**10/14/2020**

**MAT 502 - Final Project**

**Mariah Bastos and Andrea Gonzalez**

**Nutrition Study Analysis**

The current project describes application of time series models for nutrition study data modeling and forecasting. Linear models applied at the project involve also the analysis of the relationship between calorie intake, fat, and fiber within genders to determine if the cholesterol level can be predicted. We also applied these models to the analysis of the relationship between cholesterol intake and gender. Our purpose with this study is for people to be aware about the consequences that an unhealthy diet can cause. To show them which factors increase their cholesterol so they can be conscious about it, take a proactive approach to solve this problem and keep healthy. Another reason that lead us to do this project was our concern after we collected the data below from the American Heart Association:

* **46% of adults in the U.S. are estimated to suffer from hypertension.**
* **On average, someone dies from CVD every 38 seconds.**
* **Last year there were 389.4 deaths reported from a stroke each day.**

**Study objectives**

Main goal of this project is to define the relationship between the intake of fat, fiber and calories with the amount of cholesterol in men and women's blood; for people to be aware about the consequences that an unhealthy diet can cause. To analyze which factors increase cholesterol level so people can be more conscious about their food choices; analyze the data on R in order to find existing relationships between the variables.

**Data description**

Project uses a dataset obtained from Statistics: Unlocking the Power of Data by Lock, Lock, Lock, Lock, and Lock. https://www.lock5stat.com. This source is a web-based data service for the global user community. Lock5stat is a statistical website with many datasets accessible for everyone. Data frame named NutritionStudy keeps data of 315 candidates on Age, Gender, Smoke, Vitamin, Alcohol, Fat, Fibler, Cholesterol, Calories, VitaminUse and PriorSmoke.

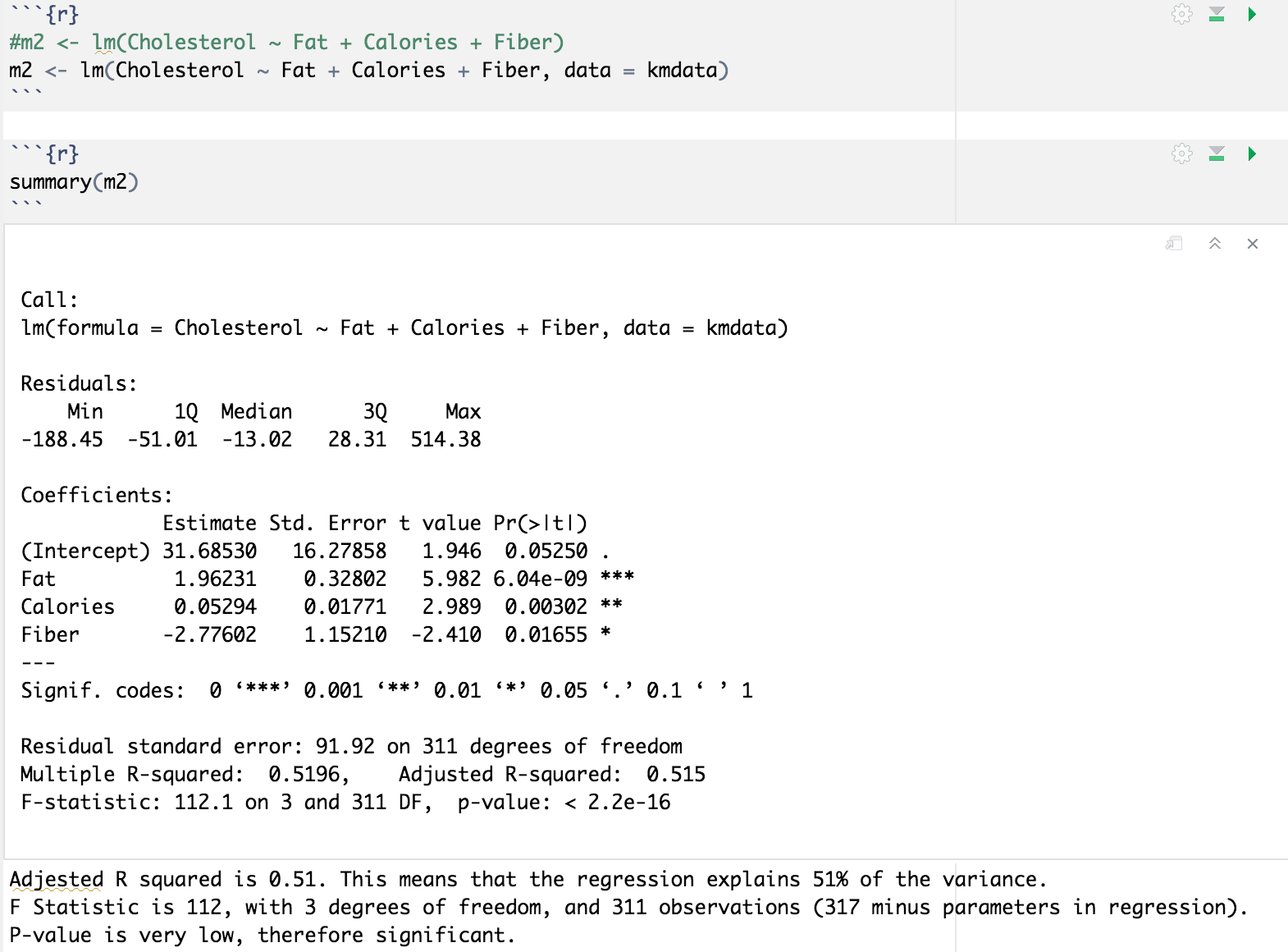
**Linear Model**

We started our project by analyzing the relationship between calorie intake, fat, and fiber within genders to determine if the cholesterol level can be predicted. We stated our null hypothesis and alternative hypothesis to be the following:

**Null Hypothesis (H0): There is a significant relationship between calories, fiber and fat intake that help us predict cholesterol level.**

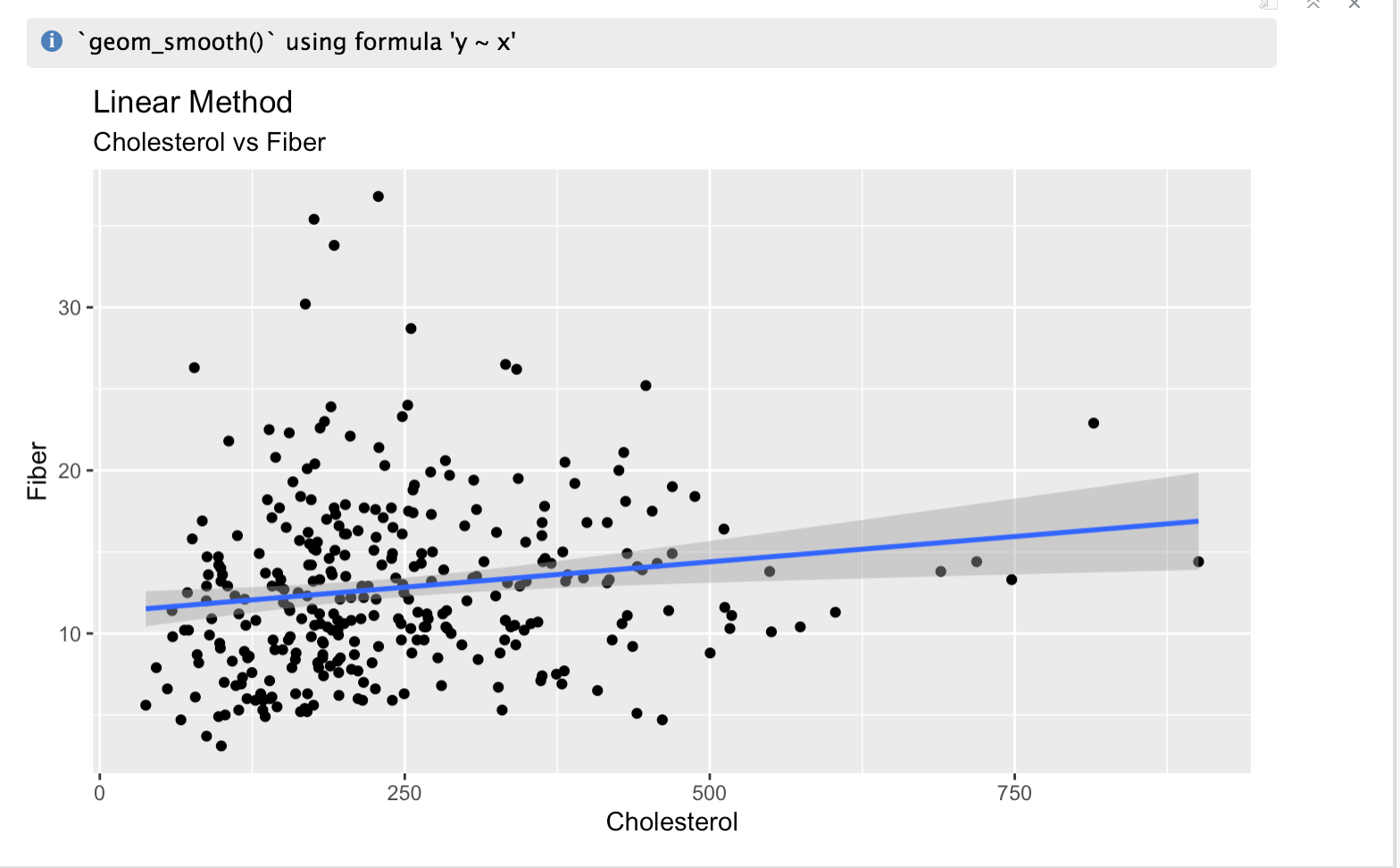
**Alternative Hypothesis (H1): There is not a significant relationship between calories, fiber and fat intake that help us predict cholesterol level**

The chunk below analyzes if there is a significant relationship between fat, calories and fiber intake.

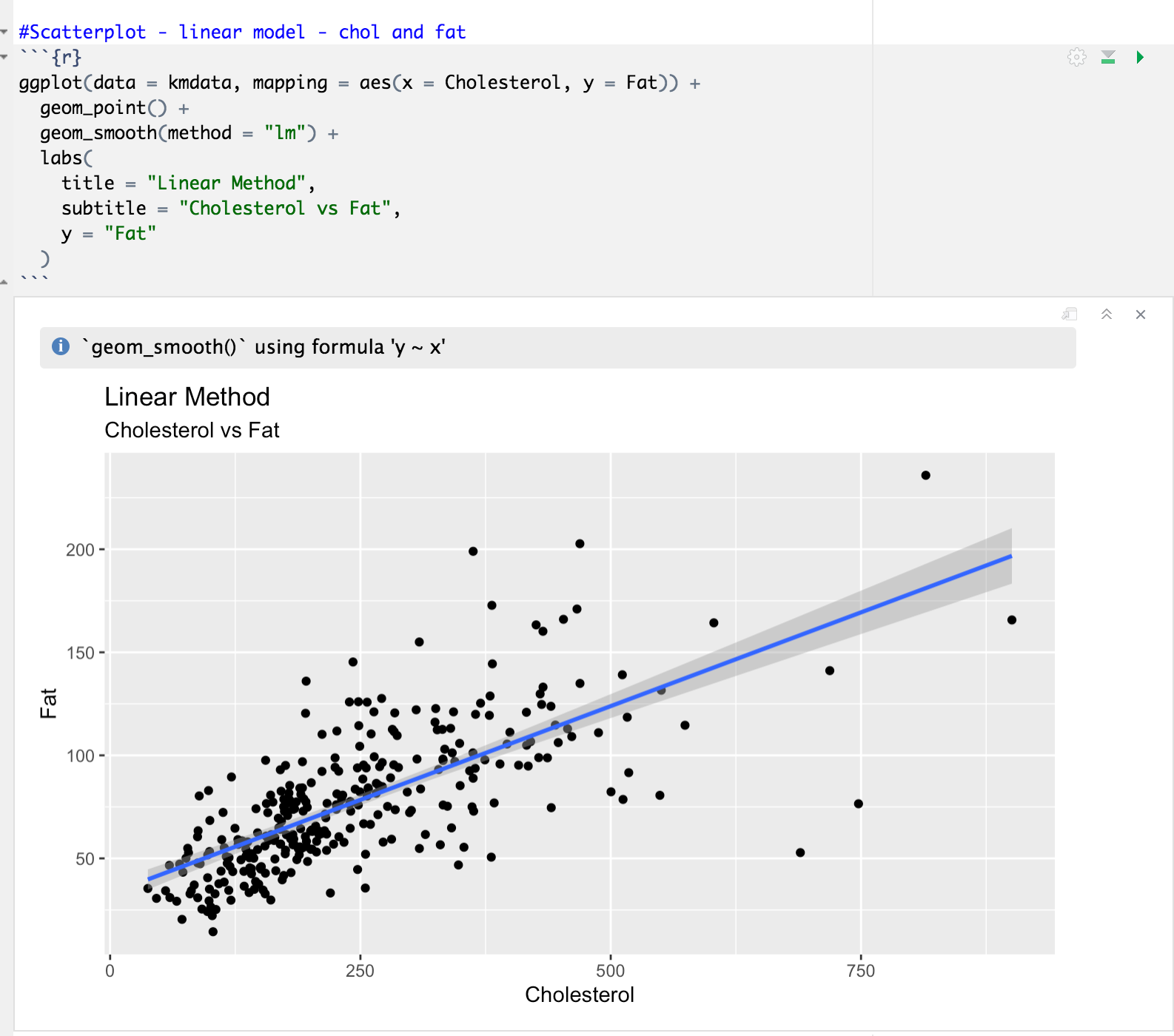


**Linear Model Visualizations**

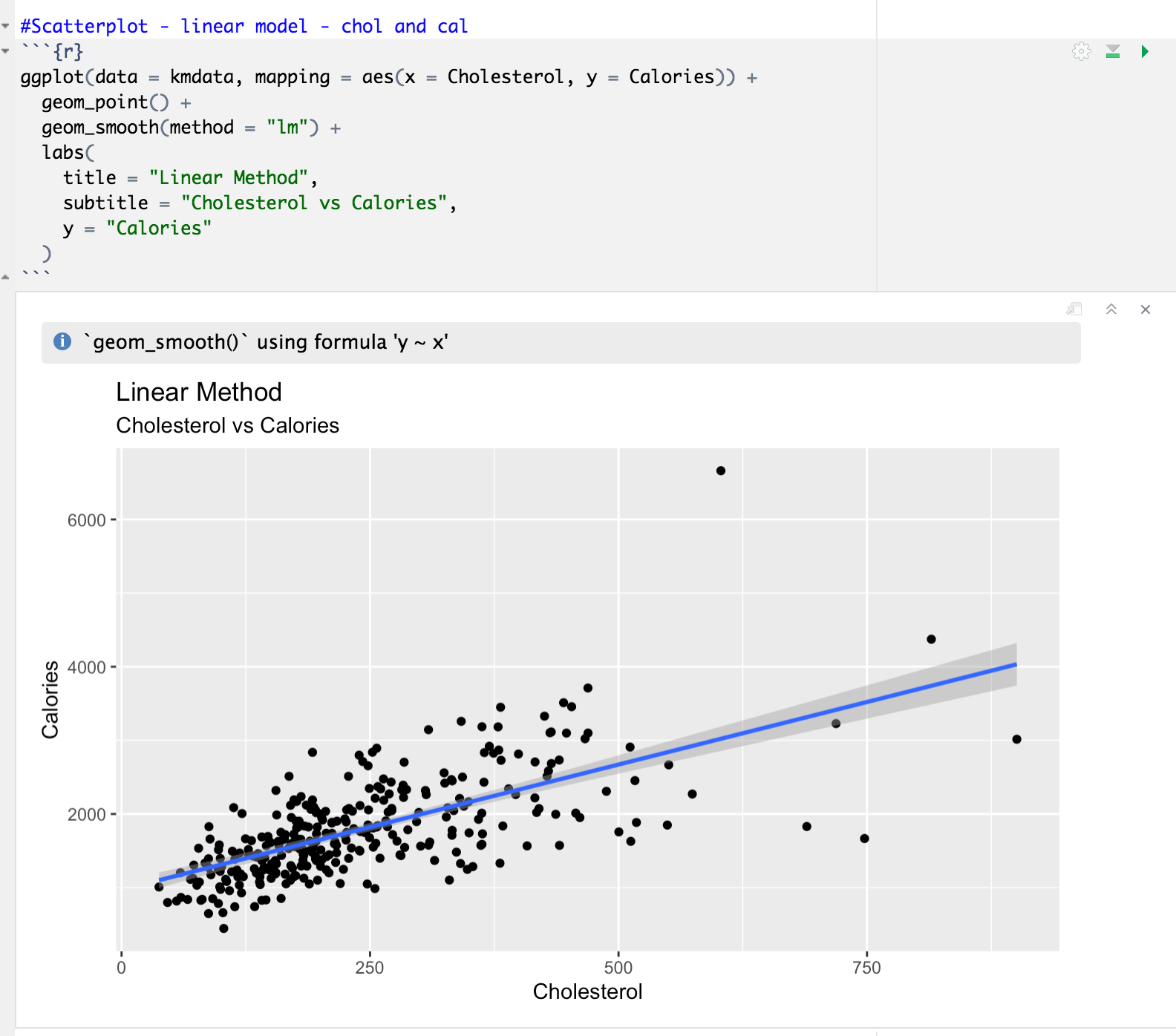
Below we can see a scatterplot specifically using a linear model method to show the relationship between cholesterol and fiber.



Below we can see a scatterplot specifically using a linear model method to show the relationship between cholesterol and fat.



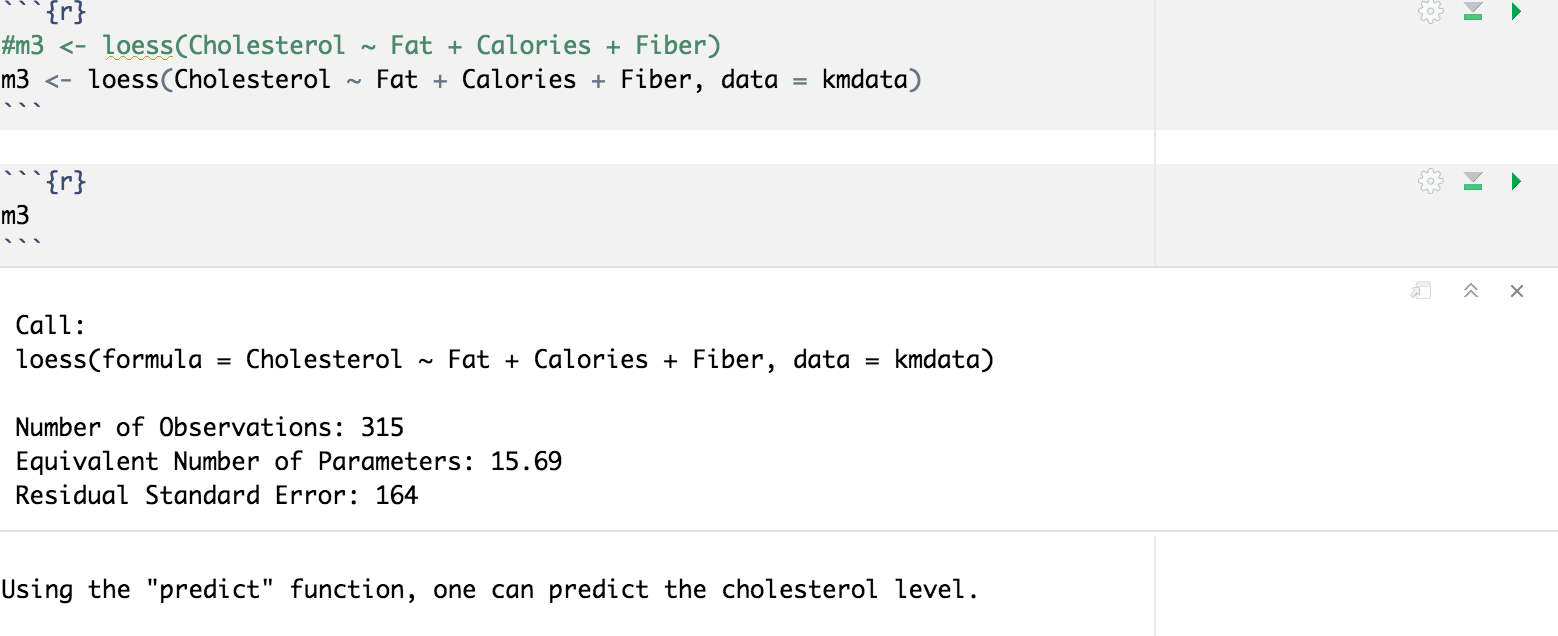
Below we can see a scatterplot specifically using a linear model method to show the relationship between cholesterol and calories.



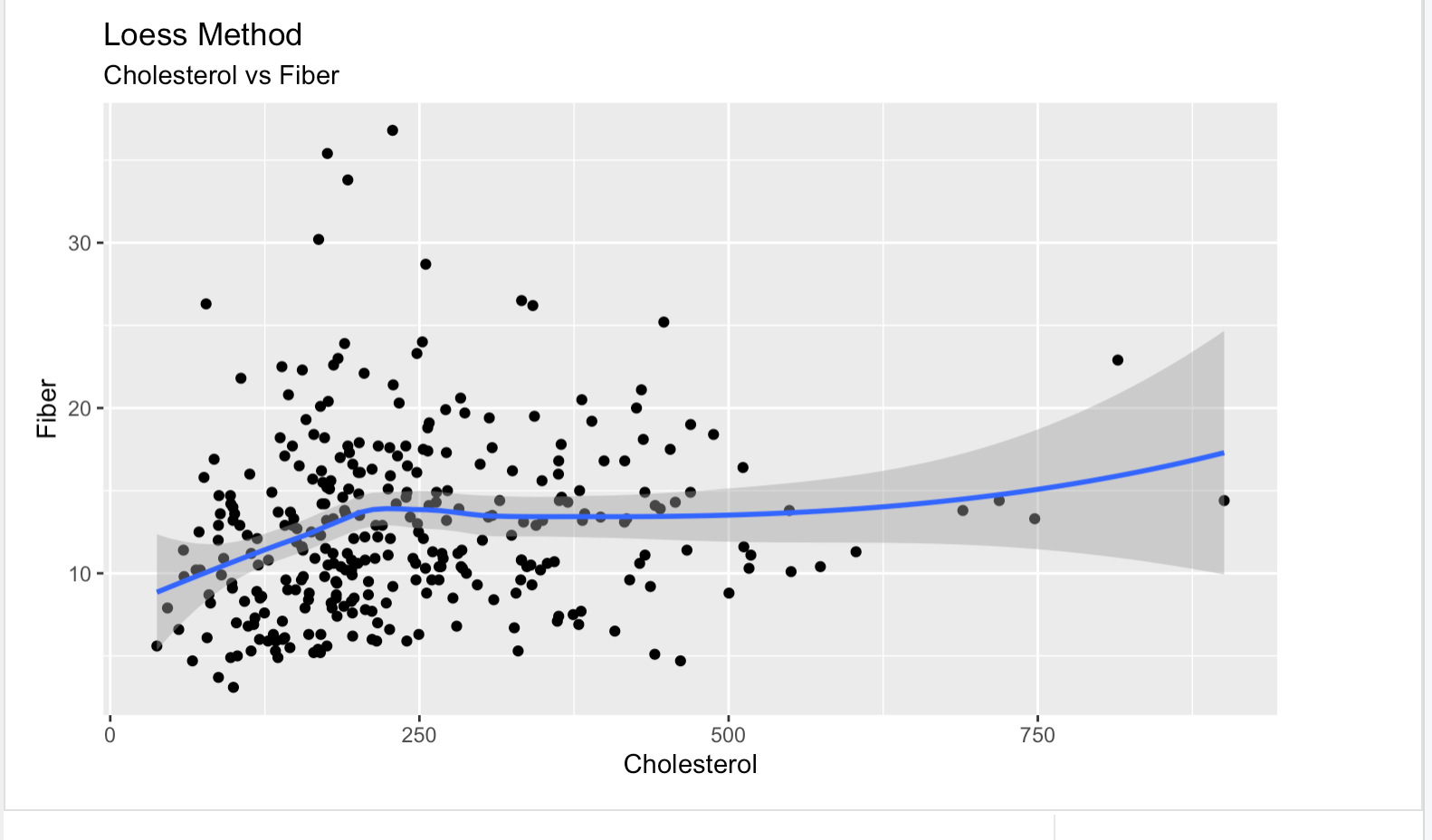
Our linear model resulted in an adjusted R squared with a value of 0.51. This means that the regression explains 51% of the variance between fat, calories and fiber. Based on the results and the plots of the Linear Models we can conclude that our P-value is less than alpha, thus we reject the null hypothesis. Therefore we say that there is not sufficient evidence to claim that there is a significant relationship between fat, calories and fiber that can help us predict the cholesterol level.

**Loess Model**

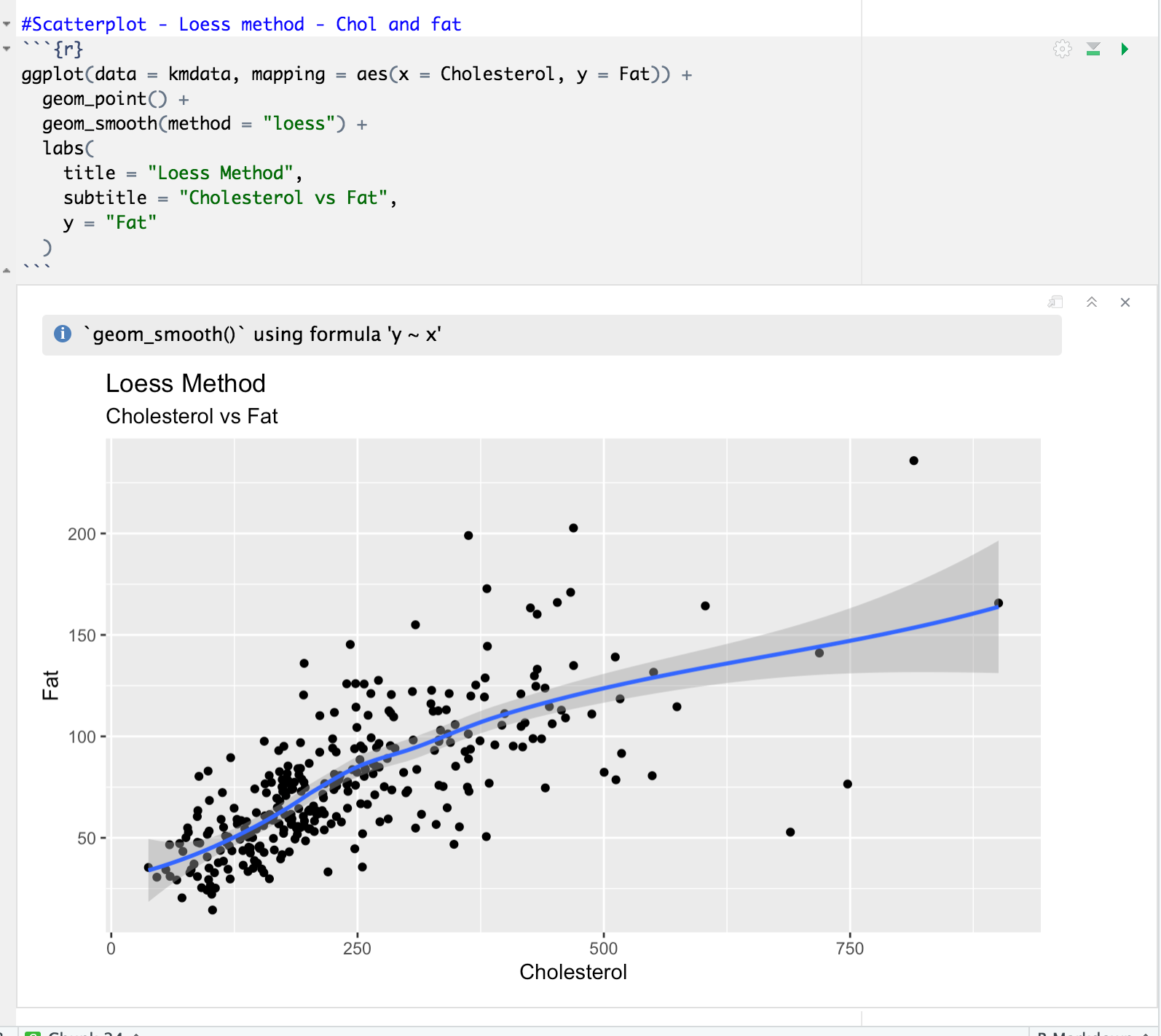
For the Loess Model we analyzed the same relationships and used the same hypothesis as we used on the Linear Model.

****

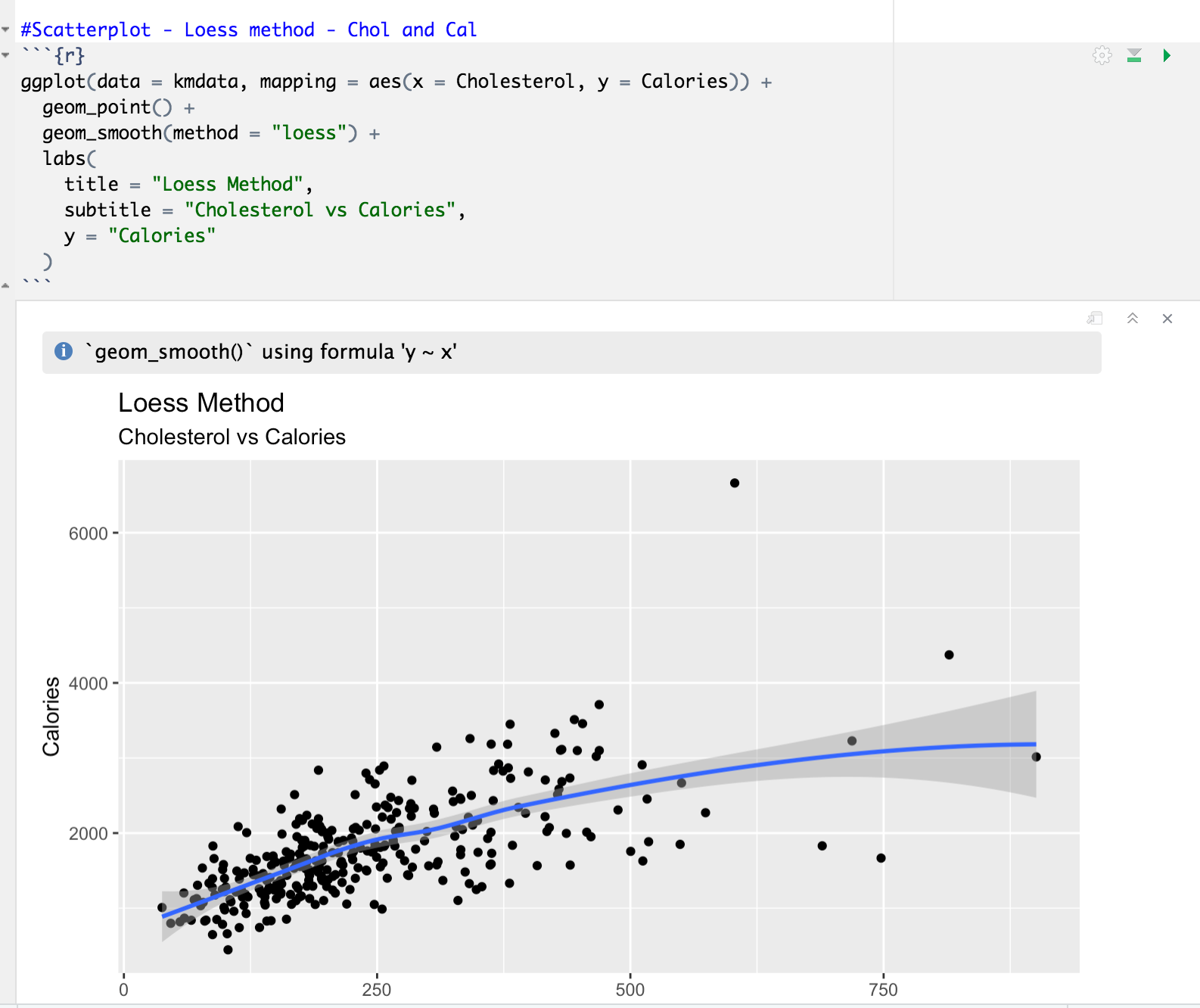
Below we can see a scatterplot specifically using a loess model method to show the relationship between cholesterol and fiber.

****

Below we can see a scatterplot specifically using a loess model method to show the relationship between cholesterol and fat.

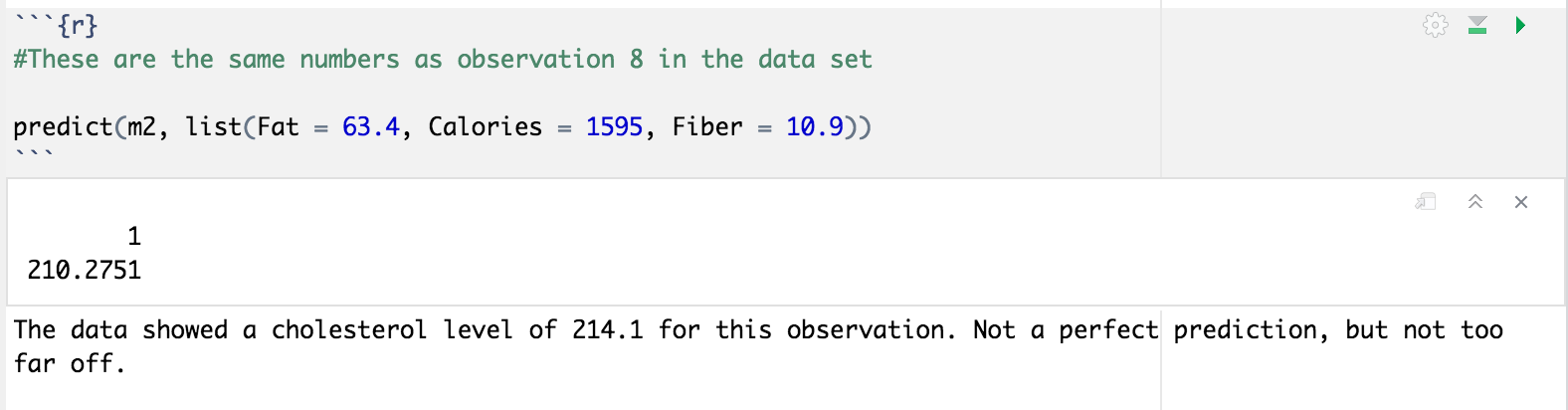
****

Below we can see a scatterplot specifically using a loess model method to show the relationship between cholesterol and calories.

****

Our Loess Model resulted with a P-value that is less than alpha, thus we reject the null hypothesis. According to the loess model we say that there is not sufficient evidence to claim that there is a significant relationship between fat, calories and fiber that can help us predict the cholesterol level.

We tried to do a prediction of our data as shown below but the result of this showed a cholesterol level of 210.27 when the data showed a cholesterol level of 214.1 for this observation. It was not a perfect prediction, but it was not too far off.

****

We then moved on to analyze the relationship between cholesterol intake with genders. We stated our null hypothesis and alternative hypothesis to be the following:

**Null Hypothesis (H0): There is a significant relationship between the intakes of cholesterol according to gender**

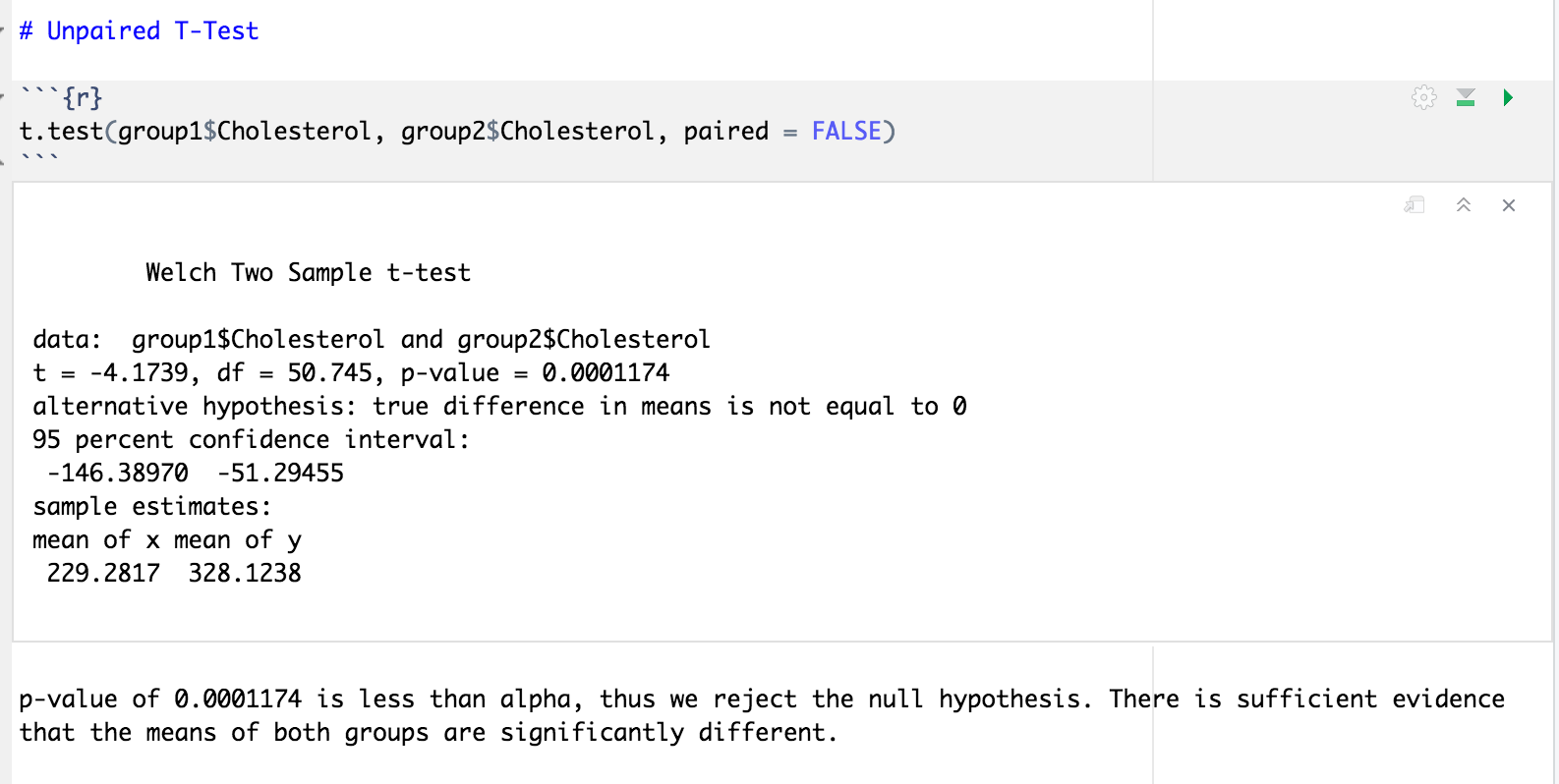
**Alternative Hypothesis (H1): There is not a significant relationship between the intakes of cholesterol according to gender**

We ran a normality test to test the relationship and accuracy of our hypothesis.



For group 1 the P-value of 6.17e-14 is less than alpha, and for group 2 the P-value of 0.2506 is greater than alpha, thus the distributions are not normally distributed.

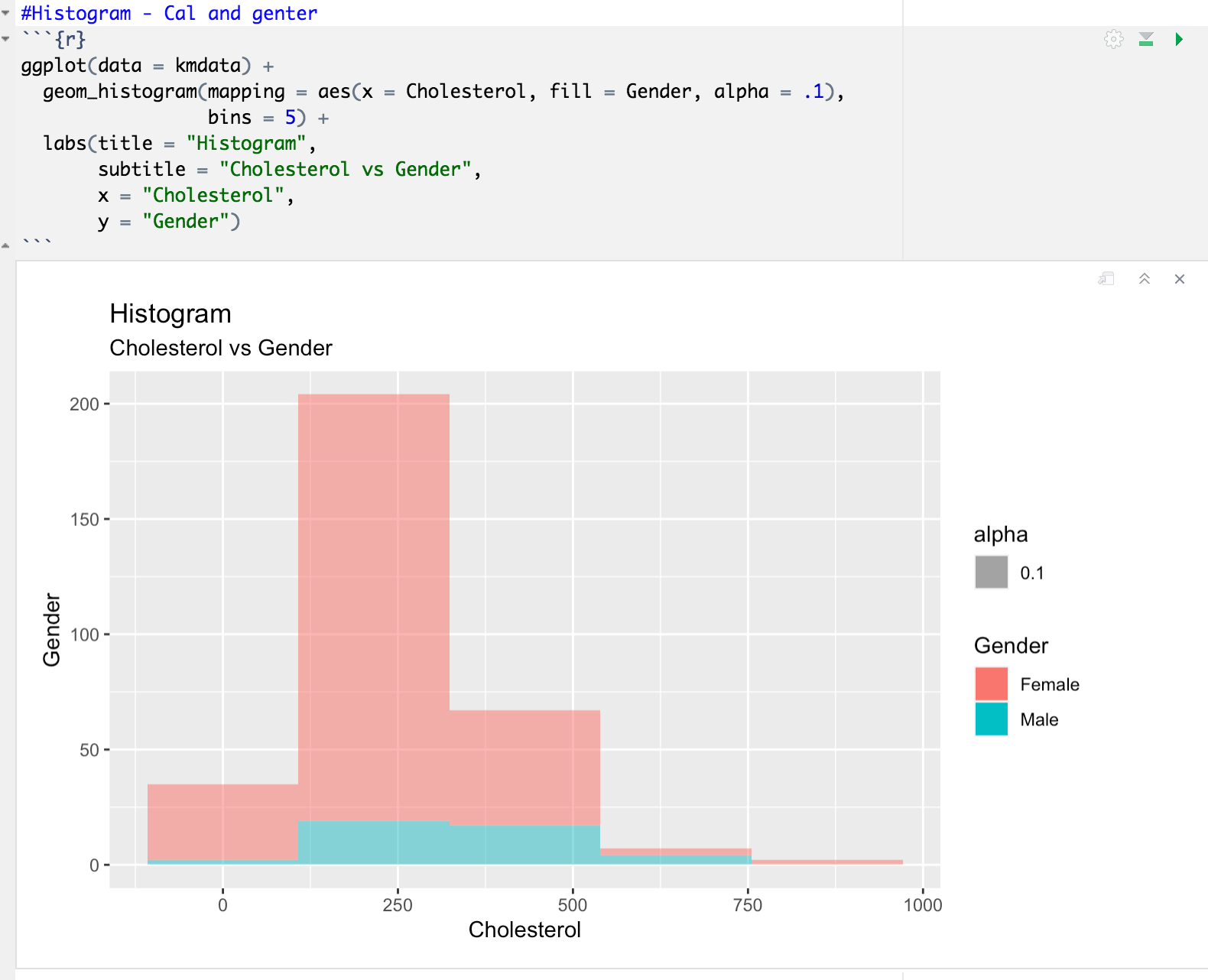
We also ran an unpaired T-Test looking for a normal distribution between our variables.

****

P-value of 0.0001174 is less than alpha, thus we reject the null hypothesis. There is sufficient evidence that the means of both groups are significantly different.

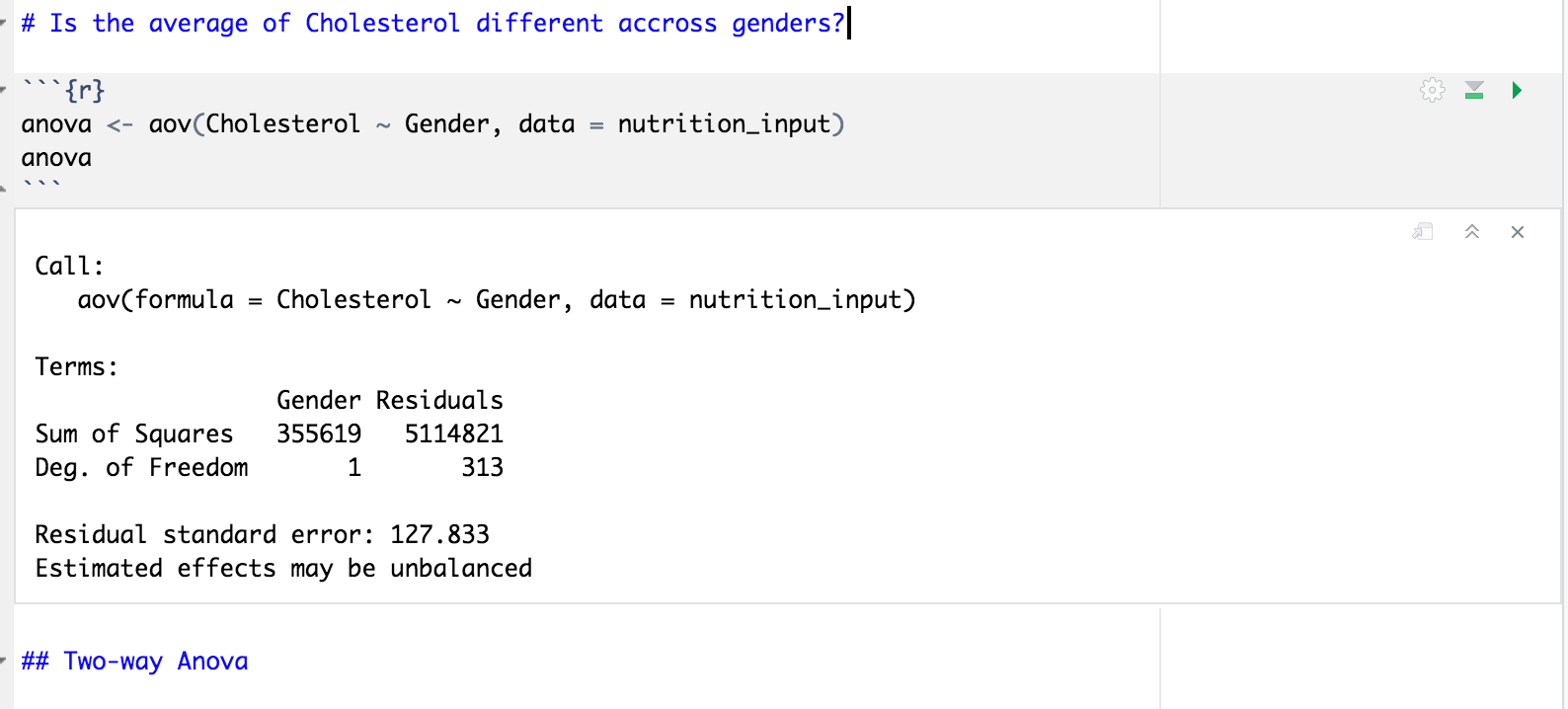
**Visualizations**

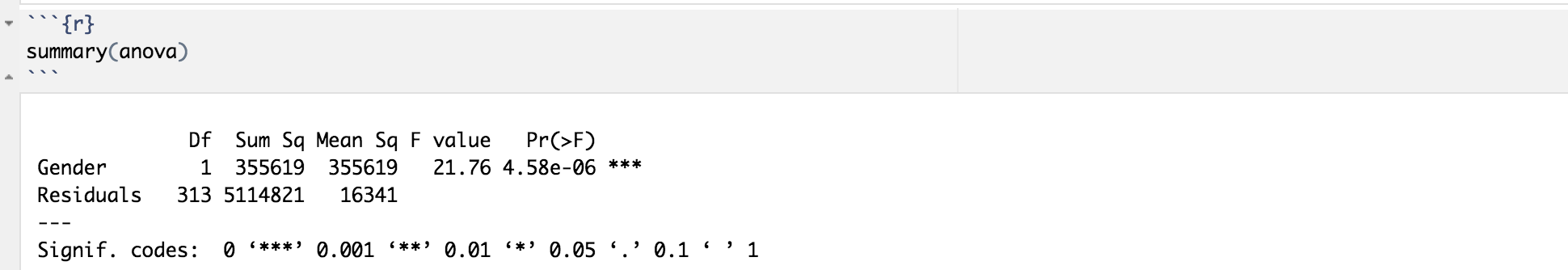
We tried a density plot and a histogram seeking for a visual approach about the relationship between cholesterol and gender.

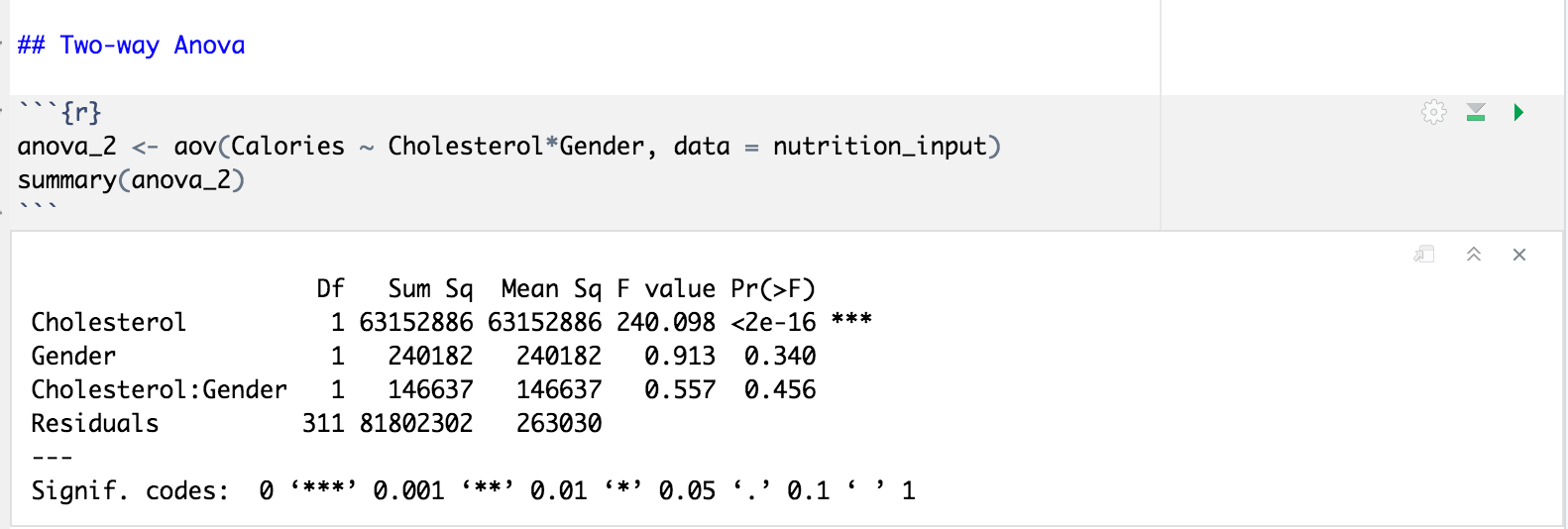
****

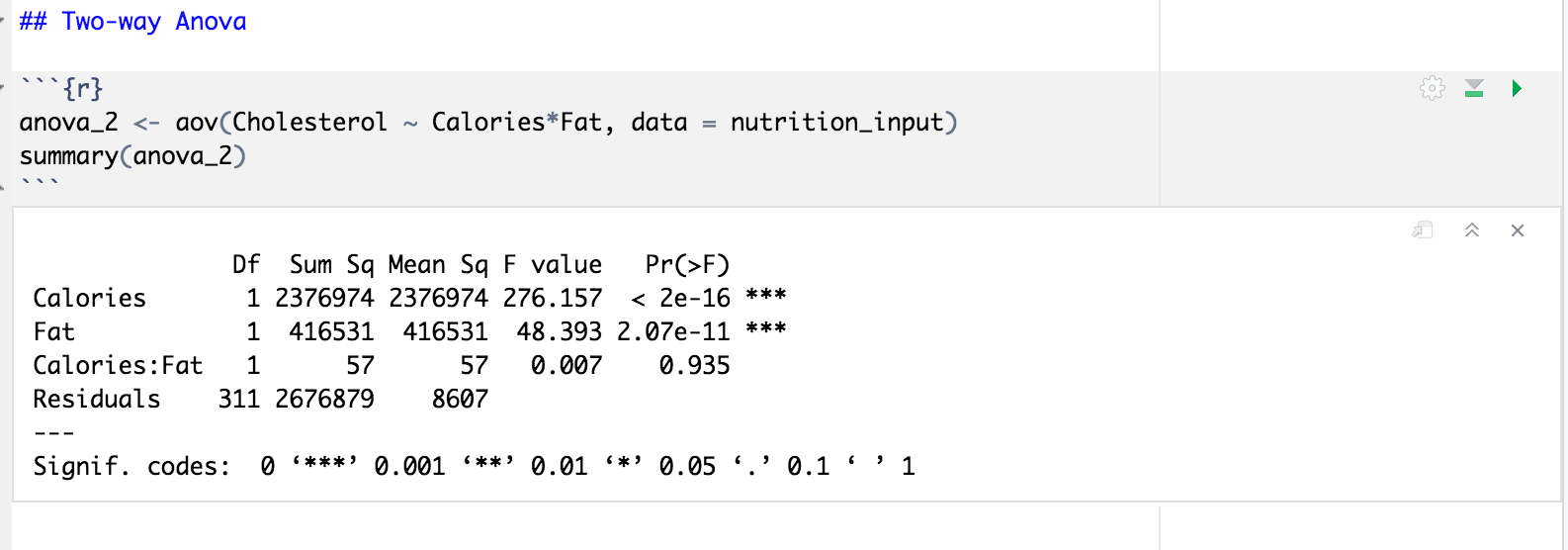
****

We finished our project by running a way one and two way Anova Test to find out if the results of our tests were significant.

****

****

****

****

According to our normality test and our variance test, P Value is less than alpha, thus we reject the null hypothesis. There is not sufficient evidence to claim that there is a significant relationship between cholesterol levels and gender.

**Conclusions**

From our first hypothesis our P-value results were lower than alpha, thus we were **not** able to prove our null hypothesis right. We conclude that there is not a significant relationship between cholesterol and the intake of fat, fiber and calories.

From our second hypothesis our P-values also resulted lower than alpha, thus we were **not** able to prove our null hypothesis right. It can also be concluded that there is not a significant relationship between cholesterol levels within gender.

A limitation of variables observed were the outliers. It would be better to leave out these observations to create a better fit and prediction.

This project put into practice what I have learned throughout the term and helped me to better understand the correlation between variables.

**References**

(2012). Lock5stat.Com. http://www.lock5stat.com/datasets/NutritionStudy.csv