Introduction to ggplot

36-290 – Statistical Research Methodology

Week 2 Thursday – Fall 2021

Preliminaries

Let's read in the data frame that we use in the `dplyr` notes set:

```
df = read.csv("http://www.stat.cmu.edu/~pfreeman/GalaxyMass.csv")
```

As a reminder, these data consist of 3456 rows and 10 columns with names

```
names(df)
```

```
##  [1] "field"  "Gini"   "M20"    "C"      "A"      "size"   "n"      "q"      "z.mode" "mass"
```

ggplot

`ggplot` (actually, and perhaps confusingly, `ggplot2`) is "a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data [frame], tell `ggplot2` how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details."

Sounds good. Let's dive in:

```
library(ggplot2)    # also loaded as part of the tidyverse
```

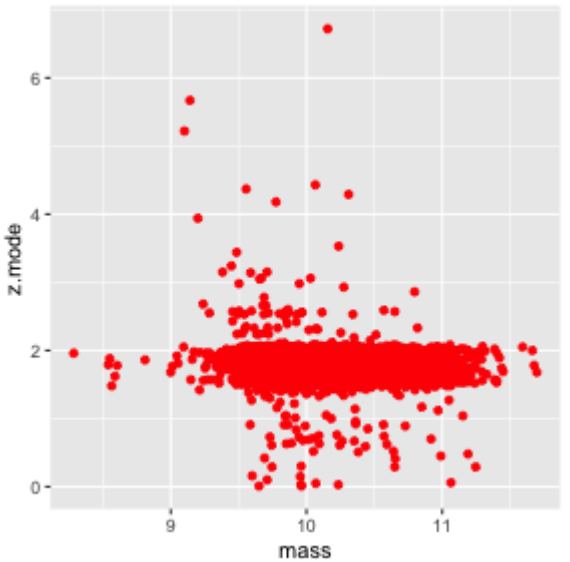
ggplot: Basic Structure

A very basic call to `ggplot()` has the following structure:

```
ggplot(data=<data frame>,mapping=aes(x=<x axis variable>,...)) + geom_<plot type>(<arguments>)
```

For instance, to plot `z.mode` vs. `mass` (and remember: we plot y vs. x):

```
ggplot(data=df,mapping=aes(x=mass,y=z.mode)) + geom_point(color="red")
```

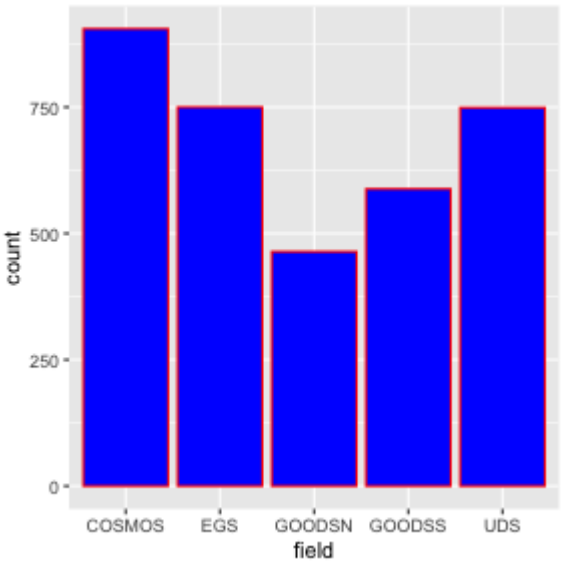


We'll talk more about two-dimensional scatter plots and such in another set of notes. Here, we'll concentrate on univariate (one-dimensional) plots.

ggplot: Bar Chart

How many galaxies are in each field?

```
ggplot(data=df,mapping=aes(x=field)) + geom_bar(color="red",fill="blue")
```

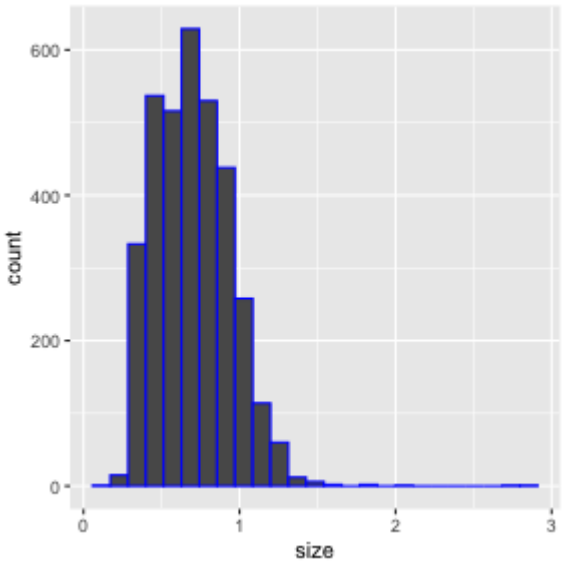


(A bar chart is appropriate when the x-axis variable is categorical and the y-axis variable is quantitative.)

ggplot: Histogram

What is the distribution of galaxy sizes?

```
ggplot(data=df,mapping=aes(x=size)) + geom_histogram(color="blue",bins=25)
```

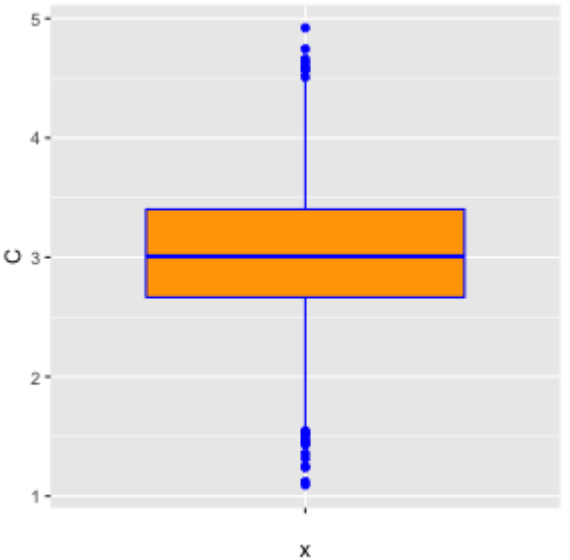


(A histogram is appropriate when the single variable in question is quantitative.)

ggplot: Boxplot

Boxplots are just a bit trickier. What is the distribution of galaxy concentrations?

```
ggplot(data=df,mapping=aes(x="",y=C)) + geom_boxplot(color="blue",fill="orange")
```

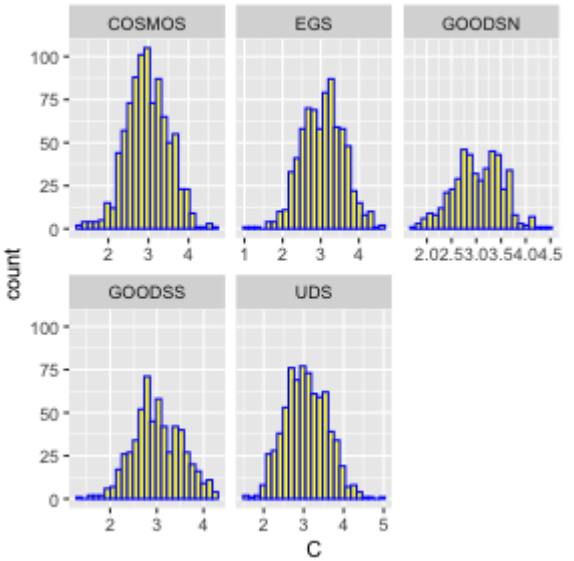


(A boxplot is also for visualizing a quantitative variable.)

ggplot: Faceting

Faceting is the act of making multiple plots at once that appear side-by-side as "facets". Faceting is something you might want to do when, e.g., you have a factor variable. Here, we show histograms of the concentration variable `C` broken up by galaxy field.

```
ggplot(data=df,mapping=aes(x=C)) + geom_histogram(color="blue",fill="yellow",bins=25) +
  facet_wrap(~field,scales='free_x')
```



ggplot: Gather

`gather()` is a function (from the `tidyr` package) that takes a data frame and realigns it. It is best illustrated via a simple example. Let's say we have the following data frame, which we'll call `df`:

```
> df
     x   y
1  0.5 0.7
2  1.2 1.9
```

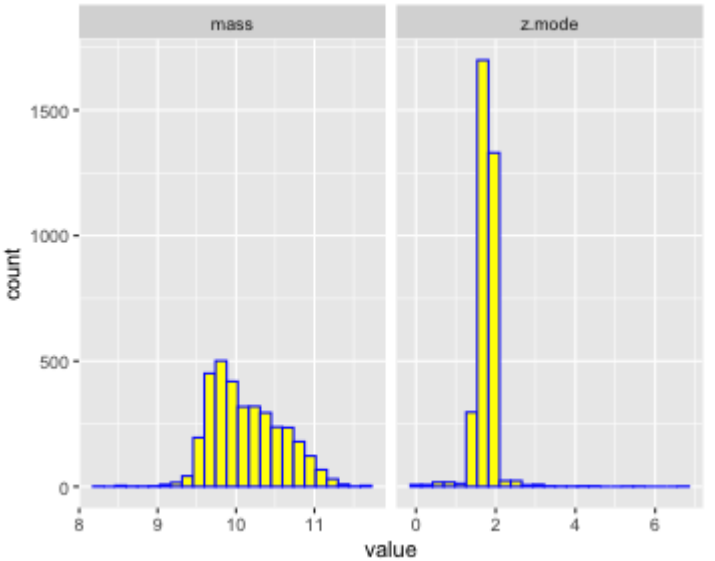
If we "gather" these data, we get the following:

```
> library(tidyr)
...
> gather(df)
      key  value
1       x    0.5
2       x    1.2
3       y    0.7
4       y    1.9
```

Combining `gather()` with faceting allows one to, e.g., visualize multiple variables at once.

ggplot: Gather (+ dplyr)

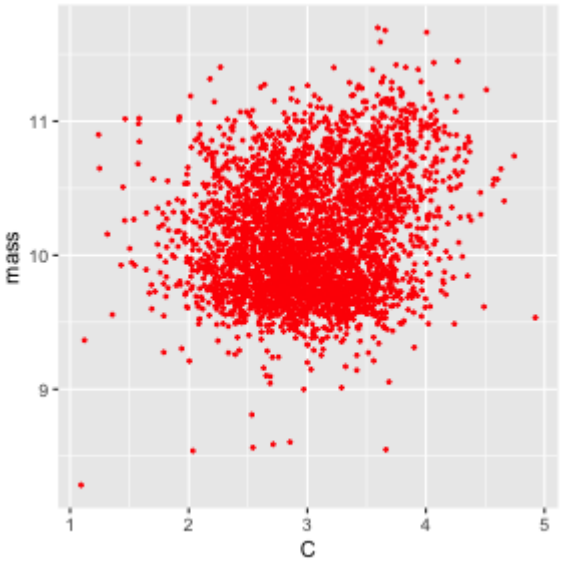
```
suppressMessages(library(tidyr))
suppressMessages(library(magrittr))
suppressMessages(library(dplyr))

df.new = df %>% select(.,z.mode,mass) %>% gather(.)
ggplot(data=df.new,mapping=aes(x=value)) + geom_histogram(color="blue",fill="yellow",bins=25) +
  facet_wrap(~key,scales='free_x')
```



ggplot: Scatter Plot

Let's get a sense of the data by plotting `mass` vs. `C` in a scatter plot:

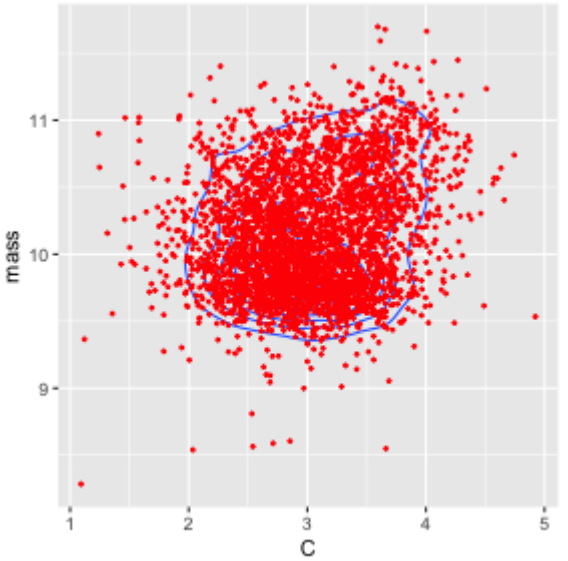
```
ggplot(data=df,mapping=aes(x=C,y=mass)) + geom_point(color="red",size=0.5)
```



ggplot: Scatter Plot with Density

By using `geom_density_2d()` we can overlay contours that indicate the density of points:

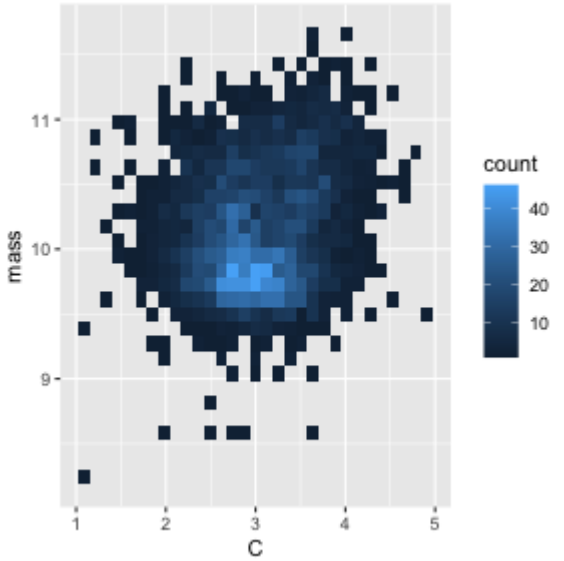
```
ggplot(data=df,mapping=aes(x=C,y=mass)) + geom_density_2d() + geom_point(color="red",size=0.5)
```



ggplot: Scatter Plot with Bins

By using `geom_bin2d` we can overlay bins whose color indicates the density of points:

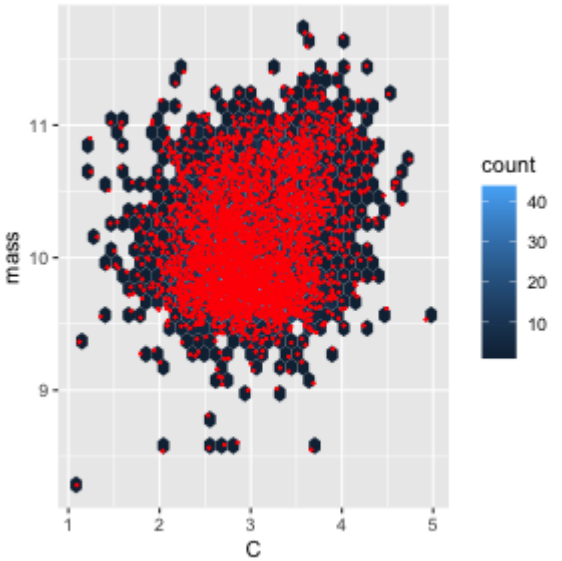
```
ggplot(data=df,mapping=aes(x=C,y=mass)) + geom_bin2d()
```



ggplot: Scatter Plot with Hexagonal Bins

What about hexagonal bins?

```
library(hexbin)
ggplot(data=df,mapping=aes(x=C,y=mass)) + geom_hex() + geom_point(color="red",size=0.25)
```

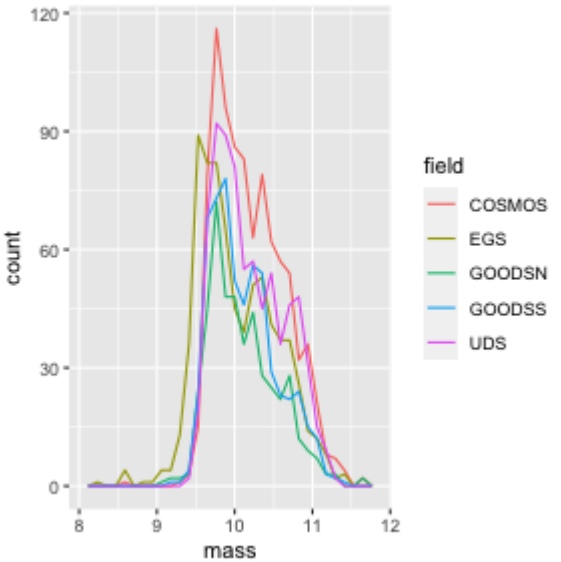


ggplot: freqpoly

`geom_freqpoly()` is a means by which to overlay empirical distributions of data in one plot pane. For instance, what is the distribution of `mass` as a function of `field`?

```
ggplot(data=df,mapping=aes(x=mass)) + geom_freqpoly(mapping=aes(color=field))
```

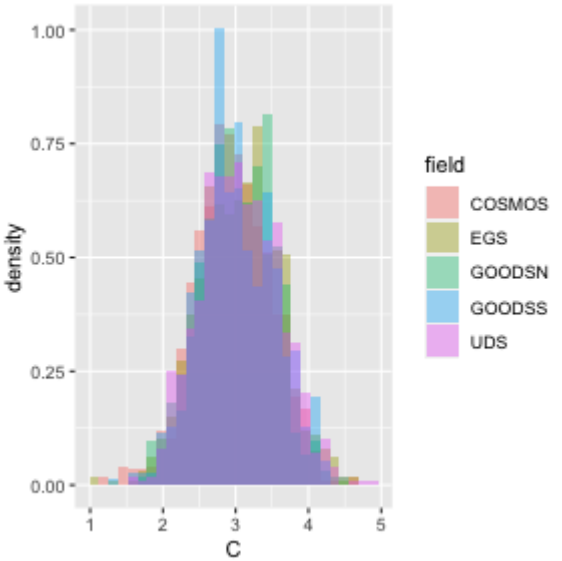
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



ggplot: Overlaid Histograms

```
ggplot(data=df,mapping=aes(x=C,fill=field,stat(density))) +
  geom_histogram(alpha=0.4,position="Identity")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In this case: not particularly helpful. `freqpoly` is better here.

ggplot: Side-By-Side Boxplots

```
ggplot(data=df,mapping=aes(x=field,y=mass)) + geom_boxplot(fill="FireBrick2")
```