Assignment 5

@author : @ruhend (Mudigonda Himansh)

AP19110010169

DATA TRANSFORMATION

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:

```
low_memory=False
```

In [4]:

```
df = pd.read_csv("Salaries.csv")
print ('dataset: %s'%(str(df.shape)))
```

```
dataset: (148654, 13)
```

In [5]:
```
df
```

Out[5]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	To
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326
...	...	...	...	...	...	...	...	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.00	
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.00	-

148654 rows × 13 columns

Decimal Scale Normalization

In [6]:
```
TotalPay = df['TotalPay']
Max = str(round(TotalPay.max()))
Len = len(Max)
df['TotalPay'] = df['TotalPay'].apply(lambda x: x/10**Len)
```

In [7]:
```
df
```

Out[7]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tot
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	0.5
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	0.5
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	0.3
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	0.3
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	0.3
...	...	...	...	...	...	...	...	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.00	0.0
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.00	-0.0

148654 rows × 13 columns

Min-Max Normalization

In [8]:
```
TotalPayBenefits = df['TotalPayBenefits']
Min = TotalPayBenefits.min()
Max = TotalPayBenefits.max()
Diff = Max-Min
df['TotalPayBenefits'] = df['TotalPayBenefits'].apply(lambda x: (x-Min)/ Di
```

In [9]:
```
df
```

Out[9]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tot
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	0.5
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	0.5
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	0.3
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	0.3
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	0.3
...	...	...	...	...	...	...	...	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.00	0.0
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.00	-0.0

148654 rows × 13 columns

Z-Score Normalization

In [11]:
```
df = pd.read_csv("Salaries.csv")
print ('dataset: %s'%(str(df.shape)))
```

```
dataset: (148654, 13)
```

In [12]:
```
TotalPay = df['TotalPay']
mean = TotalPay.mean()
std = TotalPay.std()
df['TotalPay'] = df['TotalPay'].apply(lambda x: (x-mean)/ std)
```

In [13]:
```
df
```

Out[13]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tot
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	9.7
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	9.1
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	5.1
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	5.0
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	4.9
...	...	...	...	...	...	...	...	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.00	-1.4
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	-1.4
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	-1.4
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	-1.4
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.00	-1.4

148654 rows × 13 columns