Test-1.R

lucasvalpreda

2025-08-04

```r
# Project: Employee Attrition Analysis
# Purpose: Explore and model employee attrition patterns
# Author: Lucas Valpreda
# Date: 04/08/2025

# Load packages
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.2     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.1.0
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors

library(ggplot2)
library(dplyr)
library(caret)

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift

library(corrplot)

## corrplot 0.95 loaded

library(readr)
# 1. Getting to Know the Data
df <- read_csv("Termination_Data.csv")

## Rows: 49653 Columns: 18
## ── Column specification
```

```
────────────────────────────────────────────────────────────
## Delimiter: ","
## chr (13): recorddate_key, birthdate_key, orighiredate_key,
terminationdate_k...
## dbl  (5): EmployeeID, age, length_of_service, store_name, STATUS_YEAR
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```r
head(df)
```

```
## # A tibble: 6 × 18
##   EmployeeID recorddate_key  birthdate_key orighiredate_key
terminationdate_key
##        <dbl> <chr>           <chr>         <chr>            <chr>
## 1       1318 12/31/2006 0:00 1/3/1954      8/28/1989        1/1/1900
## 2       1318 12/31/2007 0:00 1/3/1954      8/28/1989        1/1/1900
## 3       1318 12/31/2008 0:00 1/3/1954      8/28/1989        1/1/1900
## 4       1318 12/31/2009 0:00 1/3/1954      8/28/1989        1/1/1900
## 5       1318 12/31/2010 0:00 1/3/1954      8/28/1989        1/1/1900
## 6       1318 12/31/2011 0:00 1/3/1954      8/28/1989        1/1/1900
## # ℹ 13 more variables: age <dbl>, length_of_service <dbl>, city_name
<chr>,
## #   department_name <chr>, job_title <chr>, store_name <dbl>,
## #   gender_short <chr>, gender_full <chr>, termreason_desc <chr>,
## #   termtype_desc <chr>, STATUS_YEAR <dbl>, STATUS <chr>, BUSINESS_UNIT
<chr>
```

```r
str(df)
```

```
## spc_tbl_ [49,653 × 18] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ EmployeeID         : num [1:49653] 1318 1318 1318 1318 1318 ...
##  $ recorddate_key     : chr [1:49653] "12/31/2006 0:00" "12/31/2007 0:00"
"12/31/2008 0:00" "12/31/2009 0:00" ...
##  $ birthdate_key      : chr [1:49653] "1/3/1954" "1/3/1954" "1/3/1954"
"1/3/1954" ...
##  $ orighiredate_key   : chr [1:49653] "8/28/1989" "8/28/1989" "8/28/1989"
"8/28/1989" ...
##  $ terminationdate_key: chr [1:49653] "1/1/1900" "1/1/1900" "1/1/1900"
"1/1/1900" ...
##  $ age                : num [1:49653] 52 53 54 55 56 57 58 59 60 61 ...
##  $ length_of_service  : num [1:49653] 17 18 19 20 21 22 23 24 25 26 ...
##  $ city_name          : chr [1:49653] "Vancouver" "Vancouver" "Vancouver"
"Vancouver" ...
##  $ department_name    : chr [1:49653] "Executive" "Executive" "Executive"
"Executive" ...
##  $ job_title          : chr [1:49653] "CEO" "CEO" "CEO" "CEO" ...
##  $ store_name         : num [1:49653] 35 35 35 35 35 35 35 35 35 35 ...
##  $ gender_short       : chr [1:49653] "M" "M" "M" "M" ...
```

```
##  $ gender_full        : chr [1:49653] "Male" "Male" "Male" "Male" ...
##  $ termreason_desc    : chr [1:49653] "Not Applicable" "Not Applicable"
"Not Applicable" "Not Applicable" ...
##  $ termtype_desc      : chr [1:49653] "Not Applicable" "Not Applicable"
"Not Applicable" "Not Applicable" ...
##  $ STATUS_YEAR        : num [1:49653] 2006 2007 2008 2009 2010 ...
##  $ STATUS             : chr [1:49653] "ACTIVE" "ACTIVE" "ACTIVE" "ACTIVE"
...
##  $ BUSINESS_UNIT      : chr [1:49653] "HEADOFFICE" "HEADOFFICE"
"HEADOFFICE" "HEADOFFICE" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   EmployeeID = col_double(),
##   ..   recorddate_key = col_character(),
##   ..   birthdate_key = col_character(),
##   ..   orighiredate_key = col_character(),
##   ..   terminationdate_key = col_character(),
##   ..   age = col_double(),
##   ..   length_of_service = col_double(),
##   ..   city_name = col_character(),
##   ..   department_name = col_character(),
##   ..   job_title = col_character(),
##   ..   store_name = col_double(),
##   ..   gender_short = col_character(),
##   ..   gender_full = col_character(),
##   ..   termreason_desc = col_character(),
##   ..   termtype_desc = col_character(),
##   ..   STATUS_YEAR = col_double(),
##   ..   STATUS = col_character(),
##   ..   BUSINESS_UNIT = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
summary(df)
```

```
##    EmployeeID   recorddate_key     birthdate_key      orighiredate_key
##  Min.   :1318   Length:49653       Length:49653       Length:49653
##  1st Qu.:3360   Class :character   Class :character   Class :character
##  Median :5031   Mode  :character   Mode  :character   Mode  :character
##  Mean   :4859
##  3rd Qu.:6335
##  Max.   :8336
##  terminationdate_key      age        length_of_service  city_name
##  Length:49653        Min.   :19.00   Min.   : 0.00     Length:49653
##  Class :character    1st Qu.:31.00   1st Qu.: 5.00     Class :character
##  Mode  :character    Median :42.00   Median :10.00     Mode  :character
##                      Mean   :42.08   Mean   :10.43
##                      3rd Qu.:53.00   3rd Qu.:15.00
##                      Max.   :65.00   Max.   :26.00
##  department_name     job_title           store_name   gender_short
```

```
##  Length:49653       Length:49653       Min.   : 1.0   Length:49653
##  Class :character   Class :character   1st Qu.:16.0   Class :character
##  Mode  :character   Mode  :character   Median :28.0   Mode  :character
##                                        Mean   :27.3
##                                        3rd Qu.:42.0
##                                        Max.   :46.0
##  gender_full        termreason_desc    termtype_desc       STATUS_YEAR
##  Length:49653       Length:49653       Length:49653       Min.   :2006
##  Class :character   Class :character   Class :character   1st Qu.:2008
##  Mode  :character   Mode  :character   Mode  :character   Median :2011
##                                                           Mean   :2011
##                                                           3rd Qu.:2013
##                                                           Max.   :2015
##     STATUS          BUSINESS_UNIT
##  Length:49653       Length:49653
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```r
names(df)
```

```
##  [1] "EmployeeID"          "recorddate_key"      "birthdate_key"
##  [4] "orighiredate_key"    "terminationdate_key" "age"
##  [7] "length_of_service"   "city_name"           "department_name"
## [10] "job_title"           "store_name"          "gender_short"
## [13] "gender_full"         "termreason_desc"     "termtype_desc"
## [16] "STATUS_YEAR"         "STATUS"              "BUSINESS_UNIT"
```

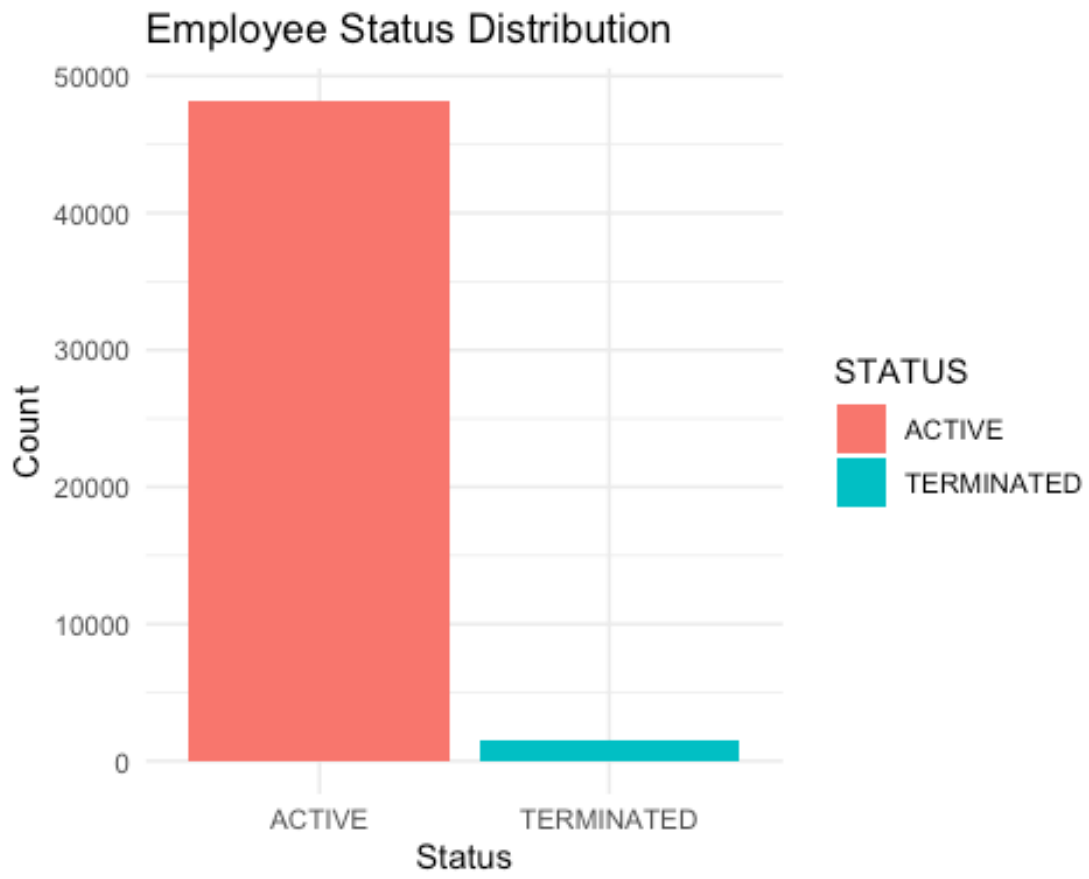
```r
# 2. Attrition status breakdown
table(df$STATUS)
```

```
##
##     ACTIVE TERMINATED
##      48168       1485
```

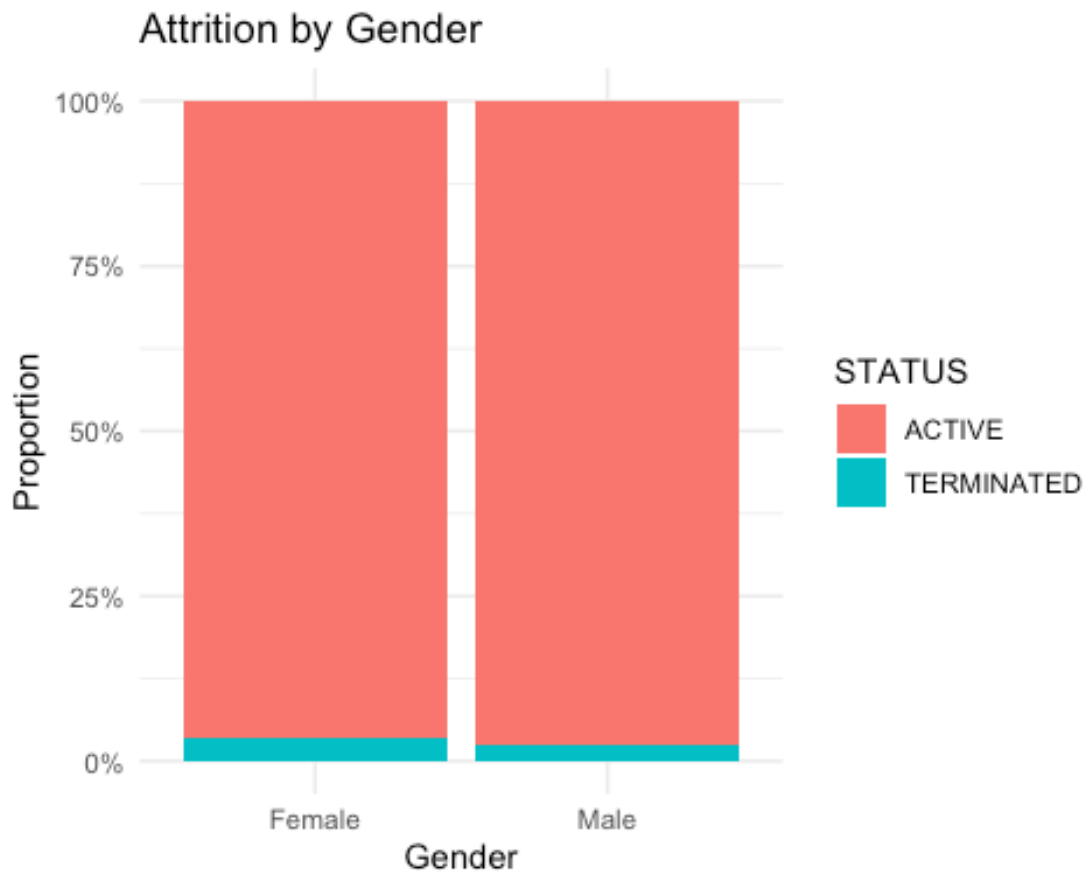
```r
round(prop.table(table(df$STATUS)) * 100, 1)
```

```
##
##     ACTIVE TERMINATED
##         97          3
```

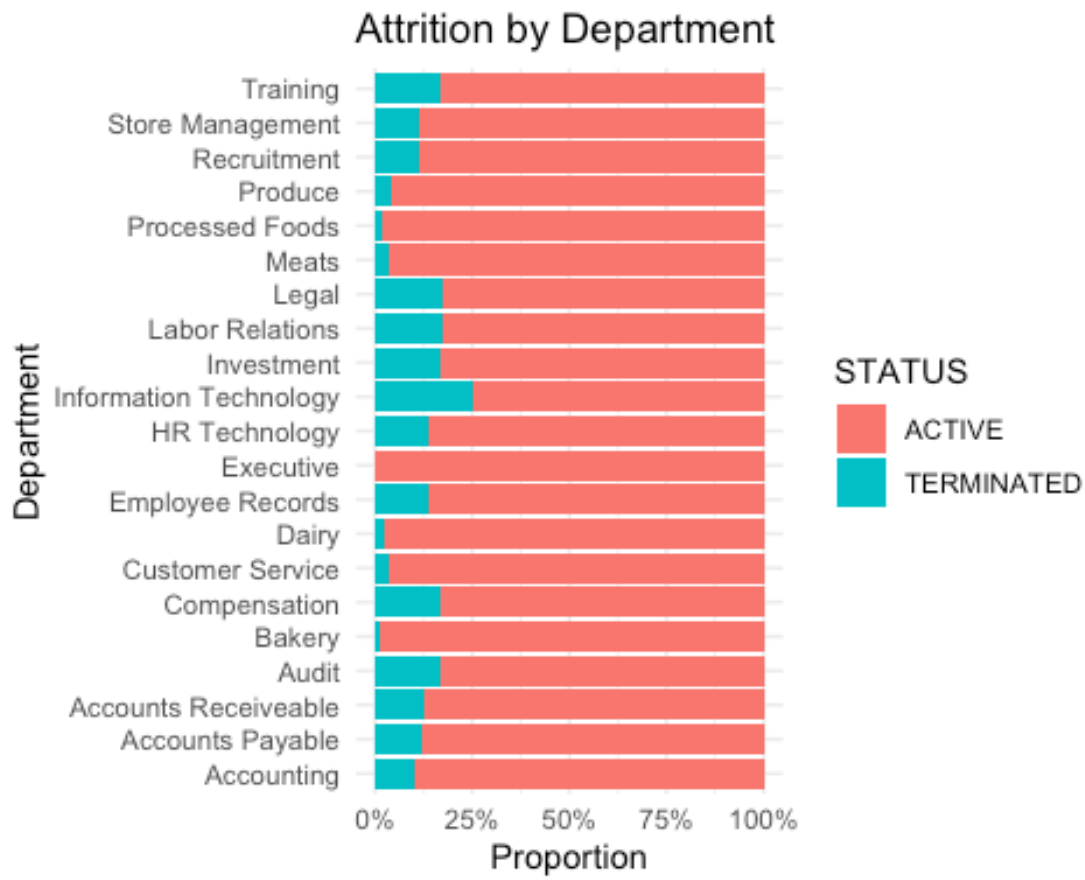
```r
library(ggplot2)
ggplot(df, aes(x = STATUS, fill = STATUS)) +
  geom_bar() +
  labs(title = "Employee Status Distribution", x = "Status", y = "Count") +
  theme_minimal()
```



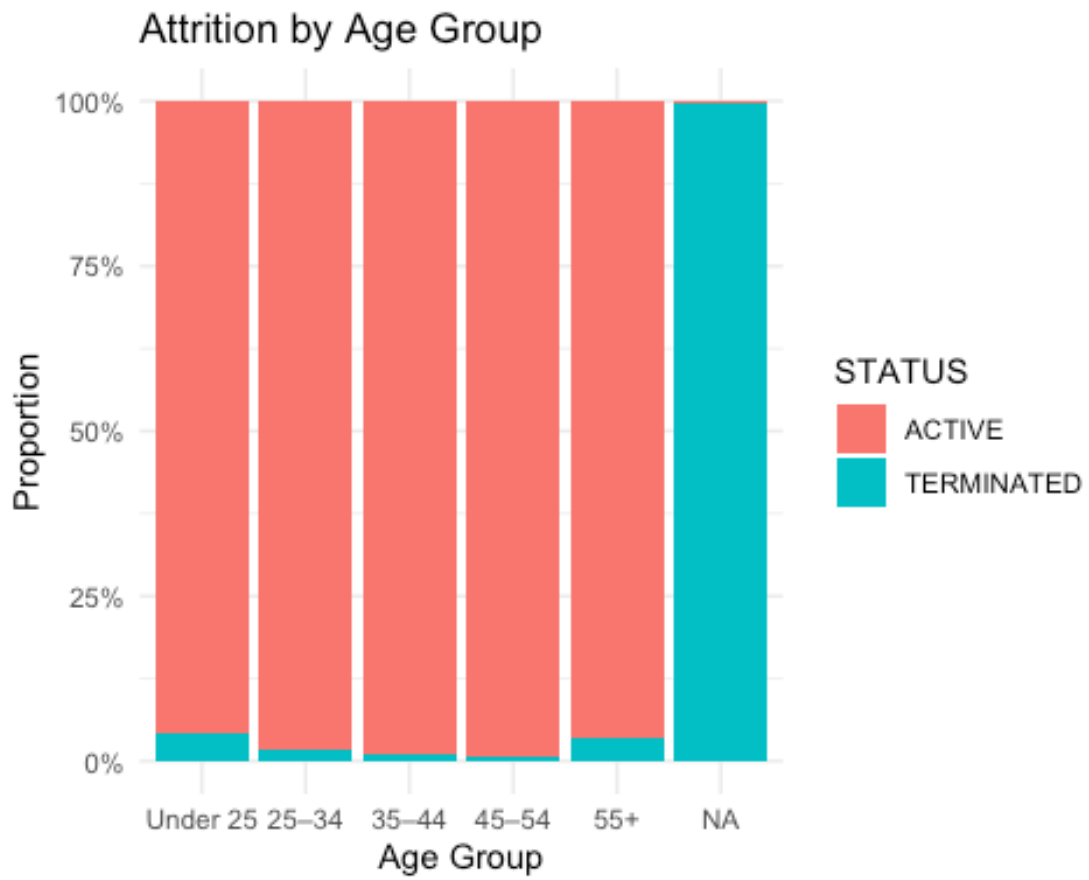

```r
ggplot(df, aes(x = gender_full, fill = STATUS)) +
  geom_bar(position = "fill") +
  labs(title = "Attrition by Gender", x = "Gender", y = "Proportion") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()
```



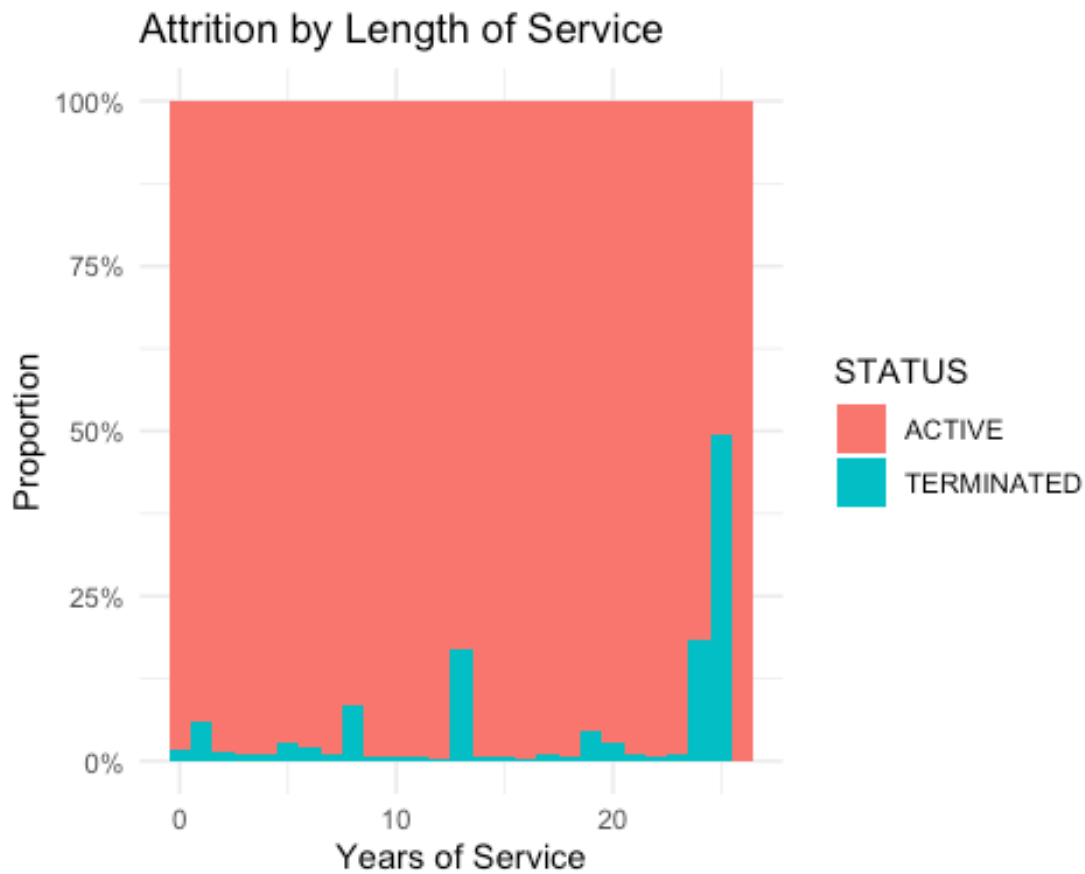

```
ggplot(df, aes(x = department_name, fill = STATUS)) +
  geom_bar(position = "fill") +
  coord_flip() +
  labs(title = "Attrition by Department", x = "Department", y = "Proportion")
+
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()
```



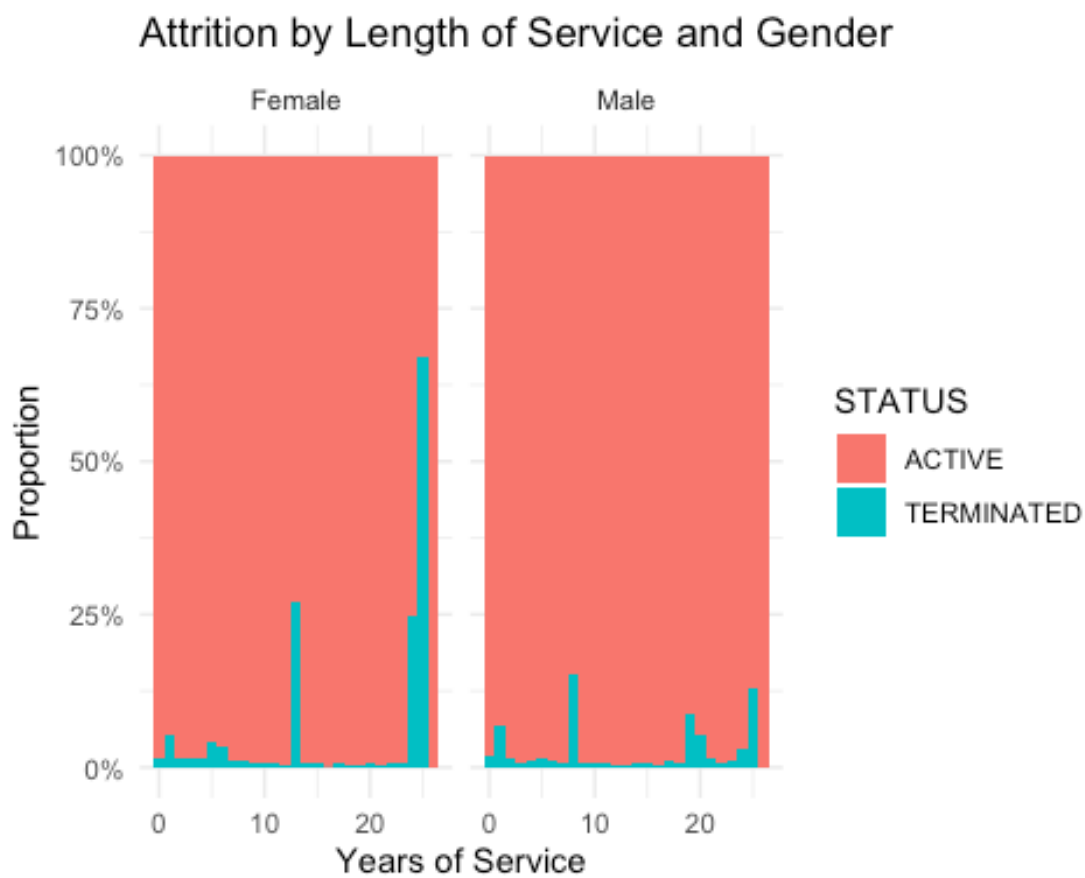

```r
df$age_group <- cut(df$age,
                    breaks = c(0, 25, 35, 45, 55, 65),
                    labels = c("Under 25", "25–34", "35–44", "45–54", "55+"),
                    right = FALSE)

ggplot(df, aes(x = age_group, fill = STATUS)) +
  geom_bar(position = "fill") +
  labs(title = "Attrition by Age Group", x = "Age Group", y = "Proportion") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()
```




```r
ggplot(df, aes(x = length_of_service, fill = STATUS)) +
  geom_histogram(binwidth = 1, position = "fill") +
  labs(title = "Attrition by Length of Service", x = "Years of Service", y =
"Proportion") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()
```




```r
ggplot(df, aes(x = length_of_service, fill = STATUS)) +
  geom_histogram(binwidth = 1, position = "fill") +
  facet_wrap(~ gender_full) +
  labs(title = "Attrition by Length of Service and Gender",
       x = "Years of Service",
       y = "Proportion") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()
```




```
df$STATUS_BINARY <- ifelse(df$STATUS == "TERMINATED", 1, 0)
model <- glm(STATUS_BINARY ~ gender_full + length_of_service +
department_name,
             data = df,
             family = binomial)
summary(model)

##
## Call:
## glm(formula = STATUS_BINARY ~ gender_full + length_of_service +
##     department_name, family = binomial, data = df)
##
## Coefficients:
##                                          Estimate Std. Error z value
Pr(>|z|)
## (Intercept)                             -2.049593   0.444318  -4.613 3.97e-
06
## gender_fullMale                         -0.350869   0.054872  -6.394 1.61e-
10
## length_of_service                       -0.006278   0.005311  -1.182
0.237241
## department_nameAccounts Payable          0.207548   0.685030   0.303
0.761907
```

```
## department_nameAccounts Receiveable     0.339455   0.644654   0.527
0.598493
## department_nameAudit                    0.754880   0.698296   1.081
0.279683
## department_nameBakery                  -2.032075   0.445091  -4.566 4.98e-
06
## department_nameCompensation             0.754645   0.698296   1.081
0.279833
## department_nameCustomer Service        -1.015467   0.442760  -2.293
0.021820
## department_nameDairy                   -1.521287   0.443055  -3.434
0.000596
## department_nameEmployee Records         0.385478   0.615571   0.626
0.531176
## department_nameExecutive              -13.217524 145.248085  -0.091
0.927493
## department_nameHR Technology            0.412826   0.561331   0.735
0.462071
## department_nameInformation Technology   1.416410   0.674796   2.099
0.035815
## department_nameInvestment               0.755407   0.698292   1.082
0.279345
## department_nameLabor Relations          0.783830   0.623956   1.256
0.209034
## department_nameLegal                    0.788564   0.769711   1.024
0.305602
## department_nameMeats                   -0.982074   0.435187  -2.257
0.024029
## department_nameProcessed Foods         -1.912339   0.450482  -4.245 2.18e-
05
## department_nameProduce                 -0.867765   0.436717  -1.987
0.046920
## department_nameRecruitment              0.134796   0.571241   0.236
0.813455
## department_nameStore Management         0.249511   0.471559   0.529
0.596722
## department_nameTraining                 0.696076   0.653273   1.066
0.286640
##
## (Intercept)                            ***
## gender_fullMale                        ***
## length_of_service
## department_nameAccounts Payable
## department_nameAccounts Receiveable
## department_nameAudit
## department_nameBakery                  ***
## department_nameCompensation
## department_nameCustomer Service        *
## department_nameDairy                   ***
## department_nameEmployee Records
```

```
## department_nameExecutive
## department_nameHR Technology
## department_nameInformation Technology *
## department_nameInvestment
## department_nameLabor Relations
## department_nameLegal
## department_nameMeats                  *
## department_nameProcessed Foods        ***
## department_nameProduce                *
## department_nameRecruitment
## department_nameStore Management
## department_nameTraining
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13349  on 49652  degrees of freedom
## Residual deviance: 12910  on 49630  degrees of freedom
## AIC: 12956
##
## Number of Fisher Scoring iterations: 14
```

```r
exp(coef(model))
```

```
##                           (Intercept)
gender_fullMale
##                          1.287872e-01                          7.040760e-
01
##                      length_of_service       department_nameAccounts
Payable
##                          9.937421e-01
1.230657e+00
##   department_nameAccounts Receiveable
department_nameAudit
##                          1.404182e+00
2.127357e+00
##                  department_nameBakery
department_nameCompensation
##                          1.310632e-01
2.126856e+00
##       department_nameCustomer Service
department_nameDairy
##                          3.622332e-01                          2.184306e-
01
##       department_nameEmployee Records
department_nameExecutive
##                          1.470317e+00                          1.818454e-
06
##          department_nameHR Technology department_nameInformation
```

```
Technology
##                          1.511082e+00
4.122293e+00
##              department_nameInvestment        department_nameLabor
Relations
##                          2.128478e+00
2.189843e+00
##                   department_nameLegal
department_nameMeats
##                          2.200236e+00                          3.745337e-
01
##         department_nameProcessed Foods
department_nameProduce
##                          1.477344e-01                          4.198891e-
01
##             department_nameRecruitment       department_nameStore
Management
##                          1.144304e+00
1.283398e+00
##                department_nameTraining
##                          2.005867e+00
```

```r
# Get predicted probabilities
df$predicted_prob <- predict(model, type = "response")
library(caret)
df$predicted_prob <- predict(model, type = "response")
df$predicted_class <- ifelse(df$predicted_prob >= 0.5, 1, 0)
df$actual <- as.factor(df$STATUS_BINARY)
df$predicted <- as.factor(df$predicted_class)
confusionMatrix(data = df$predicted, reference = df$actual, positive = "1")
```

```
## Warning in confusionMatrix.default(data = df$predicted, reference =
df$actual,
## : Levels are not in the same order for reference and data. Refactoring
data to
## match.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 48168  1485
##          1     0     0
##
##                Accuracy : 0.9701
##                  95% CI : (0.9686, 0.9716)
##     No Information Rate : 0.9701
##     P-Value [Acc > NIR] : 0.5069
##
##                   Kappa : 0
```

```
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.97009
##              Prevalence : 0.02991
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : 1
##

df$STATUS_BINARY <- as.factor(df$STATUS_BINARY)
rf_df <- df[, c("STATUS_BINARY", "gender_full", "length_of_service",
"department_name")]
set.seed(123)
```