# SOC4001 Procesamiento avanzado de bases de datos en R

Tarea 4

Ponderación: 12% de la nota final del curso

Entrega: Desde el momento de entrega, los estudiantes tienen plazo hasta el domingo 15 de Noviembre a las 23:59pm para completar esta tarea.

Formato: Desarrollar esta tarea en un RScript, agregando comentarios cuando sea necesario.

1) Carga la base de datos "Salaries" del paquete `carData` y crea un `tibble` que los contenga los datos. Llama tal objeto "datos_salarios". Lee descripción de los datos y carga la librería `tidyverse`.

Los datos deben verse así:

```
## # A tibble: 397 x 6
##    rank      discipline yrs.since.phd yrs.service sex    salary
##    <fct>     <fct>              <int>       <int> <fct>   <int>
##  1 Prof      B                     19          18 Male   139750
##  2 Prof      B                     20          16 Male   173200
##  3 AsstProf  B                      4           3 Male    79750
##  4 Prof      B                     45          39 Male   115000
##  5 Prof      B                     40          41 Male   141500
##  6 AssocProf B                      6           6 Male    97000
##  7 Prof      B                     30          23 Male   175000
##  8 Prof      B                     45          45 Male   147765
##  9 Prof      B                     21          20 Male   119250
## 10 Prof      B                     18          18 Female 129000
## # ... with 387 more rows
```

2) Usando los comandos `group_by()` y `summarise()` produce la siguiente tabla y asígnala al objeto "tabla_1":

```
## # A tibble: 6 x 6
## # Groups:   rank [3]
##   rank    discipline yrs.since.phd_med~ yrs.since.phd_m~ salary_media salary_max
##   <fct>   <fct>                   <dbl>            <int>        <dbl>      <int>
## 1 AsstPr~ A                        5.67               11       73936.      85000
## 2 AsstPr~ B                        4.79               11       84594.      97032
## 3 AssocP~ A                       17.8                49       83061.     108413
## 4 AssocP~ B                       13.8                48      101276.     126431
## 5 Prof    A                       30.5                56      119948.     205500
## 6 Prof    B                       26.2                56      133394.     231545
```

3) La siguiente base de datos ("disciplinas") contiene diferentes disciplinas con sus respectivos nombres.

```
disciplinas <- tibble(discipline = c("A","B","C"),
                      names = c("theoretical departments", "applied departments", "other") )
```

Usando algunos de los comandos `_join()` junta los datos en "tabla_1" y "disciplinas" preservando toda la información disponible en ambas bases de datos. El resultado debe verse así:

```
## # A tibble: 7 x 7
## # Groups:   rank [4]
##   rank  discipline yrs.since.phd_m~ yrs.since.phd_m~ salary_media salary_max
##   <fct> <chr>                 <dbl>            <int>        <dbl>      <int>
## 1 Asst~ A                      5.67               11       73936.      85000
## 2 Asst~ B                      4.79               11       84594.      97032
## 3 Asso~ A                     17.8                49       83061.     108413
## 4 Asso~ B                     13.8                48      101276.     126431
## 5 Prof  A                     30.5                56      119948.     205500
## 6 Prof  B                     26.2                56      133394.     231545
## 7 <NA>  C                     NA                  NA          NA          NA
## # ... with 1 more variable: names <chr>
```

4) Usando el comando `pivot_longer()` produce la siguiente tabla:

```
## # A tibble: 28 x 5
## # Groups:   rank [4]
##    rank      discipline names                   var_stat               value
##    <fct>     <chr>      <chr>                   <chr>                  <dbl>
##  1 AsstProf  A          theoretical departments yrs.since.phd_media     5.67
##  2 AsstProf  A          theoretical departments yrs.since.phd_max      11
##  3 AsstProf  A          theoretical departments salary_media        73936.
##  4 AsstProf  A          theoretical departments salary_max          85000
##  5 AsstProf  B          applied departments     yrs.since.phd_media     4.79
##  6 AsstProf  B          applied departments     yrs.since.phd_max      11
##  7 AsstProf  B          applied departments     salary_media        84594.
##  8 AsstProf  B          applied departments     salary_max          97032
##  9 AssocProf A          theoretical departments yrs.since.phd_media    17.8
## 10 AssocProf A          theoretical departments yrs.since.phd_max      49
## # ... with 18 more rows
```

5) Usando el comando `separate()` modifica la tabla producida en (4) y produce la siguiente tabla:

```
## # A tibble: 28 x 6
## # Groups:   rank [4]
##    rank      discipline names                   variable      stat     value
##    <fct>     <chr>      <chr>                   <chr>         <chr>    <dbl>
##  1 AsstProf  A          theoretical departments yrs.since.phd media     5.67
##  2 AsstProf  A          theoretical departments yrs.since.phd max      11
##  3 AsstProf  A          theoretical departments salary        media 73936.
##  4 AsstProf  A          theoretical departments salary        max   85000
##  5 AsstProf  B          applied departments     yrs.since.phd media     4.79
##  6 AsstProf  B          applied departments     yrs.since.phd max      11
##  7 AsstProf  B          applied departments     salary        media 84594.
##  8 AsstProf  B          applied departments     salary        max   97032
##  9 AssocProf A          theoretical departments yrs.since.phd media    17.8
## 10 AssocProf A          theoretical departments yrs.since.phd max      49
## # ... with 18 more rows
```

6) Usando el comando `pivot_wider()` modifica la tabla producida en (5) y produce la siguiente tabla:

```
## # A tibble: 14 x 6
## # Groups:   rank [4]
##    rank      discipline names                   variable         media    max
##    <fct>     <chr>      <chr>                   <chr>            <dbl>  <dbl>
##  1 AsstProf  A          theoretical departments yrs.since.phd     5.67     11
##  2 AsstProf  A          theoretical departments salary        73936.    85000
##  3 AsstProf  B          applied departments     yrs.since.phd     4.79     11
##  4 AsstProf  B          applied departments     salary        84594.    97032
##  5 AssocProf A          theoretical departments yrs.since.phd    17.8      49
##  6 AssocProf A          theoretical departments salary        83061.   108413
##  7 AssocProf B          applied departments     yrs.since.phd    13.8      48
##  8 AssocProf B          applied departments     salary       101276.   126431
##  9 Prof      A          theoretical departments yrs.since.phd    30.5      56
## 10 Prof      A          theoretical departments salary       119948.   205500
## 11 Prof      B          applied departments     yrs.since.phd    26.2      56
## 12 Prof      B          applied departments     salary       133394.   231545
## 13 <NA>      C          other                   yrs.since.phd    NA        NA
## 14 <NA>      C          other                   salary           NA        NA
```

7) Usando los comando para tratar valores perdidos modifica la tabla producida en (6) y produce la siguiente tabla:

```
## # A tibble: 14 x 6
## # Groups:   rank [4]
##    rank      discipline names                   variable         media    max
##    <fct>     <chr>      <chr>                   <chr>            <dbl>  <dbl>
##  1 AsstProf  A          theoretical departments yrs.since.phd     5.67     11
##  2 AsstProf  A          theoretical departments salary        73936.    85000
##  3 AsstProf  B          applied departments     yrs.since.phd     4.79     11
##  4 AsstProf  B          applied departments     salary        84594.    97032
##  5 AssocProf A          theoretical departments yrs.since.phd    17.8      49
##  6 AssocProf A          theoretical departments salary        83061.   108413
##  7 AssocProf B          applied departments     yrs.since.phd    13.8      48
##  8 AssocProf B          applied departments     salary       101276.   126431
##  9 Prof      A          theoretical departments yrs.since.phd    30.5      56
## 10 Prof      A          theoretical departments salary       119948.   205500
## 11 Prof      B          applied departments     yrs.since.phd    26.2      56
## 12 Prof      B          applied departments     salary       133394.   231545
## 13 <NA>      C          other                   yrs.since.phd     0         0
## 14 <NA>      C          other                   salary            0         0
```