FIT1043 Assignment 1

Renee Yeo Shu Ting (33518904)

Task A:

In [181…
```
import pandas as pd
import os
import matplotlib.pylab as plt
```

In [182…
```
os.getcwd()
```

Out[182]:
```
'/Users/reneeyeo/Desktop'
```

In [183…
```
os.chdir('/Users/reneeyeo/Desktop')
```

In [184…
```
salaries = pd.read_csv('salaries.csv')
```

A1:

In [185…
```
salaries.shape
```

Out[185]:
```
(3227, 11)
```

Answer: There are 3227 data instances and 11 variable exist in the given dataset.

A2:

In [186…
```
# Print out the first 8 rows
salaries.head(8)
```

Out[186]:

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

In [187…

```
# Print out the last 12 rows
salaries.tail(12)
```

Out[187]:

	work_year	experience_level	employment_type	job_title	salary	salary_currer
3215	2020	MI	FT	Data Engineer	130800	U
3216	2020	SE	FT	Machine Learning Engineer	40000	E
3217	2021	SE	FT	Director of Data Science	168000	U
3218	2021	MI	FT	Data Scientist	160000	S
3219	2021	MI	FT	Applied Machine Learning Scientist	423000	U
3220	2021	MI	FT	Data Engineer	24000	E
3221	2021	SE	FT	Data Specialist	165000	U
3222	2020	SE	FT	Data Scientist	412000	U
3223	2021	MI	FT	Principal Data Scientist	151000	U
3224	2020	EN	FT	Data Scientist	105000	U
3225	2020	EN	CT	Business Data Analyst	100000	U
3226	2021	SE	FT	Data Science Manager	7000000	I

In [188…

```
# Print out the random 6 rows of data
salaries.sample(6)
```

Out[188]:

	work_year	experience_level	employment_type	job_title	salary	salary_currer
2647	2022	SE	FT	Data Analyst	164000	U
1483	2022	MI	FT	Data Analyst	350000	G
1039	2023	EN	FT	Data Engineer	160000	U
2113	2022	SE	FT	Data Architect	66000	U
1304	2022	SE	FT	Data Science Consultant	145000	U
2039	2022	MI	FT	Data Scientist	84000	U

A3:

In [189…
```
salaries.dtypes
```

Out[189]:
```
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: The different data types for each column is stated above.

A4:

Question 1

In [190…
```
# Convert data from 'salary_in_usd' to MYR
conversion = (salaries['salary_in_usd'] * 4.47)
```

Question 2

In [203…
```
# Create new column 'salary_in_myr'
salaries['salary_in_myr'] = conversion
pd.set_option('display.max_rows', 20)
salaries
```

Out[203]:

	work_year	experience_level	employment_type	job_title	salary	salary_curren
0	2023	SE	FT	AI Scientist	1500000	I
1	2023	SE	FT	Machine Learning Engineer	216000	US
2	2023	SE	FT	Machine Learning Engineer	184000	US
3	2023	SE	FT	Data Engineer	180000	US
4	2023	SE	FT	Data Engineer	165000	US
...	...	...	...	...	...	
3222	2020	SE	FT	Data Scientist	412000	US
3223	2021	MI	FT	Principal Data Scientist	151000	US
3224	2020	EN	FT	Data Scientist	105000	US
3225	2020	EN	CT	Business Data Analyst	100000	US
3226	2021	SE	FT	Data Science Manager	7000000	IN

3227 rows × 12 columns

A5:

Question 1

In [193…

```
salaries.describe()
```

Out[193]:

	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

Question 2

Observation 1:

According to salary_in_usd, we can observed that the mean salary is 134750.29 USD, where the minimum salary and maximum salary are 5132 USD and 450000 USD. From this data, we are able to calculate the differences of mininimum and maximum salary, 450000 - 5132 = 444868. Furthermore, we can observe that the standard deviation of salary_in_usd is 62597.48. Lastly, we can see that there is more than 75 percent of jobs has a salary of 172347.50, more than 50 percent of jobs has a salary of 130026, and only 25 percent of jobs has a salary of 92350.

Observation 2:

According to remote_ratio, there are total of 3227 instances recorded where the maximum ratio is 100 and the minimum ratio is 0. The mean remote_ratio is 48.28 and the standard deviation of remote_ratio is 48.55. Moving on, we are able to observe from the data that more than 75 percent of remote ratio is 100, more than 50 percent of remote ratio is 50, and lastly more than 25 percent of remote ratio is 0.

A6:

Question 1

In [194…

```
# To find for unique job titles
unique = len(salaries['job_title'].unique())
unique
```

Out[194]:

```
85
```

Answer: There are 85 unique job titles recorded in the 'job_title' column.

Question 2

In [204…

```
# To find out the different job title
diff_jobs = salaries['job_title'].unique()

# To count the instances for each job title
diff_jobs_instance = salaries['job_title'].value_counts()

# To display all instances
pd.set_option('display.max_rows', None)

# Print
print("Answer:\n")
print("Different job titles")
print("--------------------")
print(diff_jobs)
print()
print("Instances of each job title")
print("---------------------------")
print(diff_jobs_instance)
```

```
Answer:

Different job titles
--------------------
['AI Scientist' 'Machine Learning Engineer' 'Data Engineer'
 'Data Scientist' 'Data Analyst' 'Analytics Engineer'
 'Machine Learning Scientist' 'Autonomous Vehicle Technician'
 'Applied Machine Learning Scientist' 'Lead Data Scientist'
 'Data Architect' 'Cloud Database Engineer' 'Research Engineer'
 'Data Manager' 'Data Science Manager' 'Applied Scientist'
 'Financial Data Analyst' 'Research Scientist'
 'Data Infrastructure Engineer' 'ML Engineer' 'Software Data Engineer'
 'AI Programmer' 'AI Developer' 'Lead Data Analyst'
 'Data Operations Engineer' 'BI Developer' 'Data Science Lead'
 'Data Analytics Manager' 'Deep Learning Researcher' 'BI Analyst'
 'Data Science Consultant' 'Data Analytics Specialist'
 'Machine Learning Infrastructure Engineer' 'Business Data Analyst'
 'Head of Data' 'Computer Vision Engineer' 'BI Data Analyst'
 'Head of Data Science' 'Data Quality Analyst' 'Insight Analyst'
 'Applied Machine Learning Engineer' 'Deep Learning Engineer'
 'Machine Learning Software Engineer' 'Big Data Architect'
 'Product Data Analyst' 'Computer Vision Software Engineer'
 'Director of Data Science' 'Azure Data Engineer' 'Big Data Engineer'
 'Marketing Data Engineer' 'Applied Data Scientist' 'Data Analytics Lead'
 'Data Lead' 'Data Science Engineer' 'Machine Learning Research Engineer'
 'NLP Engineer' 'Manager Data Management' 'Machine Learning Developer'
 '3D Computer Vision Researcher' 'MLOps Engineer' 'Data Specialist'
 'Principal Machine Learning Engineer' 'Machine Learning Researcher'
 'Data Analytics Engineer' 'Data Analytics Consultant'
 'Data Management Specialist' 'Data Science Tech Lead'
 'Data Scientist Lead' 'Cloud Data Engineer' 'Data Operations Analyst'
 'Marketing Data Analyst' 'Power BI Developer' 'Product Data Scientist'
 'Principal Data Architect' 'Machine Learning Manager'
 'Lead Machine Learning Engineer' 'ETL Developer' 'Cloud Data Architect'
 'Lead Data Engineer' 'Head of Machine Learning' 'Principal Data Analyst'
```

```
 'Principal Data Scientist' 'Principal Data Engineer'
 'Staff Data Scientist' 'Finance Data Analyst']

Instances of each job title
----------------------------
Data Engineer                              906
Data Scientist                             721
Data Analyst                               537
Machine Learning Engineer                  250
Data Architect                              85
Analytics Engineer                          79
Research Scientist                          69
Data Science Manager                        56
ML Engineer                                 32
Applied Scientist                           30
Machine Learning Scientist                  24
Data Science Consultant                     24
Data Manager                                23
Research Engineer                           21
Data Analytics Manager                      18
AI Scientist                                16
BI Data Analyst                             15
BI Developer                                13
Business Data Analyst                       13
Data Specialist                             12
Applied Machine Learning Scientist          12
Computer Vision Engineer                    12
Machine Learning Infrastructure Engineer    11
Big Data Engineer                           10
ETL Developer                               10
Machine Learning Software Engineer          10
Data Operations Engineer                    10
Head of Data Science                         9
BI Analyst                                   9
Lead Data Scientist                          9
Director of Data Science                     9
Data Science Lead                            8
Applied Data Scientist                       8
Head of Data                                 8
Principal Data Scientist                     7
Machine Learning Developer                   7
NLP Engineer                                 7
Data Infrastructure Engineer                 6
Lead Data Engineer                           6
Data Analytics Engineer                      6
Deep Learning Engineer                       6
Computer Vision Software Engineer            5
Machine Learning Researcher                  5
Data Science Engineer                        5
Cloud Database Engineer                      5
AI Developer                                 5
Product Data Analyst                         5
Lead Data Analyst                            4
Machine Learning Research Engineer           4
Data Operations Analyst                      4
3D Computer Vision Researcher                4
Cloud Data Engineer                          3
```

```
Financial Data Analyst                          3
Machine Learning Manager                        3
Lead Machine Learning Engineer                  3
Big Data Architect                              2
Principal Data Analyst                          2
Autonomous Vehicle Technician                   2
Marketing Data Analyst                          2
AI Programmer                                   2
Data Scientist Lead                             2
Insight Analyst                                 2
Data Analytics Consultant                       2
Data Analytics Lead                             2
Data Quality Analyst                            2
Principal Data Engineer                         2
MLOps Engineer                                  2
Data Lead                                       2
Data Analytics Specialist                       2
Software Data Engineer                          2
Staff Data Scientist                            1
Head of Machine Learning                        1
Deep Learning Researcher                        1
Cloud Data Architect                            1
Principal Machine Learning Engineer             1
Principal Data Architect                        1
Product Data Scientist                          1
Power BI Developer                              1
Data Science Tech Lead                          1
Data Management Specialist                      1
Manager Data Management                         1
Marketing Data Engineer                         1
Azure Data Engineer                             1
Applied Machine Learning Engineer               1
Finance Data Analyst                            1
Name: job_title, dtype: int64
```

Question 3

In [196…

```
# Filter data of Data Scientist
no_of_datasci = salaries[salaries['job_title'] == 'Data Scientist']

# Count for percentage of data scientist compared to all jobs
percentage = (len(no_of_datasci)/len(salaries))*100

# Print
print("Answer: The percentage of 'Data Scientist' is ", percentage,"%")
```

```
Answer: The percentage of 'Data Scientist' is  22.342733188720175 %
```

A7:

Question 1

In [197…

```
# To find out the different locations for the companies
diff_location = salaries['company_location'].unique()

# To count the instances for each location
diff_loc_instance = salaries['company_location'].value_counts()

# Print answer
pd.set_option('display.max_rows', None)
print("Answer:\n")
print("Different locations for the companies")
print("-------------------------------------")
print(diff_location)
print()
print("Instances for each location")
print("---------------------------")
print(diff_loc_instance)
```

```
Answer:

Different locations for the companies
-------------------------------------
['IL' 'US' 'IE' 'GH' 'DE' 'CA' 'GB' 'CH' 'CO' 'SG' 'IN' 'AU' 'SE' 'ES'
 'SI' 'MX' 'FR' 'BR' 'PT' 'RU' 'TH' 'HR' 'VN' 'NL' 'EE' 'AM' 'BA' 'KE'
 'GR' 'MK' 'LV' 'RO' 'PK' 'IT' 'MA' 'PL' 'AL' 'AR' 'LT' 'AS' 'CR' 'IR'
 'BS' 'HU' 'AT' 'SK' 'NG' 'CZ' 'TR' 'PR' 'FI' 'DK' 'BO' 'PH' 'BE' 'ID'
 'EG' 'AE' 'LU' 'MY' 'HN' 'JP' 'DZ' 'IQ' 'UA' 'CN' 'NZ' 'CL' 'MD' 'MT']

Instances for each location
---------------------------
US    2575
GB     159
CA      69
ES      68
IN      54
DE      53
FR      33
BR      15
AU      14
PT      14
GR      14
NL      11
MX      10
SG       6
JP       6
AT       6
TR       5
PL       5
IE       5
SI       4
IT       4
CO       4
PK       4
BE       4
DK       4
PR       4
CH       4
```

```
LV       4
CZ       3
AR       3
NG       3
RU       3
AE       3
HR       3
TH       3
AS       3
LU       3
ID       2
HU       2
IL       2
EE       2
LT       2
RO       2
KE       2
GH       2
SE       2
VN       1
MD       1
CL       1
NZ       1
CN       1
UA       1
IQ       1
DZ       1
HN       1
MY       1
CR       1
EG       1
IR       1
AM       1
PH       1
BO       1
BA       1
FI       1
MK       1
MA       1
SK       1
AL       1
BS       1
MT       1
Name: company_location, dtype: int64
```

Question 2

In [198…

```
# Filter data for company that is located in US and it is 'L' size
ttl_no_of_companies = salaries[(salaries['company_size'] == 'L') & (salar

# Print data to show total number
print("Answer: Total number of 'L' size companies in the US is", len(ttl_
```

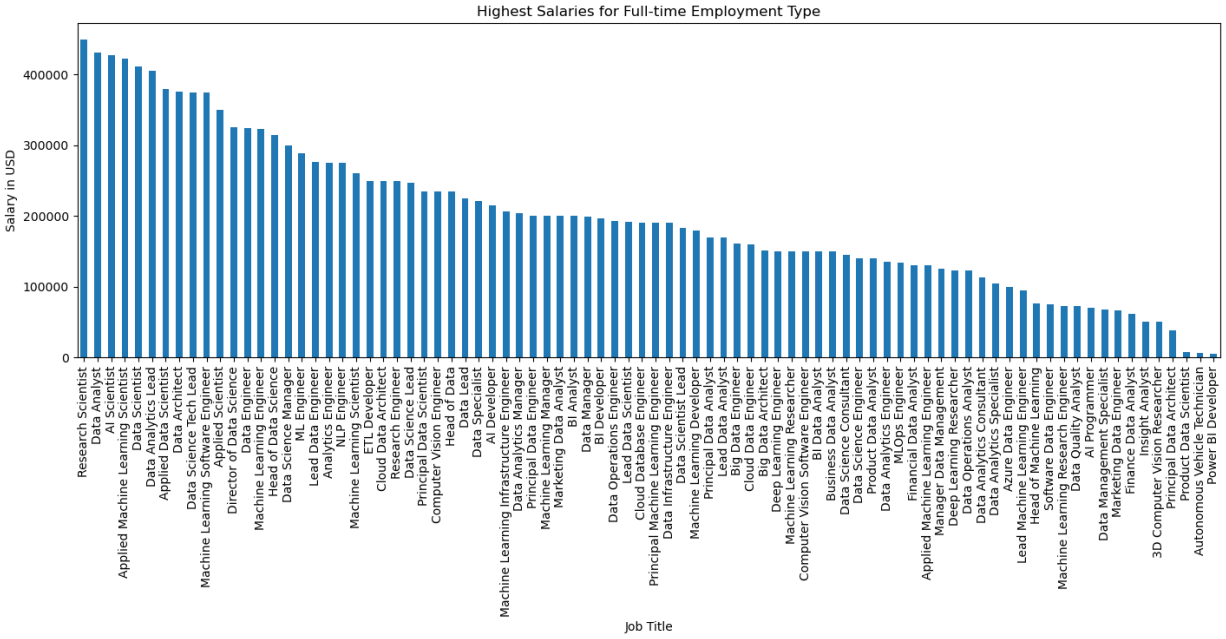
```
Answer: Total number of 'L' size companies in the US is 227
```

Task B:

B1:

Question 1

In [74]:
```
# Filter data for Full-time employment type
data1 = salaries[(salaries['employment_type'] == 'FT')]

# Group data by job title and get the maximum salary for each job
job_salary = data1.groupby('job_title')['salary_in_usd'].max()

# Sort data from highest to lowest
job_salary = job_salary.sort_values(ascending=False)

# Set graph size and labels to show graph
job_salary.plot.bar(figsize=(17,5))
plt.xlabel('Job Title')
plt.ylabel('Salary in USD')
plt.title('Highest Salaries for Full-time Employment Type')
plt.show()
```

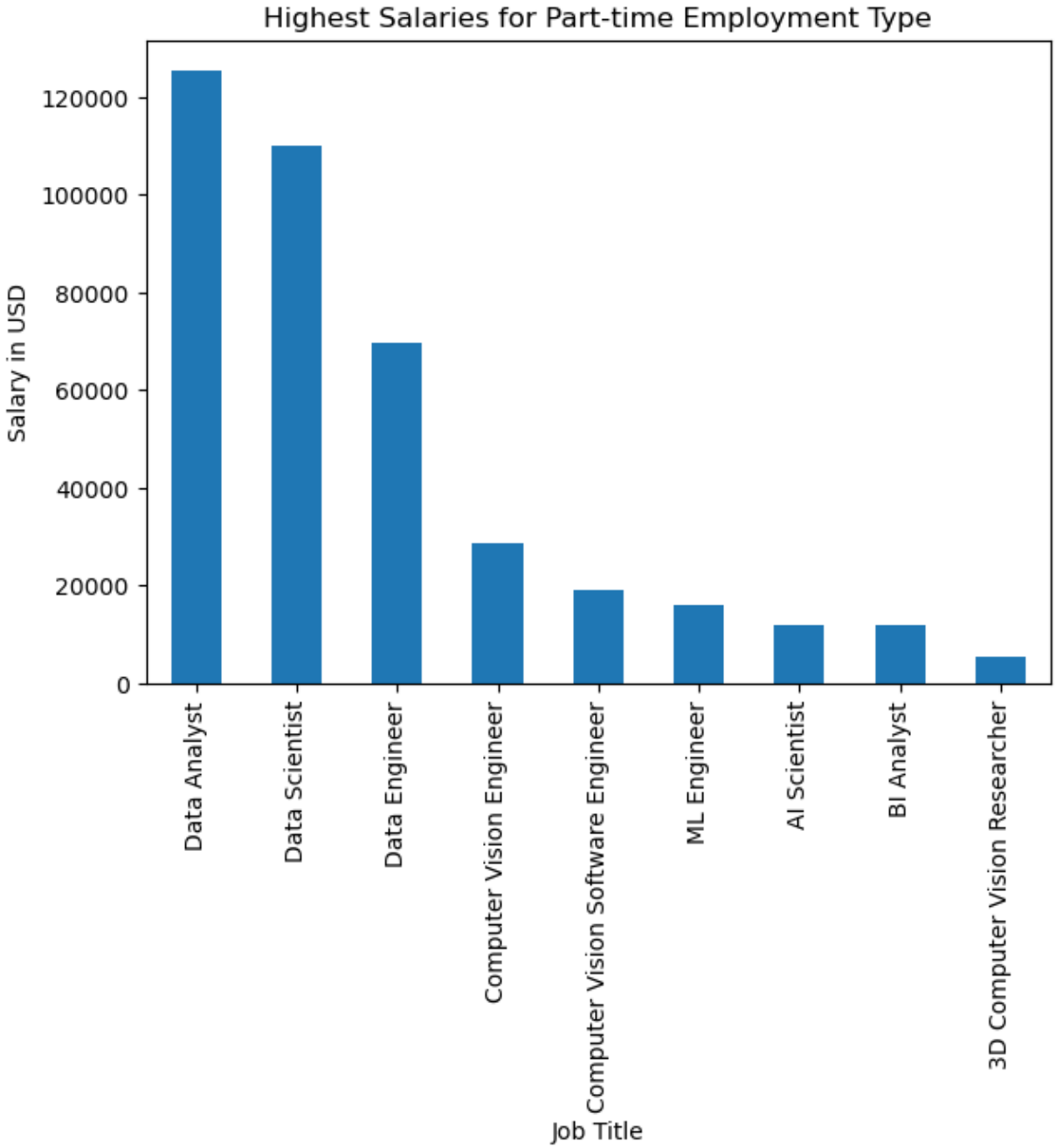


Answer: The above bar graph is showing the highest salaries in 'salaries.csv' for full-time employment type. From the bar graph, we can observe that Research Scientist has the highest salary among the others jobs. It has an estimated salary of 400000 USD and above.

Question 2

In [148…

```
# Filter data for Part-time employment type
data2 = salaries[(salaries['employment_type'] == 'PT')]

# Group data by job title and get the maximum salary for each job
job_salary2 = data2.groupby('job_title')['salary_in_usd'].max()

# Sort data from highest to lowest
job_salary2 = job_salary2.sort_values(ascending=False)

# Set graph size, labels and title
job_salary2.plot.bar(figsize=(7,5))
plt.xlabel('Job Title')
plt.ylabel('Salary in USD')
plt.title('Highest Salaries for Part-time Employment Type')
plt.show()
```



Answer: The above bar graph is showing the highest salaries in 'salaries.csv' for part-time employment type. From the bar graph, we can observe that Data Analyst has the highest salary among the others jobs. It has an estimated salary of 120000 USD and above.

Question 3

In [76]:
```
data_PT = salaries[(salaries['employment_type'] == 'PT') & (salaries['job
data_PT
```

Out[76]:

work_year	experience_level	employment_type	job_title	salary	salary_currency	salar

In [77]:
```
data_CT = salaries[(salaries['employment_type'] == 'CT') & (salaries['job
data_CT
```

Out[77]:

work_year	experience_level	employment_type	job_title	salary	salary_currency	salar

In [78]:
```
data_FL = salaries[(salaries['employment_type'] == 'FL') & (salaries['job
data_FL
```

Out[78]:

work_year	experience_level	employment_type	job_title	salary	salary_currency	salar

Observation:

From above analysis, we can observe that there is no part time(PT), contract(CT), and freelance(FL) jobs instances for Research Scientist. Hence, we cannot compare any insights for the highest salary.

B2:

Question 1

In [199…
```
# Filter data from location of company and sort to top 3 highest
records = salaries.groupby('company_location').size().sort_values(ascendi
records
```

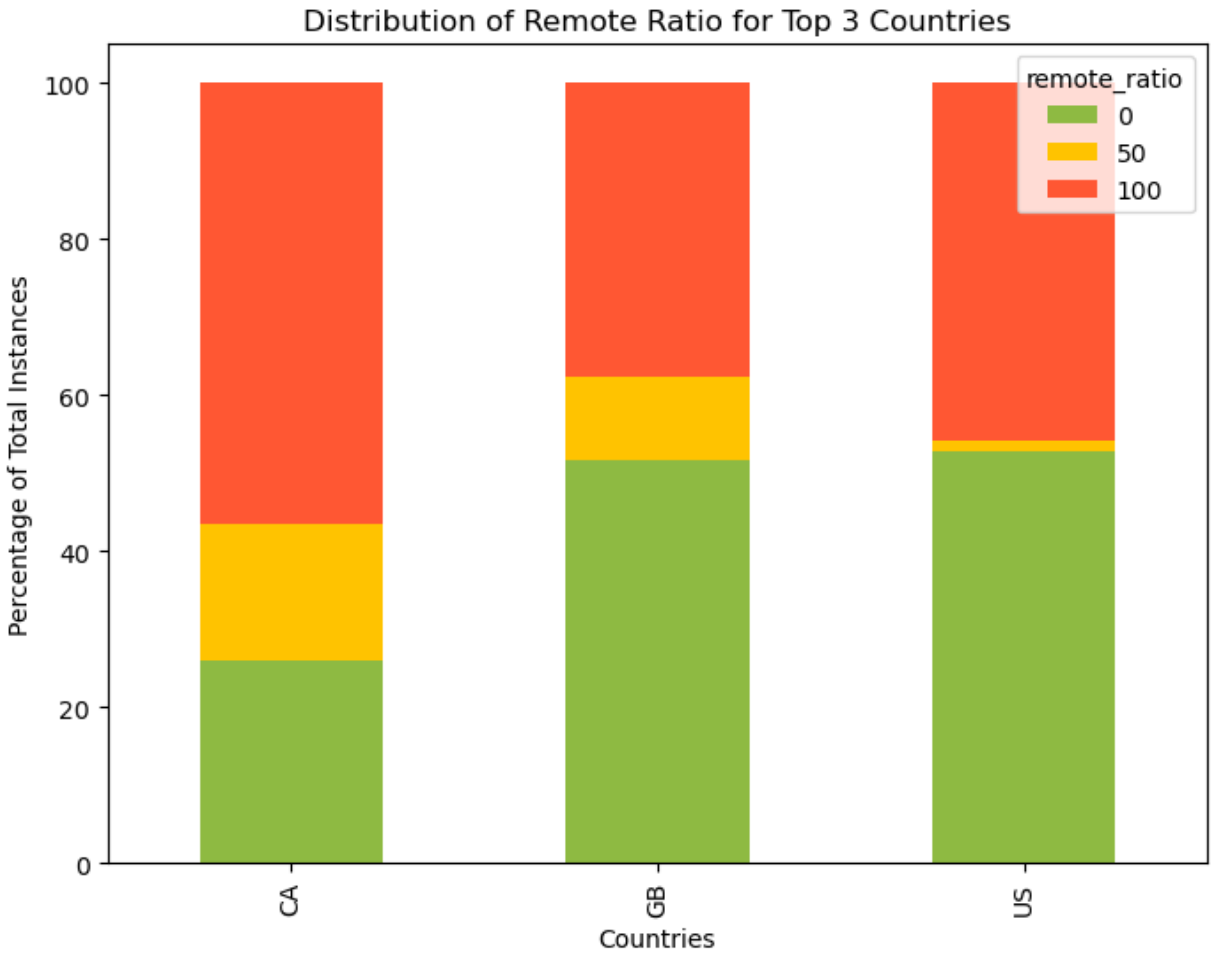
Out[199]:
```
company_location
US    2575
GB     159
CA      69
dtype: int64
```

Answer: Top three countries that has the highest recorded instances are US, GB and CA.

Question 2

In [200…

```
# Sort data
data = salaries[['remote_ratio', 'company_location']]
grouped_data = data.groupby(['remote_ratio', 'company_location']).size().
top_3_countries = grouped_data.groupby('company_location')['count'].sum()
filtered_data = salaries[salaries['company_location'].isin(top_3_countrie

# Group data by country and remote_ratio
grouped_data = filtered_data.groupby(['company_location', 'remote_ratio']
percentage = grouped_data.apply(lambda x: x / x.sum(), axis=1) * 100

# Plot the data
ax = percentage.plot(kind='bar', stacked=True, figsize=(8, 6), color=['#8
ax.set_xlabel('Countries')
ax.set_ylabel('Percentage of Total Instances')
ax.set_title('Distribution of Remote Ratio for Top 3 Countries')

# Show the plot
plt.show()
```



From the bar graph above, we can differentiate the data by three colors: green, yellow and orange, which are 0, 50 and 100 from 'remote_ratio' of 'salaries.csv'. Through this data, we are able to analyse the distribution of remote ratio of the top three countries with highest recorded instances. I have visualised the distribution through percentage of their total instances for each country.

First, we can look at the proportion of green in the bar graph. We are able to observe that US has the most distribution of remote ratio 0. This means US has a greatest distribution of less the 20% of no remote work among the rest. While Canada has the least distribution of remote ratio 0.

Moving on, we can look at the proportion of yellow in the bar graph. We can observe that Canada has the largest area of yellow area compared to the other two. This means Canada has the most distribution of remote ratio 50, in which Canada has the most percentage of its total instances that is partially remote. While US has the least distribution.

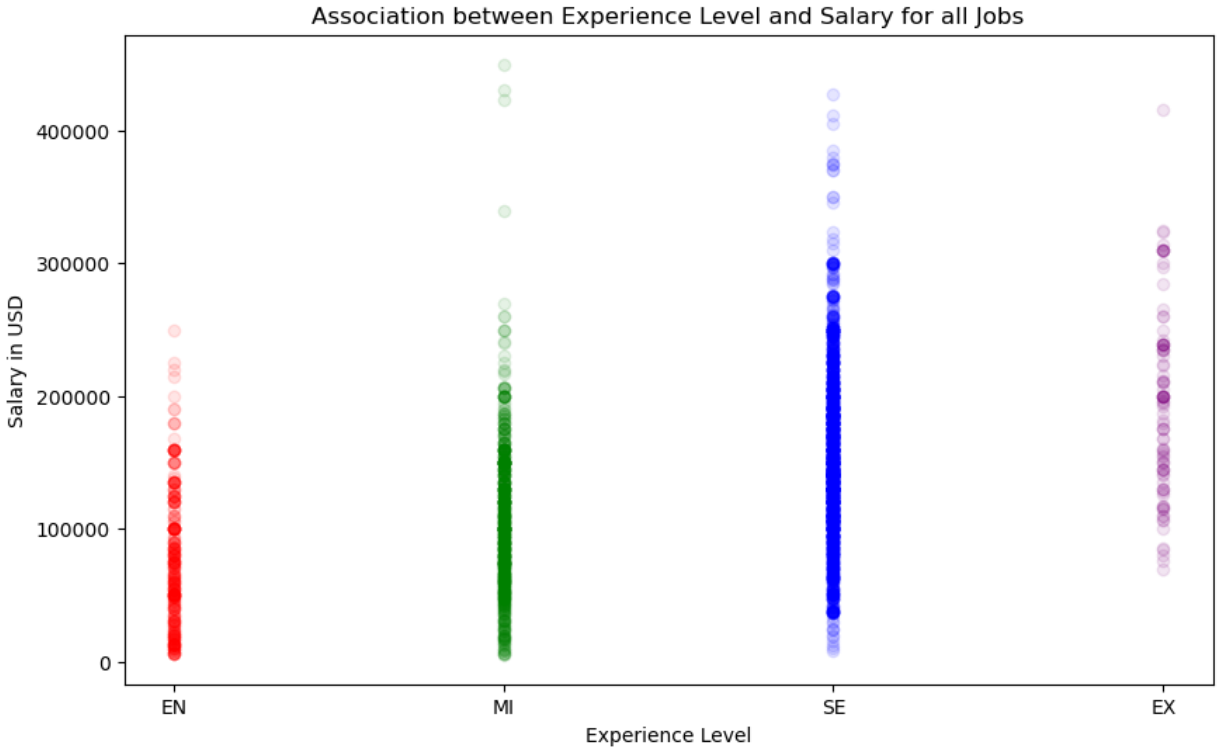
Lastly, we can look at the proportion of orange in the bar graph. We can see that Canada has the highest percentage of distribution of remote ratio 100 among its total instances. While Great Britain has the least percentage of distribution of remote ratio 100.

B3:

Question 1

In [202…

```
# Create a dictionary to map experience levels to colors
experiencelvl_color_map = {
    'EN': 'red',
    'MI': 'green',
    'SE': 'blue',
    'EX': 'purple'
}

# plot the scatter plot with different colors for each experience level
fig, ax = plt.subplots(figsize=(10,6))
for level, color in experiencelvl_color_map.items():
    level_data = salaries[salaries['experience_level']==level]
    ax.scatter(level_data['experience_level'], level_data['salary_in_usd'

# Label scatter plot and show
ax.set_xlabel('Experience Level')
ax.set_ylabel('Salary in USD')
ax.set_title('Association between Experience Level and Salary for all Job
plt.show()
```



This scatter plot shows the relationship between experience level and salary for all jobs. Each point on the plot represents a job, with the x-axis showing the experience level and the y-axis showing the salary in USD. The transparency of the points is set to 0.1 to make it easier to see areas of high density. The experience level are EN(Entry-level), MI(Mid-level), SE(Senior-level), and EX(Executive-level)

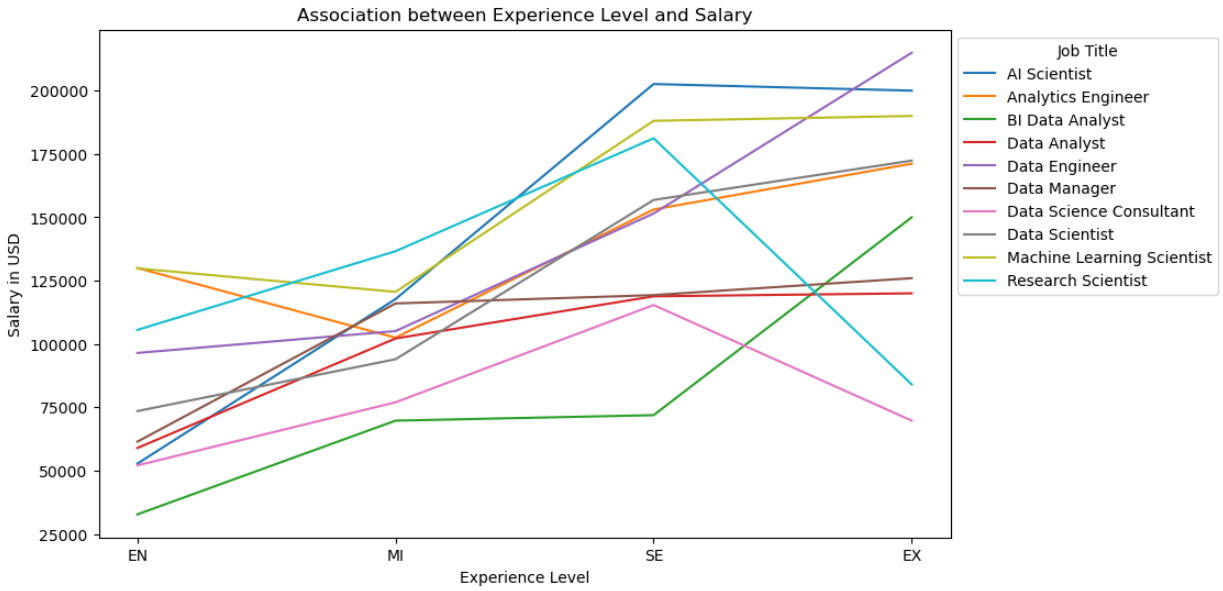
Firstly, EN has the lowest average salary among all, as we can observe through the points gathered mostly below 100000 USD. Moving on is MI, which their average salary gathered between around 100000 USD above or below. Next is SE, where the experience level is at the senior level, and their salary is mostly from 50000 USD to 250000 USD. The highest salary for MI and SE is above 400000 USD which we can assume is that they are in a well-paid company or they have really good skills. Lastly, we come to the EX category which has the least people among all jobs. Their salary is distributed mostly around 150000 USD to 250000 USD.

Overall, we are able to observe the trend of the higher the experience level, the higher the salary for average jobs. However, we could also note that there are some jobs with high salaries even at lower experience levels, indicating that experience level is not the only factor that determines salary.

Question 2

In [180…

```
# Define the order of labels on the x-axis
x_labels = ['EN', 'MI', 'SE', 'EX']

# Filter the data to include only job titles with all experience levels
unique = salaries['experience_level'].unique()
data = salaries.groupby('job_title').filter(lambda x: len(x['experience_l

# Group by job title and experience level, and calculate the mean salary
data = data.groupby(['job_title', 'experience_level'])['salary_in_usd'].m

# Create a categorical variable for the experience_level column with the
cat_x = pd.Categorical(data['experience_level'], categories=x_labels, ord
data = data.assign(cat_x=cat_x)

# Sort the data by the categorical variable and plot the line graph
data = data.sort_values(by=['job_title', 'cat_x'])
plt.figure(figsize=(10, 6))
for job_title, group in data.groupby('job_title'):
    plt.plot(group['cat_x'], group['salary_in_usd'], label=job_title)

# Set the title and axes labels
plt.title('Association between Experience Level and Salary')
plt.xlabel('Experience Level')
plt.ylabel('Salary in USD')

# Set the x-axis labels to the desired order
plt.xticks([0, 1, 2, 3], x_labels)

plt.legend(title='Job Title', bbox_to_anchor=(1,1), loc ='upper left')
plt.show()
```



According the the line graph, we are able to observe that the purple line has the highest association between Experience Level and Salary, where the job is Data Engineer. This is due to it's having a steady salary increase as experience level increases. Hence, we assume Data Engineer has a positive association between experience level and its salary.

Question 3

From the line graph above, we can observe that most of the jobs has a positive asssociation between their experience level and salary. However, there are some jobs do not have positive association for example, AI Scientisst, Research Scientist and Data Science Consultant. Three of these jobs were gradually increasing from entry level to senior level only. They are gradually decreasing in their executive level. We can assume that there are only a few jobs instances in executive level, which makes the graph has a negative association.

Finish