```
# install.packages('ggplot2')
# install.packages('car')
# install.packages('psych')
library(ggplot2)
library(car)
library(psych)
```

Dataset

```
student_df <- data.frame(
    Subject = 1:14,
    GPA = c(3.8, 4.0, 3.2, 3.5, 2.5, 3.0, 2.1, 2.8, 3.6, 4.0, 3.6, 3.4, 3.2, 2.0),
    Adaptability = c(45, 50, 45, 51, 60, 39, 42, 41, 46, 50, 53, 47, 48, 40),
    Self_Confidence = c(60, 10, 50, 25, 15, 80, 41, 14, 57, 68, 24, 95, 25, 36),
    IQ = c(105, 109, 102, 95, 92, 101, 99, 95, 94, 110, 104, 105, 98, 75),
    Gender = c("Female", "Female", "Female", "Female", "Male" , "Male", "Male", "Female", "Female", "Female",
    Economic_Condition = c("Good", "Good", "Poor", "Good", 'Poor', 'Middle', 'Poor', 'Poor', 'Middle', 'Good',
)

head(student_df)
```

A data.frame: 6 × 7

	Subject	GPA	Adaptability	Self_Confidence	IQ	Gender	Economic_Condition
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
1	1	3.8	45	60	105	Female	Good
2	2	4.0	50	10	109	Female	Good
3	3	3.2	45	50	102	Female	Poor
4	4	3.5	51	25	95	Female	Good
5	5	2.5	60	15	92	Male	Poor
6	6	3.0	39	80	101	Male	Middle

```
#Production df
production_df <- data.frame(
  Year = 2010:2021,
  Production_thousands_kg = c(60.04, 59.13, 62.52, 66.26, 63.88, 67.38, 85.95, 78.95, 82.13, 96.07, 86.39, 96
  Consumption_thousands_kg = c(57.63, 58.50, 61.19, 64.00, 67.17, 77.57, 81.64, 85.93, 90.45, 65.20, 95.02, 6
)

head(production_df)
```

A data.frame: 6 × 3

	Year	Production_thousands_kg	Consumption_thousands_kg
	<int>	<dbl>	<dbl>
1	2010	60.04	57.63
2	2011	59.13	58.50
3	2012	62.52	61.19
4	2013	66.26	64.00
5	2014	63.88	67.17
6	2015	67.38	77.57

```
# Person df
person_df <- data.frame(
  ID = 1:20,
  Age = c(48, 50, 46, 44, 60, 41, 49, 45, 55, 51, 65, 54, 44, 42, 53, 44, 44, 50, 42, 56),
```

```
  Height_in_inch = c(61, 62, 60, 57, 58, 55, 56, 62, 60, 67, 54, 72, 69, 68, 59, 72, 61, 65, 110, 70),
  IQ = c(110, 100, 102, 92, 95, 108, 92, 110, 93, 115, 115, 130, 122, 125, 130, 115, 116, 113, 110, 122),
  BMI = c(24.99, 20.16, 22.39, 20.04, 20.73, 29.72, 20.76, 24.19, 18.51, 22.44, 34.55, 23.92, 23.44, 25.14, 3
  Gender = c("Male", "Male", "Male", "Female", "Female", "Female", "Female", "Female", "Female", "Female",
             "Female", "Male", "Male", "Male", "Male", "Male", "Female", "Female", "Female", "Female"),
  Family_Type = c("Nuclear", "Nuclear", "Nuclear", "Nuclear", "Joint", "Joint", "Nuclear", "Nuclear", "Nuclea
                  "Joint", "Joint", "Nuclear", "Joint", "Nuclear", "Nuclear", "Nuclear", "Nuclear", "Joint",
  Smoking_Habit = c("Yes", "Yes", "Yes", "No", "No", "No", "No", "No", "No", "No",
                    "No", "No", "No", "Yes", "No", "Yes", "No", "No", "No", "No"),
  Disease_Suffering = c("Yes", "Yes", "No", "Yes", "No", "No", "No", "No", "Yes", "No",
                        "Yes", "No", "No", "No", "Yes", "Yes", "Yes", "No", "No", "No"),
  Psychological_Stress = c("Yes", "Yes", "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes", "No",
                           "Yes", "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "No")
)

head(person_df)
```

A data.frame: 6 × 10

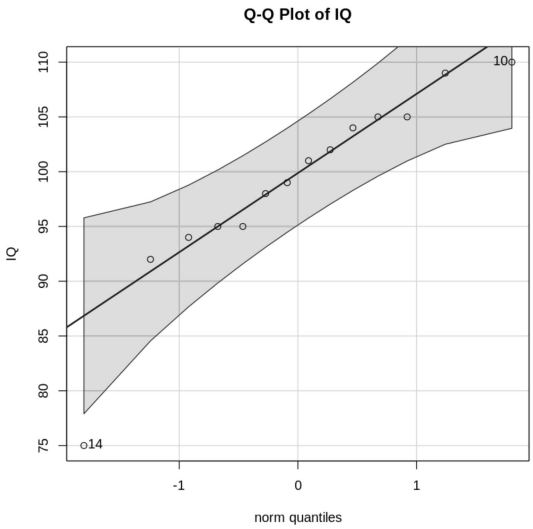
	ID	Age	Height_in_inch	IQ	BMI	Gender	Family_Type	Smoking_Habit	Disease_Suffering	Psychological_Stress
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>
1	1	48	61	110	24.99	Male	Nuclear	Yes	Yes	Yes
2	2	50	62	100	20.16	Male	Nuclear	Yes	Yes	Yes
3	3	46	60	102	22.39	Male	Nuclear	Yes	No	Yes
4	4	44	57	92	20.04	Female	Nuclear	No	Yes	Yes
5	5	60	58	95	20.73	Female	Joint	No	No	Yes
6	6	41	55	108	29.72	Female	Joint	No	No	No

1. Suggestion

1.1. Normality test for continuous data

```
#student_df$IQ
# install.packages('ggplot2')
# install.packages('car')
library(ggplot2)
library(car)

data <- student_df$IQ

## QQPlot
qqPlot(data, main="Q-Q Plot of IQ", ylab = "IQ")

# Shapiro - Wilk test
shapiro_result <- shapiro.test(data)
shapiro_value <- shapiro_result$statistic
p_value <- shapiro_result$p.value

cat("Shapiro Value is: ", shapiro_value)
cat("\nP-Value is:", p_value)
if (p_value <= 0.05){
    cat("\nReject Null hypothesis. The data set is not normally distributed")
} else {
    cat("\nCan't reject null hypothesis. The data is normally distributed")
}
```

```
14 · 10
Shapiro Value is:  0.8867281
P-Value is: 0.07253093
Can't reject null hypothesis. The data is normally distributed
```



1.2. Linear regression for GPA

```
model <- lm(GPA ~ Adaptability + Self_Confidence + IQ, data= student_df)
summary(model)
```

```
Call:
lm(formula = GPA ~ Adaptability + Self_Confidence + IQ, data = student_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98852 -0.16685  0.04994  0.25351  0.64842

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.851077   1.676988  -1.700   0.1199
Adaptability     0.022032   0.026740   0.824   0.4292
Self_Confidence  0.001559   0.005887   0.265   0.7965
IQ               0.050003   0.016754   2.984   0.0137 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4751 on 10 degrees of freedom
Multiple R-squared:  0.5843,    Adjusted R-squared:  0.4596
F-statistic: 4.685 on 3 and 10 DF,  p-value: 0.02715
```

1.3. Logistic Regression for Gender

```
student_df_encoded <- student_df
student_df_encoded$Gender <- ifelse(student_df_encoded$Gender == "Male", 1, 0)
student_df_encoded$Economic_Condition_Encoded <- ifelse(student_df_encoded$Economic_Condition == "Poor", 1, i

model <- glm(Self_Confidence ~ Gender + Economic_Condition, data= student_df)
summary(model)
```

```
Call:
glm(formula = Self_Confidence ~ Gender + Economic_Condition,
    data = student_df)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              37.4000    11.0113   3.397  0.00681 **
GenderMale              -11.4091    16.6001  -0.687  0.50753
Economic_ConditionMiddle  32.5545    18.4851   1.761  0.10871
Economic_ConditionPoor     0.6455    18.4851   0.035  0.97283
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 606.2382)

    Null deviance: 9007.7  on 13  degrees of freedom
Residual deviance: 6062.4  on 10  degrees of freedom
AIC: 134.72

Number of Fisher Scoring iterations: 2
```

1.4. Cronbach's Alpha

```
# install.packages('psych')
library(psych)

numerical_df <- student_df_encoded[, c('GPA', 'Economic_Condition_Encoded')]

alpha <- psych::alpha(numerical_df)
cat("Cronbach's Alpha Value is: ", alpha$total$raw_alpha)
```

```
Number of categories should be increased  in order to count frequencies.

Cronbach's Alpha Value is:  0.8990597
```

2022 Question

1.i) Identify Influential Factors of academic performance of students.

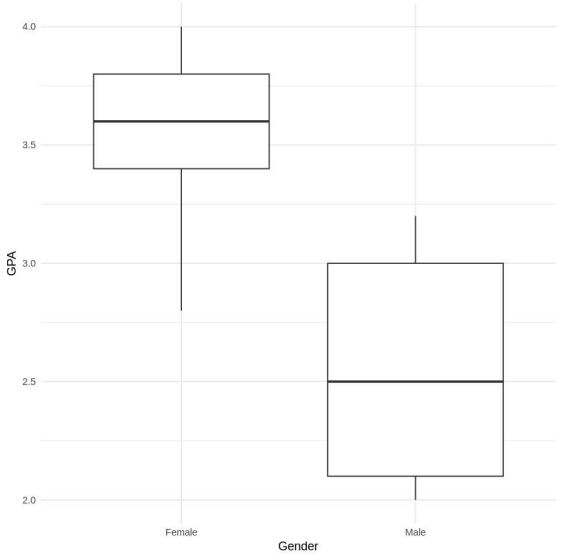
```
#Linear Regression
model <- lm(GPA ~ Adaptability + Self_Confidence + IQ , data= student_df)
summary(model)
```

```
Call:
lm(formula = GPA ~ Adaptability + Self_Confidence + IQ, data = student_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98852 -0.16685  0.04994  0.25351  0.64842

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.851077   1.676988  -1.700   0.1199
Adaptability     0.022032   0.026740   0.824   0.4292
Self_Confidence  0.001559   0.005887   0.265   0.7965
IQ               0.050003   0.016754   2.984   0.0137 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4751 on 10 degrees of freedom
Multiple R-squared:  0.5843,    Adjusted R-squared:  0.4596
F-statistic: 4.685 on 3 and 10 DF,  p-value: 0.02715
```

1.ii) Is there any variation of academic performance with respect to gender and economic condition.

```
#ttest for gender
#visualization
library('ggplot2')
ggplot(student_df, aes(x= Gender , y= GPA)) + geom_boxplot() + theme_minimal()

ttest_results <- t.test(GPA ~ Gender, data= student_df)
print(ttest_results)

cat("T-Test value is:", ttest_results$statistic)
cat("\nT-Test P-value is:", ttest_results$p.value)

if (ttest_results$p.value <= 0.05){
    cat("\nReject Null hypothesis. There is a difference")
} else {
    cat("\nFail to reject null hypothesis. There is no difference")
}
```

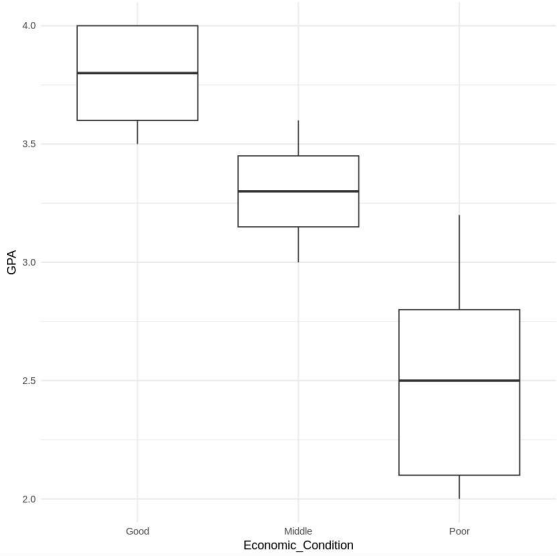
```
        Welch Two Sample t-test

data:  GPA by Gender
t = 3.6431, df = 6.3886, p-value = 0.009674
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 0.3328643 1.6360246
sample estimates:
mean in group Female   mean in group Male
            3.544444             2.560000

T-Test value is: 3.643099
T-Test P-value is: 0.009674185
Reject Null hypothesis. There is a difference
```




```
#Anova test for economic condition
ggplot(student_df, aes(x= Economic_Condition, y= GPA)) + geom_boxplot() +  theme_minimal()

anova_result <- aov(GPA ~ Economic_Condition, data = student_df)

print(summary(anova_result))
anova_value <- summary(anova_result)[[1]][['F value']][1]
p_value <- summary(anova_result)[[1]][['Pr(>F)']][1]

cat("ANOVA value is:", anova_value)
```

```
cat("\nANOVA P-Value is:", p_value)

if (p_value <= 0.05){
    cat("\nReject Null Hypothesis. There is a difference between them")
} else {
    cat("\nFail to reject null hypothesis. There is no significant difference between them")
}
```

```
                  Df Sum Sq Mean Sq F value  Pr(>F)
Economic_Condition  2  4.033  2.0166   15.89 0.00057 ***
Residuals          11  1.396  0.1269
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
ANOVA value is: 15.89045
ANOVA P-Value is: 0.000569882
Reject Null Hypothesis. There is a difference between them
```




```
# Perform Tukey HSD test
tukey_result <- TukeyHSD(anova_result)
print(tukey_result)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GPA ~ Economic_Condition, data = student_df)

$Economic_Condition
             diff       lwr        upr     p adj
Middle-Good -0.48 -1.125437  0.1654375 0.1561296
Poor-Good   -1.26 -1.868524 -0.6514757 0.0004363
Poor-Middle -0.78 -1.425437 -0.1345625 0.0190191
```

2.i) Using the best fitted model, forecast the production and consumption of tea in Bangladesh.

```
prod_model <- lm(Production_thousands_kg ~ Year, data= production_df)
cons_model <- lm(Consumption_thousands_kg ~ Year, data= production_df)

future_years <- data.frame(Year = c(2022, 2023, 2024))

prod_prediction <- predict(prod_model, newdata = future_years)
cons_prediction <- predict(cons_model, newdata = future_years)

cat("Production Prediction:", prod_prediction)
cat("\nConsumption Prediction:", cons_prediction)
```

```
Production Prediction: 98.71803 102.3002 105.8823
Consumption Prediction: 86.84939 89.06289 91.27639
```

2. ii) Is there any significant relationship between production and consumption

```
correlation_result <- cor.test(production_df$Production_thousands_kg, production_df$Consumption_thousands_kg)
correlation_value <- correlation_result$estimate
p_value <- correlation_result$p.value

cat("Correlation value is:", correlation_value)
if (p_value <= 0.05) {
  cat("\nReject Null Hypothesis. There is a significant difference between them")
} else {
  cat("\nFail to Reject Null Hypothesis. There is no difference")
}
```

```
Correlation value is: 0.4573198
Fail to Reject Null Hypothesis. There is no difference
```

2021 Question

a) State the background characteristics of the respondents by displaying a percent frequency distribution table.

```
cat_columns <- c("Gender", "Family_Type", 'Smoking_Habit',  'Disease_Suffering', 'Psychological_Stress')
for (col in cat_columns){
    cat("\nFrequency distribution for:", col )
    freq_dist <- prop.table(table(person_df[[col]])) * 100
    print(freq_dist)
}
```

```
Frequency distribution for: Gender
Female   Male
    60     40

Frequency distribution for: Family_Type
  Joint Nuclear
     30      70

Frequency distribution for: Smoking_Habit
 No Yes
 75  25

Frequency distribution for: Disease_Suffering
 No Yes
 60  40

Frequency distribution for: Psychological_Stress
 No Yes
 45  55
```

b) Explore the influential factors affecting on IQ of the respondents by applying multiple linear regression model.

```
model <- lm(IQ ~ Age + Height_in_inch   + BMI, data= person_df)
summary(model)
```

```
Call:
lm(formula = IQ ~ Age + Height_in_inch + BMI, data = person_df)

Residuals:
     Min       1Q   Median       3Q      Max
-17.2634  -7.6652  -0.2027   7.0319  17.5899

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    54.89697   28.13626   1.951   0.0688 .
Age            -0.01069    0.38470  -0.028   0.9782
Height_in_inch  0.31275    0.20579   1.520   0.1481
BMI             1.48712    0.51440   2.891   0.0106 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.22 on 16 degrees of freedom
Multiple R-squared:  0.3933,    Adjusted R-squared:  0.2795
F-statistic: 3.457 on 3 and 16 DF,  p-value: 0.04154
```

c) Determine the variation of IQ with respect to background characteristic.

```
head(person_df)
```

A data.frame: 6 × 10

	ID	Age	Height_in_inch	IQ	BMI	Gender	Family_Type	Smoking_Habit	Disease_Suffering	Psychological_Stress
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>
1	1	48	61	110	24.99	Male	Nuclear	Yes	Yes	Yes
2	2	50	62	100	20.16	Male	Nuclear	Yes	Yes	Yes
3	3	46	60	102	22.39	Male	Nuclear	Yes	No	Yes
4	4	44	57	92	20.04	Female	Nuclear	No	Yes	Yes
5	5	60	58	95	20.73	Female	Joint	No	No	Yes
6	6	41	55	108	29.72	Female	Joint	No	No	No

```
#Descriptive statistics by background characteristics
aggregate(IQ ~ Gender, data= person_df, mean)

# Mean IQ by Family Type
aggregate(IQ ~ Family_Type, data = person_df, mean)
```

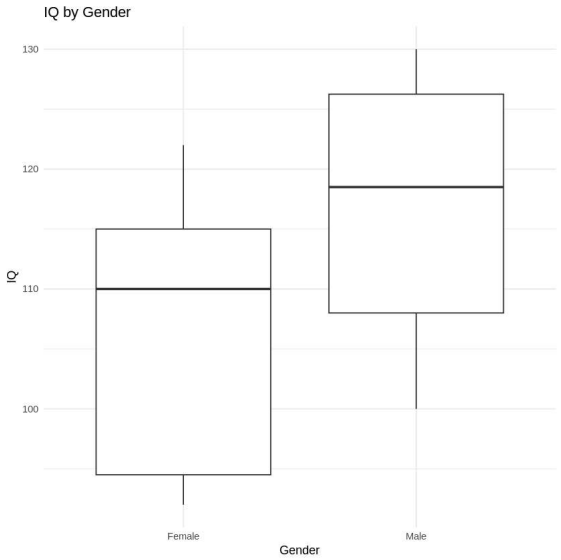
A data.frame: 2 × 2

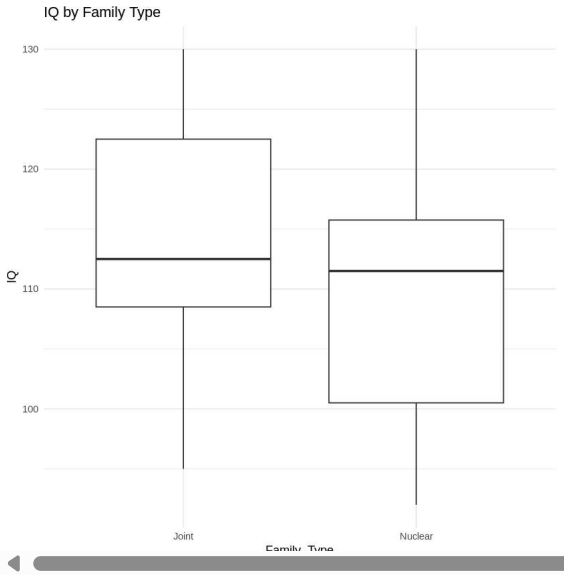
Gender	IQ
<chr>	<dbl>
Female	106.75
Male	116.75

A data.frame: 2 × 2

Family_Type	IQ
<chr>	<dbl>
Joint	113.8333
Nuclear	109.4286

```
#Boxplots to visualize IQ variation
ggplot(person_df, aes(x=Gender, y=IQ)) + geom_boxplot() + theme_minimal() + ggtitle("IQ by Gender")

ggplot(person_df, aes(x=Family_Type, y=IQ)) + geom_boxplot() + theme_minimal() + ggtitle("IQ by Family Type")
```






```
# Statistical testing
# t-test for Gender
t.test(IQ ~ Gender, data = person_df)

# t-test of Family_Type
t.test(IQ ~ Family_Type, data= person_df)
```

```
        Welch Two Sample t-test

data:  IQ by Gender
t = -1.9095, df = 14.061, p-value = 0.07681
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 -21.227373   1.227373
sample estimates:
mean in group Female   mean in group Male
              106.75              116.75

        Welch Two Sample t-test

data:  IQ by Family_Type
t = 0.72721, df = 9.1449, p-value = 0.4853
alternative hypothesis: true difference in means between group Joint and group Nuclear is not equal to 0
95 percent confidence interval:
 -9.264181 18.073705
sample estimates:
  mean in group Joint mean in group Nuclear
            113.8333              109.4286
```

d) Determine the factor affecting on psychological stress of the respondents.

```
person_df_encoded <- person_df
person_df_encoded$Psychological_Stress <- ifelse(person_df_encoded$Psychological_Stress == "Yes", 1, 0)

model <- glm(Psychological_Stress ~ Age + Height_in_inch+ IQ+ BMI, data= person_df_encoded)
summary(model)
```

```
Call:
glm(formula = Psychological_Stress ~ Age + Height_in_inch + IQ +
    BMI, data = person_df_encoded)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.414119   1.407006   2.427   0.0283 *
Age             0.002086   0.017291   0.121   0.9056
Height_in_inch -0.011563   0.009894  -1.169   0.2608
IQ             -0.018127   0.011236  -1.613   0.1275
BMI            -0.008600   0.028526  -0.301   0.7672
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2107939)

    Null deviance: 4.9500  on 19  degrees of freedom
Residual deviance: 3.1619  on 15  degrees of freedom
AIC: 31.866

Number of Fisher Scoring iterations: 2
```