CC5067NI-Smart Data Discovery

60% Individual Coursework

2023-24 Spring

Student Name: Asmi Bajracharya

London Met ID: 22068053

College ID: np01cp4a220363

Assignment Due Date: Monday, May 13, 2024

Assignment Submission Date: Sunday, May 12, 2024

Word Count: 3057

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.

Acknowledgements

I would like to express my heartfelt gratitude to Islington college for giving me this opportunity to learn about this course. I would also like to express my gratitude to our lecturer, Mr. Dipeshor Silwal, for helping me understand this course as well as the coursework.

I would also like to thank our teacher, Mr. Alish KC, for guiding and supporting me throughout this coursework. His dedication towards teaching and his willingness to answer questions made the learning experience more valuable.

I would also like to thank everyone who helped me complete this coursework on time. A special thanks to all the lecturers and teachers for not just helping me for this coursework but also making this course of valuable learning experience.

Thank you,

Asmi Bajracharya.

Abstract

This coursework presents the analysis of data science salaries. It covers various topics of data analysis and uses Python as the base programming language. This coursework consists of a huge data set of data science salaries which contains information such as job title, residence, salary, experience level, work year and more. In this coursework, we are required to understand the data and analyze it in order to find interesting patterns or trends. By using Python as the main language and libraries such as pandas and matplotlib we were able to not just analyze data but also create bar graphs, histograms, and boxplots to visualize data.

Table of Contents

Table of Figures

Table of table

Introduction

This is our first ever coursework of Smart Data Discovery. In this coursework, we are supposed to analyse the data of the data set which is given to us. The name of the data set is DataScienceSalaries which contains different information about the salaries of individuals that are involved in this field and factors that influence their salaries. We are going to analyse the dataset and understand the data given to us.

Aims

This coursework aims to analyse the salaries of individual working in data science using Python and the DataScienceSalaries dataset.

Objectives

The objective of this coursework is to:

- Analyse the dataset.
- Fix data inconsistencies and duplicates.
- Make different charts and graphs to understand the data better.
- Find out correlations between data and
- Summarize key statistics of the dataset.

1. Data understanding

For this coursework, we were given a dataset, DataScienceSalaries, which contains the data of salaries in the fields of data science. A dataset is a collection of data that is generally related to each other and is typically in a systematic format (Sheldon, 2024). Data sets are used for various purposes such as data analysis, forecasting, building AI, etc.

Dataset for this project contains various information such as work year, experience level, employment type, job title, salary, salary currency, salary in USD, employee residence, remote work ratio, company location, and company size. All these factors influence the salary levels. This dataset contains 3755 rows and 11 columns.

Having 3755 rows and 11 columns means that the data set has a large amount of data. This leaves room for duplicate data and data inconsistences. While checking all the data, the dataset does not seem perfect and seems to contain duplicates and data inconsistencies and a few unnecessary rows.

Here is the table which summarizes the dataset:

S. No	Column Name	Description	Data Type
1	work_year	This column contains the year in which data was recorded.	Integer
2	experience_level	This column contains the experience level of each individual. For example, Entry level (EN), etc.	String
3	employment_type	This column contains the type of employment of an individual. For example, full time (FT), or part time (PT) etc.	String
4	job_title	This column contains the name of the job. For example, Data Analyst, etc.	String
5	salary	This column contains the salaries of individuals.	Integer
6	salary_currency	This column contains the currency of the salaries. For example, EUR, USD, etc.	String
7	salary_in_usd	This column contains the salaries in USD.	Integer
8	employee_residence	This column contains the location of the employee's residence. For example, US, CA, etc.	String
9	remote_ratio	This column contains the ratio of work done remotely compared to onsite.	Integer
10	company_location	This column contains the location of the company.	String
11	company_size	This column contains the size of the company. For example, S for small, Medium (M), or Large (L).	String

Table 1 Data set information table.

2. Data Preparation

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

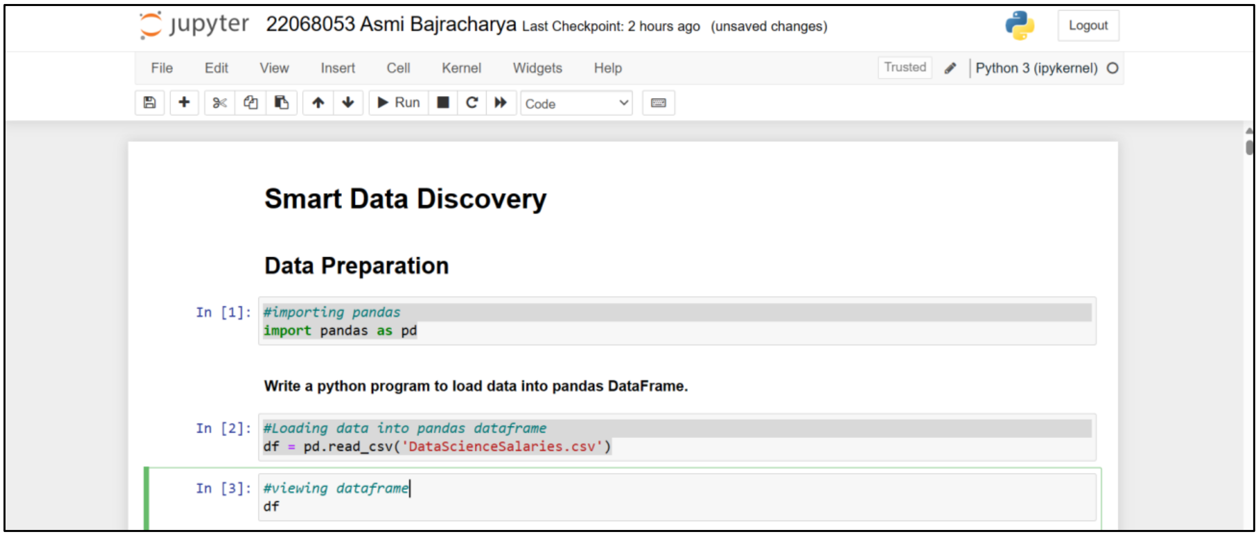
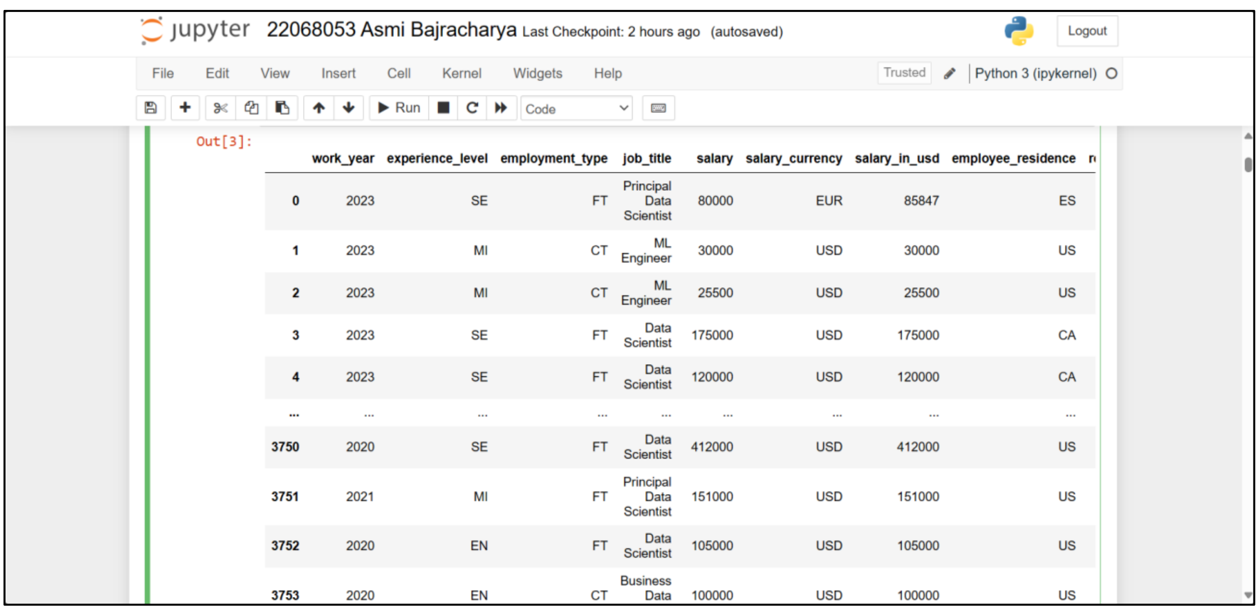


Figure 1 Importing pandas and loading data into pandas dataframe.

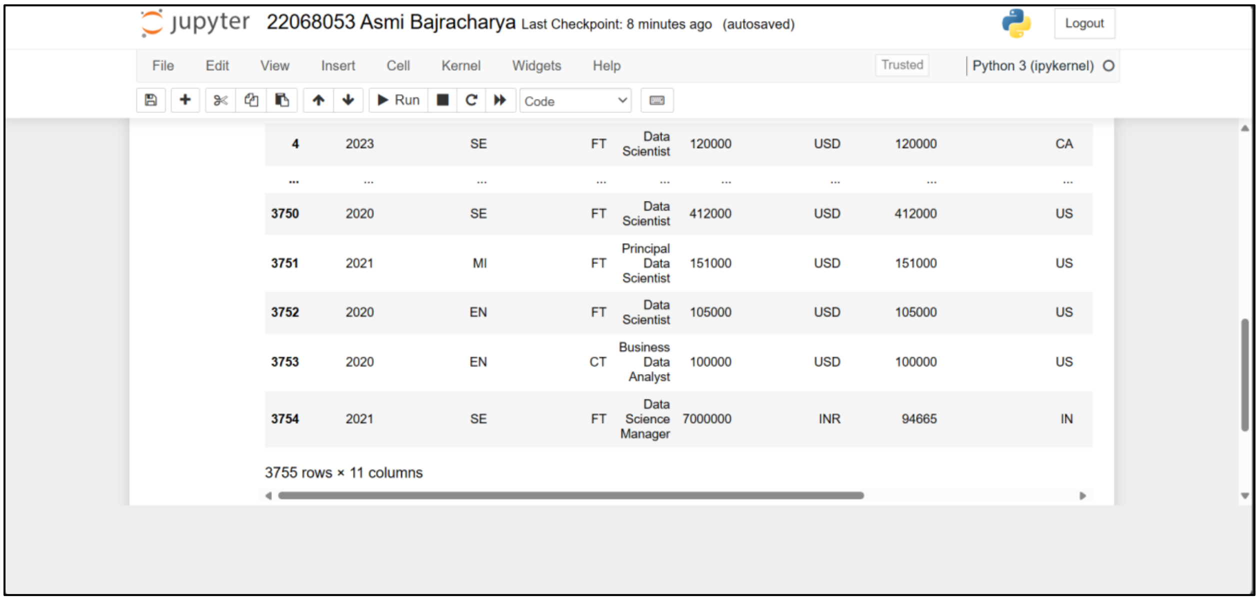
Here, the first step is to import pandas before loading it into a dataframe. The data set is loaded from a CVS file to a pandas dataframe and is stored in a variable, df. Now, the next step is to view the dataframe.



	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	r
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	
1	2023	MI	CT	ML Engineer	30000	USD	30000	US	
2	2023	MI	CT	ML Engineer	25500	USD	25500	US	
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA	
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA	
...	...	...	...	...	...	...	...	...	
3750	2020	SE	FT	Data Scientist	412000	USD	412000	US	
3751	2021	MI	FT	Principal Data Scientist	151000	USD	151000	US	
3752	2020	EN	FT	Data Scientist	105000	USD	105000	US	
3753	2020	EN	CT	Business Data	100000	USD	100000	US	

Figure 2 Data set (df)

As we can see in the figure above, this dataframe contains various columns and multiple rows containing different types of data. The dataframe consists of columns which includes work year, experience level, employment type, job title, salary, salary currency, salary in USD, employee residence, remote work ratio, company location, and company size.



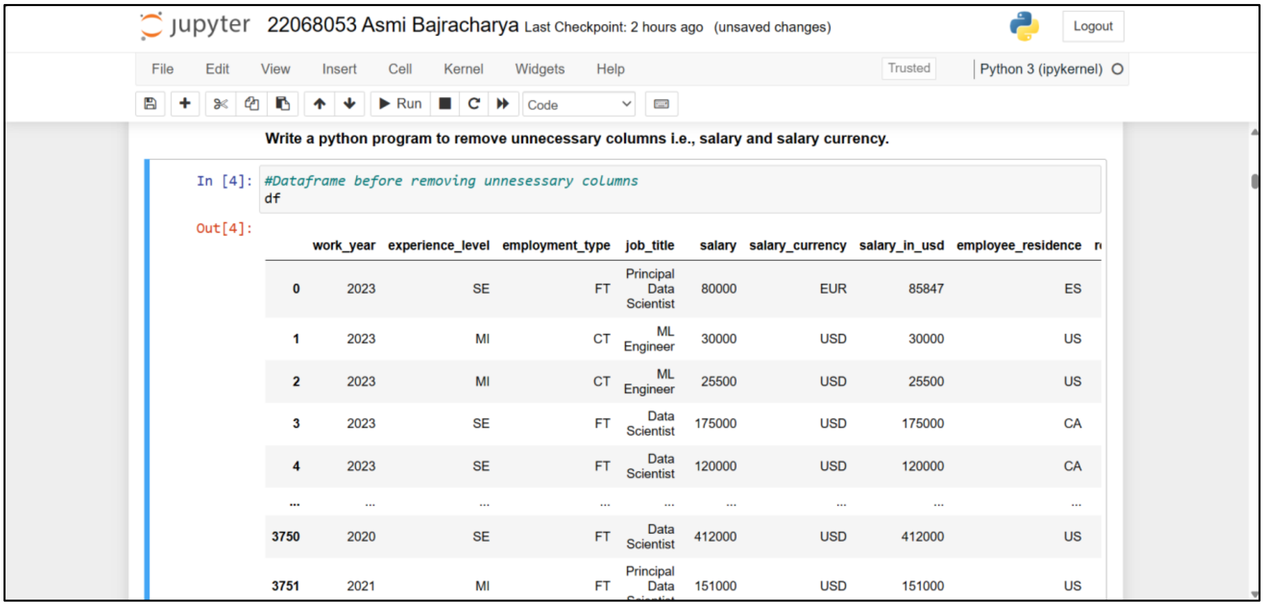
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA
...	...	...	...	...	...	...	...	...
3750	2020	SE	FT	Data Scientist	412000	USD	412000	US
3751	2021	MI	FT	Principal Data Scientist	151000	USD	151000	US
3752	2020	EN	FT	Data Scientist	105000	USD	105000	US
3753	2020	EN	CT	Business Data Analyst	100000	USD	100000	US
3754	2021	SE	FT	Data Science Manager	7000000	INR	94665	IN

3755 rows × 11 columns

Figure 3 Dataframe (df) continued.

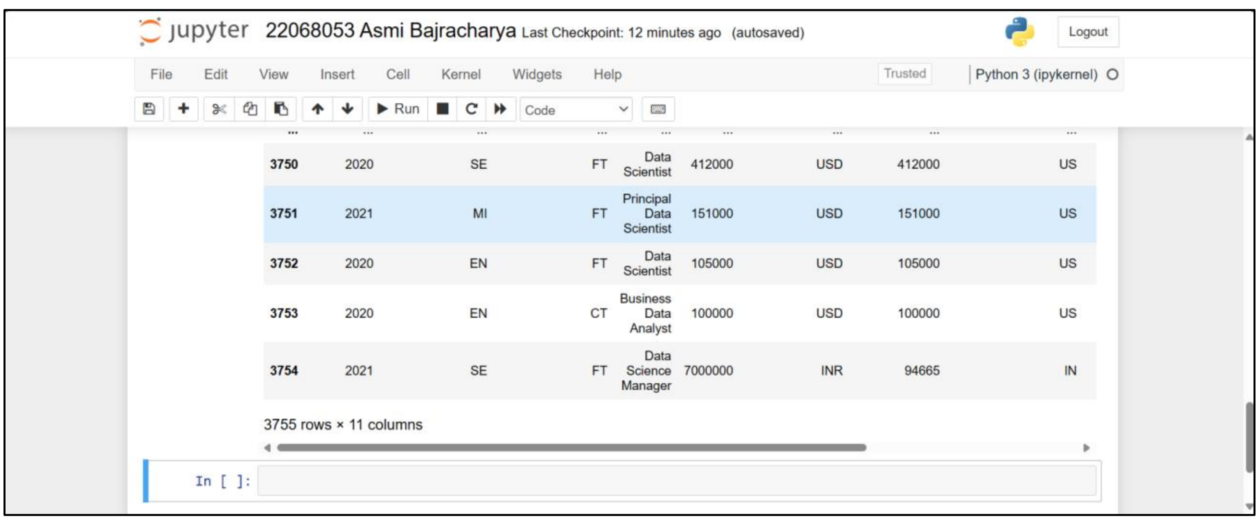
In the above figure, we can see that there are 3755 rows and 11 columns. This is the entire dataframe of DataScienceSalaries. Here, the output consists of a data frame which contains various data of the salaries of individuals and various factors that influence the salary.

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



	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES
1	2023	MI	CT	ML Engineer	30000	USD	30000	US
2	2023	MI	CT	ML Engineer	25500	USD	25500	US
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA
...	...	...	...	...	...	...	...	...
3750	2020	SE	FT	Data Scientist	412000	USD	412000	US
3751	2021	MI	FT	Principal Data	151000	USD	151000	US

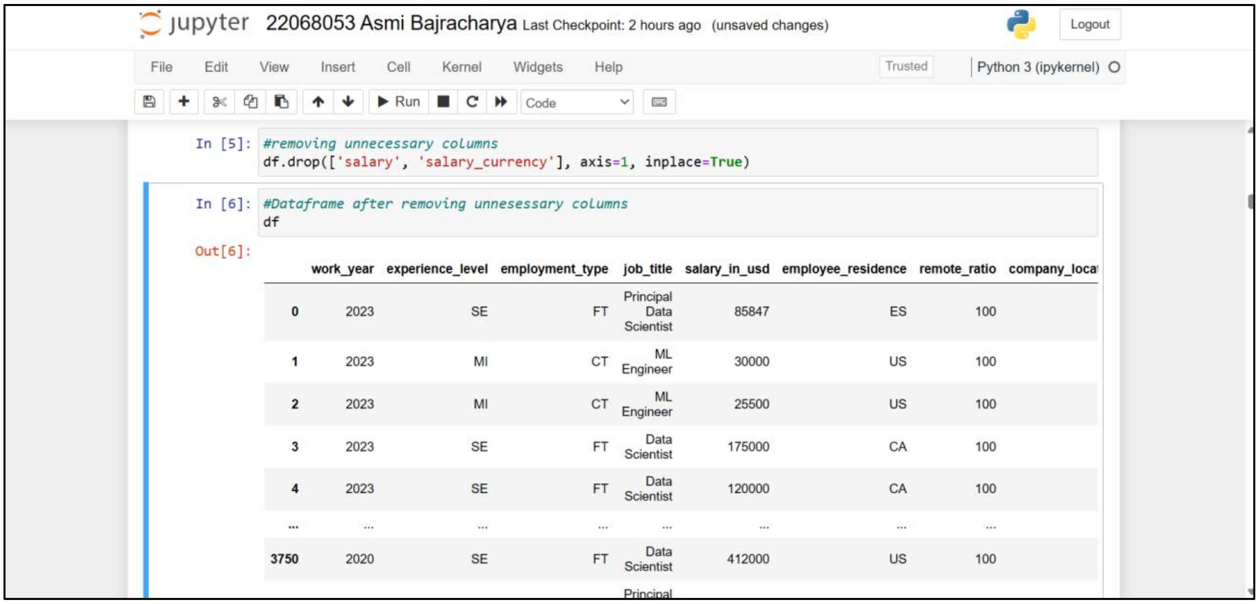
Figure 4 Dataframe before removing unnecessary columns.



...	...	...	...	...	...	...	...	...
3750	2020	SE	FT	Data Scientist	412000	USD	412000	US
3751	2021	MI	FT	Principal Data Scientist	151000	USD	151000	US
3752	2020	EN	FT	Data Scientist	105000	USD	105000	US
3753	2020	EN	CT	Business Data Analyst	100000	USD	100000	US
3754	2021	SE	FT	Data Science Manager	7000000	INR	94665	IN

Figure 5 Dataframe before removing unnecessary columns continued.

As we can see in the above figure, the same data is repeating in three different columns. Salary and salary currency is not required as a column since there already exists a column salary_in_usd which contains the salary and currency.

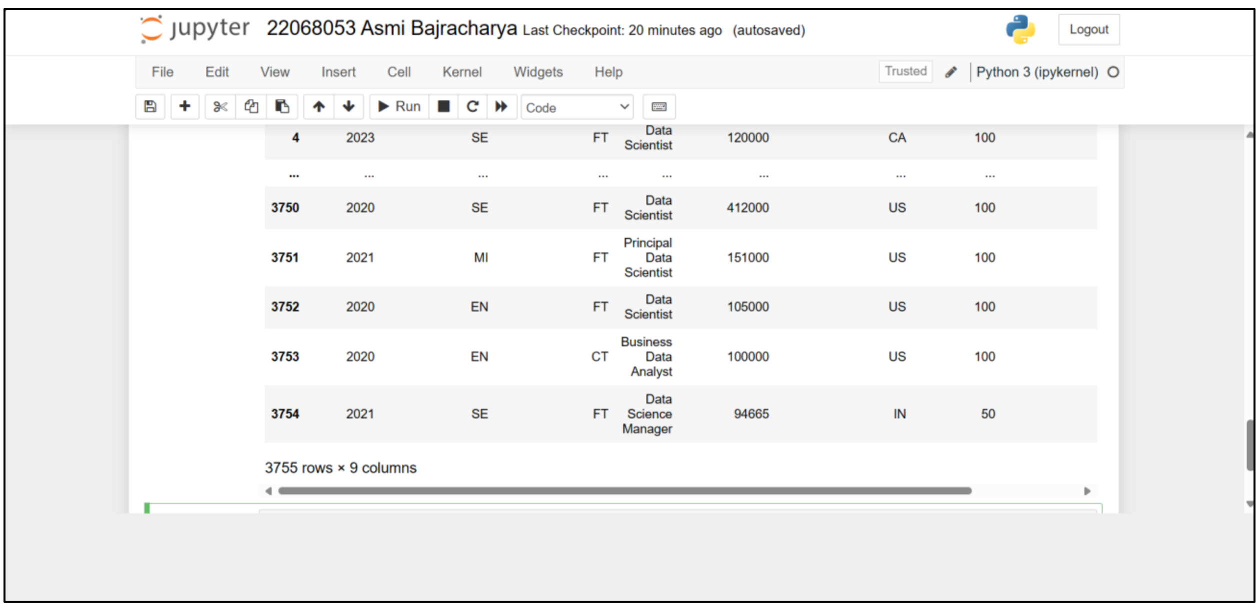



```
#removing unnecessary columns
df.drop(['salary', 'salary_currency'], axis=1, inplace=True)
```

```
#Dataframe after removing unnesessary columns
df
```

	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	
				Principal				

Figure 6 Dataframe after removing unnecessary columns.



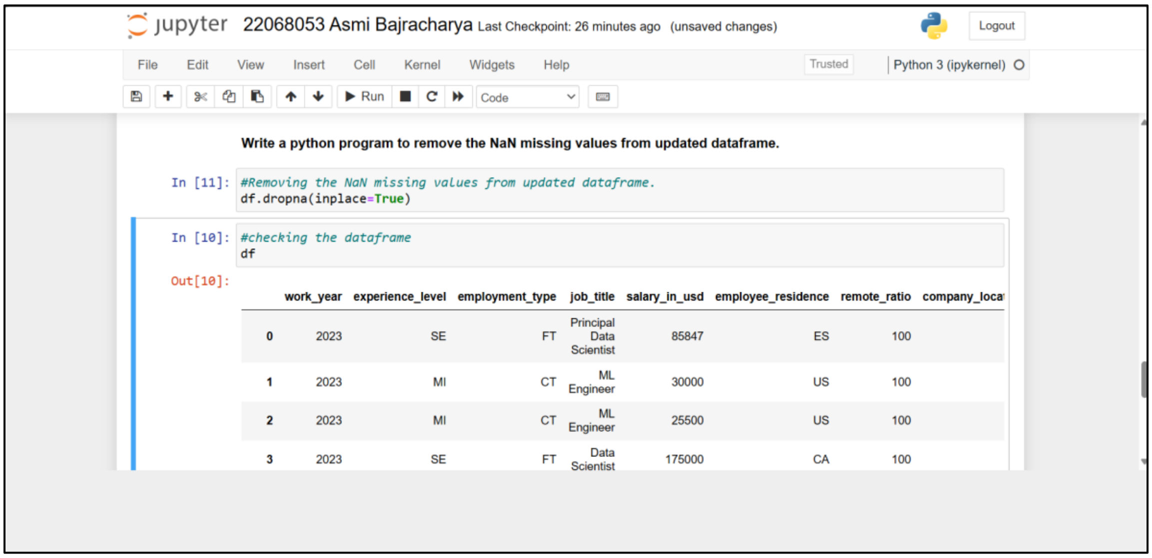
4	2023	SE	FT	Data Scientist	120000	CA	100
...	...	...	...	...	...	...	...
3750	2020	SE	FT	Data Scientist	412000	US	100
3751	2021	MI	FT	Principal Data Scientist	151000	US	100
3752	2020	EN	FT	Data Scientist	105000	US	100
3753	2020	EN	CT	Business Data Analyst	100000	US	100
3754	2021	SE	FT	Data Science Manager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

3755 rows × 9 columns

Figure 7 Dataframe after removing unnecessary columns continued.

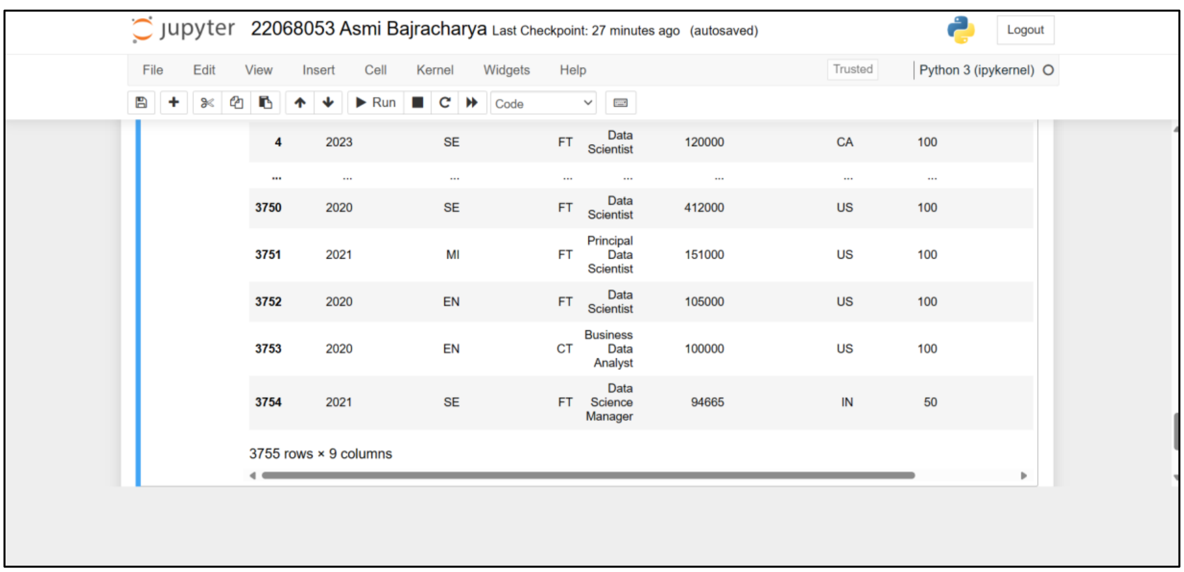
The unnecessary columns are removed by using the drop function which has removed the columns salary and salary currency from the data set. Here, in the code axis=1, refers to the column and if the column is salary or salary currency then the flag becomes true, and the column is dropped. Finally, the unnecessary columns are dropped and storage is saved.

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



	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	

Figure 8 Removing the NaN missing values from updated dataframe and checking the dataframe.



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 9 Dataframe continued.

Here, to remove the NaN missing values, we have used the dropna() function. However, this data frame does not seem to have any missing values. Hence, there are no NaN missing values in this data frame.

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

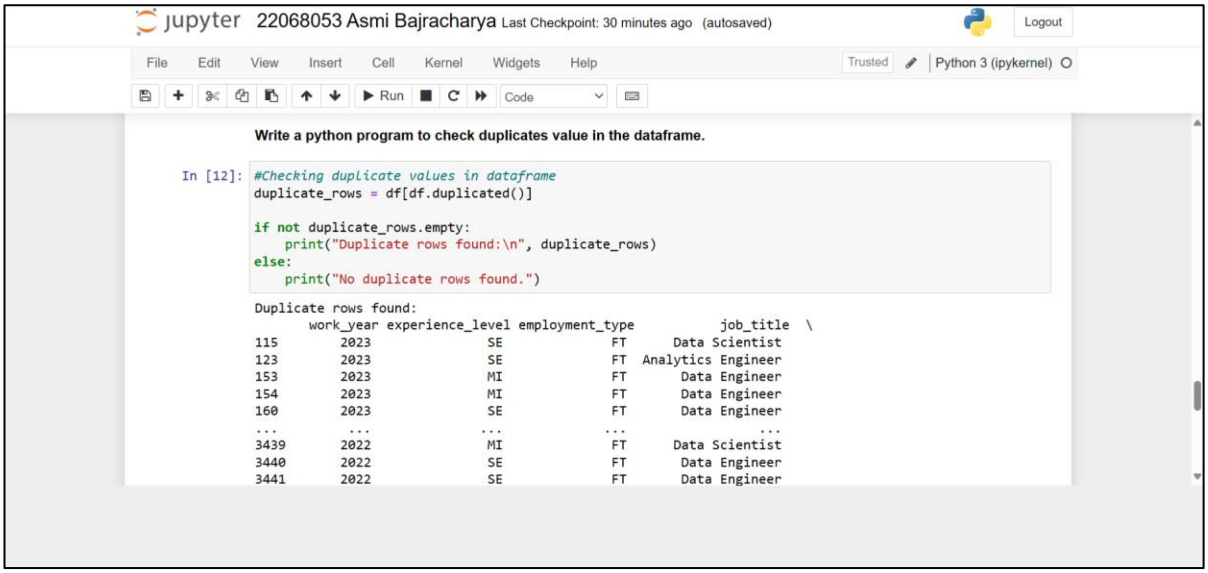


Figure 10 Checking duplicates in the dataframe.

Here, we are checking the duplicate values in the data frame. We have used the function duplicated to check the duplicate values. Since, there are a lot of duplicate rows found, the output displays all the duplicate rows found in the data frame.

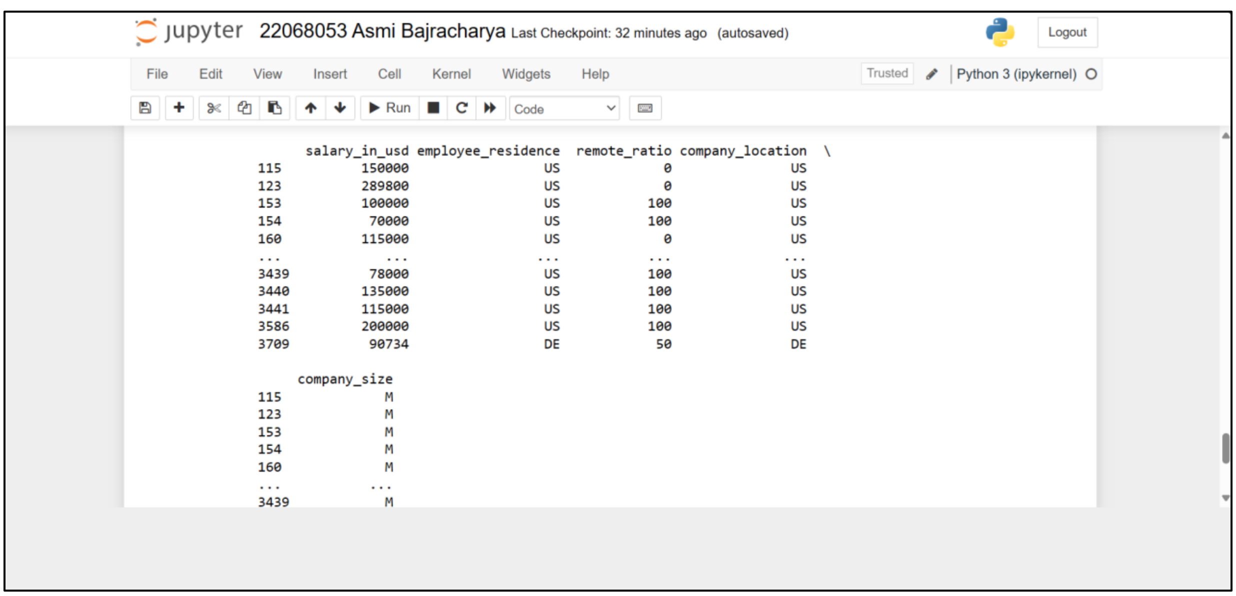


Figure 11 Duplicates in the dataframe continued.

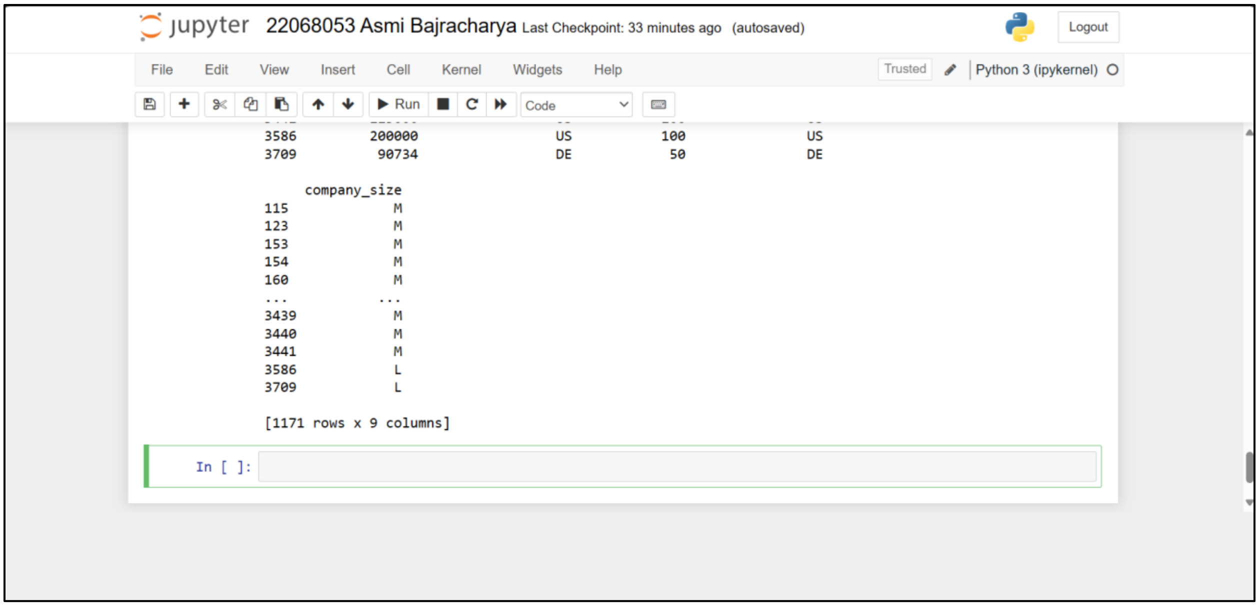


Figure 12 Duplicates in the dataframe continued.

There are a total of 1171 duplicate rows found across 9 columns in this dataframe.

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

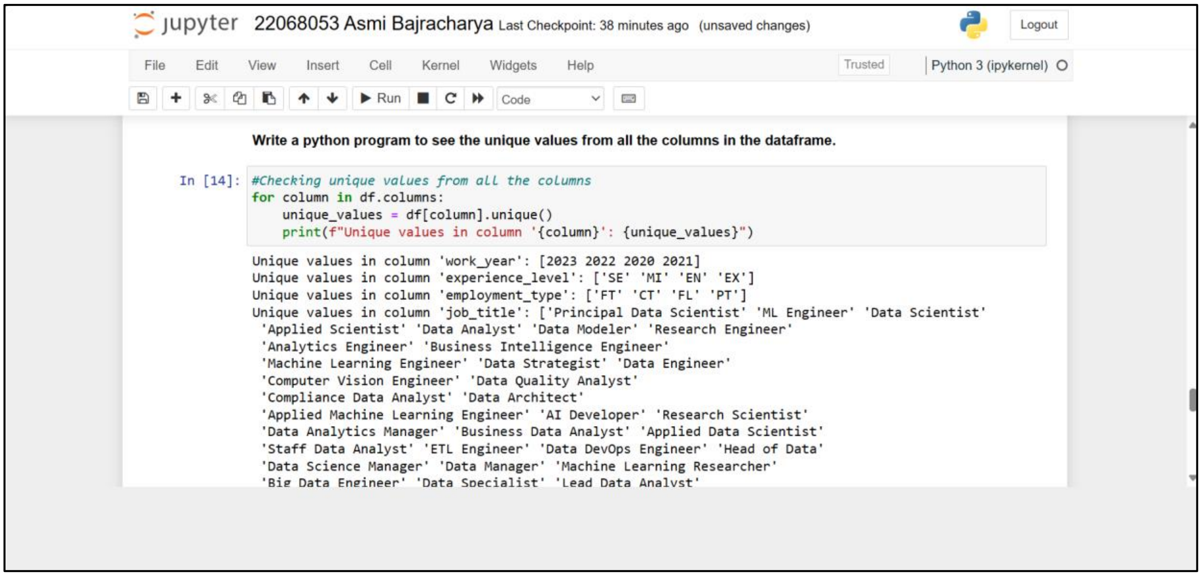


Figure 13 checking unique values from all the columns in the dataframe.

Here we use the function, unique(), to check out all the unique values from all the columns in the data frame. We found out a lot of unique values from each column however, we can see there are a lot of data inconsistency in this dataframe.

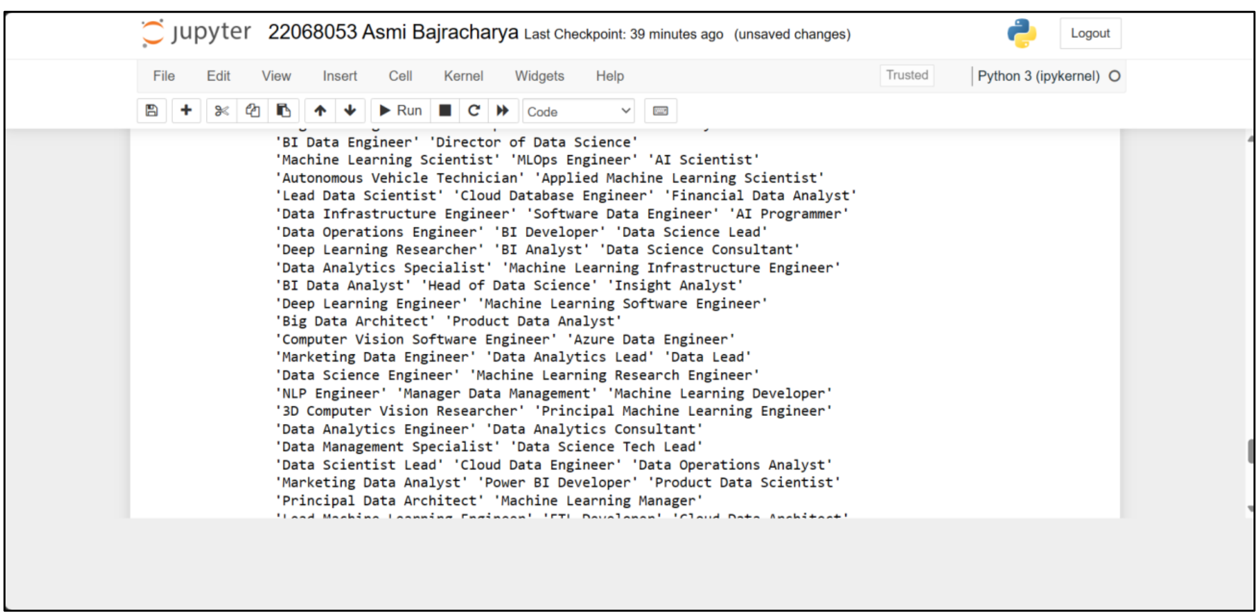


Figure 14 Unique values from all the columns in the dataframe continued.

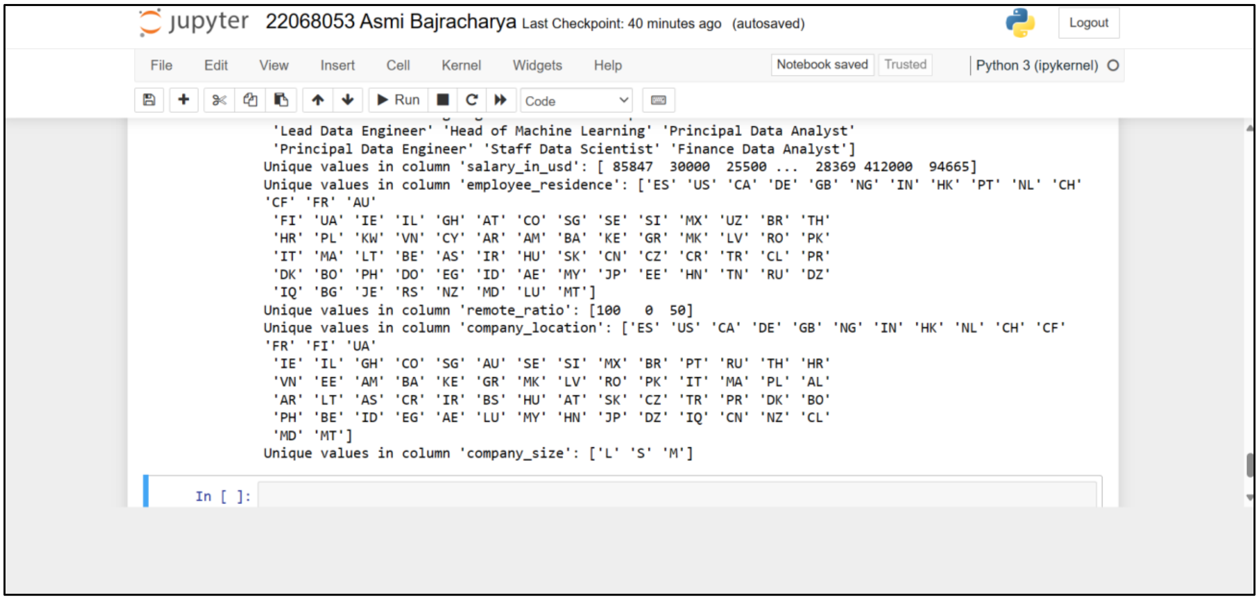


Figure 15 Unique values from all the columns in the dataframe continued.

There are different number of unique values in each column. As we can see in the figure above there are four unique values in the column work year, experience level and employment type. There are 93 unique values in the column job title, however, jobs are repeated and there is data inconsistency because of the spellings of the data. We need to fix it. There are 3 unique values remote ratio and 82 unique company location as well as employee residence, and finally there are three unique values in the company size.

2.6. Rename the experience level columns as below.

SE – Senior Level/Expert

MI – Medium Level/Intermediate

EN – Entry Level

EX – Executive Level

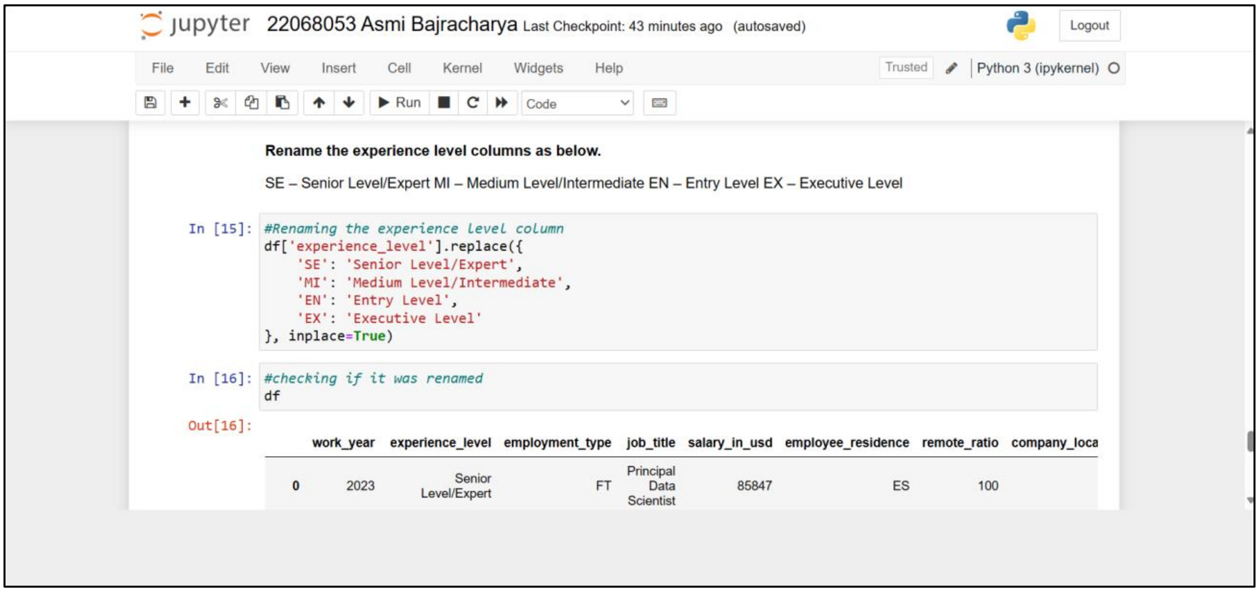
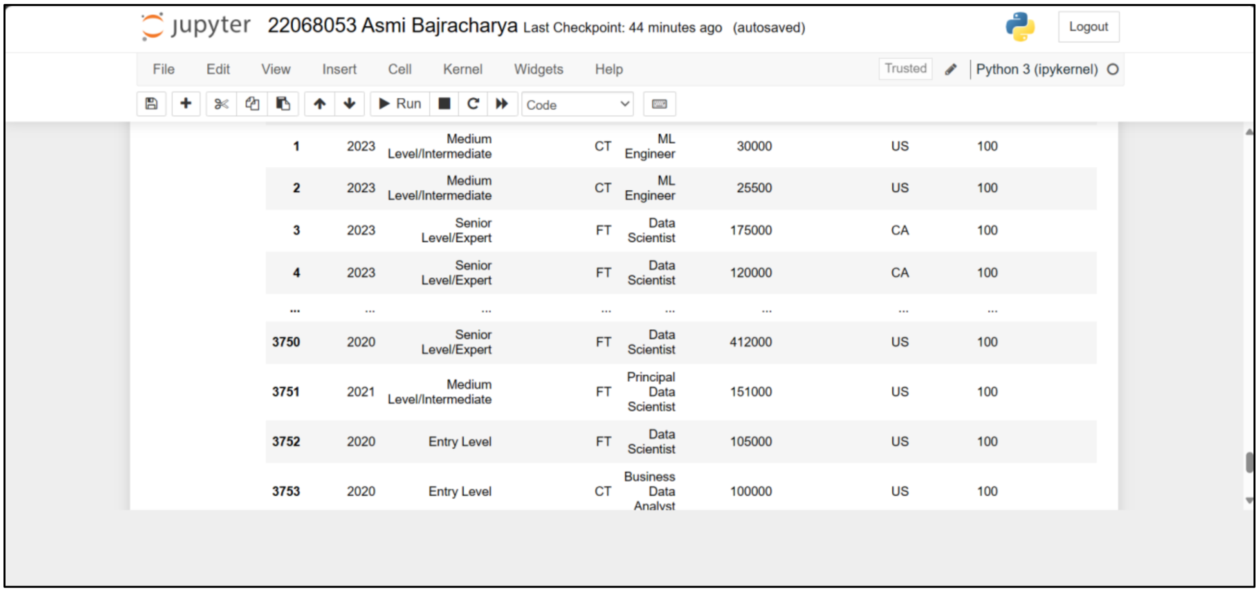


Figure 16 Renaming the experience level column.

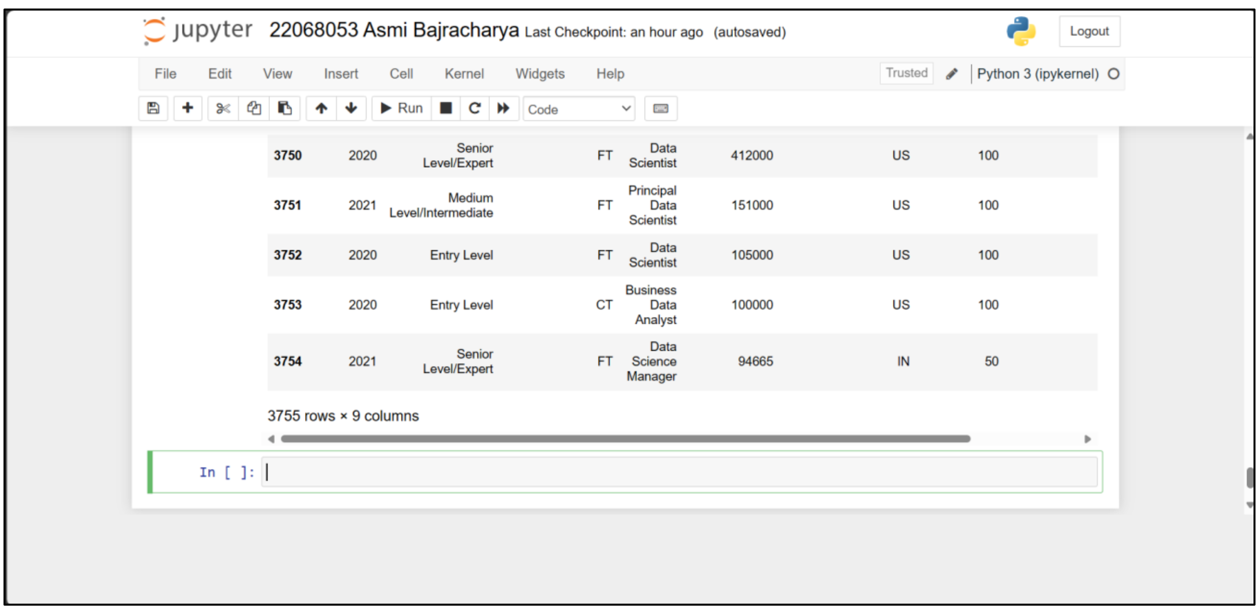
Here, we have used the replace function to rename the values of the column. We have replaced ‘SE’ as senior level/expert, ‘MI’ as medium level/intermediate, ‘EN’ as entry level and ‘EX’ as executive level. Now to check if the values were renamed, we checked the updated dataframe.



jupyter 22068053 Asmi Bajracharya Last Checkpoint: 44 minutes ago (autosaved)

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

Figure 17 Checking if the values were renamed.



jupyter 22068053 Asmi Bajracharya Last Checkpoint: an hour ago (autosaved)

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 18 Checking if the values were renamed.

As we can see each and every value of the column is renamed correctly. All the four values have been renamed and it is easier to understand what the experience level is after the values were renamed.

3. Data Analysis

Before starting with the data analysis part, we must remove all the duplicate values and remove the data inconsistency so that we get the correct values for analysis.

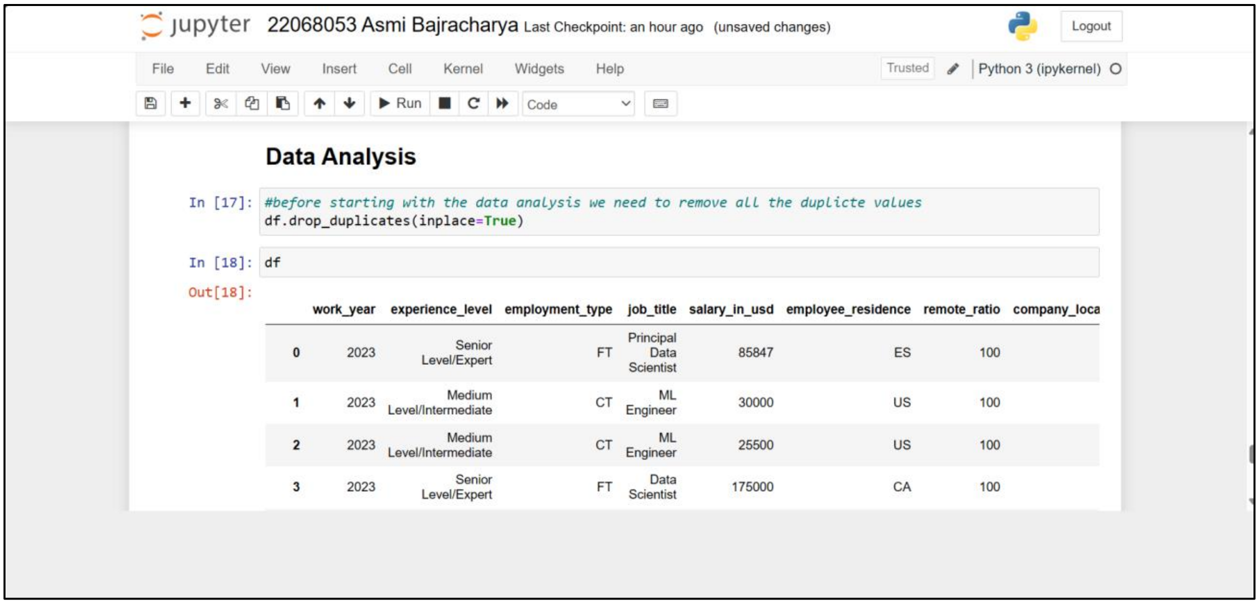


Figure 19 Removing the duplicate values.

Starting with dropping all the duplicates by using the drop function again, all the duplicates are removed. Checking the data frame to see if there are any duplicate values remaining.

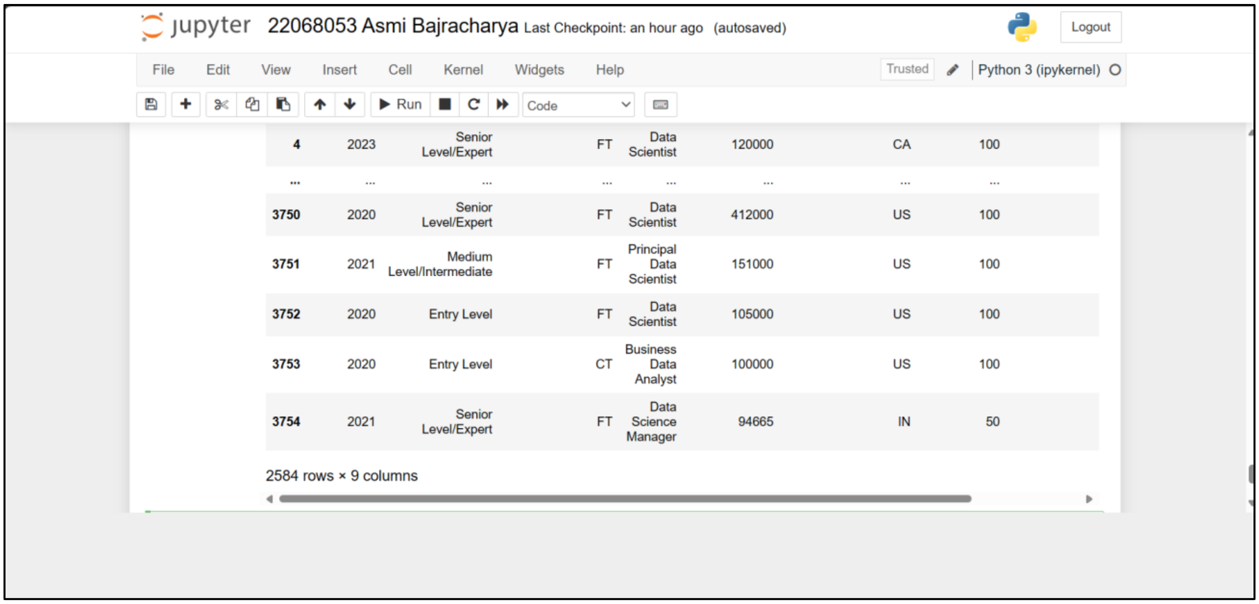


Figure 20 Checking the dataframe.

As we can see in the data frame level there are no duplicate data remaining.

Now, after removing all the duplicate values we must remove all the data inconsistencies also. As mentioned before, there is data inconsistency in the column, job title.

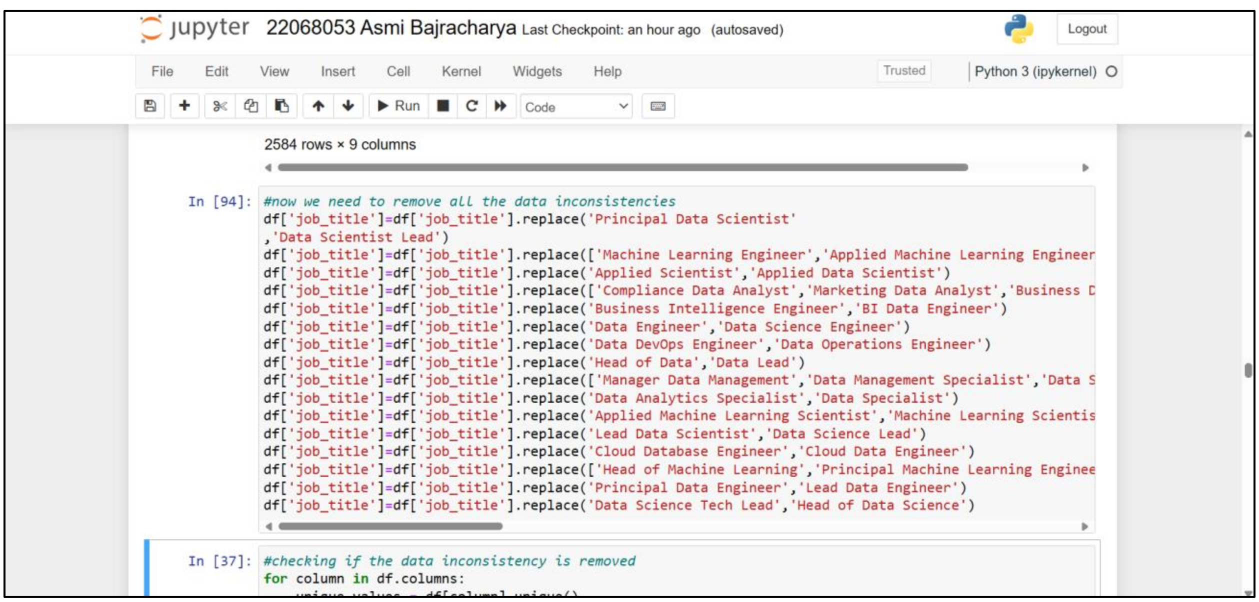
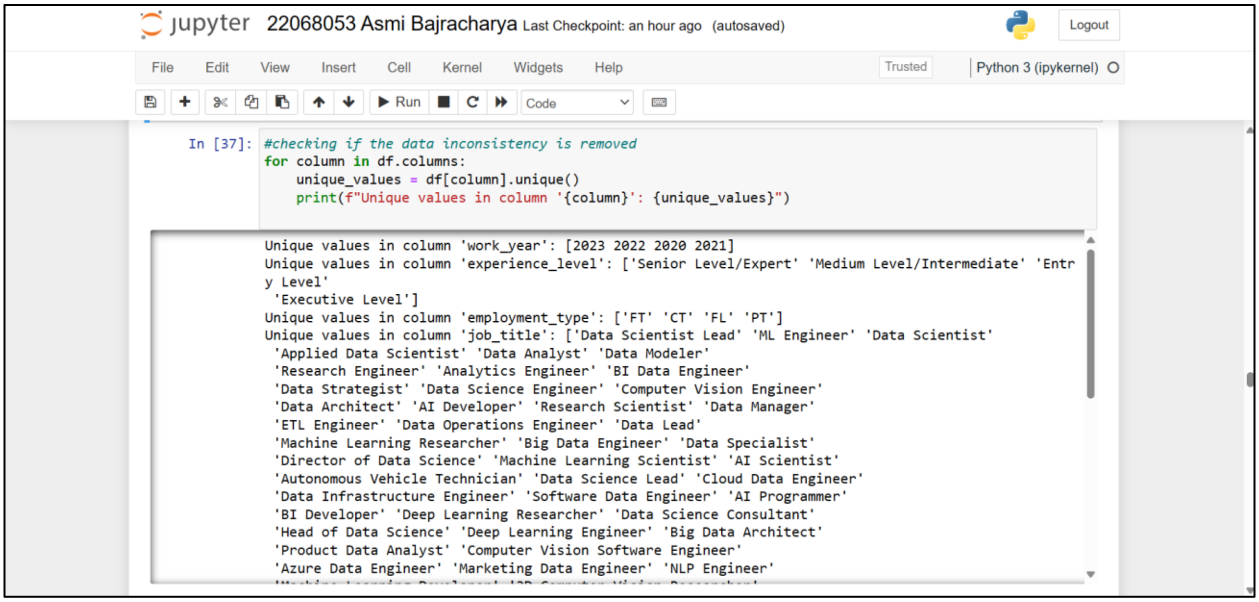


Figure 21 Removing data inconsistencies.

We have replaced similar jobs with different names into a same name. For example, principal data scientist and data scientist are both data scientists of senior level. Therefore, they can be categorized into the same job title. Similarly, all the machine learning engineers can be categorized into ML engineers. Applied scientist and applied data scientist are also similar so we have decided to categorize them as one. Business intelligence engineer and BI data engineer are both the same hence they are also categorized as one.

Similarly looking at the figure above we can see that data develop engineer and data operations engineer are quite similar, so they've been categorized as one. Head of data and data lead are also similar, so they are also categorized as one. All the data managers and data management specialists manage data hence, they are categorized as one as well. Machine learning scientists and applied machine learning scientists can also be categorized as one.

Lead data scientist and data science league sounds similar hence they're also categorized as one. All the different types of data analyst have been categorized as one for convenience. Cloud database engineer and cloud data engineer are also categorized as one. Principal data engineer and head of engineer is also categorized as one. Data science tech league and head of data science is also categorized as one.




```
#checking if the data inconsistency is removed
for column in df.columns:
    unique_values = df[column].unique()
    print(f"Unique values in column '{column}': {unique_values}")
```

```
Unique values in column 'work_year': [2023 2022 2020 2021]
Unique values in column 'experience_level': ['Senior Level/Expert' 'Medium Level/Intermediate' 'Entr
y Level'
 'Executive Level']
Unique values in column 'employment_type': ['FT' 'CT' 'FL' 'PT']
Unique values in column 'job_title': ['Data Scientist Lead' 'ML Engineer' 'Data Scientist'
 'Applied Data Scientist' 'Data Analyst' 'Data Modeler'
 'Research Engineer' 'Analytics Engineer' 'BI Data Engineer'
 'Data Strategist' 'Data Science Engineer' 'Computer Vision Engineer'
 'Data Architect' 'AI Developer' 'Research Scientist' 'Data Manager'
 'ETL Engineer' 'Data Operations Engineer' 'Data Lead'
 'Machine Learning Researcher' 'Big Data Engineer' 'Data Specialist'
 'Director of Data Science' 'Machine Learning Scientist' 'AI Scientist'
 'Autonomous Vehicle Technician' 'Data Science Lead' 'Cloud Data Engineer'
 'Data Infrastructure Engineer' 'Software Data Engineer' 'AI Programmer'
 'BI Developer' 'Deep Learning Researcher' 'Data Science Consultant'
 'Head of Data Science' 'Deep Learning Engineer' 'Big Data Architect'
 'Product Data Analyst' 'Computer Vision Software Engineer'
 'Azure Data Engineer' 'Marketing Data Engineer' 'NLP Engineer'
```

Figure 22 Checking if all the data inconsistencies are removed.

After all the data inconsistencies were removed there are a total of 59 job titles that remain.

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

Here, the chosen variable for this question is salary_in_usd.

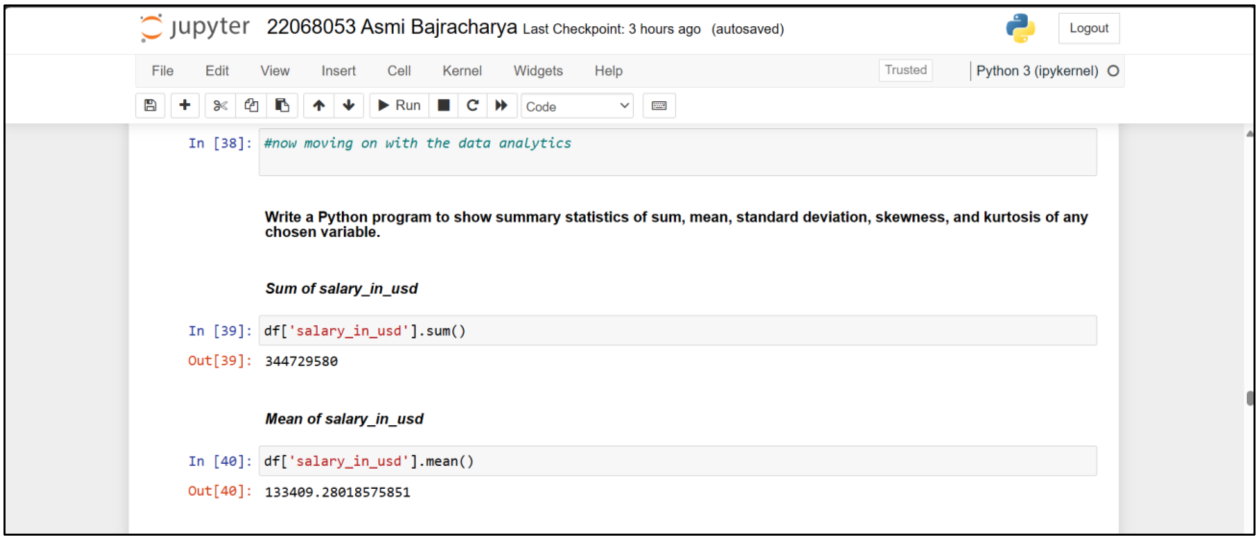


Figure 23 Python program to show the sum of salary in USD and mean of salary in USD.

Sum

We have used the sum() function to find out the sum of salary in USD of the data frame and the sum is 344729580.

Mean

Here, mean is the average salary and we have used the mean() function to find out the mean of salary in USD and the mean is 133409.28018575851.

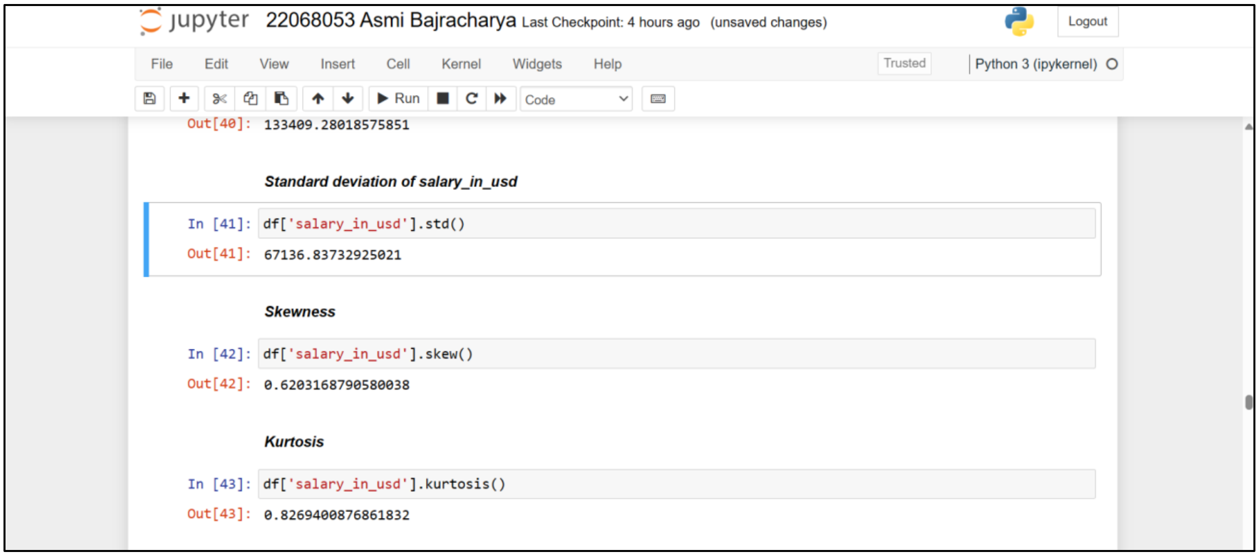


Figure 24 Python program to show the standard deviation, skewness, and kurtosis of salary in USD.

Standard deviation

Here we have used the std() function, to find out the standard deviation and the standard deviation of salary in USD is 67136.83732925021.

Skewness

The measurement of asymmetry of a distribution is known as skewness (Turney, 2022). Here we have used skew() function to find out the skewness and the skewness is 0.6203168790580038.

Kurtosis

The measurement of tailedness offer distribution is known as Kurtosis (Turney, 2022). Here we have used the Kurtosis() function to find out the Kurtosis and the Kurtosis is 0.8269400876861832.

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

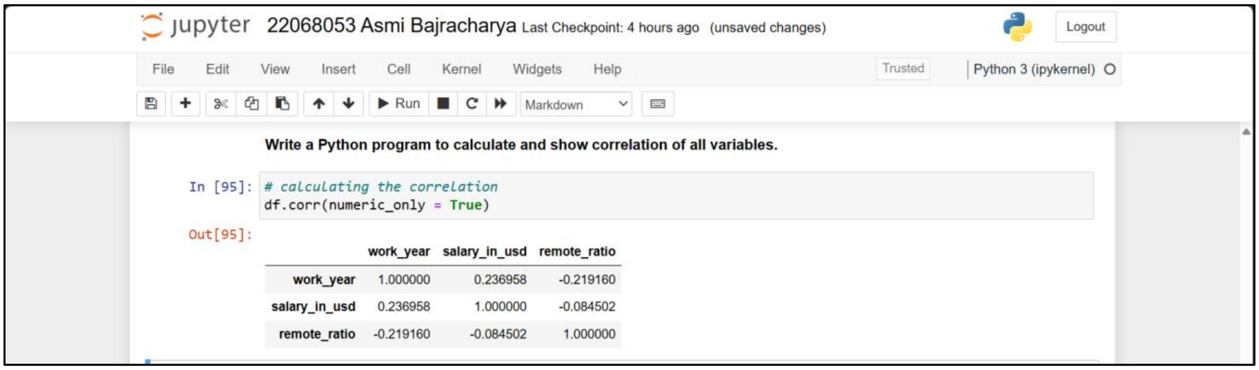


Figure 25 Python program to calculate and show correlation of all variables.

Correlation measures how to variables change together. If one variable goes up while the other also moves up, they have a positive correlation whereas if one variable goes down and other goes up then they have negative correlation.

Here, work_year and salary_in_usd have positive correlation of around 0.236958 whereas, work_year and remote_ratio have negative correlation of around -0.219160. Again, salary_in_usd and remote_ratio have negative correlation of around -0.084502 and 1 here means perfect correlation with itself.

4. Data Exploration

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

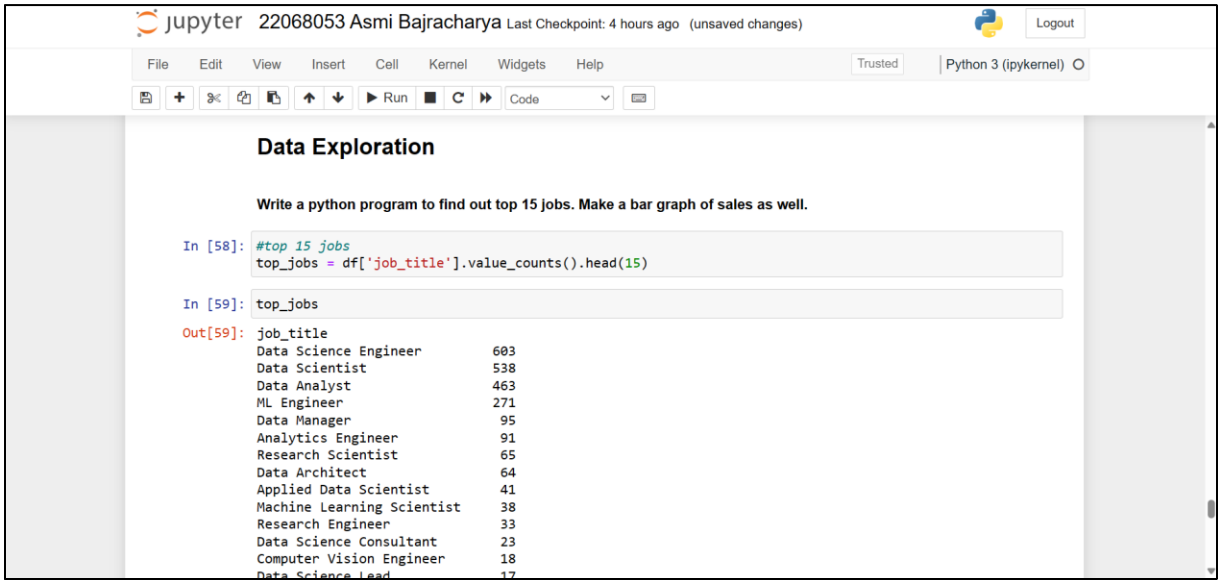


Figure 26 Python program to find out top 15 jobs.

Here, the column job_title is selected and value_counts() function counts how many times each unique job title appears in that column and finally head(15) function shows the top 15 most reoccurring jobs. The top 15 jobs in this dataframe are:

1. Data Science Engineer
2. Data Scientist
3. Data Analyst
4. ML Engineer
5. Data Manager
6. Analytics Engineer
7. Research Scientist
8. Data Architect
9. Applied Data Scientist
10. Machine Learning Scientist
11. Research Engineer

12. Data Science Consultant
13. Computer Vision Engineer
14. Data Science Lead
15. AI Scientist

Now, to meet the bar graph we need to import matplotlib.pyplot.

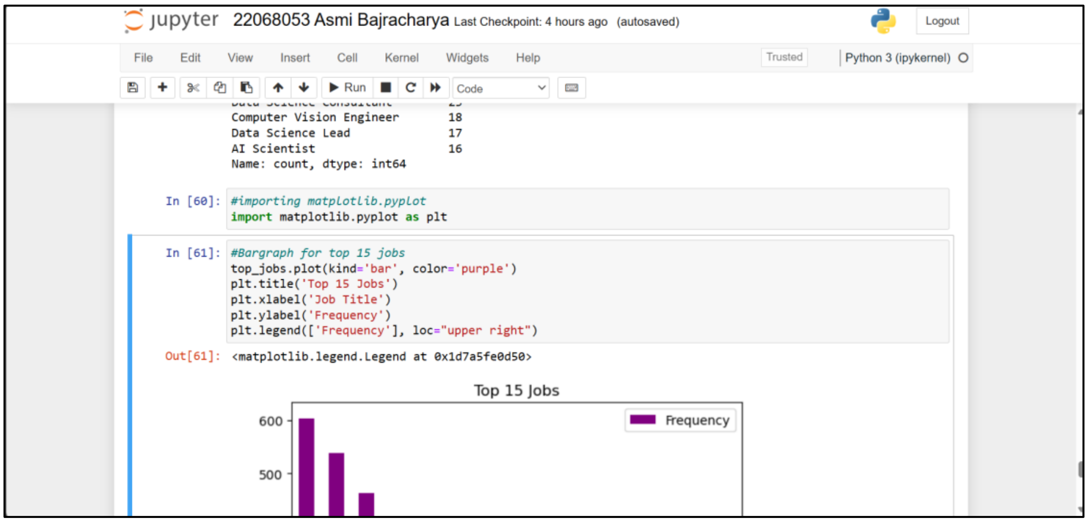


Figure 27 importing matplotlib.pyplot.

After importing matplotlib.pyplot, we are going to plot the bar graph. The topic of the paragraph is 'Top 15 jobs'. On the X axis we have the job title and on the Y axis we have the frequency of the repeated jobs.

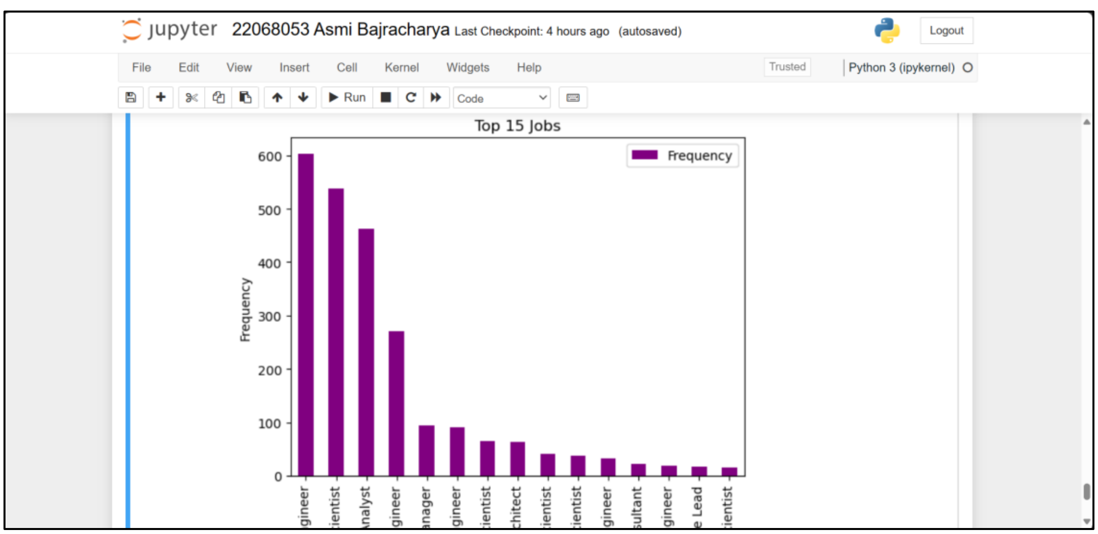


Figure 28 Bar graph of top 15 jobs.

Here, we can see that the frequency ranges from 0 to 600 and the job titles is shown in the figure below.

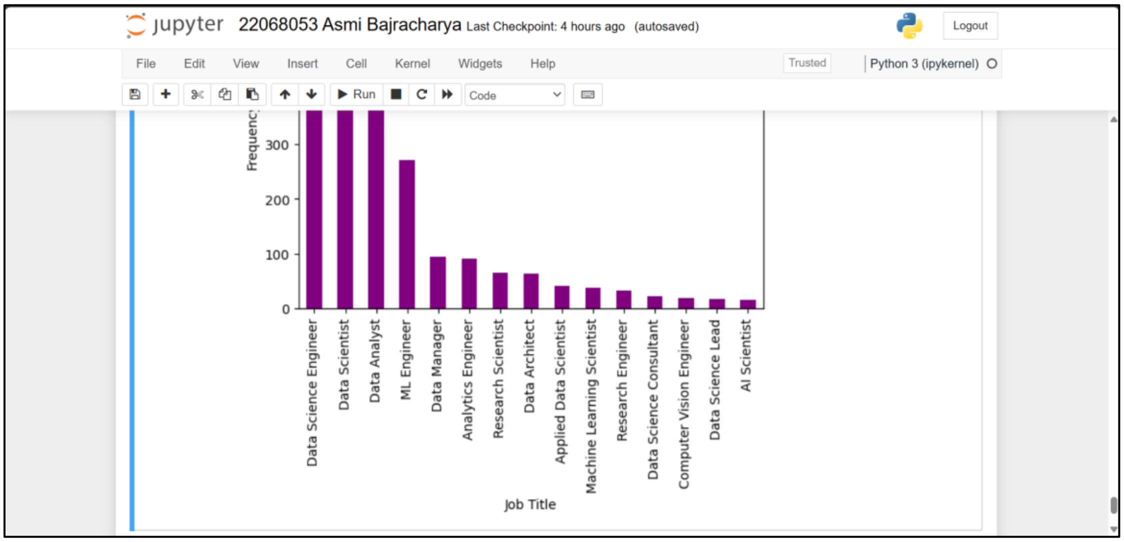
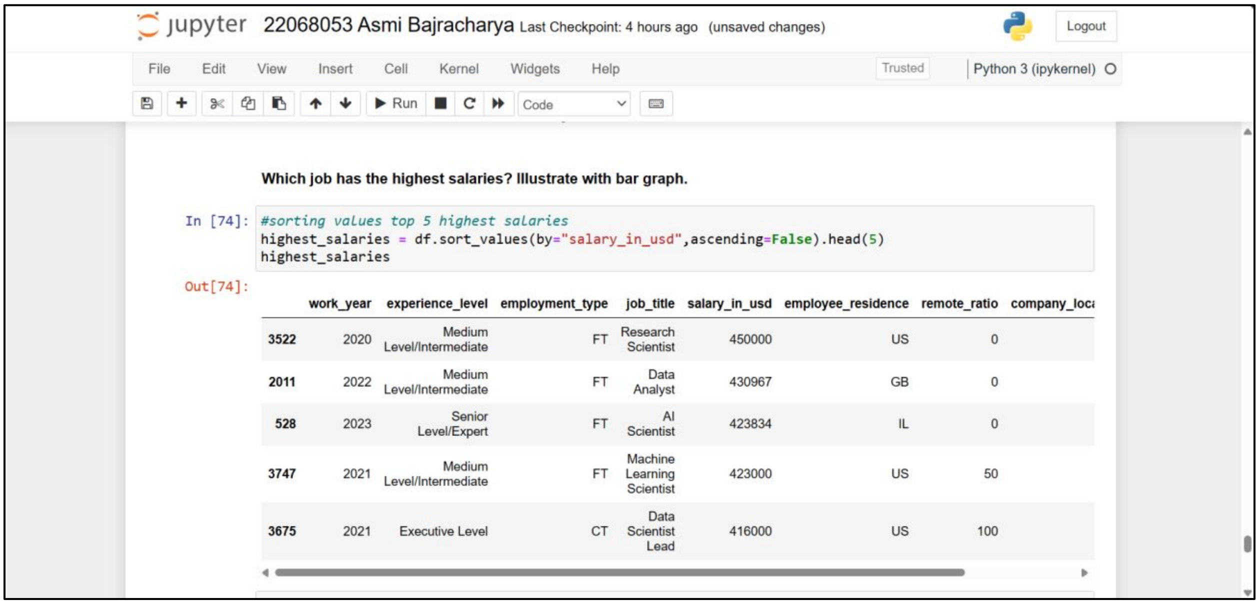


Figure 29 Bar graph of top 15 jobs x axis.

According to the bar graph data science engineer has the highest frequency of around 600 then the second is data scientist with the frequency of between 500 to 600 and similarly it is followed by data analyst. There is a significant gap between data analyst and ML engineer. And there is even more significant gap between ML engineer and data manager. Data manager is followed by research scientist, data architect, applied data scientist, machine learning scientist, research engineer, data science consultant, computer vision engineer, data science lead and finally AI scientist.

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



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

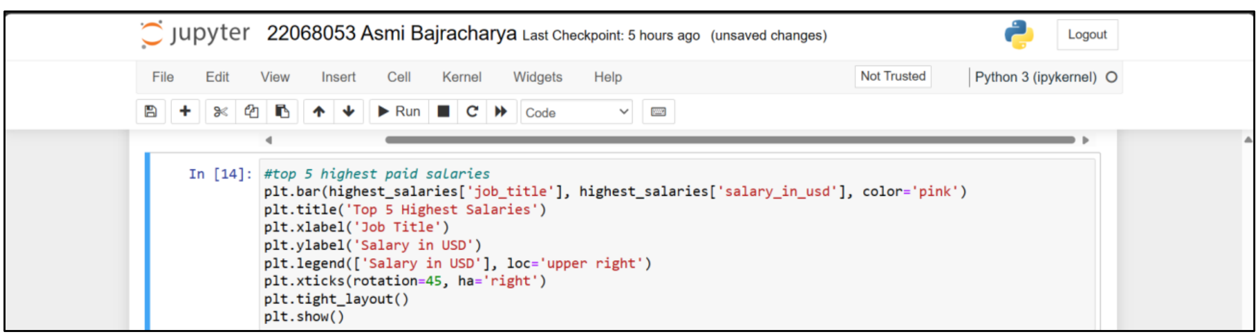
In [74]:
```
#sorting values top 5 highest salaries
highest_salaries = df.sort_values(by="salary_in_usd",ascending=False).head(5)
highest_salaries
```

Out[74]:

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_loca
3522	2020	Medium Level/Intermediate	FT	Research Scientist	450000	US	0	
2011	2022	Medium Level/Intermediate	FT	Data Analyst	430967	GB	0	
528	2023	Senior Level/Expert	FT	AI Scientist	423834	IL	0	
3747	2021	Medium Level/Intermediate	FT	Machine Learning Scientist	423000	US	50	
3675	2021	Executive Level	CT	Data Scientist Lead	416000	US	100	

Figure 30 Python program to find out the highest salaries.

Here we're finding out the top five highest salaries in the data frame. We have sorted the values in descending order and use the head(5) function to find out the top five highest salaries. The highest salary is of research scientists with 450000, then it is Data Analyst with 430967, after that it is AI Scientist with 423834, then Applied Machine Learning Scientist with 423000 and finally Principal Data Scientist with 416000.



In [14]:
```
#top 5 highest paid salaries
plt.bar(highest_salaries['job_title'], highest_salaries['salary_in_usd'], color='pink')
plt.title('Top 5 Highest Salaries')
plt.xlabel('Job Title')
plt.ylabel('Salary in USD')
plt.legend(['Salary in USD'], loc='upper right')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

Figure 31 Python program to make the bar graph for top five highest paid salaries.

Here the title of this paragraph is ‘Top 5 highest salaries’ And on the X access there are job title and on the Y axis there is salary in USD.

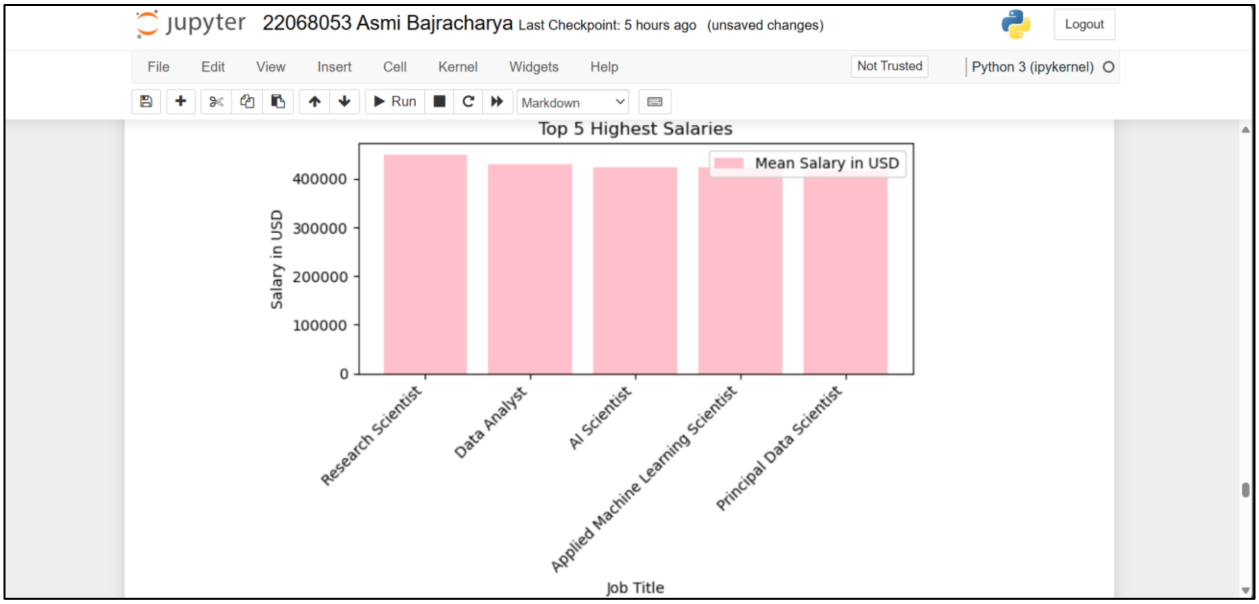


Figure 32 Bar graph of top 5 highest paid salaries.

In the above figure we can see the bar graph where research scientist has the most salary followed by data analyst call my AI scientist, applied machine learning scientist and principal data scientist. All these jobs have salaries over 400,000 USD.

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

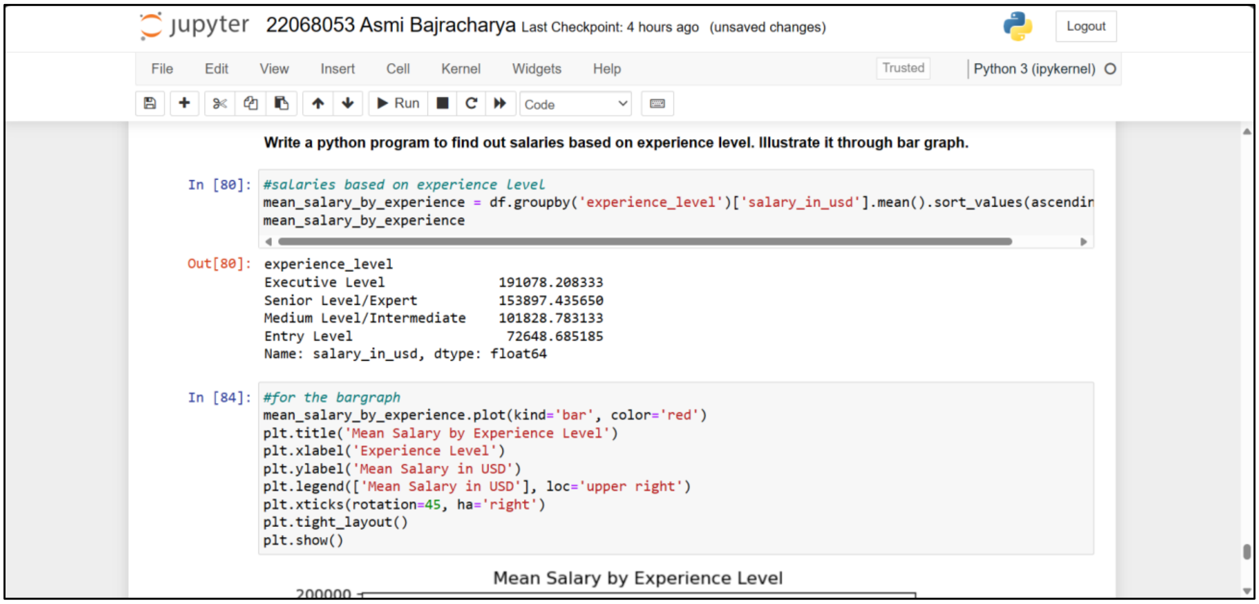


Figure 33 python program to find out salaries based on experience level.

Here, to find out the salaries based on experience level we have grouped the salary in USD by experience level and after finding out the mean, we sorted it in descending order.

We found out that the executive level earns the most an entry level earns the least. Executive level earns an average of 191078.208333, then senior level/expert earns an average of 153897.435650, then medium level/ intermediate earns an average of 101828.783133, and lastly entry level earns an average of 72648.685185.

Now to plot it we have put the experience level on X axis and mean salary in USD in Y axis.

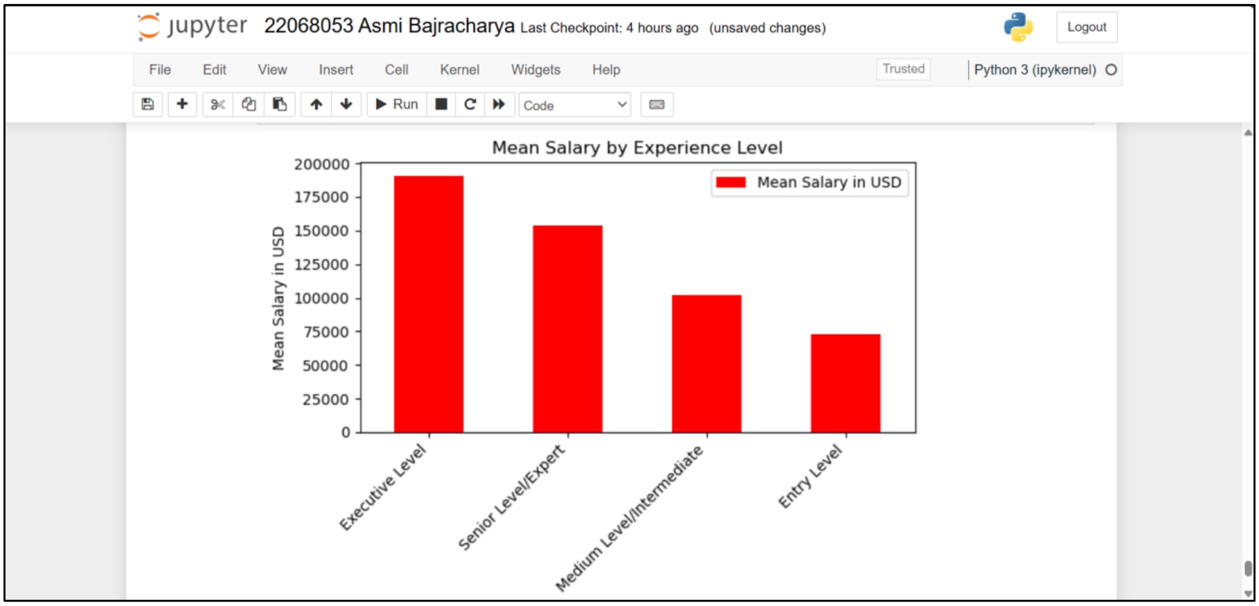


Figure 34 Bar graph of mean salary by experience level.

As we can see in the above figure, executive level earns the highest which is between 200,000 USD and 175,000 USD. Which is followed by senior level/expert then medium level/intermediate and lastly entry level.

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

The variable chosen for making this histogram is salary_in_usd.

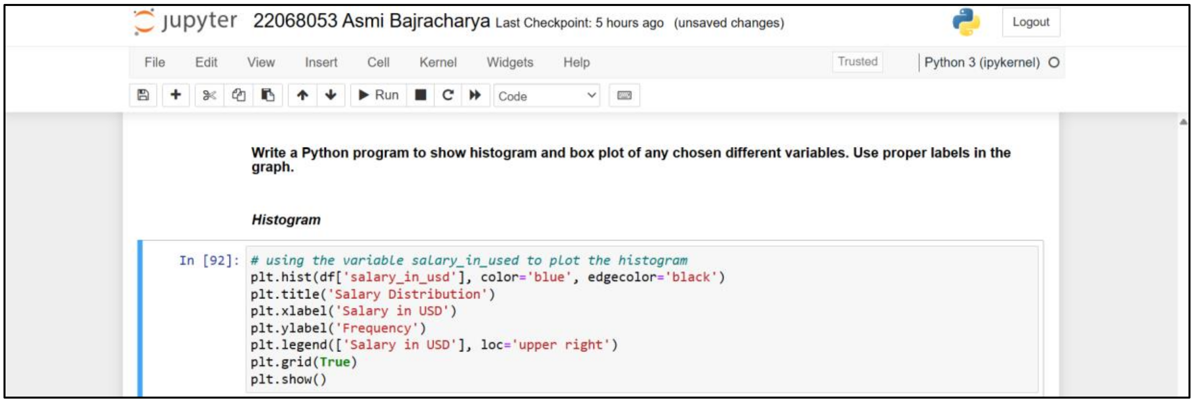


Figure 35 Python program to show histogram of salary_in_usd.

Here, the title of the histogram is salary distribution and on the X axis we have salary in USD and on the Y axis we have frequency.

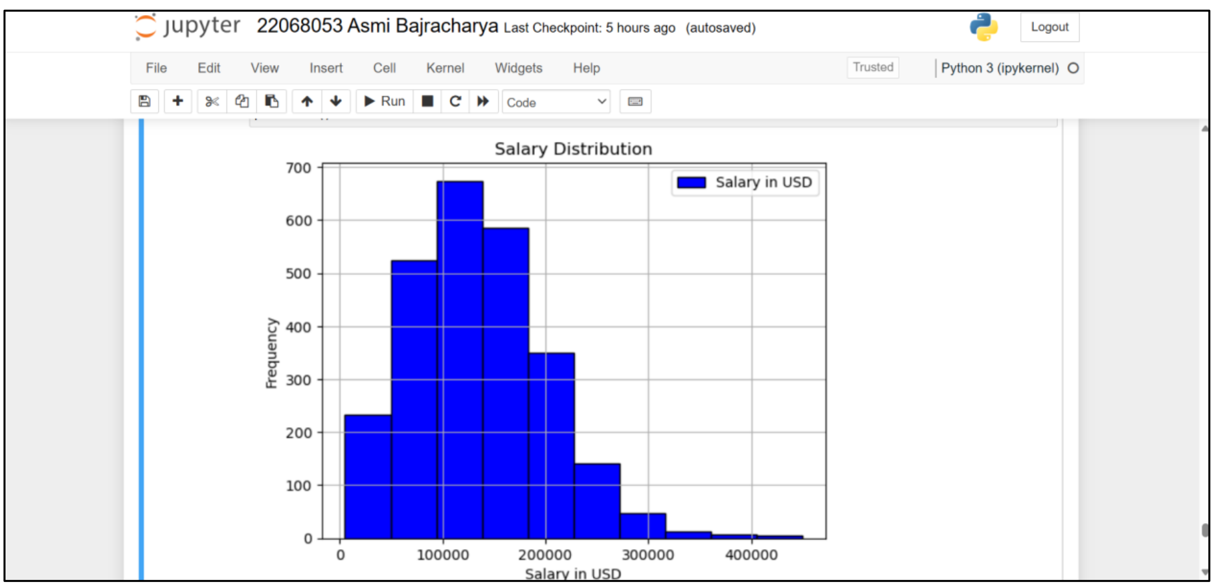


Figure 36 Histogram of salary_in_usd

Here, in this data set most people have the salary of around 100,000 USD on average. The salary of an average person ranges from 100,000 USD to 200,000 USD according to the histogram.

Salary in USD is chosen for the box plot also.

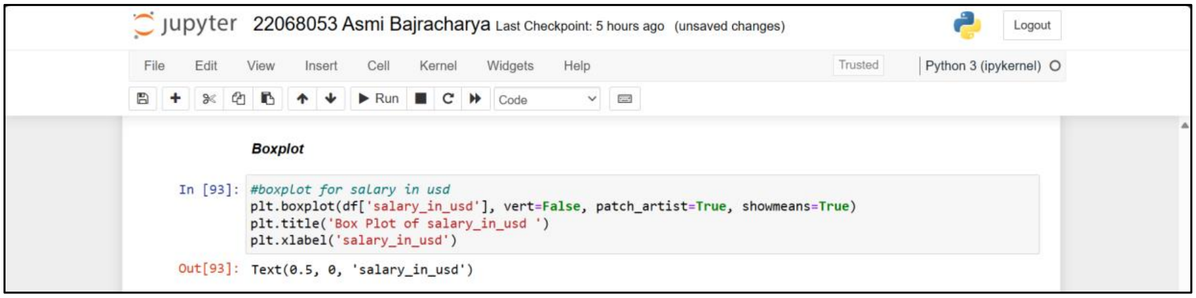


Figure 37 Box plot for salary_in_usd.

Here, the title of the box plot is 'Box Plot of salary_in_usd' and the X axis consist of the salary_in_usd.

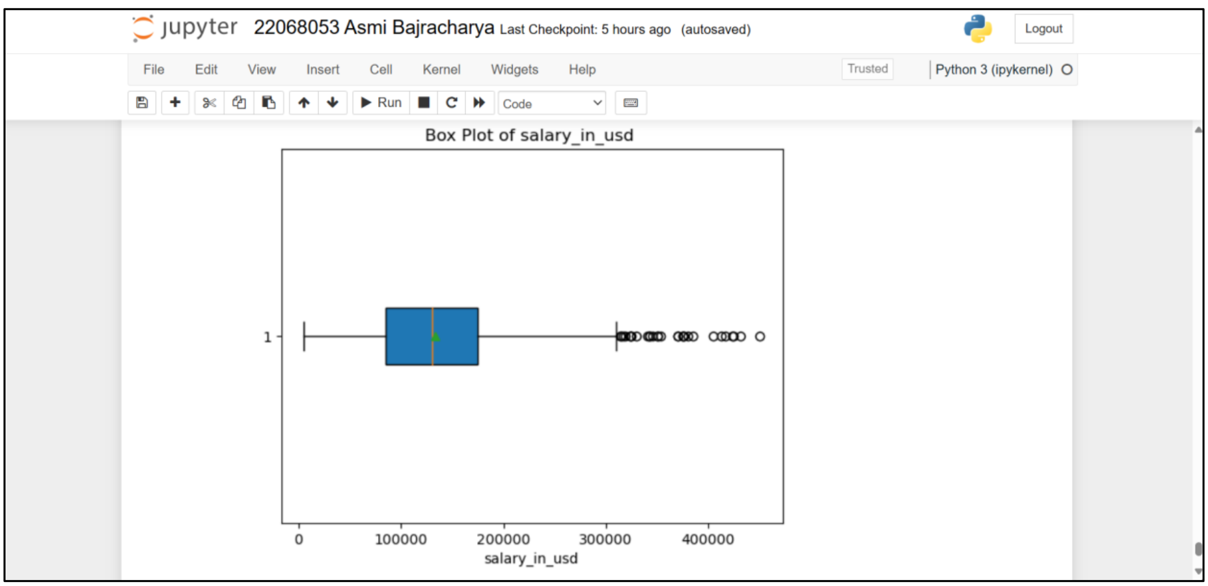


Figure 38 Box Plot of salary_in_usd

Here we can see the upper quartile, lower quartile and the mean as well as the outliers of salary_in_usd. The lower quartile ranges around 100,000 USD and the upper quartile ranges between 100,000 to 200,000 USD and the outliers lies over 300,000 USD.

Conclusion

In this course work, we got to explore data science salaries using Python and DataScienceSalaries dataset. Firstly, we read the data carefully to understand the data resources. By using python as a programing language and pandas library, we started the preparation of the data by loading it into data frames then removing unnecessary columns and missing values as well as checking for unique and duplicate values

The second part was analyzing the data where we found out the sum, mean, standard deviation, skewness, and kurtosis of the data. Then we also found out the correlation of all the numeric values of the data. We properly analyzed the data in this part.

The third part of this coursework was data exploration. We found out the top 15 jobs as well as the highest paid salaries and illustrated them in a bar graph. We also found the salaries based on experience level and showed it in a bar graph. And lastly, we also made a histogram and box plot for the salaries in the dataset.

In conclusion, it is because of this coursework we got to learn to handle data and learnt valuable information on data understanding, preparation, analysis, and exploration. This contents of this coursework is not just limited to college but is also going to be very helpful in the future and I am very grateful that I got the opportunity to learn about this course.

References

Sheldon, R., 2024. *What is a data set?.* [Online] Available at: https://www.techtarget.com/whatis/definition/data-set [Accessed 12 May 2024].

Turney, S., 2022. *Skewness* | *Definition, Examples & Formula.* [Online] Available at: https://www.scribbr.com/statistics/skewness/ [Accessed 12 May 2024].

Turney, S., 2022. *What Is Kurtosis?* | *Definition, Examples & Formula.* [Online] Available at: https://www.scribbr.com/statistics/kurtosis/#:~:text=Kurtosis%20is%20a%20measure%20of,(thin%20tails)%20are%20platykurtic.
[Accessed 12 May 2024].