# Module Visualization in R for Quantitative Variables

# Introduction and Learning Objective

Introduction

Quantitative variables are variables taking numbers as values, such as weight, height, age, and etc. Quantitative variables are also called numeric variables. Most commonly used graph for showing the relationship between two quantitative variables is scatter plot. Other commonly used graphs for showing characteristics of a quantitative variable are sunflower plot, correlation plot, box plot, line chart, histogram and bag plot. In this module, we will introduce different R functions for plotting quantitative variables.

To summarize, we need to choose appropriate graphs for a given data. A summary of graph selection is included in the PDF file of *Graph Selection Matrix* on D2l.

Learning Objectives

After this module, students will be able to

* Generate scatter plot
* Generate sunflower plot
* Generate correlation plot
* Generate box plot
* Generate histogram
* Generate line chart
* Generate bag plot

#### Learning Objective One: Construct Scatter plots

***Introduction***

A scatter plot is a plot of paired (x,y) data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the independent variable (x), and the vertical axis is used for the dependent variable (y). A scatter plot shows the relationship between two quantitative variables (x and y). Each data point represents one individual.

***Knowledge***

If a scatter plot shows a pattern of approximately a straight line, then we say there is a linear correlation between the two quantitative variables. If the straight line goes from the origin to high x- and y-values (bottom left to top right), then the two variables are said to have a positive correlation. If the straight line goes from high values on the y-axis down to low y-values but high x-values (top left to bottom right), then the variables have a negative correlation. If the data points are very close to a straight line, then we say the two variables have a strong correlation. If the data points are scattered further around a line, then we say there is a weak (or no) correlation between the two variables. Figure 1 a, b, and c show the positive, negative and no correlation.

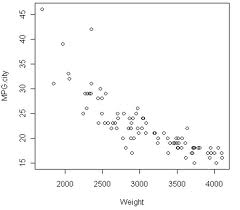
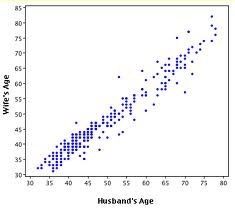
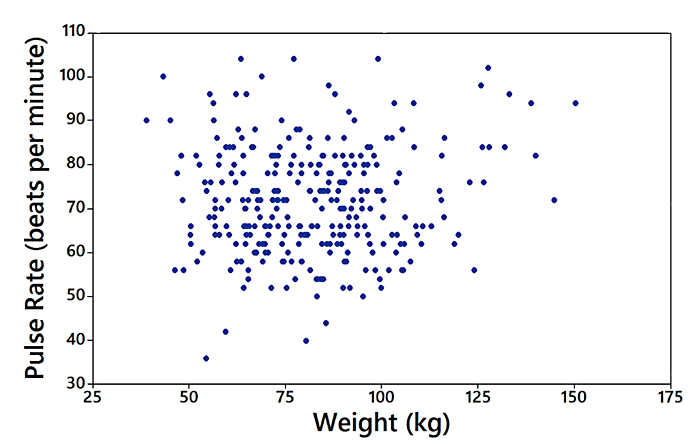
 

Figure 1.a) Figure 1.b) Figure 1.c)

Figure 1.a) A positive correlation between husband’s age and wife’s age; b) A negative correlation between cars’ weight and miles per gallon; and c) no correlation between cars’ weight and the pulse rate of the drivers

The R function to make a scatter plot is

plot(x,y, main= , xlab= , ylab= )

or

plot(y~x, data=)

* x: vector of data showing the x (independent variable) values of the individuals
* y: vector of data showing the y (dependent variable) values of the individuals
* main: title of the graph
* xlab: label of the x-axis
* ylab: label of the y-axis

A Youtube video on how to generate a scatter plot in R is here

<https://www.youtube.com/watch?v=yyXtiGCDOBo>

***Key Points to Remember***

A **scatter plot** shows the relationship between two quantitative variables. The relationship can be positive, negative or no correlation. The R function is **plot()**.

The presence of a correlation between two variables is not evidence that one of the variables causes the other. **Correlation does not imply causation**.

***Practice and Reflection***

***Practice***

Example 1: The temperature may associated with altitude. The research recorded the temperature at 7 altitudes, starting at 6327 feet. Use the scatter plot to show the relationship between altitude and temperature

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Altitude | 3 | 10 | 14 | 22 | 28 | 31 | 33 |
| Temperature | 57 | 37 | 24 | -5 | -30 | -41 | -54 |

The R code is



The scatter plot is shown in figure 2 below

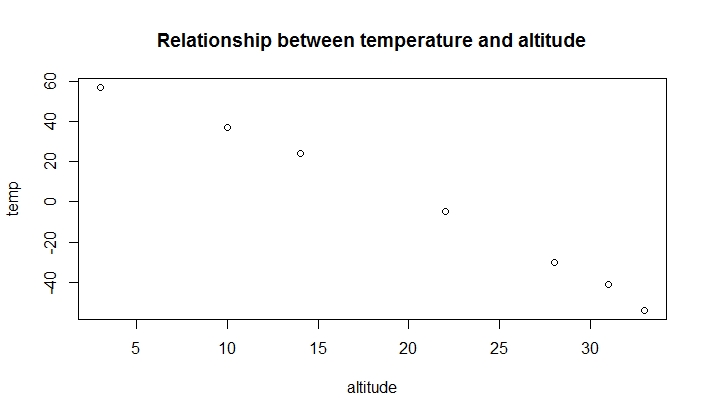


Figure 2. Scatter plot of the relationship between temperature and altitude.

Example 2: In the ChickeWeight data in R datasets, there are four Diet groups. Use the scatter plot to show the relationship between the weight and time for the four different diet groups. Show different color for the diet groups.

The R code is:



The graph is shown in figure 3.

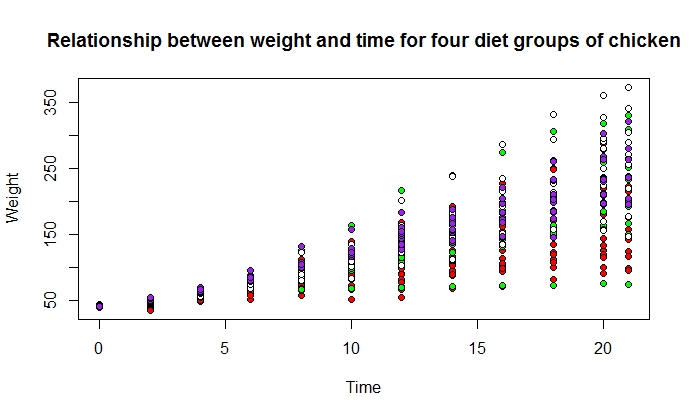


Figure 3. Scatter plot of the relationship between weight and time for the four types of diets.

***Reflection***

Activity 1: The table below lists weight and price of randomly selected six diamonds. All of the diamonds are round with rating of “very good” cut, and they all have color rating of F and clarity rating of VSI. Generate a scatter plot to show the relationship between price and weight

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Weight | 0.3 | 0.4 | 0.5 | 0.5 | 1.0 | 0.7 |
| Price | 510 | 1151 | 1343 | 1410 | 5669 | 2277 |

***Assessment***

1. The data below shows the systolic blood pressure (SBP) and age for males and females. Generate a scatter plot to show the relationship between SBP and age for males and females separately using different color

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SBP | 132 | 143 | 153 | 137 | 162 | 154 | 168 | 149 | 159 | 128 | 166 |
| Age | 52 | 59 | 67 | 54 | 73 | 64 | 74 | 61 | 65 | 46 | 72 |
| Gender | male | female | female | male | male | male | male | male | female | female | female |

1. We are interested in investigating the relationship between the percentage of children who have been immunized and the corresponding mortality rate for children under five in a country. We believe that the mortality rate for children under five in a country decreases as more children received vaccines. The table listed the data of the percentage of immunized children and the mortality rate for a randomly chosen 9 countries.

|  |  |  |
| --- | --- | --- |
| Country | Percentage immunized | Mortality rate per 1000 births |
| Bolivia | 77 | 118 |
| Cambodia | 32 | 184 |
| China | 94 | 43 |
| Czech Republic | 99 | 12 |
| Egypt | 89 | 55 |
| Ethiopia | 13 | 208 |
| Mexico | 91 | 33 |
| Senegal | 47 | 145 |
| Turkey | 76 | 87 |

Make a scatter plot to show the relationship between the percentage immunized and the mortality rate. Add title to the graph. Make the points solid and blue.

#### Learning Objective Two: Construct Sunflower graph

***Introduction***

A sunflower plot is similar to a scatter plot, but with showing the repeated points as sun flowers. In a sunflower plot, multiple overlapping points are plotted as sunflowers with multiples leaves (petals) with each leave (petal) representing one individual.

***Knowledge***

The R function for creating a sunflower plot is

sunflowerplot(x, y, cex= , size= , seg.col = , seg.lwd = )

* x and y: numeric vectors of the data
* cex: character size expansion of center points
* size: size of sunflower petals in inches
* seg.col: color to be used for the segments which make the sunflower petals
* seg.lwd: the line width for the sunflower petal segments

An example of a sunflower plot is shown in figure 4. The graph is based on a data for 100 low birth weight infants, and it shows the relationship between the gestational age and birth weight. The horizontal axis shows the gestational age (in weeks), and vertical axis shows the birth weight (in grams) of the 100 infants. Each sun flower shown in red represents that there are multiple infants with the same gestational age and same birth weight. For example, the sun flower point at the gestational age of 27 weeks and weight of 1170 g indicates that there are four infants with the same gestational age of 27 weeks and same weight of 1170 grams.

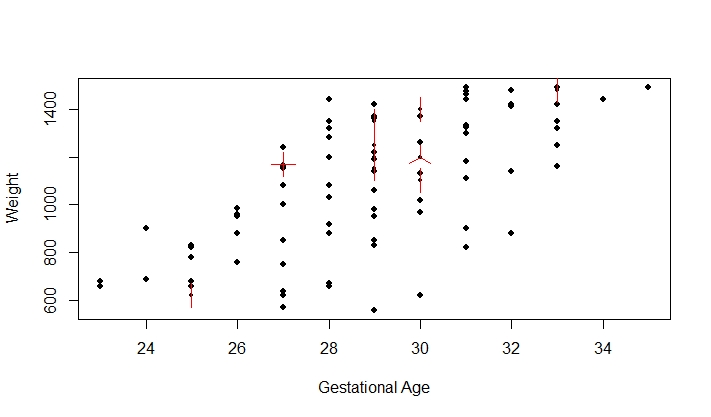


Figure 4. A sunflower plot of relationship between gestational age and birth weight for 100 infants.

***Key Points to Remember***

Both a scatter plot and **a sunflower plot** are good to show the relationship between two quantitative variables. A regular scatter plot does not show the repeated individuals as those individuals overlap, but a sunflower plot shows the repeated individuals as different petals. The R function is **sunflowerplot()**.

***Practice and Reflection***

***Practice***

Example 3: The following data shows two aspirin concentrations in urine samples of 10 patients at 1 hour. The 10 patients took Aspirin A first and had their urine samples taken, and then took Aspirin B and had their urine samples taken. The concentration of aspirin in urine samples is measured as mg%. Generate a sunflower plot to see the relationship between the urine samples from Aspirin A and Aspirin B.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Aspirin A | 15 | 26 | 15 | 28 | 17 | 26 | 15 | 20 | 18 | 15 | 26 | 15 |
| Aspirin B | 13 | 20 | 13 | 21 | 12 | 20 | 13 | 22 | 11 | 13 | 20 | 13 |

The R code is



The sunflower plot is shown in figure 5.

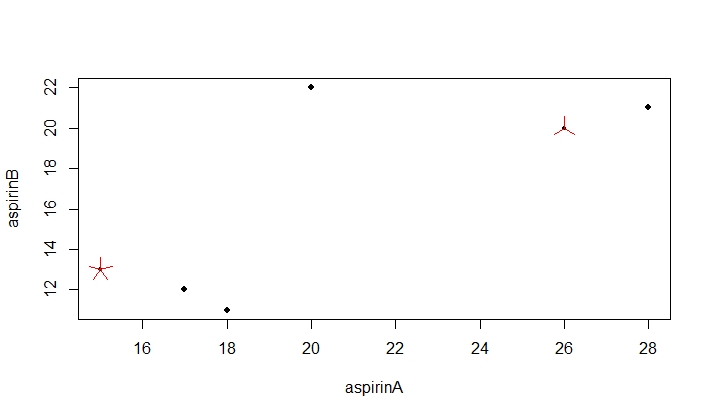


Figure 5. Sunflower plot of relationship between the Aspirin A concentration and Aspirin B concentration.

We can see that there are five patients having the same Aspirin A concentrations at 15 mg% and the same Aspirin B concentrations at 13 mg%. There are three patients having the same Aspirin A concentration at 26 mg%, and the same Aspirin B concentration at 20mg%.

***Reflection***

Activity 2: Use the R data Orange to generate a sunflower plot to show the relationship between age and circumference of the orange trees. Make the data points blue, and the sunflower green. How many sunflowers do you observe? What data points are the sunflowers represent?

***Assessment***

Use the R dataset Orange. Generate a sunflower plot to show the relationship between age and circumference. How many repeated individuals did you see?

#### Learning Objective Three: Construct Correlation plots

***Introduction***

A correlation plot is a two-dimensional plot showing the relationship between multiple pairs of numeric variables. The strength of the relationship is shown by the shape of the ellipse.

***Knowledge***

The R syntax of correlation plot is

corrplot (corr, method= , type = , col = , bg =)

* corr: the correlation matrix of variables to visualize
* method: the default method is “circle”. The areas of circles show the absolute value of corresponding correlation coefficients. If use “ellipse”, then the direction of the ellipse shows the negative or positive correlation, and the area of the ellipse shows the corresponding correlation coefficients.
* type: character in “full”, “upper” or “lower”, displaying full matrix, lower triangular or upper triangular matrix. The default setting is “full”.
* col: Color of glyphs.
* bg: Background color

A Youtube video on how to make a complex correlation plot is here

<https://www.youtube.com/watch?v=jxUiIFj2l-s>

In a correlation plot, the ellipse going from the top left corner to the bottom right corner represents a negative correlation between the two variables. The ellipse from the bottom left to the top right represents a positive correlation. The strength of the correlation is shown by the thickness of the ellipse. A very narrow ellipse shows a very strong correlation, and a very wide ellipse shows a weak correlation. If the correlation is represented by a circle, that means no correlation between the two variables.

***Key Points to Remember***

A **correlation plot** shows the strength of correlation between two quantitative variables. The R function is **corrplot()**.

***Practice and Reflection***

***Practice***

Example 4: Plot a correlation matrix between the variables in the dataset mtcars. The R code is



The graph is shown in figure 6

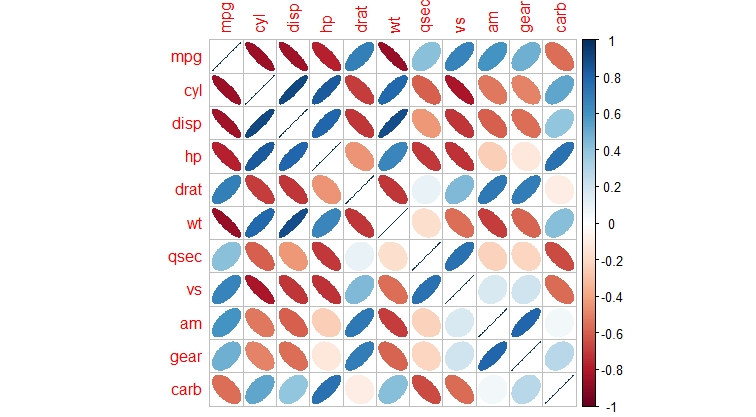


Figure 6. A correlation plot between every two numeric variables in the data set mtcars.

In figure 6, the red ellipse representing the relationship between miles per gallon (mpg) and weight (wt) shows us a negative correlation. A positive correlation is shown by an ellipse in figure 6 between number of cylinder (cyl) and display (disp).

***Reflection***

Activity 3: Use the R dataset iris and make a correlation matrix of variables containing only numeric variables in iris data. Generate a correlation plot using ellipse method.

***Assessment***

Use the R data airquality to generate a correlation plot to show the relationship between variables. Identify the positive and negative correlation.

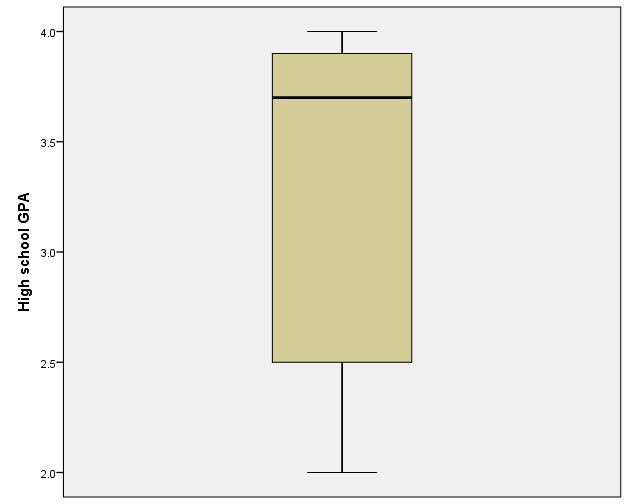
#### Learning Objective Four: Construct Box plots

***Introduction***

A box plot shows the five number summery of a quantitative variable. The five number summary includes minimum, first quartile (), median, third quartile (), and maximum. Outliers are either or more above the third quartile or or more below the first quartile.

***Knowledge***:

An example of a box plot is shown in figure 7. Outliers may be plotted as individual points.



Median

Maximum

Minimum

Figure 7. An example of a box plot.

A single box plot can be drawn for one batch of data with no distinct groups. Alternatively, multiple box plots can be drawn together to compare multiple data sets. The R function for boxplot is

boxplot(x, boxwex=, varwidth=, border=, col=, horizontal=)

or

boxplot(y~group, data=, width=, varwidth=, border=, col=, horizontal=)

* x: vector of data
* y~group: y is a numeric vector of data values to be split into groups according to the grouping variable, which is a factor.
* boxwex: a number to indicate the width of the box
* varwidth: logic value. If varwidth is TRUE, the boxes are drawn with widths proportional to the square-roots of the number of observations in the groups
* border: an optional vector of colors for the outlines of the boxplots.
* col: if col is non-null, it is assumed to contain colors to be used to color the bodies of the box plots. By default they are in the background color.
* horizontal: a logical value indicating if the boxplots should be horizontal; default FALSE means vertical boxes

***Key Points to Remember***

A **box plot** shows the five number summary of a variable. The R function is **boxplot()**.

***Practice and Reflection***

***Practice***

Example 5: Make a box plot for the ten students grades shown in the following table

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 96 | 67 | 82 | 51 | 97 | 76 | 85 | 88 | 93 | 75 |



The box plot is shown in figure 8 below

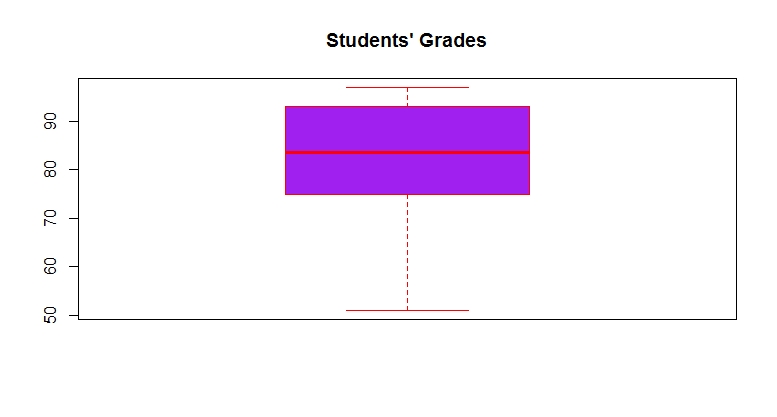


Figure 8. Box plot of students’ grades.

Example 6: Use R data mtcars to generate a box plot to show the five number summary of miles per gallon for the three types of cylinders. The R code is



The boxplot is shown in figure 9 below:

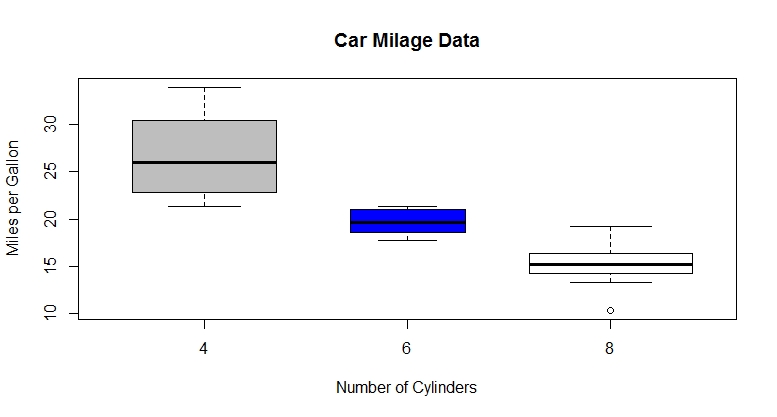


Figure 9. Box plot of miles per gallon for 4-cylinder, 6-cylinder, and 8-cylinder cars.

***Reflection***

Activity 6 Use R data ToothGrowth to generate a box plot to show the five number summary of tooth length for each dosage.

***Assessment***

1. If there are 12 students in a statistics class and the 12 students’ grades are

72 83 95 75 78 86 88 93 75 89 85 97

Generate a box plot to show the five number summary. Make the width of the box as 0.4.

1. Use the R dataset ToothGrowth to generate multiple box plots in one graph to show the five number summary of tooth length(len) for the two supplement type (supp). Add x- and y-axis labels. Add title and legend to the graph.

#### Learning Objective Five: Construct Histograms

***Introduction***

A histogram is a display of statistical information that uses bars to show the frequency of data in successive numerical intervals of equal size. It is plotted by dividing the range of observations into equal sized bins, and then counting the number of values appearing in each bin. The bars represent the bins, and the height of each bin represents the number of observations in each bin.

***Knowledge***

The R function of making a histogram is hist(x, breaks= , col= ,border= ,main= , xlab= , ylab= , xlim= , ylim= )

* x: a vector of values for which the histogram is desired
* breaks: one of the following
  + a vector giving the breakpoints between histogram bins
  + a function to compute the vector of breakpoints
  + a number giving the number of bins for the histogram
  + a function to compute the number of bins
* col: a color to be used to fill the bars; the default yields unfilled bars
* border: a color of the border of bards. The default is black
* main: title of the graph
* xlab: label of the x-axis
* ylab: label of the y-axis
* xlim: range of x values on the x-axis
* ylim: range of the y values on the y-axis

***Key Points to Remember***

A **histogram** shows frequency of data in successive numerical intervals of equal size. A bell-shape and symmetric histogram represents a normal distribution. The R function to generate a histogram is **hist()**.

***Practice and Reflection***

***Practice***

Example 7: Use the R data AirPassengers to generate a histogram to show the distribution of the number of passengers



The histogram is shown in figure 10

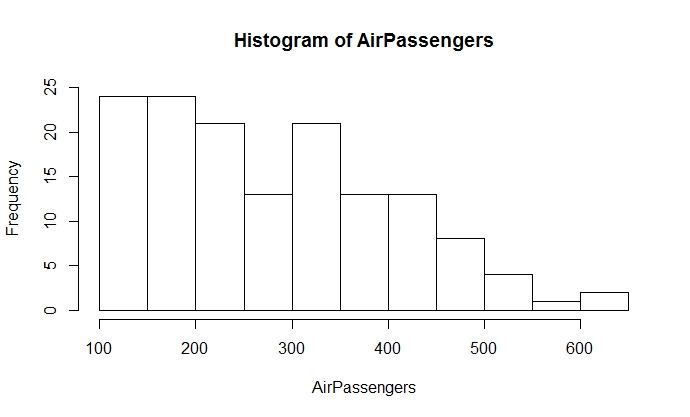


Figure 10. A histogram of air passengers

Figure 10 shows that there are about 24 flights with 100 to 150 passengers, 24 with 150 to 200 passengers, about 21 flights with 200 to 250 passengers, etc.

If we would like to make only 5 bars (bins) for the histogram, then the R code is



The histogram is shown in figure 11

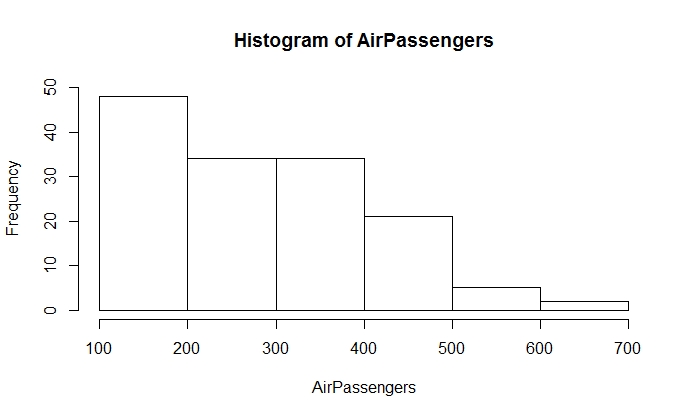


Figure 11. Histogram of airline passengers with fixed number of bins (n=5)

If we would like to add color to the histogram and make the bars grey and the borders red, and break the bins at 100, 300, 500, and 700, then the R code is



The histogram is shown in figure 12

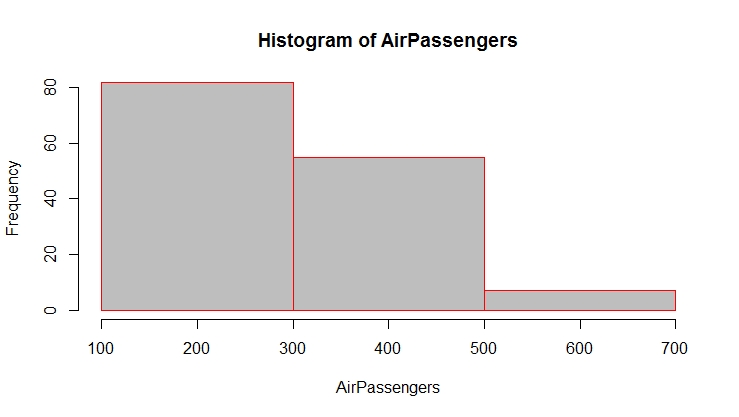


Figure 12. Histogram of airline passengers with the bins of between 100 to 300, 300 to 500, and 500 to 700. The bar color is grey, and border color is red.

***Reflection***

Activity 7: a) Use the R data iris to generate a histogram for sepal length of that dataset; b) Use “breaks=c(seq(4,8,0.4) to try the new break points for the bins and state the size of the bins and the number of the bins observed; c) make the color of the bars red and border purple.

***Assessment***

If there are 12 students in a statistics class and the 12 students’ grades are

72 83 95 75 78 86 88 93 75 89 85 97

Generate a histogram to show the frequency of grades with the bins break at 70, 80, 90 and 100; Make the bars grey and the borders black. Label the x-axis as “grades” and y-axis as “Frequency”. Title the graph as “Student Grades”.

#### Learning Objective Six: Construct Line charts

***Introduction***

A Line chart is usually used to investigate the changes over time for one or more groups. A line chart displays information as a series of data points connected by straight line segments. The horizontal axis represents time, and the vertical axis represents the measurements over time.

***Knowledge***

A line chart is similar to a scatter plot except that the measurements are ordered over time and joined with line segments. A line chart is generated based on a scatter plot in R. After plotting the scatter plot, use the R function

lines (x, type= )

to add line segments. The argument type is a character giving the type of lines desired, in which “p” is for points, “l” for lines, “b” for both lines and points, “c” for empty points joined by lines, “o” for overplotted points and lines, “s” for stair steps, and “h” for histogram like lines.

***Key Points to Remember***

A **line chart** shows the measurements through time. The R function is lines(). You may or may not plot the data points using plot() first.

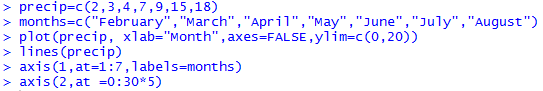
***Practice and Reflection***

***Practice***

Example 8: The following data shows the precipitation amount (in inches) from February to August in Minnesota in 2018 (based on National Weather Service). Generate a line graph to show the rain fall amount over time.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Month | February | March | April | May | June | July | August |
| Precipitation | 2 | 3 | 4 | 7 | 9 | 15 | 18 |

The R code is



The argument “axes=FALSE” in plot is to remove the x- and y-axis labels. The code

“axis(1, at=1:7, labels=months)” is to label the x-axis with the eight months, and the code

“axis (2, at = 0:30\*5)” is to label the y-axis with an interval of 5. The line graph is shown in figure 13.

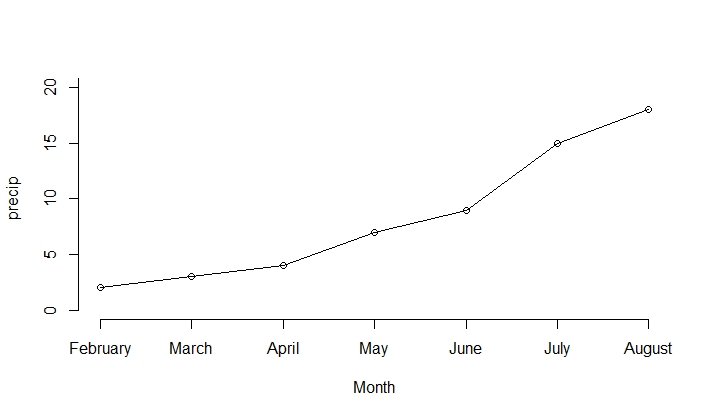


Figure 13. Line graph of precipitation from February to August in Minnesota in 2018.

Or we can use the argument *type=”b”* in the R function plot().

From the line chart, we observe that the precipitation is increasing through time.

***Reflection***: Activity 8: The table below shows the number of fatalities of law enforcement officers in the United States (Triola, 2018). Show the trend using a line chart. Where is the maximum point? What happened in that year?

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
| Number of Fatalities | 140 | 170 | 169 | 145 | 158 | 162 | 240 | 161 | 150 | 170 | 165 |

***Assessment***

The mortality rates (per 1000 births) of new born infants in a country from 2000 to 2010 are shown in the table below. Generate a line chart to show the trend of mortality rate through years.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
| 9.5 | 8 | 8.1 | 7.6 | 6.9 | 5.1 | 3.2 | 2.1 | 2.5 | 1.1 | 0.9 |

#### Learning Objective Seven: Construct Bagplots

***Introduction***

A bagplot (or starburst plot) is a two-dimensional extension of a boxplot that shows the joint distribution of two numeric variables.

***Knowledge***

The R function bagplot is in the package aplpack. You will need to install that package before run the code. The syntax of bagplot is as following:

bagplot( x, y, factor =3, approx..limit =300, show.outlier = TRUE, show.whiskers = TRUE, show. looppoints =TRUE, show.bagpoints =TRUE, show.loophull =TRUE, show.baghull =TRUE, pch =16, cex =0.4, col.loophull=”#aaccff”, col.looppoints=”#3355ff”, col.baghull=”#7799ff”, col.bagpoints=”#000088”)

* x and y: x and y values of a data set
* factor: factor defining the loop (outer light blue region); the default is 3 times the interior dark blue region
* approx.limit: if the number of data points exceeds approx.limt a sample is used to compute some of the quantities; default is 300
* show.outlier: if TRUE outlier are shown
* show.whiskers: if TRUE whiskers (red lines connecting loop points and the inner yellow region) are shown
* show.looppoints: if TRUE loop points (points between the interior dark blue region and the outer light blue region) are shown
* show.bagpoints: if TRUE bag points (points inside of the dark blue interior region) are shown
* show.loophull: if TRUE loop (outer light blue region) is shown
* show.baghull: if TRUE bag (interior dark blue region) is shown
* pch: point type, default is a solid point
* cex: size of characters
* col.loophull, col.looppoints, col.baghull, and col.bagpoints: color of the loop hull (outer light blue region), loop points, bag hull (inner dark blue region), and bag points. The default loop hull color is light blue, default bag hull color is dark blue. The default loop points and bag points color are black.

***Practice and Reflection***

***Practice***

Example 9: the data below shows the relationship between the heights of winning presidential candidates and heights of their main opponents in cm (Triola, 2018).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| President | 178 | 182 | 188 | 175 | 179 | 183 | 192 | 182 | 177 | 185 | 188 | 188 | 183 | 188 |
| Opponent | 180 | 180 | 182 | 173 | 178 | 182 | 180 | 180 | 183 | 177 | 173 | 188 | 185 | 175 |

The bagplot is shown in figure 14.

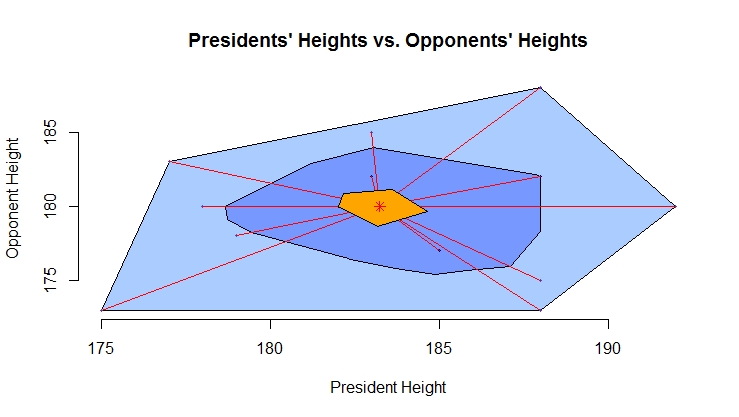
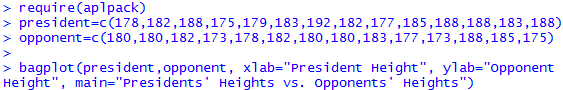


Figure 14. Bagplot of the presidents’ heights vs. opponents’ heights

The red asterisk near the center of the graph represents the median heights of the presidents and opponents. The dark blue interior region is called the bag, representing the box of the traditional boxplot, which contains 50% of the data. The light blue outer region is called the fence, and it is formed by inflating the bag by a certain factor, usually 3. Points outside the fence are the outliers. The light blue region is also called loop hull. The points outside of the dark blue region and inside of the light blue region are called loop points. The red lines connecting the central yellow region and the loop points are called whiskers.

The R code to obtain the above bagplot is



***Reflection***: Activity 9: Use the R dataset mtcars to generate a bagplot for weight and miles per gallon (mpg). Label the x- and y-axis and give the graph a title.

***Assessment***

Use the R dataset Orange to generate a bag plot to show the relationship between age and circumference.