7COM1079-0901-2024 - Team Research and Development Project

Final report title: Analysis of Employee Data

Group ID: A17

Dataset number:

Prepared by: [Mohan Sanghishetty]

[Venkatesh Kancharlapalli]

[Sri Ganesh Thotakura]

[Srikar Bakani]

[Sriram Reddy Kesireddy]

University of Hertfordshire

Hatfield, 2024

Table of Contents

[1. Introduction 4](#_Toc187096516)

[1.1 Problem Statement and Research Motivation 4](#_Toc187096517)

[1.2 Dataset 4](#_Toc187096518)

[1.3 Research Question 4](#_Toc187096519)

[1.4 Null Hypothesis and Alternative Hypothesis 4](#_Toc187096520)

[2. Background Research 5](#_Toc187096521)

[2.1 Research Papers 5](#_Toc187096522)

[2.2 RQ is of Interest 6](#_Toc187096523)

[3. Visualisation 6](#_Toc187096524)

[3.1 Appropriate Plot for RQ 6](#_Toc187096525)

[3.2 Useful Information for Data Understanding 7](#_Toc187096526)

[4. Analysis 8](#_Toc187096527)

[4.1 Statistical Test Used to Test the Hypotheses and Output 8](#_Toc187096528)

[4.2 Null Hypothesis Rejected/Not Rejected 8](#_Toc187096529)

[5. Evaluation 8](#_Toc187096530)

[5.1 What went Well 8](#_Toc187096531)

[5.2 Points for Improvement 9](#_Toc187096532)

[5.3 Group’s Time Management 9](#_Toc187096533)

[5.4 Project’s Overall Judgement 9](#_Toc187096534)

[6. Conclusions 9](#_Toc187096535)

[6.1 Results Explained 9](#_Toc187096536)

[6.2 Interpretation of Results 9](#_Toc187096537)

[6.3 Reasons for Future work, Limitations of Study 10](#_Toc187096538)

[7. References 11](#_Toc187096539)

[8. Appendices 12](#_Toc187096540)

[8.1 R Code for Analysis and Visualization 12](#_Toc187096541)

**List of Figure**

[Figure 1 Age Distribution with Trend Bell 6](#_Toc187096542)

[Figure 2 Monthly Income Distribution with Trend Bell 7](#_Toc187096543)

[Figure 3 Scatter Plot of Age vs Monthly Income 7](#_Toc187096544)

# Introduction

## Problem Statement and Research Motivation

The problem comes down to: what are the relationships between the characteristics of employees (age, etc) and their monthly income. This analysis is important to look at wage gaps to see how they shape, see if there's any trends in the income distribution, and what kind of workforce management strategy can improve the quality of work. As prior research has shown demographic factors play important role in income determination, organizations should create equitable pay structures. However, to date there have been existing studies attempting to analyze such associations using modern techniques in statistics (Shafie et al., 2024). This research seeks to fill this gap on bridging it with actionable insights that still contribute to fairness and efficiency of the organizational compensation policies powered by the employee data analytics.

## Dataset

For this project they used a complete dataset with 35 variables about icrosoft employees such as Age, MonthlyIncome and JobRole and YearsAtCompany. The data has the dimensions of demographics, and job satisfaction and performance metrics were included to look at it from a multidimensional perspective. The data was fetched from Kaggle, keeping it accurate and reliable data. By utilizing this well-structured dataset, users want to analyze the correlation between employees’ age and monthly income by discovering workforce trends and compensation patterns.

## Research Question

RQ: Is there any significant association between employees age and monthly income?

The research question asks if there is a substantial link between the age of the employees and their monthly gross income. This will be answered using statistical techniques such as Spearman rank correlation to analyze relationship between. Scatter plots and distribution charts will help us understand patterns, and validate the statistics.

## Null Hypothesis and Alternative Hypothesis

**Null-hypothesis**: There is no significant association between employee age and monthly income.

**Alternative-hypothesis**: There is a significant association between employee age and monthly income.

The null hypothesis (H0) states that there is no significant relationship between employee age and monthly income. That is, age variations do not systematically vary with monthly income variations, and any such correspondence we observe is probably random.

An alternative hypothesis (H1) was that there is a relationship between employee age and monthly income. In other words, if there is a meaningful pattern or trend for age across months, then potentially we also have a causal or a correlational relationship to be explored so this implies that as age changes, there is a relationship to the monthly income.

Based on evidence contained in the dataset, a null hypothesis will be revoked in favor of an alternative hypothesis using statistical testing.

# Background Research

## Research Papers

This study utilises a case in point dataset that fits completely with themes that have been explored in multiple research papers around employee data analytics and HR analytics.

Following the Basha et al., (2020) research on using machine learning techniques such as Multiple linear regression, there is research on HR analytics. The scope of their study on patterns and predicting human resources decision making was the datasets provided to them with variables including age, salary, and performance metrics. This is consistent with how our dataset uses demographic and income data to seek associations and trends.

In Vishesh, (2023), the performance, satisfaction, and demographic factors were analyzed with data science methodologies of employee datasets. Accordingly, our work is to actually analyze age-income relationships, and their findings emphasized the importance of predictive analytics to identify key workforce trends. In Saxena et al., (2021) which reviewed tools and techniques found in the HR analytics, it noted importance of data driven approaches to modern workforce management. Their research found that the more datasets were being used that included variables like compensation and tenure, then the more these datasets were being used to steer organizational strategy.

The relevance and applicability of these employee datasets in HR analytics is demonstrated in these studies, and the critical role of using statistical, and, more generally, machine learning techniques to extract informative actionable insights.

## RQ is of Interest

The research question is of interest because it addresses a critical aspect of workforce analysis: age and income relationship. Employee data analysis is a well-researched field; however, age has not been looked at in regard to compensation. In this way, this study fills a gap, as it examines this precise relationship and offers insights into how organizations need to design equitable pay structures (Shobhanam and Sumati, 2022). First, by looking at this connection, future work may then look at other demographic elements that impact on income, develop further approaches for managing the workforce, and generate fairness for organizational compensation policies.

# Visualisation

## Appropriate Plot for RQ

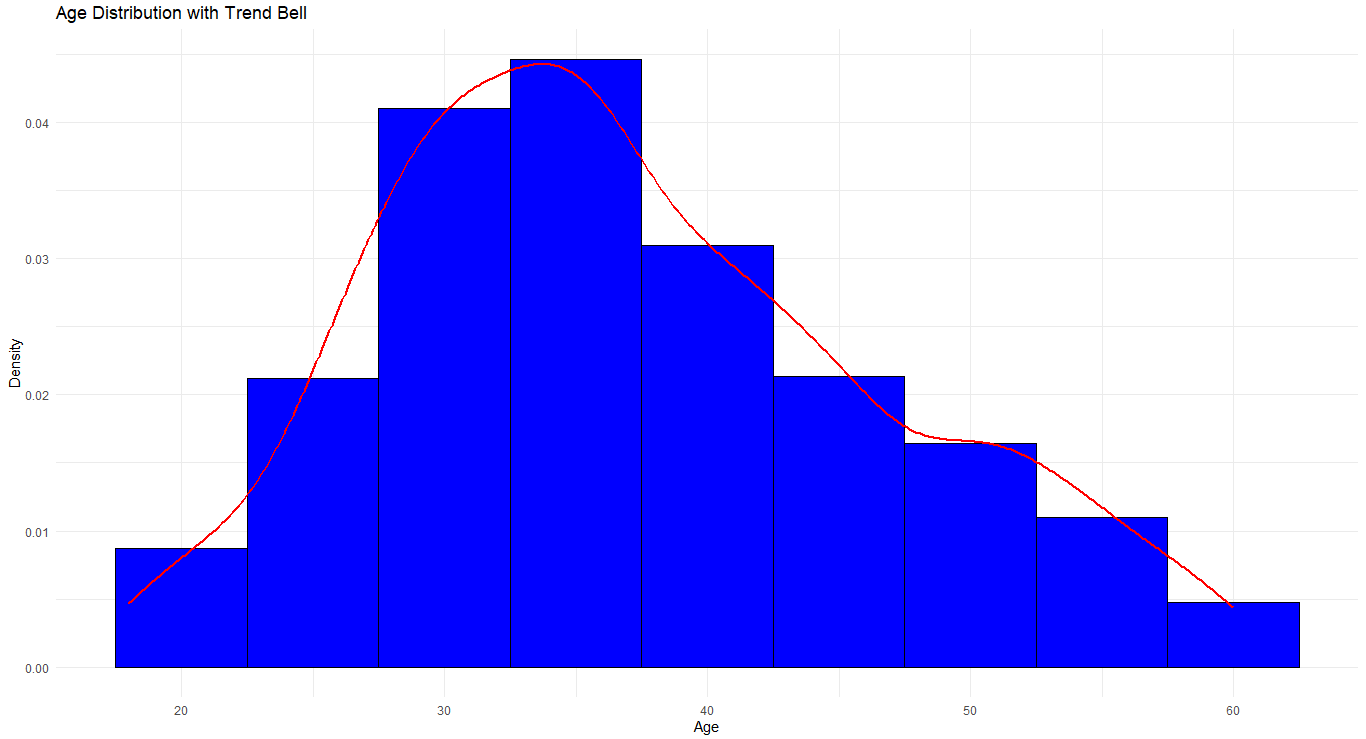


Figure 1 Age Distribution with Trend Bell

Ages are plotted on a bell curve graph, peaking at 30. Almost all are between ages 20-50, with a normal distribution around about age 30.

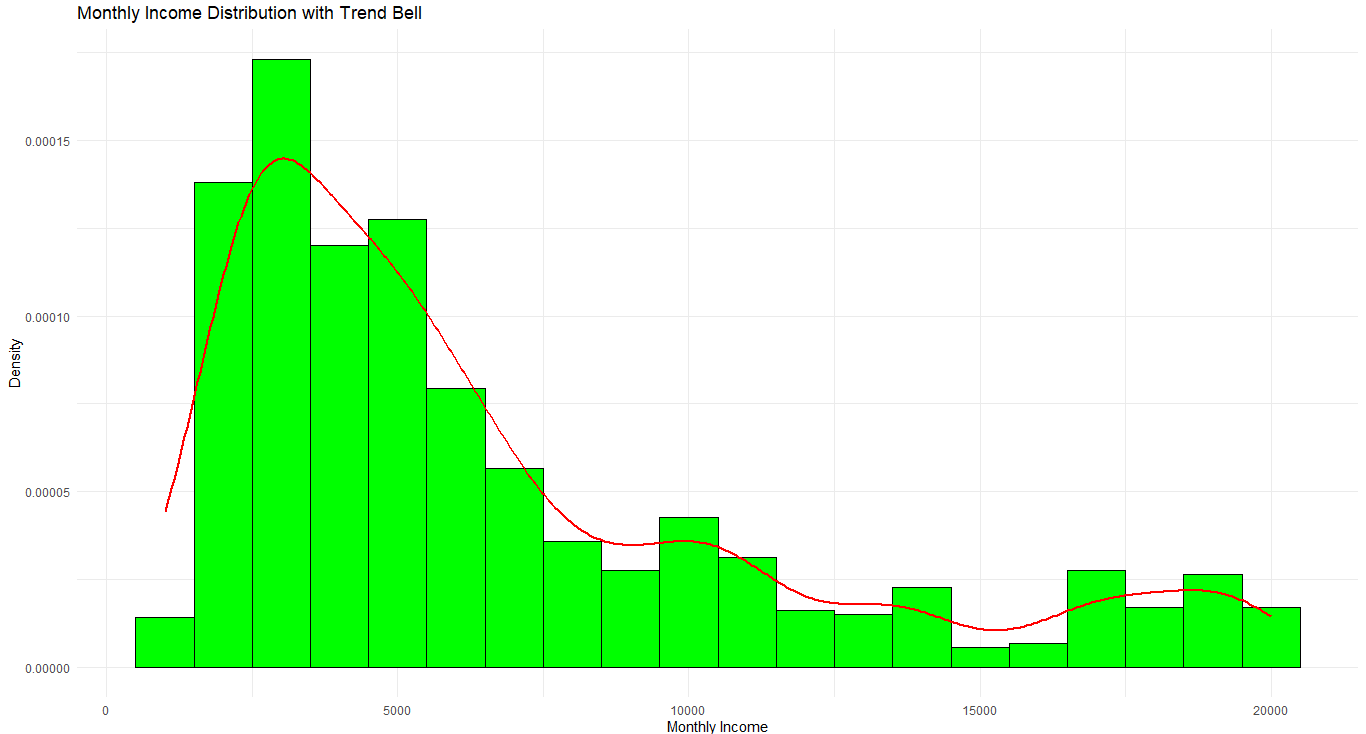


Figure 2 Monthly Income Distribution with Trend Bell

Monthly income has a right skew distribution (as shown in the graph) with a density peak at about 40,000 and a declining trend of income with larger values, reflected in the overall pattern with a linear trend line.



Figure 3 Scatter Plot of Age vs Monthly Income

The positive correlation between age (20-60) and monthly income (0-20000) is indicated by a red trend line as income increases with age.

## Useful Information for Data Understanding

Age distribution graph is a normal distribution with most employee age 20-50, which peaks around 33, implying a younger workforce. Most incomes are clustered around 40,000 with income disparity shown by the right skewness in the income distribution. The scatter plot shows that income increases with age: they are positively correlated.

# Analysis

## Statistical Test Used to Test the Hypotheses and Output

The Spearman's rank correlation test is a non-parametric test for continuous data which might not follow a normal distribution, it was selected to test the relationship between age and monthly income. The test will calculate a correlation coefficient (rho) to measure the strength of the relationship and the direction of the relationship too (McClenaghan, 2024). A moderate positive correlation is indicated by a test with a correlation coefficient of 0.4747. The p-value associated with the observed relationship is highly significant (< 2.2e-16), and so the test is appropriate for the research question.

## Null Hypothesis Rejected/Not Rejected

Our group reject the null hypothesis, that there is no significant relationship between age and monthly income because the p-value (< 2.2e-16). As presented, this very low p-value provides strong evidence against the null hypothesis. The result is positive correlation coefficient (rho = 0.4747) which means a moderate and significant relationship between age and monthly income (Tanasescu et al., 2024). Therefore, researcher find, that overall, monthly income is generally increased by age, indicating that age is a relevant explanatory variable for explaining employee compensation.

# Evaluation

## What went Well

Statistical techniques were applied by the group to analyze the dataset, and the resulting insights were meaningful. All members collaborated well, cleaning, visualizing and interpreting the data amongst themselves. For example, analysis using tools such as R meant that it became easy, smoothly and quickly to go through the analysis process, and that the analysis covered as thoroughly the research question.

## Points for Improvement

The group could improve data preprocessing by refining the handling of missing or outlier values. Filling out the workflow and decision-making processes would help make them clearer. Furthermore, the integration of the advanced analytical methods and more extensive literature review may help to obtain more robust insights for forthcoming projects.

## Group’s Time Management

Since it was a group, we kept to a schedule and that meant deadlines were met. Task delegation during early planning helped to streamline workflow though final review and interpretation of results could benefit from more time.

## Project’s Overall Judgement

The project achieved all of its objectives, namely, actionable insight into the correlation between age and income. The analysis was very thorough, findings were clearly presented and dealt with messages were strong and understood clearly statistically techniques as well as their application in analyzing of employee data.

# Conclusions

## Results Explained

Group used Spearman's rank correlation to determine a moderate positive correlation (rho = 0.4747) between employee age and monthly income, and a highly significant p-value (< 2.2e-16). This suggests that the variation in income among employees is due not only to age but also related to the fact that as employees age, their monthly income is on the increase.

## Interpretation of Results

The research question is confirmed: age is highly correlated with income. Thus, experience, often linked with age, appears to affect compensation. Organizations should consider that that these results suggest the need for effective approaches to pay structures based on demographic trends. The insight reported here also contributes towards understanding income progression across different workforce age groups.

## Reasons for Future work, Limitations of Study

Other factors such as job role or education, can be considered in future research to try and capture combined impacts on income. Limitations include reliance a single dataset, whose generalizability may be limited. In theory it’s better to further expand the dataset, or perhaps even use cross industry data to obtain deeper insights and more robust conclusions.

# References

Basha, M.Mahaboob., Srivani, M.J., Ankaiah, M.B., Dadakalandar, M.U. and Srinivaslulu, M.T. (2020). HR Analytics using R Machine Learning Algorithm: Multiple Linear Regression Analysis. International Journal of Innovative Technology and Exploring Engineering, [online] 9(5), pp.1179–1183. doi:https://doi.org/10.35940/ijitee.e2789.039520. ‌

Vishesh, S. (2023). Data Science to Analyse Employee Data. IJARCCE, [online] 12(1), p.101. doi:https://doi.org/10.17148/IJARCCE.2023.12114. ‌

Saxena, M., Bagga, T. and Gupta, S. (2021). Fearless path for human resource personnel’s through analytics: a study of recent tools and techniques of human resource analytics and its implication. International Journal of Information Technology, [online] 13(4), pp.1649–1657. doi:https://doi.org/10.1007/s41870-021-00677-z.

Shafie, M.R., Khosravi, H., Farhadpour, S., Das, S. and Ahmed, I. (2024). A cluster-based human resources analytics for predicting employee turnover using optimized Artificial Neural Networks and data augmentation. *Decision Analytics Journal*, [online] 11, p.100461. doi:https://doi.org/10.1016/j.dajour.2024.100461.

Shobhanam, K. and Sumati, S. (2022). HR Analytics: Employee Attrition Analysis using Random Forest. *International Journal of Performability Engineering*, [online] 18(4), pp.275–275. doi:https://doi.org/10.23940/ijpe.22.04.p5.275281.

Tanasescu, L.G., Vines, A., Bologa, A.R. and Vîrgolici, O. (2024). Data Analytics for Optimizing and Predicting Employee Performance. *Applied Sciences*, [online] 14(8), pp.3254–3254. doi:https://doi.org/10.3390/app14083254.

McClenaghan, E. (2024). *Spearman Rank Correlation*. [online] Technology Networks. Available at: https://www.technologynetworks.com/tn/articles/spearman-rank-correlation-385744 [Accessed 6 Jan. 2025].

# Appendices

## R Code for Analysis and Visualization

# Load necessary libraries

library(tidyverse)

# Reading dataset

EmployeeData <- read\_csv("file.csv", show\_col\_types = FALSE)

# Print the head of the dataset

head(EmployeeData)

# Print the tail of the dataset

tail(EmployeeData)

# Print column names of the dataset

colnames(EmployeeData)

# Check for missing values

colSums(is.na(EmployeeData))

# Print structure of the cleaned dataset

str(EmployeeData)

# Plotting the distribution of Age with trend bell

ggplot(EmployeeData, aes(x = Age)) +

geom\_histogram(aes(y = after\_stat(density)), binwidth = 5, fill = "blue", color = "black") +

geom\_density(color = "red", linewidth = 1) +

labs(title = "Age Distribution with Trend Bell", x = "Age", y = "Density") +

theme\_minimal()

# Plotting the distribution of Monthly Income with trend bell

ggplot(EmployeeData, aes(x = MonthlyIncome)) +

geom\_histogram(aes(y = after\_stat(density)), binwidth = 1000, fill = "green", color = "black") +

geom\_density(color = "red", linewidth = 1) +

labs(title = "Monthly Income Distribution with Trend Bell", x = "Monthly Income", y = "Density") +

theme\_minimal()

# Plotting Scatter plot between Age vs Monthly Income

ggplot(EmployeeData, aes(x = Age, y = MonthlyIncome)) +

geom\_point(color = "purple", alpha = 0.6) +

geom\_smooth(method = "lm", color = "red", se = FALSE) +

labs(title = "Scatter Plot of Age vs Monthly Income", x = "Age", y = "Monthly Income") +

theme\_minimal()

# Performing statistical tests for Spearman's rank correlation

cor.test(EmployeeData$Age, EmployeeData$MonthlyIncome, method = "spearman")