Beyond ggplot: Extra Visualization Resources

36-290 – Statistical Research Methodology

Week 3 Tuesday – Fall 2021

Preliminaries

Let's read in the data frame that we use in the `dplyr` and `ggplot` notes sets:

```
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"
##  [9] "z.mode" "mass"
```

Covariance and Correlation

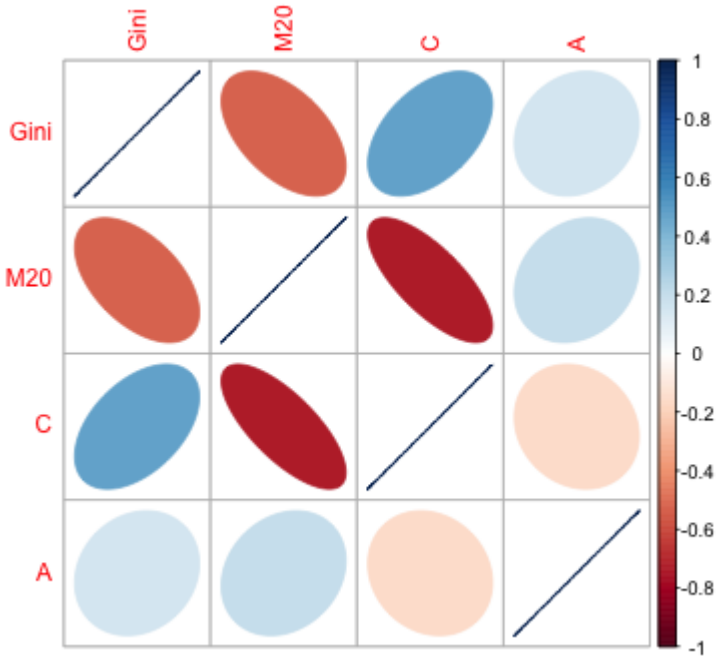
Covariance is a measure of the *linear* dependence between two variables. (To be "uncorrelated" is not the same as to be "independent"...the latter means there is no dependence, linear or otherwise, between two variables.) Correlation is a "normalized" form of covariance, that ranges from -1 (one variable linearly decreases absolutely in value while the other increases in value) through 0 (no linear dependence) to 1 (one variable linear increases absolutely while the other increases).

A good package for visualizing correlation is `corrplot`.

corrplot

Here we generate a correlation plot for the galaxy morphological variables `Gini`, `M20`, `C`, and `A`:

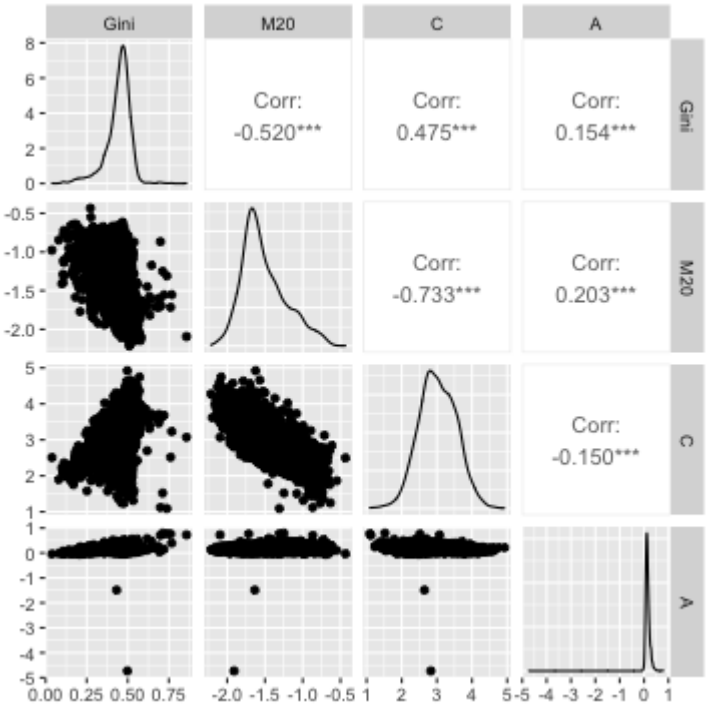
```
library(corrplot)
df %>% dplyr::select(.,Gini,M20,C,A) %>% cor(.) %>% corrplot(.,method="ellipse")
```



GGally: Pairs Plots

We step outside canonical `ggplot` plotting for now to bring you a pairs plot from the `GGally` package:

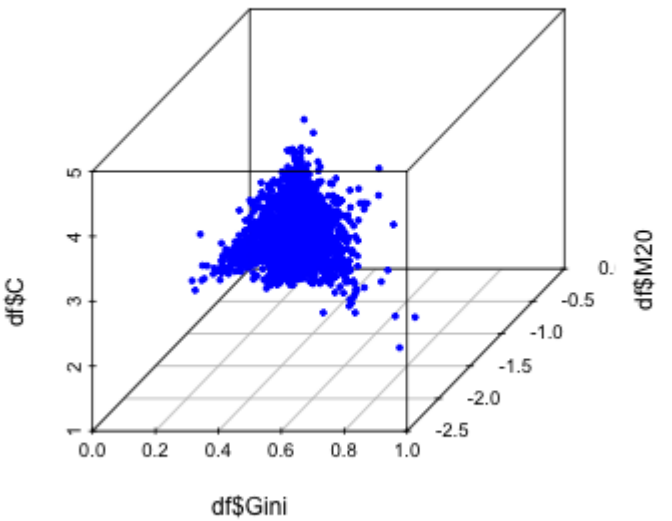
```
suppressMessages(library(GGally))
df %>% dplyr::select(.,Gini,M20,C,A) %>%
  ggpairs(.,progress=FALSE,lower=list(combo=wrap("facethist", binwidth=0.8)))
```



3D Scatter Plot

We now look at methods for visualizing data in three and more dimensions. The first is via the `scatterplot3d` package:

```
library(scatterplot3d)
scatterplot3d(x=df$Gini,y=df$M20,z=df$C,pch=19,
              color="blue",angle=45,cex.symbols=0.5)
```



Parallel Coordinates

The `parcoord()` function in the `MASS` package is one mechanism through which we can attempt to visualize more than three variables at once. Each line represents a single object.

```
suppressMessages(library(MASS))
z.color = round(64*(df$Gini-min(df$Gini))/(max(df$Gini)-min(df$Gini)))
palette(rainbow(64))
parcoord(df[,c("Gini","M20","C","A")],col=z.color,lwd=0.4)
```