Assignment 1

Name: Goh Jia Xuan
Monash ID: 33589666

Open file:

In [68]:

```
import pandas as pd
# read salaries.csv file
salary = pd.read_csv("salaries.csv")
pd.set_option('display.max_rows',10)
salary
```

Out[68]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	AI Scientist	1500000	ILS	427820	IL	0	IL	L
1	2023	SE	FT	Machine Learning Engineer	216000	USD	216000	US	100	US	M
2	2023	SE	FT	Machine Learning Engineer	184000	USD	184000	US	100	US	M
3	2023	SE	FT	Data Engineer	180000	USD	180000	US	100	US	M
4	2023	SE	FT	Data Engineer	165000	USD	165000	US	100	US	M
...	...	...	...	...	...	...	...	...	...	...	...
3222	2020	SE	FT	Data Scientist	412000	USD	412000	US	100	US	L
3223	2021	MI	FT	Principal Data Scientist	151000	USD	151000	US	100	US	L
3224	2020	EN	FT	Data Scientist	105000	USD	105000	US	100	US	S
3225	2020	EN	CT	Business Data Analyst	100000	USD	100000	US	100	US	L
3226	2021	SE	FT	Data Science Manager	7000000	INR	94665	IN	50	IN	L

3227 rows × 11 columns

Task A

A1. Dataset size

Code:

In [57]:

```
# find number of rows and columns
salary.shape
```

Out[57]:

```
(3227, 11)
```

Answer :

Data instances: 3227
Variables: 11

A2. Data Auditing

Code:

first 8 rows

```
In [58]: salary.head(8)
```

Out[58]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	AI Scientist	1500000	ILS	427820	IL	0	IL	L
1	2023	SE	FT	Machine Learning Engineer	216000	USD	216000	US	100	US	M
2	2023	SE	FT	Machine Learning Engineer	184000	USD	184000	US	100	US	M
3	2023	SE	FT	Data Engineer	180000	USD	180000	US	100	US	M
4	2023	SE	FT	Data Engineer	165000	USD	165000	US	100	US	M
5	2023	SE	FT	Data Scientist	185900	USD	185900	US	0	US	M
6	2023	SE	FT	Data Scientist	129300	USD	129300	US	0	US	M
7	2023	SE	FT	Data Engineer	145000	USD	145000	US	0	US	M

last 12 rows

```
In [59]: salary.tail(12)
```

Out[59]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
3215	2020	MI	FT	Data Engineer	130800	USD	130800	ES	100	US	M
3216	2020	SE	FT	Machine Learning Engineer	40000	EUR	45618	HR	100	HR	S
3217	2021	SE	FT	Director of Data Science	168000	USD	168000	JP	0	JP	S
3218	2021	MI	FT	Data Scientist	160000	SGD	119059	SG	100	IL	M
3219	2021	MI	FT	Applied Machine Learning Scientist	423000	USD	423000	US	50	US	L
...	...	...	...	...	...	...	...	...	...	...	...
3222	2020	SE	FT	Data Scientist	412000	USD	412000	US	100	US	L
3223	2021	MI	FT	Principal Data Scientist	151000	USD	151000	US	100	US	L
3224	2020	EN	FT	Data Scientist	105000	USD	105000	US	100	US	S
3225	2020	EN	CT	Business Data Analyst	100000	USD	100000	US	100	US	L
3226	2021	SE	FT	Data Science Manager	7000000	INR	94665	IN	50	IN	L

12 rows × 11 columns

random 6 rows

In [60]:
```
salary.sample(n=6)
```

Out[60]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
2240	2022	SE	FT	Data Scientist	191475	USD	191475	US	100	US	M
1438	2022	EN	FT	Data Analyst	55000	USD	55000	US	0	US	M
2645	2022	SE	FT	Data Engineer	155000	USD	155000	US	100	US	M
670	2023	EN	FT	Data Analyst	60000	USD	60000	US	100	US	M
16	2023	SE	FT	Data Analyst	128500	USD	128500	US	0	US	M
1278	2023	SE	FT	Data Engineer	160000	USD	160000	US	100	US	M

A3. Date Types

Code:

In [61]:
```
# get data types for each column
salary.dtypes
```

Out[61]:
```
work_year              int64
experience_level      object
employment_type       object
job_title             object
salary                 int64
                       ...  
salary_in_usd          int64
employee_residence    object
remote_ratio           int64
company_location      object
company_size          object
Length: 11, dtype: object
```

Answer:

The data type for each column is shown above. The left column represent each column name and the right column represent their data type.Data type int64 represents a 64-bit integer and data type 'object' represents strings or mixture of different data types.

A4. Conversion

1. Code:

In [62]:
```
# convert salary in usd into Malaysian ringgit
salary_in_myr= salary["salary_in_usd"] * 4.47
salary_in_myr
```

Out[62]:

```
0       1912355.40
1        965520.00
2        822480.00
3        804600.00
4        737550.00
           ...
3222    1841640.00
3223     674970.00
3224     469350.00
3225     447000.00
3226     423152.55
Name: salary_in_usd, Length: 3227, dtype: float64
```

1. Code:

In [63]:

```
# create new column store salary in myr
salary['salary_in_myr'] = salary_in_myr

# show salary in usd column and salary in myr column
salary[['salary_in_usd','salary_in_myr']]
```

Out[63]:

	salary_in_usd	salary_in_myr
0	427820	1912355.40
1	216000	965520.00
2	184000	822480.00
3	180000	804600.00
4	165000	737550.00
...	...	...
3222	412000	1841640.00
3223	151000	674970.00
3224	105000	469350.00
3225	100000	447000.00
3226	94665	423152.55

3227 rows × 2 columns

Answer:

Table above shows the employees' salary in USD and salary in Malaysian Ringgit.

A5. Descriptive Statistics

1. Calculate summary statistics

Code:

In [64]:
```
# find summary statistics of salary
salary.describe()
```

Out[64]:

	work_year	salary	salary_in_usd	remote_ratio	salary_in_myr
count	3227.000000	3.227000e+03	3227.000000	3227.000000	3.227000e+03
mean	2022.273939	1.950125e+05	134750.294391	48.280136	6.023338e+05
std	0.693571	7.226896e+05	62597.458016	48.546623	2.798106e+05
min	2020.000000	6.000000e+03	5132.000000	0.000000	2.294004e+04
25%	2022.000000	9.500000e+04	92350.000000	0.000000	4.128045e+05
50%	2022.000000	1.350000e+05	130026.000000	50.000000	5.812162e+05
75%	2023.000000	1.796375e+05	172347.500000	100.000000	7.703933e+05
max	2023.000000	3.040000e+07	450000.000000	100.000000	2.011500e+06

2. Discuss at least two observation

Answer:

From my observation,this dataset has collected employee's salary in the period of 2020 to 2023.

Regarding salary distribution within this preriod of time, the standard deviation of salary in usd is high, representing the salary range is exceptionally wide. Employee's minimum salary in USD is 5132 USD while maximum salary is up to 450000 USD. There is various factors that influence this diversity to happen, such as working experience, job title, company location and company size. For instance, employee with longer working experiences should have higher salary than others with short working experiences as they would expected the experienced one is skillful enough to handle their work and complete them prior to deadline,Workipedia (n.d).

Besides, this dataset also shows that the average of employee's remote ratio is 48.280136, indicate that approximately 48.28% of employees work from home within the period of 2020 to 2023. This value may depends on several factors such as employee's preference, distance between employee's residence and company location, not to mention the outbreak of Covid-19 around the world.For instance, an employee that live far away from company should have a relatively higher remote ratio as compare to employee that live nearby their company.

Workipedia. (n.d). *Is expected salary determined by skills, title or experience?* https://content.mycareersfuture.gov.sg/salary-determined-skills-title-experience/

A6. Exploring Job Titles

1. Number of unique job titles recorded

Code:

In [65]:
```
# find the number of rows with unique job titles
unique_job_title = salary["job_title"].value_counts()
num_job_title = len(unique_job_title)
num_job_title
```

Out[65]:
```
85
```

Answer:

1. Number of unique job titles recorded: 85

2. Show all different job titles and number of instances recorded

Code:

In [69]:

```
# result obtained from A6.1 dataframe
unique_job_title
```

Out[69]:

```
job_title
Data Engineer                        906
Data Scientist                       721
Data Analyst                         537
Machine Learning Engineer            250
Data Architect                        85
                                    ...
Manager Data Management                1
Marketing Data Engineer                1
Azure Data Engineer                    1
Applied Machine Learning Engineer      1
Finance Data Analyst                   1
Name: count, Length: 85, dtype: int64
```

Answer:

From the output above, the left column with column name "job_title" represents all different job titles and the right column represents number of instances recorded for each job title.

3. Percentage of 'Data Scientist' records

Code:

In [70]:

```
# retrieve number of instances recorded as Data Scientist
ds_job_count = unique_job_title.get('Data Scientist')
# sum up the total number of instances
total_count = unique_job_title.sum()
# calculate the percentage
percentage_ds = (ds_job_count / total_count) * 100
percentage_ds
```

Out[70]:

```
22.342733188720175
```

Answer:

Percentage of 'Data Scientist' records: 22.34%

A7. Exploring location of Companies

1. Different companies location and number of instances for each location

Code:

In [71]:

```
# group by company location and size then count the number of each location
grouped_salary = salary.groupby(["company_location",'company_size'])['company_location'].value_counts().reset_index()
```

```
# then group by company location then sum up the num of company size for each location
grouped_salary.groupby("company_location")['count'].sum().reset_index()[['company_location','count']]
```

Out[71]:

	company_location	count
0	AE	3
1	AL	1
2	AM	1
3	AR	3
4	AS	3
...	...	...
65	TH	3
66	TR	5
67	UA	1
68	US	2575
69	VN	1

70 rows × 2 columns

Answer:

The left column represents the different locations for the companies and the right column represents the number of instances for each location.

2. Total number of 'L' size companies in US

Code:

In [72]:

```
# retrive US location and L company size
grouped_salary[(grouped_salary['company_location'] == 'US') & (grouped_salary['company_size'] == 'L')]
```

Out[72]:

	company_location	company_size	count
123	US	L	227

Answer:

Total number of 'L' size companies in US: 227

Task B

B1. Investigating Employment Type

1. Job with highest salary for Full Time Employment Type

Code:

In [17]:

```
# import matplotlib library
import matplotlib.pyplot as plt
%matplotlib inline
```

In [73]:

```
# get instances with full time employement and get key columns
full_time_salary = salary[salary["employment_type"] == 'FT'][['job_title','salary_in_usd']]

# find the highest salary for each job title
grouped_salary = full_time_salary.groupby('job_title')['salary_in_usd'].max().reset_index()

# sort them in descending order based on salary
filtered_salary = grouped_salary.sort_values('salary_in_usd',ascending = False)
filtered_salary_FT = filtered_salary.head(10)
filtered_salary_FT
```

Out[73]:

	job_title	salary_in_usd
82	Research Scientist	450000
22	Data Analyst	430967
3	AI Scientist	427820
7	Applied Machine Learning Scientist	423000
42	Data Scientist	412000
25	Data Analytics Lead	405000
5	Applied Data Scientist	380000
28	Data Architect	376080
41	Data Science Tech Lead	375000
68	Machine Learning Software Engineer	375000

In [74]:

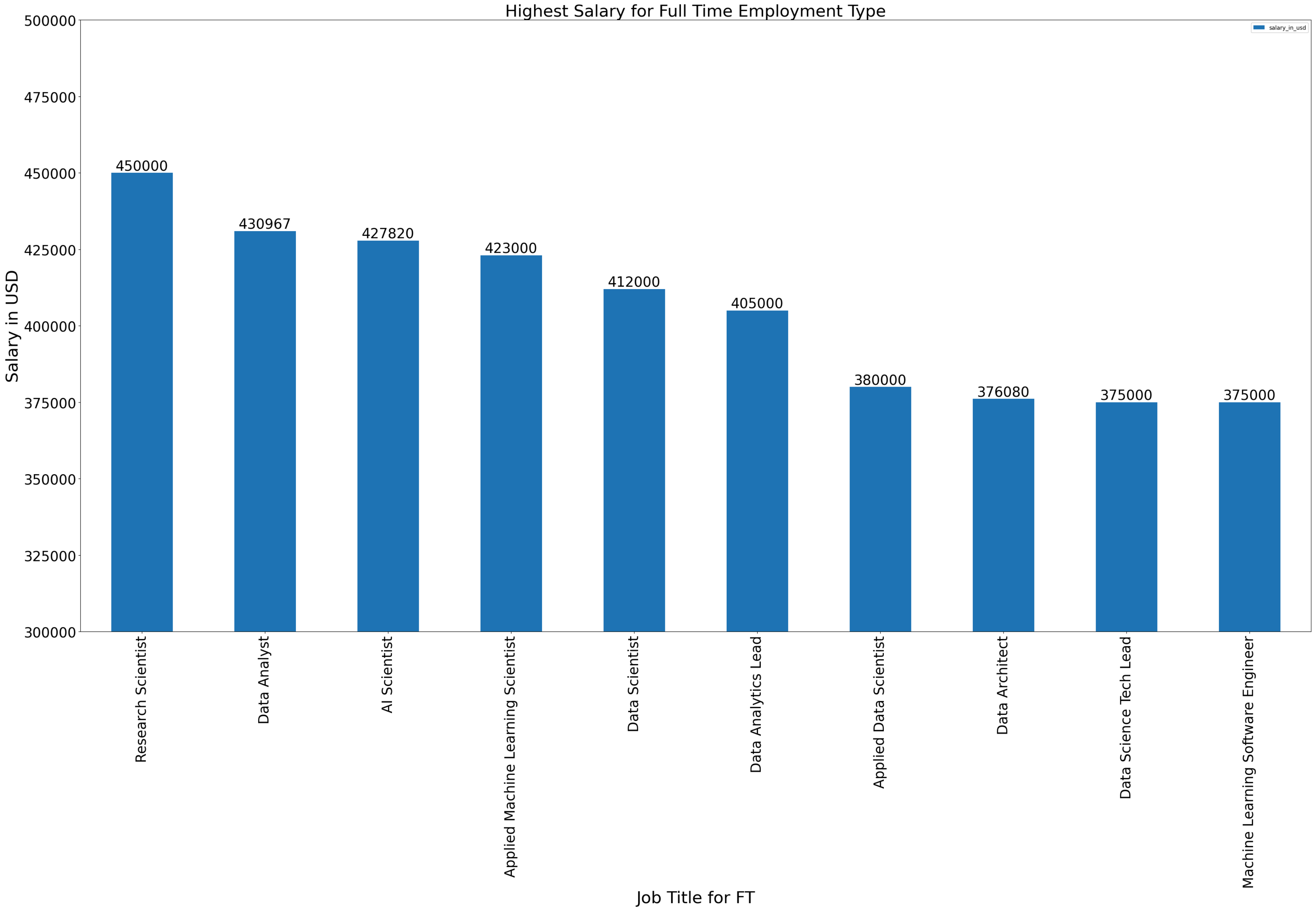
```
# plot a bar chart
ax = filtered_salary_FT.plot.bar(figsize=(40,20))
ax.set_xticklabels(filtered_salary_FT['job_title'], rotation=90)

# show the value of y axis on top of each bar
ax.bar_label(ax.containers[0],fontsize=25)
plt.xticks(fontsize=25)
plt.yticks(fontsize=25)
plt.xlabel("Job Title for FT", fontsize=30)
plt.ylabel('Salary in USD', fontsize=30)
plt.title("Highest Salary for Full Time Employment Type", fontsize=30)

# only show specific range of y-axis value
plt.ylim(300000, 500000)
```

Out[74]:

```
(300000.0, 500000.0)
```



Answer:

Highest salary job for Full Time Employment Type: Research Scientist

The bar chart above shows the highest salary of each job title for full time employment type.From the graph, it is clearly plotted that research scientist have the highest salary up to 450000 USD, which is the highest among the others.

The job position with the highest salary is influenced by several factors. According to GetEducated (n.d), research scientist is one of the fastest growing jobs, which is approximately 15% of growing potential between 2019 and 2029. This is because the utilization of computing technology is experiencing a significant increase and it lead to research scientist become high in demand. Besides, this job position requires high level of skills on understanding of research, experimentation, results analysis and etc. Therefore, they are well paid for their hard work and skills.

GetEducated. (n.d). *9 Highest paying science jobs & careers.* https://www.geteducated.com/careers/highest-paying-science-jobs/

2. Job with highest salary for Part Time Employment Type

Code:

In [75]:

```
# get instances with part time employement and get key columns
part_time_salary = salary[salary["employment_type"] == 'PT'][['job_title','salary_in_usd']]

# find the highest salary for each job title
grouped_salary = part_time_salary.groupby('job_title')['salary_in_usd'].max().reset_index()

# sort them in descending order based on salary
filtered_salary_PT = grouped_salary.sort_values('salary_in_usd',ascending = False)
filtered_salary_PT
```

Out[75]:

	job_title	salary_in_usd
5	Data Analyst	125404
7	Data Scientist	110000
6	Data Engineer	69741
3	Computer Vision Engineer	28609
4	Computer Vision Software Engineer	19073
8	ML Engineer	15966
1	AI Scientist	12000
2	BI Analyst	12000
0	3D Computer Vision Researcher	5409

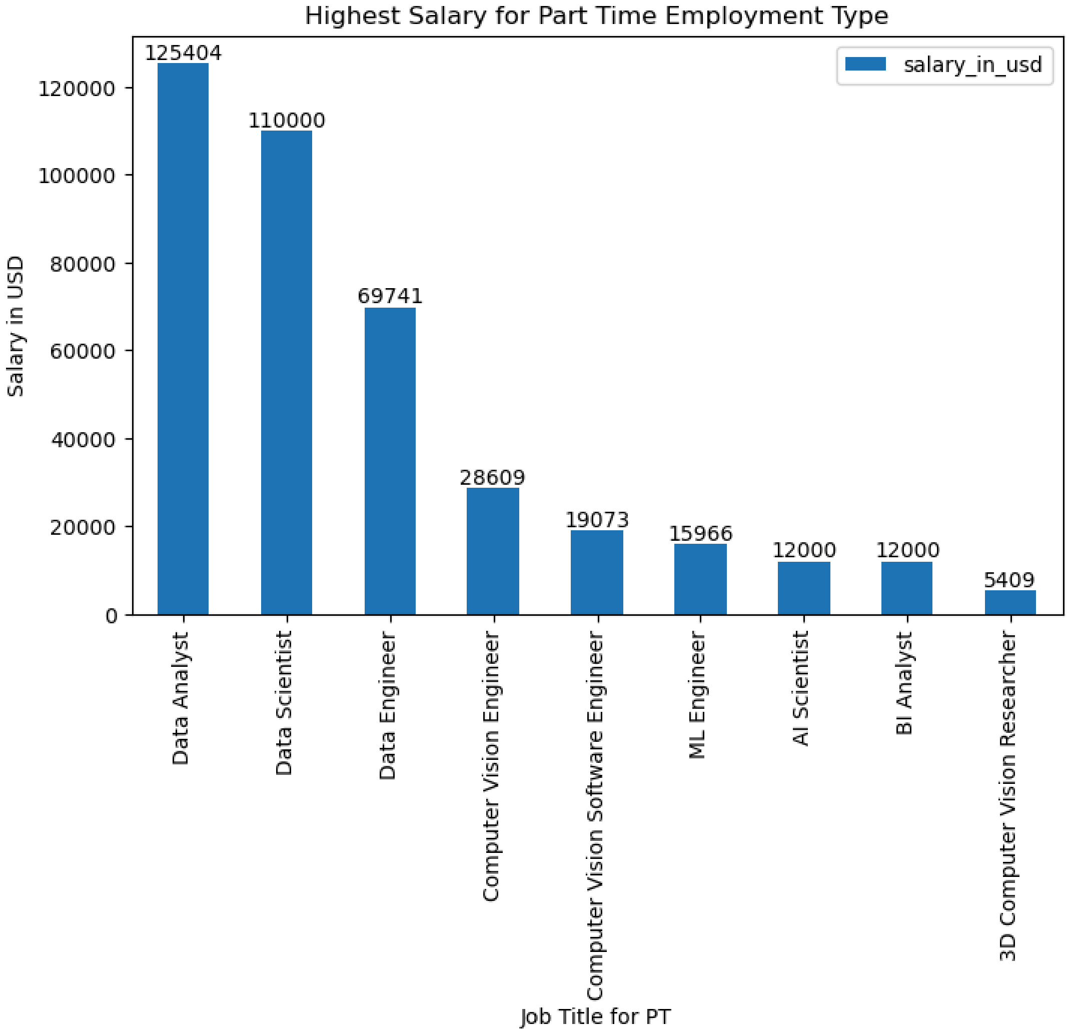
In [76]:

```
# plot a bar chart
ax = filtered_salary_PT.plot.bar(figsize=(8,5))
ax.set_xticklabels(filtered_salary_PT['job_title'], rotation=90)

# show the value of y axis on top of each bar
ax.bar_label(ax.containers[0],fontsize=10)
plt.xlabel("Job Title for PT")
plt.ylabel('Salary in USD')
plt.title("Highest Salary for Part Time Employment Type")
```

Out[76]:

```
Text(0.5, 1.0, 'Highest Salary for Part Time Employment Type')
```



Answer:

Highest salary job for Part Time Employment Type: Data Analyst

The bar chart above shows the highest salary of each job title for part time employment type.From the graph, it is clearly plotted that data analyst have the highest salary up to 125404 USD, which is the highest among the others.

The job position with the highest salary is influenced by several factors. First and foremost, Stevens.E (2023) claims that data analyst's skills can significantly affect various wordwide industries. Data is always needed and data analyst plays an important role on decision making based on huge set of data in order to drive business strategy and bring a thriving future to a company. Besides, there's also statistics prove that this job position have a high growing potential in the future as well. The U.S. BLS claims that there will be approximately 23% of growth in this role until 2031, which is far above the average (Steven.E, 2023).

Steven, E. (2023, December 12).*Am I a good fit for a career as a data analyst?* CF Blog. https://careerfoundry.com/en/blog/data-analytics/data-analyst-career-fit/#:~:text=Yes%2C%20there%20is%20a%20high,which%20is%20far%20above%20average.

3. Compare highest salary for each employment type of job from B1.1

Code:

In [77]:

```
research_scientist_salary = salary[salary['job_title'] == 'Research Scientist']
research_scientist_salary
```

Out[77]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
190	2023	SE	FT	Research Scientist	141288	USD	141288	US	0	US	M
191	2023	SE	FT	Research Scientist	94192	USD	94192	US	0	US	M
219	2023	EN	FT	Research Scientist	150000	USD	150000	US	0	US	M
220	2023	EN	FT	Research Scientist	100000	USD	100000	US	0	US	M
290	2023	MI	FT	Research Scientist	185000	USD	185000	US	100	US	M
...	...	...	...	...	...	...	...	...	...	...	...
3080	2021	MI	FT	Research Scientist	80000	CAD	63810	CA	100	CA	M
3093	2021	EN	FT	Research Scientist	100000	USD	100000	JE	0	CN	L
3125	2021	SE	FT	Research Scientist	120500	CAD	96113	CA	50	CA	L
3165	2021	MI	FT	Research Scientist	48000	EUR	56738	FR	50	FR	S
3173	2021	SE	FT	Research Scientist	50000	USD	50000	FR	100	US	S

69 rows × 11 columns

Answer:

From the dataset provided, all instances recorded for Research Scientist is a full time(FT) employment type and among all full time research scientist, the highest salary of this job position can be up to 450000 USD (found in Question B1.1)

According to this observation, we can claims that most of the research scientist would work for full time but not work as a part timer(PT), contract timer(CT) or freelancer(FL). Research scientist is responsible of creating research proposals, conducting experiments, analysing data, collaborating, writing published papers, staying up-to-date with latest scientific developments etc (Kress, n.d). Regarding working hours, research scientist typically work 35 to 40 hours a week on a full-time basis. However, due to workload, they often need to work overtime or visit the laboratory on weekends to complete certain tasks. Therefore, this might be the factor that research scientist rarely work as a part timer since they need to take part in the whole research process.

Kress, K. (n.d). *What does a research scientist fo and how do i become one>* Srg. https://www.srgtalent.com/us/blog/what-does-a-research-scientist-do-and-how-do-i-become-one

B2. Investigating Remote Ratio

1. Top 3 countries with highest recorded instances

Code:

In [78]:

```
# group the data based on company location and count the number of instances for each location
grouped_location = salary.groupby("company_location")['company_location'].value_counts().reset_index()

# sort them in descending order and retrieve the top 3 countries
country_count = grouped_location.sort_values('count',ascending = False).head(3)
country_count
```

Out[78]:

	company_location	count
68	US	2575
27	GB	159
12	CA	69

Answer:

Top three countries with highest recorded instances: US, GB, CA

2. Distribution of remote ratio

Code:

In [79]:

```
top_three_country = ['US','GB','CA']

# only obtain data where company location is from the top three countries
filtered_country = salary[salary['company_location'].isin(top_three_country)]

# group the data based on remote ratio and company location, unstack each countries into new columns
grouped_remote = filtered_country.groupby(['remote_ratio','company_location']).size().unstack().reset_index()

# rename remote ratio columns
grouped_remote['remote_ratio'].replace(0,'0-no remote work',inplace=True)
grouped_remote['remote_ratio'].replace(50,'50-partially remote',inplace=True)
grouped_remote['remote_ratio'].replace(100,'100-fully remote',inplace=True)
grouped_remote
```

Out[79]:

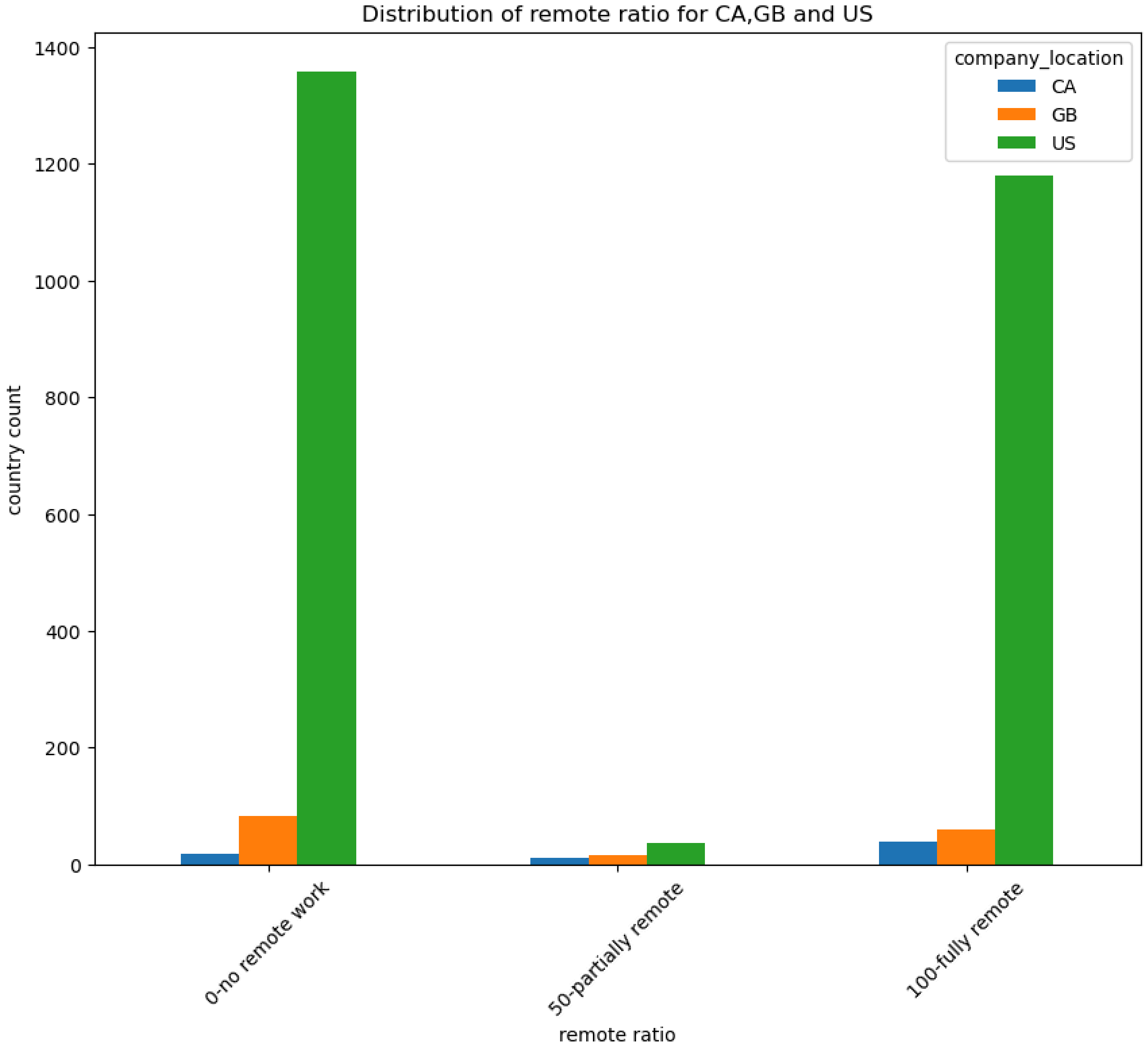
company_location	remote_ratio	CA	GB	US
0	0-no remote work	18	82	1358
1	50-partially remote	12	17	37
2	100-fully remote	39	60	1180

In [80]:

```
# plot a bar chart based on grouped_remote dataframe
ax = grouped_remote.plot.bar(figsize=(10,8))
ax.set_xticklabels(grouped_remote['remote_ratio'], rotation=45)
plt.xlabel("remote ratio")
plt.ylabel('country count')
plt.title("Distribution of remote ratio for CA,GB and US")
```

Out[80]:

```
Text(0.5, 1.0, 'Distribution of remote ratio for CA,GB and US')
```



Answer:

The bar chart suggests that the dataset predominantly comprises information from employees affiliated with companies based in US. Among CA, GB and US, employees based in US companies have the highest collection of data across all three categories, namely no remote work, partially remote and fully remote. What is more, although there's large amount of data showing that 80% of their time is working remotely(1180 records), fully in-office work still has higher records(1350 records) than working from home.

To support this result, according to Kim Parker (2023), there's statistics shows that approximately 61% of US employees do not have jobs that can be done from home. He also claims that the population of employees more likely to fall into this category is employees with lower incomes and those without a four-year college degree. Furthermore, Parker, K. (2023) also states that some employees still preferred to work in office as it helps them feel connected with collegues and have the opportunities to be mentored at work as well as give more confidence on getting work done prior to deadlines.

As a contrary, employees based in CA companies have the lowest collection of data across all categories. However, we still can see that majority of them are working from home(39 records) as compared to no remote work and partially remote. While for employees based in GB, only minority of them is partially remote(17 records) and majority of them is in-office work(82 records).

Based on this observation, there's also other sources that support to this result. There's survey found out that four in five CA public servants working remotely(Johnstone, 2023).The reason is because they find this method allows them to work more productive and have a better work-life balance. Not to mentions, we should also consider Covid-19 pandemic that happened, which involved lockdowns and stay-at-home orders to reduce interpersonal contact.

Johnstine, R. (2023). *Four in five Canadian public servants working remotely in part or in full, survey finds.* Global Government Forum. https://www.globalgovernmentforum.com/four-in-five-canadian-public-servants-working-remotely-in-part-or-in-full-survey-finds/

Parker, K. (2023).*About a third of U.S. workers who can work from homw now do so all the time.* Pew Research Center. https://www.pewresearch.org/short-reads/2023/03/30/about-a-third-of-us-workers-who-can-work-from-home-do-so-all-the-time/#:~:text=The%20majority%20of%20U.S.%20workers,to%20fall%20into%20this%20category.

B3. Investigatin Experience Level

1. Association between Experience Level and the Salary for all jobs

Code:

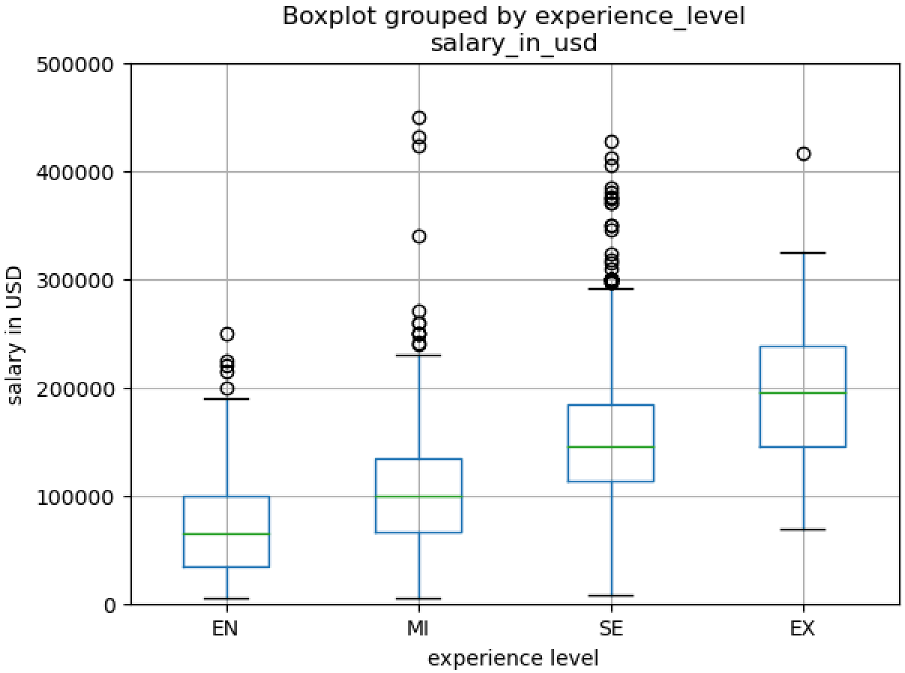
In [81]:

```
# rearrange experience level from low to high
x_order = ['EN','MI','SE','EX']
salary['experience_level'] = pd.Categorical(salary['experience_level'],categories=x_order,ordered=True)
salary.sort_values('experience_level')

# create box plot showing association between experience level and salary
salary.boxplot(column ='salary_in_usd', by='experience_level',showfliers=True)
plt.xlabel('experience level')
plt.ylabel('salary in USD')
plt.ylim(0,500000)
```

Out[81]:

```
(0.0, 500000.0)
```



Answer:

EN - Entry level/Junior

MI - Mid-level/Intermediate

SE - Senior level/Expert

EX - Executive level/Director

From the box plot created above, it can be clearly seen that all four box plots are less overlapping to each others, indicating that there is an association between experience level and employees' salary.

Based on my observation on median of each box plot, EX have the highest median salary, with an approximation of 195000 USD followed by SE which has second highest median salary, with an approximation of 150000 USD. Following the observation, the third highest is seen in MI category with an approximation of 100000 USD and lastly the lowest median salary in EN category with an approximation of 70000 USD. We can see the arrangement from highest to lowest salary as EX->SE->MI->EN.

This arrangement is also equivalent to the arrangement based on maximum value plotted on each box plot. For minimum salary in EN,MI and SE category are seemingly equivalent but EX category has significantly higher minimum salary, which is slightly higher than the median salary of EN. According to Randstad (2021), employees with higher experience level is expected to get higher paid as they are treated as mature employees that are capable of completing high level tasks based on their experiences, which indicate they have more potential to drive growth for the business. There's also another statistics from external resources that support to this statement showing that the average annual income for employees with less than one year of experience compared to those with 20+ years is on average 20404 pound vs 39199 pound (Engage Employee, n.d). This result certainly shows that experience level is one of the most prominent factors influencing the salary amount.

Apart from that, the outliers for each employment type indicate that some employees get exceptionally higher salary than the other. One of the factor could be the job title of the employee require robust skills or specialized training. Employees with in-demand skills or advanced qualification may also lead to higher compensation as compared to others with same experience level. Not to mention, outlier could also reflect the disparities in salary which could also signal the need for adjustments in salary scales.

Therefore, we can conclude that majority of employees with higher experience level are expected to have a higher salary.

Engage Employee. (n.d). https://www.engageemployee.com/blog/what-salary-should-you-be-earning-at-your-age#:~:text=Research%20shows%20that%20many%20employers,significant%20impact%20on%20your%20salary.

Randstad. (2021). https://www.randstad.com.my/career-advice/tips-and-resources/salary-based-skills-title-experience/

2. Job that Has The Highest Association between Experience Level and Salary

Code:

In [82]:

```python
# rearrange experience level from low to high
x_order = ['EN','MI','SE','EX']
salary['experience_level'] = pd.Categorical(salary['experience_level'],categories=x_order,ordered=True)
salary.sort_values('experience_level')

# only job with all 4 experience level will be used for analysis
filtered = salary.groupby('job_title')['experience_level'].nunique().reset_index()
filtered = filtered[filtered['experience_level'] == 4]
new_salary = salary[salary['job_title'].isin(filtered['job_title'])]

# find the mean salary for each job title at every experience level
new_salary = new_salary.groupby(['job_title','experience_level'])['salary_in_usd'].mean().unstack().reset_index()
new_salary
```

Out[82]:

experience_level	job_title	EN	MI	SE	EX
0	AI Scientist	52781.285714	117727.600000	202606.666667	200000.000000
1	Analytics Engineer	130000.000000	102480.230769	153088.406780	171166.666667
2	BI Data Analyst	32755.000000	69739.285714	71910.000000	150000.000000
3	Data Analyst	58956.785714	102176.109677	118822.524691	120000.000000
4	Data Engineer	96469.629630	105098.664921	151496.433657	214952.767442
5	Data Manager	61450.000000	116000.000000	119299.875000	125976.000000
6	Data Science Consultant	52098.000000	76980.400000	115375.909091	69741.000000
7	Data Scientist	73523.482143	93978.298701	156852.451292	172375.000000
8	Machine Learning Scientist	129836.000000	120566.666667	188086.363636	190000.000000
9	Research Scientist	105566.000000	136618.947368	181189.170732	84053.000000

In [83]:

```python
# only job with all 4 experience level will be used for analysis
filtered = salary.groupby('job_title')['experience_level'].nunique().reset_index()
filtered = filtered[filtered['experience_level'] == 4]
filtered = salary[salary['job_title'].isin(filtered['job_title'])]

# find the mean salary for each job title at every experience level
mean_salary = filtered.groupby(['job_title','experience_level'])['salary_in_usd'].mean().reset_index()

# find correlation of each job title
mean_salary['experience_level_num'] = mean_salary['experience_level'].replace({'EN':1,'MI':2,'SE':3,'EX':4})
grouped_salary = mean_salary.groupby('job_title').apply(lambda x:x['experience_level_num'].corr(x['salary_in_usd'])).reset_index()
grouped_salary.rename(columns={0:'correlation'},inplace=True)
```

```
# filter out correlation < 0
result = grouped_salary[grouped_salary['correlation'] > 0]
result = result.sort_values("correlation",ascending = False)
result
```

Out[83]:

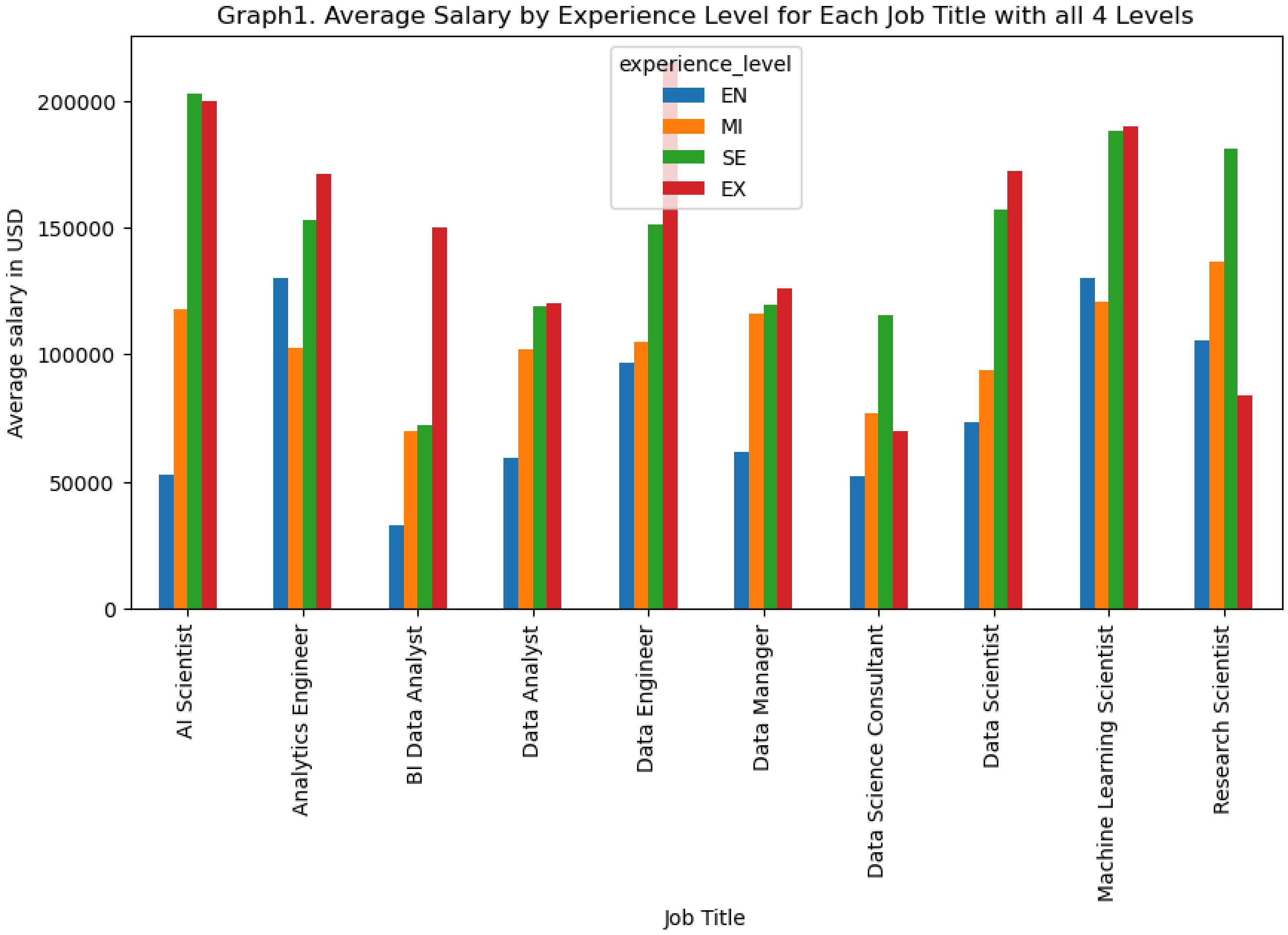
	job_title	correlation
7	Data Scientist	0.969768
4	Data Engineer	0.955317
0	AI Scientist	0.943262
2	BI Data Analyst	0.926342
3	Data Analyst	0.903683
8	Machine Learning Scientist	0.863930
5	Data Manager	0.853538
1	Analytics Engineer	0.756601
6	Data Science Consultant	0.441855

In [84]:

```
# plot bar chart for average salary
ax = new_salary.plot.bar(figsize=(10,5))
ax.set_xticklabels(new_salary['job_title'], rotation=90)
plt.xlabel("Job Title")
plt.ylabel('Average salary in USD')
plt.title("Graph1. Average Salary by Experience Level for Each Job Title with all 4 Levels")
```

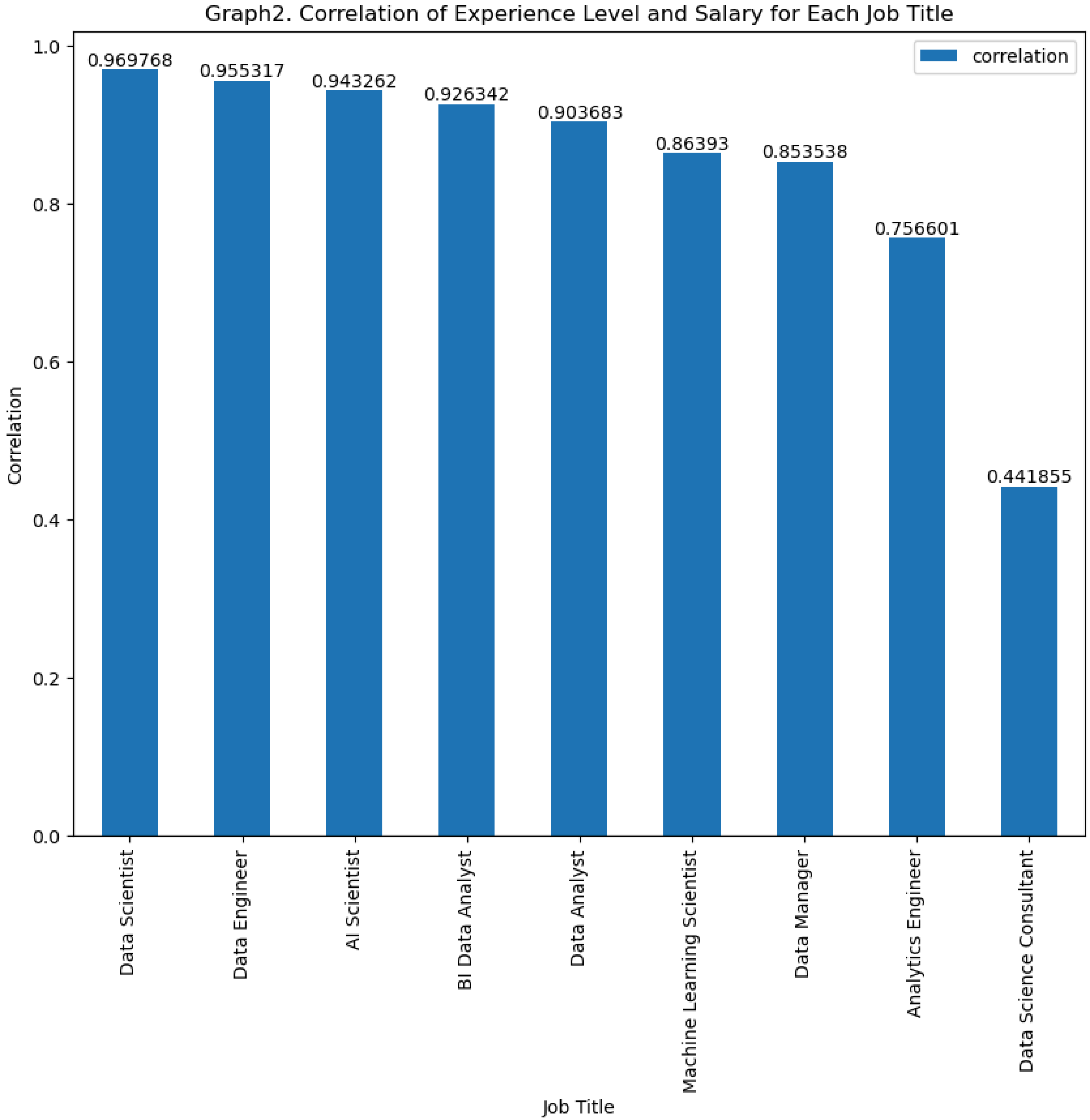
Out[84]:

```
Text(0.5, 1.0, 'Graph1. Average Salary by Experience Level for Each Job Title with all 4 Levels')
```



```
In [85]:  # plot bar chart showing the correlation of experience level and salary
           ax = result.plot.bar(figsize=(10,8))
           ax.set_xticklabels(result['job_title'], rotation=90)
           ax.bar_label(ax.containers[0])
           plt.xlabel("Job Title")
           plt.ylabel('Correlation')
           plt.title("Graph2. Correlation of Experience Level and Salary for Each Job Title")
```

```
Out[85]:  Text(0.5, 1.0, 'Graph2. Correlation of Experience Level and Salary for Each Job Title')
```



Answer:

Job with highest association between Experience Level and Salary: Data Scientist

From graph 1, it is evident that the job with the highest correlation is Data Scientist, which is approximately 0.97,showing a strong positive relationship between experience level and salary. This indicate that data scientist has the highest relevance that the more experience you have, the higher is the salary. Besides that, from graph 2, which shows the average salary by experience level for the 10 job titles with all 4 experiences level, also shows that data scientist give a positive result. A EN level data scientist has an average salary of 73523 USD, at MI level, it has an average salary of 93978 USD, at SE level it has an average salary of 156852 USD and lastly when it reaches to EX level, it can be up to an average of 172375 USD. Although there's other job title with relatively high correlation, the increment in average salary of data scientist as experience level increases is more balanced and consistent compared to other job titles. In simple word, the average salary tends to increase steadily and proportionally, without large fluctuations or inconsistencies.

Rukhaiyar, A. (2022) states that entry-level Data Scientists require 0-3 years of experience, mid-level Data Scientists require 4-8 years of experience, while senior-level Data Scientists require 9+ years of experience. According to Whitfield (2023), a junior-level Data Scientist is responsible for discovering insights through data analysis to support business development, while a senior Data Scientist has similar responsibilities but is more involved in team management and holds greater authority in long-term data-driven decisions and projects. Thus, based on the number of years they dedicate to this career, the increase in work responsibilities, and the growing demand for this job title in the future, it is certain that Data Scientist salaries are expected to increase as their experience level increases.

Rukhaiyar, A. (2022). *Data scientist job description: What to expect in 2024.* Springboard. https://www.springboard.com/blog/data-science/data-scientist-job-description/

Whitfield, B. (2023). *Senior Data Scientist.* Built-in. https://builtin.com/learn/careers/senior-data-scientist

3. Observations and comment on the distribution

Answer:

Overall, from graph plotted in Question B3.1, it can be conclude that there's an association between experience level and the salary for each job title, where higher experience level tend to have higher salary. However, we cannot prove that this statement applys on every job title. From graph 1 provided in Question B3.2, only 5 out of 10 job titles with all experience levels (which are BI Data Analyst, Data Analyst, Data Engineer, Data Scientist and Data Manager) can support on this statement while the others does not. For example, from graph 1, a data scientist consultant at senior level has higher average salary than director level and this applys on AI scientist and Research Scientist as well. From this observation, we can also says that these job position at director level is not high in demand in today society.

Furthermore, although data analyst shows the strongest positive correlation coefficient on the relationship between experience level and salary, correlation does not imply causation. Therefore, it is important to consider other factors that may influence salary amount as well. For instance, education level, certification, work performance, industry demand and negotiation skills. Additionally, as noted by Chron Contributor (2020) claims that majority of companies would implement performance-based pay strategy by giving benefits to employees that performed well in order to increase employee productivity at work. Hence, it is vital not to solely prioritize one's own experience level but also to consider other factors to maximize potential compensation and career advancement opportunities.

Chron Contributor. (2020). *How can salary influence a worker's performance in an administration?* Chron. https://work.chron.com/can-salary-influence-workers-performance-administration-25950.html

In [ ]: