Visualization as the Gateway Drug to Statistics in Week One

Xiaofei (Susan) Wang Cynthia Rush

Department of Mathematics and Statistics
Amherst College

Department of Statistics
Yale University

USCOTS 2015

The Activity

Requirements:

Time: approximately 50 minutes or more

Software: `R` (optimally bolstered by `RStudio`), accompanied by the `mosaic` package

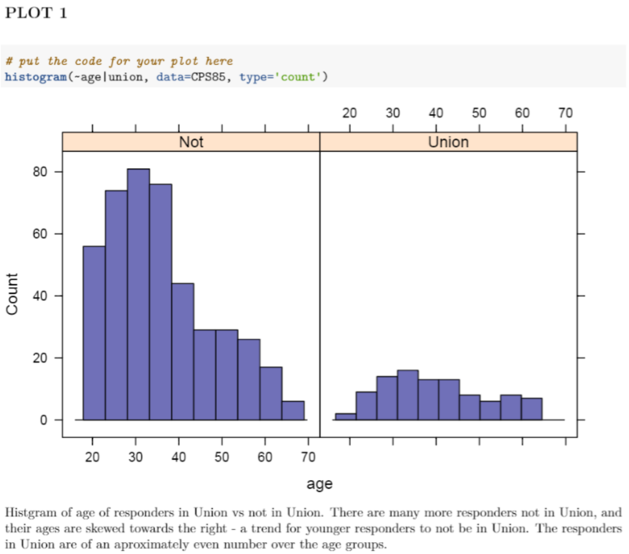
The Task

Using an RMarkdown file, produce 3 compelling graphics that tell a story about a new dataset. Write a few sentences describing what you learn from each of the three plots.

More examples of student work

`http://xiaofei-wang.com/conferences/USCOTS15/`

Student Work

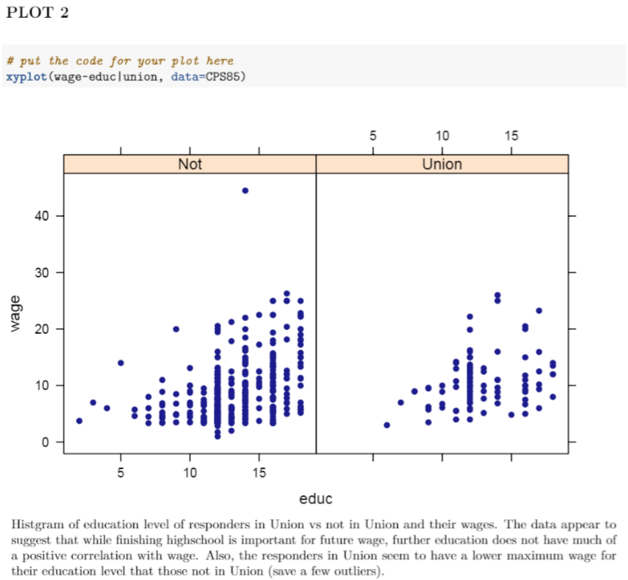


PLOT 1

```
# put the code for your plot here
histogram(~age|union, data=CPS85, type='count')
```

Histgram of age of responders in Union vs not in Union. There are many more responders not in Union, and their ages are skewed towards the right - a trend for younger responders to not be in Union. The responders in Union are of an aproximately even number over the age groups.

Student Work

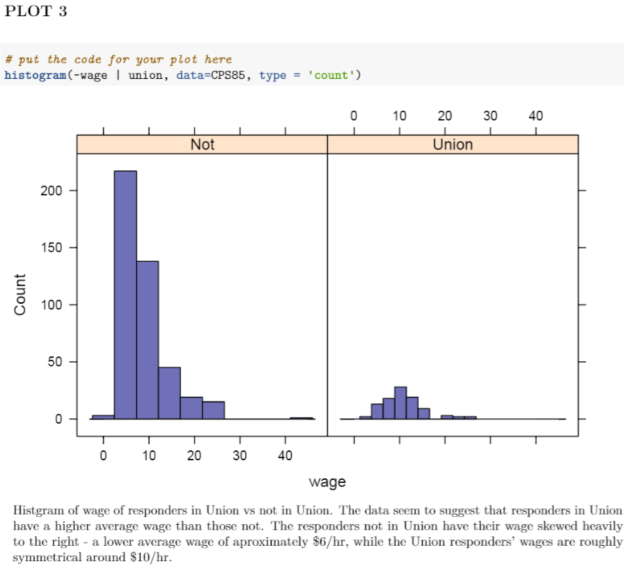


PLOT 2

```
# put the code for your plot here
xyplot(wage~educ|union, data=CPS85)
```

Histgram of education level of responders in Union vs not in Union and their wages. The data appear to suggest that while finishing highschool is important for future wage, further education does not have much of a positive correlation with wage. Also, the responders in Union seem to have a lower maximum wage for their education level that those not in Union (save a few outliers).

Student Work



PLOT 3

```
# put the code for your plot here
histogram(~wage | union, data=CPS85, type = 'count')
```

Histgram of wage of responders in Union vs not in Union. The data seem to suggest that responders in Union have a higher average wage than those not. The responders not in Union have their wage skewed heavily to the right - a lower average wage of aproximately $6/hr, while the Union responders' wages are roughly symmetrical around $10/hr.

Introductory Statistics Core Objectives

GAISE College Report (2012):

> *The desired result of all introductory statistics courses is to produce statistically educated students, which means that students should develop **statistical literacy** and the ability to **think statistically**.*

What does this mean? Students should learn to...

- ingest and digest data
- critically evaluate statistical approaches
- understand basic inferential techniques

R: Why, When, and How?

Why?

- R permits focus on **real applications** of methods to data.
- R instills the philosophy of reproducible research early on (facilitated by RMarkdown).

When?

- Starting from Week 1, R can be a powerful tool that connects textbook concepts with the real world.
- RStudio significantly reduces the shock factor encountered by those with no programming experience.

How?

- **Don't** begin with "Hello world!" or variable assignments.
- **Do** begin with compelling visualizations.

First Week Objectives

Many introductory texts start by describing what are data and how to visualize them (Diez, et al. 2012; De Veaux, et al. 2012).

After the completion of our lab activity, students will additionally have:

- authored and compiled an RMarkdown file documenting an exploratory data analysis
- applied a range of visualization tools to real data
- communicated the conclusions that can be drawn from various graphics

Handout to Students

Variable(s)	Plot	R command
1 categorical	bar graph	`bargraph( ~ x)`
1 quantitative	histogram	`histogram( ~ x)`
1 categorical vs. 1 quantitative	boxplot	`bwplot(y ~ x)`
1 quantitative vs. 1 quantitative	scatterplot	`xyplot(y ~ x)`
1 categorical vs. 1 categorical	mosaic plot	`mosaicplot(y ~ x)`

These functions can be found in the `mosaic` or `lattice` packages, or in base `R`.

mosaic Package and the Formula Notation

From Pruim (2014):

```
goal( y ~ x , data = ...)
```

The mosaic package is a helpful tool for teaching R in introductory statistics, extending the above **formula notation** to common functions.

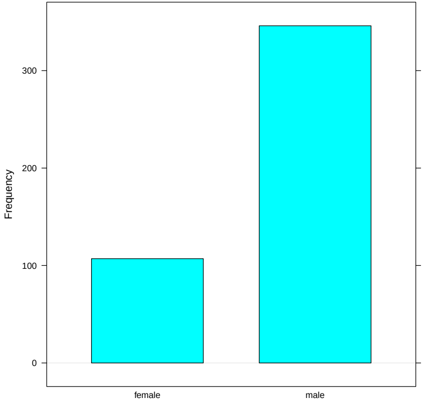
```
mean(Sepal.Width ~ Species, data=iris)

##     setosa versicolor  virginica
##      3.428      2.770      2.974

cor(Sepal.Width ~ Sepal.Length, data=iris)

## [1] -0.1175698
```

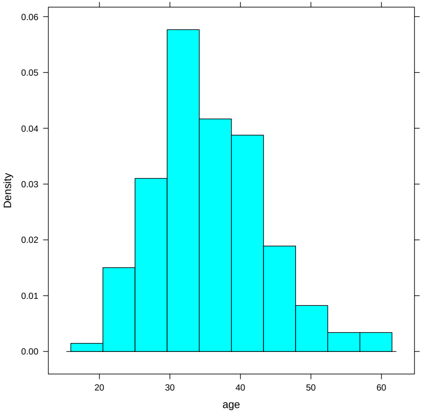
This data visualization exercise introduces students to this ubiquitous notation.

Example Plots

```
bargraph(~ sex, data=HELPrct)
```

```
histogram(~ age, data=HELPrct)
```

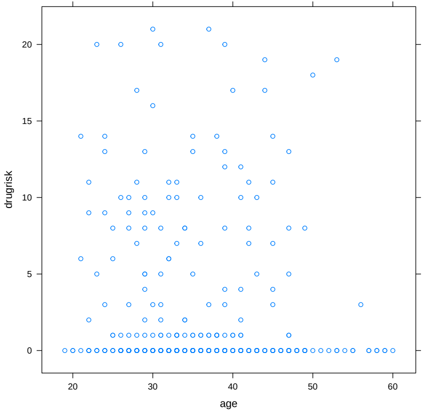


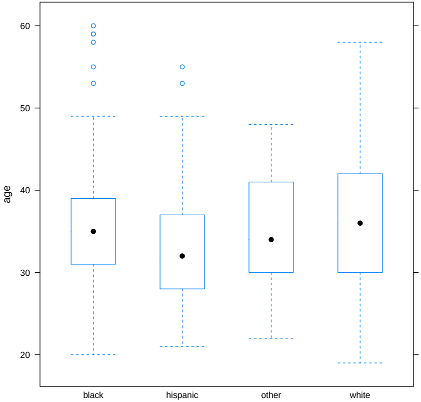


Example Plots

```
xyplot(drugrisk ~ age, data=HELPrct)
```

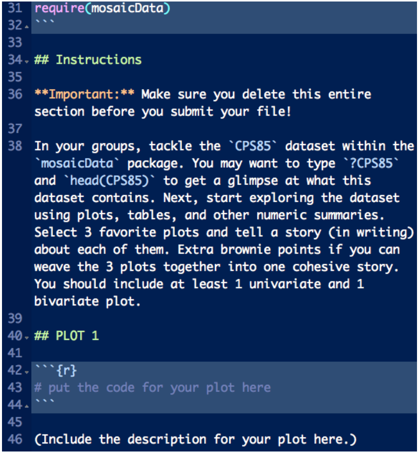
```
bwplot(age ~ racegrp, data=HELPrct)
```



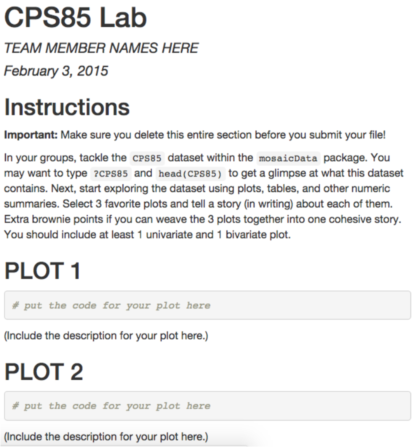


Their Turn

After providing a handout with the examples, the students are now invited to explore a different dataset.



~~~
require(mosaicData)
```

## Instructions

**Important:** Make sure you delete this entire section before you submit your file!

In your groups, tackle the `CPS85` dataset within the `mosaicData` package. You may want to type `?CPS85` and `head(CPS85)` to get a glimpse at what this dataset contains. Next, start exploring the dataset using plots, tables, and other numeric summaries. Select 3 favorite plots and tell a story (in writing) about each of them. Extra brownie points if you can weave the 3 plots together into one cohesive story. You should include at least 1 univariate and 1 bivariate plot.

## PLOT 1

```{r}
# put the code for your plot here
```

(Include the description for your plot here.)
~~~



CPS85 Lab

TEAM MEMBER NAMES HERE

February 3, 2015

Instructions

Important: Make sure you delete this entire section before you submit your file!

In your groups, tackle the `CPS85` dataset within the `mosaicData` package. You may want to type `?CPS85` and `head(CPS85)` to get a glimpse at what this dataset contains. Next, start exploring the dataset using plots, tables, and other numeric summaries. Select 3 favorite plots and tell a story (in writing) about each of them. Extra brownie points if you can weave the 3 plots together into one cohesive story. You should include at least 1 univariate and 1 bivariate plot.

PLOT 1

```
# put the code for your plot here
```

(Include the description for your plot here.)

PLOT 2

```
# put the code for your plot here
```

(Include the description for your plot here.)

Going Beyond

Ways to extend the lab:

- RPubs: with minimal work, allows students to share their work online.
- Making it competitive: in the spirit of active learning, ask students to present their graphics to the class and vote on the best presentation.

What's next?

- *Reinforce* data visualization techniques in the first homework assignment.
- Slowly ease into the nitty-gritty, like *data management skills* in `R`.

References

- 2014 ASA/MAA Guidelines for Teaching Introductory Statistics `http://magazine.amstat.org/2014/04/01/asamaaguidelines`
- Curriculum Guidelines for Undergraduate Programs in Statistical Science `http://www.amstat.org/education/pdfs/guidelines2014-11-15.pdf`
- De Veaux, R. D., Velleman, P. F., and Bock, D. E. (2012), *Stats: Data and Models* (3rd ed.), Addison-Wesley.
- Diez, D. M., Barr, C. D., Cetinkaya-Rundel, M. (2012), *OpenIntro Statistics* (2nd ed.).
- GAISE College Report. (2012). `http://www.amstat.org/education/gaise`
- Nolan, D., and Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, 64(2).
- Pruim, R. Less volume, more creativity. Workshop provided at *eCOTS 2014*. `rpruim.github.io/eCOTS2014/Workshop/LessVolumeMoreCreativity.html`