

```
%Fecha de actualización: 08/Abril/2023
%clc
%clear all
```

Portada

```
figure(15);
% tamaño de la figura
% set(gcf, 'Position', [100, 100, 800, 600]);
tamano_letra=10;
% Cargar y mostrar la primera imagen
img1 = imread('UNAMlogo.png'); % Reemplaza 'Imagen1.jpg' con el nombre de tu primera imagen
subplot(4, 4, 1); % Divide la figura en 1 fila y 2 columnas, y selecciona la primera posición
imshow(img1);

% Cargar y mostrar la segunda imagen
img2 = imread('FIlogo.jpg'); % Reemplaza 'Imagen2.jpg' con el nombre de tu segunda imagen
subplot(4, 4, 4); % Selecciona la segunda posición
imshow(img2);

% Agregar texto a un subplot específico (por ejemplo, el subplot 8)
subplot(4, 4, [2,3]);
text(0.5, 0.9, 'UNAM', 'FontSize', 14, 'HorizontalAlignment', 'center', 'Color', 'b');
text(0.5, 0.5, 'Facultad de ingeniería', 'FontSize', tamano_letra, 'FontWeight', 'bold', 'HorizontalAlignment', 'center');
text(0.5, 0.0, 'TSIBS', 'FontSize', tamano_letra, 'HorizontalAlignment', 'center');
axis off
subplot(4, 4, [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]);
text(0.5, 0.9, 'Alumna: Flores Morín María Alejandra', 'FontSize', tamano_letra, 'HorizontalAlignment', 'center');
text(0.5, 0.8, 'No. Cuenta: 315165805', 'FontSize', tamano_letra, 'HorizontalAlignment', 'center');
text(0.5, 0.6, 'Profesor: Dr. en C. Luis Antonio Aguilar Pérez ', 'FontSize', tamano_letra, 'HorizontalAlignment', 'center');
text(0.5, 0.4, 'Semestre 2024-1', 'FontSize', tamano_letra, 'HorizontalAlignment', 'center');
text(0.5, 0.3, 'tareaEjercicio1 Ejercicios de DATASTORE en MATLAB', 'FontSize', tamano_letra, 'HorizontalAlignment', 'center');
text(0.5, 0.1, 'Fecha de entrega: 26 de septiembre del 2023', 'FontSize', tamano_letra, 'HorizontalAlignment', 'center');
axis off
```



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 información dependiendo del lenguaje de programación utilizado. El formato universal y más tradicional de almacenamiento de la información es mediante archivos de tipo CSV. Un archivo de formato CSV es en realidad un archivo en formato de codificación 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:

header ("features")	"gender", "race/ethnicity", "parental level of education", "lunch", "test preparation course", "math score", "reading score", "writing score"
	"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"
	"male", "group C", "some high school", "standard", "none", "88", "75", "73"
	"female", "group C", "bachelor's degree", "standard", "completed", "57", "69", "77"

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 organización de la información sería lo más cercano a este:

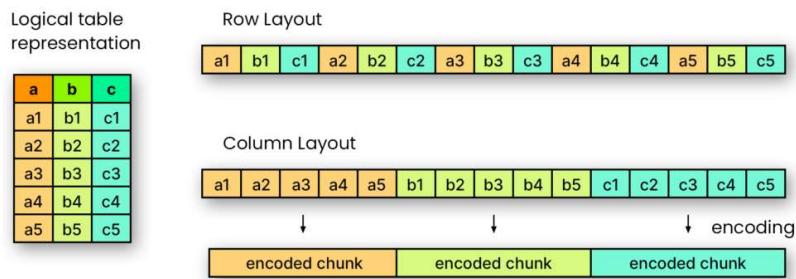


Figura 2.- Esquema tradicional de codificación tipo parquet

Finalmente, el uso y manejo de modelos de datos en MATLAB se realiza mediante la función "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 información

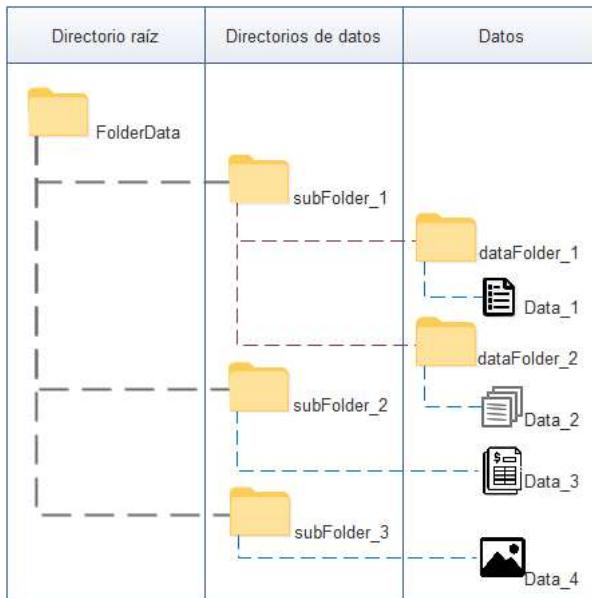


Figura 3.- Esquema de un modelo de datos

--- Pasos iniciales

```
%%
%Creacion de directorio de trabajo
%rootFolder = 'D:\TSISB_IA';
rootFolder = 'C:\Users\puma_\Documents\TSISB_IA';
workingFolder = 'practica_2';
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

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

if isEqual(fileName,0)
    disp('Se canceló la búsqueda 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 movió a la dirección ', fullfile(saveTempPath)]);
    end
end

```

```

El usuario seleccionó el archivo C:\Users\puma_\Documents\TSISB_IA\practica_2\students.csv
El cual se movió a la dirección C:\Users\puma_\Documents\TSISB_IA\practica_2\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:

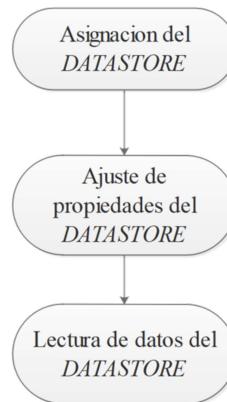


Figura 2.- Flujo de trabajo de un Datastore

La primera parte del pipeline del datastore incluye operaciones como:

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

```

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

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

%Visualización rápida de las primeras líneas del archivo de texto
%Como solo es un archivo se puede utilizar el comando dbtype
dbtype(archivoCSV, '1:5')

```

```

1 gender,race/ethnicity,parental level of education,lunch,test preparation course,math score,reading score,writing score
2 female,group D,some college,standard,completed,59,70,78
3 male,group D,associate's degree,standard,none,96,93,87
4 female,group D,some college,free/reduced,none,57,76,77
5 male,group B,some college,free/reduced,none,70,70,63

```

```

%Asignamos nuestro datastore dentro de MATLAB
%dsGeneral = datastore(archivoCSV)
%Si tuviéramos muchos archivos, se utilizaría el siguiente comando
%preview(dsGeneral)

```

```

%Desde la versión 2019, los nombres de las variables pueden incluir
%cualquier tipo de símbolos, además de no necesariamente comenzar solo

```

```

%con letras, por lo que matlab requiere el Flag "preserve" para considerar
%esta opción
%dsGeneral.VariableNamingRule='preserve';

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

%Modificación de los nombres utilizados para cada característica/feature
%values=dsGeneral.VariableNames;
%Modificación como si fuera un indexado de valores clásico
%newValues=["hola", "esta", "es", "una", "prueba", "de", "cambio", "de variables"];
%dsGeneral.VariableNames = newValues;
%preview(dsGeneral)

```

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

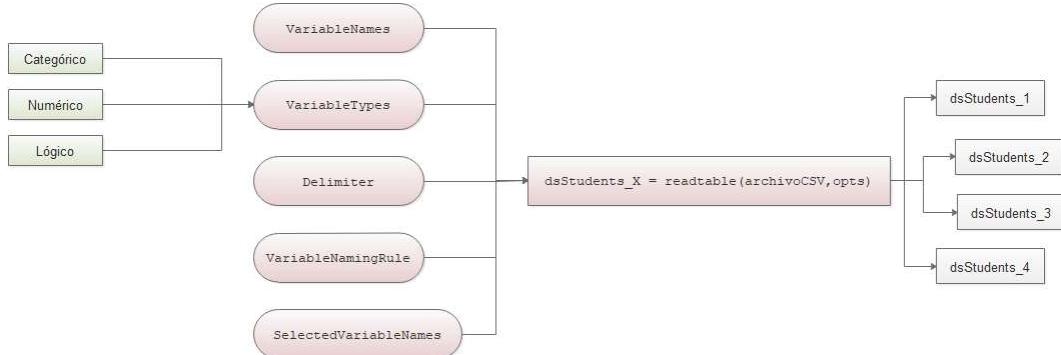


Figura 3.- Propiedades de un datastore de tipo categorico-numérico

```

%Visualización general de las opciones de importación 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:
```

```

%Visualización 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'}
```

```
%Modificación 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 versión 2019, los nombres de las variables pueden incluir
%cuálquier tipo de símbolos, además 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 =',';

%Modificación de los nombres utilizados para cada característica/feature
%Asignación a una variable particular
values=opts.VariableNames;
%Modificación 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	readingScore	writingScore
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

```
%Podemos elegir con qué características 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 categorías de datos
- Tabla 3: Datos categóricos de género y datos numéricos de las pruebas de matemáticas, lectura y escritura
- Tabla 4: Datos categóricos de raza étnica 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 categóricos contenidos en la tabla 2

Resumen de datos categóricos

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

```
grpCatGenero = 2x1 cell
'female'
'male'
```

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

```
numCatGenero = 2x1
492
508
```

```
summary(genero)
```

female	492
male	508

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

```
grpCatEtn = 5x1 cell
'group A'
'group B'
'group C'
'group D'
'group E'
```

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

```
numCatEtn = 5x1
79
198
323
257
143
```

%Vamos a reemplazar las categorías originales 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 categorías existen en:

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

Además deberás cambiar el nombre de las categorías en la columna "preparación de la prueba" de la siguiente manera

- completed - > terminado

- none -> no terminado

1) Mostrar cuantas categorias existen en el nivel de educacion de los padres.

```
nivelEducacion = categorical(dsStudents_2.parentalLevelOfEducation);
%grpCatGenero = categories(nivelEducacion)
%numCatGenero = countcats(nivelEducacion)
summary(nivelEducacion)
```

associate's degree	204
bachelor's degree	105
high school	215
master's degree	75
some college	224
some high school	177

2) Mostrar cuantas categorias existen en si contaron con un desayuno o no.

```
almuerzo = categorical(dsStudents_2.lunch);
summary(almuerzo)
```

free/reduced	340
standard	660

3) Mostrar cuantas categorias existen en si hubo una preparacion previa a la prueba

```
curso = categorical(dsStudents_2.testPreparationCourse);
summary(curso)
```

completed	344
none	656

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

```
nuevasCategorias = {'terminado',...
    'no terminado'};
nuevasCategoria = renametcats(curso,nuevasCategorias);
summary(nuevasCategoria)
```

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 = 2x5 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 = 5x5 table

	race_ethnicity	GroupCount	mean_mathScore	mean_readingScore	mean_writingScore
1	group A	79	65.6962	69.2025	67.8481
2	group B	198	64.0707	68.5303	66.7172

3	groupC	race_ethnicity	GroupCount	323	mean_mathScore	68.5108	mean_readingScore	68.6099	mean_writingScore	66.8050
4	group D			257		68.8794		70.9300		71.0584
5	group E			143		77.4266		76.6154		75.0350

%Podemos realizar operaciones por conjuntos específicos 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 = 1000x5 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
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

	gen	race	math	reading	writing
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
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
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	gen	group race	GroupCount_151	mean_math_63.0199	mean_reading_71.9934	mean_writing_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

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

```
% gen = dsStudents_2.gender;
levelOfEducation = dsStudents_2.parentalLevelOfEducation;
% math = dsStudents_2.mathScore;
% reading = dsStudents_2.readingScore;
% writing = dsStudents_2.writingScore;
auxTable1 = table(gen,levelOfEducation,math, reading, writing);
genderMeanvsLevelOfEducation = groupsummary(auxTable1,[ "gen", "levelOfEducation"], "median")
```

genderMeanvsLevelOfEducation = 12x6 table

	gen	levelOfEducation	GroupCount	median_math	median_reading	median_writing
1	female	associate's degree	101	69	76	77
2	female	bachelor's degree	52	66	78	78
3	female	high school	116	61	71	69
4	female	master's degree	33	68	75	77
5	female	some college	106	66	74	75
6	female	some high school	84	62	72.5000	72
7	male	associate's degree	103	74	70	67
8	male	bachelor's degree	53	71	67	67
9	male	high school	99	72	68	65
10	male	master's degree	42	75	70	71.5000
11	male	some college	118	72.5000	68	67
12	male	some high school	93	67	62	59

2) Cual es el promedio de la raza etnica vs si se preparó o no para la prueba

```
preparationCourse= dsStudents_2.testPreparationCourse;

auxTable2 = table(race,preparationCourse,math, reading, writing);
genderMeanvsLevelOfEducation = groupsummary(auxTable2,[ "race", "preparationCourse"], "mean")
```

genderMeanvsLevelOfEducation = 10x6 table

	race	preparationCourse	GroupCount	mean_math	mean_reading	mean_writing
1	group A	completed	31	70.0645	72.4516	73.2258
2	group A	none	48	62.8750	67.1042	64.3750
3	group B	completed	66	66.8636	73.1818	73.8939
4	group B	none	132	62.6742	66.2045	63.1288
5	group C	completed	101	66.7822	72.6139	73.4257
6	group C	none	222	64.9324	66.7883	63.7928
7	group D	completed	97	71.1753	75.5258	77.4124
8	group D	none	160	67.4875	68.1437	67.2062
9	group E	completed	49	80.8367	81.0204	81.7551
10	group E	none	94	75.6489	74.3191	71.5319

3) Cual es el promedio del genero vs si se preparó o no para la prueba

```
auxTable3 = table(gen,preparationCourse,math, reading, writing);
genderMeanvsLevelOfEducation = groupsummary(auxTable3,[ "gen", "preparationCourse"], "mean")
```

genderMeanvsLevelOfEducation = 4x6 table

	gen	preparationCourse	GroupCount	mean_math	mean_reading	mean_writing
1	female	completed	31	70.0645	72.4516	73.2258
2	female	none	48	62.8750	67.1042	64.3750
3	male	completed	66	66.8636	73.1818	73.8939
4	male	none	132	62.6742	66.2045	63.1288

1	female	gen	completed	preparationCourse	GroupCount_177	mean_math_67.5650	mean_reading_77.9288	mean_writing_79.8644
2	female		none		315	63.2063	71.0825	69.8286
3	male		completed		167	73.2695	71.5449	71.5090
4	male		none		341	69.5161	65.3519	61.7771

Modificación de las columnas en modelos de datos

Podemos realizar ligeras modificaciones a las columnas (características 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)
```

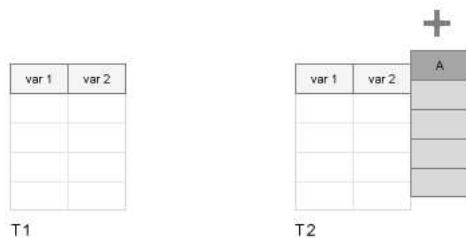


Figura 5.- Agregado de características de un modelo de datos

```
%Agregado de valores a tablas ya existentes
auxTable_2 = addvars(auxTable_1,writing);
head(auxTable_2)
```

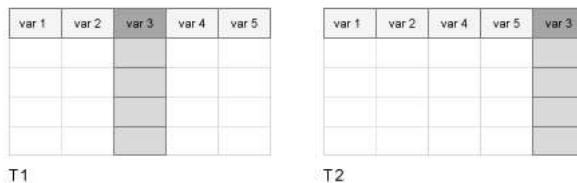


Figura 6.- Reordenamiento de características de un modelo de datos

```
%Reordenamiento de características de la tabla
auxTable_3 = movevars(auxTable_2, 'reading', 'after', 'writing');
head(auxTable_3)
```

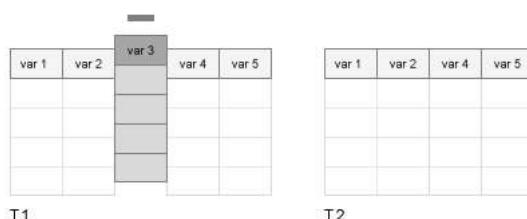


Figura 7.- Eliminación de características de un modelo de datos

```
%Eliminación de características de la tabla
auxTable_4 = removevars(auxTable_2,{'reading'});
head(auxTable_4)
```

###Filtrado de información 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

```
mathBetween_60_80 = groupfilter(auxTable, "math", @(x) min(x) >= 60 && max(x) <= 80, "math");
head(mathBetween_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(mathBetween_60_80, "gen", "max")
```

```
auxMean_Table = 2x5 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 categorías 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.

```
Ej3auxTable1 = table(preparationCourse,gen,math,reading,writing)
```

```
Ej3auxTable1 = 1000x5 table
```

	preparationCourse	gen	math	reading	writing
1	completed	female	59	70	78
2	none	male	96	93	87
3	none	female	57	76	77
4	none	male	70	70	63
5	none	female	83	85	86
6	none	male	68	57	54
7	none	female	82	83	80
8	none	female	46	61	58
9	none	male	80	75	73
10	completed	female	57	69	77
11	none	male	74	69	69
12	none	male	53	50	49
13	none	male	76	74	76
14	none	male	70	73	70
15	none	male	55	54	52

	preparationCourse	gen	math	reading	writing
16	none	male	56	46	43
17	none	female	35	47	41
18	none	female	87	92	81
19	none	female	80	82	85
20	completed	female	65	71	74
21	none	male	66	66	62
22	completed	female	67	71	76
23	none	female	70	71	71
24	none	male	89	88	86
25	completed	male	99	85	88
26	none	male	74	83	72
27	none	male	58	52	51
28	none	male	70	66	59
29	none	female	80	79	71
30	none	male	90	87	86
31	completed	female	80	81	85
32	none	female	68	76	79
33	completed	female	69	78	75
34	none	female	32	35	37
35	completed	male	82	82	82
36	none	female	57	53	54
37	none	female	69	74	75
38	completed	male	68	66	72
39	completed	male	74	85	87
40	none	male	89	85	78
41	completed	male	46	46	48
42	completed	male	76	82	77
43	none	male	86	82	72
44	none	male	69	73	67
45	none	female	52	56	54
46	none	male	63	71	65
47	completed	male	96	82	90
48	completed	male	80	76	68
49	none	female	59	52	56
50	completed	male	80	77	80
51	completed	female	65	77	74
52	completed	female	74	83	84
53	none	male	90	93	84
54	completed	female	69	72	72
55	none	male	69	67	63
56	none	female	62	64	61
57	none	female	67	75	80
58	completed	female	89	93	93
59	none	female	79	86	78
60	none	male	67	66	66
61	completed	male	82	74	75
62	completed	male	63	69	63
63	none	female	71	83	80
64	none	female	55	68	73
65	none	female	61	74	71
66	none	female	35	34	36
67	none	male	75	77	66
68	completed	female	73	91	88
69	none	female	56	62	57
70	none	male	80	70	73
71	completed	male	83	81	78
72	completed	female	64	82	80
73	none	female	23	33	33
74	completed	female	41	58	59
75	completed	male	61	49	52

	preparationCourse	gen	math	reading	writing
76	none	male	63	46	46
77	completed	male	84	91	89
78	none	male	55	61	59
79	completed	male	85	75	74
80	none	male	65	61	57
81	none	male	88	80	81
82	completed	male	91	93	95
83	none	female	51	46	42
84	completed	male	73	77	76
85	none	female	73	89	89
86	none	male	100	97	91
87	completed	female	48	68	68
88	none	male	98	79	85
89	none	male	68	65	60
90	none	male	64	62	58
91	completed	male	72	67	61
92	none	female	63	74	75
93	none	male	43	51	38
94	none	male	80	75	74
95	none	female	71	88	83
96	completed	female	91	96	97
97	completed	female	68	84	87
98	none	female	73	80	78
99	completed	female	75	90	95
100	none	male	83	62	64

:

```
head(Ej3auxTable1)
```

preparationCourse	gen	math	reading	writing
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.

```
% Ej3auxTable2 = table(preparationCourse,gen,math,reading,writing);
% head(Ej3auxTable2);
% promedio = mean([Ej3auxTable2.math, Ej3auxTable2.reading, Ej3auxTable2.writing], 2)
subjectsMeanTable = groupsummary(Ej3auxTable2,["preparationCourse","gen"],"mean")
```

subjectsMeanTable = 4x6 table

	preparationCourse	gen	GroupCount	mean_math	mean_reading	mean_writing
1	completed	female	177	67.5650	77.7288	79.8644
2	completed	male	167	73.2695	71.5449	71.5090
3	none	female	315	63.2063	71.0825	69.8286
4	none	male	341	69.5161	65.3519	61.7771

```
head(subjectsMeanTable)
```

preparationCourse	gen	GroupCount	mean_math	mean_reading	mean_writing
completed	female	177	67.565	77.729	79.864
completed	male	167	73.269	71.545	71.509
none	female	315	63.206	71.083	69.829
none	male	341	69.516	65.352	61.777

% Agregado de valores a tablas ya existentes

```
% Ej3auxTable2_1 = addvars(Ej3auxTable2,promedio);
% head(Ej3auxTable2_1)
```

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?

```
% %Reordenamiento de columna promedio de la tabla
% Ej3auxTable3_1 = movevars(Ej3auxTable2_1,'promedio','Before','preparationCourse');
% head(Ej3auxTable3_1)
% %Agrupar por promedio
% avgProm = groupsummary(Ej3auxTable3_1,[ "promedio","preparationCourse","gen"], "mean");
% head(avgProm);
% subjectsMeanTable
%Valores mayores o iguales a 60 pero menores o iguales a 80
mathBetween_60_80 = groupfilter(subjectsMeanTable,[ "mean_math","mean_reading","mean_writing"],@(x) min(x) >= 60 && max(x) <= 80, [ "mean_mat
head(mathBetween_60_80);
```

preparationCourse	gen	GroupCount	mean_math	mean_reading	mean_writing
completed	female	177	67.565	77.729	79.864
completed	male	167	73.269	71.545	71.509
none	female	315	63.206	71.083	69.829
none	male	341	69.516	65.352	61.777

```
% %Valores mayores o iguales a 60 pero menores o iguales a 80
% mathBetween_60_80 = groupfilter(avgProm,"promedio",@(x) min(x) >= 60 && max(x) <= 80,"promedio");
% head(mathBetween_60_80);
suma = sum([mathBetween_60_80.GroupCount], 1);
disp(['En total quedaron: ', num2str(suma), ' datos']);
```

En total quedaron: 1000 datos

4) Cual es el promedio general del género masculino que si se preparo para la prueba.

```
% % %Reordenamiento de columna promedio de la tabla para el ej 4
% Ej3auxTable3_4 = movevars(Ej3auxTable2_1,'promedio','after','gen');
% head(Ej3auxTable3_4);
% %Agrupar por Curso de Preparación (principalmente)
% avgProm4 = groupsummary(Ej3auxTable3_4,[ "preparationCourse","gen","promedio"], "mean");
% %Filtro //Basado en una respuesta generada por ChatGPT.
% filtro = (avgProm4.gen == 'male') & (avgProm4.preparationCourse == 'completed');
% tabla_filtrada = avgProm4(filtro, :);
% %Promedio de los estudiantes masculinos que se prepararon para la prueba
% promedio_general_masculino_preparados = mean(tabla_filtrada.promedio);
%
% disp(['Promedio general de estudiantes masculinos que se prepararon: ', num2str(promedio_general_masculino_preparados)]);
filtro = (subjectsMeanTable.gen == 'male') & (subjectsMeanTable.preparationCourse == 'completed');
tabla_filtrada_4 = subjectsMeanTable(filtro, :);
head(tabla_filtrada_4);
```

preparationCourse	gen	GroupCount	mean_math	mean_reading	mean_writing
completed	male	167	73.269	71.545	71.509

```
promedio_4 = mean([tabla_filtrada_4.mean_math, tabla_filtrada_4.mean_reading, tabla_filtrada_4.mean_writing], 2);
disp(['Promedio general de estudiantes masculinos que se prepararon: ', num2str(promedio_4)]);
```

Promedio general de estudiantes masculinos que se prepararon: 72.1078

5) Cual es el promedio general del género femenino que no se preparo para la prueba.

```
% %Reordenamiento de columna promedio de la tabla para el ej 5
% Ej3auxTable3_5 = movevars(Ej3auxTable2_1,'promedio','after','gen');
% head(Ej3auxTable3_5);
% %Agrupar por Curso de Preparación (principalmente)
% avgProm4 = groupsummary(Ej3auxTable3_5,[ "preparationCourse","gen","promedio"], "mean");
% %Filtro //Basado en una respuesta generada por ChatGPT.
% filtro = (avgProm4.gen == 'female') & (avgProm4.preparationCourse == 'none');
% tabla_filtrada = avgProm4(filtro, :);
% %Promedio de los estudiantes femeninos que no se prepararon
% promedio_general_masculino_preparados = mean(tabla_filtrada.promedio);
%
% disp(['Promedio general de estudiantes femeninos que no se prepararon: ', num2str(promedio_general_masculino_preparados)]);
filtro = (subjectsMeanTable.gen == 'female') & (subjectsMeanTable.preparationCourse == 'none');
tabla_filtrada_5 = subjectsMeanTable(filtro, :);
head(tabla_filtrada_5);
```

preparationCourse	gen	GroupCount	mean_math	mean_reading	mean_writing
none	female	315	63.206	71.083	69.829

```
promedio_4 = mean([tabla_filtrada_5.mean_math, tabla_filtrada_5.mean_reading, tabla_filtrada_5.mean_writing], 2);
disp(['Promedio general de estudiantes masculinos que se prepararon: ', num2str(promedio_4)]);
```

Promedio general de estudiantes masculinos que se prepararon: 68.0392

6) De acuerdo con los datos, ¿Existe una relación directa entre prepararse para la prueba y no prepararse para la prueba?

Si existe una relación directamente proporcional. Si observamos la tabla los primeros 5 estudiantes con la mayor calificación tomaron el curso de preparación para la prueba y en los promedio más bajos estan aquellos que no tomaron el curso de preparación. En los resultados entre 60 y 80 de promedio, hay una diferencia aproximada de 9 puntos entre aquellos que no tomaron el curso contra aquellos que si lo tomaron.

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

Gráficos

Existen distintas formas de visualizar la información.

--Histogramas

Por ejemplo podemos ver la distribucion de datos mediante un histograma

```
%Muestra el histograma de los datos que pertenecen a un genero masculino o  
%femenino  
close all  
fig01=figure;  
hold on  
histogram(genero,'BarWidth',0.5)  
ylabel("Número de muestras")  
xlabel("Categorías de género")  
hold off
```

--Swarm

```
fig02=figure;  
hold on  
%Esta gráfica nos muestra las categorias y la densidad de valores numéricos  
%asociados a cada categoría  
%En el eje X colocaremos las categorias que queremos visualizar  
x=categorical(dsStudents_2.gender);  
  
%Por su parte, en el eje y colocaremos los valores numéricos que  
%queremos visualizar, en este caso los valores de la prueba de  
%matemáticas  
y=dsStudents_2.mathScore;  
  
swarmchart(x,y);  
hold off
```

--Swarm 3D

```
fig03=figure;  
hold on  
%Podemos realizar una densidad de datos en tres dimensiones de la  
%siguiente manera  
%En el eje x tendremos nuestra primera categoría de datos  
x=categorical(dsStudents_2.gender);  
  
%En el eje y tendremos nuestra segunda categoría de datos  
y=categorical(dsStudents_2.race_ethnicity);  
  
%Finalmente en el eje z tendremos la densidad de datos de la prueba de  
%matemáticas  
z=dsStudents_2.mathScore;
```

```

swarmchart3(x,y,z,50,genero,'*')

view([60 40])
grid minor
set(gca,'FontSize' ,20, ...
    'MinorGridColor' , 'g',...
    'GridColor' , 'r',...
    'Xgrid' , 'on',...
    'Ygrid' , 'on',...
    'LineWidth' ,2)
hold off

```

---Gráficos de pay

```

%Primero creamos nuestras categorias
%El nivel de educacion de los padres y el genero del hijo
nivelEducacion=dsStudents_2.parentalLevelOfEducation;
    nivelEducacionyGeneroF = nivelEducacion(dsStudents_2.gender == 'female');
    nivelEducacionyGeneroM = nivelEducacion(dsStudents_2.gender == 'male');
grpCatEtn = categories(dsStudents_2.parentalLevelOfEducation);

fig04=figure;
hold on
t=tiledlayout(1,2,'TileSpacing','compact');
ax1 = nexttile;
    pie(nivelEducacionyGeneroF)
    title('Género femenino')
ax2 = nexttile;
    pie(nivelEducacionyGeneroM)
    title('Género masculino')
lgd = legend(grpCatEtn);
lgd.Layout.Tile = 'east';
hold off

```

---Scatter(gráfico "x" VS "y")

```

valMathScore_F = dsStudents_3.mathScore(genero == 'female');
valMathScore_M = dsStudents_3.mathScore(genero == 'male');

valReadingScore_F = dsStudents_3.readingScore(genero == 'female');
valReadingScore_M = dsStudents_3.readingScore(genero == 'male');

valWritingScore_F = dsStudents_3.writingScore(genero == 'female');
valWritingScore_M = dsStudents_3.writingScore(genero == 'male');

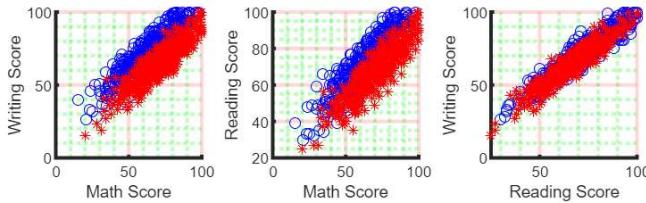
fig05=figure;
t=tiledlayout(1,3,'TileSpacing','compact');
ax1 = nexttile;
    hold on
        axis square;
        scatter(valMathScore_F,valWritingScore_F,'bo');
        scatter(valMathScore_M,valWritingScore_M,'r*');
        xlabel('Math Score')
        ylabel('Writing Score')
        grid minor
        set(gca,'MinorGridColor' , 'g',...
            'GridColor' , 'r',...
            'Xgrid' , 'on',...
            'Ygrid' , 'on',...
            'LineWidth' ,2)
    hold off
ax2 = nexttile;
    hold on
        axis square;
        scatter(valMathScore_F,valReadingScore_F,'bo');
        scatter(valMathScore_M,valReadingScore_M,'r*');
        xlabel('Math Score')
        ylabel('Reading Score')
        grid minor
        set(gca,'MinorGridColor' , 'g',...
            'GridColor' , 'r',...
            'Xgrid' , 'on',...
            'Ygrid' , 'on',...
            'LineWidth' ,2)
    hold off
ax3 = nexttile;
    hold on
        axis square;
        scatter(valReadingScore_F,valWritingScore_F,'bo');
        scatter(valReadingScore_M,valWritingScore_M,'r*');
        xlabel('Reading Score')
        ylabel('Writing Score')

```

```

grid minor
set(gca,'MinorGridColor','g',...
    'GridColor','r',...
    'Xgrid','on',...
    'Ygrid','on',...
    'LineWidth',2)
hold off

```

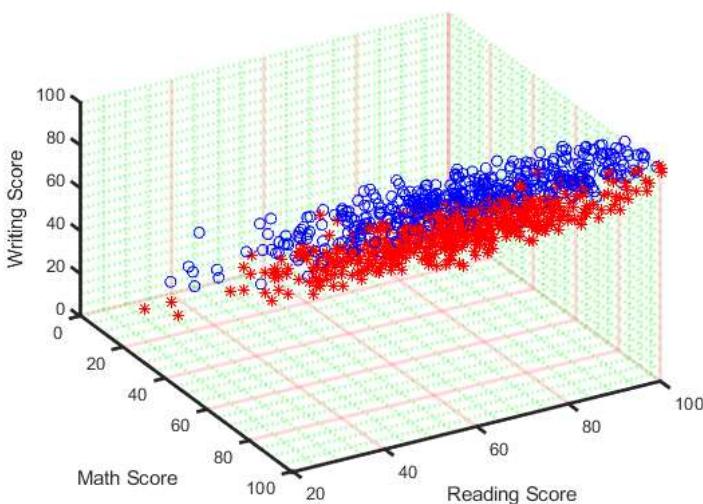


---Scatter 3D ("x" Vs "y" Vs "z")

```

fig06=figure;
hold on
scatter3(valMathScore_F,valReadingScore_F,valWritingScore_F,'bo');
scatter3(valMathScore_M,valReadingScore_M,valWritingScore_M,'r*');
xlabel('Math Score')
ylabel('Reading Score')
zlabel('Writing Score')
view([60 40])
grid minor
set(gca,'MinorGridColor','g',...
    'GridColor','r',...
    'Xgrid','on',...
    'Ygrid','on',...
    'LineWidth',2)
hold off

```



--- Group Scatter (gráfica que automatiza la distribucion "x" VS "y" dados ciertos grupos)

```

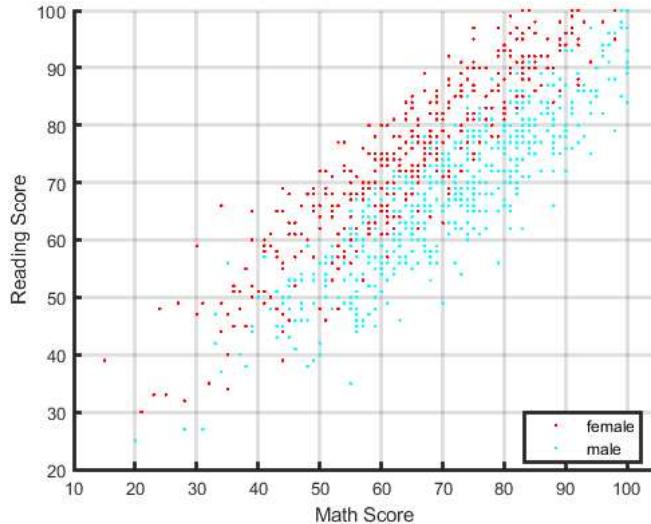
fig07 = figure;
hold on
gscatter(dsStudents_3.mathScore,dsStudents_3.readingScore,dsStudents_3.gender)
xlabel('Math Score')
ylabel('Reading Score')

```

```

grid on
%grid minor
set(gca,'Xgrid','on',...
      'Ygrid','on',...
      'LineWidth',2)
hold off

```

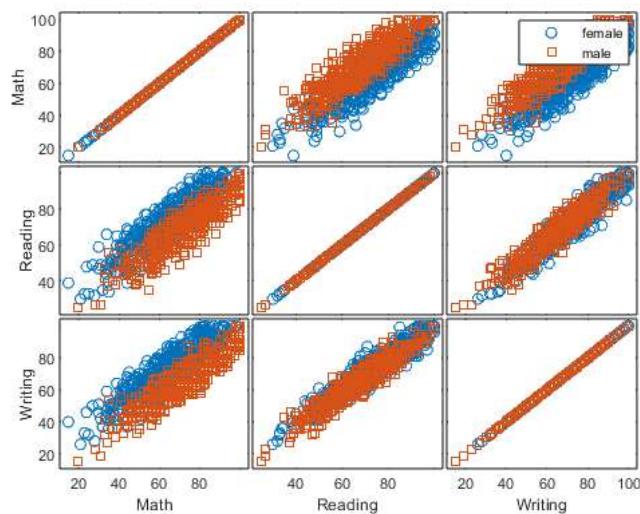


--- Matrix Scatter

```

xvars = [dsStudents_3.mathScore, dsStudents_3.readingScore, dsStudents_3.writingScore];
yvars = [dsStudents_3.mathScore, dsStudents_3.readingScore, dsStudents_3.writingScore];
colors = [0    0.4470 0.7410; ...
          0.8500 0.3250 0.0980; ...
          0.4940 0.1840 0.5560];
grp = dsStudents_3.gender;
labels = {'Math', 'Reading', 'Writing'};
fig08 = figure;
hold on
[h,ax] = gplotmatrix(xvars,yvars,grp,colors,'os*x.');
for i = 1:3
    xlabel(ax(3,i), labels{i})
    ylabel(ax(i,1), labels{i})
end
hold off

```



---Gráfico de cajas

```

fig09 = figure;
subplot(1,3,1)
hold on
axis square;
boxplot(dsStudents_2.mathScore,dsStudents_2.race_ethnicity)
xlabel('Categoria')

```

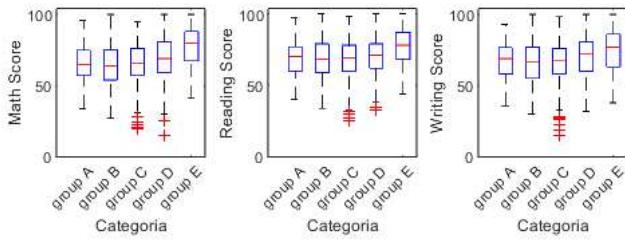
```

ylabel('Math Score')

hold off
subplot(1,3,2)
hold on
axis square;
boxplot(dsStudents_2.readingScore,dsStudents_2.race_ethnicity)
xlabel('Categoria')
ylabel('Reading Score')

hold off
subplot(1,3,3)
hold on
axis square;
boxplot(dsStudents_2.writingScore,dsStudents_2.race_ethnicity)
xlabel('Categoria')
ylabel('Writing Score')
hold off

```



---Gráfico de coordenadas paralelas

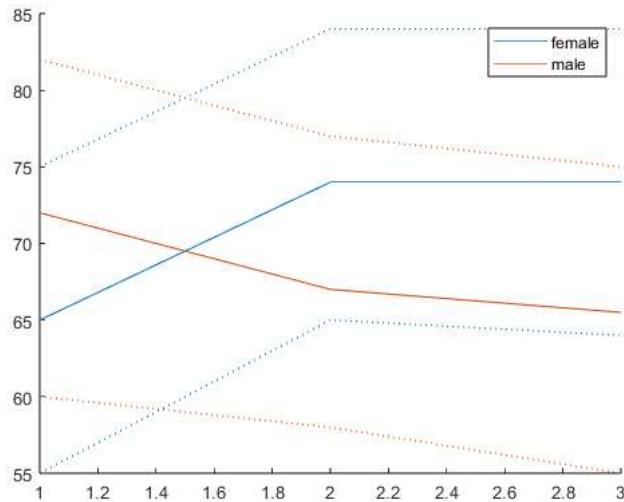
```

%Datos categoricos
gen = dsStudents_2.gender;
race = dsStudents_2.race_ethnicity;

%Datos numéricos
math = dsStudents_2.mathScore;
reading = dsStudents_2.readingScore;
writing = dsStudents_2.writingScore;

%Tabla auxiliar
auxTable = table(gen,math,reading,writing);
%auxGrpTable =
fig10 = figure;
hold on
parallelcoords(xvars,"Group",grp,"Quantile",0.25)
hold off

```



*** Ejercicio ***

Realiza las siguientes operaciones

1. Clasifica los datos por raza y por genero
2. Muestra una figura de histograma basado en esta clasificación: tres gráficas (pruebas de matemáticas, lectura y escritura) con dos categorías del histograma [femenino y masculino]
3. Muestra una figura de histograma basado en esta clasificación: cinco gráficas con tres categorías (pruebas de matemáticas, de lectura y de escritura)
4. Muestra una figura de "Swarm 3D" basado en esta clasificación: tres gráficas con cinco categorías en x (los cinco grupos de razas) seis categorías en y (correspondientes a los grados de educación de los padres) y dependiendo de la prueba realizada los datos de matemáticas, lectura y escritura
5. Muestra en una figura de pay cual es el porcentaje de genero que tuvo un curso de preparacion y cuantos de estos obtuvieron una nota promedio aprobatoria
6. Muestra en una figura de gráficos de caja, cual es la distribucion de aquellos que desayunaron vs aquellos que obtuvieron una nota promedio aprobatoria