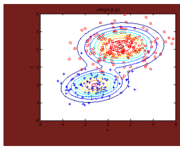


STAT 445/645 Assignment Cover Page

Student Name JooHyeok Seo

SFU Student Number 301258279

SFU email address jsa184@sfu.ca

Assignment Number Assignment5

Due Date 2024.04.04

Provide references for any data sets used in this assignment

- R. Johnson and D. Wichern, Applied Multivariate Statistics, Pearson, 6th Edition, 2018

List software used in this assignment.

I solved the problems and visulaize them using R

List **ALL** resources used to complete this assignment, including books, internet sources and people.

- I referred to the lecture notes from chapters 7 and 8

About fa function
- https://www.rdocumentation.org/packages/psych/versions/2.4.3/topics/fa

About cutoff
-https://rdrr.io/r/stats/loadings.html

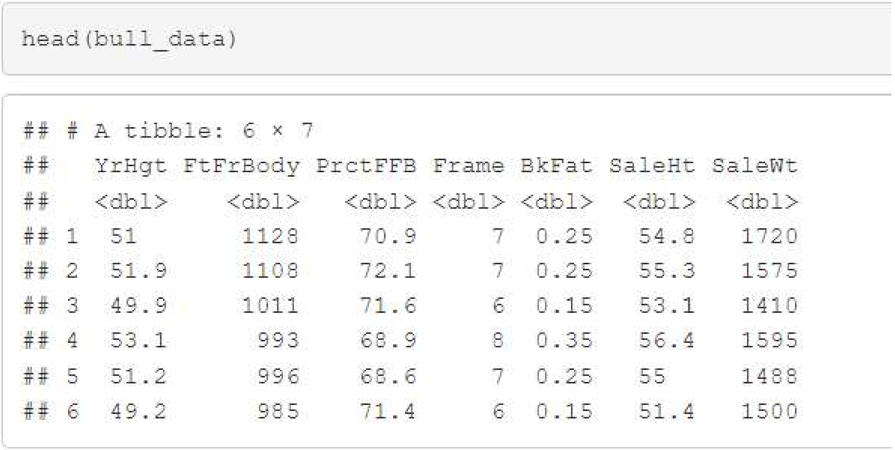
▣ I personally completed the computations and wrote the solutions submitted in this document.

Department of Statistics and Actuarial Science

SFU SIMON FRASER UNIVERSITY

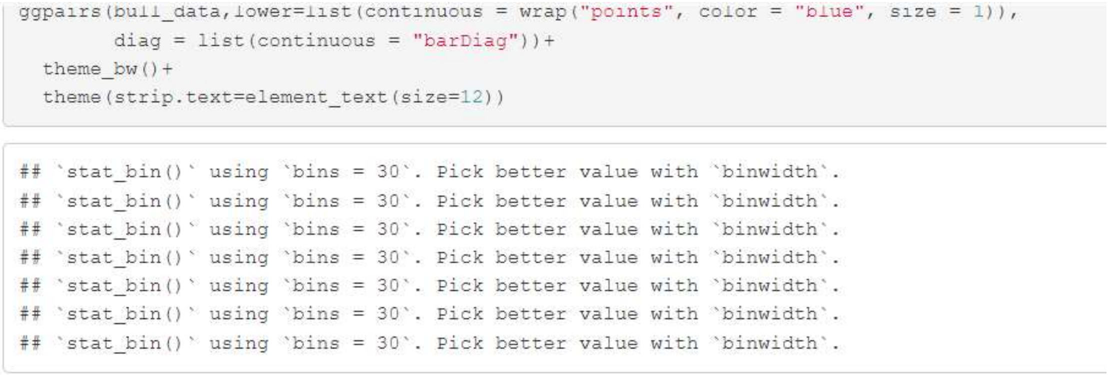
Question1

a. Carry out a shortened initial investigation



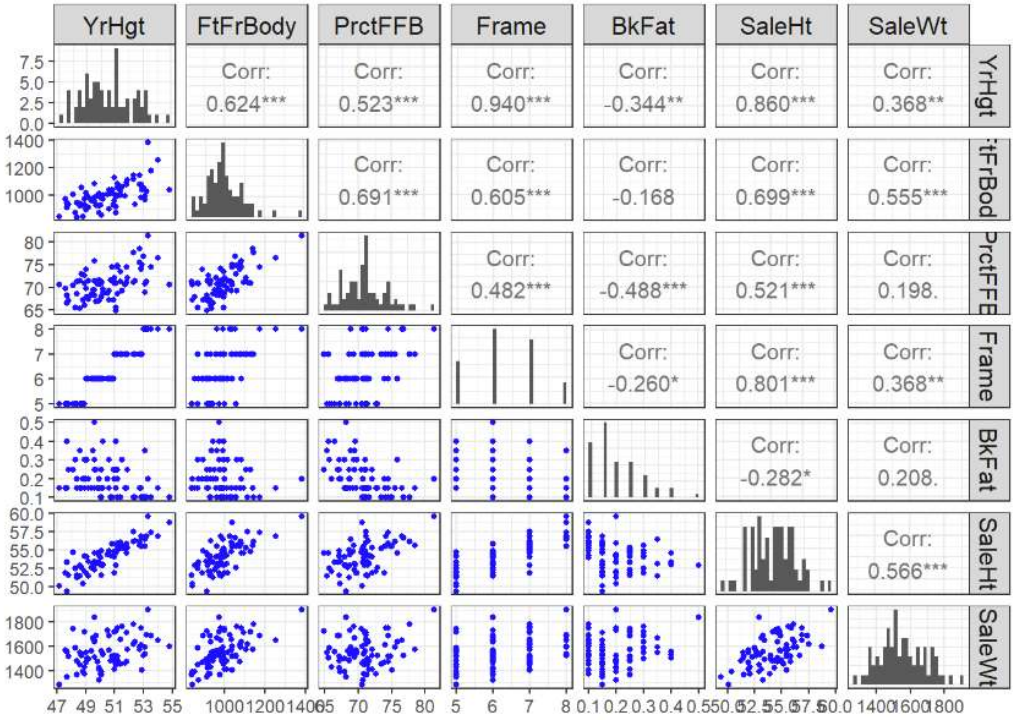
```
head(bull_data)
```

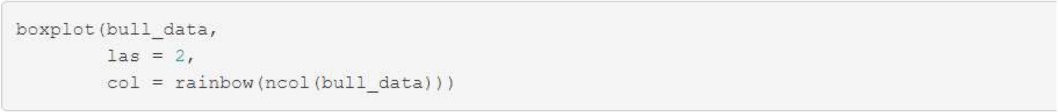
```
## # A tibble: 6 × 7
##   YrHgt FtFrBody PrctFFB Frame BkFat SaleHt SaleWt
##   <dbl>    <dbl>   <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1  51       1128    70.9     7  0.25   54.8   1720
## 2  51.9     1108    72.1     7  0.25   55.3   1575
## 3  49.9     1011    71.6     6  0.15   53.1   1410
## 4  53.1      993    68.9     8  0.35   56.4   1595
## 5  51.2      996    68.6     7  0.25   55     1488
## 6  49.2      985    71.4     6  0.15   51.4   1500
```



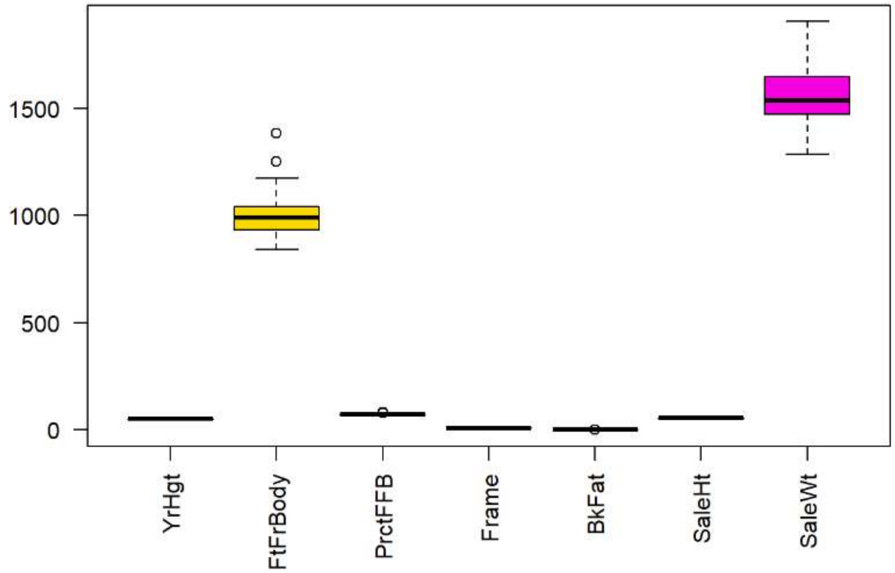
```
ggpairs(bull_data,lower=list(continuous = wrap("points", color = "blue", size = 1)),
        diag = list(continuous = "barDiag"))+
  theme_bw()+
  theme(strip.text=element_text(size=12))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





```
boxplot(bull_data,
        las = 2,
        col = rainbow(ncol(bull_data)))
```



According to Matrix Scatter Plot, a strong positive correlation was observed between Frame and YrHgt, with a correlation coefficient of 0.940, indicating a very strong linear relationship between the Frame size of the bull and its yearling height, suggesting that larger frame sizes are associated with greater heights at one year of age. This could imply that genetic or nutritional factors affecting growth are consistently influencing both traits in a similar manner.

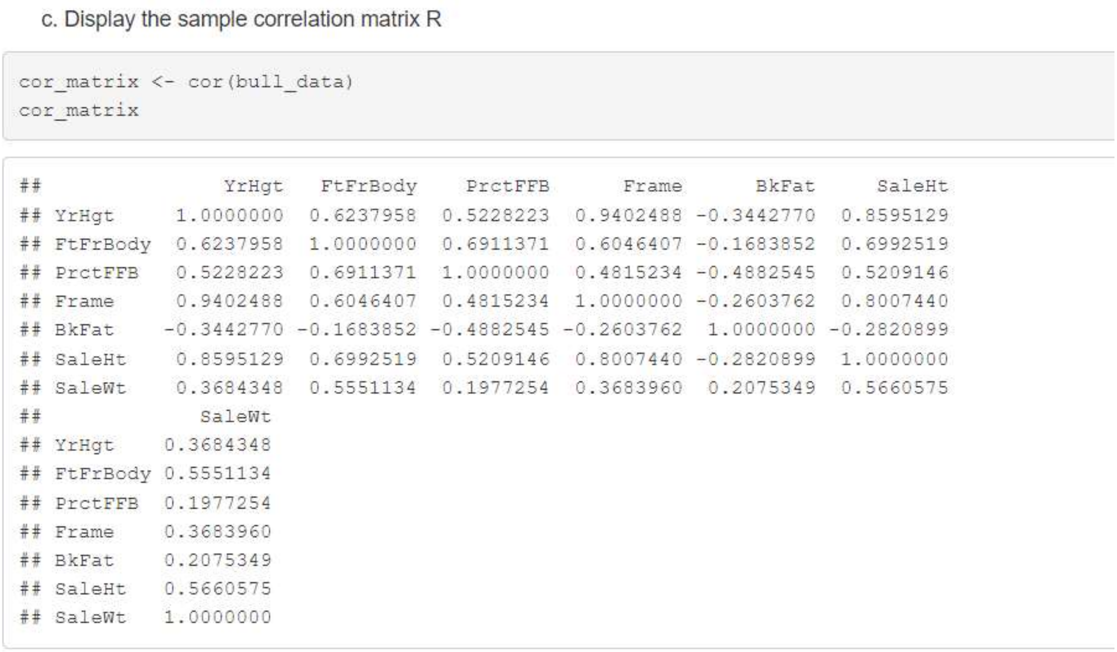
According to Box Plot, the distribution of each variable, as presented in the box plots, highlighted the presence of outlier across several measurements. Notably, variables such as SaleWt displayed outliers, which could indicate exceptional cases within the sales data, such as bulls with unusually high weights.

b) Explain why using the correlation matrix for the factor analysis is indicated

1. The correlation matrix provides insights into the linear relationships between variables

2. The correlation matrix normalizes the scale of the variables, which is critical since factor analysis assumes that the variables are on comparable scales

3. The correlation matrix is a standardized measure, meaning it is unaffected by the units of measurement, allowing for a meaningful comparison between variables that may be measured on different scales.

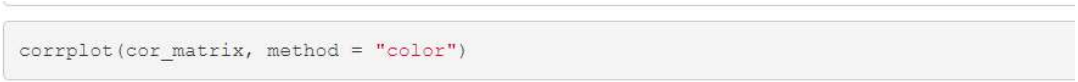


c. Display the sample correlation matrix R

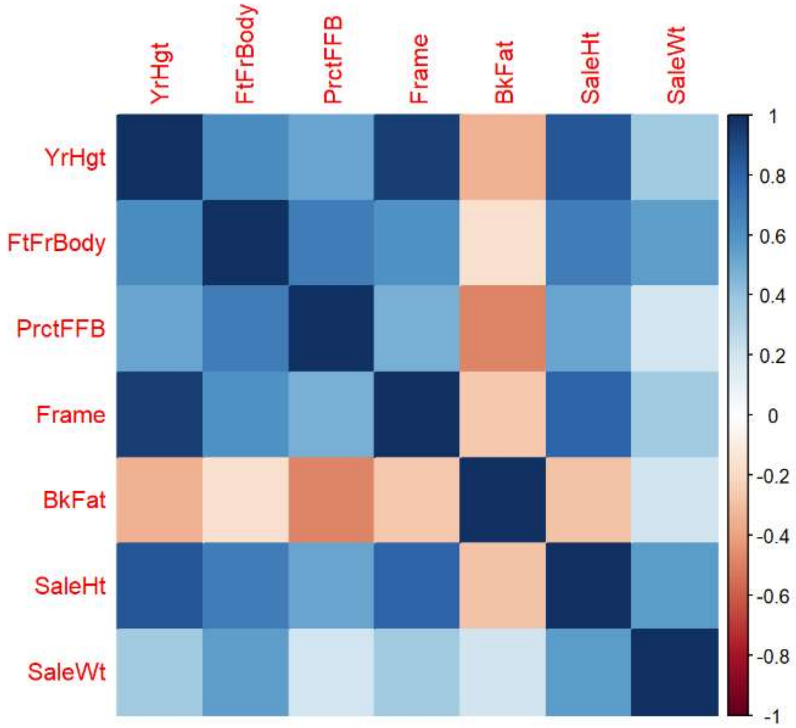
```
cor_matrix <- cor(bull_data)
cor_matrix
```

```
##                YrHgt   FtFrBody    PrctFFB      Frame      BkFat     SaleHt
## YrHgt      1.0000000  0.6237958  0.5228223  0.9402488 -0.3442770  0.8595129
## FtFrBody   0.6237958  1.0000000  0.6911371  0.6046407 -0.1683852  0.6992519
## PrctFFB    0.5228223  0.6911371  1.0000000  0.4815234 -0.4882545  0.5209146
## Frame      0.9402488  0.6046407  0.4815234  1.0000000 -0.2603762  0.8007440
## BkFat     -0.3442770 -0.1683852 -0.4882545 -0.2603762  1.0000000 -0.2820899
## SaleHt     0.8595129  0.6992519  0.5209146  0.8007440 -0.2820899  1.0000000
## SaleWt     0.3684348  0.5551134  0.1977254  0.3683960  0.2075349  0.5660575
##               SaleWt
## YrHgt     0.3684348
## FtFrBody  0.5551134
## PrctFFB   0.1977254
## Frame     0.3683960
## BkFat     0.2075349
## SaleHt    0.5660575
## SaleWt    1.0000000
```

The sample correlation matrix is $\begin{bmatrix} 1.00 & 0.62 & 0.52 & 0.94 & -0.34 & 0.86 & 0.37 \\ 0.62 & 1.00 & 0.69 & 0.60 & -0.17 & 0.70 & 0.56 \\ 0.53 & 0.69 & 1.00 & 0.48 & -0.49 & 0.52 & 0.20 \\ 0.94 & 0.60 & 0.48 & 1.00 & -0.26 & 0.80 & 0.37 \\ -0.34 & -0.17 & -0.49 & -0.26 & 1.00 & -0.28 & 0.21 \\ 0.86 & 0.70 & 0.52 & 0.80 & -0.28 & 1.00 & 0.57 \\ 0.37 & 0.56 & 0.20 & 0.37 & 0.21 & 0.57 & 1.00 \end{bmatrix}$

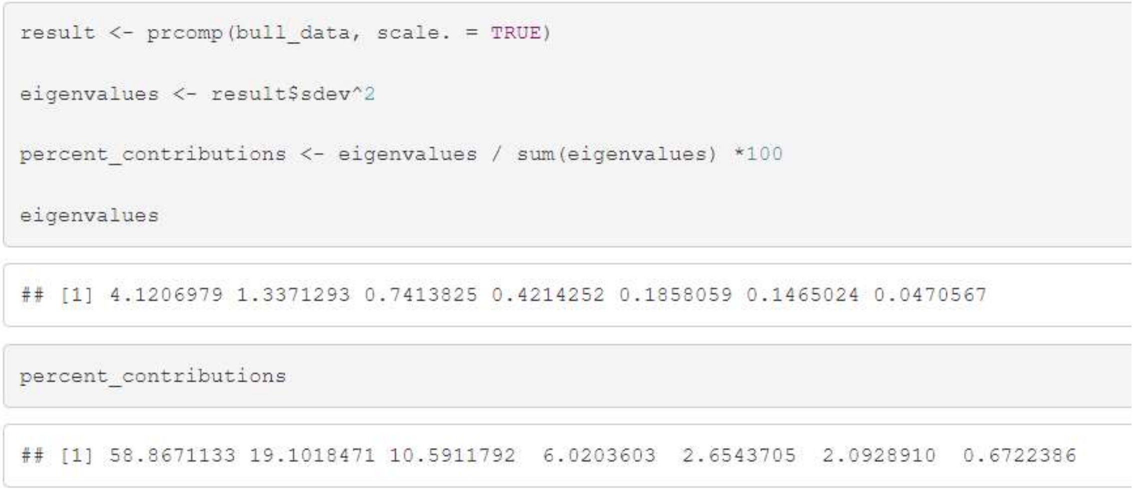


```
corrplot(cor_matrix, method = "color")
```



Strong correlations observed among sets of variables in the correlation matrix suggest the presence of common factors, indicating that fewer factors may be sufficient to explain the data. Additionally, a significant drop in eigenvalues after the first few factors in a principal component analysis implies that subsequent factors contribute less to the total variance, providing a basis for determining the number of factors to use.

d. i. List the eigenvalues and describe the percent contributions to the variance.



```
result <- prcomp(bull_data, scale. = TRUE)

eigenvalues <- result$sdev^2

percent_contributions <- eigenvalues / sum(eigenvalues) *100

eigenvalues
```

```
## [1] 4.1206979 1.3371293 0.7413825 0.4214252 0.1858059 0.1465024 0.0470567
```

```
percent_contributions
```

```
## [1] 58.8671133 19.1018471 10.5911792  6.0203603  2.6543705  2.0928910  0.6722386
```

The eigenvalue for the first principal component (PC1) is 4.12 and the first principal component explains 58.87% of the total variance.

The eigenvalue for the second principal component (PC2) is 1.34 and the second principal component explains 19.10% of the total variance.

The eigenvalue for (PC3) is 0.74 and the PC3 explains 10.59% of the total variance.

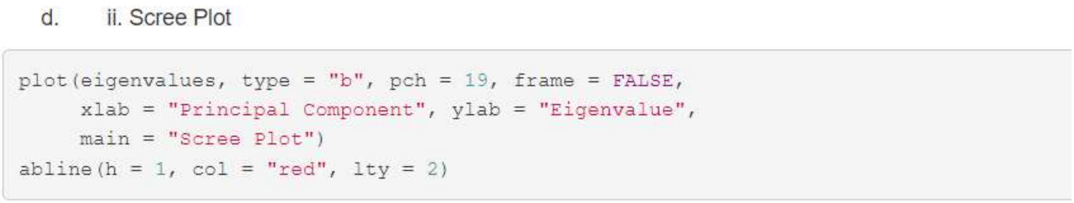
The eigenvalue for (PC4) is 0.42 and the PC4 explains 6.02% of the total variance.

The eigenvalue for (PC5) is 0.19 and the PC5 explains 2.65% of the total variance.

The eigenvalue for (PC6) is 0.15 and the PC6 explains 2.09% of the total variance.

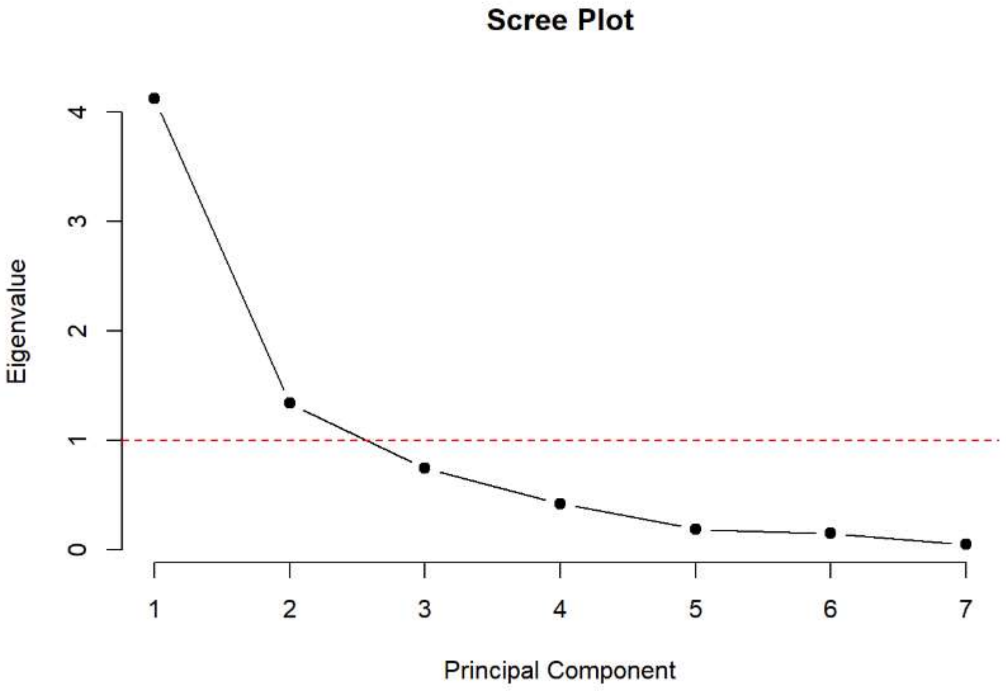
The eigenvalue for (PC7) is 0.05 and the PC7 explains 0.67% of the total variance.

The first principal component (PC1) explains 58.87% of the variance, indicating it captures the majority of the data's variability. PC2 adds 19.10%, bringing the cumulative variance explained by the first two components to nearly 78%. PC3 and PC4 explain 10.59% and 6.02%, respectively, showing diminishing contributions from subsequent components. The remaining components each contribute less than 3% to the total variance, highlighting that the most significant patterns are captured by the first few components.



d. ii. Scree Plot

```
plot(eigenvalues, type = "b", pch = 19, frame = FALSE,
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot")
abline(h = 1, col = "red", lty = 2)
```



Based on the PC analysis, the first three components explain a significant portion of the variance (88.6%).

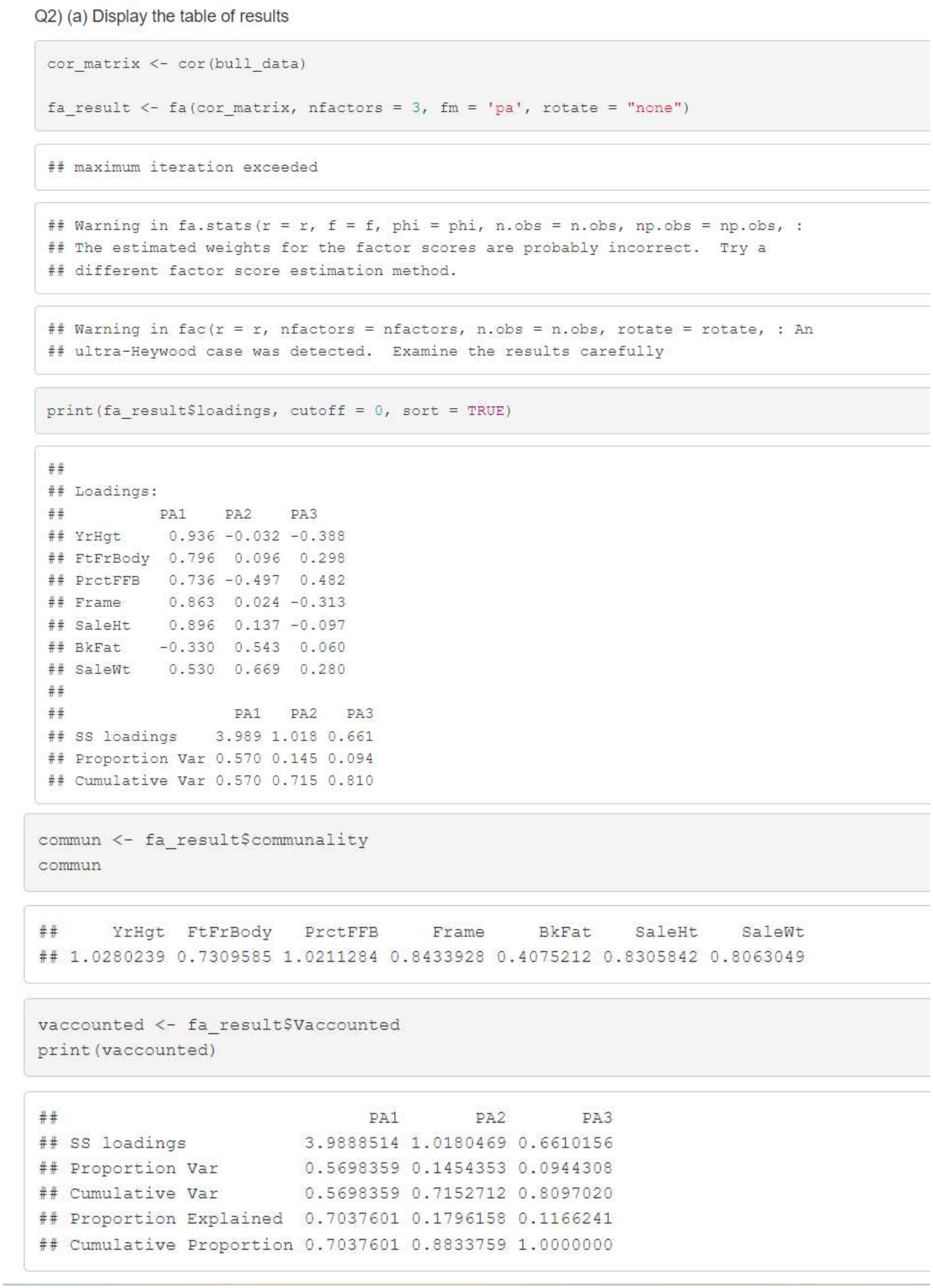
Kaiser’s Criterion suggests retaining components with eigenvalues over 1, which would include the first three components.

The Scree Plot shows a clear elbow after the third component, indicating that additional components contribute less to explaining the variance.

The cumulative Variance Explained method might suggest retaining more components if we aim for a higher percentage of explained variance, like 90% or more.

While determining the number of principal components to retain, we observe that the Kaiser's Criterion and the Scree Plot both suggest retaining the first three components, aligning well with the significant portion of the variance they explain. However, if the Cumulative Variance Explained method suggests retaining more components to cover a higher percentage of the variance, this could lead to a discrepancy between the methods. For instance, if aiming for 90% of the explained variance necessitates keeping more than three components, this would diverge from the recommendations of the Kaiser's Criterion and the Scree Plot.

Question2



Q2) (a) Display the table of results

```
cor_matrix <- cor(bull_data)

fa_result <- fa(cor_matrix, nfactors = 3, fm = 'pa', rotate = "none")
```

```
## maximum iteration exceeded
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, : An
## ultra-Heywood case was detected.  Examine the results carefully
```

```
print(fa_result$loadings, cutoff = 0, sort = TRUE)
```

```
##
## Loadings:
##          PA1    PA2    PA3   
## YrHgt     0.936 -0.032 -0.388
## FtFrBody  0.796  0.096  0.298
## PrctFFB   0.736 -0.497  0.482
## Frame     0.863  0.024 -0.313
## SaleHt    0.896  0.137 -0.097
## BkFat    -0.330  0.543  0.060
## SaleWt    0.530  0.669  0.280
##
##                  PA1   PA2   PA3
## SS loadings    3.989 1.018 0.661
## Proportion Var 0.570 0.145 0.094
## Cumulative Var 0.570 0.715 0.810
```

```
commun <- fa_result$communality
commun
```

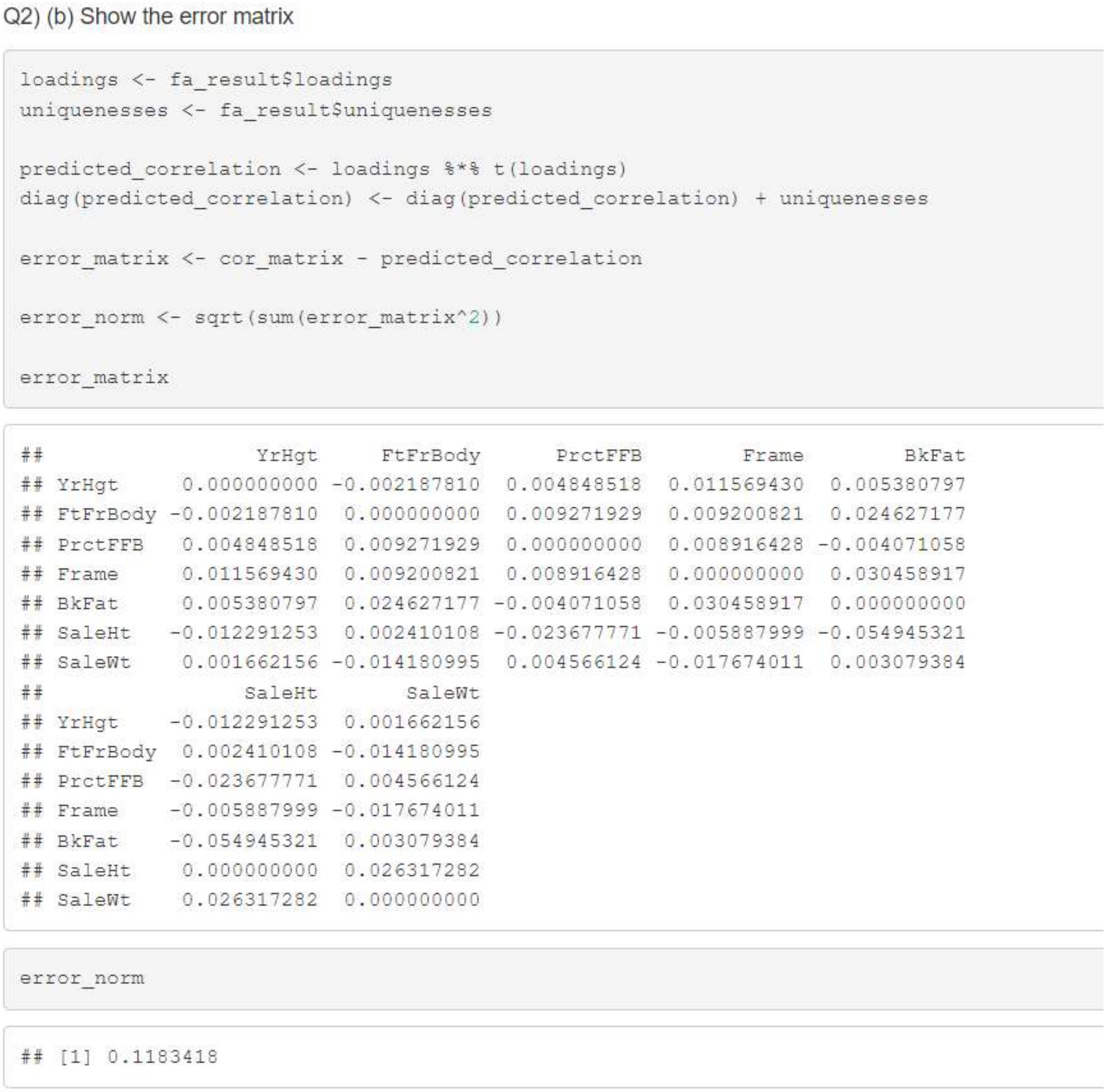
```
##     YrHgt  FtFrBody   PrctFFB     Frame     BkFat    SaleHt    SaleWt 
## 1.0280239 0.7309585 1.0211284 0.8433928 0.4075212 0.8305842 0.8063049
```

```
vaccounted <- fa_result$Vaccounted
print(vaccounted)
```

```
##                             PA1       PA2       PA3
## SS loadings           3.9888514 1.0180469 0.6610156
## Proportion Var        0.5698359 0.1454353 0.0944308
## Cumulative Var        0.5698359 0.7152712 0.8097020
## Proportion Explained  0.7037601 0.1796158 0.1166241
## Cumulative Proportion 0.7037601 0.8833759 1.0000000
```

The table of results is

Variables	load 1	load 2	load 3	commun
YrHgt	0.936	-0.032	-0.388	1.028
FtFrBody	0.796	0.096	0.298	0.731
PrctFFB	0.736	-0.497	0.482	1.021
Frame	0.863	0.024	-0.313	0.843
SaleHt	0.896	0.137	-0.097	0.831
BkFat	-0.330	0.543	0.060	0.408
SaleWt	0.530	0.669	0.280	0.806
Var. Acc. For	0.570	0.145	0.094	



Q2) (b) Show the error matrix

```
loadings <- fa_result$loadings
uniquenesses <- fa_result$uniquenesses

predicted_correlation <- loadings %*% t(loadings)
diag(predicted_correlation) <- diag(predicted_correlation) + uniquenesses

error_matrix <- cor_matrix - predicted_correlation

error_norm <- sqrt(sum(error_matrix^2))

error_matrix
```

```
##                 YrHgt     FtFrBody      PrctFFB        Frame        BkFat
## YrHgt     0.000000000 -0.002187810  0.004848518  0.011569430  0.005380797
## FtFrBody -0.002187810  0.000000000  0.009271929  0.009200821  0.024627177
## PrctFFB   0.004848518  0.009271929  0.000000000  0.008916428 -0.004071058
## Frame     0.011569430  0.009200821  0.008916428  0.000000000  0.030458917
## BkFat     0.005380797  0.024627177 -0.004071058  0.030458917  0.000000000
## SaleHt   -0.012291253  0.002410108 -0.023677771 -0.005887999 -0.054945321
## SaleWt    0.001662156 -0.014180995  0.004566124 -0.017674011  0.003079384
##                SaleHt       SaleWt
## YrHgt    -0.012291253  0.001662156
## FtFrBody  0.002410108 -0.014180995
## PrctFFB  -0.023677771  0.004566124
## Frame    -0.005887999 -0.017674011
## BkFat    -0.054945321  0.003079384
## SaleHt    0.000000000  0.026317282
## SaleWt    0.026317282  0.000000000
```

```
error_norm
```

```
## [1] 0.1183418
```

The error matrix is $\begin{bmatrix} 0.000 & -0.002 & 0.005 & 0.012 & 0.005 & -0.012 & 0.002 \\ -0.002 & 0.000 & 0.009 & 0.009 & 0.025 & 0.002 & -0.014 \\ 0.005 & 0.009 & 0.000 & 0.009 & -0.004 & -0.024 & 0.005 \\ 0.012 & 0.009 & 0.009 & 0.000 & 0.030 & -0.006 & -0.018 \\ 0.005 & 0.025 & -0.004 & 0.030 & 0.000 & -0.055 & 0.003 \\ -0.012 & 0.002 & -0.024 & -0.006 & -0.055 & 0.000 & 0.026 \\ 0.002 & -0.014 & 0.005 & -0.018 & 0.003 & 0.026 & 0.000 \end{bmatrix}$

And ||E||is 0.118

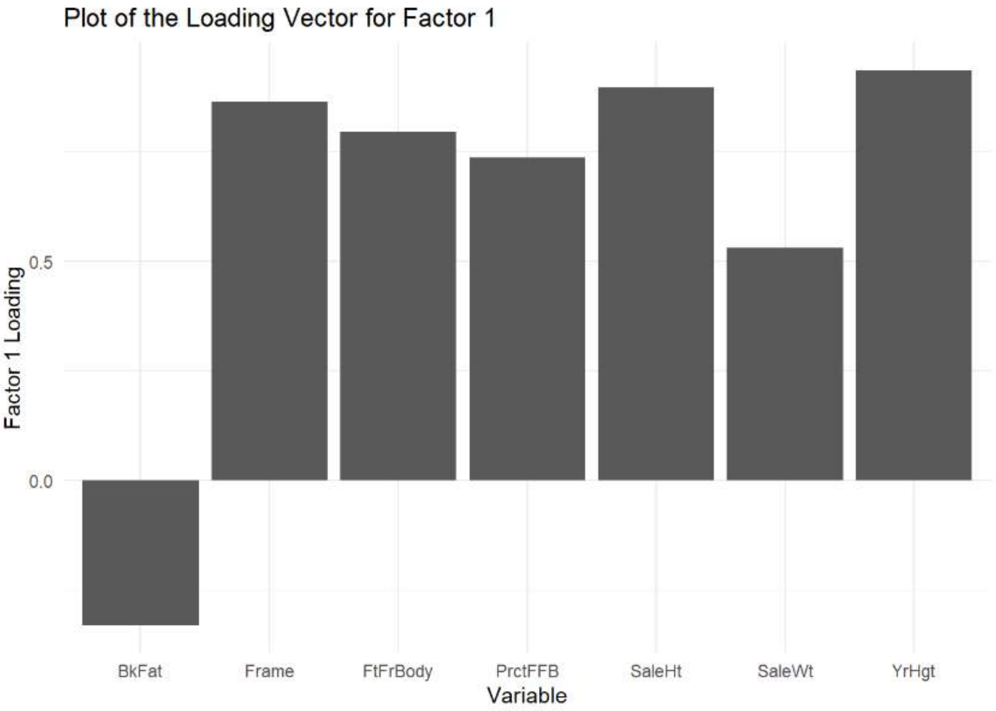
Q2) (c) Show plots of the loading vectors for the first two factors

```
loadings <- fa_result$loadings[, 1:2]
loadings
```

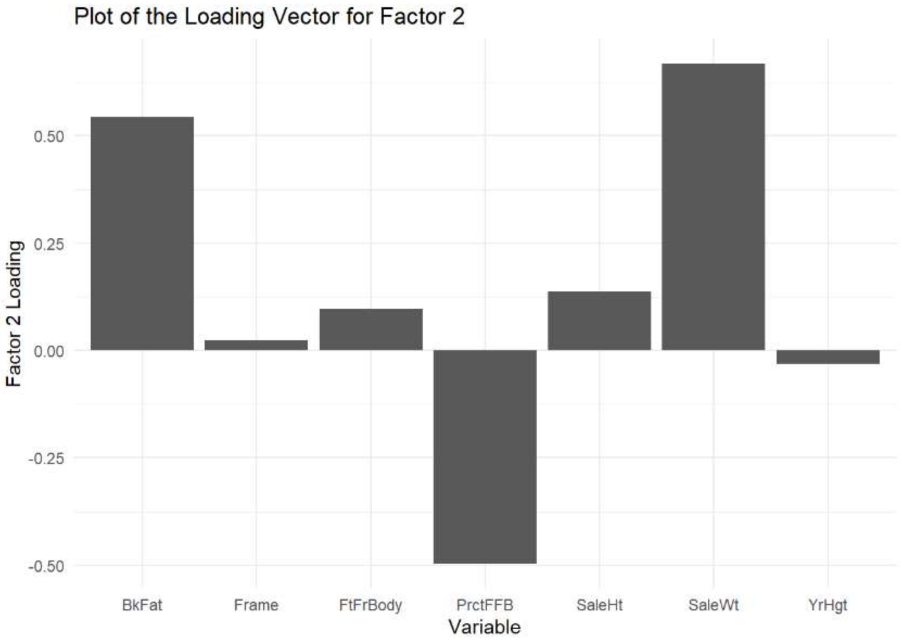
```
##                 PA1         PA2
## YrHgt      0.9361376 -0.03154950
## FtFrBody   0.7955382  0.09590139
## PrctFFB    0.7363894 -0.49662535
## Frame      0.8629621  0.02401372
## BkFat     -0.3304488  0.54291593
## SaleHt     0.8957041  0.13746342
## SaleWt     0.5302467  0.66857096
```

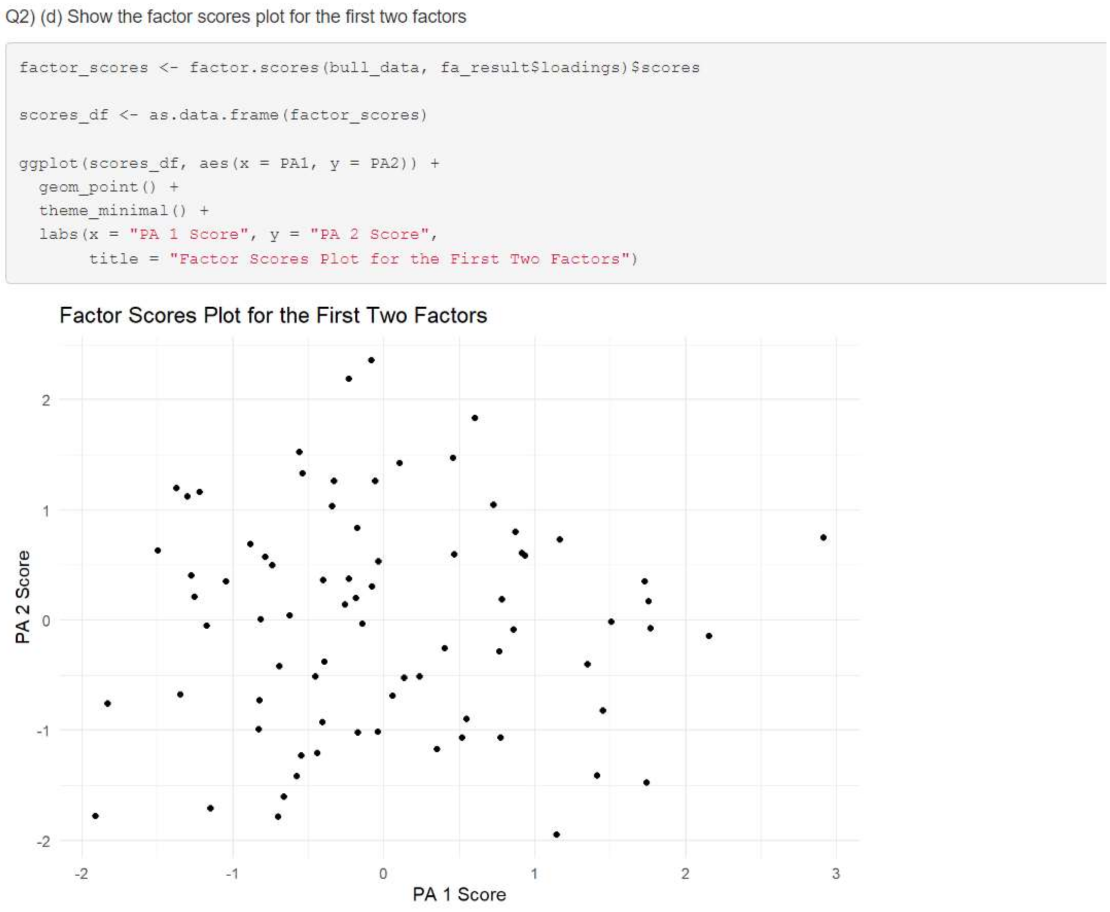
```
loading_df <- as.data.frame(loadings)
loading_df$Variable <- rownames(loading_df)

ggplot(loading_df, aes(x = Variable, y = PA1)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(y = "Factor 1 Loading",
       title = "Plot of the Loading Vector for Factor 1")
```



```r
ggplot(loading_df, aes(x = Variable, y = PA2)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(y = "Factor 2 Loading",
       title = "Plot of the Loading Vector for Factor 2")
```





Q2) (d) Show the factor scores plot for the first two factors

```
factor_scores <- factor.scores(bull_data, fa_result$loadings)$scores

scores_df <- as.data.frame(factor_scores)

ggplot(scores_df, aes(x = PA1, y = PA2)) +
  geom_point() +
  theme_minimal() +
  labs(x = "PA 1 Score", y = "PA 2 Score",
       title = "Factor Scores Plot for the First Two Factors")
```

e) The first factor (load 1) shows a high loading for YrHgt, indicating a strong association with annual height. Conversely, variables like BkFat and SaleWt exhibit higher loadings on the second factor (load 2), suggesting that sale-related characteristics are more closely related to the second factor.

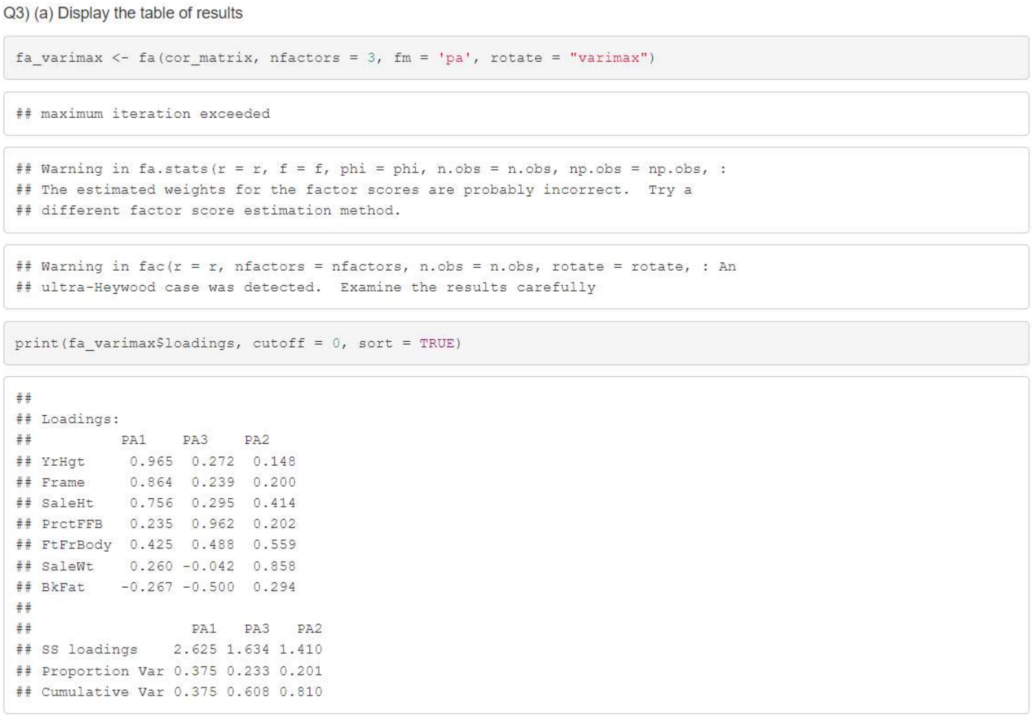
The communality for YrHgt and PrctFFB are close to 1, indicating that these variables are almost entirely explained by the factor model. This demonstrates that the factor model captures the major variability in the dataset effectively.

The first factor accounts for about 57% of the total variance, highlighting its significance in encapsulating most of the information. The second factor explains an additional 14.5% of the variance.

The error matrix and its norm of 0.118 suggest that the model reasonably fits the data well. However, further model refinement or consideration of other variables might be required for more precise fitting.

Most data points score higher on the first factor, while the distribution across the second factor is more spread out. This indicates that while the first factor captures the primary variability, the second factor represents additional structural characteristics.

Question3



Q3) (a) Display the table of results

```
fa_varimax <- fa(cor_matrix, nfactors = 3, fm = 'pa', rotate = "varimax")
```

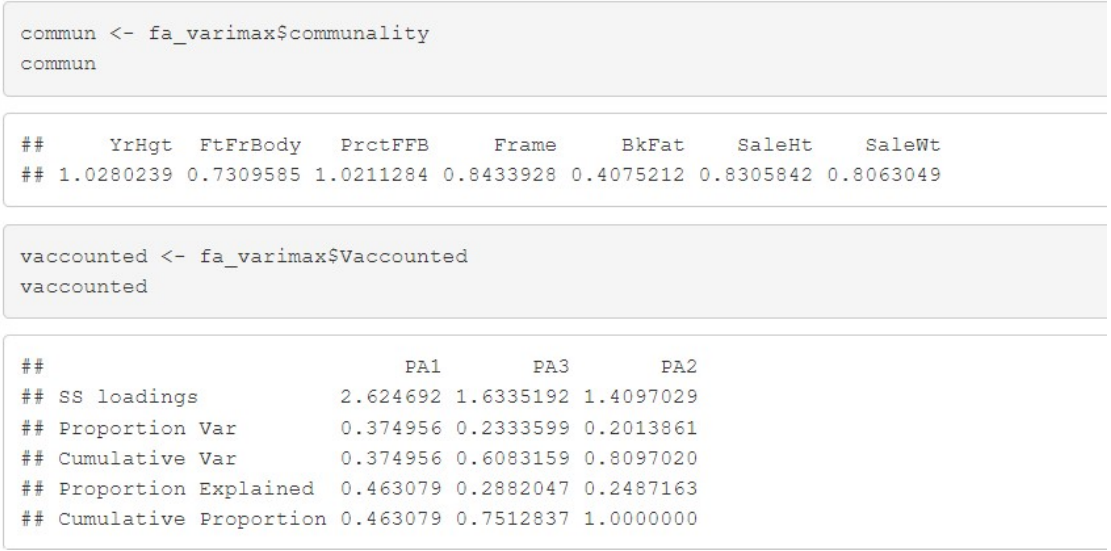
```
## maximum iteration exceeded
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, : An
## ultra-Heywood case was detected.  Examine the results carefully
```

```
print(fa_varimax$loadings, cutoff = 0, sort = TRUE)
```

```
##
## Loadings:
##          PA1    PA3    PA2
## YrHgt     0.965  0.272  0.148
## Frame     0.864  0.239  0.200
## SaleHt    0.756  0.295  0.414
## PrctFFB   0.235  0.962  0.202
## FtFrBody  0.425  0.488  0.559
## SaleWt    0.260 -0.042  0.858
## BkFat    -0.267 -0.500  0.294
##
##                  PA1   PA3   PA2
## SS loadings    2.625 1.634 1.410
## Proportion Var 0.375 0.233 0.201
## Cumulative Var 0.375 0.608 0.810
```



```
commun <- fa_varimax$communality
commun
```

```
##     YrHgt  FtFrBody   PrctFFB     Frame     BkFat    SaleHt    SaleWt
## 1.0280239 0.7309585 1.0211284 0.8433928 0.4075212 0.8305842 0.8063049
```

```
vaccounted <- fa_varimax$Vaccounted
vaccounted
```

```
##                            PA1       PA3       PA2
## SS loadings           2.624692 1.6335192 1.4097029
## Proportion Var        0.374956 0.2333599 0.2013861
## Cumulative Var        0.374956 0.6083159 0.8097020
## Proportion Explained  0.463079 0.2882047 0.2487163
## Cumulative Proportion 0.463079 0.7512837 1.0000000
```

The table of results is

Variables	load 1	load 3	load 2	commun
YrHgt	0.965	0.272	0.148	1.028
Frame	0.864	0.239	0.200	0.843
SaleHt	0.756	0.295	0.414	0.831
PrctFFB	0.235	0.962	0.202	1.021
FtFrBody	0.425	0.488	0.559	0.731
SaleWt	0.260	-0.042	0.858	0.806

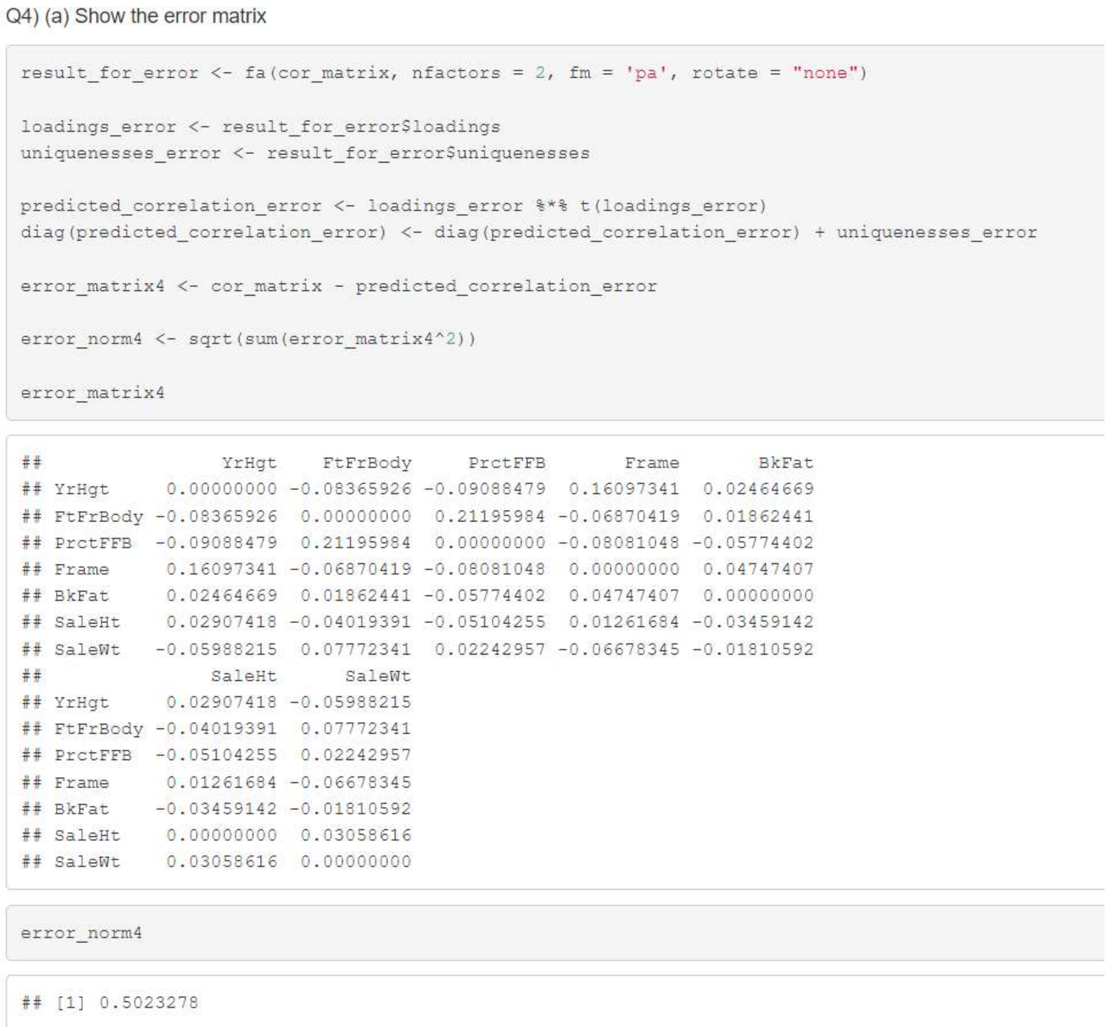
BkFat	-0.267	-0.500	0.294	0.408
Var. Acc. For	0.375	0.233	0.201	

b) The first factor (load 1) has high loadings primarily on YrHgt (0.965), Frame (0.864), and SaleHt (0.756), which can be interpreted as a factor related to the physical attributes of the bulls, hence termed the Physical Size factor. This suggests that larger physical characteristics of bulls correspond to higher loadings on this factor

The third factor (load 2), showing high loadings on SaleWt (0.806), can be interpreted as related to the weight of the bulls at the time of sale, thus can be termed the Sale Weight factor. This means that bulls that are heavier at the time of sale have higher loadings on this factor, indicating that they contribute more significantly to the Sale Weight factor.

The second factor (load 3), with a very high loading on PrctFFB (0.962), could represent the body fat percentage of the bulls, hence it can be termed the Body Fat Percentage factor. This relationship indicates that bulls with a higher percentage of body fat have higher loadings on this factor.

Question4



Q4) (a) Show the error matrix

```
result_for_error <- fa(cor_matrix, nfactors = 2, fm = 'pa', rotate = "none")

loadings_error <- result_for_error$loadings
uniquenesses_error <- result_for_error$uniquenesses

predicted_correlation_error <- loadings_error %*% t(loadings_error)
diag(predicted_correlation_error) <- diag(predicted_correlation_error) + uniquenesses_error

error_matrix4 <- cor_matrix - predicted_correlation_error

error_norm4 <- sqrt(sum(error_matrix4^2))

error_matrix4
```

```
##                 YrHgt    FtFrBody     PrctFFB       Frame       BkFat
## YrHgt      0.00000000 -0.08365926 -0.09088479  0.16097341  0.02464669
## FtFrBody  -0.08365926  0.00000000  0.21195984 -0.06870419  0.01862441
## PrctFFB   -0.09088479  0.21195984  0.00000000 -0.08081048 -0.05774402
## Frame      0.16097341 -0.06870419 -0.08081048  0.00000000  0.04747407
## BkFat      0.02464669  0.01862441 -0.05774402  0.04747407  0.00000000
## SaleHt     0.02907418 -0.04019391 -0.05104255  0.01261684 -0.03459142
## SaleWt    -0.05988215  0.07772341  0.02242957 -0.06678345 -0.01810592
##                SaleHt      SaleWt
## YrHgt      0.02907418 -0.05988215
## FtFrBody  -0.04019391  0.07772341
## PrctFFB   -0.05104255  0.02242957
## Frame      0.01261684 -0.06678345
## BkFat     -0.03459142 -0.01810592
## SaleHt     0.00000000  0.03058616
## SaleWt     0.03058616  0.00000000
```

```
error_norm4
```

```
## [1] 0.5023278
```

The error matrix is $\begin{bmatrix} 0.000 & -0.084 & -0.091 & 0.161 & 0.025 & 0.029 & -0.060 \\ -0.084 & 0.000 & 0.212 & -0.069 & 0.019 & -0.040 & 0.078 \\ -0.091 & 0.212 & 0.000 & -0.081 & -0.058 & -0.051 & 0.022 \\ 0.161 & -0.069 & -0.081 & 0.000 & 0.048 & 0.013 & -0.067 \\ 0.025 & 0.019 & -0.058 & 0.0475 & 0.000 & -0.035 & -0.018 \\ 0.029 & -0.040 & -0.051 & 0.013 & -0.035 & 0.000 & 0.031 \\ -0.060 & 0.078 & 0.023 & -0.067 & -0.018 & 0.031 & 0.000 \end{bmatrix}$

And ||E|| is 0.502.

b) The norm value of 0.502 for the error matrix indicates that the factor analysis model with two factors adequately explains the correlation structure of the data to a certain extent. A norm value closer to 0 would signify a better model fit, thus a value of 0.502 suggests that the model has a reasonable level of explanatory power, although it is not perfect.