Project 2

Jawaid Hakim

2022-09-28

Contents

1 Assignment

1.1 Data set 1 from Jhalak Das

Let’s load a CSV file containing the data.

```
df <- read.csv("jhalak_das.csv")
```

A quick look at the data frame structure shows that the CSV contains a few **wider** data characteristic:

1. **sex and age** are concatenated together in one column
2. Terms scores separate columns

```
str(df)
```

```
## ’data.frame':    10 obs. of  8 variables:
##  $ id         : num  1 2 3 4 5 1 2 3 4 5
##  $ name       : chr  "Mike   " "Linda   " " Sam     " "Esther" ...
##  $ phone      : num  134 270 210 617 114 134 270 210 617 114
##  $ sex.and.age: chr  " m_12   " "f_13    " "  m_11   " " f_12 " ...
##  $ test.number: chr  "test 1       " "test 1       " " test 1     " "   test 1  " ...
##  $ term.1     : num  76 88 78 68 65 85 87 80 70 68
##  $ term.2     : num  84 90 74 75 67 80 82 87 75 70
##  $ term.3     : num  87 73 80 74 64 90 94 80 78 63
```

```
head(df)
```

```
##   id     name phone sex.and.age   test.number term.1 term.2 term.3
## 1  1  Mike      134     m_12    test 1             76     84     87
## 2  2 Linda      270    f_13     test 1             88     90     73
## 3  3  Sam       210     m_11      test 1           78     74     80
## 4  4   Esther   617       f_12       test 1        68     75     74
## 5  5   Mary     114     f_14     test 1            65     67     64
## 6  1  Mike      134     m_12       test 2          85     80     90
```

Let's start by separating **send and age** column into **sex** and **age** columns. Note, during separation we also convert **age** into integer.

```
df <- df %>%
    separate('sex.and.age', sep = '_', into = c('sex', 'age'), convert = TRUE)
head(df)
```

```
##   id     name phone  sex age    test.number term.1 term.2 term.3
## 1  1  Mike      134    m  12 test 1            76     84     87
## 2  2 Linda      270    f  13 test 1            88     90     73
## 3  3  Sam       210    m  11   test 1          78     74     80
## 4  4   Esther   617    f  12      test 1       68     75     74
## 5  5   Mary     114    f  14  test 1           65     67     64
## 6  1  Mike      134    m  12     test 2        85     80     90
```

Next, we **pivot longer** by turning the terms

```
df <- df %>%
        pivot_longer(cols = 7:9,
                     names_to = "term",         # col with term names
                     values_to = "score")        # col with scores
head(df)
```

```
## # A tibble: 6 x 8
##      id name       phone sex     age test.number     term   score
##   <dbl> <chr>      <dbl> <chr> <dbl> <chr>           <chr>  <dbl>
## 1     1 "Mike   "    134 " m"     12 "test 1       " term.1    76
## 2     1 "Mike   "    134 " m"     12 "test 1       " term.2    84
## 3     1 "Mike   "    134 " m"     12 "test 1       " term.3    87
## 4     2 "Linda   "   270 "f"      13 "test 1       " term.1    88
## 5     2 "Linda   "   270 "f"      13 "test 1       " term.2    90
## 6     2 "Linda   "   270 "f"      13 "test 1       " term.3    73
```

Clean up strings: uppercase, trim, replace '.' with space.

```
df <- df %>%
        mutate(name = str_to_upper(str_trim(name))) %>%
        mutate(sex = str_to_upper(str_trim(sex))) %>%
        mutate(test.number = str_to_upper(str_trim(test.number))) %>%
        mutate(term = str_replace(str_to_upper(str_trim(term)), '\\.', ' '))
head(df)
```

```
## # A tibble: 6 x 8
##      id name  phone sex     age test.number term   score
##   <dbl> <chr> <dbl> <chr> <dbl> <chr>       <chr>  <dbl>
## 1     1 MIKE    134 M        12 TEST 1      TERM 1    76
## 2     1 MIKE    134 M        12 TEST 1      TERM 2    84
## 3     1 MIKE    134 M        12 TEST 1      TERM 3    87
## 4     2 LINDA   270 F        13 TEST 1      TERM 1    88
## 5     2 LINDA   270 F        13 TEST 1      TERM 2    90
## 6     2 LINDA   270 F        13 TEST 1      TERM 3    73
```

Next, we separate the table into two: first containing the student personal details (id, name, phone, sex)

```
df_personal <- df %>%
            select(id, name, phone, sex) %>%
            distinct(id, .keep_all = TRUE) %>%
            arrange(id)
head(df_personal)
```

```
## # A tibble: 5 x 4
##      id name   phone sex
##   <dbl> <chr>  <dbl> <chr>
## 1     1 MIKE     134 M
## 2     2 LINDA    270 F
## 3     3 SAM      210 M
## 4     4 ESTHER   617 F
## 5     5 MARY     114 F
```

And another containing the test scores (id, name, test.number, term, score).

```
df_test <- df %>%
            select(id, name, test.number, term, score) %>%
            distinct(id, .keep_all = TRUE) %>%
            arrange(id)
head(df_test)
```

```
## # A tibble: 5 x 5
##      id name   test.number term   score
##   <dbl> <chr>  <chr>       <chr>  <dbl>
## 1     1 MIKE   TEST 1      TERM 1    76
## 2     2 LINDA  TEST 1      TERM 1    88
## 3     3 SAM    TEST 1      TERM 1    78
## 4     4 ESTHER TEST 1      TERM 1    68
## 5     5 MARY   TEST 1      TERM 1    65
```