Sackett Building 109
State College, PA 16801
Arbelaez, Geesaman, & Ward
Stat 470W

Chocolate Chip Cookie Diameter Analysis

We R Consulting LLC
What can Statistics do for you?

1. Project Description

The client Michelle Paret, an avid baker, wants to improve her skills to bake the perfect chocolate cookie. This perfect cookie is described by our client as an almost symmetrical cookie with a short diameter and a very crunchy texture. Our client had tried to make chocolate cookies by experimenting with different factors and keeping others constant to have the least possible amount of variation.

The main goal of the study is to find through statistical analysis a way to minimize the diameter of the chocolate cookies with the elements our client used. She had 2 identical baking sheets, same scooper, same bowl, same kitchen, same temperature, and she baked 24 chocolate cookies in the same day, baking each at a time following the same recipe from the bag of chocolate cookies.

This experimental study was randomized, and it was replicated because each test was repeated 3 times. She collected 24 observations, of which 8 were unique tests and they were replicated 3 times each (as mentioned before). Our population of interest is any chocolate cookie she baked in her oven, where our parameter of interest is μ.

1.1 Research Questions

Question1: Do fat, flour or chill time affect average cookie diameter?
Question 2: For any factors (fat, flour, and chill time) that are significant, what are the best settings to minimize cookie diameter?

1.2 Variables

We have 4 different variables. Our independent variables are 3: *ChillTime, Fat, and Flour*.

- *ChillTime*, as the name says it, is the amount of time the dough was left chilling in the fridge before baking it. Even though is numerical, in this case it is a categorical variable because there are two levels: 135 minutes and 360 minutes.
- *Fat* is the type of fat that was used in the experiment. This variable is categorical, and our client used two types of fat: margarine and butter.
- *Flour* is our last independent variable, and it consists of the amount of flour in ounces that our client used for the chocolate cookie. Even though this variable (as *ChillTime*) is naturally numerical, we only have two factors: 1 ounce and 1.75 ounces, therefore is a categorical variable.
- Our response variable is *Diameter,* which ranges from 5.4 centimeters to 10 centimeters. The four variables in the dataset are summarized in Table 1 below.

Variable	Type	Description	Levels and Ranges
ChillTime	Categorical	Number of minutes that the dough was left chilling (minutes)	Two levels: 135 and 360
Fat	Categorical	Type of fat that was used (type of fat)	Two levels: Margarine and Butter
Flour	Categorical	Amount of flour (oz)	Two levels: 1 and 1.75
Diameter	Numerical	Diameter in centimeters of the diameter (cm)	5.4 to 10

Table 1: Summary of variables for the chocolate cookie study

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the crucial process of doing first investigations on data to find patterns, uncover anomalies, test hypotheses, and double-check assumptions using statistical summaries and graphical representations. This section showcases 11 different plots to provide an initial insight into Michelle's data.

Understanding the distribution of the data: The implementation of boxplots and histograms were key to identifying the data's spread and characteristics. A boxplot is a standardized way of displaying the distribution of data based on a five number summary (lowest value, first quartile, median, third quartile, and highest value), where first quartile is the value under which 25% of data points are found when they are arranged in increasing order, the median is the value which 50% of data points are found when they are arranged in the order mentioned above, and the third quartile is all above but the value being 75% of data points. The correct way to read it is reading the minimum value in line with the first line, reading the maximum value in line with the last line, reading the first quartile which is in line with the start of the box, reading the upper quartile which is in line with the end of the box, and finally reading the median which is in line with the line inside the box.

This distribution mentioned can be seen in Figure 1, where we can see the distribution of our *Diameter* variable, which ranges from 5.4 centimeters to 10 centimeters. We can also see from the boxplot shown in Figure 1 that our median (the center of our data), is around 7.15 centimeters. The other key plot that helped us identify the data's spread was the histogram. A histogram is a graph used to show the frequency distribution of data points of one variable. Figure 2 shows the histogram of *Diameter,* where we can realize that the mode (the data point that appeared the most often) was from 5 centimeters to 6 centimeters, with 10 units.

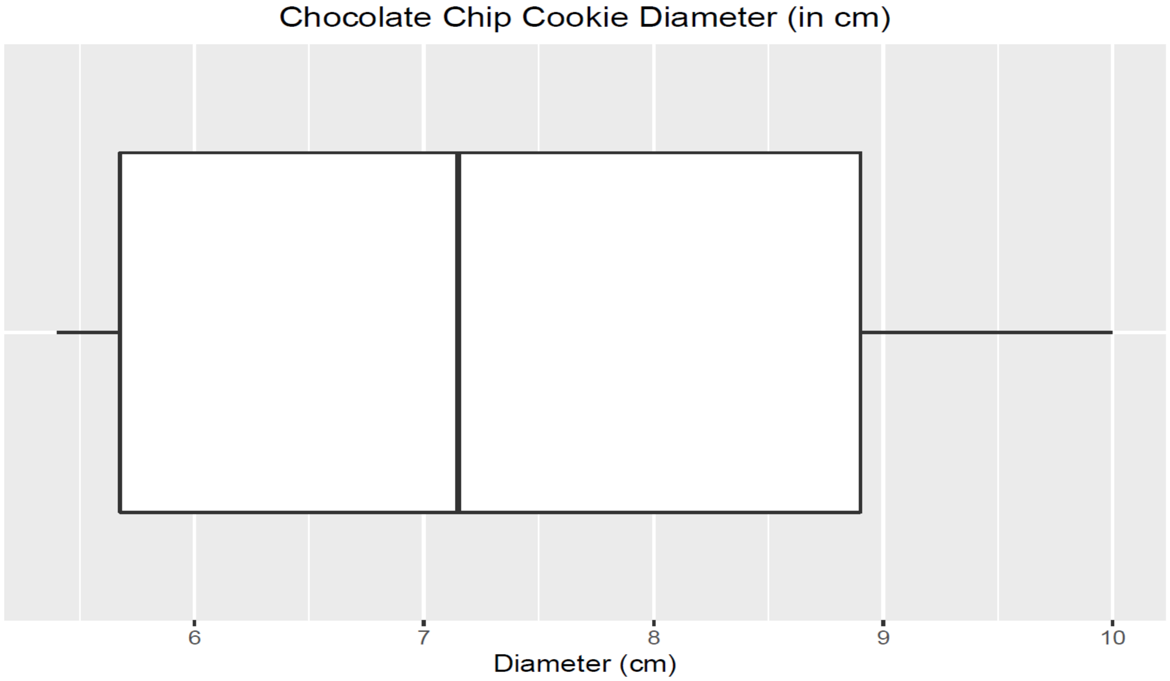


Figure 1: Boxplot of Diameter in centimeters.

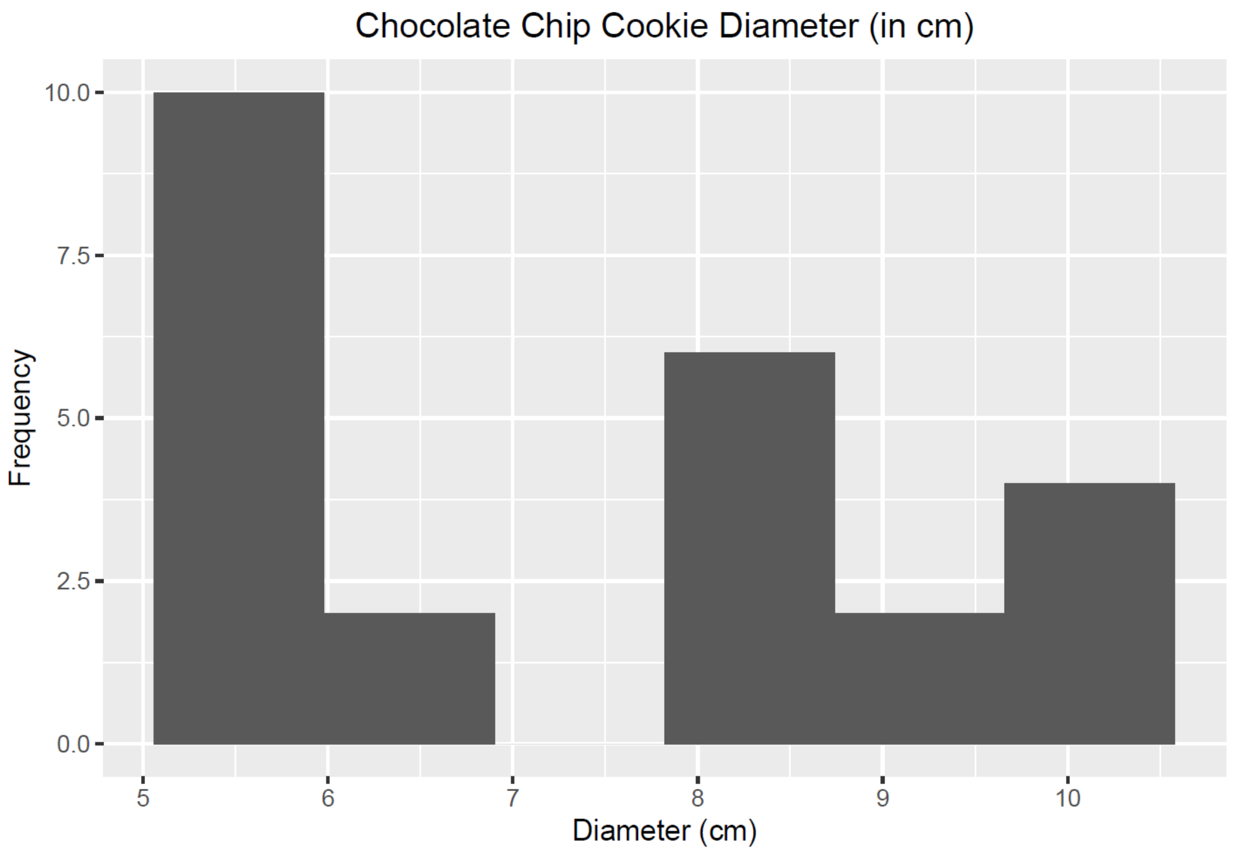
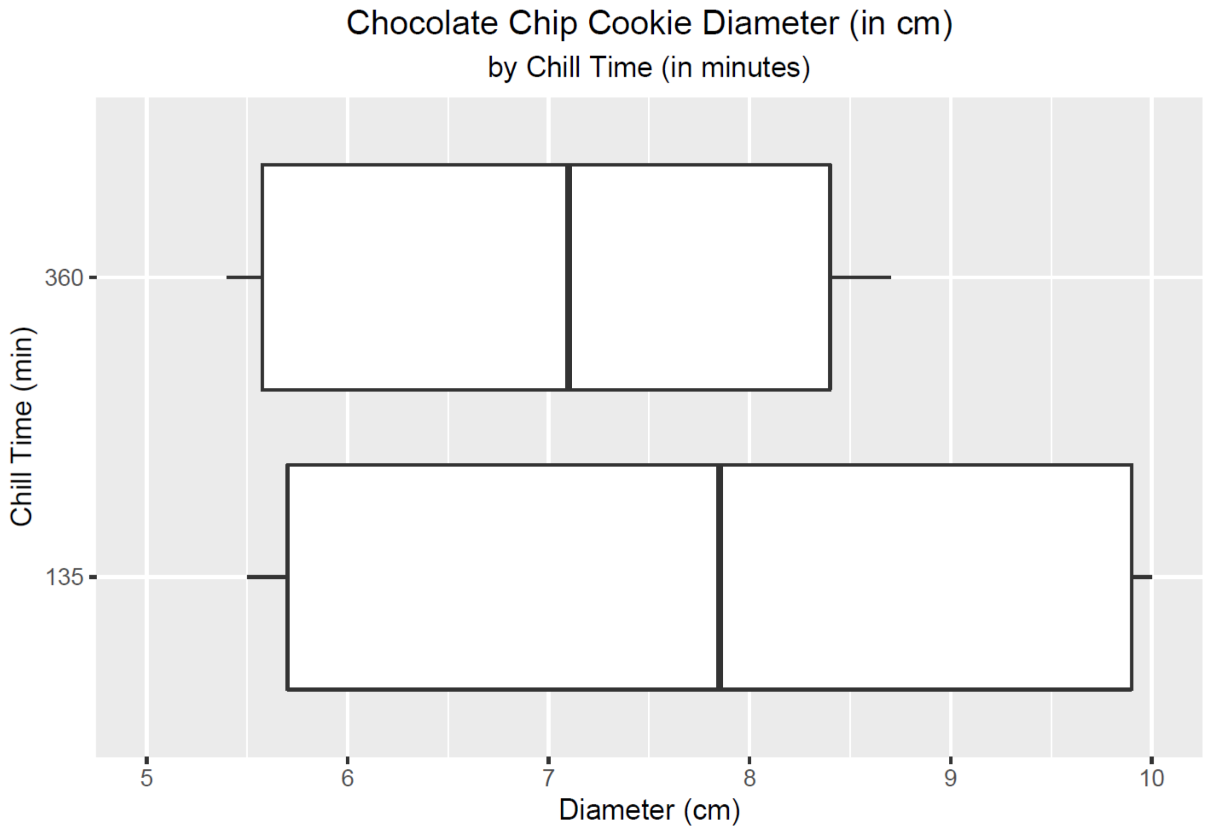


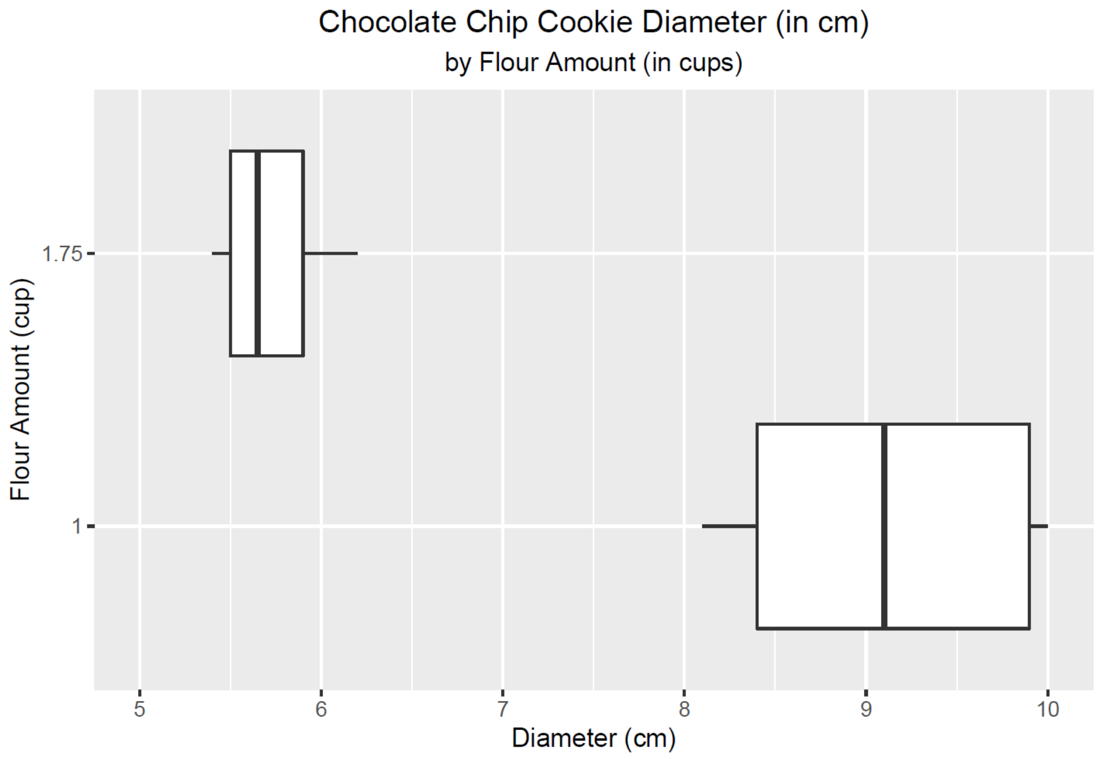
Figure 2: Histogram of Diameter in centimeters.

Once we saw the distribution at a macro level with our main variable of interest, we started to create side-by-side boxplots, which are used to display the distribution of a quantitative variable (in this case our response variable *Diameter*) along with a categorical variable (in this case we had three different categorical variables as mentioned previously in Table 1: *ChillTime, Flour*, and *Fat*). Figure 3, Figure 4, and Figure 5 show the distribution of *Diameter,* along *ChillTime, Flour*, and *Fat* respectively. In Figure 3 we can observe that the chocolate chip cookie diameter's median is slightly lower when chill time is 360 minutes and the range for the diameter of the cookie is bigger when chill time is 135 minutes. In Figure 4 we can note that the chocolate chip cookie diameter's median is lower when flour amount is 1.75 cups. The range for the diameter of the cookie is minimal for 1.75 cups of flower, which also indicates that the variation is less, and the value for diameter is more consistent. In Figure 5 we can see that the chocolate chip cookie diameter's median is roughly similar but slightly lower when the type of fat used is margarine and the range for the diameter of the cookie is roughly the same when using both types of fat. It is important to say that we **cannot conclude anything from purely observing these boxplots** because we need to do the respective hypotheses tests to verify their respective significance.



Chocolate Chip Cookie Diameter (in cm)
by Chill Time (in minutes)

Figure 3: Boxplots of Diameter in centimeters by Chill Time.



Chocolate Chip Cookie Diameter (in cm)
by Flour Amount (in cups)

Figure 4: Boxplots of Diameter in centimeters by Flour Amount.

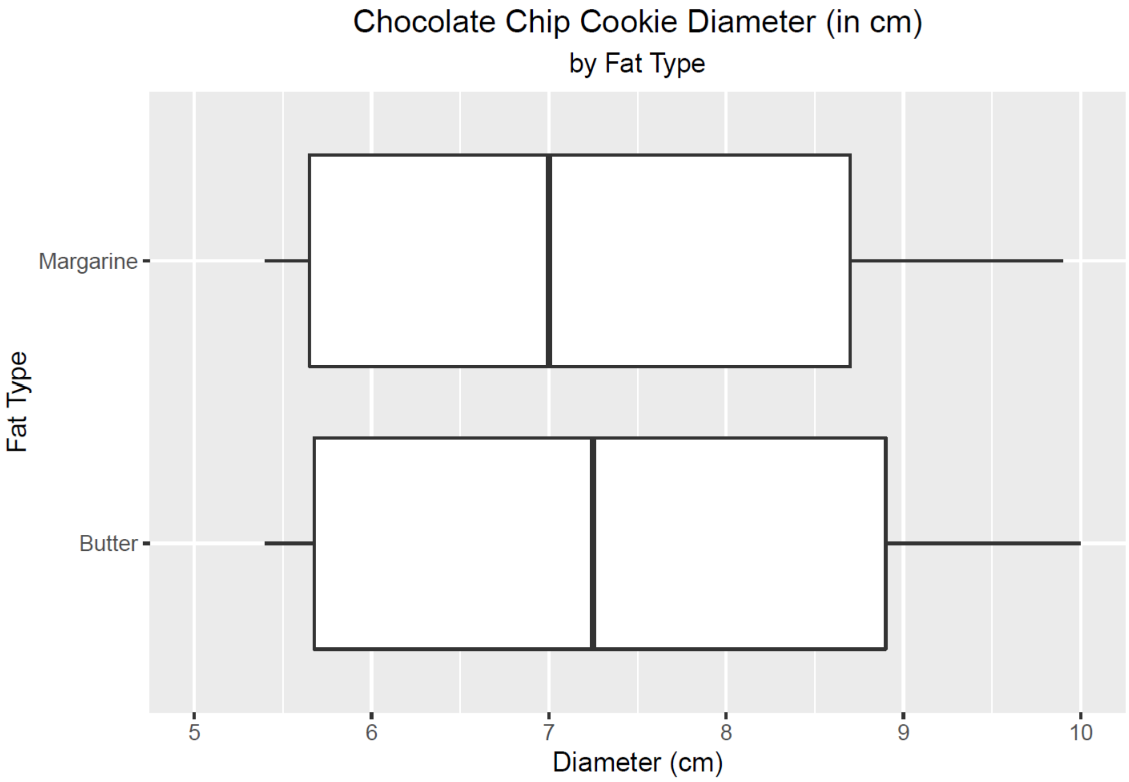


Figure 5: Boxplots of Diameter in centimeters by Fat Type.

Figures 6 and 7 both display graphs in which the predictive variables (fat type, flour amount, and chill time) are analyzed. These graphs help to initially estimate whether any two variables affect one another. Each "setting" within one predictive variable is plotted against another variable's settings. For figure 6, if these settings seem to be dependent on one another between variables, the plots will show lines with differing slopes. Otherwise, the lines will have approximately similar slopes. For figure 7, if these settings seem to be dependent on one another, the boxplots will appear more separated with spread medians. On the contrary, if the settings do not appear to influence one another, the boxplots will look similar to each other. For example, the two graphs in the top row of figure 6 appear to have no effect on one another as the lines' slopes are about the same. However, the bottom graph of figure 6 could show evidence of the flour amount and chill time variables affecting one another. Although these graphs are helpful for seeing the first picture of the data, they are not conclusive and cannot supply definite answers to these questions. Only after a statistical analysis is run can a conclusive, definite answer emerges as to whether predictive variables affect one another or not.

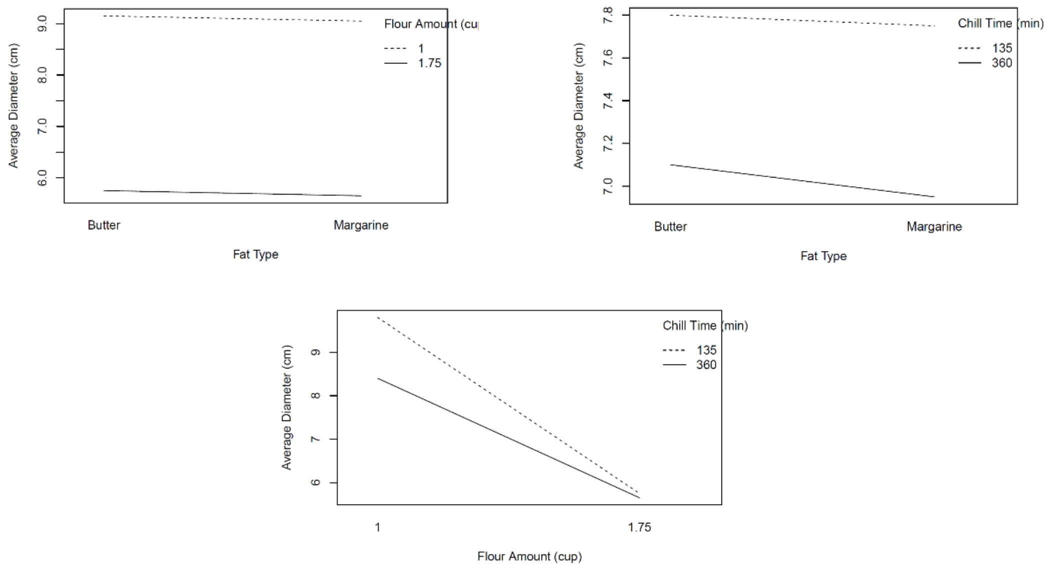


Figure 6. Interaction plots display the possible relationship two predictive variables have on one another.

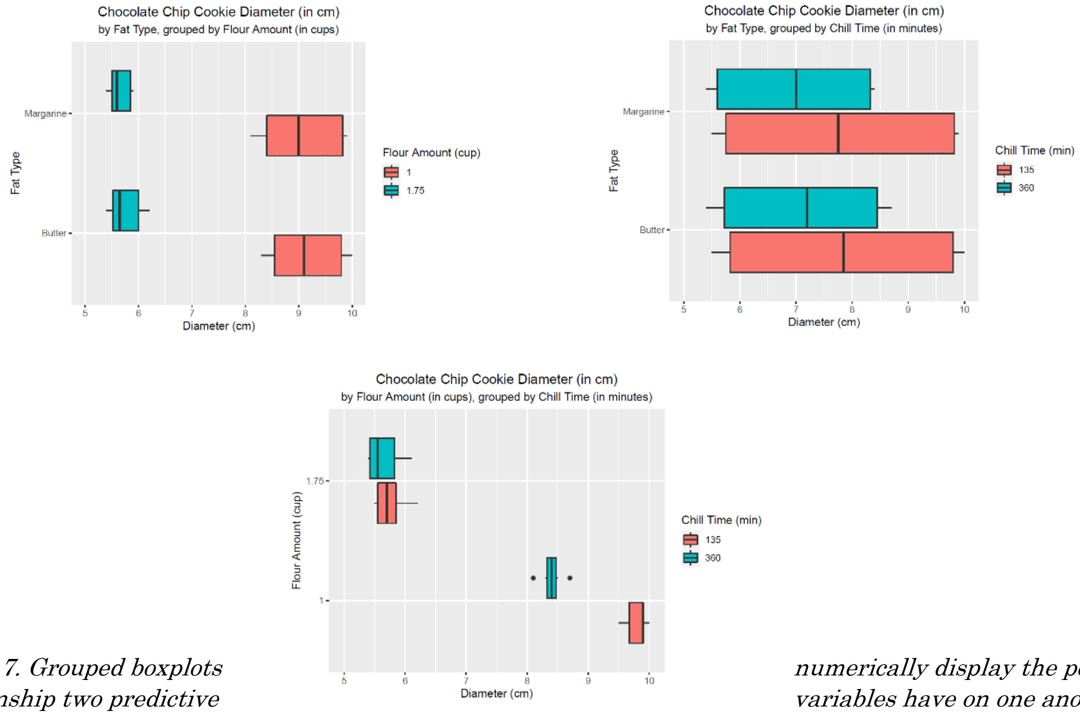


Figure 7. Grouped boxplots numerically display the possible relationship two predictive variables have on one another and how individual "settings" affect the response variable.