PCA exercises

Huey Kwik

May 16, 2016

The following solutions are adapted from the work of Huey Kwik (Signal Cohort #2).

PCA on the msq dataset

Extract columns active:scornful from the msq dataset.

```
df = msq
df = select(df, Extraversion, Neuroticism, active:scornful)
# Count number of NAs in each column
colSums(is.na(df))
```

```
## Extraversion  Neuroticism       active       afraid        alert
##            6            6            6            5           11
##        angry      anxious      aroused      ashamed   astonished
##            9         1849            6           11           13
##      at.ease      at.rest    attentive         blue        bored
##           17           17            6            5            4
##         calm     cheerful  clutched.up    confident      content
##           82         1850           23            7           22
##    delighted    depressed   determined   distressed       drowsy
##            6           17            7            8           12
##         dull       elated    energetic enthusiastic      excited
##            9           15            6            6            6
##      fearful   frustrated  full.of.pep       gloomy      grouchy
##           20           11           12           12            5
##       guilty        happy      hostile         idle     inactive
##            5           16           11         1848         1846
##     inspired      intense   interested    irritable      jittery
##            6            7           12           16            6
##       lively       lonely      nervous       placid      pleased
##           10            6           17           19           13
##        proud    quiescent        quiet      relaxed          sad
##            7          136            5            7           10
##    satisfied       scared       serene       sleepy     sluggish
##            7           10           12           16            8
##     sociable        sorry        still       strong    surprised
##            6           15           12            7            6
##        tense        tired     tranquil      unhappy        upset
##           10           10         1843            5            8
##     vigorous      wakeful  warmhearted   wide.awake        alone
##           10           10            7           12         2058
##       kindly     scornful
##         2060         2058
```

```
# Throw out columns with huge number of missing values
df = df[,colSums(is.na(df)) <= 500]
# Remove all rows with any NAs
df = df[rowSums(is.na(df)) == 0,]
```

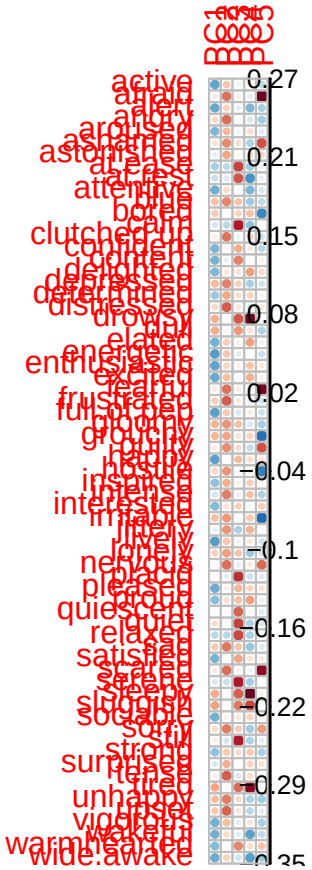
Run PCA on the remaining variables

Convenience method to print out the top 10 loadings of the nth PCA, ordered by absolute value.

```
top = function(n, p) {
  v = p$rotation[,n]
  v[order(abs(v), decreasing = TRUE)][1:10]
}
```

Run PCA on the remaining variables.

```
features = select(df, -Extraversion, -Neuroticism)
p = prcomp(features, scale = TRUE)

# Plot the PCA loadings for first 5-10 PCAs.
loadings = p$rotation[,1:5]
corrplot(loadings, is.corr = FALSE)
```



PC1

Likely interpretation: energetic happiness

```
top(1,p)
```

```
##       lively    energetic  full.of.pep        happy enthusiastic
##    0.1918454    0.1916053    0.1887963    0.1858400    0.1834828
##       active      excited        alert      pleased   wide.awake
##    0.1809240    0.1743085    0.1741707    0.1729260    0.1712993
```

PC2

Likely interpretation: negative tension

```
top(2,p)
```

```
##       tense  distressed  frustrated       upset     nervous      scared
##  -0.2103074  -0.2094405  -0.2091589  -0.2030660  -0.1937056  -0.1932364
##       angry      afraid     fearful clutched.up
##  -0.1921534  -0.1889102  -0.1877725  -0.1863390
```

PC3

Likely interpretation: calm ease

```
top(3,p)
```

```
##     serene      still       calm     placid      quiet    at.ease
## -0.2795391 -0.2753895 -0.2746859 -0.2553033 -0.2378515 -0.2342058
##    relaxed    at.rest     drowsy     sleepy
## -0.2297325 -0.2157899 -0.2125270 -0.2061455
```

PC4

Possible interpretation: tired or awake

```
top(4,p)
```

```
##     sleepy     drowsy      tired   sluggish    at.rest wide.awake
## -0.3469386 -0.3362867 -0.3151579 -0.2306464  0.2057954  0.1887401
##    wakeful      alert  attentive     elated
##  0.1873454  0.1575786  0.1572027 -0.1453231
```

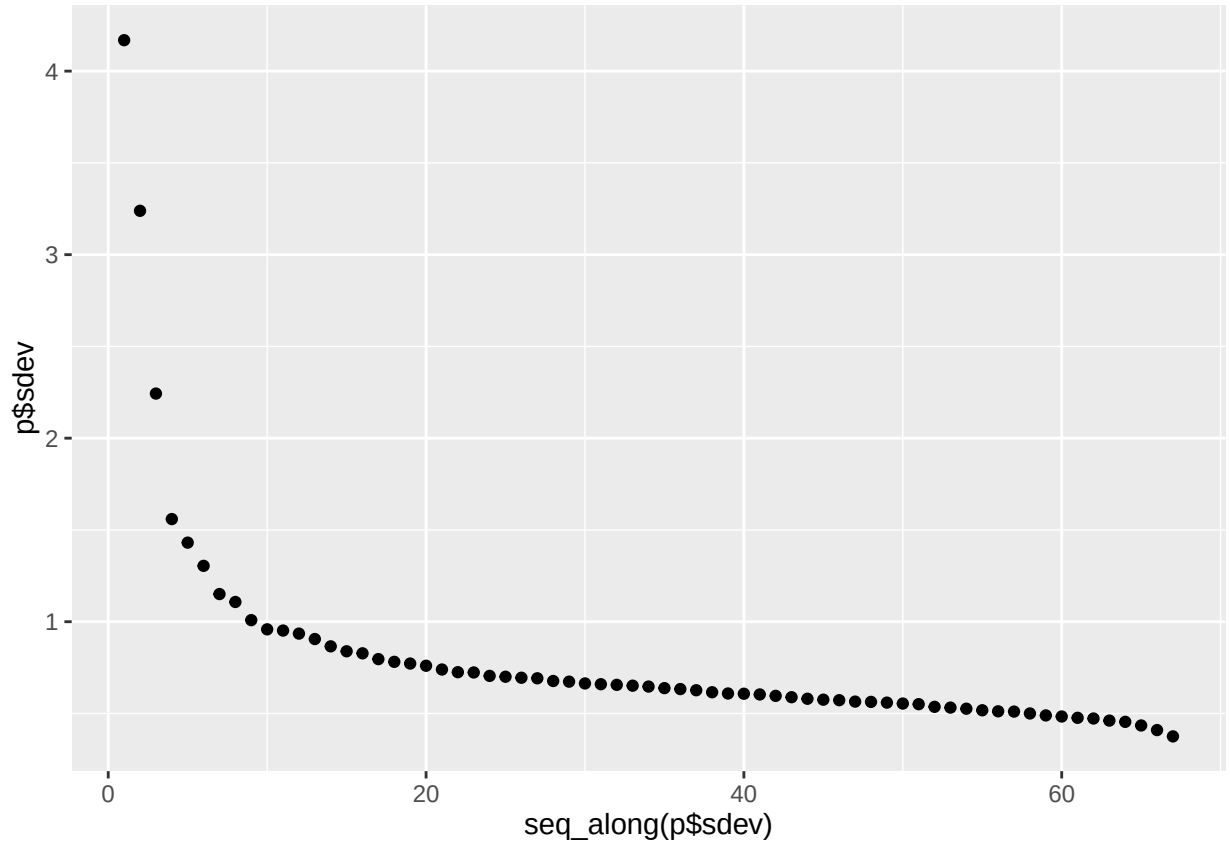
After PC4, it's difficult to see a representation of anything coherent.

Plot the eigenvalues

```
p$sdev
```

```
##  [1] 4.1685130 3.2388748 2.2430144 1.5593626 1.4308848 1.3044071 1.1507058
##  [8] 1.1076334 1.0085973 0.9585510 0.9518790 0.9352882 0.9055840 0.8658898
## [15] 0.8390250 0.8277808 0.7963318 0.7814332 0.7720824 0.7601401 0.7395894
## [22] 0.7250489 0.7235919 0.7049010 0.7002009 0.6948331 0.6917813 0.6771720
## [29] 0.6735138 0.6637272 0.6594581 0.6560347 0.6514811 0.6467385 0.6381087
## [36] 0.6330868 0.6267513 0.6157961 0.6089709 0.6076236 0.6037317 0.5960710
## [43] 0.5887581 0.5803601 0.5752520 0.5725270 0.5647907 0.5631319 0.5592432
## [50] 0.5541263 0.5507830 0.5363591 0.5321013 0.5260743 0.5175144 0.5119354
## [57] 0.5103474 0.5000729 0.4895749 0.4835413 0.4764388 0.4726846 0.4613825
## [64] 0.4545563 0.4348226 0.4095754 0.3754008
```

```
library(ggplot2)
qplot(,p$sdev)
```



Looking at the eigenvalues, the first three seem the most interpretable, and then there is a drop-off after that.

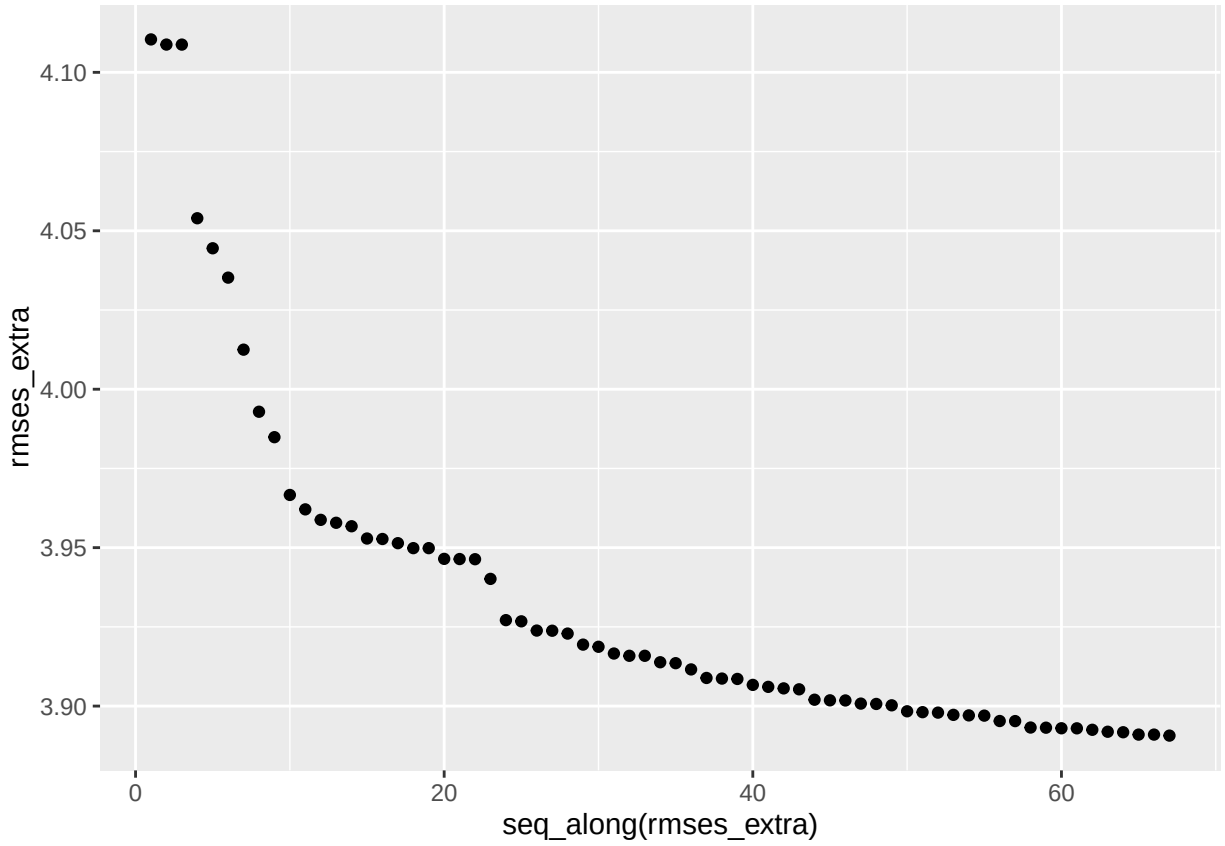
Principal component regression

Use the first n principal components to predict Extraversion and Neuroticism:

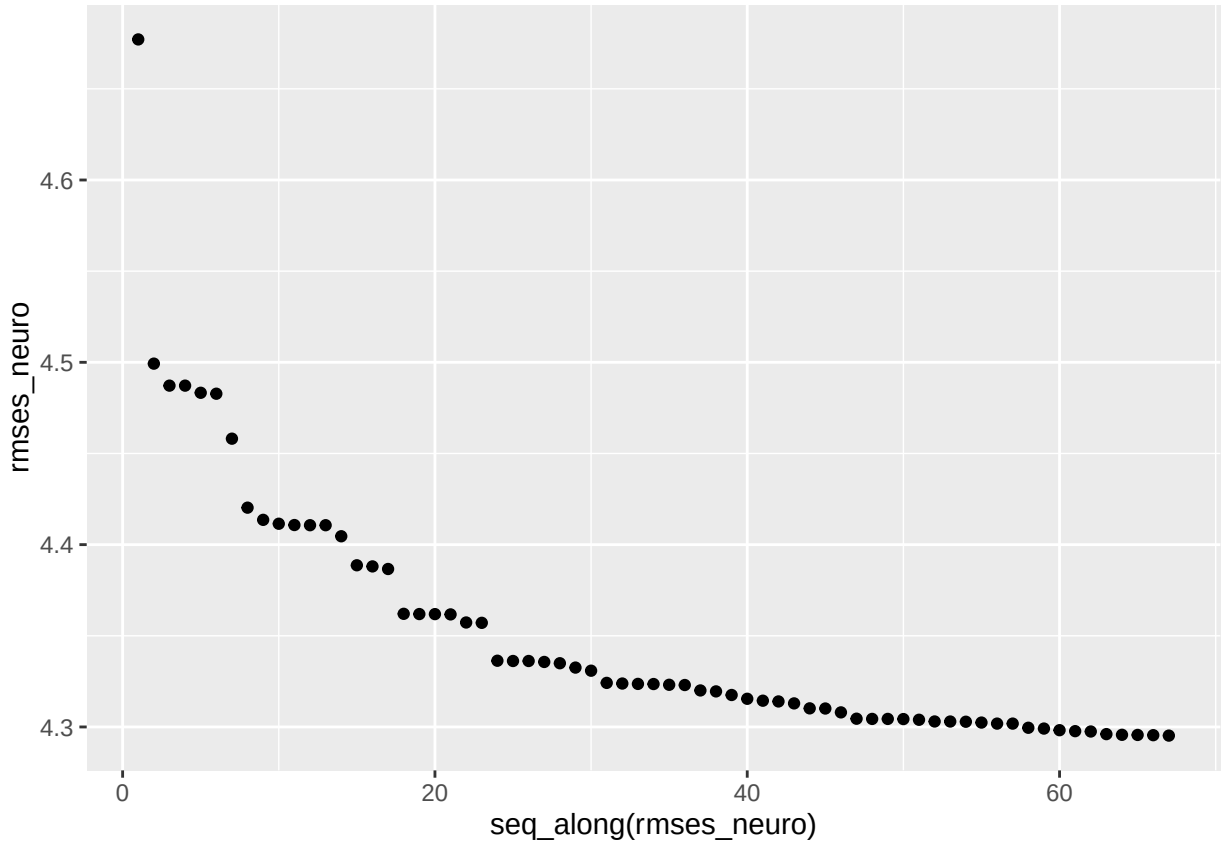
```
n = ncol(p$rotation)

rmses_extra = numeric(n)
rmses_neuro = numeric(n)

rmse = function(x, y) sqrt(mean((x - y)^2))

# What if colbind the PCAs
testDf = select(df, Extraversion:Neuroticism)

# Iterate through n
for (i in seq(n)) {
  testDf = cbind(testDf, p$x[,i])
  colnames(testDf)[ncol(testDf)] = paste0("PC", i)

  # Run a lm with PCA
  sumss = 0 # Hack to get around error message in cv.lm
  fit_extra = cv.lm(data=testDf, form.lm=formula(Extraversion~.-Neuroticism), plotit = FALSE, printit =
  fit_neuro = cv.lm(data=testDf, form.lm=formula(Neuroticism~.-Extraversion), plotit = FALSE, printit =

  # Store the RMSE
  rmses_extra[i] = rmse(fit_extra$Predicted, testDf$Extraversion)
  rmses_neuro[i] = rmse(fit_neuro$Predicted, testDf$Neuroticism)
}
```

Plot the RMSEs against n:

```
qplot(, y=rmses_extra)
```




```
qplot(, y=rmses_neuro)
```



From the self-assessment, the RMSEs for extraversion and neuroticism were 3.91 and 4.36, respectively. Note that these are similar to the RMSEs for when we use all the PCs. From looking at the plot, the RMSEs for using the first 4-5 PCs are higher.

History of Trait Theories

Neuroticism is a dimension that ranges from normal, fairly calm to one's tendency to be quite "nervous." Extraversion is a dimension that most people have a common-sense understanding of: Shy, quiet vs. out-going, loud.

```
lm(Extraversion~PC1+PC2+PC3+PC4, testDf)$coefficients
```

```
##  (Intercept)          PC1          PC2          PC3          PC4
## 13.510718492  0.138269766  0.035711657 -0.004119665 -0.428956116
```

```
lm(Neuroticism~PC1+PC2+PC3+PC4, testDf)$coefficients
```

```
##   (Intercept)           PC1           PC2           PC3           PC4
## 10.4244994111 -0.3591857893 -0.3944607788  0.1467824313  0.0006490632
```

Extraversion is positively correlated with PC1 (energetic) and PC2 (tension, slightly) and negatively correlated with PC3 (calm) and PC4 (tired). Neuroticism is negatively correlated with PC1 and PC2 and positively correlated with PC3 and PC4.

PCA on the speed dating dataset

Load the speed dating set.

```
df = read.csv("C:/Users/Andrew/Documents/Signal/curriculum/datasets/speed-dating/speeddating-aggregated
df = df[rowSums(is.na(df)) == 0,]
```

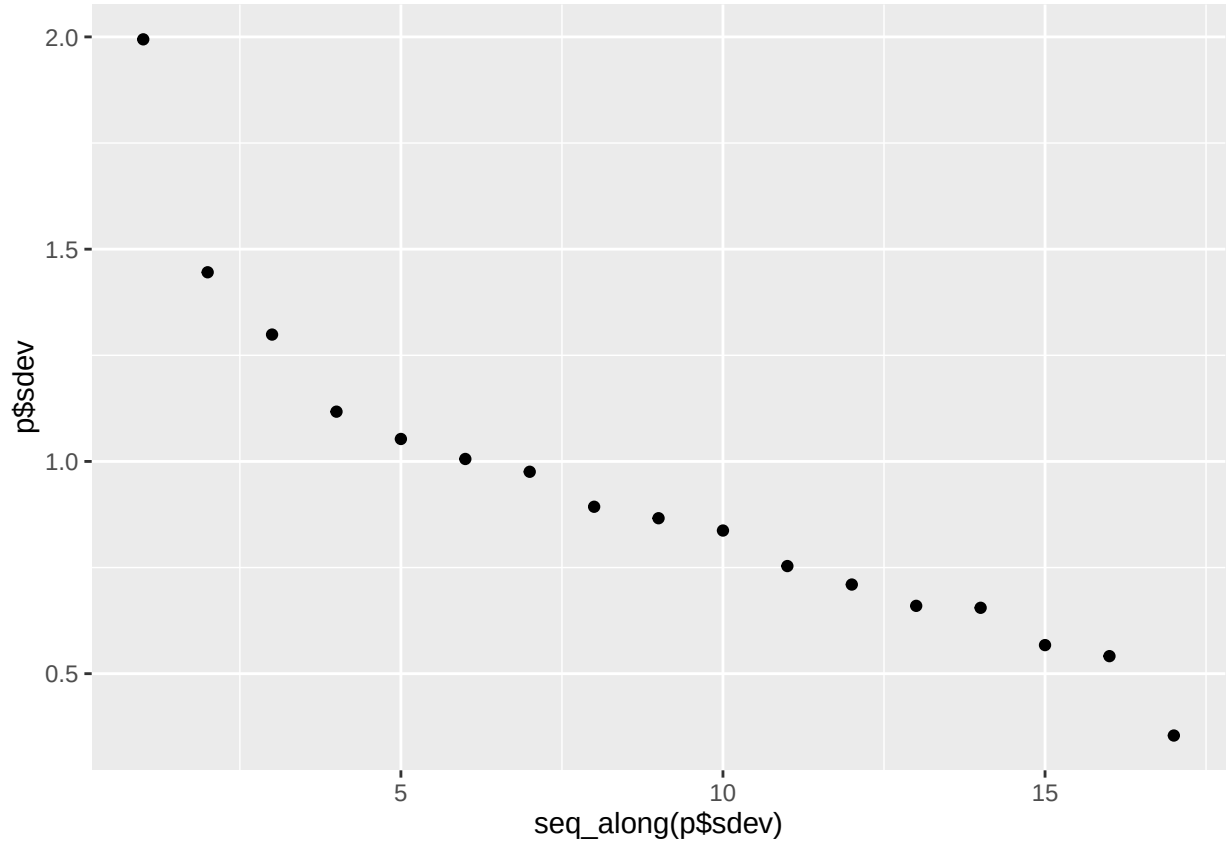
Interpreting PCA on the activities

Run PCA

```
features = select(df, sports:yoga)
p = prcomp(features, scale = TRUE)
```

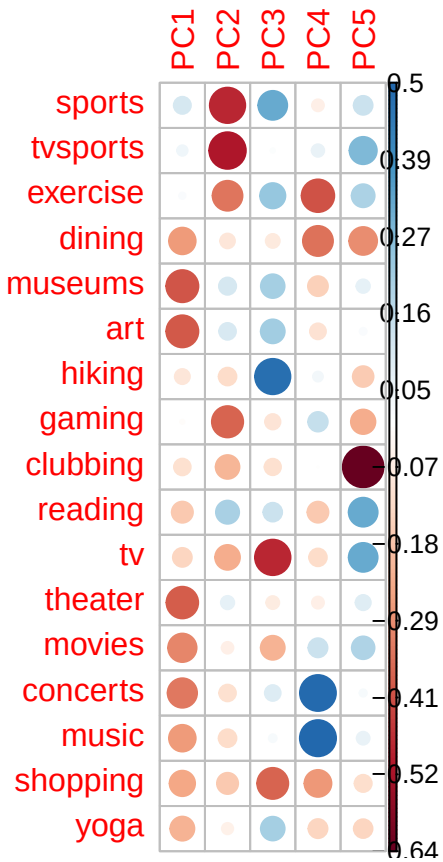
Plot the eigenvalues

```
qplot(,p$sdev)
```

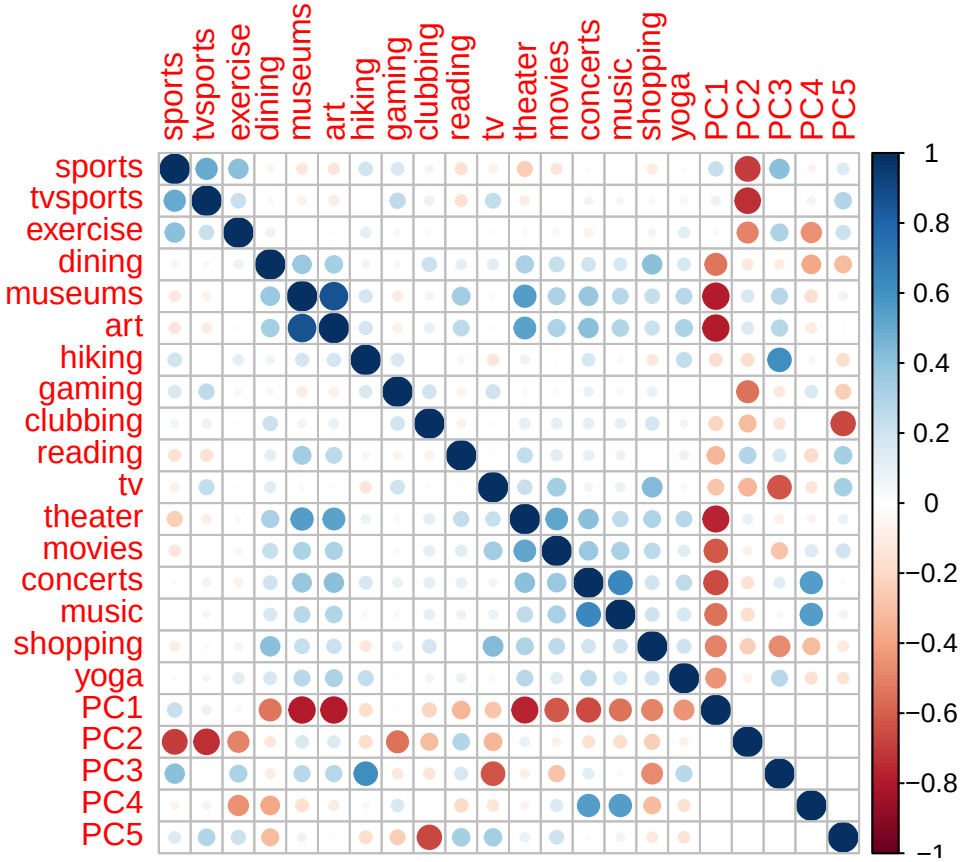


Leveling after three PCs.

```
# Plot the PCA loadings for first 5-10 PCAs.
loadings = p$rotation[,1:5]
corrplot(loadings, is.corr = FALSE)
```




```
fdf = cbind(features, p$x[,1:5])
corrplot(cor(fdf))
```



PC1: Going out/cultured (dining, museums, art, theater, movies, concerts) PC2: Sports/exercise

After PC2, the PCs appear less coherent.

Principal component regression

Predict gender, race (restrict to whites + Asians), and career code using PCAs with logistic regression

```
# Create four datasets
# gender + activities
fdf = cbind(df, p$x)
gdf = select(fdf, gender, contains("PC"))

# white/asians + activities
rdf = fdf %>% select(race, contains("PC")) %>% filter(race == 2 | race == 4)
rdf = dummy.data.frame(rdf, names=colnames(rdf)[1], sep="_")
rdf = rdf[,-1] # Just keep one dummy column.

# academia/biz+finance + activites
cdf = fdf %>% select(career_c, contains("PC")) %>% filter(career_c == 2 | career_c == 7)
cdf = dummy.data.frame(cdf, names=colnames(cdf)[1], sep="_")
cdf = cdf[,-1] # Just keep one dummy column.

# Iterate over the PCAs
for (i in 1:ncol(p$x)) {
  for (data in list(gdf, rdf, cdf)) {
```

```
    # Subset with the PCA needed
    sub = data[,1:(1+i)]
    # call glm for logistic regression
    f = paste(colnames(sub)[1], "~.")
    #print(f)
    fit = glm(formula(f), family="binomial", sub)
    print(paste(f, i))
    coefs = as.data.frame(summary(fit)$coefficients)
    probs = predict(fit, type="response")

    print(coefs[coefs[4] < 0.05, ])
    print(roc(sub[[1]], probs)$auc)
    cat("\n")
  }
}
```

```
## [1] "gender ~. 1"
##      Estimate Std. Error  z value   Pr(>|z|)
## PC1 0.4211494 0.05331597 7.899124 2.8087e-15
## Area under the curve: 0.7125
##
## [1] "race_4 ~. 1"
##               Estimate Std. Error  z value     Pr(>|z|)
## (Intercept) -0.8097108   0.104121 -7.77663 7.448191e-15
## Area under the curve: 0.5376
##
## [1] "career_c_7 ~. 1"
##      Estimate Std. Error  z value     Pr(>|z|)
## PC1 0.2167384 0.06363757 3.405825 0.0006596449
## Area under the curve: 0.6191
##
## [1] "gender ~. 2"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4281755 0.05384044  7.952674 1.825283e-15
## PC2 -0.2153452 0.06547100 -3.289169 1.004837e-03
## Area under the curve: 0.7246
##
## [1] "race_4 ~. 2"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8189397   0.105290 -7.777942 7.371357e-15
## PC2         -0.1852381   0.072887 -2.541443 1.103960e-02
## Area under the curve: 0.5831
##
## [1] "career_c_7 ~. 2"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2315009 0.06694048  3.458309 5.435776e-04
## PC2 -0.4780427 0.09748951 -4.903530 9.412959e-07
## Area under the curve: 0.7086
##
## [1] "gender ~. 3"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4331799 0.05426988  7.981959 1.440290e-15
## PC2 -0.2190267 0.06579632 -3.328860 8.720216e-04
```

```
## PC3  0.1488950 0.07260977  2.050619 4.030403e-02
## Area under the curve: 0.7293
##
## [1] "race_4 ~. 3"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8545398 0.10924489 -7.822241 5.189132e-15
## PC2         -0.1832539 0.07427367 -2.467280 1.361438e-02
## PC3         -0.3801685 0.08731901 -4.353789 1.338046e-05
## Area under the curve: 0.6525
##
## [1] "career_c_7 ~. 3"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2302367 0.06703097  3.434781 5.930339e-04
## PC2 -0.4805019 0.09810964 -4.897602 9.701344e-07
## Area under the curve: 0.7135
##
## [1] "gender ~. 4"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4497403 0.05574886  8.067256 7.189587e-16
## PC2 -0.2339834 0.06798659 -3.441611 5.782609e-04
## PC3  0.1552005 0.07386593  2.101111 3.563123e-02
## PC4  0.4426321 0.09337901  4.740167 2.135417e-06
## Area under the curve: 0.7562
##
## [1] "race_4 ~. 4"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8546825 0.10928450 -7.820711 5.252578e-15
## PC2         -0.1832278 0.07428555 -2.466534 1.364278e-02
## PC3         -0.3801818 0.08732133 -4.353825 1.337822e-05
## Area under the curve: 0.6521
##
## [1] "career_c_7 ~. 4"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2296881 0.06726611  3.414618 6.387156e-04
## PC2 -0.4724477 0.09831743 -4.805330 1.544966e-06
## Area under the curve: 0.7166
##
## [1] "gender ~. 5"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4501558 0.05575681  8.073558 6.827892e-16
## PC2 -0.2348019 0.06797066 -3.454459 5.513971e-04
## PC3  0.1553778 0.07394397  2.101292 3.561537e-02
## PC4  0.4437747 0.09349157  4.746681 2.067817e-06
## Area under the curve: 0.7563
##
## [1] "race_4 ~. 5"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8515265 0.10937555 -7.785347 6.952236e-15
## PC2         -0.1849086 0.07457046 -2.479649 1.315116e-02
## PC3         -0.3790228 0.08739321 -4.336983 1.444520e-05
## Area under the curve: 0.6522
##
## [1] "career_c_7 ~. 5"
##       Estimate Std. Error   z value     Pr(>|z|)
```

```
## PC1  0.2403297 0.06790898  3.538998 4.016490e-04
## PC2 -0.4744307 0.09972525 -4.757378 1.961240e-06
## PC5 -0.2784366 0.12670961 -2.197438 2.798915e-02
## Area under the curve: 0.7318
##
## [1] "gender ~. 6"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4623078 0.05704100  8.104833 5.281783e-16
## PC2 -0.2367125 0.06813324 -3.474258 5.122683e-04
## PC3  0.1600227 0.07449872  2.147993 3.171435e-02
## PC4  0.4463388 0.09399251  4.748663 2.047656e-06
## PC6  0.2423567 0.09858424  2.458372 1.395687e-02
## Area under the curve: 0.7667
##
## [1] "race_4 ~. 6"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8601536 0.11015241 -7.808759 5.775392e-15
## PC2         -0.1876719 0.07462293 -2.514936 1.190540e-02
## PC3         -0.3801606 0.08782365 -4.328682 1.500047e-05
## Area under the curve: 0.6648
##
## [1] "career_c_7 ~. 6"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2404894 0.06790384  3.541616 3.976833e-04
## PC2 -0.4752231 0.10021686 -4.741947 2.116735e-06
## PC5 -0.2793458 0.12718764 -2.196328 2.806848e-02
## Area under the curve: 0.7325
##
## [1] "gender ~. 7"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4707800 0.05771857  8.156474 3.449476e-16
## PC2 -0.2481296 0.06938301 -3.576230 3.485851e-04
## PC3  0.1609148 0.07589256  2.120297 3.398097e-02
## PC4  0.4655410 0.09589791  4.854548 1.206616e-06
## PC6  0.2394303 0.10051252  2.382095 1.721447e-02
## PC7  0.4612170 0.10408467  4.431171 9.372253e-06
## Area under the curve: 0.7883
##
## [1] "race_4 ~. 7"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8639743 0.11045196 -7.822172 5.191945e-15
## PC2         -0.1862364 0.07467839 -2.493846 1.263675e-02
## PC3         -0.3792264 0.08777010 -4.320679 1.555500e-05
## Area under the curve: 0.6659
##
## [1] "career_c_7 ~. 7"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2425145 0.06801301  3.565707 3.628765e-04
## PC2 -0.4751063 0.10001528 -4.750337 2.030781e-06
## PC5 -0.2801422 0.12742314 -2.198519 2.791211e-02
## Area under the curve: 0.7367
##
## [1] "gender ~. 8"
##       Estimate Std. Error   z value     Pr(>|z|)
```

```
## PC1  0.4708969 0.05773230  8.156559 3.447050e-16
## PC2 -0.2484715 0.06946086 -3.577144 3.473689e-04
## PC3  0.1611661 0.07592501  2.122700 3.377896e-02
## PC4  0.4658726 0.09597167  4.854272 1.208298e-06
## PC6  0.2392557 0.10052215  2.380129 1.730656e-02
## PC7  0.4612944 0.10408100  4.432071 9.333240e-06
## Area under the curve: 0.7881
##
## [1] "race_4 ~. 8"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8707998 0.11097704 -7.846666 4.272418e-15
## PC2         -0.1866917 0.07475124 -2.497507 1.250700e-02
## PC3         -0.3848826 0.08830363 -4.358627 1.308809e-05
## Area under the curve: 0.6652
##
## [1] "career_c_7 ~. 8"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2417174 0.06815292  3.546691 3.901015e-04
## PC2 -0.4789041 0.10034902 -4.772385 1.820573e-06
## PC5 -0.2800818 0.12751579 -2.196448 2.805989e-02
## Area under the curve: 0.7378
##
## [1] "gender ~. 9"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4728397 0.05785560  8.172756 3.014231e-16
## PC2 -0.2467936 0.06943978 -3.554066 3.793237e-04
## PC3  0.1622221 0.07607492  2.132400 3.297400e-02
## PC4  0.4676986 0.09593592  4.875114 1.087455e-06
## PC6  0.2408431 0.10061181  2.393785 1.667550e-02
## PC7  0.4645881 0.10451592  4.445142 8.783397e-06
## Area under the curve: 0.7878
##
## [1] "race_4 ~. 9"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8916929 0.11272878 -7.910073 2.572376e-15
## PC2         -0.1879017 0.07501485 -2.504860 1.224999e-02
## PC3         -0.3939179 0.08944894 -4.403830 1.063561e-05
## PC6          0.2222031 0.10902578  2.038078 4.154211e-02
## PC9          0.3486429 0.13194998  2.642235 8.236080e-03
## Area under the curve: 0.6823
##
## [1] "career_c_7 ~. 9"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2403272 0.06797797  3.535368 4.072072e-04
## PC2 -0.4814659 0.10090759 -4.771354 1.829912e-06
## PC5 -0.2870288 0.12771610 -2.247397 2.461464e-02
## Area under the curve: 0.7417
##
## [1] "gender ~. 10"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4731087 0.05786326  8.176323 2.926376e-16
## PC2 -0.2474197 0.06947349 -3.561355 3.689464e-04
## PC3  0.1631436 0.07607901  2.144397 3.200109e-02
## PC4  0.4672584 0.09604398  4.865047 1.144298e-06
```

```
## PC6  0.2429024 0.10064342  2.413496 1.580032e-02
## PC7  0.4634835 0.10434076  4.442017 8.911932e-06
## Area under the curve: 0.7892
##
## [1] "race_4 ~. 10"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8914487 0.11277544 -7.904635 2.687200e-15
## PC2         -0.1877150 0.07497434 -2.503723 1.228942e-02
## PC3         -0.3940464 0.08943920 -4.405746 1.054206e-05
## PC6          0.2226239 0.10895221  2.043317 4.102108e-02
## PC9          0.3496148 0.13195816  2.649437 8.062611e-03
## Area under the curve: 0.6828
##
## [1] "career_c_7 ~. 10"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2407016 0.06801088  3.539163 4.013982e-04
## PC2 -0.4807562 0.10090517 -4.764436 1.893828e-06
## PC5 -0.2871992 0.12769195 -2.249157 2.450252e-02
## Area under the curve: 0.7406
##
## [1] "gender ~. 11"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.4743874 0.05799375  8.179974 2.839042e-16
## PC2 -0.2491312 0.06966574 -3.576094 3.487660e-04
## PC3  0.1655475 0.07625279  2.171035 2.992852e-02
## PC4  0.4687227 0.09635572  4.864503 1.147446e-06
## PC6  0.2419742 0.10097269  2.396432 1.655556e-02
## PC7  0.4607984 0.10403413  4.429300 9.453936e-06
## Area under the curve: 0.7894
##
## [1] "race_4 ~. 11"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8957322 0.11325225 -7.909178 2.590940e-15
## PC2         -0.1902557 0.07531505 -2.526131 1.153264e-02
## PC3         -0.3987759 0.08975978 -4.442702 8.883624e-06
## PC6          0.2282099 0.10965370  2.081187 3.741675e-02
## PC9          0.3541645 0.13208047  2.681429 7.330839e-03
## Area under the curve: 0.6862
##
## [1] "career_c_7 ~. 11"
##       Estimate Std. Error   z value     Pr(>|z|)
## PC1  0.2400344 0.06809147  3.525175 4.232025e-04
## PC2 -0.4806980 0.10096233 -4.761162 1.924813e-06
## PC5 -0.2875403 0.12775415 -2.250732 2.440253e-02
## Area under the curve: 0.7409
##
## [1] "gender ~. 12"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.4981312 0.05986402  8.321045 8.719087e-17
## PC2  -0.2632816 0.07162709 -3.675726 2.371738e-04
## PC3   0.1660176 0.07787869  2.131746 3.302775e-02
## PC4   0.4839536 0.09892807  4.891975 9.982932e-07
## PC6   0.2346641 0.10255648  2.288145 2.212906e-02
## PC7   0.4838380 0.10703031  4.520571 6.167318e-06
```

```
## PC12 -0.7143413 0.15382278 -4.643924 3.418530e-06
## Area under the curve: 0.8133
##
## [1] "race_4 ~. 12"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.9007995 0.11376352 -7.918176 2.410209e-15
## PC2         -0.1926975 0.07556949 -2.549938 1.077422e-02
## PC3         -0.4003598 0.08983067 -4.456827 8.318171e-06
## PC6          0.2306040 0.10986372  2.099000 3.581686e-02
## PC9          0.3598684 0.13245080  2.716997 6.587721e-03
## Area under the curve: 0.6857
##
## [1] "career_c_7 ~. 12"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.2362921 0.06863675  3.442647 5.760505e-04
## PC2  -0.4895744 0.10249549 -4.776546 1.783318e-06
## PC5  -0.2801022 0.12831023 -2.183007 2.903528e-02
## PC12 -0.3847286 0.19093713 -2.014949 4.390995e-02
## Area under the curve: 0.7512
##
## [1] "gender ~. 13"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.5081494 0.06072300  8.368318 5.844647e-17
## PC2  -0.2688855 0.07307701 -3.679482 2.337086e-04
## PC3   0.1658321 0.07968979  2.080970 3.743663e-02
## PC4   0.5000110 0.10091589  4.954730 7.243075e-07
## PC6   0.2512644 0.10614231  2.367240 1.792129e-02
## PC7   0.5080351 0.10949231  4.639915 3.485518e-06
## PC12 -0.7470657 0.15880547 -4.704282 2.547607e-06
## PC13 -0.6987727 0.16308215 -4.284789 1.829124e-05
## Area under the curve: 0.8267
##
## [1] "race_4 ~. 13"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.9080996 0.11449513 -7.931338 2.167970e-15
## PC2         -0.1943239 0.07544108 -2.575837 9.999783e-03
## PC3         -0.4005579 0.09013458 -4.443998 8.830237e-06
## PC6          0.2301520 0.11027941  2.086990 3.688904e-02
## PC9          0.3499851 0.13240148  2.643362 8.208720e-03
## Area under the curve: 0.6905
##
## [1] "career_c_7 ~. 13"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.2353629 0.06859727  3.431082 6.011787e-04
## PC2  -0.4927233 0.10294017 -4.786502 1.697135e-06
## PC5  -0.2788296 0.12837707 -2.171958 2.985882e-02
## PC12 -0.3872656 0.19099605 -2.027610 4.260004e-02
## Area under the curve: 0.7527
##
## [1] "gender ~. 14"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.5078381 0.06075211  8.359185 6.315300e-17
## PC2  -0.2697119 0.07320902 -3.684134 2.294812e-04
## PC3   0.1661838 0.08005986  2.075745 3.791759e-02
```

```
## PC4   0.4978413 0.10101233  4.928520 8.285468e-07
## PC6   0.2483208 0.10643475  2.333081 1.964391e-02
## PC7   0.5078207 0.10969626  4.629335 3.668427e-06
## PC12 -0.7482295 0.15914291 -4.701620 2.581054e-06
## PC13 -0.6953437 0.16328160 -4.258555 2.057523e-05
## Area under the curve: 0.8281
##
## [1] "race_4 ~. 14"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.9089796 0.11462767 -7.929845 2.194201e-15
## PC2         -0.1954161 0.07585392 -2.576217 9.988800e-03
## PC3         -0.4008424 0.09026459 -4.440749 8.964616e-06
## PC6          0.2247823 0.11029359  2.038036 4.154637e-02
## PC9          0.3427399 0.13249752  2.586765 9.688164e-03
## Area under the curve: 0.6914
##
## [1] "career_c_7 ~. 14"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.2402706 0.06910903  3.476689 5.076464e-04
## PC2  -0.4891932 0.10287950 -4.755011 1.984353e-06
## PC5  -0.2927730 0.12939907 -2.262559 2.366291e-02
## PC12 -0.3889792 0.19262822 -2.019326 4.345331e-02
## Area under the curve: 0.7562
##
## [1] "gender ~. 15"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.5120785 0.06109352  8.381878 5.208967e-17
## PC2  -0.2707819 0.07320891 -3.698756 2.166585e-04
## PC3   0.1647937 0.08005265  2.058566 3.953582e-02
## PC4   0.5039782 0.10152589  4.964036 6.904303e-07
## PC6   0.2512942 0.10678542  2.353263 1.860948e-02
## PC7   0.5115415 0.10996165  4.651999 3.287326e-06
## PC12 -0.7543902 0.15937176 -4.733525 2.206538e-06
## PC13 -0.7007587 0.16313437 -4.295592 1.742278e-05
## Area under the curve: 0.8286
##
## [1] "race_4 ~. 15"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.9084524 0.11472227 -7.918710 2.399877e-15
## PC2         -0.1956021 0.07574397 -2.582411 9.811253e-03
## PC3         -0.4016339 0.09042057 -4.441841 8.919227e-06
## PC6          0.2281658 0.11054539  2.064001 3.901763e-02
## PC9          0.3465731 0.13281072  2.609527 9.066759e-03
## Area under the curve: 0.6905
##
## [1] "career_c_7 ~. 15"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.2358098 0.06950171  3.392863 6.916623e-04
## PC2  -0.4909540 0.10316874 -4.758748 1.947972e-06
## PC5  -0.2953478 0.12990828 -2.273511 2.299543e-02
## PC12 -0.4088411 0.19454317 -2.101544 3.559323e-02
## Area under the curve: 0.7584
##
## [1] "gender ~. 16"
```

```
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.5121195 0.06112479  8.378262 5.371479e-17
## PC2  -0.2708371 0.07325528 -3.697168 2.180178e-04
## PC3   0.1648040 0.08005546  2.058623 3.953038e-02
## PC4   0.5039481 0.10153312  4.963386 6.927475e-07
## PC6   0.2512527 0.10680131  2.352525 1.864646e-02
## PC7   0.5115620 0.10997059  4.651808 3.290377e-06
## PC12 -0.7544468 0.15939744 -4.733118 2.210973e-06
## PC13 -0.7006401 0.16321640 -4.292706 1.765085e-05
## Area under the curve: 0.8286
##
## [1] "race_4 ~. 16"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.9083754 0.11498583 -7.899890 2.791501e-15
## PC2         -0.1956176 0.07576090 -2.582039 9.821861e-03
## PC3         -0.4016321 0.09042288 -4.441709 8.924719e-06
## PC6          0.2281514 0.11055480  2.063695 3.904664e-02
## PC9          0.3465011 0.13300948  2.605086 9.185122e-03
## Area under the curve: 0.6907
##
## [1] "career_c_7 ~. 16"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.2337180 0.06955611  3.360136 7.790406e-04
## PC2  -0.4900112 0.10307862 -4.753762 1.996659e-06
## PC5  -0.2947196 0.12990484 -2.268735 2.328447e-02
## PC12 -0.4186217 0.19478306 -2.149169 3.162100e-02
## Area under the curve: 0.7608
##
## [1] "gender ~. 17"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.5123289 0.06115247  8.377894 5.388330e-17
## PC2  -0.2714752 0.07335615 -3.700783 2.149351e-04
## PC3   0.1649980 0.08008389  2.060314 3.936853e-02
## PC4   0.5045518 0.10148373  4.971750 6.635117e-07
## PC6   0.2497459 0.10693177  2.335563 1.951404e-02
## PC7   0.5136259 0.11034579  4.654694 3.244621e-06
## PC12 -0.7530306 0.15948521 -4.721633 2.339590e-06
## PC13 -0.7049875 0.16352138 -4.311286 1.623077e-05
## Area under the curve: 0.8296
##
## [1] "race_4 ~. 17"
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.9162701 0.11567714 -7.920927 2.357471e-15
## PC2         -0.1973424 0.07585923 -2.601429 9.283618e-03
## PC3         -0.4053677 0.09080043 -4.464381 8.030048e-06
## PC6          0.2312780 0.11080161  2.087316 3.685958e-02
## PC9          0.3534674 0.13350800  2.647537 8.108043e-03
## Area under the curve: 0.6913
##
## [1] "career_c_7 ~. 17"
##        Estimate Std. Error   z value     Pr(>|z|)
## PC1   0.2355287 0.06986843  3.371032 7.488703e-04
## PC2  -0.4847743 0.10347561 -4.684914 2.800775e-06
## PC5  -0.2911272 0.12993674 -2.240531 2.505650e-02
```

```
## PC12 -0.4230728 0.19492670 -2.170420 2.997504e-02
## Area under the curve: 0.7594
```

Comparison with stepwise regression

Run stepwise regression on activities for same predictions:

```
gdf = select(df, gender, sports:yoga)
rdf = df %>% select(race, sports:yoga) %>% filter(race == 2 | race == 4)
rdf = dummy.data.frame(rdf, names=colnames(rdf)[1], sep="_")
rdf = rdf[,-1] # Just keep one dummy column.

cdf = df %>% select(career_c, sports:yoga) %>% filter(career_c == 2 | career_c == 7)
cdf = dummy.data.frame(cdf, names=colnames(cdf)[1], sep="_")
cdf = cdf[,-1] # Just keep one dummy column.

dataList = list(gdf, rdf, cdf)

for (data in dataList) {
  resp = colnames(data)[1]

  # Perform stepwise logistic regression.
  print(resp)
  f = formula(paste(resp, "~."))
  model_init = lm(f, data)
  fit_step = step(model_init, f, direction="backward", trace=0)
  probs = predict(fit_step, type = "response")

  coefs = as.data.frame(summary(fit_step)$coefficients)
  print(coefs[coefs[4] < 0.05, ])
  print(roc(data[[1]], probs)$auc)

  print("Regularization:")

  # Perform regularization
  param_grid = expand.grid(.alpha=1:10*0.1, .lambda=10^seq(-4, -1, length.out=10))
  control = trainControl(method = "repeatedcv", number = 5, repeats = 1,
                         verboseIter = FALSE, search = "grid", classProbs = TRUE, summaryFunction = twoC

  activities = select(data, sports:yoga)

  # Might want to add twoClassSummary and probs
  target_factor = factor(data[[1]])
  levels(target_factor) = c("a","b") # Get because of default class levels not being valid R variable n
  caret_fit = train(x=scale(activities), y=target_factor, method="glmnet", tuneGrid=param_grid, trContr
  alpha = caret_fit$best[1]
  lambda = caret_fit$best[2]

  fitted = glmnet(as.matrix(activities), target_factor, family = "binomial", alpha=alpha, lambda=lambda
  probs = predict(fitted, as.matrix(activities), type="response", s=lambda)
  print(roc(data[[1]], as.numeric(probs))$auc)

  cat("\n")
```

```
}
```

```
## [1] "gender"
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  0.92545371 0.116174870  7.966040 1.014025e-14
## sports       0.04083800 0.008161581  5.003687 7.672663e-07
## exercise    -0.01790261 0.008324372 -2.150626 3.195899e-02
## hiking      -0.03056625 0.007815855 -3.910800 1.039929e-04
## gaming       0.05113418 0.007561561  6.762384 3.608013e-11
## tv          -0.01803797 0.008601797 -2.097000 3.646862e-02
## theater     -0.03778865 0.010184286 -3.710486 2.287995e-04
## shopping    -0.04844536 0.008297282 -5.838702 9.196959e-09
## yoga        -0.01637318 0.007240896 -2.261210 2.415236e-02
## Area under the curve: 0.8272
## [1] "Regularization:"
## Area under the curve: 0.8282
##
## [1] "race_4"
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  0.23289915 0.103039566  2.260289 0.0243039121
## tvsports     0.01828849 0.008145745  2.245158 0.0252666209
## exercise    -0.02751776 0.009057563 -3.038097 0.0025261924
## tv           0.01937888 0.009725292  1.992627 0.0469355253
## shopping     0.03419049 0.009447190  3.619118 0.0003308633
## Area under the curve: 0.6789
## [1] "Regularization:"
## Area under the curve: 0.6538
##
## [1] "career_c_7"
##                Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  0.40686710 0.16751969  2.428772 0.015762100
## sports       0.03103294 0.01117173  2.777810 0.005832014
## museums     -0.04463570 0.01541458 -2.895681 0.004073134
## gaming       0.03093636 0.01115271  2.773887 0.005900914
## clubbing     0.02714028 0.01177241  2.305415 0.021854038
## concerts    -0.02793743 0.01364207 -2.047888 0.041477698
## Area under the curve: 0.7493
## [1] "Regularization:"
## Area under the curve: 0.7488
```

Gender:

- PCA: 17 variables, AUC: 0.8252
- Backward Stepwise: 8 variables, AUC: 0.824 (When PCA has 8 variables, AUC: 0.78)
- Regularization: 0.821

Race:

- PCA: 17 variables, AUC: 0.6905
- Backward Stepwise: 4 variables, AUC: 0.682 (When PC has 4 variables, AUC: 0.6541)
- Regularization: 0.682

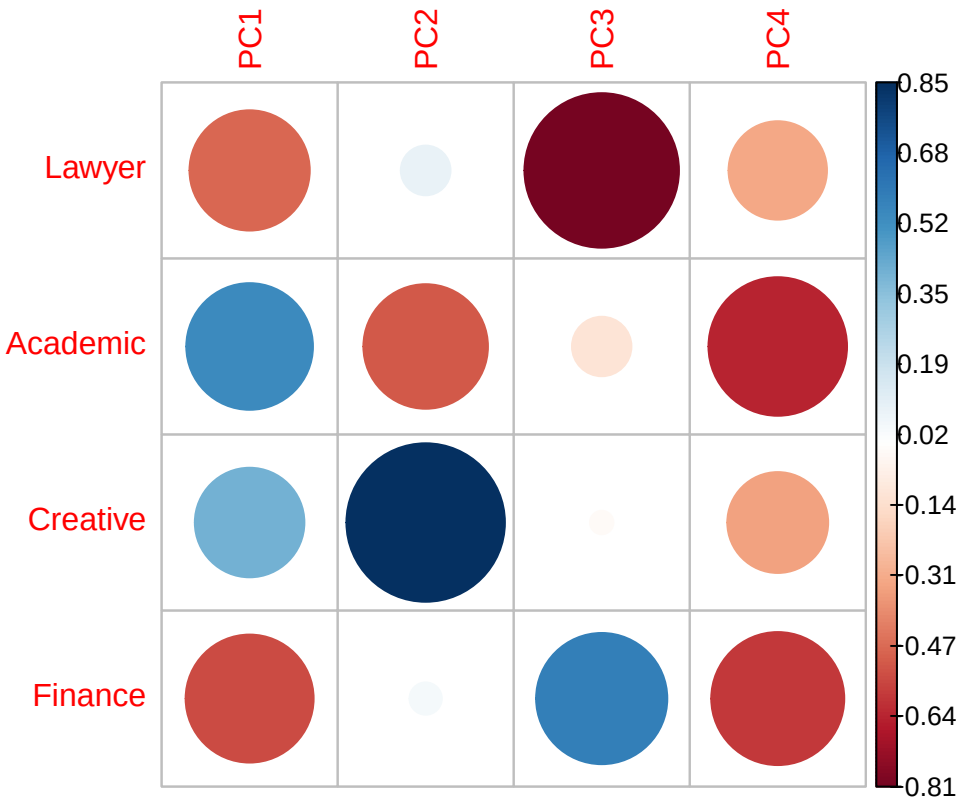
Career:

- PCA: 17 variables, AUC: 0.7594
- Backward Stepwise: 5 variables, AUC: 0.749 (When PC has 4 variables, AUC: 0.7164)
- Regularization: AUC 0.759

Multinomial logistic regression

Next, we'll use multinomial logistic regression.

```
ctable = table(df$career_c)
top4 = as.numeric(names(sort(ctable, decreasing=TRUE))[1:4])
dfc = filter(df, career_c %in% top4)
features = as.matrix(select(dfc, -career_c))
fit_career = glmnet(features, dfc$career_c, family="multinomial")
preds_career = predict(fit_career, features, s=0)
pca_preds = prcomp(scale(as.data.frame(preds_career)))
rownames(pca_preds$rotation) = c("Lawyer", "Academic", "Creative", "Finance")
corrplot(pca_preds$rotation, is.corr=FALSE)
```



The first PC is "business vs. non-business", and the second PC represents a dichotomy between academics and artists. These principal components represent the dimensions of "career variation" among the members of the dataset.