Introduction to R

Outline

- Processing Data in R
- Programming in R
- Graphical Analysis in R
- Statistical Analysis in R

Contingency Tables

- Basis for performing a statistical test on the independence of the factors used to build the table.
- table(sales$gender)

 # Build an empty character vector

- sales_group<-vector(mode="character", length= length(sales$sales_total))

 # Group the customers based upon sales amount

- sales_group[sales$sales_total<100]<-"small"
- sales_group[sales$sales_total>=100 & sales$sales_total<500]<-"medium"
- sales_group[sales$sales_total>500]<-"big“

#Create and add an ordered factor to sales data frame

- spender<-factor(sales_group, levels=c("small","medium","big"),ordered=TRUE)
- sales<-cbind(sales,spender)

Contingency Tables…

- str(sales$spender)
- head(sales$spender)

Build a continency table

- sales_table<-table(sales$gender,sales$spender)
- sales_table
- class(sales_table)
- typeof(sales_table)
- dim(sales_table)

Perform chi-squared test

- summary(sales_table)

  Number of cases in table: 10000 Number of factors: 2
  Test for independence of all factors:
  Chisq = 1.516, df = 2, p-value = 0.4686

Exploratory Data Analysis

Spotting Problems /Cleaning Dirty Data

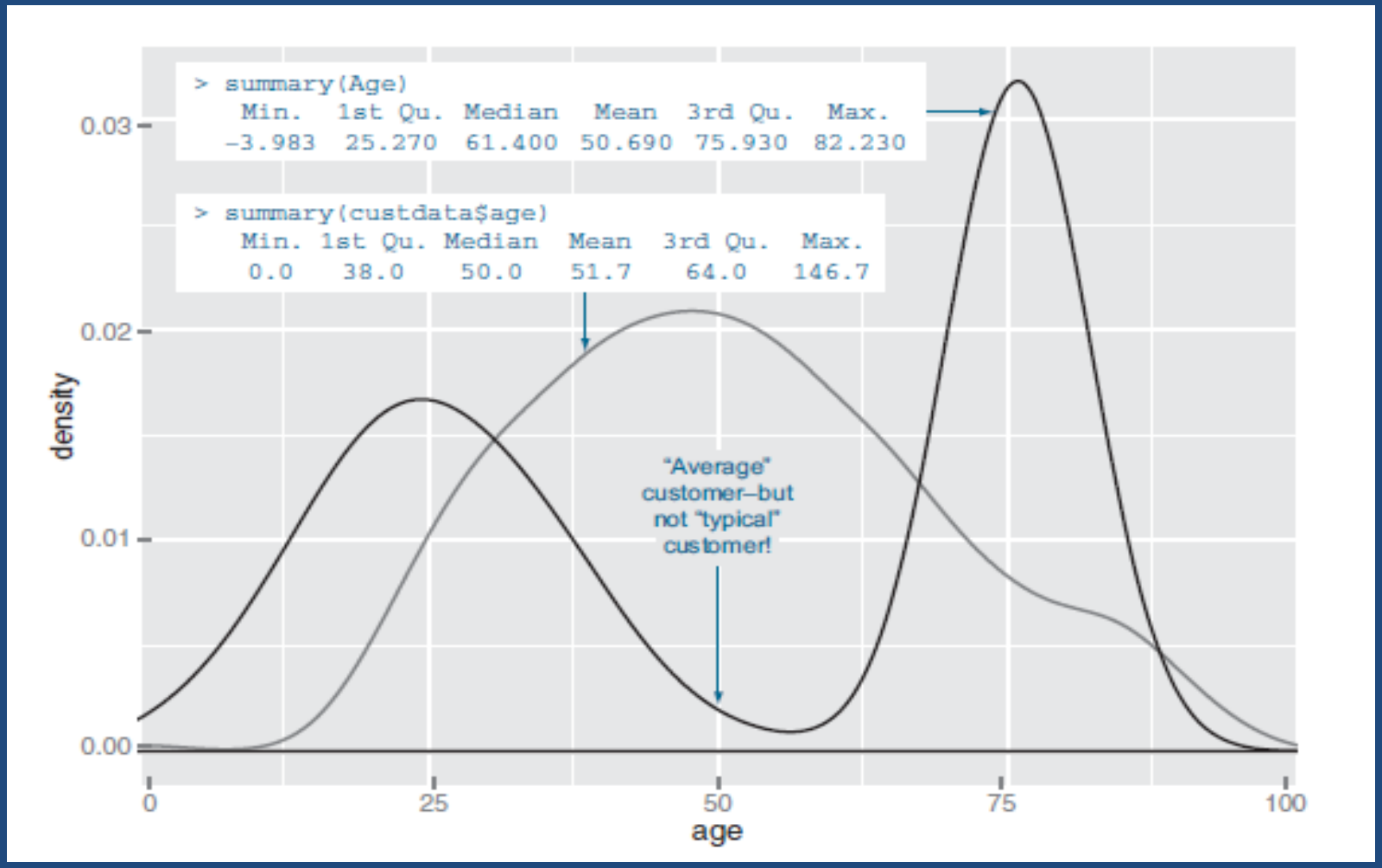
Summary()

- Typical Problems revealed:
  - Missing Values(NA)
    - How to address them? Drop / Zero / Convert/New category
  - Invalid values and outliers
    - Drop field/data point or convert
  - Data Range
    - Pretty wide / too narrow(relative)
    - Rule of thumb: (sd/mean) – very small – data isn't varying much
  - Units
    - Time – minutes/ hours/days
    - Speed-Kms per sec / miles per hour

Visualization

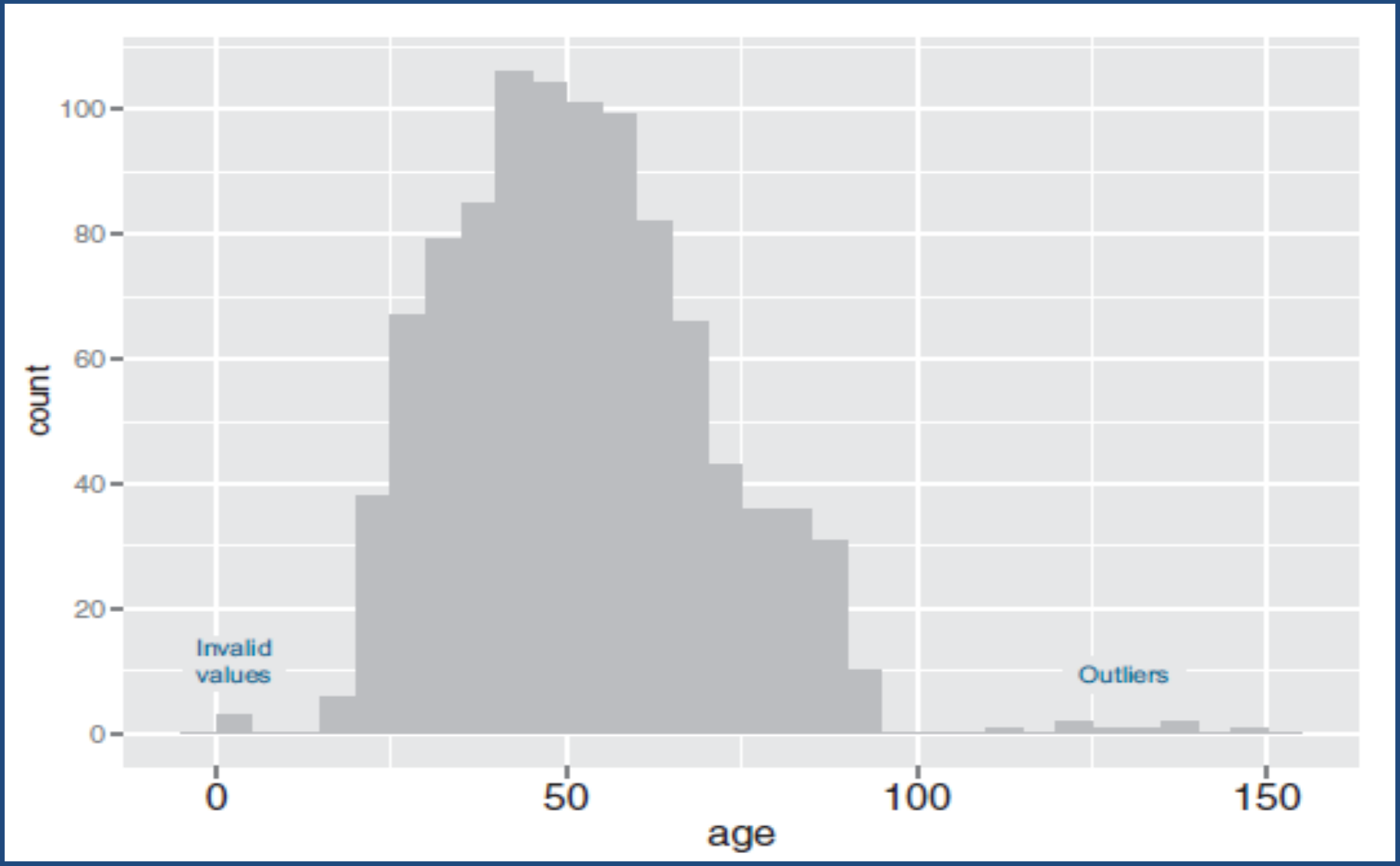
- Single Variable
  - What is the peak value of the distribution?
  - How many peaks are there in the distribution (unimodality versus bimodality)?
  - How normal (or lognormal) is the data?
  - How much does the data vary? Is it concentrated in a certain interval or in a certain category?

Unimodal / Bimodal



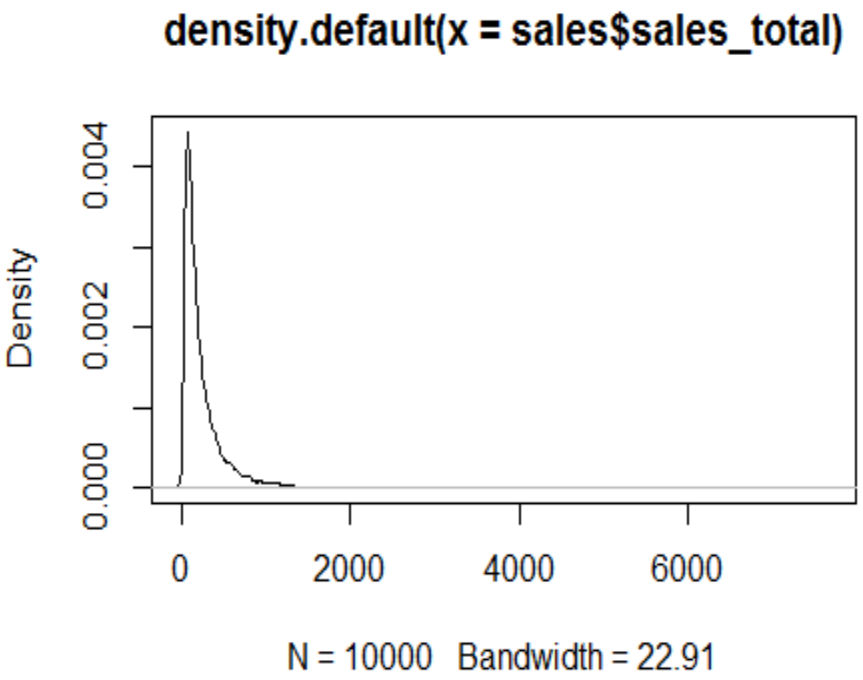
Histogram

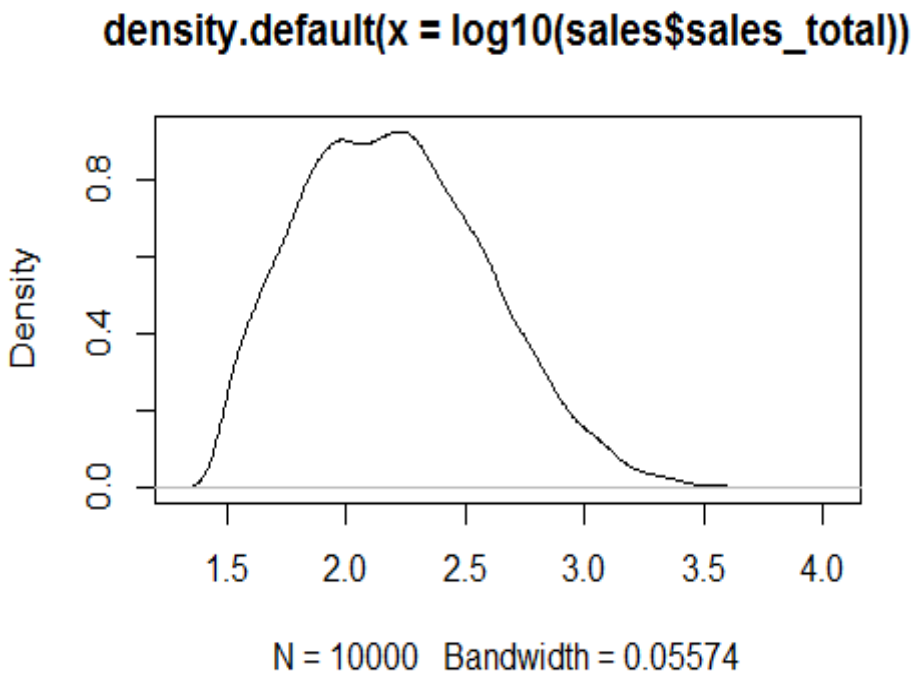
- Data concentration; outliers; anamolies



Density Plot

plot(density(sales$sales_total))
plot(density(log10(sales$sales_total)))

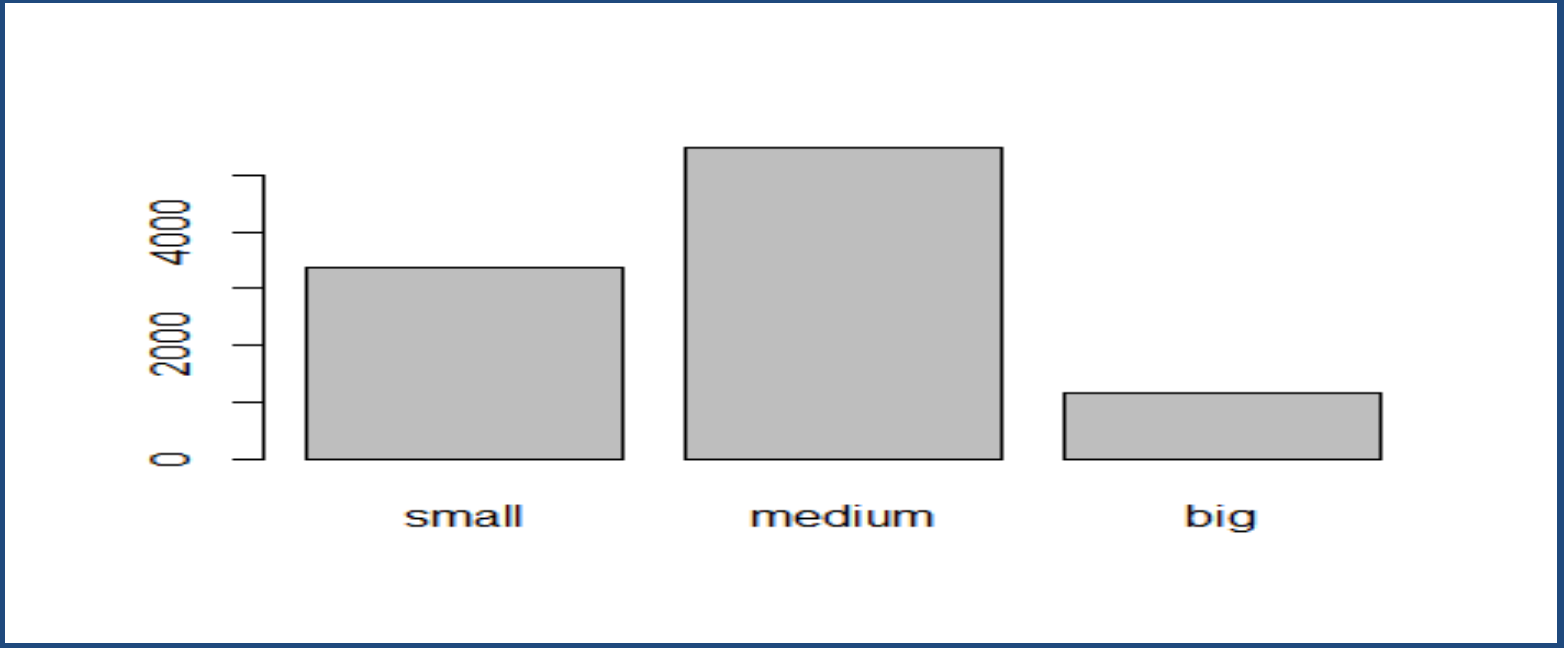




Most customers have sales_total b/w 17 to 27,000

Bar Chart

- Histogram for discrete data
- barplot(table(sales$spender))



- **A sample should enough customers from different categories**

Visualization

- Two Variables
  - Is there a relationship between the two inputs in my data?
  - What kind of relationship, and how strong?
  - Is there a relationship between the input *x* and the output *y?* How strong?

Scatter plot - regression

 # uniform distribution

- x<-runif(75,0,10)
- x<-sort(x)
- y<-200+x^3-10*x^2+x+rnorm(75,0,20)
- plot(x,y)

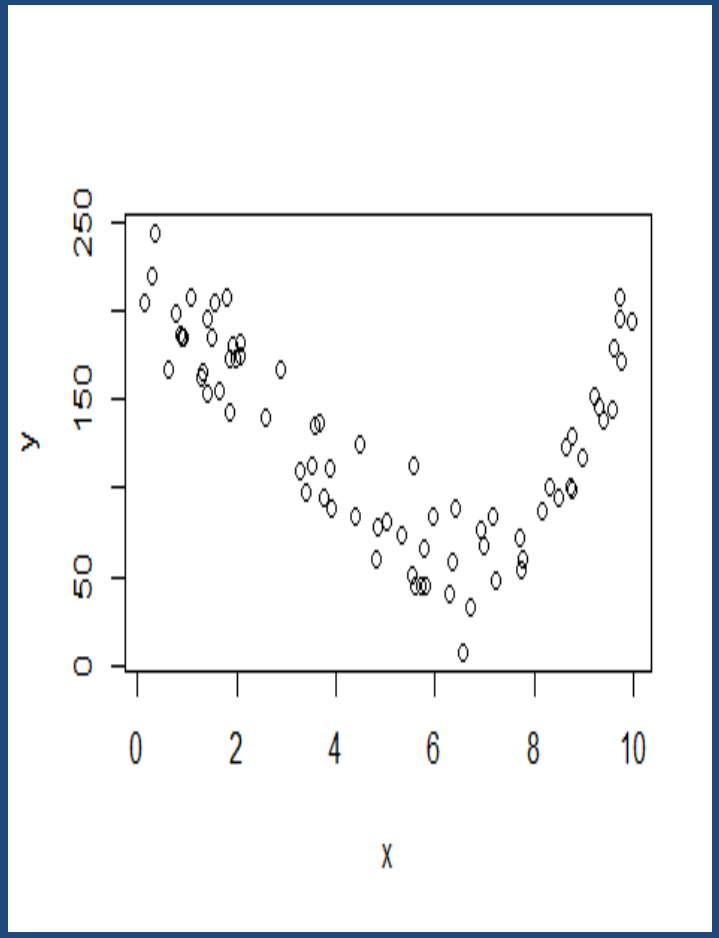
 # Linear Regression

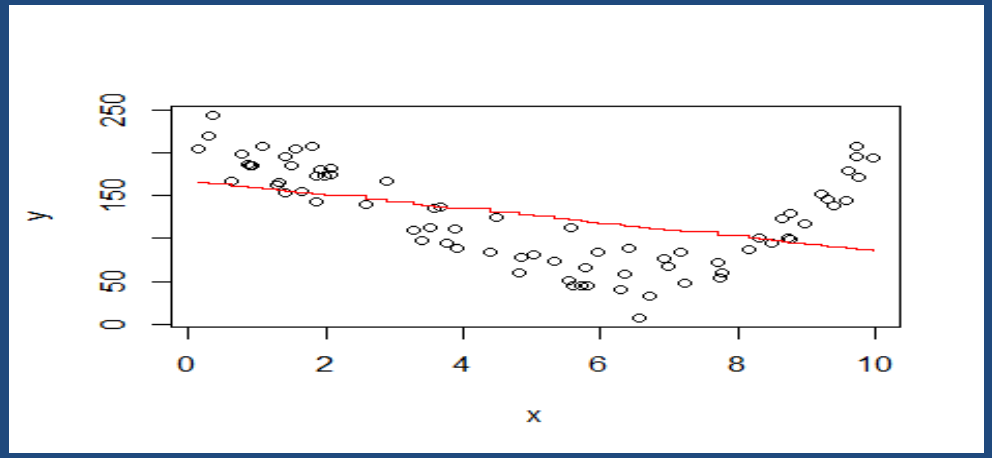
- lr<-lm(y~x)
- points(x,lr$coefficients[1] + lr$coefficients[2]* x, type="b",col=2)

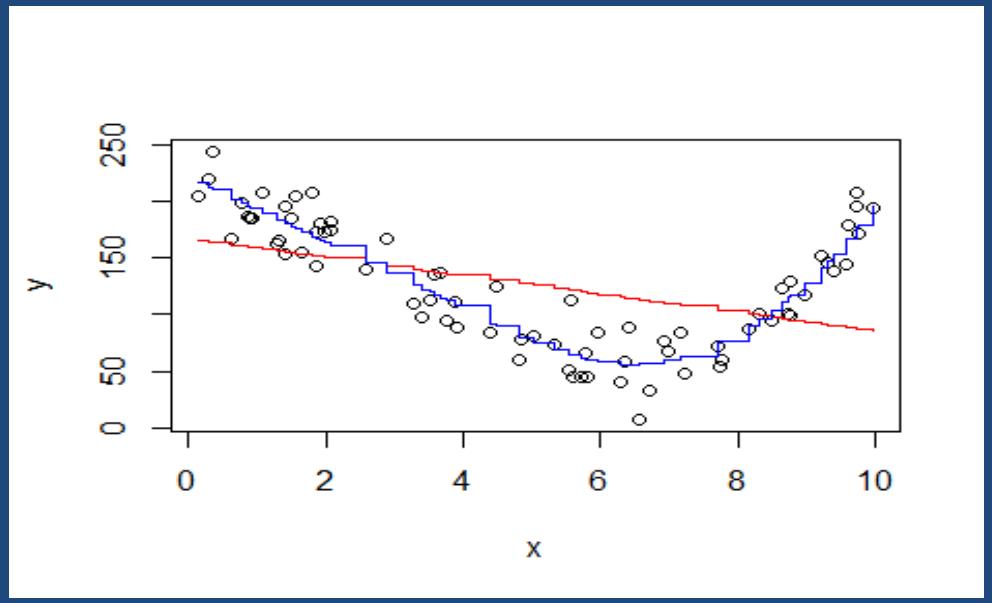
 # Non-linear Regression

- poly<-loess(y~x)
- fit<-predict(poly)
- points(x,fit, type="b", col=4)

Scatter plot – regression…





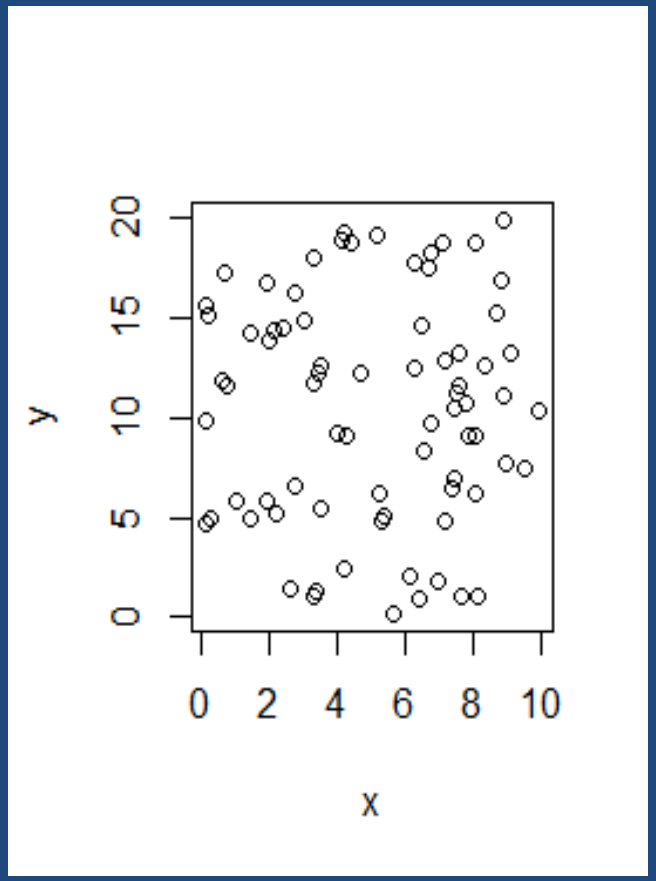


Scatter plot – regression…

- If the plot looks more like a **cluster** without a pattern, the corresponding variables may have a weak relationship

Example:

```
>x<-runif(75,0,10)
>y<-runif(75,0,20)
>plot(x,y)
```



Fixing Data Quality Problems

Data Cleaing

Missing Values: Categorical

- load("exampleData.rData")

 # NA common in three variables of 56 rows – drop?

- summary(custdata[is.na(custdata$housing.type), c("recent.move","num.vehicles")])

#NA in 1/3 rd of rows

- summary(custdata$is.employed)

 # Create a new category called Missing

```
custdata$is.employed.fix <- ifelse(is.na(custdata$is.employed),
                                   "missing",
                                   ifelse(custdata$is.employed==T,
                                          "employed",
                                          "not employed"))
summary(as.factor(custdata$is.employed.fix))
```

Missing Data: Numerical

- summary(custdata$Income)

#Find mean by removing na rows

- meanIncome <- mean(custdata$Income, na.rm=T)

#Replace na by mean of Income

- custdata$Income.fix <-
  ifelse(is.na(custdata$Income),
  meanIncome,
  custdata$Income)
- summary(custdata$Income.fix)
- Alternate Options: Categorize the attribute; Put 0

Transformation: Continuous to Discrete

#Binary: Income Less than 20000 or not?

- custdata$income.lt.20K <- custdata$income < 20000
- summary(custdata$income.lt.20K)

 # Multiple Categories: age

- brks <- c(0, 25, 65, Inf)
- custdata$age.range <- cut(custdata$age,
  breaks=brks, include.lowest=T)
- summary(custdata$age.range)

Transformation: Normalization and Rescaling

- summary(custdata$age)
- meanage <- mean(custdata$age)
- custdata$age.normalized <- custdata$age/meanage
- summary(custdata$age.normalized)

Transformation: Sampling

 # Add a sample group column to data set(no. generated uniformly between 0 and 1

- custdata$gp <- runif(dim(custdata)[1])

Test and Training Set

- testSet <- subset(custdata, custdata$gp <= 0.1)
- trainingSet <- subset(custdata, custdata$gp > 0.1)

- dim(testSet)[1]
- dim(trainingSet)[1]