Caitlin Sisilli

Homework 6

Question 1)

```
import numpy as np

import pandas as pd

data=pd.read_table('Salaries.csv',sep=',')

print(data)
```

Output:

```
         rank discipline  phd  service     sex    salary
0        Prof          B   56       49    Male  186960.0
1        Prof          A   12        6    Male   93000.0
2        Prof          A   23       20    Male  110515.0
3        Prof          A   40       31    Male  131205.0
4        Prof          B   20       18    Male  104800.0
..        ...        ...  ...      ...     ...       ...
73       Prof          B   18       10  Female  105450.0
74  AssocProf          B   19        6  Female  104542.0
75       Prof          B   17       17  Female  124312.0
76       Prof          A   28       14  Female  109954.0
77       Prof          A   23       15  Female  109646.0

[78 rows x 6 columns]
```

Question 2)

```
import numpy as np

import pandas as pd

data=pd.read_table('Salaries.csv',sep=',')

print(data.size)
```

Output:

The dimensions of the data set are: 468

Question 3)

```
import numpy as np

import pandas as pd

data=pd.read_table('Salaries.csv',sep=',')

print(data.dtypes.value_counts())

print(data.dtypes)
```

Output:

```
object     3
int64      2
float64    1
dtype: int64
rank           object
discipline     object
phd             int64
service         int64
sex            object
salary        float64
dtype: object
```

Question 4)

```
import numpy as np

import pandas as pd

data=pd.read_table('Salaries.csv',sep=',')

print(data.isnull())
```

Output:

There are many false values as shown down below. They are bases on columns.

```
rank discipline phd service sex salary 0 False False False False False
False 1 False False False False False False 2 False False False False
False False 3 False False False False False False 4 False False False
False False False .. ... ... ... ... ... ... 73 False False False False
False False 74 False False False False False False 75 False False False
False False False 76 False False False False False False 77 False False
False False False False [78 rows x 6 columns]
```

Question 5)

I replace multiple columns using the replace from panda, there is a lot of missing information in specific parts like a professor or an associate and numbers.

```
import numpy as np

import pandas as pd

data=pd.read_table('Salaries.csv',sep=',')

print(data.replace)
```

Output:

```
<bound method DataFrame.replace of          rank discipline  phd  service
sex    salary
0        Prof          B   56       49    Male  186960.0
1        Prof          A   12        6    Male   93000.0
2        Prof          A   23       20    Male  110515.0
3        Prof          A   40       31    Male  131205.0
4        Prof          B   20       18    Male  104800.0
..        ...        ...  ...      ...     ...       ...
73       Prof          B   18       10  Female  105450.0
74  AssocProf          B   19        6  Female  104542.0
75       Prof          B   17       17  Female  124312.0
```

```
76        Prof           A   28        14  Female  109954.0
77        Prof           A   23        15  Female  109646.0

[78 rows x 6 columns]>
```

Question 6)

```
import numpy as np

import pandas as pd

data=pd.read_table('Salaries.csv',sep=',', header=0)

data.loc[:,data.columns.isin(['rank','sex','salary'])]
```

Output:

	rank	sex	salary
0	Prof	Male	186960.0
1	Prof	Male	93000.0
2	Prof	Male	110515.0
3	Prof	Male	131205.0
4	Prof	Male	104800.0
...	...	...	...
73	Prof	Female	105450.0
74	AssocProf	Female	104542.0

	rank	sex	salary
75	Prof	Female	124312.0
76	Prof	Female	109954.0
77	Prof	Female	109646.0

78 rows × 3 columns

Question 7)

```
import numpy as np

import pandas as pd

data=pd.read_table('Salaries.csv',sep=',', header=0)

data.loc[:,data.columns.isin(['rank','sex','salary'])]

display(data.loc[(data['salary']>=155000)])
```

Output:

	rank	discipline	phd	service	sex	salary
0	Prof	B	56	49	Male	186960.0
13	Prof	B	35	33	Male	162200.0
27	Prof	A	45	43	Male	155865.0

	rank	discipline	phd	service	sex	salary
31	Prof	B	22	21	Male	155750.0
72	Prof	B	24	15	Female	161101.0

Question 8)

By finding the outlier you find the standard deviation of IQR then the easiest way is by a box plot graph because it be the one out of the main area of the box plot.

Question 9)

Depending on the data it can be nominal data or ordinal data. With the data you would use the panda library.

Get_dummies() for the data.

Then you would create a dictionary with key as category and values within ranks.

Then you would create a new column and map the ordinal column with the newly created dictionary.

After you would drop the original column.

Question 10)

```
import numpy as np

import pandas as pd

data=pd.read_table('Salaries.csv',sep=',', header=0)
```

```
data.loc[:,data.columns.isin(['rank','sex','salary'])]
```

```
display(data['salary'].nlargest(n=5))
```

Output:

```
0     186960.0
13    162200.0
72    161101.0
27    155865.0
31    155750.0
Name: salary, dtype: float64
```