Duncan McKinnon

West

W 11

Load Packages and Data

1).

```
# load tidyverse
suppressPackageStartupMessages({
  library(tidyverse)
})

# load skulls data
skulls <- read_csv("../Data/skulls.csv")
glimpse(skulls)
```

```
## Observations: 353
## Variables: 12
## $ GOL <dbl> 189, 182, 191, 191, 178, 194, 186, 186, 186, 189, 186, 182...
## $ NOL <dbl> 185, 178, 187, 188, 177, 191, 183, 184, 182, 189, 183, 182...
## $ BNL <dbl> 100, 102, 102, 100, 97, 106, 95, 103, 96, 101, 107, 99, 10...
## $ BBH <dbl> 135, 139, 123, 127, 128, 132, 122, 130, 134, 131, 138, 124...
## $ XCB <dbl> 143, 145, 140, 141, 138, 139, 143, 141, 141, 144, 143, 136...
## $ XFB <dbl> 120, 120, 114, 123, 117, 118, 122, 121, 123, 115, 116, 113...
## $ ZYB <dbl> 133, 137, 134, 135, 129, 136, 128, 128, 135, 130, 136, 125...
## $ AUB <dbl> 119, 125, 125, 127, 121, 128, 119, 121, 123, 125, 129, 120...
## $ WCB <dbl> 70, 66, 74, 71, 69, 76, 71, 72, 71, 63, 71, 69, 71, 70, 72...
## $ ASB <dbl> 112, 113, 112, 113, 111, 112, 108, 109, 111, 108, 113, 111...
## $ BPL <dbl> 96, 108, 102, 95, 90, 102, 95, 100, 87, 95, 96, 98, 96, 98...
## $ NPH <dbl> 66, 64, 67, 76, 67, 69, 66, 64, 72, 63, 70, 69, 70, 73, 67...
```

Data Preprocessing

2).

```
# remove na's
skulls <- skulls %>% na.omit()
dim(skulls)
```

```
## [1] 353  12
```

PCA

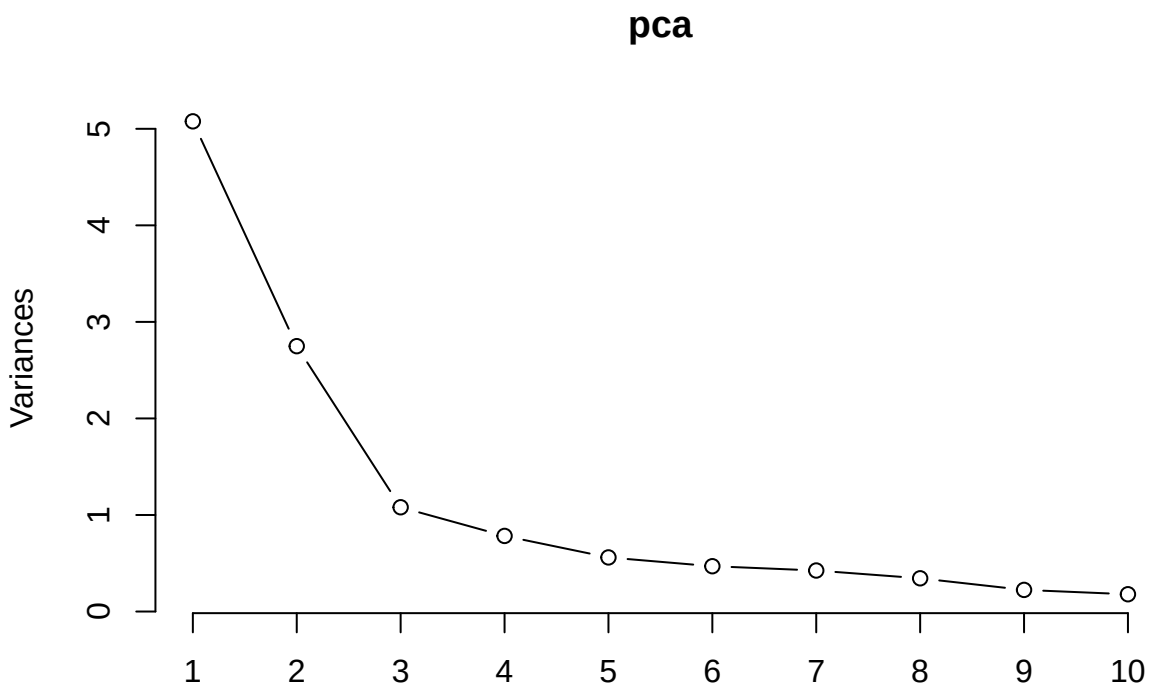
3).

```r
# run PCA with scaling
pca <- skulls %>% prcomp(., scale=T)
pca$rotation
```

```
##            PC1         PC2         PC3         PC4         PC5         PC6
## GOL -0.3184722  0.28069217  0.40153985 -0.23049483  0.23526059 -0.01940720
## NOL -0.3163243  0.26000589  0.41131042 -0.29195762  0.25452852  0.03320606
## BNL -0.3260898  0.29140992 -0.20268557 -0.02196169 -0.15297640 -0.09876013
## BBH -0.2652250  0.05251298 -0.53955698 -0.41299635  0.05585011 -0.58128200
## XCB -0.1749646 -0.48675086  0.10699235 -0.19859748 -0.20632311  0.02097648
## XFB -0.1838477 -0.46293605 -0.02715115 -0.26863794 -0.02022512  0.10596361
## ZYB -0.3691434  0.02108358 -0.17444490  0.36882471 -0.14686096 -0.00531120
## AUB -0.3656919 -0.21629736 -0.02135332  0.26951997 -0.14194340  0.01876590
## WCB -0.2728672 -0.19083849 -0.01466387  0.52103943  0.64469211 -0.15277331
## ASB -0.2795295 -0.21294150  0.43401979  0.03909103 -0.39064591 -0.22756158
## BPL -0.1771530  0.43077731 -0.01814713  0.24061123 -0.43371534  0.11727150
## NPH -0.3216349 -0.01156482 -0.32464942 -0.19849592  0.11095577  0.74158827
##             PC7         PC8         PC9        PC10        PC11
## GOL -0.043337412  0.16408067  0.08741396  0.05416082  0.03761199
## NOL -0.009345896  0.17900139 -0.00275326 -0.03434403  0.01144229
## BNL  0.213775817 -0.13029232 -0.76204263 -0.10057001 -0.27874555
## BBH -0.076819605 -0.08122933  0.29890973 -0.04741106  0.13872960
## XCB  0.220703962  0.26452761  0.14066836 -0.59826552 -0.37256145
## XFB  0.481705149  0.11270915 -0.12338528  0.61274540  0.17867539
## ZYB -0.304215558  0.42220289  0.18873547  0.38974385 -0.46289240
## AUB -0.288002261  0.25951454 -0.24499773 -0.23259820  0.67930860
## WCB  0.305297083 -0.24928917  0.08635544 -0.11044749 -0.05787170
## ASB -0.219745186 -0.63329042  0.07990751  0.14345614 -0.06862735
## BPL  0.557345410 -0.02024811  0.39926124 -0.07419880  0.21464893
## NPH -0.195096161 -0.35435665  0.13559770 -0.07655349 -0.03033051
##             PC12
## GOL  0.7133801122
## NOL -0.6917991571
## BNL  0.0531794346
## BBH -0.0293630489
## XCB  0.0580265911
## XFB  0.0006312086
## ZYB -0.0513464630
## AUB  0.0029528370
## WCB -0.0101489235
## ASB -0.0347866083
## BPL -0.0304958292
## NPH  0.0235952176
```

4).

```r
# Plot the screeplot for the PCA data
screeplot(pca, type='l')
```



5).

From the scree-plot of the principal components of the skull dataset, it appears that most of the variance in the data is captured by the first 2 principal components. In the plot there is a clear inflection point at the variance of the 3rd component, after which it appears that each subsequent component contains a similar proportion of the variance from the original dataset. Because the last 10 components seem to each contain a similar proportion of the total variance in the data, there would be little reason to prefer any one of them or expect to significantly gain from including more principal components.

6).

```
# summary of PCA object
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.2533 1.6579 1.0392 0.88486 0.74866 0.68514
## Proportion of Variance 0.4231 0.2291 0.0900 0.06525 0.04671 0.03912
## Cumulative Proportion  0.4231 0.6522 0.7421 0.80739 0.85410 0.89322
##                           PC7     PC8     PC9    PC10    PC11   PC12
## Standard deviation     0.65182 0.58680 0.47327 0.42322 0.30184 0.1340
## Proportion of Variance 0.03541 0.02869 0.01866 0.01493 0.00759 0.0015
## Cumulative Proportion  0.92863 0.95732 0.97599 0.99091 0.99850 1.0000
```

7).

If a researcher wanted to use principal components that account for 95% of the variance in the original data they would want to select the first 8 principal components because these together contain just over 95% of the variance observed in the original dataset. We can see this by looking at the cumulative proportion section of the pca summary, which shows the variance captured by selecting principal components 1:i where 1 <= i <= n.

8).

From the rotation matrix we can see that the 2nd principal component gives most significant weight to the max crainial width (XCB) and the max frontal breadth (XFB). Since the weights are similar magnitude in the same direction, we can assume that there is a significant covariation in these predictors, and the 2nd principal component is capturing that common variance.