BENNETT UNIVERSITY
THE TIMES GROUP

Enrolment No: _____ Name of Student: _____

Department/ School: _____

END TERM EXAMINATION ODD SEMESTER 2022-23

COURSE CODE – CMCA523 **MAX. DURATION: 2 HRS**
COURSE TITLE-Data Analysis using Python
COURSE CREDIT 2-0-4 **TOTAL MARKS: 35**

GENERAL INSTRUCTIONS: -

1. Do not write anything on the question paper except **name, enrolment number** and **department/school.**
2. Carrying mobile phone, smart watch and any other non-permissible materials in the examination hall is an act of **UFM.**

SECTION A

Max Marks: 35 Marks

1. Suppose you have a dataframe 'df'. You want to create groups based on the column key in the DataFrame and fill the nan values with group means using:
filling_mean = lambda g: g.fillna(g.mean()) **1 Mark**

   a. df.groupby(key).aggregate(filling_mean)
   b. df.groupby(key).filling_mean()
   c. df.groupby(key).transorm(filling_mean)
   d. df.groupby(key).apply(filling_mean)

2. Which of the following is not a valid expression to create a Pandas groupBy object from the dataframe shown below? **1 Mark**

name	Class	Breadth
Sejal	Fruit	4.24
Arjun	Fruit	2.67
Aman	Vegitable	7.60
Himanshu	Vegitable	7.10
Deepak	Vegitable	4.90

(i) grouped= df.groupby(['class','avg calories per unit']
(ii) df.groupby('Sejal')
(iii) df.groupby('class')
(iv) df.groupby('class',axis=0)

3. What will be the output of the following code? **1 Mark**

```
import pandas as pd
df1 = pd.DataFrame({'a': ['foo', 'bar'], 'b': [1, 2]})
df2 = pd.DataFrame({'a': ['foo', 'baz'], 'c': [3, 4]})
df1.merge(df2, how='inner', on='a')
```

(i)

	a	b	c
0	foo	1.0	NaN
1	bar	2.0	NaN
0	foo	NaN	3.0
1	baz	NaN	4.0

(ii)

	a	b	c
0	foo	1.0	3.0
1	bar	2.0	NaN
2	baz	NaN	4.0

(iii)

	a	b	c
0	foo	1	3

(iv) None

4. If you reject a true null hypothesis, what does this mean? **1 Mark**
(i) You have made a Type-I Error
(ii) You have made a Type-II Error
(iii) You have made a correct decision
(iv) You have increased the power of a test.

5. We have two dataframes df1 and df2 which are defined as **1 Mark**

```
import pandas as pd
data1 = {"name": ["Sally", "Mary", "John"],"age": [50, 40, 30]}
data2 = {"name": ["Sally", "Peter", "Micky"],"age": [77, 44, 22]}
df1 = pd.DataFrame(data1)
df2 = pd.DataFrame(data2)
```

Then the output of newdf = df1.merge(df2, how='right') will be
(i) Empty dataframe
(ii) None

(iii)

	name	age
0	Sally	50
1	Mary	40
2	John	30

(iv)

	name	age
0	Sally	77
1	Peter	44
2	Micky	22

6. What will be the value of c after the following code? **2 Marks**

```
import numpy as np
a = np.arange(8)
b= a[4:6]
b[:] = 40
c = a[3] + a[5]
```

7. What will be the output of the following code? **2 Marks**

```
import re
p=re.compile('Data')
r=p.match('Data Science using Python')
print(r.group(0))
```

8. What will be the output of the following code? **2 Marks**

```
import re
pattern=re.compile('CRICKET',re.I)
match=pattern.search("I watch cricket regularly")
print("Start index:", match.start())
print("End index:", match.end())
print("Tuple:", match.span())
```

9. Consider the following code. What will be the value of num? **2 Marks**

```
import re
phone='+91-2333876589 My phone number'
num=re.sub(r'\D','',phone)
num
```

10. Predict the output of the following code. **2 Marks**

```
import re
string='PQPPPPPRPPPD'
res=re.findall('P{1,2}',string)
print(len(res))
```

11. Read the dataset taxis.csv and store it in a dataframe 'df'. Use groupby function to the 'df' based on 'color, payment' and store it in another dataframe. The new dataframe should be indexed on two columns as given above and should contain the number of passengers, distance, fare, tip, tolls, total in each colour. **3 Marks**

Expected output:

		passengers	distance	fare	tip	tolls	total
payment	color						
cash	green	7	24.02	89.50	0.00	0	99.05
	yellow	1	0.79	5.00	0.00	0	9.30
credit card	green	2	11.47	35.53	1.00	0	37.83
	yellow	8	16.97	71.00	15.92	0	109.22

Using the original dataframe perform the group operation on each payment type and show the median, count, sum, and mean information of fare for each payment type and colour. **1 Mark**

Expected output:

		fare			
		median	count	sum	mean
payment	color				
cash	green	10.750	6	89.50	14.916667
	yellow	5.000	1	5.00	5.000000
credit card	green	17.765	2	35.53	17.765000
	yellow	8.250	6	71.00	11.833333

Now create another column which will contain the fare standardized color using the formula (x-mean(x)/std(x)). **1 Mark**

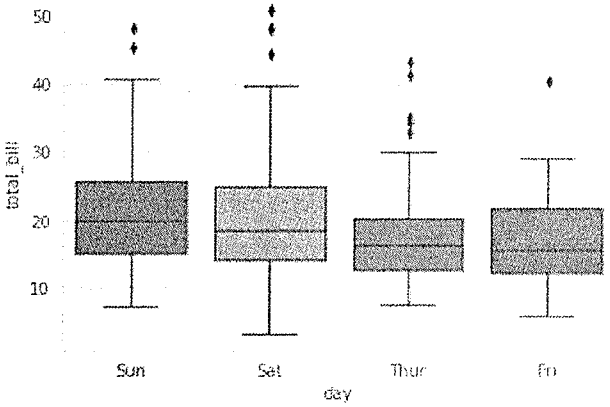
Expected output:

	fare	normalized
color		
green	125.03	0.000000e+00
yellow	76.00	-3.330669e-16

The dataset is available on the last page Appendix-A for your reference.

12. Import the tips.csv dataset. Help the owner to know the boxplot of *days* with respect *total_bill*. You can consider seaborn library with 'whitegrid' style. **5 Marks**

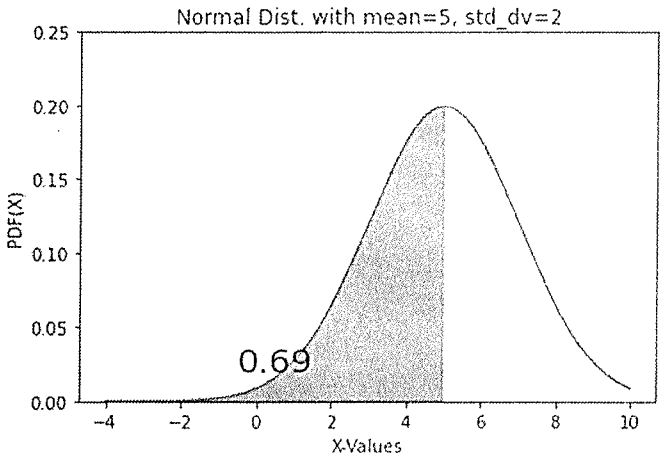
Expected Output:



The dataset is available on the last page Appendix-B for your reference

13. Find the P(X<6) using the norm.cdf() in scipy with mean=5 and s.d.=2. Generate population using np.arange function from -4 to 10 with a step 0.001. Use mean of 5 and s.d. =2 for plotting the normal distribution. Use the fill_between() to fill the region with green colour according to the condition P(X<6). **5 Marks**

Expected Output:



14. List of final marks of 14 students are given as: 183, 152, 178, 157, 194, 163, 144, 114, 178, 152, 118, 158, 172, 138. Perform t-test on the student data, whether the population mean, is less than 165 Hypothesis **5 Marks**

H0: There is no significant mean difference in marks. i.e., $\mu = 165$

H1: The population mean is less than 165. i.e., $\mu < 165$.

Appendix-A

The dataset taxis.csv is given below for reference:

pickup	dropoff	passengers	distance	fare	tip	tolls	total	color	payment	pickup_zone	dropoff_zone	pickup_borc	dropoff_borough
23-03-2019 20:21	23-03-2019 20:27	1	1.6	7	2.15	0	12.95	yellow	credit card	Lenox Hill West	UN/Turtle Bay S	Manhattan	Manhattan
04-03-2019 16:11	04-03-2019 16:19	1	0.79	5	0	0	9.3	yellow	cash	Upper West Side	Upper West Side	Manhattan	Manhattan
27-03-2019 17:53	27-03-2019 18:00	1	1.37	7.5	2.36	0	14.16	yellow	credit card	Alphabet City	West Village	Manhattan	Manhattan
10-03-2019 01:23	10-03-2019 01:49	1	7.7	27	6.15	0	36.95	yellow	credit card	Hudson Sq	Yorkville West	Manhattan	Manhattan
30-03-2019 13:27	30-03-2019 13:37	3	2.16	9	1.1	0	13.4	yellow	credit card	Midtown East	Yorkville West	Manhattan	Manhattan
11-03-2019 10:37	11-03-2019 10:47	1	0.49	7.5	2.16	0	12.96	yellow	credit card	Times Sq/Theatre	Midtown East	Manhattan	Manhattan
26-03-2019 21:07	26-03-2019 21:17	1	3.65	13	2	0	18.8	yellow	credit card	Battery Park City	Two Bridges/Se	Manhattan	Manhattan
08-03-2019 14:25	08-03-2019 15:04	1	9.84	32.5	0	0	36.05	green	cash	East Harlem Sout	Downtown Brook	Manhattan	Brooklyn
27-03-2019 10:02	27-03-2019 10:25	1	8.74	23.5	0	0	24.03	green	credit card	Woodlawn/Wakef	Hunts Point	Bronx	Bronx
02-03-2019 17:48	02-03-2019 18:10	1	8.97	26	0	0	26.8	green	cash	Hamilton Heights	Bronxdale	Manhattan	Bronx
08-03-2019 01:02	08-03-2019 01:07	1	1.05	6	0	0	7.3	green	cash	Central Harlem	Manhattanville	Manhattan	Manhattan
13-03-2019 12:29	13-03-2019 12:42	1	1.4	9.5	0	0	10.3	green	cash	Brooklyn Heights	Brooklyn Navy Y	Brooklyn	Brooklyn
06-03-2019 18:27	06-03-2019 18:43	2	2.64	12	0	0	13.8	green	cash	Downtown Brookl	Red Hook	Brooklyn	Brooklyn
03-03-2019 09:30	03-03-2019 09:43	1	2.73	12	1	0	13.8	green	credit card	East Tremont	Van Cortlandt V	Bronx	Bronx
04-03-2019 04:19	04-03-2019 04:21	1	0.12	3.5	0	0	4.8	green	cash	Hamilton Heights	Hamilton Height	Manhattan	Manhattan

Appendix-B

The dataset tips.csv is given below for reference:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns