University No:_____

THE UNIVERSITY OF HONG KONG
FACULTY OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE
COMP2501 Introduction to Data Science and Engineering
Examination

Date: December 9, 2023 Time: 2:30pm-4:30pm

INSTRUCTIONS:

a. This paper has two parts: Short Questions (60%) and Long Questions (40%)

b. Answer ALL questions.

c. Only approved calculators as announced by the Examinations Secretary can be used in this examination. It is candidates' responsibility to ensure that their calculator operates satisfactorily, and candidates must record the name and type of the calculator used on the front page of the examination script.

d. Candidates are permitted to bring five sheets of A4-sized paper with printed or written notes on both sides to the examination.

e. Write your university no. at the top of every page.

University No:_____

Part A: Short Questions (60%)

Give your answers in the answer boxes provided below each question. All questions carry equal marks.

(1)

(a) (3%) Construct a table named 'snd' in R that displays the probabilities Pr(|Z| < z) for values of z in the sequence 1.0, 1.05, ..., 3.0, where Z follows the Standard Normal Distribution. We show in the following the head, tail and the complete table of 'snd':

```
head(snd)
A tibble: 6 × 2
     z  Prob
  <dbl> <dbl>
  1     0.683
  1.05  0.706
  1.1   0.729
  1.15  0.750
  1.2   0.770
  1.25  0.789
```

```
tail(snd)
A tibble: 6 × 2
     z  Prob
  <dbl> <dbl>
  2.75  0.994
  2.8   0.995
  2.85  0.996
  2.9   0.996
  2.95  0.997
  3     0.997
```

1	1.00	0.6826895	11	1.50	0.8663856	21	2.00	0.9544997	31	2.50	0.9875807
2	1.05	0.7062819	12	1.55	0.8788585	22	2.05	0.9596356	32	2.55	0.9892277
3	1.10	0.7286679	13	1.60	0.8904014	23	2.10	0.9642712	33	2.60	0.9906776
4	1.15	0.7498561	14	1.65	0.9010571	24	2.15	0.9684448	34	2.65	0.9919508
5	1.20	0.7698607	15	1.70	0.9108691	25	2.20	0.9721931	35	2.70	0.9930661
6	1.25	0.7887005	16	1.75	0.9198817	26	2.25	0.9755511	36	2.75	0.9940405
7	1.30	0.8063990	17	1.80	0.9281394	27	2.30	0.9785518	37	2.80	0.9948897
8	1.35	0.8229640	18	1.85	0.9356865	28	2.35	0.9812266	38	2.85	0.9956281
9	1.40	0.8384867	19	1.90	0.9425669	29	2.40	0.9836049	39	2.90	0.9962684
10	1.45	0.8529415	20	1.95	0.9488239	30	2.45	0.9857144	40	2.95	0.9968223

Answer:

(b) (7%) The time X (in seconds) taken for a certain chemical reaction follows a normal distribution with a standard deviation of 5seconds. To estimate the mean time taken for the reaction, nine experiments were conducted and the reactions were recorded as follows: 11.1, 13.2, 9.8, 12.1, 10.9, 12.2, 14.3, 10.8, 11.7. Using this data, construct an 80% confidence interval for the mean reaction time. Please round your answer to 2 decimal places. You may consult the 'snd' table for your calculation.

Answer

(2) (10%) Assuming the variable `n` stores a positive number. Give a single R statement that prints a list of all prime numbers greater than 2 and smaller than `n`.
Answer:

(3)
(a) (5%) We conducted a coin flip experiment consisting of 100 flips, resulting in 61 heads heads and 39 tails. Our null hypothesis states that "The coin is fair", meaning that the probability of getting a head is 0.5, and the probability of getting a tail is also 0.5. Your task is to determine whether to accept or reject the null hypothesis. Please support your decision using the p-value of the outcome. You may consult the 'snd' table to obtain an upper bound on the p-value.
Answer

(b) (5%) Recall that the covariance of two sequences of numbers $x_1, x_{2,\ldots,} x_n$ and $y_1, y_2, \ldots, y_n$ is defined as

$$((x_1 - \mu_x)(y_1 - \mu_y) + (x_2 - \mu_x)(y_2 - \mu_y) + \cdots + (x_n - \mu_x)(y_n - \mu_y))/n$$

where μ_x and μ_y are the mean of x and y, respectively. In general, if the covariance of two sequences is significantly greater than zero, they are positively related. This means that the greater values of one variable mainly corresponding with the greater values of the other variable. If the covariance is significantly smaller than zero, they are negatively related. If their covariance is zero, the two sequences do not have any relation. However, this question challenges the notion that this intuition is always true. Your task is to complete the following two sequences x=[-3,-2,-1,...] and y=[3,2,1,...] such that (i) the two sequences are closely related, meaning that given a value of x, we can determine the corresponding value of y correctly, and (ii) the covariance of the two sequences is zero. Justify your answer.
Answer:

University No:_____

4.

(a) (5%) Let `s = "Bookname: A happy teacher, author: Peter Chan, borrower: UID12345"`. What is the output of the print statements given in (a) to (e):

a. `print(str_extract(s, "Bookname:\\s*[A-Z][a-z|\\s*]+"))`

b. `print(str_extract(s, "^(?=[^ABC][^DEF])Bookname:\\s*[A-Z][a-z|\\s*]+"))`

c.
```
tab <- str_match(s, "(borrower):\\s*UID(\\d{5})
print(tab[[2]])
```

d. `print(str_extract(s, "(?<=P)[a-zA-Z]+"))`

e.
```
s = c("5'2", "8'34", "5' 25", "5, 8", "6, ")
print(sum(str_detect(s, "^[4-7]'\\d{1,2}$")))
```

Answer

(b) (5%) Give the regular expression pattern that will accept the following two strings as valid (email addresses):

`angus12.junior@gmail.com`
`my.friend-386.email@my-school.org`

and the following string as invalid (email address)

`peterchan.hku.hk`

Answer

(5) (10%) You have developed a machine learning model for a binary classification problem where the outcome can be either 'Y' or 'N'. In a specific experiment, you have obtained the following result regarding its performance

Data point	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
True outcome	Y	N	Y	Y	N	Y	Y	Y	N	Y
Predict outcome	Y	Y	N	Y	Y	N	N	Y	Y	N

Data point	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
True outcome	Y	N	Y	Y	N	Y	Y	Y	N	Y
Predict outcome	N	N	Y	Y	N	N	N	Y	Y	N

(i) Suppose that the True outcomes and Predicted outcomes are stored in the vectors y and y_hat, respectively. Give the R statement that creates the confusion matrix of y and y_hat and assigns it to the variable CM.

(ii) What is the accuracy, sensitivity, specificity, and F1-score of the test method.

(iii) Can a model be created that guarantees100% specificity? Justify your answer.

(iv) Can a model be created that guarantees100% sensitivity? Justify your answer.

(v) Prove that if you can have a model with 100% specificity and 100% sensitivity, then it has 100% accuracy.

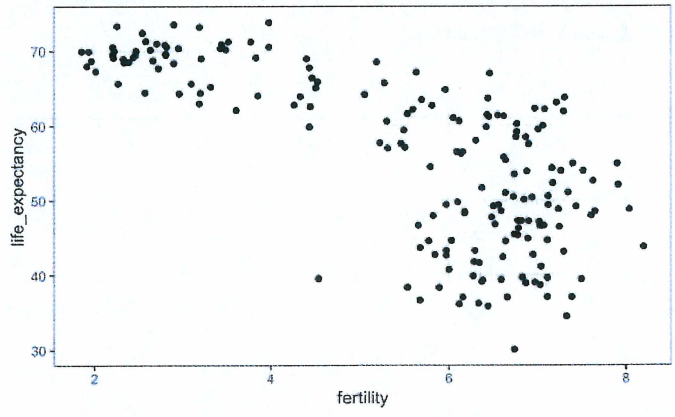
Answer

University No:_____

6. Consider the following data frame gapminder given in R's dslabs library:

```
> library(dslabs)
> data("gapminder")
> gapminder
                  country year infant_mortality life_expectancy fertility population          gdp continent                  region
1                 Albania 1960           115.40           62.87      6.19    1636054           NA    Europe          Southern Europe
2                 Algeria 1960           148.20           47.50      7.65   11124892  13828152297    Africa          Northern Africa
3                  Angola 1960           208.00           35.98      7.32    5270844           NA    Africa            Middle Africa
4      Antigua and Barbuda 1960               NA           62.97      4.43      54681           NA  Americas               Caribbean
5               Argentina 1960            59.87           65.39      3.11   20619075 108322326649  Americas           South America
6                 Armenia 1960               NA           66.86      4.55    1867396           NA      Asia            Western Asia
7                   Aruba 1960               NA           65.66      4.82      54208           NA  Americas               Caribbean
8               Australia 1960            20.30           70.87      3.45   10292328  96677859364   Oceania Australia and New Zealand
9                 Austria 1960            37.30           68.75      2.70    7065525  52392699681    Europe          Western Europe
10             Azerbaijan 1960               NA           61.33      5.57    3897889           NA      Asia            Western Asia
11                Bahamas 1960            51.00           62.00      4.50     109526   1306269490  Americas               Caribbean
12                Bahrain 1960           134.50           51.64      7.09     162501           NA      Asia            Western Asia
13             Bangladesh 1960           176.30           46.20      6.73   48200702  12767231590      Asia            Southern Asia
14               Barbados 1960            69.50           61.80      4.33     230934    784120376  Americas               Caribbean
15                Belarus 1960               NA           71.59      2.74    8190027           NA    Europe          Eastern Europe
16                Belgium 1960            29.50           69.59      2.60    9140563  68236665814    Europe          Western Europe
17                 Belize 1960               NA           60.08      6.50      92068     86532304  Americas         Central America
18                  Benin 1960           186.90           38.29      6.28    2431620    621797131    Africa          Western Africa
19                 Bhutan 1960           175.00           35.94      6.67     224108           NA      Asia            Southern Asia
20                Bolivia 1960           173.40           43.77      6.70    3693451   3001815692  Americas           South America
```

(a) (2%) Give R's ggplot2 code to generate the following scatterplots of life expectancy vs fertility rates in 1962.

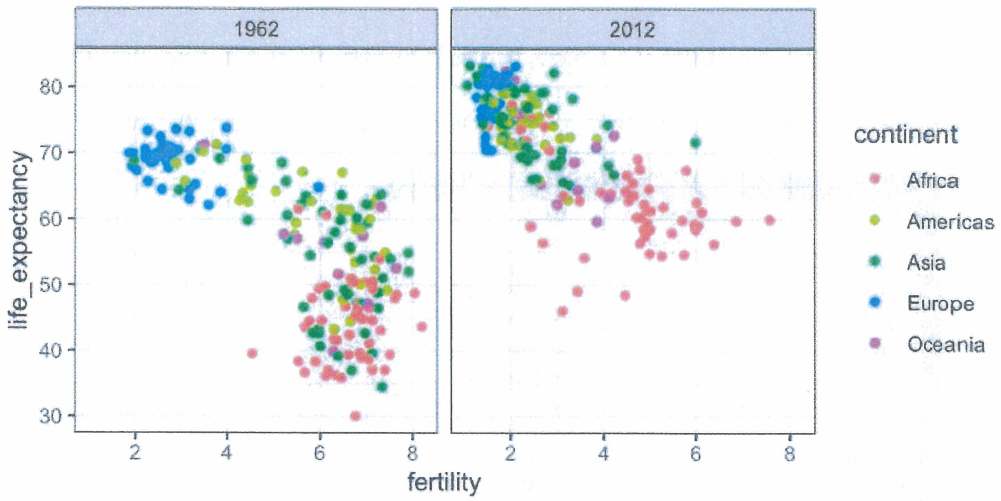


Answer

(b) (2%) Modify your code in (a) so that different colors are used for the points for different continents.

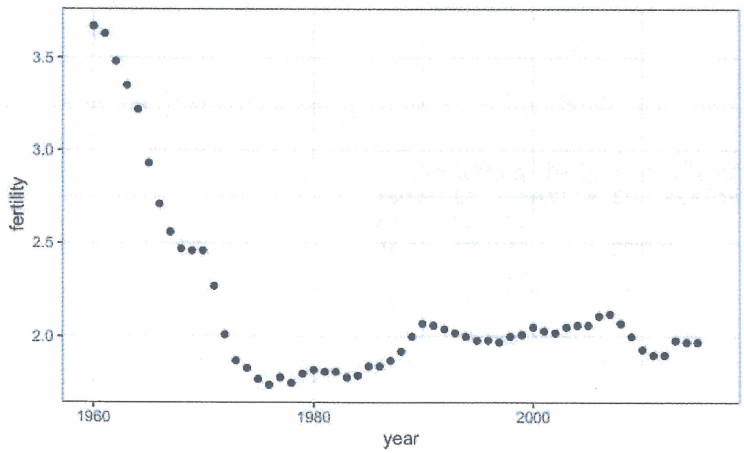
Answer

(c) (3%) Give the ggplot2 code to display the following diagram for comparing the relationship between fertility and life expectancy in the years 1962 and 2012:



Answer

(d) (3%) Give the ggplot2 code for the following plot showing the fertility rate in US over the years.



Answer

University No:_____

Part B: Long Questions (40%) Again the two questions carry equal marks.

(1) Suppose the following excel file "marks.xls" is stored in the working directory.

	A	B	C
1	UID	Subclass	Marks
2	U2022123	A	A1_34
3	U2022123	A	MidTerm_67
4	U2022123	A	A2_100
5	U2022123	A	Final_95
6	U1987136	D	A1_100
7	U1987136	D	MidTerm_76
8	U1987136	D	A2_92
9	U1987136	D	Final_86
10	U2003642	B	A1_75
11	U2020897	C	A1_86
12	U1999482	C	A1_52
13	U2005683	B	A2_84
14	U2011576	D	Final_53
15	U2020245	A	Midterm_100
16	U2021732	D	Final_75

(a) (4%) Please provide the R statements that can read the file"marks.xls" and store its contents in a data frame named 'marksdf'.

Answer

(b) (4%) Reshape and transform 'marksdf' as follows:

```
        UID Subclass  A1 Midterm  A2 Final
1 U2022123        A  34      67 100    95
2 U1987136        D 100      76  92    86
3 U2003642        B  75      84  74   100
4 U2020897        C  86      20  55    54
```

Answer

(c) (4%) The entries in the UID column should follow the format of a letter 'U' followed by 6 digits, while the entries in the Subclass column should be limited to 'A', 'B', 'C', and 'D'. Give the R statements that detect if the UID and Subclass columns in the transformed data frame 'marksdf' from (b) have invalid entries.

Answer

(d) (4%) We have observed that some values in the UID column are invalid and do not follow the format of a letter 'U' followed by 6 digits. Specifically, the invalid values have the format: Xdddddd where X is a character other than 'U' and the d's represent digits. Give the R statements to make the correction by replacing X with 'U'.

Answer

(e) (4%) Add a column called 'Weighted Mark' to the data frame. This column should calculate and store the weighted average of the marks in each row. The calculation should be based on the following weights: 0.15 * A1 + 0.15 * A2 + 0.2 * mid_term + 0.5 * final.

Answer

University No:_____

(2) This question does not have a definitive answer and is open-ended. You will be evaluated on the quality of your response, which should demonstrate insightful and logical arguments. There are no right or wrong answers, but a good grade will be awarded based on the quality of your insights and reasoning.

The *highest* kidney-cancer rates in the United States, measured by the number of cases per 100,000 of population, occur in very rural counties sprawled out across the Midwest, South and West regions of the country.

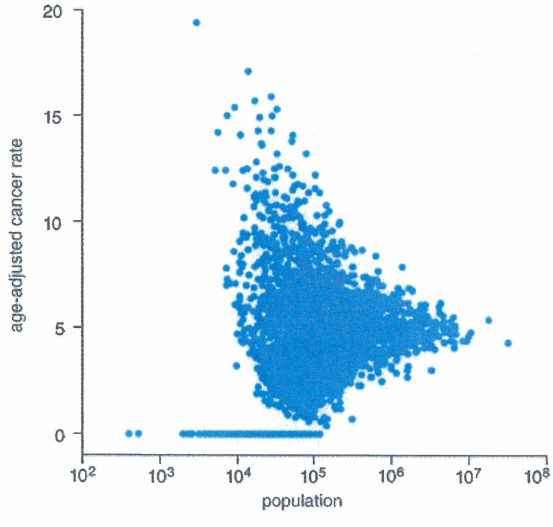
(a) (6%) Think about why this might be? Give your reasons.

Answer

(b) (7%) We observe another fact: the *lowest* kidney-cancer rates in the United State also occur in very rural counties sprawled out across the Midwest, South, and West, often neighboring the borders of the highest-rate counties. How can both (a) and (b) be true? How can two towns (or cities) with similar demographics have so very different results? Try to give your reasons.

Answer

(c) (7%) The following diagram plots the cancer rate and population for all the counties. Does it give you some idea to answer (b)?



Note: Rural counties are small towns/cities made up of rural areas, with a low population density, and usually focus on agriculture, natural resource extraction, and traditional rural industries. They may face unique challenges related to economic development, healthcare access and infrastructure maintenance due to their remote location and sparse population.

University No:_____

Answer:

-- END OF PAPER --