eda

April 26, 2025

[4]:
```
# Perfoming eda to the dataset is needed to understand its behaviour
```

[4]:
```
# import the library
import pandas as pd
```

[5]:
```
# Read data into Python
education = pd.read_csv(r"D:\kumar\learning code\eda\Data Sets\education.csv")
```

[6]:
```
#check the info of the dataset know simple information
education.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 773 entries, 0 to 772
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   datasrno  773 non-null    int64
 1   workex    773 non-null    int64
 2   gmat      773 non-null    int64
dtypes: int64(3)
memory usage: 18.2 KB
```

[7]:
```
# see the first 5 rows to know how the data is
education.head()
```

[7]:
```
   datasrno  workex  gmat
0         1      21   720
1         2     107   640
2         3      57   740
3         4      99   690
4         5     208   710
```

[8]:
```
#eda first business moment , measure of central tendancy
print(education.workex.mean())
print(education.workex.median())
print(education.workex.mode())
```

```
57.501940491591206
52.0
0    45
Name: workex, dtype: int64
```

[9]:

```
# get the mode from another library scipy(scientific calc)
from scipy import stats
```

[10]:

```
# get mode of worex
stats.mode(education["workex"])
#or
stats.mode(education.workex)
```

[10]:

```
ModeResult(mode=np.int64(45), count=np.int64(60))
```

[11]:

```
#second business moment
print("variance",education.workex.var())
print(education.workex.std())
range = max(education.workex)-min(education.workex)
print("range", range)
```

```
variance 750.0378848306511
27.386819545734973
range 270
```

[12]:

```
!pip install matplotlib
```

```
Requirement already satisfied: matplotlib in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages
(3.10.1)
Requirement already satisfied: contourpy>=1.0.1 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (1.3.2)
Requirement already satisfied: cycler>=0.10 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (4.57.0)
Requirement already satisfied: kiwisolver>=1.3.1 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (1.4.8)
Requirement already satisfied: numpy>=1.23 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (2.2.4)
Requirement already satisfied: packaging>=20.0 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (24.2)
```

```
Requirement already satisfied: pillow>=8 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (11.2.1)
Requirement already satisfied: pyparsing>=2.3.1 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (3.2.3)
Requirement already satisfied: python-dateutil>=2.7 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
matplotlib) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in
c:\users\priya\appdata\local\programs\python\python313\lib\site-packages (from
python-dateutil>=2.7->matplotlib) (1.17.0)


[notice] A new release of pip is available: 24.3.1 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

[13]:

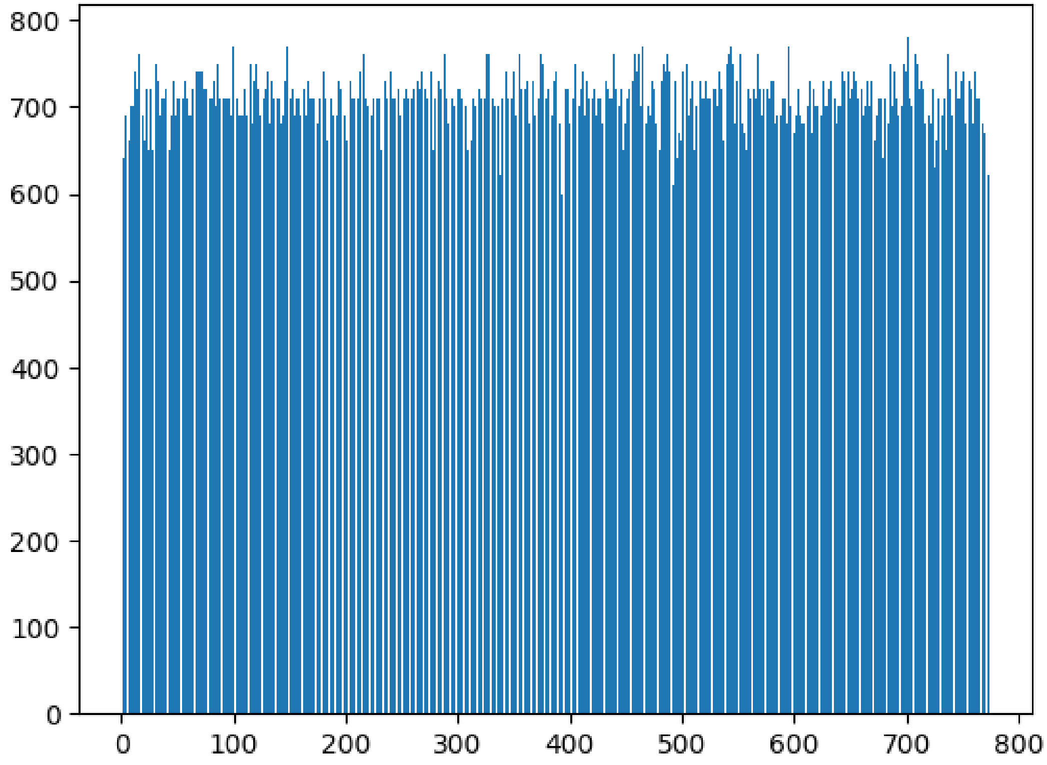
```python
# create bar plot for know skewness diagram

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

[14]:

```python
plt.bar(height = education.gmat, x = np.arange(1,774,1))
```

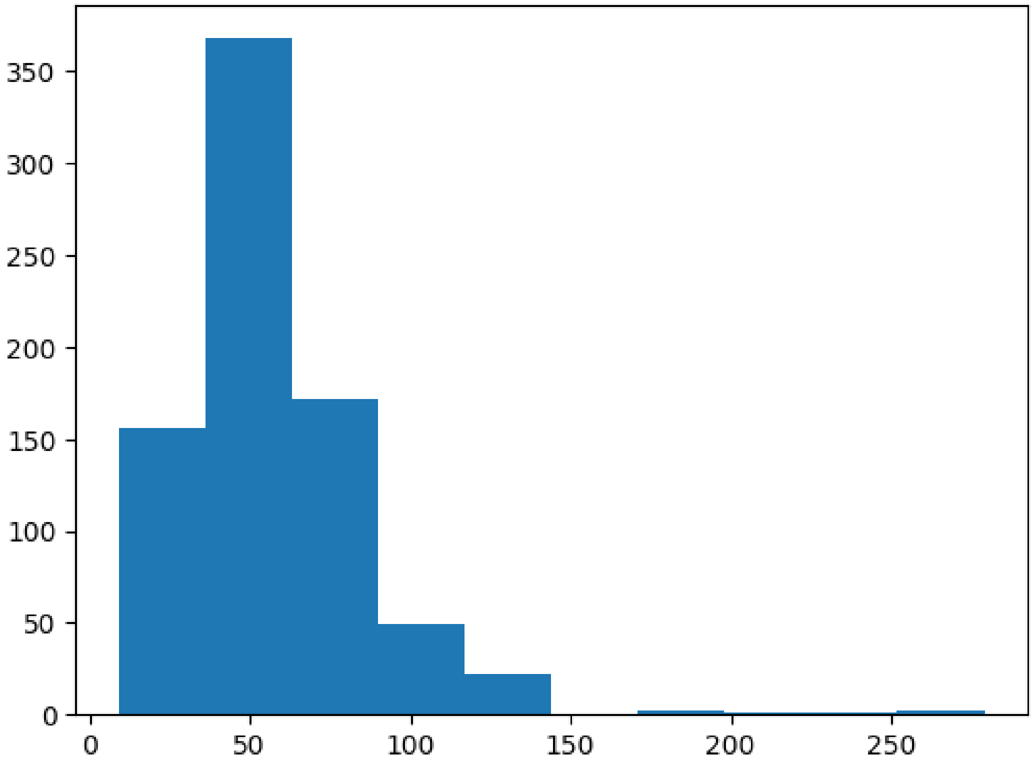
[14]:

```
<BarContainer object of 773 artists>
```



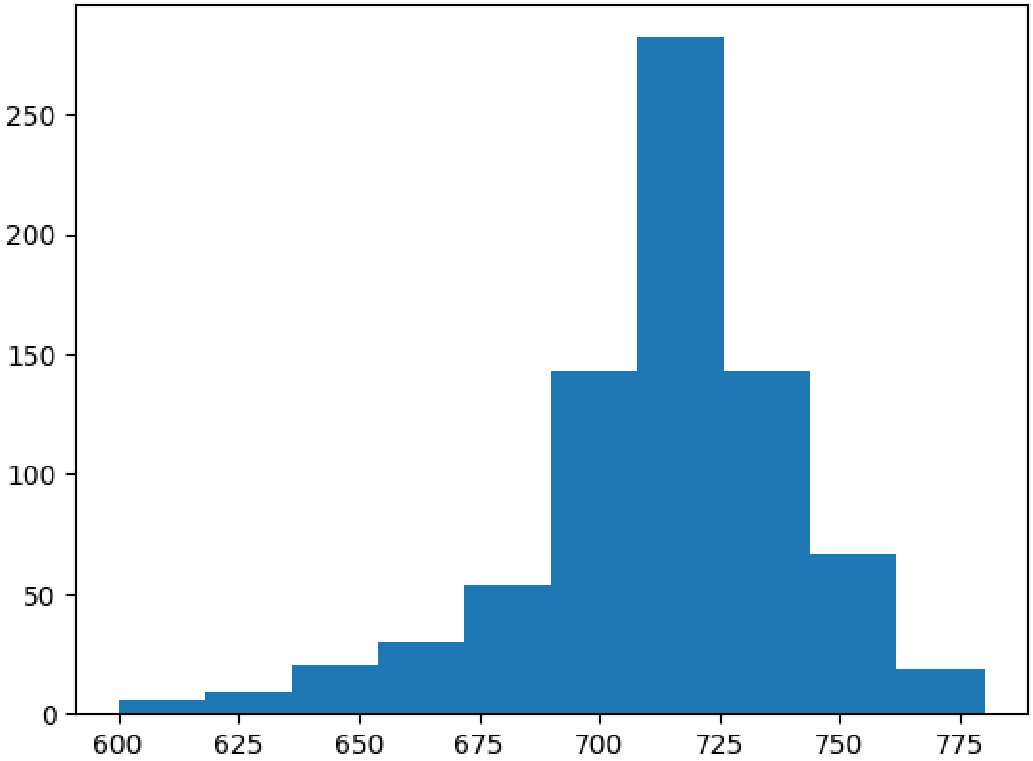
```
[15]: plt.hist(education.workex)
```

```
[15]: (array([156., 368., 172.,  49.,  22.,   0.,   2.,   1.,   1.,   2.]),
       array([  9.,  36.,  63.,  90., 117., 144., 171., 198., 225., 252., 279.]),
       <BarContainer object of 10 artists>)
```



```
[16]: plt.hist(education.gmat)
```

```
[16]: (array([  6.,   9.,  20.,  30.,  54., 143., 282., 143.,  67.,  19.]),
       array([600., 618., 636., 654., 672., 690., 708., 726., 744., 762., 780.]),
       <BarContainer object of 10 artists>)
```



[17]:

```
#third moment bussiness decision
print(education.workex.skew())
print(education.gmat.skew())
```

```
2.6085365678230614
-0.5954765248452923
```

[ ]:

```
# workex is right or positively skewd with big madnitude ...so we understand␣
  ↪data is non normal,
#in the case of gmat data is negatively or left skewd but near to normal␣
  ↪distribution
```

[18]:

```
# fourth moment business decision
education.workex.kurt()
```

[18]:

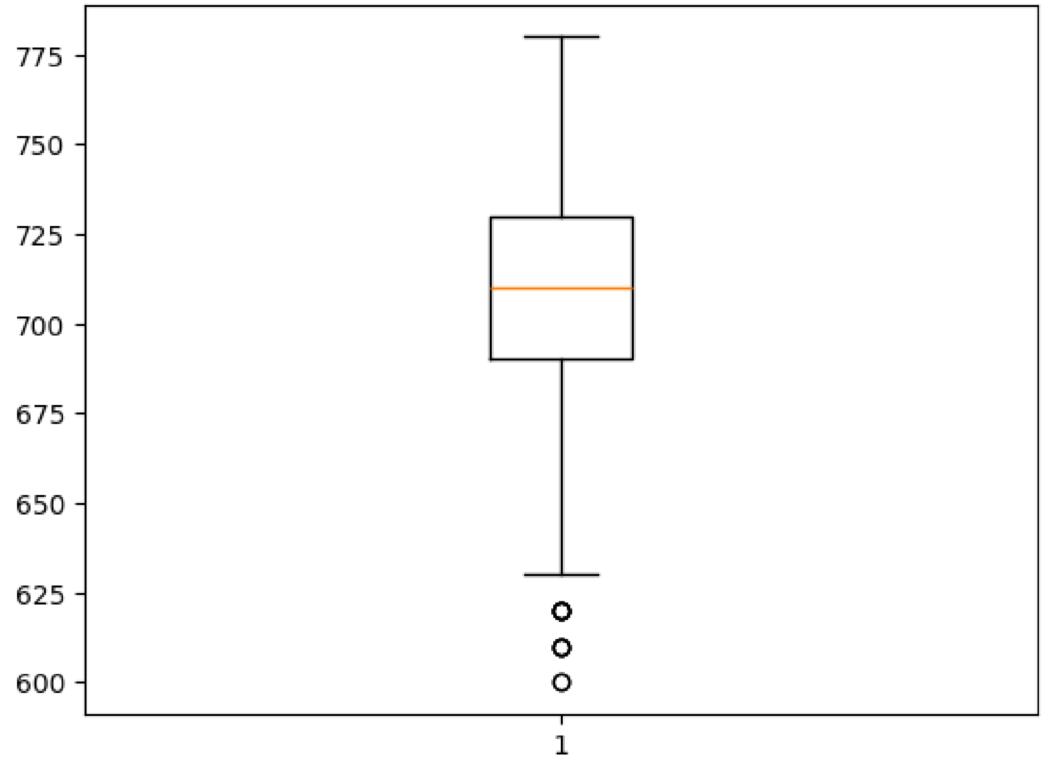
```
np.float64(13.404731601273921)
```

[22]:

```
# create boxplot to know for outliers
print(plt.boxplot(education.gmat))
```

```
{'whiskers': [<matplotlib.lines.Line2D object at 0x0000024D0188E490>,
<matplotlib.lines.Line2D object at 0x0000024D0188E5D0>], 'caps':
```

```
[<matplotlib.lines.Line2D object at 0x0000024D0188E710>,
<matplotlib.lines.Line2D object at 0x0000024D0188E850>], 'boxes':
[<matplotlib.lines.Line2D object at 0x0000024D0188E350>], 'medians':
[<matplotlib.lines.Line2D object at 0x0000024D0188E990>], 'fliers':
[<matplotlib.lines.Line2D object at 0x0000024D0188EAD0>], 'means': []}
```



```
[ ]: # by looking the plot there are some outliers , need to perform data pre␣
     ↪processing to remove them
```