CSDA1050 Capstone Project

Eugene Yong Geun Park

2019-07-17

Background

As my proposal suggests, this project focuses on this company's past 10-year employee data to discover patterns in employment.

Data Exploration

First starting with HR data which consists of employees' personal information and employment information.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(highcharter)
```

```
## Warning: package 'highcharter' was built under R version 3.5.2
```

```
## Highcharts (www.highcharts.com) is a Highsoft software product which is
```

```
## not free for commercial and Governmental use
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#importing files I will work with
hr<-read.csv("~/Desktop/CSDA-1050F18S1/eugenepark/CSDA1050HR.csv")
```

```
#Examine the HR dataset
#S.EMP. - A unique employee number assigned to each employee
#C.LEVEL - Employee's job level ranging from C1 being lowest & C6 highest
#Team - This company consists of 5 different teams/department
#Raise - # of raises received
#Term..Type - Among employees who left, voluntary leaves are shown as "Quit" & involuntary leaves
as "Terminated". Active employees are left blank.
#Dist..To.work - Distance to work from home in km
#I believe the rest are pretty self explantory
head(hr)
```

```
##     S.EMP.                    Title C.LEVEL              Team  Job.Level
## 1 60060523     Senior Art Director      C3 Creative Services Supervisor
## 2 60071662 Asso. Creative Director      C3 Creative Services Supervisor
## 3 60072318   Sr. Account Executive      C2  Account Services Supervisor
## 4 60072838   Engagement Supervisor      C3  Account Services Supervisor
## 5 60081603     Junior Art Director      C2        Production  Associate
## 6 60114364     Account Coordinator      C1  Account Services  Associate
##   BEGIN.SALARY Raise     Hire H.Year Termination T.Year TermType
## 1        70000     1 20060821   2006    20110831   2011     Quit
## 2        80000     2 20070522   2007    20110520   2011     Quit
## 3        70000     1 20070816   2007    20120323   2012     Quit
## 4        75000  -    20070910   2007    20120120   2010     Quit
## 5        35000     2 20080520   2008    20140919   2014     Quit
## 6        35000     2 20110501   2011    20140314   2014     Quit
##       Status      DOB SEX                 Education
## 1 Terminated 19740928   M          Bachelors Degree
## 2 Terminated 19660123   F          Bachelors Degree
## 3 Terminated 19800925   M          Bachelors Degree
## 4 Terminated 19731212   F          Bachelors Degree
## 5 Terminated 19811224   M          Bachelors Degree
## 6 Terminated 19830126   F College/Diploma/Associate
##                          Major DistToWork
## 1 Advertising & Graphic Design       15.0
## 2           English & Religion       40.0
## 3                                    25.0
## 4                    Sociology       20.0
## 5              Graphic Design       13.3
## 6                                    24.5
```

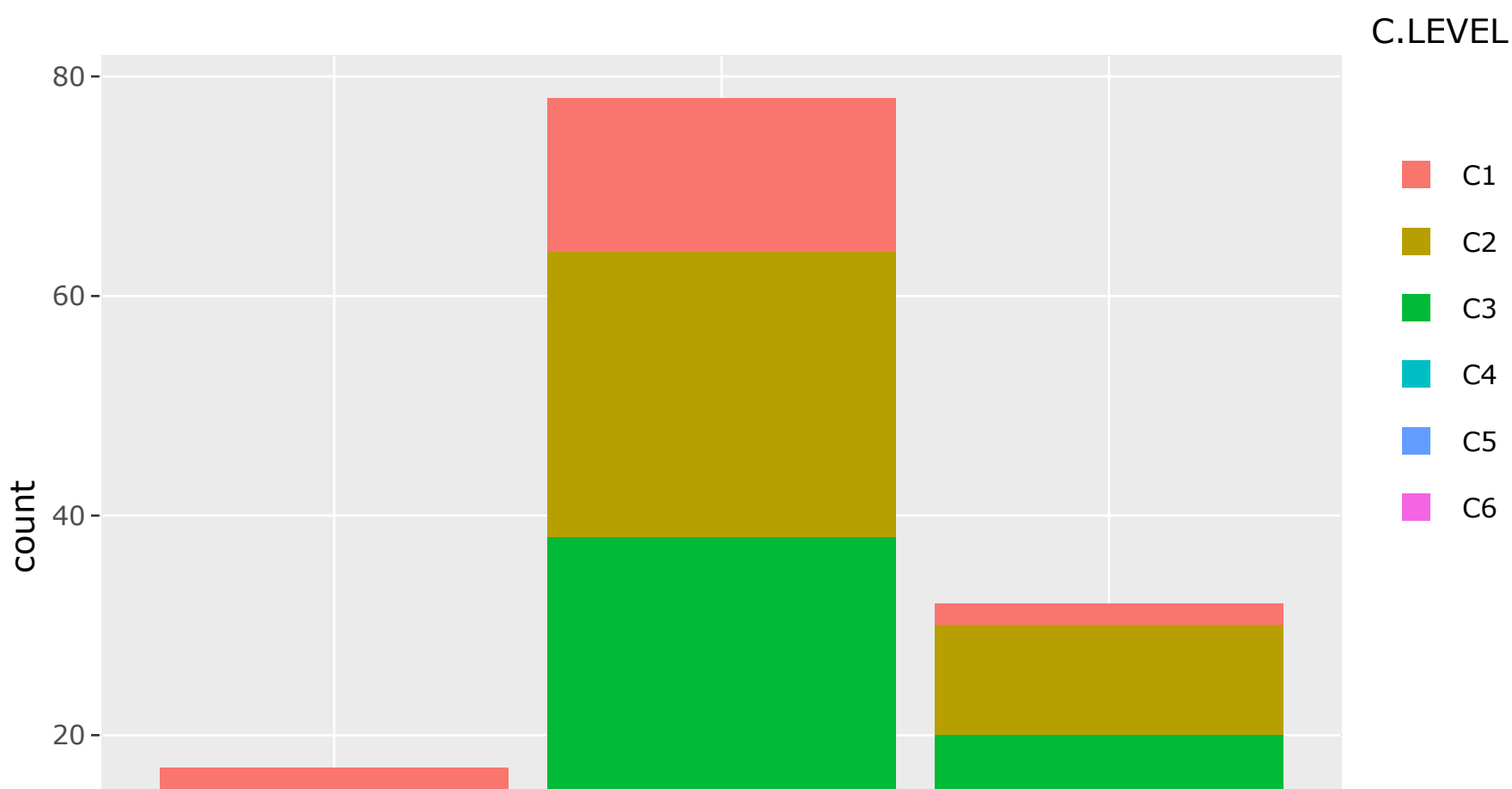
```
summary(hr)
```

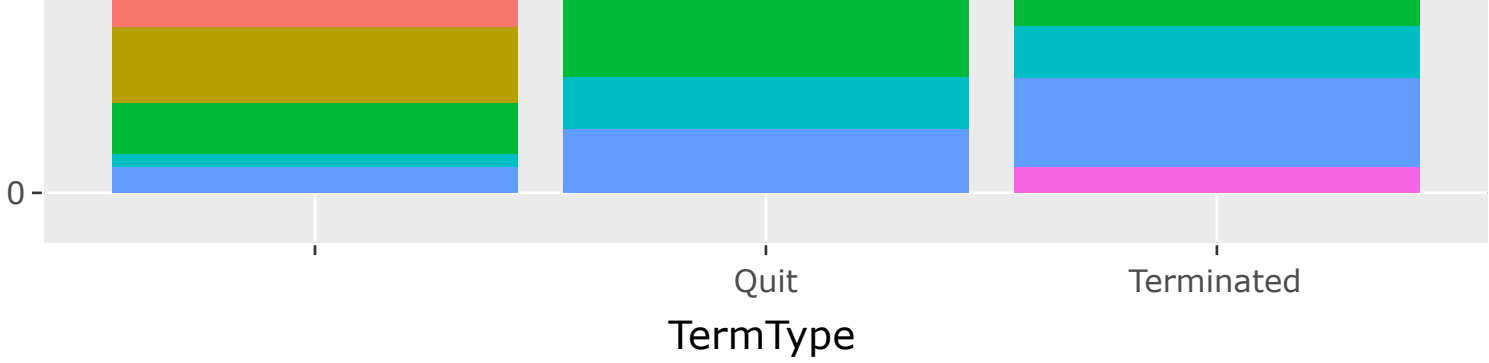
```
##      S.EMP.                           Title    C.LEVEL
##  Min.   :20001890   Account Executive  : 9   C1 :20
##  1st Qu.:60124826   ART DIRECTOR       : 8   C2 :42
##  Median :60140323   ACCOUNT SUPERVISOR : 5   C3 :40
##  Mean   :60611386   WRITER             : 5   C4 : 9
##  3rd Qu.:60160768   Account Coordinator: 4   C5 :14
##  Max.   :90150168   Account Supervisor : 4   C6 : 2
##                     (Other)            :92
##                  Team                        Job.Level   BEGIN.SALARY
##  Account Services :47   Associate              :67   Min.   : 30000
##  Corporate Mgmt   : 8   Manager                : 7   1st Qu.: 45000
##  Creative Services:38   Representative Officer: 2   Median : 60000
##  Planning         :19   Sr.Manager             :22   Mean   : 72367
##  Production       :15   Supervisor             :29   3rd Qu.: 82500
##                                                     Max.   :250000
##
##    Raise          Hire                H.Year      Termination
##   -   :76   Min.   :20060515   Min.   :2006   Min.   :20110520
##  1    :31   1st Qu.:20120410   1st Qu.:2012   1st Qu.:20140228
##  2    :15   Median :20140310   Median :2014   Median :20150329
##  3    : 4   Mean   :20137445   Mean   :2014   Mean   :20151022
##  4    : 1   3rd Qu.:20160110   3rd Qu.:2016   3rd Qu.:20170527
##             Max.   :20190612   Max.   :2019   Max.   :20190630
##                                               NA's   :17
##      T.Year            TermType          Status          DOB            SEX
##  Min.   :2010              :17   Active    : 17   Min.   :19601005   F:66
##  1st Qu.:2014   Quit       :78   Terminated:110   1st Qu.:19750516   M:61
##  Median :2015   Terminated:32                    Median :19830205
##  Mean   :2015                                    Mean   :19816494
##  3rd Qu.:2017                                    3rd Qu.:19871206
##  Max.   :2019                                    Max.   :19950216
##  NA's   :17
##                       Education                       Major
##  Bachelors Degree          :85   Advertising             :21
##  College/Diploma/Associate:33                            :10
##  Masters Degree            : 8   Design and Applied Arts: 9
##  MBA                       : 1   Marketing&Sales         : 8
##                                  Business Communications: 4
##                                  Commerce                : 4
##                                  (Other)                 :71
##    DistToWork
##  Min.   :  0.40
##  1st Qu.:  5.00
##  Median :  8.80
##  Mean   : 16.74
##  3rd Qu.: 20.30
##  Max.   :111.00
##
```

```
str(hr)
```

```
## 'data.frame':    127 obs. of  18 variables:
##  $ S.EMP.      : int  60060523 60071662 60072318 60072838 60081603 60114364 60115474 60116117
60117843 60117854 ...
##  $ Title       : Factor w/ 77 levels "Account Coordinator",..: 59 18 64 38 50 1 52 70 49 24 ..
.
##  $ C.LEVEL     : Factor w/ 6 levels "C1","C2","C3",..: 3 3 2 3 2 1 3 3 2 2 ...
##  $ Team        : Factor w/ 5 levels "Account Services",..: 3 3 1 1 5 1 4 5 3 3 ...
##  $ Job.Level   : Factor w/ 5 levels "Associate","Manager",..: 5 5 5 5 1 1 5 5 1 1 ...
##  $ BEGIN.SALARY: int  70000 80000 70000 75000 35000 35000 50000 75000 45000 45000 ...
##  $ Raise       : Factor w/ 5 levels " -   ","1","2",..: 2 3 2 1 3 3 1 2 1 1 ...
##  $ Hire        : int  20060821 20070522 20070816 20070910 20080520 20110501 20110725 20110822
20111216 20111216 ...
##  $ H.Year      : int  2006 2007 2007 2007 2008 2011 2011 2011 2011 2011 ...
##  $ Termination : int  20110831 20110520 20120323 20120120 20140919 20140314 20120328 20140815
20120413 20130405 ...
##  $ T.Year      : int  2011 2011 2012 2010 2014 2014 2012 2014 2012 2013 ...
##  $ TermType    : Factor w/ 3 levels "","Quit","Terminated": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Status      : Factor w/ 2 levels "Active","Terminated": 2 2 2 2 2 2 2 2 2 2 ...
##  $ DOB         : int  19740928 19660123 19800925 19731212 19811224 19830126 19830712 19721010
19790830 19830218 ...
##  $ SEX         : Factor w/ 2 levels "F","M": 2 1 2 1 2 1 1 2 1 2 ...
##  $ Education   : Factor w/ 4 levels "Bachelors Degree",..: 1 1 1 1 1 2 1 1 1 1 ...
##  $ Major       : Factor w/ 51 levels "","Accounting and Finance",..: 4 29 1 50 34 1 43 33 20 3
0 ...
##  $ DistToWork  : num  15 40 25 20 13.3 24.5 23 45.3 8 5.2 ...
```

```
#Plotting a bar chart to see if there's a particular group of job level that stands out
#Please note that I am using Plotly library to take advantage of its interactive functionality (i
.e. hover-over data display, etc)
#While each level can be considered as 2-3 years experience accumulatively, C2 & C3 level seem to
be harder to be retained (most "Quit"" job levels)
#Among employees terminated, again C2 & C3 level take up more than 50%. I also think it's interes
ting that C5 level takes sizable portion in "terminated" group.
a<-ggplot(hr, aes(x=TermType, fill=C.LEVEL))+
  geom_bar()
ggplotly(a)
```

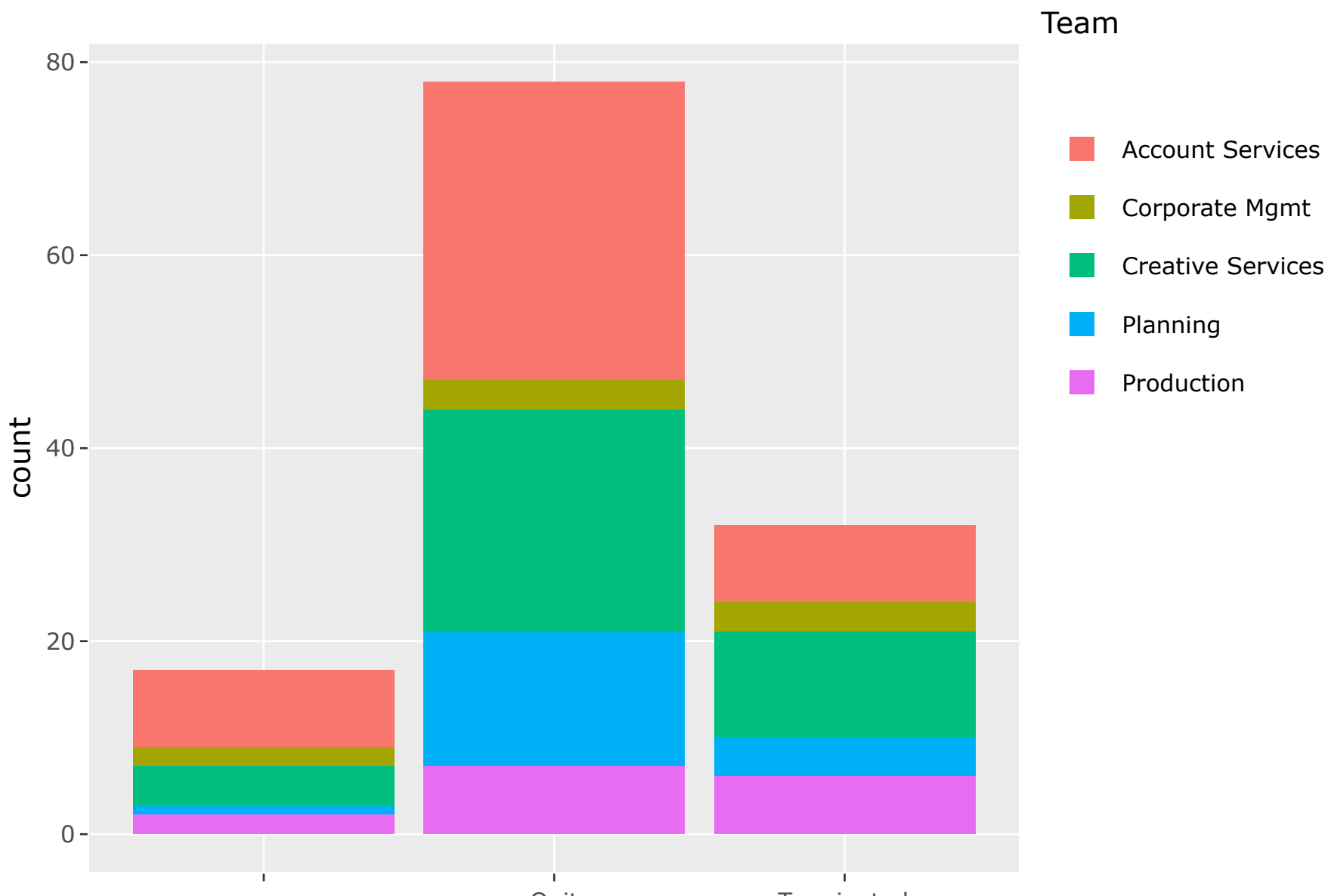





```
#C5 level has avg. 158K salary which is a significant investment from Employer's perspective. The
n the fact they are one of the biggest groups who get "terminated" (not quit) by the company may
suggests that there is an issue with hiring process.
aggregate(hr$BEGIN.SALARY, by=list(hr$C.LEVEL), FUN=mean)
```

```
##   Group.1         x
## 1      C1  36275.00
## 2      C2  52430.95
## 3      C3  72075.00
## 4     C4   91666.67
## 5      C5 158571.43
## 6      C6 167500.00
```

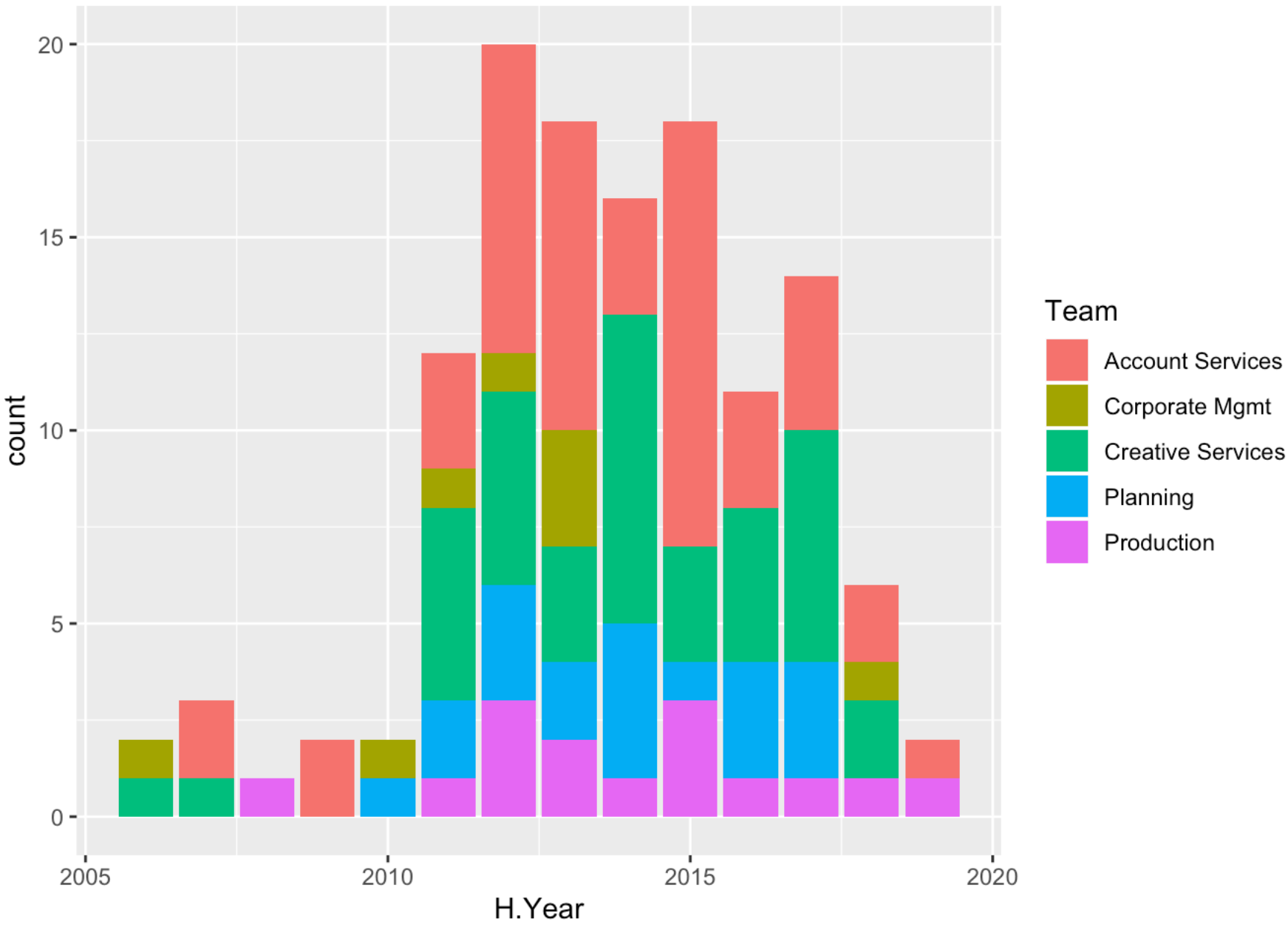
```
#Checking out the same employee status by department.
#Account and Creative teams are advertising's bread and butter type of roles.
#We seem to have more difficulty retaining Account team. Quit > Terminated
#While retention is still problematic with Creative (Quit 23), it is the department where we are
having most problem hiring the right resource (Terminated 11)
b <-ggplot(hr, aes(x=TermType, fill=Team))+
  geom_bar()
ggplotly(b)
```



Quit Terminated

TermType

```
#Checking # of hires by year and department. There must be a reason why # of hires jumped so grea
tly after 2011. This also shows that we are in constant need for Account and Creative Services th
en Planning and Production needs have grown after 2010.
ggplot(hr, aes(x=H.Year, fill=Team))+
  geom_bar()
```



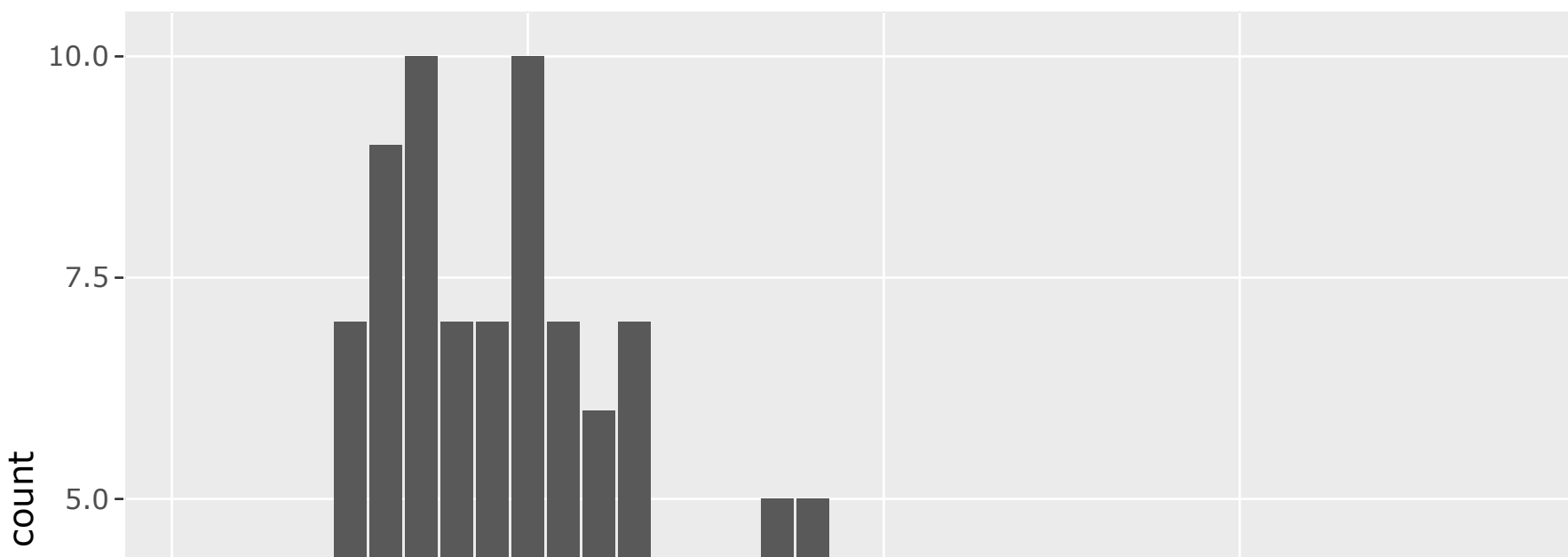

```
#Changing HireDate, TerminationDate and BirthDate to Date format for further data manipulation.
hdate <- as.character(hr$Hire)
tdate <- as.character(hr$Termination)
bdate <- as.character(hr$DOB)
hr$hdate = as.Date(hdate, format="%Y%m%d")
hr$tdate = as.Date(tdate, format="%Y%m%d")
hr$bdate = as.Date(bdate, format="%Y%m%d")
head(hr)
```

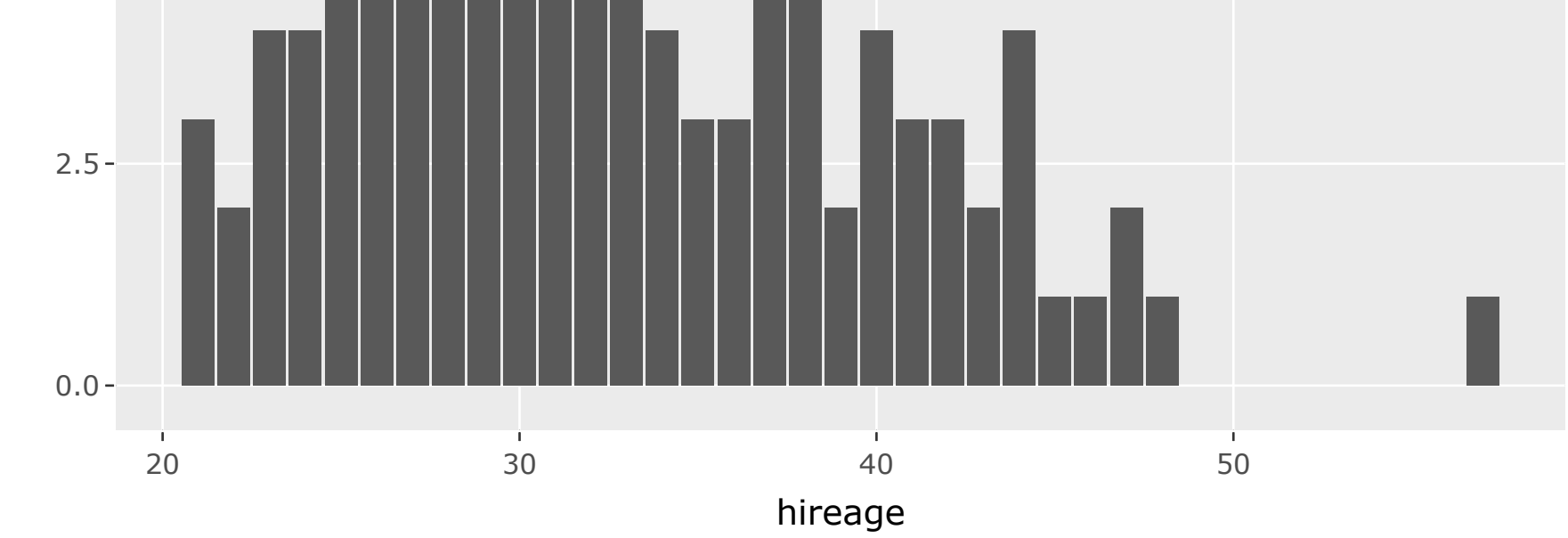
```
##     S.EMP.                    Title C.LEVEL              Team  Job.Level
## 1 60060523     Senior Art Director      C3 Creative Services Supervisor
## 2 60071662 Asso. Creative Director      C3 Creative Services Supervisor
## 3 60072318   Sr. Account Executive      C2  Account Services Supervisor
## 4 60072838   Engagement Supervisor      C3  Account Services Supervisor
## 5 60081603     Junior Art Director      C2        Production  Associate
## 6 60114364     Account Coordinator      C1  Account Services  Associate
##   BEGIN.SALARY Raise     Hire H.Year Termination T.Year TermType
## 1        70000     1 20060821   2006    20110831   2011     Quit
## 2        80000     2 20070522   2007    20110520   2011     Quit
## 3        70000     1 20070816   2007    20120323   2012     Quit
## 4        75000  -    20070910   2007    20120120   2010     Quit
## 5        35000     2 20080520   2008    20140919   2014     Quit
## 6        35000     2 20110501   2011    20140314   2014     Quit
##       Status      DOB SEX                Education
## 1 Terminated 19740928   M          Bachelors Degree
## 2 Terminated 19660123   F          Bachelors Degree
## 3 Terminated 19800925   M          Bachelors Degree
## 4 Terminated 19731212   F          Bachelors Degree
## 5 Terminated 19811224   M          Bachelors Degree
## 6 Terminated 19830126   F College/Diploma/Associate
##                          Major DistToWork      hdate      tdate      bdate
## 1 Advertising & Graphic Design       15.0 2006-08-21 2011-08-31 1974-09-28
## 2           English & Religion       40.0 2007-05-22 2011-05-20 1966-01-23
## 3                                    25.0 2007-08-16 2012-03-23 1980-09-25
## 4                    Sociology       20.0 2007-09-10 2012-01-20 1973-12-12
## 5               Graphic Design       13.3 2008-05-20 2014-09-19 1981-12-24
## 6                                    24.5 2011-05-01 2014-03-14 1983-01-26
```

```
#Figuring out each employee's Age at hire and Age at termination.
hr$hireage <- as.integer(round((hr$hdate-hr$bdate)/365, digit=0))
hr$termage <- as.integer(round((hr$tdate-hr$bdate)/365, digit=0))
head(hr)
```

```
##     S.EMP.                    Title C.LEVEL              Team  Job.Level
## 1 60060523     Senior Art Director      C3 Creative Services Supervisor
## 2 60071662 Asso. Creative Director      C3 Creative Services Supervisor
## 3 60072318   Sr. Account Executive      C2  Account Services Supervisor
## 4 60072838   Engagement Supervisor      C3  Account Services Supervisor
## 5 60081603     Junior Art Director      C2        Production  Associate
## 6 60114364     Account Coordinator      C1  Account Services  Associate
##   BEGIN.SALARY Raise     Hire H.Year Termination T.Year TermType
## 1        70000     1 20060821   2006    20110831   2011     Quit
## 2        80000     2 20070522   2007    20110520   2011     Quit
## 3        70000     1 20070816   2007    20120323   2012     Quit
## 4        75000  -    20070910   2007    20120120   2010     Quit
## 5        35000     2 20080520   2008    20140919   2014     Quit
## 6        35000     2 20110501   2011    20140314   2014     Quit
##       Status      DOB SEX                Education
## 1 Terminated 19740928   M          Bachelors Degree
## 2 Terminated 19660123   F          Bachelors Degree
## 3 Terminated 19800925   M          Bachelors Degree
## 4 Terminated 19731212   F          Bachelors Degree
## 5 Terminated 19811224   M          Bachelors Degree
## 6 Terminated 19830126   F College/Diploma/Associate
##                          Major DistToWork      hdate      tdate      bdate
## 1 Advertising & Graphic Design       15.0 2006-08-21 2011-08-31 1974-09-28
## 2           English & Religion       40.0 2007-05-22 2011-05-20 1966-01-23
## 3                                    25.0 2007-08-16 2012-03-23 1980-09-25
## 4                    Sociology       20.0 2007-09-10 2012-01-20 1973-12-12
## 5               Graphic Design       13.3 2008-05-20 2014-09-19 1981-12-24
## 6                                    24.5 2011-05-01 2014-03-14 1983-01-26
##   hireage termage
## 1      32       37
## 2      41       45
## 3      27       32
## 4      34       38
## 5      26       33
## 6      28       31
```

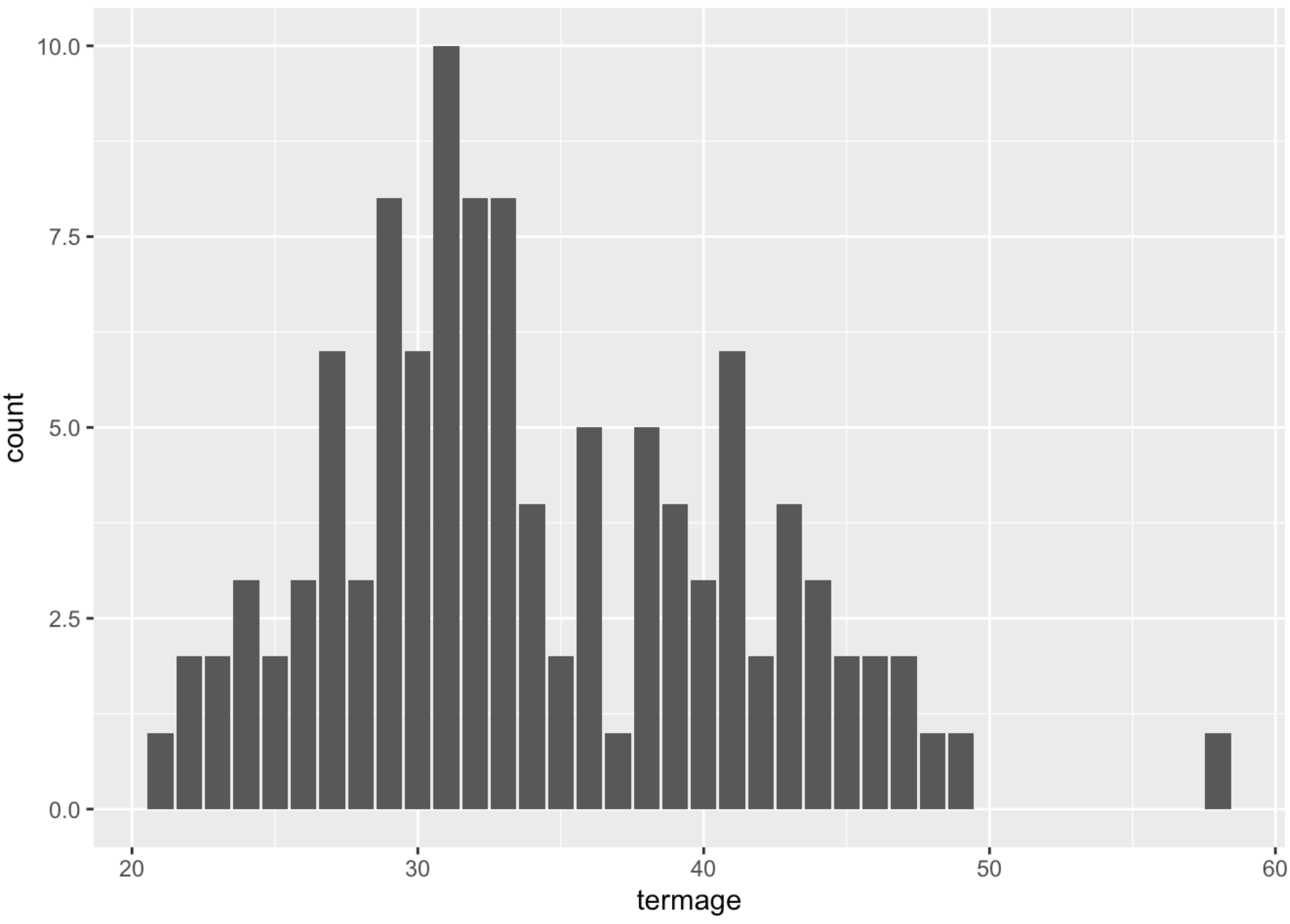
```
#Plotting a graph to see if this company is more appealing to certain age group as a career oppor
tunity. We seem to have more new hires in younger or equal to 30 group.
d<-ggplot(hr, aes(x=hireage))+
  geom_bar()
ggplotly(d)
```



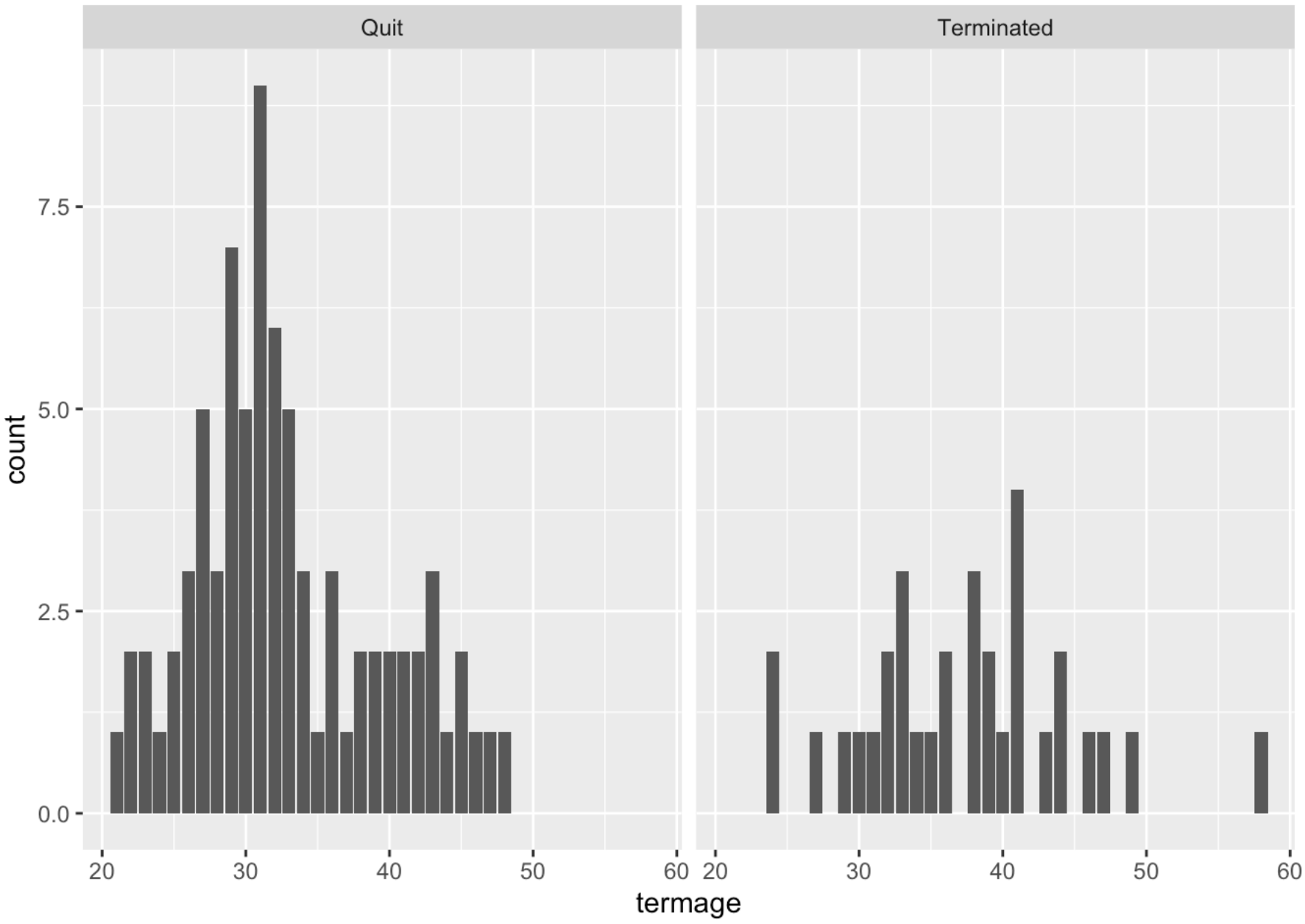



```
#Checking in what age people leave the company the most. Please note that active 17 employees are not included in this plot. It is most motable around 30. Please note that this does not different iate "Quit" and "Terminated" status. Further analysis is required.
ggplot(hr, aes(x=termage))+
  geom_bar()
```

```
## Warning: Removed 17 rows containing non-finite values (stat_count).
```

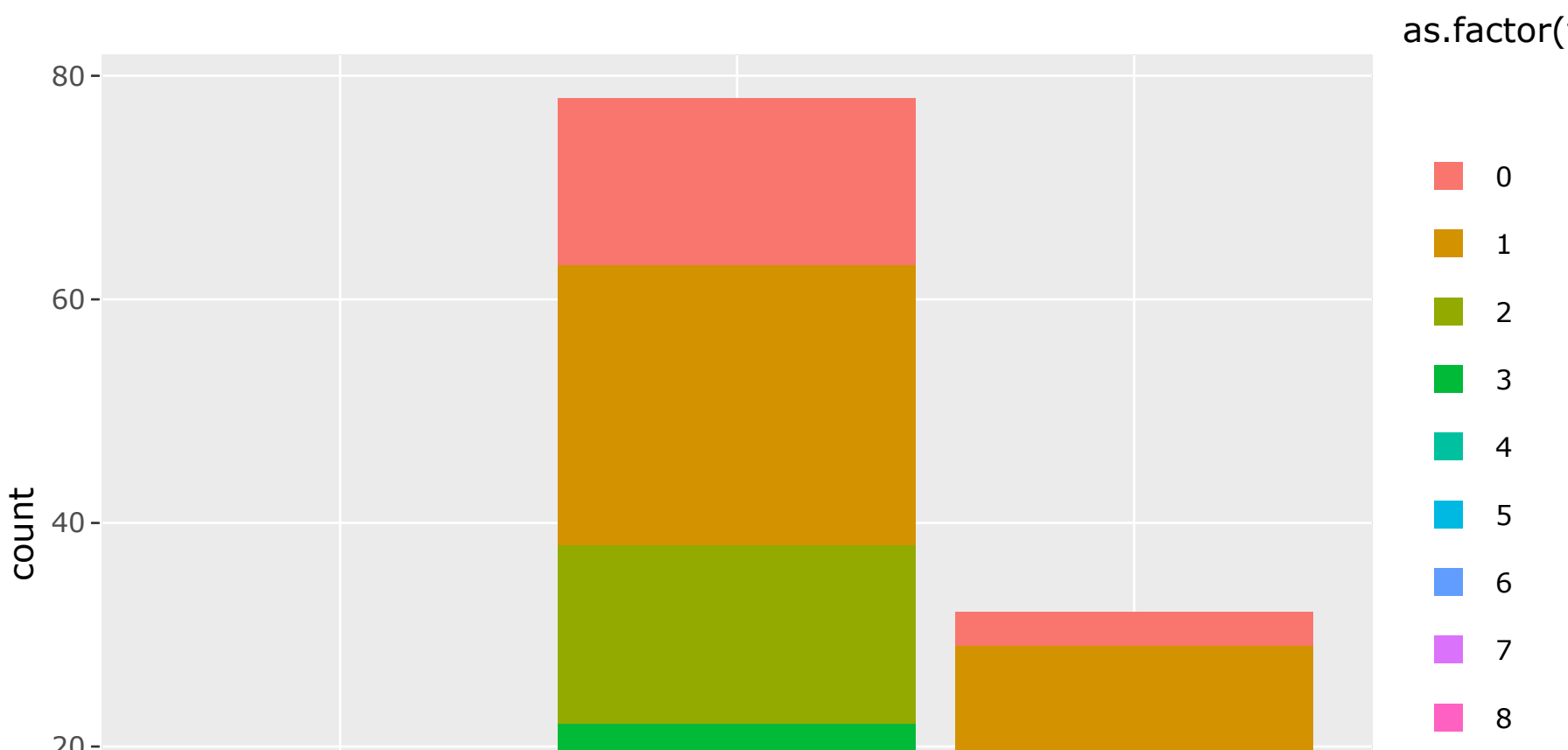


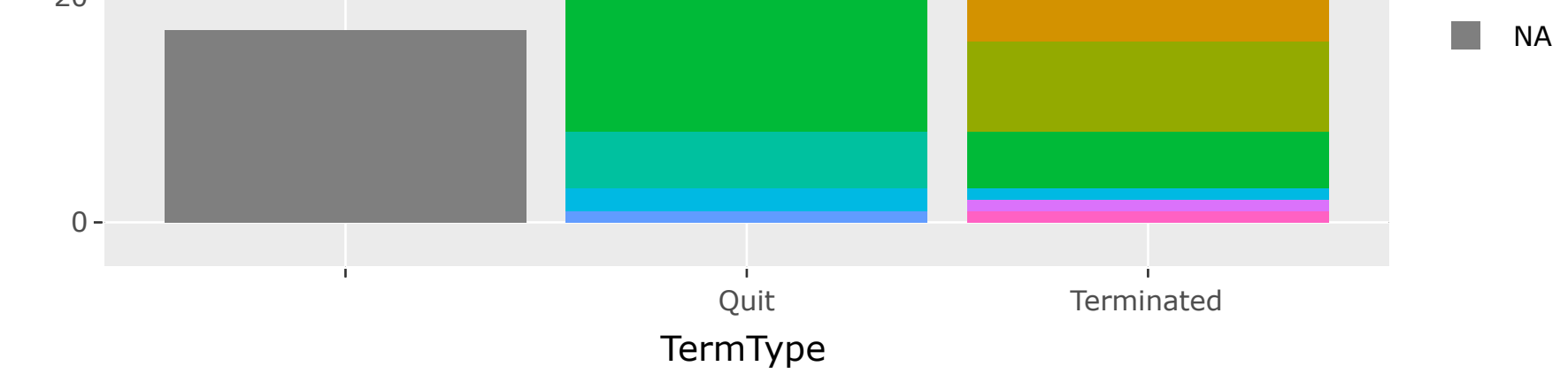

```
# Two charts that differentiate Quit and Terminiated.
hr %>% filter(TermType=="Quit" | TermType=="Terminated")%>%
ggplot(., aes(x=termage))+
  geom_bar()+
  facet_wrap(~TermType)
```



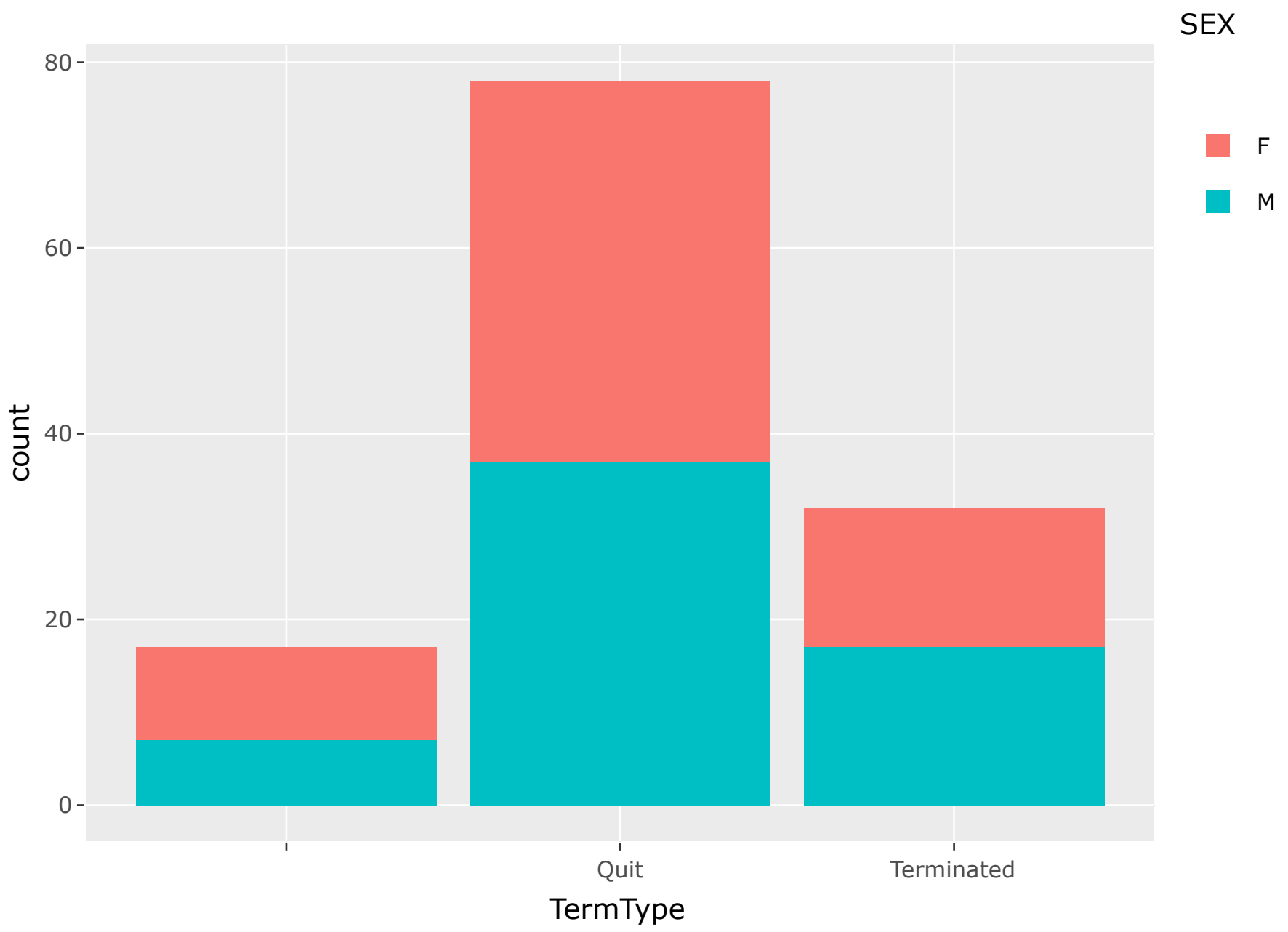

```
#Figuring out number of years employees stayed in the company.
hr$tenure <- as.integer(round((hr$tdate-hr$hdate)/365, digit=0))
```

```
#Maybe employees would feel that they have had enough and start looking after certain number of y
ears. Also employer could scrutinize employee's performance for a certain period, then maybe the
level of scrutiny gets reduced. Below plot suggests that change in staff mostly happens within 1s
t 3 years.
e<-ggplot(hr, aes(x=TermType, fill=as.factor(tenure)))+
  geom_bar()
ggplotly(e)
```



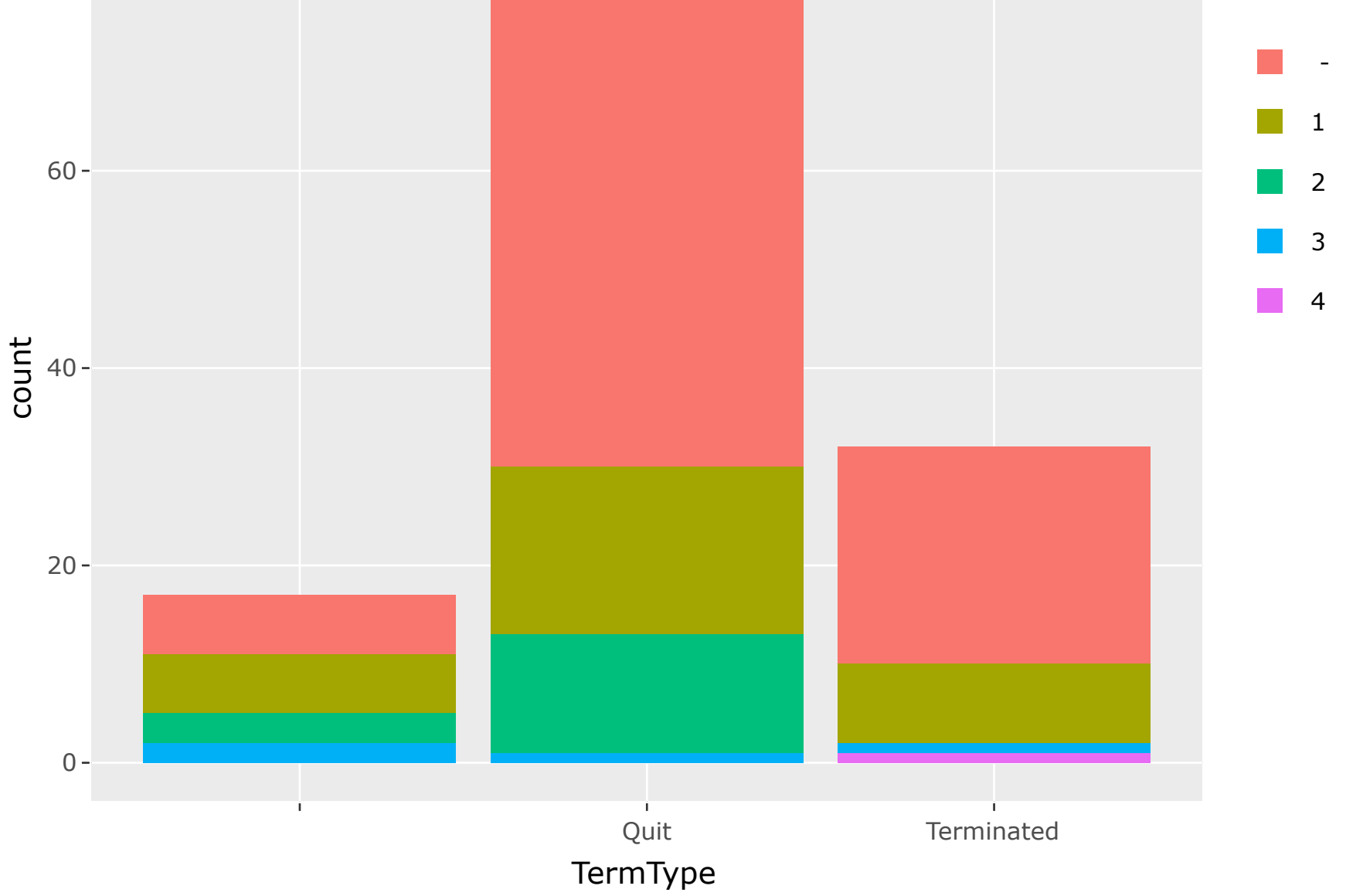



```
#Plot by gender. I don't believe this provides much insight. Might be an good illustration of gen
der equality at workplace.
g<-ggplot(hr, aes(x=TermType, fill=SEX))+
  geom_bar()
ggplotly(g)
```



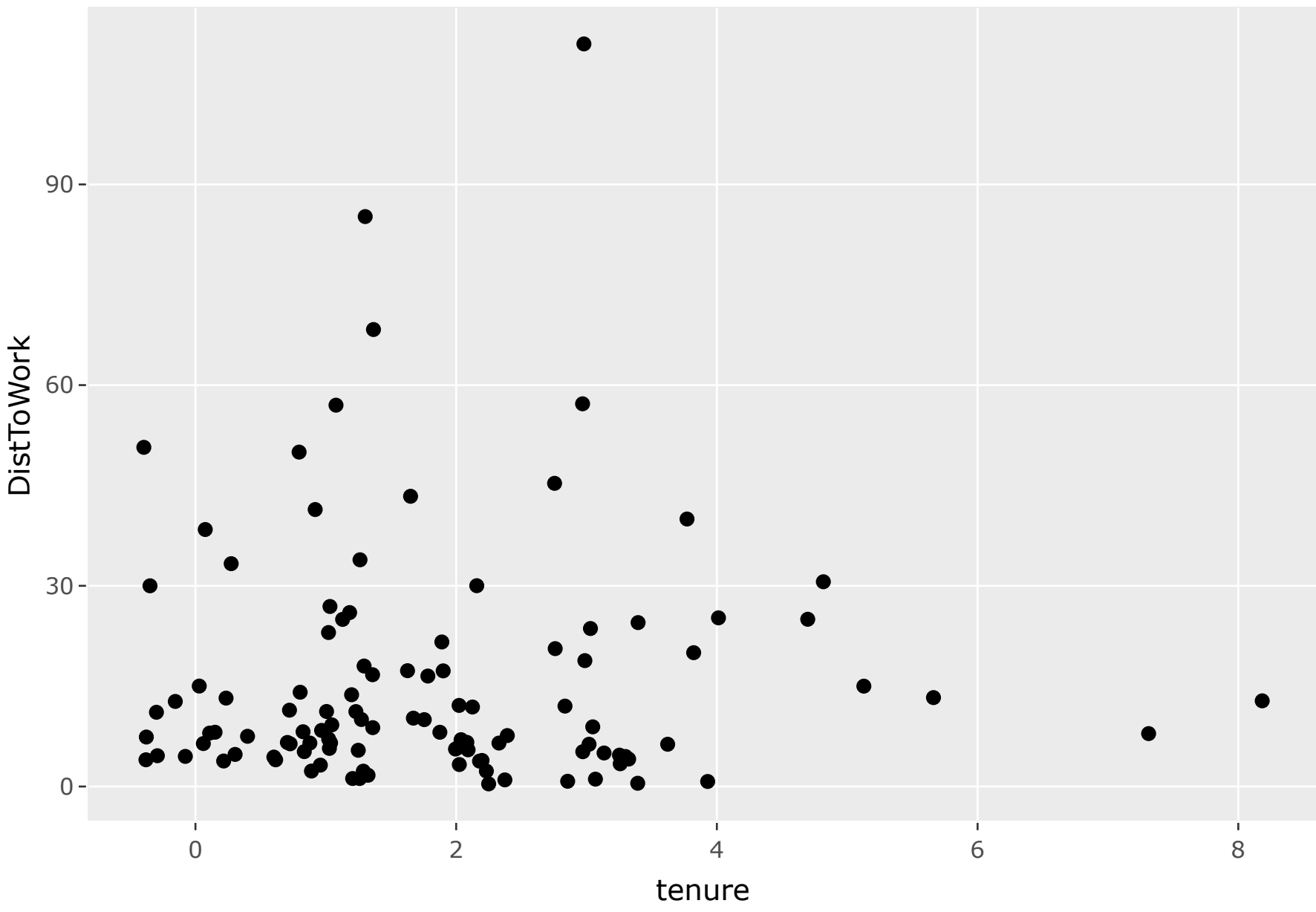

```
#Although this analysis is mostly based on employee's personal(objective), # of raises received t
hroughout the tenure is one subjective factor that can illustrate employee's performance.
#While it is obvious that employees without getting recognition (0 raise) are prone to retention
risk (both voluntarily and involuntarily), the analysis could further develop with employees with
1 or more raises.
h<-ggplot(hr, aes(x=TermType, fill=Raise))+
  geom_bar()

ggplotly(h)
```



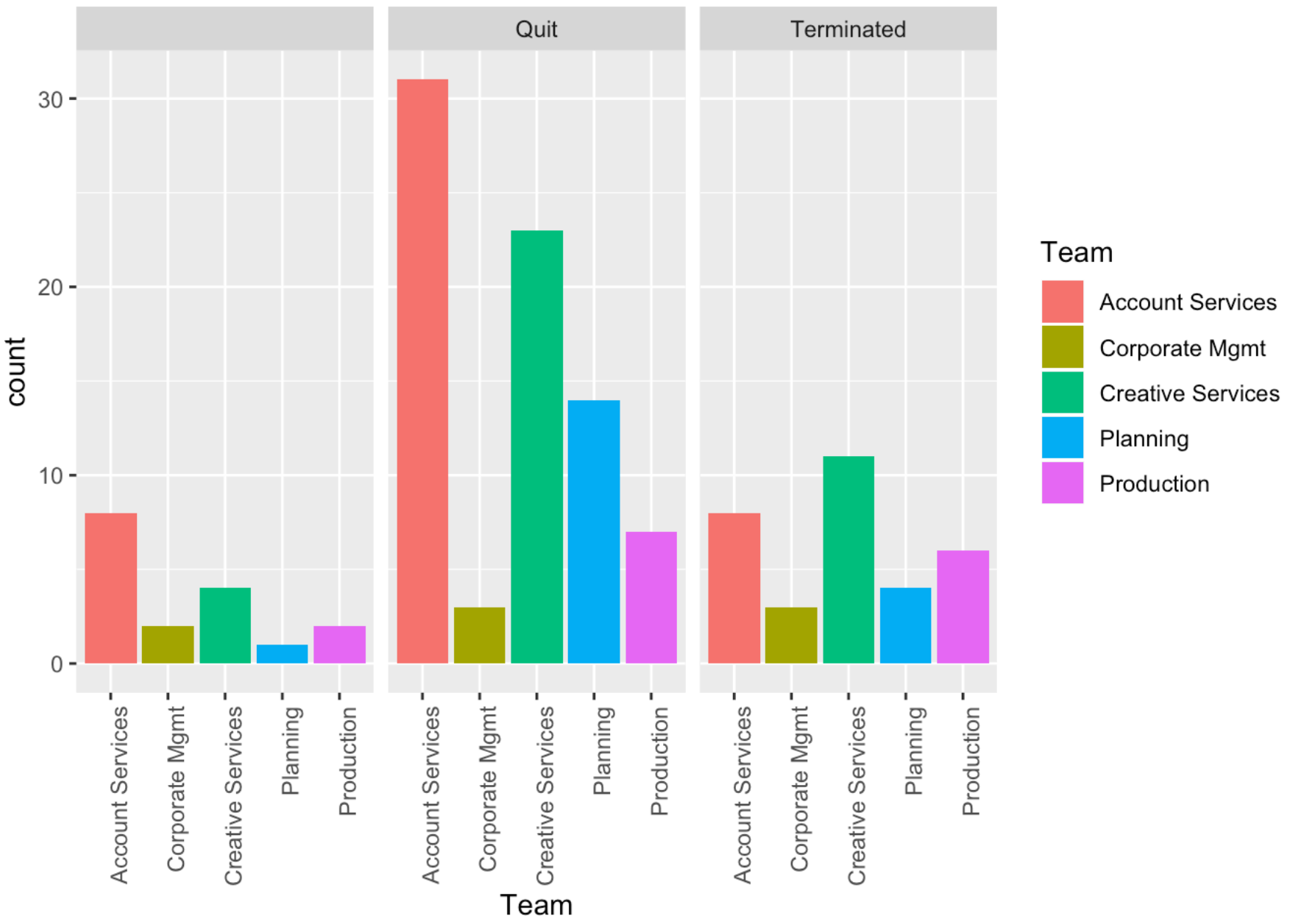

```
#Plot created to find correlation between employees' Distance to work and Tenure.

ggplotly(ggplot(hr, aes(x=tenure, y=DistToWork))+
geom_jitter())
```

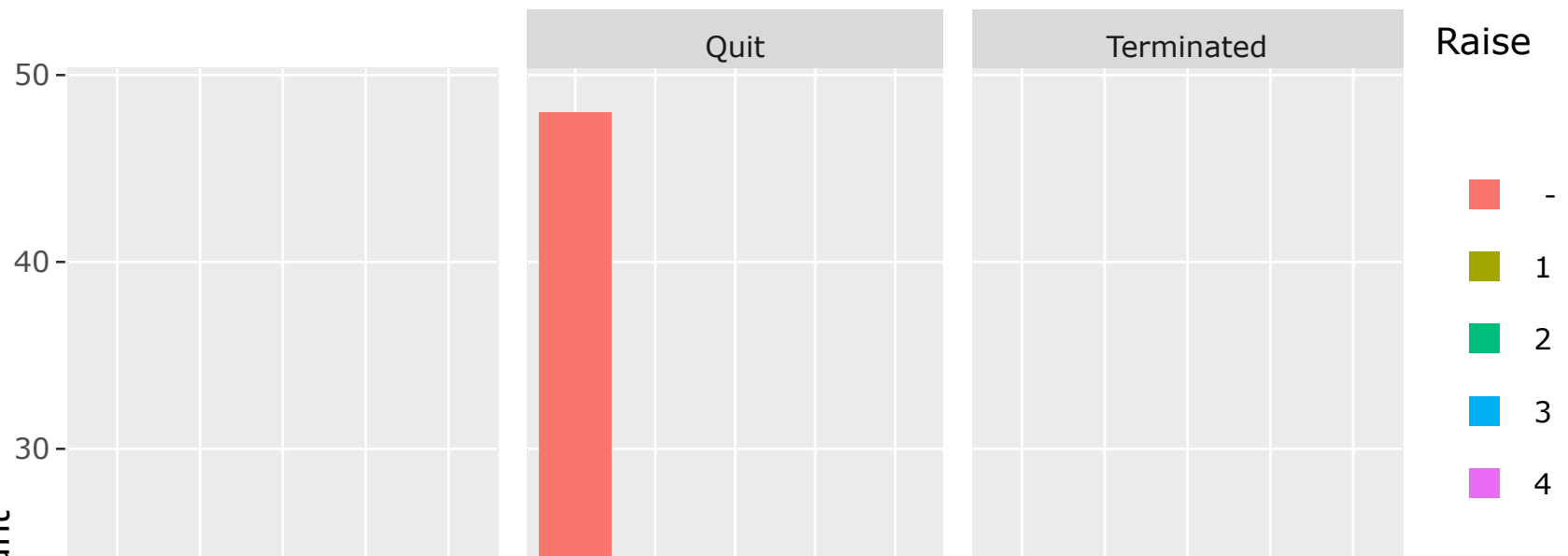


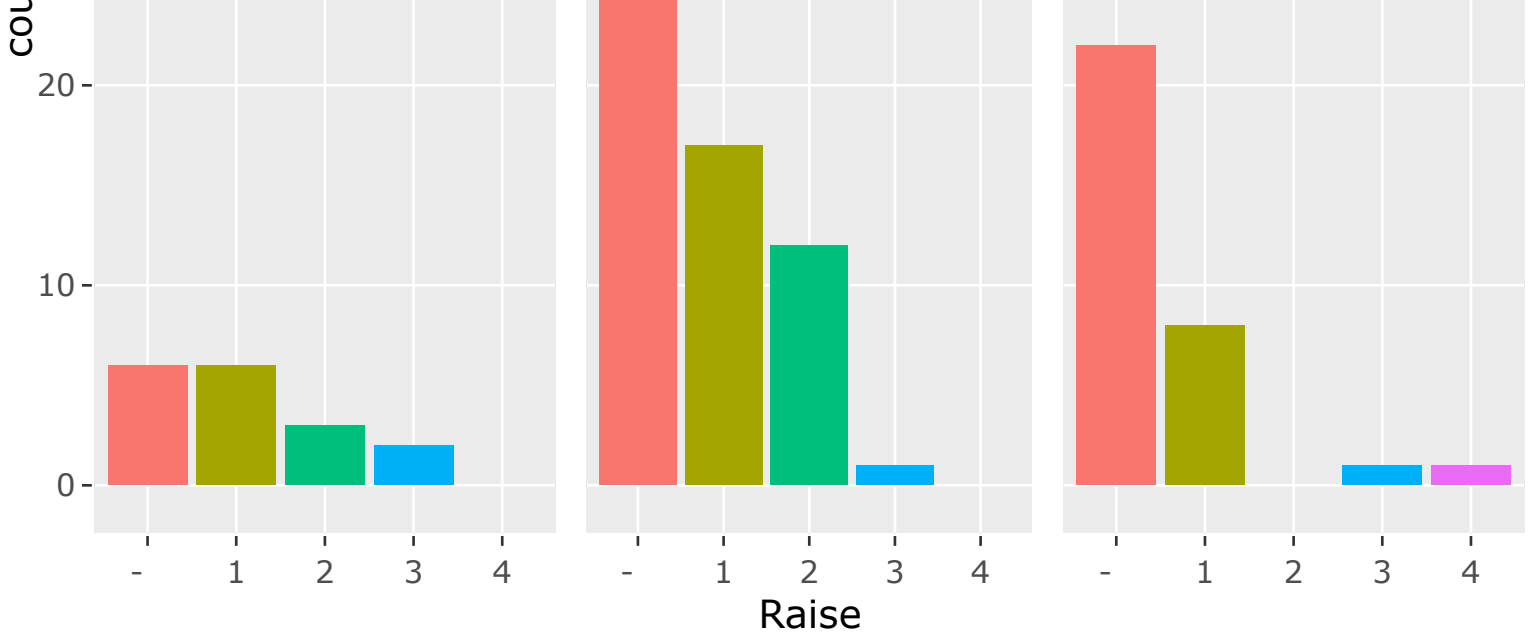
```
#Now some relationship between some variables and employee retention has been explored, I would a
lso like to explore some other visualization tools for stakeholders.

# Below is explored in the beginning but testing it in a different format to determine what is ea
sier to understand.
ggplot(hr, aes(x=Team, fill=Team))+
  geom_bar()+
  facet_wrap(~TermType)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
#Same exercise as above. Exploring different format.
p<-ggplot(hr, aes(x=Raise, fill=Raise))+
  geom_bar()+
  facet_wrap(~TermType)
ggplotly(p)
```





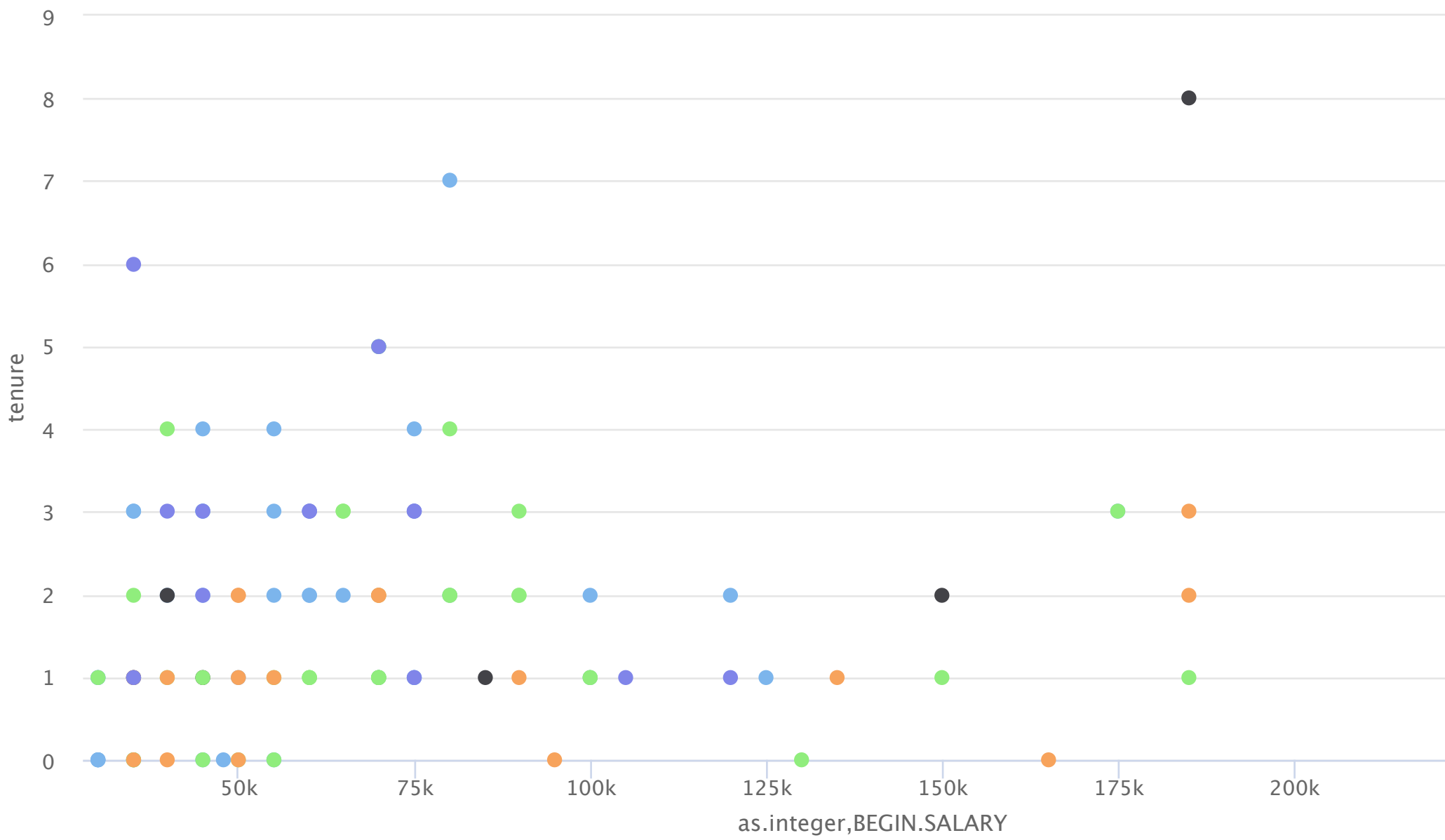

```
#Exploring HighCharter library for better visualization of analysis, in additioned to enhanced in formation delivery.
plot_ly(hr, y=hr$tenure, color=hr$Team, type="box")
```

```
## Warning: Ignoring 17 observations
```



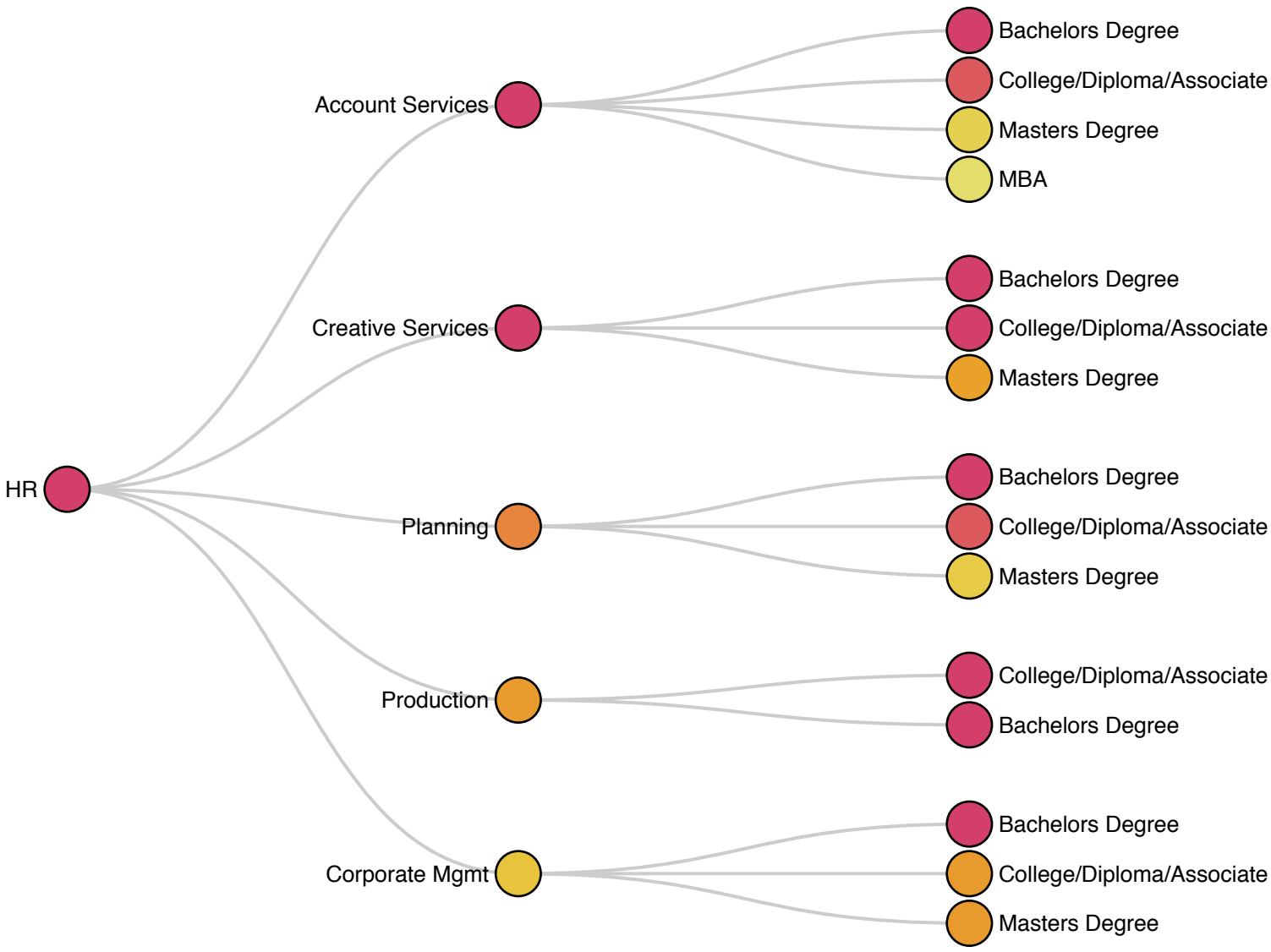
```
#Exploring HighCharter library for better visualization of analysis part 2
hchart(hr, "scatter", hcaes(x=as.integer(BEGIN.SALARY), y=tenure, group=Team))
```

```
## Warning: `parse_quosure()` is deprecated as of rlang 0.2.0.
## Please use `parse_quo()` instead.
## This warning is displayed once per session.
```




```
#A different sample utilizing collapsibleTree to add interactive function. Might be beneficial wh
en making a presentation.
library(collapsibleTree)

hr %>%
  group_by(Team, Education) %>%
  summarize('TeamNum' = n())%>%
  collapsibleTreeSummary(hr,
    hierarchy = c("Team", "Education"),
    root = "HR",
    width = 800,
    attribute = "TeamNum",
    collapsed = F,
    linkLength = 200
  )
```



Next Step

1. I believe there is some trend that can be discovered intuitively by observing data visualization. I'd like to invest some more time in developing something that can deliver information more efficiently.

2. In addition to above, I will explore a several machine learning algorithms to test if they can provide predictive power.

3. I also have company's financial data which demonstrates company's financial performance year-over-year. It will be interesting to see how its financial history reenacts with HR history.