Fatima Shamilova

Data Mining

Introduction

We selected the Data Engineer Salary dataset in 2024 for our group project.

Data engineering is a promising career choice in 2024, with high demand and lucrative salary prospects. To become a data engineer, strong programming, and database management skills, as well as knowledge of data warehousing and visualization, are essential.

On this project we will analyze the Data Engineer/Analyst Salary dataset using one of the Python IDE from Kaggle.com.

The purpose of this analysis is to explore and understand a dataset containing salary and employment-related information. The dataset includes various columns such as 'work\_year', 'job\_title', 'salary', 'employment\_type', 'company\_size', and others. The goal is to gain insights into factors affecting salaries and employment patterns.

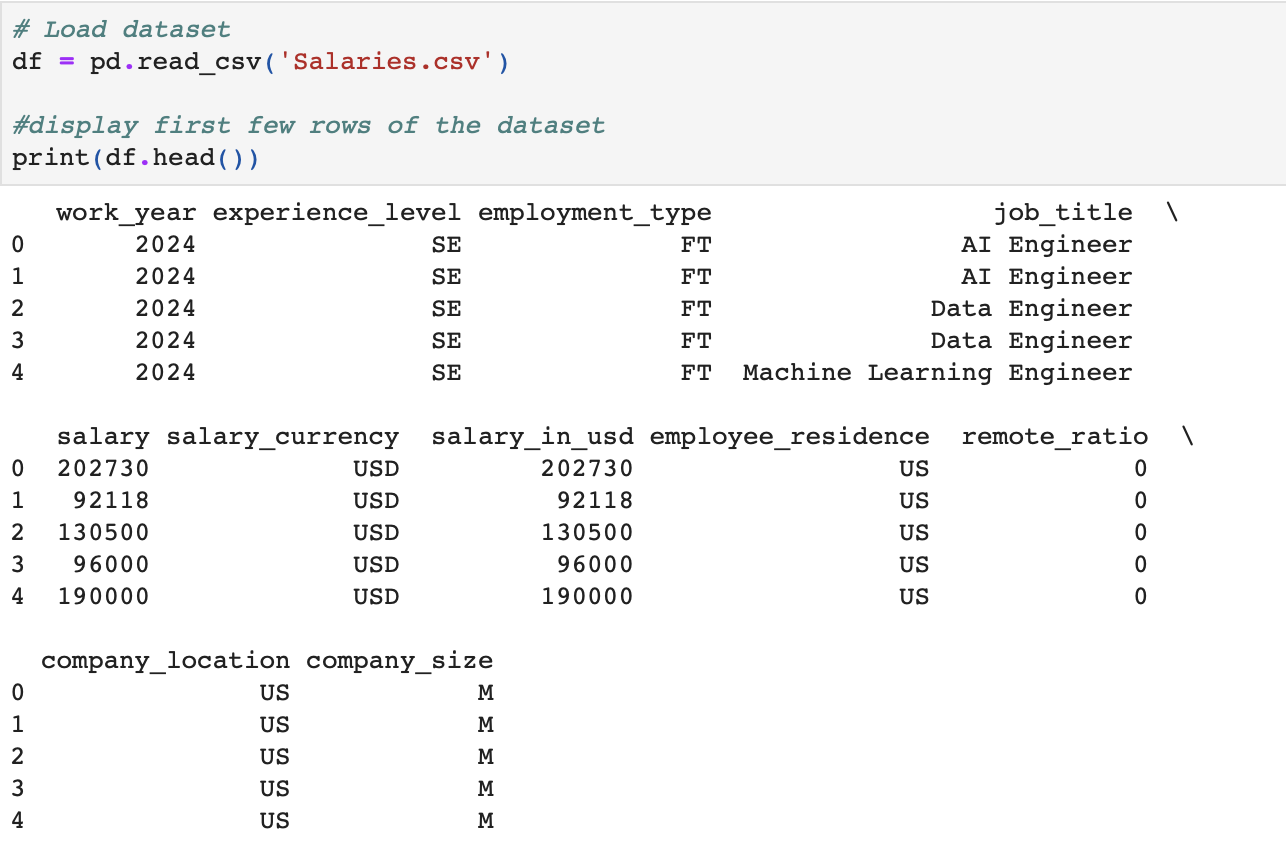
**The Python IDE selection**

The Python3 has been installed to the local machine and print(“Hello World!”) command has been ran for validation.

To install all necessary packages ***pip install*** command has been used.

The actual code was written on Jupyter Notebook

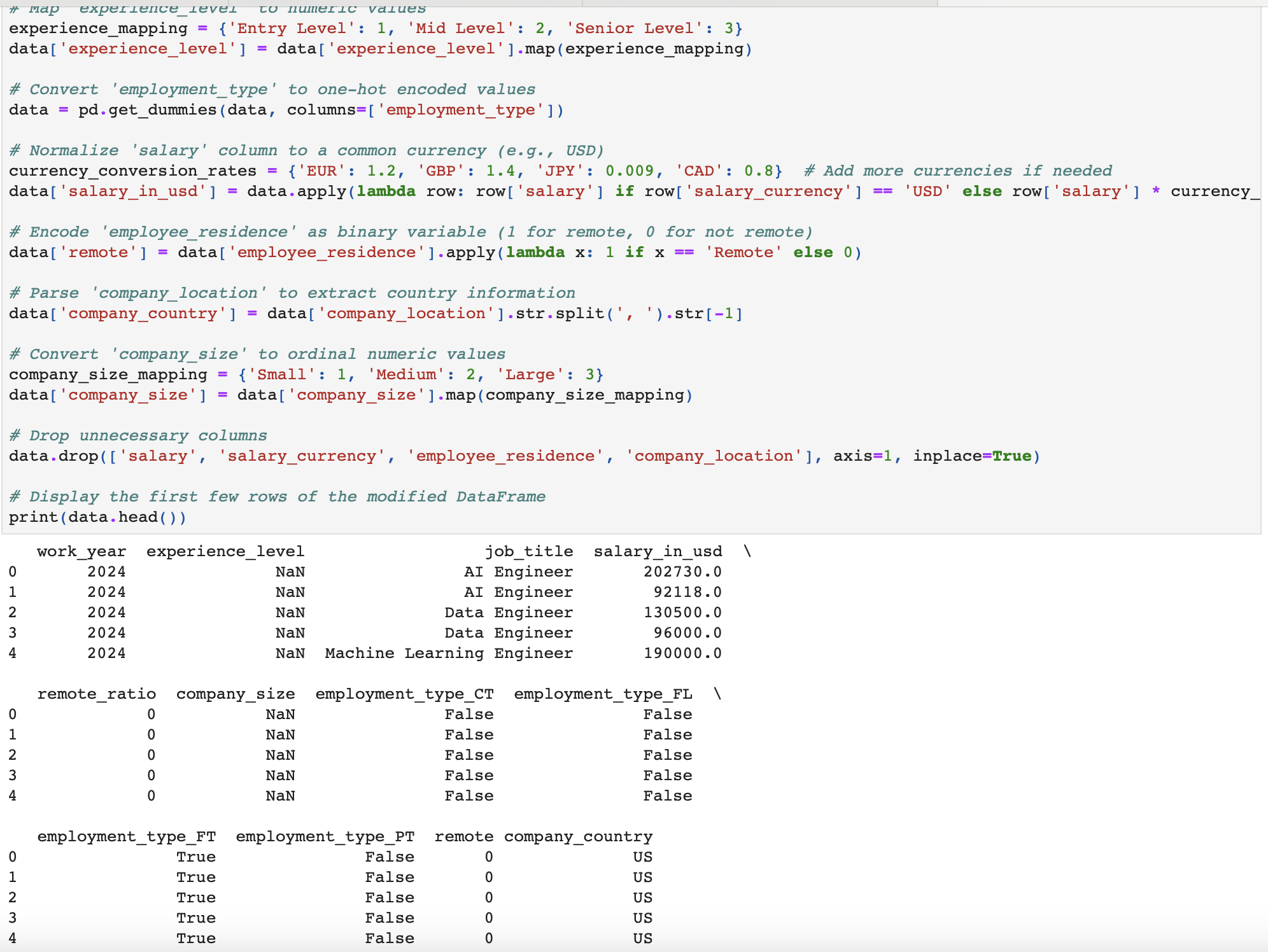
**The dataset analysis and summary**



A close-up of a text

Description automatically generated

This image displays the first rows’ values of the dataset such as work\_year, experience\_level, employment type, job\_title, salary, salary\_currency, salary in USD, employee\_residence, remote\_ration, company\_locations, comoany size.

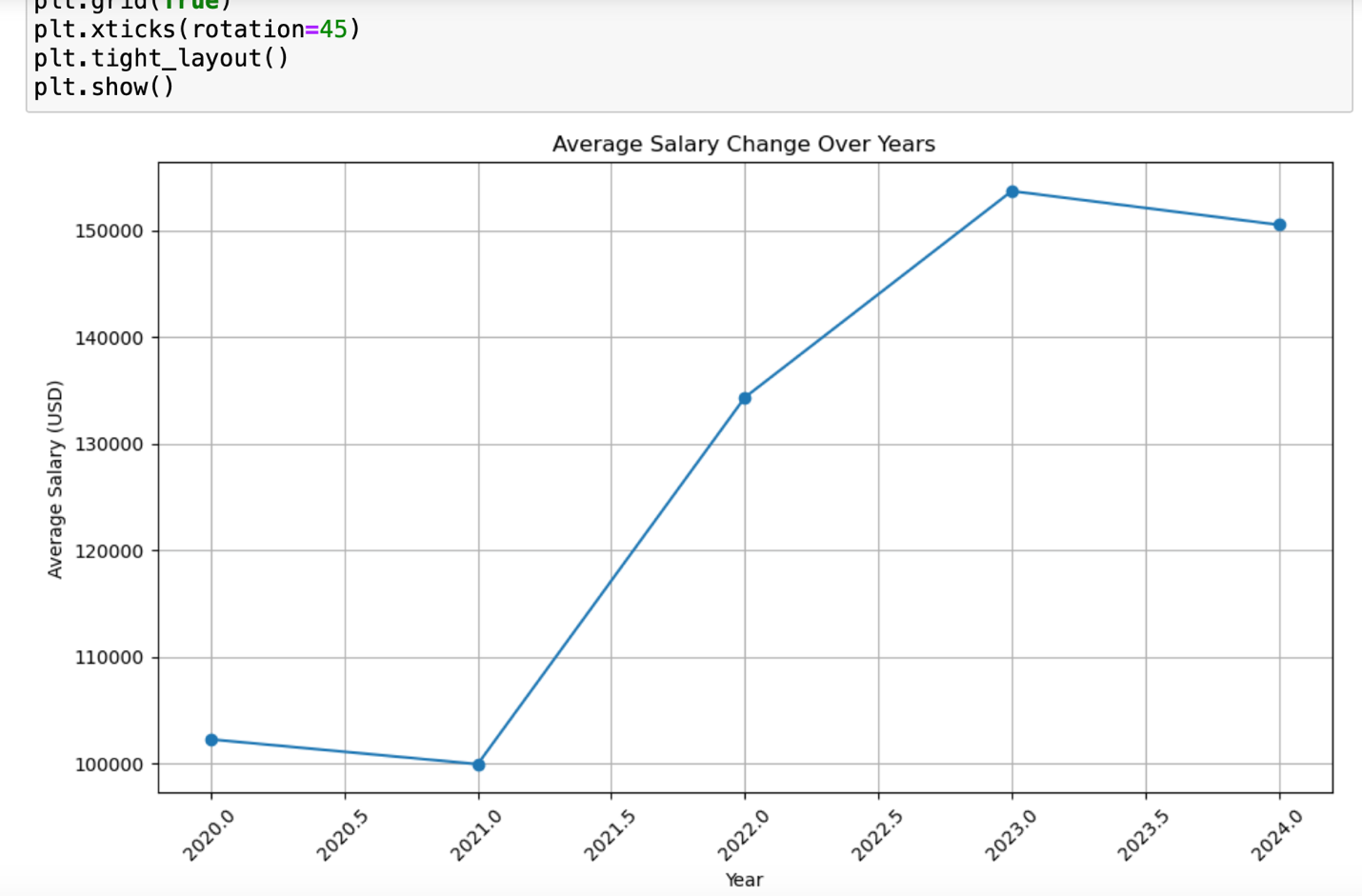


This is the result of the data modification in this version by printing the first few rows of the updated DataFrame using the print statement at the end. For particular needs, modify the print statement or any other section of the code as necessary.

A screenshot of a computer

Description automatically generated

It shows if we have any missing values in our dataset. This is important for other future coding and to get the proper result.



This line chart illustrates the changes in average salary over the years from 2020 to 2024. It indicates a consistent increase in average salary for Data Engineers up until the mid-year of 2023. However, starting from the mid-2023, there is a noticeable decline in the average salary, signifying a reversal in the previous upward trend.

**Filter data based on condition**

**A screenshot of a computer

Description automatically generated**

The filtered DataFrame contains a total of 16534 individuals who have an annual salary exceeding $100,000 USD. Most high-income earners belong to the Senior (SE) or Mid (MI) experience levels, indicating that individuals with higher levels of experience tend to earn higher salaries. The majority of high-income earners are employed in full-time positions, suggesting that full-time employment is associated with higher salaries. High-income earners hold various job titles such as AI Engineer, Data Engineer, and Machine Learning Engineer. Most high-income earners are located in the United States, as indicated by the 'employee\_residence' and 'company\_location' columns.

**Group by job title and calculate the average salary**

**A screenshot of a computer

Description automatically generated**

This summary showcases the average salary for a diverse range of job titles in the AI and data-related fields.

**Salary distribution across different employment types**

A graph of different colored boxes

Description automatically generated

**Salary distribution across different company sizes**

**A diagram of different colored shapes

Description automatically generated**

These visualizations provide insights into how salary levels vary across different employment types and company sizes, offering valuable information for job seekers and employers alike.

**Conclusion**

In this project, we analyzed the Data Engineer Salary dataset using Python. We conducted an analysis of the dataset's characteristics, including the distribution of salary across different factors such as company size, job title, and company location. Visualizations such as boxplots and violin plots were employed to explore these relationships.

Additionally, we addressed challenges associated with non-numeric data by applying encoding techniques like one-hot encoding and target encoding to convert categorical variables into numeric form.

Furthermore, we conducted a thorough check for missing values in the dataset to ensure data completeness. Fortunately, no missing values were found, which facilitated a smoother analysis process.

Throughout the analysis, we leveraged various visualization techniques and statistical methods to unravel the intricate relationships between different variables and salary outcomes. By scrutinizing factors like job titles, company sizes, and employment types, we sought to uncover patterns and correlations that could shed light on salary distributions and employment dynamics.

References

Kaggle.com. Data Engineer Salary dataset in 2024

Python3

Eric Matthes. Python Crash Course: A Hands-On, Project-Based Introduction to Programming. November 2015.