Statistics 202: Statistical Aspects of Data Mining

Professor David Mease

Tuesday, Thursday 9:00-10:15 AM Terman 156

Lecture 5 = More of chapter 3

Agenda:

1) Announce TA office hours
2) Assign chapter 3 homework
3) Lecture over more of chapter 3 (section 3.3)

Announcement:

TA office hours for (almost) the entire semester are posted at

www.stats202.com/ta.html

which is now linked from
www.stats202.com/course_info.html

which is linked from

www.stats202.com

under “Course Information”

Homework Assignment:

Chapter 3 Homework Part 1 is due Tuesday 7/17

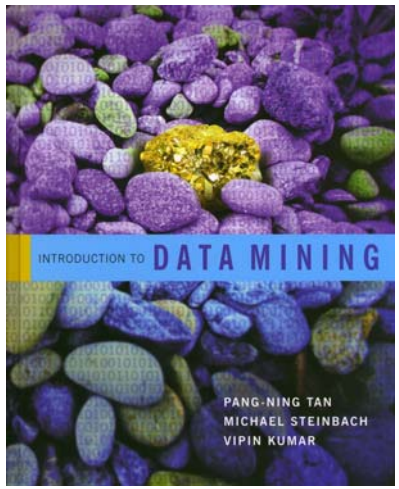
Either email to me (dmease@stanford.edu), bring it to class, or put it under my office door.

SCPD students may use email or fax or mail.

The assignment is posted at
http://www.stats202.com/homework.html

Introduction to Data Mining

by
Tan, Steinbach, Kumar



Chapter 3: Exploring Data

Exploring Data

- We can explore data visually (using tables or graphs) or numerically (using summary statistics)

- Section 3.2 deals with summary statistics

- Section 3.3 deals with visualization

- We will begin with visualization

- Note that many of the techniques you use to explore data are also useful for presenting data

Visualization

- Page 105:

"Data visualization is the display of information in a graphical or tabular format.

Successful visualization requires that the data (information) be converted into a visual format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

The goal of visualization is the interpretation of the visualized information by a person and the formation of a mental model of the information."

Example:

Below are exam scores from a course I taught once.

Describe this data.

192	160	183	136	162
165	181	188	150	163
192	164	184	189	183
181	188	191	190	184
171	177	125	192	149
188	154	151	159	141
171	153	169	168	168
157	160	190	166	150

Note, this data is at
www.stats202.com/exam_scores.csv

The Histogram

- Histogram (Page 111):

"A plot that displays the distribution of values for attributes by dividing the possible values into bins and showing the number of objects that fall into each bin."

- Page 112 – "A *Relative frequency histogram* replaces the count by the relative frequency". These are useful for comparing multiple groups of different sizes.

- The corresponding table is often called the frequency distribution (or relative frequency distribution).

- The function "hist" in R is useful.

In class exercise #7:

Make a frequency histogram in R for the exam scores using bins of width 10 beginning at 120 and ending at 200.

In class exercise #7:

Make a frequency histogram in R for the exam scores using bins of width 10 beginning at 120 and ending at 200.

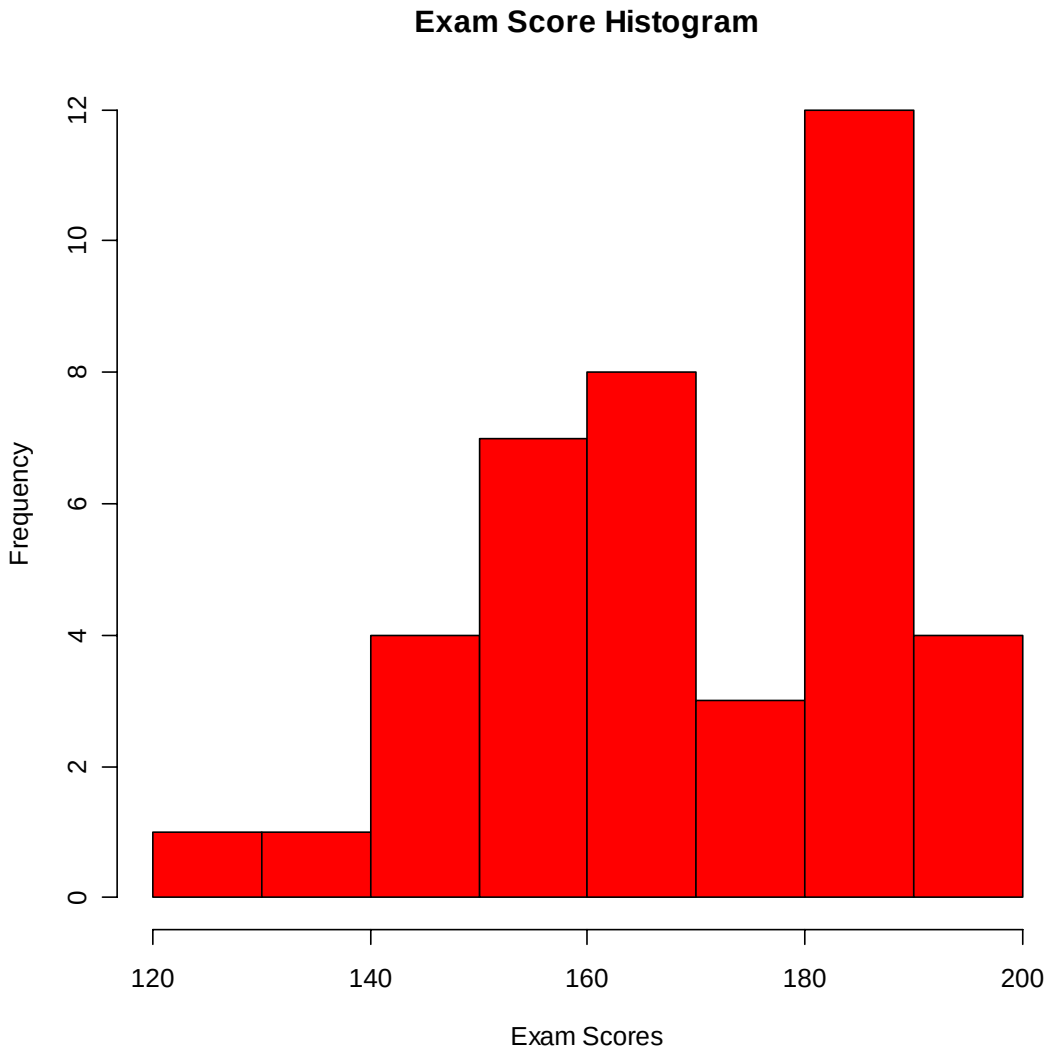
Answer:

```
> exam_scores<-
  read.csv("exam_scores.csv",header=F)


> hist(exam_scores[,1],breaks=seq(120,200,by=10),
    col="red",
    xlab="Exam Scores", ylab="Frequency",
    main="Exam Score Histogram")
```

In class exercise #7:

Make a frequency histogram in R for the exam scores using bins of width 10 beginning at 120 and ending at 200.

Answer:



The (Relative) Frequency Polygon

- Sometimes it is more useful to display the information in a histogram using points connected by lines instead of solid bars.

- Such a plot is called a (relative) frequency polygon.

- This is not in the book.

- The points are placed at the midpoints of the histogram bins and two extra bins with a count of zero are often included at either end for completeness.

In class exercise #8:

Make a frequency polygon in R for the exam scores using bins of width 10 beginning at 120 and ending at 200.

In class exercise #8:

Make a frequency polygon in R for the exam scores using bins of width 10 beginning at 120 and ending at 200.

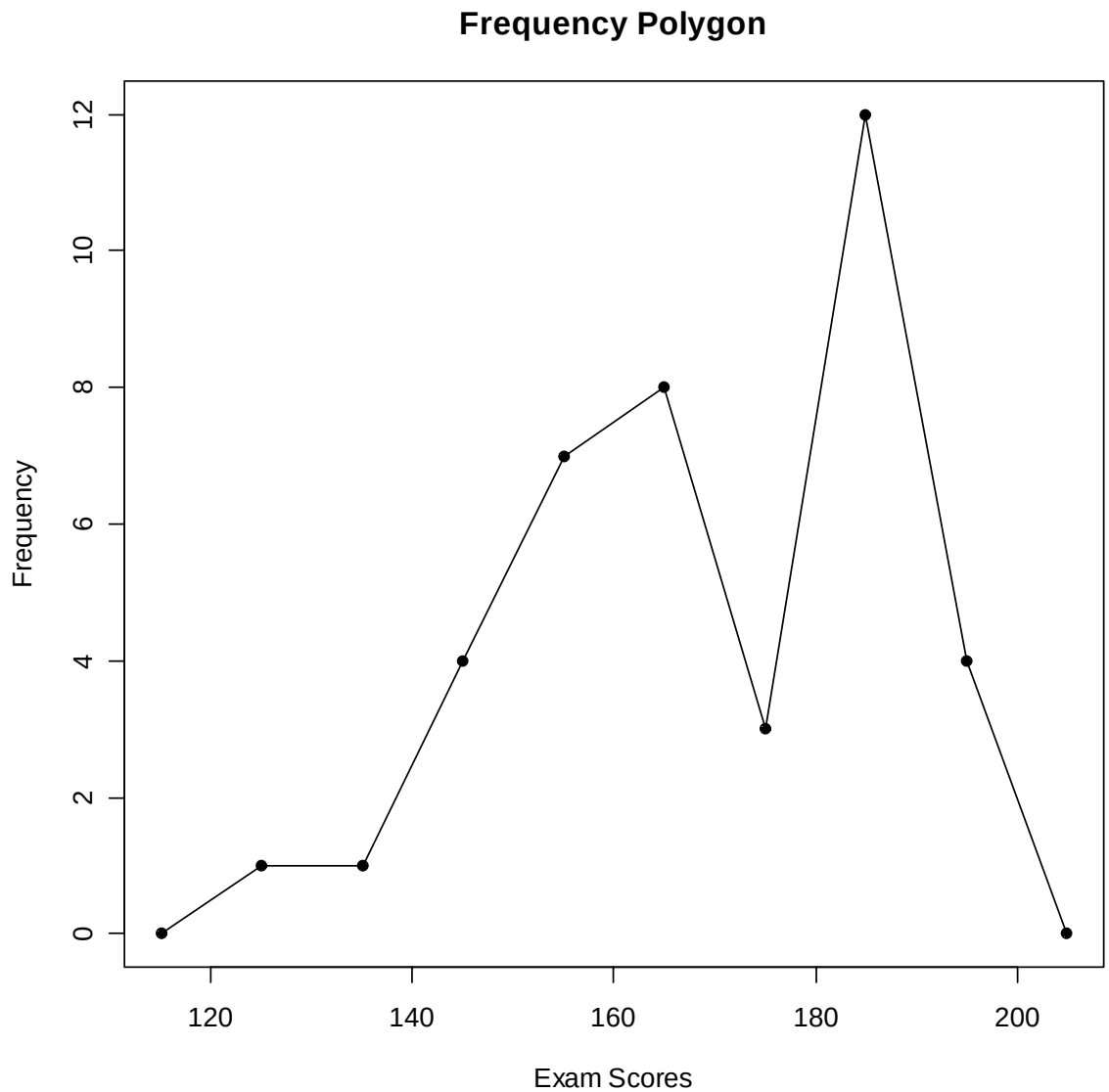
Answer:

```
> my_hist<-hist(exam_scores[,1],
  breaks=seq(120,200,by=10),plot=FALSE)
> counts<-my_hist$counts
> breaks<-my_hist$breaks
> plot(c(115,breaks+5),
  c(0,counts,0),
  pch=19,
  xlab="Exam Scores",
  ylab="Frequency",main="Frequency Polygon")
> lines(c(115,breaks+5),c(0,counts,0))
```

In class exercise #8:

Make a frequency polygon in R for the exam scores using bins of width 10 beginning at 120 and ending at 200.

Answer:



The Empirical Cumulative Distribution Function (Page 115)

- "A *cumulative distribution function* (CDF) shows the probability that a point is less than a value."

- "For each observed value, an *empirical cumulative distribution function* (ECDF) shows the fraction of points that are less than this value." (Page 116)

- A plot of the ECDF is sometimes called an *ogive*.

- The function "ecdf" in R is useful. The plotting features are poorly documented in the help(ecdf) but many examples are given.

In class exercise #9:

Make a plot of the ECDF for the exam scores using the function “ecdf” in R.

In class exercise #9:

Make a plot of the ECDF for the exam scores using the function “ecdf” in R.

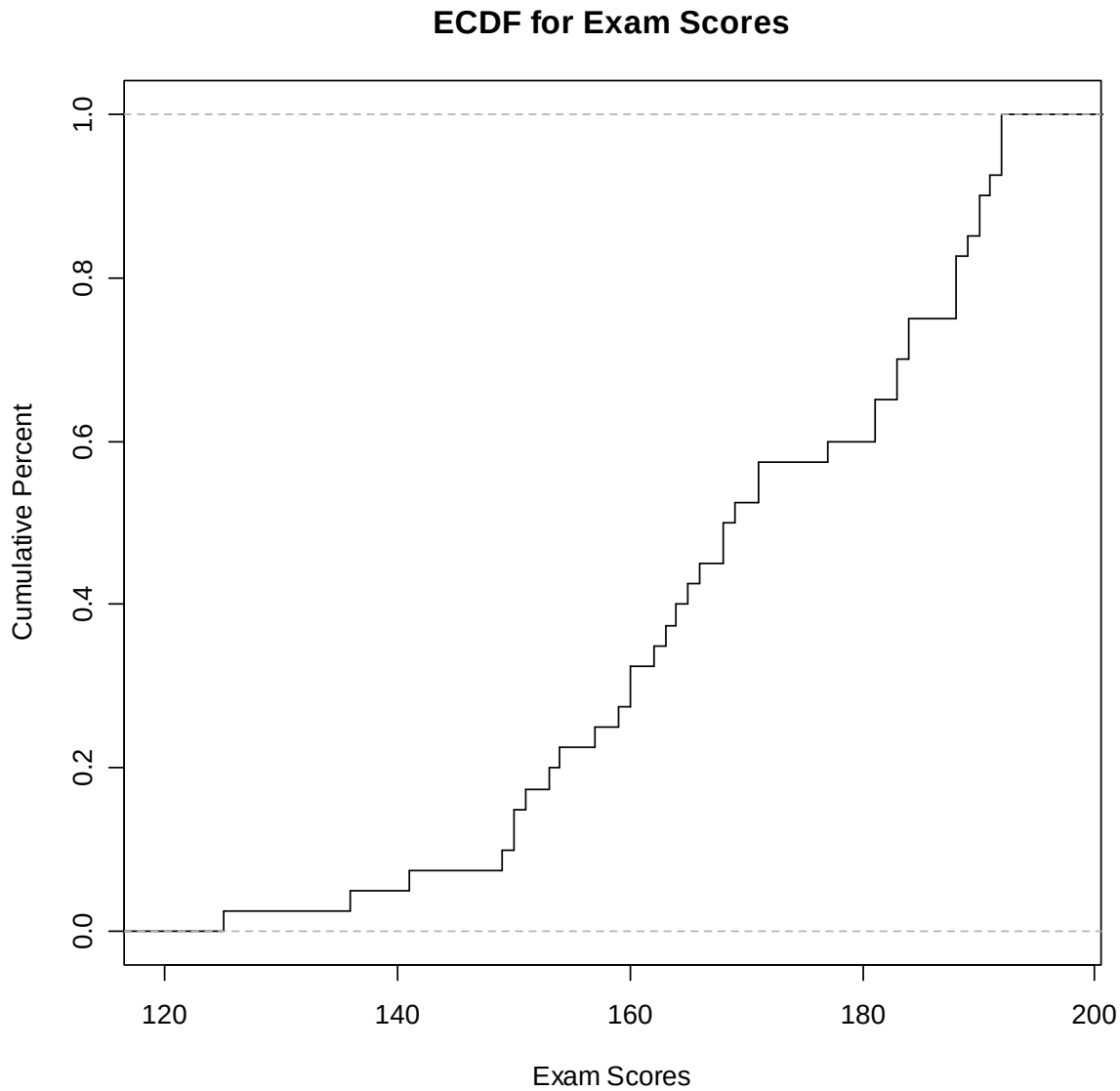
Answer:

```
> plot(ecdf(exam_scores[,1]),
  verticals= TRUE,
  do.p = FALSE,
  main ="ECDF for Exam Scores",
  xlab="Exam Scores",
  ylab="Cumulative Percent")
```

In class exercise #9:

Make a plot of the ECDF for the exam scores using the function "ecdf" in R.

Answer:



Comparing Multiple Distributions

- If there is a second exam also scored out of 200 points, how will I compare the distribution of these scores to the previous exam scores?

187	143	180	100	180
159	162	146	159	173
151	165	184	170	176
163	185	175	171	163
170	102	184	181	145
154	110	165	140	153
182	154	150	152	185
140	132			

- Note, this data is at www.stats202.com/more_exam_scores.csv

Comparing Multiple Distributions

- Histograms can be used, but only if they are relative frequency histograms.

- Relative Frequency Polygons are even better. You can use a different color/type line for each group and add a legend.

- Plots of the ECDF are often even more useful, since they can compare all the percentiles simultaneously. These can also use different color/type lines for each group with a legend.

In class exercise #10:

Plot the relative frequency polygons for both the first and second exams on the same graph. Provide a legend.

In class exercise #10:

Plot the relative frequency polygons for both the first and second exams on the same graph. Provide a legend.

Answer:

```
> more_exam_scores<-
  read.csv("more_exam_scores.csv",header=F)
> my_new_hist<- hist(more_exam_scores[,1],
  breaks=seq(100,200,by=10),plot=FALSE)
> new_counts<-my_new_hist$counts
> new_breaks<-my_new_hist$breaks
> plot(c(95,new_breaks+5),c(0,new_counts/37,0),
  pch=19,xlab="Exam Scores",
  ylab="Relative Frequency",main="Relative
  Frequency Polygons",ylim=c(0,.30))
> lines(c(95,new_breaks+5),c(0,new_counts/37,0),
  lty=2)
```

In class exercise #10:

Plot the relative frequency polygons for both the first and second exams on the same graph. Provide a legend.

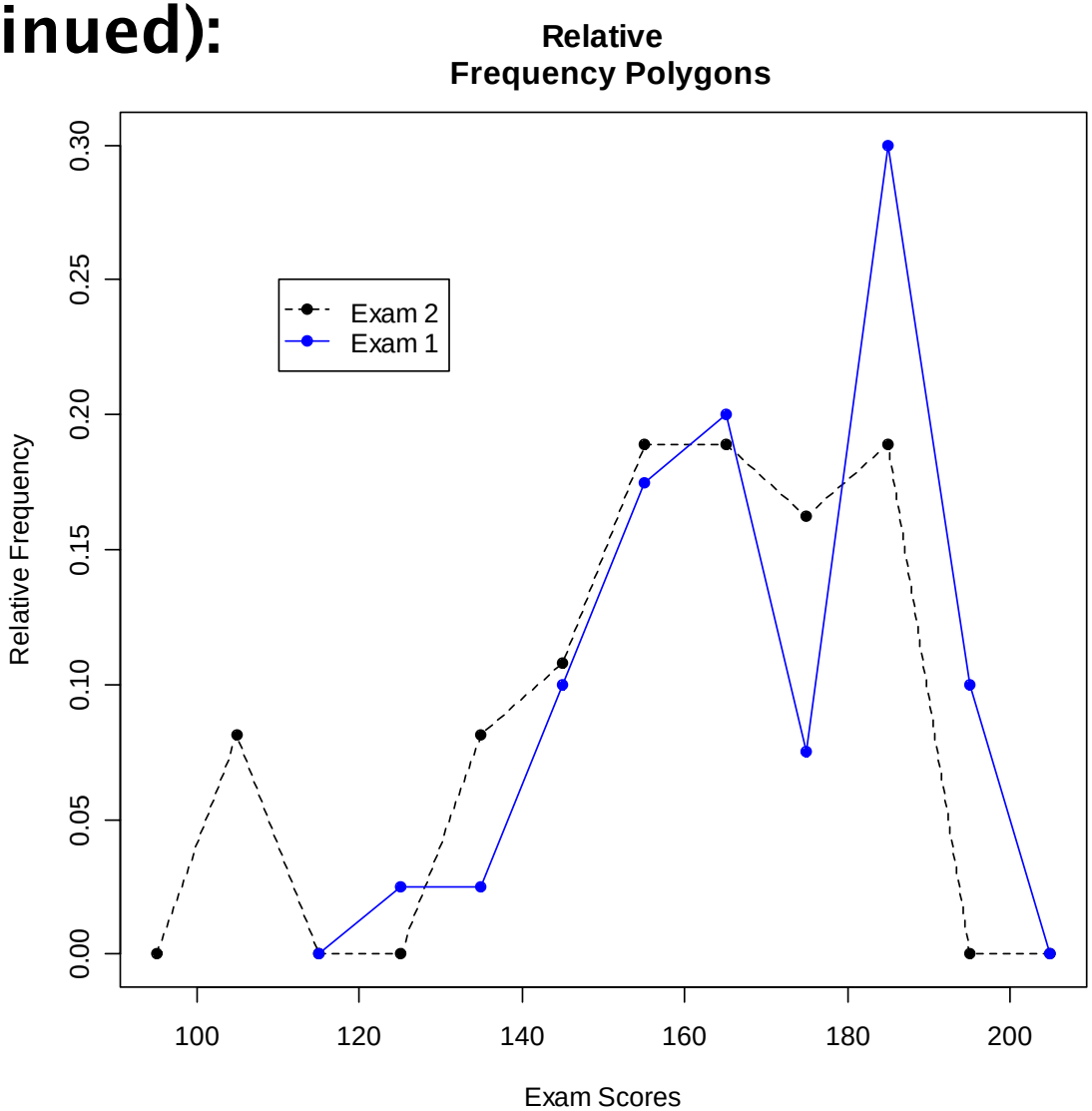
Answer (Continued):

```
> points(c(115,breaks+5),c(0,counts/40,0),
  col="blue",pch=19)

> lines(c(115,breaks+5),c(0,counts/40,0),
  col="blue",lty=1)

> legend(110,.25,c("Exam 2","Exam 1"),
  col=c("black","blue"),lty=c(2,1),pch=19)
```

In class exercise #10:

Plot the relative frequency polygons for both the first and second exams on the same graph. Provide a legend.

Answer (Continued):



In class exercise #11:

Plot the ECDF for both the first and second exams on the same graph. Provide a legend.

In class exercise #11:

Plot the ECDF for both the first and second exams on the same graph. Provide a legend.

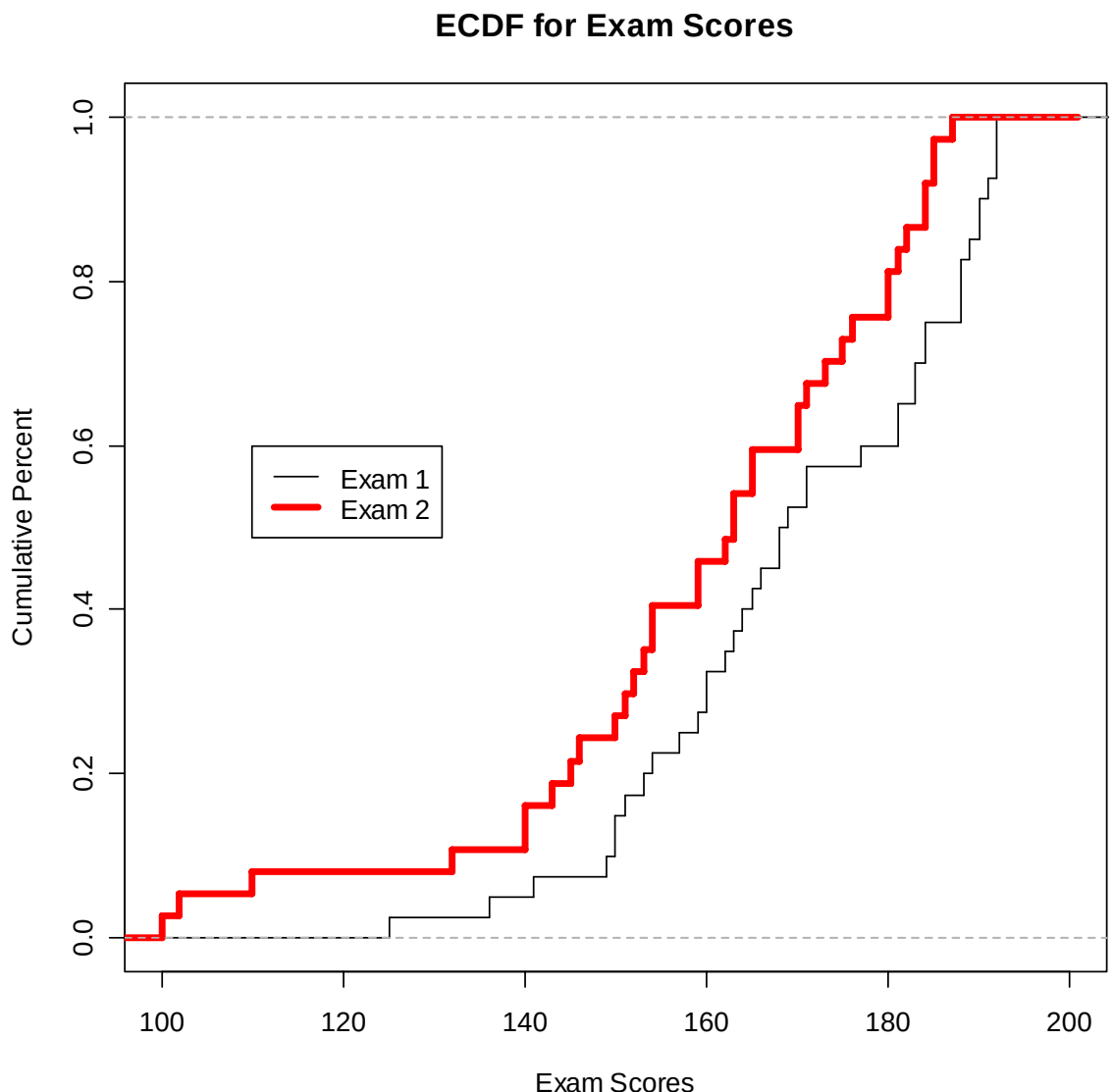
Answer:

```
> plot(ecdf(exam_scores[,1]),
  verticals= TRUE,do.p = FALSE,
  main ="ECDF for Exam Scores",
  xlab="Exam Scores",
  ylab="Cumulative Percent",
  xlim=c(100,200))

> lines(ecdf(more_exam_scores[,1]),
  verticals= TRUE,do.p = FALSE,
  col.h="red",col.v="red",lwd=4)

> legend(110,.6,c("Exam 1","Exam 2"),
  col=c("black","red"),lwd=c(1,4))
```

In class exercise #11:

Plot the ECDF for both the first and second exams on the same graph. Provide a legend.

Answer:



In class exercise #12:

Based on the plot of the ECDF for both the first and second exams from the previous exercise, which exam has lower scores in general? How can you tell from the plot?

Visualizing Paired Numeric Data

- The two sets of exam scores in the previous exercise were not paired. However, the data at www.stats202.com/exams_and_names.csv contains the same exam scores along with an identifier of the student. This data is paired.

- For visualizing paired numeric data, scatter plots (Page 116) are extremely useful. These can be produced using the plot() command in R.

- When the data set has two or more numeric attributes, examining scatter plots of all possible pairs is often useful. The function pairs() in R does this for you. The book calls this a *scatter plot matrix* (Page 116).

In class exercise #13:

Use R to make a scatter plot of the exam scores at www.stats202.com/exams_and_names.csv with the first exam on the x-axis and the second exam on the y-axis. Scale the x-axis and y-axis both from 100 to 200. Add the diagonal line (y=x) to the plot. What does this plot reveal?

In class exercise #13:

Use R to make a scatter plot of the exam scores at www.stats202.com/exams_and_names.csv with the first exam on the x-axis and the second exam on the y-axis. Scale the x-axis and y-axis both from 100 to 200. Add the diagonal line (y=x) to the plot. What does this plot reveal?

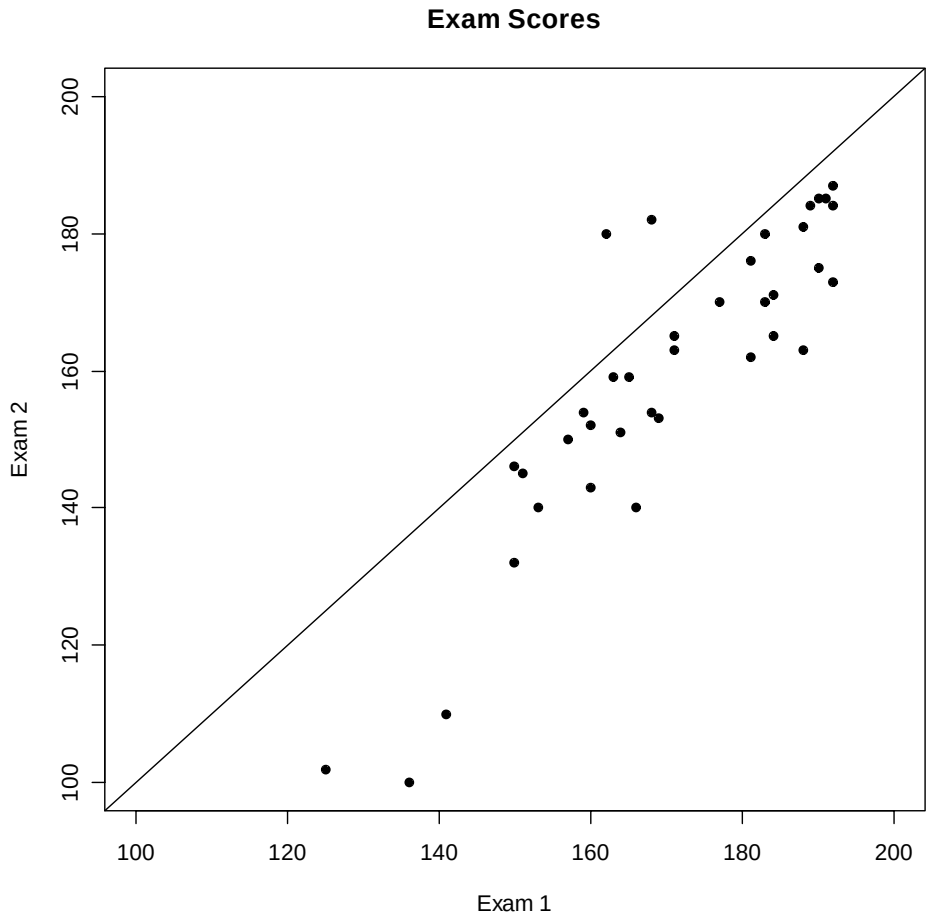
Answer:

```
data<-read.csv("exams_and_names.csv")

plot(data$Exam.1,data$Exam.2,
xlim=c(100,200),ylim=c(100,200),
pch=19,
main="Exam Scores",xlab="Exam 1",ylab="Exam 2")

abline(c(0,1))
```

In class exercise #13:

Use R to make a scatter plot of the exam scores at www.stats202.com/exams_and_names.csv with the first exam on the x-axis and the second exam on the y-axis. Scale the x-axis and y-axis both from 100 to 200. Add the diagonal line (y=x) to the plot. What does this plot reveal?

Answer:



Labeling Points on a Scatter Plot

- The R commands text() and identify() are useful for labeling points on the scatter plot.

In class exercise #14:

Use the text() command in R to label the points for the students who scored lower than 150 on the first exam. Use the identify command to label the points for the two students who did better on the second exam than the first exam. Use the first column in the data set for the labels.

In class exercise #14:

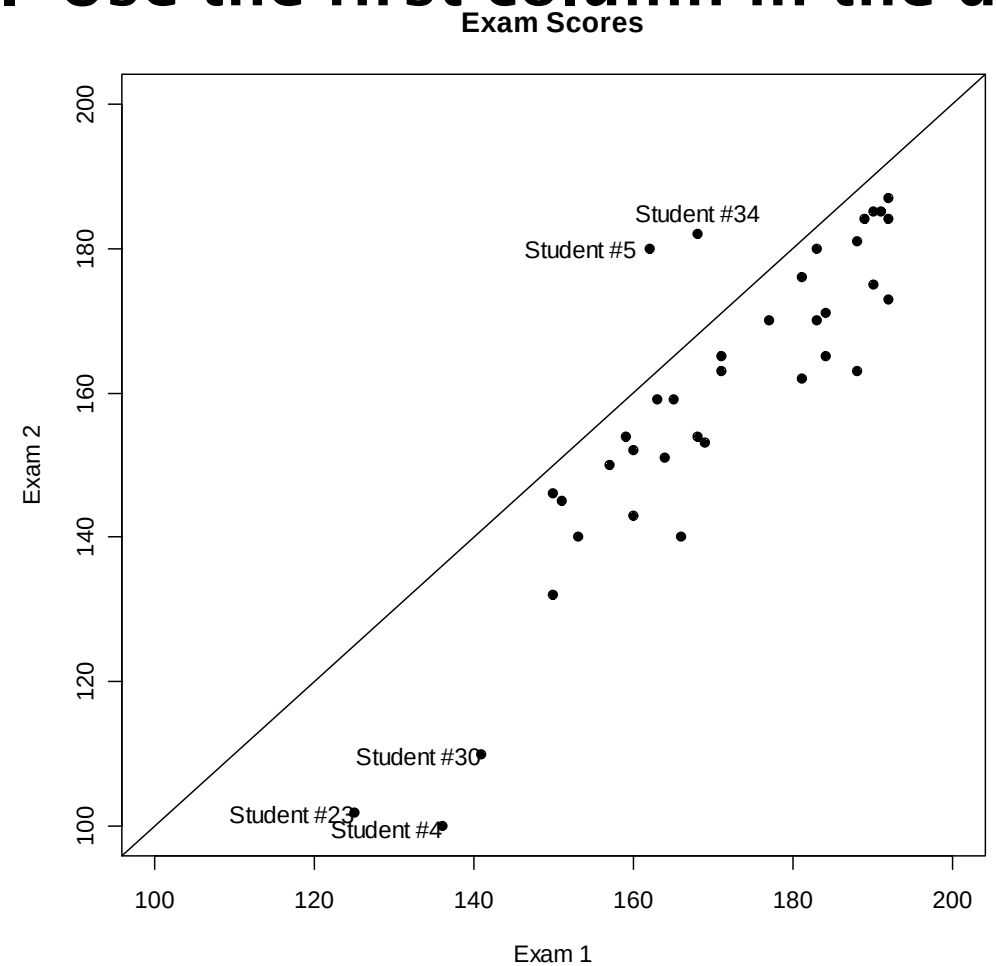
Use the text() command in R to label the points for the students who scored lower than 150 on the first exam. Use the identify command to label the points for the two students who did better on the second exam than the first exam. Use the first column in the data set for the labels.

Answer:

```
text(data$Exam.1[data$Exam.1<150],
     data$Exam.2[data$Exam.1<150],
labels=data$Student[data$Exam.1<150],adj=1)


identify(data$Exam.1,data$Exam.2,
   labels=data$Student)
```

In class exercise #14:

Use the text() command in R to label the points for the students who scored lower than 150 on the first exam. Use the identify command to label the points for the two students who did better on the second exam than the first exam. Use the first column in the data set for the labels.



Adding Noise to a Scatter Plot

- When both variables are discrete, many points in a scatter plot may be plotted over top of one another, which tends to skew the relationship.

- A solution is to add a small amount of noise to the points so that they are jittered a little bit.

- Note: If you have too many points to display cleanly on a scatter plot, sampling may also be helpful.

In class exercise #15:

Add noise uniformly distributed on the interval -0.5 to 0.5 to both the x and y values in the graph in the previous exercise.

In class exercise #15:

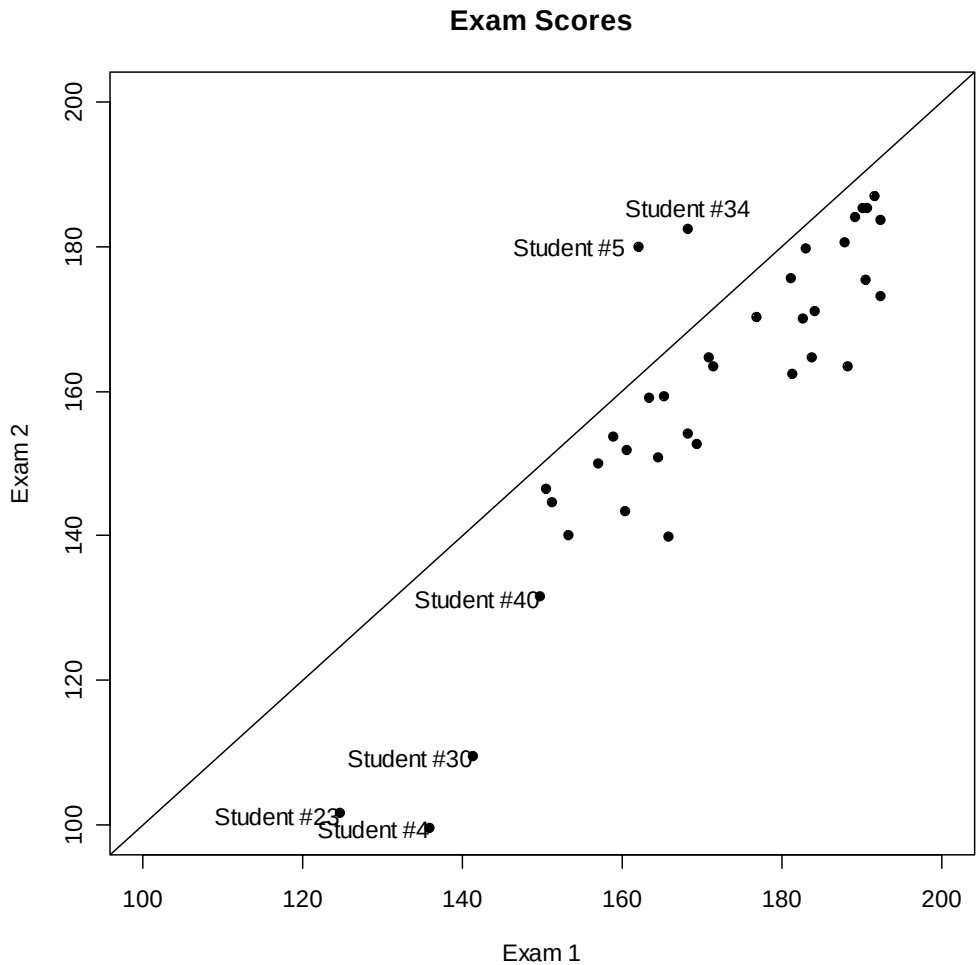
Add noise uniformly distributed on the interval -0.5 to 0.5 to both the x and y values in the graph in the previous exercise.

Answer:

```
data$Exam.1<-data$Exam.1+runif(40)-.5
data$Exam.2<-data$Exam.2+runif(40)-.5

(then same as before)
```

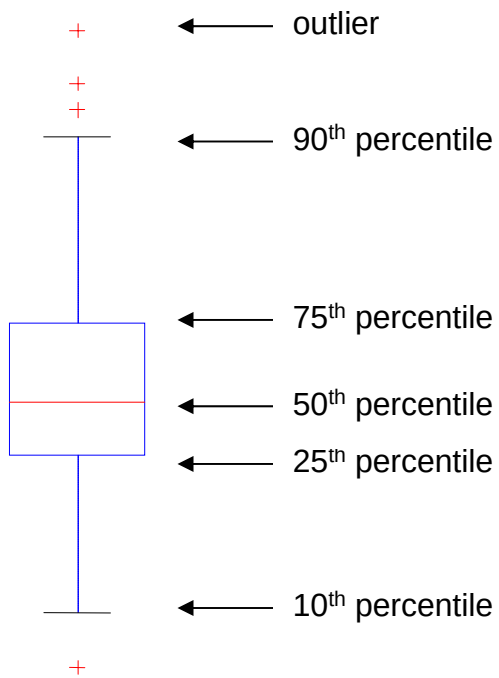
In class exercise #15:

Add noise uniformly distributed on the interval -0.5 to 0.5 to both the x and y values in the graph in the previous exercise.



Boxplots (Pages 114-115)

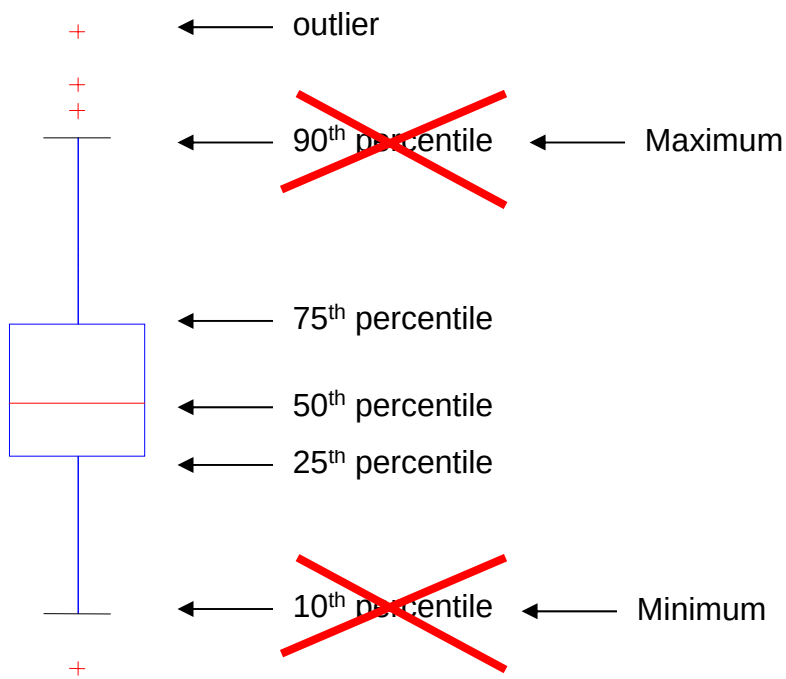
- Invented by J. Tukey
- A simple summary of the distribution of the data
- Boxplots are useful for comparing distributions of multiple attributes or the same attribute for different groups



Boxplots in R

- The function boxplot() in R plots boxplots

- By default, boxplot() in R plots the maximum and the minimum (if they are not outliers) instead of the 10th and 90th percentiles as the book describes



Boxplots (Pages 114-115)

- Boxplots help you visualize the differences in the medians of multiple attributes relative to the variation

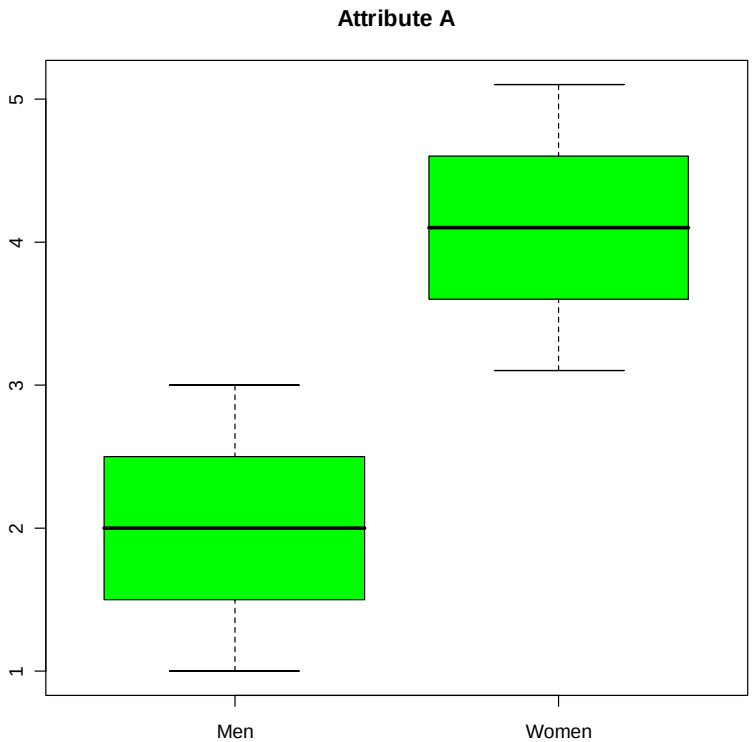
- Example: The median value of Attribute A was 2.0 for men and 4.1 for women. Is this a “big” difference?

Boxplots (Pages 114-115)

- Boxplots help you visualize the differences in the medians of multiple attributes relative to the variation

- Example: The median value of Attribute A was 2.0 for men and 4.1 for women. Is this a "big" difference?

Maybe yes:

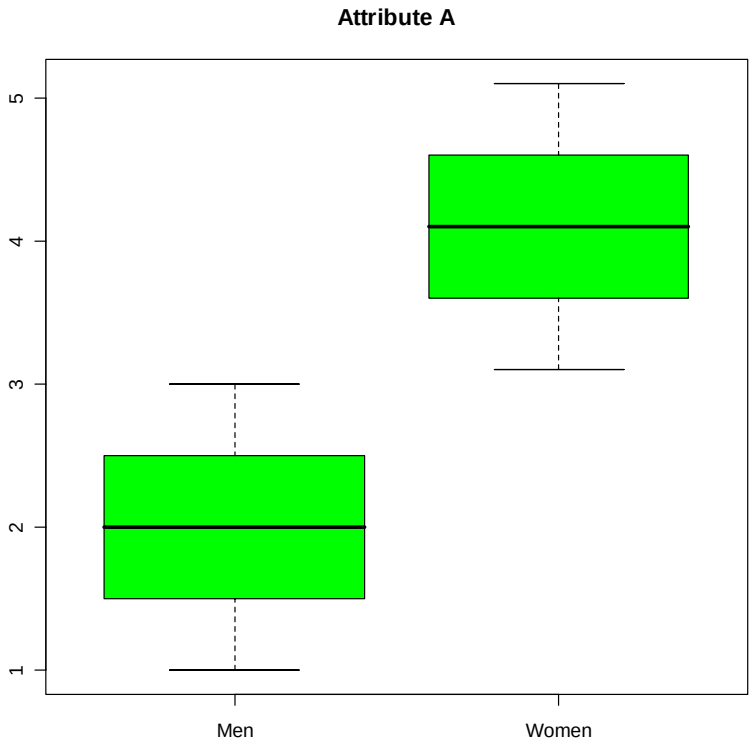


Boxplots (Pages 114-115)

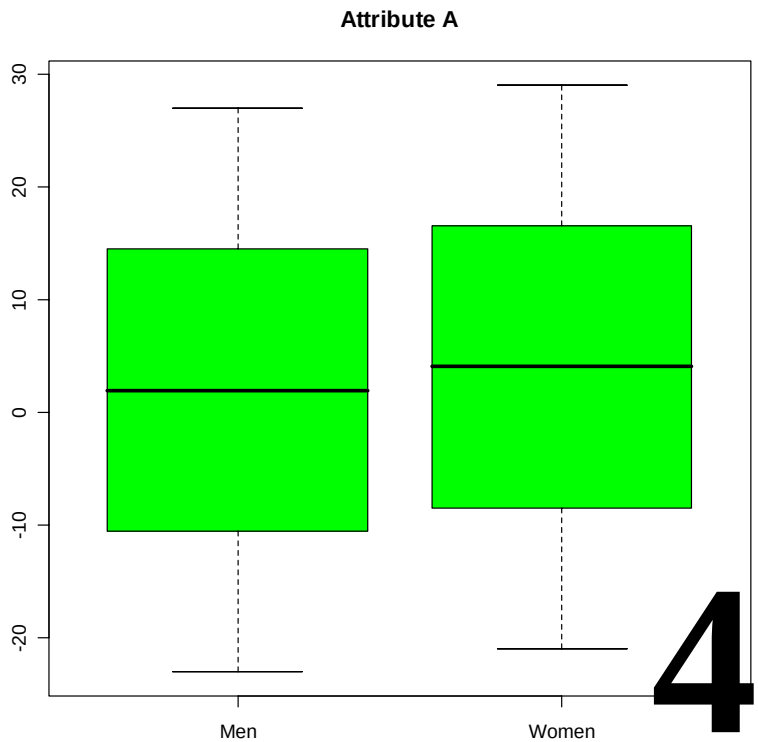
- Boxplots help you visualize the differences in the medians of multiple attributes relative to the variation

- Example: The median value of Attribute A was 2.0 for men and 4.1 for women. Is this a "big" difference?

Maybe yes:



Maybe no:



In class exercise #16:

Use boxplot() in R to make boxplots comparing the first and second exam scores in the data at
www.stats202.com/exams_and_names.csv

In class exercise #16:

Use boxplot() in R to make boxplots comparing the first and second exam scores in the data at www.stats202.com/exams_and_names.csv

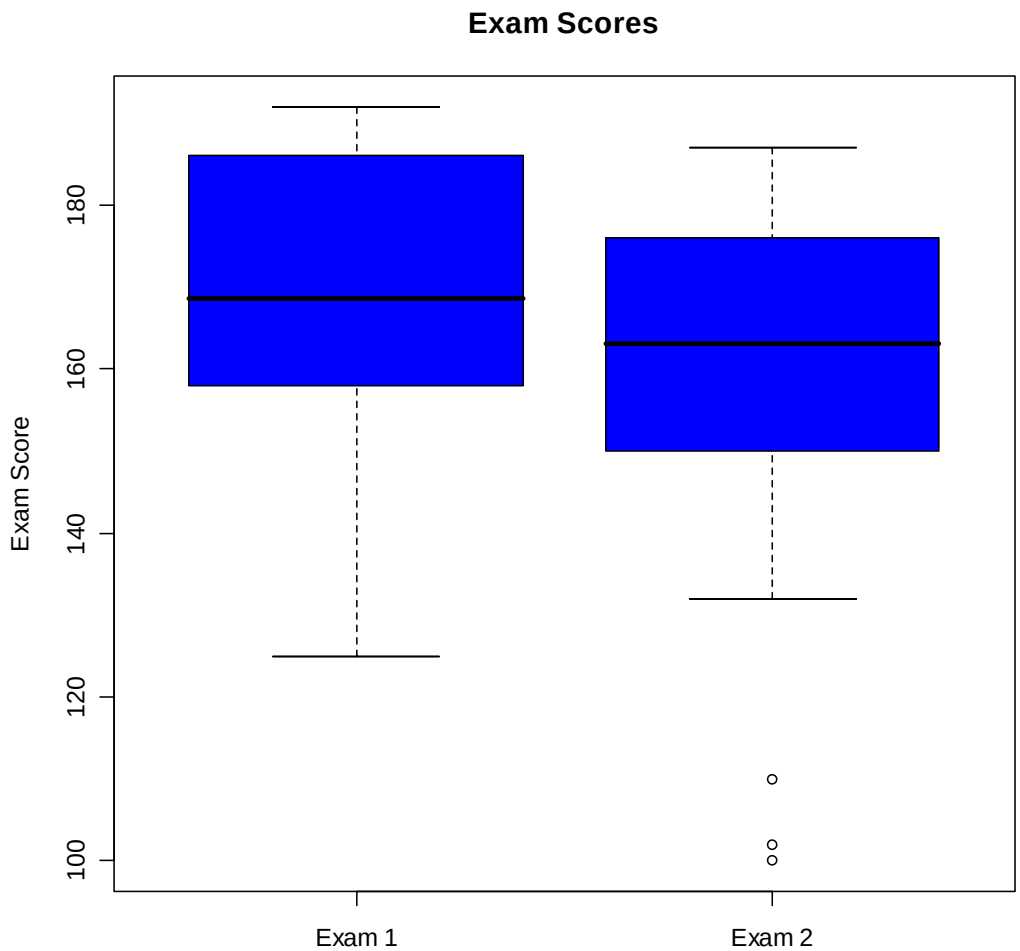
Answer:

```
data<-read.csv("exams_and_names.csv")


boxplot(data[,2],data[,3],col="blue",
main="Exam Scores",
names=c("Exam 1","Exam 2"),ylab="Exam Score")
```

In class exercise #16:

Use boxplot() in R to make boxplots comparing the first and second exam scores in the data at www.stats202.com/exams_and_names.csv

Answer:



Visualization in Excel

- Up until now, we have done all the visualization in R

- Excel also can make many different types of graphs. They are found under the "Insert" menu by selecting "Chart"

- When using Excel to make graphs which anyone will see other than yourself, I strongly encourage you to change defaults such as the grey background.

- Excel also has a nice tool for making tables and associated graphs called "PivotTable and PivotChart Report" under the "Data" menu.

In class exercise #17:

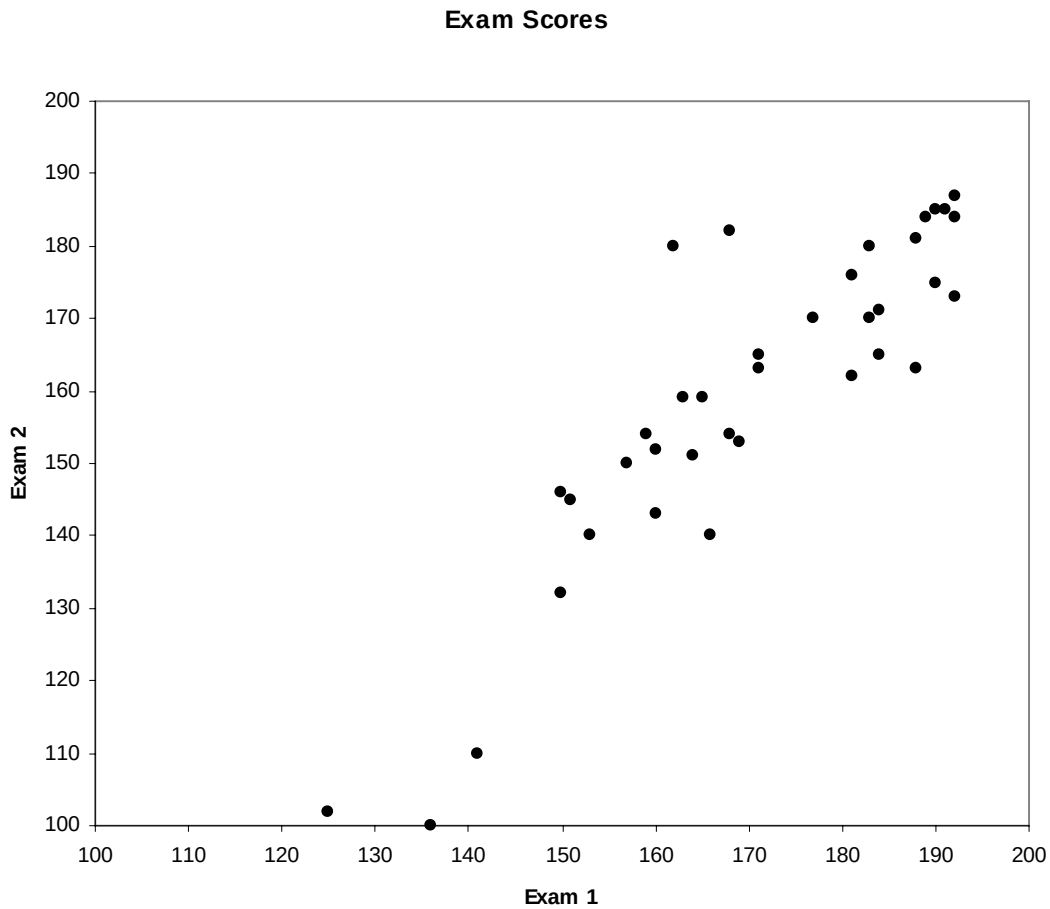
Use “Insert” > “Chart” > “XY Scatter” to make a scatter plot of the exam scores at
www.stats202.com/exams_and_names.csv
Put Exam 1 on the X axis and Exam 2 on the Y axis.

In class exercise #17:

Use "Insert" > "Chart" > "XY Scatter" to make a scatter plot of the exam scores at www.stats202.com/exams_and_names.csv

Put Exam 1 on the X axis and Exam 2 on the Y axis.

Answer:



In class exercise #18:

The data www.stats202.com/more_stats202_logs.txt contains access logs from May 7, 2007 to July 1, 2007. Use "Data" > "PivotTable and PivotChart Report" In Excel to make a table with the counts of
GET /lecture2=start-chapter-2.ppt HTTP/1.1
and
GET /lecture2=start-chapter-2.pdf HTTP/1.1
for each date. Which is more popular?

In class exercise #18:

The data www.stats202.com/more_stats202_logs.txt contains access logs from May 7, 2007 to July 1, 2007. Use "Data" > "PivotTable and PivotChart Report" In Excel to make a table with the counts of
GET /lecture2=start-chapter-2.ppt HTTP/1.1
and
GET /lecture2=start-chapter-2.pdf HTTP/1.1
for each date. Which is more popular?

Answer:

Date	GET /lecture2=start-chapter-2.pdf HTTP/1.1	GET /lecture2=start-chapter-2.ppt HTTP/1.1	Grand Total
27-Jun-07	150	17	167
28-Jun-07	247	29	276
29-Jun-07	253	53	306
30-Jun-07	77	9	86
1-Jul-07	50	7	57
Grand Total	777	115	892

In class exercise #19:

The data www.stats202.com/more_stats202_logs.txt contains access logs from May 7, 2007 to July 1, 2007. Use “Data” > “PivotTable and PivotChart Report” In Excel to make a table with the counts of the rows for each date in May.

In class exercise #19:

The data www.stats202.com/more_stats202_logs.txt contains access logs from May 7, 2007 to July 1, 2007. Use "Data" > "PivotTable and PivotChart Report" In Excel to make a table with the counts of the rows for each date in May.

Answer:

Date	Count
May-7	88
May-8	88
May-9	65
May-10	179
May-11	47
May-12	67
May-13	47
May-14	59
May-15	58
May-16	107
May-17	64
May-18	93
May-19	66
May-20	104
May-21	123
May-22	75
May-23	85
May-24	81
May-25	49
May-26	60
May-27	78
May-28	66
May-29	64
May-30	69
May-31	46

In class exercise #20:

Use “Insert” > “Chart” > “Line” In Excel to make a graph on the number of rows versus the date for the previous exercise.

In class exercise #20:

Use “Insert” > “Chart” > “Line” In Excel to make a graph on the number of rows versus the date for the previous exercise.

Answer: