Task A: Data Exploration and Auditing

A1: Dataset size

In [118…
```
import pandas as pd
salaries_report = pd.read_csv('salaries.csv')
salaries_report.shape
```

Out[118]:
```
(3227, 11)
```

Answer:

- Number of data instances (rows) : 3227
- Number of variables (columns) : 11

A2. Data auditing

In [119…
```
#printing the first 8 rows
salaries_report.head(8)
```

Out[119]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd
0	2023	SE	FT	AI Scientist	1500000	ILS	42782(
1	2023	SE	FT	Machine Learning Engineer	216000	USD	216000
2	2023	SE	FT	Machine Learning Engineer	184000	USD	184000
3	2023	SE	FT	Data Engineer	180000	USD	180000
4	2023	SE	FT	Data Engineer	165000	USD	165000
5	2023	SE	FT	Data Scientist	185900	USD	185900
6	2023	SE	FT	Data Scientist	129300	USD	129300
7	2023	SE	FT	Data Engineer	145000	USD	145000

In [120…
```
#printing the last 12 rows
salaries_report.tail(12)
```

Out[120]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in
3215	2020	MI	FT	Data Engineer	130800	USD	13
3216	2020	SE	FT	Machine Learning Engineer	40000	EUR	4
3217	2021	SE	FT	Director of Data Science	168000	USD	16
3218	2021	MI	FT	Data Scientist	160000	SGD	11
3219	2021	MI	FT	Applied Machine Learning Scientist	423000	USD	42
3220	2021	MI	FT	Data Engineer	24000	EUR	2
3221	2021	SE	FT	Data Specialist	165000	USD	16
3222	2020	SE	FT	Data Scientist	412000	USD	41
3223	2021	MI	FT	Principal Data Scientist	151000	USD	15
3224	2020	EN	FT	Data Scientist	105000	USD	10
3225	2020	EN	CT	Business Data Analyst	100000	USD	10
3226	2021	SE	FT	Data Science Manager	7000000	INR	9

In [121…

```
#printing the random 6 rows
salaries_report.sample(6)
```

Out[121]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_u
1403	2022	SE	FT	Applied Scientist	164000	USD	164(
415	2023	SE	FT	Machine Learning Engineer	126000	USD	126(
1983	2022	SE	FT	Data Engineer	150000	USD	150(
2593	2022	SE	FT	Data Scientist	160000	USD	160(
1672	2022	SE	FT	Data Engineer	40000	EUR	42(
1667	2022	MI	FT	Data Analyst	150000	USD	150(

A3. Data Types

In [122…

```
salaries_report.dtypes
```

Out[122]:

```
work_year              int64
experience_level      object
employment_type       object
job_title             object
salary                 int64
salary_currency       object
salary_in_usd          int64
employee_residence    object
remote_ratio           int64
company_location      object
company_size          object
dtype: object
```

Answer:

For the column 'work_year', 'salary', 'salary_in_usd' and 'remote_ratio' have integer data type. For the column 'experience_level', 'employment_type', 'job_title', 'salary_currency', 'employee_residence', 'company_location', 'company_size' have 'object' data type.

A4. Conversion

In [123…

```
salaries_report['salary_in_myr'] = salaries_report['salary_in_usd'] * 4.47
salaries_report.head()
```

Out[123]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd
0	2023	SE	FT	AI Scientist	1500000	ILS	42782(
1	2023	SE	FT	Machine Learning Engineer	216000	USD	216000
2	2023	SE	FT	Machine Learning Engineer	184000	USD	184000
3	2023	SE	FT	Data Engineer	180000	USD	180000
4	2023	SE	FT	Data Engineer	165000	USD	165000

A5. Descriptive Statistics

In [124…
```
salaries_report.describe()
```

Out[124]:

	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

Answer:

From the table generated above, it shows that there is a total of 3227 data in the file. This means the data is collected from all 3227 different respondent. The data collected is between year 2020 and 2023 by referring to the min and max value within work_year column. The minimum salary in USD is 5132, where it is salary of a full time NLP Engineer. The maximum salary in USD is 450000, which referring to the salary of a full time Research Scientist.

A6. Exploring Job Titles

In [125…
```
# Question 1
salaries_report['job_title'].nunique()
```

Out[125]:
```
85
```

Answer:

- 85 different job titles

In [126…
```
# Question 2
salaries_report['job_title'].value_counts()
```

Out[126]:
```
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:

The output shown above illustrate there is a total of 85 different job titles and the number of instances recorded for each job. For example, Data Engineer has 906 records.

In [127…
```
# Question 3
num_ds = salaries_report['job_title'].value_counts().get('Data Scientist')
total = salaries_report['job_title'].count()
percentage = (num_ds/total)* 100
percentage
```

Out[127]:
```
22.342733188720175
```

Answer:

The percentage of 'Data Scientist' records in 'job_title' column is approximately 22.3%.

A7. Exploring location of Companies

In [128…
```
# Question 1
salaries_report['company_location'].value_counts()
```

Out[128]:
```
company_location
US    2575
GB     159
CA      69
ES      68
IN      54
      ...
MA       1
SK       1
AL       1
BS       1
MT       1
Name: count, Length: 70, dtype: int64
```

Answer:

The output shown above illustrate there is a total of 70 different locations for the companies and the number of instances recorded for each location. For example, US have 2575 records.

In [129…

```
# Question 2
filt = (salaries_report['company_location']== 'US') & (salaries_report['company_siz
salaries_report[filt]
```

Out[129]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_
101	2023	SE	FT	Data Engineer	205600	USD	205
102	2023	SE	FT	Data Engineer	105700	USD	105
131	2023	SE	FT	Cloud Database Engineer	170000	USD	170
132	2023	SE	FT	Applied Machine Learning Scientist	90000	USD	90
173	2023	SE	FT	Applied Scientist	222200	USD	222
...	...	...	...	...	...	...	
3219	2021	MI	FT	Applied Machine Learning Scientist	423000	USD	423
3221	2021	SE	FT	Data Specialist	165000	USD	165
3222	2020	SE	FT	Data Scientist	412000	USD	412
3223	2021	MI	FT	Principal Data Scientist	151000	USD	151
3225	2020	EN	CT	Business Data Analyst	100000	USD	100

227 rows × 12 columns

Answer:

- Total number of companies = 227

Task B: Group Level Analysis and Visualisation

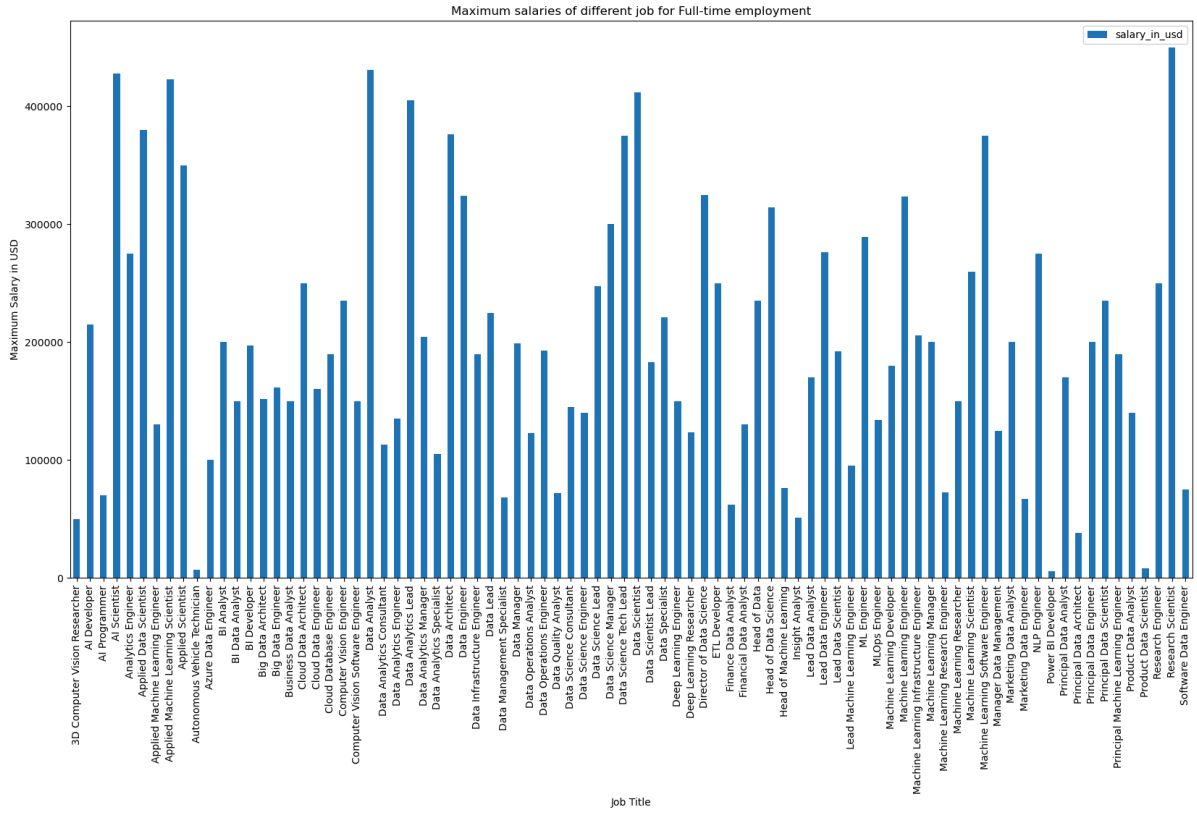
B1. Investigating Employment Type

In [130…

```
# Question 1
import matplotlib.pyplot as plt

# filter to get the Full-time employment type
```

```
filt = (salaries_report['employment_type'] == 'FT')
filt_salaries_report = salaries_report[filt]

fun = {'salary_in_usd':'max'}

# form a table which group by the job_title
max_salary = filt_salaries_report.groupby('job_title').agg(fun).reset_index()

# plot a bar graph using the table formed
ax = max_salary.plot.bar(figsize=(20,10))

# customise x axis tick labels
ax.set_xticklabels(max_salary['job_title'], rotation=90)

# renaming the x-axis, y-axis and graph title
plt.xlabel('Job Title')
plt.ylabel('Maximum Salary in USD')
plt.title('Maximum salaries of different job for Full-time employment')
```

Out[130]: Text(0.5, 1.0, 'Maximum salaries of different job for Full-time employment')



Answer:

The bar graph above displays the maximum salary of different jobs for Full-time (FT) employment type. From the graph, it can be clearly seen that 'Research Scientist' receive the highest salary among all. This is because Research Scientist are in high demand among society because they are experts in scientific literature, and their work helps advance technology, medicine, and biology. Furthermore, research scientists who specialise in advanced technologies or specialised fields may be able to get greater compensation and have an advantage over other candidates on the job market. Next, follow by Data Analyst which in charge in collect and analyze data to create information that companies can use to grow and improve to have better understanding in their trends.

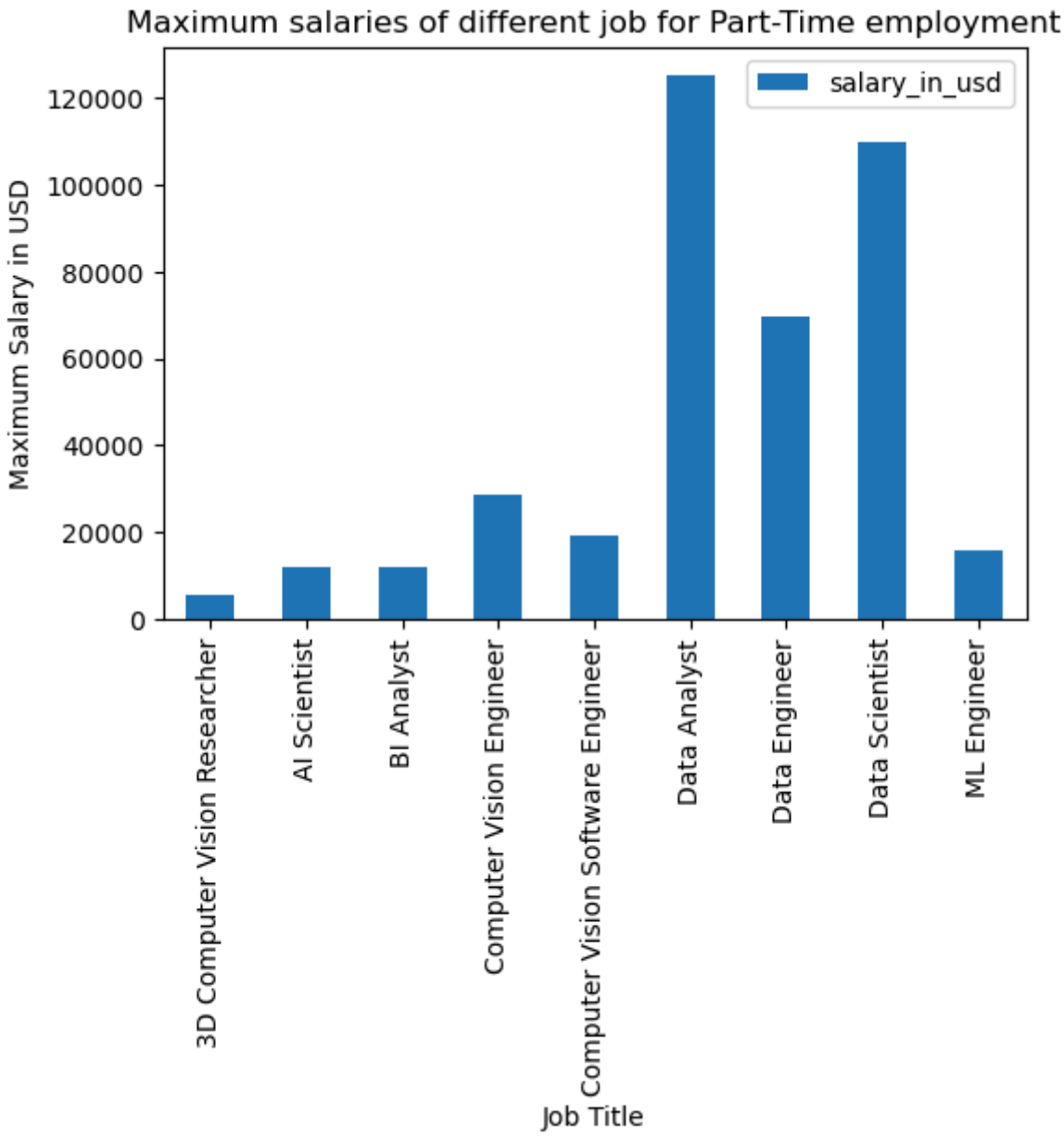
When examining the lowest salary levels, job positions such as Autonomous Vehicle Technician, Power BI Developer, and Product Data Scientist tend to fall towards the lower end of the salary spectrum. These might be due to low demand in society compared to other jobs.

In [131…

```
# Question 2

# filter to get the Part-time employment type
filt2 = (salaries_report['employment_type'] == 'PT')
filt_salaries_report2 = salaries_report[filt2]

fun2 = {'salary_in_usd':'max'}
# form a table which group by job title
max_salary2 = filt_salaries_report2.groupby('job_title').agg(fun2).reset_index()

# plot a bar graph using the table formed
ax2 = max_salary2.plot.bar(figsize=(6,4))

# renaming the x-axis, y-axis and graph title
ax2.set_xticklabels(max_salary2['job_title'], rotation=90)
plt.xlabel('Job Title')
plt.ylabel('Maximum Salary in USD')
plt.title('Maximum salaries of different job for Part-Time employment')
```

Out[131]:

```
Text(0.5, 1.0, 'Maximum salaries of different job for Part-Time employment')
```



Answer:

The bar graph above displays the maximum salary of different jobs for Part-time (PT) employment type. From the graph, it can be clearly seen that 'Data Analyst' receive the highest salary among all. This is because they are normally assigned to project-based assignments, and higher compensation is offered based on the completion of the projects. Next, followed by Data Scientist and Data Engineer. This suggests that roles within the data-related fields tend to offer higher salaries for part-time employment opportunities.

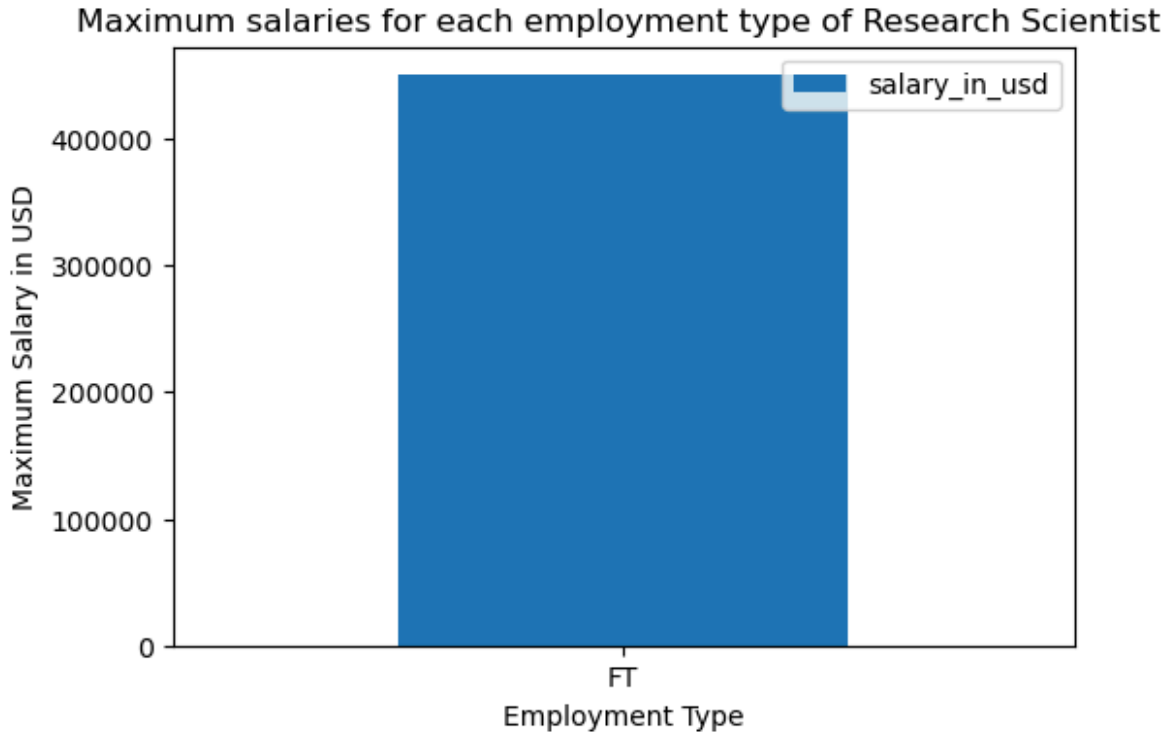
Conversely, when examining the lower end of the salary spectrum, '3D Computer Vision Researchers' offer the least salary among all positions. This may be attributed to the relatively lower demand for skills in this specialized field. Additionally, the niche nature of 3D computer vision research could result in fewer opportunities for part-time employment, potentially leading to lower salary offers.

In [132…

```
# Question 3

# filter to get the data for Research Scientist
filt3 = (salaries_report['job_title'] == 'Research Scientist')
filt_salaries_report3 = salaries_report[filt3]
```

```
fun3 = {'salary_in_usd':'max'}

# form a table which group by employment type
max_salary3 = filt_salaries_report3.groupby('employment_type').agg(fun3).reset_inde

# plot a bar graph using the table formed
ax3 = max_salary3.plot.bar(figsize=(6,4))
ax3.set_xticklabels(max_salary3['employment_type'],rotation = 0)

# log-scale to the y axis
#plt.yscale('log')

# renaming the x-axis, y-axis and graph title
plt.xlabel('Employment Type')
plt.ylabel('Maximum Salary in USD')
plt.title('Maximum salaries for each employment type of Research Scientist ')
```

Out[132]:
```
Text(0.5, 1.0, 'Maximum salaries for each employment type of Research Scientist ')
```



Answer:

The bar graph above shows the maximum salaries for each employment type of the job title 'Research Scientist'. From the graph, there is only full time employment offered for a research scientist, with no data available for other employment types such as part-time, contract, and freelance. This suggests that for Research Scientist role, companies only offer full-time positions, particularly due to the time, effort and continuous participation needed throughout research project's lifetime. Moreover, research scientist's work mode might be differ from others due to organisational policies and industry norms for their employee arrangement. These factors likely drive the preference for full-time positions, ensuring consistent commitment to research projects.

B2. Investigating Remote Ratio

In [133…
```
# Question 1
salaries_report['company_location'].value_counts()
```

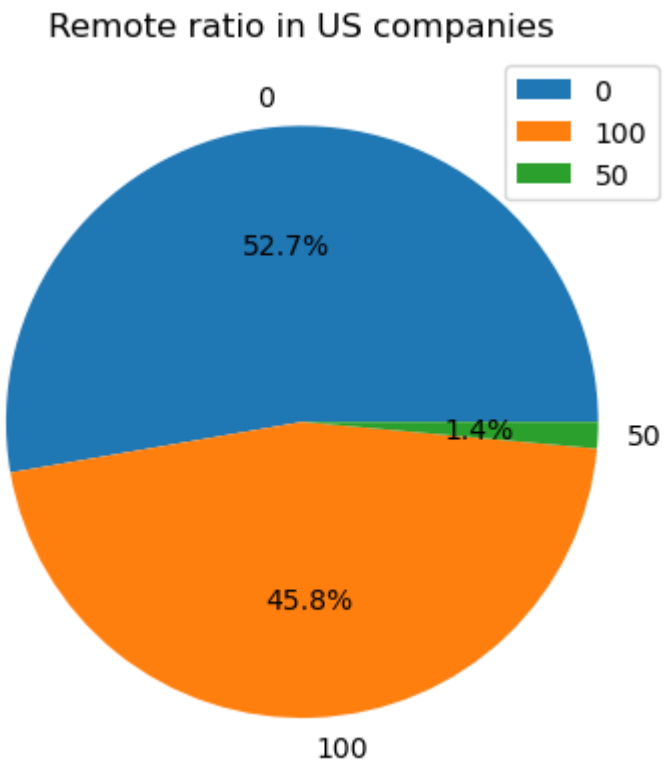
Out[133]:
```
company_location
US    2575
GB     159
CA      69
ES      68
IN      54
      ...
MA       1
SK       1
AL       1
BS       1
MT       1
Name: count, Length: 70, dtype: int64
```

Answer:

Top three countries:

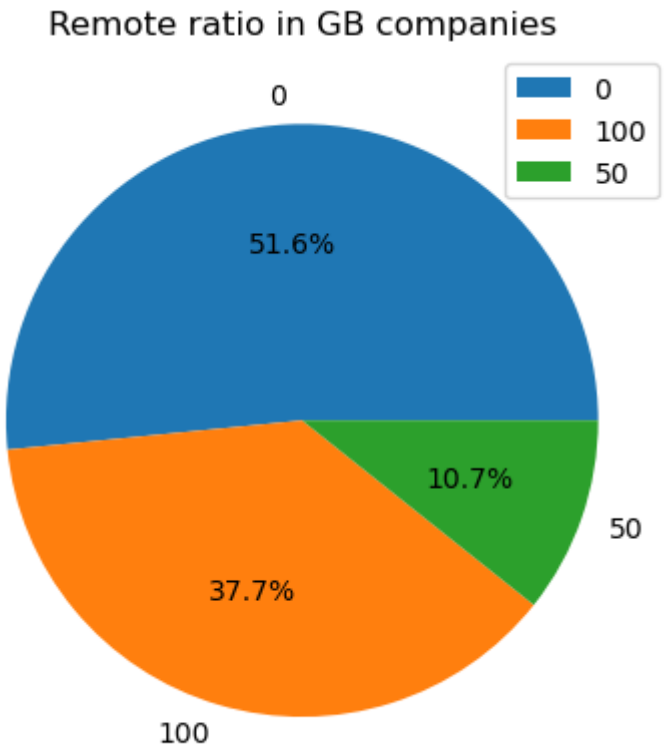
1. United States of America (US)
2. United Kingdom of Great Britain and Northern Ireland (GB)
3. Canada (CA)

In [134…
```
# Question 2

# filter to retrieve data from US company
filt4 = (salaries_report['company_location'] == 'US')
remote_US = salaries_report[filt4]

# compute the number of workers for each different remote ratio
US_table = remote_US['remote_ratio'].value_counts()

#plot the pie chart
plt.pie(US_table, labels=US_table.index, autopct='%1.1f%%')
plt.title('Remote ratio in US companies')
plt.legend(loc='upper right')
plt.show()
```



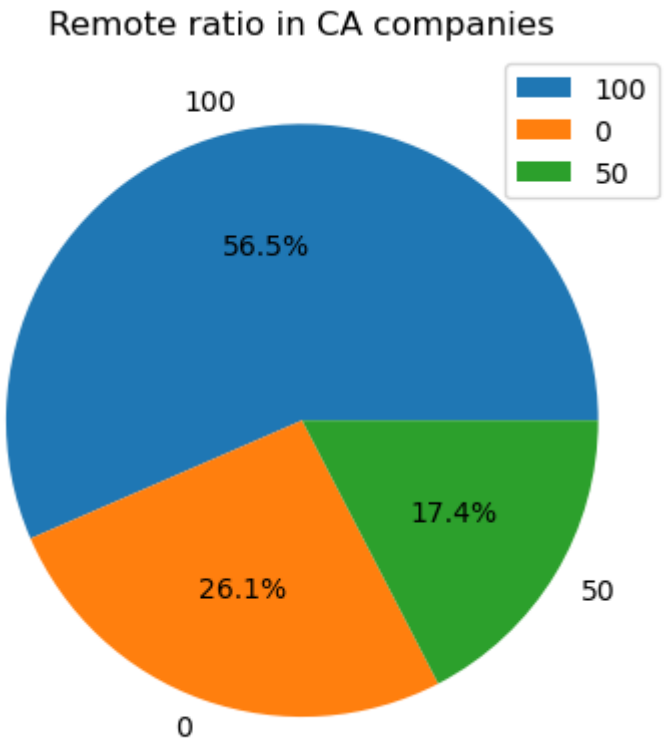
In [135…

```
# Question 2

# filter to retrieve data from US company
filt5 = (salaries_report['company_location'] == 'GB')
remote_GB = salaries_report[filt5]

# compute the number of workers for each different remote ratio
GB_table = remote_GB['remote_ratio'].value_counts()

#plot the pie chart
plt.pie(GB_table, labels= GB_table.index, autopct='%1.1f%%')
plt.title('Remote ratio in GB companies')
plt.legend(loc='upper right')
plt.show()
```



In [136…

```
# Question 2

# filter to retrieve data from CA company
filt6 = (salaries_report['company_location'] == 'CA')
remote_CA = salaries_report[filt6]

# compute the number of workers for each different remote ratio
CA_table = remote_CA['remote_ratio'].value_counts()

#plot the pie chart
plt.pie(CA_table, labels= CA_table.index, autopct='%1.1f%%')
plt.title('Remote ratio in CA companies')
plt.legend(loc='upper right')
plt.show()
```



Answer:

The three pie charts above show the remote ratio for the top three countries, United States of America (US), United Kingdom of Great Britain and Northern Ireland (GB) and Canada (CA). From the first pie chart, majority of the US workers are classified as 'No remote work'. Conversely, only 1.4% of the workers are partially remote. This shows the same pattern in GB companies, whereby 51.6% of the workers are having no remote work. For fully remote, there are 37.7% and partially remote workers' proportion are slightly greater than US by 9.3%. While in CA, most of the workers are fully remote workers. 26.1% of the workers are having no remote work and 17.4% are partially remote. This suggests a more equitable distribution between these two remote work modes in CA.

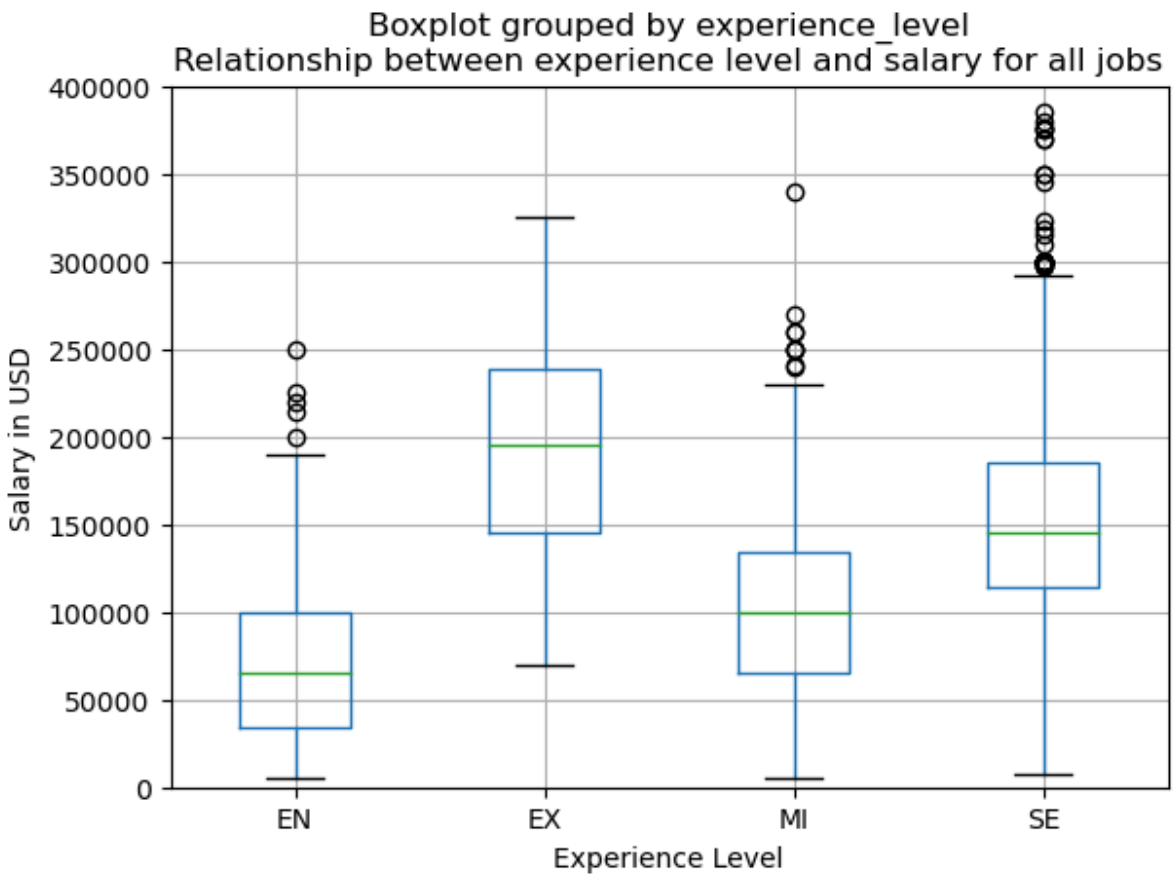
These insights highlight different remote work adoption patterns across the three countries, with US and GB showing highest proportion of workers having no remote work, CA showing most of the workers are fully remote. Overall, with the help of pie charts in providing a clear visual representation of how each category contributes to the whole, we are able to make comparison easily. This is because pie chart able to calculate the percentage directly and allow viewers to see the relative size of each category and understand its significance in relation to the whole.

B3. Investigating Experience level

In [155…

```
# Question 1

salaries_report.boxplot(column ='salary_in_usd', by = 'experience_level')
plt.xlabel('Experience Level')
plt.ylabel('Salary in USD')
plt.ylim(0,400000)
plt.title('Relationship between experience level and salary for all jobs')
```

Out[155]: Text(0.5, 1.0, 'Relationship between experience level and salary for all jobs')



Answer:

From the boxplot, there is an association between Experience Level and the Salary for all jobs. If the workers have more experience which is referring to Executive-level / Director (EX) in this case, they will get a higher pay. While Entry-level / Junior (EN) workers will get a lower pay, due to the least experience level. Moreover, by referring to the median salary of each experience level, the sorted sequence (from lowest pay to highest) are as follows:

1. Entry-level / Junior (EN)
2. Mid-level / Intermediate (MI)
3. Senior-level / Expert (SE)
4. Executive-level / Director (EX)

This act as an evidence in showing association between both experience level and salary for all jobs.

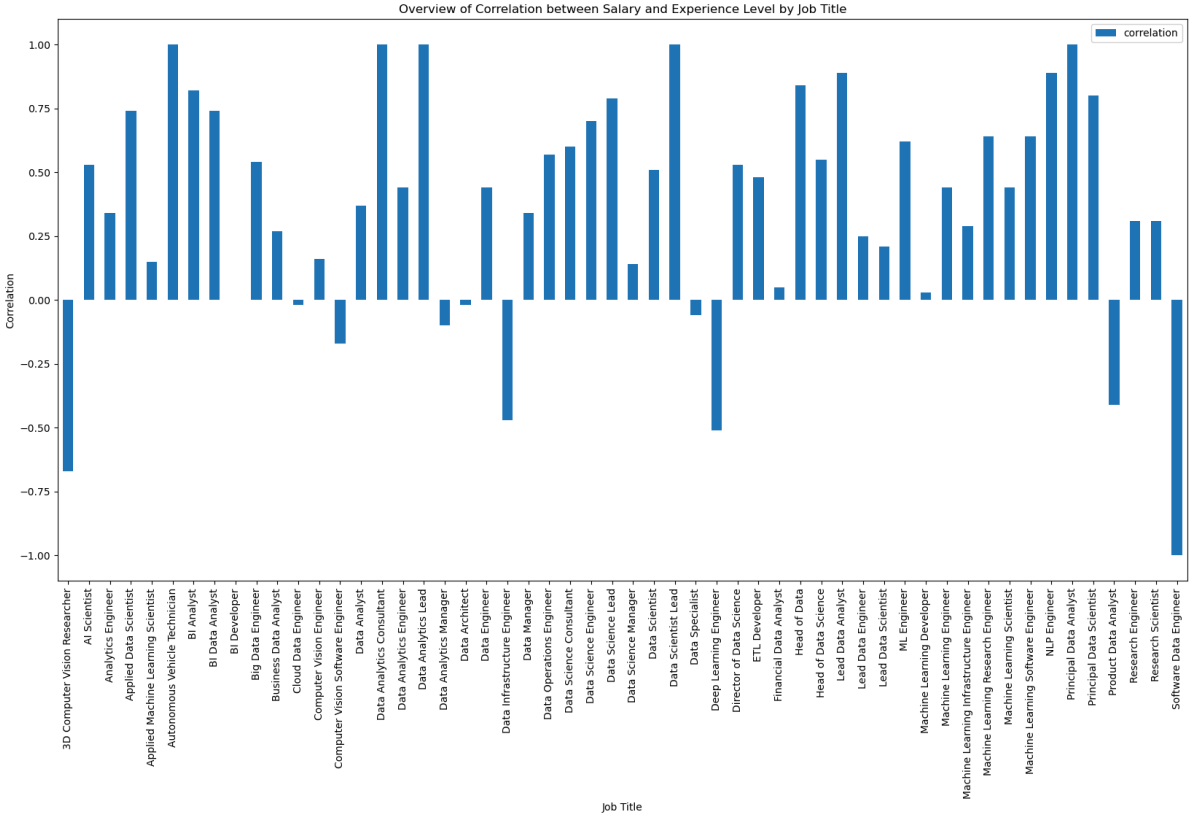
In [156…

```
# Question 2


# experience level are catrgorical ordinary data, so values are set accordingly for
experience_level_num = {'EN': 1, 'MI': 2, 'SE': 3, 'EX':4}

# extract all the columns needed
selected_columns = ['job_title', 'experience_level', 'salary_in_usd']
filtred_df = salaries_report[selected_columns]

# replacing the experience level with numbers
filtred_df.loc[:,'experience_level'] = filtred_df['experience_level'].replace(exper
```

```
# create two empty list for data storage purposes
job = []
correlation_data = []

# Group the filtered DataFrame by 'job_title'
all_group = filtred_df.groupby('job_title')

# Iterate over each group based on job title and compute correlation
for title, group_data in all_group:

    # only calculate the correlation if there is more than 1 experience level in th
    if group_data['experience_level'].nunique() > 1:
    #Compute correlation between 'salary' and 'experience_level' for the current jo
        correlation = group_data['experience_level'].corr(group_data['salary_in_usd
        job.append(title)
        correlation_data.append(correlation.round(2))

# create a new data frame tosfor graph plotting purpose
graph_correlation = pd.DataFrame({'job_title':job,'correlation':correlation_data})

# plot the graph using bar chart
ax = graph_correlation.plot.bar(figsize=(20,10))
ax.set_xticklabels(graph_correlation['job_title'], rotation=90)
plt.xlabel('Job Title')
plt.ylabel('Correlation')
plt.title('Overview of Correlation between Salary and Experience Level by Job Title
```

Out[156]:

```
Text(0.5, 1.0, 'Overview of Correlation between Salary and Experience Level by Job
Title')
```



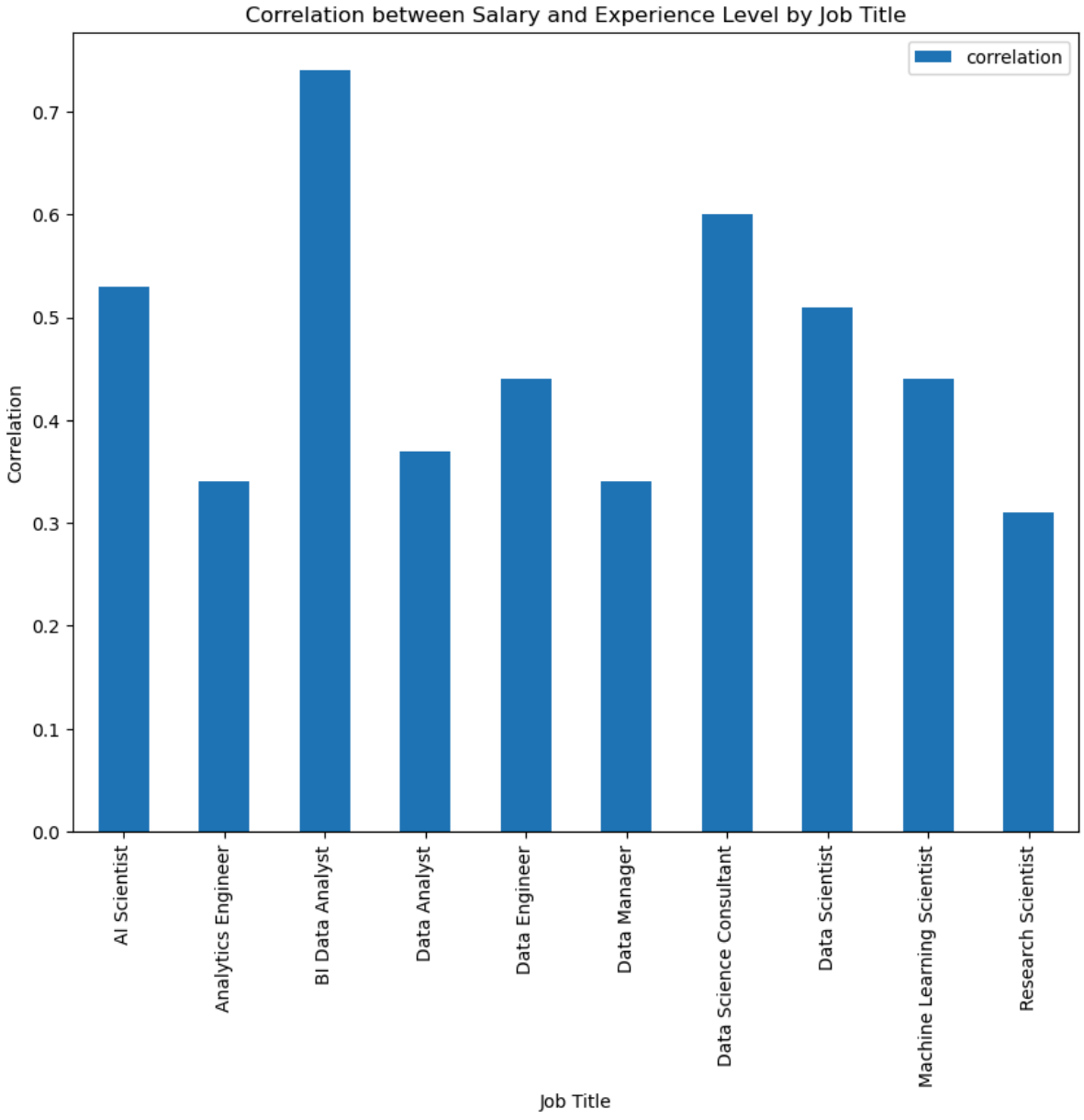
The graph above provides an overview of all jobs displaying a positive correlation between salary and experience level based on the available data. However, to get the highest association between experience level and salary, it is crucial to consider only those jobs for which data for all experience levels is available to ensure fair comparison.

In [146…

```
# create two empty list for data storage purposes
job_filter = []
```

```
correlation_data_filter = []

# Iterate over each group based on job title and compute correlation
for title, group_data in all_group:
    if group_data['experience_level'].nunique() > 3:
    #Compute correlation between 'salary' and 'experience_level' for the current jo
        correlation = group_data['experience_level'].corr(group_data['salary_in_usd
        job_filter.append(title)
        correlation_data_filter.append(correlation.round(2))

# create a new data frame tosfor graph plotting purpose
graph_correlation1 = pd.DataFrame({'job_title':job_filter,'correlation':correlation

# plot the graph using bar chart
ax = graph_correlation1.plot.bar(figsize=(10,8))
ax.set_xticklabels(graph_correlation1['job_title'], rotation=90)
plt.xlabel('Job Title')
plt.ylabel('Correlation')
plt.title('Correlation between Salary and Experience Level by Job Title')
```

Out[146]: Text(0.5, 1.0, 'Correlation between Salary and Experience Level by Job Title')



Answer:

The graph above illustrates the correlation between salary and experience level by job title. All job titles displayed have been filtered to include data for all four experience levels. This is important to ensure a more accurate correlation in order to determine the job which have the highest association between experience level and salary.

Upon analysis, the job with the highest association between experience level and salary is BI Data Analyst, showing a strong positive correlation exceeding 0.7. As they played a vital role in evaluating market strategies trough the depth analysis of the trends and products, BI Data Analysts must possess more experience to effectively navigate complex market dynamics and make informed strategic decisions. This strong correlation hightlights the importance of experience in this position. At the same time, it creates an overview for companies to focus on hiring experienced BI Data Analysts for their business goods.

In [154…

```
# plot a scatter plot which focus only on BI Data Analyst
filt = (salaries_report['job_title'] == 'BI Data Analyst')
filtered_BI = salaries_report[filt]

# replacing the experience level with numbers
experience_level_num = {'EN': 1, 'MI': 2, 'SE': 3, 'EX':4}
filtered_BI['experience_level'] = filtered_BI['experience_level'].replace(experienc

#create a new data frame with only two columns for scatter plot
selected_columns = ['experience_level','salary_in_usd']
final_filtered_BI = filtered_BI[selected_columns]

#plot the scatter plot
x = final_filtered_BI['experience_level']
y = final_filtered_BI['salary_in_usd']
plt.scatter(x, y, color='red')

#plot the regression line by computing its gradient and y-intercept using the polyf
m, c = np.polyfit(x,y, 1)
plt.plot(x, m*x + c)

plt.xlabel('Experience Level')
plt.ylabel('Salary in USD')
plt.title('Correlation between Salary and Experience Level for BI Data Analyst')
```

```
C:\Users\user\AppData\Local\Temp\ipykernel_21288\3846464864.py:7: SettingWithCopyW
arning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  filtered_BI['experience_level'] = filtered_BI['experience_level'].replace(experi
ence_level_num)
```

Out[154]:

```
Text(0.5, 1.0, 'Correlation between Salary and Experience Level for BI Data Analys
t')
```

Question 3

For B3.1, box plot has been used to show the association between Experience Level and the Salary for all jobs. This is because box plot is able to give clear representation on the distribution of the salary for each experience level. By extracting the key metrics from box plot such as median, we are able to make direct comparison within a graph. By visualisation, we can straight dispute whether there is association between both. Furthermore, the presence of outliers in the box plot allows for their identification, enabling the option to either include or exclude them from the analysis.

For B3.2, the correlation is calculated for every job title to discover which would return the highest association among all. However, it is essential to focus solely on job titles that contain data for all four different experience levels. By filtering out the job which lack of complete dataset, it ensures that the correlation calculated is accurate and able to relate to the association between both. Moreover, with the scatter plot shown, a positively sloped regression line indicates a positive correlation between salary and experience level, suggesting that as experience level increases, salary tends to increase as well. This distribution pattern aligns with the common expectation that more experienced professionals generally command higher salaries.

By looking into both questions, from B3.1 we know that there is association between both experience level and salary for all job, but if we look into each job title, we can see that there is certain jobs that shows negative correlation. For example, 3D Computer Vision Researcher ans Software Data Engineer. However, it does not affect the association between both as the correlation act as an outlier for the analysis.In short, by analyzing the distribution of salaries across different experience levels using box plots, I am able to gain a better understanding of how salaries vary within each job title and across experience levels. Similarly, by

calculating correlations for each job title, I can identify which roles exhibit the strongest association between experience level and salary.