The Tidyverse Data Wrangling & Visualizing

Ni

Objectives

- filter for particular observations
- arrange the observations in a desired order
- mutate to add ot change column

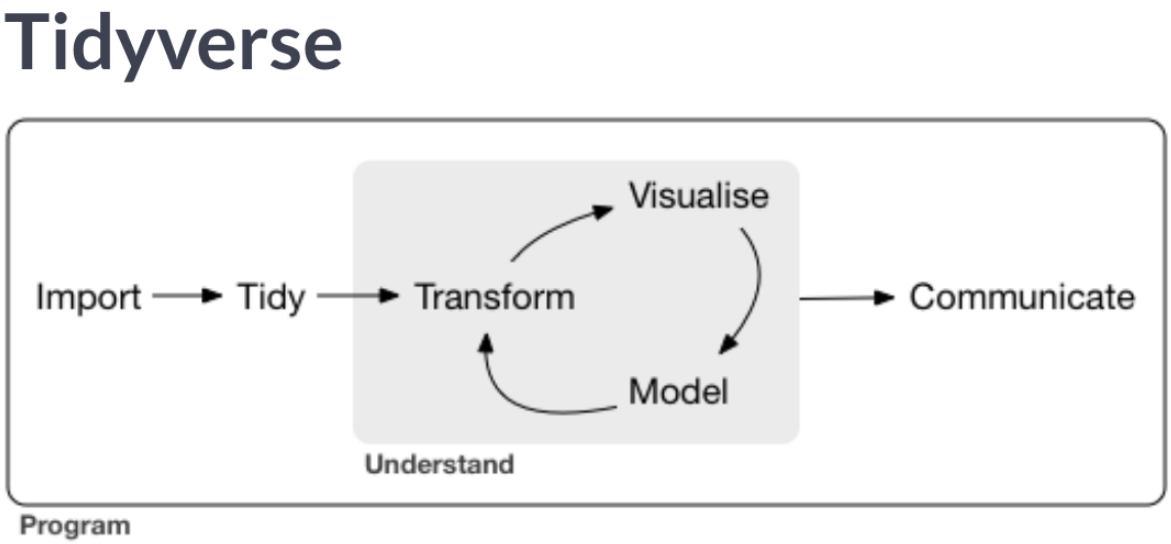


Figure 1: The tidyverse.

```
# Load the gapminder package, which contains the dataset that will be analyzed
library(gapminder)
```

```
# dplyr has tools that can transform the data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
head(gapminder)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

```
glimpse(gapminder)
```

```
## Rows: 1,704
## Columns: 6
## $ country   <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan, Afgha...
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asi...
## $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 199...
## $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 4...
## $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372,...
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.113...
```

The gapminder is a special type of data frame called a tibble

The GDP per capita is the country's total economic output (Gross Domestic Product) divided by its population, and it's common measure iof how wealthy a country is.

We can extract a few insights from this small glimpse of the data. For example, we can see that Afghanistan's life expectancy and population have both gone up from 1952 to 1997, but the GDP per capita has wavered.

We will get more insights about the social and economic history of countries around the world.

filter

- %>% means "take whatever is before it, and feed it into the next step."

```
gapminder_multiFilter<-gapminder %>%
  filter(year==2007,country=="United States")
gapminder_multiFilter
```

```
## # A tibble: 1 x 6
##   country       continent  year lifeExp       pop gdpPercap
##   <fct>         <fct>     <int>   <dbl>     <int>     <dbl>
## 1 United States Americas   2007    78.2 301139947    42952.
```

arrange

- arrange() sorts a table based on a variable
- arrange(desc()) from hight to low

```
# filter the specific year and arrange
gdpPerCap2007<-gapminder%>%
  filter(year==2007)%>%
  arrange(desc(gdpPercap))
```

mutate

- mutate changes or adds variables

```
# Combing verbs
gdp_sorted<-gapminder%>%
  mutate(gdp=gdpPercap*pop) %>%
  arrange(desc(gdp))
head(gdp_sorted)
```

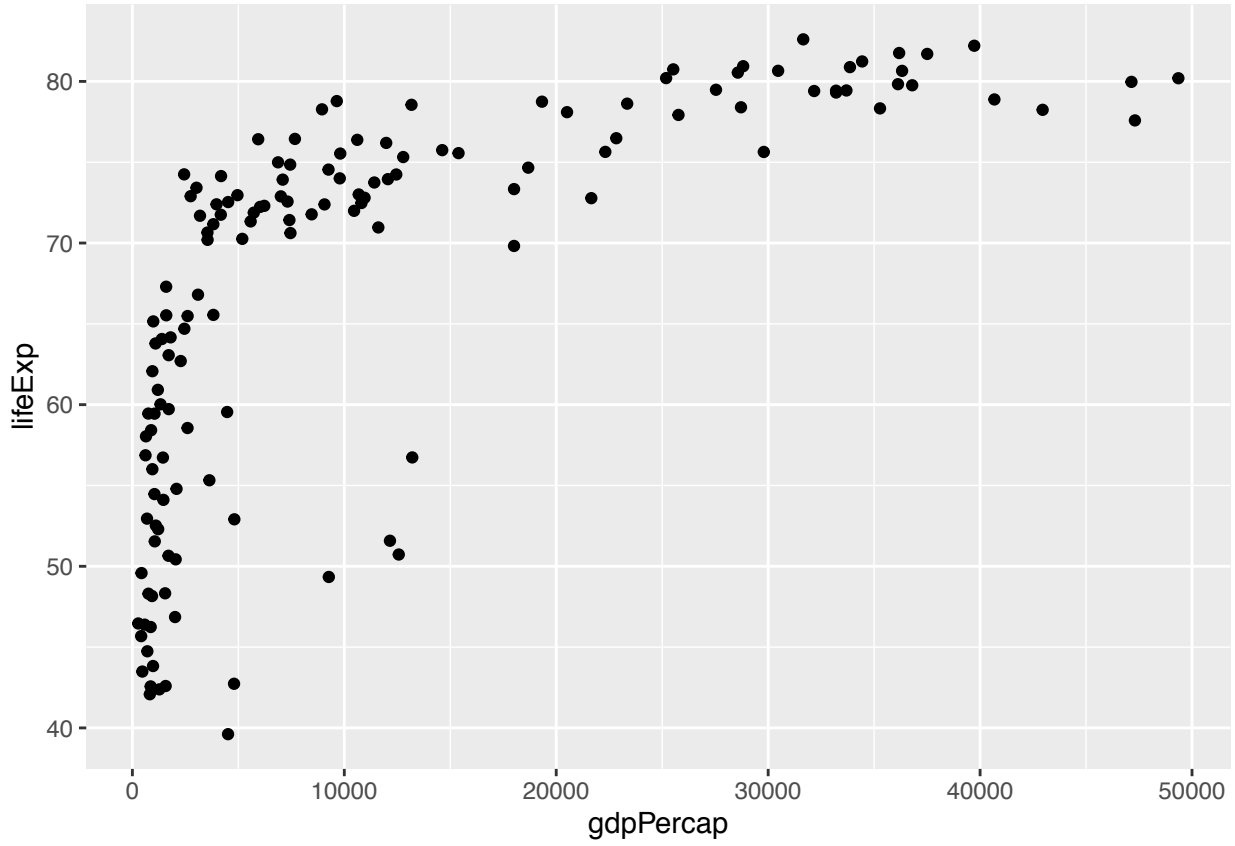
```
## # A tibble: 6 x 7
##   country       continent  year lifeExp        pop gdpPercap     gdp
##   <fct>         <fct>     <int>   <dbl>      <int>     <dbl>   <dbl>
## 1 United States Americas   2007    78.2  301139947    42952. 1.29e13
## 2 United States Americas   2002    77.3  287675526    39097. 1.12e13
## 3 United States Americas   1997    76.8  272911760    35767. 9.76e12
## 4 United States Americas   1992    76.1  256894189    32004. 8.22e12
## 5 United States Americas   1987    75.0  242803533    29884. 7.26e12
## 6 China         Asia       2007    73.0 1318683096     4959. 6.54e12
```

Visualizing with ggplot2

```
gapminder_2007<- gapminder %>%
      filter(year==2007)
head(gapminder_2007)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       2007    43.8 31889923      975.
## 2 Albania     Europe     2007    76.4  3600523     5937.
## 3 Algeria     Africa     2007    72.3 33333216     6223.
## 4 Angola      Africa     2007    42.7 12420476     4797.
## 5 Argentina   Americas   2007    75.3 40301927    12779.
## 6 Australia   Oceania    2007    81.2 20434176    34435.
```

```
ggplot(gapminder_2007,aes(x=gdpPercap,y=lifeExp))+
  geom_point()
```

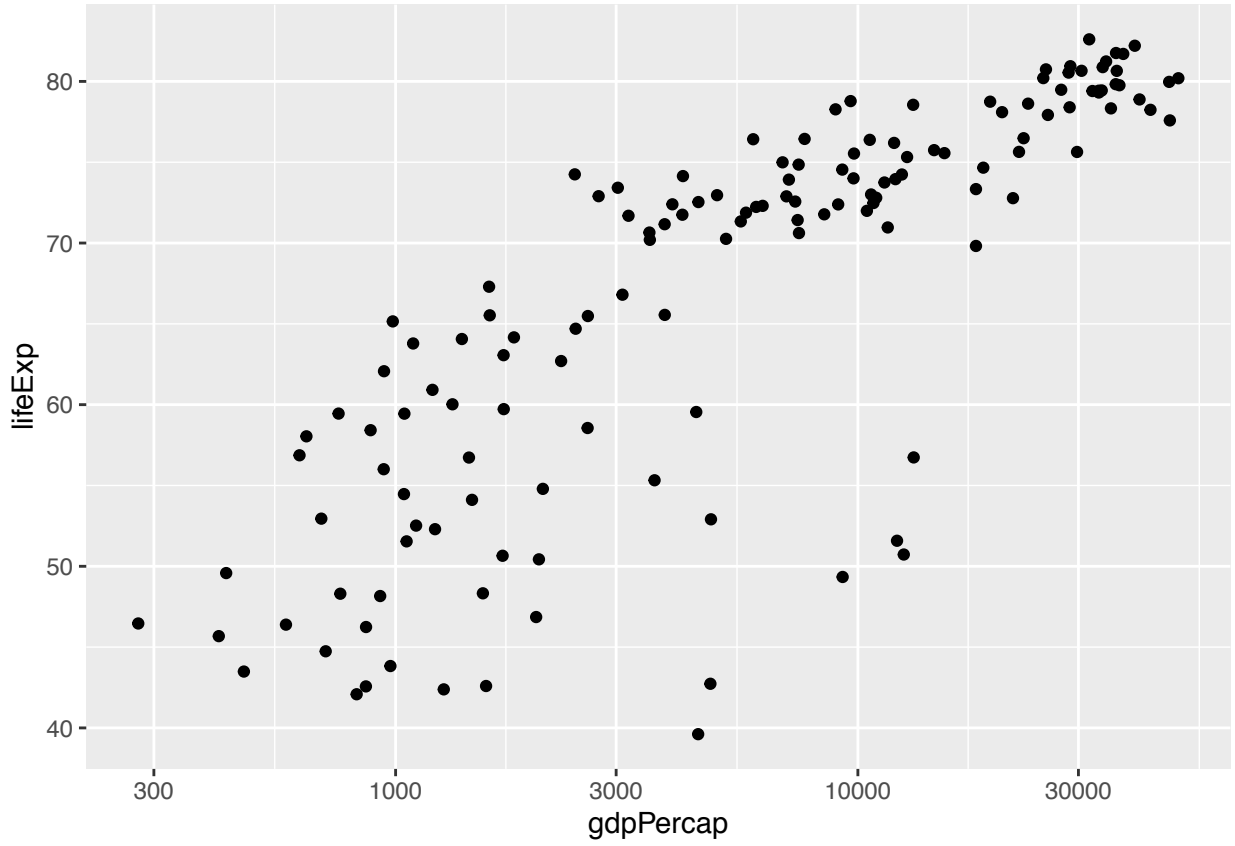


We can see that higher income countries tend to have higher life expectancy.

The problem with this plot is that a lot of countries get crammed into the leftmost part of the x-axis.This is becase the distribution of GDP per capita spans several orders of magnitude, with some countries in tens of thousands of dollars and others in hundreds.

We can use a logarithmic scale to fix distance that represents a multiplication of the value.

```
# Log scale
ggplot(gapminder_2007,aes(x=gdpPercap,y=lifeExp))+
  geom_point() +
  scale_x_log10() # specify x on a log scale
```



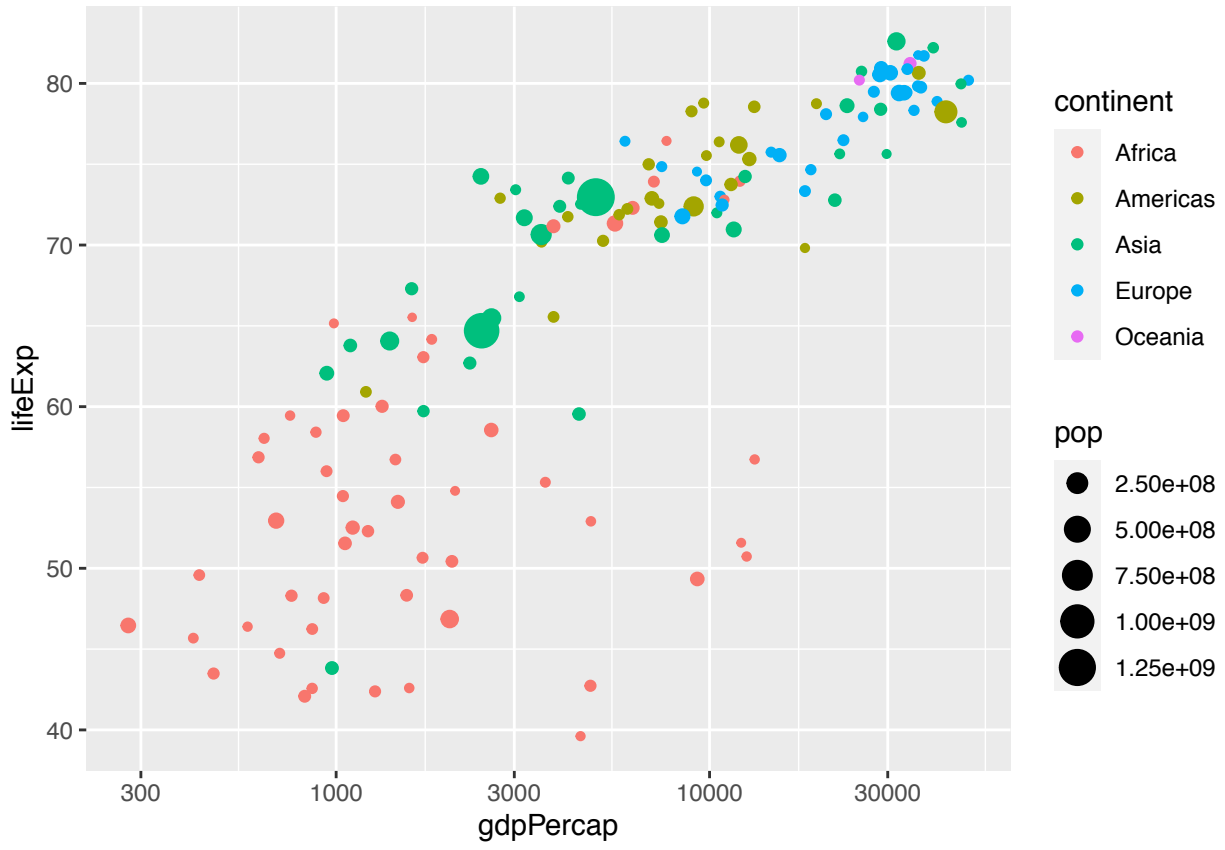
Each unit on the x-axis represents a change of 10 times the GDP.

On this scale, the relationship between GDP per capita and life expetancy looks more linear.We can moree easily distinguish the contries at the lower end of the specturm.

Additional variable

- Continent is a categorical variable, it has a few specific values such as Asia and Europe.
- A good way to represent a categorical variable is the color of the point *The color aesthestic*
- A good way to represent a numberical variable is the size *The size aesthetic*

```
ggplot(gapminder_2007,aes(x=gdpPercap,y=lifeExp,color=continent,size=pop))+
  geom_point() +
  scale_x_log10()
```

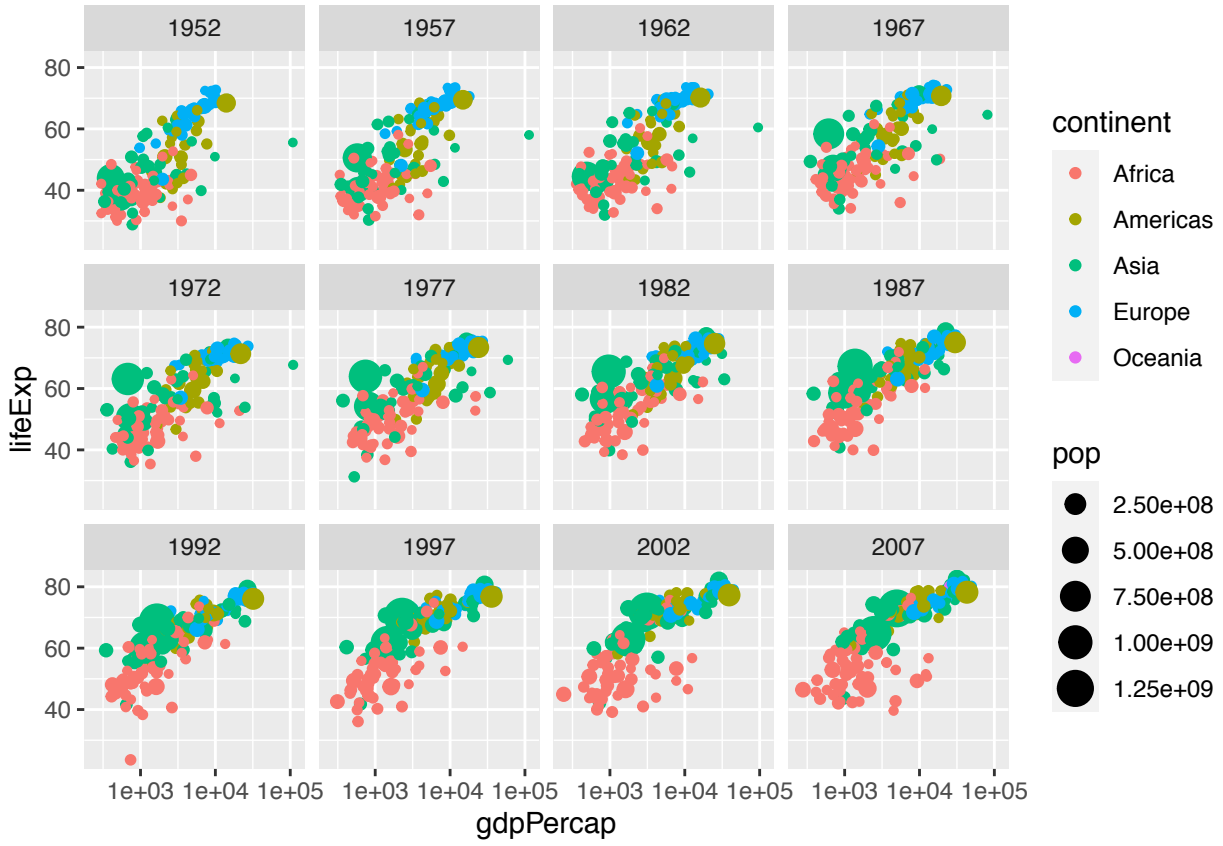


Faceting: another way to communicate relationships with the data

- dividing the plot into a subplot to get one smaller graph
- ~ the tilde symbo means "by", meaning spliting a plot by a variable

```
ggplot(gapminder_2007,aes(x=gdpPercap,y=lifeExp,color=continent,size=pop))+
  geom_point() +
  scale_x_log10()+
  facet_wrap(~continent)
```


```
ggplot(gapminder,aes(x=gdpPercap,y=lifeExp,color=continent,size=pop))+
  geom_point()+
  scale_x_log10()+
  facet_wrap(~year)
```



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.

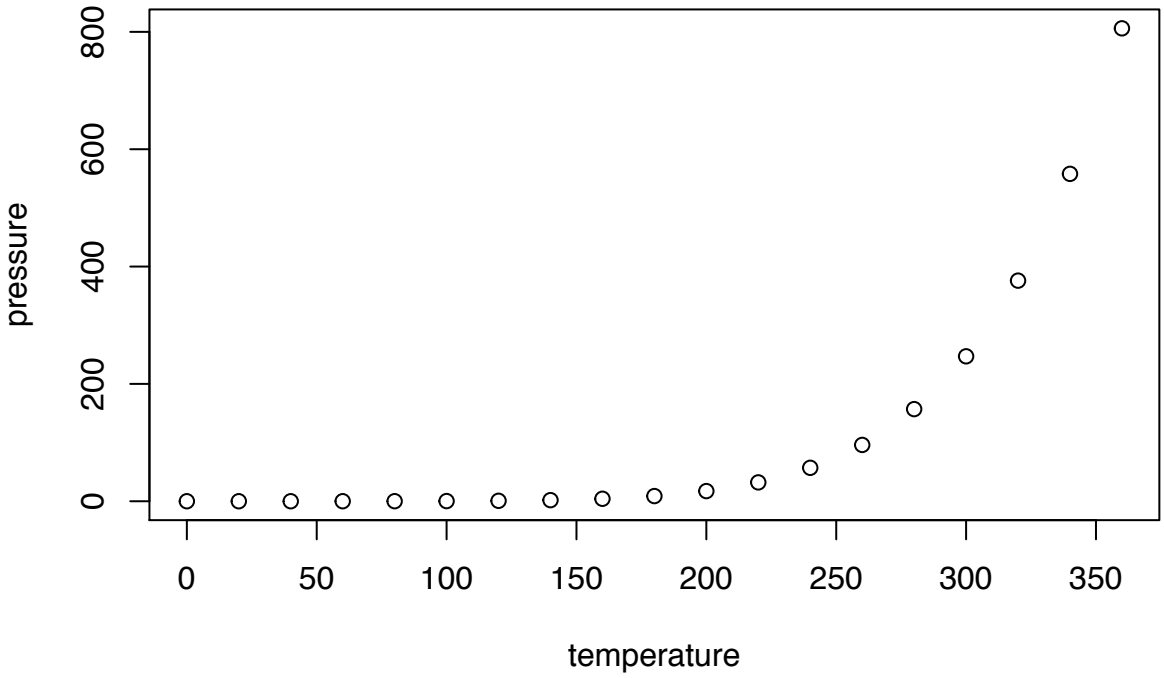
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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.