

**MH3511 Data Analysis with Computer**

**Group Project**

A Data Analytical Study of HR Metrics

| **Name** | **Matriculation No.** |
| --- | --- |
| Cheung Chun On | U2240361F |
|  |  |
|  |  |
|  |  |
|  |  |

**Content Page**

[**1. Introduction 3**](#_jbhust3dkzn4)

[**2. Data Description 3**](#_i8rj76sj5wuh)

[**3. Description and Cleaning of the Dataset 4**](#_ae38solmxugf)

[3.1. Summary statistics for the main variable of interest, Monthly Income 4](#_on4axvtw3q69)

[The following plots show the overall distribution of the variable MonthlyIncome. 4](#_xz4xf3ctxmn5)

[3.2. Summary statistics for the other variables 5](#_hpwy2s3cptgr)

[3.2.1. Age of employees, Age 5](#_mxp4mywp2v4o)

[3.2.2. Age group of employees, AgeGroup 6](#_dtt9or87iuu8)

[3.2.3. Whether Employees worked overtime, OverTime 6](#_myzgkbtq1464)

[3.2.4. Performance Level of Employees, PerformanceRating 6](#_bs5kqpru5x7i)

[3.2.5. Business Travel Frequency of Employees, BusinessTravel 7](#_px0oqob5k3zq)

[3.2.6. Job Satisfaction Level of Employees, JobSatisfaction 7](#_kaovfi9pxz0w)

[3.2.7. Departments of Employees, Department 7](#_zhl6s7gf4t3n)

[3.2.8. Number of companies worked, NumCompaniesWorked 7](#_ly5uhxdnuvl2)

[3.2.9. Distance of workplace from Employees’ Homes, DistanceFromHome 8](#_uhyb2lju9r81)

[3.2.10. Number of times employees trained last year, TrainingTimesLastYear 8](#_z03md8kgynlo)

[3.2.11. Number of years the employee been with the company, YearsAtCompany 8](#_6v42uxteliji)

[3.2.12. Number of years the employee been in the current role, YearsInCurrentRole 8](#_ex86vog48dl4)

[**4. Statistical Analysis 9**](#_hnmg7pbaywip)

[4.1. Correlation between log(MonthlyIncome) and other continuous variables 9](#_ac0v2jfu4xra)

[4.2 Statistical Tests 9](#_e2493mwppgm0)

[4.2.1. Relation between Working Overtime and Monthly Income 9](#_cngqb0s9u8p2)

[4.2.2. Correlation between Age and Monthly Income 11](#_1mxgyo8mc63b)

[4.2.3. Relation between Performance Rating and Income Group 12](#_r1odrzw7crjl)

[4.2.4. Relation between frequency of Business Travel and Monthly Income 13](#_6ziguedlie0a)

[4.2.5. Relation between Job Satisfaction and Monthly Income 14](#_y0qxygmxa9wu)

[4.2.6. The more important factors affecting Monthly Income 15](#_fwhgffv50m29)

[**5. Conclusion & Discussion 16**](#_ocgqgkb0q4g6)

[**6. Appendix 18**](#_hw58bj2eo6nj)

[**7. References 33**](#_5nyqifbdd6k)

# 1. Introduction

In the dynamic landscape of modern business, the efficacy of an organisation's human resources (HR) strategies is fundamental to its success. Recognising the pivotal role that a diverse and well-managed workforce plays, we embarked on a data analysis to garner insights into various aspects of human resources within the organisation.

In our project, a dataset containing the salaries of employees from a company is used, with other variables such as department, education and job satisfaction. Based on this dataset, we seek to answer the following questions about the factors that could affect the monthly income of an individual:

1. Is there a significant difference in the monthly income between employees who work overtime and those who do not?
2. What is the nature of the relationship between an employee's age and their monthly income?
3. Is there a significant difference in the monthly income between employees with different performance ratings?
4. Does the frequency of business travel affect the monthly income earned within each department?
5. How does the level of job satisfaction affect the monthly income of employees?
6. What are the more important factors affecting the monthly income of employees?

This report will cover the data descriptions and analysis using R language. For each of our research objectives, we performed statistical analysis and concluded the most appropriate approach, together with explanations and elaborations.

# 

# 2. Data Description

The dataset, titled “HR Analytics Dataset” is obtained from Kaggle, and seeks to offer an understanding of different facets of organisational HR management.

Before proceeding to data analysis, we perform preliminary data cleaning and manipulation, including:

* Selected relevant columns by removing unwanted ones.
* Normalised certain categorical data entries for consistency (e.g. 'TravelRarely' to 'Travel\_Rarely',’ Non-Travel’ to ‘Non\_Travel’ in the BusinessTravel column.).
* Convert specified columns to categorical data types using the factor() function.

We then conducted a monthly income data cleaning per age group.

* Filtered the salary data by age groups and selected relevant columns.
* Sorted the Monthly Income within each age group and used headTail() to examine the top and bottom incomes.
* Identified a trend across all age groups of potential underreporting at the lower end of Monthly Income, leading to a decision to trim the bottom 10% of income data for each age group.
* Created and applied a function (remove\_bottom\_10perc) to remove the bottom 10th percentile of Monthly Income data from each age group.

After all the preparation, 1329 observations (employees) with 13 variables were retained for analysis:

1. Age: Age of employee
2. AgeGroup: Age group to which the employee belongs
3. BusinessTravel: Frequency of business travel for the employee
4. Department: Department in which the employee works
5. DistanceFromHome: Distance in miles from the employee's home to the workplace
6. JobSatisfaction: Employee's satisfaction level with their job
7. MonthlyIncome: Monthly income of the employee
8. NumCompaniesWorked: Number of companies the employee has worked for in the past
9. OverTime: Whether the employee works overtime or not
10. PerformanceRating: Performance rating of the employee
11. TrainingTimesLastYear: Number of training sessions attended by the employee in the last year
12. YearsAtCompany: Number of years the employee has worked at the current company
13. YearsInCurrentRole: Number of years the employee has been in the current role

# 3. Description and Cleaning of the Dataset

## 3.1. Summary statistics for the main variable of interest, Monthly Income

## The following plots show the overall distribution of the variable *MonthlyIncome.*

|  |  |  |
| --- | --- | --- |

The variable MonthlyIncome seems highly right-skewed, as expected of salary data, which follows a log-normal distribution. Hence, we apply a log-transformation (base *e*). We noticed that the histogram of the log-transformed variable has a considerable left tail. Upon further inspection, we noticed a common trend between all age groups, where roughly the bottom 10 percentile of employees had a very low monthly income (e.g. bottom 20% of employees aged 18-25 were earning <$2000 per month, bottom 15% of employees aged 55+ were earning <$2500 per month), which we deemed unreasonable. Therefore, we trimmed the bottom 10 percentile of *MonthlySalary* for each age group.

The histogram and boxplot of the log-transformed *MonthlyIncome* variable, along with its summary statistics are shown below.

|  | |  |
| --- | --- | --- |
|  | | |

The resulting data is now more symmetrical, and with a large sample size (n = 1329), we can safely assume normality conditions for the *MonthlyIncome* variable.

## 3.2. Summary statistics for the other variables

### **3.2.1. Age of employees, Age**

|  |  |
| --- | --- |
| From the histogram, the age of the employees seems normally distributed. Plotting the box plot provides further evidence of a normal distribution, due to its symmetry. |

### **3.2.2. Age group of employees, AgeGroup**

|  | * No outlying values were removed * As expected, most employees are in their late 20s to early 30s, followed by those aged 36-45. * Employees aged 55+ constitute the lowest numbers, potentially due to nearing retirement age. |
| --- | --- |

### **3.2.3. Whether Employees worked overtime, OverTime**

|  | * No outlying values were removed * It appears that the number of employees who do not work overtime (949) far exceeds those who do (380). |
| --- | --- |

### 

### **3.2.4. Performance Level of Employees, PerformanceRating**

|  | * No outlying values were removed * As the dataset only provided performance ratings of 3 or 4, we assumed only 2 levels to this variable. * Most employees seem to receive a lower performance rating, and only a relatively small number received a higher performance rating |
| --- | --- |

### **3.2.5. Business Travel Frequency of Employees, BusinessTravel**

|  | * No outlying values were removed * It seems that most employees rarely travel |
| --- | --- |

### 

### **3.2.6. Job Satisfaction Level of Employees, JobSatisfaction**

|  | * No outlying values were removed. * Employees seem to have higher Job Satisfaction. |
| --- | --- |

### **3.2.7. Departments of Employees, Department**

|  | * No outlying values were removed. * Employees seem to be in the Research & Development Department. |
| --- | --- |

### **3.2.8. Number of companies worked, NumCompaniesWorked**

|  |  | * No outlying values were removed. * The sqrt transformation was applied due to 0 values |
| --- | --- | --- |

### **3.2.9. Distance of workplace from Employees’ Homes, DistanceFromHome**

|  |  | * No outlying values were removed. * The sqrt transformation was applied due to 0 values |
| --- | --- | --- |

### **3.2.10. Number of times employees trained last year, TrainingTimesLastYear**

|  |  | * No outlying values were removed. |
| --- | --- | --- |

### **3.2.11. Number of years the employee been with the company, YearsAtCompany**

|  |  | * No outlying values were removed. * The sqrt transformation was applied due to 0 values |
| --- | --- | --- |

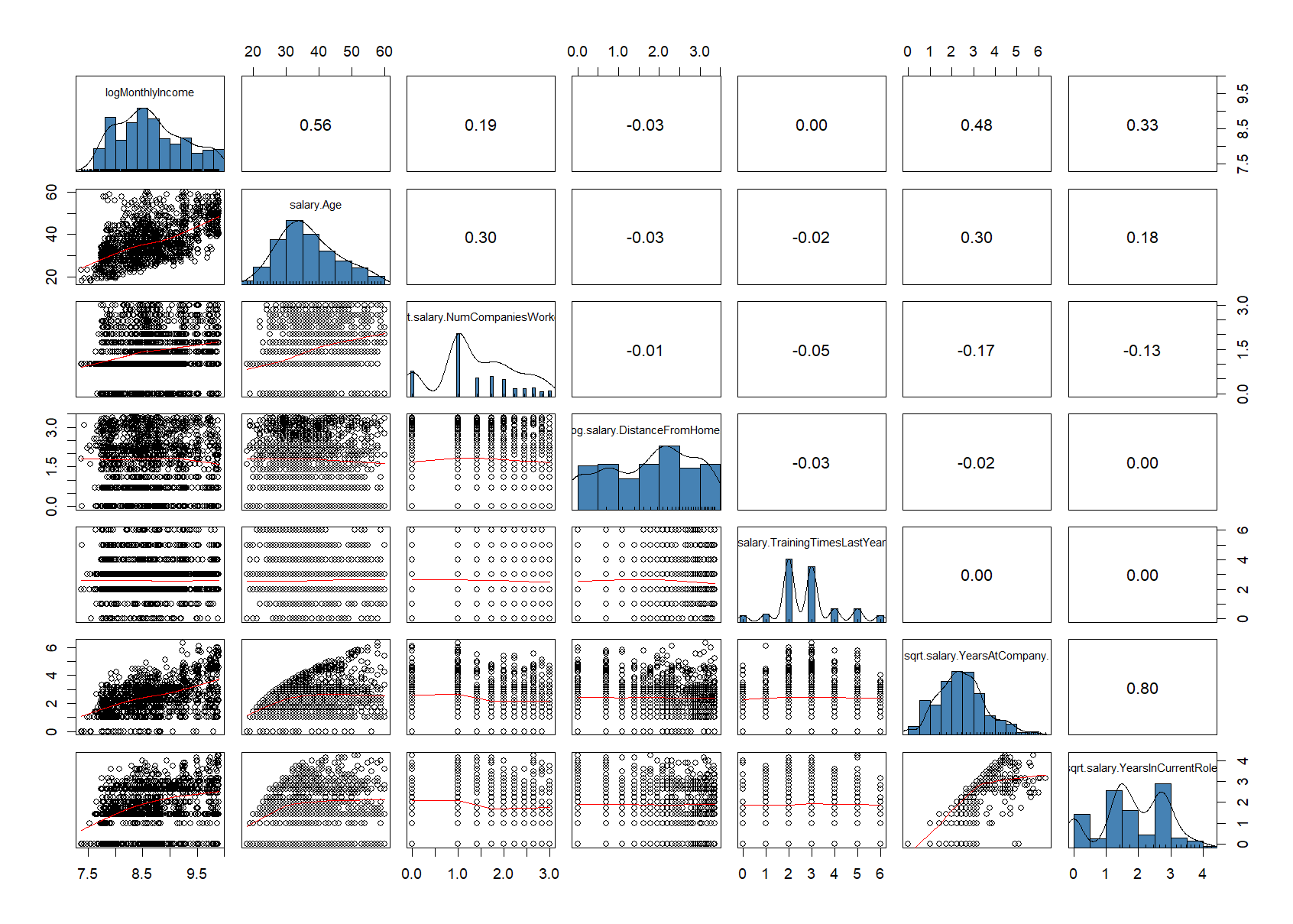
### **3.2.12. Number of years the employee been in the current role, YearsInCurrentRole**

|  |  | * No outlying values were removed. * The sqrt transformation was applied due to 0 values |
| --- | --- | --- |

# 

# 4. Statistical Analysis

## 4.1. Correlation between log(MonthlyIncome) and other continuous variables



Scatter plots and correlation coefficients are useful in studying the possible linear relationships between an employee’s salary and other numerical indicators.

From the plots, it appears that log(MonthlyIncome) is moderately correlated to Age, sqrt(YearsAtCompany) and sqrt(YearsInCurrentRole) than to other variables.

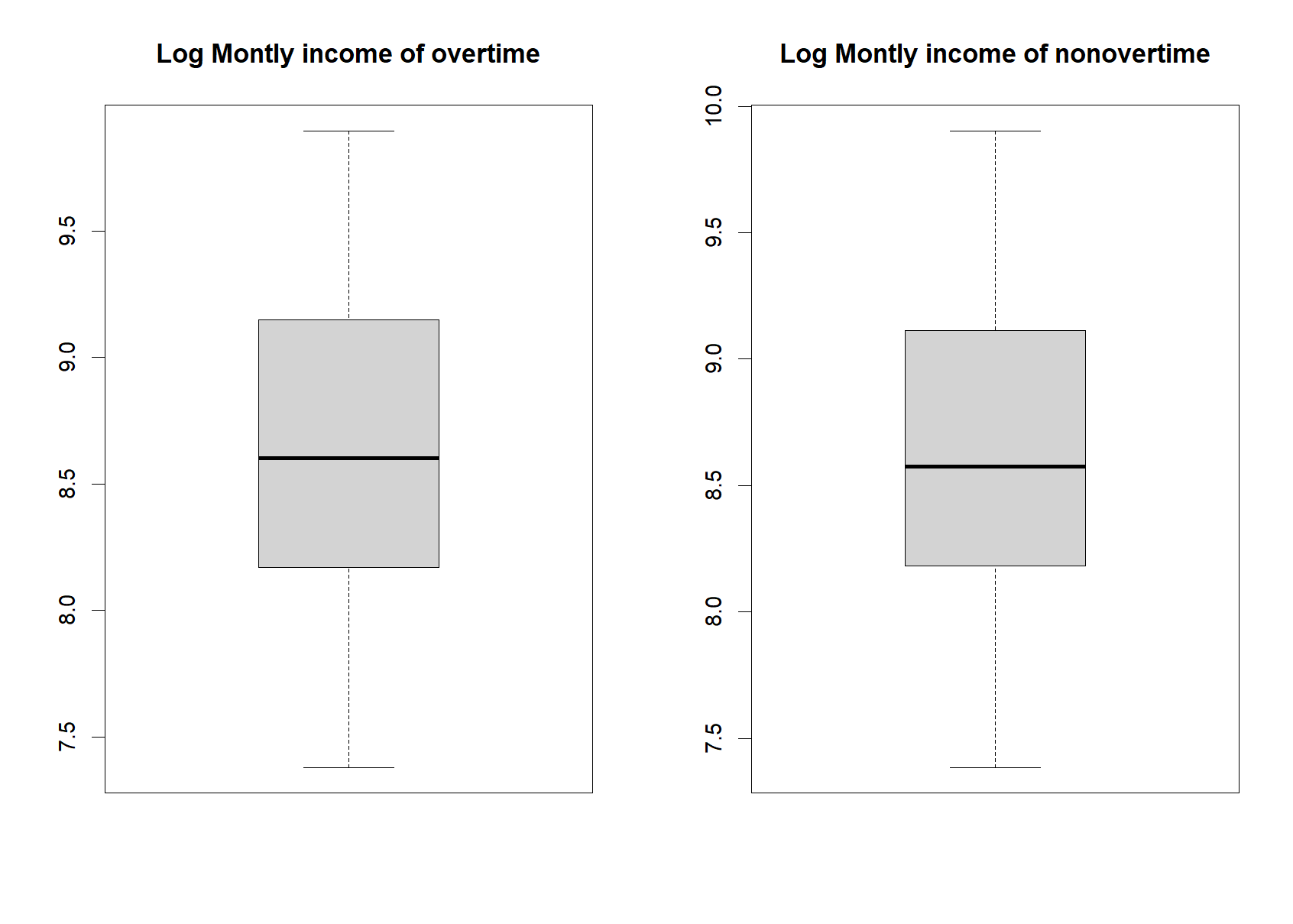
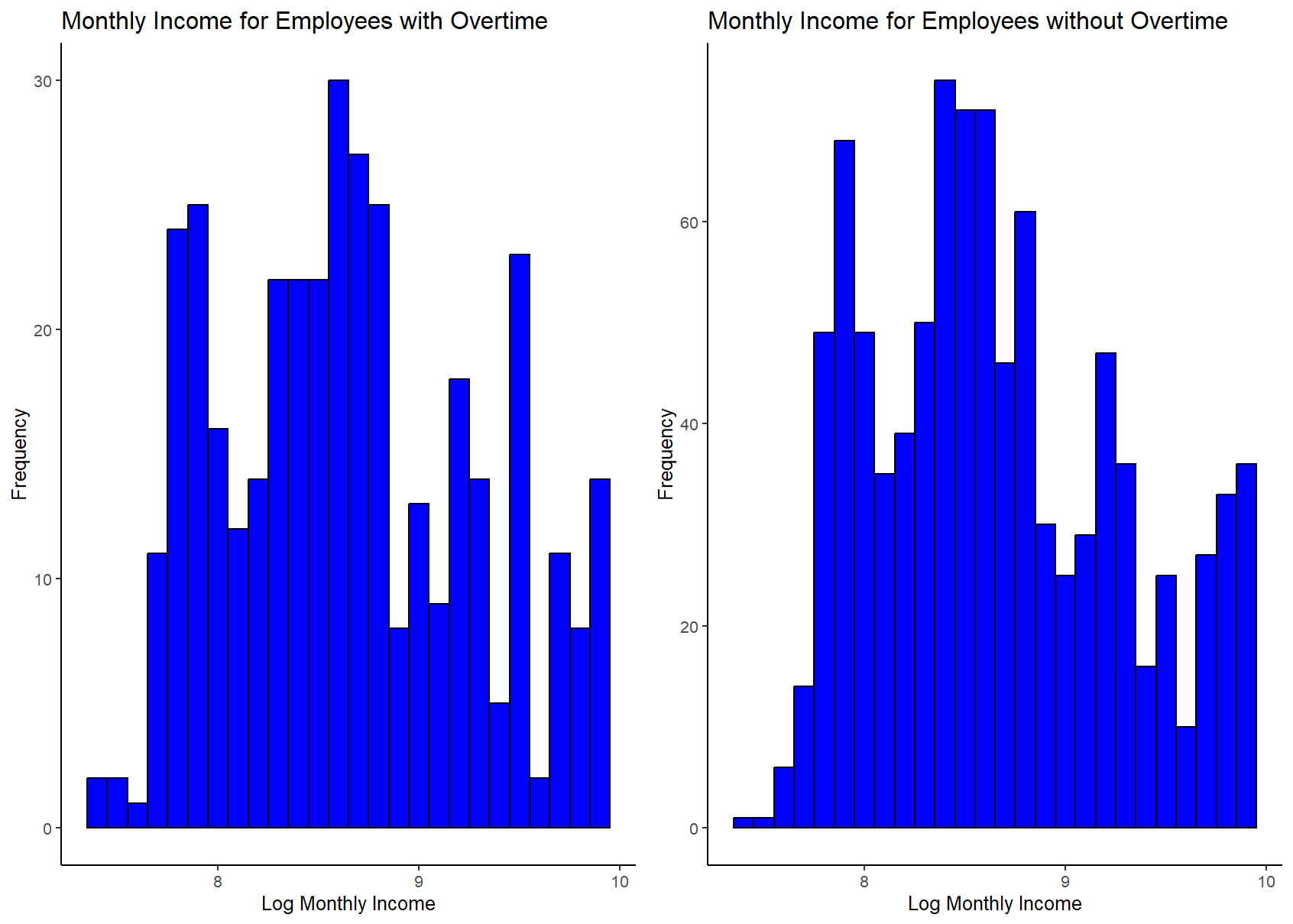
We shall perform some statistical tests to confirm some of our observations in the next section.

## 4.2 Statistical Tests

## 4.2.1. Relation between Working Overtime and Monthly Income

We want to determine if there is a statistically significant difference in the log(MonthlyIncome) of employees who work Overtime compared to those who do not work overtime. The cleaned and trimmed ‘salary’ data was segmented into two groups based on the *OverTime* variable. The ‘overtime’ data set includes employees who work overtime, while the ‘nonovertime’ data set includes those who do not.

Histograms and boxplots were plotted for both groups to visually assess the distribution of log(MonthlyIncome).



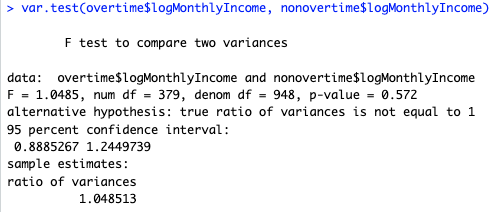
It was observed that both histograms and boxplots looked almost identical, suggesting no apparent difference in the distribution of log(MonthlyIncome) between employees who work overtime and employees who do not.

Variance test

H0: Variances of log(MonthlyIncome) of employees who work overtime and those who do not are equal;

H1: Variances of log(MonthlyIncome) of employees who work overtime and those who do not are not equal;

A variance test was conducted to check if the two samples had equal variances. The p-value of 0.572 > 0.05 indicated that there is not enough evidence to reject the null hypothesis of equal variances at the 5% significance level.

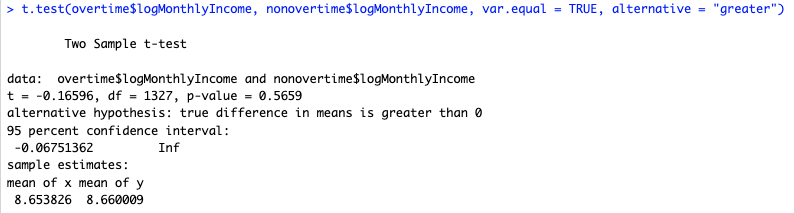


T-test

H0: The mean of log(MonthlyIncome) of employees who work overtime and those who do not are equal;

H1: The mean of log(MonthlyIncome) of employees who work overtime and those who do not are not equal;

We then conducted a one-sided t-test with equal variances to test the hypothesis. The p-value of 0.5659 > 0.05 indicated that there is not enough evidence to reject the null hypothesis of a significant difference between the average monthly income of those who work overtime and those who do not at the 5% significance level.

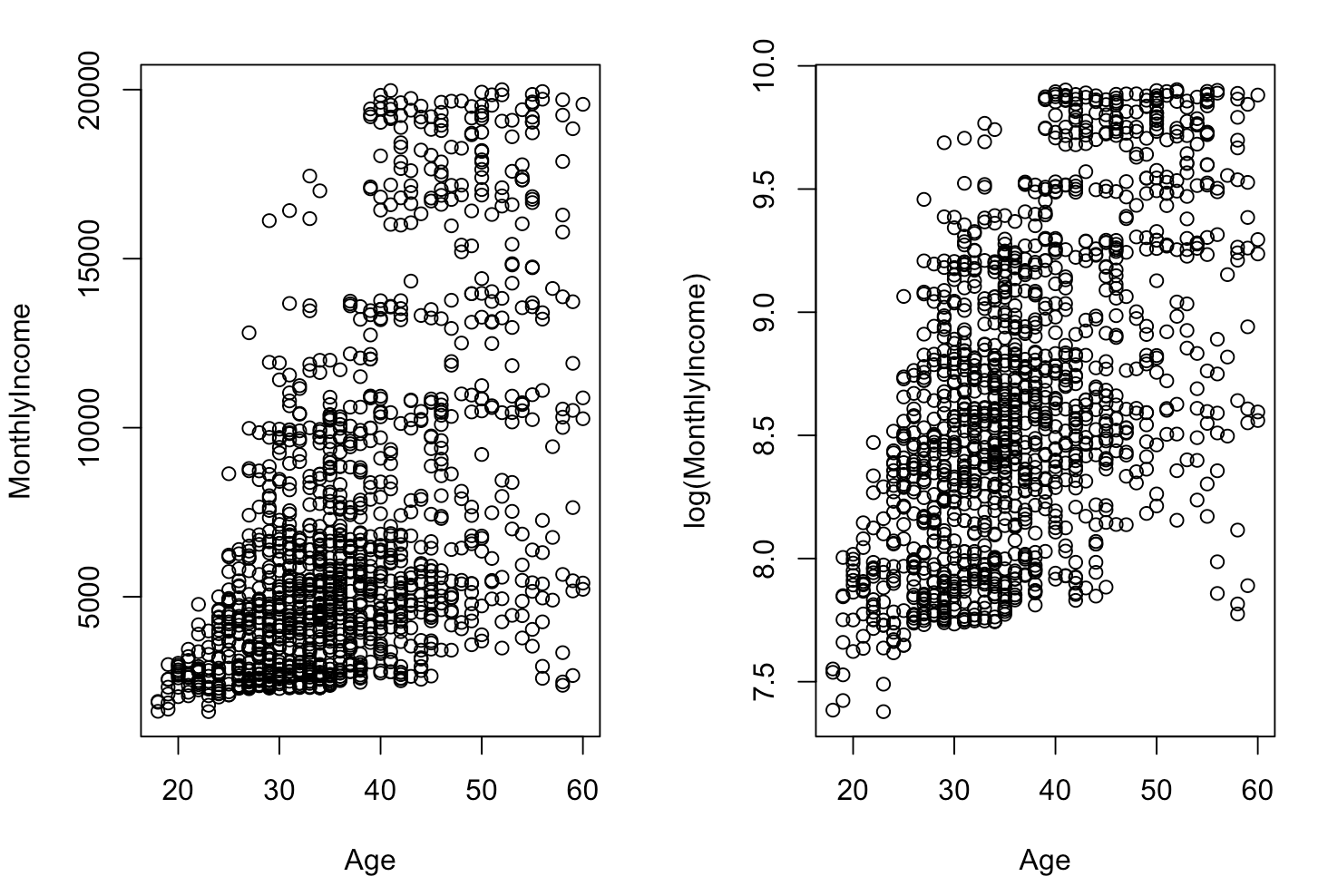


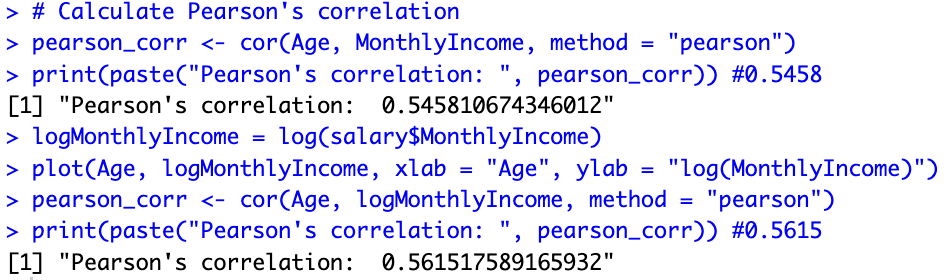
Therefore, we can conclude that the log(MonthlyIncome) of an employee is independent of whether they work overtime.

## 4.2.2. Correlation between Age and Monthly Income

We want to determine the nature of the relationship between employees' age and their Monthly Income.

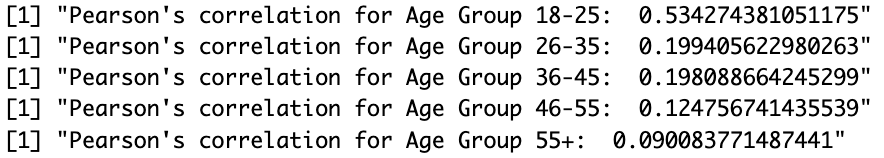
A scatter plot was generated to visualise the relationship between *Age* and *Monthly Income*. The Pearson correlation coefficient was calculated to be 0.5458, indicating a moderate positive correlation between Age and Monthly Income. This correlation became slightly stronger (0.5615) after transforming Monthly Income using the natural logarithm, suggesting that the relationship might be exponential rather than strictly linear.





Age Group-Specific Correlation

The dataset was further divided into age groups, and the correlation between age and log-transformed monthly income within each group was analysed:



Age Group 18-25: A Pearson correlation of 0.534 was observed, indicating a moderate positive correlation in this youngest age group. A similar code was used for the rest of the age groups.

Age Group 26-35: The correlation dropped significantly to 0.199, suggesting a weak positive correlation for this group.

Age Group 36-45: A weak correlation of 0.198 was observed.

Age Group 46-55: The correlation decreased further to 0.125.

Age Group 55+: The correlation is the lowest in this group at 0.090, suggesting an almost negligible relationship.

Therefore, the data suggests that the correlation between age and monthly income is most pronounced in the youngest group of employees and diminishes with increasing age categories. The strongest correlation in the youngest age group might reflect career progression and salary increments that come with initial years of experience.

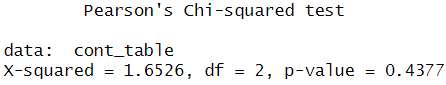
However, as employees age, this progression seems to plateau, which is evidenced by the declining correlation coefficients in the older age groups. The logarithmic transformation of *MonthlyIncome* also indicates that salary increases might not be additive with age but could potentially follow a multiplicative or exponential pattern, especially in the early career stages.

## 4.2.3. Relation between Performance Rating and Income Group

We analyse to identify if there is any significant relationship between Performance Rating and Income Group. As our main variable of interest, *MonthlyIncome* is a continuous variable, we created a new categorical variable, *IncomeGroup* to segregate employees into “Low”, “Medium” or “High” income groups by splitting the range of income into 3. Subsequently, we constructed a 2-way contingency table of our variables and performed a chi-square test based on the following hypotheses:

H0: There is no association between Performance Rating and Income Group.

H1: Some association exists between Performance Rating and Income Group.



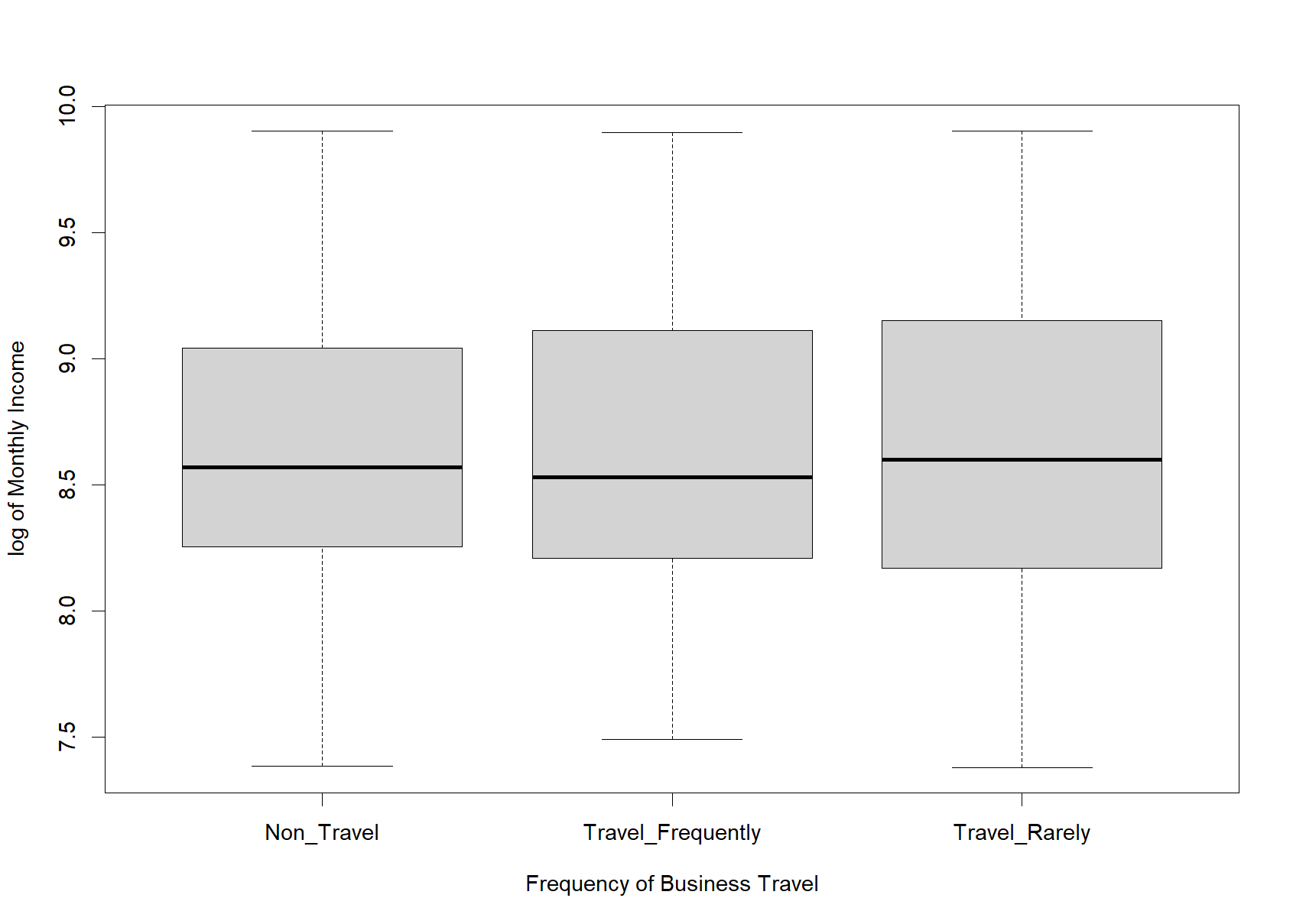
Based on our test results, as p-value = 0.4377 > 0.05, we do not reject and there is insufficient evidence to conclude that there is a significant association between the performance ratings (3 and 4) and the levels of income (Low, Medium, High).

It's crucial to note that the performance rating in this scenario seems like it should range from 1 to 5. However, the data collected only reflects ratings of 3 and 4. This skew towards the middle may be due to people's reluctance to rate individuals lower than 3 or higher than 4, possibly out of a desire to avoid being overly critical or generous. This imbalance in the ratings could significantly impact the relationship between performance ratings and income levels, as it limits the range of data available for analysis and may not accurately reflect the full spectrum of performance.

## 4.2.4. Relation between frequency of Business Travel and Monthly Income

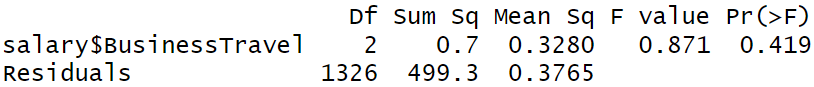
In this section, we try to answer “Does business travel frequency affect the monthly income earned within each department?”

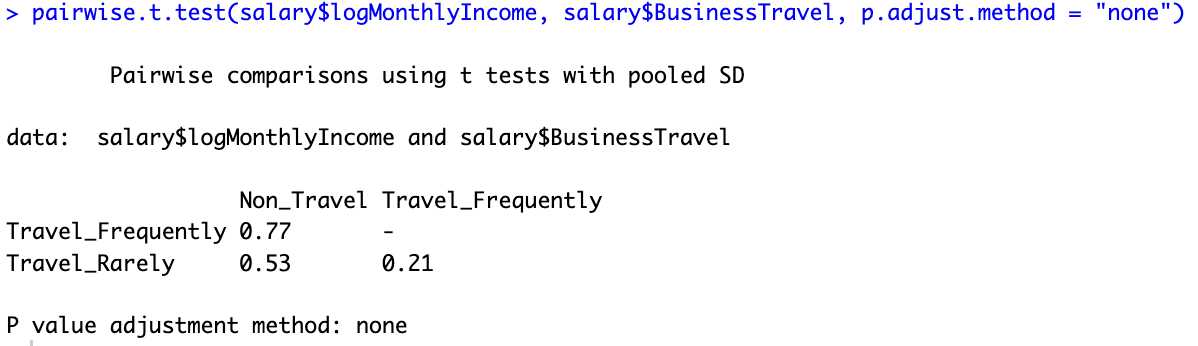
An analysis of variance (ANOVA) test will be conducted to determine whether Monthly Income is different at each frequency of Business Travel, since *BusinessTravel* is a categorical variable. The following plot illustrates the distributions of log of MonthlyIncome among the frequencies of Business Travel.



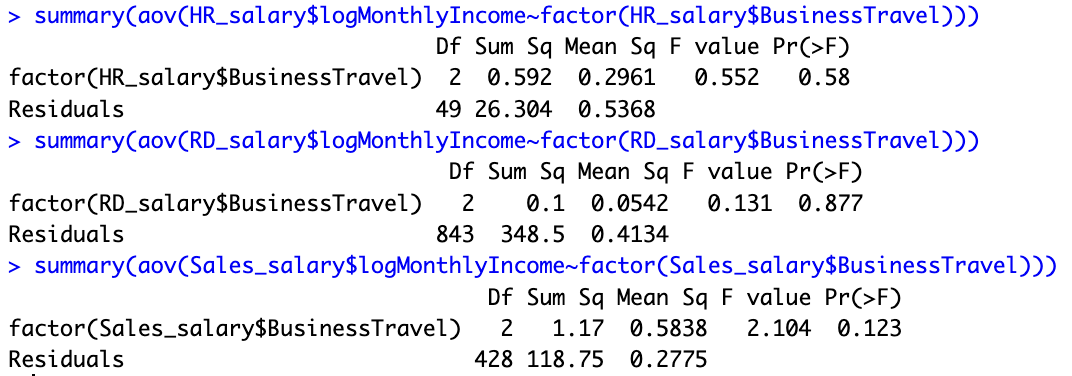
Looking at the boxplot, we see that the spread of log of Monthly Income varies for the frequency of Business Travel. We test,

H0: 𝜇NT = 𝜇TF = 𝜇TR against H1: not all 𝜇i are equal





The ANOVA test returns a p-value of 0.419, which shows that the means are not significantly different at a significance level of 0.05. The pairwise t-test also shows that there is no significant difference in the monthly income between each of the travel frequencies. Further analysis was conducted for relation between *log(MonthlyIncome)* and *BusinessTravel* within each department and it also shows that there is no relation between the two variables for all departments.

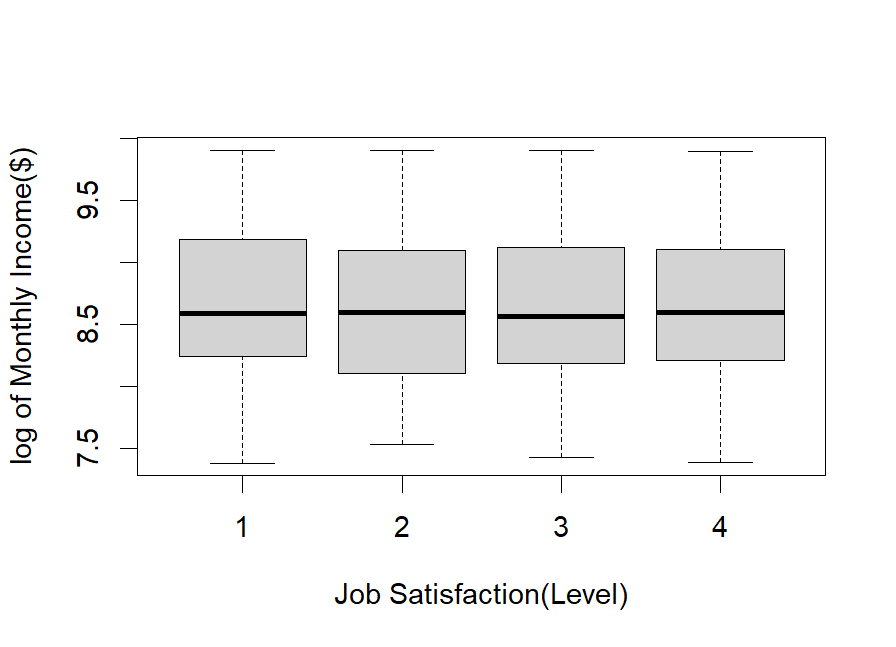


Therefore, we conclude that Monthly Income is independent of the frequency of Business Travel.

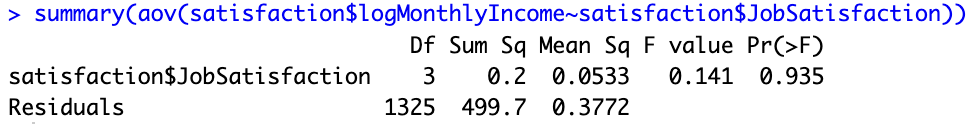
## 4.2.5. Relation between Job Satisfaction and Monthly Income

Job Satisfaction is categorised into 4 levels of satisfaction; with level 1 being the lowest and level 4 being the highest. Theoretically, having more satisfaction would mean that employees would be more willing to work harder and would seem more passionate in the eyes of their supervisors, which could lead to the employees being rewarded with higher monthly incomes.  
  
To see whether this claim is true, under the assumption that the data is normal, we use the ANOVA test. However, we can also use tests for multiple independent samples if we do not assume that the data is normal.

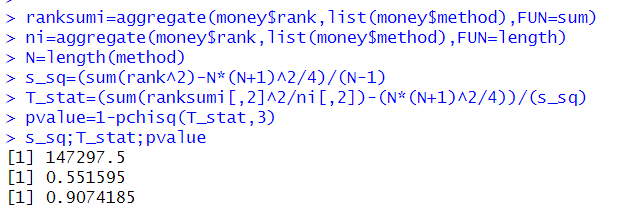
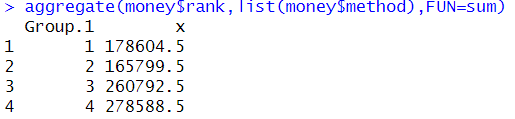
The following plot illustrates the distributions of log of Monthly Income among the levels of Job Satisfaction.



Looking at the boxplot, we see that the spread of log of MonthlyIncome are similar for all 4 levels of Job Satisfaction. We test,

H0: 𝜇1 = 𝜇2 = 𝜇3 = 𝜇4  against H1: not all 𝜇i are equal  


The test result gives a p-value of 0.907, so we do not reject H0 and accept that all the job satisfaction levels give identical means. Now we look at the usage of multiple independent sample test. We test,  
 H0: 𝜇1 = 𝜇2 = 𝜇3 = 𝜇4  against H1: not all 𝜇i are equal



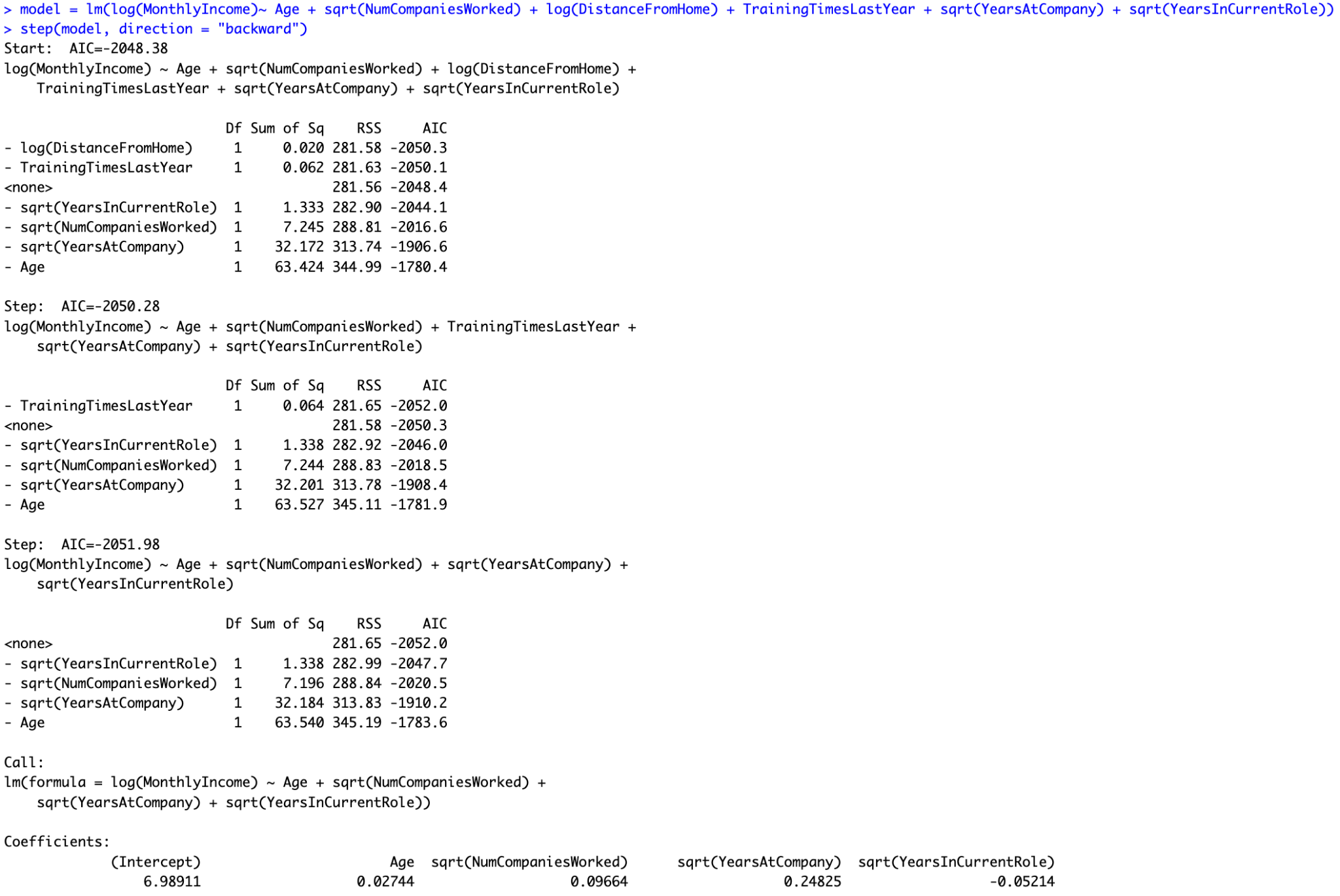
Similar to the result from the ANOVA test, we do not reject H0 since p-value = 0.907 > 0.05. Therefore, we can conclude that Job Satisfaction in fact does not affect Monthly Income.

## 4.2.6. The more important factors affecting Monthly Income

In this section, we attempt to build a multiple linear regression model for log(MonthlyIncome) based on the given measures, namely Age, sqrt(NumCompaniesWorked), log(DistanceFromHome), TrainingTimesLastYear, sqrt(YearsAtCompany) and sqrt(YearsInCurrentRole). We use a backward elimination method to select the most appropriate model. The result is shown in the R output below.

We conclude that Age, sqrt(NumCompaniesWorked), sqrt(YearsAtCompany) and sqrt(YearsInCurrentRole) are the significant measures that could be used to model log(MonthlyIncome). The fitted model is:

*log(MonthlyIncome)* = 6.99 + 0.0274\**Age* + 0.0966\**sqrt(NumCompaniesWorked)* + 0.248\**sqrt(YearsAtCompany)* - 0.052\**sqrt(YearsInCurrentRole)*



# 5. Conclusion & Discussion

Our analysis of this dataset has provided valuable information on the relationship between employees’ monthly income and other variables.

We conclude that:

* There is no significant difference in monthly income between employees who work overtime and those who do not
* The relationship between an employee's age and their monthly income is generally positive, with the strongest correlation among the younger employees aged 18 to 25 years.
* There is no significant difference in monthly income between employees with different performance ratings
* The frequency of business travel does not affect the monthly income earned within each department
* The level of job satisfaction does not affect the monthly income of employees
* Age, sqrt(NumCompaniesWorked), sqrt(YearsAtCompany) and sqrt(YearsInCurrentRole) are the more important variables in determining monthly income.

Based on the findings, the company should consider several key improvements:

Firstly, reviewing its overtime policies is essential, as working overtime does not significantly affect monthly income. Secondly, focusing on career development for younger employees could be beneficial, given the positive relationship between age and monthly income among this demographic. Thirdly, a review of the performance evaluation process may be necessary, as performance ratings were not found to impact monthly income significantly. Lastly, revisiting business travel policies to minimise costs and disruptions for employees, and enhancing overall employee well-being and job satisfaction could be beneficial.

Limitations of our report may include the reliance on a single dataset, which may not fully represent the diversity of factors influencing monthly income in the organisation. Additionally, the analysis is based on correlations and statistical associations, which do not imply causation. This means that while we can identify relationships between variables, we cannot definitively say that one variable causes changes in another. Finally, the data may be subject to errors or inconsistencies as we are unclear on the data-collecting methods, which could affect the validity of our conclusions.

# 6. Appendix

# 

# Load pacman and install required packages

if (!require("pacman")) install.packages("pacman")

pacman::p\_load(tidyverse, ggplot2, caret, corrplot, psych, stringr, ggpubr, forecast, gridExtra)

options(scipen=999) # surpress scientific notation

## Read dataset

salary = read.csv("HR\_Analytics.csv")

## Clean dataset

#========================================================= #

# Select relavant columns

salary = salary %>%

select(-c(1,4,6,9:18,20,22,23,25,27,29:32,34,37,38))

# cols remaining "Age" "AgeGroup" "BusinessTravel" "Department" "DistanceFromHome" "JobSatisfaction" "MonthlyIncome" "NumCompaniesWorked" "OverTime" "PerformanceRating" "TrainingTimesLastYear" "YearsAtCompany" "YearsInCurrentRole"

# Change (1) "TravelRarely" to "Travel\_Rarely", (2) "Non-Travel" to "Non\_Travel"

salary = salary %>%

mutate(BusinessTravel = ifelse(BusinessTravel == "TravelRarely", "Travel\_Rarely", BusinessTravel),

BusinessTravel = ifelse(BusinessTravel == "Non-Travel", "Non\_Travel", BusinessTravel))

# Change datatype to categorical data

salary = transform(salary,

AgeGroup = as.factor(AgeGroup),

BusinessTravel = as.factor(BusinessTravel),

Department = as.factor(Department),

JobSatisfaction = as.factor(JobSatisfaction),

OverTime = as.factor(OverTime),

PerformanceRating = as.factor(PerformanceRating)

)

## ---------- Cleaning Monthly Income data ---------- ##

# Age group 18-25

salary %>%

filter(AgeGroup=='18-25') %>%

select(MonthlyIncome, Department, Age) %>%

arrange(desc(MonthlyIncome)) %>%

headTail(top=5,bottom=25) # salary below 2000 seems to be a bit too low, department is R&D

# Age group 26-35

salary %>%

filter(AgeGroup=='26-35') %>%

select(MonthlyIncome, Department, Age) %>%

arrange(desc(MonthlyIncome)) %>%

headTail(top=50,bottom=25)

# Age group 36-45

salary %>%

filter(AgeGroup=='36-45') %>%

select(MonthlyIncome, Age, Department) %>%

arrange(desc(MonthlyIncome)) %>%

headTail(top=150,bottom=100)

# Age group 46-55

salary %>%

filter(AgeGroup=='46-55') %>%

select(MonthlyIncome, Age, Department) %>%

arrange(desc(MonthlyIncome)) %>%

headTail(top=5,bottom=25)

# Age group 55+

salary %>%

filter(AgeGroup=='55+') %>%

select(MonthlyIncome, Department) %>%

arrange(desc(MonthlyIncome)) %>%

headTail(top=5,bottom=25)

# Seems like there is a common trend across all age groups where there is a small portion of MonthlyIncome being underreported (maybe due to reporting error etc.)

# Hence we trim 10% off the bottom from each age group

# Function to remove bottom 10%ile of data from each age group

remove\_bottom\_10perc = function(df){

threshold = quantile(df$MonthlyIncome, 0.1)

df = df %>%

filter(MonthlyIncome > threshold)

return(df)

}

# Remove bottom 10%ile of data from each age group

salary = salary %>%

group\_by(AgeGroup) %>%

group\_modify(~ remove\_bottom\_10perc(.x)) %>%

ungroup()

## -------------------------------------------------- ##

str(salary)

glimpse(salary)

# End of Cleaning Dataset #

#========================================================== #

## Identify Categorical Data - 5 categorical data columns

# ========================================================== #

unique(salary$AgeGroup) # "18-25" "26-35" "36-45" "46-55" "55+"

unique(salary$BusinessTravel) # "Travel\_Rarely" "Travel\_Frequently" "Non\_Travel"

unique(salary$Department) # "Research & Development" "Sales" "Human Resources"

unique(salary$OverTime) # "No" "Yes"

unique(salary$PerformanceRating) # 3 4

========================================================== #

## Identify Numerical Data - 7 numerical data columns

# ========================================================== #

salary$Age

salary$DistanceFromHome

salary$MonthlyIncome

salary$NumCompaniesWorked

salary$TrainingTimesLastYear

salary$YearsAtCompany

salary$YearsInCurrentRole

# ========================================================== #

## Summary Statistics

# ========================================================== #

# ===== MAIN VARIABLE OF INTEREST - Monthly Income ===== #

hist(salary$MonthlyIncome, main = "Monthly Income", xlab = "Monthly Income", ylab = "Frequency", breaks = 20)

# Data is highly right skewed, so we take the log transformation

logMonthlyIncome = log(salary$MonthlyIncome)

hist(logMonthlyIncome, main = "log of Monthly Income", xlab = "Monthly Income", ylab = "Frequency")

# Closer to normal, with a second peak to the left of the median

# Add logMonthlyIncome to salary df

salary = salary %>% mutate(logMonthlyIncome = logMonthlyIncome)

# ---------- Checking normality ---------- #

# 1) Impose Normal pdf

xpt = seq(min(logMonthlyIncome),max(logMonthlyIncome),by=0.1) # start, end, interval

n\_den = dnorm(xpt, mean(logMonthlyIncome), sd(logMonthlyIncome)) # normal density w mean and sd of the data

ypt = n\_den \* length(logMonthlyIncome) \* 0.2 # adjust the normal density to the length of the dataset \* width of each bin (0.2 in this case)

lines(xpt, ypt, col='blue')

# 2) Shapiro test

shapiro.test(logMonthlyIncome) # p-value < 0.00000000000000022

# 3) Observe QQplot

qqnorm(logMonthlyIncome)

qqline(logMonthlyIncome) # Data look skewed

# 4) Observe boxplot

boxplot(logMonthlyIncome, main = "Boxplot of log(MonthlyIncome)")

# Whiskers length look similar, and plot is quite symmetrical around the median

# ---------- Outliers ---------- #

# Using classical method

sum(abs(logMonthlyIncome-mean(logMonthlyIncome)) > 2\*sd(logMonthlyIncome)) # 16 outliers

# Using boxplot rule

sum(logMonthlyIncome < quantile(logMonthlyIncome, 0.25) - 1.5\*IQR(logMonthlyIncome) | logMonthlyIncome > quantile(logMonthlyIncome, 0.75) + 1.5\*IQR(logMonthlyIncome)) # 0 outliers

# Here we follow the boxplot rule and determine that there are no more outliers.

# Boxplot rule because it is more lenient, giving leeway to possible individuals who may have much higher salary than expected

# ------------------------------ #

## Conclusion ##

# Even though the data is still slightly skewed after taking a log transformation, with the QQ plot and shapiro test indicating non-normality, we relieve some exceptions for normality due to the large sample size of our data (n = 1329), and also because of the symmetry of the boxplot.

# Proceed with this trimmed dataset.

#==========================================================#

## ===== Variables for Statistical Analysis ===== ##

## 1) Age of employees (Numerical)

# Set bin width and count number of observations

bw = 3.5

n\_obs = sum(!is.na(salary$Age))

# Plot histogram with normal pdf imposed

g = ggplot(salary, aes(Age)) +

geom\_histogram(aes(y = ..density..), binwidth = bw, colour = "black", fill="grey", na.rm = T) +

stat\_function(fun = dnorm, args = list(mean = mean(salary$Age), sd = sd(salary$Age))) +

theme\_classic() +

labs(title = "Distribution of Age")

ybreaks = seq(0,200,50)

# Rescale y axis

g + scale\_y\_continuous("Number of Employees", breaks = round(ybreaks / (bw \* n\_obs),3), labels = ybreaks)

ggplot(salary, aes(y=Age)) +

stat\_boxplot(geom = "errorbar", width = 0.15) +

geom\_boxplot(fill="grey") +

theme\_classic() +

theme(axis.ticks.x = element\_blank(), # Remove the xticks

axis.text.x = element\_blank(),

axis.title.y = element\_blank()) +

labs(title = "Age")

## 2) Age group of employees (Categorical)

AgeGroup\_df = data.frame(AgeGroup = salary$AgeGroup)

ggplot(AgeGroup\_df, aes(x = AgeGroup)) +

geom\_bar(fill="grey") +

theme\_classic() +

theme(panel.grid = element\_blank()) +

labs(title = "Distribution of different Age Groups", x = "Age Group", y = "Number of Employees")

## 3) Overtime (Categorical)

OverTime\_df = data.frame(OverTime = salary$OverTime)

ggplot(OverTime\_df, aes(x = OverTime)) +

geom\_bar(fill="grey") +

theme\_classic() +

theme(panel.grid = element\_blank()) +

labs(title = "Distribution of Overtime", x = "Overtime", y = "Number of Employees")

## 4) Performance rating (Categorical)

PerformanceRating\_df = data.frame(PerformanceRating = salary$PerformanceRating)

ggplot(PerformanceRating\_df, aes(x = PerformanceRating)) +

geom\_bar(fill="grey") +

theme\_classic() +

theme(panel.grid = element\_blank()) +

labs(title = "Distribution of Performance Rating", x = "Performance Rating", y = "Number of Employees")

## 5) Business Travel frequency (Categorical)

BusinessTravel\_df = data.frame(BusinessTravel = salary$BusinessTravel)

positions = c('Non\_Travel', 'Travel\_Rarely', 'Travel\_Frequently')

ggplot(BusinessTravel\_df, aes(x = BusinessTravel)) +

geom\_bar(fill="grey") +

theme\_classic() +

theme(panel.grid = element\_blank()) +

labs(title = "Distribution of Travel Frequency", x = "Travel Frequency", y = "Number of Employees") +

scale\_x\_discrete(limits = positions)

## 6) Job Satisfaction level (Categorical)

JobSatisfaction\_df = data.frame(JobSatisfaction = salary$JobSatisfaction)

ggplot(JobSatisfaction\_df, aes(x = JobSatisfaction)) +

geom\_bar(fill="grey") +

theme\_classic() +

theme(panel.grid = element\_blank()) +

labs(title = "Distribution of Job Satisfaction", x = "Job Satisfaction", y = "Number of Employees")

## 7) Department (Categorical)

Department\_df = data.frame(Department = salary$Department)

ggplot(Department\_df, aes(x = Department)) +

geom\_bar(fill="grey") +

theme\_classic() +

theme(panel.grid = element\_blank()) +

labs(title = "Distribution of Departments", x = "Department", y = "Number of Employees")

#=========== Other continuous variables ===========#

## 1) Number of companies worked

ggplot(salary, aes(NumCompaniesWorked)) +

geom\_histogram(binwidth = 2, colour = "black", fill="grey", na.rm = T, boundary=0) +

theme\_classic() +

labs(title = "Number of Companies Employees worked at before", x = "Number of Companies Worked", y = "Number of Employees") +

scale\_x\_continuous(breaks = function(x) unique(floor(pretty(seq(min(x), (max(x) + 1) \* 1.1)))))

ggplot(salary, aes(y=sqrt(NumCompaniesWorked))) +

stat\_boxplot(geom = "errorbar", width = 0.35) +

geom\_boxplot(fill="grey") +

theme\_classic() +

theme(axis.ticks.x = element\_blank(), # Remove the xticks

axis.text.x = element\_blank(),

axis.title.y = element\_blank()) +

labs(title = "sqrt(NumCompaniesWorked)")

## 2) Distance from home

ggplot(salary, aes(DistanceFromHome)) +

geom\_histogram(binwidth = 2, colour = "black", fill="grey", na.rm = T, boundary=0) +

theme\_classic() +

labs(title = "Distance of workplace from Employees' Homes", x = "Distance from Home", y = "Number of Employees") +

scale\_x\_continuous(breaks = function(x) unique(floor(pretty(seq(min(x), (max(x) + 1) \* 1.1)))))

ggplot(salary, aes(y=log(DistanceFromHome))) +

stat\_boxplot(geom = "errorbar", width = 0.35) +

geom\_boxplot(fill="grey") +

theme\_classic() +

theme(axis.ticks.x = element\_blank(), # Remove the xticks

axis.text.x = element\_blank(),

axis.title.y = element\_blank()) +

labs(title = "sqrt(DistanceFromHome)")

## 3) Training times last year

ggplot(salary, aes(TrainingTimesLastYear)) +

geom\_histogram(bins = 6, colour = "black", fill="grey", na.rm = T, boundary=0) +

theme\_classic() +

labs(title = "Number of times Employees underwent training last year", x = "Number of trainings", y = "Number of Employees") +

scale\_x\_continuous(breaks = function(x) unique(floor(pretty(seq(min(x), (max(x) + 1) \* 1.1)))))

ggplot(salary, aes(y=TrainingTimesLastYear)) +

stat\_boxplot(geom = "errorbar", width = 0.35) +

geom\_boxplot(fill="grey") +

theme\_classic() +

theme(axis.ticks.x = element\_blank(), # Remove the xticks

axis.text.x = element\_blank(),

axis.title.y = element\_blank()) +

labs(title = "TrainingTimesLastYear")

## 4) Years at company

ggplot(salary, aes(YearsAtCompany)) +

geom\_histogram(bins = 8, colour = "black", fill="grey", na.rm = T, boundary=0) +

theme\_classic() +

labs(title = "Number of years worked in the current company", x = "Number of years", y = "Number of Employees") +

scale\_x\_continuous(breaks = function(x) unique(floor(pretty(seq(min(x), (max(x) + 1) \* 1.1)))))

ggplot(salary, aes(y=sqrt(YearsAtCompany))) +

stat\_boxplot(geom = "errorbar", width = 0.35) +

geom\_boxplot(fill="grey") +

theme\_classic() +

theme(axis.ticks.x = element\_blank(), # Remove the xticks

axis.text.x = element\_blank(),

axis.title.y = element\_blank()) +

labs(title = "sqrt(YearsAtCompany)")

## 5) Years in current role

ggplot(salary, aes(YearsInCurrentRole)) +

geom\_histogram(bins = 8, colour = "black", fill="grey", na.rm = T, boundary=0) +

theme\_classic() +

labs(title = "Number of years working in the current role", x = "Number of years", y = "Number of Employees") +

scale\_x\_continuous(breaks = function(x) unique(floor(pretty(seq(min(x), (max(x) + 1) \* 1.1)))))

ggplot(salary, aes(y=sqrt(YearsInCurrentRole))) +

stat\_boxplot(geom = "errorbar", width = 0.35) +

geom\_boxplot(fill="grey") +

theme\_classic() +

theme(axis.ticks.x = element\_blank(), # Remove the xticks

axis.text.x = element\_blank(),

axis.title.y = element\_blank()) +

labs(title = "sqrt(YearsInCurrentRole)")

# End of Summary Statistics #

# ========================================================== #

## Statistical Analysis

# ========================================================== #

## Correlation plot of all the numerical data with their transformations

cont\_df = data.frame(

logMonthlyIncome,

salary$Age,

sqrt(salary$NumCompaniesWorked),

log(salary$DistanceFromHome),

salary$TrainingTimesLastYear,

sqrt(salary$YearsAtCompany),

sqrt(salary$YearsInCurrentRole)

)

corrplot(cor(cont\_df), type="upper", method="color", addCoef.col="black", number.cex=0.8)

pairs.panels(cont\_df, method = "pearson", hist.col = "steelblue",

pch = 21, density = TRUE, ellipses = FALSE ) # psych library

## ===== 4.2 Statistical Tests ===== ##

## 4.2.1 Relationship between working Overtime and Monthly Income

# 1. Conducting a hypothesis test on 2 samples

# let average MontlyIncome of people who work Overtime be (MuYes)

# let MontlyIncome of people who do not work Overtime (MuNo)

# H0: No difference between MuYes and MuNo

# H1: MuYes > MuNo (one sided)

overtime = salary %>% filter(OverTime == "Yes") %>% select(MonthlyIncome)

nonovertime = salary %>% filter(OverTime == "No") %>% select(MonthlyIncome)

# we observe that the number of people who work overtime (380) is much smaller

#than the number of people who does not work overtime (949)

# Applying log transformation to MonthlyIncome for both groups

overtime$logMonthlyIncome = log(overtime$MonthlyIncome)

nonovertime$logMonthlyIncome = log(nonovertime$MonthlyIncome)

#salary histogram of those who work overtime

# Histogram for employees with overtime

overtime\_plot\_log = ggplot(overtime, aes(x = logMonthlyIncome)) +

geom\_histogram(binwidth = 0.1,

fill = "blue",

color = "black") +

labs(title = "Histogram of Log Monthly Income for Employees with Overtime",

x = "Log Monthly Income",

y = "Frequency") +

theme\_minimal()

# Histogram for employees without overtime

nonovertime\_plot\_log = ggplot(nonovertime, aes(x = logMonthlyIncome)) +

geom\_histogram(binwidth = 0.1,

fill = "blue",

color = "black") +

labs(title = "Histogram of Log Monthly Income for Employees without Overtime",

x = "Log Monthly Income",

y = "Frequency") +

theme\_minimal()

# Arrange the log-transformed income plots side by side

grid.arrange(overtime\_plot\_log, nonovertime\_plot\_log, nrow = 1)

overtime\_income = overtime$MonthlyIncome

nonovertime\_income = nonovertime$MonthlyIncome

#comparing the boxplot

par(mfrow = c(1,2))

boxplot(overtime$logMonthlyIncome, main = "Log Monthly Income of Overtime employee",

ylab = "Log Monthly Income")

boxplot(nonovertime$logMonthlyIncome, main = "Log Monthly Income of Non-Overtime employee",

ylab = "Log Monthly Income")

# we check if they have equal variance

# H0: Equal variance

# H1: Unequal variance

var.test(overtime\_income, nonovertime\_income)

# p-value = 0.8323 > 0.05 we do not reject H0. The 2 samples have equal variance

# Checking if they have equal variance after log transformation

var.test(overtime$logMonthlyIncome, nonovertime$logMonthlyIncome)

# p-value= 0.572 > 0.05

# Conducting the t-test on log-transformed incomes

t.test(overtime$logMonthlyIncome, nonovertime$logMonthlyIncome, var.equal = TRUE, alternative = "greater")

# p-value = 0.5659

## 4.2.2 Correlation between Monthly Income and Age

MonthlyIncome = salary$MonthlyIncome

Age = salary$Age

plot(Age, MonthlyIncome, xlab = "Age", ylab = "MonthlyIncome")

cov(MonthlyIncome, Age)

var(MonthlyIncome)

var(Age)

r = cov(MonthlyIncome, Age) / sqrt(var(MonthlyIncome) \* var(Age))

r

# Calculate Pearson's correlation

pearson\_corr <- cor(Age, MonthlyIncome, method = "pearson")

print(paste("Pearson's correlation: ", pearson\_corr)) #0.5458

logMonthlyIncome = log(salary$MonthlyIncome)

plot(Age, logMonthlyIncome, xlab = "Age", ylab = "log(MonthlyIncome)")

pearson\_corr <- cor(Age, logMonthlyIncome, method = "pearson")

print(paste("Pearson's correlation: ", pearson\_corr)) #0.5615

#there is a moderate positive correlation between age and monthly income, and it appears slightly stronger when we take the logarithm of the monthly income.

# non-parametric measure of rank correlation

spearman\_corr <- cor(Age, MonthlyIncome, method = "spearman")

print(paste("Spearman's rank correlation: ", spearman\_corr)) #0.548

# we do according to the age group

# Age group 18-25

age1 = salary %>%

filter(AgeGroup=='18-25') %>%

select(MonthlyIncome, Age)

logage1Income = log(age1$MonthlyIncome)

age1Age = age1$Age

pearson\_corr\_age1 <- cor(age1Age, logage1Income, method = "pearson")

print(paste("Pearson's correlation for Age Group 18-25: ", pearson\_corr\_age1)) #0.534

# Age group 26-35

age2 = salary %>%

filter(AgeGroup=='26-35') %>%

select(MonthlyIncome, Age)

logage2Income = log(age2$MonthlyIncome)

age2Age = age2$Age

pearson\_corr\_age2 <- cor(age2Age, logage2Income, method = "pearson")

print(paste("Pearson's correlation for Age Group 26-35: ", pearson\_corr\_age2)) #0.199

# Age group 36-45

age3 = salary %>%

filter(AgeGroup=='36-45') %>%

select(MonthlyIncome, Age)

logage3Income = log(age3$MonthlyIncome)

age3Age = age3$Age

pearson\_corr\_age3<- cor(age3Age, logage3Income, method = "pearson")

print(paste("Pearson's correlation for Age Group 36-45: ", pearson\_corr\_age3)) #0.198

# Age group 46-55

age4 = salary %>%

filter(AgeGroup=='46-55') %>%

select(MonthlyIncome, Age)

logage4Income = log(age4$MonthlyIncome)

age4Age = age4$Age

pearson\_corr\_age4 <- cor(age4Age, logage4Income, method = "pearson")

print(paste("Pearson's correlation for Age Group 46-55: ", pearson\_corr\_age4)) #0.125

# Age group 55+

age5 = salary %>%

filter(AgeGroup=='55+') %>%

select(MonthlyIncome, Age)

logage5Income = log(age5$MonthlyIncome)

age5Age = age5$Age

pearson\_corr\_age5 <- cor(age5Age, logage5Income, method = "pearson")

print(paste("Pearson's correlation for Age Group 55+: ", pearson\_corr\_age5)) #0.090

print(paste("Pearson's correlation for Age Group 18-25: ", pearson\_corr\_age1)) ; print(paste("Pearson's correlation for Age Group 26-35: ", pearson\_corr\_age2)) ; print(paste("Pearson's correlation for Age Group 36-45: ", pearson\_corr\_age3)) ; print(paste("Pearson's correlation for Age Group 46-55: ", pearson\_corr\_age4)) ; print(paste("Pearson's correlation for Age Group 55+: ", pearson\_corr\_age5))

# We can see that Age correlates most with Monthly Income the most in the youngest age group, 18-25. In particular, most younger employees are earning in the lower income bracket. However, as they get older, the range of salaries that the employees earn start to increase.

## 4.2.3 Relationship between Performance Rating and Income Group

# Create income groups

range(logMonthlyIncome)

salary$IncomeGroup <- cut(logMonthlyIncome, breaks = 3, labels = c("Low", "Medium", "High"))

# Create a contingency table

cont\_table <- table(salary$IncomeGroup, salary$PerformanceRating)

colnames(cont\_table)=c("Rating:3", "Rating:4")

cont\_table

# Hypothesis Testing

# H0: There is no association between Performance Rating and Income Group.

# H1: Some association exists between Performance Rating and Income Group.

# Perform chi-square test

chisq\_test <- chisq.test(cont\_table)

chisq\_test

# p-value = 0.4377 > 0.05, do not reject H0.

# Hence, we conclude that there is no significant association between Performance Rating and Income Group

## 4.2.4 Relationship between Frequency of Business Travel and Monthly Income

# Boxplot of MonthlyIncome for each frequency of BusinessTravel

boxplot(logMonthlyIncome~salary$BusinessTravel,

col="light gray",

ylab="log of Monthly Income",

xlab="Frequency of Business Travel")

# ANOVA test

aov(salary$logMonthlyIncome~salary$BusinessTravel)

summary(aov(salary$logMonthlyIncome~salary$BusinessTravel))

# Pairwise T-test

pairwise.t.test(salary$logMonthlyIncome, salary$BusinessTravel, p.adjust.method = "none")

#Further analysis between departments

HR\_salary = salary[salary$Department=="Human Resources",]

boxplot(HR\_salary$logMonthlyIncome~factor(HR\_salary$BusinessTravel), col="light gray",

ylab="log of Monthly Income",xlab="Frequency of Business Travel",main="Human Resources")

summary(aov(HR\_salary$logMonthlyIncome~factor(HR\_salary$BusinessTravel)))

RD\_salary = salary[salary$Department=="Research & Development",]

boxplot(RD\_salary$logMonthlyIncome~factor(RD\_salary$BusinessTravel), col="light gray",

ylab="log of Monthly Income",xlab="Frequency of Business Travel",main="Research & Development")

summary(aov(RD\_salary$logMonthlyIncome~factor(RD\_salary$BusinessTravel)))

Sales\_salary = salary[salary$Department=="Sales",]

boxplot(Sales\_salary$logMonthlyIncome~factor(Sales\_salary$BusinessTravel), col="light gray",

ylab="log of Monthly Income",xlab="Frequency of Business Travel",main="Sales")

summary(aov(Sales\_salary$logMonthlyIncome~factor(Sales\_salary$BusinessTravel)))

## 4.2.5 Relationship between Job Satisfaction and Monthly Income

satisfaction = salary %>%

select(JobSatisfaction, logMonthlyIncome) %>%

arrange(JobSatisfaction)

# Perform tests with the following Hypotheses

# H0: Mean income of employees across different job satisfaction levels are identical

# H1: Not all of the means are identical

# Using parametric tests: ANOVA #

str(satisfaction)

boxplot(satisfaction$logMonthlyIncome ~ satisfaction$JobSatisfaction,

col="light gray",

ylab="log of Monthly Income($)",

xlab="Job Satisfaction(Level)")

aov(satisfaction$logMonthlyIncome~satisfaction$JobSatisfaction)

summary(aov(satisfaction$logMonthlyIncome~satisfaction$JobSatisfaction))

# p-value = 0.935 > 0.05, do not reject Null Hypothesis

# we conclude that job satisfaction levels do not impact mean income level of employees

# Now try using non-parametric test

method = satisfaction$JobSatisfaction

yield = satisfaction$logMonthlyIncome

rank = rank(yield, ties.method = "average")

money = data.frame(method, yield,rank)

head(money)

aggregate(money$rank,list(money$method),FUN=sum)

aggregate(money$rank,list(money$method),FUN=length)

ranksumi=aggregate(money$rank,list(money$method),FUN=sum)

ni=aggregate(money$rank,list(money$method),FUN=length)

N=length(method)

(s\_sq=(sum(rank^2)-N\*(N+1)^2/4)/(N-1)) # 147297.5

(T\_stat=(sum(ranksumi[,2]^2/ni[,2])-(N\*(N+1)^2/4))/(s\_sq)) # 0.551595

(pvalue=1-pchisq(T\_stat,3)) # 0.9074185

# P-value result using non-parametric test is also > 0.05 which results in the same conclusion to not reject Null Hypothesis.

##4.2.6 The more important factors affecting Monthly Income

## Multiple Linear Regression

model = lm(log(MonthlyIncome)~ Age + sqrt(NumCompaniesWorked) + log(DistanceFromHome) + TrainingTimesLastYear + sqrt(YearsAtCompany) + sqrt(YearsInCurrentRole))

step(model, direction = "backward")

# 

# 

# 7. References

Haroon. S. (2023, Oct) *HR Analytics Dataset.* <https://www.kaggle.com/datasets/saadharoon27/hr-analytics-dataset>