CC5067NI

60% Individual Coursework

2023-24 Autumn

Student Name: Amogh Man Bajracharya

London Met ID: 22067567

College ID: np0cp4a220064

Assignment Due Date: Monday, May 13, 2024

Assignment Submission Date: Monday, May 13, 2024

Word Count: 1188

I confirm that I understand my coursework needs to be submitted online via MySecondTeacher under the relevant module page before the deadline in order for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a marks of zero will be awarded.

Contents

Table of Tables

Table of Figures

1. Data Understanding

This dataset is about information of different variables that could impact salaries such as experience level job title and many more. In this dataset we find out about employment type based on job title and their salary and details about the employee and company. We dive into data cleaning, data preparation, data analysis and data exploration to prepare and generate meaningful findings and draw the conclusions. The steps in this project we will follow will help systematically prepare the data to uncover the insights.

S.no	Column Name	Description	Data Type
1	work_year	This column gives the work year of the employee	Int(64)
2	experience_level	This column give the experience level of the employee	object
3	employment_type	This column gives the employment type of the employee	object
4	job_title	This column state all the job titles of the employee	object
5	salary	This column states all the salary of given job title	Int(64)
6	salary_currency	This column describes which currency salary is received in	object
7	salary_in_usd	this column describes salary	Int(64)

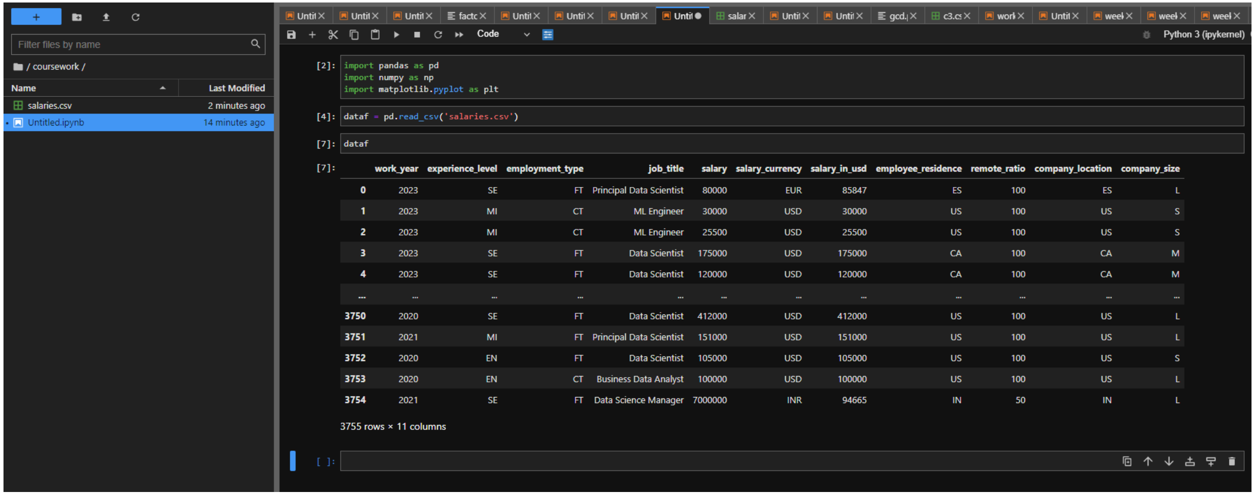
		which is received in usd	
8	Employee_residence	This column describes the residence of the employee	object
9	Remote_ratio	This column describes whether the employee is working at site or remotely	Int(64)
10	Company_location	This column describes company’s location which employee works in	object
11	Company_size	This column states company’s size.	object

Table 1 Description of every column name of the dataset

2. Data Preparation

Data preparation is the first step of cleaning and enriching raw data to help to make it ready for use in analytics and data science. Data preparation helps you to find prepare and use the prepared data faster. The idea behind data preparation is to change data into information which will be useful for data analysis. (Secoda, 2024)

2.1 Write a python program to load data into pandas DataFrame



```
[2]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt

[4]: dataf = pd.read_csv('salaries.csv')

[7]: dataf
```

[7]:

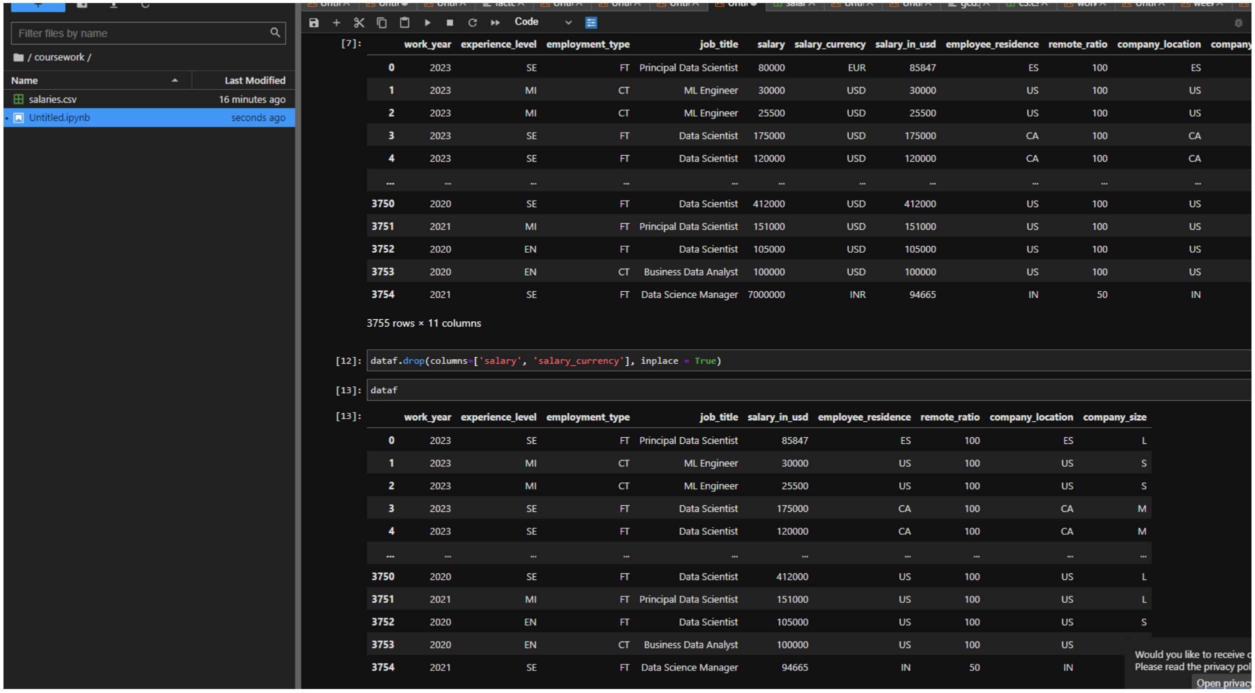
	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES	L
1	2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US	S
2	2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US	S
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA	M
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA	M
...	...	...	...	...	...	...	...	...	...	...	...
3750	2020	SE	FT	Data Scientist	412000	USD	412000	US	100	US	L
3751	2021	MI	FT	Principal Data Scientist	151000	USD	151000	US	100	US	L
3752	2020	EN	FT	Data Scientist	105000	USD	105000	US	100	US	S
3753	2020	EN	CT	Business Data Analyst	100000	USD	100000	US	100	US	L
3754	2021	SE	FT	Data Science Manager	7000000	INR	94665	IN	50	IN	L

3755 rows × 11 columns

Figure 1 Loading data into pandas DataFrame

This code imports pandas numpy and matplotlib dataframe and loads data into pandas dataframe making dataf as the name of dataframe.

2.2 Write a python program to remove unnecessary columns i.e., salary and salary currency.



[7]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES	
1	2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US	
2	2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US	
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA	
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA	
...	...	...	...	...	...	...	...	...	...	...	
3750	2020	SE	FT	Data Scientist	412000	USD	412000	US	100	US	
3751	2021	MI	FT	Principal Data Scientist	151000	USD	151000	US	100	US	
3752	2020	EN	FT	Data Scientist	105000	USD	105000	US	100	US	
3753	2020	EN	CT	Business Data Analyst	100000	USD	100000	US	100	US	
3754	2021	SE	FT	Data Science Manager	7000000	INR	94665	IN	50	IN	

3755 rows × 11 columns

```
[12]: dataf.drop(columns=['salary', 'salary_currency'], inplace = True)
```

```
[13]: dataf
```

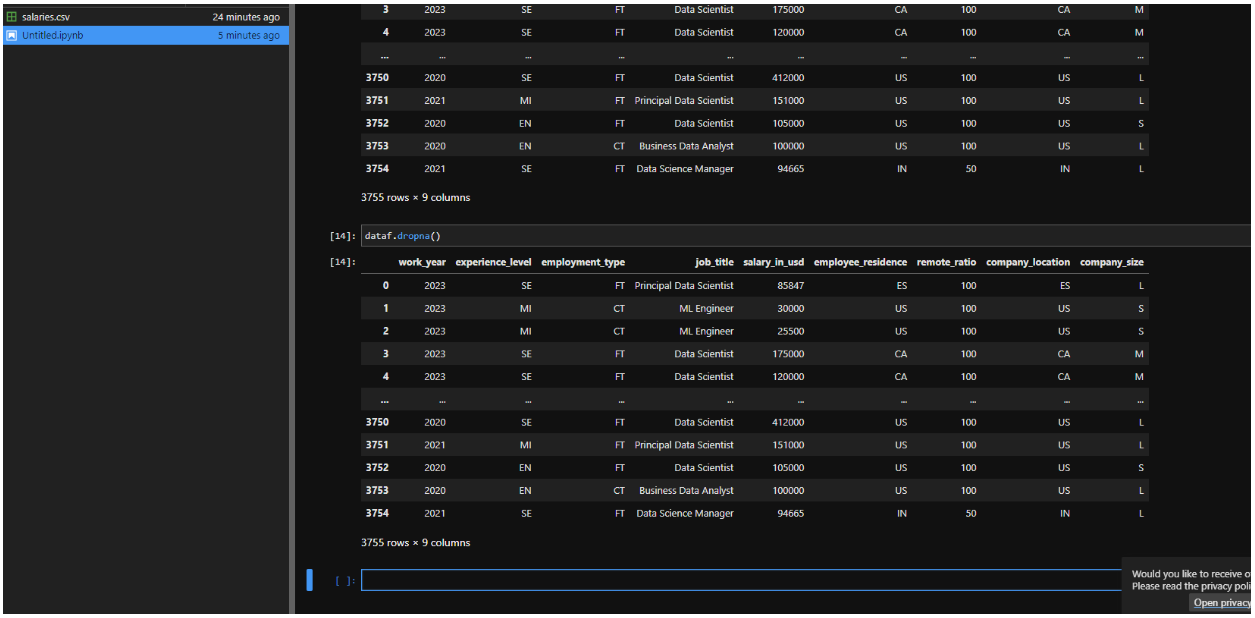
[13]:

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	Principal Data Scientist	85847	ES	100	ES	L
1	2023	MI	CT	ML Engineer	30000	US	100	US	S
2	2023	MI	CT	ML Engineer	25500	US	100	US	S
3	2023	SE	FT	Data Scientist	175000	CA	100	CA	M
4	2023	SE	FT	Data Scientist	120000	CA	100	CA	M
...	...	...	...	...	...	...	...	...	...
3750	2020	SE	FT	Data Scientist	412000	US	100	US	L
3751	2021	MI	FT	Principal Data Scientist	151000	US	100	US	L
3752	2020	EN	FT	Data Scientist	105000	US	100	US	S
3753	2020	EN	CT	Business Data Analyst	100000	US	100	US	
3754	2021	SE	FT	Data Science Manager	94665	IN	50	IN	

Figure 2 Removing Columns salary and salary currency

This code removes unnecessary columns i.e. salary and currency which is repeated by salary in USD and company's location.

2.3 Write a python program to remove the NaN missing values from updated dataframe.

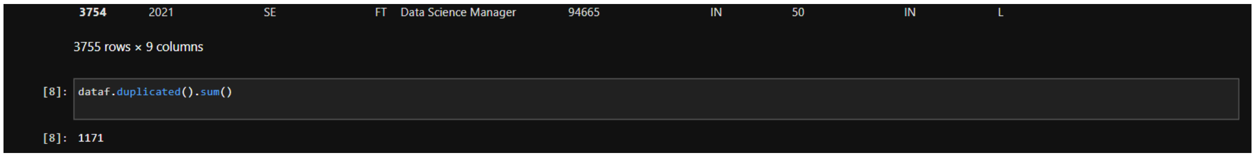


```
   3     2023   SE   FT        Data Scientist   175000   CA   100   CA   M
   4     2023   SE   FT        Data Scientist   120000   CA   100   CA   M
 ...      ...  ...  ...                   ...      ...  ...   ...  ...  ...
3750     2020   SE   FT        Data Scientist   412000   US   100   US   L
3751     2021   MI   FT  Principal Data Scientist 151000  US   100   US   L
3752     2020   EN   FT        Data Scientist   105000   US   100   US   S
3753     2020   EN   CT   Business Data Analyst  100000   US   100   US   L
3754     2021   SE   FT  Data Science Manager    94665   IN    50   IN   L

3755 rows × 9 columns

[14]: dataf.dropna()

[14]:
      work_year experience_level employment_type                job_title salary_in_usd employee_residence remote_ratio company_location company_size
   0       2023               SE              FT Principal Data Scientist         85847                 ES          100               ES            L
   1       2023               MI              CT              ML Engineer         30000                 US          100               US            S
   2       2023               MI              CT              ML Engineer         25500                 US          100               US            S
   3       2023               SE              FT           Data Scientist        175000                 CA          100               CA            M
   4       2023               SE              FT           Data Scientist        120000                 CA          100               CA            M
 ...        ...              ...             ...                      ...           ...                ...          ...              ...          ...
3750       2020               SE              FT           Data Scientist        412000                 US          100               US            L
3751       2021               MI              FT Principal Data Scientist        151000                 US          100               US            L
3752       2020               EN              FT           Data Scientist        105000                 US          100               US            S
3753       2020               EN              CT    Business Data Analyst        100000                 US          100               US            L
3754       2021               SE              FT     Data Science Manager         94665                 IN           50               IN            L

3755 rows × 9 columns

[ ]:
```

Figure 3 dropping any missing values

This code of line removes the NaN missing values to remove it from the updated dataframe with no salary and currency column.

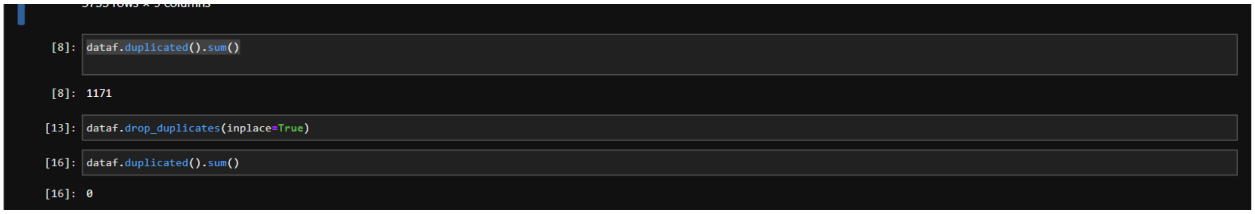
2.4 Write a python program to check duplicates value in the dataframe.



```
3754     2021     SE     FT  Data Science Manager     94665     IN     50     IN     L

3755 rows × 9 columns

[8]: dataf.duplicated().sum()

[8]: 1171
```

Figure 4 checking duplicates in dataframe

This code of line checks duplicates in dataframe

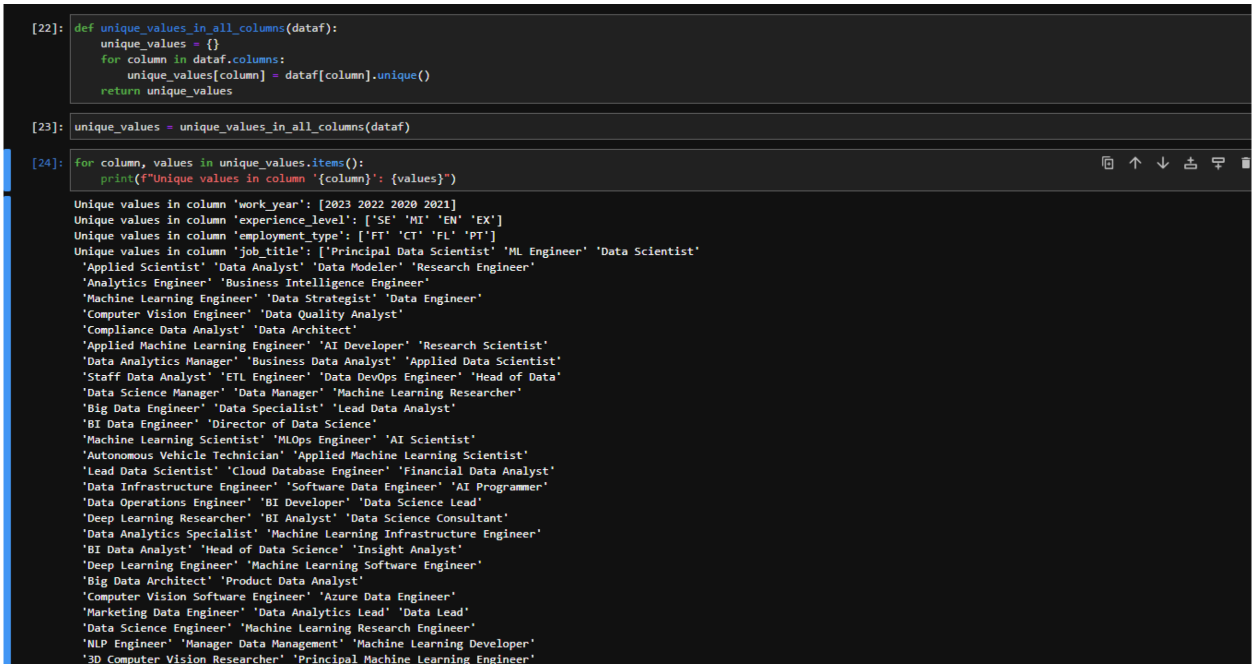


```
3755 rows × 9 columns

[8]: dataf.duplicated().sum()

[8]: 1171

[13]: dataf.drop_duplicates(inplace=True)

[16]: dataf.duplicated().sum()

[16]: 0
```

Figure 5 droping duplicated values

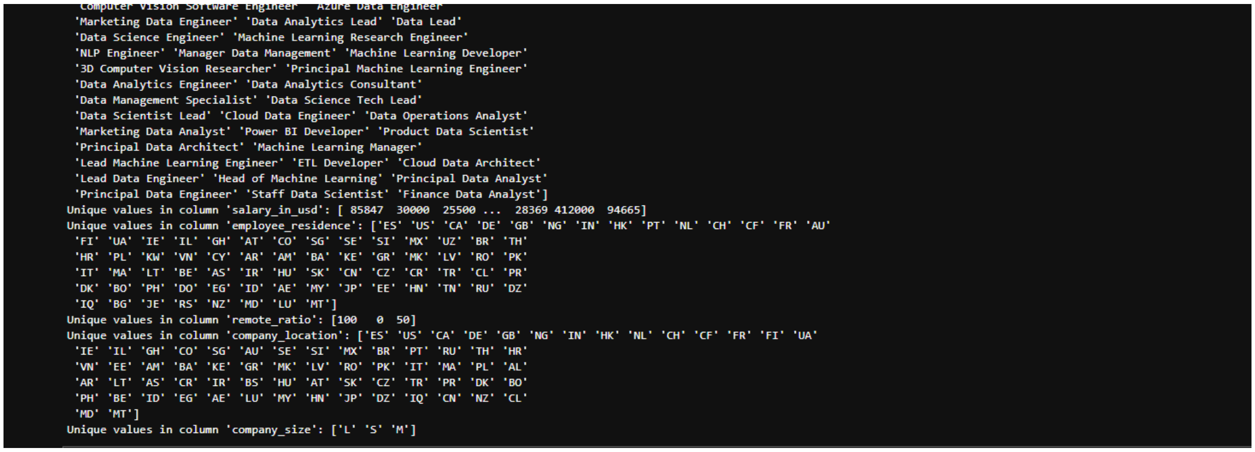
This code of line deletes duplicated values in the dataframe for better data consistency.

2.5 Write a python program to see the unique values from all the columns in the dataframe.



```
[22]: def unique_values_in_all_columns(dataf):
          unique_values = {}
          for column in dataf.columns:
              unique_values[column] = dataf[column].unique()
          return unique_values

[23]: unique_values = unique_values_in_all_columns(dataf)

[24]: for column, values in unique_values.items():
          print(f"Unique values in column '{column}': {values}")

      Unique values in column 'work_year': [2023 2022 2020 2021]
      Unique values in column 'experience_level': ['SE' 'MI' 'EN' 'EX']
      Unique values in column 'employment_type': ['FT' 'CT' 'FL' 'PT']
      Unique values in column 'job_title': ['Principal Data Scientist' 'ML Engineer' 'Data Scientist'
       'Applied Scientist' 'Data Analyst' 'Data Modeler' 'Research Engineer'
       'Analytics Engineer' 'Business Intelligence Engineer'
       'Machine Learning Engineer' 'Data Strategist' 'Data Engineer'
       'Computer Vision Engineer' 'Data Quality Analyst'
       'Compliance Data Analyst' 'Data Architect'
       'Applied Machine Learning Engineer' 'AI Developer' 'Research Scientist'
       'Data Analytics Manager' 'Business Data Analyst' 'Applied Data Scientist'
       'Staff Data Analyst' 'ETL Engineer' 'Data DevOps Engineer' 'Head of Data'
       'Data Science Manager' 'Data Manager' 'Machine Learning Researcher'
       'Big Data Engineer' 'Data Specialist' 'Lead Data Analyst'
       'BI Data Engineer' 'Director of Data Science'
       'Machine Learning Scientist' 'MLOps Engineer' 'AI Scientist'
       'Autonomous Vehicle Technician' 'Applied Machine Learning Scientist'
       'Lead Data Scientist' 'Cloud Database Engineer' 'Financial Data Analyst'
       'Data Infrastructure Engineer' 'Software Data Engineer' 'AI Programmer'
       'Data Operations Engineer' 'BI Developer' 'Data Science Lead'
       'Deep Learning Researcher' 'BI Analyst' 'Data Science Consultant'
       'Data Analytics Specialist' 'Machine Learning Infrastructure Engineer'
       'BI Data Analyst' 'Head of Data Science' 'Insight Analyst'
       'Deep Learning Engineer' 'Machine Learning Software Engineer'
       'Big Data Architect' 'Product Data Analyst'
       'Computer Vision Software Engineer' 'Azure Data Engineer'
       'Marketing Data Engineer' 'Data Analytics Lead' 'Data Lead'
       'Data Science Engineer' 'Machine Learning Research Engineer'
       'NLP Engineer' 'Manager Data Management' 'Machine Learning Developer'
       '3D Computer Vision Researcher' 'Principal Machine Learning Engineer'
```

Figure 6 python program to see unique values of all the columns.

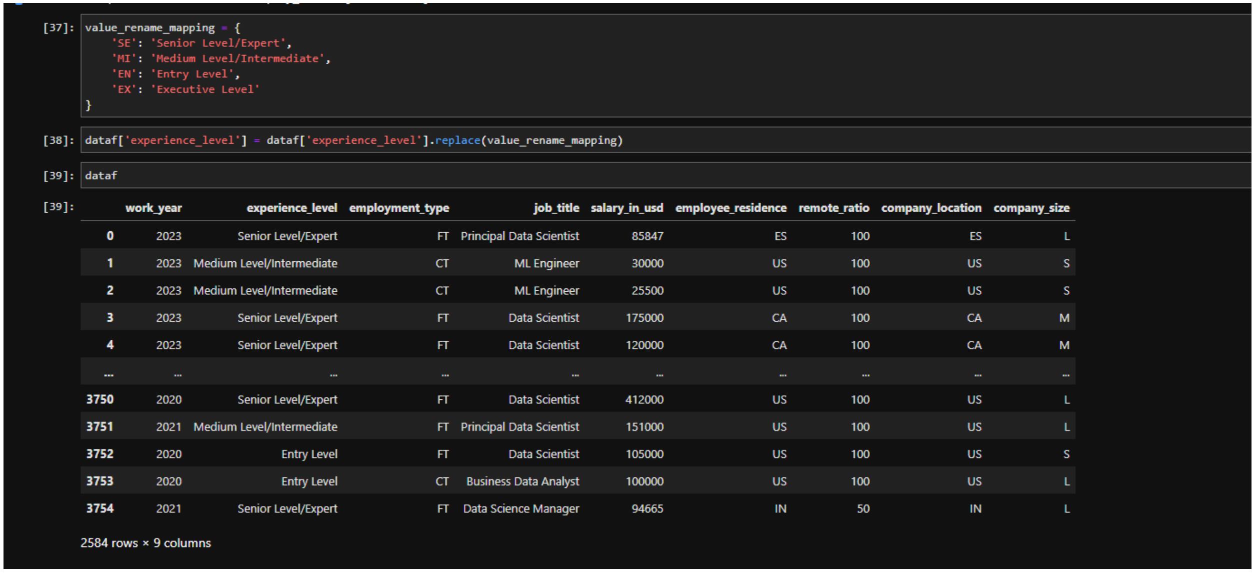


```
 'Computer Vision Software Engineer' 'Azure Data Engineer'
 'Marketing Data Engineer' 'Data Analytics Lead' 'Data Lead'
 'Data Science Engineer' 'Machine Learning Research Engineer'
 'NLP Engineer' 'Manager Data Management' 'Machine Learning Developer'
 '3D Computer Vision Researcher' 'Principal Machine Learning Engineer'
 '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 Engineer' 'Staff Data Scientist' 'Finance Data Analyst']
Unique values in column 'salary_in_usd': [ 85847  30000  25500 ...  28369 412000  94665]
Unique values in column 'employee_residence': ['ES' 'US' 'CA' 'DE' 'GB' 'NG' 'IN' 'HK' 'PT' 'NL' 'CH' 'CF' 'FR' 'AU'
 'FI' 'UA' 'IE' 'IL' 'GH' 'AT' 'CO' 'SG' 'SE' 'SI' 'MX' 'UZ' 'BR' 'TH'
 'HR' 'PL' 'KW' 'VN' 'CY' 'AR' 'AM' 'BA' 'KE' 'GR' 'MK' 'LV' 'RO' 'PK'
 'IT' 'MA' 'LT' 'BE' 'AS' 'IR' 'HU' 'SK' 'CN' 'CZ' 'CR' 'TR' 'CL' 'PR'
 'DK' 'BO' 'PH' 'DO' 'EG' 'ID' 'AE' 'MY' 'JP' 'EE' 'HN' 'TN' 'RU' 'DZ'
 'IQ' 'BG' 'JE' 'RS' 'NZ' 'MD' 'LU' 'MT']
Unique values in column 'remote_ratio': [100   0  50]
Unique values in column 'company_location': ['ES' 'US' 'CA' 'DE' 'GB' 'NG' 'IN' 'HK' 'NL' 'CH' 'CF' 'FR' 'FI' 'UA'
 'IE' 'IL' 'GH' 'CO' 'SG' 'AU' 'SE' 'SI' 'MX' 'BR' 'PT' 'RU' 'TH' 'HR'
 'VN' 'EE' 'AM' 'BA' 'KE' 'GR' 'MK' 'LV' 'RO' 'PK' 'IT' 'MA' 'PL' 'AL'
 'AR' 'LT' 'AS' 'CR' 'IR' 'BS' 'HU' 'AT' 'SK' 'CZ' 'TR' 'PR' 'DK' 'BO'
 'PH' 'BE' 'ID' 'EG' 'AE' 'LU' 'MY' 'HN' 'JP' 'DZ' 'IQ' 'CN' 'NZ' 'CL'
 'MD' 'MT']
Unique values in column 'company_size': ['L' 'S' 'M']
```

Figure 7 program to see unique values of all the columns.

These codes of lines defines a function named unique_values_in_all_columns where all the unique_values are gathered from each columns and then later called to print each and every unique values with their respective columns.

2.1 Rename the experience level columns as below.



```
[37]: value_rename_mapping = {
          'SE': 'Senior Level/Expert',
          'MI': 'Medium Level/Intermediate',
          'EN': 'Entry Level',
          'EX': 'Executive Level'
      }

[38]: dataf['experience_level'] = dataf['experience_level'].replace(value_rename_mapping)

[39]: dataf
```

[39]:

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	Senior Level/Expert	FT	Principal Data Scientist	85847	ES	100	ES	L
1	2023	Medium Level/Intermediate	CT	ML Engineer	30000	US	100	US	S
2	2023	Medium Level/Intermediate	CT	ML Engineer	25500	US	100	US	S
3	2023	Senior Level/Expert	FT	Data Scientist	175000	CA	100	CA	M
4	2023	Senior Level/Expert	FT	Data Scientist	120000	CA	100	CA	M
...	...	...	...	...	...	...	...	...	...
3750	2020	Senior Level/Expert	FT	Data Scientist	412000	US	100	US	L
3751	2021	Medium Level/Intermediate	FT	Principal Data Scientist	151000	US	100	US	L
3752	2020	Entry Level	FT	Data Scientist	105000	US	100	US	S
3753	2020	Entry Level	CT	Business Data Analyst	100000	US	100	US	L
3754	2021	Senior Level/Expert	FT	Data Science Manager	94665	IN	50	IN	L

2584 rows × 9 columns

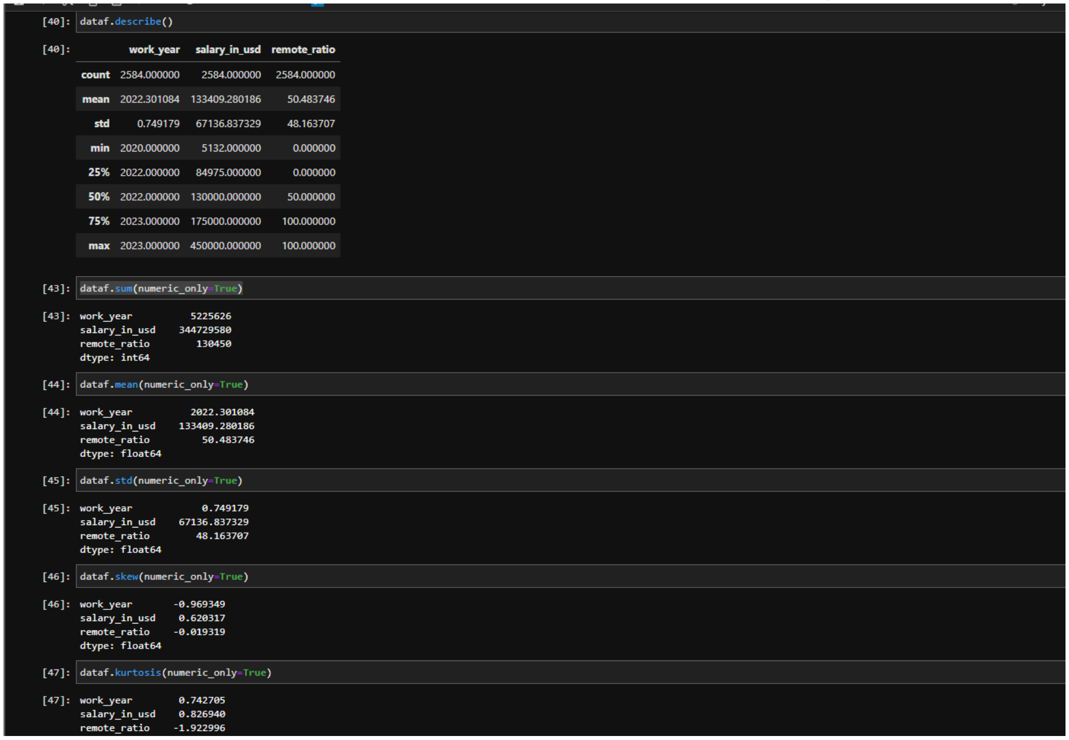
Figure 8 renaming the experience level rows.

This group of code renames experience level rows replacing and renaming every values in the column.

3. Data Analysis

Data analysis is a process of analysing, cleansing, manipulating, and modelling data in order to identify usable information and to draw conclusions which helps in decision-making. Data analysis is a process that uses a different approaches and methodologies to understand data from different sources in various formats, both structured and unstructured. (datacamp, 2023)

3.1 Write a Python program to show summary statistics of sum, mean, standard deviation, skewness, and kurtosis of any chosen variable.



```
[40]: dataf.describe()
```

[40]:

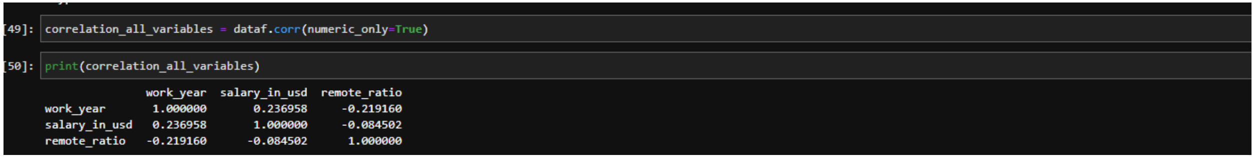
	work_year	salary_in_usd	remote_ratio
count	2584.000000	2584.000000	2584.000000
mean	2022.301084	133409.280186	50.483746
std	0.749179	67136.837329	48.163707
min	2020.000000	5132.000000	0.000000
25%	2022.000000	84975.000000	0.000000
50%	2022.000000	130000.000000	50.000000
75%	2023.000000	175000.000000	100.000000
max	2023.000000	450000.000000	100.000000

```
[43]: dataf.sum(numeric_only=True)

[43]: work_year         5225626
      salary_in_usd   344729580
      remote_ratio       130450
      dtype: int64

[44]: dataf.mean(numeric_only=True)

[44]: work_year           2022.301084
      salary_in_usd     133409.280186
      remote_ratio          50.483746
      dtype: float64

[45]: dataf.std(numeric_only=True)

[45]: work_year            0.749179
      salary_in_usd    67136.837329
      remote_ratio        48.163707
      dtype: float64

[46]: dataf.skew(numeric_only=True)

[46]: work_year       -0.969349
      salary_in_usd    0.620317
      remote_ratio    -0.019319
      dtype: float64

[47]: dataf.kurtosis(numeric_only=True)

[47]: work_year        0.742705
      salary_in_usd    0.826940
      remote_ratio    -1.922996
```

Figure 9 showing summary statistics of sum, mean , sd, skewness and kurtosis

These group of code defines the table’s summary, statistics of sum, mean, Standard Deviation, skewness and kurtosis.

3.2 Write a Python program to calculate and show correlation of all variables.



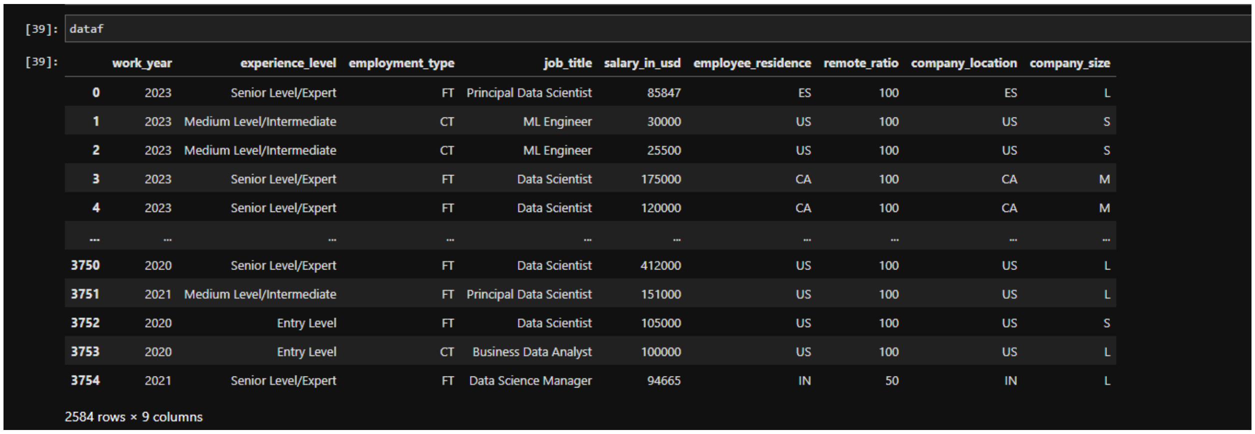
```
[49]: correlation_all_variables = dataf.corr(numeric_only=True)

[50]: print(correlation_all_variables)

                work_year  salary_in_usd  remote_ratio
work_year        1.000000       0.236958     -0.219160
salary_in_usd    0.236958       1.000000     -0.084502
remote_ratio    -0.219160      -0.084502      1.000000
```

Figure 10 Correlation of all the variables.

These lines of code defines correlation of all the variables in the table.

4. Data Cleaning

Data cleaning is the process of removing corrupted, inconsistent, or incomplete data in a dataset. It is a crucial step in the machine learning process to achieve and ensure that the data is accurate and consistent with minimum to none errors in the dataset. It is done to bring positive impact on performance of machine learning model. (Geek for Geeks, 2024)

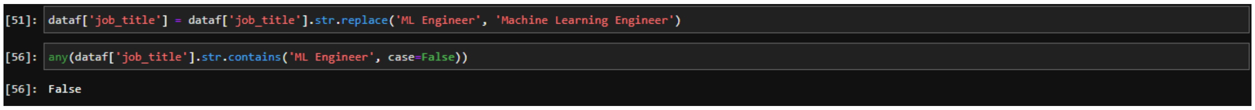


```
[39]: dataf
[39]:
```

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	Senior Level/Expert	FT	Principal Data Scientist	85847	ES	100	ES	L
1	2023	Medium Level/Intermediate	CT	ML Engineer	30000	US	100	US	S
2	2023	Medium Level/Intermediate	CT	ML Engineer	25500	US	100	US	S
3	2023	Senior Level/Expert	FT	Data Scientist	175000	CA	100	CA	M
4	2023	Senior Level/Expert	FT	Data Scientist	120000	CA	100	CA	M
...	...	...	...	...	...	...	...	...	...
3750	2020	Senior Level/Expert	FT	Data Scientist	412000	US	100	US	L
3751	2021	Medium Level/Intermediate	FT	Principal Data Scientist	151000	US	100	US	L
3752	2020	Entry Level	FT	Data Scientist	105000	US	100	US	S
3753	2020	Entry Level	CT	Business Data Analyst	100000	US	100	US	L
3754	2021	Senior Level/Expert	FT	Data Science Manager	94665	IN	50	IN	L

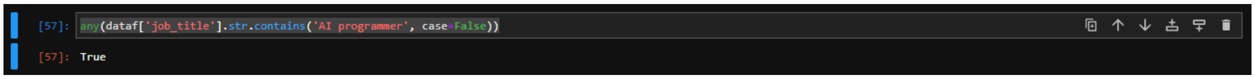
2584 rows × 9 columns

Figure 11 Before Data Cleaning



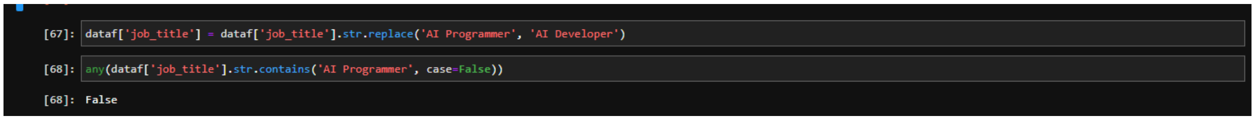
```
[51]: dataf['job_title'] = dataf['job_title'].str.replace('ML Engineer', 'Machine Learning Engineer')

[56]: any(dataf['job_title'].str.contains('ML Engineer', case=False))

[56]: False
```

Figure 12 After Data Cleaning Machine Learning



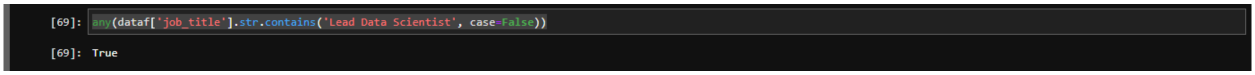
```
[57]: any(dataf['job_title'].str.contains('AI programmer', case=False))

[57]: True
```

Figure 13 Before Data Cleaning AI Programmer into AI Developer



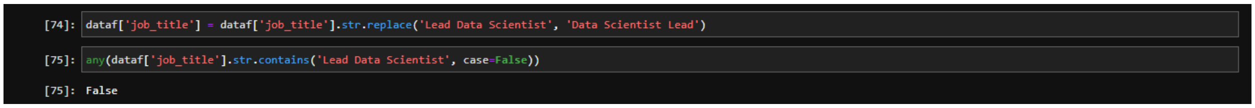
```
[67]: dataf['job_title'] = dataf['job_title'].str.replace('AI Programmer', 'AI Developer')

[68]: any(dataf['job_title'].str.contains('AI Programmer', case=False))

[68]: False
```

Figure 14 After Data Cleaning AI Programmer into AI Developer



```
[69]: any(dataf['job_title'].str.contains('Lead Data Scientist', case=False))

[69]: True
```

Figure 15 Before Data Cleaning Lead Data Scientist into Data Scientist Lead



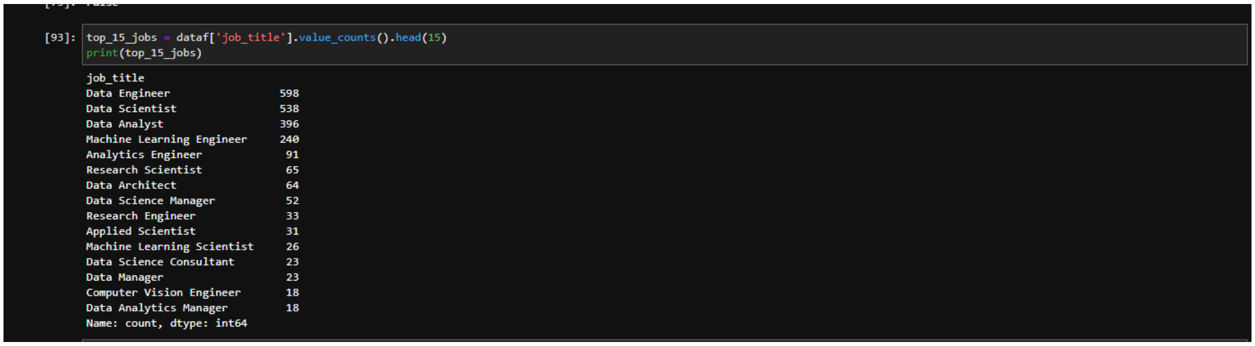
```
[74]: dataf['job_title'] = dataf['job_title'].str.replace('Lead Data Scientist', 'Data Scientist Lead')

[75]: any(dataf['job_title'].str.contains('Lead Data Scientist', case=False))

[75]: False
```

Figure 16 After Data Cleaning Lead Data Scientist into Data Scientist Lead

5. Data Exploration

Data exploration is one of the processes for machine learning which leads to reviewing of raw dataset that helps to figure out initial patterns for further analysis. As it is difficult to manage and review thousands of data elements to get proper analysis view of the dataset Data exploration helps to manage unstructured dataset and recognize patterns accordingly. (Qlik, 2024)

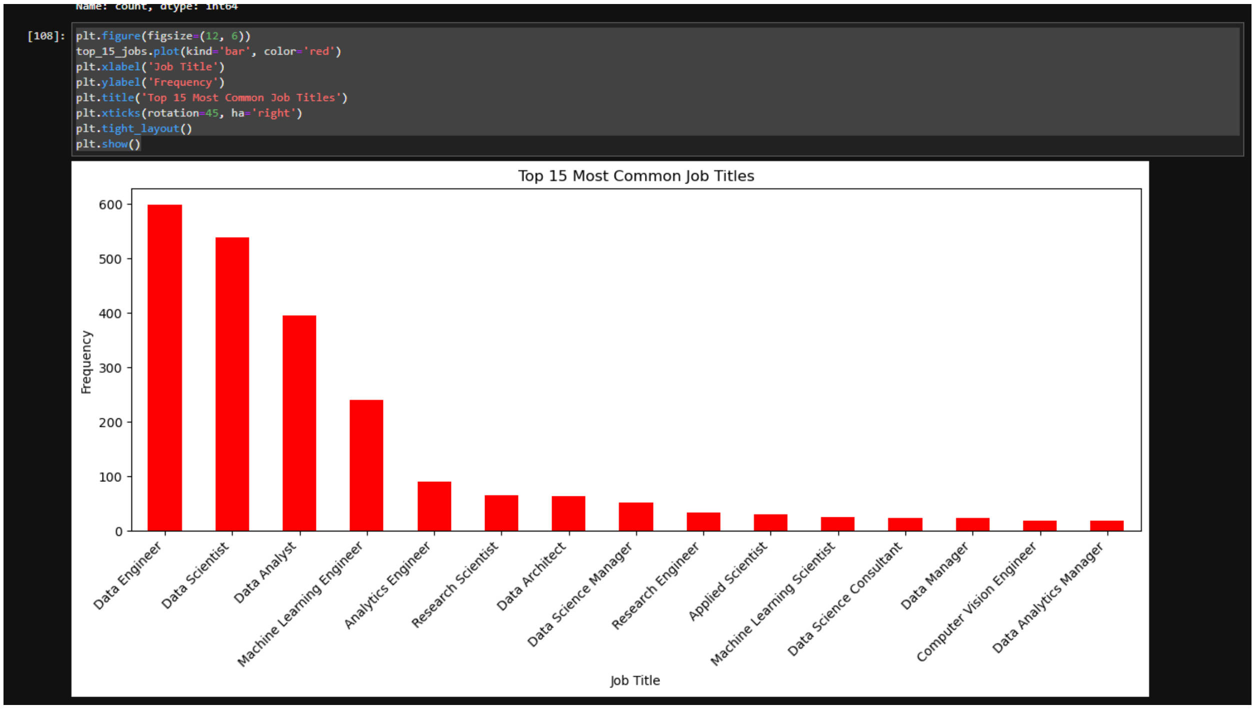
5.1 Write a python program to find out top 15 jobs. Make a bar graph of sales as well.



```
[93]: top_15_jobs = dataf['job_title'].value_counts().head(15)
      print(top_15_jobs)

      job_title
      Data Engineer                598
      Data Scientist               538
      Data Analyst                 396
      Machine Learning Engineer    240
      Analytics Engineer            91
      Research Scientist            65
      Data Architect                64
      Data Science Manager          52
      Research Engineer             33
      Applied Scientist             31
      Machine Learning Scientist    26
      Data Science Consultant       23
      Data Manager                  23
      Computer Vision Engineer      18
      Data Analytics Manager        18
      Name: count, dtype: int64
```

Figure 17 python program to find out top 15 jobs.

These lines of code print out top 15 jobs based on job title's frequency.

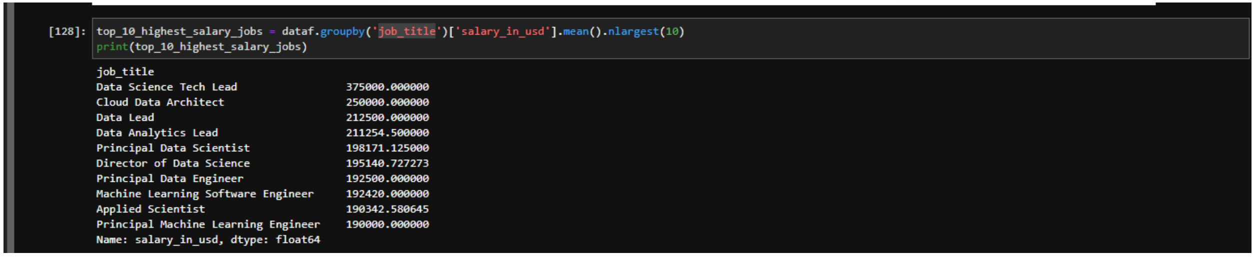


```
Name: count, dtype: int64

[108]: plt.figure(figsize=(12, 6))
       top_15_jobs.plot(kind='bar', color='red')
       plt.xlabel('Job Title')
       plt.ylabel('Frequency')
       plt.title('Top 15 Most Common Job Titles')
       plt.xticks(rotation=45, ha='right')
       plt.tight_layout()
       plt.show()
```

Figure 18 Plotting bar graph of top 15 job titles.

These lines of code plots top 15 most common job titles with proper x and y labelling and their title.

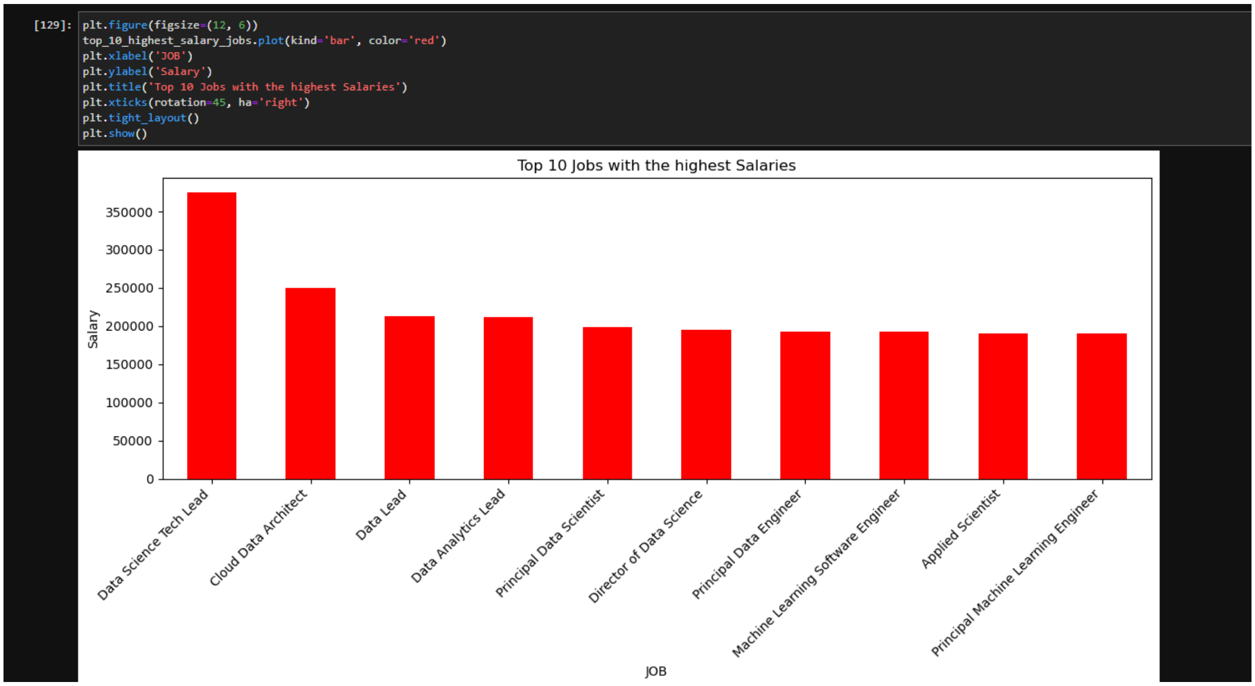
5.2 Which job has the highest salaries? Illustrate with bar graph.



```
[128]: top_10_highest_salary_jobs = dataf.groupby('job_title')['salary_in_usd'].mean().nlargest(10)
       print(top_10_highest_salary_jobs)

       job_title
       Data Science Tech Lead                  375000.000000
       Cloud Data Architect                    250000.000000
       Data Lead                               212500.000000
       Data Analytics Lead                     211254.500000
       Principal Data Scientist                198171.125000
       Director of Data Science                195140.727273
       Principal Data Engineer                 192500.000000
       Machine Learning Software Engineer      192420.000000
       Applied Scientist                       190342.580645
       Principal Machine Learning Engineer     190000.000000
       Name: salary_in_usd, dtype: float64
```

Figure 19 python program to find out highest salaries based on job titles

This line of code uses groupby() function to find out highest salaries of 10 job titles based on job titles.

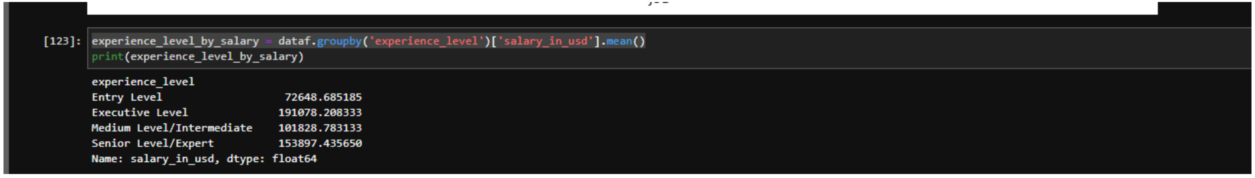


```
[129]: plt.figure(figsize=(12, 6))
       top_10_highest_salary_jobs.plot(kind='bar', color='red')
       plt.xlabel('JOB')
       plt.ylabel('Salary')
       plt.title('Top 10 Jobs with the highest Salaries')
       plt.xticks(rotation=45, ha='right')
       plt.tight_layout()
       plt.show()
```

Figure 20 bar graph to find out jobs with the highest salaries

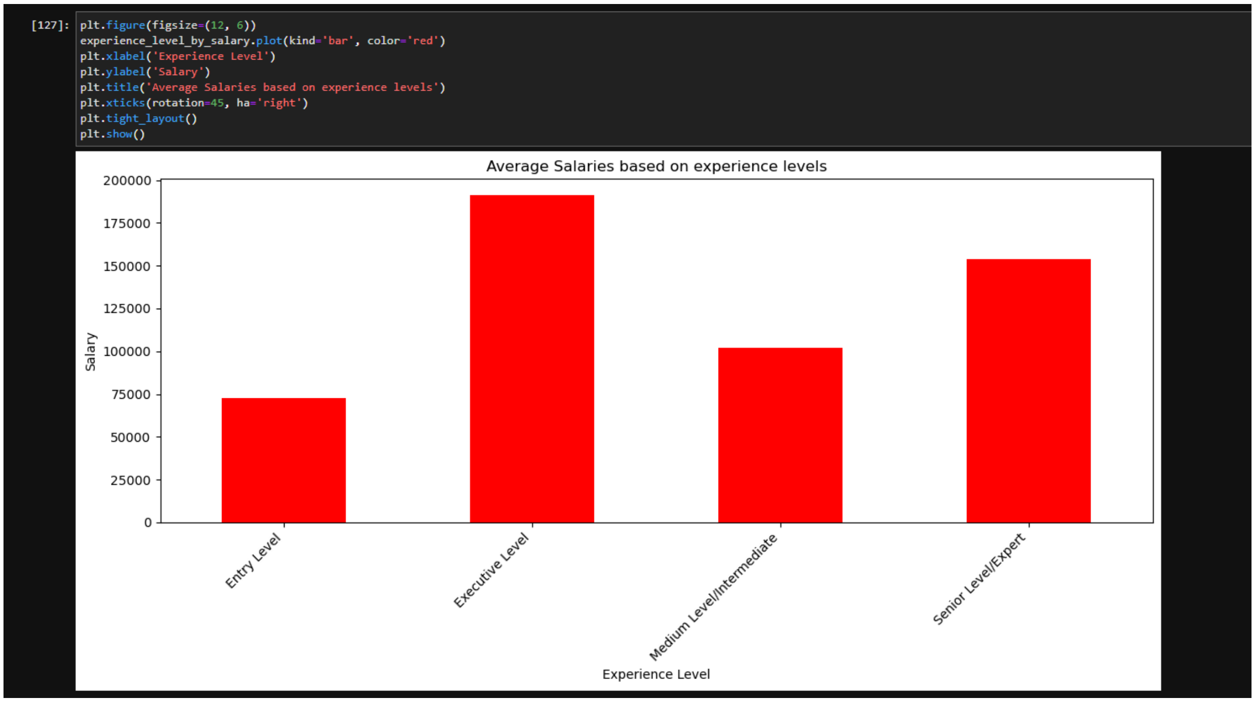
This bar graph plots different top 10 job titles which has highest average salaries.

5.3 Write a python program to find out salaries based on experience level. Illustrate it through bar graph.



```
[123]: experience_level_by_salary = dataf.groupby('experience_level')['salary_in_usd'].mean()
       print(experience_level_by_salary)

       experience_level
       Entry Level                    72648.685185
       Executive Level               191078.208333
       Medium Level/Intermediate     101828.783133
       Senior Level/Expert           153897.435650
       Name: salary_in_usd, dtype: float64
```

Figure 21 python program to find out salaries based on experience levels.

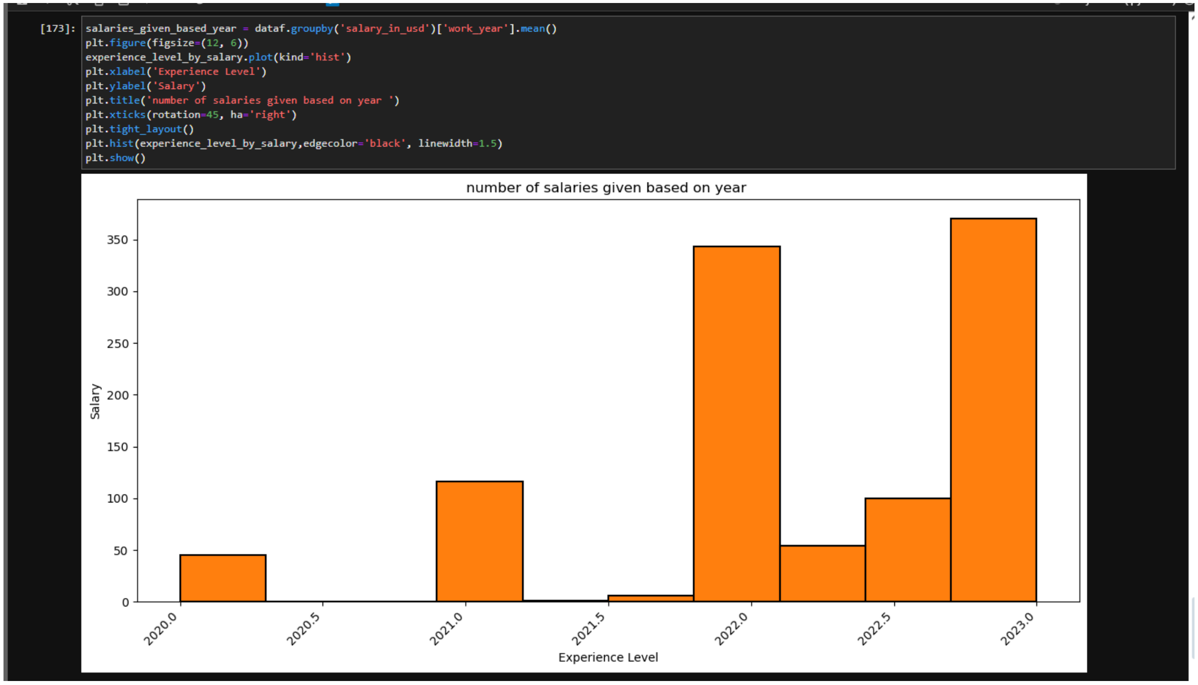


```
[127]: plt.figure(figsize=(12, 6))
       experience_level_by_salary.plot(kind='bar', color='red')
       plt.xlabel('Experience Level')
       plt.ylabel('Salary')
       plt.title('Average Salaries based on experience levels')
       plt.xticks(rotation=45, ha='right')
       plt.tight_layout()
       plt.show()
```

Figure 22 bar graph to plot average salaries based on experience levels

This code of line plots average salaries based on experience levels of the employees in the company.

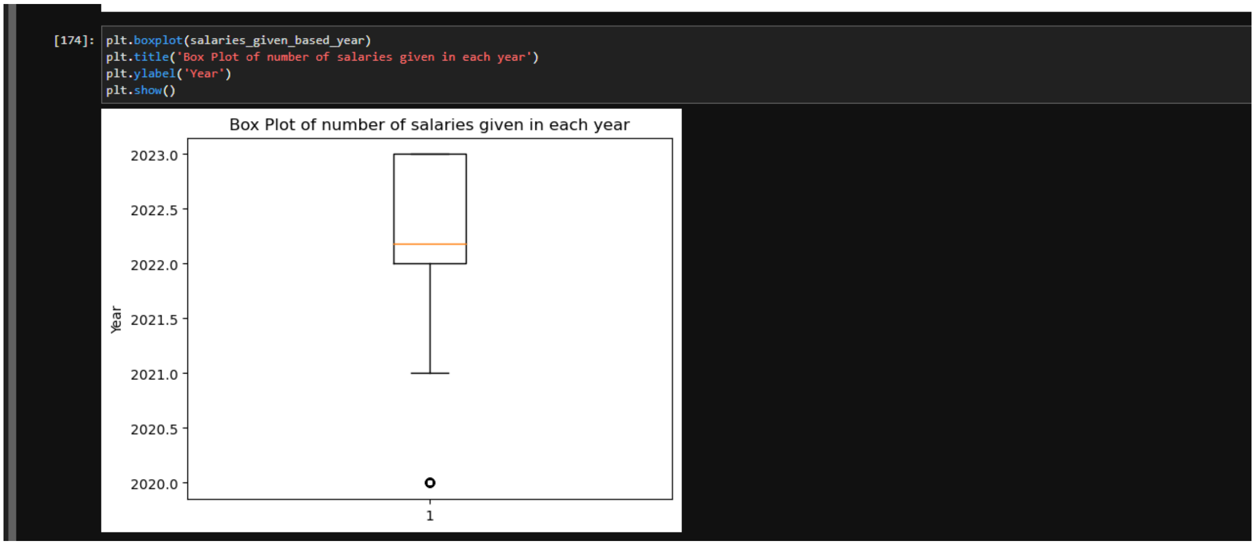
5.4 Write a Python program to show histogram and box plot of any chosen different variables. Use proper labels in the graph.



```
[173]: salaries_given_based_year = dataf.groupby('salary_in_usd')['work_year'].mean()
       plt.figure(figsize=(12, 6))
       experience_level_by_salary.plot(kind='hist')
       plt.xlabel('Experience Level')
       plt.ylabel('Salary')
       plt.title('number of salaries given based on year ')
       plt.xticks(rotation=45, ha='right')
       plt.tight_layout()
       plt.hist(experience_level_by_salary,edgecolor='black', linewidth=1.5)
       plt.show()
```

Figure 23 histogram to plot number of salaries which was given based on the year.

These code of lines plots the number of salaries which was given every year from 2020 to 2023 in the histogram.



```
[174]: plt.boxplot(salaries_given_based_year)
       plt.title('Box Plot of number of salaries given in each year')
       plt.ylabel('Year')
       plt.show()
```

Figure 24 box plot to plot the number of salaries given in each year.

These codes of lines plots in box plot to show the number of salaries given to the employees each year from 2020 to 2023.

6. References

datacamp. (2023, july). *What is Data Analysis? An Expert Guide With Examples*. Retrieved from datacamp: https://www.datacamp.com/blog/what-is-data-analysis-expert-guide

Geek for Geeks. (2024, January 25). *ML | Overview of Data Cleaning*. Retrieved from Geek for Geeks: https://www.geeksforgeeks.org/data-cleansing-introduction/

Qlik. (2024). *Data Exploration*. Retrieved from Qilk: https://www.qlik.com/us/data-analytics/data-exploration

Secoda. (2024). *What is Data Preperation?* Retrieved from Secoda: https://www.secoda.co/glossary/data-preperation-definition