

Introduction to R

Basic Graphics

Jan-Philipp Kolb

10 März, 2020

FIRST THINGS TO DO

Don't try to kiss your data on the first date; rather, you just want to get to know the data:

1. Import the data
2. Review the codebook
3. Learn about the data
4. **Quick visual understanding of the data**

A plot says more than 1000 words

Statements on graphs in R

- Graphical data analysis is great
- Good plots can contribute to a better understanding
- Generating a plot is easy
- Making a good plot can take very long
- Generating plots with R is fun
- Plots created with R have high quality
- Almost every plot type is supported by R
- A large number of export formats are available in R

\small

\normalsize

Graphic types



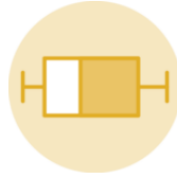
Violin



Density



Histogram



Boxplot



Ridgeline

Correlation



Scatter



Heatmap



Correlogram



Bubble



Connected scatter



Density 2d

Ranking



Barplot



Spider / Radar



Wordcloud



Parallel



Lollipop



Circular Barplot

Task View for graphics

CRAN Task View: Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization

Maintainer: Nicholas Lewin-Koh

Contact: [nikko at hailmail.net](mailto:nikko@hailmail.net)

Version: 2015-01-07

URL: <https://CRAN.R-project.org/view=Graphics>

R is rich with facilities for creating and developing interesting graphics. Base R contains functionality for many plot types including coplots, mosaic plots, biplots, and the list goes on. There are devices such as postscript, png, jpeg and pdf for outputting graphics as well as device drivers for all platforms running R. [lattice](#) and grid are supplied with R's recommended packages and are included in every binary distribution. [lattice](#) is an R implementation of William Cleveland's trellis graphics, while grid defines a much more flexible graphics environment than the base R graphics.

R's base graphics are implemented in the same way as in the S3 system developed by Becker, Chambers, and Wilks. There is a static device, which is treated as a static canvas and objects are drawn on the device through R plotting commands. The device has a set of global parameters such as margins and layouts which can be manipulated by the user using `par()` commands. The R graphics engine does not maintain a user visible graphics list, and there is no system of double buffering, so objects cannot be easily edited without redrawing a whole plot. This situation may change in R 2.7.x, where developers are working on double buffering for R devices. Even so, the base R graphics can produce many plots with extremely fine graphics in many specialized instances.

One can quickly run into trouble with R's base graphic system if one wants to design complex layouts where scaling is maintained properly on resizing, nested graphs are desired or more interactivity is needed. grid was designed by Paul Murrell to overcome some of these limitations and as a result packages like [lattice](#), [ggplot2](#), [vcd](#) or [hexbin](#) use grid for the underlying primitives. When using plots designed with grid one needs to keep in mind that grid is based on a system of viewports and graphic objects. To add objects one needs to use grid commands, e.g., `grid.polygon()` rather than `polygon()`. Also grid maintains a stack of viewports from the device and one needs to make sure the desired viewport is at the top of the stack. There is a great deal of explanatory documentation included with grid as vignettes.

The graphics packages in R can be organized roughly into the following topics, which range from the more user oriented at the top to the more developer oriented at the bottom. The categories are not mutually exclusive but are for the convenience of presentation:

<https://cran.r-project.org/web/views/Graphics.html>

The example dataset

\small

```
install.packages("AmesHousing")
```

\normalsize

\small

```
ames_df <- AmesHousing::make_ames()
```

\normalsize

Variables used in this section

- **Lot_Area**: Lot size in square feet
- **Alley**: Type of alley access to property
- **Street**: Type of road access to property

Histogram - The `hist()` function

We create a histogram of the variable `duration`:

```
\small
```

```
?hist
```

```
\normalsize
```

```
\small
```

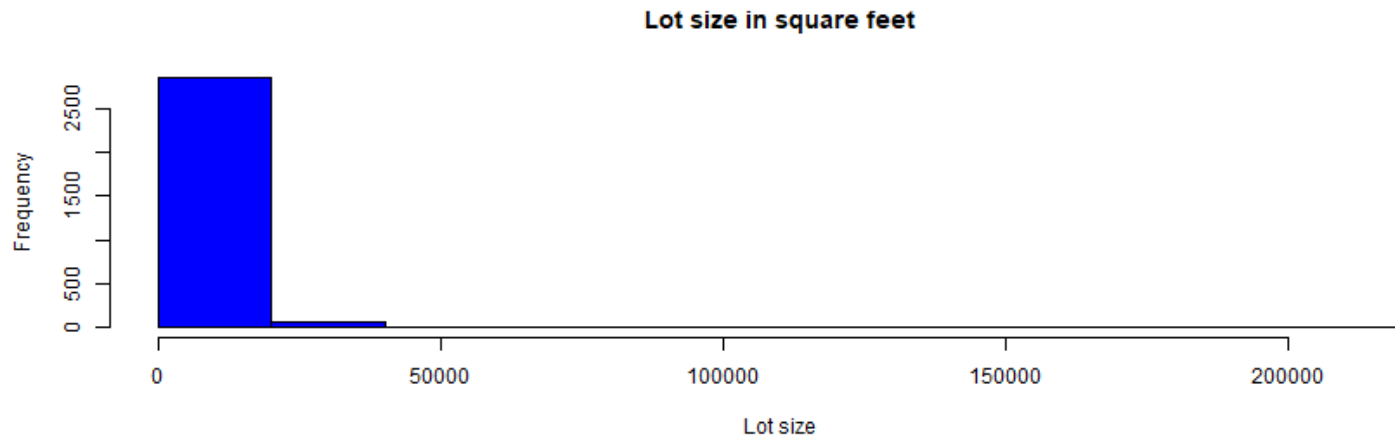
```
hist(ames_df$Lot_Area)
```

Histogram

- Command `hist()` plots a histogram
- At least one observation vector must be passed to the function
- `hist()` has many more arguments, which all have (meaningful) default values

\small

```
hist(ames_df$Lot_Area,col="blue",  
     main="Lot size in square feet",ylab="Frequency",  
     xlab="Lot size")
```



Further arguments:

- Many of the arguments are valid for all base graphics like `main` or `xlab`.
- You can see many of them if you call help on `?par`.

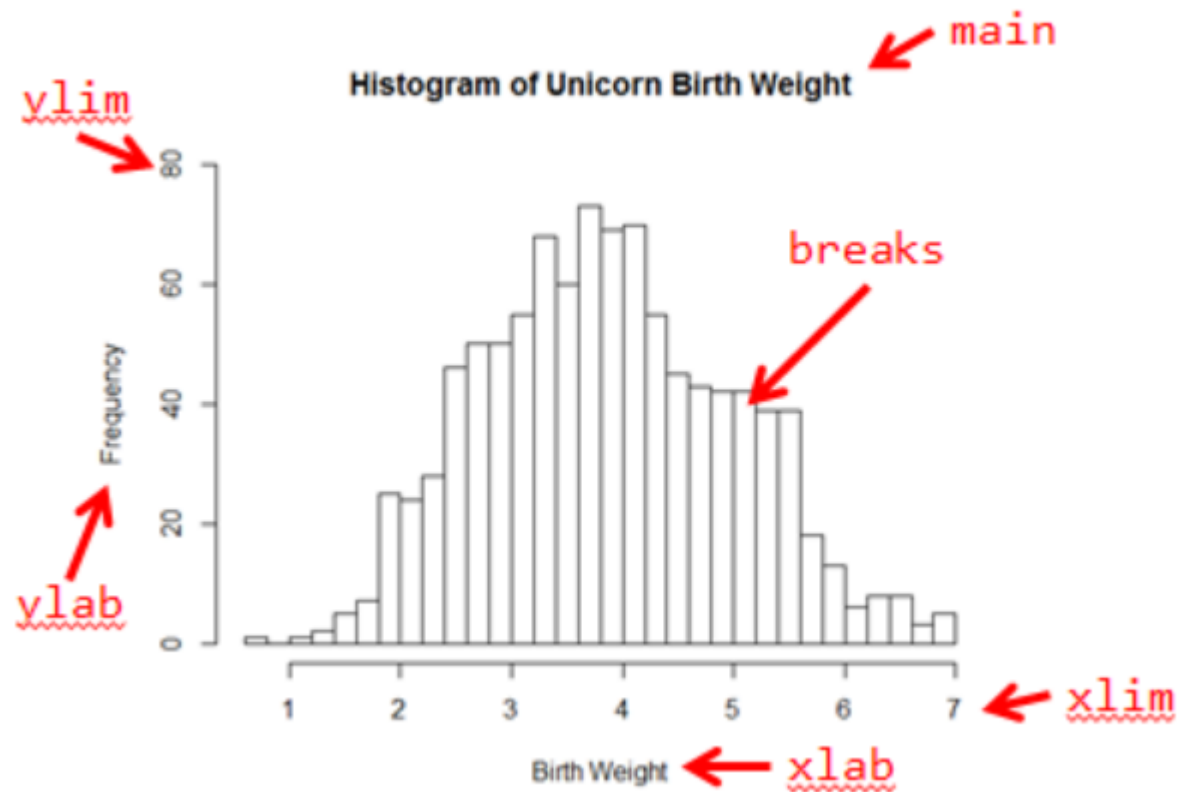
`\small`

```
?plot  
# or  
?par
```

`\normalsize`

`\small`

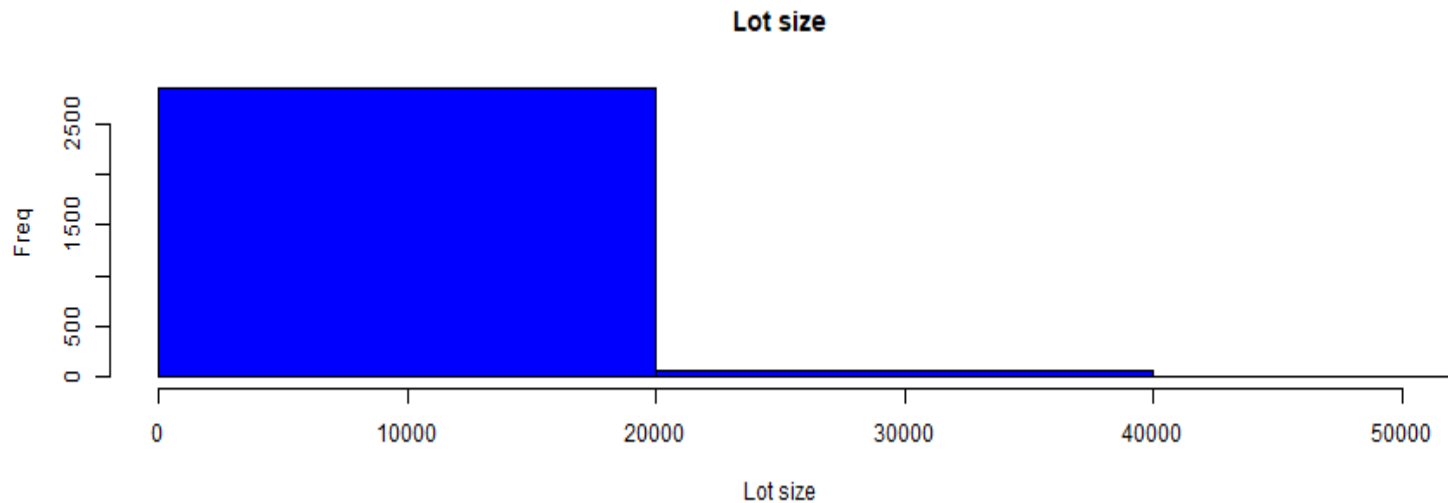
Some graphic parameters



The `xlim` argument

\small

```
hist(ames_df$Lot_Area,col="blue",  
     main="Lot size",ylab="Freq", xlab="Lot size",  
     xlim=c(0,50000))
```



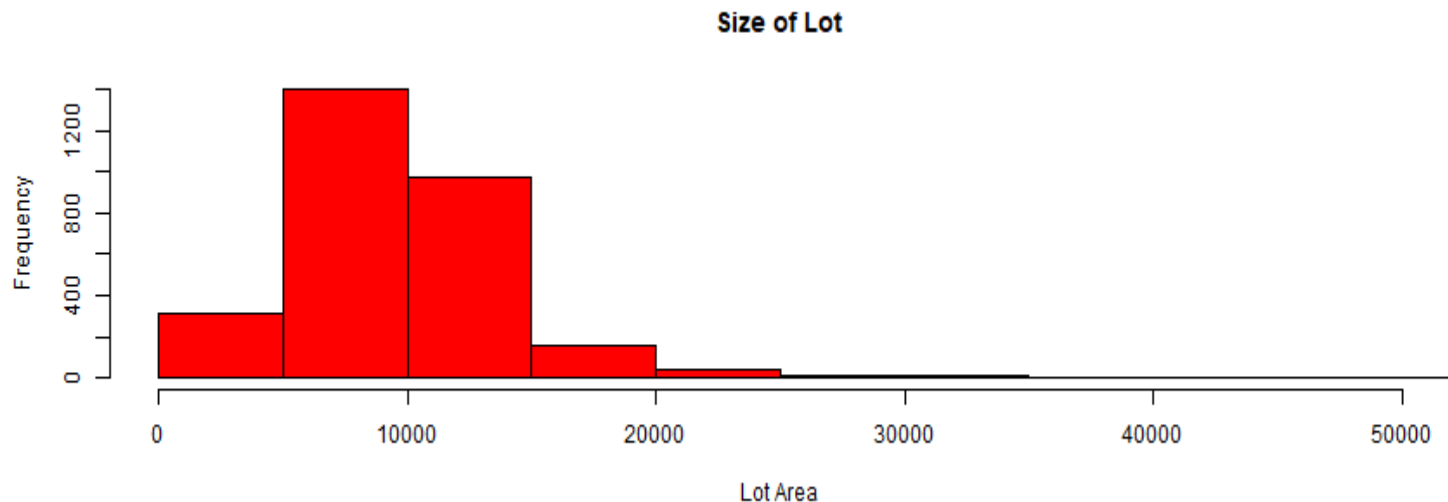
\normalsize

The `breaks` argument

- While the previous arguments are valid for many graphics functions, the following apply mainly to histograms:

\small

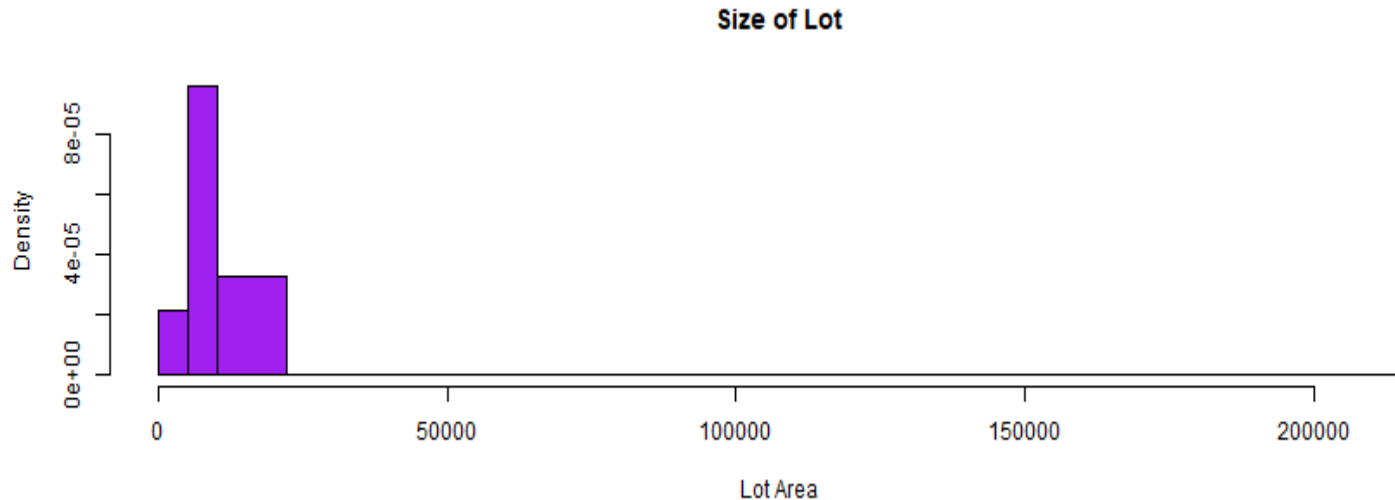
```
hist(ames_df$Lot_Area,col="red",  
     main="Size of Lot", xlab="Lot Area",  
     xlim=c(0,50000),breaks=60)
```



A vector for the breaks

\small

```
hist(ames_df$Lot_Area,col="purple",  
     main="Size of Lot", xlab="Lot Area",breaks=c(0,5000,10000,22000,
```



\normalsize

Tabulate and `barplot`

- The command `barplot()` generates a barplot from a frequency table
- We get the frequency table with the following command:

\small

```
tab_alley <- table(ames_df$Alley)
```

\normalsize

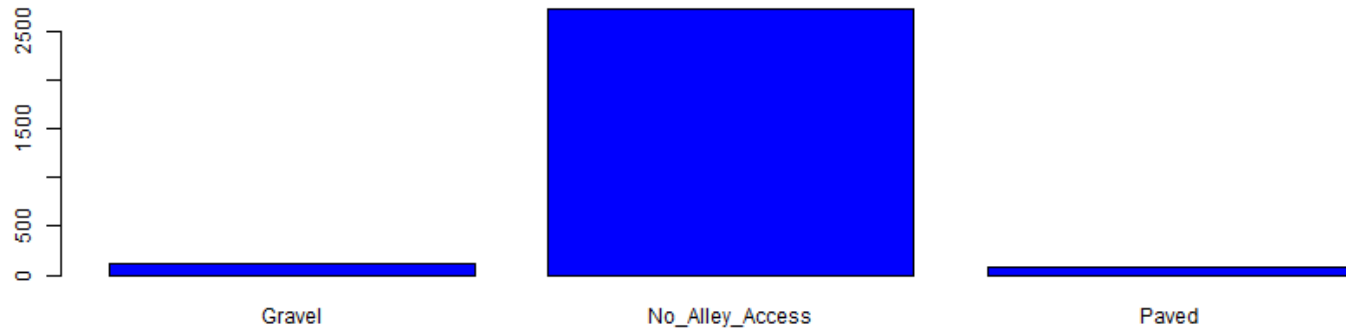
\small

```
barplot(tab_alley)
```

More colour:

\small

```
barplot(tab_alley,col=rgb(0,0,1))
```

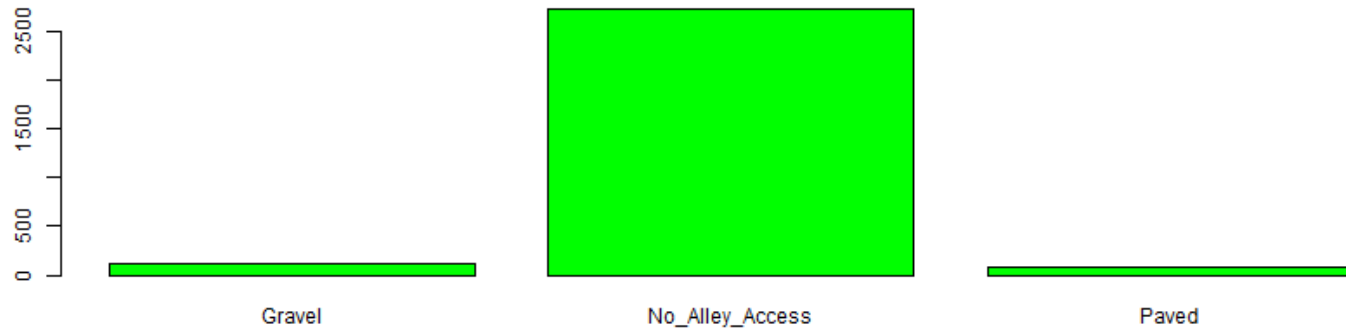


\normalsize

Green colour

\small

```
barplot(tab_alley,col=rgb(0,1,0))
```

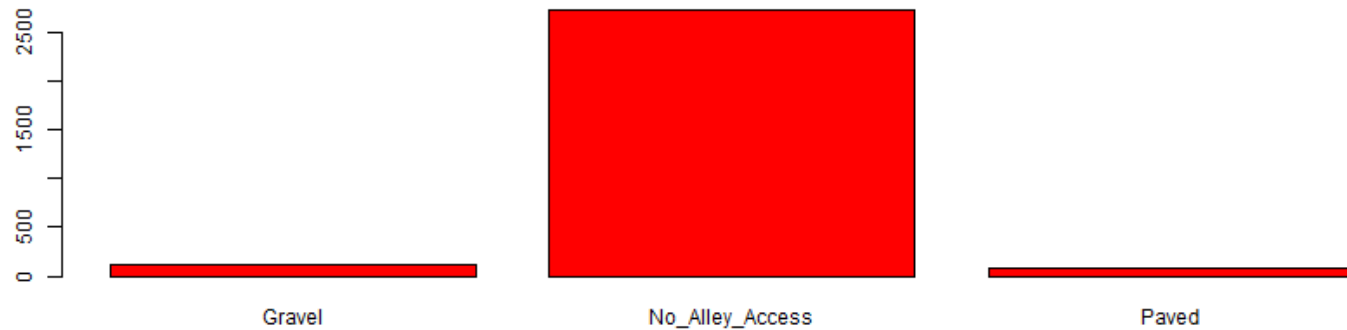


\normalsize

Red colour

\small

```
barplot(tab_alley,col=rgb(1,0,0))
```

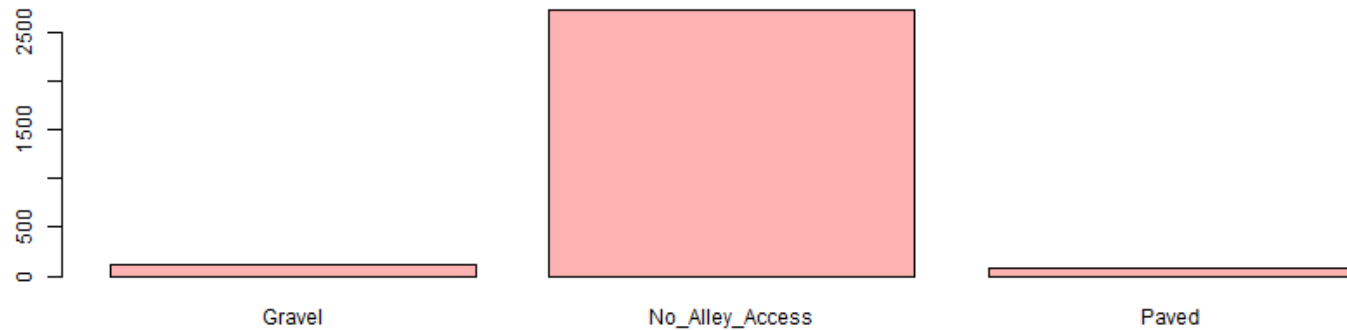


\normalsize

Transparent

\small

```
barplot(tab_alley,col=rgb(1,0,0,.3))
```



\normalsize

Rstudio addin colourpicker

\small

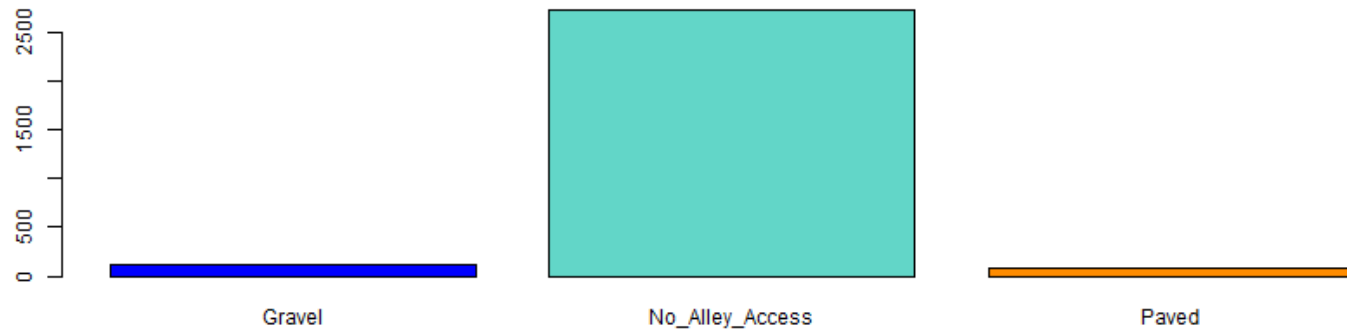
```
install.packages("colourpicker")
```

\normalsize

Set various colors

\small

```
barplot(tab_alley,col=c(20,"#62D6C8", "darkorange"))
```



\normalsize

A two dimensional table

\small

```
tab2dim <- table(ames_df$Alley,ames_df$Street)
```

\normalsize

- If the passed table object is two-dimensional, a conditional barplot is created

\small

```
tab2dim
```

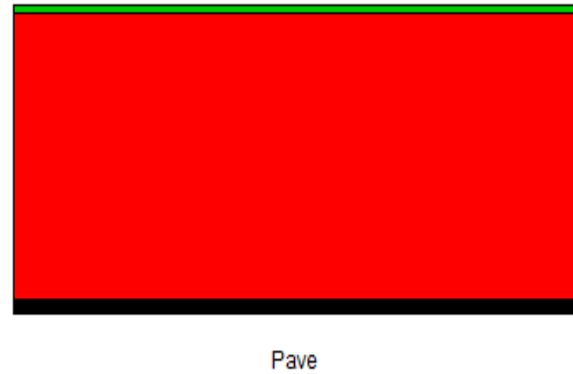
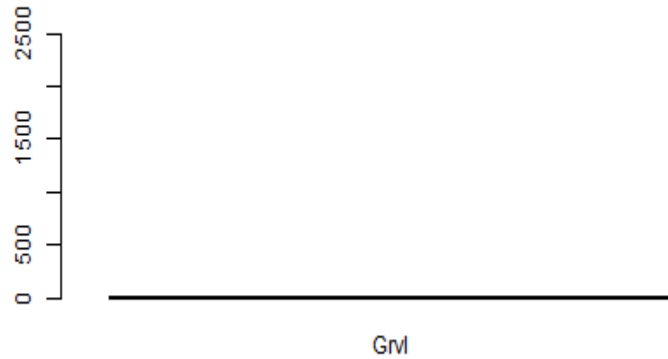
```
##
##               Grvl Pave
##   Gravel           0  120
##   No_Alley_Access  12 2720
##   Paved            0   78
```

\normalsize

Conditional barplot

\small

```
barplot(tab2dim,col=1:3)
```



\normalsize

Conditional barplot beside

\small

```
barplot(tab2dim,col=1:3,beside=T)
```

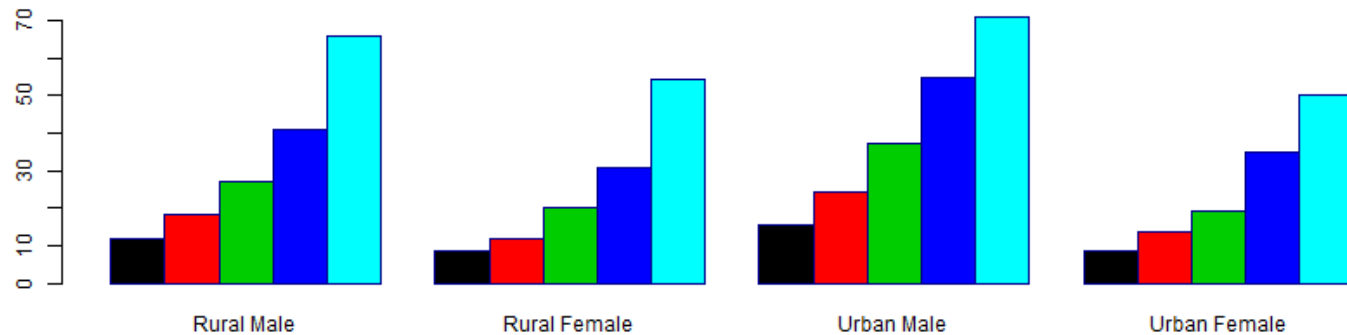


\normalsize

Exercise: simple graphics

- Load the dataset `VADeaths` and create the following plot:

`\small`



`\normalsize`

Horizontal `boxplot`

- A simple **boxplot** can be created with `boxplot()`
- For the command `boxplot()` at least one observation vector must be passed

\small

```
?boxplot
```

\normalsize

\small

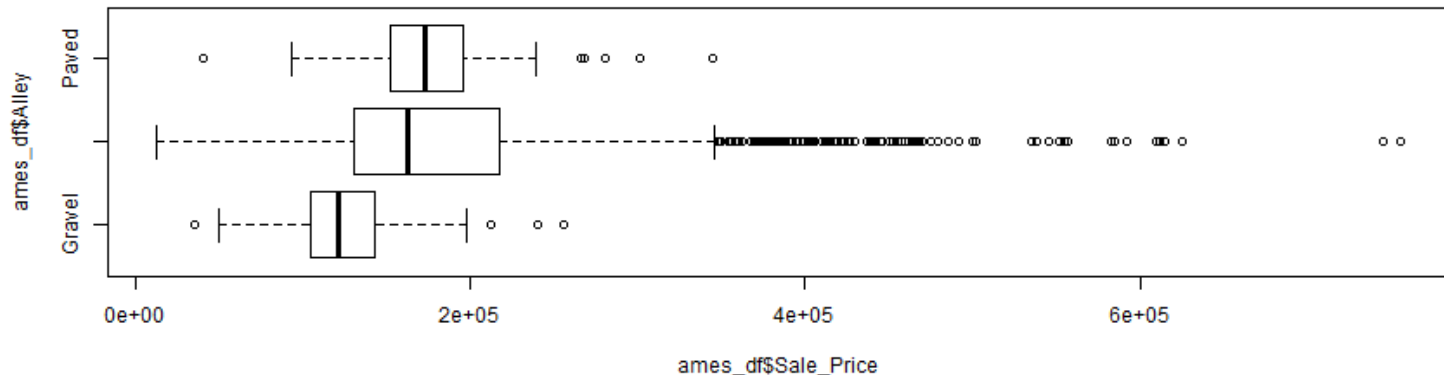
```
boxplot(ames_df$Sale_Price, horizontal=TRUE)
```

Grouped boxplots

- A very simple way to get a first impression of conditional distributions is via so-called grouped notched boxplots
- To do this, a so-called formula object must be passed to the `boxplot()` function.
- The conditional variable is located on the right side of a tilde

\small

```
boxplot(ames_df$Sale_Price~ames_df$Alley,horizontal=TRUE)
```



Boxplot alternatives - `vioplot`

- Builds on Boxplot - additional information about data density
- Density is calculated using the kernel method.
- The further the expansion, the higher the density at this point.
- White dot - median

\small

\normalsize

\small

```
library(vioplot)  
vioplot(na.omit(ames_df$Sale_Price),col="royalblue")
```

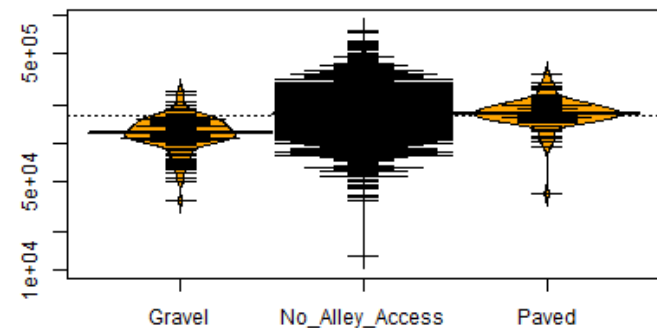
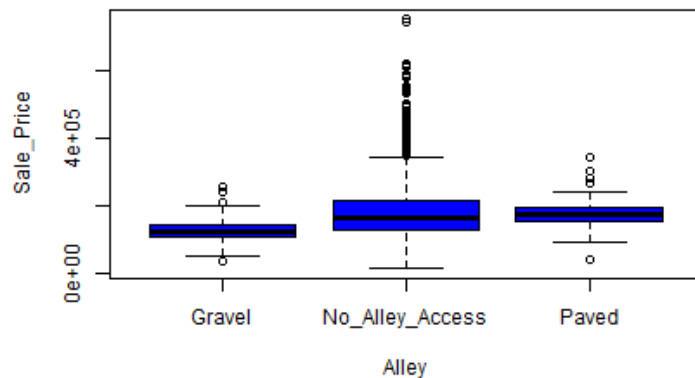
Alternatives `boxplot()`

\small

\normalsize

\small

```
library(beanplot)
par(mfrow = c(1,2))
boxplot(Sale_Price~Alley,data=ames_df,col="blue")
beanplot(Sale_Price~Alley,data=ames_df,col="orange")
```



Conditional, bi- and multivariate distribution graphics - scatterplots

- A simple two-way scatterplot can be created with the `plot()` function
- To create a scatterplot x and y observation vector must be passed
- Argument `col` to specify the color (color as character or numeric)
- Argument `pch` to specify plot symbols (plotting character) (character or numeric)
- The labels are defined with `xlab` and `ylab`.

\small

\normalsize

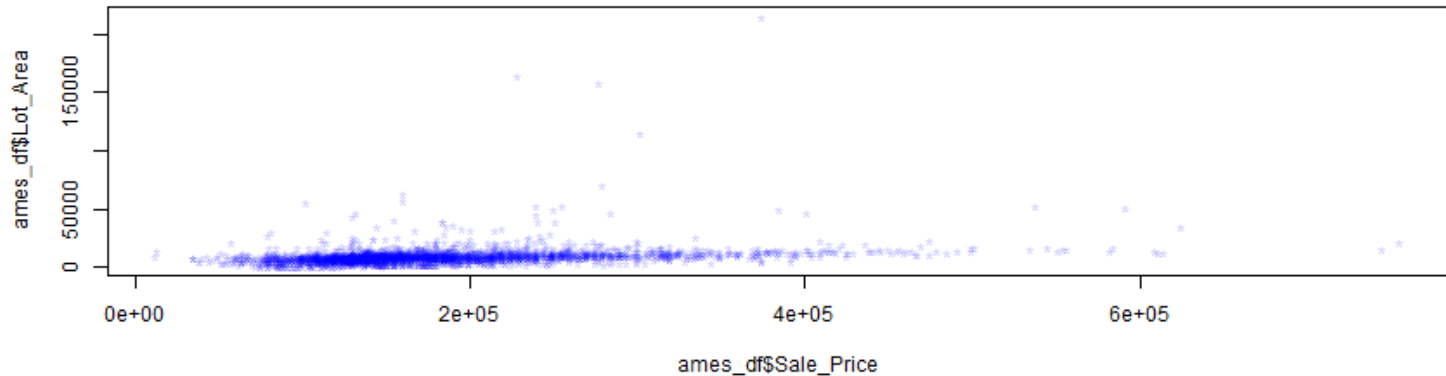
\small

```
plot(ames_df$Sale_Price,ames_df$Lot_Area)
```

Changing the plotting character

\small

```
plot(ames_df$Sale_Price, ames_df$Lot_Area,  
     pch="*", # changing the plotting symbol  
     col=rgb(0,0,1,.2)) # make the points transparent
```

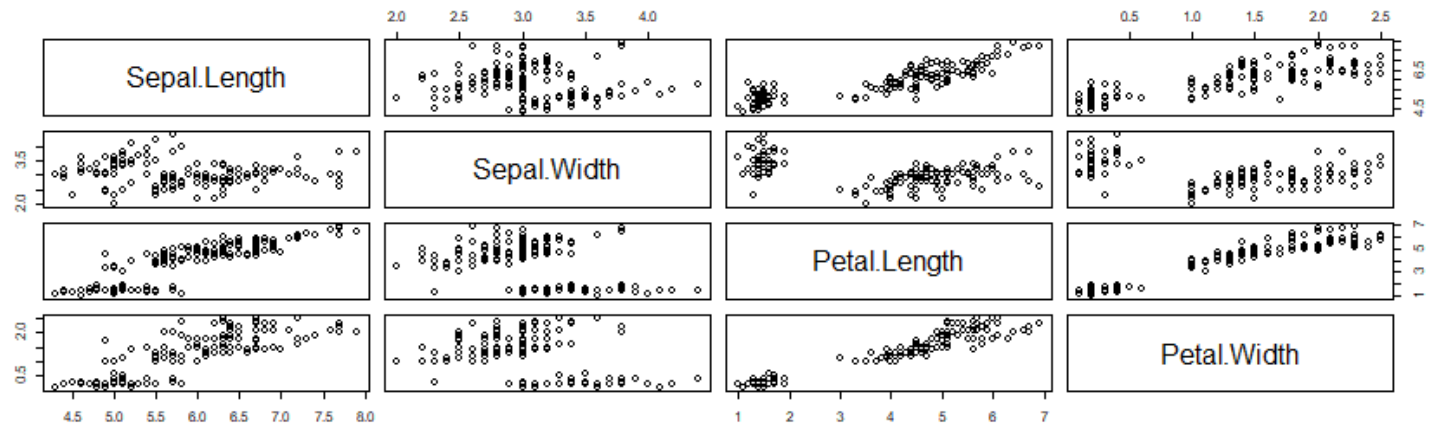


\normalsize

Relationship between variables - pairs plot

\small

```
pairs(iris[,1:4])
```



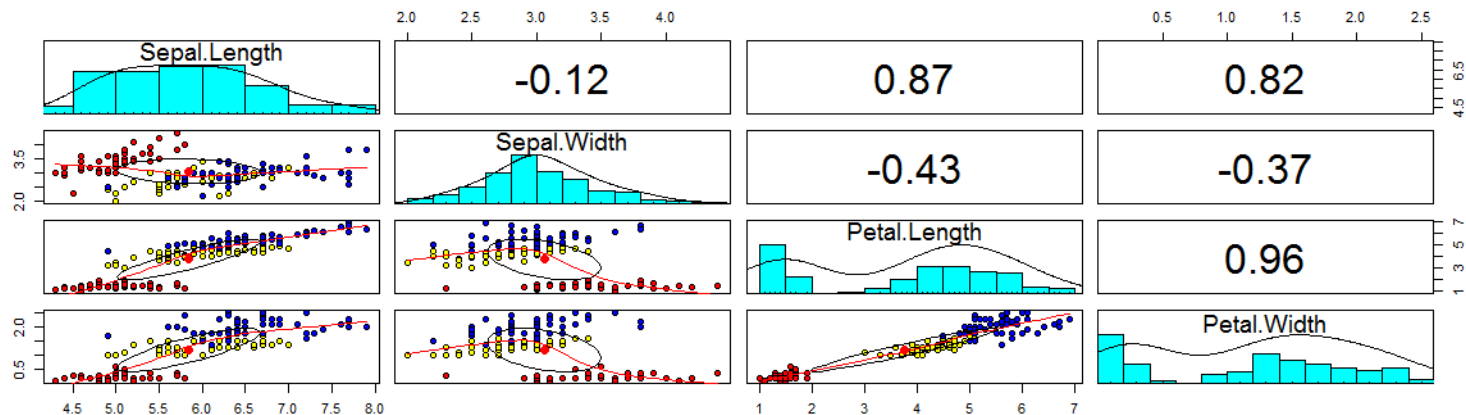
\normalsize

Enhanced multivariate plots

\small

```
library("psych")
pairs.panels(iris[,1:4],
             bg=c("red","yellow","blue")[iris$Species],
             pch=21,main="")
```

\normalsize



Tabulating

\small

```
(tab2 <- table(ames_df$Sale_Condition, ames_df$Alley))
```

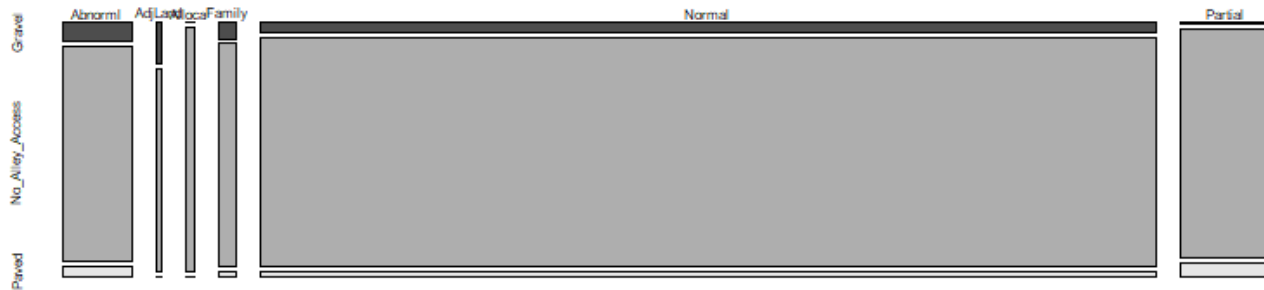
```
##
##           Gravel No_Alley_Access Paved
## Abnorml       14             167     9
## AdjLand        2              10     0
## Alloca         0              24     0
## Family         3              42     1
## Normal       100            2259    54
## Partial        1             230    14
```

\normalsize

Relationship - categorical variables

\small

```
mosaicplot(tab2, color = TRUE, main="")
```

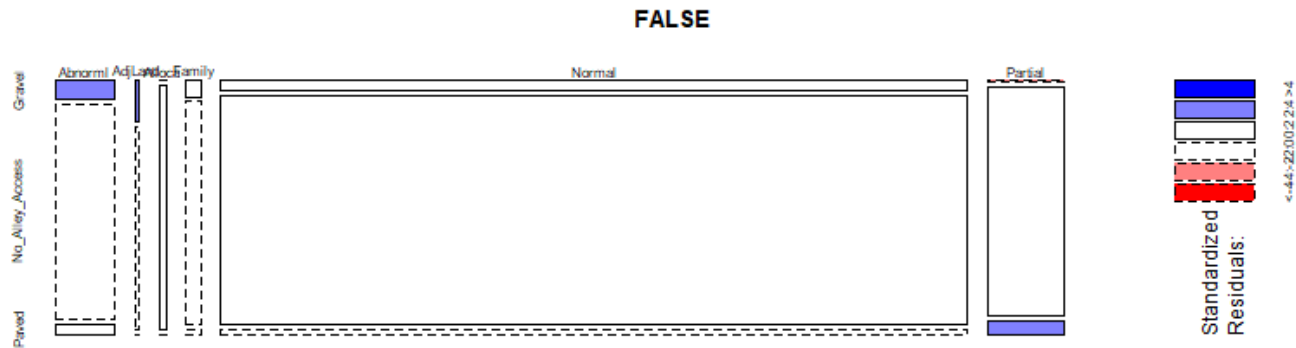


\normalsize

Surfaces are shaded according to the residuals:

\small

```
mosaicplot(tab2, main=F, shade = TRUE)
```

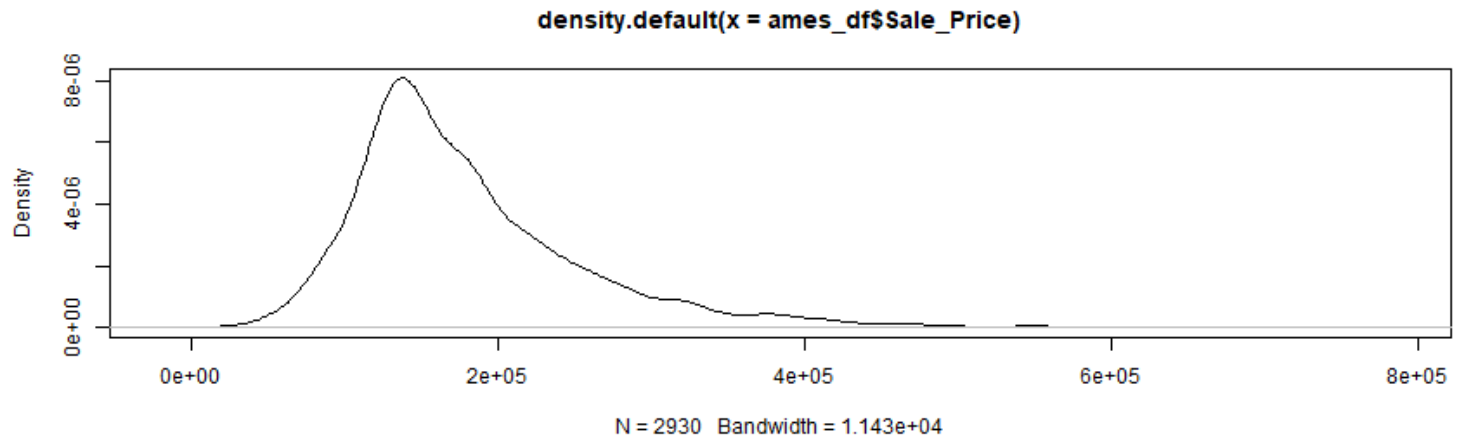


\normalsize

Density plot

\small

```
plot(density(ames_df$Sale_Price))
```



\normalsize

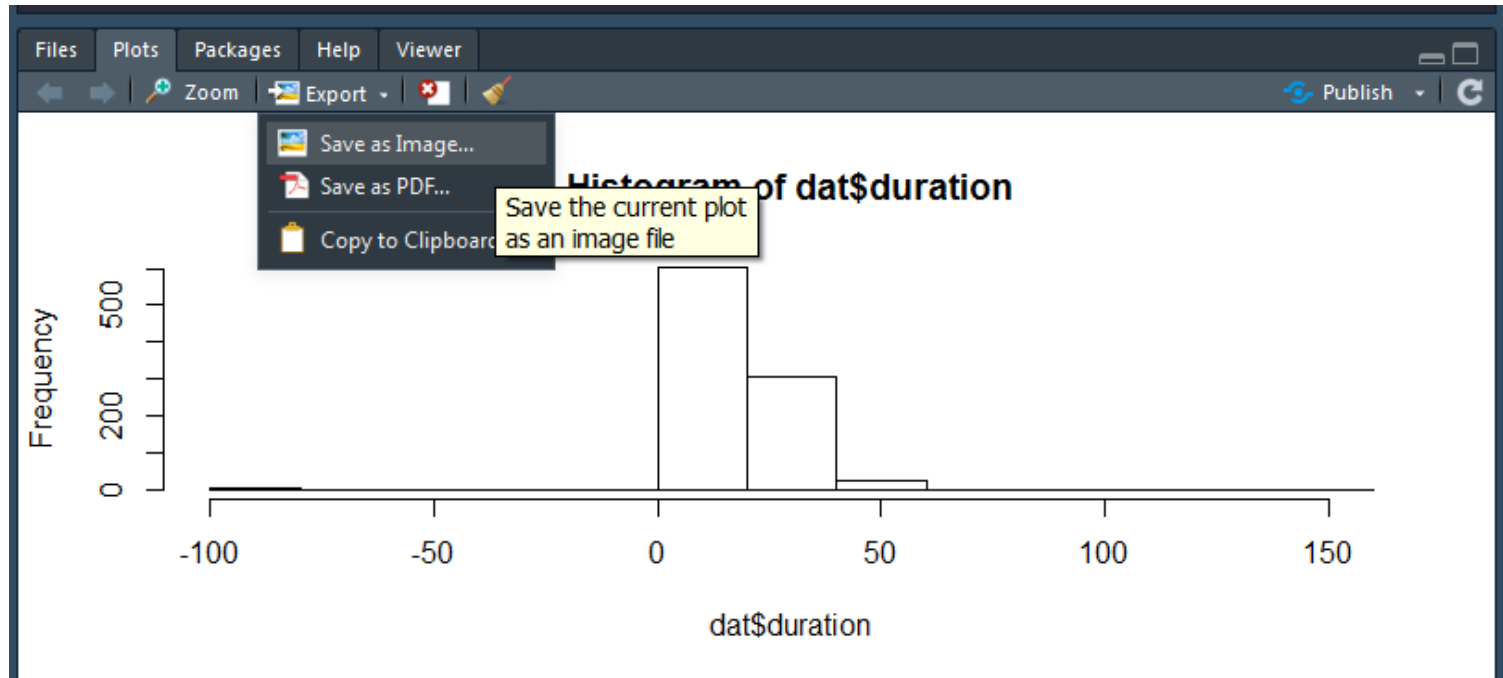
Stem and leaf plot

\small

```
data(airquality)
stem(airquality$Ozone)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    0 | 1467778999
##    1 | 01112233334444666688889
##    2 | 00001111233333334478889
##    3 | 001222455667799
##    4 | 01444556789
##    5 | 0299
##    6 | 134456
##    7 | 13367889
##    8 | 024559
##    9 | 1677
##   10 | 8
##   11 | 058
##   12 | 2
##   13 | 5
```

Export with Rstudio



Command to save graphic

- Alternatively also with the commands `png`, `pdf` or `jpeg` for example

\small

```
png("Histogramm.png")  
  hist(dat$duration)  
dev.off()
```

\normalsize

\small

```
pdf("Histogramm.pdf")  
  hist(dat$duration)  
dev.off()
```

\normalsize

\small

```
jpeg("Histogramm.jpeg")  
  hist(dat$duration)
```

Exercise: Advanced Base Graphics

Create Scatterplot with cars dataset

- a) Load the cars dataset and create a scatterplot of the data.
- b) Use the argument `las` to create a new scatterplot where more tickmarks are visible on the x and y axis.

Adjust Scatterplot

The previous plot didn't show all numbers associated to the new tickmarks. Use the argument `cex.axis` to control the size of the numbers. Make them smaller.

Exercise: Change orientation

Use the argument `las` to change the orientation of the labels from vertical to horizontal.

Exercise: Adding things to scatterplot

The `points` function

Suppose you want to add two new observations to the previous plot, and you want to be able to differentiate them. Use the `points` function to add new observations using red to identify them. The values of the new observation are speed = 23, 26 and dist = 60, 61.

Adjust x-axis range

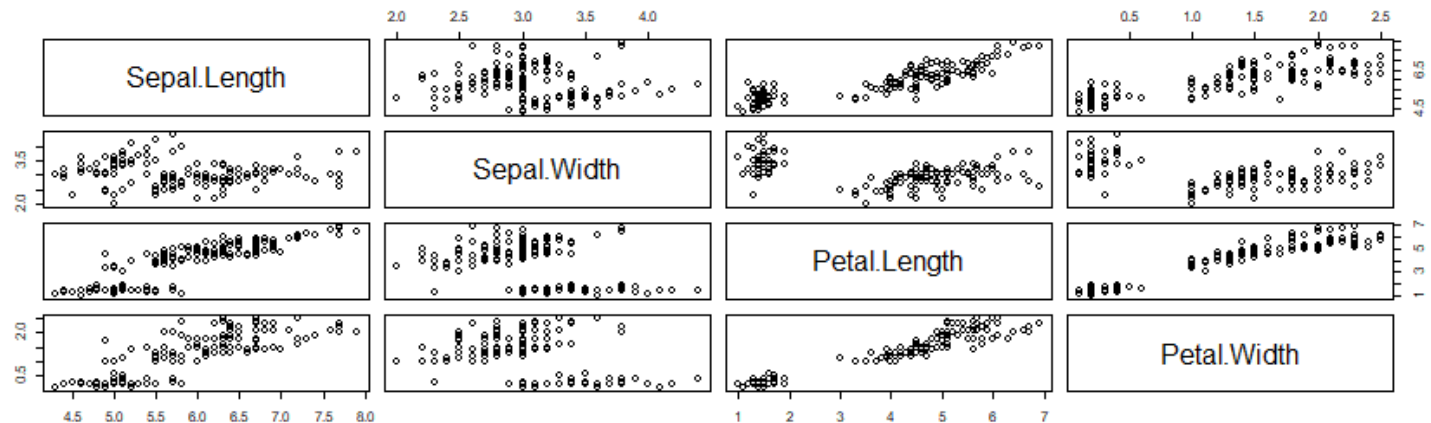
As you could see the previous plot doesn't show one of the new observations because is out the x-axis range.

- a) Create again the plot for the old observations with an x-axis range that includes all the values from 4 to 26.
- b) Add the two new observations using the `points` function.

Relationship between variables - pairs plot

\small

```
pairs(iris[,1:4])
```



\normalsize

Enhanced multivariate plots

\small

```
install.packages("psych")
```

\normalsize

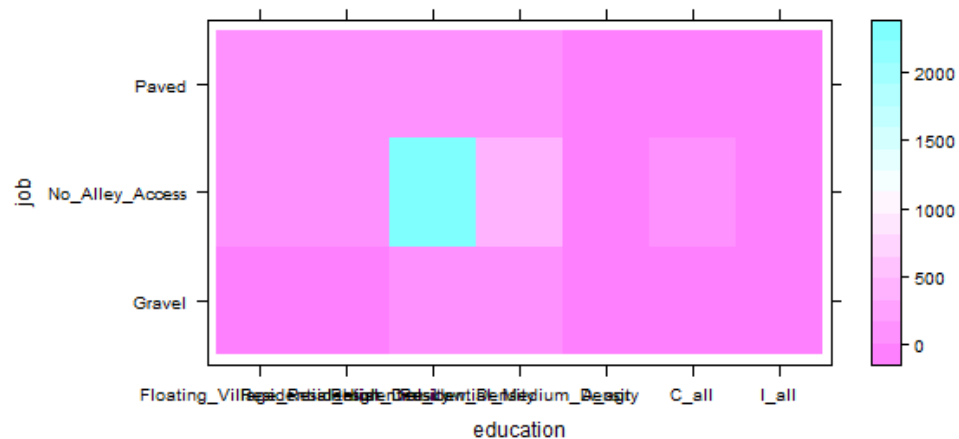
\small

```
library("psych")  
bgcol <- c("red","yellow","blue")[iris$Species]  
pairs.panels(iris[,1:4],bg=bgcol,pch=21,main="")
```

levelplot

\small

```
library("lattice")
levelplot(table(ames_df$MS_Zoning,ames_df$Alley),
           xlab="education",ylab="job")
```



\normalsize

Color palettes

\small

```
library(RColorBrewer)  
display.brewer.all()
```

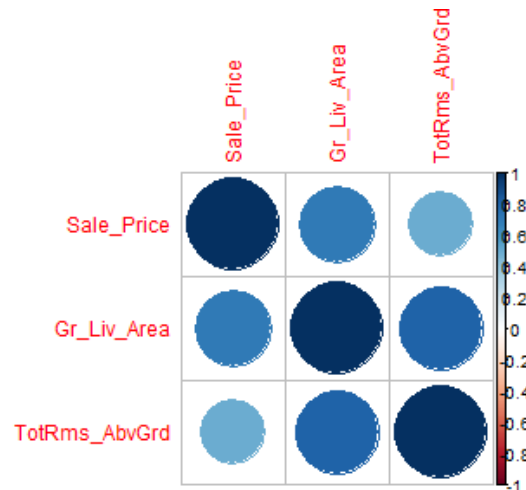
A correlation plot

\small

\normalsize

\small

```
library(corrplot)
corrplot(cor(ames_df[,c("Sale_Price",
                        "Gr_Liv_Area", "TotRms_AbvGrd")]))
```

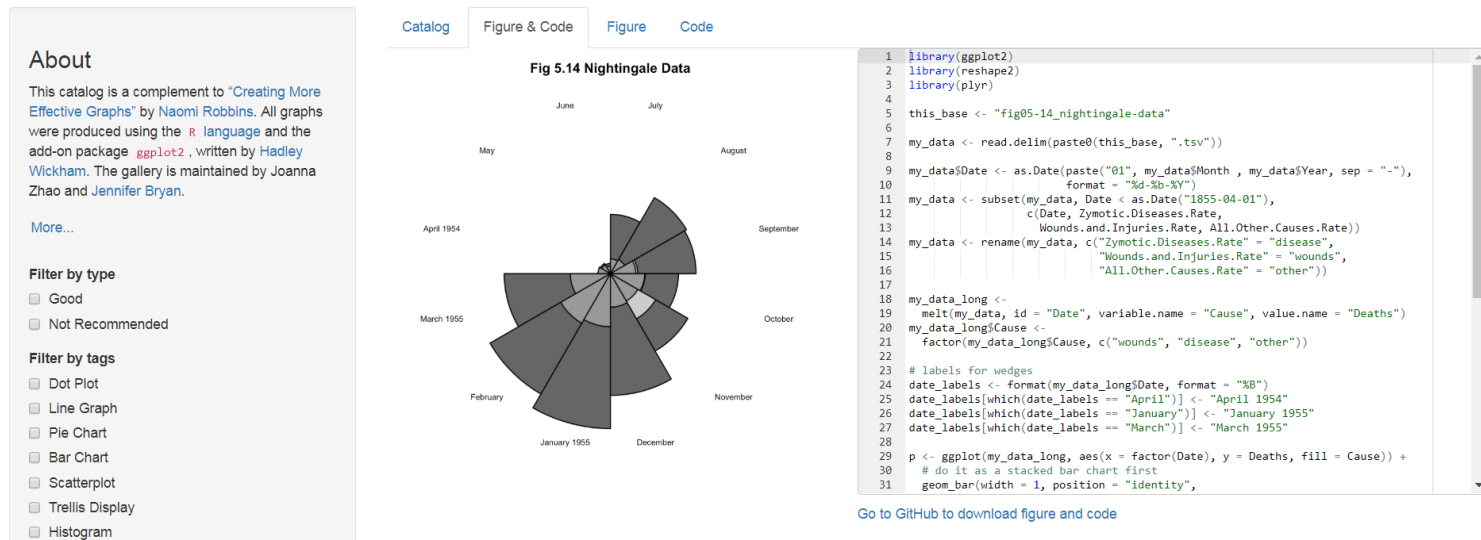


\normalsize

Shiny App - R graphs catalogue

R graphs catalogue

R Graph Catalog



<http://shinyapps.stat.ubc.ca/r-graph-catalog/>

Not all plots are the same

- The base package already includes a large number of plot functions
- Other packages like `lattice`, `ggplot2`, etc extend this functionality

Manuals that go far beyond this introduction:

- Murrell, P (2006): R Graphics.
- R Development Core Group **Graphics with R**
- Wiki on **R Programming/Graphics**
- Martin Meermeyer **Creating Reproducible Publication Quality Graphics with R: A Tutorial**
- Institute For Quantitative Social Science at Harvard - **R graphics tutorial**

Links

- **R graph gallery**
- **Bioconductor R manual** with an extensive part on graphics
- **Shiny app** for **interactive plot editing**
- **Producing Simple Graphs with R**