Salarios y trabajos en Ciencia de Datos

In [3]:
```
import pandas as pd
import numpy as np
import seaborn as sns
```

In [4]:
```
#Carga de archivo de datos
df = pd.read_csv(r'C:\Users\dcodd\OneDrive\Escritorio\Data2\jobs_in_data.csv')
```

In [5]:
```
#Visualizamos las primeras 10 filas
df.head(10)
```

Out[5]:

	work_year	job_title	job_category	salary_currency	salary	salary_in_usd	employee_residence	experience_level	employment_type
0	2023	Data DevOps Engineer	Data Engineering	EUR	88000	95012	Germany	Mid-level	Full-time
1	2023	Data Architect	Data Architecture and Modeling	USD	186000	186000	United States	Senior	Full-time
2	2023	Data Architect	Data Architecture and Modeling	USD	81800	81800	United States	Senior	Full-time
3	2023	Data Scientist	Data Science and Research	USD	212000	212000	United States	Senior	Full-time
4	2023	Data Scientist	Data Science and Research	USD	93300	93300	United States	Senior	Full-time
5	2023	Data Scientist	Data Science and Research	USD	130000	130000	United States	Senior	Full-time
6	2023	Data Scientist	Data Science and Research	USD	100000	100000	United States	Senior	Full-time

7	2023	Machine Learning Researcher	Machine Learning and AI	USD	224400	224400	United States	Mid-level	Full-time
8	2023	Machine Learning Researcher	Machine Learning and AI	USD	138700	138700	United States	Mid-level	Full-time
9	2023	Data Engineer	Data Engineering	USD	210000	210000	United States	Executive	Full-time

In [6]:
```
#Verificacion de valores nulos y ceros
print(df.isnull().sum())
print((df == 0).sum())
```

```
work_year             0
job_title             0
job_category          0
salary_currency       0
salary                0
salary_in_usd         0
employee_residence    0
experience_level      0
employment_type       0
work_setting          0
company_location      0
company_size          0
dtype: int64
work_year             0
job_title             0
job_category          0
salary_currency       0
salary                0
salary_in_usd         0
employee_residence    0
experience_level      0
employment_type       0
work_setting          0
company_location      0
```

```
company_size          0
dtype: int64
```

In [7]:
```
#Informacion sobre el data frame y algunos datos estadisticos a tener en cuenta
df.info()
df.describe()
```

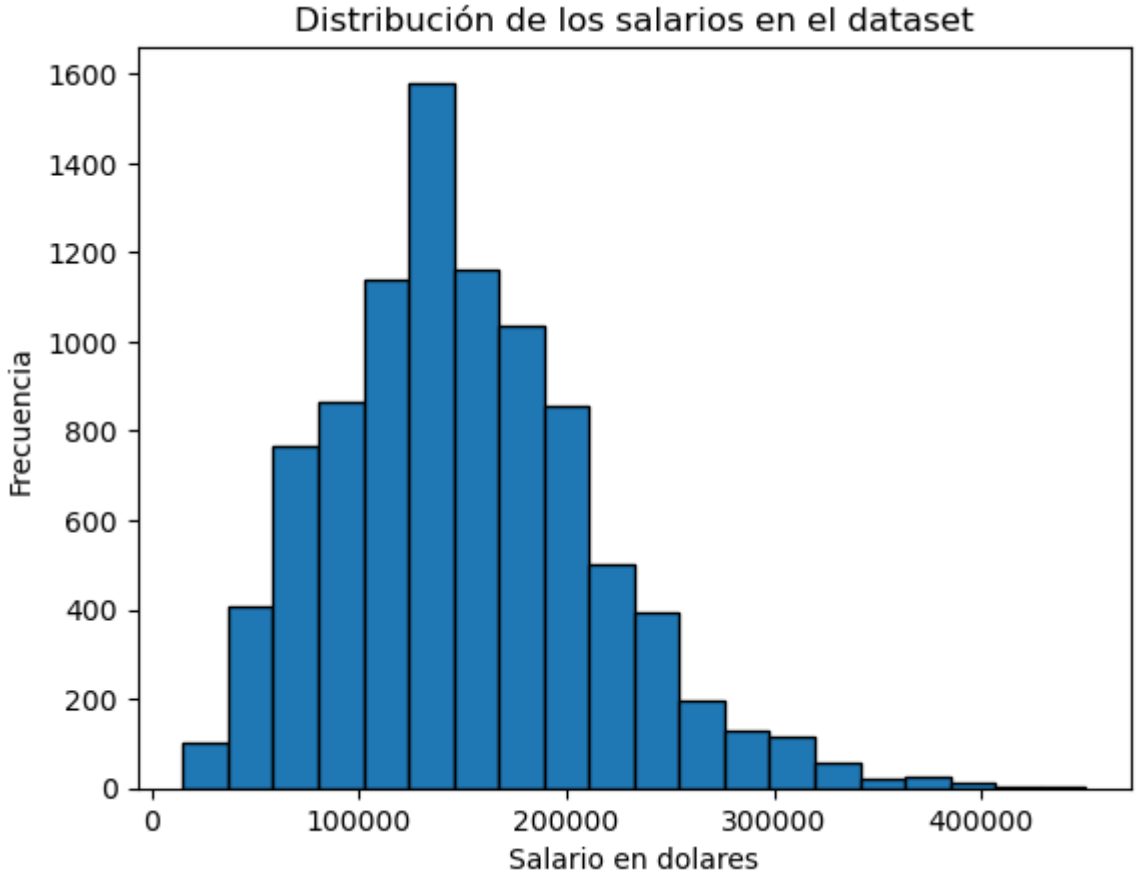
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9355 entries, 0 to 9354
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   work_year           9355 non-null   int64 
 1   job_title           9355 non-null   object
 2   job_category        9355 non-null   object
 3   salary_currency     9355 non-null   object
 4   salary              9355 non-null   int64 
 5   salary_in_usd       9355 non-null   int64 
 6   employee_residence  9355 non-null   object
 7   experience_level    9355 non-null   object
 8   employment_type     9355 non-null   object
 9   work_setting        9355 non-null   object
 10  company_location    9355 non-null   object
 11  company_size        9355 non-null   object
dtypes: int64(3), object(9)
memory usage: 877.2+ KB
```

Out[7]:

	work_year	salary	salary_in_usd
count	9355.000000	9355.000000	9355.000000
mean	2022.760449	149927.981293	150299.495564
std	0.519470	63608.835387	63177.372024
min	2020.000000	14000.000000	15000.000000
25%	2023.000000	105200.000000	105700.000000
50%	2023.000000	143860.000000	143000.000000
75%	2023.000000	187000.000000	186723.000000

max	2023.000000	450000.000000	450000.000000

In [8]:
```
#Grafico para visualizar la distribuicion de los salarios en el dataset
import matplotlib.pyplot as plt
df['salary_in_usd'].plot(kind='hist', bins=20, edgecolor='black')
plt.xlabel('Salario en dolares')
plt.ylabel('Frecuencia')
plt.title('Distribución de los salarios en el dataset')
plt.show()
```



Basandonos en este grafico y la informacion obtenida utilizando el metodo describe(), se pueden obtener algunas conclusiones como.

La media de los salarios en dolares ronda los 150000 dolares mientras que la mediana es de aproximadamente 143000 dolares. Al ser la media mayor que la mediana se sugiere que la distribuicion de los salarios esta sesgada hacia la derecha, lo que indica que hay mas valores altos que bajos.

El valor maximo del salario en dolares es de 450000 mientras que el valor mas bajo es de 15000

El desvio estandar de los salarios es de 63177.37, lo que indica una gran variabilidad en este conjunto de datos

In [9]:
```
#Promedios de salarios por tipo de trabajo
promedio_por_categoria = df.groupby('job_category')['salary_in_usd'].mean().map(lambda x: f'${x:.2f}').sort_v
df_promedio_por_categoria = promedio_por_categoria.reset_index()
df_promedio_por_categoria.columns = ['job_category', 'salary_in_usd']
display(df_promedio_por_categoria)
```

	job_category	salary_in_usd
0	Machine Learning and AI	$178925.85
1	Data Science and Research	$163758.58
2	Data Architecture and Modeling	$156002.36
3	Cloud and Database	$155000.00
4	Data Engineering	$146197.66
5	Leadership and Management	$145476.02
6	BI and Visualization	$135092.10
7	Data Analysis	$108505.72
8	Data Management and Strategy	$103139.93
9	Data Quality and Operations	$100879.47

Teniendo en cuenta estos datos podemos observar una gran diferencia significativa en los salarios promedio entre las distintas categorias. Esto puede deberse a factores como a demanda de habilidades específicas, la complejidad del trabajo, la experiencia requerida y el nivel de responsabilidad del trabajo.

In [21]:
```
df_sin_columnas_no_numericas = df.select_dtypes(include=[np.number])
correlacion = df_sin_columnas_no_numericas.corr()
print(correlacion)
```

```
                work_year    salary  salary_in_usd
work_year        1.000000  0.160708       0.166003
salary           0.160708  1.000000       0.991309
salary_in_usd    0.166003  0.991309       1.000000
```

Lo que hice aca fue calculare la correlacion entre las variables año de trabajo, salario, y salario en dolares.

La correlacion entre salario y salario en dolares es 1 ya que son las mismas variables expresadas en diferentes monedas. La correlacion entre año de trabajo y el salario es positiva baja, lo cual nos indica que el salario tiende a aumentar ligeramente a medida que se tienen mas años de trabajo (deberia agregar mas datos con otros años de trabajo para establecer mejor esto).

In [18]:
```
#Promedio por nivel de experiencia
promedio_por_nivel_de_experiencia = df.groupby('experience_level')['salary_in_usd'].mean().sort_values(ascend
print(promedio_por_nivel_de_experiencia)
```

```
experience_level
Executive      189462.914591
Senior         162356.126099
Mid-level      117523.918138
Entry-level     88534.776210
Name: salary_in_usd, dtype: float64
```

Aca analice el promedio de salario por nivel de experiencia, con esta informacion se pueden obtener algunas conclusiones como:

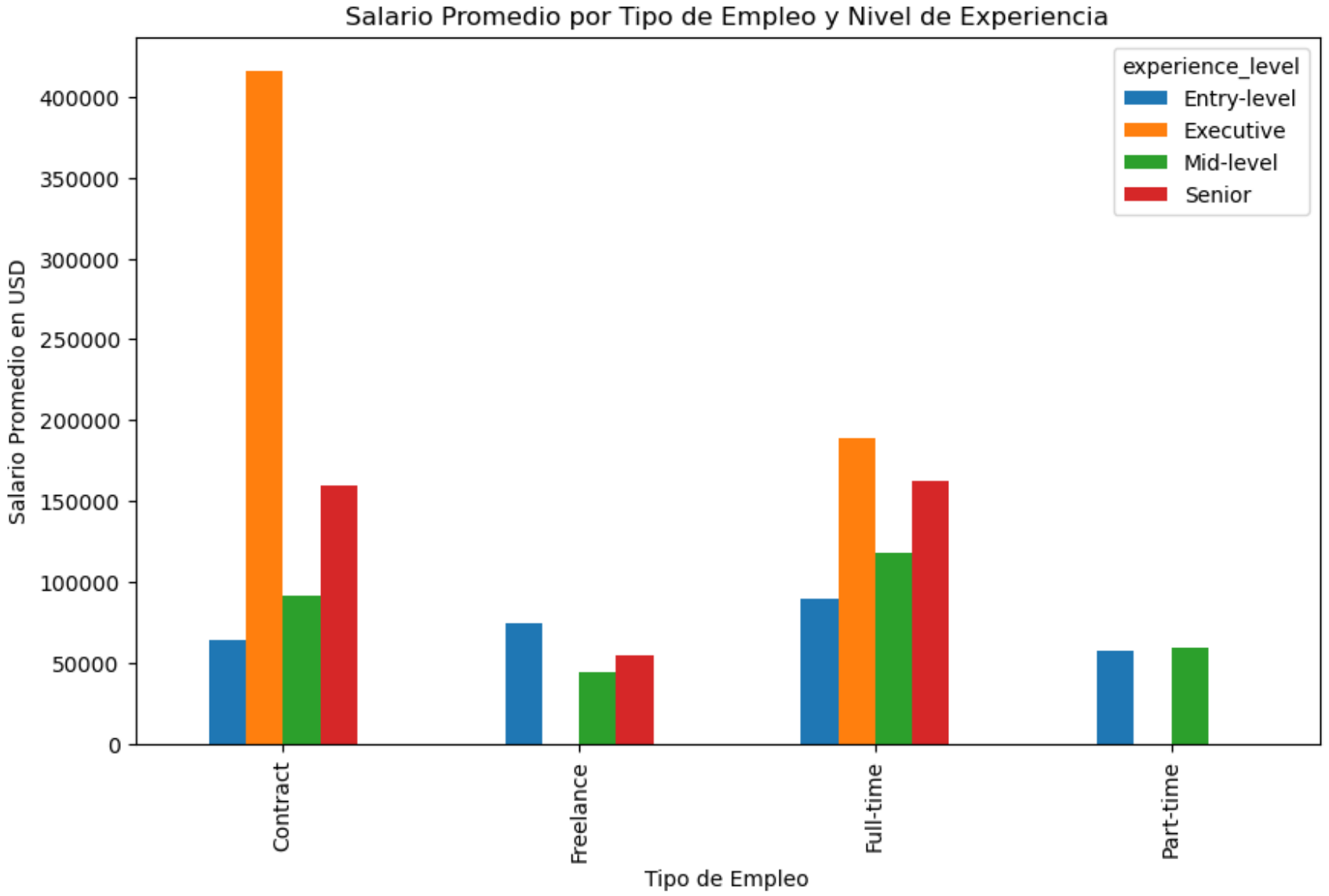
1. La demanda de cientificos de datos a aumentado significativamente en los ultimos años por el crecimiento de la industria, lo cual lleva a que las empresas esten dispuestas a pagar salarios mas altos para atraer a mejores talentos.
2. Tambien podriamos suponer que la oferta de cientificos de datos no ha aumentado al mismo ritmo que la demanda.

3. Esto tambien resalta que los cientificos de datos con mas experiencia pueden tener ciertas habilidades que los mas nuevos no poseen, lo cual, los hace mas valiosos para las empresas. Esto tambien nos puede llevar a una conclusion de que los cientificos de datos con mas experiencia son mas productivos y eficientes en su trabajo.

In [27]:
```
df['employee_residence'] = pd.Categorical(df['employee_residence'])
df['employee_residence'] = df['employee_residence'].cat.codes
correlacion_salario_residencia = df['salary_in_usd'].corr(df['employee_residence'])
print(correlacion_salario_residencia)
```

```
0.19430437861418492
```

In [29]:
```
promedio_por_tipo_de_empleo_y_nivel_de_experiencia = df.groupby(['employment_type', 'experience_level'])['sal
promedio_por_tipo_de_empleo_y_nivel_de_experiencia = promedio_por_tipo_de_empleo_y_nivel_de_experiencia.unsta
promedio_por_tipo_de_empleo_y_nivel_de_experiencia.plot(kind='bar', figsize=(10, 6))
plt.title('Salario Promedio por Tipo de Empleo y Nivel de Experiencia')
plt.xlabel('Tipo de Empleo')
plt.ylabel('Salario Promedio en USD')
plt.show()
```



El gráfico muestra el salario promedio en dólares para diferentes tipos de empleo y niveles de experiencia.

Los tipos de empleo se dividen en tres categorías: contrato, freelance y part-time. Se puede observar que el tipo de empleo con el salario promedio más alto es el contrato, seguido por freelance y part-time, que tienen el salario promedio más bajo. Esto sugiere que los empleados con contrato tienden a ganar más en promedio que los empleados freelance y part-time. Además, se puede observar que el salario promedio aumenta a medida que el nivel de experiencia aumenta, independientemente del tipo de empleo. Esto sugiere que la experiencia puede tener un impacto significativo en el salario de un empleado, independientemente del tipo de empleo que tenga.

Conclusiones generales sobre el analisis:

Los salarios promedio varían significativamente entre diferentes categorías de trabajo, con Machine Learning y AI como la categoría con el salario promedio más alto y Data Quality y Operations como la categoría con el salario promedio más bajo.

Tambien se observa que el salario promedio aumenta a medida que el nivel de experiencia aumenta, independientemente del tipo de empleo. Esto sugiere que la experiencia puede tener un impacto significativo en el salario de un empleado, independientemente del tipo de empleo que tenga.

Estos datos pueden ser de utilidad para tener un enfoque sobre las tendencias en la industria de la ciencia de datos.