```
%Fecha de actualización: 21/septiembre/2023
clc
clear all
```

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

PROFESOR: Luis Antonio Aguilar Pérez

SEMESTRE: 2024-1

GRUPO: 02

TAREA: Ejercicios DataStore en Matlab

ALUMNA: Margarita Méndez Aguilar

FECHA ENTREGA: 27 Septiembre 2023

Manipulación de archivos de datos

Un modelo de datos o DataFrame es una estructura de datos tabular, la cual organiza la información en filas y columnas de manera similar a una hoja de cálculo. En particular cada columna del DataFrame representa una variable o feature (atributo), mientras que cada fila representa una observación o registro realizado. Los DataFrames son utilizados comúnmente en análisis de datos para manipular y analizar grandes conjuntos de datos de forma eficiente. Existen diversos métodos de almacenar esta informacion dependiendo del lenguaje de programacion utilizado. El formato universal y más tradicional de almacenamiento de la informacion es mediante archivos de tipo CSV. Un archivo de formato CSV es en realidad un archivo en formato de

codificacion de texto plano, donde cada línea del archivo representa una fila de datos, y los valores de cada columna están separados por un carácter delimitador, siendo generalmente este una coma. Es un formato popular debido a su simplicidad y fácil manipulación por programas y hojas de cálculo. Además, muchos sistemas y aplicaciones pueden exportar datos en formato CSV, lo que lo hace fácilmente intercambiable entre diferentes plataformas y herramientas. En particular un archivo de tipo CSV se visualiza de la siguiente manera:

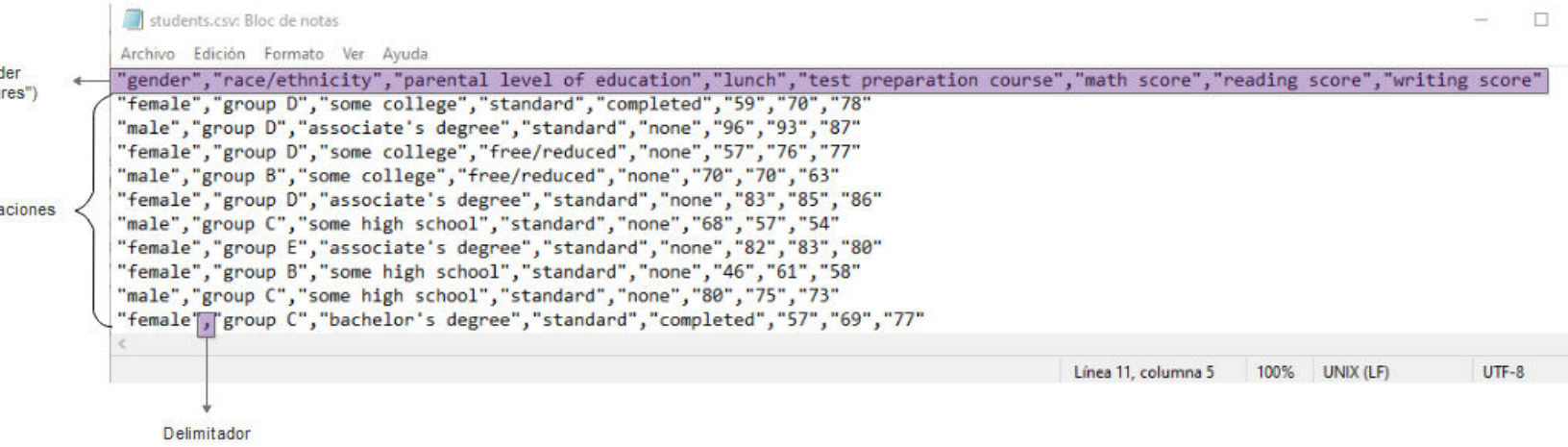

Figura 1.- Archivo de tipo CSV visualizado como texto plano

Existen otros formatos de archivos como el parquet, el cual es un formato de archivo de almacenamiento de datos de código abierto utilizado para almacenar datos tabulares en una estructura de columna, en lugar de una estructura de fila tradicional. Este formato está diseñado para ser eficiente en términos de almacenamiento y procesamiento, y permite una lectura y escritura más rápida de grandes conjuntos de datos. En particular, al estar los datos encriptados y codificados en este tipo de archivos, no es posible visualizarlos de manera tradicional, aunque el esquema de organizacion de la información seria lo más cercano a este:

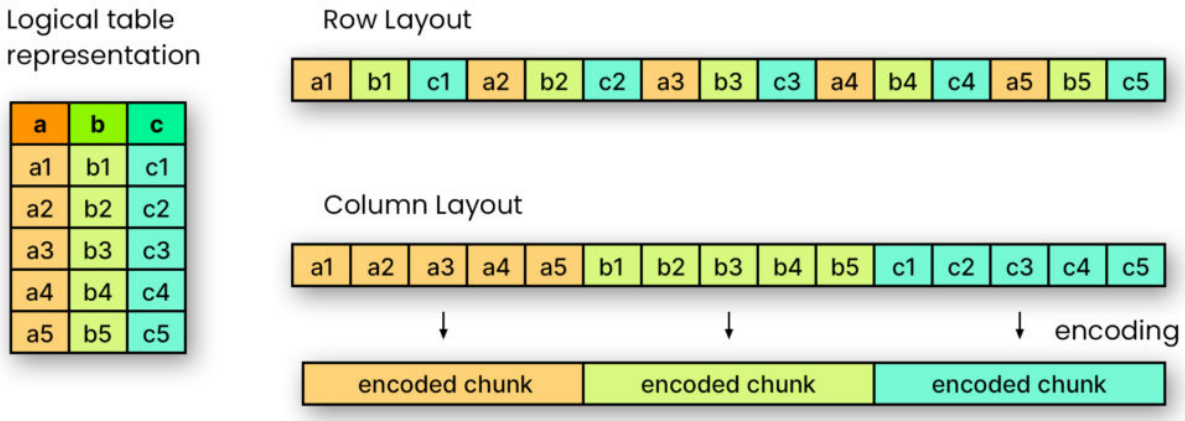

Figura 2.- Esquema tradicional de codificacion tipo parquet

Finalmente, el uso y manejo de modelos de datos en MATLAB se realiza mediante la funcion "Datastore". Esta es una función que proporciona una interfaz rápida para acceder a grandes conjuntos de datos, como archivos de imágenes, archivos de audio o archivos de texto, sin cargarlos en la memoria. Permite la lectura de datos de manera eficiente y escalable, ya que lee y procesa los datos de forma incremental a medida que se necesitan, lo que permite trabajar con grandes conjuntos de datos sin tener que cargarlos por completo en la memoria. Además, la función "datastore" permite realizar operaciones de preprocesamiento y manipulación de datos, como filtrado y transformación de datos, de manera eficiente y fácil. Estas bases de datos tienen la siguiente estructura de informacion

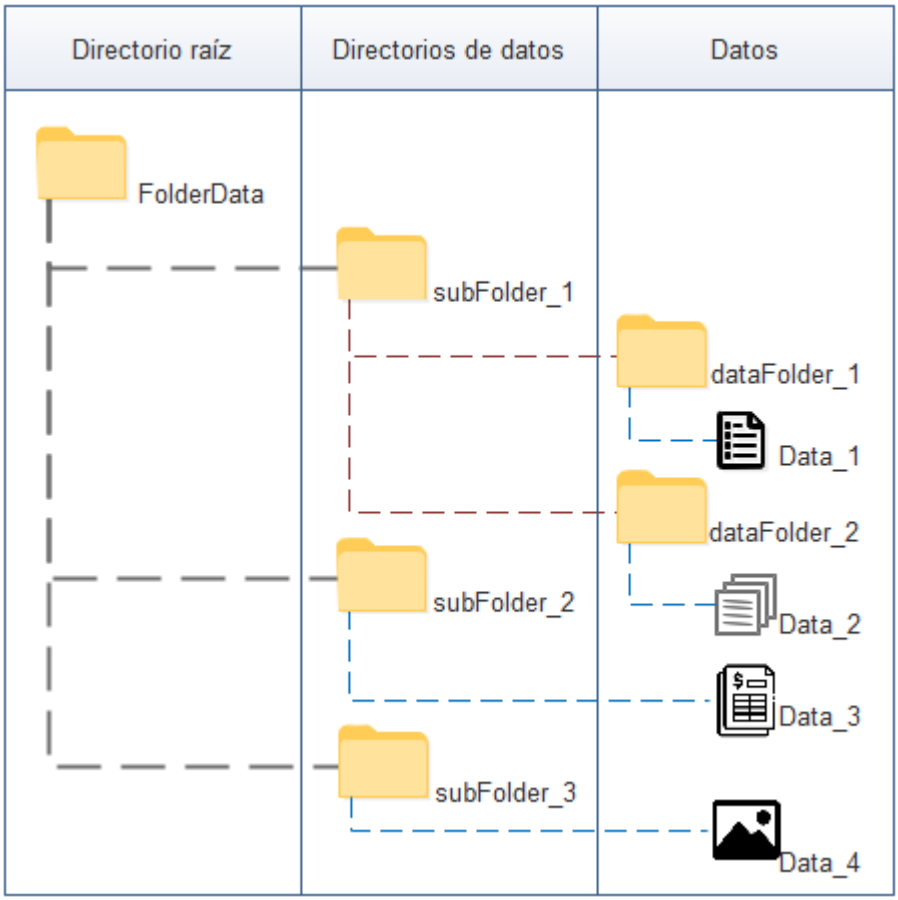

Figura 3.- Esquema de un modelo de datos

--- Pasos iniciales

```
%%%
%Creacion de directorio de trabajo
%rootFolder = 'D:\TSISB_IA';
rootFolder ='C:\Users\Maggie\Documents\MATLAB\TSISB_IA';
workingFolder = 'tareaejercicio_1';
tempFolder = 'temp';
savePath = fullfile(rootFolder,workingFolder);
saveTempPath = fullfile(rootFolder,workingFolder,tempFolder);

prefix = ['\' 'students'];
sufix = '.csv';
newName = [prefix, sufix];
```

```
%Despues de correr esta celda una vez, crea un bloque de comentarios a partir de
esta linea

if ~exist(savePath,'dir')
    [status, message, ~] = mkdir(savePath);
    if status == 0
        disp(message)
    end
end

if ~exist(saveTempPath,'dir')
    [status, message, ~] = mkdir(saveTempPath);
    if status == 0
        disp(message)
    end
end

%%%
%Organizacion y copia de archivos
[fileName, pathFileName] = uigetfile('C:\','*.txt' );

if isequal(fileName,0)
    disp('Se canceló la busqueda de archivos');
else
    disp(['El usuario seleccionó el archivo ', fullfile(pathFileName,fileName)]);
    [status, message, ~] = copyfile([pathFileName,fileName],[saveTempPath,newName]);
    if status == 0
        disp(message)
    else
        disp(['El cual se movio a la dirección ', fullfile(saveTempPath)]);
    end
end
```

```
El usuario seleccionó el archivo C:\Users\Maggie\Downloads\students.csv
El cual se movio a la dirección C:\Users\Maggie\Documents\MATLAB\TSISB_IA\tareaejercicio_1\temp
```

```
%}
```

°°°Modelo de datos mediante un solo archivo

Para lograr el manejo de grandes conjuntos de información, es necesario establecer un "pipeline" o flujo de trabajo de la función datastore. Este se muestra a continuación:

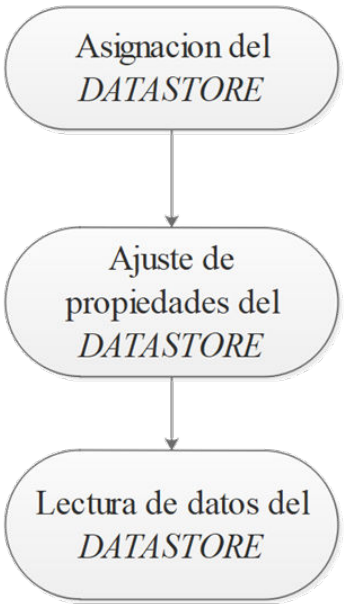

Figura 2.- Flujo de trabajo de un Datastore

La primer parte del pipeline del datastore incluye operaciones como:

- Detección de propiedades de los archivos del dataframe/modelo de datos
- Normalizacion de datos

```
%****
%En este ejercicio solo estaremos trabajando con un archivo
%contenido dentro del datastore
%****

%Se determina la direccion de los archivos que integran el "Datastore"
archivoCSV = [saveTempPath, newName];

%Visualización rápida de las primeras lineas del archivo de texto
%Como solo es un archivo se puede utilizar el comando dbtype
dbtype(archivoCSV,'10:15')
```

```
10    "male","group C","some high school","standard","none","80","75","73"
11    "female","group C","bachelor's degree","standard","completed","57","69","77"
12    "male","group B","some high school","standard","none","74","69","69"
13    "male","group B","master's degree","standard","none","53","50","49"
14    "male","group B","bachelor's degree","free/reduced","none","76","74","76"
15    "male","group A","some college","standard","none","70","73","70"
```

```
%Asignamos nuestro datastore dentro de MATLAB
dsGeneral = datastore(archivoCSV)
```

```
Warning: Table variable names that were not valid MATLAB identifiers have been modified. Since table
variable names must be unique, any table variable names that happened to match the new identifiers also
have been modified.
dsGeneral =
```

```
  TabularTextDatastore with properties:

                     Files: {
                            ' ...\Maggie\Documents\MATLAB\TSISB_IA\tareaejercicio_1\temp\students.csv'
                            }
                   Folders: {
                            'C:\Users\Maggie\Documents\MATLAB\TSISB_IA\tareaejercicio_1\temp'
                            }
              FileEncoding: 'UTF-8'
   AlternateFileSystemRoots: {}
         VariableNamingRule: 'modify'
          ReadVariableNames: true
              VariableNames: {'gender', 'race_ethnicity', 'parentalLevelOfEducation' ... and 5 more}
             DatetimeLocale: en_US

  Text Format Properties:
             NumHeaderLines: 0
                  Delimiter: ','
               RowDelimiter: '\r\n'
             TreatAsMissing: ''
               MissingValue: NaN

  Advanced Text Format Properties:
            TextscanFormats: {'%q', '%q', '%q' ... and 5 more}
                   TextType: 'char'
          ExponentCharacters: 'eEdD'
                CommentStyle: ''
                  Whitespace: ' \b\t'
    MultipleDelimitersAsOne: false

  Properties that control the table returned by preview, read, readall:
      SelectedVariableNames: {'gender', 'race_ethnicity', 'parentalLevelOfEducation' ... and 5 more}
            SelectedFormats: {'%q', '%q', '%q' ... and 5 more}
                   ReadSize: 20000 rows
                 OutputType: 'table'
                   RowTimes: []

  Write-specific Properties:
     SupportedOutputFormats: ["txt"    "csv"    "xlsx"    "xls"    "parquet"    "parq"]
        DefaultOutputFormat: "txt"
```

```
%Si tuvieramos muchos archivos, se utilizaria el siguiente comando
preview(dsGeneral)
```

```
ans = 8×8 table
```

...

	gender	race_ethnicity	parentalLevelOfEducation	lunch
1	'female'	'group D'	'some college'	'standard'
2	'male'	'group D'	'associate's degree'	'standard'
3	'female'	'group D'	'some college'	'free/reduced'
4	'male'	'group B'	'some college'	'free/reduced'
5	'female'	'group D'	'associate's degree'	'standard'
6	'male'	'group C'	'some high school'	'standard'
7	'female'	'group E'	'associate's degree'	'standard'
8	'female'	'group B'	'some high school'	'standard'

```
%Desde la version 2019, los nombres de las variables pueden incluir
%cualquier tipo de simbolos, ademas de no necesariamente comenzar solo
%con letras, por lo que matlab requiere el Flag "preserve" para considerar
%esta opción
dsGeneral.VariableNamingRule='preserve'
```

```
dsGeneral =
  TabularTextDatastore with properties:

                     Files: {
                            ' ...\Maggie\Documents\MATLAB\TSISB_IA\tareaejercicio_1\temp\students.csv'
                            }
                   Folders: {
                            'C:\Users\Maggie\Documents\MATLAB\TSISB_IA\tareaejercicio_1\temp'
                            }
              FileEncoding: 'UTF-8'
   AlternateFileSystemRoots: {}
         VariableNamingRule: 'preserve'
          ReadVariableNames: true
              VariableNames: {'gender', 'race_ethnicity', 'parentalLevelOfEducation' ... and 5 more}
             DatetimeLocale: en_US

  Text Format Properties:
            NumHeaderLines: 0
                 Delimiter: ','
               RowDelimiter: '\r\n'
            TreatAsMissing: ''
              MissingValue: NaN

  Advanced Text Format Properties:
            TextscanFormats: {'%q', '%q', '%q' ... and 5 more}
                   TextType: 'char'
          ExponentCharacters: 'eEdD'
                CommentStyle: ''
                  Whitespace: ' \b\t'
    MultipleDelimitersAsOne: false

  Properties that control the table returned by preview, read, readall:
      SelectedVariableNames: {'gender', 'race_ethnicity', 'parentalLevelOfEducation' ... and 5 more}
            SelectedFormats: {'%q', '%q', '%q' ... and 5 more}
                   ReadSize: 20000 rows
                 OutputType: 'table'
                   RowTimes: []

  Write-specific Properties:
     SupportedOutputFormats: ["txt"    "csv"    "xlsx"    "xls"    "parquet"    "parq"]
        DefaultOutputFormat: "txt"
```

```
%Selección del delimitador de texto
dsGeneral.Delimiter=",";

%Modificacion de los nombres utilizados para cada caracteristica/feature
values=dsGeneral.VariableNames;
%Modificacion como si fuera un indexado de valores clásico
newValues=["hola", "esta", "es", "una", "prueba", "de", "cambio", "de etiquetas de
columna"];
dsGeneral.VariableNames = newValues;
```

```
preview(dsGeneral)
```

```
ans = 8×8 table
```

...

	hola	esta	es	una	prueba	de	cambio
1	'female'	'group D'	'some college'	'standard'	'completed'	'59'	'70'
2	'male'	'group D'	'associate's degree'	'standard'	'none'	'96'	'93'
3	'female'	'group D'	'some college'	'free/reduced'	'none'	'57'	'76'
4	'male'	'group B'	'some college'	'free/reduced'	'none'	'70'	'70'
5	'female'	'group D'	'associate's degree'	'standard'	'none'	'83'	'85'
6	'male'	'group C'	'some high school'	'standard'	'none'	'68'	'57'
7	'female'	'group E'	'associate's degree'	'standard'	'none'	'82'	'83'
8	'female'	'group B'	'some high school'	'standard'	'none'	'46'	'61'

°°°Detección y modificación de propiedades de los datos contenidos en el archivo

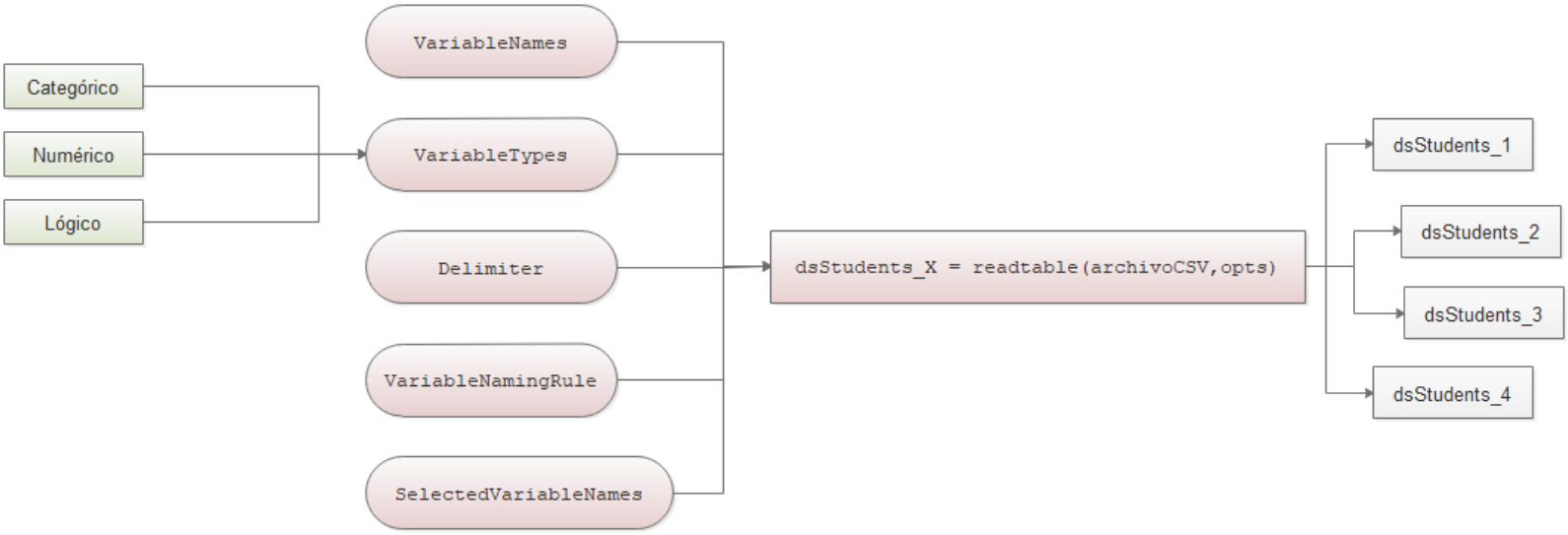

Figura 3.- Propiedades de un **datastore de tipo categorico-numérico**

```
%Visualizacion general de las opciones de importacion de los archivos
%dentro del datastore
opts = detectImportOptions(archivoCSV)
```

```
opts =
  DelimitedTextImportOptions with properties:

    Format Properties:
                    Delimiter: {','}
                   Whitespace: '\b\t '
                   LineEnding: {'\n'  '\r'  '\r\n'}
                CommentStyle: {}
     ConsecutiveDelimitersRule: 'split'
         LeadingDelimitersRule: 'keep'
        TrailingDelimitersRule: 'ignore'
```

```
              EmptyLineRule: 'skip'
                   Encoding: 'UTF-8'

  Replacement Properties:
                MissingRule: 'fill'
             ImportErrorRule: 'fill'
           ExtraColumnsRule: 'addvars'

  Variable Import Properties: Set types by name using setvartype
               VariableNames: {'gender', 'race_ethnicity', 'parentalLevelOfEducation' ... and 5 more}
               VariableTypes: {'char', 'char', 'char' ... and 5 more}
      SelectedVariableNames: {'gender', 'race_ethnicity', 'parentalLevelOfEducation' ... and 5 more}
            VariableOptions: Show all 8 VariableOptions
   Access VariableOptions sub-properties using setvaropts/getvaropts
         VariableNamingRule: 'modify'

  Location Properties:
                  DataLines: [2 Inf]
           VariableNamesLine: 1
              RowNamesColumn: 0
           VariableUnitsLine: 0
    VariableDescriptionsLine: 0
   To display a preview of the table, use preview
```

```
%Visualizacion de los nombres y tipo de variables de las columnas
disp([opts.VariableNames' opts.VariableTypes'])
```

```
    {'gender'                  }    {'char'  }
    {'race_ethnicity'          }    {'char'  }
    {'parentalLevelOfEducation'}    {'char'  }
    {'lunch'                   }    {'char'  }
    {'testPreparationCourse'   }    {'char'  }
    {'mathScore'               }    {'double'}
    {'readingScore'            }    {'double'}
    {'writingScore'            }    {'double'}
```

```
%Modificacion del tipo de variables
opts = setvaropts(opts,
{'race_ethnicity','gender','parentalLevelOfEducation','lunch','testPreparationCourse
'},'Type','categorical');
disp([opts.VariableNames' opts.VariableTypes'])
```

```
    {'gender'                  }    {'categorical'}
    {'race_ethnicity'          }    {'categorical'}
    {'parentalLevelOfEducation'}    {'categorical'}
    {'lunch'                   }    {'categorical'}
    {'testPreparationCourse'   }    {'categorical'}
    {'mathScore'               }    {'double'     }
    {'readingScore'            }    {'double'     }
    {'writingScore'            }    {'double'     }
```

```
%Desde la version 2019, los nombres de las variables pueden incluir
%cualquier tipo de simbolos, ademas de no necesariamente comenzar solo
%con letras, por lo que matlab requiere el Flag "preserve" para considerar
%esta opción
opts.VariableNamingRule='preserve';
```

```
%Selección del delimitador de texto
opts.Delimiter =',';

%Modificacion de los nombres utilizados para cada caracteristica/feature
%Asignacion a una variable particular
values=opts.VariableNames;
%Modificacion como si fuera un indexado de valores clásico
newValues={'hola', 'esta', 'es', 'una', 'prueba', 'de', 'cambio', 'de variables'};
opts.VariableNames = newValues;
dsStudents_1 = readtable(archivoCSV,opts);
head(dsStudents_1)
```

hola	esta	es	una	prueba	de	cambio	de variables
female	group D	some college	standard	completed	59	70	78
male	group D	associate's degree	standard	none	96	93	87
female	group D	some college	free/reduced	none	57	76	77
male	group B	some college	free/reduced	none	70	70	63
female	group D	associate's degree	standard	none	83	85	86
male	group C	some high school	standard	none	68	57	54
female	group E	associate's degree	standard	none	82	83	80
female	group B	some high school	standard	none	46	61	58

```
%regresamos los valores originales
opts.VariableNames = values;
dsStudents_2 = readtable(archivoCSV,opts);
head(dsStudents_2)
```

gender	race_ethnicity	parentalLevelOfEducation	lunch	testPreparationCourse	mathScore	re
female	group D	some college	standard	completed	59	
male	group D	associate's degree	standard	none	96	
female	group D	some college	free/reduced	none	57	
male	group B	some college	free/reduced	none	70	
female	group D	associate's degree	standard	none	83	
male	group C	some high school	standard	none	68	
female	group E	associate's degree	standard	none	82	
female	group B	some high school	standard	none	46	

```
%Podemos elegir con que caracteristicas queremos trabajar nuestra base de datos
opts.SelectedVariableNames = {'gender','mathScore','readingScore', 'writingScore'};
dsStudents_3 = readtable(archivoCSV,opts);
head(dsStudents_3)
```

gender	mathScore	readingScore	writingScore
female	59	70	78
male	96	93	87
female	57	76	77
male	70	70	63
female	83	85	86
male	68	57	54

```
female       82            83            80
female       46            61            58
```

```
%y crear distintos tipos de tablas a partir del datastore original
opts.SelectedVariableNames = {'race_ethnicity','mathScore','readingScore',
'writingScore'};
dsStudents_4 = readtable(archivoCSV,opts);
head(dsStudents_4)
```

race_ethnicity	mathScore	readingScore	writingScore
group D	59	70	78
group D	96	93	87
group D	57	76	77
group B	70	70	63
group D	83	85	86
group C	68	57	54
group E	82	83	80
group B	46	61	58

°°°Visualizaciones exploratorias de datos

En este momento se han creado 4 tablas que contienen los siguientes datos

- Tabla 1: No la usaremos
- Tabla 2: Todas las categorias de datos
- Tabla 3: Datos categóricos de genero y datos numéricos de las pruebas de matemáticas, lectura y escritura
- Tabla 4: Datos categóricos de raza etnica y datos numéricos de las pruebas de matemáticas, lectura y escritura

Vamos a realizar un análisis exploratorio de los datos, primero visualicemos un resumen de los datos categoricos contenidos en la tabla 2

Resumen de datos categoricos

```
genero = categorical(dsStudents_2.gender);
grpCatGenero = categories(genero)
```

```
grpCatGenero = 2×1 cell
'female'
'male'
```

```
numCatGenero = countcats(genero)
```

```
numCatGenero = 2×1
   492
   508
```

```
summary(genero)
```

```
female      492
male        508
```

```
etnia = categorical(dsStudents_2.race_ethnicity);
grpCatEtn = categories(etnia)
```

```
grpCatEtn = 5×1 cell
'group A'
'group B'
'group C'
'group D'
'group E'
```

```
numCatEtn = countcats(etnia)
```

```
numCatEtn = 5×1
    79
   198
   323
   257
   143
```

```
%Vamos a reemplazar las categorias origniales por otros nombres
nuevasEtnias = {'Latino',...
                'Afroamericano',...
                'Americano',...
                'Asiatico',...
                'Europeo'};
nuevaEtnia = renamecats(etnia,nuevasEtnias);
summary(etnia)
```

```
    group A       79
    group B      198
    group C      323
    group D      257
    group E      143
```

```
summary(nuevaEtnia)
```

```
    Latino             79
    Afroamericano     198
    Americano         323
    Asiatico          257
    Europeo           143
```

```
```

*** Ejercicio ***

Deberás mostrar cuantas categorias existen en:

- el nivel de educacion de los padres
- si contaron con un desayuno o no
- si hubo una preparacion previa a la prueba

```
%Categorías en "Nivel de educación de los padres"
```

```
niveledu = categorical(dsStudents_2.parentalLevelOfEducation);
grpCatNiv = categories(niveledu)
```

```
grpCatNiv = 6×1 cell
'associate's degree'
'bachelor's degree'
'high school'
'master's degree'
'some college'
'some high school'
```

```
numCatNiv = countcats(niveledu)
```

```
numCatNiv = 6×1
   204
   105
   215
    75
   224
   177
```

```
%Categorías en "Lunch"
desayuno = categorical(dsStudents_2.lunch);
grpCatDes = categories(desayuno)
```

```
grpCatDes = 2×1 cell
'free/reduced'
'standard'
```

```
numCatDes = countcats(desayuno)
```

```
numCatDes = 2×1
   340
   660
```

```
%Categorías en "Preparación previa a la prueba"
preparacion = categorical(dsStudents_2.testPreparationCourse);
grpCatPrep = categories(preparacion)
```

```
grpCatPrep = 2×1 cell
'completed'
'none'
```

```
numCatPrep = countcats(preparacion)
```

```
numCatPrep = 2×1
   344
   656
```

Además deberas de cambiar el nombre de las categorias en la columna "preparacion de la prueba" de la siguiente manera

- completed - > terminado
- none -> no terminado

```
%Reemplazo de los nombres de categorías dentro de "Preparación de la
```

```
%prueba"
nuevasTest = {'terminado',...
              'no terminado',...
             };
nuevaPrep = renamecats(preparacion,nuevasTest);
summary(preparacion)
```

```
completed      344
none           656
```

```
summary(nuevaPrep)
```

```
terminado         344
no terminado      656
```

Resumen de datos numericos

Ahora vamos a realizar un resumen de los datos numéricos, lo cual puede incluir encontrar la media de los datos de pruebas numericas

---Resumen general de una categoria

Los distintos tipos de operaciones numéricas que podemos realizar son las siguientes

Method	Description
"sum"	Sum
"mean"	Mean
"median"	Median
"mode"	Mode
"var"	Variance
"std"	Standard deviation
"min"	Minimum
"max"	Maximum
"range"	Maximum minus minimum
"nummissing"	Number of missing elements
"nnz"	Number of nonzero and non-NaN elements
"all"	All computations previously listed

Figura 4.- Operaciones numéricas de una tabla

```
%Primero veremos el score promedio dependiendo del genero, utilizando la
%tabla 3
genderMean = groupsummary(dsStudents_3,"gender","mean")
```

genderMean = 2×5 table

	gender	GroupCount	mean_mathScore	mean_readingScore	mean_writingScore
1	female	492	64.7744	73.4736	73.4390
2	male	508	70.7500	67.3878	64.9764

```
%Podemos visualizar tambien el promedio dependiendo del grupo etnico
%utilizando la tabla 4
raceMean = groupsummary(dsStudents_4,"race_ethnicity","mean")
```

raceMean = 5×5 table

...

	race_ethnicity	GroupCount	mean_mathScore	mean_readingScore
1	group A	79	65.6962	69.2025
2	group B	198	64.0707	68.5303
3	group C	323	65.5108	68.6099
4	group D	257	68.8794	70.9300
5	group E	143	77.4266	76.6154

```
%Podemos realizar operaciones por conjuntos especificos por ejemplo de la
%siguiente manera
gen = dsStudents_2.gender;
race = dsStudents_2.race_ethnicity;
math = dsStudents_2.mathScore;
reading = dsStudents_2.readingScore;
writing = dsStudents_2.writingScore;
auxTable = table(gen,race,math, reading, writing)
```

auxTable = 1000×5 table

	gen	race	math	reading	writing
1	female	group D	59	70	78
2	male	group D	96	93	87
3	female	group D	57	76	77
4	male	group B	70	70	63
5	female	group D	83	85	86
6	male	group C	68	57	54
7	female	group E	82	83	80
8	female	group B	46	61	58
9	male	group C	80	75	73
10	female	group C	57	69	77
11	male	group B	74	69	69
12	male	group B	53	50	49
13	male	group B	76	74	76
14	male	group A	70	73	70
15	male	group C	55	54	52
16	male	group E	56	46	43
17	female	group C	35	47	41
18	female	group C	87	92	81
19	female	group E	80	82	85

	gen	race	math	reading	writing
20	female	group D	65	71	74
21	male	group C	66	66	62
22	female	group D	67	71	76
23	female	group B	70	71	71
24	male	group E	89	88	86
25	male	group D	99	85	88
26	male	group B	74	83	72
27	male	group D	58	52	51
28	male	group D	70	66	59
29	female	group E	80	79	71
30	male	group D	90	87	86
31	female	group B	80	81	85
32	female	group D	68	76	79
33	female	group B	69	78	75
34	female	group D	32	35	37
35	male	group D	82	82	82
36	female	group A	57	53	54
37	female	group E	69	74	75
38	male	group D	68	66	72
39	male	group C	74	85	87
40	male	group E	89	85	78
41	male	group C	46	46	48
42	male	group C	76	82	77
43	male	group B	86	82	72
44	male	group D	69	73	67
45	female	group B	52	56	54
46	male	group C	63	71	65
47	male	group A	96	82	90
48	male	group C	80	76	68
49	female	group E	59	52	56
50	male	group D	80	77	80
51	female	group E	65	77	74
52	female	group E	74	83	84

	gen	race	math	reading	writing
53	male	group D	90	93	84
54	female	group B	69	72	72
55	male	group C	69	67	63
56	female	group C	62	64	61
57	female	group D	67	75	80
58	female	group E	89	93	93
59	female	group C	79	86	78
60	male	group C	67	66	66
61	male	group D	82	74	75
62	male	group C	63	69	63
63	female	group D	71	83	80
64	female	group C	55	68	73
65	female	group B	61	74	71
66	female	group B	35	34	36
67	male	group C	75	77	66
68	female	group B	73	91	88
69	female	group C	56	62	57
70	male	group D	80	70	73
71	male	group C	83	81	78
72	female	group D	64	82	80
73	female	group C	23	33	33
74	female	group D	41	58	59
75	male	group E	61	49	52
76	male	group B	63	46	46
77	male	group B	84	91	89
78	male	group C	55	61	59
79	male	group A	85	75	74
80	male	group B	65	61	57
81	male	group C	88	80	81
82	male	group D	91	93	95
83	female	group A	51	46	42
84	male	group C	73	77	76
85	female	group D	73	89	89

	gen	race	math	reading	writing
86	male	group D	100	97	91
87	female	group D	48	68	68
88	male	group E	98	79	85
89	male	group B	68	65	60
90	male	group C	64	62	58
91	male	group C	72	67	61
92	female	group C	63	74	75
93	male	group C	43	51	38
94	male	group D	80	75	74
95	female	group C	71	88	83
96	female	group C	91	96	97
97	female	group D	68	84	87
98	female	group B	73	80	78
99	female	group B	75	90	95
100	male	group C	83	62	64

⋮

```
gender_raceMean = groupsummary(auxTable,["gen","race"],"mean")
```

gender_raceMean = 10×6 table

	gen	race	GroupCount	mean_math	mean_reading	mean_writing
1	female	group A	41	62.9512	72.2683	71.7073
2	female	group B	112	62.2679	72.3839	71.8929
3	female	group C	151	63.0199	71.9934	71.4768
4	female	group D	118	63.8390	72.8729	74.1610
5	female	group E	70	75.2143	80.1286	79.9429
6	male	group A	38	68.6579	65.8947	63.6842
7	male	group B	86	66.4186	63.5116	59.9767
8	male	group C	172	67.6977	65.6395	62.7035
9	male	group D	139	73.1583	69.2806	68.4245
10	male	group E	73	79.5479	73.2466	70.3288

*** Ejercicio ***

Realiza las siguientes operaciones

- Cual es la media del grupo genero vs nivel de estudios de los padres

- Cual es el promedio de la raza etnica vs si se preparó o no para la prueba
- Cual es el promedio del genero vs si se preparó o no para la prueba

```
%Media del genero vs. nivel de estudios de los padres
genero = dsStudents_2.gender;
nivel = dsStudents_2.parentalLevelOfEducation;
auxTable = table(genero,nivel)
```

auxTable = 1000×2 table

	genero	nivel
1	female	some college
2	male	associate's degree
3	female	some college
4	male	some college
5	female	associate's degree
6	male	some high school
7	female	associate's degree
8	female	some high school
9	male	some high school
10	female	bachelor's degree
11	male	some high school
12	male	master's degree
13	male	bachelor's degree
14	male	some college
15	male	master's degree
16	male	master's degree
17	female	some college
18	female	high school
19	female	associate's degree
20	female	associate's degree
21	male	high school
22	female	associate's degree
23	female	some college
24	male	associate's degree
25	male	associate's degree
26	male	some college
27	male	high school

	genero	nivel
28	male	some high school
29	female	associate's degree
30	male	associate's degree
31	female	associate's degree
32	female	associate's degree
33	female	high school
34	female	master's degree
35	male	high school
36	female	some high school
37	female	some college
38	male	associate's degree
39	male	associate's degree
40	male	master's degree
41	male	associate's degree
42	male	associate's degree
43	male	high school
44	male	some college
45	female	high school
46	male	bachelor's degree
47	male	associate's degree
48	male	some college
49	female	high school
50	male	some high school
51	female	high school
52	female	master's degree
53	male	some high school
54	female	some college
55	male	high school
56	female	some college
57	female	master's degree
58	female	some high school
59	female	bachelor's degree
60	male	some high school

	genero	nivel
61	male	some high school
62	male	some high school
63	female	some college
64	female	associate's degree
65	female	high school
66	female	associate's degree
67	male	high school
68	female	some high school
69	female	high school
70	male	associate's degree
71	male	some high school
72	female	some college
73	female	some high school
74	female	some high school
75	male	some college
76	male	bachelor's degree
77	male	associate's degree
78	male	some college
79	male	associate's degree
80	male	high school
81	male	associate's degree
82	male	master's degree
83	female	some college
84	male	some college
85	female	some college
86	male	master's degree
87	female	some high school
88	male	some college
89	male	master's degree
90	male	bachelor's degree
91	male	associate's degree
92	female	bachelor's degree
93	male	some college

	genero	nivel
94	male	some college
95	female	some college
96	female	associate's degree
97	female	some college
98	female	associate's degree
99	female	high school
100	male	some college

⋮

```
gender_levelMean = groupsummary(auxTable,["genero","nivel"],"mean")
```

gender_levelMean = 12×3 table

	genero	nivel	GroupCount
1	female	associate's degree	101
2	female	bachelor's degree	52
3	female	high school	116
4	female	master's degree	33
5	female	some college	106
6	female	some high school	84
7	male	associate's degree	103
8	male	bachelor's degree	53
9	male	high school	99
10	male	master's degree	42
11	male	some college	118
12	male	some high school	93

```
%Promedio de la raza etnica vs. si se preparó o no para la prueba
raza = dsStudents_2.race_ethnicity;
prueba = dsStudents_2.testPreparationCourse;
auxTable = table(raza,prueba)
```

auxTable = 1000×2 table

	raza	prueba
1	group D	completed
2	group D	none
3	group D	none
4	group B	none
5	group D	none

	raza	prueba
6	group C	none
7	group E	none
8	group B	none
9	group C	none
10	group C	completed
11	group B	none
12	group B	none
13	group B	none
14	group A	none
15	group C	none
16	group E	none
17	group C	none
18	group C	none
19	group E	none
20	group D	completed
21	group C	none
22	group D	completed
23	group B	none
24	group E	none
25	group D	completed
26	group B	none
27	group D	none
28	group D	none
29	group E	none
30	group D	none
31	group B	completed
32	group D	none
33	group B	completed
34	group D	none
35	group D	completed
36	group A	none
37	group E	none
38	group D	completed

	raza	prueba
39	group C	completed
40	group E	none
41	group C	completed
42	group C	completed
43	group B	none
44	group D	none
45	group B	none
46	group C	none
47	group A	completed
48	group C	completed
49	group E	none
50	group D	completed
51	group E	completed
52	group E	completed
53	group D	none
54	group B	completed
55	group C	none
56	group C	none
57	group D	none
58	group E	completed
59	group C	none
60	group C	none
61	group D	completed
62	group C	completed
63	group D	none
64	group C	none
65	group B	none
66	group B	none
67	group C	none
68	group B	completed
69	group C	none
70	group D	none
71	group C	completed

	raza	prueba
72	group D	completed
73	group C	none
74	group D	completed
75	group E	completed
76	group B	none
77	group B	completed
78	group C	none
79	group A	completed
80	group B	none
81	group C	none
82	group D	completed
83	group A	none
84	group C	completed
85	group D	none
86	group D	none
87	group D	completed
88	group E	none
89	group B	none
90	group C	none
91	group C	completed
92	group C	none
93	group C	none
94	group D	none
95	group C	none
96	group C	completed
97	group D	completed
98	group B	none
99	group B	completed
100	group C	none

⋮

```
race_testMean = groupsummary(auxTable,["raza","prueba"],"mean")
```

```
race_testMean = 10×3 table
```

	raza	prueba	GroupCount
1	group A	completed	31
2	group A	none	48
3	group B	completed	66
4	group B	none	132
5	group C	completed	101
6	group C	none	222
7	group D	completed	97
8	group D	none	160
9	group E	completed	49
10	group E	none	94

```
%Promedio del genero vs. si se preparó o no para la prueba
genero = dsStudents_2.gender;
prueba = dsStudents_2.testPreparationCourse;
auxTable = table(genero,prueba)
```

auxTable = 1000×2 table

	genero	prueba
1	female	completed
2	male	none
3	female	none
4	male	none
5	female	none
6	male	none
7	female	none
8	female	none
9	male	none
10	female	completed
11	male	none
12	male	none
13	male	none
14	male	none
15	male	none
16	male	none
17	female	none

	genero	prueba
18	female	none
19	female	none
20	female	completed
21	male	none
22	female	completed
23	female	none
24	male	none
25	male	completed
26	male	none
27	male	none
28	male	none
29	female	none
30	male	none
31	female	completed
32	female	none
33	female	completed
34	female	none
35	male	completed
36	female	none
37	female	none
38	male	completed
39	male	completed
40	male	none
41	male	completed
42	male	completed
43	male	none
44	male	none
45	female	none
46	male	none
47	male	completed
48	male	completed
49	female	none
50	male	completed

	genero	prueba
51	female	completed
52	female	completed
53	male	none
54	female	completed
55	male	none
56	female	none
57	female	none
58	female	completed
59	female	none
60	male	none
61	male	completed
62	male	completed
63	female	none
64	female	none
65	female	none
66	female	none
67	male	none
68	female	completed
69	female	none
70	male	none
71	male	completed
72	female	completed
73	female	none
74	female	completed
75	male	completed
76	male	none
77	male	completed
78	male	none
79	male	completed
80	male	none
81	male	none
82	male	completed
83	female	none

	genero	prueba
84	male	completed
85	female	none
86	male	none
87	female	completed
88	male	none
89	male	none
90	male	none
91	male	completed
92	female	none
93	male	none
94	male	none
95	female	none
96	female	completed
97	female	completed
98	female	none
99	female	completed
100	male	none

⋮

```
gender_testMean = groupsummary(auxTable,["genero","prueba"],"mean")
```

gender_testMean = 4×3 table

	genero	prueba	GroupCount
1	female	completed	177
2	female	none	315
3	male	completed	167
4	male	none	341

Modificación de las columnas en modelos de datos

Podemos realizar ligeras modificaciones a las columnas (caracteristicas de nuestro modelo de datos) realizando las siguientes operaciones

- Crear tablas a partir de datos específicos
- Agregar una columna de datos
- Mover columnas de datos
- Remover columnas de datos

```
%Creacion de tablas a partir de datos originales
auxTable_1 = table(gen, math, reading);
head(auxTable_1)
```

gen	math	reading
female	59	70
male	96	93
female	57	76
male	70	70
female	83	85
male	68	57
female	82	83
female	46	61

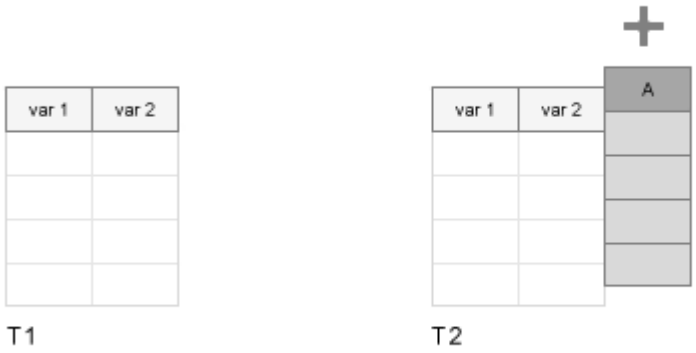

Figura 5.- Agregado de caracteristicas de un modelo de datos

```
%Agregado de valores a tablas ya existentes
auxTable2 = addvars(auxTable_1,writing);
head(auxTable2)
```

gen	math	reading	writing
female	59	70	78
male	96	93	87
female	57	76	77
male	70	70	63
female	83	85	86
male	68	57	54
female	82	83	80
female	46	61	58

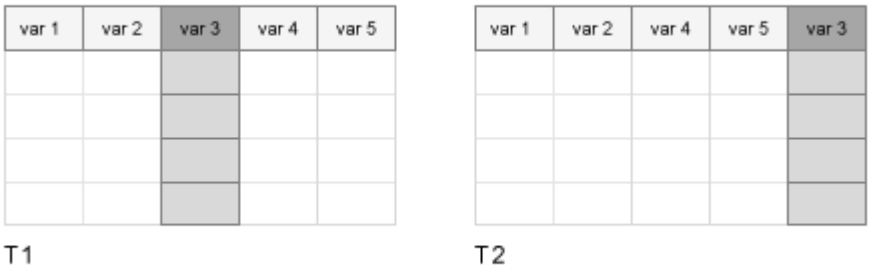

Figura 6.- Reordenamiento de caracteristicas de un modelo de datos

```
%Reordenamiento de caracteristicas de la tabla
auxTable_3 = movevars(auxTable2,'reading','after','writing');
head(auxTable_3)
```

gen	math	writing	reading
female	59	78	70
male	96	87	93
female	57	77	76
male	70	63	70
female	83	86	85
male	68	54	57
female	82	80	83
female	46	58	61

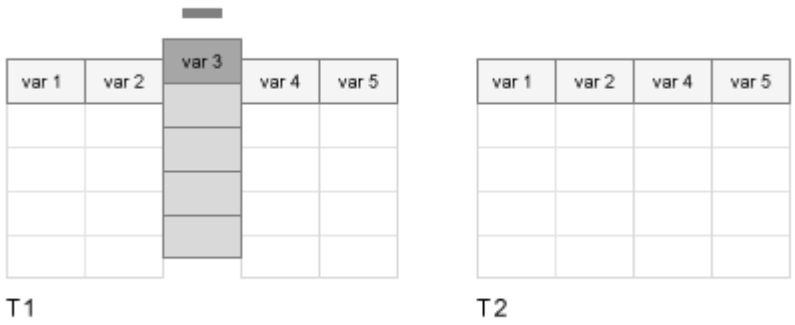

Figura 7.- Eliminacion de caracteristicas de un modelo de datos

```
%Eliminacion de caracteristicas de la tabla
auxTable_4 = removevars(auxTable2,{'reading'});
head(auxTable_4)
```

gen	math	writing

```
female     59       78
male       96       87
female     57       77
male       70       63
female     83       86
male       68       54
female     82       80
female     46       58
```

###Filtrado de informacion basado en reglas de valores

```
%Podemos tambien filtrar los datos por grupos de informacion
%filtrado de la informacion de la tabla auxTable_2 para resultados de datos
%de prueba Matemáticas mayores a 60 pero menores a 80
auxTable = table(gen,math,reading,writing);
head(auxTable)
```

gen	math	reading	writing
female	59	70	78
male	96	93	87
female	57	76	77
male	70	70	63
female	83	85	86
male	68	57	54
female	82	83	80
female	46	61	58

```
promBetween_60_80 = groupfilter(auxTable,"math",@(x) min(x) >= 60 && max(x) <=
80,"math");
head(promBetween_60_80)
```

gen	math	reading	writing
male	70	70	63
male	68	57	54
male	80	75	73
male	74	69	69
male	76	74	76
male	70	73	70
female	80	82	85
female	65	71	74

```
auxMean_Table = groupsummary(promBetween_60_80,"gen","max")
```

auxMean_Table = 2×5 table

	gen	GroupCount	max_math	max_reading	max_writing
1	female	243	80	97	99
2	male	245	80	88	87

*** Ejercicio ***

Realiza las siguientes operaciones

- Crea una tabla que contenga las siguientes categorias en este orden: Preparacion de la prueba, genero, Resultado de matemáticas, Resultado de lectura, Resultado de escritura
- Calcula el promedio de los tres valores de las pruebas y coloca el valor al final de las columnas
- Agrupa los resultados por promedio y clasificalos para valores mayores o iguales a 60 pero menores o iguales a 80. ¿Cuantos datos quedaron en total?
- Cual es el promedio general del género masculino que si se preparo para la prueba
- Cual es el promedio general del género femenino que no se preparo para la prueba
- De acuerdo con los datos, ¿Existe una relación directa entre prepararse para la prueba y no prepararse para la prueba?

```
% 1. Crea una tabla que contenga las siguientes categorias en este orden:
Preparacion de la prueba, genero, Resultado de matemáticas, Resultado de lectura,
Resultado de escritura
PreparacionPrueba = dsStudents_2.testPreparationCourse
```

```
PreparacionPrueba = 1000×1 categorical
completed
none
none
none
none
none
none
none
none
completed
    ⋮
```

```
Genero = dsStudents_2.gender;
ResultadoMatematicas = dsStudents_2.mathScore;
ResultadoLectura = dsStudents_2.readingScore;
ResultadoEscritura = dsStudents_2.writingScore;
auxTable1 =
table(PreparacionPrueba,Genero,ResultadoMatematicas,ResultadoLectura,ResultadoEscrit
ura);
head(auxTable1)
```

PreparacionPrueba	Genero	ResultadoMatematicas	ResultadoLectura	ResultadoEscritura
completed	female	59	70	78
none	male	96	93	87
none	female	57	76	77
none	male	70	70	63
none	female	83	85	86
none	male	68	57	54
none	female	82	83	80
none	female	46	61	58

```
% 2. Calcula el promedio de los tres valores de las pruebas y coloca el valor al
final de las columnas
promedio = mean([auxTable1.ResultadoMatematicas, auxTable1.ResultadoLectura,
auxTable1.ResultadoEscritura],2);
```

```
auxTable2= addvars(auxTable1,promedio);
head(auxTable2)
```

PreparacionPrueba	Genero	ResultadoMatematicas	ResultadoLectura	ResultadoEscritura	promedio
completed	female	59	70	78	69
none	male	96	93	87	92
none	female	57	76	77	70
none	male	70	70	63	67.667
none	female	83	85	86	84.667
none	male	68	57	54	59.667
none	female	82	83	80	81.667
none	female	46	61	58	55

```
% 3. Agrupa los resultados por promedio y clasificalos para valores mayores o
iguales a 60 pero menores o iguales a 80. ¿Cuantos datos quedaron en total?
testsBetween_60_80 = groupfilter(auxTable2,
["ResultadoMatematicas","ResultadoLectura","ResultadoEscritura","promedio"],@(x)
min(x) >= 60 && max(x) <= 80,
["ResultadoMatematicas","ResultadoLectura","ResultadoEscritura","promedio"]);
head(testsBetween_60_80)
```

PreparacionPrueba	Genero	ResultadoMatematicas	ResultadoLectura	ResultadoEscritura	promedio
none	male	70	70	63	67.667
none	male	80	75	73	76
none	male	74	69	69	70.667
none	male	76	74	76	75.333
none	male	70	73	70	71
completed	female	65	71	74	70
none	male	66	66	62	64.667
completed	female	67	71	76	71.333

```
auxMean_Table = groupsummary(testsBetween_60_80,
["ResultadoMatematicas","ResultadoLectura","ResultadoEscritura","promedio"])
```

auxMean_Table = 300×5 table

...

	ResultadoMatematicas	ResultadoLectura	ResultadoEscritura	promedio
1	60	65	67	64
2	60	66	64	63.3333
3	60	67	61	62.6667
4	60	67	74	67
5	60	69	70	66.3333
6	60	73	68	67
7	60	74	68	67.3333
8	60	74	72	68.6667
9	60	75	72	69
10	60	80	76	72

	ResultadoMatematicas	ResultadoLectura	ResultadoEscritura	promedio
11	61	63	70	64.6667
12	61	64	63	62.6667
13	61	66	71	66
14	61	68	67	65.3333
15	61	69	62	64
16	61	69	66	65.3333
17	61	69	69	66.3333
18	61	69	73	67.6667
19	61	70	73	68
20	61	73	70	68
21	61	73	74	69.3333
22	61	74	71	68.6667
23	62	62	68	64
24	62	64	61	62.3333
25	62	64	63	63
26	62	64	64	63.3333
27	62	65	61	62.6667
28	62	65	65	64
29	62	65	69	65.3333
30	62	67	68	65.6667
31	62	67	70	66.3333
32	62	68	65	65
33	62	68	74	68
34	62	68	75	68.3333
35	62	69	60	63.6667
36	62	69	67	66
37	62	70	64	65.3333
38	62	71	61	64.6667
39	62	71	62	65
40	62	71	74	69
41	62	72	67	67
42	62	74	71	69
43	62	74	74	70

	ResultadoMatematicas	ResultadoLectura	ResultadoEscritura	promedio
44	62	75	73	70
45	62	75	79	72
46	62	78	71	70.3333
47	62	78	74	71.3333
48	63	61	64	62.6667
49	63	63	63	63
50	63	69	63	65
51	63	69	66	66
52	63	69	68	66.6667
53	63	71	65	66.3333
54	63	72	60	65
55	63	72	78	71
56	63	73	74	70
57	63	74	75	70.6667
58	63	76	67	68.6667
59	63	78	74	71.6667
60	64	62	61	62.3333
61	64	63	64	63.6667
62	64	66	62	64
63	64	66	69	66.3333
64	64	66	72	67.3333
65	64	67	72	67.6667
66	64	67	75	68.6667
67	64	68	60	64
68	64	70	72	68.6667
69	64	71	68	67.6667
70	64	72	66	67.3333
71	64	72	79	71.6667
72	64	77	77	72.6667
73	64	77	79	73.3333
74	65	62	63	63.3333
75	65	64	61	63.3333
76	65	64	73	67.3333

	ResultadoMatematicas	ResultadoLectura	ResultadoEscritura	promedio
77	65	65	62	64
78	65	66	61	64
79	65	66	66	65.6667
80	65	68	67	66.6667
81	65	71	74	70
82	65	72	76	71
83	65	73	77	71.6667
84	65	76	70	70.3333
85	65	77	70	70.6667
86	65	77	74	72
87	65	79	73	72.3333
88	66	60	64	63.3333
89	66	62	66	64.6667
90	66	63	60	63
91	66	63	63	64
92	66	65	76	69
93	66	66	62	64.6667
94	66	66	75	69
95	66	68	62	65.3333
96	66	70	67	67.6667
97	66	72	74	70.6667
98	66	73	69	69.3333
99	66	73	72	70.3333
100	66	77	80	74.3333

⋮

Después de filtar la información según lo establecido, quedaron 300 datos.

```
% 4. Cual es el promedio general del género masculino que si se preparo para la
prueba
promedio_masculino_preparado = mean(testsBetween_60_80(testsBetween_60_80.
('Genero') == 'male' & testsBetween_60_80.('PreparacionPrueba') == 'completed', :).
('promedio'));
% 5. Cual es el promedio general del género femenino que no se preparo para la
prueba
promedio_femenino_no_preparado = mean(testsBetween_60_80(testsBetween_60_80.
('Genero') == 'female' & testsBetween_60_80.('PreparacionPrueba') == 'none', :).
('promedio'));
```

```
% 6. De acuerdo con los datos, ¿Existe una relación directa entre prepararse para
la prueba y no prepararse para la prueba?
num_datos_60_80 = height(testsBetween_60_80);
disp(['Total de datos con promedio entre 60 y 80: ' num2str(num_datos_60_80)]);
```

```
Total de datos con promedio entre 60 y 80: 303
```

```
disp(['Promedio general masculino preparado: '
num2str(promedio_masculino_preparado)]);
```

```
Promedio general masculino preparado: 70.3838
```

```
disp(['Promedio general femenino no preparado: '
num2str(promedio_femenino_no_preparado)]);
```

```
Promedio general femenino no preparado: 70.3129
```

```
% Según los datos, no hay una relación directa entre prepararse para la
% prueba y no hacerlo porque la variación es de centésimas.
```

NOTA IMPORTANTE:

Utilicé códigos hechos por ChatGPT.

Ingresé todas las instrucciones del ejercicio al chat pero sólo utilicé los códigos a partir del punto 4; para lograr obtener el punto 3 di una instrucción específica: *como sacar el promedio de tres columnas y visualizarlo en una tabla;* de esta manera pude obtener el promedio directamente en una tabla y no en una subtabla o que tuviera datos categóricos; el código lo modifique a mi beneficio para que MATLAB pudiera correr el código sin problemas.

Discretizacion de datos en categorias

```
%Vamos a establecer los siguientes valores de calificaciones en el formato
%americano de calificaciones donde
%Valor | calificacion
%0 59  | F
%60 69 | D
%70 79 | C
%80 89 | B
%90 99 | A
%100   | A+

%Primero estableceremos nuestras nuevas categorias de datos
%etiquetas
calAme = {'F', 'D', 'C', 'B', 'A', 'A+'};

%Limites de calificaciones
limit = [0 59 69 79 89 99 100];
auxVal = dsStudents_3.mathScore; %Solo para guardar el valor
```

```
dsStudents_3.mathScore = discretize(dsStudents_3.mathScore, limit, 'categorical',
calAme);
summary(dsStudents_3.mathScore)
```

```
    F       256
    D       256
    C       219
    B       186
    A        70
    A+       13
```

```
head(dsStudents_3)
```

gender	mathScore	readingScore	writingScore
female	D	70	78
male	A	93	87
female	F	76	77
male	C	70	63
female	B	85	86
male	D	57	54
female	B	83	80
female	F	61	58

```
dsStudents_3.mathScore = auxVal; %Regreso el valor original para no entorpecer la BD
head(dsStudents_3)
```

gender	mathScore	readingScore	writingScore
female	59	70	78
male	96	93	87
female	57	76	77
male	70	70	63
female	83	85	86
male	68	57	54
female	82	83	80
female	46	61	58