ALUPE UNIVERSITY

COLLEGE

...Bastion of Knowledge....

P. O.Box 845-50400 **Busia(K)**
principal@auc.ac.ke
Tel: +254 741 217 185
+254 736 044 469
off Busia-Malaba road

OFFICE OF THE DEPUTY PRINCIPAL
ACADEMICS, STUDENT AFFAIRS AND RESEARCH

UNIVERSITY EXAMINATIONS

2019 /2020 ACADEMIC YEAR

THIRD YEAR SECOND SEMESTER REGULAR EXAMINATION

FOR THE DEGREE OF BACHELOR OF SCIENCE (COMPUTER SCIENCE)

COURSE CODE: **COM 325**

COURSE TITLE: **COMPUTER APPLICATION II**

DATE: 4TH NOVEMBER, 2020 TIME: 0900 – 1200 HRS

INSTRUCTION TO CANDIDATES

- **SEE INSIDE**

THIS PAPER CONSISTS OF 4 PRINTED PAGES PLEASE TURN OVER

REGULAR-MAIN EXAM

COM 325 E: COMPUTER APPLICATION II

STREAM: COM **DURATION: 3 Hours**

INSTRUCTION TO CANDIDATES

Answer **ALL** questions from section A and any **THREE** from section B.

SECTION A [24 MARKS]. ANSWER ALL QUESTIONS.

QUESTION ONE [12 Marks]

a) Define clearly the following terms as used in data analysis [6 Marks]

i) Variable

ii) Outliers.

iii) Statistical software.

iv Data mining.

v) Data Wangling

vi) Pandas

b) Explain 4 main stages of data processing cycle [4 Marks]

c) Data cleaning is the initial screening the collected raw data made to assess its validity and usefulness. Identify and discuss briefly two objectives of this. [2 Marks]

QUESTION TWO [12 Marks]

a) How does a database differ from a spread sheet? [2 Marks]

b) In a study on relationship between rates of marriage and affairs, data was obtained from a study (Ray, 1978); summary of the variables is given below:

```
Number of observations: 6366
    Number of variables: 9
    Variable name definitions:

    rate_marriage  : How rate marriage, 1 = very poor, 2 = poor, 3 = fair,
                            4 = good, 5 = very good
    age            : Age
    yrs_married    : No. years married. Interval approximations. See
                            original paper for detailed explanation.
    children       : No. children
    religious      : How relgious, 1 = not, 2 = mildly, 3 = fairly,
                            4 = strongly
    educ           : Level of education, 9 = grade school, 12 = high
                    school, 14 = some college, 16 = college graduate,
                            17 = some graduate school, 20 = advanced degree
    occupation     : 1 = student, 2 = farming, agriculture; semi-skilled,
                      or unskilled worker; 3 = white-colloar; 4 = teacher
                      counselor social worker, nurse; artist, writers;
                               technician, skilled worker, 5 = managerial,
            administrative, business, 6 = professional with
                            advanced degree
    occupation_husb: Husband's occupation. Same as occupation.
    affairs        : measure of time spent in extramarital affairs
```

Write python Codes to perform;

i) Pearson correlation coefficient and Kendall rank correlation coefficient [6 marks]

ii) Ordinary Linear regression for predicting affairs using how the individual rates marriage [4 marks]

SECTION B [36 Marks] Answer any THREE questions]

QUESTION THREE [12 Marks]

The following data gives the hourly numbers of units produced and the hourly number of items spoiled for 10 press operators.

Operator	1	2	3	4	5	6	7	8	9	10
Units produced per hour (X)	23	11	32	16	19	25	19	29	28	12
Units spoiled per hour (Y)	22	13	22	32	39	13	43	21	47	16

Using the above data, calculate;

a) Karl Pearson coefficient of correlation. [6 Marks]

b) Spearman's rank correlation coefficient. [6 Marks]

QUESTION FOUR [12 Marks]

In an annual study on body weight (BW) in kilograms and resting metabolic rate (RMR) levels in kcal/24hours for seven individuals the following data for BW and RMR respectively. (60.0, 1330), (72.8, 1382), (57.6, 1325), (64.9, 1365), (59.2, 1342), (77.1, 1439) and (82.0, 1475)

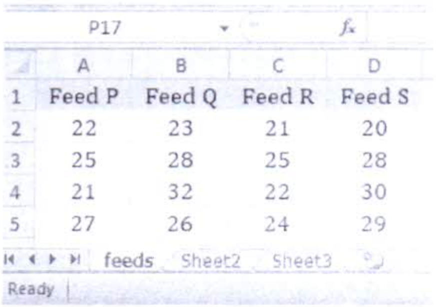
i) Plot a scatter diagram and comment on the relationship [3 Marks]

ii) Develop a regression equation used to predict the RMR levels using BW. [9 Marks]

QUESTION FIVE [12 Marks]

a) What is multicollinearity, problems it causes and how can this be corrected in regression analysis [3 Marks]

b) An experiment was conducted to investigate the effectiveness of various feed supplements on the growth rate (weight) using a completely randomized design. Sixteen pigs were randomly selected and fed on four different feed supplements P, Q, R and S and their growth rate were as entered in the .csv file below



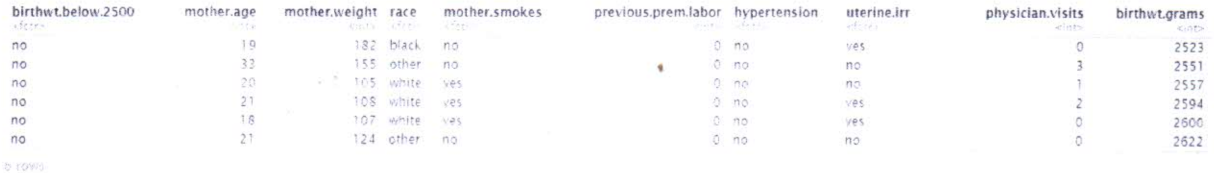
	A	B	C	D
1	Feed P	Feed Q	Feed R	Feed S
2	22	23	21	20
3	25	28	25	28
4	21	32	22	30
5	27	26	24	29

Write R code for data input procedure and perform ANOVA test [9 Marks]

QUESTION SIX [12 MARKS]

a) Explain any advantage of using R over Python in data analysis [2 Marks]

b) A researcher is interested in comparing an outcome between two groups; birth weight between smoking and non-smoking mothers. The head of the dataset is shown below;



birthwt.below.2500	mother.age	mother.weight	race	mother.smokes	previous.prem.labor	hypertension	uterine.irr	physician.visits	birthwt.grams
<fctr>	<int>	<int>	<fctr>	<fctr>	<int>	<fctr>	<fctr>	<int>	<int>
no	19	182	black	no	0	no	yes	0	2523
no	33	155	other	no	0	no	no	3	2551
no	20	105	white	yes	0	no	no	1	2557
no	21	108	white	yes	0	no	yes	2	2594
no	18	107	white	yes	0	no	yes	0	2600
no	21	124	other	no	0	no	no	0	2622

6 rows

Provide an R code to;

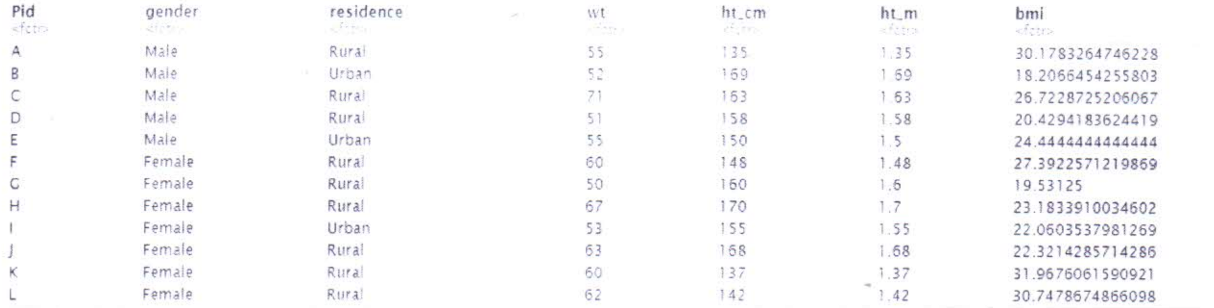
i) Plot a box plot for the birth weight between smoking and non-smoking mothers. Put appropriate labelling and colourings. [4 marks]

ii) Assess whether this difference is statistically significant? [3 Marks]

iii) Compute the confidence interval for the two groups [3 marks]

QUESTION SEVEN [12 MARKS]

In a medical study to determine the body mass index of patients a sample of twelve patients were randomly selected, their gender, heights, weights and residence were recorded and captured as shown in the screenshot below.



Pid	gender	residence	wt	ht_cm	ht_m	bmi
<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>
A	Male	Rural	55	135	1.35	30.1783264746228
B	Male	Urban	52	169	1.69	18.2066454255803
C	Male	Rural	71	163	1.63	26.7228725206067
D	Male	Rural	51	158	1.58	20.4294183624419
E	Male	Urban	55	150	1.5	24.4444444444444
F	Female	Rural	60	148	1.48	27.3922571219869
G	Female	Rural	50	160	1.6	19.53125
H	Female	Rural	67	170	1.7	23.1833910034602
I	Female	Urban	53	155	1.55	22.0603537981269
J	Female	Rural	63	168	1.68	22.3214285714286
K	Female	Rural	60	137	1.37	31.9676061590921
L	Female	Rural	62	142	1.42	30.7478674866098

Write commands in R and give a brief explanation that would;

i) Assign the patients id to the 'pid' attribute of this vector. Also create a vector 'gender', 'residence, 'wt_kg', 'ht_cm','ht_m' that corresponds to gender, residence, weight, heights in centimeters and heights in meters respectively. [4 Marks]

ii) Compute patient's body mass index (BMI),where $bmi = \frac{wt\_kg}{ht\_m^2}$ [1 Mark]

iii) Bind all the variables in i and ii above [3 Marks]

iv) Generate table above, and sot BMI descriptive statistics [4 Marks]