

PROJECT REPORT

Sisir Lunawat ([lunawat.s@husky.neu.edu](mailto:lunawat.s@husky.neu.edu))

Saketh Vommina Venkata(venkata.s@husky.neu.edu)

Topic 1: Analysing different diet regimes using ANOVA

Topic 2: Predicting student’s Math Score on various Socioeconomic factors using regression

Professor: Nasser Fard

IE7280: Statistical Method in Engineering

**Topic 1:**

**Introduction:**

Sixty-two percent of adult Americans are overweight or obese, according to the Centres for Disease Control. More than 9 million children and teenagers are suffering the same problems. An increase in fast food consumption, coupled with low physical activity, largely contribute to the problem. Given those facts, it's no surprise that dieting and weight loss are common issues among people of all ages.

Up to 50 percent of women are on a diet at any given time, according to Judy Mahle Lutter in her book "The Body wise Woman." Up to 90 percent of teenagers diet regularly, and up to 50 percent of younger kids have tried a diet at some point. Americans spend more money in dieting, dieting products and weight loss surgery than any other people in the world. The numbers are expected to continue to grow significantly every year.

A diet is best described as a fixed plan of eating and drinking where the type and amount of food are planned out to achieve weight loss or follow a lifestyle. Many of the beliefs related to dieting and weight loss are erroneous. For example, it makes no difference whether you eat your calories at night or during the day. It's the total number of calories you consume throughout the day that makes or breaks your diet. Another common myth is that carbohydrates are the enemy. The truth is that complex carbs such as whole grains, brown rice and unsweetened cereals are excellent sources of fiber and can help with weight loss. Simple carbs like sugars and highly processed grains and flours are the problem. There are many weight loss diets; some focus on reducing one’s appetite, while others restrict calories, carbs or fats. Since all of them claim to be superior, it can be hard to know which ones are worth trying.

With such statistics, it is very natural that one might want to know which diet plan gives better results and if they are effective.

In this project, we have considered three types of diets – “Atkins”, “Ketogenic” and “Vegan”.

We will analyse their effects and deduce if they all have the same effects or if one is better than the rest.

**Problem Statement:**

In this study, we want to analyse different diets and compute the weight losses for each of them and answer the following questions: -

1. Are all the diets same or do they produce different results?
2. Do the diets have same effect on either sex or do they differ based on the sex?
3. Do the diets have same effect on any age group?

**Statistical Procedure:**

This study applies ANOVA on analysing weight loss of different people and their diets based on the data collected by the University of Sheffield website. Tukey’s test would be performed while the effect is significant. We also perform the test of normality of the residuals as well as test for the homogeneity of the variances. Finally, we will check for interaction between the factors after performing two-way ANOVA.

**Data Description:**

The dataset contains 76 observations with 7 variables. The 7 variables used in this study are:

1. Person – person ID in a numerical format
2. Gender – male/female
3. Age – age in years
4. Height – height in centimeters
5. Pre\_weight – weight in kgs before the person started the diet
6. Diet Type – three different diet types – Atkins/Ketogenic/Vegan
7. Weight after 6 weeks – weight in kgs after 6 weeks of following the diet

**Data Preprocessing:**

The raw data had few of its columns such as “gender”, “diet.type” in either “numerical” or “String” format. These have been converted into categorical variables and a new column called “weight\_loss” has been computed to indicate the weight lost in kgs for each of the records.

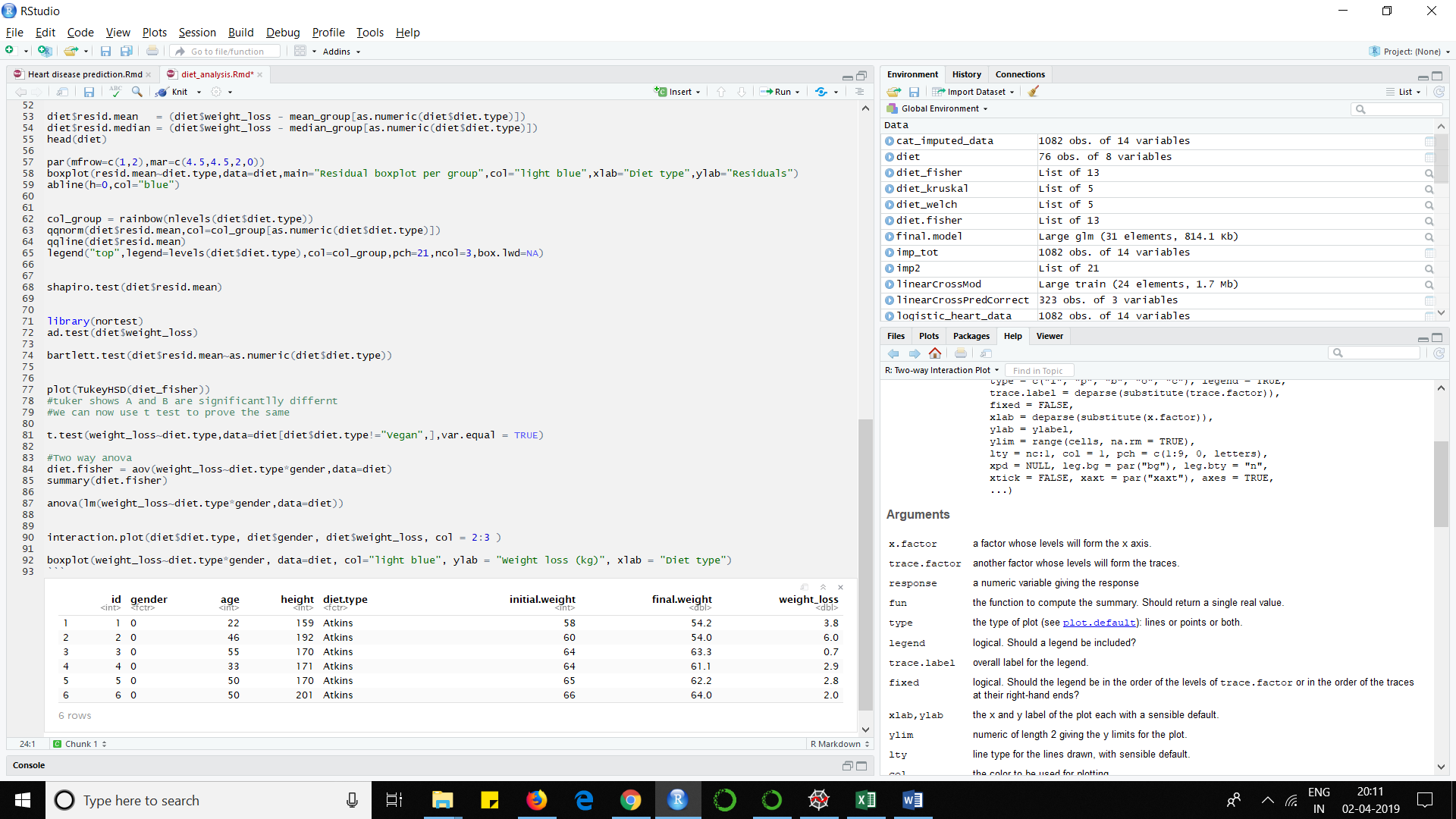


Figure 1: Partial display of the data set

**Data Analysis:**

First, we implement a box plot to visualize the effect of each of the diet types. It can be clearly be seen from Figure 2 that the diet type “Vegan” has a bigger impact compared to that of “atkins” and “ketogenic”. While “Atkins” and “Ketogenic” have their median around 3 and 3.5, the diet type “Vegan” has its median at close to 5.

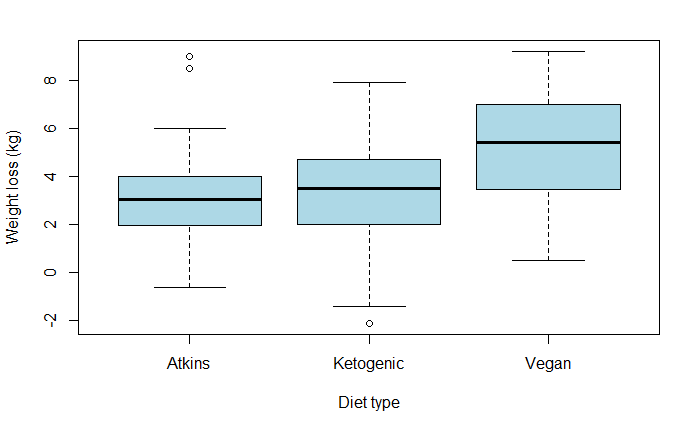


Figure 2: Weight loss per diet type using box plot

After performing one- way ANOVA using the Fisher’s test by means of aov() function in R, we get the following ANOVA table:

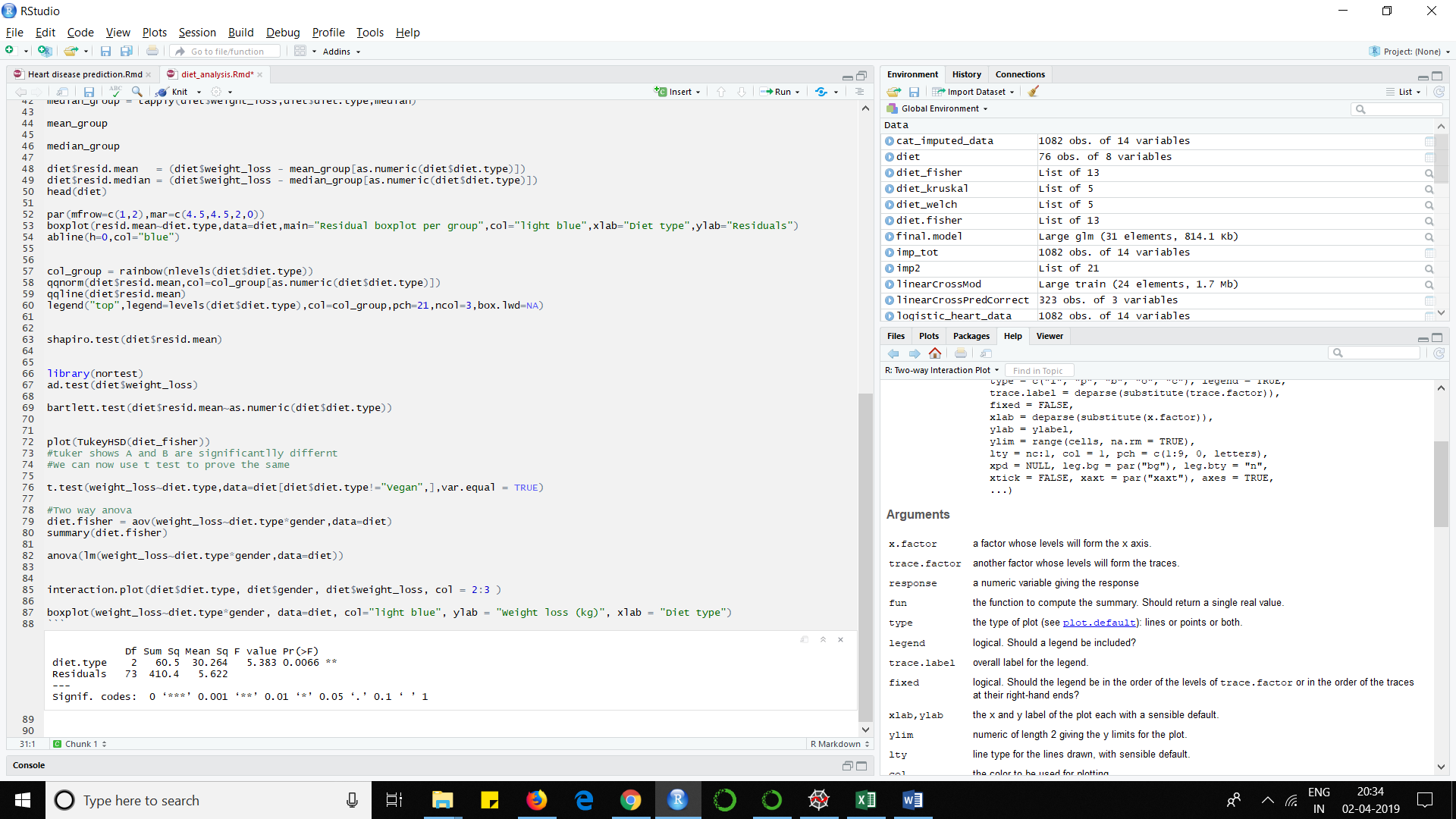


Figure 3: Summary of the analysis of variance using Fisher’s test

The p-value turns out to be a very low value of 0.0066 which means that the variable “diet type” is significant for weight loss and thus indicate that there is difference amongst the groups.

ANOVA has three main assumptions:

1. Independence of the cases- To satisfy this condition we ensured that all the data is converted to categorical to be able to perform the ANOVA test.
2. Normality – The distributions of the residuals are normal.
3. Homoscedasticity – equality or homogeneity of variances

Tests have been performed so that the Normality and Homoscedasticity assumptions are indeed satisfied.

The residuals of the records have been calculated by computing the means of the diet types and then subtracting the mean of each group to the weight loss of the corresponding participants. The following data was then obtained:

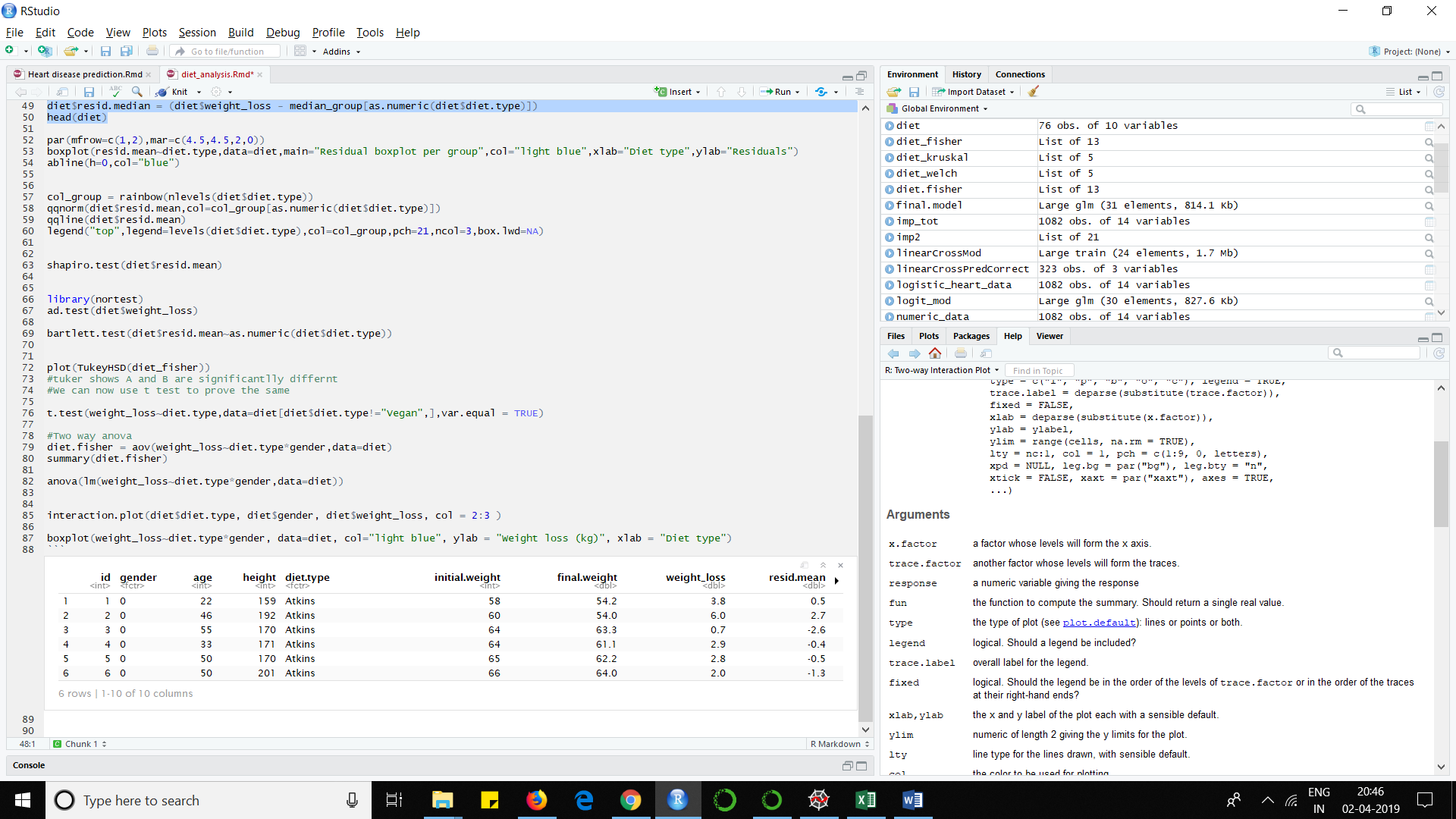


Figure 4: Data after computing residuals

The below figure displays box plots as well as QQ plot of the residuals. These are used to draw insights if the assumptions are indeed true for the given datasets.

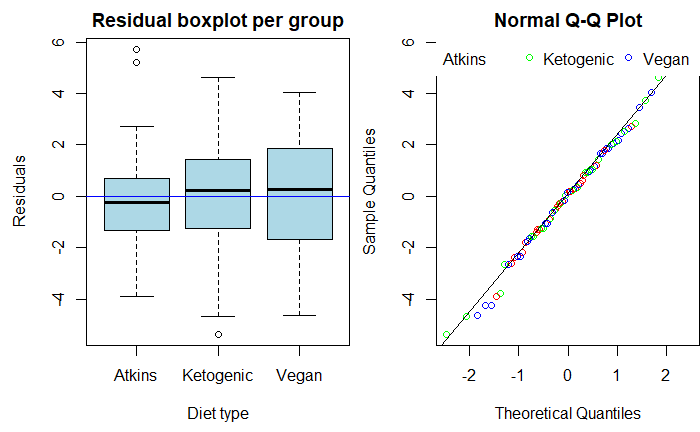


Figure 5: Data after computing residuals

From figure 5, the box plot shows that the residual medians are close to zero, with few outliers present. Also, the Quantile-Quantile plot indicates that the points seem to fall about the line confirming the datasets is from the same distribution. Both these plots confirm that the residuals are normal. To prove this assumption statistically Shapiro-Wilk test is performed, the results of which is shown below:

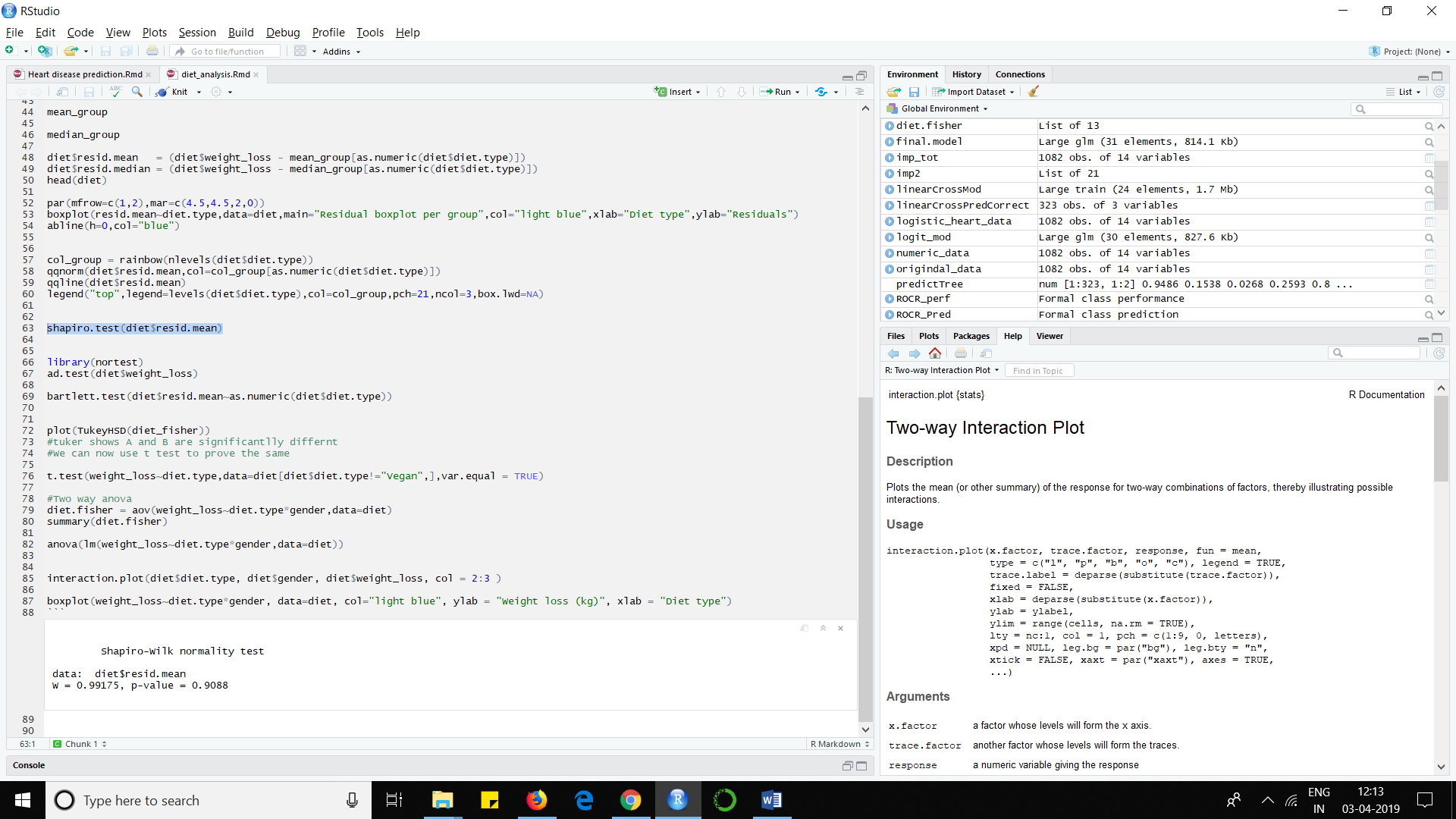


Figure 6: Shapiro-wilk test output

The high p-value of 0.9088 means that we fail to reject the null hypothesis and conclude that the residuals are indeed are normal.

The third assumption i.e. homogeneity of the variances is tested by using the Bartlett’s Test, whose results are shown below:

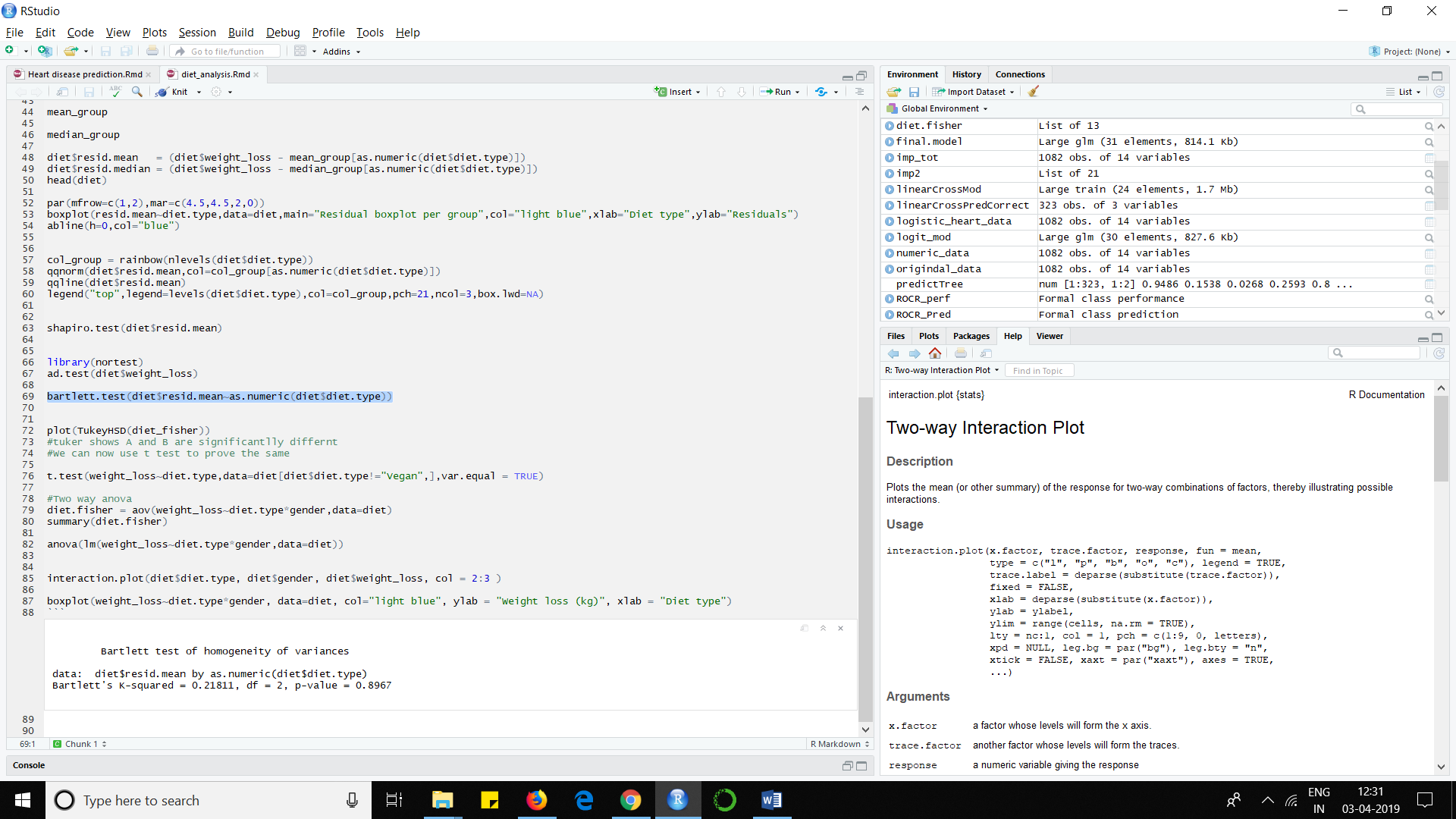


Figure 7: Bartlett’s test output

The above results indicate a high p-value of 0.8967 which means that null hypothesis is accepted and conclude that the variances are indeed homogeneous.

**Post hoc test:**

The ANOVA tests indicate if there is an overall difference amongst the groups but fail to tell which specific group differed. Post hoc tests tell this information and can confirm where the differences occurred between the groups. Tukey’s honestly significant difference (HSD) has been applied in this study to analyse the groups and the output is as follows:

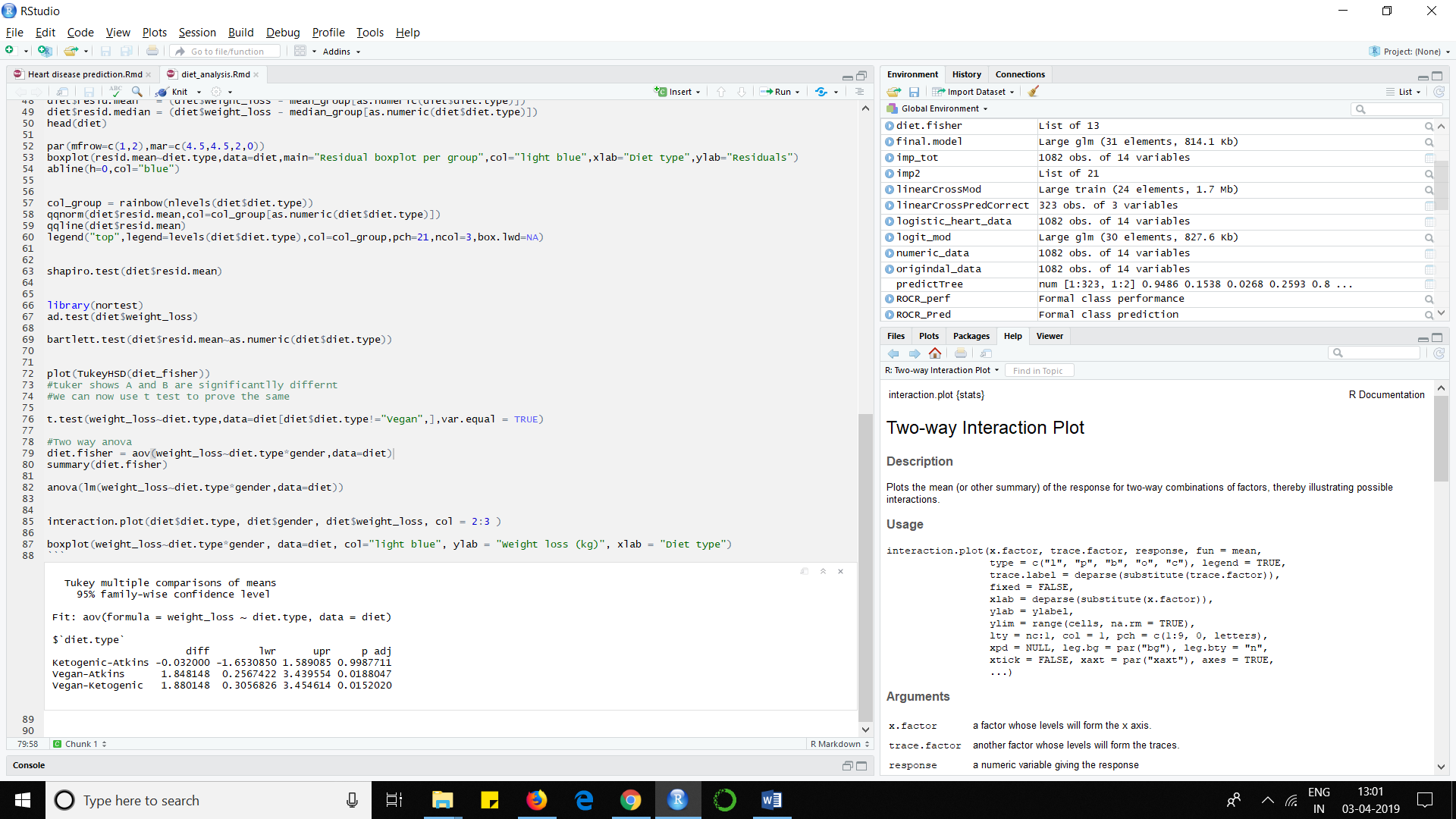


Figure 8: Tukey HSD test output

The “diff” column in the above output indicates the difference between the two groups and a negative value indicates that the second value is higher than the first. The p adj value indicates that the values “Ketogenic – Atkins” are quite similar to each other whereas they differ significantly when compared with “Vegan” diet. The same can be confirmed visually as shown below:



Figure 9: Plot of Tukey HSD test output

**Two – way ANOVA:**

The study also analyses the effect of gender on weight loss along with the effect of interaction between gender and diet type on weight loss using the two – way ANOVA. Performing the two – way ANOVA, the following table is formed:

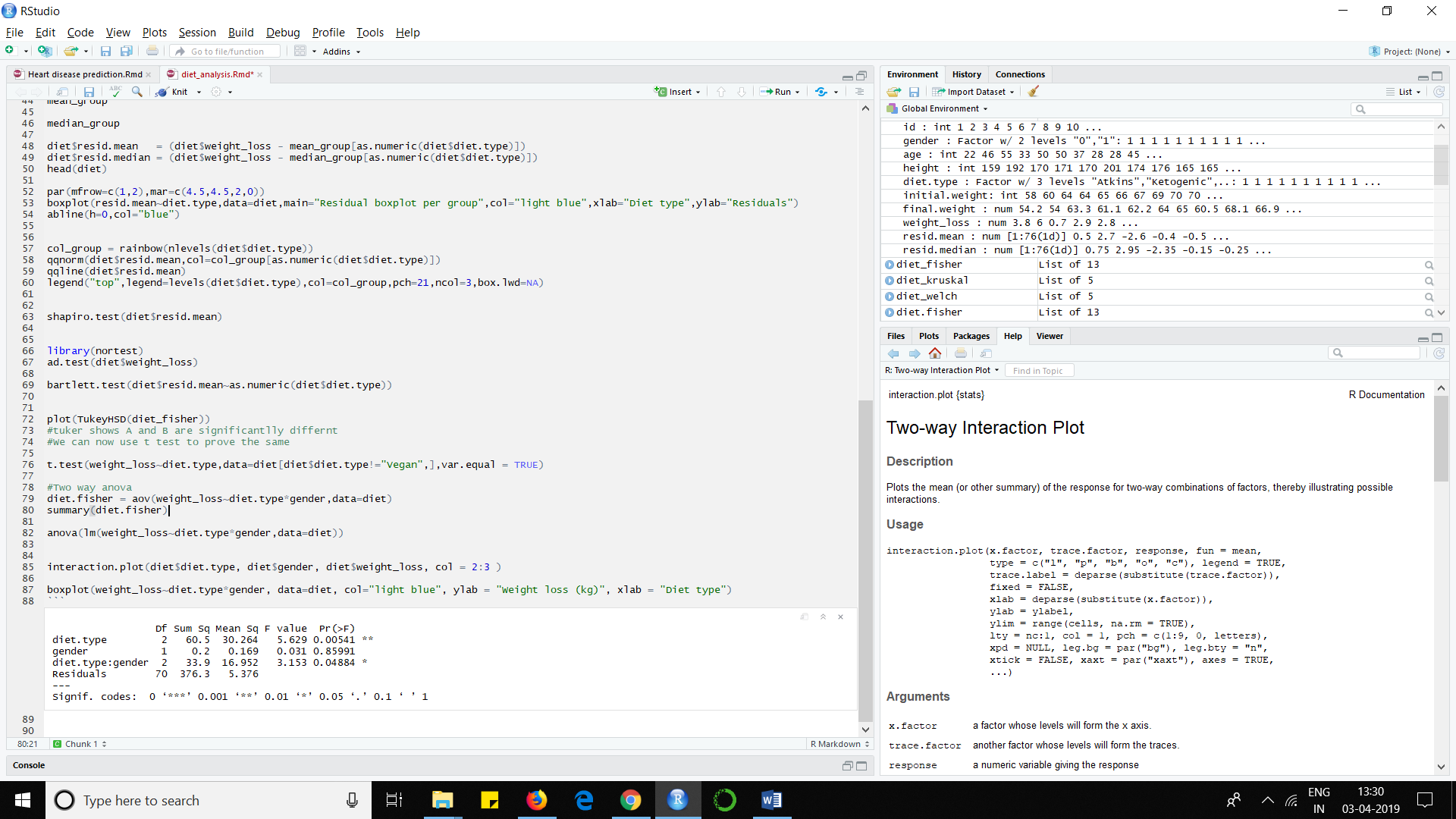


Figure 10: Two – way ANOVA table

The p value of “diet.types” indicate that it is significant and it’s values affects the column “weight loss”. The output also shows that the variable “gender” is not significant but the interaction between the two variables”Diet.type : gender” prove to be significant and thus we can conclude that gender and diet type affects weight loss. We can also show the same through interaction plot as follows:

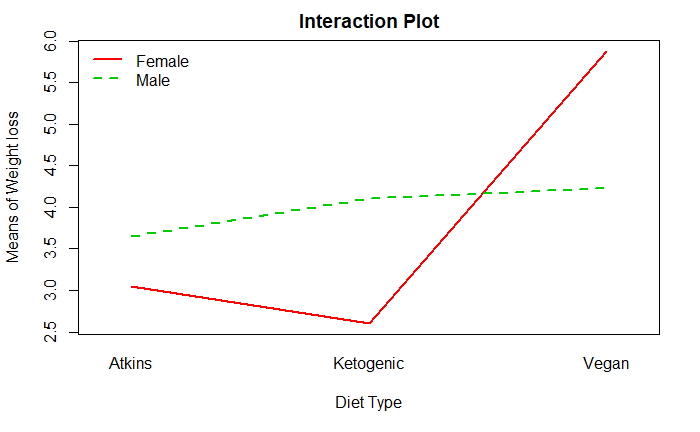


Figure 11: Interaction plot

The interaction plot shows that for a “male” all the three diet types of similar effect and “Vegan” diet as the best results on “males”. Whereas for females the results are the worst for “Ketogenic” diet and is by far the best when “Vegan” diet is followed.

Similarly, the study then analyses the effect of age on diet types. The summary of the ANOVA is shown below when age is also one of the factors: -

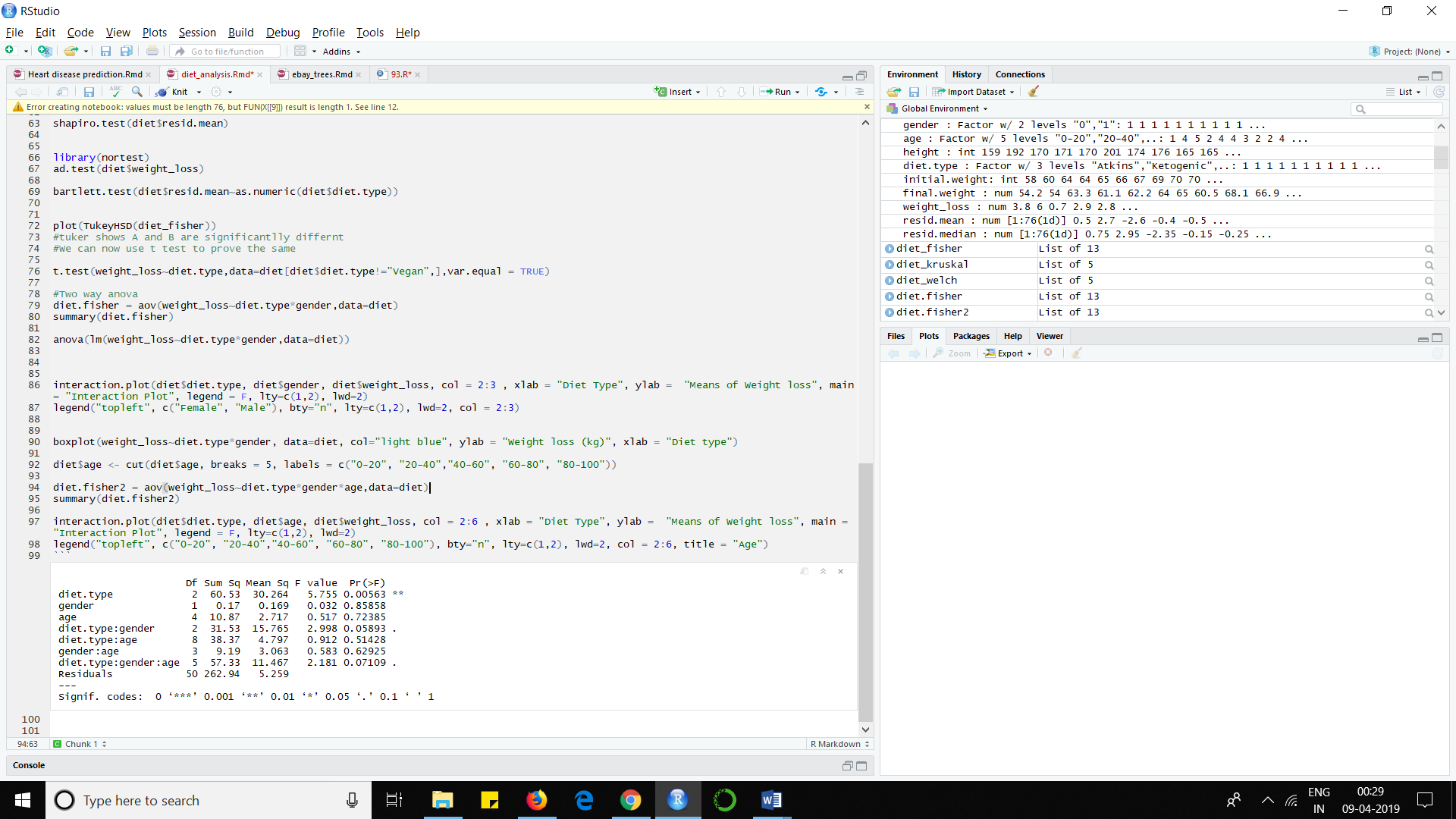


Figure 12: Summary of the ANOVA table

Analysing the table, we can conclude that the interaction between the three factors “gender”, “age, and “diet.types” is not significant on the weight loss. We can conclude that age is not a significant factor to influence weight loss when following one of these three diet types under study.

**Project Outcome:**

After analysing the data, the results conclude that the three diets considered indeed have significantly different results when followed, the best results shown by “Vegan” diet and similar results shown by “Ketogenic” and “Atkins”.

The results also show diet types are in turn affected based on genders. In general, the data concludes that “Ketogenic” has better results than “Atkins” by a minute margins on both males and females. “Vegan” diet has significantly better results than other two diets on females and on males its influence is slightly better than the other two diet types.

**Topic 2:**

**Introduction:**

Education provides individual children with the knowledge and skills necessary to advance themselves and their nation economically. Socioeconomic factors, such as family income level, parents' level of education, race and gender, all influence the quality and availability of education as well as the ability of education to improve life circumstances.

**Family Income Level**

A family's financial status influences several factors that can help or hinder a child in gaining an education. Wealthy families have the financial resources to send a son or daughter to high-quality schools, hire tutors and obtain supplemental education sources. In some countries, students from low-income families may not even be able to attend school; in the U.S., low-income families are limited mostly to public schools while wealthier families can afford to send their children to private schools. Financial stress on the parents can cause a child to leave school early to work. Worries about financial hardship at home can negatively affect low-income children's ability to learn.

**Parents' Level of Education**

Parents' education level directly correlates to the importance and influence of education in their children's lives. Educated parents can assess a son or daughter's academic strengths and weaknesses to help that child improve overall academic performance. The educated parent also sets expectations of academic performance that propel students forward in their achievement levels. However, even if educated, parents who struggled academically and do not think highly of formalized education may have negative attitudes toward education that can still hinder the child academically.

**Gender**

The availability of education to girls and women varies by country. Restrictions on education for girls and women are based on gender bias prevalent with the culture. Some cultures will allow education for girls and women but limit the content of the education or skew the education to prepare them for a limited number of social roles. In the United States, the availability of education to girls and women expanded to become coeducational in most schools within the 20th century.

**Race**

While race is not a predictor of how a student will perform in school, African American students have trailed behind European American students in reading and mathematics. This phenomenon may occur less because of race and more because of family income level. The No Child Left Behind Act of 2002 seeks to improve academic performance for students in predominantly African American or Hispanic schools by placing an emphasis on teacher quality and performance.

**Problem Statement:**

In this study we want to analyse and predict how different socioeconomic factors affects a child performance in a math exam; try to understand what are the factors that heavily affects the child’s performance.

**Statistical Procedure:**

This study is performed using Multiple Linear Regression for predicting the math score based on different factors such as reading score, writing score, gender, race, parents education etc , Data collected would have to pass a health check to see if all the rows have values and don’t contain nulls also the rows should contain meaningful data for the regression. Collinearity is checked using the VIF test to determine highly collinear variables. Dummies are created for different groups to model data. Finally, we divide the entire data into training and testing dataset to predict the new math scores.

**Data Description:**

The dataset contains 1000 observations with 8 variables. The 7 variables used in this study are:

1. Gender – Gender of the student i.e., Male or Female
2. Race/Ethnicity – Race or ethnicity of the student. It is divided into 5 groups i.e., Group A, Group B, Group C, Group D, Group E
3. Parental level of education – Level of education attained by the parent. There are 6 different level of parental education i.e., High School Dropout, High School, Associate Degree, ﻿Master’s degree, College Dropout, Bachelor’s Degree.
4. Lunch Type – Type of lunch the student has access to. Standard or Free/Reduced
5. Test Preparation – Kind of preparation the student has for the test. None or Completed
6. Reading Score – It is a continuous variable and it signifies the reading score of the student
7. Writing Score – It is a continuous variable and it signifies the writing score of the student
8. Math Score – It is a continuous variable and it signifies the math score of the student

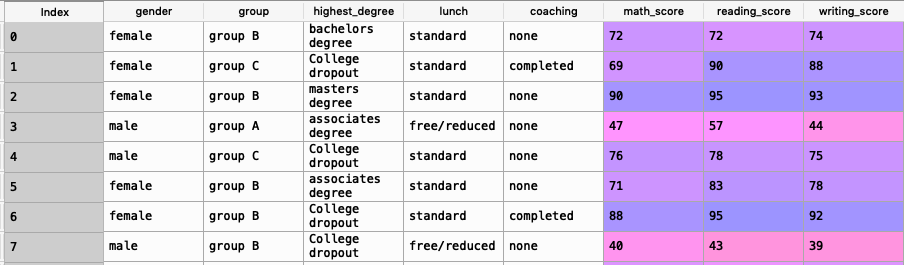
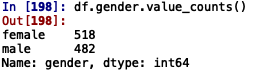


Figure 1: A snapshot of the data used for Regression

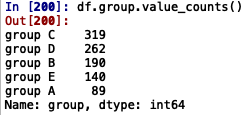
**Exploratory Analysis:**

To do an exploratory analysis we should know the data better and understand how structured it is to perform any further analysis.

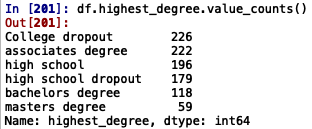
1. No. of male and female students



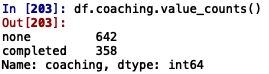
1. Number of students belonging to each Race/Ethnicity



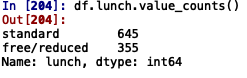
1. Number of students under each Parental level of education



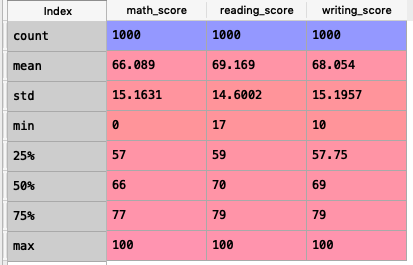
1. Number of students with different Coaching provided



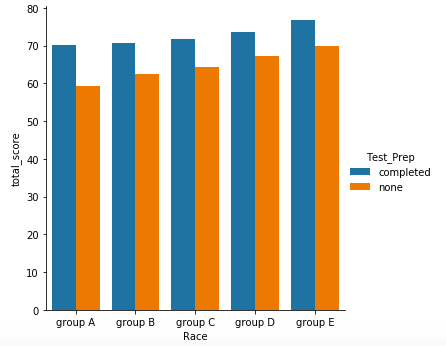
1. Number of students under each Lunch type



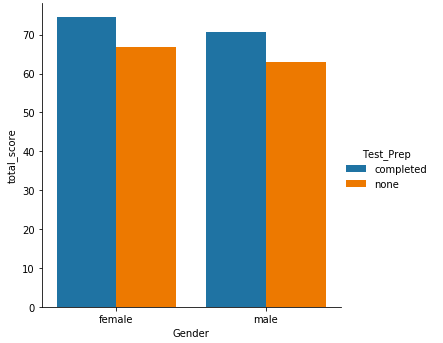
1. Summary of Math Score, Reading Score and Writing Score in the data



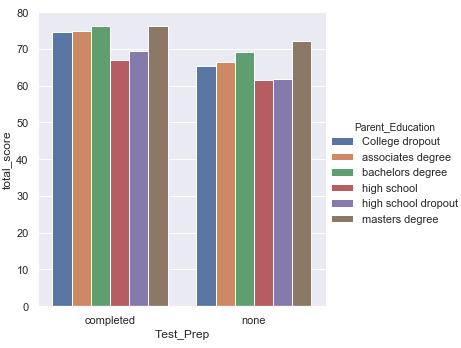
1. Mean Score obtained by students of different Races given the kind of Test Prep they had.



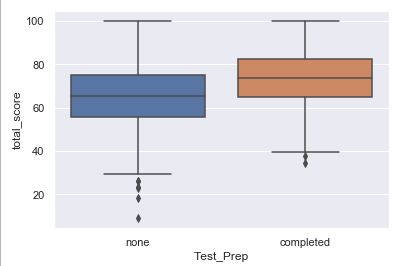
1. Mean Score obtained by students of different Gender given the kind of Test prep they had.



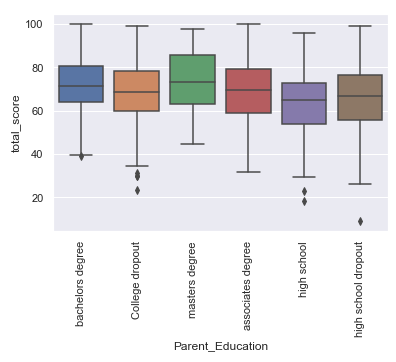
1. Mean Score obtained by students with different level of parental education given the kind of Test prep they had.



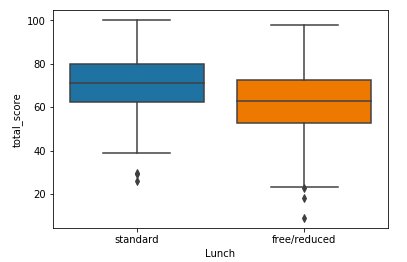
1. Boxplot of Test Prep vs the Total Score.



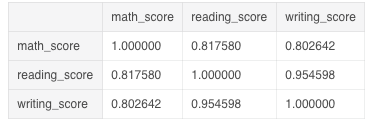
1. Boxplot of Parents Education vs Total Score

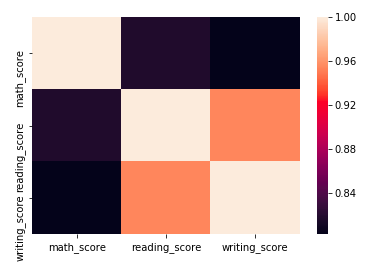


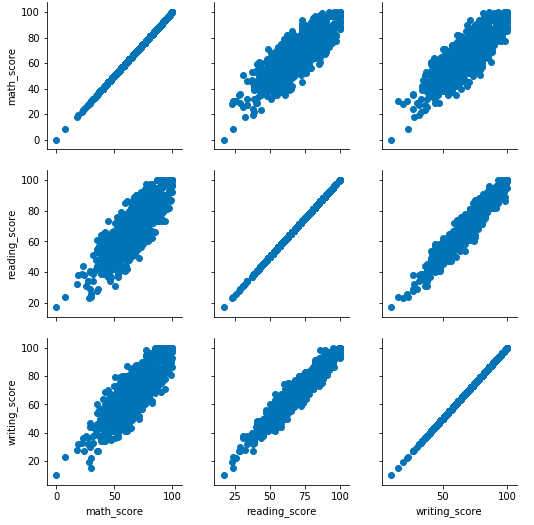
1. Boxplot of Kind of lunch provided vs Total Score



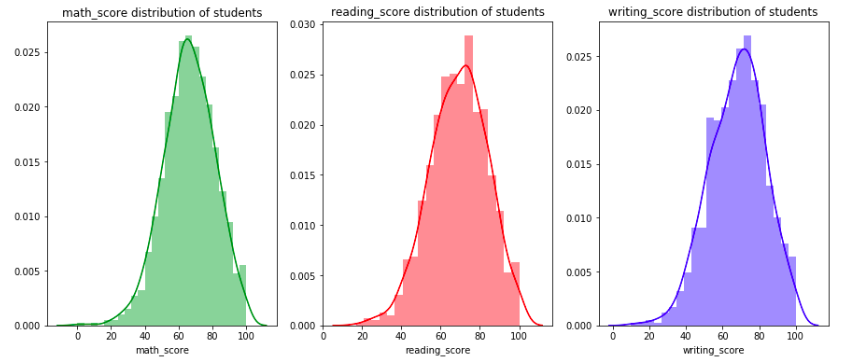
1. Correlation between Match Score, Reading Score and Writing Score







1. Distribution of Math Score, Reading Score and Writing Score



**Data Pre-processing**

Data obtained from Kaggle had to be processed to use it for the regression model since columns such as Gender, Group, Highest degree, Lunch and Coaching contains text data which could not be used for performing the regression so these variables containing text data were converted to Categorical Integer variables and were assigned dummy values of 0’s and 1’s.

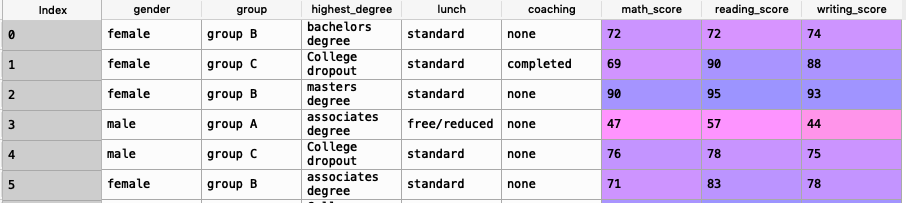


Figure 2: Data with features containing Text Data

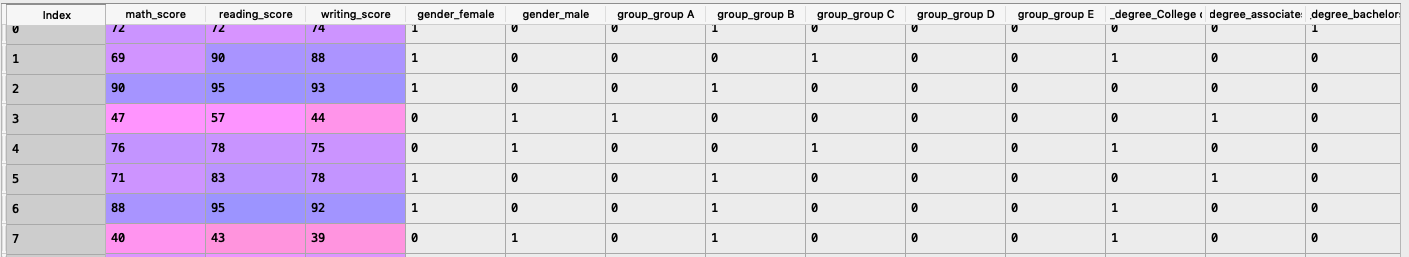


Figure 3: Data with features containing dummy data

**Multiple Linear Regression**

In this study we are predicting the Math Score based on features such as Gender, Group, Highest Degree, Lunch, Coaching, Reading Score, Writing Score.

No. of Students = 1000

Training Data = 400

Testing Data = 600

**1st Approach:**

In this approach columns such as Gender, Group, Highest Degree, Lunch were created into dummy variables and given values of 0’s and 1’s, along with writing score and reading score the above-mentioned dummy variables were used to predict the Math Score.

We obtained the following results using this approach:

Below are the regression coefficients after using the training data:

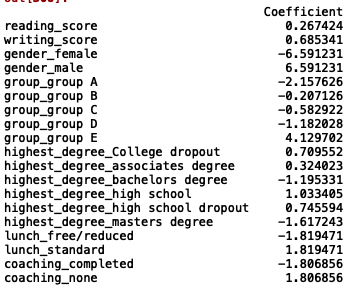


Figure 4: Regression Coefficients after using training data

R-Squared Value:



This shows that 87.02% changes in Math score can be attributed to the predictor variables used in the model

Below table shows a comparison of Actual and Predicted Math Scores:

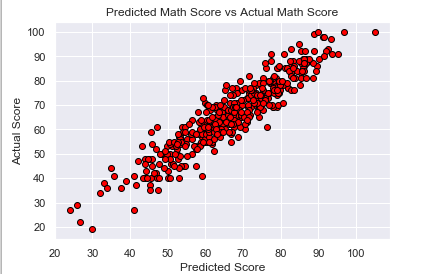


Figure 5: Scatter Plot of Actual Math Score and Predicted Math Score

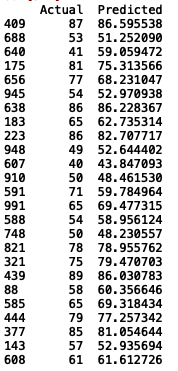


Figure 6: Comparison of Actual vs Predicted Math Scores

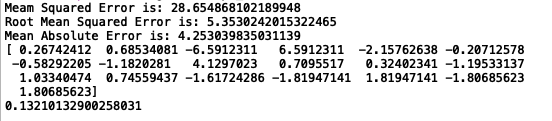


Figure 7: Regression Outcomes

From the above figure we get the following values:

Mean Squared Error: ﻿28.654868102189948

Comment: We are getting a very high value for Mean Squared Error because our data is getting skewed because of converting the categorical data to dummy variables.

Root Mean Squared Error: ﻿5.3530242015322465

Comment: We are getting a very high value for Root Mean Squared Error because our data is getting skewed because of converting the categorical data to dummy variables.

﻿Mean Absolute Error is: 4.253039835031139

Comment: We are getting a very high value for Mean Absolute Error because our data is getting skewed because of converting the categorical data to dummy variables.

Regression Coefficients:

﻿[ 0.26742412 0.68534081 -6.5912311 6.5912311 -2.15762638 -0.20712578

-0.58292205 -1.1820281 4.1297023 0.7095517 0.32402341 -1.19533137

1.03340474 0.74559437 -1.61724286 -1.81947141 1.81947141 -1.80685623

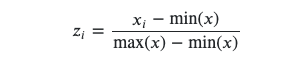
1.80685623]

Intercept: 0.13210132900258031

What is the solution to above problem?

Since after converting the text variables to dummy variables the scale has changed for the categorical variables and Reading Score, Writing Score. So to overcome this issue data has to be normalized in order to achieve the required results and better accuracy.

Normalization is done using the formula:



**2nd Approach**

In this approach all the predictor variables are normalized using the formula mentioned above and the data looks something like this:



Figure 8: Snapshot of normalized data

We obtained the following results using this approach:

Below are the regression coefficients after using the training data:

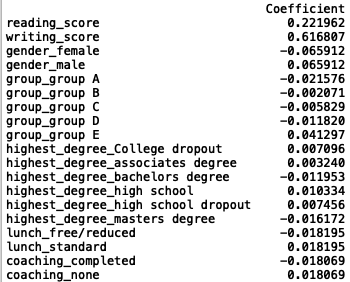


Figure 9: Regression Coefficients after using Training Data

R-Squared Value:



This shows that 87.02% changes in Math score can be attributed to the predictor variables used in the model

Below table shows a comparison of Actual and Predicted Math Scores:

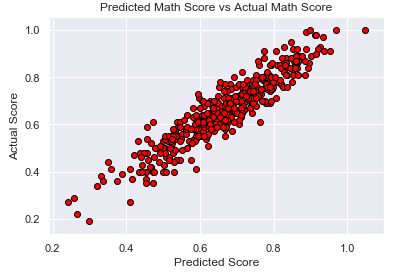


Figure 10: Scatter Plot of Actual Math Score and Predicted Math Score

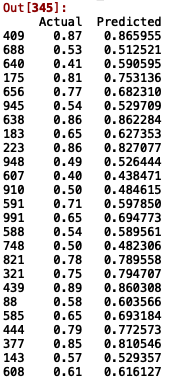


Figure 11: Comparison of Actual vs Predicted Math Scores

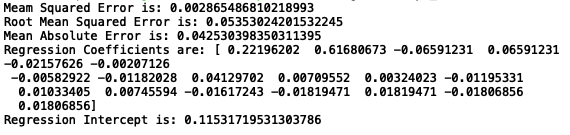


Figure 12: Regression Outcomes

From the above figure we get the following values:

Mean Squared Error: ﻿ ﻿0.002865486810218993

Comment: We are getting a very low/good value for Mean Squared Error because our data is now normalized, and all the parameters are on the same scale so it is giving us the best accuracy with lowest error.

Root Mean Squared Error: ﻿ ﻿0.05353024201532245

Comment: We are getting a very low/good value for Root Mean Squared Error because our data is now normalized, and all the parameters are on the same scale so it is giving us the best accuracy with lowest error.

﻿Mean Absolute Error is: ﻿0.042530398350311395

Comment: We are getting a very low/good value for Root Mean Absolute Error because our data is now normalized, and all the parameters are on the same scale so it is giving us the best accuracy with lowest error.

Regression Coefficients:

﻿[ 0.22196202 0.61680673 -0.06591231 0.06591231 -0.02157626 -0.00207126

-0.00582922 -0.01182028 0.04129702 0.00709552 0.00324023 -0.01195331

0.01033405 0.00745594 -0.01617243 -0.01819471 0.01819471 -0.01806856

0.01806856]

Intercept: ﻿0.11531719531303786

**Project Outcome:**

After analysing the data, we have come to a strong conclusion that socioeconomic factors hugely impact a student’s Math Score. Also, it can be concluded that a student who has good reading abilities would in fact perform good on his Math Exam as well.

**APPENDIX:**

ANOVA with R: analysis of the diet dataset

library(plyr)  
diet <- read.csv("C:/Users/Dr. Suresh babu/Desktop/Stats Project/diet.csv")  
  
levels(diet$gender)

## [1] "Female" "Male"

diet$gender <- revalue(diet$gender, c("Female"=0, "Male"=1))  
diet$gender

levels(diet$diet.type)

diet$diet.type <- revalue(diet$diet.type, c("A"= "Atkins", "B"= "Ketogenic", "C"="Vegan"))  
diet$diet.type

diet$weight\_loss = diet$initial.weight - diet$final.weight  
head(diet)

boxplot(weight\_loss~diet.type, data=diet, col="light blue", ylab = "Weight loss (kg)", xlab = "Diet type")

diet\_fisher = aov(weight\_loss~diet.type, data = diet)  
summary(diet\_fisher)

diet\_welch = oneway.test(weight\_loss~diet.type, data = diet)  
print(diet\_welch)

diet\_kruskal = kruskal.test(weight\_loss~diet.type, data = diet)  
print(diet\_kruskal)

mean\_group = tapply(diet$weight\_loss,diet$diet.type,mean)  
median\_group = tapply(diet$weight\_loss,diet$diet.type,median)  
  
mean\_group

median\_group

#diet$resid.mean = (diet$weight\_loss - mean\_group[as.numeric(diet$diet.type)])  
diet$resid.median = (diet$weight\_loss - median\_group[as.numeric(diet$diet.type)])  
head(diet)

par(mfrow=c(1,2),mar=c(4.5,4.5,2,0))   
boxplot(resid.mean~diet.type,data=diet,main="Residual boxplot per group",col="light blue",xlab="Diet type",ylab="Residuals")  
abline(h=0,col="blue")  
  
  
col\_group = rainbow(nlevels(diet$diet.type))  
qqnorm(diet$resid.mean,col=col\_group[as.numeric(diet$diet.type)])  
qqline(diet$resid.mean)  
legend("top",legend=levels(diet$diet.type),col=col\_group,pch=21,ncol=3,box.lwd=NA)

shapiro.test(diet$resid.mean)

library(nortest)

ad.test(diet$weight\_loss)

bartlett.test(diet$resid.mean~as.numeric(diet$diet.type))

plot(TukeyHSD(diet\_fisher))  
#tuker shows A and B are significantlly differnt  
#We can now use t test to prove the same  
  
t.test(weight\_loss~diet.type,data=diet[diet$diet.type!="Vegan",],var.equal = TRUE)

#Two way anova  
diet.fisher = aov(weight\_loss~diet.type\*gender,data=diet)  
summary(diet.fisher)

anova(lm(weight\_loss~diet.type\*gender,data=diet))

interaction.plot(diet$diet.type, diet$gender, diet$weight\_loss, col = 2:3 , xlab = "Diet Type", ylab = "Means of Weight loss", main = "Interaction Plot", legend = F, lty=c(1,2), lwd=2)  
legend("topleft", c("Female", "Male"), bty="n", lty=c(1,2), lwd=2, col = 2:3)

boxplot(weight\_loss~diet.type\*gender, data=diet, col="light blue", ylab = "Weight loss (kg)", xlab = "Diet type")  
  
diet$age <- cut(diet$age, breaks = 5, labels = c("0-20", "20-40","40-60", "60-80", "80-100"))  
  
diet.fisher2 = aov(weight\_loss~diet.type\*gender\*age,data=diet)  
summary(diet.fisher2)

interaction.plot(diet$diet.type, diet$age, diet$weight\_loss, col = 2:6 , xlab = "Diet Type", ylab = "Means of Weight loss", main = "Interaction Plot", legend = F, lty=c(1,2), lwd=2)  
legend("topleft", c("0-20", "20-40","40-60", "60-80", "80-100"), bty="n", lty=c(1,2), lwd=2, col = 2:6, title = "Age")