|  |  |
| --- | --- |
|  | Data set Exploration Part 2 |
|  |  |
|  |  |

In continuation of Dataset Exploration Part 1, the same dataset, "Employee Attrition," has been chosen for this analysis. This dataset fulfills the requirements with a population of 1471 records and 36 columns, providing an extensive mix of both qualitative and quantitative data types. This makes it an appropriate dataset for our study.

**Data set Introduction:**

The "Employee Attrition" dataset is a synthetic dataset created by IBM for the purpose of simulating HR analytics data. It is vital to highlight that this dataset is entirely fictional and does not represent actual employee data. It contains a variety of employee-related variables, such as demographics, work-related data, compensation-related information, and employee satisfaction metrics.

It described its origin, purpose, and inherent limitations, mainly emphasizing its fictional nature and how this could diverge from real-world employee data. This understanding forms the basis of our analysis.

You can access the dataset on Kaggle using the following reference link: <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

**Data Dictionary:**

Here is a data dictionary for the dataset, including statistical variable types, descriptions, and, where applicable, ranges and limitations:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | **Variable Name** | **Data Type** | **Description** | **Range / Limitations** |
| 0 | Age | Numeric | Age of the employee | Typically ranges from 18 to 65 years. |
| 1 | Attrition | Categorical | Employee attrition (Yes/No) | Binary: 'Yes' for attrition, 'No' for no attrition |
| 2 | BusinessTravel | Categorical | Frequency of business travel | Categorical: 'Travel\_Rarely', 'Travel\_Frequently', 'Non-Travel'. |
| 3 | DailyRate | Numeric | Daily rate of pay | Varies based on the company's pay structure. |
| 4 | Department | Categorical | Department where the employee works | Categorical: 'Sales', 'Research & Development', 'Human Resources' |
| 5 | DistanceFromHome | Numeric | Distance from home to workplace | Measured in miles. |
| 6 | Education | Numeric | Employee's education level | Numeric scale: 1 (Below College) to 5 (Doctorate). |
| 7 | EducationField | Categorical | Field of education of the employee | Categorical: 'Life Sciences', 'Medical', 'Marketing', etc. |
| 9 | Gender | Categorical | Gender of the employee | Categorical: 'Male' or 'Female'. |
| 10 | HourlyRate | Numeric | Hourly rate of pay | Varies based on the company's pay structure. |
| 12 | JobLevel | Numeric | Level of the employee's job | Numeric scale: 1 (Entry Level) to 5 (Manager/Director). |
| 13 | JobRole | Numeric | Role or position of the employee | Categorical: 'Sales Executive', 'Research Scientist', etc. |
| 14 | JobSatisfaction | Numeric | Satisfaction level with the job | Numeric scale: 1 (Low) to 4 (Very High). |
| 15 | MaritalStatus | Categorical | Marital status of the employee | Categorical: 'Single', 'Married', 'Divorced'. |
| 16 | MonthlyIncome | Numeric | Monthly income of the employee | Varies widely based on job roles, experience, etc. |
| 18 | NumCompaniesWorked | Numeric | Number of companies the employee has worked for | Typically ranges from 0 (current company is first) to a higher value. |
| 19 | OverTime | Categorical | Overtime work (Yes/No) | Binary: 'Yes' for overtime, 'No' for no overtime. |
| 20 | PerformanceRating | Numeric | Performance rating of the employee | Numeric scale: 3 (Average) to 4 (Excellent). |
| 21 | MonthlyRate | Numeric | Monthly rate of pay | Varies based on the company's pay structure. |
| 22 | PercentSalaryHike | Numeric | Percentage salary hike in the last year | Measured as a percentage. |
| 23 | RelationshipSatisfaction | Numeric | Satisfaction level with work relationships | Numeric scale: 1 (Low) to 4 (Very High) |
| 24 | StockOptionLevel | Numeric | Level of stock options held by the employee | Numeric scale: 0 (None) to 3 (High). |
| 25 | TotalWorkingYears | Numeric | Total years of work experience | Varies based on employee's career duration |
| 26 | TrainingTimesLastYear | Numeric | Number of training times attended last year | Typically ranges from 0 to a higher value. |
| 27 | WorkLifeBalance | Numeric | Work-life balance satisfaction | Numeric scale: 1 (Bad) to 4 (Very Good). |
| 28 | YearsAtCompany | Numeric | Years at the current company | Varies based on employee's tenure. |
| 29 | YearsInCurrentRole | Numeric | Years in the current job role | Varies based on employee's role stability. |
| 30 | YearsSinceLastPromotion | Numeric | Years since the last promotion | Varies based on promotion history. |
| 31 | YearsWithCurrManager | Numeric | Years with the current manager | Varies based on managerial relationship. |

**Dataset Assumptions, Extra Data Requirements:**

Based on the nature of the dataset, several key assumptions have been identified:

* The dataset is entirely fictitious, and the exact collection process and timing cannot be verified.
* It is unclear whether the dataset is a subset of a larger dataset or represents an entire employee population. This has implications for generalizing our results.
* Measurement units for variables like "Age," "MonthlyIncome," and "HourlyRate" are not explicitly defined.
* Assumptions have been made about missing data: that missing data is not highly skewed and is absent randomly.
* No additional data is required for this analysis, as the provided dataset is sufficient for our goals.

**Univariate Statistics Performed:**

In this phase, I conducted univariate descriptive statistics on a selection of eight variables from the dataset, striking a balance between quantitative and categorical types.

Quantitative Variables:

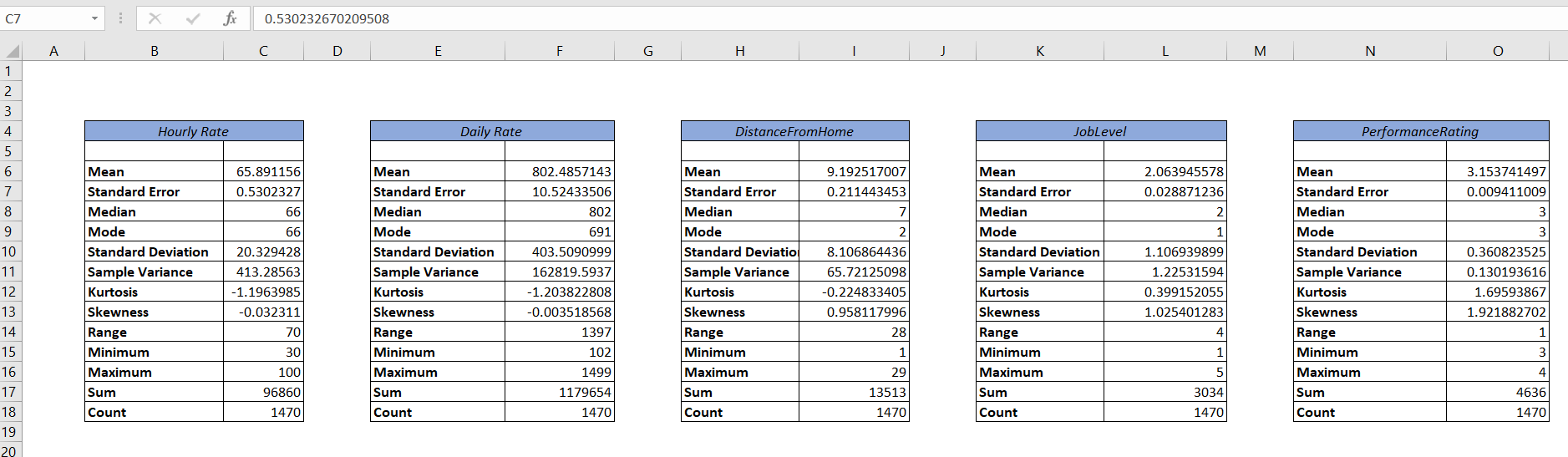
* HourlyRate
* DailyRate
* DistanceFromHome
* JobLevel
* PerformanceRating

Categorical Variables:

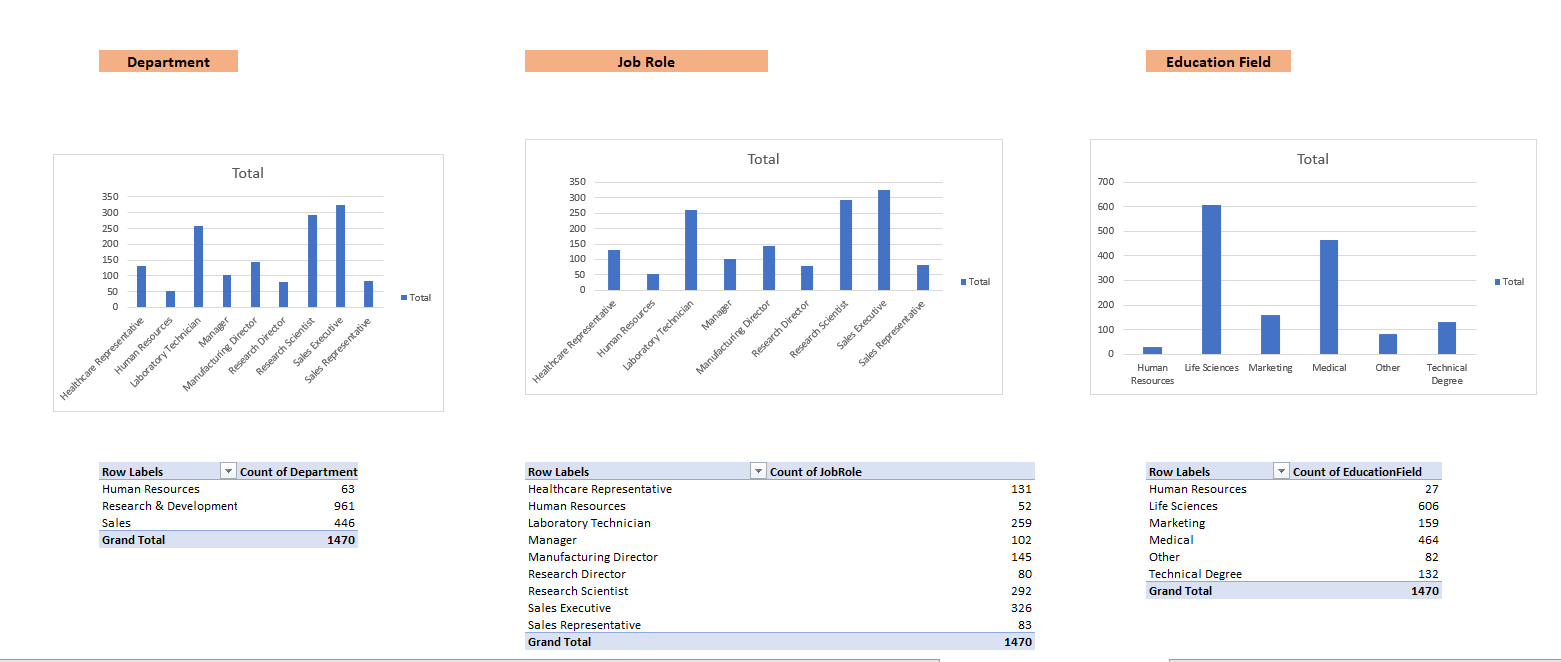
* Department
* JobRole
* EducationField

Here are some screenshots of excel file I have done

* For Quantitative Variables



* For Categorical Variables



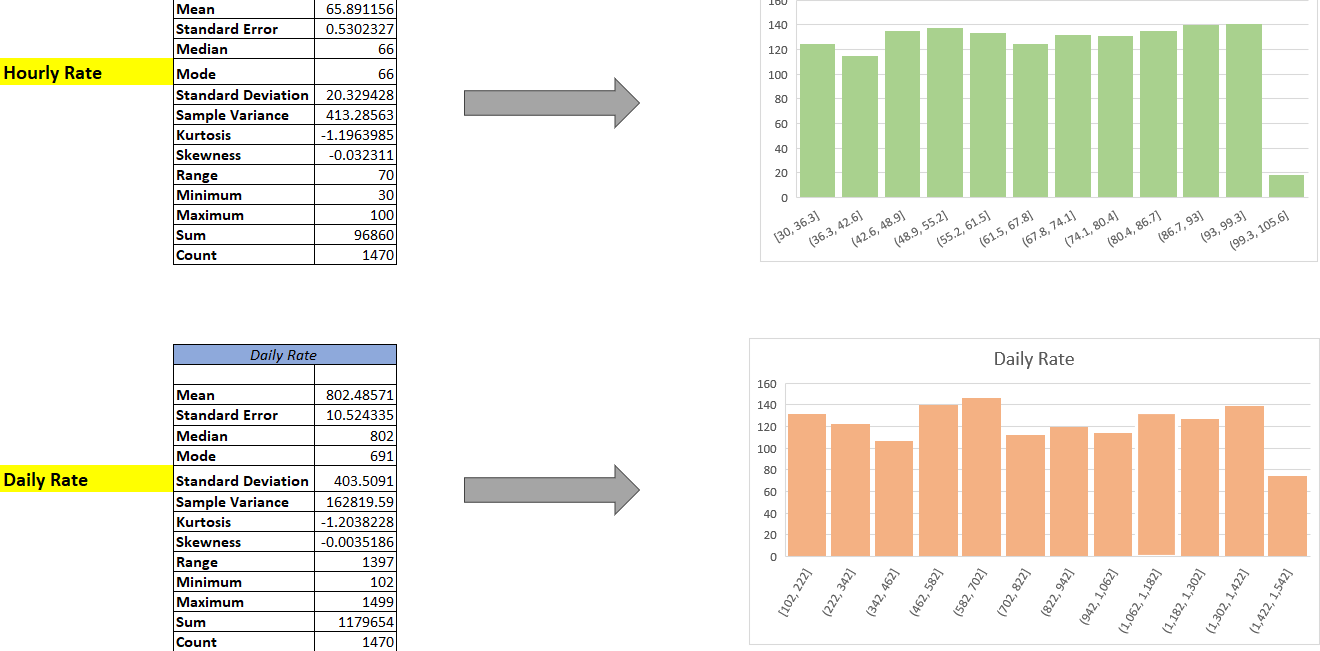
**Univariate Statistics Performed:**

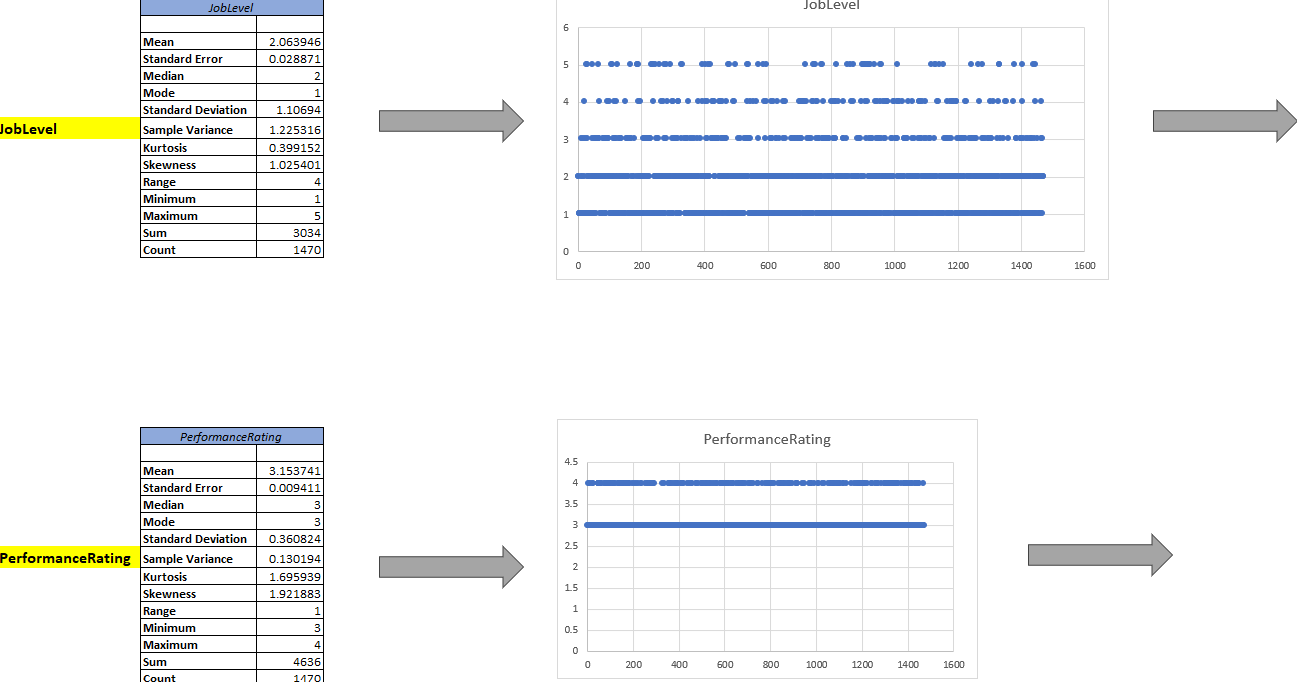
All univariate descriptive statistics have been carefully presented in the report using appropriate visualizations, including tables, histograms, and scatter plots. The summary of statistics is depicted in the following tables and figures:

Include tables, histograms, and scatter plots displaying the univariate statistics for each of the chosen variables, ensuring that the presentation is clear and informative.

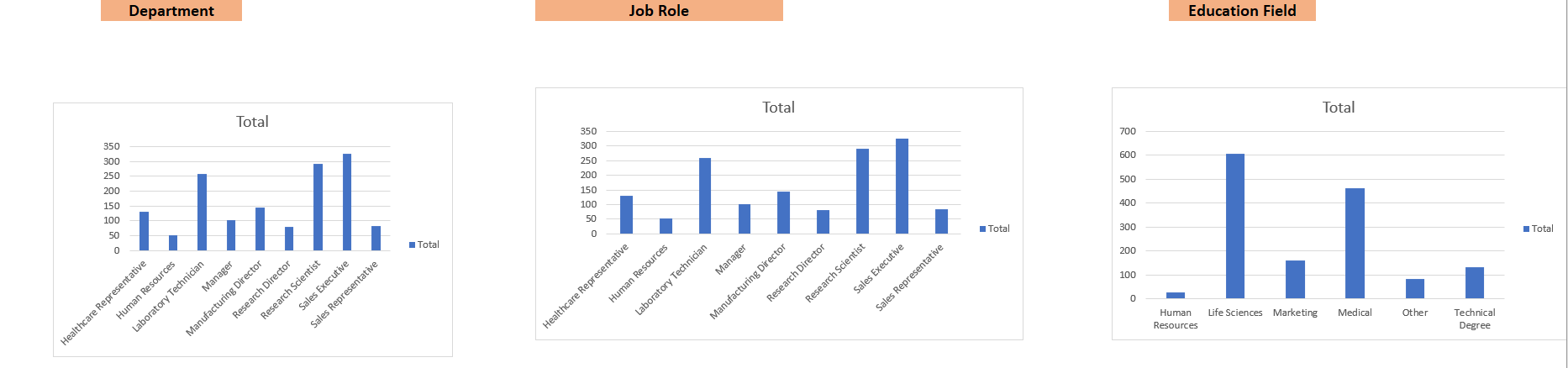
Here are some screenshots of excel file I have done

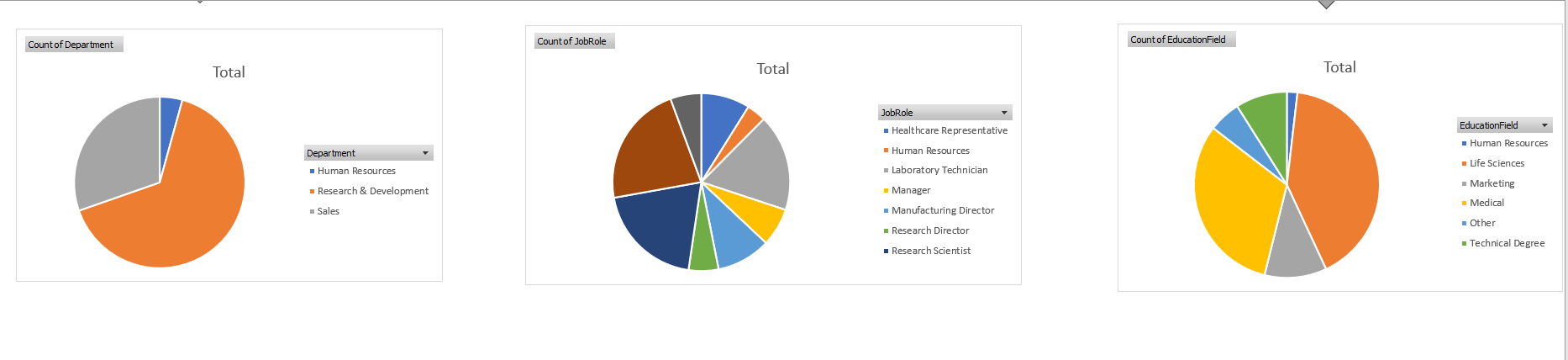
* For Quantitative Variables





* For Categorical Variables





**Suggested Outliers Identified:**

During the analysis, I reviewed the quantitative variables for potential outliers. However, it's noteworthy that no significant outliers were identified within the dataset. While some variables showed slight skewness, no values were deemed as outliers requiring exclusion from the analysis.

This mean my dataset provided by IBM is too ideal.

Here are some screenshots of excel file I have done to check outlier



**Cleaning Decisions Explained:**

a. Handling Missing Data:

Missing values were addressed in Dataset Exploration Part 1, where we removed rows with missing values in key variables, ensuring that our analysis is based on complete data.

b. Handling Invalid Data:

The dataset was subjected to a thorough review to identify any anomalies or data entries that violate the data dictionary. Decisions were made to correct, remove, or investigate records that did not conform to the dataset's specifications. These decisions were documented, ensuring transparency in data handling.

**Coding or Categorization Performed:**

Categorization was applied to the "Age" variable. I divided Age it into three categories: "Young," "Middle-aged," and "Senior," based on defined age ranges. A new column, "Age Class," was created to reflect these categories.

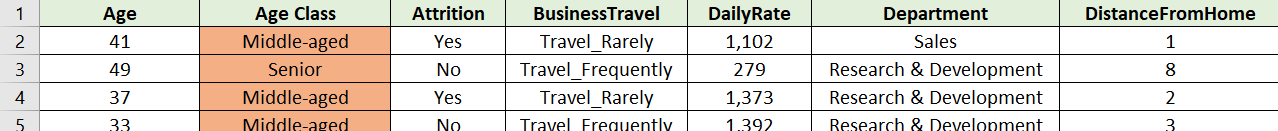
Based on my project I can create categories like "Young," "Middle-aged," and "Senior."

* Create a new column, e.g., "Age Class."
* Use Excel's IF statements to categorize the data =IF(A2<=30, "Young", IF(A2<=45, "Middle-aged", "Senior"))
* Apply this formula to each row in my dataset, replacing "A2" with the appropriate cell for the age value.

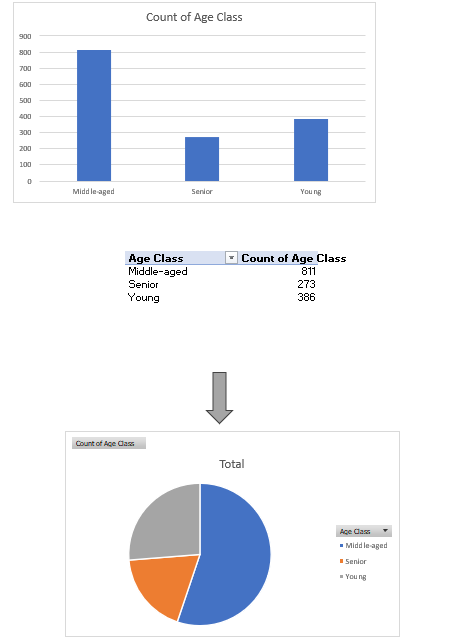
**Helpful Coding or Categorization:**

This coding adds value to the analysis by providing a clearer perspective on the age distribution within the dataset, which will be beneficial for answering research questions.

After creating this new column in my dataset



I apply unriveted analysis and present in pie chart to know the contribution employee attrition



**FINER Questions**

* What is the gender-based age distribution, and are there any notable disparities between males and females in terms of age?
* What is the gender-based salary distribution, and can you provide the gender-specific employee headcount for each department?
* Is there a connection between an employee's job level and their performance rating, and does this connection differ among various job roles?
* Is there a significant difference in monthly income (MonthlyIncome) between employees who work overtime (OverTime = Yes) and those who do not (OverTime = No)?

**Tracking**

This analysis process has been diligently tracked. I have maintained detailed records of our activities, including data manipulations, decision-making steps, and thought processes. Assumptions and decisions regarding data cleaning, categorization, and outlier handling have been documented to ensure transparency and replicability. Potential next steps in our analysis have also been noted for reference in future stages.

This report summarizes the actions taken in Dataset Exploration Part 2, ensuring that it adheres to the provided evaluation criteria. It lays the groundwork for further analysis as we progress through the project.

The Excel spreadsheet containing the dataset and additional details is attached.