

In []:

ANALYSIS OF JOBS IN DATA SCIENCE IN RECENT TIMES

INTRODUCTION

In recent years, the field of data science has experienced a rapid **and** profound evolution, transforming the way business is conducted across various industries. The proliferation of digital technologies, coupled **with** the exponential growth of data, has fueled the demand **for** skilled professionals in data science roles across various industries.

The introduction of data science jobs has reshaped traditional business models, leading to a paradigm shift where data is no longer just a resource but a strategic asset. These roles encompass a wide range of responsibilities, **from** collecting, cleaning, **and** preparing data to extracting insights **and** predictions through advanced analytics **and** machine learning techniques.

ACKNOWLEDGEMENT:

We Omprava Mandal, Supritha Pal, Payel Ghosh, **and** Pallabi Dutta would like to express our sincere gratitude **and** appreciation to all the members of our team who have contributed to the successful execution of this data science job analysis project. We extend our thanks to each member for their hard work, contributions, **and** dedication, which significantly enriched the project's **outcomes**.

The csv file **is** being found, downloaded, **and** then cleaned **and** the project **is** being executed by our team members **with** the support of Suramya Biswas sir. We are deeply thankful to Suramya Biswas sir **for** the guidance, support, **and** invaluable insights provided throughout the project. We are grateful to Anudip Foundation **for** granting access to essential resources, datasets, **and** tools crucial to conducting this analysis. We acknowledge the creators **and** contributors of Python, Word that were instrumental **in** data collection, analysis, **and** presentation.

This project would **not** have been possible without the collective efforts **and** support of each individual **and** entity mentioned above. Their contributions have been invaluable **in** shaping the outcomes **and** enhancing the quality of this endeavor.

OBJECTIVES:

Here are some potential objectives:

Trend Analysis: Understand the overall trend **in** demand **for** data science jobs over a specific period. In our project, the demand **is** increasing **in** an exponential height, especially **in** foreign countries.



```
In [ ]: Geographical Distribution: The geographical distribution of data science jobs has been widespread in recent times, with hubs in Silicon Valley, New York, London, and Bangalore. However, the demand for data scientists has been increasing in cities like Berlin, Singapore, and Tel Aviv. Remote work options have also increased, contributing to the field from various locations.
```

Countries vs job graph



In []: Data science jobs have seen significant demand across various industries. Technology companies, finance, healthcare, continue to heavily invest in data science talent. Additionally, industries such as manufacturing, energy, and logistics science for process optimization and decision-making. The interdisciplinary nature of data science allows it to permeate a crucial skillset in today's job market.

Salary Trends: As of my last knowledge update in January 2022, data science salaries continued to be robust, with varying experience, location, and industry. Generally, data scientists with advanced degrees and several years of experience often command higher salaries. Tech hubs like Silicon Valley, New York, and Seattle often offered higher compensation packages for these professionals.

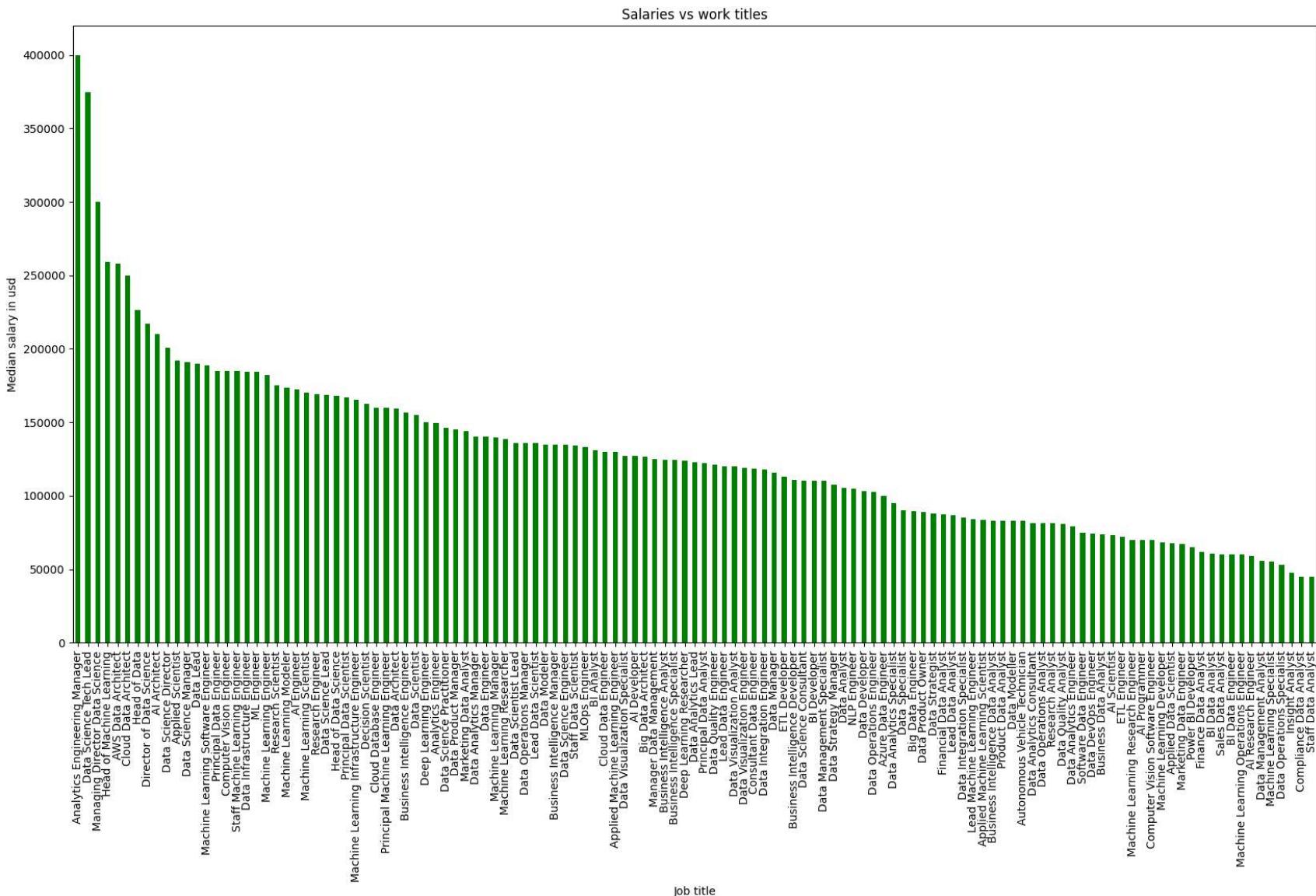
Salaries vs job titles

```
In [55]: csv_file_path = 'jobs_in_data_science.csv'
df = pd.read_csv(csv_file_path)
plt.figure(figsize=(20,10))
salary_by_job_title = df.groupby('job_title')['salary_in_usd'].median().sort_values(ascending=False)
salary_by_job_title.plot(kind='bar', color='green')
```

```

plt.title('Salaries vs work titles')
plt.xlabel('Job title')
plt.ylabel('Median salary in usd')
plt.show()

```



In []: However, salary trends can change over time based on market dynamics, so it's advisable to consult the latest industry salary surveys for the most up-to-date information on data science salaries in 2024.

Comparison Across Companies: Different companies may emphasize distinct aspects of data science jobs based on their size and industry. Tech giants like Google, Amazon, and Microsoft often focus on large-scale data processing, machine learning, and artificial intelligence. Financial institutions like JPMorgan Chase or Goldman Sachs may emphasize risk management and financial modeling.

Startups often require data scientists to wear multiple hats, engaging in both analysis and implementation. Healthcare organizations may prioritize data scientists working on healthcare analytics and predictive modeling.

It's essential to consider the industry, company size, and specific business goals when comparing data science job requirements. A company organization tailors data science roles to address its unique challenges and objectives.

LIBRARIES AND TOOLS:

- **NumPy:** For numerical computations and statistical analysis.
- **Matplotlib:** For creating visualizations and plots.
- **Pandas:** For data manipulation and analysis.
- **Basic Python:** For handling data structures like arrays, tuples, and basic operations.

DATA COLLECTION STEPS:

Data sources is being identified from Kaggle. The CSV file is downloaded and the information and dataset is being collected.

Define Data Structure:

Determining the information for collecting for each job posting: job title, company, location, description, required experience, and salary.

Data Collection Execution:

The extracted data is stored in a structured format i.e, csv file. It is being formatted, cleaned, and executed.

EXPLORATORY DATA ANALYSIS (EDA)-

is crucial for understanding the characteristics and patterns within the collected dataset.

Here's a structured approach using Python libraries like Pandas and Matplotlib:

Exploratory Data Analysis Steps:

Data Loading: We are using Pandas to load the dataset (CSV) collected during the data collection phase. The dataset is loaded and read.

In [5]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
def pythonProject():
    csv_file_path = 'jobs_in_data_science.csv'
    df = pd.read_csv(csv_file_path)
```

```
    print(df)
pythonProject()
```

	work_year		job_title		job_category	\
0	2023	Data DevOps Engineer		Data Engineering		
1	2023	Data Architect	Data Architect	Data Architecture and Modeling		
2	2023	Data Architect	Data Architect	Data Architecture and Modeling		
3	2023	Data Scientist	Data Scientist	Data Science and Research		
4	2023	Data Scientist	Data Scientist	Data Science and Research		
...	
9350	2021	Data Specialist	Data Management and Strategy			
9351	2020	Data Scientist	Data Science and Research			
9352	2021	Principal Data Scientist	Data Science and Research			
9353	2020	Data Scientist	Data Science and Research			
9354	2020	Business Data Analyst		Data Analysis		
	salary_currency	salary	salary_in_usd	employee_residence	\	
0	EUR	88000	95012	Germany		
1	USD	186000	186000	United States		
2	USD	81800	81800	United States		
3	USD	212000	212000	United States		
4	USD	93300	93300	United States		
...	
9350	USD	165000	165000	United States		
9351	USD	412000	412000	United States		
9352	USD	151000	151000	United States		
9353	USD	105000	105000	United States		
9354	USD	100000	100000	United States		
	experience_level	employment_type	work_setting	company_location	\	
0	Mid-level	Full-time	Hybrid	Germany		
1	Senior	Full-time	In-person	United States		
2	Senior	Full-time	In-person	United States		
3	Senior	Full-time	In-person	United States		
4	Senior	Full-time	In-person	United States		
...	
9350	Senior	Full-time	Remote	United States		
9351	Senior	Full-time	Remote	United States		
9352	Mid-level	Full-time	Remote	United States		
9353	Entry-level	Full-time	Remote	United States		
9354	Entry-level	Contract	Remote	United States		
	company_size					
0	L					
1	M					

```
2          M
3          M
4          M
...
9350        L
9351        L
9352        L
9353        S
9354        L
```

```
[9355 rows x 12 columns]
```

```
In [ ]: Dataset info:
```

```
In [4]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
print(df.info())
```

```
<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
None
```

```
In [ ]: Check for missing values :
```

```
In [5]: csv_file_path = 'jobs in data science.csv'  
df = pd.read_csv(csv_file_path)  
print(df.isnull().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
```

```
In [ ]: Drop columns with any missing values:
```

```
In [6]: csv_file_path = 'jobs in data science.csv'  
df = pd.read_csv(csv_file_path)  
df.dropna(axis=1, inplace=True)  
print(df)
```

	work_year		job_title		job_category	\
0	2023	Data DevOps Engineer		Data Engineering		
1	2023	Data Architect	Data Architect	Data Architecture and Modeling		
2	2023	Data Architect	Data Architect	Data Architecture and Modeling		
3	2023	Data Scientist	Data Scientist	Data Science and Research		
4	2023	Data Scientist	Data Scientist	Data Science and Research		
...	
9350	2021	Data Specialist	Data Management and Strategy			
9351	2020	Data Scientist	Data Science and Research			
9352	2021	Principal Data Scientist	Data Science and Research			
9353	2020	Data Scientist	Data Science and Research			
9354	2020	Business Data Analyst		Data Analysis		
	salary_currency	salary	salary_in_usd	employee_residence	\	
0	EUR	88000	95012	Germany		
1	USD	186000	186000	United States		
2	USD	81800	81800	United States		
3	USD	212000	212000	United States		
4	USD	93300	93300	United States		
...	
9350	USD	165000	165000	United States		
9351	USD	412000	412000	United States		
9352	USD	151000	151000	United States		
9353	USD	105000	105000	United States		
9354	USD	100000	100000	United States		
	experience_level	employment_type	work_setting	company_location	\	
0	Mid-level	Full-time	Hybrid	Germany		
1	Senior	Full-time	In-person	United States		
2	Senior	Full-time	In-person	United States		
3	Senior	Full-time	In-person	United States		
4	Senior	Full-time	In-person	United States		
...	
9350	Senior	Full-time	Remote	United States		
9351	Senior	Full-time	Remote	United States		
9352	Mid-level	Full-time	Remote	United States		
9353	Entry-level	Full-time	Remote	United States		
9354	Entry-level	Contract	Remote	United States		
	company_size					
0	L					
1	M					

```
2          M  
3          M  
4          M  
...        ...  
9350         L  
9351         L  
9352         L  
9353         S  
9354         L
```

[9355 rows x 12 columns]

```
In [ ]: Drop rows with missing values:
```

```
In [7]: csv_file_path = 'jobs in data science.csv'  
df = pd.read_csv(csv_file_path)  
df.dropna(inplace=True)  
print(df)
```

	work_year		job_title		job_category	\
0	2023	Data DevOps Engineer		Data Engineering		
1	2023	Data Architect	Data Architect	Data Architecture and Modeling		
2	2023	Data Architect	Data Architect	Data Architecture and Modeling		
3	2023	Data Scientist	Data Scientist	Data Science and Research		
4	2023	Data Scientist	Data Scientist	Data Science and Research		
...	
9350	2021	Data Specialist	Data Management and Strategy			
9351	2020	Data Scientist	Data Science and Research			
9352	2021	Principal Data Scientist	Data Science and Research			
9353	2020	Data Scientist	Data Science and Research			
9354	2020	Business Data Analyst		Data Analysis		
	salary_currency	salary	salary_in_usd	employee_residence	\	
0	EUR	88000	95012	Germany		
1	USD	186000	186000	United States		
2	USD	81800	81800	United States		
3	USD	212000	212000	United States		
4	USD	93300	93300	United States		
...	
9350	USD	165000	165000	United States		
9351	USD	412000	412000	United States		
9352	USD	151000	151000	United States		
9353	USD	105000	105000	United States		
9354	USD	100000	100000	United States		
	experience_level	employment_type	work_setting	company_location	\	
0	Mid-level	Full-time	Hybrid	Germany		
1	Senior	Full-time	In-person	United States		
2	Senior	Full-time	In-person	United States		
3	Senior	Full-time	In-person	United States		
4	Senior	Full-time	In-person	United States		
...	
9350	Senior	Full-time	Remote	United States		
9351	Senior	Full-time	Remote	United States		
9352	Mid-level	Full-time	Remote	United States		
9353	Entry-level	Full-time	Remote	United States		
9354	Entry-level	Contract	Remote	United States		
	company_size					
0	L					
1	M					

```
2          M  
3          M  
4          M  
...        ...  
9350         L  
9351         L  
9352         L  
9353         S  
9354         L
```

[9355 rows x 12 columns]

In []: Data (after cleaning)

```
In [8]: csv_file_path = 'jobs in data science.csv'  
df = pd.read_csv(csv_file_path)  
df.to_csv('cleaned_dataset.csv', index=False)  
print(df)
```

	work_year		job_title		job_category	\
0	2023	Data DevOps Engineer		Data Engineering		
1	2023	Data Architect	Data Architect	Data Architecture and Modeling		
2	2023	Data Architect	Data Architect	Data Architecture and Modeling		
3	2023	Data Scientist	Data Scientist	Data Science and Research		
4	2023	Data Scientist	Data Scientist	Data Science and Research		
...	
9350	2021	Data Specialist	Data Management and Strategy			
9351	2020	Data Scientist	Data Science and Research			
9352	2021	Principal Data Scientist	Data Science and Research			
9353	2020	Data Scientist	Data Science and Research			
9354	2020	Business Data Analyst		Data Analysis		
	salary_currency	salary	salary_in_usd	employee_residence	\	
0	EUR	88000	95012	Germany		
1	USD	186000	186000	United States		
2	USD	81800	81800	United States		
3	USD	212000	212000	United States		
4	USD	93300	93300	United States		
...	
9350	USD	165000	165000	United States		
9351	USD	412000	412000	United States		
9352	USD	151000	151000	United States		
9353	USD	105000	105000	United States		
9354	USD	100000	100000	United States		
	experience_level	employment_type	work_setting	company_location	\	
0	Mid-level	Full-time	Hybrid	Germany		
1	Senior	Full-time	In-person	United States		
2	Senior	Full-time	In-person	United States		
3	Senior	Full-time	In-person	United States		
4	Senior	Full-time	In-person	United States		
...	
9350	Senior	Full-time	Remote	United States		
9351	Senior	Full-time	Remote	United States		
9352	Mid-level	Full-time	Remote	United States		
9353	Entry-level	Full-time	Remote	United States		
9354	Entry-level	Contract	Remote	United States		
	company_size					
0	L					
1	M					

```
2          M
3          M
4          M
...
9350         L
9351         L
9352         L
9353         S
9354         L
```

```
[9355 rows x 12 columns]
```

```
In [ ]: Statistical summary of the dataset:
```

```
In [6]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
print(df['salary_in_usd'].describe())
```

```
count    9355.000000
mean    150299.495564
std     63177.372024
min     15000.000000
25%    105700.000000
50%    143000.000000
75%    186723.000000
max     450000.000000
Name: salary_in_usd, dtype: float64
```

```
In [ ]: Dataset Description with the titles:
```

```
In [7]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
print(df.nunique())
```

```
work_year          4
job_title         125
job_category       10
salary_currency    11
salary            1507
salary_in_usd     1786
employee_residence 83
experience_level    4
employment_type     4
work_setting        3
company_location    70
company_size         3
dtype: int64
```

```
In [ ]:
```

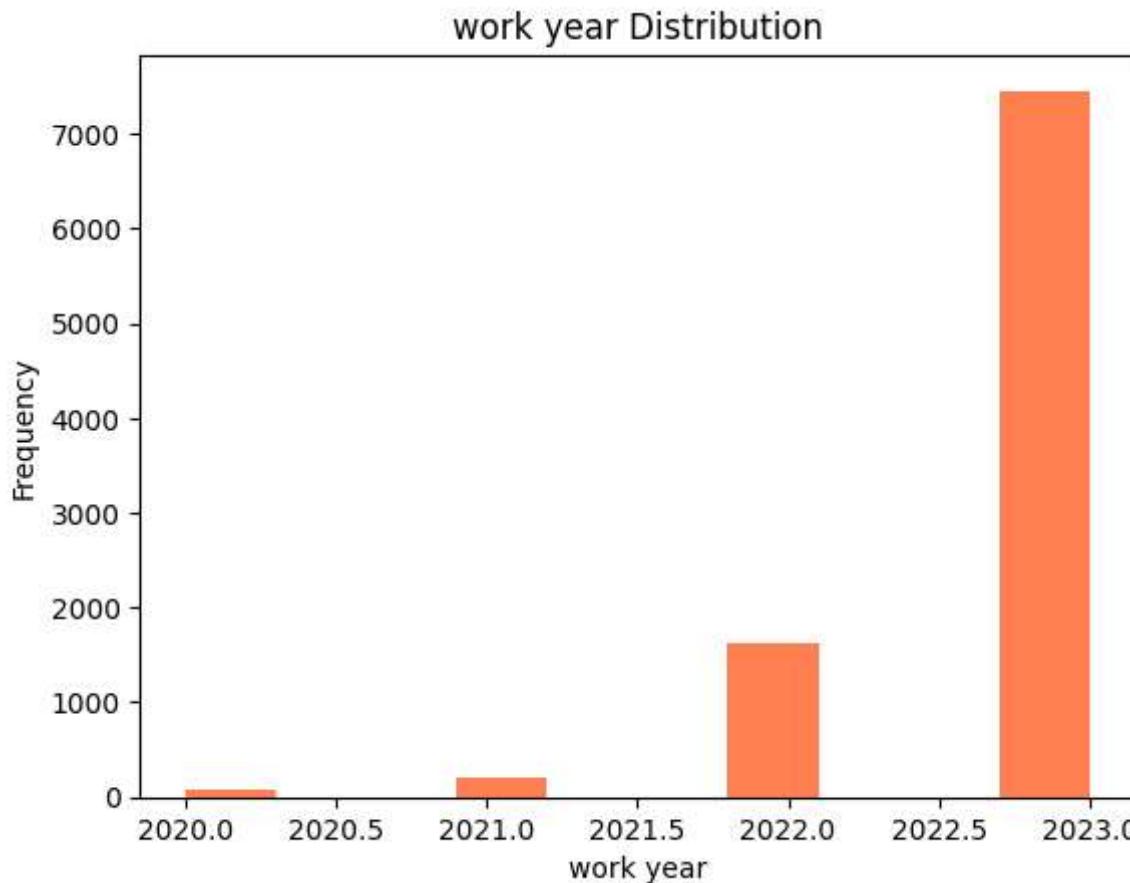
```
In [ ]: CONTEXT QUESTIONS
```

Project Scope **and** Objectives

❓ How does this project show the leniency towards data science field?

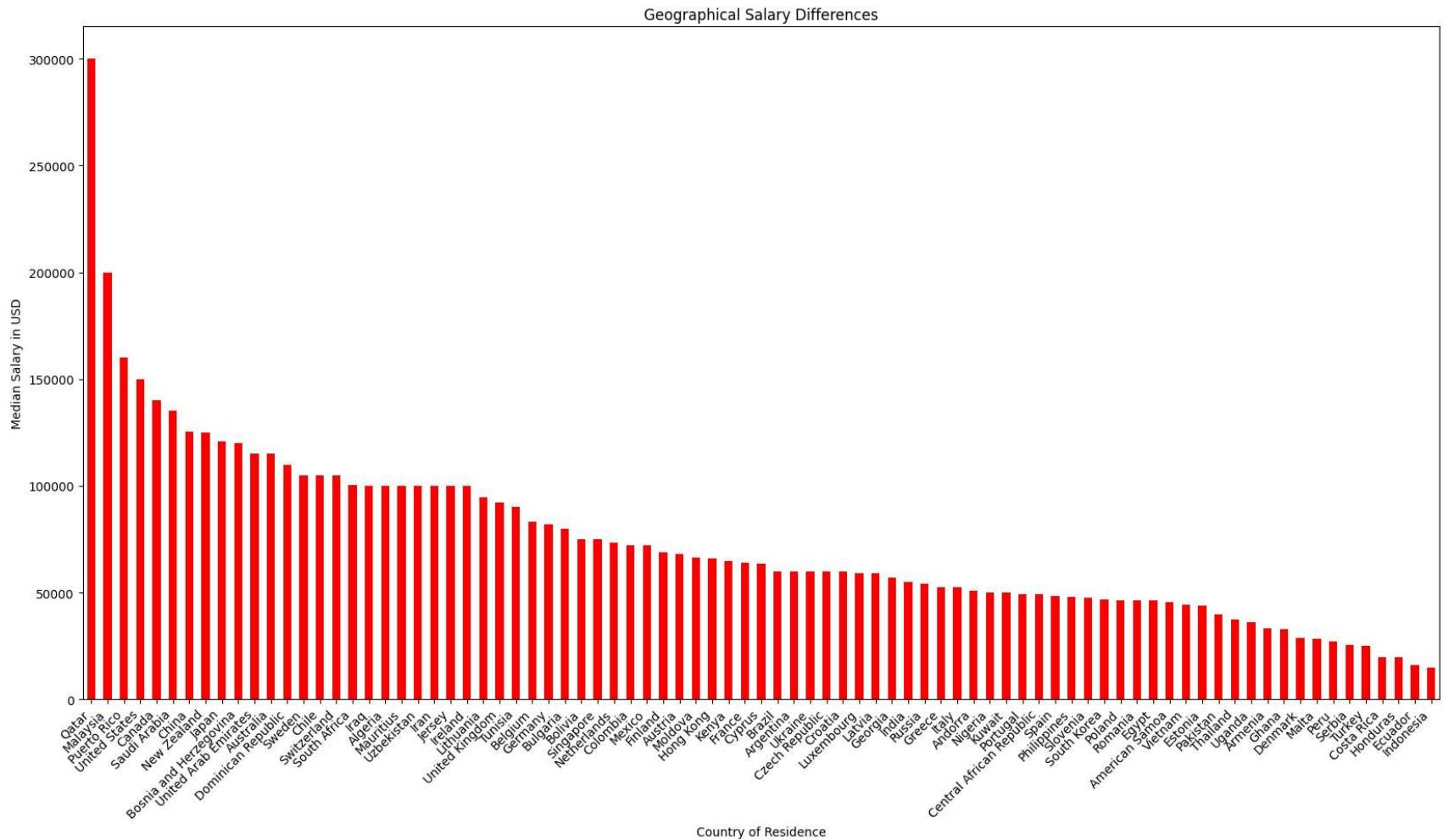
```
In [8]:
```

```
csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
plt.hist(df['work_year'], bins=10, color='coral')
plt.xlabel('work year')
plt.ylabel('Frequency')
plt.title('work year Distribution')
plt.show()
```



In []: Which countries are currently showing a high demand **for** data science professionals?

```
In [37]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
plt.figure(figsize=(20,10))
salary_by_residence = df.groupby('employee_residence')['salary_in_usd'].median().sort_values(ascending=False)
salary_by_residence.plot(kind='bar', color='red')
plt.title('Geographical Salary Differences')
plt.xlabel('Country of Residence')
plt.ylabel('Median Salary in USD')
plt.xticks(rotation=45,ha='right')
plt.show()
```



In [20]: In what ways has the industry focus **for** salaries **for** data science jobs evolved over the past few years?

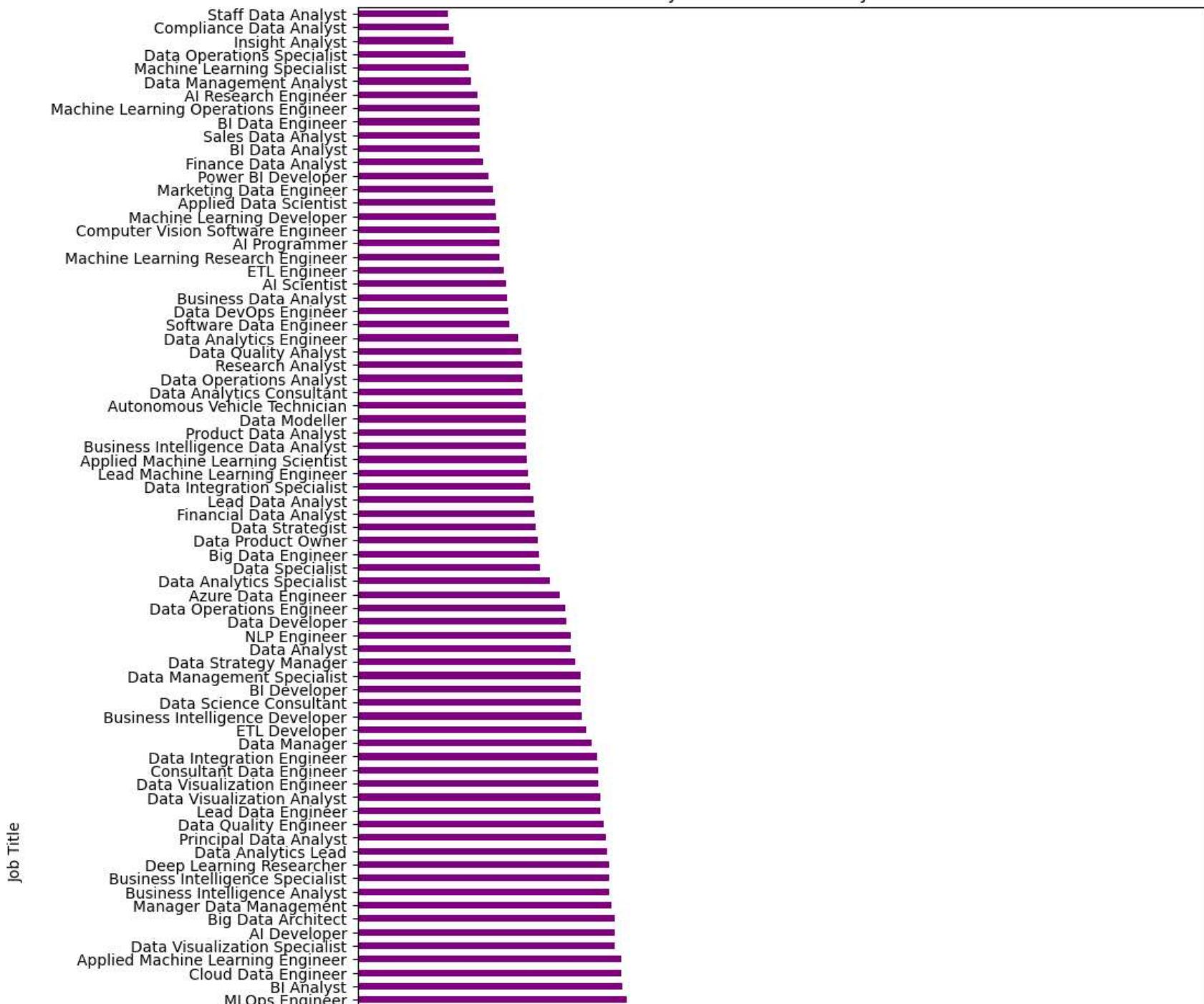
Object `years` not found.

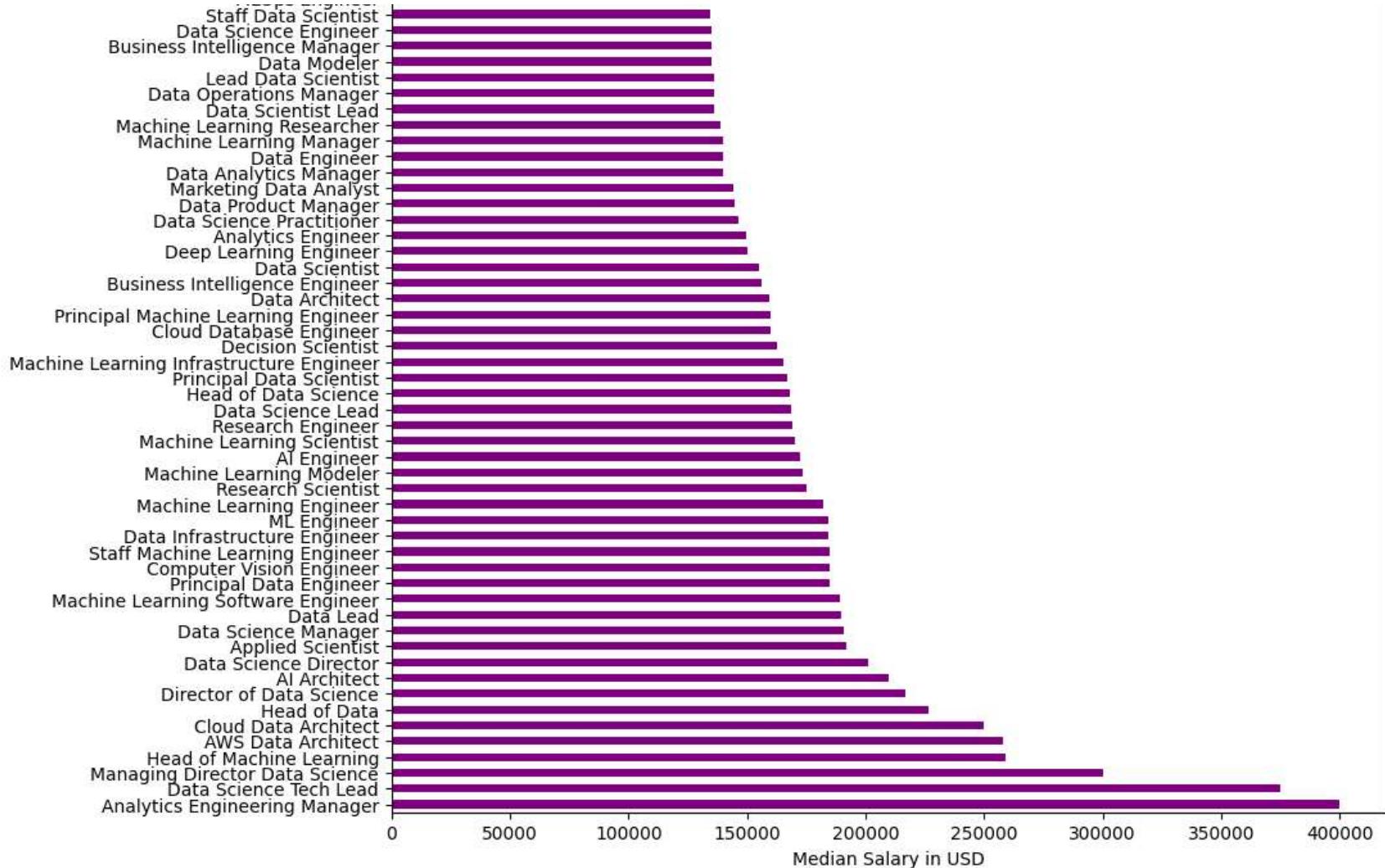
In [36]:

```
csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
job_title_salary_distribution = df.groupby('job_title')['salary_in_usd'].median().sort_values(ascending=False)
plt.figure(figsize=(10,20))
job_title_salary_distribution.plot(kind='barh', color='purple')
plt.title('Salary Distribution Across Job Titles')
plt.xlabel('Median Salary in USD')
plt.ylabel('Job Title')
```

```
plt.show()
```

Salary Distribution Across Job Titles





In []: Data Collection and Sources:

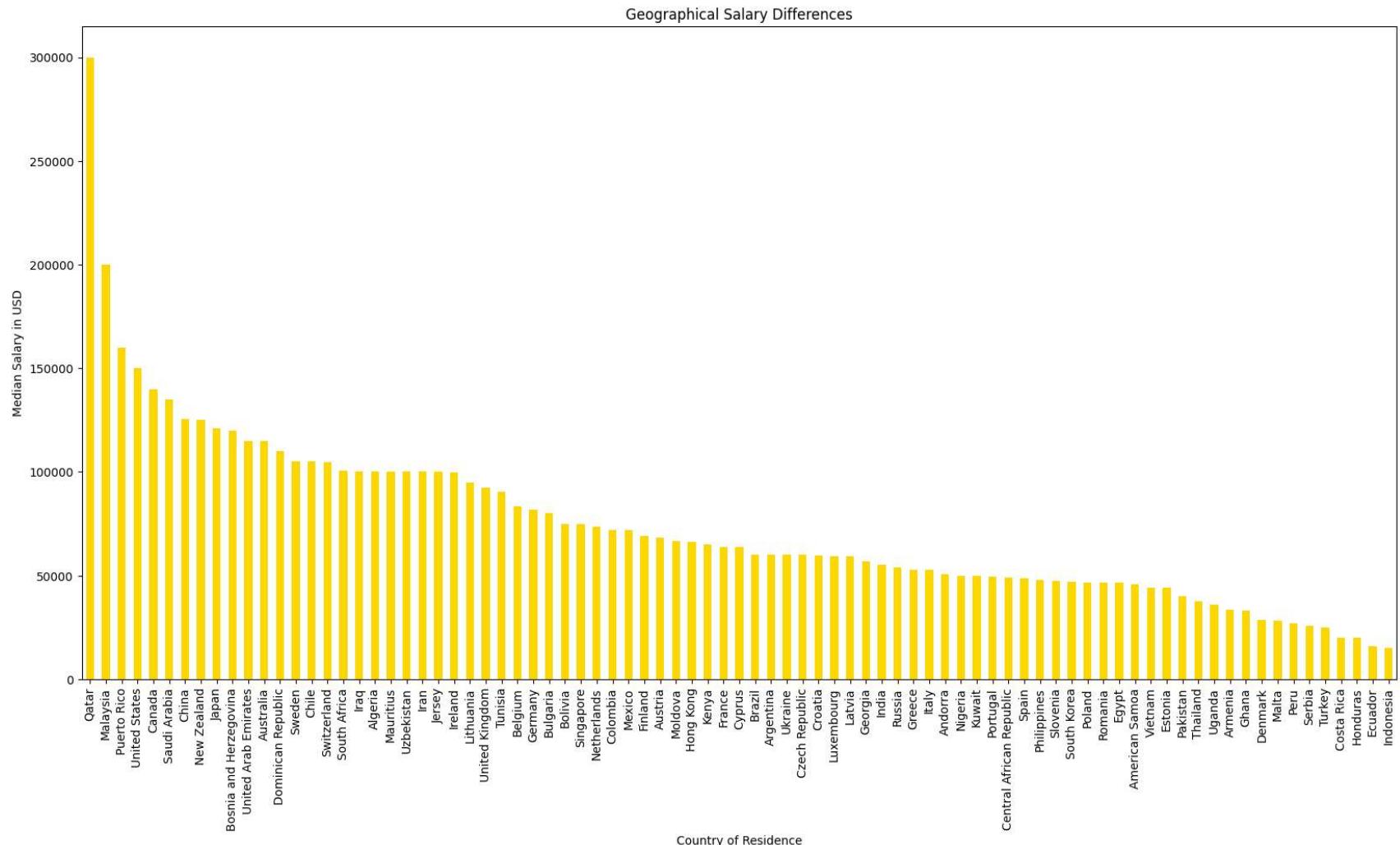
- How was the data collected, and what steps were taken to ensure its accuracy and relevance?

Data has been collected from Kaggle, ChatGPT, and YouTube and after that it is being utilized in the analysis of the

Dataset Questions:

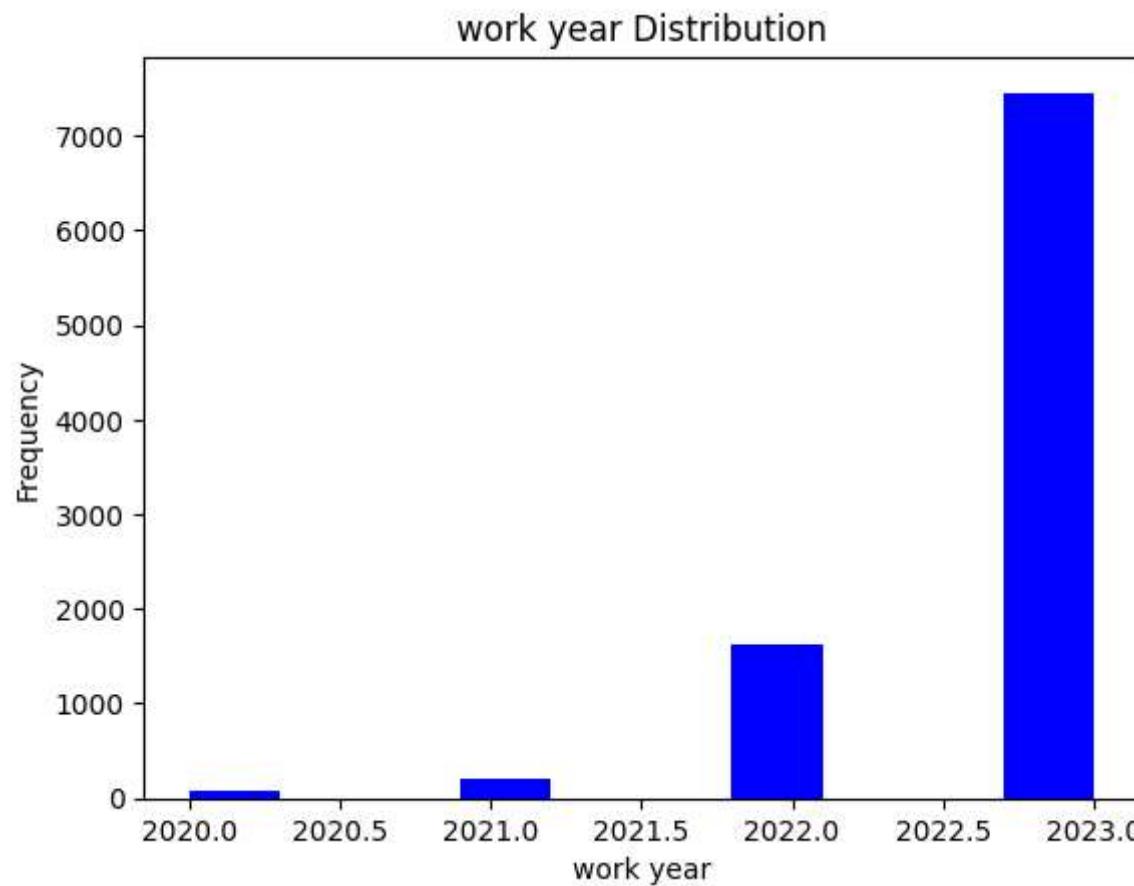
- Which countries recruits and pays maximum salaries?

```
In [38]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
plt.figure(figsize=(20,10))
salary_by_residence = df.groupby('employee_residence')['salary_in_usd'].median().sort_values(ascending=False)
salary_by_residence.plot(kind='bar', color='gold')
plt.title('Geographical Salary Differences')
plt.xlabel('Country of Residence')
plt.ylabel('Median Salary in USD')
plt.show()
```



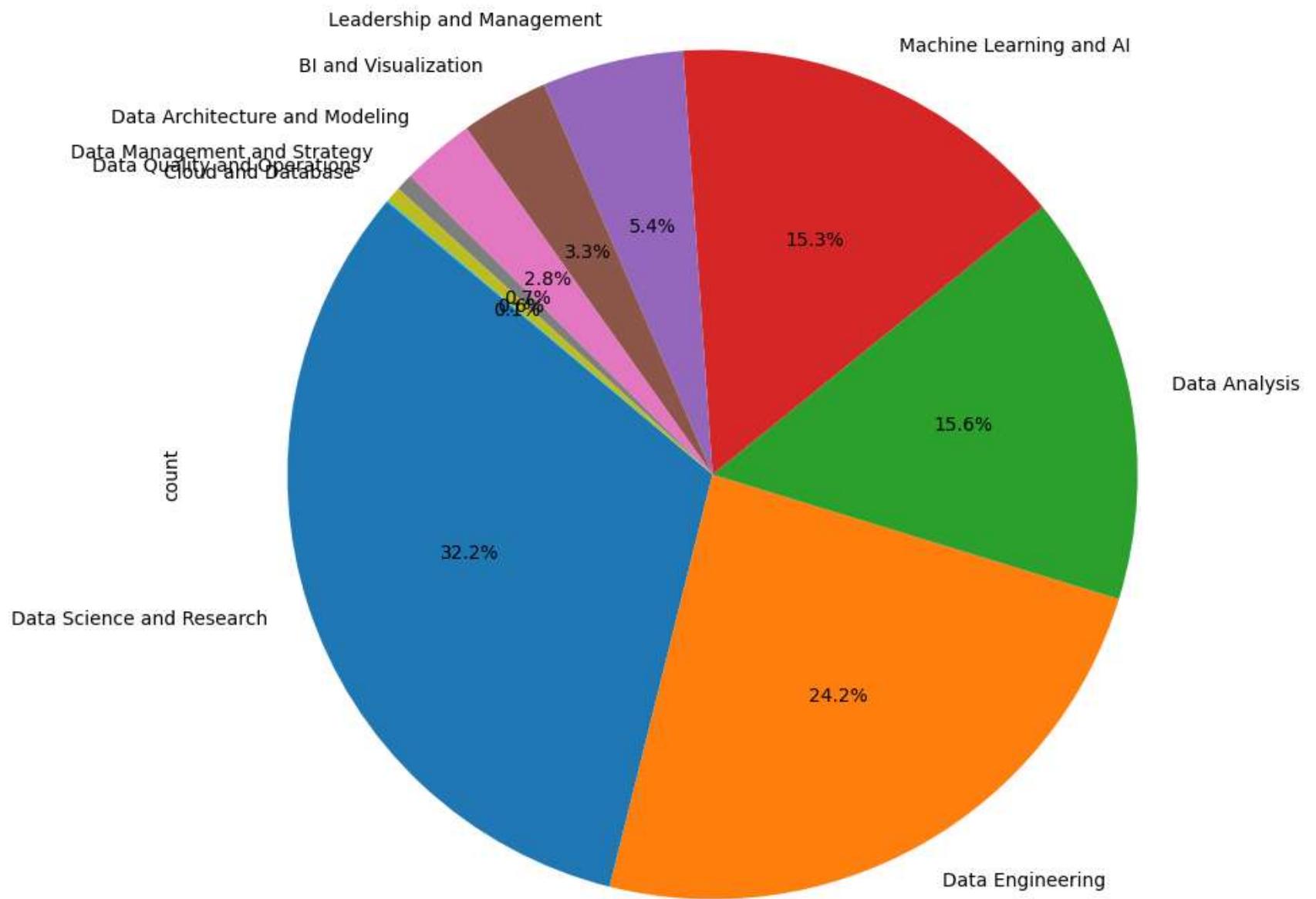
```
In [ ]: 2.      Which year the jobs for data science field is maximum?
```

```
In [40]: csv_file_path = 'jobs in data science.csv'  
df = pd.read_csv(csv_file_path)  
plt.hist(df['work_year'], bins=10, color='blue')  
plt.xlabel('work year')  
plt.ylabel('Frequency')  
plt.title('work year Distribution')  
plt.show()
```



```
In [ ]: 3.      What is the part of data science jobs in all fields of work?
```

```
In [49]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
plt.figure(figsize=(10,9))
df['job_category'].value_counts().plot(kind='pie', autopct='%1.1f%%', startangle=140);
plt.axis('equal')
plt.show()
```



In []: 4. What is the salary distribution based on experience level of employees?

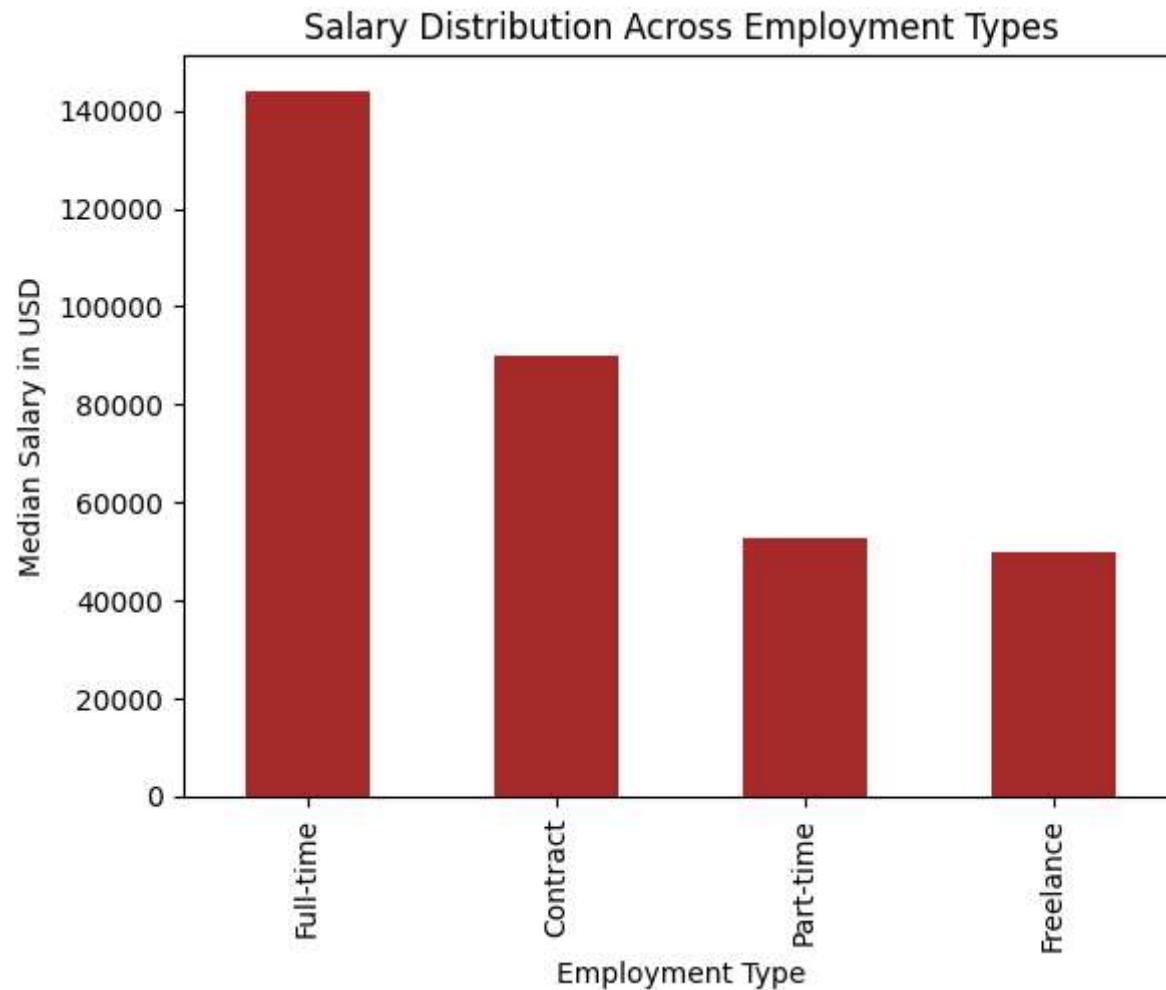
In [50]: `csv_file_path = 'jobs in data science.csv'`
`df = pd.read_csv(csv_file_path)`

```
salary_by_experience = df.groupby('experience_level')['salary_in_usd'].median().sort_values(ascending=False)
salary_by_experience.plot(kind='bar', color='yellow')
plt.title('Salary Distribution Across Experience Levels')
plt.xlabel('Experience Level')
plt.ylabel('Median Salary in USD')
plt.show()
```



In []: 5. What **is** the effect of employee type on salary structure?

```
In [51]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
salary_by_employment_type = df.groupby('employment_type')['salary_in_usd'].median().sort_values(ascending=False)
salary_by_employment_type.plot(kind='bar', color='brown')
plt.title('Salary Distribution Across Employment Types')
plt.xlabel('Employment Type')
plt.ylabel('Median Salary in USD')
plt.show()
```

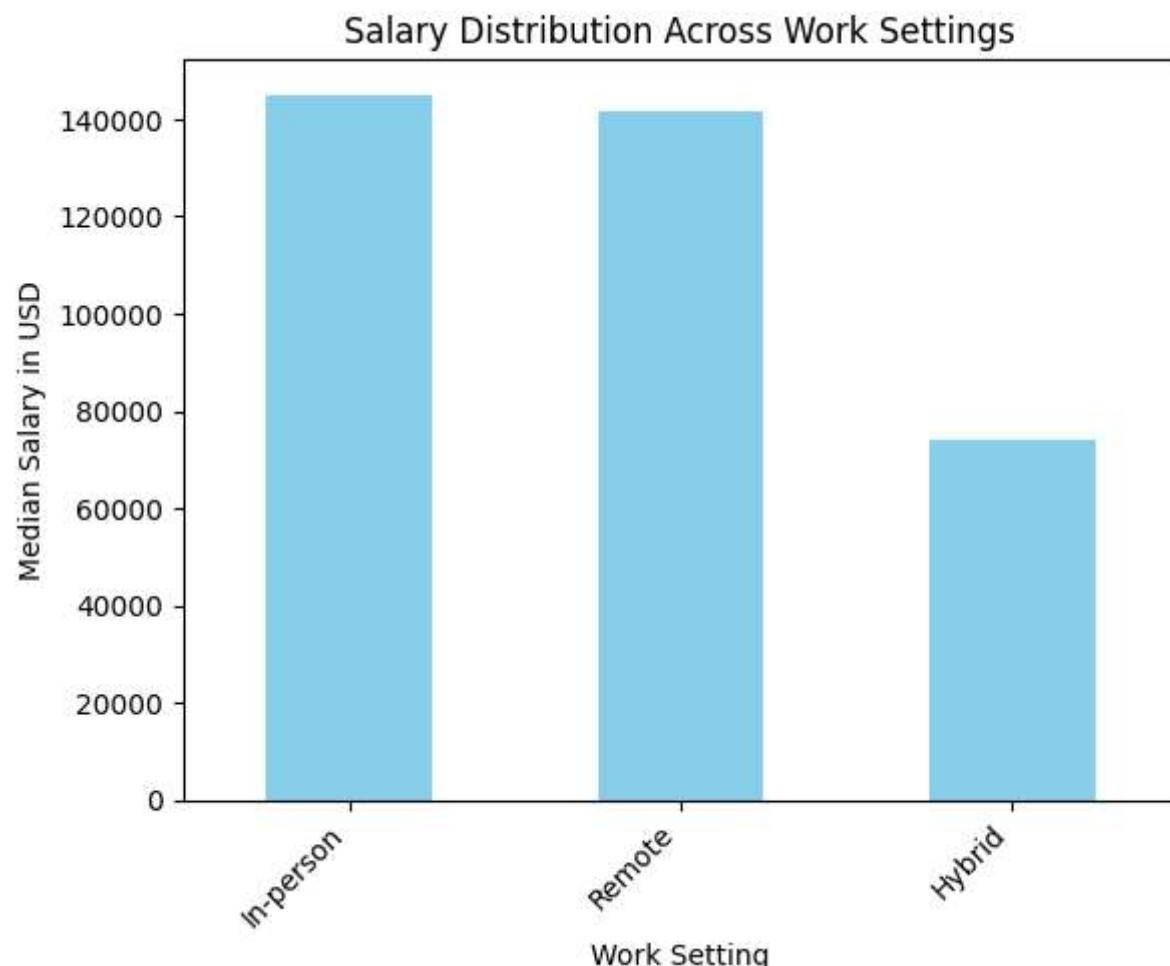


```
In [ ]: Analysis result:
```

What **is** impact of work setting on salary levels?

In [58]:

```
csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
salary_by_work_setting = df.groupby('work_setting')['salary_in_usd'].median().sort_values(ascending=False)
salary_by_work_setting.plot(kind='bar', color='skyblue')
plt.title('Salary Distribution Across Work Settings')
plt.xlabel('Work Setting')
plt.ylabel('Median Salary in USD')
plt.xticks(rotation=45, ha='right')
plt.show()
```



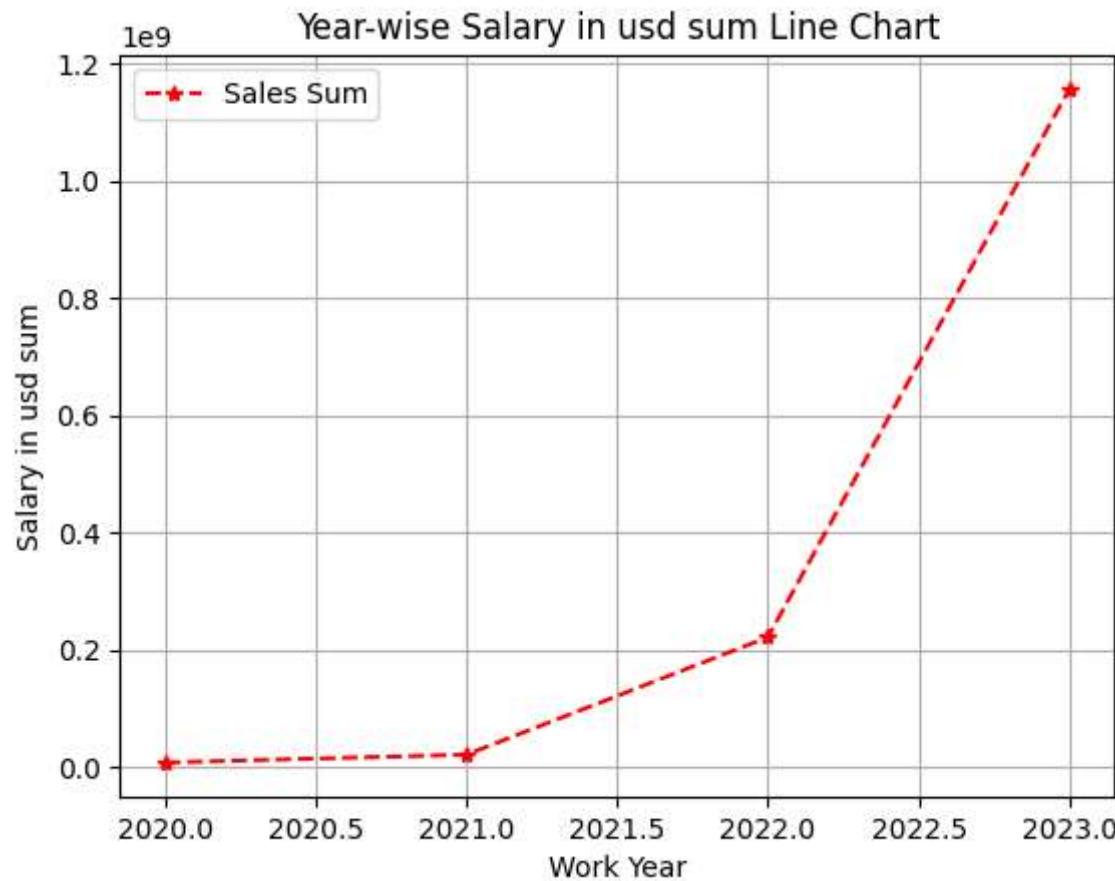
```
In [ ]: What is the relationship between company size and the salary distribution?
```

```
In [59]: csv_file_path = 'jobs in data science.csv'  
df = pd.read_csv(csv_file_path)  
salary_by_company_size = df.groupby('company_size')['salary_in_usd'].median().sort_values(ascending=False)  
salary_by_company_size.plot(kind='bar', color='orange')  
plt.title('Salary Distribution Across Company Sizes')  
plt.xlabel('Company Size')  
plt.ylabel('Median Salary in USD')  
plt.show()
```



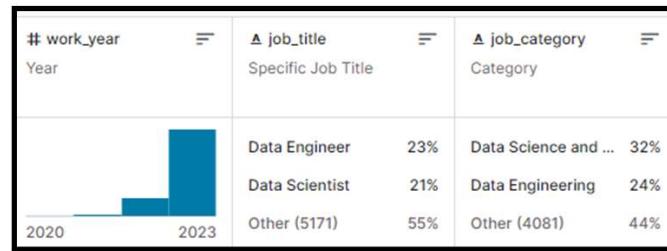
```
In [ ]: How the salary trend changes with recent years ?
```

```
In [60]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
yearly_salary = df.groupby('work_year')['salary_in_usd'].sum().reset_index()
plt.plot(yearly_salary['work_year'], yearly_salary['salary_in_usd'], marker='*', linestyle='--', color='r', label='Sales Sum')
plt.xlabel('Work Year')
plt.ylabel('Salary in usd sum')
plt.title('Year-wise Salary in usd sum Line Chart')
plt.grid(True)
plt.legend()
plt.show()
```

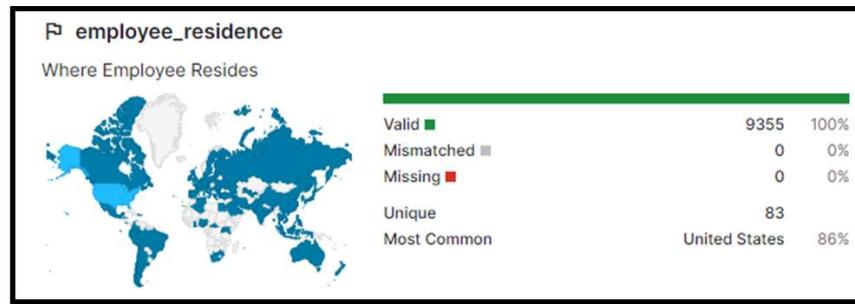


```
In [ ]: DATA VISUALIZATION:
```

Work year showing different job category



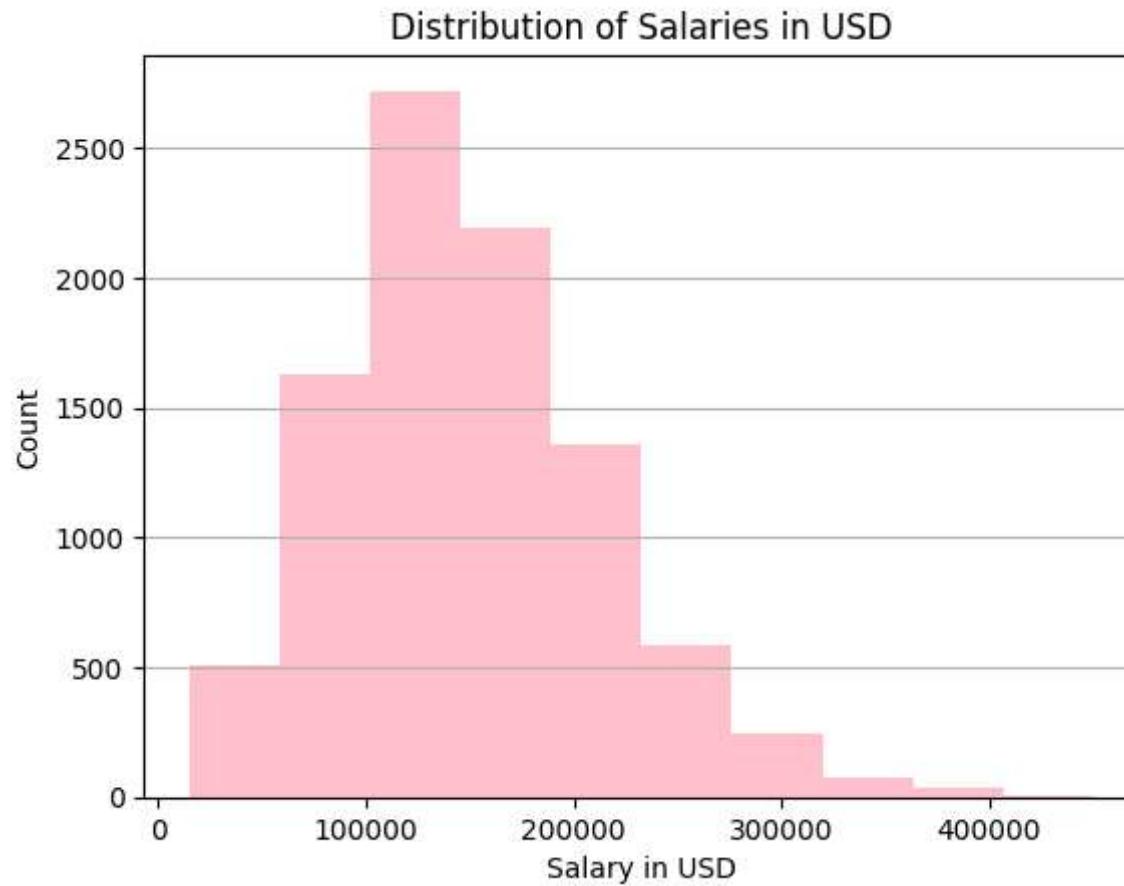
In []: Geographical Analysis:



In []: Job distribution in a graph is visualized.

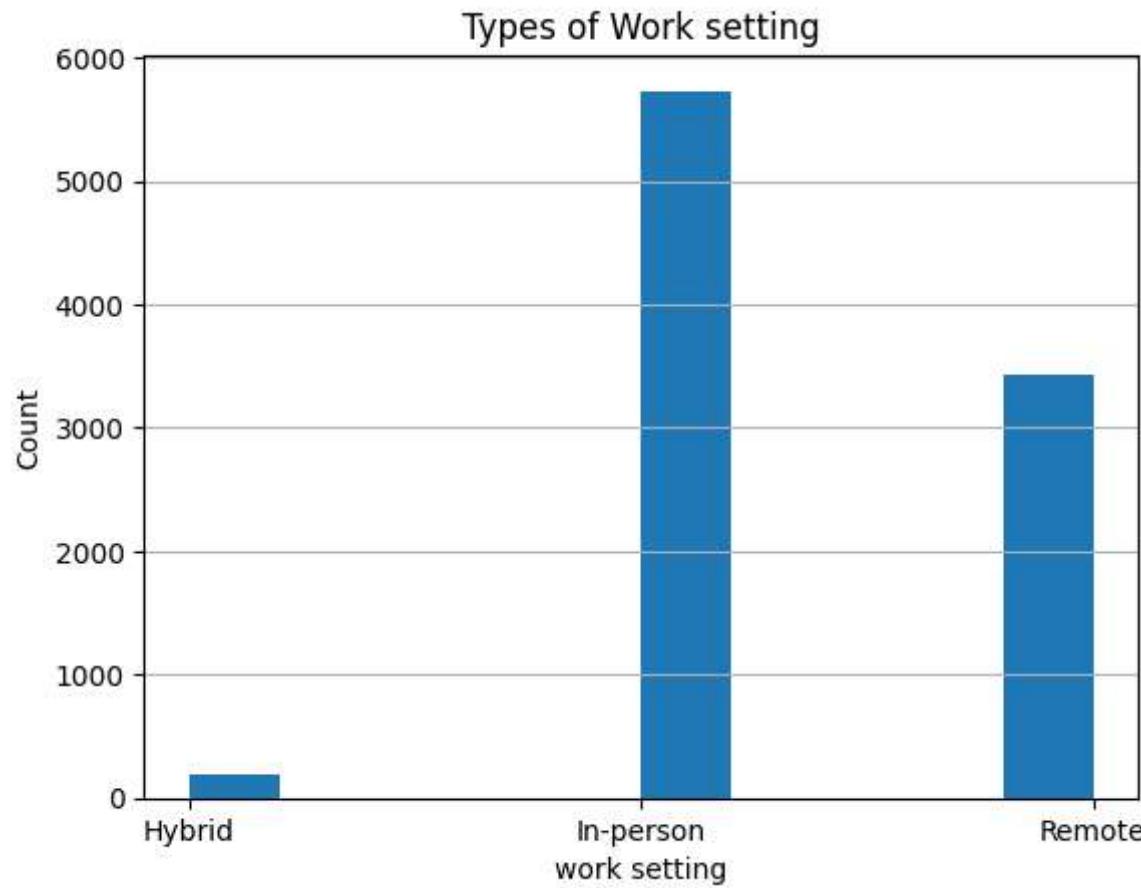
Salary Trends:

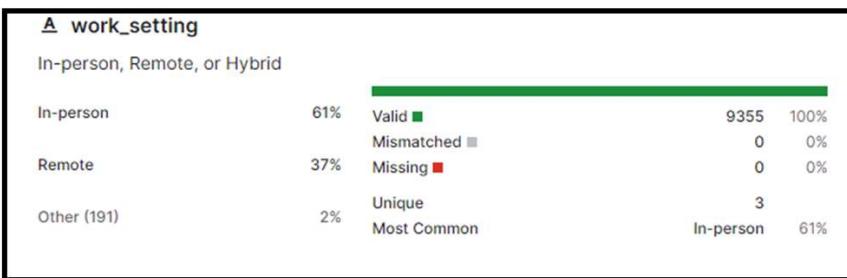
```
In [63]: csv_file_path = 'jobs_in_data_science.csv'
df = pd.read_csv(csv_file_path)
plt.hist(df['salary_in_usd'], bins=10, color='pink')
plt.xlabel('Salary in USD')
plt.ylabel('Count')
plt.title('Distribution of Salaries in USD')
plt.grid(axis='y')
plt.show()
```



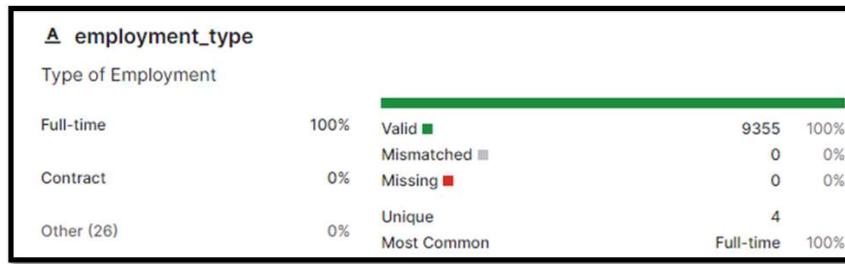
In []: Work Setting

```
In [69]: csv_file_path = 'jobs in data science.csv'
df = pd.read_csv(csv_file_path)
plt.hist(df['work_setting'],bins=10)
plt.xlabel('work setting')
plt.ylabel('Count')
plt.title('Types of Work setting')
plt.grid(axis='y')
plt.show()
```





In []: Employment type:



In []: CONCLUSION:

The demand **for** data science talent continues to outpace supply, leading to a competitive job market where organizations value technical proficiency **and** domain-specific knowledge. As businesses increasingly rely on data to drive decision-making, it becomes integral across sectors like finance, healthcare, retail, **and** technology.

In conclusion, the rise of data science jobs represents a fundamental shift **in** how businesses harness information to drive innovation. The evolution of these roles reflects a growing recognition of the value of data-driven insights **and** the pivotal role played by data scientists in extracting meaningful information **from** vast datasets to drive innovation **and** strategic decision-making.

FUTURISTIC APPROACH:

The field of data analytics **is** experiencing a rapid expansion, **with** a strong demand **for** skilled professionals who can analyze vast amounts of data.

As we progress into **2023**, a number of emerging trends are reshaping the landscape of data analytics careers. Hence, one can expect to see a significant increase in the demand for data analysts, data engineers, and machine learning specialists. The arena of jobs **and** careers **in** both remote **and** foreign areas.