BGGN 213
Data visualization with R
Lecture 5
Barry Grant
UC San Diego
http://thegrantlab.org/bggn213
Recap From Last Time:

- What is R and why should we use it?
- Familiarity with R’s basic syntax.
- Familiarity with major R data structures namely vectors and data.frames.
- Understand the basics of using functions (arguments, vectorizion and re-cycling).
- Appreciate how you can use R scripts to aid with reproducibility.

DataCamp Homework Reminder!!
Today’s Learning Goals

- Appreciate the major elements of exploratory data analysis and why it is important to visualize data.

- Be conversant with data visualization best practices and understand how good visualizations optimize for the human visual system.

- Be able to generate informative graphical displays including scatterplots, histograms, bar graphs, boxplots, dendrograms and heatmaps and thereby gain exposure to the extensive graphical capabilities of R.

- Appreciate that you can build even more complex charts with ggplot and additional R packages such as rgl.
Today’s Learning Goals

• Appreciate the major elements of **exploratory data analysis** and why it is important to visualize data.

• Be conversant with **data visualization best practices** and understand how good visualizations optimize for the human visual system.

• Be able to generate informative graphical displays including **scatterplots**, **histograms**, **bar graphs**, **boxplots**, **dendrograms** and **heatmaps** and thereby gain exposure to the extensive graphical capabilities of R.

• Appreciate that you can build even more complex charts with **ggplot** and additional R packages such as **rgl**.
Why visualize at all?
# Over-the-Counter

## National Market System

The companies listed below reflect the volume in 100's of shares on a daily basis and the closing price and net change are reflected from the previous day's close on trades as reported under the NASDAQ National Market System.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Sales Class</th>
<th>Chg</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued in columns)}
<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>4.18</td>
<td>2.75</td>
</tr>
<tr>
<td>3</td>
<td>1.98</td>
<td>4.59</td>
</tr>
<tr>
<td>4</td>
<td>-0.86</td>
<td>4.92</td>
</tr>
<tr>
<td>5</td>
<td>-3.43</td>
<td>3.64</td>
</tr>
<tr>
<td>6</td>
<td>-4.86</td>
<td>1.16</td>
</tr>
<tr>
<td>7</td>
<td>-4.70</td>
<td>-1.70</td>
</tr>
<tr>
<td>8</td>
<td>-2.99</td>
<td>-4.01</td>
</tr>
<tr>
<td>9</td>
<td>-0.30</td>
<td>-4.99</td>
</tr>
<tr>
<td>10</td>
<td>2.49</td>
<td>-4.34</td>
</tr>
<tr>
<td>11</td>
<td>4.46</td>
<td>-2.25</td>
</tr>
<tr>
<td>12</td>
<td>4.97</td>
<td>0.57</td>
</tr>
<tr>
<td>13</td>
<td>3.84</td>
<td>3.20</td>
</tr>
<tr>
<td>14</td>
<td>1.45</td>
<td>4.79</td>
</tr>
<tr>
<td>15</td>
<td>-1.42</td>
<td>4.79</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Min.</td>
<td>-4.86</td>
<td>-4.99</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>-2.21</td>
<td>-1.98</td>
</tr>
<tr>
<td>Median</td>
<td>1.45</td>
<td>1.16</td>
</tr>
<tr>
<td>Mean</td>
<td>0.65</td>
<td>0.87</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>4.01</td>
<td>4.12</td>
</tr>
<tr>
<td>Max.</td>
<td>5.00</td>
<td>4.92</td>
</tr>
</tbody>
</table>
Exploratory Data Analysis

- ALWAYS look at your data!
- If you can’t see it, then don’t believe it!
- Exploratory Data Analysis (EDA) allows us to:
  1. Visualize distributions and relationships
  2. Detect errors
  3. Assess assumptions for confirmatory analysis
- EDA is the first step of data analysis!
Exploratory Data Analysis 1977

- Based on insights developed at Bell Labs in the 60's
- Techniques for visualizing and summarizing data
- What can the data tell us? (in contrast to "confirmatory" data analysis)
- Introduced many basic techniques:
  - 5-number summary, box plots, stem and leaf diagrams, ...
- 5 Number summary:
  - extremes (min and max)
  - median & quartiles
  - More robust to skewed & longtailed distributions
Side-note: How to summarize data?

```r
x <- rnorm(1000)
```
Side-note: Mean & standard deviation

Fine for normally distributed data

```r
x <- rnorm(1000)
mean(x)
sd(x)
```
Side-note: 5 number summary
Minimum, Q1, Q2, Q3, and maximum

```r
x <- rnorm(1000)
mean(x)
sd(x)
summary(x)
```
Side-note: **5 number summary**

Minimum, Q1, Q2, Q3, and maximum

```r
x <- rnorm(1000)
mean(x)
sd(x)
summary(x)
```
Side-note: **5 number summary**

Minimum, Q1, Q2, Q3, and maximum

\[ x \leftarrow \text{rnorm}(1000) \]

\[ \text{mean}(x) \]

\[ \text{sd}(x) \]

\[ \text{summary}(x) \]
Side-note: 5 number summary
Minimum, Q1, Q2, Q3, and maximum

x <- rnorm(1000)
mean(x)
sd(x)
summary(x)
Side-note: **5 number summary**

Minimum, Q1, Q2, Q3, and maximum

\[ \begin{align*}
\text{minimum} & \quad 2 \\
\text{Q1} & \quad \text{median} \\
\text{Q2} & \quad \text{median} \\
\text{Q3} & \quad \text{interquartile range} \\
\text{maximum} & \quad \text{IQR} = \text{interquartile range} \\
\end{align*} \]

\[ x \leftarrow \text{rnorm}(1000) \]

\[ \text{mean}(x) \]

\[ \text{sd}(x) \]

\[ \text{summary}(x) \]
Side-note: **boxplot**

Graphical form of the 5 number summary!

Also called box-and-whisker plots; See also violin plots etc.
Side-note: **boxplot**

Graphical form of the 5 number summary!

```r
x <- rnorm(1000)
mean(x)
sd(x)
summary(x)
boxplot(x)
hist(x)
rug(x)
```

Also called **box-and-whisker plots**; See also violin plots etc.
The Trouble with Summary Stats

<table>
<thead>
<tr>
<th>Set A</th>
<th></th>
<th>Set B</th>
<th></th>
<th>Set C</th>
<th></th>
<th>Set D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
<td>6.58</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
<td>5.76</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
<td>13</td>
<td>12.74</td>
<td>8</td>
<td>7.71</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
<td>9</td>
<td>7.11</td>
<td>8</td>
<td>8.84</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
<td>8.47</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
<td>14</td>
<td>8.84</td>
<td>8</td>
<td>7.04</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
<td>5.25</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.1</td>
<td>4</td>
<td>5.39</td>
<td>19</td>
<td>12.5</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.11</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
<td>5.56</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
<td>7</td>
<td>6.42</td>
<td>8</td>
<td>7.91</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
<td>5</td>
<td>5.73</td>
<td>8</td>
<td>6.89</td>
</tr>
</tbody>
</table>

Summary Statistics Linear Regression

\[
u_X = 9.0 \quad \sigma_X = 3.317 \quad Y = 3 + 0.5X \quad [\text{Anscombe 73}]
\]

\[
u_Y = 7.5 \quad \sigma_Y = 2.03 \quad R^2 = 0.67
\]
Looking at Data

https://en.wikipedia.org/wiki/Anscombe%27s_quartet
**Key point:** You need to visualize your data!

https://github.com/stephlocke/datasauRus
Today’s Learning Goals

• Appreciate the major elements of exploratory data analysis and why it is important to visualize data.

• Be conversant with data visualization best practices and understand how good visualizations optimize for the human visual system.

• Be able to generate informative graphical displays including scatterplots, histograms, bar graphs, boxplots, dendrograms and heatmaps and thereby gain exposure to the extensive graphical capabilities of R.

• Appreciate that you can build even more complex charts with ggplot and additional R packages such as rgl.
The Elements of Graphing Data

William S. Cleveland
Key Point:
Good visualizations optimize for the human visual system.
Key Point: The most important measurement should exploit the highest ranked encoding possible

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or Color saturation/hue
Key Point: The most important measurement should exploit the highest ranked encoding possible

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or Color saturation/hue
Key Point: The most important measurement should exploit the highest ranked encoding possible

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or **Color saturation/hue**
Observation: Alphabetical is almost never the correct ordering of a categorical variable.
If we did not have the legend would you know which was low or high mpg?
The most important measurement should exploit the highest ranked encoding possible.

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- **Area**
- Volume or Density or Color saturation/hue
The most important measurement should exploit the highest ranked encoding possible.

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- **Angle or Slope**
- Area
- Volume or Density or Color saturation/hue
If growth (slope) is important, plot it directly.
The most important measurement should exploit the highest ranked encoding possible.

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- **Angle** or Slope
- Area
- Volume or Density or Color saturation/hue
Observation: Pie charts are ALWAYS a mistake. Apart from MPAs :-(
Piecharts are the information visualization equivalent of a roofing hammer to the frontal lobe. They have no place in the world of grownups, and occupy the same semiotic space as short pants, a runny nose, and chocolate smeared on one’s face. They are as professional as a pair of assless chaps.

Piecharts are the information visualization equivalent of a roofing hammer to the frontal lobe. They have no place in the world of grownups, and occupy the same semiotic space as short pants, a runny nose, and chocolate smeared on one’s face. They are as professional as a pair of assless chaps.

Who do you think did a better job in tonight’s debate?

Among Republicans

Donald Trump 47%

Among Democrats

Hillary Clinton 53%
Who do you think did a better job in tonight’s debate?

Among Republicans

- Donald Trump: 1%

Among Democrats

- Hillary Clinton: 99%
Tables are preferable to graphics for many small data sets. A table is nearly always better than a dumb pie chart; the only thing worse than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies... Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.

-Edward Tufte, The Visual Display of Quantitative Information
Tables are preferable to graphics for many small data sets. A table is nearly always better than a dumb pie chart; the only thing worse than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies... Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.

-Edward Tufte, The Visual Display of Quantitative Information
Who do you think did a better job in tonight’s debate?

<table>
<thead>
<tr>
<th></th>
<th>Clinton</th>
<th>Trump</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among Democrats</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>Among Republicans</td>
<td>53%</td>
<td>47%</td>
</tr>
</tbody>
</table>
All good pie charts are jokes...
The most important measurement should exploit the highest ranked encoding possible.

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or Color saturation/hue
Toyota Corolla
Fiat 128
Lotus Europa
Honda Civic
Fiat X1-9
Porsche 914-2
Merc 240D
Merc 230
Datsun 710
Toyota Corona
Volvo 142E
Hornet 4 Drive
Mazda RX4 Wagon
Mazda RX4
Ferrari Dino
Pontiac Firebird
Merc 280
Hornet Sportabout
Valiant
Merc 280C
Merc 450SL
Merc 450SE
Ford Pantera L
Dodge Challenger
Merc 450SLC
AMC Javelin
Maserati Bora
Chrysler Imperial
Duster 360
Camaro Z28
Lincoln Continental
Cadillac Fleetwood

11 mpg
Stacked anything is nearly always a mistake.
This is much better...
The most important measurement should exploit the highest ranked encoding possible.

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or Color saturation/hue
Observation: Comparison is trivial on a common scale.
Today’s Learning Goals

• Appreciate the major elements of **exploratory data analysis** and why it is important to visualize data.

• Be conversant with **data visualization best practices** and understand how good visualizations optimize for the human visual system.

• Be able to generate informative graphical displays including **scatterplots**, **histograms**, **bar graphs**, **boxplots**, **dendrograms** and **heatmaps** and thereby gain exposure to the extensive graphical capabilities of R.

• Appreciate that you can build even more complex charts with **ggplot** and additional R packages such as **rgl**.
**Different graphs for different purposes**

**Exploratory graphs**: many images for a narrow audience (you!)

**Presentation graphs**: single image for a large audience

---

**Presentation**

** Exploration**
Core R Graph Types

- Scatter plot
- Bar Chart
- Histogram
- Boxplot
- Dot Chart
- Stripchart
- Pie Chart
- Smooth Scatter
- Cluster Dendrogram
The R Painters Model

Side-Note: “Red and green should never be seen”
Do it Yourself!

Hands-on

Section 1 only please

➡ Create a new **RStudio Project** for this class,
➡ **Download** the example data files and move them to your project directory,
➡ Focus on **Sections 1A & 1B** in the **handout**.
Hands-on

Section 1 only please

➡ Create a new **RStudio Project** for this class,
➡ **Download** the example data files and move them to your project directory,
➡ Focus on **Sections 1A & 1B** in the **handout**.
Common Options

- **Axis scales**
  - `xlim c(min, max)`
  - `ylim c(min, max)`

- **Axis labels**
  - `xlab(text)`
  - `ylab(text)`

- **Plot titles**
  - `main(text)`
  - `sub(text)`

- **Plot characters**
  - `pch(number)`
  - `cex(number)`

- Local options to change a specific plot
- Global options to affect all graphs
Plot Characters

```r
plot( 1:5, pch=1:5, cex=1:5 )
```
Plot Type Specific Options
Plot (scatterplots and line graphs)

- **Input:** Almost anything. 2 x Vectors
- **Output:** Nothing
- **Options:**
  - `type` l=line, p=point, b=line+point
  - `lwd` line width (thickness)
  - `lty` line type (1=solid, 2=dashed, 3=dotted etc.)

```r
plot( c(1:10)^2, typ="b", lwd=4, lty=3 )
```
Hands-on
Section 2 Notes

➡ Focus on Sections 2A & 2B in the lab handout.
➡ Try Section 2C if you have time.
➡ See notes on the following slides...
**Section 2B: Barplot (a.k.a. bar graphs)**

- **Input:** Vector (single) or Matrix (stack or group)
- **Output:** Bar centre positions
- **Options:**
  - `names.arg` Bar labels (if not from data)
  - `horiz=TRUE` Plot horizontally
  - `beside=TRUE` Plot multiple series as a group not stacked

```r
barplot(VADeaths, beside = TRUE)
```
Controlling plot area options with `par`
Par

- The `par()` function controls global parameters affecting all plots in the current plot area.

- Changes affect all subsequent plots.

- Many `par` options can also be passed to individual plots.

`?par`
Par examples

• Reading current value
  ‣ `old.par <- par()$mar`

• Setting a new value
  ‣ `par(mar=c(4,11,2,1))` # Do plot

• Restoring old value after you are done
  ‣ `par(mar=old.par)`
Margin values are set with a 4 element vector (bottom, left, top, right)

par( mar=c(5,4,4,2))
barplot(x)

par( mar=c(2,10,10,1))
barplot(x)
Par options

- **Margins**
  - `mai` (set margins in inches)
  - `mar` (set margins in number of lines)
  - `mex` (set lines per inch)
  - 4 element vector (bottom, left, top, right)

- **Warning**
  - Error in plot.new() : figure margins too large

```r
par( mar=c(2, 10, 1, 1) )
```
Par options

• Multi-panel
  • \texttt{par( mfrow=c(rows, cols) )}
Hands-on
Section 3 only please
Using Color
Specifying colors

• Controlled names
  ▶ `col=c("red", "green")` etc.
  ▶ `see colors()`

• Color by number
  ▶ `col=c(1, 2, 3)`
  ▶ Will give black, red, green etc.

• Hexadecimal strings string
  ▶ Of the form "#RRGGBB" where each of the pairs RR, GG, BB consists of two hexadecimal digits giving a value in the range 00 to FF:
    ▶ `#FF0000` (red)
    ▶ `#0000FF` (blue)
Built in color schemes

• Functions to generate colors
• Pass in the number of colors you want, e.g., to get 7 different colors:
  ‣ rainbow(7)
  ‣ heat.colors(7)
  ‣ cm.colors(7)
  ‣ terrain.colors(7)
  ‣ topo.colors(7)
  ‣ Etc.
Color Packages

• Color Brewer
  - Set of pre-defined, optimized palettes
  - `library(RColorBrewer)`
  - `brewer.pal(n_colours, palette)`

• ColorRamps
  - Create smooth palettes for ramped color
  - Generates a function to make actual color vectors
  - `colorRampPalette(c("red","white","blue"))`
  - `colorRampPalette(c("red","white","blue"))(5)`
Applying Color to Plots

• Vector of numbers or specified colors passed to the `col` parameter of a plot function

• Vector of factors used to divide the data
  – Colors will be taken from the set color `palette`
  – Can read or set using `palette()` function
    • `palette()`
    • `palette(brewer.pal(9,"Set1"))`

```r
plot(1:5, col=1:5, pch=15, cex=2)
```
Dynamic use of color

• Coloring by density
  – Pass data and palette to `densCols()`
  – Vector of colors returned

• See Lab Supplement (online):
  • Plotting with color in R

https://www.rdocumentation.org/packages/grDevices/versions/3.4.3/topics/densCols
Make a lab report!

• Open your previous class05 RStudio project (and your saved R script)

• Can you source your class05.R file to regenerate all your plots without error?

• If so you can now generate a nice HTML report of your work to date...

[Take 2-3 minutes]
Homework!

New DataCamp Assignments

• **RStudio IDE (Pt 1)**
• **Intermediate R**
  • Conditionals and Control Flow
  • Functions
  • Loops

Muddy Point Assessment Form Link

Useful new website: [https://www.data-to-viz.com/](https://www.data-to-viz.com/)