Answer Sheet
Final Project Spring 2022
STA 201L

Name:___Omar Al – Khazali b00088673_____

Question One:

I cleaned my data in two ways, one using Excel, and once using an R program. Both had the same conceptual conditions which included; removing classifications of qualitative variables that are outside of their required bounds, such as the Academic column which had a record holding the value “5”, that is outside the bounds of the variable (1 to 4), thus is removed.

In addition, the five number summary could easily be found with the summary() function in R. Thus, providing the minimum, maximum, mean, median, quartile 1, and quartile 2. Afterwards, I repeated the process for every relevant column, such as GPA, Height, Weight, Allowance, and Studying Time. All of which needed calculations for interquartile range, and holding bounds for outlier detection. Otherwise, all of the other data was parsed through R by only allowing values which were within the bounds given to each variable, such as GPA existing between 0 and 4, or college as an integer between 1 and 4. The majority of the cleaning was done in R, which took away unnecessary values automatically without needing to sift through them.

Cleaning Code:

```
library(readxl)
library(writexl)

data <- read_excel("./ProjectData.xlsx")
length(data$ID)

Q1_of_height        <-  quantile(na.omit(data$Height), prob=c(.25))
median_of_height    <-  quantile(na.omit(data$Height), prob=c(.5))
Q3_of_height        <-  quantile(na.omit(data$Height), prob=c(.75))

lower_bound_height  <-  Q1_of_height - (1.5 * Q3_of_height - Q1_of_height)
upper_bound_height  <-  Q3_of_height + (1.5 * Q3_of_height - Q1_of_height)

Q1_of_weight        <-  quantile(na.omit(data$Weight), prob=c(.25))
median_of_weight    <-  quantile(na.omit(data$Weight), prob=c(.5))
Q3_of_weight        <-  quantile(na.omit(data$Weight), prob=c(.75))

lower_bound_weight  <-  Q1_of_weight - (1.5 * Q3_of_weight - Q1_of_weight)
upper_bound_weight  <-  Q3_of_weight + (1.5 * Q3_of_weight - Q1_of_weight)
```

```r
Q1_of_allowance     <-  quantile(na.omit(data$Allowance), prob=c(.25))
median_of_allowance <-  quantile(na.omit(data$Allowance), prob=c(.5))
Q3_of_allowance     <-  quantile(na.omit(data$Allowance), prob=c(.75))

lower_bound_allowance  <-  Q1_of_allowance - (1.5 * Q3_of_allowance - Q1_of_allowance)
upper_bound_allowance  <-  Q3_of_allowance + (1.5 * Q3_of_allowance - Q1_of_allowance)

Q1_of_StudyTime     <-  quantile(na.omit(data$Sudying_Time), prob=c(.25))
median_of_StudyTime <-  quantile(na.omit(data$Sudying_Time), prob=c(.5))
Q3_of_StudyTime     <-  quantile(na.omit(data$Sudying_Time), prob=c(.75))

lower_bound_StudyTime  <-  Q1_of_StudyTime - (1.5 * Q3_of_StudyTime - Q1_of_StudyTime)
upper_bound_StudyTime  <-  Q3_of_StudyTime + (1.5 * Q3_of_StudyTime - Q1_of_StudyTime)

data <- na.omit(data)
data
new_data <- data[ which((data$Gender <= 2 | data$Gender >= 1)
                      &(data$College >=1 | data$College <=4)
                      &(data$Academic>=1 | data$Academic<=4)
                      &(data$Origin>= 1 | data$Origin <= 4)
                      &(data$Living ==1 | data$Living ==2)
                      &(data$GPA %in% c(0:4))
                      &(data$Height %in% c(lower_bound_height:upper_bound_height))
                      &(data$Weight %in% c(lower_bound_weight:upper_bound_weight))
                      &(data$Allowance %in% c(lower_bound_allowance:upper_bound_allowance))
                      &(data$Sudying_Time %in% c(lower_bound_StudyTime:upper_bound_StudyTime))), ]

new_data
write_xlsx(new_data, "NewProjectData.xlsx")
```

Q2: (b)
Code & Output:

```
> library(readxl)
>
> data <- read_excel(file.choose())

>
> data$Gender <- as.factor(data$Gender)
> data$College <- as.factor(data$College)
> data$Academic <- as.factor(data$Academic)
> data$Origin <- as.factor(data$Origin)
> data$Living <- as.factor(data$Living)
>
```

Q3: (c)

Code & Output:

```
> library(readxl)
>
> data <- read_excel(file.choose())
>
> data$Gender <- as.factor(data$Gender)
> data$College <- as.factor(data$College)
> data$Academic <- as.factor(data$Academic)
> data$Origin <- as.factor(data$Origin)
> data$Living <- as.factor(data$Living)
>
> set.seed(88673)
> sample_data<-data.frame(data[sample(1:dim(data)[1], size=100),])
>
```

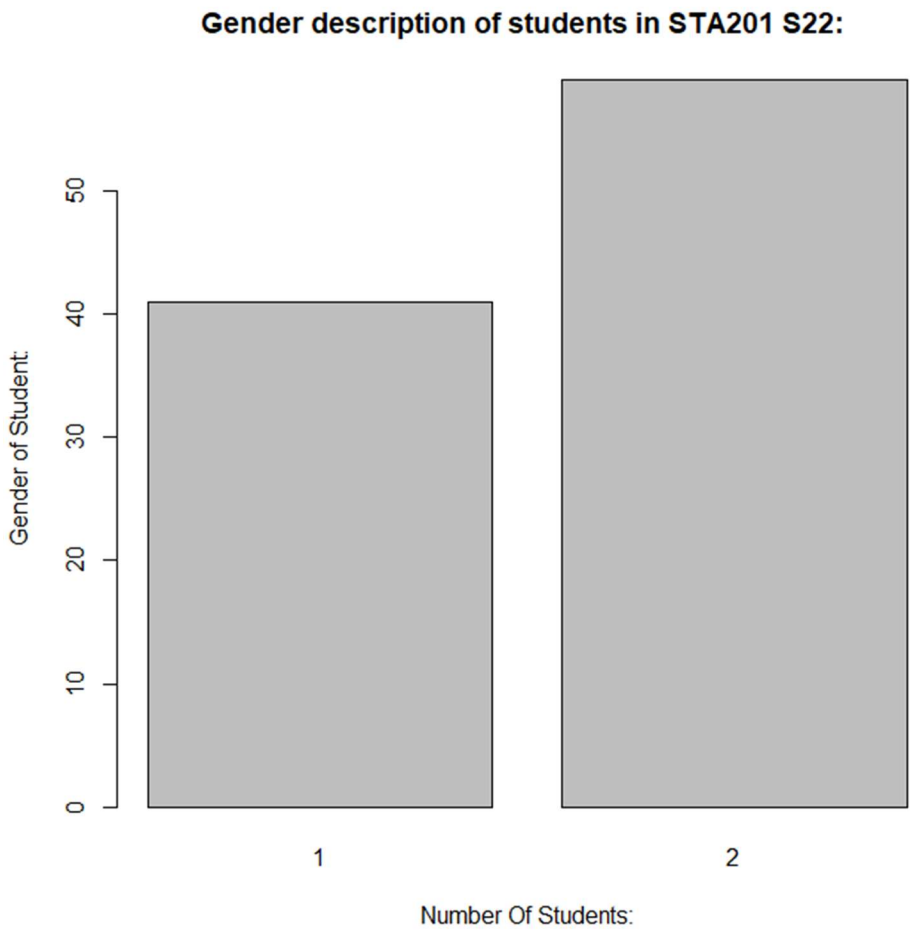
Q4: (d)(i)

Code & Output:

```
> library(readxl)
>
> data <- read_excel(file.choose())
>
> data$Gender <- as.factor(data$Gender)
> data$College <- as.factor(data$College)
> data$Academic <- as.factor(data$Academic)
> data$Origin <- as.factor(data$Origin)
> data$Living <- as.factor(data$Living)
>
> set.seed(88673)
> sample_data<-data.frame(data[sample(1:dim(data)[1], size=100),])
>
> ################## Gender ######################
> print("Gender: [M=1;F=2]:")
[1] "Gender: [M=1;F=2]:"
> print("Barplot of Gender")
[1] "Barplot of Gender"
> summary(sample_data$Gender)
 1  2
41 59
> barplot(table(sample_data$Gender), main="Gender description of students in STA201 S22:
",xlab="Number Of Students:",ylab="Gender of Student:")
>
> print("Percentage of Females over Males")
[1] "Percentage of Females over Males"
> print(41)
[1] 41
> print("Percentage of Males over Females")
[1] "Percentage of Males over Females"
> print(100 - 41)
[1] 59
>
>
> ################## Academic Status ######################
>
> print("Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: ")
[1] "Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: "
> print("Summary statistics for Academic Status: ")
[1] "Summary statistics for Academic Status: "
> summary(sample_data$Academic)
 1  2  3  4
35 19 28 18
> print("Barchart for Academic status: ")
[1] "Barchart for Academic status: "
> barplot(table(sample_data$Academic), main = "Academic Status of Students in STA 201
S22",xlab="Academic Status", ylab = "Number of Students")
>
> print("Percentage of Freshman:")
[1] "Percentage of Freshman:"
> print(35)
[1] 35
> print("Percentage of Sophomores:")
[1] "Percentage of Sophomores:"
> print(19)
[1] 19
> print("Percentage of Juniors:")
[1] "Percentage of Juniors:"
> print(28)
[1] 28
> print("Percentage of Seniors:")
```

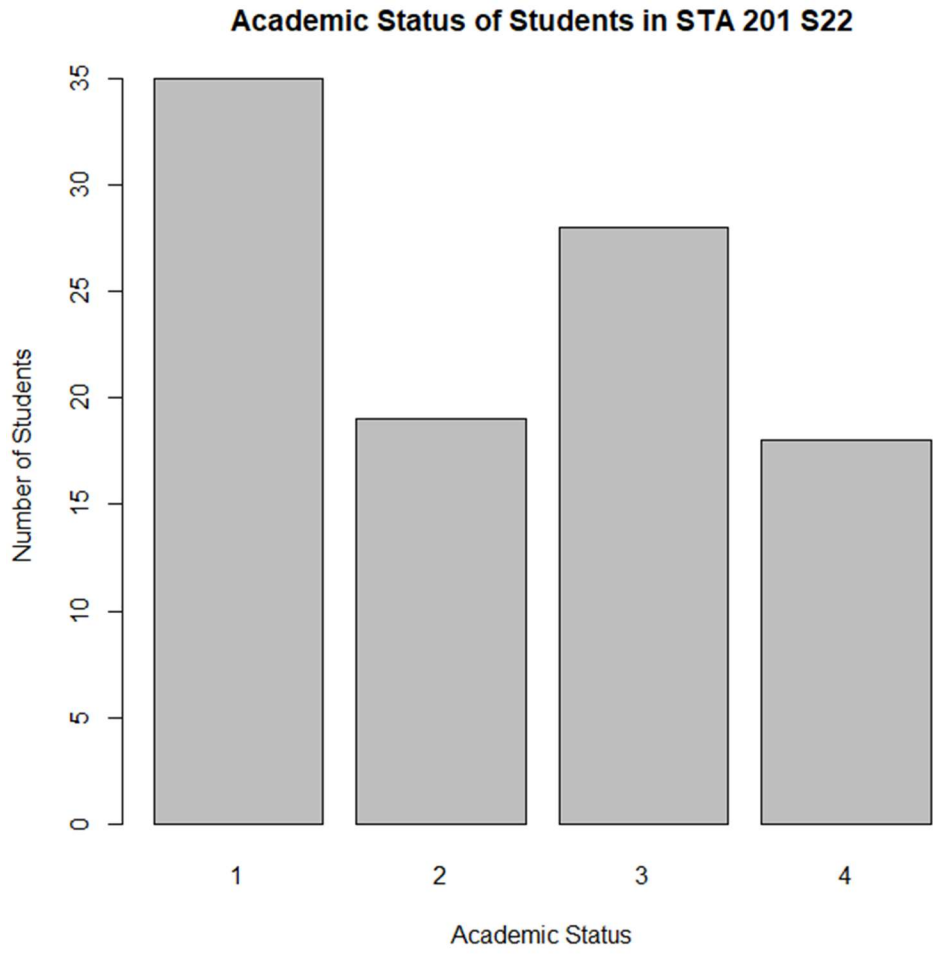
```
[1] "Percentage of Seniors:"
> print(18)
[1] 18
>
>
> ################## Region of Origin ######################
> print("Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:")
[1] "Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:"
> print("Summary statistics for Region of Living:")
[1] "Summary statistics for Region of Living:"
> summary(sample_data$Origin)
 1  2  3  4
27 21 38 14
>
>
> print("Barplot for Origin:")
[1] "Barplot for Origin:"
> barplot(table(sample_data$Origin),main="Region of Living of students in STA 201 S22",xlab="Region
of Living", ylab="Number of students")
>
> print("Percentage of GCC:")
[1] "Percentage of GCC:"
> print(27)
[1] 27
> print("Percentage of Africa:")
[1] "Percentage of Africa:"
> print(21)
[1] 21
> print("Percentage of ME:")
[1] "Percentage of ME:"
> print(38)
[1] 38
> print("Percentage of Other:")
[1] "Percentage of Other:"
> print(14)
[1] 14
>
>
> ################## Place of Living ######################
> print("Place of Living: [On campus=1;Off campus=2]")
[1] "Place of Living: [On campus=1;Off campus=2]"
> print("Summary Statistics for Place of Living:")
[1] "Summary Statistics for Place of Living:"
> summary(sample_data$Living)
 1  2
43 57
> print("Barplot for Place of Living")
[1] "Barplot for Place of Living"
> barplot(table(sample_data$Living), main = "Place of living of students in STA 201 S22",xlab="Place
of living", ylab="Number of studnets")
>
> print("Percentage of Students on Campus")
[1] "Percentage of Students on Campus"
> print(43)
[1] 43
> print("Percentage of Students off Campus")
[1] "Percentage of Students off Campus"
> print(57)
[1] 57
>
```

Gender:



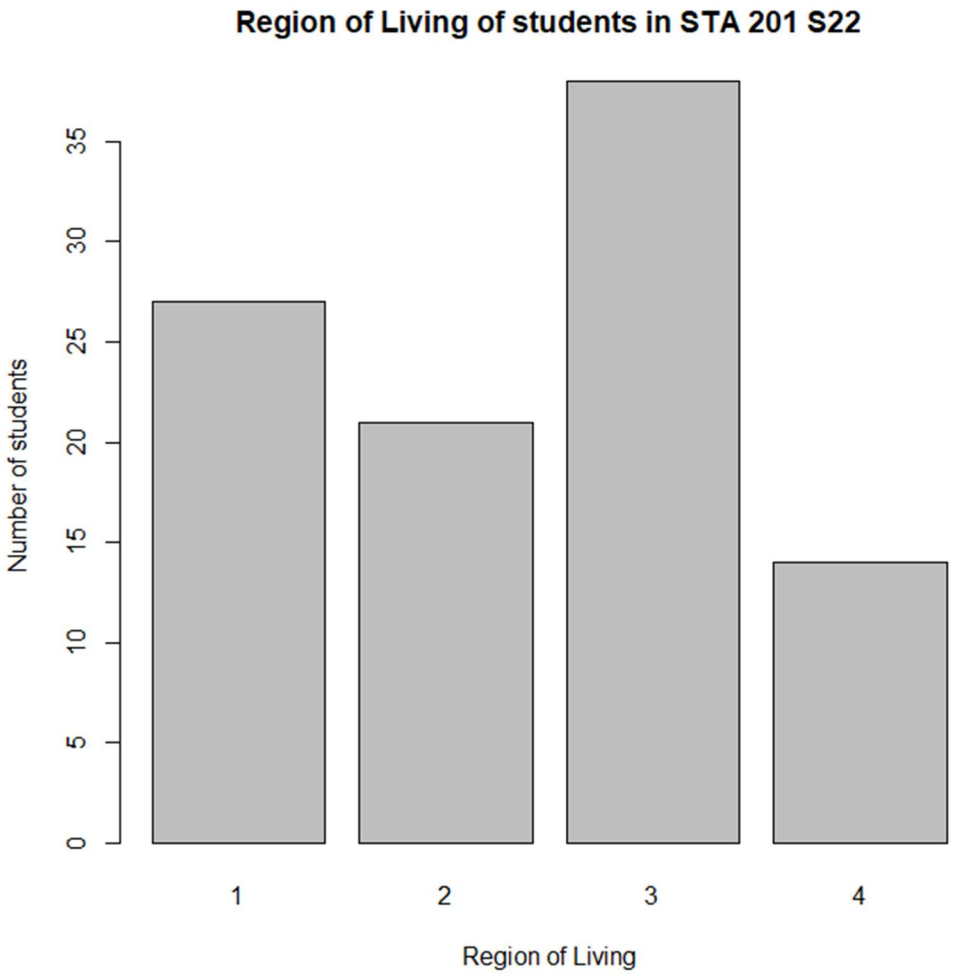
A random sample of 100 people were taken from Female and Male students enrolled in STA 201 S22 at AUS. In this dataset, it is known that the value 1 refers to males, and the value 2 refers to females. The summary statistics shown in the output and the barplot above both show that the proportion of males is 41 out of 100, and proportion of females is 59 out of 100. Thus showing that there are 18 more female students than male students in the class.

Academic Status:



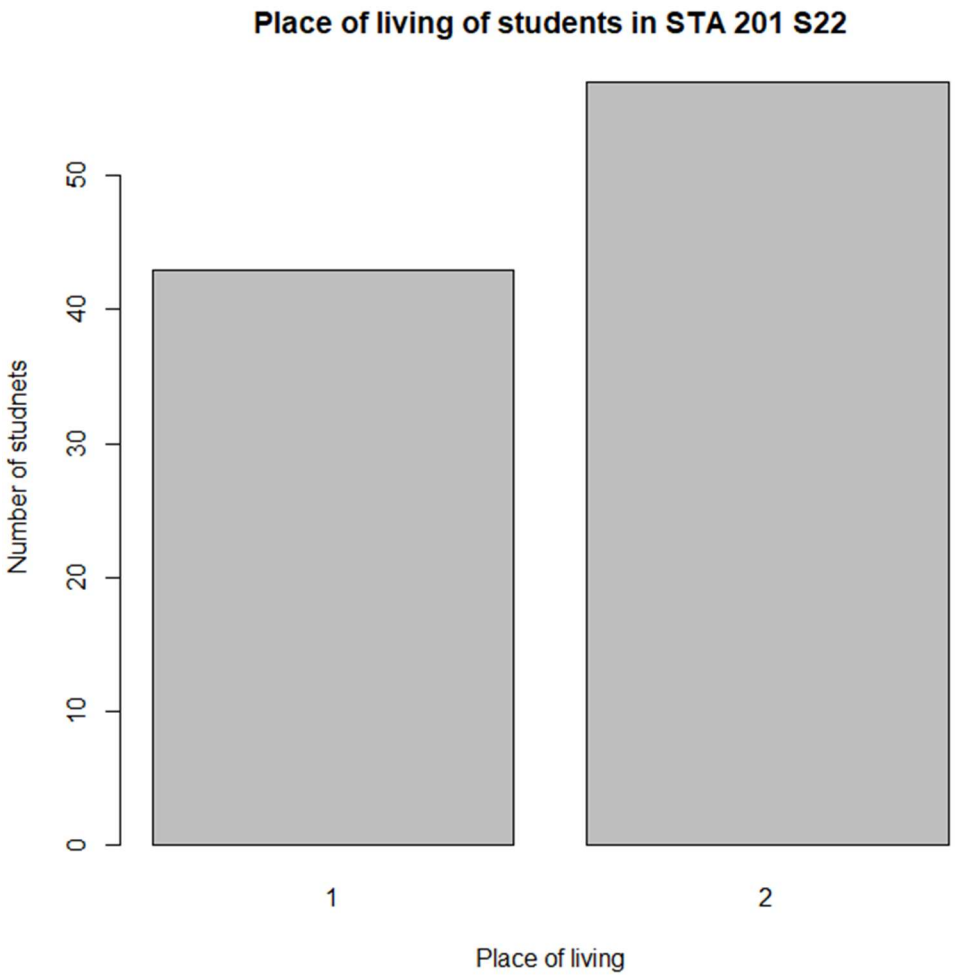
A random sample of 100 students were taken from Freshmen, Sophomore, Junior, and Senior years who were enrolled in STA 201 S22. The summary statistics and the bar chart shown above show that the randomly sampled set contains mostly freshmen, then juniors, then sophomores, then seniors in descending order. The statistics show that the number of freshmen is 35, the number of sophomores is 19, the number of juniors is 28, and finally the number of seniors is 18. Since it is out of 100 sampled students, the percentage value is the same as the proportion value.

Region of Living:



A random sample of 100 people were taken from GCC, Africa, Middle East, and Other countries in STA 201 S22. In the dataset, GCC=1 Africa=2 Middle East=3 Other=4. The summary statistics and bar chart both show that the randomly sampled number of Middle East students are the most enrolled in the courses. In descending order, the statistics show that the number of students in the relevant countries are from the Middle East, GCC, Africa, and Other in descending order. With the Middle East variable taking the value of 38, GCC taking the value 27, Africa taking the value 21, and Other taking the value of 14.

Place of Living:



A random sample of 100 people were taken from students living On Campus and students living Off Campus enrolled in STA 201 S22at AUS. In the dataset, On Campus = 1 and Off Campus = 2. The summary statistics and bar chart both show that the randomly sampled number of Off Campus students enrolled in STA 201 S22 at AUS are greater than the randomly sampled number of On Campus students. The number of Off Campus students is 57 and the number of On Campus is 43, which means there is a higher percentage of Off Campus students (57%) than On Campus (43%). There are 14 more Off Campus students than On Campus students enrolled in the courses at AUS during the specified time period.

Q5: (d)(ii)

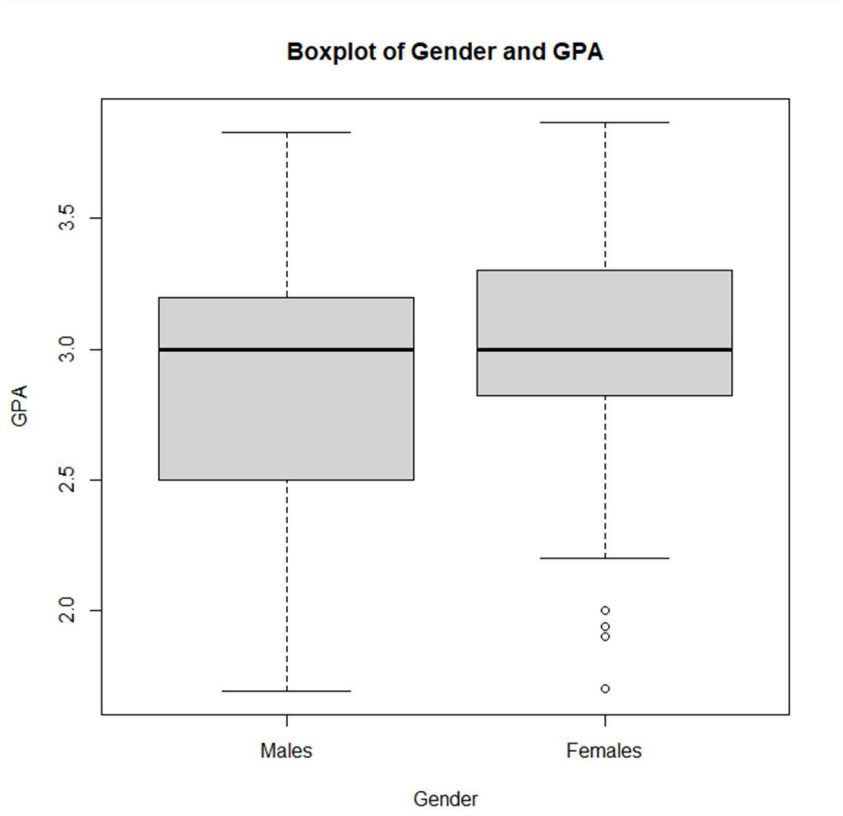
Code & Output:

```
> library(readxl)
>
> data <- read_excel(file.choose())
>
> data$Gender <- as.factor(data$Gender)
> data$College <- as.factor(data$College)
> data$Academic <- as.factor(data$Academic)
> data$Origin <- as.factor(data$Origin)
> data$Living <- as.factor(data$Living)
>
> set.seed(88673)
> sample_data<-data.frame(data[sample(1:dim(data)[1], size=100),])
>
> ################## Gender ######################
> print("Gender: [M=1;F=2]:")
[1] "Gender: [M=1;F=2]:"
> print("Barplot of Gender")
[1] "Barplot of Gender"
> summary(sample_data$Gender)
 1  2
41 59
> barplot(table(sample_data$Gender), main="Gender description of students in STA201 S22:
",xlab="Number Of Students:",ylab="Gender of Student:")
>
> print("Percentage of Females over Males")
[1] "Percentage of Females over Males"
> print(41)
[1] 41
> print("Percentage of Males over Females")
[1] "Percentage of Males over Females"
> print(100 - 41)
[1] 59
>
>
> ################## Academic Status ######################
>
> print("Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: ")
[1] "Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: "
> print("Summary statistics for Academic Status: ")
[1] "Summary statistics for Academic Status: "
> summary(sample_data$Academic)
 1  2  3  4
35 19 28 18
> print("Barchart for Academic status: ")
[1] "Barchart for Academic status: "
> barplot(table(sample_data$Academic), main = "Academic Status of Students in STA 201
S22",xlab="Academic Status", ylab = "Number of Students")
>
> print("Percentage of Freshman:")
[1] "Percentage of Freshman:"
> print(35)
[1] 35
> print("Percentage of Sophomores:")
[1] "Percentage of Sophomores:"
> print(19)
[1] 19
> print("Percentage of Juniors:")
[1] "Percentage of Juniors:"
> print(28)
[1] 28
> print("Percentage of Seniors:")
```

```
[1] "Percentage of Seniors:"
> print(18)
[1] 18
>
>
> ################## Region of Origin ######################
> print("Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:")
[1] "Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:"
> print("Summary statistics for Region of Living:")
[1] "Summary statistics for Region of Living:"
> summary(sample_data$Origin)
 1  2  3  4
27 21 38 14
>
>
> print("Barplot for Origin:")
[1] "Barplot for Origin:"
> barplot(table(sample_data$Origin),main="Region of Living of students in STA 201 S22",xlab="Region
of Living", ylab="Number of students")
>
> print("Percentage of GCC:")
[1] "Percentage of GCC:"
> print(27)
[1] 27
> print("Percentage of Africa:")
[1] "Percentage of Africa:"
> print(21)
[1] 21
> print("Percentage of ME:")
[1] "Percentage of ME:"
> print(38)
[1] 38
> print("Percentage of Other:")
[1] "Percentage of Other:"
> print(14)
[1] 14
>
>
> ################## Place of Living ######################
> print("Place of Living: [On campus=1;Off campus=2]")
[1] "Place of Living: [On campus=1;Off campus=2]"
> print("Summary Statistics for Place of Living:")
[1] "Summary Statistics for Place of Living:"
> summary(sample_data$Living)
 1  2
43 57
> print("Barplot for Place of Living")
[1] "Barplot for Place of Living"
> barplot(table(sample_data$Living), main = "Place of living of students in STA 201 S22",xlab="Place
of living", ylab="Number of studnets")
>
> print("Percentage of Students on Campus")
[1] "Percentage of Students on Campus"
> print(43)
[1] 43
> print("Percentage of Students off Campus")
[1] "Percentage of Students off Campus"
> print(57)
[1] 57
>
>
>
>
> ################## Comparing GPAs ######################
> gpa_males   <- sample_data$GPA[sample_data$Gender == 1]
```

```
> gpa_females <- sample_data$GPA[sample_data$Gender == 2]
>
> print("Boxplot of Males and Female GPA")
[1] "Boxplot of Males and Female GPA"
> boxplot(gpa_males, gpa_females, names = c("Males", "Females"), main = "Boxplot of Gender and GPA",
xlab = "Gender", ylab = "GPA")
>
> print("Summary of Male GPA:")
[1] "Summary of Male GPA:"
> summary(gpa_males)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.690   2.500   3.000   2.835   3.200   3.830
> print("Summary of Female GPA: ")
[1] "Summary of Female GPA: "
> summary(gpa_females)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.700   2.820   3.000   2.988   3.300   3.870
> print("Test for homogeneity of Variance")
[1] "Test for homogeneity of Variance"
> var.test(gpa_females, gpa_males)

        F test to compare two variances

data:  gpa_females and gpa_males
F = 0.95621, num df = 58, denom df = 40, p-value = 0.864
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5288953 1.6762724
sample estimates:
ratio of variances
         0.9562122

> print("Two Samples t-test between Male and Female GPA")
[1] "Two Samples t-test between Male and Female GPA"
> t.test(gpa_females, gpa_males, conf.level=0.95, alternative = "greater", var.equal=F)

        Welch Two Sample t-test

data:  gpa_females and gpa_males
t = 1.4794, df = 84.952, p-value = 0.07137
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.01898072         Inf
sample estimates:
mean of x mean of y
 2.988305  2.835366
```

Box Plot:



The boxplot for female GPA is wider than the boxplot for male GPA (Male GPA: 3.83-1.69=2.14; Female GPA: 3.87 – 1.70=2.17). This shows there is more variability in the GPA for females compared to males. It can also be seen from the boxplot (and as confirmed by summary statistics) that the Median for both is the same (3.0). The first quartile of Male GPA (2.5) is lower than Female GPA (2.8) and the third quartile of Male GPA (3.2) is less than Female GPA (3.3). The summary statistics show that the maximum Male GPA (3.83) is slightly lower than the maximum Female GPA (3.87) whereas the minimum Male GPA is 1.69, which is slightly lower than the minimum Female GPA at (1.7).

There are at least 4 outliers in the Female GPA dataset, while there are none for the male dataset. There are two different large (n =100 > 30 → this is for both; therefore assume normality for both) sample populations with unknown population standard deviations so a 2-sample t-distribution must be used. The Null Hypothesis is σFemaleGPA <= σMaleGPA and the Alternative Hypothesis is σFemaleGPA > σMaleGPA. While testing for equality of variance, p-value = 0.864 > 5% thus we fail to reject that the variances are equal, therefore we assume they're equal. The p-value (0.0714) is greater than alpha = 5% therefore we fail to reject the Null

Hypothesis. We are 95% confident that we have enough evidence that the difference in average GPA between Females and Males lies between -0.01898072 and Inf. 0 can be found in the interval thus there is no significant difference between Female GPA and Male GPA. Therefore, we conclude that there is no major statistical difference, thus not enough evidence to show a difference between the average male and female GPA with 95% confidence.

Q6: (d)(iii)

Code & Output:

```
> library(readxl)
>
> data <- read_excel(file.choose())
>
> data$Gender <- as.factor(data$Gender)
> data$College <- as.factor(data$College)
> data$Academic <- as.factor(data$Academic)
> data$Origin <- as.factor(data$Origin)
> data$Living <- as.factor(data$Living)
>
> set.seed(88673)
> sample_data<-data.frame(data[sample(1:dim(data)[1], size=100),])
>
> ################## Gender ######################
> print("Gender: [M=1;F=2]:")
[1] "Gender: [M=1;F=2]:"
> print("Barplot of Gender")
[1] "Barplot of Gender"
> summary(sample_data$Gender)
 1  2
41 59
> barplot(table(sample_data$Gender), main="Gender description of students in STA201 S22:
",xlab="Number Of Students:",ylab="Gender of Student:")
>
> print("Percentage of Females over Males")
[1] "Percentage of Females over Males"
> print(41)
[1] 41
> print("Percentage of Males over Females")
[1] "Percentage of Males over Females"
> print(100 - 41)
[1] 59
>
>
> ################## Academic Status ######################
>
> print("Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: ")
[1] "Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: "
> print("Summary statistics for Academic Status: ")
[1] "Summary statistics for Academic Status: "
> summary(sample_data$Academic)
 1  2  3  4
35 19 28 18
> print("Barchart for Academic status: ")
[1] "Barchart for Academic status: "
> barplot(table(sample_data$Academic), main = "Academic Status of Students in STA 201
S22",xlab="Academic Status", ylab = "Number of Students")
>
> print("Percentage of Freshman:")
[1] "Percentage of Freshman:"
> print(35)
[1] 35
> print("Percentage of Sophomores:")
[1] "Percentage of Sophomores:"
> print(19)
[1] 19
> print("Percentage of Juniors:")
[1] "Percentage of Juniors:"
> print(28)
[1] 28
> print("Percentage of Seniors:")
```

```
[1] "Percentage of Seniors:"
> print(18)
[1] 18
>
>
> ################## Region of Origin ######################
> print("Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:")
[1] "Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:"
> print("Summary statistics for Region of Living:")
[1] "Summary statistics for Region of Living:"
> summary(sample_data$Origin)
 1  2  3  4
27 21 38 14
>
>
> print("Barplot for Origin:")
[1] "Barplot for Origin:"
> barplot(table(sample_data$Origin),main="Region of Living of students in STA 201 S22",xlab="Region
of Living", ylab="Number of students")
>
> print("Percentage of GCC:")
[1] "Percentage of GCC:"
> print(27)
[1] 27
> print("Percentage of Africa:")
[1] "Percentage of Africa:"
> print(21)
[1] 21
> print("Percentage of ME:")
[1] "Percentage of ME:"
> print(38)
[1] 38
> print("Percentage of Other:")
[1] "Percentage of Other:"
> print(14)
[1] 14
>
>
> ################## Place of Living ######################
> print("Place of Living: [On campus=1;Off campus=2]")
[1] "Place of Living: [On campus=1;Off campus=2]"
> print("Summary Statistics for Place of Living:")
[1] "Summary Statistics for Place of Living:"
> summary(sample_data$Living)
 1  2
43 57
> print("Barplot for Place of Living")
[1] "Barplot for Place of Living"
> barplot(table(sample_data$Living), main = "Place of living of students in STA 201 S22",xlab="Place
of living", ylab="Number of studnets")
>
> print("Percentage of Students on Campus")
[1] "Percentage of Students on Campus"
> print(43)
[1] 43
> print("Percentage of Students off Campus")
[1] "Percentage of Students off Campus"
> print(57)
[1] 57
>
>
>
>
> ################## Comparing GPAs ######################
> gpa_males   <- sample_data$GPA[sample_data$Gender == 1]
```

```
> gpa_females <- sample_data$GPA[sample_data$Gender == 2]
>
> print("Boxplot of Males and Female GPA")
[1] "Boxplot of Males and Female GPA"
> boxplot(gpa_males, gpa_females, names = c("Males", "Females"), main = "Boxplot of Gender and GPA",
xlab = "Gender", ylab = "GPA")
>
> print("Summary of Male GPA:")
[1] "Summary of Male GPA:"
> summary(gpa_males)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.690   2.500   3.000   2.835   3.200   3.830
> print("Summary of Female GPA: ")
[1] "Summary of Female GPA: "
> summary(gpa_females)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.700   2.820   3.000   2.988   3.300   3.870
> print("Test for homogeneity of Variance")
[1] "Test for homogeneity of Variance"
> var.test(gpa_females, gpa_males)

        F test to compare two variances

data:  gpa_females and gpa_males
F = 0.95621, num df = 58, denom df = 40, p-value = 0.864
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5288953 1.6762724
sample estimates:
ratio of variances
         0.9562122

> print("Two Samples t-test between Male and Female GPA")
[1] "Two Samples t-test between Male and Female GPA"
> t.test(gpa_females, gpa_males, conf.level=0.95, alternative = "greater", var.equal=F)

        Welch Two Sample t-test

data:  gpa_females and gpa_males
t = 1.4794, df = 84.952, p-value = 0.07137
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.01898072         Inf
sample estimates:
mean of x mean of y
 2.988305  2.835366

>
>
>
>
> ################### Average number of Hours studied for on vs off campus ########################
>
> studytime_on      <- sample_data$Sudying_Time[sample_data$Living == 1]
> studytime_off     <- sample_data$Sudying_Time[sample_data$Living == 2]
>
> print("Test for Homogeneity of variance:")
[1] "Test for Homogeneity of variance:"
> var.test(studytime_on, studytime_off)

        F test to compare two variances

data:  studytime_on and studytime_off
F = 2.3036, num df = 42, denom df = 56, p-value = 0.003666
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
 1.314720 4.132825
sample estimates:
ratio of variances
          2.303581


>
> print("Two sample t-test between students on vs off campus on Studying Time:")
[1] "Two sample t-test between students on vs off campus on Studying Time:"
> t.test(studytime_on, studytime_off, conf.level=0.95,var.equal=F)

        Welch Two Sample t-test

data:  studytime_on and studytime_off
t = 1.758, df = 68.503, p-value = 0.08321
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5956549  9.4263363
sample estimates:
mean of x mean of y
 18.47674  14.06140
```

There are two different large (n=100>30 → for both, thus assuming normality) sample populations with some unknown standard deviation. Therefore, a 2-sample t-distribution must be used. The null hypothesis is study time on campus = study time off campus, and the alternative would be study time on campus != study time off campus. The test for equality of variance shows a p-value = 0.00366 < 5% thus we can conclude that the two variances are different, however homogeneity is assumed to run the t-test. The t-test p-value = 0.08321 > 5% thus failing to reject the null hypothesis, implying that we assume the study time on campus is equal to the study time off campus, showing no major statistical difference on average. We are 95% confident that the difference in average studying time between people on campus vs people off campus lies between -0.597 and 9.426. 0 Can be found in the interval which prevents a conclusion of difference, thus no group appears to study more. Therefore, we conclude that there is no major statistical difference, thus not enough evidence to show a difference between the study time on campus vs study time off campus with 95% confidence.

Q7: (d)(iv)

Code & Output:

```
> library(readxl)
>
> data <- read_excel(file.choose())
>
> data$Gender <- as.factor(data$Gender)
> data$College <- as.factor(data$College)
> data$Academic <- as.factor(data$Academic)
> data$Origin <- as.factor(data$Origin)
> data$Living <- as.factor(data$Living)
>
> set.seed(88673)
> sample_data<-data.frame(data[sample(1:dim(data)[1], size=100),])
>
> ################## Gender ######################
> print("Gender: [M=1;F=2]:")
[1] "Gender: [M=1;F=2]:"
> print("Barplot of Gender")
[1] "Barplot of Gender"
> summary(sample_data$Gender)
 1  2
41 59
> barplot(table(sample_data$Gender), main="Gender description of students in STA201 S22:
",xlab="Number Of Students:",ylab="Gender of Student:")
>
> print("Percentage of Females over Males")
[1] "Percentage of Females over Males"
> print(41)
[1] 41
> print("Percentage of Males over Females")
[1] "Percentage of Males over Females"
> print(100 - 41)
[1] 59
>
>
> ################## Academic Status ######################
>
> print("Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: ")
[1] "Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: "
> print("Summary statistics for Academic Status: ")
[1] "Summary statistics for Academic Status: "
> summary(sample_data$Academic)
 1  2  3  4
35 19 28 18
> print("Barchart for Academic status: ")
[1] "Barchart for Academic status: "
> barplot(table(sample_data$Academic), main = "Academic Status of Students in STA 201
S22",xlab="Academic Status", ylab = "Number of Students")
>
> print("Percentage of Freshman:")
[1] "Percentage of Freshman:"
> print(35)
[1] 35
> print("Percentage of Sophomores:")
[1] "Percentage of Sophomores:"
> print(19)
[1] 19
> print("Percentage of Juniors:")
[1] "Percentage of Juniors:"
> print(28)
[1] 28
> print("Percentage of Seniors:")
```

```
[1] "Percentage of Seniors:"
> print(18)
[1] 18
>
>
> ################## Region of Origin ######################
> print("Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:")
[1] "Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:"
> print("Summary statistics for Region of Living:")
[1] "Summary statistics for Region of Living:"
> summary(sample_data$Origin)
 1  2  3  4
27 21 38 14
>
>
> print("Barplot for Origin:")
[1] "Barplot for Origin:"
> barplot(table(sample_data$Origin),main="Region of Living of students in STA 201 S22",xlab="Region
of Living", ylab="Number of students")
>
> print("Percentage of GCC:")
[1] "Percentage of GCC:"
> print(27)
[1] 27
> print("Percentage of Africa:")
[1] "Percentage of Africa:"
> print(21)
[1] 21
> print("Percentage of ME:")
[1] "Percentage of ME:"
> print(38)
[1] 38
> print("Percentage of Other:")
[1] "Percentage of Other:"
> print(14)
[1] 14
>
>
> ################## Place of Living ######################
> print("Place of Living: [On campus=1;Off campus=2]")
[1] "Place of Living: [On campus=1;Off campus=2]"
> print("Summary Statistics for Place of Living:")
[1] "Summary Statistics for Place of Living:"
> summary(sample_data$Living)
 1  2
43 57
> print("Barplot for Place of Living")
[1] "Barplot for Place of Living"
> barplot(table(sample_data$Living), main = "Place of living of students in STA 201 S22",xlab="Place
of living", ylab="Number of studnets")
>
> print("Percentage of Students on Campus")
[1] "Percentage of Students on Campus"
> print(43)
[1] 43
> print("Percentage of Students off Campus")
[1] "Percentage of Students off Campus"
> print(57)
[1] 57
>
>
>
>
> ################## Comparing GPAs ######################
> gpa_males   <- sample_data$GPA[sample_data$Gender == 1]
```

```
> gpa_females <- sample_data$GPA[sample_data$Gender == 2]
>
> print("Boxplot of Males and Female GPA")
[1] "Boxplot of Males and Female GPA"
> boxplot(gpa_males, gpa_females, names = c("Males", "Females"), main = "Boxplot of Gender and GPA",
xlab = "Gender", ylab = "GPA")
>
> print("Summary of Male GPA:")
[1] "Summary of Male GPA:"
> summary(gpa_males)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.690   2.500   3.000   2.835   3.200   3.830
> print("Summary of Female GPA: ")
[1] "Summary of Female GPA: "
> summary(gpa_females)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.700   2.820   3.000   2.988   3.300   3.870
> print("Test for homogeneity of Variance")
[1] "Test for homogeneity of Variance"
> var.test(gpa_females, gpa_males)

        F test to compare two variances

data:  gpa_females and gpa_males
F = 0.95621, num df = 58, denom df = 40, p-value = 0.864
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5288953 1.6762724
sample estimates:
ratio of variances
         0.9562122

> print("Two Samples t-test between Male and Female GPA")
[1] "Two Samples t-test between Male and Female GPA"
> t.test(gpa_females, gpa_males, conf.level=0.95, alternative = "greater", var.equal=F)

        Welch Two Sample t-test

data:  gpa_females and gpa_males
t = 1.4794, df = 84.952, p-value = 0.07137
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.01898072         Inf
sample estimates:
mean of x mean of y
 2.988305  2.835366

>
>
>
>
> ################### Average number of Hours studied for on vs off campus ########################
>
> studytime_on      <- sample_data$Sudying_Time[sample_data$Living == 1]
> studytime_off     <- sample_data$Sudying_Time[sample_data$Living == 2]
>
> print("Test for Homogeneity of variance:")
[1] "Test for Homogeneity of variance:"
> var.test(studytime_on, studytime_off)

        F test to compare two variances

data:  studytime_on and studytime_off
F = 2.3036, num df = 42, denom df = 56, p-value = 0.003666
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
 1.314720 4.132825
sample estimates:
ratio of variances
          2.303581

>
> print("Two sample t-test between students on vs off campus on Studying Time:")
[1] "Two sample t-test between students on vs off campus on Studying Time:"
> t.test(studytime_on, studytime_off, conf.level=0.95,var.equal=F)

        Welch Two Sample t-test

data:  studytime_on and studytime_off
t = 1.758, df = 68.503, p-value = 0.08321
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5956549  9.4263363
sample estimates:
mean of x mean of y
 18.47674  14.06140

>
>
> ################### 0.75 of students from non GCC origin ########################
>
> origins_of_gcc <- sample_data$Origin[sample_data$Origin == 1]
> origins_of_nongcc <- 100 - length(origins_of_gcc)
> print("Number of people with GCC origin:")
[1] "Number of people with GCC origin:"
> length(origins_of_gcc)
[1] 27
> print("Number of people without GCC origin:")
[1] "Number of people without GCC origin:"
> length(origins_of_nongcc)
[1] 1
> print("One sample t-test for Region of Origin within GCC")
[1] "One sample t-test for Region of Origin within GCC"
> prop.test(origins_of_nongcc, n=100, p=0.75, alternative = "greater", conf.level = 0.95, correct =
"FALSE")

        1-sample proportions test without continuity correction

data:  origins_of_nongcc out of 100, null probability 0.75
X-squared = 0.21333, df = 1, p-value = 0.6779
alternative hypothesis: true p is greater than 0.75
95 percent confidence interval:
 0.6516303 1.0000000
sample estimates:
   p
0.73
```

There is only one sample population which has a large sample size (np & n(1-p) >5) therefore, normality is assumed. The null hypothesis is that the percentage of students with GCC origin is less than or equal to 75% while the alternative hypothesis is that the percentage of students with GCC origin is greater than 75% (H0 = GCC <= 75%; Ha = GCC > 75%). The test shows a p-value = 0.6779 > 5% which is greater than 5% which implies that we fail to reject the null hypothesis,

meaning that we assume the null hypothesis. In conclusion, we are 95% confident that the proportion of people with GCC origin is less than or equal to 75%, which means that there is not enough evidence to show that more than 75% of the student body is of GCC origin with a 95% confidence level.

Q8: (d)(IV)

Code & Output:

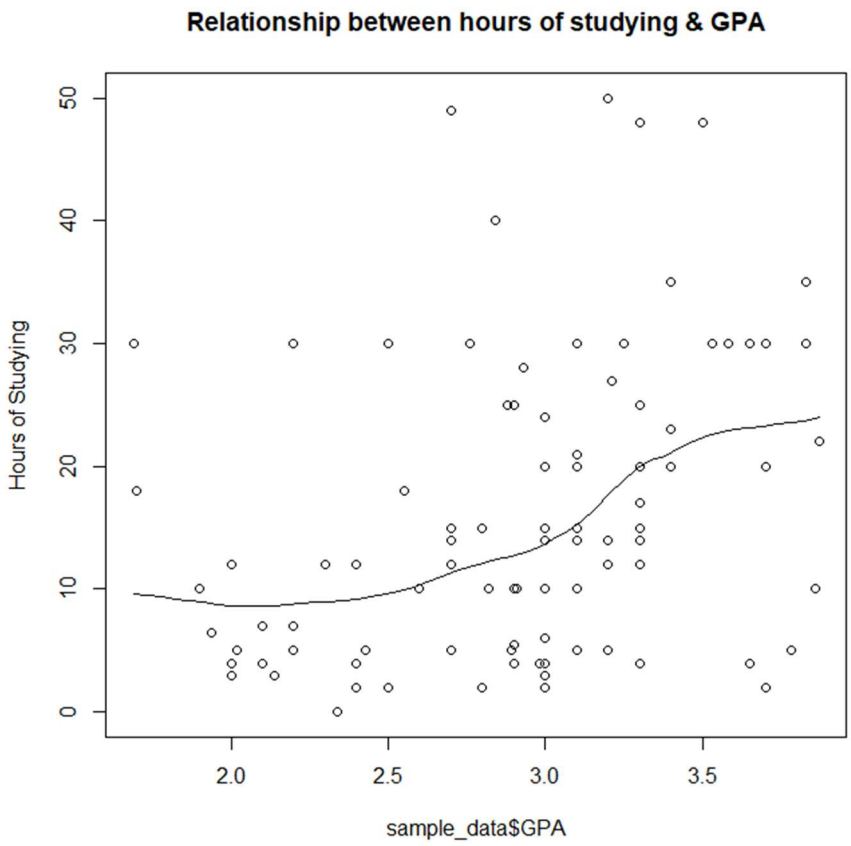
```
> library(readxl)
>
> data <- read_excel(file.choose())
>
> data$Gender <- as.factor(data$Gender)
> data$College <- as.factor(data$College)
> data$Academic <- as.factor(data$Academic)
> data$Origin <- as.factor(data$Origin)
> data$Living <- as.factor(data$Living)
>
> set.seed(88673)
> sample_data<-data.frame(data[sample(1:dim(data)[1], size=100),])
>
> ################## Gender ######################
> print("Gender: [M=1;F=2]:")
[1] "Gender: [M=1;F=2]:"
> print("Barplot of Gender")
[1] "Barplot of Gender"
> summary(sample_data$Gender)
 1  2
41 59
> barplot(table(sample_data$Gender), main="Gender description of students in STA201 S22:
",xlab="Number Of Students:",ylab="Gender of Student:")
>
> print("Percentage of Females over Males")
[1] "Percentage of Females over Males"
> print(41)
[1] 41
> print("Percentage of Males over Females")
[1] "Percentage of Males over Females"
> print(100 - 41)
[1] 59
>
>
> ################## Academic Status ######################
>
> print("Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: ")
[1] "Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: "
> print("Summary statistics for Academic Status: ")
[1] "Summary statistics for Academic Status: "
> summary(sample_data$Academic)
 1  2  3  4
35 19 28 18
> print("Barchart for Academic status: ")
[1] "Barchart for Academic status: "
> barplot(table(sample_data$Academic), main = "Academic Status of Students in STA 201
S22",xlab="Academic Status", ylab = "Number of Students")
>
> print("Percentage of Freshman:")
[1] "Percentage of Freshman:"
> print(35)
[1] 35
> print("Percentage of Sophomores:")
[1] "Percentage of Sophomores:"
> print(19)
[1] 19
> print("Percentage of Juniors:")
[1] "Percentage of Juniors:"
> print(28)
[1] 28
> print("Percentage of Seniors:")
```

```
[1] "Percentage of Seniors:"
> print(18)
[1] 18
>
>
> ################## Region of Origin ######################
> print("Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:")
[1] "Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:"
> print("Summary statistics for Region of Living:")
[1] "Summary statistics for Region of Living:"
> summary(sample_data$Origin)
 1  2  3  4
27 21 38 14
>
>
> print("Barplot for Origin:")
[1] "Barplot for Origin:"
> barplot(table(sample_data$Origin),main="Region of Living of students in STA 201 S22",xlab="Region
of Living", ylab="Number of students")
>
> print("Percentage of GCC:")
[1] "Percentage of GCC:"
> print(27)
[1] 27
> print("Percentage of Africa:")
[1] "Percentage of Africa:"
> print(21)
[1] 21
> print("Percentage of ME:")
[1] "Percentage of ME:"
> print(38)
[1] 38
> print("Percentage of Other:")
[1] "Percentage of Other:"
> print(14)
[1] 14
>
>
> ################## Place of Living ######################
> print("Place of Living: [On campus=1;Off campus=2]")
[1] "Place of Living: [On campus=1;Off campus=2]"
> print("Summary Statistics for Place of Living:")
[1] "Summary Statistics for Place of Living:"
> summary(sample_data$Living)
 1  2
43 57
> print("Barplot for Place of Living")
[1] "Barplot for Place of Living"
> barplot(table(sample_data$Living), main = "Place of living of students in STA 201 S22",xlab="Place
of living", ylab="Number of studnets")
>
> print("Percentage of Students on Campus")
[1] "Percentage of Students on Campus"
> print(43)
[1] 43
> print("Percentage of Students off Campus")
[1] "Percentage of Students off Campus"
> print(57)
[1] 57
>
>
>
>
> ################## Comparing GPAs ######################
> gpa_males   <- sample_data$GPA[sample_data$Gender == 1]
```

```
> gpa_females <- sample_data$GPA[sample_data$Gender == 2]
>
> print("Boxplot of Males and Female GPA")
[1] "Boxplot of Males and Female GPA"
> boxplot(gpa_males, gpa_females, names = c("Males", "Females"), main = "Boxplot of Gender and GPA",
xlab = "Gender", ylab = "GPA")
>
> print("Summary of Male GPA:")
[1] "Summary of Male GPA:"
> summary(gpa_males)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.690   2.500   3.000   2.835   3.200   3.830
> print("Summary of Female GPA: ")
[1] "Summary of Female GPA: "
> summary(gpa_females)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.700   2.820   3.000   2.988   3.300   3.870
> print("Test for homogeneity of Variance")
[1] "Test for homogeneity of Variance"
> var.test(gpa_females, gpa_males)

        F test to compare two variances

data:  gpa_females and gpa_males
F = 0.95621, num df = 58, denom df = 40, p-value = 0.864
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5288953 1.6762724
sample estimates:
ratio of variances
         0.9562122

> print("Two Samples t-test between Male and Female GPA")
[1] "Two Samples t-test between Male and Female GPA"
> t.test(gpa_females, gpa_males, conf.level=0.95, alternative = "greater", var.equal=F)

        Welch Two Sample t-test

data:  gpa_females and gpa_males
t = 1.4794, df = 84.952, p-value = 0.07137
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.01898072         Inf
sample estimates:
mean of x mean of y
 2.988305  2.835366

>
>
>
>
> ################### Average number of Hours studied for on vs off campus ########################
>
> studytime_on      <- sample_data$Sudying_Time[sample_data$Living == 1]
> studytime_off     <- sample_data$Sudying_Time[sample_data$Living == 2]
>
> print("Test for Homogeneity of variance:")
[1] "Test for Homogeneity of variance:"
> var.test(studytime_on, studytime_off)

        F test to compare two variances

data:  studytime_on and studytime_off
F = 2.3036, num df = 42, denom df = 56, p-value = 0.003666
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
 1.314720 4.132825
sample estimates:
ratio of variances
          2.303581

>
> print("Two sample t-test between students on vs off campus on Studying Time:")
[1] "Two sample t-test between students on vs off campus on Studying Time:"
> t.test(studytime_on, studytime_off, conf.level=0.95,var.equal=F)

        Welch Two Sample t-test

data:  studytime_on and studytime_off
t = 1.758, df = 68.503, p-value = 0.08321
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5956549  9.4263363
sample estimates:
mean of x mean of y
 18.47674  14.06140

>
>
> ################## 0.75 of students from non GCC origin ######################
>
> origins_of_gcc <- sample_data$Origin[sample_data$Origin == 1]
> origins_of_nongcc <- 100 - length(origins_of_gcc)
> print("Number of people with GCC origin:")
[1] "Number of people with GCC origin:"
> length(origins_of_gcc)
[1] 27
> print("Number of people without GCC origin:")
[1] "Number of people without GCC origin:"
> length(origins_of_nongcc)
[1] 1
> print("One sample t-test for Region of Origin within GCC")
[1] "One sample t-test for Region of Origin within GCC"
> prop.test(origins_of_nongcc, n=100, p=0.75, alternative = "greater", conf.level = 0.95, correct =
"FALSE")

        1-sample proportions test without continuity correction

data:  origins_of_nongcc out of 100, null probability 0.75
X-squared = 0.21333, df = 1, p-value = 0.6779
alternative hypothesis: true p is greater than 0.75
95 percent confidence interval:
 0.6516303 1.0000000
sample estimates:
   p
0.73

>
>
>
> ################## Comparing Studying Time & GPA ######################
>
> summary(lm(sample_data$Sudying_Time~sample_data$GPA))

Call:
lm(formula = sample_data$Sudying_Time ~ sample_data$GPA)

Residuals:
    Min      1Q  Median      3Q     Max
-19.754  -7.125  -2.268   7.326  34.728
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.927      6.639  -0.893  0.37411
sample_data$GPA    7.481      2.236   3.346  0.00116 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.33 on 98 degrees of freedom
Multiple R-squared:  0.1025,    Adjusted R-squared:  0.09338
F-statistic:  11.2 on 1 and 98 DF,  p-value: 0.001163

>
> cor.test(sample_data$Sudying_Time, sample_data$GPA,
alternative="greater",method="pearson",conf.level=0.95)

        Pearson's product-moment correlation

data:  sample_data$Sudying_Time and sample_data$GPA
t = 3.3462, df = 98, p-value = 0.0005814
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.1633994 1.0000000
sample estimates:
      cor
0.3202152

>
> scatter.smooth(x = sample_data$GPA, y = sample_data$Sudying_Time, ylab="Hours of Studying", main =
"Relationship between hours of studying & GPA")
>
>
```



To compare the hours of study vs the GPA correlation, a hypothesis test with (H0 = R <= 0; Ha = R > 0) where R refers to the linear correlation between the hours of study vs the GPA correlation. The null hypothesis adopts the statement that the correlation value R is less than or equal to 0 which means that there is either no correlation or a negative correlation of the two variables. While the alternative hypothesis adopts the statement that there is purely a positive correlation between the two variables. The correlation test shown above has a p-value = 0.0005814 < 5% which implies a rejection of the null hypothesis, thus adopting the alternative hypothesis. Therefore, the test proves that there is a positive correlation between the hours of study vs the GPA. Thus, there is enough evidence to conclude that there is a positive correlation between hours of study and GPA with 95% confidence. In addition, there is no need to evaluate linearity as the question at hand only requests an evaluation of correlation, not linearity.

Appendix:

```
library(readxl)

data <- read_excel(file.choose())

data$Gender <- as.factor(data$Gender)
data$College <- as.factor(data$College)
data$Academic <- as.factor(data$Academic)
data$Origin <- as.factor(data$Origin)
data$Living <- as.factor(data$Living)

set.seed(88673)
sample_data<-data.frame(data[sample(1:dim(data)[1], size=100),])

################## Gender ######################
print("Gender: [M=1;F=2]:")
print("Barplot of Gender")
summary(sample_data$Gender)
barplot(table(sample_data$Gender), main="Gender description of students in STA201 S22:
",xlab="Number Of Students:",ylab="Gender of Student:")

print("Percentage of Females over Males")
print(41)
print("Percentage of Males over Females")
print(100 - 41)


################## Academic Status ######################

print("Academic Status: [Freshman=1;Sophomore=2;Junior=3;Senior=4]: ")
print("Summary statistics for Academic Status: ")
summary(sample_data$Academic)
print("Barchart for Academic status: ")
barplot(table(sample_data$Academic), main = "Academic Status of Students in STA 201
S22",xlab="Academic Status", ylab = "Number of Students")

print("Percentage of Freshman:")
print(35)
print("Percentage of Sophomores:")
print(19)
print("Percentage of Juniors:")
print(28)
print("Percentage of Seniors:")
print(18)


################## Region of Origin ######################
print("Region of Living: [GCC=1;Africa=2;ME=3;Other=4]:")
print("Summary statistics for Region of Living:")
summary(sample_data$Origin)


print("Barplot for Origin:")
barplot(table(sample_data$Origin),main="Region of Living of students in STA 201 S22",xlab="Region of
Living", ylab="Number of students")

print("Percentage of GCC:")
print(27)
print("Percentage of Africa:")
print(21)
print("Percentage of ME:")
print(38)
print("Percentage of Other:")
print(14)
```

```r
################## Place of Living ######################
print("Place of Living: [On campus=1;Off campus=2]")
print("Summary Statistics for Place of Living:")
summary(sample_data$Living)
print("Barplot for Place of Living")
barplot(table(sample_data$Living), main = "Place of living of students in STA 201 S22",xlab="Place
of living", ylab="Number of studnets")

print("Percentage of Students on Campus")
print(43)
print("Percentage of Students off Campus")
print(57)




################## Comparing GPAs ######################
gpa_males   <- sample_data$GPA[sample_data$Gender == 1]
gpa_females <- sample_data$GPA[sample_data$Gender == 2]

print("Boxplot of Males and Female GPA")
boxplot(gpa_males, gpa_females, names = c("Males", "Females"), main = "Boxplot of Gender and GPA",
xlab = "Gender", ylab = "GPA")

print("Summary of Male GPA:")
summary(gpa_males)
print("Summary of Female GPA: ")
summary(gpa_females)
print("Test for homogeneity of Variance")
var.test(gpa_females, gpa_males)
print("Two Samples t-test between Male and Female GPA")
t.test(gpa_females, gpa_males, conf.level=0.95, alternative = "greater", var.equal=F)




################## Average number of Hours studied for on vs off campus ######################

studytime_on      <- sample_data$Sudying_Time[sample_data$Living == 1]
studytime_off     <- sample_data$Sudying_Time[sample_data$Living == 2]

print("Test for Homogeneity of variance:")
var.test(studytime_on, studytime_off)

print("Two sample t-test between students on vs off campus on Studying Time:")
t.test(studytime_on, studytime_off, conf.level=0.95,var.equal=F)


################## 0.75 of students from non GCC origin ######################

origins_of_gcc <- sample_data$Origin[sample_data$Origin == 1]
origins_of_nongcc <- 100 - length(origins_of_gcc)
print("Number of people with GCC origin:")
length(origins_of_gcc)
print("Number of people without GCC origin:")
length(origins_of_nongcc)
print("One sample t-test for Region of Origin within GCC")
prop.test(origins_of_nongcc, n=100, p=0.75, alternative = "greater", conf.level = 0.95, correct =
"FALSE")



################## Comparing Studying Time & GPA ######################
```

```
summary(lm(sample_data$Sudying_Time~sample_data$GPA))

cor.test(sample_data$Sudying_Time, sample_data$GPA,
alternative="greater",method="pearson",conf.level=0.95)

scatter.smooth(x = sample_data$GPA, y = sample_data$Sudying_Time, ylab="Hours of Studying", main =
"Relationship between hours of studying & GPA")
```