***INDEX***

### *Introduction*

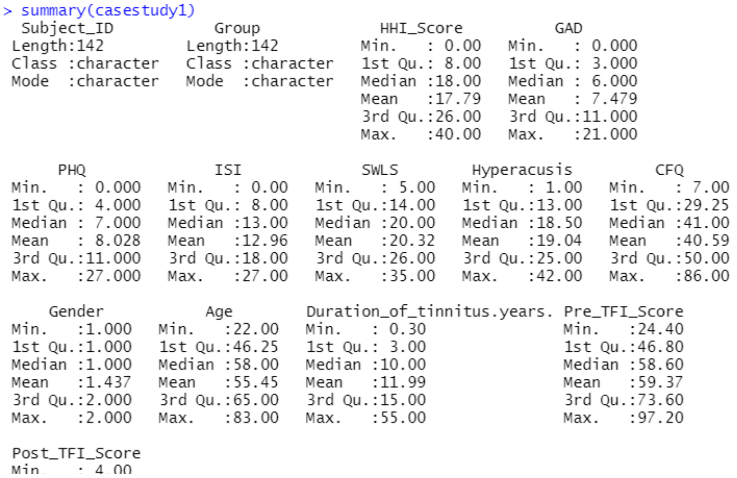
Tinnitus can be described as a sensation of hearing ringing or buzzing sounds even in the absence of external sounds. Often, tinnitus interferes with daily tasks resulting in psychological distress. To help tinnitus patients cope with the psychological effects of tinnitus, an internet based cognitive behavioral therapy (ICBT) was developed. For this case study, a data set containing the information of 142 subjects experiencing Tinnitus has been provided. Information provided by the data set are: pre-study and post-study tinnitus scores (Pre_TFI and Post_TFI scores), clinical information such as a hearing survey score (HHI_Score), Generalized Anxiety Disorder score (GAD), depression sum (PHQ), Insomnia (ISI), satisfaction with life (SWLS), hyperacusis, cognitive failures (CFQ), duration of tinnitus, and demographic information such as Gender and Age.

### *Methodology*

To conduct this study, R Software (Version 1.4.1717) was used. In R software the following libraries were used: tidyverse to remove variables as needed; corrplot for correlation matrix, caret for data partition; leaps, olsrr, and MASS for model selection; nortest for Anderson-Darling normality test; splines and gam for Generalized Additive Model (GAM).

### *Exploring the data.*

After importing the .csv file into R, we proceed to do some data exploration. To familiarize ourselves with the data, we do some descriptive statistics. Using the View() function, we are able to visualize the data in its rows and columns. Here, we can obtain the number of subjects (observations) by focusing on the number of rows, and to get an idea of the number of predictors we can observe the number of columns. Using the dim() function, we observe that the data set contains 142 observations (rows), and 14 predictors(columns). Despite this initial observation, the number of predictors is subject to change since not all predictors bring substantial information to what we will try to estimate.



```
> summary(casestudy1)
  Subject_ID           Group             HHI_Score          GAD        
 Length:142         Length:142         Min.   : 0.00   Min.   : 0.000  
 Class :character   Class :character   1st Qu.: 8.00   1st Qu.: 3.000  
 Mode  :character   Mode  :character   Median :18.00   Median : 6.000  
                                       Mean   :17.79   Mean   : 7.479  
                                       3rd Qu.:26.00   3rd Qu.:11.000  
                                       Max.   :40.00   Max.   :21.000  
                                                                       
      PHQ              ISI             SWLS        Hyperacusis         CFQ       
 Min.   : 0.000   Min.   : 0.00   Min.   : 5.00   Min.   : 1.00   Min.   : 7.00  
 1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:14.00   1st Qu.:13.00   1st Qu.:29.25  
 Median : 7.000   Median :13.00   Median :20.00   Median :18.50   Median :41.00  
 Mean   : 8.028   Mean   :12.96   Mean   :20.32   Mean   :19.04   Mean   :40.59  
 3rd Qu.:11.000   3rd Qu.:18.00   3rd Qu.:26.00   3rd Qu.:25.00   3rd Qu.:50.00  
 Max.   :27.000   Max.   :27.00   Max.   :35.00   Max.   :42.00   Max.   :86.00  
                                                                                 
     Gender           Age        Duration_of_tinnitus.years. Pre_TFI_Score  
 Min.   :1.000   Min.   :22.00   Min.   : 0.30               Min.   :24.40  
 1st Qu.:1.000   1st Qu.:46.25   1st Qu.: 3.00               1st Qu.:46.80  
 Median :1.000   Median :58.00   Median :10.00               Median :58.60  
 Mean   :1.437   Mean   :55.45   Mean   :11.99               Mean   :59.37  
 3rd Qu.:2.000   3rd Qu.:65.00   3rd Qu.:15.00               3rd Qu.:73.60  
 Max.   :2.000   Max.   :83.00   Max.   :55.00               Max.   :97.20  
                                                                            
 Post_TFI_Score 
 Min.   : 4.00  
```

*Figure 1: Summary of the predictors in CaseStudy1 data set. Here, we can check for descriptive statistics, implied classes of the predictors, and missing values if any.*

To draw some insight about the nature of our predictors, we call the summary() function in Fig 1. Through the summary function, we can observe that the predictors Subject_ID and Group are qualitative since we are given some information on the class, length, and mode. However, the predictors from HHI_Score through Post_TFI_Score are quantitative. For the quantitative, predictors, we can get some insight on their mean, median, quartiles, minimum and maximum values. Additionally, we can observe that the predictor Post_TFI_Score has a total of 86 missing

(NA) values. The summary function is a useful tool that can aid us on the next steps of our data management. For example, before proceeding with our linear fit, we know we must omit our qualitative predictors Subject_ID and Group, and that we can use the Mean to substitute the missing values in Post_TFI_Score.

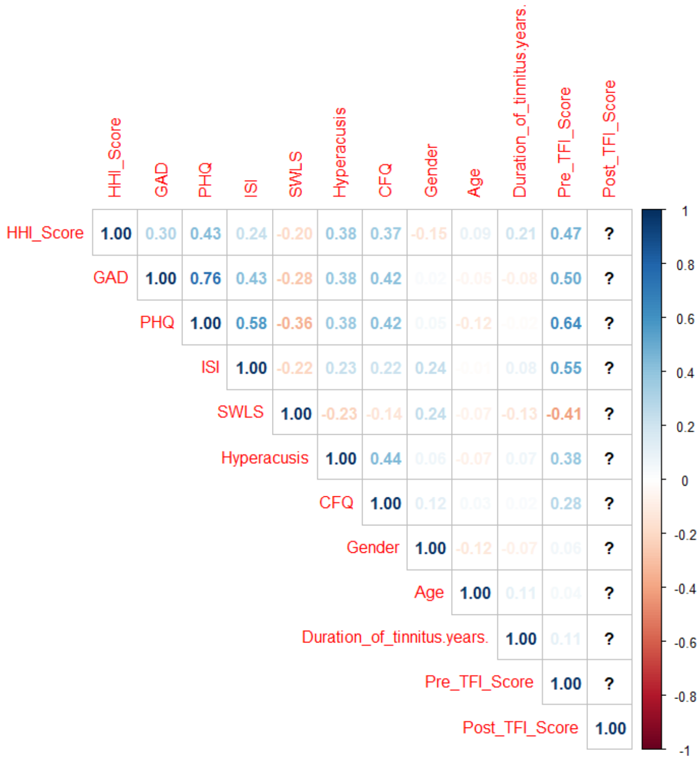


*Figure 2: Correlation Matrix of the variables.*

Additionally, we check if there is any correlation between the variables. There are different ways to check for correlation. However, a visually helpful, and intuitive way to check for correlation is using a correlation matrix given by the library corrplot (Fig. 2). As shown in the correlation matrix in Fig. 2, we can visualize the correlation between two predictors, we have both its quantifiable measure, and an intuitive color-coded way to interpret it. In the given color key, dark blue corresponds to heavily positively correlated variables, and dark red corresponds to negatively correlated variables. The gradient between dark blue and dark red is a relative measure of either positive or negative correlated variables. Given the pastel tones in the coded numbers, the variables are not correlated in its majority. However, we can see a positive correlation between GAD and PHQ(0.76), PHQ and ISI(0.58), PHQ and Pre_TFI_Score (0.64), ISI and Pre_TFI_Score (0.55), and GAD and Pre_TFI_Score(0.50). The relationship between GAD, PHQ, and ISI can first serve as evidence on how Generalized Anxiety Disorders and Insomnia, negatively affect the patient’s satisfaction with life which is linked to a high depression sum measured by the variable PHQ. The latter correlations, PHQ, ISI, and GAD to Pre_TFI_Score, show that high levels of depression, insomnia, and generalized anxiety disorders lead to a high score in a patient’s Pre-Tinnitus Functional Index measured by the variable Pre_TFI_Score.

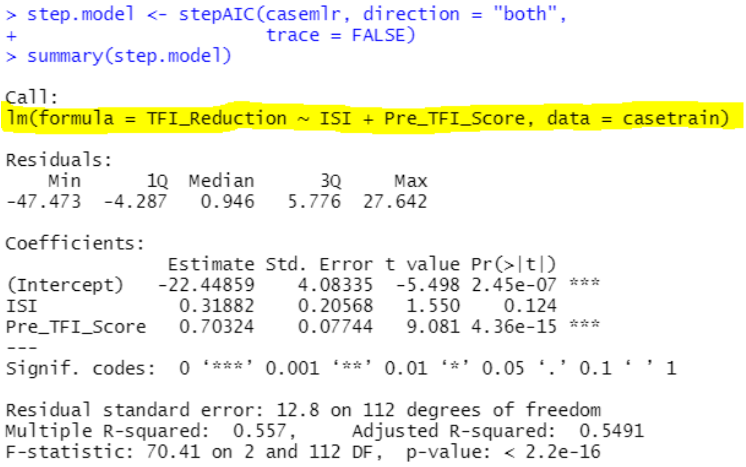
### *Data Cleaning and Preparation*

The data was cleaned and prepared before proceeding to the models. Using the library tidyverse, the variables Subject_ID and Group were removed. A new data frame titled numcasestudy1 containing all predictors except for Subject_ID, and Group was created and then used to create the correlation matrix and in the models. Recalling to Fig 1., it was detected that Post_TFI_Score contained a total of 86 missing observations (NAs). To solve for this, the mean of Post_TFI_Score was used to replace the missing observations. In numcasestudy1, a new predictor containing the difference between Pre_TFI_Score and Post_TFI_Score was created and named TFI_Reduction. With the removal of the qualitative predictors and the addition of TFI_Reduction, numcasestudy1 has a total of 13 variables and still 142 observations. Additionally, the is.na command was used in all 13 predictors, to confirm that no other observations were missing.

After the data was cleaned, the data in numcasestudy1 was partitioned using the library caret and the random number generator set.seed(123). For the data partitioning, 80% of the data was used for training samples and stored in the data frame casetrain, and 20% of the data was used for testing samples and stored in the data frame castest.

### *Multiple Linear Regression and Model Selection*

Using the data in casetrain, a multiple linear regression model named casemlr was fitted. Since a goal in this study is to analyze the psychological improvement in Tinnitus patients due to ICBT, it is natural to make TFI_Reduction our response variable. Afterall, the psychological improvement in tinnitus patients can be measured by the change of Pre_TFI_Score vs Post_TFI_Score. In casemlr, TFI_Reduction is the response variable, and the predictors are all the remaining variables with the exception of Post_TFI_Score. A glance at the p-values of the predictors in casemlr given by the summary function can give us insight on statistical significance. Here, Pre_TFI_Score is the predictor that looks to be the most statistically significant since it has a p-value < 0.05. Regardless, several selection methods were used to explore the best linear regression model.



```
> step.model <- stepAIC(casemlr, direction = "both",
+                       trace = FALSE)
> summary(step.model)

Call:
lm(formula = TFI_Reduction ~ ISI + Pre_TFI_Score, data = casetrain)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.473  -4.287   0.946   5.776  27.642 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -22.44859    4.08335  -5.498 2.45e-07 ***
ISI             0.31882    0.20568   1.550    0.124    
Pre_TFI_Score   0.70324    0.07744   9.081 4.36e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.8 on 112 degrees of freedom
Multiple R-squared:  0.557,	Adjusted R-squared:  0.5491 
F-statistic: 70.41 on 2 and 112 DF,  p-value: < 2.2e-16
```

*Figure 3: Stepwise Selection Method from MASS library suggesting ISI and Pre_TFI_Score*

For selection methods; leaps, olsrr, and MASS libraries were used. Using the best subset selection method contained in the leaps library, and looking at the adjusted R squared, CP, BIC, and AIC values, the linear model containing two predictors (ISI, and Pre_TFI_Score) was selected. In alignment with the best subset selection in the leaps library, stepwise selection method in MASS, and forward and backward selection in olsrr library also suggested a linear model containing ISI and Pre_TFI_Score as the best model (Fig. 3).

### *Linear Regression Models and Analysis*

A linear regression model containing ISI and Pre_TFI_Score was fitted and named bmlr. To seek for outliers and leverage points, the plot of bmlr was also used. However, given current understanding, a computation of the observations Cook’s distance, standard residuals, and hat values resulted more useful. From the plots and computations, the model was a promising candidate to make predictions since it had no outliers or leverage points (Fig. 4).

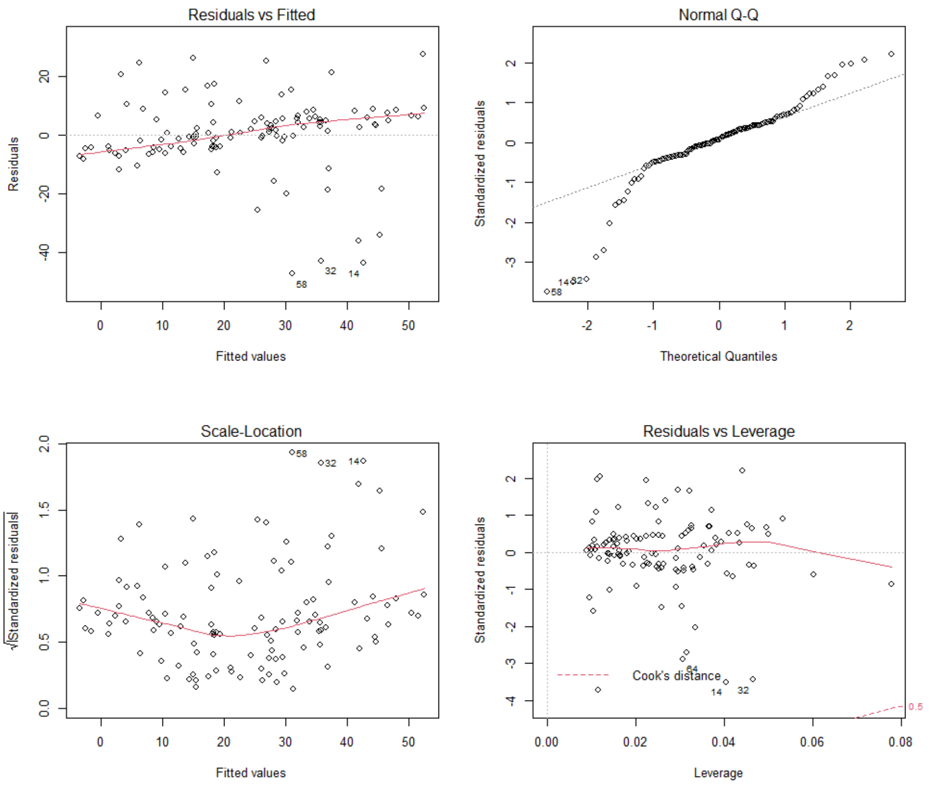


*Figure 4: Plot of bmlr with ISI and Pre_TFI_Score as predictors. Data does not seem to have outliers or leverage points.*

However, the bmlr model failed normality tests. To test for normality, the Shapiro-Wilk normality test, and Anderson-Darling normality test were used. The model bmlr had a p-value < 0.05 indicating a failure in normality. To seek a better model, a new linear regression model titled bmlr1 containing only Pre_TFI_Score as suggested by stepwise selection in oslrr library was fitted. However, bmlr1 also failed normality tests.

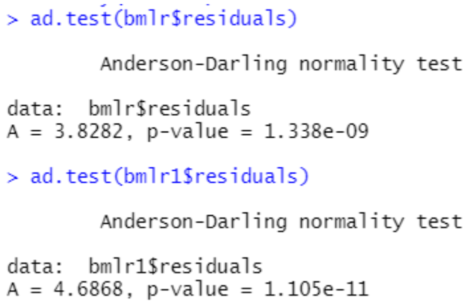

```
> ad.test(bmlr$residuals)

        Anderson-Darling normality test

data:  bmlr$residuals
A = 3.8282, p-value = 1.338e-09

> ad.test(bmlr1$residuals)

        Anderson-Darling normality test

data:  bmlr1$residuals
A = 4.6868, p-value = 1.105e-11
```

*Figure 5: Anderson-Darling normality test on bmlr and bmlr1*

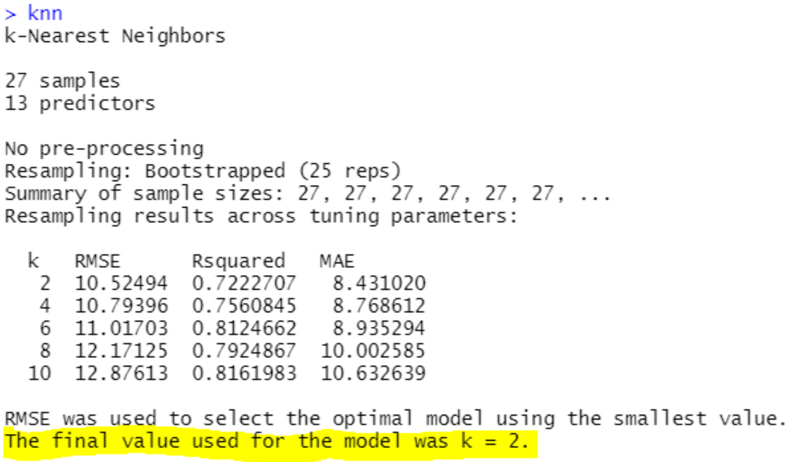
### *Generalized Additive Model (GAM) and Analysis*

Generalized Additive Model (GAM) is an alternative to linear regression since we do not need the normality assumption to be satisfied. Since both linear regression models bmlr and bmlr1 failed the normality tests, a Generalized Additive Model (GAM) was used.

To use GAM model in R, the libraries splines and gam were used. Two GAM models using our training data were fitted. The first model, gam.fit1 uses natural splines and contains

TFI_Reduction as the response variable, and Pre_TFI_Score with 4 degrees of freedom, and ISI with 5 degrees of freedom as predictors. The second GAM model gam.fit2, was fitted using splines, and also contains TFI_Reduction as the response variable, and Pre_TFI_Score and ISI as predictors using the same degrees of freedom as gam.fit1. Predictions for both models were done using the testing data and the predict function and their root mean square error (RMSE) calculated. For both gam.fit1 and gam.fit2 the RMSE was calculated to be 2.0038.

### *K-Mean Regression and Analysis*



```
> knn
k-Nearest Neighbors

27 samples
13 predictors

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 27, 27, 27, 27, 27, 27, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared   MAE
   2  10.52494  0.7222707   8.431020
   4  10.79396  0.7560845   8.768612
   6  11.01703  0.8124662   8.935294
   8  12.17125  0.7924867  10.002585
  10  12.87613  0.8161983  10.632639

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 2.
```

*Figure 6: k-value selection in K-mean regression.*

Alternative to linear regression and GAM, K-mean regression was also performed and its RMSE compared to the GAM model. Using the test data set, k-mean regression was used using the train function and specifying knn as the method, and stored in the variable knn. For k-mean regression, the k-values used were 2,4,6,8, and 10. After calling knn, the final value used for the model was of k = 2, as it contains the lowest root mean square error (10.52) compared to the other k's. Consecutively, predictions on the knn model were done using the testing data and its root mean square calculated. The RMSEfor the K-Mean regression model was calculated to be: 0.2817.

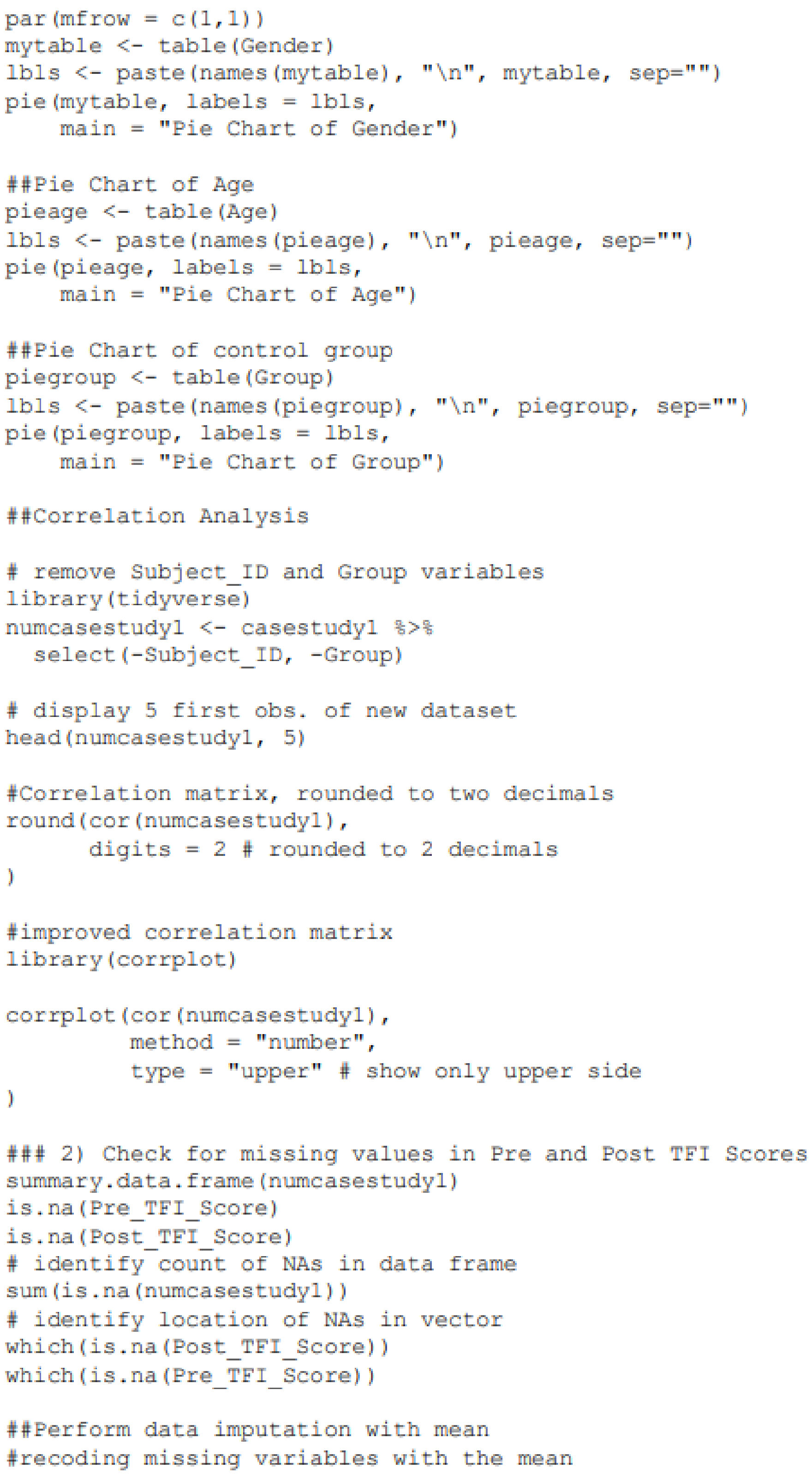
### *Conclusions*

Comparing the RMSE of the GAM model (2.0038) to the K-Mean Regression (0.287). The method that had the lowest RMSE in this case study was K-Mean Regression.

Possible future improvements and analysis to this case study are investigating further on the reasons why linear regression models failed the normality test, researching data normalization techniques, and peer collaboration to solve the data normality issue. Additionally, further investigation and analysis on GAM models to enrich discussion can be added.

# APPENDIX

```
#### CASE STUDY #1 ####

###Set working directory to visualize the data
setwd("C:/Users/betyv/Documents/Fall 2021/Statistical Learning/Case Studies/
Case Study 1")
getwd()

###Read the data
casestudy1 <- read.csv(file = 'CaseStudy1.csv')
View(casestudy1)
dim(casestudy1)
attach(casestudy1)

### MULTIPLE REGRESSION ANALYSIS ###

### 1) Do descriptive analysis: a.k.a Pie Charts, Histograms, Correlation
Analysis, etc.

##check for mode / mean of certain variables.
summary(casestudy1)

##check which variables are quantitative and which are qualitative
  class(Subject_ID)
  class(Group)
  class(HHI_Score)
  class(GAD)
  class(PHQ)
  class(ISI)
  class(SWLS)
  class(Hyperacusis)
  class(CFQ)
  class(Gender)
  class(Age)
  class(Duration_of_tinnitus.years.)
  class(Pre_TFI_Score)
  class(Post_TFI_Score)


## perfroming a histogram for
par(mfrow = c(4,3))
hist(Age)
hist(HHI_Score)
hist(GAD)
hist(PHQ)
hist(ISI)
hist(SWLS)
hist(Hyperacusis)
hist(CFQ)
hist(Gender)
hist(Duration_of_tinnitus.years.)
hist(Pre_TFI_Score)
hist(Post_TFI_Score)

##Pie Chart Gender
```



```
par(mfrow = c(1,1))
mytable <- table(Gender)
lbls <- paste(names(mytable), "\n", mytable, sep="")
pie(mytable, labels = lbls,
    main = "Pie Chart of Gender")

##Pie Chart of Age
pieage <- table(Age)
lbls <- paste(names(pieage), "\n", pieage, sep="")
pie(pieage, labels = lbls,
    main = "Pie Chart of Age")

##Pie Chart of control group
piegroup <- table(Group)
lbls <- paste(names(piegroup), "\n", piegroup, sep="")
pie(piegroup, labels = lbls,
    main = "Pie Chart of Group")

##Correlation Analysis

# remove Subject_ID and Group variables
library(tidyverse)
numcasestudy1 <- casestudy1 %>%
  select(-Subject_ID, -Group)

# display 5 first obs. of new dataset
head(numcasestudy1, 5)

#Correlation matrix, rounded to two decimals
round(cor(numcasestudy1),
      digits = 2 # rounded to 2 decimals
)

#improved correlation matrix
library(corrplot)

corrplot(cor(numcasestudy1),
         method = "number",
         type = "upper" # show only upper side
)

### 2) Check for missing values in Pre and Post TFI Scores
summary.data.frame(numcasestudy1)
is.na(Pre_TFI_Score)
is.na(Post_TFI_Score)
# identify count of NAs in data frame
sum(is.na(numcasestudy1))
# identify location of NAs in vector
which(is.na(Post_TFI_Score))
which(is.na(Pre_TFI_Score))

##Perform data imputation with mean
#recoding missing variables with the mean
```

```
numcasestudy1$Post_TFI_Score[is.na(numcasestudy1$Post_TFI_Score)] <-
mean(numcasestudy1$Post_TFI_Score, na.rm = TRUE)
Post_TFI_Score
View(numcasestudy1)
hist(numcasestudy1$Post_TFI_Score)

##Create new variable 'TFI_Reduction' by subtrating Post_TFI_Score by
Pre_TFI_Score. TFI_Reduction will be the response in the MLR model.
numcasestudy1$TFI_Reduction <- numcasestudy1$Pre_TFI_Score -
numcasestudy1$Post_TFI_Score

hist(numcasestudy1$TFI_Reduction)

### 3) Use mean to impute data for numerical measurements and mode for
categorical measurements.
which(is.na(Group))
which(is.na(Subject_ID))
which(is.na(HHI_Score))
which(is.na(GAD))
which(is.na(PHQ))
which(is.na(ISI))
which(is.na(SWLS))
which(is.na(Hyperacusis))
which(is.na(CFQ))
which(is.na(Gender))
which(is.na(Age))
which(is.na(Duration_of_tinnitus.years.))
which(is.na(casestudy1$TFI_Reduction))



### 4) Partition the data set (obtained in step 2 / I assume is the
TFI_Reduction data)
## 80% train data / 20% test data. (Hint. Use set.seed(123)) and
createDataPartition() in Caret package.
library(caret)
set.seed(123)
training.samples <- numcasestudy1$TFI_Reduction %>%
  createDataPartition(p = .8, list = FALSE)

casetrain <- numcasestudy1[training.samples,]
castest <- numcasestudy1[-training.samples,]

### 5) Perform MLR. TFI_Reduction is the response. Comment on findings. Tip:
use best subset selection/fwd/bckwd selection to select the best MLR model
with lm().
casemlr <- lm(TFI_Reduction ~.-Post_TFI_Score, data = casetrain)
summary(casemlr)

##Best subset selection leaps
library(leaps)
bss.leaps <- regsubsets(TFI_Reduction~.-Post_TFI_Score, data = casetrain,
really.big = TRUE, nvmax = 12)
bss.leaps
bss.leaps.sum <- summary(bss.leaps)
bss.leaps.sum
names(bss.leaps.sum)
bss.leaps.sum$rsq

##Best subset with different library
install.packages("olsrr")
library(olsrr)

bss.olsrr <- ols_step_best_subset(casemlr)
bss.olsrr
names(bss.olsrr)

##Stepwise selection
library(MASS)
step.model <- stepAIC(casemlr, direction = "both",
                      trace = FALSE)
summary(step.model)
step.model

##Stepwise selection using olsrr library
ols_step_both_p(casemlr)
?ols_step_both_p

##Forward selection olsrr lib
ols_step_forward_p(casemlr)

##Backward selection olsrr lib
library(olsrr)
back.model <- ols_step_backward_aic(casemlr)
back.model
back.model$model


### 6) Model Diagnostics
## 6.0 - Assess model with a high prediction power (use multiple metrics like
Adjusted R2, AIC, BIC to select the best model.)
data.frame(
  Adj.R2 = which.max(bss.leaps.sum$adjr2),
  CP = which.min(bss.leaps.sum$cp),
  BIC = which.min(bss.leaps.sum$bic)
)
## 6.1 - Correct violation in the model assusmptions, if any. (E.g for U shape
in residual plot? inlcude quadratic term. Influential point? create two
regression models; one with and one without the data point to ssee how the
estimates ands std errors get impacted).
#best model according to tests includes Pre_TFI_Score and Age as predictors

#According to best subset, new fit is:
bmlr <- lm(TFI_Reduction ~ ISI + Pre_TFI_Score, data = casetrain)
summary(bmlr)
par(mfrow = c(2,2))
plot(bmlr)
```

```
bmlr1 <- lm(TFI_Reduction ~ Pre_TFI_Score, data = casetrain)


##Normality tests
#shapiro test
shapiro.test(bmlr$residuals)
shapiro.test(bmlr1$residuals)

install.packages('nortest')
library(nortest)
ad.test(bmlr$residuals)
ad.test(bmlr1$residuals)

#checking for outliers and lvg points
#cooks distance
cooksd <- as.data.frame(cooks.distance(bmlr))
view(cooksd)

# (4/(n-p-1)
4/(115-13-1)

#standarized residuals
stanres <-as.data.frame(rstandard(bmlr))
stanres
view(stanres)

#hat values
hatv <- as.data.frame(hatvalues(bmlr))
hatv
view(hatv)

#2(p+1)/n
14*2
28/115

which.max(hatv)
?filter_if


##Since linear model failed linearity test, use GAM
install.packages(gam)
install.packages(splines)
library(splines)
library(gam)

#GAM with natural splines
gam.fit1 <- lm(TFI_Reduction ~ ns(Pre_TFI_Score, 4) + ns(ISI, 5) , data =
casetrain)
gam.fit1
summary(gam.fit1)

#GAM with splines
gam.fit2 <- gam(TFI_Reduction ~ s(Pre_TFI_Score, 4) + s(ISI, 5), data =
casetrain)
gam.fit2
summary(gam.fit2)

#DO ANOVA
anova(gam.fit1, gam.fit2)



### 7) After you clarifying that there is no any issue with model assumptions,
use that model to find out the factors which highly influence the reduction in
TFI score. Comment on your findings.
#use predictors from best subset


### 8) Make predictions on the test data set. Comment on the mean square error
on the testing data set.
##RMSE for GAM

predict_gam1 <- predict(object = gam.fit1, newdata=castest)
MSEpred1 <- sqrt(mean(castest$TFI_Reduction-predict_gam1)^2)
MSEpred1

predict_gam2 <- predict(object = gam.fit1, newdata=castest)
MSEpred2 <- sqrt(mean(castest$TFI_Reduction-predict_gam2)^2)
MSEpred2

### K-MEAN Regression ###

### 9) Use K-mean regression to train regression models multiple k values
(K=2,4,6,8, and 10)
knn<-train(castest,castest$TFI_Reduction,method='knn',tuneGrid =
expand.grid(k=c(2,4,6,8,10)))
knn

### 10) Make prediction on the testing data set and obtain MSE.
predict_knn <- predict(object = knn, castest)
MSE.knn <- sqrt(mean(castest$TFI_Reduction-predict_knn)^2)
MSE.knn

### 11) Select the best k that gives the lowest mean square error
knn$bestTune

### 12) Compare MSE in 11) to the MSE of the MLR computed in 8). Which method
gives the lowest MSE?


savehistory()
```

## ***References***

Choueiry, George. “Understand Best Subset Selection.” *Quantifying Health*, 2021, quantifyinghealth.com/best-subset-selection.

“Descriptive Statistics in R.” *Stats and R*, 22 Jan. 2020, statsandr.com/blog/descriptive-statistics-in-r.

*An Introduction to Statistical Learning: With Applications in R*. Springer, 2017.

“Quick-R: Pie Charts.” *Quick R by Data Camp*, 2017, www.statmethods.net/graphs/pie.html.

Taiyun, Wei, and Simko Viliam. “An Introduction to Corrplot Package.” *CRAN Project*, 30 June 2021, cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html.