Employee Churn

Code

Hide

```
#loading the required libraries
library(mlr)
library(survival)
library(pec)
library(survAUC)
library(dplyr)
library(reshape2)
library(ggplot2)
library(plyr)
library(reshape2)
library(plotly)
library(corrplot)
library(ggcorrplot)
library(randomForestSRC)
```

step 1: Loading the data

Hide

```
Employees=read.csv("turnover.csv")
Employeess = Employees
Employees
```

stag	event	gender	age	industry	profession	traffic	coach
<dbl>	<int>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>
7.0308008	1	m	35.00000	Banks	HR	rabrecNErab	no
22.9650924	1	m	33.00000	Banks	HR	empjs	no
15.9342916	1	f	35.00000	PowerGeneration	HR	rabrecNErab	no
15.9342916	1	f	35.00000	PowerGeneration	HR	rabrecNErab	no
8.4106776	1	m	32.00000	Retail	Commercial	youjs	yes
8.9691992	1	f	42.00000	manufacture	HR	empjs	yes
8.9691992	1	f	42.00000	manufacture	HR	empjs	yes
120.4435318	1	f	28.00000	Retail	HR	referal	no
8.6078029	1	f	29.00000	Banks	HR	empjs	no
4.4353183	1	f	30.00000	Consult	Marketing	youjs	yes

1-10 of 1,129 rows | 1-8 of 16 columns Previous **1** 2 3 4 5 6 ... 100 Next

step 2: Data manipulation and Exploratory Data Analysis

Hide

```
#removing null values
Employees=na.omit(Employees)
```

Hide

```
# Summary statistics
summary(Employees)
```

```
     stag              event           gender               age          industry          profession          t
raffic
 Min.   :  0.3942   Min.   :0.0000   Length:1129        Min.   :18.00   Length:1129        Length:1129        Len
gth:1129
 1st Qu.: 11.7289   1st Qu.:0.0000   Class :character   1st Qu.:26.00   Class :character   Class :character   Cla
ss :character
 Median : 24.3450   Median :1.0000   Mode  :character   Median :30.00   Mode  :character   Mode  :character   Mod
e  :character
 Mean   : 36.6275   Mean   :0.5058                      Mean   :31.07
 3rd Qu.: 51.3183   3rd Qu.:1.0000                      3rd Qu.:36.00
 Max.   :179.4497   Max.   :1.0000                      Max.   :58.00
    coach           head_gender          greywage             way             extraversion       independ        s
elfcontrol
 Length:1129        Length:1129        Length:1129        Length:1129        Min.   : 1.000   Min.   : 1.000   Mi
n.   : 1.000
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 4.600   1st Qu.: 4.100   1s
t Qu.: 4.100
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 5.400   Median : 5.500   Me
dian : 5.700
                                                                              Mean   : 5.592   Mean   : 5.478   Me
an   : 5.597
                                                                              3rd Qu.: 7.000   3rd Qu.: 6.900   3r
d Qu.: 7.200
                                                                              Max.   :10.000   Max.   :10.000   Ma
x.   :10.000
    anxiety          novator
 Min.   : 1.700   Min.   : 1.00
 1st Qu.: 4.800   1st Qu.: 4.40
 Median : 5.600   Median : 6.00
 Mean   : 5.666   Mean   : 5.88
 3rd Qu.: 7.100   3rd Qu.: 7.50
 Max.   :10.000   Max.   :10.00
```

Hide

```
# Structure of the data
str(Employees)
```

```
'data.frame':   1129 obs. of  16 variables:
 $ stag        : num  7.03 22.97 15.93 15.93 8.41 ...
 $ event       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ gender      : chr  "m" "m" "f" "f" ...
 $ age         : num  35 33 35 35 32 42 42 28 29 30 ...
 $ industry    : chr  "Banks" "Banks" "PowerGeneration" "PowerGeneration" ...
 $ profession  : chr  "HR" "HR" "HR" "HR" ...
 $ traffic     : chr  "rabrecNErab" "empjs" "rabrecNErab" "rabrecNErab" ...
 $ coach       : chr  "no" "no" "no" "no" ...
 $ head_gender : chr  "f" "m" "m" "m" ...
 $ greywage    : chr  "white" "white" "white" "white" ...
 $ way         : chr  "bus" "bus" "bus" "bus" ...
 $ extraversion: num  6.2 6.2 6.2 5.4 3 6.2 6.2 3.8 8.6 5.4 ...
 $ independ    : num  4.1 4.1 6.2 7.6 4.1 6.2 6.2 5.5 6.9 5.5 ...
 $ selfcontrol : num  5.7 5.7 2.6 4.9 8 4.1 4.1 8 2.6 3.3 ...
 $ anxiety     : num  7.1 7.1 4.8 2.5 7.1 5.6 5.6 4 4 7.9 ...
 $ novator     : num  8.3 8.3 8.3 6.7 3.7 6.7 6.7 4.4 7.5 8.3 ...
```

Hide

```
# converting the data type to int
Employees$age <- as.integer(Employees$age)
head(Employees)
```

	stag <dbl>	event <int>	gender <chr>	a... <int>	industry <chr>	profession <chr>	traffic <chr>	coach <chr>	head_gender <chr>
1	7.030801	1	m	35	Banks	HR	rabrecNErab	no	f
2	22.965092	1	m	33	Banks	HR	empjs	no	m
3	15.934292	1	f	35	PowerGeneration	HR	rabrecNErab	no	m
4	15.934292	1	f	35	PowerGeneration	HR	rabrecNErab	no	m
5	8.410678	1	m	32	Retail	Commercial	youjs	yes	f

6	8.969199	1	f	42	manufacture	HR	empjs	yes	m

6 rows | 1-10 of 16 columns

Hide

```
attach(Employees)
table(gender)
```

```
gender
  f   m 
853 276 
```

Hide

```
table(event)
```

```
event
  0   1 
558 571 
```

Hide

```
table(industry)
```

```
industry
         HoReCa     Agriculture           Banks        Building         Consult             etc              IT 
manufacture 
             11              15             114              41              74              94             122 
145 
         Mining          Pharma PowerGeneration      RealEstate          Retail           State         Telecom 
transport 
             24              20              38              13             289              55              36 
38 
```

Hide

```
table(profession)
```

```
profession
         Accounting BusinessDevelopment          Commercial             Consult            Engineer 
etc             Finance 
                 10                  27                  23                  25                  15 
37                  17 
                 HR                  IT                 Law              manage           Marketing 
PR                Sales 
                757                  74                   7                  22                  31 
6                   66 
           Teaching 
                 12 
```

Hide

```
table(greywage)
```

```
greywage
 grey white 
  127  1002 
```

Hide

```
table(way)
```

```
way
 bus  car foot 
 681  331  117 
```

Hide

```
detach(Employees)
```

Label encoding to change the categorical to numerical to feed into our model.

Hide

```
# Gender: Male/Female
Employees$gender=revalue(Employees$gender,c('m' = 0, 'f' = 1))
Employees$gender=as.numeric((Employees$gender))

# Industry: Describes what industry they belong to
Employees$industry=revalue(Employees$industry,c('Retail'= 10, 'manufacture'= 14, 'IT'= 5, 'Banks'= 2, 'etc'= 13,
'Consult'= 4, 'State'= 11, 'Building'= 3, 'PowerGeneration'= 8, 'transport'= 15, 'Telecom'= 12, 'Mining'= 6, 'Pha
rma'= 7, 'Agriculture'= 1, 'RealEstate'= 9, ' HoReCa'= 0))
Employees$industry=as.numeric((Employees$industry))

# Profession: Describes their respective profession
Employees$profession=revalue(Employees$profession,c('HR'=6, 'IT'= 7, 'Sales'= 11, 'etc'= 13, 'Marketing'= 9, 'Bus
inessDevelopment'= 1, 'Consult'= 3, 'Commercial'= 2, 'manage'= 14, 'Finance'= 5, 'Engineer'= 4, 'Teaching'= 12, '
Accounting'= 0, 'Law'= 8, 'PR'= 10))
Employees$profession=as.numeric((Employees$profession))

# Traffic: Describes what pipeline the employee came into the company
Employees$traffic=revalue(Employees$traffic,c('youjs'= 7, 'empjs'= 2, 'rabrecNErab'= 4, 'friends'= 3, 'referal'=
6, 'KA'= 0, 'recNErab'= 5, 'advert'= 1))
Employees$traffic=as.numeric((Employees$traffic))

# Coach: Describes if they had a coach in their probation period
Employees$coach=revalue(Employees$coach,c('no'= 1, 'my head'= 0, 'yes'= 2))
Employees$coach=as.numeric((Employees$coach))

# Head Gender: Gender of their coach during probation.
Employees$head_gender=revalue(Employees$head_gender,c('m' = 0, 'f' = 1))
Employees$head_gender=as.numeric((Employees$head_gender))

# Grey wage: white - taxed, grey - not taxed
Employees$greywage=revalue(Employees$greywage,c('white'= 1, 'grey'= 0))
Employees$greywage=as.numeric((Employees$greywage))

# Way: Describes the way employee travels to office.
Employees$way=revalue(Employees$way,c(
'bus'= 0, 'car'= 1, 'foot'= 2))
Employees$way=as.numeric((Employees$way))

# Stag: Experience in months, now converted to years
#Employees$stag = Employees$stag/12
```

Hide

```
head(Employees, 10)
```

	stag <dbl>	event <int>	gender <dbl>	age <int>	industry <dbl>	profession <dbl>	traffic <dbl>	coach <dbl>	head_gender <dbl>
1	7.030801	1	0	35	2	6	4	1	1
2	22.965092	1	0	33	2	6	2	1	0
3	15.934292	1	1	35	8	6	4	1	0
4	15.934292	1	1	35	8	6	4	1	0
5	8.410678	1	0	32	10	2	7	2	1
6	8.969199	1	1	42	14	6	2	2	0
7	8.969199	1	1	42	14	6	2	2	0
8	120.443532	1	1	28	10	6	6	1	0
9	8.607803	1	1	29	2	6	2	1	1
10	4.435318	1	1	30	4	9	7	2	0

1-10 of 10 rows | 1-10 of 16 columns

Hide

```
Employees <- Employees %>%
    rename(
        supervisor = coach,
        supervisor_gender = head_gender,
        independence = independ,
        innovator = novator
        )

head(Employees)
```

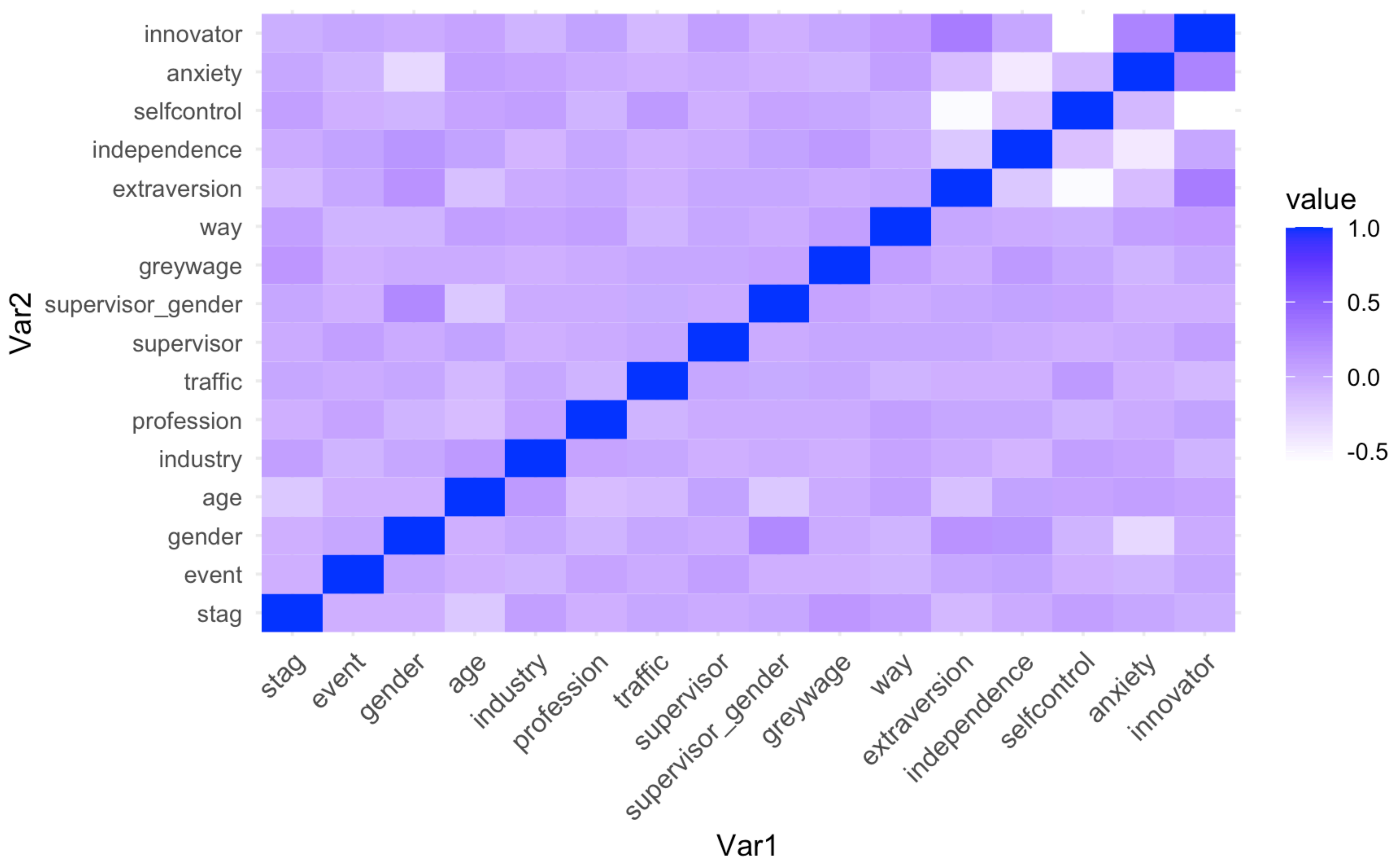
	stag <dbl>	event <int>	gender <dbl>	a... <int>	industry <dbl>	profession <dbl>	traffic <dbl>	supervisor <dbl>	supervisor_gender <dbl>
1	7.030801	1	0	35	2	6	4	1	1
2	22.965092	1	0	33	2	6	2	1	0
3	15.934292	1	1	35	8	6	4	1	0
4	15.934292	1	1	35	8	6	4	1	0
5	8.410678	1	0	32	10	2	7	2	1
6	8.969199	1	1	42	14	6	2	2	0

6 rows | 1-10 of 16 columns

EAD

Hide

```
# Correlation Plot
var1=Employees[, !colnames(Employees) %in% "event"]
var2=Employees$event
cor_matrix=cor(Employees)
reshaping=melt(cor_matrix)
ggplot(reshaping, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient(low="white", high="blue") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 10, hjust = 1))
```

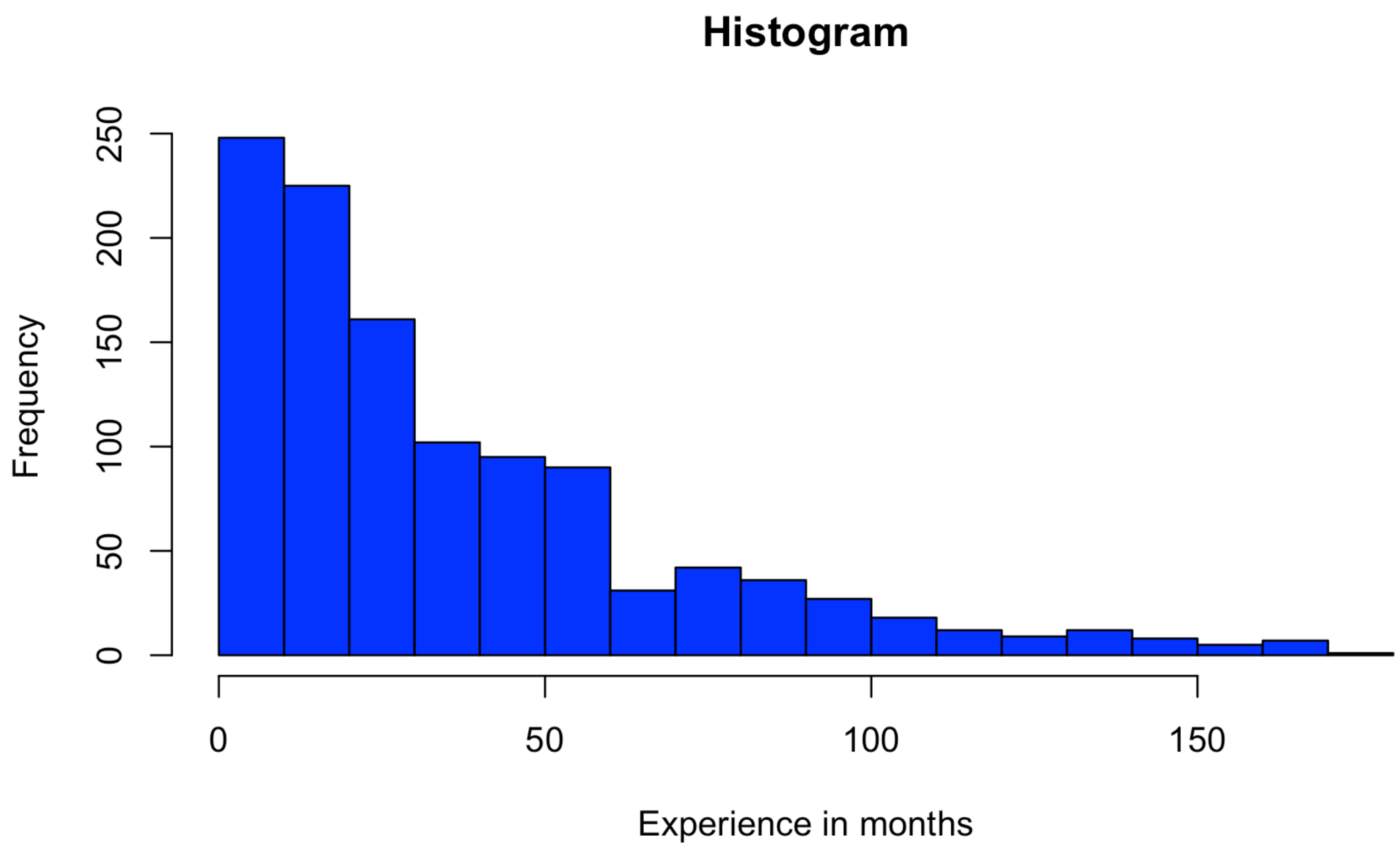


Hide

```
#corrplot(Employees, method = 'color')
# corr <- round(cor(Employees), 1)
# ggcorrplot(corr, method = 'square')
```

Hide

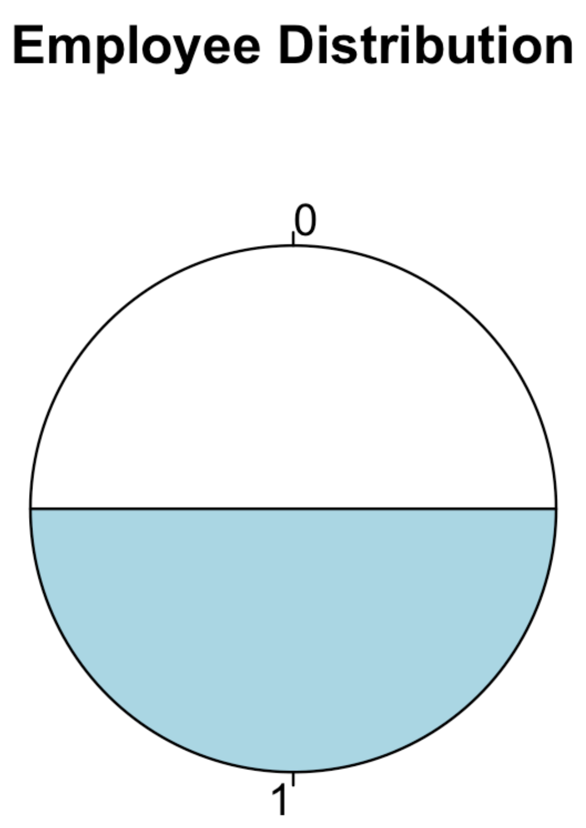
```
#distribution of employees as per their experience in months
num_bins <- 16
hist(Employees$stag, breaks = num_bins, main = "Histogram", xlab = "Experience in months", ylab = "Frequency",col
= 'blue')
```



We could see there are more employees with experience less than 50 months.

Hide

```
# Lets see the distribution of employee resigning or not

# create a frequency table of the "fruit" column
df <- table(unique(Employees$event))

# plot the frequency table as a pie chart
pie(df, labels = names(df), main = "Employee Distribution")
```



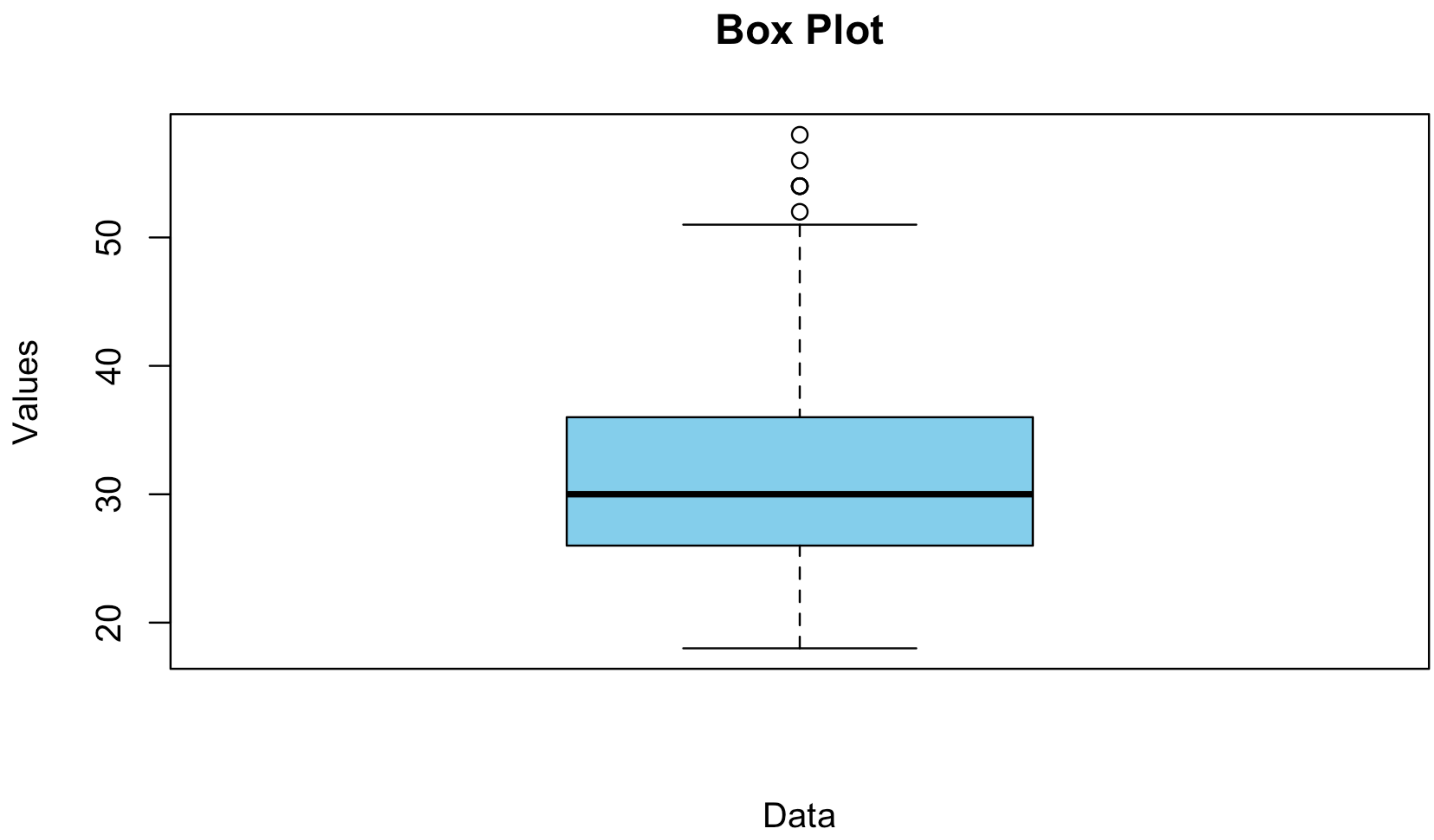
Hide

```
#we could see the distribution is almost equal
```

We could see the distribution is almost equal.

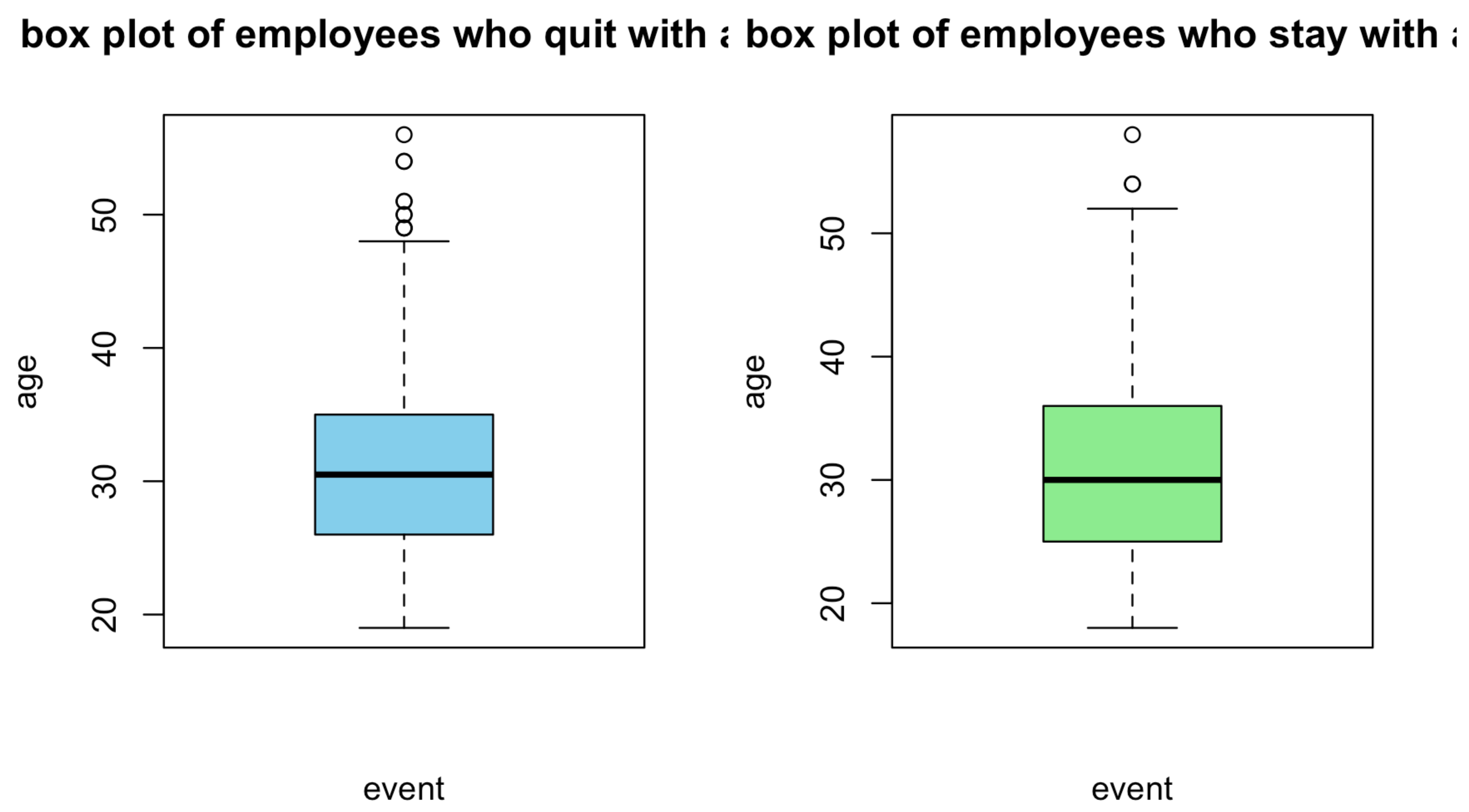
Hide

```
# Create a box plot with customization
df=data.frame(Employees$age)
boxplot(df,
  main = "Box Plot",
  xlab = "Data",
  ylab = "Values",
  col = "skyblue",
  border = "black",
  notchwidth = 0.5,
  horizontal = FALSE
)
```



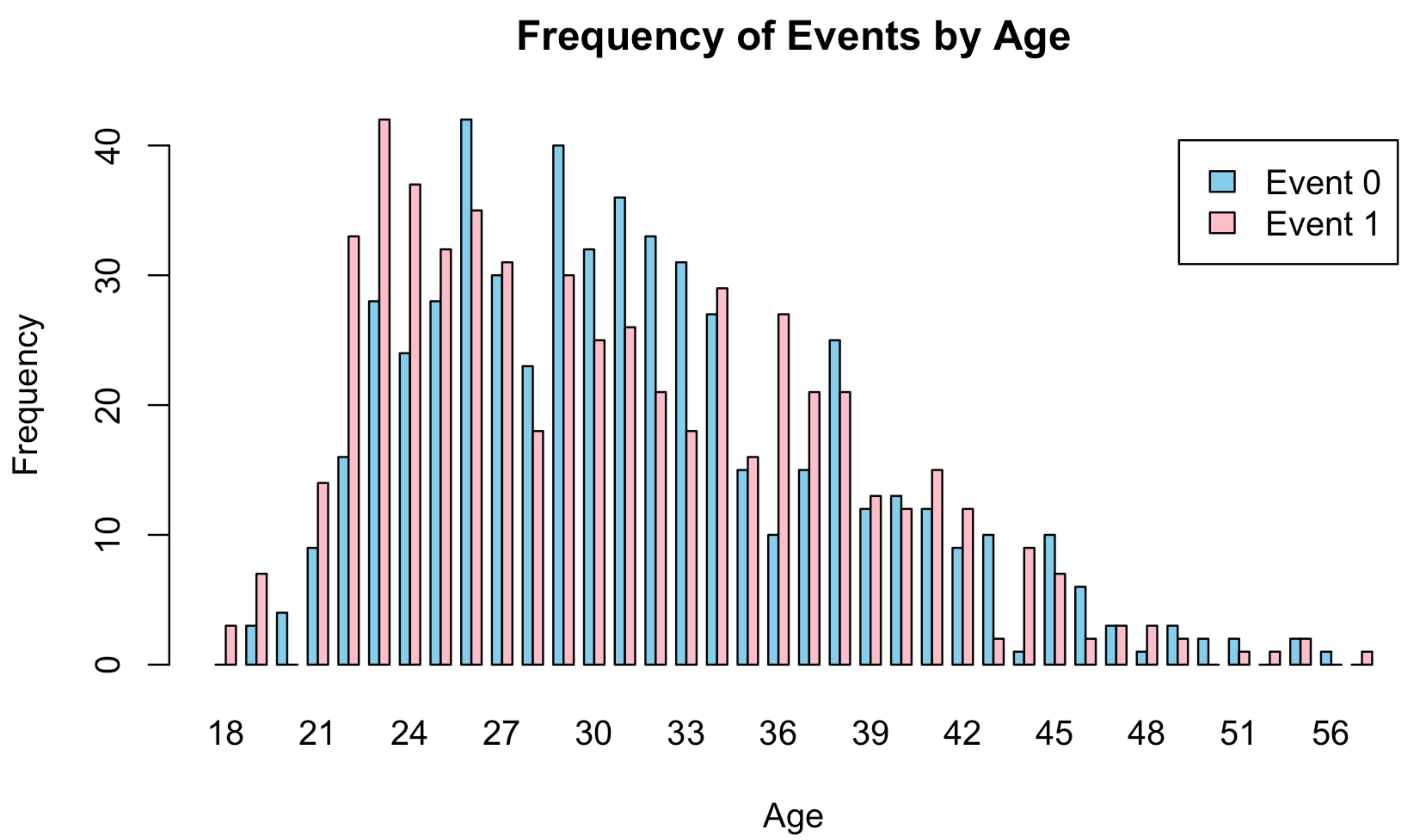
Hide

```
# Filter data for quitting
data_event0 <- subset(Employees, event == 0)

# Filter data for not quitting
data_event1 <- subset(Employees, event == 1)

# Create box plots for event 0 and event 1
par(mfrow = c(1, 2))  # Set up a 1x2 layout for side-by-side plots
boxplot(age ~ event, data = data_event0, col = "skyblue", main = "box plot of employees who quit with age")
boxplot(age ~ event, data = data_event1, col = "lightgreen", main = "box plot of employees who stay with age")
```



Hide

```
#seeing if age influence quitting
event_freq <- table(Employees$event,Employees$age)
my_colors <- c("skyblue", "pink")
# Create a bar plot
barplot(event_freq, beside = TRUE, legend.text = c("Event 0", "Event 1"),
        xlab = "Age", ylab = "Frequency", main = "Frequency of Events by Age",col = my_colors)
```



Hide

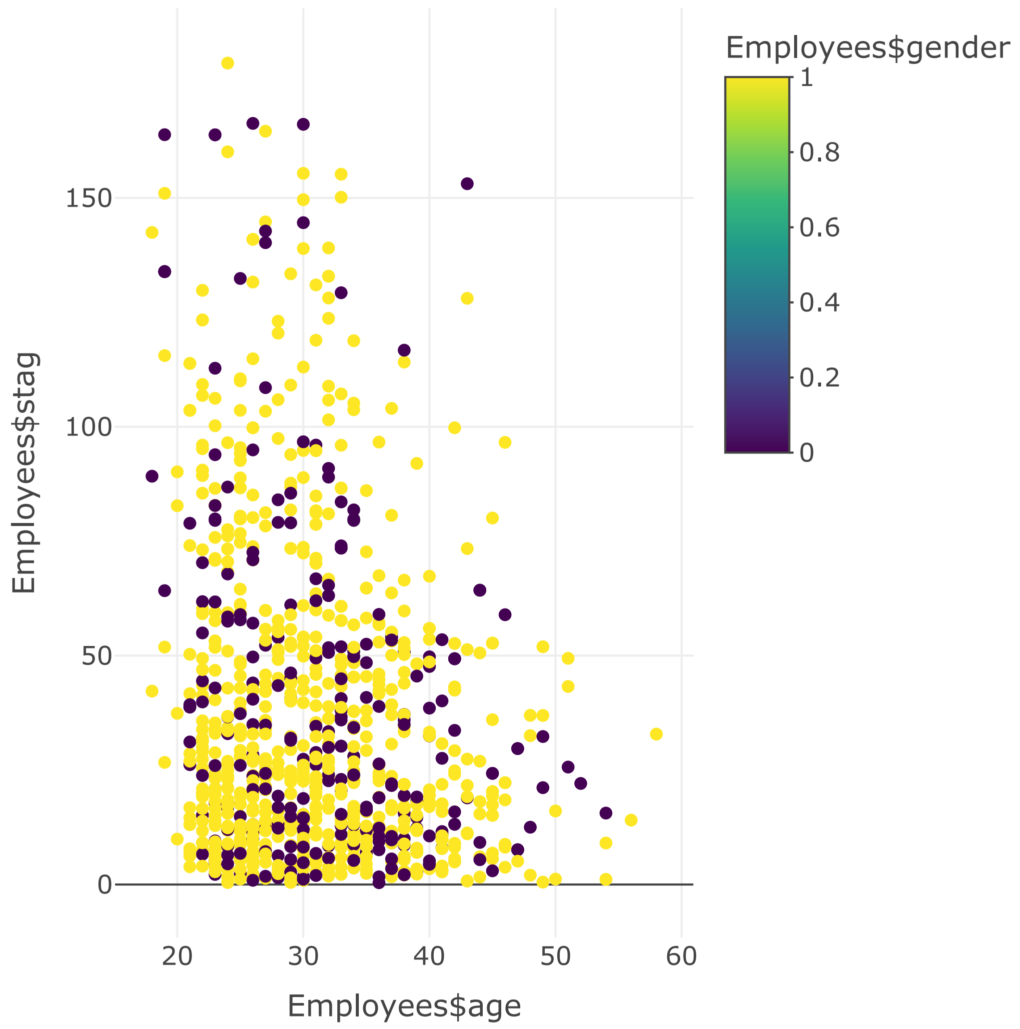
```
#we can see that employees from age 27-30 years tend to quit more often
```

We can see that employees from age 27-30 years tend to quit more often.

Hide

```
# Create a scatter plot with colors based on gender
p5 <-plot_ly(data = Employees, x = ~Employees$age, y = ~Employees$stag, color = ~Employees$gender)
p5
```

```
No trace type specified:
  Based on info supplied, a 'scatter' trace seems appropriate.
  Read more about this trace type -> https://plotly.com/r/reference/#scatter
No scatter mode specifed:
  Setting the mode to markers
  Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
No trace type specified:
  Based on info supplied, a 'scatter' trace seems appropriate.
  Read more about this trace type -> https://plotly.com/r/reference/#scatter
No scatter mode specifed:
  Setting the mode to markers
  Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```



Hide

```
pca_fit <- prcomp(select(Employees, -c("event")), scale. = TRUE)
pca_fit
```

```
Standard deviations (1, .., p=15):
 [1] 1.4293570 1.3118447 1.1627485 1.1140314 1.0583506 1.0305625 0.9980794 0.9906903 0.9551433 0.9419927 0.921007
5 0.8155224 0.7450777
[14] 0.6458678 0.4448485

Rotation (n x k) = (15 x 15):
                         PC1          PC2         PC3         PC4         PC5          PC6          PC7
PC8         PC9
stag               0.11107144  0.012015423 -0.34877227  0.49577623 -0.04395068  0.236629384  0.154203593  0.18951
708 -0.49025763
gender            -0.04578407 -0.489180206 -0.04117965 -0.23820358  0.11935219  0.346059424 -0.077581624 -0.09019
062  0.02631502
age                0.03561515  0.209491108  0.63900952 -0.09879427  0.06576115  0.294390654 -0.008070935 -0.04444
247  0.13392075
industry           0.09494245  0.121340334 -0.05765063 -0.07964117  0.52645897  0.451471593  0.368995002 -0.10637
865 -0.20687495
profession        -0.09333242  0.001971635 -0.09308453  0.14375383  0.47386001 -0.636666886  0.138319419 -0.31178
586  0.01051548
traffic            0.15418091 -0.066627864 -0.24040686 -0.15414259 -0.38454895 -0.002143075  0.451827683 -0.07567
782  0.39989661
supervisor        -0.07235582  0.016981392  0.12170878  0.04335960 -0.40865417  0.049183426 -0.068897796 -0.78172
567 -0.41803122
supervisor_gender  0.03799404 -0.271786807 -0.34221184 -0.03803960  0.13990470  0.141699457 -0.659770926 -0.11016
710  0.11151052
greywage           0.03248106 -0.108419400 -0.03129770  0.54296761 -0.22631816  0.156925365  0.078311959 -0.03533
436  0.37091931
way               -0.08495950  0.137906157  0.01661898  0.42829574  0.28220367  0.212216581 -0.057010472 -0.34409
067  0.43472418
extraversion      -0.49540436 -0.159210431 -0.20689294 -0.18523285  0.01314493  0.109649848  0.258694889 -0.10030
939  0.06919987
independence       0.02725154 -0.441405489  0.43183090  0.32032910  0.02297193 -0.125176077 -0.027353106  0.19590
048 -0.12828159
selfcontrol        0.60911671  0.095720451 -0.05290176 -0.04694869  0.02653949 -0.011512663 -0.072548941 -0.14122
042  0.04935976
anxiety           -0.13073054  0.594733381 -0.17237271 -0.01660031 -0.07967463  0.054908646 -0.289810476  0.11885
486  0.01962038
innovator         -0.54181910  0.097564941  0.05978231  0.12632113 -0.06881386  0.073544570 -0.065932841  0.11971
822 -0.01944254
                         PC10        PC11        PC12        PC13        PC14         PC15
stag              -0.075482326  0.20368618  0.23674245 -0.39446570 -0.05421891 -0.018274338
gender            -0.022192650  0.06075502  0.67602568  0.26004107 -0.13380409 -0.067843406
age                0.055141582 -0.07256200  0.20626571 -0.60597346 -0.08584823 -0.019000217
industry           0.373581605 -0.21467109 -0.26253694  0.19894857  0.02951384  0.017759126
profession         0.193071776 -0.13667826  0.33409986 -0.20212876 -0.06724184 -0.051537468
traffic            0.499941856  0.29967411  0.03086168 -0.16955670 -0.04644295 -0.030872515
supervisor         0.093580120 -0.01822360 -0.04901220  0.06818069 -0.03970323 -0.007578446
supervisor_gender  0.309891923 -0.07022223 -0.25117420 -0.37414299  0.02077301  0.046305178
greywage           0.008609911 -0.67260777  0.08103375  0.09282492 -0.04700983  0.026846684
way               -0.213839165  0.53112594 -0.09094661  0.13452818 -0.01959072  0.012922377
extraversion      -0.355705736 -0.13057396 -0.23243247 -0.26940548 -0.05604776 -0.533684834
independence       0.317384538  0.17734182 -0.22392330  0.12210868 -0.17385121 -0.464857277
selfcontrol       -0.148475666 -0.05152735  0.10005906  0.01399600  0.49635722 -0.552316523
anxiety            0.244325040 -0.03218289  0.18160720  0.19750366 -0.42314370 -0.423768012
innovator          0.323104779  0.05510539  0.18817020  0.04843840  0.70770695 -0.046300214
```

Hide

```
summary(pca_fit)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5    PC6     PC7     PC8     PC9    PC10    PC11    PC
12    PC13    PC14    PC15
Standard deviation     1.4294 1.3118 1.16275 1.11403 1.05835 1.0306 0.99808 0.99069 0.95514 0.94199 0.92101 0.815
52 0.74508 0.64587 0.44485
Proportion of Variance 0.1362 0.1147 0.09013 0.08274 0.07467 0.0708 0.06641 0.06543 0.06082 0.05916 0.05655 0.044
34 0.03701 0.02781 0.01319
Cumulative Proportion  0.1362 0.2509 0.34107 0.42380 0.49848 0.5693 0.63569 0.70112 0.76194 0.82110 0.87765 0.921
99 0.95900 0.98681 1.00000
```

Hide

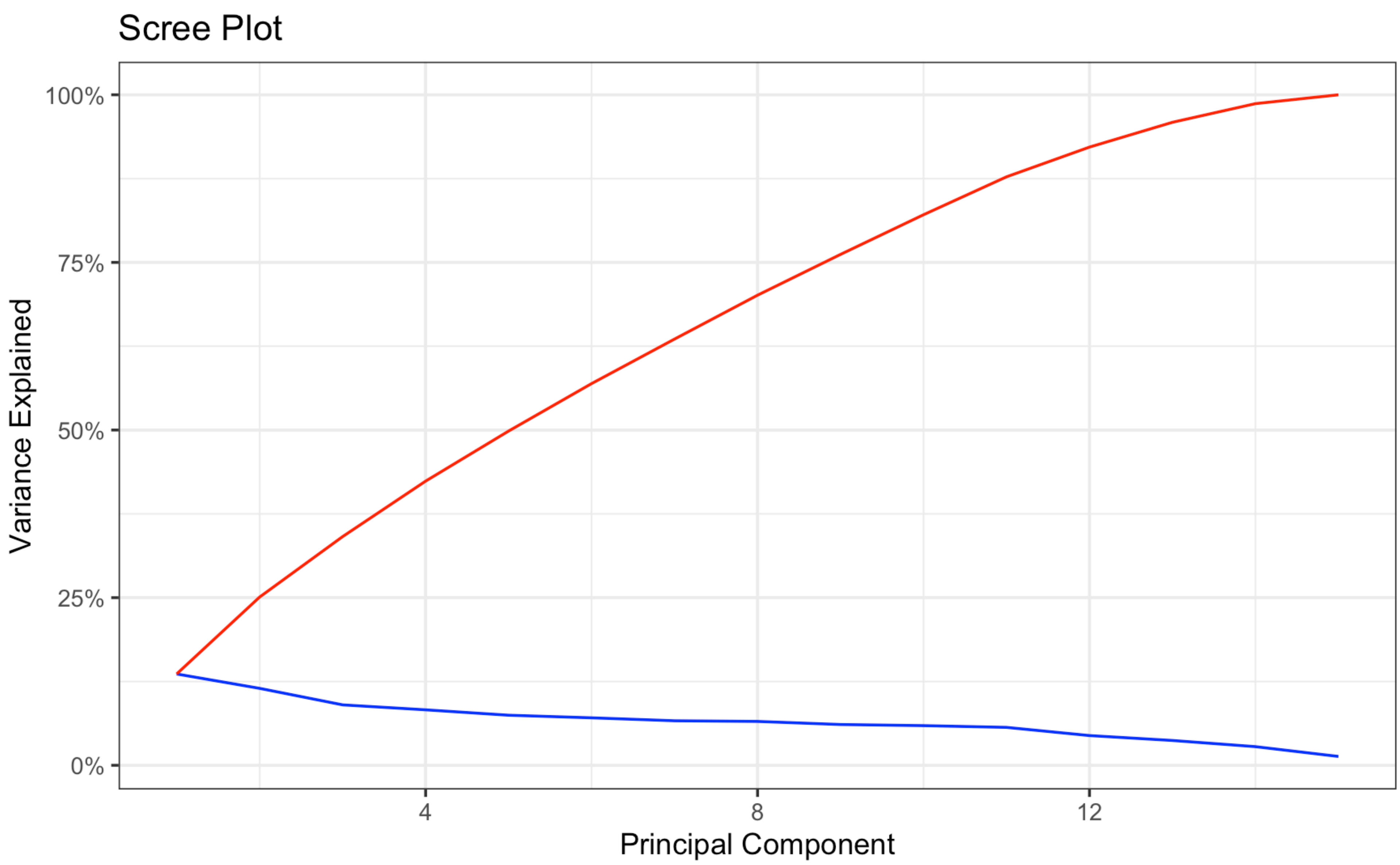
```
var_explained <- (pca_fit$sdev)^2 / sum(pca_fit$sdev^2)
round(var_explained,3)
```

```
[1] 0.136 0.115 0.090 0.083 0.075 0.071 0.066 0.065 0.061 0.059 0.057 0.044 0.037 0.028 0.013
```

Hide

```
cum_var <- cumsum(var_explained)
ggplot(data = data.frame(PC = 1:15, var_explained, cum_var), aes(x = PC)) +
  geom_line(aes(y = var_explained), color = "blue") +
  geom_line(aes(y = cum_var), color = "red") +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0, 1) +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```

```
Scale for y is already present.
Adding another scale for y, which will replace the existing scale.
```



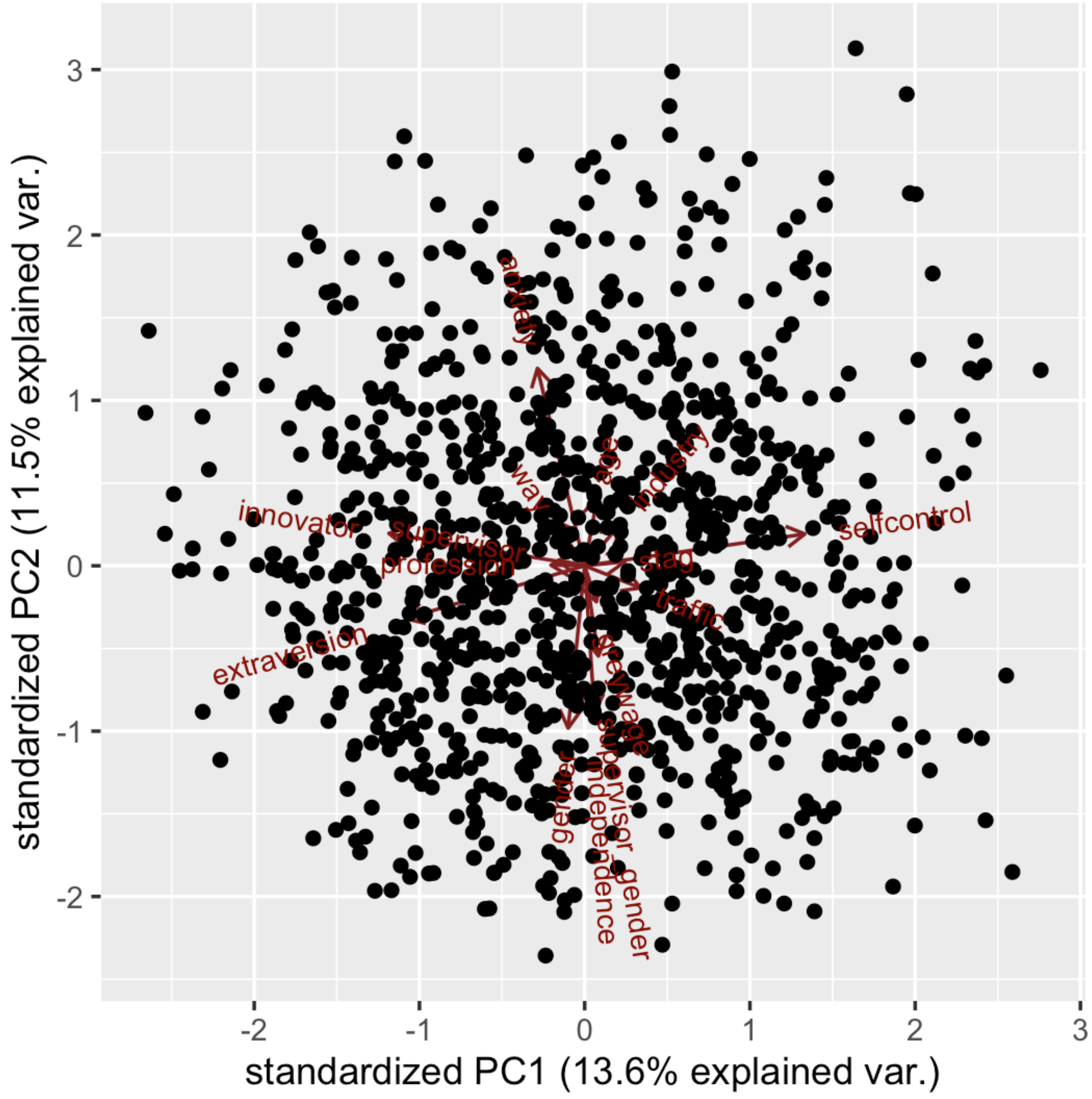
PCA can be used to reduce the dimensionality of a dataset while retaining most of its original variability. By projecting the original data onto a smaller number of dimensions, PCA can help identify underlying patterns and relationships between variables that may not be apparent in the original data.

Based on the plot, we can infer that the first principal component explains the most variance (0.136), followed by the second component (0.115), the third component (0.090), and so on.

Using the elbow method we can infer that almost all the PCs would be required to capture a significant amount of variance and hence wouldn't be of much use in this data.

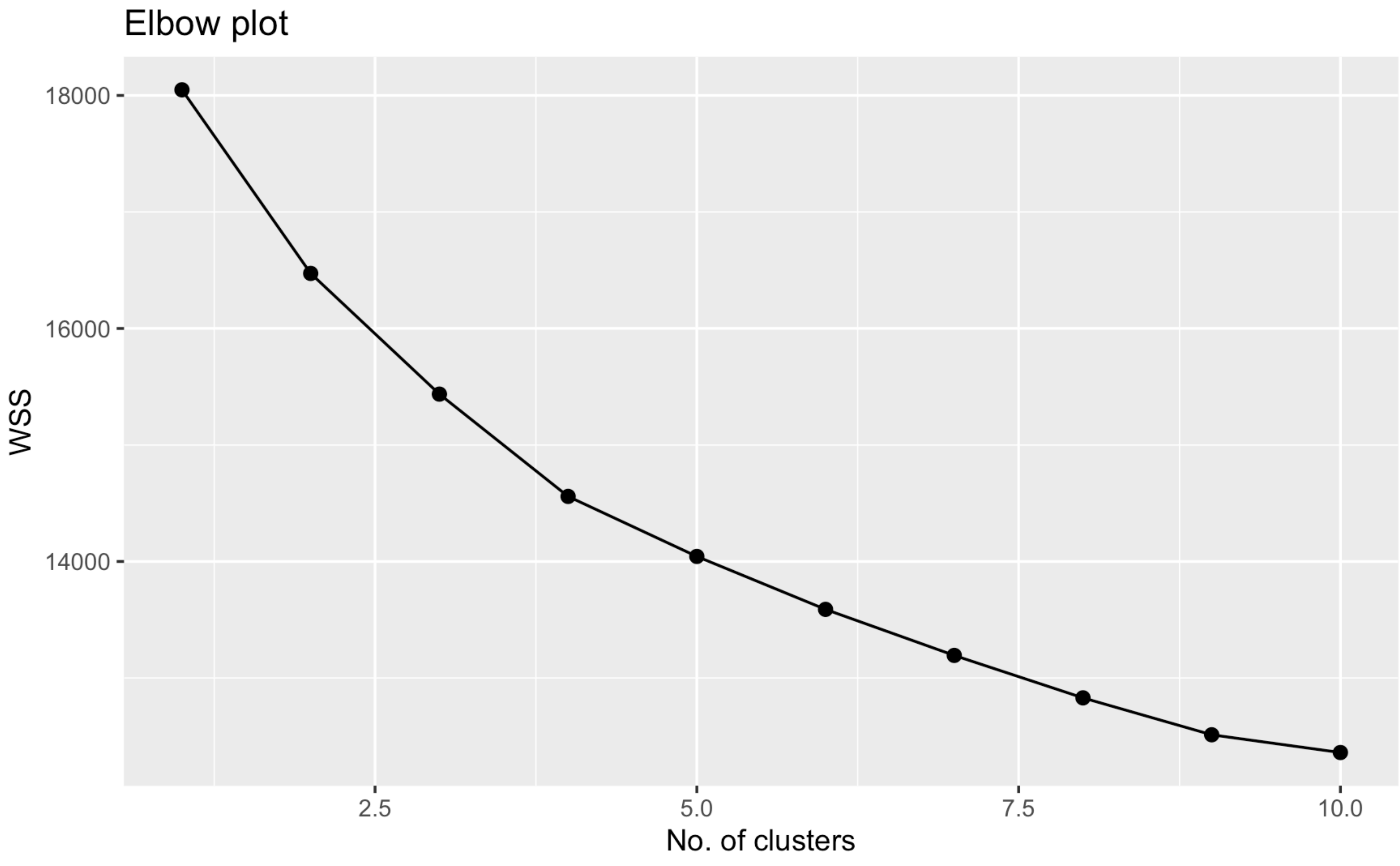
Hide

```
library(ggbiplot)
ggbiplot(pca_fit)
```



Hide

```
set.seed(2)
cluster_max <- 10
df_scale <- scale(Employees)
wss <- sapply(1:cluster_max, function(k){kmeans(df_scale, k, nstart=10 )$tot.withinss})
ggplot(data.frame(k=1:cluster_max, WSS=wss), aes(x=k, y=WSS)) +
  geom_point(size=2) +
  geom_line() +
  labs(title="Elbow plot", x="No. of clusters", y="WSS")
```



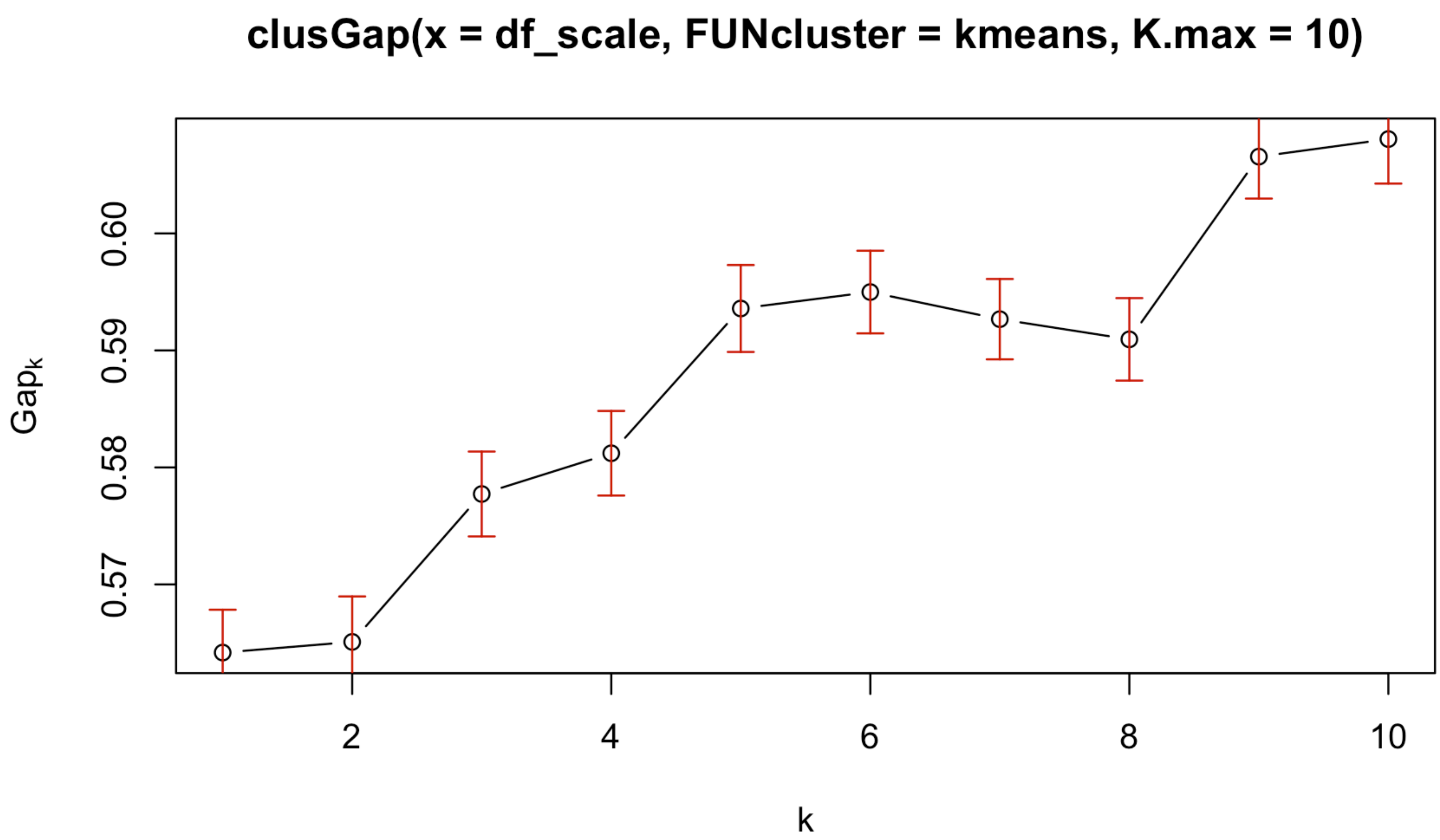
Hide

```
library(cluster)
gap_stat <- clusGap(df_scale, FUNcluster = kmeans, K.max = 10)
```

```
Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
.................................................. 50
.................................................. 100
```

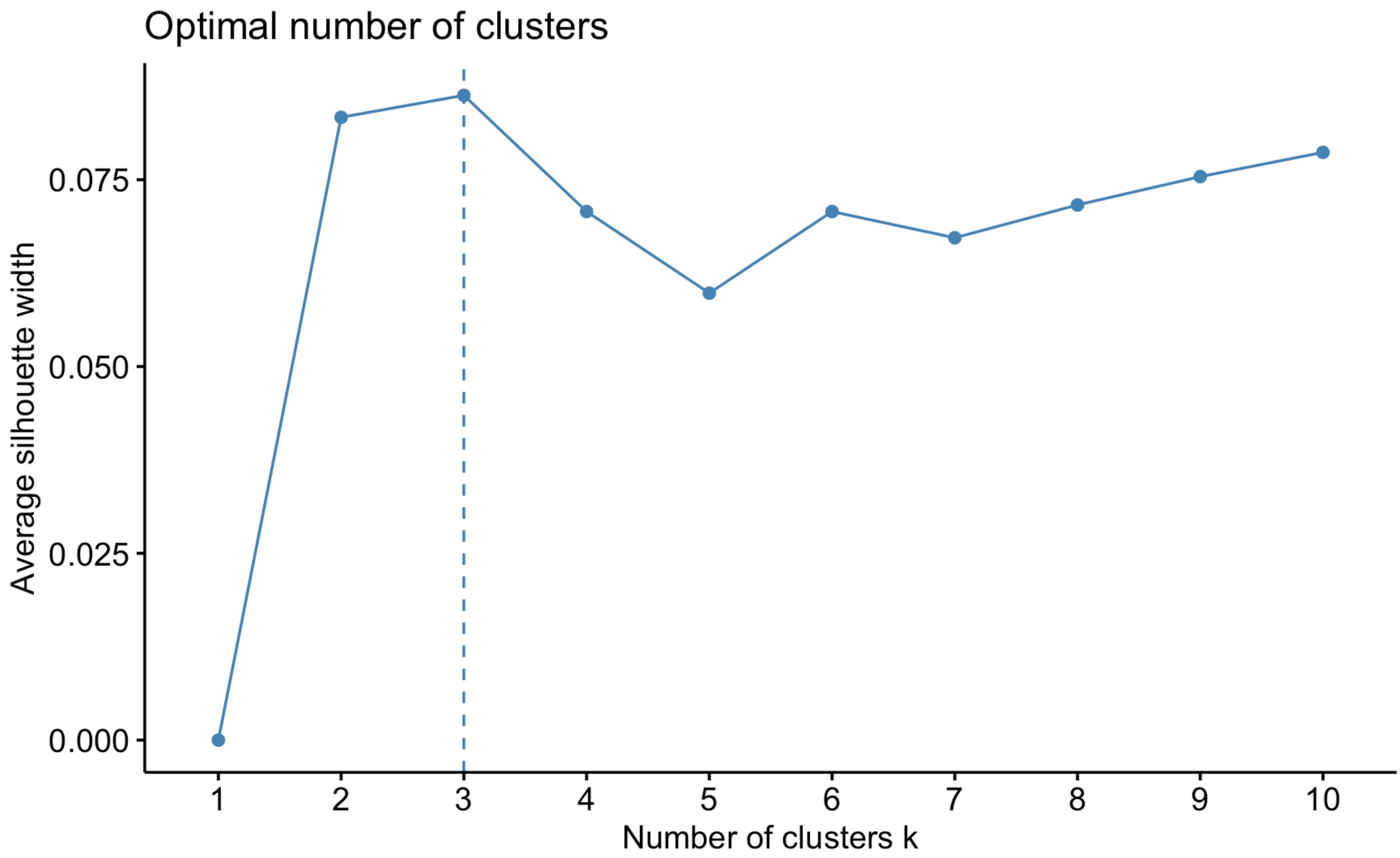
Hide

```
plot(gap_stat)
```



Hide

```
library(factoextra)
fviz_nbclust(df_scale,kmeans,method="silhouette")
```



Taking K=3 as 3 clusters.

Hide

```
km_out <- kmeans(df_scale, 3)
km_out
```

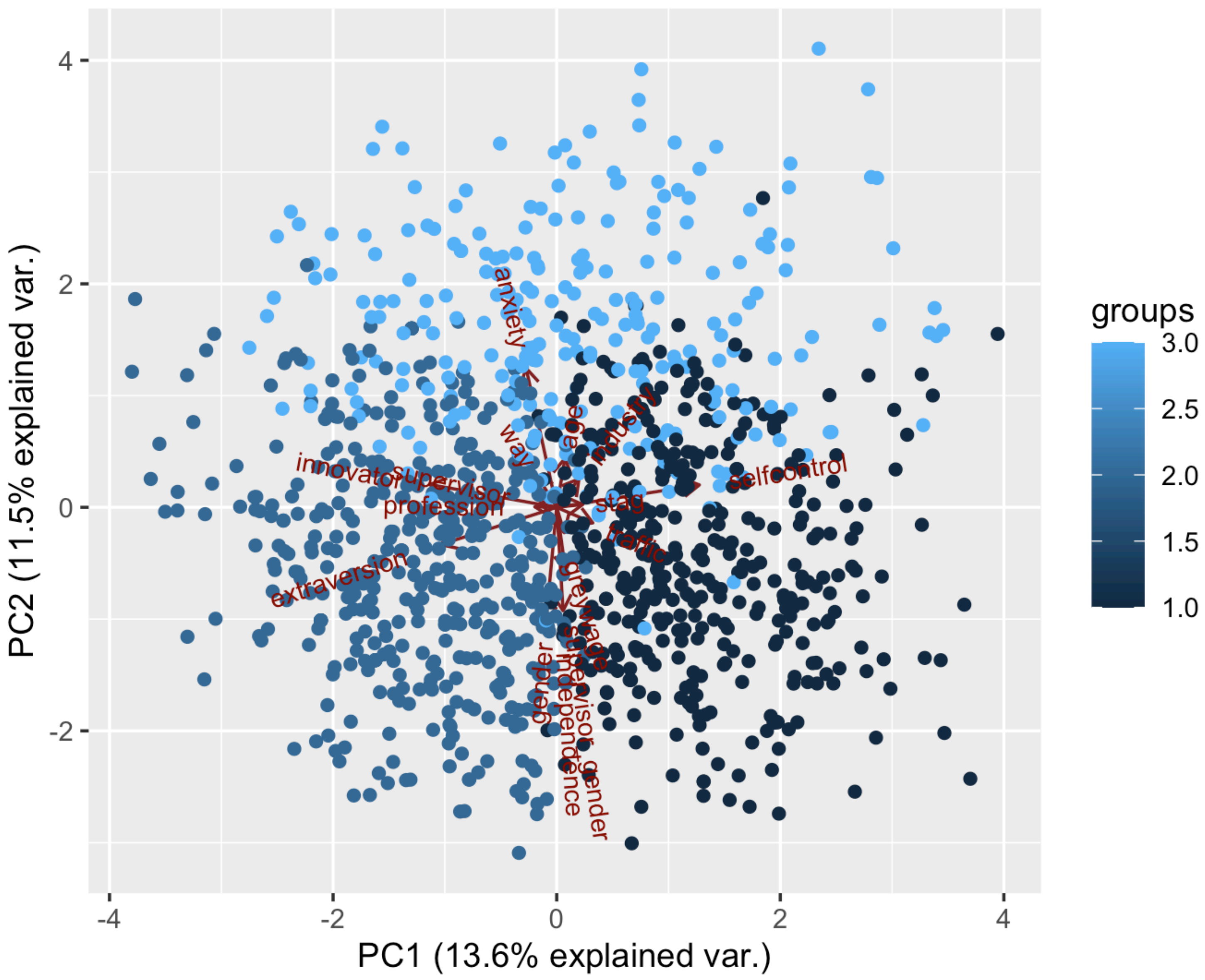
```
K-means clustering with 3 clusters of sizes 402, 463, 264

Cluster means:
         stag       event     gender         age    industry   profession    traffic  supervisor supervisor_gende
r     greywage         way
1  0.08656047 -0.02145678  0.5222902 -0.04466913  0.21502058 -0.102296783  0.3125969 -0.09903216        0.2286088
8  0.009604948 -0.14773631
2 -0.12913758  0.07269125  0.5183415 -0.03165001 -0.15644380  0.008082362 -0.2203708  0.09821655        0.0107874
1 -0.026767500  0.06160094
3  0.09467193 -0.09481222 -1.7043666  0.12352631 -0.05304847  0.141595353 -0.0895161 -0.02145202       -0.3670278
0  0.032318800  0.11692713
  extraversion independence selfcontrol    anxiety   innovator
1   -0.5257387 -0.003022397   0.7639490 -0.3051783 -0.72511112
2    0.6339926  0.162648949  -0.7598591 -0.1058803  0.60702920
3   -0.3113319 -0.280649469   0.1693457  0.6503950  0.03954602

Clustering vector:
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22   2
3   24   25   26   27   28
   3    3    2    2    3    2    2    1    2    2    2    3    2    2    2    2    2    2    2    2    1    2
2    1    1    2    2    3
  29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50   5
1   52   53   54   55   56
   1    1    2    3    1    2    2    1    2    2    2    2    1    1    1    2    1    1    2    2    2    1
2    3    2    2    2    3
  57   58   59   60   61   62   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77   78   7
9   80   81   82   83   84
   3    2    2    3    2    2    1    1    1    2    1    3    3    2    2    3    2    2    2    3    2    2
2    1    1    2    2    2
  85   86   87   88   89   90   91   92   93   94   95   96   97   98   99  100  101  102  103  104  105  106  10
7  108  109  110  111  112
   2    2    1    1    2    2    2    2    3    3    1    1    1    1    1    1    2    2    2    2    2    2
1    1    1    1    2    2
 113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128  129  130  131  132  133  134  13
5  136  137  138  139  140
   3    1    3    1    1    3    3    2    1    1    2    1    2    1    1    2    2    3    3    1    3    1
```

```
3   2   2   1   2   2
 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 16
3 164 165 166 167 168
  2   2   2   1   1   1   1   1   1   3   2   1   2   2   3   1   2   3   3   3   2   1
1   3   1   3   2   3
 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 19
1 192 193 194 195 196
  3   1   3   2   2   1   2   2   2   1   1   2   3   1   3   3   3   2   2   3   1   1
1   1   2   2   1   2
 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 21
9 220 221 222 223 224
  1   3   3   1   1   1   2   2   2   1   2   1   2   2   2   1   1   2   3   2   1   1
3   3   3   3   1   2
 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 24
7 248 249 250 251 252
  1   3   2   2   2   1   1   1   2   2   1   1   1   1   1   1   2   3   2   1   1   1
2   1   1   2   2   2
 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 27
5 276 277 278 279 280
  1   3   1   2   1   3   3   2   3   1   2   1   1   1   1   2   2   1   1   2   1   3
1   1   2   2   2   2
 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 30
3 304 305 306 307 308
  1   2   3   2   2   2   1   1   2   2   2   1   2   1   3   2   1   3   2   3   3   2
1   1   1   1   1   3
 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 33
1 332 333 334 335 336
  1   1   1   1   3   1   1   1   1   2   2   1   1   2   3   1   3   1   1   1   1   1
1   3   1   1   1   2
 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 35
9 360 361 362 363 364
  3   2   2   2   1   2   3   2   2   2   3   2   2   3   1   3   3   1   1   1   1   2
2   2   2   1   2   2
 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 38
7 388 389 390 391 392
  2   2   1   3   1   2   2   3   3   1   1   2   2   1   3   3   3   2   2   1   2   2
2   2   2   3   1   2
 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 41
5 416 417 418 419 420
  1   2   1   3   1   3   2   3   2   2   1   1   2   2   3   2   3   3   2   2   3   2
3   3   1   1   1   1
 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 44
3 444 445 446 447 448
  2   2   1   2   2   3   1   1   1   2   2   2   2   3   2   3   3   1   2   3   2   2
1   1   1   2   2   1
 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 47
1 472 473 474 475 476
  3   1   1   3   2   2   2   3   3   2   2   1   2   1   1   2   2   2   1   2   2   2
2   2   2   2   1   1
 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 49
9 500 501 502 503 504
  1   1   1   1   3   1   1   2   2   1   1   2   3   2   2   2   1   2   1   3   2   1
3   1   1   3   2   2
 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 52
7 528 529 530 531 532
  1   2   2   3   2   2   1   2   2   1   3   3   2   2   2   1   2   2   1   1   1   1
1   3   2   2   1   1
 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 55
5 556 557 558 559 560
  2   3   3   2   2   1   2   2   2   1   1   1   3   1   1   1   1   1   1   1   3   3
1   3   1   3   1   1
 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 58
3 584 585 586 587 588
  3   2   1   1   2   1   2   1   3   3   1   2   3   2   2   3   2   2   3   2   3   2
2   1   2   2   2   3
 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 61
1 612 613 614 615 616
  2   2   2   3   3   1   2   3   2   1   1   2   1   1   1   3   1   2   2   1   2   1
2   2   2   3   2   2
 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 63
9 640 641 642 643 644
  1   2   1   2   2   2   2   2   2   2   3   2   2   1   1   3   2   3   1   2   1   2
2   3   2   2   1   1
 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 66
7 668 669 670 671 672
```

```
   3    2    2    1    1    2    3    3    2    2    1    1    1    1    1    2    2    2    1    1    1    1 
2    2    1    1    1    1 
 673  674  675  676  677  678  679  680  681  682  683  684  685  686  687  688  689  690  691  692  693  694  69
5  696  697  698  699  700 
   3    3    3    3    2    1    1    2    2    3    2    1    1    1    1    1    2    2    2    2    2    1 
2    1    1    3    3    1 
 701  702  703  704  705  706  707  708  709  710  711  712  713  714  715  716  717  718  719  720  721  722  72
3  724  725  726  727  728 
   2    3    2    2    2    2    2    1    1    2    2    2    3    2    2    1    1    3    3    2    2    1 
2    2    2    1    2    1 
 729  730  731  732  733  734  735  736  737  738  739  740  741  742  743  744  745  746  747  748  749  750  75
1  752  753  754  755  756 
   2    3    1    2    3    3    3    3    1    1    1    3    2    1    3    3    2    2    2    2    1    2 
3    3    1    3    1    3 
 757  758  759  760  761  762  763  764  765  766  767  768  769  770  771  772  773  774  775  776  777  778  77
9  780  781  782  783  784 
   3    3    3    3    3    3    3    3    3    2    2    3    3    1    3    1    3    3    1    1    3    3 
3    1    1    3    1    2 
 785  786  787  788  789  790  791  792  793  794  795  796  797  798  799  800  801  802  803  804  805  806  80
7  808  809  810  811  812 
   2    1    2    1    3    2    2    1    1    2    2    2    1    2    2    2    1    2    2    2    1    1 
2    2    2    2    2    2 
 813  814  815  816  817  818  819  820  821  822  823  824  825  826  827  828  829  830  831  832  833  834  83
5  836  837  838  839  840 
   1    2    1    2    1    3    2    1    1    1    1    1    1    2    3    1    3    3    1    1    2    2 
2    1    3    2    2    2 
 841  842  843  844  845  846  847  848  849  850  851  852  853  854  855  856  857  858  859  860  861  862  86
3  864  865  866  867  868 
   3    3    3    1    2    1    3    2    3    2    1    1    1    2    1    3    3    2    2    2    2    1 
1    3    3    1    1    2 
 869  870  871  872  873  874  875  876  877  878  879  880  881  882  883  884  885  886  887  888  889  890  89
1  892  893  894  895  896 
   1    2    3    2    1    1    2    3    1    1    1    3    2    2    2    3    3    2    3    2    1    1 
2    2    2    2    1    3 
 897  898  899  900  901  902  903  904  905  906  907  908  909  910  911  912  913  914  915  916  917  918  91
9  920  921  922  923  924 
   2    2    1    1    1    2    2    1    1    1    1    2    1    1    2    1    2    1    1    2    2    1 
1    3    1    2    1    1 
 925  926  927  928  929  930  931  932  933  934  935  936  937  938  939  940  941  942  943  944  945  946  94
7  948  949  950  951  952 
   2    1    1    3    1    1    1    1    2    2    2    2    2    3    1    1    3    2    2    1    1    1 
1    1    3    2    1    2 
 953  954  955  956  957  958  959  960  961  962  963  964  965  966  967  968  969  970  971  972  973  974  97
5  976  977  978  979  980 
   3    1    3    1    2    1    2    2    3    3    1    1    2    2    1    2    3    1    3    1    1    1 
3    1    2    1    3    3 
 981  982  983  984  985  986  987  988  989  990  991  992  993  994  995  996  997  998  999 1000 
   2    1    1    3    1    3    3    3    1    2    2    2    1    2    3    2    3    3    2    2 
 [ reached getOption("max.print") -- omitted 129 entries ]

Within cluster sum of squares by cluster:
[1] 5448.113 6100.411 3889.063
 (between_SS / total_SS =  14.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "ite
r"         "ifault"      
```

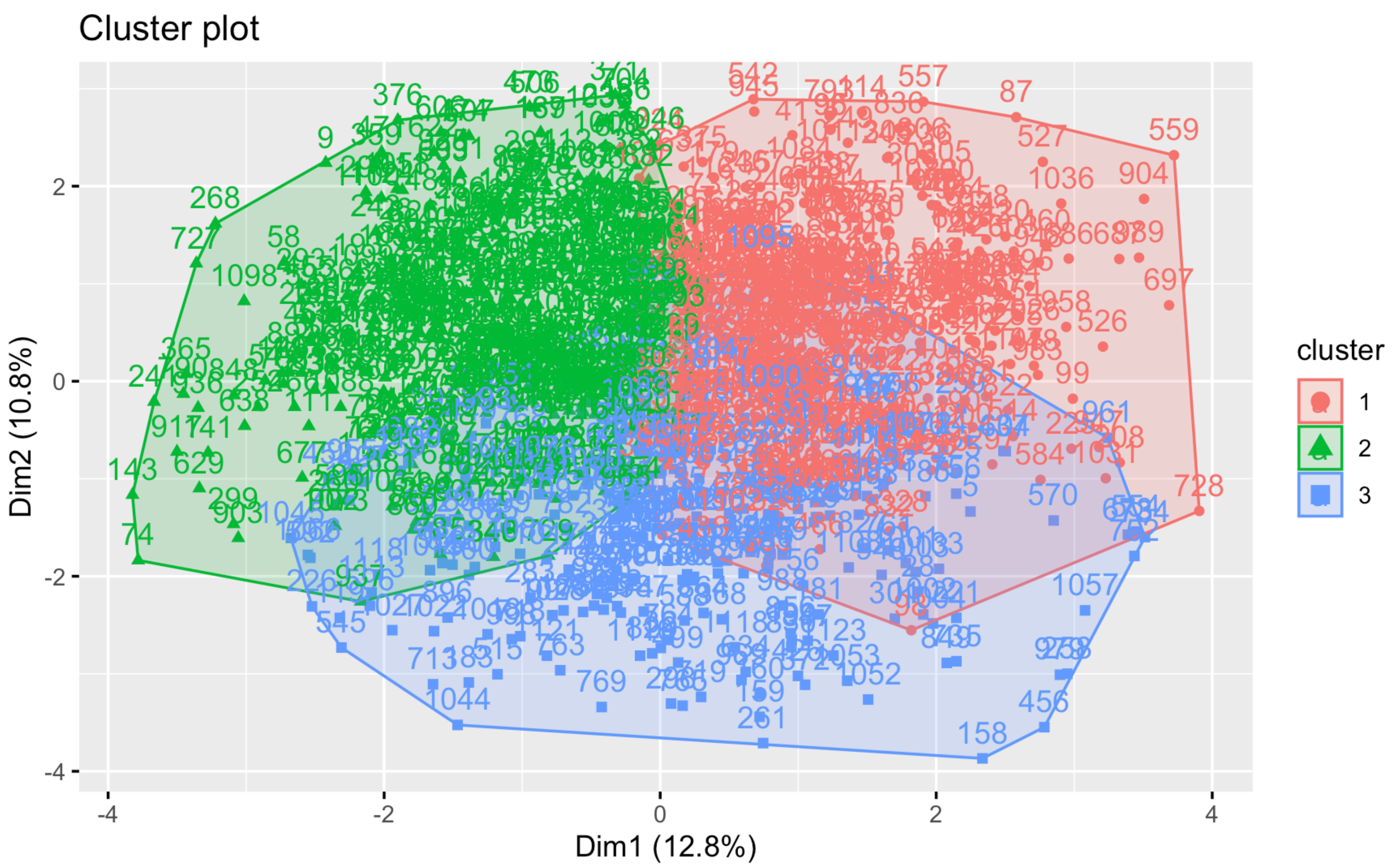
Hide

```
ggbiplot(pca_fit,groups=km_out$cluster,scale=0)
```



Hide

```
fviz_cluster(km_out, data=df_scale)
```



Clustering is used to group similar observations together based on their similarity.The clusters shows us different sub-groups in our data.

Based on the clusters, we can see this trend in our data:

-Cluster 1 has a relatively higher proportion of female employees, and they are relatively younger and have a lower wage. They also tend to have higher extraversion and innovation scores, but lower self-control and anxiety scores. Additionally, they are less likely to have a supervisor, and if they do, their supervisor is more likely to be male. Employees in this cluster are more likely to quit compared to those in the other clusters.

-Cluster 2 has a higher proportion of male employees and they are relatively older with a higher wage. They tend to have higher self-control and anxiety scores but lower extraversion and innovation scores. They are less likely to have a female supervisor. Employees in this cluster are less likely to quit compared to those in Cluster 1 but more likely to quit compared to those in Cluster 3.

-Cluster 3 has a relatively higher proportion of female employees, and they are relatively older with a higher wage. They tend to have lower extraversion and innovation scores but higher self-control and anxiety scores. They are more likely to have a female supervisor. Employees in this cluster are less likely to quit compared to those in the other clusters.

Hide

```
X <- subset(Employeess, select = -event)
y <- Employeess$event
head(X)
```

	stag <dbl>	gender <chr>	a... <dbl>	industry <chr>	profession <chr>	traffic <chr>	coach <chr>	head_gender <chr>	greywage <chr>
1	7.030801	m	35	Banks	HR	rabrecNErab	no	f	white
2	22.965092	m	33	Banks	HR	empjs	no	m	white
3	15.934292	f	35	PowerGeneration	HR	rabrecNErab	no	m	white
4	15.934292	f	35	PowerGeneration	HR	rabrecNErab	no	m	white
5	8.410678	m	32	Retail	Commercial	youjs	yes	f	white
6	8.969199	f	42	manufacture	HR	empjs	yes	m	white

6 rows | 1-10 of 15 columns

Hide

```
#Kaplan-Meier survival curve

library(survival)
fit.surv <- survfit(Surv(stag, event) ~ 1, data=Employeess)
summary(fit.surv)
```

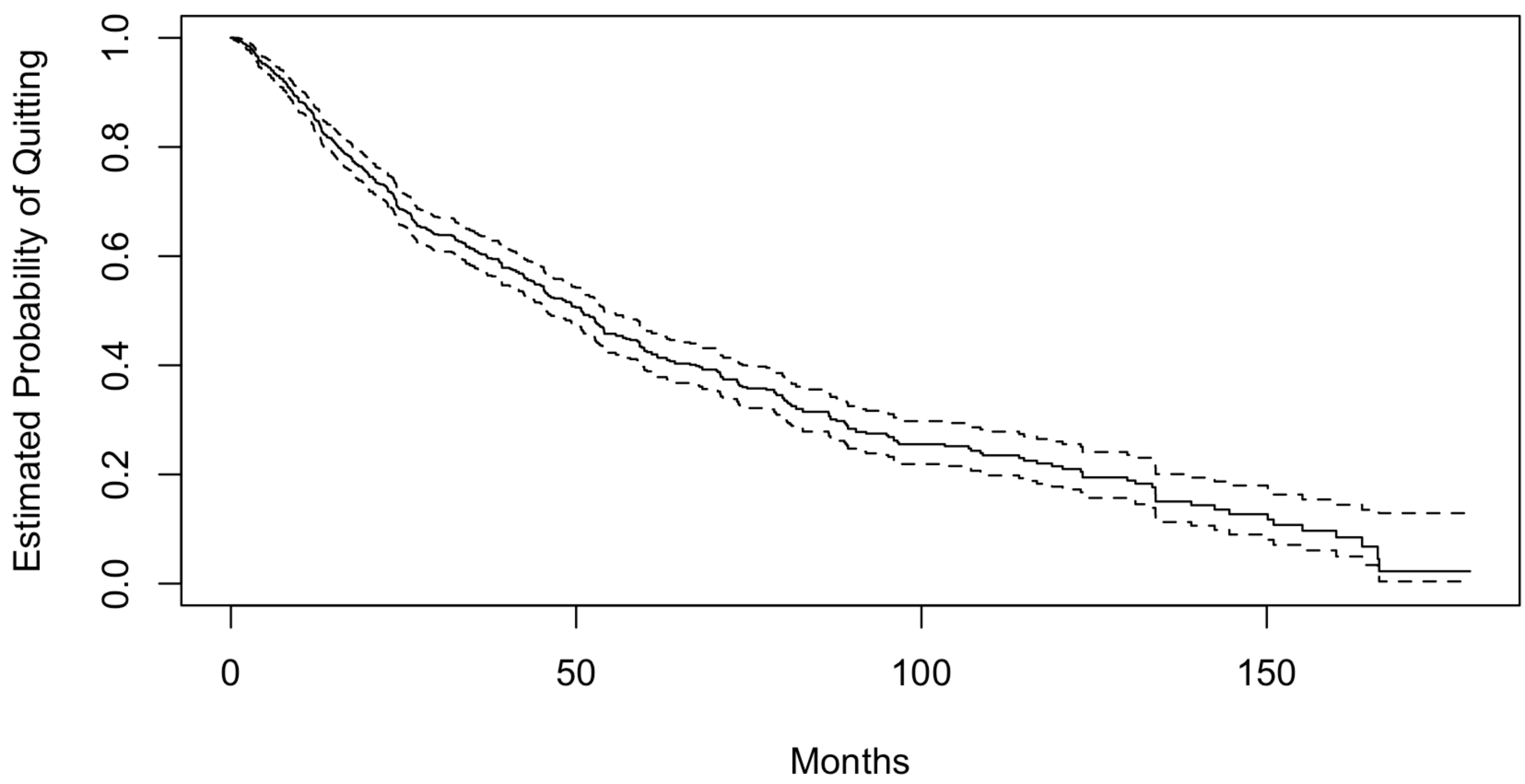
```
Call: survfit(formula = Surv(stag, event) ~ 1, data = Employeess)

    time n.risk n.event survival  std.err lower 95% CI upper 95% CI
   0.394   1129       1   0.9991 0.000885      0.99738        1.000
   0.427   1128       1   0.9982 0.001252      0.99578        1.000
   0.756   1122       1   0.9973 0.001534      0.99434        1.000
   0.920   1119       1   0.9964 0.001773      0.99298        1.000
   1.117   1118       2   0.9947 0.002172      0.99042        0.999
   1.413   1112       1   0.9938 0.002347      0.98918        0.998
   1.478   1110       1   0.9929 0.002510      0.98797        0.998
   1.643   1109       1   0.9920 0.002663      0.98677        0.997
   1.708   1107       1   0.9911 0.002807      0.98560        0.997
   1.741   1105       1   0.9902 0.002944      0.98443        0.996
   1.807   1104       1   0.9893 0.003075      0.98328        0.995
   2.037   1097       1   0.9884 0.003202      0.98213        0.995
   2.201   1086       2   0.9866 0.003445      0.97984        0.993
   2.267   1082       1   0.9857 0.003560      0.97870        0.993
   2.431   1079       1   0.9847 0.003672      0.97757        0.992
   2.760   1076       1   0.9838 0.003781      0.97644        0.991
   2.793   1073       2   0.9820 0.003990      0.97420        0.990
   2.858   1069       2   0.9802 0.004189      0.97198        0.988
   2.990   1067       1   0.9792 0.004285      0.97088        0.988
   3.121   1066       2   0.9774 0.004469      0.96868        0.986
   3.253   1062       3   0.9746 0.004732      0.96541        0.984
   3.351   1058       1   0.9737 0.004817      0.96432        0.983
   3.417   1057       1   0.9728 0.004899      0.96324        0.982
   3.450   1056       1   0.9719 0.004981      0.96216        0.982
   3.483   1054       1   0.9710 0.005061      0.96109        0.981
   3.581   1051       4   0.9673 0.005368      0.95679        0.978
   3.811   1045       2   0.9654 0.005515      0.95466        0.976
   3.975   1041       6   0.9598 0.005933      0.94829        0.972
   4.008   1035       1   0.9589 0.005999      0.94723        0.971
   4.074   1033       1   0.9580 0.006065      0.94617        0.970
   4.337   1031       1   0.9571 0.006129      0.94512        0.969
   4.370   1030       1   0.9561 0.006194      0.94407        0.968
   4.435   1028       2   0.9543 0.006320      0.94196        0.967
   4.534   1024       2   0.9524 0.006443      0.93986        0.965
```

4.961	1019	1	0.9515	0.006504	0.93881	0.964
5.027	1014	2	0.9496	0.006626	0.93670	0.963
5.257	1006	1	0.9486	0.006686	0.93564	0.962
5.322	1004	1	0.9477	0.006746	0.93458	0.961
5.388	1002	1	0.9468	0.006805	0.93352	0.960
5.421	1001	1	0.9458	0.006863	0.93246	0.959
5.651	996	1	0.9449	0.006922	0.93139	0.959
5.749	994	1	0.9439	0.006980	0.93033	0.958
5.782	991	1	0.9430	0.007038	0.92927	0.957
5.815	990	1	0.9420	0.007095	0.92820	0.956
5.848	989	2	0.9401	0.007207	0.92608	0.954
6.144	982	1	0.9391	0.007263	0.92502	0.953
6.177	981	3	0.9363	0.007428	0.92183	0.951
6.275	977	2	0.9344	0.007535	0.91970	0.949
6.439	974	1	0.9334	0.007588	0.91864	0.948
6.538	973	1	0.9324	0.007641	0.91758	0.948
6.604	972	1	0.9315	0.007693	0.91652	0.947
6.669	971	1	0.9305	0.007745	0.91546	0.946
7.031	966	4	0.9267	0.007949	0.91122	0.942
7.129	962	1	0.9257	0.007998	0.91016	0.942
7.326	955	1	0.9247	0.008049	0.90909	0.941
7.589	953	5	0.9199	0.008294	0.90377	0.936
7.819	948	1	0.9189	0.008341	0.90271	0.935
7.951	940	2	0.9170	0.008437	0.90057	0.934
8.016	937	1	0.9160	0.008485	0.89950	0.933
8.082	936	1	0.9150	0.008532	0.89843	0.932
8.148	935	1	0.9140	0.008579	0.89736	0.931
8.181	933	3	0.9111	0.008718	0.89416	0.928
8.279	928	1	0.9101	0.008763	0.89309	0.927
8.312	927	1	0.9091	0.008809	0.89202	0.927
8.411	925	1	0.9081	0.008854	0.89095	0.926
8.575	923	1	0.9072	0.008899	0.88988	0.925
8.608	922	1	0.9062	0.008943	0.88881	0.924
8.641	921	2	0.9042	0.009032	0.88667	0.922
8.772	918	1	0.9032	0.009075	0.88560	0.921
8.871	914	1	0.9022	0.009119	0.88453	0.920
8.936	913	1	0.9012	0.009162	0.88346	0.919
8.969	912	3	0.8983	0.009291	0.88025	0.917
9.035	909	1	0.8973	0.009333	0.87918	0.916
9.101	905	2	0.8953	0.009417	0.87704	0.914
9.199	901	1	0.8943	0.009459	0.87596	0.913
9.265	900	1	0.8933	0.009500	0.87489	0.912
9.528	896	1	0.8923	0.009542	0.87381	0.911
9.593	895	1	0.8913	0.009583	0.87274	0.910
9.626	894	1	0.8903	0.009624	0.87166	0.909
9.791	890	6	0.8843	0.009866	0.86519	0.904
9.823	884	1	0.8833	0.009906	0.86412	0.903
9.889	883	1	0.8823	0.009945	0.86304	0.902
10.349	878	1	0.8813	0.009984	0.86196	0.901
10.480	876	1	0.8803	0.010023	0.86088	0.900
10.645	869	1	0.8793	0.010063	0.85979	0.899
10.776	868	2	0.8773	0.010141	0.85762	0.897
10.842	866	1	0.8763	0.010180	0.85653	0.896
10.908	865	1	0.8752	0.010219	0.85545	0.896
10.940	864	1	0.8742	0.010257	0.85436	0.895
11.006	863	1	0.8732	0.010295	0.85327	0.894
11.039	862	2	0.8712	0.010370	0.85110	0.892
11.072	860	1	0.8702	0.010407	0.85002	0.891
11.236	859	1	0.8692	0.010444	0.84894	0.890
11.302	857	1	0.8682	0.010481	0.84785	0.889
11.499	856	2	0.8661	0.010555	0.84568	0.887
11.696	850	3	0.8631	0.010664	0.84242	0.884
11.828	842	1	0.8620	0.010700	0.84132	0.883
11.893	839	1	0.8610	0.010737	0.84023	0.882
11.959	837	2	0.8590	0.010809	0.83803	0.880
11.992	835	3	0.8559	0.010916	0.83474	0.878
12.025	832	2	0.8538	0.010987	0.83255	0.876
12.057	830	1	0.8528	0.011021	0.83146	0.875
12.090	829	1	0.8518	0.011056	0.83036	0.874
12.123	828	1	0.8507	0.011090	0.82927	0.873
12.222	827	1	0.8497	0.011125	0.82817	0.872
12.353	822	1	0.8487	0.011159	0.82707	0.871
12.386	821	1	0.8476	0.011193	0.82598	0.870
12.682	817	2	0.8456	0.011262	0.82377	0.868
12.780	815	3	0.8424	0.011363	0.82047	0.865

```
  12.813    812       1    0.8414 0.011396      0.81937        0.864
  12.879    811       2    0.8393 0.011462      0.81717        0.862
  12.977    809       2    0.8373 0.011527      0.81497        0.860
  13.010    807       2    0.8352 0.011591      0.81277        0.858
  13.076    805       1    0.8341 0.011623      0.81167        0.857
  13.109    804       3    0.8310 0.011718      0.80838        0.854
  13.142    801       1    0.8300 0.011749      0.80728        0.853
  13.175    800       2    0.8279 0.011811      0.80509        0.851
  13.273    798       1    0.8269 0.011842      0.80400        0.850
  13.339    797       1    0.8258 0.011872      0.80290        0.849
  13.405    796       2    0.8238 0.011933      0.80071        0.847
  13.569    794       1    0.8227 0.011963      0.79962        0.847
  13.667    793       1    0.8217 0.011993      0.79852        0.846
  13.864    791       1    0.8207 0.012022      0.79743        0.845
  13.897    790       3    0.8175 0.012111      0.79415        0.842
  14.226    785       1    0.8165 0.012140      0.79305        0.841
  14.587    777       2    0.8144 0.012199      0.79083        0.839
  14.620    775       1    0.8133 0.012229      0.78973        0.838
  14.719    774       1    0.8123 0.012258      0.78862        0.837
  14.850    773       2    0.8102 0.012316      0.78641        0.835
  14.949    771       1    0.8091 0.012345      0.78531        0.834
  14.982    770       1    0.8081 0.012373      0.78420        0.833
  15.113    768       1    0.8070 0.012402      0.78309        0.832
  15.146    765       1    0.8060 0.012431      0.78199        0.831
  15.211    762       1    0.8049 0.012459      0.78087        0.830
  15.310    761       1    0.8039 0.012488      0.77976        0.829
  15.343    760       1    0.8028 0.012516      0.77865        0.828
  15.409    758       1    0.8018 0.012544      0.77754        0.827
  15.540    757       1    0.8007 0.012572      0.77643        0.826
  15.573    756       2    0.7986 0.012628      0.77420        0.824
  15.737    749       1    0.7975 0.012656      0.77309        0.823
  15.934    747       2    0.7954 0.012712      0.77085        0.821
  16.000    744       2    0.7932 0.012767      0.76860        0.819
 [ reached getOption("max.print") -- omitted 274 rows ]
```

Hide

```
plot(fit.surv, xlab = "Months",
    ylab = "Estimated Probability of Quitting")
```



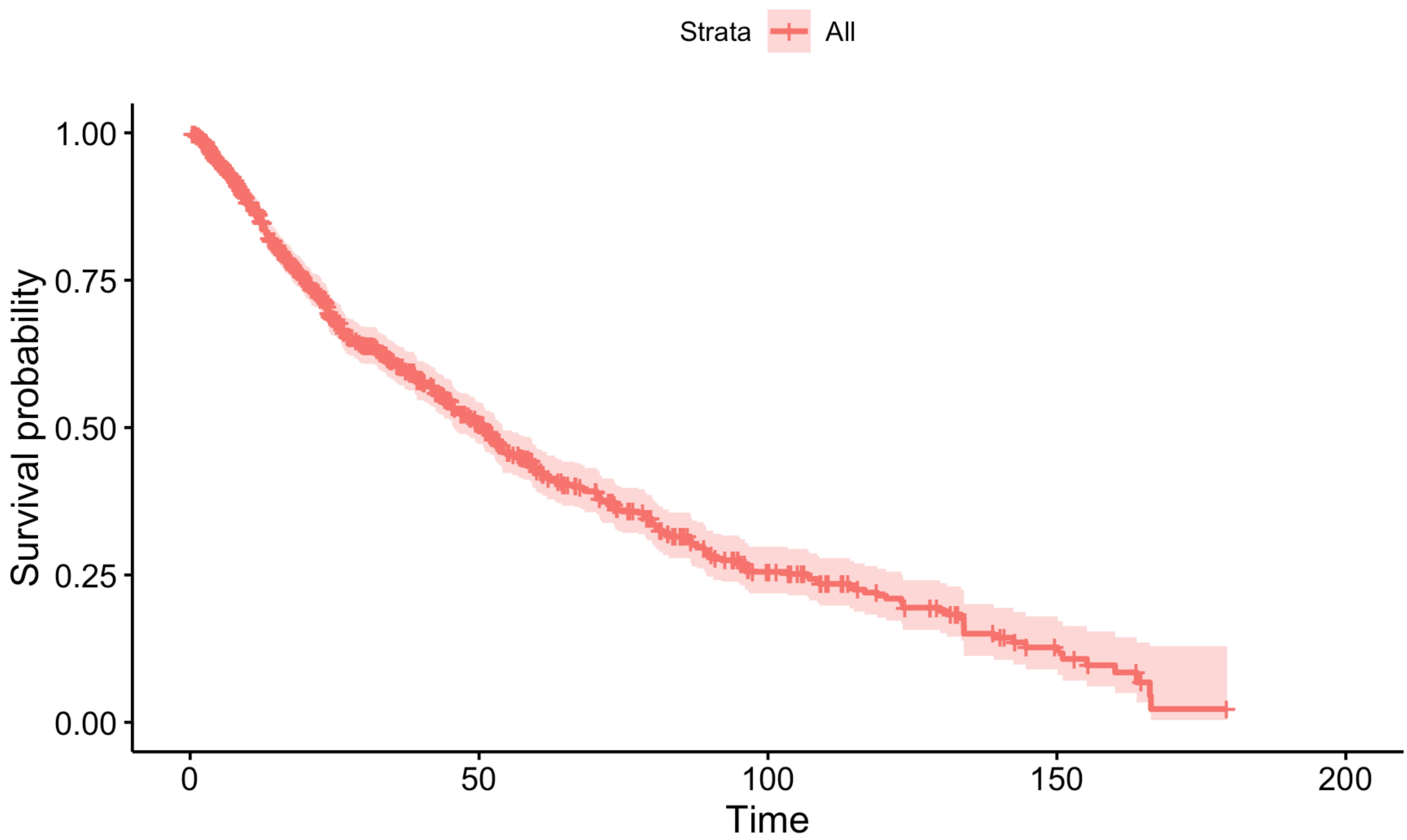
Hide

```
library(survminer)
```

```
Loading required package: ggpubr

Attaching package: 'ggpubr'

The following object is masked from 'package:plyr':

    mutate


Attaching package: 'survminer'

The following object is masked from 'package:survival':

    myeloma
```

Hide

```
ggsurvplot(fit = fit.surv)
```

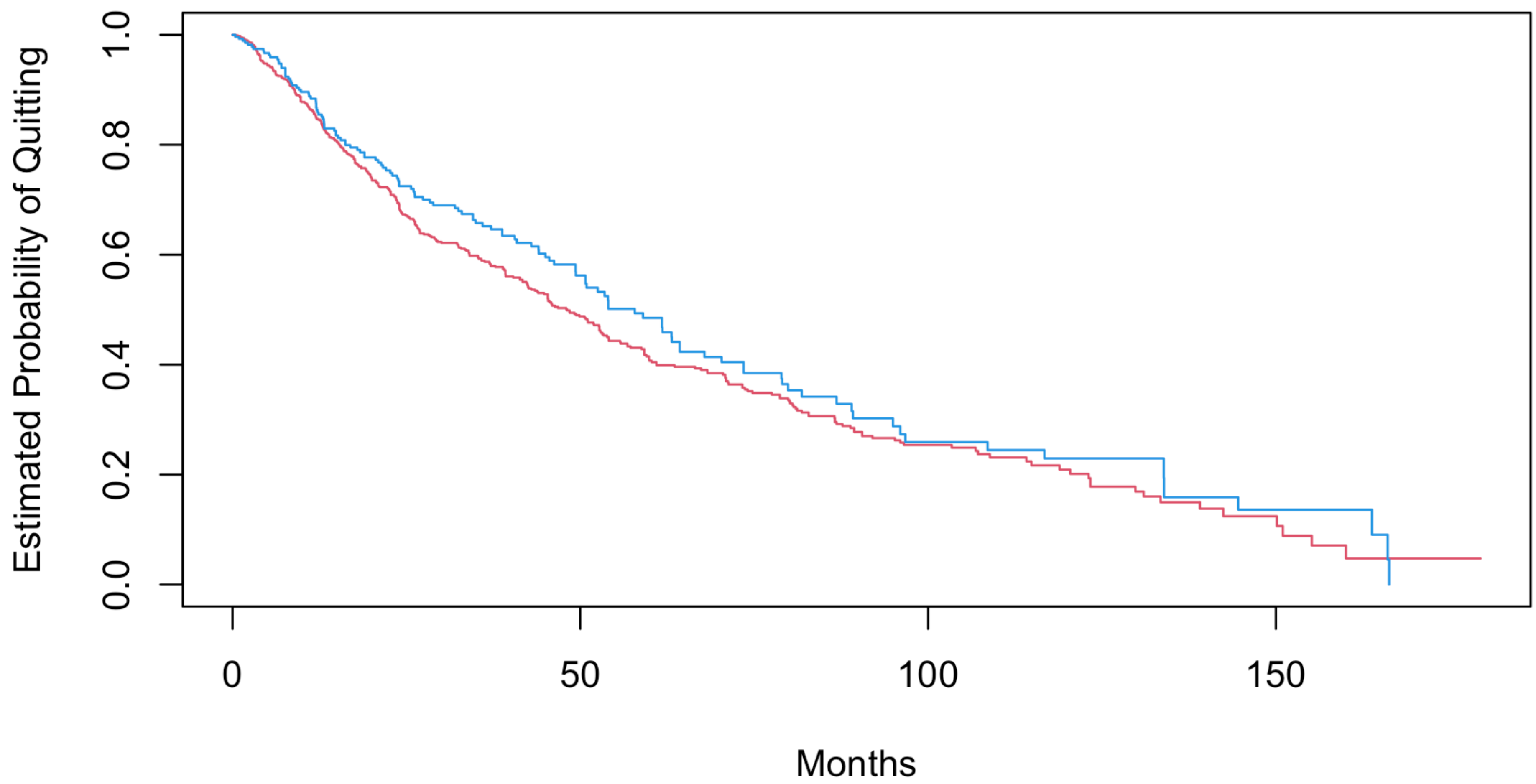


From the Kaplan-Meier curve above, we can say that with time the probability of an employee decreases. We can see that it does not decrease rapidly over time.

From the graph the median survival time of an employee seems to be around 50 months.

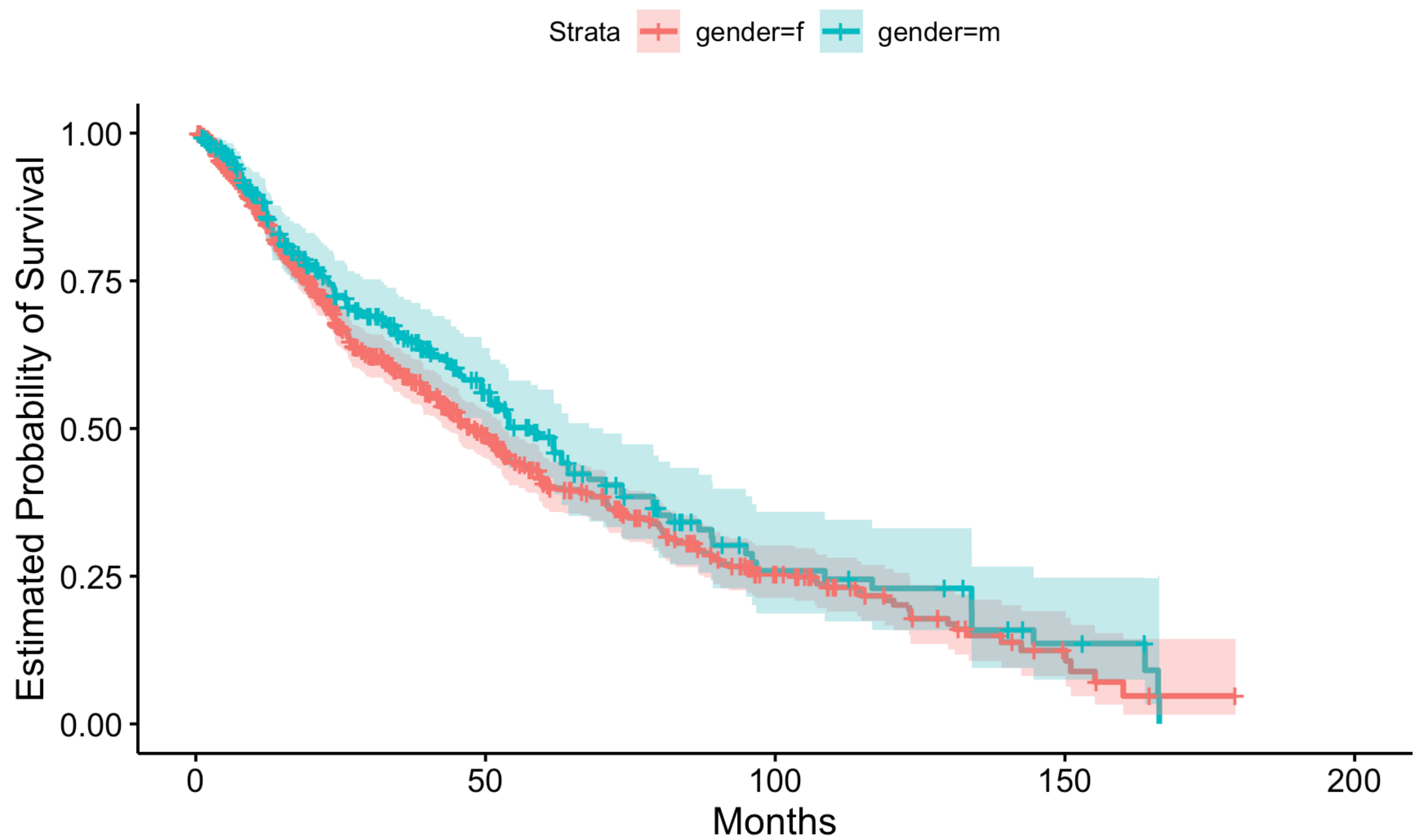
Hide

```
#K-M curve stratified by gender
fit.sex <- survfit(Surv(stag, event) ~ gender, data=Employeess)
plot(fit.sex, xlab = "Months",
    ylab = "Estimated Probability of Quitting", col = c(2,4))
```



Hide

```
ggsurvplot(fit.sex,
           conf.int =T,
           xlab = "Months",
    ylab = "Estimated Probability of Survival")
```



Hide

```
#log-rank test to compare the survival of males to females, using the
#`survdiff()` function.
logrank.test <- survdiff(Surv(stag, event) ~ gender, data=Employeess)
logrank.test
```

```
Call:
survdiff(formula = Surv(stag, event) ~ gender, data = Employeess)

           N Observed Expected (O-E)^2/E (O-E)^2/V
gender=f 853      436      420     0.614      2.35
gender=m 276      135      151     1.706      2.35

 Chisq= 2.3  on 1 degrees of freedom, p= 0.1
```

Hide

```
logrank.test$pvalue
```

```
[1] 0.1254498
```

Hide

```
#Next, we fit  Cox proportional hazards models using the `coxph()`  function.
fit.cox <- coxph(Surv(stag, event) ~ gender, data=Employeess)
summary(fit.cox)
```

```
Call:
coxph(formula = Surv(stag, event) ~ gender, data = Employeess)

  n= 1129, number of events= 571

            coef exp(coef) se(coef)      z Pr(>|z|)
genderm -0.15133   0.85956  0.09898 -1.529    0.126

        exp(coef) exp(-coef) lower .95 upper .95
genderm    0.8596      1.163     0.708     1.044

Concordance= 0.516  (se = 0.01 )
Likelihood ratio test= 2.4  on 1 df,   p=0.1
Wald test            = 2.34  on 1 df,   p=0.1
Score (logrank) test = 2.34  on 1 df,   p=0.1
```

Hide

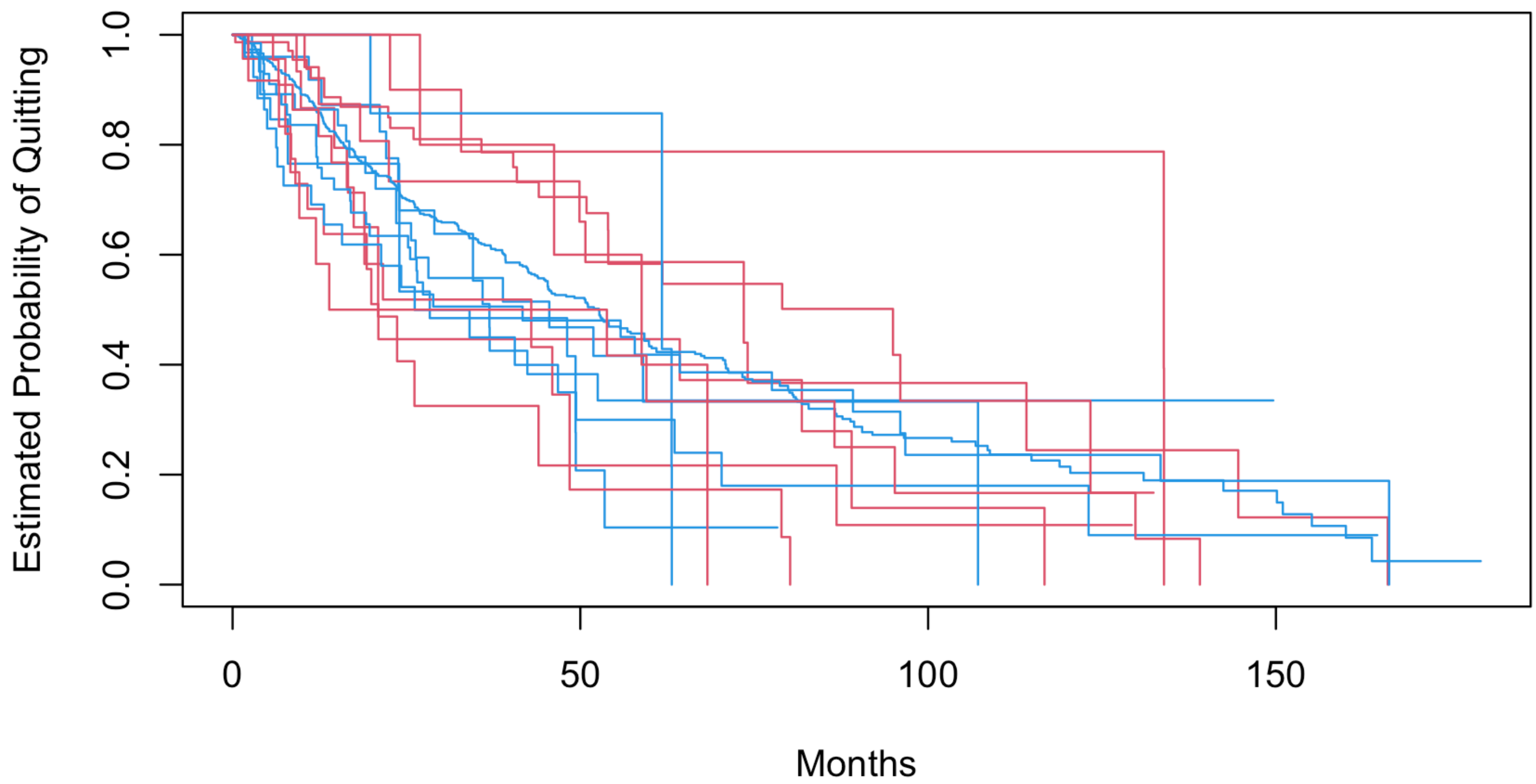
```
#Regardless of which test we use, we see that there is no clear evidence for a
#difference in survival between males and females.
```

Above we plotted a K-M curve stratified by gender and we can infer from the curve that there is not much difference between the probabilty of quitting between males and females over time.

Upon further performing a logrank test to compare survival rates of both genders, we can infer from the outcome that survival analysis of employee churn is not affected by the gender of the employee.

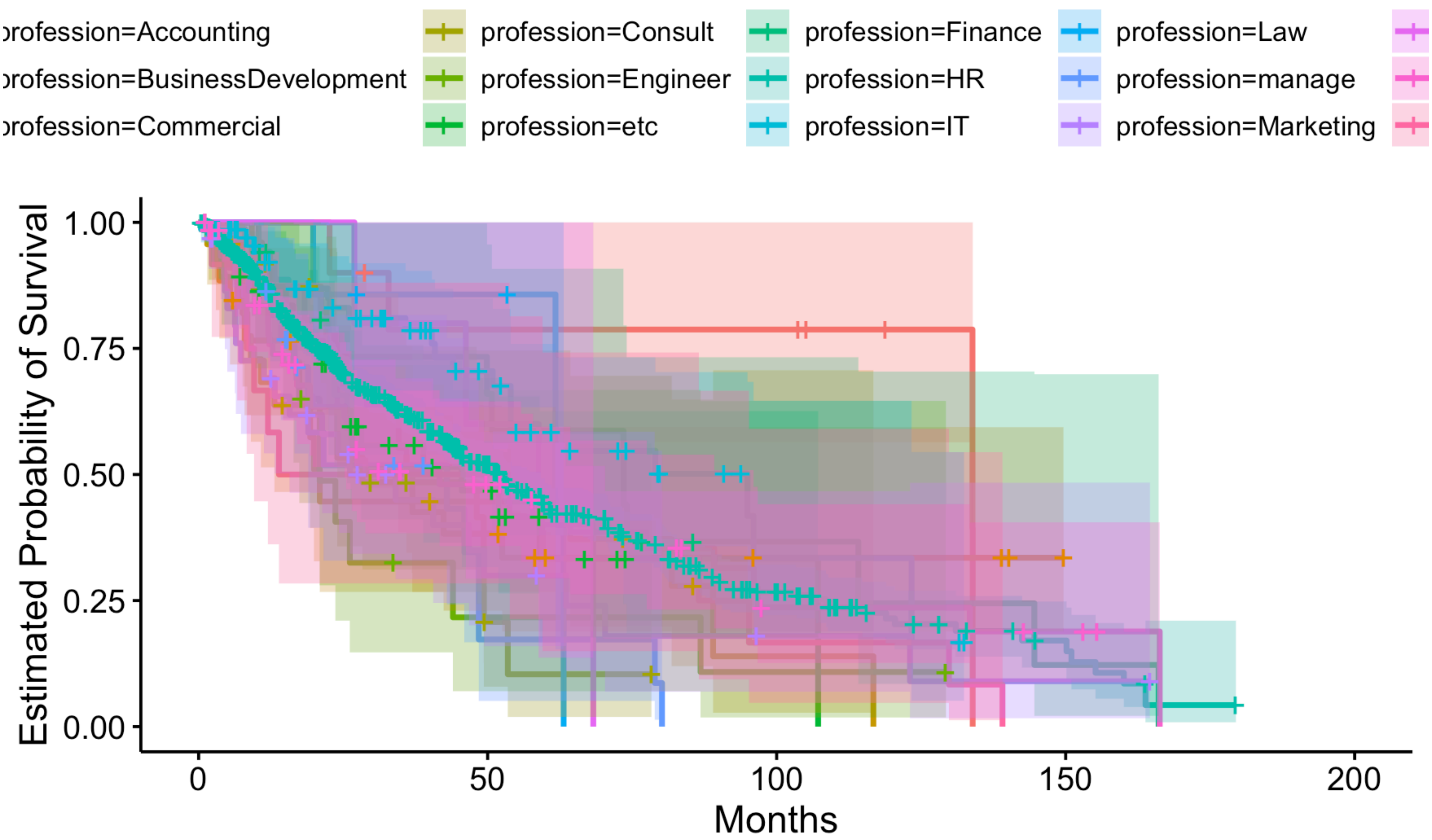
Hide

```
#K-M curve stratified by profession
fit.pr <- survfit(Surv(stag, event) ~ profession, data=Employeess)
plot(fit.pr, xlab = "Months",
    ylab = "Estimated Probability of Quitting", col = c(2,4))
```



Hide

```
ggsurvplot(fit.pr,
           conf.int =T,
           xlab = "Months",
    ylab = "Estimated Probability of Survival")
```



Hide

```
#log-rank test to compare the survival of different professions , using the `survdiff()` function.
plogrank.test <- survdiff(Surv(stag, event) ~ profession, data=Employeess)
plogrank.test
```

```
Call:
survdiff(formula = Surv(stag, event) ~ profession, data = Employeess)

                                N Observed Expected (O-E)^2/E (O-E)^2/V
profession=Accounting          10        6    12.83  3.64e+00  3.81e+00
profession=BusinessDevelopment  27       16    15.90  6.28e-04  6.51e-04
profession=Commercial          23       15     9.99  2.52e+00  2.57e+00
profession=Consult             25       16    10.32  3.13e+00  3.21e+00
profession=Engineer            15       11     6.56  3.00e+00  3.05e+00
profession=etc                 37       20    16.96  5.46e-01  5.66e-01
profession=Finance             17       12    14.51  4.34e-01  4.54e-01
profession=HR                 757      357   370.58  4.98e-01  1.43e+00
profession=IT                  74       25    39.42  5.27e+00  5.69e+00
profession=Law                  7        5     4.99  4.07e-05  4.12e-05
profession=manage              22       15     8.45  5.08e+00  5.18e+00
profession=Marketing           31       21    14.32  3.11e+00  3.21e+00
profession=PR                   6        5     3.75  4.18e-01  4.23e-01
profession=Sales               66       35    34.13  2.23e-02  2.43e-02
profession=Teaching            12       12     8.30  1.65e+00  1.69e+00

 Chisq= 29.6  on 14 degrees of freedom, p= 0.009
```

Hide

```
plogrank.test$pvalue
```

```
[1] 0.008684665
```

Hide

```
#Next, we fit  Cox proportional hazards models using the `coxph()`  function.
pfit.cox <- coxph(Surv(stag, event) ~ profession, data=Employeess)
summary(pfit.cox)
```

```
Call:
coxph(formula = Surv(stag, event) ~ profession, data = Employeess)

  n= 1129, number of events= 571

                               coef exp(coef) se(coef)     z Pr(>|z|)
professionBusinessDevelopment 0.7726    2.1653   0.4797 1.611  0.10726
professionCommercial          1.1801    3.2547   0.4851 2.433  0.01498 *
professionConsult             1.2214    3.3918   0.4824 2.532  0.01135 *
professionEngineer            1.2919    3.6396   0.5094 2.536  0.01121 *
professionetc                 0.9431    2.5678   0.4686 2.012  0.04417 *
professionFinance             0.5747    1.7767   0.5019 1.145  0.25219
professionHR                  0.7363    2.0881   0.4140 1.779  0.07532 .
professionIT                  0.3167    1.3727   0.4565 0.694  0.48779
professionLaw                 0.7786    2.1784   0.6082 1.280  0.20048
professionmanage              1.3543    3.8741   0.4864 2.784  0.00537 **
professionMarketing           1.1542    3.1714   0.4649 2.483  0.01304 *
professionPR                  1.0649    2.9007   0.6081 1.751  0.07989 .
professionSales               0.7921    2.2080   0.4435 1.786  0.07412 .
professionTeaching            1.1303    3.0965   0.5005 2.258  0.02393 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                              exp(coef) exp(-coef) lower .95 upper .95
professionBusinessDevelopment     2.165     0.4618    0.8457     5.544
professionCommercial              3.255     0.3072    1.2578     8.422
professionConsult                 3.392     0.2948    1.3176     8.731
professionEngineer                3.640     0.2748    1.3411     9.878
professionetc                     2.568     0.3894    1.0249     6.433
professionFinance                 1.777     0.5629    0.6643     4.752
professionHR                      2.088     0.4789    0.9276     4.700
professionIT                      1.373     0.7285    0.5610     3.359
professionLaw                     2.178     0.4591    0.6614     7.175
professionmanage                  3.874     0.2581    1.4932    10.051
professionMarketing               3.171     0.3153    1.2751     7.887
professionPR                      2.901     0.3447    0.8808     9.552
professionSales                   2.208     0.4529    0.9257     5.267
professionTeaching                3.097     0.3229    1.1610     8.258

Concordance= 0.558  (se = 0.011 )
Likelihood ratio test= 28.28  on 14 df,   p=0.01
Wald test            = 28.54  on 14 df,   p=0.01
Score (logrank) test = 29.5  on 14 df,   p=0.009
```

Above we plotted a K-M curve stratified by profession of the employee. We can see from the curves that employees from different professions have different probability of quitting over time, where some are decreasing rapidly (like IT, Law), some are decreasing at a normal rate over time.

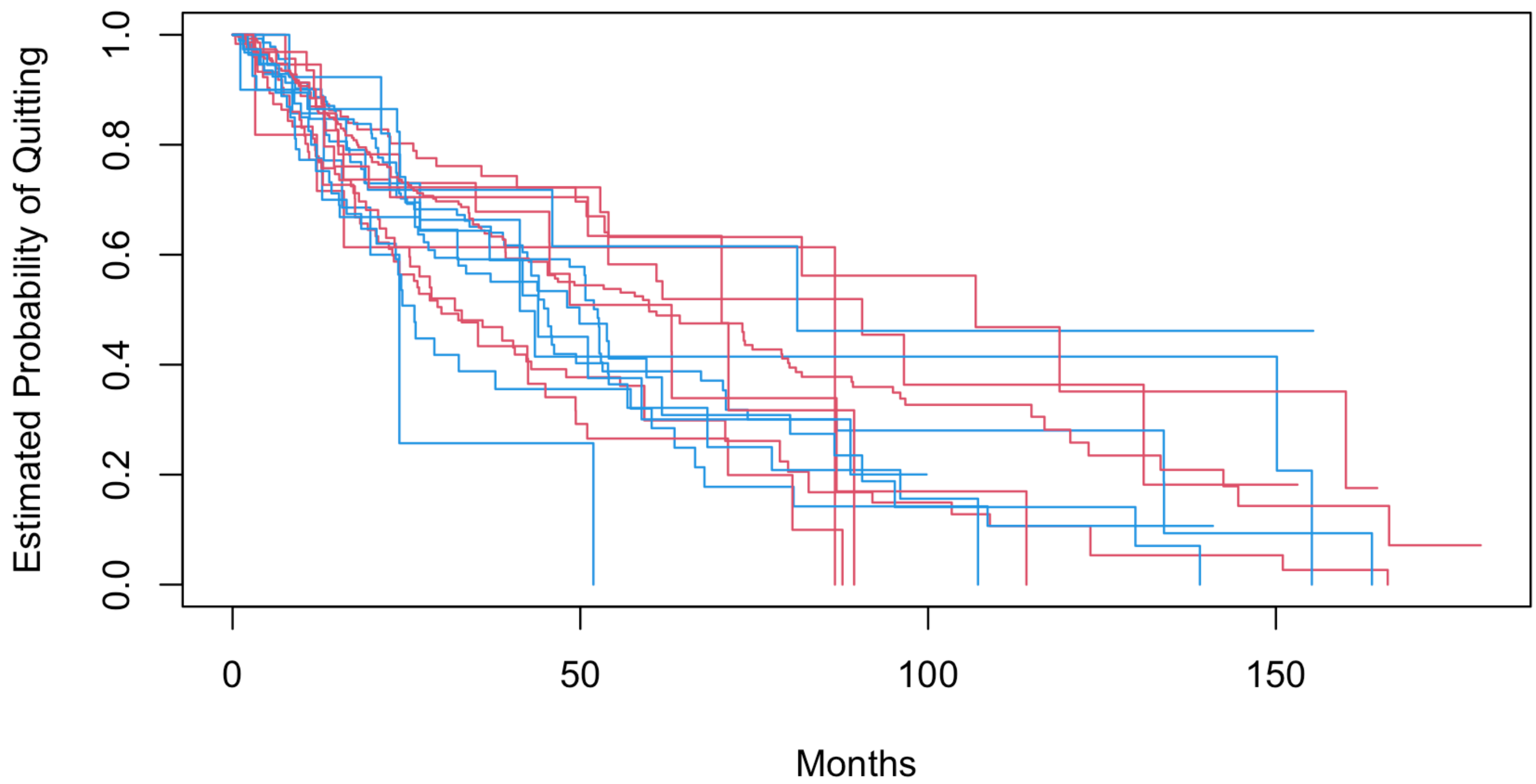
The p-value (0.0087) observed from the log-rank test tells us that profession does help in determining the survival rate of the employee as the p-value is way below 0.05.

On fitting the Cox-proportional hazard model, it will help identify the variables that are significantly associated with the survival outcome. From the summary of the model we can see the coefficients and the p-value of different professions and infer that (the larger the coefficient and lower the p-value, the variable has more impact on the final outcome). Hence we can say that, employees from management, marketing, consulting, engineering and teaching have a higher risk of quitting compared to others.

Here the above outcomes are not completely accurate, because above we saw the number of people for each profession are not distributed equally as employees from HR are considerably more than employees from other profession.

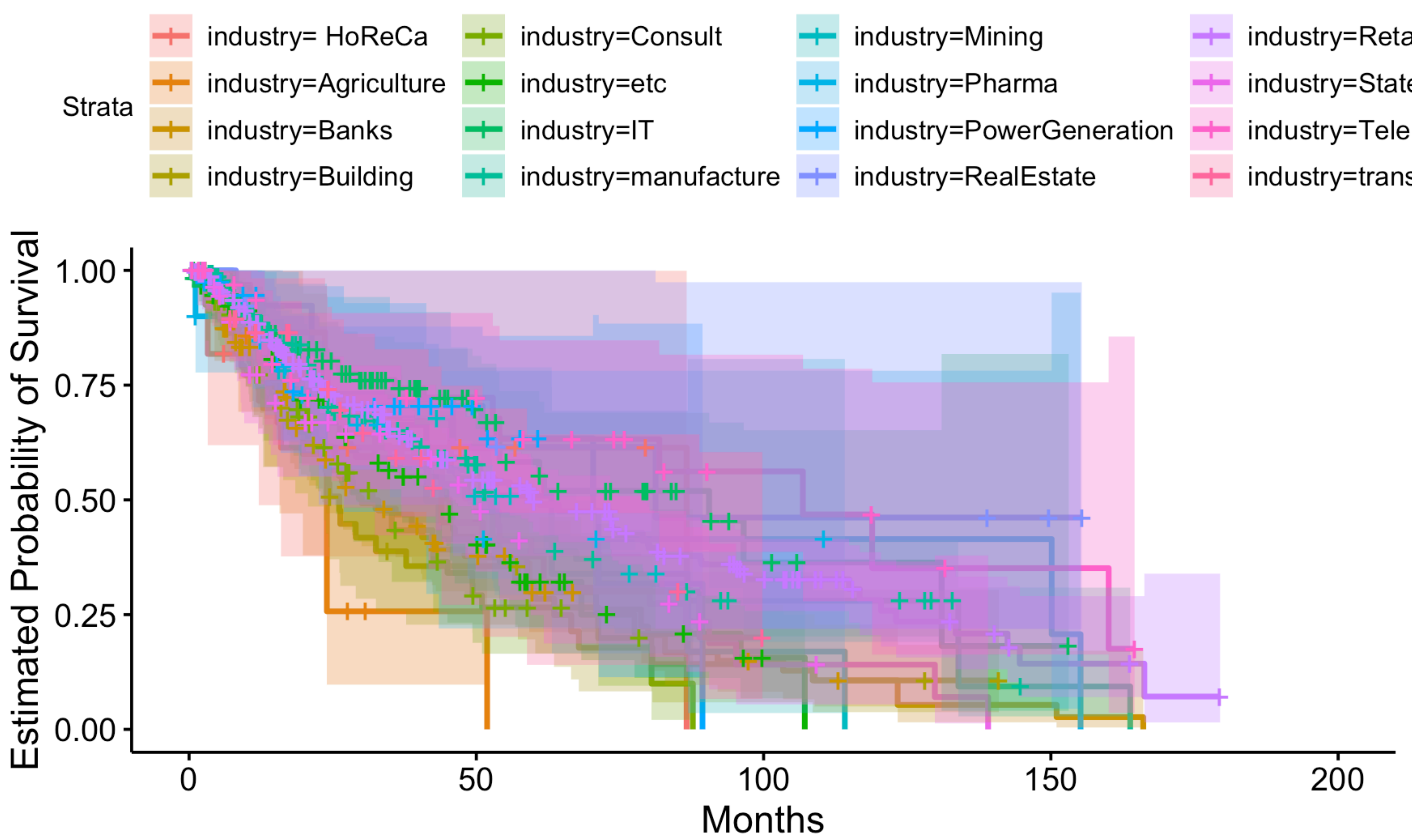
Hide

```
#K-M curve stratified by industry
fit.ind <- survfit(Surv(stag, event) ~ industry, data=Employeess)
plot(fit.ind, xlab = "Months",
    ylab = "Estimated Probability of Quitting", col = c(2,4))
```



Hide

```
ggsurvplot(fit.ind,
           conf.int =T,
           xlab = "Months",
    ylab = "Estimated Probability of Survival")
```



Hide

```
#log-rank test to compare the survival various industries, using the `survdiff()` function.
ilogrank.test <- survdiff(Surv(stag, event) ~ industry, data=Employeess)
ilogrank.test
```

```
Call:
survdiff(formula = Surv(stag, event) ~ industry, data = Employeess)

                           N Observed Expected (O-E)^2/E (O-E)^2/V
industry= HoReCa          11        6     5.79   0.00779   0.00789
industry=Agriculture      15       10     4.16   8.21470   8.32018
industry=Banks           114       75    51.11  11.17213  12.32365
industry=Building         41       31    20.59   5.26931   5.49241
industry=Consult          74       45    28.17  10.06064  10.69575
industry=etc              94       54    43.61   2.47783   2.70780
industry=IT              122       34    53.53   7.12660   7.90274
industry=manufacture     145       70    76.37   0.53160   0.61763
industry=Mining           24       14    13.60   0.01187   0.01220
industry=Pharma           20       11    11.93   0.07234   0.07466
industry=PowerGeneration  38       15    18.51   0.66687   0.69303
industry=RealEstate       13        5    10.57   2.93788   3.04576
industry=Retail          289      136   165.06   5.11503   7.30153
industry=State            55       35    27.97   1.76601   1.86333
industry=Telecom          36       14    24.50   4.50264   4.77265
industry=transport        38       16    15.54   0.01341   0.01385

 Chisq= 60.9  on 15 degrees of freedom, p= 2e-07
```

Hide

```
ilogrank.test$pvalue
```

```
[1] 1.740932e-07
```

Hide

```
#Next, we fit  Cox proportional hazards models using the `coxph()`  function.
ifit.cox <- coxph(Surv(stag, event) ~ industry, data=Employeess)
summary(ifit.cox)
```

```
Call:
coxph(formula = Surv(stag, event) ~ industry, data = Employeess)

  n= 1129, number of events= 571

                              coef  exp(coef)   se(coef)      z Pr(>|z|)
industryAgriculture      0.8687904  2.3840255  0.5179957  1.677   0.0935 .
industryBanks            0.3451907  1.4122593  0.4247577  0.813   0.4164
industryBuilding         0.3681991  1.4451297  0.4466633  0.824   0.4098
industryConsult          0.4463360  1.5625763  0.4352886  1.025   0.3052
industryetc              0.1862531  1.2047271  0.4307377  0.432   0.6654
industryIT              -0.4878455  0.6139477  0.4431102 -1.101   0.2709
industrymanufacture     -0.1262863  0.8813625  0.4259822 -0.296   0.7669
industryMining          -0.0043689  0.9956406  0.4881063 -0.009   0.9929
industryPharma          -0.1375448  0.8714953  0.5091819 -0.270   0.7871
industryPowerGeneration -0.2379344  0.7882544  0.4834113 -0.492   0.6226
industryRealEstate      -0.8176897  0.4414504  0.6076191 -1.346   0.1784
industryRetail          -0.2386062  0.7877251  0.4176653 -0.571   0.5678
industryState            0.1857274  1.2040939  0.4421011  0.420   0.6744
industryTelecom         -0.6189377  0.5385162  0.4890460 -1.266   0.2057
industrytransport        0.0003095  1.0003096  0.4791301  0.001   0.9995
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                        exp(coef) exp(-coef) lower .95 upper .95
industryAgriculture        2.3840     0.4195    0.8638     6.580
industryBanks              1.4123     0.7081    0.6143     3.247
industryBuilding           1.4451     0.6920    0.6022     3.468
industryConsult            1.5626     0.6400    0.6658     3.667
industryetc                1.2047     0.8301    0.5179     2.802
industryIT                 0.6139     1.6288    0.2576     1.463
industrymanufacture        0.8814     1.1346    0.3824     2.031
industryMining             0.9956     1.0044    0.3825     2.592
industryPharma             0.8715     1.1475    0.3213     2.364
industryPowerGeneration    0.7883     1.2686    0.3056     2.033
industryRealEstate         0.4415     2.2653    0.1342     1.452
industryRetail             0.7877     1.2695    0.3474     1.786
industryState              1.2041     0.8305    0.5062     2.864
industryTelecom            0.5385     1.8570    0.2065     1.404
industrytransport          1.0003     0.9997    0.3911     2.558

Concordance= 0.58  (se = 0.014 )
Likelihood ratio test= 57.67  on 15 df,   p=6e-07
Wald test            = 58.23  on 15 df,   p=5e-07
Score (logrank) test = 60.97  on 15 df,   p=2e-07
```

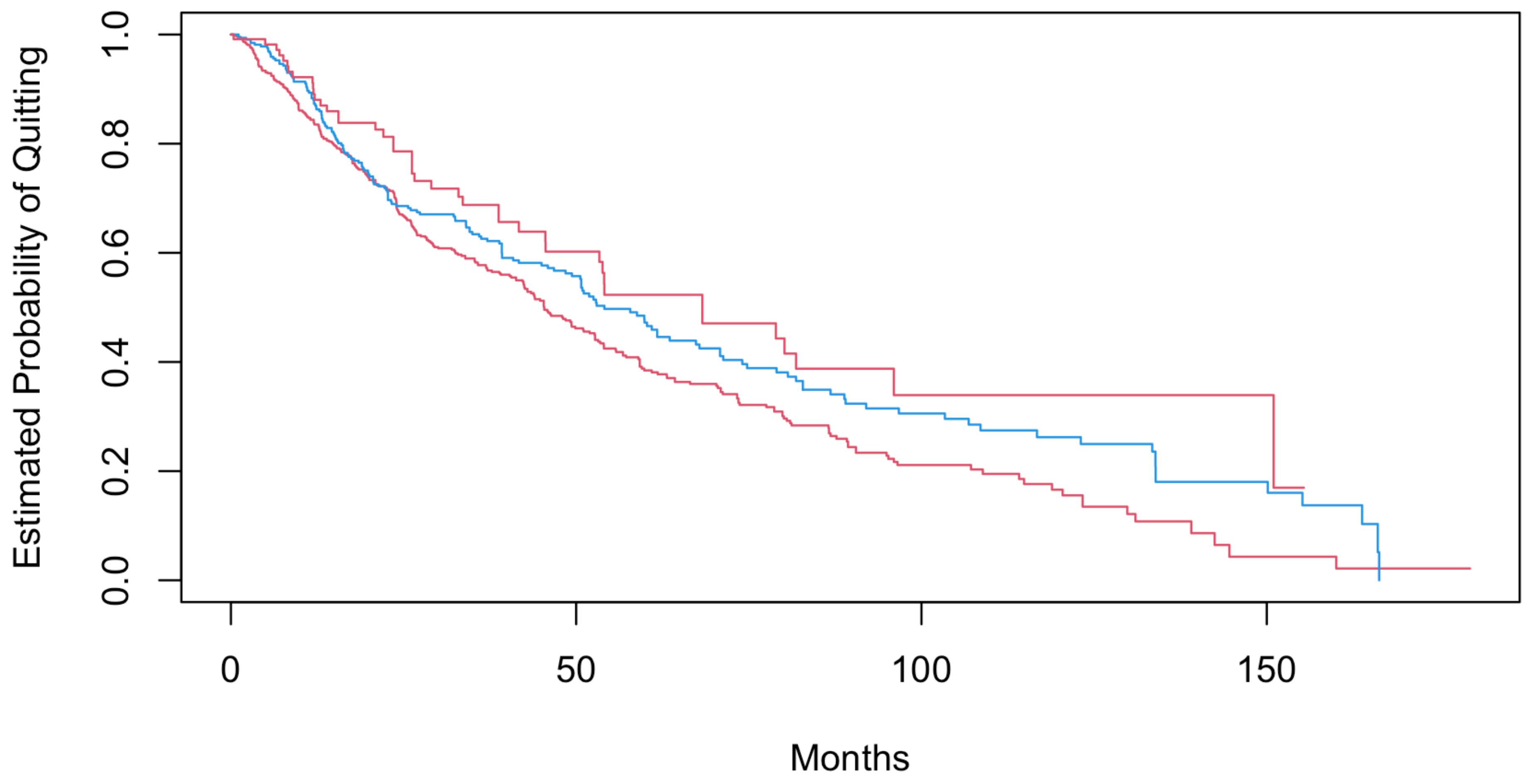
Above we plotted a K-M curve stratified by industry of the employee. We can see from the curves that employees from different industries have different probability of quitting over time, where some are decreasing rapidly (Agriculture), some are decreasing at a normal rate over time and some industries (Retail) remain constant after some time period.

The p-value (1.740932e-07) observed from the log-rank test tells us that industry does help in determining the survival rate of the employee as the p-value is way below 0.05.

On fitting the Cox-proportional hazard model, it will help identify the variables that are significantly associated with the survival outcome. From the summary of the model we can see the coefficients and the p-value of different industries and infer that (the larger the coefficient and lower the p-value, the variable has more impact on the final outcome). Hence we can say that, employees from Real Estate, Telecom and Retail industry do not have a higher risk of quitting compared to other industries.

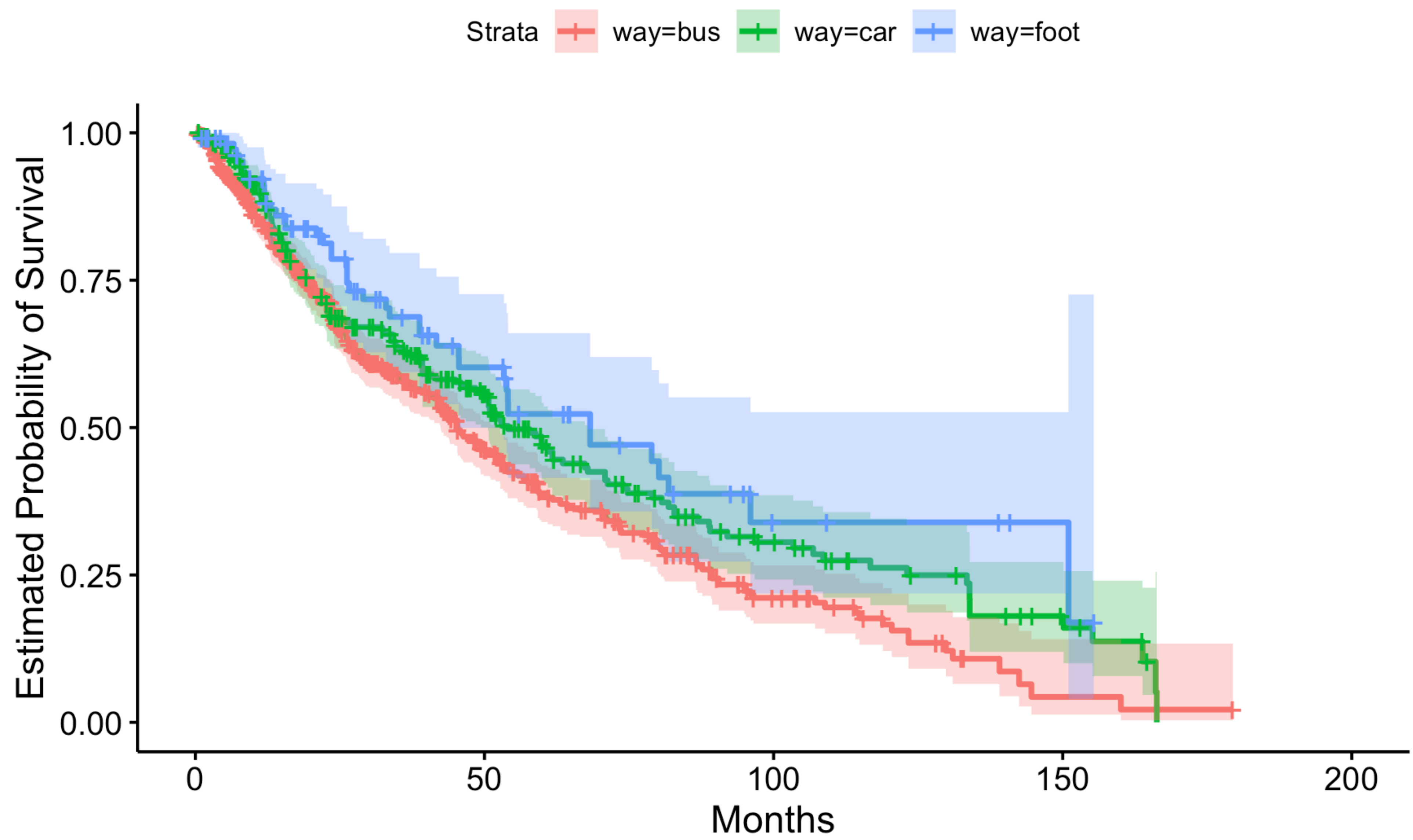
Hide

```
#K-M curve stratified by way of transportation
fit.way <- survfit(Surv(stag, event) ~ way, data=Employeess)
plot(fit.way, xlab = "Months",
    ylab = "Estimated Probability of Quitting", col = c(2,4))
```



Hide

```
ggsurvplot(fit.way,
           conf.int =T,
           xlab = "Months",
    ylab = "Estimated Probability of Survival")
```



Hide

```
#log-rank test to compare the survival of males to females, using the `survdiff()` function.
wlogrank.test <- survdiff(Surv(stag, event) ~ way, data=Employeess)
wlogrank.test
```

```
Call:
survdiff(formula = Surv(stag, event) ~ way, data = Employeess)

           N Observed Expected (O-E)^2/E (O-E)^2/V
way=bus  681      354    316.7      4.39     10.06
way=car  331      174    194.0      2.06      3.19
way=foot 117       43     60.3      4.95      5.55

 Chisq= 11.6  on 2 degrees of freedom, p= 0.003
```

Hide

```
wlogrank.test$pvalue
```

```
[1] 0.003061535
```

Hide

```
#Next, we fit  Cox proportional hazards models using the `coxph()`  function.
wfit.cox <- coxph(Surv(stag, event) ~ way, data=Employeess)
summary(wfit.cox)
```

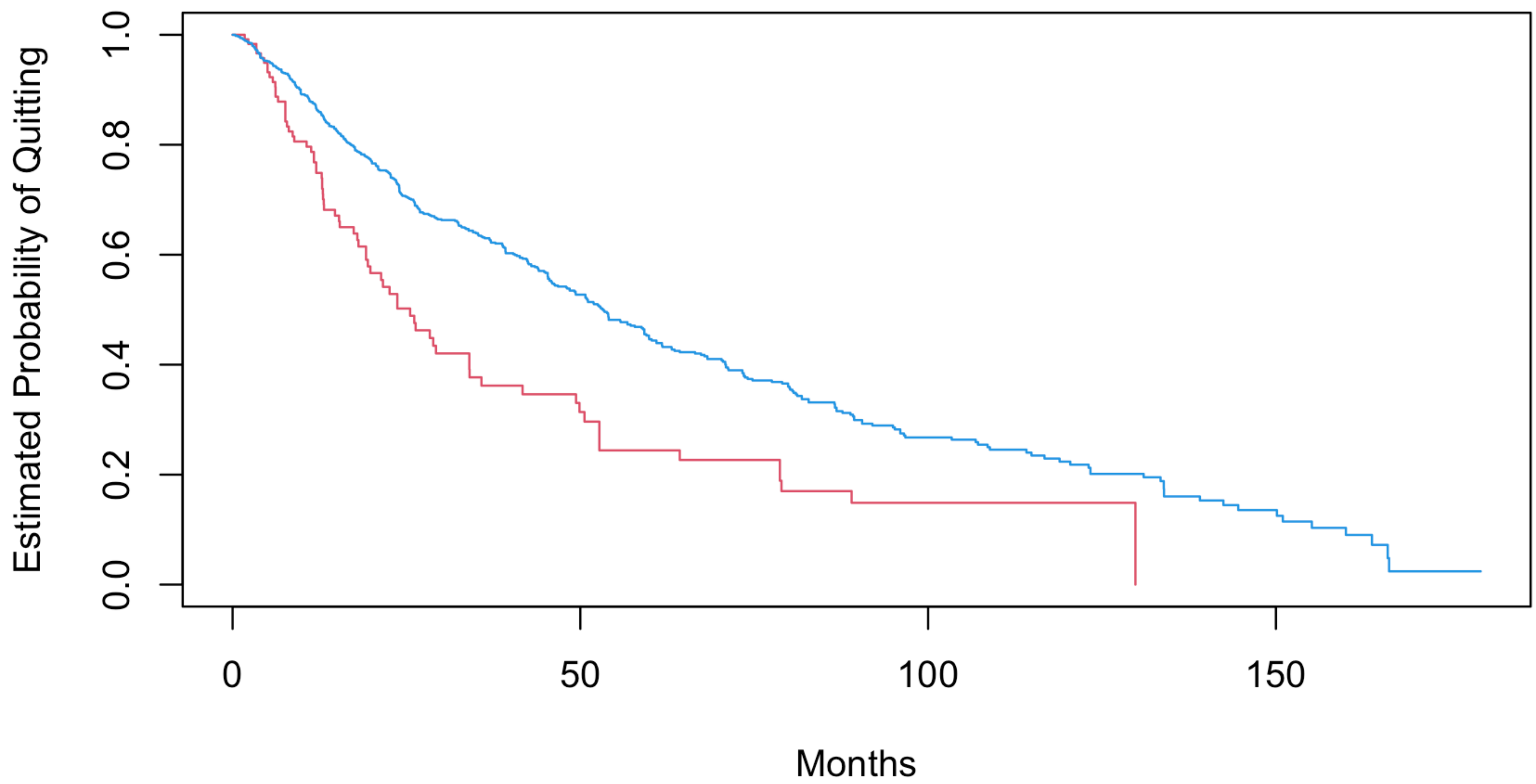
```
Call:
coxph(formula = Surv(stag, event) ~ way, data = Employeess)

  n= 1129, number of events= 571

            coef exp(coef) se(coef)      z Pr(>|z|)
waycar  -0.22575   0.79792  0.09363 -2.411  0.01590 *
wayfoot -0.45301   0.63571  0.16180 -2.800  0.00511 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
waycar     0.7979      1.253    0.6641    0.9586
wayfoot    0.6357      1.573    0.4629    0.8729

Concordance= 0.533  (se = 0.012 )
Likelihood ratio test= 12.06  on 2 df,   p=0.002
Wald test            = 11.47  on 2 df,   p=0.003
Score (logrank) test = 11.59  on 2 df,   p=0.003
```

Above we plotted a K-M curve stratified by way of transportation of the employee. We can see from the curves that employees having different ways of transportation do not have much difference in rate of survival probability reduction.

The p-value (0.003) observed from the log-rank test tells us that way of transportation does help in determining the survival rate of the employee as the p-value is below 0.05.

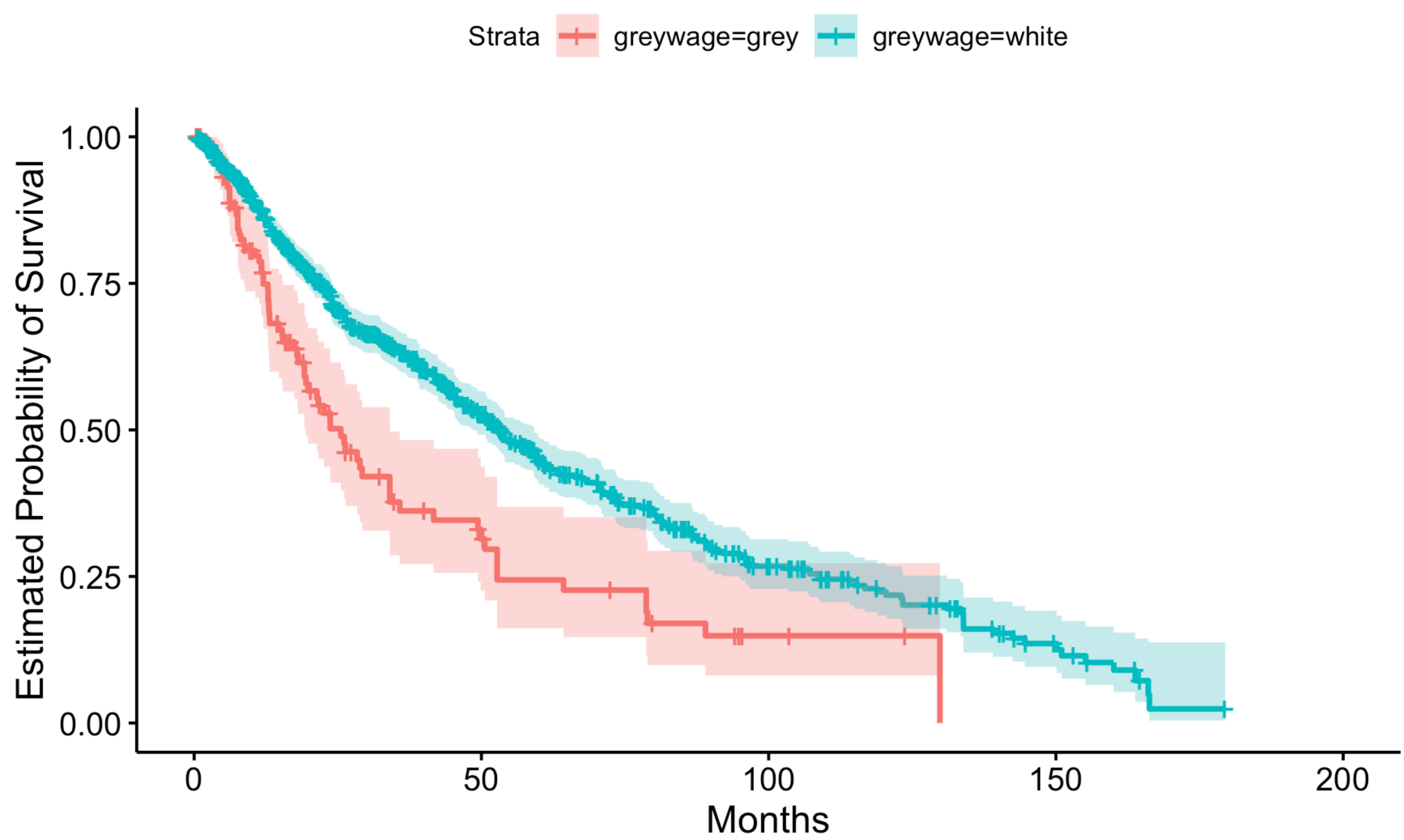
Hide

```
#K-M curve stratified by employee wages.
fit.wage <- survfit(Surv(stag, event) ~ greywage, data=Employeess)
plot(fit.wage, xlab = "Months",
    ylab = "Estimated Probability of Quitting", col = c(2,4))
```



Hide

```
ggsurvplot(fit.wage,
           conf.int =T,
           xlab = "Months",
    ylab = "Estimated Probability of Survival")
```



Hide

```
#log-rank test to compare the survival employee wage, using the `survdiff()` function.
wglogrank.test <- survdiff(Surv(stag, event) ~ greywage, data=Employeess)
wglogrank.test
```

```
Call:
survdiff(formula = Surv(stag, event) ~ greywage, data = Employeess)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
greywage=grey   127       73     43.2     20.47      22.3
greywage=white 1002      498    527.8      1.68      22.3

 Chisq= 22.3  on 1 degrees of freedom, p= 2e-06
```

Hide

```
wglogrank.test$pvalue
```

```
[1] 2.27932e-06
```

Above we plotted a K-M curve stratified by employee wage type. We can see from the curves that employees having grey wage and white wage have different rate of probability of survival reduction over time. We can see greywage employees quitting earlier than white wage employees.

The p-value (2.27932e-06) observed from the log-rank test tells us that employee wage does help in determining the survival rate of the employee as the p-value is below 0.05.

Hide

```
fit.all <- coxph(Surv(stag, event) ~ gender + profession + industry + way, data=Employeess)
summary(fit.all)
```

```
Call:
coxph(formula = Surv(stag, event) ~ gender + profession + industry + 
    way, data = Employeess)

  n= 1129, number of events= 571 

                                 coef exp(coef) se(coef)      z Pr(>|z|)    
genderm                      -0.17541   0.83912  0.11854 -1.480 0.138963    
professionBusinessDevelopment  0.85119   2.34244  0.49125  1.733 0.083150 .  
professionCommercial          1.27599   3.58223  0.49325  2.587 0.009684 ** 
professionConsult             0.67774   1.96942  0.50732  1.336 0.181578    
professionEngineer            1.07415   2.92750  0.51979  2.067 0.038781 *  
professionetc                 0.69128   1.99627  0.47975  1.441 0.149609    
professionFinance             0.30228   1.35294  0.51265  0.590 0.555430    
professionHR                  0.49326   1.63765  0.42469  1.161 0.245454    
professionIT                  0.30359   1.35471  0.47031  0.646 0.518597    
professionLaw                 0.54049   1.71684  0.63843  0.847 0.397225    
professionmanage              1.29775   3.66104  0.49796  2.606 0.009158 ** 
professionMarketing           0.92119   2.51228  0.47484  1.940 0.052377 .  
professionPR                  0.94254   2.56649  0.62793  1.501 0.133346    
professionSales               0.78789   2.19874  0.45264  1.741 0.081743 .  
professionTeaching            0.68350   1.98079  0.55739  1.226 0.220106    
industryAgriculture           0.70407   2.02196  0.53819  1.308 0.190803    
industryBanks                 0.34657   1.41421  0.42548  0.815 0.415336    
industryBuilding              0.36855   1.44564  0.44934  0.820 0.412100    
industryConsult               0.41081   1.50804  0.44026  0.933 0.350758    
industryetc                   0.07040   1.07294  0.43573  0.162 0.871638    
industryIT                   -0.46401   0.62876  0.44492 -1.043 0.296995    
industrymanufacture          -0.12459   0.88286  0.42886 -0.291 0.771424    
industryMining               -0.06952   0.93285  0.50304 -0.138 0.890090    
industryPharma               -0.20312   0.81618  0.51063 -0.398 0.690781    
industryPowerGeneration      -0.29852   0.74192  0.48512 -0.615 0.538329    
industryRealEstate           -0.93382   0.39305  0.62429 -1.496 0.134700    
industryRetail               -0.33339   0.71649  0.42035 -0.793 0.427707    
industryState                 0.03227   1.03280  0.46616  0.069 0.944802    
industryTelecom              -0.66314   0.51523  0.49254 -1.346 0.178186    
industrytransport            -0.13737   0.87165  0.48201 -0.285 0.775646    
waycar                       -0.17417   0.84015  0.10104 -1.724 0.084749 .  
wayfoot                      -0.57211   0.56433  0.16791 -3.407 0.000656 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
                             exp(coef) exp(-coef) lower .95 upper .95
genderm                         0.8391     1.1917    0.6651    1.0586
professionBusinessDevelopment   2.3424     0.4269    0.8944    6.1351
professionCommercial            3.5822     0.2792    1.3624    9.4191
professionConsult               1.9694     0.5078    0.7286    5.3232
professionEngineer              2.9275     0.3416    1.0569    8.1085
professionetc                   1.9963     0.5009    0.7796    5.1119
professionFinance               1.3529     0.7391    0.4953    3.6953
professionHR                    1.6377     0.6106    0.7124    3.7646
professionIT                    1.3547     0.7382    0.5389    3.4055
professionLaw                   1.7168     0.5825    0.4912    6.0002
professionmanage                3.6610     0.2731    1.3795    9.7157
professionMarketing             2.5123     0.3980    0.9906    6.3716
professionPR                    2.5665     0.3896    0.7496    8.7868
professionSales                 2.1987     0.4548    0.9055    5.3390
professionTeaching              1.9808     0.5048    0.6643    5.9059
industryAgriculture             2.0220     0.4946    0.7041    5.8061
industryBanks                   1.4142     0.7071    0.6142    3.2560
industryBuilding                1.4456     0.6917    0.5992    3.4877
industryConsult                 1.5080     0.6631    0.6363    3.5740
industryetc                     1.0729     0.9320    0.4568    2.5204
industryIT                      0.6288     1.5904    0.2629    1.5038
industrymanufacture             0.8829     1.1327    0.3809    2.0461
industryMining                  0.9328     1.0720    0.3480    2.5003
industryPharma                  0.8162     1.2252    0.3000    2.2204
industryPowerGeneration         0.7419     1.3479    0.2867    1.9200
industryRealEstate              0.3930     2.5442    0.1156    1.3361
industryRetail                  0.7165     1.3957    0.3143    1.6331
industryState                   1.0328     0.9682    0.4142    2.5752
industryTelecom                 0.5152     1.9409    0.1962    1.3529
industrytransport               0.8716     1.1473    0.3389    2.2420
waycar                          0.8402     1.1903    0.6892    1.0242
wayfoot                         0.5643     1.7720    0.4061    0.7843

Concordance= 0.612  (se = 0.013 )
Likelihood ratio test= 97.32  on 32 df,   p=2e-08
Wald test            = 96.36  on 32 df,   p=2e-08
Score (logrank) test = 100.7  on 32 df,   p=5e-09
```

Above code helps us in understanding and identifying the variables that are significantly associated with the survival outcome.