The University of Texas Rio Grande Valley

School of Mathematical and Statistical Science

# Statistical Learning (MATH 6392)

## Case Study 1

**Name:** Sergio Soto Quintero

**UTRGV ID:** 0375494

Summer II / August 2020

## 1. Introduction

Tinnitus is a conscious awareness of a sound in the ears or head that is not due to an external noise. People who experience this symptom described the noise as a roaring, buzzing, hissing or whizzing noise. Tinnitus affect 10-15% of the world's population; and is especially common in people over age 55 as well as it is associated with hearing loss [1]. A significant number of people who experience tinnitus symptoms do become severely distressed by the sounds as it significantly affect quality of life. Unfortunately, there is no cure for chronic tinnitus, but there are several ways to help tune out the noise and minimize its impact. The aim of this study is to investigate the outcome of a new treatment for Tinnitus and its influence on the person's quality of life.

## 2. Literature Review

### *Causes*

A common cause of tinnitus is inner ear hair cell damage. Tiny, delicate hairs in your inner ear move in relation to the pressure of sound waves. This triggers cells to release an electrical signal through a nerve from your ear (auditory nerve) to your brain. Your brain interprets these signals as sound. If the hairs inside your inner ear are bent or broken, they can "leak" random electrical impulses to your brain, causing tinnitus [2].

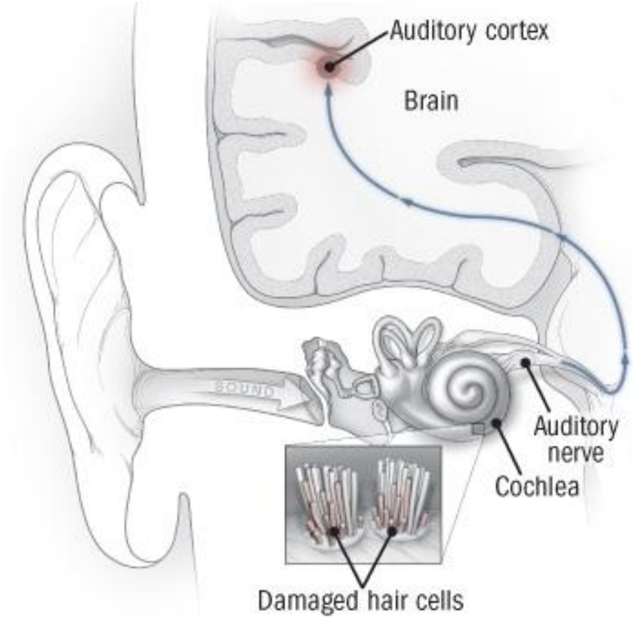


*Fig. 1 Sound waves travel through the ear canal to the middle and inner ear, where hair cells in part of the cochlea help transform sound waves into electrical signals that then travel to the brain's auditory cortex via the auditory nerve. When hair cells are damaged — by loud noise or ototoxic drugs, for example — the circuits in the brain don't receive the signals they're expecting. This stimulates abnormal activity in the neurons, which results in the illusion of sound, or tinnitus. [1]*

### ***Risk factors and complications***

Anyone can experience tinnitus, but some factors may increase your risk of experiencing it. For example:

- Loud noise exposure. Prolonged exposure to loud noise can damage the hair cells in your ear. People who work in noisy environments, like factories, or construction workers, musicians, and soldiers are particularly at risk.
- Age. The number of functioning nerve fibers in your ears decline as you age.
- Sex. Men are more likely to experience tinnitus.
- Smoking. Smokers have higher risk of developing tinnitus.
- Cardiovascular problems. Conditions that affect blood flow, such as blood pressure or narrowed arteries.

Although, tinnitus can affect people in different ways, the most common side effects or complications include fatigue, stress, sleep problems, trouble concentrating, memory problems, depression, anxiety and irritability.

Because of these complications, several studies have been done and found out a strong association between hearing loss (which is a strong predictor of tinnitus) and mental illnesses [3].

### ***Preventions***

In many cases, tinnitus is the result that something that can’t be prevented. However, some precautions can help prevent certain kind of tinnitus. For example, using hearing protection equipment (especially if working in a industry that uses loud machinery or firearms), turn down the volume (especially if exposed to amplified music or thought headphones), and taking care of your cardiovascular health, as healthy blood vessels can prevent tinnitus linked to blood vessels disorders [2].

### ***Treatment***

There is currently no scientifically proven cure for chronic tinnitus. The search for a definitive cure is ongoing and real progress is being made, but there is currently no clinically proven way to fully eliminate the perception of tinnitus [4]. In fact the most effective approaches are behavioral strategies and sound-generating devices, often used in combination. [1]

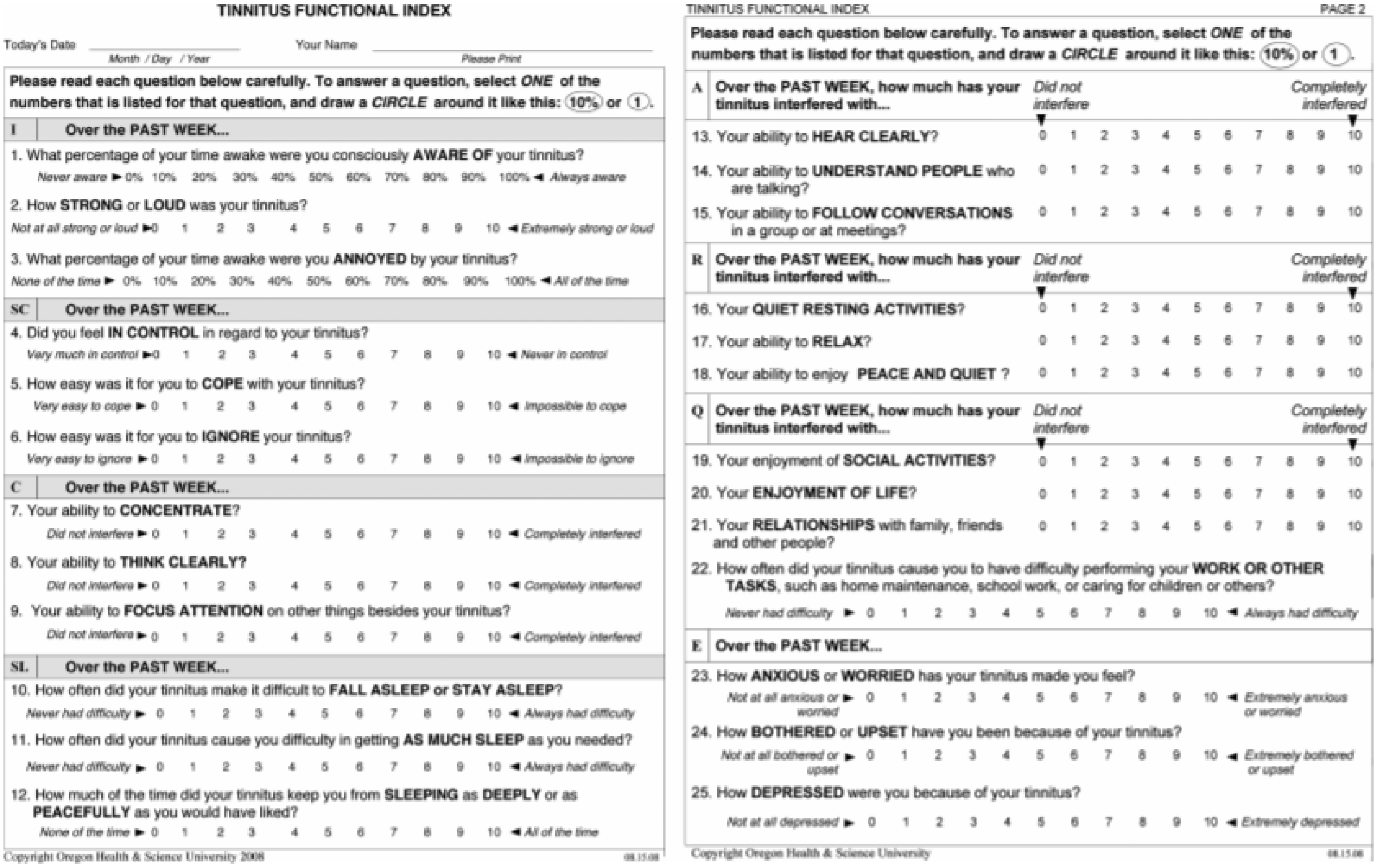
There are, however, excellent tools to help patients manage their condition; treatments that reduce the perceived intensity, omnipresence, and burden of tinnitus. These currently available treatments are not “cures”, they neither repair the underlying causes of tinnitus, nor eliminate the tinnitus signal in the brain. Instead, they address the attentional, emotional, and cognitive impact of tinnitus. They help patients live better, more fulfilling, and more productive lives, even if the perception of tinnitus remains [4].

The treatments that are available, however, receive a lot of criticism, because of the different methods used to assess patients. The challenge medical practitioners faced when assessing the effectiveness of treatments is very straightforward: how can you accurately assess a patient response to a treatment if the symptoms are mostly based on the patient’s perception.

The most effective tinnitus treatment tools address the aspects of tinnitus that so often make the condition feel burdensome: anxiety, stress, social isolation, sound sensitivity, hearing difficulties, and perceived volume.

One of the most used treatments approaches include the use of surveys with questions relating to hearing, attitude, emotion and sleep to gauge the severity, or perceived level of distress from their tinnitus. These questionnaires contain a series of questions from which patients select a response from the given choices which is usually recorded as a grade scale. In this way the severity of the tinnitus is graded. However, some questionnaires were not designed to measure the effectiveness of tinnitus interventions and treatments.

Most recently, in 2012, a group of researchers from the Oregon Health and Science University proposed a new self-report type of questionnaire called the Tinnitus Functional Index (TFI) which scales the severity and negative impact of tinnitus. This new questionnaire emphasizes on the effect sizes, content validly and response scaling which enables the detection of changes in condition of the patients during treatment [5].



**TINNITUS FUNCTIONAL INDEX**

Today's Date ____________ Your Name ____________
*Month / Day / Year* *Please Print*

**Please read each question below carefully. To answer a question, select *ONE* of the numbers that is listed for that question, and draw a *CIRCLE* around it like this: 10% or 1.**

I **Over the PAST WEEK...**

1. What percentage of your time awake were you consciously **AWARE OF** your tinnitus?
*Never aware* ► 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% ◄ *Always aware*

2. How **STRONG** or **LOUD** was your tinnitus?
*Not at all strong or loud* ►0 1 2 3 4 5 6 7 8 9 10 ◄ *Extremely strong or loud*

3. What percentage of your time awake were you **ANNOYED** by your tinnitus?
*None of the time* ► 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% ◄ *All of the time*

SC **Over the PAST WEEK...**

4. Did you feel **IN CONTROL** in regard to your tinnitus?
*Very much in control* ►0 1 2 3 4 5 6 7 8 9 10 ◄ *Never in control*

5. How easy was it for you to **COPE** with your tinnitus?
*Very easy to cope* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Impossible to cope*

6. How easy was it for you to **IGNORE** your tinnitus?
*Very easy to ignore* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Impossible to ignore*

C **Over the PAST WEEK...**

7. Your ability to **CONCENTRATE**?
*Did not interfere* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Completely interfered*

8. Your ability to **THINK CLEARLY?**
*Did not interfere* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Completely interfered*

9. Your ability to **FOCUS ATTENTION** on other things besides your tinnitus?
*Did not interfere* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Completely interfered*

SL **Over the PAST WEEK...**

10. How often did your tinnitus make it difficult to **FALL ASLEEP or STAY ASLEEP**?
*Never had difficulty* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Always had difficulty*

11. How often did your tinnitus cause you difficulty in getting **AS MUCH SLEEP** as you needed?
*Never had difficulty* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Always had difficulty*

12. How much of the time did your tinnitus keep you from **SLEEPING** as **DEEPLY** or as **PEACEFULLY** as you would have liked?
*None of the time* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *All of the time*

Copyright Oregon Health & Science University 2008 08.15.08

TINNITUS FUNCTIONAL INDEX PAGE 2

**Please read each question below carefully. To answer a question, select *ONE* of the numbers that is listed for that question, and draw a *CIRCLE* around it like this: 10% or 1.**

A **Over the PAST WEEK, how much has your tinnitus interfered with...** *Did not interfere* ▼ *Completely interfered* ▼

13. Your ability to **HEAR CLEARLY**? 0 1 2 3 4 5 6 7 8 9 10

14. Your ability to **UNDERSTAND PEOPLE** who are talking? 0 1 2 3 4 5 6 7 8 9 10

15. Your ability to **FOLLOW CONVERSATIONS** in a group or at meetings? 0 1 2 3 4 5 6 7 8 9 10

R **Over the PAST WEEK, how much has your tinnitus interfered with...** *Did not interfere* ▼ *Completely interfered* ▼

16. Your **QUIET RESTING ACTIVITIES**? 0 1 2 3 4 5 6 7 8 9 10

17. Your ability to **RELAX**? 0 1 2 3 4 5 6 7 8 9 10

18. Your ability to enjoy **PEACE AND QUIET** ? 0 1 2 3 4 5 6 7 8 9 10

Q **Over the PAST WEEK, how much has your tinnitus interfered with...** *Did not interfere* ▼ *Completely interfered* ▼

19. Your enjoyment of **SOCIAL ACTIVITIES**? 0 1 2 3 4 5 6 7 8 9 10

20. Your **ENJOYMENT OF LIFE**? 0 1 2 3 4 5 6 7 8 9 10

21. Your **RELATIONSHIPS** with family, friends and other people? 0 1 2 3 4 5 6 7 8 9 10

22. How often did your tinnitus cause you to have difficulty performing your **WORK OR OTHER TASKS**, such as home maintenance, school work, or caring for children or others?
*Never had difficulty* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Always had difficulty*

E **Over the PAST WEEK...**

23. How **ANXIOUS** or **WORRIED** has your tinnitus made you feel?
*Not at all anxious or worried* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Extremely anxious or worried*

24. How **BOTHERED** or **UPSET** have you been because of your tinnitus?
*Not at all bothered or upset* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Extremely bothered or upset*

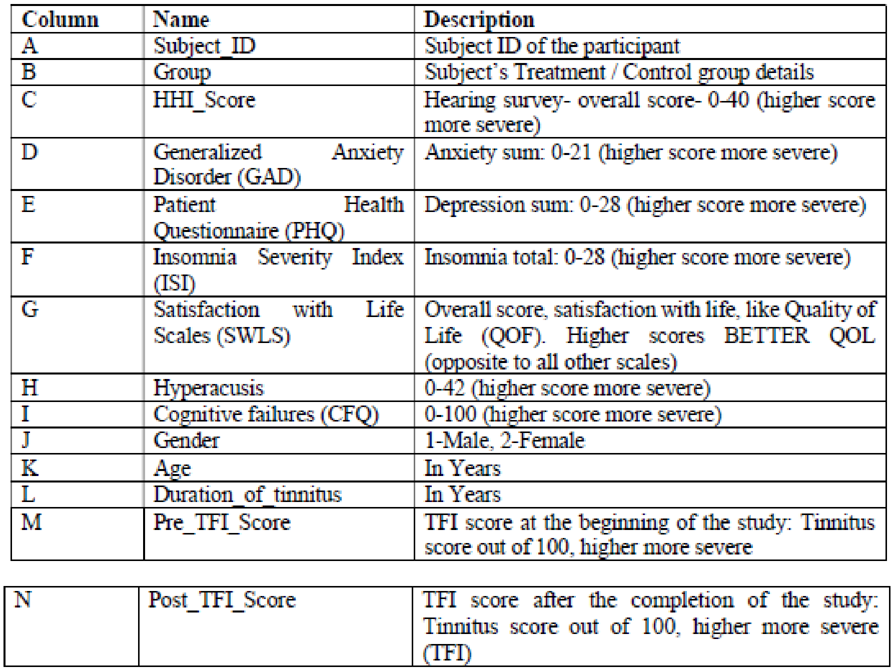
25. How **DEPRESSED** were you because of your tinnitus?
*Not at all depressed* ► 0 1 2 3 4 5 6 7 8 9 10 ◄ *Extremely depressed*

Copyright Oregon Health & Science University 08.15.08

*Fig. 2 Final version of the TFI questionnaire proposed by a group of researches from the Oregon Health and Science University in 2012.*

## 3. Methods and Analysis

In this case study we will investigate the effectiveness of an internet based cognitive behavioral therapy (referred as the *Treatment* in the data set) intervention which has been developed in UK to improve the access to evidence-based tinnitus treatment. The Tinnitus Functional Index (**TFI_Scores**) has been used as the primary assessment measure to quantify tinnitus distress prior and after the treatment. We will be performing a Multiple Linear Regression and a KNN Regression Analysis on a dataset ("CaseStudy1.csv)" containing the **pre** and **post TFI_Scores**, and clinical and demographic factors related to a pre-post interventional study of 142 subjects with tinnitus. For the computations we will be using R. The specific column details of the data set are the following:



| Column | Name | Description |
|---|---|---|
| A | Subject_ID | Subject ID of the participant |
| B | Group | Subject's Treatment / Control group details |
| C | HHI_Score | Hearing survey- overall score- 0-40 (higher score more severe) |
| D | Generalized Anxiety Disorder (GAD) | Anxiety sum: 0-21 (higher score more severe) |
| E | Patient Health Questionnaire (PHQ) | Depression sum: 0-28 (higher score more severe) |
| F | Insomnia Severity Index (ISI) | Insomnia total: 0-28 (higher score more severe) |
| G | Satisfaction with Life Scales (SWLS) | Overall score, satisfaction with life, like Quality of Life (QOF). Higher scores BETTER QOL (opposite to all other scales) |
| H | Hyperacusis | 0-42 (higher score more severe) |
| I | Cognitive failures (CFQ) | 0-100 (higher score more severe) |
| J | Gender | 1-Male, 2-Female |
| K | Age | In Years |
| L | Duration_of_tinnitus | In Years |
| M | Pre_TFI_Score | TFI score at the beginning of the study: Tinnitus score out of 100, higher more severe |
| N | Post_TFI_Score | TFI score after the completion of the study: Tinnitus score out of 100, higher more severe (TFI) |

*Table 1. Complete list of variables for the data set "CaseStudy1.csv"*

We begin our analysis by looking at the data gathered. From the 142 surveyed patients,  73 were in the control group and 69 were in the treatment group. Also, 80 identified themselves as males and 62 as females.

Control
Treatment
Male
Female

*Fig. 3 Pie charts for Control vs Treatment and Male vs Female*

In Fig. 4, we can observe the different values of TFI: before treatment (left plot), after treatment (middle) and the difference between the two (right plot). These graphs were obtained after the initial data cleaning. The dataset was missing 86 values from the **Post_TFI_Score** predictor and to fix this situation we proceed to a data imputation with mean. Also, the variable **TFI_Reduction** was created and added to the dataset to serve as the variable response for the upcoming analysis.

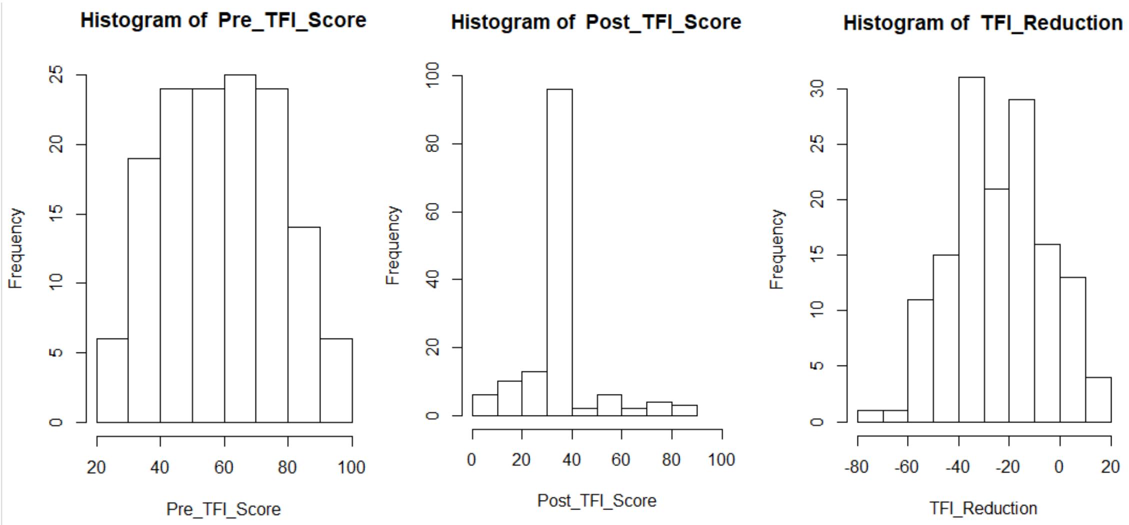
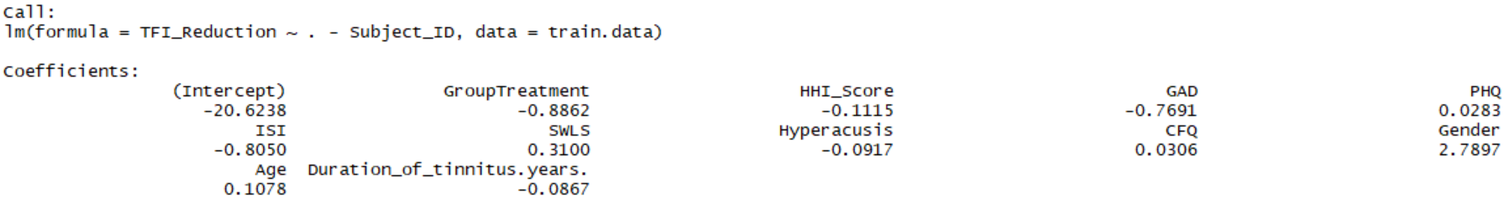


*Fig. 4. Perceibable values of TFI from patients before treatment, after treatment and their difference.*

## Multiple Regression Analysis

To proceed with the Multiple Regression Analysis mix the dataset, remove the **Pre_TFI_Score**, **Post_TFI_Score** and **Subject_ID** predictors from the analysis and partition the dataset using the caret package. The partition of the data splits the 142 observation into two subgroups randomly. One group of observations represent the 80% and it represents the training set. The remaining 20% becomes the subgroup called the testing set.

We perform the Multiple Linear Regression Analysis on the training set (80% observations) using variables B,C,D,E,F,G,H,I,J,K,L as the predictors (see Table.1) and **TFI_Reduction** as the response to the model. We obtain the following coefficient results:



```
Call:
lm(formula = TFI_Reduction ~ . - Subject_ID, data = train.data)

Coefficients:
          (Intercept)              GroupTreatment                  HHI_Score                        GAD                        PHQ
             -20.6238                     -0.8862                    -0.1115                    -0.7691                     0.0283
                  ISI                        SWLS                Hyperacusis                        CFQ                     Gender
              -0.8050                      0.3100                    -0.0917                     0.0306                     2.7897
                  Age   Duration_of_tinnitus.years.
               0.1078                     -0.0867
```

*Table 2 Coefficients for the Multiple Linear Regression Analysis.*

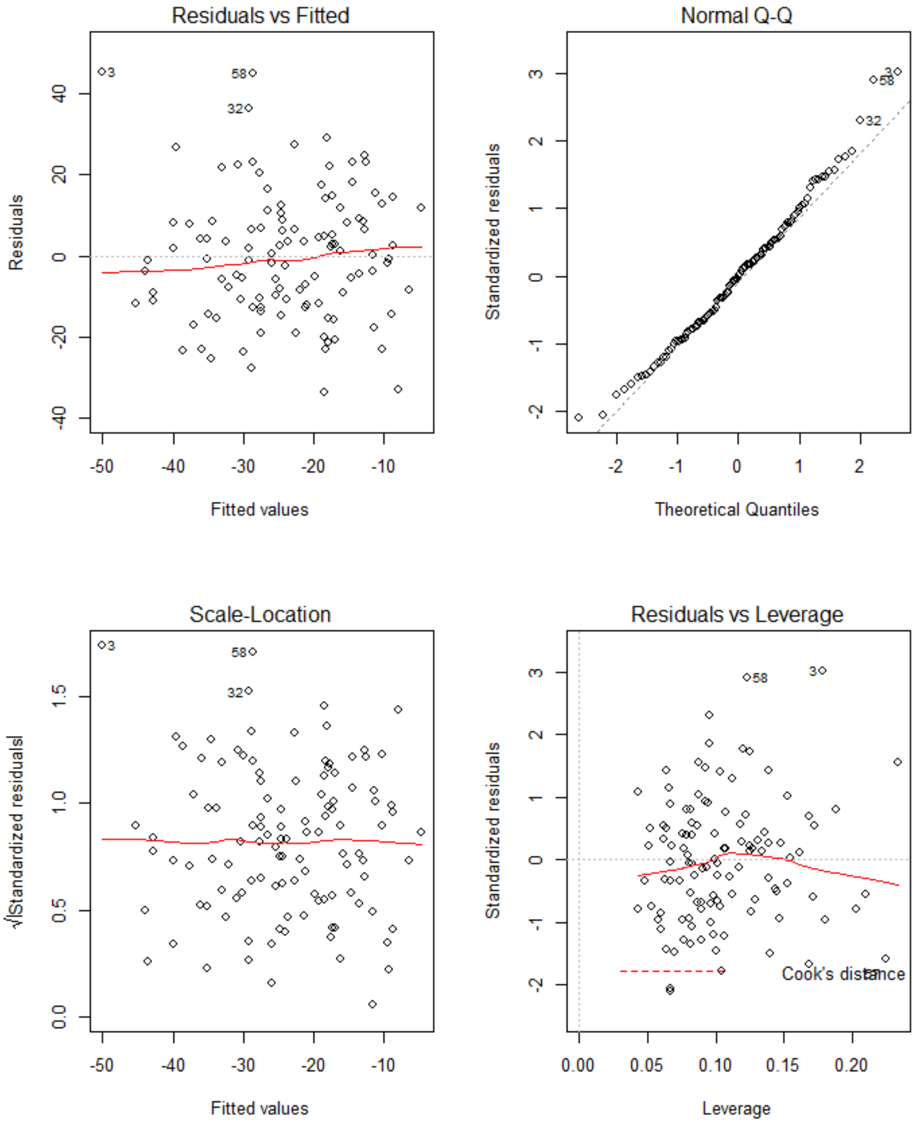


*Fig. 5. Plots of Residuals vs Fitted (or predicted) values (top and bottom left) show little pattern in the residuals, which is ideally what we want for the linear model. The linearity of the normal Q-Q plot (top right) suggests the data is normally distributed. Observation 3 seems to be a high leverage point. In the Residual vs Leverage (bottom right) plot we see observation 3 and 58 are three standard deviations away from the mean.*

We will leave observations 3 and 58 in our analysis since they are a natural part of the population we are studying. The observations are well within the range of each individual's predictor's values, but they are an unusual combination in term of the full set of predictors. Ideally, we would have to perform a series of analysis on the impact of the outliners (leverage statistic) to figure out the true impact on the least square fit.

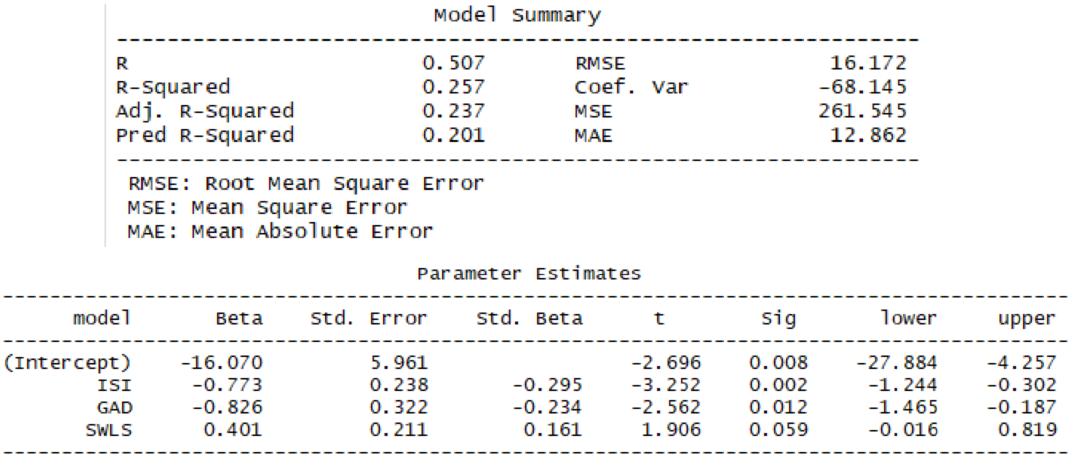
We will now use an alternative model to least square fit, because our dataset contains p=10 predictors. A model with a high number of variables might include irrelevant predictors which can lead to a needlessly complex model. The purpose of using another fitting procedure is to yield a better prediction accuracy and model interpretability [6].

## Forward and Backward Selection

The analysis of forward and backward selections yield the same result. The best model for both approaches contain predictors **ISI** , **GAD** and **SWLS** which represent the factors that highly influence the reduction in TFI**_Score**. In mathematical terms, the predicted response, $\boldsymbol{\hat{y}}$ **(**or **TFI_Response)** that best fits the model according to forward and backward selection is:

$$\boldsymbol{\hat{y}} = \widehat{\beta_0} + \widehat{\beta_1}(\boldsymbol{ISI}) + \widehat{\beta_2}(\boldsymbol{GAD}) + \widehat{\beta_3}(\boldsymbol{SWLS}) \qquad (1)$$

Where $\widehat{\beta_0}$ is the predicted y-intercept of the model, $\widehat{\beta_1}$ is the predicted coefficient of **ISI**, $\widehat{\beta_2}$ is the predicted coefficient of **GAD** and $\widehat{\beta_3}$ is the predicted coefficient of **SWLS**.



```
                        Model Summary
-------------------------------------------------------------
R                       0.507       RMSE               16.172
R-Squared               0.257       Coef. Var         -68.145
Adj. R-Squared          0.237       MSE               261.545
Pred R-Squared          0.201       MAE                12.862
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              Parameter Estimates
-----------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta       t      Sig      lower    upper
-----------------------------------------------------------------------------------
(Intercept)    -16.070         5.961                -2.696    0.008   -27.884   -4.257
        ISI     -0.773         0.238       -0.295   -3.252    0.002    -1.244   -0.302
        GAD     -0.826         0.322       -0.234   -2.562    0.012    -1.465   -0.187
       SWLS      0.401         0.211        0.161    1.906    0.059    -0.016    0.819
-----------------------------------------------------------------------------------
```

*Table 3 Model Summary and Parameter Estimates of Forward and Backward Selection alike. Both methods yielded the same results.*

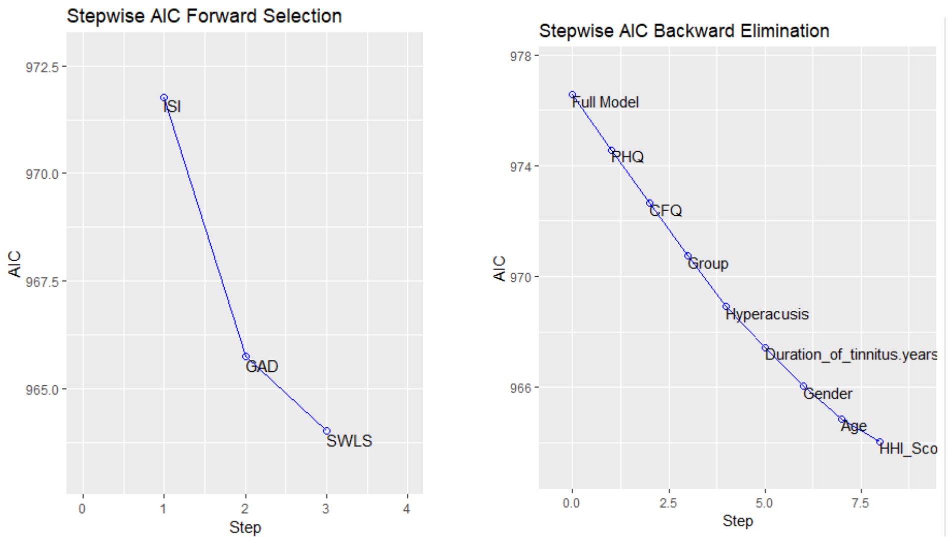
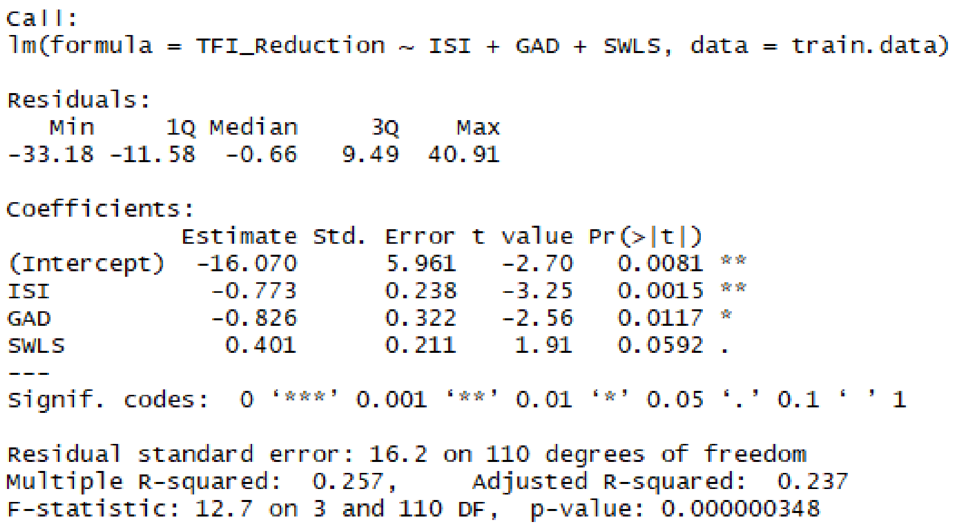


*Fig.6 AIC vs. Steps in forward and backward elimination respectively.*

Now, that we know which are the predictors that highly influence the reduction in **TFI_Score**, we perform and linear regression analysis considering only the predictors: **ISI** , **GAD** and **SWLS.**

The summary of the linear regression model and diagnostic graphs are shown below.



```
Call:
lm(formula = TFI_Reduction ~ ISI + GAD + SWLS, data = train.data)

Residuals:
   Min     1Q Median     3Q    Max 
-33.18 -11.58  -0.66   9.49  40.91 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -16.070      5.961   -2.70   0.0081 **
ISI           -0.773      0.238   -3.25   0.0015 **
GAD           -0.826      0.322   -2.56   0.0117 * 
SWLS           0.401      0.211    1.91   0.0592 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.2 on 110 degrees of freedom
Multiple R-squared:  0.257,	Adjusted R-squared:  0.237 
F-statistic: 12.7 on 3 and 110 DF,  p-value: 0.000000348
```

*Table 4. Summary of the linear regression including ISI,GAD and SWLS as the only predictors and TFI_Score as the response; p- values (below 0.05) show high correlation between predictors and response.*

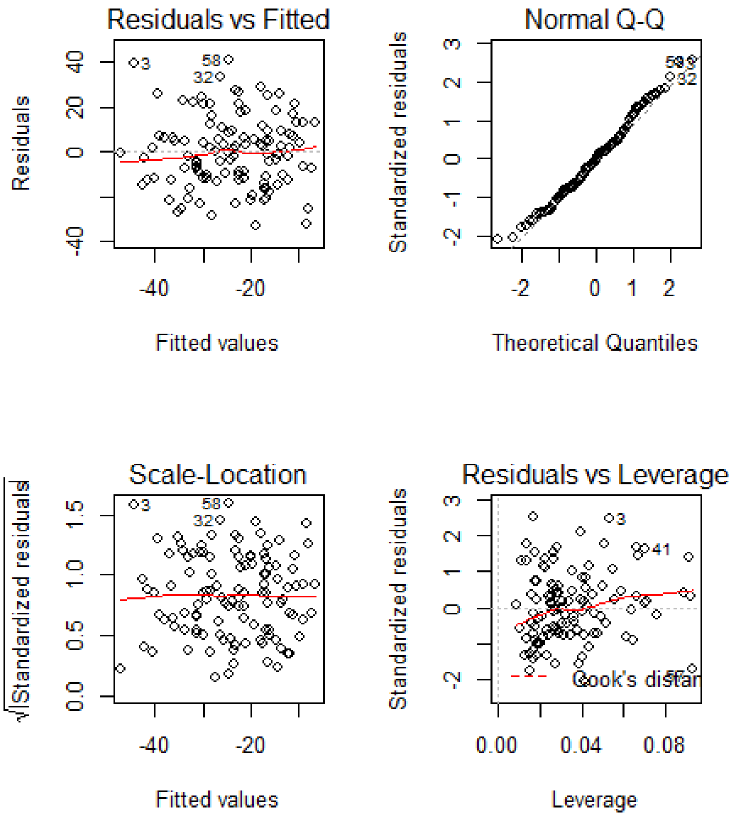


*Fig. 7. Graphs show a very similar outcome as the graphs from Fig. 5.*

Then, we make predictions on the test dataset (20% observations) to compare how accurate was the proposed model that we found using the training dataset (80% observations). To compare both models we run a Mean Square Error (MSE) calculations on both datasets.

*Multiple Linear Regression MSE:*

$$MSE_{training} = 252.4$$

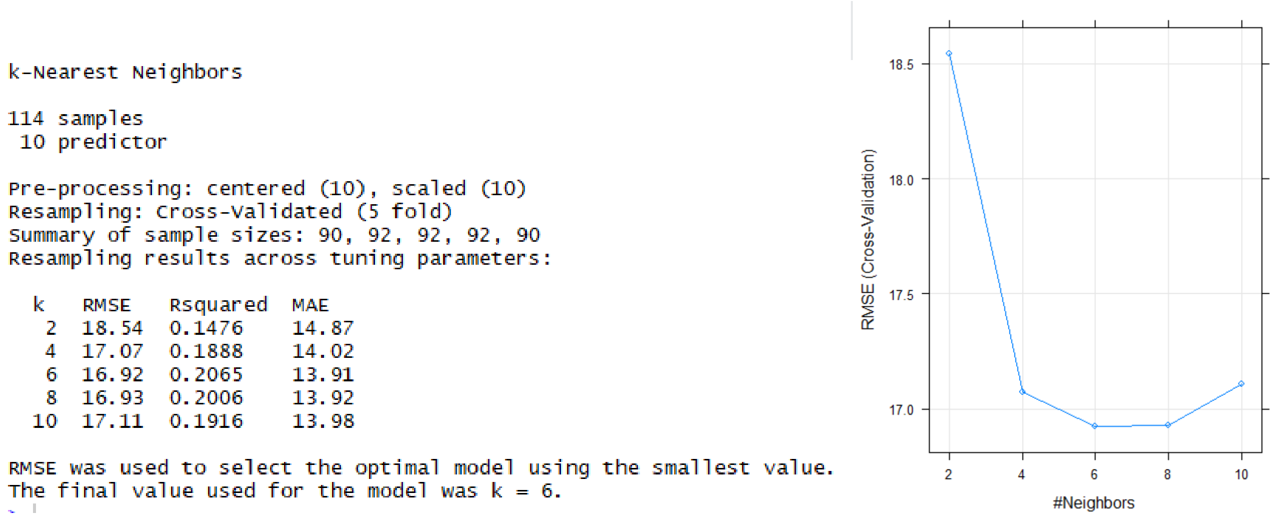
$$MSE_{testing} = 256$$

## K- Mean Regression

We will now perform a K-Mean Regression analysis using the *caret* package in R. For the analysis, we first perform the training of the model by using the same training dataset (80%) that we used for the Multiple Regression Analysis. The predictors used for this analysis included C,D,E,F,G,H,I,J,K,L (Table.1) and **TFI_Reduction** for the response.

We run a KNN regression analysis for different values of k, which represent the number of nearest neighbor's labels used to assign a label to the current point. We perform the analysis for k = 2,4,6,8,10.

The K-nearest neighbors analysis tells us that for our particular sample, k = 6 results in the lowest RMSE (Root Mean Square Error), so we will use this to run a prediction on the testing dataset.



```
k-Nearest Neighbors

114 samples
 10 predictor

Pre-processing: centered (10), scaled (10)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 90, 92, 92, 92, 90
Resampling results across tuning parameters:

  k   RMSE   Rsquared  MAE
   2  18.54  0.1476    14.87
   4  17.07  0.1888    14.02
   6  16.92  0.2065    13.91
   8  16.93  0.2006    13.92
  10  17.11  0.1916    13.98

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 6.
```

*Fig. 8. KNN analysis for k values of 2,4,6,8,10. Graph shows that k=6 and k=8 have pretty much the same RMSE. Summary (left) shows that RMSE for k=6 is 16.92 and RMSE for k=8 is 16.93.*

We run the prediction on the testing dataset (20%) and calculate the MSE for the training and testing dataset.

*K-Nearest Neighbors MSE:*

$$MSE_{training} = 286.4$$

$$MSE_{testing} = 306.2$$

## 4. Conclusion

On the basis of the findings, several conclusions concerning the interaction between a reduction in TFI and some quality life indicators can be drawn. The findings of this study suggest that the Multiple Regression Analysis results in a more accurate model than the KNN Regression Analysis by comparing both model's MSE. The results of the Multiple Regression Analysis indicate strong correlation of three particular variables with the reduction of TFI. These variables are the Insomnia Severity Index (ISM), the Generalized Anxiety Disorder (GAD), and the Satisfaction with Life Scales (SWLS). The study suggest than by focusing in therapies and treatments for variables mentioned above will translate in a reduction of TFI and a better quality of life for the patient.

5. Appendix

## Appendix A. References

[1] Harvard Health Publishing Harvard Medical School. (2020, April 8). *Tinnitus: Ringing in the ears and what to do about it.* Retrieved June 29, 2020, from https://www.health.harvard.edu/diseases-and-conditions/tinnitus-ringing-in-the-ears-and-what-to-do-about-it

[2] Mayo Foundation for Medical Education and Research (MFMER). (n.d.). *Tinnitus*. Retrieved June 29th, 2020, from https://www.mayoclinic.org/diseases-conditions/tinnitus/symptoms-causes/syc-20350156

[3] Adeyi A. Adoga and Taiwo J. Obindo (January 16th 2013). *The Association Between Tinnitus and Mental Illnesses*, Mental Disorders - Theoretical and Empirical Perspectives, Robert Woolfolk and Lesley Allen, IntechOpen, DOI: 10.5772/52755. Available from: https://www.intechopen.com/books/mental-disorders-theoretical-and-empirical-perspectives/the-association-between-tinnitus-and-mental-illnesses

[4] American Tinnitus Association. (n.d.). *Managing your tinnitus.* Retrieved June 29th, 2020, from https://www.ata.org/managing-your-tinnitus/treatment-options

[5] Meikle MB, Stewart BJ, Griest SE, Henry JA.Tinnitus Outcomes Assessments, Trends, Amplif..2008; 12: 223-235.

[6] G. James et al., *An Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7, Springer Science+Business Media New York 2013.

## Appendix B. Code

```r
# clear plots
if(!is.null(dev.list())) dev.off()

# clear console
cat("\014")

# clean workspace
rm(list=ls())

setwd("C:/Users/Checo/Desktop/Statistical Learning/Case Study")
getwd()

#library(RColorBrewer)
tin.rdata = read.csv("CaseStudy1.csv",header=T)
fix(tin.rdata)

#### Multiple Linear Regression Analysis
## 1 - Descriptive Analysis of the data

# Data cleaning - Combining levels (the categories) in the "Group" predictor
levels(tin.rdata$Group) = c("Control","Control", "Treatment")

# Pie chart
pie.GROUP = pie(table(tin.rdata$Group), labels=c('Control', 'Treatment'))
pie.GENDER = pie(table(tin.rdata$Gender), labels=c('Male','Female'))

#Histograms
hist.GROUP= plot(as.factor(tin.rdata$Group),xlab = names(tin.rdata)[2], main =
paste("Histogram of ",names(tin.rdata)[2]), ylim=c(0,80) )
hist.HHI  = hist(as.numeric(tin.rdata[,3]), xlab = names(tin.rdata)[3], main =
paste("Histogram of ",names(tin.rdata)[3]) )
hist.GAD = hist(as.numeric(tin.rdata[,4]), xlab = names(tin.rdata)[4], main =
paste("Histogram of ",names(tin.rdata)[4]) )
hist.PHQ= hist(as.numeric(tin.rdata[,5]), xlab = names(tin.rdata)[5], main =
paste("Histogram of ",names(tin.rdata)[5]),ylim=c(0,60))
hist.ISI= hist(as.numeric(tin.rdata[,6]), xlab = names(tin.rdata)[6], main =
paste("Histogram of ",names(tin.rdata)[6]))
hist.SWLS= hist(as.numeric(tin.rdata[,7]), xlab = names(tin.rdata)[7], main =
paste("Histogram of ",names(tin.rdata)[7]),ylim=c(0,40))
hist.HYP= hist(as.numeric(tin.rdata[,8]), xlab = names(tin.rdata)[8], main =
paste("Histogram of ",names(tin.rdata)[8]),ylim=c(0,40))
hist.CFQ= hist(as.numeric(tin.rdata[,9]), xlab = names(tin.rdata)[9], main =
paste("Histogram of ",names(tin.rdata)[9]),ylim=c(0,40))
hist.GENDER= plot(as.factor(tin.rdata$Gender),xlab =
paste(names(tin.rdata)[10],"(1 = Man , 2 = Female)"), main = paste("Histogram
of ",names(tin.rdata)[10]) )
hist.AGE= hist(as.numeric(tin.rdata[,11]), xlab = names(tin.rdata)[11], main =
paste("Histogram of ",names(tin.rdata)[11]),ylim=c(0,35))
hist.DURATION= hist(as.numeric(tin.rdata[,12]), xlab = names(tin.rdata)[12],
main = paste("Histogram of ",names(tin.rdata)[12]))
hist.PRE= hist(as.numeric(tin.rdata[,13]), xlab = names(tin.rdata)[13], main =
paste("Histogram of ",names(tin.rdata)[13]))

#Summary
summary(tin.rdata)
```

```
## 2 - Data cleaning

# Data imputation with mean for Pre and Post TFI scores
tin.rdata$Post_TFI_Score[is.na(tin.rdata$Post_TFI_Score)] =
mean(tin.rdata$Post_TFI_Score, na.rm = TRUE)

# Create response variable "TFI_Reduction"
TFI_Reduction = tin.rdata$Post_TFI_Score - tin.rdata$Pre_TFI_Score
tin.rdata$TFI_Reduction = TFI_Reduction
View(tin.rdata)

# Histogram of Post_TFI_Score and TFI_Reduction
hist.POST= hist(as.numeric(tin.rdata[,14]), xlab = names(tin.rdata)[14], main =
paste("Histogram of ",names(tin.rdata)[14]),ylim=c(0,100), xlim=c(0,100))
hist.RED= hist(as.numeric(tin.rdata[,15]), xlab = names(tin.rdata)[15], main =
paste("Histogram of ",names(tin.rdata)[15]))
par(mfrow=c(1,1))

## 3
# Data imputation for numerical values using mean and data imputation for
categorical values using mode

## 4 - Partition the data set

# Mix Data Set
set.seed(123)
mix.data = runif(nrow(tin.rdata)) # Randomly genertes 142 values
tin.rdata_mix = tin.rdata[order(mix.data),]
sset.data1 = subset(tin.rdata_mix,select = -c(Pre_TFI_Score,Post_TFI_Score))
#Removing Pre and Post TFI variables

#Partition the data set into a trainning set (80%)  and test set (20%)
library(caret)
set.seed(123)
train.index =createDataPartition(sset.data1$TFI_Reduction,p=.8,list = F,
times=1)
train.data = sset.data1[train.index,]
test.data = sset.data1[-train.index,]
names(sset.data1)

## 5 - Multiple Regression Analysis
# Remove the "Subject ID" indicator column and perform Multiple regression
analysis
data.fit = lm(TFI_Reduction ~ . -Subject_ID  ,data = train.data)
data.fit
options(scipen="100",digits = "4")
summary(data.fit)

par(mfrow=c(2,2))
plot(data.fit)

# Forward selection model
# Forward Selection (Approach A)
# install.packages("olsrr")
library(olsrr)
fws_a.aic = ols_step_forward_aic(data.fit,details=T)
plot(fws_a.aic)
```

```
#Forward Selection (Approach B)
# fit.start = lm(TFI_Reduction ~ 1 , data = train.data)  # Specify Starting
list of coefficients
# fws_b = step(fit.start, direction ="forward", scope=formula(data.fit))
# summary(fws_b)

# Backward selection model
# Backward selection (Approach A)
bws_a.aic = ols_step_backward_aic(data.fit,details=T)
bws_a.aic
plot(bws_a.aic)
betas = bws_a.aic$model$coefficients
betas

# Backward Selection (Approach B)
# bws_b = step(data.fit, direction ="backward")
# summary(bws_b)

## 6 - Model diagnostic and correcttion of the model assumptions (if needed)
## 7 - Factors that influence the reduction in TFI

train.response = lm(TFI_Reduction ~ ISI+GAD+SWLS , data = train.data )
summary(train.response)
plot(train.response)
summary(influence.measures(train.response))


## 8 - Make predictions on the test data set and calculate the mean square
error
test.response = predict(train.response,newdata = test.data)
test.response
summary(test.response)
plot(test.response)

# Train Model Mean square error
sum((train.data$TFI_Reduction - train.response$fitted)^2)/nrow(train.data)

#Test Model Mean Square Error
sum((test.data$TFI_Reduction - test.response)^2)/nrow(test.data)


#### K-mean Regression
## 9,10,11,12 - K-mean regression to train several regression models (k=
2,4,6,8,10)
set.seed(123)
model.knn = train(train.data[,3:12],train.data[,13], method ="knn" ,
                  trControl=trainControl(method="cv", number =5),
                  preProcess = c("center","scale"),
                  tuneGrid = expand.grid(k=seq(2,10, by =2)))
model.knn
plot(model.knn)
model.knn$bestTune


# Make prediction on the testing data set and find the Mean Square Error
prediction.knn = predict(object=model.knn ,newdata = test.data[,2:12])
prediction.knn
summary(prediction.knn)
if(!is.null(dev.list())) dev.off()
plot(prediction.knn)
```

```r
# Calculate RMSE and MSE
library(Metrics)
rmse = model.knn$results$RMSE
rmse.train = min(rmse)
rmse.train
mse.train = (rmse.train)^2
mse.train

rmse.test = rmse(test.data[,13],prediction.knn)
rmse.test
mse.test = (rmse.test)^2
mse.test
```