ggpairs vignette

Oscar A. Trevizo

2023-May-03

Contents

Load the libraries

```
library(ggplot2)
library(GGally)
```

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

```
knitr::opts_chunk$set(echo = TRUE)
```

Get the data

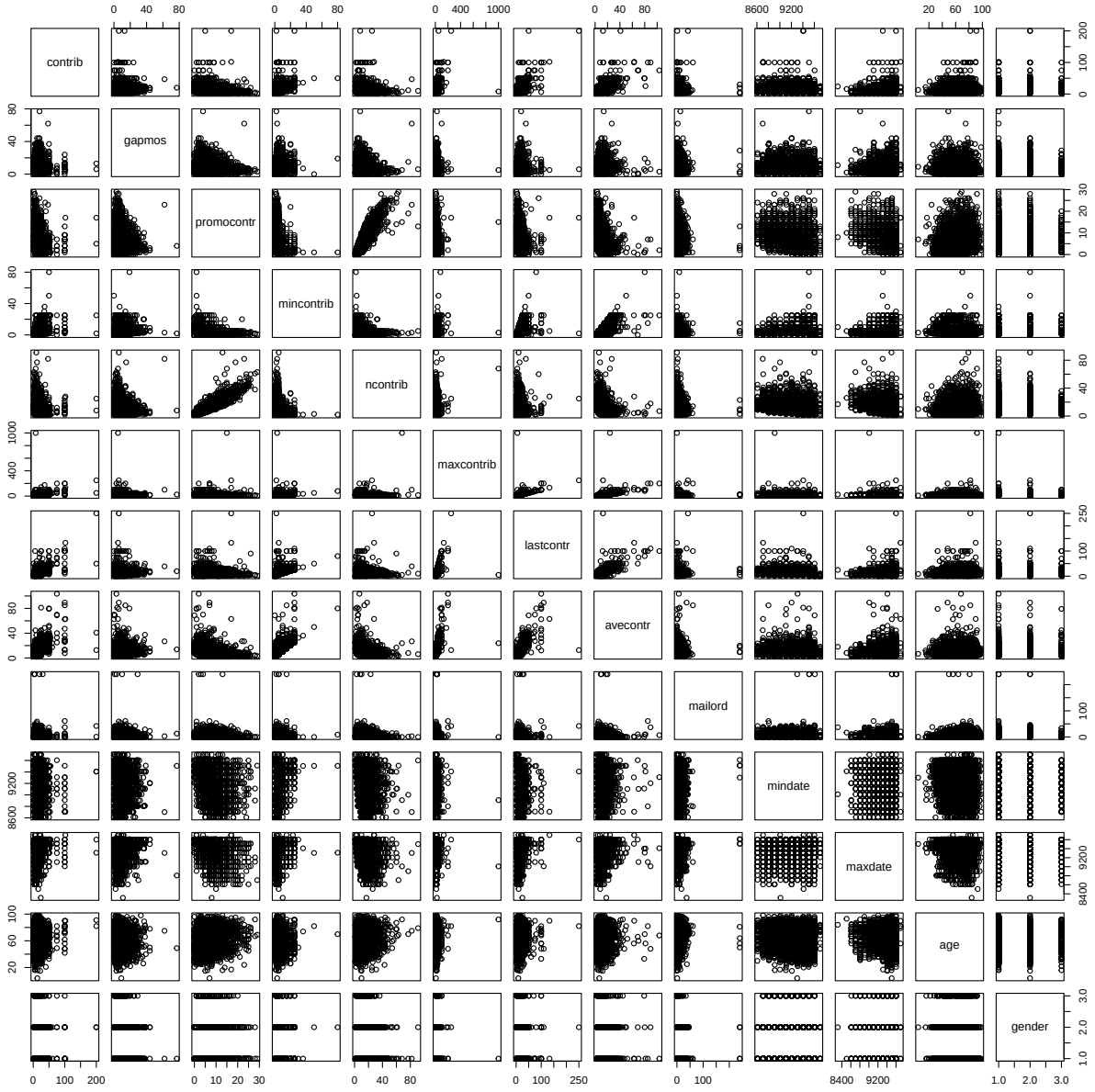
Reference: Harvard CSCI E-63c Statistical Learning, 2021.

```
####
#
#
# Data:
# 1. file "fund-raising.csv"
# 2. file "fund-raising-notes.txt"
#
```

```
# contrib     -- the outcome that the model is expected to predict: the donation
#                  amount ($) associated with the response to the mailing campaign of interest
#                  Type: num
#                  Units: $USD
# gapmos      -- number of months between first and second contributions from this donor
#                  Type: int
#                  Units: Months
# promocontr  -- total number of contributions to promotional mailing campaigns
#                  Type: int
#                  Units: number of contributions
# mincontrib  -- dollar amount of this donor smallest contribution to date
#                  Type: num
#                  Units: $USD
# ncontrib    -- total number of contributions by this donor
#                  Type: int
#                  Units: Number of contributions
# maxcontrib  -- dollar amount of this donor largest contribution to date
#                  Type: num
#                  Units: $USD
# lastcontr   -- dollar amount of this donor most recent contribution
#                  Type: num
#                  Units: $USD
# avecontr    -- average dollar amount of this donor contribution
#                  Type: num
#                  Units: $USD
# mailord     -- total number of times the donor has responded to a mail order
#                  offer other than the study sponsor's
#                  Type: int
#                  Units: Number of contributions
# mindate     -- date (YYMM) of this donor smallest contribution
#                  Type: int
#                  Units: Number of contributions
# maxdate     -- date (YYMM) of this donor largest contribution
#                  Type: int
#                  Units: Number of contributions
# age         -- donor age
#                  Type: int
#                  Units: Number of years
# gender      -- donor gender --> Categorical
#                  Type: Factor
#                  Units: Category Female "F", male "M", Unspecified "U"
##

frcDat <- read.table("../data/fund-raising.csv",sep=",",header=TRUE, as.is = FALSE)
dim(frcDat)
```

```
## [1] 3470   13
```

Classic pairs()



ggpairs()

The following charts and matrices indicate that the most robustly correlated variable with the target contributions (`contrib`) is the last contribution (`lastcontr`): I.e. **the "dollar amount of this donor most recent contribution**."

To start this analysis, the following three sets of *ggpairs()* plots help us to explore patterns:

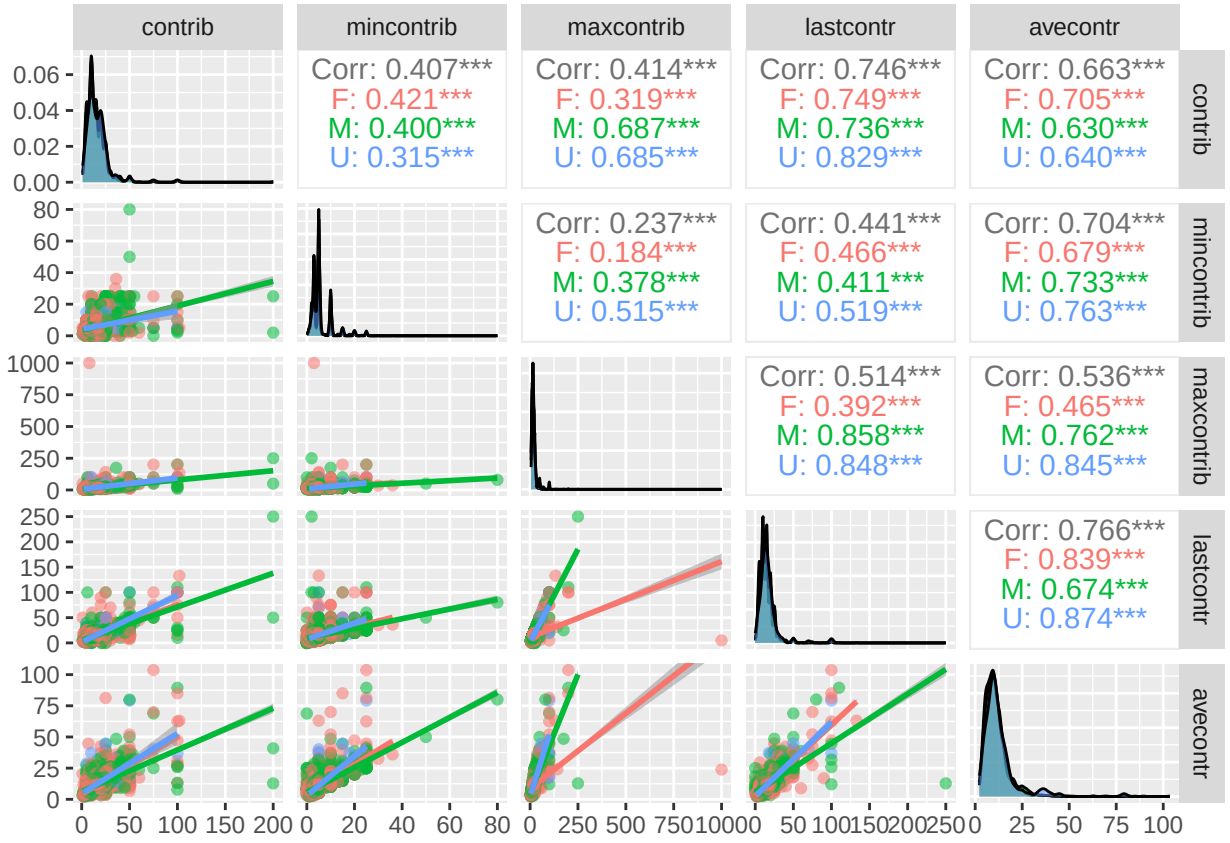
One *ggpairs()* plot shows *dollar* type of variables.

A second *ggpairs()* plot shows variables that count (e.g. number of contributions) vs `contrib`

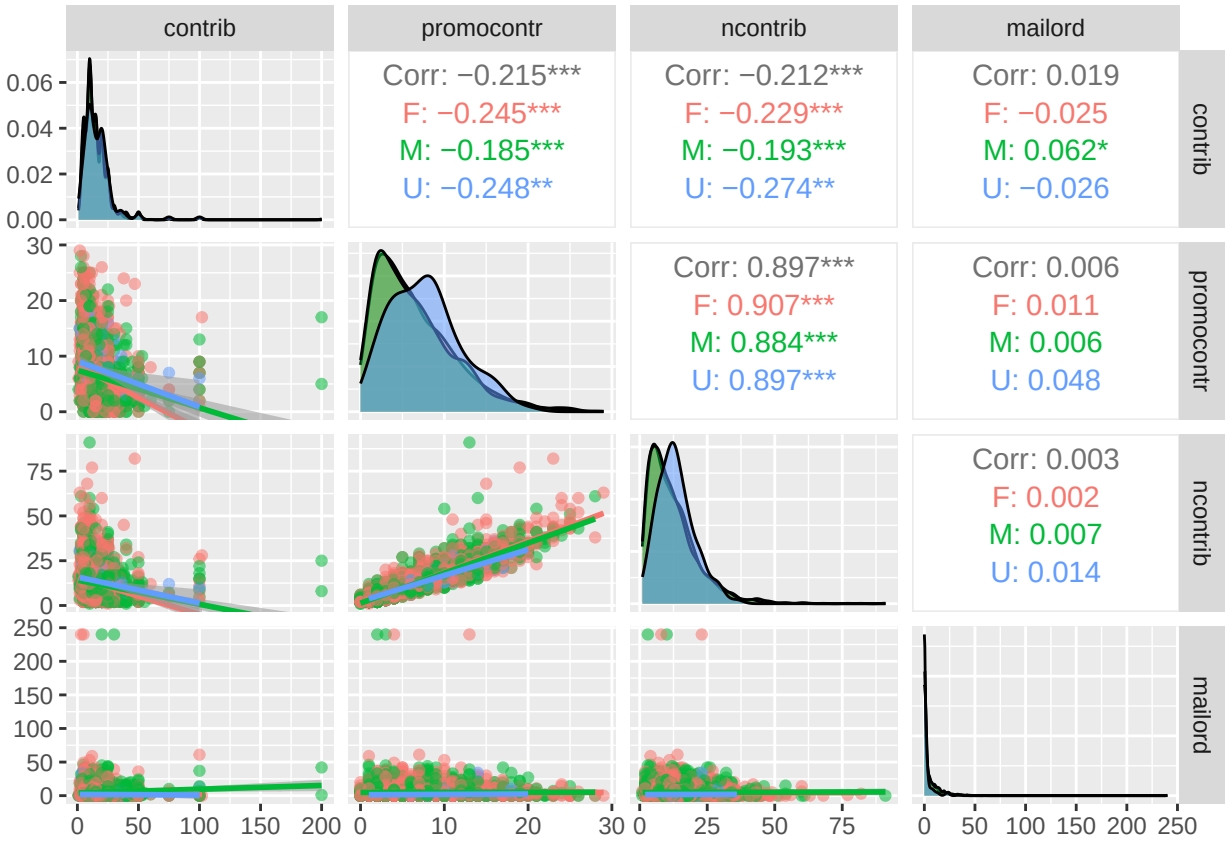
And a third *ggpairs()* plot to show variables that involve time or dates vs `contrib`.

The three sets of *ggpairs()* plots will distinguish gender and will include trend lines on a per-gender basis.

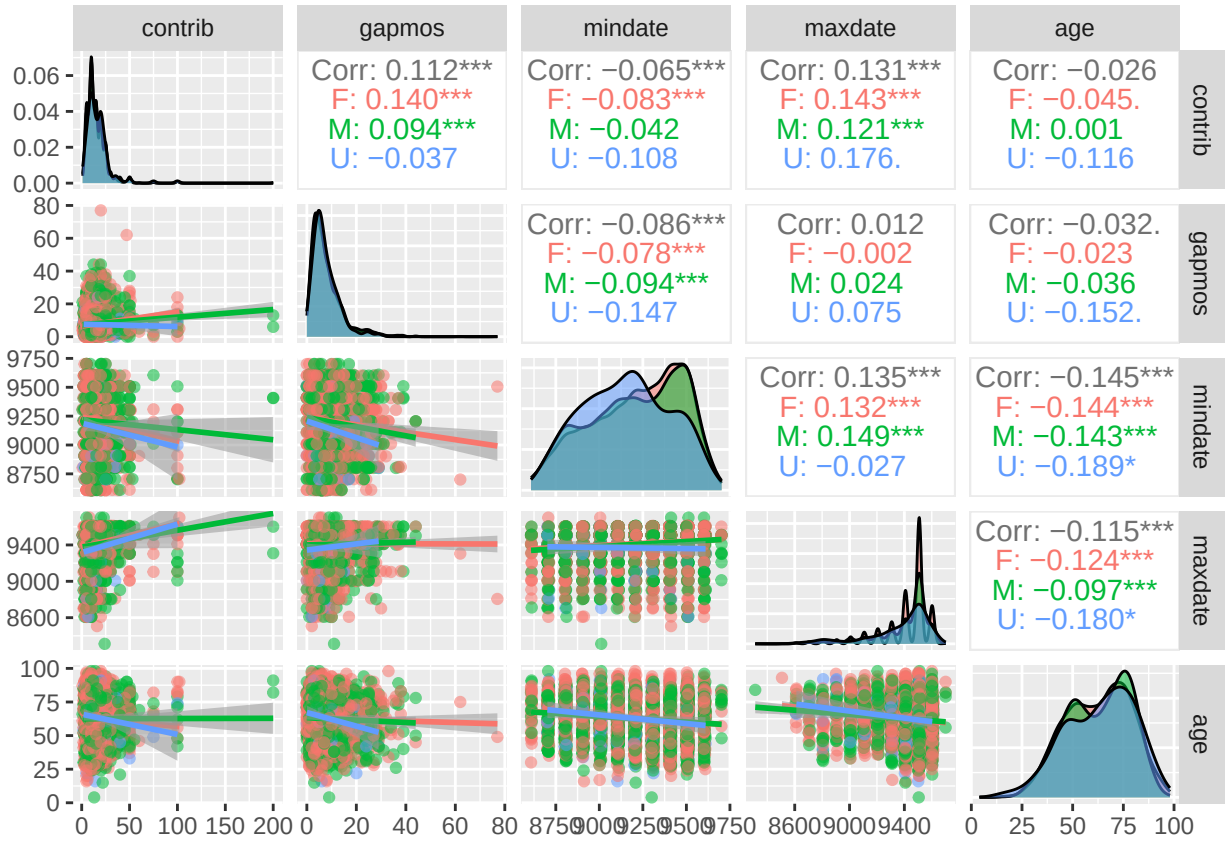
```
par(mfrow=c(3, 1))

# Let ggpairs plot $USD continuous variables before analyzing
ggpairs(frcDat, columns = c("contrib", "mincontrib", "maxcontrib", "lastcontr", "avecontr"),
        upper = list(continuous = wrap("cor", size = 4)),
        lower = list(continuous = "smooth"),
        aes(color = gender,
            alpha = 0.5)
        )
```




```
# Let ggpairs plot INT (number of) continuous variables vs "contrib" before analyzing
ggpairs(frcDat, columns = c("contrib", "promocontr", "ncontrib", "mailord"),
        upper = list(continuous = wrap("cor", size = 4)),
        lower = list(continuous = "smooth"),
        aes(color = gender,
            alpha = 0.5)
        )
```




```
# Let ggpairs plot time, date or age related continuous variables vs "contrib" before analyzing
ggpairs(frcDat, columns = c("contrib", "gapmos", "mindate", "maxdate", "age"),
        upper = list(continuous = wrap("cor", size = 4)),
        lower = list(continuous = "smooth"),
        aes(color = gender,
            alpha = 0.5)
        )
```



The first *ggpairs()* set of plots for *$USD dollar* variables shows a level of correlation with `contrib`. The second set of *ggpairs()* plots based on counts also shows a level of correlation with `contrib`. The third set shows little correlation. The plots also show outliers that may impact the following results. The plots also show some outliers that may impact the results.

References

1. Book (ISLR) "An Introduction to Statistical Learning with Applications in R" by Gareth James et al
2. Book "R for Data Science" by Hadley Wickham and Garrett Grolemund
3. Book "R Graphics Cookbook" by Winston Chang
4. LinkedIn Learning "Wrangling and Visualizing Data" by Barton Poulson