HR Analytics – Predicting Employee Churn

Paul Bedu-Osei

2025-12-18


```
# ==============================================================
# Setup
# ==============================================================
# Knit-safe options
options(mc.cores = 1)

library(readr)
library(dplyr)
library(Information)
library(caret)
library(car)
library(tidypredict)
library(ggplot2)
library(lubridate)
```

1. Business Context

This analysis was conducted for a HR client to **understand employee turnover** and **predict churn risk** among employees. The goal is to identify key drivers of attrition and quantify the potential **ROI of targeted retention strategies**.

---

2. Data Loading

```
org    <- read_csv("~/Desktop/employee_data/org.csv",show_col_types = FALSE)
rating <- read_csv("~/Desktop/employee_data/rating.csv",show_col_types = FALSE)
survey <- read_csv("~/Desktop/employee_data/survey.csv",show_col_types = FALSE)
```

---

3. Exploratory Data Analysis (EDA)

Workforce Overview

```
glimpse(org)
```

```
## Rows: 2,291
## Columns: 14
## $ emp_id            <chr> "E11061", "E1031", "E6213", "E5900", "E3044", "E4008…
## $ status            <chr> "Inactive", "Inactive", "Inactive", "Inactive", "Ina…
## $ turnover          <dbl> 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0…
## $ location          <chr> "New York", "New York", "New York", "New York", "Flo…
## $ level             <chr> "Analyst", "Analyst", "Analyst", "Analyst", "Analyst…
## $ date_of_joining   <chr> "22/03/2012", "09/03/2012", "06/01/2012", "22/03/201…
## $ date_of_birth     <chr> "22/03/1992", "10/01/1992", "06/02/1992", "19/12/199…
## $ last_working_date <chr> "11/09/2014", "05/06/2014", "30/04/2014", "09/04/201…
## $ gender            <chr> "Male", "Female", "Female", "Female", "Female", "Fem…
## $ department        <chr> "Customer Operations", "Customer Operations", "Custo…
## $ mgr_id            <chr> "E1712", "E10524", "E4443", "E3638", "E3312", "E1393…
## $ cutoff_date       <chr> "31/12/2014", "31/12/2014", "31/12/2014", "31/12/201…
## $ generation        <chr> "Millennials", "Millennials", "Millennials", "Millen…
## $ emp_age           <dbl> 22.5, 22.4, 22.2, 22.3, 22.1, 23.0, 23.0, 23.4, 23.0…
```

```
org %>% count(status)
```

```
## # A tibble: 2 × 2
##   status       n
##   <chr>    <int>
## 1 Active    1881
## 2 Inactive   410
```

```
org %>% summarise(avg_turnover_rate = mean(turnover, na.rm = TRUE))
```

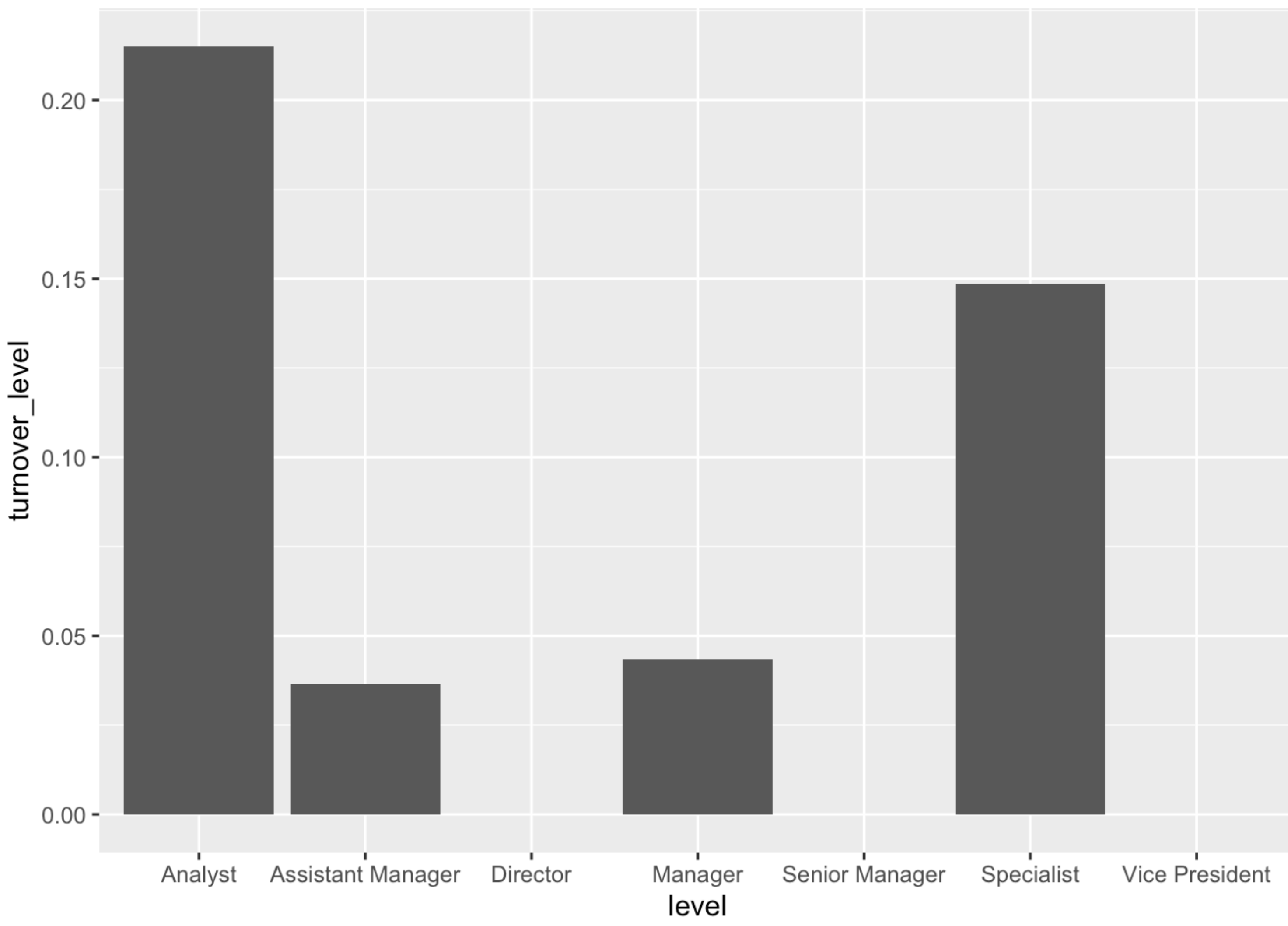
```
## # A tibble: 1 × 1
##   avg_turnover_rate
##               <dbl>
## 1             0.179
```

Turnover by Level

```
df_level <- org %>%
  group_by(level) %>%
  summarise(turnover_level = mean(turnover, na.rm = TRUE))

df_level
```

```
## # A tibble: 7 × 2
##   level             turnover_level
##   <chr>                      <dbl>
## 1 Analyst                   0.215
## 2 Assistant Manager         0.0365
## 3 Director                  0
## 4 Manager                   0.0435
## 5 Senior Manager            0
## 6 Specialist                0.149
## 7 Vice President            0
```

```
ggplot(df_level, aes(level, turnover_level)) + geom_col()
```

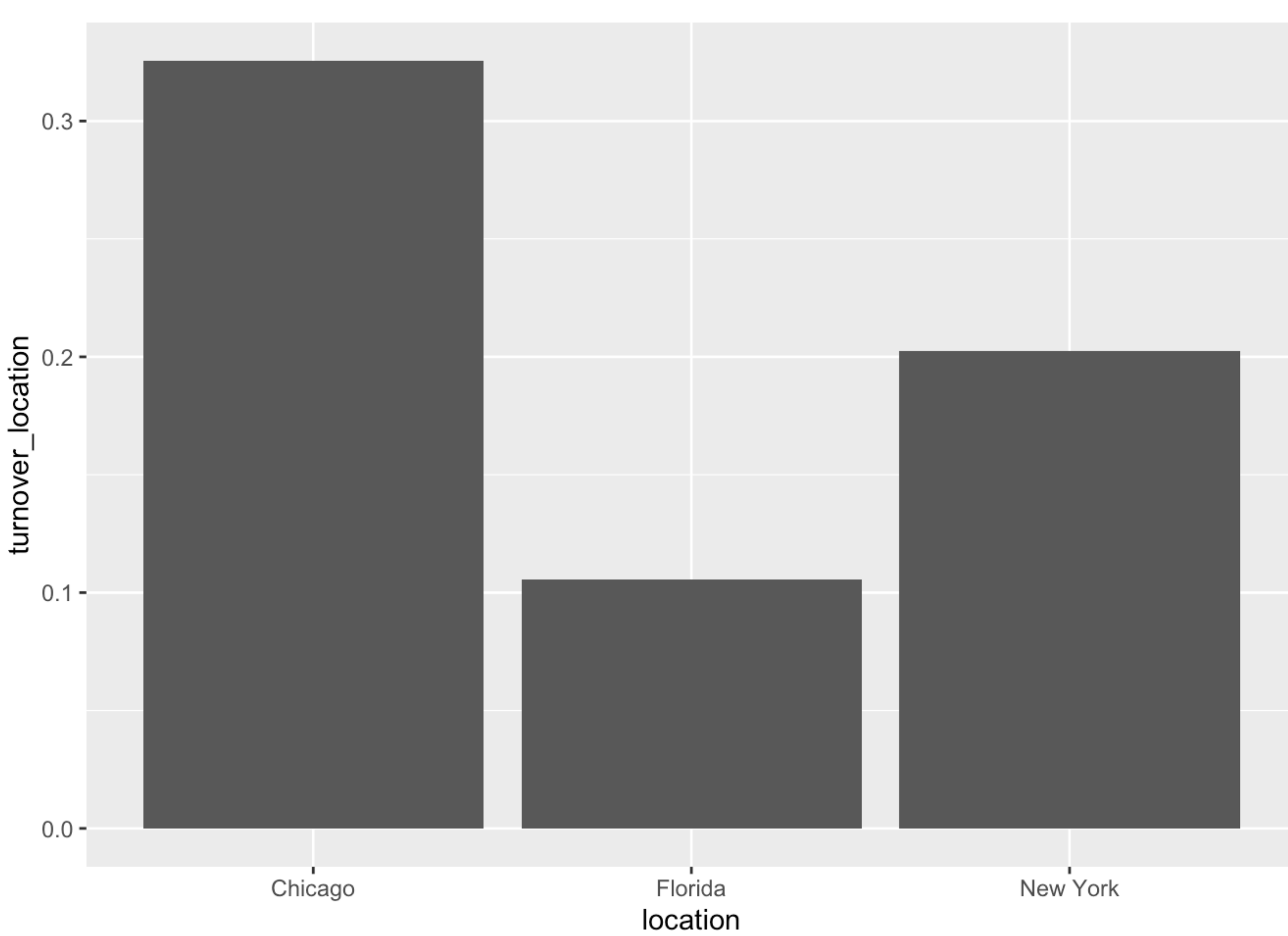


Turnover by Location

```
df_location <- org %>%
  group_by(location) %>%
  summarise(turnover_location = mean(turnover, na.rm = TRUE))

df_location
```

```
## # A tibble: 3 × 2
##   location turnover_location
##   <chr>                <dbl>
## 1 Chicago              0.326
## 2 Florida              0.106
## 3 New York             0.203
```

```
ggplot(df_location, aes(location, turnover_location)) + geom_col()
```



4. Data Preparation & Feature Engineering

Filter Relevant Roles

```
org2 <- org %>% filter(level %in% c("Analyst", "Specialist"))
org2 %>% count(level)
```

```
## # A tibble: 2 × 2
##   level          n
##   <chr>      <int>
## 1 Analyst     1604
## 2 Specialist   350
```

Join Performance & Survey Data

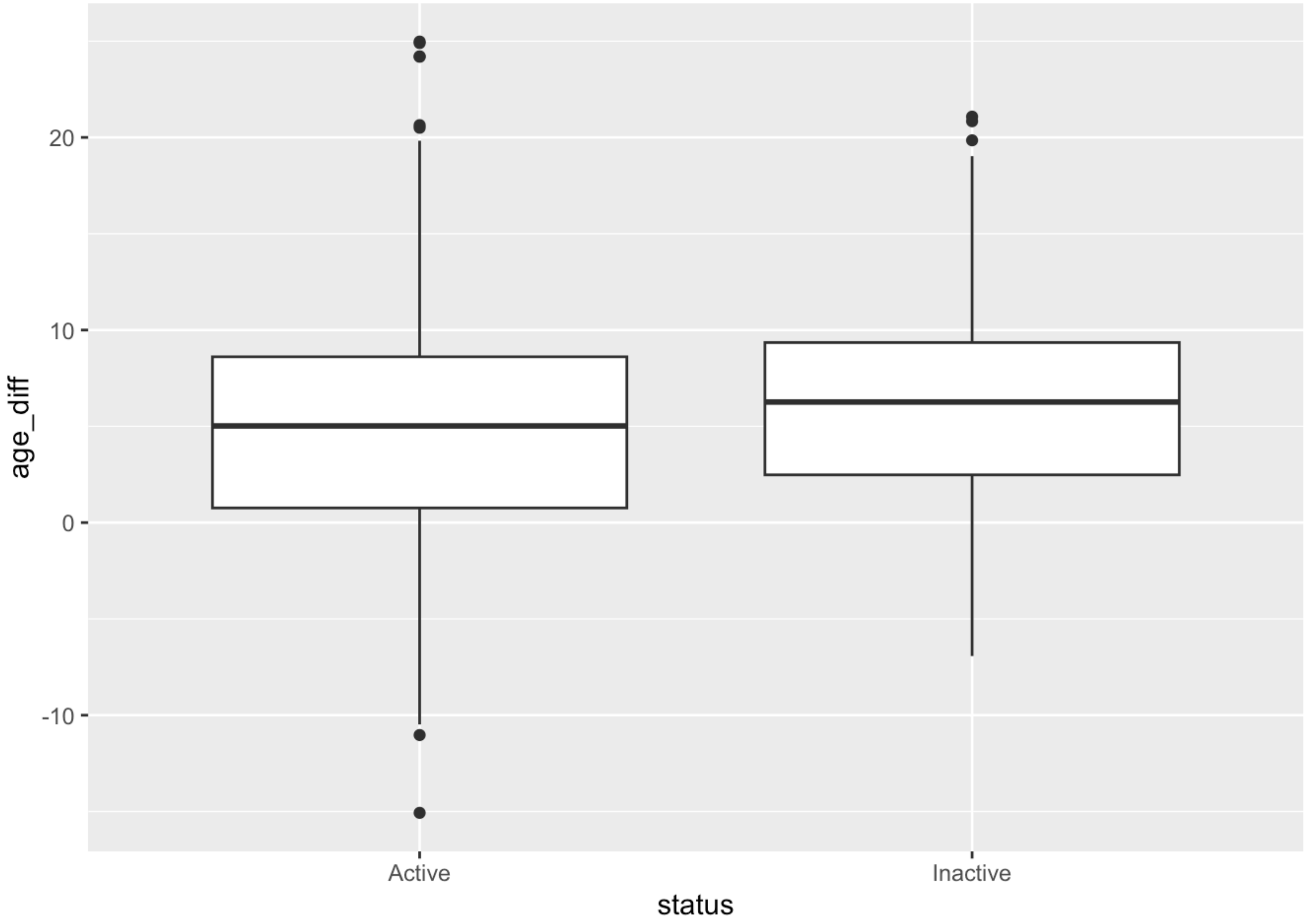
```
org3 <- left_join(org2, rating, by = "emp_id")
org_final <- left_join(org3, survey, by = "mgr_id")
```

Engineer New Features

```
org_final1 <- read_csv("~/Desktop/employee_data/org_final.csv",show_col_types = FALSE)

# View the structure of updated org final dataset
glimpse(org_final1)
```

```
## Rows: 1,954
## Columns: 34
## $ emp_id                      <chr> "E10012", "E10025", "E10027", "E10048", "…
## $ status                      <chr> "Active", "Active", "Active", "Active", "…
## $ location                    <chr> "New York", "Chicago", "Orlando", "Chicag…
## $ level                       <chr> "Analyst", "Analyst", "Specialist", "Spec…
## $ gender                      <chr> "Female", "Female", "Female", "Male", "Ma…
## $ emp_age                     <dbl> 25.09, 25.98, 33.40, 24.55, 31.23, 31.98,…
## $ rating                      <chr> "Above Average", "Acceptable", "Acceptabl…
## $ mgr_rating                  <chr> "Acceptable", "Excellent", "Above Average…
## $ mgr_reportees               <dbl> 9, 4, 6, 10, 11, 19, 21, 9, 12, 22, 17, 1…
## $ mgr_age                     <dbl> 44.07, 35.99, 35.78, 26.70, 34.28, 34.82,…
## $ mgr_tenure                  <dbl> 3.17, 7.92, 4.38, 2.87, 12.95, 10.88, 4.0…
## $ compensation                <dbl> 64320, 48204, 85812, 49536, 75576, 56904,…
## $ percent_hike                <dbl> 10, 8, 11, 8, 12, 8, 12, 9, 9, 6, 11, 7, …
## $ hiring_score                <dbl> 70, 70, 77, 71, 70, 75, 72, 70, 70, 70, 7…
## $ hiring_source               <chr> "Consultant", "Job Fairs", "Consultant", …
## $ no_previous_companies_worked <dbl> 0, 9, 3, 5, 0, 8, 9, 6, 1, 3, 3, 6, 2, 6,…
## $ distance_from_home          <dbl> 14, 21, 15, 9, 25, 23, 17, 16, 22, 22, 18…
## $ total_dependents            <dbl> 2, 2, 5, 3, 4, 5, 2, 5, 2, 5, 5, 5, 4, 5,…
## $ marital_status              <chr> "Single", "Single", "Single", "Single", "…
## $ education                   <chr> "Bachelors", "Bachelors", "Bachelors", "B…
## $ promotion_last_2_years      <chr> "No", "No", "Yes", "Yes", "No", "No", "No…
## $ no_leaves_taken             <dbl> 2, 10, 18, 19, 25, 15, 10, 20, 22, 23, 24…
## $ total_experience            <dbl> 6.86, 4.88, 8.55, 4.76, 8.06, 13.72, 5.81…
## $ monthly_overtime_hrs        <dbl> 1, 5, 3, 8, 1, 7, 2, 10, 2, 10, 8, 3, 1, …
## $ date_of_joining             <chr> "06/03/2011", "23/09/2009", "02/11/2005",…
## $ last_working_date           <chr> NA, NA, NA, NA, NA, "11/12/2014", NA, NA,…
## $ department                  <chr> "Customer Operations", "Customer Operatio…
## $ mgr_id                      <chr> "E9335", "E6655", "E13942", "E7063", "E56…
## $ cutoff_date                 <chr> "31/12/2014", "31/12/2014", "31/12/2014",…
## $ turnover                    <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1,…
## $ mgr_effectiveness           <dbl> 0.730, 0.581, 0.770, 0.240, 0.710, 0.574,…
## $ career_satisfaction         <dbl> 0.73, 0.72, 0.85, 0.42, 0.78, 0.88, 0.68,…
## $ perf_satisfaction           <dbl> 0.73, 0.84, 0.80, 0.33, 0.67, 0.81, 0.57,…
## $ work_satisfaction           <dbl> 0.75, 0.85, 0.87, 0.85, 0.80, 0.86, 0.75,…
```

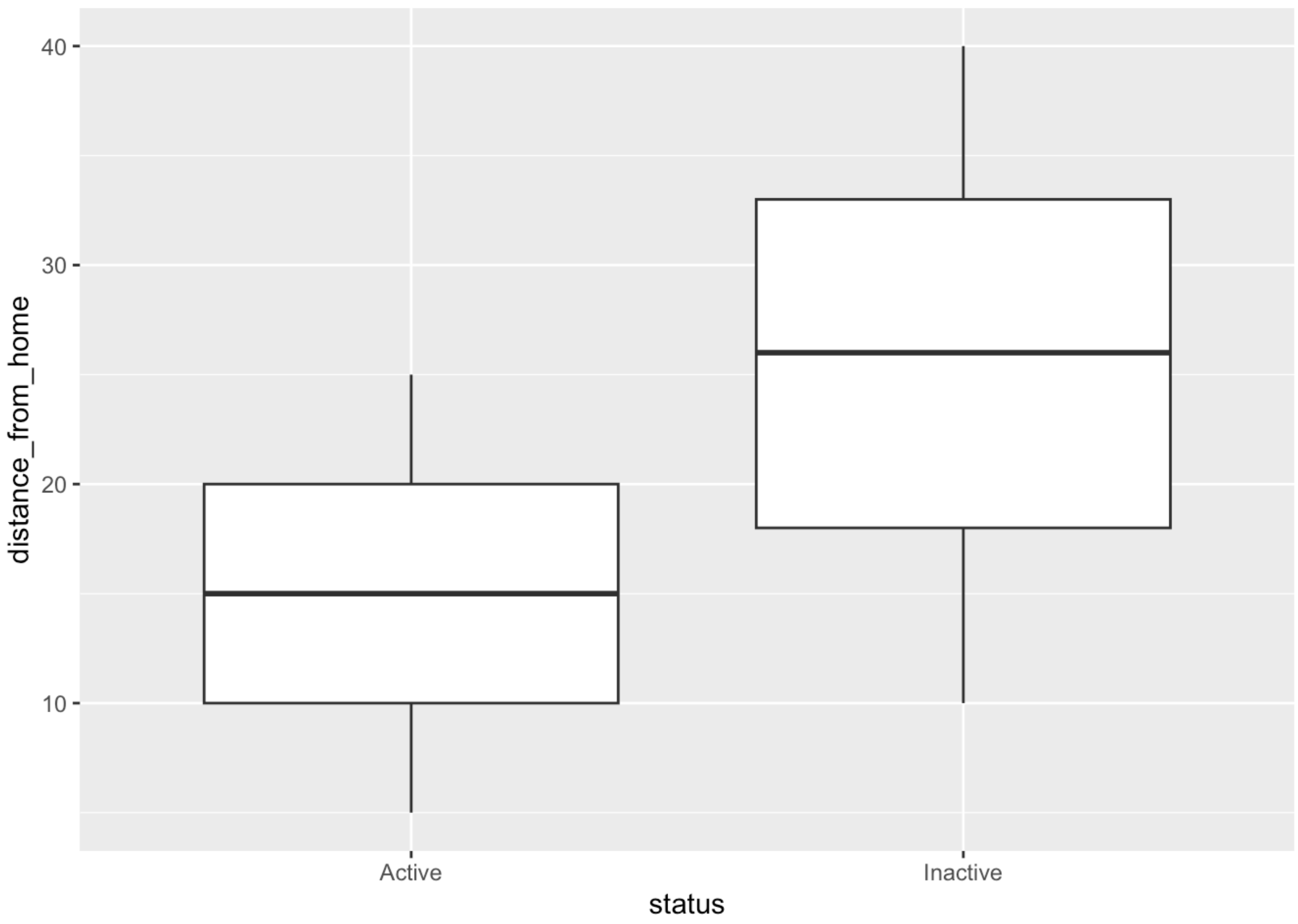
```
# Add age_diff
emp_age_diff <- org_final1 %>%
  mutate(age_diff = mgr_age - emp_age)

# Plot the distribution of age difference
ggplot(emp_age_diff, aes(x = status, y = age_diff)) +
  geom_boxplot()
```



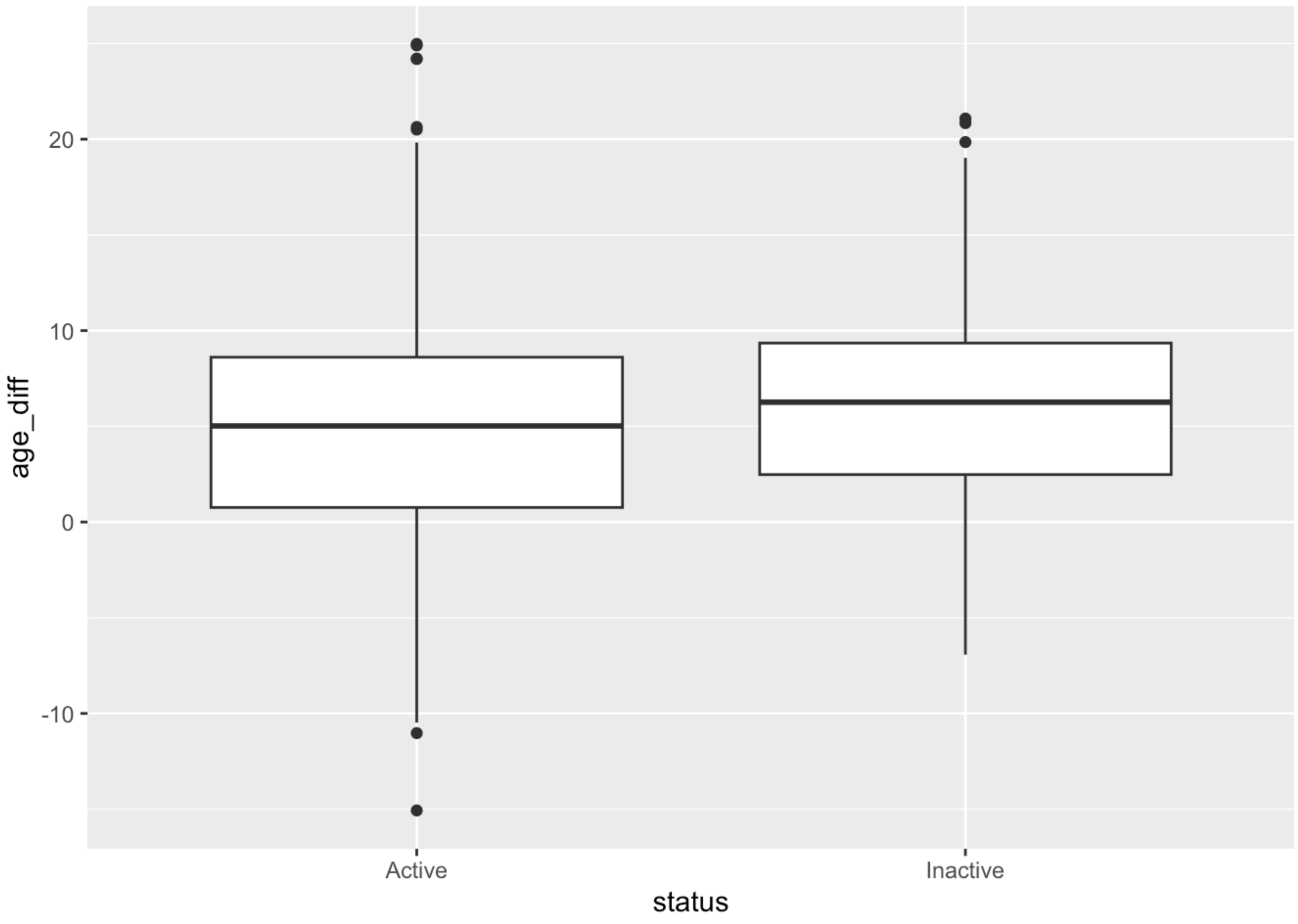

```
emp_features <- org_final1 %>%
  mutate(
    age_diff = mgr_age - emp_age,
    job_hop_index = if_else(no_previous_companies_worked > 0,
                            total_experience / no_previous_companies_worked,
                            NA_real_),
    tenure = ifelse(
      status == "Active",
      time_length(interval(date_of_joining, cutoff_date), "years"),
      time_length(interval(date_of_joining, last_working_date), "years")
    )
  )
```

```
## Warning: There were 4 warnings in `mutate()`.
## The first warning was:
## ℹ In argument: `tenure = ifelse(...)`.
## Caused by warning:
## ! All formats failed to parse. No formats found.
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
```

```
# Compare the travel distance of Active and Inactive employees
ggplot(org_final1, aes(x = status, y = distance_from_home)) +
  geom_boxplot()
```

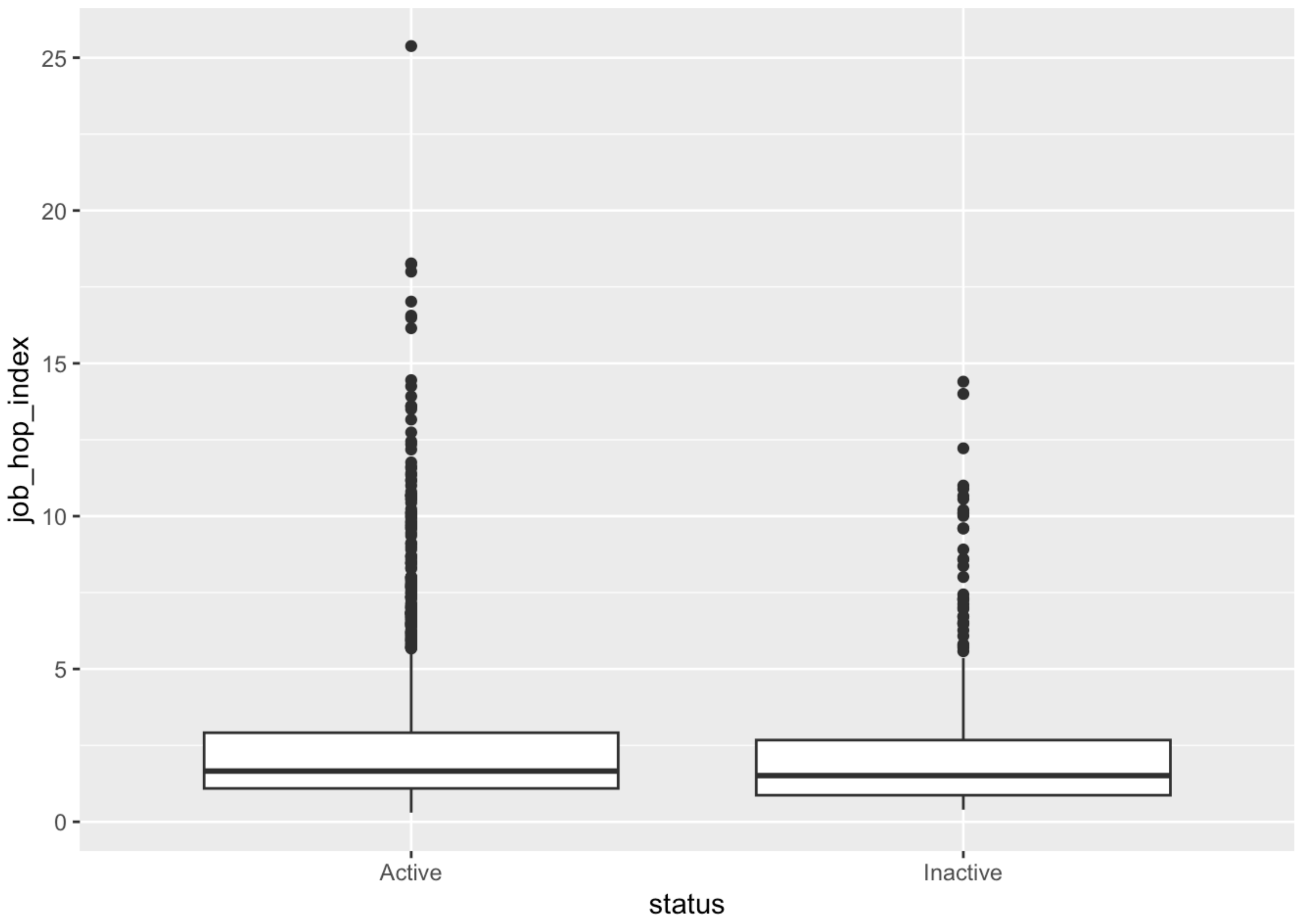



```
# Plot the distribution of age difference
ggplot(emp_features, aes(x = status, y = age_diff)) +
  geom_boxplot()
```



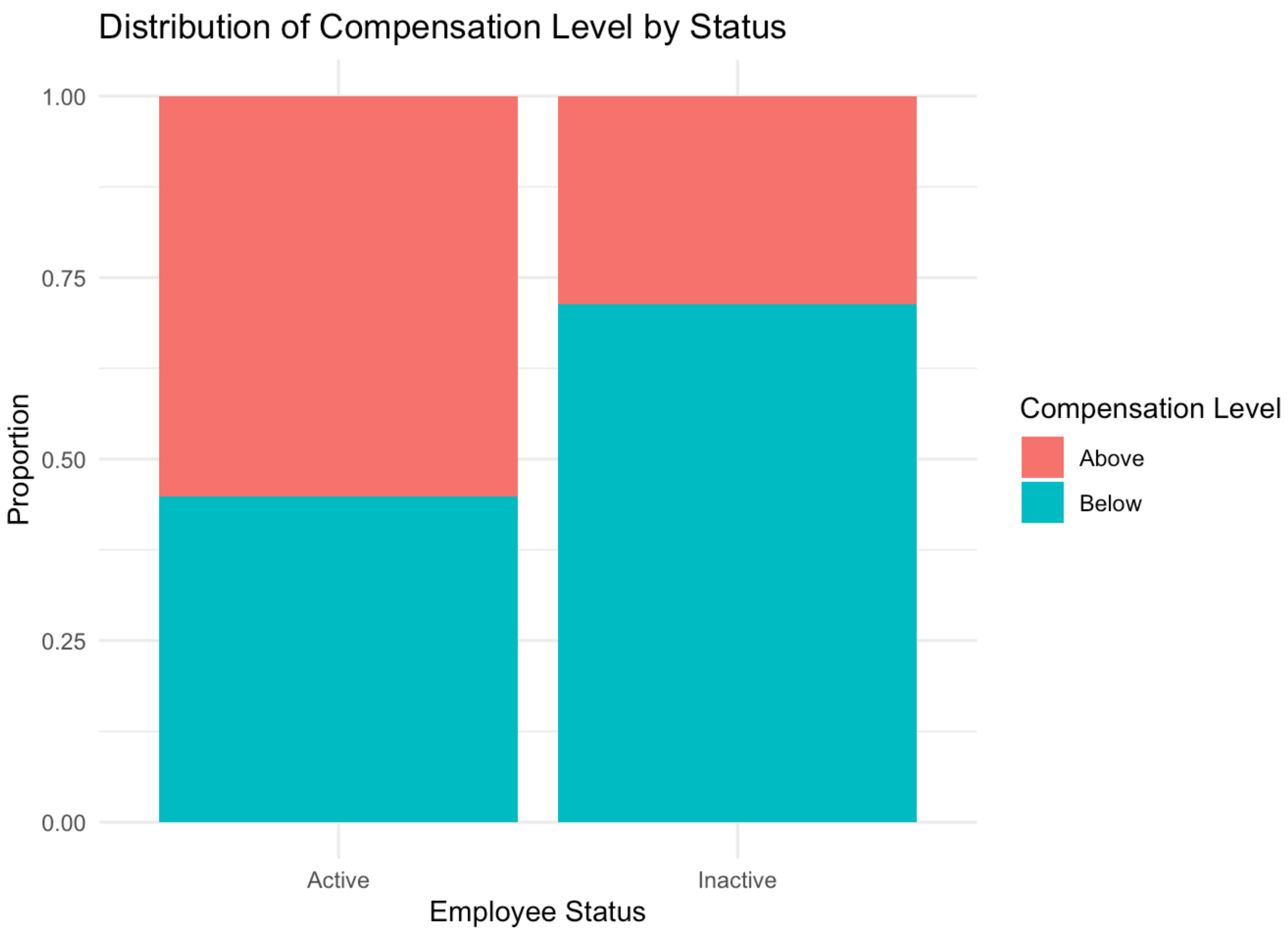

```
# Compare job hopping index of Active and Inactive employees
ggplot(emp_features, aes(x = status, y = job_hop_index)) +
  geom_boxplot()
```

```
## Warning: Removed 186 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

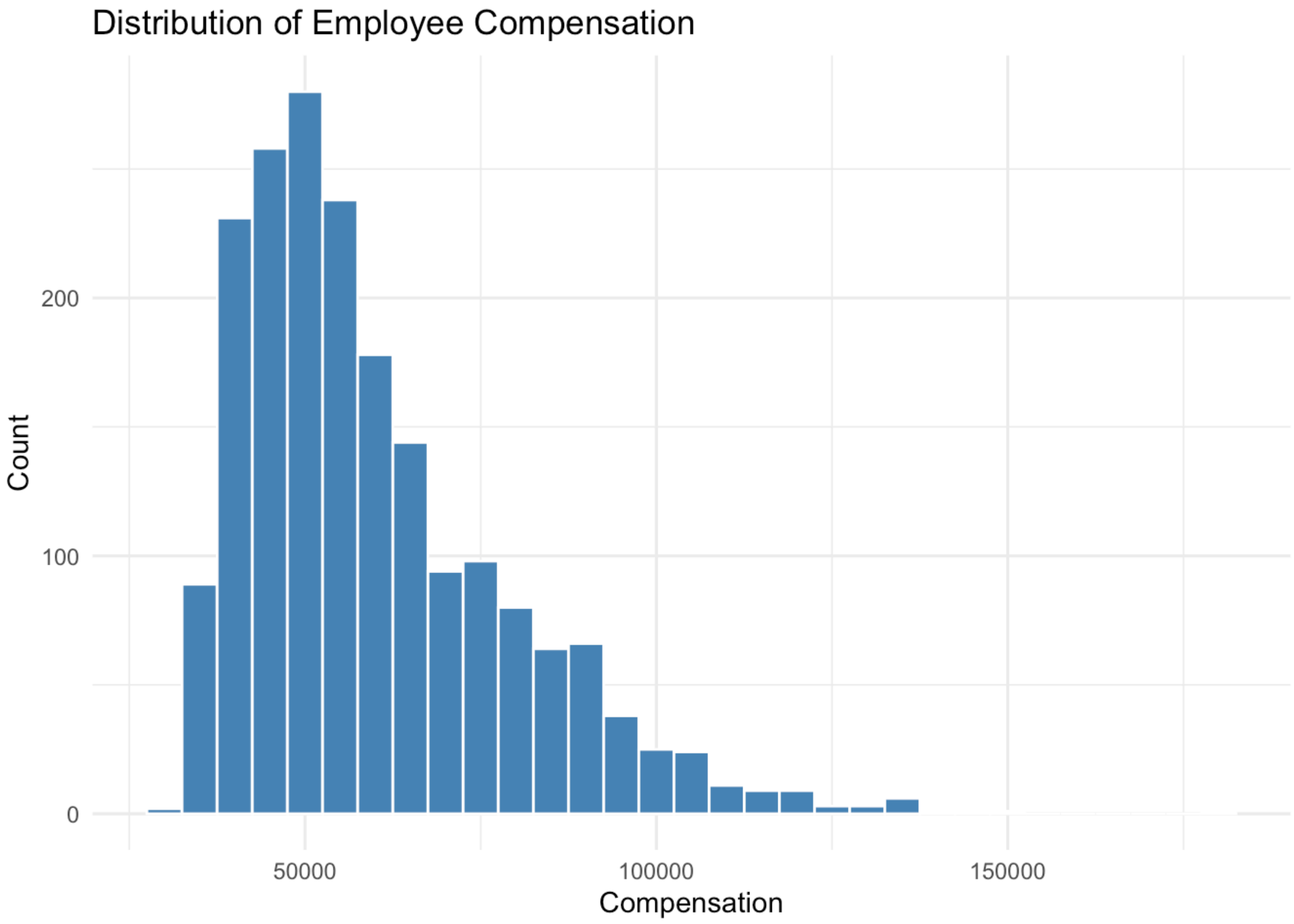


5. Compensation Analysis

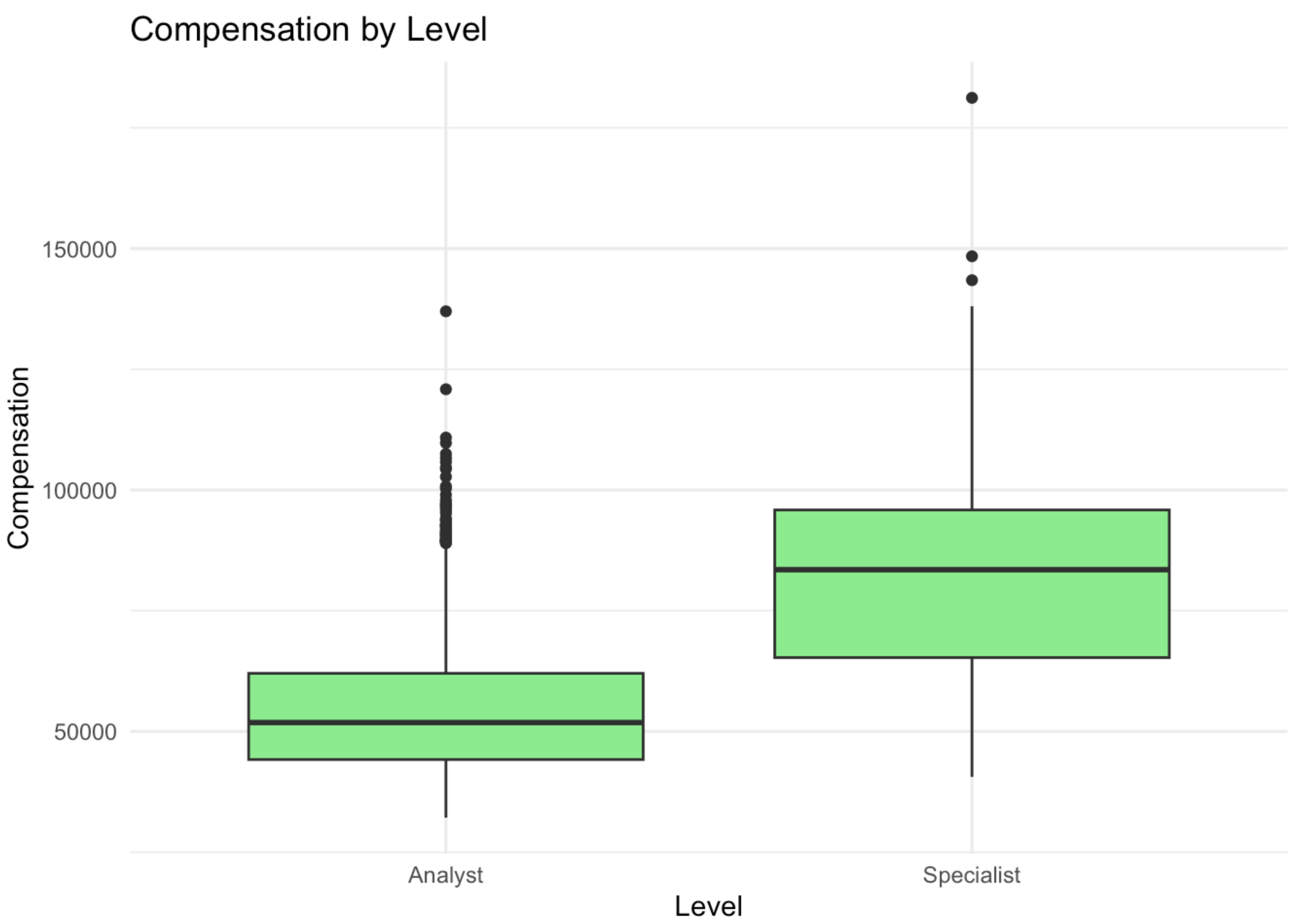
```
# Calculate median compensation and compa_ratio, then classify compa_level
emp_compa <- emp_features %>%
  group_by(level) %>%
  mutate(
    median_compensation = median(compensation, na.rm = TRUE),
    compa_ratio = compensation / median_compensation,
    compa_level = factor(if_else(compa_ratio > 1, "Above", "Below"))
  ) %>%
  ungroup()

# Plot the distribution of compa_level across status
ggplot(emp_compa, aes(x = status, fill = compa_level)) +
  geom_bar(position = "fill") +
  labs(
    title = "Distribution of Compensation Level by Status",
    x = "Employee Status",
    y = "Proportion",
    fill = "Compensation Level"
  ) +
  theme_minimal()
```



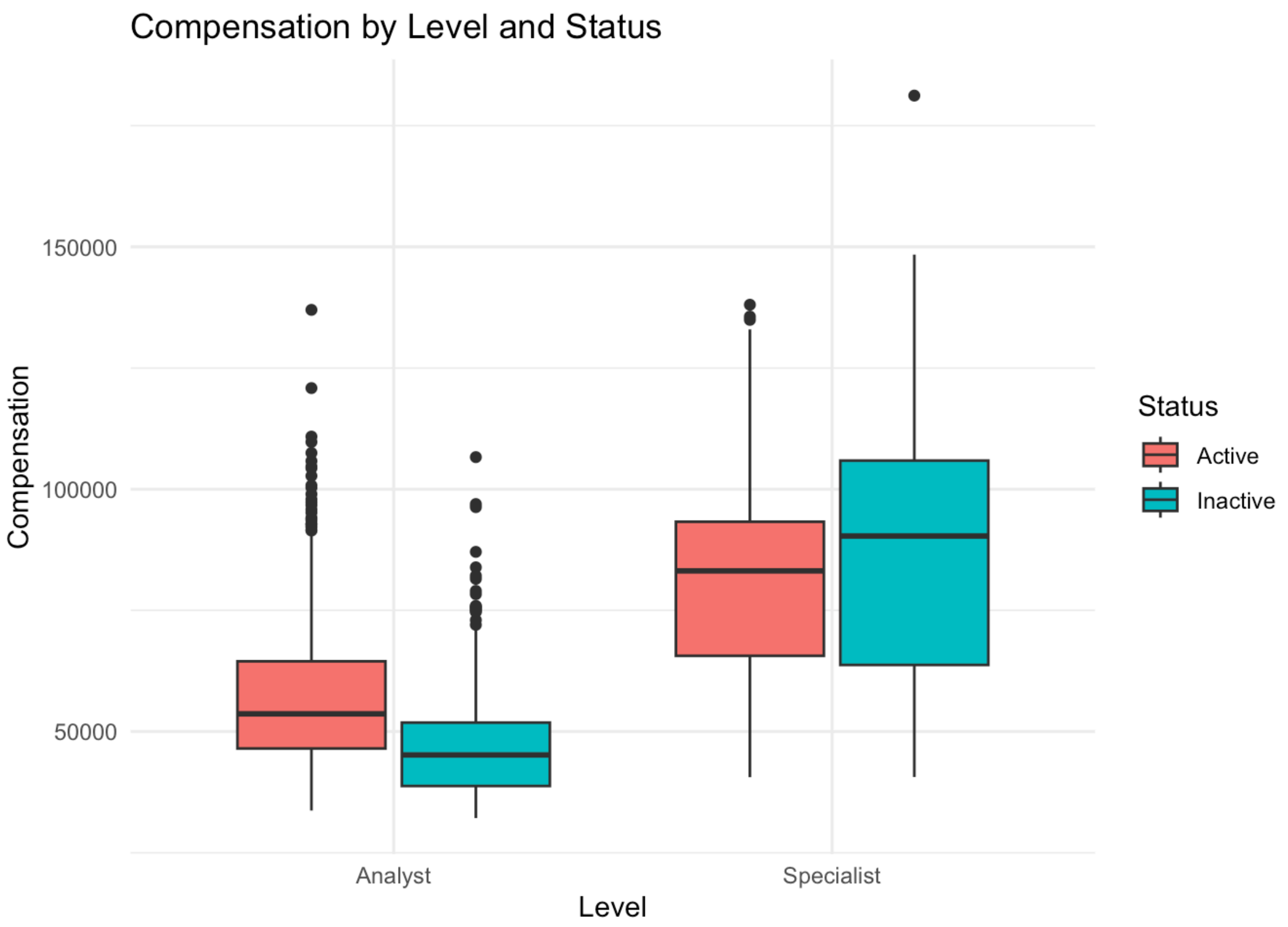

```
# Plot the distribution of compensation
ggplot(emp_features, aes(x = compensation)) +
  geom_histogram(binwidth = 5000, fill = "steelblue", color = "white") +
  labs(
    title = "Distribution of Employee Compensation",
    x = "Compensation",
    y = "Count"
  ) +
  theme_minimal()
```



```
# Plot the distribution of compensation across levels
ggplot(emp_features, aes(x = level, y = compensation)) +
  geom_boxplot(fill = "lightgreen") +
  labs(
    title = "Compensation by Level",
    x = "Level",
    y = "Compensation"
  ) +
  theme_minimal()
```




```
# Compare compensation of Active and Inactive employees across levels
ggplot(emp_features, aes(x = level, y = compensation, fill = status)) +
  geom_boxplot() +
  labs(
    title = "Compensation by Level and Status",
    x = "Level",
    y = "Compensation",
    fill = "Status"
  ) +
  theme_minimal()
```




```
# Add median_compensation and compa_ratio
emp_compa_ratio <- emp_features %>%
  group_by(level) %>%
  mutate(
    median_compensation = median(compensation, na.rm = TRUE),
    compa_ratio = compensation / median_compensation
  )

# Look at the median compensation for each level
emp_compa_ratio %>%
  distinct(level, median_compensation)
```

```
## # A tibble: 2 × 2
## # Groups:   level [2]
##   level      median_compensation
##   <chr>                    <dbl>
## 1 Analyst                  51840
## 2 Specialist               83496
```

```
# Add compa_level
emp_final <- emp_compa_ratio %>%
  mutate(compa_level = case_when(
    compa_ratio > 1 ~ "Above",
    TRUE ~ "Below"
  ))
```

6. Information Value (Feature Strength)

```
IV <- create_infotables(emp_compa, y = "turnover", parallel = FALSE)
```

```
## [1] "Variable emp_id was removed because it is a non-numeric variable with >1000 categories"
## [1] "Variable department was removed because it has only 1 unique value"
## [1] "Variable cutoff_date was removed because it has only 1 unique value"
## [1] "Variable tenure was removed because it has only 1 unique level"
```

```
IV$Summary
```

```
##                         Variable           IV
## 12                  percent_hike 1.144784e+00
## 17              total_dependents 1.088645e+00
## 21               no_leaves_taken 9.404533e-01
## 27             mgr_effectiveness 6.830020e-01
## 11                  compensation 6.074885e-01
## 34                   compa_ratio 4.768892e-01
## 24               date_of_joining 4.330804e-01
## 6                         rating 3.869373e-01
## 23          monthly_overtime_hrs 3.786644e-01
## 8                  mgr_reportees 3.620543e-01
## 2                       location 2.963023e-01
## 35                   compa_level 2.940446e-01
## 26                        mgr_id 2.820235e-01
## 5                        emp_age 2.275477e-01
## 16            distance_from_home 1.470549e-01
## 30             work_satisfaction 1.378953e-01
## 22              total_experience 1.345781e-01
## 19                     education 1.253865e-01
## 20       promotion_last_2_years 9.979915e-02
## 9                        mgr_age 9.816205e-02
## 29             perf_satisfaction 7.099511e-02
## 13                  hiring_score 6.684727e-02
## 31                      age_diff 6.634065e-02
## 32                 job_hop_index 6.605312e-02
## 10                    mgr_tenure 5.918048e-02
## 28           career_satisfaction 3.539857e-02
## 3                          level 2.726491e-02
## 33           median_compensation 2.726491e-02
## 18                marital_status 2.588063e-02
## 7                     mgr_rating 2.172222e-02
## 15 no_previous_companies_worked 1.729893e-02
## 14                 hiring_source 8.773529e-03
## 4                         gender 3.959968e-05
## 1                         status 0.000000e+00
## 25             last_working_date 0.000000e+00
```

7. Modeling Approach

Train / Test Split

```
set.seed(567)
index_train <- createDataPartition(emp_compa$turnover, p = 0.7, list = FALSE)

train_set <- emp_compa[index_train, ]
test_set  <- emp_compa[-index_train, ]
```

Logistic Regression Model

```
# Calculate turnover proportion in train_set
train_set %>%
  count(status) %>%
  mutate(prop = n / sum(n))
```

```
## # A tibble: 2 × 3
##   status       n  prop
##   <chr>    <int> <dbl>
## 1 Active    1094 0.800
## 2 Inactive   274 0.200
```

```
# Calculate turnover proportion in test_set
test_set %>%
  count(status) %>%
  mutate(prop = n / sum(n))
```

```
## # A tibble: 2 × 3
##   status       n  prop
##   <chr>    <int> <dbl>
## 1 Active     463 0.790
## 2 Inactive   123 0.210
```

```
# Taking some columns from the dataset
train_set_multi <- emp_final %>% select( (-c("emp_id", "mgr_id","date_of_joining", "last_working_
date", "cutoff_date", "mgr_age", "emp_age","median_compensation","department","status",, "tenur
e")))
train_set_multi
```

```
## # A tibble: 1,954 × 29
## # Groups:   level [2]
##    location level gender rating mgr_rating mgr_reportees mgr_tenure compensation
##    <chr>    <chr> <chr>  <chr>  <chr>              <dbl>      <dbl>        <dbl>
##  1 New York Anal… Female Above… Acceptable             9       3.17        64320
##  2 Chicago  Anal… Female Accep… Excellent              4       7.92        48204
##  3 Orlando  Spec… Female Accep… Above Ave…             6       4.38        85812
##  4 Chicago  Spec… Male   Accep… Acceptable            10       2.87        49536
##  5 Orlando  Anal… Male   Accep… Acceptable            11      13.0         75576
##  6 Orlando  Anal… Male   Below… Above Ave…            19      10.9         56904
##  7 Chicago  Anal… Male   Accep… Above Ave…            21       4.01        38772
##  8 Orlando  Anal… Male   Above… Above Ave…             9       4.21        52320
##  9 New York Anal… Female Accep… Acceptable            12       1.27        50940
## 10 New York Anal… Male   Accep… Acceptable            22       4.87        40380
## # ℹ 1,944 more rows
## # ℹ 21 more variables: percent_hike <dbl>, hiring_score <dbl>,
## #   hiring_source <chr>, no_previous_companies_worked <dbl>,
## #   distance_from_home <dbl>, total_dependents <dbl>, marital_status <chr>,
## #   education <chr>, promotion_last_2_years <chr>, no_leaves_taken <dbl>,
## #   total_experience <dbl>, monthly_overtime_hrs <dbl>, turnover <dbl>,
## #   mgr_effectiveness <dbl>, career_satisfaction <dbl>, …
```

```
# Build a simple logistic regression model
simple_log <- glm(turnover~percent_hike,
                  family = "binomial", data = train_set_multi)

# Print summary
summary(simple_log)
```

```
##
## Call:
## glm(formula = turnover ~ percent_hike, family = "binomial", data = train_set_multi)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.49061    0.18647   7.994 1.31e-15 ***
## percent_hike -0.30700    0.02031 -15.113  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1972.6  on 1953  degrees of freedom
## Residual deviance: 1681.2  on 1952  degrees of freedom
## AIC: 1685.2
##
## Number of Fisher Scoring iterations: 5
```

```r
# Build a multiple logistic regression model
multi_log <- glm(
  # Manually list variables, *omitting* 'compa_level' and 'job_hop_index'
  turnover ~ location + level + gender + rating + mgr_rating + mgr_reportees +
    mgr_tenure + compensation + percent_hike + hiring_score + hiring_source +
    no_previous_companies_worked + distance_from_home + total_dependents +
    marital_status + education + promotion_last_2_years + no_leaves_taken +
    total_experience + monthly_overtime_hrs + mgr_effectiveness +
    career_satisfaction + perf_satisfaction + work_satisfaction +
    age_diff + compa_ratio ,
  family = "binomial",
  data = train_set_multi,
  na.action = na.omit
)
# Print summary
summary(multi_log)
```

```
##
## Call:
## glm(formula = turnover ~ location + level + gender + rating +
##     mgr_rating + mgr_reportees + mgr_tenure + compensation +
##     percent_hike + hiring_score + hiring_source + no_previous_companies_worked +
##     distance_from_home + total_dependents + marital_status +
##     education + promotion_last_2_years + no_leaves_taken + total_experience +
##     monthly_overtime_hrs + mgr_effectiveness + career_satisfaction +
##     perf_satisfaction + work_satisfaction + age_diff + compa_ratio,
##     family = "binomial", data = train_set_multi, na.action = na.omit)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -8.814e+00  3.095e+00  -2.848 0.004401 **
## locationNew York               9.358e-01  3.520e-01   2.658 0.007851 **
## locationOrlando               -1.128e+00  2.985e-01  -3.778 0.000158 ***
## levelSpecialist                1.359e+01  5.128e+02   0.026 0.978863
## genderMale                     4.328e-01  2.576e-01   1.680 0.092916 .
## ratingAcceptable              -3.796e-01  2.929e-01  -1.296 0.194993
## ratingBelow Average           -2.383e+00  5.520e-01  -4.317 1.58e-05 ***
## ratingExcellent               -6.915e-01  7.573e-01  -0.913 0.361161
## ratingUnacceptable            -3.834e+00  9.385e-01  -4.085 4.40e-05 ***
## mgr_ratingAcceptable           1.034e-01  2.809e-01   0.368 0.712673
## mgr_ratingBelow Average       -8.151e-01  5.001e-01  -1.630 0.103142
## mgr_ratingExcellent           -1.099e-01  3.893e-01  -0.282 0.777802
## mgr_ratingUnacceptable         1.041e+00  1.028e+00   1.012 0.311418
## mgr_reportees                  8.033e-02  2.286e-02   3.514 0.000442 ***
## mgr_tenure                    -8.668e-02  3.330e-02  -2.603 0.009246 **
## compensation                   8.527e-05  3.315e-05   2.572 0.010112 *
## percent_hike                  -5.585e-01  6.208e-02  -8.996  < 2e-16 ***
## hiring_score                   6.084e-02  3.666e-02   1.659 0.097018 .
## hiring_sourceConsultant       -4.977e-01  4.359e-01  -1.142 0.253562
## hiring_sourceEmployee Referral -1.470e-01  4.616e-01  -0.318 0.750212
## hiring_sourceJob Boards       -3.201e-01  4.474e-01  -0.715 0.474377
## hiring_sourceJob Fairs        -4.014e-01  4.421e-01  -0.908 0.363968
## hiring_sourceSocial Media     -2.775e-01  4.556e-01  -0.609 0.542429
## hiring_sourceWalk-In          -2.917e-01  4.473e-01  -0.652 0.514253
## no_previous_companies_worked  -1.855e-02  4.014e-02  -0.462 0.643973
## distance_from_home             2.078e-01  1.841e-02  11.286  < 2e-16 ***
## total_dependents               7.302e-01  8.642e-02   8.450  < 2e-16 ***
## marital_statusSingle           1.786e+00  4.112e-01   4.344 1.40e-05 ***
## educationMasters               1.450e+00  4.372e-01   3.316 0.000914 ***
## promotion_last_2_yearsYes     -1.643e+01  5.128e+02  -0.032 0.974441
## no_leaves_taken                1.089e-01  1.588e-02   6.858 6.98e-12 ***
## total_experience              -2.852e-02  5.350e-02  -0.533 0.594040
## monthly_overtime_hrs           2.232e-01  3.282e-02   6.802 1.03e-11 ***
## mgr_effectiveness             -8.308e+00  1.072e+00  -7.750 9.21e-15 ***
## career_satisfaction            4.492e+00  1.175e+00   3.822 0.000132 ***
## perf_satisfaction              9.963e-02  1.066e+00   0.093 0.925565
## work_satisfaction             -6.256e-02  1.208e+00  -0.052 0.958709
## age_diff                       7.161e-02  2.853e-02   2.510 0.012079 *
## compa_ratio                   -6.451e+00  2.106e+00  -3.063 0.002194 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1972.64  on 1953  degrees of freedom
## Residual deviance:  550.16  on 1915  degrees of freedom
## AIC: 628.16
##
## Number of Fisher Scoring iterations: 17
```

```
# Check for multicollinearity
vif(multi_log)
```

```
##                                      GVIF Df GVIF^(1/(2*Df))
## location                    1.887326e+00  2        1.172092
## level                       2.899331e+06  1     1702.742150
## gender                      1.184032e+00  1        1.088132
## rating                      3.448039e+00  4        1.167338
## mgr_rating                  1.919310e+00  4        1.084909
## mgr_reportees               1.240958e+00  1        1.113983
## mgr_tenure                  1.239445e+00  1        1.113304
## compensation                3.772625e+01  1        6.142169
## percent_hike                2.934121e+00  1        1.712928
## hiring_score                1.101045e+00  1        1.049307
## hiring_source               1.458813e+00  6        1.031969
## no_previous_companies_worked 1.070218e+00  1        1.034514
## distance_from_home          1.242416e+00  1        1.114637
## total_dependents            1.805121e+00  1        1.343548
## marital_status              2.026817e+00  1        1.423663
## education                   1.190016e+00  1        1.090878
## promotion_last_2_years      2.899317e+06  1     1702.738057
## no_leaves_taken             1.140525e+00  1        1.067954
## total_experience            2.143385e+00  1        1.464030
## monthly_overtime_hrs        1.212496e+00  1        1.101134
## mgr_effectiveness           2.591719e+00  1        1.609882
## career_satisfaction         2.695242e+00  1        1.641719
## perf_satisfaction           2.843518e+00  1        1.686273
## work_satisfaction           1.638172e+00  1        1.279911
## age_diff                    1.830167e+00  1        1.352836
## compa_ratio                 2.308654e+01  1        4.804845
```

```
# Which variable has the highest VIF?
highest <- "level"

# Taking level out of the model
model_1 <- glm(
  # Manually list variables, *omitting* 'compa_level' and 'job_hop_index'
  turnover ~ location  + gender + rating + mgr_rating + mgr_reportees +
    mgr_tenure + percent_hike + hiring_score + hiring_source +
    no_previous_companies_worked + distance_from_home + total_dependents +
    marital_status + education + promotion_last_2_years + no_leaves_taken +
    total_experience + monthly_overtime_hrs + mgr_effectiveness +
    career_satisfaction + perf_satisfaction + work_satisfaction +
    age_diff + compa_ratio ,
  family = "binomial",
  data = train_set_multi,
  na.action = na.omit
)
# Check for multicollinearity again
vif(model_1)
```

```
##                                 GVIF Df GVIF^(1/(2*Df))
## location                    1.887871  2        1.172177
## gender                      1.181810  1        1.087111
## rating                      3.404203  4        1.165472
## mgr_rating                  1.869051  4        1.081316
## mgr_reportees               1.249027  1        1.117599
## mgr_tenure                  1.245626  1        1.116076
## percent_hike                2.967240  1        1.722568
## hiring_score                1.104158  1        1.050789
## hiring_source               1.423408  6        1.029858
## no_previous_companies_worked 1.065198  1        1.032084
## distance_from_home          1.226450  1        1.107452
## total_dependents            1.844494  1        1.358121
## marital_status              2.019323  1        1.421029
## education                   1.194325  1        1.092852
## promotion_last_2_years      1.135748  1        1.065715
## no_leaves_taken             1.130859  1        1.063419
## total_experience            2.049912  1        1.431752
## monthly_overtime_hrs        1.201337  1        1.096055
## mgr_effectiveness           2.614241  1        1.616861
## career_satisfaction         2.742578  1        1.656073
## perf_satisfaction           2.866921  1        1.693198
## work_satisfaction           1.627506  1        1.275738
## age_diff                    1.787225  1        1.336871
## compa_ratio                 1.551656  1        1.245655
```

```
# Which variable has the highest VIF?
highest <- "compensation"

#Taking Compensation out
model_2 <-  glm(
  # Manually list variables, *omitting* 'compa_level' and 'job_hop_index'
  turnover ~ location  + gender + rating + mgr_rating + mgr_reportees +
    mgr_tenure + percent_hike + hiring_score + hiring_source +
    no_previous_companies_worked + distance_from_home + total_dependents +
    marital_status + education + promotion_last_2_years + no_leaves_taken +
    total_experience + monthly_overtime_hrs + mgr_effectiveness +
    career_satisfaction + perf_satisfaction + work_satisfaction +
    age_diff + compa_ratio ,
  family = "binomial",
  data = train_set_multi,
  na.action = na.omit
)

# Check for multicollinearity again to see if we dealt with multicolinearity
vif(model_2)
```

```
##                                 GVIF Df GVIF^(1/(2*Df))
## location                    1.887871  2        1.172177
## gender                      1.181810  1        1.087111
## rating                      3.404203  4        1.165472
## mgr_rating                  1.869051  4        1.081316
## mgr_reportees               1.249027  1        1.117599
## mgr_tenure                  1.245626  1        1.116076
## percent_hike                2.967240  1        1.722568
## hiring_score                1.104158  1        1.050789
## hiring_source               1.423408  6        1.029858
## no_previous_companies_worked 1.065198  1        1.032084
## distance_from_home          1.226450  1        1.107452
## total_dependents            1.844494  1        1.358121
## marital_status              2.019323  1        1.421029
## education                   1.194325  1        1.092852
## promotion_last_2_years      1.135748  1        1.065715
## no_leaves_taken             1.130859  1        1.063419
## total_experience            2.049912  1        1.431752
## monthly_overtime_hrs        1.201337  1        1.096055
## mgr_effectiveness           2.614241  1        1.616861
## career_satisfaction         2.742578  1        1.656073
## perf_satisfaction           2.866921  1        1.693198
## work_satisfaction           1.627506  1        1.275738
## age_diff                    1.787225  1        1.336871
## compa_ratio                 1.551656  1        1.245655
```

```
# Check for multicollinearity again to see if we dealt with multicolinearity
vif(model_2)
```


```
##                                 GVIF Df GVIF^(1/(2*Df))
## location                    1.887871  2        1.172177
## gender                      1.181810  1        1.087111
## rating                      3.404203  4        1.165472
## mgr_rating                  1.869051  4        1.081316
## mgr_reportees               1.249027  1        1.117599
## mgr_tenure                  1.245626  1        1.116076
## percent_hike                2.967240  1        1.722568
## hiring_score                1.104158  1        1.050789
## hiring_source               1.423408  6        1.029858
## no_previous_companies_worked 1.065198  1        1.032084
## distance_from_home          1.226450  1        1.107452
## total_dependents            1.844494  1        1.358121
## marital_status              2.019323  1        1.421029
## education                   1.194325  1        1.092852
## promotion_last_2_years      1.135748  1        1.065715
## no_leaves_taken             1.130859  1        1.063419
## total_experience            2.049912  1        1.431752
## monthly_overtime_hrs        1.201337  1        1.096055
## mgr_effectiveness           2.614241  1        1.616861
## career_satisfaction         2.742578  1        1.656073
## perf_satisfaction           2.866921  1        1.693198
## work_satisfaction           1.627506  1        1.275738
## age_diff                    1.787225  1        1.336871
## compa_ratio                 1.551656  1        1.245655
```

```
# Build the final logistic regression model
final_log <- glm(turnover ~location  + gender + rating + mgr_rating + mgr_reportees +
                   mgr_tenure + percent_hike + hiring_score + hiring_source +
                   no_previous_companies_worked + distance_from_home + total_dependents +
                   marital_status + education + promotion_last_2_years + no_leaves_taken +
                   total_experience + monthly_overtime_hrs + mgr_effectiveness +
                   career_satisfaction + perf_satisfaction + work_satisfaction +
                   age_diff + compa_ratio+ job_hop_index,
                 family = "binomial",
                 data = train_set_multi)

# Print summary
summary(final_log )
```

```
##
## Call:
## glm(formula = turnover ~ location + gender + rating + mgr_rating +
##     mgr_reportees + mgr_tenure + percent_hike + hiring_score +
##     hiring_source + no_previous_companies_worked + distance_from_home +
##     total_dependents + marital_status + education + promotion_last_2_years +
##     no_leaves_taken + total_experience + monthly_overtime_hrs +
##     mgr_effectiveness + career_satisfaction + perf_satisfaction +
##     work_satisfaction + age_diff + compa_ratio + job_hop_index,
##     family = "binomial", data = train_set_multi)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -10.07634    3.17964  -3.169 0.001530 **
## locationNew York                1.00660    0.36420   2.764 0.005712 **
## locationOrlando                -1.11615    0.30515  -3.658 0.000254 ***
## genderMale                      0.38843    0.26390   1.472 0.141055
## ratingAcceptable               -0.37748    0.29764  -1.268 0.204707
## ratingBelow Average            -2.35675    0.56576  -4.166 3.10e-05 ***
## ratingExcellent                -0.85783    0.80773  -1.062 0.288225
## ratingUnacceptable             -3.66539    0.95232  -3.849 0.000119 ***
## mgr_ratingAcceptable            0.07275    0.28303   0.257 0.797151
## mgr_ratingBelow Average        -0.79244    0.51869  -1.528 0.126570
## mgr_ratingExcellent             0.03224    0.39932   0.081 0.935651
## mgr_ratingUnacceptable          1.06152    1.04248   1.018 0.308552
## mgr_reportees                   0.08588    0.02472   3.475 0.000511 ***
## mgr_tenure                     -0.07924    0.03458  -2.292 0.021920 *
## percent_hike                   -0.54459    0.06331  -8.602  < 2e-16 ***
## hiring_score                    0.07695    0.03736   2.059 0.039453 *
## hiring_sourceConsultant        -0.61317    0.44335  -1.383 0.166657
## hiring_sourceEmployee Referral  -0.06618    0.47067  -0.141 0.888185
## hiring_sourceJob Boards        -0.40841    0.45394  -0.900 0.368273
## hiring_sourceJob Fairs         -0.28860    0.45524  -0.634 0.526106
## hiring_sourceSocial Media      -0.27320    0.47429  -0.576 0.564608
## hiring_sourceWalk-In           -0.60182    0.47145  -1.277 0.201771
## no_previous_companies_worked   -0.07310    0.07022  -1.041 0.297894
## distance_from_home              0.21028    0.01929  10.901  < 2e-16 ***
## total_dependents                0.68860    0.08762   7.859 3.87e-15 ***
## marital_statusSingle            1.50268    0.42619   3.526 0.000422 ***
## educationMasters                1.47863    0.45309   3.263 0.001101 **
## promotion_last_2_yearsYes      -0.51170    0.31974  -1.600 0.109515
## no_leaves_taken                 0.10297    0.01615   6.377 1.80e-10 ***
## total_experience               -0.02237    0.06238  -0.359 0.719910
## monthly_overtime_hrs            0.21717    0.03343   6.497 8.21e-11 ***
## mgr_effectiveness              -8.33452    1.10710  -7.528 5.14e-14 ***
## career_satisfaction             4.38810    1.24565   3.523 0.000427 ***
## perf_satisfaction               0.59388    1.07285   0.554 0.579882
## work_satisfaction              -0.14881    1.27436  -0.117 0.907043
## age_diff                        0.05620    0.02914   1.929 0.053738 .
## compa_ratio                    -1.31114    0.56371  -2.326 0.020023 *
## job_hop_index                  -0.06145    0.07975  -0.771 0.440936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1787.02  on 1767  degrees of freedom
## Residual deviance:  526.25  on 1730  degrees of freedom
##   (186 observations deleted due to missingness)
## AIC: 602.25
##
## Number of Fisher Scoring iterations: 7
```

8. Model Evaluation

```
pred_test <- predict(final_log, newdata = test_set, type = "response")
pred_class <- ifelse(pred_test > 0.5, 1, 0)
confusionMatrix(table(pred_class, test_set$turnover))
```

```
## Confusion Matrix and Statistics
##
##
## pred_class   0   1
##          0 411  22
##          1  10  90
##
##                Accuracy : 0.94
##                  95% CI : (0.9163, 0.9586)
##     No Information Rate : 0.7899
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8117
##
##  Mcnemar's Test P-Value : 0.05183
##
##             Sensitivity : 0.9762
##             Specificity : 0.8036
##          Pos Pred Value : 0.9492
##          Neg Pred Value : 0.9000
##              Prevalence : 0.7899
##          Detection Rate : 0.7711
##    Detection Prevalence : 0.8124
##       Balanced Accuracy : 0.8899
##
##        'Positive' Class : 0
##
```

9. Employee Risk Scoring

```
emp_risk <- emp_compa %>%
  filter(status == "Active") %>%
  tidypredict_to_column(final_log)

emp_risk %>%
  select(emp_id, fit) %>%
  slice_max(fit, n = 5)
```

```
## # A tibble: 5 × 2
##   emp_id   fit
##   <chr>  <dbl>
## 1 E13342 0.911
## 2 E9878  0.907
## 3 E6037  0.851
## 4 E1236  0.846
## 5 E6574  0.845
```

10. Business Impact & ROI

```
median_salary_analyst <- 51840
turnover_cost <- 40000
ROI <- ((turnover_cost * 0.17) / (median_salary_analyst * 0.05)) * 100
cat(paste0("The estimated return on investment is ", round(ROI), "%"))
```

```
## The estimated return on investment is 262%
```