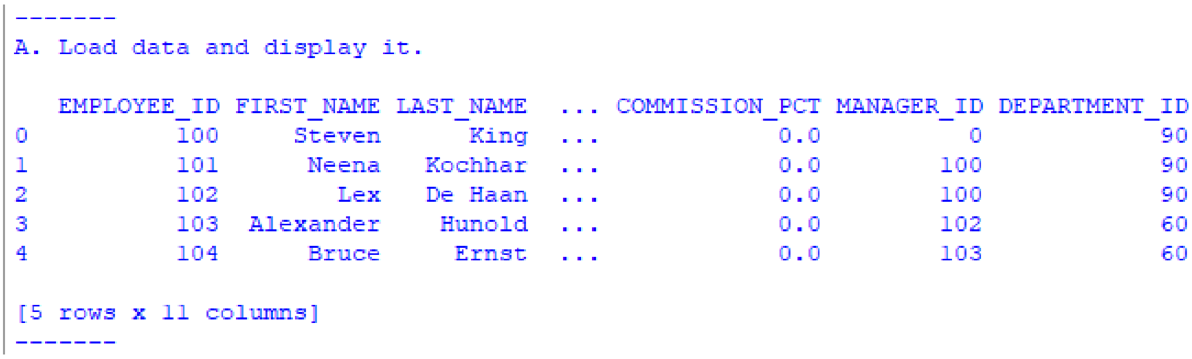
Institute of Computer Technology
B. Tech Computer Science and Engineering
Sub: Data Mining and Warehousing (2CSE60E27)

Practical 3: Advanced Data Exploration (GroupBy)

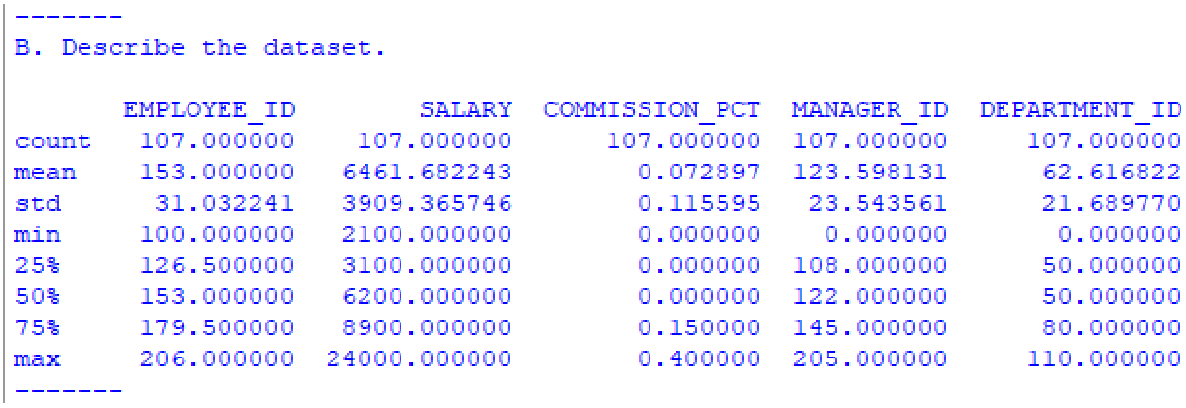
Consider the given dataset of the employee of Zee organization. It has the details of the employee working for that organization. You need to find out the below mentioned information from the given dataset.

1. Load data and display it.



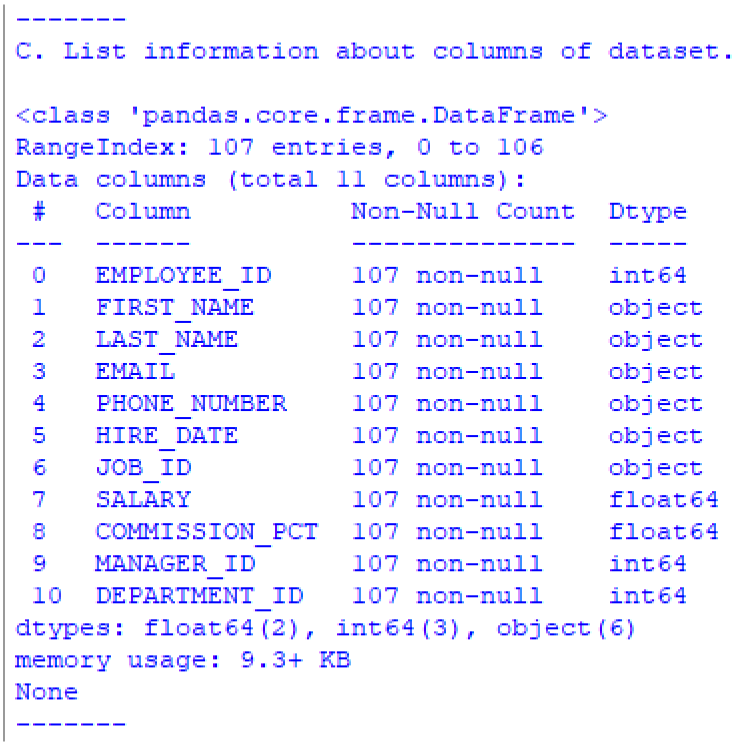
```
-------
A. Load data and display it.

   EMPLOYEE_ID FIRST_NAME LAST_NAME  ... COMMISSION_PCT MANAGER_ID DEPARTMENT_ID
0          100     Steven      King  ...            0.0          0            90
1          101      Neena   Kochhar  ...            0.0        100            90
2          102        Lex   De Haan  ...            0.0        100            90
3          103  Alexander   Hunold  ...            0.0        102            60
4          104      Bruce     Ernst  ...            0.0        103            60

[5 rows x 11 columns]
-------
```

2. Describe the dataset.



```
-------
B. Describe the dataset.

       EMPLOYEE_ID        SALARY  COMMISSION_PCT  MANAGER_ID  DEPARTMENT_ID
count   107.000000    107.000000      107.000000  107.000000     107.000000
mean    153.000000   6461.682243        0.072897  123.598131      62.616822
std      31.032241   3909.365746        0.115595   23.543561      21.689770
min     100.000000   2100.000000        0.000000    0.000000       0.000000
25%     126.500000   3100.000000        0.000000  108.000000      50.000000
50%     153.000000   6200.000000        0.000000  122.000000      50.000000
75%     179.500000   8900.000000        0.150000  145.000000      80.000000
max     206.000000  24000.000000        0.400000  205.000000     110.000000
-------
```

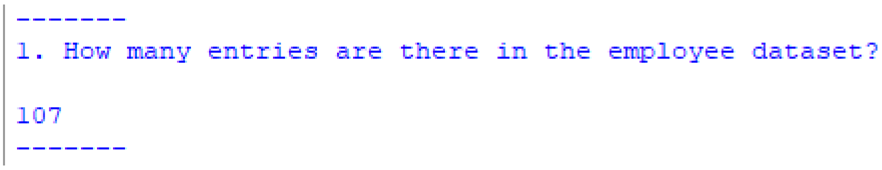
3. List information about columns of dataset.



```
-------
C. List information about columns of dataset.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 107 entries, 0 to 106
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   EMPLOYEE_ID     107 non-null    int64
 1   FIRST_NAME      107 non-null    object
 2   LAST_NAME       107 non-null    object
 3   EMAIL           107 non-null    object
 4   PHONE_NUMBER    107 non-null    object
 5   HIRE_DATE       107 non-null    object
 6   JOB_ID          107 non-null    object
 7   SALARY          107 non-null    float64
 8   COMMISSION_PCT  107 non-null    float64
 9   MANAGER_ID      107 non-null    int64
 10  DEPARTMENT_ID   107 non-null    int64
dtypes: float64(2), int64(3), object(6)
memory usage: 9.3+ KB
None
-------
```

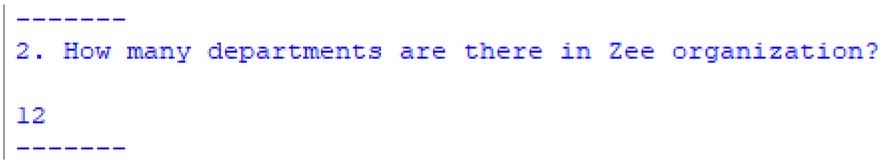
Explore the below queries:

I. How many entries are there in the employee dataset?



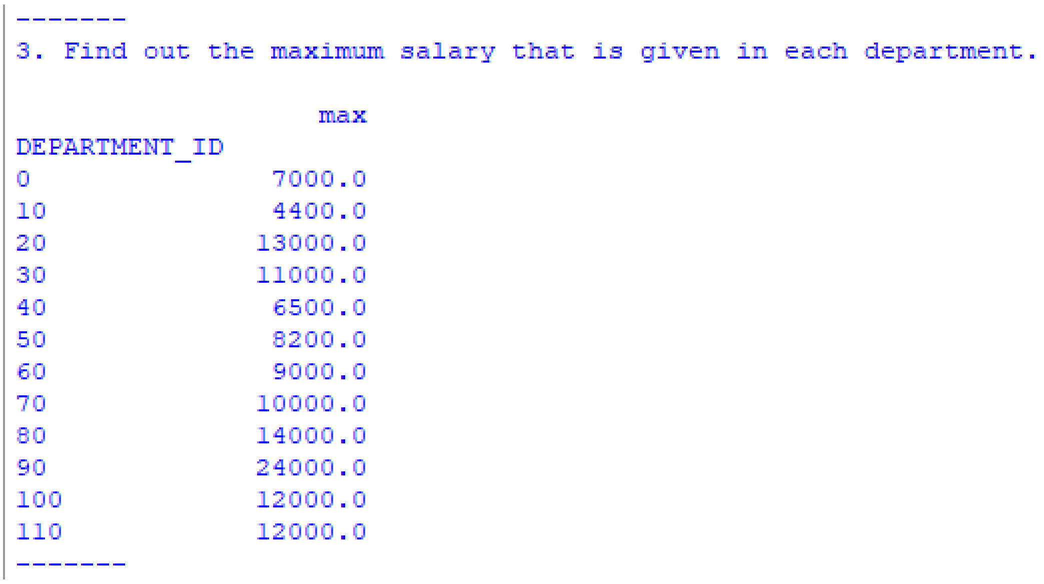
```
-------
1. How many entries are there in the employee dataset?

107
-------
```

II. How many departments are there in Zee organization?



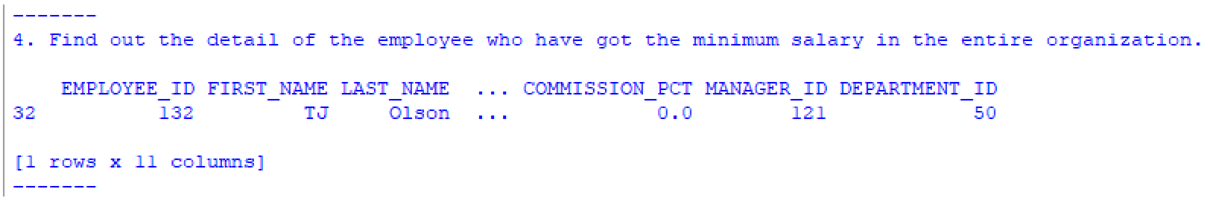
```
-------
2. How many departments are there in Zee organization?

12
-------
```

III. Find out the maximum salary that is given in each department?



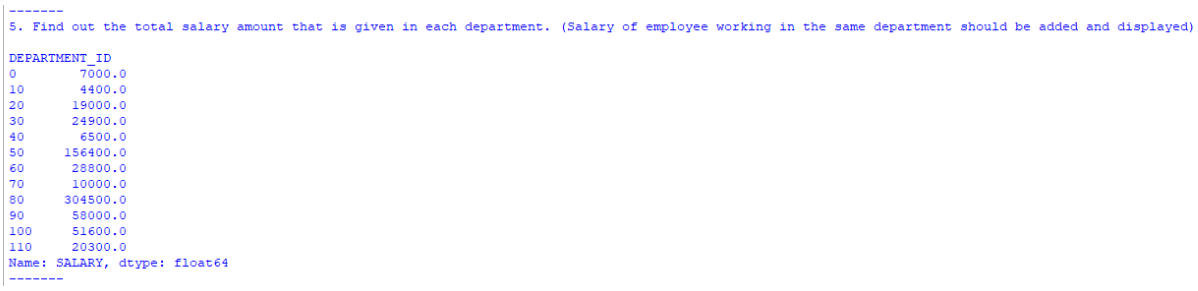
```
-------
3. Find out the maximum salary that is given in each department.

                   max
DEPARTMENT_ID
0                7000.0
10               4400.0
20              13000.0
30              11000.0
40               6500.0
50               8200.0
60               9000.0
70              10000.0
80              14000.0
90              24000.0
100             12000.0
110             12000.0
-------
```

IV. Find out the detail of the employee who have got the minimum salary in the entire organization?



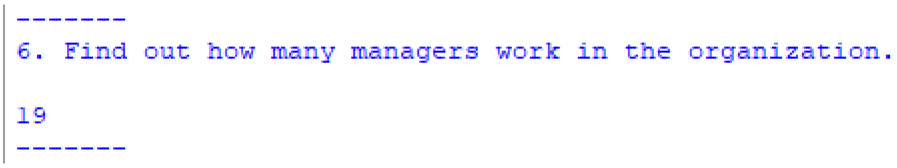
```
-------
4. Find out the detail of the employee who have got the minimum salary in the entire organization.

    EMPLOYEE_ID FIRST_NAME LAST_NAME  ... COMMISSION_PCT MANAGER_ID DEPARTMENT_ID
32          132         TJ     Olson  ...            0.0        121            50

[1 rows x 11 columns]
-------
```

V. Find out the total salary amount that is given in each department? (Salary of employee working in the same department should be added and displayed)



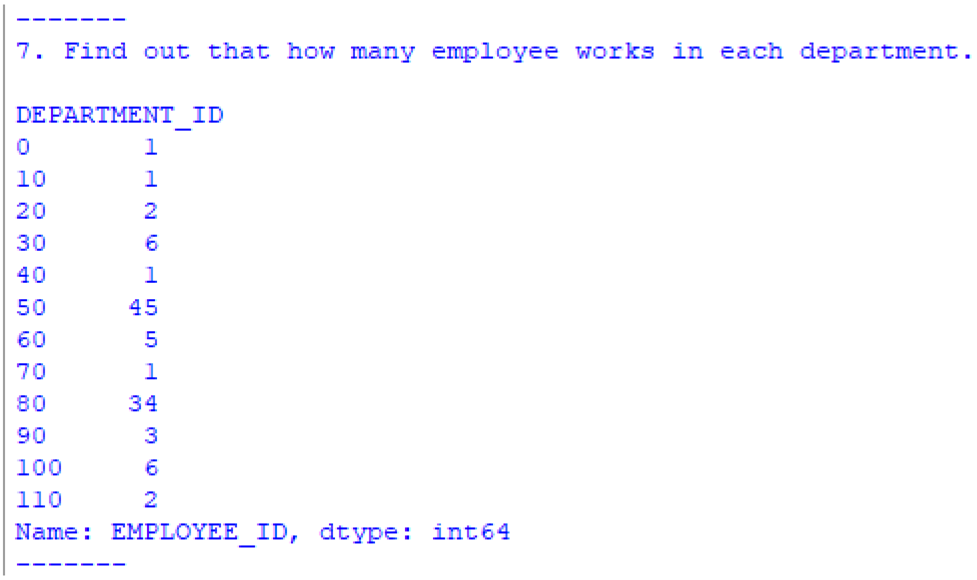
```
-------
5. Find out the total salary amount that is given in each department. (Salary of employee working in the same department should be added and displayed)

DEPARTMENT_ID
0         7000.0
10        4400.0
20       19000.0
30       24900.0
40        6500.0
50      156400.0
60       28800.0
70       10000.0
80      304500.0
90       58000.0
100      51600.0
110      20300.0
Name: SALARY, dtype: float64
-------
```

VI. Find out how many managers work in the organization?



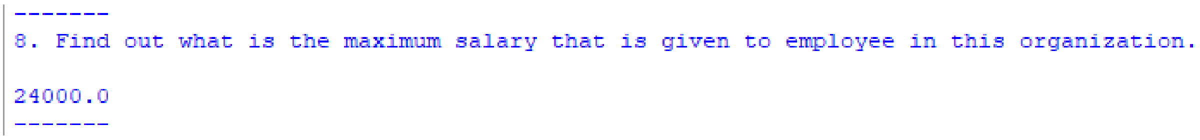
```
-------
6. Find out how many managers work in the organization.

19
-------
```

VII. Find out that how many employee works in each department?



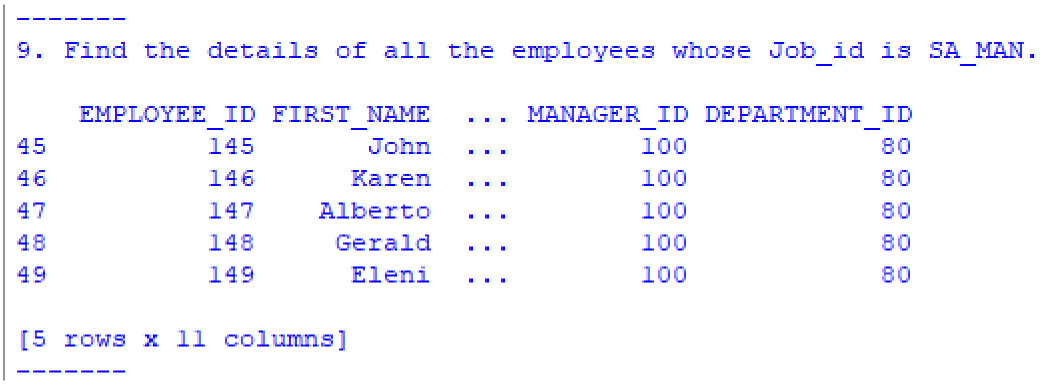
```
-------
7. Find out that how many employee works in each department.

DEPARTMENT_ID
0         1
10        1
20        2
30        6
40        1
50       45
60        5
70        1
80       34
90        3
100       6
110       2
Name: EMPLOYEE_ID, dtype: int64
-------
```

VIII. Find out what is the maximum salary that is given to employee in this organization?



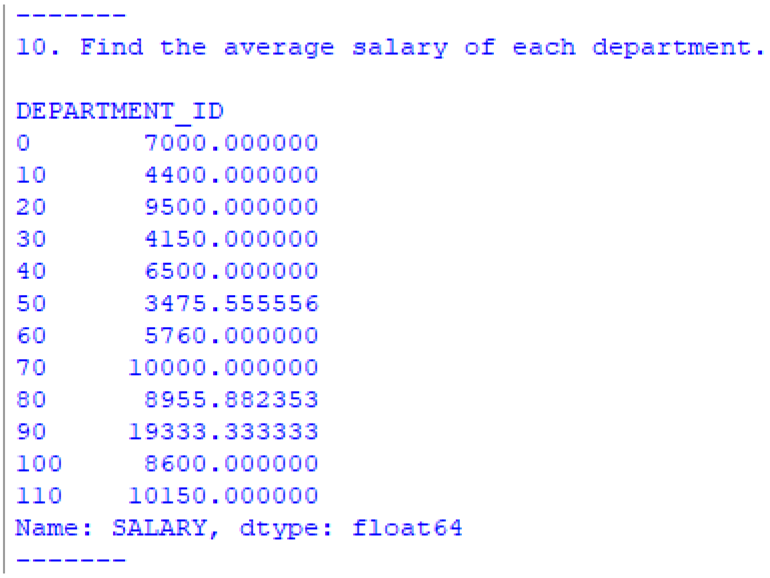
```
-------
8. Find out what is the maximum salary that is given to employee in this organization.

24000.0
-------
```

IX. Find the details of all the employees whose Job_id is "SA_MAN".



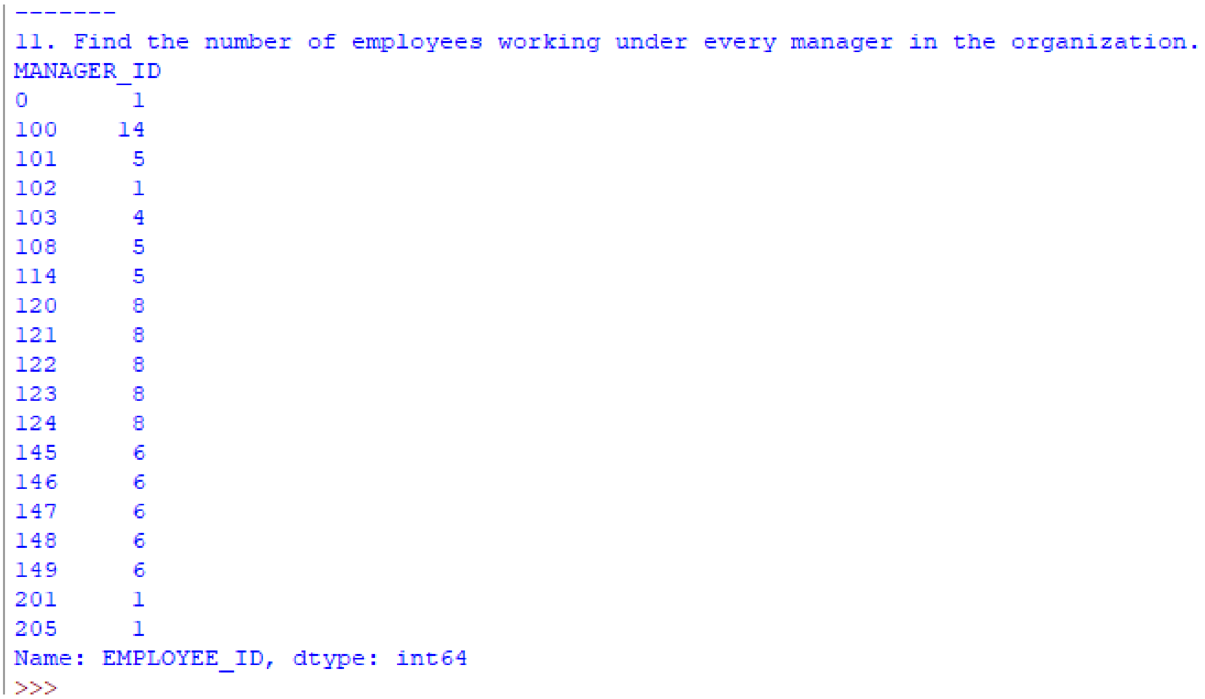
```
-------
9. Find the details of all the employees whose Job_id is SA_MAN.

    EMPLOYEE_ID FIRST_NAME  ... MANAGER_ID DEPARTMENT_ID
45          145       John  ...        100            80
46          146      Karen  ...        100            80
47          147    Alberto  ...        100            80
48          148     Gerald  ...        100            80
49          149      Eleni  ...        100            80

[5 rows x 11 columns]
-------
```

X. Find the average salary of each department?



```
-------
10. Find the average salary of each department.

DEPARTMENT_ID
0        7000.000000
10       4400.000000
20       9500.000000
30       4150.000000
40       6500.000000
50       3475.555556
60       5760.000000
70      10000.000000
80       8955.882353
90      19333.333333
100      8600.000000
110     10150.000000
Name: SALARY, dtype: float64
-------
```

XI. Find the number of employees working under every manager in the organization.



```
-------
11. Find the number of employees working under every manager in the organization.
MANAGER_ID
0       1
100    14
101     5
102     1
103     4
108     5
114     5
120     8
121     8
122     8
123     8
124     8
145     6
146     6
147     6
148     6
149     6
201     1
205     1
Name: EMPLOYEE_ID, dtype: int64
>>>
```

Code:

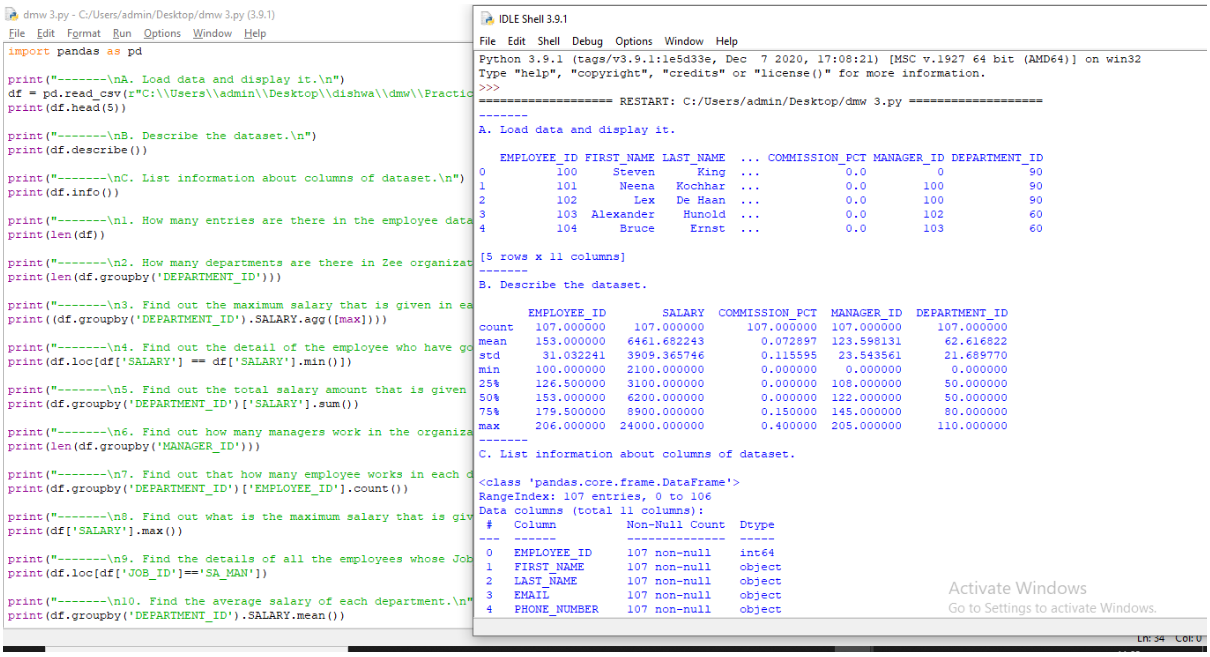
import pandas as pd

print("-------\nA. Load data and display it.\n")
df = pd.read_csv(r"C:\\Users\\admin\\Desktop\\dishwa\\dmw\\Practical 3\\employees.csv", delimiter = ';', on_bad_lines='skip')
print(df.head(5))

print("-------\nB. Describe the dataset.\n")
print(df.describe())

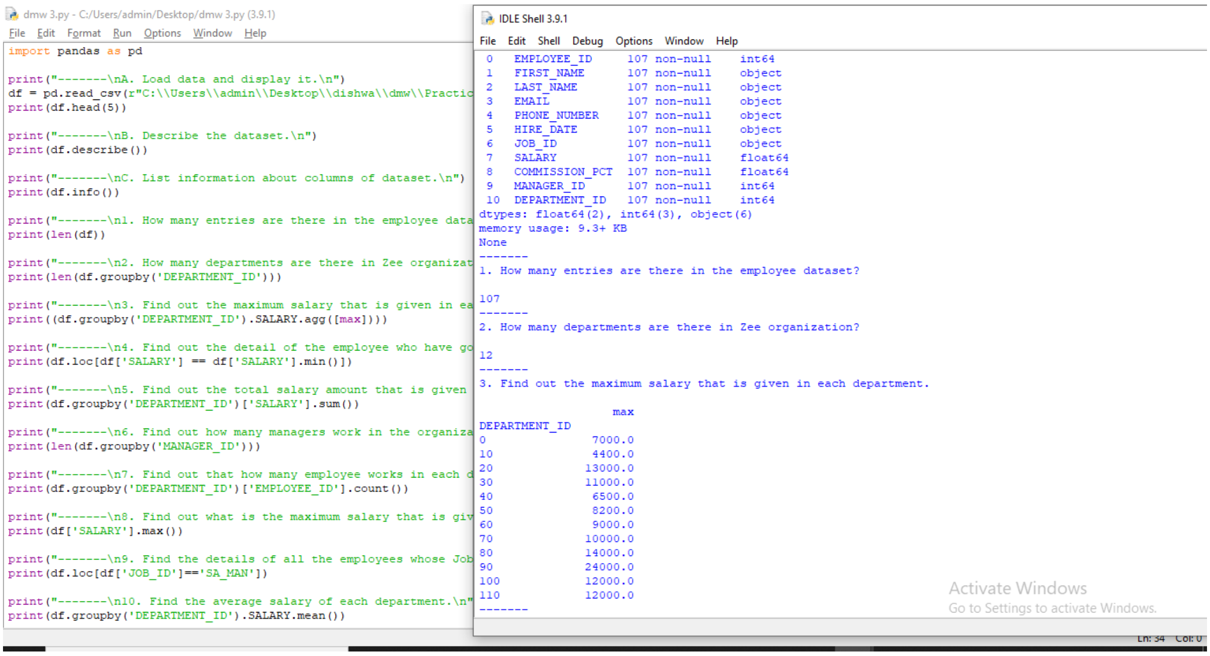
```
print("-------\nC. List information about columns of dataset.\n")
print(df.info())

print("-------\n1. How many entries are there in the employee dataset?\n")
print(len(df))

print("-------\n2. How many departments are there in Zee organization?\n")
print(len(df.groupby('DEPARTMENT_ID')))

print("-------\n3. Find out the maximum salary that is given in each department.\n")
print((df.groupby('DEPARTMENT_ID').SALARY.agg([max])))

print("-------\n4. Find out the detail of the employee who have got the minimum salary in the entire organization.\n")
print(df.loc[df['SALARY'] == df['SALARY'].min()])

print("-------\n5. Find out the total salary amount that is given in each department. (Salary of employee working in the same department should be added and displayed)\n")
print(df.groupby('DEPARTMENT_ID')['SALARY'].sum())

print("-------\n6. Find out how many managers work in the organization.\n")
print(len(df.groupby('MANAGER_ID')))

print("-------\n7. Find out that how many employee works in each department.\n")
print(df.groupby('DEPARTMENT_ID')['EMPLOYEE_ID'].count())

print("-------\n8. Find out what is the maximum salary that is given to employee in this organization.\n")
print(df['SALARY'].max())

print("-------\n9. Find the details of all the employees whose Job_id is SA_MAN.\n")
print(df.loc[df['JOB_ID']=='SA_MAN'])

print("-------\n10. Find the average salary of each department.\n")
print(df.groupby('DEPARTMENT_ID').SALARY.mean())

print("-------\n11. Find the number of employees working under every manager in the organization.")
print(df.groupby('MANAGER_ID')['EMPLOYEE_ID'].count())
```

Output:


dmw 3.py - C:/Users/admin/Desktop/dmw 3.py (3.9.1)
File Edit Format Run Options Window Help
IDLE Shell 3.9.1
File Edit Shell Debug Options Window Help
Activate Windows
Go to Settings to activate Windows.
Ln: 34 Col: 0

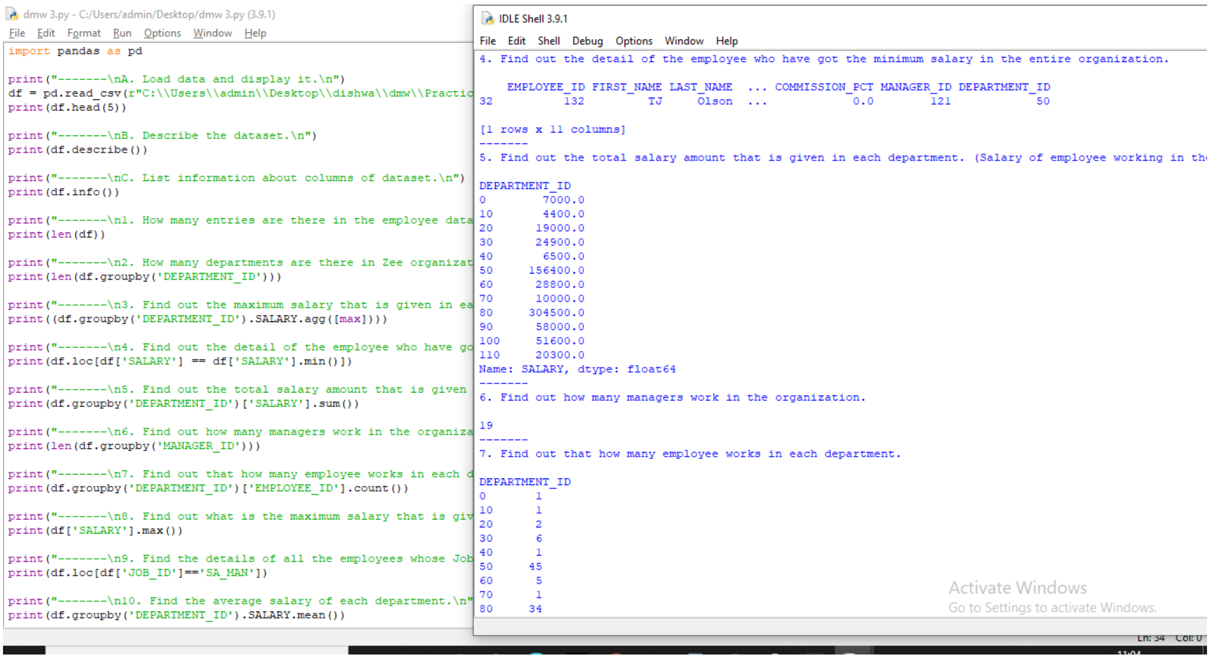
```python
import pandas as pd

print("-------\nA. Load data and display it.\n")
df = pd.read_csv(r"C:\\Users\\admin\\Desktop\\dishwa\\dmw\\Practic
print(df.head(5))

print("-------\nB. Describe the dataset.\n")
print(df.describe())

print("-------\nC. List information about columns of dataset.\n")
print(df.info())

print("-------\n1. How many entries are there in the employee data
print(len(df))

print("-------\n2. How many departments are there in Zee organizat
print(len(df.groupby('DEPARTMENT_ID')))

print("-------\n3. Find out the maximum salary that is given in ea
print((df.groupby('DEPARTMENT_ID').SALARY.agg([max])))

print("-------\n4. Find out the detail of the employee who have go
print(df.loc[df['SALARY'] == df['SALARY'].min()])

print("-------\n5. Find out the total salary amount that is given
print(df.groupby('DEPARTMENT_ID')['SALARY'].sum())

print("-------\n6. Find out how many managers work in the organiza
print(len(df.groupby('MANAGER_ID')))

print("-------\n7. Find out that how many employee works in each d
print(df.groupby('DEPARTMENT_ID')['EMPLOYEE_ID'].count())

print("-------\n8. Find out what is the maximum salary that is giv
print(df['SALARY'].max())

print("-------\n9. Find the details of all the employees whose Job
print(df.loc[df['JOB_ID']=='SA_MAN'])

print("-------\n10. Find the average salary of each department.\n"
print(df.groupby('DEPARTMENT_ID').SALARY.mean())
```

```
Python 3.9.1 (tags/v3.9.1:1e5d33e, Dec  7 2020, 17:08:21) [MSC v.1927 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> 
=================== RESTART: C:/Users/admin/Desktop/dmw 3.py ===================
-------
A. Load data and display it.

   EMPLOYEE_ID FIRST_NAME LAST_NAME  ... COMMISSION_PCT MANAGER_ID DEPARTMENT_ID
0          100     Steven      King  ...            0.0          0            90
1          101      Neena   Kochhar  ...            0.0        100            90
2          102        Lex   De Haan  ...            0.0        100            90
3          103  Alexander    Hunold  ...            0.0        102            60
4          104      Bruce     Ernst  ...            0.0        103            60

[5 rows x 11 columns]
-------
B. Describe the dataset.

       EMPLOYEE_ID        SALARY  COMMISSION_PCT  MANAGER_ID  DEPARTMENT_ID
count   107.000000    107.000000      107.000000  107.000000     107.000000
mean    153.000000   6461.682243        0.072897  123.598131      62.616822
std      31.032241   3909.365746        0.115595   23.543561      21.689770
min     100.000000   2100.000000        0.000000    0.000000       0.000000
25%     126.500000   3100.000000        0.000000  108.000000      50.000000
50%     153.000000   6200.000000        0.000000  122.000000      50.000000
75%     179.500000   8900.000000        0.150000  145.000000      80.000000
max     206.000000  24000.000000        0.400000  205.000000     110.000000
-------
C. List information about columns of dataset.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 107 entries, 0 to 106
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   EMPLOYEE_ID     107 non-null    int64  
 1   FIRST_NAME      107 non-null    object 
 2   LAST_NAME       107 non-null    object 
 3   EMAIL           107 non-null    object 
 4   PHONE_NUMBER    107 non-null    object 
```


dmw 3.py - C:/Users/admin/Desktop/dmw 3.py (3.9.1)
File Edit Format Run Options Window Help
IDLE Shell 3.9.1
File Edit Shell Debug Options Window Help
Activate Windows
Go to Settings to activate Windows.
Ln: 34 Col: 0

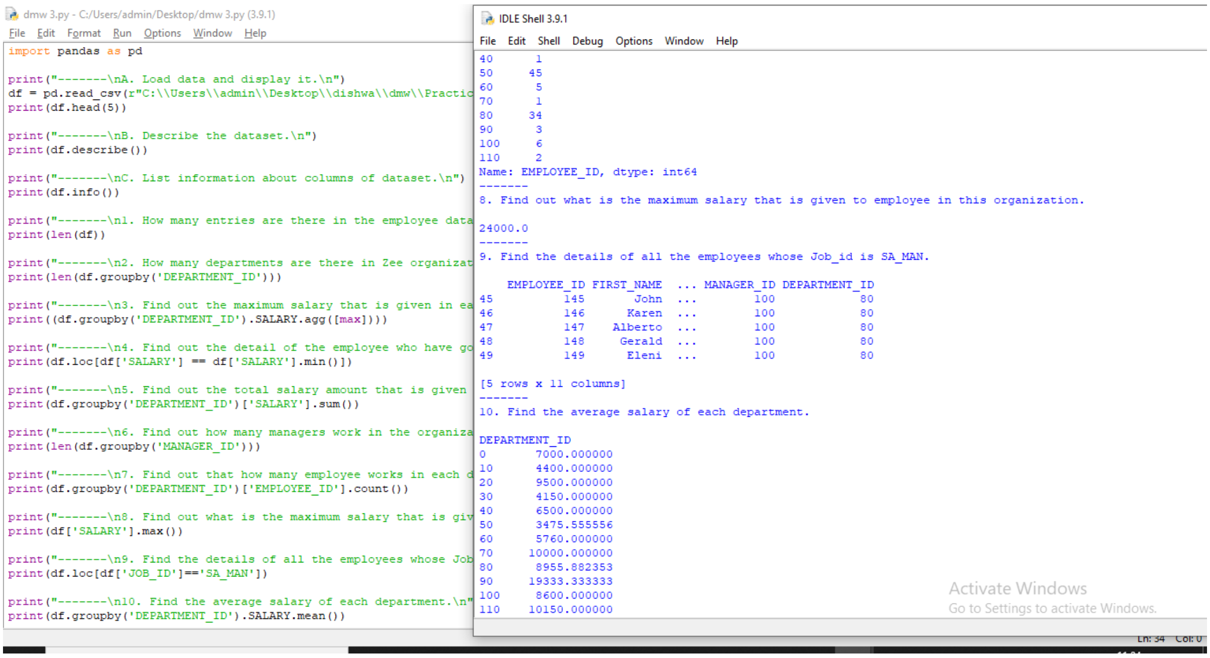
```python
import pandas as pd

print("-------\nA. Load data and display it.\n")
df = pd.read_csv(r"C:\\Users\\admin\\Desktop\\dishwa\\dmw\\Practic
print(df.head(5))

print("-------\nB. Describe the dataset.\n")
print(df.describe())

print("-------\nC. List information about columns of dataset.\n")
print(df.info())

print("-------\n1. How many entries are there in the employee data
print(len(df))

print("-------\n2. How many departments are there in Zee organizat
print(len(df.groupby('DEPARTMENT_ID')))

print("-------\n3. Find out the maximum salary that is given in ea
print((df.groupby('DEPARTMENT_ID').SALARY.agg([max])))

print("-------\n4. Find out the detail of the employee who have go
print(df.loc[df['SALARY'] == df['SALARY'].min()])

print("-------\n5. Find out the total salary amount that is given
print(df.groupby('DEPARTMENT_ID')['SALARY'].sum())

print("-------\n6. Find out how many managers work in the organiza
print(len(df.groupby('MANAGER_ID')))

print("-------\n7. Find out that how many employee works in each d
print(df.groupby('DEPARTMENT_ID')['EMPLOYEE_ID'].count())

print("-------\n8. Find out what is the maximum salary that is giv
print(df['SALARY'].max())

print("-------\n9. Find the details of all the employees whose Job
print(df.loc[df['JOB_ID']=='SA_MAN'])

print("-------\n10. Find the average salary of each department.\n"
print(df.groupby('DEPARTMENT_ID').SALARY.mean())
```

```
 0   EMPLOYEE_ID     107 non-null    int64  
 1   FIRST_NAME      107 non-null    object 
 2   LAST_NAME       107 non-null    object 
 3   EMAIL           107 non-null    object 
 4   PHONE_NUMBER    107 non-null    object 
 5   HIRE_DATE       107 non-null    object 
 6   JOB_ID          107 non-null    object 
 7   SALARY          107 non-null    float64
 8   COMMISSION_PCT  107 non-null    float64
 9   MANAGER_ID      107 non-null    int64  
 10  DEPARTMENT_ID   107 non-null    int64  
dtypes: float64(2), int64(3), object(6)
memory usage: 9.3+ KB
None
-------
1. How many entries are there in the employee dataset?

107
-------
2. How many departments are there in Zee organization?

12
-------
3. Find out the maximum salary that is given in each department.

                   max
DEPARTMENT_ID         
0               7000.0
10              4400.0
20             13000.0
30             11000.0
40              6500.0
50              8200.0
60              9000.0
70             10000.0
80             14000.0
90             24000.0
100            12000.0
110            12000.0
-------
```


dmw 3.py - C:/Users/admin/Desktop/dmw 3.py (3.9.1)
File Edit Format Run Options Window Help
IDLE Shell 3.9.1
File Edit Shell Debug Options Window Help
Activate Windows
Go to Settings to activate Windows.
Ln: 34 Col: 0

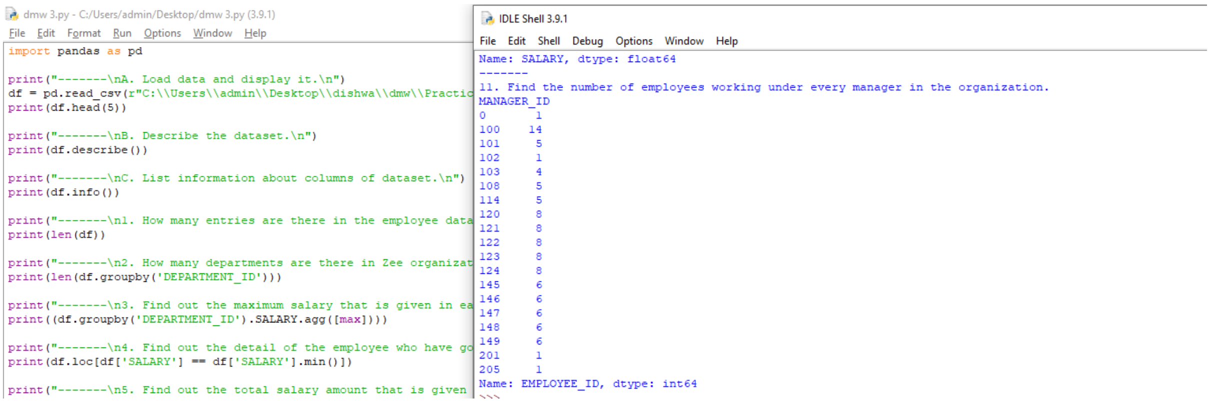
```
import pandas as pd

print("-------\nA. Load data and display it.\n")
df = pd.read_csv(r"C:\\Users\\admin\\Desktop\\dishwa\\dmw\\Practic
print(df.head(5))

print("-------\nB. Describe the dataset.\n")
print(df.describe())

print("-------\nC. List information about columns of dataset.\n")
print(df.info())

print("-------\n1. How many entries are there in the employee data
print(len(df))

print("-------\n2. How many departments are there in Zee organizat
print(len(df.groupby('DEPARTMENT_ID')))

print("-------\n3. Find out the maximum salary that is given in ea
print((df.groupby('DEPARTMENT_ID').SALARY.agg([max])))

print("-------\n4. Find out the detail of the employee who have go
print(df.loc[df['SALARY'] == df['SALARY'].min()])

print("-------\n5. Find out the total salary amount that is given
print(df.groupby('DEPARTMENT_ID')['SALARY'].sum())

print("-------\n6. Find out how many managers work in the organiza
print(len(df.groupby('MANAGER_ID')))

print("-------\n7. Find out that how many employee works in each d
print(df.groupby('DEPARTMENT_ID')['EMPLOYEE_ID'].count())

print("-------\n8. Find out what is the maximum salary that is giv
print(df['SALARY'].max())

print("-------\n9. Find the details of all the employees whose Job
print(df.loc[df['JOB_ID']=='SA_MAN'])

print("-------\n10. Find the average salary of each department.\n"
print(df.groupby('DEPARTMENT_ID').SALARY.mean())
```

```
4. Find out the detail of the employee who have got the minimum salary in the entire organization.

    EMPLOYEE_ID FIRST_NAME LAST_NAME  ... COMMISSION_PCT MANAGER_ID DEPARTMENT_ID
32          132         TJ     Olson  ...            0.0        121            50

[1 rows x 11 columns]
-------
5. Find out the total salary amount that is given in each department. (Salary of employee working in th

DEPARTMENT_ID
0         7000.0
10        4400.0
20       19000.0
30       24900.0
40        6500.0
50      156400.0
60       28800.0
70       10000.0
80      304500.0
90       58000.0
100      51600.0
110      20300.0
Name: SALARY, dtype: float64
-------
6. Find out how many managers work in the organization.

19
-------
7. Find out that how many employee works in each department.

DEPARTMENT_ID
0       1
10      1
20      2
30      6
40      1
50     45
60      5
70      1
80     34
```


dmw 3.py - C:/Users/admin/Desktop/dmw 3.py (3.9.1)
File Edit Format Run Options Window Help
IDLE Shell 3.9.1
File Edit Shell Debug Options Window Help
Activate Windows
Go to Settings to activate Windows.
Ln: 34 Col: 0

```
import pandas as pd

print("-------\nA. Load data and display it.\n")
df = pd.read_csv(r"C:\\Users\\admin\\Desktop\\dishwa\\dmw\\Practic
print(df.head(5))

print("-------\nB. Describe the dataset.\n")
print(df.describe())

print("-------\nC. List information about columns of dataset.\n")
print(df.info())

print("-------\n1. How many entries are there in the employee data
print(len(df))

print("-------\n2. How many departments are there in Zee organizat
print(len(df.groupby('DEPARTMENT_ID')))

print("-------\n3. Find out the maximum salary that is given in ea
print((df.groupby('DEPARTMENT_ID').SALARY.agg([max])))

print("-------\n4. Find out the detail of the employee who have go
print(df.loc[df['SALARY'] == df['SALARY'].min()])

print("-------\n5. Find out the total salary amount that is given
print(df.groupby('DEPARTMENT_ID')['SALARY'].sum())

print("-------\n6. Find out how many managers work in the organiza
print(len(df.groupby('MANAGER_ID')))

print("-------\n7. Find out that how many employee works in each d
print(df.groupby('DEPARTMENT_ID')['EMPLOYEE_ID'].count())

print("-------\n8. Find out what is the maximum salary that is giv
print(df['SALARY'].max())

print("-------\n9. Find the details of all the employees whose Job
print(df.loc[df['JOB_ID']=='SA_MAN'])

print("-------\n10. Find the average salary of each department.\n"
print(df.groupby('DEPARTMENT_ID').SALARY.mean())
```

```
40      1
50     45
60      5
70      1
80     34
90      3
100     6
110     2
Name: EMPLOYEE_ID, dtype: int64
-------
8. Find out what is the maximum salary that is given to employee in this organization.

24000.0
-------
9. Find the details of all the employees whose Job_id is SA_MAN.

    EMPLOYEE_ID FIRST_NAME  ... MANAGER_ID DEPARTMENT_ID
45          145       John  ...        100            80
46          146      Karen  ...        100            80
47          147    Alberto  ...        100            80
48          148     Gerald  ...        100            80
49          149      Eleni  ...        100            80

[5 rows x 11 columns]
-------
10. Find the average salary of each department.

DEPARTMENT_ID
0       7000.000000
10      4400.000000
20      9500.000000
30      4150.000000
40      6500.000000
50      3475.555556
60      5760.000000
70     10000.000000
80      8955.882353
90     19333.333333
100     8600.000000
110    10150.000000
```


dmw 3.py - C:/Users/admin/Desktop/dmw 3.py (3.9.1)
File Edit Format Run Options Window Help
IDLE Shell 3.9.1
File Edit Shell Debug Options Window Help

```
import pandas as pd

print("-------\nA. Load data and display it.\n")
df = pd.read_csv(r"C:\\Users\\admin\\Desktop\\dishwa\\dmw\\Practic
print(df.head(5))

print("-------\nB. Describe the dataset.\n")
print(df.describe())

print("-------\nC. List information about columns of dataset.\n")
print(df.info())

print("-------\n1. How many entries are there in the employee data
print(len(df))

print("-------\n2. How many departments are there in Zee organizat
print(len(df.groupby('DEPARTMENT_ID')))

print("-------\n3. Find out the maximum salary that is given in ea
print((df.groupby('DEPARTMENT_ID').SALARY.agg([max])))

print("-------\n4. Find out the detail of the employee who have go
print(df.loc[df['SALARY'] == df['SALARY'].min()])

print("-------\n5. Find out the total salary amount that is given
```

```
Name: SALARY, dtype: float64
-------
11. Find the number of employees working under every manager in the organization.
MANAGER_ID
0        1
100     14
101      5
102      1
103      4
108      5
114      5
120      8
121      8
122      8
123      8
124      8
145      6
146      6
147      6
148      6
149      6
201      1
205      1
Name: EMPLOYEE_ID, dtype: int64
>>>
```