Data Visualization using BoxPlot and Outlier Detection

By

Dr. Prashant Singh Rana

Computer Science and Engineering Department

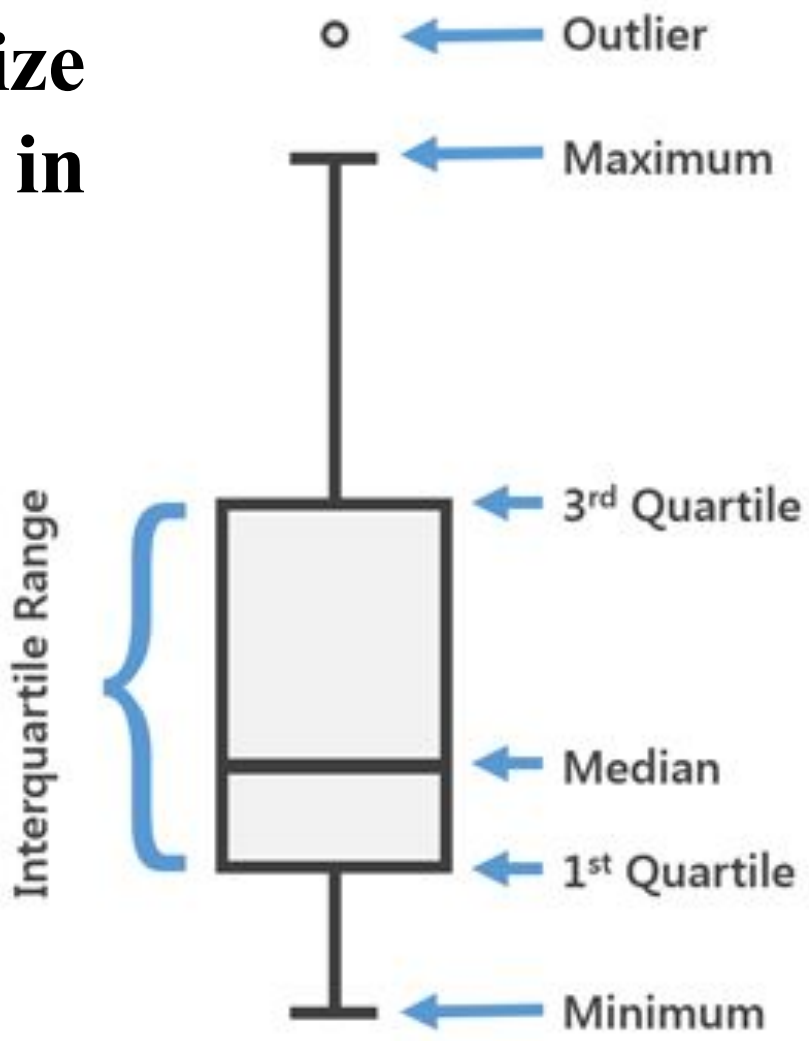
Thapar Institute of Engineering & Technology

Patiala, Punjab.

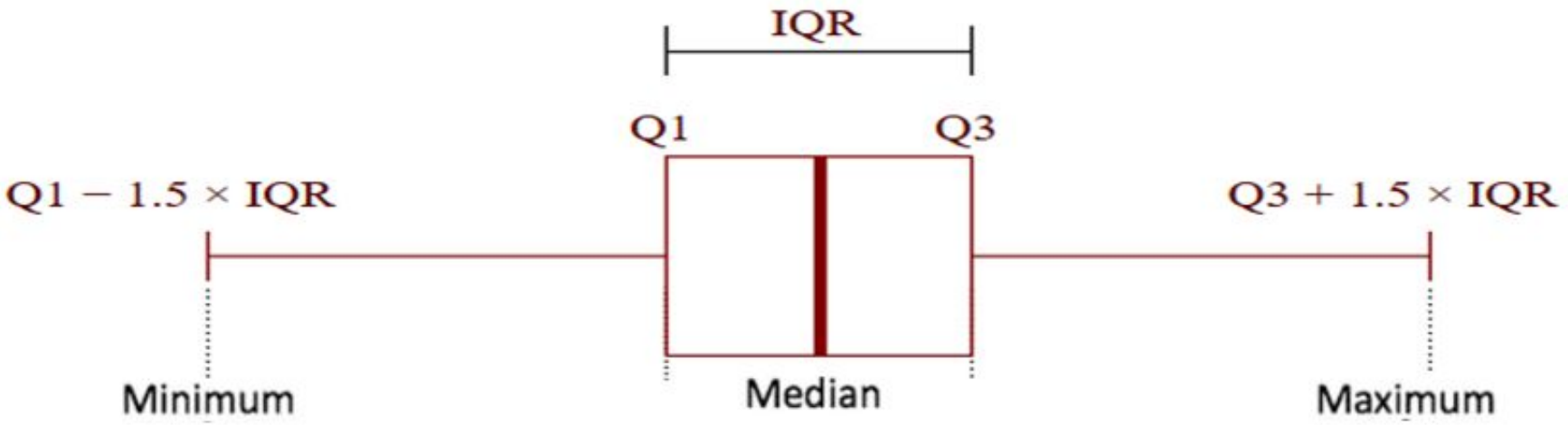
www.psrana.com

Box Plot

Box plots is used to visualize one or many distributions in a dataset.



Steps



1. Sort the list
2. Find Q1 (25^{th} %), Q3, (75^{th} %), IQR (Q3-Q1)
3. Find Min (Q1 – 1.5 × IQR), Max (Q3 + 1.5 × IQR), Median
4. Plot each values on Plot along with outliers

Example

Draw the box plot for given data

1. Sort the list
2. Find Q1, Q3, IQR
3. Find Min, Max, Median
4. Plot
5. Find the number of Outliers ?

Runs	Accuracy
1	90
2	51
3	68
4	53
5	51
6	41
7	60
8	2
9	44
10	55

Example

1. Sort the list
2. Q1=(41+44)/2 = 42.5
3. Q3=(55+60)/2 = 57.5
4. IQR = (57.5 - 42.5) = 15
5. Min = 42.5 – 1.5 × 15 =20
6. Max = 57.5 + 1.5×15 = 80
7. Median = (51+53)/2 = 52
8. Plot
9. Find the number of Outliers ?

Runs	Accuracy
1	2
2	41
3	44
4	51
5	51
6	53
7	55
8	60
9	68
10	90

Question

Draw the box plot for given data

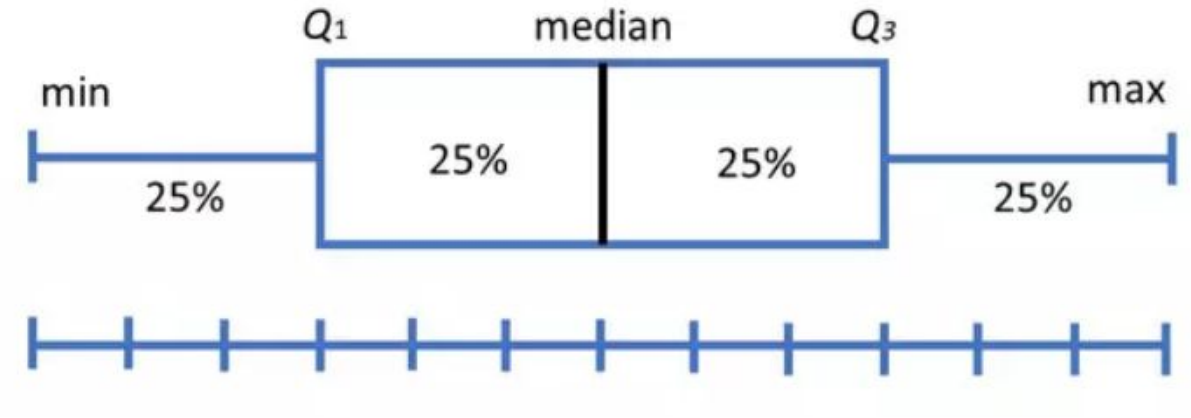
1. Sort the list
2. Find Q1, Q3, IQR
3. Find Min, Max, Median
4. Plot
5. Find the number of Outliers ?

Runs	Accuracy
1	28
2	30
3	52
4	60
5	63
6	67
7	81
8	84
9	86
10	88

Why Box Plot is useful ???

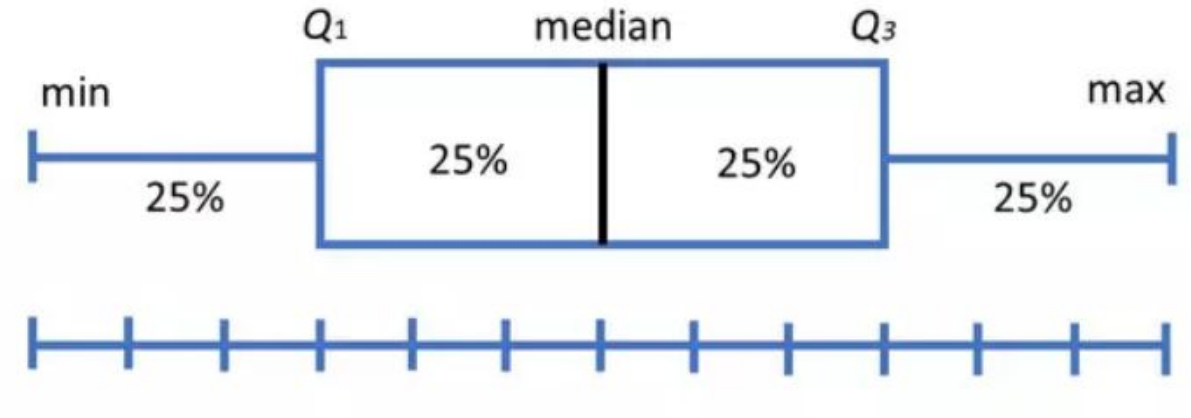
Why Box Plot is useful ???

Box plots divide the data into sections that each contain approximately 25% of the data in that set.



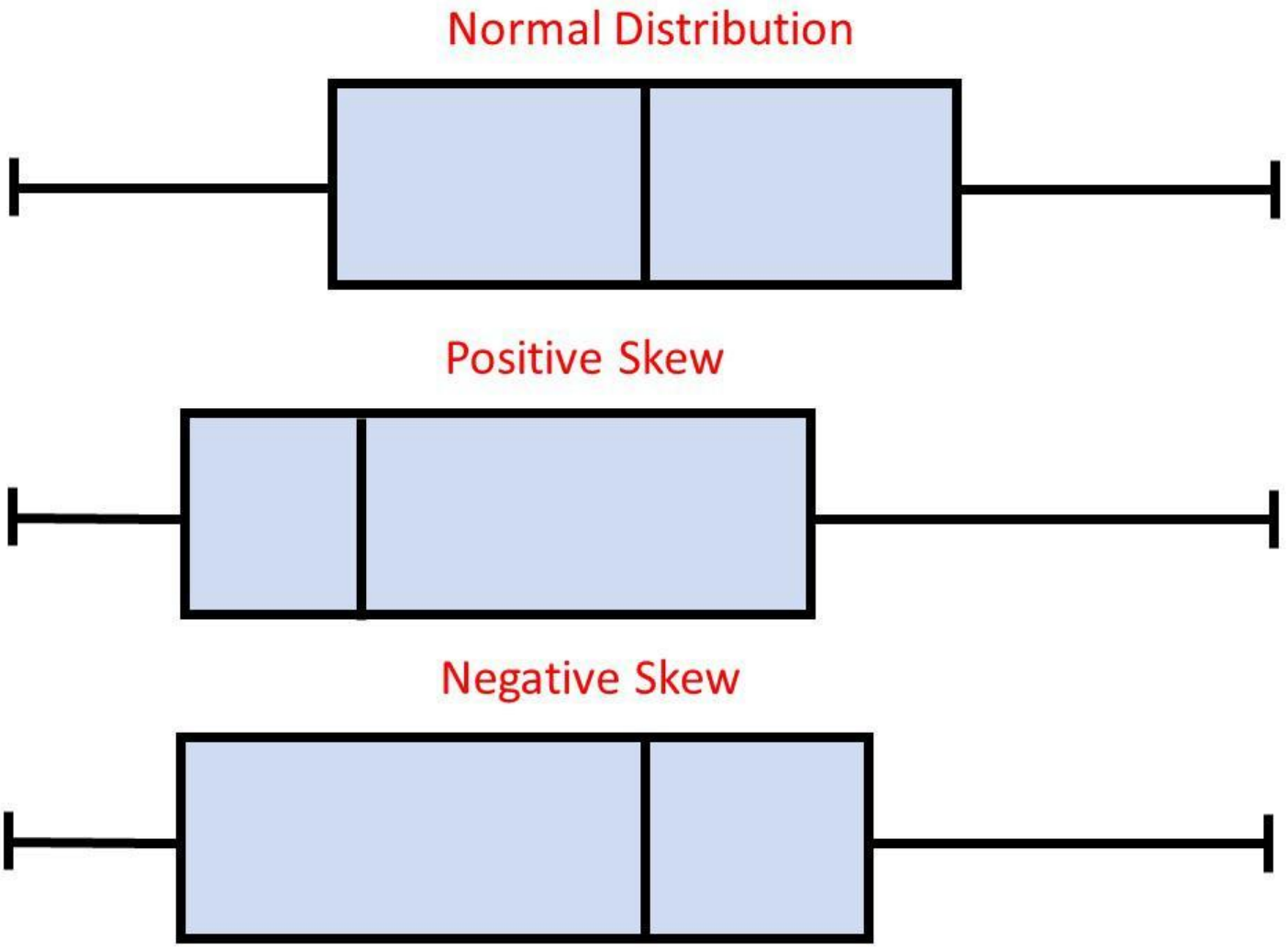
Why Box Plot is useful ???

Box plots divide the data into sections that each contain approximately 25% of the data in that set.



Box plots are useful as they provide a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set, and signs of skewness.

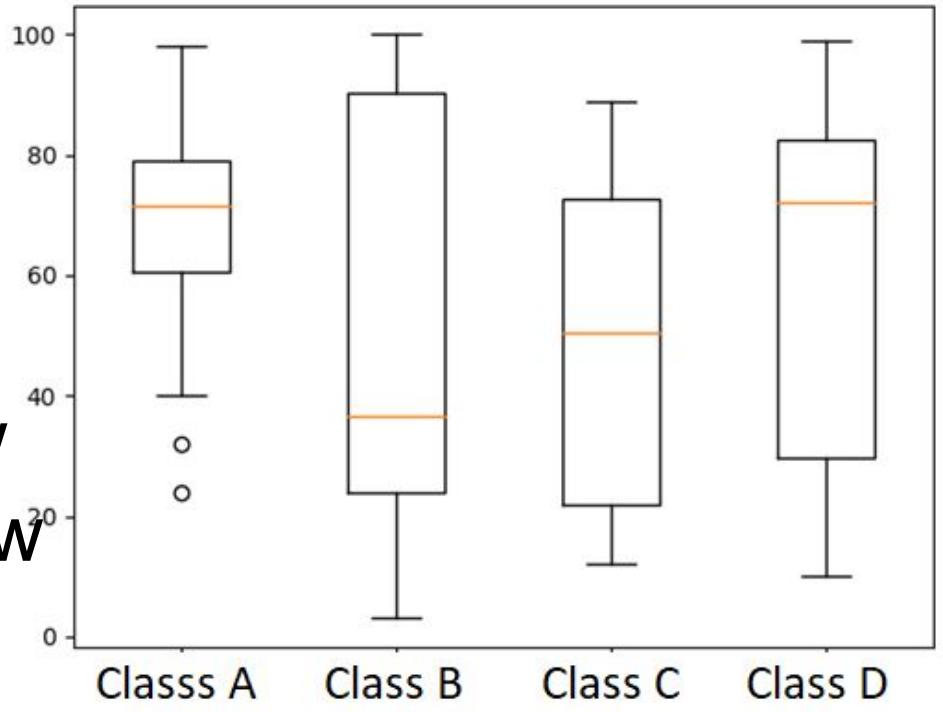
Analysis of Box Plot



Comparison of Boxplots

Student performance in various classes

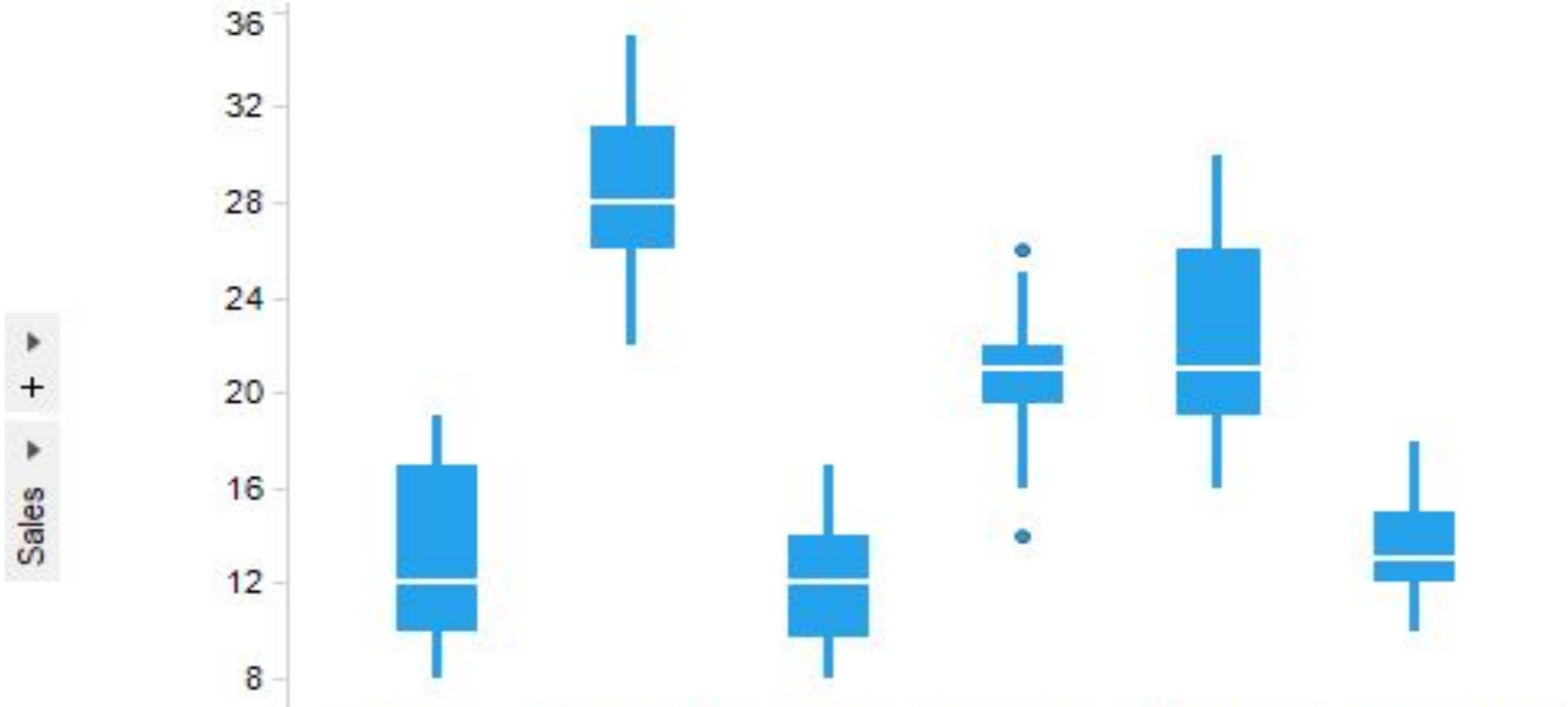
1. Class C is normal distributed
2. Class B positive skew
3. Class D negative skew



…….Analyze more??

Comparison of Boxplots

Sales of fruits



	Apples	Bananas	Cucumber	Lettuce	Pears	Tomatoes
Count	36	36	36	36	36	36
Median	12	28	12	21	21	13
Outliers	0	0	0	2	0	0

Type + +

Application of Box plot in ML

Application of Box plot in ML

- Feature selection techniques are used to select the important columns in a dataset. Feature selection reduce the data column wise.

- Box plot can be used to reduce data row wise.

Application of Box plot in ML

Removal of outliers from the dataset
i.e. Remove the complete row from the dataset.

Before (10 x 5)

Class	F1	F2	F3	F4
0	102	23	19	80
1	10	21	37	16
0	52	21	43	80
1	67	5	48	37
0	63	53	79	20
1	28	17	29	12
0	60	66	99	25
0	30	58	41	46
0	88	23	56	71
1	86	71	21	1

After (5 x 5)

Class	F1	F2	F3	F4
0	52	21	43	80
0	63	53	79	20
1	28	17	29	12
0	30	58	41	46
0	88	23	56	71

Tips & Tricks

1. Feature selection should be apply when dataset have large number of features.

2. Feature selection is not useful when number of features are small.

3. Row removal (using outlier) will be useful when the number of rows are too large.

4. Row removal (using outlier) will be not useful when the number of rows small i.e. 100, 200, 300.

5. Which one to apply first:
   Column wise removal (Feature Selection)
   or
   Row wise (Outlier removal)

Project Work

1. Learn the outlier using pandas from the given below link.
   https://www.youtube.com/watch?v=rzR_cKnkD18
2. Implement the outlier removal method (row removal) in python and develop a cmd line solution (i.e. run through cmd line), handle all the exception:
   Usages:
   python outlier.py <InputDataFile> <OutputDataFile>
   Example:
   python outlier.py myData.csv newData.csv
   Output also show the number of rows are removed.
3. Develop a python package.
4. Test and validate it on different dataset
5. Upload the package on pypi.org

Thanks

Learning by Doing

www.psrana.com