```
#Q2:

R <- cor(`stock.(1)`[, 1:5])
R
          V1             V2            V3              V4          V5
V1 1.0000000 0.6322878 0.5104973 0.1146019 0.1544628
V2 0.6322878 1.0000000 0.5741424 0.3222921 0.2126747
V3 0.5104973 0.5741424 1.0000000 0.1824992 0.1462067
V4 0.1146019 0.3222921 0.1824992 1.0000000 0.6833777
V5 0.1544628 0.2126747 0.1462067 0.6833777 1.0000000

eig <- eigen(R)
eig$vectors
           [,1]            [,2]             [,3]             [,4]             [,5]
[1,] -0.4690832  0.3680070 -0.60431522  0.3630228  0.38412160
[2,] -0.5324055  0.2364624 -0.13610618 -0.6292079 -0.49618794
[3,] -0.4651633  0.3151795  0.77182810  0.2889658  0.07116948
[4,] -0.3873459 -0.5850373  0.09336192 -0.3812515  0.59466408
[5,] -0.3606821 -0.6058463 -0.10882629  0.4934145 -0.49755167
 eig$values
[1] 2.4372731 1.4070127 0.5005127 0.4000316 0.2551699
```

```
#a
PC<- princomp(`stock.(1)`[, 1:5], cor = TRUE)

Call:
princomp(x = `stock.(1)`[, 1:5], cor = TRUE)

Standard deviations:
  Comp.1       Comp.2      Comp.3       Comp.4     Comp.5
1.5611768   1.1861756    0.7074693.  0.6324805   0.5051434

 5  variables and  103 observations.
```

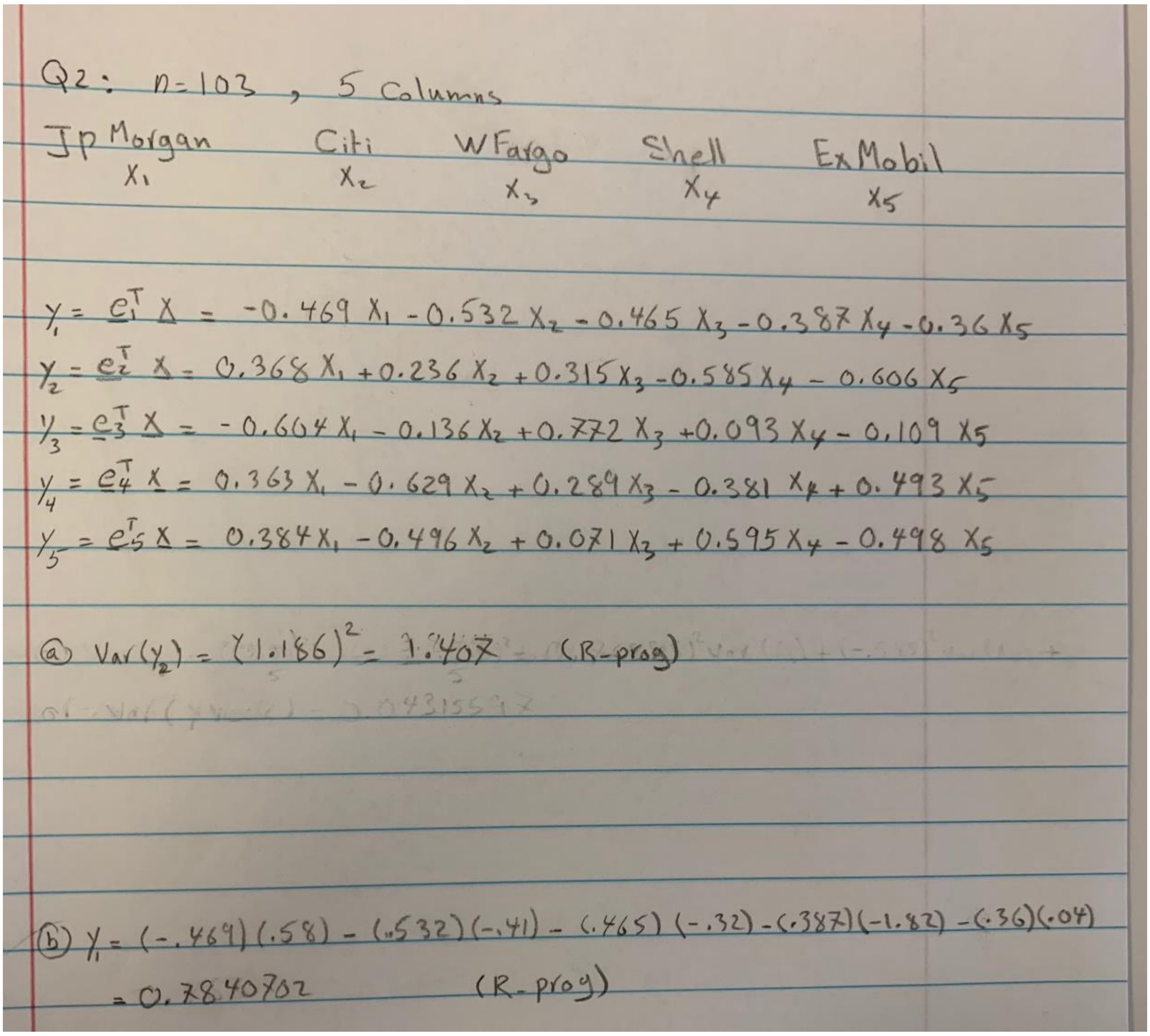
So the estimate of the variance of Y2 is (1.1861756)^2 = 1.407

```
#b

 S <- cov(`stock.(1)`[, 1:5])
 centerDat <- t(t(`stock.(1)`[, 1:5]) - apply(`stock.(1)`[, 1:5], 2, mean))
 stanDat <- as.matrix(centerDat) %*% diag(sqrt(1/diag(S)))
 obsVec1 <- as.numeric(stanDat[1,])
y1<- eig$vectors[,1]
yVec1 <- y1 %*% obsVec1
  0.7840702
```

```
#c
#cov(x2,y1)= e12 * lambda1
eig$values
[1] 2.4372731 1.4070127 0.5005127 0.4000316 0.2551699
> eig$vectors[,1]
[1] -0.4690832 -0.5324055 -0.4651633 -0.3873459 -0.3606821

-0.5324055 *2.4372731
= -1.297618
```



Q2: $n = 103$, 5 Columns

Jp Morgan	Citi	WFargo	Shell	ExMobil
X_1	X_2	X_3	X_4	X_5

$y_1 = \underline{e}_1^T \underline{X} = -0.469 X_1 - 0.532 X_2 - 0.465 X_3 - 0.387 X_4 - 0.36 X_5$

$y_2 = \underline{e}_2^T \underline{X} = 0.368 X_1 + 0.236 X_2 + 0.315 X_3 - 0.585 X_4 - 0.606 X_5$

$y_3 = \underline{e}_3^T \underline{X} = -0.664 X_1 - 0.136 X_2 + 0.772 X_3 + 0.093 X_4 - 0.109 X_5$

$y_4 = \underline{e}_4^T \underline{X} = 0.363 X_1 - 0.629 X_2 + 0.289 X_3 - 0.381 X_4 + 0.493 X_5$

$y_5 = \underline{e}_5^T \underline{X} = 0.384 X_1 - 0.496 X_2 + 0.071 X_3 + 0.595 X_4 - 0.498 X_5$

ⓐ $Var(y_2) = (1.186)^2 = 1.407$ (R-prog)

ⓑ $y_1 = (-.469)(.58) - (.532)(-.41) - (.465)(-.32) - (.387)(-1.82) - (.36)(.04)$

$= 0.7840702$ (R-prog)

#d

In the correlation matrix the average of the eigen values is 1.

```
 aveigvalues<- sum(eig$values) / length(eig$values)
 aveigvalues
[1] 1
eig$values
#[1] 2.4372731 1.4070127 0.5005127 0.4000316 0.2551699
```

We retain the first two principle component because the eigen values are greater than the average of the eigen values.

Or by the percentage of the total variance:

```
> cumsum(eig$values) / 5
[1] 0.4874546 0.7688572 0.8689597 0.9489660 1.0000000
```

If we retain the first three principle components the amount of variability is approximately 87% which I think is better.

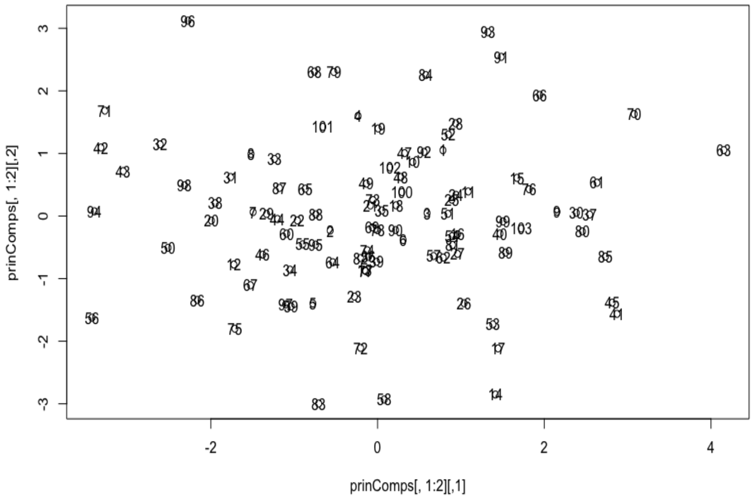
#e

First, we need to determine at what level the correlation is importance. Here a correlation above 0.45 is important.

The first Principal component coefficients is approximately weighted subtract of the five stocks. So, the first PC is strongly correlated with the three banks. The first PC increases with the decreasing rates of return for the three banks (JPMorgan, Citi and WFargo). This suggests that these three banks vary together. If one decrease, then the remaining ones tend to decrease.

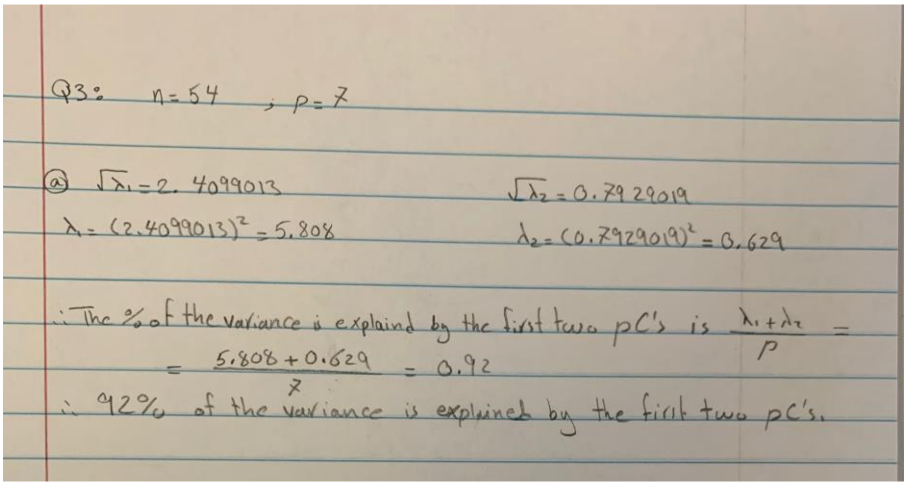
The second principal component coefficients represent a contrast between the banking stocks (JP Morgan, Citi, WFargo) and the oil stocks (Shell, EXMobil). This PC is strongly correlated with the two oil companies. The second PC increases with the decreasing rates of return for the two oil companies (Shell and ExMobile). This suggests that these two companies vary together. If one decreases the other company tend to decrease. Also, the second PC is the direction of the most important after we have accounted for the direction of the first most important.

```
#f
#plot first two pc's
plot(prinComps[,1:2])
text(prinComps[,1:2], labels = 1:103)
```



As we see in this plot, that each dot in this plot represents a week. Looking at the dot number 63 out by itself to the right, we may conclude that this particular dot has a very high value for the first principal component and we would expect this week to have high decrease values for the rates of return for the three banks. And kind of low value in the second PC.

#Q3



Q3: $n = 54$; $p = 7$

ⓐ $\sqrt{\lambda_1} = 2.4099013$ $\sqrt{\lambda_2} = 0.7929019$

$\lambda_1 = (2.4099013)^2 = 5.808$ $\lambda_2 = (0.7929019)^2 = 0.629$

∴ The % of the variance is explaind by the first two pC's is $\frac{\lambda_1 + \lambda_2}{p} =$

$= \frac{5.808 + 0.629}{7} = 0.92$

∴ 92% of the variance is explained by the first two pC's.

#b

The loadings (eigen vectors) are coefficients in linear combination predicting variables by the “standardized” components.

The eigen vectors when multiplied with vector X yields a scalar value which is Y.

Comp.1: The direction of the most important (the first largest variance as possible).Also, if all the records increases the first PC decreases. An overall measure, high values on this component indicate slower runner.

Comp.2: The direction of the most important after we have accounted for the direction of the first most important (the largest variance as possible and is orthogonal to the first component). Also contrast long and short races if we take the short races (100m, 200m and 400m) and the long races are (800m, 1500m, 3000m and Marathon). Small values indicate faster on short races than long ones. Large values indicate slower on short races than long ones. Value near zero means that tend to be similar on short and long races (could be slow, fast or somewhere in between on all races).

#c

The first PC for Kenya is y1= 0.926. which is a measure of the time spent in all of the records. If we say that the important correlation is above 0.3. This component is associated with the high decreasing on all these variables. They are all negatively related to PCA1 because they all have negative signs.

The second PC for Kenya is y2= -1.395. This component is associated with the first three tracks (100m, 200m and 400m) the less time they spent the longer time they spend in last three tracks (1500m, 3000m and Marathon) after accounting for the first PC.

So overall, the performance of women in Kenya for the track records with 200m is better than in 100m because they spent less time in the 200m which mean faster.