CS 5135/6035 Learning Probabilistic Models

Lecture 2b: Visualizing data with Gadfly

Gowtham Atluri

August 28, 2018

Visualization

- Histograms, Density plots
- Box plots, Violin plots
- Scatter plots
- Line plots
- Heatmaps
- Bar charts

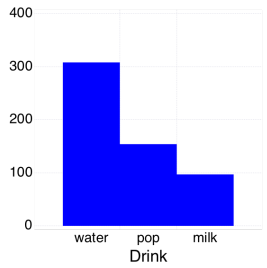
Gadfly - A Visualization Package

```
#Pkg.add("Gadfly"); #A plotting framework, based on ggplot2 fo
#Pkg.add("Cairo"); #Graphics package for output
#Pkg.add("Fontconfig");

using Gadfly, Cairo, Fontconfig;

white_panel = Theme(
    panel_fill=colorant"white",
    default_color=colorant"blue",
    major_label_font_size=26pt,
    minor_label_font_size=22pt,
    major_label_color=colorant"black",
    minor_label_color=colorant"black"
)
```

```
## Gadfly.Theme(RGB{N0f8}(0.0,0.0,1.0), 0.9mm, 0.45mm, 1.8mm,
```

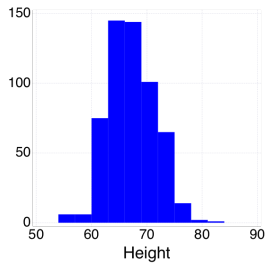

Histogram of categorical data: Drink

```
myplot = Gadfly.plot(data, x="Drink",
                     Geom.histogram, white_panel);
draw(PNG("./figs/hist_plot_a.png", 6inch, 6inch), myplot)
```



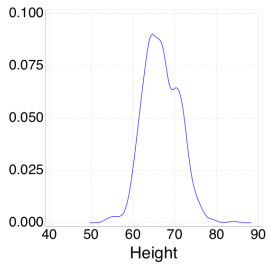
Histogram of continuous data: Height

```
myplot = Gadfly.plot(data, x="Height",
                     Geom.histogram(bincount=10),
                     white_panel);
draw(PNG("./figs/hist_plot_b.png", 6inch, 6inch), myplot)
```



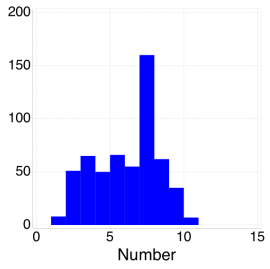
Density of continuous data: Height

```
myplot = Gadfly.plot(data, x="Height",
                     Geom.density,
                     white_panel);
draw(PNG("./figs/density_plot.png", 6inch, 6inch), myplot)
```



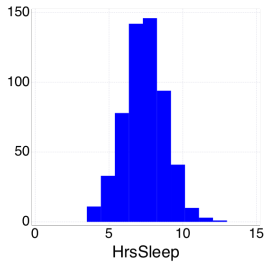
Histogram of continuous random data: Number

```
myplot = Gadfly.plot(data, x="Number",
                     Geom.histogram(bincount=10),
                     white_panel);
draw(PNG("./figs/hist_plot_c.png", 6inch, 6inch), myplot)
```



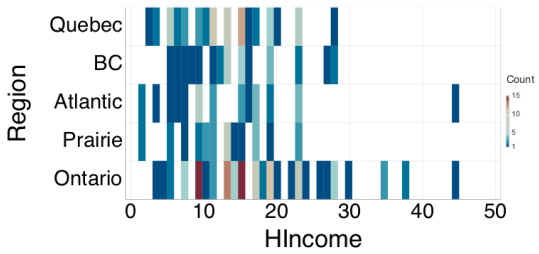
Histogram of hours of sleep

```
myplot = Gadfly.plot(data, x="HrsSleep",
                     Geom.histogram(bincount=10),
                     white_panel);
draw(PNG("./figs/hist_plot_d.png", 6inch, 6inch), myplot)
```



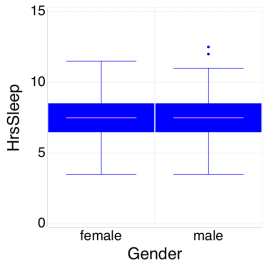
2D Histogram

```
#Pkg.add("RDatasets");
using RDatasets;
womenlf = dataset("car", "Womenlf");
myplot = plot(womenlf, x="HIncome", y="Region",
    Geom.histogram2d,white_panel);
draw(PNG("./figs/2d_hist_plot.png", 8inch, 4inch), myplot);
```



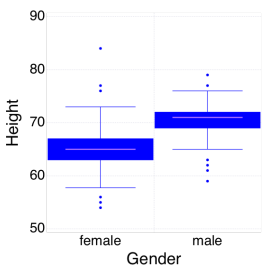
Box plot comparing hours of sleep for M and F

```
myplot = plot(data, x="Gender", y="HrsSleep",
              Geom.boxplot, white_panel);
draw(PNG("./figs/box_plot_a.png", 6inch, 6inch), myplot)
```



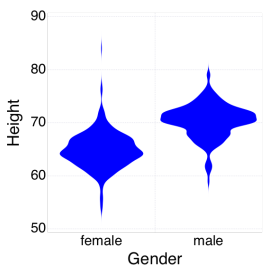
Box plot comparing height for M and F

```
myplot = plot(data, x="Gender", y="Height",
              Geom.boxplot, white_panel);
draw(PNG("./figs/box_plot_b.png", 6inch, 6inch), myplot)
```



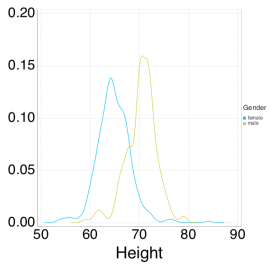
Violin plot comparing height for M and F

```
myplot = plot(data, x="Gender", y="Height",
              Geom.violin, white_panel);
draw(PNG("./figs/violin_plot.png", 6inch, 6inch), myplot)
```



Density plot for comparing height for M and F

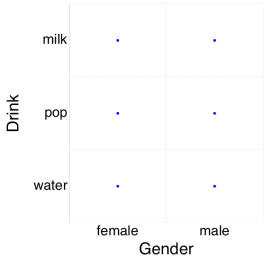
```
myplot = plot(data, x="Height", color="Gender",
              Geom.density, white_panel);
draw(PNG("./figs/box_plot_c_density.png", 6inch, 6inch), myplot)
```



Scatter plot - binary values

- Relationship between :Gender and :Drink

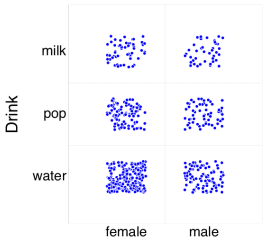
```
myplot = Gadfly.plot(data, x=:Gender, y=:Drink,
                     Geom.point, white_panel);
draw(PNG("./figs/scatter_plot_a.png", 6inch, 6inch), myplot)
```



Scatter plot - binary values (jitter)

- Relationship between :Gender and :Drink

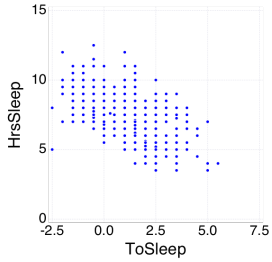
```
myplot = Gadfly.plot(data, x=:Gender, y=:Drink,
                     Stat.x_jitter(range=0.5,seed=10),
                     Stat.y_jitter(range=0.5,seed=25),
                     Geom.point, white_panel);
draw(PNG("./figs/scatter_plot_b.png", 6inch, 6inch), myplot)
```



Scatter plot - continuous values

- Relationship between :ToSleep and :HrsSleep

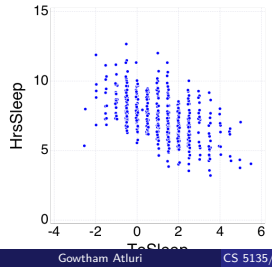
```
myplot = Gadfly.plot(data, x=:ToSleep, y=:HrsSleep,
                     Geom.point, white_panel);
draw(PNG("./figs/scatter_plot_c.png", 6inch, 6inch), myplot)
```



Scatter plot - continuous values (jitter)

- Relationship between :ToSleep and :HrsSleep

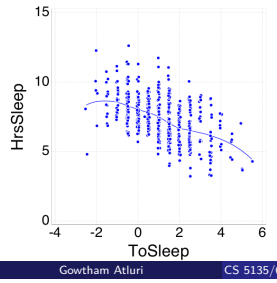
```
myplot = Gadfly.plot(data, x=:ToSleep, y=:HrsSleep,
                     Stat.x_jitter(range=0.5),
                     Stat.y_jitter(range=0.5),
                     Geom.point, white_panel);
draw(PNG("./figs/scatter_plot_d.png", 6inch, 6inch), myplot)
```



Scatter plot - continuous values (smooth fit)

- Relationship between :ToSleep and :HrsSleep

```
# Geom.smooth - a smooth function estimated from the data
myplot = Gadfly.plot(data, x=:ToSleep, y=:HrsSleep,
                     Stat.x_jitter(range=0.5,seed=10),
                     Stat.y_jitter(range=0.5,seed=20),
                     Geom.point, Geom.smooth, white_panel);
draw(PNG("./figs/scatter_plot_e.png", 6inch, 6inch), myplot)
```



Scatter plot - continuous values (line fit, layers in Gadfly)

- Relationship between :ToSleep and :HrsSleep

```
#Pkg.add("GLM")
using GLM;
ols = lm(@formula(HrsSleep ~ ToSleep), data)
```

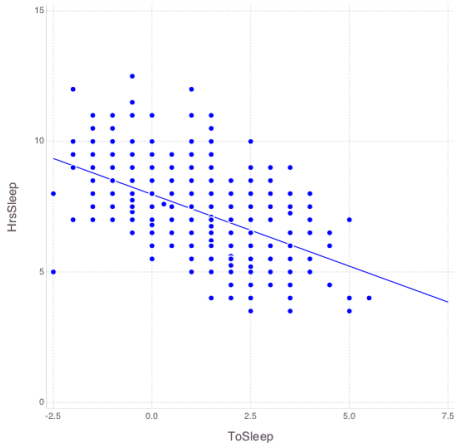
```
## StatsModels.DataFrameRegressionModel{GLM.LinearModel{GLM.Lm
##
## Formula: HrsSleep ~ 1 + ToSleep
##
## Coefficients:
##              Estimate Std.Error  t value Pr(>|t|)
## (Intercept)   7.97442 0.0681373  117.035   <1e-99
## ToSleep      -0.55519 0.0397884 -13.9536   <1e-37
```


Scatter plot - continuous values (line fit, layers in Gadfly)

- Relationship between :ToSleep and :HrsSleep

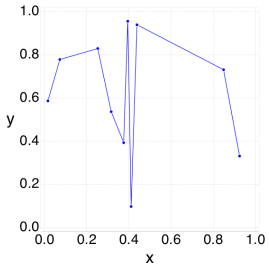
```
t = [-2.5;7.5];
myplot = plot(
  layer(data,x=:ToSleep, y=:HrsSleep, Geom.point,white_panel),
  layer(data,x=t, y= 7.97 - 0.55 * t , Geom.line,white_panel),
  Theme(panel_fill="white")
);
draw(PNG("./figs/scatter_plot_f.png", 6inch, 6inch), myplot)
```


Scatter plot - continuous values (line fit, layers in Gadfly)



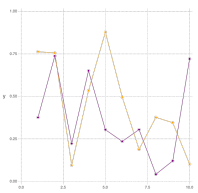
Line plots

```
myplot = plot(x=rand(10), y=rand(10), Geom.point,
              Geom.line, white_panel);
draw(PNG("./figs/line_plot_a.png", 6inch, 6inch), myplot)
```



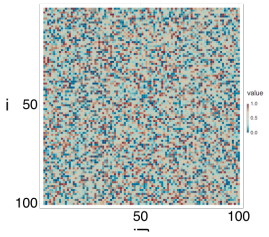
Line plots

```
myplot = plot(layer(x=collect(1:10), y=rand(10),Geom.point,
            Geom.line, Theme(default_color=colorant"orange")
            ),
      layer(x=collect(1:10), y=rand(10),Geom.point,
            Geom.line, Theme(default_color=colorant"purple")
            )
      );
draw(PNG("./figs/line_plot_b.png", 6inch, 6inch), myplot)
```



Heatmap

```
rand_data = rand(100,100);
myplot = spy(rand_data, white_panel);
draw(PNG("./figs/spy_rand.png", 6inch, 6inch), myplot);
```



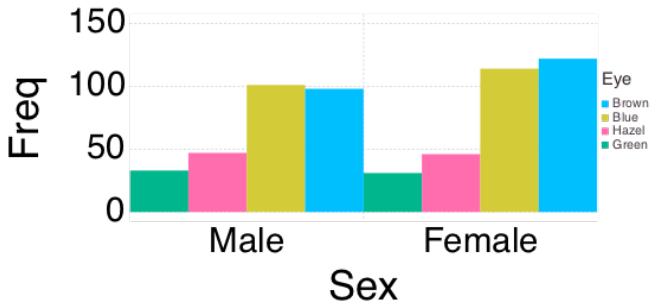
Bar plot

```
df = by(dataset("datasets","HairEyeColor"),[:Eye,:Sex], d->sum(d[:Freq]))
```

```
## 8×3 DataFrames.DataFrame
##   Row   Eye     Sex      x1
##
##   1     Brown   Male     98
##   2     Blue    Male     101
##   3     Hazel   Male     47
##   4     Green   Male     33
##   5     Brown   Female   122
##   6     Blue    Female   114
##   7     Hazel   Female   46
##   8     Green   Female   31
```

```
myplot = plot(df, color="Eye", y="x1", x="Sex",
    Geom.bar(position=:dodge), Guide.ylabel("Freq"),
    white_panel);
draw(PNG("./figs/bar_plot_a.png", 6inch, 3inch), myplot);
```


Bar plot



Bar plot - vertical stacking

```
D = by(dataset("datasets","HairEyeColor"),
       [:Eye,:Sex], d->sum(d[:Freq]))
```

```
## 8×3 DataFrames.DataFrame
##   Row   Eye     Sex      x1
##
##   1     Brown   Male     98
##   2     Blue    Male     101
##   3     Hazel   Male     47
##   4     Green   Male     33
##   5     Brown   Female   122
##   6     Blue    Female   114
##   7     Hazel   Female   46
##   8     Green   Female   31
```

```
rename!(D, :x1, :Frequency)
```


Bar plot - vertical stacking

```
palette=["brown","blue","tan","green"];
myplot= plot(D, x=:Sex, y=:Frequency, color=:Eye,
             Geom.bar(position=:stack),
             Scale.color_discrete_manual(palette...),
             white_panel);
draw(PNG("./figs/bar_plot_b.png", 6inch, 6inch), myplot);
```


Bar plot - vertical stacking