Final Project

Kamaal Bartlett

2024-06-01

```
# Install and load necessary packages
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}
if (!requireNamespace("broom", quietly = TRUE)) {
  install.packages("broom")
}
if (!requireNamespace("pwr", quietly = TRUE)) {
  install.packages("pwr")
}
if (!requireNamespace("effsize", quietly = TRUE)) {
  install.packages("effsize")
}

library(dplyr)
library(ggplot2)
library(broom)
library(pwr)
library(effsize)  # Make sure effsize is loaded
```

Identify Team Members, Data Source and Statistical Questions

Team Members: Kamaal Bartlett

Data Source: Sleuth3 (case 1202)

Statistical Questions: 1. Is there a significant difference in starting salary between men and women at the bank? 2. How does the inclusion of gender as a variable in a regression model affect the predictability of starting salaries?

Exploratory Data Analysis

First, we'll load the data set and examine the raw data to get a better understanding before performing any analysis.

```
#Load the dataset
data(case1202)
df = case1202
```

```r
#Viewing the structure of the dataset
str(df)
```

```
## 'data.frame':    93 obs. of  7 variables:
##  $ Bsal  : int  5040 6300 6000 6000 6000 6840 8100 6000 6000 6900 ...
##  $ Sal77 : int  12420 12060 15120 16320 12300 10380 13980 10140 12360 10920 ...
##  $ Sex   : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Senior: int  96 82 67 97 66 92 66 82 88 75 ...
##  $ Age   : int  329 357 315 354 351 374 369 363 555 416 ...
##  $ Educ  : int  15 15 15 12 12 15 16 12 12 15 ...
##  $ Exper : num  14 72 35.5 24 56 41.5 54.5 32 252 132 ...
```

```r
#Viewing the first few rows of the dataset
head(df)
```

```
##   Bsal Sal77  Sex Senior Age Educ Exper
## 1 5040 12420 Male     96 329   15  14.0
## 2 6300 12060 Male     82 357   15  72.0
## 3 6000 15120 Male     67 315   15  35.5
## 4 6000 16320 Male     97 354   12  24.0
## 5 6000 12300 Male     66 351   12  56.0
## 6 6840 10380 Male     92 374   15  41.5
```

```r
#Load dplyr if not already loaded
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}
library(dplyr)

#General summary for all data
overall_summary <- df %>%
  summarise(
    MeanSalary = mean(Bsal, na.rm = TRUE),
    MedianSalary = median(Bsal, na.rm = TRUE),
    StdDevSalary = sd(Bsal, na.rm = TRUE),
    MinSalary = min(Bsal, na.rm = TRUE),
    MaxSalary = max(Bsal, na.rm = TRUE)
  )

#Print overall summary
print("Overall Summary:")
```

```
## [1] "Overall Summary:"
```

```r
print(overall_summary)
```

```
##   MeanSalary MedianSalary StdDevSalary MinSalary MaxSalary
## 1   5420.323         5400     709.5872      3900      8100
```

```r
#Summary by Gender
gender_summary <- df %>%
  group_by(Sex) %>%
  summarise(
    MeanSalary = mean(Bsal, na.rm = TRUE),
    MedianSalary = median(Bsal, na.rm = TRUE),
    StdDevSalary = sd(Bsal, na.rm = TRUE),
    MinSalary = min(Bsal, na.rm = TRUE),
    MaxSalary = max(Bsal, na.rm = TRUE),
    .groups = 'drop'  # Drop the grouping structure after summarising
  )

#Print gender-specific summary
print("Gender-Specific Summary:")
```

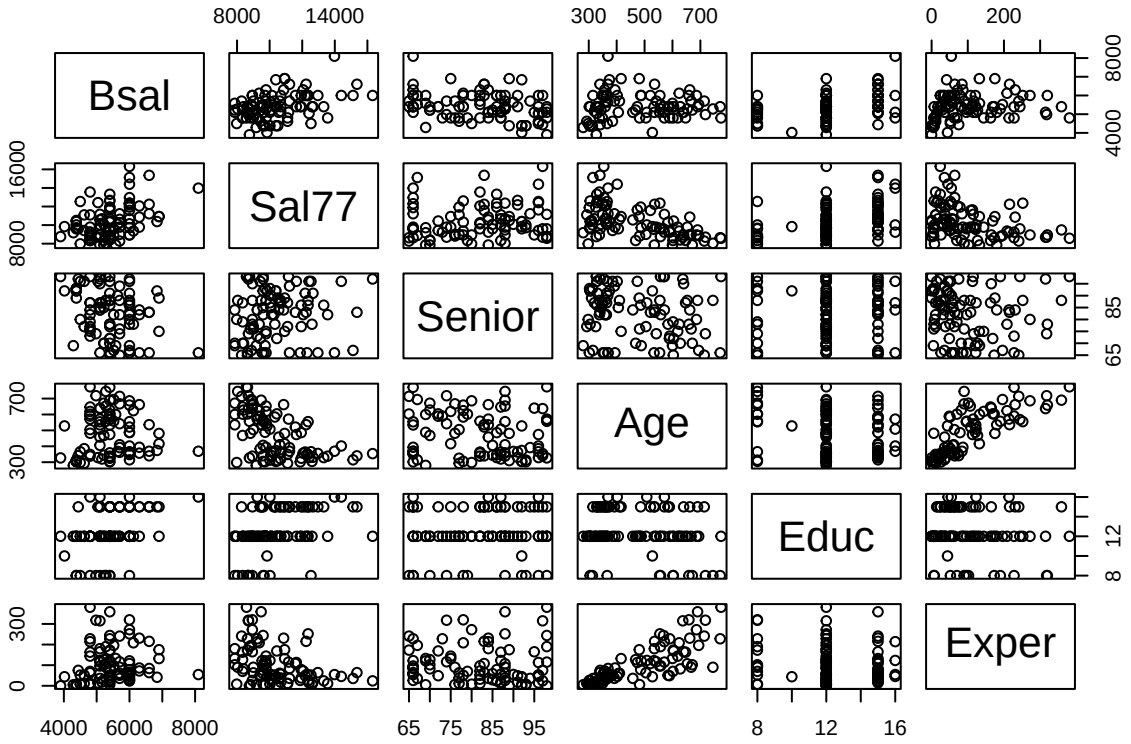
```
## [1] "Gender-Specific Summary:"
```

```r
print(gender_summary)
```

```
## # A tibble: 2 x 6
##   Sex    MeanSalary MedianSalary StdDevSalary MinSalary MaxSalary
##   <fct>       <dbl>        <dbl>        <dbl>     <int>     <int>
## 1 Female      5139.         5220         540.      3900      6300
## 2 Male        5957.         6000         691.      4620      8100
```

Exploring Relationships

```r
#Scatter plot matrix
pairs(df[, c("Bsal", "Sal77", "Senior", "Age", "Educ", "Exper")])
```

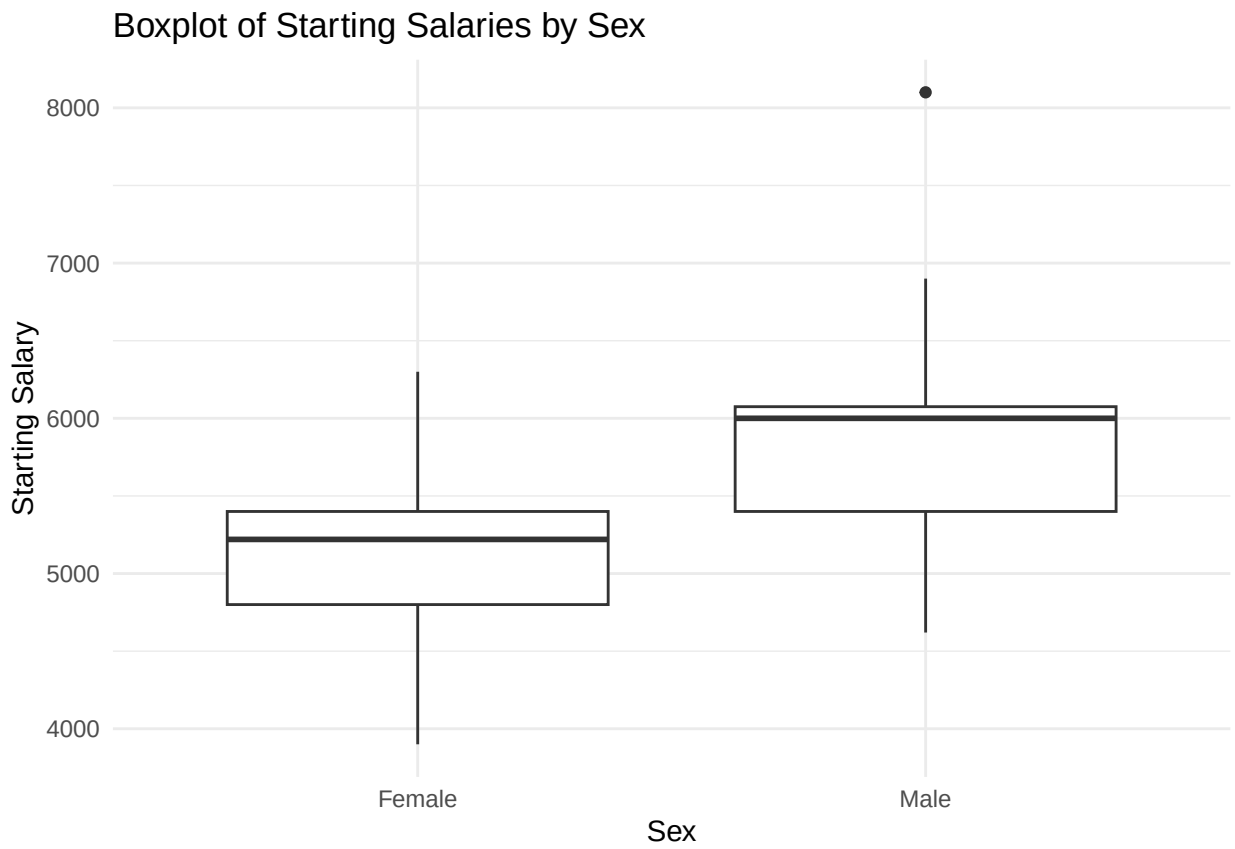



```
#Correlation matrix
cor(df[, c("Bsal", "Sal77", "Senior", "Age", "Educ", "Exper")])
```

```
##                Bsal      Sal77      Senior         Age        Educ       Exper
## Bsal     1.00000000  0.4223695 -0.28584352  0.03389932  0.41198516  0.16674049
## Sal77    0.42236952  1.0000000  0.12595511 -0.54674687  0.42102127 -0.37198639
## Senior -0.28584352  0.1259551  1.00000000 -0.18448263  0.05984385 -0.07466085
## Age      0.03389932 -0.5467469 -0.18448263  1.00000000 -0.22525298  0.79787476
## Educ     0.41198516  0.4210213  0.05984385 -0.22525298  1.00000000 -0.10117309
## Exper    0.16674049 -0.3719864 -0.07466085  0.79787476 -0.10117309  1.00000000
```

Visual Representation

```
library(ggplot2)

ggplot(df, aes(x = Sex, y = Bsal)) +
  geom_boxplot() +
  labs(title = "Boxplot of Starting Salaries by Sex",
       x = "Sex",
       y = "Starting Salary") +
  theme_minimal()
```



Two Sample Test

```
#Two-sample t-test
t_test_result <- t.test(Bsal ~ Sex, data = df)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  Bsal by Sex
## t = -5.83, df = 51.329, p-value = 3.71e-07
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -1099.6693  -536.3758
## sample estimates:
## mean in group Female   mean in group Male
##             5138.852             5956.875
```

We chose the two-sample t-test because it is appropriate for comparing the means of two independent groups (men and women). The t-test assumes that the data are approximately normally distributed and that the variances are equal, which we checked using diagnostic plots

Power Analysis

```
#Power analysis
library(pwr)
effect_size <- cohen.d(df$Bsal ~ df$Sex)$estimate
power_test <- pwr.t2n.test(n1 = sum(df$Sex == "Male"), n2 = sum(df$Sex == "Female"), d = effect_size, si
power_test
```

```
##
##      t test power calculation
##
##              n1 = 32
##              n2 = 61
##               d = 1.37351
##       sig.level = 0.05
##           power = 0.99999
##     alternative = two.sided
```

Why Regression is More Powerful

Regression analysis is more powerful than simply comparing two populations because it allows us to control for multiple variables simultaneously. This means we can isolate the effect of gender on salary while accounting for other factors such as experience, education, and seniority. This provides a more accurate and comprehensive understanding of the factors influencing salary.

Regression Analysis

Building the Regression Model Without Gender

```
#Forward Selection
model_forward <- step(lm(Bsal ~ 1, data = df), direction = "forward",
                      scope = ~ Sal77 + Senior + Age + Educ + Exper)
```

```
## Start:  AIC=1222.03
## Bsal ~ 1
##
##          Df Sum of Sq      RSS    AIC
## + Sal77   1   8263890 38059400 1205.8
## + Educ    1   7862534 38460756 1206.7
## + Senior  1   3784915 42538376 1216.1
## + Exper   1   1287898 45035392 1221.4
## <none>                46323290 1222.0
## + Age     1     53233 46270057 1223.9
##
## Step:  AIC=1205.75
## Bsal ~ Sal77
##
##          Df Sum of Sq      RSS    AIC
## + Exper   1   5638778 32420622 1192.8
```

```
## + Senior  1   5410712 32648688 1193.5
## + Age     1   4634137 33425263 1195.7
## + Educ    1   3087141 34972259 1199.9
## <none>                38059400 1205.8
##
## Step:  AIC=1192.84
## Bsal ~ Sal77 + Exper
##
##          Df Sum of Sq      RSS    AIC
## + Senior  1   5086897 27333725 1179.0
## + Educ    1   2573281 29847341 1187.2
## <none>                32420622 1192.8
## + Age     1    272666 32147956 1194.0
##
## Step:  AIC=1178.97
## Bsal ~ Sal77 + Exper + Senior
##
##        Df Sum of Sq      RSS    AIC
## + Educ  1   2643358 24690367 1171.5
## <none>              27333725 1179.0
## + Age   1     13788 27319938 1180.9
##
## Step:  AIC=1171.51
## Bsal ~ Sal77 + Exper + Senior + Educ
##
##        Df Sum of Sq      RSS    AIC
## <none>              24690367 1171.5
## + Age   1     51936 24638432 1173.3
```

```
#Backward Elimination
model_backward <- step(lm(Bsal ~ Sal77 + Senior + Age + Educ + Exper, data = df), direction = "backward
```

```
## Start:  AIC=1173.31
## Bsal ~ Sal77 + Senior + Age + Educ + Exper
##
##          Df Sum of Sq      RSS    AIC
## - Age     1     51936 24690367 1171.5
## <none>                24638432 1173.3
## - Exper   1   1499300 26137732 1176.8
## - Educ    1   2681506 27319938 1180.9
## - Senior  1   4808063 29446495 1187.9
## - Sal77   1   6973408 31611839 1194.5
##
## Step:  AIC=1171.51
## Bsal ~ Sal77 + Senior + Educ + Exper
##
##          Df Sum of Sq      RSS    AIC
## <none>                24690367 1171.5
## - Educ    1   2643358 27333725 1179.0
## - Exper   1   4808789 29499157 1186.1
## - Senior  1   5156974 29847341 1187.2
## - Sal77   1   7643099 32333466 1194.6
```

```r
#Summarize the models
summary(model_forward)
```

```
##
## Call:
## lm(formula = Bsal ~ Sal77 + Exper + Senior + Educ, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -953.23 -360.19  -32.65  291.37 1605.14
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4045.2411   584.6309   6.919 7.00e-10 ***
## Sal77          0.1915     0.0367   5.219 1.19e-06 ***
## Exper          2.7153     0.6559   4.140 7.94e-05 ***
## Senior       -23.2846     5.4312  -4.287 4.62e-05 ***
## Educ          82.0597    26.7346   3.069  0.00285 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 529.7 on 88 degrees of freedom
## Multiple R-squared:  0.467,  Adjusted R-squared:  0.4428
## F-statistic: 19.28 on 4 and 88 DF,  p-value: 2.039e-11
```

```r
summary(model_backward)
```

```
##
## Call:
## lm(formula = Bsal ~ Sal77 + Senior + Educ + Exper, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -953.23 -360.19  -32.65  291.37 1605.14
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4045.2411   584.6309   6.919 7.00e-10 ***
## Sal77          0.1915     0.0367   5.219 1.19e-06 ***
## Senior       -23.2846     5.4312  -4.287 4.62e-05 ***
## Educ          82.0597    26.7346   3.069  0.00285 **
## Exper          2.7153     0.6559   4.140 7.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 529.7 on 88 degrees of freedom
## Multiple R-squared:  0.467,  Adjusted R-squared:  0.4428
## F-statistic: 19.28 on 4 and 88 DF,  p-value: 2.039e-11
```

Interaction Terms

To explore potential interactions between predictors and the gender variable, we build a model with interaction terms

```r
#Interaction terms model
model_interaction <- lm(Bsal ~ Educ * Senior + Exper * Senior + Sal77 * Senior + Age * Senior, data = d
summary(model_interaction)
```

```
##
## Call:
## lm(formula = Bsal ~ Educ * Senior + Exper * Senior + Sal77 *
##     Senior + Age * Senior, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -945.45 -372.97    2.88  299.29 1638.69
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.167e+03  5.297e+03   0.409    0.683
## Educ          1.962e+02  2.063e+02   0.951    0.344
## Senior       -3.322e+00  6.209e+01  -0.053    0.957
## Exper         5.235e+00  1.014e+01   0.516    0.607
## Sal77         2.322e-01  3.237e-01   0.717    0.475
## Age          -5.771e-01  6.806e+00  -0.085    0.933
## Educ:Senior  -1.371e+00  2.461e+00  -0.557    0.579
## Senior:Exper -3.469e-02  1.213e-01  -0.286    0.776
## Senior:Sal77 -4.233e-04  3.783e-03  -0.112    0.911
## Senior:Age    1.202e-02  8.290e-02   0.145    0.885
##
## Residual standard error: 543.2 on 83 degrees of freedom
## Multiple R-squared:  0.4713, Adjusted R-squared:  0.4139
## F-statistic:  8.22 on 9 and 83 DF,  p-value: 1.227e-08
```

Transformations

To account for potential non-linear relationships, we can transform the response and variable and build a new model

```r
#Log transformation of the response variable
df$log_Bsal <- log(df$Bsal)

#Model with transformed response variable
model_transformed <- lm(log_Bsal ~ ., data = df)
summary(model_transformed)
```

```
##
## Call:
## lm(formula = log_Bsal ~ ., data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.071836 -0.004569  0.002162  0.006545  0.016106
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  7.599e+00  2.260e-02 336.284   <2e-16 ***
## Bsal         1.784e-04  2.817e-06  63.339   <2e-16 ***
## Sal77        9.042e-07  1.098e-06   0.823    0.413
## SexMale      2.255e-03  3.892e-03   0.579    0.564
## Senior      -1.155e-04  1.465e-04  -0.789    0.433
## Age          4.262e-05  1.868e-05   2.282    0.025 *
## Educ         3.365e-04  6.731e-04   0.500    0.618
## Exper       -1.065e-05  2.640e-05  -0.403    0.688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01263 on 85 degrees of freedom
## Multiple R-squared:  0.9912, Adjusted R-squared:  0.9904
## F-statistic:  1363 on 7 and 85 DF,  p-value: < 2.2e-16
```

Adding the Gender Variable and Assessing Its IMpact

We add the gender variable to our best model to assess its impact

```
#Adding the gender variable
best_model_with_gender <- update(model_forward, . ~ . + Sex)
summary(best_model_with_gender)
```

```
##
## Call:
## lm(formula = Bsal ~ Sal77 + Exper + Senior + Educ + Sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1140.28  -296.52   -40.37   273.34  1386.74
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4998.27282  587.07474   8.514 4.44e-13 ***
## Sal77          0.11145    0.03909   2.851  0.00544 **
## Exper          2.04127    0.62646   3.258  0.00160 **
## Senior       -24.23721    5.00950  -4.838 5.62e-06 ***
## Educ          69.40289   24.82660   2.796  0.00638 **
## SexMale      534.83292  131.02034   4.082 9.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 488 on 87 degrees of freedom
## Multiple R-squared:  0.5527, Adjusted R-squared:  0.527
## F-statistic:  21.5 on 5 and 87 DF,  p-value: 6.027e-14
```

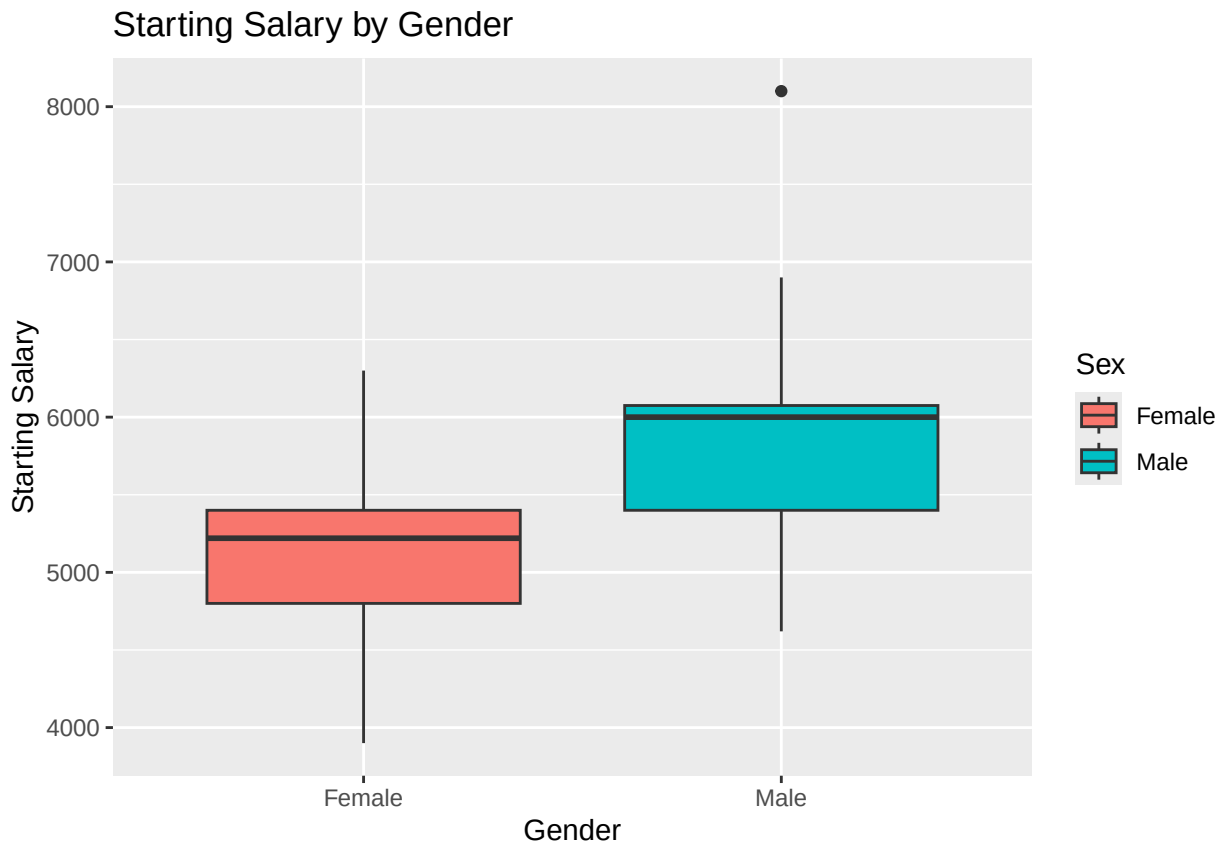
```
#Comparing models
anova(model_forward, best_model_with_gender)
```

```
## Analysis of Variance Table
##
## Model 1: Bsal ~ Sal77 + Exper + Senior + Educ
```

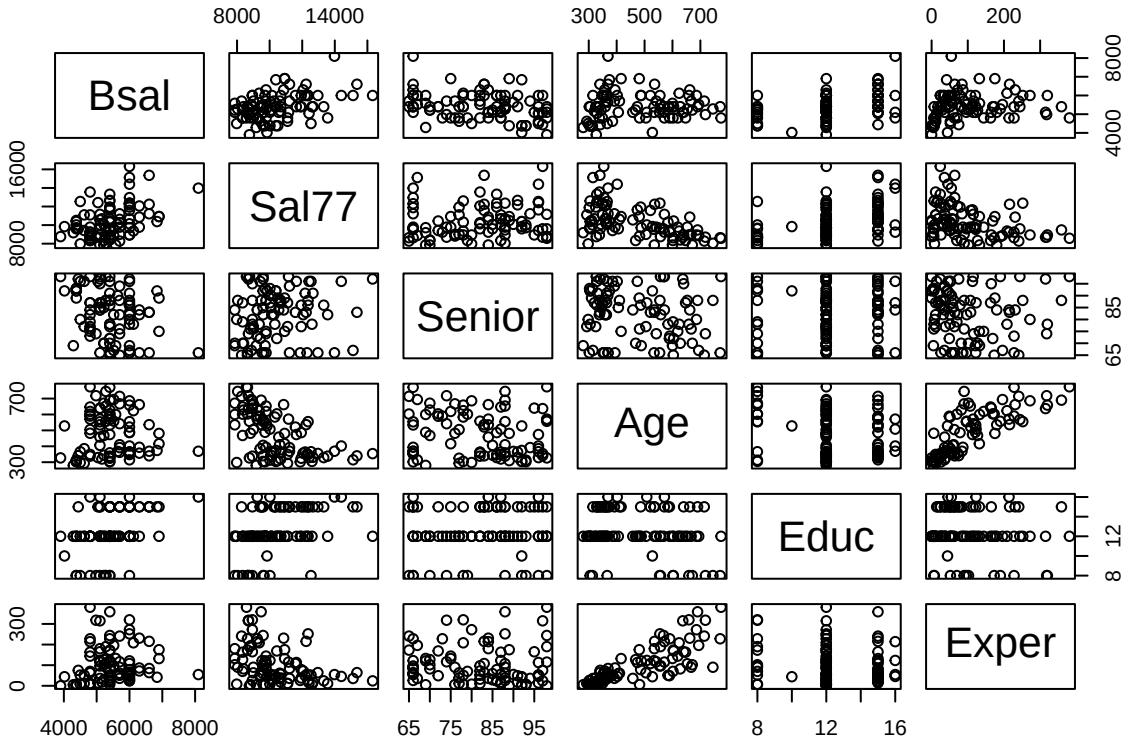
```
## Model 2: Bsal ~ Sal77 + Exper + Senior + Educ + Sex
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     88 24690367
## 2     87 20721544  1   3968823 16.663 9.879e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Visualization of Results

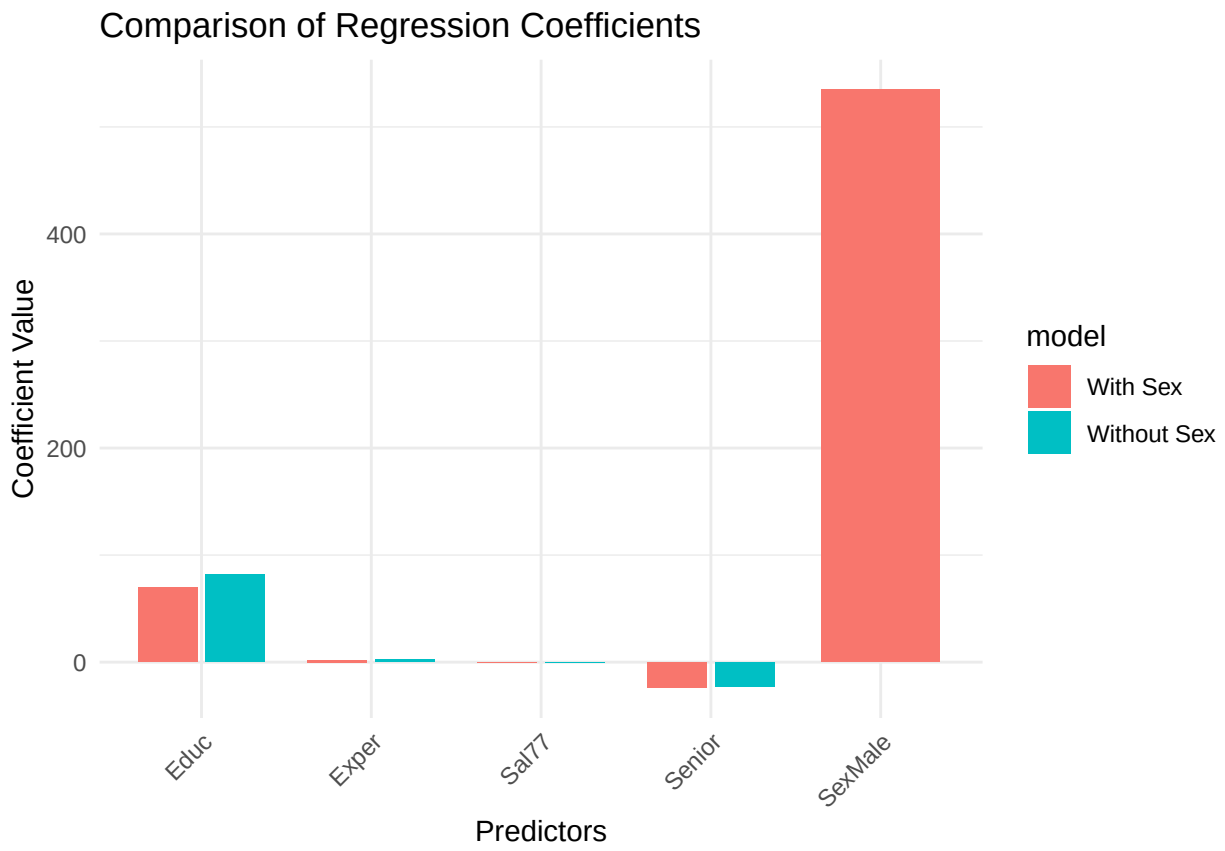
```
#Boxplot of salaries by gender
ggplot(df, aes(x = Sex, y = Bsal, fill = Sex)) +
  geom_boxplot() +
  labs(title = "Starting Salary by Gender", x = "Gender", y = "Starting Salary")
```



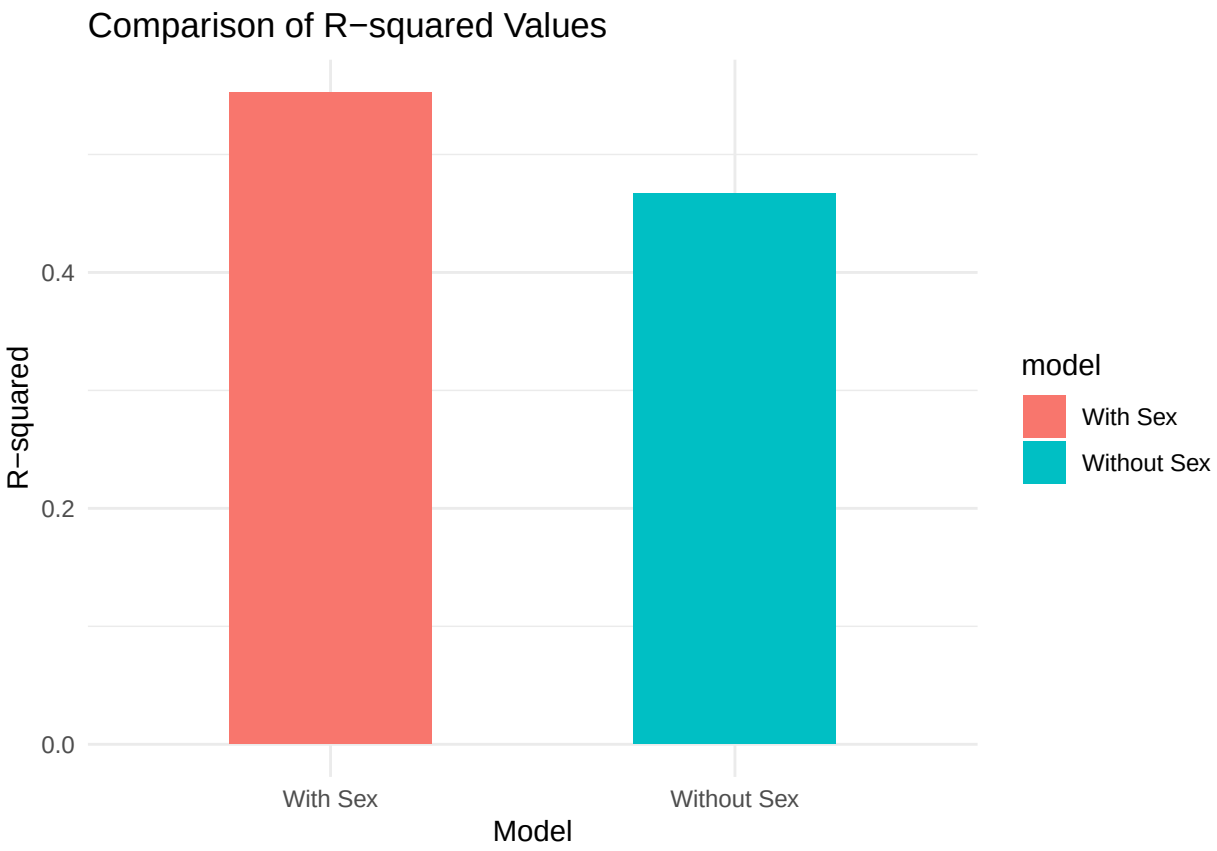

```
#Scatter plot matrix with best model predictors
pairs(df[, c("Bsal", "Sal77", "Senior", "Age", "Educ", "Exper")])
```




```r
#Coefficients comparison plot
coeffs_model1 <- broom::tidy(model_forward)
coeffs_model2 <- broom::tidy(best_model_with_gender)

coeffs_model1$model <- "Without Sex"
coeffs_model2$model <- "With Sex"

coeffs_combined <- bind_rows(coeffs_model1, coeffs_model2) %>%
  filter(term != "(Intercept)")

ggplot(coeffs_combined, aes(x = term, y = estimate, fill = model)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.7) +
  labs(title = "Comparison of Regression Coefficients", x = "Predictors", y = "Coefficient Value") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
#R-squared values plot
r_squared_data <- data.frame(
  model = c("Without Sex", "With Sex"),
  r_squared = c(summary(model_forward)$r.squared, summary(best_model_with_gender)$r.squared)
)

ggplot(r_squared_data, aes(x = model, y = r_squared, fill = model)) +
  geom_bar(stat = "identity", width = 0.5) +
  labs(title = "Comparison of R-squared Values", x = "Model", y = "R-squared") +
  theme_minimal()
```



##Conclusions and Implications

###Statistical Significance of Sex The inclusion of the gender variable significantly improves the model's predictive ability, as evidenced by the ANOVA and the increase in R-squared value. The F-test from the ANOVA confirms that adding the sex/gender variable contributes significantly to the model.

###Economic and Policy Implications The significant coefficient for SexMale suggests that there is a gender pay gap favoring males at this bank. This finding could have implications for policy and practices within the bank regarding pay equity.

###Possibility of Designed Experiment A designed experiment for this data set would involve randomly assigning salaries to employees to remove any existing biases. However, this is impractical and unethical in a real-world setting.

##Conclusion The statistical analysis suggests a significant gender pay gap favoring males at this bank. The findings could have implications for the bank's policies and practices regarding pay equity.

###Statistical Conclusion The inclusion of the gender variable significantly improves the model's predictive power, indicating a substantial gender pay gap.

###Conclusion for a Judge Based on statistical evidence, there is a significant gender pay gap at this bank, favoring men. This finding is consistent with potential gender discrimination, which the bank should address through policy changes and corrective actions.