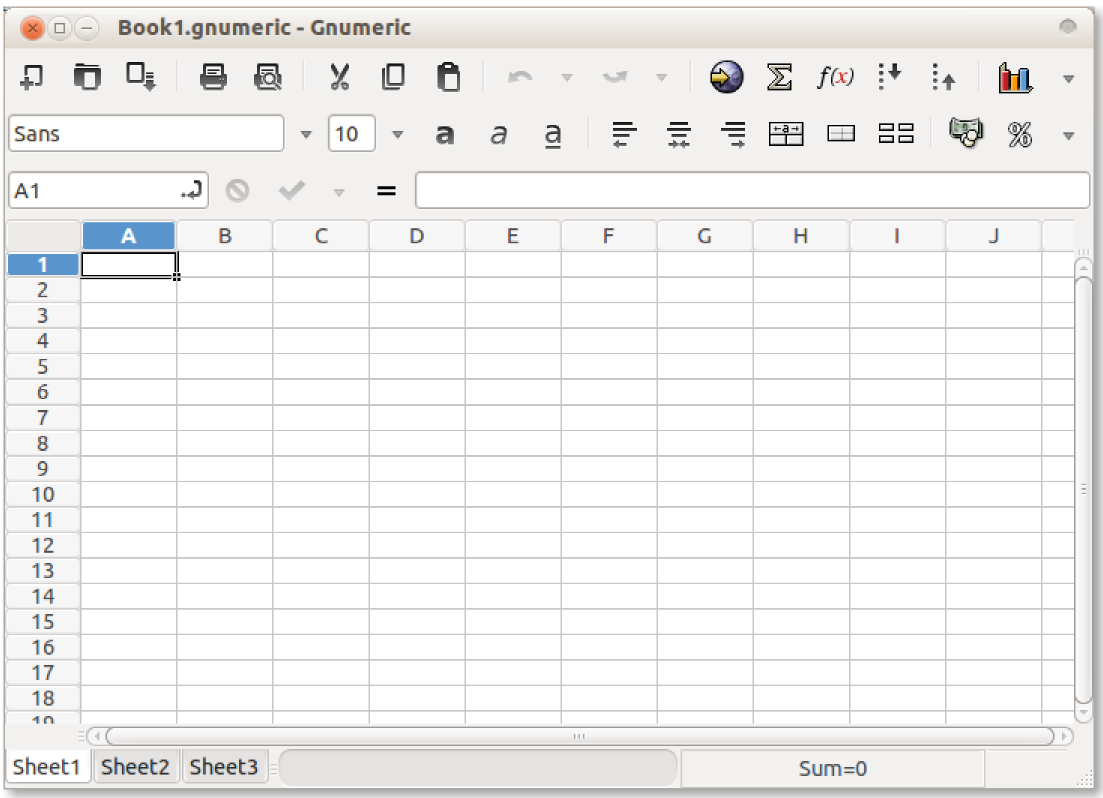
Statistical Analysis Using Gnumeric

There are many software packages that will analyse data. For casual analysis, a spreadsheet may be an appropriate tool. Popular spreadsheets include Microsoft Excel, LibreOffice Calc, and Gnumeric.

While most operations below can be carried out using any spreadsheet, they often lack certain features. Excel, for example, has no native box-and-whisker plot capabilities. Gnumeric, however, has a wide range of statistical analysis tools built into it. Gnumeric is free software, and runs on multiple operating systems. You can download Gnumeric at **`http://projects.gnome.org/gnumeric`**. You can even download a portable version that you can run off of a USB drive from **`http://portableapps.com`**.

Entering Data and Calculating Values

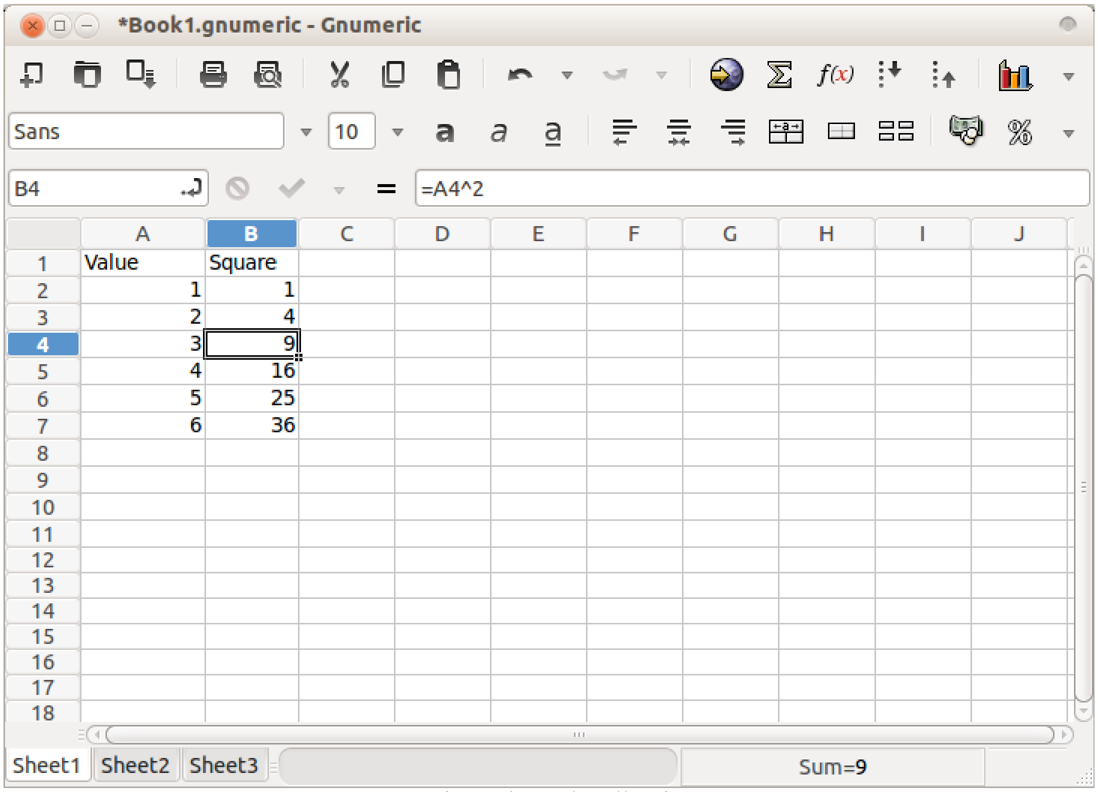
When you first start Gnumeric, the main screen will look something like this.



Initial screen after starting Gnumeric

The area is divided into a rectangular grid of *cells*. Each cell has a *reference* made up of its column letter and row number. For example, the reference of the *active cell* in the screenshot above is A1.

Data can be entered by typing in a value or a formula in a cell. Typically, data from each variable is entered downward in columns rather than across in rows. A descriptive *column header* is recommended, so that it is clear what the data represent. A simple example is shown on the following page.



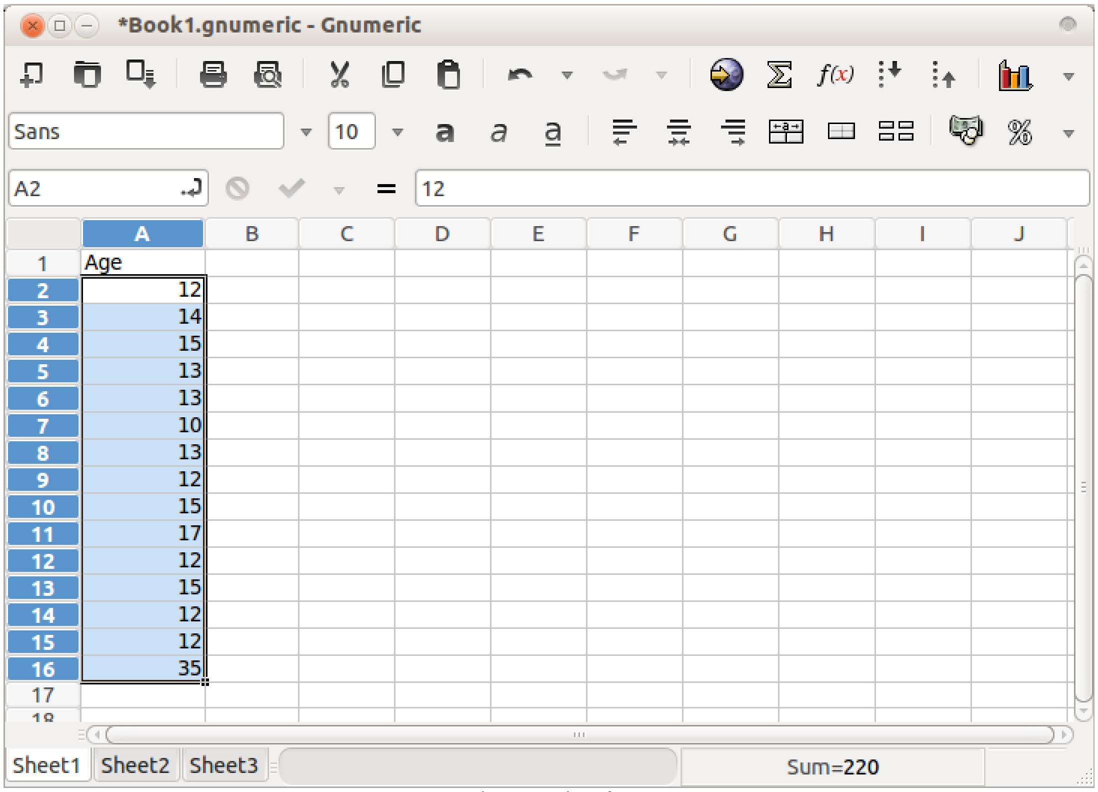
Using a formula with cell references

Formulae may be entered in the formula box, next to the equals sign at the top of the screen, or directly in an empty cell. All formulae begin with **=**, to indicate that they should be evaluated. Formulae may use values, cell references and mathematical operators in them. In the screenshot above, the value of 9 in cell B4 is the square of the value of 3 in A4, and was calculated using the formula `=A4^2`. Cell references are encouraged, since all calculations based on the value of a referenced cell will update if the value is updated.

In addition to mathematical operators, *functions* can also be used in formulae. Common functions are **sum()** to add a range of values, **average()** to calculate the mean, **median()** to find the central datum in a set, and so on. See the Help menu option to discover the full range of functions available.

Obtaining Basic Statistics About Data

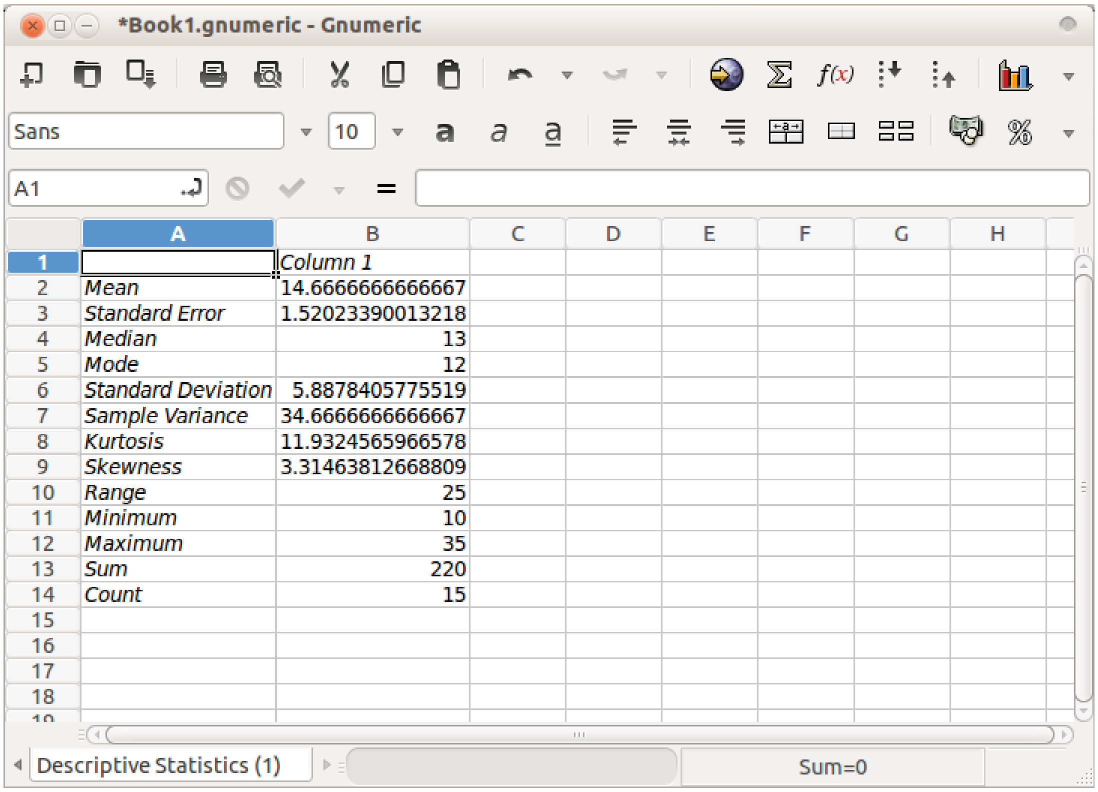
Assume that the following sample of ages of various players of an online game has been entered.



A simple sample of 15 ages

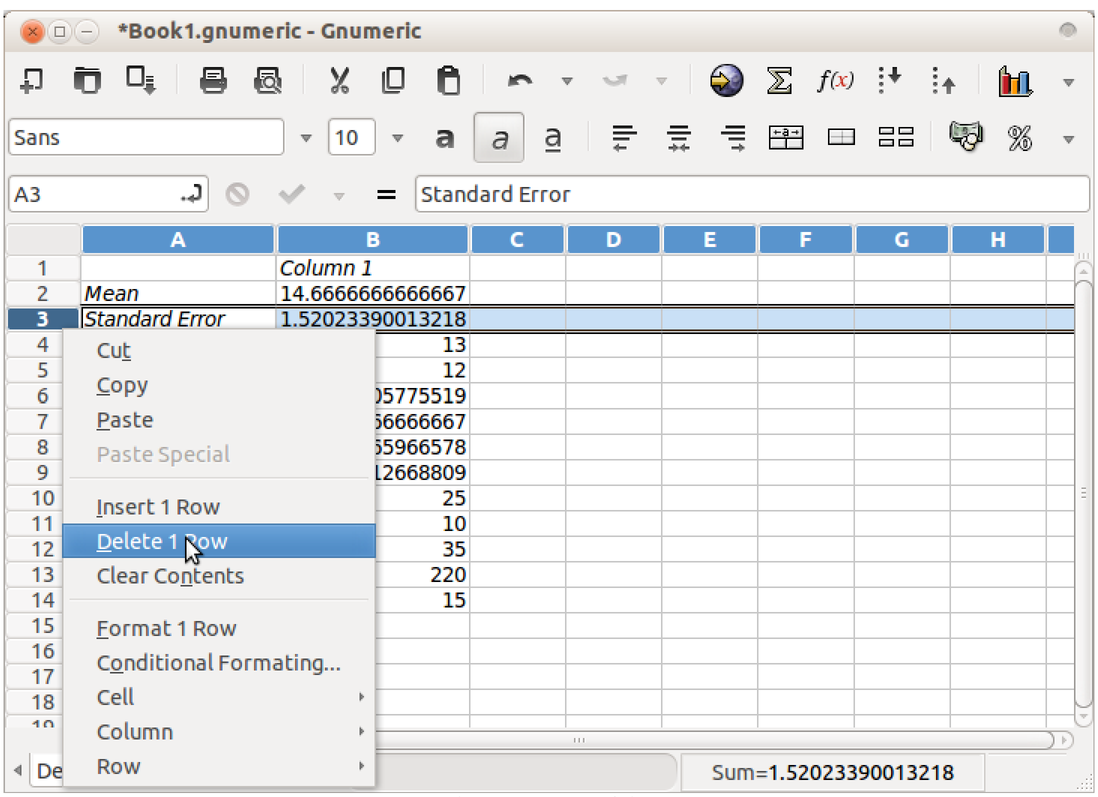
It is possible to calculate different statistics by using Gnumeric's functions. For example, the standard deviation of the sample can be calculated using **`=stdev(A2:A16)`**, or the mode using **`=mode(A2:A16)`**. While effective, there is a better method.

A quick way to calculate basic statistics is to select **Statistics** → **Descriptive Statistics** → **Descriptive Statistics** from the main menu. This calculates common measures of central tendency (mean, median and mode) and measures of spread (standard deviation and variance) for a *sample*. Doing this with the earlier data produces the following information on a separate sheet.



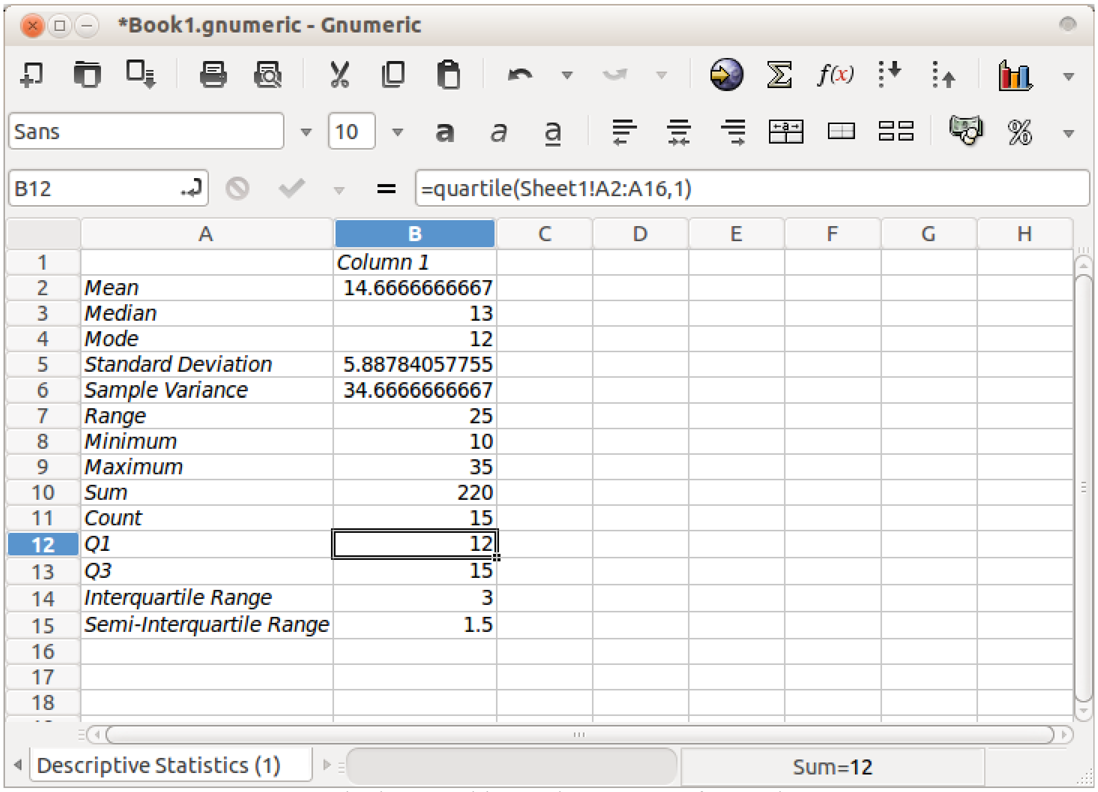
Generating common statistics

Statistics that are not of interest (such as kurtosis, skewness, or standard error) can be deleted by right-clicking a row number and selecting **Delete Row**.



Deleting a row of data

Additional statistics, including the first and third quartiles and the interquartile range, must be calculated manually. To calculate the first quartile, enter the label Q1 in cell A12, and enter the following formula in cell B12: **=quartile(Sheet1!A2:A16,1)**. In this formula, **Sheet1!** refers to the data on Sheet1 of the workbook, **A2:A16** specifies the cells containing the data, and **1** refers to the first quartile. The result is shown below. The third quartile can be calculated in the same way, using **3** in place of **1**. This is shown in row 13.



	A	B
1		Column 1
2	Mean	14.6666666667
3	Median	13
4	Mode	12
5	Standard Deviation	5.88784057755
6	Sample Variance	34.6666666667
7	Range	25
8	Minimum	10
9	Maximum	35
10	Sum	220
11	Count	15
12	Q1	12
13	Q3	15
14	Interquartile Range	3
15	Semi-Interquartile Range	1.5

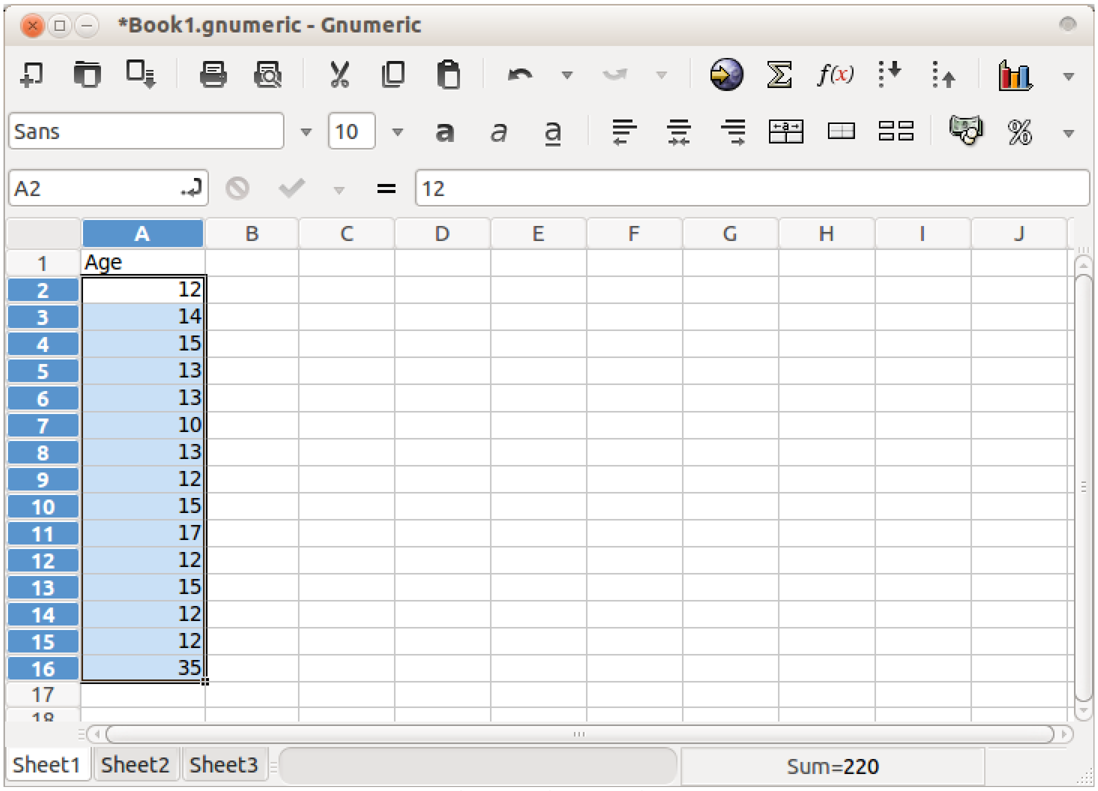
Calculating additional measures of spread

The interquartile range is the range of the data between the first and third quartiles. In cell B14, the formula **=B13-B12** was used. The semi-interquartile range is half the value of the interquartile range, and is calculated using **=B14/2**.

Besides calculating values of common measures of central tendency or spread, it is possible to create various types of graphs. The next few sections describe how to create box-and-whisker plots and histograms for single-variable data.

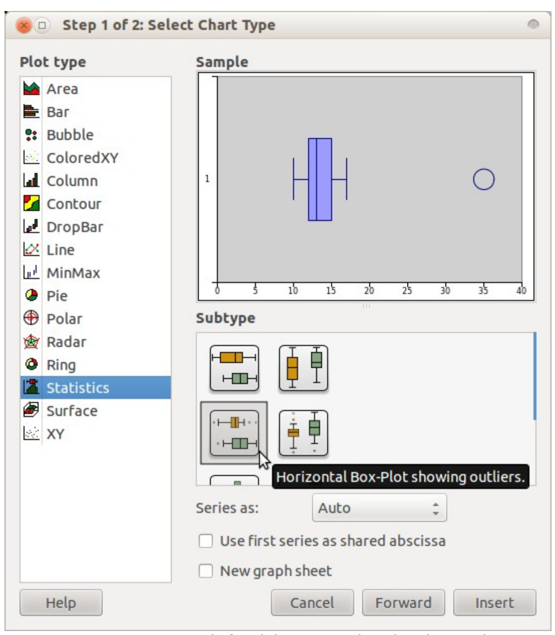
Creating Box-and-Whisker Plots

To create a box-and-whisker plot for single-variable data, enter the data in a column and select the values (not including the column header) using either the mouse of the shift key.



Selecting data to plot

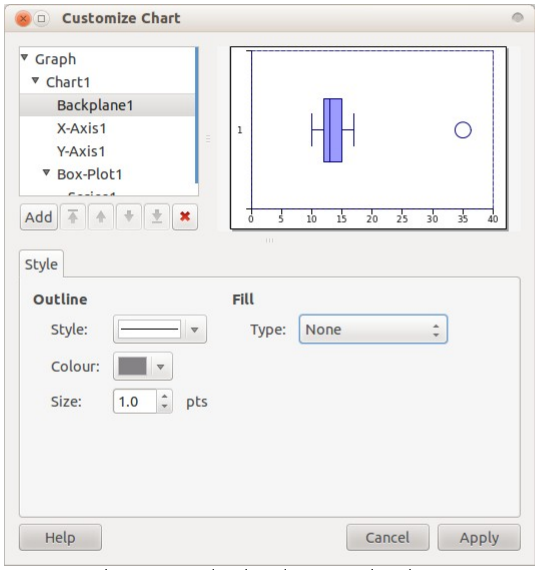
Select **Insert** → **Chart** from the main menu, or click the Insert Chart icon on the menu bar. This will bring up a dialogue box with various options. Select **Statistics** and you should see four types of box-and-whisker plots: both standard modified box-and-whisker plots, in both horizontal or vertical orientations.



Creating a modified box-and-whisker plot

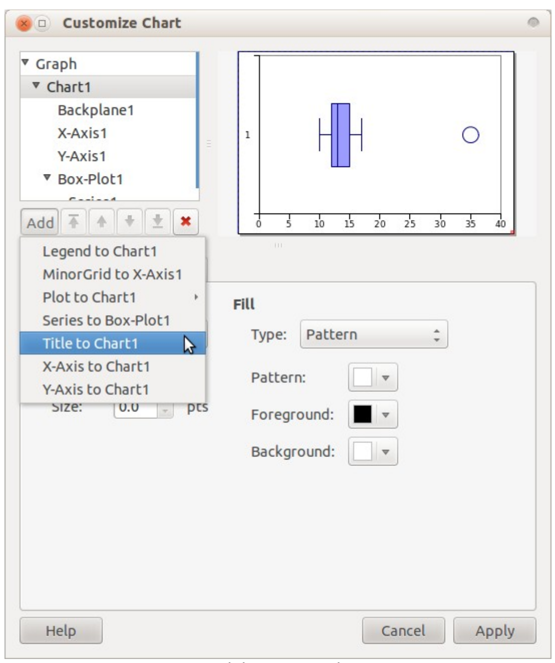
The mouse cursor should change from an arrow to cross-hairs, allowing you to select where on your sheet you would like to place the graph. Click anywhere on the sheet to create the plot.

The appearance of a box-and-whisker plot can be changed by right-clicking on the plot and selecting **Properties**. For example, to remove the default grey background, click B*ackplane*, and change the *Fill Type* to **None**.



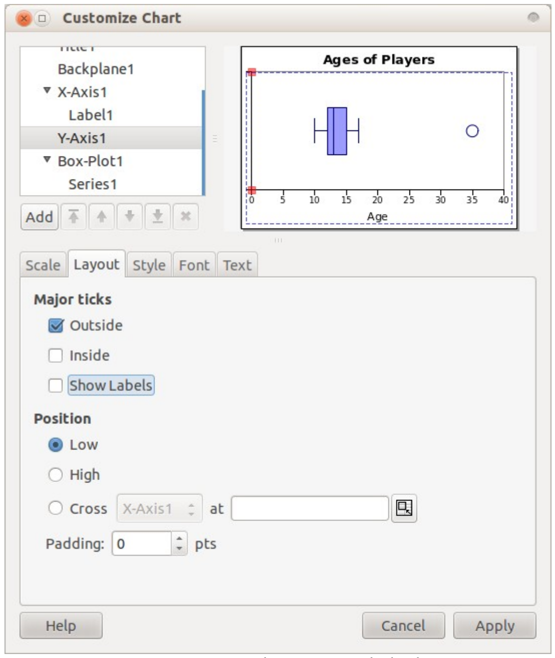
Changing the background colour

To create a title for the box-and-whisker plot, click *Chart1* in the list. Click the Add button, then select **Add Title** as shown below. Similarly, a label can be added to the x-axis by selecting *X-Axis*, then the **Add** button, then **Label**.



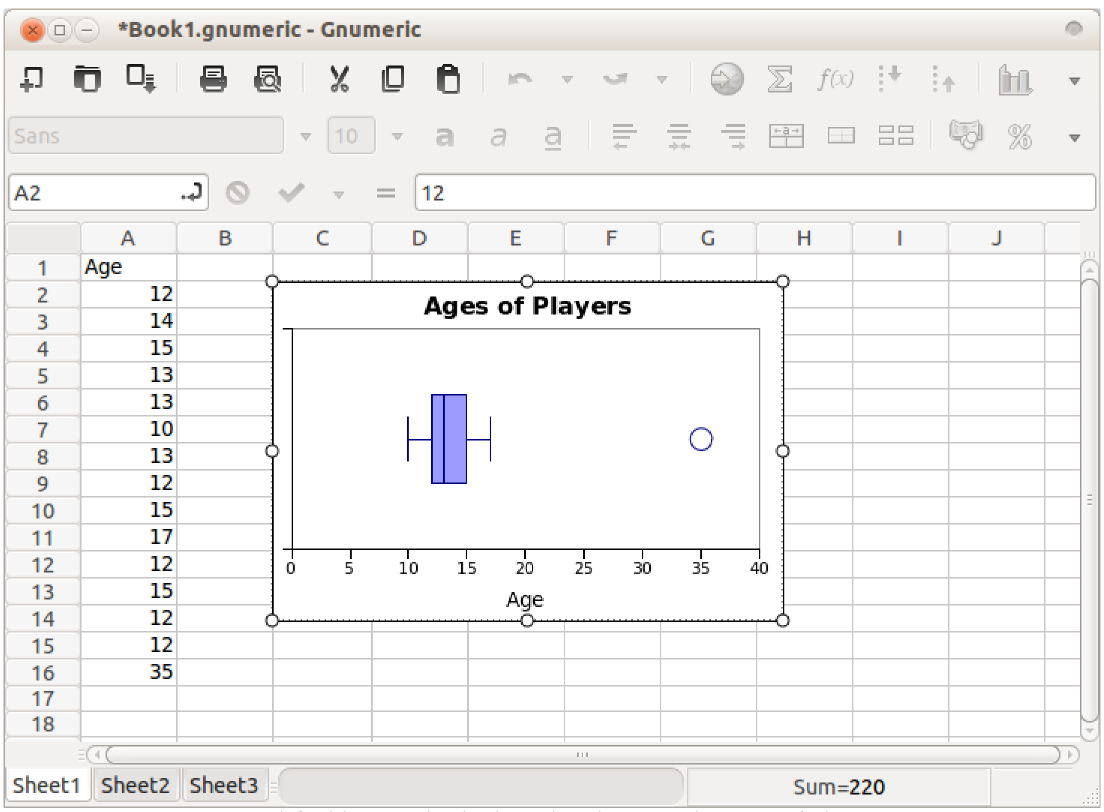
Adding a title

The label on the y-axis is meaningless, and may be removed by selecting *Y-Axis* from the list, then unchecking **Show Label** from the **Layout** tab.



Removing the y-axis label

The final product is displayed below.

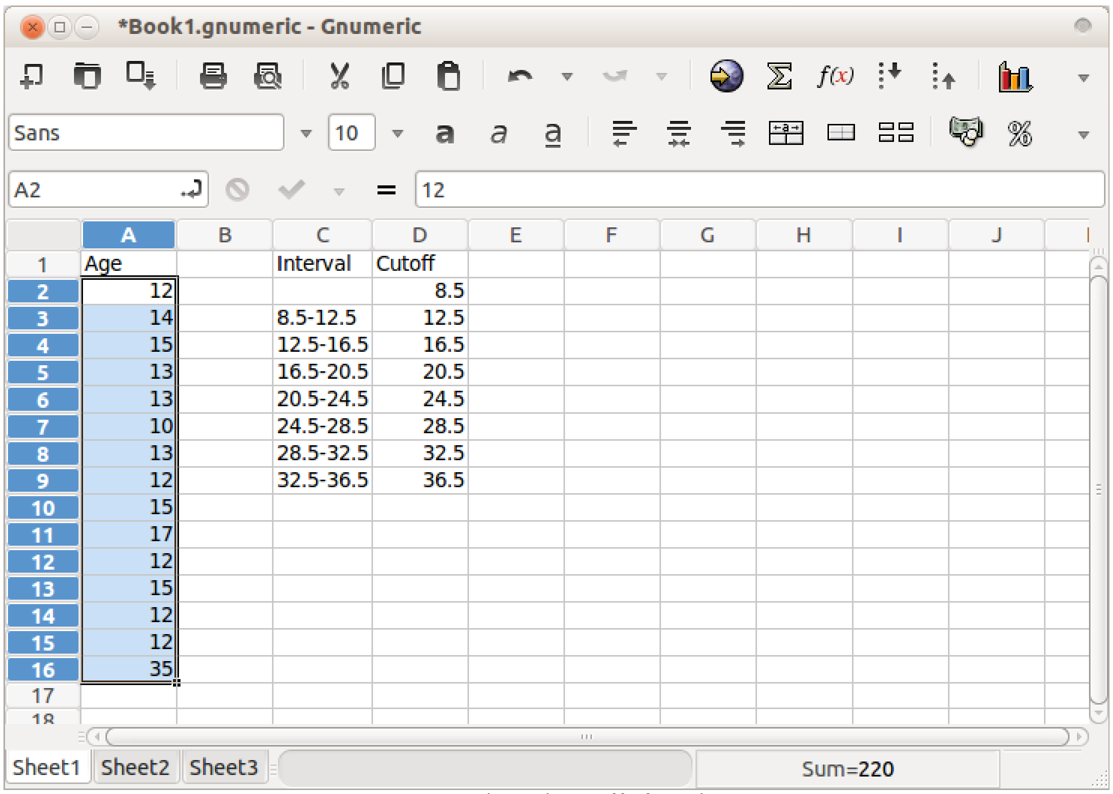


A modified box-and-whisker plot showing the ages of players

Creating Histograms and Frequency Polygons

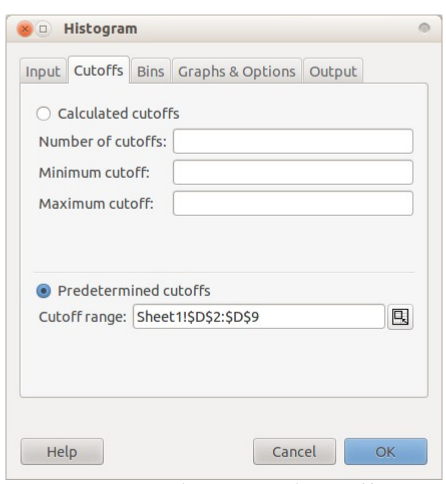
A histogram is similar to a bar chart, where the area of each bar is equal to the frequency of all data in a given interval, and the height of each bar is equal to the interval frequency divided by its width. It may, however, be easier to visualize data when the height of each bar corresponds directly to its frequency. The procedure below produces a hybrid histogram/bar chart with this property instead.

Before generating a frequency table to use with our histogram, create two columns. The first column will correspond to the intervals used to group the data. The second column will be the upper cutoff of the interval. Also include the lower cutoff of the first interval, as shown in the screenshot below.



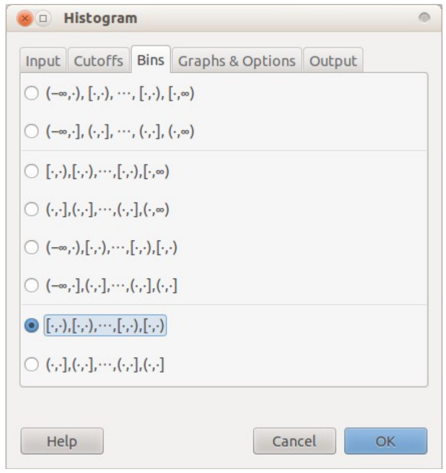
Creating intervals and cutoffs for a histogram

Highlight the data, then select **Statistics** → **Descriptive Statistics** → **Frequency Tables** → **Histogram**. A dialogue box will appear with the input range pre-filled. Click the *Cutoffs* tab and choose **Predetermined cutoffs**. Use the selection tool to highlight the cutoffs created earlier.



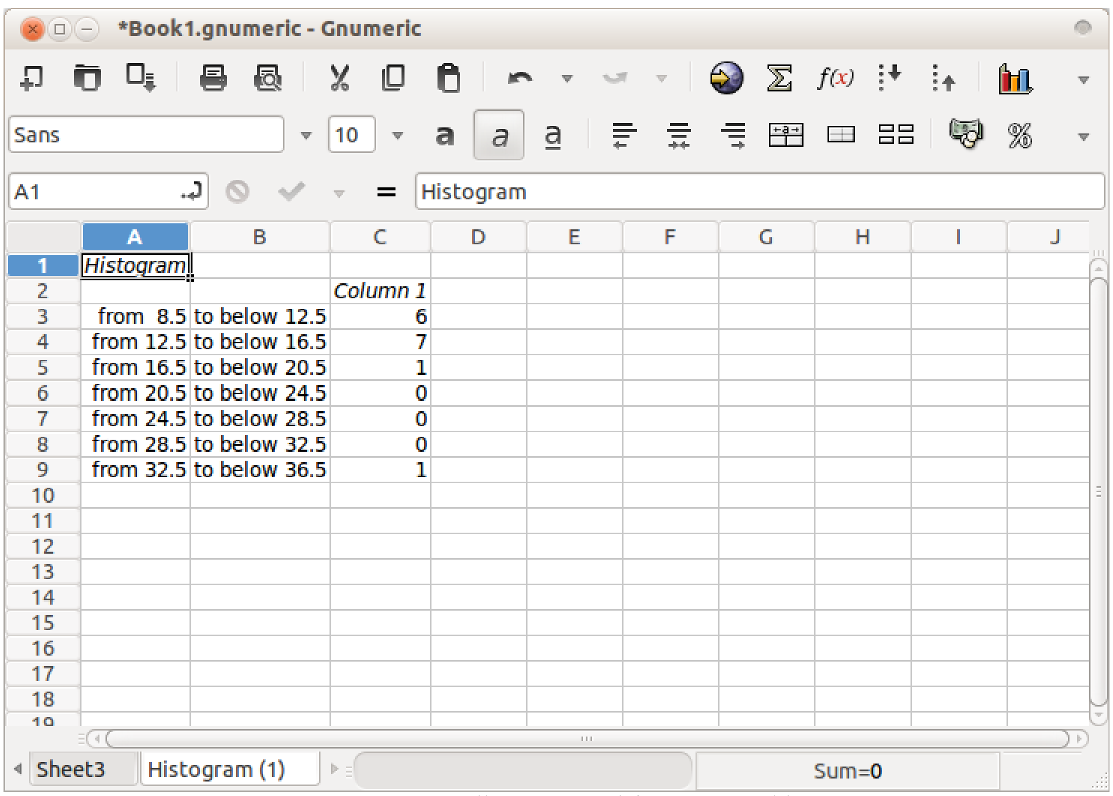
Using predetermined cutoffs

The *Bins* tab provides various options as to how the data should be grouped into intervals. The notation used is standard for describing intervals, but for those unfamiliar with this notation, select the second-last option as shown below. This option places a datum into an interval if it is greater than or equal to the lower cutoff, and strictly less than the upper cutoff.



Defining intervals

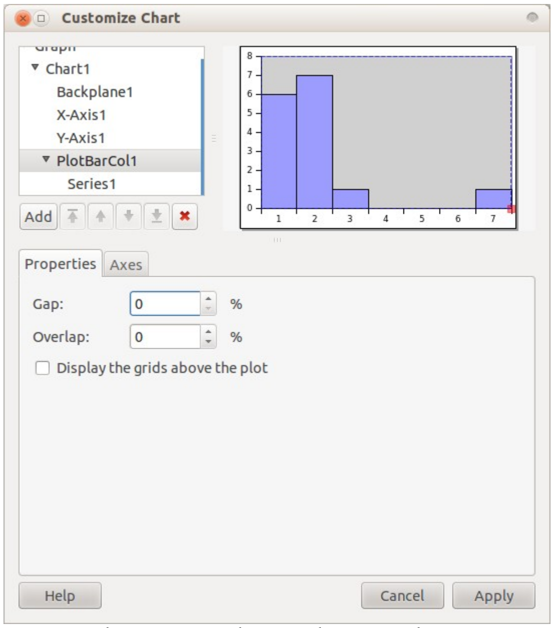
By default, the histogram tool does not produce a graph, and produces a frequency table on a separate sheet. Since a graph will be created manually, click **OK** to create the table. The result is below.



An automatically-generated frequency table

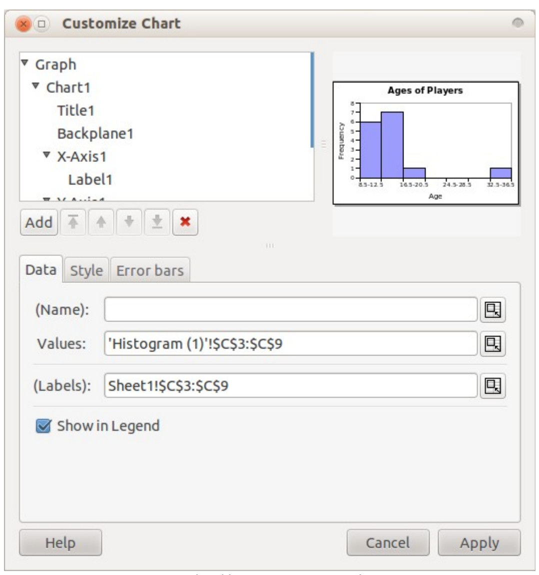
Verify that the data has been properly tallied. Once convinced, select all of the data in column C and click **Insert** → **Chart**. Select a **Column** chart, and place it on the worksheet. Right-click and select **Properties** to make some changes.

Since a histogram typically represents continuous data, a column graph may not be appropriate. The gaps between bars can be eliminated by selecting the *Series* and changing the **Gap** to 0%.



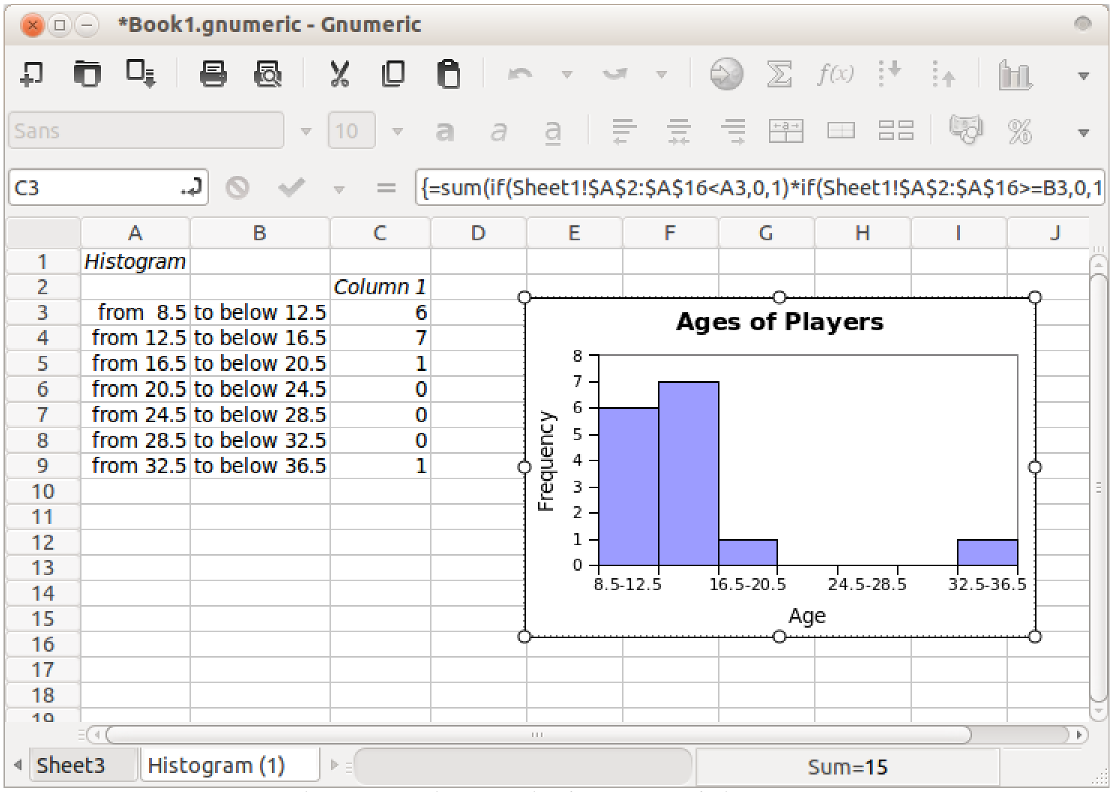
Eliminating the gap between bars

The default labels on the *x*-axis do not really make sense. The intervals defined on Sheet1 can be used in place of 1-7. To do this, select *Series1* from the list and click on the **Data** tab. Next to the Labels field, use the tool to select the intervals from Sheet1 of the workbook. The labels on the *x*-axis should change to those that were created earlier.



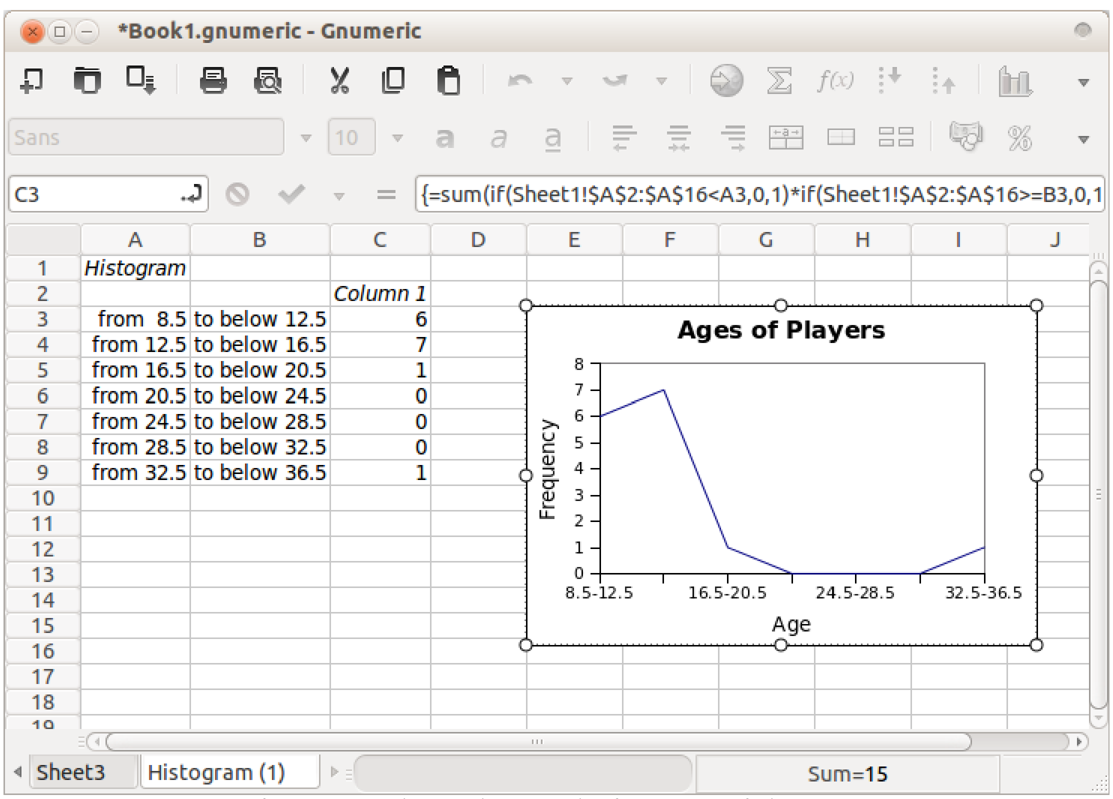
Labelling intervals

As with box-and-whisker plots, it is possible to add a title, label the axes, change colours, and so forth. A completed histogram is shown on the next page.



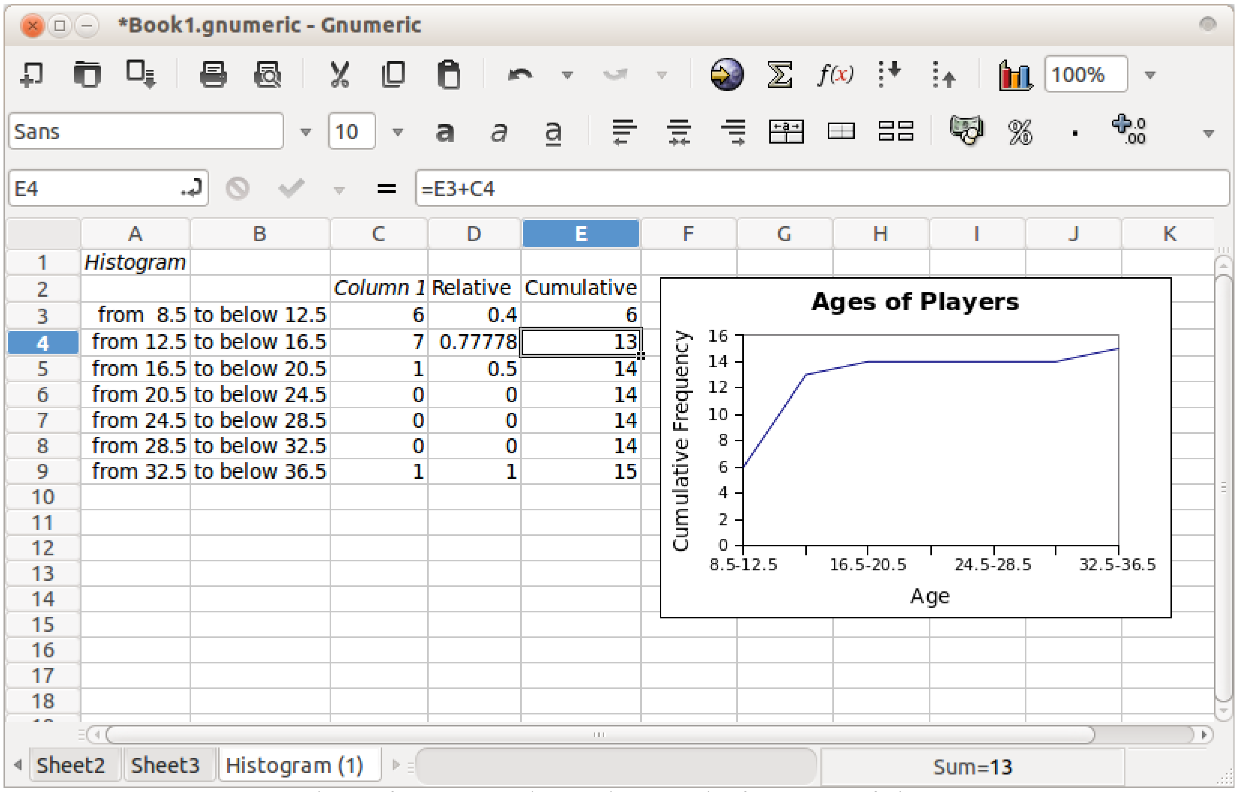
A histogram showing the frequency of player ages

A frequency polygon is a line graph. Using the data generated from the Histogram tool, it is possible to make a frequency polygon by selecting **Line** from the chart options, as shown below. Compare it to the histogram above.



A frequency polygon showing the frequency of player ages

Both histograms and frequency polygons can display cumulative or relative frequencies if desired. The cumulative frequency polygon below was generated using the formula in column E.

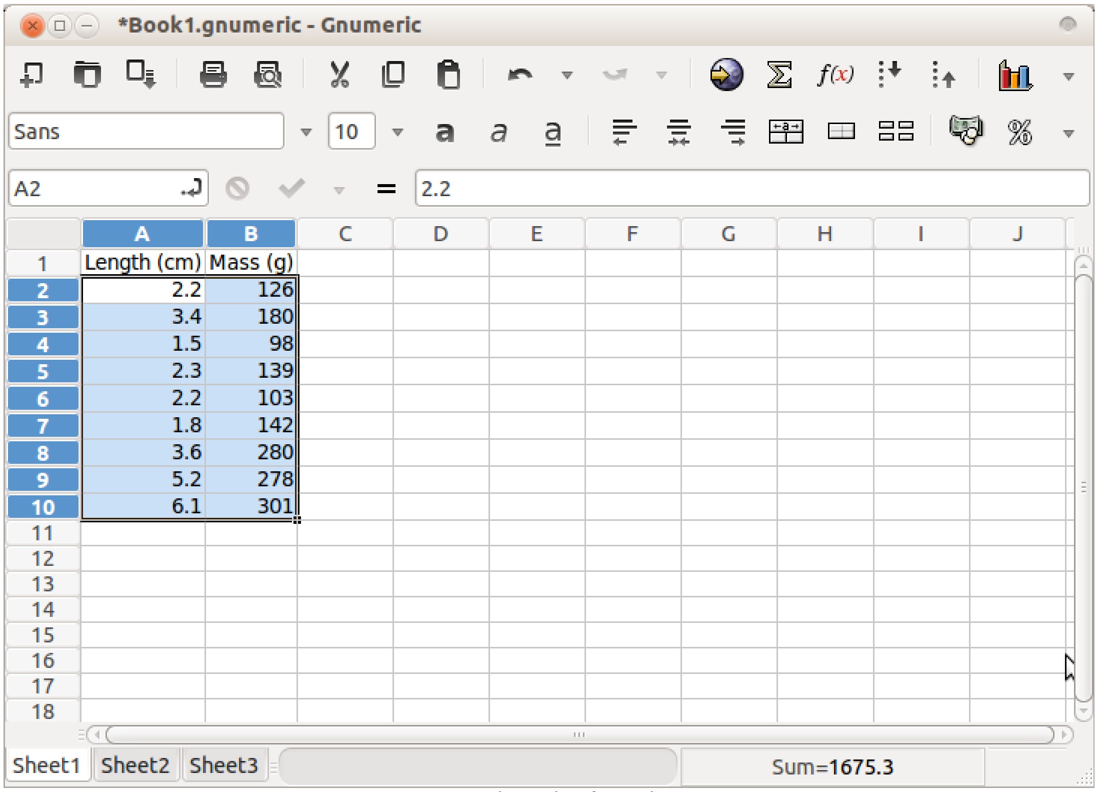


A cumulative frequency polygon showing the frequency of player ages

Similarly, the relative frequencies in column D were obtained by dividing each frequency in column C by the sum of all of the frequencies. For example, cell D1 contains the formula `=C1/sum(C3:C9)`.

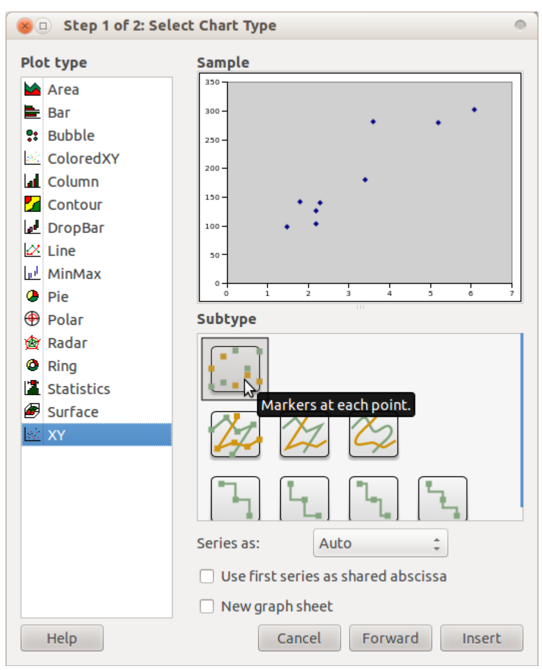
Scatter Plots and Regression Analysis

Consider the following data relating the length of an object, in centimetres, to its mass, in kilograms.



Mass vs. length of an object

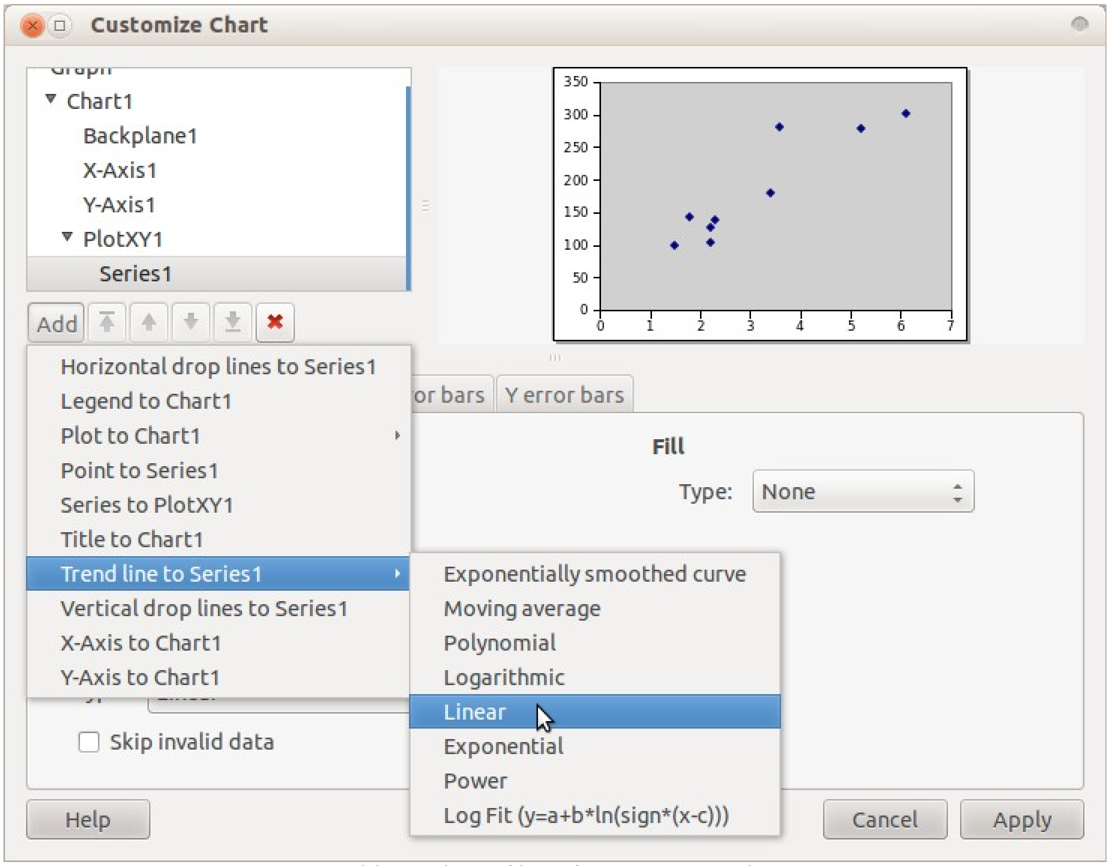
Select the data in columns A and B as shown, then insert a chart. Select the **XY** chart, as shown.



Creating a scatter plot

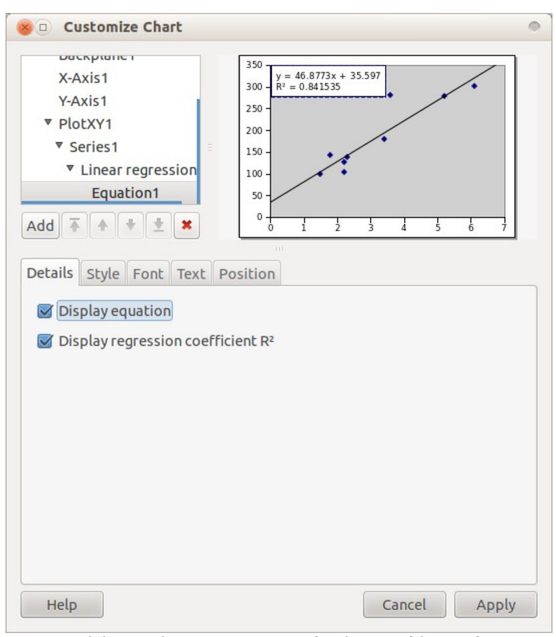
Place the chart on the worksheet, and right-click and select **Properties** to access the various options.

Select *Series1* from the list, and click the **Add** button. From this submenu, move down to **Trend Line**. There should be various types of trend lines, including linear, polynomial, and exponential. In this case, select a **Linear** trend line and you will see *Linear Regression* appear beneath *Series1* in the list, and a line of best fit on the scatter plot preview.



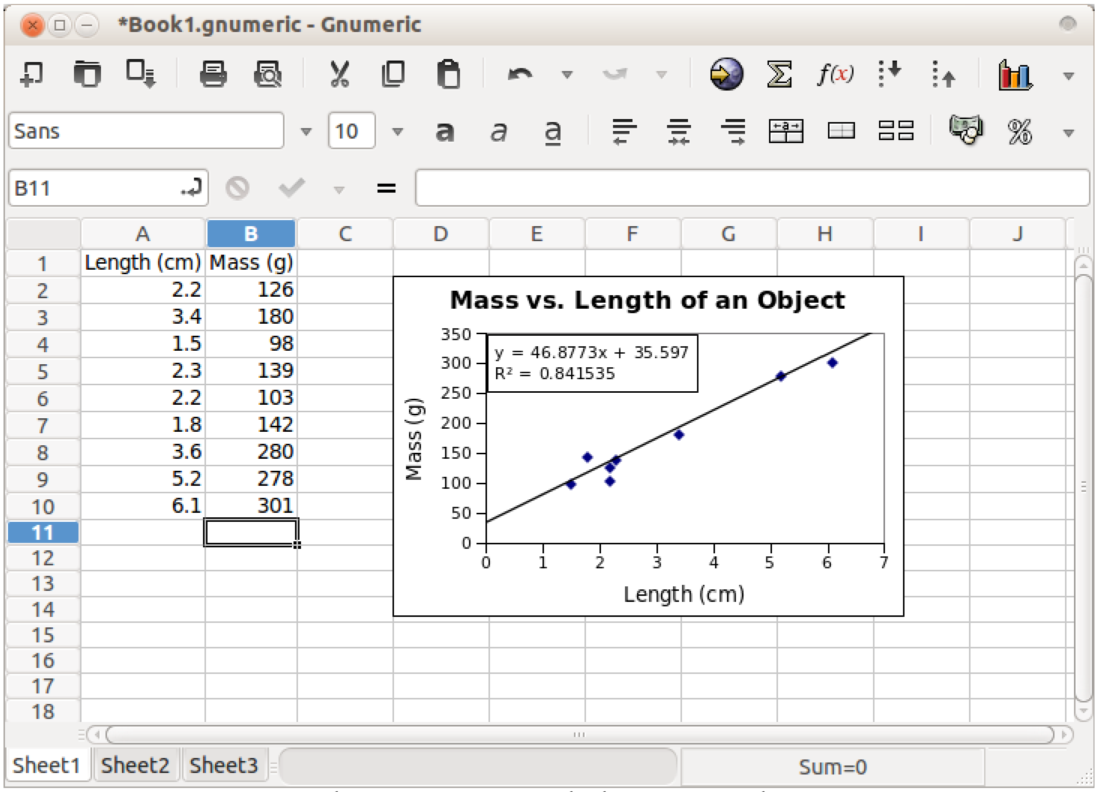
Adding a line of best fit to a scatter plot

To add the equation of the line of best fit (and the correlation coefficient, R^2), click **Add** again and select **Equation**. The result should look something like this.



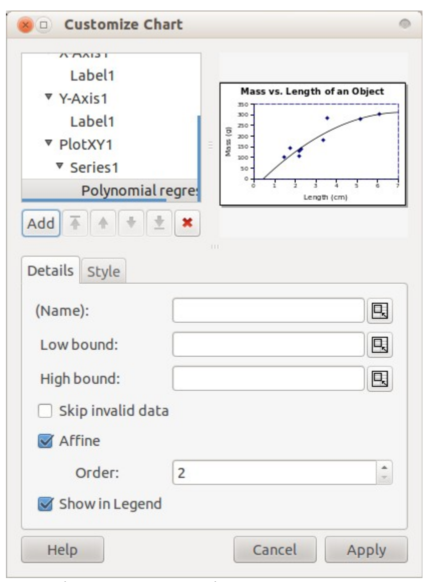
Adding the equation of a line of best fit

As with any chart, it is possible to label and decorate a scatter plot. A sample is below.



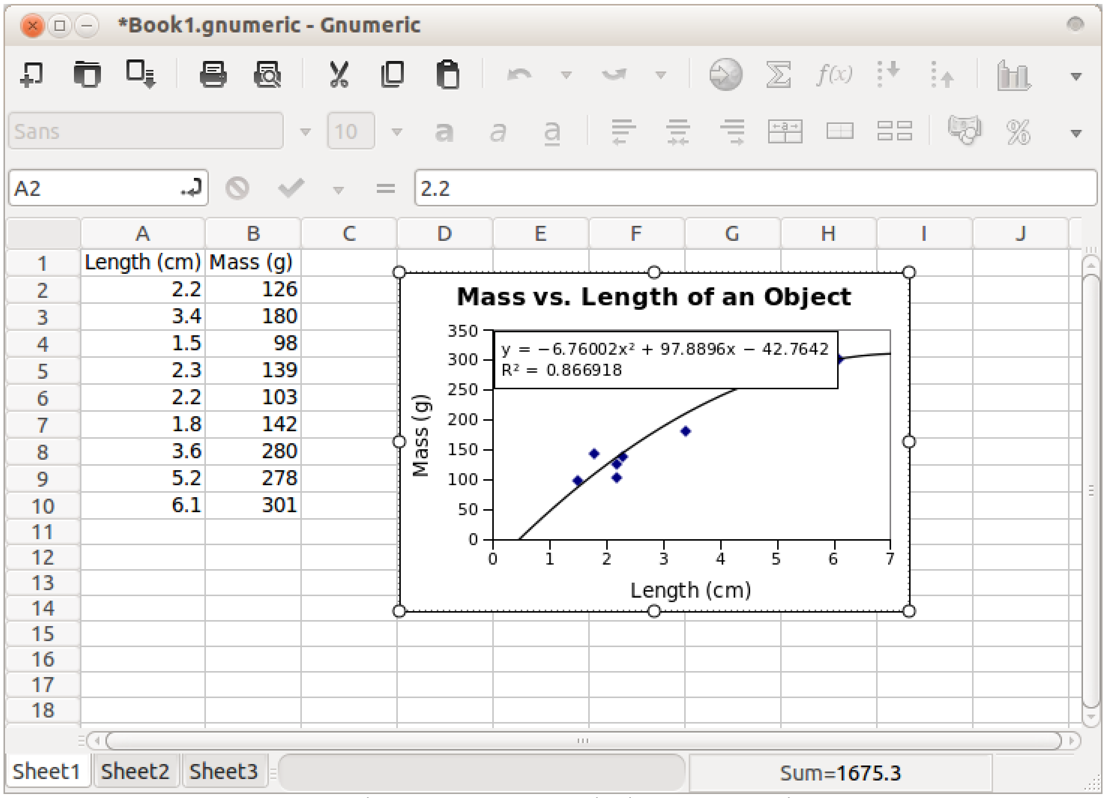
A linear regression applied to a scatter plot

To perform a different type of regression, select a different trend line from the available options. For example, to perform a quadratic regression instead, select a **Polynomial** trend line with an *Order* of 2, as depicted below.



Selecting a quadratic regression

Compare this curve of best fit to the line of best fit generated earlier. Note that the chart should probably be enlarged and the position of the equation changed so that it does not obscure the graph.



A quadratic regression applied to a scatter plot

Exercises

The file *1000cars.gnumeric* contains data regarding the age (in years) of 1000 cars observed on a highway during a weekend in June.

1. Determine the mean, median and mode of the data. Which measure of central tendency best describes the data? Explain.
2. Determine the range, the interquartile range, and the semi-interquartile range.
3. Are there any outliers? Explain.
4. Create a modified box-and-whisker plot for the data. Add a title, and labels to the axes.
5. Use the Histogram tool to produce a frequency table for the data.
6. Create a histogram for the data. Add a title and labels to the axes.
7. Create a *relative* frequency histogram for the data. Add a title and label to the axes.

The file *500marks.gnumeric* contains data regarding the marks (as percentages) of 500 students who wrote a standardized examination.

1. Determine the mean, median and mode of the data. Which measure of central tendency best describes the data? Explain.
2. Determine the range, the interquartile range, and the semi-interquartile range.
3. Are there any outliers? Explain.
4. Create a modified box-and-whisker plot for the data. Add a title, and labels to the axes.
5. Use the Histogram tool to produce a frequency table for the data.
6. Create a frequency polygon for the data. Add a title and labels to the axes.
7. Create a *cumulative* frequency polygon for the data. Add a title and labels to the axes.

The file *100bacteria.gnumeric* contains data regarding the number of bacteria observed in a petri dish after a fixed number of hours.

1. Using the data, perform a linear regression. Create a fully-labelled chart.
2. Calculate the correlation coefficient. Is the correlation strong, moderate or weak?
3. Perform an exponential regression. Create a fully-labelled chart.
4. Compare the coefficients of determination from both the linear and exponential regressions. Which model seems to be a better fit to the data?
5. Which model seems more reasonable to describe the data? Justify your choice.
6. Are there any extraneous variables that could affect the data? Explain.

The file *20exams.gnumeric* contains data regarding the final examination marks of 20 students after a fixed number of hours studying.

1. Using the data, perform a linear regression. Create a fully-labelled chart.
2. Calculate the correlation coefficient. Is the correlation strong, moderate or weak?
3. Determine the polynomial regression of lowest degree that results in a perfect correlation. Create a fully-labelled chart.
4. Does the polynomial regression provide a reasonable model to describe the data? Explain.
5. Are there any extraneous variables that could affect the data? Explain.