Scatterplot and Correlation in R

Scatterplot

In these notes, we are going to reproduce Figure 12.4 from the Clark, Golder, and Golder (2013, pp. 477-478) reading. The necessary variables are in the data set `gamson.rds`.

```
# load data
gamson <- readRDS("data/gamson.rds")
# note: make sure the file 'taiwan.rds' is in the 'data' subdirectory
#   and your working directory is set appropriately.

# quick look at data
tibble::glimpse(gamson)
```

```
## Observations: 826
## Variables: 2
## $ seat_share      <dbl> 0.02424242, 0.46060607, 0.51515150, 0.47204968...
## $ portfolio_share <dbl> 0.09090909, 0.36363637, 0.54545456, 0.45454547...
```

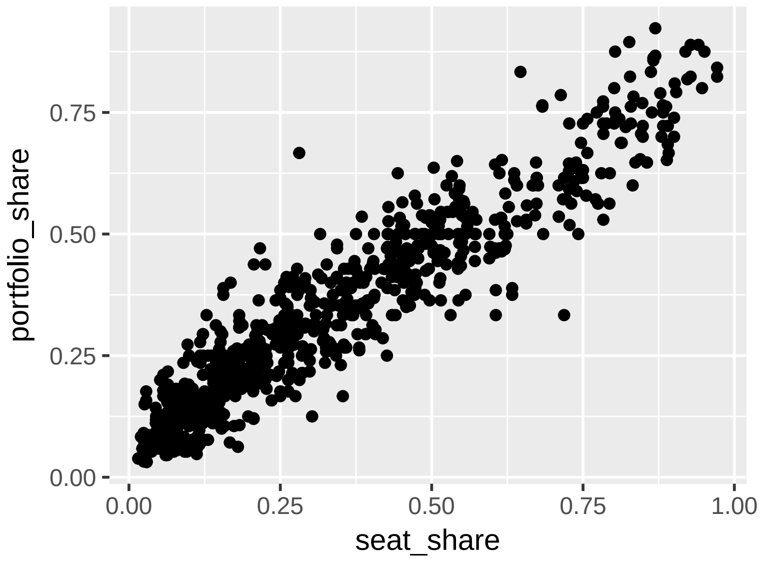
You can see that the data frame `gamson` has two variables: `seat_share`, and `portfolio`.

Of course, we're going to use ggplot, and we know we'll use `gamson` as the data frame.

In a scatterplot, we usually plot the variable *doing the causing* along the x-axis and the variable *being caused* along the y-axis. It makes sense here that seat shares cause portfolio shares because portfolio shares are determined well after seat shares are determined. Based on this rule, we can quickly figure out what the x and y aesthetics will be: `x = seat_share` and `y = portfolio_share`.

All that's left is the geometry. To create a scatterplot, we use `geom_point()`.

```
# load packages
library(ggplot2)

# create scatterplot
ggplot(gamson, aes(x = seat_share, y = portfolio_share)) +
  geom_point()
```



The Size and Color Aesthetics

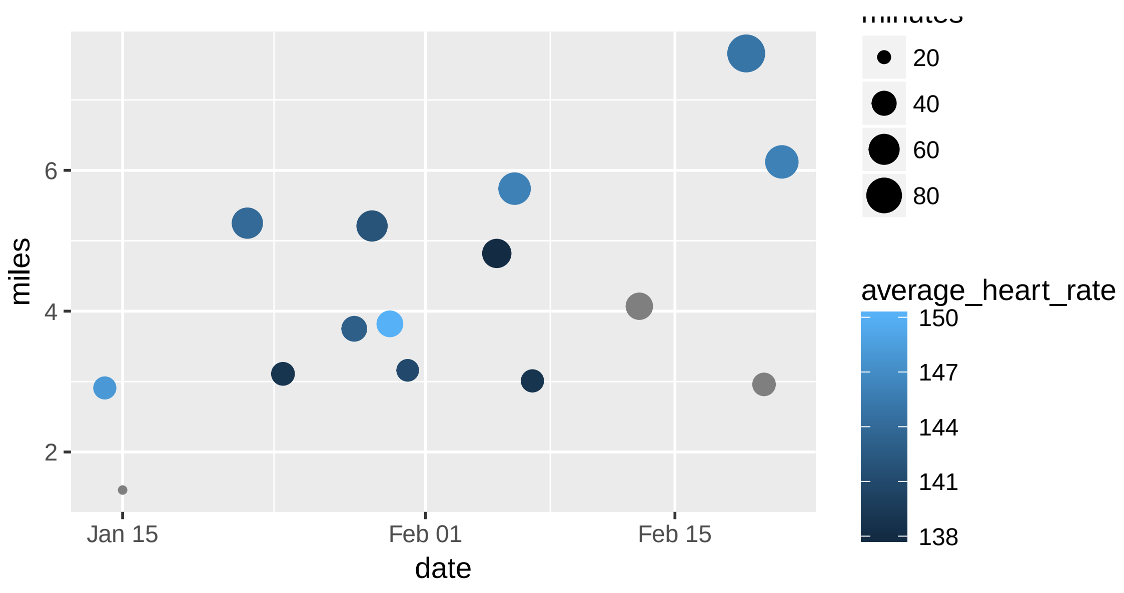
I enjoy triathlons. Running is by far my worst event, so I've been working hard at it this year. I log my runs in a Google Sheet, which we can read into R using the code below, which I've also posted on the course webpage, so that you can copy-and-paste more easily.

```r
# data loading and tidying
# ########################
# Note: we haven't talked about how to do this, and I don't expect you to
# understand the code below.
library(dplyr)  # useful function for cleaning up data
library(lubridate)  # useful for working with dates and durations
library(magrittr)  # useful for data manipulation
library(googlesheets)  # used to load the google sheet data
sheet <- gs_key("19aZA5_xnMiNfrpJpVur1FxNWGxUe3l-LlJ8vPp7_uBc") # register google sheet
runs <- gs_read(sheet)  # load sheet
runs <- runs %>%
  mutate(date = mdy(date),
         minutes = as.numeric(time)/60) %>%
  rename(average_heart_rate = av_hr) %>%
  select(-time)
# done with tidying
# #################

# quick look at data
tibble::glimpse(runs)
```

```
## Observations: 15
## Variables: 4
## $ date               <date> 2017-01-14, 2017-01-15, 2017-01-22, 2017-0...
## $ miles              <dbl> 2.91, 1.46, 5.25, 3.11, 3.75, 5.21, 3.82, 3...
## $ average_heart_rate <int> 148, NA, 144, 139, 143, 142, 150, 141, 138,...
## $ minutes            <dbl> 34.81667, 18.13333, 60.88333, 36.58333, 42....
```

You can see that the cleaned-up data frame `runs` has four variables, all of which we might be interested in. Since a scatterplot only has two spatial aesthetics (horizontal and vertical positioning), we'll have to use other aesthetics. Color and size are two options.

```r
# create scatterplot with color and size aesthetics
ggplot(runs, aes(x = date, y = miles, color = average_heart_rate, size = minutes)) +
  geom_point()
```



Review Exercises

1. In creating a scatterplot, what variable do we typically place along the x- and y-axes?
2. What geometry creates a scatterplot?
3. Experiment with the x, y, size, and color aesthetics for the runs data. What combination produces the most useful plot?
4. Many authors argue that district magnitude (the number of legislative seats in a districts) causes turnout. In particular, they argue that increasing district magnitude leads to an increase in turnout. The taiwan data set has information that might be useful in testing this hypothesis. Using the data set `taiwan.rds`, create the appropriate scatterplot and evaluate whether the data are consistent with the claim that district magntidue has a large, positive effect on turnout.

Correlation

In order to compute a correlation in R, we use the `cor()` function. The first argument `x` is the first of the two variables for which we would like to calculate a correlation. The second argument `y` is the second of the two variables. `cor()` is not designed to work with data frames, so we have to use the `data$variable` syntax.

```
# calculate a correlation
cor(gamson$seat_share, gamson$portfolio_share)
```

```
## [1] 0.9423176
```

For the data frame `runs` we have some missing data in the variable `average_heart_rate`. I simply forgot to wear my monitor on these days. In order to drop the incomplete pairs (either x is missing, y is missing, or both), we just supply the argument `use = "pairwise.complete.obs"` to the `cor()` function.

```
# calculate a correlation
cor(runs$minutes, runs$average_heart_rate) # returns NA
```

```
## [1] NA
```

```
cor(runs$minutes, runs$average_heart_rate, use = "pairwise.complete.obs") # returns NA
```

```
## [1] 0.2631646
```

Review Exercises

1. What function do we use to calculate a correlation in R?

2. If some observations are missing, what argument to we use to drop those observations?
3. Does `cor()` take a `data` argument? If not, how do we calculate correlations for variables in data frames?
4. Using the data set `taiwan.rds`, calculate a correlation to assess whether the data are consistent with the claim that district magntidue has a large, positive effect on turnout.

Correlation Matrix

Sometimes we want to calculate correlations for many variables at the same time. Each of the correlations is computed in the usual way, except they are presented in a correlation matrix. The `cor` function allows us to compute a correlation matrix quickly if the first argument is a data frame rather than a vector. If the first argument is a data frame, then `cor()` computes correlations between every varibles in the data frame.

Note: every variable in the data frame must be numeric.

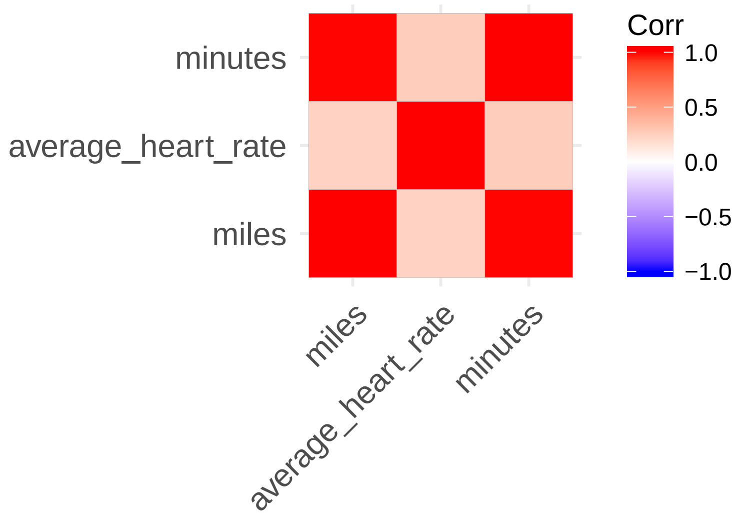
If we don't want to compute correlations for every variable in the data frame, then we can use the `select()` function from the package dplyr to create a new data frame that includes only certain variables from the original data frame. The first argument to `select()` is the orginal data frame. The remaining arguments are the variables we want to keep.

```
# load package
library(dplyr)  # for select()

# keep only the miles, average_heart_rate, and minutes variables
numeric_vars <- select(runs, miles, average_heart_rate, minutes)

# calculate the correlations between every variables in our new data frame
cor(numeric_vars, use = "pairwise.complete.obs")
```

```
##                        miles average_heart_rate   minutes
## miles              1.0000000          0.2340291 0.9970443
## average_heart_rate 0.2340291          1.0000000 0.2631646
## minutes            0.9970443          0.2631646 1.0000000
```

We can also store these correlations as an object and create a ggplot that graphically communicates the correlations. The function `ggcorrplot()` in the package ggcorrplot does this automatically without us having to specify a data frame, aesthetics, or geometry. `ggcorrplot()` needs only one argument–the output of the `cor()` function.

```
# calculate the correlations between every variables in our new data frame
cors <- cor(numeric_vars, use = "pairwise.complete.obs")

# use ggcorrplot()
library(ggcorrplot)  # for ggcorrplot()
ggcorrplot(cors)
```



The data frame `health.rds` contains a lot of variables for which we might like to calculate correlations, so let's take a look.

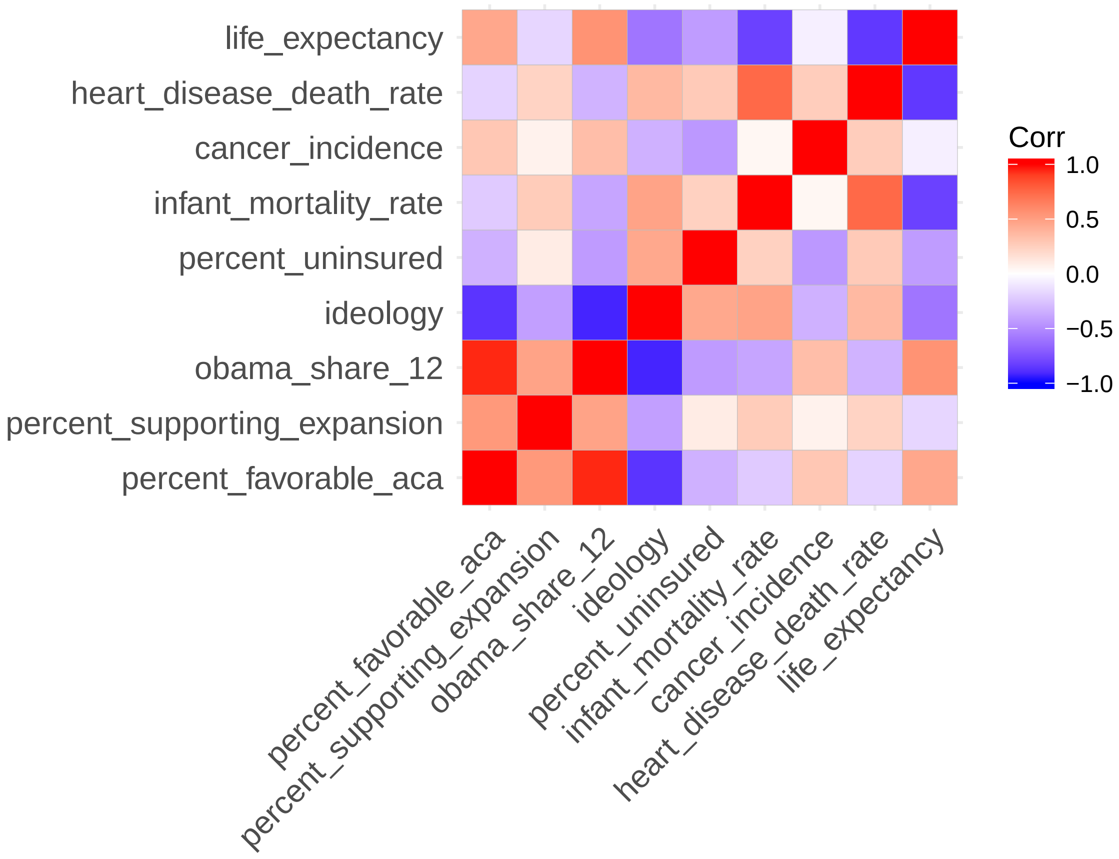
```
# load data
health <- readRDS("data/health.rds")

# quick look
tibble::glimpse(health)
```

```
## Observations: 50
## Variables: 17
## $ state                          <chr> "Alabama", "Alaska", "Arizona", "...
## $ state_abbr                     <chr> "AL", "AK", "AZ", "AR", "CA", "CO...
## $ gov_party                      <fctr> Repubican Governor, Repubican Go...
## $ sen_party                      <fctr> Republican Senate, Republican Se...
## $ house_party                    <fctr> Republican House, Republican Hou...
## $ percent_favorable_aca          <dbl> 38.27111, 37.44285, 39.67216, 36....
## $ percent_supporting_expansion   <dbl> 57.76161, 47.42469, 53.21254, 54....
## $ obama_share_12                 <dbl> 38.78377, 42.68471, 45.38662, 37....
## $ ideology                       <dbl> 0.24404363, 0.04723307, 0.1048642...
## $ percent_uninsured              <int> 14, 19, 18, 18, 19, 15, 8, 10, 21...
## $ infant_mortality_rate          <dbl> 9.2, 6.5, 6.4, 7.6, 5.1, 6.2, 6.1...
## $ cancer_incidence               <dbl> 472.9, 451.4, 387.1, 426.7, 434.0...
## $ heart_disease_death_rate       <dbl> 236.0, 151.5, 146.7, 222.5, 161.9...
## $ life_expectancy                <dbl> 75.4, 78.3, 79.6, 76.0, 80.8, 80....
## $ leg_party                      <fctr> Unified Republican Legislature, ...
## $ health_score                   <dbl> -2.09998657, 0.04841030, 0.644463...
## $ health_score_cat               <fctr> Bottom Tercile, Middle Tercile, ...
```

Let's use the `select()` function to create a new data frame with only numeric variables and then plot a correlation matrix for all of them.

```
# keep only the miles, average_heart_rate, and minutes variables
numeric_vars <- select(health, percent_favorable_aca, percent_supporting_expansion,
                       obama_share_12, ideology, percent_uninsured,
                       infant_mortality_rate, cancer_incidence,
                       heart_disease_death_rate, life_expectancy)

# calculate the correlations between every variables in our new data frame
```

```r
cors <- cor(numeric_vars, use = "pairwise.complete.obs")

# use ggcorrplot()
ggcorrplot(cors)
```



Review Exercises

1. How can we use the `cor()` function to compute many correlations at once?
2. What does the `select()` function do? What is the first argument? What are the subsequent arguments? What does it output?
3. What function can we use to plot a correlation matrix? What argument does it take?
4. Take a look at the very last figure in this document–the correlation matrix for the health data set. Why is the diagonal completely red?
5. Looking at that same figure, how do the measures of state health (i.e., infant mortality rate, cancer incidence, heart disease death rate, and life expectancy) correlate with support for the ACA? Which correlations are in the expected direction?