# SOC4001 Procesamiento avanzado de bases de datos en R

## Tarea 3

Ponderación: 12% de la nota final del curso

Entrega: Desde el momento de entrega, los estudiantes tienen plazo hasta el domingo 19 de Octubre a las 23:59pm para completar esta tarea.

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

El código a continuación carga la Base de Datos Histórica Proyectos Adjudicados ANID (ex-conicyt) y extrae una selección de variables que son almacenados en el objecto `data_anid`.

```
library("tidyverse")
library("readr")

path <- url("https://raw.githubusercontent.com/ANID-GITHUB/Historico-de-Proyectos-Adjudicados/da63cab4fa

data_anid <- read_delim(path, delim = ";")

data_anid <- data_anid %>% rename(codigo_proyecto = CODIGO_PROYECTO, anno = ANO_FALLO, sexo = SEXO, area
```

Descripción de los datos: La Agencia Nacional de Investigación y Desarrollo (ANID) cada año adjudica financiamiento para proyectos en Ciencia y Tecnología a través de sus diferentes concursos. La base de datos denominada "BDH_Proyectos" contiene la información disponible de proyectos adjudicados por la Agencia (antes del 2020, CONICYT) desde el año 1982 hasta el 2020, con fecha de corte al 31 de diciembre del 2020. Cada fila representa una iniciativa adjudicada. Los datos deben verse así:

```
## # A tibble: 6 x 5
##   codigo_proyecto  anno sexo   area                           monto
##             <dbl> <dbl> <chr>  <chr>                          <dbl>
## 1         1820005  1982 HOMBRE CIENCIAS NATURALES               300
## 2         1820006  1982 HOMBRE CIENCIAS MEDICAS Y DE LA SALUD   130
## 3         1820009  1982 HOMBRE CIENCIAS NATURALES               506
## 4         1820010  1982 HOMBRE HUMANIDADES                      335
## 5         1820015  1982 HOMBRE CIENCIAS NATURALES               260
## 6         1820043  1982 HOMBRE CIENCIAS AGRICOLAS               464
```

1) Usando los comandos `group_by()` y `summarise()` produce la siguiente tabla y asígnala al objeto `tabla_1`. El resultado debe verse así:

```
## # A tibble: 6 x 4
## # Groups:   area, anno [4]
##   area               anno sexo   monto_promedio
##   <chr>             <dbl> <chr>           <dbl>
## 1 CIENCIAS AGRICOLAS  1982 HOMBRE           408.
## 2 CIENCIAS AGRICOLAS  1982 MUJER            549
## 3 CIENCIAS AGRICOLAS  1983 HOMBRE           382.
## 4 CIENCIAS AGRICOLAS  1984 HOMBRE           355.
## 5 CIENCIAS AGRICOLAS  1984 MUJER            373
## 6 CIENCIAS AGRICOLAS  1985 HOMBRE           414.
```

2) Carga la base de datos con el IPC anual y guárdala en un objecto llamado `datos_ipc`. Para los años con valores perdidos en la variable `datos_ipc$ipc`, usa la función `fill()` para asignales el valor correspondiente al año siguiente. Conserva sólo las variables `anno` e `ipc`. Los datos deben verse así:

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   anno = col_double(),
##   ipc = col_double()
## )
```

```
## # A tibble: 6 x 2
##    anno   ipc
##   <dbl> <dbl>
## 1  1982  2.79
## 2  1983  8.98
## 3  1984  6.53
## 4  1985 10.1
## 5  1986  6.45
## 6  1987  6.54
```

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

```
## # A tibble: 6 x 5
## # Groups:   area, anno [4]
##   area                anno sexo   monto_promedio   ipc
##   <chr>              <dbl> <chr>           <dbl> <dbl>
## 1 CIENCIAS AGRICOLAS  1982 HOMBRE           408.  2.79
## 2 CIENCIAS AGRICOLAS  1982 MUJER            549   2.79
## 3 CIENCIAS AGRICOLAS  1983 HOMBRE           382.  8.98
## 4 CIENCIAS AGRICOLAS  1984 HOMBRE           355.  6.53
## 5 CIENCIAS AGRICOLAS  1984 MUJER            373   6.53
## 6 CIENCIAS AGRICOLAS  1985 HOMBRE           414. 10.1
```

4) Crea la nueva variable `monto_precios2021` multiplicando las variables `monto` e `ipc`. Posteriormente remueve las variables `monto` e `ipc`. El resultado debe verse así:

```
## # A tibble: 608 x 4
## # Groups:   area, anno [277]
##    area                anno sexo   monto_precios2021
##    <chr>              <dbl> <chr>              <dbl>
##  1 CIENCIAS AGRICOLAS  1982 HOMBRE             1141.
##  2 CIENCIAS AGRICOLAS  1982 MUJER              1534.
##  3 CIENCIAS AGRICOLAS  1983 HOMBRE             3428.
##  4 CIENCIAS AGRICOLAS  1984 HOMBRE             2317.
##  5 CIENCIAS AGRICOLAS  1984 MUJER              2437.
##  6 CIENCIAS AGRICOLAS  1985 HOMBRE             4198.
##  7 CIENCIAS AGRICOLAS  1986 HOMBRE             7220.
##  8 CIENCIAS AGRICOLAS  1986 MUJER              3161.
##  9 CIENCIAS AGRICOLAS  1987 HOMBRE            15665.
## 10 CIENCIAS AGRICOLAS  1987 MUJER              9688.
## # ... with 598 more rows
```

5) Usando el comando `pivot_wider()` transforma los datos de la siguiente manera.

```
## # A tibble: 277 x 5
## # Groups:   area, anno [277]
##    area                anno   HOMBRE   MUJER `SIN INFORMACION`
##    <chr>              <dbl>    <dbl>   <dbl>              <dbl>
##  1 CIENCIAS AGRICOLAS  1982    1141.   1534.                 NA
##  2 CIENCIAS AGRICOLAS  1983    3428.     NA                  NA
##  3 CIENCIAS AGRICOLAS  1984    2317.   2437.                 NA
##  4 CIENCIAS AGRICOLAS  1985    4198.     NA                  NA
##  5 CIENCIAS AGRICOLAS  1986    7220.   3161.                 NA
##  6 CIENCIAS AGRICOLAS  1987   15665.   9688.                 NA
##  7 CIENCIAS AGRICOLAS  1988   41485.     NA                  NA
##  8 CIENCIAS AGRICOLAS  1989   44183.     NA                  NA
##  9 CIENCIAS AGRICOLAS  1990   77329.  84307.                 NA
## 10 CIENCIAS AGRICOLAS  1991 1118740. 462418.                 NA
## # ... with 267 more rows
```

6) Usa la función `replace_na()` para reemplazar los valores perdidos en las variables `HOMBRE` y `MUJER` por ceros. El resultado debe verse así:

```
## # A tibble: 277 x 5
## # Groups:   area, anno [277]
##    area                anno   HOMBRE   MUJER `SIN INFORMACION`
##    <chr>              <dbl>    <dbl>   <dbl>              <dbl>
##  1 CIENCIAS AGRICOLAS  1982    1141.   1534.                 NA
##  2 CIENCIAS AGRICOLAS  1983    3428.      0                  NA
##  3 CIENCIAS AGRICOLAS  1984    2317.   2437.                 NA
##  4 CIENCIAS AGRICOLAS  1985    4198.      0                  NA
##  5 CIENCIAS AGRICOLAS  1986    7220.   3161.                 NA
##  6 CIENCIAS AGRICOLAS  1987   15665.   9688.                 NA
##  7 CIENCIAS AGRICOLAS  1988   41485.      0                  NA
##  8 CIENCIAS AGRICOLAS  1989   44183.      0                  NA
##  9 CIENCIAS AGRICOLAS  1990   77329.  84307.                 NA
## 10 CIENCIAS AGRICOLAS  1991 1118740. 462418.                 NA
## # ... with 267 more rows
```

7) Crea una nueva variable llamada `dif_hombremujer` que mida la diferencia entre el monto asignado a hombres y mujeres `= HOMBRE - MUJER`. Posteriormente conserva sólo las variables `anno`, `area` y `dif_hombremujer`. El resultado debe verse así:

```
## # A tibble: 277 x 3
## # Groups:   area, anno [277]
##     anno area               dif_hombremujer
##    <dbl> <chr>                        <dbl>
##  1  1982 CIENCIAS AGRICOLAS           -394.
##  2  1983 CIENCIAS AGRICOLAS           3428.
##  3  1984 CIENCIAS AGRICOLAS           -120.
##  4  1985 CIENCIAS AGRICOLAS           4198.
##  5  1986 CIENCIAS AGRICOLAS           4060.
##  6  1987 CIENCIAS AGRICOLAS           5978.
##  7  1988 CIENCIAS AGRICOLAS          41485.
##  8  1989 CIENCIAS AGRICOLAS          44183.
##  9  1990 CIENCIAS AGRICOLAS          -6978.
## 10  1991 CIENCIAS AGRICOLAS         656322.
## # ... with 267 more rows
```

8) Usando el comando `pivot_wider()` modifica la tabla producida en (7) y produce la siguiente tabla:

```
## # A tibble: 39 x 10
## # Groups:   anno [39]
##     anno `CIENCIAS AGRICOLAS` `CIENCIAS MEDIC~ `CIENCIAS NATUR~ `CIENCIAS SOCIA~
##    <dbl>                <dbl>            <dbl>            <dbl>            <dbl>
##  1  1982                -394.            -147.             76.6             570.
##  2  1983                3428.            3425.           -344.             -600.
##  3  1984                -120.             305.             63.6             -99.6
##  4  1985                4198.             189.            389.             -245.
##  5  1986                4060.            1025.           4238.              719.
##  6  1987                5978.           16343.           -509.             1888.
##  7  1988               41485.           11794.          -6313.             1096.
##  8  1989               44183.           14803.          -1604.            -2292.
##  9  1990               -6978.           30467.           8112.           -10033.
## 10  1991              656322.           18343.         117489.             4355.
## # ... with 29 more rows, and 5 more variables: HUMANIDADES <dbl>,
## #   INGENIERIA Y TECNOLOGIA <dbl>, MULTIDISCIPLINARIO <dbl>, NO APLICA <dbl>,
## #   SIN INFORMACION <dbl>
```

9) Elige el valor correspondiente a una celda cualquiera y describe la información que comunica.