

# Data Visualization

# Example: Boston\_housing Data

**This dataset contains information on neighborhoods in Boston.**

**Response: the median value of a housing unit in the neighborhood (medv)**

## **Variable Information:**

crim: per capita crime rate by town  
zn: proportion of residential land zoned for lots over 25,000 sq.ft.  
indus: proportion of non-retail business acres per town  
chas: Charles River variable (= tract if tract bounds river; = other if otherwise)  
nox: nitric oxides concentration (parts per 10 million)  
rm: average number of rooms per dwelling  
age: proportion of owner-occupied units built prior to 1940  
dis: weighted distances to five Boston employment centres  
rad: index of accessibility to radial highways  
tax: full-value property-tax rate per \$10,000  
ptratio: pupil-teacher ratio by town  
black:  $1000(B_k - 0.63)^2$  where  $B_k$  is the proportion of blacks by town  
lstat: % lower status of the population  
medv: Median value of owner-occupied homes in \$1000's

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.00632	18	2.31	other	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	other	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	other	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	other	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	other	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	other	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	other	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	other	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
0.21124	12.5	7.87	other	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	other	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
0.22489	12.5	7.87	other	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
0.11747	12.5	7.87	other	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
0.09378	12.5	7.87	other	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
0.62976	0	8.14	other	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
0.63796	0	8.14	other	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
0.62739	0	8.14	other	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9
1.05393	0	8.14	other	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
0.7842	0	8.14	other	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5
0.80271	0	8.14	other	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2
0.7258	0	8.14	other	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2
1.25179	0	8.14	other	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	13.6

# Read Data from CSV file

Read data from drive C

```
> data1 <- read.csv("C:/MA 299/R/Boston_housing.csv")
```

Ask for columns' names

```
> colnames(data1)
```

R will show the following columns' names

```
[1] "crim"  "zn"    "indus" "chas"  "nox"   "rm"    "age"  
[8] "dis"   "rad"   "tax"   "ptratio" "black" "lstat" "medv"
```

```
> summary(data1)
```

crim		zn	indus		chas	
Min.	: 0.00632	Min.	: 0.00	Min.	: 0.46	other:471
1st Qu.:	0.08204	1st Qu.:	0.00	1st Qu.:	5.19	tract: 35
Median	: 0.25651	Median	: 0.00	Median	: 9.69	
Mean	: 3.61352	Mean	: 11.36	Mean	:11.14	
3rd Qu.:	3.67708	3rd Qu.:	12.50	3rd Qu.:	18.10	
Max.	:88.97620	Max.	:100.00	Max.	:27.74	

nox		rm	age	dis	
Min.	:0.3850	Min.	:3.561	Min.	: 1.130
1st Qu.:	0.4490	1st Qu.:	5.886	1st Qu.:	2.100
Median	:0.5380	Median	:6.208	Median	: 3.207
Mean	:0.5547	Mean	:6.285	Mean	: 3.795
3rd Qu.:	0.6240	3rd Qu.:	6.623	3rd Qu.:	5.188
Max.	:0.8710	Max.	:8.780	Max.	:12.127

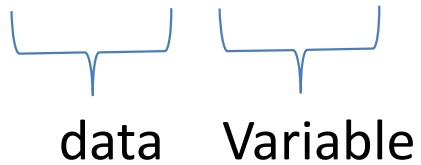
rad		tax	ptratio	black	
Min.	: 1.000	Min.	:187.0	Min.	: 0.32
1st Qu.:	4.000	1st Qu.:	279.0	1st Qu.:	375.38
Median	: 5.000	Median	:330.0	Median	:391.44
Mean	: 9.549	Mean	:408.2	Mean	:356.67
3rd Qu.:	24.000	3rd Qu.:	666.0	3rd Qu.:	396.23
Max.	:24.000	Max.	:711.0	Max.	:396.90

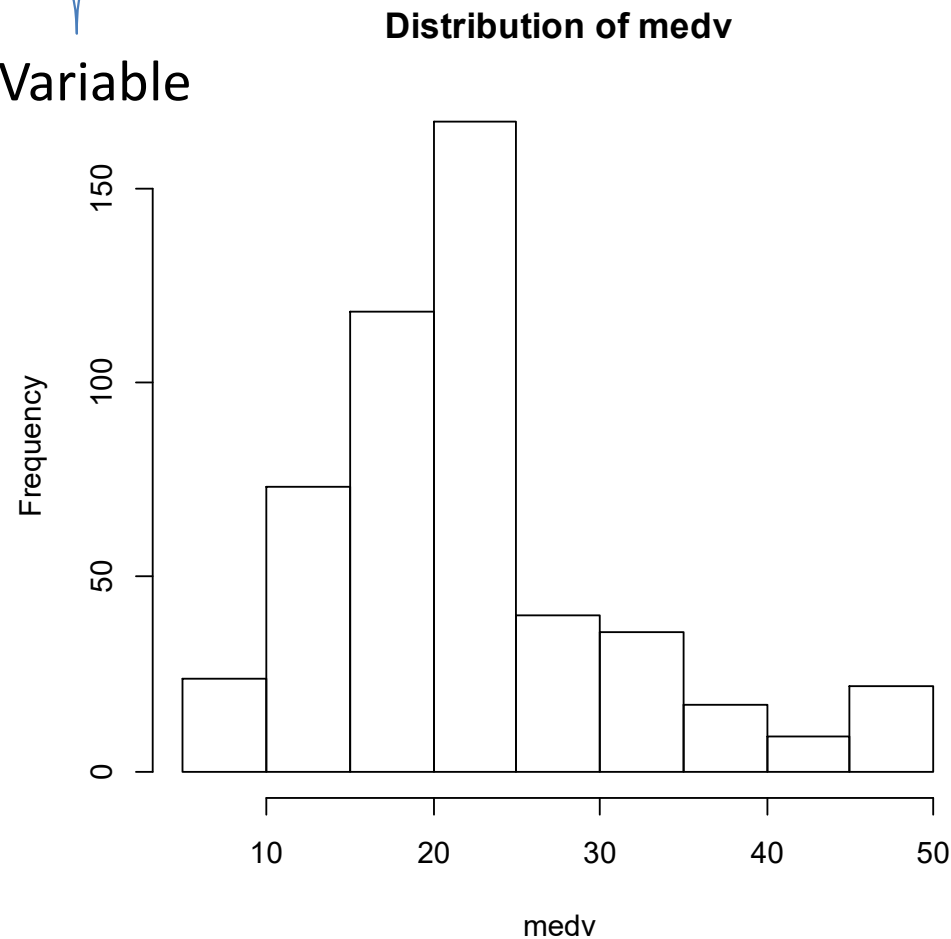
  

lstat		medv	
Min.	: 1.73	Min.	: 5.00
1st Qu.:	6.95	1st Qu.:	17.02
Median	:11.36	Median	:21.20
Mean	:12.65	Mean	:22.53
3rd Qu.:	16.95	3rd Qu.:	25.00
Max.	:37.97	Max.	:50.00

# Histogram

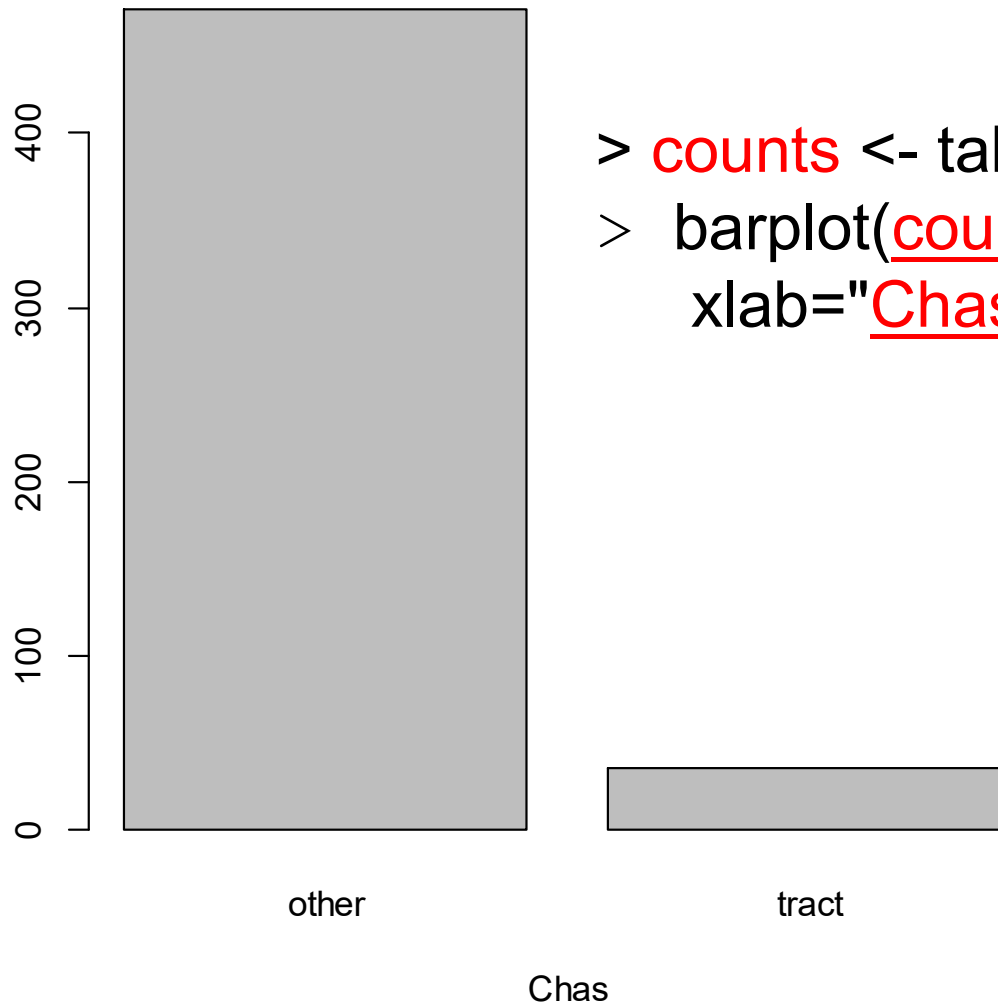
```
> hist(data1$medv, main="Distribution of medv", xlab="medv")
```

  
data    Variable



# Bar Chart

Bar chart of chas

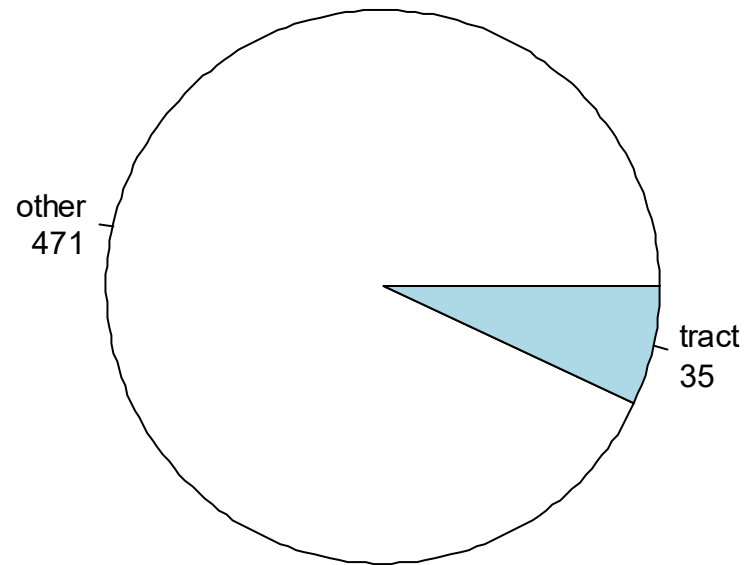


```
> counts <- table(data1$chas)  
> barplot(counts, main="Bar chart of chas",  
          xlab="Chas")
```

# Pie chart

```
> mytable <- table(data1$chas)  
> lbls <- paste(names(mytable), "\n", mytable, sep="")  
> pie(mytable, labels = lbls, main="Pie Chart of chas")
```

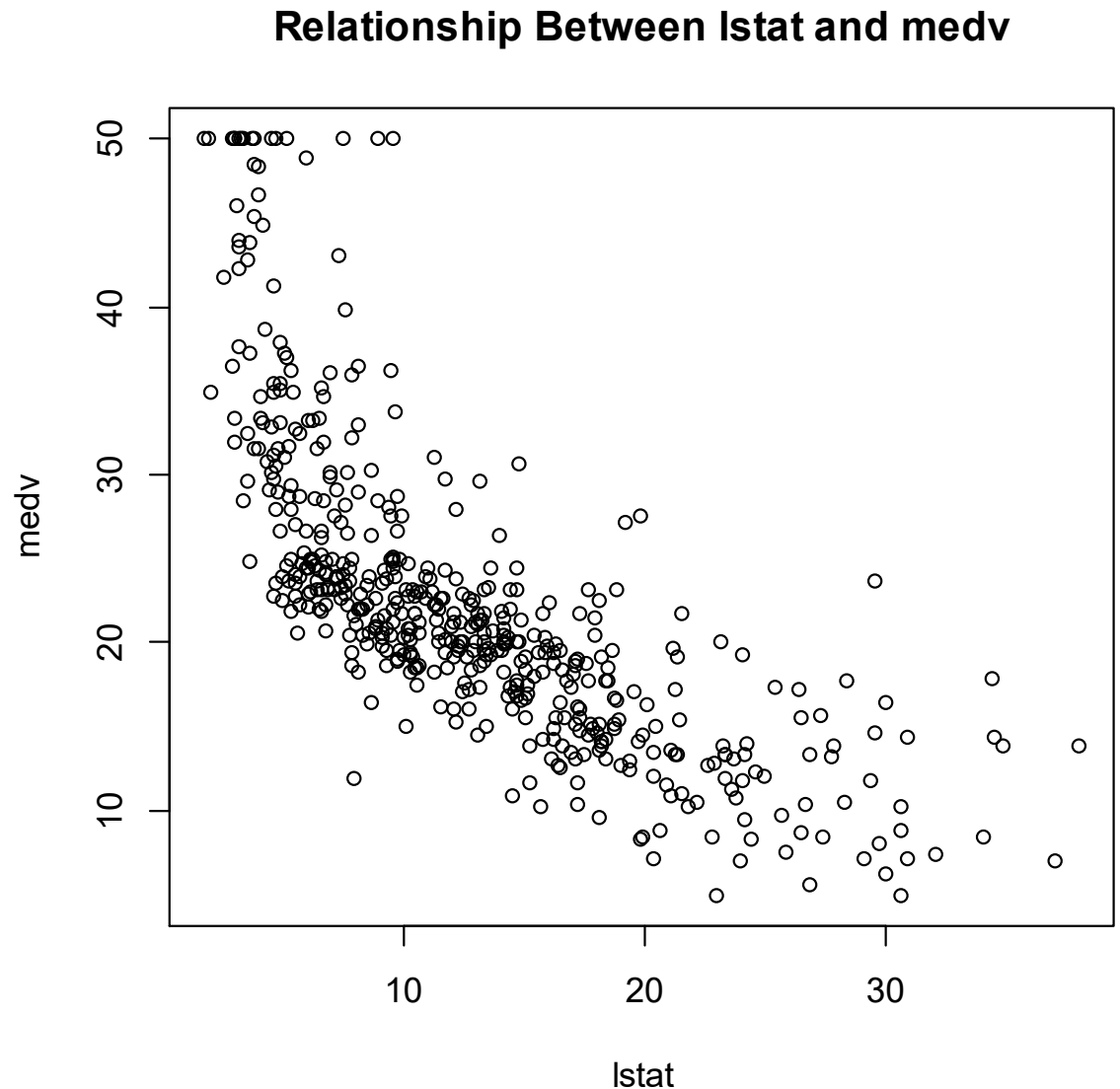
Pie Chart of chas





# Scatter Plot

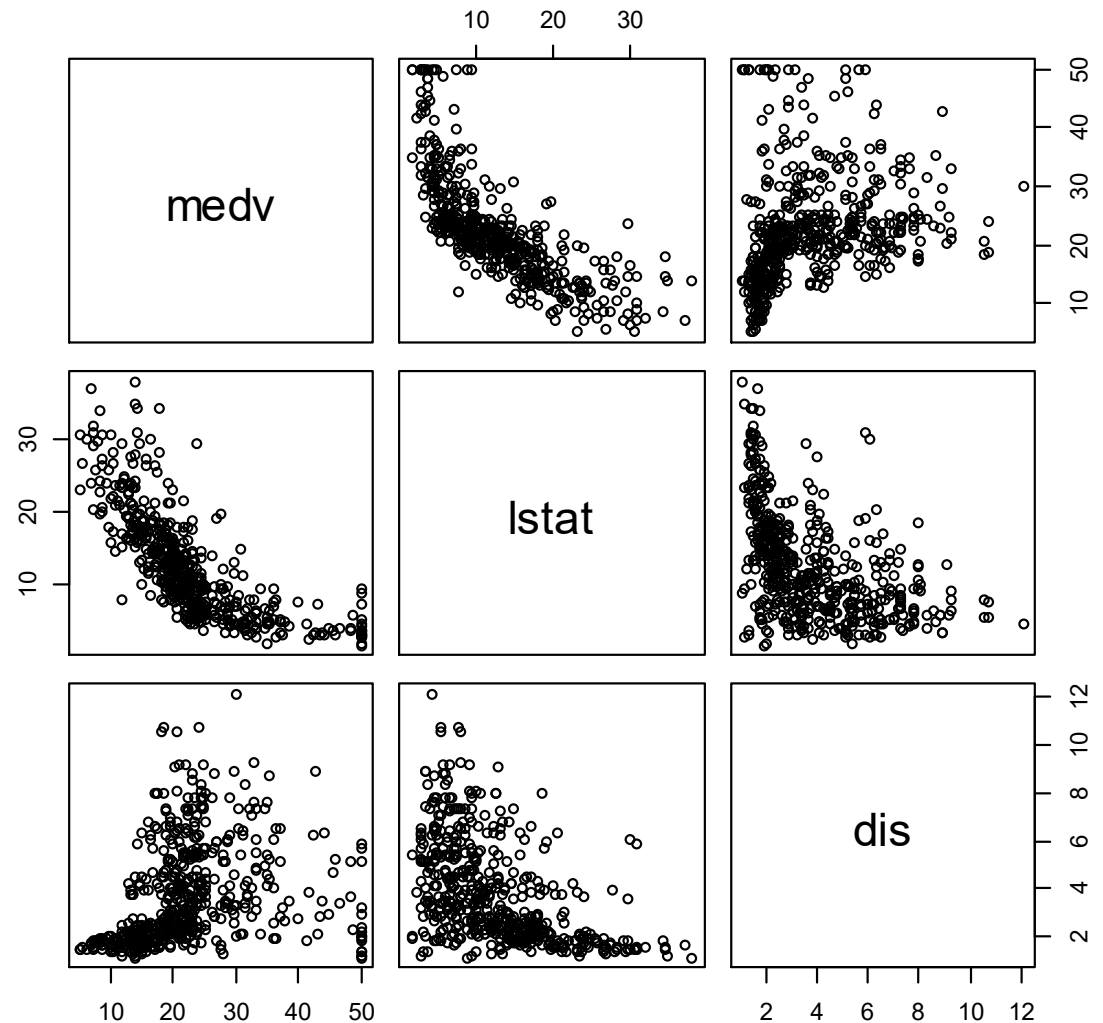
```
> plot(data1$lstat,data1$medv,  
      main="Relationship Between lstat and medv",  
      xlab="lstat",  
      ylab="medv")
```



```
> pairs(~medv+lstat+dis,data=data1,  
main="Simple Scatterplot Matrix")
```

# Scatter Plot

Simple Scatterplot Matrix



# Box Plot

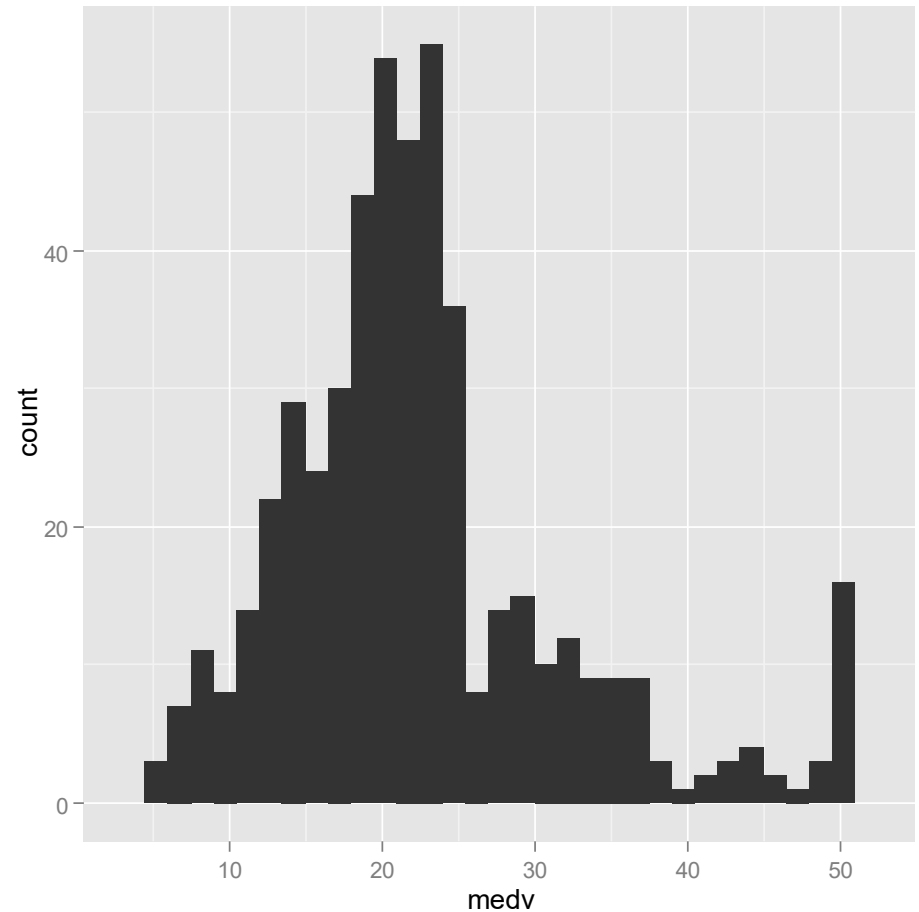
```
> boxplot(medv~chas,data=data1, main="Housing Data",  
xlab="chas", ylab="medv")
```



# Data Visualization (with ggplot2)

# Histogram

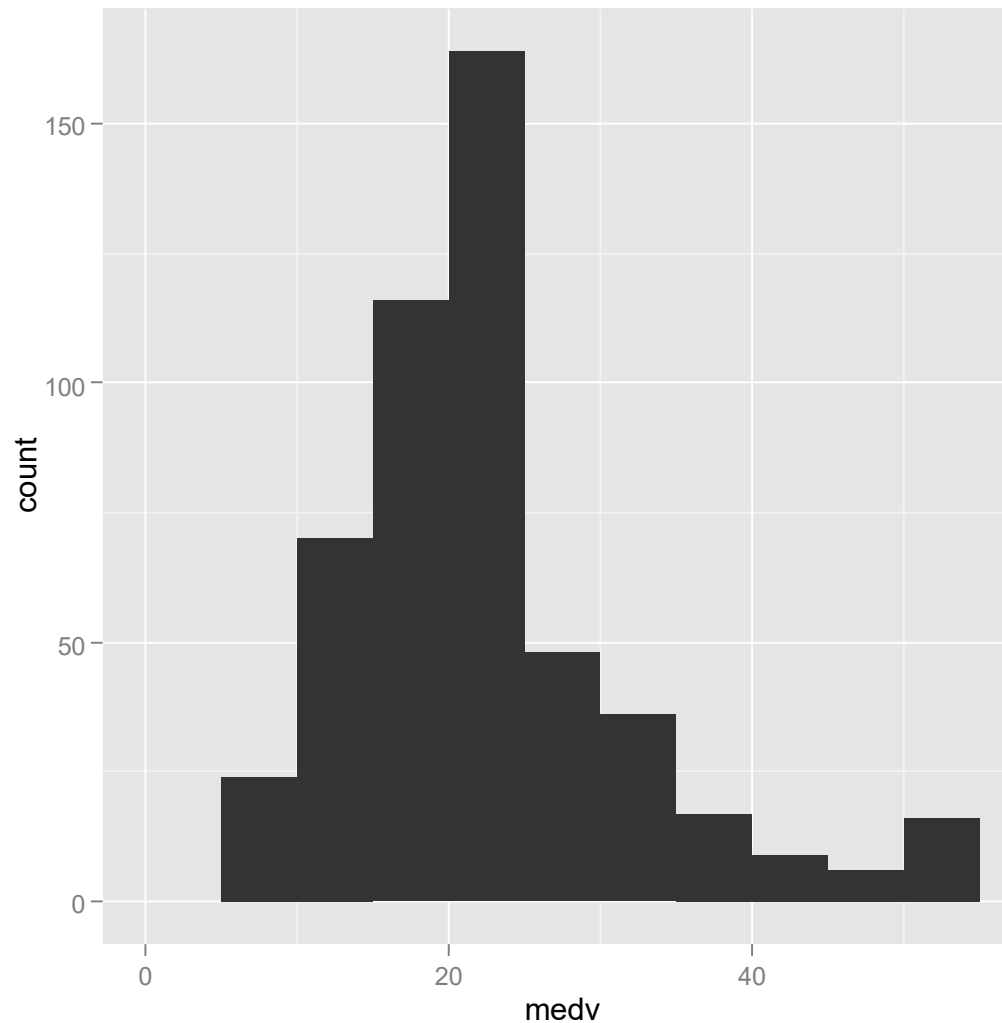
```
> install.packages("ggplot2") # only once on your computer  
> library(ggplot2) # load routines from the ggplot2 library  
> qplot(data = data1, x = medv)
```



##stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.

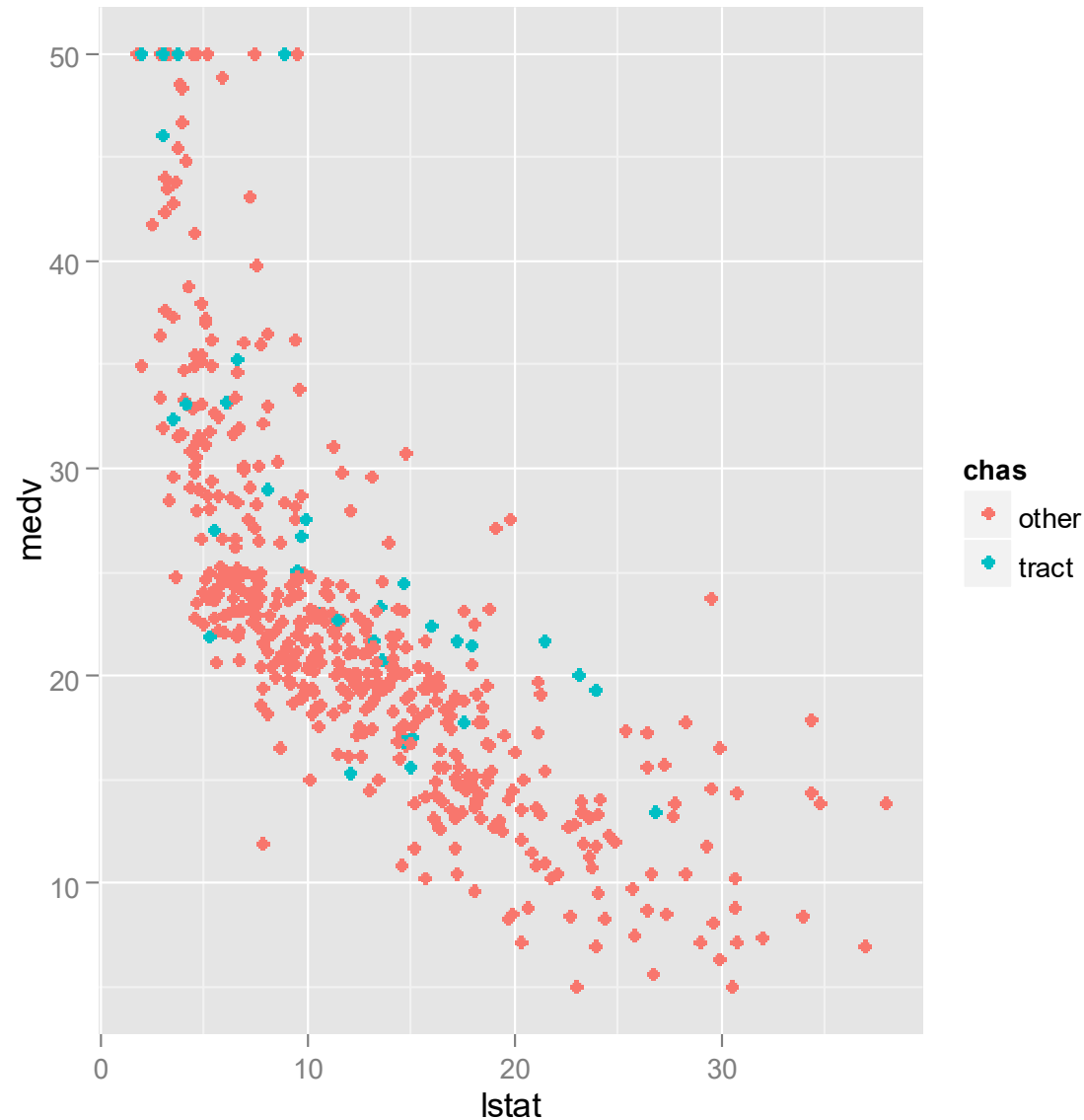
# Histogram (adjusting bin width)

```
> qplot(data = data1, x = medv, binwidth = 5)
```



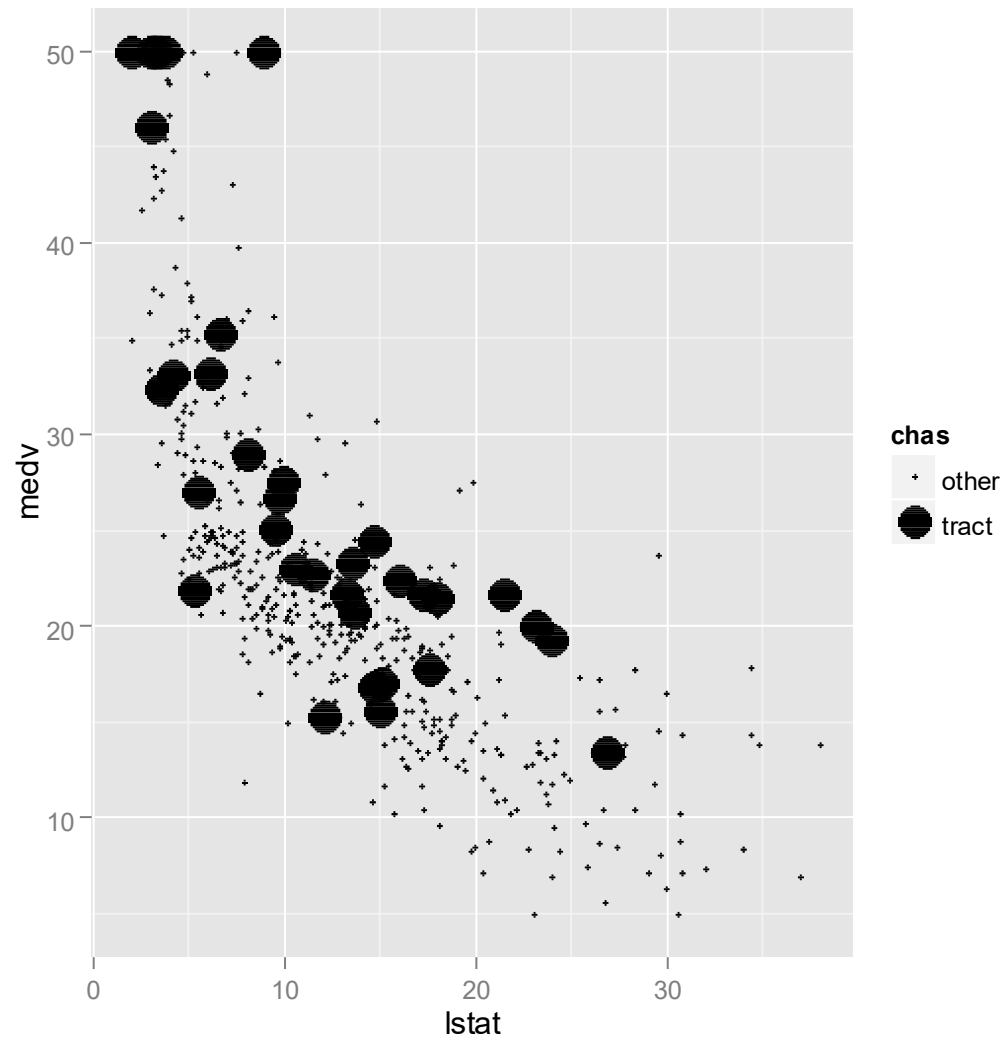
# Scatter Plot

```
> qplot(data = data1, x = lstat, y = medv, color = chas)
```



# Scatter Plot

```
> qplot(data = data1, x = lstat, y = medv, size = chas)
```

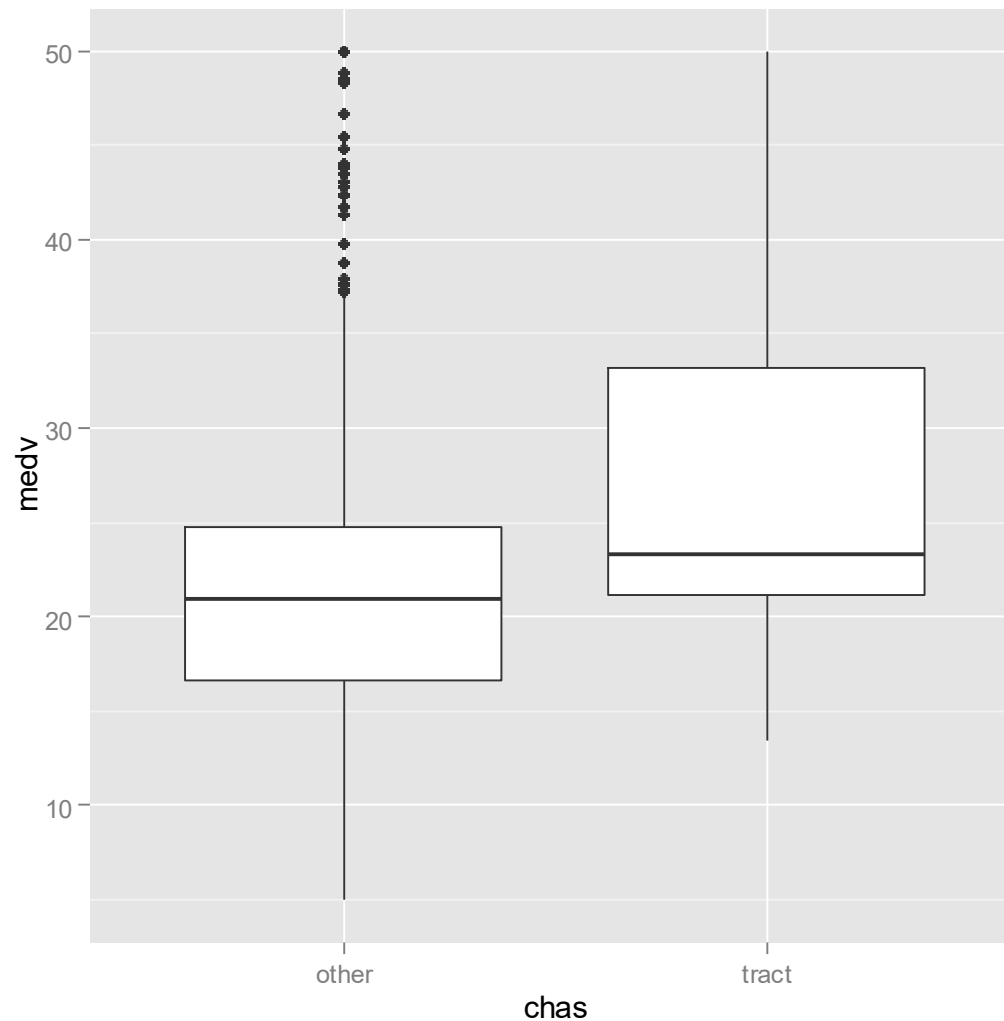


See <https://www.r-graph-gallery.com/320-the-basis-of-bubble-plot.html> for more examples.



# Boxplot

```
> qplot(x = chas, y = medv, data = data1, geom = "boxplot")
```



## Example: Acceptance of Personal Loan

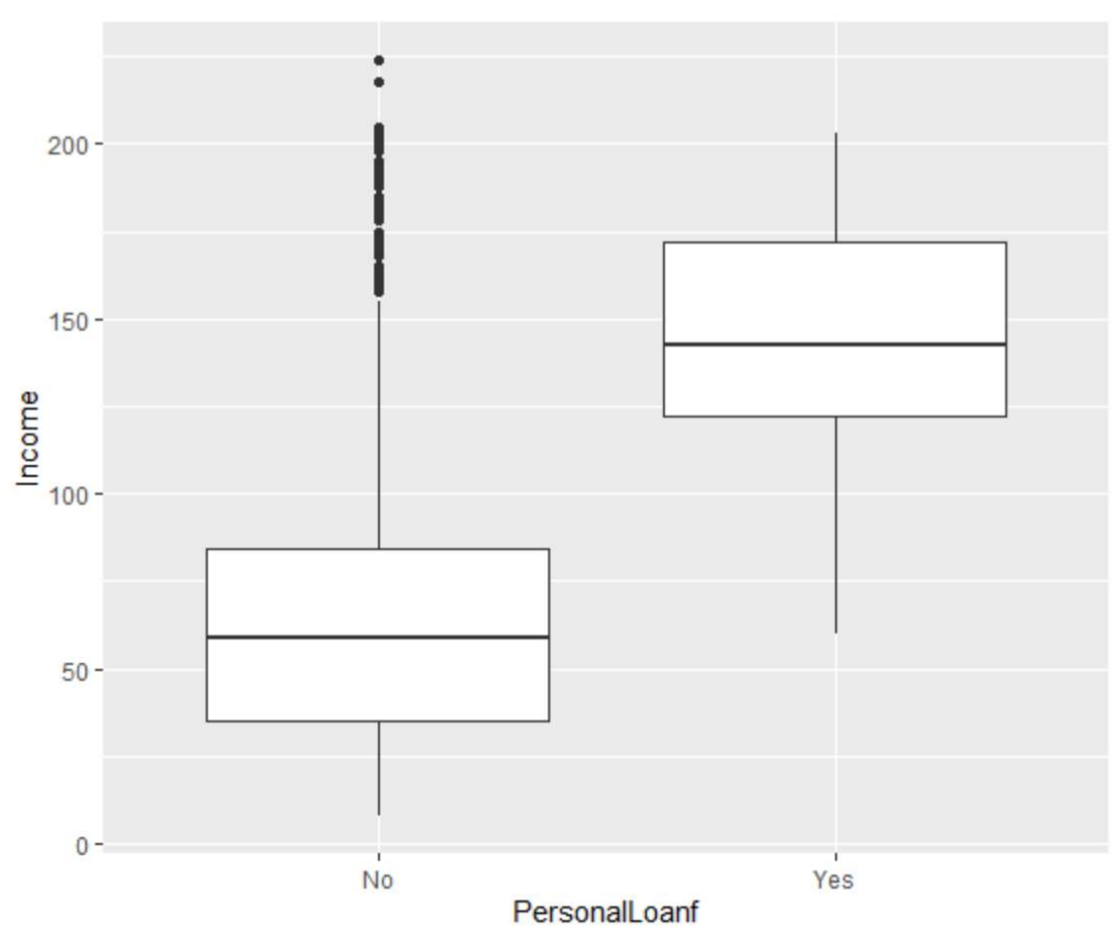
- The bank's dataset includes data on 5000 customers. The data include customer demographic information (age, Income, etc.), customer response to the last personal loan campaign (Personal Loan), and the customer's relationship with the bank (mortgage, securities account, etc.).
- Among there 5000 customers, only 480 (=9.6%) accepted the personal loan that was offered to them in a previous campaign.
- The goal is to find characteristics of customers who are most likely to accept the loan offer in future mailings.

## Predictors and Response

Age	Customer's age in completed years
Experience	#years of professional experience
Income	Annual income of the customer (\$000)
Family	Family size of the customer
CCAvg	Avg. spending on credit cards per month (\$000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. (\$000)
Securitiesf	Does the customer have a securities account with the bank? Yes, No
CD	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by UniversalBank?
PersonalLoanf	Did this customer accept the personal loan offered in the last campaign? Yes, No

Age	Experience	Income	Family	CCAvg	Mortgage	EducProf	EducGrad	Securitiesf	CDf	Onlinef	CreditCardf	PersonalLoanf
25	1	49	4	1.6	0	0	0	Yes	No	No	No	No
45	19	34	3	1.5	0	0	0	Yes	No	No	No	No
39	15	11	1	1	0	0	0	No	No	No	No	No
35	9	100	1	2.7	0	0	1	No	No	No	No	No
35	8	45	4	1	0	0	1	No	No	No	Yes	No
37	13	29	4	0.4	155	0	1	No	No	Yes	No	No
53	27	72	2	1.5	0	0	1	No	No	Yes	No	No
50	24	22	1	0.3	0	1	0	No	No	No	Yes	No
35	10	81	3	0.6	104	0	1	No	No	Yes	No	No
34	9	180	1	8.9	0	1	0	No	No	No	No	Yes
65	39	105	4	2.4	0	1	0	No	No	No	No	No
29	5	45	3	0.1	0	0	1	No	No	Yes	No	No
48	23	114	2	3.8	0	1	0	Yes	No	No	No	No
59	32	40	4	2.5	0	0	1	No	No	Yes	No	No
67	41	112	1	2	0	0	0	Yes	No	No	No	No
60	30	22	1	1.5	0	1	0	No	No	Yes	Yes	No
38	14	130	4	4.7	134	1	0	No	No	No	No	Yes
42	18	81	4	2.4	0	0	0	No	No	No	No	No
46	21	193	2	8.1	0	1	0	No	No	No	No	Yes
55	28	21	1	0.5	0	0	1	Yes	No	No	Yes	No
56	31	25	4	0.9	111	0	1	No	No	Yes	No	No
57	27	63	3	2	0	1	0	No	No	Yes	No	No
29	5	62	1	1.2	260	0	0	No	No	Yes	No	No
44	18	43	2	0.7	163	0	0	Yes	No	No	No	No

```
BankL <- read.csv("C:/MA 299/R/UniversalBankLogisticglm.csv")  
attach(BankL)  
library(ggplot2)  
qplot(x = PersonalLoanf, y = Income, data = BankL, geom = "boxplot")
```



```
counts <- table(CDf, PersonalLoanf)
```

```
counts
```

	PersonalLoanf	
CDf	No	Yes
No	4358	340
Yes	162	140

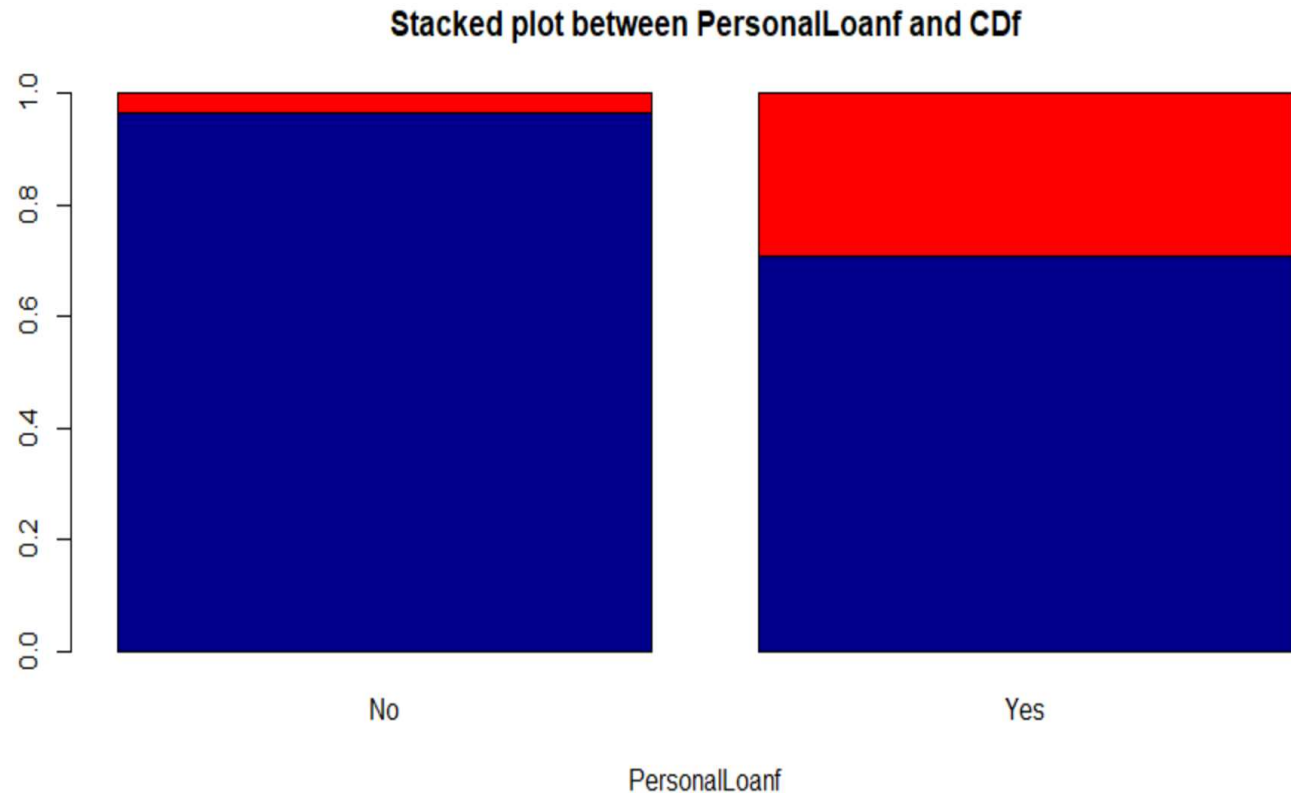
```
mytable<-prop.table(counts, 2)
```

```
mytable
```

	PersonalLoanf	
CDf	No	Yes
No	0.96415929	0.70833333
Yes	0.03584071	0.29166667

# Stacked Plots

```
barplot(mytable, main="Stacked plot between PersonalLoanf  
and CDf", xlab="PersonalLoanf", col=c("darkblue","red"))
```



# Stacked Plots (an alternative)

```
BankL <- read.csv("C:/MA 299/R/UniversalBankLogisticglm.csv")
attach(BankL)
library(ggplot2)
library(ggthemes)
library(extrafont)
library(plyr)
library(scales)
library(reshape2)
counts <- table(CDf, PersonalLoanf)
mytable<-prop.table(counts, 2)
mydata<-as.data.frame(mytable)
plots <- ggplot() + geom_bar(aes(y = Freq, x = PersonalLoanf, fill = CDf),
                             data = mydata, stat="identity")
```

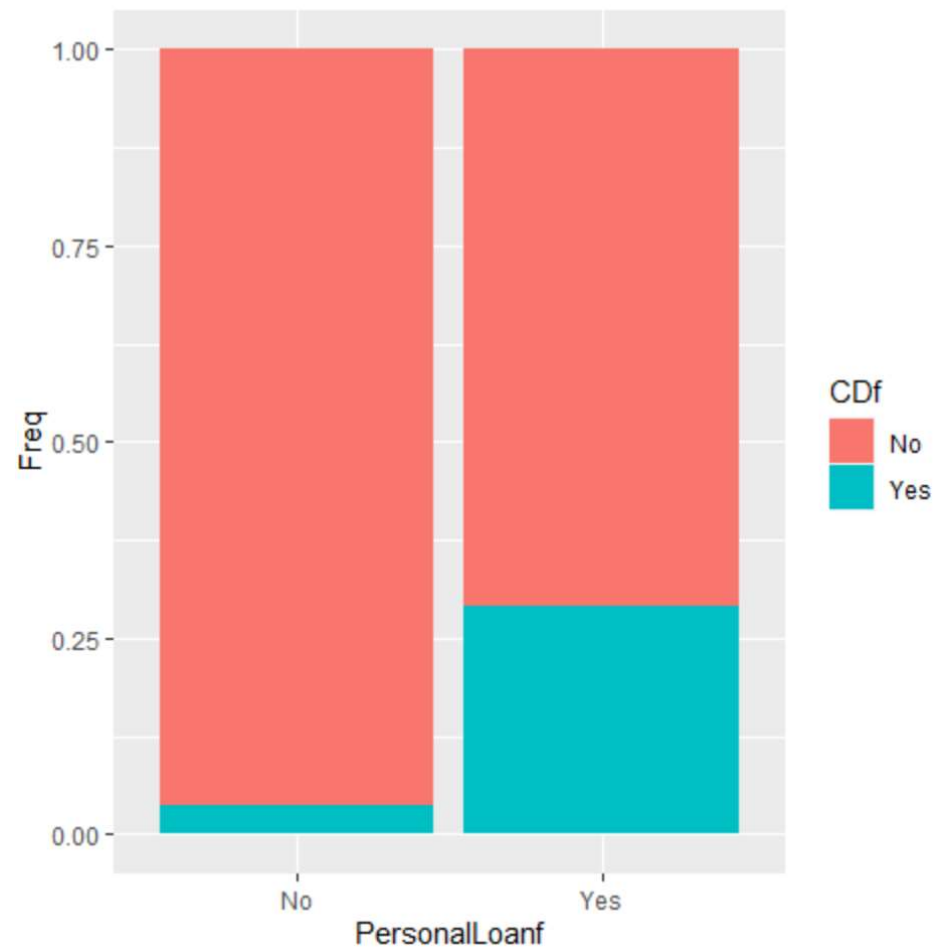


```

> counts
      PersonalLoanf
CDf      No  Yes
No  4358  340
Yes  162  140
> mytable<-prop.table(counts, 2)
> mytable
      PersonalLoanf
CDf      No      Yes
No  0.96415929 0.70833333
Yes 0.03584071 0.29166667
> mydata<-as.data.frame(mytable)
> mydata
  Cdf PersonalLoanf      Freq
1  No              No 0.96415929
2  Yes              No 0.03584071
3  No              Yes 0.70833333
4  Yes              Yes 0.29166667

```

```
> plots <- ggplot() + geom_bar(aes(y = Freq, x = PersonalLoanf, fill = Cdf),  
+                               data = mydata, stat="identity")  
> plots
```



# Data with different coding (BOSTON.HOUSING)

**This dataset contains information on neighborhoods in Boston.**

**Response: the median value of a housing unit in the neighborhood (MEDV)**

## **Variable Information:**

CRIM:	per capita crime rate by town
ZN:	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS:	proportion of non-retail business acres per town
CHAS:	Charles River variable (= 1 if tract bounds river; = 0 if otherwise)
NOX:	nitric oxides concentration (parts per 10 million)
RM:	average number of rooms per dwelling
AGE:	proportion of owner-occupied units built prior to 1940
DIS:	weighted distances to five Boston employment centres
RAD:	index of accessibility to radial highways
TAX:	full-value property-tax rate per \$10,000
PTRATIO:	pupil-teacher ratio by town
LSTAT:	% lower status of the population
MEDV:	Median value of owner-occupied homes in \$1000's
CAT.MEDV:	Is median value of owner-occupied homes in tract above \$30,000 (CAT.MEDV = 1) or not (CAT.MEDV = 0)

**Note: Compare to Boston.Housing.csv, this dataset does not contain variable 'black'.**

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	4.98	24	0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	9.14	21.6	0
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	4.03	34.7	1
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	2.94	33.4	1
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	5.33	36.2	1
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	5.21	28.7	0
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	12.43	22.9	0
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	19.15	27.1	0
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	29.93	16.5	0
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	17.1	18.9	0
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	20.45	15	0
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	13.27	18.9	0
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	15.71	21.7	0
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	8.26	20.4	0
0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	10.26	18.2	0
0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	8.47	19.9	0
1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	6.58	23.1	0
0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	14.67	17.5	0
0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	11.69	20.2	0

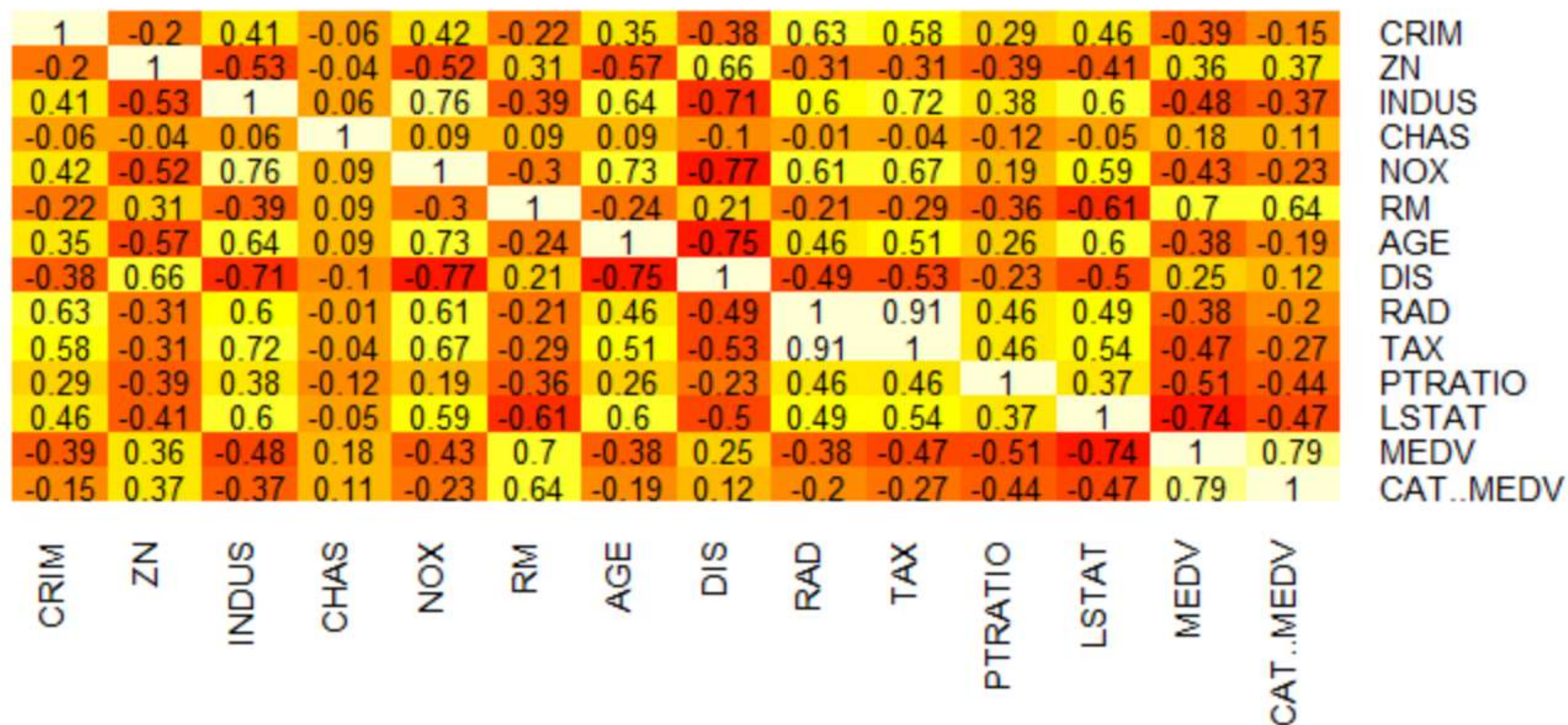
# Heatmaps

A heatmap is a graphical display of numerical data where color is used to denote values.

## ## heatmap with values

```
library(gplots)
```

```
heatmap.2(cor(housing.df), Rowv = FALSE, Colv = FALSE, dendrogram = "none",
  cellnote = round(cor(housing.df),2),
  notecol = "black", key = FALSE, trace = 'none', margins = c(10,10))
```

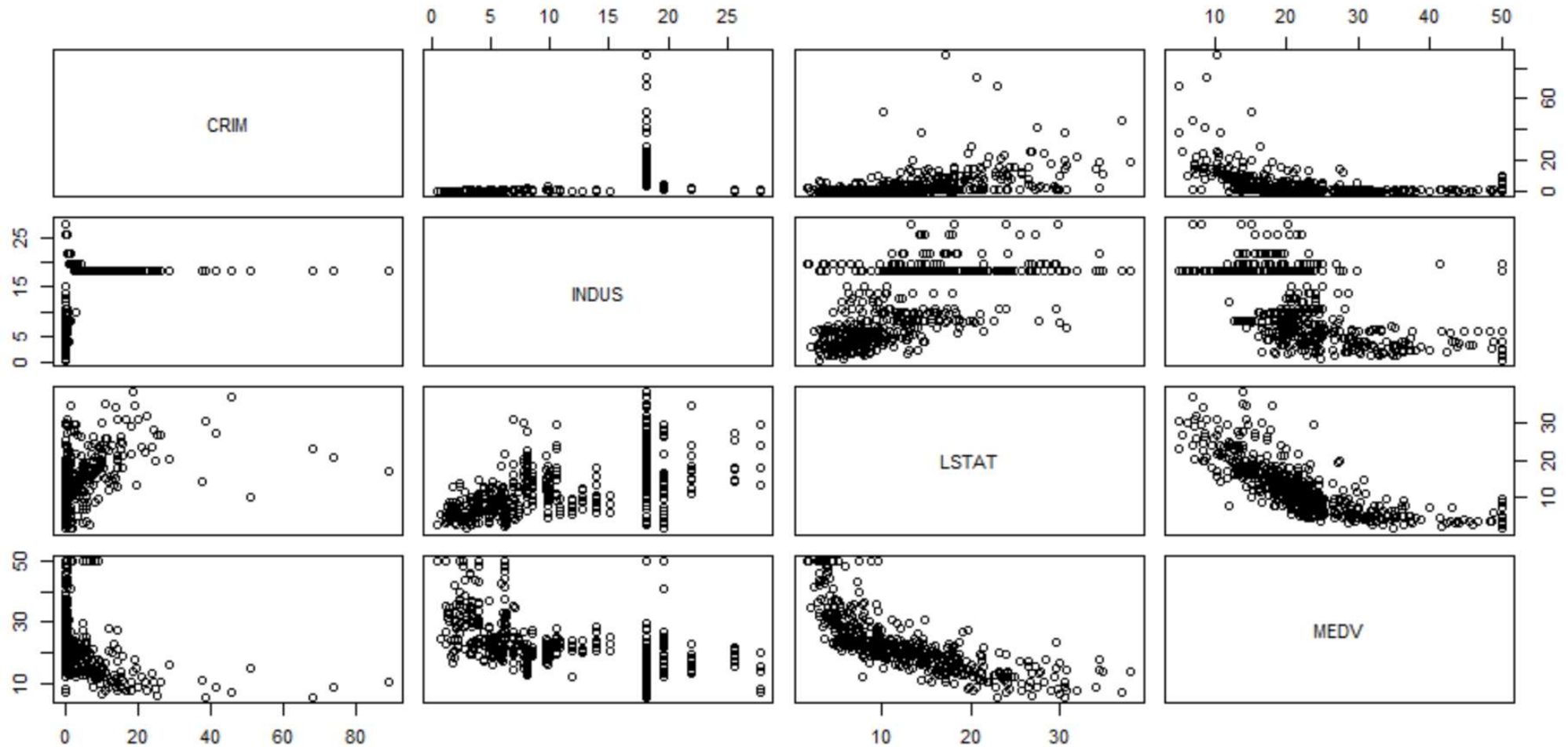


## ## pairwise scatter plots

# use plot() to generate a matrix of 4X4 panels with variable name on the diagonal,

# and scatter plots in the remaining panels.

```
plot(housing.df[, c(1, 3, 12, 13)])
```

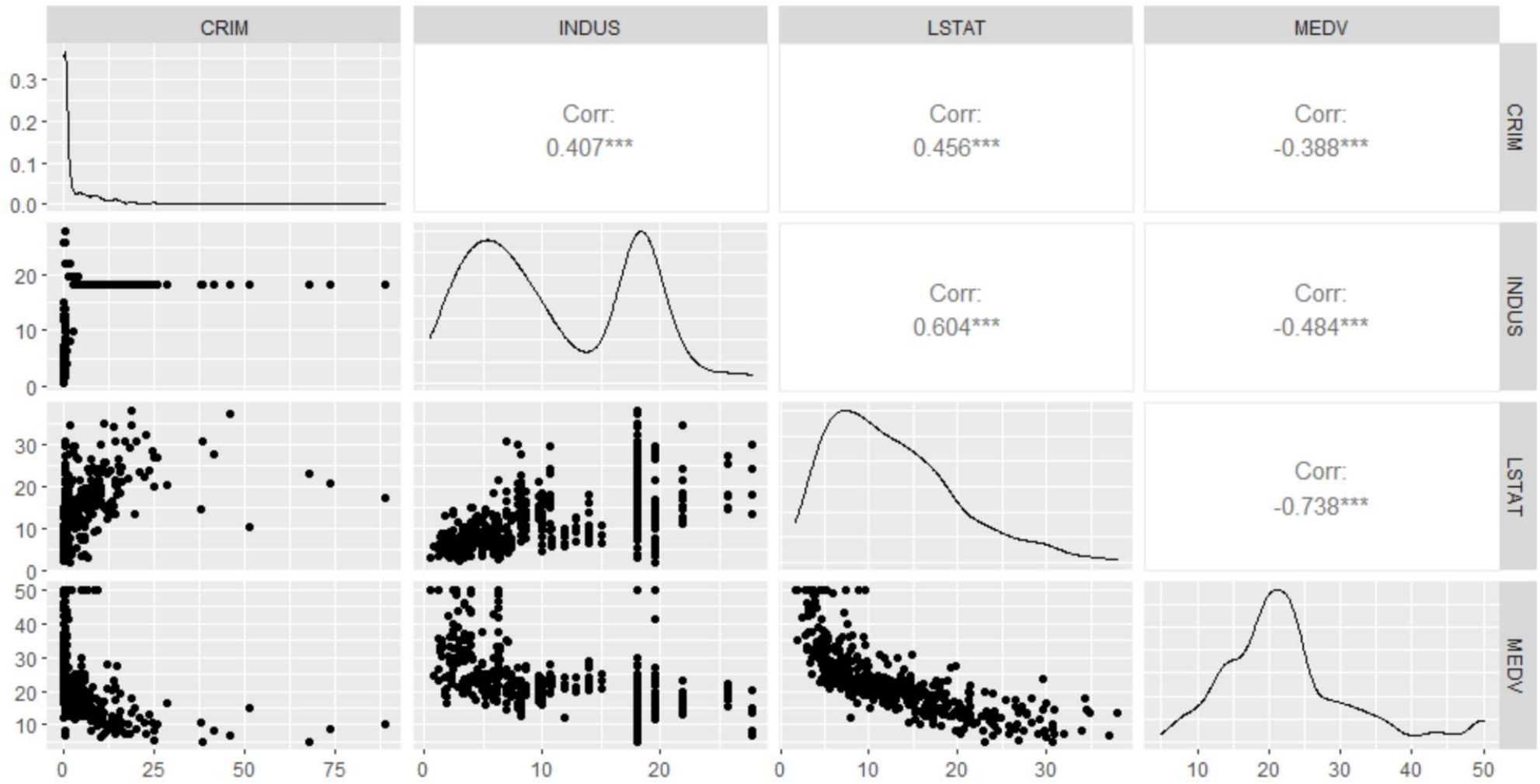




## # alternative plots

```
library(GGally)
```

```
ggpairs(housing.df[, c(1, 3, 12, 13)])
```



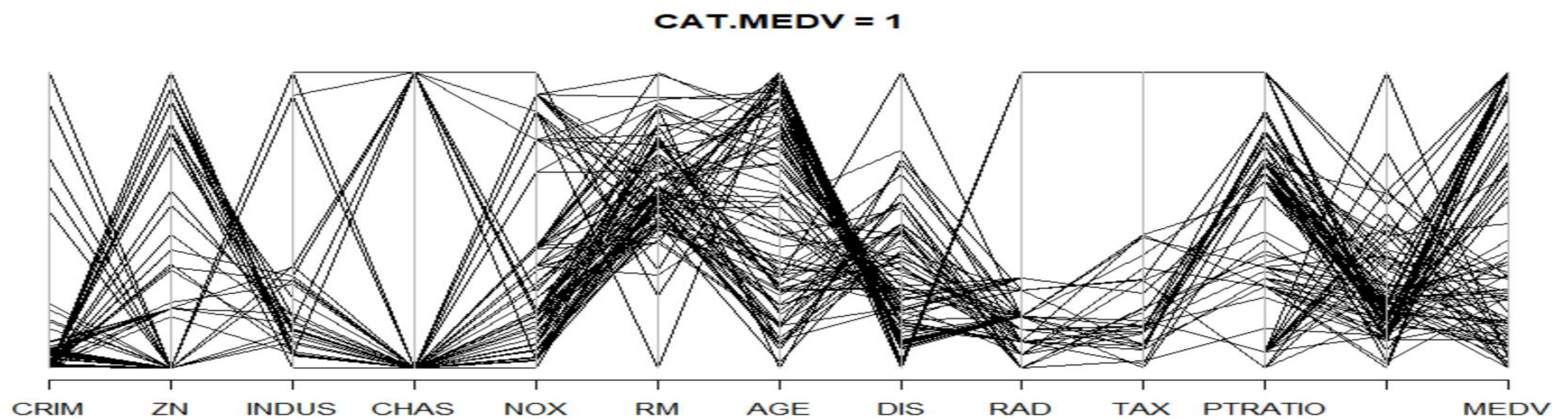
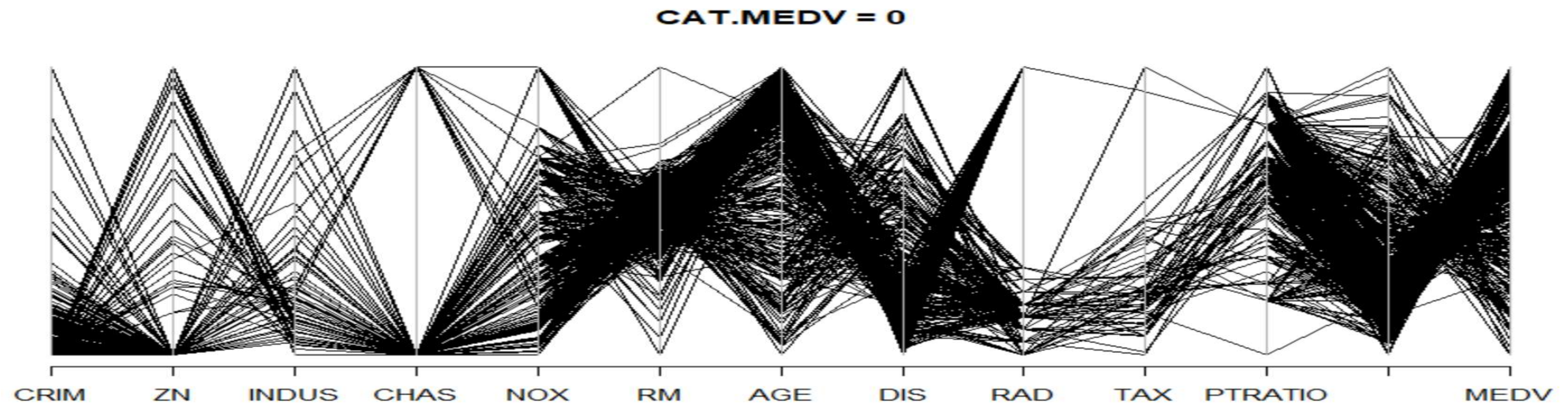


## ## Parallel coordinates plots

```
library(MASS)
```

```
parcoord(housing.df[housing.df$CAT..MEDV == 0, -14], main = "CAT.MEDV = 0")
```

```
parcoord(housing.df[housing.df$CAT..MEDV == 1, -14], main = "CAT.MEDV = 1")
```



# Utility Dataset

Company	Fixed	RoR	Cost	Load	Demand	Sales	Nuclear	Fuel Cost
Arizona	1.06	9.2	151	54.4	1.6	9077	0	0.628
Boston	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
Central	1.43	15.4	113	53	3.4	9212	0	1.058
Commonwealth	1.02	11.2	168	56	0.3	6423	34.3	0.7
Consolidated	1.49	8.8	192	51.2	1	3300	15.6	2.044
Florida	1.32	13.5	111	60	-2.2	11127	22.5	1.241
Hawaiian	1.22	12.2	175	67.6	2.2	7642	0	1.652
Idaho	1.1	9.2	245	57	3.3	13082	0	0.309
Kentucky	1.34	13	168	60.4	7.2	8406	0	0.862
Madison	1.12	12.4	197	53	2.7	6455	39.2	0.623
Nevada	0.75	7.5	173	51.5	6.5	17441	0	0.768
New England	1.13	10.9	178	62	3.7	6154	0	1.897
Northern	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
Oklahoma	1.09	12	96	49.8	1.4	9673	0	0.588
Pacific	0.96	7.6	164	62.2	-0.1	6468	0.9	1.4
Puget	1.16	9.9	252	56	9.2	15991	0	0.62
San Diego	0.76	6.4	136	61.9	9	5714	8.3	1.92
Southern	1.05	12.6	150	56.7	2.7	10140	0	1.108
Texas	1.16	11.7	104	54	-2.1	13507	0	0.636
Wisconsin	1.2	11.8	148	59.9	3.5	7287	41.1	0.702
United	1.04	8.6	204	61	3.5	6650	0	2.116
Virginia	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

### ## scatter plot with labelled points

```
utilities.df <- read.csv("C:/MA 299/R/Utilities.csv")  
plot(utilities.df$Fuel.Cost ~ utilities.df$Sales,  
     xlab = "Sales", ylab = "Fuel Cost", xlim = c(2000, 20000))  
text(x = utilities.df$Sales, y = utilities.df$Fuel.Cost,  
     labels = utilities.df$Company, pos = 4, cex = 0.8, srt = 20, offset = 0.2)
```

