Exploratory Data Analysis with R

Introduction to EDA

Xuemao Zhang
East Stroudsburg University

September 2, 2022

What's covered in this lecture?

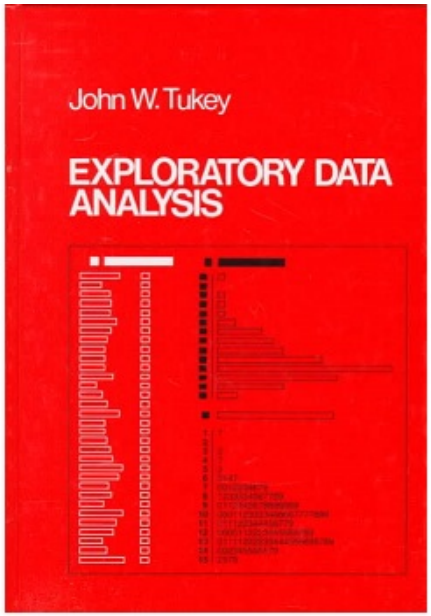
- Introduction to EDA
- A first look at EDA
- EDA with R
- A representative R session

Introduction to EDA

- Why do we analyze/summarize data?
  - to understand what has happened or what is happening;
  - to predict what is likely to happen, either in the future or in other circumstances we haven't seen yet;
  - to guide us in making decisions.
- We focus on data visualizations in this course. Predictions will require statistical models.

Introduction to EDA

- John W. Tukey (1977; Exploratory Data Analysis): "The greatest value of a picture is when it forces us to notice what we never expected to see.''



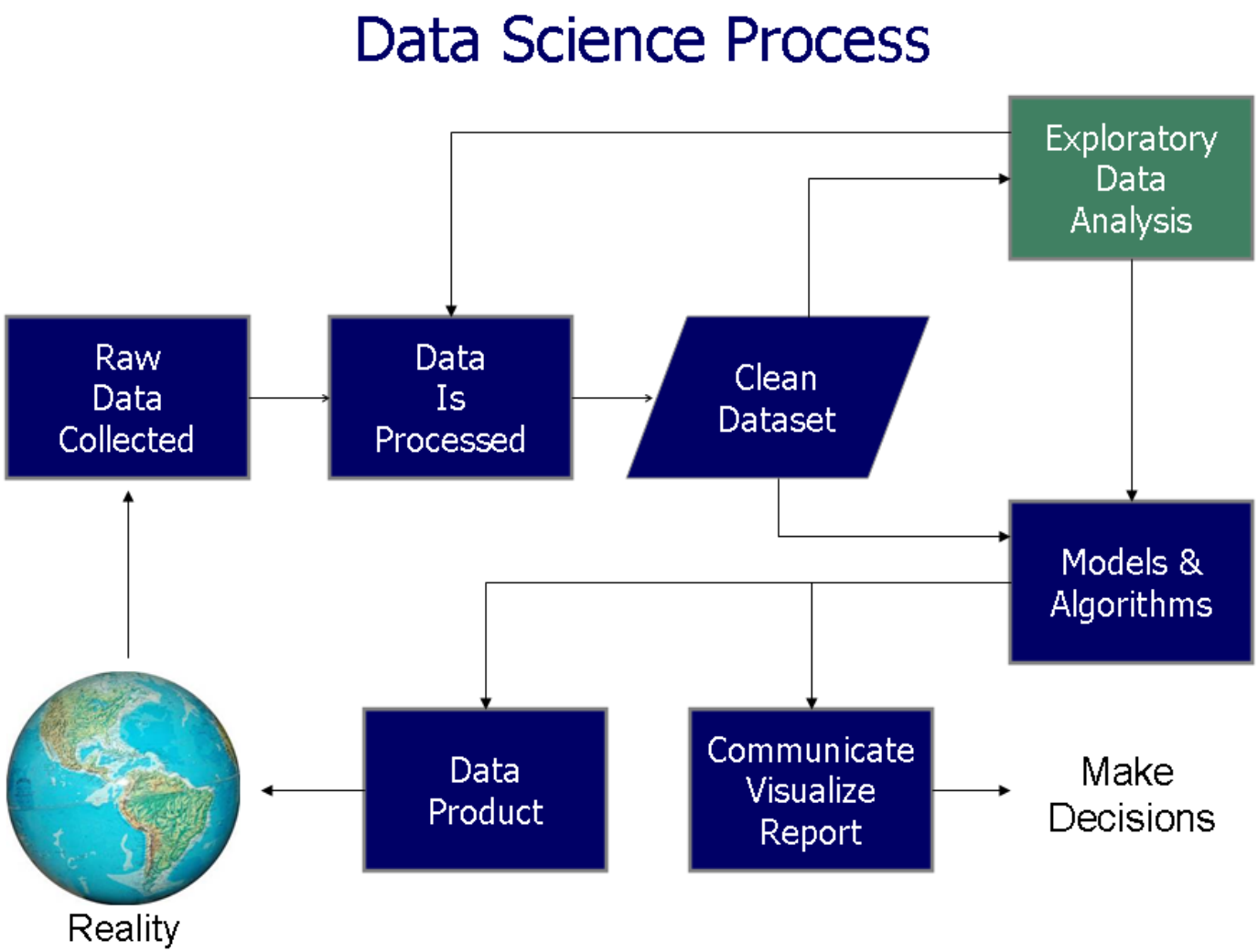
- John W. Tukey coined terms: Boxplot, Stem-and-Leaf plot, ANOVA (Analysis of Variance); "Bit" and "Software".

A First Look at EDA

- The EDA is a statistical approach to make sense of data by using a variety of techniques (mostly graphical). It may help
  - Assess assumption about variables distribution
  - Identify relationship between variables
  - Extract important variables
  - Suggest use of appropriate models
  - Detect problems of collected data (e.g. outliers, missing data, measurement errors)

A First Look at EDA

Data science process flowchart

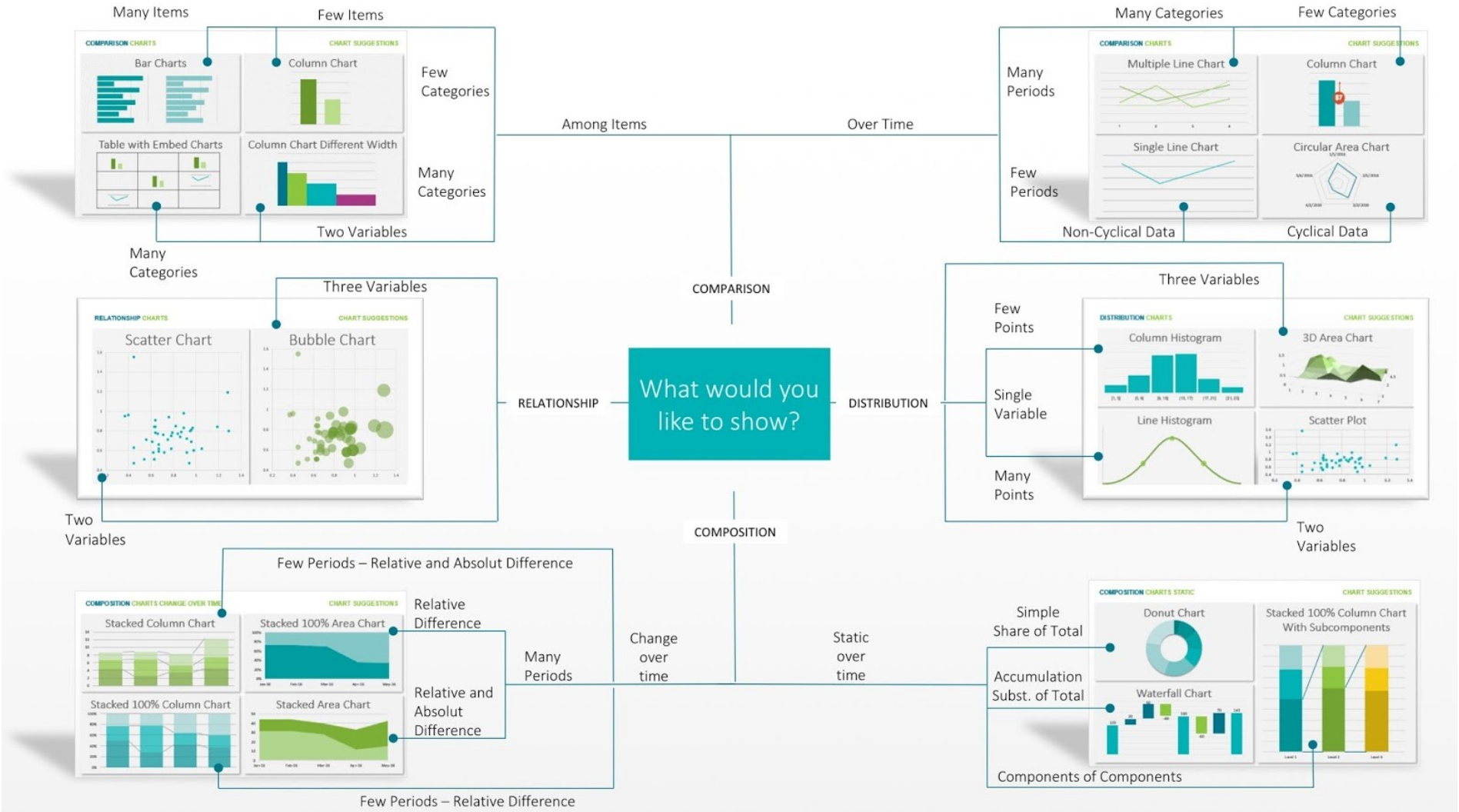


A First Look at EDA - Statistical Graphics

- **Univarite**
  - Histogram, Stem-and-Leaf, Dot, Q-Q, Density plots
  - Box-and-whisker, Violin
  - Bar, Pie, Polar, Waterfall charts
- **Bivariate**
  - Scatter, Line, Area, Bubble charts
- **Trivariate**
  - 3D Scatter, Contour, Level/Heatmap, Surface plots

A First Look at EDA - Statistical Graphics

Which chart to use?



EDA with R - R console

- R console
  - You can enter commands one at a time at the command prompt (>). For example,

```
3+5
```

```
## [1] 8
```

The above is the form of code in our presentations: The first line is what I typed into the console; the second line is my result.

Note: # is the comment symbol in R.

When you run the code 3+5 in your local console, you type after the command prompt > and it will look like this:

```
> 3+5
[1] 8
```

- Or you can type your R code in an R chunk in an .Rmd file.

An R session

- R packages
  - installing an R package using `install.packages()` if it is not installed
  - You must use the `library()` function to load the required packages into your current R session.

```
library(MASS)
```

- The command loads the MASS package which contains the `whiteside` data set which is an R data frame.
- An R data frame is a rectangular array of n records/**observations** each represented as a row with p fields per record, each representing a value of a particular **variable** for that record.

An R session

- This structure may be seen by applying the head function to the `whiteside` data frame, which displays its first few records:

```
head(whiteside);
```

```
##    Insul Temp Gas
## 1 Before -0.8 7.2
## 2 Before -0.7 6.9
## 3 Before  0.4 6.4
## 4 Before  2.5 6.0
## 5 Before  2.9 5.8
## 6 Before  3.2 5.8
```

```
tail(whiteside);   #show the last several rows
```

```
##    Insul Temp Gas
## 51 After  7.2 2.8
## 52 After  7.5 2.6
## 53 After  8.0 2.7
## 54 After  8.7 2.8
## 55 After  8.8 1.3
## 56 After  9.7 1.5
```

An R session

- A more detailed view of this data frame is provided by the `str` function, which returns structural characterizations of essentially any R object.

```
str(whiteside);
```

```
## 'data.frame':    56 obs. of  3 variables:
##  $ Insul: Factor w/ 2 levels "Before","After": 1 1 1 1 1 1 1 1 1
##  $ Temp : num  -0.8 -0.7 0.4 2.5 2.9 3.2 3.6 3.9 4.2 4.3 ...
##  $ Gas  : num  7.2 6.9 6.4 6 5.8 5.8 5.6 4.7 5.8 5.2 ...
```

An R session

- to see the dimensions of the data

```
nrow(whiteside);     # number of data rows
```

```
## [1] 56
```

```
ncol(whiteside);     # number of columns
```

```
## [1] 3
```

```
dim(whiteside);      # dimensions of the data
```

```
## [1] 56  3
```

An R session

- The $ sign.
  - It is an operator in R which extract,replace or add parts of an R object

```
whiteside$Temp;
```

```
##  [1] -0.8 -0.7  0.4  2.5  2.9  3.2  3.6  3.9  4.2  4.3  5.4  6.0
## [16]  6.3  6.9  7.0  7.4  7.5  7.5  7.6  8.0  8.5  9.1 10.2 -0.7
## [31]  1.5  1.6  2.3  2.5  2.5  3.1  3.9  4.0  4.0  4.2  4.3  4.6
## [46]  4.9  5.0  5.3  6.2  7.1  7.2  7.5  8.0  8.7  8.8  9.7
```

```
whiteside$v4=1;
head(whiteside);
```

```
##    Insul Temp Gas v4
## 1 Before -0.8 7.2  1
## 2 Before -0.7 6.9  1
## 3 Before  0.4 6.4  1
## 4 Before  2.5 6.0  1
## 5 Before  2.9 5.8  1
## 6 Before  3.2 5.8  1
```

An R session

- summary is a generic function in R, which returns a relatively simple characterization of the values each variable can assume.

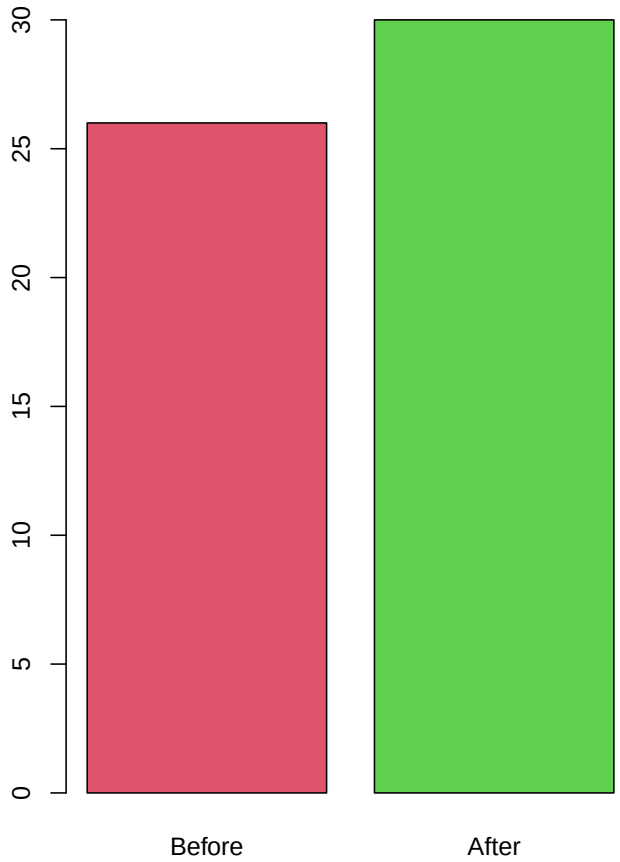
```
summary(whiteside);
```

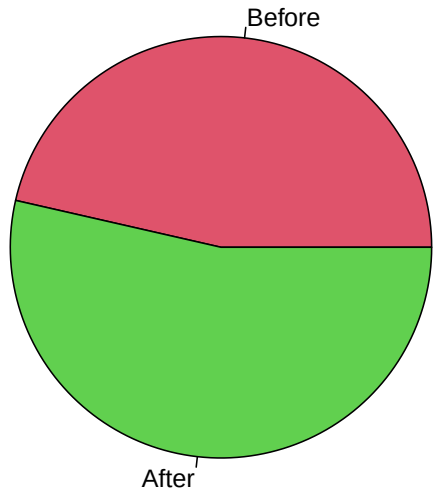
```
##     Insul          Temp             Gas              v4
##  Before:26   Min.   :-0.800   Min.   :1.300   Min.   :1
##  After :30   1st Qu.: 3.050   1st Qu.:3.500   1st Qu.:1
##              Median : 4.900   Median :3.950   Median :1
##              Mean   : 4.875   Mean   :4.071   Mean   :1
##              3rd Qu.: 7.125   3rd Qu.:4.625   3rd Qu.:1
##              Max.   :10.200   Max.   :7.200   Max.   :1
```

An R session

- The variable `Insul` is a factor variable with two levels: `Before` and `After`.

```
par(mfrow=c(1,2));
barplot(table(whiteside$Insul), col=c(2,3));
pie(table(whiteside$Insul), col=c(2,3));
```

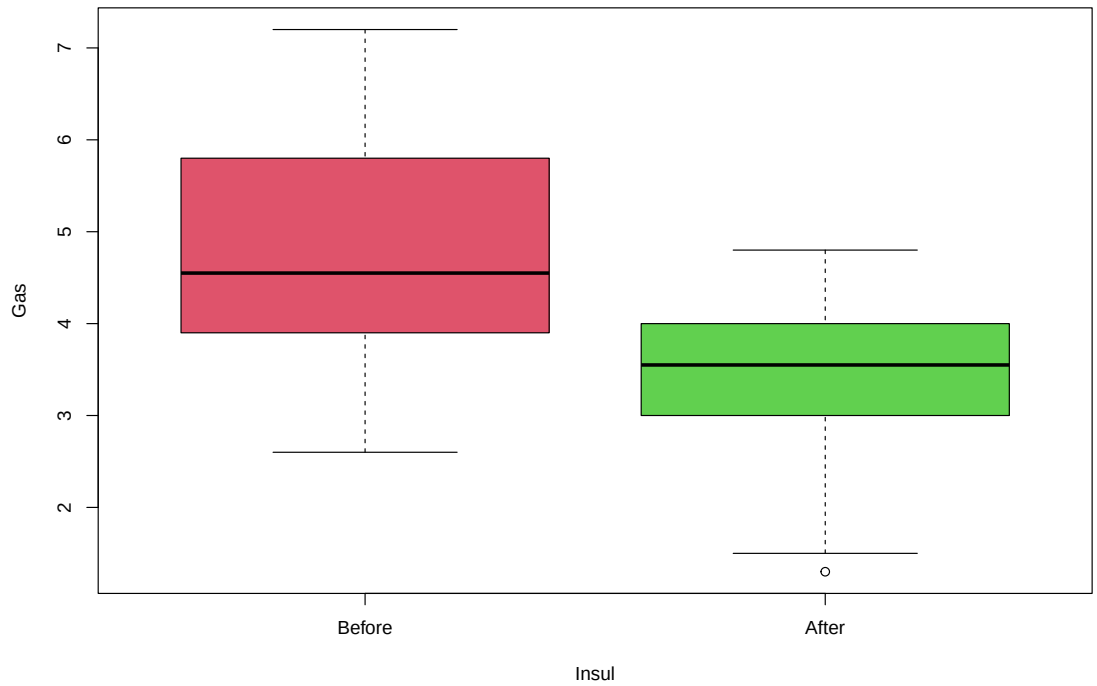




An R session

- Side-by-side boxplot comparison of the Before and After subsets of the Gas values from the whiteside data frame.

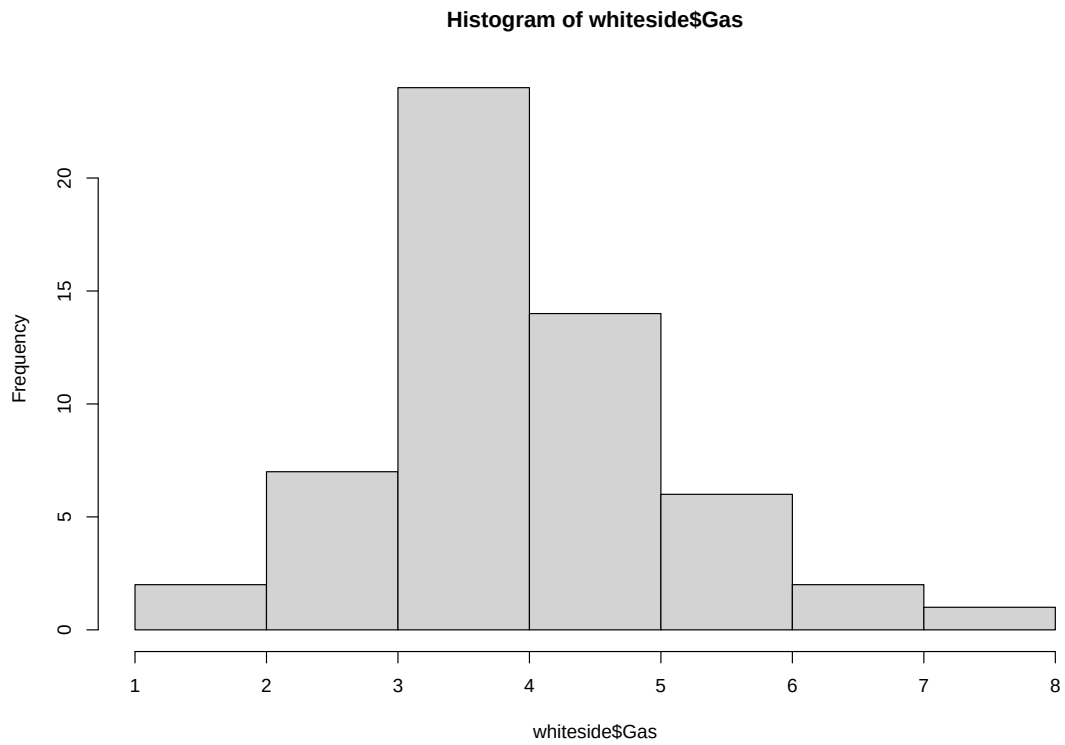
```
boxplot(Gas ~ Insul, data = whiteside, col=c(2,3));
```



An R session

- Distribution of a single variable
  - histogram

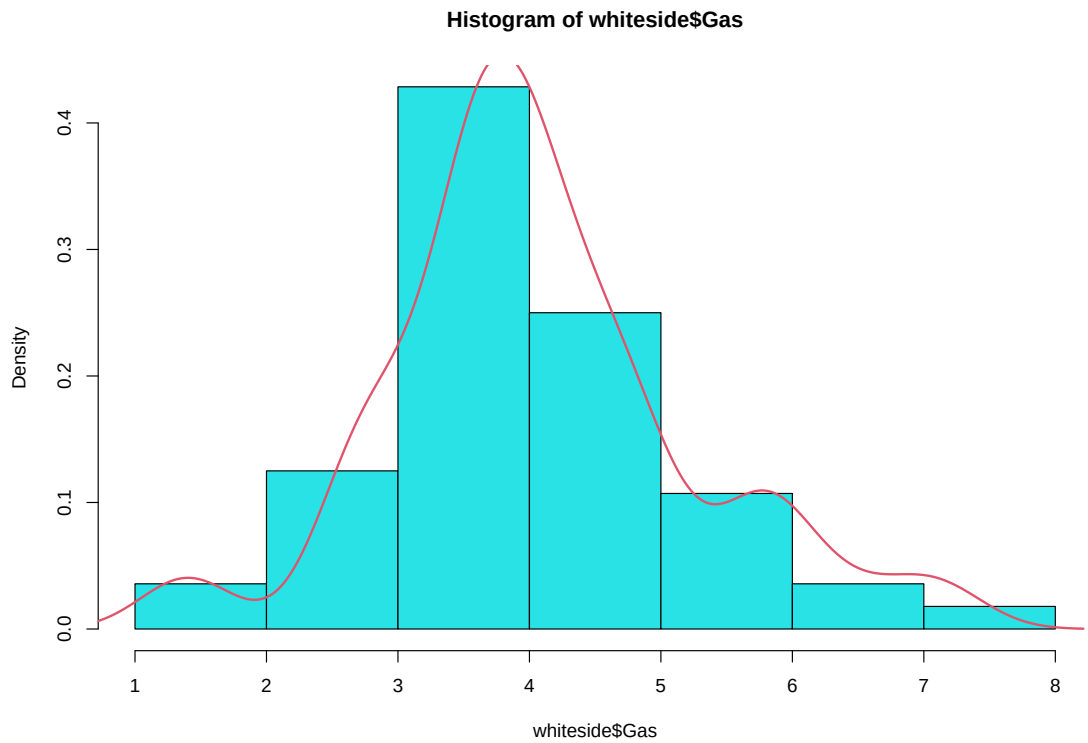
```
hist(whiteside$Gas)
```



EDA with R - an R session

- Distribution of a single variable
  - relative histogram with estimated density function

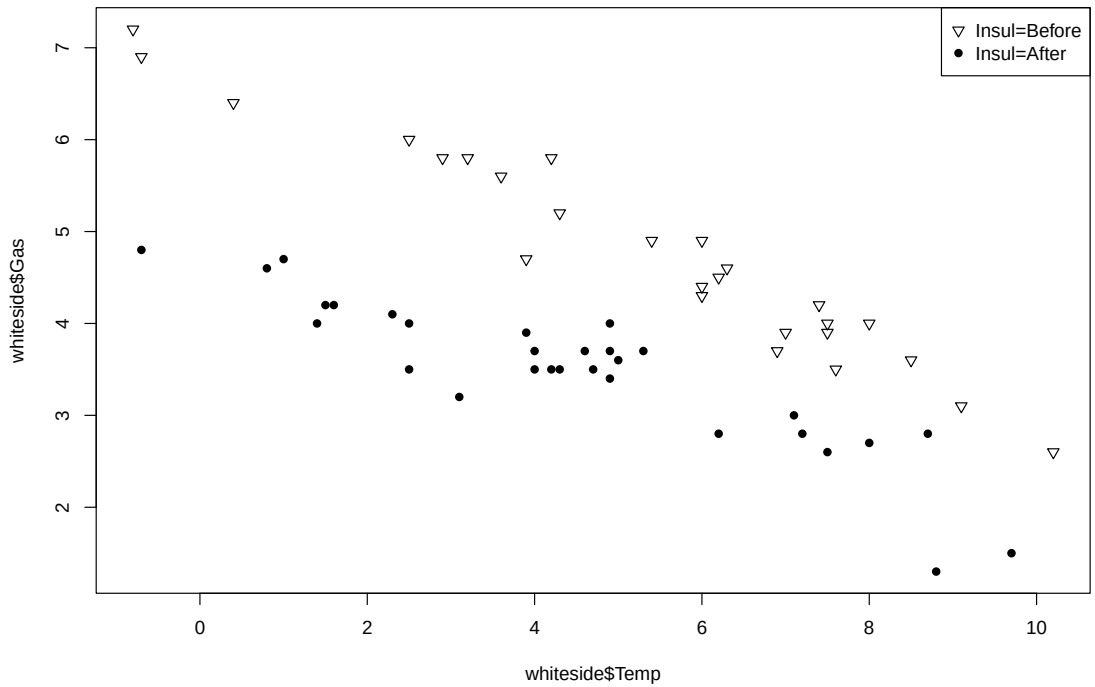
```
hist(whiteside$Gas, col=5,prob=TRUE);#show relative frequencies
lines(density(whiteside$Gas),col=2,lwd=2); # density plot
```



An R session

- Check the relationship between the two numerical variables
  - plot with different symbols for the two heating seasons (i.e., Before and After).

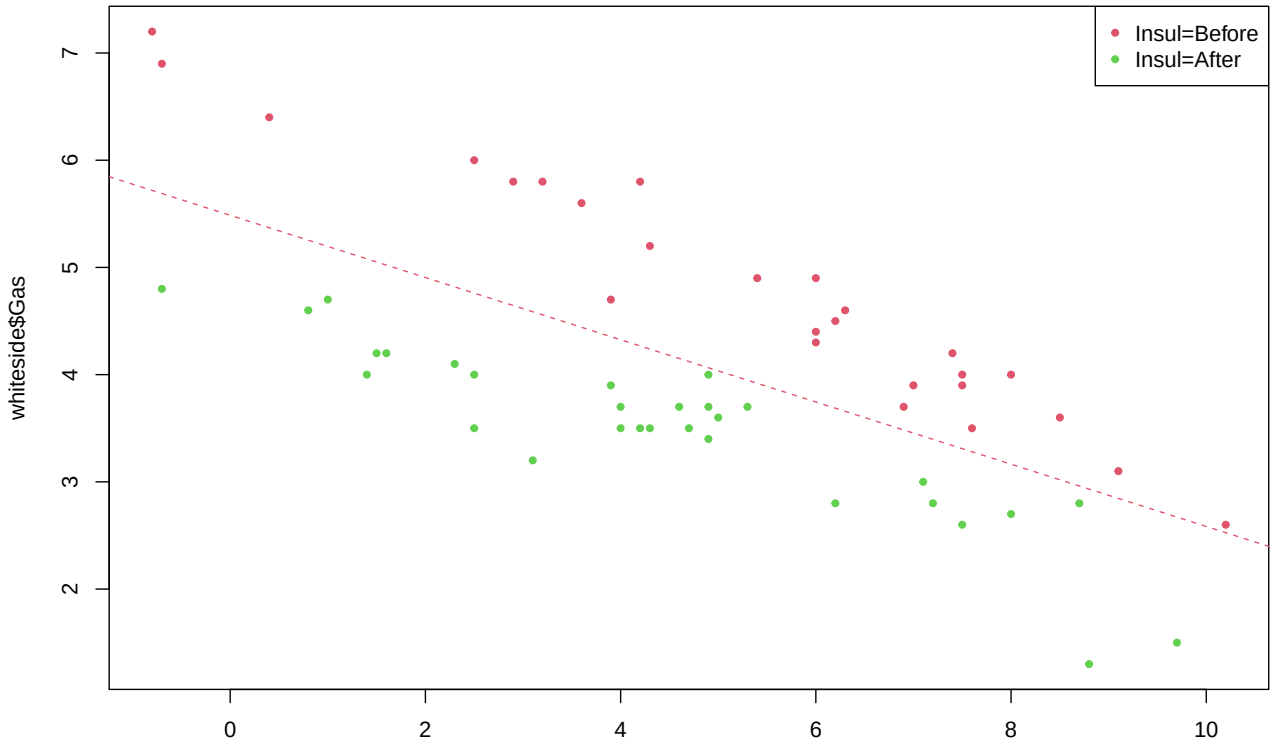
```
plot(whiteside$Temp, whiteside$Gas, pch=c(6,16)[whiteside$Insul]);
legend(x="topright",legend=c("Insul=Before","Insul=After"),
       pch=c(6,16));
```



An R session

- Fit the data using a simple linear regression model

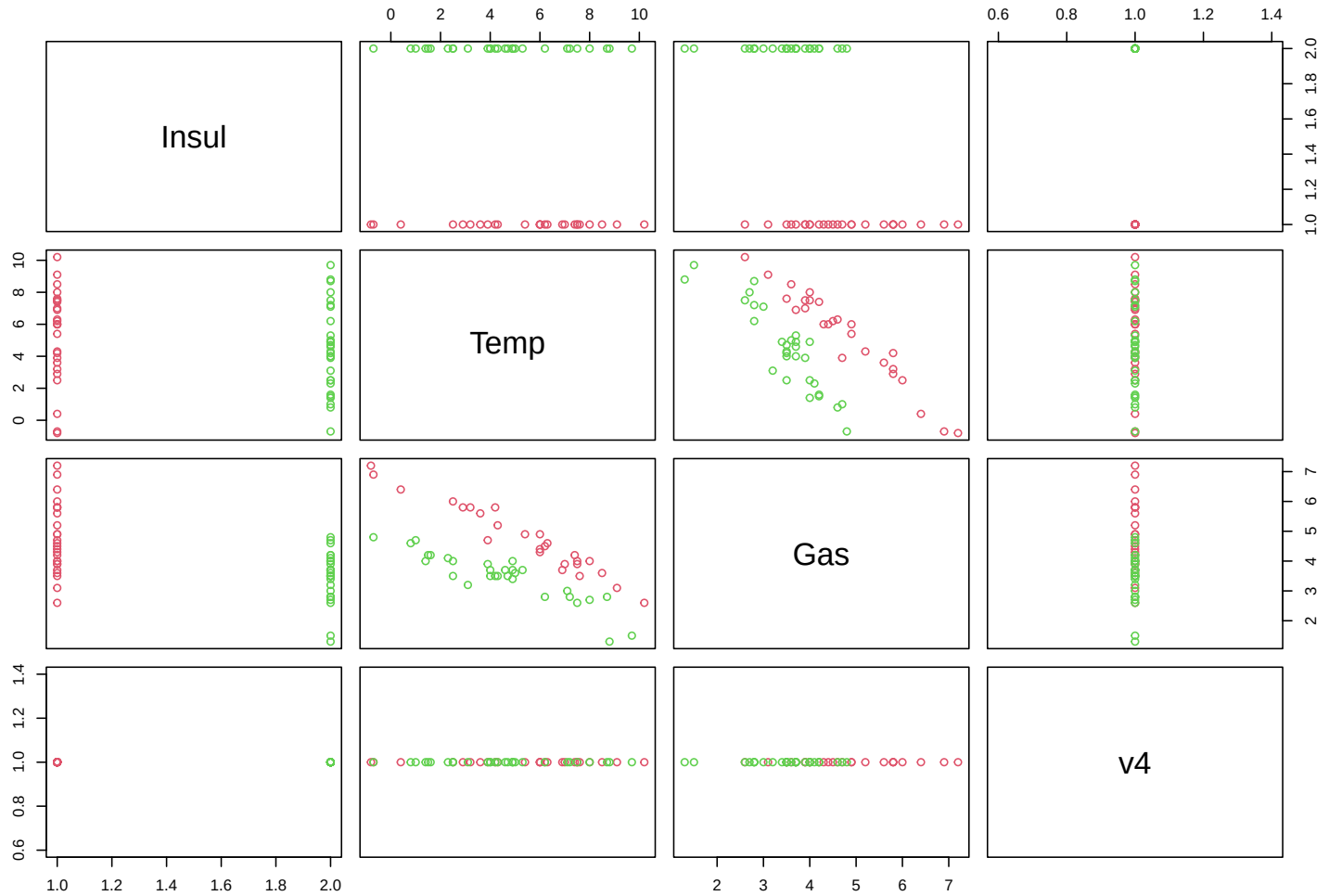
```
plot(whiteside$Temp, whiteside$Gas, pch=20,
     col=c(2,3)[whiteside$Insul]);
abline(lm(whiteside$Gas~whiteside$Temp), col=c(2,3), lty=2);
legend(x="topright",legend=c("Insul=Before","Insul=After"),
       pch=20,col=c(2,3));
```



An R session

- Applying the `plot` function to the `whiteside` data frame: it generates a matrix of scatterplots, showing how each variable relates to the others.

```
plot(whiteside,col=c(2,3)[whiteside$Insul]);
```



License