**Employees Data Analysis**

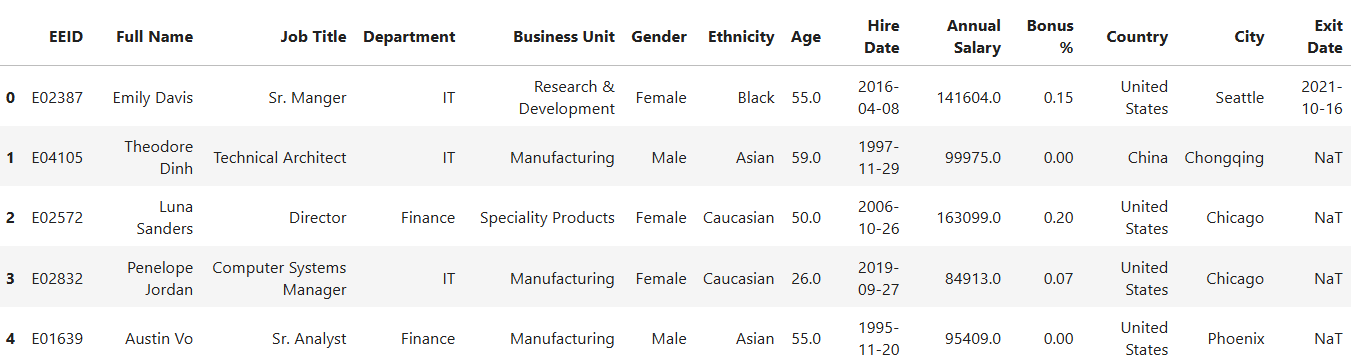
**Task 1 (EDA) Report – Estarta AI training**

**Prepared Nashat Al-Farajat**

**Table of Content**

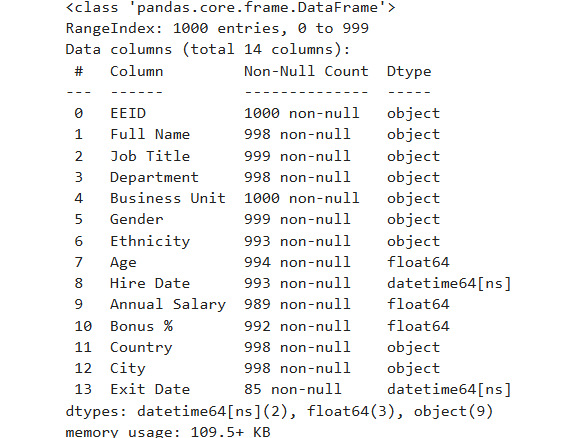
1. Introduction (About the Dataset) ...............................................................
2. Initial Data Exploration ....................................................................
3. Data Cleaning ..................................................................................  
    3.1 Handling Missing Values, Methods Used ..............................................  
    3.2 Handling Duplicates .......................................................................  
    3.3 Handling Outliers ............................................................................
4. Changing the First Five Rows ...............................................................
5. Data Aggregation and Analysis ..........................................................  
    5.1 Row with Largest Salary .................................................................  
    5.2 Group by Department: Average Salary and Age ......................  
    5.3 Group by Department + Ethnicity: Max/Min Age, Median Salary
6. Visualizations and Insights .................................................................
7. Conclusion and Recommendations...............................................................

**Introduction (about the Data Set):** This dataset contains detailed employee information from a company, with 1,000 records. Each entry includes details such as employee ID, name, job title, department, unit, gender, age, salary, bonus %, hire and exit dates, country and city. The data can be used to explore employee details, pay levels, and how teams are set up within the company.

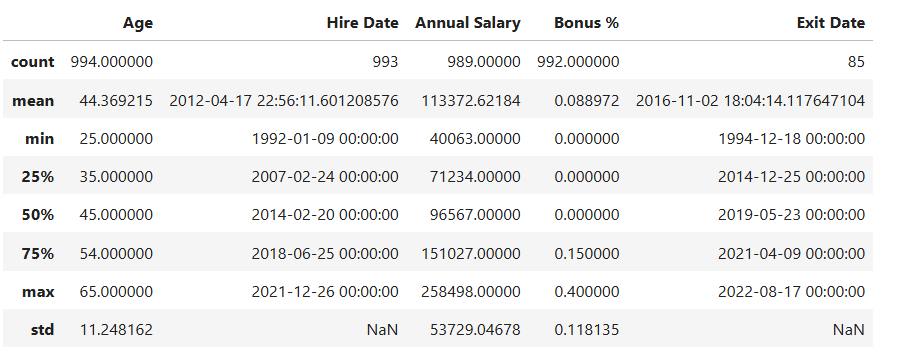
****

**Initial Data Exploration:** Before cleaning and processing, a basic exploration was conducted to understand the structure and content of the dataset.

* **Dataset Shape**: The dataset contains **1,000 rows** and **14 columns**.
* **Data Types** and Non-Null Counts (Several columns contain missing values, which will be addressed in the data cleaning stage).

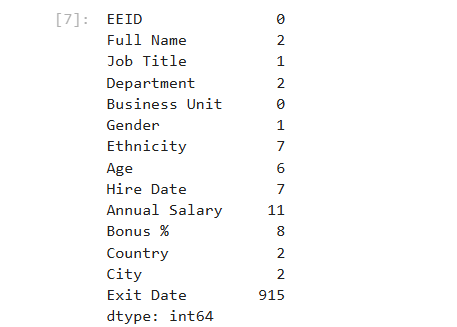


* **Summary Statistics** (Selected Numerical and Date Columns)



**Data Cleaning**

**Handling Missing Values:** The table below summarizes the number of missing values in each column.

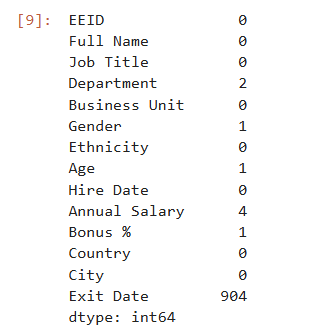
****

**Note:** Missing values in **Exit Date** are normal because they indicate employees who are still working at the company.

**Methods used**:

1. **Dropping the rows with missing values**: Rows with missing values in **Full Name**, **Job Title**, **Ethnicity**, **Hire Date**, **Country**, and **City** were removed because these fields are essential for accurate identification and analysis. For example, when **City** is missing, **Country** is often missing too, making it impossible to reliably fill these values, Initially, the dataset contained 1,000 rows. After removing records with missing values in critical fields, 993 rows remain, meaning **only 0.7% of the data was removed.**

Number of missing values in each column after dropping the rows.

****

#### 2 - Manual Filling of Missing Departments and Gender:

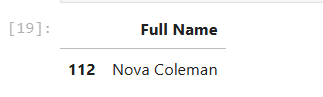
For the small number of missing values in the **Department** column (2 rows), the departments were manually filled by referencing the corresponding **Job Title**. This approach leverages the logical relationship between an employee’s role and their department to ensure accurate completion of missing data.

For the two employees with missing **Department** values, the **Job Titles** were **System Administrator** and **Business Partner**. Based on these roles, the missing departments were manually filled as follows:

* **System Administrator** → **IT Department.**
* **Business Partner** → **Human Resources Department.**

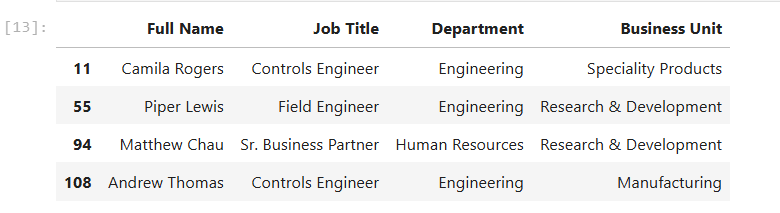
This ensured the dataset remains accurate and consistent.

For the single missing value in **Gender**, the employee’s name was **Nova**, which is typically a female name. Based on this, the missing gender was manually filled as **Female**.

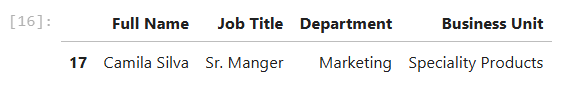


1. **Filling by Mean for department(same department, business unit and job title), bonus % and age**:

Missing values in **Annual Salary** were filled using the **mean salary** of employees who share the same **Department**, **Business Unit**, and **Job Title**.  
This method provides a more accurate estimation by considering employees in similar roles and organizational context, rather than using a general company-wide average.



Missing values in **Bonus %** were filled using the same method as **Annual Salary** — by calculating the average bonus of employees with the same **Department**, **Business Unit**, and **Job Title**. This helps maintain consistency within similar employee groups.

****

For the only missing values in the **Age** column, the overall **mean age** of all employees (~44) was used. Since age is less tied to specific roles or departments, a global average was considered appropriate.

**Handling duplicates:**

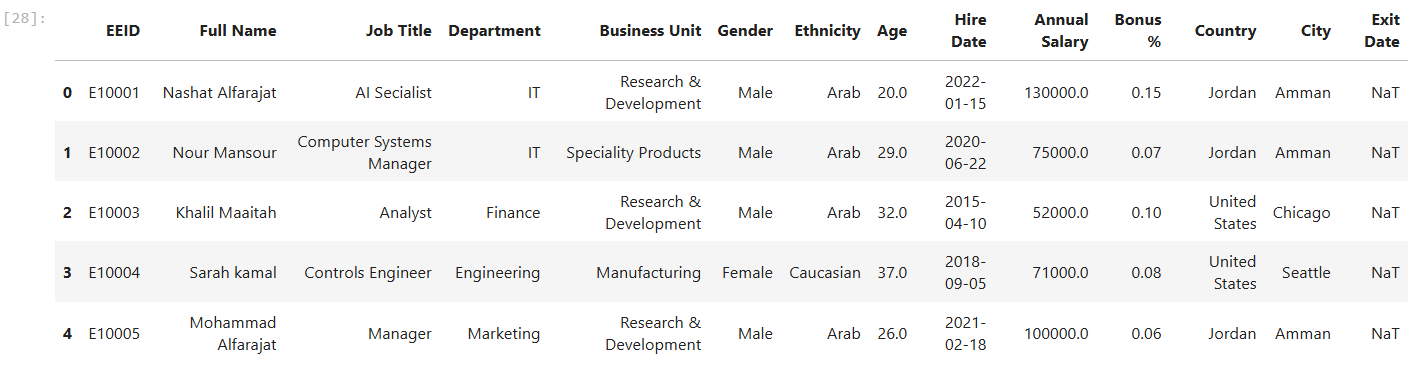
The dataset was checked for duplicate rows, and **no duplicates were found**. Therefore, no action was needed in this step.

**Handling outliers:**

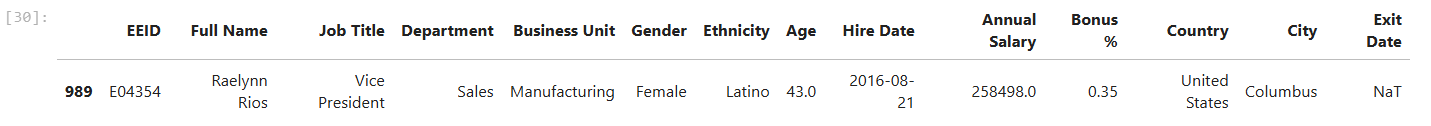
The **Annual Salary** and **Age** columns were examined for outliers using **interquartile range (IQR)** and **box plot** analysis. No significant outliers were detected, so no records were removed or modified in this step.

**Bonus %** column was also examined for outliers, there were outliers by calculating but logically it was normal since all outliers were for employees with the job title vice president, so it makes sense that they have a high very bonus.

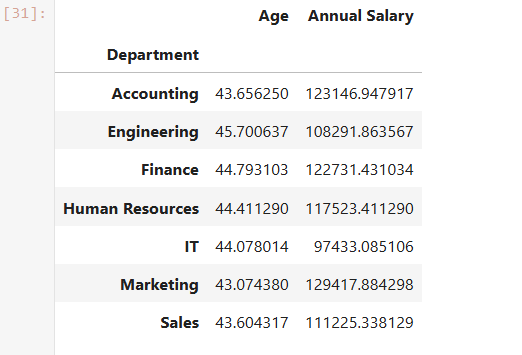
**Changing the first five rows:**

****

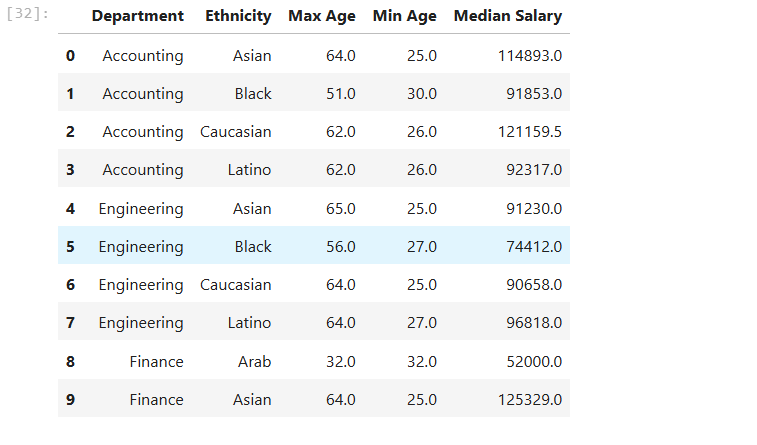
**Row with largest salary:**

****

**Group by Department, getting average salary and average age:**

****

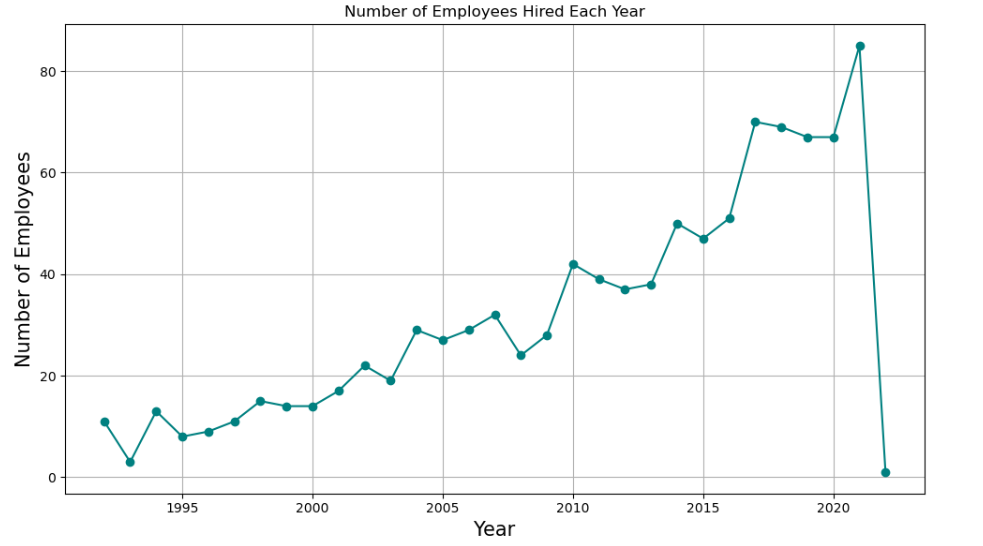
**Group by Department + Ethnicity, maximum age, minimum age, median salary:**

****

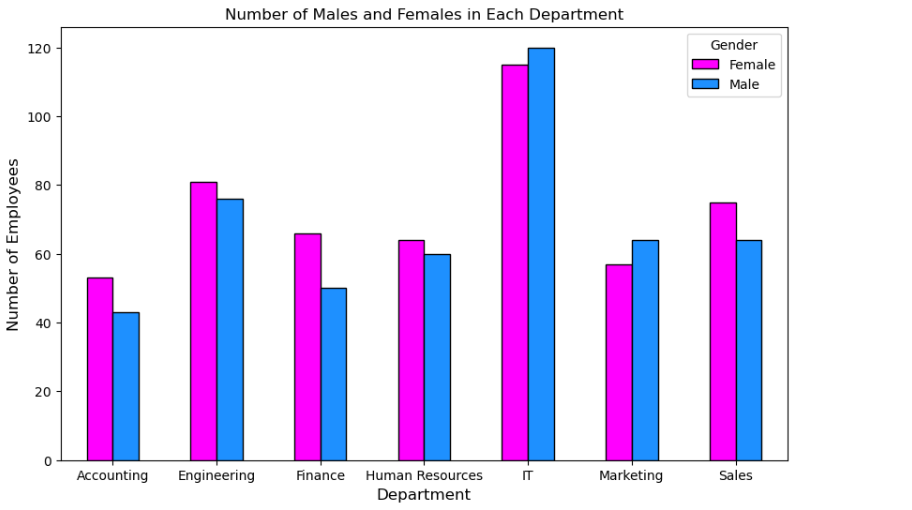
Note: the cleaned and prepared dataset is saved to a new excel file.

**Plotting and Gaining Insights**

1. **Line chart showing the number of employees hired each year**

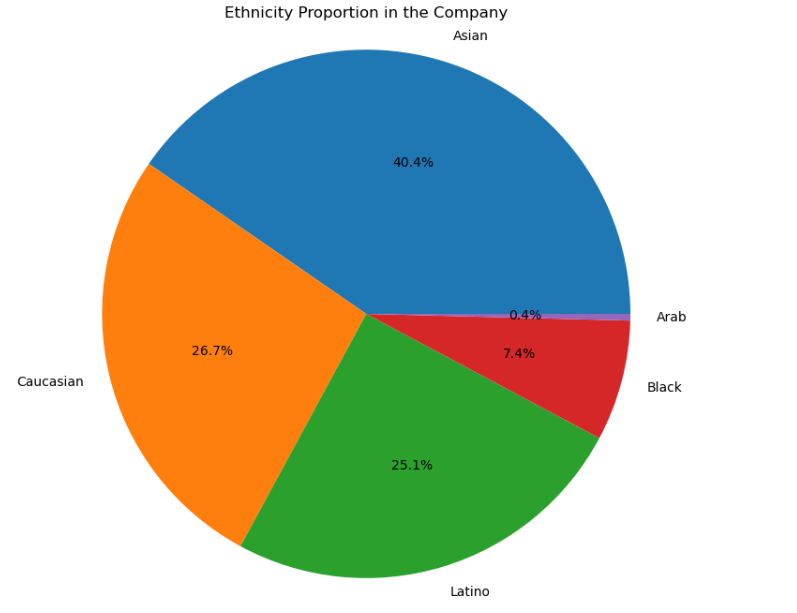
****

The Line chart shows that the number of employees hired has increased each year, with **2021** being the year with the highest number of new hires.

**2- Bar chart showing the number of employees in each department, based on gender**

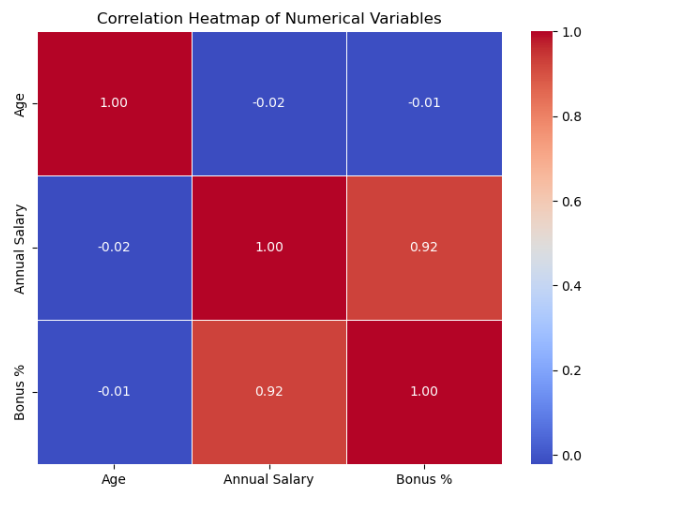
The **bar chart** shows the number of employees in each department, categorized by **gender**. In most departments, the number of **female employees is slightly higher** than males. However, in the **IT** and **Marketing** departments, **male employees slightly outnumber females**. The chart also shows that the **IT department** has had the **highest total number of employees** over the years.

**3- Pie chart showing the ethnicity proportion in the company**

****

The **pie chart** shows the **ethnicity** of employees in the company. **Asian** employees make up the largest group at **40.4%**, followed by **Caucasian** (26.7%) and **Latina** (25.1%). **Black** employees are 7.4%, and **Arab** employees are the smallest group at 0.4%.

**4- Correlation Heatmap of Numeric Variables**

****

The **correlation heatmap** shows the relationships between the numeric columns: **Annual Salary**, **Age**, and **Bonus %**. It shows that **Age** has little to no correlation with **Salary** or **Bonus %**, while **Bonus %** and **Annual Salary** are **strongly correlated.**

**Conclusion and Recommendations:**

The data shows that the **average employee age is 44**, which is relatively high. This suggests an experienced workforce but also highlights the need to attract younger talent.

**Bonus percentages are significantly higher for Vice Presidents** compared to other employees, indicating a gap in compensation distribution.

On a positive note, the company has been **growing steadily over the years**, with a clear increase in hiring. This reflects strong organizational development and expansion.

**Recommendations:**

1. **Attract younger professionals** to balance the age distribution and support long-term growth.
2. **Review bonus policies** to ensure fair and motivating rewards across all job levels.
3. **Promote diversity and gender balance**, especially in underrepresented groups and departments.
4. **Continue supporting growth** through strategic hiring and employee development programs.