University No:_____

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

Date: December 10, 2021 Time: 9:30am-11:30am

INSTRUCTIONS:

a. This paper has three parts: Part A Multiple Choice Questions (25%), Part B Short Questions (40%) and Part C Long Question (35%)

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

University No:_____

Part A: Multiple Choice Questions (25%)

There are 25 questions in this part. For each question, exactly one of the options a, b, c, d is correct, and you will get one mark if you choose the correct option, lose one mark if you choose some incorrect one, and get or lose nothing if you do not choose any.

Give you answers in the answer boxes below. Put a cross at the boxes for your answers. For example, if your answer for Question 1 is a, then put a cross (x) at the first column (for Q1) and first row (for a). If you don't know the answer, you may leave the column blank.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
a										
b										
c										
d										

	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
a										
b										
c										
d										

	Q21	Q22	Q23	Q24	Q25
a					
b					
c					
d					

1. Which of the following operators could not be used on Boolean indexing?
   a. &
   b. |
   c. !
   d. None of above

2. Provided that we have the DataFrame df with the following data:

	Class	Year	Students
0	1	1	29
1	2	1	30
2	3	2	25
3	4	2	28
4	5	3	30

What will be displayed if the last statement of a Juypter notebook is
df.groupby(['Year'])['Students'].mean()

a. None

b.

```
Year
1    29.5
2    26.5
3    30.0
```

c.

Year	Class	Students
1	1.5	29.5
2	3.5	26.5
3	5.0	30.0

d.

	Class	Year	Students
0	1	1	29
1	2	1	30
2	3	2	25
3	4	2	28
4	5	3	30

3. When we consider "the average weight of the students in the University of Hong Kong", "The weights of all students in the University of Hong Kong" belongs to _____.

   a. Population
   b. Population parameter
   c. Confident interval
   d. None of above

4. Which of the following method generates descriptive statistics of numerical features of a DataFrame?

   a. plot.box
   b. mean
   c. median
   d. describe

5. Which of the following parameters can help to set the chart title?

   a. set_xlabel
   b. set_ylabel
   c. set_title
   d. title

6. Which of the following is NOT a function of Seaborn?
   a. catplot
   b. distplot
   c. relplot
   d. pyplot

7. Which of the following is NOT a common problem of a default plot?

   a. Figure too small
   b. Ticks not readable
   c. No title
   d. Too much information

8. Which of the following methods can be used for sorting two columns in descending order?

   a. sort_values(by=['Mark'])
   b. sort_values(by=['Mark', 'Name'])
   c. sort_values(by=['Mark'], ascending=False)
   d. sort_values(by=['Mark', 'Name'], ascending=False)

9. Which of the following pandas methods could help to change an original column name A to column name B?
   a. replace
   b. rename
   c. dropna
   d. None of above

10. Which of the following statements about matplotlib.pyplot is incorrect?
   a. We may set the title of a plot by label().
   b. We may plot one dimensional data.
   c. We may resize the plot area.
   d. We may use show() to make a plot appear.

11. Which of the following methods is used to show the gridlines of a plot in matplotlib.pyplot?
   a. legend
   b. plot
   c. marker
   d. grid

12. Which of the following statements about matplotlib.pyplot is correct?
   a. In scatter plot (scatter(x,y)), we can have two different size array as the x and y.
   b. We must set the width in a bar chart.
   c. We must use color instead of c to define color in a bar chart.
   d. We can plot the horizontal bars by plt.bar()

13. Which of the following python libraries is an interactive, open-source plotting library that supports many chart types covering a wide range of statistical, financial, geographic, scientific and 3-dimensional use-cases?
   a. pyplot
   b. geoplot
   c. plotly
   d. matplotlib

14. What is a correct syntax to use a conda env named "comp2501"?
   a. conda comp2501
   b. conda activate comp2501
   c. conda comp2501 activate
   d. activate comp2501

15. What is a correct syntax to install numpy using pip?
   a. pip install numpy
   b. pip installing numpy
   c. pip numpy
   d. pip install -r numpy

16. Which of the following methods is not supported by Pandas?
    a. read_csv()
    b. read_images()
    c. read_excel()
    d. read_table()

17. What is a correct syntax to check the data type an array arr?
    a. arr.dtype
    b. arr.datatype
    c. type(arr)
    d. arr.type()

18. What is a correct syntax to print the number 3 from the array below?

        arr = np.array([[1,2,3], [6,7,8]])

    a. print(arr[2,0])
    b. print(arr[0, 2])
    c. print(arr[0, 1])
    d. None of above

19. What is a correct syntax to print the number [6,7] from the array below?

        arr = np.array([[1,2,3], [6,7,8]])

    a. arr[1, :2]
    b. arr[1, :3]
    c. arr[:3, 0]
    d. arr[:2, 1]

20. What is a correct way to convert a numpy array data type?
    a. arr.convert_to(np.int8)
    b. arr.to(np.int8)
    c. arr.type(np.int8)
    d. arr.astype(np.int8)

21. When using the NumPy random module, how can you return a Normal Data Distrbution with 1000 numbers, concentrated around the number 50, with a standard deviation of 0.2?
    a. random.normal(size=1000, loc=50, scale=0.2)
    b. random.normal(size=1000, mean=50, deviation=0.2)
    c. random.normal(size=1000, normal=50, s=0.1)
    d. random.norm(size=1000, mean=50, deviation=0.2)

22. If a dimension is given as ____ in a reshaping operation, the other dimensions are automatically calculated.
    a. Zero
    b. One
    c. Negative one
    d. Infinite

23. How we can change the shape of the Numpy array in python?
    a. By Shape()
    b. By reshape()
    c. By ord()
    d. By change()

24. What is the output shape of A * B, whose shape are (15, 1, 3) and (1, 4, 1)
    a. (15, 4, 3)
    b. (15, 1, 3)
    c. (1, 4, 1)
    d. (15, 1, 3, 1, 4, 1)

25. Which of the following is a scraper library?
    a. Requests
    b. Beautiful Soup
    c. Numpy
    d. Matplotlib

University No:_____

Part B: Short Questions (40%)

Give you answers at the answer boxes following the questions. All questions carry equal marks (2 marks).

1. We flip a coin 10 times and get 3 heads and 7 tails. Our null hypothesis is
   "The coin is fair",
   i.e., the probability of getting a head is 0.5, and the probability of getting a tail is also 0.5. Determine whether we should reject the null hypothesis by considering its confidence interval with confidence level 95%.

   Answer

2. Repeat the above problem by considering the p-value of the observation instead of the confidence interval.

   Answer

3. Briefly explain the sampling technique for estimating the population mean, and how to use the central limit theorem to show the technique gives good estimate with high probability.

   Answer

University No:_____

4. Suppose that the content of the DataFrame answers is as follows:

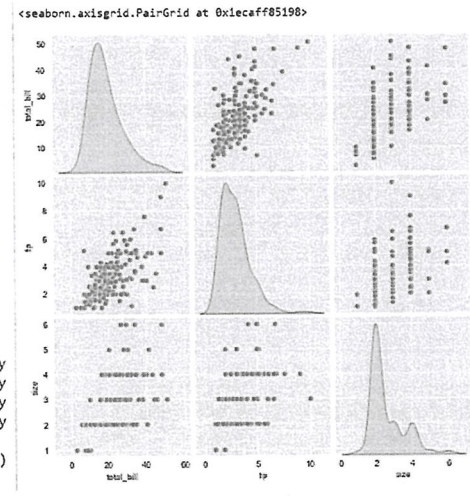
	Answers
0	a
1	a
2	b
3	c
4	d
5	c
6	c
7	e

Describe what the output of the statement "df.groupby(['Answer']).size()" would be.

Answer

5. Suppose the DataFrame df has the columns listed below. Given the Python snippet that generate the diagram on the right. Your snippet should include all the necessary libraries.

```
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   total_bill  244 non-null    float64
 1   tip         244 non-null    float64
 2   sex         244 non-null    category
 3   smoker      244 non-null    category
 4   day         244 non-null    category
 5   time        244 non-null    category
 6   size        244 non-null    int64
dtypes: category(4), float64(2), int64(1)
memory usage: 7.3 KB
```



Answer

6. Consider the following DataFrame df:

	a	b	c
0	1	2	3
1	4	5	6
2	7	8	3
3	1	5	3

Given the python statement that generate the following table

1	2
7	1
4	1
..	

Answer

7. Given the list S = [3, 34, 23, 10, 27, 42, 66, 34, 58, 150, 120, 76], what is the 5-number summary of this list?

Answer

8. For the list S in the previous question, does it have any outliers? If yes, what are they? Justify your answer.

Answer

9. Give the re expression 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

University No:_____

10. What is the output of the following program snippet:

```
1. import re
2. def text_match(text):
3. regexp1 = r"BookName:\s*[A-Z][a-z]*\s+"
4. regexp2 = r"([A-Z][a-z]*\s+)*"
5. regexp3 = r"[A-Z][a-z]*"
6. result = re.search(regexp1+regexp2+regexp3, text)
7. return result.group()

8. text = "BookName: A boy, AuthorName: Peter W Chan,
   UID:S12345"
9. print(text_match(text))
```

Answer

For Question 11 to Question 21:

In one of our lectures, we downloaded from the web site of Eurostat the file "eduData.csv", which contains the educational funding by the member states of the European Union. There are some entries in file whose values are the character ':', which indicates that the actual values are missing.

For each of the questions below, give a program snippet for a Jupyter notebook to carry out the task specified by that question.

11. Read the file "eduData.csv" and store its content in a dataframe with name edu. In edu, all the missing values should be filled with the Pandas' default missing value marker NaN. Your snippet should include all the necessary libraries. The following table shows a few rows in the dataframe edu.

	TIME	GEO	INDIC_ED	Value	Flag and Footnotes
0	2000	European Union (28 countries)	Total public expenditure on education as % of ...	NaN	NaN
1	2001	European Union (28 countries)	Total public expenditure on education as % of ...	NaN	NaN
2	2002	European Union (28 countries)	Total public expenditure on education as % of ...	5.00	e
3	2003	European Union (28 countries)	Total public expenditure on education as % of ...	5.03	e
4	2004	European Union (28 countries)	Total public expenditure on education as % of ...	4.95	e
...	...	...	...	...	...
379	2007	Finland	Total public expenditure on education as % of ...	5.90	NaN
380	2008	Finland	Total public expenditure on education as % of ...	6.10	NaN
381	2009	Finland	Total public expenditure on education as % of ...	6.81	NaN
382	2010	Finland	Total public expenditure on education as % of ...	6.85	NaN
383	2011	Finland	Total public expenditure on education as % of ...	6.76	NaN

Answer

University No:_____

12. Print the following information about edu.

	TIME	Value
count	384.000000	361.000000
mean	2005.500000	5.203989
std	3.456556	1.021694
min	2000.000000	2.880000
25%	2002.750000	4.620000
50%	2005.500000	5.060000
75%	2008.250000	5.660000
max	2011.000000	8.810000

Answer

13. Print the following information about edu:

	GEO	INDIC_ED	Flag and Footnotes
count	384	384	165
unique	32	1	6
top	Italy	Total public expenditure on education as % of ...	e
freq	12	384	70

Answer

14. The columns "INDIC_ED" and "Flag and Footnotes" are not relevant to our analysis. Remove these two columns from edu.

Answer

15. For each column with numeric values, replace any NaN value in that column by the average of the non-NaN values in that column. Then, remove all the rows that have NaN in any of its non-numeric columns.

Answer

16. Print the 'Value" column and the 'GEO' column of edu from row 10 to row 13.

Answer

17. Print the countries in 'GEO' whose average 'Value' is greater than 5.5.

Answer

18. Add the column "NValue" to edu whose values are obtained by normalizing the values in the 'Value" column, i.e., if the row has 'Value' = x, then its 'NValue' is equal to (x-minvalue)/(maxvalue – minvalue) where maxvalue and minvalue are respectively the maximum and minimum values found in 'Value'.

Answer

19. Print a sorted version of edu in which it is sorted in ascending order of 'GEO' and for each country in 'GEO', it is sorted in descending order of 'Value'.

Answer

University No:_____

20. Generate a table in which its rows are labeled by the countries in 'GEO', its columns labeled by the years in 'TIME', and the entry at row A and column B and the 'Value' of country A and year B in edu. For example, in the following table, the first entry has value 5.66 because in edu, the row with 'GEO'=Austria and 'TIME'=2000 has the value 5.66 in the 'Value' column.

TIME	2000	2001	2002	2003	2004	2005	2006	2007	2008
GEO									
Austria	5.66	5.74	5.68	5.53	5.48	5.44	5.40	5.33	5.47
Belgium	NaN	5.99	6.09	6.02	5.95	5.92	5.98	6.00	6.43
Bulgaria	3.88	3.70	3.94	4.09	4.40	4.25	4.04	3.88	4.44
Cyprus	5.42	5.98	6.60	7.37	6.77	6.95	7.02	6.95	7.45
:ech Republic	3.83	3.93	4.15	4.32	4.20	4.08	4.42	4.05	3.92

Answer

University No:_____

Part C: Long Question (35%)

Write a Python program that inputs the name of a stock, and its current price, and then predicts and prints its price on the next day as follows: using Yahoo Finance's history() method to get the daily prices of the stock from 2010 Jan 2 to 2020 Dec 31, and uses them to train a ML model (you may choose any model you like), and then use it to make the prediction. You may assume that your program has a predefined dictionary ticker_list so that ticket_list[x] will give you the ticker for the stock x. Remember to include all the necessary libraries.

Answer

--END OF PAPER--