Lab1

Name: Devi Sri Swetha Tanuku Student ID: N01623362

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 [6]:
```
#Write your code here
import pandas as pd
df_data = pd.read_csv('Salaries.csv')
df_data.head()
```

Out[6]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tota
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	56759
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	53890
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	33527
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	33234
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	32637

Use the dtypes attribute to view how each column is stored

In [8]:
```
#Write your code here
df_data.dtypes
```

Out[8]:
```
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 [10]:
```
#Write you code here
result = df_data.loc[0:,['Id','EmployeeName']]
result
```

Out[10]:

	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 called result_2.

In [12]:
```
#Write you code here
result_2 = df_data.loc[0:1]
result_2
```

Out[12]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tota
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	5675
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	5389

Slice the first four rows and the first five columns and store the result in a variable called result_3.

In [14]:
```
#Write you code here
result_3 = df_data.loc[0:3,'Id':'OvertimePay']
result_3
```

Out[14]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71

Slice rows 0,4,6 and select two columns randomly and store the result in variable called result_4.

In [16]:
```
#Write you code here
result_4 = df_data.loc[[0,4,6],['EmployeeName','TotalPay']]
result_4
```

Out[16]:

	EmployeeName	TotalPay
0	NATHANIEL FORD	567595.43
4	PATRICK GARDNER	326373.19
6	ALSON LEE	315981.05

Store the number rows in a variable called num_rows.

In [ ]:
```
#Write you code here
num_rows = df_data.index
num_rows
```

Print out the last row of the data to dataframe.

In [ ]:
```
#Write you code here
df_data.tail(1)
```

In [ ]:
```

```

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

In [ ]:
```
#Write your code here
avg_TotalPay = df_data.TotalPay.mean()
avg_TotalPay
```

In [ ]:
```
max_TotalPay = df_data.TotalPay.max()
max_TotalPay
```

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

In [34]:
```
#Write your code here
df_data["final"] = df_data["BasePay"]*2
df_data.head()
```

Out[34]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tota
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	56759
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	53890
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	33527
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	33234
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	32637

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

In [38]:
```
#Write your code here
df_data.drop(["OvertimePay"], axis = 1, inplace = False)
```

Out[38]:

	Id	EmployeeName	JobTitle	BasePay	OtherPay	Benefits	TotalPay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	400184.25	NaN	567595.43
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	137811.38	NaN	538909.28
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	16452.60	NaN	335279.91
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	198306.90	NaN	332343.61
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	182234.59	NaN	326373.19
...	...	...	...	...	...	...	...
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.0	0.00
148650	148651	Not provided	Not provided	NaN	NaN	NaN	0.00
148651	148652	Not provided	Not provided	NaN	NaN	NaN	0.00
148652	148653	Not provided	Not provided	NaN	NaN	NaN	0.00
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	-618.13	0.0	-618.13

148654 rows × 13 columns

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 [71]:
```
import pandas as pd
#Write your code here
df_titanic = pd.read_csv("Titanic.csv")
df_titanic.head()
```

Out[71]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Em
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 [73]:
```
#Write your code here
df_titanic.rename(columns = {"Name":"Passenger_Name","Ticket":"Ticket_Num"}, inplac
```

Out[73]:

	PassengerId	Pclass	Passenger_Name	Sex	Age	SibSp	Parch	Ticket_Num	Fa
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.82
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.00
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.68
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.66
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.28
...	...	...	...	...	...	...	...	...	
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.05
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.90
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.25
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.05
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.35

418 rows × 11 columns

Select the name of passenger 896

In [161…

```
#Write your code here
df_titanic.loc[df_titanic['PassengerId']== 896, 'Name'].values[0]
```

Out[161…

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

How many missing entries are there in the Age column?

In [77]:

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

Out[77]:

```
86
```

Compute the avg age of passengers ignoring the missing data.

In [79]:

```
#Write your code here
df_titanic.Age.mean()
```

Out[79]: 30.272590361445783

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

In [163…
```
#Write your code here
df_titanic['Age'].fillna(df_titanic['Age'].mean())
```

Out[163…
```
0      34.50000
1      47.00000
2      62.00000
3      27.00000
4      22.00000
         ...
413    30.27259
414    39.00000
415    38.50000
416    30.27259
417    30.27259
Name: Age, Length: 418, dtype: float64
```

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 [187…
```
#Write your code here
Older_Pass = df_titanic.sort_values(by='Age', ascending = False)
Indx_reset = df_titanic.reset_index(drop = True)
Top5_Older_Pass = Indx_reset.head(5)
Top5_Avarage = Top5_Older_Pass['Age'].mean()
print(Top5_Avarage)
```

```
38.5
```