CourseworkSAP

Thitirat Meekaewkunchorn

10 August 2022

Preamble

```
# Loading relevant libraries
library("here")
library("tidyverse")
library("magrittr")
library("janitor")
library("lubridate")
library("gridExtra")
library("glmnet")
library("readxl")
library("lindia")
library("lme4")
library("caret")
library("pROC")
library("sandwich")
```

Loading and cleaning data

```
# Reading in the data, cleaning column names, make missing values identifiable in dif
ferent ways
hr_data <- clean_names(read_csv(here("data/HRDataset_v14.csv"),
                       na = c("?", "", "NA", "N/A", "Na"), ))
```

```
# Making adjustments to some column names to make them standardised
hr_data <- hr_data %>%
  rename(term_id = termd,
         date_of_hire = dateof_hire,
         date_of_termination = dateof_termination,
         date_of_birth = dob,
         zip_code = zip)
```

Dealing with missing values

```
# Checking missing values
hr_data %>% sapply(function(x) sum(is.na(x)))
```

```
##                employee_name                          emp_id
##                            0                               0
##                   married_id               marital_status_id
##                            0                               0
##                    gender_id                   emp_status_id
##                            0                               0
##                      dept_id                    perf_score_id
##                            0                               0
##   from_diversity_job_fair_id                          salary
##                            0                               0
##                      term_id                     position_id
##                            0                               0
##                     position                           state
##                            0                               0
##                     zip_code                   date_of_birth
##                            0                               0
##                          sex                    marital_desc
##                            0                               0
##                 citizen_desc                 hispanic_latino
##                            0                               0
##                    race_desc                    date_of_hire
##                            0                               0
##          date_of_termination                     term_reason
##                          207                               0
##            employment_status                      department
##                            0                               0
##                 manager_name                      manager_id
##                            0                               8
##           recruitment_source               performance_score
##                            0                               0
##            engagement_survey                emp_satisfaction
##                            0                               0
##       special_projects_count last_performance_review_date
##                            0                               0
##              days_late_last30                        absences
##                            0                               0
```

```
# Checking all incomplete cases
view(hr_data %>% filter(!complete.cases(.)))
```

There are a total of 207 incomplete cases with missing values in columns “manager_id” and “date_of_termination”

- We can see that the missing values in the column “manager_id” is for the manager named “Webster Butler”, upon further investigation, we can see that Webster Butler’s “manager_id” is 39, which we will replace the missing value with.
- Missing values in the “date_of_termination” indicates the employee is still employed at the company when checking against “term_reason”, therefore we will keep the incomplete cases as this will not affect our analysis to keep them, otherwise we will have very little data to work with which could lead to less accuracy for prediction. Additionally, we will not be looking at the “date_of_termination” as a predictor variable.

```r
# Replace the missing values or NA in manager_id with 39
hr_data <- hr_data %>%
  mutate(manager_id = replace_na(manager_id, 39))
```

```r
# Checking if the missing values in the "manager_id" column has been changed.
hr_data %>% sapply(function(x) sum(is.na(x)))
```

```
##               employee_name                       emp_id 
##                           0                            0 
##                  married_id            marital_status_id 
##                           0                            0 
##                   gender_id                emp_status_id 
##                           0                            0 
##                     dept_id                perf_score_id 
##                           0                            0 
##   from_diversity_job_fair_id                       salary 
##                           0                            0 
##                     term_id                  position_id 
##                           0                            0 
##                    position                        state 
##                           0                            0 
##                    zip_code                date_of_birth 
##                           0                            0 
##                         sex                marital_desc 
##                           0                            0 
##                citizen_desc              hispanic_latino 
##                           0                            0 
##                   race_desc                 date_of_hire 
##                           0                            0 
##         date_of_termination                  term_reason 
##                         207                            0 
##           employment_status                   department 
##                           0                            0 
##                manager_name                   manager_id 
##                           0                            0 
##          recruitment_source            performance_score 
##                           0                            0 
##           engagement_survey             emp_satisfaction 
##                           0                            0 
##       special_projects_count last_performance_review_date 
##                           0                            0 
##             days_late_last30                     absences 
##                           0                            0
```

Converting variables to appropriate data type

```
# Converting categorical variables to factors, date variables to datetime and zip var
iable to numerical type
hr_data <- hr_data %>%
  mutate_at(vars(contains("_id"), contains("_desc"),
                 position, state, sex, hispanic_latino, term_reason, employment_statu
s,
                 department, recruitment_source, performance_score, emp_satisfaction,
                 employee_name, manager_name),
            list(factor)) %>%
  mutate_at(vars(contains("date")),
            lubridate::dmy) %>%
  mutate(zip_code = as.numeric(zip_code))
```

Removing duplicates

```
hr_data <- hr_data %>% distinct()
```

Recoding variables

```
# Checking each factor levels
levels(hr_data$hispanic_latino)
```

```
## [1] "no"  "No"  "yes" "Yes"
```

```
# Recoding the "hispanic_latino" variable to make the level names consistent, recodin
g the levels for "sex" to "Male" and "Female" and "term_id" levels to "Active" and 'T
erminated" for readability.
hr_data <- hr_data %>%
  mutate(hispanic_latino = recode(hispanic_latino,
                                            "yes" = "Yes",
                                            "no" = "No"))%>%
  mutate(sex = recode(sex,
                      "M" = "Male",
                      "F" = "Female"))  %>%
  mutate(term_id = recode(term_id,
                          "0" = "Active",
                          "1" = "Terminated"))
```

Task 1 - Carefully constructed numerical and graphical summaries (using ggplot) of 5 relevant variables. [10 points]

```
# Selecting 5 relevant variables
hr_5vars <- hr_data %>%
  select(absences, days_late_last30, performance_score, engagement_survey, employment
_status)

# Numerical summary using the summary() function
summary(hr_5vars)
```

```
##     absences     days_late_last30         performance_score engagement_survey
##  Min.   : 1.00   Min.   :0.0000   Exceeds           : 37     Min.   :1.12
##  1st Qu.: 5.00   1st Qu.:0.0000   Fully Meets       :243     1st Qu.:3.69
##  Median :10.00   Median :0.0000   Needs Improvement: 18     Median :4.28
##  Mean   :10.24   Mean   :0.4148   PIP               : 13     Mean   :4.11
##  3rd Qu.:15.00   3rd Qu.:0.0000                              3rd Qu.:4.70
##  Max.   :20.00   Max.   :6.0000                              Max.   :5.00
##               employment_status
##  Active                 :207
##  Terminated for Cause  : 16
##  Voluntarily Terminated: 88
##
##
##
```

```
print('standard deviation for absences:')
```

```
## [1] "standard deviation for absences:"
```

```
sd(hr_5vars$absences)
```

```
## [1] 5.852596
```

```
print('standard deviation for days_late_last30:')
```

```
## [1] "standard deviation for days_late_last30:"
```

```
sd(hr_5vars$days_late_last30)
```

```
## [1] 1.294519
```

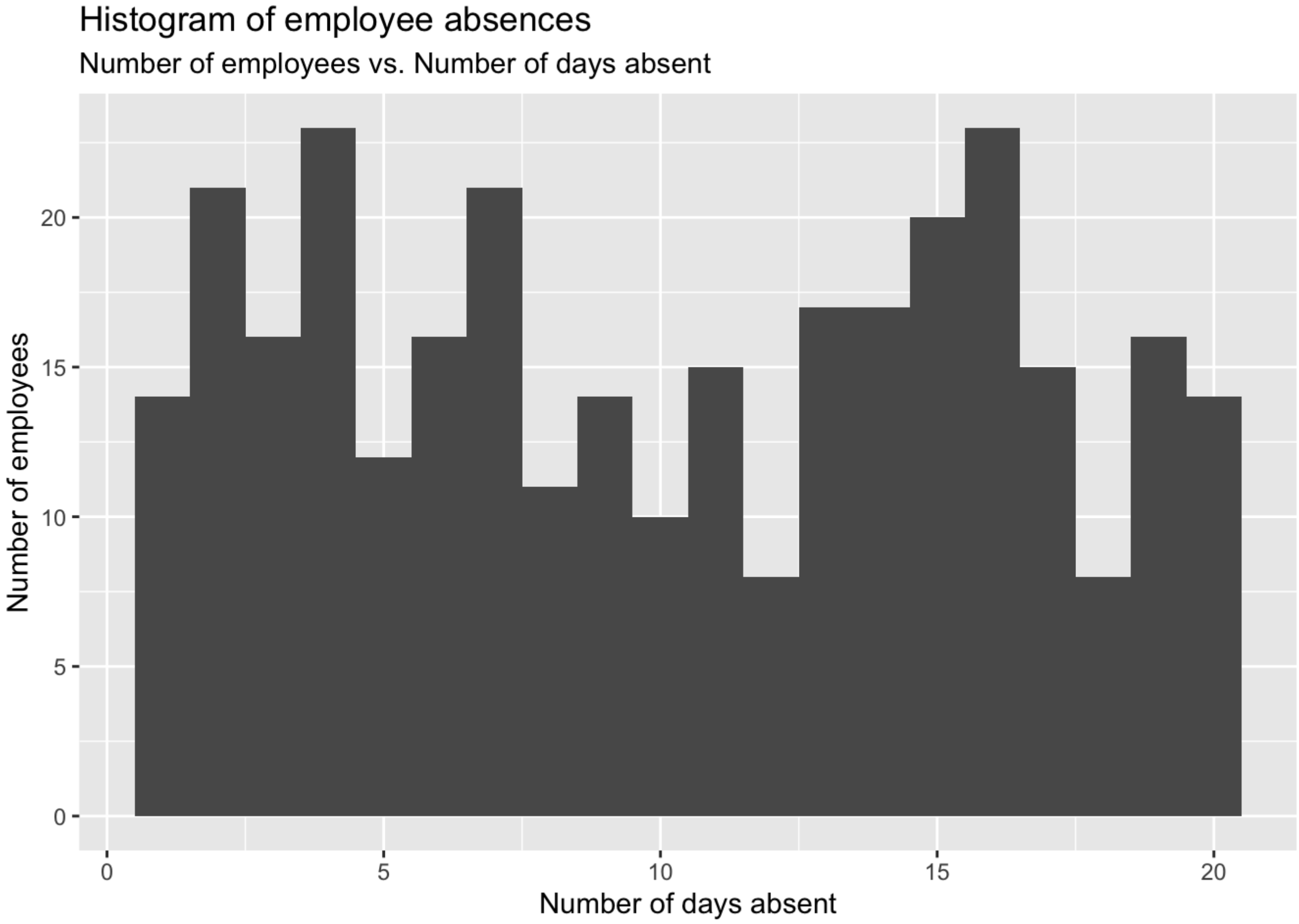
```
print('standard deviation for engagement_survey:')
```

```
## [1] "standard deviation for engagement_survey:"
```

```
sd(hr_5vars$engagement_survey)
```

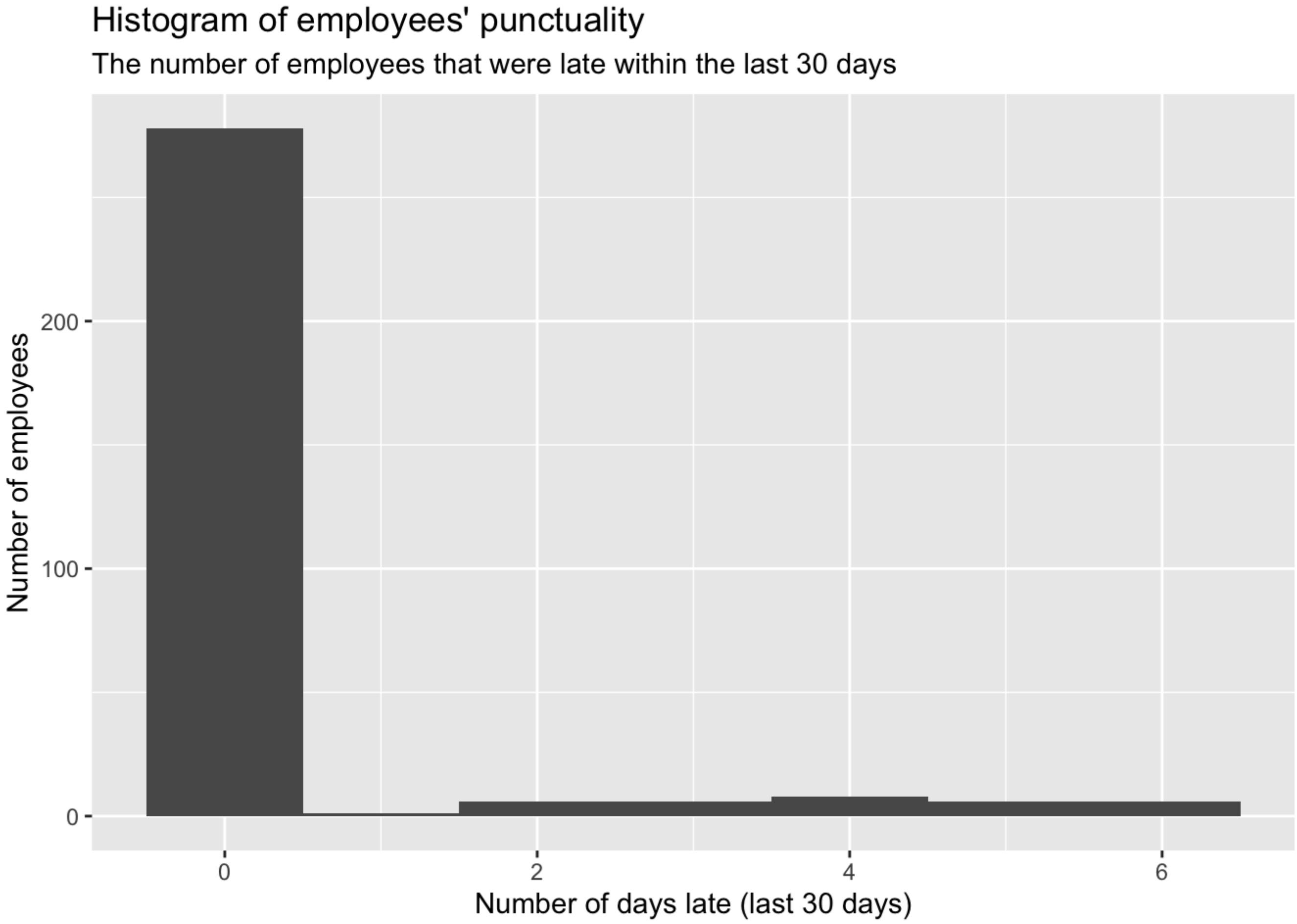
```
## [1] 0.7899375
```

```
# Graphical summary for absences
hr_5vars %>%
  ggplot(aes(absences)) +
  geom_histogram(binwidth = 1) +
    labs(x = "Number of days absent",
         y = "Number of employees",
       title="Histogram of employee absences",
       subtitle="Number of employees vs. Number of days absent")
```



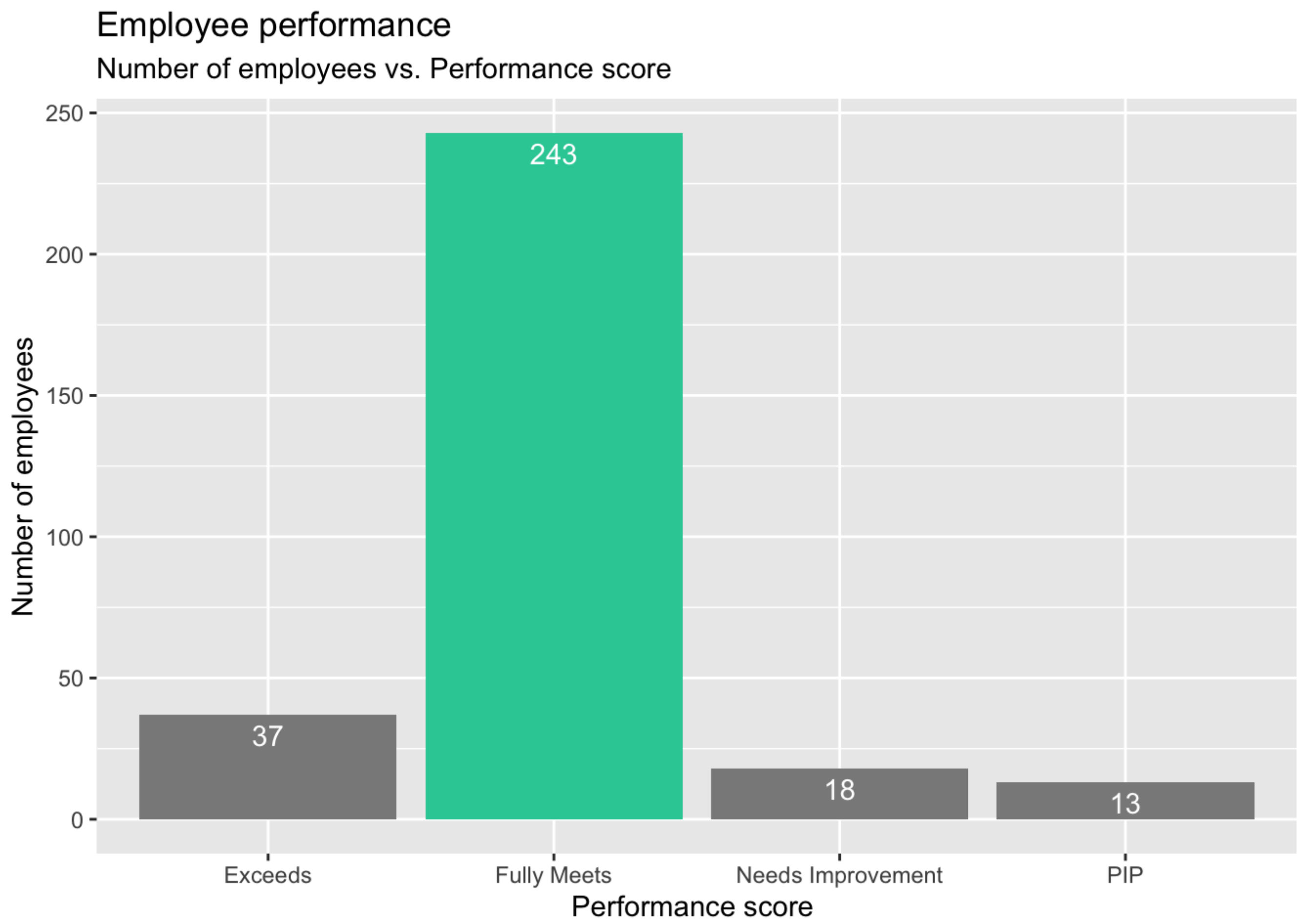
Analysis: The histogram of employee absences shows a random distribution, with no obvious pattern. This is expected due to reasons of absences and their duration will vary by nature. For example people falling ill or taking their entitled annual leave (holidays) at different times of year for different lengths of time.

```
# Graphical summary for days_late_last30
hr_5vars %>%
  ggplot(aes(days_late_last30)) +
  geom_histogram(binwidth = 1) +
    labs(x = "Number of days late (last 30 days)",
         y = "Number of employees",
       title="Histogram of employees' punctuality",
       subtitle="The number of employees that were late within the last 30 days")
```



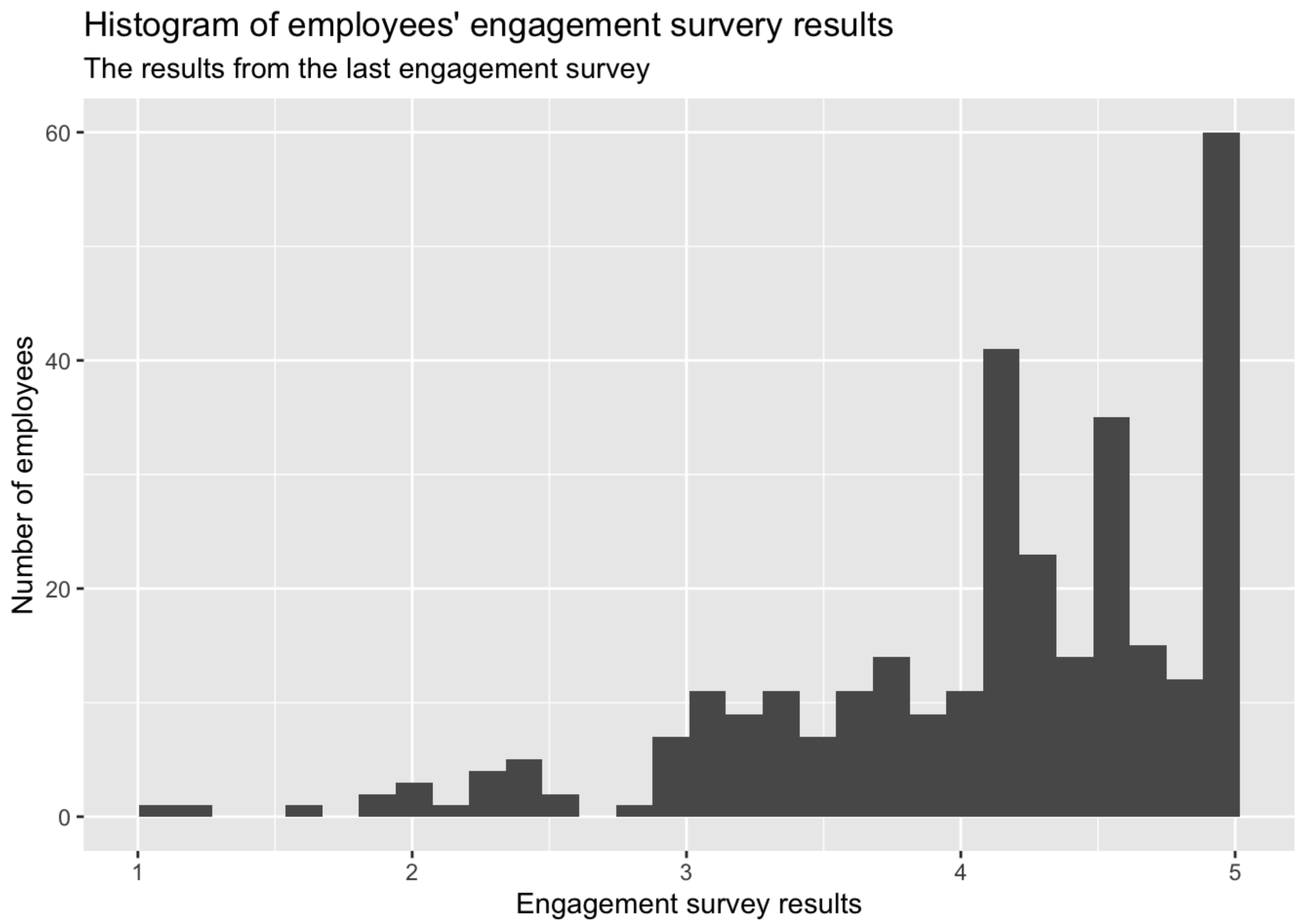
Analysis: 278 employees were punctual within the last 30 days, meaning they had 0 days late. 1 employee was late once within the last 30 days, 6 employees were late twice within the last 30 days, 6 were late three times, 8 employees were late 4 times, 6 were late 5 times and 6 were late 6 days within the last 30 days. We should be mindful regarding employees who have been late at least 4 times within 30 days as this would make them on average late once a week. Punctuality is important as it reflects reliable, dependable and consistent employees.

```
# Graphical summary for performance_score
hr_5vars %>%
  ggplot(aes(performance_score, fill = performance_score)) +
  geom_bar() +
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, colour = "white") +
  theme(legend.position = "none") +
  scale_fill_manual(values=c("#8A8A8A", "#2CCDA4", "#8A8A8A", "#8A8A8A")) +
  labs(x="Performance score ",
       y="Number of employees",
       title="Employee performance",
       subtitle="Number of employees vs. Performance score")
```



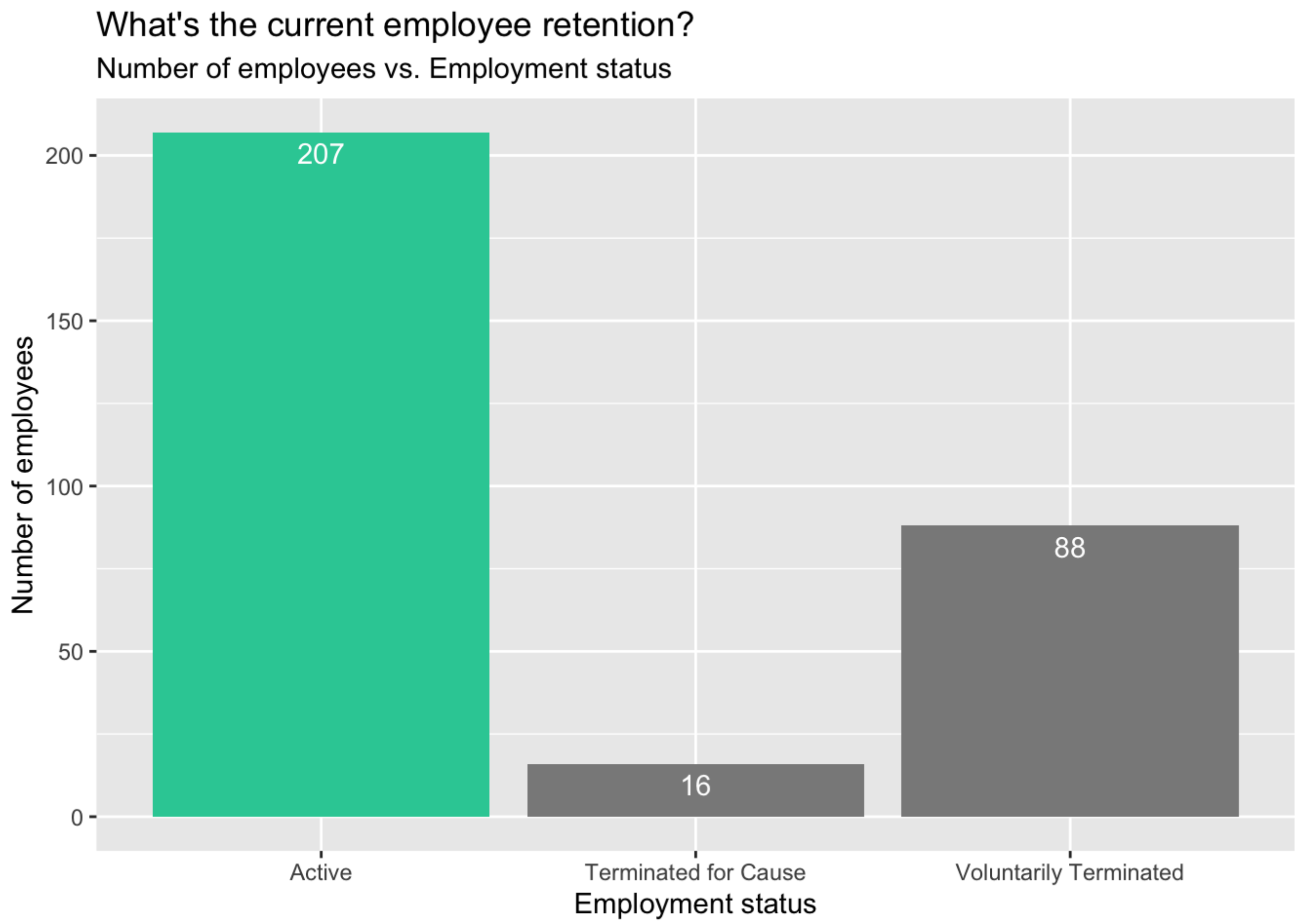
Analysis: 280 or around 90% of all employees in the data set fully meets or has exceeded in their performance which is a good sign, however 31 employees or around 10% needs improvement or are on a Performance Improvement Plan (PIP), these are the employees that may require more support during their tenure to help achieve their performance goals.

```
# Graphical summary for engagement_survey
hr_5vars %>%
  ggplot(aes(engagement_survey)) +
  geom_histogram() +
    labs(x = "Engagement survey results",
         y = "Number of employees",
       title="Histogram of employees' engagement survery results",
       subtitle="The results from the last engagement survey")
```



Analysis: This histogram is skewed left showing a most employees gave a score of between 4 and 5, with fewer scoring between 3 and 4 and less between 1 and 3.

```
# Graphical summary for employment_status
hr_5vars %>%
  ggplot(aes(employment_status, fill = employment_status)) +
  geom_bar() +
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, colour = "white") +
  theme(legend.position = "none") +
  scale_fill_manual(values=c("#2CCDA4", "#8A8A8A", "#8A8A8A")) +
  labs(x="Employment status",
       y="Number of employees",
       title="What's the current employee retention?",
       subtitle="Number of employees vs. Employment status")
```



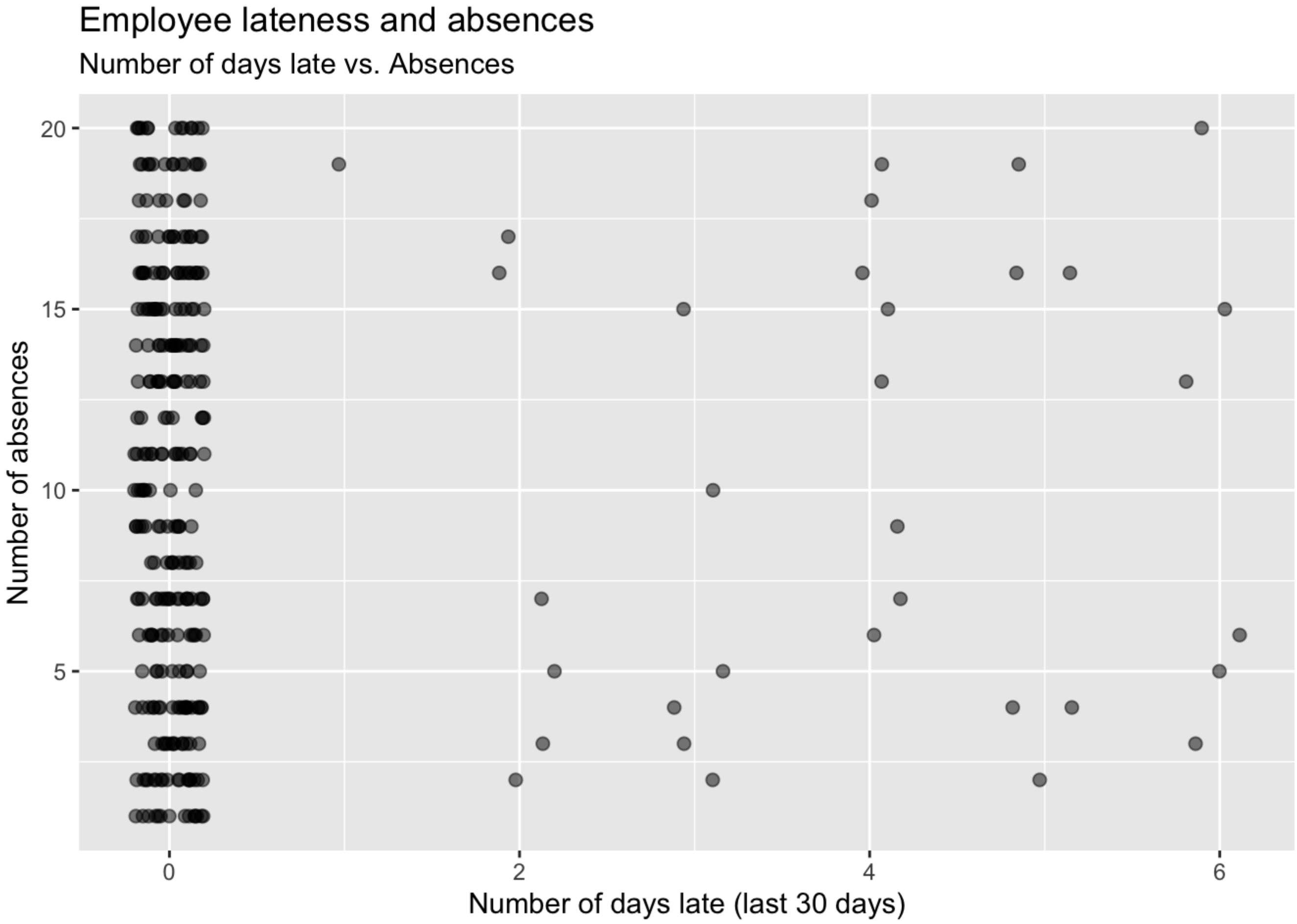
Analysis: The graphs shows that we currently have 207 out of 311 employees still with us, which means we still have around 67% employees active, 16 or around 5% of the employees were terminated for cause and 88 employees or around 28% who have voluntarily terminated. It would be interesting to see the different contributing factors that causes of the different terminations.

Using ggplot, construct five plots depicting the relationship between: [20 points, 4 points per plot]

i. Plot 1: Two quantitative variables

```
late_vs_absence <- hr_data %>%
  group_by(employee_name, emp_id, absences, days_late_last30, employment_status) %>%
  summarise()

late_vs_absence %>%
  ggplot(aes(jitter(days_late_last30), absences)) +
  geom_point(alpha=0.5, size=2) +
  labs(x="Number of days late (last 30 days)",
       y="Number of absences",
       title="Employee lateness and absences",
       subtitle="Number of days late vs. Absences")
```

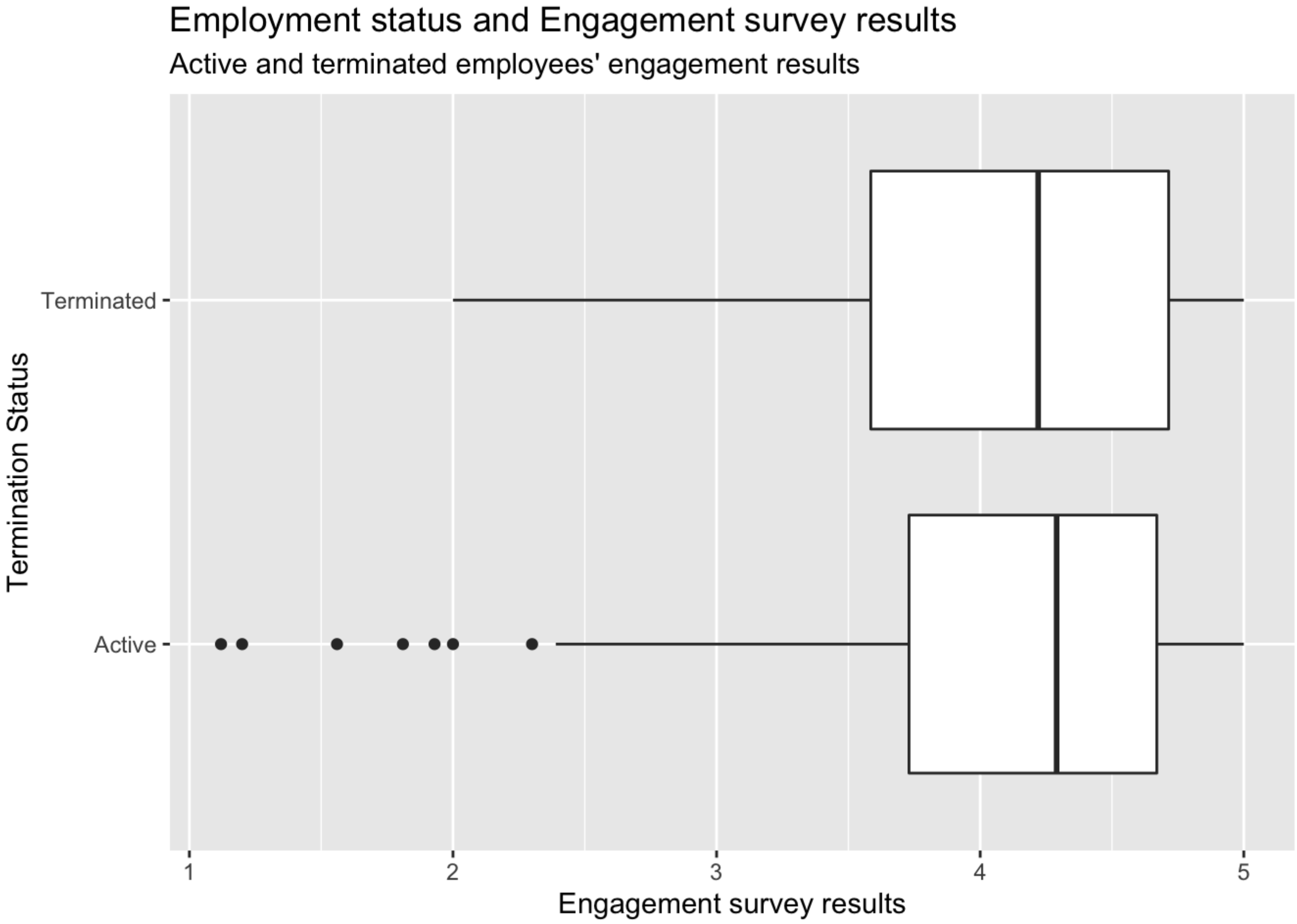


Analysis: This scatterplot shows that there are more punctual employees with 0 days late in the last 30 days, however all employees recorded in this dataset have had at least one day absent from work. Since the reason for absence recorded were not specified, e.g. illness, holiday or other, we cannot know if the higher number of absences is a cause for concern. Nevertheless we should be mindful of employees who have had at least 4 days late in the last 30 days as this means they have been late on average of one day a week, in usual workplace circumstances this does not reflect positively on employees' professionalism.

ii. Plot 2: A factor and a quantitative variable

```
# term_id vs. engagement_survey
termination_vs_engagement <- hr_data %>%
  group_by(term_id) %>%
  summarise(engagement_survey)

termination_vs_engagement %>%
  ggplot(aes(engagement_survey, term_id)) +
  geom_boxplot() +
    labs(x="Engagement survey results",
       y="Termination Status",
       title="Employment status and Engagement survey results",
       subtitle="Active and terminated employees' engagement results")
```



Analysis: The range of the engagement results is 4, where the: median engagement survey results for terminated employees is around 4.2 and for active employees is around 4.3. The interquartile range of engagement survey results for terminated employees is around 1.1 and for active employees is around 0.9.

There is evidence of outliers for Active employees' engagement survey results being low, around 1.1 to around 2.3, which means despite scoring the a low engagement result these employees are still employed at the orgnisation. Another thing to point out is that employees who are no longer at the organisation or terminated have a minimum engagament result of 2, a higher minimum than the current employees, and the there is not much difference in the average and median results between the terminated and active employees. We would have first assumed that higher

```
terminated <- hr_data[hr_data$term_id == "Terminated",]
summary(terminated$engagement_survey)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.585   4.220   4.090   4.715   5.000
```

```
IQR(terminated$engagement_survey)
```

```
## [1] 1.13
```

```
active <- hr_data[hr_data$term_id == "Active",]
summary(active$engagement_survey)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.12    3.73    4.29    4.12    4.67    5.00
```

```
IQR(active$engagement_survey)
```

```
## [1] 0.94
```

iii. Plot 3: Two factors

```
# manager_name vs. performance_score
manager_vs_performance <- hr_data %>%
  group_by(manager_name, performance_score) %>%
  summarise(total_count = n())

manager_vs_performance %>%
  ggplot(aes(manager_name, total_count, fill=performance_score)) +
  geom_col(position=position_dodge2(preserve = "single")) +
  theme(axis.text.x=element_text(angle=90)) +
      labs(x="Managers",
       y="Number of employees",
       title="Bar graph of manager and employee performance",
       subtitle="Employees performance scores given by each manager",
       fill = "Performance Score")
```

Analysis: Board of directors manages just 2 employees who and given both “Fully Meets” in their performance scores.

David Stanley manages 21 employees, 20 of whom were given “Fully Meets” or “Exceeds” their in performance with just 1 on a PIP. Suggests that around 95% of David Stanley’s direct reports are performing well under his management or he is generous with his feedback.

Kelley Spirea, Elijiah Gray and Ketsia Liebig also has around 95% of direct reports who “Fully Meets” or “Exceeds” in their performance.

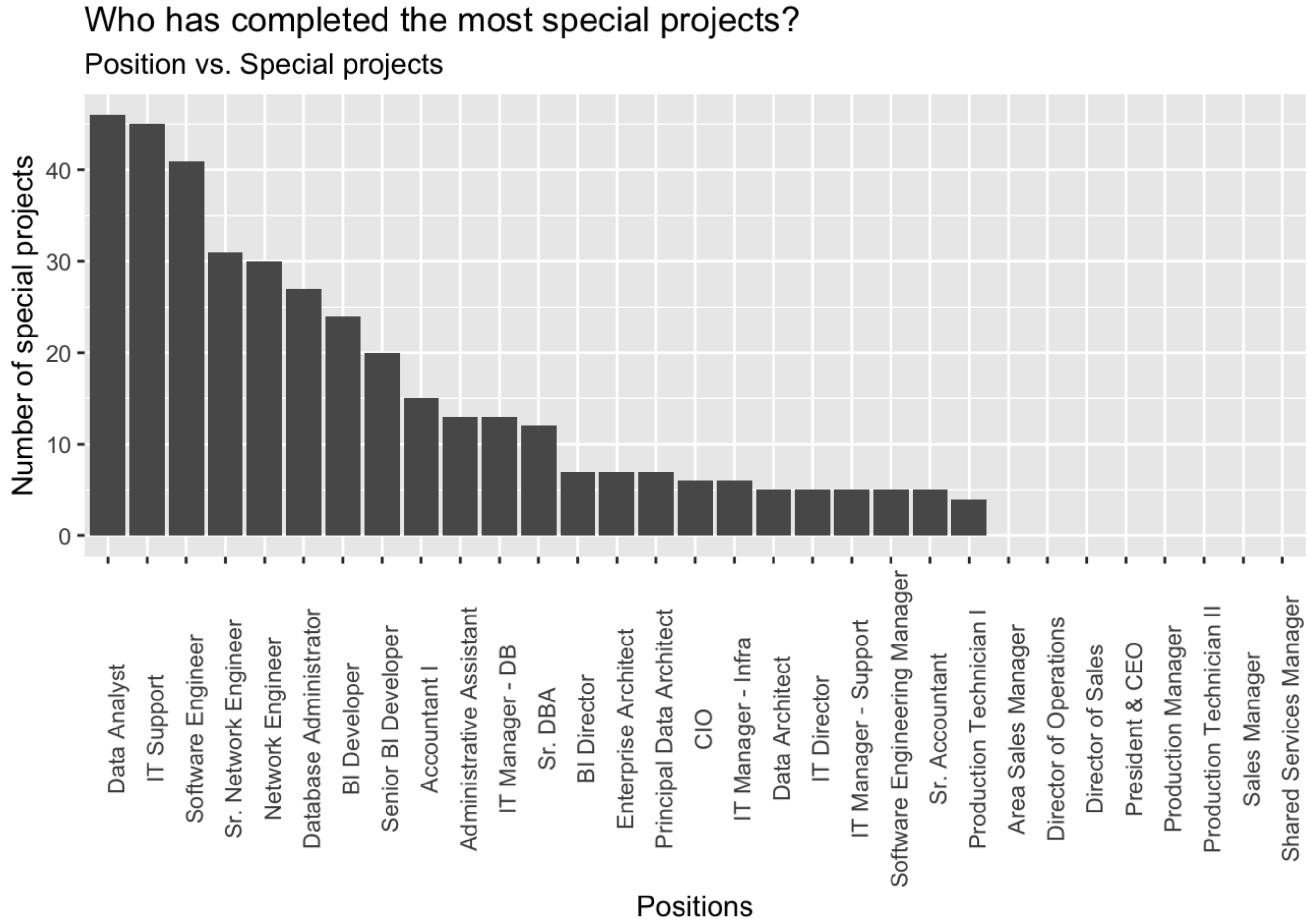
Whilst 17 or 77% of Brannon Miller’s direct reports “Fully Meets” of “Exceeds” in their performance, 5 or around 23% are either on a “PIP” or “Needs Improvement”. Suggesting that many employees under his management are either under-performing or he is stricter in giving performance scores/feedback.

iv. Plot 4: A count variable and a factor

```
projects_vs_position <- hr_data %>%
  group_by(position) %>%
  summarise(total_projects = sum(special_projects_count))

projects_vs_position <- projects_vs_position %>%
  arrange(desc(total_projects))

# special_projects_count vs. position
# reorganise bars
projects_vs_position %>%
  ggplot(aes(reorder(position, -total_projects), total_projects)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x=element_text(angle=90)) +
  labs(x="Positions",
       y="Number of special projects",
       title="Who has completed the most special projects?",
       subtitle="Position vs. Special projects")
```

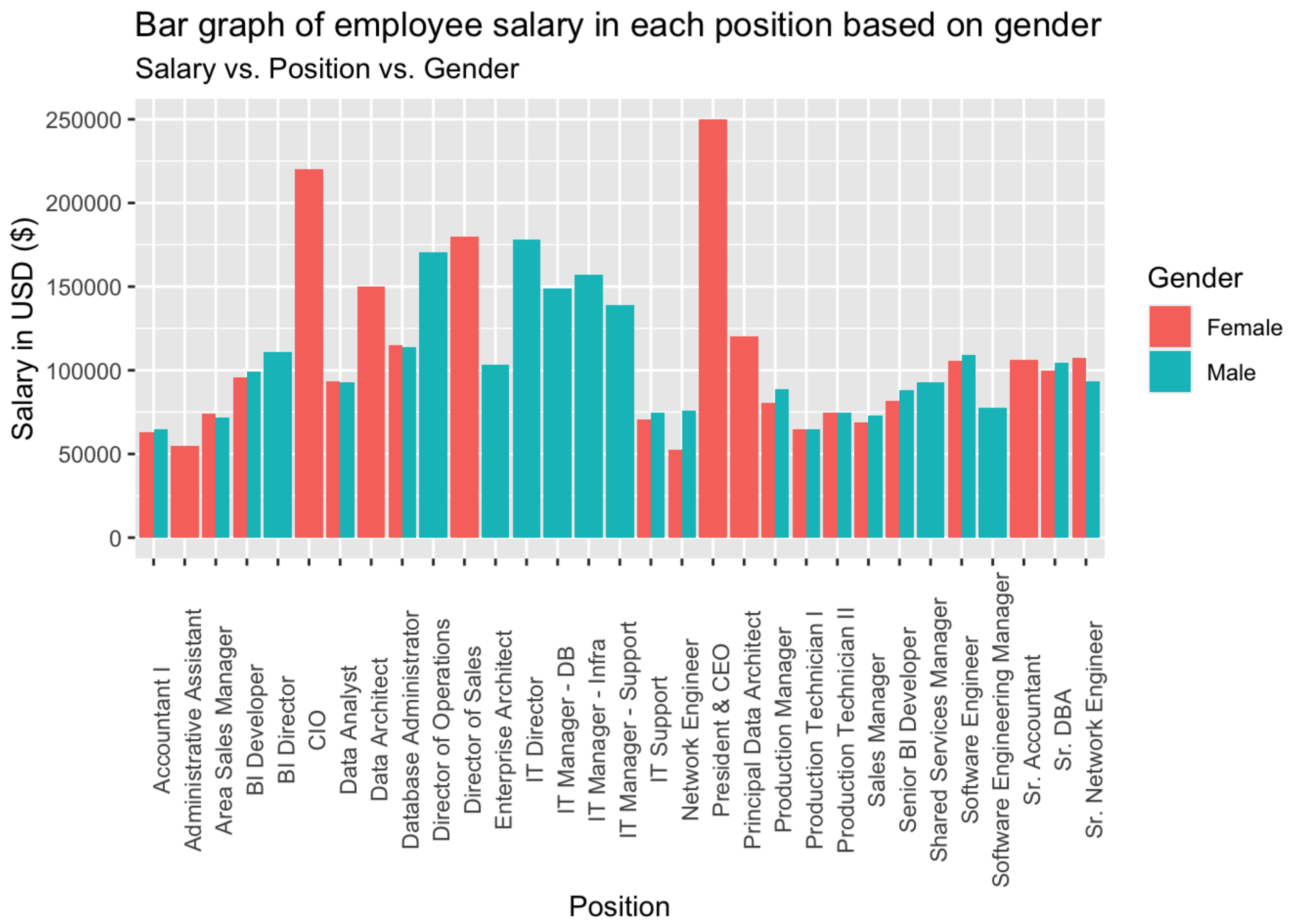


Analysis: Data Analysts, IT Support and Software Engineers are the top 3 positions who have completed the most number of special projects with over 40 projects per each position. However, 8 positions within the company have not completed any special projects, this could be due to the nature and responsibilities of the roles not having special projects especial in senior or managerial roles.

v. Plot 5: Three different variables (of any type).

```
salary_vs_position_vs_sex <- hr_data %>%
  group_by(position, sex) %>%
  summarise(salary)

salary_vs_position_vs_sex %>%
  ggplot(aes(position, salary, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme(axis.text.x=element_text(angle=90)) +
  labs(x="Position",
       y="Salary in USD ($)",
       title="Bar graph of employee salary in each position based on gender",
       subtitle="Salary vs. Position vs. Gender",
       fill = "Gender")
```



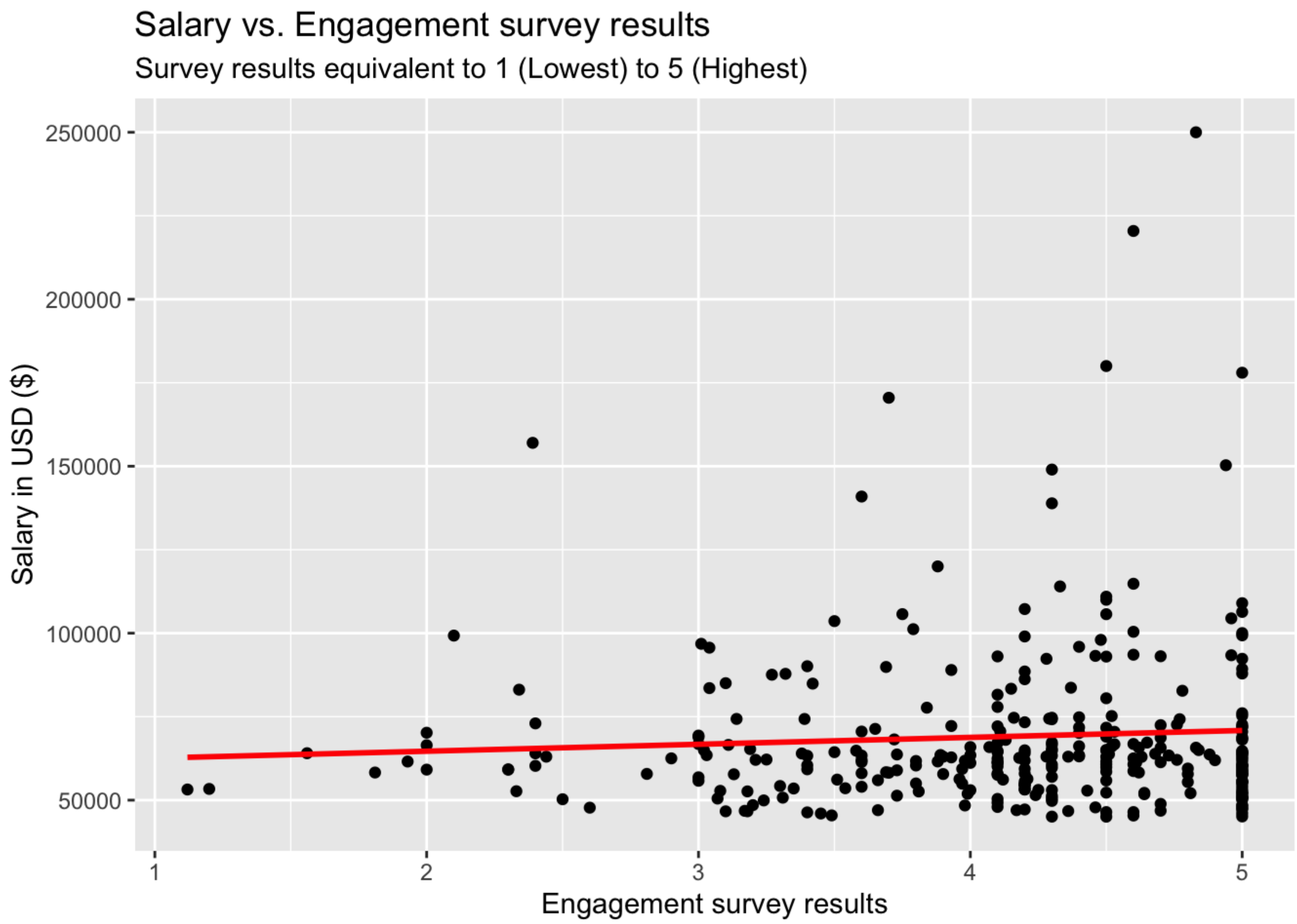
Analysis: This graph shows that the highest earner is the female President & CEO at 250,000 USD, followed by the female CIO who earns 220,450 USD. Overall female employees' collective earnings is higher than those of men but that could be due to the fact that there are more female employees than male employees in this data set.

Fit, evaluate, interpret, and compare two linear models.

Both models should focus on the same response variable and include at least two predictors. [10 points]

```
# linearmodel1 - salary ~ engagement_survey + emp_satisfaction

# examine the relationship of the covariates and the response variable
hr_data %>%
  ggplot(aes(engagement_survey, salary)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, colour = "red") +
  labs(x = "Engagement survey results",
       y = "Salary in USD ($)",
       title = "Salary vs. Engagement survey results",
       subtitle = "Survey results equivalent to 1 (Lowest) to 5 (Highest)")
```



There seems to be a linear relationship between engagement survey results and salary.

The model we will fit is in the form:

$$
\begin{aligned}
\text{mean salary} \sim N(b_0 + b_1 \times engagement\_survey \\
+b_2 \times emp\_satisfaction2 \\
+b_3 \times emp\_satisfaction3 \\
+b_4 \times emp\_satisfaction4 \\
+b_5 \times emp\_satisfaction5, \sigma)
\end{aligned}
$$

```
# Fitting the model salary ~ engagement_survey, emp_satisfaction
linearmodel1 <- lm(salary ~ engagement_survey + emp_satisfaction, data=hr_data)
summary(linearmodel1)
```

```
##
## Call:
## lm(formula = salary ~ engagement_survey + emp_satisfaction, data = hr_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27842 -12933  -6369   1792 178808
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          53348      18488   2.885  0.00419 **
## engagement_survey     1770       1926   0.919  0.35872
## emp_satisfaction2     2502      19647   0.127  0.89875
## emp_satisfaction3     9294      18143   0.512  0.60883
## emp_satisfaction4     4798      18224   0.263  0.79253
## emp_satisfaction5    11574      18184   0.636  0.52493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25130 on 305 degrees of freedom
## Multiple R-squared:  0.01814,    Adjusted R-squared:  0.002043
## F-statistic: 1.127 on 5 and 305 DF,  p-value: 0.346
```

- The base level for “emp_satisfaction” is “emp_satisfaction1”
- For each additional engagement survey result, the salary increases on average by $1770.
- For each additional employee satisfaction score of 2, the salary increases on average by $2502
- For each additional employee satisfaction score of 3, the salary increases on average by $9294
- For each additional employee satisfaction score of 4, the salary increases on average by $4798
- For each additional employee satisfaction score of 5, the salary increases on average by $11574
- The Multiple R-Squared for linearmodel1 is 0.01814 or 1.81% of the variance in salary is explained by the differences in engagement_survey, emp_satisfaction2, emp_satisfaction3, emp_satisfaction4 and emp_satisfaction5. As the percentage is so close to zero, this suggests that the linearmodel1 is not very explanatory of the variation in salary around its mean.

If we were to predict the average salary of an employee who had an engagement result of 1 and employee satisfaction score of 1:

```
# Predicting with the linearmodel1
predict(linearmodel1, data.frame(engagement_survey = 1, emp_satisfaction ="1"))
```

```
##        1
## 55118.22
```

```
# Confidence interval
predict(linearmodel1, data.frame(engagement_survey = 1, emp_satisfaction ="1"), inter
val = "confidence")
```

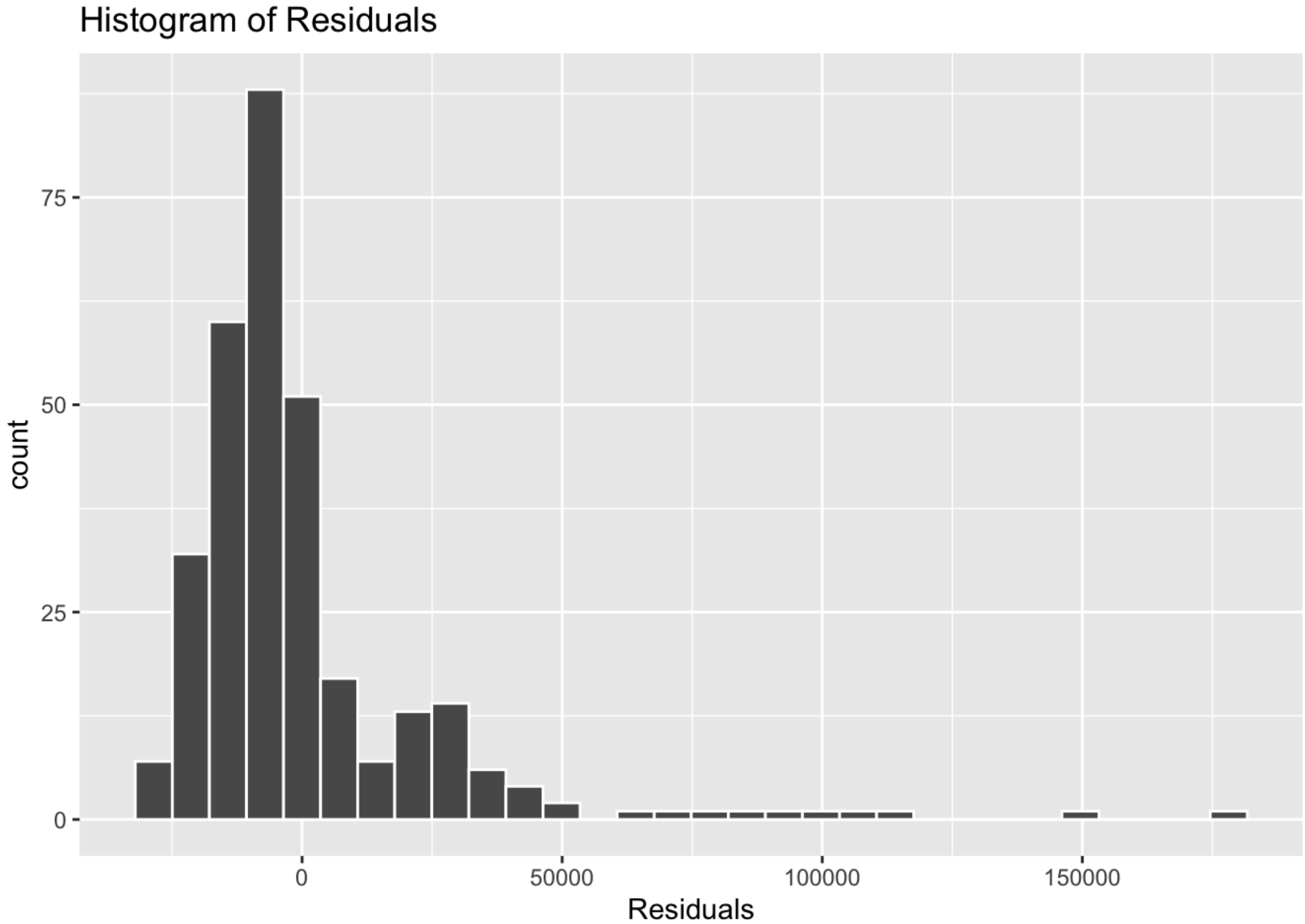
```
##        fit      lwr      upr
## 1 55118.22 19595.79 90640.65
```

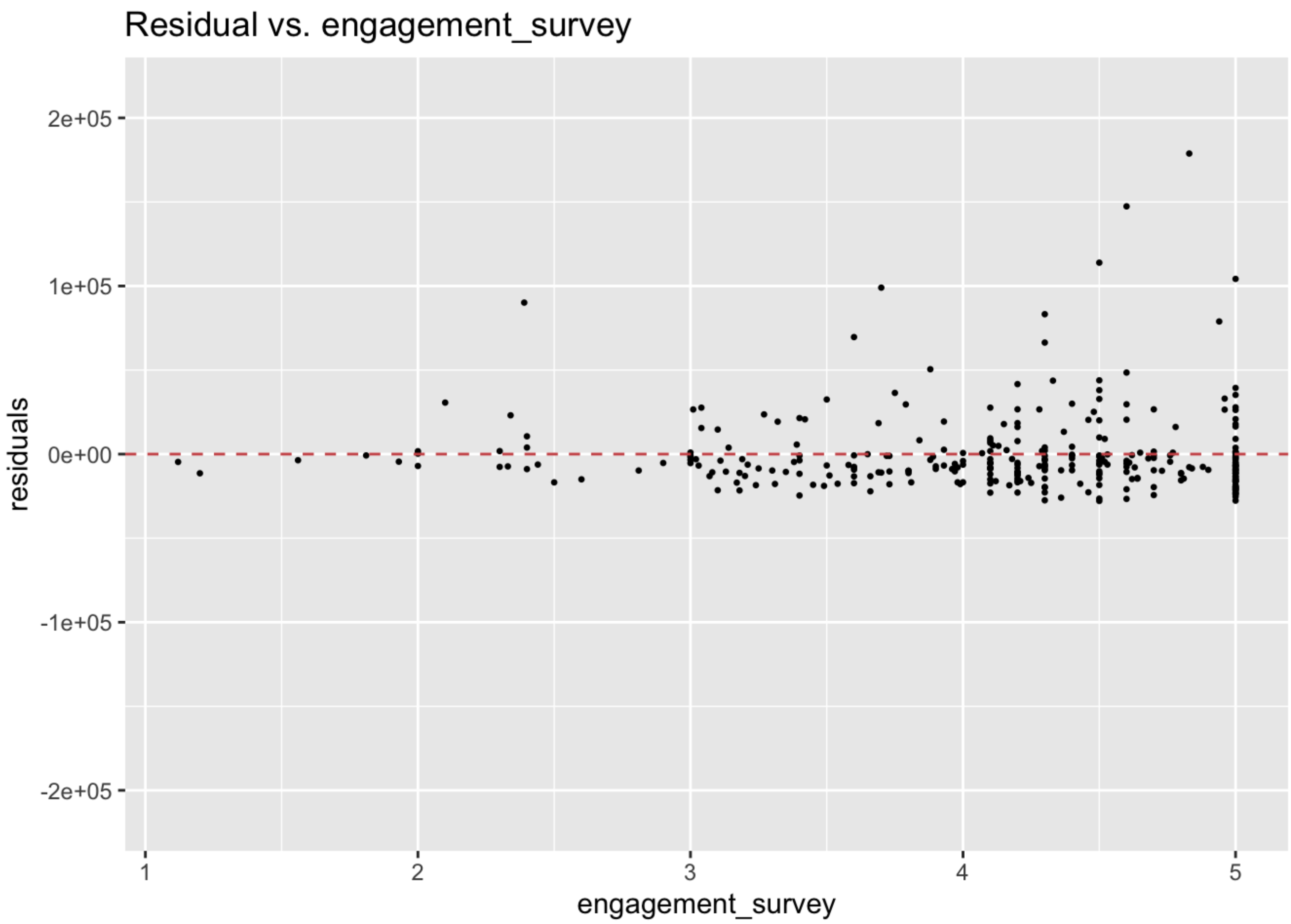
```
# Prediction interval
predict(linearmodel1, data.frame(engagement_survey = 1, emp_satisfaction ="1"), inter
val = "prediction")
```

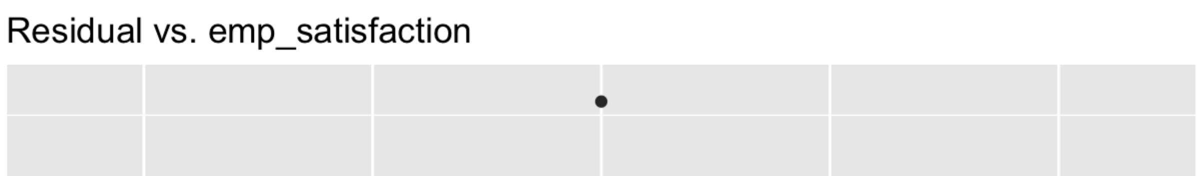
```
##        fit       lwr      upr
## 1 55118.22 -5769.687 116006.1
```

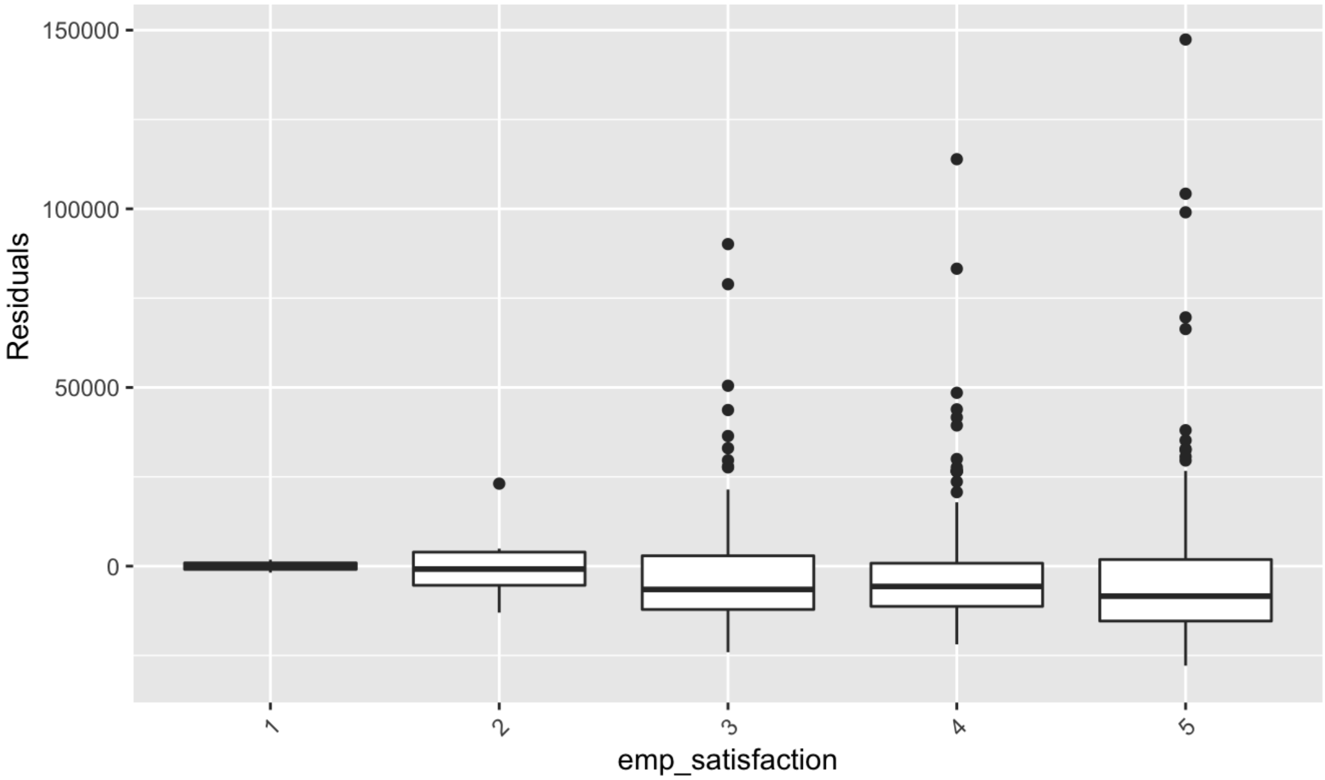
With this model we can expect the employee to earn an average salary of $55118.22 with these characteristics. The prediction interval suggests an average salary between -$5769.69 and $116006.10. However, the confidence interval does suggests that the average salary can be between $19595.79 and $90640.65 with these characteristics.

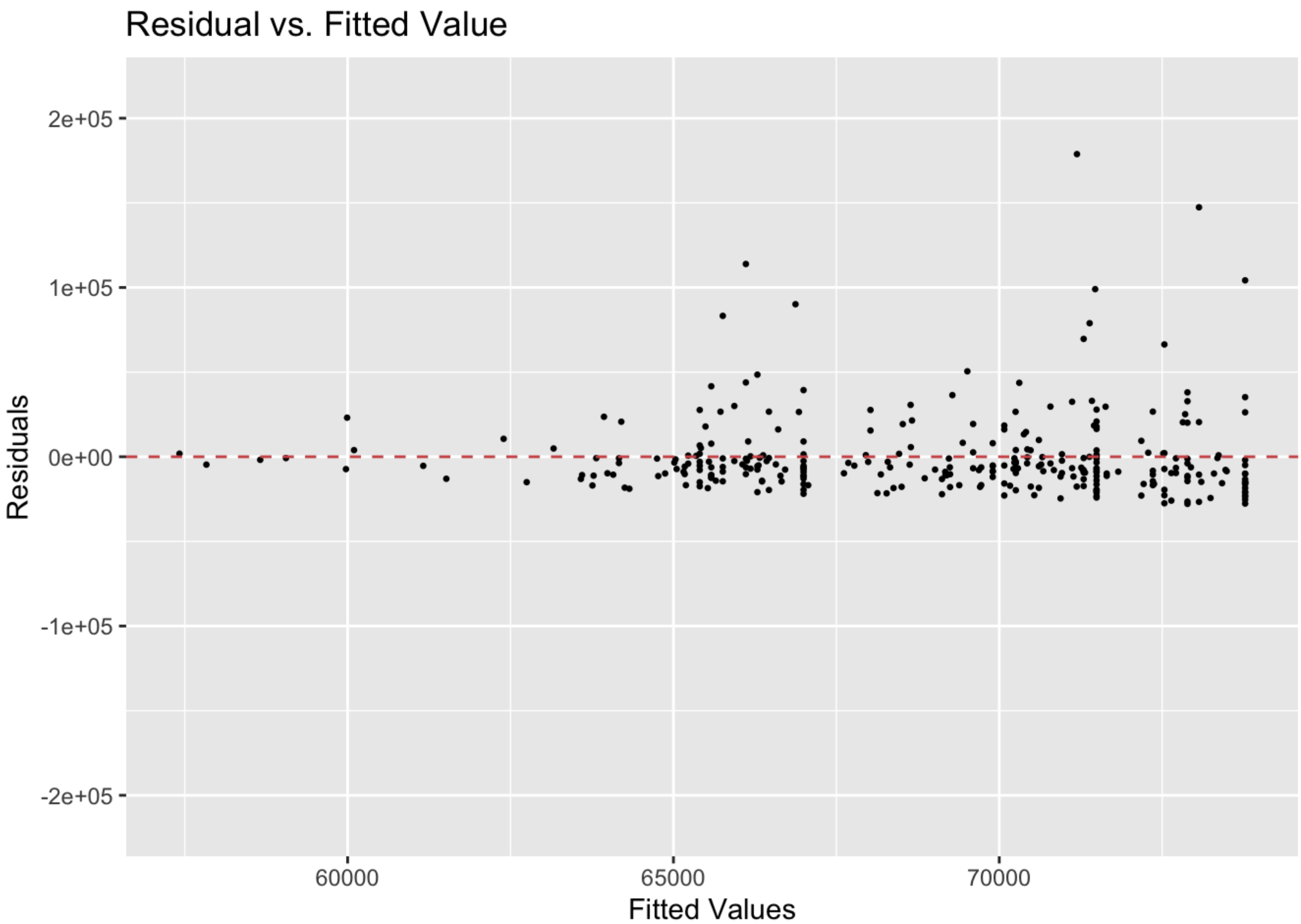
```
# Examine the diagnostic plots of the model
linearmodel1 %>%
  gg_diagnose(max.per.page = 1)
```

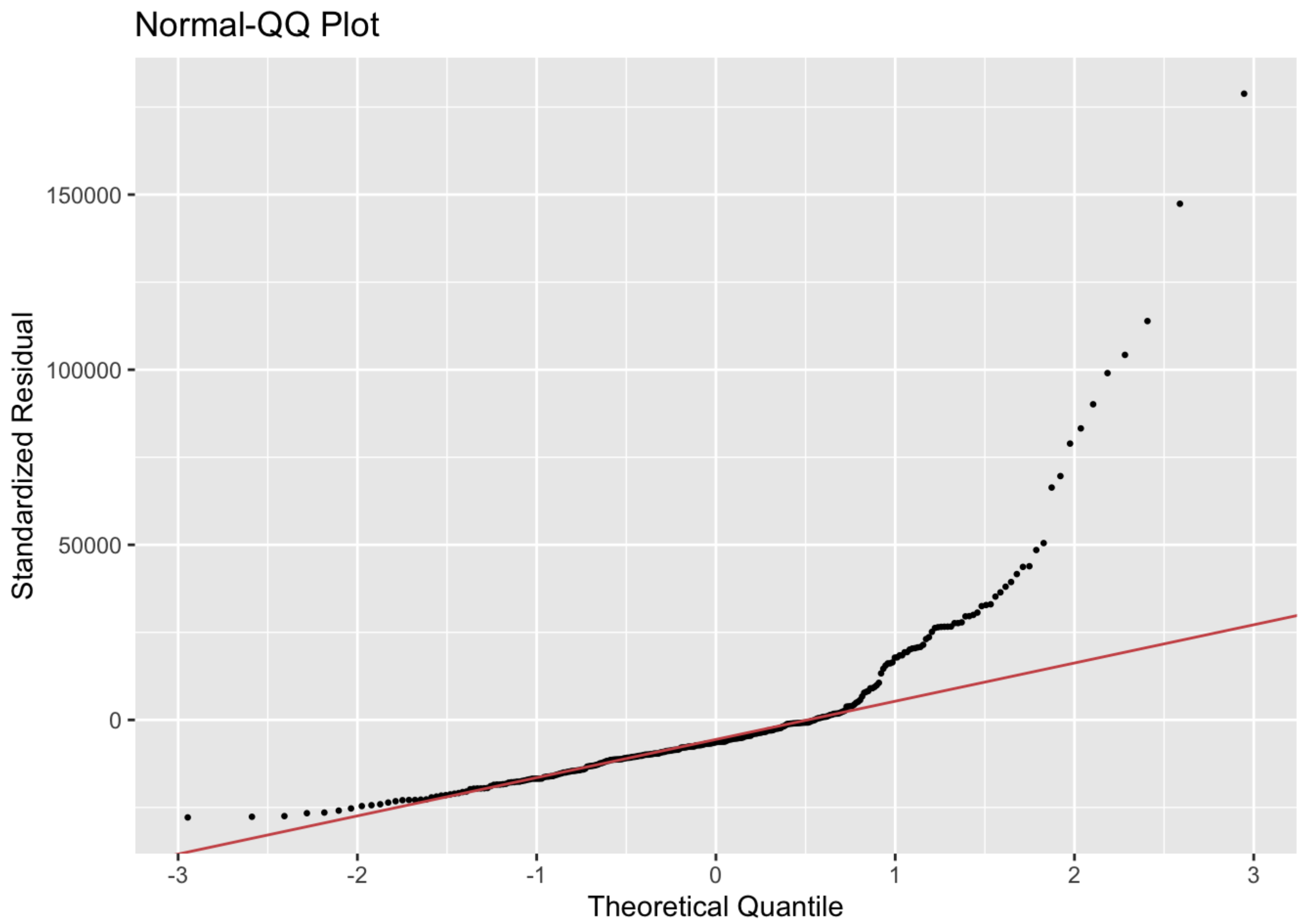

Histogram of Residuals
count
75
50
25
0
0
50000
100000
150000
Residuals

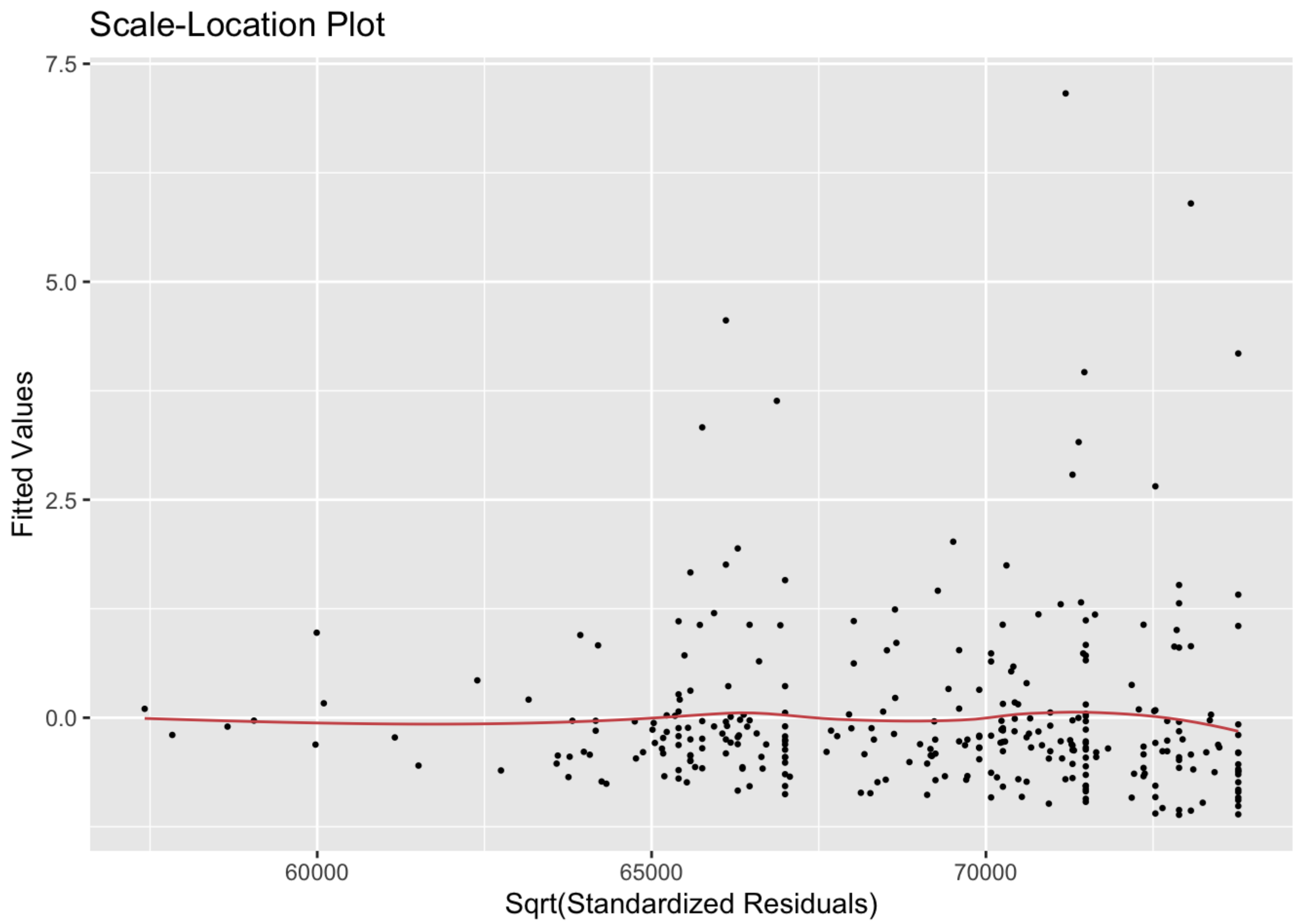

Residual vs. engagement_survey
residuals
2e+05
1e+05
0e+00
-1e+05
-2e+05
1
2
3
4
5
engagement_survey


Residual vs. emp_satisfaction

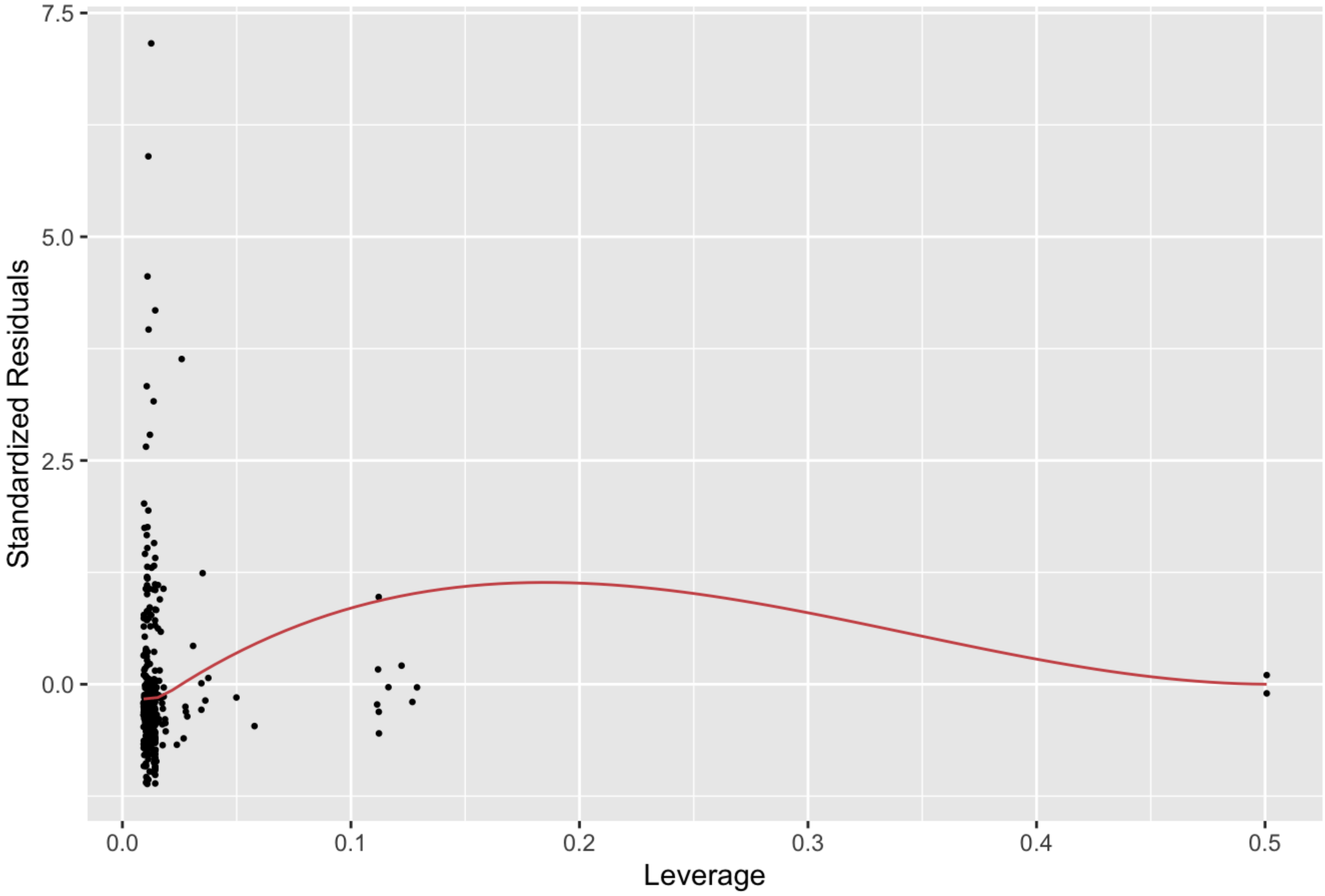

150000
100000
50000
0
Residuals
1
2
3
4
5
emp_satisfaction


Residual vs. Fitted Value
2e+05
1e+05
0e+00
-1e+05
-2e+05
Residuals
60000
65000
70000
Fitted Values

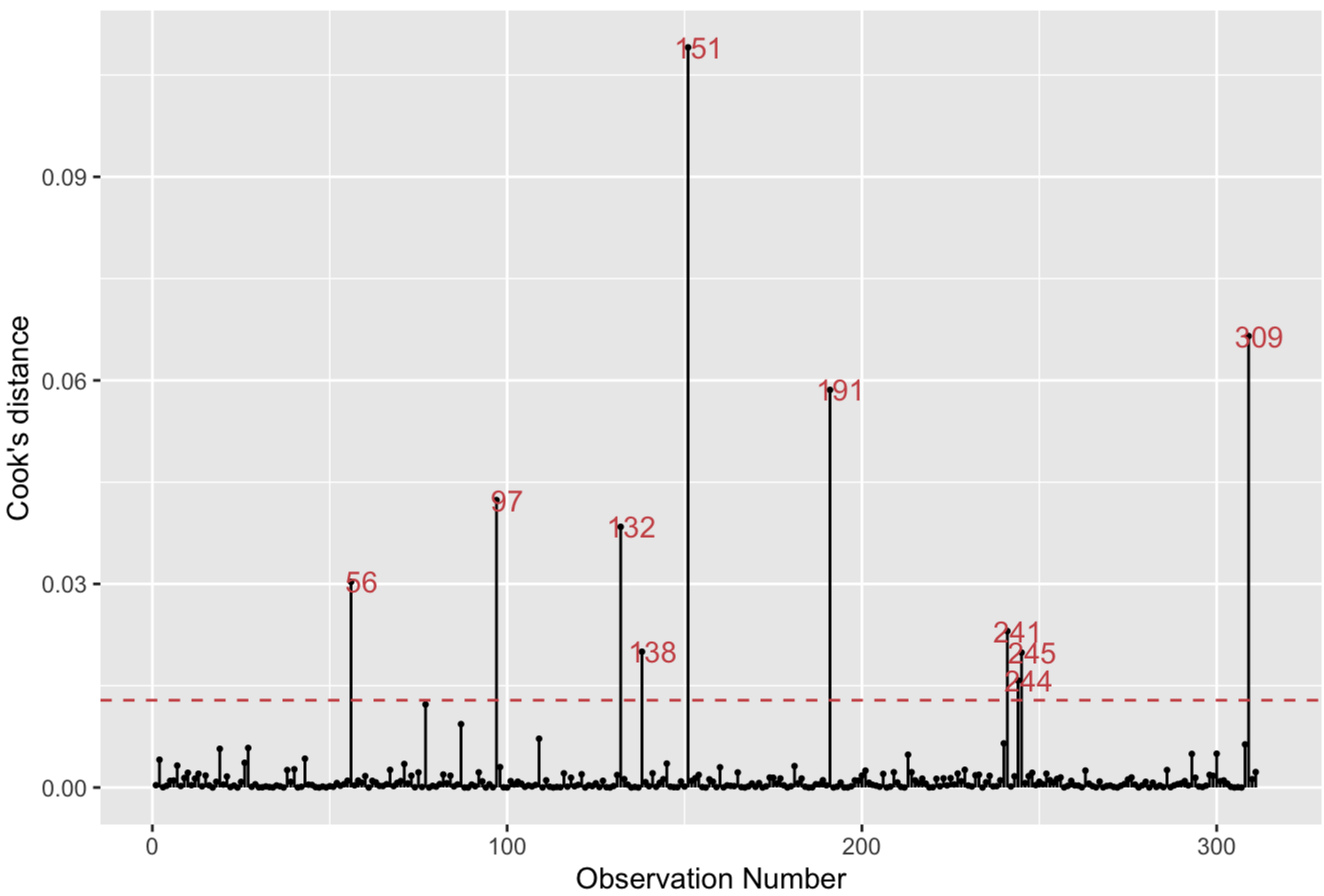

Normal-QQ Plot
Standardized Residual
150000
100000
50000
0
-3
-2
-1
0
1
2
3
Theoretical Quantile


Scale-Location Plot
Fitted Values
7.5
5.0
2.5
0.0
60000
65000
70000
Sqrt(Standardized Residuals)

Residual vs. Leverage



Cook's Distance Plot



$$
\begin{aligned}
\text{mean salary} \sim N(53348 + 1770 \times engagement\_survey \\
+2502 \times emp\_satisfaction2 \\
+9294 \times emp\_satisfaction3 \\
+4798 \times emp\_satisfaction4 \\
+11574 \times emp\_satisfaction5, 25130)
\end{aligned}
$$

Evaluating the assumptions:

Linearity: The scatterplot of the salary versus engagement survey results is roughly linear.

Homoscedasticity: The scatter of residuals versus engagement survey results is roughly the same width, although there seems to be a slight increase in spread as the engagement survey result increases.

Normality: The histogram of residuals have a roughly Gaussian distribution.

Normality: The QQ-plot of residuals is roughly linear and skewed right with an increase in spread toward the right.

Independence: As we do not know the order which the data was collected in we cannot evaluate the assumption of independence.

The second model, linearmodel2 will fit in the form:

$$
\begin{aligned}
\text{mean salary} \sim N(b_0 + b_1 \times engagement\_survey \\
+b_2 \times emp\_satisfaction2 \\
+b_3 \times emp\_satisfaction3 \\
+b_4 \times emp\_satisfaction4 \\
+b_5 \times emp\_satisfaction5 \\
+b_6 \times Executive\ Office \\
+b_7 \times IT/IS \\
+b_8 \times Production \\
+b_9 \times Sales \\
+b_{10} \times SoftwareEngineering, \sigma)
\end{aligned}
$$

```
# Fitting the model salary ~ engagement_survey + emp_satisfaction _ department
linearmodel2 <- lm(salary ~ engagement_survey + emp_satisfaction + department, data=h
r_data)
summary(linearmodel2)
```

```
##
## Call:
## lm(formula = salary ~ engagement_survey + emp_satisfaction +
##     department, data = hr_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50513  -7697  -1348   4618 120277
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       62291      14925   4.174 3.93e-05 ***
## engagement_survey                  1296       1402   0.925 0.355761
## emp_satisfaction2                  4690      14207   0.330 0.741539
## emp_satisfaction3                  3935      13114   0.300 0.764361
## emp_satisfaction4                  2636      13167   0.200 0.841488
## emp_satisfaction5                  7217      13126   0.550 0.582840
## departmentExecutive Office       177513      19126   9.281  < 2e-16 ***
## departmentIT/IS                   24702       6623   3.730 0.000229 ***
## departmentProduction             -12173       6206  -1.962 0.050727 .
## departmentSales                   -3201       6973  -0.459 0.646558
## departmentSoftware Engineering    22243       8218   2.706 0.007191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18070 on 300 degrees of freedom
## Multiple R-squared:  0.5005, Adjusted R-squared:  0.4838
## F-statistic: 30.06 on 10 and 300 DF,  p-value: < 2.2e-16
```

- The base level for “emp_satisfaction” is “emp_satisfaction1” and the base level for “department” is “departmentAdmin Office”
- For each additional engagement survey result, the salary on average by $1296
- For each additional employee satisfaction score of 2, the salary increases on average by $4,690
- For each additional employee satisfaction score of 3, the salary increases on average by $3,935
- For each additional employee satisfaction score of 4, the salary increases on average by $2,636
- For each additional employee satisfaction score of 5, the salary increases on average by $7,217
- For each Execeutive Office department, the salary increases on average by $177,513
- For employee from IT/IS department, the salary increases on average by $24,702
- For employee from Production department, the salary decreases on average by $12,173
- For employee from Sales department, the salary decreases on average by $3,201
- For employee from Software Engineering department, the salary increases on average by $22,243
- The Multiple R-Squared for linearmodel2 is 0.5005 or 50.05% of the variance in salary is explained by the differences in engagement_survey, emp_satisfaction2, emp_satisfaction3, emp_satisfaction4, emp_satisfaction5, departmentExecutive Office, departmentIT/IS, departmentProduction, departmentSales and departmentSoftware Engineering. This suggests that lineamodel2 is explanatory.

If we were to predict the average salary of an employee from the Production department who had an engagement result of 1 and employee satisfaction score of 1:

```
# Predicting with the linearmodel2
predict(linearmodel2, data.frame(engagement_survey = 1, emp_satisfaction ="1", depart
ment = "Production"))
```

```
##       1
## 51413.9
```

```
# Confidence interval
predict(linearmodel2, data.frame(engagement_survey = 1, emp_satisfaction ="1", depart
ment = "Production"), interval = "confidence")
```

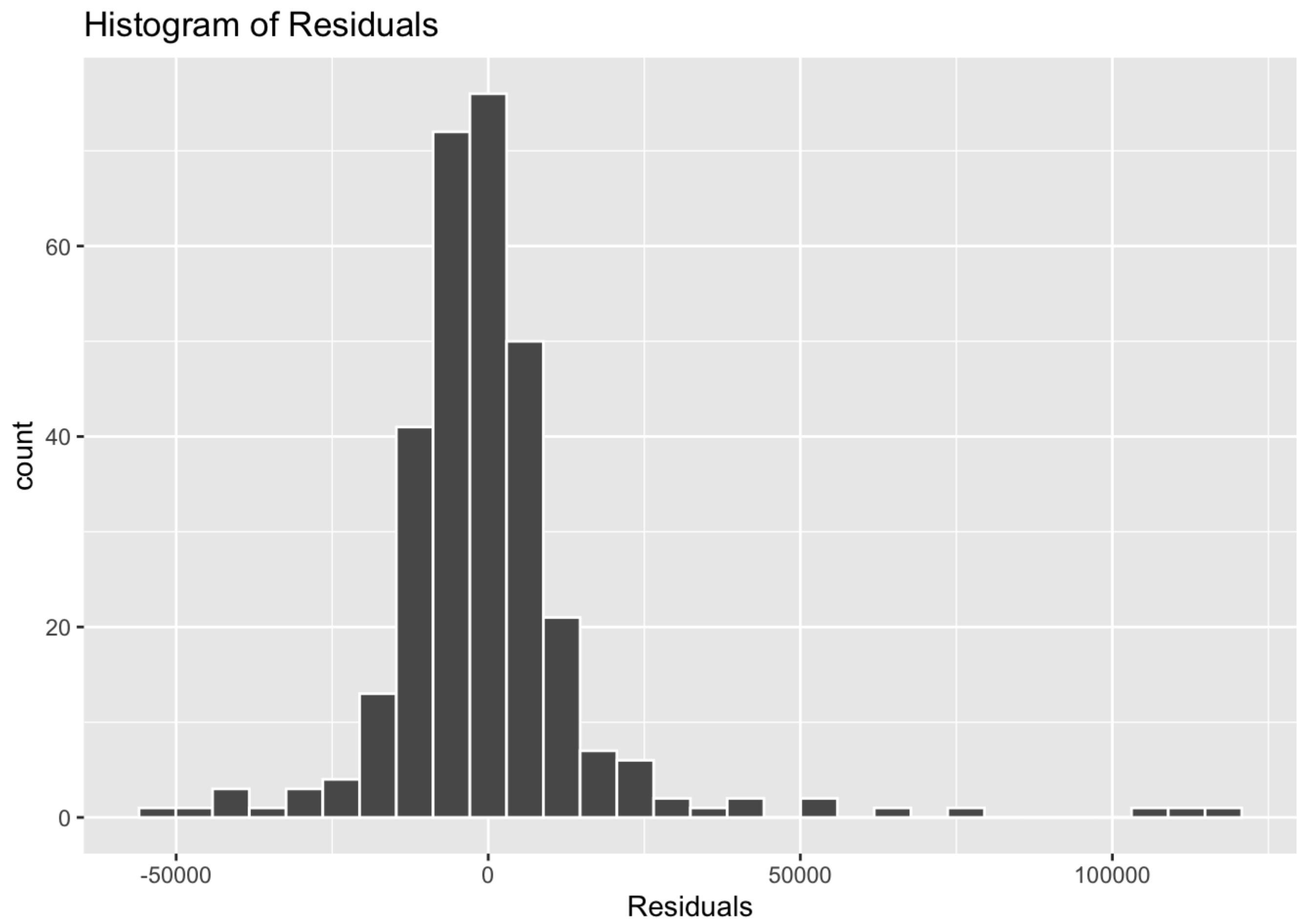
```
##       fit      lwr      upr
## 1 51413.9 25551.81 77275.98
```

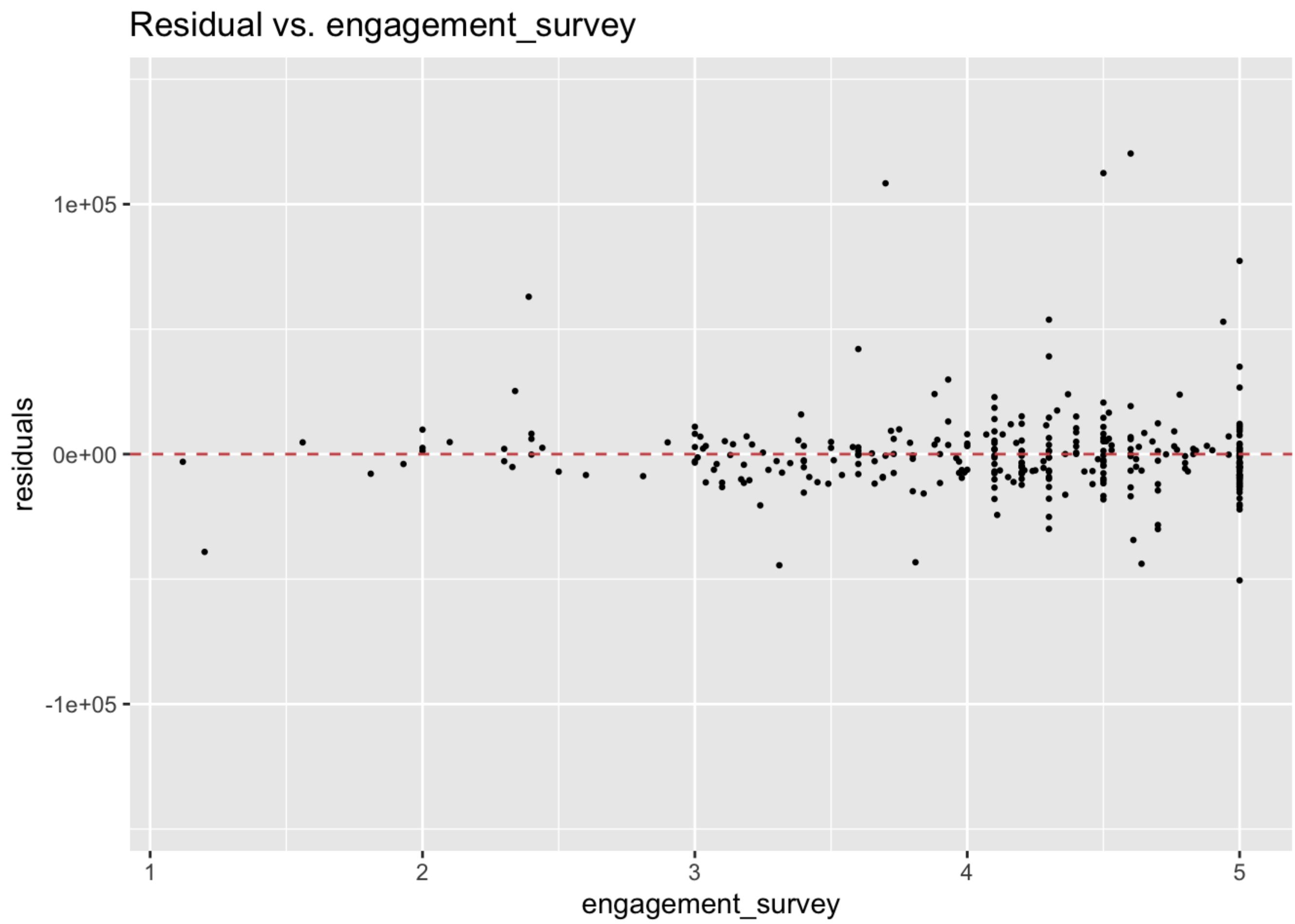
```
# Prediction interval
predict(linearmodel2, data.frame(engagement_survey = 1, emp_satisfaction ="1", depart
ment = "Production"), interval = "prediction")
```

```
##       fit      lwr      upr
## 1 51413.9 7438.014 95389.78
```

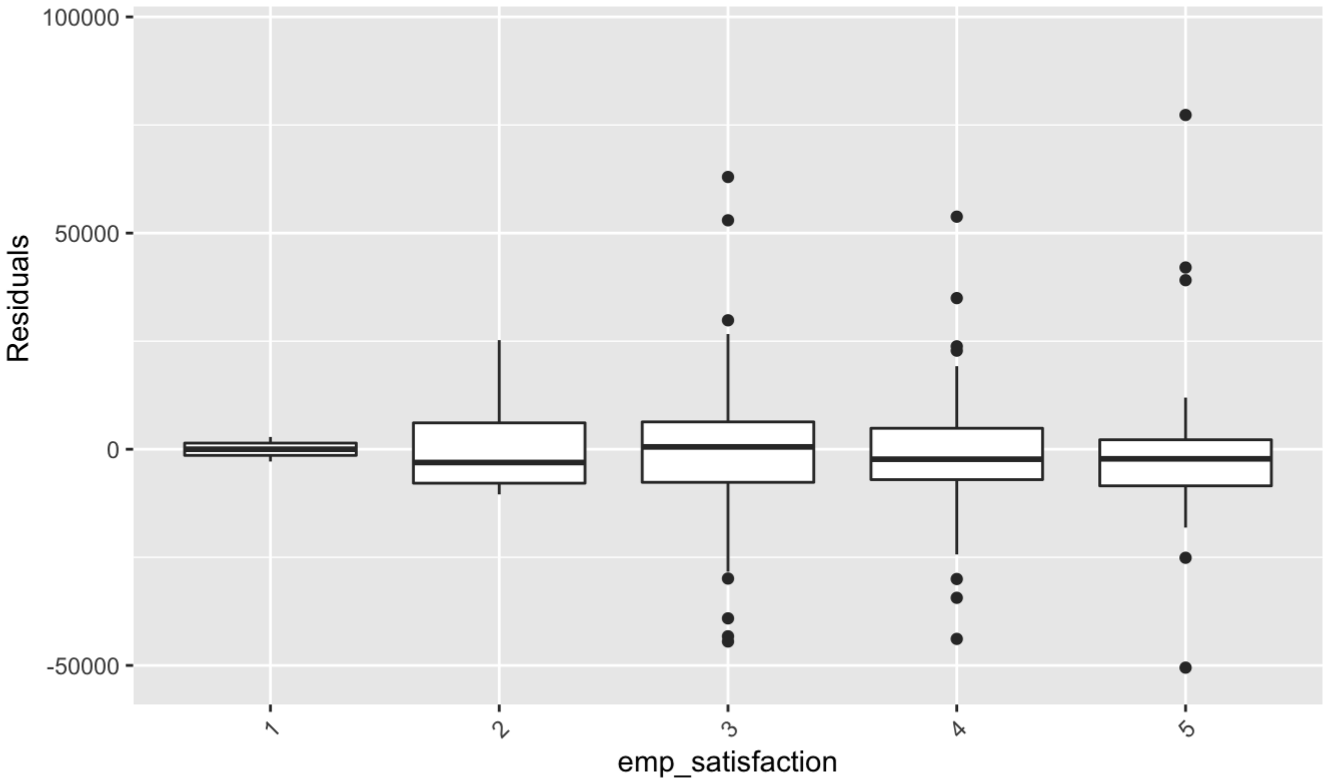
With this model, we can expect the employee to earn an average salary of $51413.90 with these characteristics, with the prediction of interval of between $7438.01 and $95389.78. However, the confidence interval does suggests that the average salary can be between $25551.81 and $77275.98 with these characteristics.

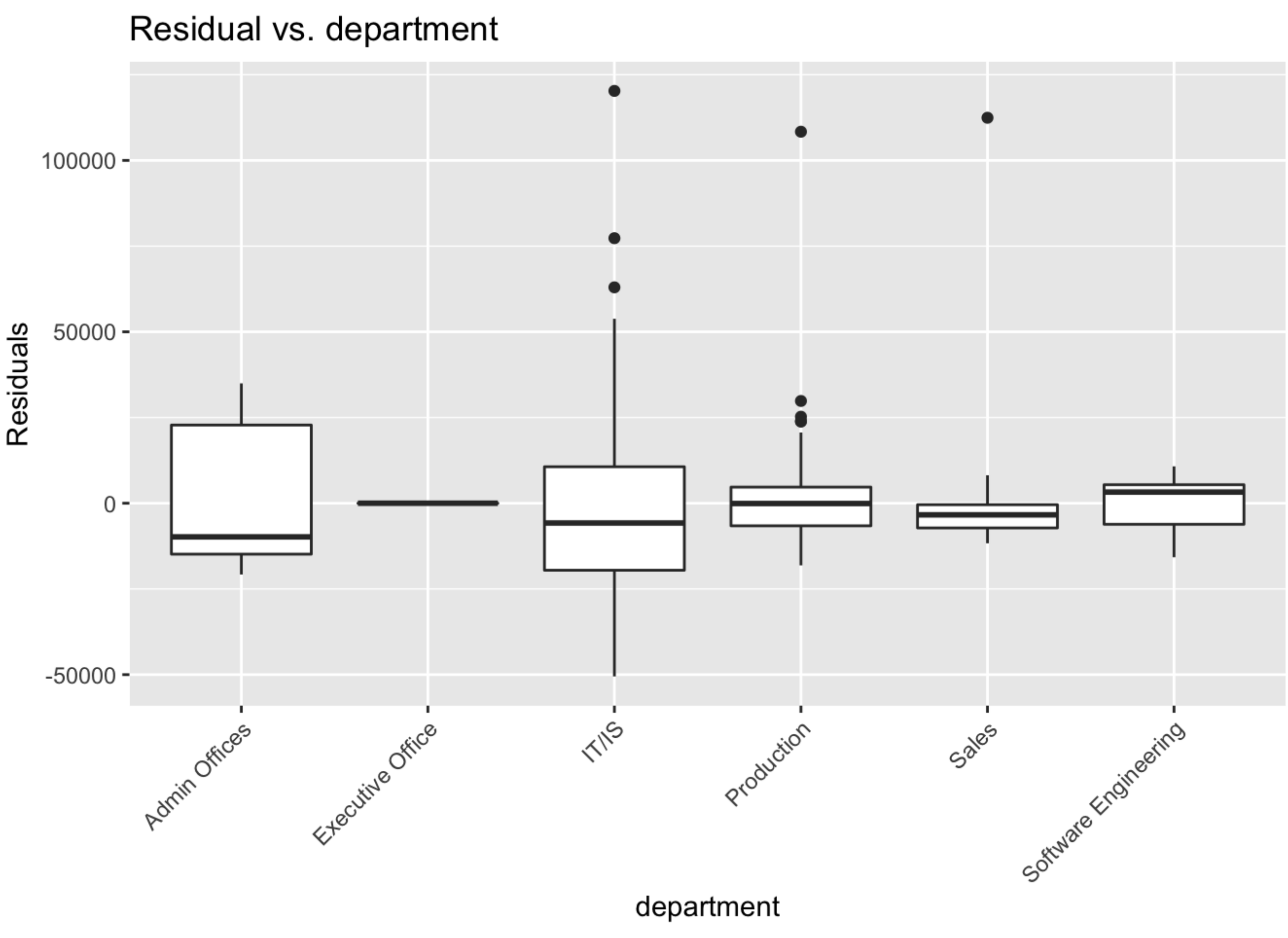
```
# Examine the diagnostic plots of the model
linearmodel2 %>%
  gg_diagnose(max.per.page = 1)
```

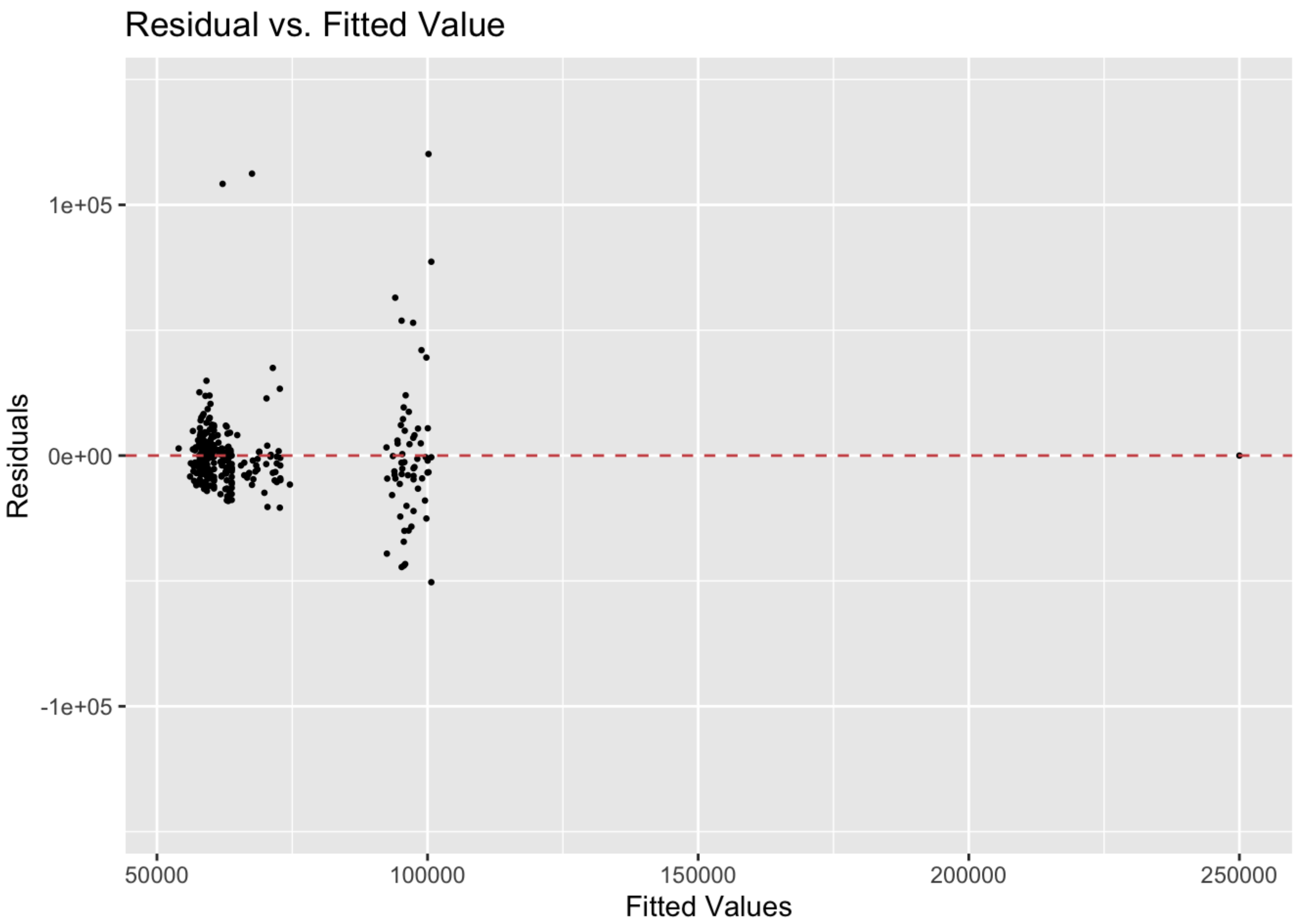

Histogram of Residuals
count
Residuals
60
40
20
0
-50000
0
50000
100000

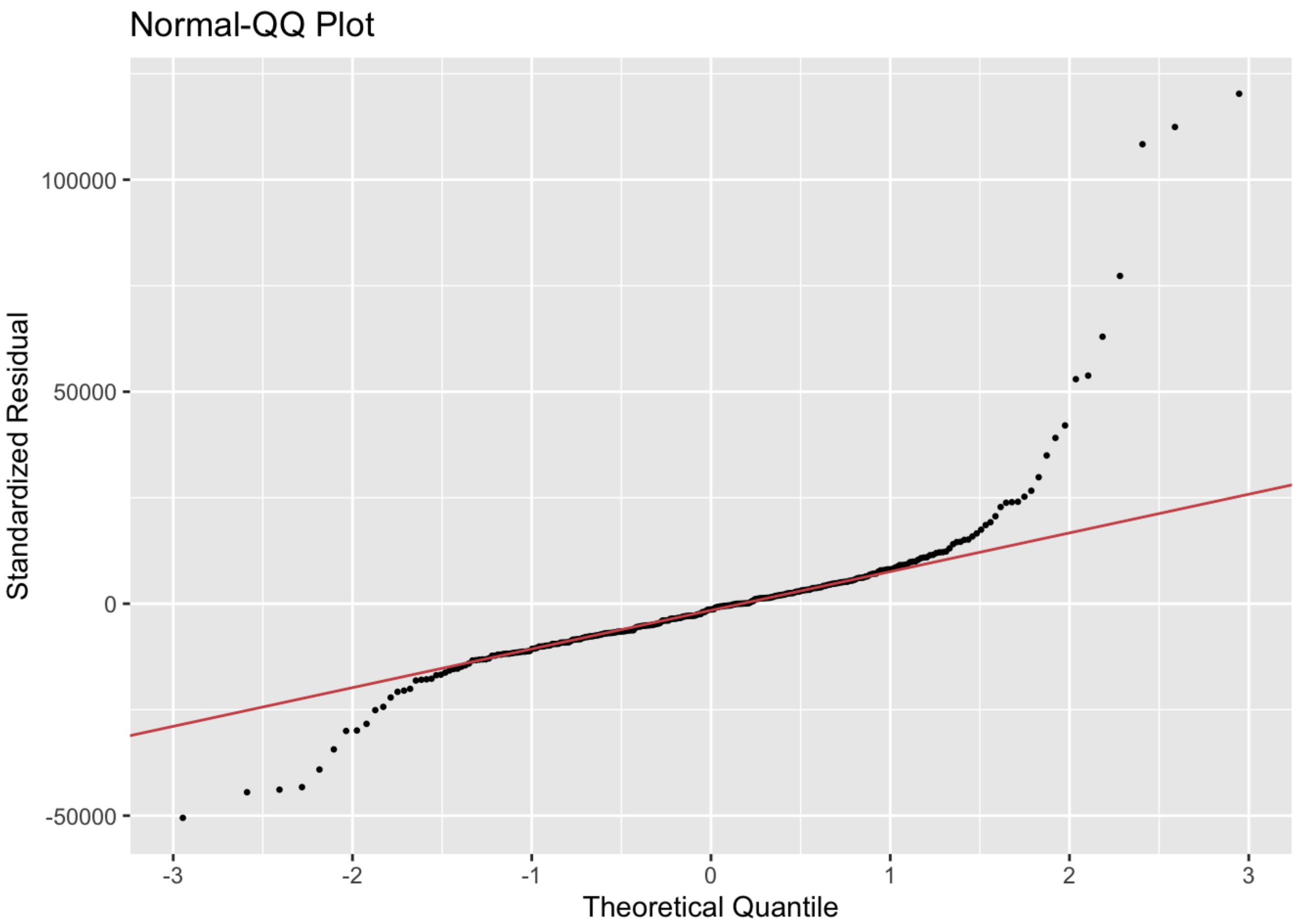

Residual vs. engagement_survey
residuals
engagement_survey
1e+05
0e+00
-1e+05
1
2
3
4
5

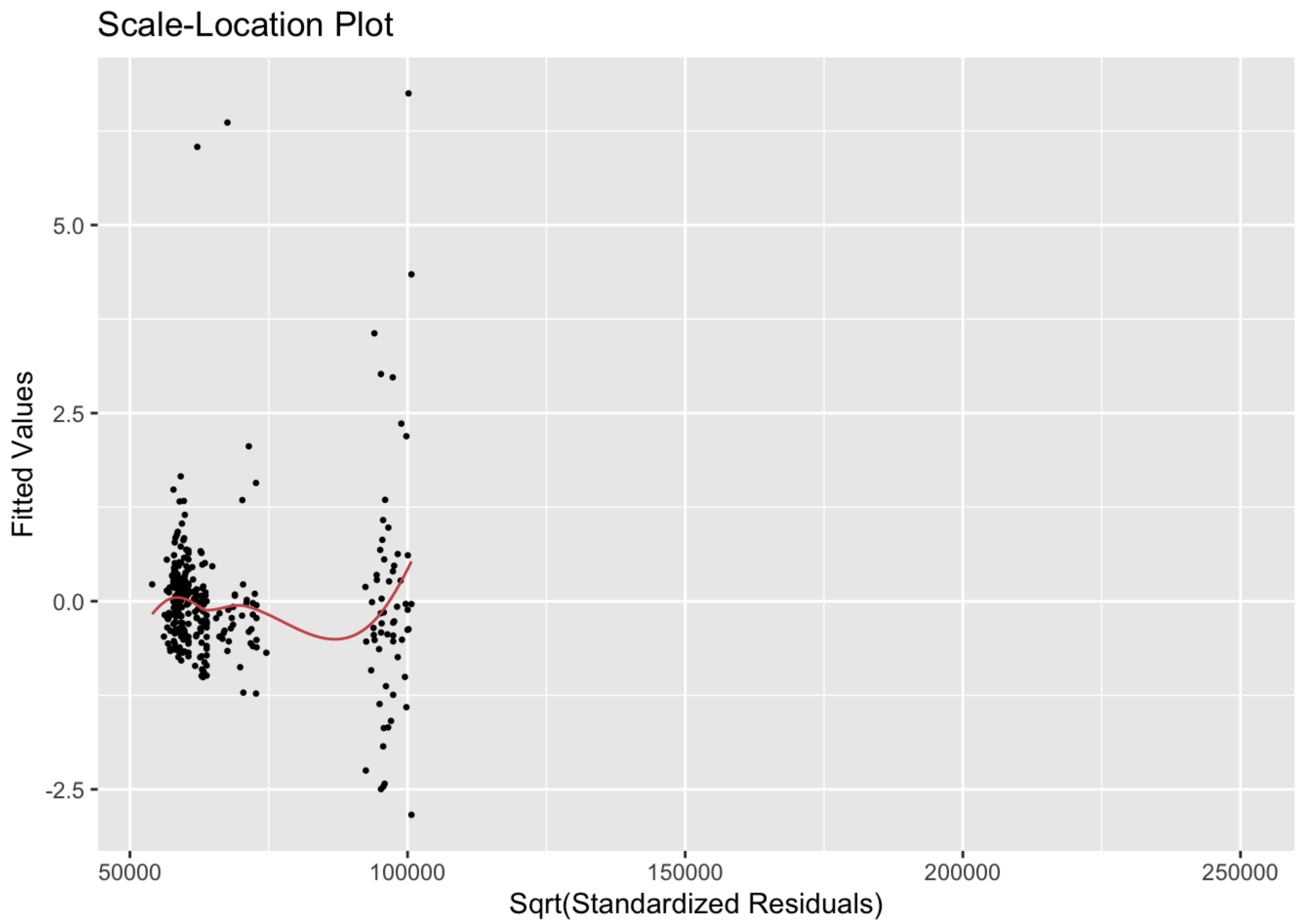

Residual vs. emp_satisfaction


100000
50000
0
-50000
Residuals
1
2
3
4
5
emp_satisfaction

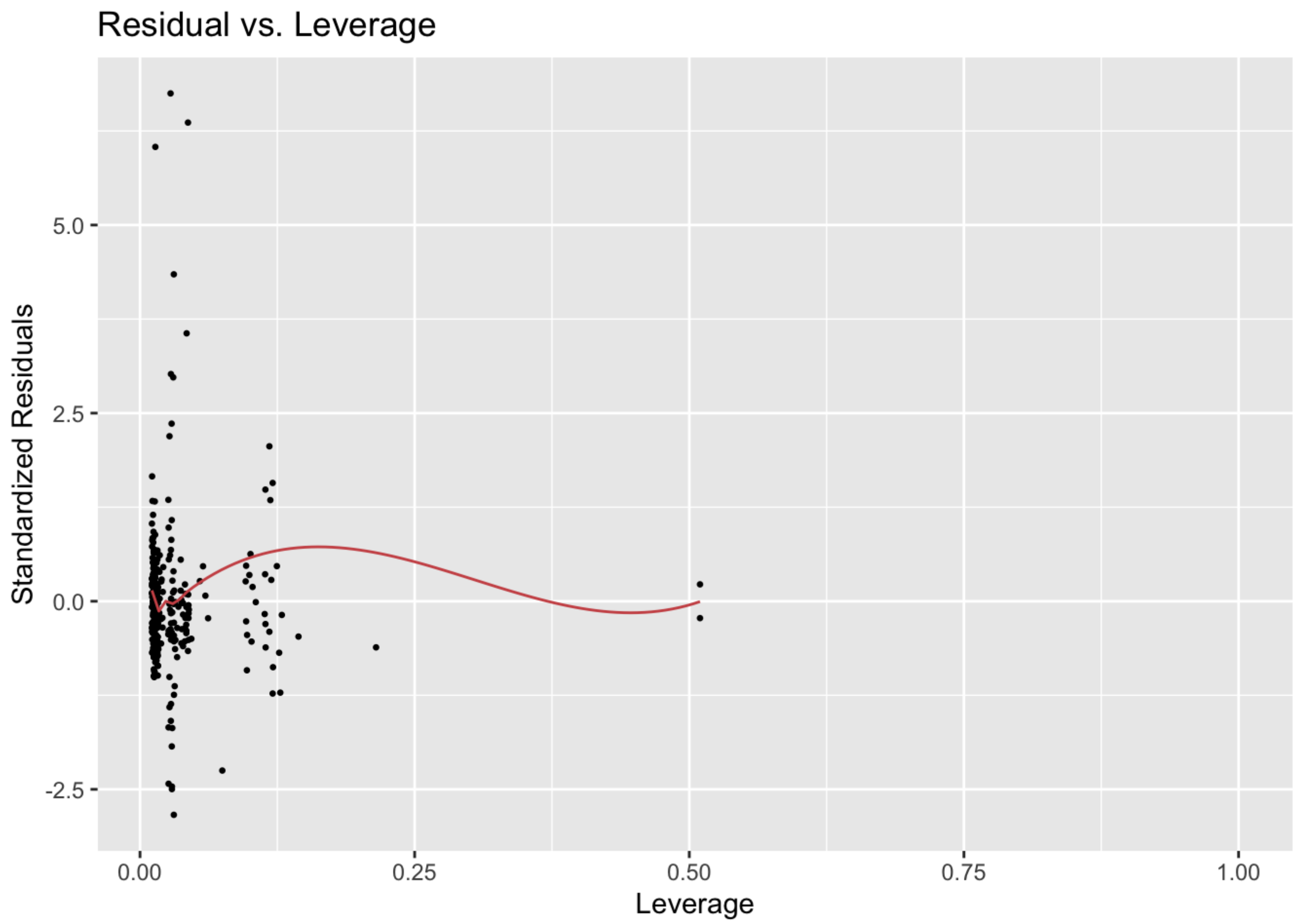

Residual vs. department
100000
50000
0
-50000
Residuals
Admin Offices
Executive Office
IT/IS
Production
Sales
Software Engineering
department


Residual vs. Fitted Value
Residuals
Fitted Values
1e+05
0e+00
-1e+05
50000
100000
150000
200000
250000

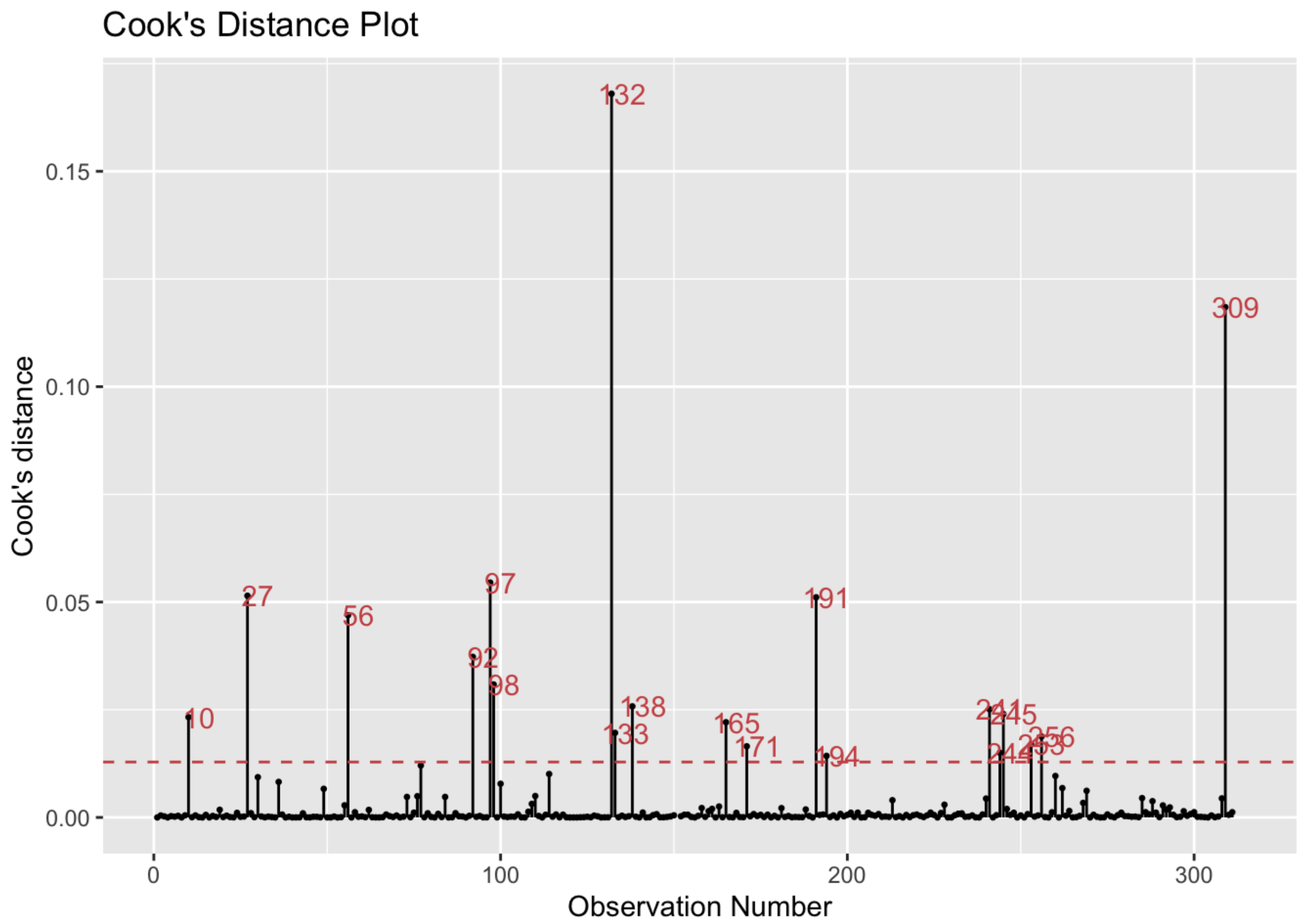

Normal-QQ Plot
Standardized Residual
Theoretical Quantile
100000
50000
0
-50000
-3
-2
-1
0
1
2
3



Scale-Location Plot
Fitted Values
5.0
2.5
0.0
-2.5
50000
100000
150000
200000
250000
Sqrt(Standardized Residuals)



Residual vs. Leverage
Standardized Residuals
5.0
2.5
0.0
-2.5
0.00
0.25
0.50
0.75
1.00
Leverage



Evaluating the assumptions:

Linearity: The scatter plot of the salary versus engagement survey results is roughly linear.

Homoscedasticity: The scatter of residuals versus engagement survey results is roughly the same width.

Normality: The histogram of residuals has a Gaussian distribution.

Normality: The QQ-plot of residuals is roughly linear and skewed right with an increase in spread toward the right.

Independence: As we do not know the order which the data was collected in we cannot evaluate the assumption of independence.

Final model form:

$$\begin{aligned} \text{mean salary} \sim N(62291 + 1296 \times engagement\_survey \\ +4690 \times emp\_satisfaction2 \\ +3935 \times emp\_satisfaction3 \\ +2636 \times emp\_satisfaction4 \\ +7217 \times emp\_satisfaction5 \\ +177513 \times Executive\ Office \\ +24702 \times IT/IS \\ -12173 \times Production \\ -3201 \times Sales \\ +22243 \times SoftwareEngineering, 18070) \end{aligned}$$

```
# Model comparison of the nested models with anova()
anova(linearmodel1, linearmodel2)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ engagement_survey + emp_satisfaction
## Model 2: salary ~ engagement_survey + emp_satisfaction + department
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    305 1.9263e+11
## 2    300 9.7997e+10  5 9.4629e+10 57.938 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing models

- Adjusted R-Squared in linearmodel1 was 0.002043 or 0.20% whilst the Adjusted R-squared in linearmodel2 was 0.4838 or 48.38%, suggests that the fuller linearmodel2 is a lot more explanatory, as the Adjust R-Squared will increase when another predictor variable and its parameters that are added improves the model. It is a better comparison than the Multiple R-Squared where an irrelevant variable added increases the Multiple R-Squared value.
- Model 2 (linearmodel2) has 5 additional parameters with a very small p-value of 0.00000000000000022 (< 0.001), suggests that adding the variable 'department' led to a significantly improved model, than Model 1 (linearmodel1).

Fit, evaluate, interpret, and compare two logistic models.

Both models should focus on the same response variable and include at least two predictors. [10 points]

```
# Fitting the model term_id ~ absences + days_late_last30 + engagement_survey + perfo
rmance_score
logisticmodel1 <-glm(term_id ~ absences + days_late_last30, family = "binomial", data
= hr_data)
summary(logisticmodel1)
```

```
##
## Call:
## glm(formula = term_id ~ absences + days_late_last30, family = "binomial",
##     data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5414  -0.9153  -0.7981   1.3993   1.6776
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.16269    0.25829  -4.502 6.75e-06 ***
## absences          0.03638    0.02095   1.737   0.0825 .
## days_late_last30  0.20992    0.09012   2.329   0.0198 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 396.37  on 310  degrees of freedom
## Residual deviance: 387.88  on 308  degrees of freedom
## AIC: 393.88
##
## Number of Fisher Scoring iterations: 4
```

The model is in the form of:

$$logit(P(term\_id)) = -1.16269 + 0.03638 \times absences + 0.20992 \times days\_late\_last30$$

If we were to predict the probability of the employee who has had 10 days of absences and have been late 4 times in the last 30 days will be terminated:

```
# Example prediction with logisticmodel1
predict(logisticmodel1, newdata = data.frame(absences = 10, days_late_last30 = 4), ty
pe = "response")
```

```
##         1
## 0.5102003
```

We can expect about 51% chance that this employee would be terminated with these characteristics. If to classify this employee, the employee would be terminated, however the probability of the employee being an active employee is about a 49% chance. Please note the term ‘terminated’ includes employees that have been both terminated for a cause and voluntarily terminated.

Evaluating the predictions with confusion matrix, since linearity is not a meaningful concept when performing model analytics for logistic regression as it does not make sense, we will evaluate the predictions with confusion matrix to see how good the predictions are.

```
# Confusion matrix
new_data = hr_data[is.na(hr_data$days_late_last30) == FALSE,]
new_data$pred_term<-ifelse(predict(logisticmodel1, newdata = new_data, type = "respon
se") >= 0.5, "Terminated", "Active")
table(new_data$pred_term, as.factor(new_data$term_id))
```

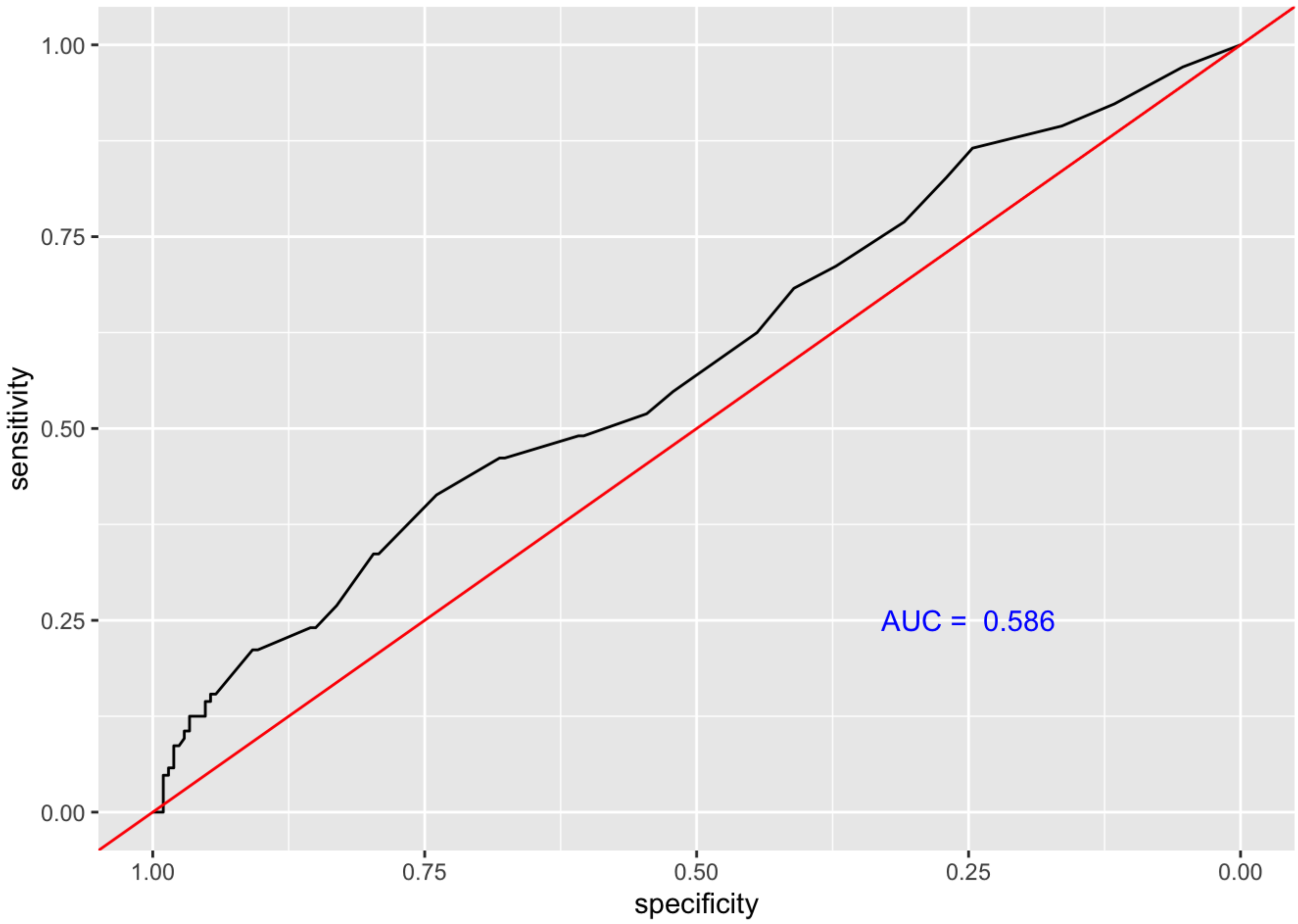
```
##
##              Active Terminated
##   Active        200         93
##   Terminated      7         11
```

This model identified 200 out of 207 Active employees correctly, whereas it identified only 11 out of 104 Terminated employees correctly this suggests that data set is unbalanced and there is not an equal number in Active (207) and Terminated (104) categories. Therefore, this model is not a very good model with the probability cutoff set at 0.5, sensitivity is low at 0.1057692 but specificity is quite good at 0.9661836.

$$Sensitivity(0.5) + Specificity(0.5) = 11/104 + 200/207 = 1.071953$$

This means that there is only a 7% improvement in correct classifications than with no model.

```
# ROC curve
prob_term <- predict(logisticmodel1, newdata = new_data, type = "response")
roc_term <- roc(response = new_data$term_id, predictor = prob_term, auc = TRUE)
ggroc(roc_term) +
  geom_abline(aes(intercept = 1, slope = 1), colour = "red") +
  annotate(geom = "text", x = 0.25, y = 0.25, label = paste("AUC = ", round(auc(roc_t
erm), 3) ), colour = "blue")
```




```
# Finding Youden's index
youden_term <- coords(roc_term, "b", best.method = "youden", transpose = TRUE)
youden_term
```

```
##   threshold specificity sensitivity
##   0.3546303   0.7391304   0.4134615
```

```
youden_term[2] + youden_term [3]
```

```
## specificity
##    1.152592
```

```
roc_term$auc
```

```
## Area under the curve: 0.5856
```

This suggests that the best threshold for this model is around 0.4 with the sensitivity + specificity of 1.15. The area under the ROC curve (AUC) of around 0.59 or around 59% suggests that the quality of the classification model is close to a random model and not quite acceptable.

```
# Fitting the model
logisticmodel2 <-glm(term_id ~ absences + days_late_last30 + recruitment_source, fami
ly = "binomial", data = hr_data)
logisticmodel2
```

```
##
## Call:  glm(formula = term_id ~ absences + days_late_last30 + recruitment_source,
##     family = "binomial", data = hr_data)
##
## Coefficients:
##                              (Intercept)
##                                  -0.6481
##                                 absences
##                                   0.0391
##                         days_late_last30
##                                   0.2054
##      recruitment_sourceDiversity Job Fair
##                                   0.2977
##       recruitment_sourceEmployee Referral
##                                  -1.4501
##           recruitment_sourceGoogle Search
##                                   0.6363
##                  recruitment_sourceIndeed
##                                  -1.0073
##                recruitment_sourceLinkedIn
##                                  -1.0287
## recruitment_sourceOn-line Web application
##                                  15.1751
##                   recruitment_sourceOther
##                                   0.2962
##                 recruitment_sourceWebsite
##                                  -2.3108
##
## Degrees of Freedom: 310 Total (i.e. Null);  300 Residual
## Null Deviance:       396.4
## Residual Deviance: 346.2     AIC: 368.2
```

```
summary(logisticmodel2)
```

```
##
## Call:
## glm(formula = term_id ~ absences + days_late_last30 + recruitment_source,
##     family = "binomial", data = hr_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6114  -0.7683  -0.6238   1.0499   2.3457
##
## Coefficients:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                               -0.64812    0.49002  -1.323   0.1860
## absences                                   0.03910    0.02264   1.727   0.0842
## days_late_last30                           0.20542    0.09583   2.144   0.0321
## recruitment_sourceDiversity Job Fair       0.29771    0.56849   0.524   0.6005
## recruitment_sourceEmployee Referral       -1.45014    0.65006  -2.231   0.0257
## recruitment_sourceGoogle Search            0.63626    0.51918   1.226   0.2204
## recruitment_sourceIndeed                  -1.00733    0.49432  -2.038   0.0416
## recruitment_sourceLinkedIn                -1.02869    0.50439  -2.039   0.0414
## recruitment_sourceOn-line Web application 15.17509  882.74351   0.017   0.9863
## recruitment_sourceOther                    0.29622    1.48753   0.199   0.8422
## recruitment_sourceWebsite                 -2.31084    1.12692  -2.051   0.0403
##
## (Intercept)
## absences                                 .
## days_late_last30                         *
## recruitment_sourceDiversity Job Fair
## recruitment_sourceEmployee Referral      *
## recruitment_sourceGoogle Search
## recruitment_sourceIndeed                 *
## recruitment_sourceLinkedIn               *
## recruitment_sourceOn-line Web application
## recruitment_sourceOther
## recruitment_sourceWebsite                *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 396.37  on 310  degrees of freedom
## Residual deviance: 346.21  on 300  degrees of freedom
## AIC: 368.21
##
## Number of Fisher Scoring iterations: 13
```

The model is in the form of:

$$\begin{aligned} logit(P(term\_id)) = -0.64812 + 0.03910 \times absences \\ +0.20542 \times days\_late\_last30 \\ +0.29771 \times Diversity\ Job\ Fair \\ -1.45014 \times Employee\ Referral \\ +0.63626 \times Google\ Search \\ -1.00733 \times Indeed \\ -1.02869 \times LinkedIn \\ +15.17509 \times Online\ Web\ application \\ +0.29622 \times Other \\ -2.31084 \times Website \end{aligned}$$

If we were to predict the probability of the employee who has had 10 days of absences, has been late 4 times in the last 30 days and was recruited from LinkedIn:

```
# Example prediction with logisticmodel2
predict(logisticmodel2, newdata = data.frame(absences = 10, days_late_last30 = 4, rec
ruitment_source = "LinkedIn"), type = "response")
```

```
##        1
## 0.386009
```

We can expect about 39% chance that this employee would be terminated with these characteristics. If to classify, the employee would be Active, as the probability of with these characteristics being an Active employee is about a 61% chance. Please note the term ‘terminated’ includes employees that have been both terminated for a cause and voluntarily terminated.

Evaluating the predictions with confusion matrix:

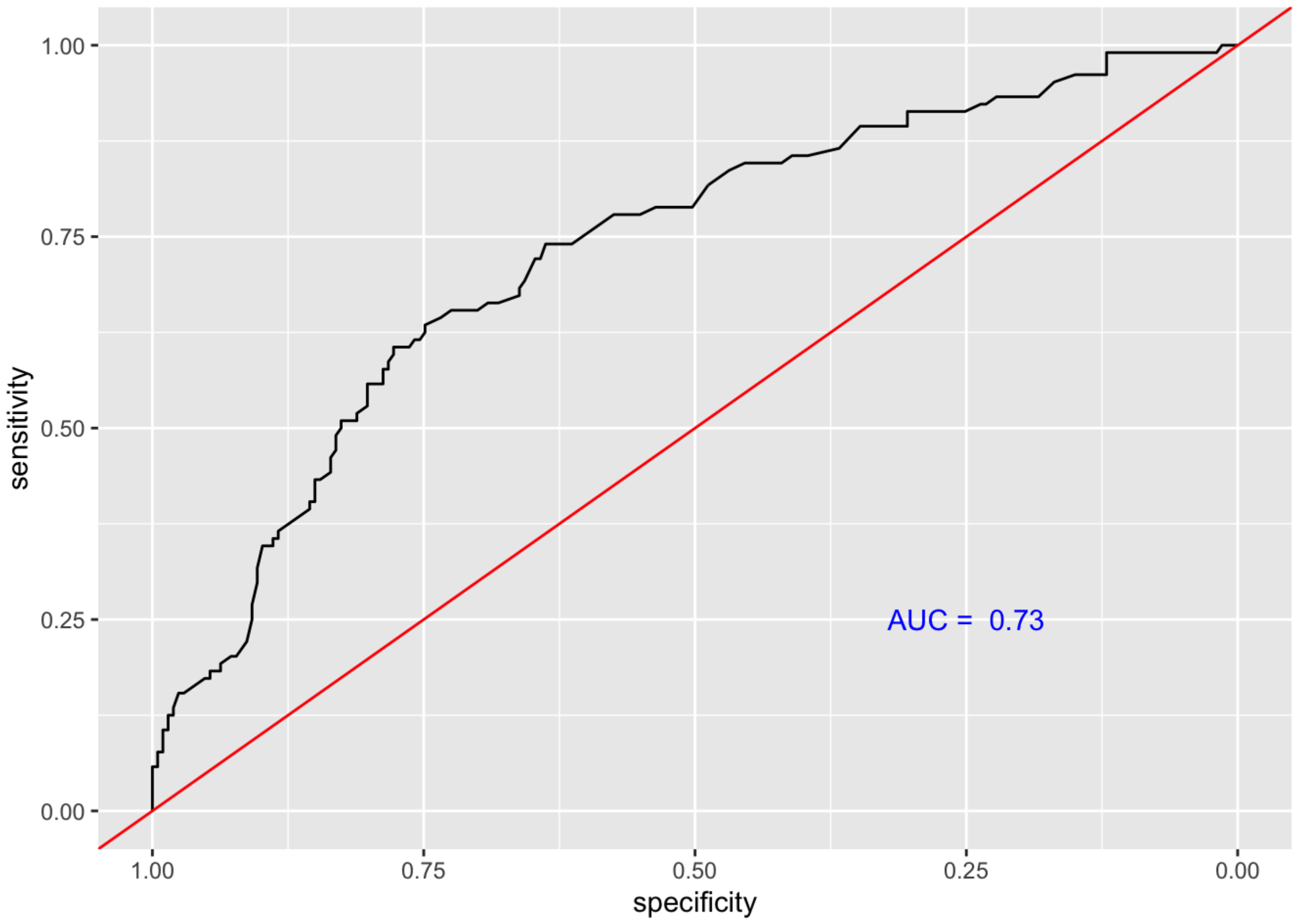
```
# Confusion matrix
new_data2 = hr_data[is.na(hr_data$recruitment_source) == FALSE,]
new_data2$pred_term<-ifelse(predict(logisticmodel2, newdata = new_data2, type = "resp
onse") >= 0.5, "Terminated", "Active")
table(new_data2$pred_term, as.factor(new_data2$term_id))
```

```
##
##              Active Terminated
##   Active        172         55
##   Terminated     35         49
```

This model indentified 172 out of 207 employee Active employees correctly and identified 49 out of 104 Terminated employees correctly. The sensitivity is low but not too bad at 0.4711538 whilst we have a very good specificity at 0.8309179. At the 0.5 threshold the sensitivity + specificty is 30% in correct classifications than with no model, but let’s see if we can improve this further through the ROC curves and Youden’s index.

$$Sensitivity(0.5) + Specificity(0.5) = 49/104 + 172/207 = 1.302072$$

```
# ROC curves
prob_term2 <- predict(logisticmodel2, newdata = new_data2, type = "response")
roc_term2 <- roc(response = new_data2$term_id, predictor = prob_term2, auc = TRUE)
ggroc(roc_term2) +
  geom_abline(aes(intercept = 1, slope = 1), colour = "red") +
  annotate(geom = "text", x = 0.25, y = 0.25, label = paste("AUC = ", round(auc(roc_t
erm2), 3) ), colour = "blue")
```




```
# Finding Youden's index
youden_term2 <- coords(roc_term2, "b", best.method = "youden", transpose = TRUE)
youden_term2
```

```
##   threshold specificity sensitivity
##   0.4027709   0.7777778   0.6057692
```

```
youden_term2[2] + youden_term2 [3]
```

```
## specificity
##    1.383547
```

```
roc_term2$auc
```

```
## Area under the curve: 0.7301
```

The best threshold is around 0.4, the sensitivy + specificty is around 1.39 or 39% better than no model which is an improvement compared to the threshold at 0.5 where we got 30% in correct classifications. The AUC at 0.73 suggests that this model is an acceptable model.

Comparing the models:

- If we compare the models through AIC, logisticmodel1 model was at 393.88 with predictor variables absences and days_late_last30 and logisticmodel2 model was at 368.21 with an additional predictor variable recruitment_source and its 9 parameters. The model logisticmodel2 has the lower AIC score which indicates that there is less information loss making it a more explanatory model.

5) Using only the numerical variable in your dataset and the clustering methods seen in class,

explore whether the data tend to form clusters. [10 points]

Question not required to be answered as previously agreed by the department “If you choose to sit the original piece of work the marker will take into account the content included in the brief which was not taught in your module.” - SCI ProgAdmin

6) Again, using only the numerical variable in your dataset, determine whether this set of variables

could be summarized into a smaller number of representative variables (i.e., principal components) [10 points]

Question not required to be answered as previously agreed by the department “If you choose to sit the original piece of work the marker will take into account the content included in the brief which was not taught in your module.” - SCI ProgAdmin