Práctica Dirigida 6

Mg. Sc. J. Eduardo Gamboa U.

Pregunta 1

Los datos corresponden a la matrícula estudiantil de una universidad peruana correspondientes al año 2023. La fuente de los datos es Kaggle.

El archivo **PD6 - peru_student_enrollment_data_2023.csv** contiene los datos correspondientes a las siguientes variables:

- ENROLLMENT: Tipo de matrícula del estudiante:
  - Nuevo: Estudiante que se matricula por primera vez.
  - Reincorporado: Estudiante que continúa sus estudios sin interrupción.
  - Reinscrito: Estudiante que regresa después de un periodo de inactividad.
- TUITION PAYMENT MARCH 2022: Indica si el estudiante pagó la matrícula en marzo de 2022 (0 = No, 1 = Sí).
- TUITION PAYMENT MARCH 2023: Indica si el estudiante pagó la matrícula en marzo de 2023 (0 = No, 1 = Sí).
- GENDER: Género del estudiante (M, F, U, 1 $\rightarrow$ (M), 2 $\rightarrow$ (F)).
- TYPE.OF.EDUCATIONAL.INSTITUTION: Tipo de institución de la que proviene el estudiante (Colegio, Instituto, etc.).
- EDUCATIONAL.INSTITUTION: Institución educativa de procedencia
- INSTITUTION.STATUS: Condición de la institución (Pública o Privada).
- DEPARTMENT: Departamento donde reside o estudia el estudiante.
- PROVINCE: Provincia donde reside o estudia el estudiante.
- DISTRICT: Distrito donde reside o estudia el estudiante.
- CLASSIFICATION: Tipo de carrera que estudia el alumno.

- CAMPUS: Sede en la que estudia el alumno.
- FACULTY: Facultad en la que estudia el alumno.
- PROGRAM.MAJOR: Programa académico o especialidad en la que está matriculado el estudiante.
- SHIFT.SCHEDULE: Horario de estudios (Mañana, Tarde, Noche, Mixto).
- BENEFIT.DISCOUNTS: Indica si el estudiante recibe beneficios económicos o descuentos.
- STUDY.MODE: Modalidad de estudio:
  - On-site: Classes held at a physical campus
  - Online: Fully online classes.
  - Remote: Online classes with some in-person activities.
  - To be determined: Study mode not yet selected.
- AGE.RANGE.OF.ENROLLED.STUDENT: Rango de edad de los estudiantes matriculados.
- DISABILITY: Indica si la persona presenta alguna discapacidad o no.
- NUMBER.OF.ENROLLED.COURSES: Número de cursos en los que el estudiante está matriculado.
- AT.RISK.COURSE: Indica la cantidad de cursos en riesgo de desaprobación.

1. Leer el archivo de datos.

```
datos = read.csv2('PD6 - peru_student_enrollment_data_2023.csv', na.strings = "")
```

2. Ejecutar las siguientes tareas de preprocesamiento:

a. Modificar el nombre de las columnas, asignando nombres cortos en español.

```
library(dplyr)
datos |> rename(TIPO_MAT  = 1,
                PAGO22    = 2,
                PAGO23    = 3,
                GENERO    = 4,
                TIPO_IE   = 5,
                IE        = 6,
                STAT_IE   = 7,
                DEPTO     = 8,
                PROV      = 9,
                DISTRITO  = 10,
                TIPO_CAR  = 11,
                FACULTAD  = 13,
                CARRERA   = 14,
                HORARIO   = 15,
                BENEFIC   = 16,
                MODALIDAD = 17,
                RANG_EDAD = 18,
                DISCAPAC  = 19,
                NCURSOS   = 20,
                RIESGO    = 21) -> datos
```

b. Corregir los valores de la variable GENDER.

```
datos |> count(GENERO)
```

```
  GENERO     n
1      1    67
2      2   102
3      F 16149
4      M 21061
5      U   201
6   <NA>     2
```

```
datos |> mutate(GENERO = case_when(GENERO %in% c(1) ~ "M",
                                   GENERO %in% c(2) ~ "F",
                                   TRUE ~ GENERO)) -> datos
datos |> count(GENERO)
```

```
  GENERO     n
1      F 16251
2      M 21128
3      U   201
4   <NA>     2
```

c. Crear una nueva variable que se llame PAGO2 que tome el valor SI si el estudiante pagó matrícula en 2022 y 2023, y NO en caso contrario.

```
datos |>
  mutate(PAGO2 = ifelse(PAGO22==1 & PAGO23 ==1, "SI","NO")) -> datos
```

2. Luego de Lima, ¿cuáles son los 3 departamentos donde hay más alumnos que residen o estudian?

```
datos |> count(DEPTO) |> arrange(-n) |> head(4) |> pull(DEPTO)
```

```
[1] "LIMA"       "AREQUIPA"   "LAMBAYEQUE" "ICA"
```

3. Filtrar los registros de las estudiantes que llevan al menos 4 cursos. Almacenar en datos1 y mostrar las 2 primeras filas.

```
datos |> filter(NCURSOS>=4) -> datos1
datos1 |> head(2)
```

```
  TIPO_MAT PAGO22 PAGO23 GENERO TIPO_IE   IE STAT_IE DEPTO PROV   DISTRITO
1    Nuevo      1      1      M    <NA> <NA>    <NA>  LIMA LIMA LOS OLIVOS
2    Nuevo      0      0      F    <NA> <NA>    <NA>  LIMA LIMA      RIMAC
           TIPO_CAR          CAMPUS              FACULTAD               CARRERA
1      Carreras PPE UTP Lima Centro Fac. Ing. Ind. Y Mec. INGENIERIA INDUSTRIAL
2 Carreras Pregrado UTP Lima Centro      Fac. Adm. Y Neg.      ADM. Y MARKETING
  HORARIO       BENEFIC  MODALIDAD RANG_EDAD DISCAPAC NCURSOS RIESGO PAGO2
1   NOCHE SIN BENEFICIO Presencial   5. >=30       No       4      0    SI
2   MIXTO SIN BENEFICIO Presencial  2. 19-20       No       5      0    NO
```

4. Filtrar los registros de estudiantes nuevos que no pagaron la matrícula en marzo 2022. Almacenar en datos2 y mostrar las 6 últimas filas.

```
datos |> filter(TIPO_MAT == "Nuevo" & PAGO22 == 0) -> datos2
datos2 |> tail(6)
```

```
    TIPO_MAT PAGO22 PAGO23 GENERO   TIPO_IE                          IE
849    Nuevo      0      0      F   COLEGIO                ALAS PERUANAS
850    Nuevo      0      0      M      <NA>                       ESSUMIN
851    Nuevo      0      0      M      <NA> JUAN BAUTISTA SCARSI VALDIVIA
852    Nuevo      0      0      M      <NA>            LA CATOLICA - LIMA
853    Nuevo      0      0      F INSTITUTO                DEL ALTIPLANO
854    Nuevo      0      0      F INSTITUTO          SAN MARTIN DE PORRES
    STAT_IE    DEPTO           PROV          DISTRITO                  TIPO_CAR
849 PRIVADA     PUNO      SAN ROMAN           JULIACA Carreras Pregrado Virtual
850    <NA> AREQUIPA       AREQUIPA          AREQUIPA Carreras Pregrado Virtual
851    <NA> MOQUEGUA MARISCAL NIETO           SAMEGUA Carreras Pregrado Virtual
852    <NA>     LIMA           LIMA VILLA EL SALVADOR Carreras Pregrado Virtual
853 PRIVADA     PUNO           PUNO              PUNO Carreras Pregrado Virtual
854 PRIVADA     LIMA         CANETE    NUEVO IMPERIAL Carreras Pregrado Virtual
        CAMPUS                      FACULTAD
849 UTP Virtual              Fac. Hum y CC Soc
850 UTP Virtual          Fac. Ing. Ind. Y Mec.
851 UTP Virtual       Fac. Ing. Sist. Y Elect.
852 UTP Virtual          Fac. Ing. Ind. Y Mec.
853 UTP Virtual          Fac. Ing. Ind. Y Mec.
854 UTP Virtual Fac. Der. Cienc. Polit. Y RRII
                                    CARRERA HORARIO       BENEFIC MODALIDAD
849                     PSICOLOGIA (VIRTUAL)   NOCHE SIN BENEFICIO   Virtual
850                ING. INDUSTRIAL (VIRTUAL)   MIXTO SIN BENEFICIO   Virtual
851 ING. DE SISTEMAS E INFORMÁTICA (VIRTUAL)   MIXTO SIN BENEFICIO   Virtual
852                ING. INDUSTRIAL (VIRTUAL)   MIXTO SIN BENEFICIO   Virtual
853                ING. INDUSTRIAL (VIRTUAL)   MIXTO SIN BENEFICIO   Virtual
854                        DERECHO (VIRTUAL)   NOCHE SIN BENEFICIO   Virtual
    RANG_EDAD DISCAPAC NCURSOS RIESGO PAGO2
849  4. 24-29       No       3      0    NO
850   5. >=30       No       3      0    NO
851  4. 24-29       No       3      0    NO
852  4. 24-29       No       3      0    NO
853  4. 24-29       No       3      3    NO
854   5. >=30       No       3      0    NO
```

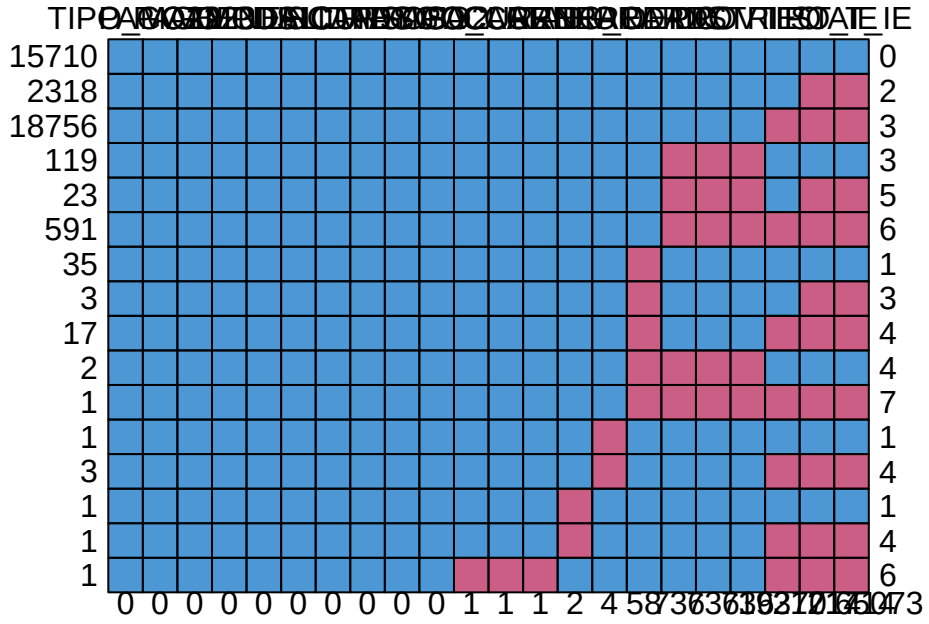
5. Seleccionar a los estudiantes que se encuentran en riesgo o están matriculados en más de 4 cursos. Luego, seleccionar las 3 primeras y las 2 últimas filas.

```
datos |> filter(RIESGO > 0 | NCURSOS > 4) -> datos3
library(psych)
datos |> headTail(3,2)
```

```
        TIPO_MAT PAGO22 PAGO23 GENERO     TIPO_IE                          IE
1          Nuevo      0      0      M   INSTITUTO                       IDAT
2          Nuevo      1      0      M     COLEGIO                 COLEGIO SISE
3          Nuevo      1      1      F        <NA>                        <NA>
...          <NA>    ...    ...   <NA>        <NA>                        <NA>
37581 Reinscrito      1      1      M UNIVERSIDAD PONTIFICIAUNIVERSIDADCATÓLIC
37582 Reinscrito      1      1      F        <NA>                        <NA>
      STAT_IE DEPTO PROV                DISTRITO                TIPO_CAR
1     PRIVADA  LIMA LIMA                   BRENA         Carreras Pregrado
2     PRIVADA  LIMA LIMA VILLA MARIA DEL TRIUNFO         Carreras Pregrado
3        <NA>  LIMA LIMA             JESUS MARIA         Carreras Pregrado
...      <NA>  <NA> <NA>                    <NA>                      <NA>
37581 PRIVADA  LIMA LIMA              CHORRILLOS Carreras Pregrado Virtual
37582    <NA>  LIMA LIMA                   COMAS Carreras Pregrado Virtual
                CAMPUS                    FACULTAD
1     UTP Lima Centro       Fac. Ing. Sist. Y Elect.
2     UTP Lima Centro       Fac. Ing. Sist. Y Elect.
3     UTP Lima Centro Fac. Der. Cienc. Polit. Y RRII
...              <NA>                          <NA>
37581     UTP Virtual          Fac. Ing. Ind. Y Mec.
37582     UTP Virtual Fac. Der. Cienc. Polit. Y RRII
                            CARRERA HORARIO       BENEFIC  MODALIDAD RANG_EDAD
1     ING. DE REDES Y COMUNICACIONES   MIXTO SIN BENEFICIO Presencial   5. >=30
2                  ING. DE SISTEMAS   NOCHE SIN BENEFICIO Presencial  4. 24-29
3                           DERECHO   NOCHE SIN BENEFICIO Presencial  4. 24-29
...                            <NA>    <NA>          <NA>       <NA>      <NA>
37581      ING. INDUSTRIAL (VIRTUAL)   MIXTO SIN BENEFICIO    Virtual   5. >=30
37582              DERECHO (VIRTUAL)   MIXTO SIN BENEFICIO     Remoto   5. >=30
      DISCAPAC NCURSOS RIESGO PAGO2
1           No       0      0    NO
2           No       3      0    NO
3           No       1      0    SI
...       <NA>     ...    ...  <NA>
37581       No       3      0    SI
37582       No       3      0    SI
```

6. Realizar el análisis de valores perdidos usando las funciones `md.pattern` y `gg_miss_upset`.

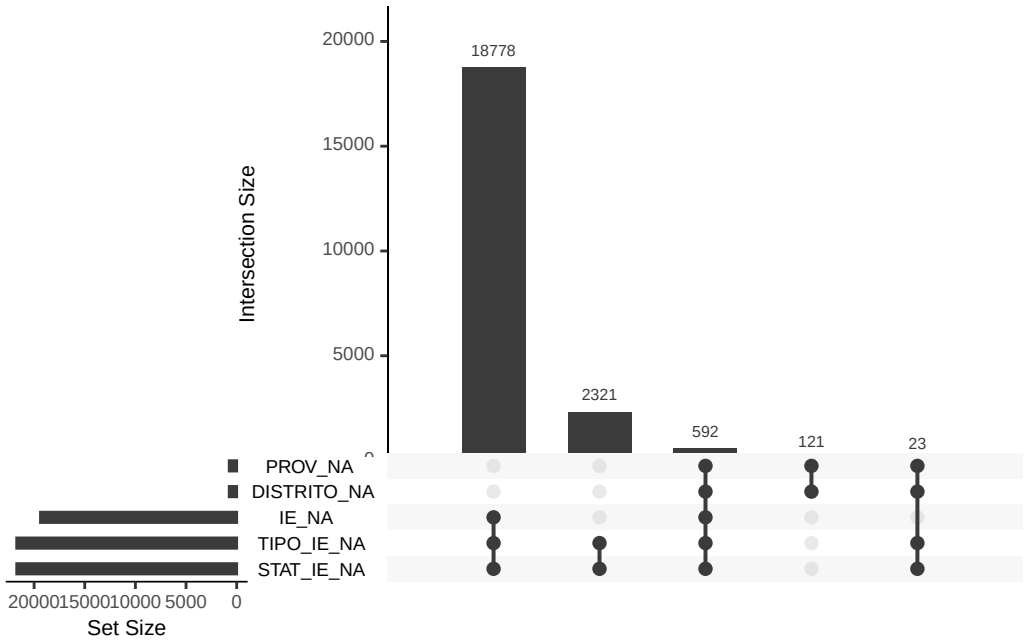
```
library(mice)
datos |> md.pattern()
```



	TIPO_MAT	PAGO22	PAGO23	CAMPUS	BENEFIC	MODALIDAD	DISCAPAC	NCURSOS	RIESGO	
15710	1	1	1	1	1	1	1	1	1	0
2318	1	1	1	1	1	1	1	1	1	2
18756	1	1	1	1	1	1	1	1	1	3
119	1	1	1	1	1	1	1	1	1	3
23	1	1	1	1	1	1	1	1	1	5
591	1	1	1	1	1	1	1	1	1	6
35	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	3
17	1	1	1	1	1	1	1	1	1	4
2	1	1	1	1	1	1	1	1	1	4
1	1	1	1	1	1	1	1	1	1	7
1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	4
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	4
1	1	1	1	1	1	1	1	1	1	6
	0	0	0	0	0	0	0	0	0	

```
1            1      1      1      1       1         1        1       1      1
             0      0      0      0       0         0        0       0      0
      PAGO2 TIPO_CAR FACULTAD CARRERA GENERO RANG_EDAD HORARIO DEPTO PROV
15710     1        1        1       1      1         1       1     1    1
2318      1        1        1       1      1         1       1     1    1
18756     1        1        1       1      1         1       1     1    1
119       1        1        1       1      1         1       1     0    0
23        1        1        1       1      1         1       1     0    0
591       1        1        1       1      1         1       1     0    0
35        1        1        1       1      1         1       0     1    1
3         1        1        1       1      1         1       0     1    1
17        1        1        1       1      1         1       0     1    1
2         1        1        1       1      1         1       0     0    0
1         1        1        1       1      1         1       0     0    0
1         1        1        1       1      1         0       1     1    1
3         1        1        1       1      1         0       1     1    1
1         1        1        1       1      0         1       1     1    1
1         1        1        1       1      0         1       1     1    1
1         1        0        0       0      1         1       1     1    1
          0        1        1       1      2         4      58   736  736
      DISTRITO    IE TIPO_IE STAT_IE
15710        1     1       1       1     0
2318         1     1       0       0     2
18756        1     0       0       0     3
119          0     1       1       1     3
23           0     1       0       0     5
591          0     0       0       0     6
35           1     1       1       1     1
3            1     1       0       0     3
17           1     0       0       0     4
2            0     1       1       1     4
1            0     0       0       0     7
1            1     1       1       1     1
3            1     0       0       0     4
1            1     1       1       1     1
1            1     0       0       0     4
1            1     0       0       0     6
           736 19370   21714   21714 65073
```

```
library(naniar)
datos |> gg_miss_upset()
```



Pregunta 2

Un proceso de selección para el puesto de Analista de Logística consta de tres etapas: evaluación del currículum vitae, entrevista con el área de Recursos Humanos y entrevista con el gerente del área de Logística. Las calificaciones de los postulantes se encuentran en el archivo **PD6 - seleccion.xlsx**. La primera hoja corresponde a la evaluación del CV, la segunda a la entrevista con Recursos Humanos, y la tercera a la entrevista con el gerente de Logística. Cabe señalar que la cantidad de postulantes disminuye en cada etapa.

1. Identificar al postulante que obtuvo el puntaje total más alto, es decir, la mayor suma de las calificaciones en las tres etapas. ¿Quién cumple con este criterio?
2. A continuación, se debe generar un data frame que contenga únicamente tres columnas:

- Nombre completo del o de la postulante (NOM)
- Tipo de evaluación (EVAL)
- Calificación obtenida (CALIF)

3. Guardar este objeto en un archivo llamado CONSOLIDADO.xlsx, utilizando el paquete writexl.