

Reto empleados

1. se comprueba la correcta lectura del csv

In [146...]

```
import pandas as pd

EmpleadosAttrition = pd.read_csv('empleadosRETO.csv')

print(f"data frame: {EmpleadosAttrition}")
```

	Age	BusinessTravel	Department	DistanceFromHome	\
0	50	Travel_Rarely	Research & Development	1 km	
1	36	Travel_Rarely	Research & Development	6 km	
2	21	Travel_Rarely	Sales	7 km	
3	52	Travel_Rarely	Research & Development	7 km	
4	33	Travel_Rarely	Research & Development	15 km	
..
395	33	Travel_Rarely	Research & Development	14 km	
396	31	Travel_Rarely	Sales	20 km	
397	37	Travel_Frequently	Research & Development	11 km	
398	38	Travel_Rarely	Research & Development	4 km	
399	33	Travel_Rarely	Research & Development	14 km	
	Education	EducationField	EmployeeCount	EmployeeNumber	\
0	2	Medical	1	997	
1	2	Medical	1	178	
2	1	Marketing	1	1780	
3	4	Life Sciences	1	1118	
4	1	Medical	1	582	
..
395	3	Medical	1	325	
396	3	Life Sciences	1	175	
397	3	Other	1	306	
398	2	Medical	1	1687	
399	3	Medical	1	252	
	EnvironmentSatisfaction	Gender	...	PercentSalaryHike	\
0	4	Male	...	22	
1	2	Male	...	20	
2	2	Male	...	13	
3	2	Male	...	19	
4	2	Male	...	12	
..
395	3	Male	...	13	
396	2	Female	...	11	
397	2	Male	...	14	
398	4	Female	...	19	
399	4	Female	...	13	
	PerformanceRating	RelationshipSatisfaction	StandardHours		\
0	4		3	80	
1	4		4	80	
2	3		2	80	
3	3		4	80	
4	3		4	80	
..
395	3		3	80	
396	3		3	80	
397	3		3	80	
398	3		4	80	
399	3		4	80	
	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance		\
0	32		1	2	
1	7		0	3	
2	1		3	3	

```
3          18          4          3
4          15          2          4
..
395         ...
396         4          2          1
397         10         1          3
398         7          5          2
399         8          5          3
```

```
YearsInCurrentRole YearsSinceLastPromotion Attrition
0                  4                      1      No
1                  2                      0      No
2                  0                      1      Yes
3                  6                      4      No
4                  6                      7      Yes
..
395         ...
396         2                      2      Yes
397         8                      0      No
398         0                      0      No
399         7                      1      No
```

[400 rows x 30 columns]

2. Eliminacion de columnas que no tienen relacion con la salida.

```
In [147]: EmpleadosAttrition = EmpleadosAttrition.drop(["EmployeeCount", "EmployeeNumber", "Ove
print(f"EmpleadosAttrition")
```

	Age	BusinessTravel	Department	DistanceFromHome	\
0	50	Travel_Rarely	Research & Development	1 km	
1	36	Travel_Rarely	Research & Development	6 km	
2	21	Travel_Rarely	Sales	7 km	
3	52	Travel_Rarely	Research & Development	7 km	
4	33	Travel_Rarely	Research & Development	15 km	
..
395	33	Travel_Rarely	Research & Development	14 km	
396	31	Travel_Rarely	Sales	20 km	
397	37	Travel_Frequently	Research & Development	11 km	
398	38	Travel_Rarely	Research & Development	4 km	
399	33	Travel_Rarely	Research & Development	14 km	

	Education	EducationField	EnvironmentSatisfaction	Gender	\
0	2	Medical	4	Male	
1	2	Medical	2	Male	
2	1	Marketing	2	Male	
3	4	Life Sciences	2	Male	
4	1	Medical	2	Male	
..
395	3	Medical	3	Male	
396	3	Life Sciences	2	Female	
397	3	Other	2	Male	
398	2	Medical	4	Female	
399	3	Medical	4	Female	

	JobInvolvement	JobLevel	...	Overtime	PercentSalaryHike	\
0	3	4	...	No	22	
1	3	2	...	No	20	
2	3	1	...	No	13	
3	3	3	...	No	19	
4	3	3	...	Yes	12	
..
395	3	1	...	Yes	13	
396	1	2	...	Yes	11	
397	3	3	...	Yes	14	
398	3	1	...	No	19	
399	3	1	...	No	13	

	PerformanceRating	RelationshipSatisfaction	TotalWorkingYears	\
0	4	3	32	
1	4	4	7	
2	3	2	1	
3	3	4	18	
4	3	4	15	
..
395	3	3	8	
396	3	3	4	
397	3	3	10	
398	3	4	7	
399	3	4	8	

	TrainingTimesLastYear	WorkLifeBalance	YearsInCurrentRole	\
0	1	2	4	
1	0	3	2	
2	3	3	0	

```
3           4           3           6
4           2           4           6
..
395        ...         ...         ...
396        2           1           4
396        2           3           2
397        1           3           8
398        5           2           0
399        5           3           7
```

```
YearsSinceLastPromotion Attrition
0                      1      No
1                      0      No
2                      1      Yes
3                      4      No
4                      7      Yes
..
395                   ...    ...
396                   0      Yes
396                   2      Yes
397                   0      No
398                   0      No
399                   1      No
```

[400 rows x 26 columns]

3. se obtienen los años que el empleado lleva trabajando en la empresa

```
In [148]: EmpleadosAttrition["Year"] = EmpleadosAttrition["HiringDate"].str.split("/").str[2]

EmpleadosAttrition["YearsAtCompany"] = 2018 - EmpleadosAttrition["Year"]

print(f"EmpleadosAttrition")
```

	Age	BusinessTravel	Department	DistanceFromHome	\
0	50	Travel_Rarely	Research & Development	1 km	
1	36	Travel_Rarely	Research & Development	6 km	
2	21	Travel_Rarely	Sales	7 km	
3	52	Travel_Rarely	Research & Development	7 km	
4	33	Travel_Rarely	Research & Development	15 km	
..
395	33	Travel_Rarely	Research & Development	14 km	
396	31	Travel_Rarely	Sales	20 km	
397	37	Travel_Frequently	Research & Development	11 km	
398	38	Travel_Rarely	Research & Development	4 km	
399	33	Travel_Rarely	Research & Development	14 km	
	Education	EducationField	EnvironmentSatisfaction	Gender	\
0	2	Medical	4	Male	
1	2	Medical	2	Male	
2	1	Marketing	2	Male	
3	4	Life Sciences	2	Male	
4	1	Medical	2	Male	
..
395	3	Medical	3	Male	
396	3	Life Sciences	2	Female	
397	3	Other	2	Male	
398	2	Medical	4	Female	
399	3	Medical	4	Female	
	JobInvolvement	JobLevel	... PerformanceRating		\
0	3	4	...	4	
1	3	2	...	4	
2	3	1	...	3	
3	3	3	...	3	
4	3	3	...	3	
..
395	3	1	...	3	
396	1	2	...	3	
397	3	3	...	3	
398	3	1	...	3	
399	3	1	...	3	
	RelationshipSatisfaction	TotalWorkingYears	TrainingTimesLastYear		\
0	3	32	1		
1	4	7	0		
2	2	1	3		
3	4	18	4		
4	4	15	2		
..
395	3	8	2		
396	3	4	2		
397	3	10	1		
398	4	7	5		
399	4	8	5		
	WorkLifeBalance	YearsInCurrentRole	YearsSinceLastPromotion	Attrition	\
0	2	4	1	No	
1	3	2	0	No	
2	3	0	1	Yes	

3	3	6	4	No
4	4	6	7	Yes
..
395	1	4	0	Yes
396	3	2	2	Yes
397	3	8	0	No
398	2	0	0	No
399	3	7	1	No

	Year	YearsAtCompany
0	2013	5
1	2015	3
2	2017	1
3	2010	8
4	2011	7
..
395	2013	5
396	2016	2
397	2008	10
398	2018	0
399	2010	8

[400 rows x 28 columns]

4. se renombra DistanceFromHome a DistanceFromHome_km y se crea una nueva variable DistanceFromHome de tipo int

```
In [149...]: EmpleadosAttrition = EmpleadosAttrition.rename(columns={"DistanceFromHome": "DistanceFromHome_km"})
EmpleadosAttrition["DistanceFromHome"] = EmpleadosAttrition["DistanceFromHome_km"].astype(int)
print(f"EmpleadosAttrition")
```

	Age	BusinessTravel	Department	DistanceFromHome_km	\
0	50	Travel_Rarely	Research & Development	1 km	
1	36	Travel_Rarely	Research & Development	6 km	
2	21	Travel_Rarely	Sales	7 km	
3	52	Travel_Rarely	Research & Development	7 km	
4	33	Travel_Rarely	Research & Development	15 km	
..
395	33	Travel_Rarely	Research & Development	14 km	
396	31	Travel_Rarely	Sales	20 km	
397	37	Travel_Frequently	Research & Development	11 km	
398	38	Travel_Rarely	Research & Development	4 km	
399	33	Travel_Rarely	Research & Development	14 km	

	Education	EducationField	EnvironmentSatisfaction	Gender	\
0	2	Medical	4	Male	
1	2	Medical	2	Male	
2	1	Marketing	2	Male	
3	4	Life Sciences	2	Male	
4	1	Medical	2	Male	
..
395	3	Medical	3	Male	
396	3	Life Sciences	2	Female	
397	3	Other	2	Male	
398	2	Medical	4	Female	
399	3	Medical	4	Female	

	JobInvolvement	JobLevel	...	RelationshipSatisfaction	\
0	3	4	...	3	
1	3	2	...	4	
2	3	1	...	2	
3	3	3	...	4	
4	3	3	...	4	
..
395	3	1	...	3	
396	1	2	...	3	
397	3	3	...	3	
398	3	1	...	4	
399	3	1	...	4	

	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	\
0	32	1	2	
1	7	0	3	
2	1	3	3	
3	18	4	3	
4	15	2	4	
..
395	8	2	1	
396	4	2	3	
397	10	1	3	
398	7	5	2	
399	8	5	3	

	YearsInCurrentRole	YearsSinceLastPromotion	Attrition	Year	\
0	4	1	No	2013	
1	2	0	No	2015	
2	0	1	Yes	2017	

```
3           6           4       No  2010
4           6           7      Yes  2011
..
395         ...          0      Yes  2013
396         2           2      Yes  2016
397         8           0      No   2008
398         0           0      No   2018
399         7           1      No   2010
```

```
YearsAtCompany  DistanceFromHome
0              5             1
1              3             6
2              1             7
3              8             7
4              7            15
..
395          ...            ...
396          2            20
397          10            11
398          0             4
399          8            14
```

[400 rows x 29 columns]

5. se removeran las siguientes variables: Year, HiringDate y DistanceFromHome_km

```
In [150]: EmpleadosAttrition = EmpleadosAttrition.drop(["Year", "HiringDate", "DistanceFromHome"])
print(f"EmpleadosAttrition")
```

	Age	BusinessTravel	Department	Education	EducationField	\
0	50	Travel_Rarely	Research & Development	2	Medical	
1	36	Travel_Rarely	Research & Development	2	Medical	
2	21	Travel_Rarely	Sales	1	Marketing	
3	52	Travel_Rarely	Research & Development	4	Life Sciences	
4	33	Travel_Rarely	Research & Development	1	Medical	
..	
395	33	Travel_Rarely	Research & Development	3	Medical	
396	31	Travel_Rarely	Sales	3	Life Sciences	
397	37	Travel_Frequently	Research & Development	3	Other	
398	38	Travel_Rarely	Research & Development	2	Medical	
399	33	Travel_Rarely	Research & Development	3	Medical	
	EnvironmentSatisfaction	Gender	JobInvolvement	JobLevel	JobRole	\
0		4 Male	3	4	Research Director	
1		2 Male	3	2	Manufacturing Director	
2		2 Male	3	1	Sales Representative	
3		2 Male	3	3	Healthcare Representative	
4		2 Male	3	3	Manager	
..	
395		3 Male	3	1	Laboratory Technician	
396		2 Female	1	2	Sales Executive	
397		2 Male	3	3	Research Director	
398		4 Female	3	1	Laboratory Technician	
399		4 Female	3	1	Research Scientist	
	JobRole	PerformanceRating	RelationshipSatisfaction	TotalWorkingYears	TrainingTimesLastYear	\
0	Research Director	4	3	32	1	
1	Manufacturing Director	4	4	7	0	
2	Sales Representative	3	2	1	3	
3	Healthcare Representative	3	4	18	4	
4	Manager	3	4	15	2	
..	
395	Laboratory Technician	3	3	8	2	
396	Sales Executive	3	3	4	2	
397	Research Director	3	3	10	1	
398	Laboratory Technician	3	4	7	5	
399	Research Scientist	3	4	8	5	
	WorkLifeBalance	YearsInCurrentRole	YearsSinceLastPromotion	Attrition	YearsAtCompany	\
0	2	4	4	1	No	
1	3	2	2	0	No	
2	3	0	1	1	Yes	

```

3          3          6          4          No
4          4          6          7          Yes
..
395        1          4          0          Yes
396        3          2          2          Yes
397        3          8          0          No
398        2          0          0          No
399        3          7          1          No

```

	YearsAtCompany	DistanceFromHome
0	5	1
1	3	6
2	1	7
3	8	7
4	7	15
..
395	5	14
396	2	20
397	10	11
398	0	4
399	8	14

[400 rows x 26 columns]

6. Creación de la nueva tabla SueldoPromedioDepto

```
In [151...]: SueldoPromedioDepto = EmpleadosAttrition.groupby("Department")["MonthlyIncome"].mean()
SueldoPromedioDepto = SueldoPromedioDepto.rename(columns={"MonthlyIncome": "SueldoPromedio"})
print(SueldoPromedioDepto)
```

	Department	SueldoPromedio
0	Human Resources	6239.888889
1	Research & Development	6804.149813
2	Sales	7188.250000

7. Escalacion de MonthlyIncome para tener valores entre 0 y 1

```
In [152...]: EmpleadosAttrition["MonthlyIncome"] = ((EmpleadosAttrition["MonthlyIncome"] - EmpleadosAttrition["MonthlyIncome"].min()) / (EmpleadosAttrition["MonthlyIncome"].max() - EmpleadosAttrition["MonthlyIncome"].min()))
print(EmpleadosAttrition)
```

	Age	BusinessTravel	Department	Education	EducationField	\
0	50	Travel_Rarely	Research & Development	2	Medical	
1	36	Travel_Rarely	Research & Development	2	Medical	
2	21	Travel_Rarely	Sales	1	Marketing	
3	52	Travel_Rarely	Research & Development	4	Life Sciences	
4	33	Travel_Rarely	Research & Development	1	Medical	
..	
395	33	Travel_Rarely	Research & Development	3	Medical	
396	31	Travel_Rarely	Sales	3	Life Sciences	
397	37	Travel_Frequently	Research & Development	3	Other	
398	38	Travel_Rarely	Research & Development	2	Medical	
399	33	Travel_Rarely	Research & Development	3	Medical	

	EnvironmentSatisfaction	Gender	JobInvolvement	JobLevel	\
0	4	Male	3	4	
1	2	Male	3	2	
2	2	Male	3	1	
3	2	Male	3	3	
4	2	Male	3	3	
..	
395	3	Male	3	1	
396	2	Female	1	2	
397	2	Male	3	3	
398	4	Female	3	1	
399	4	Female	3	1	

	JobRole	...	PerformanceRating	\
0	Research Director	...	4	
1	Manufacturing Director	...	4	
2	Sales Representative	...	3	
3	Healthcare Representative	...	3	
4	Manager	...	3	
..	
395	Laboratory Technician	...	3	
396	Sales Executive	...	3	
397	Research Director	...	3	
398	Laboratory Technician	...	3	
399	Research Scientist	...	3	

	RelationshipSatisfaction	TotalWorkingYears	TrainingTimesLastYear	\
0	3	32	1	
1	4	7	0	
2	2	1	3	
3	4	18	4	
4	4	15	2	
..	
395	3	8	2	
396	3	4	2	
397	3	10	1	
398	4	7	5	
399	4	8	5	

	WorkLifeBalance	YearsInCurrentRole	YearsSinceLastPromotion	Attrition	\
0	2	4	1	No	
1	3	2	0	No	
2	3	0	1	Yes	

```

3          3          6          4      No
4          4          6          7     Yes
..
395        1          4          0     Yes
396        3          2          2     Yes
397        3          8          0     No
398        2          0          0     No
399        3          7          1     No

   YearsAtCompany  DistanceFromHome
0              5                  1
1              3                  6
2              1                  7
3              8                  7
4              7                 15
..
395        ...                14
396        2                 20
397        10                11
398        0                  4
399        8                 14

[400 rows x 26 columns]

```

8. Modificar variables categoricas a numeros

```
In [153...]: EmpleadosAttrition["BusinessTravel"] = EmpleadosAttrition["BusinessTravel"].astype("category")
EmpleadosAttrition["Department"] = EmpleadosAttrition["Department"].astype("category")
EmpleadosAttrition["EducationField"] = EmpleadosAttrition["EducationField"].astype("category")
EmpleadosAttrition["Gender"] = EmpleadosAttrition["Gender"].astype("category").cat.add_categories(["Other"], inplace=True)
EmpleadosAttrition["JobRole"] = EmpleadosAttrition["JobRole"].astype("category").cat.add_categories(["Manager"], inplace=True)
EmpleadosAttrition["MaritalStatus"] = EmpleadosAttrition["MaritalStatus"].astype("category")
EmpleadosAttrition["Attrition"] = EmpleadosAttrition["Attrition"].astype("category")

print(EmpleadosAttrition["Attrition"])

0      0
1      0
2      1
3      0
4      1
..
395    1
396    1
397    0
398    0
399    0
Name: Attrition, Length: 400, dtype: int8
```

9. Creacion de correlacion con Attrition

```
In [154...]: columns_evaluar = [
    "Attrition",
    "BusinessTravel",
```

```

    "Department",
    "EducationField",
    "Gender",
    "JobRole",
    "MaritalStatus"
]

EvaluacionCorr = EmpleadosAttrition[columns_evaluar]

Correlacion = EvaluacionCorr.corr()["Attrition"].sort_values(ascending=False)

columns_finales = Correlacion[Correlacion.abs() >= 0.1].index

EmpleadosAttritionFinal = EmpleadosAttrition[columns_finales]
print(EmpleadosAttritionFinal)

```

	Attrition	MaritalStatus
0	0	0
1	0	0
2	1	2
3	0	2
4	1	1
..
395	1	1
396	1	1
397	0	0
398	0	1
399	0	1

[400 rows x 2 columns]

11. Creacion de dataframe EmpleadosAttritionPCA

In [155...]

```

from sklearn.decomposition import PCA
import numpy as np

X = EmpleadosAttritionFinal.drop("Attrition", axis=1)

pca = PCA()
EmpleadosAttritionPCA = pca.fit_transform(X)
print(EmpleadosAttritionPCA[:,0])

```

12. Minimo numero de componentes principales de columnas

```
In [156...]: varianza_acumulada = np.cumsum(pca.explained_variance_ratio_)
print(varianza_acumulada)
numero_componentes = np.argmax(varianza_acumulada >= 0.80) + 1
print(f"numero de componentes: {numero_componentes}")
for i in range(numero_componentes):
    EmpleadosAttritionFinal = EmpleadosAttritionFinal.assign(**{f"C{i}": EmpleadosA
```

[1.]
numero de componentes: 1

13. Creación de archivo csv

```
In [157...]: EmpleadosAttritionFinal.to_csv("EmpleadosAttritionFinal.csv", index=False)
```