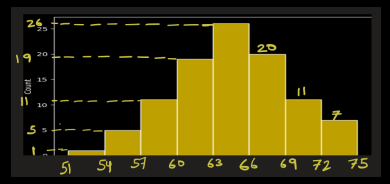
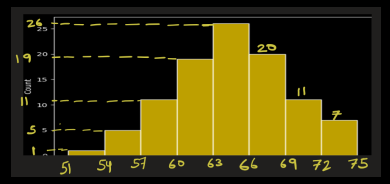


Histogram

Data = 51, 55, 55, 55, 56, 56, 57, 57, 57, 57, 58, 58, 59, 59, 59, 59, 59, 60, 60, 60, 61, 61, 61, 61, 61, 61, 62, 62, 62, 62, 62, 62, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 63, 63, 63, 63, 63, 64, 64, 64, 64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 67, 67, 67, 68, 68, 68, 68, 69, 69, 69, 69, 69, 69, 69, 70, 70, 70, 71, 72, 72, 72, 72, 72, 72, 75

- Histogram turns **raw numbers into a shape** that helps us “see” the data. It is tool to perform univariate analysis
- A histogram groups data into bins.
- The height of each bar shows frequency or density.
- We use histogram
  - to check frequency or count of data between some range (say from 60 – 65)
  - to identify if data distribution is ****normal, skewed, uniform, bimodal,**** etc. In other words, histogram shows how data values are spread across different intervals (bins).
  - to compare multiple groups (e.g. scores of students from 2 schools)
  - to compare variance of 2 data (e.g. to find which of 2 stocks is more volatile (riskier))
  - to detect **outliers or anomalies** (for e.g. to detect fraud in banks)



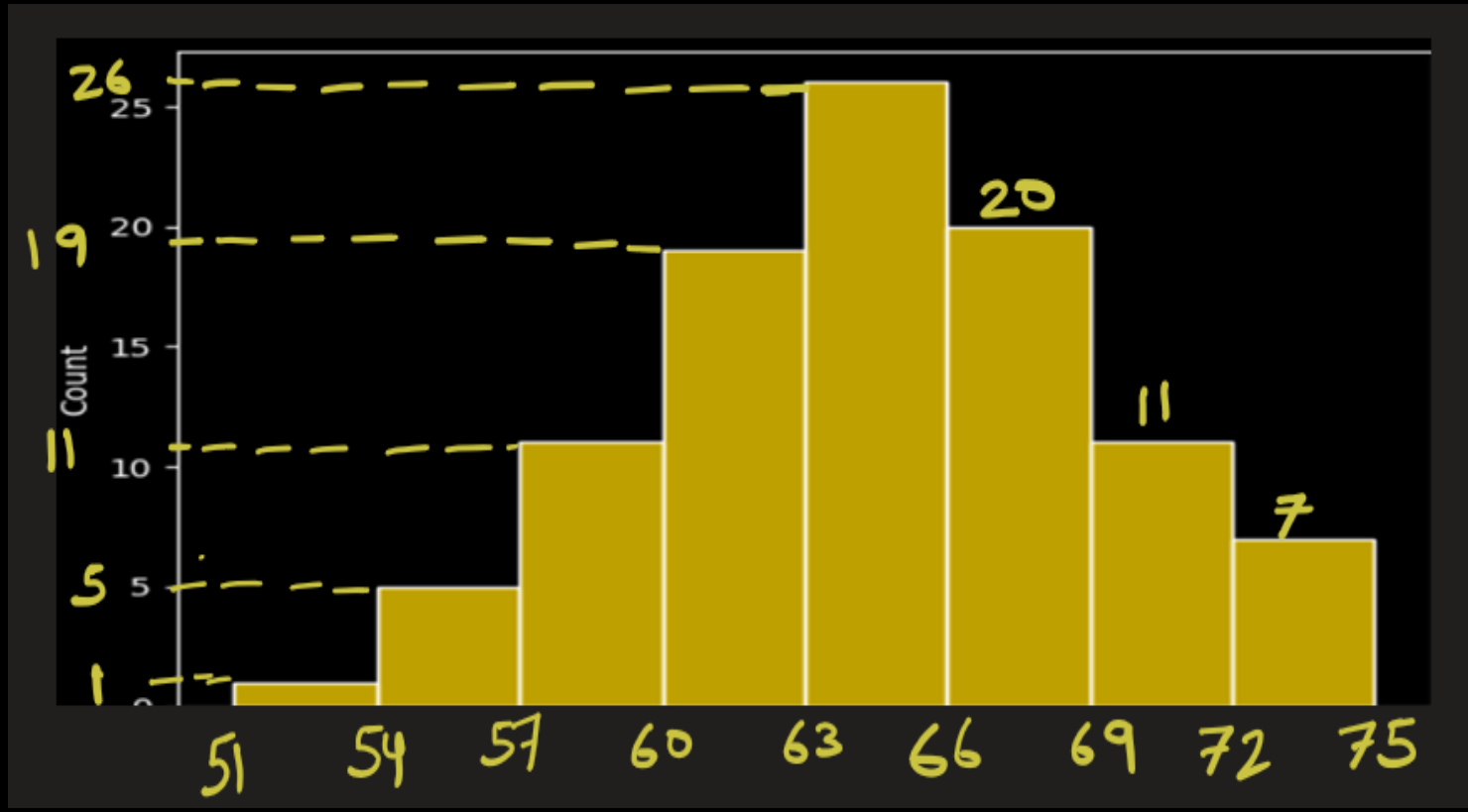
Histogram

Data = **51**, 55, 55, 55, 56, 56, 57, 57, 57, 57, 58, 58, 59, 59, 59, 59, 59, 60, 60, 60, 61, 61, 61, 61, 61, 61, 62, 62, 62, 62, 62, 62, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 63, 63, 63, 63, 63, 64, 64, 64, 64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 67, 67, 67, 68, 68, 68, 68, 69, 69, 69, 69, 69, 69, 69, 70, 70, 70, 71, 72, 72, 72, 72, 72, 72, **75**

Step1) First decide how many bins.
I want 8 bins. So, bin width
= (max-min) / bins
= (75 - 51) / 8 = **3**

Step2) Calculate bin edges:
[51, 54, 57, 60, 63, 66, 69, 72, 75]
The bins are [51,54), [54,57), [57,60),…

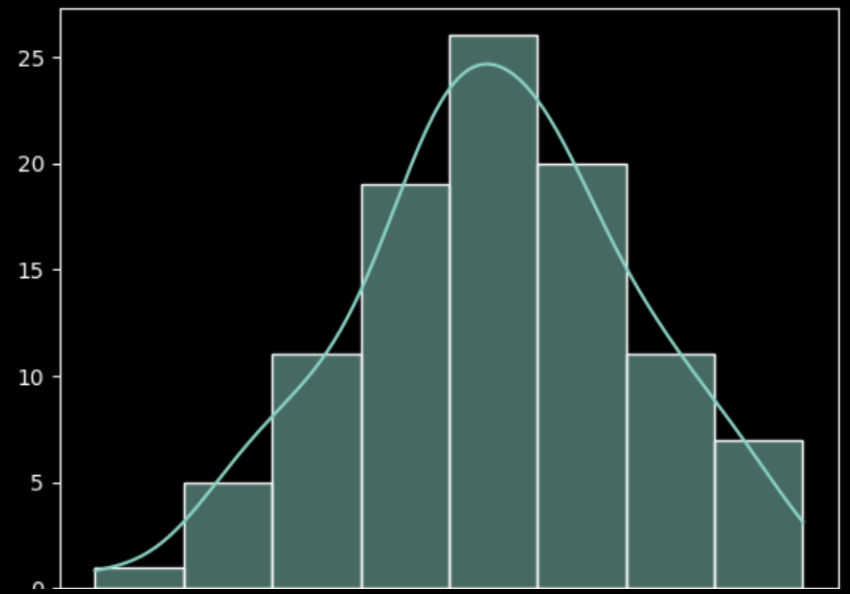
Step3) From the data, find bin counts:
bins= [51,54), [54,57), [57,60), …
counts= [1, 5, 11, 19, 26, 20, 11, 7]



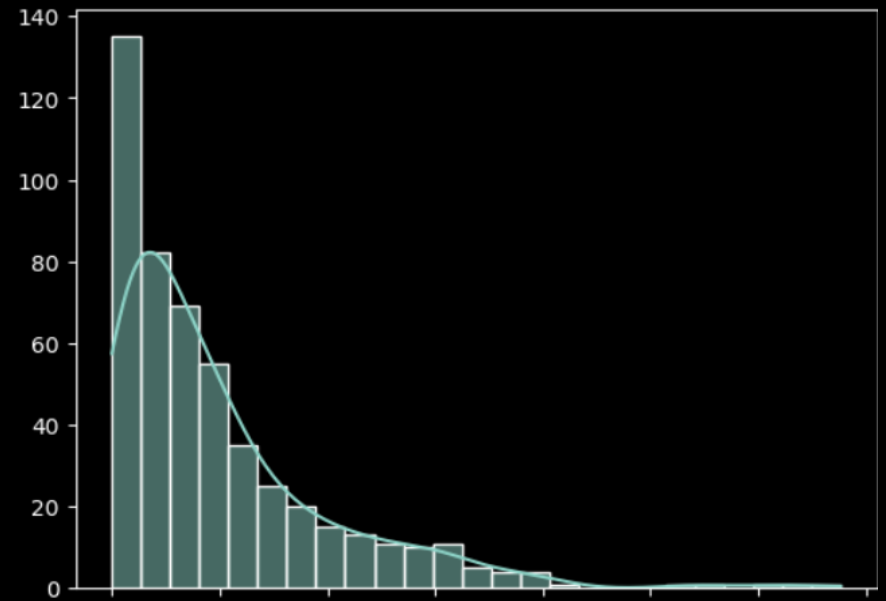
Uses of histogram

Histogram helps in identifying type of distribution

Normal

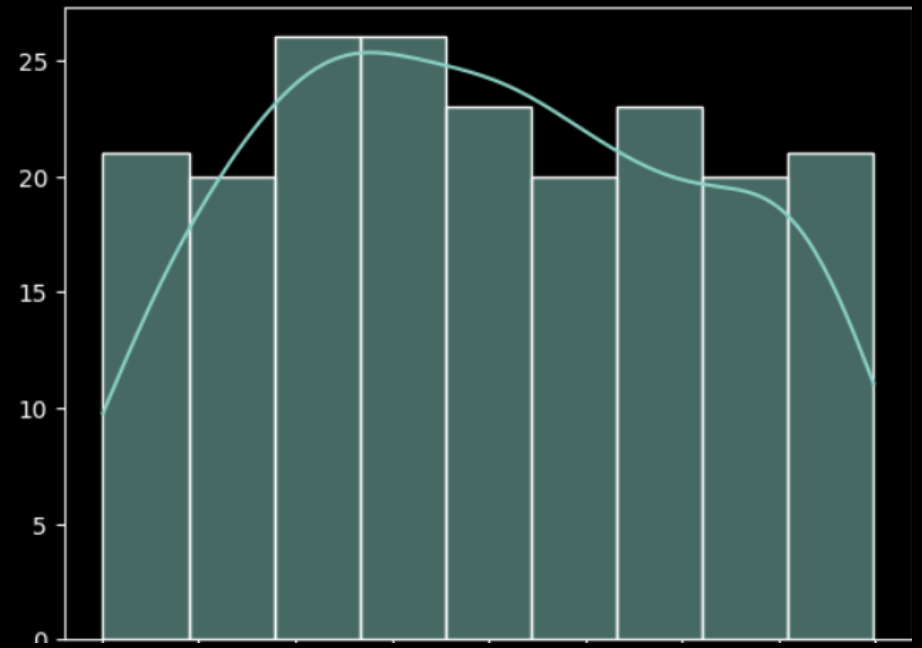


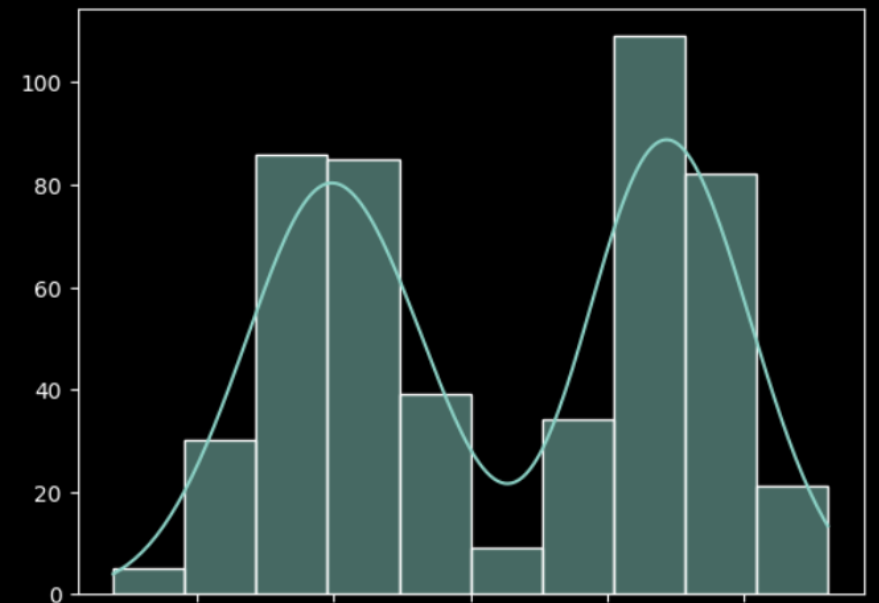
Skewed Distribution



Right-Skewed

Uniform



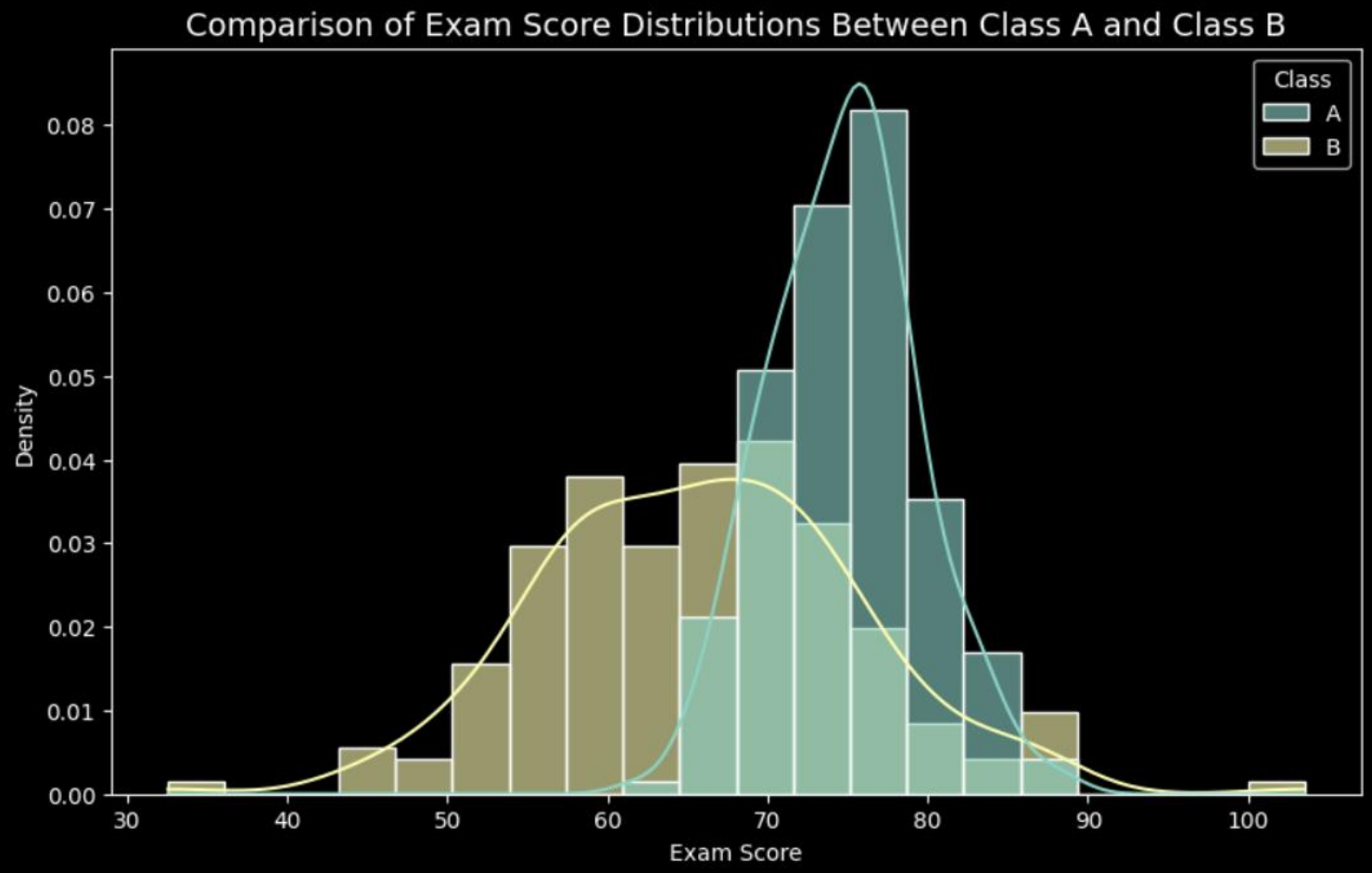


Bimodal

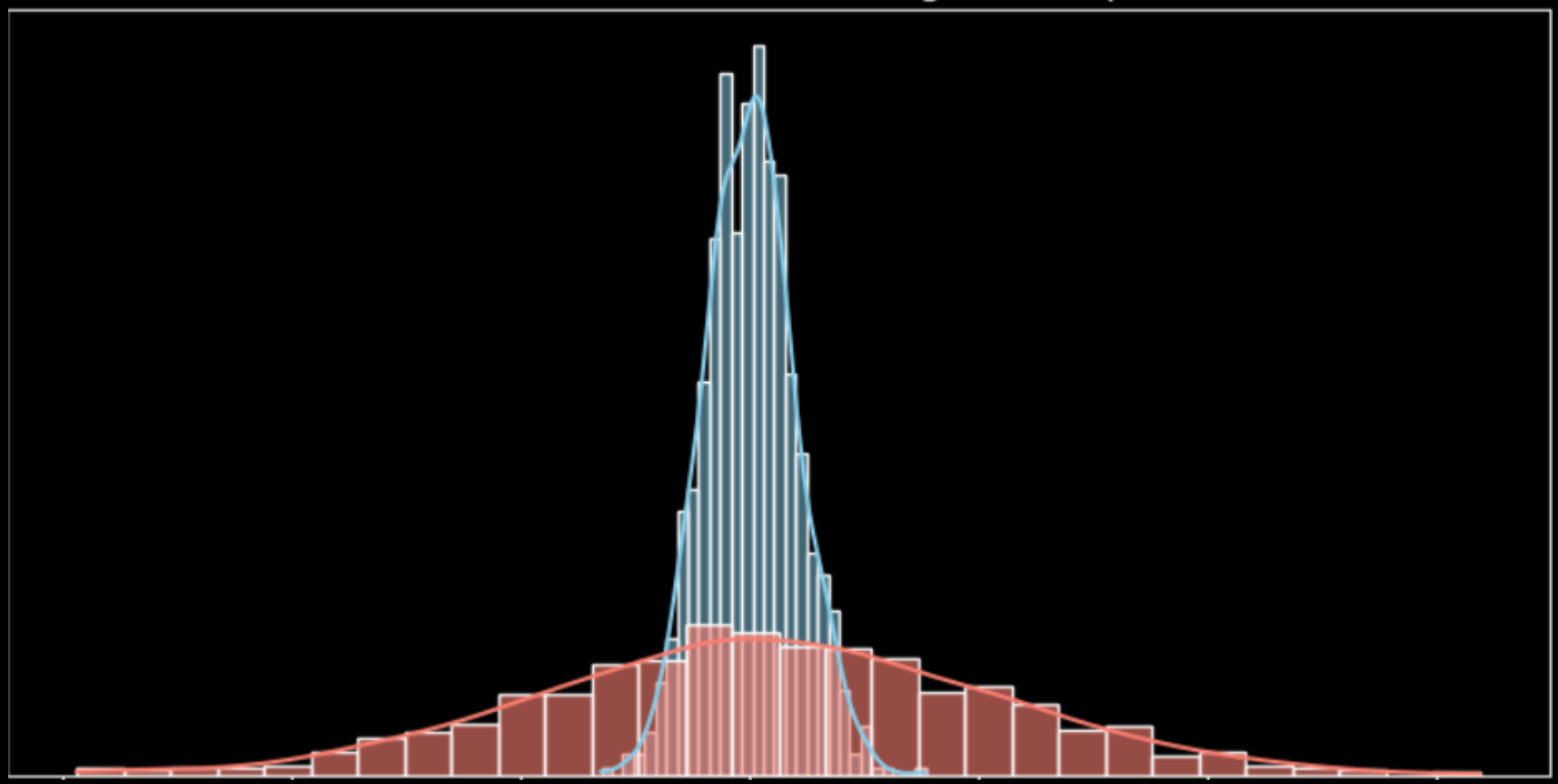
Histogram helps to compare multiple groups

For example, on right is the histogram plot of exam score for class A and class B.

- Class A (blue curve) is centered around 75 and tightly grouped → high-performing, consistent students.
- Class B (orange curve) is centered around 65 and more spread → lower mean, higher variation.
- Both KDE curves make the comparison smooth and intuitive.



Histogram allows you to compare variance of 2 data

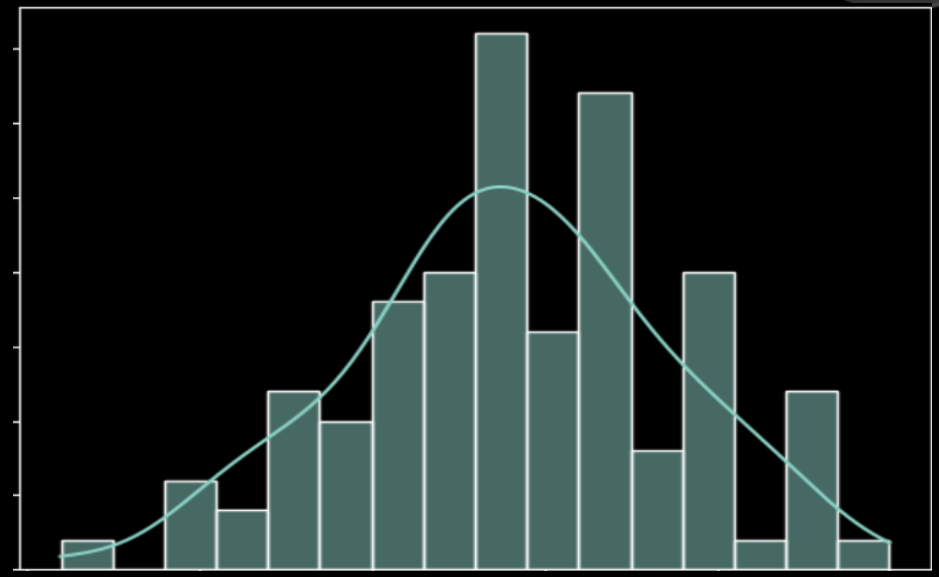


A narrow, tall histogram → low variance (data tightly clustered around mean)
A wide, flat histogram → high variance (data widely spread around mean)
The red curve has more variance

Histogram help to detect Outliers or Anomalies

Data

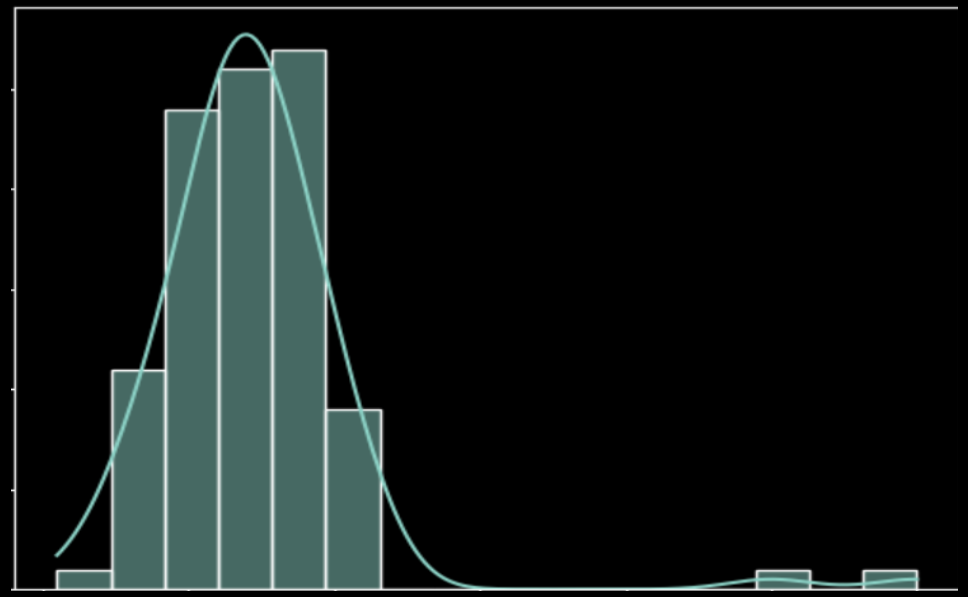
51, 55, 55, 55, 56, 56, 57, 57, 57, 57, 58, 58, 59, 59, 59, 59, 59,
60, 60, 60, 61, 61, 61, 61, 61, 61, 62, 62, 62, 62, 62, 62, 62, 62,
62, 62, 63, 63, 63, 63, 63, 63, 63, 63, 63, 63, 63, 64, 64, 64,
64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 66, 66, 66, 66,
66, 66, 66, 66, 66, 66, 66, 67, 67, 67, 68, 68, 68, 68, 69, 69, 69,
69, 69, 69, 69, 70, 70, 70, 71, 72, 72, 72, 72, 72, 72, 75



Now modify last 2 elements with outliers and plot the histogram.

Data with outliers

51, 55, 55, 55, 56, 56, 57, 57, 57, 57, 58, 58, 59, 59, 59, 59, 59,
60, 60, 60, 61, 61, 61, 61, 61, 61, 62, 62, 62, 62, 62, 62, 62, 62,
62, 62, 63, 63, 63, 63, 63, 63, 63, 63, 63, 63, 63, 64, 64, 64,
64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 66, 66, 66, 66,
66, 66, 66, 66, 66, 66, 66, 67, 67, 67, 68, 68, 68, 68, 69, 69, 69,
69, 69, 69, 69, 70, 70, 70, 71, 72, 72, 72, 72, 72, **110, 100**



Hfdslfds

fksdljfsD|]

fkdslfa