Expository graphs

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Why do we use graphs in data analysis?

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To "debug" analyses
- To communicate results

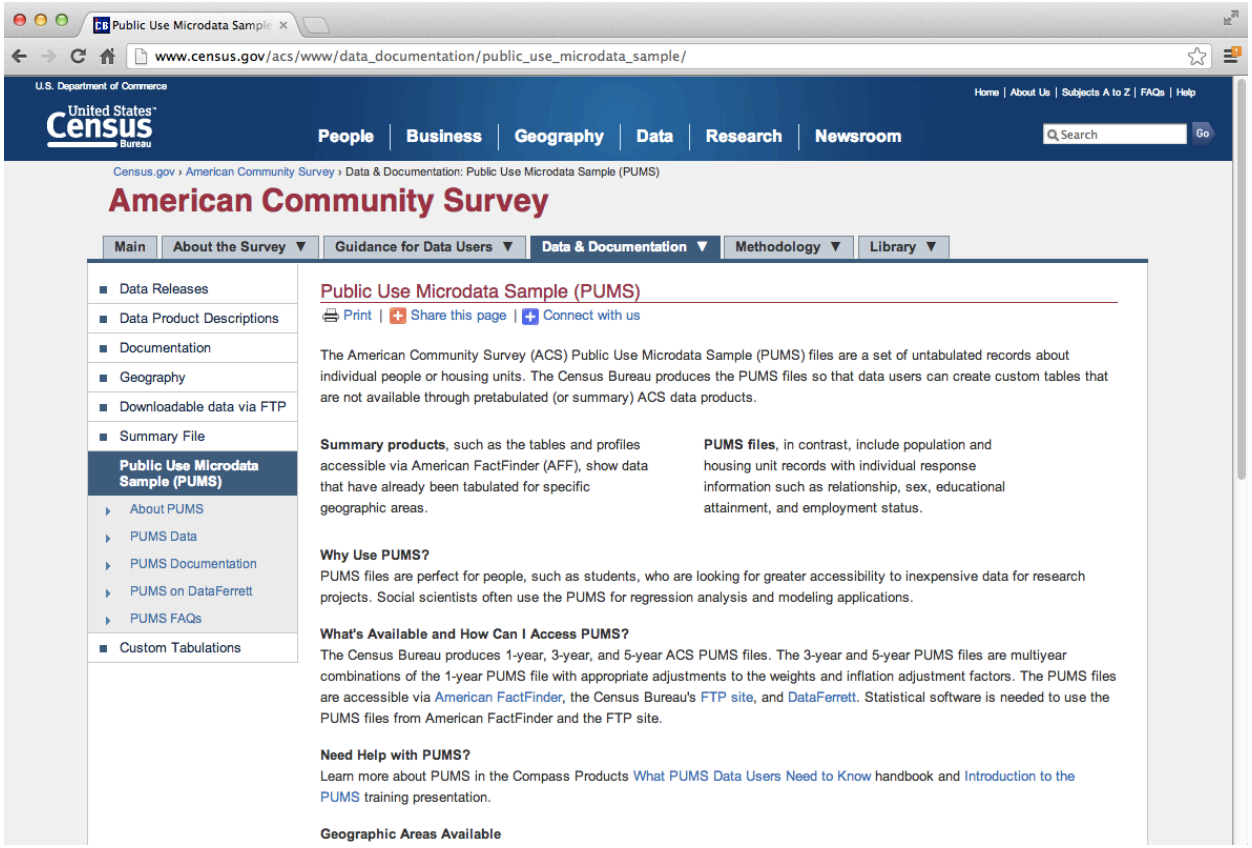
Expository graphs

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To "debug" analyses
- To communicate results

Characteristics of expository graphs

- The goal is to communicate information
- Information density is generally good
- Color/size are used both for aesthetics and communication
- Expository figures have understandable axes, titles, and legends

Housing data



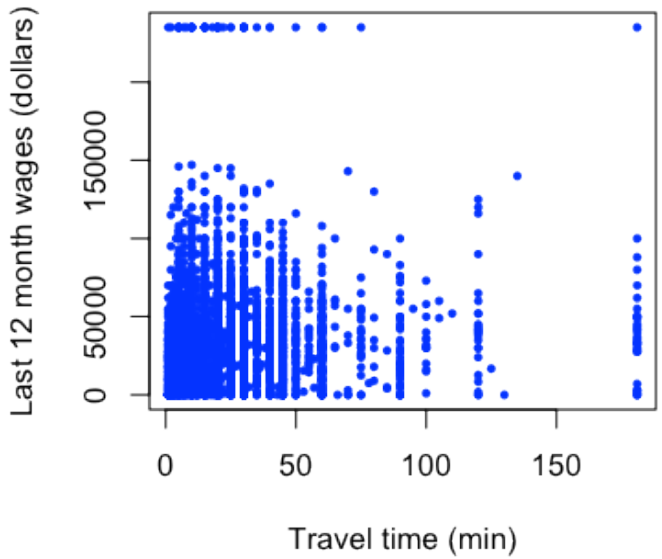

```
pData <- read.csv("./data/ss06pid.csv")
```

5/21

Axes

Important parameters: *xlab*,*ylab*,*cex.lab*,*cex.axis*

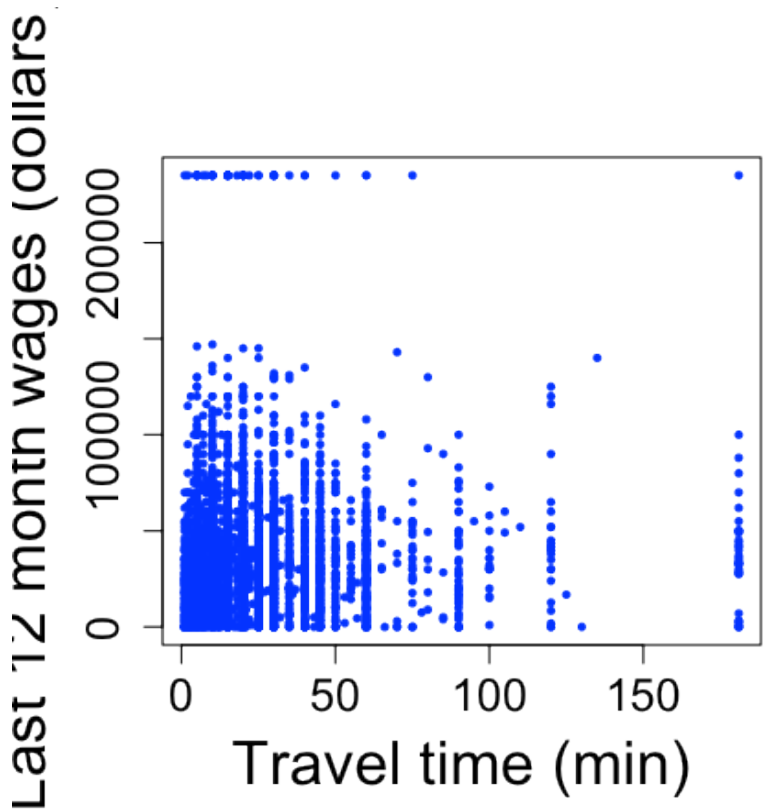
```
plot(pData$JWMNP,pData$WAGP,pch=19,col="blue",cex=0.5,
     xlab="Travel time (min)",ylab="Last 12 month wages (dollars)")
```



6/21

Axes

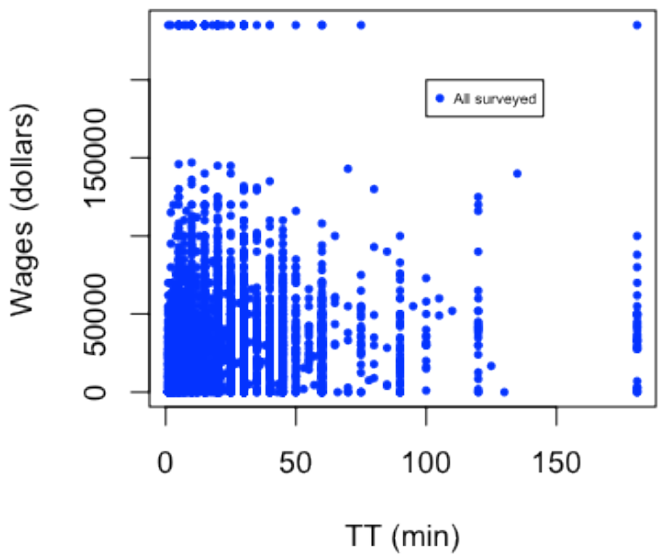
```
plot(pData$JWMNP,pData$WAGP,pch=19,col="blue",cex=0.5,
     xlab="Travel time (min)",ylab="Last 12 month wages (dollars)",cex.lab=2,cex.axis=1.5)
```



Legends

- Important paramters: *x*,*y*,*legend*, *other plotting parameters*

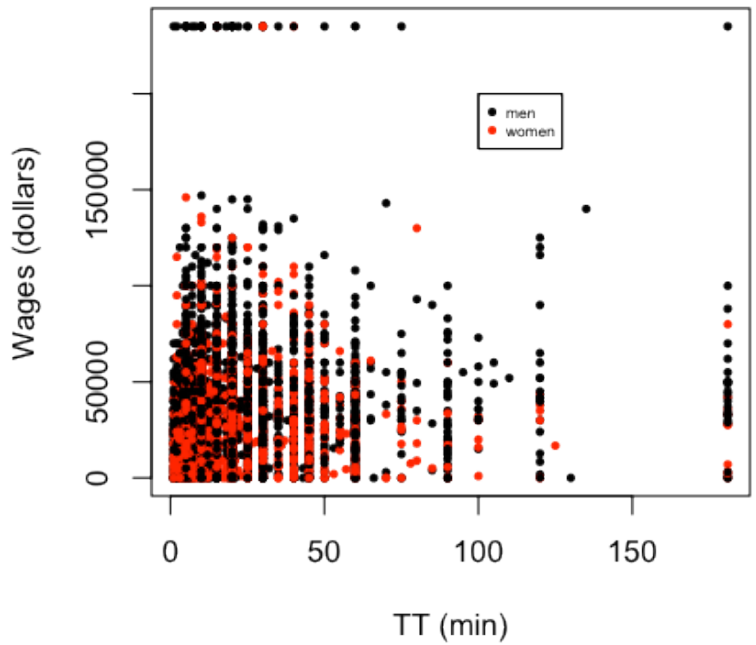
```
plot(pData$JWMNP,pData$WAGP,pch=19,col="blue",cex=0.5,xlab="TT (min)",ylab="Wages (dollars)")
legend(100,200000,legend="All surveyed",col="blue",pch=19,cex=0.5)
```



8/21

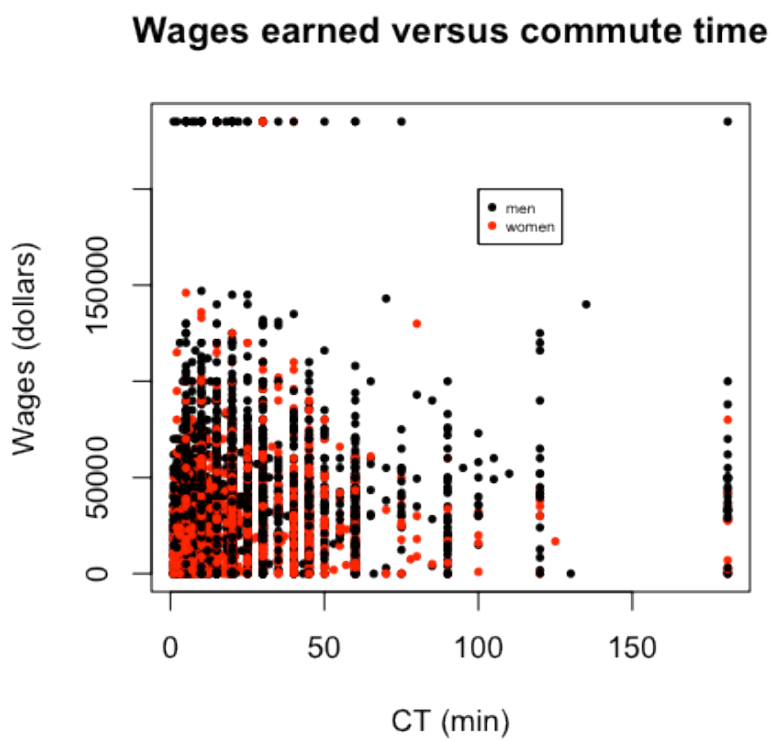
Legends

```
plot(pData$JWMNP,pData$WAGP,pch=19,cex=0.5,xlab="TT (min)",ylab="Wages (dollars)",col=pData$SEX)
legend(100,200000,legend=c("men","women"),col=c("black","red"),pch=c(19,19),cex=c(0.5,0.5))
```



Titles

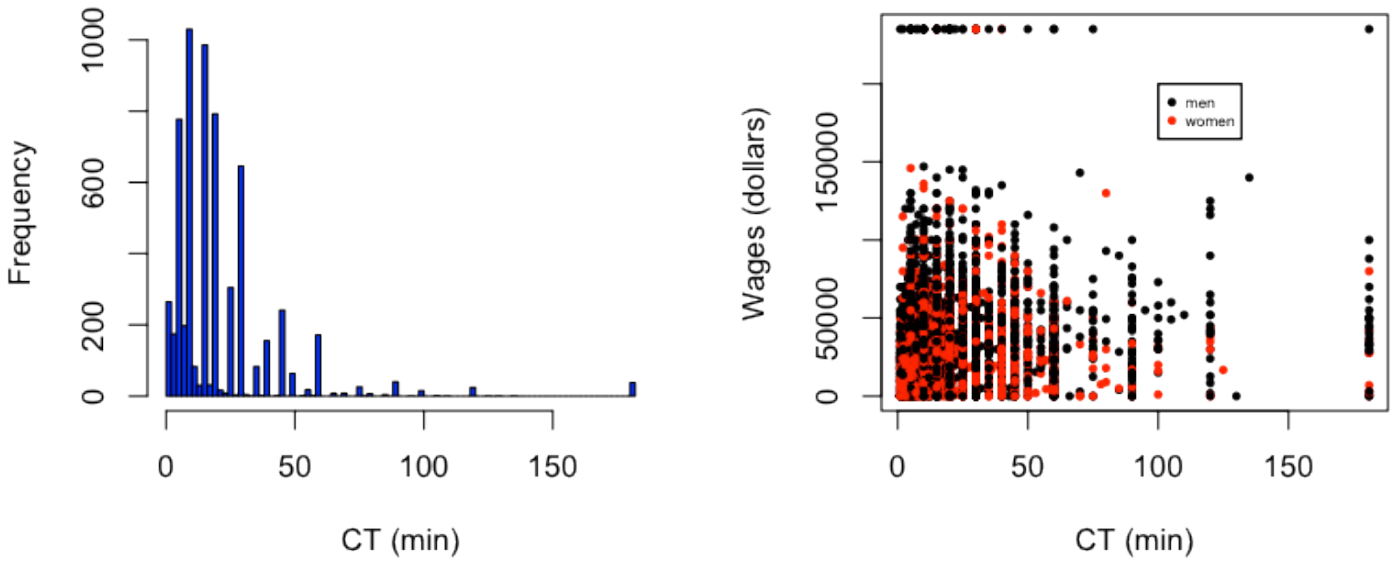
```
plot(pData$JWMNP,pData$WAGP,pch=19,cex=0.5,xlab="CT (min)",
     ylab="Wages (dollars)",col=pData$SEX,main="Wages earned versus commute time")
legend(100,200000,legend=c("men","women"),col=c("black","red"),pch=c(19,19),cex=c(0.5,0.5))
```



10/21

Multiple panels

```
par(mfrow=c(1,2))
hist(pData$JWMNP,xlab="CT (min)",col="blue",breaks=100,main="")
plot(pData$JWMNP,pData$WAGP,pch=19,cex=0.5,xlab="CT (min)",ylab="Wages (dollars)",col=pData$SEX)
legend(100,200000,legend=c("men","women"),col=c("black","red"),pch=c(19,19),cex=c(0.5,0.5))
```



Adding text

```
par(mfrow=c(1,2))
hist(pData$JWMNP,xlab="CT (min)",col="blue",breaks=100,main="")
mtext(text="(a)",side=3,line=1)
plot(pData$JWMNP,pData$WAGP,pch=19,cex=0.5,xlab="CT (min)",ylab="Wages (dollars)",col=pData$SEX)
legend(100,200000,legend=c("men","women"),col=c("black","red"),pch=c(19,19),cex=c(0.5,0.5))
mtext(text="(b)",side=3,line=1)
```

Figure captions

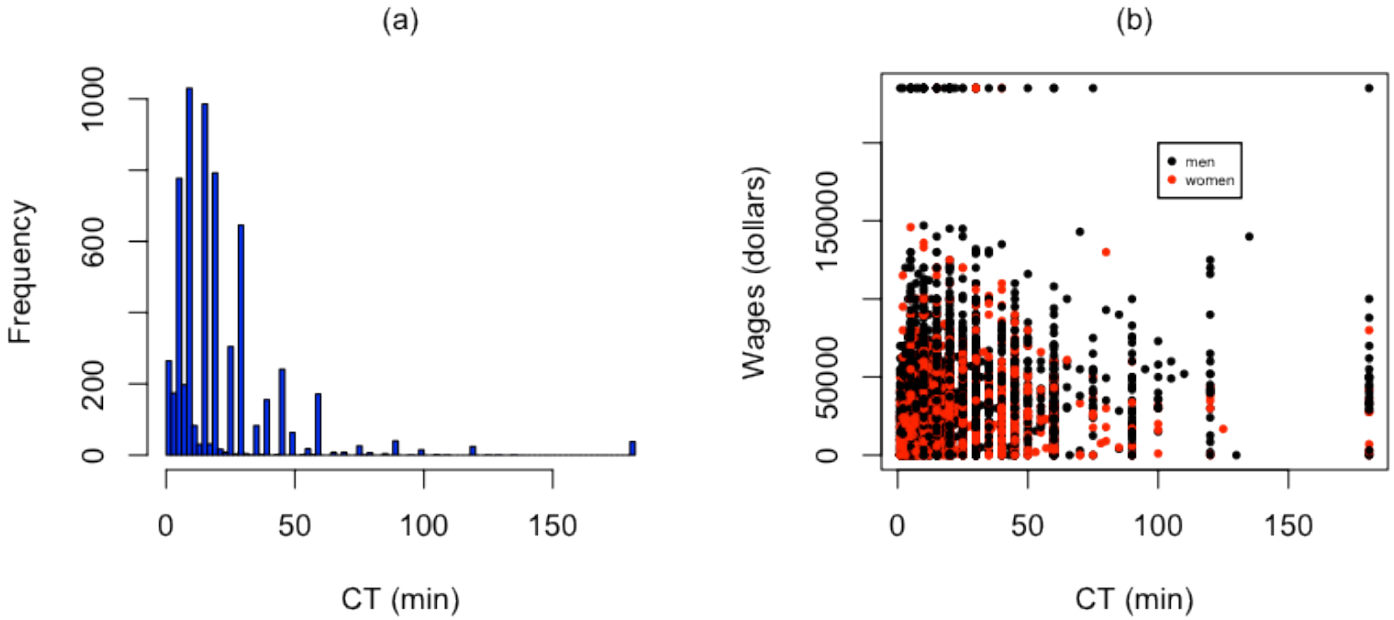
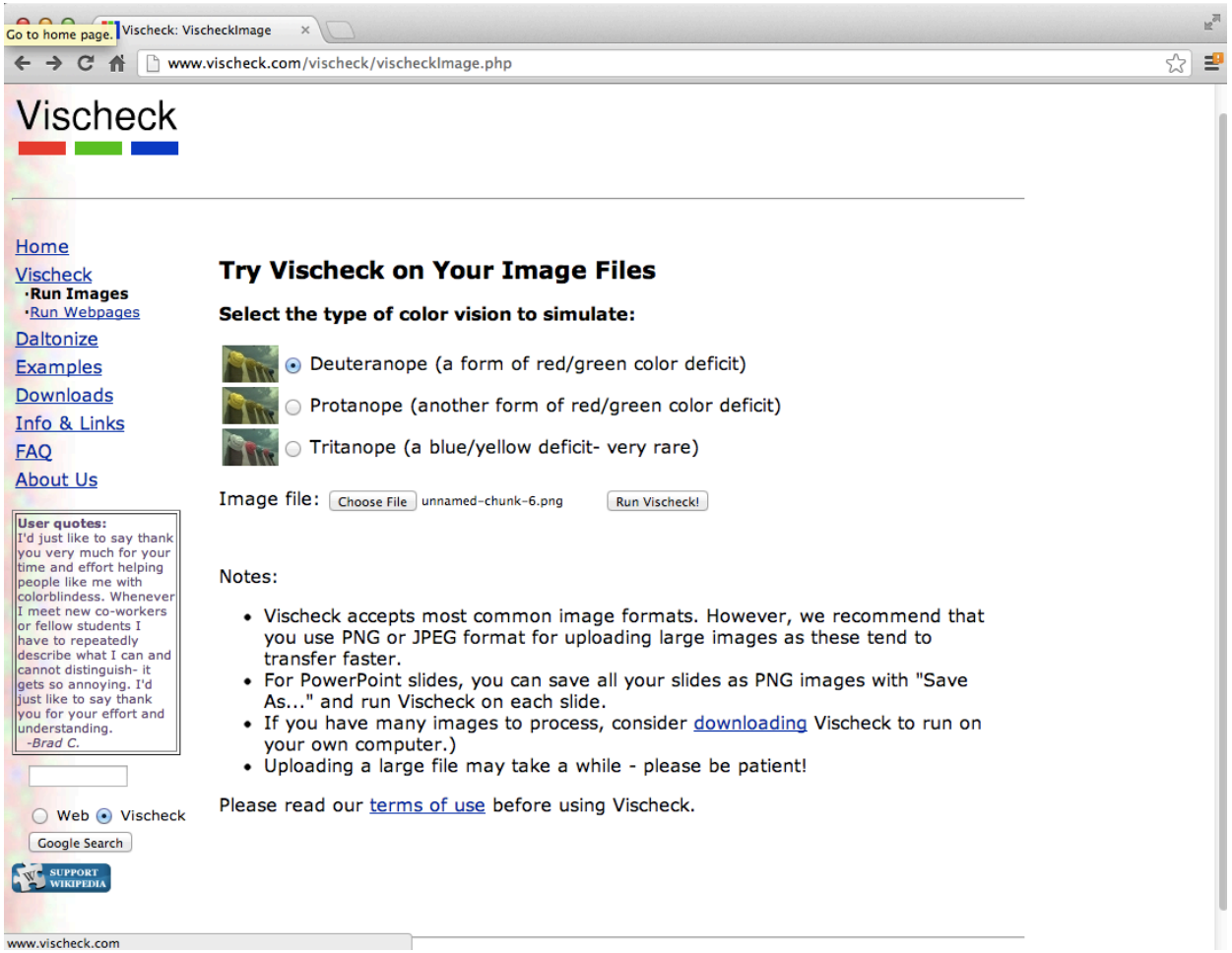


Figure 1. Distribution of commute time and relationship to wage earned by sex (a) Commute times in the American Community Survey (ACS) are right skewed. **(b)** Commute times do not appear to be strongly correlated with wage for either sex.

Colorblindness



http://www.vischeck.com/

Graphical workflow

- Start with a rough plot
- Tweak it to make it expository
- Save the file
- Include it in presentations

Saving files in R is done with graphics *devices*. Use the command ?Devices to see a list. Here we will go over the most popular devices.

pdf

- Important parameters: *file*, *height*,*width*

```
pdf(file="twoPanel.pdf",height=4,width=8)
par(mfrow=c(1,2))
hist(pData$JWMNP,xlab="CT (min)",col="blue",breaks=100,main="")
mtext(text="(a)",side=3,line=1)
plot(pData$JWMNP,pData$WAGP,pch=19,cex=0.5,xlab="CT (min)",ylab="Wages (dollars)",col=pData$SEX)
legend(100,200000,legend=c("men","women"),col=c("black","red"),pch=c(19,19),cex=c(0.5,0.5))
mtext(text="(b)",side=3,line=1)

dev.off()
```

```
pdf
  2
```

png

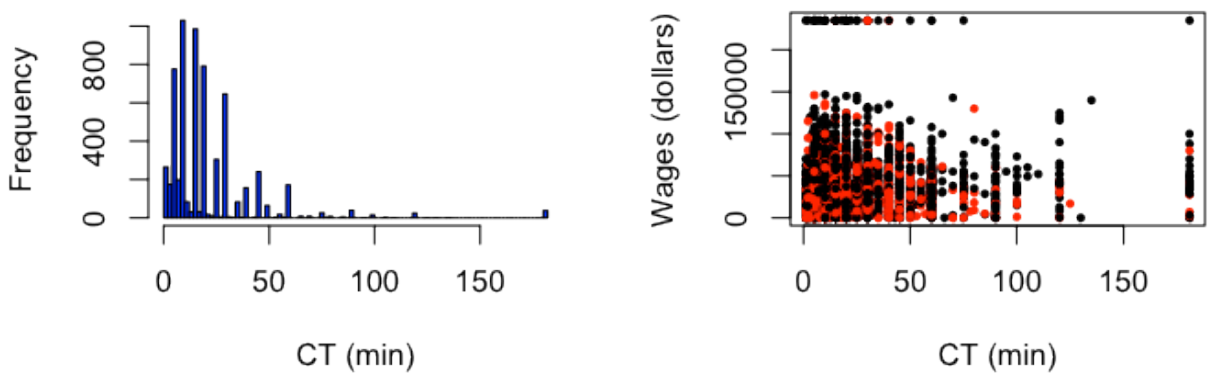
- Important parameters: *file*, *height*,*width*

```
png(file="twoPanel.png",height=480,width=(2*480))
par(mfrow=c(1,2))
hist(pData$JWMNP,xlab="CT (min)",col="blue",breaks=100,main="")
mtext(text="(a)",side=3,line=1)
plot(pData$JWMNP,pData$WAGP,pch=19,cex=0.5,xlab="CT (min)",ylab="Wages (dollars)",col=pData$SEX)
legend(100,200000,legend=c("men","women"),col=c("black","red"),pch=c(19,19),cex=c(0.5,0.5))
mtext(text="(b)",side=3,line=1)
dev.off()
```

```
pdf
  2
```

dev.copy2pdf

```
par(mfrow=c(1,2))
hist(pData$JWMNP,xlab="CT (min)",col="blue",breaks=100,main="")
plot(pData$JWMNP,pData$WAGP,pch=19,cex=0.5,xlab="CT (min)",ylab="Wages (dollars)",col=pData$SEX)
```

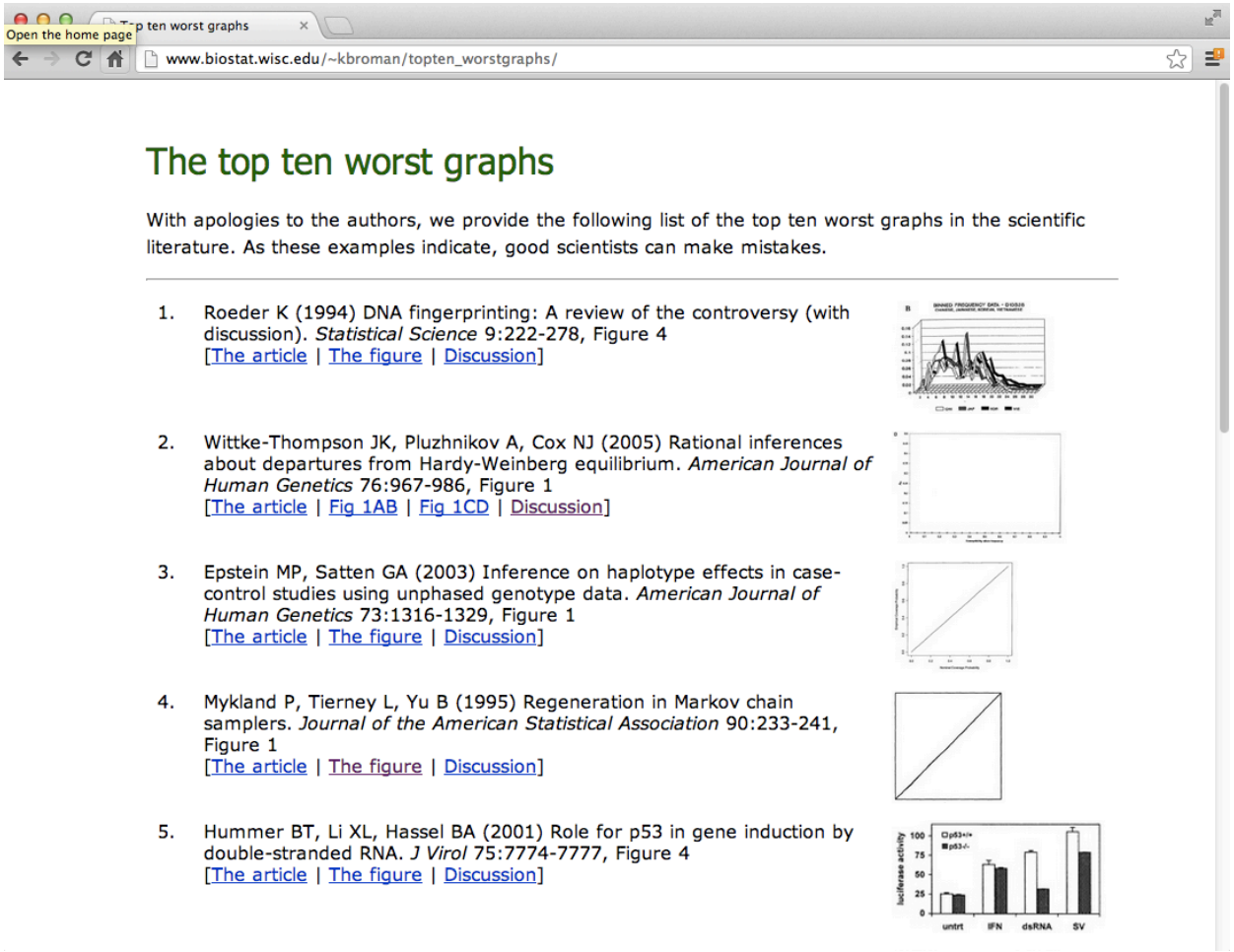



```
dev.copy2pdf(file="twoPanelv2.pdf")
```

```
pdf
  2
```

18/21

Something to avoid



http://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

19/21

Something to aspire to

http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

20/21

Further resources

- How to display data badly
- The visual display of quantitative information
- Creating more effective graphs
- R Graphics Cookbook
- ggplot2: Elegant Graphics for Data Analysis
- Flowing Data