HR Dashboard

Context: This dataset comprises extensive data on company employees, encompassing their professional experience, personal characteristics, position they hold and employment-related variables. It offers a wealth of information for conducting diverse analyses in the realms of HR and workforce management, such as employee retention assessments, salary structure evaluations, diversity and inclusion investigations, and examinations of their feedback scores. The source of the dataset is not mentioned, we can assume that it was created to gain a comprehensive view into the dynamics of a particular workplace.

There are about 200 rows with 11 different columns.

Varianbles

Name: (Character type) Name of the employee

Age: (Character type) The age of each employee, providing demographic insights.

Gender: (Character type) Gender identity of employees, promoting diversity analysis.

Projects.Completed: (Integer type) Number of projects completed by each employee during their time at the company.

Productivity(%): (Integer type) Productivity rating of employee in percentage.

Satisfaction.Rate(%): (Integer type) rating of satisfactory work done by employee in percentage.

Feedback.Score: (Double type) Feedback value of employee.

Department: (Character type) department he/she is working in at the company.

Position: (Character type) position of the employee.

Joining.Date: (Character type) Joining month and year of the employee.

Salary: (Integer type) Salary of the employee in USD.

Research Question:

What are the key defining factors behind salary?

In [27]:
```
library(dplyr)
library(ggplot2)
library(tidyverse)
library(repr)
library(infer)
```

```
library(cowplot)
library(broom)
library(GGally)
```

In [28]:
```
employee <- read.csv("hr_dashboard_data.csv")
```

In [29]:
```
head(employee)
```

A data.frame: 6 × 11

	Name	Age	Gender	Projects.Completed	Productivity....	Satisfaction.Rate....	F
	<chr>	<int>	<chr>	<int>	<int>	<int>	
1	Douglas Lindsey	25	Male	11	57	25	
2	Anthony Roberson	59	Female	19	55	76	
3	Thomas Miller	30	Male	8	87	10	
4	Joshua Lewis	26	Female	1	53	4	
5	Stephanie Bailey	43	Male	14	3	9	
6	Jonathan King	24	Male	5	63	33	

In [30]:
```
names(employee)
```

'Name' · 'Age' · 'Gender' · 'Projects.Completed' · 'Productivity....' · 'Satisfaction.Rate....' · 'Feedback.Score' · 'Department' · 'Position' · 'Joining.Date' · 'Salary'

In [31]:
```
summary(employee)
```

```
     Name                Age            Gender          Projects.Completed
 Length:200         Min.   :22.00   Length:200         Min.   : 0.00
 Class :character   1st Qu.:26.00   Class :character   1st Qu.: 6.00
 Mode  :character   Median :32.00   Mode  :character   Median :11.00
                    Mean   :34.65                      Mean   :11.46
                    3rd Qu.:41.00                      3rd Qu.:17.00
                    Max.   :60.00                      Max.   :25.00
 Productivity.... Satisfaction.Rate.... Feedback.Score   Department
 Min.   : 0.00    Min.   :  0.00        Min.   :1.000   Length:200
 1st Qu.:23.00    1st Qu.: 25.75        1st Qu.:1.900   Class :character
 Median :45.00    Median : 50.50        Median :2.800   Mode  :character
 Mean   :46.76    Mean   : 49.94        Mean   :2.883
 3rd Qu.:70.00    3rd Qu.: 75.25        3rd Qu.:3.900
 Max.   :98.00    Max.   :100.00        Max.   :4.900
   Position          Joining.Date           Salary
 Length:200         Length:200         Min.   : 30231
 Class :character   Class :character   1st Qu.: 53080
 Mode  :character   Mode  :character   Median : 80540
                                       Mean   : 76619
                                       3rd Qu.:101108
                                       Max.   :119895
```

In [32]:
```
# Rename variable to have a humman readable.
new_col_name <- c("name", "age", "gender", "projects", "productivity", "sati

names(employee) <- new_col_name
```

In [33]:
```
head(employee)
```

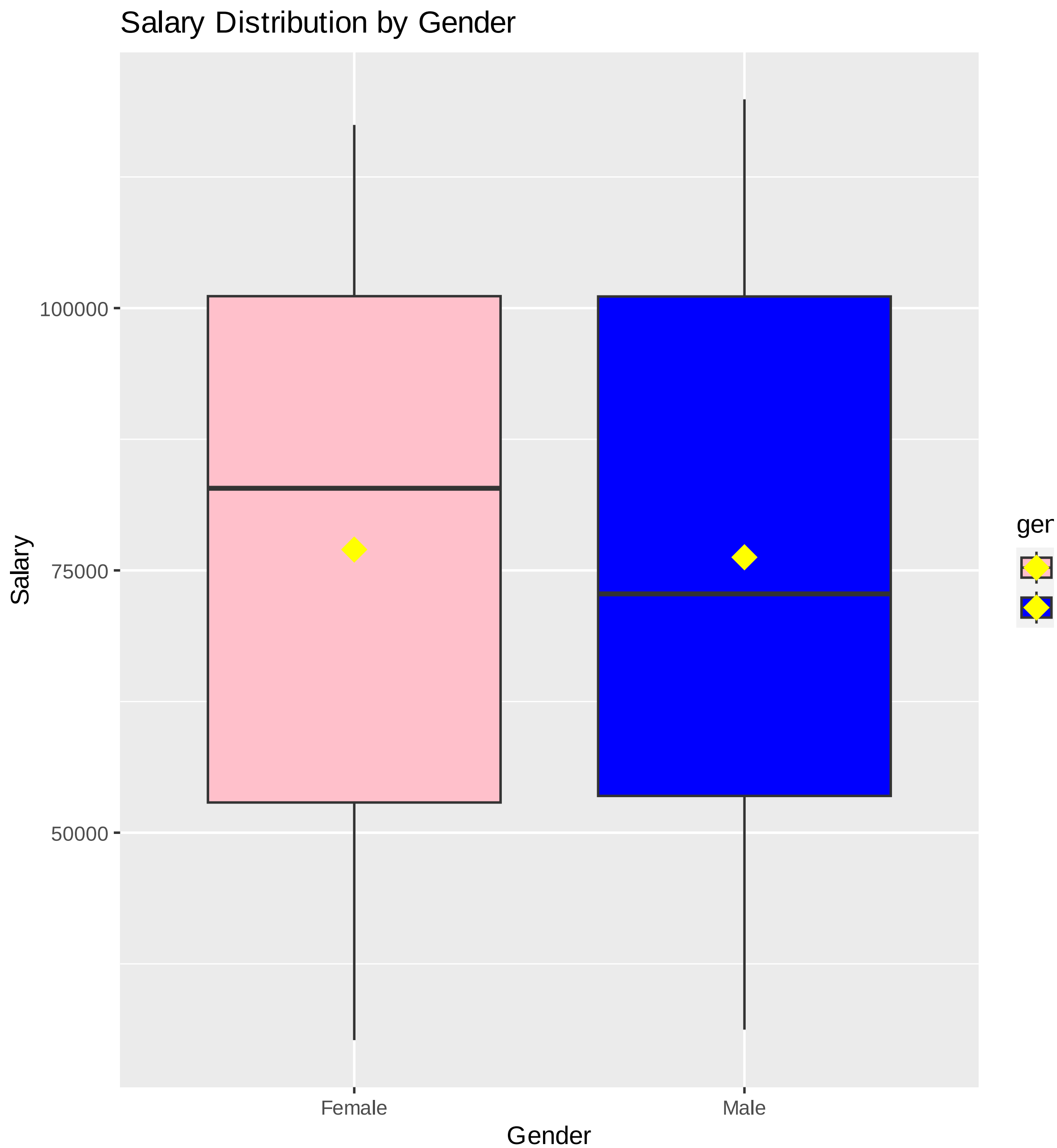
A data.frame: 6 × 11

	name	age	gender	projects	productivity	satisfaction	feedback	department
	<chr>	<int>	<chr>	<int>	<int>	<int>	<dbl>	<chr>
1	Douglas Lindsey	25	Male	11	57	25	4.7	Marketing
2	Anthony Roberson	59	Female	19	55	76	2.8	IT
3	Thomas Miller	30	Male	8	87	10	2.4	IT
4	Joshua Lewis	26	Female	1	53	4	1.4	Marketing
5	Stephanie Bailey	43	Male	14	3	9	4.5	IT
6	Jonathan King	24	Male	5	63	33	4.2	Sales

Boxplot Visualization

In [34]:
```
Salary_betweenGender_boxplot <- ggplot(employee, aes(x = gender, y = salary,
  geom_boxplot() +
```

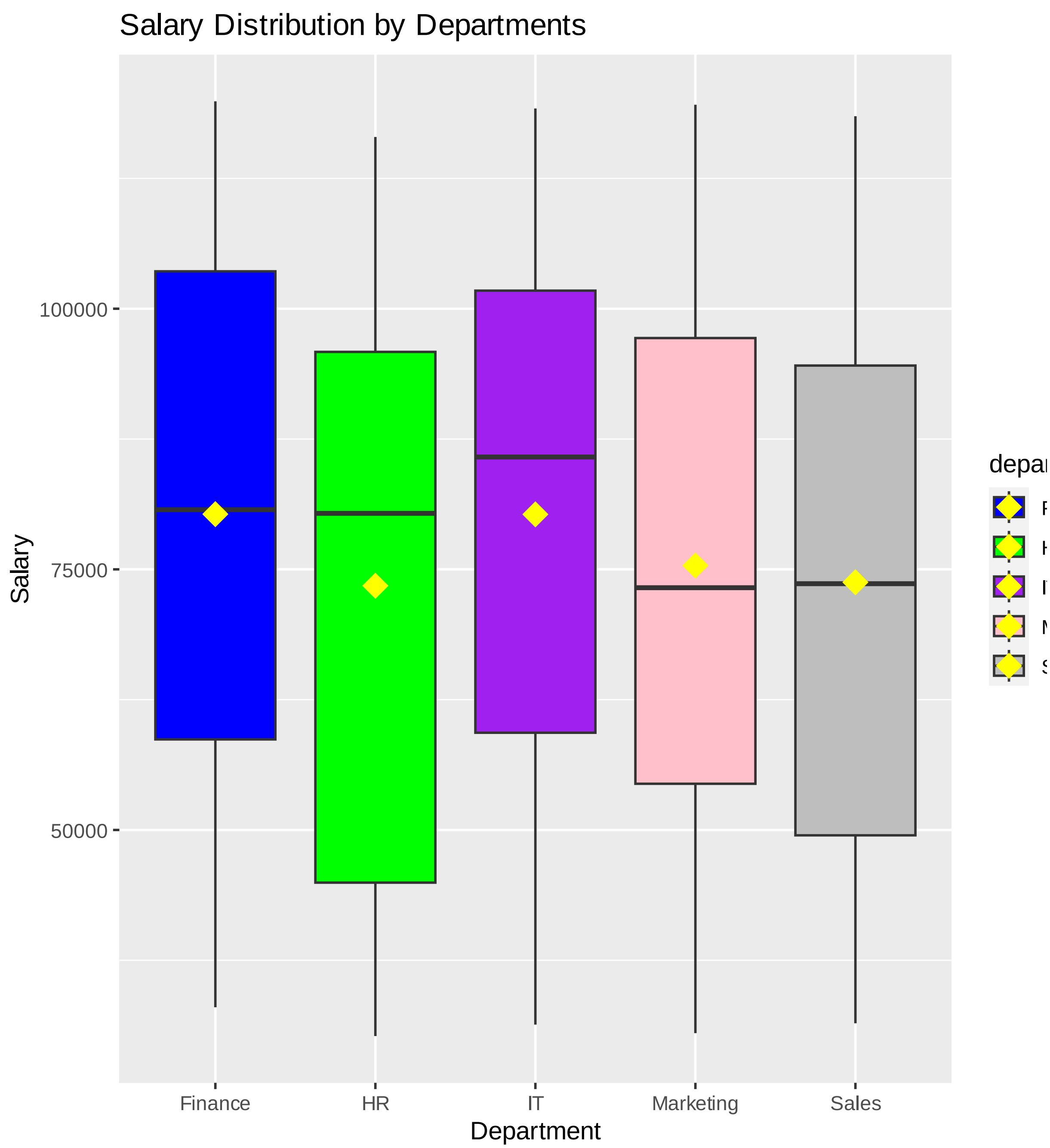
```
  labs(x = "Gender", y = "Salary", title = "Salary Distribution by Gender")
  scale_fill_manual(values = c("Male" = "blue", "Female" = "pink"))+
stat_summary(aes(gender, salary, fill = gender),
    fun = mean, colour = "yellow", geom = "point",
    shape = 18, size = 5
  )
Salary_betweenGender_boxplot
```



Analysis of Salary Distribution by Gender Boxplot

As the Salary_betweenGender_boxplot shows that the two boxplot almost perfectly overlapped and the mean salary of the male and female are almost same. The only obvious difference is the meidan value of salary that is the female median salary is approximately 10000 higher than the median male salary. It might mean there is almost no asscoication bewtween the salary and gender.

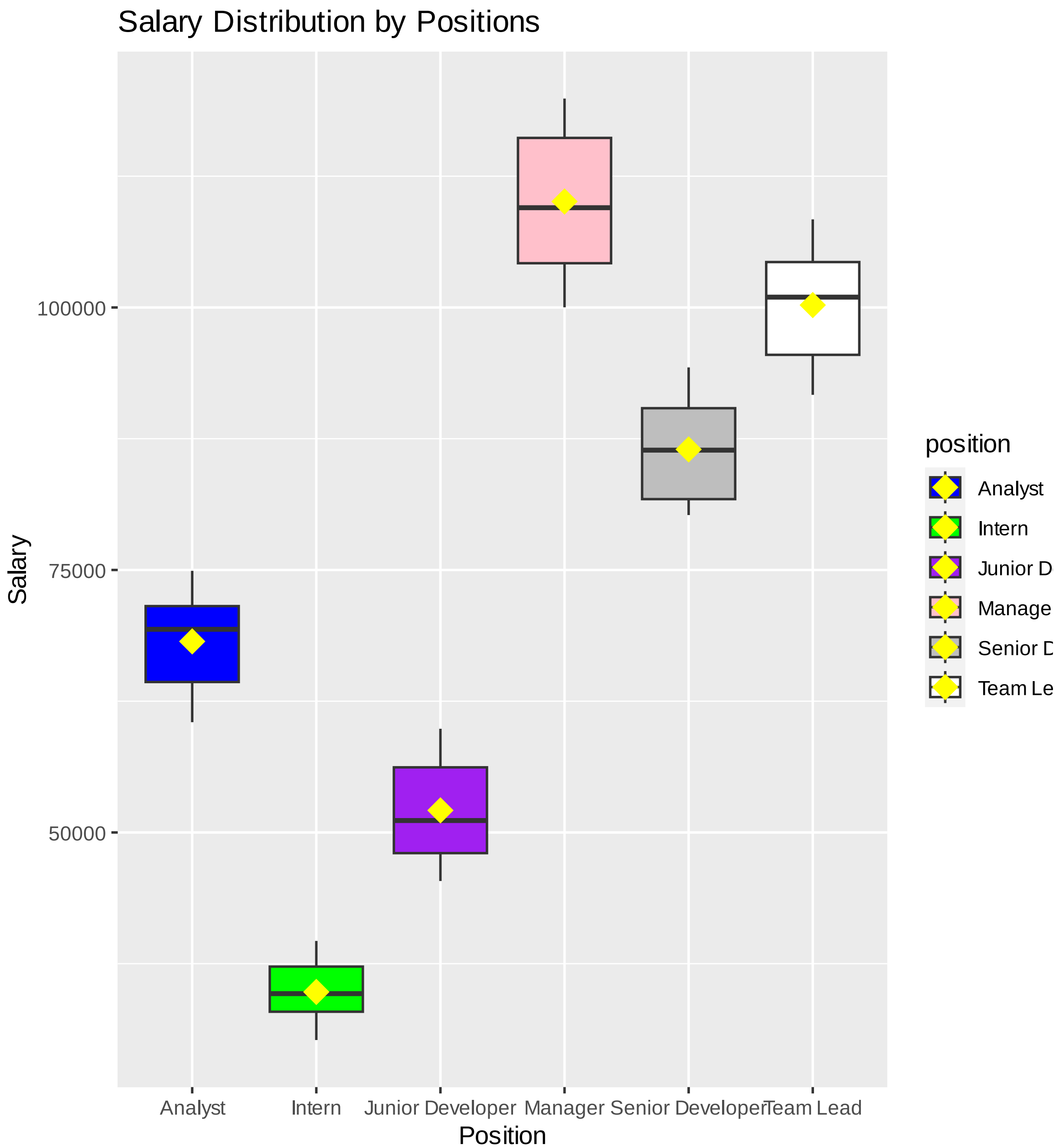
In [35]:
```
Salary_amongdepartment_boxplot <- ggplot(employee, aes(x = department, y = s
  geom_boxplot() +
  labs(x = "Department", y = "Salary", title = "Salary Distribution by Depar
  scale_fill_manual(values = c("Finance" = "blue", "HR" = "Green", "IT" = "p
stat_summary(aes(department, salary, fill = department),
    fun = mean, colour = "yellow", geom = "point",
    shape = 18, size = 5
  )
Salary_amongdepartment_boxplot
```



Analysis of Salary Distribution by Departments Boxplot

The side-by-side boxplot of Salary Distribution by Departments illustartes that there is not much difference in the quantile except the HR department. The lower quantile of the HR department is respectively much lower that the other department. In addition to that, there are also some differences among the median salary among the different departments. The meidan salary of the IT is the highest, then the finace and HR meidan salary is the secondly highest(They are approximately same). Lastly, the marketing and sales department are approximately same and they have the lowest meidan salary. In conclusin, the plot shows that there might be association between the salary and the department.

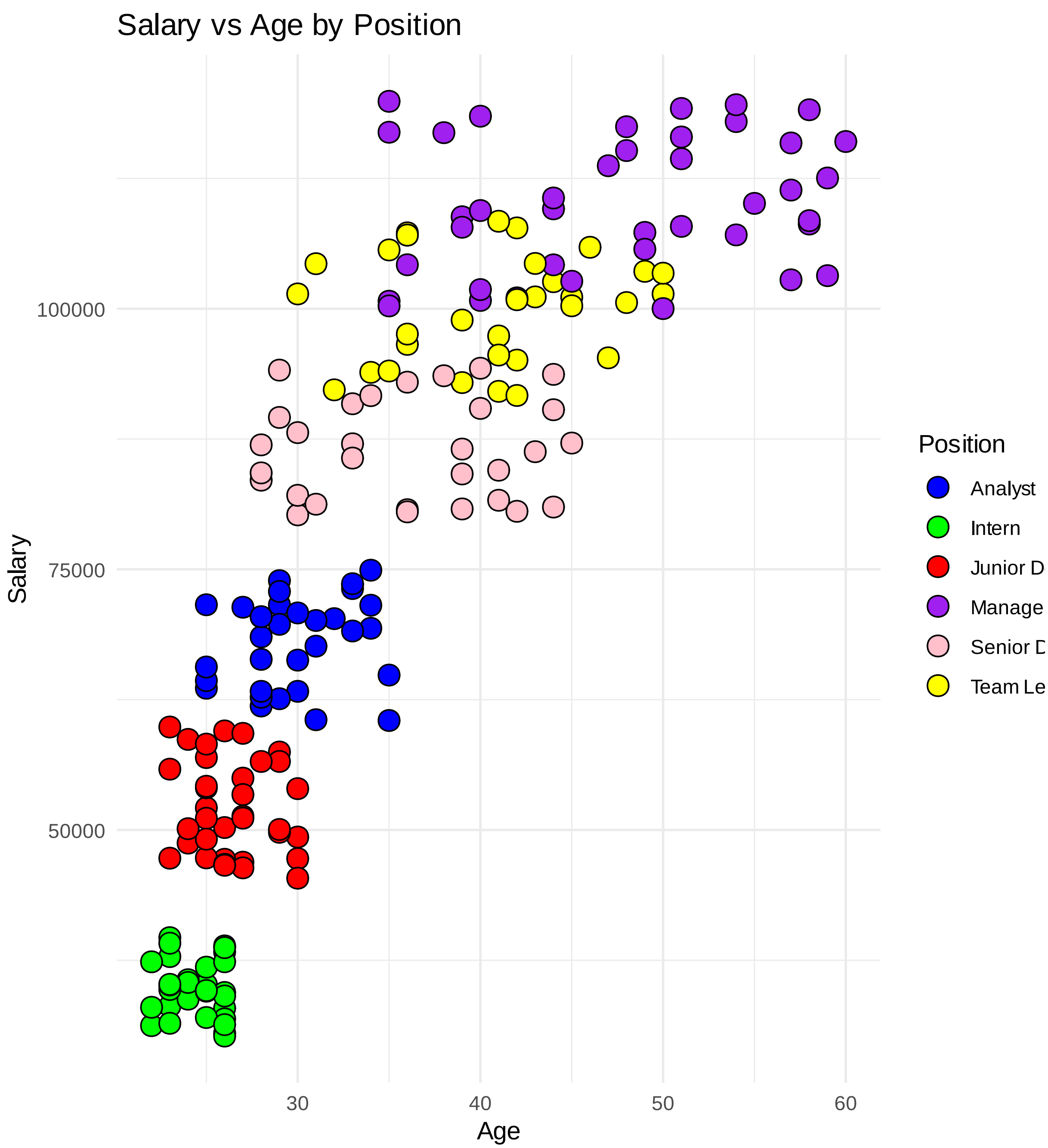
In [36]:
```
Salary_amongPostions_boxplot <- ggplot(employee, aes(x = position, y = salar
    geom_boxplot() +
    labs(x = "Position", y = "Salary", title = "Salary Distribution by Posit
    scale_fill_manual(values = c("Analyst" = "blue", "Intern" = "Green", "Ju
                                 "Manager" = "pink", "Senior Developer" = "G
    stat_summary(aes(position, salary, fill = position),
    fun = mean, colour = "yellow", geom = "point",
    shape = 18, size = 5
  )
Salary_amongPostions_boxplot
```



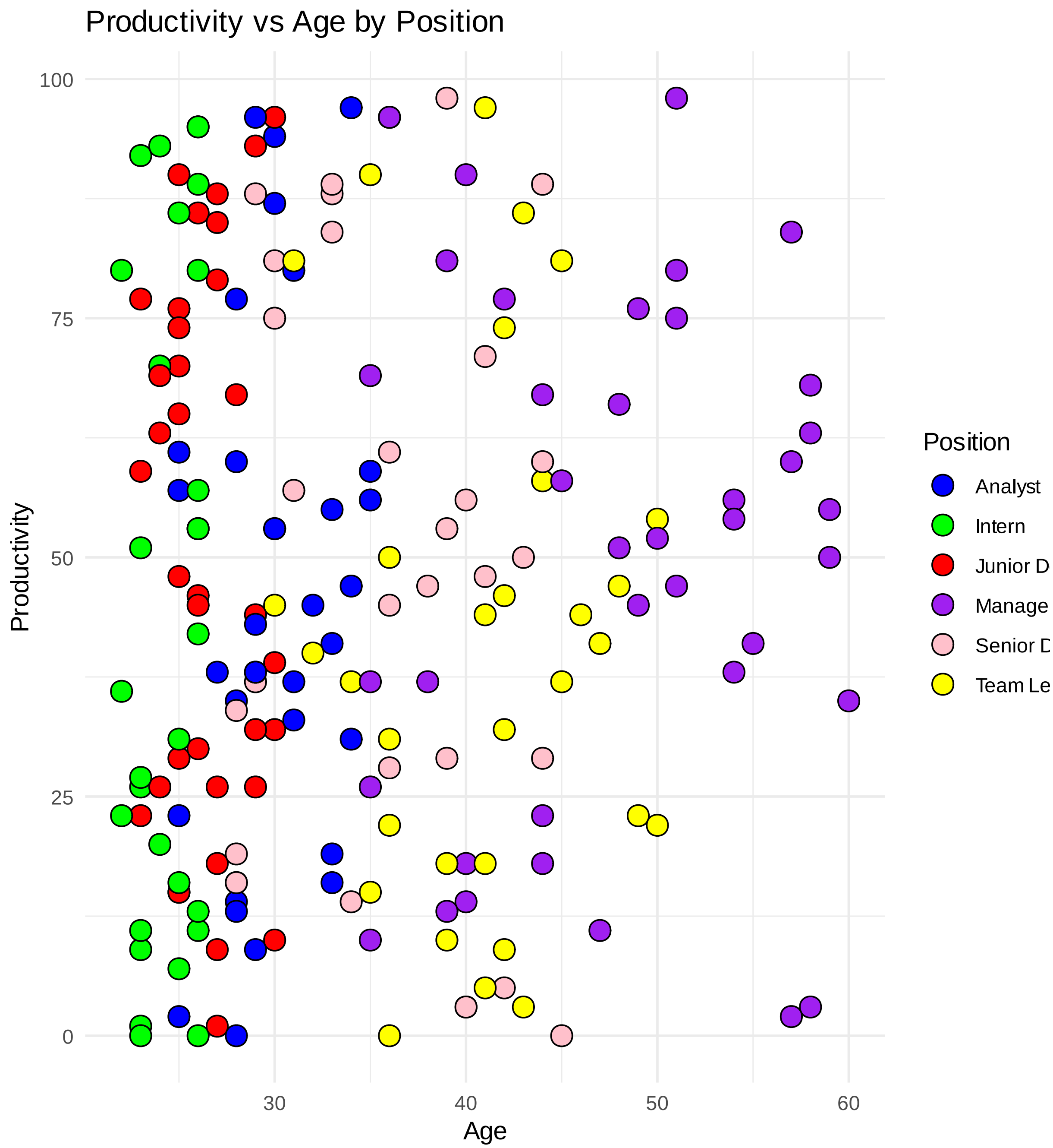
Analysis of Salary Distribution by Positions Boxplot Comparing to previous two boxplot. The side-by-side Boxplots of Salary of different Positions shows that there is almost no overlapping among these positions which means there might be a strong association beween the salary and position. In the futrue analysis, the association between the position and salary should be analyed deeper.

In [37]:
```
Salary_age_scatterplot <- ggplot(employee, aes(x = age, y = salary, fill = p
  geom_point(shape = 21, size = 4, color = "black") +
  scale_fill_manual(values = c("blue", "green", "red", "purple", "pink", "ye
  labs(title = "Salary vs Age by Position",
       x = "Age",
       y = "Salary",
       fill = "Position") +
```

```
        theme_minimal()
Salary_age_scatterplot
```



In [38]:
```
complete_age_scatterplot <- ggplot(employee, aes(x = age, y = productivity,
  geom_point(shape = 21, size = 4, color = "black") +
  scale_fill_manual(values = c("blue", "green", "red", "purple", "pink", "ye
  labs(title = "Productivity vs Age by Position",
       x = "Age",
       y = "Productivity",
       fill = "Position") +
        theme_minimal()
complete_age_scatterplot
```



In [40]:
```
numeric_vars <- employee %>%
  select_if(is.numeric)

# Calculate the correlation matrix
cor_matrix <- round(cor(numeric_vars),2)
cor_matrix
```

A matrix: 6 × 6 of type dbl

	age	projects	productivity	satisfaction	feedback	salary
age	1.00	0.76	0.02	0.04	0.01	0.83
projects	0.76	1.00	0.06	-0.01	0.08	0.87
productivity	0.02	0.06	1.00	0.05	-0.01	0.03
satisfaction	0.04	-0.01	0.05	1.00	0.01	-0.02
feedback	0.01	0.08	-0.01	0.01	1.00	0.03
salary	0.83	0.87	0.03	-0.02	0.03	1.00

In [41]:
```
library(car)
vif(lm(salary~ projects + age, employee))
```

```
Loading required package: carData


Attaching package: ‘car’


The following object is masked from ‘package:purrr’:

    some


The following object is masked from ‘package:dplyr’:

    recode

```

projects: 2.35358589273836 **age:** 2.35358589273836

Method and Plan

Question of Interest

What is the impact of age, projects and other relevant features on the salary of employee? Proposed Methods: Multiple Linear Regression, Logistic Regression, Stepwise Selection, Lasso Regression, Ridge Regression, Cross Validation.

1. Data Preparation:

- Clean the dataset, removing the missing dataset.
- Transform categorical variables into int if needed.

2. Exploratory Data Analysis (EDA):

- Conduct EDA to initially explore the relationship among the variables, identify the pattern among the variables such as linearity, and find out if there is a great difference in salary in different group in different categorical variables.

- Initially visualizing the data:

1. Create the boxplot to compare the distribution of different groups in the categorical variables.
2. Using the correlation heat map to peak at the correlation between the variables.
3. Getting the scatter plot of different variables by using ggpairs.

3. Multiple Linear Regression:

Fit a model to understand the linear relationship between asking prices and predictors. Assess the significance and impact of each predictor variable.

4. Stepwise Forward Selection:

After the exploratory data analysis, the analyzed predictor variables should be confirmed. Then the forward selection will be implemented with regsubsets() function.

5. Regularization Penalty:

After each predictor variable is added, apply the regularization penalty() to each model. In some of the case the coefficients might be zero because of Lasso regression. We will use the Ridge Regression and Lasso Regression These two regression both help shrinking our model coefficients which address the multicollinearity between the variables. And in r we could use glmnet() function for Ridge and Lasso regression.

VIF

VIF is also a measure of the amount of multicollinearity in regression analysis.

VIF < 5: Low collinearity 5 < VIF < 10: Moderate collinearity VIF > 10: High collinearity

6. Cross-Validation:

Cross validation technique will be performed to evaluate the performance for the model of each step. That will help prevent the overfitting situation and ensures the model fitting well the data. Our data will be split into 4 sets. One by one, each set will be selected as the test set and remaining will be used as training data. And finally the testing set with best parameter will be used to fit the model.

Limitation:

There are overfitting risks arising from the combination of regression techniques, multicollinearity, and stepwise selection. Additionally, stepwise forward selection biases and potential loss of interpretability in Ridge and Lasso regression are concerns. And there are some assumption of linear model which are needed to be considered in the model fitting such as linearity, independece, homoscedasticity and constant variance...

Pre-steps of Assignment4

In [42]:
```
head(numeric_vars)
head(employee)
```

A data.frame: 6 × 6

	age	projects	productivity	satisfaction	feedback	salary
	<int>	<int>	<int>	<int>	<dbl>	<int>
1	25	11	57	25	4.7	63596
2	59	19	55	76	2.8	112540
3	30	8	87	10	2.4	66292
4	26	1	53	4	1.4	38303
5	43	14	3	9	4.5	101133
6	24	5	63	33	4.2	48740

A data.frame: 6 × 11

	name	age	gender	projects	productivity	satisfaction	feedback	department
	<chr>	<int>	<chr>	<int>	<int>	<int>	<dbl>	<chr>
1	Douglas Lindsey	25	Male	11	57	25	4.7	Marketing
2	Anthony Roberson	59	Female	19	55	76	2.8	IT
3	Thomas Miller	30	Male	8	87	10	2.4	IT
4	Joshua Lewis	26	Female	1	53	4	1.4	Marketing
5	Stephanie Bailey	43	Male	14	3	9	4.5	IT
6	Jonathan King	24	Male	5	63	33	4.2	Sales

In [43]:
```
employee$gender <- ifelse(employee$gender == "Male", 1, 0)
```

In [44]:
```
head(employee)
```

A data.frame: 6 × 11

	name	age	gender	projects	productivity	satisfaction	feedback	department
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<dbl>	<chr>
1	Douglas Lindsey	25	1	11	57	25	4.7	Marketing
2	Anthony Roberson	59	0	19	55	76	2.8	IT
3	Thomas Miller	30	1	8	87	10	2.4	IT
4	Joshua Lewis	26	0	1	53	4	1.4	Marketing
5	Stephanie Bailey	43	1	14	3	9	4.5	IT
6	Jonathan King	24	1	5	63	33	4.2	Sales

In [52]:
```
# Lasso
library(rsample)
library(glmnet)
library(car)
employee_raw <- employee %>%
                select(-name, -joiningdate, -department, -position)

employee_split <- initial_split(employee_raw, prop = .6, strata = salary)
employee_selection <- training(employee_split)
employee_inference <- testing(employee_split)

lasso_model <-
    cv.glmnet(employee_selection %>% select(-salary) %>% as.matrix(),
              employee_selection %>% select(salary) %>% as.matrix(),
              alpha = 1)

lasso_model
beta_lasso <- coef(lasso_model, s = "lambda.min")
beta_lasso

lasso_selected_covariates <- as_tibble(
        as.matrix(beta_lasso),
        rownames='covariate') %>%
        filter(covariate != '(Intercept)' & abs(s1) !=0) %>%
        pull('covariate')

lasso_selected_covariates

lasso_variables_vif <-
   vif(lm(salary ~age + projects + gender, employee_selection))

lasso_variables_vif
inference_model <- lm(lm(salary ~age + projects, employee_inference))
summary(inference_model)
```

```
Call:  cv.glmnet(x = employee_selection %>% select(-salary) %>% as.matrix(),
y = employee_selection %>% select(salary) %>% as.matrix(),      alpha = 1)

Measure: Mean-Squared Error

    Lambda Index   Measure       SE Nonzero
min    984    35 134380153 11847278       3
1se   2738    24 144711816 15627614       2
7 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept)  10011.872
age           1157.480
gender        3185.792
projects      2271.734
productivity     .
satisfaction     .
feedback         .
```

'age' · 'gender' · 'projects'

age: 2.39914610028948 **projects:** 2.40864405163273 **gender:** 1.01768194226264

```
Call:
lm(formula = lm(salary ~ age + projects, employee_inference))

Residuals:
   Min     1Q Median     3Q    Max
-17017  -8495  -1177   7041  30510

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12750.3     4468.6   2.853  0.00555 **
age            969.5      174.3   5.562 3.71e-07 ***
projects      2511.9      266.8   9.415 1.91e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10430 on 77 degrees of freedom
Multiple R-squared:  0.8571,	Adjusted R-squared:  0.8534
F-statistic: 230.9 on 2 and 77 DF,  p-value: < 2.2e-16
```

In [51]:

```
head(employee)
employee_for <- employee %>%
                select(-name, -joiningdate, -department, -position)
training_employee <- sample_n(employee_for, size = nrow(employee_for) * 0.70
  replace = FALSE
)
testing_employee <- anti_join(employee_for,
  training_employee
)
head(training_employee)
head(testing_employee)


# Calculate the correlation matrix
cor_matrix <- round(cor(employee_for),2)
cor_matrix
```

A data.frame: 6 × 11

	name	age	gender	projects	productivity	satisfaction	feedback	department
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<dbl>	<chr>
1	Douglas Lindsey	25	1	11	57	25	4.7	Marketing
2	Anthony Roberson	59	0	19	55	76	2.8	IT
3	Thomas Miller	30	1	8	87	10	2.4	IT
4	Joshua Lewis	26	0	1	53	4	1.4	Marketing
5	Stephanie Bailey	43	1	14	3	9	4.5	IT
6	Jonathan King	24	1	5	63	33	4.2	Sales

```
Joining with `by = join_by(age, gender, projects, productivity, satisfactio
n,
feedback, salary)`
```

A data.frame: 6 × 7

	age	gender	projects	productivity	satisfaction	feedback	salary
	<int>	<dbl>	<int>	<int>	<int>	<dbl>	<int>
1	40	1	22	18	68	4.7	100795
2	44	0	11	29	17	1.3	90310
3	25	1	4	31	90	2.8	32010
4	29	1	18	88	8	1.8	89571
5	23	0	9	59	11	4.9	55833
6	48	1	17	66	4	2.0	115170

A data.frame: 6 × 7

	age	gender	projects	productivity	satisfaction	feedback	salary
	<int>	<dbl>	<int>	<int>	<int>	<dbl>	<int>
1	59	0	19	55	76	2.8	112540
2	23	0	4	92	68	2.8	39670
3	25	0	2	15	97	1.8	35169
4	36	1	12	22	66	2.2	107279
5	23	1	2	1	17	4.4	37855
6	25	1	10	29	73	2.0	52122

A matrix: 7 × 7 of type dbl

	age	gender	projects	productivity	satisfaction	feedback	salary
age	1.00	-0.08	0.76	0.02	0.04	0.01	0.83
gender	-0.08	1.00	-0.08	0.13	-0.05	-0.11	-0.01
projects	0.76	-0.08	1.00	0.06	-0.01	0.08	0.87
productivity	0.02	0.13	0.06	1.00	0.05	-0.01	0.03
satisfaction	0.04	-0.05	-0.01	0.05	1.00	0.01	-0.02
feedback	0.01	-0.11	0.08	-0.01	0.01	1.00	0.03
salary	0.83	-0.01	0.87	0.03	-0.02	0.03	1.00

In [48]:

```
library(leaps)
employee_forward_sel <- regsubsets(
  salary~., nvmax = 6 ,
  data = training_employee,
  method = "forward"
)
employee_fwd_summary <- summary(employee_forward_sel)
employee_fwd_summary
employee_fwd_summary <- tibble(
   n_input_variables = 1:6,
   RSS = employee_fwd_summary$rss,
   BIC = employee_fwd_summary$bic,
   Cp = employee_fwd_summary$cp
)
employee_fwd_summary
cp_min = which.min(employee_fwd_summary$Cp)
cp_min
names(coef(employee_forward_sel, cp_min))
selected_var <- names(coef(employee_forward_sel, cp_min))[-1]
selected_var
```

```
Subset selection object
Call: regsubsets.formula(salary ~ ., nvmax = 6, data = training_employee,
    method = "forward")
6 Variables  (and intercept)
             Forced in Forced out
age              FALSE      FALSE
gender           FALSE      FALSE
projects         FALSE      FALSE
productivity     FALSE      FALSE
satisfaction     FALSE      FALSE
feedback         FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: forward
         age gender projects productivity satisfaction feedback
1  ( 1 ) " " " "    "*"      " "          " "          " "
2  ( 1 ) "*" " "    "*"      " "          " "          " "
3  ( 1 ) "*" "*"    "*"      " "          " "          " "
4  ( 1 ) "*" "*"    "*"      " "          "*"          " "
5  ( 1 ) "*" "*"    "*"      "*"          "*"          " "
6  ( 1 ) "*" "*"    "*"      "*"          "*"          "*"
```

A tibble: 6 × 4

n_input_variables	RSS	BIC	Cp
<int>	<dbl>	<dbl>	<dbl>
1	23352113074	-192.7138	58.300552
2	17055993325	-231.7582	7.913878
3	16228078147	-233.7828	3.025234
4	16099646017	-229.9535	3.956619
5	16000438625	-225.8772	5.131167
6	15984674296	-221.0736	7.000000

3

'(Intercept)' · 'age' · 'gender' · 'projects'

'age' · 'gender' · 'projects'

Assignment 4 Computational Code and Output

Implementation of a proposed model"

In [68]:

```
employee_raw <- employee %>%
                select(-name, -joiningdate, -department, -position)

employee_split <- initial_split(employee_raw, prop = .6, strata = salary)
employee_selection <- training(employee_split)
employee_inference <- testing(employee_split)
```

```
lasso_model <-
    cv.glmnet(employee_selection %>% select(-salary) %>% as.matrix(),
              employee_selection %>% select(salary) %>% as.matrix(),
              alpha = 1)

lasso_model
beta_lasso <- coef(lasso_model, s = "lambda.min")
beta_lasso

lasso_selected_covariates <- as_tibble(
        as.matrix(beta_lasso),
        rownames='covariate') %>%
        filter(covariate != '(Intercept)' & abs(s1) !=0) %>%
        pull('covariate')

lasso_selected_covariates

lasso_variables_vif <-
   vif(lm(salary ~age + projects + gender, employee_selection))

lasso_variables_vif
inference_model <- lm(salary ~age + projects + gender, employee_inference)
inference_model_summary <- summary(inference_model)
```

```
Call:  cv.glmnet(x = employee_selection %>% select(-salary) %>% as.matrix(),
y = employee_selection %>% select(salary) %>% as.matrix(),      alpha = 1)

Measure: Mean-Squared Error

    Lambda Index   Measure       SE Nonzero
min   1171    33 131717755 21835904       3
1se   4309    19 149600300 23334297       2
7 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)  14477.74352
age           1104.56002
gender           .
projects      2155.58848
productivity     .
satisfaction   -18.13531
feedback         .
```

'age' · 'projects' · 'satisfaction'

age: 2.2372865772822 **projects:** 2.21839590173824 **gender:** 1.02671234952738

Result

In [72]:
```
tidy(inference_model_summary)
```

A tibble: 4 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	7865.418	4847.0313	1.622729	1.087886e-01
age	1014.513	193.4678	5.243834	1.371549e-06
projects	2664.665	319.3427	8.344217	2.411556e-12
gender	5609.154	2444.5790	2.294527	2.452226e-02

The Interpretation of Result.

The result above is derived from lasso regression on the training set and applied to the inference set, predicts salary based on age, projects, and gender. The coefficients suggest that, on average ,for constant other variables, each additional year of age is associated with a salary increase of 1014.513 unit, and each additional completed project is associated with a salary increase of 2664.665 unit. This is just as expected, it is shown in the previous part that the correlation between age and salary and the correlation between projects and salary are relatively high compared to other variables. Surprisingly, the p-value of gender is 2.45e-05(less than .05) implying that the gender is statistically significant to the response variable salary, contrary to initial expectations from exploratory data analysis (EDA).