THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Class Project II

Intro to Big Data & Analytics
CSCI_6444_80

Team – 9

Bhavya Sree Gudiseva - G41949795
Shejal Shankar - G32395894

Prof. Stephen Kaisler
26th February 2024

Installation of all required packages:

To install the igraph package, you can use the following command in R:

> install.packages("dplyr")
> install.packages("ggplot2")
> install.packages("readr")
> install.packages("tidyr")
> install.packages("caret")
> install.packages("cluster")
> install.packages("corrplot")
> install.packages("glmnet")
> install.packages("rstatix")
> install.packages("gmodels")
> install.packages("psych")
> install.packages("nnet")

Load necessary packages:

```
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)
library(caret)
library(cluster)
library(corrplot)
library(glmnet)
library(rstatix)
library(gmodels)
library(psych)
library(nnet)
```

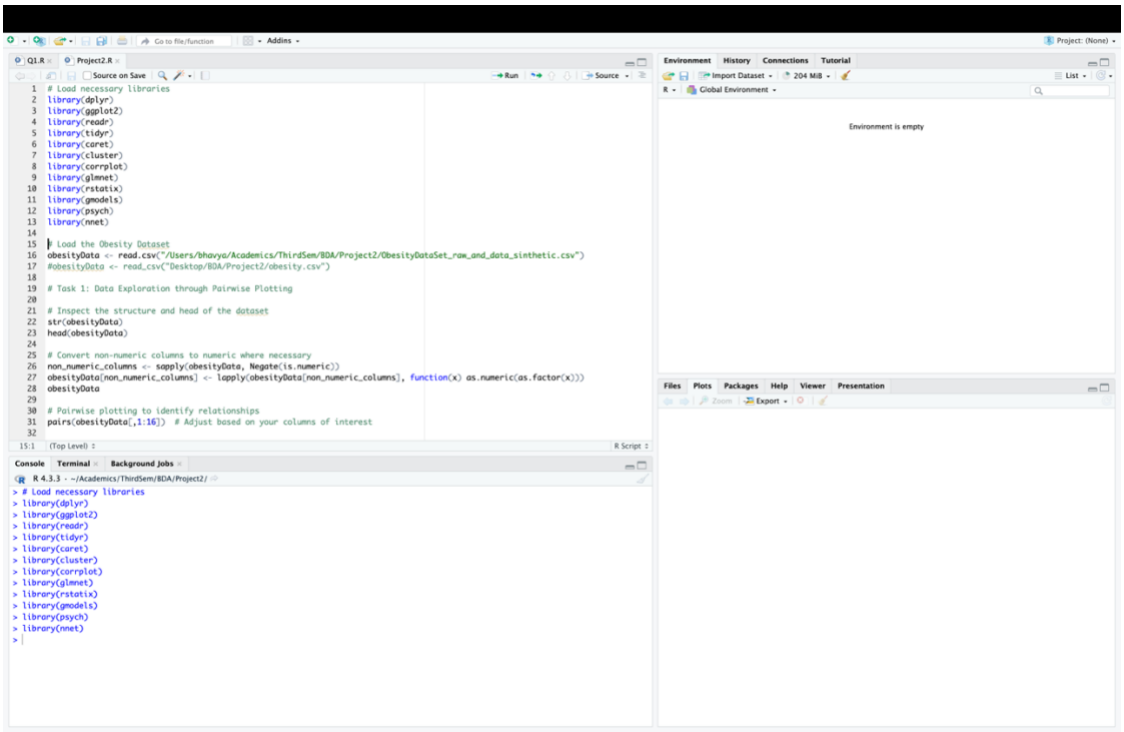

Fig 1 – Loading necessary packages.

The summary provides numerical descriptions of each attribute in the dataset, while the plot is a visual tool. Here's an analysis based on the summary statistics:

1.Gender: Binary variable (1 or 2), possibly indicating male or female. Not a continuous variable, so correlation in the traditional sense isn't applicable.

2.Age: Seems to have a wide range from 14 to 61 years. Age could correlate with features such as weight or obesity level.

3. Height: Ranges from 1.45 to 1.98 meters. Height and weight are often correlated (as height increases, weight also tends to increase, although not always corresponding to obesity levels).

4. Weight: Ranges from 39 to 173 kg. Likely to have a strong correlation with the obesity level (NObeyesdad).

5.Family History with Overweight (family_history_with_overweight): A binary variable, indicating the presence (2) or absence (1) of family history with overweight. This may have a significant correlation with the obesity level.

6. FAVC (Frequency of consumption of high caloric food: Binary variable, can correlate with obesity levels as frequent high-caloric food consumption can lead to higher obesity levels.

7. FCVC (Frequency of consumption of vegetables): This might negatively correlate with obesity levels; higher vegetable consumption could correspond to lower obesity levels.

8.NCP (Number of main meals): Could have a correlation with obesity levels, but not necessarily straightforward without knowing meal content.

9.CAEC (Consumption of food between meals): The higher the consumption between meals, there might be a tendency towards higher obesity levels.

10.SMOKE: Smoking status is a binary variable, and its impact on weight can be complex. It's not traditionally correlated with weight or obesity directly.

11.CH2O (Consumption of water daily): Might not correlate strongly with obesity levels, but it's important for general health.

12.SCC (Calories consumption monitoring): Those who monitor their calorie consumption might have lower obesity levels, suggesting a possible negative correlation.

13.FAF (Physical activity frequency: Likely to be negatively correlated with obesity levels. Higher physical activity generally corresponds to lower obesity levels.

14.TUE (Time using technology devices): Could be positively correlated with obesity levels due to sedentary behaviour.

15.CALC (Consumption of alcohol): The relationship with obesity levels can be complex; not a clear direct correlation.

16.MTRANS (Transportation used): This categorical variable could have some correlation with obesity levels depending on the physical activity involved in the mode of transport.

To address the second part of your question, from the matrix plot, we would look for the following:

Scatter plots shows a clear upward or downward trend, indicating a linear relationship.
Histograms: (diagonal line of plots) that might show skewed distributions, which could suggest potential transformations to normalize the data.

Using both the summary statistics and the matrix plot, you can draw conclusions on which attributes might be most relevant to obesity levels (NObeyesdad). For example, based on common knowledge and the data provided,

attributes like weight, physical activity frequency, and consumption of high caloric food might be strongly related to obesity levels.

Import Dataset:
Now, let's import the specified dataset.
obesityData <- read.csv("/BDA/Project2/ObesityDataSet_raw_and_data_sinthetic.csv")

Task 1: Data Exploration through Pairwise Plotting

# Inspect the structure and head of the dataset
str(obesityData)
head(obesityData)

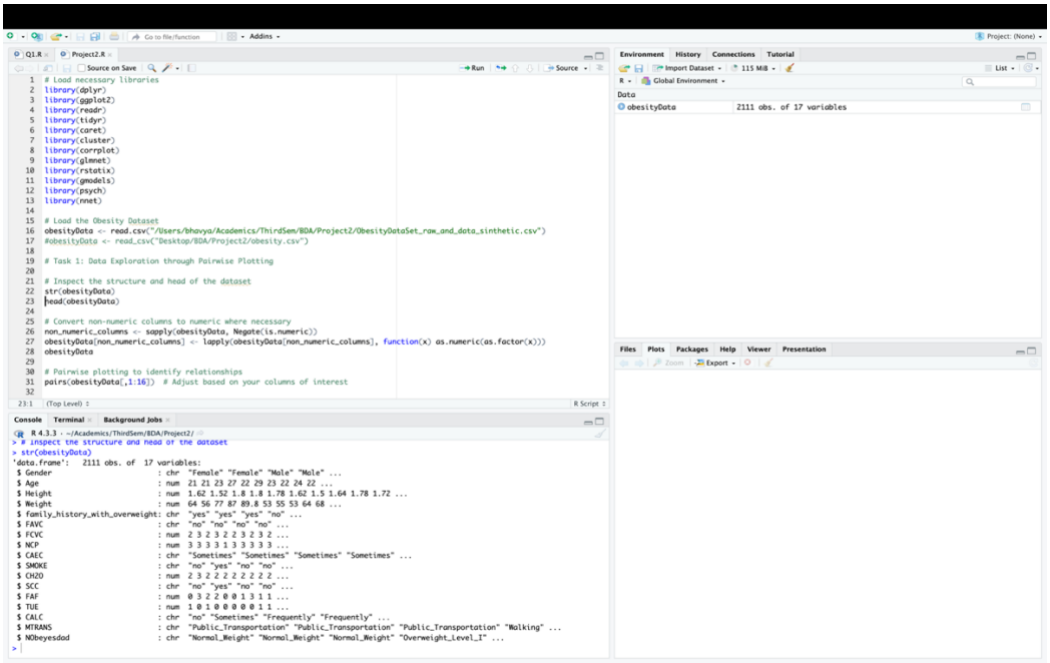

Fig 2 – Read the dataset and view it.

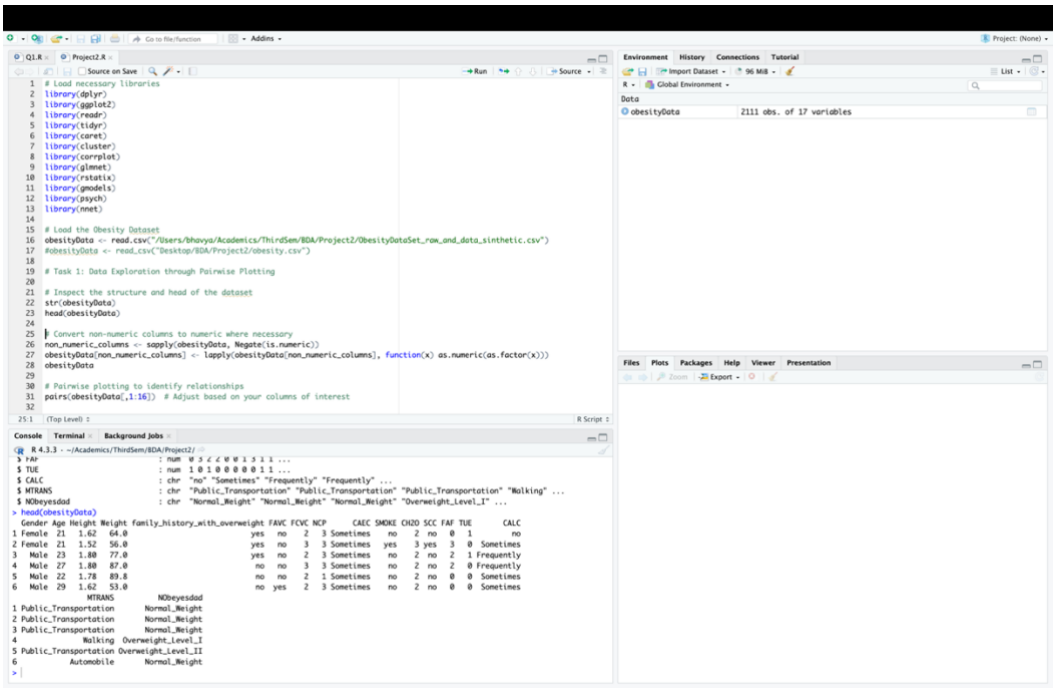

Fig 3 – Data Exploration

```
# Convert non-numeric columns to numeric where necessary
non_numeric_columns <- sapply(obesityData, Negate(is.numeric))
obesityData[non_numeric_columns] <-  lapply(obesityData[non_numeric_columns], function(x)
as.numeric(as.factor(x)))
```

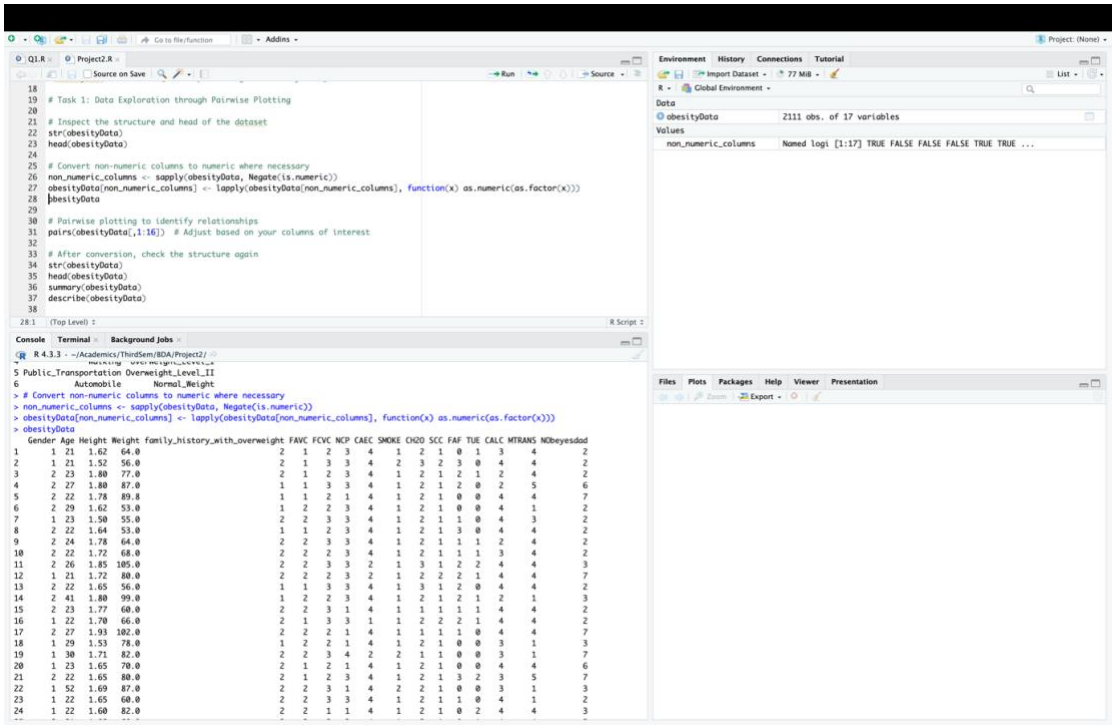

Fig 4 – Preparing the dataset.

```
# Pairwise plotting to identify relationships between the features and the target.
pairs(obesityData)
```

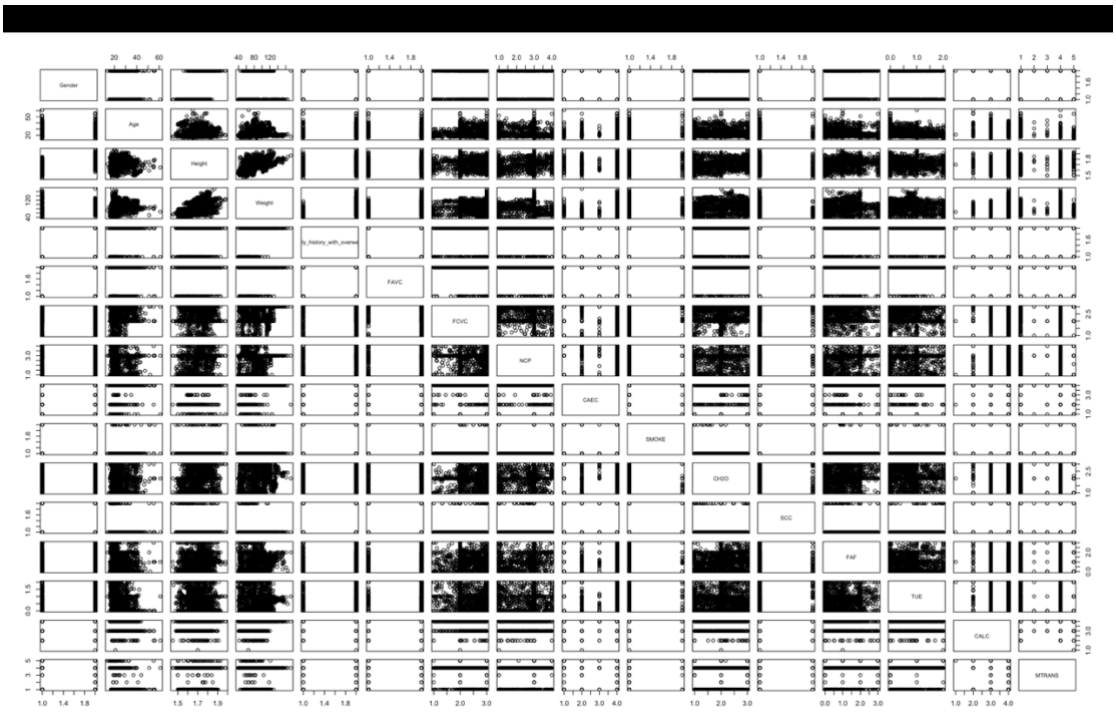

Fig 5 – Pairwise plotting.

```
# After conversion, check the structure again
str(obesityData)
head(obesityData)
summary(obesityData)
describe(obesityData)
```

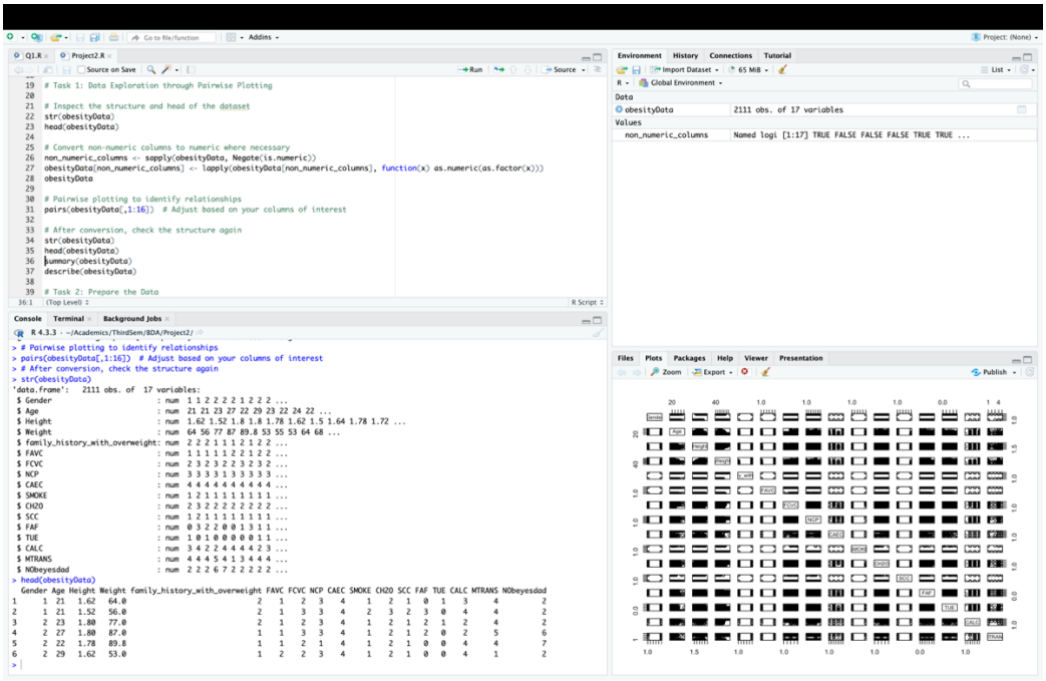

Fig 6 – Structure of data.

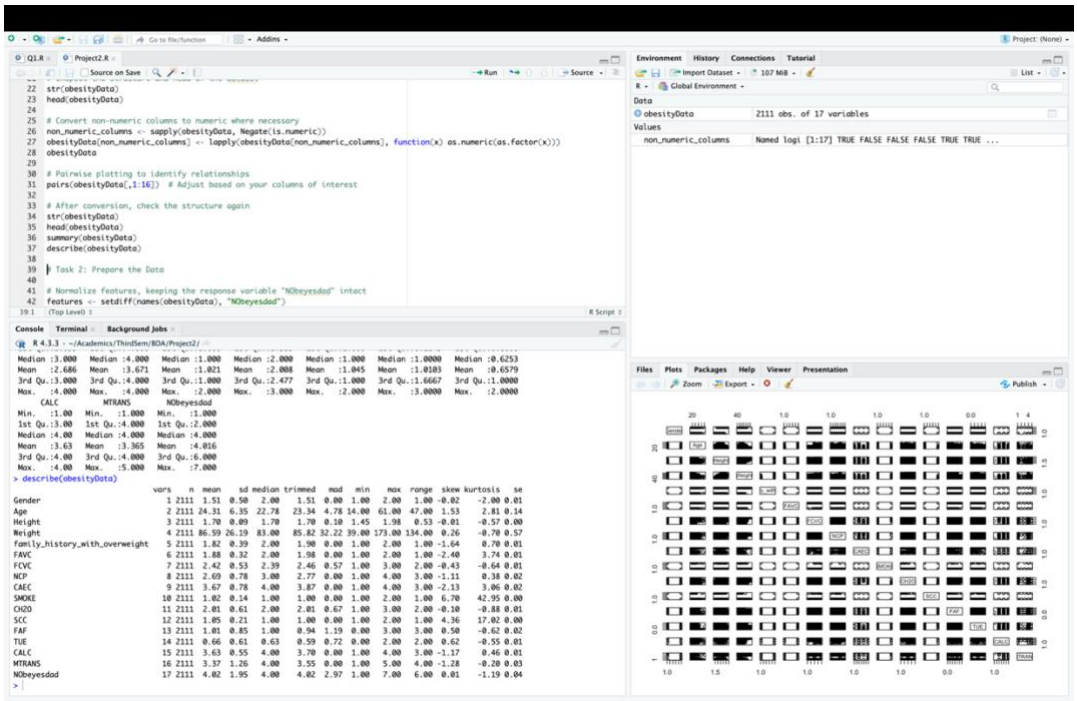

Fig 7 - Summary and description of the dataset.

Based on the summary and describe outputs, here's an analysis of the dataset:

1)Gender: Appears to be binary (1 or 2), likely representing male or female. The mean is close to the median, indicating a fairly even distribution between the categories.

2)Age: Ranges from 14 to 61 years with a mean slightly higher than the median, suggesting a slight skew towards older ages. The standard deviation is moderate, indicating a varied age distribution among the individuals.

3)Height: Fairly normally distributed around a mean of 1.70 meters. There isn't much deviation from the mean, indicating that most individuals' height is close to the average.

4)Weight: The range is quite wide, from 39 to 173 kg. The mean is greater than the median, suggesting a positive skew — there are more individuals with a weight above the median value.

5)Family History with Overweight: Most individuals have a family history of overweight (mode is 2), which might be a significant factor in the study of obesity levels.

6)FAVC (Frequency of consumption of high caloric food): Most individuals frequently consume high caloric food, with a mean close to 2.

7)FCVC (Frequency of consumption of vegetables): On average, individuals consume vegetables more than once but less than three times per day.

8)NCP (Number of main meals): Most individuals consume around three main meals per day, which is typical for many diets.

9)CAEC (Consumption of food between meals): The high mean and median values suggest that most individuals frequently consume food between meals.

10)SMOKE: Most individuals do not smoke, indicated by the mean and median being close to 1.

11)CH2O (Consumption of water daily): Individuals, on average, consume around 2 units of water daily. The measure of skewness suggests a fairly symmetrical distribution around the mean.

12)SCC (Calories consumption monitoring): The majority of the individuals do not monitor their calorie consumption.

13)FAF (Physical activity frequency) The low mean suggests that on average, the frequency of physical activity is low among the individuals.

14)TUE (Time using technology devices): On average, individuals spend a moderate amount of time using technology, with a slight skew towards higher usage

15)CALC (Consumption of alcohol): Most individuals consume alcohol less than three times per week.

16)MTRANS (Transportation used): Most individuals use motorized transportation.

NObeyesdad (Obesity Level): The variable of interest has a mean around 4, which may correspond to "Overweight Level II" if assuming a scale where 1 is "Insufficient Weight" and 7 is "Obesity Type III". The distribution appears to be even across different levels of obesity.

The describe function adds additional details such as skewness and kurtosis, which provide insight into the shape of the distribution for each variable. For example, skewness values far from zero indicate asymmetry in the distribution, while kurtosis values significantly greater than zero suggest heavier tails than a normal distribution. Most variables show skewness close to zero, suggesting a symmetric distribution around the mean. However, variables like SMOKE and SCC show high skewness, indicating that most values cluster around a single value with fewer individuals reporting smoking and calorie counting, respectively.

From these statistics, it's possible to infer which factors may have significant relationships with obesity levels, such as age, weight, family history of overweight, consumption habits, physical activity frequency, and use of technology. These insights could be useful for predictive modelling and understanding the factors associated with obesity.

Task 2: Prepare the Data

```
# Normalize features, keeping the response variable "NObeyesdad" intact

features <- setdiff(names(obesityData), "NObeyesdad")

obesityData.features <- obesityData[features]

normalize <- function(x) {((x-min(x))/(max(x)-min(x)))}

obesityData.features.norm <- as.data.frame(lapply(obesityData.features, normalize))

obesityData.norm <- cbind(obesityData.features.norm, NObeyesdad = obesityData$NObeyesdad)


# Correlation analysis on the normalized features

correlation <- cor(obesityData.features.norm)


# Plot the correlation matrix

corrplot(correlation, method = "circle")
```

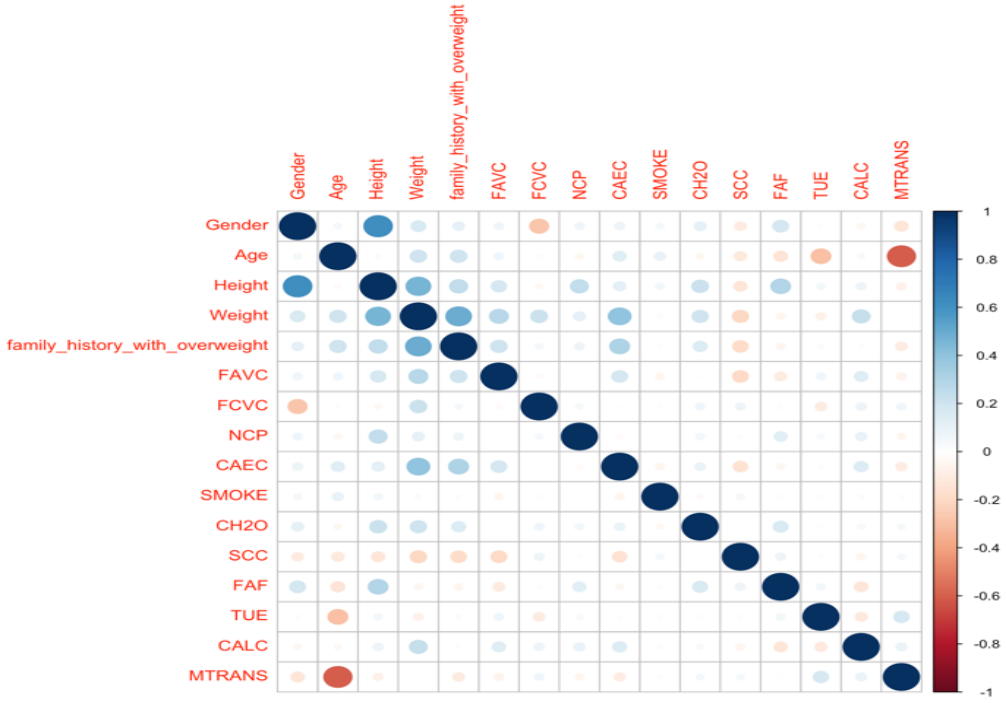

Fig 8 – Corelation matrix

```
# Split the dataset into Training and Test Sets (70-30%, 60-40%, 50-50%)
set.seed(123)  # Ensures reproducibility
# For 70-30 Split
split_ratio_70_30 <- 0.7
obesityData_norm_rows_70_30 <- nrow(obesityData.norm)
obesityData_rows_70_30 <- round(split_ratio_70_30 * obesityData_norm_rows_70_30)
obesityData_train_index_70_30 <- sample(obesityData_norm_rows_70_30, obesityData_rows_70_30)
obesityData_train_70_30 <- obesityData.norm[obesityData_train_index_70_30, ]
obesityData_test_70_30 <- obesityData.norm[-obesityData_train_index_70_30, ]
cat("70-30 Split: Training set has", nrow(obesityData_train_70_30), "rows. Test set has",
nrow(obesityData_test_70_30), "rows.\n")


# For 60-40 Split
split_ratio_60_40 <- 0.6
obesityData_norm_rows_60_40 <- nrow(obesityData.norm)
obesityData_rows_60_40 <- round(split_ratio_60_40 * obesityData_norm_rows_60_40)
obesityData_train_index_60_40 <- sample(obesityData_norm_rows_60_40, obesityData_rows_60_40)
obesityData_train_60_40 <- obesityData.norm[obesityData_train_index_60_40, ]
obesityData_test_60_40 <- obesityData.norm[-obesityData_train_index_60_40, ]
```

```
cat("60-40 Split: Training set has", nrow(obesityData_train_60_40), "rows. Test set has",
nrow(obesityData_test_60_40), "rows.\n")


# For 50-50 Split

split_ratio_50_50 <- 0.5

obesityData_norm_rows_50_50 <- nrow(obesityData.norm)

obesityData_rows_50_50 <- round(split_ratio_50_50 * obesityData_norm_rows_50_50)

obesityData_train_index_50_50 <- sample(obesityData_norm_rows_50_50, obesityData_rows_50_50)

obesityData_train_50_50 <- obesityData.norm[obesityData_train_index_50_50, ]

obesityData_test_50_50 <- obesityData.norm[-obesityData_train_index_50_50, ]

cat("50-50 Split: Training set has", nrow(obesityData_train_50_50), "rows. Test set has",
nrow(obesityData_test_50_50), "rows.\n")
```

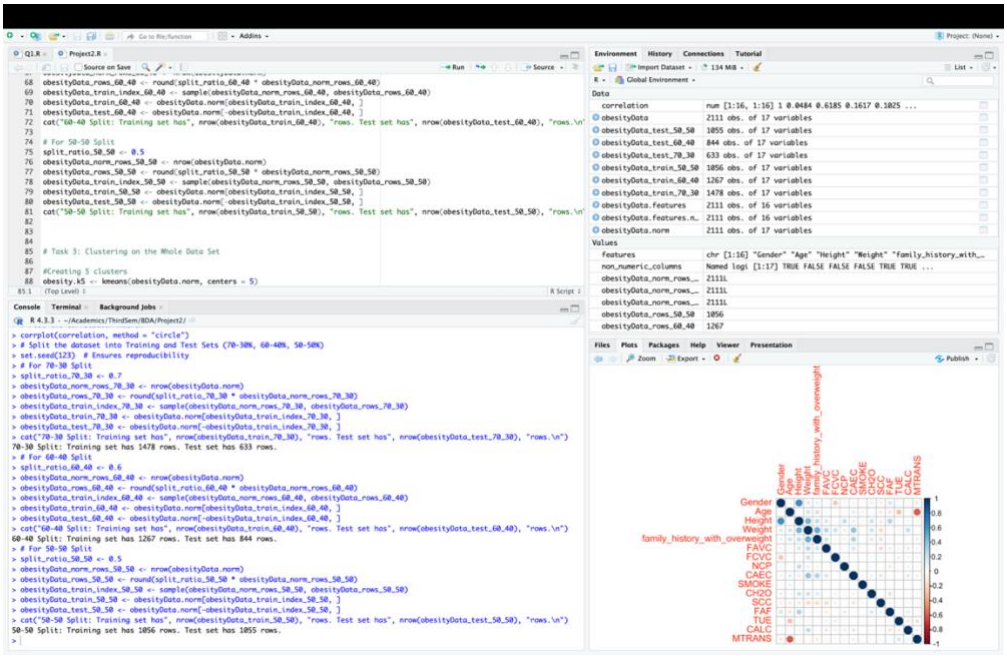

Fig 9 – Splitting of Dataset

Task 3: Clustering on the Whole Data Set

#Creating 5 clusters

```
obesity.k5 <- kmeans(obesityData.norm, centers = 5)

str(obesity.k5)

obesity.k5
```

```
factoextra::fviz_cluster(obesity.k5,obesityData.norm)
obesityData[194,]
obesityData[578,]
```

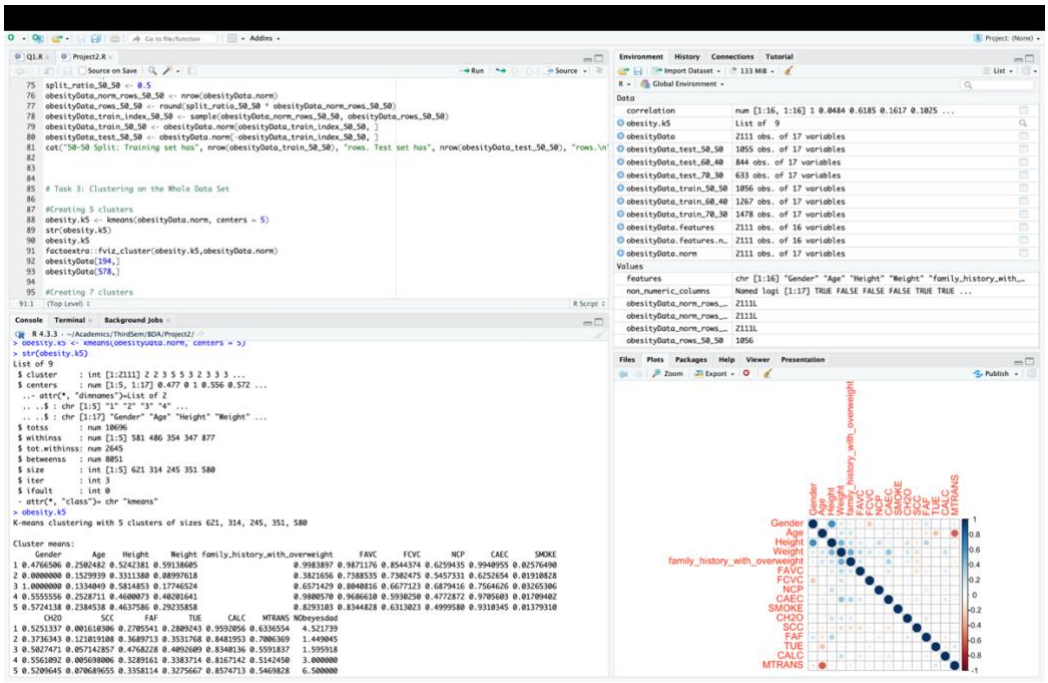

Fig 10 – Creating 5 clusters.

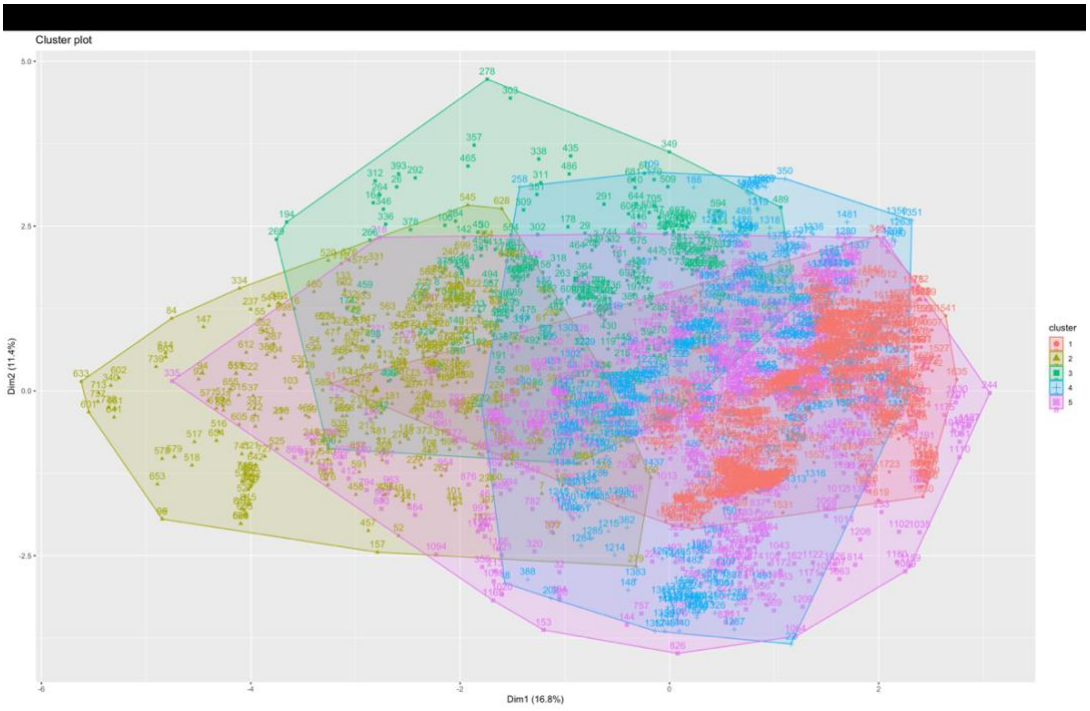

Fig 11 – Cluster 5 plot.

#Creating 7 clusters

```
obesity.k7 <- kmeans(obesityData.norm, centers = 7)
str(obesity.k7)
```

obesity.k7

factoextra::fviz_cluster(obesity.k7,obesityData.norm)

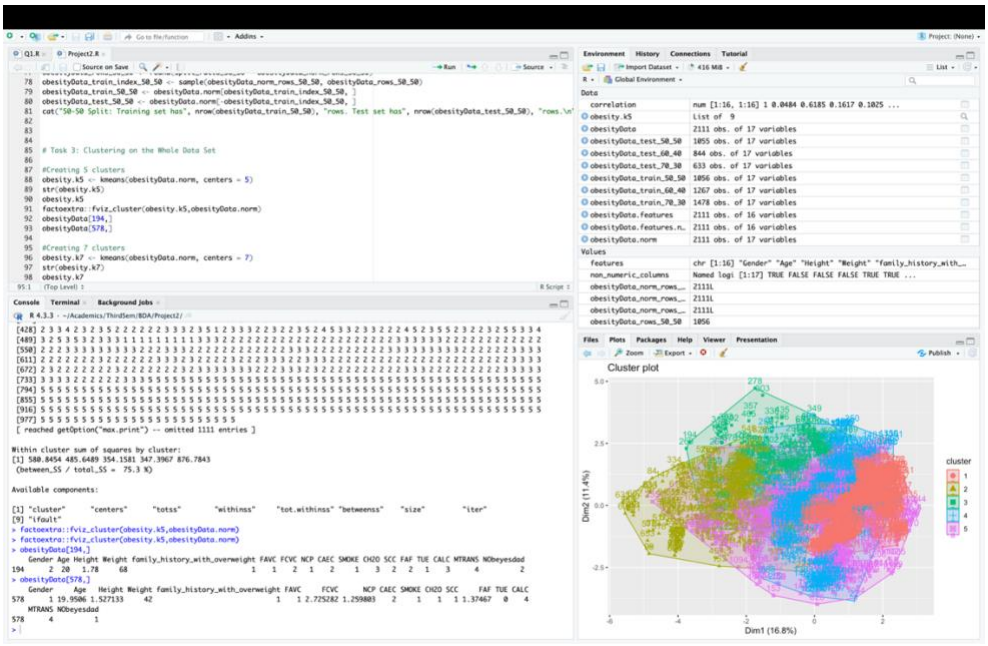

Fig 12 – Creating 7 clusters.

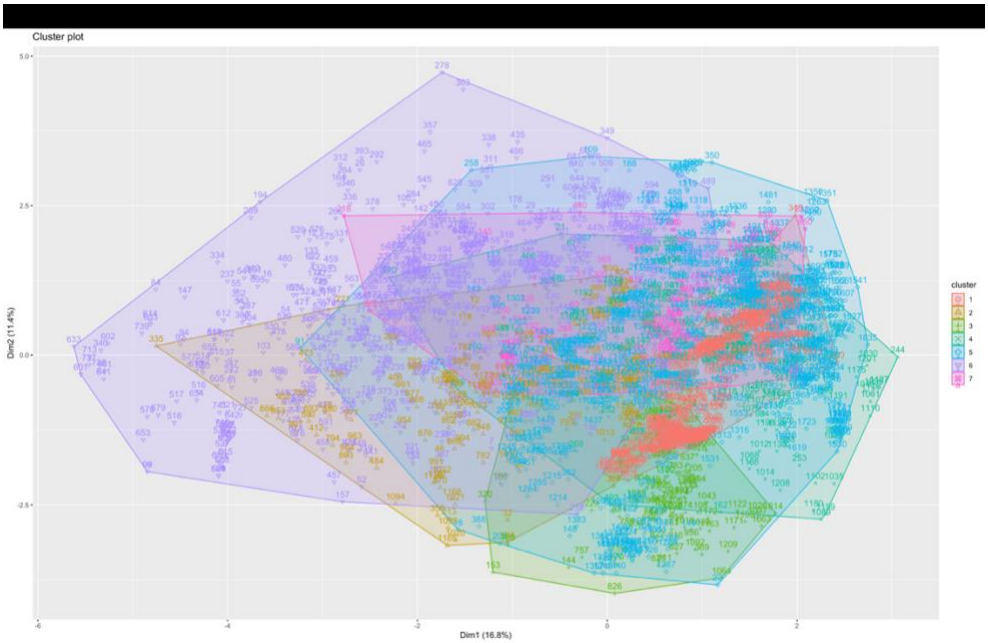

Fig 13 – Plot of 7 clusters

#Creating 9 clusters

obesity.k9 <- kmeans(obesityData.norm, centers = 9)

str(obesity.k9)

obesity.k9

factoextra::fviz_cluster(obesity.k9,obesityData.norm)

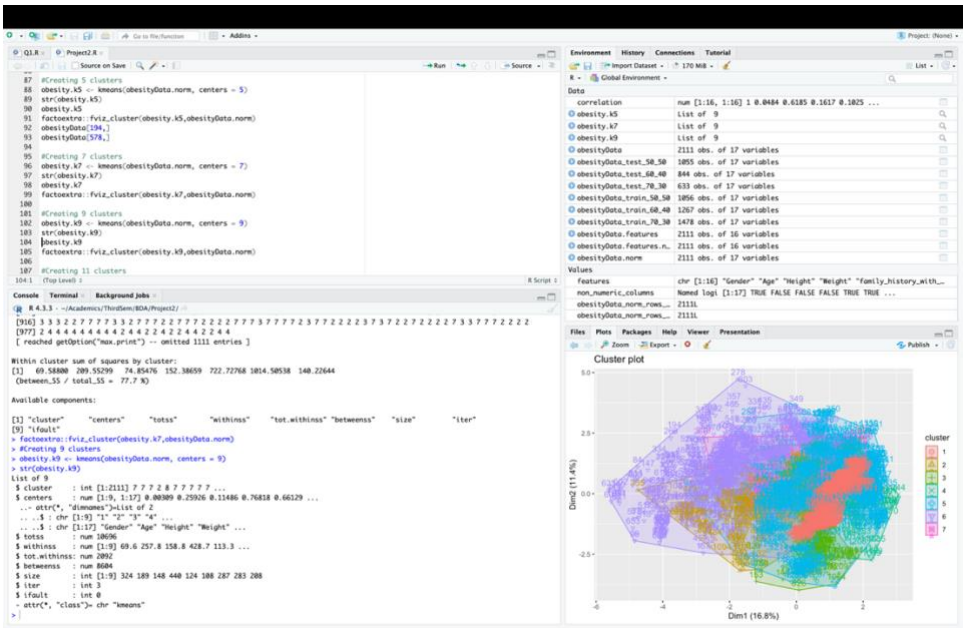

Fig 14 – Creating 9 clusters.

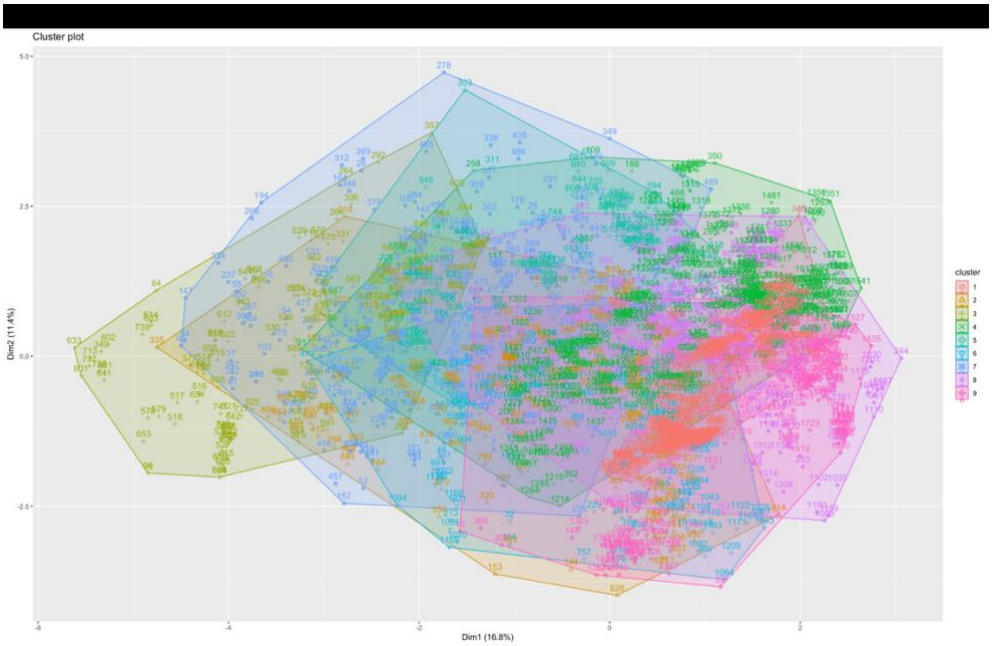

Fig 15 – Plot of 9 clusters.

#Creating 11 clusters

obesity.k11 <- kmeans(obesityData.norm, centers = 11)

obesity.k11

str(obesity.k11)

factoextra::fviz_cluster(obesity.k11,obesityData.norm)

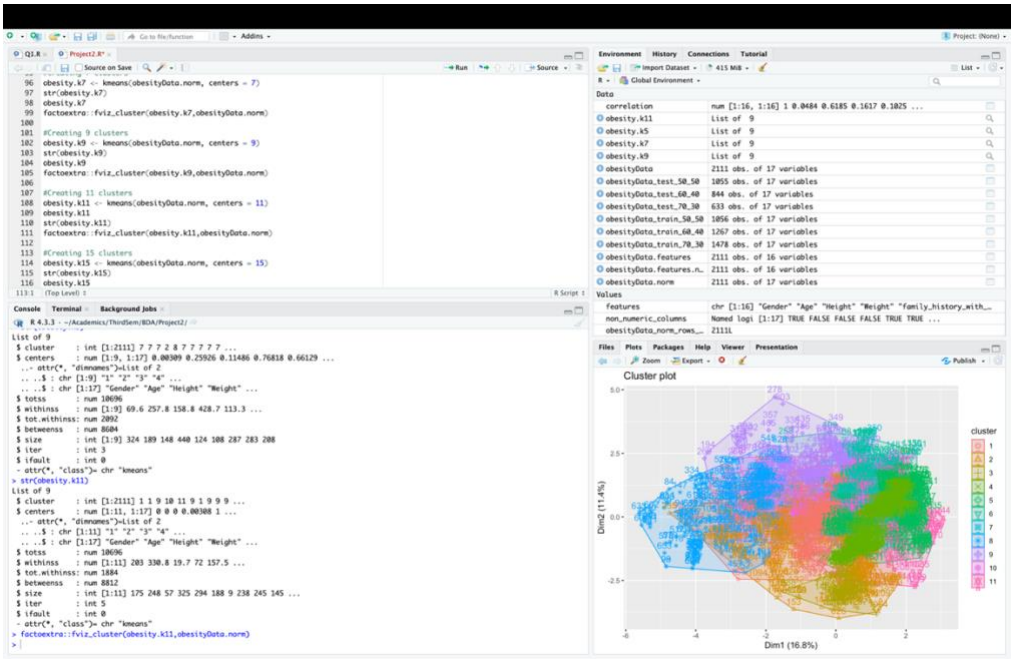

Fig 16 – Creating 11 clusters.

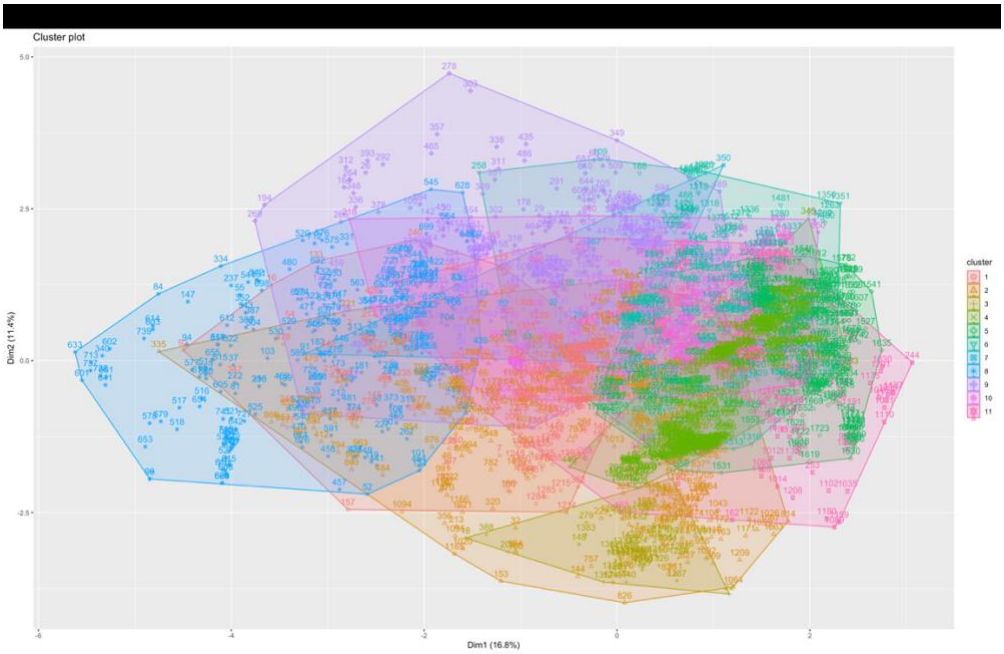

Fig 17 – Plot of 11 clusters.

#Creating 15 clusters

obesity.k15 <- kmeans(obesityData.norm, centers = 15)

str(obesity.k15)

obesity.k15

factoextra::fviz_cluster(obesity.k15,obesityData.norm)

obesityData[269,]

obesityData[164,]

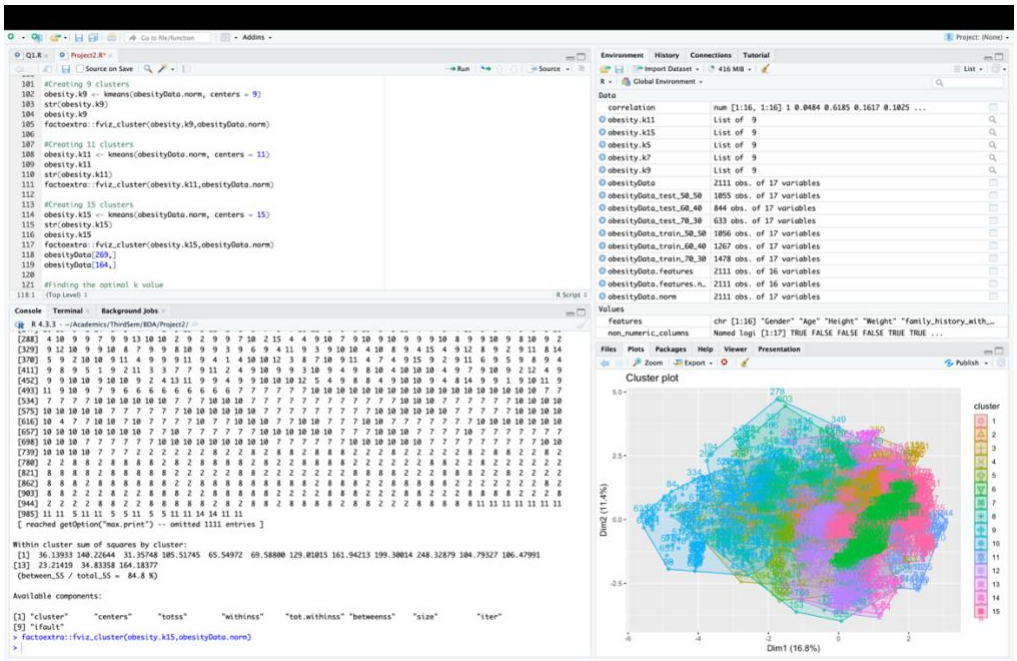

Fig 18 – Creating 15 clusters.

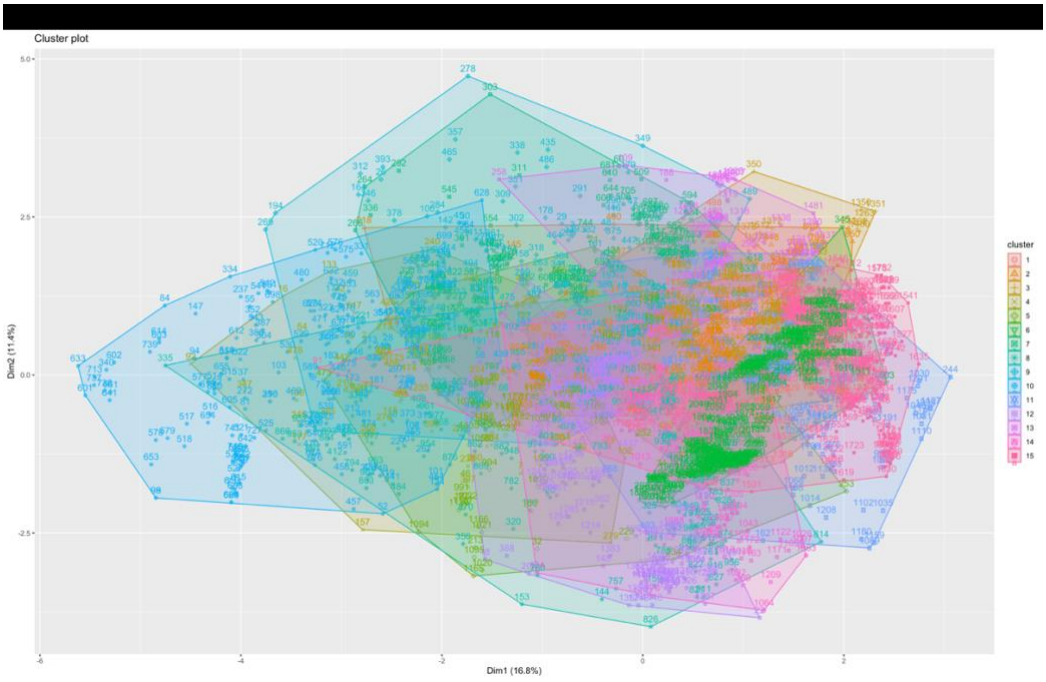

Fig 19 – Plot of 15 clusters.

#Finding the optimal k value

factoextra::fviz_nbclust(obesityData,FUNcluster = kmeans,method = "wss",k.max=20,verbose = TRUE)

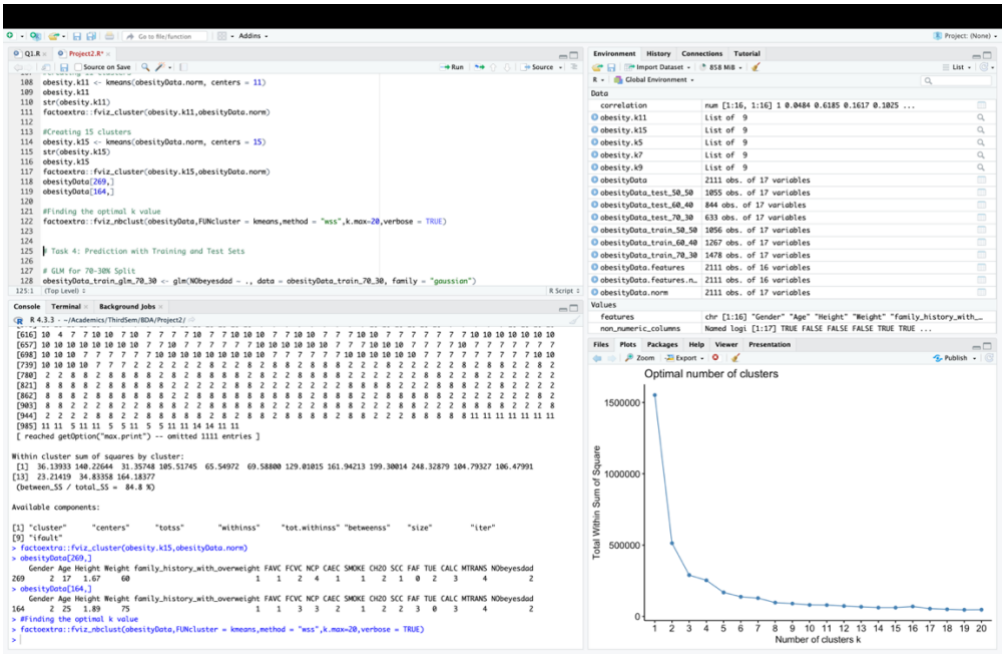

Fig 20 – Finding the optimal K value.

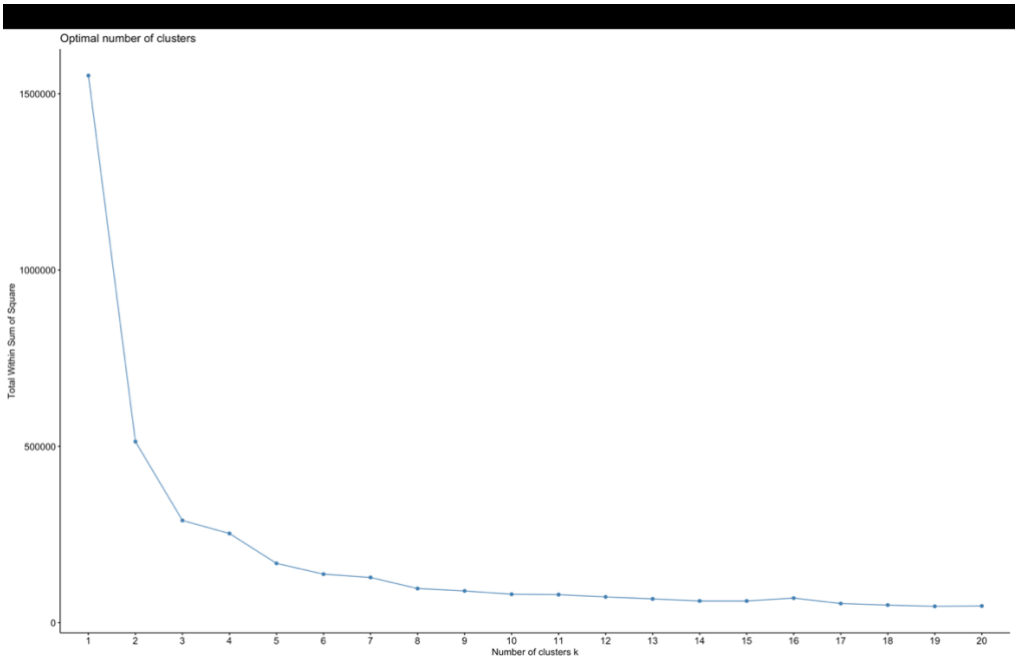

Fig 21 – Plot to show optimal K value.

Task 4: Prediction with Training and Test Sets

GLM for 70-30% Split

obesityData_train_glm_70_30 <- glm(NObeyesdad ~ ., data = obesityData_train_70_30, family = "gaussian")

#obesityData_train_glm_70_30 <- glm(formula = obesityData.train$NObeyesdad ~ obesityData.train$Age+obesityData.train$Height+obesityData.train$Weight+obesityData.train$MTRANS,family = gaussian, data=obesityData.train)

obesityData_test_pred_70_30 <- predict(obesityData_train_glm_70_30, newdata = obesityData_test_70_30, type = "response")

Adjusting predictions and generating true labels

obesityData_test_pred_class_70_30 <- ifelse(obesityData_test_pred_70_30 > 0.5, 1, 0) # Adjust according to your outcome

true_labels_70_30 <- obesityData_test_70_30$NObeyesdad

Accuracy Calculation

accuracy_70_30 <- mean(true_labels_70_30 == obesityData_test_pred_class_70_30)

cat("Accuracy for GLM predictions (70-30 split):", accuracy_70_30, "\n")

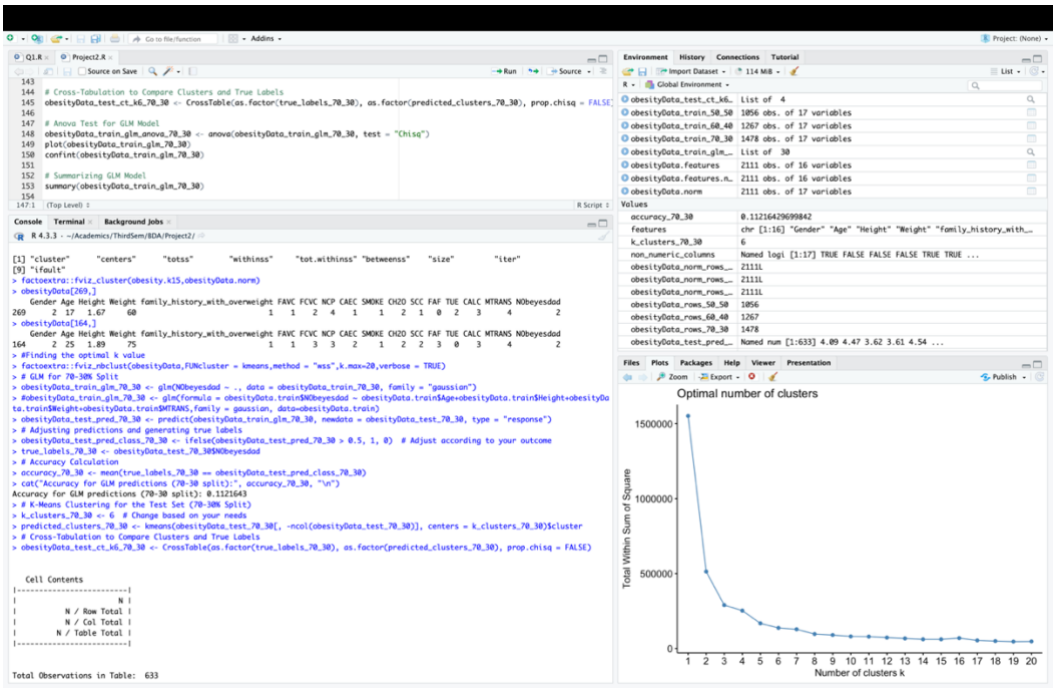

Fig 22 – Accuracy for the glm model.

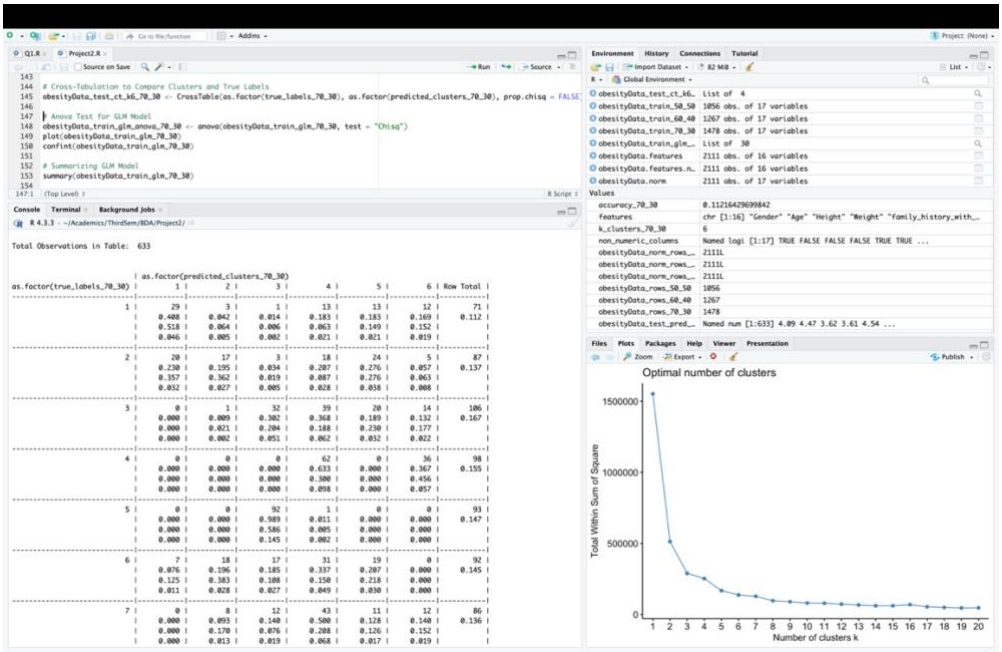

Fig 23 –glm model.

```
# K-Means Clustering for the Test Set (70-30% Split)

k_clusters_70_30 <- 6 # Change based on your needs

predicted_clusters_70_30 <- kmeans(obesityData_test_70_30[, -ncol(obesityData_test_70_30)], centers =
k_clusters_70_30)$cluster


# Cross-Tabulation to Compare Clusters and True Labels

obesityData_test_ct_k6_70_30 <- CrossTable(as.factor(true_labels_70_30),
as.factor(predicted_clusters_70_30), prop.chisq = FALSE)


# Anova Test for GLM Model

obesityData_train_glm_anova_70_30 <- anova(obesityData_train_glm_70_30, test = "Chisq")

plot(obesityData_train_glm_70_30)

confint(obesityData_train_glm_70_30)
```

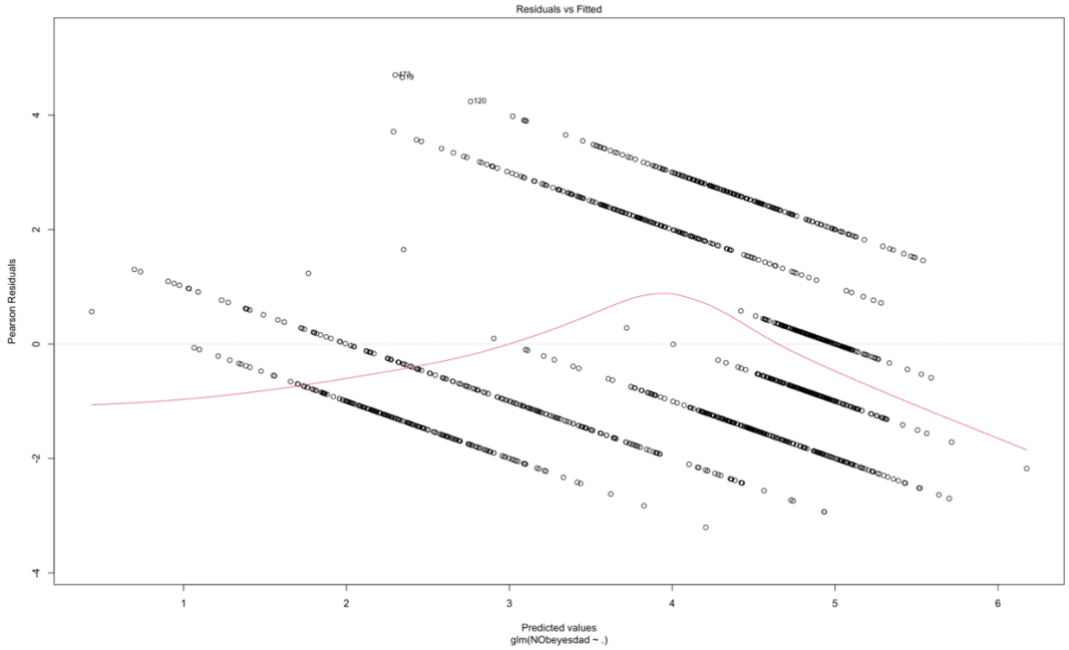

Fig 24 - Residuals versus fitted_values plot.

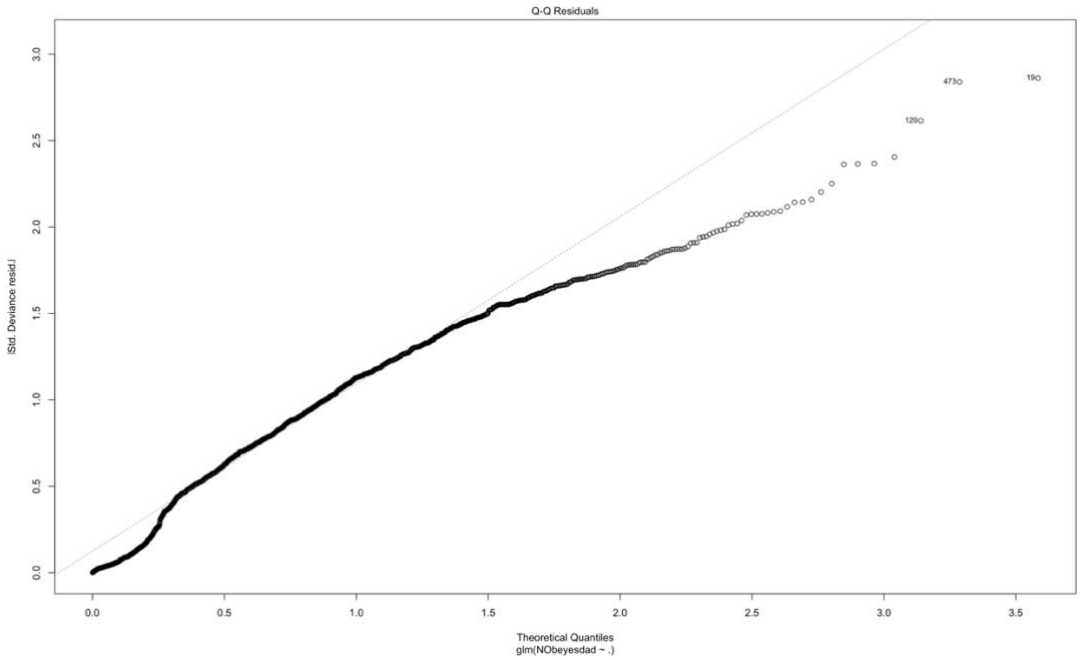

Fig 25 – Q-Q residuals plot.

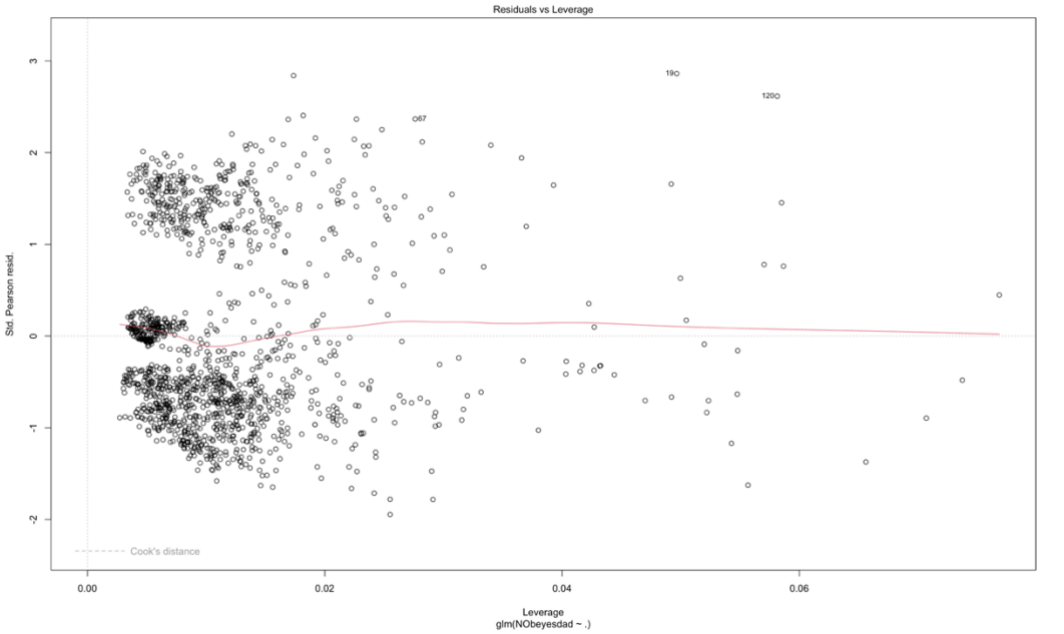

Fig 26 – Residuals vs Leverage

Summarizing GLM Model

summary(obesityData_train_glm_70_30)

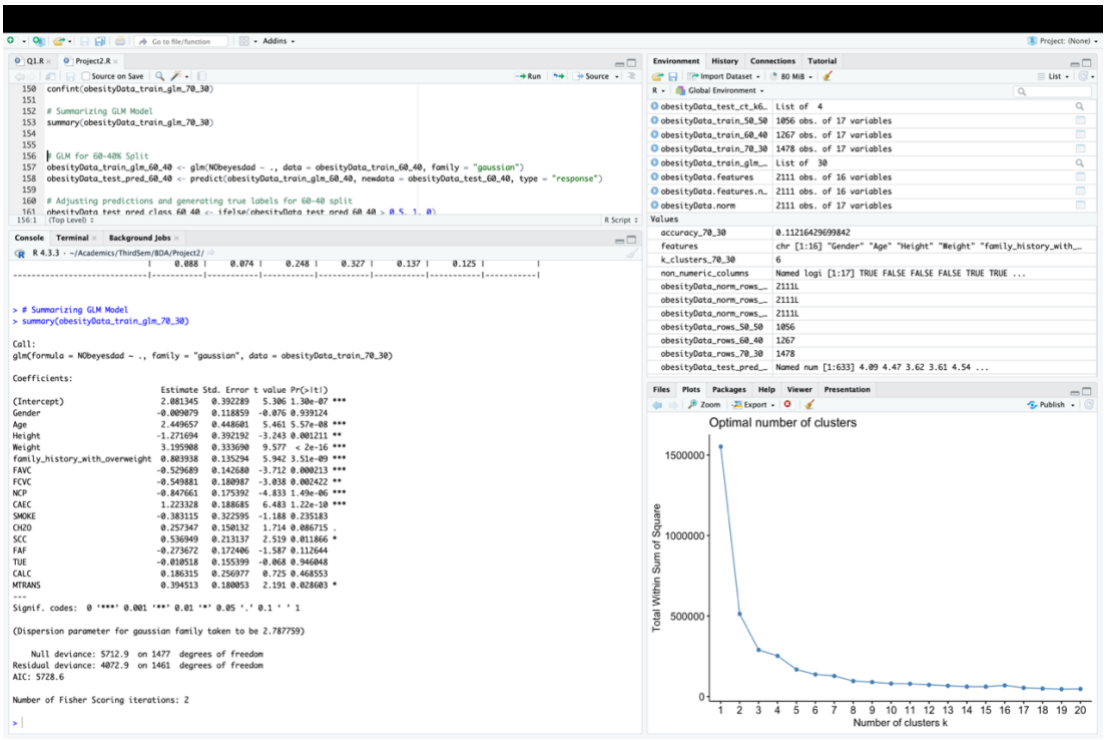

Fig 27 – GLM with 70-30 split

```
# GLM for 60-40% Split

obesityData_train_glm_60_40 <- glm(NObeyesdad ~ ., data = obesityData_train_60_40, family = "gaussian")

obesityData_test_pred_60_40 <- predict(obesityData_train_glm_60_40, newdata = obesityData_test_60_40, type = "response")


# Adjusting predictions and generating true labels for 60-40 split

obesityData_test_pred_class_60_40 <- ifelse(obesityData_test_pred_60_40 > 0.5, 1, 0)

true_labels_60_40 <- obesityData_test_60_40$NObeyesdad


# Accuracy Calculation for 60-40 split

accuracy_60_40 <- mean(true_labels_60_40 == obesityData_test_pred_class_60_40)

cat("Accuracy for GLM predictions (60-40 split):", accuracy_60_40, "\n")


# K-Means Clustering for the Test Set (60-40% Split)

k_clusters_60_40 <- 6  # Adjust based on your needs

predicted_clusters_60_40 <- kmeans(obesityData_test_60_40[, -ncol(obesityData_test_60_40)], centers = k_clusters_60_40)$cluster

# Cross-Tabulation to Compare Clusters and True Labels for 60-40 split

obesityData_test_ct_k6_60_40 <- CrossTable(x = as.factor(true_labels_60_40), y = as.factor(predicted_clusters_60_40), prop.chisq = FALSE)
```

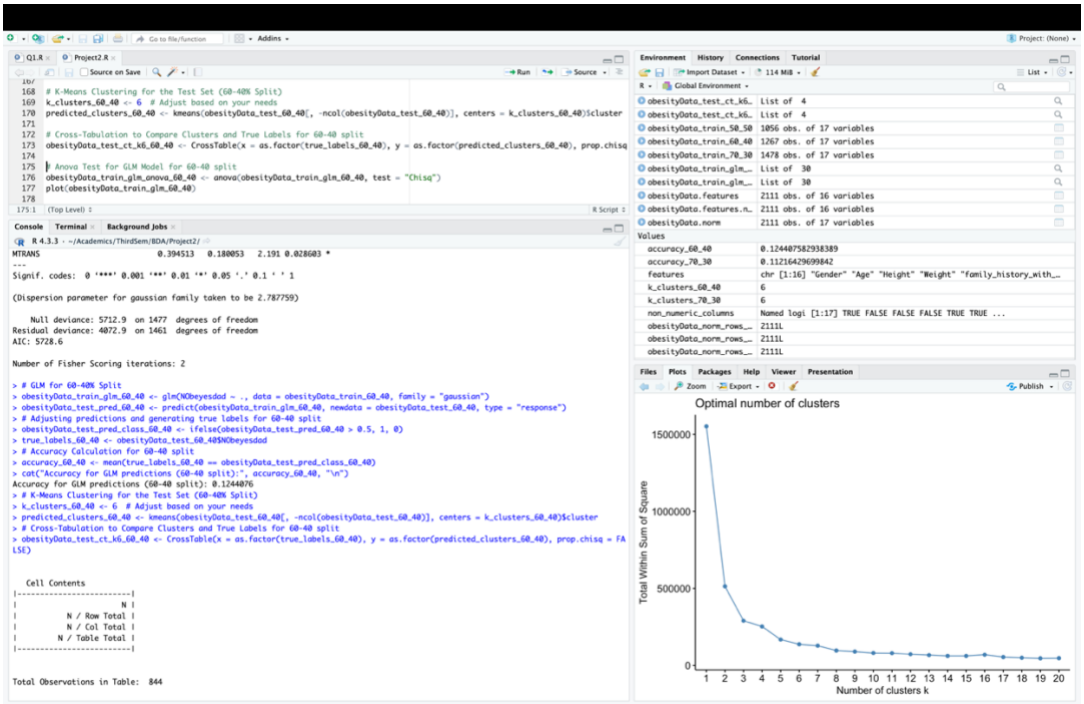

Fig 28 – Glm and K-means for 60-40 split

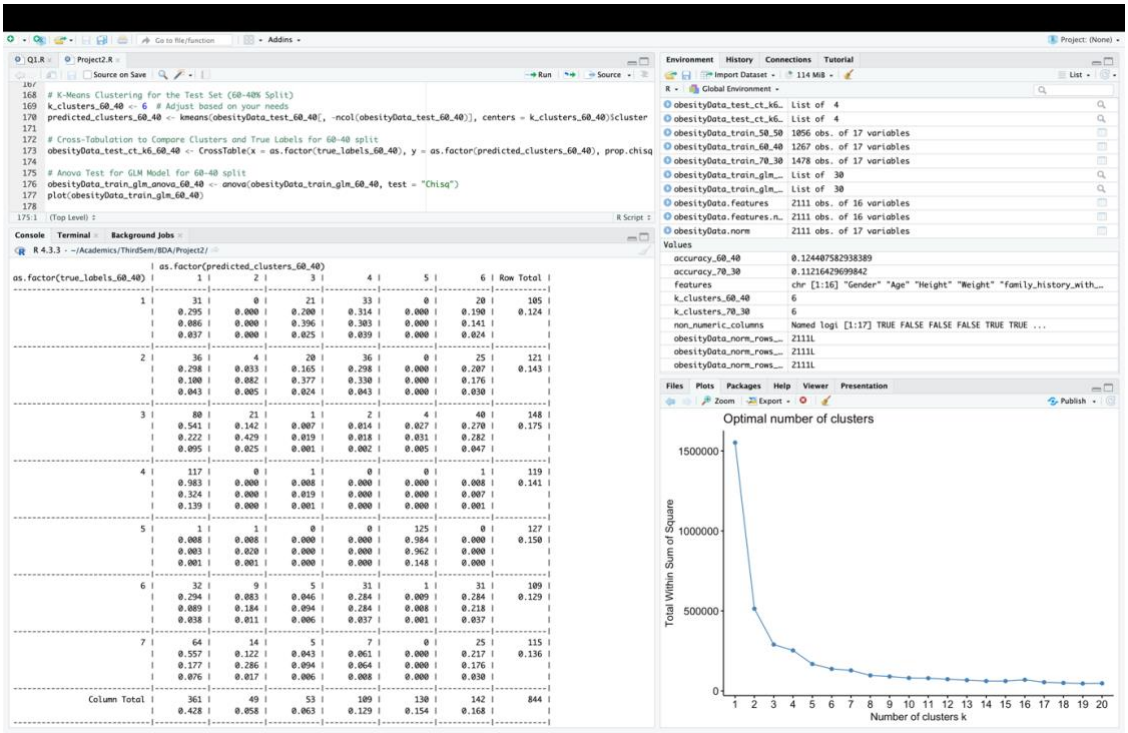

Fig 29 – Anova test for GLM 60-40 split.

```
# Anova Test for GLM Model for 60-40 split
obesityData_train_glm_anova_60_40 <- anova(obesityData_train_glm_60_40, test = "Chisq")
plot(obesityData_train_glm_60_40)
```

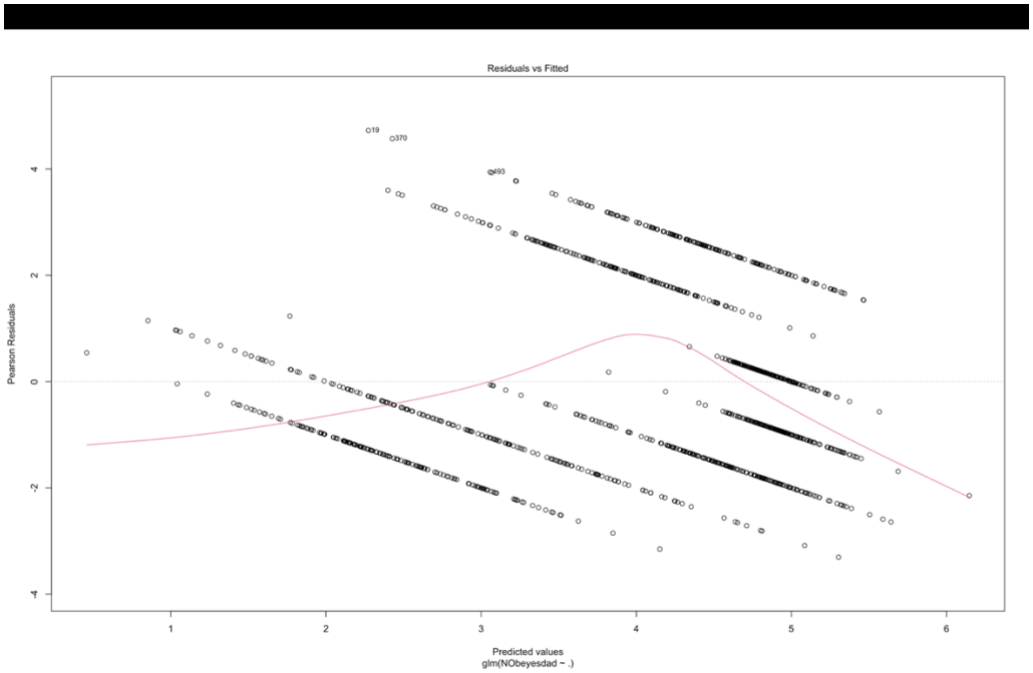

Fig 30 – Residuals vs Fitted.

[redacted]

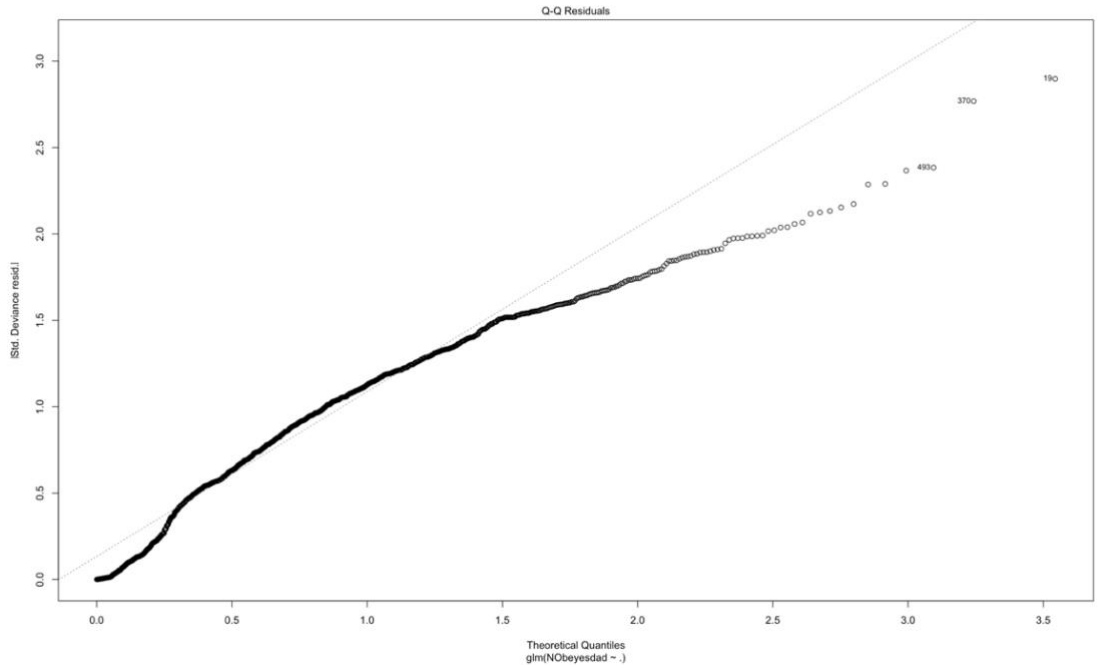

Fig 31– Q-Q Residuals.

[redacted]

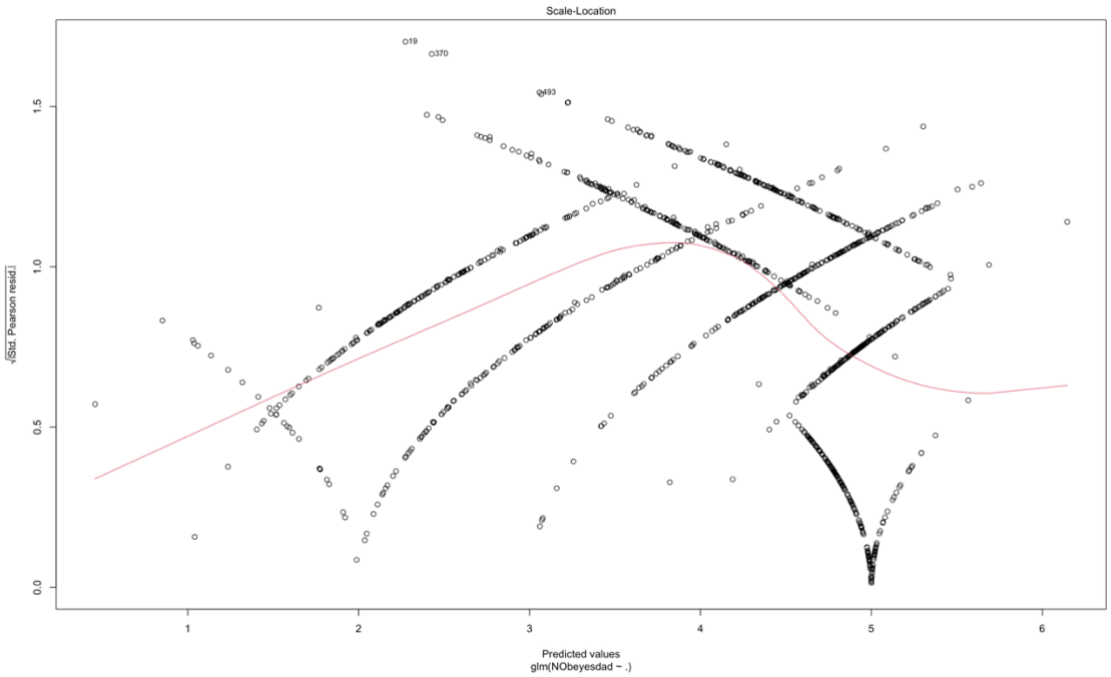

Fig 32 – Scale-location.

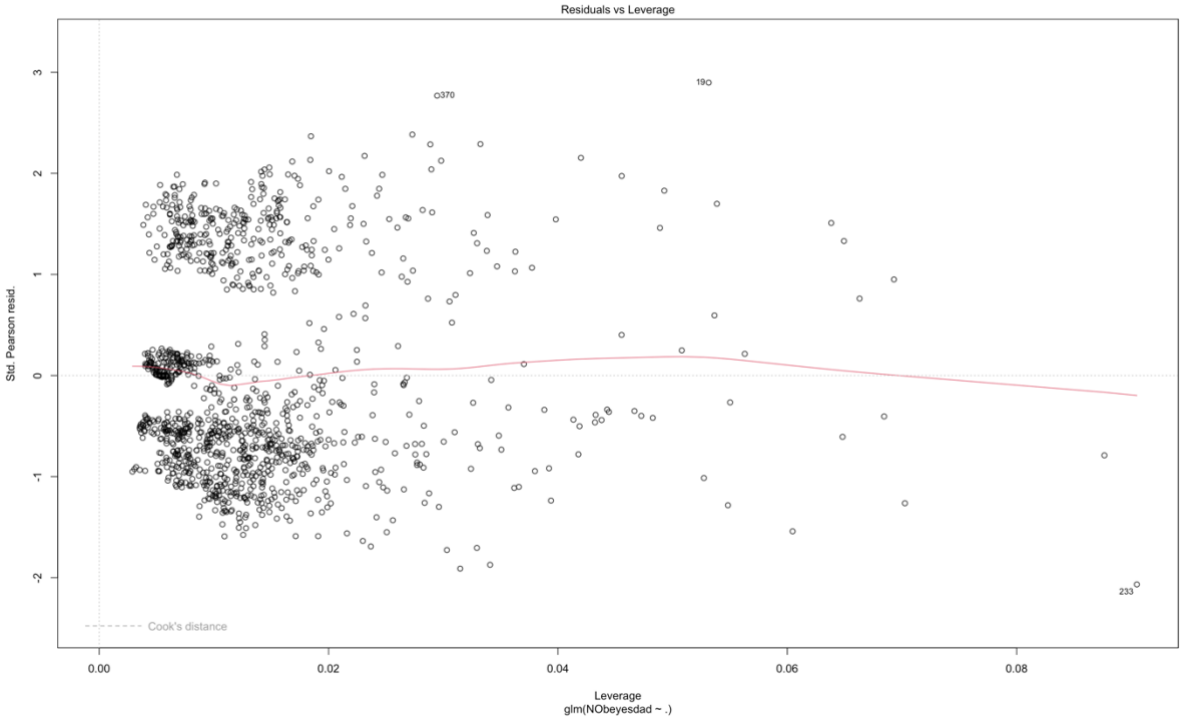

Fig 33 – Residuals vs Leverage.

```
# Summarizing GLM Model for 60-40 split
summary(obesityData_train_glm_60_40)
# GLM for 50-50% Split
obesityData_train_glm_50_50 <- glm(NObeyesdad ~ ., data = obesityData_train_50_50, family = "gaussian")
obesityData_test_pred_50_50 <- predict(obesityData_train_glm_50_50, newdata = obesityData_test_50_50,
type = "response")

# Adjusting predictions and generating true labels for 50-50 split
obesityData_test_pred_class_50_50 <- ifelse(obesityData_test_pred_50_50 > 0.5, 1, 0)
true_labels_50_50 <- obesityData_test_50_50$NObeyesdad

# Accuracy Calculation for 50-50 split
accuracy_50_50 <- mean(true_labels_50_50 == obesityData_test_pred_class_50_50)
cat("Accuracy for GLM predictions (50-50 split):", accuracy_50_50, "\n")

# K-Means Clustering for the Test Set (50-50% Split)
k_clusters_50_50 <- 6  # Adjust based on your needs
predicted_clusters_50_50 <- kmeans(obesityData_test_50_50[, -ncol(obesityData_test_50_50)], centers =
k_clusters_50_50)$cluster
```

Cross-Tabulation to Compare Clusters and True Labels for 50-50 split

obesityData_test_ct_k6_50_50 <- CrossTable(x = as.factor(true_labels_50_50), y = as.factor(predicted_clusters_50_50), prop.chisq = FALSE)

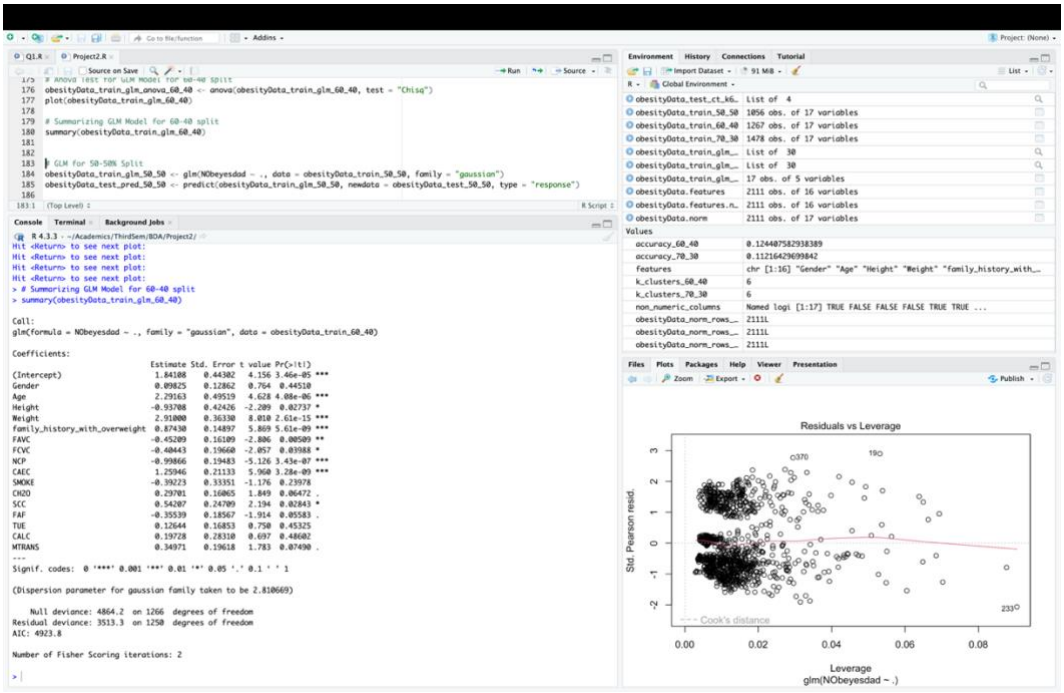

Fig 34 – Cross Tabulation results.

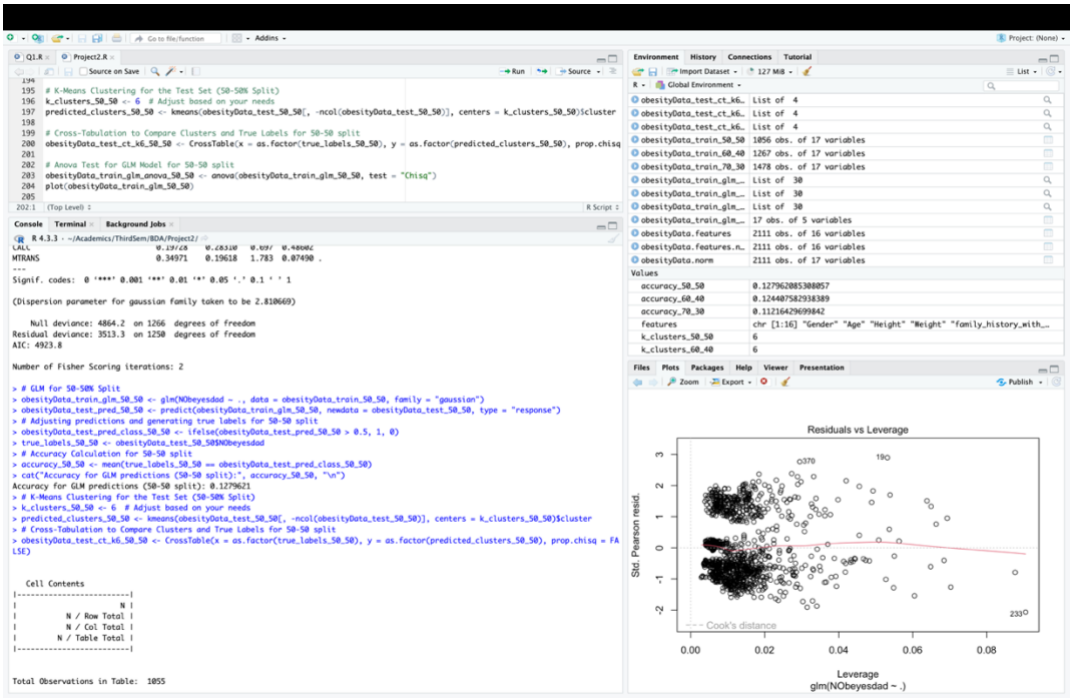

Fig 35 – Cross Tabulation results with comparisons

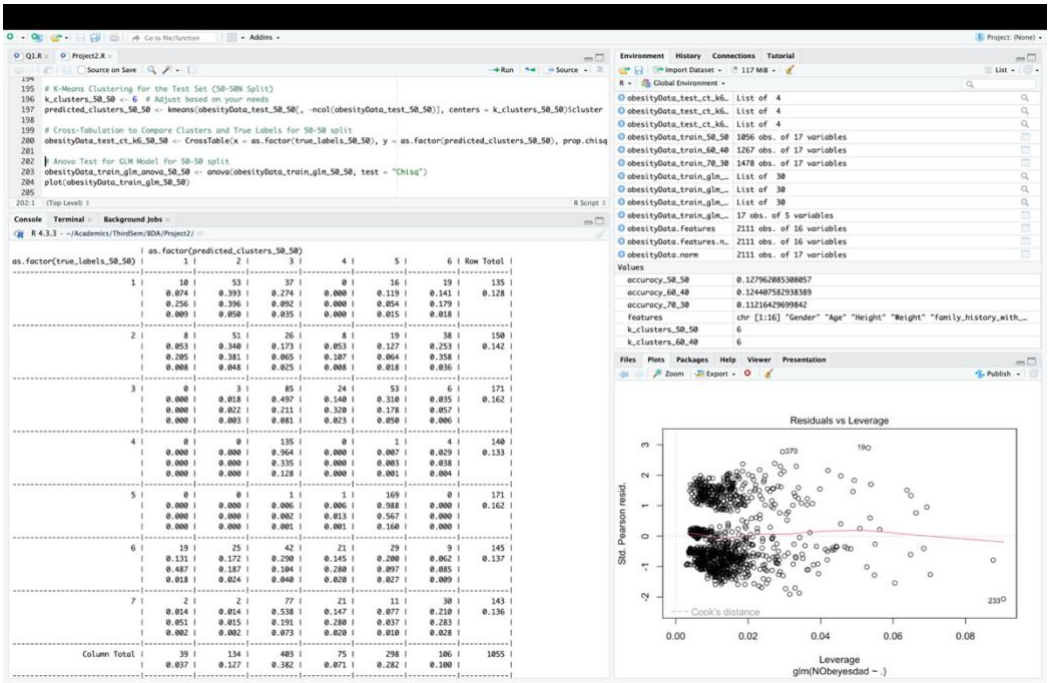

Fig 36 – Cross Tabulation results.

Anova Test for GLM Model for 50-50 split

obesityData_train_glm_anova_50_50 <- anova (obesityData_train_glm_50_50, test = "Chisq")

plot(obesityData_train_glm_50_50)

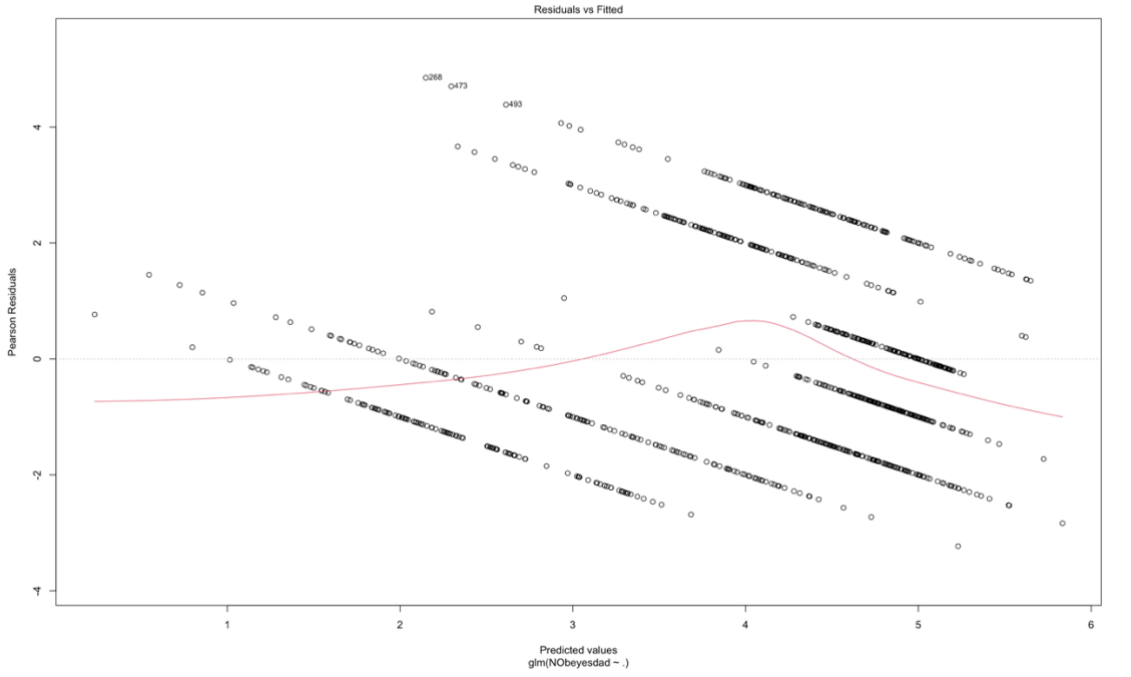

Fig 37 – Anova test for 50-50 split.

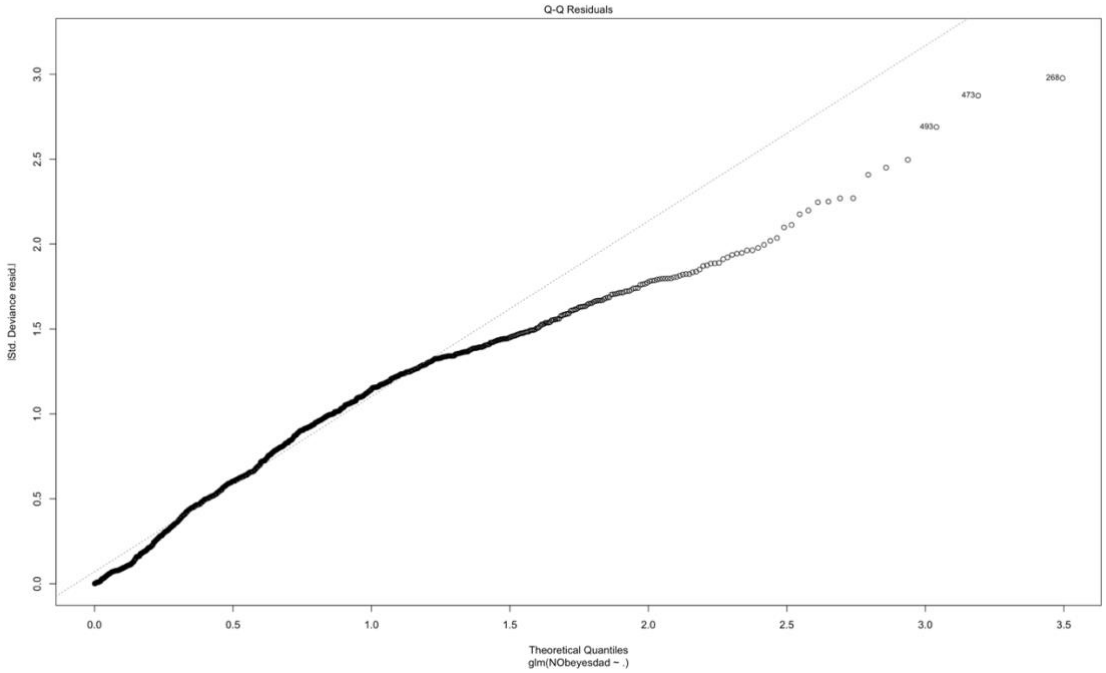

Fig 38 – Q-Q residuals.

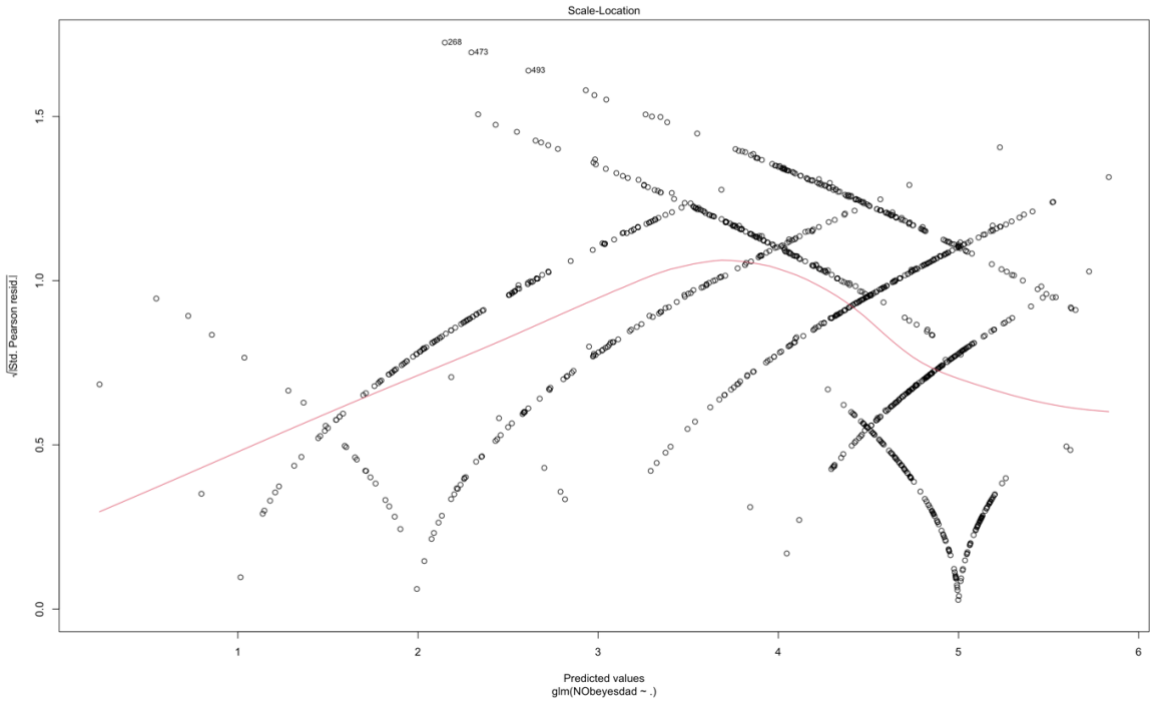

Fig 39 – Scale-location plot.

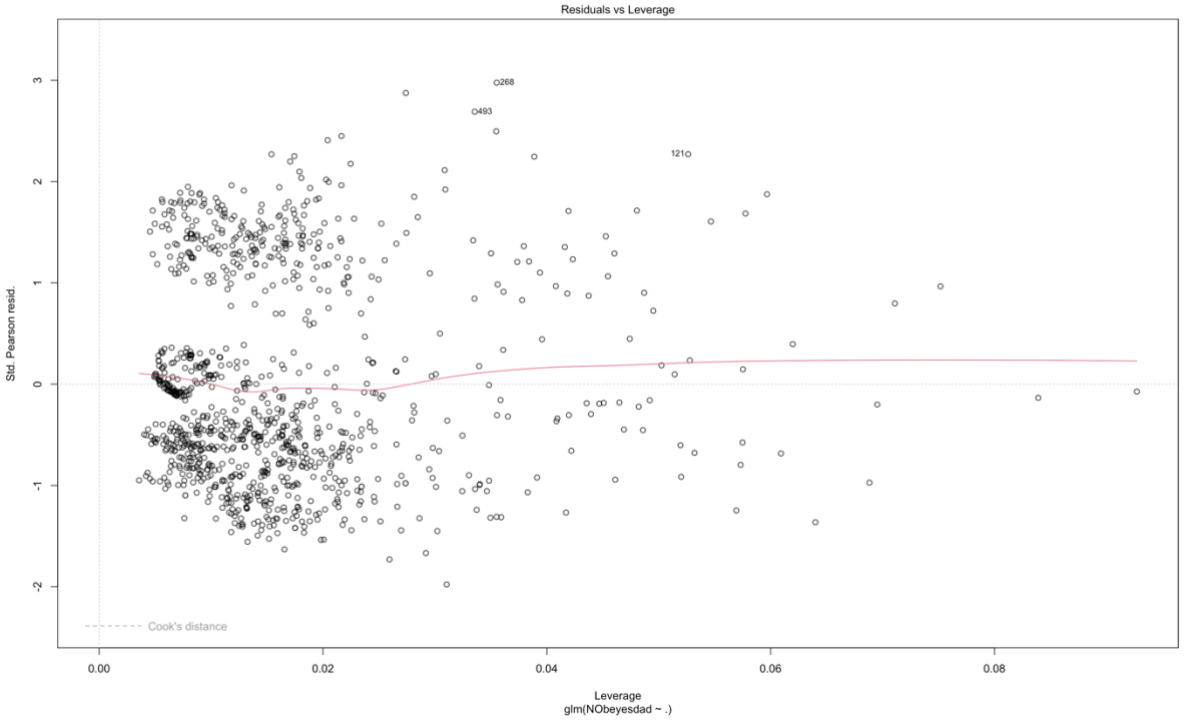

Fig 40 – Residual vs Leverage.

Summarizing GLM Model for 50-50 split

summary(obesityData_train_glm_50_50)]

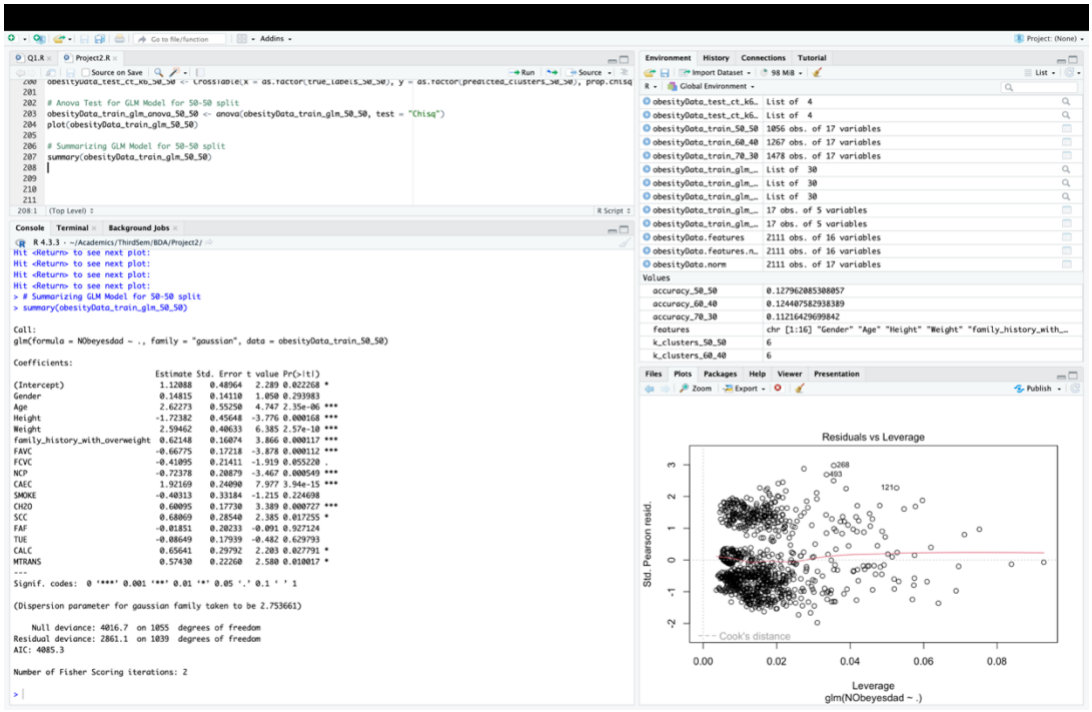

Fig 41 – Summarizing GLM model for 50-50 split.

Task 6

Analysis of what this project helped us learn about Data Science

1)Exploratory Data Analysis (EDA): EDA is a crucial step in understanding the dataset's characteristics, identifying patterns, and formulating hypotheses. Analysing the relationships between variables such as diet, exercise, socio-economic status, and obesity rates in the three countries can help in gaining insights into the factors influencing obesity.

2)Data Visualization: Visualizing data through graphs, charts, and maps is essential for communicating findings effectively. Creating visualizations to depict obesity trends over time or across different regions within each country can aid in understanding the data more intuitively.

3)Statistical Analysis: Employing statistical techniques to analyse correlations or clustering methods can help in identifying significant factors contributing to obesity levels. For instance, understanding the correlation between age and obesity and between height and obesity.

4)Machine Learning Models: Building predictive models using machine learning algorithms to forecast obesity rates based on various features such as transportation, physical activity, and socio-economic factors can be another aspect of the analysis. This involves model selection, feature engineering, training, and evaluation.

5)Interpretation and Insights: Understanding the implications of the analysis results for public health policies, personalized interventions, or targeted education campaigns is crucial in Data Science.