Homework 3

Elena Volpi

#Problem 1 Monthly temperature data for 20 different weather stations within 100 miles of Corvallis were obtained for the period 1950to2009. From this data, decade averages were computed for each station and are given in the TempData.csv file available on Canvas. Let μ_1,μ_2,μ_3,μ_4,μ_5,μ_6 denote the average temperature for decades 1950s, 1960s, 1970s, 1980s, 1990s, 2000s respectively.

```
temp <- read.csv("~/Desktop/Multivariate/TempData.csv")
colnames(temp) <- c('1950s','1960s','1970s','1980s','1990s','2000s')
```

a)Test the null hypothesis H_0:μ_1 =μ_2=μ_3=μ_4=μ_5=μ_6 vs. H_A : not all μ_j are equalat level $\alpha = 0.05$ using Hotelling's T2 test. Explain how you performed this test. Based on the result of this hypothesis test, would you conclude that the average temperature around Corvallis has stayed constant over the past 60 years?

```
contrast_matrix <- as.matrix(cbind(rep(1,5),diag(-1,5,5)))  #from lecture 9, repeated measures
T2.test(t(contrast_matrix%*%t(temp)))
```

```
##
##  One-sample Hotelling test
##
## data:  t(contrast_matrix %*% t(temp))
## T2 = 394.273, F = 62.254, df1 = 5, df2 = 15, p-value = 1.701e-09
## alternative hypothesis: true mean vector is not equal to (0, 0, 0, 0, 0)'
##
## sample estimates:
##                  [,1]    [,2]     [,3]    [,4]    [,5]
## mean x-vector -0.1606 -0.0237 -0.30205 -0.7301 -0.6103
```

I performed the Hotelling's T^2 test using a constant matrix C,

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

and $Y = (CX^T)^T, where X is temperature$. At an $\alpha = 0.5$ significance level with a p-value of $1.701x10^{-9}$, we have strong evidence to reject the null hypothesis that the average difference between temperature in 1950's and following decades is equal. We have evidence to support that at least one of the average differences of temperature between the 1950's and a later decade was not zero, meaning that the temperature in Corvallis has not been the same on average over time.

b) Construct simultaneous 95% Bonferroni confidence intervals for $\mu_2 - \mu_1, \mu_3 - \mu_1, \mu_4 - \mu_1, \mu_5 - \mu_1$ and $\mu_6 - \mu_1$. Do any of these confidence intervals include 0? What would you conclude based on these confidence intervals?

```
difference_mu <- data.frame(temp[,2:6]-temp[,1])

mu_dif_means <- colMeans(difference_mu)
mu_dif_sd <- apply(difference_mu,2,sd)
n <- nrow(temp)
p <- ncol(temp)
alpha <- 0.05

for(i in 1:5){
   print(mu_dif_means[i]+ c(-1,1)*qt(1-(alpha/10),n-1)*(mu_dif_sd[i]/sqrt(n)))
}
```

```
## [1] 0.09156047 0.22963953
## [1] -0.06986223  0.11726223
## [1] 0.1690805 0.4350195
## [1] 0.5837893 0.8764107
## [1] 0.4298272 0.7907728
```

Our 95% confidence intervals are as follows:

$$\mu_2 - \mu_1 = (0.916, 0.230)$$
$$\mu_3 - \mu_1 = (-0.070, 0.117)$$
$$\mu_4 - \mu_1 = (0.169, 0.435)$$
$$\mu_5 - \mu_1 = (0.584, 0.876)$$
$$\mu_6 - \mu_1 = (0.430, 0.791)$$

The only interval that contains zero $\mu_3 - \mu_1$, the difference between temperature in the 1970's and 1950's. The difference in average temperature between the 60's, 80's, 90's,2000's and the 50's, shows that with the exception of the 1970's, we are seeing rising temperatures in each decade.

#Problem 2

```
SkullData <- read_csv("SkullData.csv",show_col_types = FALSE)
```

a) Compute and compare the covariance matrices for each time period. Do they seem approximately similar?

```
sub_4000BC <- SkullData %>% filter(Year == -4000)
cov_4000BC <- cov(sub_4000BC[,2:5])
cov_4000BC
```

```
##           MB         BH         BL         NH
## MB 26.309195  4.1517241  0.4540230  7.2459770
## BH  4.151724 19.9724138 -0.7931034  0.3931034
## BL  0.454023 -0.7931034 34.6264368 -1.9195402
## NH  7.245977  0.3931034 -1.9195402  7.6367816
```

```
sub_3300BC <- SkullData %>% filter(Year == -3300)
cov_3300BC <- cov(sub_3300BC[,2:5])
cov_3300BC
```

```
##           MB        BH         BL        NH
## MB 23.136782  1.010345  4.7678161 1.8425287
## BH  1.010345 21.596552  3.3655172 5.6241379
## BL  4.767816  3.365517 18.8919540 0.1908046
## NH  1.842529  5.624138  0.1908046 8.7367816
```

```
sub_1850BC <- SkullData %>% filter(Year == -1850)
cov_1850BC <- cov(sub_1850BC[,2:5])
cov_1850BC
```

```
##            MB          BH         BL          NH
## MB 12.1195402  0.78620690 -0.7747126  0.89885057
## BH  0.7862069 24.78620690  3.5931034 -0.08965517
## BL -0.7747126  3.59310345 20.7229885  1.67011494
## NH  0.8988506 -0.08965517  1.6701149 12.59885057
```

```
sub_200BC <- SkullData %>% filter(Year == -200)
cov_200BC <- cov(sub_200BC[,2:5])
cov_200BC
```

```
##           MB        BH        BL       NH
## MB 15.362069 -5.534483 -2.172414 2.051724
## BH -5.534483 26.355172  8.110345 6.148276
## BL -2.172414  8.110345 21.085057 5.328736
## NH  2.051724  6.148276  5.328736 7.964368
```

```
#I'm going to assume 180 ad was a typo since that doesnt exist and do 150 AD
sub_150AD <- SkullData %>% filter(Year == 150)
cov_150AD <- cov(sub_150AD[,2:5])
cov_150AD
```

```
##            MB         BH         BL         NH
## MB 28.6264368 -0.2298851 -1.8793103 -1.9942529
## BH -0.2298851 24.7126437 11.7241379  2.1494253
## BL -1.8793103 11.7241379 25.5689655  0.3965517
## NH -1.9942529  2.1494253  0.3965517 13.8264368
```

The covariance matrices of skull sizes throughout time do not seem similiar.

b)Perform a level $\alpha = 0.05$ test of the hypothesis that population mean vectors for all of thesetime periods are the same (assume equal covariance matrices). Based on this hypothesis test, does there seem to be evidence of interbreeding (if the researchers' theory that skull size change indicates interbreeding is correct)?

```
Wilks.test(SkullData[,1:4],grouping=as.factor(SkullData$Year))
```

```
##
##  One-way MANOVA (Bartlett Chi2)
##
## data:  x
## Wilks' Lambda = 1.1752e-32, Chi2-Value = 10624, DF = 16, p-value <
## 2.2e-16
```

```
## sample estimates:
##        Year       MB       BH       BL
## -1850 -1850 134.4667 133.8000 96.03333
## -200   -200 135.5000 132.3000 94.53333
## -3300 -3300 132.3667 132.7000 99.06667
## -4000 -4000 131.3667 133.6000 99.16667
## 150     150 136.1667 130.3333 93.50000
```

Our p-value is near zero so with 95% confidence, we have strong evidence that at least one period has has difference in the mean skull size which provides evidence for inbreeding.

c) Perform separate univariate ANOVAs for each variable at level $\alpha^* = \frac{\alpha}{p}$. Are any of these univariate ANOVAs significant? If we reject $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ if any of the univariate ANOVAs are significant at level $\alpha^* = \frac{\alpha}{p}$, will the overall probability of a Type I error(the overall significance level) be controlled at level α (that is, will it be $\leq \alpha$)? Explain.

```
alpha_star <- 0.04/p
anova(lm(MB~as.factor(Year),data=SkullData))
```

```
## Analysis of Variance Table
##
## Response: MB
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## as.factor(Year)    4  502.83 125.707  5.9546 0.0001826 ***
## Residuals        145 3061.07  21.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(BH~as.factor(Year),data=SkullData))
```

```
## Analysis of Variance Table
##
## Response: BH
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(Year)    4  229.9  57.477  2.4474 0.04897 *
## Residuals        145 3405.3  23.485
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(BL~as.factor(Year),data=SkullData))
```

```
## Analysis of Variance Table
##
## Response: BL
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(Year)    4  803.3 200.823  8.3057 4.636e-06 ***
## Residuals        145 3506.0  24.179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(NH~as.factor(Year),data=SkullData))
```

```
## Analysis of Variance Table
##
## Response: NH
##                   Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Year)    4   61.2  15.300   1.507 0.2032
## Residuals        145 1472.1  10.153
```

The univariate ANOVA for Basialveolar Length of Skull and Maximal Breadth of Skull are significant at the $\alpha^* = \frac{\alpha}{4}$ level. Type I error $< \alpha^*$ with Bonferroni correction.

#Problem 3

```
PollutionData <- read_csv("PollutionData.csv",show_col_types = FALSE)
```

a)Perform a multivariate multiple regression analysis using both responses $\mathbf{Y_1} = NO_2$ and $\mathbf{Y_2} = O_3$ and predictors X_1 = Wind and X_2=Solar Radiation. Test the null hypothesis that $\beta_\mathbf{2} = 0$. What would you conclude based on this test?

```
mod_full <- lm(as.matrix(PollutionData[,3:4]) ~as.matrix(PollutionData[,1:2]))
```

```
wind_mod <- lm(as.matrix(PollutionData[,3:4]) ~as.matrix(PollutionData[,1]))
```

```
anova(mod_full,wind_mod)
```

```
## Analysis of Variance Table
##
## Model 1: as.matrix(PollutionData[, 3:4]) ~ as.matrix(PollutionData[, 1:2])
## Model 2: as.matrix(PollutionData[, 3:4]) ~ as.matrix(PollutionData[, 1])
##   Res.Df Df Gen.var.   Pillai approx F num Df den Df Pr(>F)
## 1     39      17.834
## 2     40  1   18.297 0.096851   2.0375      2     38 0.1444
```

We fail to reject the null hypothesis at an $\alpha = 0.5$ level.Our partial model without solar radiation performs as well as the full model.
b)

```
solar_mod <- lm(as.matrix(PollutionData[,3:4]) ~as.matrix(PollutionData[,2]))
anova(mod_full,solar_mod)
```

```
## Analysis of Variance Table
##
## Model 1: as.matrix(PollutionData[, 3:4]) ~ as.matrix(PollutionData[, 1:2])
## Model 2: as.matrix(PollutionData[, 3:4]) ~ as.matrix(PollutionData[, 2])
##   Res.Df Df Gen.var.  Pillai approx F num Df den Df Pr(>F)
## 1     39      17.834
## 2     40  1   17.931 0.05962   1.2046      2     38  0.311
```

We fail to reject the null hypothesis that $\beta_1 = 0$ with a p.value of 0.311. Our solar model that does not account for wind performs as well as our full model.

c)

```
intercept_mod <- lm(as.matrix(PollutionData[,3:4]) ~1)
anova(mod_full,intercept_mod)
```

```
## Analysis of Variance Table
##
## Model 1: as.matrix(PollutionData[, 3:4]) ~ as.matrix(PollutionData[, 1:2])
## Model 2: as.matrix(PollutionData[, 3:4]) ~ 1
##   Res.Df Df Gen.var.  Pillai approx F num Df den Df Pr(>F)
## 1     39      17.834
## 2     41  2   18.500 0.15921   1.6865      4     78 0.1615
```

Pillai's shows none of the predictors have a linear effect on either of the response variables, which is consistent with above.

#Problem 4

```
TrackData <- read_csv("TrackData.csv",show_col_types = FALSE)
```

a) btain the sample covariance matrix **S** and the sample correlation matrix **R** for the distances based on this data. Which of these matrices would you find more interesting/appropriate to use for a principal component analysis of this data, and why?

```
TrackData_standardized <- TrackData %>%
             mutate(`100m.s` = 100/`100m.s`,
                    `200m.s` = 200/`200m.s`,
                    `400m.s` = 400/`400m.s`,
                    `800m.m` = 800/60*`800m.m`,
                    `1500m.m` = 1500/60*`1500m.m`,
                    `5000m.m` = 5000/60*`5000m.m`,
                    `10000m.m` = 10000/60*`10000m.m`,
                    Marathon.m = 42195/60*Marathon.m)

#now all standardized and in seconds, but not sure if I should have done this or not.
```

```
S <- cov(TrackData_standardized[,3:10])
```

```
R <- cor(TrackData_standardized[,3:10])
```

Since the original data was not standardized, we should use the correlation matrix.

b) Determine the eigenvalues and eigenvectors of **S**

```
S.eigen <- eigen(S)
S.eigen
```

```
## eigen() decomposition
## $values
## [1] 4.219063e+07 1.037976e+04 2.087752e+02 1.854719e+00 1.362390e-01
```

```
## [6] 4.890646e-02 1.000796e-02 5.889759e-03
##
## $vectors
##               [,1]         [,2]          [,3]          [,4]          [,5]
## [1,] -2.361921e-05 -0.001262393  0.0007749685  0.0701807927 -3.860673e-01
## [2,] -2.604134e-05 -0.001124472  0.0018734108  0.0718326966 -3.110839e-01
## [3,] -2.800880e-05 -0.000918563  0.0010586730  0.0540374216 -2.462532e-01
## [4,]  1.054588e-04  0.002732588 -0.0043379264 -0.1904515339  8.028610e-01
## [5,]  5.195442e-04  0.013689195 -0.0207724476 -0.9747150048 -2.212517e-01
## [6,]  9.583632e-03  0.191673075 -0.9811270252  0.0236369757 -7.276536e-05
## [7,]  4.375977e-02  0.980335070  0.1921801405  0.0097379905 -2.181096e-04
## [8,]  9.989960e-01 -0.044788616  0.0010053571 -0.0001212479  1.642267e-05
##               [,6]          [,7]          [,8]
## [1,]  0.6027932870  2.133953e-01  6.611560e-01
## [2,]  0.5274271179  2.434764e-01 -7.487336e-01
## [3,]  0.2184817668 -9.416453e-01 -4.480241e-02
## [4,]  0.5572396648 -9.211464e-02  1.072017e-02
## [5,] -0.0145025543 -8.974906e-04 -1.216780e-02
## [6,] -0.0003774857  3.091152e-05 -7.009979e-04
## [7,]  0.0003077459 -6.509869e-05  2.275223e-04
## [8,] -0.0000270156 -2.262907e-06 -3.187148e-06
```

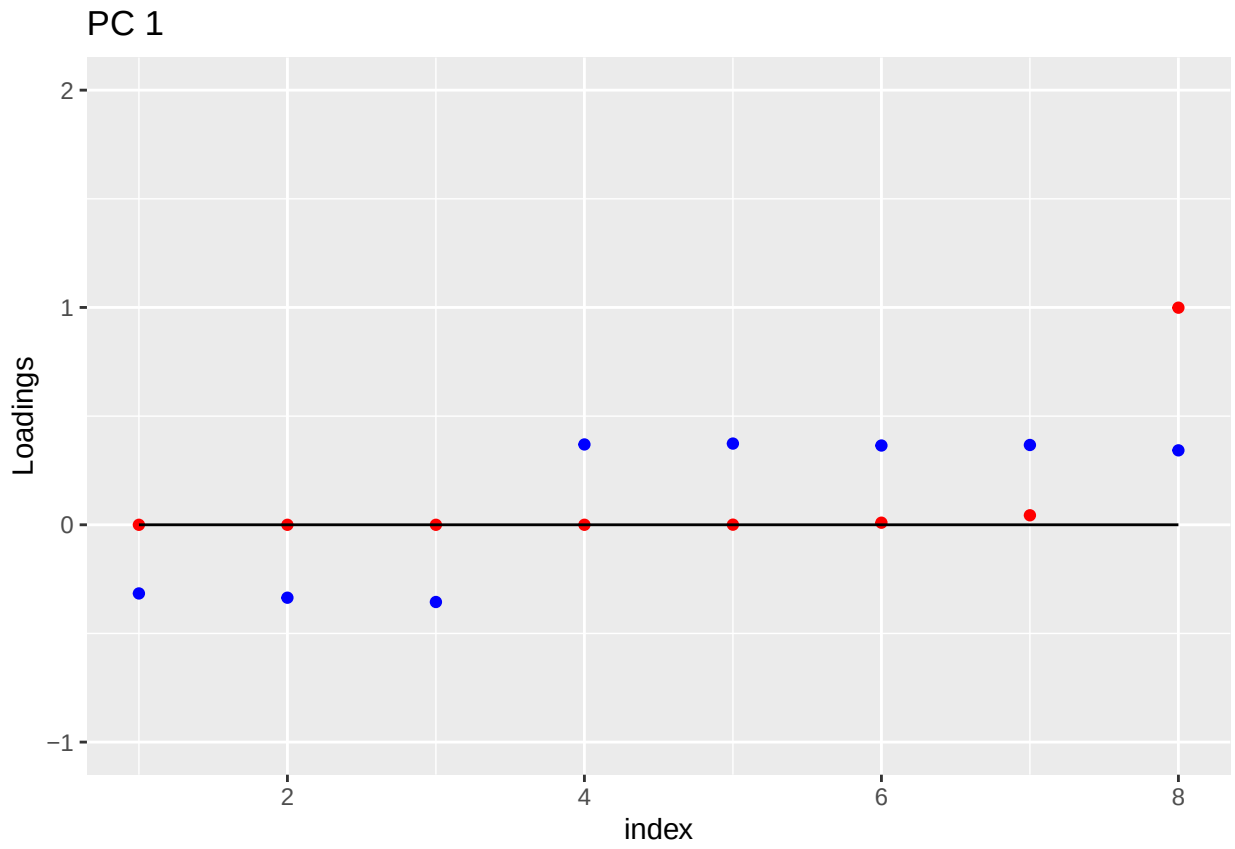
c) Determine the eigenvalues and eigenvectors of $\mathbf{R}$

```
#I looked at both the standardized and unstandardized because the plots looked funky and I wanted to co
R.eigen <- eigen(R)
R.eigen
```

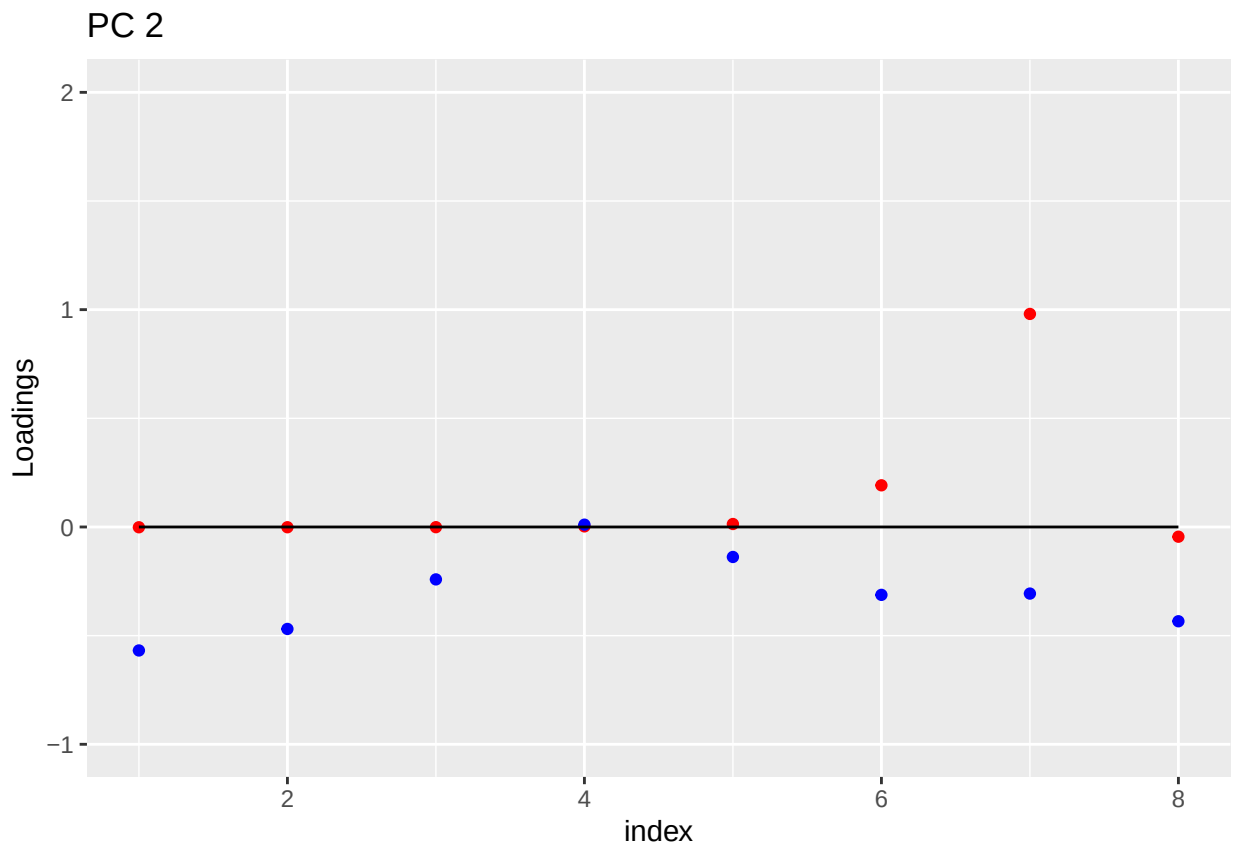
```
## eigen() decomposition
## $values
## [1] 6.59370447 0.89877566 0.16465142 0.12254351 0.08093998 0.07035242 0.04658469
## [8] 0.02244786
##
## $vectors
##            [,1]        [,2]       [,3]        [,4]        [,5]        [,6]
## [1,] -0.3159367 -0.56815830  0.3334671  0.05035536 -0.37010917  0.54597022
## [2,] -0.3355385 -0.46921305  0.3597304 -0.10213950  0.29275457 -0.64623127
## [3,] -0.3553139 -0.24119156 -0.6300813  0.62085065  0.18115169 -0.01052047
## [4,]  0.3695899  0.01066167  0.4770203  0.55281849  0.49520094  0.15190230
## [5,]  0.3736906 -0.13796124  0.1082725  0.41920009 -0.47737557 -0.18500231
## [6,]  0.3646881 -0.31238850 -0.1845862 -0.03952520 -0.22977097 -0.18089128
## [7,]  0.3671723 -0.30643246 -0.1769490 -0.08065234 -0.08828218 -0.25264829
## [8,]  0.3424845 -0.43385192 -0.2402850 -0.33492131  0.45912763  0.36102356
##             [,7]         [,8]
## [1,] -0.12787942  0.110574485
## [2,]  0.11552424 -0.103675603
## [3,] -0.01128094  0.008172889
## [4,] -0.24453732  0.045123867
## [5,]  0.61082952 -0.136544380
## [6,] -0.59052167 -0.547311124
## [7,] -0.17503811  0.794483887
## [8,]  0.39737684 -0.159759857
```

d)

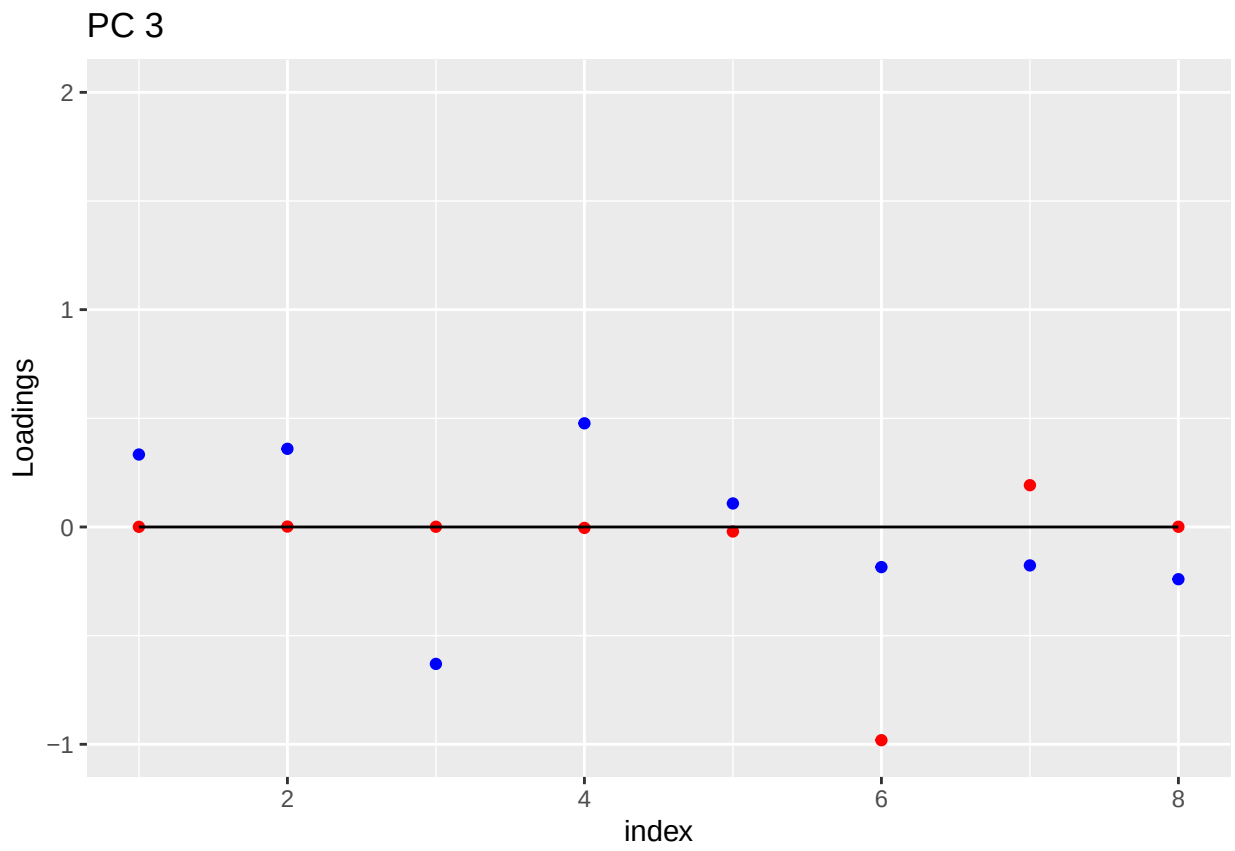
```r
#All the standardized PCS
cov_loadings <- data.frame(S.eigen$vectors)
cor_loadings <- data.frame(R.eigen$vectors)
loadings <- cbind(seq(1,length(S.eigen$values)),cov_loadings,cor_loadings)
colnames(loadings) <- c('index', 'cov1', 'cov2', 'cov3', 'cov4', 'cov5', 'cov6', 'cov7', 'cov8', 'cor1'

p1<- ggplot(loadings)+
  geom_point(aes(x=index,y=cov1), color="red")+
  geom_point(aes(x=index,y=cor1), color="blue")+
  geom_line(aes(x=index,y=0))+
  ylim(c(-1,2))+
  labs(y="Loadings",title="PC 1 ")
p1
```



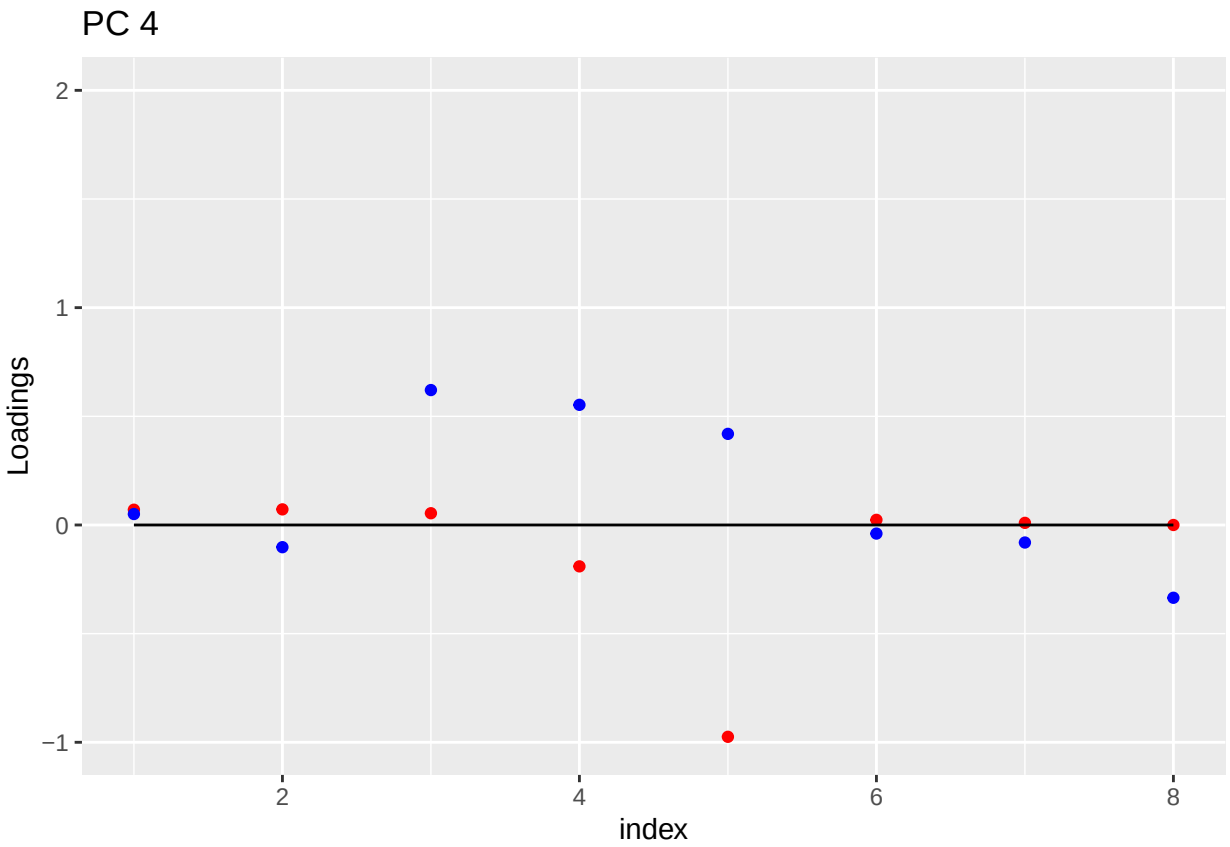

```r
p2<- ggplot(loadings)+
  geom_point(aes(x=index,y=cov2), color="red")+
  geom_point(aes(x=index,y=cor2), color="blue")+
  geom_line(aes(x=index,y=0))+
  ylim(c(-1,2))+
  labs(y="Loadings",title="PC 2")
p2
```




```
p3<- ggplot(loadings)+
  geom_point(aes(x=index,y=cov3), color="red")+
  geom_point(aes(x=index,y=cor3), color="blue")+
  geom_line(aes(x=index,y=0))+
  ylim(c(-1,2))+
  labs(y="Loadings",title="PC 3")
p3
```




```
p4<- ggplot(loadings)+
  geom_point(aes(x=index,y=cov4), color="red")+
  geom_point(aes(x=index,y=cor4), color="blue")+
  geom_line(aes(x=index,y=0))+
  ylim(c(-1,2))+
  labs(y="Loadings",title="PC 4")
p4
```



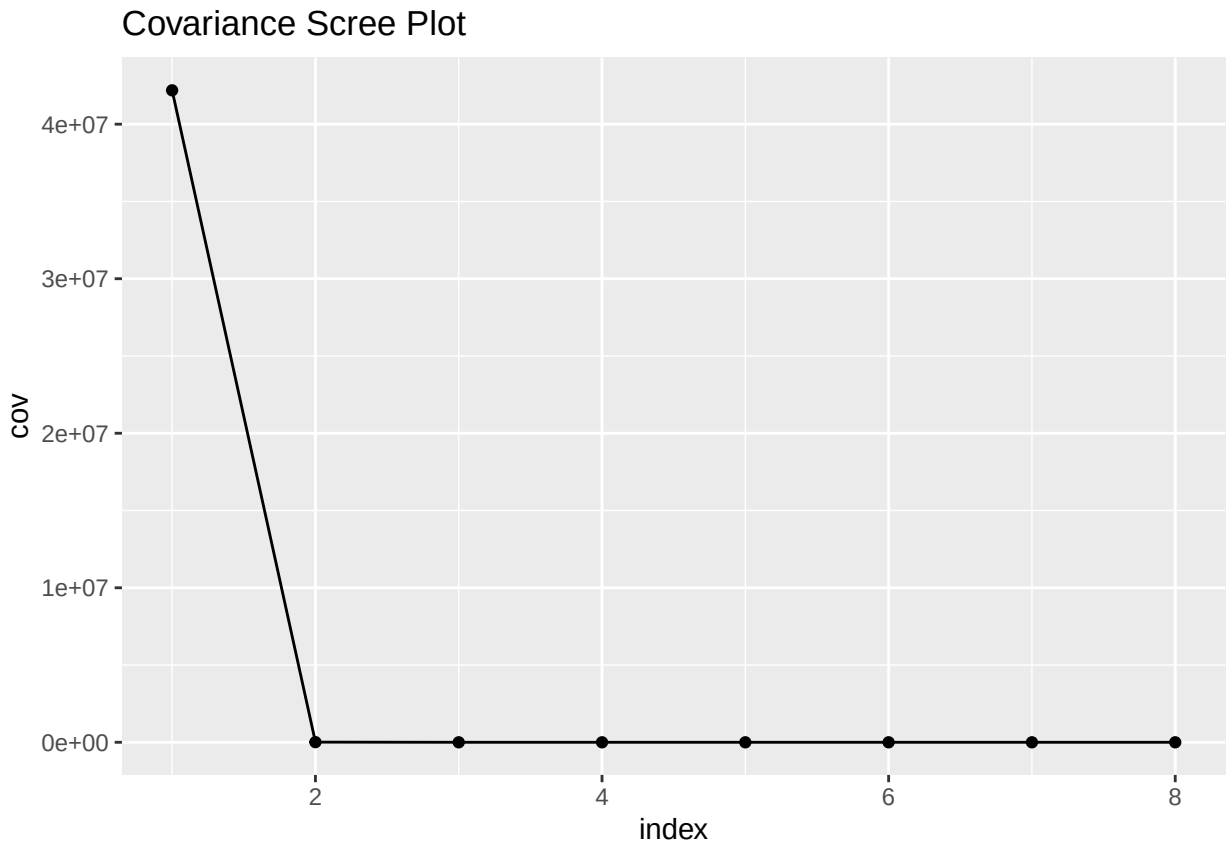
e) Principal Component 1 for **S** is show in plot 1 titled "PC 1" in red. We see that the points are fairly equal and close to zero. The first principal component represents the the performance across all race lengths. (?)

f) Principal Component 1 for **R** is show in plot 1 titled "PC 1" in blue. The weights for each race are consistent. PC1 is is giving the performance of all race lengths. (?)

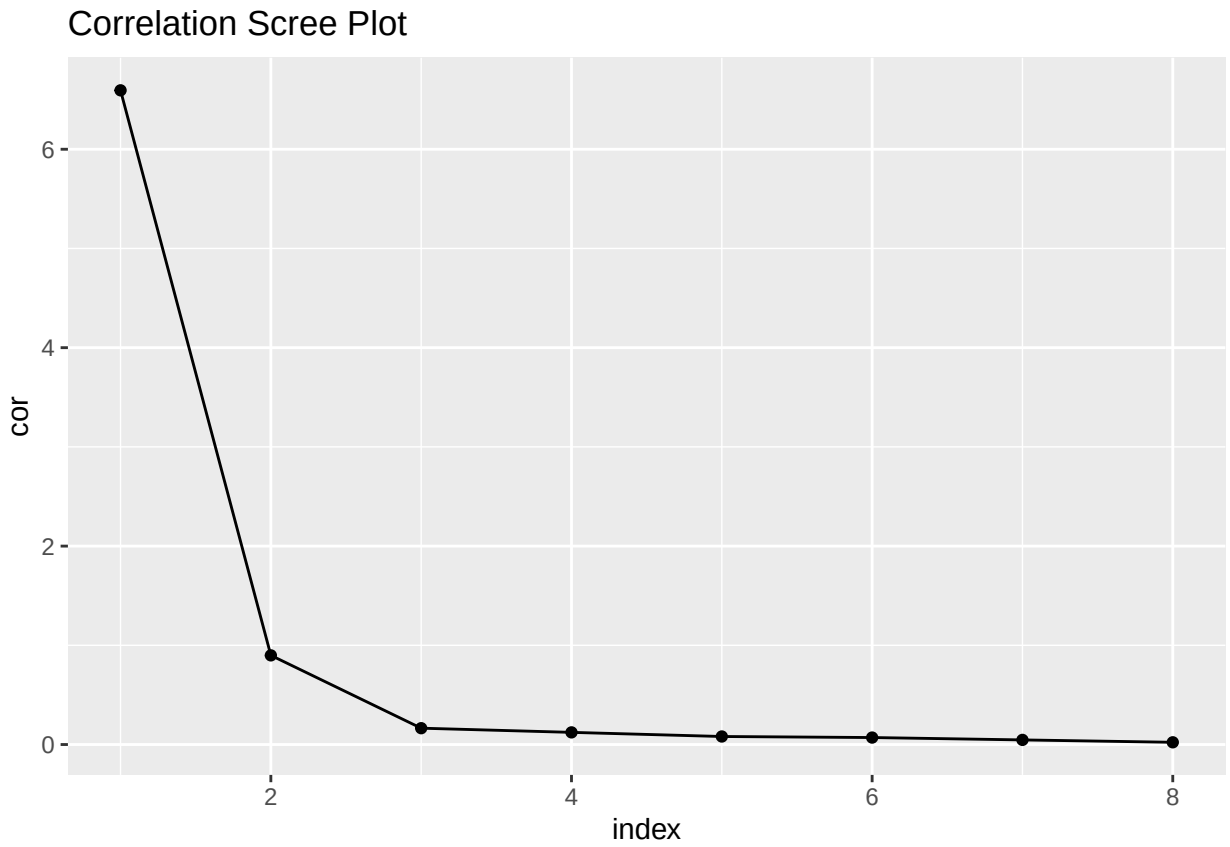
g) Principal Component 2 for **R** is show in plot 1 titled "PC 2" in blue. (?)

h)

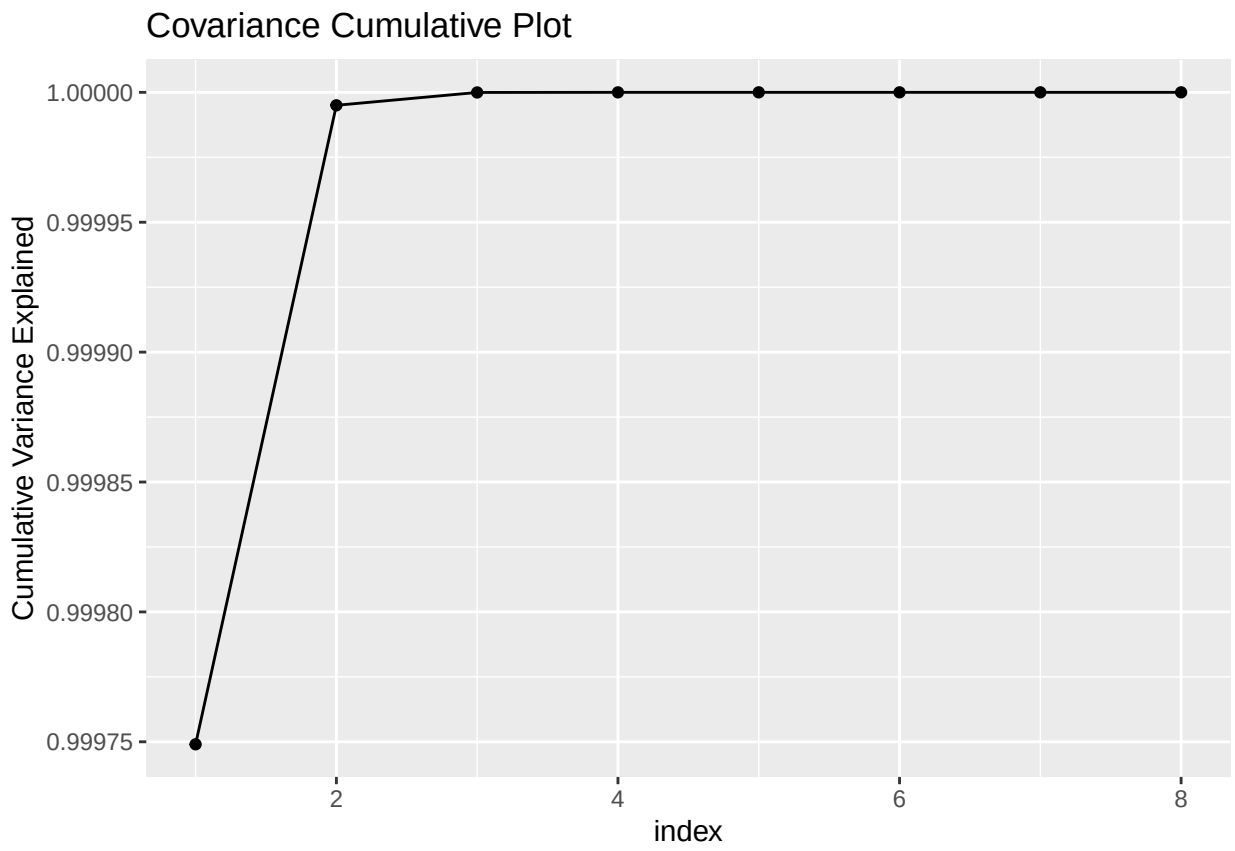
```
variance <- data.frame( cbind(seq(1,length(S.eigen$values)),S.eigen$values,R.eigen$values))
colnames(variance) <- c('index', 'cov', 'cor')

scree_cov <- ggplot(variance, aes(x=index, y=cov))+
  geom_point()+
  geom_line()+
  ggtitle("Covariance Scree Plot")
scree_cov
```



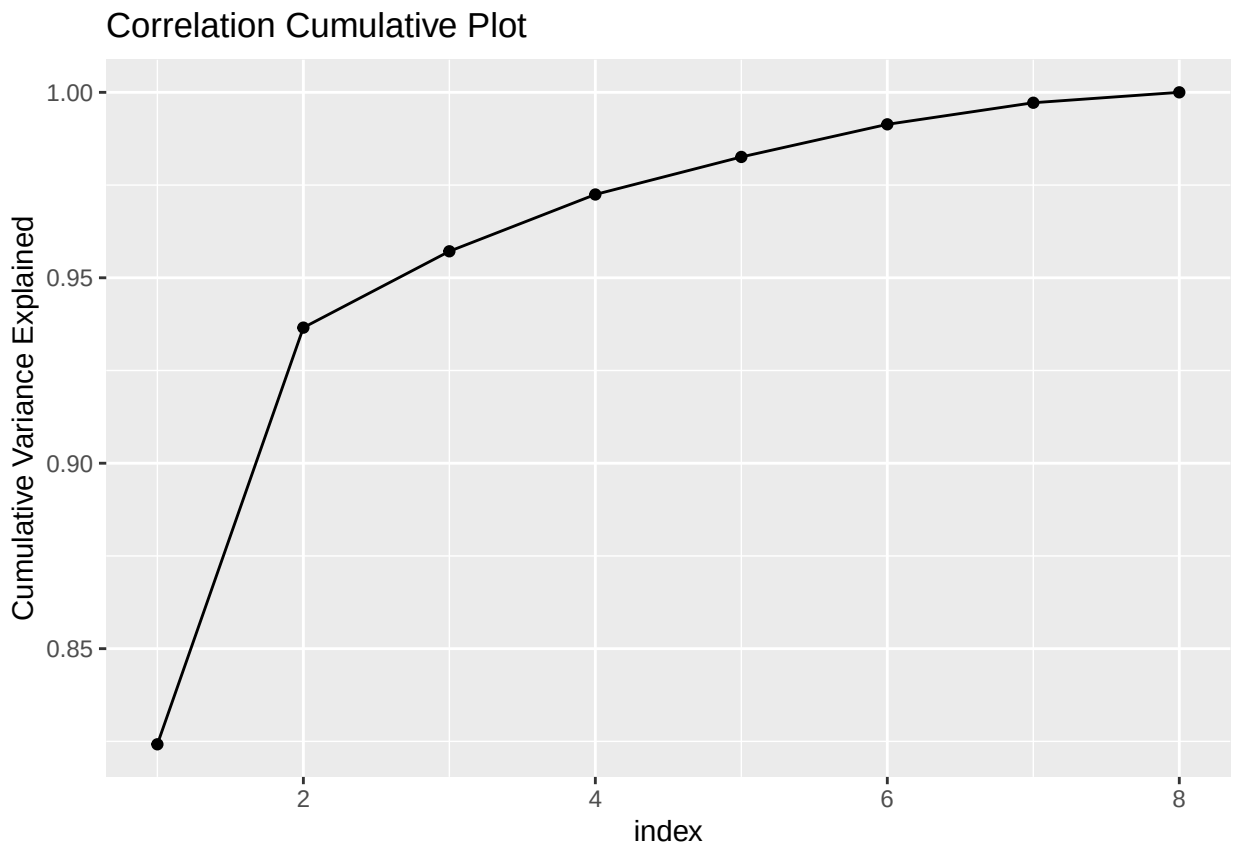

```
scree_cor <- ggplot(variance, aes(x=index, y=cor))+
  geom_point()+
  geom_line()+
  ggtitle("Correlation Scree Plot")
scree_cor
```




```
cum_cov <- ggplot(variance, aes(x=index, y=cumsum(cov)/sum(cov)))+
  geom_point()+
  geom_line()+
  labs(y="Cumulative Variance Explained", title="Covariance Cumulative Plot")
cum_cov
```




```
cum_cor <- ggplot(variance, aes(x=index, y=cumsum(cor)/sum(cor)))+
  geom_point()+
  geom_line()+
  labs(y="Cumulative Variance Explained", title="Correlation Cumulative Plot")
cum_cor
```



We should keep the first and second principal components. We can see this for where the plots 'elbow' in the plots for the correlation matrix.

#Problem 5

```
NYSEData <- read_csv("NYSEData.csv",show_col_types = FALSE)
```

a)

```
S <- cov(NYSEData)
nyse_pca <- prcomp(NYSEData)
nyse_pca
```

```
## Standard deviations (1, .., p=5):
## [1] 0.03698213 0.02647942 0.01593118 0.01194163 0.01090352
##
## Rotation (n x k) = (5 x 5):
##                         PC1        PC2         PC3        PC4         PC5
## JPMorgan        -0.2228228  0.6252260 -0.32611218  0.6627590 -0.11765952
## Citibank        -0.3072900  0.5703900  0.24959014 -0.4140935  0.58860803
## WellsFargo      -0.1548103  0.3445049  0.03763929 -0.4970499 -0.78030428
## RoyalDutchShell -0.6389680 -0.2479475  0.64249741  0.3088689 -0.14845546
## ExxonMobil      -0.6509044 -0.3218478 -0.64586064 -0.2163758  0.09371777
```

b)

```
summary(nyse_pca)
```

```
## Importance of components:
##                             PC1     PC2     PC3     PC4     PC5
## Standard deviation     0.03698 0.02648 0.01593 0.01194 0.01090
## Proportion of Variance 0.52926 0.27133 0.09822 0.05518 0.04601
## Cumulative Proportion  0.52926 0.80059 0.89881 0.95399 1.00000
```

The first three components cumulatively make up 89.88% of the variance. Individually, PC1 accounts for 52.93% of the variance, PC2 accounts for 27.13% of the variance, and PC3 9.8% of the variance.

c)