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`.

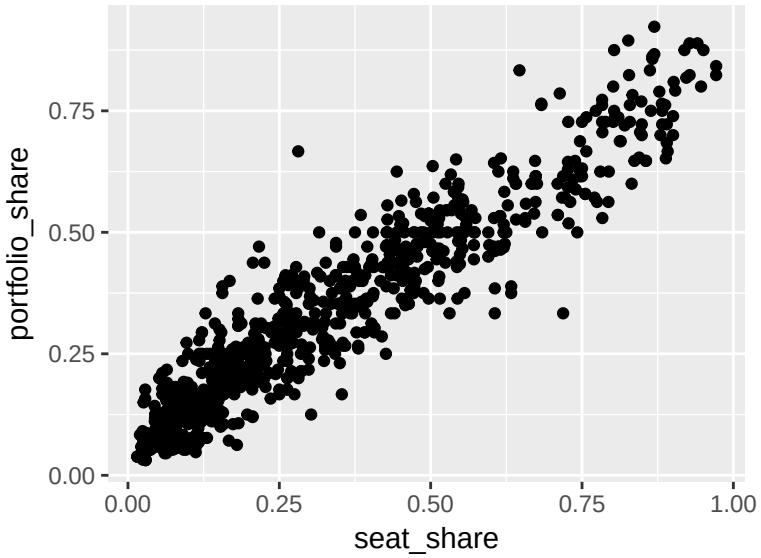
Of course, we're going to use ggplot, and we know we'll use `gamson` as the data frame.

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

The data set health has several variables capturing the health outcomes and politics of the 50 U.S. states, such as the infant mortality rate and percent that have a favorable attitude toward the ACA, respectively.

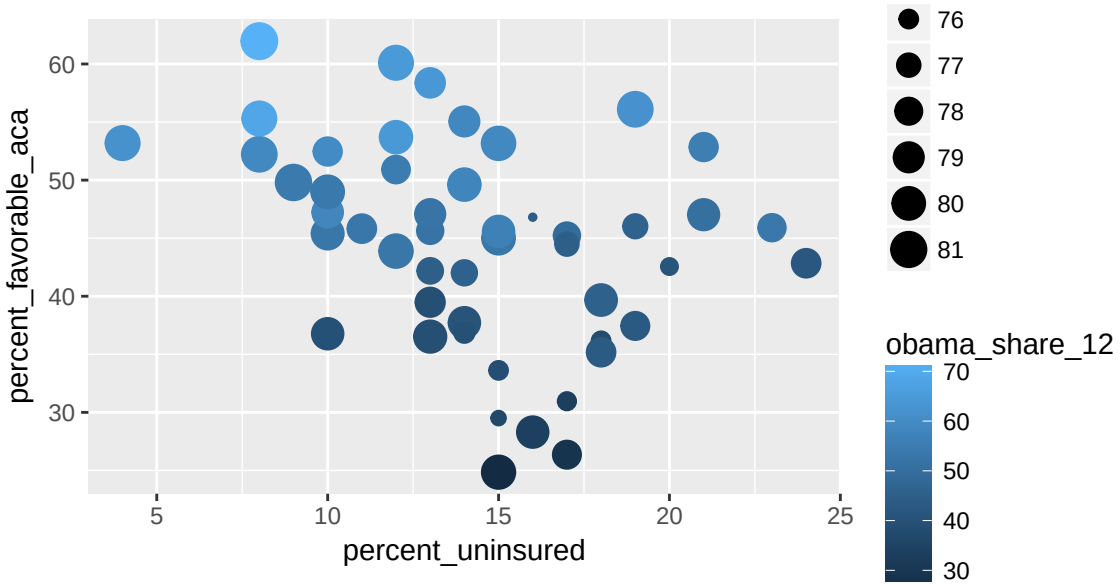
```
# load data
health <- readr::read_csv("data/health.csv")

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

```
## Observations: 50
## Variables: 17
## $ state                         <chr> "Alabama", "Alaska", "Arizona", "...
## $ state_abbr                    <chr> "AL", "AK", "AZ", "AR", "CA", "CO...
## $ gov_party                     <chr> "Repubican Governor", "Repubican ...
## $ sen_party                     <chr> "Republican Senate", "Republican ...
## $ house_party                   <chr> "Republican House", "Republican H...
## $ percent_favorable_aca         <dbl> 38.2711, 37.4428, 39.6722, 36.162...
## $ percent_supporting_expansion <dbl> 57.7616, 47.4247, 53.2125, 54.438...
## $ obama_share_12                <dbl> 38.7838, 42.6847, 45.3866, 37.845...
## $ ideology                      <dbl> 0.2440440, 0.0472331, 0.1048640, ...
## $ 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                     <chr> "Unified Republican Legislature",...
## $ health_score                  <dbl> -2.0999900, 0.0484103, 0.6444630,...
## $ health_score_cat              <chr> "Bottom Tercile", "Middle Tercile...
```

You can see that the data frame **health** has several variables, many 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.

```
# create scatterplot with color and size aesthetics
ggplot(health, aes(x = percent_uninsured, y = percent_favorable_aca, color = obama_share_12, size = life
  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 health data. What combination produces the most useful or interesting 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 magnitude 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.

Remember that the textbook defines a correlation as $r =$ average of (x in standard units)$\times$(y in standard units). For reasons similar to the SD, most computer programs, including R, divide by the number of entries $ - 1$ rather than the number of entries. For this reason, you'll see small differences between a correlation computed by R and a correlation computed by hand, especially when the number of observations is small. For practical purposes, though, the two approaches are equivalent, especially when we have many observations.

```
# 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(health$percent_uninsured, health$infant_mortality_rate) # returns NA
```

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

```r
cor(health$percent_uninsured, health$infant_mortality_rate, use = "pairwise.complete.obs") # returns NA
```

```
## [1] 0.2431223
```

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 magnitude has a large, positive effect on turnout.