Outliers’ detection and treatment

**Outliers:** An outlier is an observation that lies at an abnormal distance from other values in a random sample from a population. These are extreme or unusual observations in the data.

**Possible causes:** recording error, data entry/etl errors, noise.

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer** | **Age** | **Income** | **Gender** |
| Ralph | 400 | 110000 | M |
| Kate | |  | | --- | | 25 | | 103000 | F |
| Mike | 31 | 96500 | M |
| Kevin Young | 42 | 150000 | M |
| Peter | 22 | 64000 | M |
| Sophie | 28 | 96000 | F |
| Sarah | 45 | 50000000 | F |
|  |  | |  | | --- | |  | |  |
|  | Invalid | Valid |  |

**Detection of outliers:**

1. Univariate methods
2. Multivariate methods

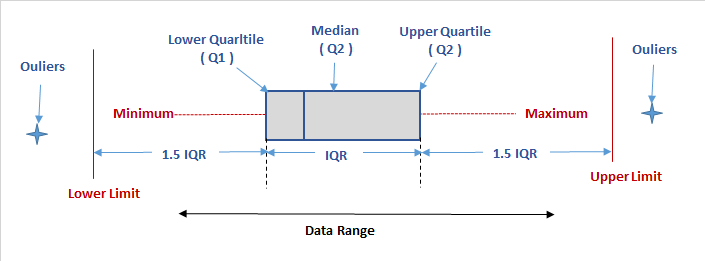
**Univariate Methods:**

1. Statistics – simple min – max
2. Box plots
3. Histograms
4. Z-scores

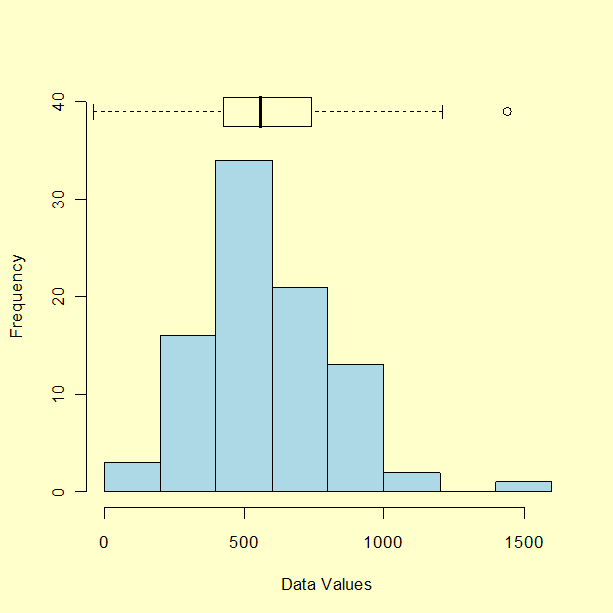
**Box Plots:**

Any values greater than or less than 1.5 times of IQR are outliers.

IQR = Q3 – Q1



**Histograms:**



**Z-score method:**

The Z-score measures how many standard deviation away from the mean an observation lies for a specific variable.

Z = (X – μ)/σ

Outliers are defined when Z >3(sometimes 2.5)

Note: Mean of the Z-scores is 0 and standard deviation is 1.

|  |  |
| --- | --- |
| **Age** | **Income** |
| 22 | 4000 |
| 52 | 2200 |
| 48 | 2000 |
| 34 | 1800 |
| 64 | 1500 |
| 32 | 3800 |
| 46 | 2300 |
| 50 | 3600 |
| 52 | 3750 |

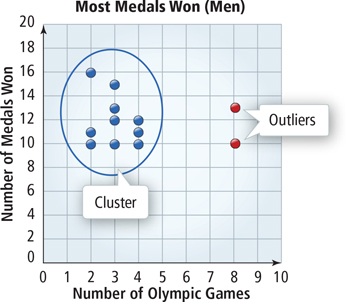
Outliers can be hidden in one dimensional views:

**Multivariate Methods:**

1. Clustering
2. Regression

**Clustering:**

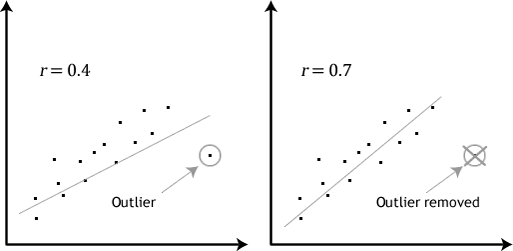
Look for the observations away from clusters



**Regression**

Fit a regression line and look for observations with large errors. Inspect residual plots.

r is a statistical measure of how close the data are to the fitted regression line



**Treating outliers:**

**Treating invalid outliers:**

1. Keep
2. Delete
3. Replace

Keep: Keep the values and use some tags to identify the outliers

Delete: Completely delete the row.

Replace: replace with something like mean/ median for numeric variables and mode for categorical variables.

**Treating valid outliers:**

1. Truncation based on Z-scores
2. Truncation based on IQR
3. Truncation using sigmoid

**Truncation** is also called **winsorizing** or **capping**

**Truncation based on Z-scores:**

Replace all the variable values having Z-score of >3 with the mean +3 times the standard deviation.

Replace all the variable values having Z-score of < -3 with the mean -3 times the standard deviation.

**Truncation based on IQR:**

Truncate to M ± 2S

M: Median  
S = IQR/ (2\*0.6745)

**Truncation using sigmoid:**

Sigmoid transformation f(x) = 1/ (1+ e^​x)