CWRU DSCI351-451: Exploratory Multi-variate Pair-wise Correlation Plots

Roger H. French, JiQi Liu

18 October, 2018

Contents

8.2.2.1 Reading, Homeworks, Projects, SemProjects . . . 1
8.2.2.2 Textbooks . . . 1
8.2.2.3 Syllabus . . . 1
8.2.2.4 Pair Coding . . . 3
8.2.2.5 Everything is a variable . . . 3
8.2.2.6 Scatter Plot Matrices . . . 3
8.2.2.6.1 Using the iris dataset in R as an example . . . 3
8.2.2.6.2 so lets use a pairwise linear correlation plot . . . 3
8.2.2.6.3 Lets make a better pairs plot . . . 4
8.2.2.6.4 even easier to do with ggplot and GGally . . . 5
8.2.2.7 Lets take another run at scatterplots . . . 6
8.2.2.7.1 Simple Scatterplots . . . 6
8.2.2.7.2 Now onto pairs plots, i.e. scatterplot matrices . . . 9
8.2.2.7.3 High Density scatterplots with Binning . . . 12
8.2.2.7.4 There is a 3D scatterplot package . . . 15
8.2.2.7.5 Correlograms . . . 18
8.2.2.7.6 Psych package is very popular in our group . . . 20
8.2.2.8 Citations . . . 21

8.2.2.1 Reading, Homeworks, Projects, SemProjects

- Readings:
  - OIS5, Sections 1-4 "Inference for Numerical Data"
- Homeworks
  - HW5 Numerical Inference
- Data Science Projects:
  - Project 2
- 451 SemProjects:
  - 2nd Report Outs, Tuesday October 30th
- Friday Comm. Hour
  -

8.2.2.2 Textbooks

- Peng: R Programming for Data Science
- Peng: Exploratory Data Analysis with R
- Open Intro Stats, v3
- Wickham: R for Data Science
- Hastie: Intro to Statistical Learning with R

8.2.2.3 Syllabus

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1,**
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

Figure 1: DSCI351-451 Syllabus

8.2.2.4 Pair Coding

- Reading posted in the Class Repo
- What is Code Review
- 11 Best Practices for Peer Code Review

8.2.2.5 Everything is a variable

And in EDA

- Finding relationships among variables
  - Starting with scatterplots
- And continuing with linear correlations
  - Is a good way to go

Pairs plots are a fast way to EDA for relationships

- These may be expected, or unexpected
- They don't necesarily mean causality

8.2.2.6 Scatter Plot Matrices

8.2.2.6.1 Using the iris dataset in R as an example

Lets load the iris dataset, check out its background

And then look at correlation coefficients among variables: numerically

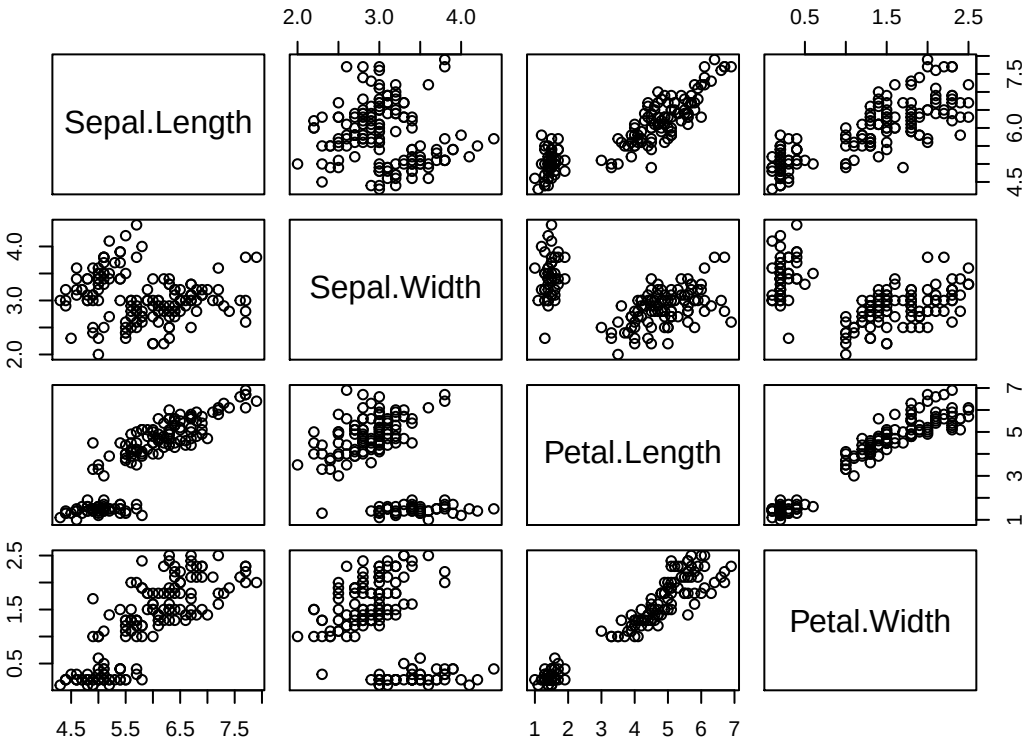
```
data(iris)
?iris
cor(iris[,1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Tabular data doesn't communicate to us very well

8.2.2.6.2 so lets use a pairwise linear correlation plot

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



This is a nice example of un-biased analytics

- We can visually see if relationships are present
- but not necesarily what their origin or nature is

The upper right and lower left quadrants are identical

- the diagonal is the variable names
- I find it best to read the lower left quadrant

8.2.2.6.3 Lets make a better pairs plot

With the corrrelation coefficients and p values

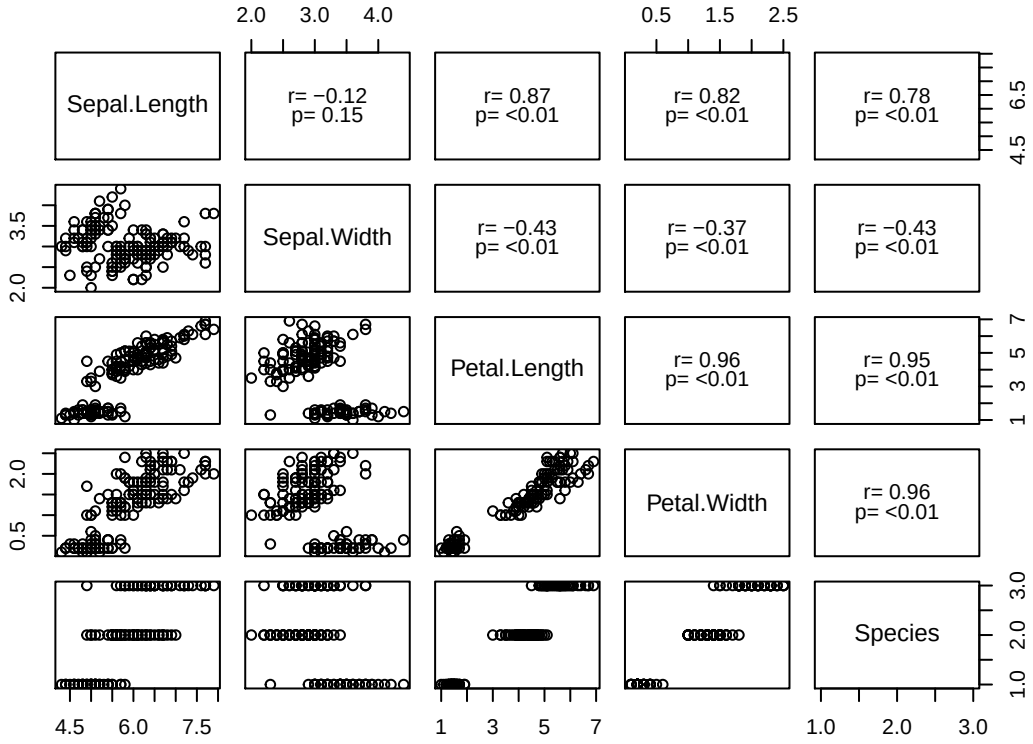
- make r = correlation coefficients
- make p = p values for the correlation test
- and lets make this into a function we can use later also.

```
panel.cor <- function(x, y, digits = 2, cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  # correlation coefficient
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste("r= ", txt, sep = "")
  text(0.5, 0.6, txt)

  # p-value calculation
  p <- cor.test(x, y)$p.value
  txt2 <- format(c(p, 0.123456789), digits = digits)[1]
  txt2 <- paste("p= ", txt2, sep = "")
  if (p < 0.01) txt2 <- paste("p= ", "<0.01", sep = "")
```

```
  text(0.5, 0.4, txt2)
}

pairs(iris, upper.panel = panel.cor)
```

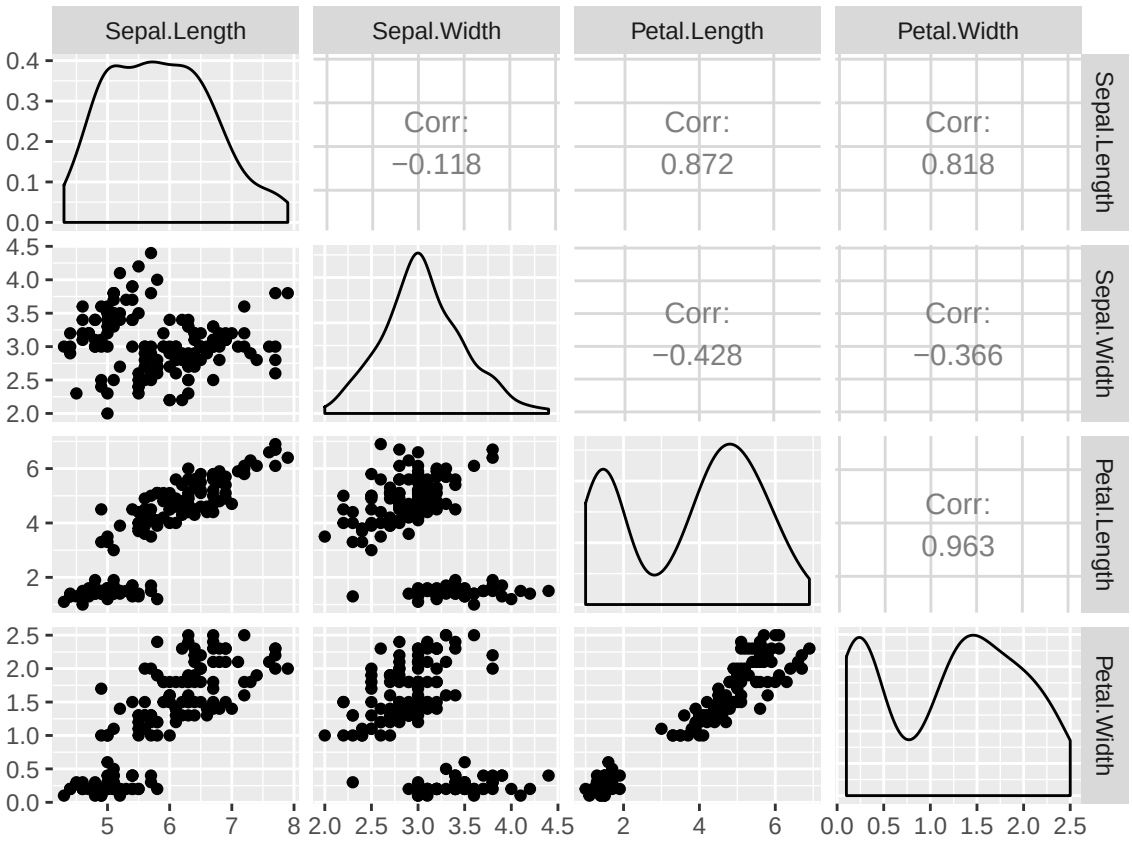


8.2.2.6.4 even easier to do with ggplot and GGally

```
library(GGally)
```

```
## Loading required package: ggplot2
```

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



Some tuning of the x-axis labels required!

8.2.2.7 Lets take another run at scatterplots

Lets use the mtcars dataset in R

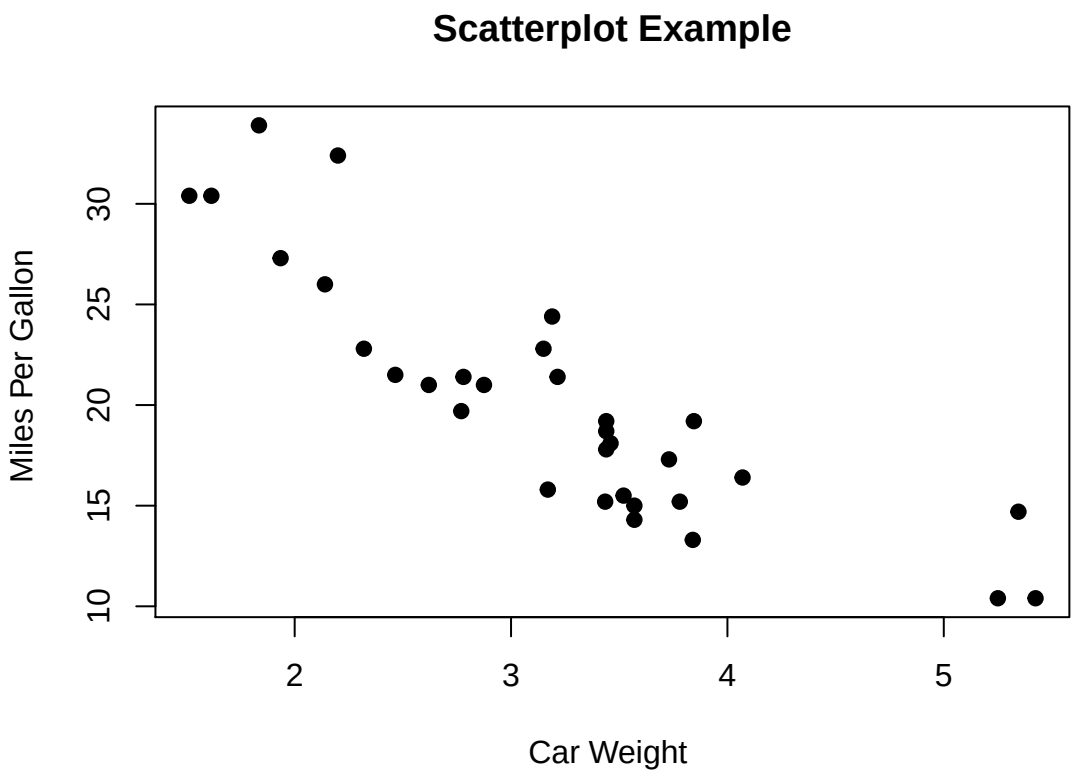
- Motor Trend Car Road Tests for 32, 1973-4 models

8.2.2.7.1 Simple Scatterplots

```
attach(mtcars)
```

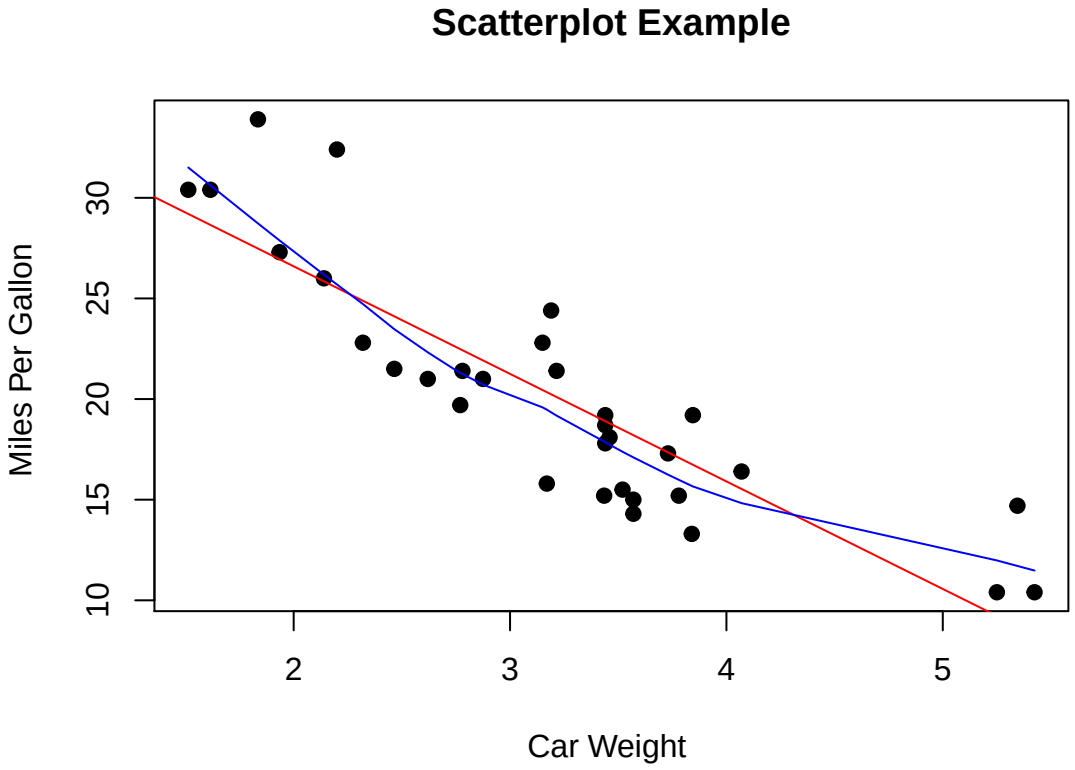
```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
?mtcars
plot(wt, mpg, main = "Scatterplot Example",
     xlab = "Car Weight ", ylab = "Miles Per Gallon ", pch = 19)
```



Add fit lines

```
# Add fit lines
plot(wt, mpg, main = "Scatterplot Example",
    xlab = "Car Weight ", ylab = "Miles Per Gallon ", pch = 19)
abline(lm(mpg~wt), col = "red") # regression line (y~x)
lines(lowess(wt,mpg), col = "blue") # lowess line (x,y)
```



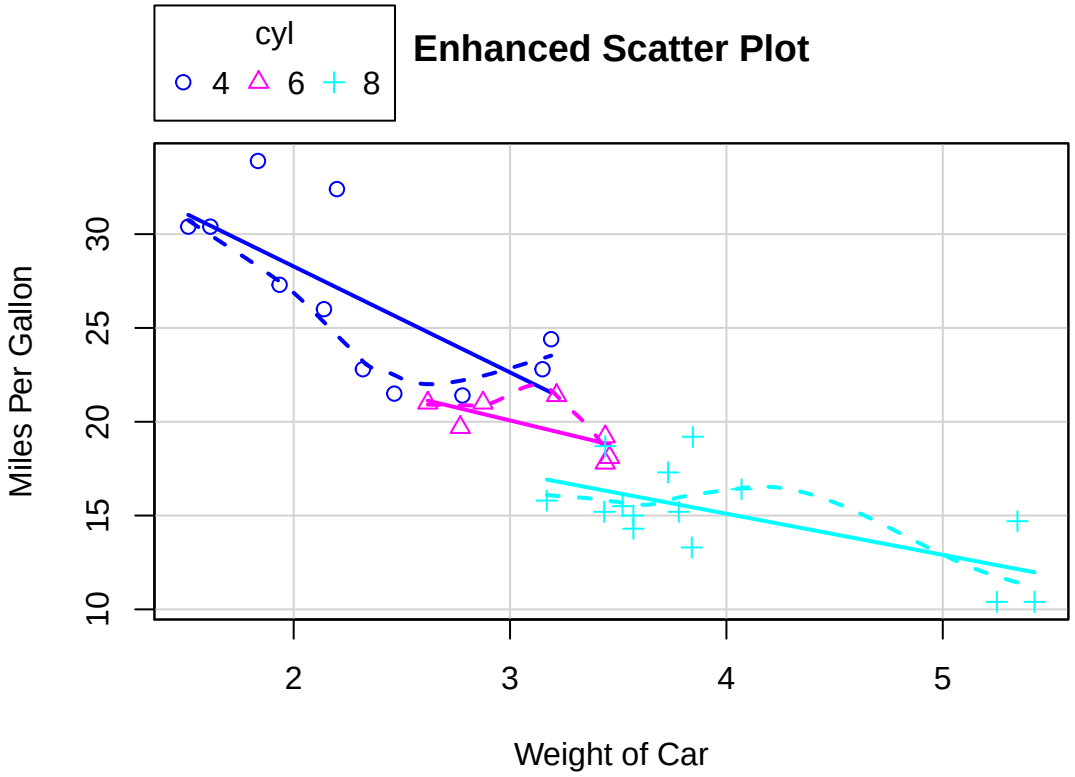
Try scatterplot function in the car package

```
# Enhanced Scatterplot of MPG vs. Weight
# by Number of Car Cylinders
library(car)
```

```
## Loading required package: carData
```

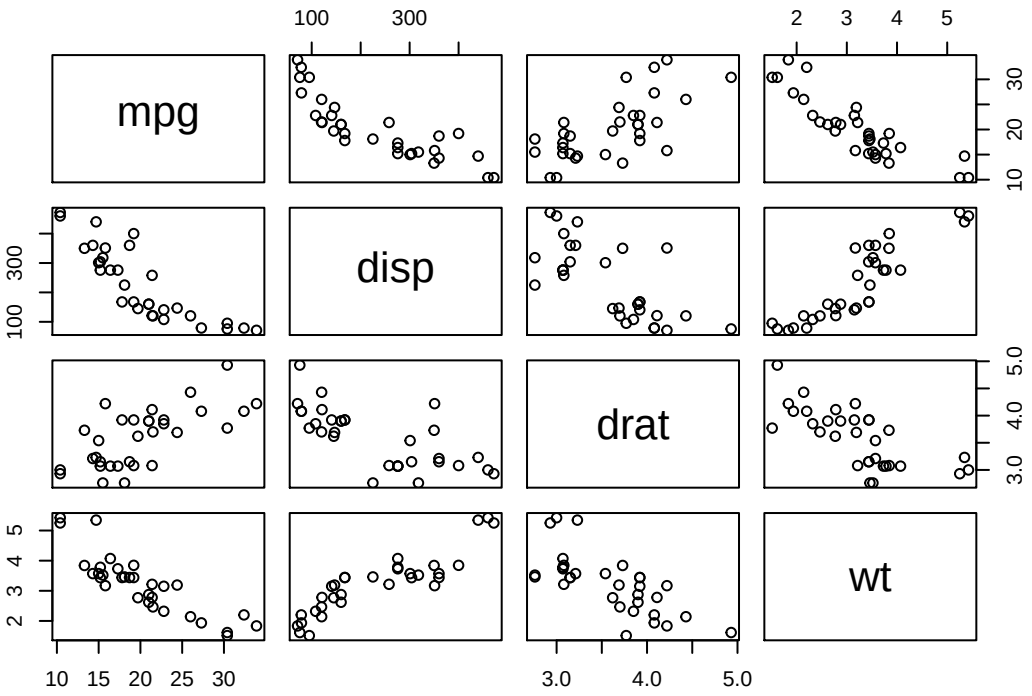
```
??car
scatterplot(mpg ~ wt | cyl, data = mtcars, xlab = "Weight of Car",
            ylab = "Miles Per Gallon", main = "Enhanced Scatter Plot",
            legend = row.names(mtcars))
```

```
## Warning in applyDefaults(legend, defaults = list(), type = "legend"):
## unnamed legend arguments, will be ignored
```



8.2.2.7.2 Now onto pairs plots, i.e. scatterplot matrices

```
# Basic Scatterplot Matrix
pairs(~mpg+disp+drat+wt,data = mtcars,
      main = "Simple Scatterplot Matrix")
```



and using the car package

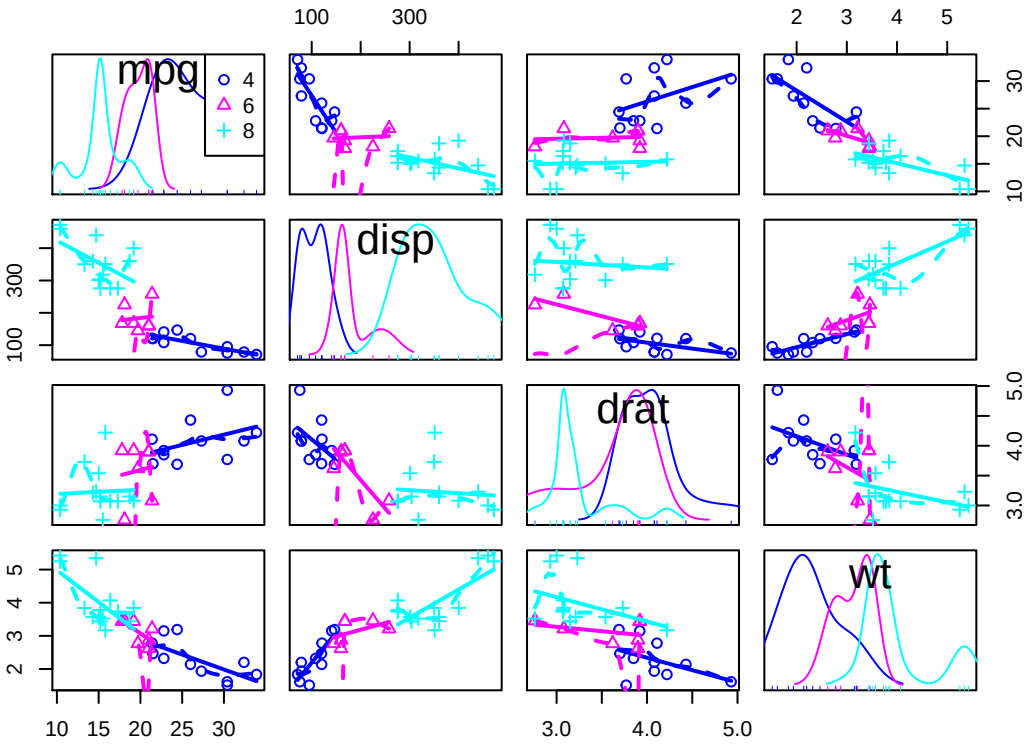
```
# Scatterplot Matrices from the car Package
library(car)
attach(mtcars)
```

```
## The following objects are masked from mtcars (pos = 5):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
scatterplotMatrix(~ mpg + disp + drat + wt | cyl, data = mtcars,
                  legend = "Three Cylinder Options")
```

```
## Warning in applyDefaults(legend, defaults = list(coords = NULL), type =
## "legend"): unnamed legend arguments, will be ignored
```



and the gclus package

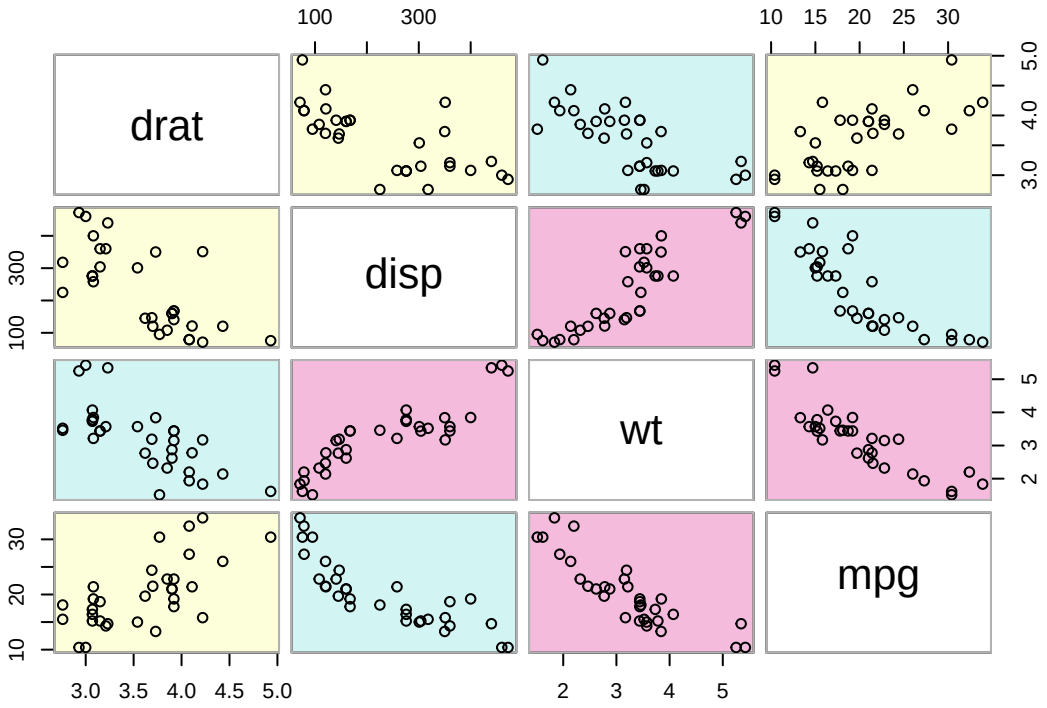
The gclus package gives pairs plots

- colored by magnitude of the correlation coefficient
- very useful "signatures"

```
# Scatterplot Matrices from the gclus Package
library(gclus)
```

```
## Loading required package: cluster
```

```
??gclus
dta <- mtcars[c(1,3,5,6)] # get data
dta.r <- abs(cor(dta)) # get correlations
dta.col <- dmat.color(dta.r) # get colors
# reorder variables so those with highest correlation
# are closest to the diagonal
dta.o <- order.single(dta.r)
cpairs(dta, dta.o, panel.colors = dta.col, gap = 0.5,
       main = "Variables Ordered and Colored by Correlation" )
```



8.2.2.7.3 High Density scatterplots with Binning

using hexbin package

```
# High Density Scatterplot with Binning
library(hexbin)
x <- rnorm(1000)
y <- rnorm(1000)
bin<-hexbin(x, y, xbins=50)
plot(bin, main="Hexagonal Binning")
```

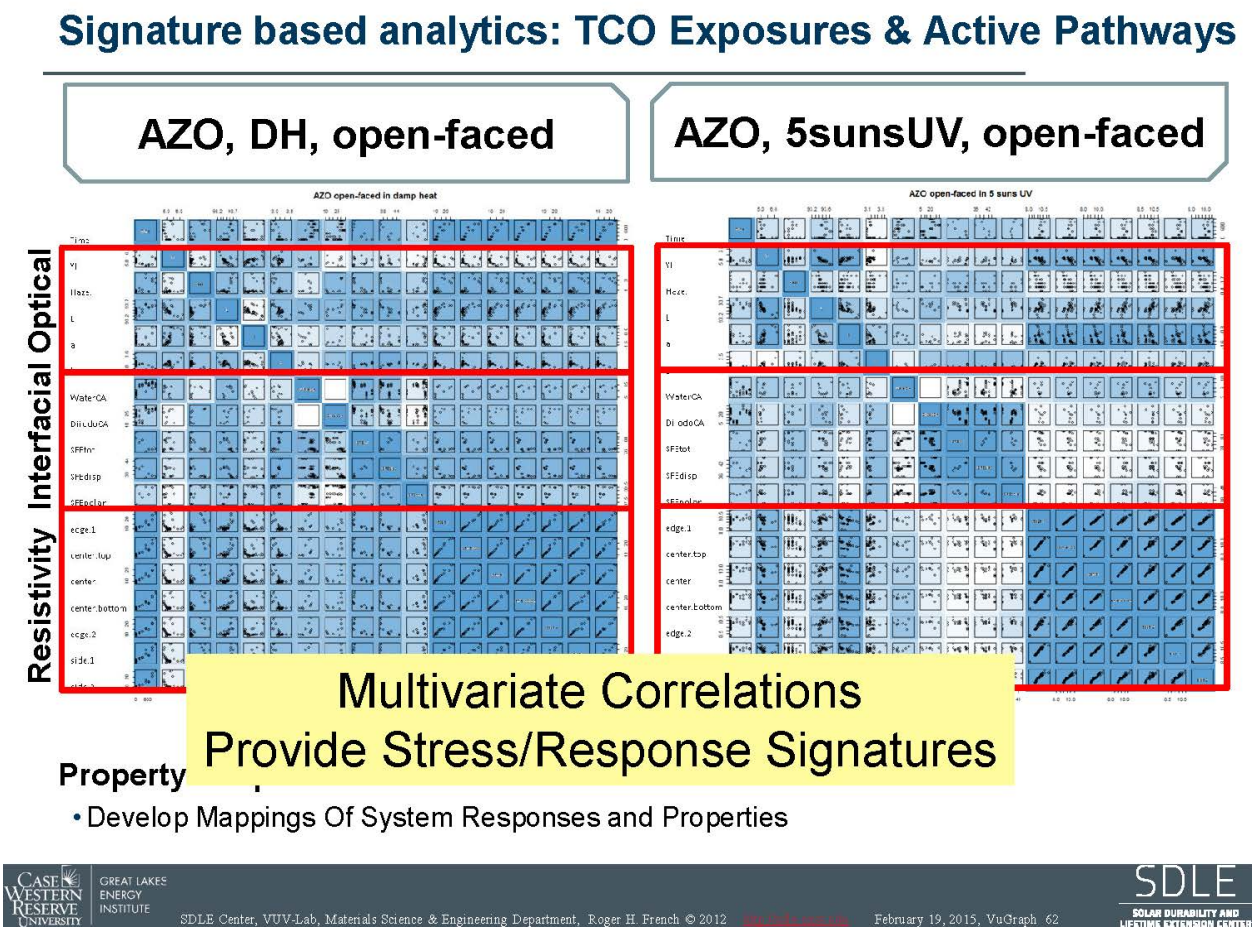
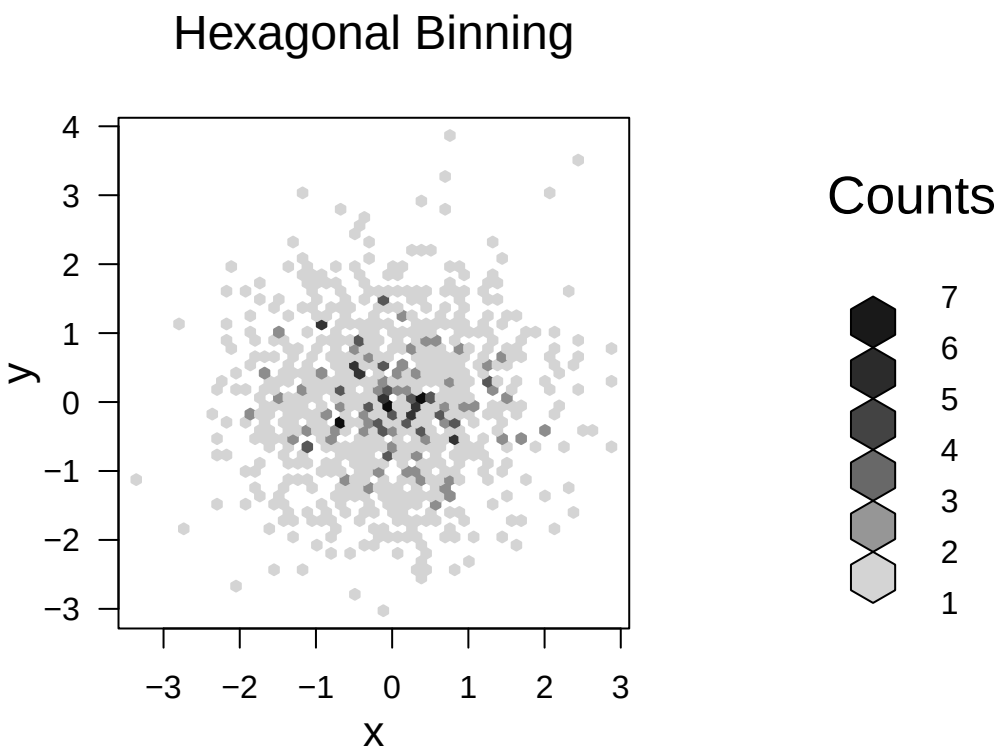
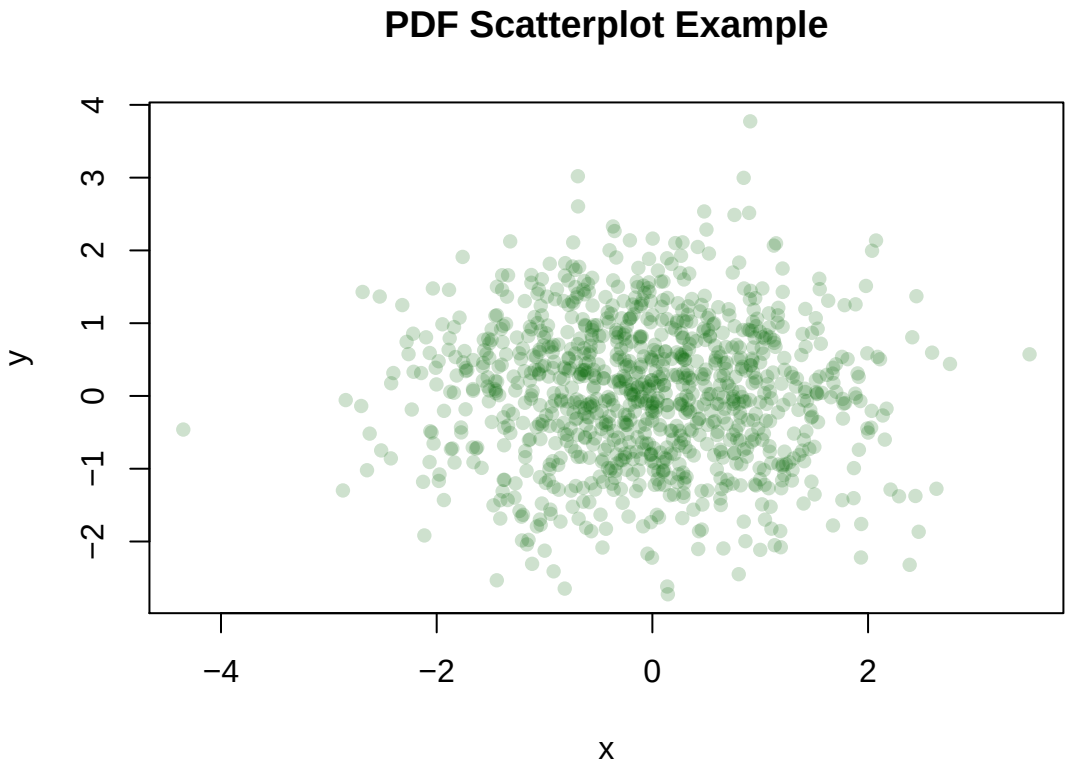


Figure 2: Degradation Signatures of Transparent Conductive Oxides



with sunflowerpot if the points overlap

```
# High Density Scatterplot with Color Transparency
x <- rnorm(1000)
y <- rnorm(1000)
plot(x,y, main = "PDF Scatterplot Example",
     col = rgb(0,100,0,50,maxColorValue = 255), pch = 16)
```



8.2.2.7.4 There is a 3D scatterplot package

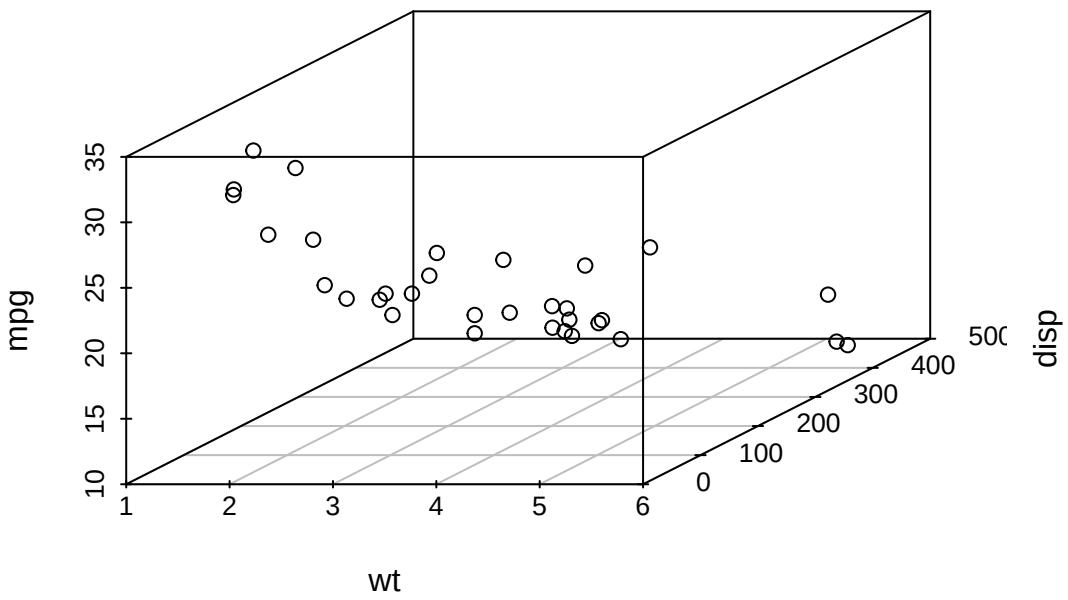
```
# 3D Scatterplot
library(scatterplot3d)
attach(mtcars)
```

```
## The following objects are masked from mtcars (pos = 7):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following objects are masked from mtcars (pos = 10):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
scatterplot3d(wt,disp,mpg, main = "3D Scatterplot")
```




```
# 3D Scatterplot with Coloring and Vertical Drop Lines
library(scatterplot3d)
attach(mtcars)
```

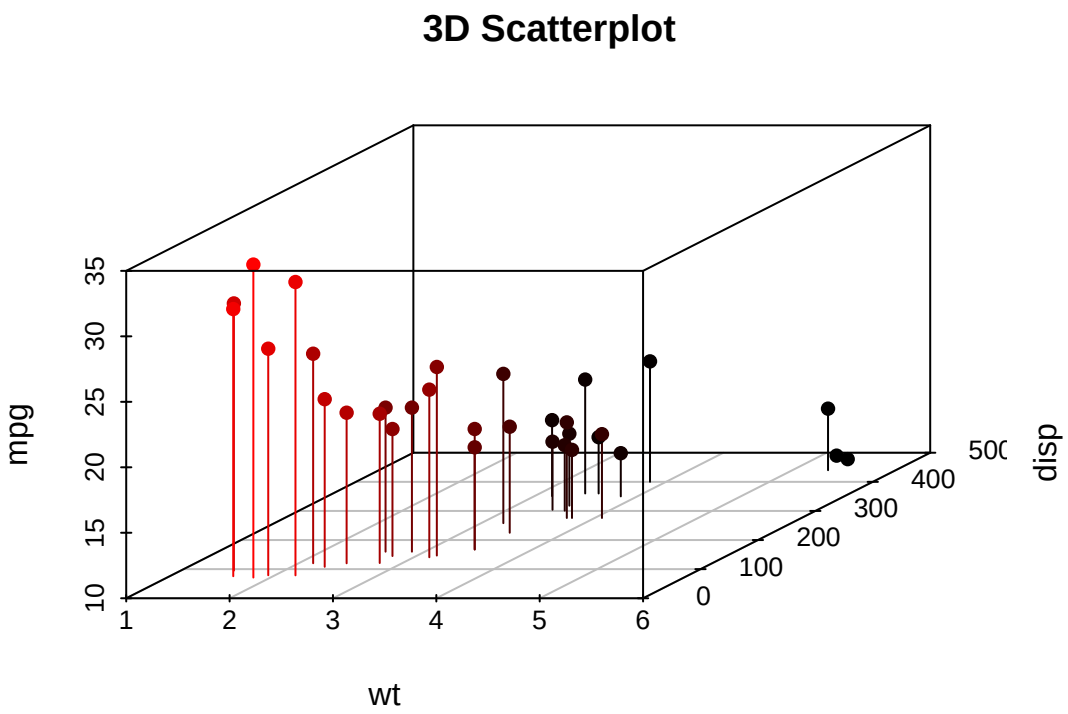
```
## The following objects are masked from mtcars (pos = 3):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following objects are masked from mtcars (pos = 8):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following objects are masked from mtcars (pos = 11):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
scatterplot3d(wt,disp,mpg, pch = 16, highlight.3d = TRUE,
              type = "h", main = "3D Scatterplot")
```




```
# 3D Scatterplot with Coloring and Vertical Lines
# and Regression Plane
library(scatterplot3d)
attach(mtcars)
```

```
## The following objects are masked from mtcars (pos = 3):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

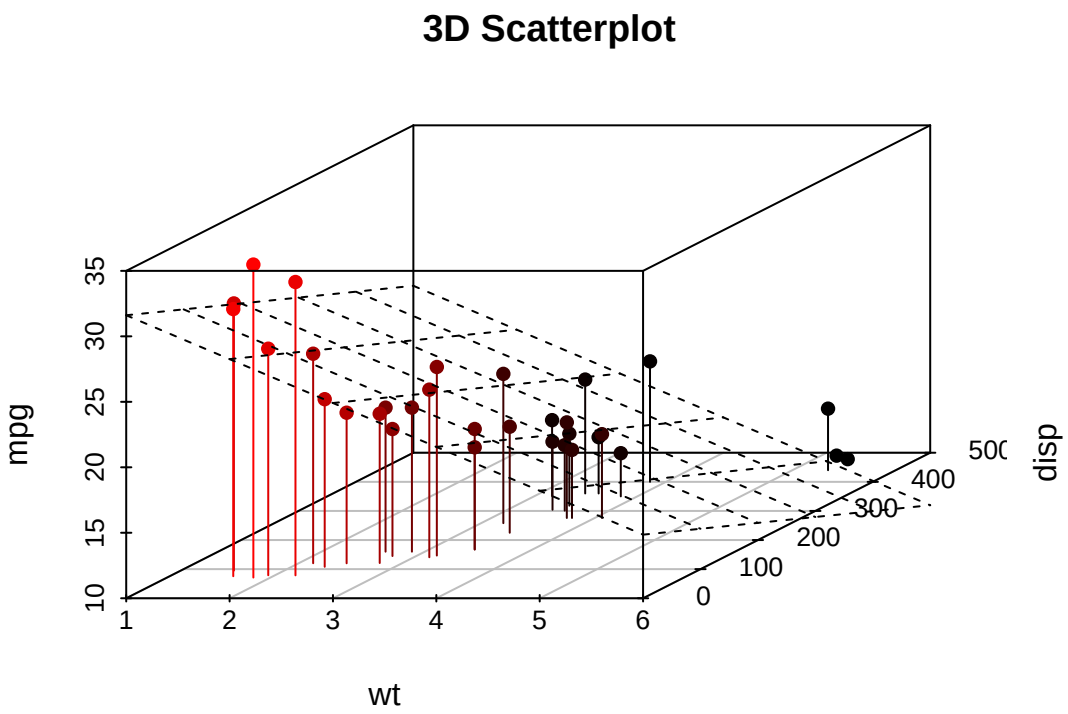
```
## The following objects are masked from mtcars (pos = 4):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following objects are masked from mtcars (pos = 9):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following objects are masked from mtcars (pos = 12):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
s3d <- scatterplot3d(wt, disp, mpg, pch = 16, highlight.3d = TRUE,
                     type = "h", main = "3D Scatterplot")
fit <- lm(mpg ~ wt+disp)
s3d$plane3d(fit)
```



3D spinning scatterplots using rgl or Rcmdr packages

```
# Spinning 3d Scatterplot
library(rgl)

plot3d(wt, disp, mpg, col = "red", size = 3)
```

8.2.2.7.5 Correlograms

Many statistical tools exist for analyzing their structure, but, surprisingly,

- there are few techniques for exploratory visual display,
  - and for depicting the patterns of relations among variables
  - in such matrices directly,
  - particularly when the number of variables is moderately large.

This describes a set of techniques we subsume under the name corrgram, based on two main schemes:

- (a) rendering the value of a correlation to depict its sign and magnitude.
  - We consider some of the properties of several iconic representations,
  - in relation to the kind of task to be performed.

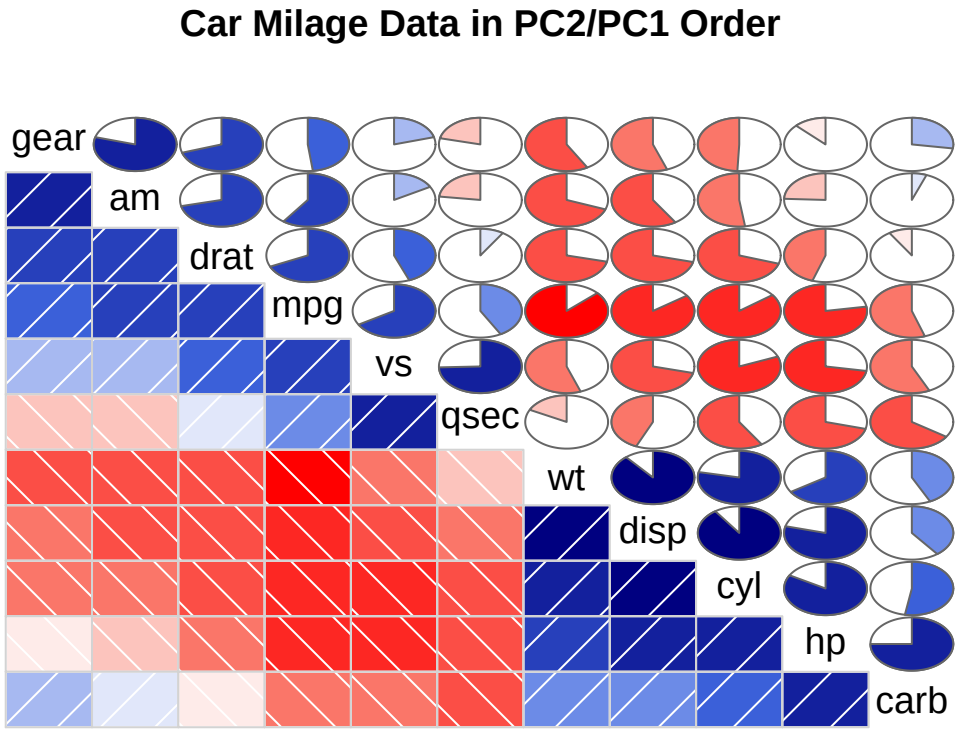
(b) re-ordering the variables in a correlation matrix - so that "similar" variables are positioned adjacently, - facilitating perception.

```
# First Correlogram Example
library(corrgram)
```

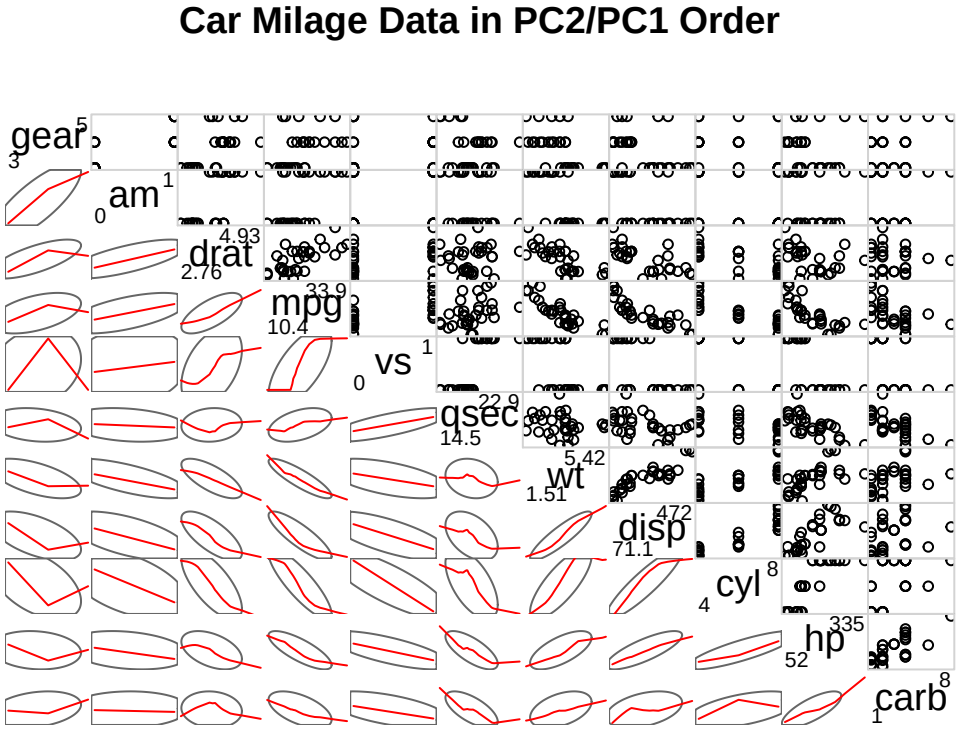
```
##
## Attaching package: 'corrgram'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     panel.cor
```

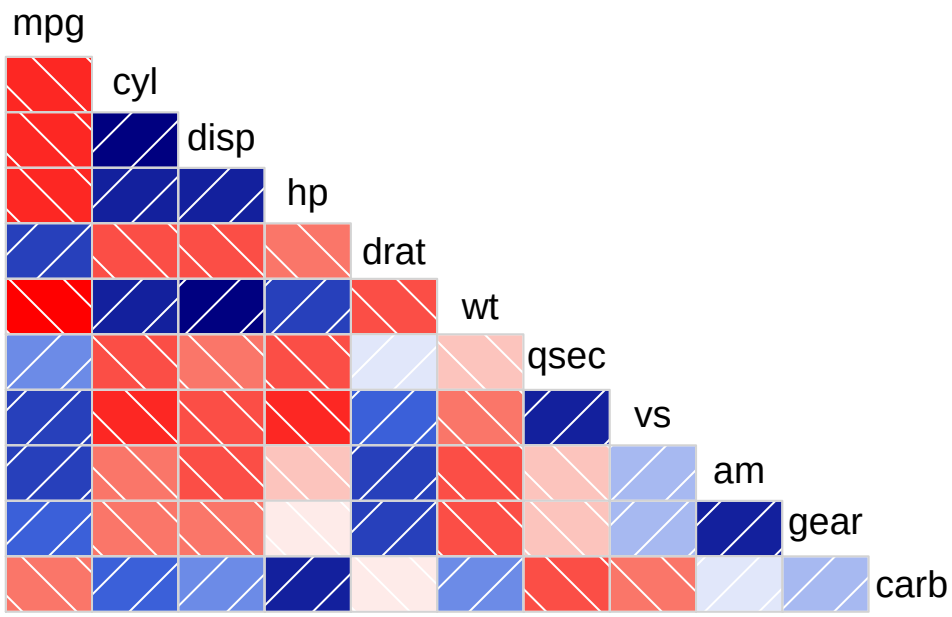
```
corrgram(mtcars, order = TRUE, lower.panel = panel.shade,
         upper.panel = panel.pie, text.panel = panel.txt,
         main = "Car Milage Data in PC2/PC1 Order")
```




```
# Second Correlogram Example
library(corrgram)
corrgram(mtcars, order = TRUE, lower.panel = panel.ellipse,
         upper.panel = panel.pts, text.panel = panel.txt,
         diag.panel = panel.minmax,
   main = "Car Milage Data in PC2/PC1 Order")
```




```
# Third Correlogram Example
library(corrgram)
corrgram(mtcars, order = NULL, lower.panel = panel.shade,
         upper.panel = NULL, text.panel = panel.txt,
         main = "Car Milage Data (unsorted)")
```



8.2.2.7.6 Psych package is very popular in our group

From Degradation Science COSSMS review

Fig. 9. (a) Schematic diagram of PV module and microinverter setup. (b) Comparison of actual microinverter temperature and fitted microinverter temperature for the microinverters connected to four different PV module brands during noon time on a typical cloudy day. (c) Pairs plot and correlation coefficient between different environmental and application stressors. Irradiance, wind speed and ambient temperature (Ambient.T) are the environmental stressors. PV module temperature (Module.T), PV module brand(Brand), AC power (Power) and microinverter temperature (Micro.T) are application stressors.

8.2.2.8 Citations

1. Scatter Plot Matrices in R
2. Simple Scatterplot
3. Correlograms + Michael Friendly Corrgrams: Exploratory displays for correlation matrices
4. Nice pairs plots with correlation coefficients in the upper quandrant + psych: Procedures for Psychological, Psychometric, and Personality Research + Using R and psych for personality and psychological research

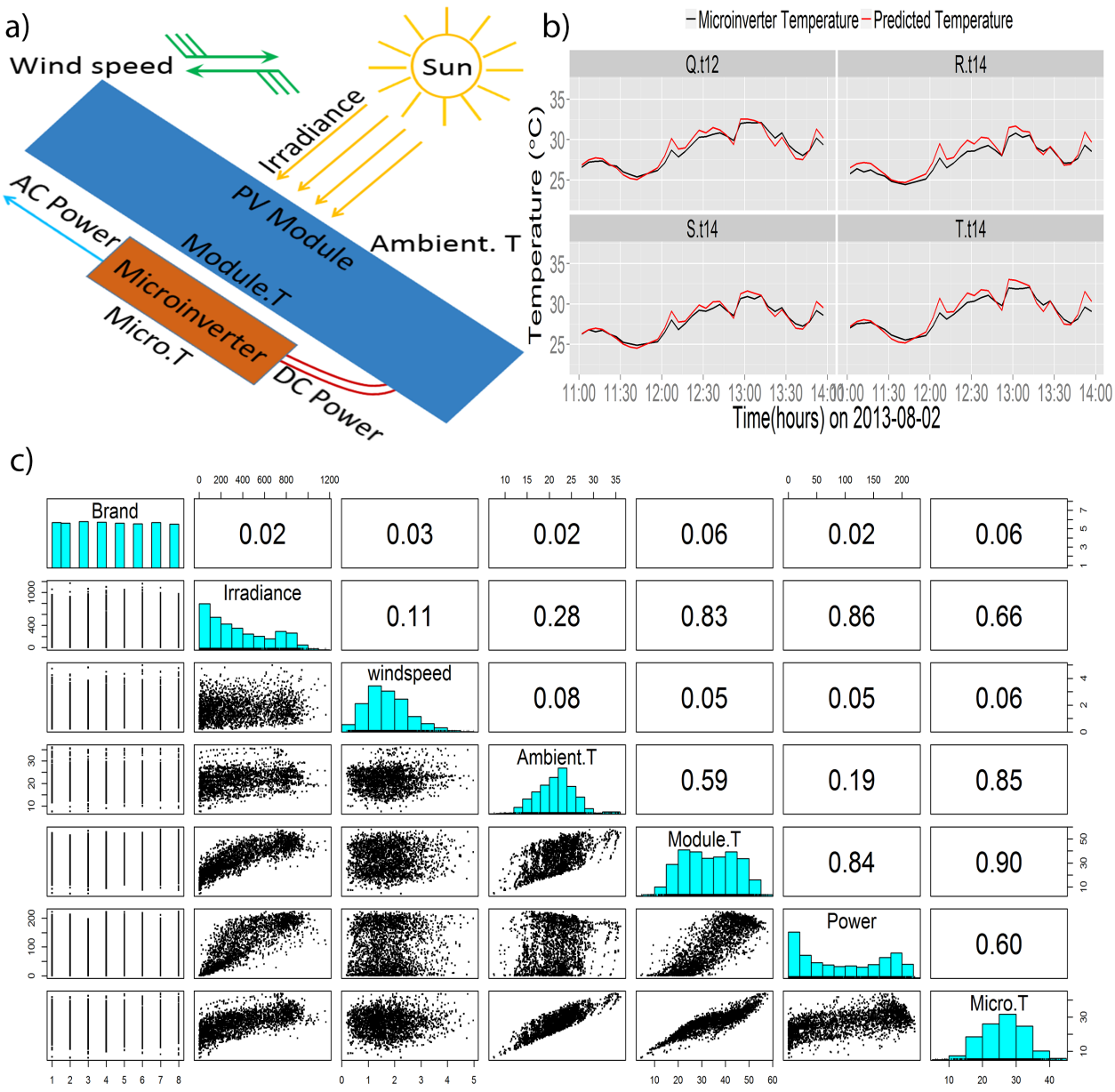


Figure 3: Figure 9.