Eric Pitman Summer Workshop in Computational Science

5. Visualizing Data
DATA science

- Statistics
- Visualisation
- Pattern recognition
- Machine Learning
- Neurocomputing
- Data Mining
- & Database & Data process
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

factor(cut): Fair, Good, Very Good, Premium, Ideal

count

color: D, E, F, G, H, I, J
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

bw=1

bw=0.1

Density

Density

Density

x too small

x too big

x just right
Kernel Density Plot

plot(density(Cars93$RPM),
xlab="engine RPM",
main="density plot of engine RPM",
col="red")
Density Plot

Distribution of Gas Mileage with Number of Gears

Miles Per Gallon

Density

gear
- 3 gears
- 4 gears
- 5 gears
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")
• 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 $1.5 \times (\text{IQR})$
  – *Outliers*: beyond $1.5 \times (\text{IQR})$
Boxplot: Data Symmetry?

Mileage by Gear Number

- 3 gears
- 4 gears
- 5 gears

Miles per Gallon

35
30
25
20
15
10
5
0

3 gears
4 gears
5 gears
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!)

```
pie(table(Cars93$Horsepower))
```
Interlude

Complete plotting exercises.

Open in the RStudio source editor:

<workshop>/exercises/5-exercises-plotting-basic.R
Interlude++

When you have done your plotting exercises...

Read flowing-data’s “Rules not to Break”

http://flowingdata.com/2015/08/11/real-chart-rules-to-follow
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/