WORKSHEET – Data Handling Using Pandas

1	**What will be the output of following code-** **import pandas as pd** **s1=pd.Series([1,2,2,7,'Sachin',77.5])** **print(s1.head())** **print(s1.head(3))** Ans: 0 1 1 2 2 2 3 7 4 Sachin dtype: object 0 1 1 2 2 2 dtype: object

2	**Write a program in python to find maximum value over index in Data frame.**
	
	Ans:
	# importing pandas as pd
	import pandas as pd
	
	# Creating the dataframe
	df = pd.DataFrame({"A":[4, 5, 2, 6],
	"B":[11, 2, 5, 8],
	"C":[1, 8, 66, 4]})
	
	# Print the dataframe
	df
	# applying idxmax() function.
	df.idxmax(axis = 0)

3	**What are the purpose of following statements-** **1. df.columns** **2. df.iloc[: , :-5]** **3. df[2:8]** **4. df[:]** **5. df.iloc[: -4 , :]** Ans: 1. It displays the names of columns of the Dataframe. 2. It will display all columns except the last 5 columns.

3. It displays all columns with row index 2 to 7.
4. It will display entire dataframe with all rows and columns.
5. It will display all rows except the last 4 four rows.

4 **Write a python program to sort the following data according to ascending order of Age.**

Name	Age	Designation
Sanjeev	37	Manager
Keshav	42	Clerk
Rahul	38	Accountant

Ans:

```
import pandas as pd
name=pd.Series(['Sanjeev','Keshav','Rahul'])
age=pd.Series([37,42,38])
designation=pd.Series(['Manager','Clerk','Accountant'])
d1={'Name':name,'Age':age,'Designation':designation}
df=pd.DataFrame(d1)
print(df)
df1=df.sort_values(by='Age')
print(df1)
```

5 **Write a python program to sort the following data according to descending order of Name.**

Name	Age	Designation
Sanjeev	37	Manager
Keshav	42	Clerk
Rahul	38	Accountant

Ans:

```
import pandas as pd
name=pd.Series(['Sanjeev','Keshav','Rahul'])
age=pd.Series([37,42,38])
designation=pd.Series(['Manager','Clerk','Accountant'])
d1={'Name':name,'Age':age,'Designation':designation}
df=pd.DataFrame(d1)
print(df)
```

	df2=df.sort_values(by='Name',ascending=0) print(df2)
6	**Which of the following thing can be data in Pandas?** 1. A python dictionary 2. An nd array 3. A scalar value 4. All of above Ans: **5. All the above**
7	**All pandas data structure are _____mutable but not always _____mutable.** 1. Size, value 2. Semantic , size 3. Value, size 4. None of the above Ans: 3. Value,size
8	**Data and index in an nd array must be of same length-** **1.** True **2.** False Ans: **1. True**
9 **3.**	**What is the output of the following program?** import pandas as pd df=pd.DataFrame(index=[0,1,2,3,4,5],columns=['one','two']) print df['one'].sum() Ans: It will produce an error.
10	**What will be the output of following code:** Users.groupby('occupation').age.mean() 1. Get mean age of occupation 2. Groups users by mean age 3. Groups user by age and occupation 4. None Ans: 1. Get mean age of occupation
11	**Which object do you get after reading a CSV file using pandas.read_csv()?** 1. Dataframe 2. Nd array 3. Char Vector

4. None

Ans:

1. Dataframe

12 **What will be the output of df.iloc[3:7,3:6]?**

Ans:
It will display the rows with index 3 to 6 and columns with index 3 to 5 in a dataframe 'df'

13 **How to select the rows where where age is missing?**

1. df[df['age'].isnull]
2. df[df['age']==NaN]
3. df[df['age']==0]
4. None

Ans:

4. None As the right answer is **df[df['age'].isnull()]**

14 **Consider the following record in dataframe IPL**

Player	Team	Category	BidPrice	Runs
Hardik Pandya	Mumbai Indians	Batsman	13	1000
KL Rahul	Kings Eleven	Batsman	12	2400
Andre Russel	Kolkata Knight riders	Batsman	7	900
Jasprit Bumrah	Mumbai Indians	Bowler	10	200
Virat Kohli	RCB	Batsman	17	3600
Rohit Sharma	Mumbai Indians	Batsman	15	3700

Retrieve first 2 and last 3 rows using python program.

Ans:

```
d={'Player':['Hardik Pandya','K L Rahul','AndreRussel','Jasprit Bumrah','Virat Kohli','Rohit Sharma'],

  'Team':['Mumbai Indians','Kings Eleven','Kolkata Knight Riders','Mumbai Indians','RCB','Mumbai Indians'],
'Category':['Batsman','Batsman','Batsman','Bowler','Batsman','Batsman'] ,

'Bidprice':[13,12,7,10,17,15],
'Runs':[1000,2400,900,200,3600,3700]}
df=pd.DataFrame(d)
print(df)
print(df.iloc[:2,:])
print(df.iloc[-3:,:])
```

15 **Write a command to Find most expensive Player.**

Ans:

```
print(df[df['BidPrice']==df['BidPrice'].max()])
```

16 **Write a command to Print total players per team.**

	Ans: print(df.groupby('Team').Player.count())
17	**Write a command to Find player who had highest BidPrice from each team.** **Ans:** val=df.groupby('Team') print(val['Player','BidPrice'].max())
18	**Write a command to Find average runs of each team.** **Ans:** print(df.groupby(['Team']).Runs.mean())
19	**Write a command to Sort all players according to BidPrice.** **Ans:** print(df.sort_values(by='BidPrice'))
20	**We need to define an index in pandas-** **1.** True **2.** False Ans: 2 False
21	**Who is data scientist?** 1. Mathematician 2. Statistician 3. Software Programmer 4. All of the above Ans: 4 All the above
22	**What is the built-in database used for python?** 1. Mysql 2. Pysqlite 3. Sqlite3 4. Pysqln Ans: 3 Sqlite3
23	**How can you drop columns in python that contain NaN?** Ans: df1.dropna(axis=1)

24	**How can you drop all rows that contains NaN?** Ans: df1.dropna(axis=0)
25	**A Series is _____ array, which is labelled and _____ type.** Ans: One dimensional array, homogeneous
26	**Minimum number of arguments we require to pass in pandas series –** 1. 0 2. 1 3. 2 4. 3 Ans: 1. 0
27	**What we pass in data frame in pandas?** **1.** Integer **2.** String **3.** Pandas series **4.** All Ans: 4 All
28	**How many rows the resultant data frame will have?** import pandas as pd df1=pd.DataFrame({'key':['a','b','c','d'], 'value':[1,2,3,4]}) df2=pd.DataFrame({'key':['a','b','e','b'], 'value':[5,6,7,8]}) df3=df1.merge(df2, on='key', how='outer') 1. 5 2. 4 3. 2 4. 6 Ans: 4.6
29	**How many rows the resultant data frame will have?** import pandas as pd df1=pd.DataFrame({'key':['a','b','c','d'], 'value':[1,2,3,4]}) df2=pd.DataFrame({'key':['a','b','e','b'], 'value':[5,6,7,8]}) df3=df1.merge(df2, on='key', how='inner') 1. 3 2. 4 3. 5 4. 6 Ans: 1. 3
30	**How many rows the resultant data frame will have?**

	import pandas as pd df1=pd.DataFrame({'key':['a','b','c','d'], 'value':[1,2,3,4]}) df2=pd.DataFrame({'key':['a','b','e','b'], 'value':[5,6,7,8]}) df3=df1.merge(df2, on='key', how='right') 1. 3 2. 4 3. 5 4. 6 Ans: **2. 4**
31	**How many rows the resultant data frame will have?** import pandas as pd df1=pd.DataFrame({'key':['a','b','c','d'], 'value':[1,2,3,4]}) df2=pd.DataFrame({'key':['a','b','e','b'], 'value':[5,6,7,8]}) df3=df1.merge(df2, on='key', how='left') 1. 3 2. 4 3. 5 4. 6 Ans: 3. 5
32	**_____method is used to delete the series and also return the series as a result.** **Ans:** pop()
33	**A _____ is an interactive way to quickly summarize large amount of data.** **Ans:** Pivoting
34	**_____Method is used to rename the existing indexes in a data frame.** **Ans:** rename
35	**_____ Attribute that can prohibit to create a new data frame in sort_values() method.** **Ans:** Inplace
36	**Write a program in python to calculate the sum of marks in CS subject in a given dataset-** 'CS':[45,55,78,95,99,97], 'IP':[87,89,98,94,78,77] **Ans:** d1={ 'CS':[45,55,78,95,99,97], 'IP':[87,89,98,94,78,77] } df=pd.DataFrame(d1) print(df['CS'].sum())

37 **Write a python program to create a data frame with headings (CS and IP) from the list given below-**
[[79,92][86,96],[85,91],[80,99]]

Ans:

```
l=[[10,20],[20,30],[30,40]]
df=pd.DataFrame(l,columns=['CS','IP'])
print(df)
```

38 **How you can find the total number of rows and columns in a data frame.**

Ans:

df.shape

39

MaxTemp	MinTemp	City	RainFall
45	30	Delhi	25.6
34	24	Guwahati	41.5
48	34	Chennai	36.8
32	22	Bangluru	40.2
44	29	Mumbai	38.5
39	37	Jaipur	24.9

Consider the above data frame as df-

1. **Write command to compute sum of every column of the data frame.**

Ans:
print(df.sum(axis=0))

40 **Based on the above data frame df, Write a command to compute mean of column MaxTemp.**

Ans:
Print(df['MaxTemp'].mean())

41 **Based on the above data frame df, Write a command to compute average MinTemp, RainFall for first 4 rows.**

Ans:
df[['MinTemp', 'Rainfall']][:4].mean()

42 **Which method is used to read the data from MySQL database through Data Frame?**

Ans:
read_sql_query()

43 **Which method is used to perform a query in MySQL through Data Frame?**

Ans:

execute()

44 **What will be the output of following code?**

	import pandas as pd df = pd.DataFrame([45,50,41,56], index = [True, False, True, False]) print(df.iloc[True]) **Ans:** It will display error message like- Cannot index by location index with a non-integer key **because iloc accept only integer index.**
45	**Write a program in python to join two data frame.** **Ans:** xiia={'sub':['eng','mat','ip','phy','che','bio'],'id':['302','041','065','042','043','044']} xiic={'sub':['eng','mat','ip','acc','bst','eco'],'id':['302','041','065','055','056','057']} df1=pd.DataFrame(xiia) print(df1) df2=pd.DataFrame(xiic) print(df2) print(df1.merge(df2,on='id')) print(df1.merge(df2,on='id',how='outer'))
46.	**What is a Series? Explain with the help of an example.**
	Pandas Series is a one-dimensional labeled array capable of holding data of any type (integer, string, float, python objects etc.). The axis labels are collectively called index. e.g. import pandas as pd data =pd.Series([1,2,3,4,5]) print(datAns:
47.	**Hitesh wants to display the last four rows of the dataframe df and has written the following code:** **df.tail()** **But last 5 rows are being displayed. Identify the error and rewrite the correct code so that last 4 rows get displayed.**
	If tail() doesn't receive any argument, then by default last 5 rows will be displayed. Correct Code is: df.tail(4)
48.	**Write the command to add a new column in the last place(3rd place) named "Salary" from the list of values, Sal=[10000,15000,20000] in an existing dataframe named EMP, assume already having 2 columns.**
	EMP['Salary']=Sal
49.	**Consider the following python code and write the output:** **import pandas as pd** **K=pd.series([2,4,6,8,10,12,14])** **print(K.quantile([0.50,0.75]))**
	0.50 8.0 0.75 11.0
50.	**Write a small python code to drop a row from dataframe labeled as 0.**
	df=df.drop(0)
51.	**What is Pivoting? Name any two functions of Pandas which support pivoting.**
	Pivoting is a technique to quickly summarize large amount of data so that data can be viewed in a different perspective. Pivot table in pivoting can be used to apply aggregate function like-count.

	Two functions for pivoting are: pivot() and pivot_table()
52.	**Write a python code to create a dataframe with appropriate headings from the list given below:** **['S101', 'Amy', 70], ['S102', 'Risha', 69], ['S104', 'Susan', 75], ['S105','George', 82]**
	import pandas as pd L=[['S101','Amy',70], ['S102','Risha',69], ['S104','Susan',75], ['S105','George',82]] df=pd.DataFrame(L,index=[1,2,3,4],columns=['ID','Name','Points']) print(df)
53.	**Consider the following dataframe, and answer the questions given below:** **import pandas as pd** **df = pd.DataFrame({"Quarter1":[2000, 4000, 5000, 4400, 10000],** **"Quarter2":[5800, 2500, 5400, 3000, 2900],** **"Quarter3":[20000, 16000, 7000, 3600, 8200],** **"Quarter4":[1400, 3700, 1700, 2000, 6000]})** **Write the code to find mean value from above dataframe df over the index and column axis. (Skip NaN value)**
	print(df.mean(axis=0,skipna=True)) print(df.mean(axis=1,skipna=True))
54.	**Use sum() function to find the sum of all the values over the index axis.**
	print(df.sum(axis=0))
55.	**Find the median of the dataframe df.**
	print(df.median())
56.	**Find the output of the following code:** **import pandas as pd** **data = [{'a': 10, 'b': 20},{'a': 6, 'b': 32, 'c': 22}]** **df1 = pd.DataFrame(data,columns=['a','b'])** **df2 = pd.DataFrame(data,columns=['a','b1'])** **print(df1)** **print(df2)**
	a b 0 10 20 1 6 32 a b1 0 10 NaN 1 6 NaN
57.	**Write the code in pandas to create the following dataframes:** df1 df2 mark1 mark2 mark1 mark2 0 10 150 30 20 1 40 451 20 25 2 15 302 20 30 3 40 703 50 30
	import pandas as pd x1=[[10,150],[40,451],[15,302],[40,703]] df1=pd.DataFrame(x1,columns=['mark1','mark2']) x2=[[30,20],[20,25],[20,30],[5,30]] df2=pd.DataFrame(x2,columns=['mark1','mark2 ']) print(df1) print(df2)

58.	**To add dataframes df1 and df2.**
	print(df1.add(df2))
59.	**To subtract df2 from df1**
	print(df1.sub(df2))
60.	**To change index label of df1 from 0 to zero and from 1 to one.**
	df1=df1.rename(index={0:'zero',1:'one'})
61.	**What will be the output of the following python code?** **import pandas as pd** **d={'Student':['Ali','Ali','Tom','Tom'],** **'House':['Red','Red','Blue','Blue'],** **'Points':[50,70,60,80]}** **df =pd.DataFrame(d)** **df1=df.pivot_table(index='Student',columns='House',values='Points',aggfunc='sum')** **print(df1)**
	House Blue Red Student Ali NaN 120.0 Tom 140.0 NaN
62.	**For the given code fill in the blanks so that we get the desired output with maximum value for Quantity and Average Value for Cost:** **import pandas as pd** **import numpy as np** **d={'Product':['Apple','Pear','Banana','Grapes'],'Quantity':[100,150,200,250],** **'Cost':[1000,1500,1200,900]}** **df = pd.DataFrame(d)** **df1 =_____** **print(df1)** **Quantity 250.0** **Cost 1150.0** **dtype: float64**
	df1=pd.DataFrame([df['Quantity'].max(),df['Cost'].mean()],index=['Quantity','Cost'])
63.	**Find Output for the following program code:**

```
import pandas as pd
df1=pd.DataFrame({'Icecream':['Vanila','ButterScotch','Caramel'] ,
'Cookies':['Goodday','Britannia', 'Oreo']})
df2=pd.DataFrame({'Chocolate':['DairyMilk','Kitkat'],'Icecream':['Vanila','ButterScotc
h'],'Cookies':['Hide and Seek','Britannia'}
df2.reindex_like(df1)
print(df2)
```

```
     Chocolate      Icecream          Cookies
0    DairyMilk        Vanila    Hide and Seek
1       Kitkat   ButterScotch        Britannia
```

64. **A dictionary Smarks contains the following data:**
Smarks={'name':['rashmi','harsh','priya'],'grade':['A1','A2','1']}
Write a statement to create DataFrame called df.
Assume that pandas has been imported as pd.

```
df=pd.DataFrame(Smarks,index=[1,2,3])
```

65. **In pandas, S is a series with the following result:**
S=pd.Series([5,10,15,20,25])
The series object is automatically indexed as 0,1,2,3,4. Write a statement to assign the series as a, b, c, d,e index explicitly.

```
S=pd.Series([5,10,15,20,25],index=['a','b','c','d','e'])
```

66. **Write python statement to delete the 3rd and 5th rows from dataframe df.**

```
df1=df.drop(index=[2,4],axis=0)
        or
df1=df.drop([2,4])
```

67. **Given the two dataframes df1 and df2 as given below:**

df1

	First	Second	Third
0	10	4	30
1	20	5	40
2	30	7	50
3	40	9	70

df2

	First	Second	Third
0	17	14	13
1	18	15	14
2	19	17	15
3	20	19	17

Write the commands to do the following on the dataframe:
To add dataframes df1 and df2.

```
print(df1.add(df2))
```

68.	**To sort df1 by Second column in descending order.**
	df1=df1.sort_values(by='Second',ascending=False)
69.	**To change the index of df2 from 0,1,2,3 to a,b,c,d**
	df2=df2.rename(index={0:'a',1:'b',2:'c',3:'d'})
70.	**To display those rows in df1 where value of third column is more than 45.**
	print(df1[df1['Third']>45])
71.	**Consider the following dataframe: student_df** **Name class marks** **Anamay XI 95** **Aditi XI 82** **Mehak XI 65** **Kriti XI 45** **Write a statement to get the minimum value of the column marks**
	print(student_df['Marks'].min())
72.	**Write a small python code to add a row to a dataframe.**
	import pandas as pd student_df=pd.DataFrame({'Name':['Ananmay','Aditi','Mehak','Kriti'],'Class':['XI','XI','XI','XI'],'Marks':[95,82,65,45]},index=[1,2,3,4]) data={'Name':'Sohail','Class':'XII','Marks':77} newstd=pd.DataFrame(data,index=[5]) student_df=student_df.append(newstd)
73.	**Jitesh wants to sort a DataFrame df. He has written the following code.** **df=pd.DataFrame({"a":[13, 24, 43, 4],"b":[51, 26, 37, 48]})** **print(df)** **df.sort_values('a')** **print(df)** **He is getting an output which is showing original DataFrame and not the sorted DataFrame. Identify the error and suggest the correction so that the sorted DataFrame is printed.**
	The possible reason is that the original dataframe is not modified. The correct answer is: df.sort_values('a',inplace=True)
74.	**Write a command to display the name of the company and the highest car price from DataFrame having data about cars.**
	import pandas as pd car={'Name':['Innova','Tavera','Royal','Scorpio'],'Price':[300000,800000,250000,650000]} df=pd.DataFrame(car,index=[1,2,3,4]) print(df[df.Price==df.Price.max()])
75.	**Write a command in python to Print the total number of records in the DataFrame.**
	print(df1.count())

76.	**Consider a DataFrame 'df' created using the dictionary given below, answer the questions given below:** **exam_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily',** **'Michael','Matthew', 'Lara', 'Kevin', 'Jonas'],** **'score': [12.5, 9, 16.5, np.NaN, 9, 20,14.5, np.NaN, 8, 19],** **'attempts' : [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],** **qualify': ['yes','no','yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}** **Write command to remove the rows having NaN values.**
	df=df.dropna()
77.	**Write a command to create a pivot table based on 'qualify' column and display sum of the score and attempt columns.**
	print(df.pivot_table(columns=['qualify'],values=['score','attempts'],aggfunc='sum'))
78.	**Write a command to display the names of students who have qualified.**
	print(df[df['qualify']=='yes'].name)
79.	**Consider the following DataFrame df and answer the questions given below:** **Write command to change the indices to 'zero','one','two',three' and 'four' respectively.**
	df=df.rename(index={0:'Zero',1:'One',2:'Two',3:'Three'})
80.	**Write command to compute mean of every column of the data frame.**
	print(df.mean(axis=0))
81.	**Write command to add one more row to the data frame with data [5,12,33,3]**
	df2 = {'col1': 5, 'col2': 12, 'col3': 33, 'col4':3} df = df.append(df2, ignore_index=True)

	A	B	C	D
0	12	5	20	14
1	4	2	16	3
2	5	54	7	17
3	44	3	3	2
4	1	2	8	6

82.

Emp_ID	Name	Dept	Salary	Status
100	Kabir	IT	34000	Regular
110	Rishav	Finance	28500	Regular
120	Seema	IT	13500	Contract
130	David	IT	41000	Regular
140	Ruchi	HRD	17000	Contract

Consider the above Data frame as df.
Write a Python Code to calculate the average salary of the Regular employees and the Contract employees separately.

```
print(df.groupby('Status').mean().Salary)
```

83. **Write a Python Code to print the dataframe in the descending order of Salary.**

```
df=df.sort_values(by='Salary',ascending=False)
print(df)
```

84. **Write a Python Code to update the Salary of all Contract employees to Rs 19000**

```
df.Salary[df.Status=='Contract']=19000
```

85. **Write a Python Code to count the total number of employees in each department.**

```
print(df.groupby('Dept').count().Name)
```

86. **Write a Python Code to display the maximum salary of the "Contract" staff.**

```
print(df[df['Status']=='Contract'].max().Salary)
```

87. **Write a Python Code to display the 4^{tn} Record.**

```
print(df.iloc[3:4,:])
```

88. **Write a Python Code to delete the column Status.**

```
del df['Status']
```

89. **Write a Python Code to display the maximum salary of all employees in the 'IT' department.**

```
print(df[df.Dept=='IT'].max().Salary)
```

90. **Write a Python Code to delete the 1st and the last record.**

```
df=df.drop([0,4])
```

91. **Consider a dataframe as follows:**

	A	B	C
1	56	71	-13
2	-29	-63	34
3	83	-60	71

Write a Python Code to : Replace all negative numbers with 0

```
df[df<0]=0
```

92. **Count the number of elements which are greater than 50**

print(df[df>50].count().sum())

93. **Write Python Code to count the number of even numbers and number of odd numbers in the dataframe.**

```
print('No of Even Numbers:',df[df%2==0].count().sum())
print('No of Odd Numbers:',df[df%2==1].count().sum())
```

94. **Consider the above data frame df.**

employee	sales	Quarter	State
Sahay	125600	1	Delhi
George	235600	1	Tamil Nadu
Priya	213400	1	Kerala
Manila	189000	1	Haryana
Raina	456000	1	West Bengal
Manila	172000	2	Haryana
Priya	201400	2	Kerala

Write Python Program to create the above dataframe.

```
import pandas as pd
data={'employee':['Sahay','George','Priya','Manila','Raina','Manila','Priya'],
'Sales':[125600,235600,213400,189000,456000,172000,201400],
'Quarter':[1,1,1,1,1,2,2],'State':['Delhi','TamilNadu','Kerala','Haryana','West
Bengal','Haryana','Kerala']}
 df=pd.DataFrame(data)
print(df)
```

95. **Write Python Program to find total sales per state.**

```
print(df.groupby('State').sum().Sales)
```

96. **Write Python Program to find total sales per employee.**

```
print(df.groupby('employee').sum().Sales)
```

97. **Write Python Program to find average sales on both employee and state wise.**

```
print(df.groupby(['employee','State']).sum().Sales)
```

98. **Write Python Program to find mean,median and minimum sale statewise.**

```
print(df.groupby('State').mean().Sales)
print(df.groupby('State').median().Sales)
print(df.groupby('State').min().Sales)
```

99. **Write Python Program to find maximum sales quarter-wise.**

```
print(df.groupby('Quarter').max().Sales)
```

100. **Write Python Program to create a Pivot Table with State as the index, Sales as the values and calculating the maximum Sales in each State.**

```
print(df.pivot_table(index='State',values='Sales',aggfunc='max'))
```