

**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**

[1. Introduction 1](#_Toc166492468)

[2. Data Understanding 2](#_Toc166492469)

[3. Data Preparation 5](#_Toc166492470)

[3.1 Write a python program to load data into pandas DataFrame. 5](#_Toc166492471)

[3.2 Write a python program to remove unnecessary columns i.e., salary and salary currency. 7](#_Toc166492472)

[3.3 Write a python program to remove the NaN missing values from updated dataframe. 8](#_Toc166492473)

[3.4 Write a python program to check duplicates value in the dataframe. 9](#_Toc166492474)

[3.5 Write a python program to see the unique values from all the columns in the dataframe. 10](#_Toc166492475)

[3.6 Rename the experience level columns as below. 12](#_Toc166492476)

[4. Data Analysis 13](#_Toc166492477)

[4.1 Write a Python program to show summary statistics of sum, mean, standard deviation, skewness, and kurtosis of any chosen variable. 13](#_Toc166492478)

[4.2 Write a Python program to calculate and show correlation of all variables. 15](#_Toc166492479)

[5. Data Exploration 16](#_Toc166492480)

[5.1 Write a python program to find out top 15 jobs. Make a bar graph of sales as well. 16](#_Toc166492481)

[5.2 Which job has the highest salaries? Illustrate with bar graph. 18](#_Toc166492482)

[5.3 Write a python program to find out salaries based on experience level. Illustrate it through bar graph. 19](#_Toc166492483)

[5.4 Write a Python program to show histogram and box plot of any chosen different variables. Use proper labels in the graph. 20](#_Toc166492484)

[6. Conclusion 22](#_Toc166492485)

[7. References 23](#_Toc166492486)

**Table of figures**

[Figure 1: importing pandas and NumPy. 5](#_Toc166492527)

[Figure 2: Data preparation -1. 5](#_Toc166492528)

[Figure 3: Data preparation -1 (Output). 6](#_Toc166492529)

[Figure 4: Data preparation -2. 7](#_Toc166492530)

[Figure 5: Data preparation -2 (output). 7](#_Toc166492531)

[Figure 6: Data preparation -3. 8](#_Toc166492532)

[Figure 7: Dat preparation -3 (Output). 8](#_Toc166492533)

[Figure 8: Data preparation -4. 9](#_Toc166492534)

[Figure 9: Data preparation -4 (Output). 9](#_Toc166492535)

[Figure 10: Data preparation -5. 10](#_Toc166492536)

[Figure 11: Data preparation -5 (Output). 10](#_Toc166492537)

[Figure 12: Counting the unique value of each column. 11](#_Toc166492538)

[Figure 13: Data preparation -6. 12](#_Toc166492539)

[Figure 14: Data preparation -6 (Output). 12](#_Toc166492540)

[Figure 15: showing the summary statistics. 13](#_Toc166492541)

[Figure 16: showing the summary statistics (Output). 13](#_Toc166492542)

[Figure 17: Data Analysis -1. 14](#_Toc166492543)

[Figure 18: Data Analysis -1 (Output). 14](#_Toc166492544)

[Figure 19: Data Analysis -2. 15](#_Toc166492545)

[Figure 20: Data Analysis -2 (Output). 15](#_Toc166492546)

[Figure 21: Showing columns. 15](#_Toc166492547)

[Figure 22: first, showing top 15 jobs. 16](#_Toc166492548)

[Figure 23: Data Exploration -1 (Bar graph). 17](#_Toc166492549)

[Figure 24: Data Exploration -1 (Bar graph: Output). 17](#_Toc166492550)

[Figure 25: Data Exploration -2 (code). 18](#_Toc166492551)

[Figure 26: Data Exploration -2 (Output). 18](#_Toc166492552)

[Figure 27: Data Exploration -3 (Code). 19](#_Toc166492553)

[Figure 28: Data Exploration -3 (Output). 19](#_Toc166492554)

[Figure 29: Data Exploration -4 (Histogram Plot: Code). 20](#_Toc166492555)

[Figure 30: Data Exploration -4 (Histogram Plot: Output). 20](#_Toc166492556)

[Figure 31: Data Exploration -4 (box Plot: Code). 21](#_Toc166492557)

[Figure 32: Data Exploration -4 (box Plot: Output). 21](#_Toc166492558)

**Table of tables**

[Table 1: datasets columns details. 4](#_Toc166489148)

# 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.

# 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 |
|  | work\_year | Indicates the year in which the job was completed. | Integer |
|  | experience\_level | Defines the employees’ level of experience. SE that stands for Senior, ML for Middle-level and EN for Entry-level respectively. | Object |
|  | 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 |
|  | Job\_title | Describes the name of the job performed by the employee. | object |
|  | salary | Indicates the employee’s salary. | integer |
|  | Salary\_currency | Defines the currency in which the salary is being paid. | object |
|  | Salary\_in\_usd | Contains the salary translated to usd. | integer |
|  | Employee\_residence | Indicates the country where the staff members belong. | objects |
|  | Remote\_ratio | Defines the employer’s permitted ratio of remote worker where 100 for completely remote and 0 for no remote work. | integer |
|  | Company\_location | Indicates the place where the corporate employees are working. | object |
|  | 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.

# 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)

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

A close-up of a screen

Description automatically generated

Figure 1: importing pandas and NumPy.



Figure 2: Data preparation -1.

A screen shot of a computer

Description automatically generated

Figure 3: Data preparation -1 (Output).

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

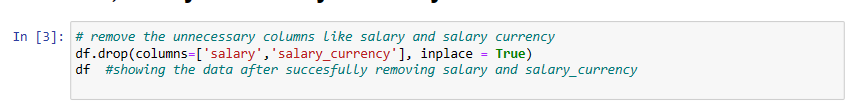


Figure 4: Data preparation -2.

A screenshot of a computer

Description automatically generated

Figure 5: Data preparation -2 (output).

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



Figure 6: Data preparation -3.

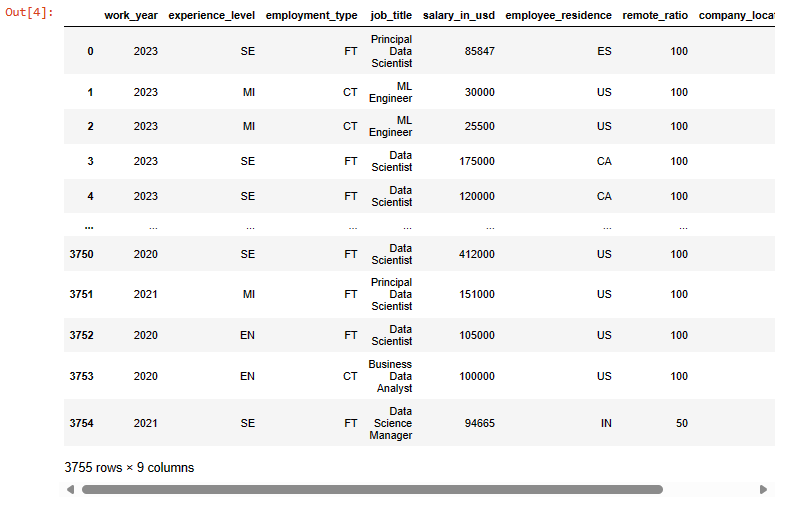


Figure 7: Dat preparation -3 (Output).

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



Figure 8: Data preparation -4.

A screenshot of a computer

Description automatically generated

Figure 9: Data preparation -4 (Output).

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

A white rectangular object with black text

Description automatically generated

Figure 10: Data preparation -5.

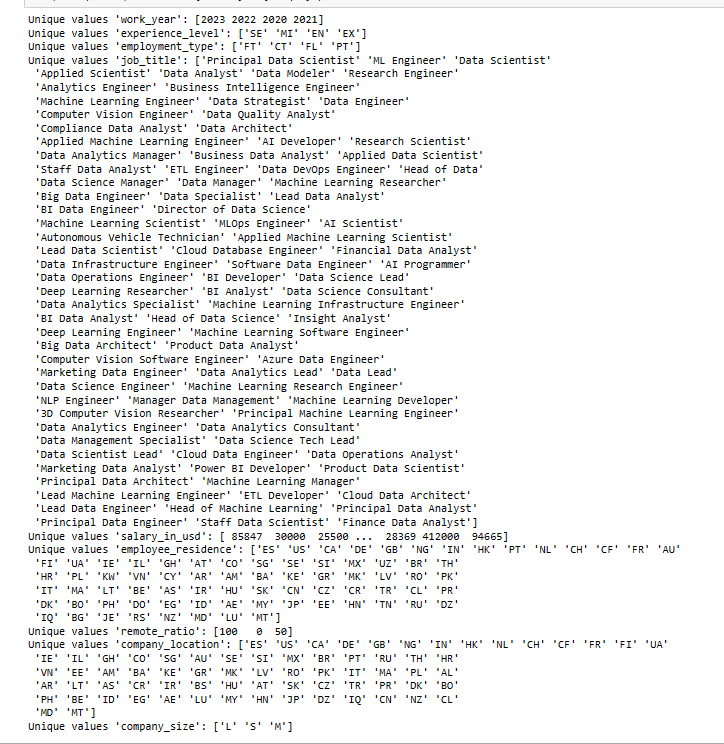


Figure 11: Data preparation -5 (Output).

A screenshot of a computer code

Description automatically generated

Figure 12: Counting the unique value of each column.

## Rename the experience level columns as below.

**SE – Senior Level/Expert**

**MI – Medium Level/Intermediate**

**EN – Entry Level**

**EX – Executive Level**

**A screen shot of a computer code

Description automatically generated**

Figure 13: Data preparation -6.

A screenshot of a computer screen

Description automatically generated

Figure 14: Data preparation -6 (Output).

# 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)

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

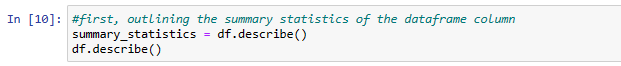


Figure 15: showing the summary statistics.

A table with numbers and text

Description automatically generated

Figure 16: showing the summary statistics (Output).

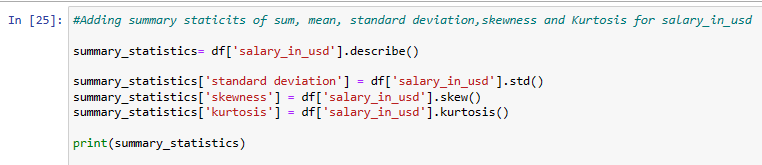


Figure 17: Data Analysis -1.

A screenshot of a computer

Description automatically generated

Figure 18: Data Analysis -1 (Output).

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

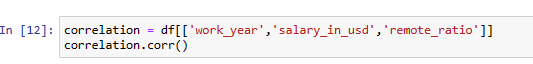


Figure 19: Data Analysis -2.

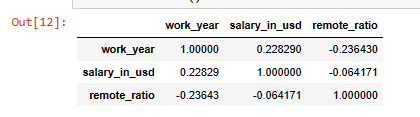


Figure 20: Data Analysis -2 (Output).

A screenshot of a computer

Description automatically generated

Figure 21: Showing columns.

# 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)

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

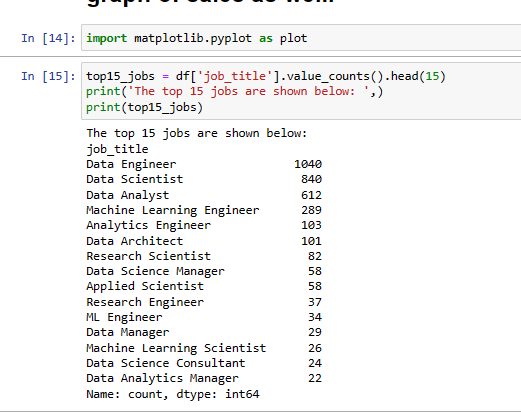


Figure 22: first, showing top 15 jobs.



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

A graph of jobs with text

Description automatically generated with medium confidence

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.

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



Figure 25: Data Exploration -2 (code).

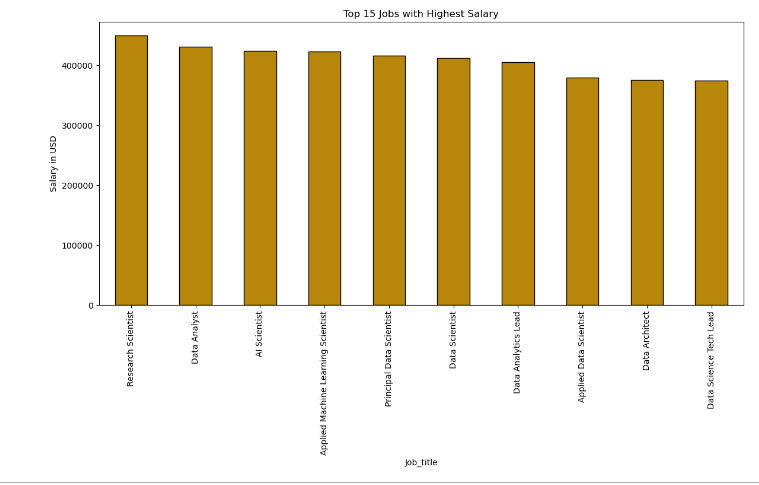


Figure 26: Data Exploration -2 (Output).

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.

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

A screen shot of a computer code

Description automatically generated

Figure 27: Data Exploration -3 (Code).

A graph of a bar chart

Description automatically generated

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.

## 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:

A white background with text

Description automatically generated

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

A graph of a graph

Description automatically generated

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:

A screen shot of a computer code

Description automatically generated

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

A diagram of a box

Description automatically generated

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.

# 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.

# 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**



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated