

BENEMERITA UNIVERSIDAD AUTONOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACION

INGENERIA EN CIENCIAS DE DATOS

Materia: Introducción a la Ciencias de Datos

Profesor: Jaime Alejandro Romero Sierra

Alumno: José Alejandro Mejía Hernández

GRUPO: 1

PROYECTO 1

En esta práctica se trabaja con un dataset de transacciones bancarias de Kaggle, modificado por el profesor para simular datos sucios.

El objetivo es **limpiar y preparar los datos** para su posterior análisis exploratorio, aplicando las funciones básicas de pandas vistas en clase

## Código:

### EN ESTA IMAGEN MUESRA UNA PARTE DE MI BASE DE DATOS CON MAS DE 4MIL REGISTROS

```
1 home_team,away_team,home_goals,away_goals,result,season
2 Sheffield United,Liverpool,1,0,1,14,D,2006-2007
3 Arsenal,Desconocido,1,0,1,0,D,2006-2007
4 Everton,Watford,2,0,1,0,H,2006-2007
5 Newcastle United,Wigan Athletic,2,0,1,0,H,2006-2007
6 Portsmouth,Blackburn Rovers,3,0,0,0,H,2006-2007
7 Reading,Middlesbrough,3,0,2,0,H,2006-2007
8 West Ham United,Charlton Athletic,3,0,1,0,H,Desconocido
9 Bolton Wanderers,Auto%%,2,0,0,0,H,2006-2007
10 Manchester United,Fulham,5,0,1,0,H,2006-2007
11 Chelsea,Manchester City,3,0,0,0,H,Desconocido
12 Watford,West Ham United,1,0,1,0,D,2006-2007
13 Tottenham Hotspur,Sheffield United,2,0,0,0,H,2006-2007
14 Aston Villa,Reading,2,0,1,0,H,2006-2007
15 Manchester City,Portsmouth,0,0,0,0,D,2006-2007
16 Blackburn Rovers,Everton,1,0,1,0,D,Desconocido
17 Charlton Athletic,Manchester United,0,0,3,0,A,2006-2007
18 Fulham,Bolton Wanderers,1,0,1,0,D,2006-2007
19 Middlesbrough,Chesterfield,2,0,1,0,H,Desconocido
20 Liverpool,West Ham United,2,0,1,0,H,2006-2007
21 Charlton Athletic,Auto%%,2,0,0,0,H,Auto%%
22 Fulham,Sheffield United,1,0,0,0,H,2006-2007
23 Tottenham Hotspur,Everton,0,0,2,0,A,2006-2007
24 Watford,Manchester United,1,0,2,0,A,2006-2007
25 Wigan Athletic,Desconocido,1,0,1,14,H,2006-2007
26 Manchester City,Arsenal,1,0,0,0,H,2006-2007
27 Aston Villa,Newcastle United,2,0,0,0,Desconocido,2006-2007
28 Blackburn Rovers,Chesterfield,0,0,2,0,A,2006-2007
29 Middlesbrough,Portsmouth,0,0,4,0,A,2006-2007
30 Everton,Liverpool,3,0,0,0,H,2006-2007
31 Arsenal,Middlesbrough,1,0,1,0,D,Desconocido
32 Bolton Wanderers,Watford,1,0,0,0,H,2006-2007
```

```
2447 West Ham United,Everton,1,0,2,0,A,2012-2013
2458 Liverpool,Fulham,4,0,0,0,H,2012-2013
2459 Swansea City,Manchester United,1,0,1,0,D,2012-2013
2460 Chelsea,Aston Villa,8,0,0,0,H,2012-2013
2461 Everton,Wigan Athletic,2,0,1,0,H,2012-2013
2462 Fulham,Southampton,1,0,1,0,D,2012-2013
2463 Manchester United,Newcastle United,4,0,3,0,H,2012-2013
2464 Norwich City,Chelsea,0,0,1,0,A,2012-2013
2465 Queens Park Rangers,West Bromwich Albion,1,0,2,0,A,2012-2013
2466 Reading,Swansea City,0,0,0,0,D,2012-2013
2467 Sunderland,Manchester City,1,0,0,0,H,2012-2013
2468 Aston Villa,Tottenham Hotspur,0,0,4,0,A,2012-2013
2469 Stoke City,Liverpool,3,0,1,0,H,2012-2013
2470 Sunderland,Tottenham Hotspur,1,0,2,0,A,2012-2013
2471 Aston Villa,Wigan Athletic,0,0,3,0,A,2012-2013
2472 Fulham,Swansea City,1,0,2,0,A,2012-2013
2473 Manchester United,West Bromwich Albion,2,0,0,0,H,2012-2013
2474 Norwich City,Manchester City,3,0,4,0,A,2012-2013
2475 Reading,West Ham United,1,0,0,0,H,2012-2013
2476 Stoke City,Southampton,3,0,3,0,D,2012-2013
2477 Arsenal,Newcastle United,7,0,3,0,H,2012-2013
2478 Auto%%,Chelsea,1,0,2,0,A,2012-2013
2479 Queens Park Rangers,Liverpool,0,0,3,0,A,2012-2013
2480 Desconocido,Fulham,1,0,2,0,A,2012-2013
2481 Manchester City,Stoke City,3,0,0,0,H,2012-2013
2482 Swansea City,Aston Villa,2,0,2,0,D,2012-2013
2483 Tottenham Hotspur,Reading,3,0,1,0,H,2012-2013
2484 West Ham United,Norwich City,2,0,1,14,H,2012-2013
2485 Wigan Athletic,Manchester United,0,0,4,0,A,2012-2013
2486 Southampton,Arsenal,1,0,1,0,D,2012-2013
2487 Chelsea,Queens Park Rangers,0,0,1,0,A,2012-2013
2488 Desconocido,Sunderland,3,0,0,0,H,2012-2013
```

▲ 4

Lín. 1, col. 1 Espacios: 4 UTF-8 CRLF () Texto sin formato

4930	Newcastle United,Crystal Palace,1.0,0.0,H,2017-2018
4931	West Bromwich Albion,Chelsea,2.0,3.0,A,2015-2016
4932	Watford,Manchester City,1.0,1.0,D,2006-2007
4933	AFC Bournemouth,Stoke City,2.0,2.0,Desconocido,2016-2017
4934	Newcastle United,Fulham,2.0,0.0,H,2007-2008
4935	Newcastle United,Manchester City,1.0,1.0,D,2015-2016
4936	Aston Villa,Birmingham City,0.0,0.0,D,2010-2011
4937	Manchester United,Cardiff City,2.0,0.0,H,2013-2014
4938	Queens Park Rangers,Liverpool,0.0,3.0,A,2012-2013
4939	Liverpool,Swansea City,2.0,3.0,A,2016-2017
4940	Reading,Fulham,0.0,2.0,A,2007-2008
4941	Hull City,Manchester United,0.0,0.0,D,2014-2015
4942	Tottenham Hotspur,Southampton,1.0,2.0,A,2015-2016
4943	Stoke City,Liverpool,1.0,0.0,H,2011-2012
4944	Sheffield United,Manchester United,1.0,2.0,A,Desconocido
4945	Crystal Palace,West Bromwich Albion,0.0,1.0,A,2016-2017
4946	West Bromwich Albion,Wigan Athletic,2.0,3.0,A,Desconocido
4947	West Ham United,Desconocido,1.0,2.0,A,2014-2015
4948	Swansea City,West Bromwich Albion,2.0,1.0,H,2016-2017
4949	Sunderland,Wigan Athletic,1.0,0.0,H,2012-2013
4950	Manchester United,Everton,1.0,1.0,D,2016-2017
4951	Burnley,Bolton Wanderers,1.0,1.0,D,Desconocido
4952	Bolton Wanderers,Reading,1.0,3.0,A,2006-2007
4953	Manchester City,Reading,2.0,1.0,H,2007-2008
4954	Manchester United,Tottenham Hotspur,1.0,0.0,H,2006-2007
4955	Crystal Palace,Brighton and Hove Albion,3.0,2.0,H,Desconocido
4956	Chelsea,Swansea City,1.0,0.0,H,2013-2014
4957	Arsenal,Hull City,2.0,0.0,Desconocido,2016-2017
4958	Newcastle United,Wigan Athletic,3.0,0.0,H,2012-2013
4959	Manchester City,Manchester United,1.0,0.0,H,2014-2015
4960	Manchester City,Swansea City,4.0,0.0,H,2011-2012
4961	Aston Villa,Liverpool,0.0,1.0,A,2013-2014

JAQUI6123 PLANTA BAJA

DURANTE NUESTRA CLASE DE INTRODUCCIÓN ENSUCIAMOS NUESTRA BASE. AQUÍ UN EJEMPLO

USANDO LA LIBRERÍA DE pandas

```
import pandas as pd
df= pd.read_csv("df_sucio.csv")
df
```

[46]

	home_team	away_team	home_goals	away_goals	result	season
0	Sheffield United	Liverpool	1.0	NaN	D	2006-2007
1	Arsenal	NaN	1.0	1.0	D	2006-2007
2	Everton	Watford	2.0	1.0	H	2006-2007
3	Newcastle United	Wigan Athletic	2.0	1.0	H	2006-2007
4	Portsmouth	Blackburn Rovers	3.0	0.0	H	2006-2007
...	...	...	...	...	...	...
4957	Manchester City	Manchester United	1.0	0.0	H	2014-2015
4958	Manchester City	Swansea City	4.0	0.0	H	2011-2012
4959	Aston Villa	Liverpool	0.0	1.0	A	2013-2014
4960	Stoke City	Blackpool	NaN	1.0	A	2010-2011
4961	Newcastle United	Manchester United	2.0	2.0	D	2006-2007

4962 rows × 6 columns

#Veo cantidad de filas y columnas

Spaces: 4 LF ⌂ Celda 27 de 27 ⌂

EL shape, sirve para ver la cantidad de filas y columnas,

En mi base me mostró (4962, 6)

```
▷ ▾ #Ver cantidad de filas y columnas
df.shape
47] Python
.. (4962, 6)

HACER DIAGNOSTICO GENERAL

48] df.columns
Python
.. Index(['home_team', 'away_team', 'home_goals', 'away_goals', 'result',
       'season'],
       dtype='object')

# INFORMACIÓN GENERAL, PARA HACER UN DIAGNOSTICO INICIAL
df.info()
df.describe()
49] Python
.. <class 'pandas.core.frame.DataFrame'>
Spaces: 4 ⚖ Celda 27 de 27 ⌂
```

Aquí ocupamos columns, para hacer un “Diagnóstico general”

# HACER DIAGNOSTICO GENERAL

Generar Código Markdown Agregar celda de código

```
[48] df.columns
Python
.. Index(['home_team', 'away_team', 'home_goals', 'away_goals', 'result',
       'season'],
       dtype='object')
```

```
▷ ▾ # INFORMACIÓN GENERAL, PARA HACER UN DIAGNOSTICO INICIAL
df.info()
df.describe()
49] Python
.. <class 'pandas.core.frame.DataFrame'>
RangeIndex: 4962 entries, 0 to 4961
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
 --- 
 0   home_team   4811 non-null   object 
 1   away_team   4811 non-null   object 
 2   home_goals  4811 non-null   int64  
 3   away_goals  4811 non-null   int64  
 4   result      4811 non-null   object 
 5   season      4811 non-null   int64 
```

Aquí verificamos cuantos NaN tiene

```
home_team      148
away_team      148
home_goals     148
away_goals     148
result         148
season         148
dtype: int64
```

Agregar una celda de Markdown

## VERIFICAMOS CUANTOS NaN TIENE

```
#Cantidad de Null
df.isnull().sum()

[50]
```

Python

```
... home_team      148
    away_team      148
    home_goals     148
    away_goals     148
    result         148
    season         148
    dtype: int64
```

Aquí ocupamos un ciclo for:

```
#Con este ciclo for es posible identificar la cantidad de auto%# y cuales
#son los valores en cada columna
df2=df.copy()
lista_col=df2.columns
for p in lista_col:
    print(f"En la Columna {p} hay: ")
    print(f"Los Auto%# son: {df2[df2[p] == 'Auto%#'].shape[0]} ")
    print(f"Hay {df2[p].nunique()} valores en la columna")
    print(df2[p].unique())
    print(f"_____")
```

En la Columna home\_team hay:  
Los Auto%# son: 97  
Hay 40 valores en la columna  
['Sheffield United' 'Arsenal' 'Everton' 'Newcastle United' 'Portsmouth'  
'Reading' 'West Ham United' 'Bolton Wanderers' 'Manchester United'  
'Chelsea' 'Watford' 'Tottenham Hotspur' 'Aston Villa' 'Manchester City'  
'Blackburn Rovers' 'Charlton Athletic' 'Fulham' 'Middlesbrough'  
'Liverpool' 'Wigan Athletic' 'Auto%' nan 'Sunderland' 'Derby County'  
'Birmingham City' 'Hull City' 'Stoke City' 'West Bromwich Albion'  
'Wolverhampton Wanderers' 'Burnley' 'Blackpool' 'Queens Park Rangers'  
'Swansea City' 'Norwich City' 'Southampton' 'Crystal Palace'  
'Cardiff City' 'Leicester City' 'AFC Bournemouth'  
'Brighton and Hove Albion' 'Huddersfield Town']

```
En la Columna away_team hay:  
Los Auto%# son: 97  
Hay 40 valores en la columna  
['Liverpool' nan 'Watford' 'Wigan Athletic' 'Blackburn Rovers'  
'Middlesbrough' 'Charlton Athletic' 'Auto%' 'Fulham' 'Manchester City'  
'West Ham United' 'Sheffield United' 'Reading' 'Portsmouth' 'Everton'  
'Manchester United' 'Bolton Wanderers' 'Chelsea' 'Arsenal'  
'Newcastle United' 'Tottenham Hotspur' 'Aston Villa' 'Birmingham City'  
'Sunderland' 'Derby County' 'West Bromwich Albion' 'Stoke City'  
'Hull City' 'Burnley' 'Wolverhampton Wanderers' 'Blackpool'  
'Norwich City' 'Swansea City' 'Queens Park Rangers' 'Southampton'  
...  
['2006-2007' nan 'Auto%' '2007-2008' '2008-2009' '2009-2010' '2010-2011'  
'2011-2012' '2012-2013' '2013-2014' '2014-2015' '2015-2016' '2016-2017'  
'2017-2018']
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
for c in lista_col:
    unicos = df[c].nunique()
    print(f'Columna {c} tiene {unicos} valores únicos')
```

Columna home\_team tiene 40 valores únicos  
Columna away\_team tiene 40 valores únicos  
Columna home\_goals tiene 10 valores únicos  
Columna away\_goals tiene 7 valores únicos  
Columna result tiene 3 valores únicos  
Columna season tiene 13 valores únicos

Ocupé un duplicated para mostrar renglones duplicados

## SE CREA UN BOOLEANO PARA MOSTRAR RENGLONES DUPLICADOS

```
df2.duplicated()
```

[53]

Python

```
... 0    False  
1    False  
2    False  
3    False  
4    False  
...  
4957  True  
4958  True  
4959  True  
4960  False  
4961  True  
Length: 4962, dtype: bool
```

Length: 4962, dtype: bool

```
df2.duplicated().sum()
```

[54]

Python

```
... np.int64(264)
```

```
df3=df2.drop_duplicates()  
df3.head()
```

[55]

Python

	home_team	away_team	home_goals	away_goals	result	season
0	Sheffield United	Liverpool	1.0	NaN	D	2006-2007
1	Arsenal	Nan	1.0	1.0	D	2006-2007
2	Everton	Watford	2.0	1.0	H	2006-2007
3	Newcastle United	Wigan Athletic	2.0	1.0	H	2006-2007
4	Portsmouth	Blackburn Rovers	3.0	0.0	H	2006-2007

```
df3.shape
```

[56]

Python

```
... (4698, 6)
```

```
df3.shape  
[6] (4698, 6) Python  
df3.shape  
[7] (4698, 6) Python
```

Generar Código Markdown

Aquí como se muestra remplazamos los valores nulos

## REEMPLAZAMOS VALORES NULOS (COLUMNAS NUMERICAS)

```
column_float = [ 'away_goals' ]  
df3 = df.copy()  
  
# Aplicar los cambios solo al nuevo DataFrame  
for columna in column_float:  
    promedio = df3[col].mean()  
    promedio_redondeado = round(promedio, 2) # (ROUND) Redondea el promedio para mantener consistencia y evitar decimales largos  
    df3[columna] = df3[columna].fillna(promedio_redondeado)  
    print(f'Para columna: {columna} /// Valores nulos reemplazados por: {promedio_redondeado}')  
[58] ... Para columna: away_goals /// Valores nulos reemplazados por: 1.14 Python
```

Ocupé un fillna para borrar mis datos nulos:

## PARA BORRAR DATOS NULOS

```
[59] df3 = df3.fillna("Desconocido")
```

Python

```
[60] df3.head()
```

Python

```
[61] ...
```

	home_team	away_team	home_goals	away_goals	result	season
0	Sheffield United	Liverpool	1.0	1.14	D	2006-2007
1	Arsenal	Desconocido	1.0	1.00	D	2006-2007
2	Everton	Watford	2.0	1.00	H	2006-2007
3	Newcastle United	Wigan Athletic	2.0	1.00	H	2006-2007
4	Portsmouth	Blackburn Rovers	3.0	0.00	H	2006-2007

Aquí verificamos si nuestra base ya no nos maracara errores

## VERIFICAMOS SI NUESTRA BASE YA QUEDO LIMPIA

```
[61] df3.isnull().sum()
```

Python

```
[62] ...
```

```
home_team      0
away_team      0
home_goals     0
away_goals     0
result         0
season         0
dtype: int64
```

```
[62] df3.shape
```

Python

```
[62] ...
```

```
(4962, 6)
```

```
df3.info()
df3.describe()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4962 entries, 0 to 4961
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   home_team    4962 non-null   object  
 1   away_team    4962 non-null   object  
 2   home_goals   4962 non-null   object  
 3   away_goals   4962 non-null   float64 
 4   result       4962 non-null   object  
 5   season       4962 non-null   object  
dtypes: float64(1), object(5)
memory usage: 232.7+ KB

          away_goals
count    4962.000000
mean     1.142225
std      1.119140
min      0.000000
25%     0.000000
50%     1.000000
75%     2.000000
max      6.000000
```

## # CONCLUSIONES

```
# 1.Se reemplazaron valores faltantes con la mediana o "desconocido".
# 2.No se utilizó dropna(), cumpliendo con la consigna del profesor.
# 3.La base final no presenta nulos ni duplicados, y está lista para análisis.

# Aprendimos de nuestros errores y pusimos en práctica lo que vimos en la clase
```