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 \( Q_1 \) where, \( \frac{25}{100} \cdot (\# \text{ points in data set}) = \text{position of } Q_1 \text{ in set.} \)
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** \( (Q_2) \) 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 \( Q_3 \) where, \( \frac{75}{100} \cdot (\# \text{ points in data set}) = \text{position of } Q_3 \text{ in set.} \)
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.

\[
\begin{align*}
\text{Interquartile Range} \text{ (IQR)} &= Q_3 - Q_1 \\
\text{Lower Outlier Boundary} &= Q_1 - 1.5 \text{ IQR} \\
\text{Upper Outlier Boundary} &= Q_3 + 1.5 \text{ IQR}
\end{align*}
\]

**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.

<table>
<thead>
<tr>
<th>Team</th>
<th>Payroll</th>
<th>Team</th>
<th>Payroll</th>
<th>Team</th>
<th>Payroll</th>
<th>Team</th>
<th>Payroll</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Padres</td>
<td>55</td>
<td>9</td>
<td>Rockies</td>
<td>78</td>
<td>17</td>
<td>Mets</td>
</tr>
<tr>
<td>2</td>
<td>Athletics</td>
<td>55</td>
<td>10</td>
<td>Indians</td>
<td>78</td>
<td>18</td>
<td>Twins</td>
</tr>
<tr>
<td>3</td>
<td>Astros</td>
<td>61</td>
<td>11</td>
<td>Nationals</td>
<td>81</td>
<td>19</td>
<td>Dodgers</td>
</tr>
<tr>
<td>4</td>
<td>Royals</td>
<td>61</td>
<td>12</td>
<td>Orioles</td>
<td>81</td>
<td>20</td>
<td>W Sox</td>
</tr>
<tr>
<td>5</td>
<td>Pirates</td>
<td>63</td>
<td>13</td>
<td>Mariners</td>
<td>82</td>
<td>21</td>
<td>Brewers</td>
</tr>
<tr>
<td>6</td>
<td>Rays</td>
<td>64</td>
<td>14</td>
<td>Reds</td>
<td>82</td>
<td>22</td>
<td>Cardinals</td>
</tr>
<tr>
<td>7</td>
<td>D Backs</td>
<td>74</td>
<td>15</td>
<td>Braves</td>
<td>83</td>
<td>23</td>
<td>Giants</td>
</tr>
<tr>
<td>8</td>
<td>Blue Jays</td>
<td>75</td>
<td>16</td>
<td>Cubs</td>
<td>88</td>
<td>24</td>
<td>Marlins</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Minimum</th>
<th>( Q_1 ) Position</th>
<th>Median</th>
<th>( Q_3 ) Position</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>( \frac{25}{100} \times 30 ) = 7.5 → 8\text{th} Position</td>
<td>( \frac{83+88}{2} ) = 85.5</td>
<td>( \frac{75}{100} \times 30 ) = 22.5 → 23\text{rd} Position</td>
<td>198</td>
</tr>
</tbody>
</table>

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** –

\[
\text{Lower Outlier Boundary} = Q_1 - 1.5 \times IQR = 75 - 1.5 \times 43 = 10.5 \\
\text{Upper Outlier Boundary} = Q_3 + 1.5 \times IQR = 118 + 1.5 \times 43 = 182.5
\]

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.
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

<table>
<thead>
<tr>
<th>8.2</th>
<th>8.8</th>
<th>9.2</th>
<th>10.6</th>
<th>12.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.4</td>
<td>9.0</td>
<td>9.7</td>
<td>11.6</td>
<td>14.0</td>
</tr>
<tr>
<td>8.5</td>
<td>9.2</td>
<td>10.4</td>
<td>11.8</td>
<td>15.9</td>
</tr>
<tr>
<td>8.8</td>
<td>9.2</td>
<td>10.5</td>
<td>12.6</td>
<td>16.1</td>
</tr>
</tbody>
</table>

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: