Python Project - Data Analysis

Overview

This project aims to explore and analyze an employee dataset to uncover key trends, patterns, and correlations. The analysis covers various aspects, such as the distribution of employees across teams and positions, salary expenditure by team and position, and the correlation between age and salary. Graphical representations are used throughout to visualize the findings and make the results more comprehensible.

Project Structure

1. Data Preprocessing

Data preprocessing is the first crucial step in the analysis, ensuring the dataset is clean and ready for analysis. The preprocessing steps included:

- **Data Cleaning**: Ensured there were no missing or null values in the dataset.
- **Data Type Conversion**: Converted columns to appropriate data types (e.g., Age and Salary columns were ensured to be in numerical format).
- **Duplicate Removal**: Removed any duplicate rows that might affect the analysis.
- **Handling Inconsistent Data**: Addressed any inconsistencies or anomalies in the data such as removing invalid characters or formatting issues.

---

2. Analysis Tasks

After preprocessing, the following key analysis tasks were performed:

a. Distribution of Employees Across Teams

We grouped employees by their respective teams and calculated the total number of employees in each team. This helped in understanding how the workforce is spread across different teams and identifying any discrepancies or trends in the team sizes.

b. Position-wise Employee Distribution

Employees were grouped by their positions within the company. This task allowed us to see how many employees are assigned to each position and how the workforce is structured in terms of roles and responsibilities.

c. Salary Expenditure by Team and Position

We calculated the total salary expenditure for each team and each position. This analysis highlighted which teams and positions have the highest salary expenses, offering insights into the cost distribution within the organization.

- **Top 5 Teams with Highest Salary Expenditure**: The teams with the highest salary expenditures were identified, including the Cleveland Cavaliers, Los Angeles Clippers, and others.
- **Top 5 Positions with Highest Salary Expenditure**: The positions with the highest salary expenditures were also identified, shedding light on roles that command the highest salaries.

d. Correlation Between Age and Salary

We analyzed the relationship between age and salary to determine whether older employees tend to earn higher salaries. The correlation coefficient was calculated, and a scatter plot was created to visualize this relationship. The correlation was found to be weak, suggesting that while age may play a role in salary, other factors like position, experience, and performance are more influential.

---

3. Graphical Representations

To visualize the findings, various graphical representations were used:

- **Bar Plots**: Used to display the top 5 teams and positions with the highest salary expenditures. This allowed for easy comparison between different teams and positions.
- **Scatter Plot**: Created to show the relationship between age and salary, highlighting the weak positive correlation between the two variables.
- **Pie Charts**: Represented the distribution of employees across teams and positions, providing a clear visual of how the workforce is allocated.

---

4. Insights Gained

From the analysis, several key insights were uncovered:

- **Team Size and Salary Expenditure**: Some teams have a significantly larger salary expenditure, which could indicate their importance in the organization. Teams with higher salaries likely have more experienced or highly skilled members.
- **Position-wise Salary Distribution**: Senior roles or higher-level positions tend to have higher salary expenditures. These positions typically require more experience and specialized skills, contributing to their higher pay.
- **Age and Salary Correlation**: While there is a weak positive correlation between age and salary, other factors, such as experience, job performance, and team contributions, are likely more significant in determining salary. Age alone does not strongly predict salary.

---

5. Additional Information

- **Tools Used**: The analysis was performed using Python, with libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, and NumPy for statistical operations.
- **Future Enhancements**: Future work could involve analyzing the correlation between salary and other variables such as education level, years of experience, or performance metrics.

Data Story and Insights:

This dataset provides a comprehensive overview of the employees in a company (in this case, presumably an NBA team roster), including details like Name, Team, Position, Age, Height, Weight, College, and Salary. The following steps were taken to analyze the data, uncover insights, and make important business decisions:

---

1. Data Cleaning and Preprocessing:

- **Missing Values**:
    - The dataset initially had missing values in the "College" and "Salary" columns.
    - For the "College" column, the mode (most frequent value) was used to fill the missing values, which was 'Kentucky'.
    - For the "Salary" column, the mean salary value was used to fill in the missing entries, which was approximately $4.83 million.
    - After handling the missing data, there were no more missing values, ensuring that the data was clean for further analysis.
- **Unrealistic Values**:
    - No unrealistic values were found in the dataset (such as negative age, height, or weight).
- **Data Types**:
    - The dataset contains a mix of integer, object (string), and float data types, and the 'Height' column was updated from an object to an integer type to facilitate calculations.

---

2. Distribution of Employees Across Teams:

We examined the distribution of employees (presumably players) across different teams. The results showed the following insights:

- **Team Size Distribution**:
    - The team with the most employees is the **New Orleans Pelicans**, which has 19 employees (4.15% of the total).
    - The team with the least employees is **Minnesota Timberwolves** and **Orlando Magic**, both with 14 employees (3.06% of the total).

The teams are fairly evenly distributed, with the majority having 15 or more employees, and this ensures a broad player roster.

Visualization: A bar chart was created to visually show the distribution of employees across teams.

---

3. Segregation of Employees Based on Positions:

Next, we grouped the employees by their positions. The insights derived from the distribution of positions are:

- **Position Count Distribution**:
    - The **Shooting Guard (SG)** position has the most employees, with 102 employees (22.27%).
    - The **Power Forward (PF)** position follows with 100 employees (21.83%).
    - The **Center (C)** position has 79 employees (17.25%).
    - The **Point Guard (PG)** and **Small Forward (SF)** positions have 92 (20.09%) and 85 (18.56%) employees, respectively.

This suggests that the **SG** and **PF** positions are more predominant, while the **C** position is less populated.

Visualization: A bar chart was used to show the count of employees by position, helping in understanding the position distribution within the company.

---

4. Age Group Distribution:

We analyzed the age distribution among employees, classifying them into age groups. The predominant age group was identified as follows:

- **Age Group Distribution**:
    - The **20-29** age group is the largest, with 346 employees (75.5% of the total).
    - The **30-39** age group has 91 employees (19.9%).
    - No employees fall in the **40-49**, **50-59**, or **60+** age ranges.

This indicates that the workforce is primarily composed of younger employees, most likely in the early stages of their careers.

Visualization: A bar chart was plotted to show the distribution of employees across age groups, highlighting the prominence of the **20-29** group.

---

5. Highest Salary Expenditure by Team and Position:

The next step was to discover which team and position had the highest salary expenditures. Here's what we found:

- **Team Salary Expenditure**:
    - The **Cleveland Cavaliers** has the highest salary expenditure, with a total of **$106,988,689**.
    - Other teams like **Charlotte Hornets**, **Chicago Bulls**, and **Atlanta Hawks** also have significant expenditures.

- **Position Salary Expenditure**:
    - The **Point Guards (PG)** and **Shooting Guards (SG)** positions, which have a higher number of employees, generally account for higher total salary expenditures.

This information could help in assessing which teams or positions have a higher budget allocation and whether this correlates with performance metrics.

Correlation between Age and Salary

We investigated the relationship between age and salary in the dataset, visualizing it with a scatter plot and a regression line. The correlation coefficient calculated is 0.21, which suggests a **weak positive correlation**. This means that, while salary slightly increases as age increases, the relationship is not strong.

Insights:

- **Weak Positive Correlation (0.21):** Age and salary are not strongly related in this dataset. While there is a slight upward trend in salary with age, it is not significant enough to suggest a consistent pattern.
- **Visualization:** The scatter plot, complemented with a regression line, helps visualize the mild relationship between the two variables.

Conclusion:

Through the exploration of this dataset, we uncovered a wealth of valuable insights:

1. **Employee Distribution**: Teams have a relatively balanced number of employees, with the **New Orleans Pelicans** having the most.
2. **Position Insights**: **Shooting Guards** and **Power Forwards** are the most common positions, while **Centers** have fewer players.
3. **Age Distribution**: The company is predominantly made up of younger employees, especially in the **20-29** age group.
4. **Salary Insights**: The **Cleveland Cavaliers** have the highest salary expenditure, and positions like **Point Guard (PG)** and **Shooting Guard (SG)** tend to have higher salary allocations.

   By visualizing these insights, decision-makers can have a clearer understanding of team compositions, budget allocations, and workforce distribution across positions and age groups.

Data Preprocessing Steps

In this project, we applied several data preprocessing steps to ensure data quality and prepare the dataset for analysis. These steps include:

1. **Handling Missing Values**
    - Identified and resolved missing values within the dataset to maintain data integrity.
    - Used suitable imputation techniques or removed entries based on the context and requirements of the data.
2. **Correcting Data Types**
    - Ensured each column had the correct data type. For instance, numerical columns such as 'Age' and 'Salary' were converted to appropriate numeric types (integers or floats).
    - Categorical columns were converted to category types to enhance performance and facilitate analysis.
3. **Height Standardization**
    - Verified and standardized the 'Height' column values to ensure a uniform measurement unit across the dataset.
    - Addressed any inconsistencies in the data entries, resulting in accurate and consistent height measurements.
4. **Data Consistency Checks**
    - Conducted thorough checks to identify and correct inconsistencies across columns, ensuring uniform data entry and eliminating discrepancies.
5. **Outlier Detection and Removal**
    - Analyzed numerical columns, particularly 'Salary,' to identify outliers that could distort analysis.
    - Removed or adjusted outliers based on statistical analysis, thus improving the dataset's representativeness.

These preprocessing steps established a robust, clean dataset ready for detailed analysis and insights.