Eric Pitman Summer Workshop in Computational Science

5. Visualizing Data
Plotting Data

Plotting is another way to explore a dataset, visually:

- What's in the dataset?
- What does it mean?
- What if there's a lot of it?
Some Plot Types

- **Pie Chart**
  - Display proportions of different values for a variable
- **Bar Plot**
  - Display counts of values for a categorical variable
- **Histogram, Density Plot**
  - Display counts of values for a binned, numeric variable
- **Scatter Plot**
  - y vs. x
- **Box Plot**
  - Display distributions over different values of a variable
# Barplot: Counts of Categorical Values

<table>
<thead>
<tr>
<th>Color</th>
<th>Fair</th>
<th>Good</th>
<th>Very Good</th>
<th>Premium</th>
<th>Ideal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note: The barplot shows the count of categorical values across different colors and quality levels.*
ideal = diamonds$diamonds$cut == "Ideal", "color"

barplot(table(ideal),
    xlab="color",
    ylab="count",
    main="Ideal cut diamonds by Color",
    col="hotpink")
Histogram: Frequencies of Numeric Values

hist(Cars93$RPM, 
  xlab="engine RPM", 
  main="histogram of engine RPM", 
  col="red")
Histogram and Density Binning

bw=0.01

\( x \)

too small

Density

bw=1

\( x \)

too big

Density

bw=0.1

\( x \)

just right

Density
Kernel Density Plot

plot(density(Cars93$RPM),
xlab="engine RPM",
main="density plot of engine RPM",
col="red")
Scatterplot: Numeric Data

Price = Dependent Variable ↑

Carats = Independent variable →
Scatterplot with Regression Lines

Regression of MPG on Weight

Miles per Gallon

Weight
Scatterplot: Numeric Data, y vs. x

plot(formula=price~carat, 
data=diamonds, 
col="darkblue", 
pch=20, 
main="Diamond Price with Size")
Box (and Whisker) Plot

- The box extends from Q1 to Q3
- The median, Q2, is marked inside the box
- The whiskers extend to the min and max
  - Whiskers: required to lie within $1.5 \times (IQR)$
  - Outliers: beyond $1.5 \times (IQR)$
Boxplot: Data Symmetry?

Mileage by Gear Number

3 gears
4 gears
5 gears

Miles per Gallon

30
25
20
15
10
10
15
20
25
30
35
Box (and Whisker) Plot

```r
boxplot(formula=mpg~gear,
data=mtcars,
main="Mileage by Gear Number",
xlab="Number of Gears",
ylab="Miles Per Gallon",
col=c("red","green","blue"))
```
GIS plot

2011 Hospital Admissions by NYS County

admissions as % of 2000 population
Approach to Plotting

- Remember, you're getting to know your data.
- Don't be afraid to tinker and play.
- Sometimes the outcomes are silly (make sure you learn something!)
Interlude

Complete plotting exercises.

Open in the RStudio source editor:

<workshop>/exercises/exercises-plotting-basic.R
If you want to experiment further with R and RStudio, you can install them on your favorite operating system at home.

First, install R:

http://cran.r-project.org/

Then, install the Rstudio IDE:

http://www.rstudio.com/ide/