LONDON METROPOLITAN UNIVERSITY

ITAHARI INTERNATIONAL COLLEGE

**Module Code & Module Title**

**CC5067NT Smart Data Discovery**

**Assessment Weightage & Type**

**60% Individual Coursework**

**Year and Semester**

**2023 Autmn**

**Student Name: Aashish Neupane**

**Group: L2C2**

**London Met ID: 22072025**

**College ID: NP05CP4A220004**

**Assignment Due Date: 13th May 2024**

**Assignment Submission Date: 13th May 2024**

*I confirm that I understand my coursework needs to be submitted online via My second teacher 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.*

## Table of Contents

## Table of figures

## Table of tables

# 1. Introduction

This coursework involves the task based on data exploration and research to better understanding the factors that influence salaries in the data science field. The dataset at hand includes a variety of factors such as experience, work level, job title and many more, all potential impacting salary levels.

So, my work is to gain a better understanding of the elements that impact the salaries of data scientists and to identify any patterns or tendencies in the data. I have utilized a variety of tools to finish my task, such as:

- **Pandas:** Pandas is a widely used python library for data science and analysis. It is mostly popular for activities like reading data from CSV or excel files, arranging and analyzing it, and finding relevant information. We can also use pandas to clean up messy data, target on specific portions and produce helpful visuals. (greeksforgreeks, 2023)

- **Jupyter Notebook:** Jupyter Notebook is a open source web tool where we can use it to create and share documents with code, math equations, graph, and text. This application originated form the IPython project, which has its own notebook project. It is managed by the same people who work on Project Jupyter. (Driscoll, 2024)

- **MS Word:** MS Word is a word processing application created by the Microsoft team that is a popular tool for writing and generating documents. It meets all the requirements for documentation. So, I used to create documentation for our project.

## 2. Data Understanding

Data understanding involves understanding what the data is for, the demands it will meet, its content and where it is located. There are not any physical tools for data understanding because it is expressed in business glossaries, dictionaries of data, models, and other place where information about the data is maintained. (Ladley, 2016)

The main objective of the analysis is to gain a proper understanding of the factors that impacts the salaries of data scientists as well as to identify any underlying structures or pattern within the dataset. The provided dataset consists of 3755 rows and 11 columns respectively which mainly focuses on data science job information. It examines many aspects such as experience level, job title, company size and so on. As there are 11 column which contained different information is fully described below in table form. This dataset is useful resource to examine developments and patterns in the field of data science.

In data analysis and exploration, summary statistics and correlation analyses are the key tools utilized alongside the creation of visualizations like bar graphs and histograms. These tools help to provide deeper insights into different aspects like finding top 15 jobs, finding the highest salaries of jobs based on chosen variable, facilitating a more effective understanding of the data.

In conclusion, data understanding is an important phase in both data science and machine learning. The given dataset includes work year, experience level of the employee, employment type, job title, employee salary, salary currency, salary in usd, employee residence, remote ratio, company location, company size. To fully understand sales patterns, summary statistics, correlation analysis, and visual representations like as bar graphs and histograms are used.

The dataset column details are described below:

| S.N. | Columns name | Description | Data Type |
|---|---|---|---|
| 1. | work_year | Indicates the year in which the job was completed. | Integer |
| 2. | experience_level | Defines the employees' level of experience. SE that stands for Senior, ML for Middle-level and EN for Entry-level respectively. | Object |
| 3. | employment_type | Specifies the type of employment that the individual is performing where FT mean Full-Time, CT means Contract-Time, PT means Part-Time and FL means Freelance respectively. | Object |
| 4. | Job_title | Describes the name of the job performed by the employee. | object |
| 5. | salary | Indicates the employee's salary. | integer |
| 6. | Salary_currency | Defines the currency in which the salary is being paid. | object |
| 7. | Salary_in_usd | Contains the salary translated to usd. | integer |
| 8. | Employee_residence | Indicates the country where the staff members belong. | objects |
| 9. | Remote_ratio | Defines the employer's permitted ratio of remote worker where 100 for completely remote and 0 for no remote work. | integer |

| | | | |
|---|---|---|---|
| 10. | Company_location | Indicates the place where the corporate employees are working. | object |
| 11. | Company_size | Defines the company size where letters S, M, and L stand for Small, Medium, and Large respectively. | object |

*Table 1: datasets columns details.*

## 3. Data Preparation

Data preparation can be defined as the process of modifying raw data in order to make that respective data ready for analysis and processing. Raw data contains errors, duplication, and missing values, which has an impact on the accuracy of data as well as data-driven decision-making. It is very crucial since it may be responsible for up to 80% of the effort required in a machine learning project. It is critical to use specialized data preparation tools to speed and enhance this process.

The importance of Data Preparation is:

- Improving Data Quality.
- Enhancing the value.
- Enabling Data Analysis.
- Improving Data Consumption.
- Extracting Unstructured Data. (Khan, 2024)

### 3.1 Write a python program to load data into pandas DataFrame.


*Figure 1: importing pandas and NumPy.*


*Figure 2: Data preparation -1.*

Out[2]:

| | work_year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio | company_location | c |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **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

*Figure 3: Data preparation -1 (Output).*

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

```
In [3]: # remove the unnecessary columns like salary and salary currency
        df.drop(columns=['salary','salary_currency'], inplace = True)
        df  #showing the data after succesfully removing salary and salary_currency
```

*Figure 4: Data preparation -2.*

Out[3]:

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

3755 rows × 9 columns

*Figure 5: Data preparation -2 (output).*

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

```
In [4]: df.dropna(inplace = True)
        df
```

*Figure 6: Data preparation -3.*

Out[4]:

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

3755 rows × 9 columns

*Figure 7: Dat preparation -3 (Output).*

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

```
In [5]: df.duplicated() #to find the duplicated value
```

*Figure 8: Data preparation -4.*

```
Out[5]: 0       False
        1       False
        2       False
        3       False
        4       False
                ...
        3750    False
        3751    False
        3752    False
        3753    False
        3754    False
        Length: 3755, dtype: bool
```

*Figure 9: Data preparation -4 (Output).*

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

```
In [6]: for column in df.columns:
            unique=df[column].unique()
            print(f"Unique values '{column}': {unique}")
```

*Figure 10: Data preparation -5.*

```
Unique values 'work_year': [2023 2022 2020 2021]
Unique values 'experience_level': ['SE' 'MI' 'EN' 'EX']
Unique values 'employment_type': ['FT' 'CT' 'FL' 'PT']
Unique values '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'
 '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 'salary_in_usd': [ 85847  30000  25500 ...  28369 412000  94665]
Unique values '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 'remote_ratio': [100   0  50]
Unique values '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 'company_size': ['L' 'S' 'M']
```

*Figure 11: Data preparation -5 (Output).*

```
In [7]: df.nunique() # to count the unique values of columns

Out[7]: work_year              4
        experience_level       4
        employment_type        4
        job_title             93
        salary_in_usd       1035
        employee_residence    78
        remote_ratio           3
        company_location      72
        company_size           3
        dtype: int64
```

*Figure 12: Counting the unique value of each column.*

### 3.6 Rename the experience level columns as below.

**SE – Senior Level/Expert**

**MI – Medium Level/Intermediate**

**EN – Entry Level**

**EX – Executive Level**

```
In [8]: mapping = {
            'SE':'Senior level/Expert',
            'MI':'Medium level/Intermediate',
            'EN':'Entry Level',
            'EX': 'Executive Level'
        }
        df['experience_level']= df['experience_level'].replace(mapping)

In [9]: df
```

*Figure 13: Data preparation -6.*

Out[9]:

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

3755 rows × 9 columns

*Figure 14: Data preparation -6 (Output).*

# 4. Data Analysis

Data Analysis is the process of inspecting, filtering, cleaning, and arranging data in order to get insights and make better decisions. As a data analyst the role belongs to them are analyzing extensive datasets, identifying hidden patterns, and converting statistics into actionable information. (Simplilearn, 2024)

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

```
In [10]: #first, outlining the summary statistics of the dataframe column
summary_statistics = df.describe()
df.describe()
```

*Figure 15: showing the summary statistics.*

Out[10]:

| | work_year | salary_in_usd | remote_ratio |
|---|---|---|---|
| count | 3755.000000 | 3755.000000 | 3755.000000 |
| mean | 2022.373635 | 137570.389880 | 46.271638 |
| std | 0.691448 | 63055.625278 | 48.589050 |
| min | 2020.000000 | 5132.000000 | 0.000000 |
| 25% | 2022.000000 | 95000.000000 | 0.000000 |
| 50% | 2022.000000 | 135000.000000 | 0.000000 |
| 75% | 2023.000000 | 175000.000000 | 100.000000 |
| max | 2023.000000 | 450000.000000 | 100.000000 |

*Figure 16: showing the summary statistics (Output).*

```
In [25]: #Adding summary staticits of sum, mean, standard deviation,skewness and Kurtosis for salary_in_usd

summary_statistics= df['salary_in_usd'].describe()

summary_statistics['standard deviation'] = df['salary_in_usd'].std()
summary_statistics['skewness'] = df['salary_in_usd'].skew()
summary_statistics['kurtosis'] = df['salary_in_usd'].kurtosis()

print(summary_statistics)
```

*Figure 17: Data Analysis -1.*

```
count                3755.000000
mean               137570.389880
std                 63055.625278
min                  5132.000000
25%                 95000.000000
50%                135000.000000
75%                175000.000000
max                450000.000000
standard deviation  63055.625278
skewness                0.536401
kurtosis                0.834006
Name: salary_in_usd, dtype: float64
```

*Figure 18: Data Analysis -1 (Output).*

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

```
In [12]: correlation = df[['work_year','salary_in_usd','remote_ratio']]
         correlation.corr()
```

*Figure 19: Data Analysis -2.*

Out[12]:

| | work_year | salary_in_usd | remote_ratio |
|---|---|---|---|
| **work_year** | 1.00000 | 0.228290 | -0.236430 |
| **salary_in_usd** | 0.22829 | 1.000000 | -0.064171 |
| **remote_ratio** | -0.23643 | -0.064171 | 1.000000 |

*Figure 20: Data Analysis -2 (Output).*

```
In [13]: print(df.columns)

Index(['work_year', 'experience_level', 'employment_type', 'job_title',
       'salary_in_usd', 'employee_residence', 'remote_ratio',
       'company_location', 'company_size'],
      dtype='object')
```

*Figure 21: Showing columns.*

## 5. Data Exploration

Data exploration is defined as the first step in data analysis, during which analysts uses techniques like data visualization and statistics to define dataset parameters such as volume, size, and accuracy respectively. This helps to improve understanding of the data nature. (heavy.AI, 2024)

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

```
In [14]: import matplotlib.pyplot as plot

In [15]: top15_jobs = df['job_title'].value_counts().head(15)
         print('The top 15 jobs are shown below: ',)
         print(top15_jobs)

         The top 15 jobs are shown below:
         job_title
         Data Engineer                 1040
         Data Scientist                 840
         Data Analyst                   612
         Machine Learning Engineer      289
         Analytics Engineer             103
         Data Architect                 101
         Research Scientist              82
         Data Science Manager            58
         Applied Scientist               58
         Research Engineer               37
         ML Engineer                     34
         Data Manager                    29
         Machine Learning Scientist      26
         Data Science Consultant         24
         Data Analytics Manager          22
         Name: count, dtype: int64
```

*Figure 22: first, showing top 15 jobs.*

```
In [16]: plot.figure(figsize=(10, 6))
         top15_jobs.plot(kind='bar',  color = 'darkgoldenrod', edgecolor= 'black')
         plot.title('Top 15 jobs')
         plot.xlabel('Jobs_List')
         plot.ylabel('Frequency')
         plot.xticks(rotation=90, ha='right')
         plot.tight_layout()
         plot.show()
```

*Figure 23: Data Exploration -1 (Bar graph).*

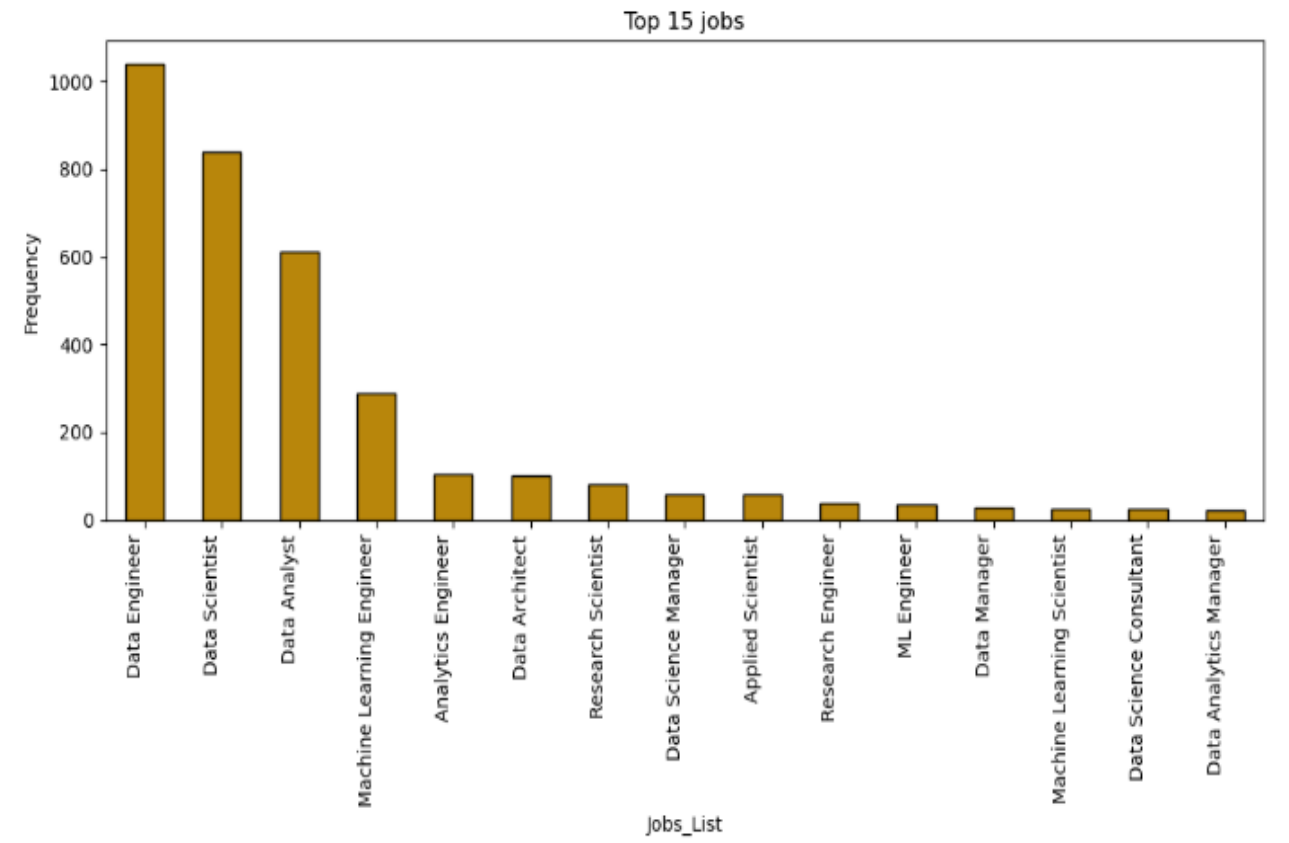


*Figure 24: Data Exploration -1 (Bar graph: Output).*

The code calculates the instances of every job list included in the dataset's 'job_title' column, identifying the top 15 most often appearing titles. The top 15 titles are then displayed in a bar chart, with the titles themselves listed on the x-axis and their frequency represented on the y-axis. The Data engineer ranks as the most popular job title followed by Data Scientist and so on, and the resulting graphic clearly represents the most common job titles. The top 15 job titles have been split using the 'head (15)' function.

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

```
In [17]: highest_salary = df.groupby('job_title')['salary_in_usd'].max().sort_values(ascending=False).head(10)

plot.figure(figsize=(12, 8))
highest_salary.plot(kind='bar',  color = 'darkgoldenrod', edgecolor= 'black')
plot.title('Highest Salary by Job_title')
plot.xlabel('Job_title')
plot.ylabel('Salary in USD')
plot.title('Top 15 Jobs with Highest Salary')
plot.xticks(rotation=90)
plot.tight_layout()
plot.show()
```

*Figure 25: Data Exploration -2 (code).*

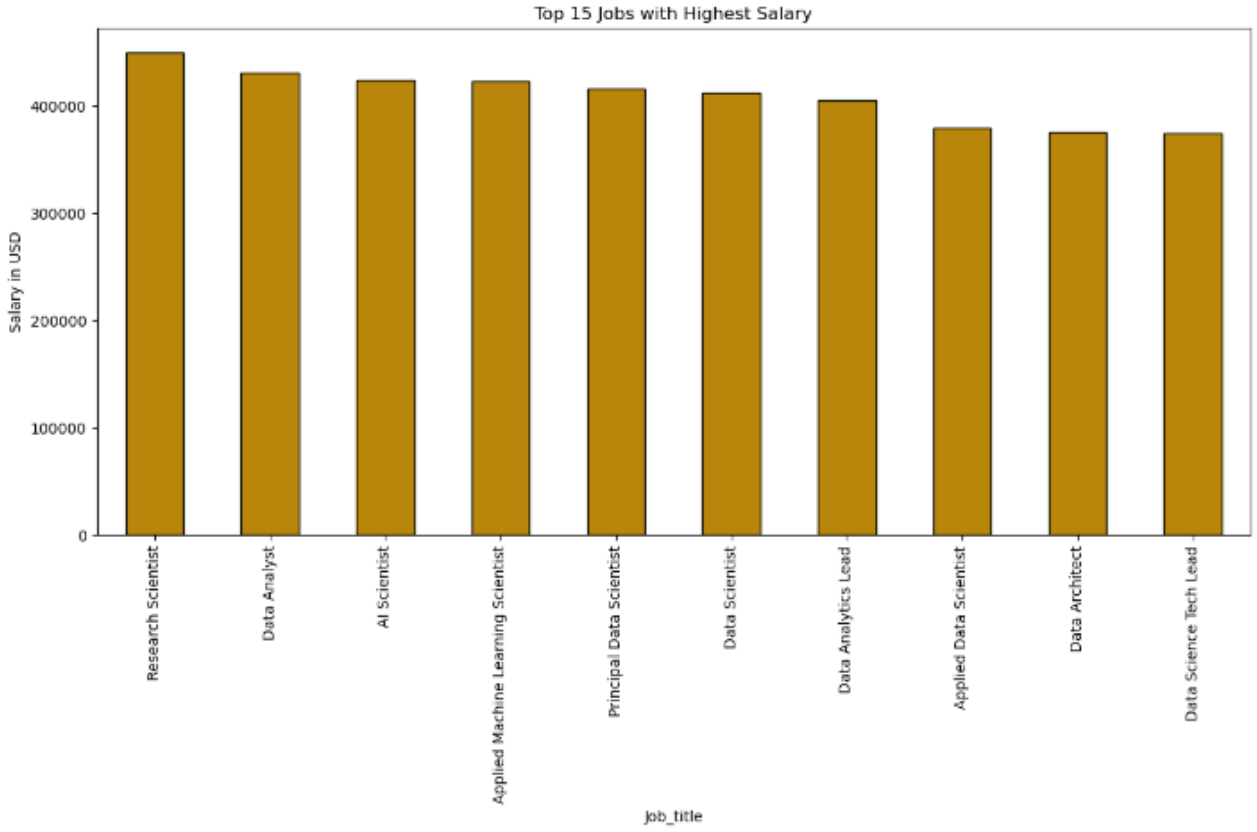


*Figure 26: Data Exploration -2 (Output).*

The above code displays the highest salaries of each job where I show the top 10 jobs with highest salaries respectively. The output is display in the bar graph on the x- axis with job title and their corresponding highest salaries in USD on y-axis. The graph is plotted successively.

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

```
In [18]: ex_label = df.groupby('experience_level')['salary_in_usd'].max()
top_salaries = ex_label.sort_values(ascending=False).head(15)
plot.figure(figsize=(12, 10))
top_salaries.plot(kind='bar',  color = 'darkgoldenrod', edgecolor= 'black')
plot.xlabel('Experience Level')
plot.ylabel('Max Salary in usd')
plot.title('Top Highest Salaries based on Experience Level')
plot.xticks(rotation=0)
plot.tight_layout()
plot.show()
```

*Figure 27: Data Exploration -3 (Code).*

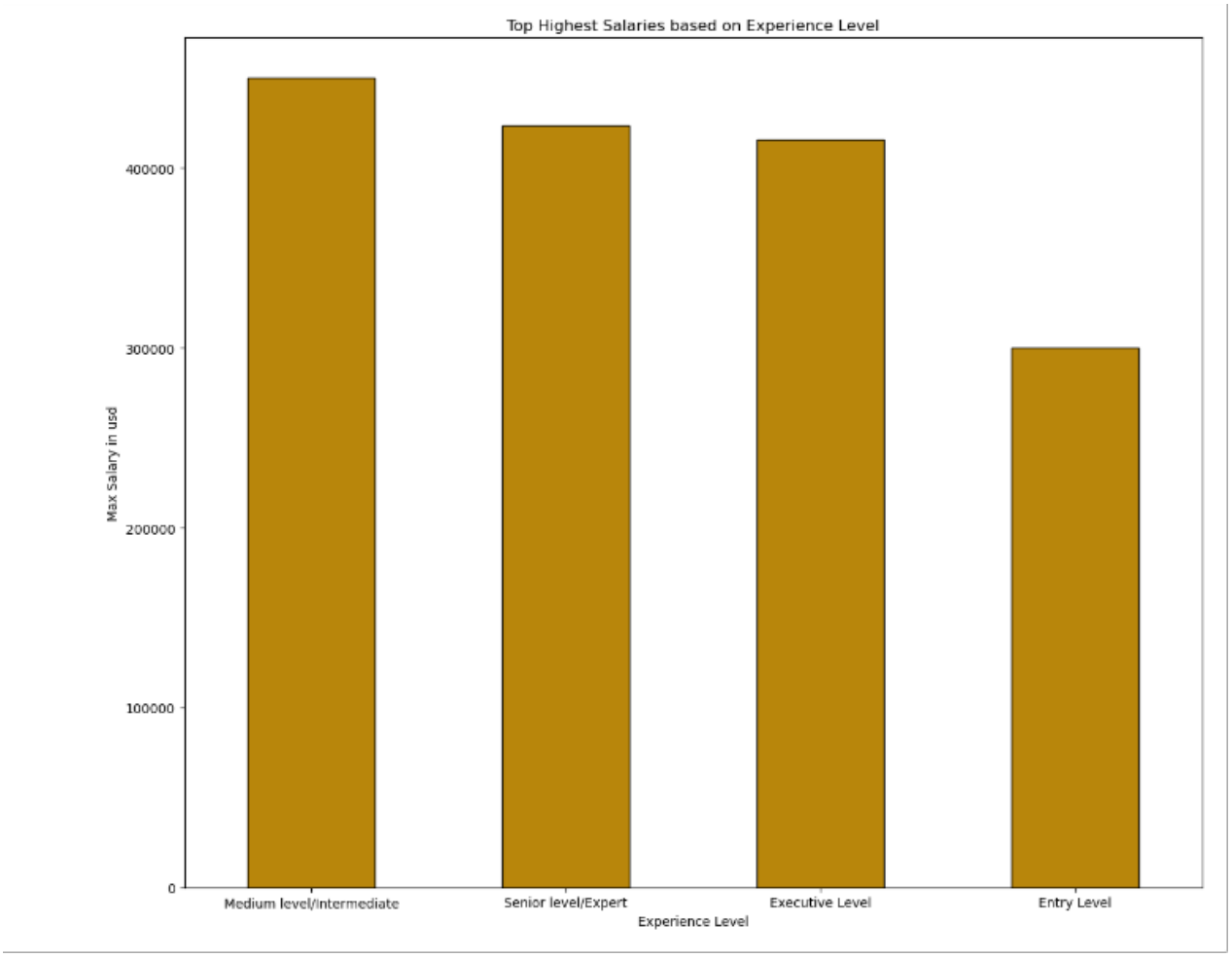


*Figure 28: Data Exploration -3 (Output).*

The code clearly displays the highest salaries based on the Experience Level. The output is displayed in the bar graph where x-axis represent the Experience level and y-axis shows the Max salaries in usd respectively.

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

- Histogram Plot for salary_in_usd:

```
In [24]: #histogram plot
plot.figure(figsize=(12,8))
plot.hist(df['salary_in_usd'], bins = 10, color = 'darkgoldenrod', edgecolor= 'black')

#add labels and title
plot.xlabel('salary_in_usd')
plot.ylabel('Frequency')
plot.title('Histogram plot of salary_in_usd')

#showing output
plot.show()
```

*Figure 29: Data Exploration -4 (Histogram Plot: Code).*

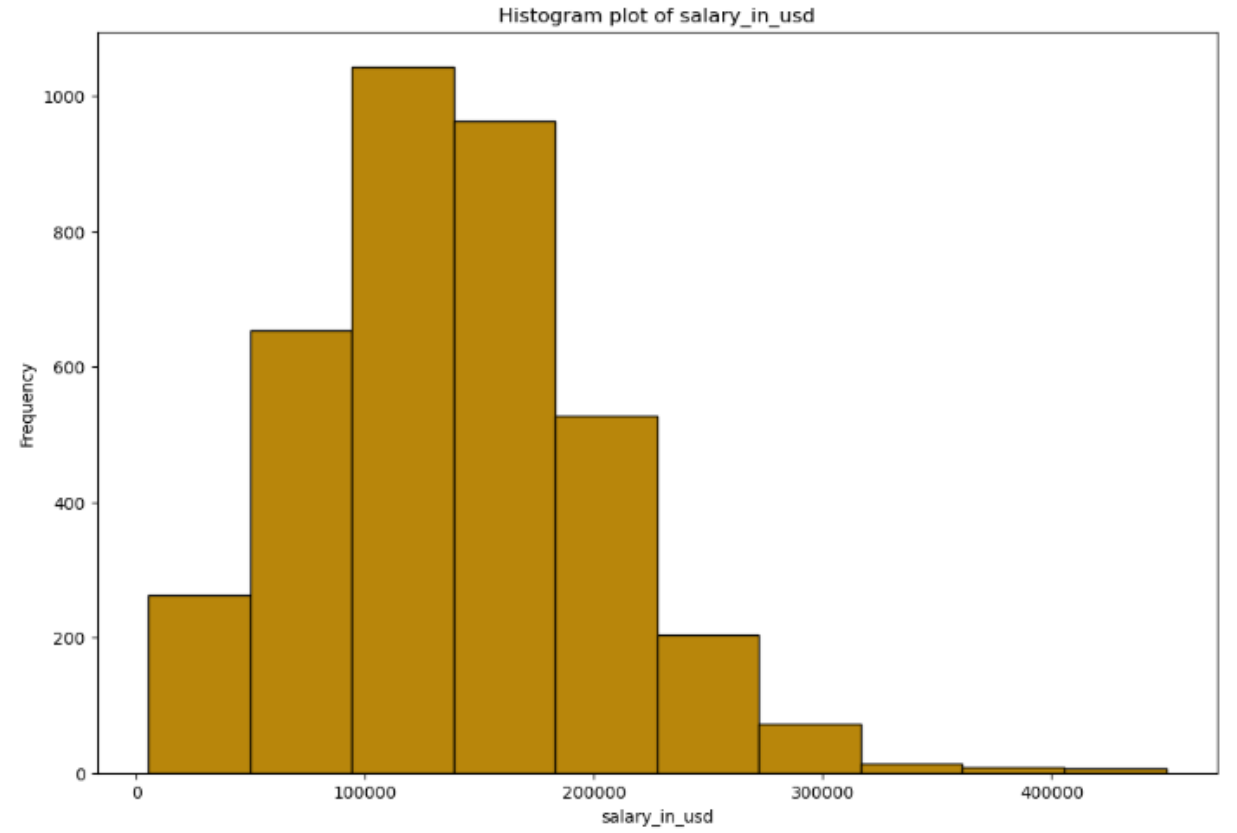


*Figure 30: Data Exploration -4 (Histogram Plot: Output).*

The above code is done to display the histogram plot for 'salary_in_usd' variable in the dataset which the histogram clearly displays the frequency distribution of respective variable values. The histogram is done successfully.

- Box plot for salary_in usd:

```
In [22]: #box plot
plot.figure(figsize=(10,6))
plot.boxplot(df['salary_in_usd'])
plot.title('box plot of salary in usd')
plot.ylabel('salary_in_usd')
plot.show()
```

*Figure 31: Data Exploration -4 (box Plot: Code).*

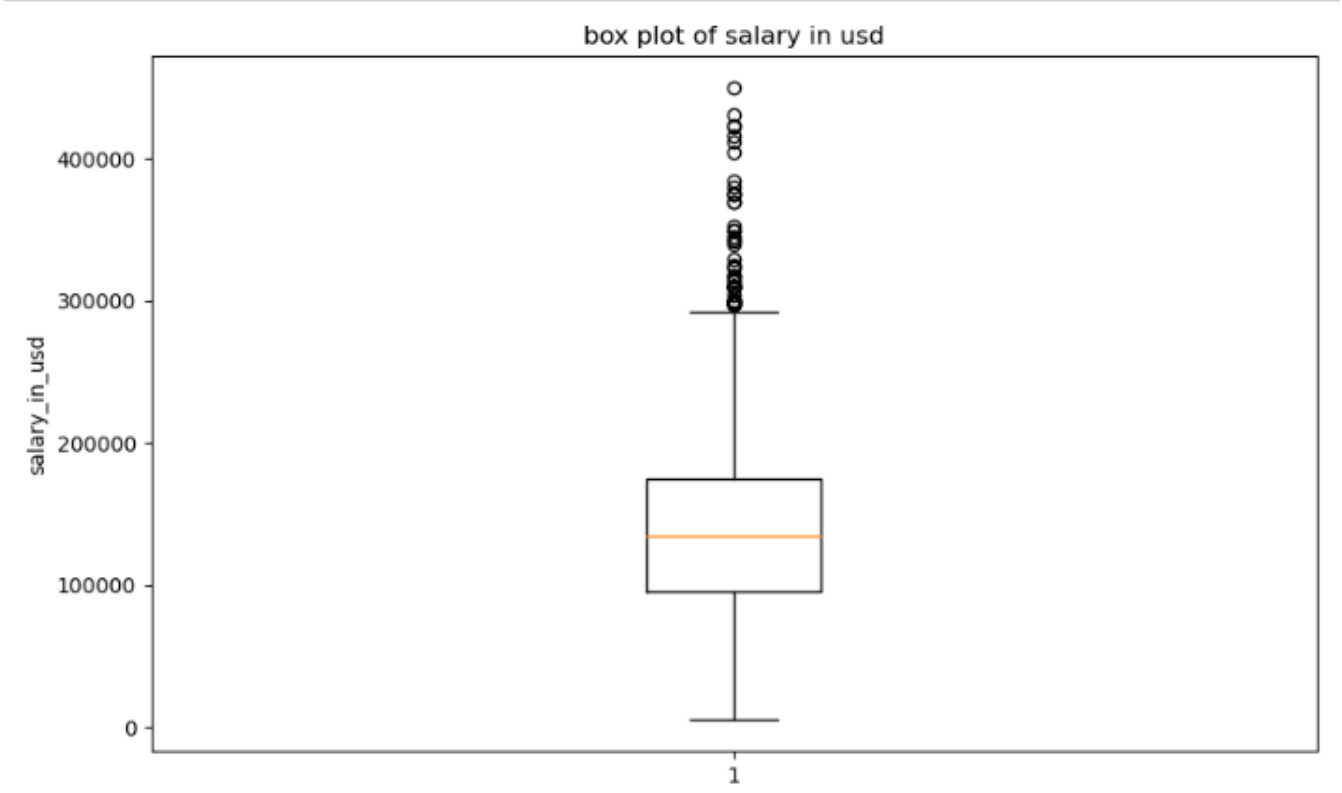


*Figure 32: Data Exploration -4 (box Plot: Output).*

Also, the box plot diagram for the 'salary_in_usd' variable is drawn, providing insights into the central tendency, and spread of the data. The box plot is successfully generated.

# 6. Conclusion

In conclusion, the given coursework is a significant individual assessment which accounts for 60% of the module grade, focusing on the application of programming skills to data analysis problems. The main goal is to evaluate the salary information for data scientists using python programming and technical report writing respectively, finding factors that influence salaries and discovering trends within the datasets. Here, required data understanding, preparation, analysis, and exploration of the data are included in the coursework materials. Completing this assessment is going to enable me to showcase my analytical, problem-solving, and critical assessment abilities. Also, this project allows me to learn important programming skills and utilize them to tackle actual data analysis challenges.

I had several challenges and mistakes, but I am grateful to our module leader, Mr. Projesh Basnet sir, for leading us positively, motivating us, and clarifying the importance of this project in a real-life scenario. I also viewed a YouTube video to understand more about the problem-solving concept. Also, many thanks to our college for providing us with this module, which helps to improve the knowledge of these topics, as well as access to MySecondTeacher, which allows us to track down the daily material of the course and access it from anywhere for self-study.

## 7. References

Driscoll, M., 2024. *real python.* [Online] Available at: https://realpython.com/jupyter-notebook-introduction/ [Accessed 25 April 2024].

greeksforgreeks, 2023. *greeksforgreeks.* [Online] Available at: https://www.geeksforgeeks.org/pandas-functions-in-python/ [Accessed 25 April 2024].

heavy.AI, 2024. *heavy.AI.* [Online] Available at: https://www.heavy.ai/heavyeco/heavyeco-overview [Accessed 06 May 2024].

Khan, F., 2024. *Astera.* [Online] Available at: https://www.astera.com/type/blog/data-preparation/ [Accessed 26 April 2024].

Ladley, J., 2016. *CIO.* [Online] Available at: https://www.cio.com/article/238649/mastering-and-managing-data-understanding.html [Accessed 26 April 2024].

Simplilearn, 2024. *siplielearn.com.* [Online] Available at: https://www.simplilearn.com/data-analysis-methods-process-types-article [Accessed 06 May 2024].

# Appendix

## Originality report

COURSE NAME
CC5067 Smart Data Discovery

STUDENT NAME
AASHISH NEUPANE

FILE NAME
22072025 Aashish Neupane (1).docx

REPORT CREATED
May 13, 2024

### Summary

| | | |
|---|---|---|
| Flagged passages | 19 | 18% |
| Cited/quoted passages | 0 | 0% |

Web matches

| | | |
|---|---|---|
| chegg.com | 12 | 12% |
| coursehero.com | 4 | 4% |
| kaggle.com | 1 | 1% |
| heavy.ai | 1 | 0.7% |
| geeksforgeeks.org | 1 | 0.7% |

1 of 19 passages

Student passage FLAGGED

**I confirm that I understand my coursework needs to be submitted online via** My second teacher **under the relevant module page before the deadline in order for my assignment to be accepted and marked**...

Top web match

Individual Coursework 2022-23 Student Name: Mahima Chaudhary London Met ID: College ID: Assignment Due Date: Thursday, May 4, 2023 Assignment Submission Date: Wednesday, May 3, 2023 Word Count: 2229 **I**...

21039883 Mahima Chaudhary.docx - CC5067NI-Smart Data... https://www.coursehero.com/file/233030575/21039883-Mahima-Chaudharydocx/

2 of 19 passages

Student passage FLAGGED

**2. Data Preparation**2.1 **Write a python program to load data into pandas DataFrame.**

Top web match

1 0 Marks ) **2 . Data Preparation Write a python program to load data into pandas DataFrame** ( 5 Marks ) Write a python program to remove unnecessary columns i

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

3 of 19 passages

Student passage FLAGGED

**Write a python program to remove the NaN missing values from updated dataframe**.2.4 **Write a python program to check duplicates value in the dataframe**.2.5 **Write a python program to see the unique values**...

Top web match

**Write a python program to remove the NaN missing values from updated dataframe**...**Write a python program to check duplicates value in the dataframe**...**Write a python program to see the unique values from**...

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

4 of 19 passages

Student passage FLAGGED

**3. Data Analysis**3.1 **Write a Python program to show summary statistics of sum, mean, standard deviation, skewness, and kurtosis of any chosen variable**.3.2 **Write a Python program to calculate and show**...

Top web match

**3 . Data Analysis Write a Python program to show summary statistics of sum, mean, standard deviation, skewness, and kurtosis of any chosen variable**...**Write a Python program to calculate and show**...

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

5 of 19 passages

Student passage FLAGGED

**Write a python program to find out salaries based on experience level. Illustrate it through bar graph**.4.4 **Write a Python program to show histogram and box plot of any chosen different variables. Use**...

Top web match

**Write a python program to find out salaries based on experience level. Illustrate it through bar graph**… **Write a Python program to show histogram and box plot of any chosen different variables. Use**…

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

6 of 19 passages

Student passage FLAGGED

**The** dataset at hand includes a variety of **factors such as experience, work level, job title and many more**, all potential impacting salary levels.So, my work **is to** gain **a better understanding of the**…

Top web match

dont use AI tools for plagarism free and write the whole python code Data Set DescriptionThe data contains **the** information about various **factors** which can influence salary levels **such as experience,**…

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

7 of 19 passages

Student passage FLAGGED

…Jupyter Notebook: **Jupyter Notebook is** a **open source web** tool where we can use it **to create and share documents** with **code**, math **equations**, graph, **and text**

Top web match

The **Jupyter Notebook is** an **open-source web** application that allows you **to create and share documents** that contain live **code, equations**, visualizations **and** narrative **text**. Uses include data cleaning…

Using Matplotlib with Jupyter Notebook - GeeksforGeeks https://www.geeksforgeeks.org/using-matplotlib-with-jupyter-notebook/

---

8 of 19 passages

Student passage FLAGGED

**The** main **objective of** the **analysis is to** gain **a** proper **understanding of the** factors **that** impacts **the salaries of data scientists** as well as to identify **any** underlying structures **or** pattern **within the**

Top web match

**The objective of** this **analysis is to** obtain **a** better **understanding of the** elements **that** influence **the salaries of data scientists** and discover **any** regularities **or** tendencies **within the** data.

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

9 of 19 pas ages
Student assa

**given dataset includes work year, experience level** of the employee, **employment type, job title**, employee **salary, salary currency, salary in usd, employee residence, remote ratio, company location,**...

Top web match

This strength is particularly useful **given** the diverse nature of our **dataset**, which **includes** features such as **work year , experience level , employment type , job title , salary , salary currency ,**...

DS Salary : Full EDA | Geo | Cluster + XGboost - Kaggle https://www.kaggle.com/code/tumpanjawat/ds-salary-full-eda-geo-cluster-xgboost

---

10 of 19 passages
Student passage FLAGGED

2.1 **Write a python program to load data into pandas DataFrame.**

Top web match

Data Preparation **Write a python program to load data into pandas DataFrame** ( 5 Marks ) Write a python program to remove unnecessary columns i

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

11 of 19 passages
Student passage FLAGGED

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

Top web match

4 2.2. **Write a python program to remove the NaN missing values from updated dataframe**... 5 2.3.

21039883 Mahima Chaudhary.docx - CC5067NI-Smart Data... https://www.coursehero.com/file/233030575/21039883-Mahima-Chaudharydocx/

---

12 of 19 passages
Student passage FLAGGED

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

Top web match

5 Marks ) **Write a python program to check duplicates value in the dataframe.** ( 5 Marks

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

13 of 19 pa sages
Student assa

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

Top web match

5 Marks ) **Write a python program to see the unique values from all the columns in the dataframe.** ( 5 Marks

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

14 of 19 passages
Student passage FLAGGED

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

Top web match

8 2.6. **Write a Python program to show summary statistics of sum, mean, standard deviation, skewness, and kurtosis of any chosen variable**... 9 2.7.

21039883 Mahima Chaudhary.docx - CC5067NI-Smart Data... https://www.coursehero.com/file/233030575/21039883-Mahima-Chaudharydocx/

---

15 of 19 passages
Student passage FLAGGED

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

Top web match

9 2.7. **Write a Python program to calculate and show correlation of all variables**... 10 3.

21039883 Mahima Chaudhary.docx - CC5067NI-Smart Data... https://www.coursehero.com/file/233030575/21039883-Mahima-Chaudharydocx/

---

16 of 19 passages
Student passage FLAGGED

**Data exploration** is defined as **the** first **step in data analysis**, during **which analysts** uses techniques like **data visualization and** statistics **to** define **dataset** parameters **such as** volume, **size, and**...

Top web match

Data exploration definition: **Data exploration** refers to **the** initial **step in data analysis** in **which** data **analysts** use **data visualization and** statistical techniques **to** describe **dataset**...

Data Exploration - A Complete Introduction - HEAVY.AI https://www.heavy.ai/learn/data-exploration

17 of 19 pa sages

Student assa

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

Top web match

Data Exploration **Write a python program to find out top** 1 5 jobs. **Make a bar graph of sales as well**. ( 1 0 Marks

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

18 of 19 passages

Student passage FLAGGED

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

Top web match

1 0 Marks ) **Write a python program to find out salaries based on experience level**. **Illustrate it through bar graph**. ( 1 0 Marks

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---

19 of 19 passages

Student passage FLAGGED

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

Top web match

1 0 Marks ) **Write a Python program to show histogram and box plot of any chosen different variables**. **Use proper labels in the graph**. ( 1 0 Marks

Solved dont use AI tools for plagarism free and write the | Chegg.com https://www.chegg.com/homework-help/questions-and-answers/dont-use-ai-tools-plagarism-free-write-whole-python-code-data-set-description-data-contain-q170004344

---