Lab3

In [ ]:

```
Name: Meet Hiteshkumar Trivedi
Student ID: N01520331
```

Read the Salaries.csv into a dataframe called df_data and use the head() method to check that you have read in the data correctly. Make sure you import pandas.

In [2]:

```
import numpy as np
import pandas as pd

#Write your code here
df_data=pd.read_csv('E:\Programming\Humber college\Humber Sem 2\Data Analytics\Week-3\Datas
df_data.head()
```

Out[2]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	335279.91
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	332343.61
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	326373.19

Use the dtypes attribute to view how each column is stored

In [4]:

```
#Write your code here
df_data.dtypes
```

Out[4]:

```
Id                   int64
EmployeeName        object
JobTitle            object
BasePay            float64
OvertimePay        float64
OtherPay           float64
Benefits           float64
TotalPay           float64
TotalPayBenefits   float64
Year                 int64
Notes              float64
Agency              object
Status             float64
dtype: object
```

Slice the first two columns using .loc and store the result in a variable.

In [8]:

```
#Write you code here
res = df_data.loc[:,"Id":"EmployeeName"]
res
```

Out[8]:

	Id	EmployeeName
0	1	NATHANIEL FORD
1	2	GARY JIMENEZ
2	3	ALBERT PARDINI
3	4	CHRISTOPHER CHONG
4	5	PATRICK GARDNER
...	...	...
148649	148650	Roy I Tillery
148650	148651	Not provided
148651	148652	Not provided
148652	148653	Not provided
148653	148654	Joe Lopez

148654 rows × 2 columns

Slice the first two rows using .loc and store the result in a variable

In [9]:

```
#Write you code here
res = df_data.loc[0:1, :]
res
```

Out[9]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28

In [ ]:

```
```

Compute the average and max TotalPay. Store the results in variables called avg_TotalPay and max_TotalPay

In [19]:

```
#Write your code here
avg_TotalPay = df_data["TotalPay"].mean()
max_TotalPay = df_data["TotalPay"].max()
res = "average is {0}and max is {1}".format(avg_TotalPay,max_TotalPay)
res
```

Out[19]:

```
'average is 74768.321971703and max is 567595.43'
```

Create a column called "final", which is BasePay*2.

In [20]:

```
#Write your code here
df_data["final"] = df_data["BasePay"]*2
df_data.head()
```

Out[20]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	335279.91
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	332343.61
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	326373.19

Use the drop() method to delete the column OvertimePay from the dataframe df_data.

In [22]:

```
#Write your code here
df_data.drop(["OvertimePay"], inplace = True, axis=1)
df_data.head()
```

Out[22]:

	Id	EmployeeName	JobTitle	BasePay	OtherPay	Benefits	TotalPay	TotalPayBene
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	400184.25	NaN	567595.43	56759
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	137811.38	NaN	538909.28	53890
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	16452.60	NaN	335279.91	33527
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	198306.90	NaN	332343.61	33234
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	182234.59	NaN	326373.19	32637

In this set of practice exercises, we will be working with a demographic data regarding the passengers aboard the Titanic. Read in the data frame and use the head() method to check that it was read in correctly.

In [3]:

```
import pandas as pd
#Write your code here
df_tin = pd.read_csv('E:\Programming\Humber college\Humber Sem 2\Data Analytics\Week-3\Data
df_tin.head()
```

Out[3]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emba
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	

Use the rename method to change the column "Name" to "Passenger_Name" and the column "Ticket" to "Ticket_Num".

In [4]:

```
#Write your code here
df_tin.rename(columns={"Name":"Passenger_Name", "Ticket":"Ticket_Num"}, inplace=True)

df_tin.head()
```

Out[4]:

	PassengerId	Pclass	Passenger_Name	Sex	Age	SibSp	Parch	Ticket_Num	Fare	C
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	

Select the name of passenger 896

In [26]:

```
#Write your code here
df_tin.loc[4,"Passenger_Name"]
```

Out[26]:

```
'Hirvonen, Mrs. Alexander (Helga E Lindqvist)'
```

How many missing entries are there in the Age column?

In [28]:

```
#Write you code here
df_tin.isnull().sum()["Age"]
```

Out[28]:

```
86
```

Compute the avg age of passengers ignoring the missing data.

In [ ]:

```
#Write your code here
```

Using the fillna() method replace the missing values in the Age column with the mean.

In [5]:

```
#Write your code here
df_tin.fillna(df_tin.mean())
```

Out[5]:

	PassengerId	Pclass	Passenger_Name	Sex	Age	SibSp	Parch	Ticket_Num	
0	892	3	Kelly, Mr. James	male	34.50000	0	0	330911	7
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.00000	1	0	363272	7
2	894	2	Myles, Mr. Thomas Francis	male	62.00000	0	0	240276	9
3	895	3	Wirz, Mr. Albert	male	27.00000	0	0	315154	8
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.00000	1	1	3101298	12
...	...	...	...	...	...	...	...	...	
413	1305	3	Spector, Mr. Woolf	male	30.27259	0	0	A.5. 3236	8
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.00000	0	0	PC 17758	108
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.50000	0	0	SOTON/O.Q. 3101262	7
416	1308	3	Ware, Mr. Frederick	male	30.27259	0	0	359309	8
417	1309	3	Peter, Master. Michael J	male	30.27259	1	1	2668	22

418 rows × 11 columns

In [ ]:

```
#Bonus: for students who wants to practice more
```

What is the average age of the 5 oldest passengers? The reset_index method will be helpful here.

In [ ]:

```
#Write your code here
```

In [ ]:

```
```