Exercise 5

```
import pandas as pd
sal = pd.read_csv("Salaries.csv")

sal.head()
```

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	

Use the .info() method to find out how many entries there are.

```
sal.info(verbose=True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 148654 entries, 0 to 148653
Data columns (total 13 columns):
Id                  148654 non-null int64
EmployeeName        148654 non-null object
JobTitle            148654 non-null object
BasePay             148045 non-null float64
OvertimePay         148650 non-null float64
OtherPay            148650 non-null float64
Benefits            112491 non-null float64
TotalPay            148654 non-null float64
TotalPayBenefits    148654 non-null float64
Year                148654 non-null int64
Notes               0 non-null float64
Agency              148654 non-null object
Status              0 non-null float64
dtypes: float64(8), int64(2), object(3)
memory usage: 14.7+ MB
```

What is the average BasePay?

```
sal["BasePay"].mean()
```

```
66325.4488404877
```

What is the highest amount of OvertimePay in the dataset ?

```
sal['OvertimePay'].max()
```

```
245131.88
```

What is the job title of JOSEPH DRISCOLL ? Note: Use all caps, otherwise you may get an answer that doesn't match up (there is also a lowercase Joseph Driscoll).

```
sal[sal['EmployeeName']=='JOSEPH DRISCOLL']['JobTitle']
```

```
24    CAPTAIN, FIRE SUPPRESSION
Name: JobTitle, dtype: object
```

How much does JOSEPH DRISCOLL make (including benefits)?

```
sal[sal['EmployeeName']=='JOSEPH DRISCOLL']['TotalPayBenefits']
```

```
24    270324.91
Name: TotalPayBenefits, dtype: float64
```

What is the name of highest paid person (including benefits)?

```
sal[sal['TotalPayBenefits'] == sal['TotalPayBenefits'].max()]
```

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	T
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN	167411.18	0.0	400184.25	NaN	56

What is the name of lowest paid person (including benefits)? Do you notice something strange about how much he or she is paid?

```
sal[sal['TotalPayBenefits'] == sal['TotalPayBenefits'].min()]
```

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	T
148653	148654	Joe Lopez	Counselor, Log Cabin	0.0	0.0	-618.13	0.0	

Her salary is a negative value.

What was the average (mean) BasePay of all employees per year? (2011-2014)?

```
sal.groupby('Year').mean()['BasePay']
```

```
Year
2011    63595.956517
2012    65436.406857
2013    69630.030216
2014    66564.421924
Name: BasePay, dtype: float64
```

How many unique job titles are there?

```
sal['JobTitle'].nunique()
```

```
2159
```

What are the top 5 most common jobs?

```
sal['JobTitle'].value_counts().head()
```

```
Transit Operator                7036
Special Nurse                   4389
Registered Nurse                3736
Public Svc Aide-Public Works    2518
Police Officer 3                2421
Name: JobTitle, dtype: int64
```

How many Job Titles were represented by only one person in 2013? (e.g. Job Titles with only one occurence in 2013?)

```
sum(sal[sal['Year'] == 2013]['JobTitle'].value_counts() == 1)
```

```
202
```

How many people have the word Chief in their job title? (This is pretty tricky)

```
sal['JobTitle'].apply(lambda str:('chief' in str.lower())).sum()
```

```
627
```

Is there a correlation between length of the Job Title string and Salary?

```
sal['title_len'] = sal['JobTitle'].apply(len)
sal[['title_len','TotalPayBenefits']].corr()
```

	title_len	TotalPayBenefits
title_len	1.000000	-0.036878
TotalPayBenefits	-0.036878	1.000000

Generate a histogram plot of base salary with 20 bins?

```
import pylab as pl
sal['BasePay'].hist(bins=20)
pl.suptitle("BasePay")
```

```
Text(0.5,0.98,'BasePay')
```