DATA-RRHH

April 9, 2025

[5]:
```
#Importamos las librerías que usaremos
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

#Leemos el archivo CSV
df = pd.read_csv('./HRDataset_v14.csv')
#Mostramos el dataframe
df.head(5)
```

[5]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	…
0	Adinolfi, Wilson K	10026	0	0	1	
1	Ait Sidi, Karthikeyan	10084	1	1	1	
2	Akinkuolie, Sarah	10196	1	1	0	
3	Alagbe,Trina	10088	1	1	0	
4	Anderson, Carol	10069	0	2	0	

	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	Salary	…
0	1	5	4	0	62506	…
1	5	3	3	0	104437	…
2	5	5	3	0	64955	…
3	1	5	3	0	64991	…
4	5	5	3	0	50825	…

	ManagerName	ManagerID	RecruitmentSource	PerformanceScore
0	Michael Albert	22.0	LinkedIn	Exceeds
1	Simon Roup	4.0	Indeed	Fully Meets
2	Kissy Sullivan	20.0	LinkedIn	Fully Meets
3	Elijiah Gray	16.0	Indeed	Fully Meets
4	Webster Butler	39.0	Google Search	Fully Meets

	EngagementSurvey	EmpSatisfaction	SpecialProjectsCount
0	4.60	5	0
1	4.96	3	6
2	3.02	3	0
3	4.84	5	0
4	5.00	4	0

```
  LastPerformanceReview_Date DaysLateLast30 Absences
0                   1/17/2019              0        1
1                   2/24/2016              0       17
2                   5/15/2012              0        3
3                    1/3/2019              0       15
4                    2/1/2016              0        2

[5 rows x 36 columns]
```

[6]:
```
#Mostramos la información del dataframe
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 311 entries, 0 to 310
Data columns (total 36 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Employee_Name               311 non-null    object
 1   EmpID                       311 non-null    int64
 2   MarriedID                   311 non-null    int64
 3   MaritalStatusID             311 non-null    int64
 4   GenderID                    311 non-null    int64
 5   EmpStatusID                 311 non-null    int64
 6   DeptID                      311 non-null    int64
 7   PerfScoreID                 311 non-null    int64
 8   FromDiversityJobFairID      311 non-null    int64
 9   Salary                      311 non-null    int64
 10  Termd                       311 non-null    int64
 11  PositionID                  311 non-null    int64
 12  Position                    311 non-null    object
 13  State                       311 non-null    object
 14  Zip                         311 non-null    int64
 15  DOB                         311 non-null    object
 16  Sex                         311 non-null    object
 17  MaritalDesc                 311 non-null    object
 18  CitizenDesc                 311 non-null    object
 19  HispanicLatino              311 non-null    object
 20  RaceDesc                    311 non-null    object
 21  DateofHire                  311 non-null    object
 22  DateofTermination           104 non-null    object
 23  TermReason                  311 non-null    object
 24  EmploymentStatus            311 non-null    object
 25  Department                  311 non-null    object
 26  ManagerName                 311 non-null    object
 27  ManagerID                   303 non-null    float64
 28  RecruitmentSource           311 non-null    object
 29  PerformanceScore            311 non-null    object
 30  EngagementSurvey            311 non-null    float64
```

```
 31  EmpSatisfaction             311 non-null    int64
 32  SpecialProjectsCount        311 non-null    int64
 33  LastPerformanceReview_Date  311 non-null    object
 34  DaysLateLast30              311 non-null    int64
 35  Absences                    311 non-null    int64
dtypes: float64(2), int64(16), object(18)
memory usage: 87.6+ KB
```

[7]:
```
df.describe()
```

[7]:

	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID
count	311.000000	311.000000	311.000000	311.000000	311.000000
mean	10156.000000	0.398714	0.810289	0.434084	2.392283
std	89.922189	0.490423	0.943239	0.496435	1.794383
min	10001.000000	0.000000	0.000000	0.000000	1.000000
25%	10078.500000	0.000000	0.000000	0.000000	1.000000
50%	10156.000000	0.000000	1.000000	0.000000	1.000000
75%	10233.500000	1.000000	1.000000	1.000000	5.000000
max	10311.000000	1.000000	4.000000	1.000000	5.000000

	DeptID	PerfScoreID	FromDiversityJobFairID	Salary
count	311.000000	311.000000	311.000000	311.000000
mean	4.610932	2.977492	0.093248	69020.684887
std	1.083487	0.587072	0.291248	25156.636930
min	1.000000	1.000000	0.000000	45046.000000
25%	5.000000	3.000000	0.000000	55501.500000
50%	5.000000	3.000000	0.000000	62810.000000
75%	5.000000	3.000000	0.000000	72036.000000
max	6.000000	4.000000	1.000000	250000.000000

	Termd	PositionID	Zip	ManagerID	EngagementSurvey
count	311.000000	311.000000	311.000000	303.000000	311.000000
mean	0.334405	16.845659	6555.482315	14.570957	4.110000
std	0.472542	6.223419	16908.396884	8.078306	0.789938
min	0.000000	1.000000	1013.000000	1.000000	1.120000
25%	0.000000	18.000000	1901.500000	10.000000	3.690000
50%	0.000000	19.000000	2132.000000	15.000000	4.280000
75%	1.000000	20.000000	2355.000000	19.000000	4.700000
max	1.000000	30.000000	98052.000000	39.000000	5.000000

	EmpSatisfaction	SpecialProjectsCount	DaysLateLast30	Absences
count	311.000000	311.000000	311.000000	311.000000
mean	3.890675	1.218650	0.414791	10.237942
std	0.909241	2.349421	1.294519	5.852596
min	1.000000	0.000000	0.000000	1.000000
25%	3.000000	0.000000	0.000000	5.000000
50%	4.000000	0.000000	0.000000	10.000000

```
75%          5.000000              0.000000        0.000000    15.000000
max          5.000000              8.000000        6.000000    20.000000
```

[8]:
```
#Verificamos si hay valores nulos
df.isnull().sum()
```

[8]:
```
Employee_Name                0
EmpID                        0
MarriedID                    0
MaritalStatusID              0
GenderID                     0
EmpStatusID                  0
DeptID                       0
PerfScoreID                  0
FromDiversityJobFairID       0
Salary                       0
Termd                        0
PositionID                   0
Position                     0
State                        0
Zip                          0
DOB                          0
Sex                          0
MaritalDesc                  0
CitizenDesc                  0
HispanicLatino               0
RaceDesc                     0
DateofHire                   0
DateofTermination          207
TermReason                   0
EmploymentStatus             0
Department                   0
ManagerName                  0
ManagerID                    8
RecruitmentSource            0
PerformanceScore             0
EngagementSurvey             0
EmpSatisfaction              0
SpecialProjectsCount         0
LastPerformanceReview_Date   0
DaysLateLast30               0
Absences                     0
dtype: int64
```

[9]:
```
#Verificamos si hay duplicados
df.duplicated().sum()
```

[9]:
```
np.int64(0)
```

[10]:
```
#Eliminamos las columnas que no usaremos
columnas_a_eliminar = [
    'MarriedID',
    'MaritalStatusID',
    'GenderID',
    'PerfScoreID',
    'EmpStatusID',
    'DeptID',
    'PositionID',
    'Termd',
    'FromDiversityJobFairID',
    'ManagerID',
    'Zip'
]

# Eliminar las columnas
df = df.drop(columns=columnas_a_eliminar)
df
```

[10]:

	Employee_Name	EmpID	Salary	Position	State	…
0	Adinolfi, Wilson K	10026	62506	Production Technician I	MA	
1	Ait Sidi, Karthikeyan	10084	104437	Sr. DBA	MA	
2	Akinkuolie, Sarah	10196	64955	Production Technician II	MA	
3	Alagbe,Trina	10088	64991	Production Technician I	MA	
4	Anderson, Carol	10069	50825	Production Technician I	MA	
..	…	…	…	…	…	
306	Woodson, Jason	10135	65893	Production Technician II	MA	
307	Ybarra, Catherine	10301	48513	Production Technician I	MA	
308	Zamora, Jennifer	10010	220450	CIO	MA	
309	Zhou, Julia	10043	89292	Data Analyst	MA	
310	Zima, Colleen	10271	45046	Production Technician I	MA	

	DOB	Sex	MaritalDesc	CitizenDesc	HispanicLatino	…	\
0	07/10/83	M	Single	US Citizen	No	…	
1	05/05/75	M	Married	US Citizen	No	…	
2	09/19/88	F	Married	US Citizen	No	…	
3	09/27/88	F	Married	US Citizen	No	…	
4	09/08/89	F	Divorced	US Citizen	No	…	
..	…	..	…	…	…	…	
306	05/11/85	M	Single	US Citizen	No	…	
307	05/04/82	F	Single	US Citizen	No	…	
308	08/30/79	F	Single	US Citizen	No	…	
309	02/24/79	F	Single	US Citizen	No	…	
310	08/17/78	F	Widowed	US Citizen	No	…	

```
            Department     ManagerName  RecruitmentSource PerformanceScore  \
0    Production         Michael Albert           LinkedIn          Exceeds
1               IT/IS      Simon Roup             Indeed      Fully Meets
2    Production         Kissy Sullivan           LinkedIn      Fully Meets
3    Production           Elijiah Gray             Indeed      Fully Meets
4    Production         Webster Butler      Google Search      Fully Meets
..                 …              …                 …              …
306  Production         Kissy Sullivan           LinkedIn      Fully Meets
307  Production         Brannon Miller      Google Search              PIP
308             IT/IS      Janet King  Employee Referral          Exceeds
309             IT/IS      Simon Roup  Employee Referral      Fully Meets
310  Production          David Stanley           LinkedIn      Fully Meets

    EngagementSurvey EmpSatisfaction SpecialProjectsCount  \
0               4.60               5                    0
1               4.96               3                    6
2               3.02               3                    0
3               4.84               5                    0
4               5.00               4                    0
..               …              …                 …
306             4.07               4                    0
307             3.20               2                    0
308             4.60               5                    6
309             5.00               3                    5
310             4.50               5                    0

    LastPerformanceReview_Date DaysLateLast30  Absences
0                    1/17/2019              0         1
1                    2/24/2016              0        17
2                    5/15/2012              0         3
3                     1/3/2019              0        15
4                     2/1/2016              0         2
..                         …            …       …
306                  2/28/2019              0        13
307                   9/2/2015              5         4
308                  2/21/2019              0        16
309                   2/1/2019              0        11
310                  1/30/2019              0         2

[311 rows x 25 columns]
```

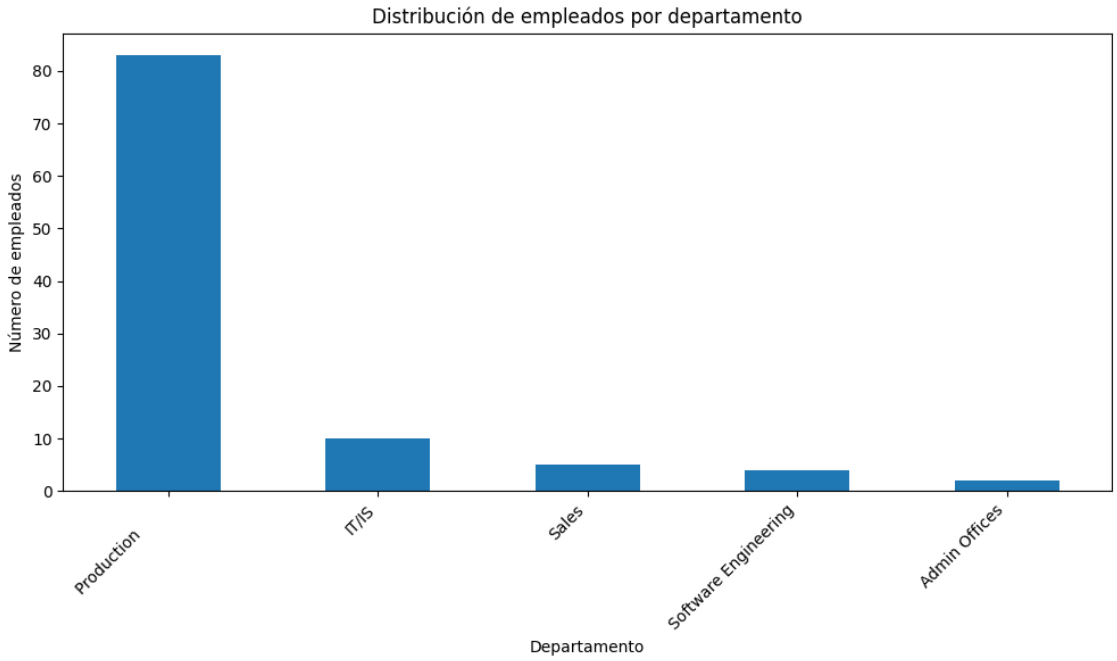
[11]:
```
#Eliminamos los registros nulos
df_sin_nulos = df.dropna()
df_sin_nulos.isnull().sum()
```

[11]:
```
Employee_Name                0
EmpID                        0
```

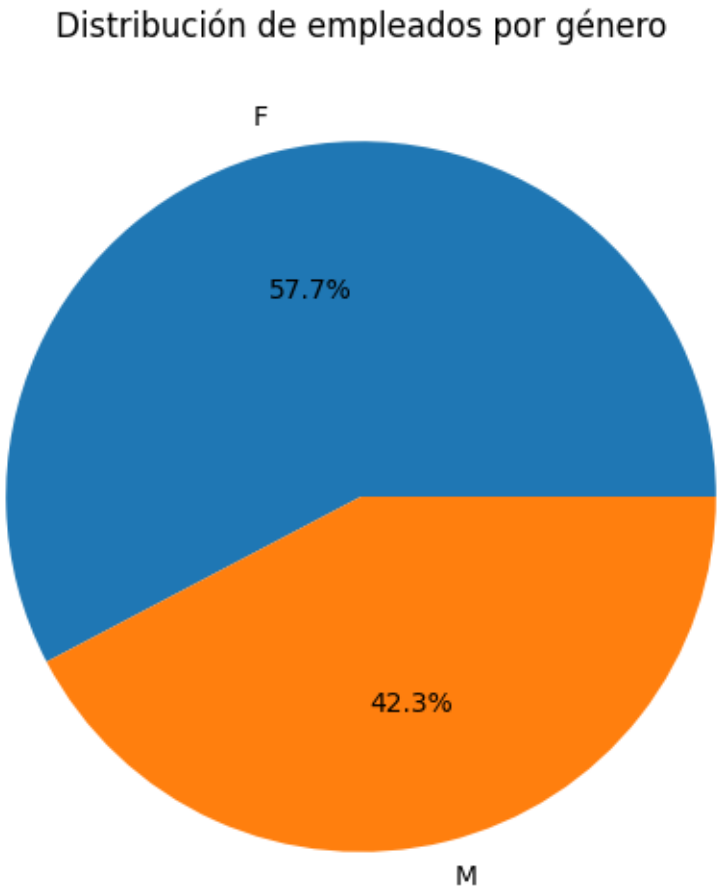
```
Salary                          0
Position                        0
State                           0
DOB                             0
Sex                             0
MaritalDesc                     0
CitizenDesc                     0
HispanicLatino                  0
RaceDesc                        0
DateofHire                      0
DateofTermination               0
TermReason                      0
EmploymentStatus                0
Department                      0
ManagerName                     0
RecruitmentSource               0
PerformanceScore                0
EngagementSurvey                0
EmpSatisfaction                 0
SpecialProjectsCount            0
LastPerformanceReview_Date      0
DaysLateLast30                  0
Absences                        0
dtype: int64
```

[12]:
```
#Exportamos los datos limpios a un excel
df_sin_nulos.to_excel('data_RH.xlsx' , index=False)
```

[14]:
```
# Distribución de empleados por departamento
plt.figure(figsize=(10, 6))
df_sin_nulos['Department'].value_counts().plot(kind='bar')
plt.title('Distribución de empleados por departamento')
plt.xlabel('Departamento')
plt.ylabel('Número de empleados')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

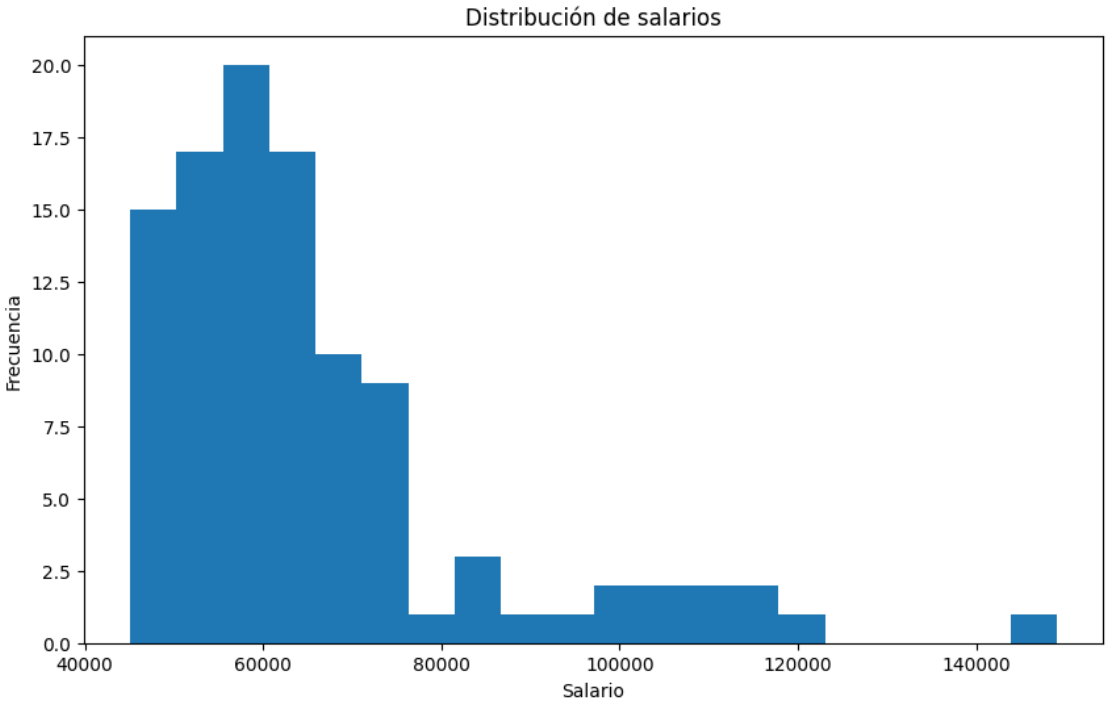


[15]:
```
# Distribución de empleados por género
plt.figure(figsize=(6, 6))
df_sin_nulos['Sex'].value_counts().plot(kind='pie', autopct='%1.1f%%')
plt.title('Distribución de empleados por género')
plt.ylabel('')  # Ocultar la etiqueta del eje y
plt.show()
```



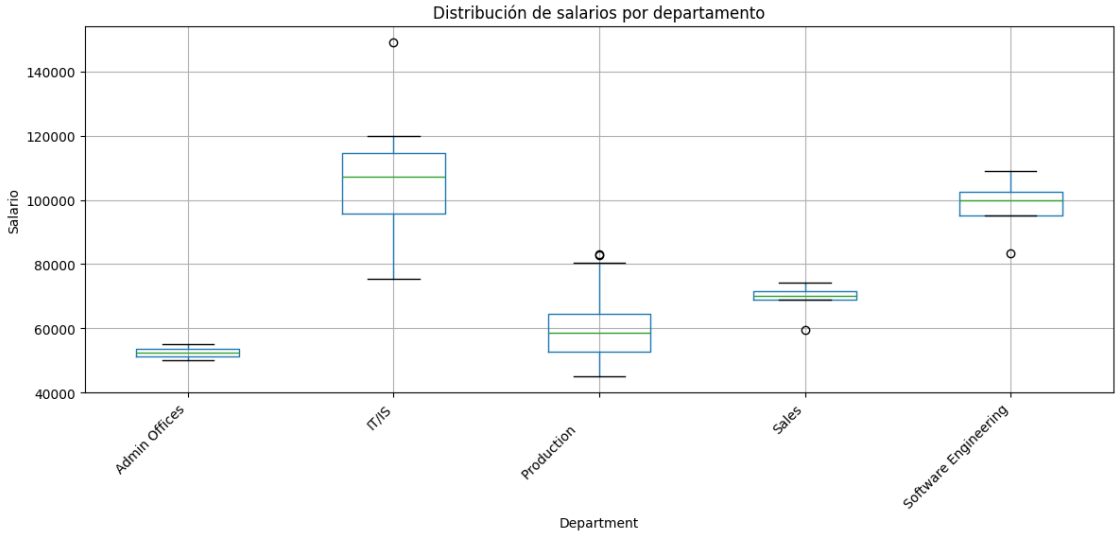
[16]:
```
# Relación entre salario y desempeño
plt.figure(figsize=(10, 6))
plt.scatter(df_sin_nulos['Salary'], df_sin_nulos['PerformanceScore'])
plt.title('Relación entre salario y desempeño')
plt.xlabel('Salario')
plt.ylabel('Puntuación de desempeño')
plt.show()
```

[17]:
```
# Histograma de salarios
plt.figure(figsize=(10, 6))
plt.hist(df_sin_nulos['Salary'], bins=20)
plt.title('Distribución de salarios')
plt.xlabel('Salario')
plt.ylabel('Frecuencia')
plt.show()
```



[18]:
```
# Boxplot de salarios por departamento
plt.figure(figsize=(10, 6))
df_sin_nulos.boxplot(column='Salary', by='Department', figsize=(12, 6))
plt.title('Distribución de salarios por departamento')
plt.suptitle('') # Quitar el título generado automáticamente por boxplot
plt.ylabel('Salario')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```


Distribución de salarios por departamento
Salario
140000
120000
100000
80000
60000
40000
Admin Offices
IT/IS
Production
Sales
Software Engineering
Department

[ ]: