R6 Exercises: Import data

Contents

Packages used in this notebook:

---

1 Import text files

Try to read in the four text files `data1.txt`, `data2.txt`, `data3.txt` and `data4.txt` in `data.zip` using `read_delim()`. Comment on (solvable/unsolvable) problems during reading in the data.

1.1 `data1.txt` (3P)

Why is it not possible to read in `data1.txt` correctly?

```
## # A tibble: 93 x 5
##      row   col expected   actual     file
##    <int> <int> <chr>      <chr>      <chr>
##  1     2    34 28 columns 34 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  2     3    32 28 columns 32 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  3     4    33 28 columns 33 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  4     5    33 28 columns 33 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  5     6    32 28 columns 32 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  6     7    33 28 columns 33 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  7     8    33 28 columns 33 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  8     9    33 28 columns 33 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  9    10    34 28 columns 34 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
## 10    11    31 28 columns 31 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
## # ... with 83 more rows
```

1.2 data2.txt (3P)

```
## # A tibble: 0 x 5
## # ... with 5 variables: row <int>, col <int>, expected <chr>, actual <chr>,
## #   file <chr>
```

1.3 data3.txt (3P)

```
## # A tibble: 6 x 10
##   X1           X2      X3      X4        X5    X6        X7    X8    X9    X10
##   <chr>        <chr>   <chr>   <chr>     <chr> <chr>     <chr> <chr> <chr> <chr>
## 1 Manufacturer Model   Type    Min Price Price Max Price MPG ~ MPG ~ AirB~ Driv~
## 2 Acura        Integra Small   12.9      15.9  18.8      25    31    None  Front
## 3 Acura        Legend  Midsize 29.2      33.9  38.7      18    25    Driv~ Front
## 4 Audi         90      Compact 25.9      29.1  32.3      20    26    Driv~ Front
## 5 Audi         100     Midsize 30.8      37.7  44.6      19    26    Driv~ Front
## 6 BMW          535i    Midsize 23.7      30.0  36.2      22    30    Driv~ Rear
```

1.4 data4.txt (3P)

```
## # A tibble: 93 x 5
##      row   col expected   actual     file
##    <int> <int> <chr>      <chr>      <chr>
##  1     2    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  2     3    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  3     4    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  4     5    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  5     6    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  6     7    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  7     8    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  8     9    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
##  9    10    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
## 10    11    28 27 columns 28 columns C:/Users/reschrei/Documents/Teaching/_DE_M~
## # ... with 83 more rows
```

Using `read.table()` command instead of `read_delim()`:

```
##   Manufacturer   Model    Type Min.Price Price Max.Price MPG.city MPG.highway
## 1        Acura Integra   Small      12.9  15.9      18.8       25          31
## 2        Acura  Legend Midsize      29.2  33.9      38.7       18          25
## 3         Audi      90 Compact      25.9  29.1      32.3       20          26
## 4         Audi     100 Midsize      30.8  37.7      44.6       19          26
## 5          BMW    535i Midsize      23.7  30.0      36.2       22          30
##              AirBags DriveTrain Cylinders EngineSize Horsepower  RPM
## 1               None      Front         4        1.8        140 6300
## 2 Driver & Passenger      Front         6        3.2        200 5500
## 3        Driver only      Front         6        2.8        172 5500
## 4 Driver & Passenger      Front         6        2.8        172 5500
## 5        Driver only       Rear         4        3.5        208 5700
##   Rev.per.mile Man.trans.avail Fuel.tank.capacity Passengers Length Wheelbase
## 1         2890             Yes               13.2          5    177       102
## 2         2335             Yes               18.0          5    195       115
```

```
## 3         2280            Yes               16.9          5    180       102
## 4         2535            Yes               21.1          6    193       106
## 5         2545            Yes               21.1          4    186       109
##   Width Turn.circle Rear.seat.room Luggage.room Weight  Origin          Make
## 1    68          37           26.5           11   2705 non-USA Acura Integra
## 2    71          38             30           15   3560 non-USA  Acura Legend
## 3    67          37             28           14   3375 non-USA       Audi 90
## 4    70          37             31           17   3405 non-USA      Audi 100
## 5    69          39             27           13   3640 non-USA      BMW 535i
```

2 Separating and joining columns (6P)

Using the data below, transform the birth date into the format `YYYY-MM-DD`. Try to pad days and months with a leading 0, so that, e.g., `1.1.1988` becomes `1988-01-01`. (Hint: use `mutate()` with `str_pad()`).

```
tribble(~Name, ~Birthdate,
        "Susan", "29.10.1966",
        "Will", "1.1.1988",
        "Chris", "10.10.1977")
```

```
## # A tibble: 3 x 2
##   Name  Birtday
##   <chr> <chr>
## 1 Susan 1966-10-29
## 2 Will  1988-01-01
## 3 Chris 1977-10-10
```

3 Missing data (7P)

Using the data below, find all rows with missing data, impute missing invitations with 0, missing ages with the average age and remove all rows with other missing data.

```
tribble(~Name, ~Age, ~Invitations, ~Phone,
        "Tim", 20, 0, "123 345",
        "Mary", 30, 12, "321 999",
        "Chris", 25, NA, "444 324",
        "Lilly", NA, 0, "453 424",
        "Will", 20, 0, NA
)
```

All rows with missing data:

```
## # A tibble: 3 x 4
##   Name    Age Invitations Phone
##   <chr> <dbl>       <dbl> <chr>
## 1 Chris    25          NA 444 324
## 2 Lilly    NA           0 453 424
## 3 Will     20           0 <NA>
```

Impute missing invitations with 0 values:

```
## # A tibble: 5 x 4
##   Name    Age Invitations Phone
##   <chr> <dbl>       <dbl> <chr>
## 1 Tim      20           0 123 345
## 2 Mary     30          12 321 999
## 3 Chris    25           0 444 324
## 4 Lilly    NA           0 453 424
## 5 Will     20           0 <NA>
```

Impute missing ages with the average age:

```
## # A tibble: 5 x 4
##   Name    Age Invitations Phone
##   <chr> <dbl>       <dbl> <chr>
## 1 Tim    20             0 123 345
## 2 Mary   30            12 321 999
## 3 Chris  25             0 444 324
## 4 Lilly  23.8           0 453 424
## 5 Will   20             0 <NA>
```

Remove all rows with other missing data:

```
## # A tibble: 4 x 4
##   Name    Age Invitations Phone
##   <chr> <dbl>       <dbl> <chr>
## 1 Tim    20             0 123 345
## 2 Mary   30            12 321 999
## 3 Chris  25             0 444 324
## 4 Lilly  23.8           0 453 424
```