Filas: 1614

Columnas: 41

Empresa: 1980

EEUU

IA y machine learning

1. **Age**: La edad del empleado.

- cambiamos

- no hay nulos

- tipo de dato: object (str) à cambiar a int

- Count: 1614

- Unique: 54

- Top: 35

- Freq: 84

- hay números escritos con palabras 🡪 los cambiamos a números (thirty-two, fifty-eight, twenty-six, thirty-seven, thirty-one, thirty, fifty-two, fifty-five, thirty-six, forty-seven, twenty-four)

- cambiado tipo de dato a int64

- media: 36.9206

- std: 9.10

- min: 18

- max: 60

1. **Attrition**: Indica si el empleado ha dejado la empresa ("No" significa que no ha dejado la empresa y "Yes" significa que ha dejado la empresa).

- no hay nulos

- tipo de dato: object (str)

- Count: 1614

- Unique: 2

- Top: NO

- Freq: 1355

- No 🡪 1355

- Yes 🡪 259

3. **BusinessTravel**: Describe la frecuencia de los viajes relacionados con el trabajo del empleado (por ejemplo, "Travel\_Rarely" para raramente).

- hay NaN

- 772 nulos

- tipo de dato: object (str)

- Count: 842

- Unique: 3

- Top: Travel rarely

- Freq: 586

- ‘travel\_rarely’: 586

- ‘travel\_frecuently’: 165

- ‘non-travel’: 91

- cambiamos ‘non-travel’ por ‘non\_travel’

- cambiamos los Nan por ‘desconocido’

4. **DailyRate**: La tarifa diaria del empleado.

- sacamos $ para convertirlo a float

- 124 nulos

- tipo de dato: object (str) à float

- Count: 1614

- Unique: 849

- Top: nan$

- Freq: 124

- ‘nan’ a np.nan para pasar la columna a float64

5. **Department**: El departamento en el que trabaja el empleado (por ejemplo, "Research & Development", "Sales", etc.).

- 1312 nulos

- object

- Count: 302

- Unique: 3

- Top: Research & Development

- Freq: 196

- Research & Development: 196

- Sales: 91

- Human Resources: 15

- cambiamos:

- ‘Research & Development’ por ‘research\_development’

- ‘Sales’ por ‘sales’

- Human Resources por ‘human\_resources’

- nulos a ‘desconocido’

6. **DistanceFromHome**: La distancia desde el hogar del empleado hasta su lugar de trabajo.

- Datos negativos – hay que eliminar los valores negativos – pasarlo a nulo

- no nulos

- Int64

- media: 4.5273

- std: 14.59

- min: -49

- max: 29

- cambiamos negativos a positivos

- media: 11.1419

- std: 10.45

- min: 1

- max: 49

- negativos a nulos: 178

7. **Education**: Nivel de educación del empleado (generalmente en una escala del 1 al 5).

- no nulos

- int64

- media: 2.9250

- std: 1.02

- min: 1.0

- max: 5.

- count:

- 1: 180 eso

- 2: 314 fp

- 3: 621 bachillerato/college

- 4: 445 university degree

- 5: 54 phd

8. **EducationField**: El campo de educación del empleado.

- Nan

- 745 nulos

- object

- Count: 869

- Unique: 6

- Top: Life Sciences

- Freq: 349

- counts:

- Life sciences: 349

- Medical: 276

- Maketing: 104

- Technical degree: 69

- Other: 59

- Human Resources: 12

- cambiamos a minúsculas y un \_ en los espacios de los valores

9. **EmployeeCount**: Un contador que generalmente es 1 y se usa para contar empleados.

- no nulos

- int64

- media: 1.0000

- std: 0.0000

- min: 1.0

- max: 1.

- eliminamos columna

10. **EmployeeNumber**: Un número de identificación único para el empleado.

- transformar a inbt? Completar vacíos (NaN)

- añadir números autoincrementales en nulos (id) y eliminar esta o no tenerla en cuenta

- 431 nulos

- object

- Count: 1183

- Unique; 1079

- Top: 482

- Freq: 2

11. **EnvironmentSatisfaction**: Nivel de satisfacción del empleado en relación con su entorno de trabajo. Con valores que están comprendidos entre el 1 y el 4, siendo el 4 el nivel de máxima satisfacción.

- no nulos

- int64

- media: 4.2949

- std: 6.99

- min: 1.00

- max: 49.

- máximo debería ser 4, gestionar 🡪 cambiamos los elementos de 2 dígitos por el primer digito, ya

12. **Gender**: El género del empleado. Donde 0 corresponde con "hombre" y 1 con "mujer".

- cambiar a string 🡪 0 por ‘hombre’ y 1 por 1 por ‘mujer

- no nulos

- int64

- min: 0

- max: 1

- 0: 971 🡪 cambiado a male

- 1: 643 🡪 cambiado a female

13. **HourlyRate**: La tarifa por hora del empleado.

- no nulos

- object à cambia a int?

- Count: 1614

- Unique; 72

- Top: Not available

- Freq: 84

Not Available: 84

42: 33

66: 32

48: 30

57: 29

84: 29

46: 28

98: 28

54: 28

87: 27

92: 27

96: 27

72: 27

79: 26

32: 26

56: 26

81: 25

61: 25

52: 25

86: 25

51: 24

82: 24

62: 24

83: 24

94: 23

74: 23

41: 23

45: 23

43: 23

88: 23

77: 23

73: 23

78: 22

95: 22

60: 22

99: 21

80: 21

76: 21

59: 21

67: 21

97: 20

49: 20

44: 20

85: 20

75: 20

90: 20

91: 20

65: 19

64: 19

71: 19

100: 19

37: 19

55: 19

39: 19

58: 18

40: 18

36: 18

93: 17

89: 17

70: 17

35: 17

30: 17

33: 17

31: 16

63: 16

69: 16

50: 15

47: 15

53: 13

68: 13

38: 12

34: 11

- not available a np.nan y type a numeric pasa

- falta gestinonar nulos

14. **JobInvolvement**: Nivel de implicación del empleado en su trabajo.

- no nulos

- int64

- media: 2.7397

- std: 0.71

- min: 1

- max: 4

- valoraciones del 0 al 4 🡪 correct

3 - 955

2 - 406

4 - 164

1 - 89

15. **JobLevel**: Nivel jerárquico del empleado en la empresa. Del mínimo 1 al máximo 5

- no nulos

- int64

- media: 2.0681

- std: 1.10

- min: 1

- max: 5

2 – 597

1 - 586

3 - 242

4 – 113

5 – 76

16. **JobRole**: El rol o puesto de trabajo del empleado.

- tendremos que hacer un lower(), cada elemento escrito diferente y después el value\_counts

- no nulos

- object

- Count: 1614

- Unique:1579

- Top: Manager

- Freq: 4

sales executive - 369 research scientist - 314 laboratory technician - 278 manufacturing director - 158 healthcare representative - 149 manager - 111 sales representative - 90 research director - 88 human resources - 57

- lower, strip y \_ entre palabras

17. **JobSatisfaction**: Nivel de satisfacción del empleado con su trabajo.

- no nulos

- int64

- media: 2.7385

- std: 1.10

- min: 1

- max: 4

- valoraciones del 0 al 4 🡪 correcto

4 - 514

3 - 481

1 - 317

2 - 302

18. **MaritalStatus**: El estado civil del empleado (por ejemplo, "Single", "Married", etc.).

- NaN

- 651 nulos

- object

- Count: 963

- Unique: 5

- Top: Married

- Freq: 404

Married - 404

Single - 325

Divorced - 188

Marreid - 35

Divorced – 11

- gestión nulos - desconocido

19. **MonthlyIncome**: Ingresos mensuales del empleado.

- 843 nulos

- object a int

- Count: 771

- Unique: 668

- Top: 6347,0

- Freq: 4

6347,0 - 4

5304,0 - 4

2657,0 - 3

2258,0 - 3

2070,0 - 2

..

4556,0 - 1

4230,0 - 1

4859,0 - 1

6322,0 - 1

7639,0 – 1

20. **MonthlyRate**: Tasa mensual del empleado.

- no nulos

- int64

- media: 14284.4956

- std: 7110.41

- min: 2094.0

- max: 26999.0

9150 - 4

6069 - 3

11737 - 3

4156 - 3

25326 - 3

..

19982 - 1

12449 - 1

5829 - 1

17940 - 1

16130 1

21. **NumCompaniesWorked**: Número de compañías en las que el empleado ha trabajado.(0-9)

- no nulos

- int64

- cambiar nombre de columna minúsculas

- media: 2.6734

- std: 2.50

- min: 0

- max: 9

1 - 573

0 - 226

3 - 169

4 - 157

2 - 156

7 - 84

6 - 73

5 - 66

9 - 59

8 - 51

22. **Over18**: Indica si el empleado es mayor de 18 año (eliminamos, se puede

- (Y)

- Nan

- 901 nulos

- object

- Count: 713

- Unique: 1

- Top: Y

- Freq: 713

- Y 713

23. **OverTime**: Indica si el empleado trabaja horas extra ("Yes" para sí o "No" para no).

- 676 nulos -> cambiamos por desconocido porque son muchos nulos

- object

- Count: 938

- Unique: 2

- Top: no

- Freq: 682

No - 682

Yes - 256

24. **PercentSalaryHike**: El porcentaje de aumento salarial del empleado.

- no nulos

- int64

- media: 15.1654

- std: 3.64

- min: 11

- max: 25

11 232

13 230

12 225

14 220

15 110

18 98

17 88

16 86

19 82

20 60

22 59

21 51

23 29

24 25

25 19

25. **PerformanceRating**: Calificación de rendimiento del empleado.

- decidir puntuación todas en int o todas en float

- 195 nulos – cambiamos por la moda

- object

- Count: 1419

- Unique: 2

- Top: 3,0

- Freq: 1205

3,0 - 1205

4,0 - 214

26. **RelationshipSatisfaction**: Nivel de satisfacción en las relaciones interpersonales del empleado.

- no nulos

- int64

- media: 2.7044

- std: 1.07

- min: 1

- max: 4

- valoraciones del 0 al 4 🡪 correcto

3 - 504

4 - 468

2 - 339

1 - 303

27. **StandardHours**: Las horas estándar de trabajo.( la eliminamos porque noj es útil si todos los datos son iguales, no aporta información)

- 1195 nulos

- object

- Count: 419

- Unique: 1

- Top: 80,0

- Freq: 419

80,0 - 419

28. **StockOptionLevel**: Nivel de opciones de compra de acciones del empleado.

- no nulos

- int64

- media: 2.7044

- std: 1.07

- min: 1

- max: 3

0 - 687

1 - 666

2 - 172

3 – 89

29. **TotalWorkingYears**: Total de años de experiencia laboral del empleado.

- 526 nulos

- object -- float

- Count: 1088

- Unique: 40

- Top:10,0

- Freq:144

10,0 - 144

8,0 - 86

6,0 - 84

9,0 - 69

5,0 - 66

7,0 - 56

4,0 - 54

1,0 - 53

12,0 - 34

3,0 - 32

13,0 - 30

14,0 - 30

11,0 - 29

16,0 - 28

15,0 - 28

20,0 - 28

18,0 - 27

21,0 - 23

17,0 22

2,0 21

22,0 18

19,0 17

24,0 14

28,0 13

23,0 13

0,0 8

26,0 8

33,0 6

36,0 6

29,0 6

25,0 6

27,0 5

37,0 5

31,0 4

32,0 3

40,0 3

30,0 3

35,0 3

34,0 2

38,0 1

30. **TrainingTimesLastYear**: Número de veces que el empleado recibió capacitación el año pasado.

- no nulos

- int 64

- media: 2.8097

- std: 1.29

- min: 0

- max: 6

2 598

3 534

4 137

5 136

1 77

6 72

0 60

31. **WorkLifeBalance**: Equilibrio entre trabajo y vida personal del empleado.

- 108 nulos – por al moda

- object -- float

- nombre columna todo en mayúscula

- Count:1506

- Unique: 4

- Top: 3,0

- Freq: 913

3,0 913

2,0 359

4,0 155

1,0 79

32. **YearsAtCompany**: Años que el empleado ha trabajado en la empresa actual.

- no nulos

- int64

- media: 7.1325

- std: 6.12

- min: 0

- max: 40

5 208

1 171

2 141

3 141

10 133

7 115

4 114

8 106

9 94

6 78

0 44

11 36

20 29

13 26

15 21

14 19

22 17

18 15

12 15

21 14

16 14

19 12

17 9

24 7

33 5

25 5

26 4

31 3

32 3

27 3

36 3

23 2

40 2

29 2

37 1

30 1

34 1

33. **YearsInCurrentRole**: Años que el empleado ha estado en su puesto actual. (34 datos no son relevantes, los borramos)

- 1580 nulos

- object

- Count: 34

- Unique: 10

- Top: 2,0

- Freq: 11

2,0 11

7,0 5

0,0 4

4,0 3

1,0 3

11,0 2

6,0 2

3,0 2

12,0 1

13,0 1

34. **YearsSinceLastPromotion**: Años desde la última promoción del empleado.

- no nulos

- int64

- media: 2.2459

- std: 3.23

- min: 0

- max: 15

0 625

1 384

2 177

7 93

4 67

3 62

5 53

6 37

11 26

8 20

9 18

15 15

12 11

14 10

13 10

10 6

35. **YearsWithCurrManager**: Años que el empleado ha estado bajo la supervisión del actual gerente.

- no nulos

- int64

- media: 4.2205

- std: 3.56

- min: 0

- max: 17

2 380

0 270

7 267

3 148

8 115

4 104

1 84

9 70

5 36

10 31

6 30

11 22

12 20

13 16

17 8

15 5

14 5

16 3

36. **SameAsMonthlyIncome**: Ingresos mensuales del empleado.

- comprobar para errores y eliminar. Eliminada, es igual que el monthly income

- 843 nulos

- object

- Count: 771

- Unique: 668

- Top: 6347,0

- Freq: 4

6347,0 4

5304,0 4

2657,0 3

2258,0 3

2070,0 2

..

4556,0 1

4230,0 1

4859,0 1

6322,0 1

7639,0 1

37. **DateBirth**: Año de nacimiento del empleado (teniendo en cuenta que los datos fueron recogidos en el 2023)

- no nulos

- int64 🡪 cambiar a datetime?

- media: 1986.0762

- std: 9.10

- min: 1963

- max: 2005

date\_birth

1988 84

1989 83

1992 83

1994 78

1987 75

1991 65

1993 64

1985 64

1990 61

1983 60

1986 55

1996 54

1995 53

1981 50

1978 48

1982 46

1997 46

1984 45

1980 40

1977 38

1979 34

1973 31

1999 27

1976 27

1998 27

1974 25

1968 25

1972 20

1969 20

1970 20

1975 20

1971 19

1967 17

2002 16

2001 16

2000 15

1965 15

2003 11

1964 10

2004 9

2005 8

1963 5

1966 5

38. **Salary**: Salario de los empleados.

- todos son iguales, calcularlo a partir del mensual, diario, ….(no aporta info, eliminados

- no nulos

- object

- Count: 1614

- Unique: 1

- Top: 1000000000$

- Freq: 1614

1000000000$ 1614

39. **RoleDepartament**: El departamento y el rol del empleado.

- usar el rol para completar errores del JobRole y eliminar (dejar solo departamento). Eliminamos porque tenemos los mismos datos en department y job\_role

- hacer value counts despues de un lower

- 1312 nulos

- object

- Count: 302

- Unique: 301

- Top: Manager sales

- Freq: 2

40. **NumberChildren**: Número de hijos de los empleados

- todo nulos (eliminar columna)

- float64

41. **RemoteWork**: Si el empleado puede teletrabajar o no.

- False, 0 ,1, … decidir tipo de dato o a no, 1 a yes, true a yes

- no nulos

- object

- Count: 1614

- Unique: 5

- Top: 1

- Freq: 360

1 360

True 345

0 309

False 305

Yes 295

Procedimiento:

- Estudiar dataframe: nulos, tipos de dato x columna, datos atípicos que llamen la atención

- cambiamos nombres de columna a minúsculas y con \_ entre palabras

Pasar a int:

* Employee\_number
* Hourly\_rate
* Total\_working\_years
* work\_life\_balance