Salary Insights: Exploratory Data Analysis Report

Aim

The analysis was to explore and analyze the salary dataset. The dataset includes features such as age, gender, education level, job title, years of experience, race, and country. The primary objective of this analysis was to understand how different factors influence salary and whether there are any significant differences in salaries based on different groups.

Data Exploration and Preparation

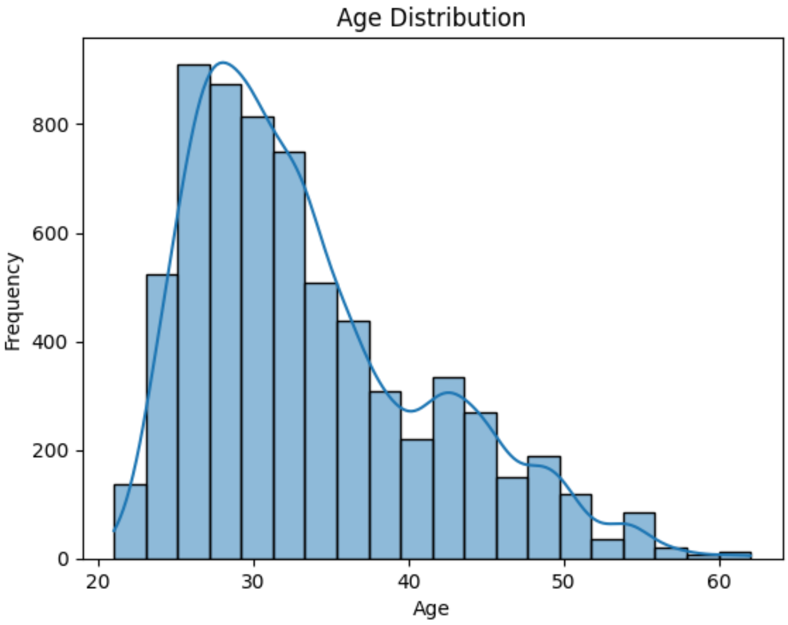
The dataset was first explored to understand its structure and characteristics. Missing values were handled appropriately, either by imputing mean or by dropping rows with missing data, depending on the context.

Data Visualization

Libraries Used: pandas, matplotlib, seaborn, scipy, plotly

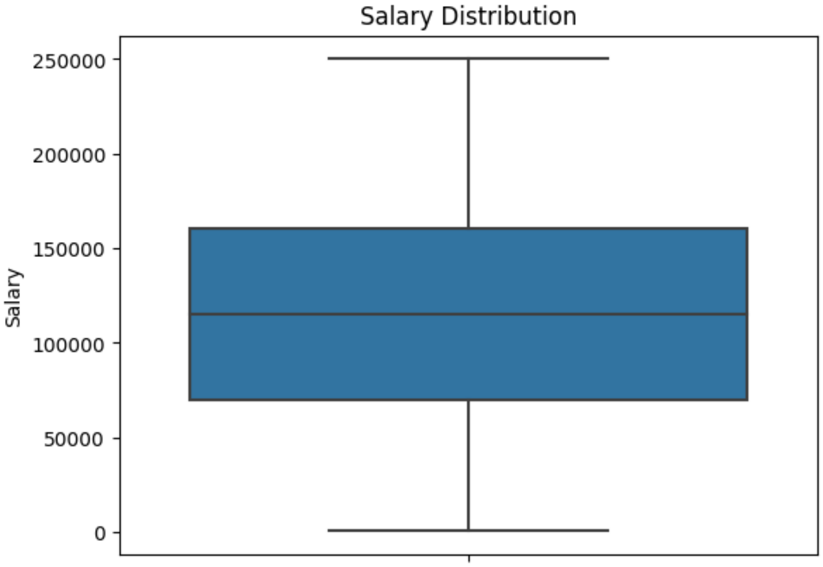
Plots and Interpretation

Histogram showing Age Distribution



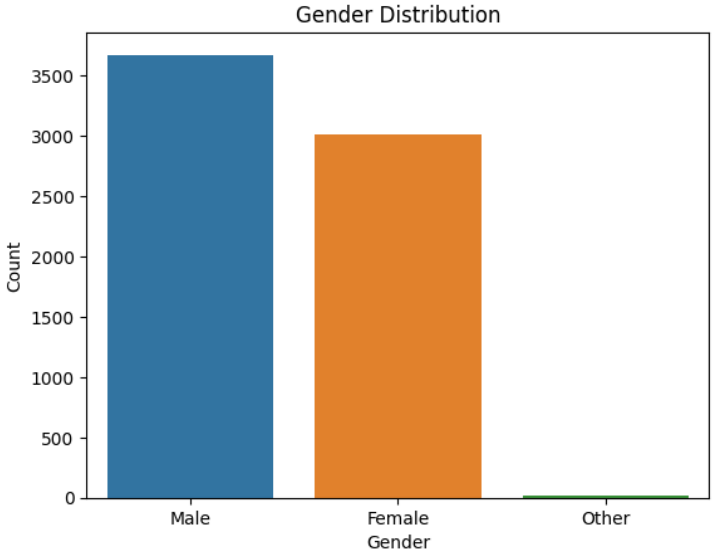
- Dependent Variable: Frequency (the number of occurrences or counts of each age group)
- Independent Variable: Age (the variable being observed and grouped into bins)
- The number of bins is 20, which means it is divided into 20 intervals or groups to display the distribution effectively
- Each bin represents a specific age range, and the height of each bar in the histogram represents the frequency of ages falling within that range.
- Within the dataset, a significant number of individuals fall within the age range of 22 to 35 years.

Boxplot showing Salary Distribution



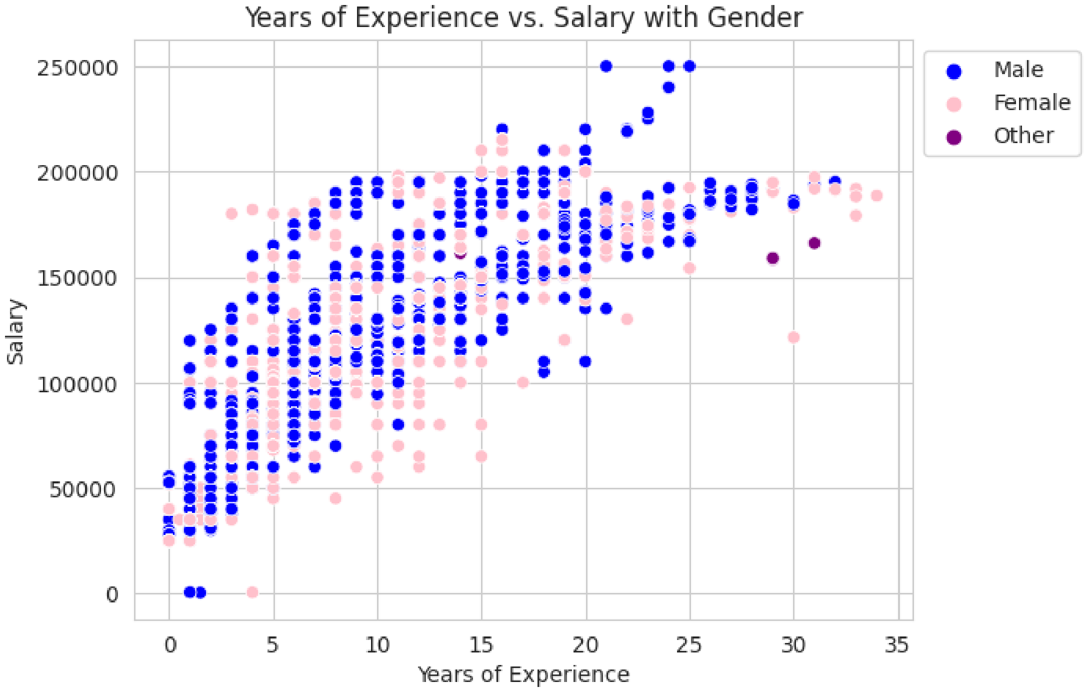
- The median salary falls between 100,000 to 150,000
- The box plot is skewed toward the lower values, indicating that the median (middle line inside the box) is closer to the lower end of the data distribution.
- The box's width being very wide indicates that the middle 50% of the data (IQR) is spread over a broader range of values. The data distribution shows a significant spread within the middle 50% of the data.
- The long whiskers which extend to min and max represent a great range for the sample. It is an indicator of highly probable values. It suggests that the data has a substantial range and includes outliers.

Countplot showing Gender Distribution



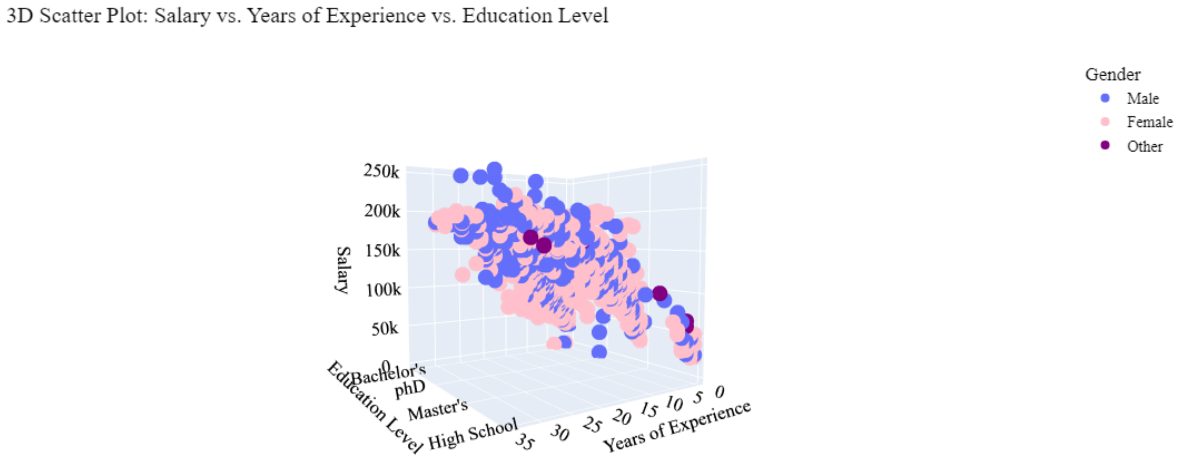
- The x-axis represents the categories of the categorical variable, i.e. Gender. The categories are Female, Male, and Other,
- The y-axis represents the count or frequency of each category.
- The highest frequency is of Male, followed by Female. The sample consists of very few data entries of Other.

Scatterplot showing Years of Experience influencing Salary for Different Genders



- Years of experience is on the X-axis and Salary earned is on the Y-axis. The gender is the color of the points. The blue dots indicate male, pink indicate female, and purple represent others.
- From the scatterplot, one can infer that in most cases, as the years of experience in any profession increases, the salary earned also increases. This is true for all genders in the dataset.
- This was plotted using matplotlib and seaborn.

3D Scatterplot showing Education Level, Years of Experience and Gender’s influence on Salary



- The education levels are High School, Bachelor’s Degree, Master’s Degree, and phD.
- Years of Experience ranges from 0 to 35 years.
- Gender is the color of the points, male being blue, female pink, and others purple.
- The plot suggests that the highest levels of salaries are earned by individuals with a higher degree, i.e. phD or Master’s, and most years of experience. This is true for all three categories of genders represented.
- This plot was created using plotly.

Statistical Testing

Welch’s T-Tests

1. T-test to compare the salaries between male and female employees

- Grouped the data by gender to compute the mean and standard deviation of salaries for men and women separately.
- Imputed missing values in the "Salary" column with the overall mean salary.
- Used Welch's T-test, as the standard deviations of salaries for men and women were different.
- Results:
    - The T-test results showed a t-statistic of 0, suggesting weak evidence against the null hypothesis of similar salaries for both genders.
    - The p-value obtained was 1, indicating that any observed difference in salaries is likely due to random chance.
    - Thus, the null hypothesis cannot be rejected.

2. T-test to compare salaries between employees with a Master’s Degree and a phD Degree

- The mean salary for employees with Ph.D. degrees (mean_phd) was calculated, as well as the mean salary for employees with Master's degrees (mean_mast).
- The calculated difference (diff) between the mean salaries of the two groups shows the extent of the salary disparity between Ph.D. and Master's degree holders.
- Results:
    - The T-test resulted in a t-statistic of 26.19643266234008, indicating a significant difference in salaries between the two groups.
    - The p-value obtained was 2.5827013430995746e-137, which is very close to zero, suggesting strong evidence against the null hypothesis of equal salaries for Ph.D. and Master's degree holders.
    - Therefore, the null hypothesis is rejected.

One-Way ANOVA Tests

1. Country
    - The variance of salaries for each country group was calculated as ANOVA tests depend on variance to assess differences.
    - The dataset was grouped based on the "Country" column, and arrays of salaries for each country (Australia, Canada, China, UK, and USA) were created.
    - The ANOVA test was performed using the f_oneway() function, comparing the salaries among the different countries' groups.
    - Results:
        - The ANOVA test resulted in an F-statistic of 1.0151015795355185. A smaller F-statistic suggests that the means of the groups are similar.
        - The p-value obtained was 0.3979757356026538, which is relatively high. The high p-value indicates that there is a significant probability of observing the observed difference in means if there were no real difference between the countries.
        - The interesting result was that the null hypothesis cannot be rejected, indicating that there is no statistically significant difference in salaries among the tested countries.

2. Years of Experience
    - The "Years of Experience" data was binned into different ranges (0-5, 5-10, 10-15, 15-20, 20-25, 25-30) using the pd.cut() function, and a new column named 'Experience Range' was created to store the bin labels.
    - The ANOVA test was performed using the f_oneway() function to compare the salaries among the different groups based on years of experience.
    - Results:
        - The ANOVA test resulted in an F-statistic of 2892.665336919322, indicating a high value for the statistic. A high F-statistic suggests that there is more difference between the group means than what would be expected by random chance alone.

        - The p-value obtained was 0.0, which is very close to zero. The low p-value suggests strong evidence against the null hypothesis of no difference in salaries among the experience groups.
        - Based on these results, the null hypothesis is rejected, indicating that there is a statistically significant difference in salaries among the different years of experience groups.

3. Levels of Education
    - The dataset was grouped based on the 'Education Level' column, and arrays of salaries for each education level group were created.
    - Groups with different names but the same degree, such as phD and PhD or Master’s and Master’s Degree, were merged to obtain the Salary dataset.
    - The ANOVA test was performed using the f_oneway() function to compare the salaries among the different education level groups.
    - Results:
        - The ANOVA test resulted in an F-statistic of 1630.9393525726196, indicating a high value for the statistic. A high F-statistic suggests that there is more difference between the group means than what would be expected by random chance alone.
        - The p-value obtained was 0.0, which is very close to zero. The low p-value suggests strong evidence against the null hypothesis of no difference in salaries among the education level groups.
        - Thus, the null hypothesis is rejected, indicating that there is a statistically significant difference in salaries among the different education levels.

4. Race

- The dataset was then grouped based on the 'Race' column, and arrays of salaries for each race group were created.
- The variance of salaries for each race group was calculated as part of the ANOVA test.
- Race groups that were compared : "Asian," "Australian," "Chinese," "Hispanic," "Korean," "Mixed," "Welsh," and "White."

- Results:
    - The ANOVA test resulted in an F-statistic of 1.1907321123205628, which indicates a relatively low value for the statistic.
    - The p-value obtained was 0.3000853257762943, which is not very low. The higher p-value suggests that the probability of observing the observed difference in means if there were no real difference between the race groups is relatively high.
    - Based on these results, you cannot reject the null hypothesis, and there is no statistically significant difference in salaries among the different race groups.