**LETSUPGRADE ASSIGNMENT**

**ATTRITION**

**STEP: 1**

Launching the packages like pandas, numpy , matplotlib.

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

**STEP: 2**

dataset1=pd.read\_excel("3. Descriptive Statistics.xlsx",sheet\_name=0)

dataset1.head()

Out[5]:

ID Gender Birth Date ... Job Time Prev Exep Minority

0 3 f 07/26/1929 ... 98 381 0

1 4 f 04/15/1947 ... 98 190 0

2 8 f 1966-06-05 00:00:00 ... 98 0 0

3 9 f 01/23/1946 ... 98 115 0

4 10 f 02/13/1946 ... 98 244 0

[5 rows x 11 columns]

dataset1.tail()

Out[6]:

ID Gender Birth Date ... Job Time Prev Exep Minority

469 464 m 03/20/1962 ... 64 27 0

470 465 m 07/20/1962 ... 64 106 0

471 470 m 01/22/1964 ... 64 69 1

472 471 m 1966-03-08 00:00:00 ... 64 32 1

473 472 m 02/21/1966 ... 63 46 0

[5 rows x 11 columns]

dataset1.info

Out[7]:

<bound method DataFrame.info of ID Gender Birth Date ... Job Time Prev Exep Minority

0 3 f 07/26/1929 ... 98 381 0

1 4 f 04/15/1947 ... 98 190 0

2 8 f 1966-06-05 00:00:00 ... 98 0 0

3 9 f 01/23/1946 ... 98 115 0

4 10 f 02/13/1946 ... 98 244 0

.. ... ... ... ... ... ... ...

469 464 m 03/20/1962 ... 64 27 0

470 465 m 07/20/1962 ... 64 106 0

471 470 m 01/22/1964 ... 64 69 1

472 471 m 1966-03-08 00:00:00 ... 64 32 1

473 472 m 02/21/1966 ... 63 46 0

[474 rows x 11 columns]>

dataset1.columns

Out[8]:

Index(['ID', 'Gender', 'Birth Date', 'Education', 'JobCategory',

'CurrentSalary', 'After6Months', 'SalBegin', 'Job Time', 'Prev Exep',

'Minority'],

dtype='object')

dataset1.index

Out[9]: RangeIndex(start=0, stop=474, step=1)

**STEP: 3**

dataset1.isnull()

Out[10]:

ID Gender Birth Date ... Job Time Prev Exep Minority

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

.. ... ... ... ... ... ... ...

469 False False False ... False False False

470 False False False ... False False False

471 False False False ... False False False

472 False False False ... False False False

473 False False False ... False False False

[474 rows x 11 columns]

dataset1.duplicated()

Out[11]:

0 False

1 False

2 False

3 False

4 False

469 False

470 False

471 False

472 False

473 False

Length: 474, dtype: bool

dataset1.drop\_duplicates()

Out[12]:

ID Gender Birth Date ... Job Time Prev Exep Minority

0 3 f 07/26/1929 ... 98 381 0

1 4 f 04/15/1947 ... 98 190 0

2 8 f 1966-06-05 00:00:00 ... 98 0 0

3 9 f 01/23/1946 ... 98 115 0

4 10 f 02/13/1946 ... 98 244 0

.. ... ... ... ... ... ... ...

469 464 m 03/20/1962 ... 64 27 0

470 465 m 07/20/1962 ... 64 106 0

471 470 m 01/22/1964 ... 64 69 1

472 471 m 1966-03-08 00:00:00 ... 64 32 1

473 472 m 02/21/1966 ... 63 46 0

**STEP: 4**

Statistical Functions

dataset2=dataset1[['ID', 'Gender', 'Birth Date', 'Education', 'JobCategory','CurrentSalary', 'After6Months', 'SalBegin', 'Job Time', 'Prev Exep','Minority']].mean()

dataset2

Out[15]:

ID 237.500000

Education 13.491561

JobCategory 1.411392

CurrentSalary 34419.567511

After6Months 25717.827004

SalBegin 17016.086498

Job Time 81.109705

Prev Exep 95.860759

Minority 0.219409

dtype: float64

dataset2=dataset1[['ID', 'Gender', 'Birth Date', 'Education', 'JobCategory','CurrentSalary', 'After6Months', 'SalBegin', 'Job Time', 'Prev Exep','Minority']].median()

dataset2

Out[17]:

ID 237.5

Education 12.0

JobCategory 1.0

CurrentSalary 28875.0

After6Months 21900.0

SalBegin 15000.0

Job Time 81.0

Prev Exep 55.0

Minority 0.0

dtype: float64

dataset2=dataset1[['ID', 'Gender', 'Birth Date', 'Education', 'JobCategory','CurrentSalary', 'After6Months', 'SalBegin', 'Job Time', 'Prev Exep','Minority']].var()

dataset2

Out[20]:

ID 1.876250e+04

Education 8.322339e+00

JobCategory 5.978403e-01

CurrentSalary 2.915782e+08

After6Months 1.475236e+08

SalBegin 6.194694e+07

Job Time 1.012226e+02

Prev Exep 1.093828e+04

Minority 1.716309e-01

dtype: float64

dataset2=dataset1[['ID', 'Gender', 'Birth Date', 'Education', 'JobCategory','CurrentSalary', 'After6Months', 'SalBegin', 'Job Time', 'Prev Exep','Minority']].std()

dataset2

Out[22]:

ID 136.976275

Education 2.884846

JobCategory 0.773201

CurrentSalary 17075.661465

After6Months 12145.928474

SalBegin 7870.638154

Job Time 10.060945

Prev Exep 104.586236

Minority 0.414284

dtype: float64

dataset2=dataset1[['ID', 'Gender', 'Birth Date', 'Education', 'JobCategory','CurrentSalary', 'After6Months', 'SalBegin', 'Job Time', 'Prev Exep','Minority']].skew()

dataset2

Out[24]:

ID 0.000000

Education -0.114107

JobCategory 1.455978

CurrentSalary 2.124606

After6Months 2.267649

SalBegin 2.852856

Job Time -0.052570

Prev Exep 1.509984

Minority 1.360322

dtype: float64

dataset2=dataset1[['ID', 'Gender', 'Birth Date', 'Education', 'JobCategory','CurrentSalary', 'After6Months', 'SalBegin', 'Job Time', 'Prev Exep','Minority']].kurt()

dataset2

Out[26]:

ID -1.200000

Education -0.265000

JobCategory 0.267547

CurrentSalary 5.377822

After6Months 6.823250

SalBegin 12.390215

Job Time -1.152594

Prev Exep 1.695953

Minority -0.150174

dtype: float64

dataset1.describe()

Out[27]:

ID Education JobCategory ... Job Time Prev Exep Minority

count 474.000000 474.000000 474.000000 ... 474.000000 474.000000 474.000000

mean 237.500000 13.491561 1.411392 ... 81.109705 95.860759 0.219409

std 136.976275 2.884846 0.773201 ... 10.060945 104.586236 0.414284

min 1.000000 8.000000 1.000000 ... 63.000000 0.000000 0.000000

25% 119.250000 12.000000 1.000000 ... 72.000000 19.250000 0.000000

50% 237.500000 12.000000 1.000000 ... 81.000000 55.000000 0.000000

75% 355.750000 15.000000 1.000000 ... 90.000000 138.750000 0.000000

max 474.000000 21.000000 3.000000 ... 98.000000 476.000000 1.000000

[8 rows x 9 columns]

**STEP: 4**

Outliers

plt.boxplot(dataset1.CurrentSalary)

Out[31]:

{'whiskers': [<matplotlib.lines.Line2D at 0x1c11d8fa508>,

<matplotlib.lines.Line2D at 0x1c11dfa0d88>],

'caps': [<matplotlib.lines.Line2D at 0x1c11dfa0e88>,

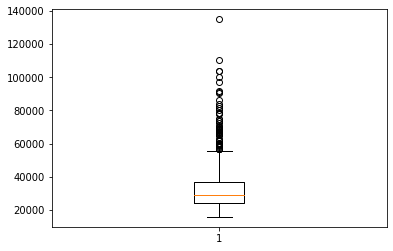
<matplotlib.lines.Line2D at 0x1c11dfa0f48>],

'boxes': [<matplotlib.lines.Line2D at 0x1c11df91f08>],

'medians': [<matplotlib.lines.Line2D at 0x1c11dfaef08>],

'fliers': [<matplotlib.lines.Line2D at 0x1c11dfaebc8>],

'means': []}



a=dataset1.After6Months

a

Out[36]:

0 16725.0

1 17550.0

2 15825.0

3 20325.0

4 18750.0

469 40275.0

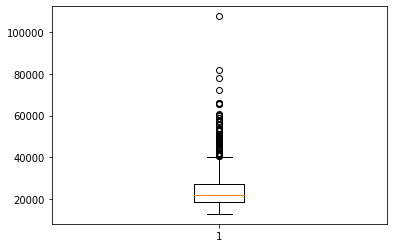
470 25200.0

471 21000.0

472 21075.0

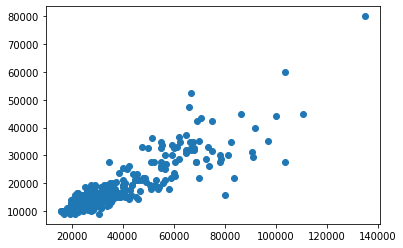
473 27450.0

Name: After6Months, Length: 474, dtype: float64



**plt.scatter(dataset1.CurrentSalary,dataset1.SalBegin)**

**Out[45]: <matplotlib.collections.PathCollection at 0x1c11f207ac8>**

****