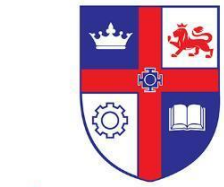
LONDON METROPOLITAN UNIVERSITY



islington college
(इस्लिङ्टन कलेज)

CC5067NI-Smart Data Discovery

60% Individual Coursework

2023-24 Spring

Student Name: Sumit Shrestha

London Met ID: 22085637

College ID: np01cp4s230046

Assignment Due Date: Monday, May 13, 2024

Assignment Submission Date: Sunday, May 12, 2024

Word Count: 2232

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.

Table of Content

Table of Figures

Table Of Content

1. Introduction

Data science is the methods and techniques for extracting useful information and knowledge from data for personal growth or business growth. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

Data science is important because it combines tools, methods, and technology to generate meaning from data. Modern organizations are inundated with data; there is a proliferation of devices that can automatically collect and store information. Online systems and payment portals capture more data in the fields of e-commerce, medicine, finance, and every other aspect of human life. We have text, audio, video, and image data available in vast quantities.

The given CSV file is the data of the salaries of data scientist whole world. This assignment is assigned to year 2, second semester. Data science has become increasingly important in today's data-driven world, which is defined by an abundance of digital data coming from various sources, including social media, online transactions, sensor networks, and Internet of Things devices. Data science is used by modern companies in a wide range of industries, including e-commerce, healthcare, finance, and more, to get actionable insights, streamline operations, and gain a competitive edge in a market that is becoming more and more crowded.

This assignment is evidence of the increasing importance of data science in both academia and industry. It provides a hands-on opportunity for second-year students in their second semester to apply fundamental data science principles to real-world datasets. The complexities of data preparation, analysis, and visualization are navigated by students, who obtain vital expertise in utilizing.

2. Data Understanding

Data understanding is the fundamental step in our analysis, providing insights into the structure, quality, and characteristics of the dataset. This phase encompasses data collection, exploration, cleaning, and preprocessing setting the stage for subsequent analysis and modeling. The most important activities of data understanding are to identify potential data sources (transactional databases, spreadsheets, CSV, text files, web logs, web services, etc.). Data collection efforts focused on gathering information from multiple sources relevant to the project objectives. The first step in data understanding involves identifying and assessing potential data sources. Data collection efforts are focused on gathering information from multiple sources relevant to the project objectives. This involves systematically retrieving data from identified sources and consolidating it into a unified dataset for analysis. Once the data is collected, thorough exploration is conducted to understand its underlying structure and properties. Exploring data analysis techniques to understand its underlying structure and properties.

Exploratory data analysis techniques such as summary statistics, data visualization, and correlation analysis are employed to uncover patterns, trends, and potential issues within the dataset. Data cleaning and preprocessing are essential steps to address inconsistencies, errors, and missing values in the dataset. Techniques such as handling missing data, removing duplicates, correcting errors, and transforming variables are applied to ensure data integrity and quality (ScienceDirect, 2016). The main goal of data understanding is to gain general insights about the data that will potentially be helpful for further steps in the data analysis process, but data understanding should not be driven exclusively by the goals and methods to be applied in later steps.

Although these requirements should be kept in mind during data understanding, one should approach the data from a neutral point of view. Never trust any data as long as you have not carried out some simple plausibility checks. Methods for such plausibility checks will be discussed in this chapter. At the end of the data understanding phase, we know much better whether the assumptions we made during the project understanding phase concerning representativeness, informativeness, data quality, and the presence or absence of external factors are justified (Berthold, 2018).

SN	Column Name	Description	Data Type
1	work_year	Year of individual got enrolled.	Integer
2	experience_level	Level of experience an individual has.	String
3	employment_type	Duration of job an individual doing.	String
4	job_title	Type of job of an individual.	String
5	salary	Amount individual getting paid.	Integer
6	Salary_currency	Currency individual getting paid.	String
7	Salary_in_USD	US dollar Currency individual getting paid.	Integer
8	employee_residenc y	The location of the individual.	String
9	remote_ratio	Remote working ratio of an individual.	Integer
10	company_location	Location of the company an individual is working for.	String

Table 1Data Understanding

3. Data Preparation

Data preparation is a critical phase in the data science workflow, where raw data is transformed, cleaned, and structured to make it suitable for analysis and modeling. This phase encompasses a series of activities aimed at ensuring that the data is accurate, complete, and formatted appropriately for the chosen analysis techniques. Effective data preparation lays the foundation for meaningful insights and accurate predictions. The most important activities of data preparation are preparing an analytics sandbox, a central repository environment separate from the production environment. (Data warehouse, data lake, and big data platform).

The repository should collect all kinds of data (summary-level aggregated data, structured data, raw data feeds, and unstructured text data from call logs or weblogs. To get data into the sandbox by performing a combination of extract, transform, and load activities. Learning or understanding the data to clarify what data is accessible Highlight gaps by identifying datasets that are useful but not accessible. Identify external datasets that might be useful to obtain through APIs, data sharing, or purchasing.

3.1 Write a Python program to load data into Pandas Data Frame.

Pandas Data Frame is a common task in data analysis and manipulation using Python. Panda is a powerful library that provides data structures and functions for working with structured data, such as tabular data, intuitively and efficiently.

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt
```

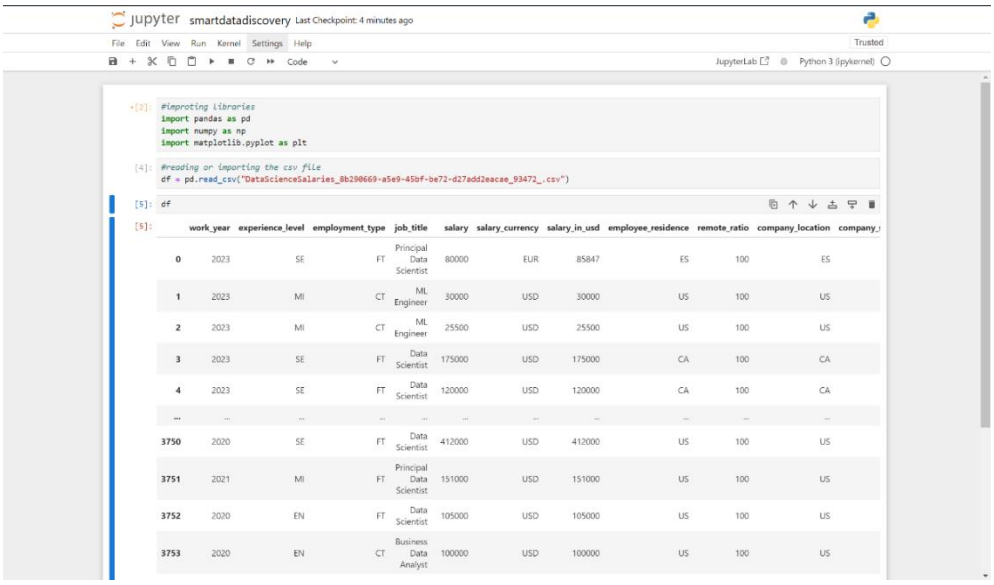


Figure 1 Importing Library

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

It is frequently required to remove unnecessary features from the dataset that have minimal bearing on the main study to achieve analytical clarity and efficiency. Analysts reduce the number of duplicate or unnecessary columns in the dataset such as salary and salary currency which reduces the possibility of dimensionality-related

complications and improves the readability of further analysis.

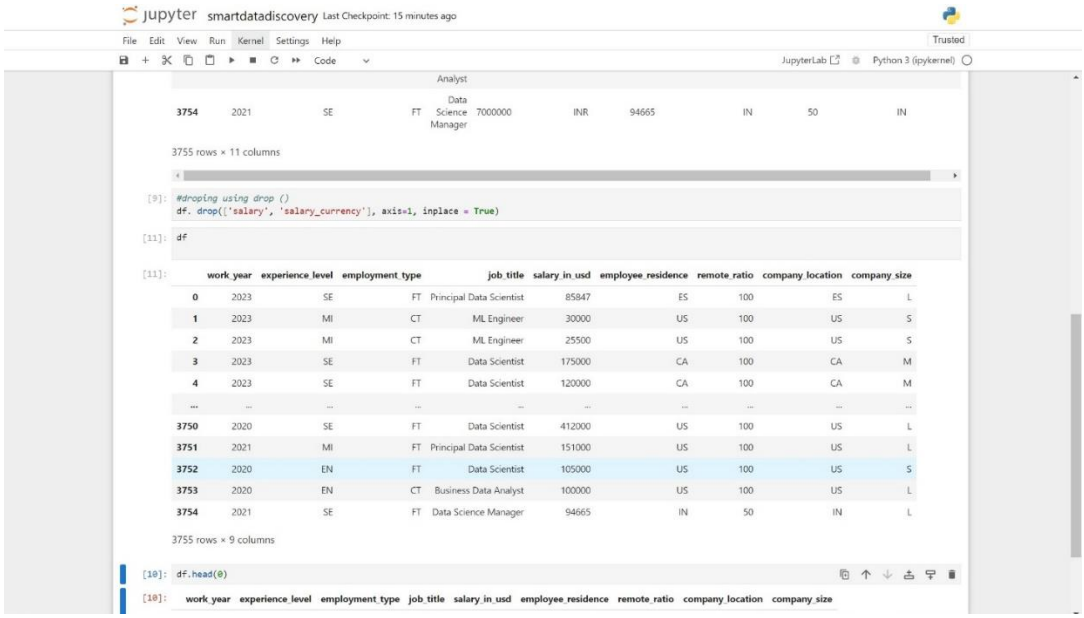


Figure 2Removing unnecessary columns.

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

The existence of missing values, indicated as NaN (Not a Number), significantly compromises the dataset's dependability and integrity. To address this problem, analysts utilize methods to detect and remove NaN values, guaranteeing the accuracy and consistency of the dataset. Analysts can increase the accuracy and usefulness of their conclusions by carefully managing missing data, which strengthens the validity of later analyses.

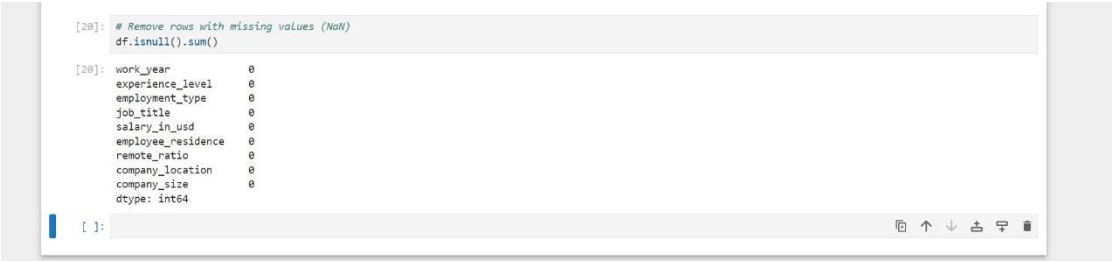


Figure 3removing NaN value

3.4 Write a python program to check duplicate values in the data frame.

Within the dataset, duplicate items have the potential to distort analytical findings, resulting in incorrect conclusions and poor decision-making. To prevent this problem in the first place, analysts put processes in place to identify and eliminate duplicate values, protecting the integrity of the dataset and guaranteeing the accuracy of ensuing studies. Analysts create a more accurate and representative representation of the underlying data distribution by removing entries that are duplicates.

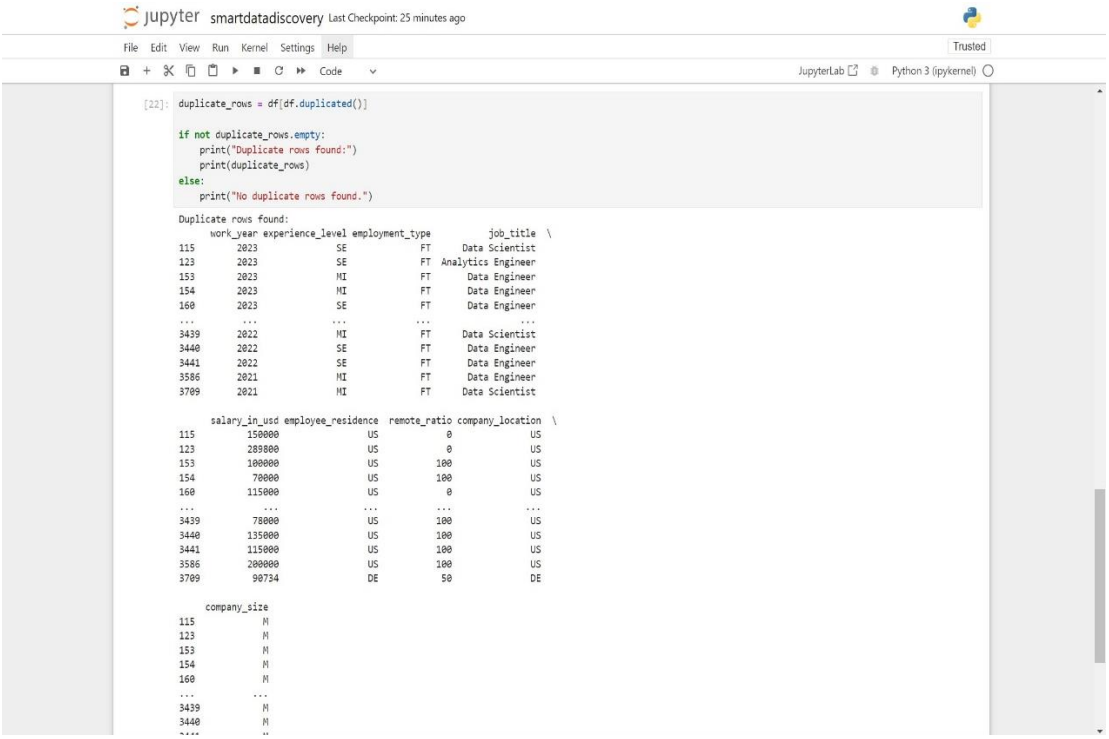


Figure 4Check Duplicate

2.5 Write a Python program to see the unique values from all the columns in the data frame.

Finding distinct values in each dataset column is a fundamental first step in exploring and comprehending the data. Analysts can obtain insights into the categorical distribution of the dataset by counting the unique values in each column. This allows them to detect both common categories and potential abnormalities. This makes it easier to comprehend the fundamental structure of the dataset and opens the door to more complex analysis and interpretations.

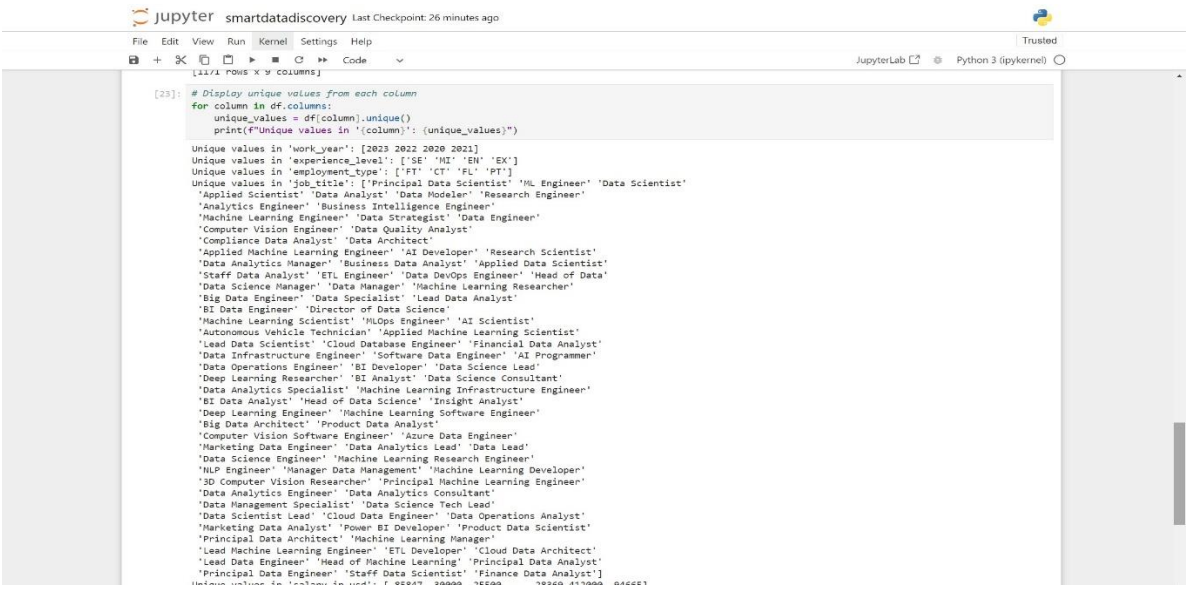


Figure 5Program to see the unique values

2.6 Rename the experience level columns as below.

SE – Senior Level/Expert

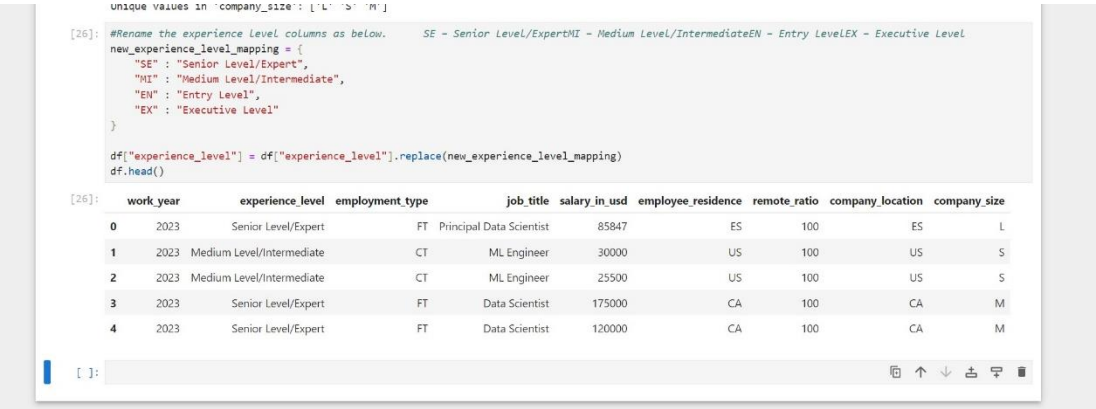
MI – Medium Level/Intermediate

EN – Entry Level

EX – Executive Level

The semantics and importance of each characteristic in the dataset are clarified in large part by the column names. Analysts improve the accessibility and interpretability of the dataset by standardizing and elucidating column names. This promotes easy

communication and information sharing within multidisciplinary teams. Furthermore, analysts can more effectively communicate domain-specific information by renaming columns, which promotes a more nuanced comprehension of the dataset's complexities.



```
[26]: #Rename the experience level columns as below.      SE - Senior Level/ExpertMI - Medium Level/IntermediateEN - Entry LeveLEX - Executive Level
new_experience_level_mapping = {
    "SE" : "Senior Level/Expert",
    "MI" : "Medium Level/Intermediate",
    "EN" : "Entry Level",
    "EX" : "Executive Level"
}

df["experience_level"] = df["experience_level"].replace(new_experience_level_mapping)
df.head()
```

[26]:

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	Senior Level/Expert	FT	Principal Data Scientist	85847	ES	100	ES	L
1	2023	Medium Level/Intermediate	CT	ML Engineer	30000	US	100	US	S
2	2023	Medium Level/Intermediate	CT	ML Engineer	25500	US	100	US	S
3	2023	Senior Level/Expert	FT	Data Scientist	175000	CA	100	CA	M
4	2023	Senior Level/Expert	FT	Data Scientist	120000	CA	100	CA	M

[ ]:

Figure 6Renaming the names of experience-level columns

4. Data Analysis

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

A broad overview of the distributional properties of the dataset is provided by summary statistics, which shed light on shape, dispersion, and central tendencies. Analysts can identify trends, spot outliers, and create hypotheses for more research by calculating summary statistics like sum, mean, standard deviation, skewness, and kurtosis for selected variables. This gives them a thorough understanding of the dataset's underlying distributional properties.

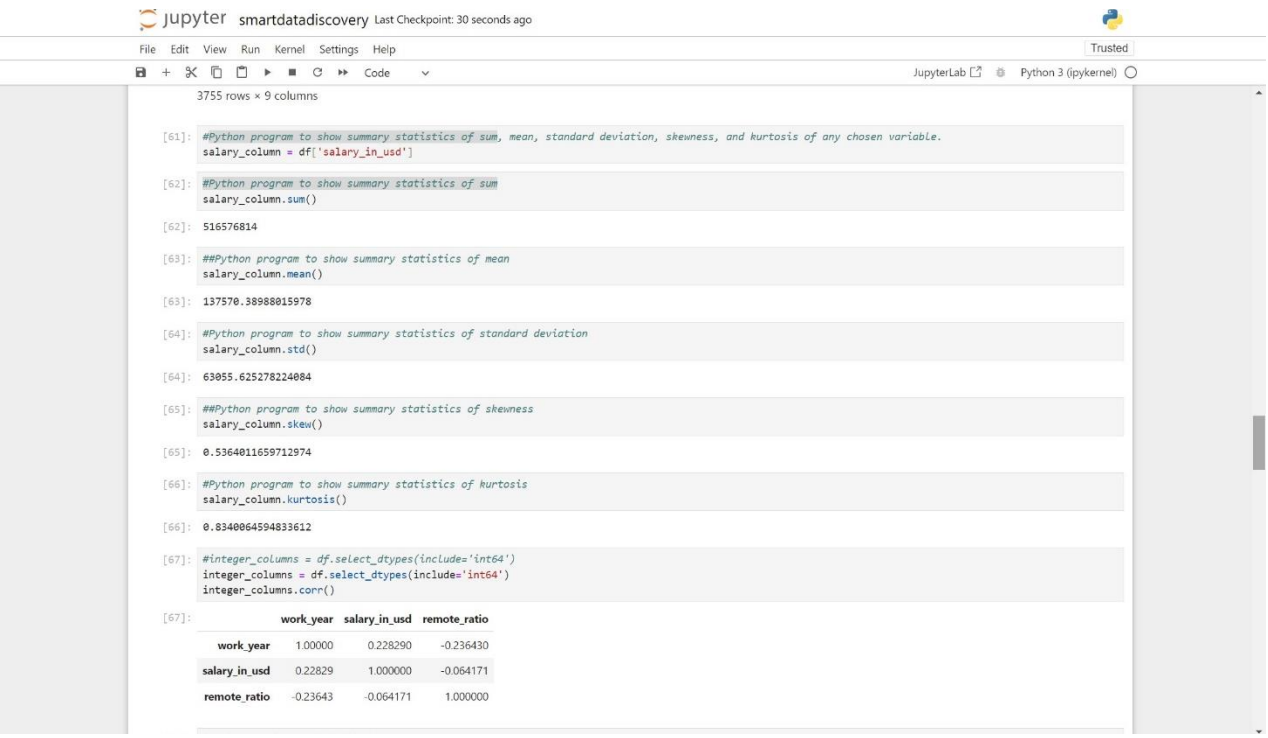


Figure 7Python program to show summary statistics of sum, mean, standard deviation, skewness, and kurtosis of any chosen variable.

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

The foundation of exploratory data analysis is correlation analysis, which helps analysts find connections and dependencies among the variables in the dataset. Analysts detect patterns of association by calculating correlation coefficients between all variables and identifying pairings of variables that show co-occurrence or reciprocal influence. This makes it easier to find predictive associations and provides guidance for later modeling efforts that result in data-driven strategies and interventions.

Figure 8Calculate and show correlation of all variables.

5. Data Exploration

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

The dataset's top 15 jobs are identified, offering an overview of the occupational landscape and illuminating popular job categories and employment patterns. Through the use of a bar graph to illustrate the distribution of job categories, analysts are able to determine the occupational makeup of the dataset and identify the most common job titles along with their respective frequencies.

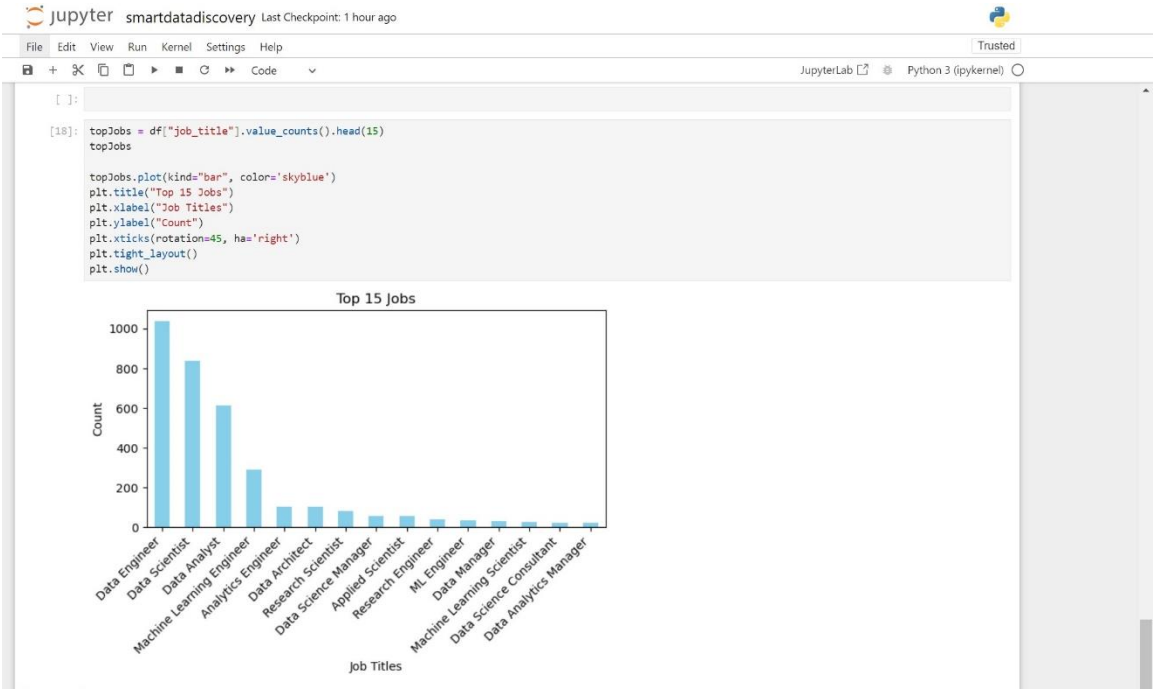


Figure 9Python program to find out top 15 jobs

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

Finding the highest paid job in the dataset provides an insight into the compensation landscape by highlighting high-paying jobs and in-demand skill sets. Analysts discover outliers and abnormalities in the wage distributions across various job titles by displaying the data as a bar graph, which helps them uncover highly sought-after occupations that come with premium compensation packages.

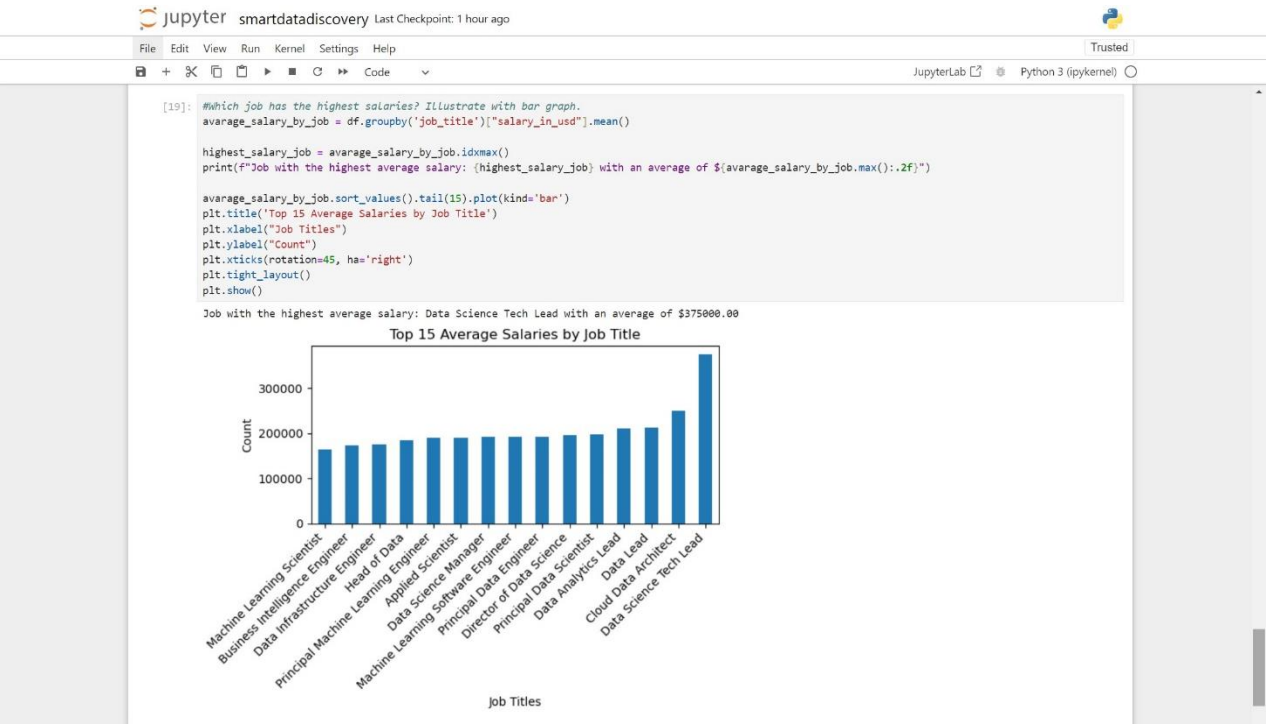


Figure 10Job has the highest salaries

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

Experience level-based salary analysis reveals complex patterns of compensation at various career stages and sheds light on the relationship between tenure and compensation. Through the use of a bar graph to display the wage distributions across different experience levels, analysts are able to discover trends and inequalities, as well as entry-level positions, mid-career milestones, and senior leadership roles.

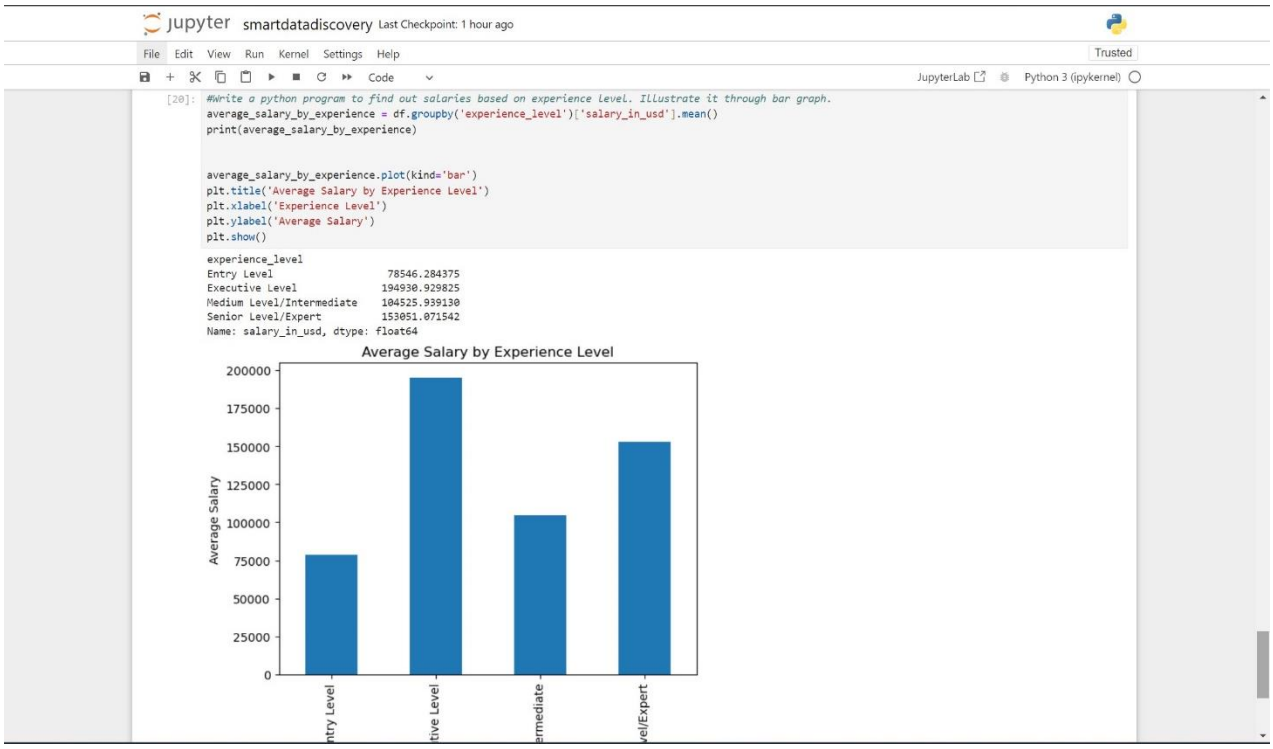


Figure 11Python program to find out salaries based on experience level

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

Strong methods for displaying the distributional properties of numerical variables inside the dataset are histograms and box plots. Plotting histograms and box plots for selected data helps analysts identify outliers, anomalies, and underlying trends by providing insights into the variables' form, central tendencies, and dispersion. Additionally, analysts ensure that the graphical representations are useful as communication tools by adding appropriate labels and annotations to improve the graphical representations' interpretability and clarity.

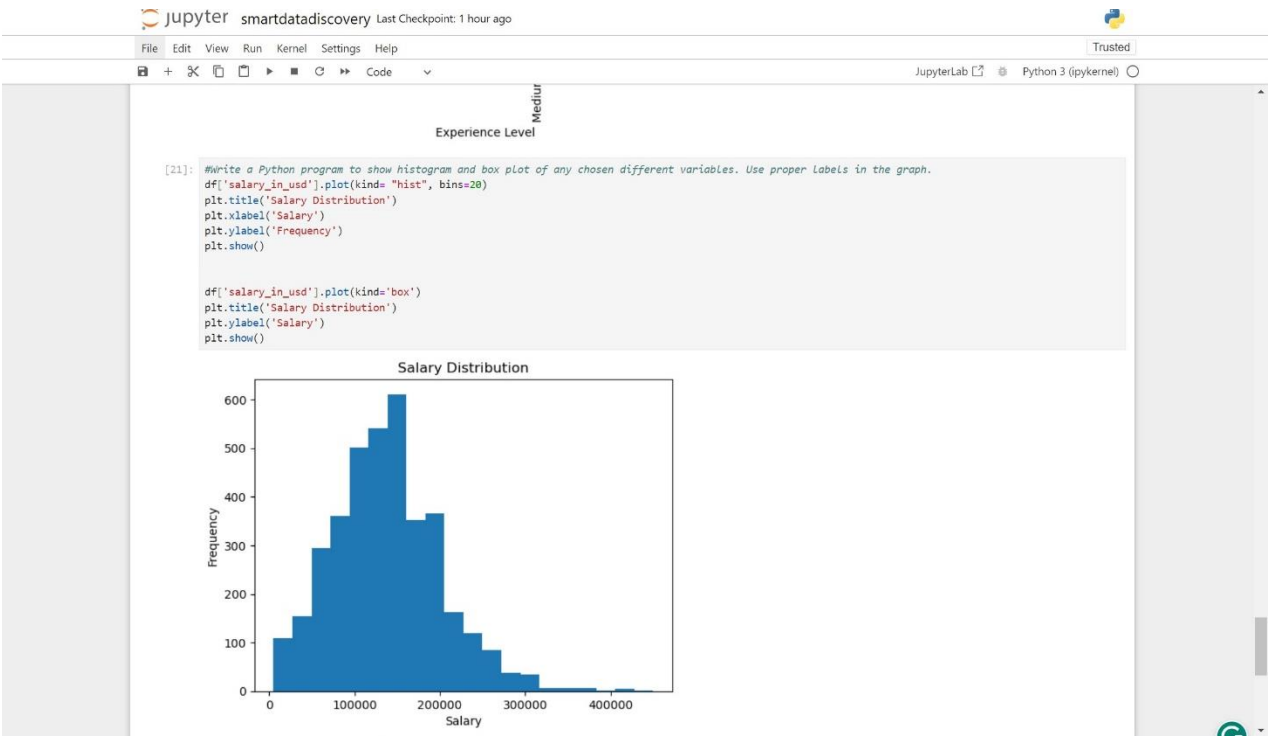


Figure 12Python program to show histogram

6. Conclusion

In conclusion, this assignment has provided us with valuable insights into various aspects of data manipulation, analysis, and retrieval from diverse data sources. Through hands-on experience with Python programming and libraries like Pandas and Matplotlib, we have honed our skills in handling real-world datasets effectively. By delving into tasks such as data preparation, analysis, and visualization, we have gained a deeper understanding of the intricacies involved in working with data. Moving forward, the knowledge and expertise gained from this assignment will serve as a solid foundation for tackling more complex data science challenges and driving informed decision-making in the future.

References

Berthold, P. D. (2018). *Data Understanding.* Chicago: Springer.

ScienceDirect. (2016). Data Understanding. *ScienceDirect*, 1.