exercise level at 0.05 significance level using the R output as shown in Figure 19. (5)

```
        Pearson's Chi-squared test

data:  stu_data
X-squared = 5.4885, df = 6, p-value = 0.4828
```

Figure 19: Chisquare

Write the R codes for finding the following

(c) Range. (2)

(d) Total number of columns in the dataset. (2)

In detail explain

(e) features found in the R Studio integrated development environment intrerface. (4)

(f) the difference betwean R core team and R Studio. (4)