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

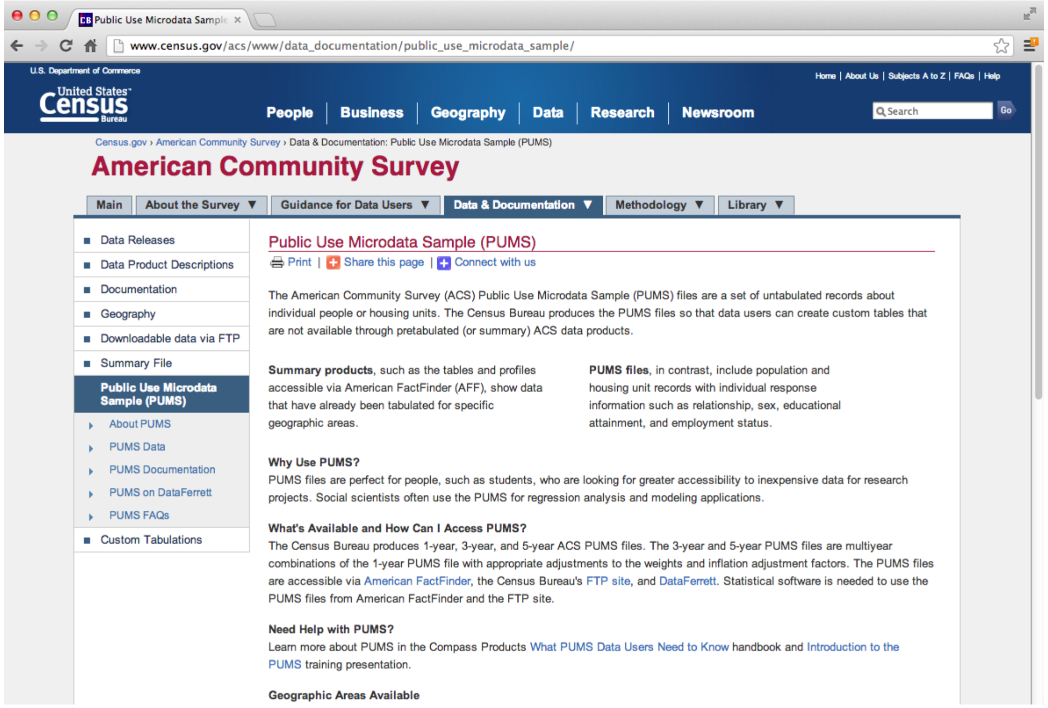
Expository graphs

- 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


Public Use Microdata Sample
www.census.gov/acs/www/data_documentation/public_use_microdata_sample/
U.S. Department of Commerce
United States Census Bureau
Home | About Us | Subjects A to Z | FAQs | Help
People
Business
Geography
Data
Research
Newsroom
Search
Go
Census.gov › American Community Survey › Data & Documentation: Public Use Microdata Sample (PUMS)
Main
About the Survey ▼
Guidance for Data Users ▼
Data & Documentation ▼
Methodology ▼
Library ▼
Data Releases
Data Product Descriptions
Documentation
Geography
Downloadable data via FTP
Summary File
Public Use Microdata Sample (PUMS)
About PUMS
PUMS Data
PUMS Documentation
PUMS on DataFerrett
PUMS FAQs
Custom Tabulations
Print | Share this page | Connect with us

American Community Survey

Public Use Microdata Sample (PUMS)

The American Community Survey (ACS) Public Use Microdata Sample (PUMS) files are a set of untabulated records about individual people or housing units. The Census Bureau produces the PUMS files so that data users can create custom tables that are not available through pretabulated (or summary) ACS data products.

Summary products, such as the tables and profiles accessible via American FactFinder (AFF), show data that have already been tabulated for specific geographic areas.

PUMS files, in contrast, include population and housing unit records with individual response information such as relationship, sex, educational attainment, and employment status.

Why Use PUMS?

PUMS files are perfect for people, such as students, who are looking for greater accessibility to inexpensive data for research projects. Social scientists often use the PUMS for regression analysis and modeling applications.

What's Available and How Can I Access PUMS?

The Census Bureau produces 1-year, 3-year, and 5-year ACS PUMS files. The 3-year and 5-year PUMS files are multiyear combinations of the 1-year PUMS file with appropriate adjustments to the weights and inflation adjustment factors. The PUMS files are accessible via American FactFinder, the Census Bureau's FTP site, and DataFerrett. Statistical software is needed to use the PUMS files from American FactFinder and the FTP site.

Need Help with PUMS?

Learn more about PUMS in the Compass Products What PUMS Data Users Need to Know handbook and Introduction to the PUMS training presentation.

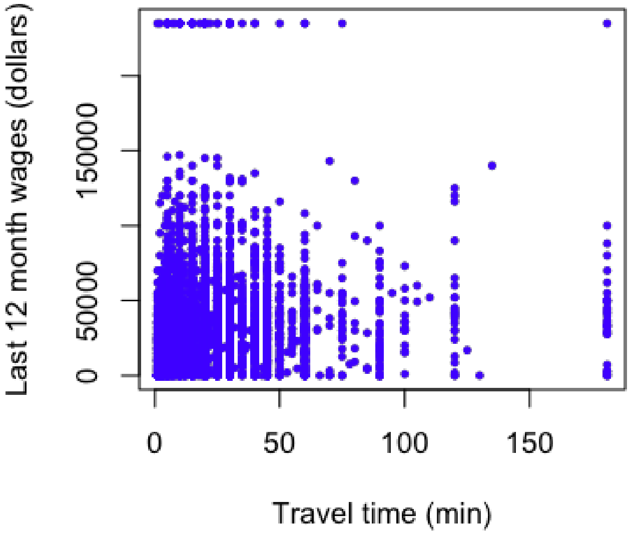
Geographic Areas Available

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

Axes

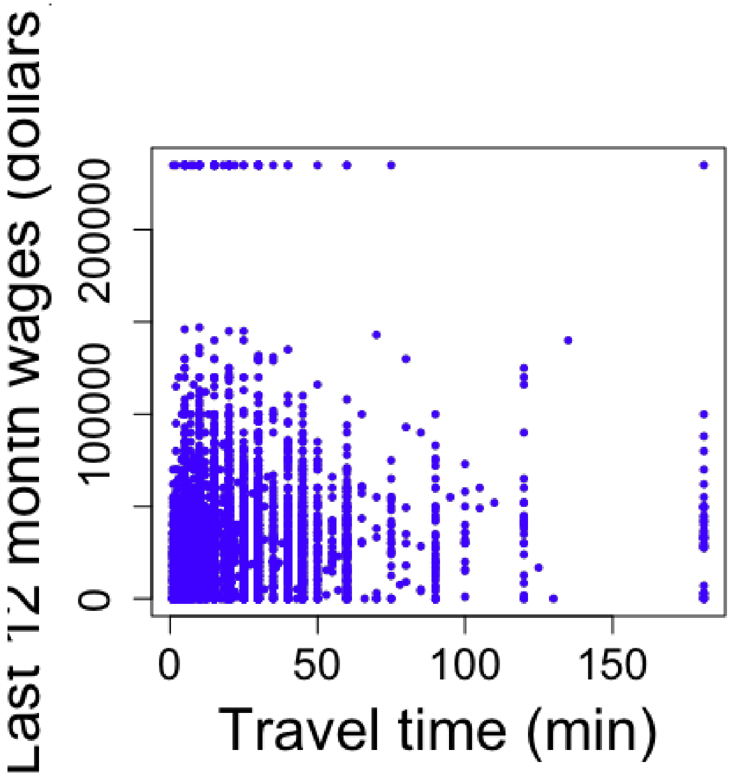
Important parameters: *xlab,ylab,cex.lab,cex.axis*

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



Axes

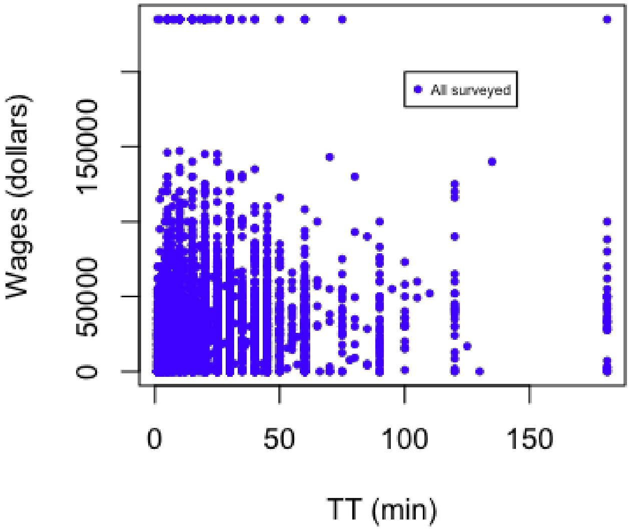
```
plot(pData$JWMNP,pData$WAGP,pch=19,col="blue",cex=0.5,
     xlab="Travel time (min)",ylab="Last 12 month wages (dollars)",cex.lab=2,cex.axis=1.5)
```



Legends

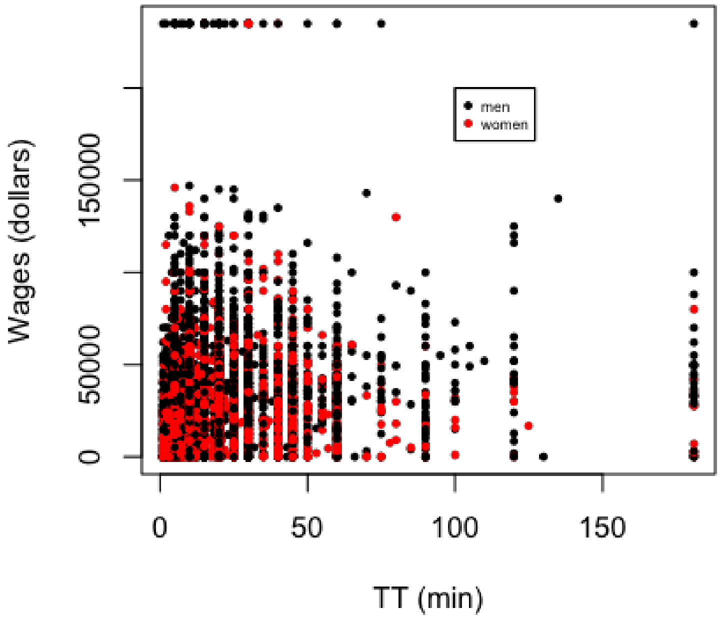
- Important paramters: *x*,*y*,*legend*, *other plotting parameters*

```
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)
```



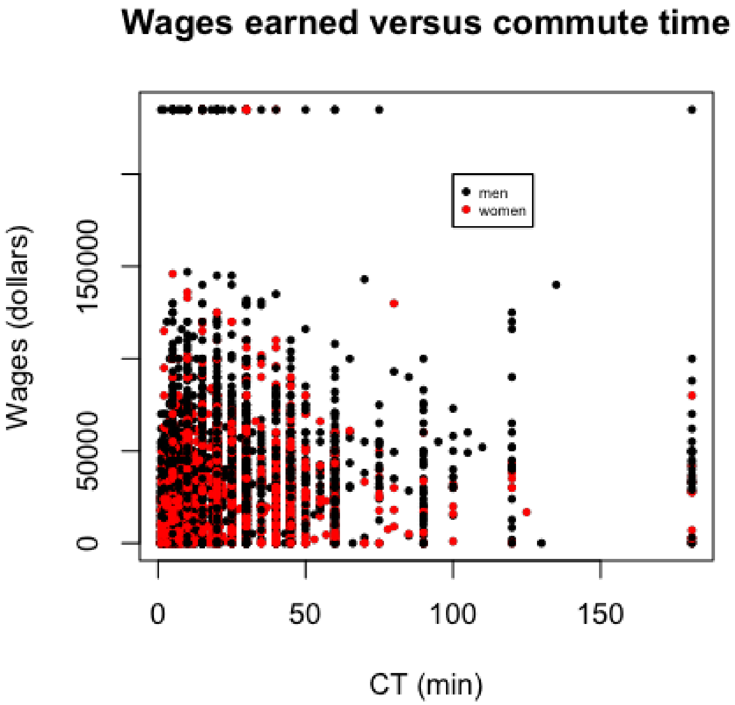
Legends

```
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))
```



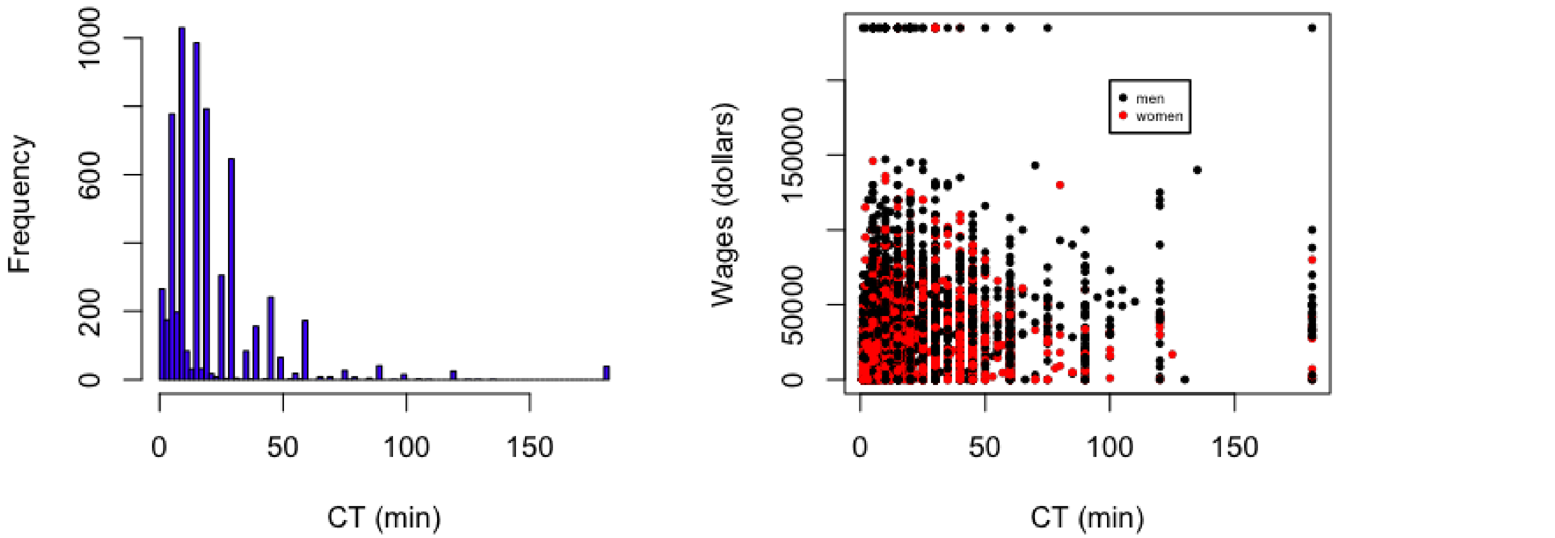
Titles

```
plot(pDataSJWMNP,pDataSWAGP,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))
```



Multiple panels

```
par(mfrow=c(1,2))
hist(pDataSJWMNP,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)
```

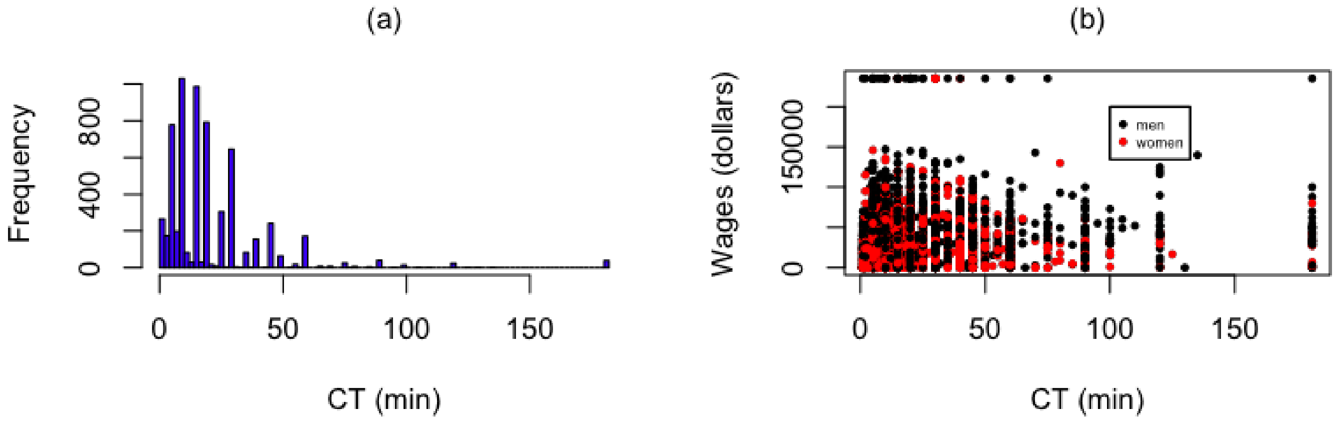


Figure captions

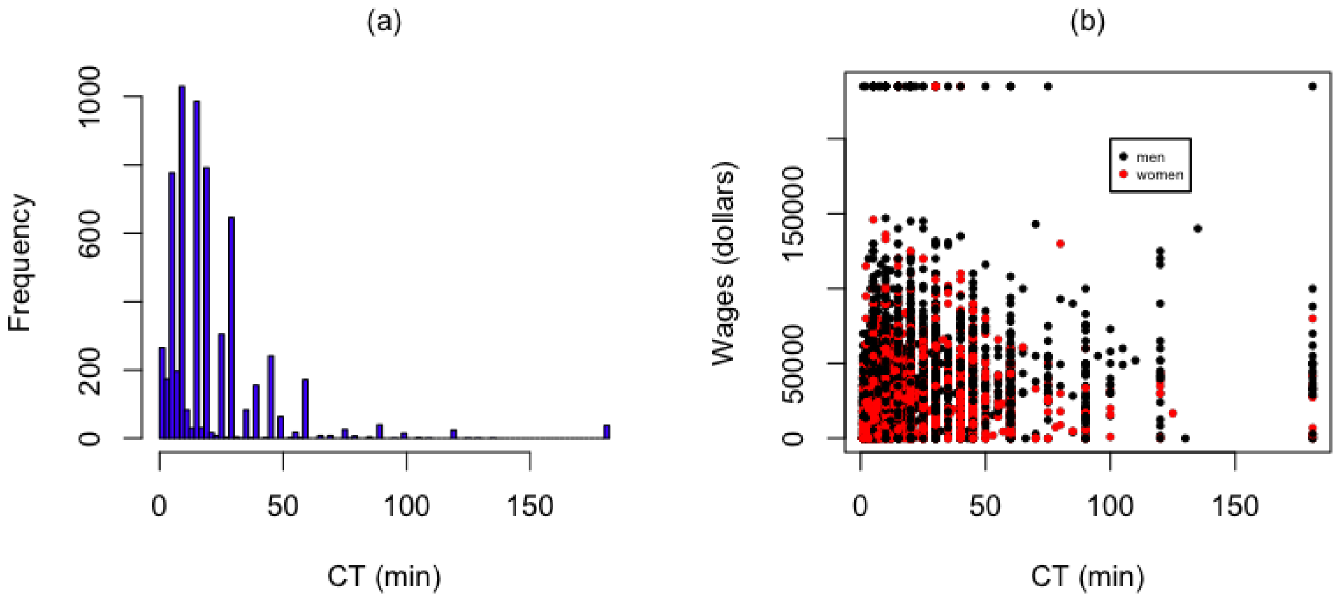
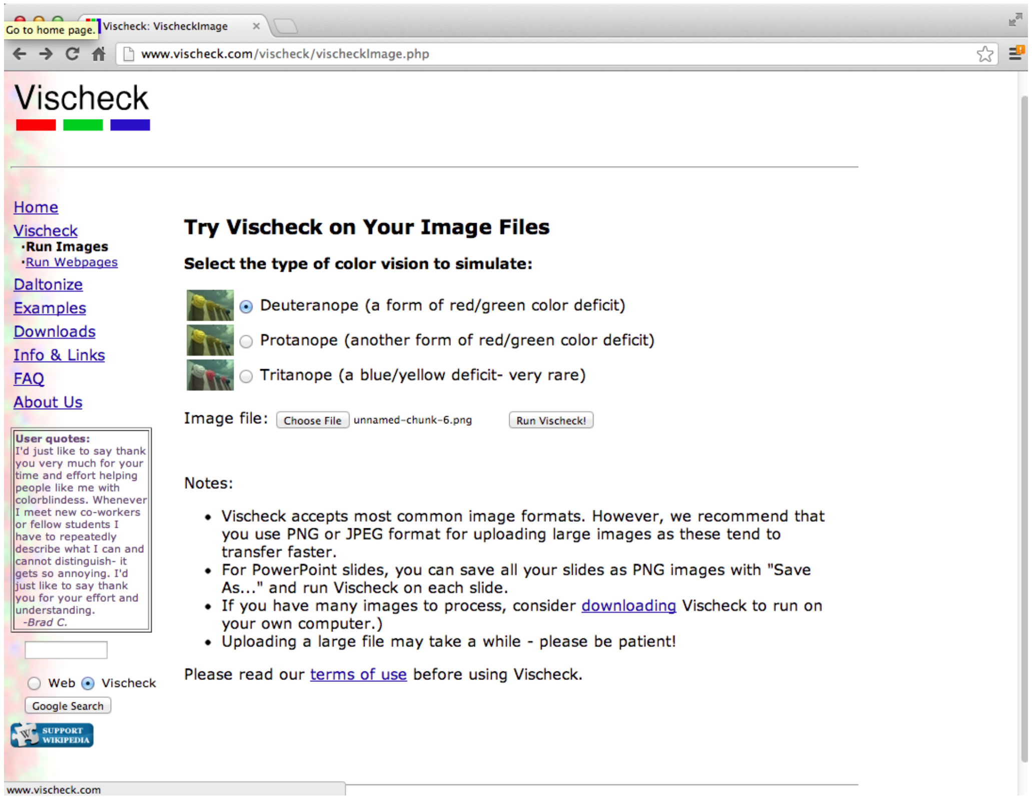


Figure 1. Distribution of commute time and relationship to wage earned by sex (a) Commute times in the American Community Survey (ACS) are right skewed. **(b)** Commute times do not appear to be strongly correlated with wage for either sex.

Colorblindness



User quotes:
I'd just like to say thank you very much for your time and effort helping people like me with colorblindness. Whenever I meet new co-workers or fellow students I have to repeatedly describe what I can and cannot distinguish- it gets so annoying. I'd just like to say thank you for your effort and understanding.
-Brad C.

Try Vischeck on Your Image Files

Select the type of color vision to simulate:

● Deuteranope (a form of red/green color deficit)
○ Protanope (another form of red/green color deficit)
○ Tritanope (a blue/yellow deficit- very rare)

Image file: Choose File unnamed-chunk-6.png Run Vischeck!

Notes:

- Vischeck accepts most common image formats. However, we recommend that you use PNG or JPEG format for uploading large images as these tend to transfer faster.
- For PowerPoint slides, you can save all your slides as PNG images with "Save As..." and run Vischeck on each slide.
- If you have many images to process, consider downloading Vischeck to run on your own computer.)
- Uploading a large file may take a while - please be patient!

Please read our terms of use before using Vischeck.

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()
```

png

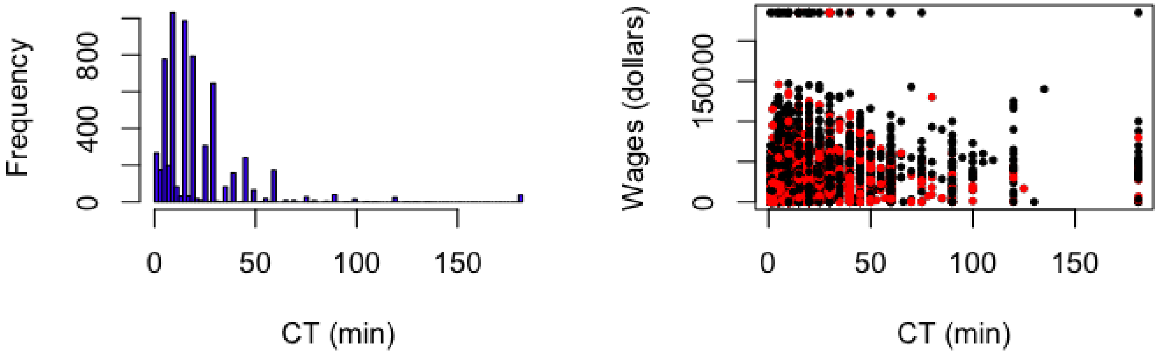
- Important parameters: *file*, *height*,*width*

```
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()
```

```
RStudioGD
        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)
```

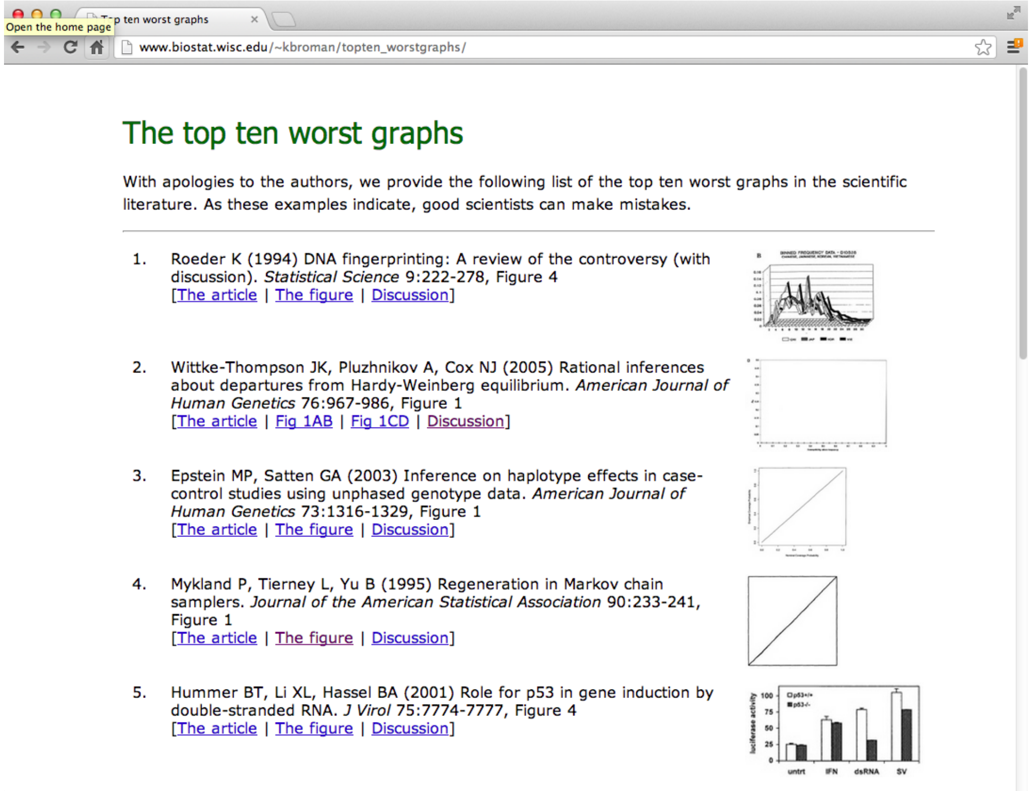


dev.copy2pdf

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

```
RStudioGD
        2
```

Something to avoid



The top ten worst graphs

With apologies to the authors, we provide the following list of the top ten worst graphs in the scientific literature. As these examples indicate, good scientists can make mistakes.

1. Roeder K (1994) DNA fingerprinting: A review of the controversy (with discussion). *Statistical Science* 9:222-278, Figure 4
[The article | The figure | Discussion]

2. Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76:967-986, Figure 1
[The article | Fig 1AB | Fig 1CD | Discussion]

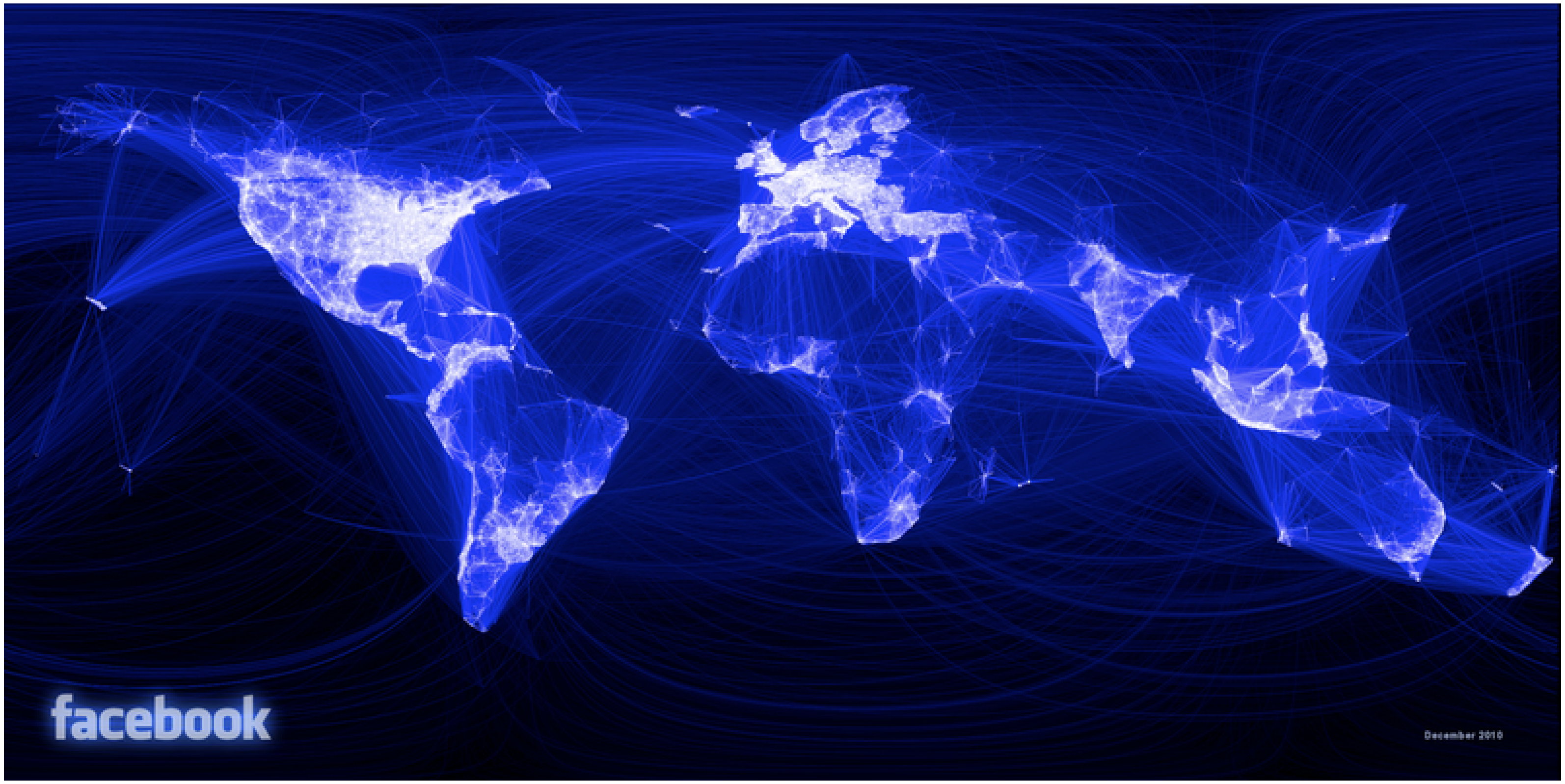
3. Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* 73:1316-1329, Figure 1
[The article | The figure | Discussion]

4. Mykland P, Tierney L, Yu B (1995) Regeneration in Markov chain samplers. *Journal of the American Statistical Association* 90:233-241, Figure 1
[The article | The figure | Discussion]

5. Hummer BT, Li XL, Hassel BA (2001) Role for p53 in gene induction by double-stranded RNA. *J Virol* 75:7774-7777, Figure 4
[The article | The figure | Discussion]

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

Something to aspire to



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

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