PYTHON LAB – 27 PANDAS IO AND CLEANING DATA

NAME: Keerthana K R

ID: AF0363623

QUESTIONS

1. Write a Pandas program to detect missing values of a given DataFrame.
2. Write a Pandas program to drop the rows where at least one element is missing in a given DataFrame.
3. Write a Pandas program to drop the rows where all elements are missing in a given DataFrame.
4. Write a Pandas program to drop those rows from a given DataFrame in which specific columns have missing values.

1. Write a Pandas program to detect missing values of a given DataFrame.
Input: df = pd.DataFrame({ 'ord_no':[70001,np.nan,70002,70004,np.nan,70005,np.nan,70010,70003,70012,np.na n,70013], 'purch_amt' : [150.5, 270.65, 65.26, 110.5, 948.5, 2400.6, 5760, 1983.43, 2480.4, 250.45, 75.29, 3045.6], 'ord_date': ['2012-10-05','2012-09-10',np.nan,'2012-08-17','2012-09-10','2012-07-27','2012-09-10' ,'2012-10-10','2012-10-10','2012-06-27','2012-08-17','2012-04-25'], 'customer_id' : [3002, 3001, 3001, 3003, 3002, 3001, 3001, 3004,3003,3002,3001,3001], 'salesman_id' : [5002, 5003, 5001, np.nan, 5002,5001,5001,np.nan,5003,5002,5003,np.n an]})

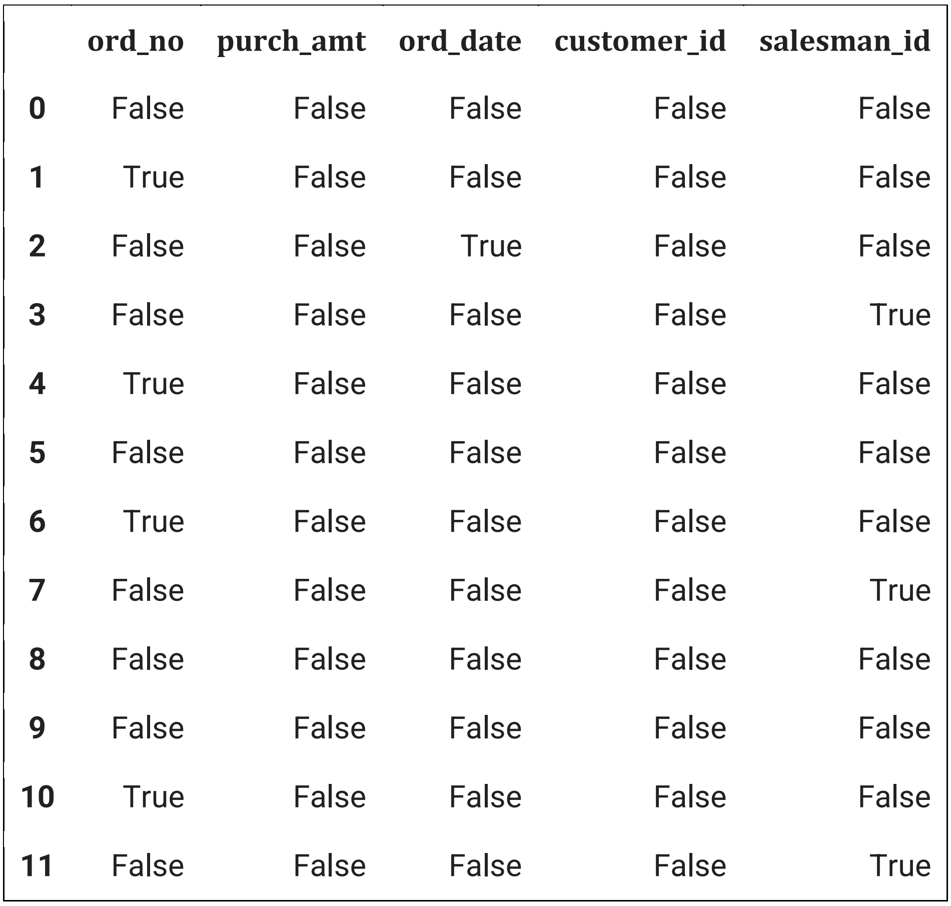
Code :

```
# Import necessary package
import numpy as np
import pandas as pd

# Input data
df = pd.DataFrame({
'ord_no':[70001,np.nan,70002,70004,np.nan,70005,np.nan,70010,7
0003,70012,np.nan,70013],
'purch_amt' : [150.5, 270.65, 65.26, 110.5, 948.5, 2400.6,
5760, 1983.43, 2480.4, 250.45, 75.29, 3045.6],
'ord_date': ['2012-10-05','2012-09-10',np.nan,'2012-08-
17','2012-09-10','2012-07-27','2012-09-10' ,'2012-10-
10','2012-10-10','2012-06-27','2012-08-17','2012-04-25'],
'customer_id' : [3002, 3001, 3001, 3003, 3002, 3001, 3001,
3004,3003,3002,3001,3001],
'salesman_id' : [5002, 5003, 5001, np.nan, 5002, 5001, 5001,
np.nan, 5003, 5002,5003,np.nan]})

# Detecting missing values
df.isna()
```

Output :



	ord_no	purch_amt	ord_date	customer_id	salesman_id
0	False	False	False	False	False
1	True	False	False	False	False
2	False	False	True	False	False
3	False	False	False	False	True
4	True	False	False	False	False
5	False	False	False	False	False
6	True	False	False	False	False
7	False	False	False	False	True
8	False	False	False	False	False
9	False	False	False	False	False
10	True	False	False	False	False
11	False	False	False	False	True

2. Write a Pandas program to drop the rows where at least one element is missing in a given DataFrame.

Input: df = pd.DataFrame({ 'ord_no' : [70001, np.nan, 70002, 70004, np.nan, 70005,np.nan,70010,70003,70012,np.na n,70013], 'purch_amt' : [150.5,270.65,65.26,110.5,948.5,2400.6,5760,1983.43,2480.4,250.45, 75.29,3045.6], 'ord_date': ['2012-10-05','2012-09-10',np.nan,'2012-08-17','2012-09-10','2012-07-27','2012-09-10' ,'2012-10-10','2012-10-10','2012-06-27','2012-08-17','2012-04-25'], 'customer_id' : [3002, 3001, 3001,3003,3002,3001,3001,3004,3003,3002,3001,3001], 'salesman_id':[5002,5003,5001,np.nan,5002,5001,5001,np.nan,5003,5002, 5003,np.n an]})

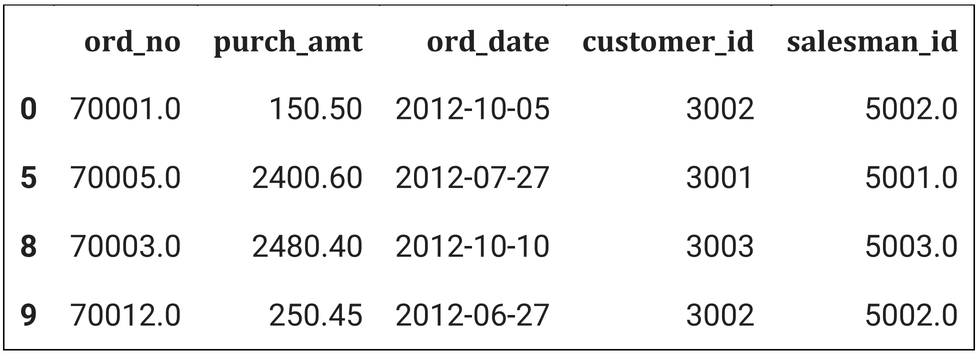
Code :

```
# Import necessary package
import numpy as np
import pandas as pd

# Input data
df = pd.DataFrame({ 'ord_no' : [70001, np.nan, 70002,
70004,  np.nan, 70005,np.nan,70010,70003,70012,np.nan,70013],
'purch_amt' : [150.5, 270.65, 65.26, 110.5, 948.5, 2400.6,
5760, 1983.43,2480.4,250.45, 75.29,3045.6],
'ord_date': ['2012-10-05','2012-09-10',np.nan,'2012-08-
17','2012-09-10','2012-07-27','2012-09-10' ,'2012-10-
10','2012-10-10','2012-06-27','2012-08-17','2012-04-25'],
'customer_id' : [3002, 3001, 3001, 3003, 3002, 3001, 3001,
3004, 3003,3002,3001,3001],
'salesman_id':[5002,5003,5001,np.nan,5002,5001,5001,np.nan,500
3,5002,5003,np.nan]})

# Removing missing value rows and printing other rows
df.dropna()
```

Output :



```
    ord_no   purch_amt   ord_date    customer_id   salesman_id
0   70001.0     150.50   2012-10-05         3002        5002.0
5   70005.0    2400.60   2012-07-27         3001        5001.0
8   70003.0    2480.40   2012-10-10         3003        5003.0
9   70012.0     250.45   2012-06-27         3002        5002.0
```

3. Write a Pandas program to drop the rows where all elements are missing in a given DataFrame.

df = pd.DataFrame({ 'ord_no' : [np.nan, np.nan, 70002, 70004, np.nan, 70005, np.nan,70010,70003,70012,np.n an,70013], 'purch_amt' : [np.nan, 270.65, 65.26, 110.5,948.5,2400.6,5760,1983.43,2480.4,250.45, 75.29,3045.6], 'ord_date': [np.nan,'2012-09-10',np.nan,'2012-08-17','2012-09-10','2012-07-27','2012-09-10','201 2-10-10','2012-10-10','2012-06-27','2012-08-17','2012-04-25'], 'customer_id' : [np.nan, 3001, 3001,3003,3002,3001,3001,3004,3003,3002,3001,3001]})

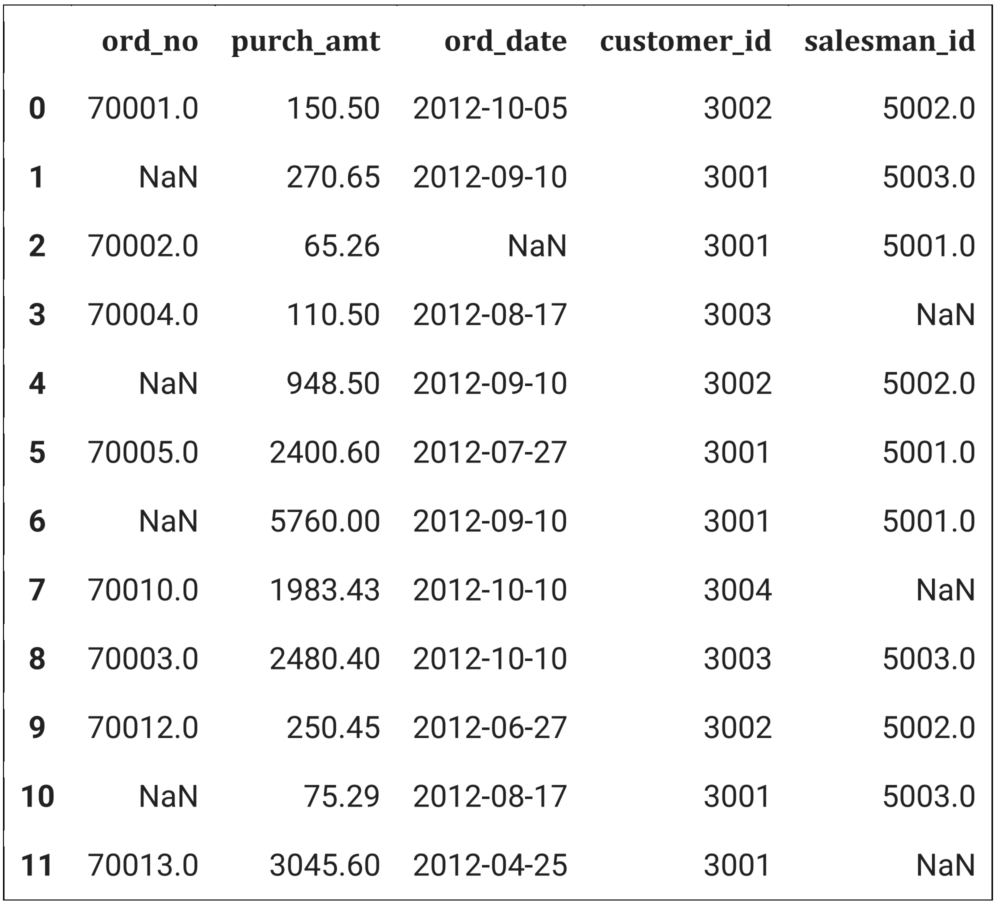
Code :

```
# Import necessary package
import numpy as np
import pandas as pd

# Input data
df = pd.DataFrame({ 'ord_no' : [70001, np.nan, 70002,
70004,  np.nan, 70005,np.nan,70010,70003,70012,np.nan,70013],
                      'purch_amt' :
[150.5,270.65,65.26,110.5,948.5,2400.6,5760,1983.43,2480.4,250.45,
75.29,3045.6],
                       'ord_date': ['2012-10-05','2012-09-
10',np.nan,'2012-08-17','2012-09-10','2012-07-27','2012-09-10'
,'2012-10-10','2012-10-10','2012-06-27','2012-08-17','2012-04-25'],
                       'customer_id' : [3002, 3001,
3001,3003,3002,3001,3001,3004,3003,3002,3001,3001],
                       'salesman_id':[5002,5003,5001,np.nan,5002,5001,5
001,np.nan,5003,5002,5003,np.nan]})

# Dropping the rows where all elements are missing and printing
other rows
df.dropna(how='all')
```

Output :



	ord_no	purch_amt	ord_date	customer_id	salesman_id
0	70001.0	150.50	2012-10-05	3002	5002.0
1	NaN	270.65	2012-09-10	3001	5003.0
2	70002.0	65.26	NaN	3001	5001.0
3	70004.0	110.50	2012-08-17	3003	NaN
4	NaN	948.50	2012-09-10	3002	5002.0
5	70005.0	2400.60	2012-07-27	3001	5001.0
6	NaN	5760.00	2012-09-10	3001	5001.0
7	70010.0	1983.43	2012-10-10	3004	NaN
8	70003.0	2480.40	2012-10-10	3003	5003.0
9	70012.0	250.45	2012-06-27	3002	5002.0
10	NaN	75.29	2012-08-17	3001	5003.0
11	70013.0	3045.60	2012-04-25	3001	NaN

4. Write a Pandas program to drop those rows from a given DataFrame in which specific columns have missing values. Input:

df = pd.DataFrame({ 'ord_no' : [np.nan, np.nan, 70002, np.nan, np.nan, 70005, np.nan,70010,70003,70012,np.n an,np.nan], 'purch_amt' : [np.nan , 270.65, 65.26,np.nan,948.5,2400.6,5760,1983.43,2480.4,250.45, 75.29, np.nan], 'ord_date': [np.nan,'2012-09-10',np.nan,np.nan,'2012-09-10' , '2012-07-27','2012-09-10','2012-10- 10','2012-10-10','2012-06-27','2012-08-17',np.nan], 'customer_id': [np.nan, 3001, 3001, np.nan, 3002, 3001,3001,3004,3003,3002,3001,np.na n]})

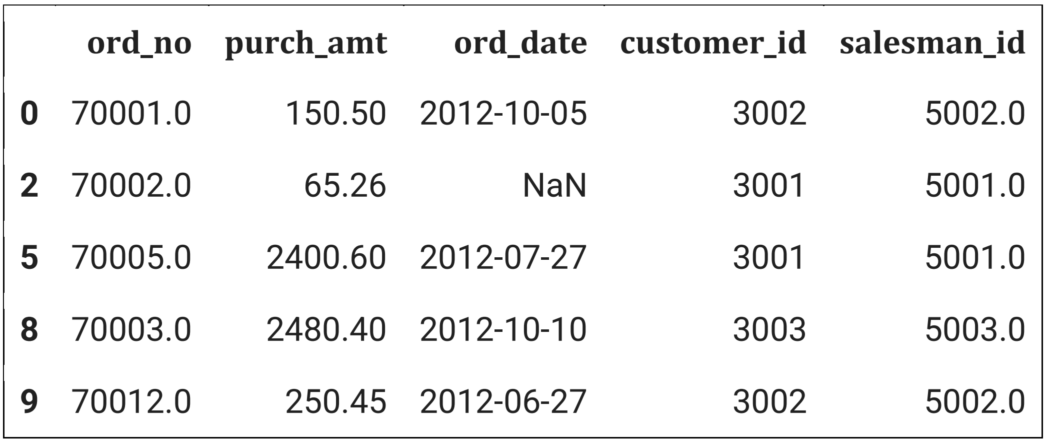
Code :

```
# Import necessary package
import numpy as np
import pandas as pd

# Input data
df = pd.DataFrame({ 'ord_no' : [70001, np.nan, 70002,
70004,  np.nan, 70005,np.nan,70010,70003,70012,np.nan,70013],
                   'purch_amt' :
[150.5,270.65,65.26,110.5,948.5,2400.6,5760,1983.43,2480.4,250
.45, 75.29,3045.6],
                    'ord_date': ['2012-10-05','2012-09-
10',np.nan,'2012-08-17','2012-09-10','2012-07-27','2012-09-10'
,'2012-10-10','2012-10-10','2012-06-27','2012-08-17','2012-04-
25'],
                    'customer_id' : [3002, 3001,
3001,3003,3002,3001,3001,3004,3003,3002,3001,3001],
                    'salesman_id':[5002,5003,5001,np.nan,5002,
5001,5001,np.nan,5003,5002,5003,np.nan]})

# Dropping the rows in which specific columns have missing
values and printing other rows
check_columns = ['ord_no','salesman_id']
df.dropna(subset=check_columns)
```

Output :



	ord_no	purch_amt	ord_date	customer_id	salesman_id
0	70001.0	150.50	2012-10-05	3002	5002.0
2	70002.0	65.26	NaN	3001	5001.0
5	70005.0	2400.60	2012-07-27	3001	5001.0
8	70003.0	2480.40	2012-10-10	3003	5003.0
9	70012.0	250.45	2012-06-27	3002	5002.0