EBTM 350 Pilot project (Lab 9)

18 November, 2024

Background or Introduction

Labor market success has historically been linked with human capital aspects such as education and job training. There has also been interested in examining the impact of non-cognitive factors such as workers' sex, age, and personality trait. We will rely on the Net Promoter Score (NPS) to understand customer loyalty at Cassius Weatherby. Focusing on the NPS of their sales professionals to identify factors influencing these scores. What are the key variables that significantly influence the salary of sales representatives at Cassius Weatherby? What are the most determinants for predicting the salary? What is the gender gap in the salary of sales representatives? The insights from this research will guide hiring and training decisions at the company

Exploratory analysis

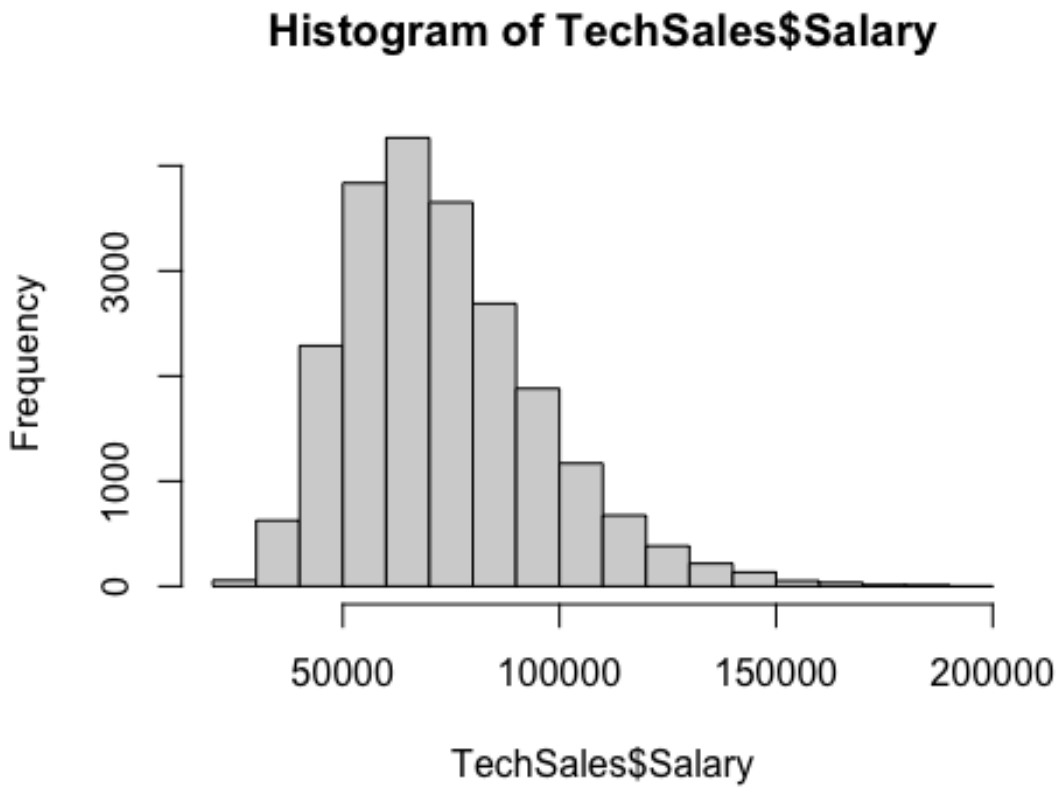
1. Exploring the range and distribution

```
# Loading the dataset
library(readxl)
TechSales <- read_excel("TechSales.xlsx")
summary(TechSales)
```

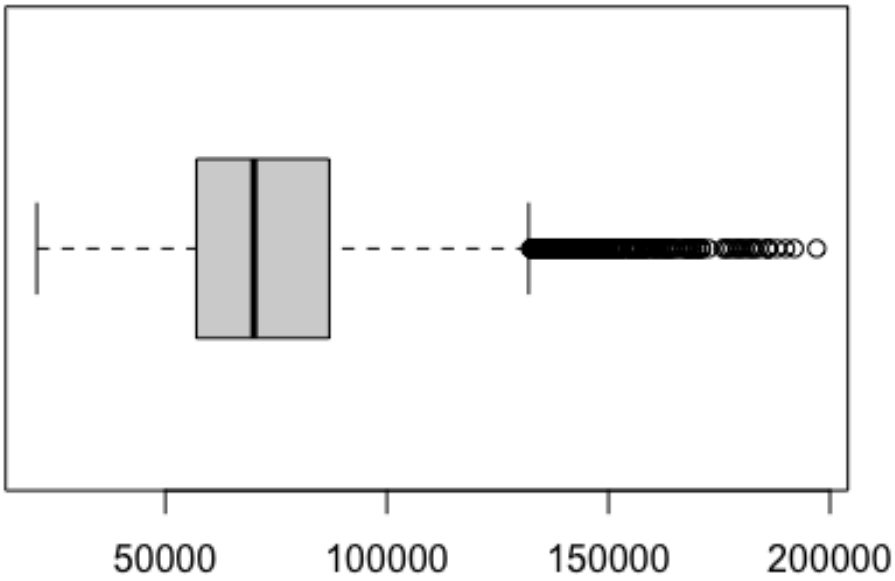
```
##    Sales_Rep       Business              Age           Female
##  Min.   :    1   Length:21990       Min.   :21.0   Min.   :0.000
##  1st Qu.: 5498   Class :character   1st Qu.:32.0   1st Qu.:0.000
##  Median :10996   Mode  :character   Median :41.0   Median :0.000
##  Mean   :10996                      Mean   :41.5   Mean   :0.383
##  3rd Qu.:16493                      3rd Qu.:51.0   3rd Qu.:1.000
##  Max.   :21990                      Max.   :65.0   Max.   :1.000
##      Years          College          Personality         Certficates
##  Min.   : 1.000   Length:21990       Length:21990       Min.   :0.000
##  1st Qu.: 1.000   Class :character   Class :character   1st Qu.:1.000
##  Median : 2.000   Mode  :character   Mode  :character   Median :2.000
##  Mean   : 2.646                                        Mean   :2.612
##  3rd Qu.: 2.000                                        3rd Qu.:4.000
##  Max.   :13.000                                        Max.   :6.000
##     Feedback         Salary            NPS
##  Min.   :1.080   Min.   : 21000   Min.   : 1.000
##  1st Qu.:1.990   1st Qu.: 57000   1st Qu.: 5.000
##  Median :2.660   Median : 70000   Median : 6.000
##  Mean   :2.665   Mean   : 73674   Mean   : 6.278
##  3rd Qu.:3.390   3rd Qu.: 87000   3rd Qu.: 8.000
##  Max.   :4.000   Max.   :197000   Max.   :10.000
```

The age range starts at the minimum of 21 years to a maximum of 65 years and the average age is 41.5 years. For the NPS, the third quartile (8) and the minimum are (1) which shows that,50% of scores are clustered within this range. The median salary is $70,000. The average salary is approximately $73,674, indicating a few higher salaries influencing the average. The first quartile is $57,000 and the third quartile is $87,000, suggesting that 50% of salaries fall within this range. The maximum salary also shows that you can earn up to $197,000.

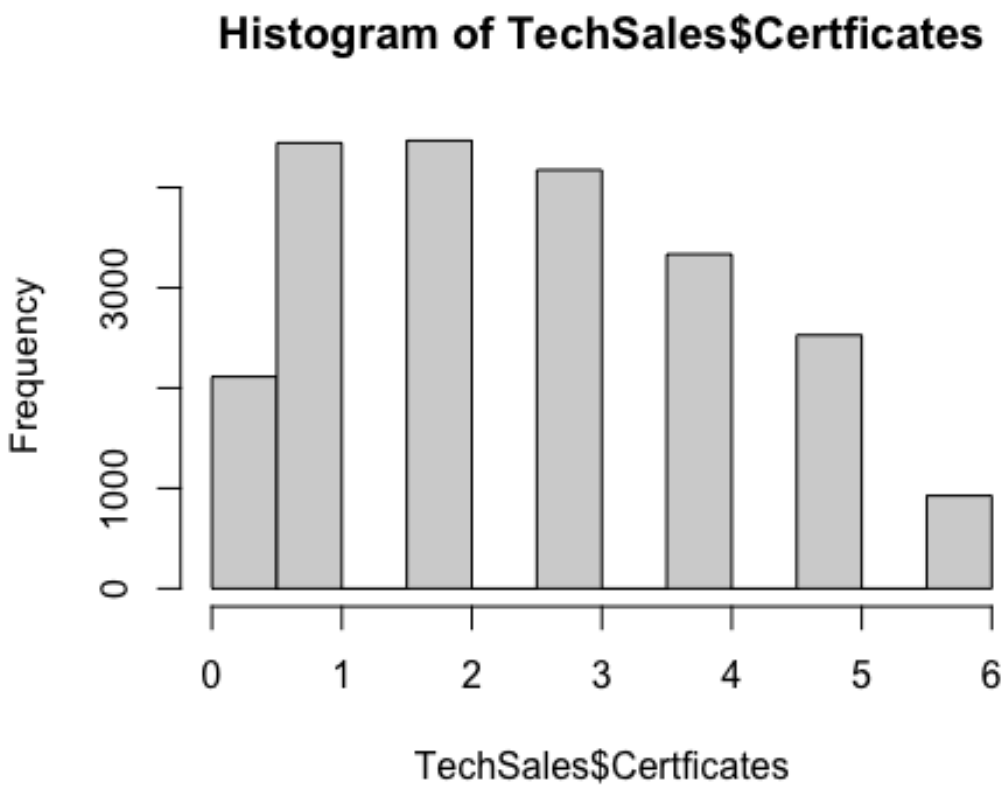
```
# Understand the distribution of these variables
hist(TechSales$Salary)
```



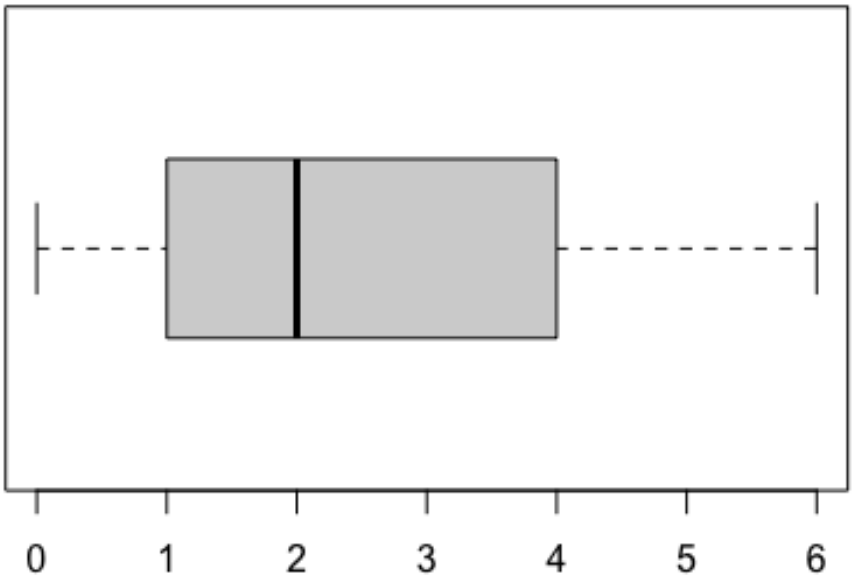

```
boxplot(TechSales$Salary, horizontal = TRUE)
```



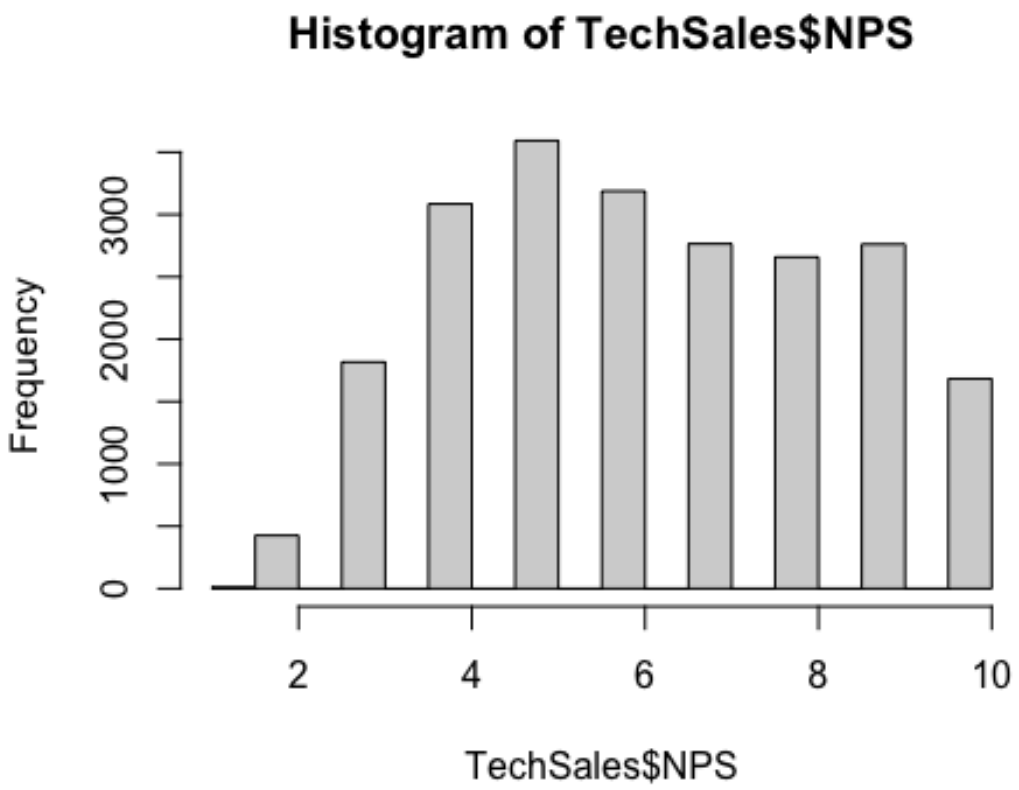

```r
# Try to explore other variables
hist(TechSales$Certficates)
```



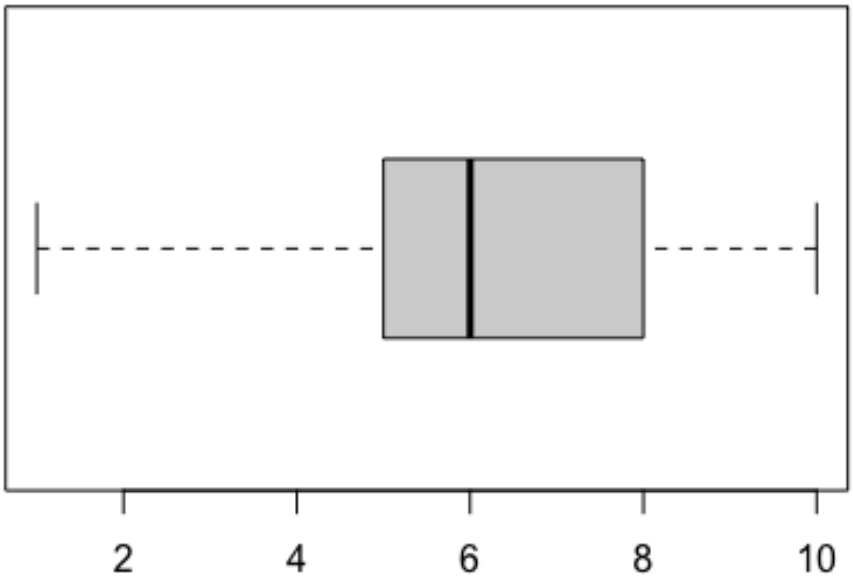

```r
boxplot(TechSales$Certficates, horizontal = TRUE)
```



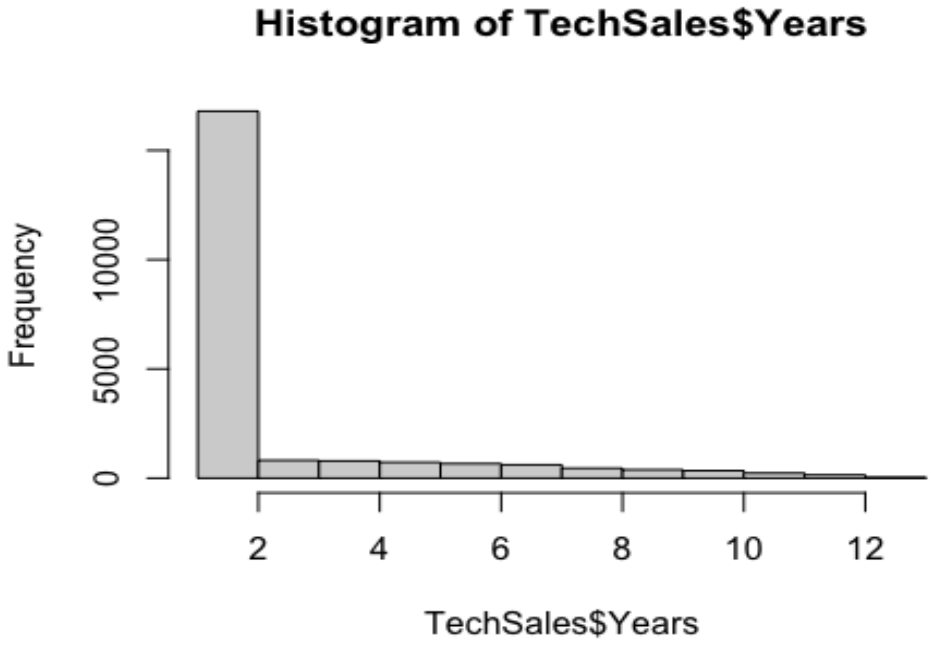

```
hist(TechSales$NPS)
```




```
summary(TechSales$NPS)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   6.000   6.278   8.000  10.000

boxplot(TechSales$NPS, horizontal = TRUE)
```




```r
hist(TechSales$Years)
```

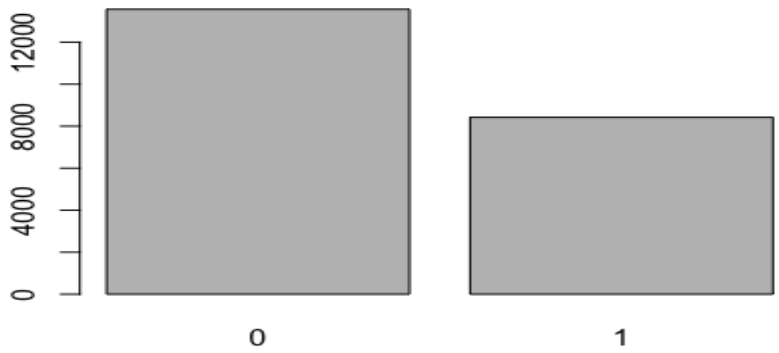


The population average salary @ 95% confidence level is around 73.3K

```
# Hint: how to construct confidence interval in R? Is it t.test() function?
t.test(TechSales$Salary, conf.level = 0.95)

##
##  One Sample t-test
##
## data:  TechSales$Salary
## t = 479.8, df = 21989, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  73372.81 73974.75
## sample estimates:
## mean of x
##  73673.78
```

Exploring the main variables to study: For the most significant variables we can visualize a histogram for the NPS Score and the Salary, the higher the NPS Score the higher the salary, followed by a box plot for both Variables. The number of certifications by a representative also affects NPS and the salary of the sales representative.

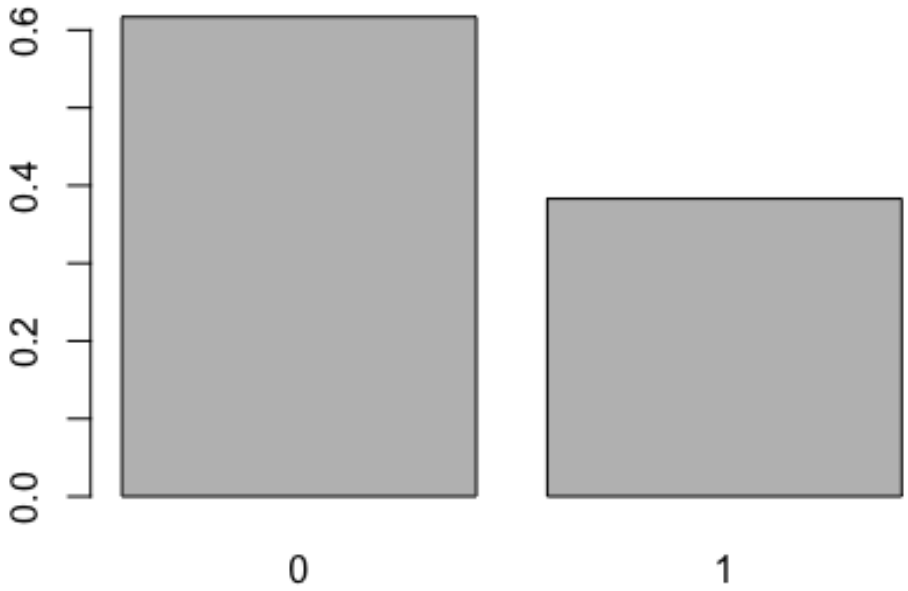
```
# For categorical variables, how should you tabulate or visualize them?
Histogram doesn't work
table(TechSales$Female)

##
##     0     1
## 13567  8423

tab1 <- table(TechSales$Female)
barplot(table(TechSales$Female))
```




```
tab2 <- prop.table(tab1)
tab2
```

```
##
##         0         1
## 0.6169623 0.3830377
```

```
barplot(tab2)
```




```
table(TechSales$Female)
```

```
##
##     0     1
## 13567  8423
```

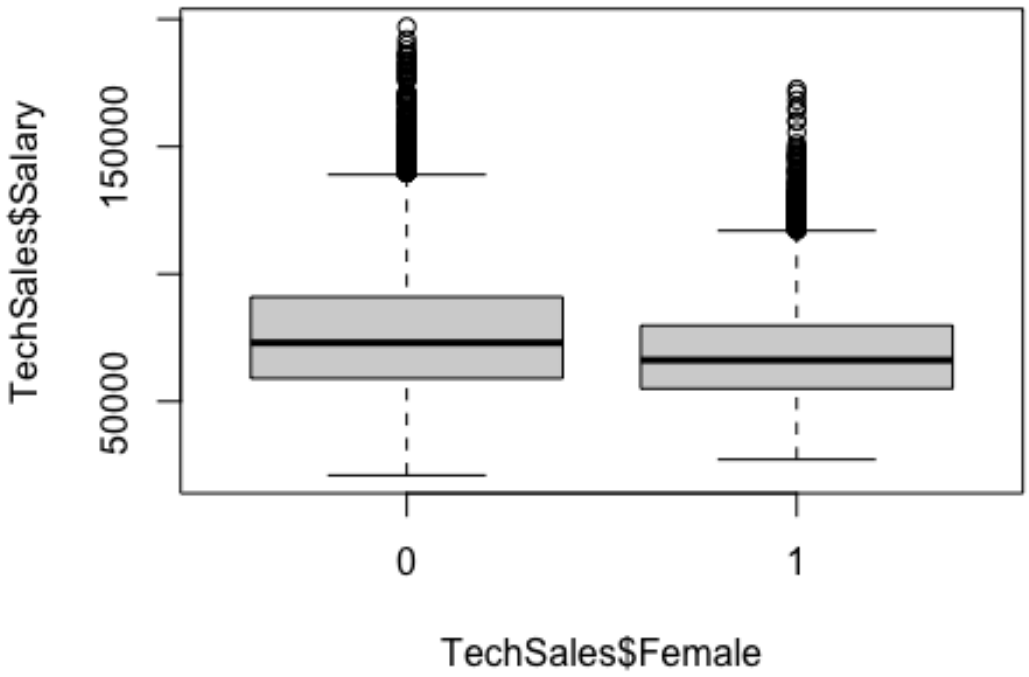
```
tab3 <- table(TechSales$Personality)
tab3
```

```
##
##  Analyst Diplomat Explorer Sentinel
##     2659     7849     8200     3282
```

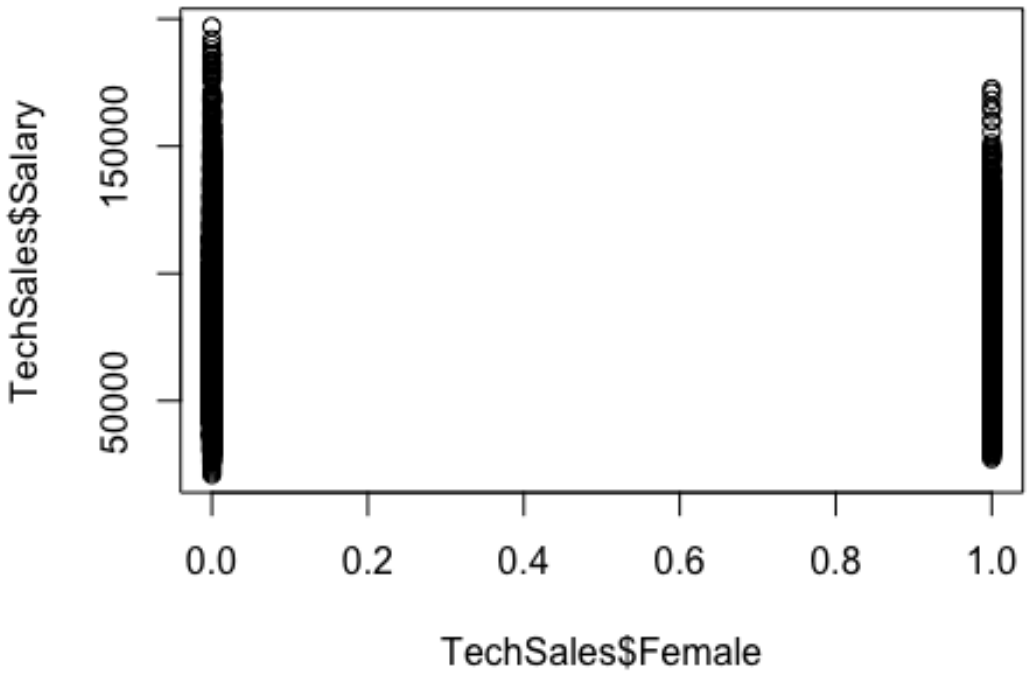
```
table(TechSales$Business)
```

```
##
## Hardware Software
##     9860    12130
```

```r
boxplot(TechSales$Salary ~ TechSales$Female)
```




```r
plot(TechSales$Salary ~ TechSales$Female)
```



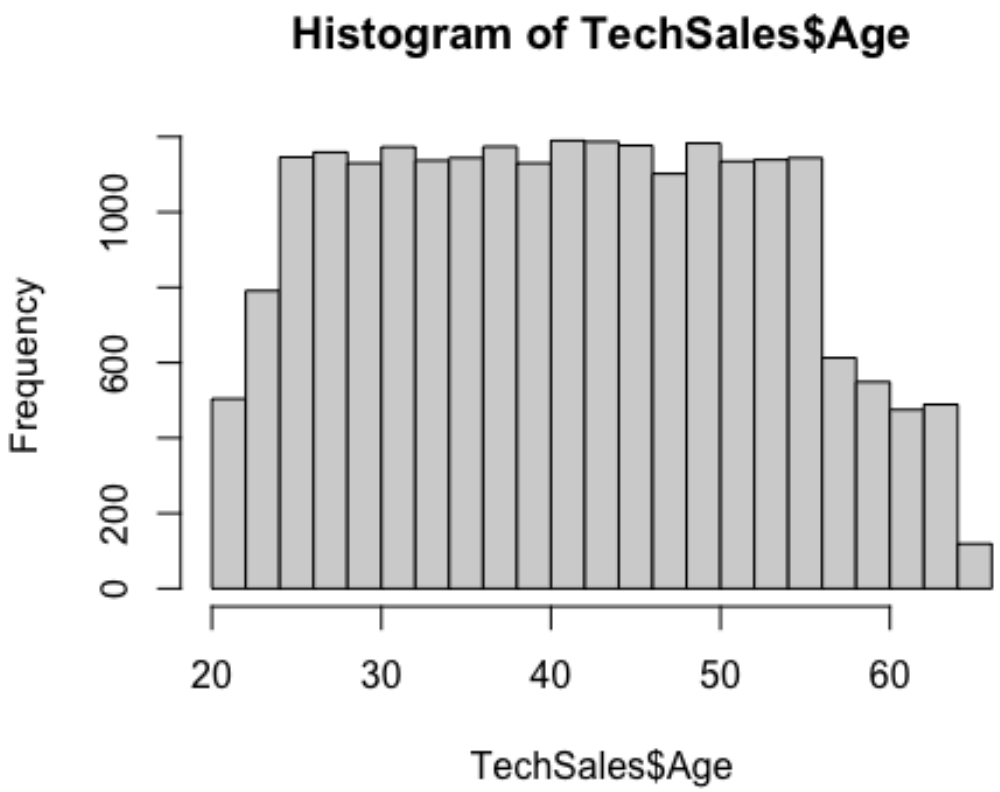

```
result <- aov(Salary ~ Personality, data = TechSales)
summary(result)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## Personality     3 9.885e+11 3.295e+11   695.8 <2e-16 ***
## Residuals   21986 1.041e+13 4.736e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Try to do this for other categorical variables
table(TechSales$Certficates)
```

```
##
##    0    1    2    3    4    5    6
## 2113 4445 4468 4175 3335 2528  926
```

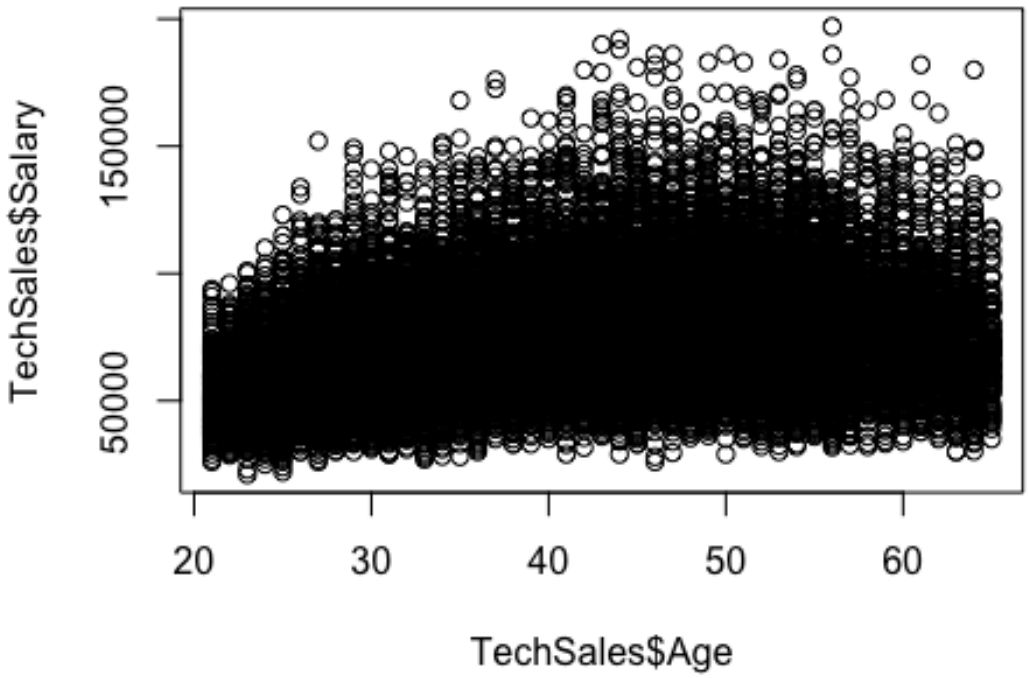
```
hist(TechSales$Age)
```



62% of the sales representatives are males and 38% are females, where male sales averages approximately 13,357, while for females (represented as 1), it's around 8,423. Average salaries among the four personality types. Sales professionals identified as Diplomats and Explorers have higher average salaries, approximately $78,597 and $77,899 respectively, compared to Analysts and Sentinels, who earn around $62,994 and $62,388 respectively.

2. Exploring the association between Earnings and other numerical variables

```
# 2. Exploring the association between Salary and other numerical variables
# Visualization tool
plot(TechSales$Salary ~ TechSales$Age)
```



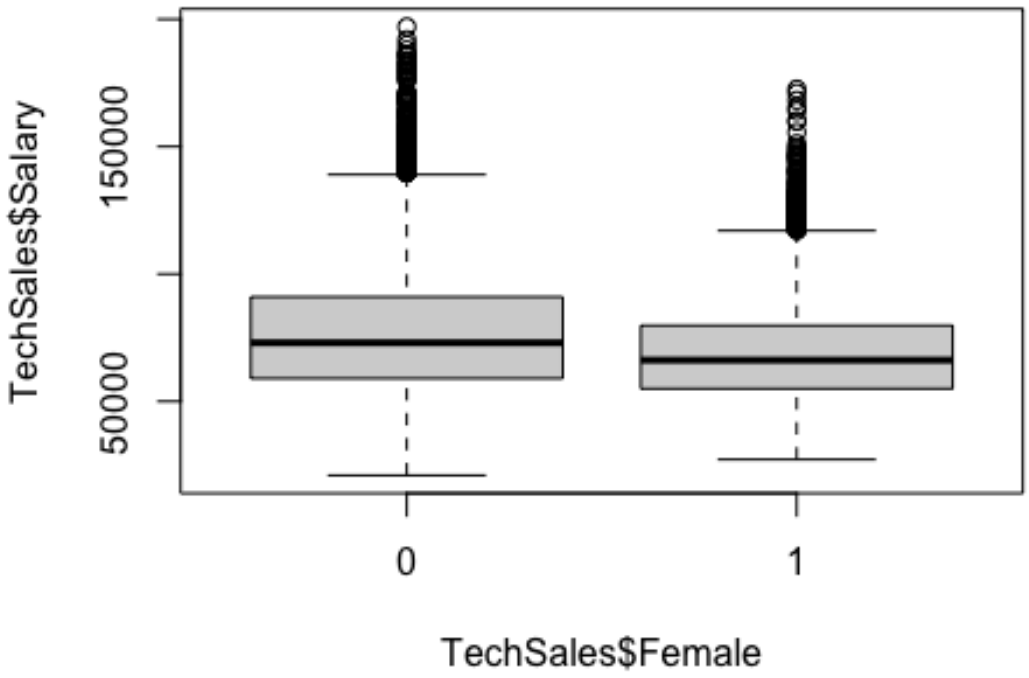

```
# Quantification using correlation coefficient
cor(TechSales$Salary, TechSales$Years)

## [1] 0.09308011

cor(TechSales$Salary, TechSales$Certficates)

## [1] 0.4584402

cor(TechSales$Salary, TechSales$NPS)

## [1] 0.5499314
```

The relationship between Sales and certifications and NPS is strong as it is approximately closer to 1 as compared to years of working which has a week relationship.

3. Exploring the association between Earnings and categorical variables

```
# 3. Exploring the association between Salary and categorical variables
tapply(TechSales$Salary, TechSales$Female, mean)

##        0        1
## 76597.70 68964.19
```

```r
# How to visualize this relationship using boxplot tool
boxplot(TechSales$Salary ~ TechSales$Female)
```



62% of the sales representatives are males and 38% are females, where male earning averages approximately $76,598, while for females (represented as 1), it's around $68,964.

4.Building models

```r
# Build simple models based on your intuition
TechSales$Sales_Rep <- NULL
# If you want to predict salary using: Age, Female, Years, College, how?
SalaryModel1 <- lm(Salary ~ Age + Female +Years+ College, data = TechSales)
summary(SalaryModel1, digits = 3)
```

```
##
## Call:
## lm(formula = Salary ~ Age + Female + Years + College, data = TechSales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59642 -14841  -2753  11850 112918
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 43441.39     631.75   68.76   <2e-16 ***
## Age           516.92      12.45   41.52   <2e-16 ***
## Female      -7463.47     291.85  -25.57   <2e-16 ***
## Years         738.76      58.38   12.65   <2e-16 ***
## CollegeYes  12157.41     352.52   34.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21030 on 21985 degrees of freedom
## Multiple R-squared:  0.1468, Adjusted R-squared:  0.1466
## F-statistic: 945.4 on 4 and 21985 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(Salary ~ Age+Female+Certficates, data = TechSales)
summary(model2)
```

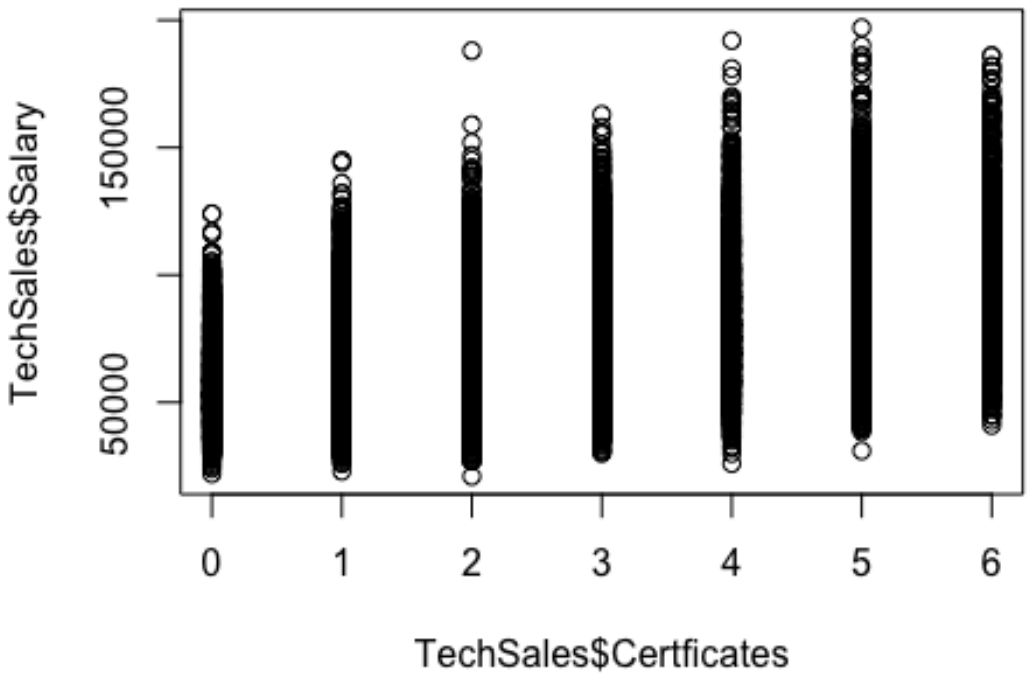
```
##
## Call:
## lm(formula = Salary ~ Age + Female + Certficates, data = TechSales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56649 -13083  -2027  11011 114043
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37257.58     536.67   69.42   <2e-16 ***
## Age           542.00      11.17   48.51   <2e-16 ***
## Female      -7466.69     262.26  -28.47   <2e-16 ***
## Certficates  6425.83      77.36   83.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18900 on 21986 degrees of freedom
## Multiple R-squared:  0.3109, Adjusted R-squared:  0.3108
## F-statistic:  3307 on 3 and 21986 DF,  p-value: < 2.2e-16
```

```
model3 <- lm(Salary ~ Certficates+Feedback, data = TechSales)
summary(model3)
```

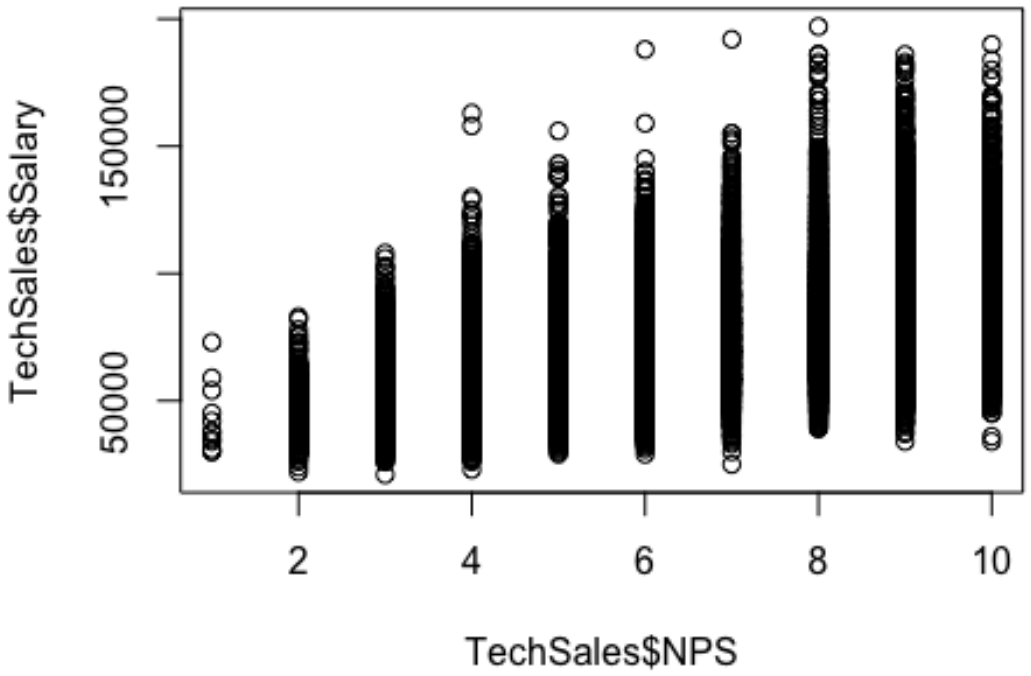
```
##
## Call:
## lm(formula = Salary ~ Certficates + Feedback, data = TechSales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64993 -13064  -1646  11203 109867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34274.79     472.39   72.56   <2e-16 ***
## Certficates  6352.67      77.44   82.03   <2e-16 ***
```

```
## Feedback     8558.61     152.54   56.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18930 on 21987 degrees of freedom
## Multiple R-squared:  0.3091, Adjusted R-squared:  0.309
## F-statistic:  4918 on 2 and 21987 DF,  p-value: < 2.2e-16
```

```
plot(TechSales$Salary ~ TechSales$Certficates)
lines(lowess(TechSales$Age, TechSales$Salary), col = "blue")
```



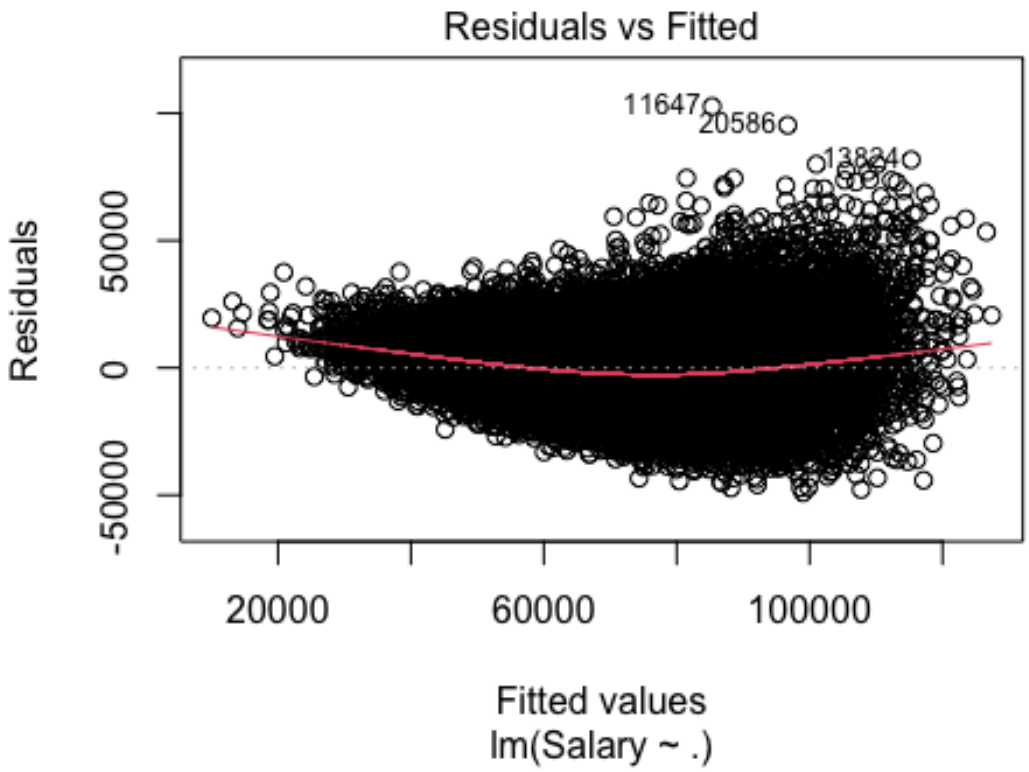

```
plot(TechSales$Salary ~ TechSales$NPS)
```

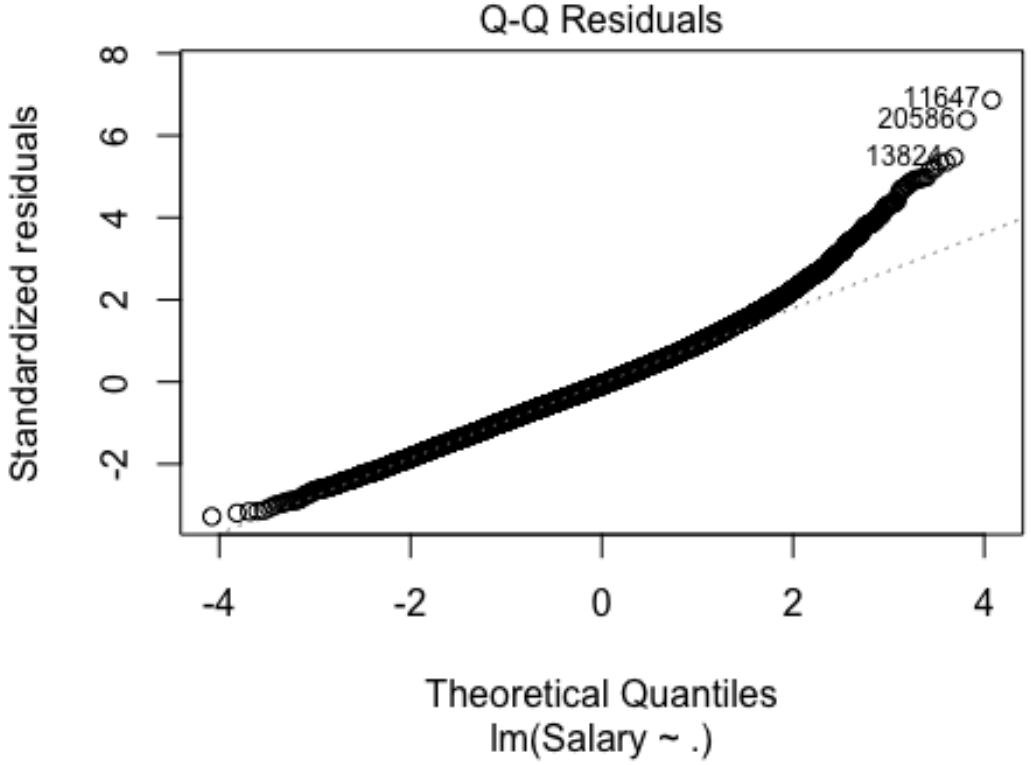


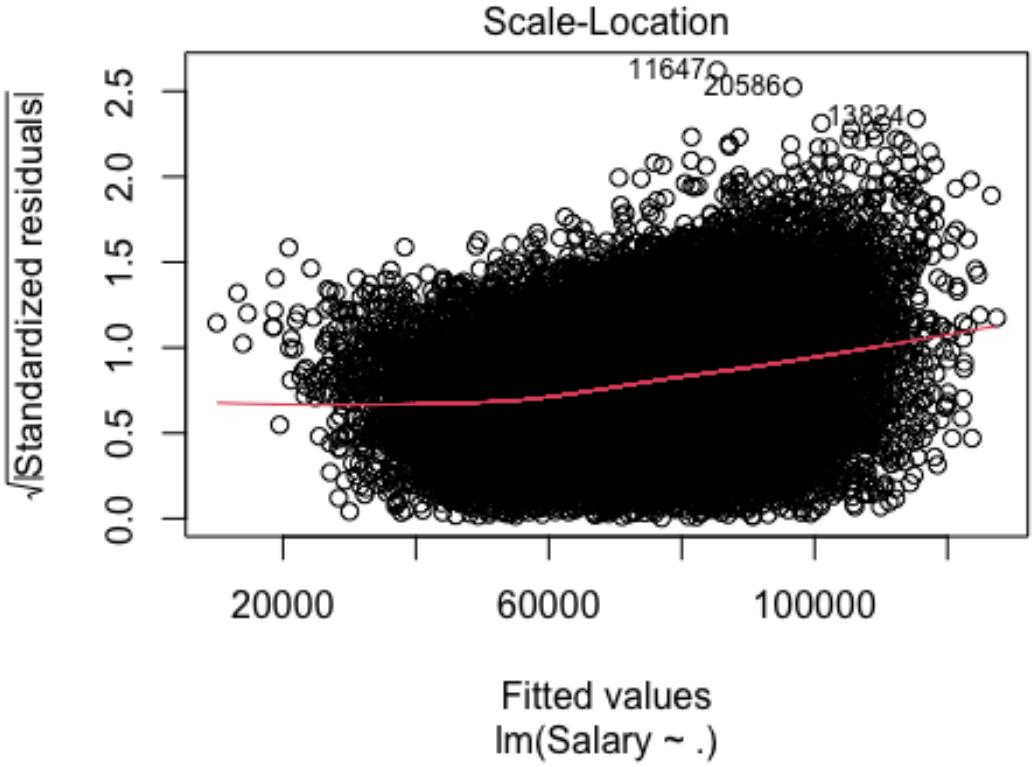

```
full_model <- lm(Salary ~ ., data = TechSales)
summary(full_model)

##
## Call:
## lm(formula = Salary ~ ., data = TechSales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49027  -9758  -1007   8658 102669
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -7156.280    667.630 -10.719  < 2e-16 ***
## BusinessSoftware    -1623.446    211.592  -7.673 1.76e-14 ***
## Age                   508.906      9.123  55.784  < 2e-16 ***
## Female              -7654.444    208.179 -36.769  < 2e-16 ***
## Years                 386.313     43.306   8.920  < 2e-16 ***
## CollegeYes          11235.487    252.571  44.484  < 2e-16 ***
## PersonalityDiplomat 11483.177    357.370  32.132  < 2e-16 ***
## PersonalityExplorer 11698.514    355.437  32.913  < 2e-16 ***
## PersonalitySentinel  -275.133    390.116  -0.705    0.481
## Certficates          5351.422     73.021  73.286  < 2e-16 ***
```

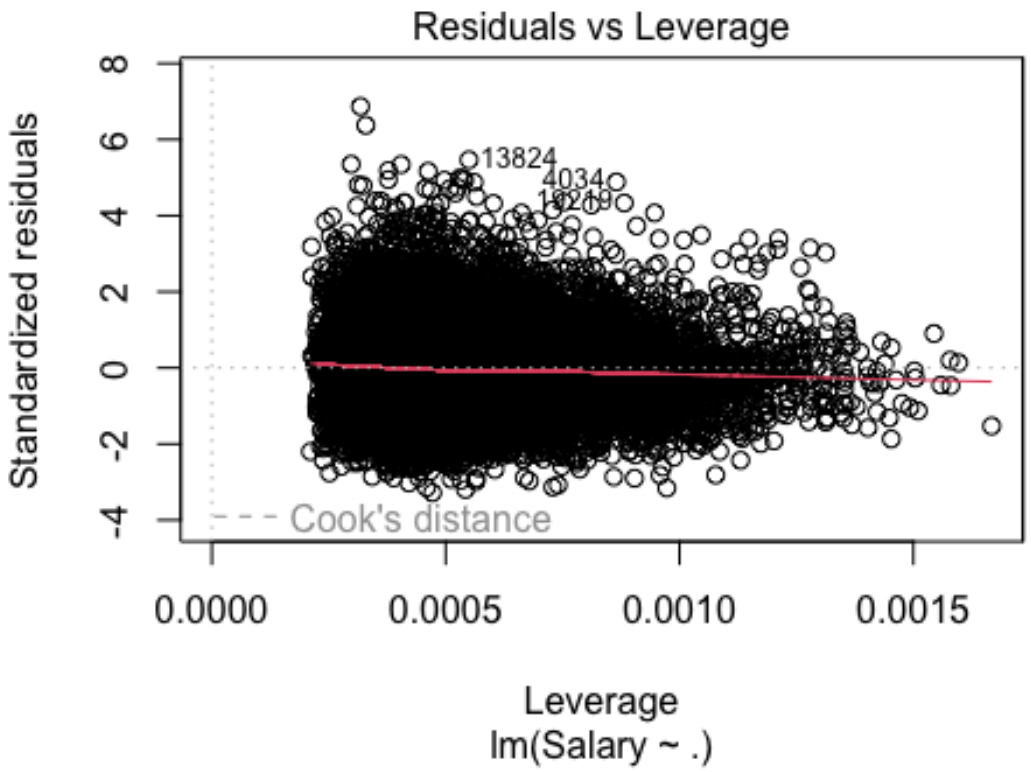
```
## Feedback             7233.601    129.304  55.942  < 2e-16 ***
## NPS                  1894.261     65.084  29.105  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14950 on 21978 degrees of freedom
## Multiple R-squared:  0.5691, Adjusted R-squared:  0.5689
## F-statistic:  2639 on 11 and 21978 DF,  p-value: < 2.2e-16
```

```
plot(full_model)
```


Residuals vs Fitted
11647
20586
13824
Residuals
50000
0
-50000
20000
60000
100000
Fitted values
lm(Salary ~ .)


Q-Q Residuals
Standardized residuals
8
6
4
2
0
-2
11647
20586
13824
-4
-2
0
2
4
Theoretical Quantiles
lm(Salary ~ .)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
11647
20586
13824
20000
60000
100000
Fitted values
lm(Salary ~ .)




```
model4 <- lm(Salary ~ NPS+ Certficates + Feedback, data = TechSales)

# If you want to predict NPS using: Age, Female, Years, College, how?
#NPSModel1 <- lm(NPS ~ , data = TechSales)
#summary(NPSModel1, digits = 3)

# If you're interested in a different model using Age, Female, Years,
College, Certificates, Business, how?
#SalaryModel2 <- lm(Salary ~ , data = TechSales)
#summary(SalaryModelFull, digits = 3)

# What if you want to use all available variables for your linear regression
model, how?
#SalaryModelFull <- lm(Salary ~ , data = TechSales)
#summary(SalaryModelFull, digits = 3)
```

Interpreting models

model1: being everything constant, the coefficient for female is negative, indicating that being female is associated with a decrease in salary of approximately $7463.47 compared to males. The R-squared is 0.1468, indicating that about 1.5% of the variation in salary can be explained by age, gender,years and college.

model2 for each additional year of experience, there is an average salary increase of $542.00, a figure that does not hold statistical significance, suggesting a weak positive correlation between age and earnings. For each certification, salary increases by 6425.83, indicating a strong relationship between certification and salary. Multiple R-squared values of 0.3109, meaning nearly 31.09% of the salary variation is accounted for by the second model.

full_model R-squared value is 0.5691, indicating that about 56.92% of the variation in salary can be explained by all the predictors in the model. The t-tests show that most are significant, with p-values well below the 0.05 threshold, indicating a strong likelihood that these variables have a true effect on salary. Notably, the coefficients for age, college education, personality types (Diplomat and Explorer), certificates, feedback, and NPS are all positive, suggesting that higher values in these predictors are associated with higher salaries.

Conclusion

Upon examining the variables, we discovered that the number of Certificates obtained, and the Feedback score, NPS are the variables most strongly correlated with the Salary of sales representatives at Cassius Weatherby. These variables significantly contribute to explaining the salary variations, which is why Model 4 (constructed as model4 <- lm(Salary ~ NPS+ Certificates + Feedback, data = TechSales)) performs comparably to the full model in terms of the Multiple R-squared and Adjusted R-squared values.