KARNAVATI UNIVERSITY

UNITED WORLD SCHOOL OF COMPUTATIONAL INTELLIGENCE (USCI)

Summative Assessment (SA)

Submitted BY

KHUSHI PAGARIA

(Enrl. No.: 20220701059)

B.Sc. (Hons.) Computer Science / Data Science / AIML

III Semester – July – Nov 2023

USCI

Nov/Dec 2023

INDEX

Introduction

Aim of the Project

An R programming project aims to leverage the capabilities of the R programming language for statistical computing and data analysis. This includes tasks such as data exploration, statistical modeling, data visualization, automation, report generation, package development, integration with other tools, data cleaning, preprocessing, optimization, and collaboration. The goal is to use R as a powerful tool to analyze and derive meaningful insights from data in various domains.

Intended Outcome

Data Analysis and Exploration:

- Perform descriptive statistics to summarize and explore the main characteristics of a dataset.
- Identify patterns, trends, and outliers within the data.
- Conduct inferential statistics to make predictions or draw conclusions about a population based on a sample.

Statistical Modeling:

- Develop and implement statistical models to understand relationships within the data.
- Apply machine learning algorithms for predictive modeling or classification tasks.

Data Visualization:

- Create informative and visually appealing plots and charts to communicate data insights.
- Use packages like ggplot2 to generate static or interactive visualizations

Report Generation:

- Generate reports or documents summarizing the findings, analyses, and visualizations.
- Utilize tools like R Markdown to create dynamic reports that integrate code, results, and narrative.

Package Development:

- Contribute to or create R packages to extend the functionality of R for specific tasks.

- Package development can involve creating functions, documentation, and unit tests.

Integration with Other Tools:

- Integrate R with other tools and platforms for broader data analysis and reporting.
- Use R in conjunction with databases, web applications, or other programming languages.

Collaboration:

- Facilitate collaboration by using version control systems (e.g., Git) and collaborating on projects with others.

Dataset

Sr no.	age	Gender	Education Level	Job Title	Years of Experience	Salary
1	32	Male	Bachelor's	Software Engineer	5	90000
2	28	Female	Master's	Data Analyst	3	65000
3	45	Male	PhD	Senior Manager	15	150000
4	36	Female	Bachelor's	Sales Associate	7	60000
5	52	Male	Master's	Director	20	200000
6	29	Male	Bachelor's	Marketing Analyst	2	55000
7	42	Female	Master's	Product Manager	12	120000
8	31	Male	Bachelor's	Sales Manager	4	80000
9	26	Female	Bachelor's	Marketing Coordinator	1	45000
10	38	Male	PhD	Senior Scientist	10	110000
11	29	Male	Master's	Software Developer	3	75000
12	48	Female	Bachelor's	HR Manager	18	140000
13	35	Male	Bachelor's	Financial Analyst	6	65000
14	40	Female	Master's	Project Manager	14	130000
15	27	Male	Bachelor's	Customer Service Rep	2	40000
16	44	Male	Bachelor's	Operations Manager	16	125000
17	33	Female	Master's	Marketing Manager	7	90000
18	39	Male	PhD	Senior Engineer	12	115000
19	25	Female	Bachelor's	Data Entry Clerk	0	35000
20	51	Male	Bachelor's	Sales Director	22	180000

21	34	Female	Master's	Business Analyst	5	80000
22	47	Male	Master's	VP of Operations	19	190000
23	30	Male	Bachelor's	IT Support	2	50000
24	36	Female	Bachelor's	Recruiter	9	60000
25	41	Male	Master's	Financial Manager	13	140000
26	28	Female	Bachelor's	Social Media Specialist	3	45000
27	37	Female	Master's	Software Manager	11	110000
28	24	Male	Bachelor's	Junior Developer	1	40000
29	43	Female	PhD	Senior Consultant	15	140000
30	33	Male	Master's	Product Designer	6	90000
31	50	Male	Bachelor's	CEO	25	250000
32	31	Female	Bachelor's	Accountant	4	55000
33	29	Male	Master's	Data Scientist	3	75000
34	39	Female	Bachelor's	Marketing Specialist	10	65000
35	46	Male	PhD	Senior Manager	20	170000
36	27	Male	Bachelor's	Technical Writer	2	45000
37	35	Female	Bachelor's	HR Generalist	7	60000
38	42	Male	Master's	Project Engineer	14	115000
39	26	Female	Bachelor's	Customer Success Rep	1	40000
40	49	Male	Bachelor's	Sales Executive	21	160000
41	34	Female	Master's	UX Designer	5	80000
42	48	Male	Master's	Operations Director	18	190000
43	30	Male	Bachelor's	Network Engineer	3	60000
44	36	Female	Bachelor's	Administrative Assistant	8	45000
45	41	Male	Master's	Strategy Consultant	13	130000
46	28	Female	Bachelor's	Copywriter	2	40000
47	32	Male	Bachelor's	Account Manager	5	75000
48	45	Female	Master's	Director of Marketing	16	180000
49	38	Male	PhD	Senior Scientist	11	120000
50	25	Male	Bachelor's	Help Desk Analyst	0	35000
51	51	Female	Bachelor's	Customer Service Manager	22	130000
52	33	Male	Master's	Business Intelligence Analyst	7	85000
53	40	Female	Bachelor's	Event Coordinator	12	60000
54	47	Male	Master's	VP of Finance	19	200000
55	29	Female	Bachelor's	Graphic Designer	3	50000
56	36	Male	Bachelor's	Sales Manager	9	95000
57	27	Female	Master's	UX Researcher	2	65000
58	43	Male	PhD	Senior Engineer	17	140000
59	30	Female	Bachelor's	Social Media Manager	4	55000
60	35	Male	Master's	Product Manager	7	105000

Dataset Description

This dataset contains information about the salaries of employees at a company. Each row represents a different employee, and the columns include information such as age, gender, education level, job title, years of experience, and salary.

Columns:

Age: This column represents the age of each employee in years. The values in this column are numeric.

Gender: This column contains the gender of each employee, which can be either male or female. The values in this column are categorical.

Education Level: This column contains the educational level of each employee, which can be high school, bachelor's degree, master's degree, or PhD. The values in this column are categorical.

Job Title: This column contains the job title of each employee. The job titles can vary depending on the company and may include positions such as manager, analyst, engineer, or administrator. The values in this column are categorical.

Years of Experience: This column represents the number of years of work experience of each employee. The values in this column are numeric.

Salary: This column represents the annual salary of each employee in US dollars. The values in this column are numeric and can vary depending on factors such as job title, years of experience, and education level.

This dataset appears to be a tabular dataset containing information about individuals, possibly related to their professional and employment characteristics. Each row represents a unique individual, and the columns provide various attributes such as age, gender, education level, job title, years of experience, and salary. The dataset seems to include a diverse range of roles and positions across different industries, with information about both technical and non-technical job titles. The dataset may be useful for exploring relationships between factors like education, experience, and salary, or for conducting analyses related to workforce demographics and compensation.

STATISTICAL ANALYSIS

INPUT

```
df=read.csv("C:/Users/HP/Desktop/R project work.csv")

print(df)
```

OUTPUT

```
> df=read.csv("C:/Users/HP/Desktop/R project work.csv")
>
> print(df)
   Sr.no.  age Gender Education.Level                Job.Title Years.of.Experience Salary
1       1   32   Male      Bachelor's     Software Engineer                  5  90000
2       2   28 Female        Master's          Data Analyst                  3  65000
3       3   45   Male             PhD        Senior Manager                 15 150000
4       4   36 Female      Bachelor's       Sales Associate                  7  60000
5       5   52   Male        Master's              Director                 20 200000
6       6   29   Male      Bachelor's     Marketing Analyst                  2  55000
7       7   42 Female        Master's       Product Manager                 12 120000
8       8   31   Male      Bachelor's         Sales Manager                  4  80000
9       9   26 Female      Bachelor's  Marketing Coordinator                  1  45000
10     10   38   Male             PhD      Senior Scientist                 10 110000
11     11   29   Male        Master's     Software Developer                  3  75000
12     12   48 Female      Bachelor's            HR Manager                 18 140000
13     13   35   Male      Bachelor's     Financial Analyst                  6  65000
14     14   40 Female        Master's       Project Manager                 14 130000
15     15   27   Male      Bachelor's  Customer Service Rep                  2  40000
16     16   44   Male      Bachelor's    Operations Manager                 16 125000
17     17   33 Female        Master's     Marketing Manager                  7  90000
18     18   39   Male             PhD       Senior Engineer                 12 115000
19     19   25 Female      Bachelor's      Data Entry Clerk                  0  35000
20     20   51   Male      Bachelor's        Sales Director                 22 180000
21     21   34 Female        Master's      Business Analyst                  5  80000
22     22   47   Male        Master's      VP of Operations                 19 190000
23     23   30   Male      Bachelor's            IT Support                  2  50000
24     24   36 Female      Bachelor's             Recruiter                  9  60000
25     25   41   Male        Master's     Financial Manager                 13 140000
26     26   28 Female      Bachelor's Social Media Specialist                  3  45000
27     27   37 Female        Master's      Software Manager                 11 110000
28     28   24   Male      Bachelor's       Junior Developer                  1  40000
```

29	29	43	Female	PhD	Senior Consultant	15	140000
30	30	33	Male	Master's	Product Designer	6	90000
31	31	50	Male	Bachelor's	CEO	25	250000
32	32	31	Female	Bachelor's	Accountant	4	55000
33	33	29	Male	Master's	Data Scientist	3	75000
34	34	39	Female	Bachelor's	Marketing Specialist	10	65000
35	35	46	Male	PhD	Senior Manager	20	170000
36	36	27	Male	Bachelor's	Technical Writer	2	45000
37	37	35	Female	Bachelor's	HR Generalist	7	60000
38	38	42	Male	Master's	Project Engineer	14	115000
39	39	26	Female	Bachelor's	Customer Success Rep	1	40000
40	40	49	Male	Bachelor's	Sales Executive	21	160000
41	41	34	Female	Master's	UX Designer	5	80000
42	42	48	Male	Master's	Operations Director	18	190000
43	43	30	Male	Bachelor's	Network Engineer	3	60000
44	44	36	Female	Bachelor's	Administrative Assistant	8	45000
45	45	41	Male	Master's	Strategy Consultant	13	130000
46	46	28	Female	Bachelor's	Copywriter	2	40000
47	47	32	Male	Bachelor's	Account Manager	5	75000
48	48	45	Female	Master's	Director of Marketing	16	180000
49	49	38	Male	PhD	Senior Scientist	11	120000
50	50	25	Male	Bachelor's	Help Desk Analyst	0	35000
51	51	51	Female	Bachelor's	Customer Service Manager	22	130000
52	52	33	Male	Master's	Business Intelligence Analyst	7	85000
53	53	40	Female	Bachelor's	Event Coordinator	12	60000
54	54	47	Male	Master's	VP of Finance	19	200000
55	55	29	Female	Bachelor's	Graphic Designer	3	50000
56	56	36	Male	Bachelor's	Sales Manager	9	95000
57	57	27	Female	Master's	UX Researcher	2	65000
58	58	43	Male	PhD	Senior Engineer	17	140000
59	59	30	Female	Bachelor's	Social Media Manager	4	55000
60	60	35	Male	Master's	Product Manager	7	105000

INPUT

```r
# Assuming the dataset is stored in a data frame named 'df'

# Extracting the 'Salary' column
salary <- df$Salary

# Calculating mean, median, variance, minimum, maximum, standard deviation, range, and quantiles
mean_salary <- mean(salary)
median_salary <- median(salary)
variance_salary <- var(salary)
min_salary <- min(salary)
max_salary <- max(salary)
sd_salary <- sd(salary)
range_salary <- range(salary)
quantiles_salary <- quantile(salary)

# Print the results
cat("Mean Salary:", mean_salary, "\n")
cat("Median Salary:", median_salary, "\n")
cat("Variance of Salary:", variance_salary, "\n")
cat("Minimum Salary:", min_salary, "\n")
cat("Maximum Salary:", max_salary, "\n")
cat("Standard Deviation of Salary:", sd_salary, "\n")
cat("Range of Salary:", range_salary, "\n")
cat("Quantiles of Salary:", quantiles_salary, "\n")
```

OUTPUT

```
> # Extracting the 'Salary' column
> salary <- df$Salary
>
> # Calculating mean, median, variance, minimum, maximum, standard deviation, range, and quantiles
> mean_salary <- mean(salary)
> median_salary <- median(salary)
> variance_salary <- var(salary)
> min_salary <- min(salary)
> max_salary <- max(salary)
> sd_salary <- sd(salary)
> range_salary <- range(salary)
> quantiles_salary <- quantile(salary)
>
> # Print the results
> cat("Mean Salary:", mean_salary, "\n")
Mean Salary: 97250
> cat("Median Salary:", median_salary, "\n")
Median Salary: 90000
> cat("Variance of Salary:", variance_salary, "\n")
Variance of Salary: 2600357143
> cat("Minimum Salary:", min_salary, "\n")
Minimum Salary: 30000
> cat("Maximum Salary:", max_salary, "\n")
Maximum Salary: 250000
> cat("Standard Deviation of Salary:", sd_salary, "\n")
Standard Deviation of Salary: 50993.7
> cat("Range of Salary:", range_salary, "\n")
Range of Salary: 30000 250000
> cat("Quantiles of Salary:", quantiles_salary, "\n")
Quantiles of Salary: 30000 55000 90000 130000 250000
```

INPUT

```
> # Assuming the dataset is stored in a data frame named 'df'
>
> # Extracting the 'Years of Experience' and 'Age' columns
> experience <- df$Years.of.Experience
> age <- df$age
>
> # Function to print summary statistics
> print_summary_stats <- function(data, name) {
+   cat("Summary statistics for", name, "\n")
+   cat("Mean:", mean(data), "\n")
+   cat("Median:", median(data), "\n")
+   cat("Variance:", var(data), "\n")
+   cat("Minimum:", min(data), "\n")
+   cat("Maximum:", max(data), "\n")
+   cat("Standard Deviation:", sd(data), "\n")
+   cat("Range:", range(data), "\n")
+   cat("Quantiles:", quantile(data), "\n\n")
+ }
>
> # Print summary statistics for 'Years of Experience'
> print_summary_stats(experience, "Years of Experience")
Summary statistics for Years of Experience
Mean: 10.09583
Median: 9
Variance: 49.32897
Minimum: 0
Maximum: 25
Standard Deviation: 7.023459
Range: 0 25
Quantiles: 0 3.75 9 15.25 25
```

OUTPUT

```
>
> # Print summary statistics for 'Age'
> print_summary_stats(age, "Age")
Summary statistics for Age
Mean: 37.01667
Median: 36
Variance: 61.6972
Minimum: 23
Maximum: 52
Standard Deviation: 7.854756
Range: 23 52
Quantiles: 23 30 36 43 52
```

INPUT

```
summary(df)
```

OUTPUT

```
> summary(df)
     Sr.no.            age           Gender          Education.Level     Job.Title
 Min.   :  1.00   Min.   :23.00   Length:120         Length:120         Length:120
 1st Qu.: 30.75   1st Qu.:30.00   Class :character   Class :character   Class :character
 Median : 60.50   Median :36.00   Mode  :character   Mode  :character   Mode  :character
 Mean   : 60.50   Mean   :37.02
 3rd Qu.: 90.25   3rd Qu.:43.00
 Max.   :120.00   Max.   :52.00
 Years.of.Experience     Salary
 Min.   : 0.00       Min.   : 30000
 1st Qu.: 3.75       1st Qu.: 55000
 Median : 9.00       Median : 90000
 Mean   :10.10       Mean   : 97250
 3rd Qu.:15.25       3rd Qu.:130000
 Max.   :25.00       Max.   :250000
>
```

INPUT

```
1  library(psych)
2  describe(df)
```

OUTPUT

```
Console  Background Jobs
R 4.2.3 · ~/
> library(psych)

Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha

> describe(df)
                   vars   n     mean       sd  median  trimmed      mad   min    max  range  skew
Sr.no.                1 120    60.50    34.79    60.5    60.50    44.48     1    120    119  0.00
age                   2 120    37.02     7.85    36.0    36.76     8.90    23     52     29  0.25
Gender*               3 120     1.54     0.50     2.0     1.55     0.00     1      2      1 -0.17
Education.Level*      4 120     1.57     0.69     1.0     1.47     0.00     1      3      2  0.78
Job.Title*            5 120    52.03    28.91    53.5    52.12    35.58     1    103    102 -0.04
Years.of.Experience   6 120    10.10     7.02     9.0     9.69     8.90     0     25     25  0.39
Salary                7 120 97250.00 50993.70 90000.0 91822.92 51891.00 30000 250000 220000  0.84
                   kurtosis      se
Sr.no.                -1.23    3.18
age                   -1.08    0.72
Gender*               -1.99    0.05
Education.Level*      -0.62    0.06
Job.Title*            -1.16    2.64
Years.of.Experience   -1.07    0.64
Salary                 0.03 4655.07
>
```

DATASET VISUALISATION

INPUT

Source on Save

```r
# Load the required library
library(ggplot2)
library(psych)

# Assuming the dataset is stored in a data frame named 'df'



# Create a scatter plot using ggplot2
ggplot(df, aes(x = Years.of.Experience, y = Salary)) +
  geom_point() +
  labs(title = "Scatter Plot of Years of Experience vs. Salary",
       x = "Years of Experience",
       y = "Salary")
```

OUTPUT

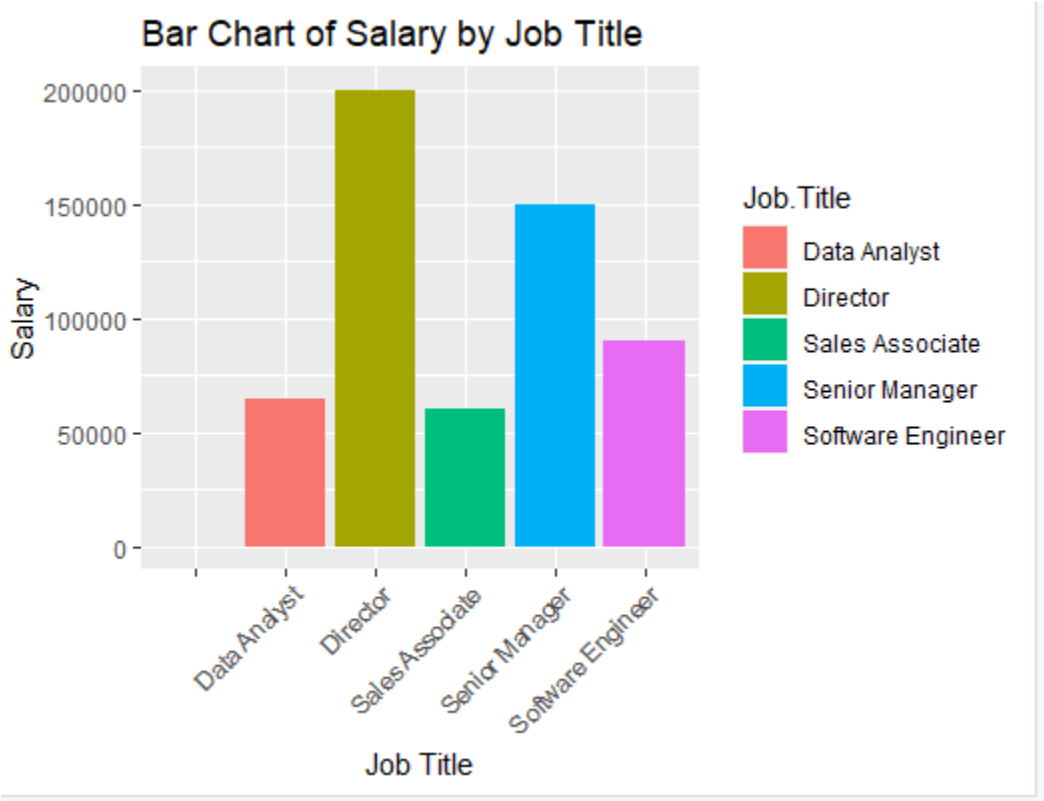
INPUT

```
# Load the required library
library(ggplot2)

# Assuming the dataset is stored in a data frame named 'df'


# Create a bar chart for Salary
ggplot(df, aes(x = Job.Title, y = Salary, fill = Job.Title)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Chart of Salary by Job Title",
       x = "Job Title",
       y = "Salary") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

OUTPUT



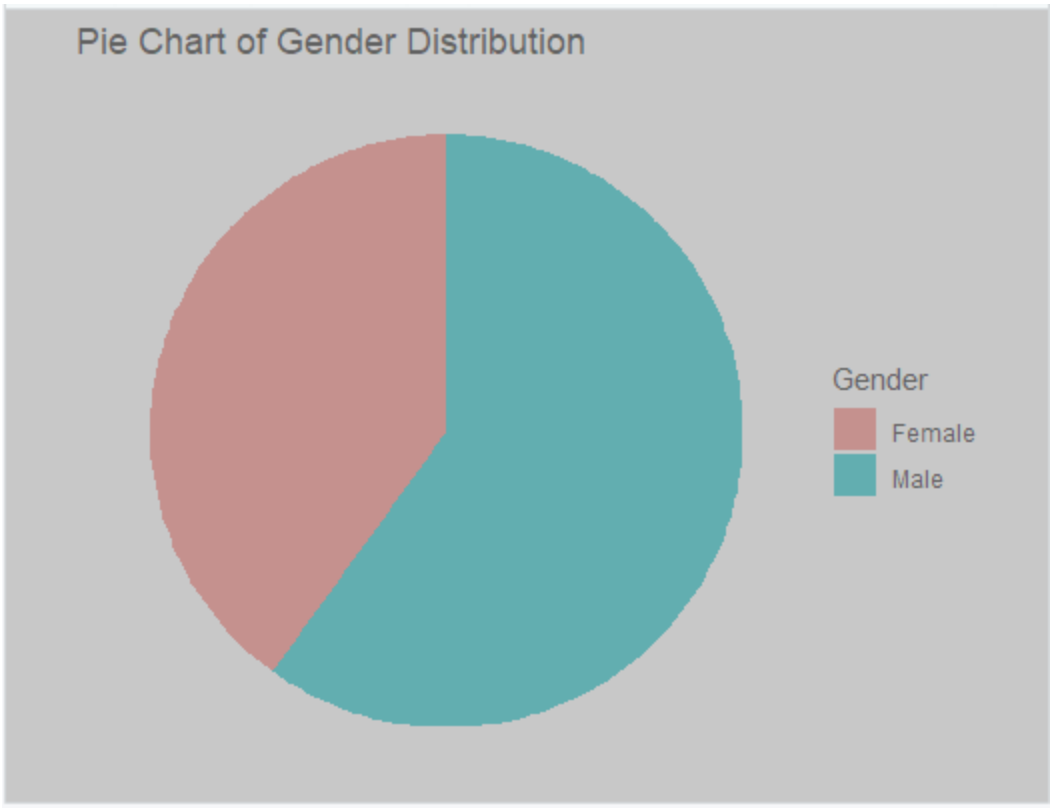
INPUT

```
# Load the required library
library(ggplot2)

# Assuming the dataset is stored in a data frame named 'df'

# Create a pie chart for Gender distribution
ggplot(df, aes(x = "", fill = Gender)) +
  geom_bar(stat = "count", width = 1) +
  coord_polar("y") +
  labs(title = "Pie Chart of Gender Distribution") +
  theme_void()  # Removes unnecessary elements like axis labels and ticks
```

OUTPUT



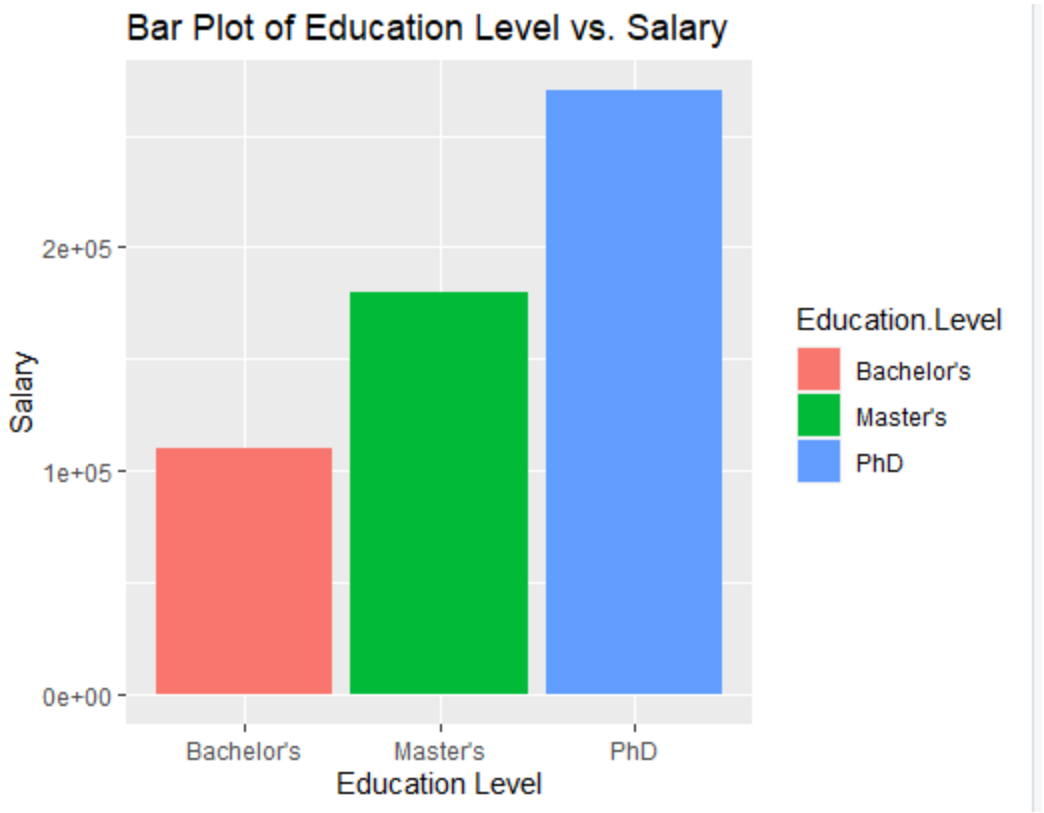
INPUT

```
# Assuming df is your dataset
# Make sure 'df' contains columns 'Education.Level' and 'Salary'

# Create a bar plot for Education Level vs. Salary
library(ggplot2)

ggplot(df, aes(x = Education.Level, y = Salary, fill = Education.Level)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Plot of Education Level vs. Salary",
       x = "Education Level",
       y = "Salary")
```

OUTPUT



INSIGHTS

Insights and data visualizations often depend on specific questions or goals you have for your analysis. However, based on the dataset provided (featuring columns like "Age," "Gender," "Education Level," "Job Title," "Years of Experience," and "Salary"), I can suggest some general insights and visualizations that might be of interest:

Age Distribution:

- Visualize the distribution of ages in the dataset using a histogram or a density plot. This can provide insights into the age demographics of the workforce.

Gender Distribution:

- Create a pie chart or a bar plot to show the gender distribution in the dataset. This helps in understanding the gender balance within the sample.

Education Level vs. Salary:

- Use a bar plot or box plot to compare the average salaries across different education levels. This can reveal if there's a correlation between education level and salary.

Job Title vs. Salary:

- Create a bar plot or a box plot to visualize the salary distribution across different job titles. This can provide insights into the salary range for different roles.

Experience vs. Salary:

- Explore the relationship between years of experience and salary using a scatter plot or a line plot. This can help identify trends in how salary changes with increased experience.

Salary Distribution:

- Use a histogram or a density plot to visualize the distribution of salaries in the dataset. This provides insights into the overall salary structure.

Correlation Matrix:

- Generate a correlation matrix to quantify relationships between numerical variables like age, years of experience, and salary. This can be helpful in understanding how variables are interrelated.

Job Title Breakdown:

- Create a bar plot or a pie chart to show the distribution of job titles in the dataset. This gives an overview of the different roles present.

Salary by Gender:

- Investigate if there is a gender pay gap by comparing the average salary for males and females. A grouped bar plot can be useful for this visualization.

Experience by Job Title:

- Use a box plot or violin plot to visualize the distribution of years of experience across different job titles. This can provide insights into the typical experience levels required for specific roles.

RECOMMENDATIONS

These recommendations focus on potential areas of improvement, further analysis, and actions that can be taken:

Gender Pay Gap Analysis:

- Conduct a detailed analysis of the gender pay gap within the organization. This may involve a more in-depth exploration of salary differences between male and female employees across different job titles and experience levels.

Employee Retention Strategies:

- Explore the relationship between years of experience and job satisfaction. Identify trends or patterns that might indicate potential areas for improvement in employee retention or engagement.

Performance Analysis by Job Title:

- Consider a performance analysis by job title to understand if there are variations in performance metrics or outcomes. This can help in recognizing high-performing teams or areas for improvement.

Professional Development Opportunities:

- Assess the correlation between education level and job title. Identify areas where additional education or professional development opportunities may contribute to career advancement.

Succession Planning:

- Analyze the distribution of years of experience across different job titles. This information can be valuable for succession planning and identifying potential gaps in leadership roles.

Diversity and Inclusion Initiatives:

- Evaluate the diversity within different job titles and departments. Consider implementing or enhancing diversity and inclusion initiatives to foster a more inclusive workplace.

Salary Benchmarking:

- Conduct salary benchmarking to compare the organization's salary levels with industry standards. This can help ensure that the organization remains competitive in attracting and retaining talent.

Employee Satisfaction Survey:

- Consider conducting an employee satisfaction survey to gather qualitative insights into factors affecting job satisfaction, work-life balance, and overall workplace culture.

Career Path Planning:

- Develop clear career paths for employees, especially those in entry-level positions. This can contribute to increased employee motivation and retention.

Training Programs:

- Identify areas where additional training programs or skill development initiatives can benefit employees and align with organizational goals.