MAASAI MARA UNIVERSITY

REGULAR UNIVERSITY EXAMINATIONS

2021/ 2022 ACADEMIC YEAR

THIRD YEAR FIRST SEMESTER

SCHOOL OF PURE, APPLIED AND HEALTH SCIENCES/SAHSSCI.
DEGREE IN APPLIED STATISTICS, CRIMINOLOGY, CMD AND SOCIAL WORKS.

COURSE CODE: STA 3125

COURSE TITLE: STATISTICAL METHODS AND DATA ANALYSIS.

DATE: 31st March, 2022 **TIME: 0830-1030**

INSTRUCTIONS TO CANDIDATES

Answer Question ONE and any other TWO questions

This paper consists of FOUR printed pages. Please turn over.

PART ONE

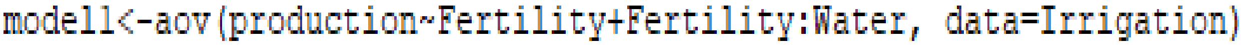
Question 1

a. Give an elaborate meaning of the following terms as used in programming
   i. Algorithm. (1 mark)
   ii. Compiler. (1 mark)

b. Below is a line of code extracted from **R** program. Briefly explain what the code does (2 marks)

   Data <-read.csv(file.choose() , header=TRUE)

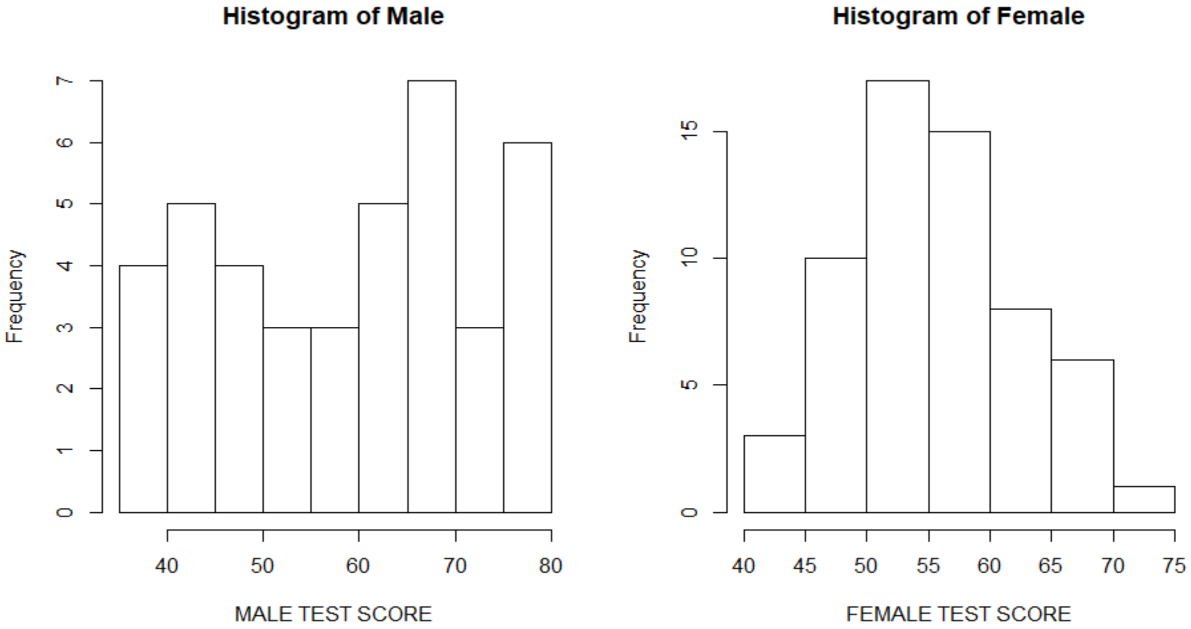
c. Below is an extract of an **R** code for an experimental design model fitting



```
modell<-aov(production~Fertility+Fertility:Water, data=Irrigation)
```

   Use it to answer the questions below;
   i. Which type of experimental design is being fitted in the model? (1 mark)
   ii. Describe what the experiment is investigating. (2 marks)
   iii. What is the line of code that should be added in order to view the results? (1 mark)
   iv. State one assumption that must be satisfied in order for the results given by the model to be valid. (1 mark)
   v. Give the appropriate diagnostic test that must be carried out to verify the assumption in (iv) together with the null and alternative hypothesis. (3 marks)
   vi. If the test in (v) shows that the assumption in (iv) is not satisfied, what is the next course of action? (2 marks)

d. The figure below shows the distribution of scores for students in a class, use it to answer the questions that follows;



   i. Describe the distribution of the male and female test scores. (2 marks)

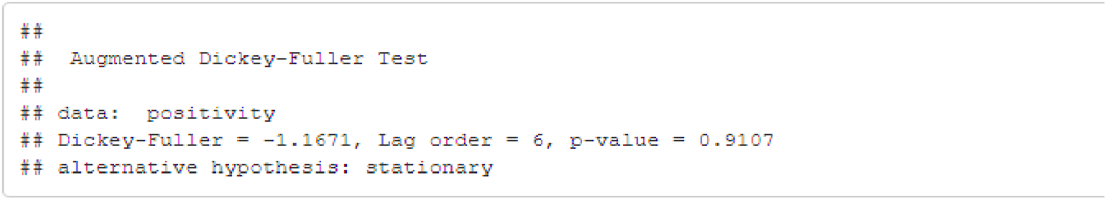
ii. Can we use independent sample t-test to compare the scores for males and female? Kindly justify your reasoning. (2 marks)

iii. Write down the R code that can be used to fit the above plot for female test scores given the data set for female test scores is called "Female". (2 marks)

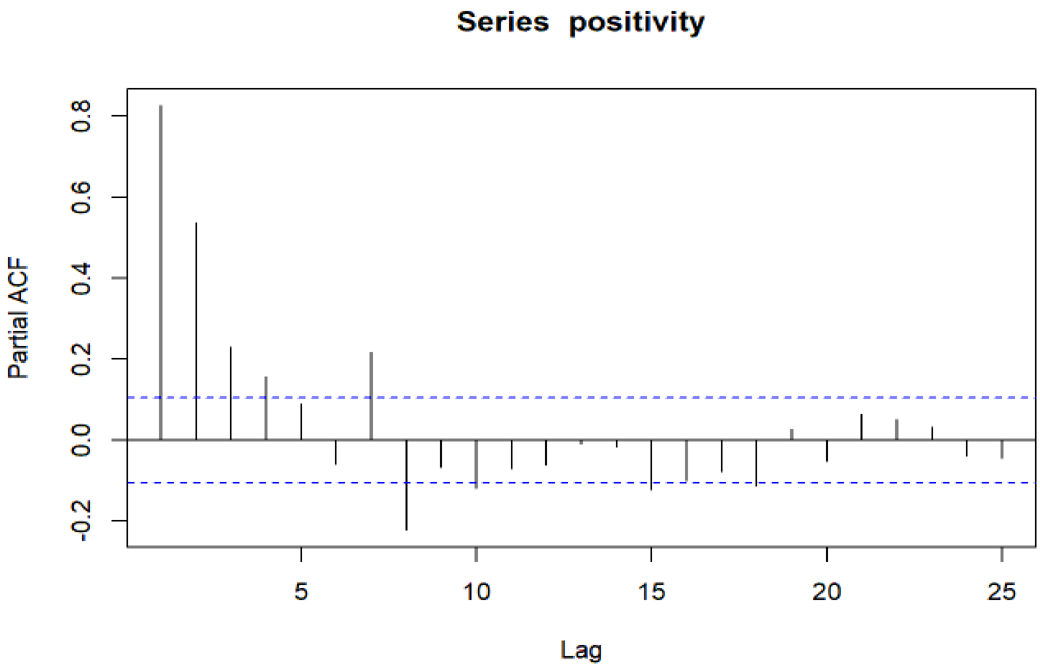
iv. From the above plot which group of students do you think performed well? Kindly justify your reasoning. (2 marks)

v. Give the R code for computing the appropriate measure of central tendency and dispersion for the male test score. (2 marks)

e. Below are results of a time series analysis, use it to answer the questions that follows;



```
##
##  Augmented Dickey-Fuller Test
##
## data:  positivity
## Dickey-Fuller = -1.1671, Lag order = 6, p-value = 0.9107
## alternative hypothesis: stationary
```



Based on the results above;

i. Which form of test is being carried out on the series above and what is the conclusion? (2 marks)

ii. Is there anything that should be done on the series with regard to the test result in (i)? If, yes, then what is it that should be done? If No, why? (2 marks)

iii. What defect is being illustrated on the series by the above plot? (1 mark)

iv. How can the defect in (iv) be corrected? (1 mark)

Question 2

a. Discuss the three control structures used in programming. (6 marks)

b. Write an R code that will accept coefficients of a quadratic equation from the user and use the coefficients given to determine the roots of the quadratic equation. (10 marks)

c. Write an algorithm for computing the surface area of an open cone. (4 marks)

Question 3

a. Below is a system of linear equation. Write down a sequence of R code that would be used to solve the linear system of equations using matrix algebra. (6 marks)

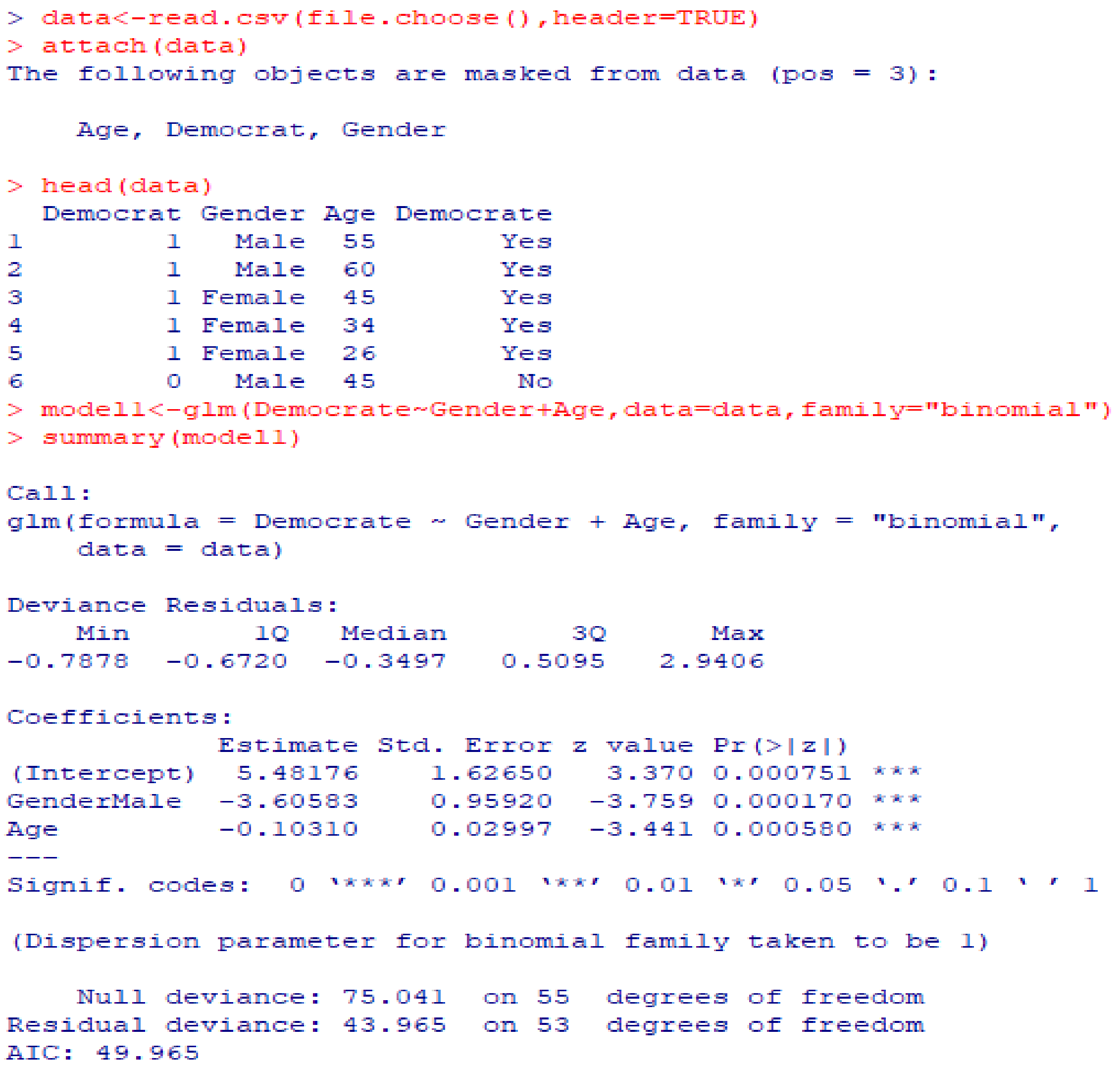
$$2x+3y-4z+6w=180$$

$$x+14y+2z-3w=236$$

$$9x-2y-3z+12w=350$$

$$7x+y+3z-8w=45$$

b. Below is an extract of R analysis, use it to answer the questions that follows;



```
> data<-read.csv(file.choose(),header=TRUE)
> attach(data)
The following objects are masked from data (pos = 3):

    Age, Democrat, Gender

> head(data)
  Democrat Gender Age Democrate
1        1   Male  55       Yes
2        1   Male  60       Yes
3        1 Female  45       Yes
4        1 Female  34       Yes
5        1 Female  26       Yes
6        0   Male  45        No
> model1<-glm(Democrate~Gender+Age,data=data,family="binomial")
> summary(model1)

Call:
glm(formula = Democrate ~ Gender + Age, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7878  -0.6720  -0.3497   0.5095   2.9406

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.48176    1.62650   3.370 0.000751 ***
GenderMale  -3.60583    0.95920  -3.759 0.000170 ***
Age         -0.10310    0.02997  -3.441 0.000580 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 75.041  on 55  degrees of freedom
Residual deviance: 43.965  on 53  degrees of freedom
AIC: 49.965
```

i. Write an R code that will be used to extract only Females from the data set.

(1 mark)

ii. Write an R code that will be used to filter out Males who are democrats from the data set. (1 mark)

iii. Write an R code that will be used to compute the Average age of individuals who are Democrats or Male. (3 marks)

iv. Which sort of model is fitted in the output above. (1 mark)

v. Give one an assumption of the model stated in (iv). (1 mark)

vi. What is the aim of the analysis in the output above? (1 mark)

vii. Compute the odds ratio for the independent variables and interpret the results.

(4 marks)

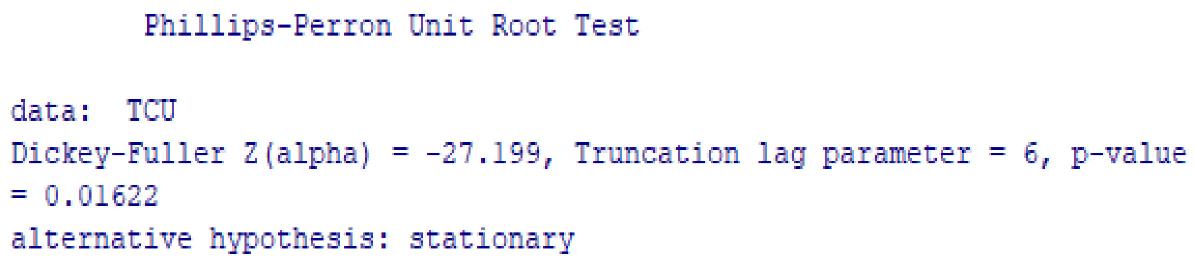
viii. Determine the probability of a female being a democrat from the above analysis.

(2 marks)

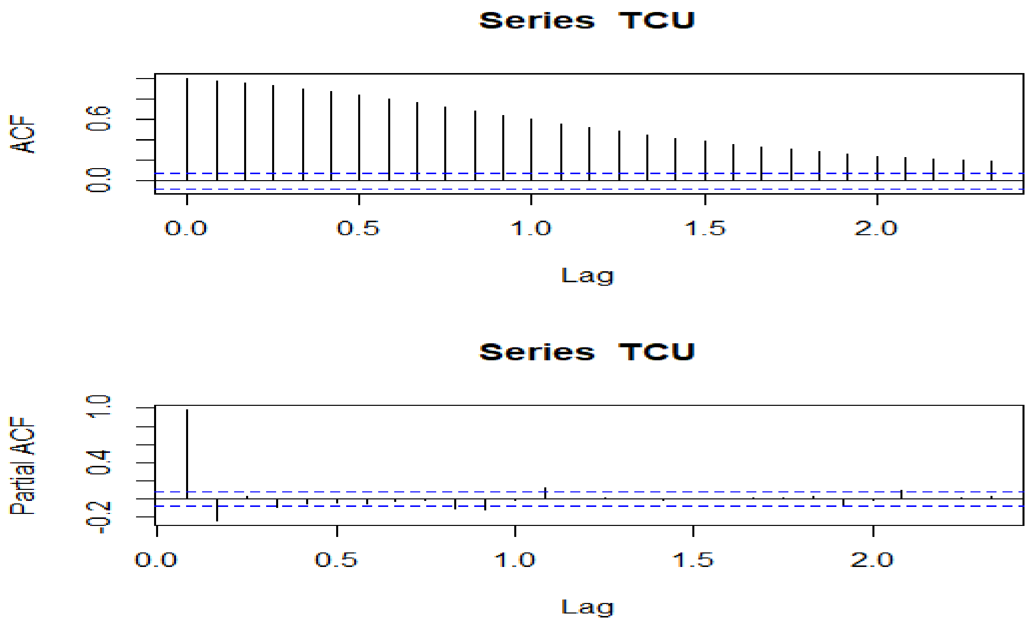
Question 4

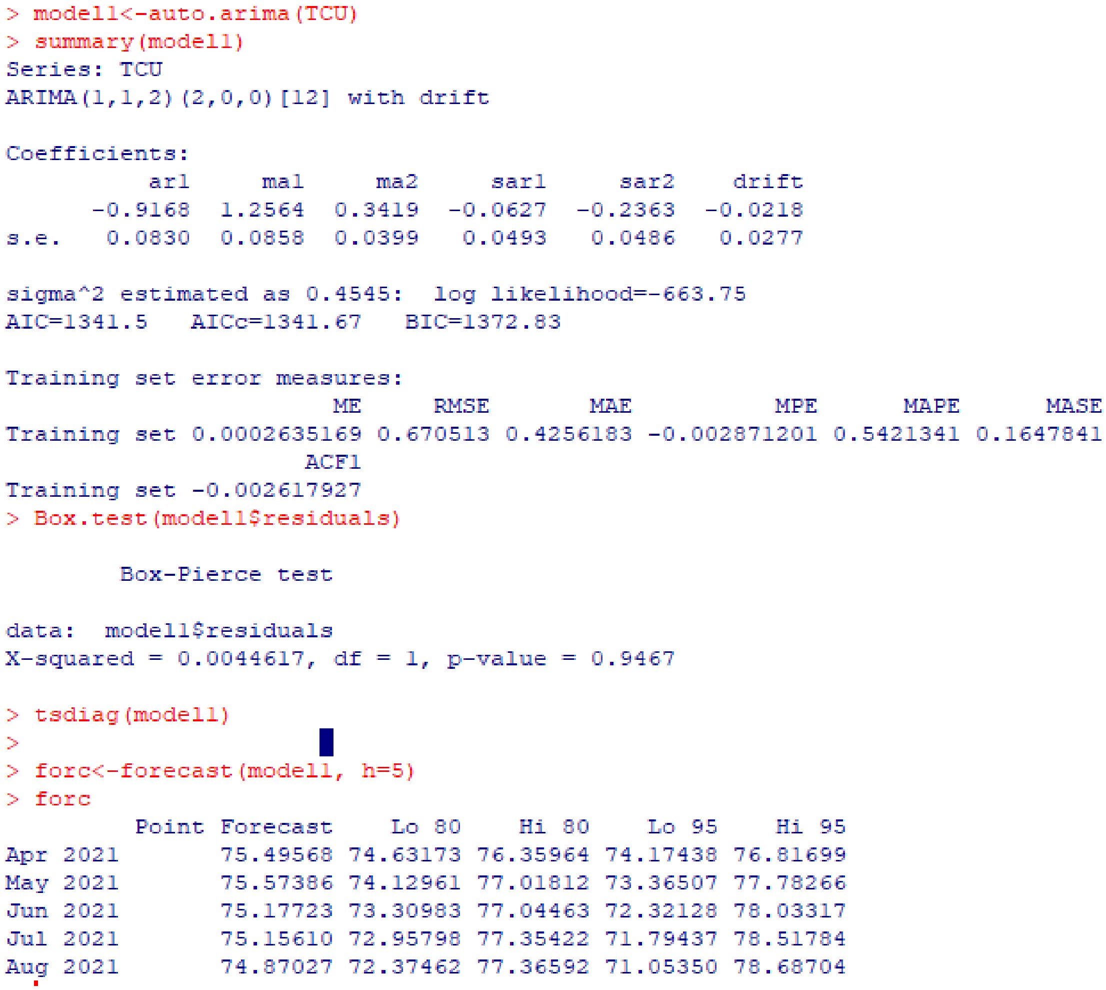
a. Given the following data:
X: 5, 7, 11, 10, 3, 12, 8.
Y: 3.5, 8, 13, 12, 5, 15, 9
Write a code to Calculate:
i. Pearson's Rank correlation coefficient between X and Y (6 marks)
ii. Assume X measures the weight of children which is estimated to have a mean of 6.5. Perform a one sample t test**.** (4 marks)

b. Extract below is part of an R Analysis Use it to answer the questions that follows;

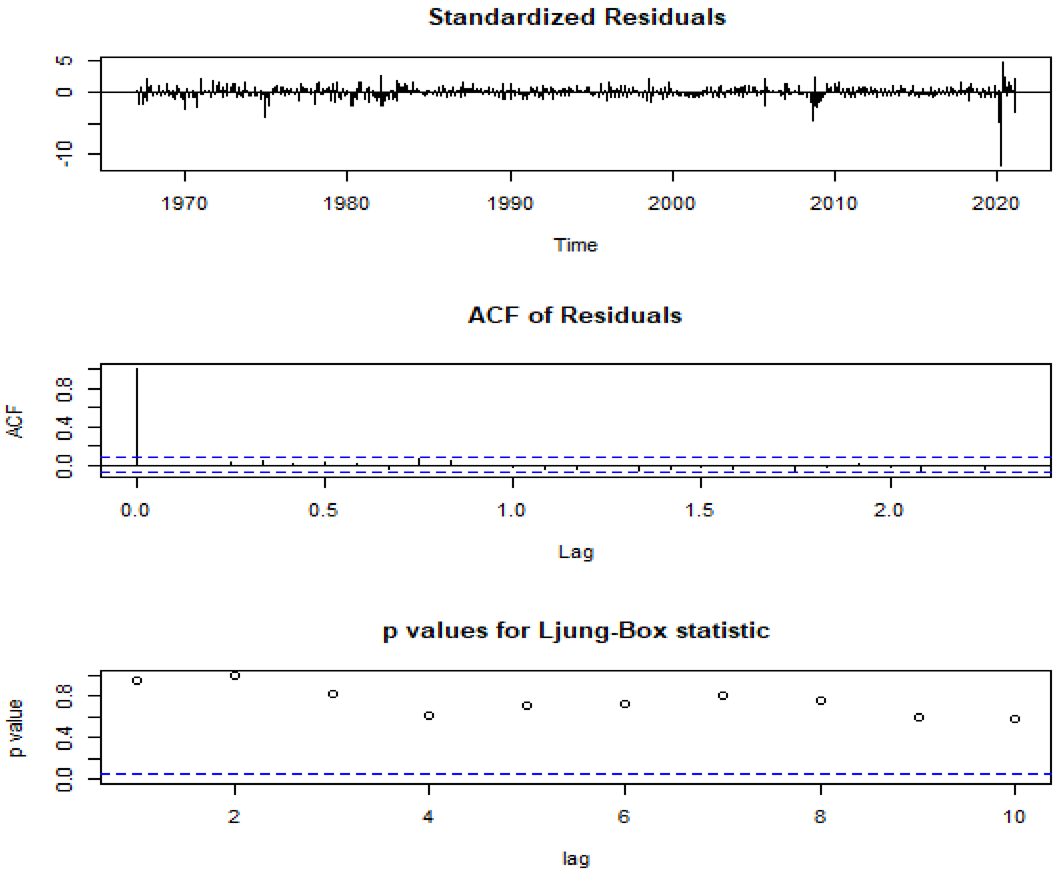


```
        Phillips-Perron Unit Root Test

data:  TCU
Dickey-Fuller Z(alpha) = -27.199, Truncation lag parameter = 6, p-value
= 0.01622
alternative hypothesis: stationary
```





```
> modell<-auto.arima(TCU)
> summary(modell)
Series: TCU 
ARIMA(1,1,2)(2,0,0)[12] with drift 

Coefficients:
          ar1     ma1     ma2     sar1     sar2    drift
      -0.9168  1.2564  0.3419  -0.0627  -0.2363  -0.0218
s.e.   0.0830  0.0858  0.0399   0.0493   0.0486   0.0277

sigma^2 estimated as 0.4545:  log likelihood=-663.75
AIC=1341.5   AICc=1341.67   BIC=1372.83

Training set error measures:
                       ME     RMSE       MAE          MPE      MAPE      MASE
Training set 0.0002635169 0.670513 0.4256183 -0.002871201 0.5421341 0.1647841
                     ACF1
Training set -0.002617927
> Box.test(modell$residuals)

        Box-Pierce test

data:  modell$residuals
X-squared = 0.0044617, df = 1, p-value = 0.9467

> tsdiag(modell)
> 
> forc<-forecast(modell, h=5)
> forc
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Apr 2021       75.49568 74.63173 76.35964 74.17438 76.81699
May 2021       75.57386 74.12961 77.01812 73.36507 77.78266
Jun 2021       75.17723 73.30983 77.04463 72.32128 78.03317
Jul 2021       75.15610 72.95798 77.35422 71.79437 78.51784
Aug 2021       74.87027 72.37462 77.36592 71.05350 78.68704
```


Standardized Residuals

ACF of Residuals

p values for Ljung-Box statistic

i. Give two issues that have been identified in the time series TCU based on the results above. (2 marks)

ii. What corrective action was taken by the analyst to correct the issues identified in (i)? (1 mark)

iii. Was the corrective measure employed by the analyst in (ii) adequate enough in addressing the issue? Kindly give adequate evidence from the results above to justify your reasoning. (3 marks)

iv. What does the numbers (1,1,2) in the ARIMA model stands for? (3 marks)

v. Kindly make a conclusion regarding the 95% confidence interval for the predicted TCU values in the month of June 2021. (1 mark)

PART TWO

Question One

a. Differentiate the following terms as used in statistics
   i. Type I and Type II error **(2 marks)**
   ii. Snowball sampling and quota sampling **(2 marks)**

b. Give three assumptions for binomial random process **(3 marks)**

c. Give two differences between regression and correlation **(2 marks)**

d. The incidence of malaria in Maasai Mara University is that students have 30% chance of suffering from it. What is the probability that out of 12 students three or more will contract the disease? **(3 marks)**

e. The average rate of vehicles arriving randomly at a petrol station is 20 per minute. 10% of the vehicles are trucks. Compute the probability that:
   i. 50 vehicles arrive within 2 minutes **(2 marks)**
   ii. 40 cars arrive within 2 minutes **(2 marks)**

f. Three groups of sociologists contain respectively 3 women and 1 man, 2 women and 2 men, 1 woman and 3 men. One economist is selected at random from each group. Calculate the chance that the three selected consists of 1 woman and 2 men **(4 marks)**

g. There are 10 numbers; 0 through to 9, which are to be used in code group of four to identify an item of clothing in a boutique shop e.g. code 1083 is to identify blue blouse, code 1030 is identify a pair of socks and so on. How many codes can you generate if repetition of numbers is not permitted **(3 marks)**

h. The mean lifetime of a sample of 100 light tubes produced by a company is found to be 1620 hours with standard deviation of 72 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1600 hours. **(3 marks)**

i. A random sample of 16 items is taken and is found to have a mean weight of 48 grams and a standard deviation of 9 grams. What is the mean weight of population:
   a. With 95% confidence? **(2 marks)**
   b. With 99% confidence? **(2 marks)**

Question Two

a. Two different types of drugs A and B were tried on certain patients for increasing weight. 6 persons were given drug A and 8 persons were given drug B. The increase in weight (in pounds) is given below.

Drug A	8	12	13	9	3	10		
Drug B	10	8	12	15	6	8	11	13

Do the two drugs differ significantly with regard to their effect in increasing weight? **(7 marks)**

b. A sample of 300 students with a particular disease was selected. Out of these, 150 were given a drug and the other were not given any drug. The results are as follows

	Drug	No drug
Cured	87	80
Not cured	63	70

Test whether the drug is effective or not **(7 marks)**

c. A survey was conducted on the newspapers readership of three dailies: Nation Daily (D), the Standard (S) and the Kenya times (K) in the University and the following data was obtained. The number that read: D and K =19, S and D =17, S and K= 11, D, K and S =6 only, D = 65, S = 51, K= 47

Determine the number of people who read:

i) Daily nation only **(2 marks)**

ii) The Kenya times only **(2 marks)**

iii) Standard or Kenya times but not Daily nation **(2 marks)**

Question Three

a. In a certain town, male and female each form 50 percent of the population. It is known that 20 percent of the males and 5 percent of the female are unemployed. A research student studying the employment situation selects an unemployment person at random. Using Bayes' Theorem what is the probability that the person selected is (a) Male (b) Female? **(6 marks)**

b. In the following table are recorded data showing the test score made by salesmen on intelligence test and their weekly sales.

Salesmen	1	2	3	4	5	6	7	8	9	10
Test score	40	68	50	64	80	53	76	40	66	60
Sales (000 Sh.)	2.5	6.4	4.0	5.5	4.0	2.5	5.5	3.0	4.5	3.7

Calculate:

i. Regression line of sales on test score **(8 marks)**

ii. Estimate the probable weekly sales volume if a salesman makes a score of 120 **(2 marks)**

iii. Calculate the correlation coefficient between the two variables under study **(4 marks)**

Question Four

a. Differentiate using examples between descriptive and inferential statistics **(4 marks)**

b. A researcher wanted to investigate if the average number of crimes reported per day is different between Narok, Nakuru and Kericho Counties. The researcher recorded the number of crimes reported in the three town in a single week. A one-way analysis of variance test was carried out on the data set and the results were as illustrated below;

Descriptives

Number of crimes reported

County	N	Mean	Std.	Std. Error	95%

			Deviation		Confidence Interval for Mean	
					L. B	U. B.
Narok	7	11.00	2.449	.926	8.73	13.27
Nakuru	7	21.29	1.799	.680	19.62	22.95
Kericho	7	16.14	1.574	.595	14.69	17.60
Total	21	16.14	4.693	1.024	14.01	18.28

ANOVA				
Source of variatio	Sum of Squares	df	Mean Squares	F
Between groups	370.286			
Within groups				
Total	440.571			

i) State the null and alternative hypothesis for the study **(2 marks)**

ii) Complete the ANOVA table above **(7 marks)**

iii) Based on the test results above, at 95% level of confidence is there sufficient evidence to show that the number of crime reported in the three towns is different. **(3 marks)**

iv) State two assumptions that must be satisfied in order to carry out the above test **(2 marks)**

v) Give the corrective measure that would be taken to test the hypothesis made in the study if each of the assumptions in (iv) are violated **(2 marks)**

/////END/////