Intro R - Importing Data

Step 1: Load the `tidyverse` package.

Step2: Download the file `titanic.csv` from the datasets directory on Google Drive. Read it into R using `read_csv()` from the readr package. Explore the data to make sure it was read in properly.

```
titanic <- read_csv("../../datasets/titanic.csv")
```

```
## Rows: 1309 Columns: 15
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): Survived, Name, Sex, Ticket #, Cabin, Port, Lifeboat, Home / Destin...
## dbl (7): Passenger Class, Age, Siblings and Spouses, Parents and Children, F...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(titanic)
```

```
## spec_tbl_df [1,309 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Passenger Class     : num [1:1309] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Survived            : chr [1:1309] "Yes" "Yes" "No" "No" ...
##  $ Name                : chr [1:1309] "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevo
##  $ Sex                 : chr [1:1309] "female" "male" "female" "male" ...
##  $ Age                 : num [1:1309] 29 0.917 2 30 25 ...
##  $ Siblings and Spouses: num [1:1309] 0 1 1 1 1 0 1 0 2 0 ...
##  $ Parents and Children: num [1:1309] 0 2 2 2 2 0 0 0 0 0 ...
##  $ Ticket #            : chr [1:1309] "24160" "113781" "113781" "113781" ...
##  $ Fare                : num [1:1309] 211 152 152 152 152 ...
##  $ Cabin               : chr [1:1309] "B5" "C22 C26" "C22 C26" "C22 C26" ...
##  $ Port                : chr [1:1309] "S" "S" "S" "S" ...
##  $ Lifeboat            : chr [1:1309] "2" "11" NA NA ...
##  $ Body                : num [1:1309] NA NA NA 135 NA NA NA NA NA 22 ...
##  $ Home / Destination  : chr [1:1309] "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ
##  $ Midpoint age        : num [1:1309] 27.5 2.5 2.5 32.5 27.5 47.5 62.5 37.5 52.5 72.5 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Passenger Class` = col_double(),
##   ..   Survived = col_character(),
##   ..   Name = col_character(),
##   ..   Sex = col_character(),
##   ..   Age = col_double(),
##   ..   `Siblings and Spouses` = col_double(),
##   ..   `Parents and Children` = col_double(),
##   ..   `Ticket #` = col_character(),
##   ..   Fare = col_double(),
##   ..   Cabin = col_character(),
##   ..   Port = col_character(),
```

```
##   ..   Lifeboat = col_character(),
##   ..   Body = col_double(),
##   ..   `Home / Destination` = col_character(),
##   ..   `Midpoint age` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(titanic)
```

```
## # A tibble: 6 x 15
##   `Passenger Class` Survived Name  Sex      Age `Siblings and S~ `Parents and Ch~
##               <dbl> <chr>    <chr> <chr>  <dbl>            <dbl>            <dbl>
## 1                 1 Yes      Alle~ fema~ 29                   0                0
## 2                 1 Yes      Alli~ male   0.917               1                2
## 3                 1 No       Alli~ fema~  2                   1                2
## 4                 1 No       Alli~ male  30                   1                2
## 5                 1 No       Alli~ fema~ 25                   1                2
## 6                 1 Yes      Ande~ male  48                   0                0
## # ... with 8 more variables: Ticket # <chr>, Fare <dbl>, Cabin <chr>,
## #   Port <chr>, Lifeboat <chr>, Body <dbl>, Home / Destination <chr>,
## #   Midpoint age <dbl>
```

```
summary(titanic)
```

```
##  Passenger Class   Survived            Name               Sex
##  Min.   :1.000   Length:1309        Length:1309        Length:1309
##  1st Qu.:2.000   Class :character   Class :character   Class :character
##  Median :3.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2.295
##  3rd Qu.:3.000
##  Max.   :3.000
##
##       Age          Siblings and Spouses Parents and Children   Ticket #
##  Min.   : 0.1667   Min.   :0.0000       Min.   :0.000        Length:1309
##  1st Qu.:21.0000   1st Qu.:0.0000       1st Qu.:0.000        Class :character
##  Median :28.0000   Median :0.0000       Median :0.000        Mode  :character
##  Mean   :29.8811   Mean   :0.4989       Mean   :0.385
##  3rd Qu.:39.0000   3rd Qu.:1.0000       3rd Qu.:0.000
##  Max.   :80.0000   Max.   :8.0000       Max.   :9.000
##  NA's   :263
##       Fare            Cabin              Port             Lifeboat
##  Min.   :  0.000   Length:1309        Length:1309        Length:1309
##  1st Qu.:  7.896   Class :character   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 33.295
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
##       Body       Home / Destination  Midpoint age
##  Min.   :  1.0   Length:1309        Min.   : 2.50
##  1st Qu.: 72.0   Class :character   1st Qu.:22.50
##  Median :155.0   Mode  :character   Median :27.50
##  Mean   :160.8                      Mean   :30.41
##  3rd Qu.:256.0                      3rd Qu.:37.50
##  Max.   :328.0                      Max.   :82.50
```

```
##  NA's   :1188                      NA's   :263
```

```
table(titanic$Sex)
```

```
##
## female   male
##    466    843
```

```
table(titanic$Survived)
```

```
##
##  No Yes
## 809 500
```

Step 3: Compare the file when you read it in using `read.csv()` (base `R`) versus `read_csv()` (`readr`). After exploring, what do you see different?

```
titanic_other <- read.csv("../../datasets/titanic.csv")
head(titanic_other)
```

```
##   Passenger.Class Survived                                          Name
## 1               1      Yes                 Allen, Miss. Elisabeth Walton
## 2               1      Yes                Allison, Master. Hudson Trevor
## 3               1       No                  Allison, Miss. Helen Loraine
## 4               1       No          Allison, Mr. Hudson Joshua Creighton
## 5               1       No Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
## 6               1      Yes                           Anderson, Mr. Harry
##      Sex     Age Siblings.and.Spouses Parents.and.Children Ticket..     Fare
## 1 female 29.0000                    0                    0    24160 211.3375
## 2   male  0.9167                    1                    2   113781 151.5500
## 3 female  2.0000                    1                    2   113781 151.5500
## 4   male 30.0000                    1                    2   113781 151.5500
## 5 female 25.0000                    1                    2   113781 151.5500
## 6   male 48.0000                    0                    0    19952  26.5500
##     Cabin Port Lifeboat Body              Home...Destination Midpoint.age
## 1      B5    S        2   NA                    St Louis, MO         27.5
## 2 C22 C26    S       11   NA Montreal, PQ / Chesterville, ON          2.5
## 3 C22 C26    S           NA Montreal, PQ / Chesterville, ON          2.5
## 4 C22 C26    S          135 Montreal, PQ / Chesterville, ON         32.5
## 5 C22 C26    S           NA Montreal, PQ / Chesterville, ON         27.5
## 6     E12    S        3   NA                    New York, NY         47.5
```

Step 4: Download the file `manufacturing.xlsx` from the datasets directory on Google Drive. Read it into R using `read_xlsx()` from the `readxl` package. Explore the data to make sure it was read in properly.

```
manufacturing <- read_xlsx("../../datasets/manufacturing.xlsx",
                           sheet = 2)
head(manufacturing)
```

```
## # A tibble: 6 x 5
##   Output NumEmp PlantAge Product Shift
##    <dbl>  <dbl>    <dbl> <chr>   <chr>
## 1   2793    199        5 B       Night
## 2   1912    147       13 A       Day
## 3   2663    184        7 B       Night
## 4   2389    174        8 A       Day
## 5   2517    185        7 A       Day
## 6   2793    196        9 B       Night
```

```
manufacturing
```

```
## # A tibble: 159 x 5
##    Output NumEmp PlantAge Product Shift
##     <dbl>  <dbl>    <dbl> <chr>   <chr>
##  1   2793    199        5 B       Night
##  2   1912    147       13 A       Day
##  3   2663    184        7 B       Night
##  4   2389    174        8 A       Day
##  5   2517    185        7 A       Day
##  6   2793    196        9 B       Night
##  7   2542    178       10 A       Day
##  8   2420    191        4 A       Day
##  9   2500    182        8 A       Day
## 10   2440    175        8 A       Day
## # ... with 149 more rows
```

```
summary(manufacturing)
```

```
##      Output          NumEmp         PlantAge       Product
##  Min.   :1912   Min.   :147.0   Min.   : 2.00   Length:159
##  1st Qu.:2326   1st Qu.:175.0   1st Qu.: 6.00   Class :character
##  Median :2516   Median :185.0   Median : 7.00   Mode  :character
##  Mean   :2499   Mean   :185.8   Mean   : 7.27
##  3rd Qu.:2652   3rd Qu.:195.0   3rd Qu.: 9.00
##  Max.   :3325   Max.   :257.0   Max.   :13.00
##     Shift
##  Length:159
##  Class :character
##  Mode  :character
##
##
##
```

```
table(manufacturing$Product)
```

```
##
##   A   B
## 104  55
```

```
table(manufacturing$Shift)
```

```
##
##   Day Night
##    97    62
```