LESSON HANDOUT

# Plotting and interpreting boxplots

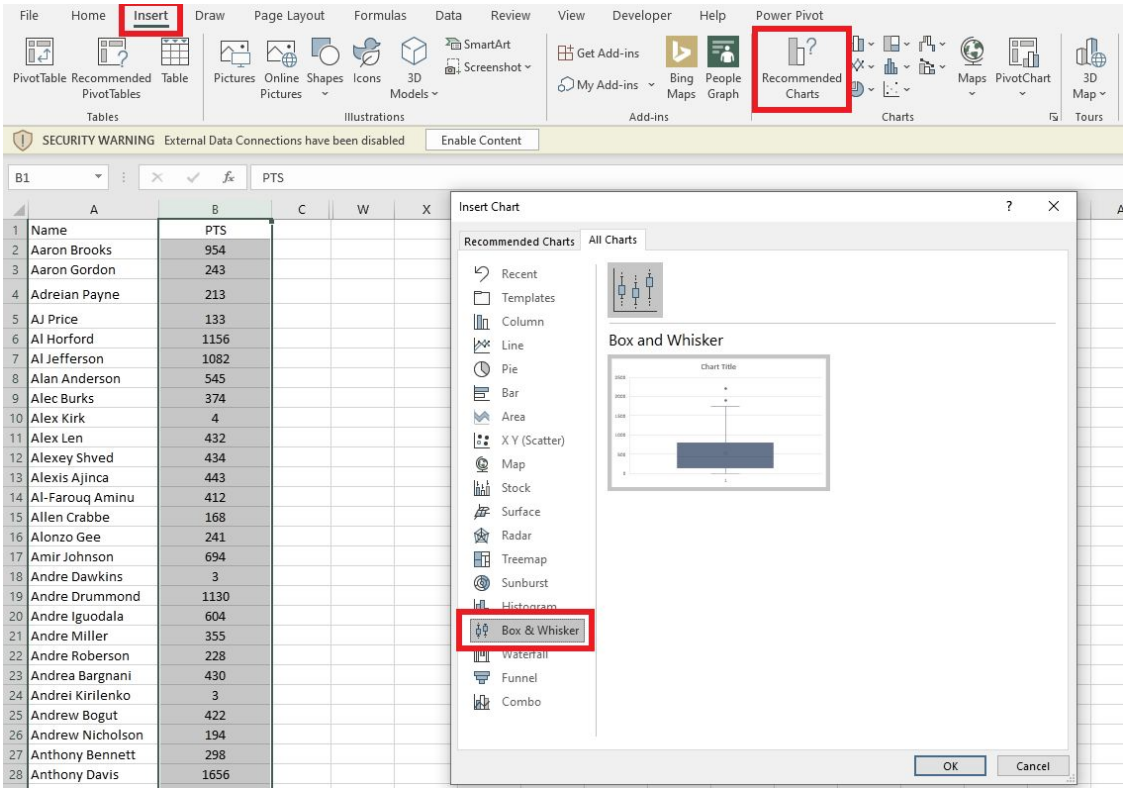
## Why

Boxplots are a way of visualising the spread of your dataset and to identify outliers at a glance.

## How

To plot a boxplot in excel:

- Highlight data > Insert > Recommended Charts > Box and whisker

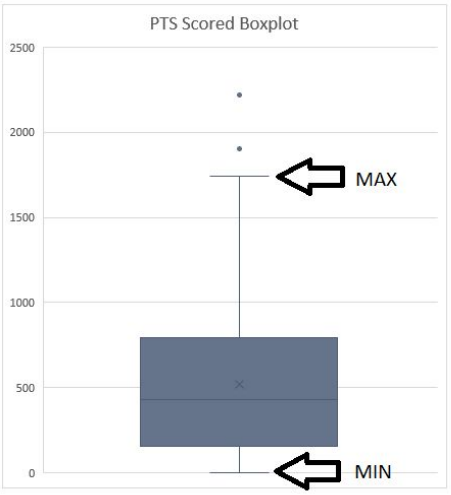



| | A | B |
|---|---|---|
| 1 | Name | PTS |
| 2 | Aaron Brooks | 954 |
| 3 | Aaron Gordon | 243 |
| 4 | Adreian Payne | 213 |
| 5 | AJ Price | 133 |
| 6 | Al Horford | 1156 |
| 7 | Al Jefferson | 1082 |
| 8 | Alan Anderson | 545 |
| 9 | Alec Burks | 374 |
| 10 | Alex Kirk | 4 |
| 11 | Alex Len | 432 |
| 12 | Alexey Shved | 434 |
| 13 | Alexis Ajinca | 443 |
| 14 | Al-Farouq Aminu | 412 |
| 15 | Allen Crabbe | 168 |
| 16 | Alonzo Gee | 241 |
| 17 | Amir Johnson | 694 |
| 18 | Andre Dawkins | 3 |
| 19 | Andre Drummond | 1130 |
| 20 | Andre Iguodala | 604 |
| 21 | Andre Miller | 355 |
| 22 | Andre Roberson | 228 |
| 23 | Andrea Bargnani | 430 |
| 24 | Andrei Kirilenko | 3 |
| 25 | Andrew Bogut | 422 |
| 26 | Andrew Nicholson | 194 |
| 27 | Anthony Bennett | 298 |
| 28 | Anthony Davis | 1656 |

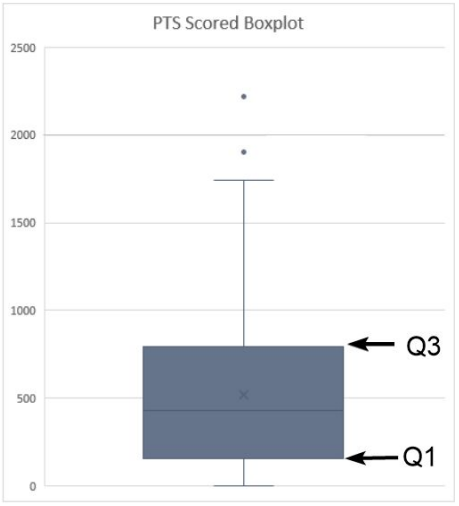
## Components of a boxplot

The boxplot has the following components:

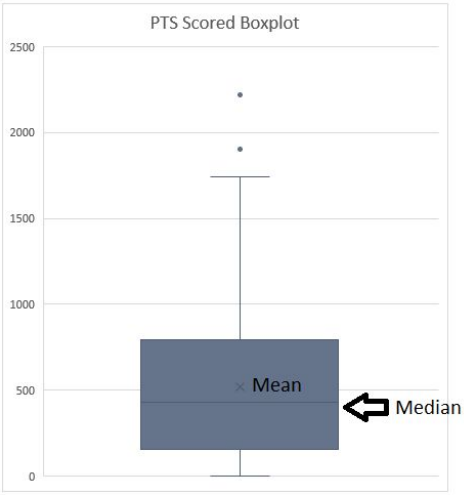
- The upper and lower whiskers represent the **minimum** and **maximum** values (not including outliers)



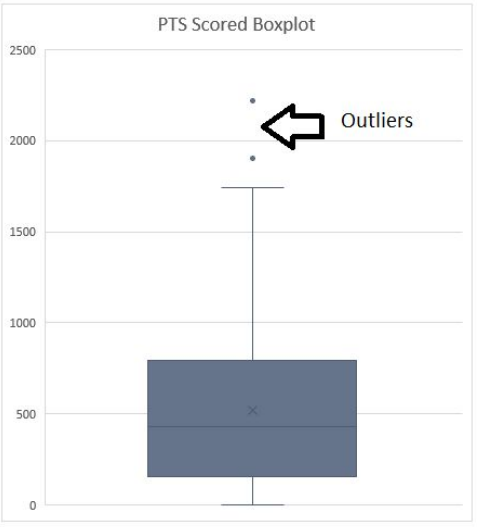

- The bottom of the box represents Quartile 1 and the top of the box Quartile 3




- The middle line represents the median value and X represents the mean



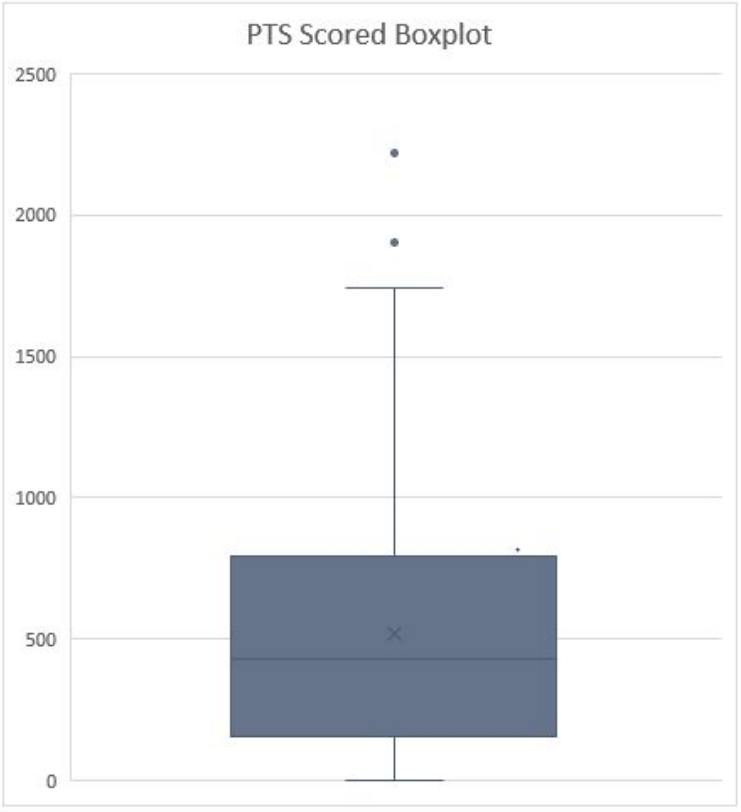
- Any point above or below the whiskers represent outliers i.e. those points that lie more than 1.5 x the interquartile range from Q3 or below Q1




## Interpreting a boxplot

A boxplot shows, at a glance, how spread apart data is within your quartiles. For example from the below boxplot we could deduce:

- The upper quarter of this data has a much greater range than the lower quarter.
- Upper outliers exist and the mean is higher than the median
- This dataset is therefore *right-skewed*



## Using quartiles and IQR to segment data

A useful exercise in understanding our data is segmenting (or grouping) our data, then examining those segments. We can do this by splitting our data around the quartiles and IQR.

- Those values below 1.5 x IQR from Q1 or above 1 x IQR from Q3 are classified as very low and very high values respectively
- Those values in the lower and upper quarters are classified as low and high, and;
- Those in the middle two quarters are classified as medium.

- We can assign these values to our data using the *approximate match* in **VLOOKUP** (assigning approximate rather than exact match at the end of the VLOOKUP formula).
  - Step 1. Create a **classification table**. Approximate match works in this case by looking up the value in the second column when it is equal to the value in the first column.

**Classification using approximate match**

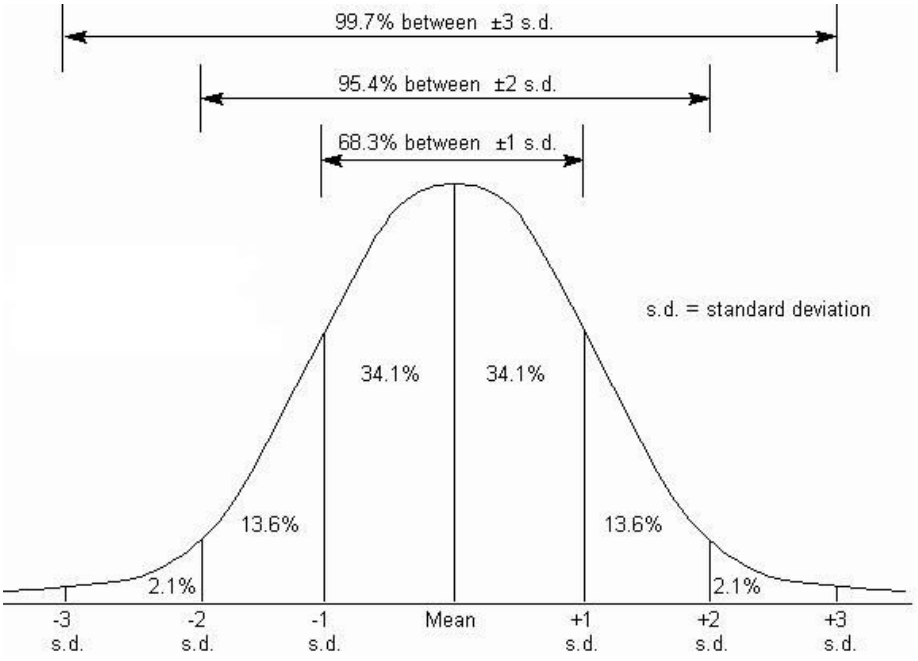
| Value | Classification |
|---|---|
| Min value | Very low |
| Lower outlier bound (Q3 + 1.5 x IQR) | Low |
| Quartile 1 | Medium |
| Quartile 3 | High |
| Upper outlier bound (Q3 + 1.5 x IQR) | Very high |

  - Step 2. Create the VLOOKUP formula using an approximate match.
    - Example: =VLOOKUP(Value, Array, Column number, TRUE)

## Using percentile rank to check an individual value's significance

Percentile rank is useful in determining an individual value's significance in the context of the whole.

- For example, if an NBA players **Z scores** for the number of points scored this season is 3 (i.e. 3 STD deviations from the mean), then this would indicate that the players' PTS scores are significantly high. In fact, in a normal distribution, this Z score would place this player in the top 0.3 percent of point scorers.




- To calculate **Z score**, divide the individual values distance to the mean by the STD deviation.
    - Example. If Player A is 30cm taller than the mean and the STD Deviation is only 15cm, Player A's Z score would be 2.