Boxplots, Interquartile Range, and Outliers

Boxplots provide a visual representation of a data set that can be used to determine whether the data set is symmetric or skewed. Constructing a boxplot requires calculation of the "5 number summary", the interquartile range (IQR), and the presence of any outliers.

5 Number Summary – The 5 number summary for a data set includes the following, which are listed in order from smallest to largest –

1. **Minimum** - The smallest value in the data set.
2. **First Quartile** - Separates the lowest 25% of the data in a set from the highest 75%. It is typically denoted as $\boldsymbol{Q_1}$ *where*, $\frac{25}{100} \cdot (\#\ points\ in\ data\ set) = position\ of\ Q_1\ in\ set.$
3. **Median** – The middle value in a sorted (smallest to largest) data set. If there is an even number of values, it is calculated by averaging the two middle values. The Median is also referred to as the **Second Quartile** $(\boldsymbol{Q_2})$ because it separates the lower 50% of data in a set from the upper 50%.
4. **Third Quartile** - Separates the lowest 75% of the data in a set from the highest 25%. It is typically denoted as $\boldsymbol{Q_3}$ *where*, $\frac{75}{100} \cdot (\#\ points\ in\ data\ set) = position\ of\ Q_3\ in\ set.$
5. **Maximum** – The largest value in the data set.

IQR - The Interquartile Range is a measure of spread used to calculate the lower and upper outlier boundaries. These boundaries are then used to determine whether a data set has any actual **outliers**.

$$\boldsymbol{Interquartile\ Range}\ (IQR) = Q_3 - Q_1$$

$$\boldsymbol{Lower}\ Outlier\ Boundary = Q_1 - 1.5\ IQR$$

$$\boldsymbol{Upper}\ Outlier\ Boundary = Q_3 + 1.5\ IQR$$

Outliers - Outliers are data points that are *considerably smaller or larger* than most of the other values in a data set. Data values that are smaller than the lower outlier boundary **or** larger than the upper outlier boundary are outliers. Some data sets do not have any outliers. Outliers that are determined to be the *result of an error should be removed* from the data set.

Example – For the following data set *(2012 data for MLB team payrolls in millions)*, find **a)** the 5 number summary, **b)** the IQR, **c)** the upper and lower outlier boundaries, and **d)** any outliers. *Note – data should be sorted from lowest to highest if it is not provided that way. This allows the easy identification of the min, max, median, and individual data positions within the set.*

	Team	Payroll		Team	Payroll		Team	Payroll		Team	Payroll
1	Padres	55	**9**	Rockies	78	**17**	Mets	93	**25**	Rangers	121
2	Athletics	55	**10**	Indians	78	**18**	Twins	94	**26**	Tigers	132
3	Astros	61	**11**	Nationals	81	**19**	Dodgers	95	**27**	Angels	154
4	Royals	61	**12**	Orioles	81	**20**	W Sox	97	**28**	Red Sox	173
5	Pirates	63	**13**	Mariners	82	**21**	Brewers	98	**29**	Phillies	175
6	Rays	64	**14**	Reds	82	**22**	Cardinals	110	**30**	Yankees	198
7	D Backs	74	**15**	Braves	83	**23**	Giants	118			
8	Blue Jays	75	**16**	Cubs	88	**24**	Marlins	118			

a) 5 Number Summary – These values can be calculated by hand (shown below) **OR** they can be found using the "1-Var Stats" button from the Stat Menu on a TI-83 or TI-84 calculator.

Represents 25th percentile

# of data points in set

Average of 2 middle data points in set

Represents 75th percentile

# of data points in set

Minimum	$\boldsymbol{Position\ Q_1} = \frac{25}{100}(30)$	**Median**	$\boldsymbol{Position\ Q_3} = \frac{75}{100}(30)$	**Maximum**
55	= 7.5 → *8th Position*	$= \frac{83+88}{2}$	= 22.5 → *23rd Position*	**198**
	= **75**	= **85.5**	= **118**	

If the "position" calculation results in a decimal, round up to the next whole number to determine the position. If the calculation results in a whole number, average that position's data value with the next data value

b) IQR → $IQR = Q_3 - Q_1 = 118 - 75 = 43$

c) Upper and Lower Outlier Boundaries –

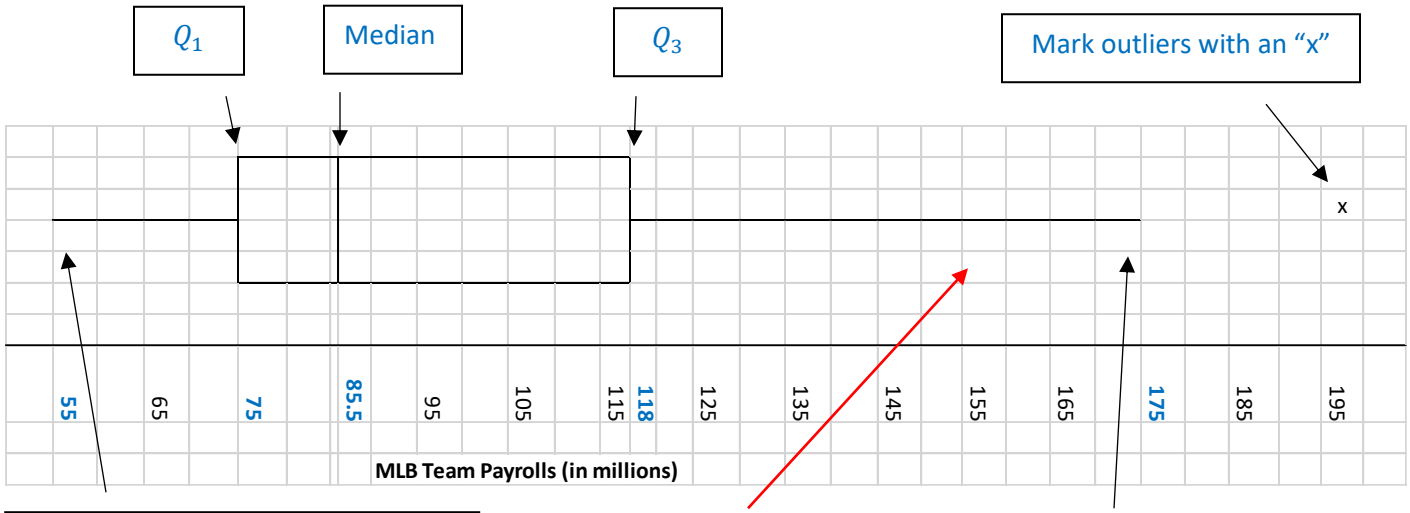
$$Lower\ Outlier\ Boundary = Q_1 - 1.5\,IQR = 75 - 1.5\,(43) = 10.5$$

$$Upper\ Outlier\ Boundary = Q_3 + 1.5\,IQR = 118 + 1.5\,(43) = 182.5$$

d) Outliers – Lower Outliers → None *(There are no individual data points smaller than the lower boundary of 10.5.)*

Upper Outliers → 198 (Yankees) *(This data value is bigger than the upper boundary of 182.5.)*

Constructing a Box Plot – Construct a Boxplot for the data set in the previous example. Determine whether the data set is symmetric or skewed.



Draw the **whisker** out to the smallest data value that is larger than the lower boundary

This data set is *Skewed RIGHT*

Draw the **whisker** out to the largest data value that is smaller than the upper boundary

Try this on your own - Construct a Boxplot for the following data set by finding the 5 number summary, the IQR, the outlier boundaries, and any outliers (if they exist.).

Data Set

8.2	8.8	9.2	10.6	12.7
8.4	9.0	9.7	11.6	14.0
8.5	9.2	10.4	11.8	15.9
8.8	9.2	10.5	12.6	16.1

Answers:

5 Number Summary →

$Min = 8.2$

$Q_1 = 8.9$

$Median = 10.05$

$Q_3 = 12.2$

$Max = 16.1$

$IQR = 3.3$

$Lower\ Outlier\ Boundary =$ 3.95

$Upper\ Outlier\ Boundary =$ 17.15

$Outliers =$ None

Box Plot: