Employee Turnover Rates

Jacob Miller, Sarah Tappin, Jacqueline Zhang

11/30/2020

```
library(readr)
library(survival)
library(survminer)
library(ggplot2)
library(mclust)
library(factoextra)
library(gridExtra)
library(purrr)
library(cluster)
library(GGally)
library(plotly)
library(ggbiplot)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   stag = col_double(),
##   event = col_double(),
##   gender = col_character(),
##   age = col_double(),
##   industry = col_character(),
##   profession = col_character(),
##   traffic = col_character(),
##   coach = col_character(),
##   head_gender = col_character(),
##   greywage = col_character(),
##   way = col_character(),
##   extraversion = col_double(),
##   independ = col_double(),
##   selfcontrol = col_double(),
##   anxiety = col_double(),
##   novator = col_double()
## )
```

Overview

Data Description

We sourced our Employee Turnover data set from Kaggle. It was provided by Edward Babushkin, a Russian blogger.

The Employee Turnover data set, aims to predict an employee's risk of quitting. Some of the attributes included are stag (experience time in months), event (employee turnover), gender, age, industry, profession, traffic (how the employee came to the company), coach (whether or not there's a supervisor/mentor), head gender (gender of manager/supervisor), greywage (a mix of taxed and untaxed wages). The final attributes are personality based, including scores on extraversion, independence, self control, anxiety, and novator.

Variables:

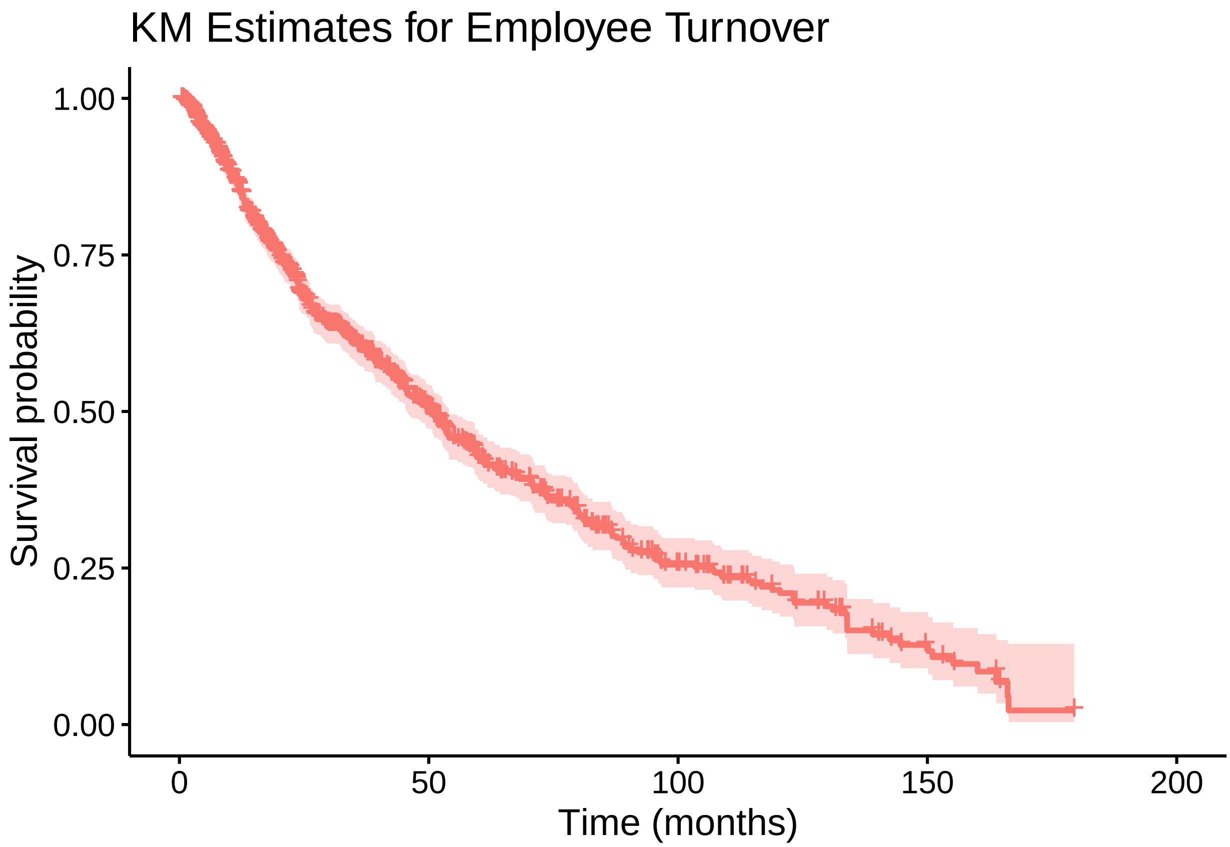
- stag - experience
- event - staying or quitting
- gender
- age
- industry
- profession
- traffic - From what pipeline candidate came to the company
- coach - presence of a mentor during the probation period
- head_gender - interpreted to mean the gender of the supervisor
- greywage - Portion of the salary that is paid in cash
- way - how an employee gets to work (by foot, by bus, etc.)
- Personality Traits:
  - extraversion
  - independ
  - selfcontrol
  - anxiety
  - novator

Objective

We are interested in determining which attributes strongly influence employee turnover. In order to do this we are utilizing survival analysis modeling and BIC as a criterion for model selection. In order to further explore the personlaity attributes which are unique to this data set, we used machine learning techniques such as Principal Componenet Analysis (PCA) and Clustering to determine if certain personality traits had an effect on the individual choosing to quit their job.

Survival Analysis Model

```r
turn.surv<-Surv(turnover$stag,turnover$event)
turn.fit<-surv_fit(turn.surv~1, data = turnover)
ggsurvplot(turn.fit, legend = 'none') +
  ggtitle("KM Estimates for Employee Turnover") + xlab('Time (months)')
```



Here we will start forward selection to find a model

```
full.model<-coxph(turn.surv~gender + age + industry + profession + traffic + coach +
                    head_gender + greywage + way + extraversion + independ +
                    selfcontrol + anxiety + novator, data = turnover)
red.model<-coxph(turn.surv~1, data = turnover)
n<-length(turn.surv)
(best_model<- step(red.model, scope = list(lower = red.model, upper = full.model),
                  direction = 'forward', trace = 0, k =log(n)))
```

```
## Call:
## coxph(formula = turn.surv ~ greywage + age + extraversion, data = turnover)
##
##                    coef exp(coef)  se(coef)      z        p
## greywagewhite -0.580522  0.559606  0.126429 -4.592  4.4e-06
## age            0.023327  1.023602  0.006176  3.777 0.000159
## extraversion   0.072045  1.074703  0.023326  3.089 0.002011
##
## Likelihood ratio test=39.59  on 3 df, p=1.299e-08
## n= 1129, number of events= 571
```

This gives us a good look at what our model using BIC will look like, but in order to take a closer look at why it chose those covariates, we will go through the process step by step. For the sake of not having compious amounts of code, we only included the code for the last step in the selection processs.

```r
step_4_basis<-BIC(
model8.2.10.1<-coxph(turn.surv~greywage + age + extraversion + gender, data = turnover),
model8.2.10.3<-coxph(turn.surv~greywage + age + extraversion + industry, data = turnover),
model8.2.10.4<-coxph(turn.surv~greywage + age + extraversion + profession, data = turnover),
model8.2.10.5<-coxph(turn.surv~greywage + age + extraversion + traffic, data = turnover),
model8.2.10.6<-coxph(turn.surv~greywage + age + extraversion + coach, data = turnover),
model8.2.10.7<-coxph(turn.surv~greywage + age + extraversion + head_gender,
                     data = turnover),
model8.2.10.9<-coxph(turn.surv~greywage + age + extraversion + way, data = turnover),
model8.2.10.11<-coxph(turn.surv~greywage + age + extraversion + independ, data = turnover),
model8.2.10.12<-coxph(turn.surv~greywage + age + extraversion + selfcontrol,
                      data = turnover),
model8.2.10.13<-coxph(turn.surv~greywage + age + extraversion + anxiety, data = turnover),
model8.2.10.14<-coxph(turn.surv~greywage + age + extraversion + novator, data = turnover)
)
step_4_basis$BIC
```

```
##  [1] 6925.220 6957.388 6984.436 6943.002 6929.736 6926.360 6922.034 6923.498
##  [9] 6924.772 6923.078 6926.373
```

```r
which(step_4_basis$BIC==min(step_4_basis$BIC))
```

```
## [1] 7
```

```r
#model8.2.10.9 with greywage + age + extraversion + way has highest BIC of these
```

If we continued the step process after our 3 covariate model given by the step function, model8.2.10.9 with the covariates greywage, age, extraversion, and way was the next step using BIC so we want to take a closer look at this model.

```r
test_model<-model8.2.10.9
AIC(best_model,test_model)
```

```
##            df      AIC
## best_model  3 6907.008
## test_model  5 6900.297
```

```r
anova(best_model, test_model) #test suggests we utilize the more complex model
```

```
## Analysis of Deviance Table
##  Cox model: response is  turn.surv
##  Model 1: ~ greywage + age + extraversion
##  Model 2: ~ greywage + age + extraversion + way
##    loglik Chisq Df P(>|Chi|)
## 1 -3450.5
## 2 -3445.1 10.71  2  0.004723 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model8.2.10.9 actually has a lower AIC score, so it is reasonable to compare the two models using a Likelihood Ratio Test to determine the best model. The result of the LRT tells us that the better model is in fact the one that includes the covariate ‘way’, so we will continue forward with our analysis using this larger model.Before this, however, we beg the question, if we utilized AIC instead of BIC, would our model be different? Let’s find out.

```
step(red.model, scope = list(lower = red.model, upper = full.model), direction = 'forward', trace = 0)
```

```
## Call:
## coxph(formula = turn.surv ~ industry + greywage + traffic + age +
##     selfcontrol + way + anxiety + profession, data = turnover)
##
##                                 coef exp(coef)  se(coef)      z        p
## industryBanks              -0.310101  0.733373  0.362537 -0.855 0.392350
## industryBuilding           -0.269891  0.763462  0.388213 -0.695 0.486920
## industryConsult            -0.453030  0.635699  0.368894 -1.228 0.219419
## industryetc                -0.641162  0.526680  0.368615 -1.739 0.081968
## industryHoReCa             -0.789502  0.454071  0.539769 -1.463 0.143559
## industryIT                 -1.247336  0.287269  0.385138 -3.239 0.001201
## industrymanufacture        -0.875538  0.416638  0.366679 -2.388 0.016952
## industryMining             -0.650265  0.521908  0.442790 -1.469 0.141952
## industryPharma             -0.937419  0.391637  0.461943 -2.029 0.042428
## industryPowerGeneration    -0.980404  0.375159  0.430639 -2.277 0.022809
## industryRealEstate         -1.811964  0.163333  0.580516 -3.121 0.001801
## industryRetail             -1.057573  0.347298  0.353997 -2.988 0.002813
## industryState              -0.725829  0.483923  0.401387 -1.808 0.070559
## industryTelecom            -1.250450  0.286376  0.437479 -2.858 0.004259
## industrytransport          -0.859110  0.423539  0.421849 -2.037 0.041697
## greywagewhite              -0.497038  0.608330  0.132217 -3.759 0.000170
## trafficempjs                0.837217  2.309929  0.309496  2.705 0.006828
## trafficfriends              0.055687  1.057267  0.334141  0.167 0.867639
## trafficKA                   0.124603  1.132698  0.347010  0.359 0.719539
## trafficrabrecNErab          0.466196  1.593919  0.305273  1.527 0.126725
## trafficrecNErab            -0.104370  0.900892  0.373424 -0.279 0.779866
## trafficreferal              0.296130  1.344646  0.320974  0.923 0.356216
## trafficyoujs                0.599363  1.820959  0.306205  1.957 0.050301
## age                         0.022213  1.022461  0.006357  3.494 0.000476
## selfcontrol                -0.064847  0.937211  0.022513 -2.880 0.003971
## waycar                     -0.183878  0.832037  0.101025 -1.820 0.068738
## wayfoot                    -0.358152  0.698967  0.171172 -2.092 0.036407
## anxiety                    -0.057459  0.944160  0.025891 -2.219 0.026468
## professionBusinessDevelopment  0.589555  1.803187  0.500417  1.178 0.238746
## professionCommercial        0.981333  2.668010  0.497617  1.972 0.048602
## professionConsult           0.570490  1.769134  0.508310  1.122 0.261723
## professionEngineer          0.971260  2.641270  0.525810  1.847 0.064723
## professionetc               0.462411  1.587898  0.482048  0.959 0.337426
## professionFinan\xf1e        0.086383  1.090223  0.516640  0.167 0.867212
## professionHR                0.246210  1.279168  0.423664  0.581 0.561143
## professionIT                0.060864  1.062755  0.473306  0.129 0.897679
## professionLaw               0.333430  1.395747  0.640200  0.521 0.602491
## professionmanage            1.281883  3.603418  0.497135  2.579 0.009922
## professionMarketing         0.720965  2.056417  0.477451  1.510 0.131036
## professionPR                0.845109  2.328232  0.637613  1.325 0.185030
## professionSales             0.532110  1.702521  0.454873  1.170 0.242082
## professionTeaching             0.637788  1.892291  0.567332  1.124 0.260933
##
## Likelihood ratio test=165.5  on 42 df, p=< 2.2e-16
## n= 1129, number of events= 571
```

Safe to say that if we utilized AIC, we would end up with a very different model. This larger model has a much higher likelihood, but it has significantly more covariates than our other two models. This would make interpretation and utilization of the model very difficult and confusing, so we will stick to utilizing our test_model from above. To start, we will check to see if cox proportional hazards is a reasonable assumption for the model. The first step will be a visualization of the log-log plots of our covariates. In order to get a comprehensible plot of our extraversion and age covariates, we need to first group the data.

```
turnover$extra.cat<-factor(ceiling((turnover$extraversion)/5))
levels(turnover$extra.cat) <- c('1-5', '5+')
table(turnover$extra.cat)
```

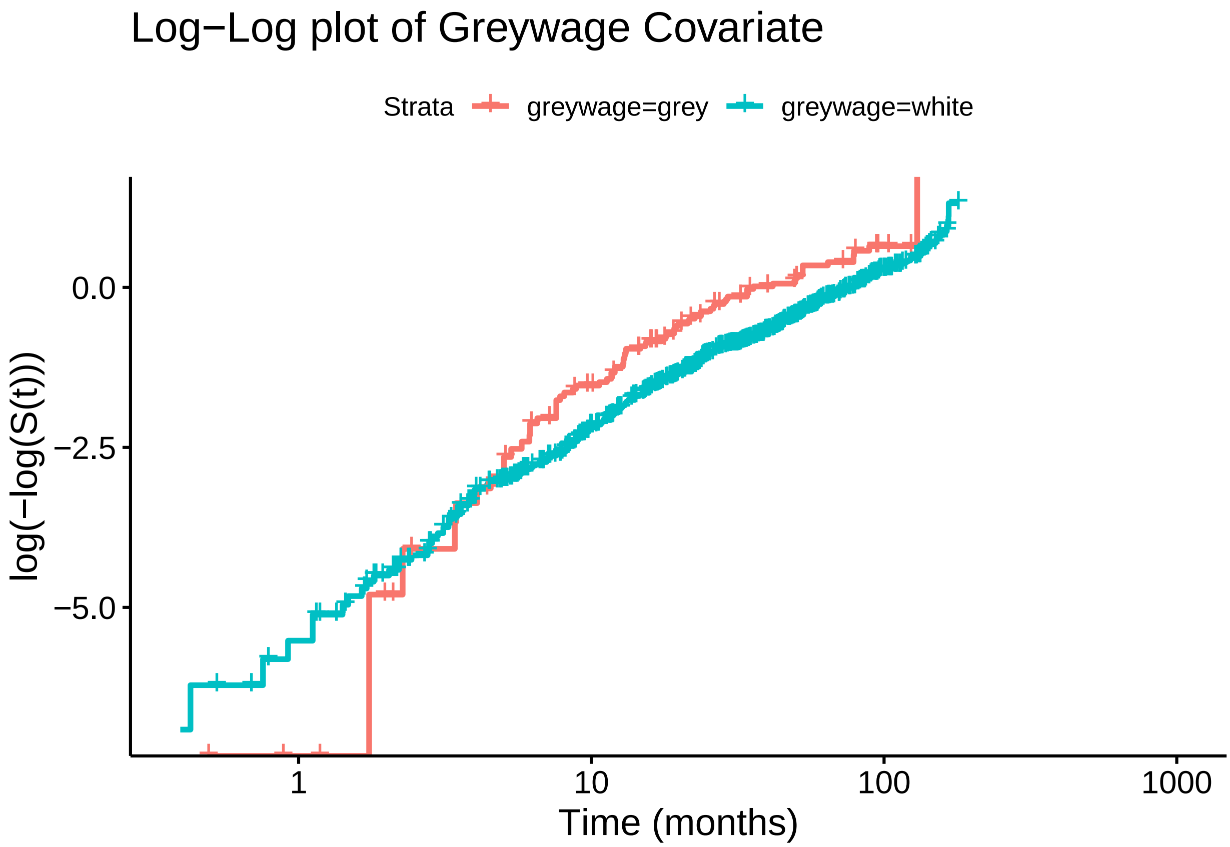
```
##
## 1-5  5+
## 431 698
```

```
turnover$age.cat<-factor(ceiling((turnover$age)/30))
levels(turnover$age.cat)<-c('18-30','30-58')
table(turnover$age.cat)
```

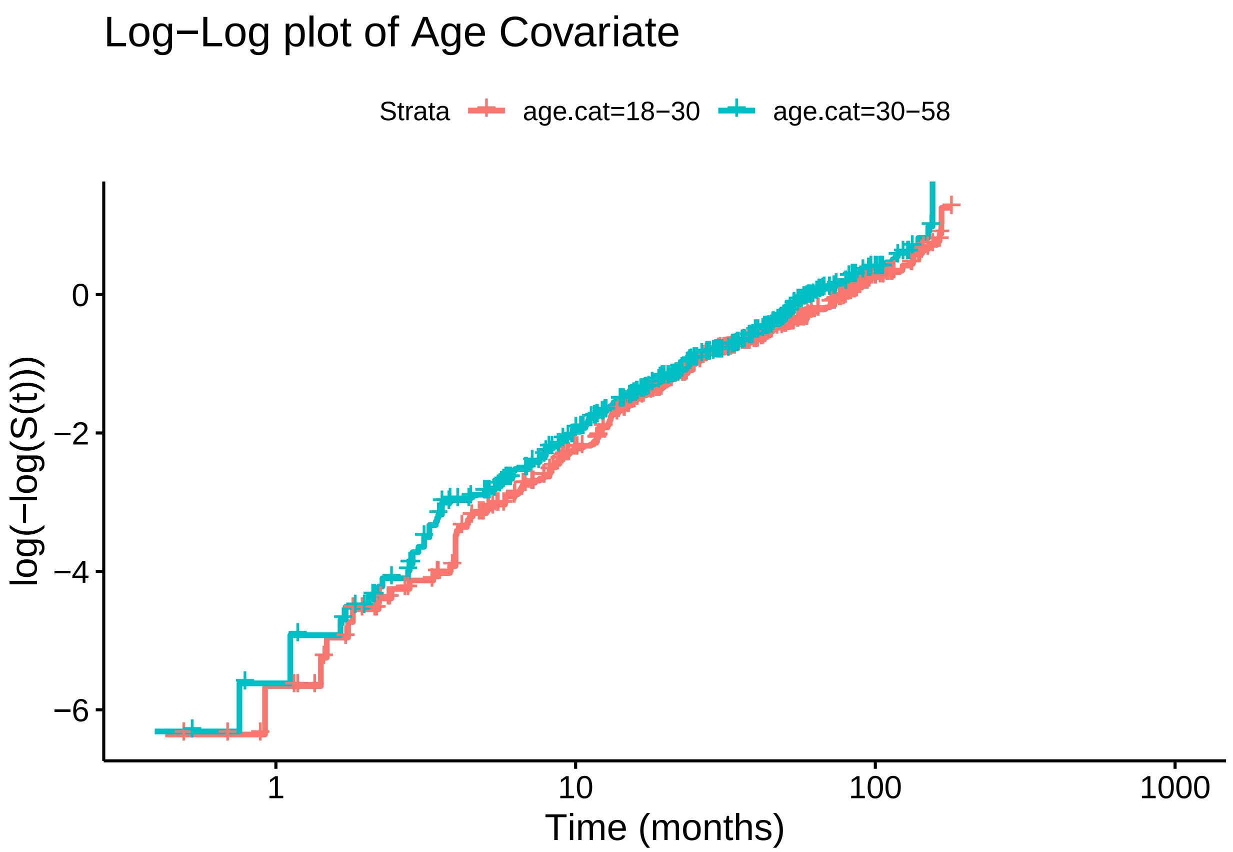
```
##
## 18-30 30-58
##   577   552
```

With our covariates adjusted into reasonable groups we can more clearly see the log-log plots.

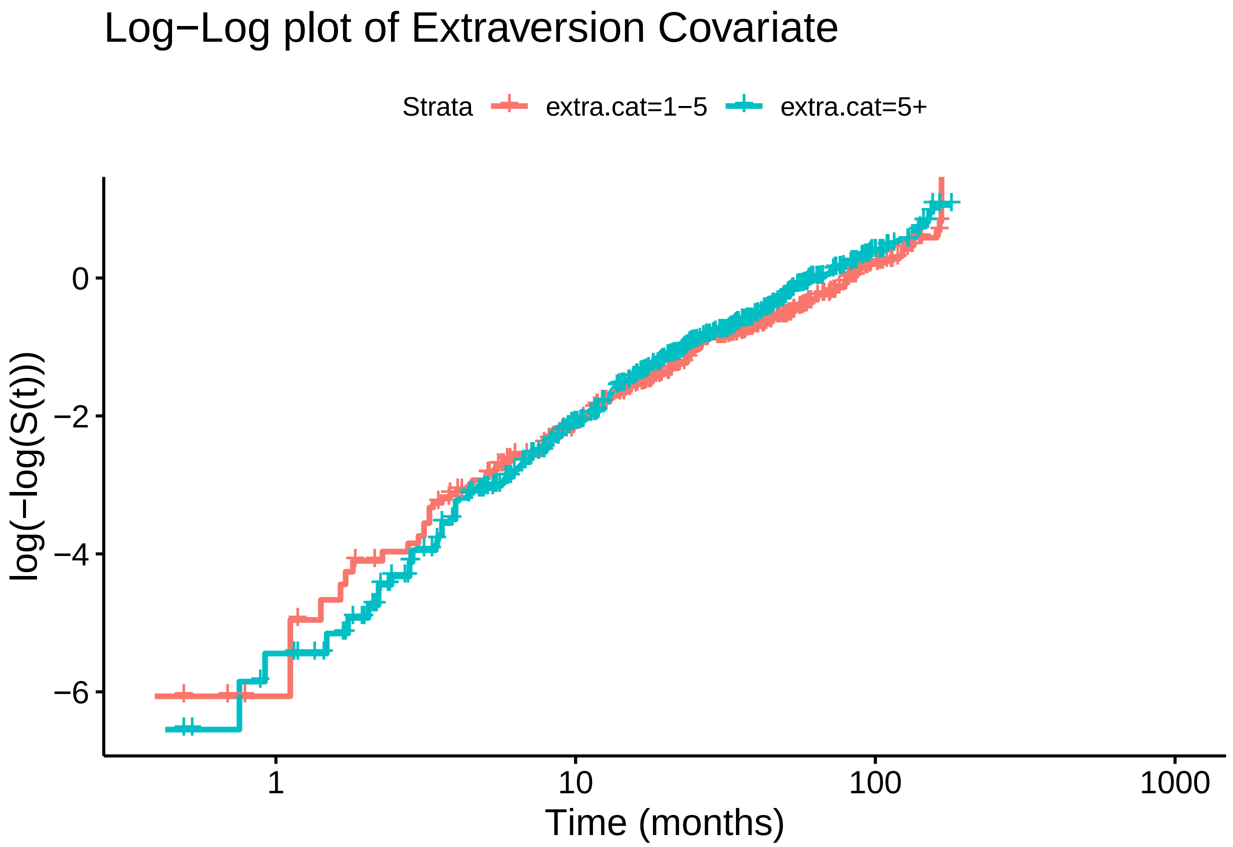
```
test_fit1<-surv_fit(turn.surv~ greywage, data = turnover)
test_fit2<-surv_fit(turn.surv~ age.cat, data = turnover)
test_fit3<-surv_fit(turn.surv~ extra.cat, data = turnover)
test_fit4<-surv_fit(turn.surv~ way, data = turnover)

ggsurvplot(test_fit1, fun = 'cloglog') + ggtitle('Log-Log plot of Greywage Covariate') +
  xlab('Time (months)')
```



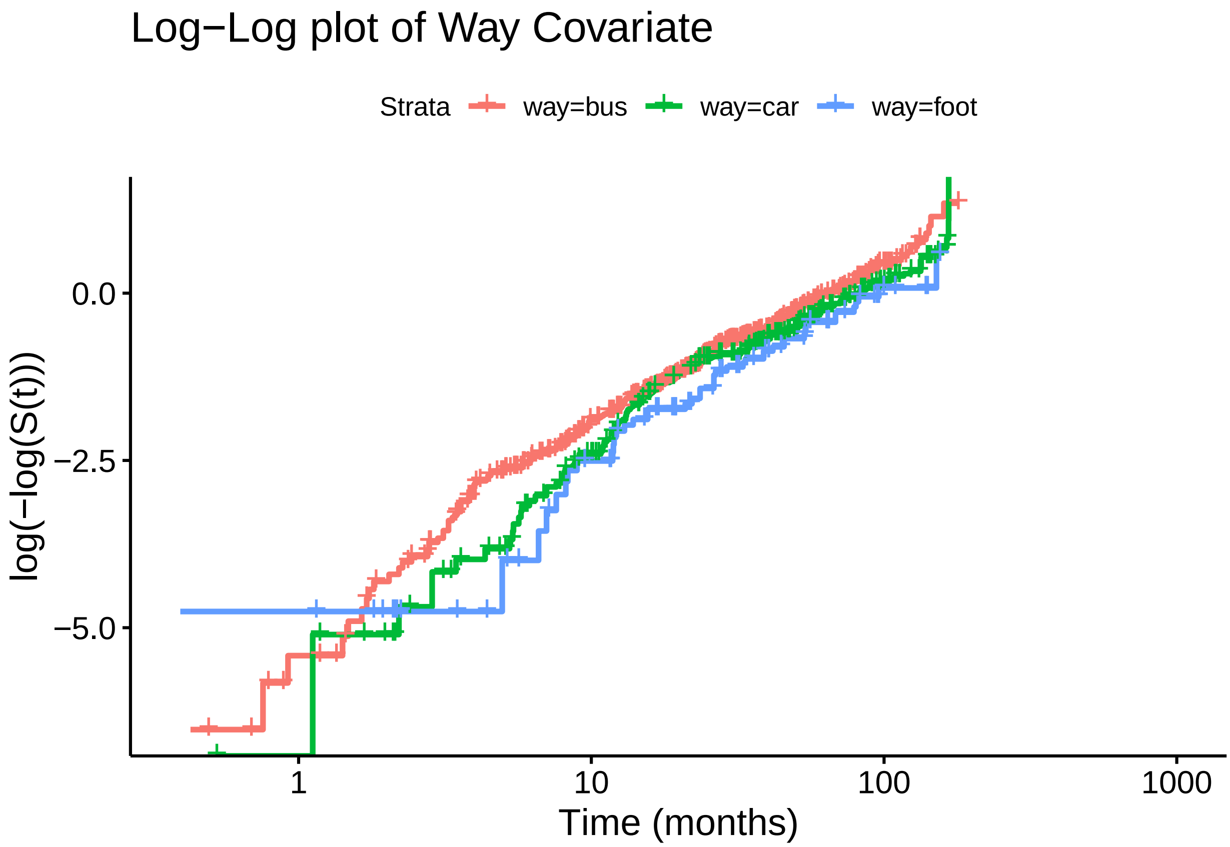
```r
ggsurvplot(test_fit2, fun = 'cloglog') + ggtitle('Log-Log plot of Age Covariate') +
  xlab('Time (months)')
```



```r
ggsurvplot(test_fit3, fun = 'cloglog') + ggtitle('Log-Log plot of Extraversion Covariate') +
  xlab('Time (months)')
```



```
ggsurvplot(test_fit4, fun = 'cloglog') + ggtitle('Log-Log plot of Way Covariate') +
  xlab('Time (months)')
```

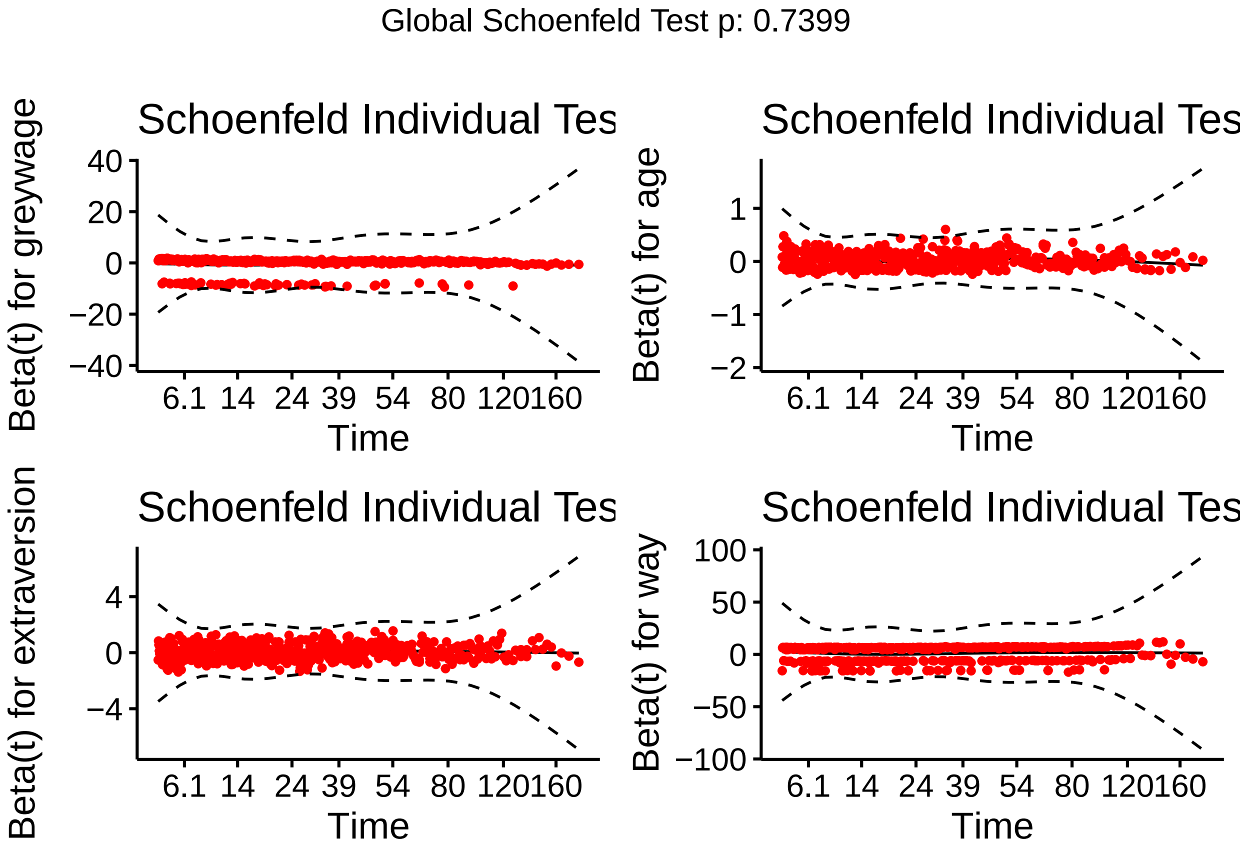


When looking at these log-log plots, there is a bit of concern in every plot from $t = 0$ to $t = 10$ and after $t = 100$ but between 10 months and 100 months, our plots, for the most part, look very good for our cox proportional hazards assumptions. That does not mean that there are not concerns in each plot, however. Our biggest concern with the greywage covariate plot is that it appears 'grey' has a curve that would eventually cross 'white' if the data continued in its trajectory. For the age covariate, there are no real concerns that stand out besides how close the two lines appear to be to each other, but this is not indicative of proportional hazards being violated since the two curves appear to be mostly parallel throughout. extraversion has a similar concern to age with the curves being very close to each other, but more importantly, the curves appear to cross at $t = 1$, after $t = 10$, and again at the end of the data a little after $t = 100$. The way covariate does not appear to have an issue of intersecting curves, at least not after $t = 10$, so the only real concern would be if all the curves are truly parallel as the distance between the curves appears to be inconsistent through time. Overall, the concerns we have with the log-log plots appear mostly at the very beginning or the end of our data where there are less data points meaning that it is possible that these violations are simply due to chance. To double check our assumptions we will utilize cox.zph to determine if we should be worried about these violations.

```
(cox_test<-cox.zph(test_model))
```

```
##              chisq df    p
## greywage     1.170  1 0.28
## age          0.363  1 0.55
## extraversion 1.218  1 0.27
## way          0.314  2 0.85
## GLOBAL       2.741  5 0.74
```

```r
ggcoxzph(cox_test)
```



As we can see from the cox.zph results, none of our concerns were statistically significant. We can also note that the two covariates that we had the largest concerns for were the most significant in the cox.zph test which indicates we had a good interpretation of our plots. Since we now know that our model meets cox proportional hazards assumptions, we do not need to look at possible stratified or time varying models. Instead, we feel that it may be a good idea to take a look at possible interaction terms that may be useful to our model.

```r
interaction.1.model<-coxph(turn.surv~ age + extraversion + way*greywage, data = turnover)
interaction.2.model<-coxph(turn.surv~ extraversion + way + age*greywage, data = turnover)
interaction.3.model<-coxph(turn.surv~ age + way + extraversion*greywage, data = turnover)
interaction.4.model<-coxph(turn.surv~ greywage + extraversion + way*age, data = turnover)
interaction.5.model<-coxph(turn.surv~ greywage + way + extraversion*age, data = turnover)
interaction.6.model<-coxph(turn.surv~ age + greywage + way*extraversion, data = turnover)

anova(interaction.1.model) #definitely don't need interaction
```

```
## Analysis of Deviance Table
##  Cox model: response is turn.surv
## Terms added sequentially (first to last)
##
##                 loglik   Chisq Df Pr(>|Chi|)
## NULL           -3470.3
## age            -3464.0 12.6361  1  0.0003784 ***
## extraversion -3459.7  8.6208  1  0.0033234 **
## way          -3453.2 12.8631  2  0.0016100 **
## greywage     -3445.1 16.1850  1  5.745e-05 ***
## way:greywage -3445.1  0.0620  2  0.9694788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(interaction.2.model)
```

```
## Analysis of Deviance Table
##  Cox model: response is turn.surv
## Terms added sequentially (first to last)
##
##                loglik   Chisq Df Pr(>|Chi|)
## NULL          -3470.3
## extraversion -3467.3  6.0142  1   0.014192 *
## way          -3461.2 12.1389  2   0.002312 **
## age          -3453.2 15.9670  1  6.446e-05 ***
## greywage     -3445.1 16.1850  1  5.745e-05 ***
## age:greywage -3444.8  0.6227  1   0.430037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(interaction.3.model)
```

```
## Analysis of Deviance Table
##  Cox model: response is turn.surv
## Terms added sequentially (first to last)
##
##                         loglik   Chisq Df Pr(>|Chi|)
## NULL                   -3470.3
## age                    -3464.0 12.6361  1  0.0003784 ***
## way                    -3457.6 12.7505  2  0.0017032 **
## extraversion           -3453.2  8.7334  1  0.0031244 **
## greywage               -3445.1 16.1850  1  5.745e-05 ***
## extraversion:greywage -3445.0  0.3133  1  0.5756922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(interaction.4.model)
```

```
## Analysis of Deviance Table
##  Cox model: response is turn.surv
## Terms added sequentially (first to last)
##
##                loglik   Chisq Df Pr(>|Chi|)
## NULL          -3470.3
## greywage     -3460.9 18.8210  1  1.436e-05 ***
## extraversion -3457.4  6.9239  1  0.0085053 **
## way          -3452.4 10.0369  2  0.0066146 **
## age          -3445.1 14.5232  1  0.0001384 ***
## way:age      -3442.8  4.6163  2  0.0994472 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(interaction.5.model)
```

```
## Analysis of Deviance Table
##  Cox model: response is turn.surv
## Terms added sequentially (first to last)
##
##                    loglik   Chisq Df Pr(>|Chi|)
## NULL              -3470.3
## greywage          -3460.9 18.8210  1  1.436e-05 ***
## way               -3456.0  9.8622  2  0.0072186 **
## extraversion      -3452.4  7.0986  1  0.0077144 **
## age               -3445.1 14.5232  1  0.0001384 ***
## extraversion:age -3444.5  1.2131  1  0.2707122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(interaction.6.model)
```

```
## Analysis of Deviance Table
##  Cox model: response is turn.surv
## Terms added sequentially (first to last)
##
##                    loglik   Chisq Df Pr(>|Chi|)
## NULL              -3470.3
## age               -3464.0 12.6361  1  0.0003784 ***
## greywage          -3455.2 17.4778  1  2.907e-05 ***
## way               -3450.0 10.4612  2  0.0053502 **
## extraversion      -3445.1  9.7299  1  0.0018130 **
## way:extraversion -3444.7  0.9915  2  0.6091269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the analysis of deviance table of all the possible interaction models, we can see that none of the interaction terms are significant at a $\alpha = 0.05$ significance level. The only interaction that possibly seems viable is the interaction between way and age, but it also appears to decrease our log likelihood, so we will leave it out of our model.With our definite final model set, we will look at how each covariate affects the hazard rate of employment.

```
test_model
```

```
## Call:
## coxph(formula = turn.surv ~ greywage + age + extraversion + way,
##     data = turnover)
##
##                     coef exp(coef)  se(coef)      z        p
## greywagewhite -0.544968  0.579860  0.127015 -4.291 1.78e-05
## age            0.023756  1.024040  0.006132  3.874 0.000107
## extraversion   0.072605  1.075305  0.023205  3.129 0.001755
## waycar        -0.234916  0.790637  0.093639 -2.509 0.012116
## wayfoot       -0.400204  0.670183  0.162536 -2.462 0.013807
##
## Likelihood ratio test=50.31  on 5 df, p=1.2e-09
## n= 1129, number of events= 571
```

```
model.summary<-summary(test_model)
intervals<-model.summary$conf.int
intervals[,c(1,3,4)]
```

```
##               exp(coef) lower .95 upper .95
## greywagewhite 0.5798602 0.4520721 0.7437704
## age           1.0240399 1.0118061 1.0364217
## extraversion  1.0753054 1.0274942 1.1253415
## waycar        0.7906373 0.6580692 0.9499113
## wayfoot       0.6701834 0.4873521 0.9216042
```

The first thing that stands out about these intervals of the hazard ratios is that none of them contain 1. This means that all of our covariates have significant hazard ratios compared to the baseline hazard rate. Specifically, when looking at greywage, employees with white wage, meaning a 100% legal wage, had a hazard rate between 45.2% and 74.4% of the hazard rate for employees with a grey wage, meaning they had a mix of legal and illegal wage.This indicates that employees were much more likely to stay at the job if their wage was 100% legal. This could be for a variety of reasons ranging from employees not being comfortable with receiving illegal compensation for work, to the idea that companies intend for grey wage jobs to be more temporary. For age we see that the interval for the hazard ration is strictly greater than 1 which indicates that as age increases, the hazard rate for the employee also increases meaning they are more likely to leave their job as they get older. The same type of interpretation can be given for extraversion where the more extroverted an employee is, the more likely they are to leave their job. When looking at the way covariate, we need to interpret the intervals for 'car' and 'foot' compared to employees who took the bus.First, when looking at 'car' we notice that the hazard rate for employees that drove a car to work was between 65.8% and 95% of the hazard rate for employees that took the bus. Second, employees that walked to work have a hazard rate that is between 48.7% and 92.2% of the hazard rate for employees that took the bus.From both these intervals we can interpret that employees that take the bus to work are more likely to leave their job

Personality Analysis

With the personality score data in our dataset, we were interested to see if and how people's personalities affected their decision of quitting.

Focusing on the personality columns (i.e. extraversion, independence, selfcontrol, anxiety, novator), these variables were scored on a scale of 1 to 10 (i.e. 10 in extraversion means a very extroverted person, 1 means an introverted person).

Assumptions were made about what personality attributes contribute more to someone quitting or staying at a job:

- Highly independent people tend to quit more, since they are less likely to rely on things like jobs or conform to company rules or hierarchy. Because independent people are assumed to be more confident, they are more likely to feel brave enough to quit.
- Looking at extraversion, we believed an extroverted person would be more likely to quit. Extroverted people are typically more comfortable voicing their opinions and tend to show more of their emotions.
- Looking at self control, people with a low rating are more likely to quit, since they are prone to making rash decisions.
- A person with high anxiety would be more prone to staying at a job, instead of quitting. Someone with high anxiety is assumed to be more scared to quit a job, possibility due to a lack of steady income, inability to pay bills, and other stresses caused by unemployment.

- Looking at novator, someone who is more innovative is more likely to quit their job because they crave new and exciting things. Their old job may become repetitive and tedious or they do not see potential for growth in their current company.

Principal Components Analysis on Personality Ratings

```
# Personality columns
personality <- turnover[12:16]
```

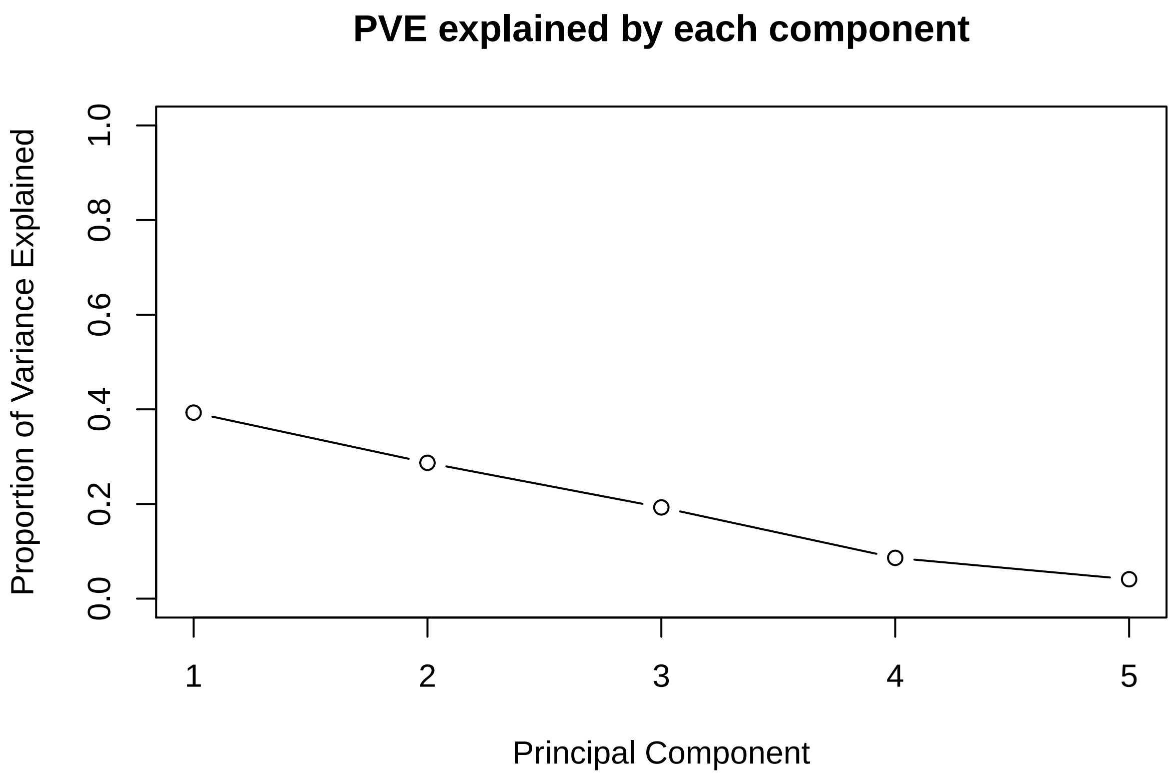
Running PCA, we obtained 5 principal components, PC1-PC5.

```
pr.out = prcomp(personality, scale=TRUE)
summary(pr.out)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.4019 1.1979 0.9818 0.65631 0.45257
## Proportion of Variance 0.3931 0.2870 0.1928 0.08615 0.04096
## Cumulative Proportion  0.3931 0.6801 0.8729 0.95904 1.00000
```

Each explain a percentage of the total variation in the dataset. PC1 explains about 39% of the total variance, PC2 explains about 29% of the total variance, and so forth. The values in the plots correspond with the values seen above from *prcomp*.

```
pr.var=pr.out$sdev^2
pve=pr.var/sum(pr.var)
plot(pve, main="PVE explained by each component", xlab="Principal Component",
     ylab="Proportion of Variance Explained", ylim=c(0,1), type="b")
```



Since PC1 and PC2 had the highest proportion of variance explained, we focused on these two principal components.

The center and scale components of our PCA object correspond to the means and standard deviations of the variables that were used prior to using PCA. It appeared that all of the personality ratings had similar means.

```
pr.out$center
```

```
## extraversion     independ  selfcontrol      anxiety      novator
##     5.592383     5.478034     5.597254     5.665633     5.879628
```

Ultimately, *novator* had the highest mean, so we saw that people only have a slightly higher rating for *novator* versus the other attributes.

We saw that all the variables have similar standard deviations as well.

```
pr.out$scale
```

```
## extraversion     independ  selfcontrol      anxiety      novator
##     1.851637     1.703312     1.980101     1.709176     1.904016
```

Ultimately, *selfcontrol* had the highest standard deviation, so there is more variation between ratings when it comes to self control.

The rotation matrix gives the principal component loadings, with each column containing the corresponding principal component loading vector. This shows the relationship between the initial variables and the principal components.

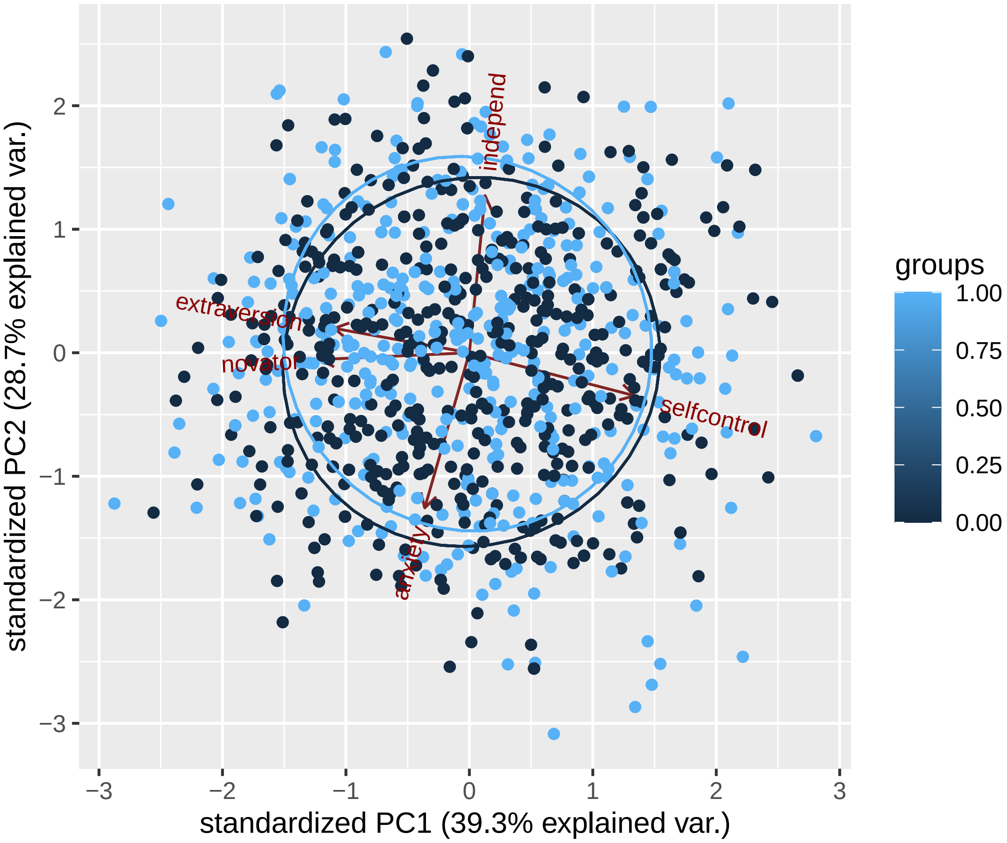
```
pr.out$rotation
```

```
##                      PC1         PC2        PC3         PC4         PC5
## extraversion -0.51020478  0.10788824  0.6500410 -0.08788515 -0.54568687
## independ      0.05922318  0.69415891 -0.4933164 -0.22558418 -0.46945369
## selfcontrol   0.62383167 -0.18979361  0.1162372  0.49532234 -0.56210064
## anxiety      -0.16855764 -0.68526874 -0.4127052 -0.41111425 -0.40330454
## novator      -0.56445201 -0.02981082 -0.3876197  0.72596751 -0.05681016
```

Looking at PC1 loadings, we saw that extroversion, self control, and novator had the heaviest weights. Looking at PC2 loadings, independence and anxiety had the heaviest weights. In the Cox Model, we saw that extraversion had the third largest effect on employee turnover. Since PC1 explained the most variance and extraversion had a higher weight in that principal component, we can say that extraversion plays a large role in employee turnover, thus following our Cox Proportional Hazards Model.

Using a biplot with PC1 and PC2, we visualized how the samples relate to each other in our PCA, while revealing how each variable contributes to each principal component.

Specifically, we created ellipses based on the *event* of the dataset, allowing us to create two groups: people who quit and people who stayed at their job.



Looking at the legend, 1.00 (light blue) represents the value of “1” in the *event* column of the dataset, marking the event of quitting. 0.00 (dark blue) represents the value of “0” in the *event* column, marking the event of not quitting. The light blue dots represent the group of people who quit, and the dark blue dots represent the those who did not quit.

The two ellipses illustrate the overall trend the two groups have. The light blue circle, representing the people who quit, is slighter higher in placement compared to the dark blue circle, representing the people who did not quit. This illustrates that people who quit trend towards certain attributes and people who stayed lean towards other attributes.

Independence is pointed upward and extraversion is pointing left and slightly upward. Since the light blue circle is higher than the dark blue circle, this means that people who quit tend to be more independent and slightly more extroverted. Following our previous assumptions, this makes sense since independent and extroverted people are more confident in their own abilities or brave enough to voice concerns and emotions. Since unemployment may cause someone to feel insecure or vulnerable, independence and extraversion may contribute to someone feeling more comfortable voicing their opinions and confident enough to quit. This coincided with our Cox Proportional Hazards model, seeing that people with high ratings in extraversion had higher hazard proportions.

Anxiety is pointed downward, self control is pointing right and slightly downward, and novator is pointing left and slightly downward. Representing people who stayed at a job, the dark blue circle is lower in placement, going towards the anxiety, novator, and self control arrows. Unlike our previous assumption, someone with high ratings in novator was actually more likely to stay at a job. Coinciding with our other previous assumptions, people who stayed at a job tend to have higher ratings in anxiety and self control. Since people with high anxiety would be too nervous to quit, possibily worried about the repercussions of unemployment. Someone with more self control is able to regulate emotions, thoughts, or behavior, ultimately less likely to make a rash or impulsive decision like quitting.

From Principal Component Analysis, we saw that attributes like extraversion, novator, independence, self control, and anxiety impacted employee turnover. Ultimately, we are able to conclude that people’s personalities have an affect on their decision of quitting a job.

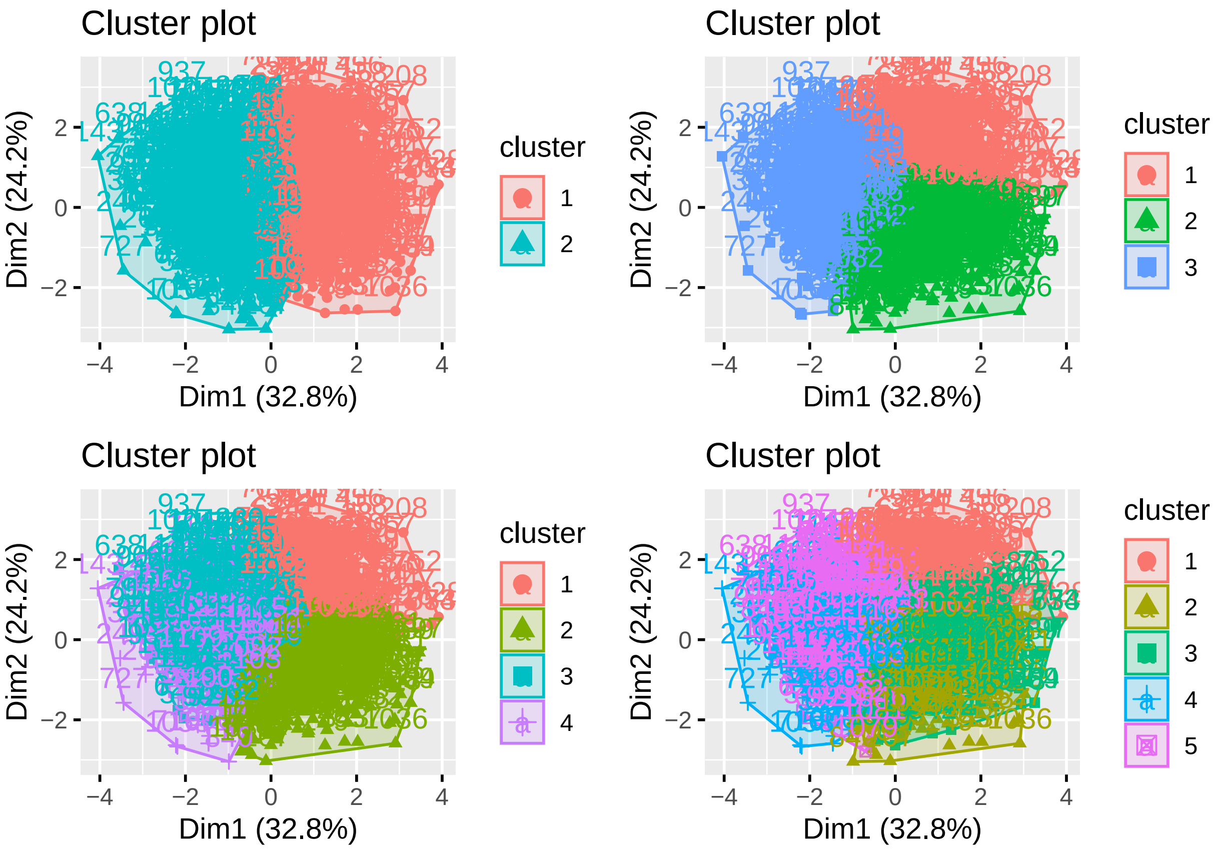
K-means clustering Analysis

When deciding which method of clustering analysis to use, I considered using hierarchical clustering, but due to the sheer number of observations and not knowing what levels of each personality type were normal for the company, it made it difficult to prune the dendrograms to illicit comprehensible conclusions. Moreover, since the personality traits were all measured on a numeric scale form 0 to 10, it was easier to use k-means as a method of cluster analysis.

```r
# Standardize the variables by subtracting mean and divided by standard deviation
sturnover = scale(turnover[, -c(1,3:11)], center=TRUE, scale=TRUE)
```

Before running tests to determine the optimal k-value, I did a cursory analysis by creating clustering models using 2, 3, 4, and 5 different centroids to see what the clustering distribution looked like visually.

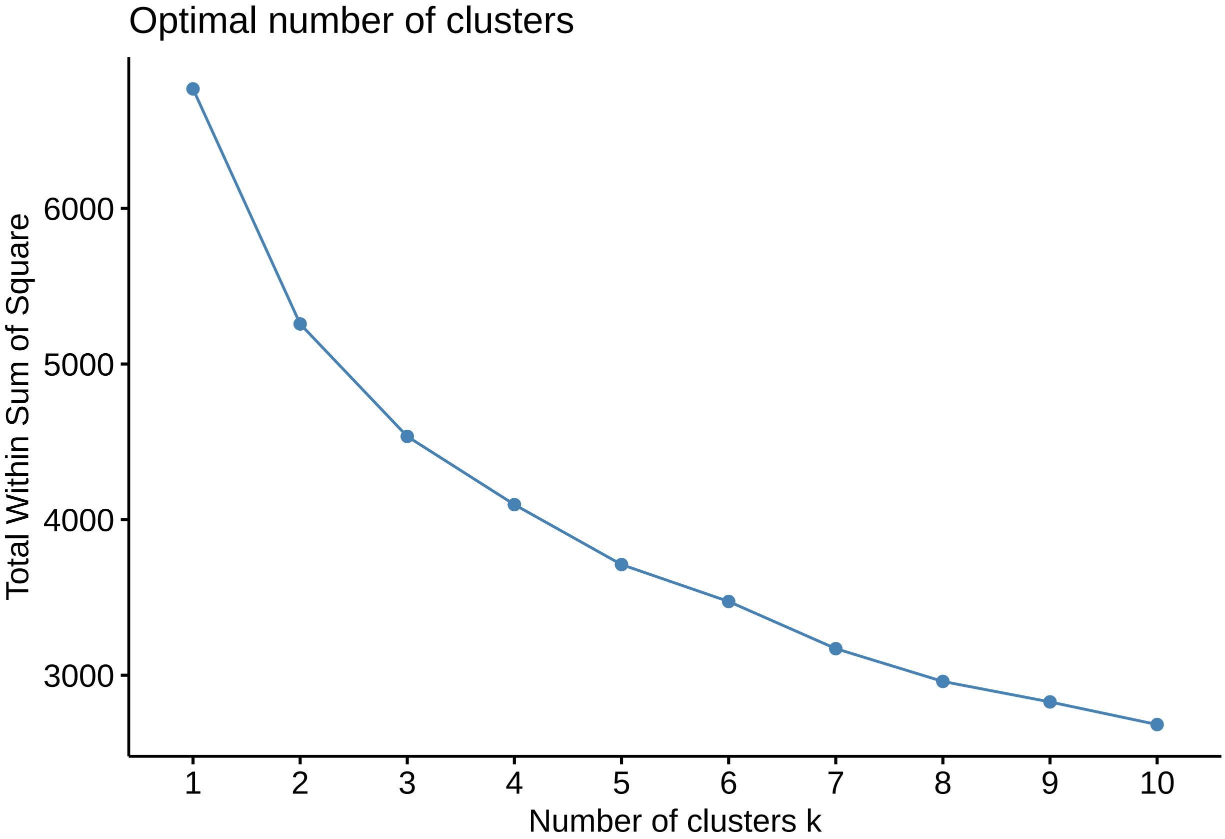
```r
# Tried a few different types of k-values to see what looked
# correct visually and create comparison models
set.seed(1)
km2 = kmeans(sturnover, centers = 2, nstart = 25)
km3 = kmeans(sturnover, centers = 3, nstart = 25)
km4 = kmeans(sturnover, centers = 4, nstart = 25)
km5 = kmeans(sturnover, centers = 5, nstart = 25)
# set nstart = 25 to have multiple initialization to
# ensure that the centroid value is accuarate
```



Picking the Optimal Number of Clusters

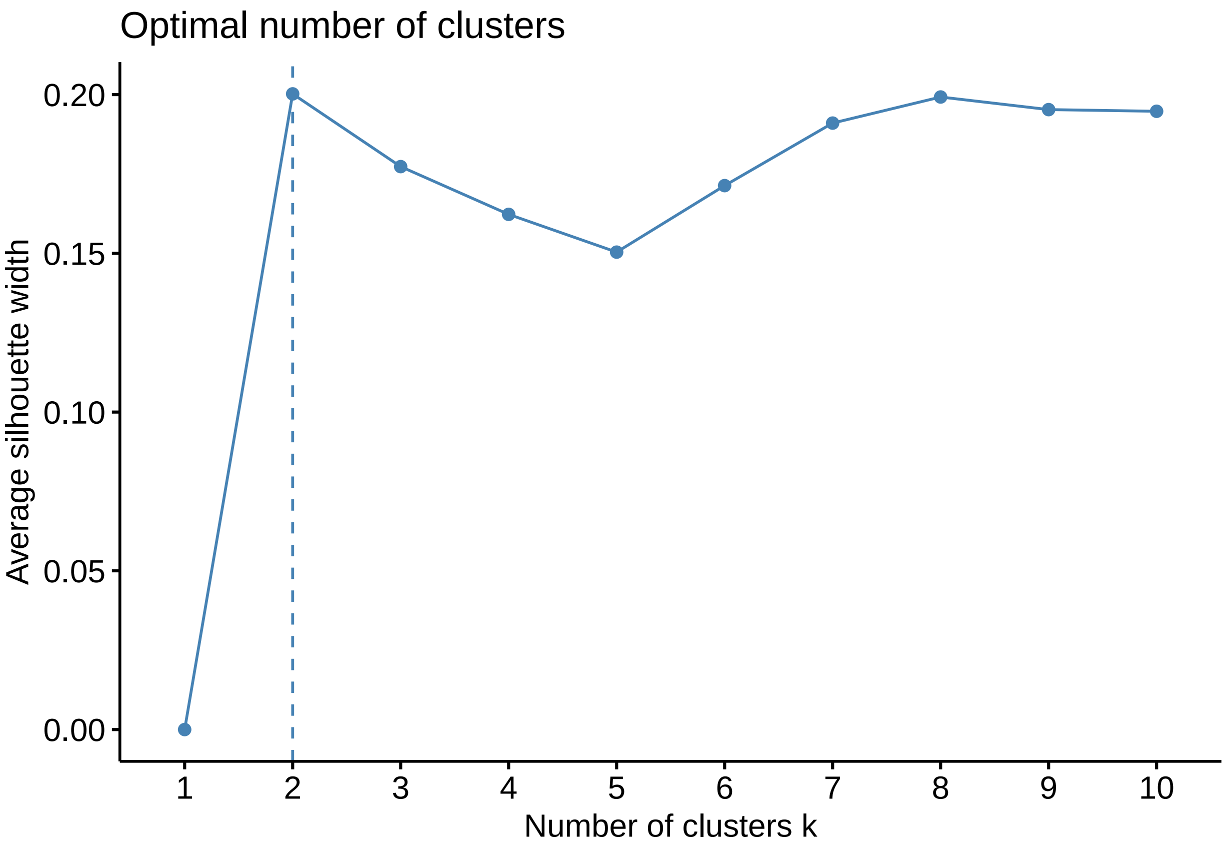
In order to pick the optimal number of centroids, or the optimal k-value, I used two verification methods, the elbow method and the silhouette method. The elbow method works by calculating the within sum of squares, or within cluster variation, which measures the compactness of data points within the cluster. As with normal variation, we want to minimize the within sum of squares value. The optimal k-value indicated graphically by the elbow method is visually apparent by seeing where there is a bend in the plot. This bend can also be observed by seeing at which point the change in the total within sum of squares value becomes less drastic and the graph tapers off.

Using the elbow method, we produce the following graph:



Looking at the graph, we see that there is little change going from 5 clusters to 6 clusters, suggesting that 5 clusters might be the optimum number of centroids. Given that there isn't a particularly well defined elbow in the graph, we opted to have 5 centers to match the 5 personality traits that we are comparing on.

The second verification method used is the silhouette method. This method measures the quality of the cluster or how well each observation fits into the cluster. A high average silhouette width indicates good clustering, therefore it suggest an optimal number of clusters by selecting the one which produces the largest average silhouette width value indicated by a vertical dashed line.



Using the silhouette method observed in the above graph, we find that the optimal number of clusters is 2, with 8 clusters being the next optimal amount.

I tested doing clustering using 2 centroids, but due to the number of personality traits being 5, it made it difficult to understand which personality traits had the largest effect on an employee's decision to quit. Using 8 clusters seemed unnecessarily large since this surpasses our number of personality traits. Because of this, I chose to use 5 centroids, one for each personality trait, in the final model as it made the results more comprehensible to understand.

```r
# creating the final model using the found optimal k-value
set.seed(1)
final<- kmeans(sturnover, 5, nstart = 25)
final
```

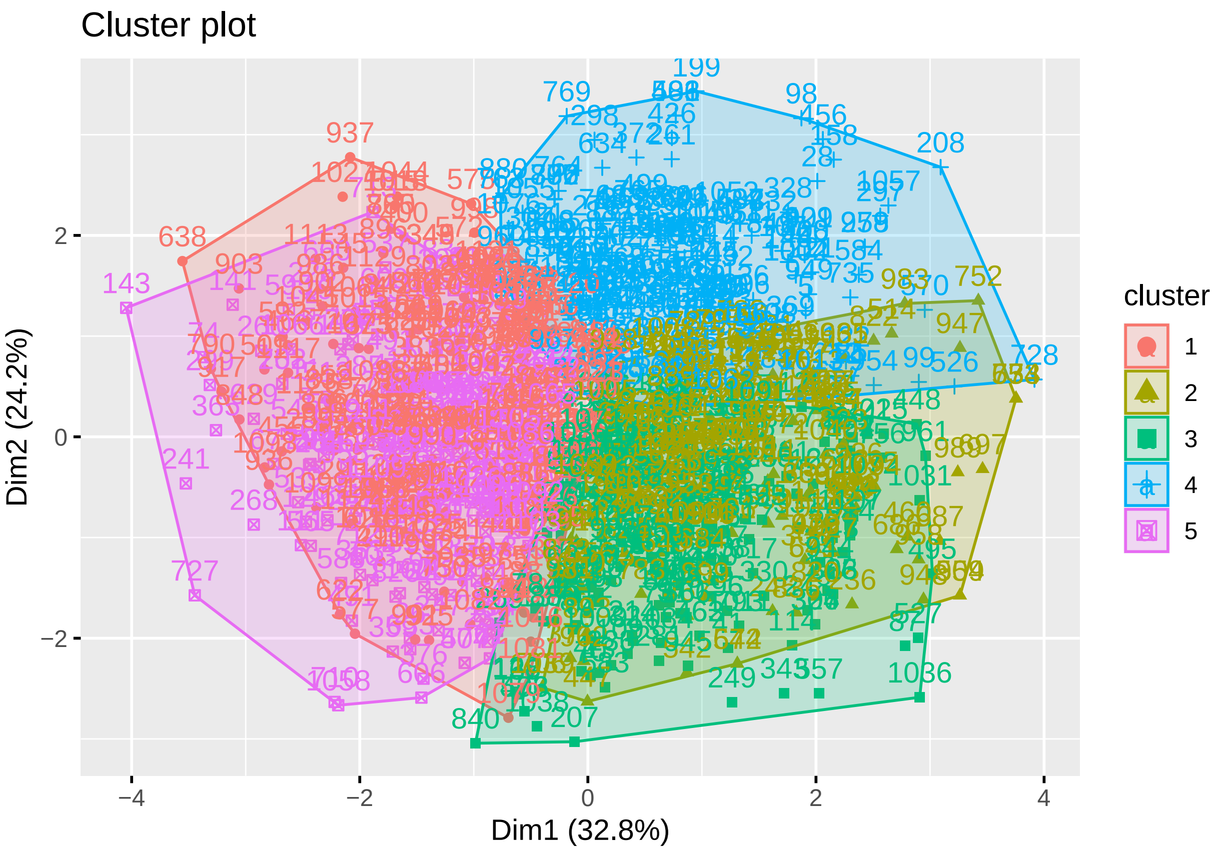
```
## K-means clustering with 5 clusters of sizes 256, 208, 238, 206, 221
##
## Cluster means:
##         event extraversion    independ selfcontrol    anxiety    novator
## 1 -1.01113358    0.5922759 -0.03663959  -0.7533886  0.2580234  0.7086487
## 2 -1.01113358   -0.5120615  0.42611830   0.6909398 -0.5925362 -0.7912493
## 3  0.98811303   -0.4025216  0.63994938   0.3563894 -0.6623191 -0.4708126
## 4  0.07583545   -0.5778955 -1.08234278   0.7905492  0.9070957 -0.3786156
## 5  0.98811303    0.7680216 -0.03890593  -0.8982887  0.1265330  0.7837738
##
## Clustering vector:
##    [1] 5 5 5 3 4 5 5 3 5 5 5 3 3 5 5 5 5 5 5 4 3 5 3 4 5 4 4 3 3 5 4 3 5 5 4 3
##   [38] 3 3 3 3 3 3 3 3 3 5 3 3 3 3 4 5 5 5 4 5 5 5 5 5 5 4 3 3 3 4 4 4 3 5 3 5 5
##   [75] 5 3 1 1 1 3 4 3 5 3 3 3 3 3 5 5 5 5 4 3 3 3 3 4 4 3 5 5 5 5 5 5 4 4 4 3 5
##  [112] 5 3 3 5 3 3 5 5 5 3 4 3 3 3 3 3 3 5 5 4 3 3 3 3 5 3 4 3 5 5 5 5 4 4 4 4 3
##  [149] 3 3 3 3 5 5 4 3 5 4 4 4 5 3 2 5 3 3 3 3 5 3 3 5 5 3 5 3 5 3 3 3 4 3 5 3 4
##  [186] 3 5 5 4 3 3 3 5 4 3 3 3 4 4 3 3 3 5 5 3 3 3 4 5 5 5 4 4 3 5 5 4 4 2 1 4 4
##  [223] 5 1 3 5 1 1 2 4 3 3 3 5 3 2 2 3 4 2 5 5 5 3 2 2 5 3 3 5 5 5 2 4 3 5 4 4 3
##  [260] 5 4 4 5 2 2 3 2 5 3 4 4 5 3 5 4 4 3 5 5 5 3 1 5 1 5 5 3 3 1 1 1 4 1 2 5 1
##  [297] 4 4 5 5 4 1 3 3 3 3 4 2 1 2 4 4 3 2 4 3 3 1 5 3 3 5 4 3 2 3 3 4 4 3 3 3 2
##  [334] 2 3 1 4 5 5 1 4 5 3 5 1 1 4 1 1 4 1 4 2 2 3 4 4 4 5 1 1 3 5 5 5 5 2 4 4 5
##  [371] 2 4 3 3 2 5 5 2 5 4 3 2 1 2 5 3 1 2 1 4 3 3 3 3 3 4 3 4 5 1 5 1 4 4 3 3 1
##  [408] 1 4 3 1 1 5 3 2 1 2 2 2 2 1 1 2 1 1 4 2 4 4 1 1 1 1 3 1 3 2 4 1 3 1 1 2 4
##  [445] 3 1 2 3 5 3 2 1 1 3 1 4 5 5 5 2 1 3 3 5 5 1 2 5 5 5 5 1 3 5 2 3 3 4 2 4 4
##  [482] 3 4 5 1 4 4 1 4 5 5 1 3 5 3 3 5 4 4 2 3 1 3 1 2 3 5 3 1 1 2 3 5 2 1 1 1 1
##  [519] 1 2 1 1 3 3 3 4 3 2 5 1 4 4 5 2 1 1 5 3 5 5 5 2 2 2 1 2 2 1 2 2 2 3 4 2 3
##  [556] 2 3 3 2 2 3 5 3 4 5 3 4 4 4 4 4 1 1 5 1 1 1 4 4 5 3 5 3 4 1 5 5 4 1 5 5 1
##  [593] 5 3 5 1 1 2 2 1 4 4 2 2 2 5 5 2 1 3 5 4 1 2 1 1 3 1 4 3 1 1 5 2 1 1 3 1 5
##  [630] 3 3 2 5 4 2 3 3 1 1 3 2 1 2 2 2 2 1 4 2 2 1 1 5 1 4 4 4 3 2 5 1 5 4 4 2 2
##  [667] 1 5 2 2 2 2 2 1 2 1 5 2 3 5 1 5 1 2 1 2 2 4 1 5 3 5 3 1 2 2 2 1 5 2 1 4 5
##  [704] 3 5 2 5 2 2 5 1 5 5 5 1 2 2 1 4 1 1 2 1 5 5 3 5 4 1 1 3 5 1 2 4 2 2 3 4 4
##  [741] 1 3 1 2 2 1 1 5 4 1 1 2 1 1 2 3 2 1 5 4 2 2 4 4 4 1 5 5 4 4 3 2 1 2 2 2 4
##  [778] 4 4 4 2 3 2 3 1 2 1 2 2 1 1 4 3 1 1 1 2 1 5 1 2 1 1 5 3 4 1 1 1 1 5 3 1 5
##  [815] 4 2 3 4 1 2 2 2 2 3 3 5 2 3 4 4 2 4 1 1 5 2 1 1 1 3 2 1 4 2 1 4 4 1 4 1 3
##  [852] 3 3 1 4 4 1 1 1 1 1 1 2 4 4 4 4 1 2 1 2 2 4 2 1 4 2 2 2 4 1 3 5 5 5 1 4 1
##  [889] 2 4 1 1 2 1 3 1 1 5 2 4 4 1 1 2 2 2 2 1 2 2 5 2 2 3 2 1 1 2 2 2 1 2 2 2 1
##  [926] 2 2 2 4 4 2 2 1 5 1 1 1 1 2 2 2 1 5 3 3 2 2 2 4 1 2 5 2 4 5 2 1 2 1 4 3 2
##  [963] 2 2 4 5 4 1 4 3 1 4 3 2 4 2 1 2 4 1 2 2 2 1 2 1 4 4 2 1 1 1 2 1 1 3 1 1 1
## [1000] 5 2 4 2 4 2 1 2 2 3 4 3 4 4 4 1 1 1 1 4 1 1 1 1 1 5 1 1 5 1 3 4 2 1 1 3
## [1037] 2 3 2 2 5 5 1 1 1 1 3 1 4 2 4 4 4 5 4 4 4 5 3 3 4 4 1 1 5 5 1 1 2 2 3 3 1
## [1074] 4 1 4 1 1 1 3 1 1 1 3 4 5 1 1 1 2 3 5 3 2 3 5 1 1 1 5 4 4 5 2 5 2 3 4 5 4
## [1111] 2 2 1 2 1 1 1 4 1 4 1 2 4 1 1 1 2 2 1
## 
## Within cluster sum of squares by cluster:
## [1] 804.6864 643.4102 764.8607 818.0442 627.7269
##  (between_SS / total_SS =  45.9 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# prints out a summary of the clusters, the cluster means
# for each personality trait, and the within cluster sum of squares
```

Obtaining the summary of our final clustering model, we find that the sizes of all the clusters are relatively similar which indicates that the clusters aren't disproportionate.

Since our event variable is binary with 0 being didn't quit and 1 being that the individual did quit, we are using how close the expected value of event is to determine which clusters contain individuals that are more likely to quit and those that are not. From our mean event values found in the model summary, we can see that clusters 1, 2, and 4 have average event values closer to 0 indicating that they are less likely to quit than those who are in clusters 3 and 5, which have average event values closer to 1. Using this observation, we can draw conclusions about the personality types of those who are more likely to stay with the company by observing the personality trait analysis associated with the observations in clusters 1, 2, and 4 and those who are more likely to quit as seen with the observations in clusters 3 and 5.

Final Cluster Plot



Interactive Clustering Plot

```r
turnover$cluster <- as.factor(final$cluster)
p <- ggparcoord(data = turnover, columns = c(12:16), groupColumn = "cluster",
                scale = "std") + labs(x = "personality traits",
                                      y = "value (in standard-deviation units)",
                                      title = "Clustering")
ggplotly(p)
```

```r
# code for interactive plot found from Towards Data Science
```

The graph above reports its results utilizing standard deviations from the mean. Keeping this in mind, we can make generalizations about the personality traits of 5 general types of individuals as denoted by the clusters by seeing how much they deviate from the average reported value associated with each trait. From the interactive graph above, we see that for the most part, employees who are less likely to quit which are those included in clusters 1, 2, and 4, tend to be introverted, independent, have slightly more self control than average, are more anxious, and slightly more innovative than those who are more likely to quit, as observed in clusters 3 and 5.

For the most part, these findings aligned with our original assumptions, with the exception of independence, self-control, and innovation (novator).

Conclusion

Using various analysis methods, we attempted to determine which attributes had an effect on employee turnover rate. In terms of all of our covariates, during our model selction process, we discovered that the most important attributes overall were greywage, extraversion, age, and way, or the method of commuting used by the individual. When we focused our analysis on personality traits, using PCA we confirmed that extraversion has a large impact on employee turnover and also determined that novator, independence, self control, and anxiety also affected employee turnover. Finally, using k-means clustering analysis, we were able to generalize a set of personality traits that are more conducive to an individual quitting. Specifically, we found that individuals who are more extraverted, have lower than average self-control, and are not anxious are more likely to quit. This assertion is supported by the findings of both machine learning techniques used in the personality analysis.

References

- Employee Turnover Dataset: https://www.kaggle.com/davinwijaya/employee-turnover
- PCA biplot graph: https://www.datacamp.com/community/tutorials/pca-analysis-r?utm_source=adwords_ppc&utm_campaignid=1565261270&utm_adgroupid=67750485268&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=332661264371&utm_targetid=aud-522010995285:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=9032048&gclid=CjwKCAiA8Jf-BRB-EiwAWDtEGqk_UZHaEqMCnyaI7D2zKAr6_wdpCoQjf6OZXeHbT3b8gnPLFeXjjxoCt5EQAvD_BwE
- Github Repository for ggbiplot function: https://github.com/vqv/ggbiplot
- Interactive clustering plot: https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967