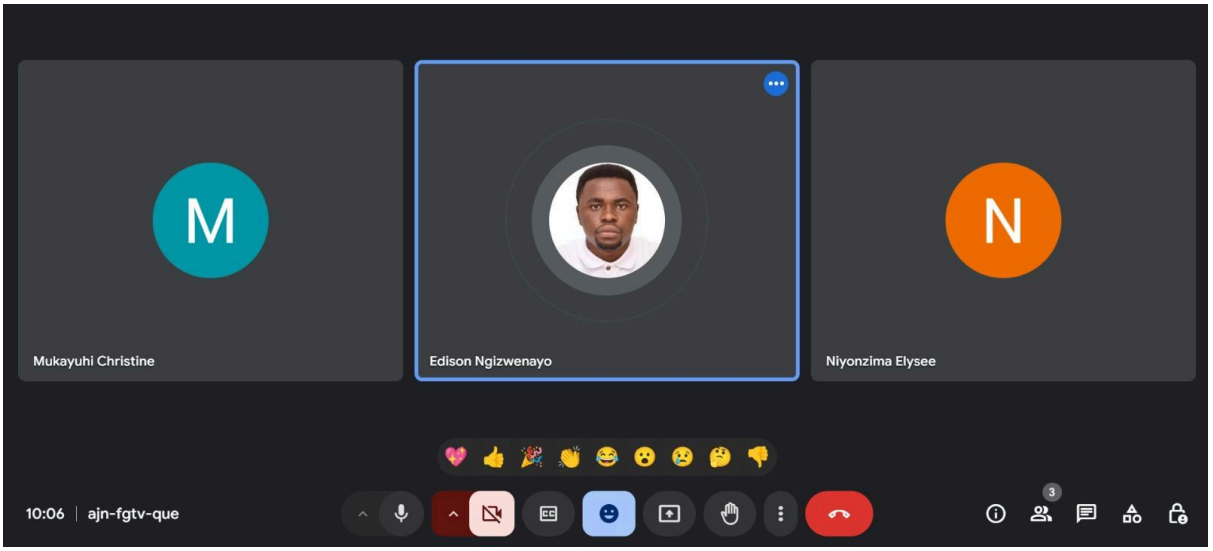
GROUP MEMBERS

Names	Registration Number
Edison Ngizwenayo	**217029434**
Christine Mukayuhi	**10104747**
Elysee Niyonzima	**224019528**

INTRODUCTION

The primary goal of this exploratory data analysis (EDA) was to demonstrate the application of R programming concepts to analyze a dataset effectively. The analysis involved creating and manipulating data frames, performing slicing and sampling operations, applying summary statistics, and utilizing visualizations to derive insights. The dataset used for this analysis consisted of six columns and 914 rows, including both numerical and categorical variables. And this random dataset was generated from website called https://www.mockaroo.com/

As a team, we collaborated effectively using Google Meet, and the following screenshot illustrates our virtual teamwork in action.



Key Observations

1. Dataset Overview

- The dataset was loaded using a CSV file, containing 6 columns and 914 rows.

```
This dataframe has 914 rows and 6 columns
```

- Top rows of the dataset revealed columns such as age, salary, score, year, and categorical variables like Department and gender.

age	salary	score	gender	Department	year
38	475971	53	Female	Human Resources	2013
54	205066	64	Female	Accounting	2013
52	616716	77	Female	Marketing	2013

2. Slicing Operations

For better understanding slicing we have perform the following operation

- Filtered rows where salary was greater than 600,000, identifying high-income employees. Below is screenshot with sample salary greater than 600,000

age	salary	score	gender	Department	year
52	616716	77	Female	Marketing	2013
52	825918	69	Female	Product Management	2013
24	759520	60	Male	Engineering	2013
56	999627	54	Male	Engineering	2012
30	604431	54	Female	Accounting	2012
43	903310	54	Female	Marketing	2012
20	966957	63	Female	Accounting	2012
21	982015	75	Female	Accounting	2012

- Isolated youth population by filtering rows where age was between 18 and 35, revealing demographic insights. Below is screenshot with age was between 18 and 35 which represent youth in Rwanda context

age	salary	score	gender	Department	year
24	759520	60	Male	Engineering	2013
29	430063	77	Male	Engineering	2012
30	604431	54	Female	Accounting	2012
20	966957	63	Female	Accounting	2012
21	321184	59	Female	Human Resources	2012
32	282752	66	Male	Engineering	2012
21	982015	75	Female	Accounting	2012

- Identified top performers in the year 2013 with scores above 80.The data reveal that the top performers in 2013 is Female with score of 89

age	salary	score	gender	Department	year
52	616716	89	Female	Marketing	2013

3. Sampling Operations

- Randomly sampled 10 rows from the dataset to observe variability and confirm representativeness.

```
age salary score gender          Department year
 31 714077    74 Female Product Management 2010
 30 780756    62   Male     Human Resources 1999
 58 303118    69 Female          Accounting 2006
 22 901776    59 Female     Human Resources 1991
 55 541856    67 Female     Human Resources 1993
 60 677288    66   Male     Human Resources 2005
 34 785475    67 Female           Marketing 2007
 46 454294    66 Female           Marketing 2008
 47 641599    61   Male     Human Resources 1990
 46 517512    69   Male          Accounting 2013
```

4. Apply-family Functions

- Applied apply() to compute column-wise mean values for numerical columns. We used apply to compute column mean

```
      age       salary        score
39.27790 595206.60722     64.56893
```

- Used lapply() to calculate the range (minimum and maximum) for each numerical column. We used range function returns a list with min and max for each column and then apply it to our dataframe using lapply

```
$age
[1] 18 60

$salary
[1] 200247 999998

$score
[1] 50 89
```

- Leveraged sapply() to compute summary statistics (mean, median, min, max) for numerical columns.

```
            age   salary    score
mean   39.2779 595206.6 64.56893
median 39.0000 606205.0 64.00000
min    18.0000 200247.0 50.00000
max    60.0000 999998.0 89.00000
```

- Utilized mapply() to generate a combined metric based on age and salary. Below is the screenshot of the result

```
  [1] 11863.775 17104.752 17447.625 24574.284 11097.020 31005.975 58907.311 53300.352 12975.120 14654.211  9974.748 17106.963
 [13] 12579.980 16901.532 19477.950 54001.805 32615.660 14901.516 52008.794 10038.194 27807.934 40936.056 14624.816 13311.500
 [25] 15059.733 11455.320 56840.580 41673.528 31175.244  6310.071 31136.618 13079.450 14644.960 17078.970 13588.260 12927.564
 [37] 17688.069 42133.249 36521.365 37180.131  9411.960 34523.711 24664.528 31053.869  9985.716 18034.578 25092.801  8105.286
 [49] 22604.976 17999.964 15096.528 16707.263 39341.343 20025.712 18093.040 47139.976 12340.549 26687.250  9765.713 14596.214
 [61] 27905.384 13929.375 44537.062 11065.776 38088.076 28720.494 36275.232 13675.374 34956.675 33499.626 30222.514 53602.503
 [73] 24868.402 22638.900 26077.324 40789.710 17823.070 16552.588 12517.434 47628.412 57312.758 13901.545 15154.090 25384.080
 [85]  4412.940 16182.250 18031.218 22233.425 11838.480 50273.496 16367.890 43081.965 47387.466 38059.112 14453.760 30045.636
 [97] 22165.660 13620.486 41927.683 18868.330  7761.291 17299.920 11034.912 38954.208 18371.803 37808.150  8695.070 16646.601
[109] 13310.768 10390.226 29590.560 40408.658 10659.363 49994.122 22347.452 45042.360 30155.153 20252.253 56834.820 17839.030
[121] 18402.034 10560.205 21774.324 15196.400 24921.558 20039.425 14833.908 13996.052 25349.144  7784.244 21001.806  9651.440
[133] 34254.264 48323.222 34743.566 23665.048 22630.740 11348.180 47793.024 19839.072 45620.292  9963.360 10436.605 14859.768
[145] 30799.065 30660.300 10753.497 11395.254 21220.885 13518.180  5420.541 35851.119 26649.379 31205.880 28678.824 39274.895
```

5. Summary Operations

- Calculated the mean and sum of numerical columns, excluding the year column, to highlight overall trends.

age_mean	age_sum	salary_mean	salary_sum	score_mean	score_sum
39.2779	35900	595206.6	544018839	64.56893	59016

- Tabulated employee counts by Department and gender to understand categorical distributions.

Distribution by department

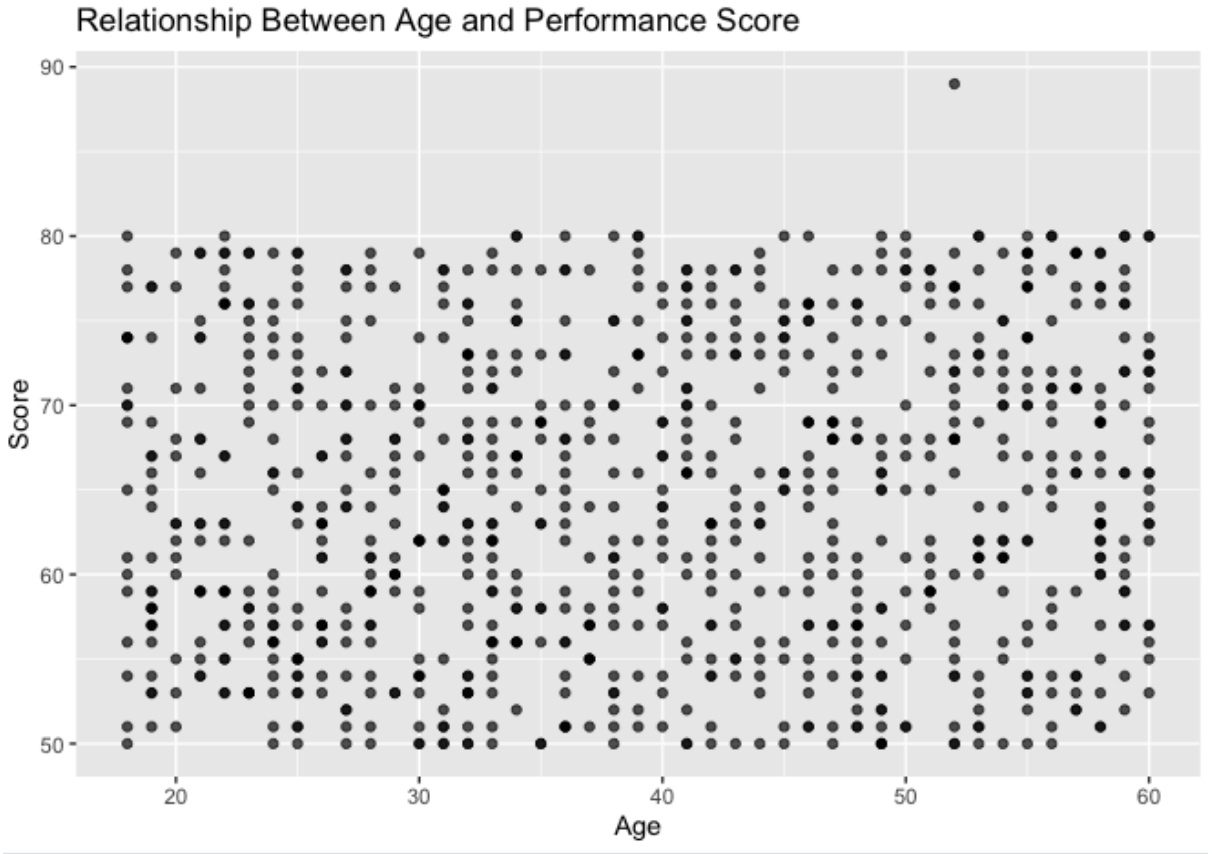
Accounting	Engineering	Human Resources	Marketing	Product Management
133	231	323	152	75

Distribution by Gender

Female	Male
458	456

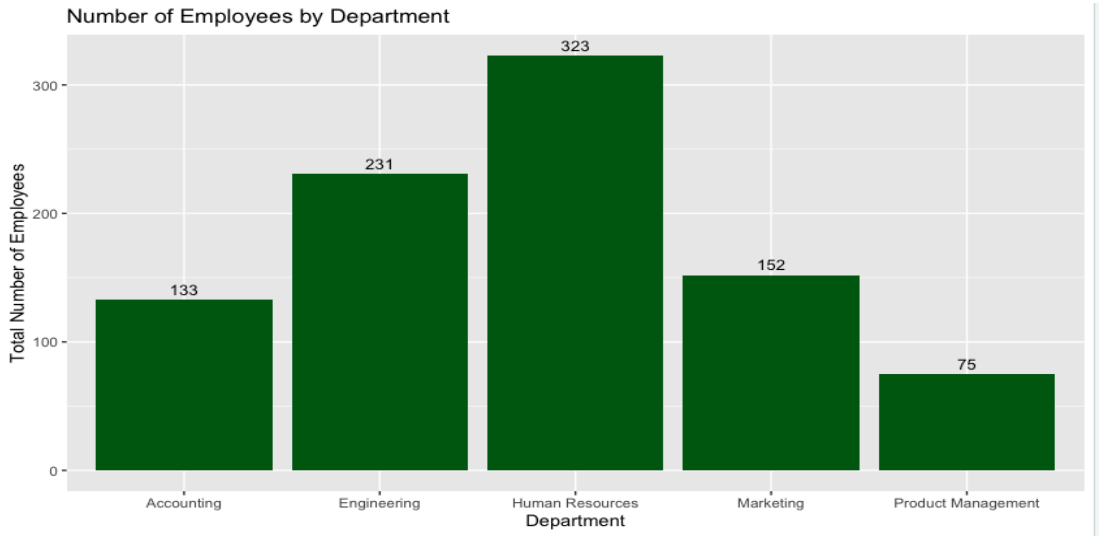
1. Scatter Plot: Age vs. Score

- Visualized the relationship between age and performance scores.The illustrations indicate that as age increases, employees tend to achieve better scores. However, there appears to be an outlier for an employee aged 89



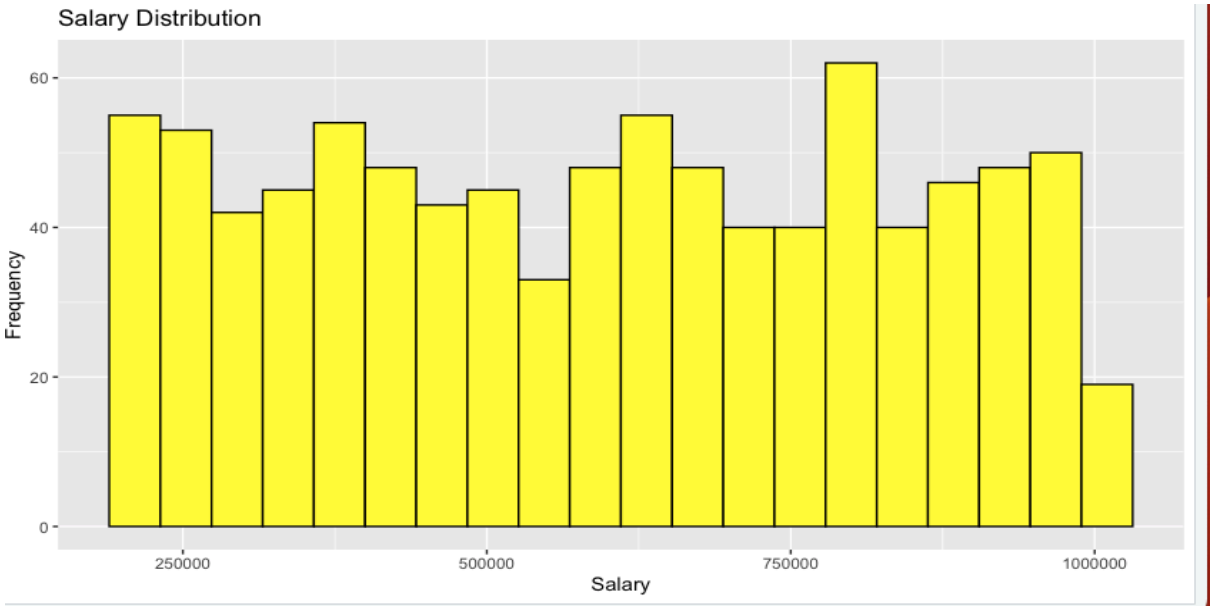
2. Bar Plots

- **Employee Count by Department**: Revealed the distribution of employees across various departments, with Human resources being significantly larger followed by the Engineering department, and also Product management showing a lower, a very low number.



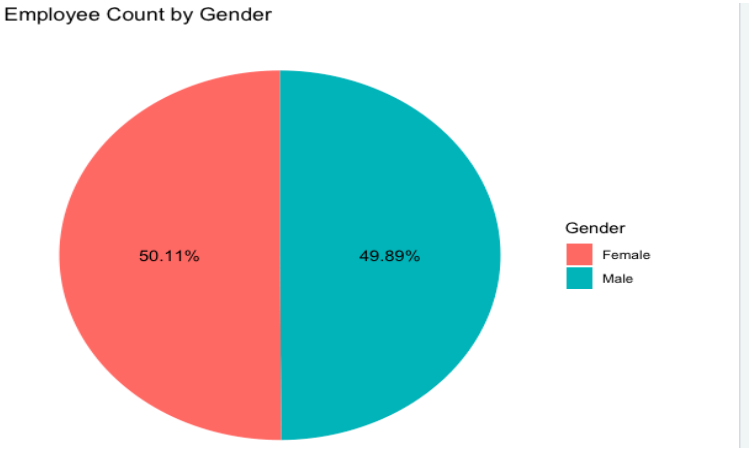
3. Histogram: Salary Distribution

- Highlighted the salary distribution among employees, identifying peaks and outliers.



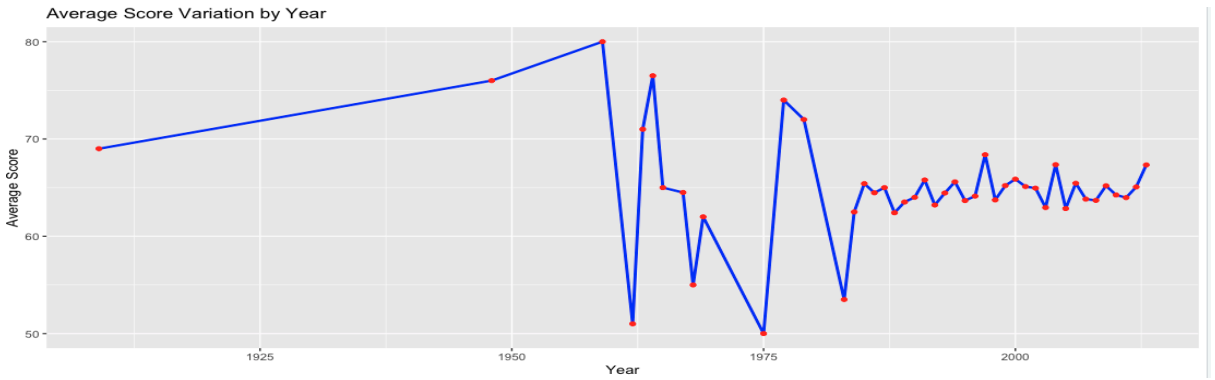
4. Pie chart Presenting Gender

The illustration shows that the number of female employees exceeds that of male employees



5. Line Plot: Average Score Over Years

- Tracked the variation in average performance scores across different years.The illustration depicts fluctuations in the average score over time, showing a gradual upward trend from 1983 to 2013



Conclusion

This EDA demonstrated the power of R in exploring and analyzing datasets effectively. Key takeaways include:

- The ability to slice and filter data provided focused insights, such as identifying high-income earners and top performers.
- Summary statistics and apply-family functions helped compute descriptive metrics efficiently.
- Visualizations provided a clear understanding of trends, distributions, and relationships in the data.

By leveraging these techniques, we gained actionable insights into employee demographics, performance, and salary distribution, setting the stage for further analyses and data-driven decision-making.

Refer to the following github link for to fully access script, and dataset
https://github.com/cavani12345/R-Data-analysis-Assignment-Evening-Session-Group-3