

# Principal Components Analysis and Factor Analysis

Long Pham

## 1. EDA:

```
dfraw = read.csv("health.csv", header=TRUE)
str(dfraw)
```

```
## 'data.frame':    450 obs. of  17 variables:
## $ Q1      : int  15 15 15 15 15 16 15 14 15 13 ...
## $ Q2      : int  14 15 15 15 16 16 15 14 15 13 ...
## $ Q3      : int  14 14 15 14 15 16 15 14 15 13 ...
## $ Q4      : int  15 15 15 14 17 16 15 14 15 14 ...
## $ Q5      : int  13 15 15 15 15 16 16 14 14 13 ...
## $ Q6      : int  12 13 11 12 11 12 11 12 12 12 ...
## $ Q7      : int  12 13 12 12 12 12 11 12 12 12 ...
## $ Q8      : int  13 12 11 12 12 13 11 12 12 12 ...
## $ Q9      : int  12 13 11 12 11 12 11 12 12 12 ...
## $ Q10     : int  13 12 12 11 12 11 11 11 12 13 ...
## $ Q11     : int  10 10 11 11 9 11 11 10 10 9 ...
## $ Q12     : int  10 10 11 11 9 10 11 10 10 9 ...
## $ Q13     : int  9 10 11 11 8 10 12 10 10 9 ...
## $ Q14     : int  10 10 10 11 10 10 12 10 11 9 ...
## $ Q15     : int  10 9 12 10 9 11 11 9 11 9 ...
## $ diet     : chr  "Vegan" "Vegan" "Vegan" "Non-Vegetarian" ...
## $ body_type: chr  "Ectomorph" "Ectomorph" "Ectomorph" "Ectomorph" ...
```

```
df = dfraw[, -c(16:17)]
```

```
# View the first few rows of the dataset
head(dfraw)
```

```
##   Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15      diet body_type
## 1 15 14 14 15 13 12 12 13 12 13 10 10 9 10 10      Vegan Ectomorph
## 2 15 15 14 15 15 13 13 12 13 12 10 10 10 10 9      Vegan Ectomorph
## 3 15 15 15 15 15 11 12 11 11 12 11 11 11 10 12      Vegan Ectomorph
## 4 15 15 14 14 15 12 12 12 12 11 11 11 11 11 10 Non-Vegetarian Ectomorph
## 5 15 16 15 17 15 11 12 12 11 12 9 9 8 10 9      Vegan Ectomorph
## 6 16 16 16 16 16 12 12 13 12 11 11 10 10 10 11 Non-Vegetarian Ectomorph
```

```
# Summary statistics for the entire dataset
summary(dfraw)
```

```
##           Q1           Q2           Q3           Q4           Q5
## Min.      : 8.0      Min.      : 8.00      Min.      : 8.0      Min.      : 8.00      Min.      : 8.00
## 1st Qu.:11.0      1st Qu.:11.00      1st Qu.:10.0      1st Qu.:10.00      1st Qu.:10.00
## Median :12.0      Median :12.00      Median :12.0      Median :12.00      Median :12.00
## Mean      :12.3      Mean      :12.32      Mean      :12.3      Mean      :12.31      Mean      :12.31
## 3rd Qu.:14.0      3rd Qu.:14.00      3rd Qu.:14.0      3rd Qu.:15.00      3rd Qu.:14.00
## Max.      :17.0      Max.      :18.00      Max.      :17.0      Max.      :17.00      Max.      :18.00
##           Q6           Q7           Q8           Q9
## Min.      : 8.00      Min.      : 8.00      Min.      : 8.00      Min.      : 8.00
## 1st Qu.:10.00      1st Qu.:10.00      1st Qu.:10.00      1st Qu.:10.00
## Median :12.00      Median :12.00      Median :12.00      Median :12.00
## Mean      :12.33      Mean      :12.35      Mean      :12.34      Mean      :12.35
## 3rd Qu.:14.00      3rd Qu.:14.00      3rd Qu.:14.00      3rd Qu.:14.00
## Max.      :18.00      Max.      :17.00      Max.      :17.00      Max.      :17.00
##           Q10          Q11          Q12          Q13
## Min.      : 8.00      Min.      : 8.00      Min.      : 8.00      Min.      : 8.00
## 1st Qu.:10.00      1st Qu.:10.25      1st Qu.:11.00      1st Qu.:11.00
## Median :12.00      Median :12.00      Median :12.00      Median :12.00
## Mean      :12.35      Mean      :12.37      Mean      :12.38      Mean      :12.38
## 3rd Qu.:14.00      3rd Qu.:15.00      3rd Qu.:15.00      3rd Qu.:15.00
## Max.      :18.00      Max.      :17.00      Max.      :17.00      Max.      :17.00
##           Q14          Q15          diet          body_type
## Min.      : 8.00      Min.      : 8.00      Length:450      Length:450
## 1st Qu.:11.00      1st Qu.:11.00      Class :character  Class :character
## Median :12.00      Median :12.00      Mode  :character  Mode  :character
## Mean      :12.36      Mean      :12.36
## 3rd Qu.:15.00      3rd Qu.:15.00
## Max.      :17.00      Max.      :17.00
```

## Checking NA values

```
# Check for missing values
sum(is.na(df))
```

```
## [1] 0
```

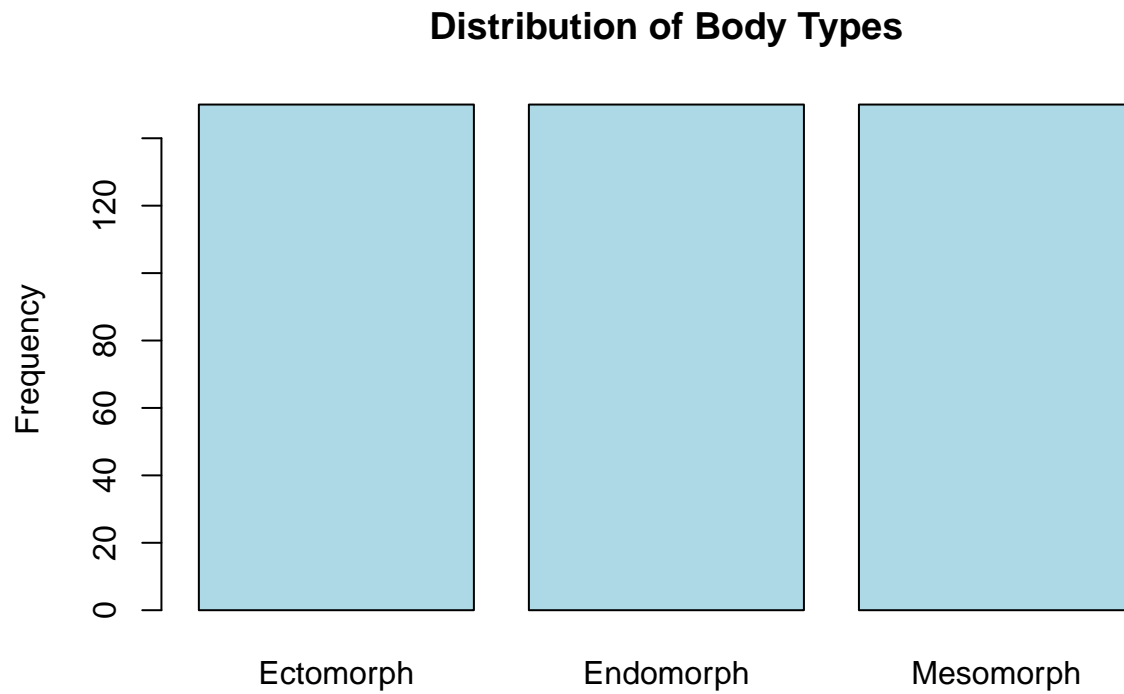
```
# Find the number of missing values per column
colSums(is.na(df))
```

```
##  Q1  Q2  Q3  Q4  Q5  Q6  Q7  Q8  Q9 Q10 Q11 Q12 Q13 Q14 Q15
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

```
# Frequency of each body type
table(dfraw$body_type)
```

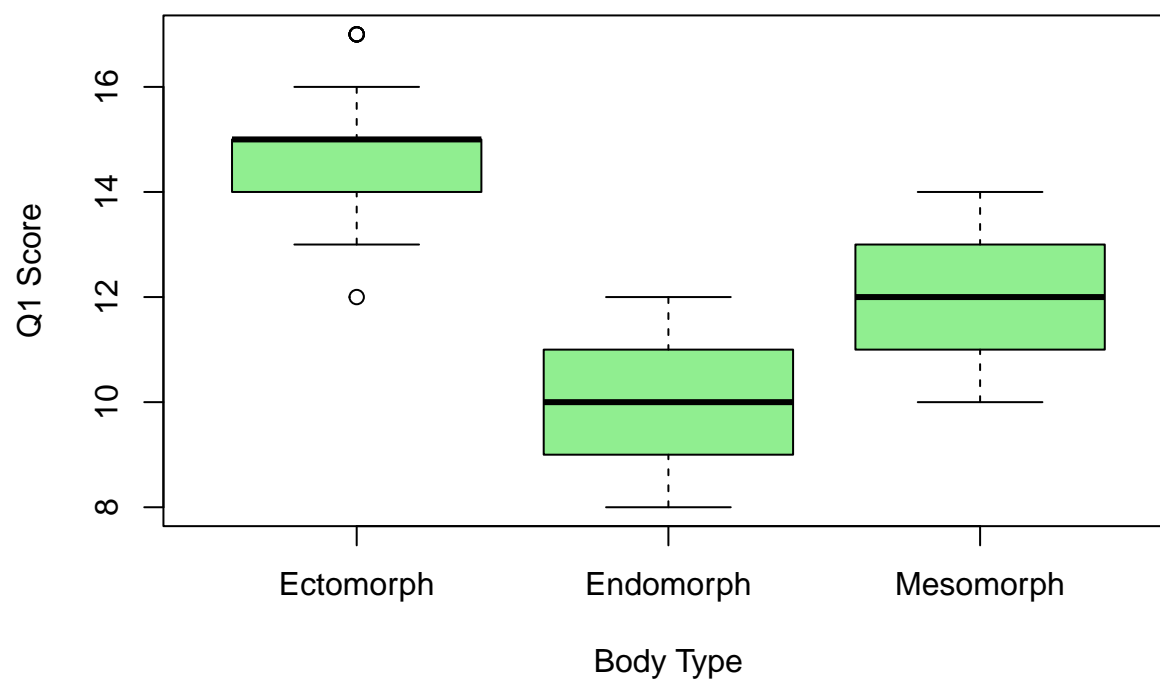
```
##
## Ectomorph Endomorph Mesomorph
##      150      150      150
```

```
# Bar plot of body types
barplot(table(dfraw$body_type), main = "Distribution of Body Types", col = "lightblue", ylab = "Frequency")
```



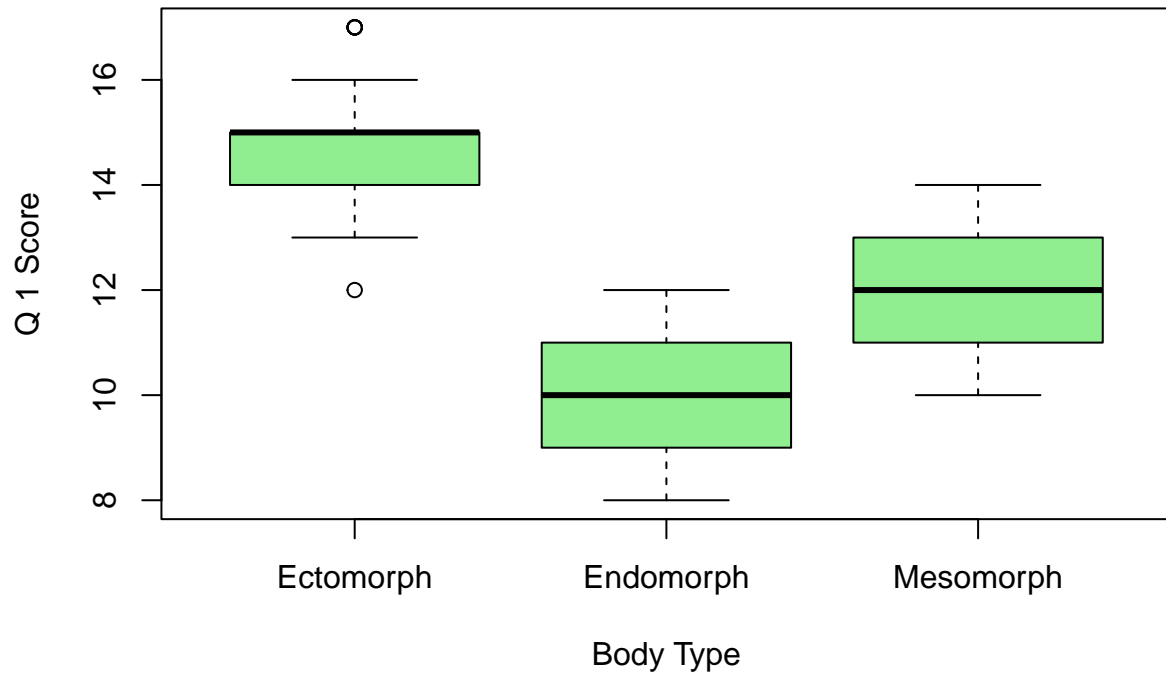
```
# Create boxplots for the first question by body type (Q1 as an example)
boxplot(dfraw$Q1 ~ dfraw$body_type, main = "Responses to Q1 by Body Type",
        xlab = "Body Type", ylab = "Q1 Score", col = "lightgreen")
```

## Responses to Q1 by Body Type

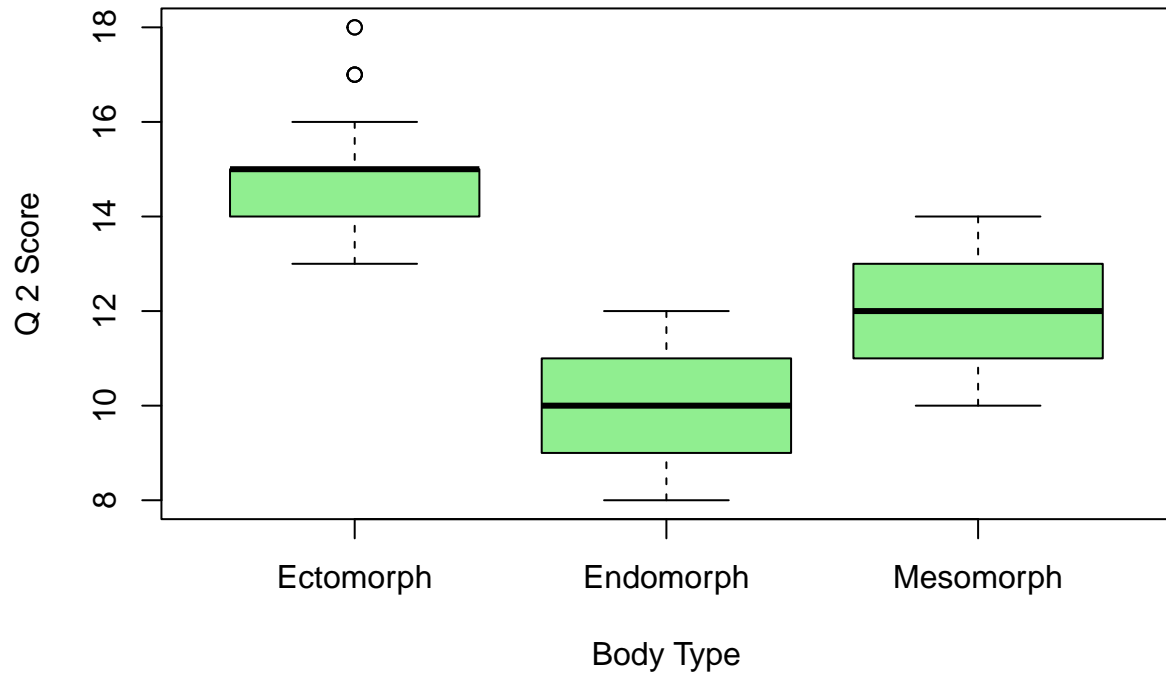


```
for (i in 1:15) {  
  boxplot(dfraw[, i] ~ dfraw$body_type, main = paste("Responses to Q", i, "by Body Type"),  
    xlab = "Body Type", ylab = paste("Q", i, "Score"), col = "lightgreen")  
}
```

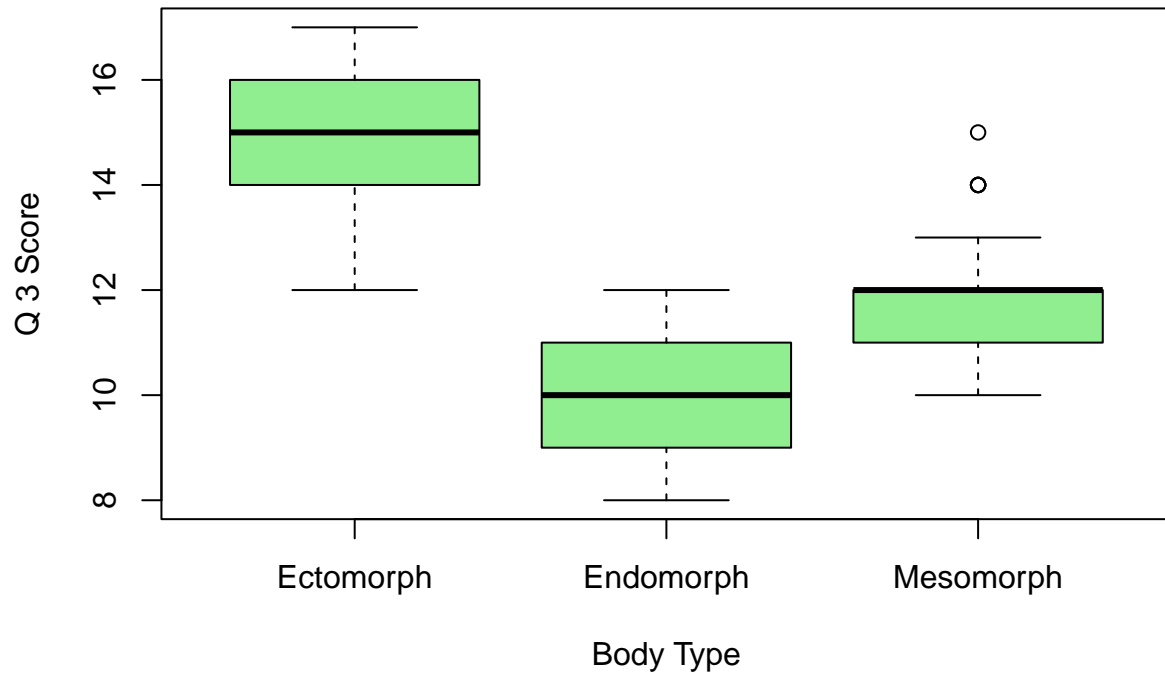
**Responses to Q 1 by Body Type**



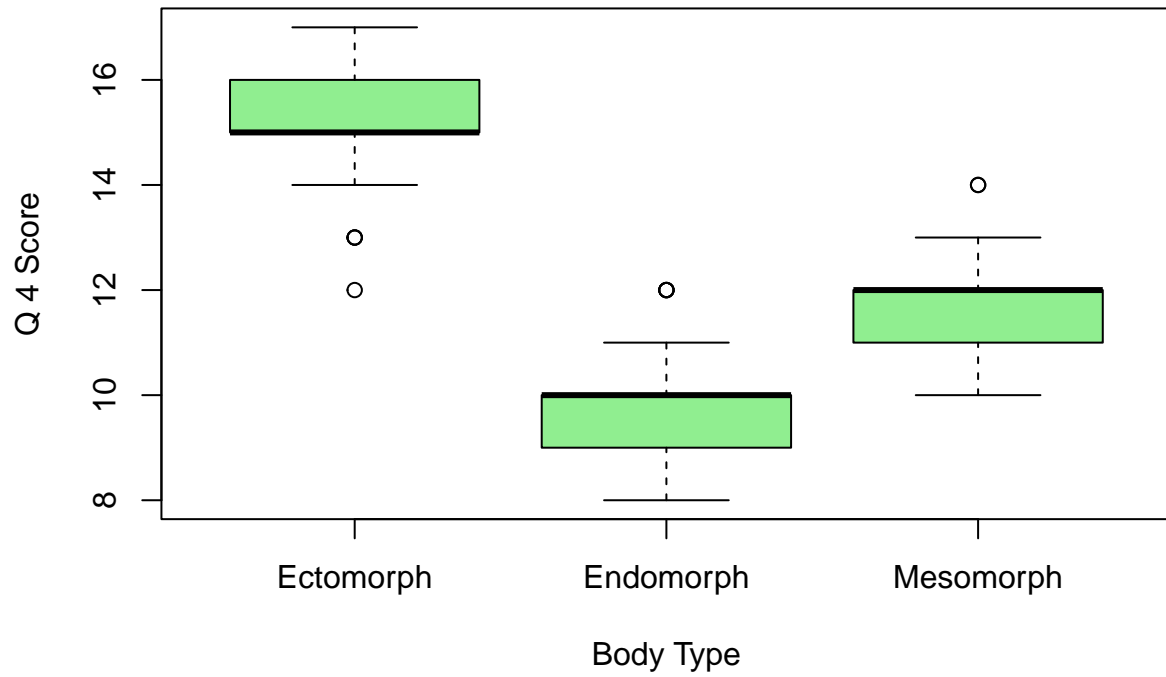
**Responses to Q 2 by Body Type**



**Responses to Q 3 by Body Type**

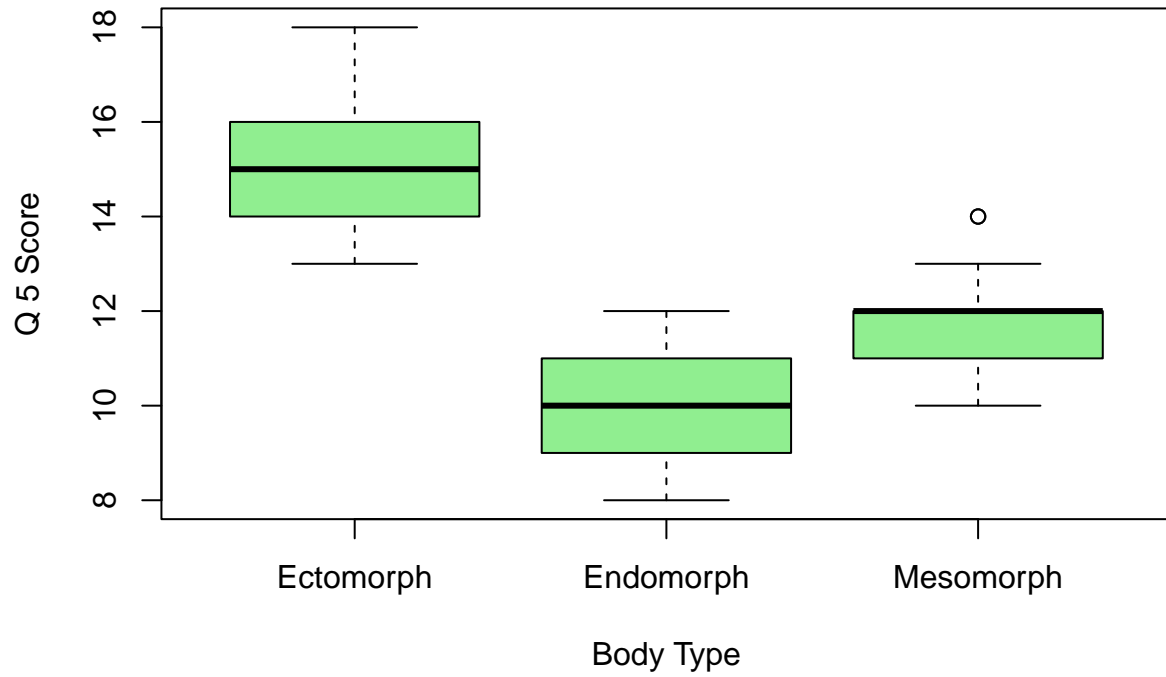


**Responses to Q 4 by Body Type**

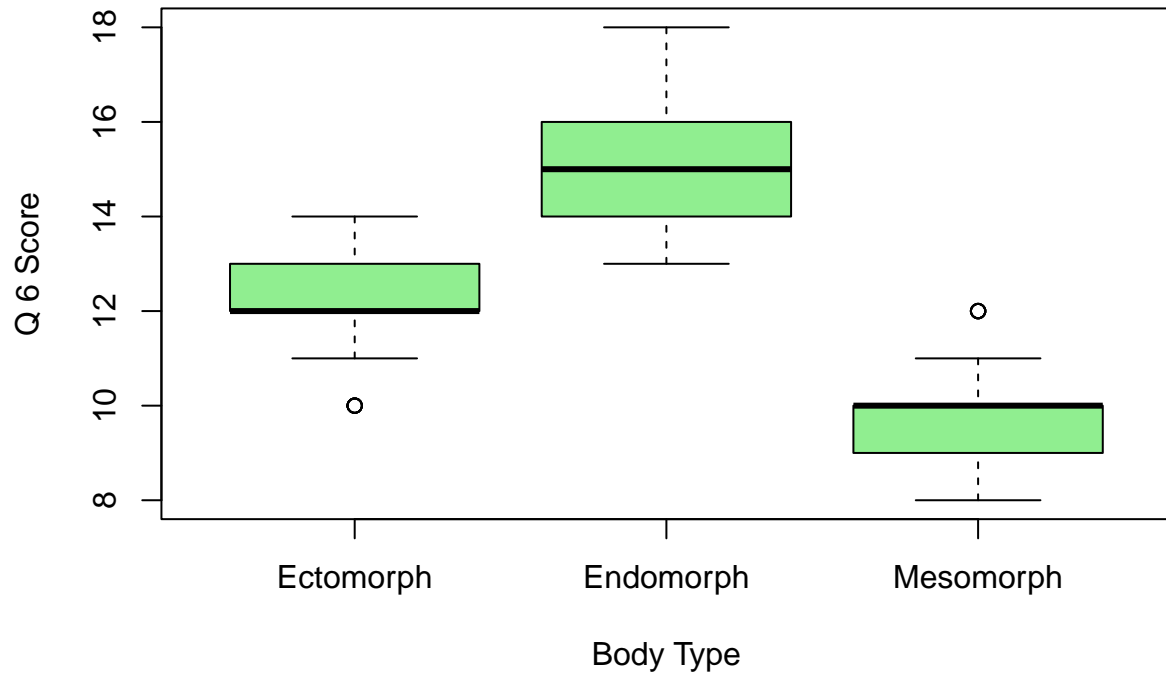




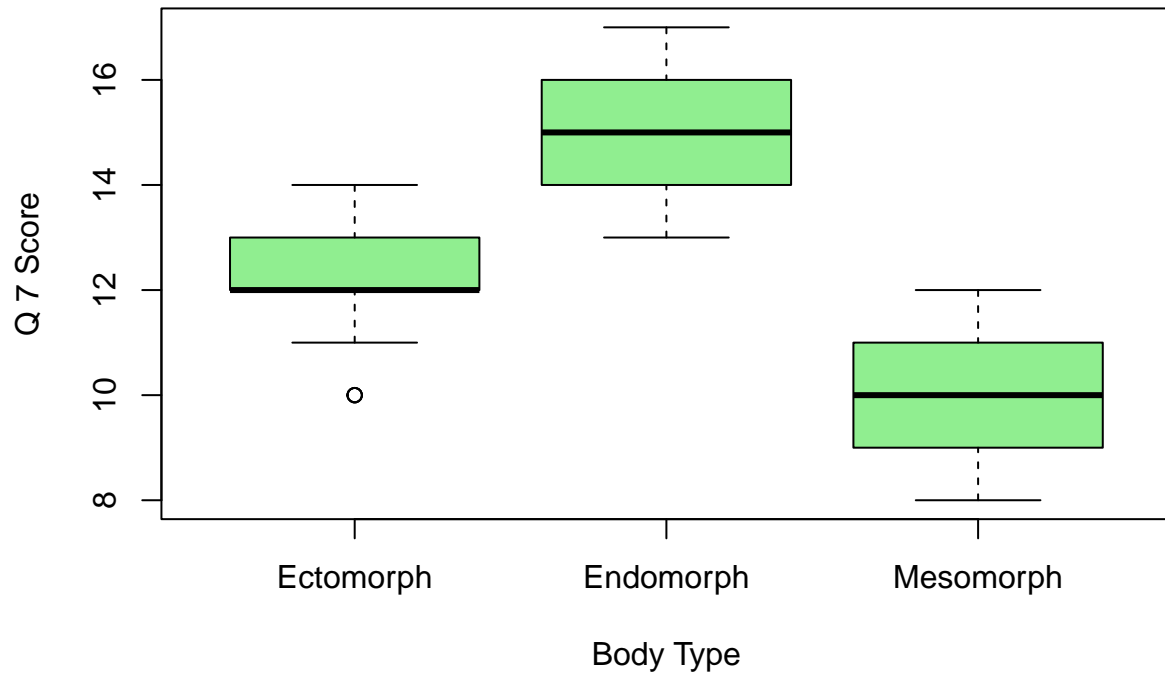
**Responses to Q 5 by Body Type**



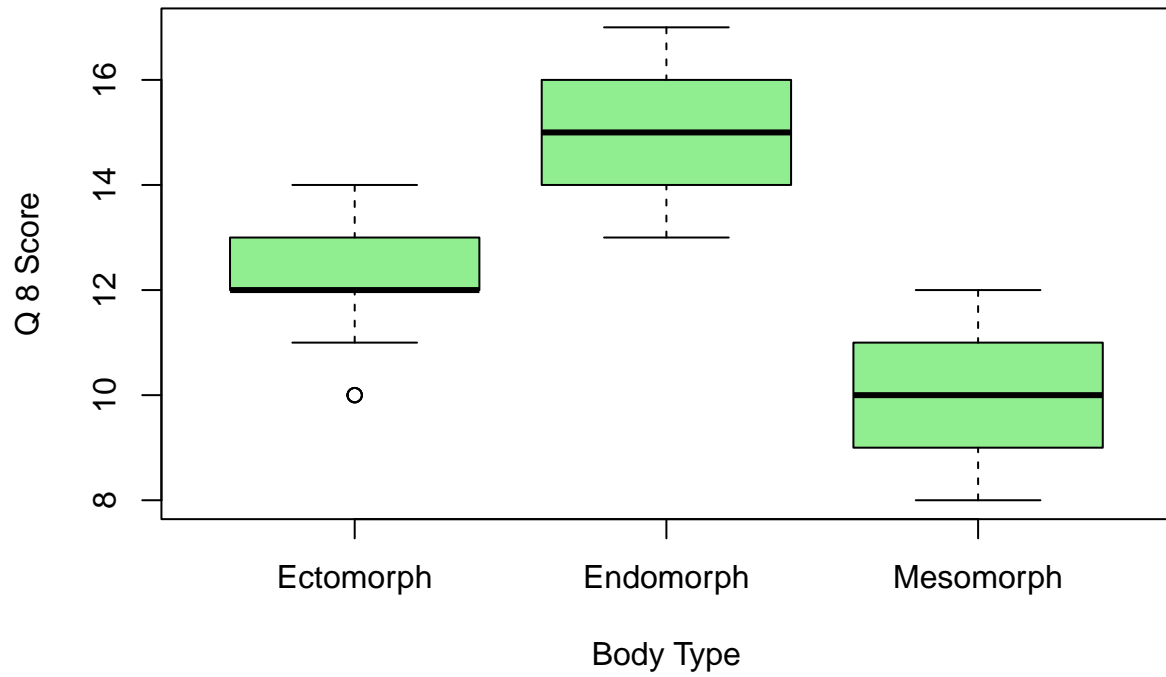
**Responses to Q 6 by Body Type**



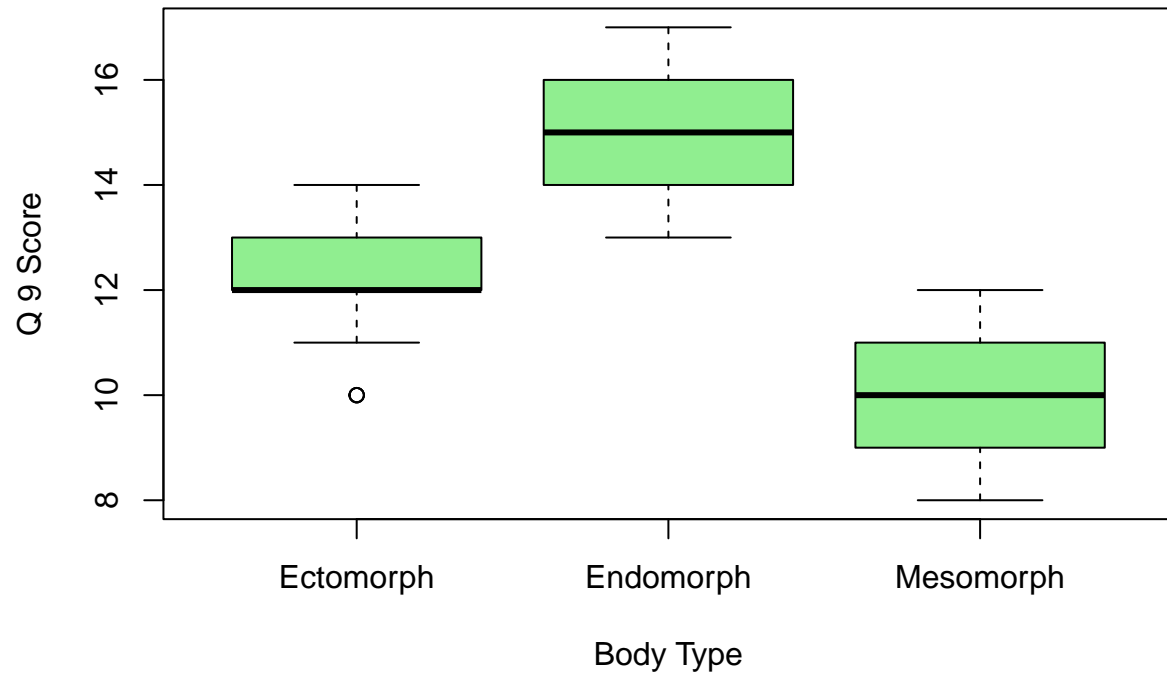
**Responses to Q 7 by Body Type**



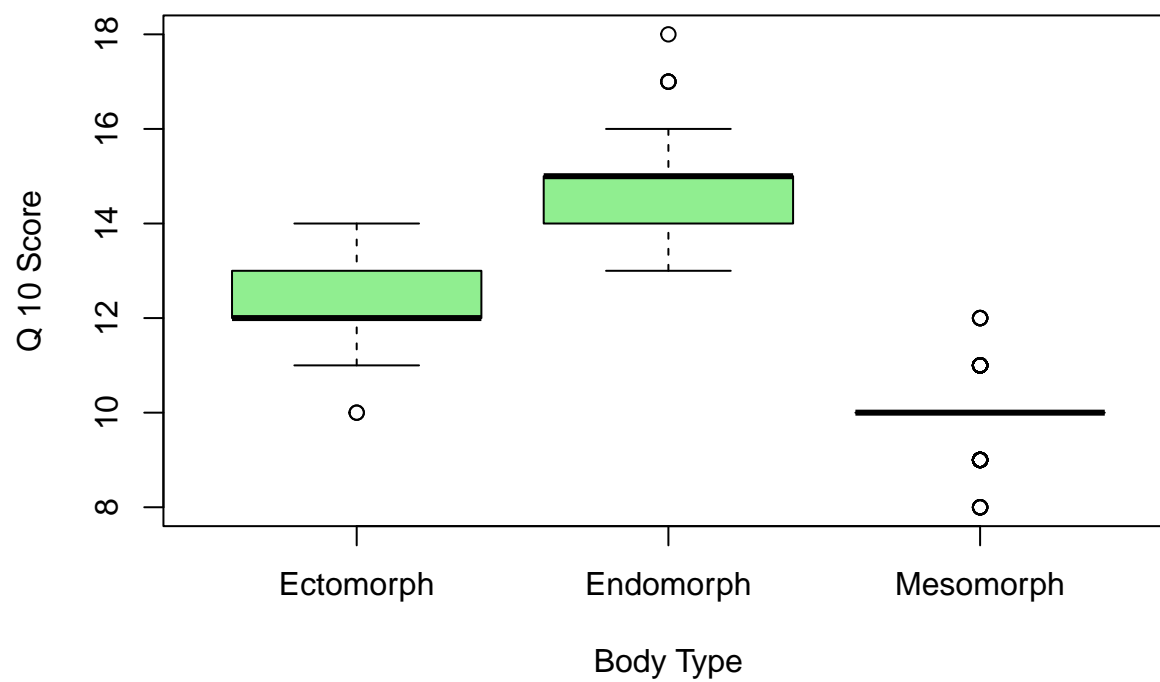
**Responses to Q 8 by Body Type**



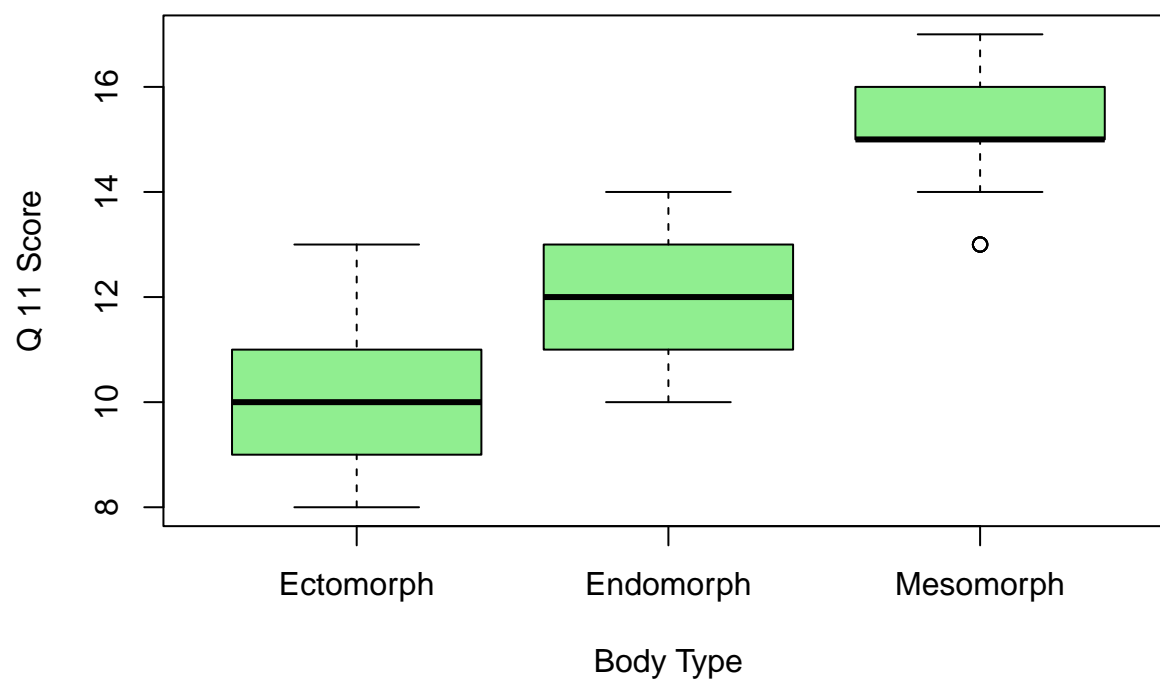
**Responses to Q 9 by Body Type**



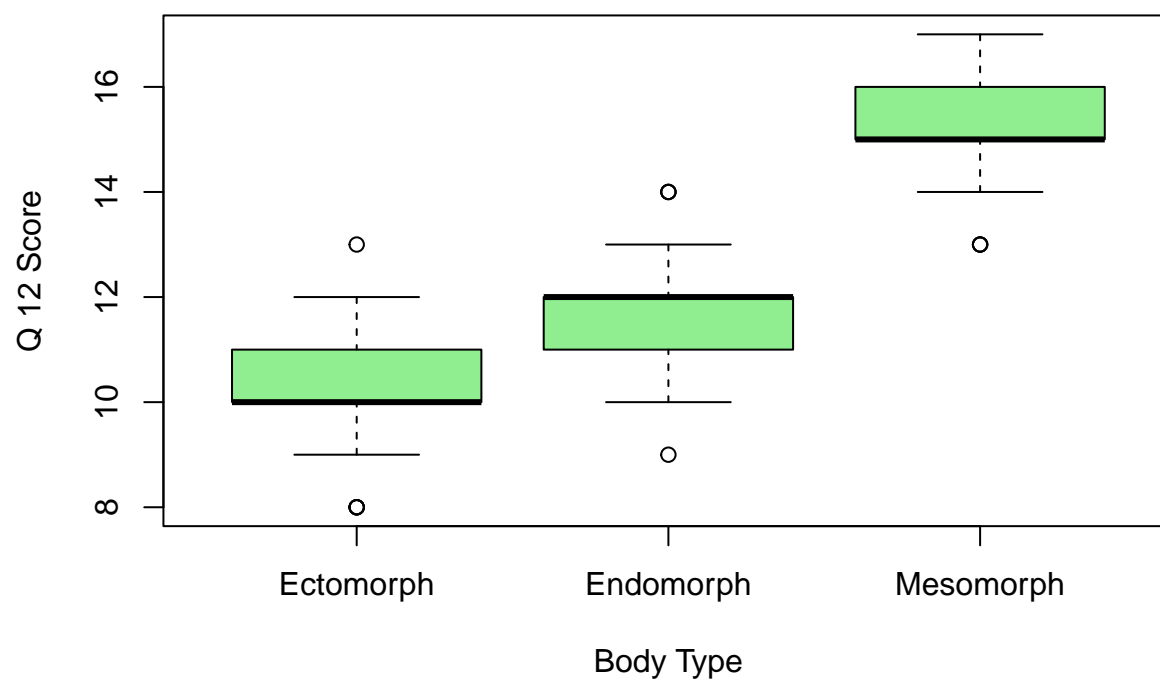
**Responses to Q 10 by Body Type**



**Responses to Q 11 by Body Type**

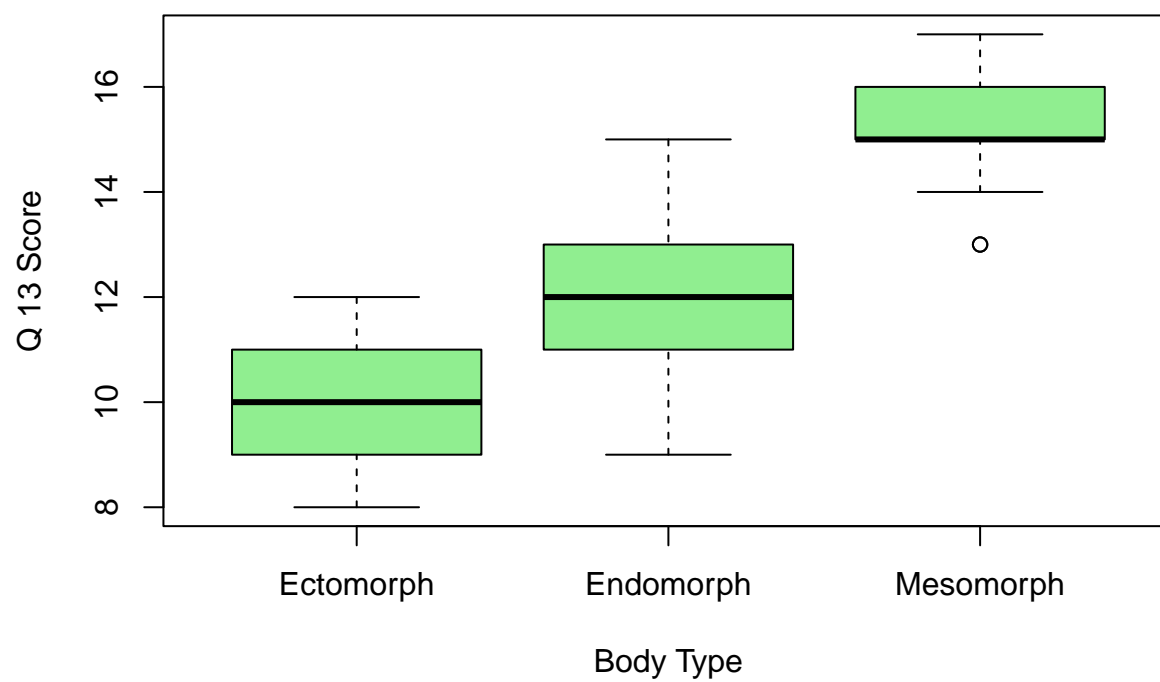


**Responses to Q 12 by Body Type**

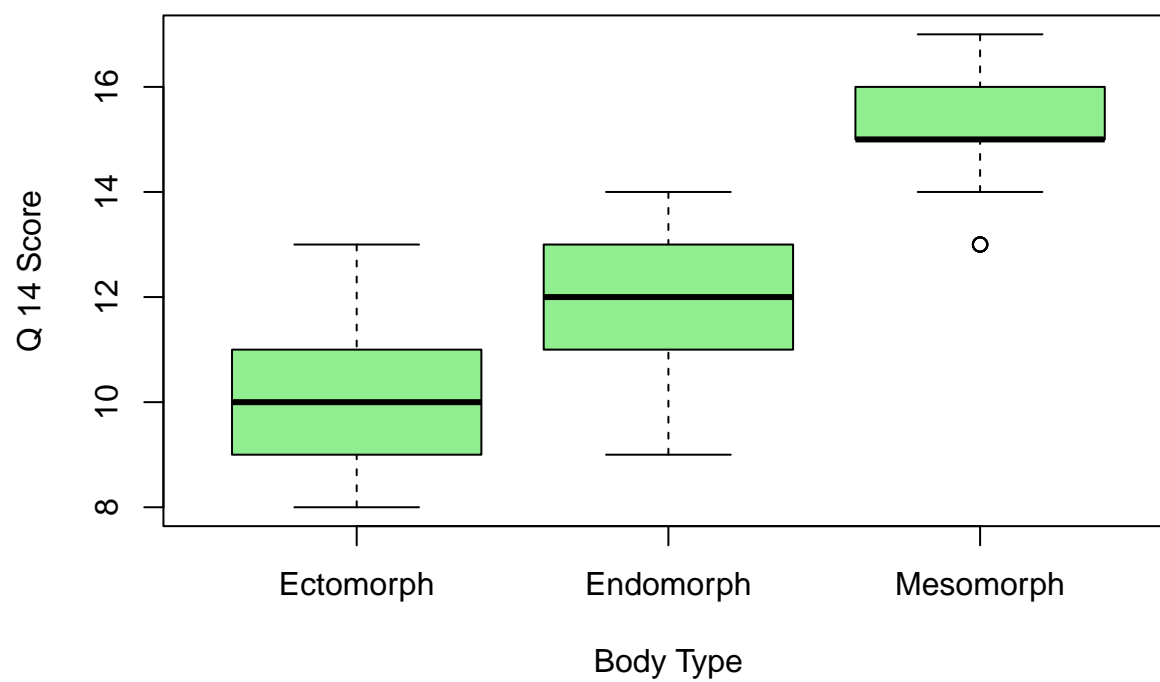




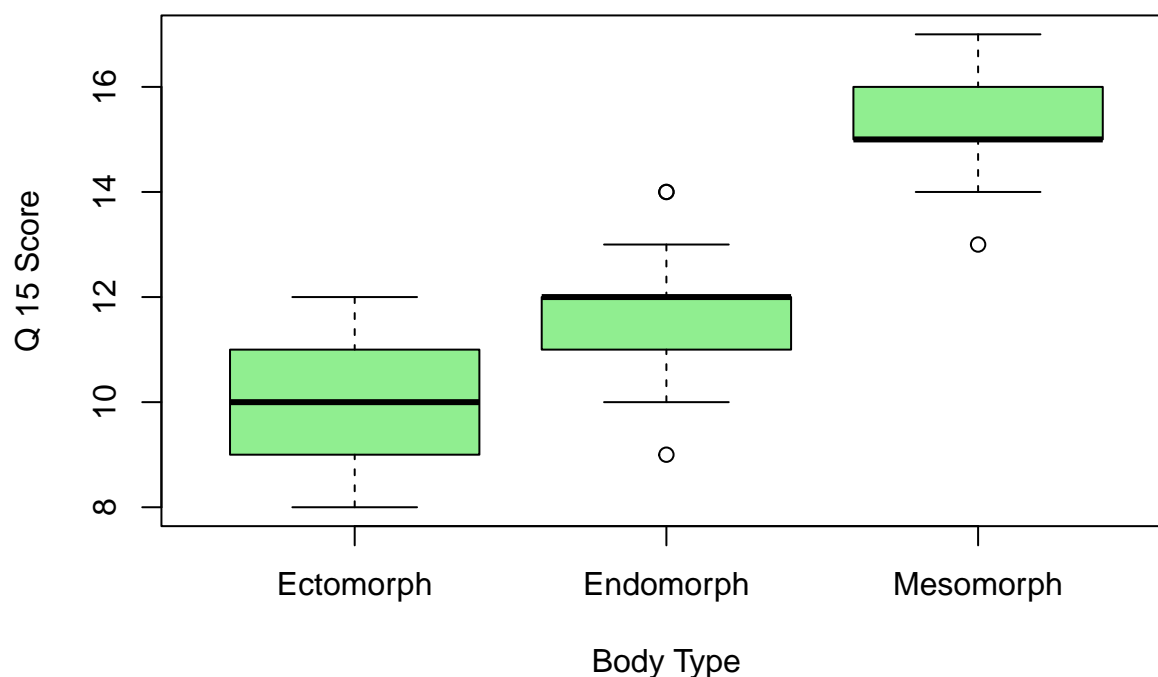
**Responses to Q 13 by Body Type**



**Responses to Q 14 by Body Type**



## Responses to Q 15 by Body Type



```
# Calculate and visualize the correlation matrix
cor_matrix <- cor(df, use = "complete.obs")
print(cor_matrix)
```

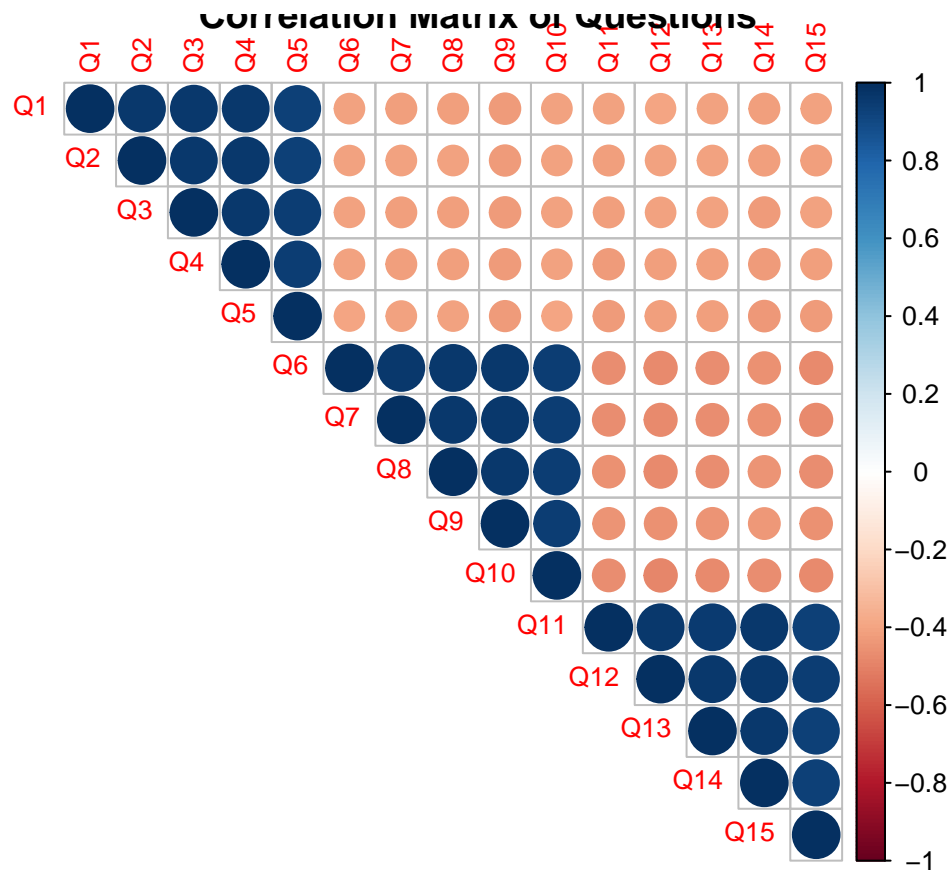
```
##          Q1          Q2          Q3          Q4          Q5          Q6
## Q1  1.0000000  0.9626820  0.9606543  0.9638353  0.9361172 -0.4046188
## Q2  0.9626820  1.0000000  0.9628250  0.9646802  0.9381620 -0.4000133
## Q3  0.9606543  0.9628250  1.0000000  0.9649138  0.9420422 -0.4009673
## Q4  0.9638353  0.9646802  0.9649138  1.0000000  0.9443898 -0.4049952
## Q5  0.9361172  0.9381620  0.9420422  0.9443898  1.0000000 -0.3994513
## Q6 -0.4046188 -0.4000133 -0.4009673 -0.4049952 -0.3994513  1.0000000
## Q7 -0.4138825 -0.4060603 -0.4115136 -0.4120667 -0.4054543  0.9630028
## Q8 -0.4115690 -0.4065140 -0.4119267 -0.4123606 -0.4027596  0.9656202
## Q9 -0.4278577 -0.4237208 -0.4282447 -0.4252717 -0.4203089  0.9631932
## Q10 -0.4057092 -0.4015274 -0.4015220 -0.4005533 -0.3932587  0.9418512
## Q11 -0.4069944 -0.4128256 -0.4135808 -0.4239000 -0.4254353 -0.4645074
## Q12 -0.3975180 -0.4047199 -0.4036354 -0.4151552 -0.4182281 -0.4795217
## Q13 -0.4021549 -0.4044108 -0.4038173 -0.4169008 -0.4112192 -0.4690827
## Q14 -0.4131046 -0.4171617 -0.4211308 -0.4279657 -0.4331026 -0.4542989
## Q15 -0.4012451 -0.4107141 -0.4092613 -0.4191208 -0.4231135 -0.4739785
##          Q7          Q8          Q9          Q10          Q11          Q12
## Q1  -0.4138825 -0.4115690 -0.4278577 -0.4057092 -0.4069944 -0.3975180
## Q2  -0.4060603 -0.4065140 -0.4237208 -0.4015274 -0.4128256 -0.4047199
## Q3  -0.4115136 -0.4119267 -0.4282447 -0.4015220 -0.4135808 -0.4036354
## Q4  -0.4120667 -0.4123606 -0.4252717 -0.4005533 -0.4239000 -0.4151552
## Q5  -0.4054543 -0.4027596 -0.4203089 -0.3932587 -0.4254353 -0.4182281
```

```
## Q6 0.9630028 0.9656202 0.9631932 0.9418512 -0.4645074 -0.4795217
## Q7 1.0000000 0.9666687 0.9671768 0.9427553 -0.4612617 -0.4748120
## Q8 0.9666687 1.0000000 0.9668347 0.9443801 -0.4585041 -0.4759503
## Q9 0.9671768 0.9668347 1.0000000 0.9433006 -0.4432835 -0.4591134
## Q10 0.9427553 0.9443801 0.9433006 1.0000000 -0.4641952 -0.4830867
## Q11 -0.4612617 -0.4585041 -0.4432835 -0.4641952 1.0000000 0.9646631
## Q12 -0.4748120 -0.4759503 -0.4591134 -0.4830867 0.9646631 1.0000000
## Q13 -0.4624290 -0.4609844 -0.4463451 -0.4706861 0.9591575 0.9644050
## Q14 -0.4552115 -0.4493657 -0.4378962 -0.4659248 0.9634438 0.9643319
## Q15 -0.4730040 -0.4665481 -0.4566030 -0.4768678 0.9389646 0.9425124
##      Q13      Q14      Q15
## Q1 -0.4021549 -0.4131046 -0.4012451
## Q2 -0.4044108 -0.4171617 -0.4107141
## Q3 -0.4038173 -0.4211308 -0.4092613
## Q4 -0.4169008 -0.4279657 -0.4191208
## Q5 -0.4112192 -0.4331026 -0.4231135
## Q6 -0.4690827 -0.4542989 -0.4739785
## Q7 -0.4624290 -0.4552115 -0.4730040
## Q8 -0.4609844 -0.4493657 -0.4665481
## Q9 -0.4463451 -0.4378962 -0.4566030
## Q10 -0.4706861 -0.4659248 -0.4768678
## Q11 0.9591575 0.9634438 0.9389646
## Q12 0.9644050 0.9643319 0.9425124
## Q13 1.0000000 0.9637051 0.9389123
## Q14 0.9637051 1.0000000 0.9385831
## Q15 0.9389123 0.9385831 1.0000000
```

```
# Visualize the correlation matrix
library(corrplot)
```

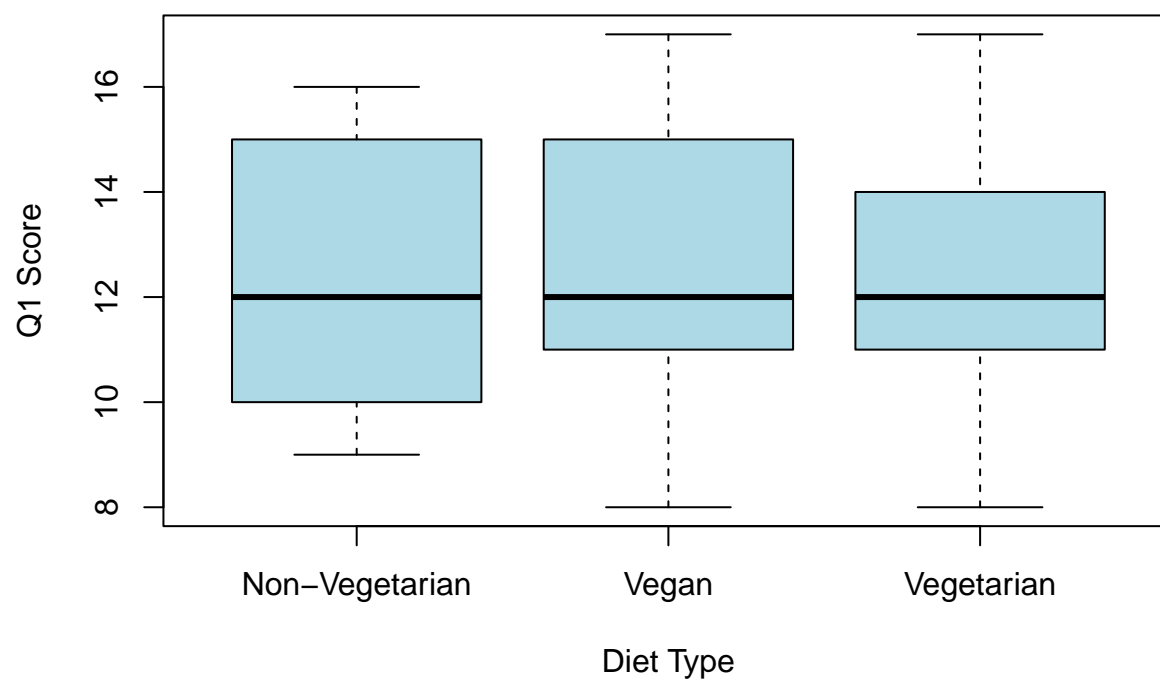
```
## corrplot 0.94 loaded
```

```
corrplot(cor_matrix, method = "circle", type = "upper", tl.cex = 0.8, title = "Correlation Matrix of Qu
```



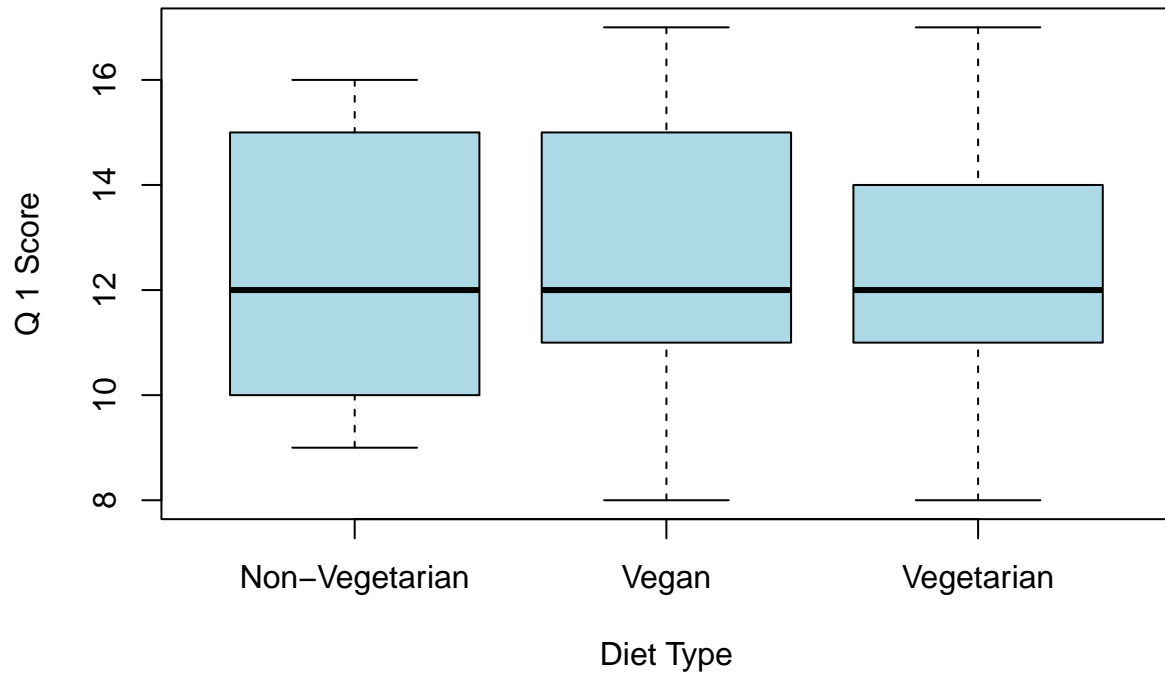
```
# Create boxplots for the first question by diet type
boxplot(dfraw$Q1 ~ dfraw$diet, main = "Responses to Q1 by Diet Type",
        xlab = "Diet Type", ylab = "Q1 Score", col = "lightblue")
```

## Responses to Q1 by Diet Type

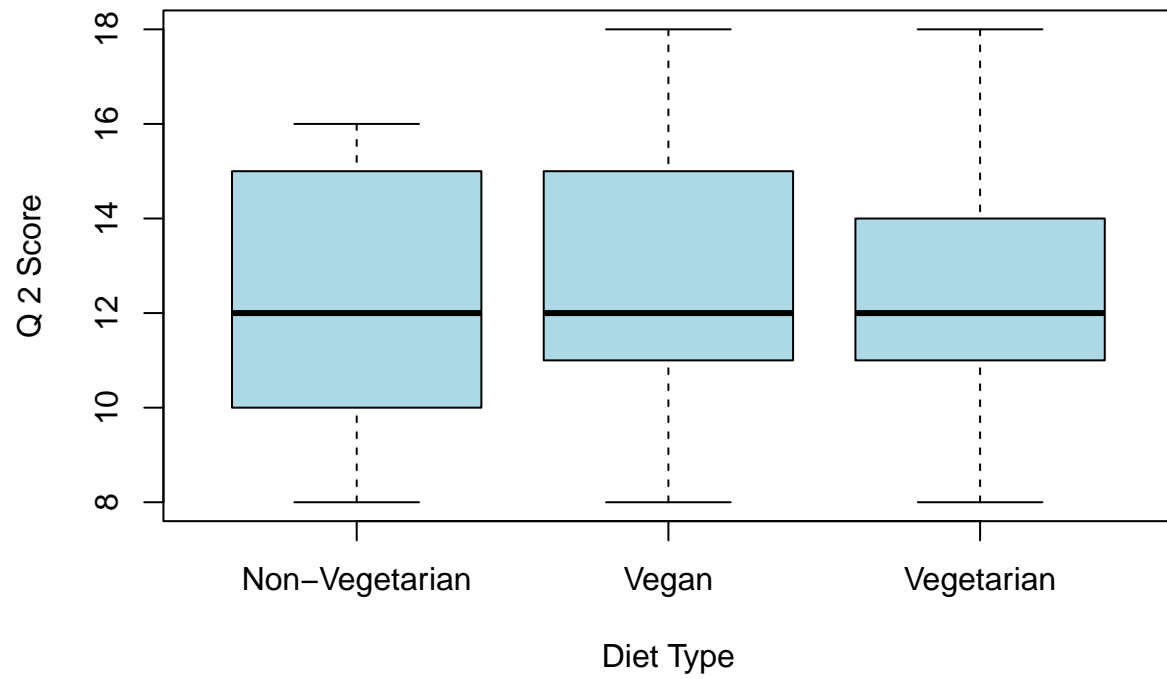


```
# Loop through all questions for boxplots by diet type
for (i in 1:15) {
  boxplot(dfraw[, i] ~ dfraw$diet, main = paste("Responses to Q", i, "by Diet Type"),
    xlab = "Diet Type", ylab = paste("Q", i, "Score"), col = "lightblue")
}
```

**Responses to Q 1 by Diet Type**

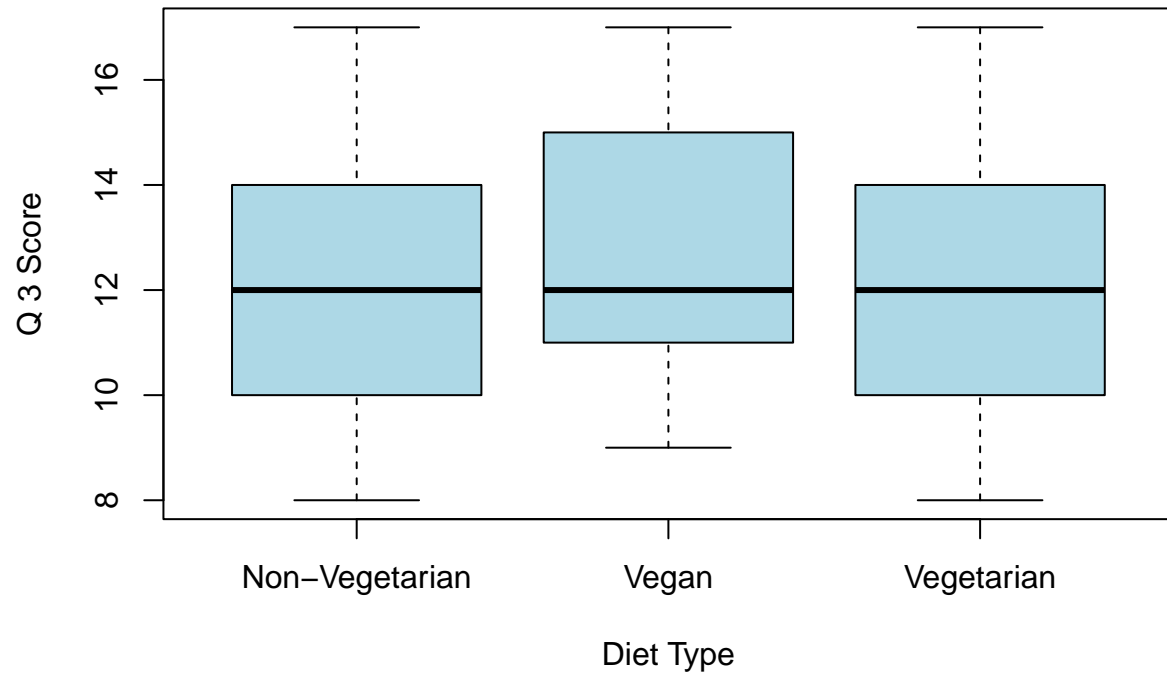


**Responses to Q 2 by Diet Type**

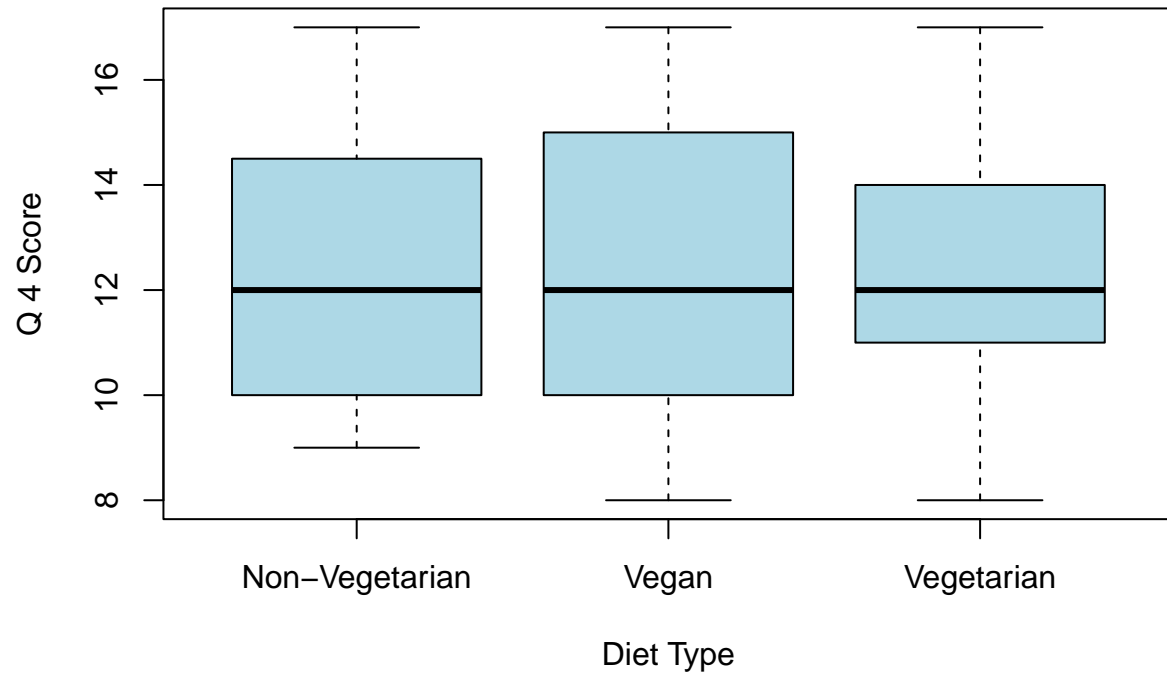




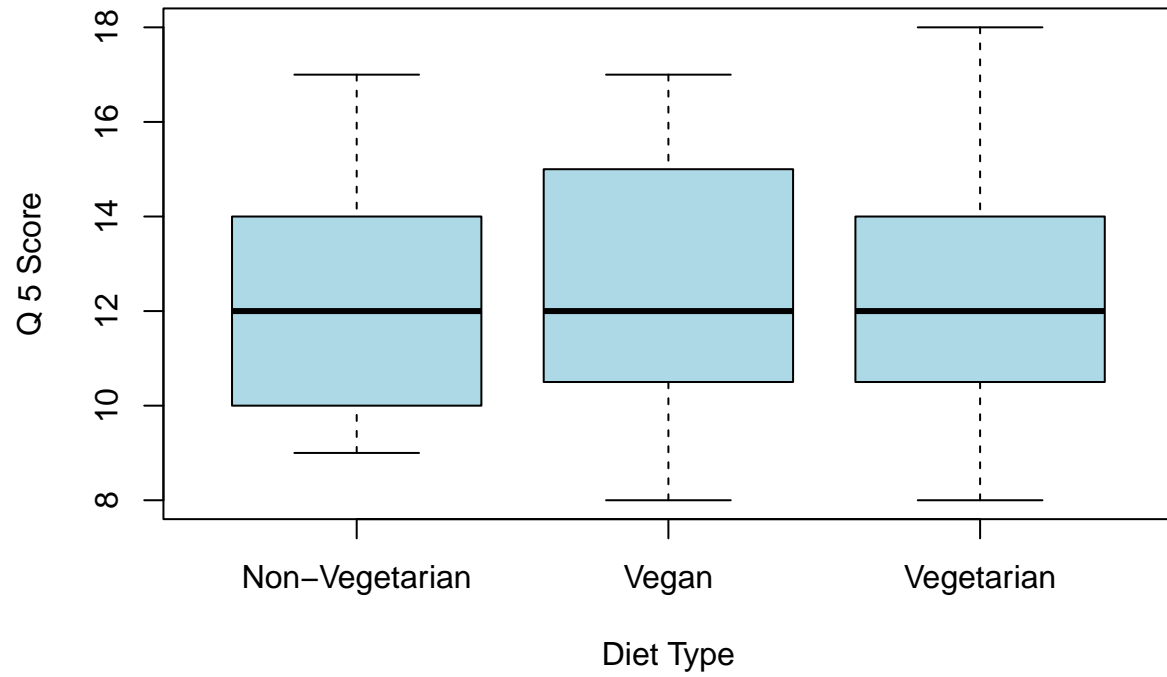
**Responses to Q 3 by Diet Type**



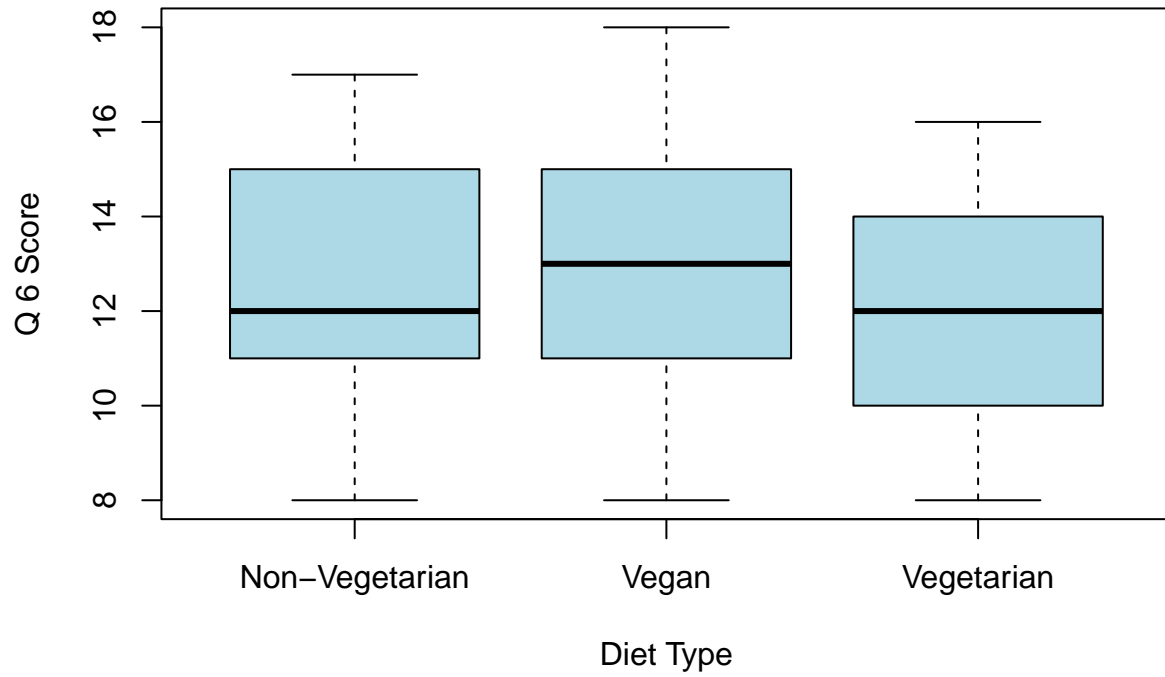
**Responses to Q 4 by Diet Type**



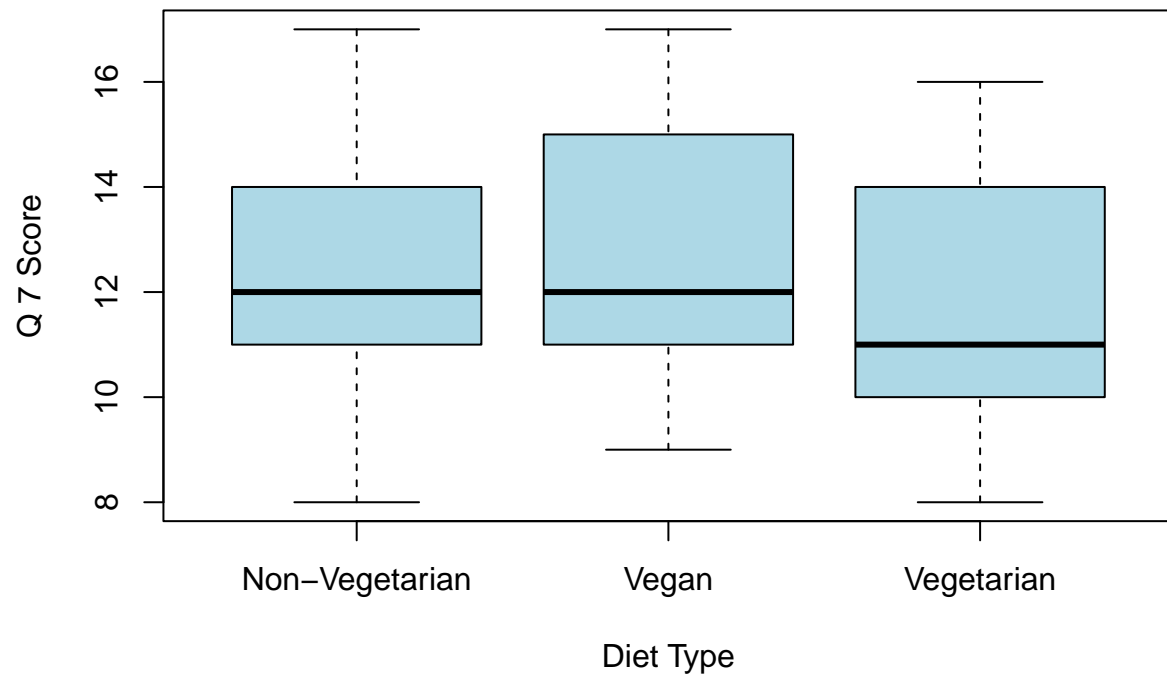
**Responses to Q 5 by Diet Type**



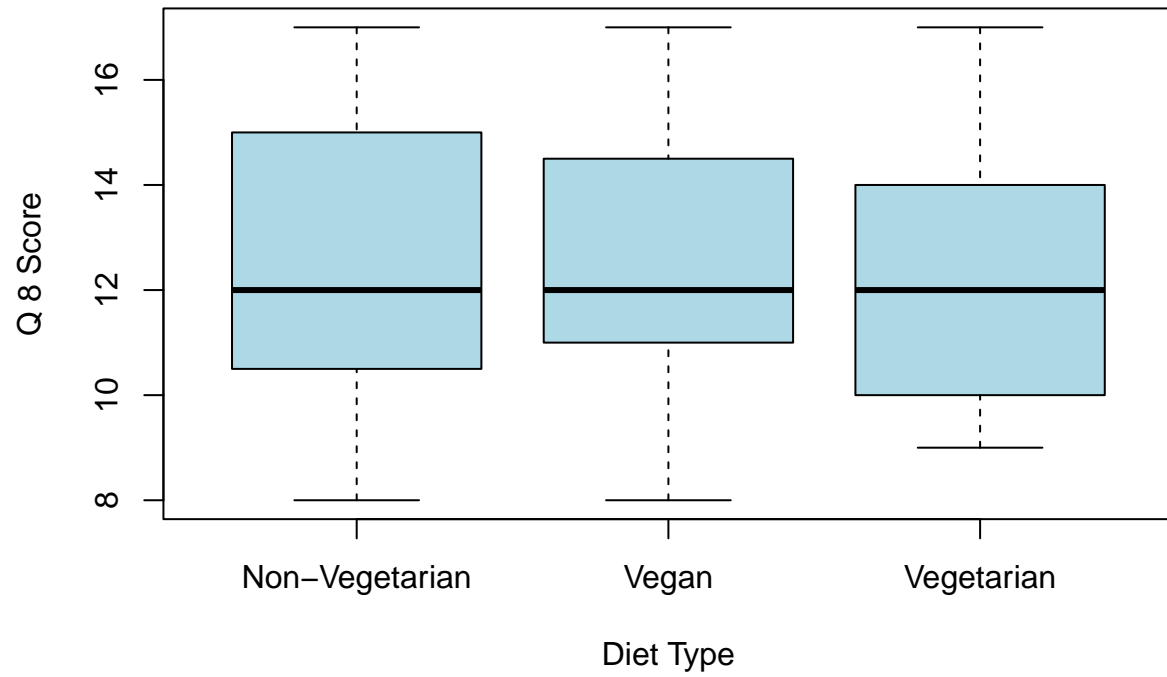
**Responses to Q 6 by Diet Type**



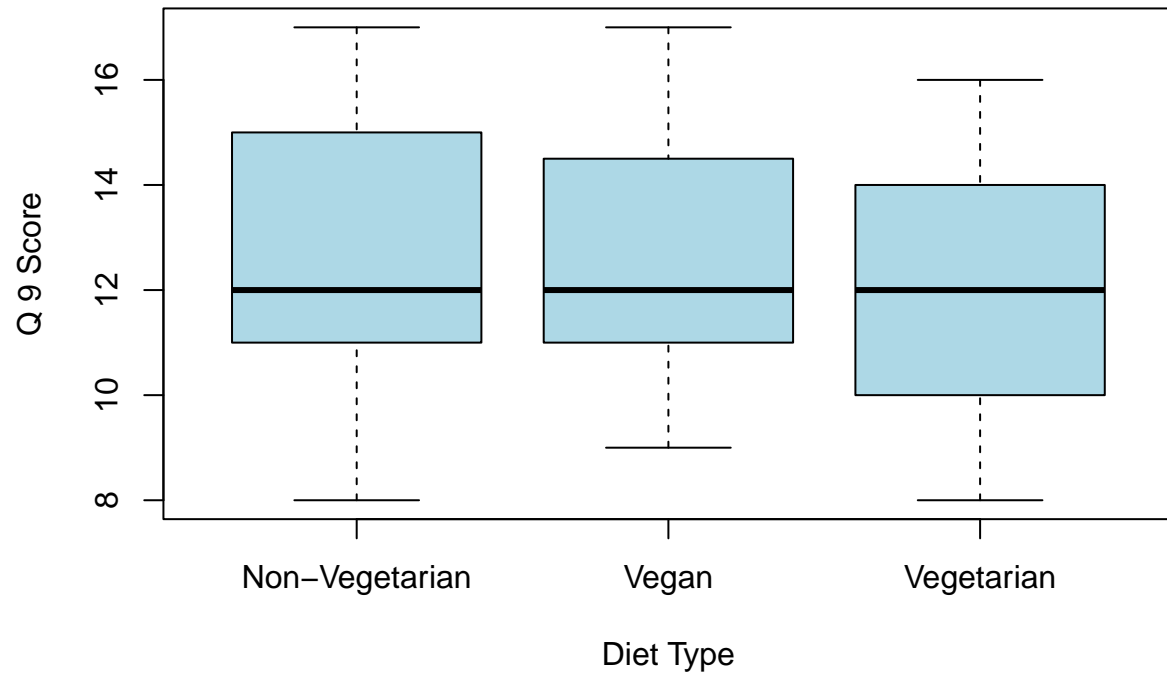
**Responses to Q 7 by Diet Type**



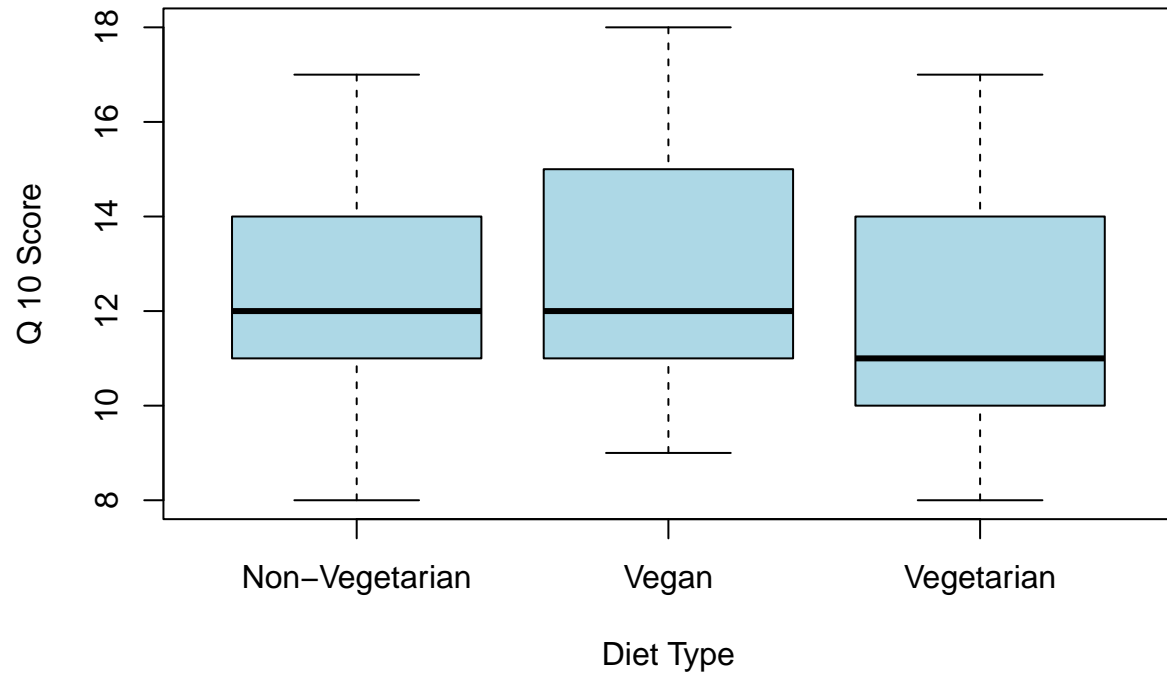
**Responses to Q 8 by Diet Type**



**Responses to Q 9 by Diet Type**

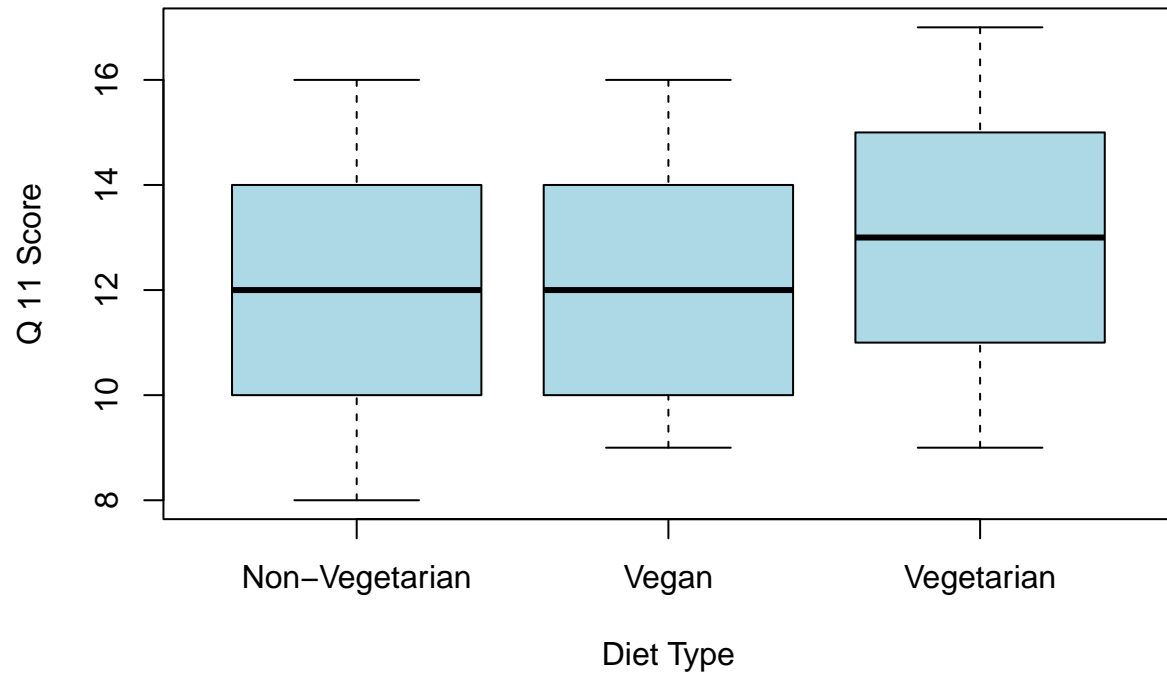


**Responses to Q 10 by Diet Type**

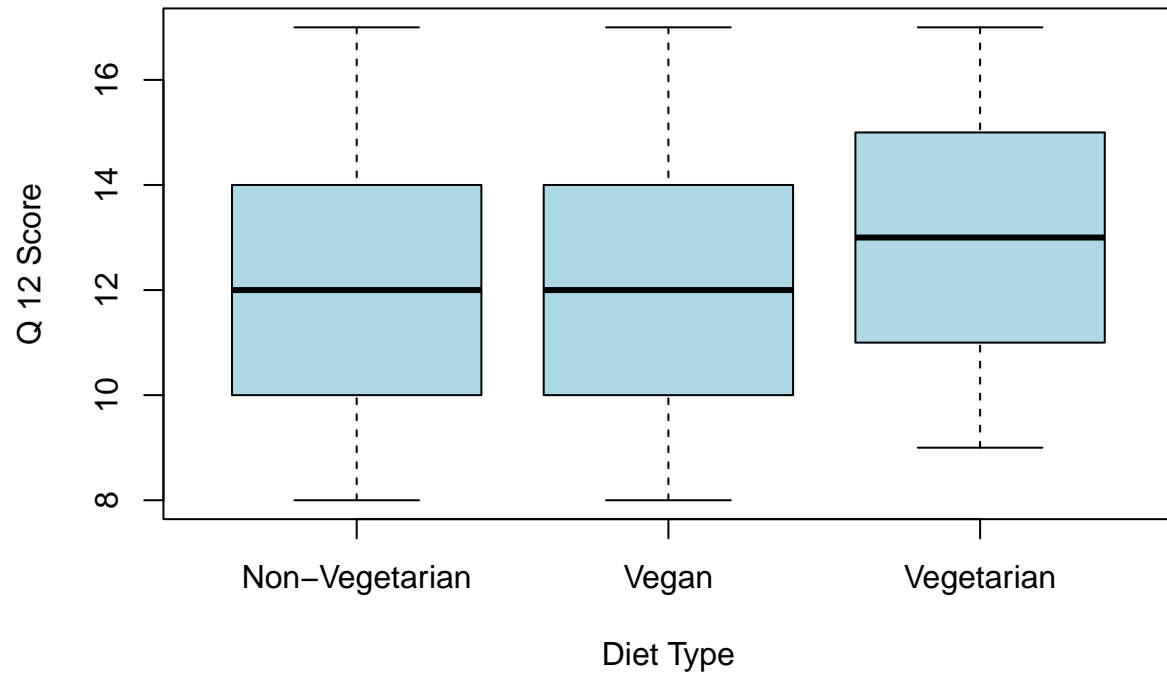




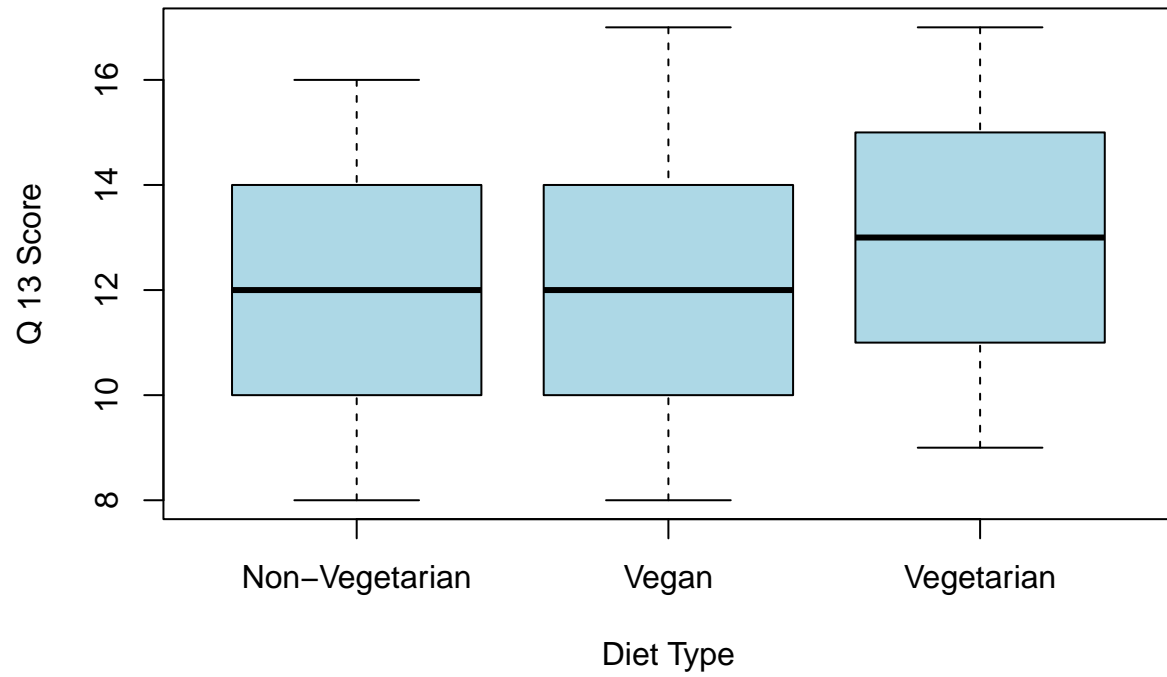
**Responses to Q 11 by Diet Type**



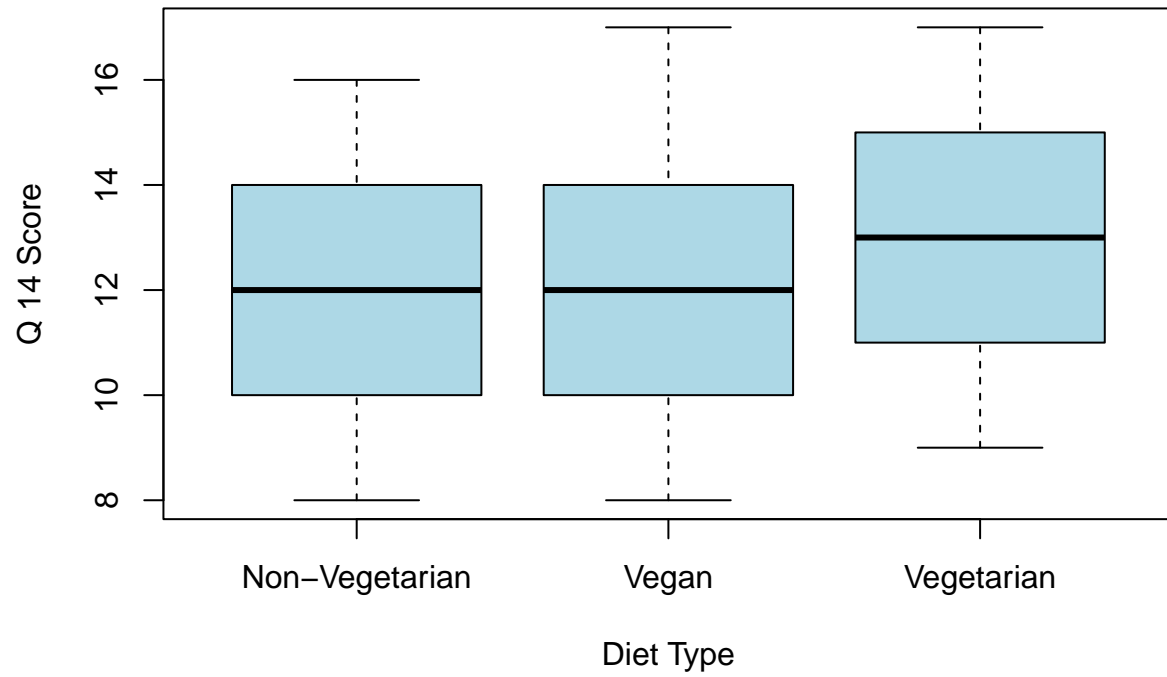
**Responses to Q 12 by Diet Type**



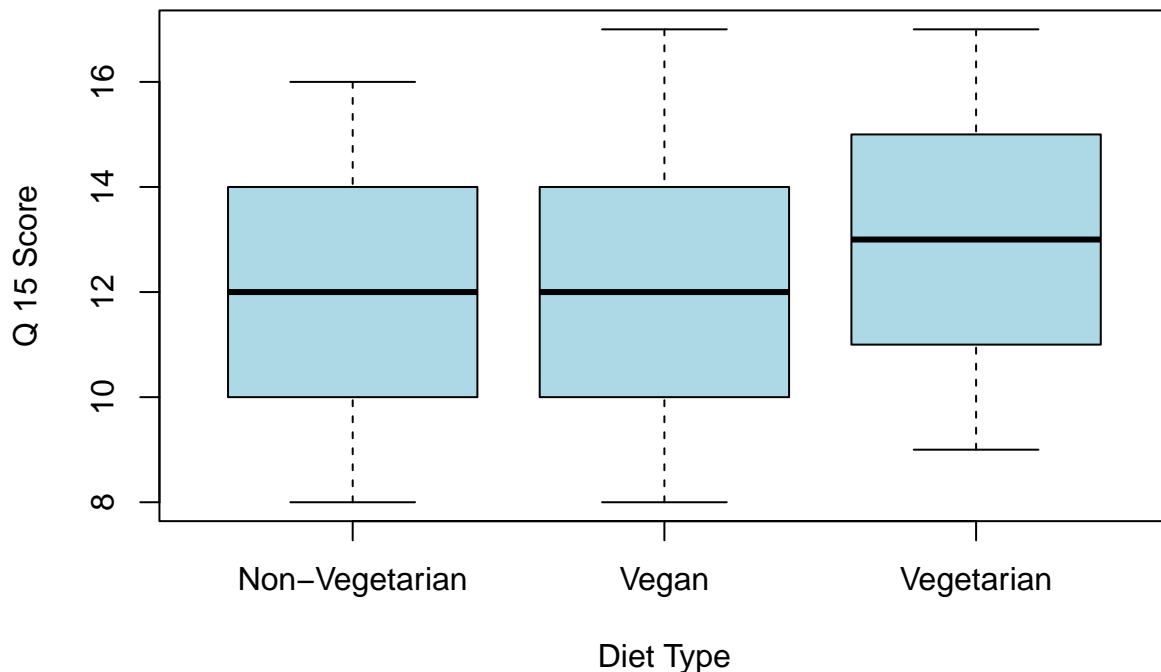
**Responses to Q 13 by Diet Type**



**Responses to Q 14 by Diet Type**



## Responses to Q 15 by Diet Type



### Conclusion about EDA:

- The data is clean and has no outlier
- The differences in diet type based in the answer of 15 questions are not large
- The differences in body type based in the answer are noticeable:
- Ectomorph body type answered highly in average for the first five questions, meaning in average, they find themselves are physically fit enough
- Endomorph body type answered highly in average for the question 6 - 10, meaning in average, they find themselves are healthy mentally.
- Mesomorph body type answered highly in average for the question 11 - 15, meaning in average, they find meaning connections and socially accepted.

## 2. Using PCA analysis:

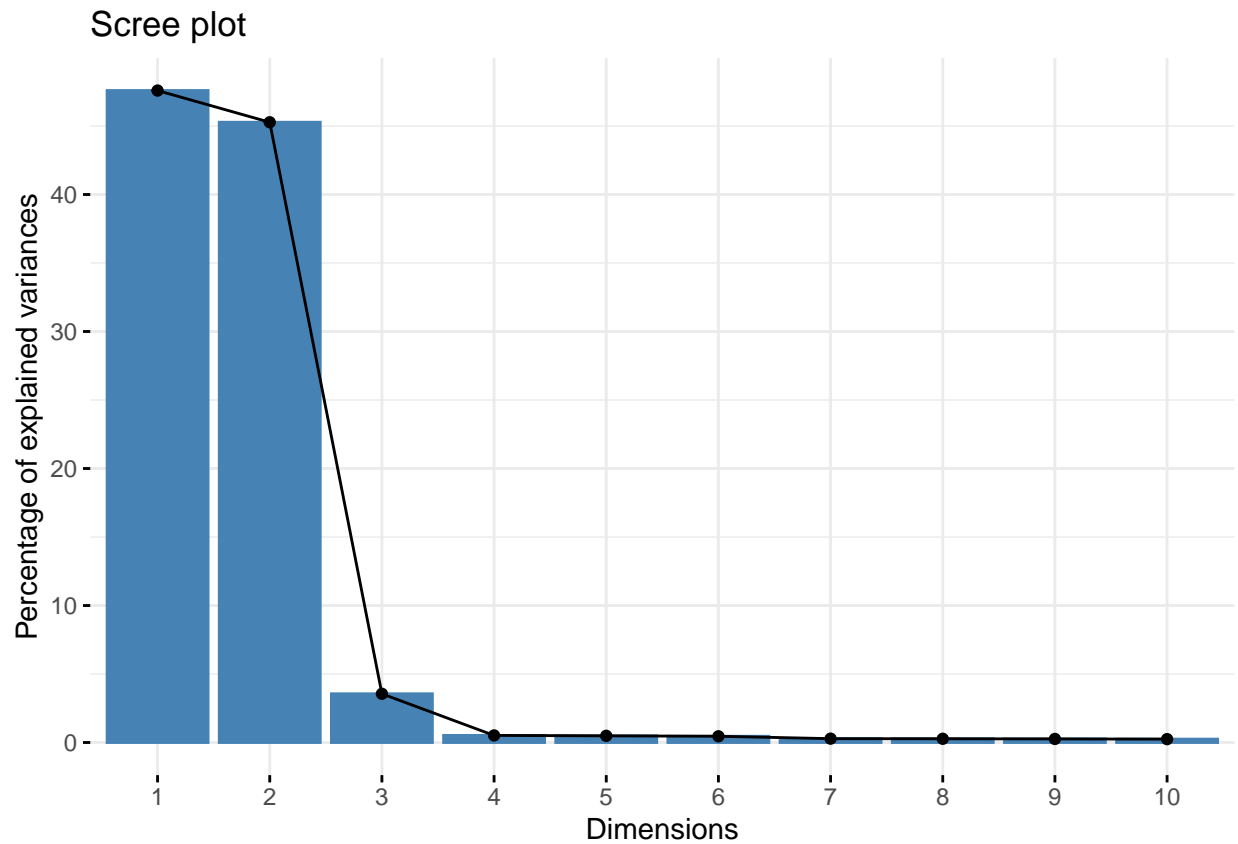
```
pca_result <- prcomp(df, scale. = TRUE)
```

```
library("factoextra")
```

```
## Loading required package: ggplot2
```

## Welcome! Want to learn more? See two factoextra-related books at <https://goo.gl/ve3WBa>

```
fviz_eig(pca_result)
```



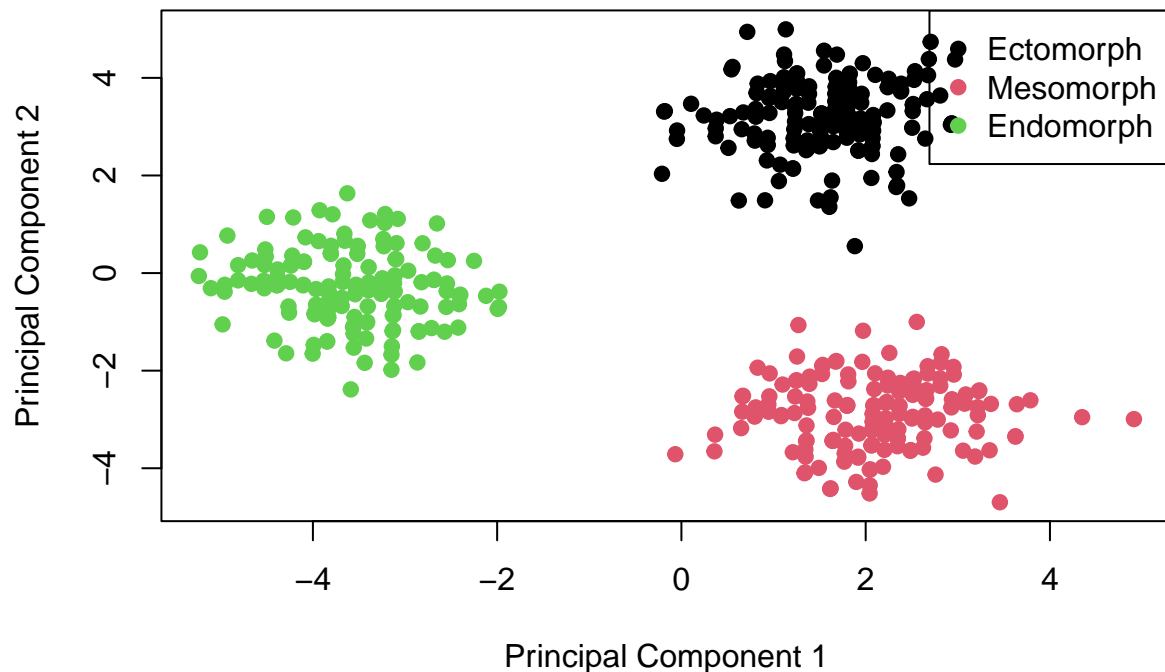
*Conclusion:* Keeping 2 Principal Components

```
body_types <- as.factor(dfraw$body_type)
unique_body_types <- unique(body_types)
colors <- 1:length(unique_body_types)

plot(pca_result$x[, 1:2], col = colors[body_types], pch = 19,
     xlab = "Principal Component 1", ylab = "Principal Component 2",
     main = "PCA of Body Types based on 15 Questions")

legend("topright", legend = unique_body_types, col = colors, pch = 19)
```

## PCA of Body Types based on 15 Questions



*Conclusion:* Based on the PC Mapping, we can say that each body types are very unique to each other, and very difficult to group them. If we isolate each component, we can explain like this:

- If we consider mainly Principal Component 1, Ectomorph and Mesomorph body types are more similar to Endomorph
- If we consider mainly Principal Component 2, we have three body types are equally different.

As a result, Ectomorph and Mesomorph body types are more similar to Endomorph

### 3. Comprehensive Factor Analysis

#### Checking data set

#### Correlation Test and sample adequacy test

$H_0$ : There are no significant correlation (Correlation matrix = identity matrix)  $H_a$ : There are significant correlations.

```
library(psych)
```

```
##  
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':  
##  
##    %+%, alpha
```

```
library(GPArotation)
```

```
##  
## Attaching package: 'GPArotation'
```

```
## The following objects are masked from 'package:psych':  
##  
##    equamax, varimin
```

```
correlations = cor(df)  
cortest.bartlett(correlations, n = nrow(df))
```

```
## $chisq  
## [1] 15399.6  
##  
## $p.value  
## [1] 0  
##  
## $df  
## [1] 105
```

*Test interpretation:* With p value < 0.05, FA analysis might be useful for this data set.

**Test Sampling adequacy using KMO test and interpret the results.**

$H_0$ : No significant factors in data  $H_a$ : Significant factor.

```
km = KMO(correlations)  
km$MSA
```

```
## [1] 0.9541395
```

**Interpretation:** Overall MSA = 0.74 > 0.7 meaning there are significant factors in the data set

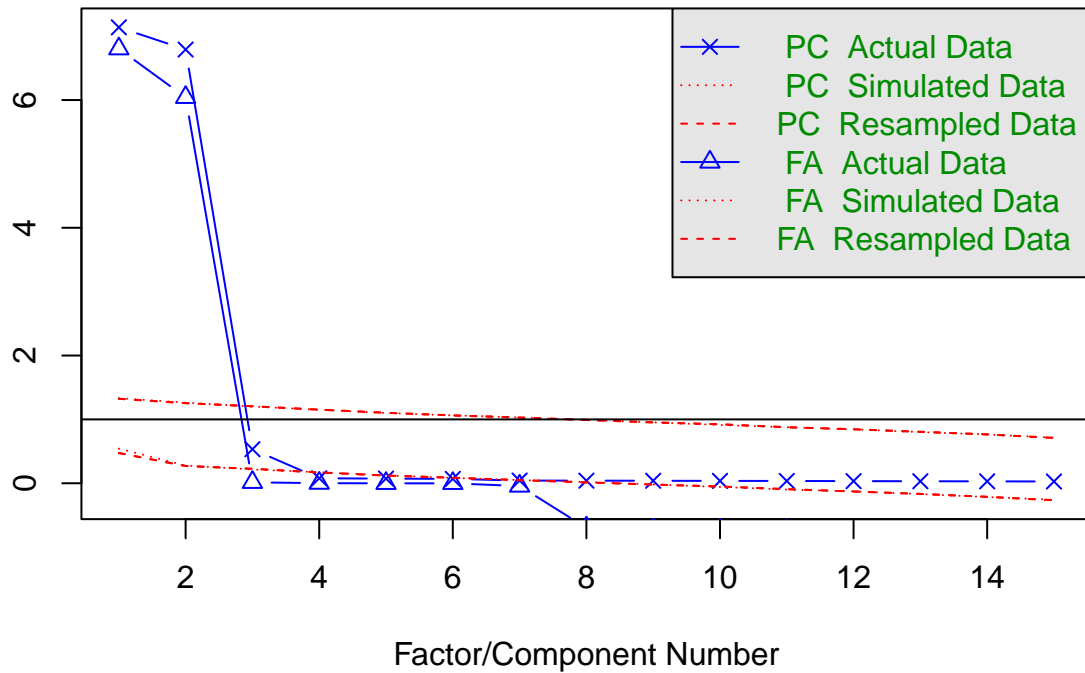
**Number of factors**

```
plot.new()  
nofactors <- fa.parallel(df, fm = "ml")
```



eigenvalues of principal components and factor analysis

## Parallel Analysis Scree Plots



## Parallel analysis suggests that the number of factors = 2 and the number of components = 2

```
nofactors$fa.values
```

```
## [1] 6.802961333 6.041569327 0.014701449 0.001901216 -0.001116841
## [6] -0.001368754 -0.043553237 -0.697969450 -0.733807136 -0.737597222
## [11] -0.747849607 -0.759354826 -0.785517896 -0.790332836 -0.791739711
```

```
sum(nofactors$fa.values > .7) ## kaiser criterion
```

```
## [1] 2
```

**Conclusion:** There would be a total of 2 factors

```
fa_health = fa(df, nfactors=2, rotate = "Varimax", fm="ml")
```

```
fa_health$loadings
```

```
##
## Loadings:
##      ML1      ML2
## Q1 -0.183  0.957
## Q2 -0.176  0.960
```

```
## Q3 -0.180 0.960
## Q4 -0.177 0.965
## Q5 -0.173 0.944
## Q6 0.944 -0.250
## Q7 0.945 -0.258
## Q8 0.945 -0.258
## Q9 0.939 -0.275
## Q10 0.928 -0.249
## Q11 -0.675 -0.587
## Q12 -0.689 -0.579
## Q13 -0.677 -0.579
## Q14 -0.670 -0.592
## Q15 -0.681 -0.581
##
##                ML1    ML2
## SS loadings    6.880 6.617
## Proportion Var 0.459 0.441
## Cumulative Var 0.459 0.900
```

```
loadings <- as.matrix(fa_health$loadings) # Alternatively, try unclass(fa_health$loadings)

loadings_filtered <- ifelse(abs(loadings) < 0.4, NA, loadings)

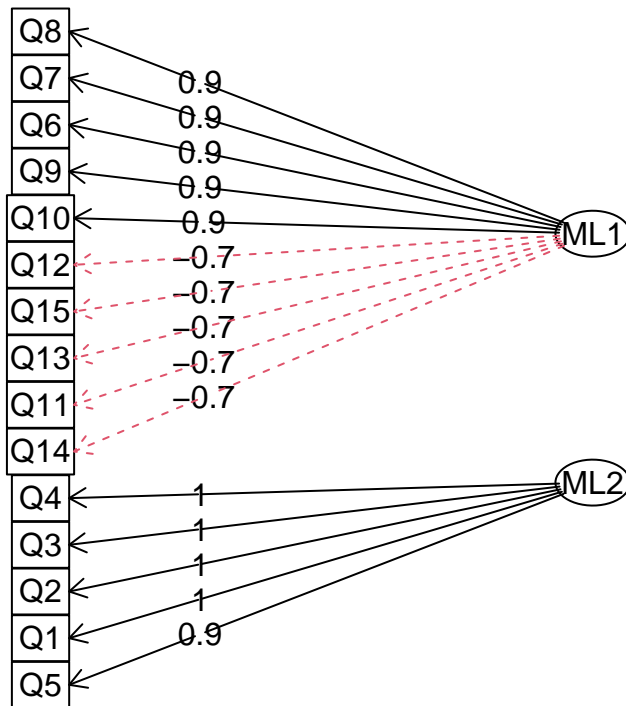
print(loadings_filtered, digits = 3)
```

```
##          ML1    ML2
## Q1         NA 0.957
## Q2         NA 0.960
## Q3         NA 0.960
## Q4         NA 0.965
## Q5         NA 0.944
## Q6 0.944      NA
## Q7 0.945      NA
## Q8 0.945      NA
## Q9 0.939      NA
## Q10 0.928     NA
## Q11 -0.675 -0.587
## Q12 -0.689 -0.579
## Q13 -0.677 -0.579
## Q14 -0.670 -0.592
## Q15 -0.681 -0.581
```

## Diagram of Factors

```
fa.diagram(fa_health)
```

## Factor Analysis



### Name of latent factors:

According to the content of the questions, we can say that:

- Latent Factor ML2: Questions related to physical health
- Latent Factor ML1: Questions related to emotional health and connections

## 4. Report the mean and standard deviation of the average scores of subjects for the factors identified above

```
vegetarian_ml1 <- dfraw[dfraw$diet == "Vegetarian", 1:5]
vegetarian_ml2 <- dfraw[dfraw$diet == "Vegetarian", 6:15]

nonvege_ml1 <- dfraw[dfraw$diet == "Non-Vegetarian", 1:5]
nonvege_ml2 <- dfraw[dfraw$diet == "Non-Vegetarian", 6:15]
```

```
# Calculate and print column means and standard deviations for each subset
# For Vegetarian_ml1
colMeans(vegetarian_ml1)
```

```
##      Q1      Q2      Q3      Q4      Q5
## 12.16774 12.26452 12.14194 12.18710 12.19355
```

```
apply(vegetarian_ml1, 2, sd)
```

```
##      Q1      Q2      Q3      Q4      Q5
## 2.167707 2.171472 2.196333 2.197363 2.104716
```

```
# For Vegetarian_ml2
colMeans(vegetarian_ml2)
```

```
##      Q6      Q7      Q8      Q9      Q10      Q11      Q12      Q13
## 11.95484 11.94194 12.00645 11.99355 11.92258 12.87742 12.87097 12.85161
##      Q14      Q15
## 12.88387 12.83871
```

```
apply(vegetarian_ml2, 2, sd)
```

```
##      Q6      Q7      Q8      Q9      Q10      Q11      Q12      Q13
## 2.242859 2.180905 2.139598 2.178695 2.234719 2.322492 2.295421 2.255710
##      Q14      Q15
## 2.233031 2.190546
```

```
# For NonVegetarian_ml1
colMeans(nonvege_ml1)
```

```
##      Q1      Q2      Q3      Q4      Q5
## 12.33553 12.30921 12.35526 12.34211 12.32895
```

```
apply(nonvege_ml1, 2, sd)
```

```
##      Q1      Q2      Q3      Q4      Q5
## 2.180419 2.211450 2.220936 2.261410 2.251641
```

```
# For NonVegetarian_ml2
colMeans(nonvege_ml2)
```

```
##      Q6      Q7      Q8      Q9      Q10      Q11      Q12      Q13
## 12.42105 12.45395 12.42105 12.42105 12.47368 12.05263 12.05921 12.11842
##      Q14      Q15
## 12.06579 12.15789
```

```
apply(nonvege_ml2, 2, sd)
```

```
##      Q6      Q7      Q8      Q9      Q10      Q11      Q12      Q13
## 2.250905 2.240030 2.323295 2.250905 2.219562 2.140092 2.175217 2.177349
##      Q14      Q15
## 2.139725 2.183943
```