Lab4_26Feb

Aicha

2023-02-26

Data Wrangling

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.2.1     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0

## Warning: package 'tibble' was built under R version 4.2.3

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

The data we will use is presenditial dataset which is built in tidyverse

```
data <- presidential
head(data) #print first 5 rows
```

```
## # A tibble: 6 x 4
##   name       start      end        party
##   <chr>      <date>     <date>     <chr>
## 1 Eisenhower 1953-01-20 1961-01-20 Republican
## 2 Kennedy    1961-01-20 1963-11-22 Democratic
## 3 Johnson    1963-11-22 1969-01-20 Democratic
## 4 Nixon      1969-01-20 1974-08-09 Republican
## 5 Ford       1974-08-09 1977-01-20 Republican
## 6 Carter     1977-01-20 1981-01-20 Democratic
```

```
data # prints all dataframe
```

```
## # A tibble: 12 x 4
##    name       start      end        party
##    <chr>      <date>     <date>     <chr>
##  1 Eisenhower 1953-01-20 1961-01-20 Republican
##  2 Kennedy    1961-01-20 1963-11-22 Democratic
```

```
##  3 Johnson    1963-11-22 1969-01-20 Democratic
##  4 Nixon      1969-01-20 1974-08-09 Republican
##  5 Ford       1974-08-09 1977-01-20 Republican
##  6 Carter     1977-01-20 1981-01-20 Democratic
##  7 Reagan     1981-01-20 1989-01-20 Republican
##  8 Bush       1989-01-20 1993-01-20 Republican
##  9 Clinton    1993-01-20 2001-01-20 Democratic
## 10 Bush       2001-01-20 2009-01-20 Republican
## 11 Obama      2009-01-20 2017-01-20 Democratic
## 12 Trump      2017-01-20 2021-01-20 Republican
```

select()

The first method that we will use to process the data is **select()** which is used to manipulate columns

```
select(data, name, party)
```

```
## # A tibble: 12 x 2
##    name       party
##    <chr>      <chr>
##  1 Eisenhower Republican
##  2 Kennedy    Democratic
##  3 Johnson    Democratic
##  4 Nixon      Republican
##  5 Ford       Republican
##  6 Carter     Democratic
##  7 Reagan     Republican
##  8 Bush       Republican
##  9 Clinton    Democratic
## 10 Bush       Republican
## 11 Obama      Democratic
## 12 Trump      Republican
```

```
#select range of columns
select(data, name:end)
```

```
## # A tibble: 12 x 3
##    name       start      end
##    <chr>      <date>     <date>
##  1 Eisenhower 1953-01-20 1961-01-20
##  2 Kennedy    1961-01-20 1963-11-22
##  3 Johnson    1963-11-22 1969-01-20
##  4 Nixon      1969-01-20 1974-08-09
##  5 Ford       1974-08-09 1977-01-20
##  6 Carter     1977-01-20 1981-01-20
##  7 Reagan     1981-01-20 1989-01-20
##  8 Bush       1989-01-20 1993-01-20
##  9 Clinton    1993-01-20 2001-01-20
## 10 Bush       2001-01-20 2009-01-20
## 11 Obama      2009-01-20 2017-01-20
## 12 Trump      2017-01-20 2021-01-20
```

Also, we cam use select to reorder the columns in data frame

```
data <- select(data, name, party, start, end)
```

We can also hnage the name of a variable using select functions as the following

```
select(data, president = name, startdate=start, enddate=end)
```

```
## # A tibble: 12 x 3
##    president  startdate  enddate
##    <chr>      <date>     <date>
##  1 Eisenhower 1953-01-20 1961-01-20
##  2 Kennedy    1961-01-20 1963-11-22
##  3 Johnson    1963-11-22 1969-01-20
##  4 Nixon      1969-01-20 1974-08-09
##  5 Ford       1974-08-09 1977-01-20
##  6 Carter     1977-01-20 1981-01-20
##  7 Reagan     1981-01-20 1989-01-20
##  8 Bush       1989-01-20 1993-01-20
##  9 Clinton    1993-01-20 2001-01-20
## 10 Bush       2001-01-20 2009-01-20
## 11 Obama      2009-01-20 2017-01-20
## 12 Trump      2017-01-20 2021-01-20
```

select helps us to **drop** some columns as well

```
select(data, -end)
```

```
## # A tibble: 12 x 3
##    name       party      start
##    <chr>      <chr>      <date>
##  1 Eisenhower Republican 1953-01-20
##  2 Kennedy    Democratic 1961-01-20
##  3 Johnson    Democratic 1963-11-22
##  4 Nixon      Republican 1969-01-20
##  5 Ford       Republican 1974-08-09
##  6 Carter     Democratic 1977-01-20
##  7 Reagan     Republican 1981-01-20
##  8 Bush       Republican 1989-01-20
##  9 Clinton    Democratic 1993-01-20
## 10 Bush       Republican 2001-01-20
## 11 Obama      Democratic 2009-01-20
## 12 Trump      Republican 2017-01-20
```

several methods can be used within select. We will check some of them below

```
select(data, contains("ar")) # columns that contain ar
```

```
## # A tibble: 12 x 2
##    party      start
##    <chr>      <date>
```

```
##  1 Republican 1953-01-20
##  2 Democratic 1961-01-20
##  3 Democratic 1963-11-22
##  4 Republican 1969-01-20
##  5 Republican 1974-08-09
##  6 Democratic 1977-01-20
##  7 Republican 1981-01-20
##  8 Republican 1989-01-20
##  9 Democratic 1993-01-20
## 10 Republican 2001-01-20
## 11 Democratic 2009-01-20
## 12 Republican 2017-01-20
```

```
select(data, starts_with("s")) # columns that start with s
```

```
## # A tibble: 12 x 1
##    start
##    <date>
##  1 1953-01-20
##  2 1961-01-20
##  3 1963-11-22
##  4 1969-01-20
##  5 1974-08-09
##  6 1977-01-20
##  7 1981-01-20
##  8 1989-01-20
##  9 1993-01-20
## 10 2001-01-20
## 11 2009-01-20
## 12 2017-01-20
```

```
select(data, ends_with("y")) # columns that ends with y
```

```
## # A tibble: 12 x 1
##    party
##    <chr>
##  1 Republican
##  2 Democratic
##  3 Democratic
##  4 Republican
##  5 Republican
##  6 Democratic
##  7 Republican
##  8 Republican
##  9 Democratic
## 10 Republican
## 11 Democratic
## 12 Republican
```

```
# begin with the stated colun and the print everything
select(data, party,everything())
```

```
## # A tibble: 12 x 4
##    party      name       start      end
##    <chr>      <chr>      <date>     <date>
##  1 Republican Eisenhower 1953-01-20 1961-01-20
##  2 Democratic Kennedy    1961-01-20 1963-11-22
##  3 Democratic Johnson    1963-11-22 1969-01-20
##  4 Republican Nixon      1969-01-20 1974-08-09
##  5 Republican Ford       1974-08-09 1977-01-20
##  6 Democratic Carter     1977-01-20 1981-01-20
##  7 Republican Reagan     1981-01-20 1989-01-20
##  8 Republican Bush       1989-01-20 1993-01-20
##  9 Democratic Clinton    1993-01-20 2001-01-20
## 10 Republican Bush       2001-01-20 2009-01-20
## 11 Democratic Obama      2009-01-20 2017-01-20
## 12 Republican Trump      2017-01-20 2021-01-20
```

```
# columns that match the given regular expression
select(data, matches("^s"))
```

```
## # A tibble: 12 x 1
##    start
##    <date>
##  1 1953-01-20
##  2 1961-01-20
##  3 1963-11-22
##  4 1969-01-20
##  5 1974-08-09
##  6 1977-01-20
##  7 1981-01-20
##  8 1989-01-20
##  9 1993-01-20
## 10 2001-01-20
## 11 2009-01-20
## 12 2017-01-20
```

Filter

Filter() function used to select soe rows from data frame (filter for rows, select for columns)

```
republican_pres <- filter(data, party == "Republican")
```

arrange()

This function is used to sort the dataframe based on some variables (ascending or descending using (desc))

```
arrange(data, desc(name))
```

```
## # A tibble: 12 x 4
##    name       party      start      end
##    <chr>      <chr>      <date>     <date>
##  1 Trump      Republican 2017-01-20 2021-01-20
```

```
##  2 Reagan     Republican 1981-01-20 1989-01-20
##  3 Obama      Democratic 2009-01-20 2017-01-20
##  4 Nixon      Republican 1969-01-20 1974-08-09
##  5 Kennedy    Democratic 1961-01-20 1963-11-22
##  6 Johnson    Democratic 1963-11-22 1969-01-20
##  7 Ford       Republican 1974-08-09 1977-01-20
##  8 Eisenhower Republican 1953-01-20 1961-01-20
##  9 Clinton    Democratic 1993-01-20 2001-01-20
## 10 Carter     Democratic 1977-01-20 1981-01-20
## 11 Bush       Republican 1989-01-20 1993-01-20
## 12 Bush       Republican 2001-01-20 2009-01-20
```

mutate()

we use mutate function to create new variables or columns

```
data2 <- mutate(data, duration=end-start,
                years=as.integer(duration/365),
                months=as.integer((duration-(years*365))/30))
data2
```

```
## # A tibble: 12 x 7
##    name       party      start      end        duration  years months
##    <chr>      <chr>      <date>     <date>     <drtn>    <int>  <int>
##  1 Eisenhower Republican 1953-01-20 1961-01-20 2922 days     8      0
##  2 Kennedy    Democratic 1961-01-20 1963-11-22 1036 days     2     10
##  3 Johnson    Democratic 1963-11-22 1969-01-20 1886 days     5      2
##  4 Nixon      Republican 1969-01-20 1974-08-09 2027 days     5      6
##  5 Ford       Republican 1974-08-09 1977-01-20  895 days     2      5
##  6 Carter     Democratic 1977-01-20 1981-01-20 1461 days     4      0
##  7 Reagan     Republican 1981-01-20 1989-01-20 2922 days     8      0
##  8 Bush       Republican 1989-01-20 1993-01-20 1461 days     4      0
##  9 Clinton    Democratic 1993-01-20 2001-01-20 2922 days     8      0
## 10 Bush       Republican 2001-01-20 2009-01-20 2922 days     8      0
## 11 Obama      Democratic 2009-01-20 2017-01-20 2922 days     8      0
## 12 Trump      Republican 2017-01-20 2021-01-20 1461 days     4      0
```

transmutate

we can also create column but not add them to the table => this column prints true if president was during cold war

```
transmute(data, CW = start < "1990-03-11")
```

```
## # A tibble: 12 x 1
##    CW
##    <lgl>
##  1 TRUE
##  2 TRUE
##  3 TRUE
##  4 TRUE
```

```
##  5 TRUE
##  6 TRUE
##  7 TRUE
##  8 TRUE
##  9 FALSE
## 10 FALSE
## 11 FALSE
## 12 FALSE
```

summurize

summuarize can be used to summarize the table in one row as per some functions => this summary provide us with the average duration of all president, the max and mean durations, the total duration, and the number of president which is calculated by counting number of rows. The result is grouped by party.

```
data3 <- group_by(data2, party)
summarize(data3, averageDays = mean(duration),
          maxDuration = max(duration),
          minDuration = min(duration),
          total=sum(duration),
          presidentNumber = n())
```

```
## # A tibble: 2 x 6
##   party      averageDays   maxDuration minDuration total      presidentNumber
##   <chr>      <drtn>        <drtn>      <drtn>      <drtn>                 <int>
## 1 Democratic 2045.400 days 2922 days   1036 days   10227 days                 5
## 2 Republican 2087.143 days 2922 days    895 days   14610 days                 7
```

```
#n_distinct counts distinct values
# n() counts number of rows
```

Other Functions

we have other useful functions

```
#adds column with count of repitions of this value
add_count(data, party)
```

```
## # A tibble: 12 x 5
##    name       party      start      end            n
##    <chr>      <chr>      <date>     <date>     <int>
##  1 Eisenhower Republican 1953-01-20 1961-01-20     7
##  2 Kennedy    Democratic 1961-01-20 1963-11-22     5
##  3 Johnson    Democratic 1963-11-22 1969-01-20     5
##  4 Nixon      Republican 1969-01-20 1974-08-09     7
##  5 Ford       Republican 1974-08-09 1977-01-20     7
##  6 Carter     Democratic 1977-01-20 1981-01-20     5
##  7 Reagan     Republican 1981-01-20 1989-01-20     7
##  8 Bush       Republican 1989-01-20 1993-01-20     7
##  9 Clinton    Democratic 1993-01-20 2001-01-20     5
## 10 Bush       Republican 2001-01-20 2009-01-20     7
```

```
## 11 Obama      Democratic 2009-01-20 2017-01-20     5
## 12 Trump      Republican 2017-01-20 2021-01-20     7
```

```
# we change the values of the rows for example
#from string to num for better processing in ml
mutate(data, party=recode(party, "Republican"=1, "Democratic"=2))
```

```
## # A tibble: 12 x 4
##    name       party start      end
##    <chr>      <dbl> <date>     <date>
##  1 Eisenhower     1 1953-01-20 1961-01-20
##  2 Kennedy        2 1961-01-20 1963-11-22
##  3 Johnson        2 1963-11-22 1969-01-20
##  4 Nixon          1 1969-01-20 1974-08-09
##  5 Ford           1 1974-08-09 1977-01-20
##  6 Carter         2 1977-01-20 1981-01-20
##  7 Reagan         1 1981-01-20 1989-01-20
##  8 Bush           1 1989-01-20 1993-01-20
##  9 Clinton        2 1993-01-20 2001-01-20
## 10 Bush           1 2001-01-20 2009-01-20
## 11 Obama          2 2009-01-20 2017-01-20
## 12 Trump          1 2017-01-20 2021-01-20
```

```
pull(data, party) # transform column to vector useful in ml
```

```
##  [1] "Republican" "Democratic" "Democratic" "Republican" "Republican"
##  [6] "Democratic" "Republican" "Republican" "Democratic" "Republican"
## [11] "Democratic" "Republican"
```

```
glimpse(data3)
```

```
## Rows: 12
## Columns: 7
## Groups: party [2]
## $ name     <chr> "Eisenhower", "Kennedy", "Johnson", "Nixon", "Ford", "Carter"~
## $ party    <chr> "Republican", "Democratic", "Democratic", "Republican", "Repu~
## $ start    <date> 1953-01-20, 1961-01-20, 1963-11-22, 1969-01-20, 1974-08-09, ~
## $ end      <date> 1961-01-20, 1963-11-22, 1969-01-20, 1974-08-09, 1977-01-20, ~
## $ duration <drtn> 2922 days, 1036 days, 1886 days, 2027 days, 895 days, 1461 d~
## $ years    <int> 8, 2, 5, 5, 2, 4, 8, 4, 8, 8, 8, 4
## $ months   <int> 0, 10, 2, 6, 5, 0, 0, 0, 0, 0, 0, 0
```

```
slice(data, 1:5)
```

```
## # A tibble: 5 x 4
##   name       party      start      end
##   <chr>      <chr>      <date>     <date>
## 1 Eisenhower Republican 1953-01-20 1961-01-20
## 2 Kennedy    Democratic 1961-01-20 1963-11-22
## 3 Johnson    Democratic 1963-11-22 1969-01-20
## 4 Nixon      Republican 1969-01-20 1974-08-09
## 5 Ford       Republican 1974-08-09 1977-01-20
```

```r
sample_n(data, 4)
```

```
## # A tibble: 4 x 4
##   name    party      start      end
##   <chr>   <chr>      <date>     <date>
## 1 Bush    Republican 1989-01-20 1993-01-20
## 2 Clinton Democratic 1993-01-20 2001-01-20
## 3 Kennedy Democratic 1961-01-20 1963-11-22
## 4 Bush    Republican 2001-01-20 2009-01-20
```

```r
training <- sample_frac(data, 0.8)
testing <- sample_frac(data, 0.2)
training
```

```
## # A tibble: 10 x 4
##    name       party      start      end
##    <chr>      <chr>      <date>     <date>
##  1 Reagan     Republican 1981-01-20 1989-01-20
##  2 Obama      Democratic 2009-01-20 2017-01-20
##  3 Johnson    Democratic 1963-11-22 1969-01-20
##  4 Bush       Republican 1989-01-20 1993-01-20
##  5 Trump      Republican 2017-01-20 2021-01-20
##  6 Carter     Democratic 1977-01-20 1981-01-20
##  7 Nixon      Republican 1969-01-20 1974-08-09
##  8 Bush       Republican 2001-01-20 2009-01-20
##  9 Eisenhower Republican 1953-01-20 1961-01-20
## 10 Kennedy    Democratic 1961-01-20 1963-11-22
```

```r
testing
```

```
## # A tibble: 2 x 4
##   name       party      start      end
##   <chr>      <chr>      <date>     <date>
## 1 Trump      Republican 2017-01-20 2021-01-20
## 2 Eisenhower Republican 1953-01-20 1961-01-20
```

```r
# changing values in rows depending on certain values using boolean
mutate(data2,
       duration = case_when(duration == 2922 ~ "Two ters",
                            duration == 1461 ~ "One term",
                            TRUE ~ "Special Case"))
```

```
## # A tibble: 12 x 7
##    name       party      start      end        duration     years months
##    <chr>      <chr>      <date>     <date>     <chr>        <int>  <int>
##  1 Eisenhower Republican 1953-01-20 1961-01-20 Two ters         8      0
##  2 Kennedy    Democratic 1961-01-20 1963-11-22 Special Case     2     10
##  3 Johnson    Democratic 1963-11-22 1969-01-20 Special Case     5      2
##  4 Nixon      Republican 1969-01-20 1974-08-09 Special Case     5      6
##  5 Ford       Republican 1974-08-09 1977-01-20 Special Case     2      5
##  6 Carter     Democratic 1977-01-20 1981-01-20 One term         4      0
```

```
##  7 Reagan     Republican 1981-01-20 1989-01-20 Two ters         8      0
##  8 Bush       Republican 1989-01-20 1993-01-20 One term         4      0
##  9 Clinton    Democratic 1993-01-20 2001-01-20 Two ters         8      0
## 10 Bush       Republican 2001-01-20 2009-01-20 Two ters         8      0
## 11 Obama      Democratic 2009-01-20 2017-01-20 Two ters         8      0
## 12 Trump      Republican 2017-01-20 2021-01-20 One term         4      0
```

```
data4 <- arrange(data, start)
data4
```

```
## # A tibble: 12 x 4
##    name       party      start      end
##    <chr>      <chr>      <date>     <date>
##  1 Eisenhower Republican 1953-01-20 1961-01-20
##  2 Kennedy    Democratic 1961-01-20 1963-11-22
##  3 Johnson    Democratic 1963-11-22 1969-01-20
##  4 Nixon      Republican 1969-01-20 1974-08-09
##  5 Ford       Republican 1974-08-09 1977-01-20
##  6 Carter     Democratic 1977-01-20 1981-01-20
##  7 Reagan     Republican 1981-01-20 1989-01-20
##  8 Bush       Republican 1989-01-20 1993-01-20
##  9 Clinton    Democratic 1993-01-20 2001-01-20
## 10 Bush       Republican 2001-01-20 2009-01-20
## 11 Obama      Democratic 2009-01-20 2017-01-20
## 12 Trump      Republican 2017-01-20 2021-01-20
```

```
#lag gives previous values by number of
#tows here we have n=1 whci means just the previous row
mutate(data4, previous = lag(name, n=1))
```

```
## # A tibble: 12 x 5
##    name       party      start      end        previous
##    <chr>      <chr>      <date>     <date>     <chr>
##  1 Eisenhower Republican 1953-01-20 1961-01-20 <NA>
##  2 Kennedy    Democratic 1961-01-20 1963-11-22 Eisenhower
##  3 Johnson    Democratic 1963-11-22 1969-01-20 Kennedy
##  4 Nixon      Republican 1969-01-20 1974-08-09 Johnson
##  5 Ford       Republican 1974-08-09 1977-01-20 Nixon
##  6 Carter     Democratic 1977-01-20 1981-01-20 Ford
##  7 Reagan     Republican 1981-01-20 1989-01-20 Carter
##  8 Bush       Republican 1989-01-20 1993-01-20 Reagan
##  9 Clinton    Democratic 1993-01-20 2001-01-20 Bush
## 10 Bush       Republican 2001-01-20 2009-01-20 Clinton
## 11 Obama      Democratic 2009-01-20 2017-01-20 Bush
## 12 Trump      Republican 2017-01-20 2021-01-20 Obama
```

Piping

allows us to nest output of one function into the other

```
data5 <- data %>% mutate(duration = end-start) %>%
  mutate(terms = case_when(duration == 2922 ~ "Two",
                           duration == 1461 ~ "One",
                           TRUE ~ "Abnormal")) %>%
  filter(terms == "Two")
data5
```

```
## # A tibble: 5 x 6
##   name       party      start      end        duration  terms
##   <chr>      <chr>      <date>     <date>     <drtn>    <chr>
## 1 Eisenhower Republican 1953-01-20 1961-01-20 2922 days Two
## 2 Reagan     Republican 1981-01-20 1989-01-20 2922 days Two
## 3 Clinton    Democratic 1993-01-20 2001-01-20 2922 days Two
## 4 Bush       Republican 2001-01-20 2009-01-20 2922 days Two
## 5 Obama      Democratic 2009-01-20 2017-01-20 2922 days Two
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

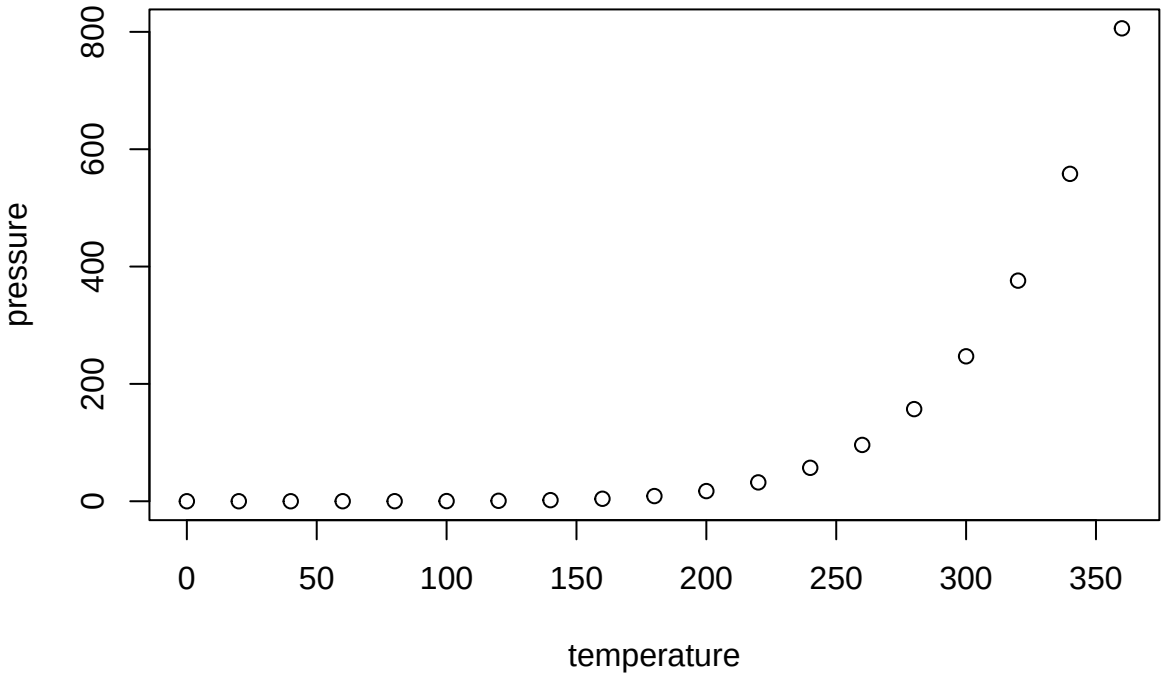
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.