Variation on Scatterplots

Diardano Raihan

2022-07-19

Import the Libraries and Dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(ggthemes)
library(RColorBrewer)
cces <- read_csv("cces_sample_coursera.csv")
```

```
## Rows: 1000 Columns: 25

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (25): caseid, region, gender, educ, edloan, race, hispanic, employ, mars...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
cel <- read_csv("cel_dataset_coursera.csv")
```

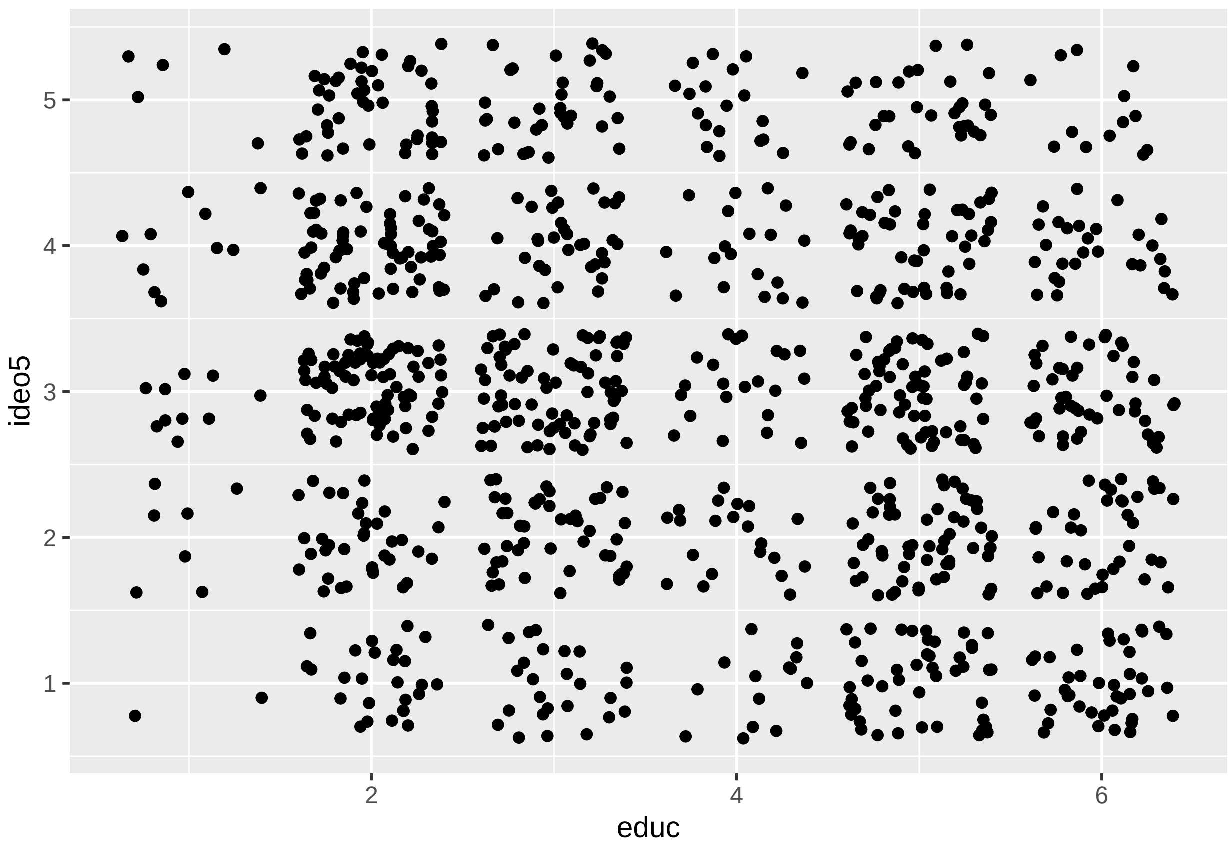
```
## Rows: 10262 Columns: 38
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): thomas_name, st_name
## dbl (36): thomas_num, icpsr, congress, year, cd, dem, elected, female, votep...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Variations on Scatter Plots

Add a best-fit line with geom_smooth

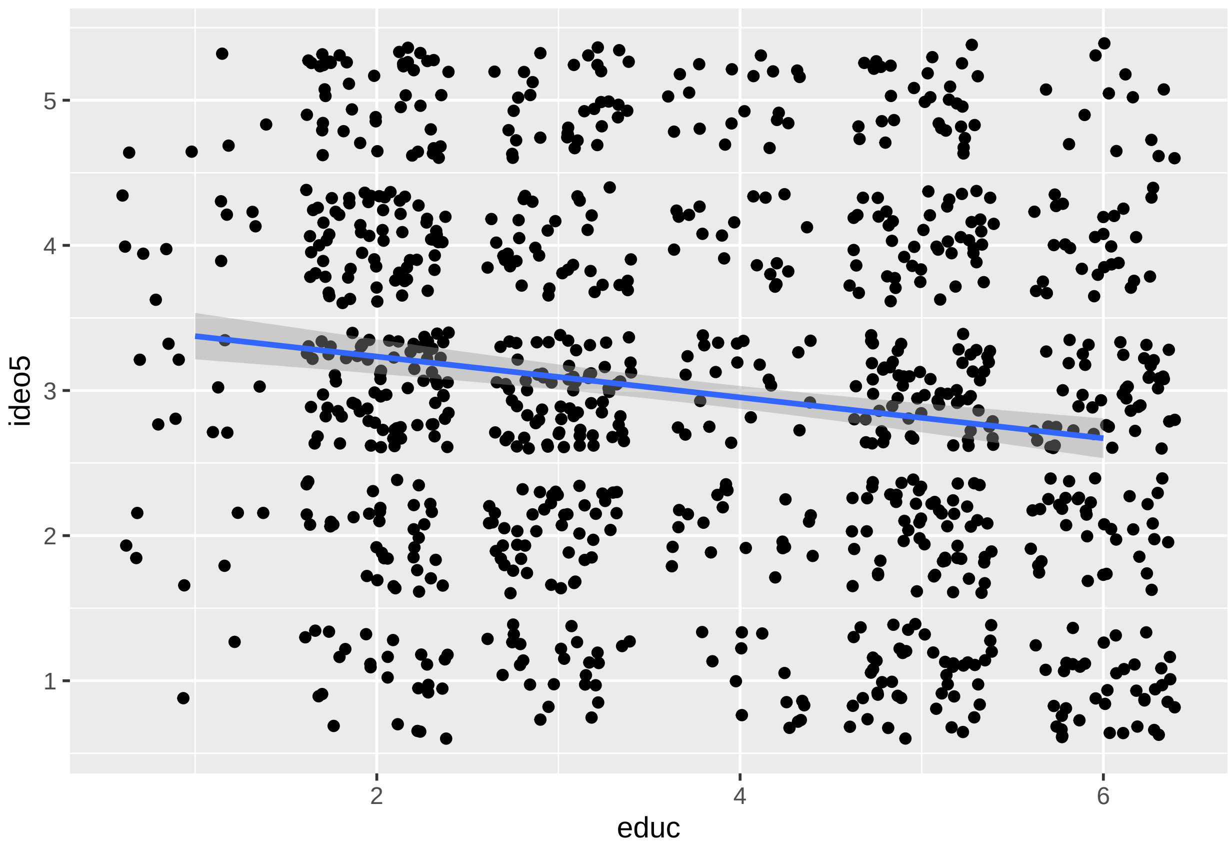
Reference: - Adapted from https://r-graphics.org/recipe-scatter-fitlines - Adapted from https://ggplot2.tidyverse.org/reference/geom_smooth.html

```
# Let's see the normal scatter plot look like
ggplot(cces, aes(x = educ, y = ideo5))+
  geom_jitter()
```



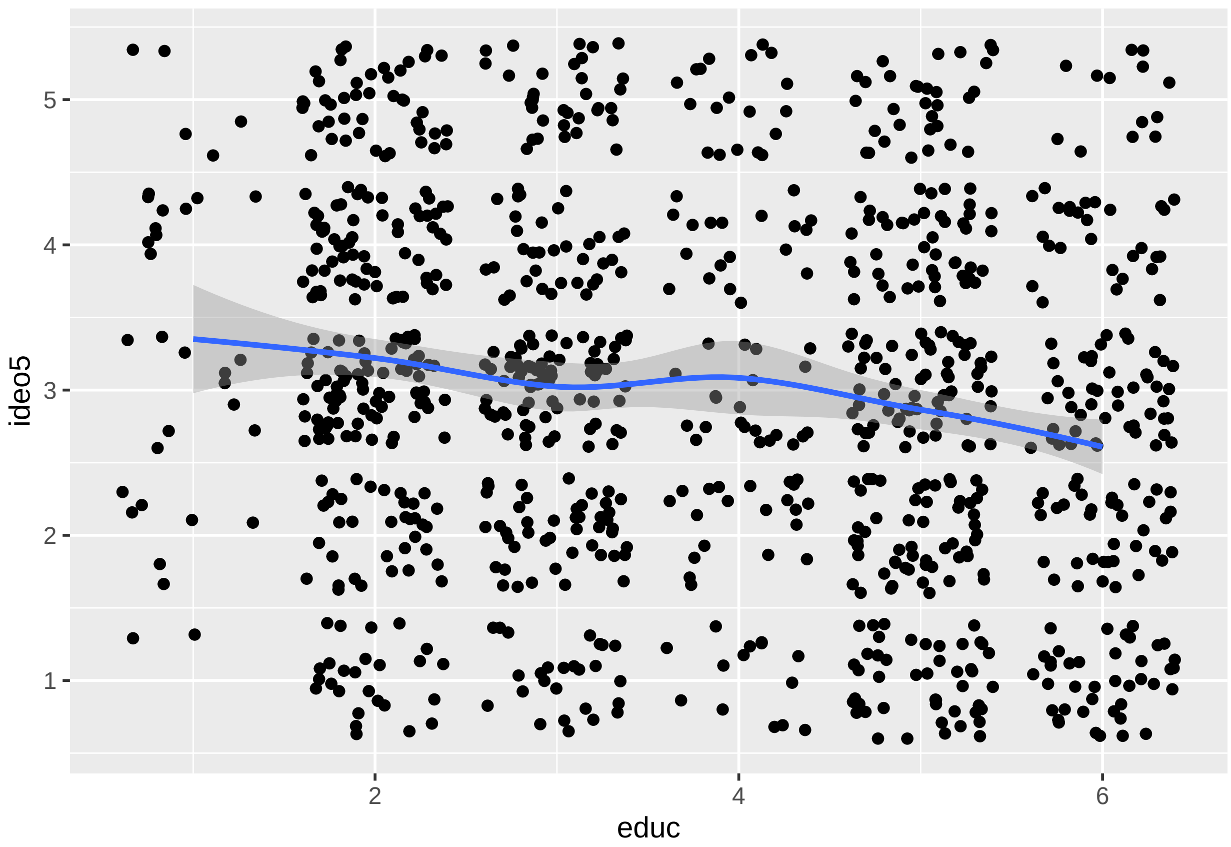
```
# Add a linear model "lm" that fit the plot
ggplot(cces, aes(x = educ, y = ideo5))+
  geom_jitter()+
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



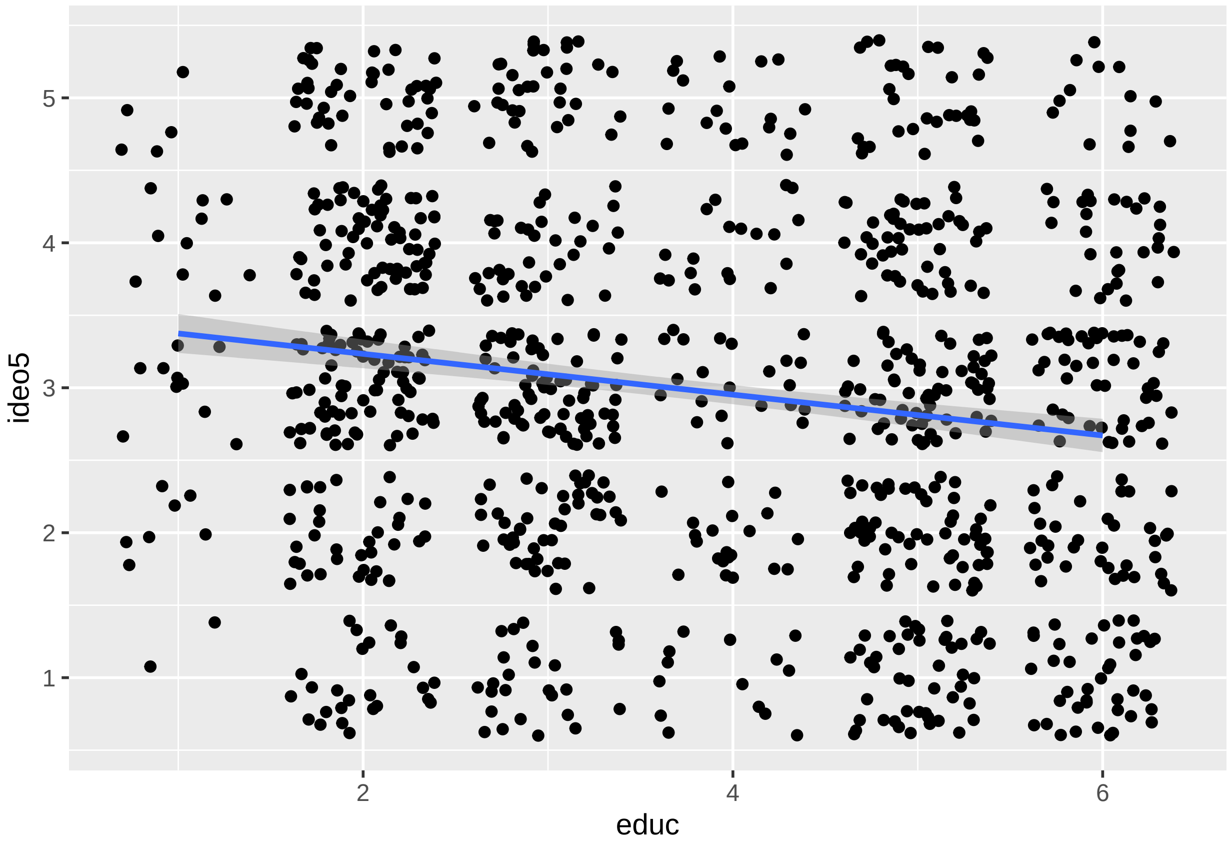
```
# Add a loess curve moder "lm" that fit the plot
ggplot(cces, aes(x = educ, y = ideo5))+
  geom_jitter()+
  geom_smooth(method = "loess")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```r
# Change the confidence level
ggplot(cces, aes(x = educ, y = ideo5))+
  geom_jitter()+
  geom_smooth(method = "lm", level = .9)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



- Be very careful with changing the confidence level if you're not statistically inclined, because doing statistical inference with this kind of approach is pretty dicey.
- Exploratory analysis is fine so if you want to draw some best fit lines to your scatter plot to get a sense for the data just sort of an intuition is an okay thing to do.

Scatter Plot Matrix

- Another use of scatter plot for EDA is to use a scatter plot matrix which will show you the bivariate relationship several different variables in your data at once.
- We can achieve this by using `ggplot` function.
  - Reference: https://rkabacoff.github.io/datavis/Other.html#scatterplot-matrix

```
# install.packages("GGally")
library(GGally)
```

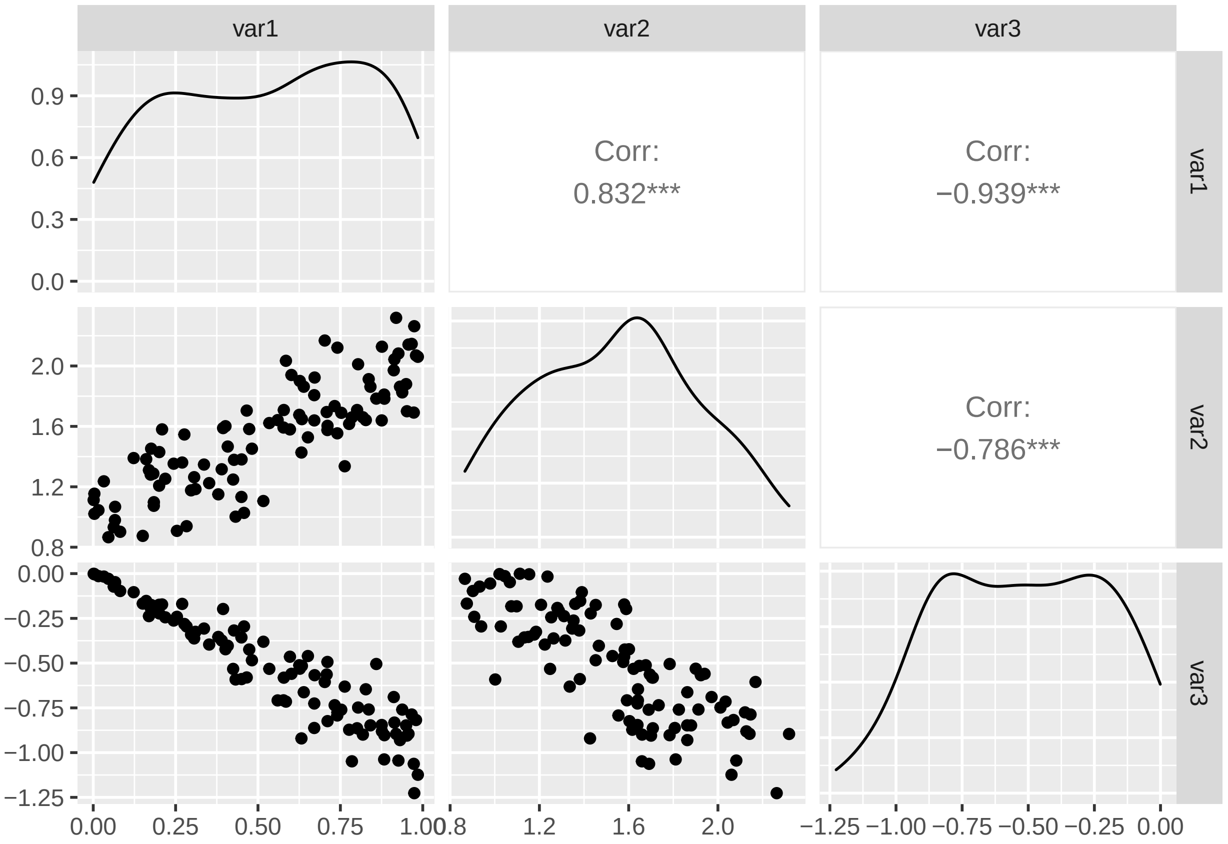
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
# Make up some numeric data.
# Two variables will be positively correlated, and the third will be negatively correlated with the fir
var1 <- runif(100, 0, 1)
```

```
var2 <- var1 + rnorm(100, 1, .2)
var3 <- var1*(-rnorm(100, 1, .2))

df <- tibble(var1, var2, var3)
df
```

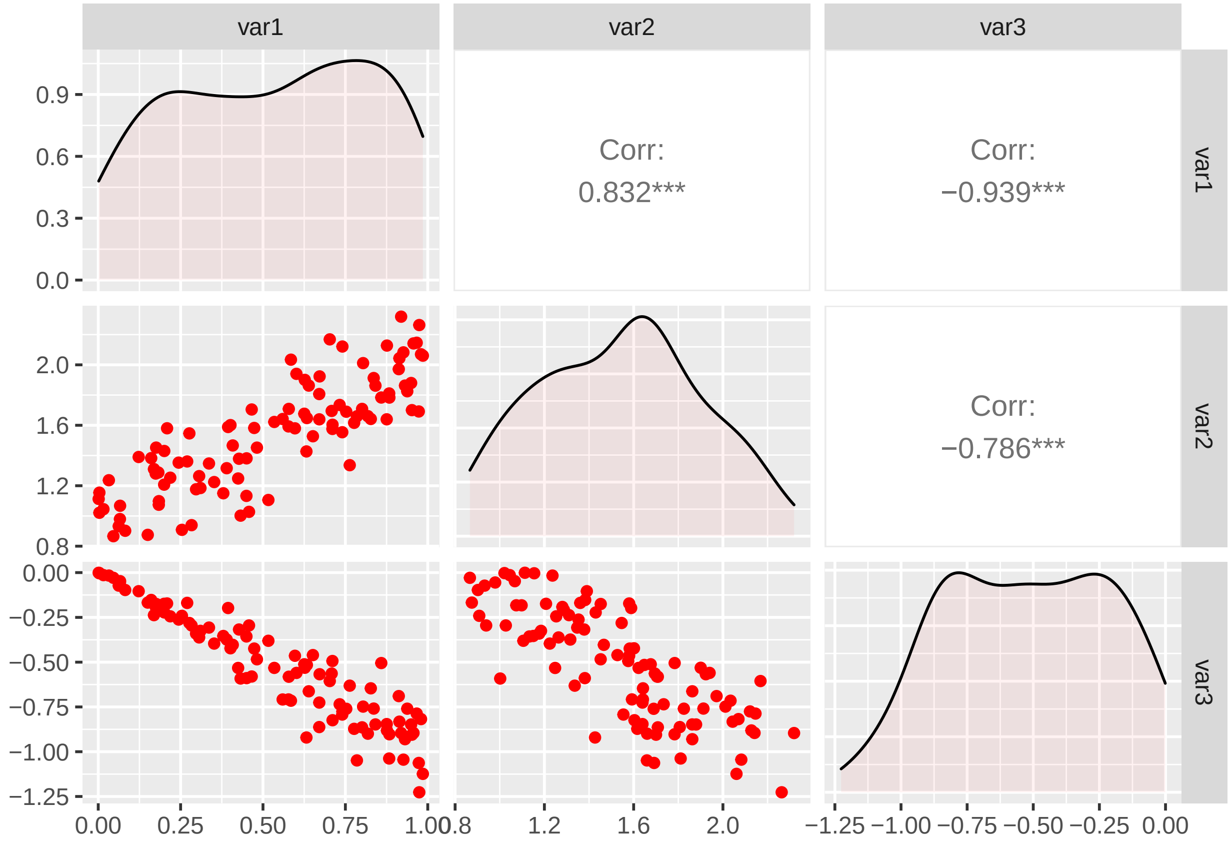
```
## # A tibble: 100 x 3
##      var1  var2    var3
##     <dbl> <dbl>   <dbl>
##  1 0.639  1.86  -0.663
##  2 0.270  1.36  -0.169
##  3 0.534  1.62  -0.532
##  4 0.883  1.81  -1.04
##  5 0.0460 0.866 -0.0290
##  6 0.652  1.53  -0.460
##  7 0.577  1.59  -0.708
##  8 0.450  1.13  -0.357
##  9 0.818  1.66  -0.899
## 10 0.801  1.71  -0.864
## # ... with 90 more rows
```

```
# Let's plot the scatterplot matrix using ggpair()
ggpairs(df)
```



Customize the Matrix Figures by Custom Function

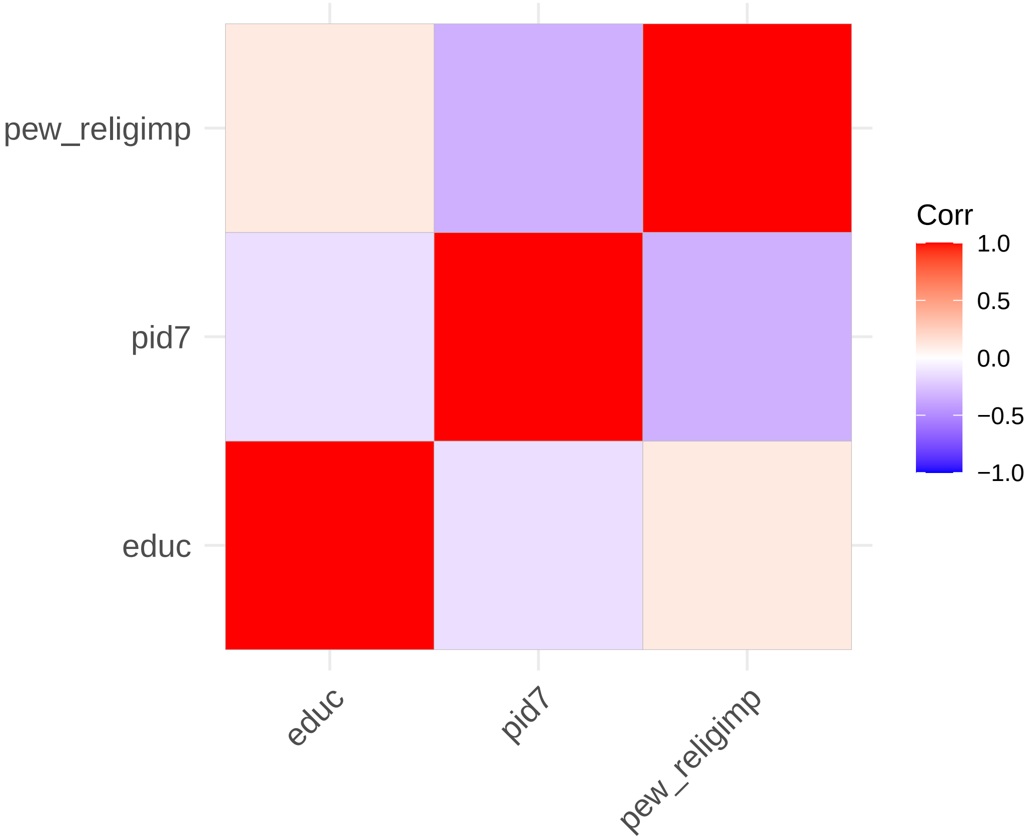
```r
# Write your own function for the scatterplot
my_scatter <- function(data, mapping){
  ggplot(data = data, mapping = mapping)+
    geom_jitter(color = "red")
}

# Write your own function for density plot
my_density <- function(data, mapping){
  ggplot(data = data, mapping = mapping)+
    geom_density(alpha = .05,
                 fill = "red")
}

# Substitute your functions for the functions that ggpairs() uses to draw the figures
ggpairs(df,
        lower = list(continuous = my_scatter),
        diag = list(continuous = my_density))
```



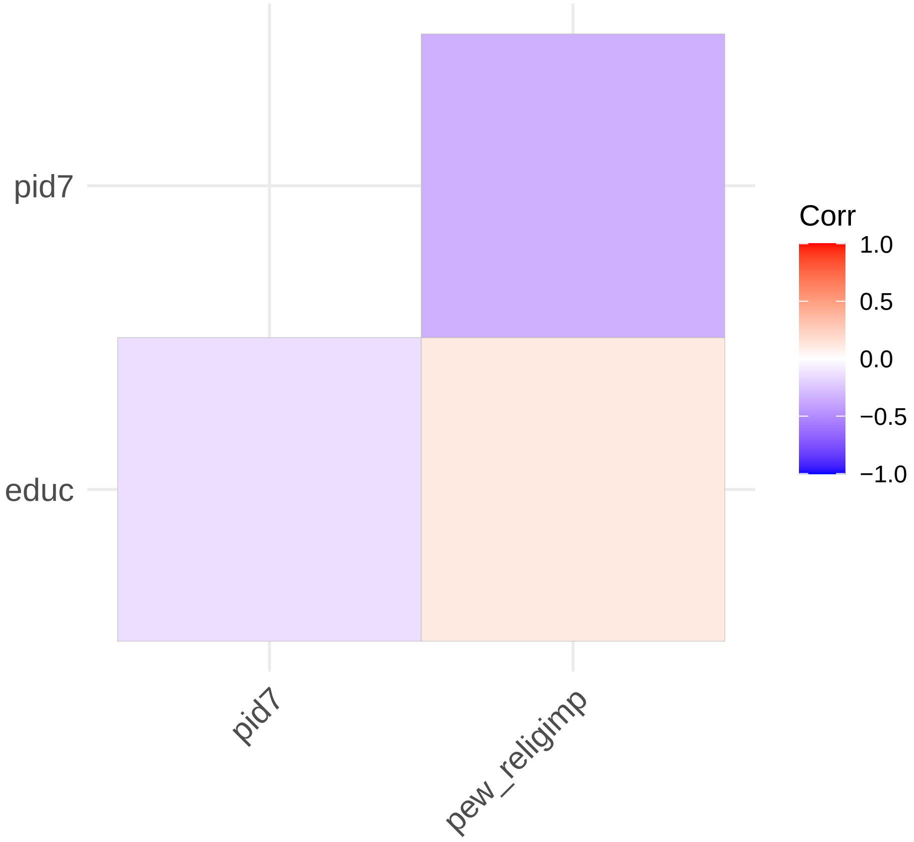
Correlation Plots

- Reference: https://rkabacoff.github.io/datavis/Models.html#Corrplot

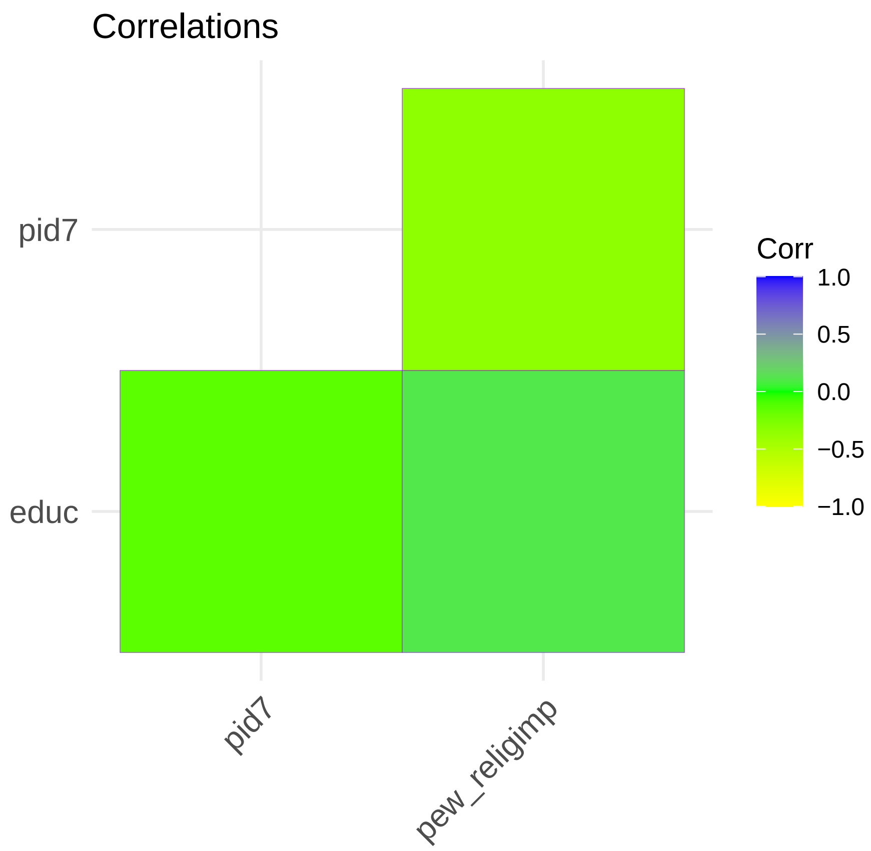
```r
# install.packages("ggcorrplot")
library(ggcorrplot)
library(ggthemes)

# Let's use some example survey data
df <- cces %>% select("educ", "pid7", "pew_religimp")

# Calculate correlation coefficients
r <- cor(df, use = 'complete.obs')
?cor

# generate the correlation plot
ggcorrplot(r)
```



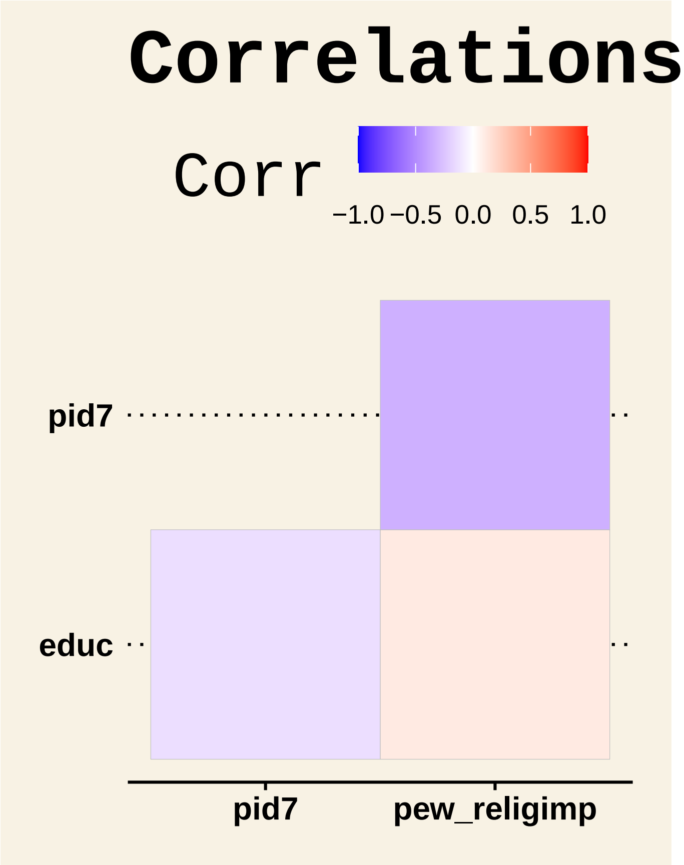
```r
# show just the lower part of the figure (to avoid 1 correlations on the diagonal)
ggcorrplot(r, type = "lower")
```



```
# modify some visual elements
ggcorrplot(r, type = "lower",
           title = "Correlations",
           colors = c("yellow", "green", "blue"),
           outline.color = "purple")
```



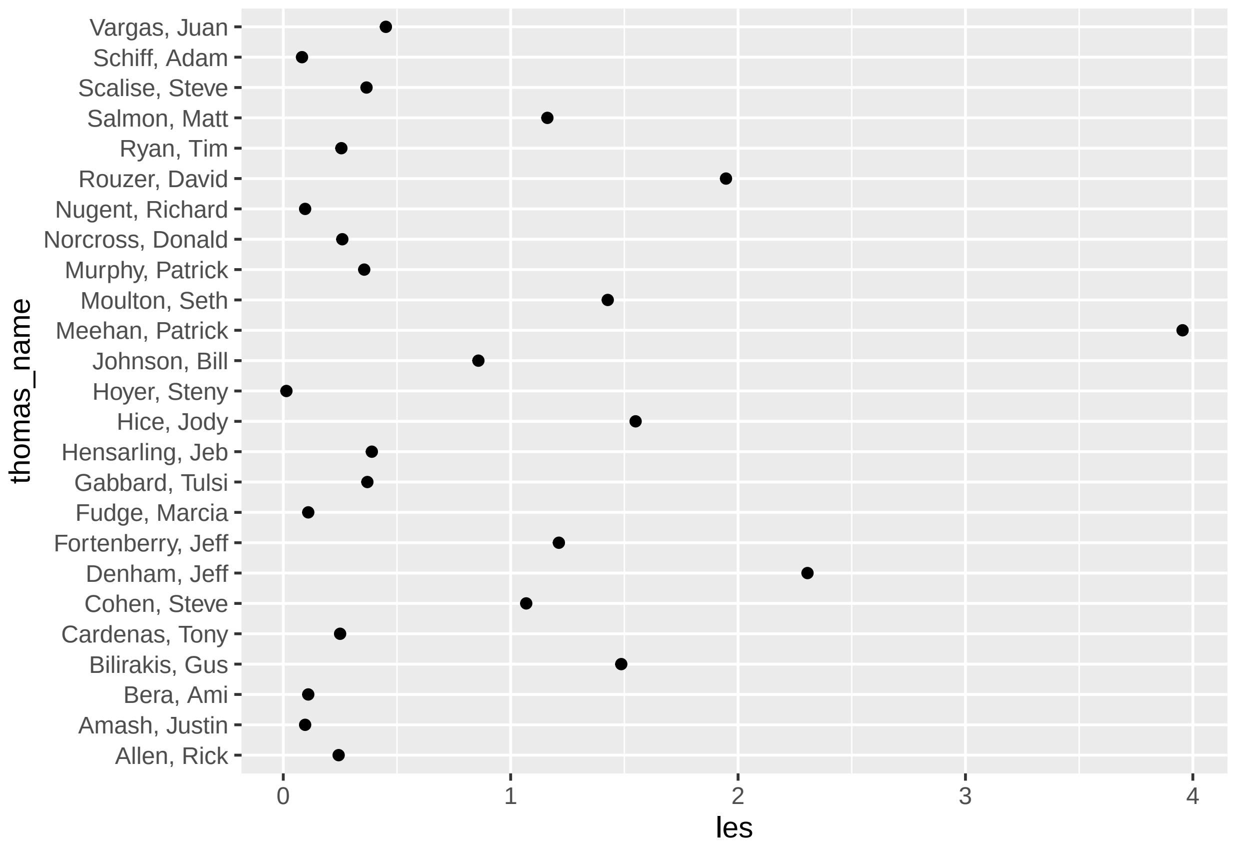
```
# use ggtheme is you want
ggcorrplot(r, type = "lower",
           title = "Correlations")+
  theme_wsj()
```



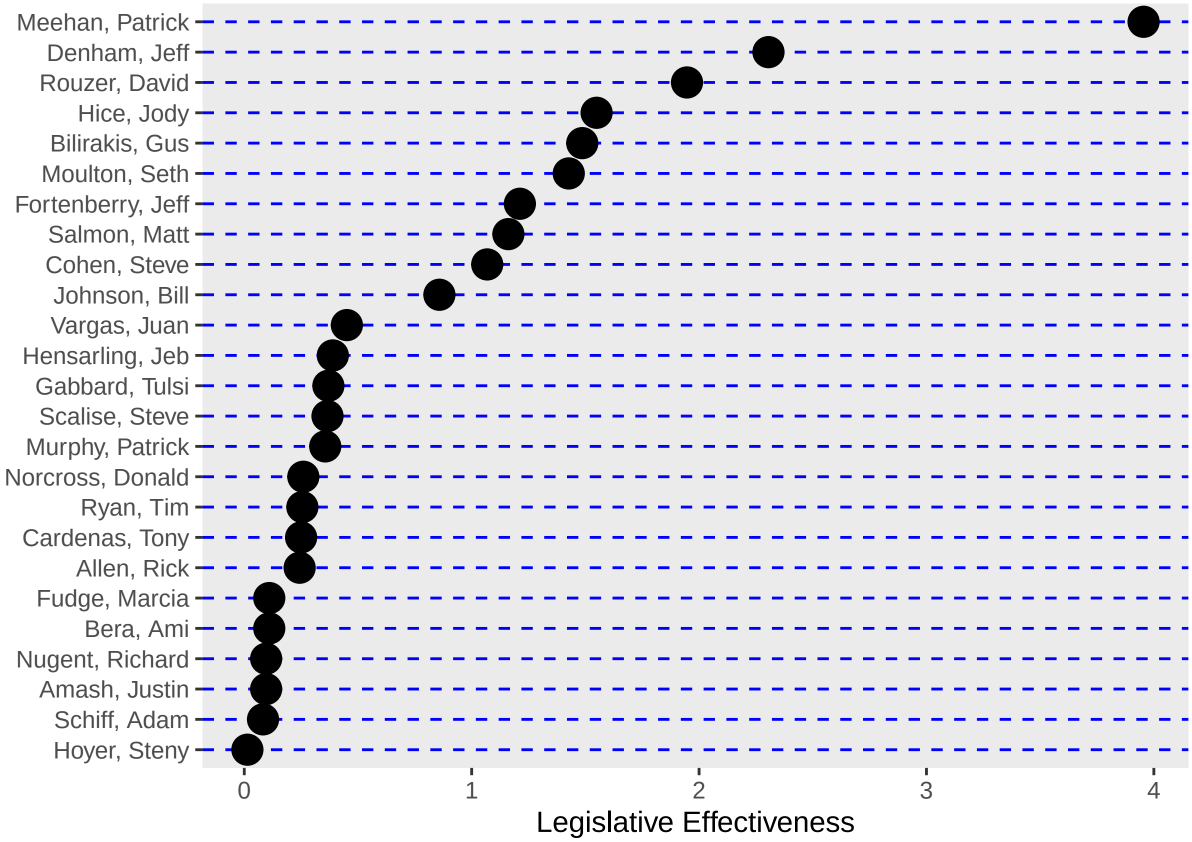
Cleveland Dot Plots

- Adapted from https://r-graphics.org/recipe-bar-graph-dot-plot

```
# Use some of the congress data
cel_114 <- cel %>% filter(congress == 114)

members <- sample_n(cel_114, 25)

# points only
ggplot(members, aes(x = les, y = thomas_name))+
  geom_point()
```



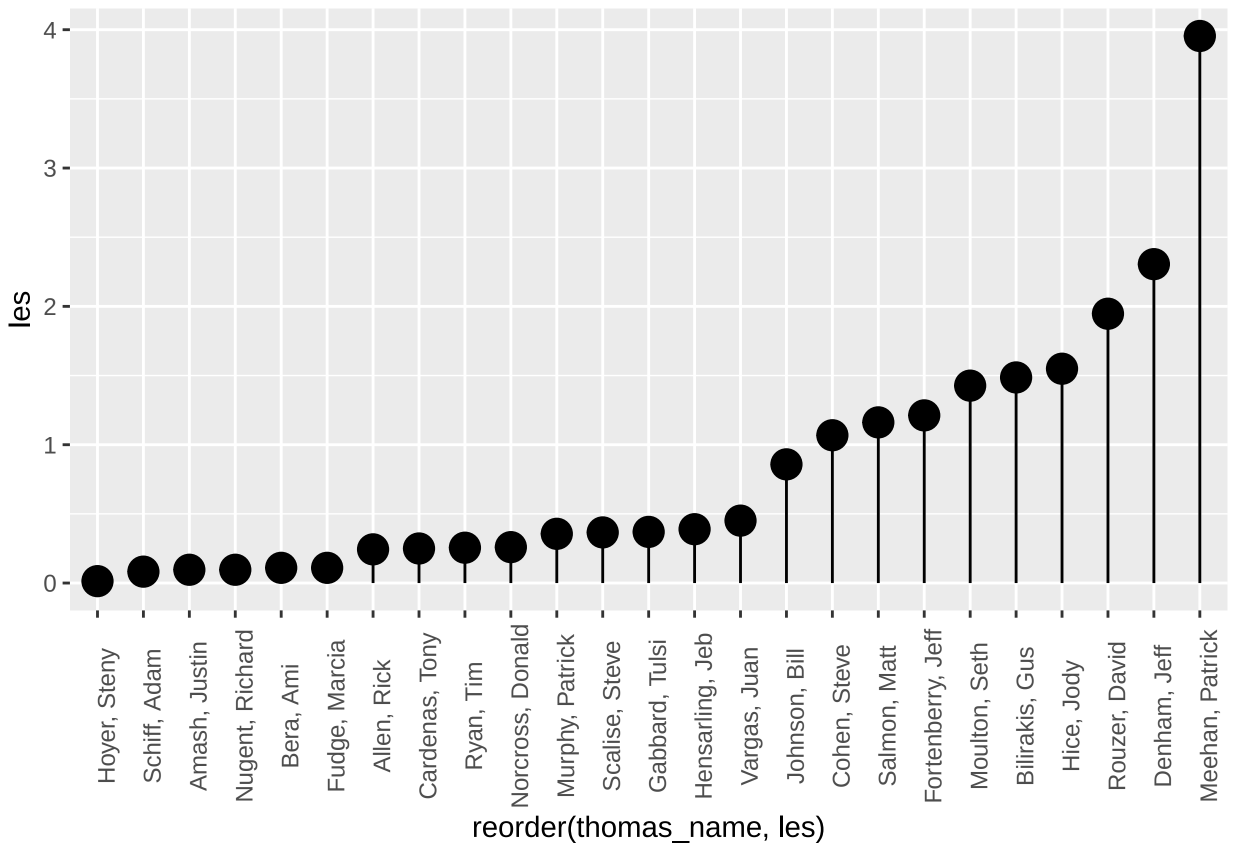
```
# some refinements
ggplot(members, aes(x = les, y = reorder(thomas_name, les)))+ # reorder in descending
  geom_point(size = 5)+
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(linetype = "dashed", color = "blue"))+
  labs(x = "Legislative Effectiveness", y = "")
```



Lollipop Figure

- Although it falls under the variations of scatterplot, it is also just the very refined simple way of making a bar plot essentially.
- https://www.r-graph-gallery.com/300-basic-lollipop-plot.html

```
ggplot(members, aes(x = reorder(thomas_name, les), y = les))+ # reorder in descending
  geom_point(size = 5)+
  geom_segment(aes(x = thomas_name, xend = thomas_name, y = 0, yend = les))+
  theme(axis.text.x = element_text(angle = 90))
```


les
4
3
2
1
0
Hoyer, Steny
Schiff, Adam
Amash, Justin
Nugent, Richard
Bera, Ami
Fudge, Marcia
Allen, Rick
Cardenas, Tony
Ryan, Tim
Norcross, Donald
Murphy, Patrick
Scalise, Steve
Gabbard, Tulsi
Hensarling, Jeb
Vargas, Juan
Johnson, Bill
Cohen, Steve
Salmon, Matt
Fortenberry, Jeff
Moulton, Seth
Bilirakis, Gus
Hice, Jody
Rouzer, David
Denham, Jeff
Meehan, Patrick
reorder(thomas_name, les)