STAC51 Case Study Report

Mahek Prasad, Hammad Bin Arif, Naivil Patel, Tan Lin

03/04/2022

Factors That Are Most Likely To Activate the Presence of Brown Fat in Humans
Group 3

Mahek Prasad (1005877696): Exploratory Data Analysis and Presentation

Hammad Bin Arif (1006209050): Multiple Factor Analysis

Tan Lin (1005102871): Model Building

Naivil Patel (1005569369): Model Validation and Diagnostics

Word Count: 1764

The Libraries Used In This Research

library(readxl)

library(ggplot2)

library(tidyverse)

library(FactoMineR)

library(factoextra)

library(caret)

library("pROC")

library (ROCR)

library(corrplot)

library(ggpubr)

Background and Significance

Abstract

Brown fat, also called brown adipose tissue, is a special type of body fat that is activated when you get cold. Brown fat produces heat to help maintain your body temperature in cold conditions. This capacity to burn fuel and produce heat when exposed to cold temperatures can help treat obesity and other metabolic disorders if we are able to replicate the cold exposure on a cellular level (Reynolds, 2021).There can be many internal and external factors affecting the existence and volume of BrownFat. From our analysis we have found the major factors affecting BrownFat are Age, Sex, External Temp, Weight, and LBW.

This case study aims to provide a better insight into the key factors affecting BrownFat presence in patients so we can better understand the tissue and its relationship to other factors determining its existence and volume.

Research Question: What Factors Are Most Likely To Activate the Presence of Brown Fat in Humans?

Variable Description

- Sex: sex of the patient (Female=1, Male=2)
- Diabetes: (No=0, Yes=1)
- Age: Age of the patient in years
- Day: Day of the year
- Month: Month of the exam
- Ext_Temp: External Temperature
- 2D_Temp: Average temperature of last 2 days.
- 3D_Temp: Average temperature of last 3 days.
- 7D_Temp: Average temperature of last 7 days.
- 1M_Temp: Average temperature of last month.
- Season: Spring=1, Summer=2, Automn=3, Winter=4.
- Duration_Sunshine: Sunshine duration in seconds.
- Weight: in Kgs
- Size: height in cms.
- BMI: Body Mass index. BMI = weight/(height)^2
- Glycemia: the concentration of blood sugar
- Lean Body Weight: total body weight - body fat weight (ie. the weight of everything except fat)
- Cancer_Status: (No=0, Yes=1).
- Cancer_Type: (No=0, lung=1, digestive=2, Oto-Rhino-Laryngology=3, breast=4, gynaecological(female)=5, genital (male)=6, urothelial=7, kidney=8, brain=9, skin=10, thyroid=11, prostate=12, non-Hodgkin lymphoma=13, Hodgkin=14, Kaposi=15, Myeloma=16, Leukemia=17, other=18).
- TSH
- BrownFat: (No=0, Yes=1).
- Total_Vol: Total volume of Brown Fat.

Loading the Data

```
set.seed(1005102871)
library("readxl")
data = read_excel("BrownFat.xls")
```

Examining the Data

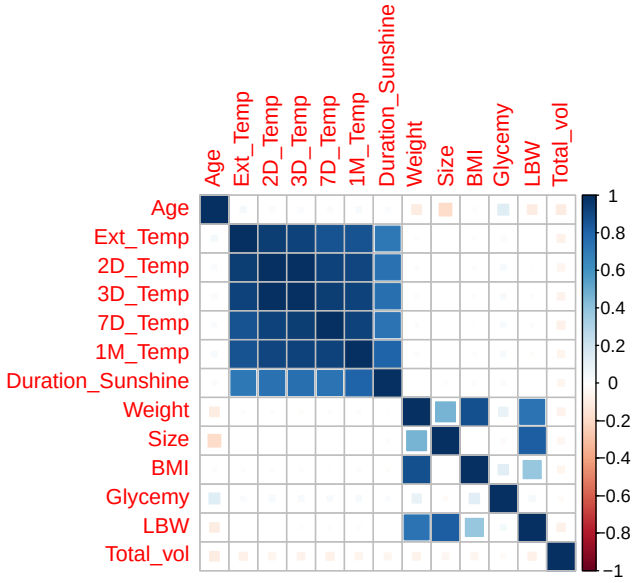
```
# Renaming Weigth to Weight
colnames(data)[14] = "Weight"
#unique(data$Sex) # No bad entries
```

```
#unique(data$Diabetes) # No bad entries
#unique(data$Season) # No bad entries
#Checking for N/A values
#if (!any(is.na(data))) {
#print("No NA values were found in the dataset")
#}

# More Cleaning
data <- data[,!(names(data) %in% c("TSH"))]
data <- data[data$Cancer_Type != "NA",]

#Removing Strange Ages
#data[data$Age < 10,] #Found weird data when looking at the summary statistics.
data = data[!(data$Id %in% c(2489,6875)),]
```

Exploratory Data Analysis / Data Visualization

Analysis of Quantitative Variables and MultiCollinearity

```
# correlation matrix
drop <- c("Month","Sex", "Id", "Diabetes", "Season", "Day", "Cancer_Status", "Cancer_Type", "BrownFat")
df = data[,!(names(data) %in% drop)]
corrplot(cor(df), method = "square")
```



Initially we examined the correlation between the quantitative variables and the total volume of Brown Fat present, to our surprise we did not notice any significant correlation. However, we noted down the high correlation between Weight/Size/BMI/LBW and among the temperature variables to later fix the Multi-Collinearity issue.

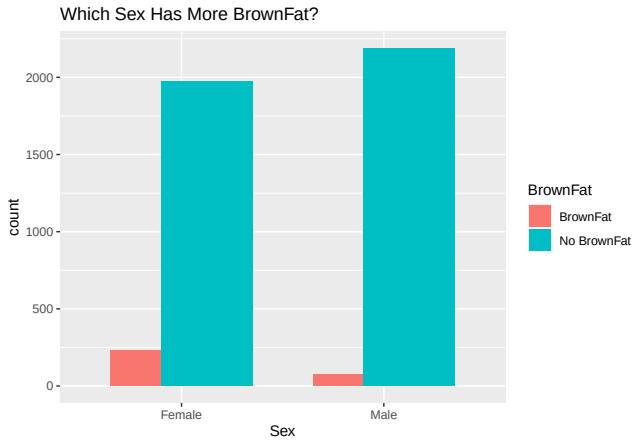
Analysis of Qualitative Variables

```
data$Sex[data$Sex==1] <- "Female"
data$Sex[data$Sex==2] <- "Male"
```

```
data$BrownFat[data$BrownFat==0] <- "No BrownFat"
data$BrownFat[data$BrownFat==1] <- "BrownFat"

data$Diabetes[data$Diabetes==0] <- "Not Diabetic"
data$Diabetes[data$Diabetes==1] <- "Diabetic"
```

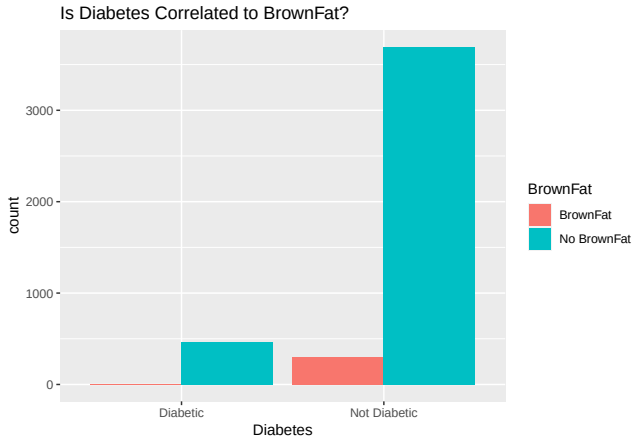
Which Sex Has More BrownFat?

```
  ggplot(data) +
  geom_bar(aes(x = Sex, fill = BrownFat), position = position_dodge(width=0.5)) +
  ggtitle("Which Sex Has More BrownFat?")
```



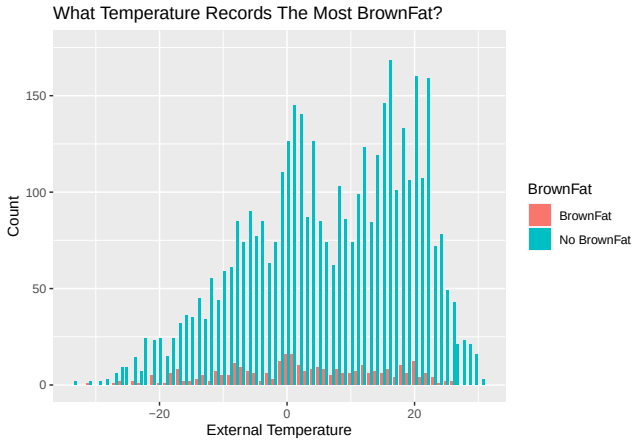
We can see a clear indication of higher brown fat presence in females, indicating sex has some significant correlation with brown fat.

Do People With Diabetes Tend to Have BrownFat?



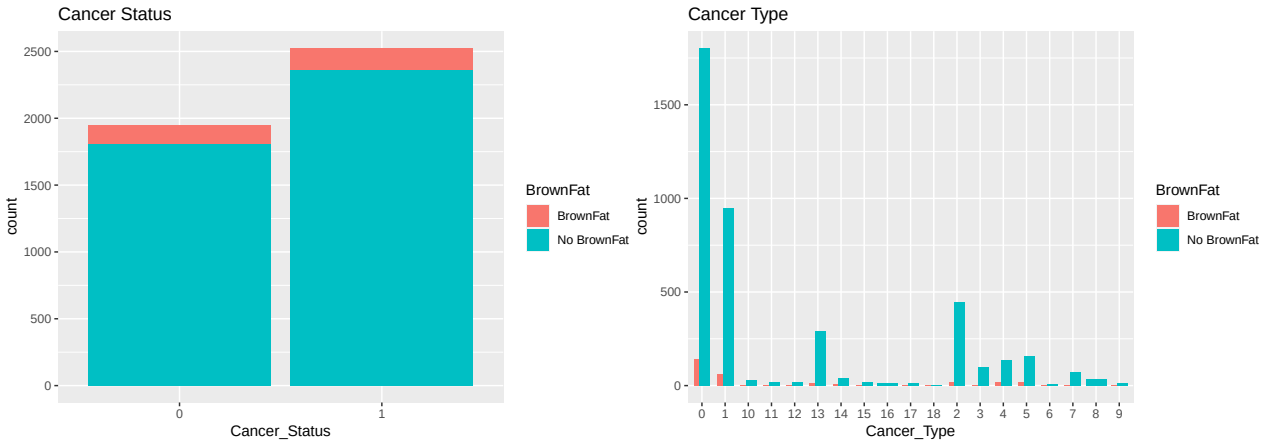
Although we notice a higher count of Non-Diabetic patients with BrownFat, we noticed the proportion of is still quiet small, so we concluded that there may not be much correlation between the two variables.

Which Temperature Records The Most BrownFat?



We checked which temperature would record the most cases of Brown Fat, and as expected we saw that the colder temperatures would record more brown fat (seen by proportion to those tested).

Does Cancer and Cancer Type have any correlation with BrownFat?



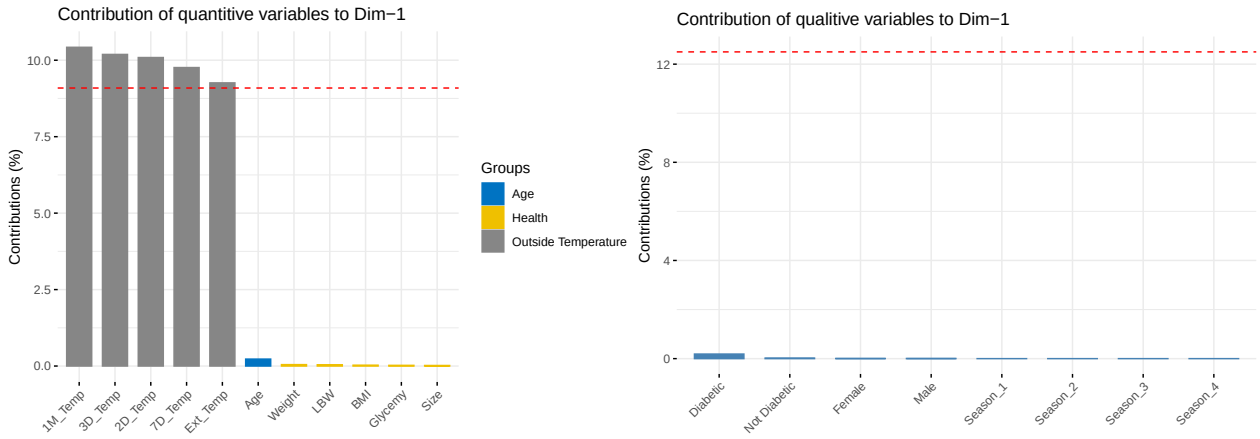
We notice the count of patients with BrownFat is similar however there is a higher proportion among Non-Cancer patients. To reinforce our findings, we will check BrownFat correlation within Cancer Patients to see if there is a correlation between Cancer Types and BrownFat. Plotting the observations, we see that cancer is not a factor of BrownFat.

Multifactor Analysis

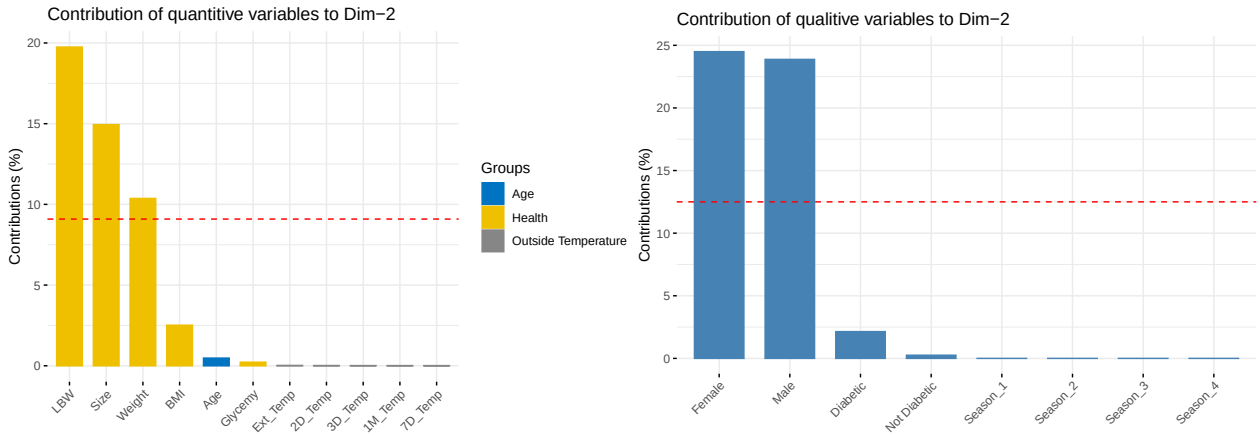
Considering the variables in this model are a mix of categorical and quantitative, the first technique that was used in attempting to reduce the number of variables was multifactor analysis (MFA). MFA involves grouping variables together in meaningful categorical and quantitative groups and applying multiple component analysis (MCA) to categorical variables and principal component analysis (PCA) to quantitative variables. PCA and MCA both involve transforming the data by finding independent components within the groups which can be used to better explain the data than the observed variables (Chavent, 2013). This transforms the data into dimensions. In R, the FactoMineR package was used in order to analyze the data with MFA. As confirmed by our earlier heat map, most of the quantitative ones appeared to be very related to each other namely in the Health and Outside Temperature groups.

In order to get a better idea of the dimensional groupings we will look at the contributions of both qualitative and quantitative variables to each dimension.

```
#Contributions to dimension 1
fviz_contrib(res, choice = "quanti.var", axes = 1, top = 20,
             palette = "jco")
fviz_contrib(res, choice = "quali.var", axes = 1, top = 20,
             palette = "jco")
```




```
# Contributions to dimension 2
fviz_contrib(res, choice = "quanti.var", axes = 2, top = 20,
             palette = "jco")
fviz_contrib(res, choice = "quali.var", axes = 2, top = 20,
             palette = "jco")
```



We can now see that Dim1 is the outside weather as the temperatures are strong contributors to Dim1 from the quantitative variables. Dim2 however, is getting most of its contribution from the health and sex variables from both qualitative and quantitative analysis. This can now help us determine the key variables to include in the model.

Analysis Summary

Based on the correlation matrix all the temperature variables as well as Duration_Sunlight are very highly correlated to one another, with all of their values being >0.7. By multicollinearity, it is adequate to pick one of the 6 variables as it will sufficiently explain the other ones. We will see which of the 6 will give us the lowest AIC value in our later analysis.

From our correlation matrix we also see that all the health variables (Weight, BMI, Size, LBW, Glycemy and Total_Vol) have high correlation, and we will see which of these variables will give us the lowest AIC value later on.

Cancer_Type was another variable that we decided to remove. We came to this conclusion from the graphs in the explanatory data analysis: brown fat relating to cancer was explained much more clearly using Cancer_Status than Cancer_Type and thus the 18 factors that Cancer_Type included were redundant when they could be reduced to 2: having cancer or not having cancer.

Model Selection and Diagnosis

Split Data

```
columns = c("Id","Sex","Diabetes","Age","Ext_Temp","2D_Temp","3D_Temp","7D_Temp","1M_Temp","Weight","Si
data2 = data[,columns]

data2$BrownFat<-ifelse(data$BrownFat=="No BrownFat",0,1)
data2$Sex = as.factor(data2$Sex)
data2$Diabetes = as.factor(data2$Diabetes)

sample = sample(1:length(data$Id),floor(length(data$Id)*0.8),0)
data.train = data2[sample,]
data.test = data2[-sample,]
```

We started by splitting the data into training and testing subsets. We factored the numerical Brown Fat, Sex, Diabetes since they are qualitative, not quantitative, variables.

Single Predictor Models

```
#Bodily Proportion Models
fit.weight = glm(BrownFat~Weight,data=data.train, family=poisson)
fit.size = glm(BrownFat~Size,data=data.train, family=poisson)
fit.bmi = glm(BrownFat~BMI,data=data.train, family=poisson)
fit.lbw = glm(BrownFat~LBW,data=data.train, family=poisson)
```

We chose to examine the bodily proportion models together, because they are high correlated with each other. We tried to select only one or two predictors from this group. Out of the bodily proportion models, Weight and LBW had the lowest AICs. BMI had a slightly higher AIC and size has a high AIC.

```
# Temperature Models
fit.ext_temp = glm(BrownFat~Ext_Temp,data=data.train, family=poisson)
fit.2d_temp = glm(BrownFat~`2D_Temp`,data=data.train, family=poisson)
fit.3d_temp = glm(BrownFat~`3D_Temp`,data=data.train, family=poisson)
fit.7d_temp = glm(BrownFat~`7D_Temp`,data=data.train, family=poisson)
fit.1m_temp = glm(BrownFat~`1M_Temp`,data=data.train, family=poisson)
```

We chose to examine the temperature variables together since they are extremely correlated since they are used to calculate each other. We would have to select a single predictor from these variables. Out of the temperature models, External Temperature had the lowest AIC.

```
# Miscellaneous  Models
fit.sex = glm(BrownFat~Sex,data=data.train, family=poisson)
fit.age = glm(BrownFat~Age,data=data.train, family=poisson)
fit.cancer = glm(BrownFat~Cancer_Status,data=data.train, family=poisson)
fit.glycemy = glm(BrownFat~Glycemy,data=data.train, family=poisson)
fit.diabetes = glm(BrownFat~Diabetes,data=data.train, family=poisson)
```

Out of the miscellaneous models, Sex and Age had the lowest AICs. Diabetes had a higher AIC while Cancer Status and Glycemy had much higher AICs.

Takeaways From Single Predictor Models

We didn't include the date variables because we believed that date only affects Brown Fat via temperature. We divided the remaining predictors into 3 groups: Bodily Proportions, Temperature, and Miscellaneous. From our investigation of these groups we found that Age, Sex, External Temperature, Weight, and LBW had the lowest AICs. We also included Diabetes We chose these variables as the predictors in our main effect model.

Combined Models

```
fit.main = glm(BrownFat ~ Age + Sex + Ext_Temp + Weight + LBW + Diabetes,
               data=data.train, family=poisson)
```

We created our main effect model with the predictors that we believe captured the most relevant information. Every variable is significant though LBW is only significant at 0.01 level. ## Automated Model Selection

```
fit.sat = glm(BrownFat ~ Age * Sex * Ext_Temp * Weight * LBW,data=data.train, family=poisson)
fit = glm(BrownFat~1,data=data.train, family=poisson)
fit.forw = step(fit, scope=list(upper=fit.sat,lower=fit), direction="forward",trace=0)
fit.back = step(fit.sat, direction="backward",trace=0)
fit.both = step(fit, scope=list(upper=fit.sat,lower=fit), direction="both",trace=0)
```

We created the saturated model then used step for automated model selection. We compared all three directions: forward, backward, and both. During this process we had ran into trouble when we tried to include Diabetes. It would cause backwards regression to break. We decided to only use Diabetes when we manually selected models. Each model had the same AIC but we chose the model from backwards step since it had a lower residual deviance.

Manual Model Selection

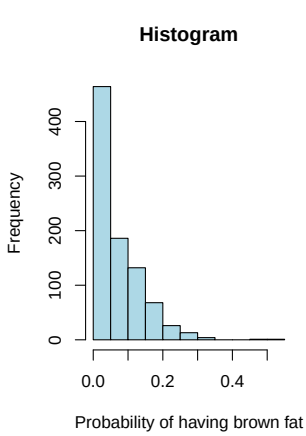
```
fit.manual = glm(BrownFat ~ Age + Sex + Ext_Temp + LBW + Weight + Diabetes + Age:Sex
                 + I(LBW/Weight) + Diabetes:Age,data=data.train, family=poisson)
```

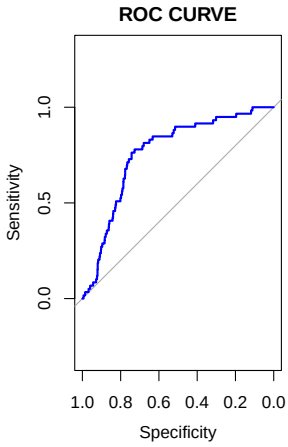
We experimented with adding Diabetes and found that it improved our model. We also considered how LBW and Weight might interact. We thought that the percentage of LBW would be significant and we were proven right. We ended up with a lower AIC by incorporating Diabetes manually.

```
#Our Final Model:
fit.selected = fit.manual
```

Model Validation / Diagnostics

```
y_hat = as.vector(predict(fit.selected,data.test,type="response")) #get the probabilities
men<- mean(y_hat)
predictions <- ifelse(y_hat>0.09, 1, 0)
library("pROC")
library(VGAM)
bf = data.test$BrownFat
par(mfrow=c(1,2))
hist(y_hat, main="Histogram",xlab="Probability of having brown fat", col="light blue")
roc(data.test$BrownFat~y_hat, plot = TRUE, main = "ROC CURVE", col= "blue")
```






```
##
## Call:
## roc.formula(formula = data.test$BrownFat ~ y_hat, plot = TRUE,     main = "ROC CURVE", col = "blue")
##
## Data: y_hat in 836 controls (data.test$BrownFat 0) < 59 cases (data.test$BrownFat 1).
## Area under the curve: 0.7621
```

```
auc(data.test$BrownFat~y_hat)
```

```
## Area under the curve: 0.7621
```

```
confusionMatrix(as.factor(predictions),as.factor(bf))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 601  13
##          1 235  46
##
##                Accuracy : 0.7229
##                  95% CI : (0.6923, 0.752)
##     No Information Rate : 0.9341
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1814
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7189
##             Specificity : 0.7797
##          Pos Pred Value : 0.9788
##          Neg Pred Value : 0.1637
##              Prevalence : 0.9341
##          Detection Rate : 0.6715
##    Detection Prevalence : 0.6860
##       Balanced Accuracy : 0.7493
##
##        'Positive' Class : 0
##
```

```
anova(fit.selected,fit.sat,test="LRT")
```

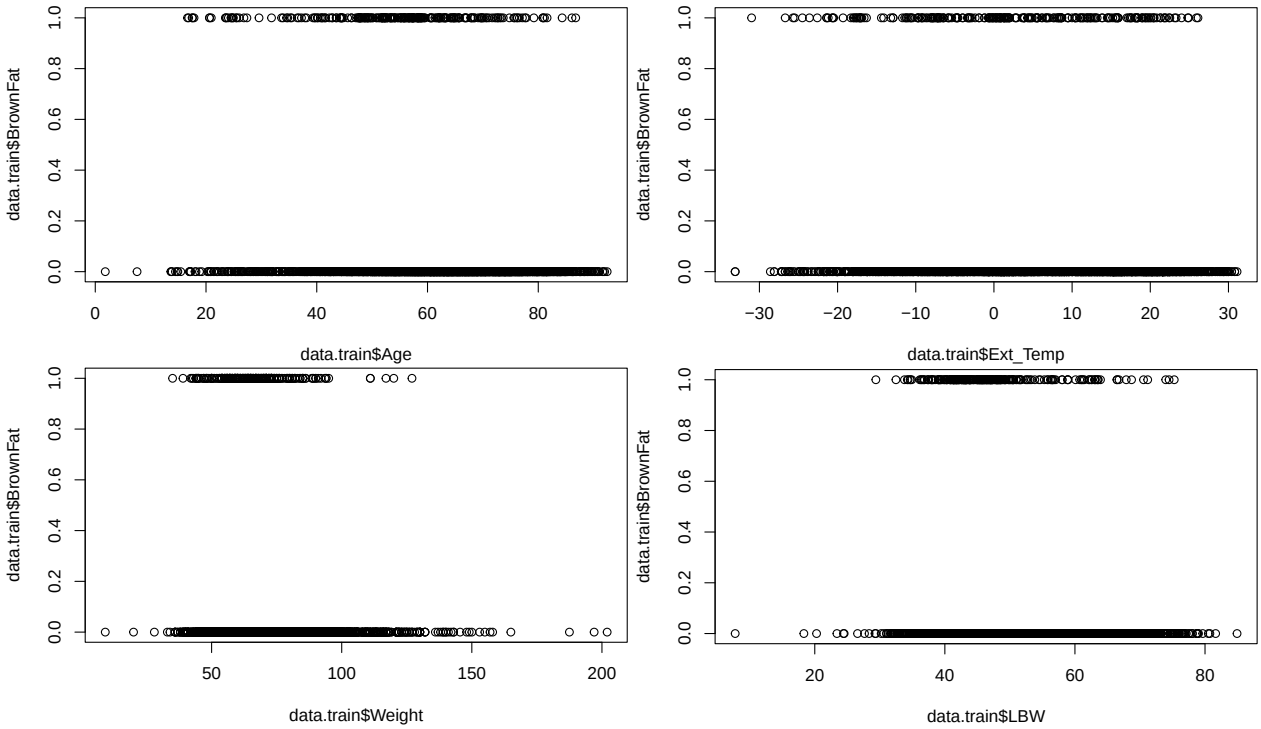
```
## Analysis of Deviance Table
##
## Model 1: BrownFat ~ Age + Sex + Ext_Temp + LBW + Weight + Diabetes + Age:Sex +
##     I(LBW/Weight) + Diabetes:Age
## Model 2: BrownFat ~ Age * Sex * Ext_Temp * Weight * LBW
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3566     1117.1
## 2      3544     1114.0 22   3.0157        1
```

```
anova(fit.selected,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: BrownFat
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          3575     1330.3
## Age            1   78.140      3574     1252.1 < 2.2e-16 ***
## Sex            1   57.648      3573     1194.5 3.134e-14 ***
## Ext_Temp       1   20.887      3572     1173.6 4.873e-06 ***
## LBW            1    5.156      3571     1168.4 0.0231663 *
## Weight         1   21.590      3570     1146.8 3.376e-06 ***
## Diabetes       1   13.554      3569     1133.3 0.0002317 ***
## I(LBW/Weight)  1    3.449      3568     1129.8 0.0632732 .
## Age:Sex        1    8.933      3567     1120.9 0.0028005 **
## Age:Diabetes   1    3.855      3566     1117.1 0.0495983 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Goodness of fit test based on deviance gives a p value of 1, which suggests the model fits the data well. It also shows that the model created by the backward process is better than the saturated model. Testing the model with a drop1() function with the Chisq test shows that all variables are needed as they are not equal to 0. Our area under the ROC curve was 0.75 which shows that it is a fairly good model. The precision of the test is 16.4%, which shows that a large number of positive predictions done by the model are wrong. However due to predicting a large number of positives, we were able to correctly predict 78% of the true positive cases which means our recall is higher. For our model, we want the recall to be higher, due to the fact that we want to correctly predict people having brown fat and false positives do not matter as much thus having a higher recall is better for this model.

Discussion/Conclusion



The goal of this report was to identify the factors that are likely to activate the presence of BrownFat in humans. From our analysis, we can conclude that Age, Sex, External Temperature, Weight, and Lean Body Weight are the factors that will most likely activate the presence of brown fat. We also went a step further and found that from our data set, those from age 45 - 70, with weight between 40-90 kgs, lean body weight between 35 - 65 kgs and females are more likely to have brown fat during external temperatures of -15 to 25 degrees celsius. Lastly, an interesting fact we found was that for every 1 kg increase in lean body weight, the probability of the person having brown fat increases by 25%. Analyzing factors that affect BrownFat in humans allow for significant research in this field, and can advance healthcare in various ways such as treating obesity and other metabolic disorders.

However, one limitation from our analysis would be that the dataset the model was built upon had a negatively skewed data set on brown fat, as it contained substantially more negatives than positives. This would lead our model to generate more negative predictions than positive, and since the test data will also reflect this, any new data with covariates that are positively skewed on BrownFat would not work very well with this model. For future research, it would be interesting to know by how much the factors we identified can increase brown fat. In other words, while we were able to identify factors that cause BrownFat, which factors are able to increase the amount of BrownFat in humans?

References

Reynolds, S. (2021, May 11). Uncovering the origins of Brown fat | National Institutes of Health (NIH). Nih. Retrieved from https://www.nih.gov/news-events/nih-research-matters/uncovering-origins-brown-fat

Chavent, M. (2013). Multiple Correspondance Analysis (MCA) | Bordeaux University. Retrieved from http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/ACM-M2.pdf