**SHREE RAMSWAROOP MEMORIAL UNIVERSITY**



**PROJECT TITLE:**

**Employee Attrition Analysis using IBM HR Dataset.**

**Submitted by:** Nistha Arora – 202410101150078

Divyata Maurya- 202410101150082

**Course:** B.Tech CS (DS+AI)

**Group:** 33,34

**Subject:** Data Analytics and Reporting

**Submitted to:** Ms. Deepika Tiwari

**Introduction**

In today’s data-driven world, organizations rely heavily on analytics to make informed decisions and improve workforce management. One of the critical challenges faced by Human Resource (HR) departments is employee attrition — the rate at which employees leave the organization. Understanding the reasons behind attrition helps companies enhance employee satisfaction, improve retention, and reduce recruitment costs.

This project, “Employee Attrition Analysis using IBM HR Dataset”, focuses on exploring and analyzing HR data to uncover patterns and factors that contribute to employee turnover.

The analysis was carried out using a combination of tools and techniques:

**Python (Pandas & Matplotlib):** for data cleaning, exploration, and visualization.

**Microsoft Excel:** for additional data analysis using pivot tables, filtering, and conditional formatting.

Through this project, we aimed to gain insights into the key attributes influencing attrition, such as job role, satisfaction level, salary, and working environment. The results of this analysis can help HR departments develop data-backed strategies to reduce employee turnover and improve organizational efficiency.

Ultimately, this analysis serves as a foundation for data-driven decision-making in HR management, enabling organizations to proactively address attrition and foster a more stable and motivated workforce.

**USING PYTHON**

Data Import:

The first step of the analysis involved importing the IBM HR dataset into Python. The dataset was loaded using the Pandas library, which provides powerful tools for handling structured data. The data was read into a DataFrame using the read\_csv() function for easier manipulation and analysis.

**import pandas as pd**

**df = pd.read\_csv("IBM\_HR.csv")**

After loading, the first few rows of the dataset were viewed using df.head() to get an overview of the data and its structure, including column names, data types, and initial records.

Data Cleaning:

Data cleaning is an essential step in ensuring the accuracy and reliability of the analysis. The dataset was checked for missing, duplicate, etc.

**Checking for Null Values:**

The isnull().sum() function was used to identify columns containing missing values.

**Handling Missing Values:**

Missing values were either removed using dropna() or filled with appropriate values using fillna(), depending on their importance.

**Removing Duplicates:**

Duplicate records were identified and removed using the drop\_duplicates() method to maintain dataset integrity.

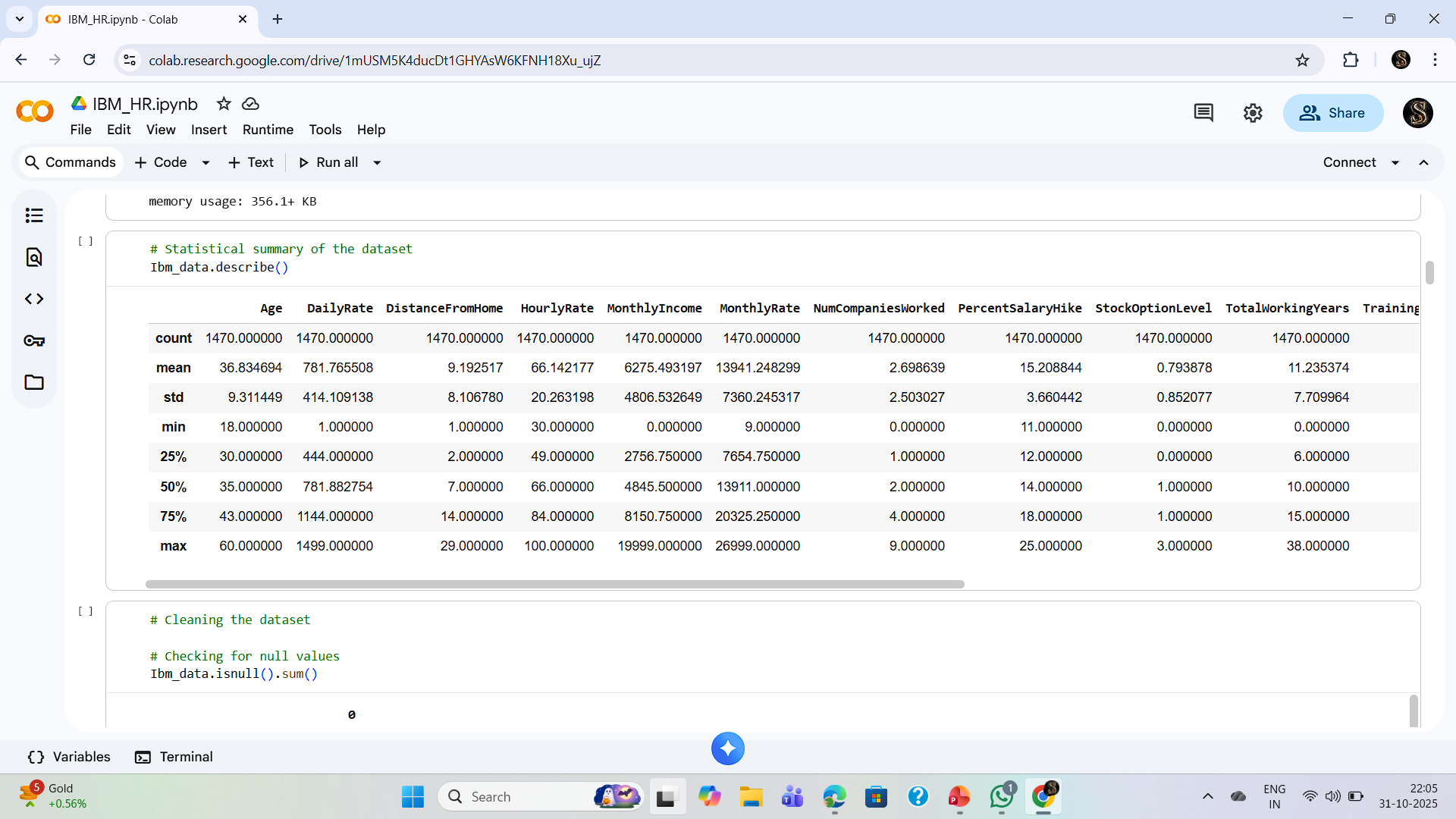
**Sorting and Filtering:**

Data was sorted using sort\_values() and filtered to focus on relevant employee information for analysis.

Statistical Summary of the Dataset:

A statistical summary was generated to understand the overall distribution and characteristics of the data. The describe() function in Pandas was used to display key metrics such as mean, median, standard deviation, minimum, and maximum values.

This summary provided insights into employee demographics, job satisfaction, and work-related factors — forming the foundation for further analysis and visualization in subsequent stages.



The Python-based preprocessing stage laid a strong foundation for accurate data analysis. By carefully importing, cleaning, and summarizing the IBM HR dataset, we ensured that the data was free from errors and inconsistencies.

This process not only enhanced the quality of insights but also prepared the dataset for meaningful visual exploration and trend identification in the next phase of the project. The cleaned and structured data thus served as the cornerstone for reliable employee attrition analysis, enabling precise and actionable results.

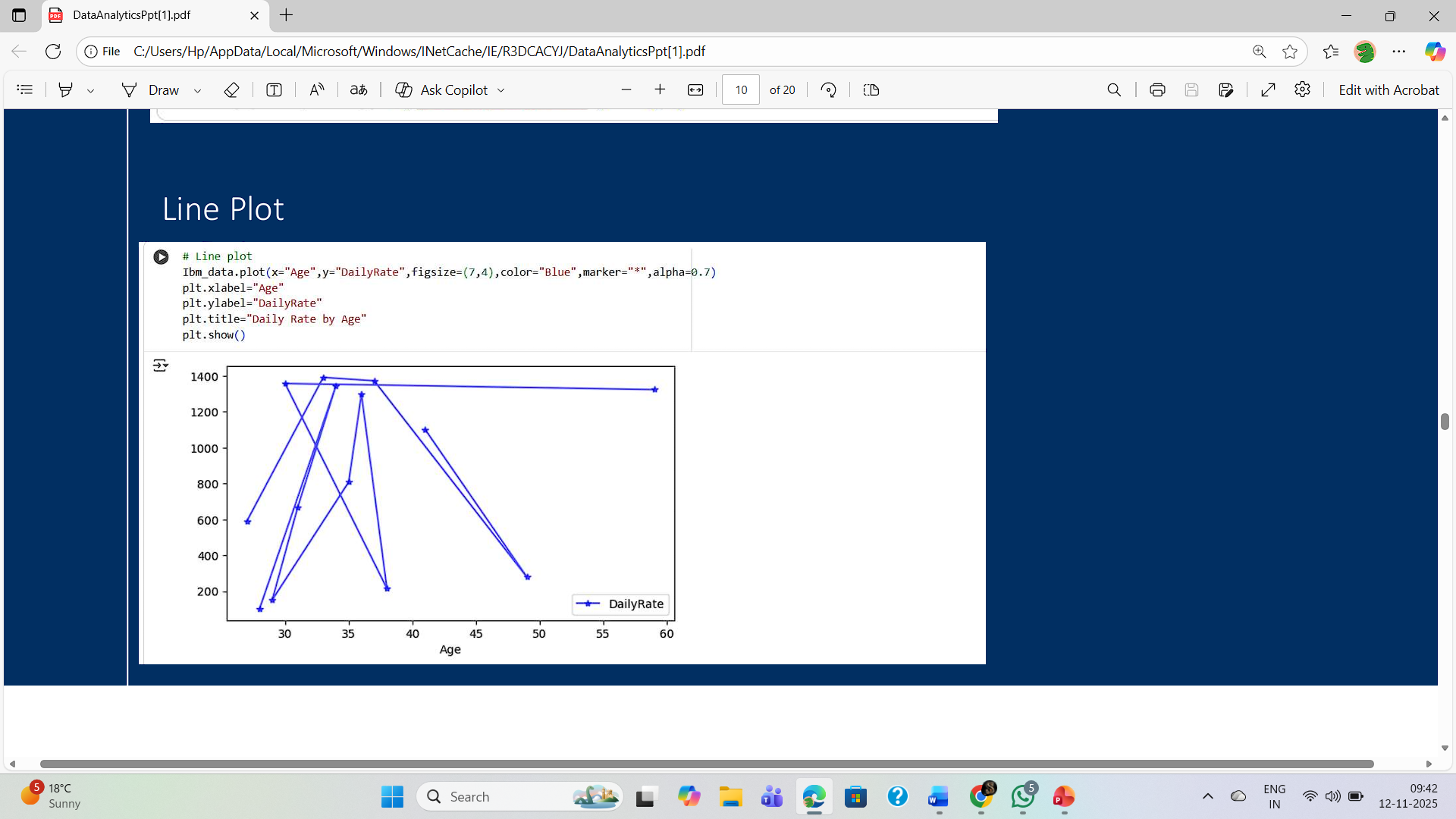
**DATA VISUALIZATION**

(Using Matplotlib)

Data visualization plays a key role in understanding patterns and trends within the dataset. Using the Matplotlib library in Python, different types of plots were created to visually explore relationships between variables affecting employee attrition.

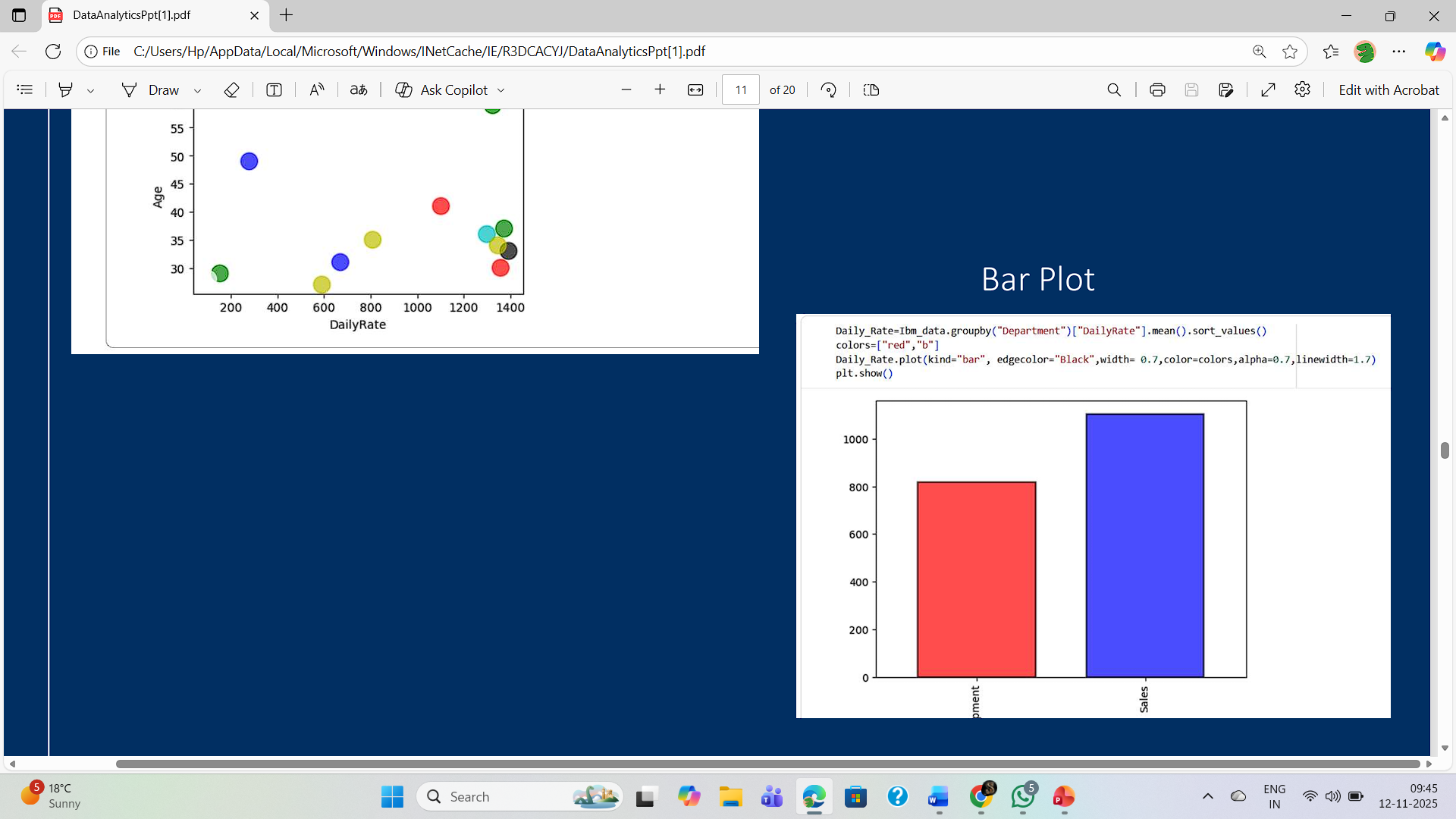
**Line Plot:**

Used to observe general trends and patterns over continuous variables.



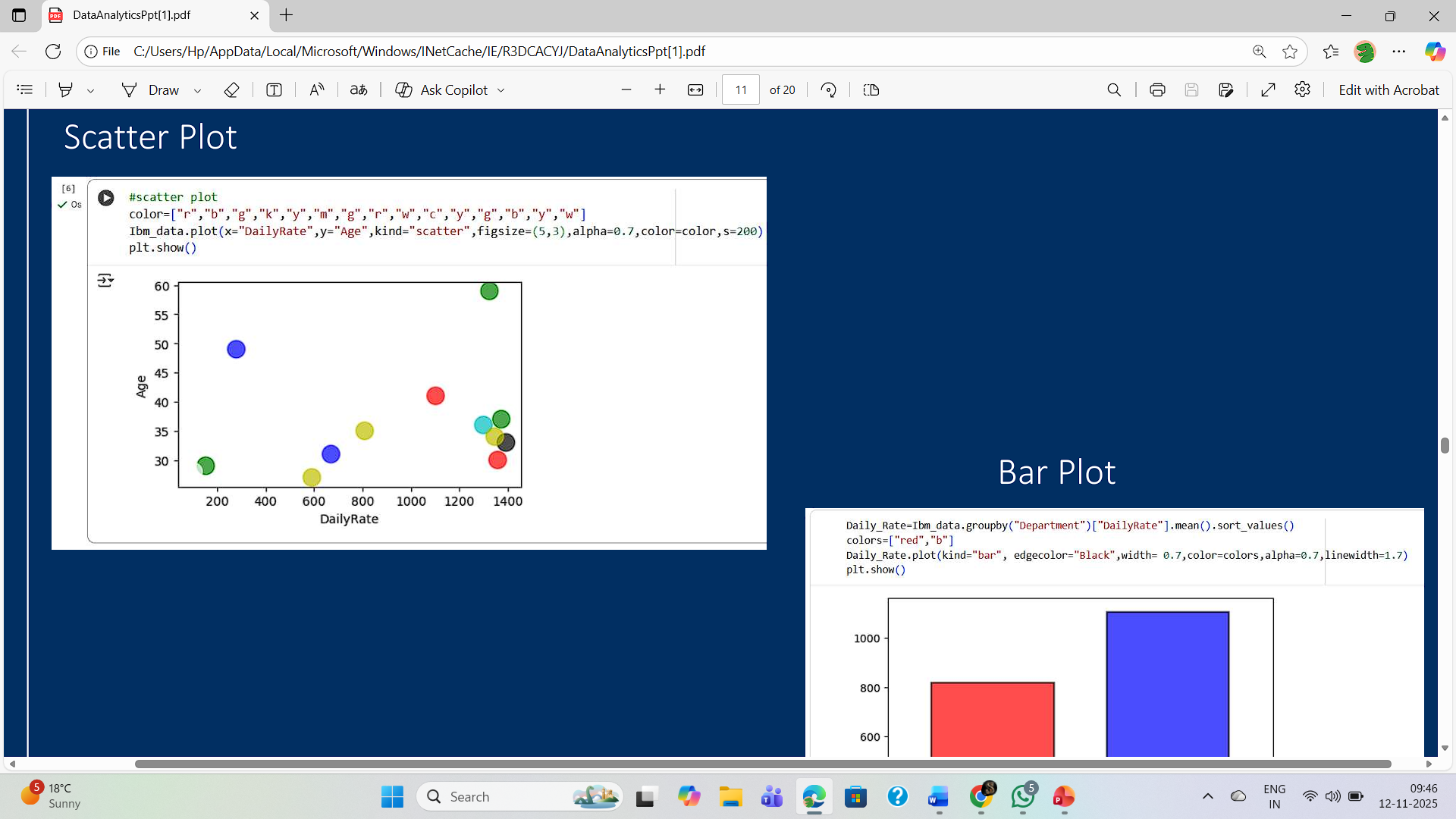
**Bar Plot:**

Bar Plot displayed comparisons of categorical data such as job roles and departments.



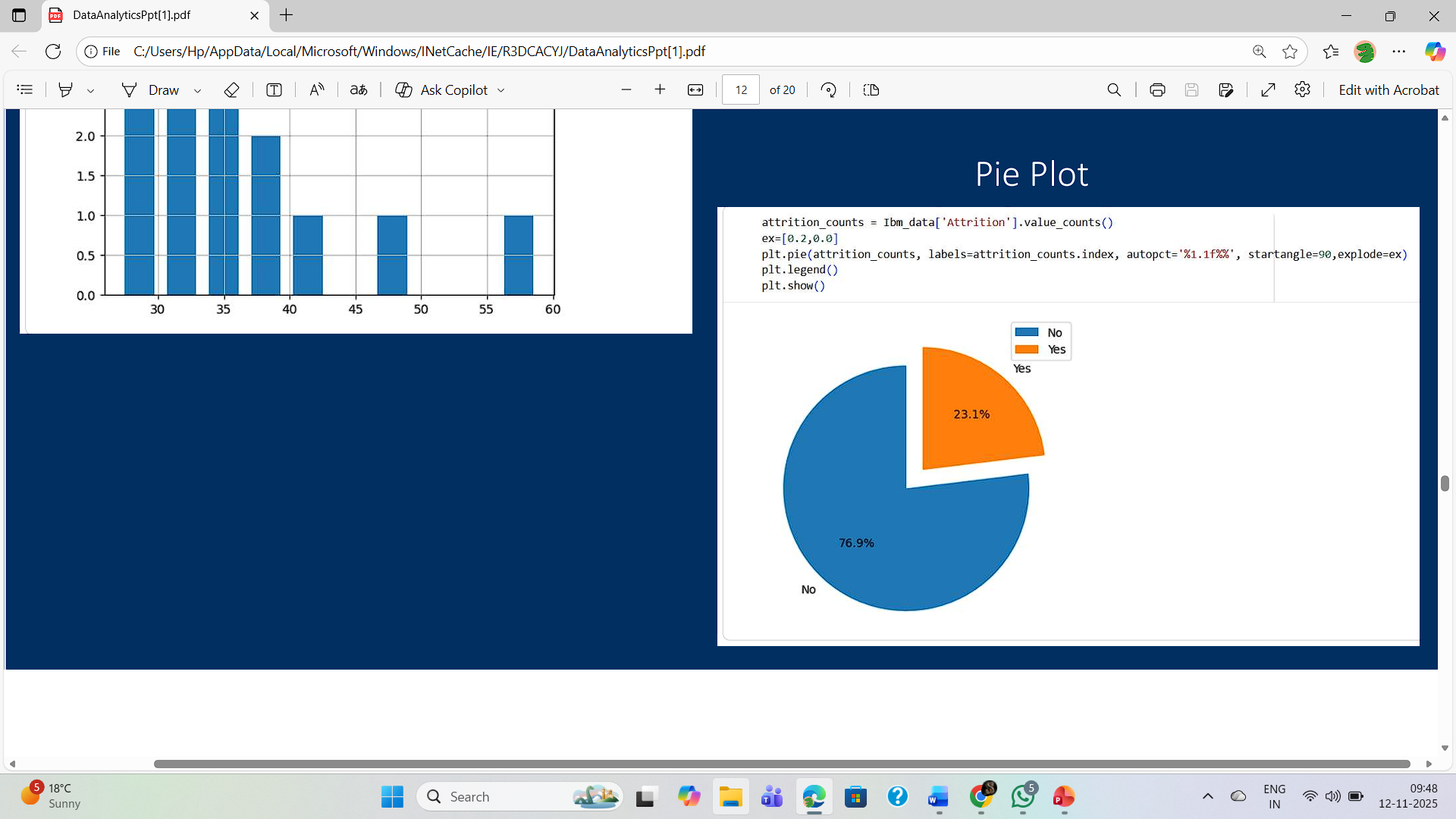
**Scatter Plot:**

Scatter plot help identify relationships between numerical features like age, salary and attrition.



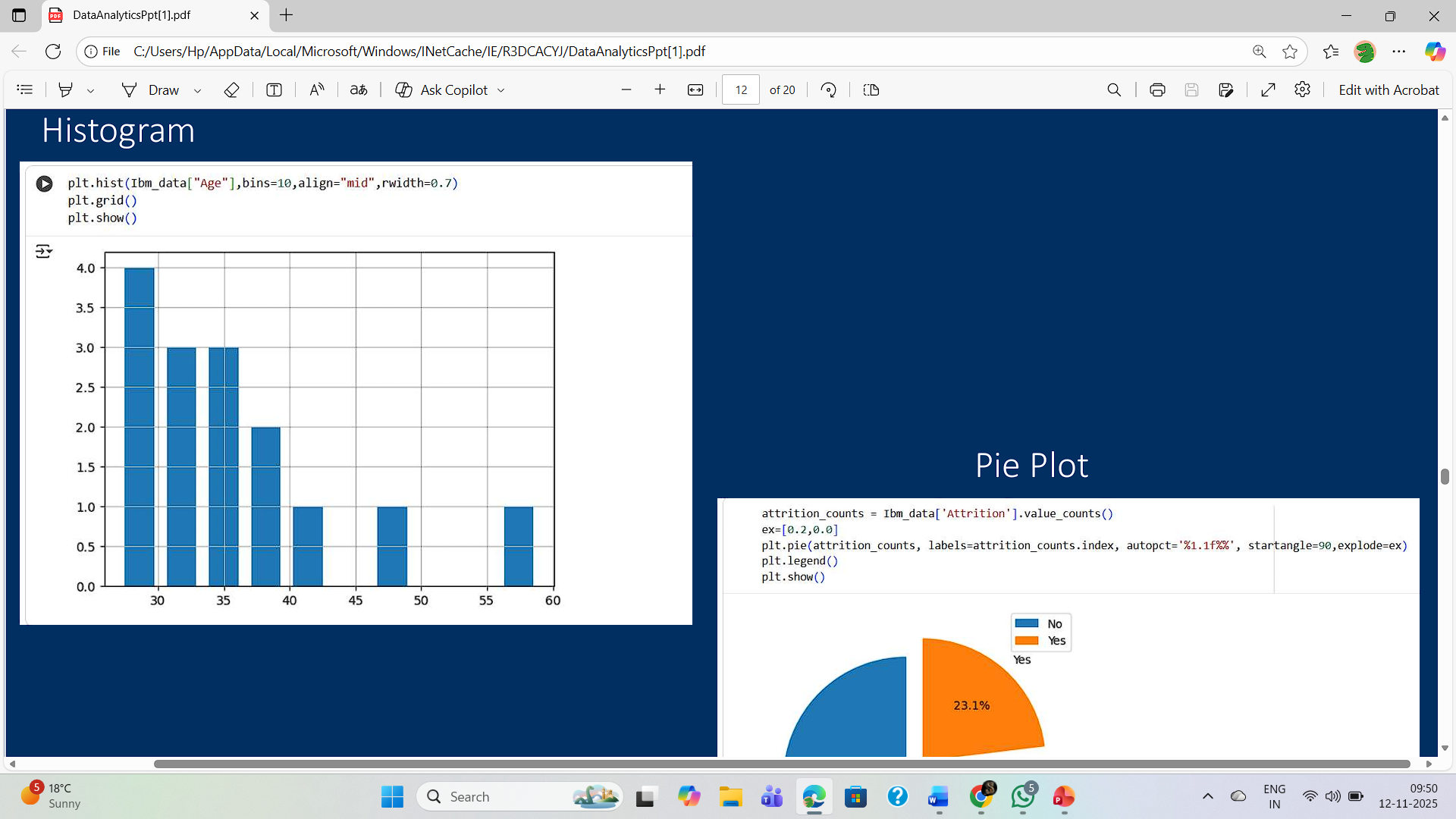
**Pie Plot:**

Represented the proportion of employees who stayed versus those who left the company.



**Histogram Plot:**

Showed the distribution of employees satisfaction and performance ratings.

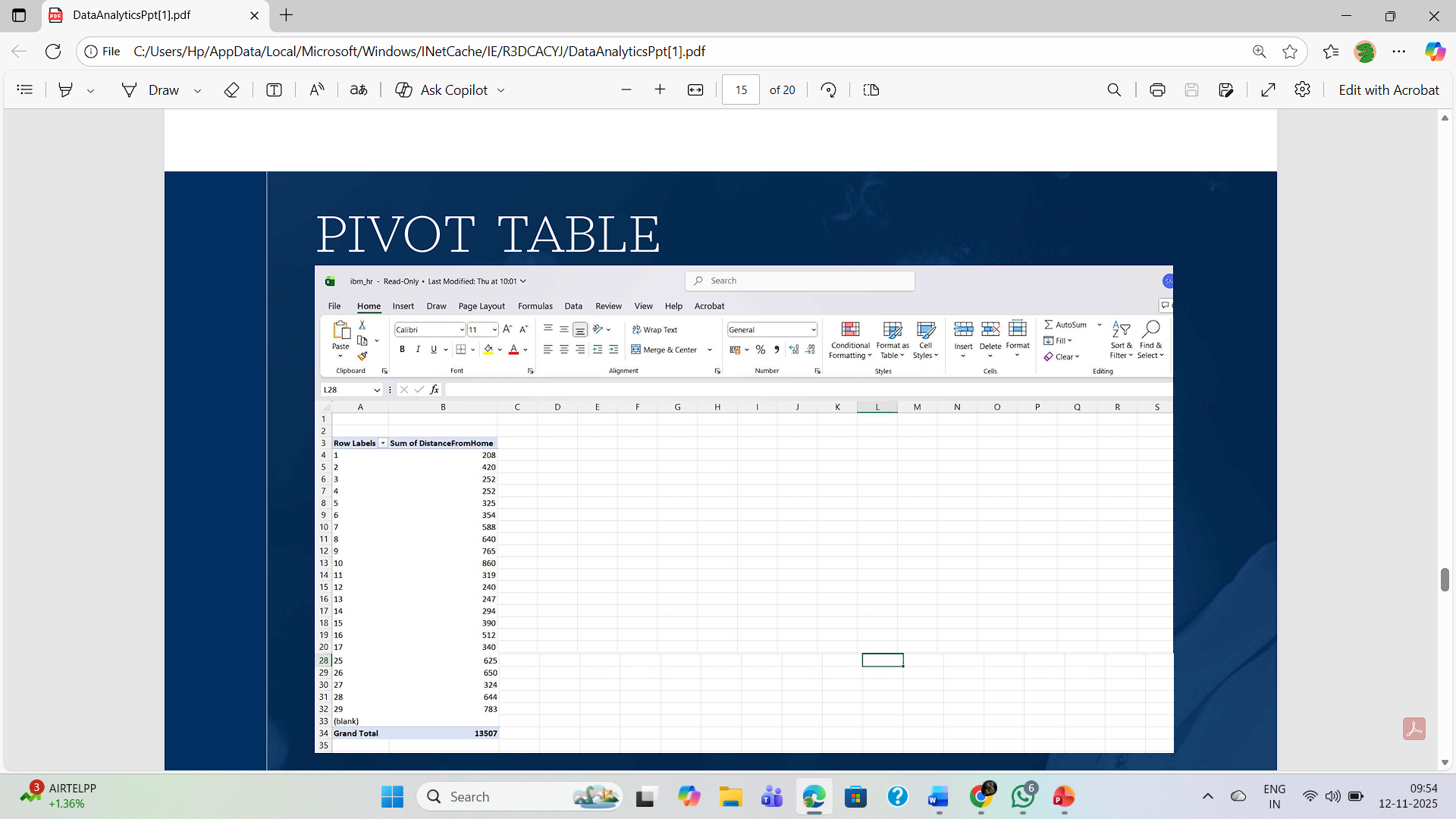


**USING EXCEL**

Microsoft Excel was also used to perform additional analysis and visualization of the IBM HR dataset. Excel provided a simple and interactive environment to explore data trends and summarize employee information efficiently.

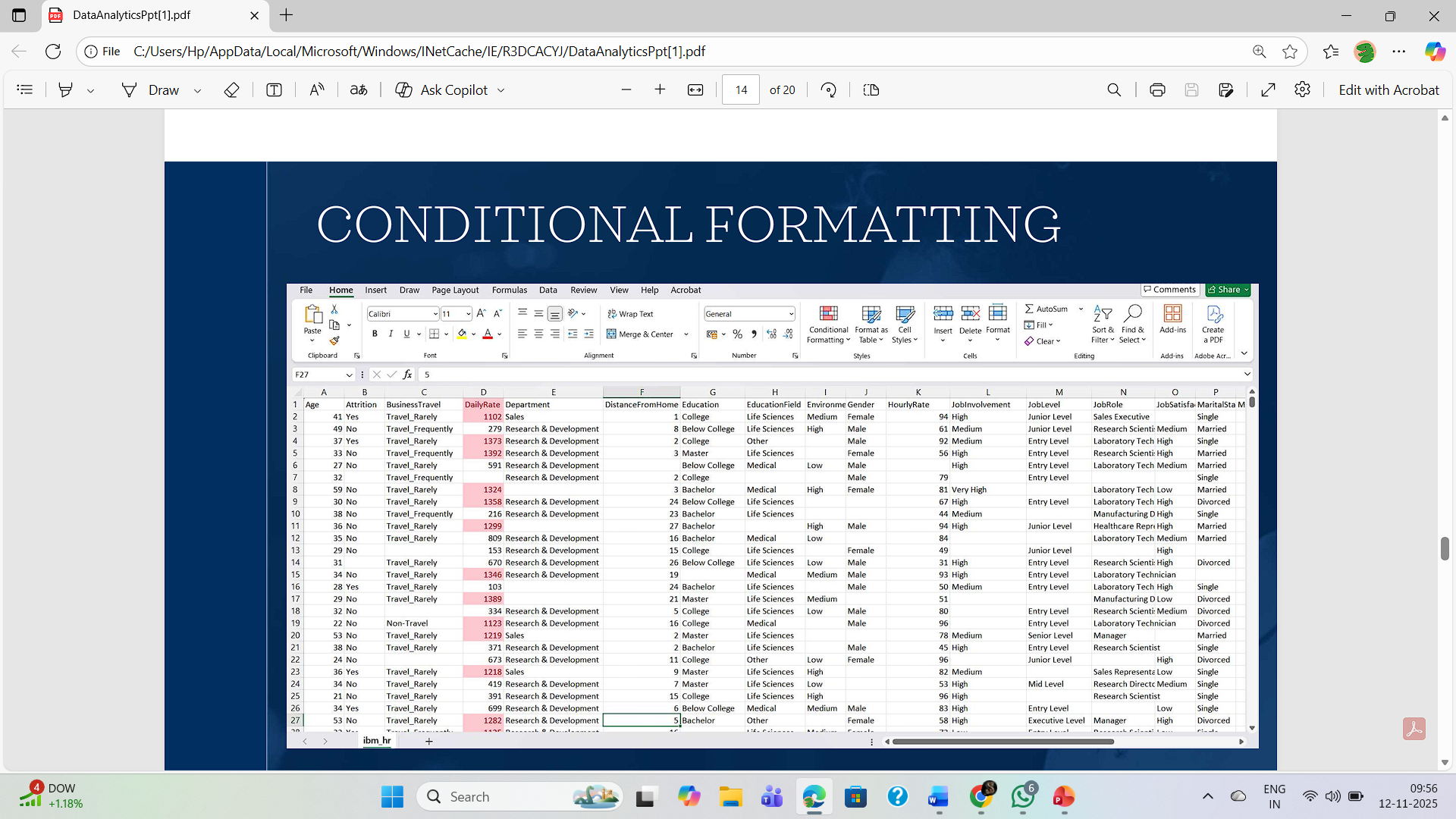
**Pivot Tables:**

Pivot tables were created to summarize employee details such as department-wise attrition rate, average age, salary, and job satisfaction.



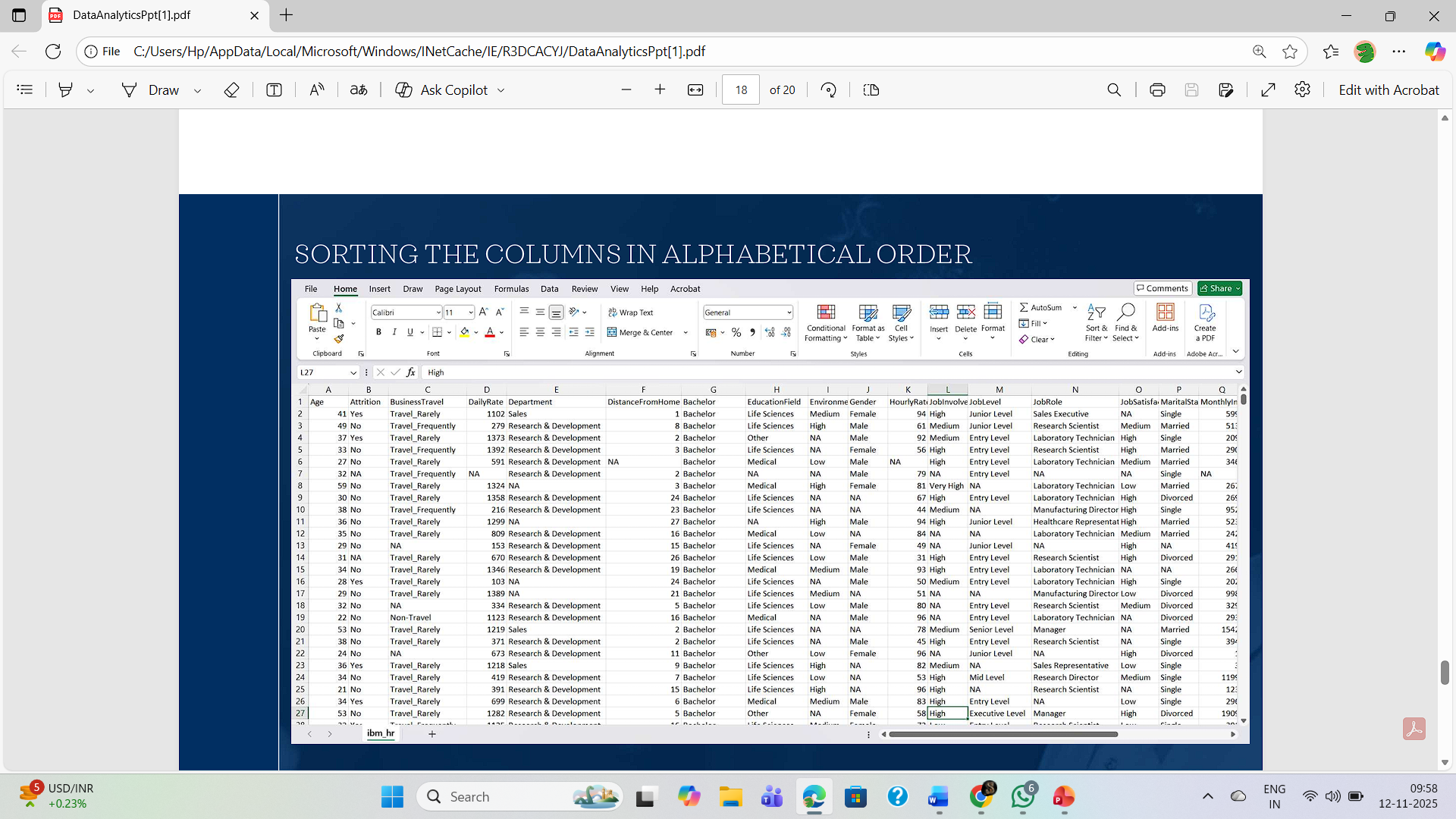
**Conditional Formatting:**

Conditional formatting was used to highlight important patterns — for example, high attrition rates, low satisfaction scores, or salary ranges — making the data visually easier to interpret.



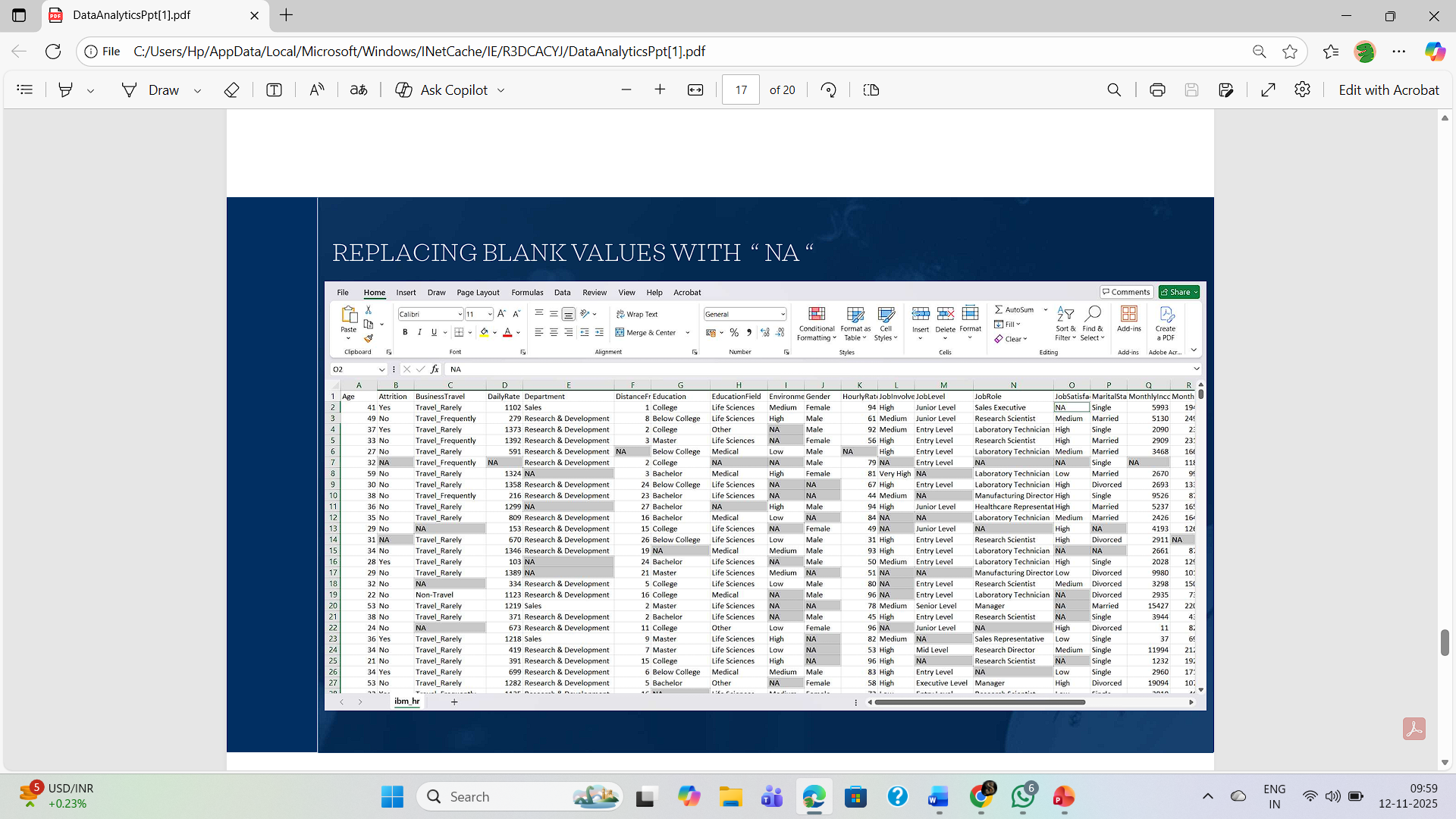
**Sorting the Dataset in Excel:**

Sorting the dataset in Excel is an important step to organize and analyze employee information effectively. It allows data to be arranged in a specific order based on one or more columns, helping to identify trends and patterns easily.



**Highlighting Blank Values and Replacing with “NA” :**

During data cleaning, it is essential to identify and handle missing values to maintain the accuracy and consistency of the dataset. Blank cells can lead to incorrect analysis and must be replaced with appropriate values.



**CONCLUSION**

The analysis of the IBM HR dataset provided valuable insights into the factors influencing employee attrition. Through data cleaning, visualization, and interpretation using Python and Excel, patterns such as job satisfaction, income level, and work-life balance were identified as key contributors to employee turnover.

Using Python, the dataset was efficiently cleaned, summarized, and visualized through meaningful charts and graphs. Excel further supported the analysis through pivot tables, sorting, and conditional formatting, helping to cross-verify results and present them in a simplified manner.

This project highlighted the importance of data-driven decision-making in Human Resource Management. By identifying the causes of attrition, organizations can implement targeted retention strategies, improve employee engagement, and reduce recruitment costs.

Overall, this project emphasized the crucial role of analytics in predicting and managing employee behavior. By systematically exploring the IBM HR dataset, we learned how various demographic and professional factors influence attrition rates. The findings can help organizations design better work environments, enhance job satisfaction, and promote long-term employee retention. This analysis not only strengthened our understanding of data handling and visualization but also demonstrated how analytical tools can contribute to effective business decision-making.