**Forecasting Walmart Sales**



STAT 471

Final Project

Fall 2022

Daniel Garcia

Nicole Pham

**Introduction**

This dataset came from Kaggle.com which includes 6,435 observations following the popular general merchandise store Walmart. The data was collected from February 5, 2010 to October 26, 2012. Variables include the store number, date of the sales week, weekly sales, holiday, temperature, fuel price, consumer price index(CPI), and unemployment rate. Holidays only included the Super Bowl, Labor Day, Thanksgiving, and Christmas. During our analysis, we found that extremely high sales were made from the weeks of Thanksgiving to Christmas. Since Walmart has such a large chain, we wanted to know what factors contributed to weekly sales the most.

**Questions of Interest**

1. Is there a model that fits the data in such a way that can accurately predict weekly sales for given conditions?
2. Are there any outlier weeks with unusually high or low weekly sales? If so, is there a trend in the timing of these weeks?
3. If CPI got higher, would that cause customers to spend less?
4. Are weekly sales influenced by an increase in unemployment rate and temperature?
5. What range can we expect average weekly sales to fall in?

**Analysis**

1. **Is there a model that fits the data in such a way that can accurately predict weekly sales for given conditions?**
2. Important details of the analysis:

The “Walmart” data set includes 6,435 observations with 8 variables: store number, date, holiday indicator, temperature(°F), fuel price($), consumer price index(CPI), unemployment rate(%), and weekly sales($). After visualizing the distribution of weekly sales with a histogram, we noticed there was a right skew in the distribution. This meant lower weekly sales were more common while higher weekly sales were rather uncommon. To confirm weekly sales was not normally distributed, we performed the Shapiro-Wilk normality test which showed a significant difference in the distribution.

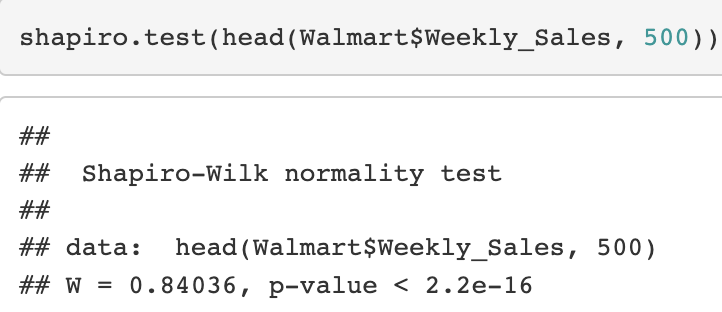
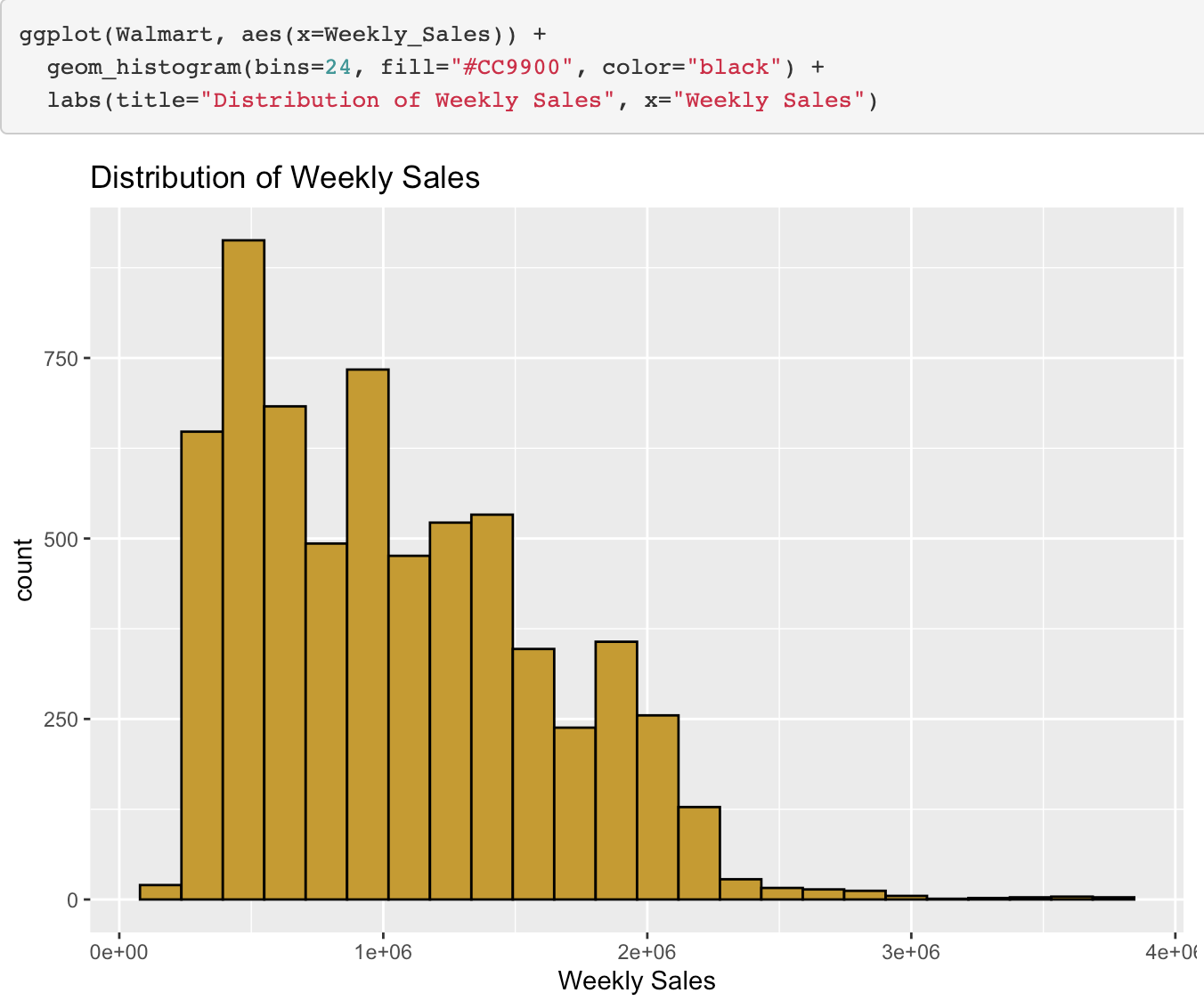
Since there was a right skew to account for, we thought a Gamma regression model would be the most appropriate. First, the variable “Store” and “Holiday\_Flag” had to be releveled as a factor using the relevel() and factor() functions to set a reference level which were saved as “Store\_” and “Holiday\_Flag\_” respectively. Using the new releveled variables, we fit a Gamma regression model using “Weekly\_Sales” as the response variable and “Temperature”, “CPI”, “Unemployment”, and the releveled variables as the predictor variables. Only the variable “Store\_” was not a significant predictor of “Weekly\_Sales.”

To check if the Gamma regression model was a good fit, we fit a null model to conduct a hypothesis test. The null hypothesis states that the null model has a better fit while the alternative hypothesis states that the fitted Gamma regression model has a better fit. After finding the deviance and p-value of the deviance test, we reject the null hypothesis and conclude that the Gamma regression model has a better fit.

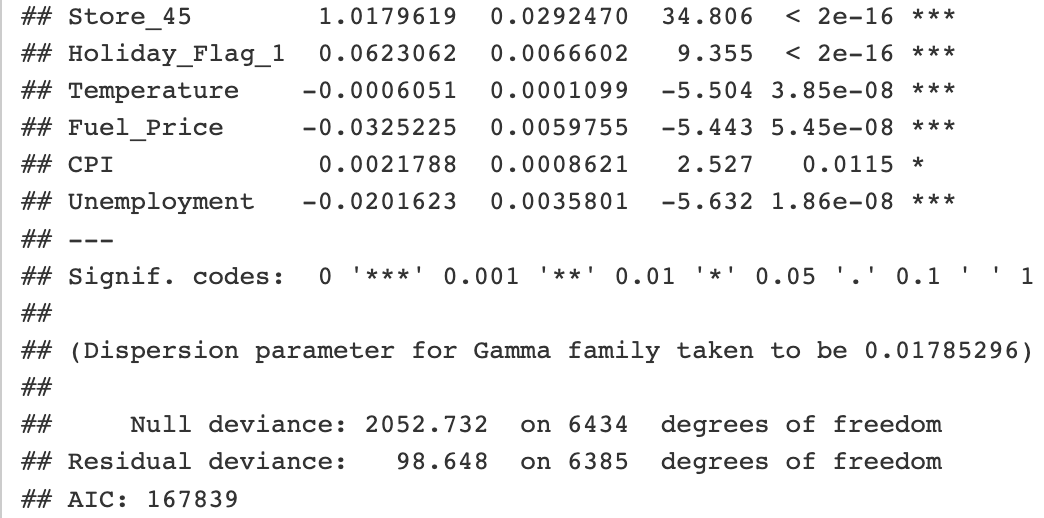
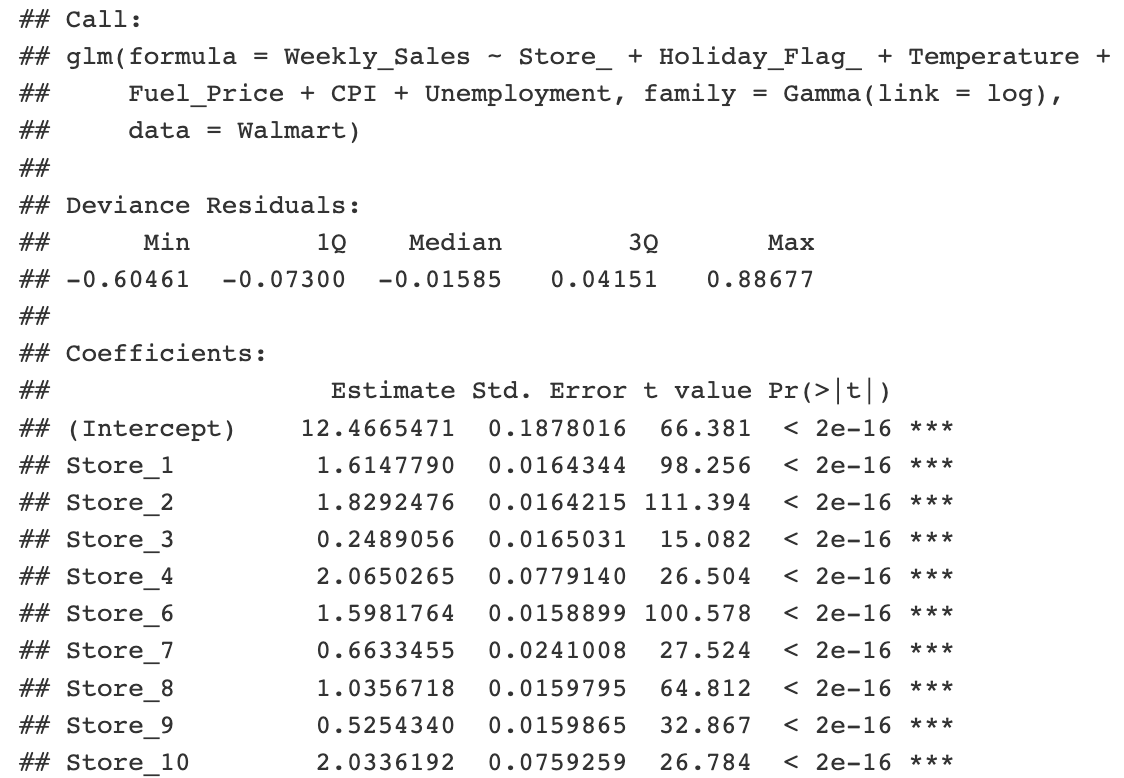
Putting the model to use, we generated a random sample to test the model. The first prediction of weekly sales was $2,114,918 while the actual weekly sales was $1,607,343. The second prediction of weekly sales was $1,495,565 while the actual weekly sales was $1,135,578.

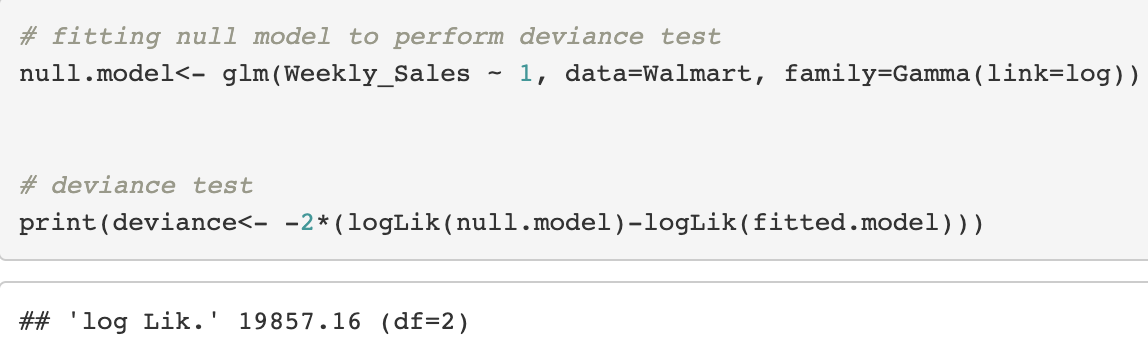
1. Coding:

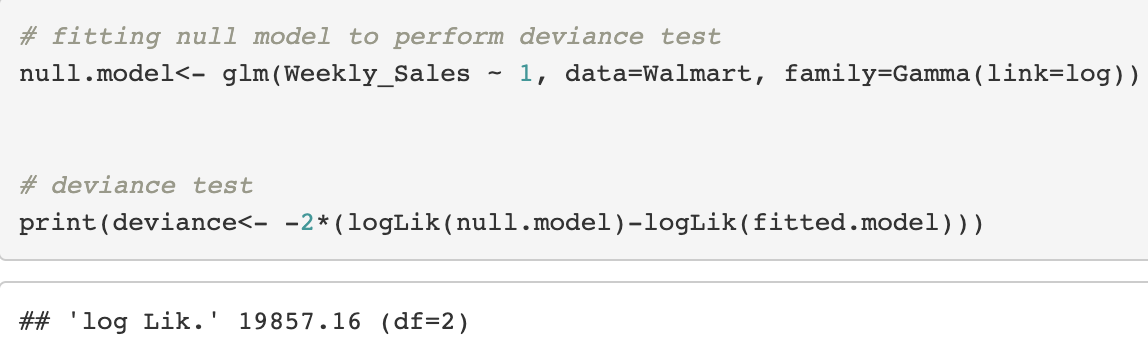
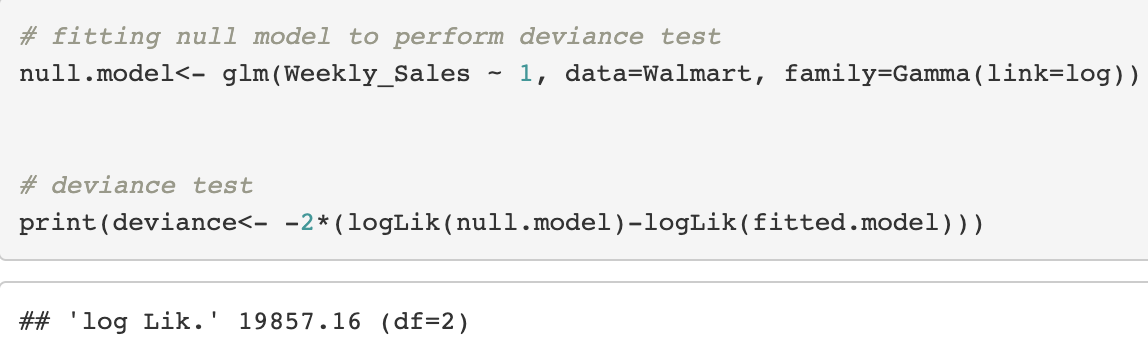
*Code Figure 1*

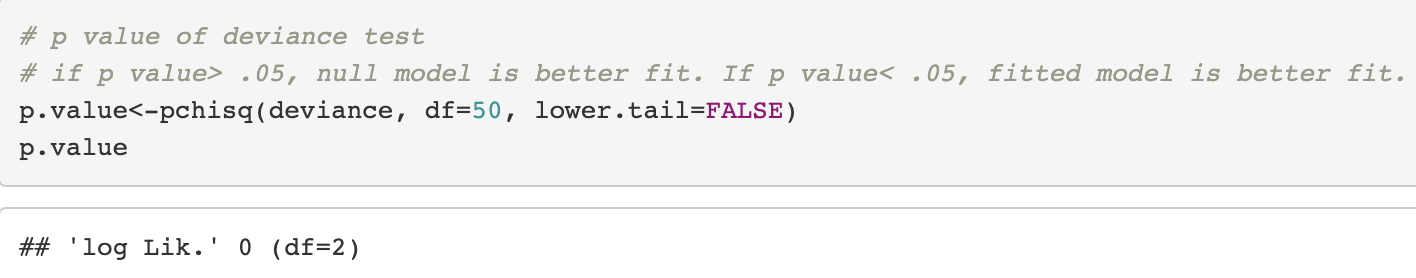
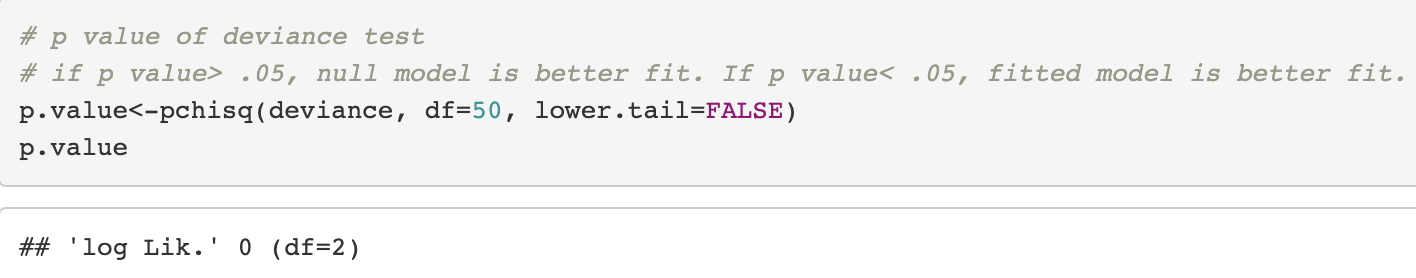


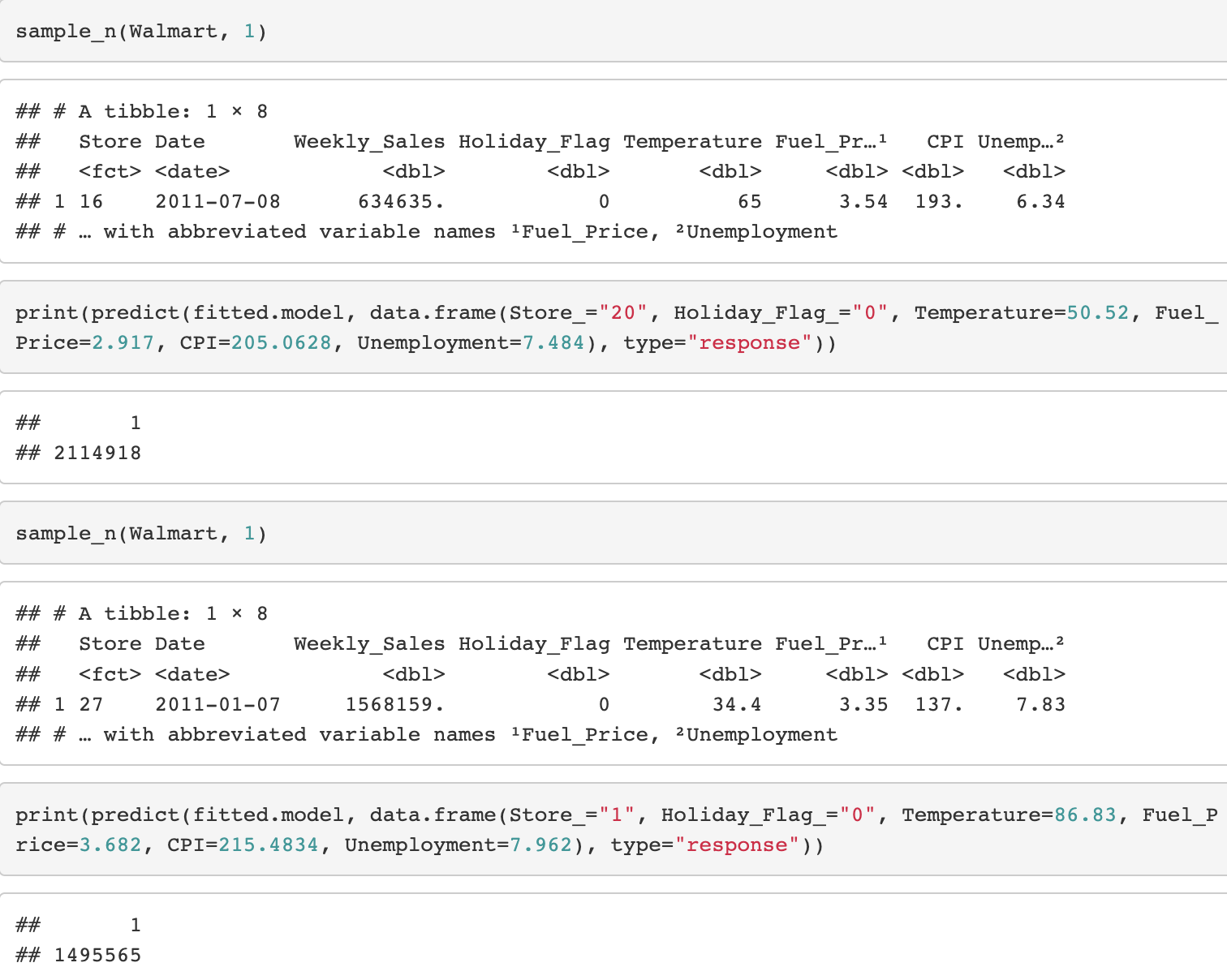
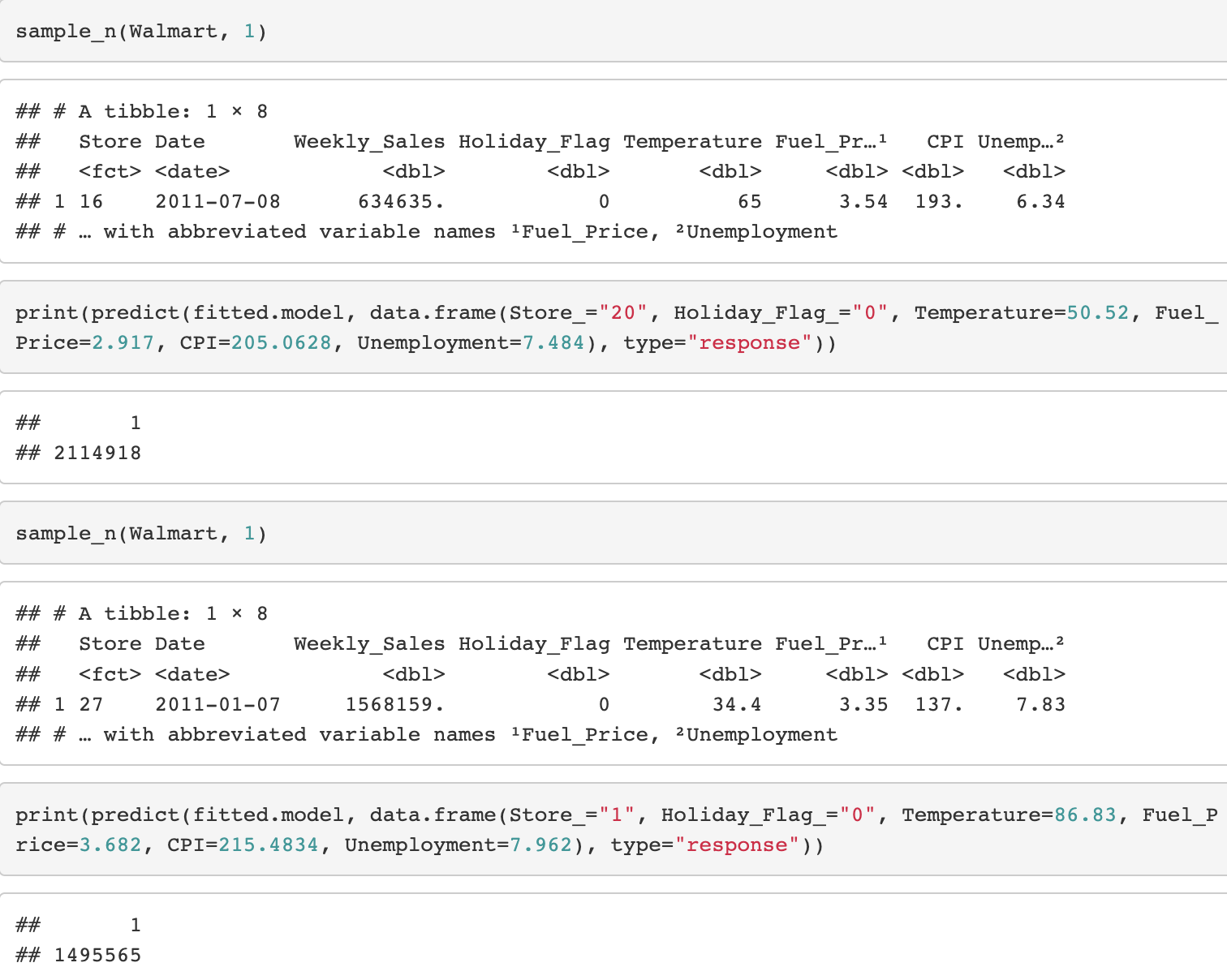
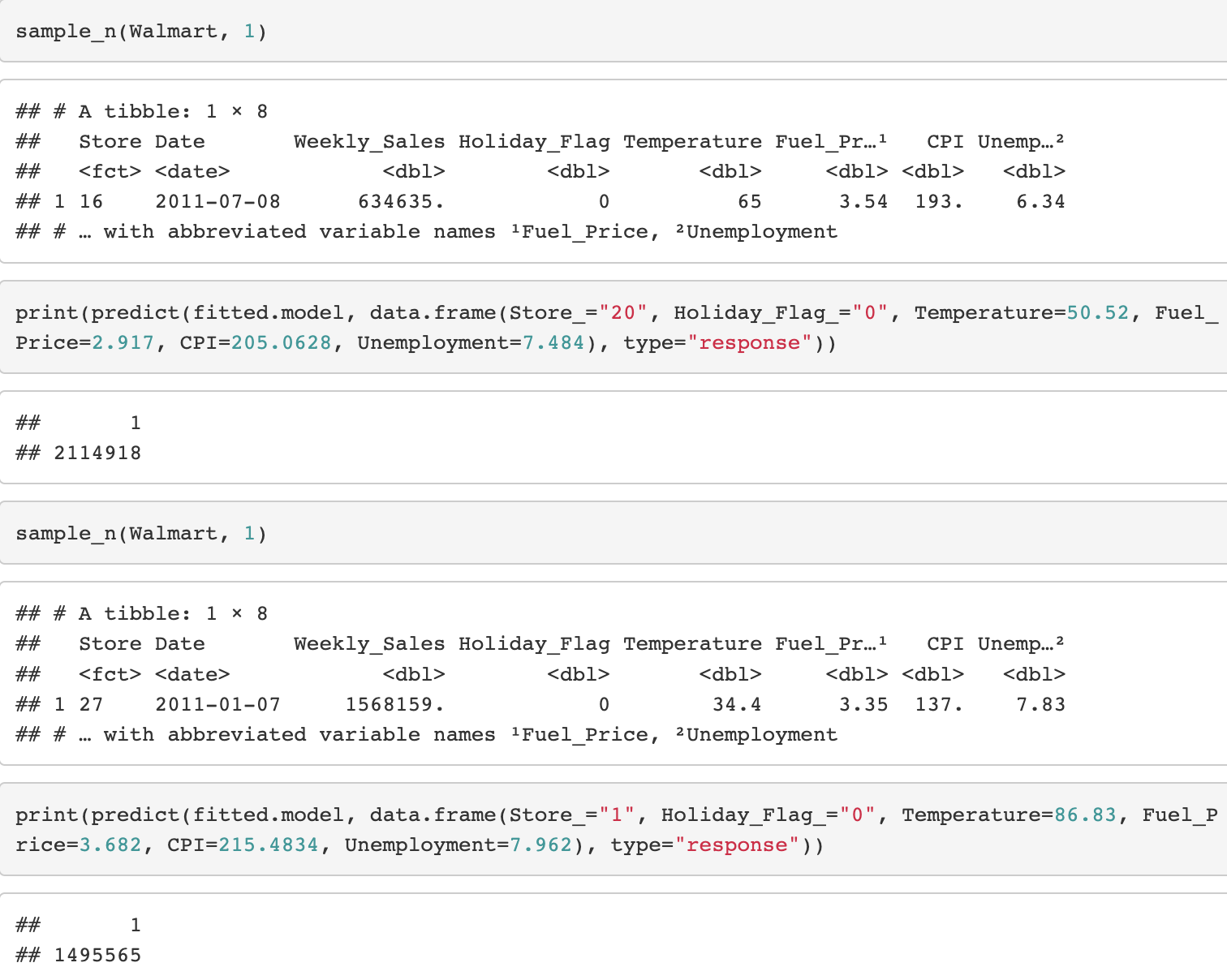
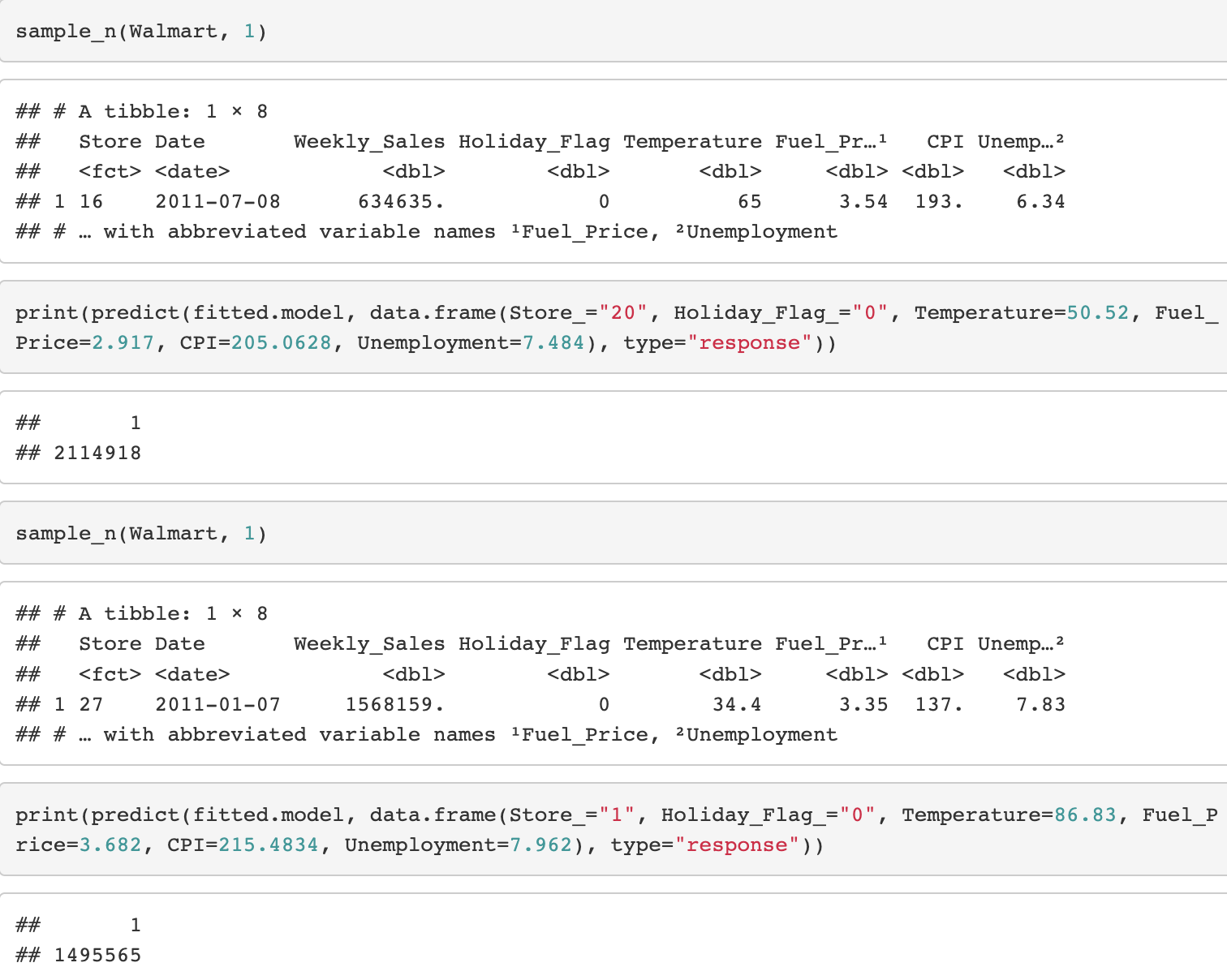
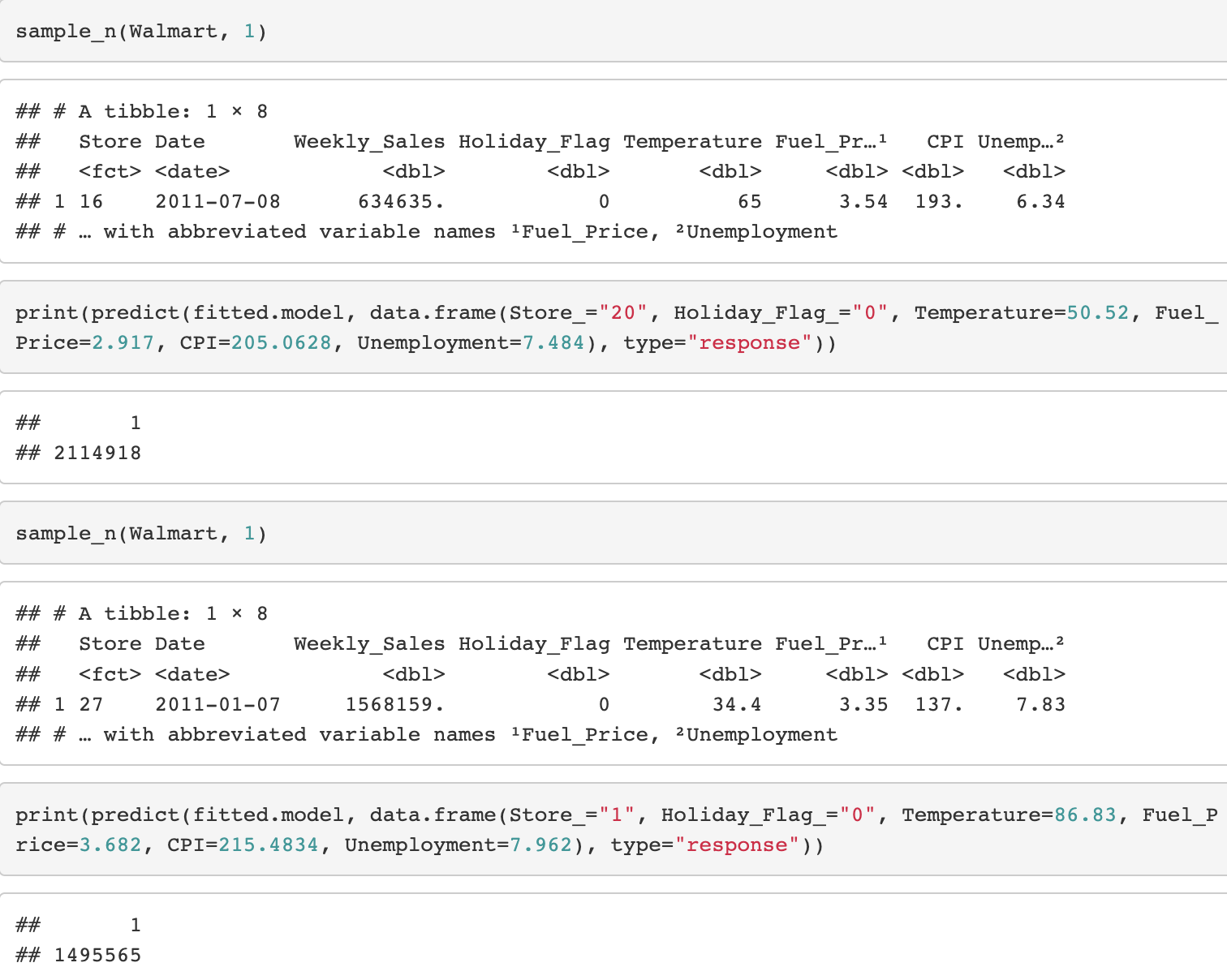
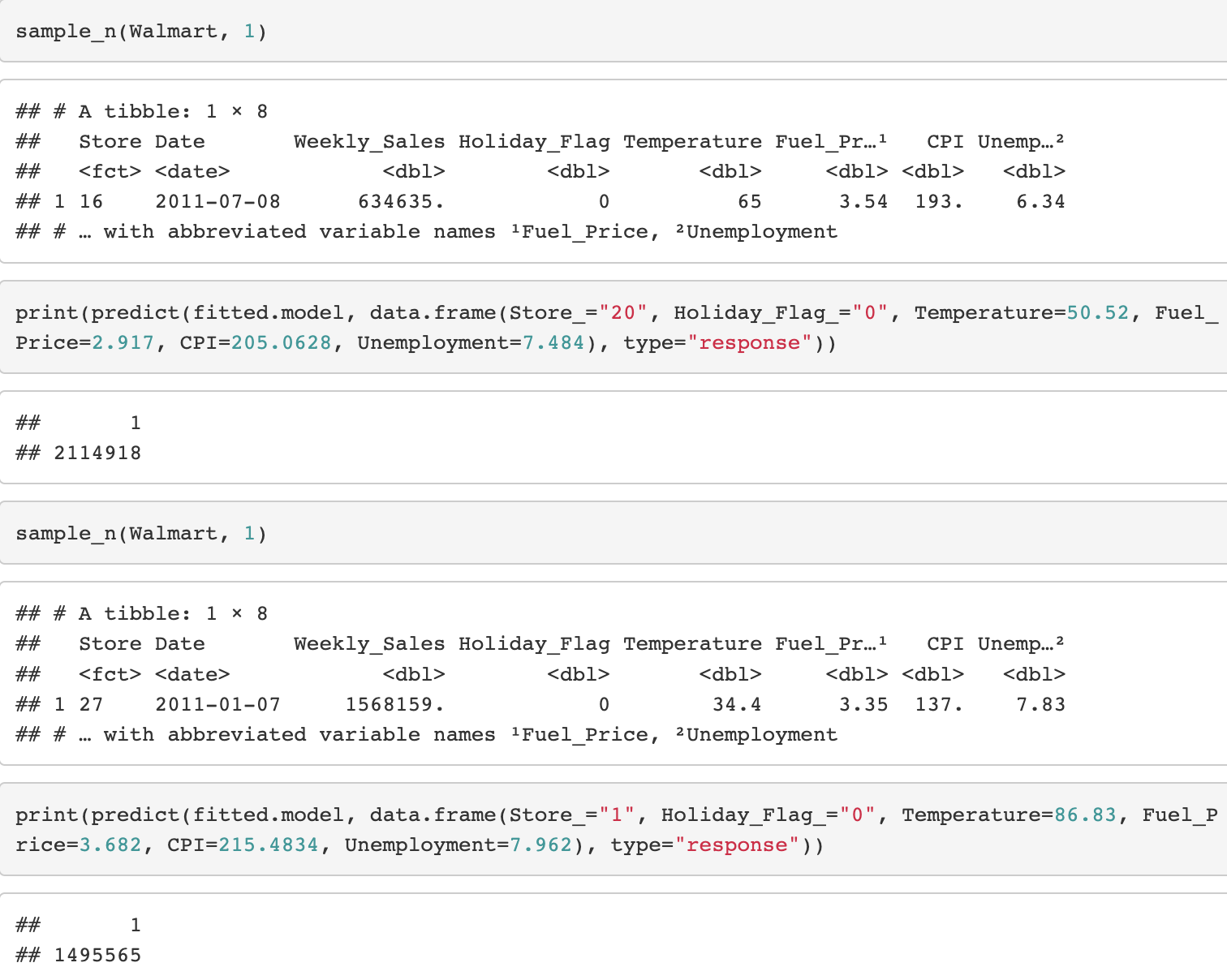
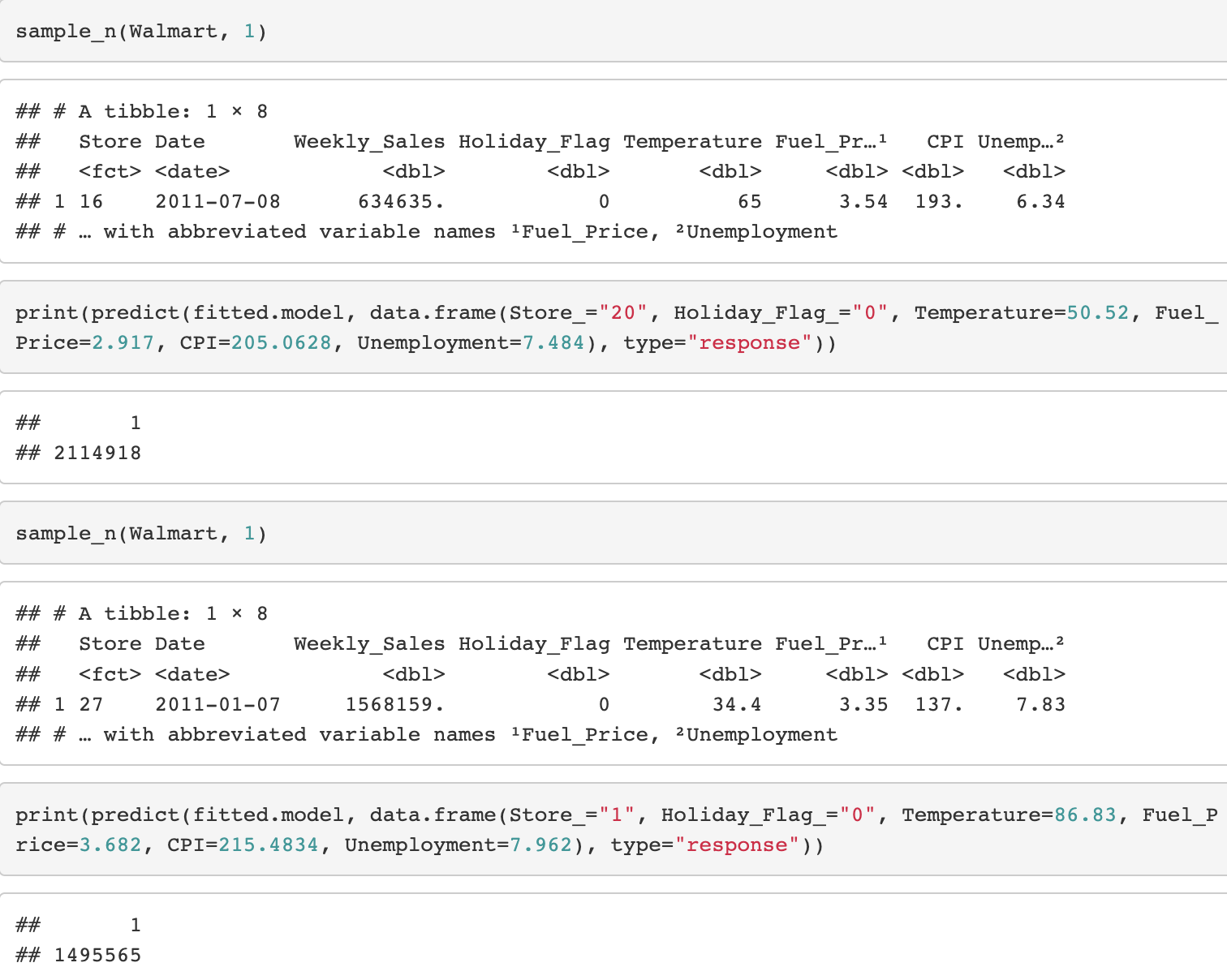
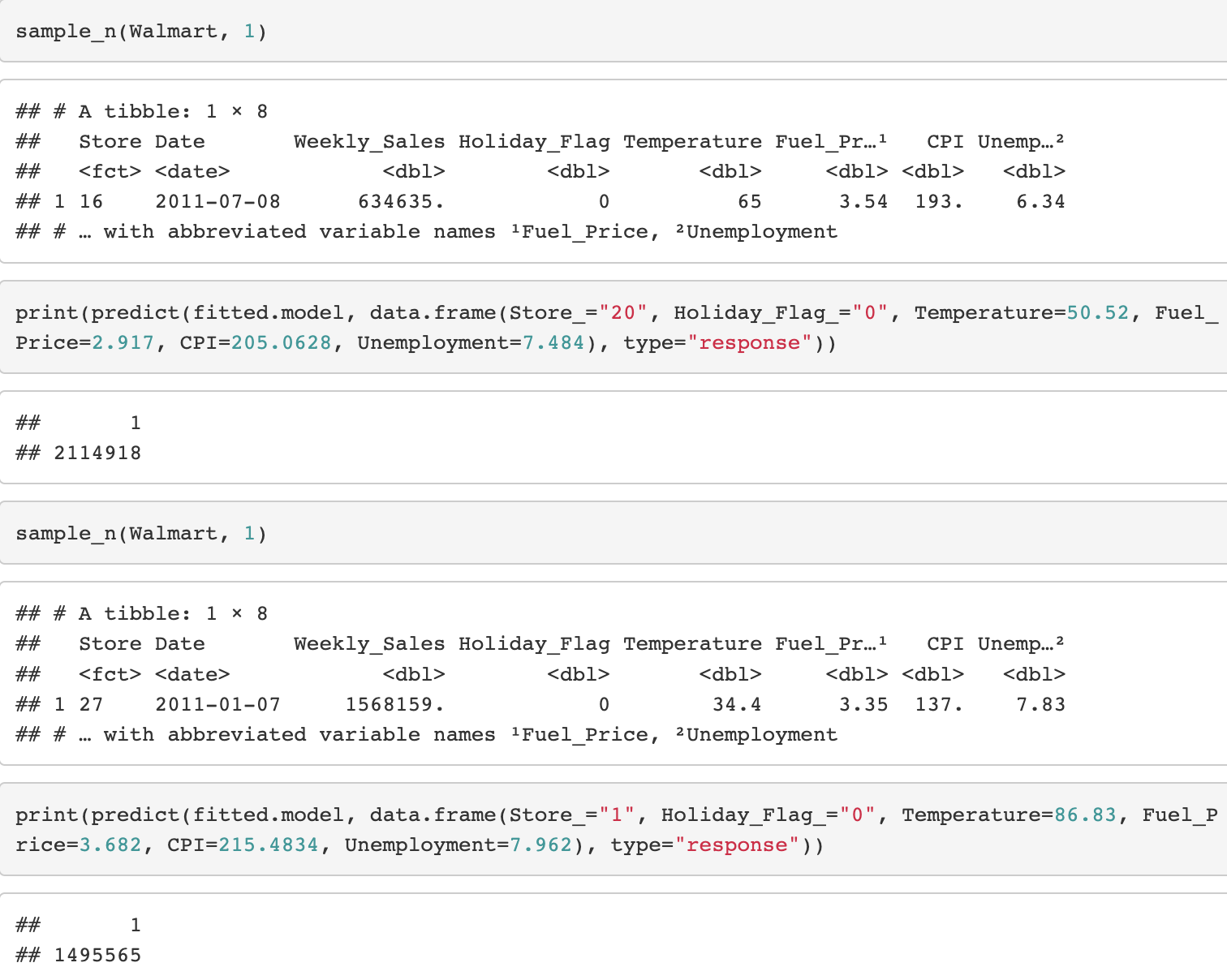
*Code Figure 2*











1. Interpretation:

The Gamma regression model was a good fit for the “Walmart” data set since it accounted for the right skew of the distribution of weekly sales. After fitting the model, we put it to use by doing some predictions. In the first prediction, the fitted model estimated weekly sales by over 30% compared to the actual weekly sales. The second prediction also estimated weekly sales by over 30%. While the model was a good fit, perhaps there were too many high valued outliers that affected the predictions making the model somewhat unreliable.

1. **Are there any outlier weeks with unusually high or low weekly sales? If so, is there a trend in the timing of these weeks?**
2. Important details of the analysis:

This question arose from visualizing boxplots of the average weekly sales and seeing the many outlier values within the boxplot. The next questions naturally were: When do these outliers occur? And are they possibly during holiday times?

To further analyze this question, we first needed to group our data by common dates and holiday flags. To accomplish this, we formatted the original dataset “Walmart” to only include month and day—which signifies the start of each sales week. We then passed the group\_by() function in the tidyverse package into our newly formatted dataset and took the mean of each week. This gave us a tibble of 143 observations (representing the 143 weeks of our data set) and 3 columns (date, holiday\_flag, and weekly sales). We called this transformed data set “Mean\_wkly\_sales”.

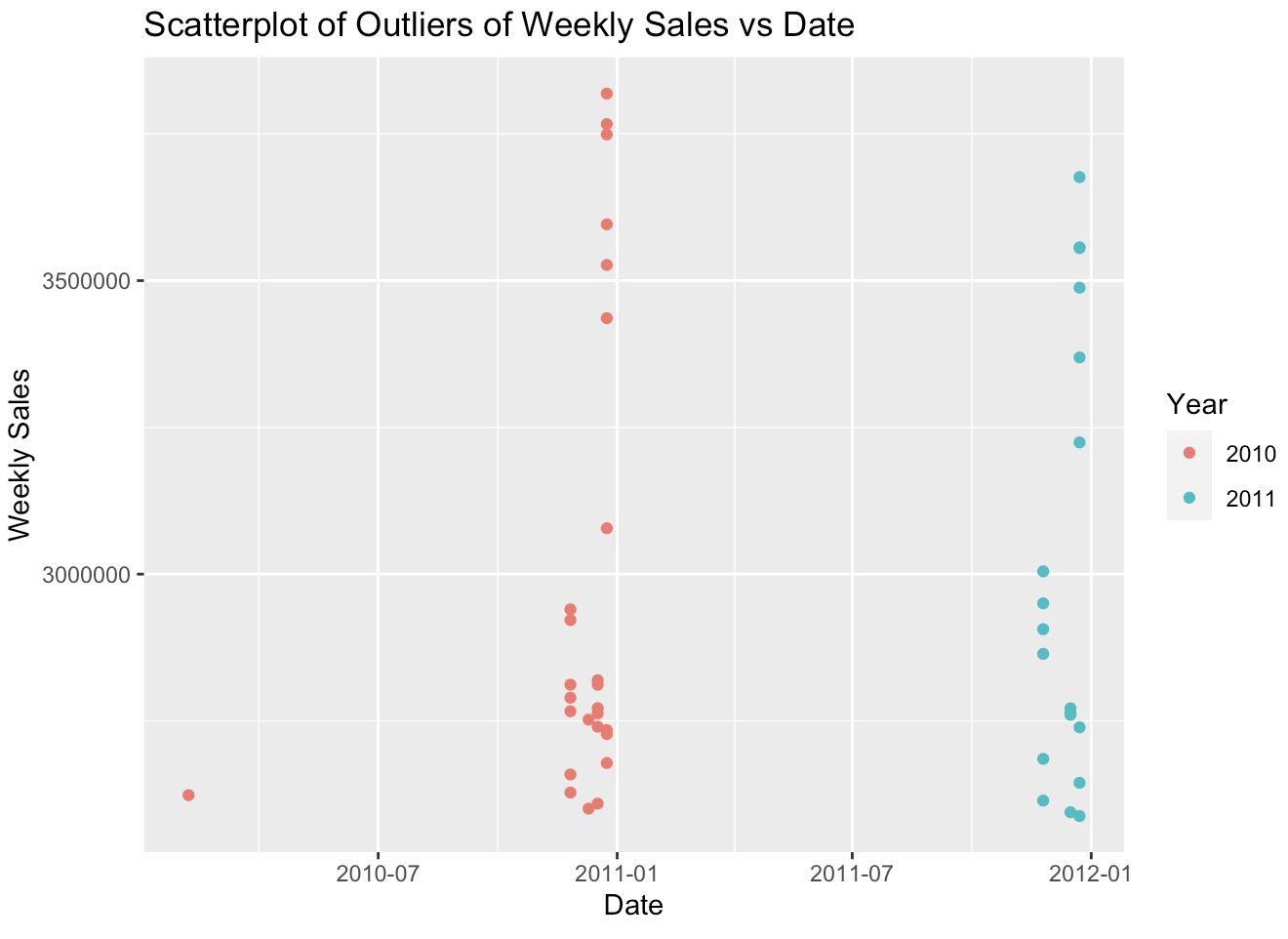
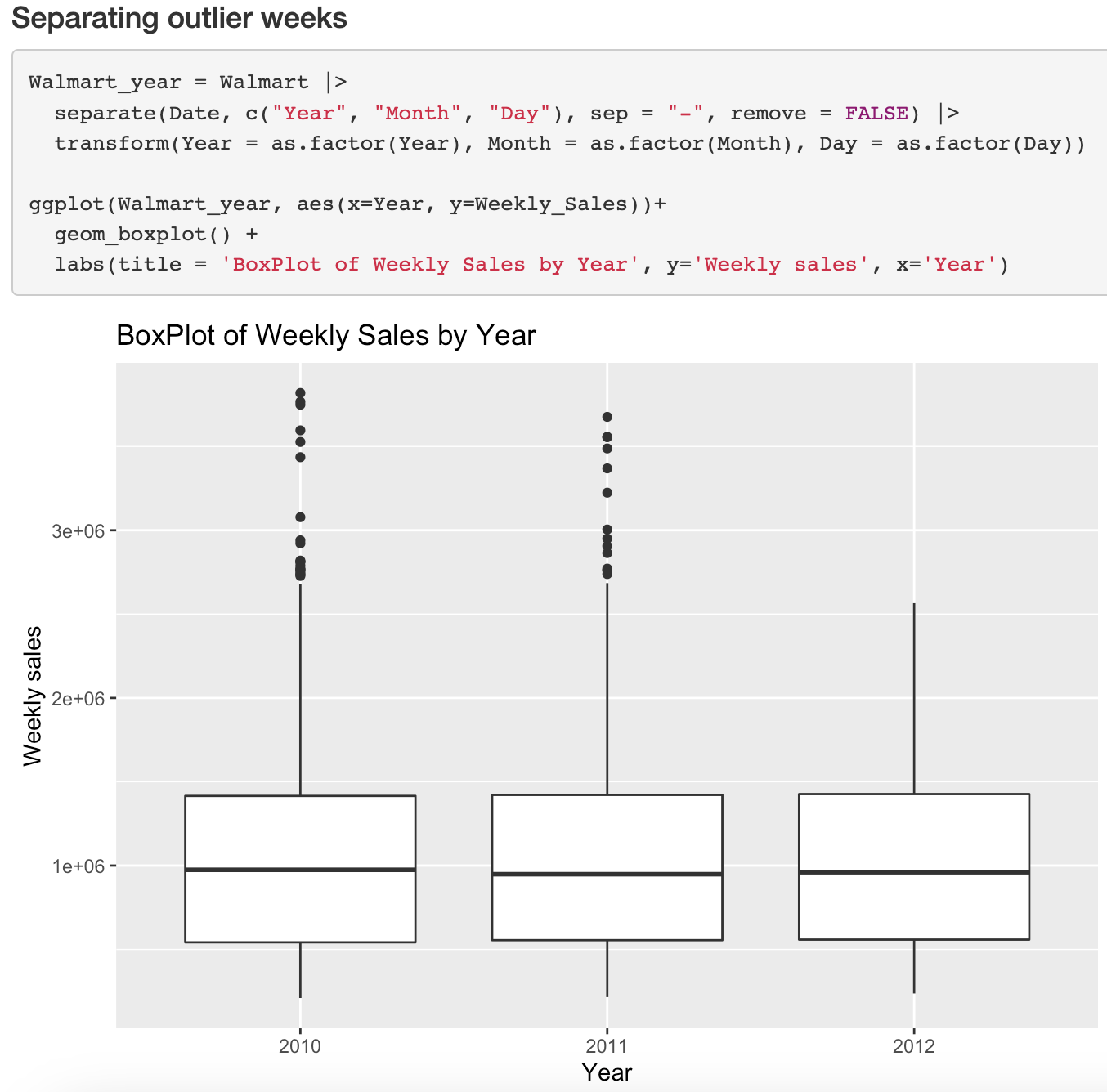
Next step was to visualize the relationship between the average weekly sales and time in a time series plot. Understanding the univariate statistics from our earlier explorations of average weekly sales, a reasonable outlier can be defined as those weeks that surpass $1,200,000. So when we executed the time series plot, we highlighted all those observations that exceeded $1,200,000 in average weekly sales. In doing so, we found all outlier observations to come in clusters, specifically, late 2010 and late 2011.

To get a more definitive answer as to when these outliers occurred. We passed the filter() function into the “Mean\_wkly\_sales” dataset to include only those observations that exceeded $1,200,000 in average weekly sales. The results show that all outlier observations occur between thanksgiving and Christmas.

To better represent our findings, we modified the original dataset “Walmart” to consider all weeks within thanksgiving to Christmas as holiday weeks. The methods we used were to filter those weeks from our original dataset, mutate the “Holiday\_Flag” to represent 1 in those weeks, and set it as a variable equal to “x”. We then deleted the old observations that still valued those weeks as 0 in our original “Walmart” dataset and called the variable “Walmart\_new”. Lastly, we used the function “rbind()” to merge “x” and “Walmart\_new”. We then performed the same operations as before to get the average weekly sales with the modified holiday weeks. When we graphed “Walmart\_new” as a time series graph and set the points to represent holiday weeks, we saw that all outliers were now during holiday times.

1. Coding:

*Code Figure 3 Code Figure 4*



*Code Figure 5: Mean\_wkly\_sales Code Figure 6*

Graphical user interface, text

Description automatically generatedChart

Description automatically generated

*Code Figure 7: all\_extremes Code Figure 8: x*

Table

Description automatically generated with medium confidenceGraphical user interface

Description automatically generated with medium confidence

*Code Figure 9*

Chart, scatter chart

Description automatically generated

1. Interpretation:

The timing of peak sales is drastically affected by holiday times. When we redefined which weeks are considered holiday weeks, we saw from the code and from the visualization that all outlier weeks lie within thanksgiving to Christmas. Further analysis of the visualization suggests that the week succeeding thanksgiving shows a sharp decline in weekly sales, followed by successive increases in sales until Christmas.

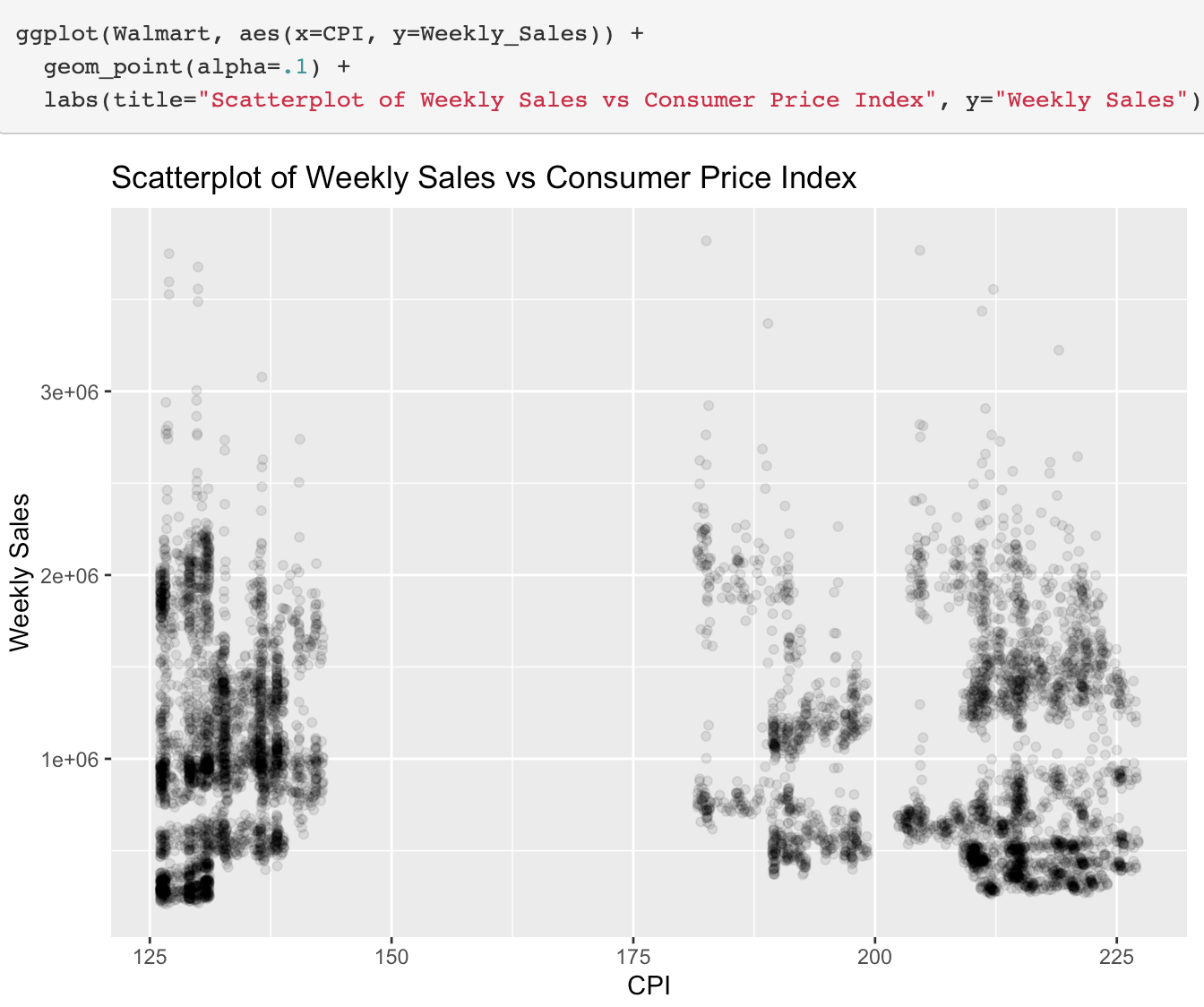
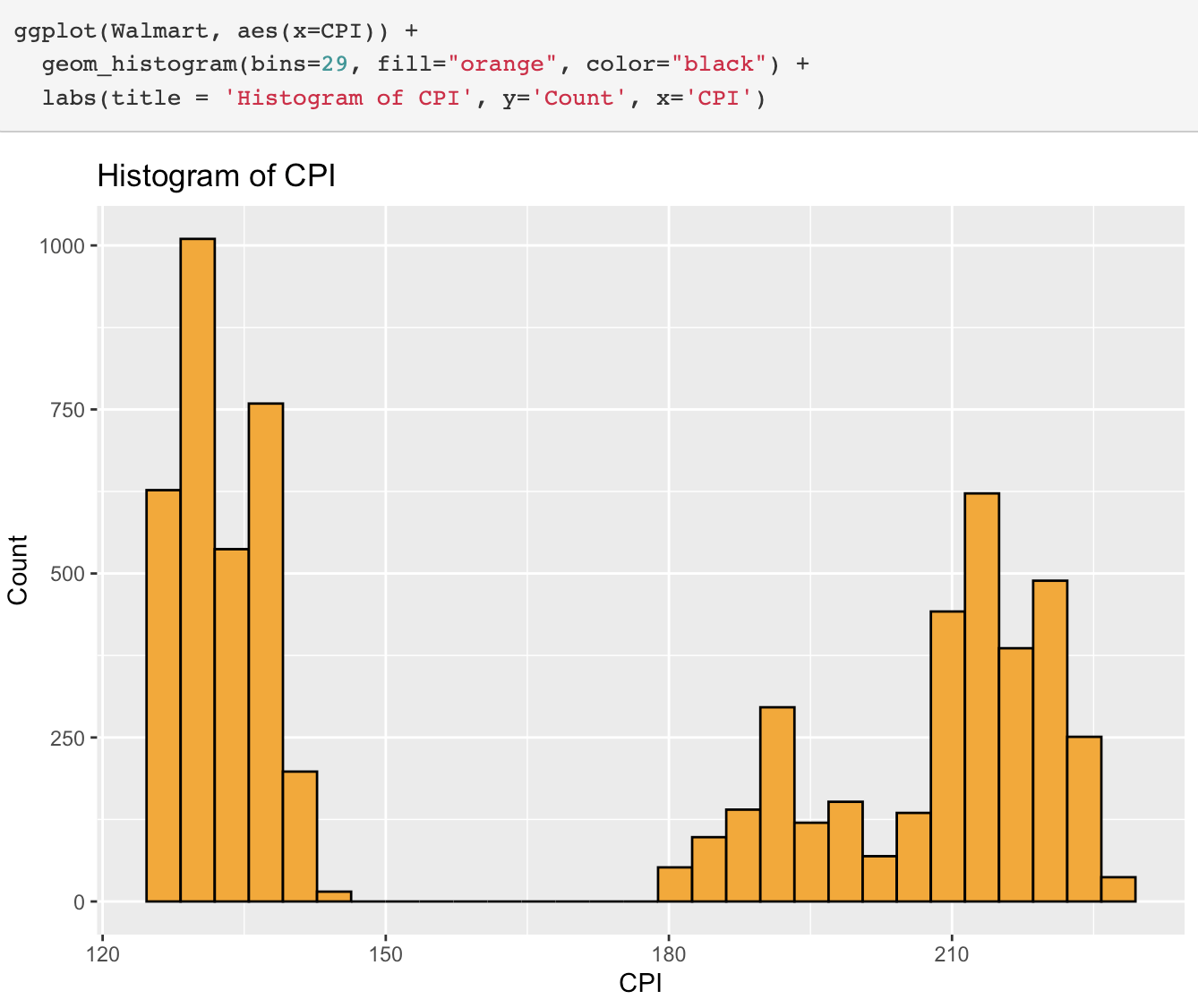
1. **If CPI got higher, would that cause customers to spend less?**
2. Important details of the analysis:

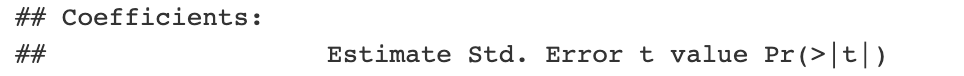
Looking at the distribution of CPI, the histogram shows most of the data was in the intervals of 120-150 or 180-225 CPI. There was not an obvious pattern so we plotted weekly sales by CPI. Looking at the scatterplot, CPI does not seem to affect weekly sales as it has a weak association. Weekly sales also look equally distributed across the CPI intervals and do not show an underlying distribution.

However, after fitting the “Walmart” data set into a Gamma regression model, we have found that CPI was indeed a significant predictor of weekly sales as evidenced by a p-value less than 0.05. The fitted model suggested that as CPI increased by one unit, the estimated weekly sales increased by 0.22%.

1. Coding:

*Code Figure 10 Code Figure 11*







1. Interpretation:

Doing an initial exploratory data analysis on CPI and Weekly Sales by CPI did not show anything meaningful. However, the Gamma regression model showed that CPI was significant in predicting weekly sales by 0.22%. Even though CPI increased, customers were spending more instead of less as seen by the increase in weekly sales. A higher consumer price index does not mean customers will spend less.

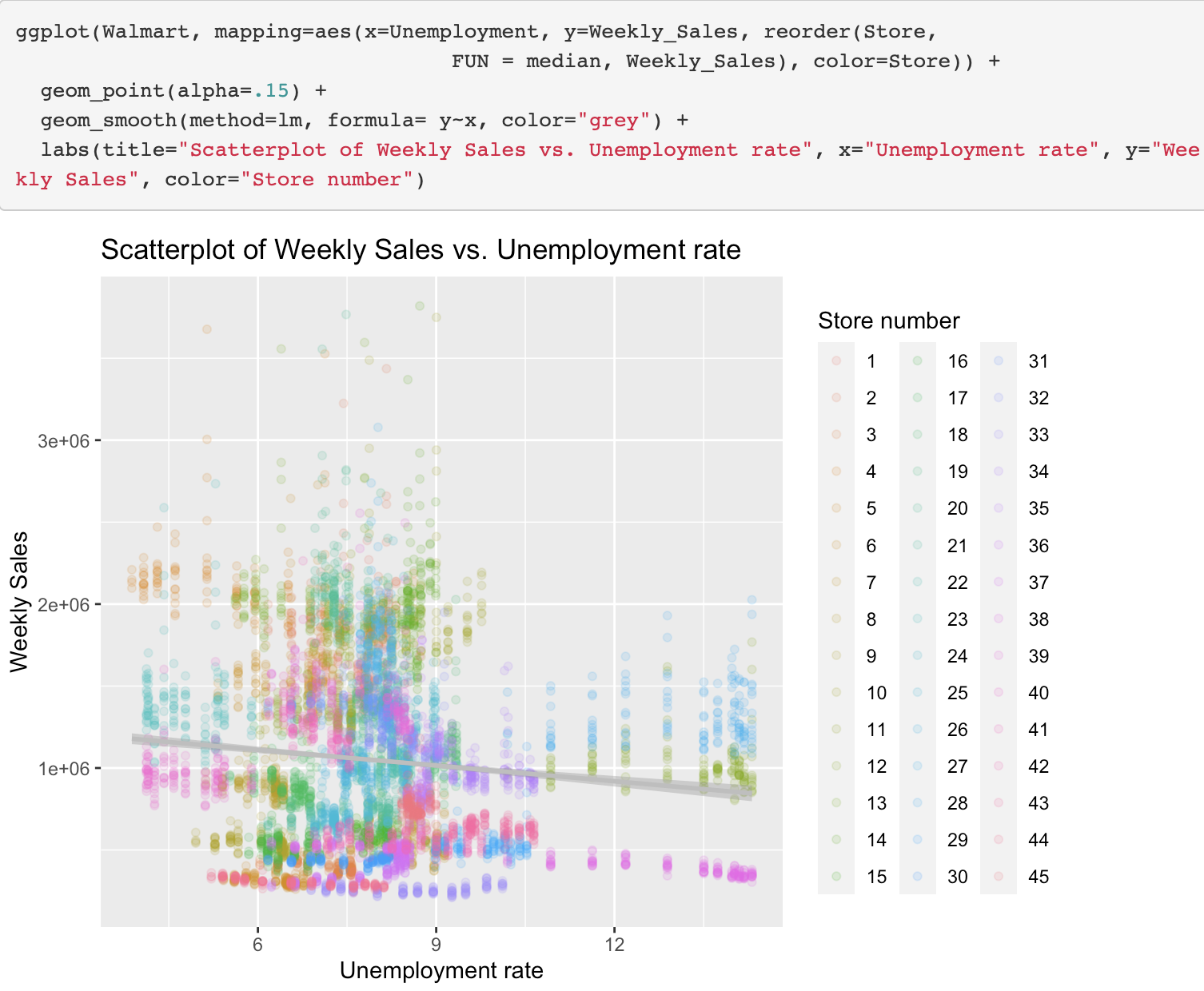
1. **Are weekly sales influenced by an increase in unemployment rate and temperature?**
2. Important details of the analysis:

From the “Scatterplot of Weekly Sales vs. Unemployment Rate,” there appeared to be a trend of decreasing weekly sales as unemployment rate increases. We confirmed that unemployment rate was a significant predictor of weekly sales by checking its p-value of the estimated coefficients from the Gamma regression model. Using the fitted model, we see that the estimated weekly sales decreased by 2% as the unemployment rate increased by 1%.

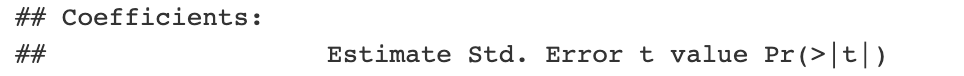
Both the “Scatterplot of Weekly Sales vs. Temperature by Holiday Weeks” and “Scatterplot of Weekly Sales vs. Temperature by Year” show a slight decrease in weekly sales as temperature increases. Temperature was also a significant predictor of weekly sales as evidenced by its low p-value. The fitted model suggested that weekly sales decreased by 0.0605% as temperature increases by 1°F.

1. Coding:

*Code Figure 12*

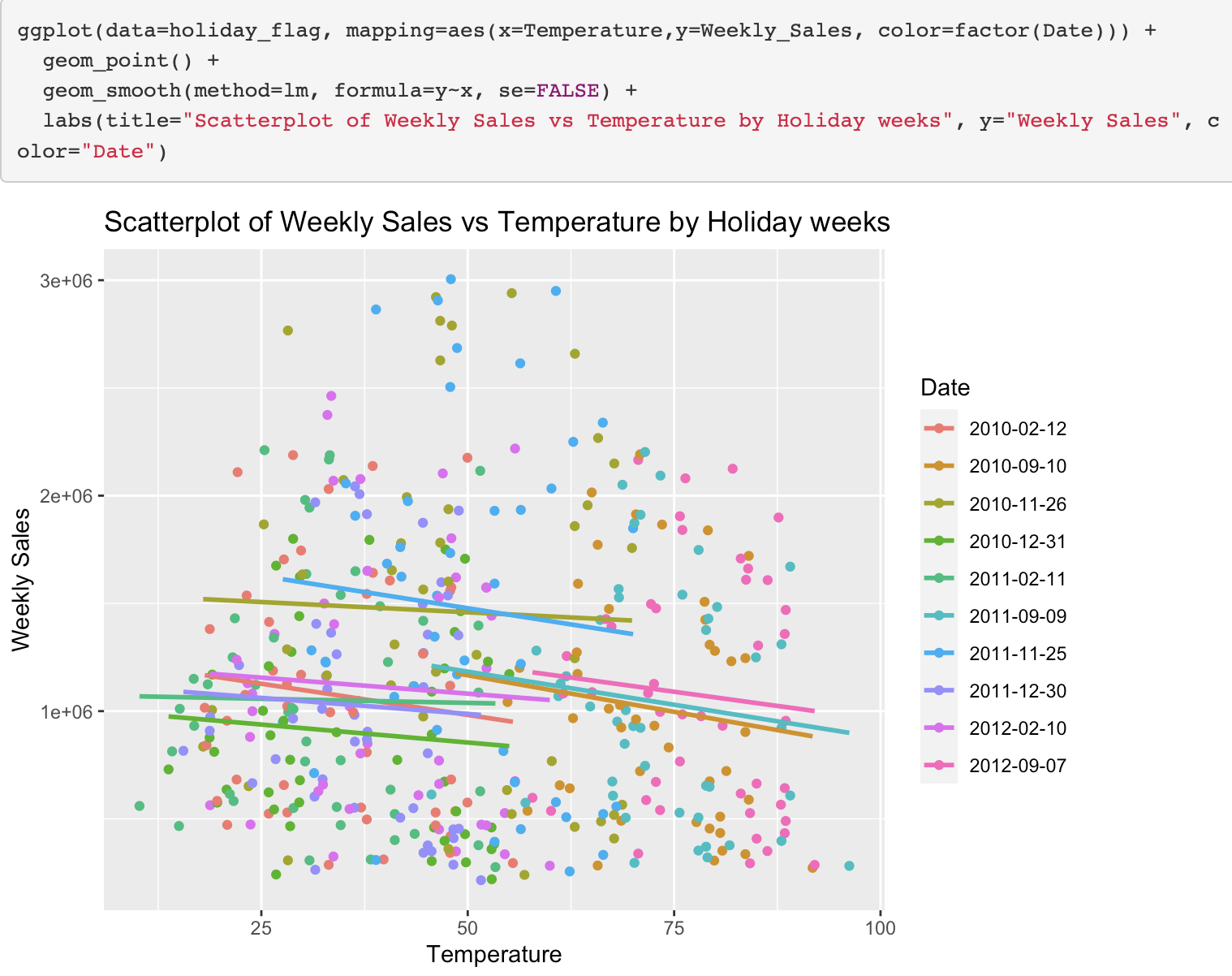


*Code Figure 2*

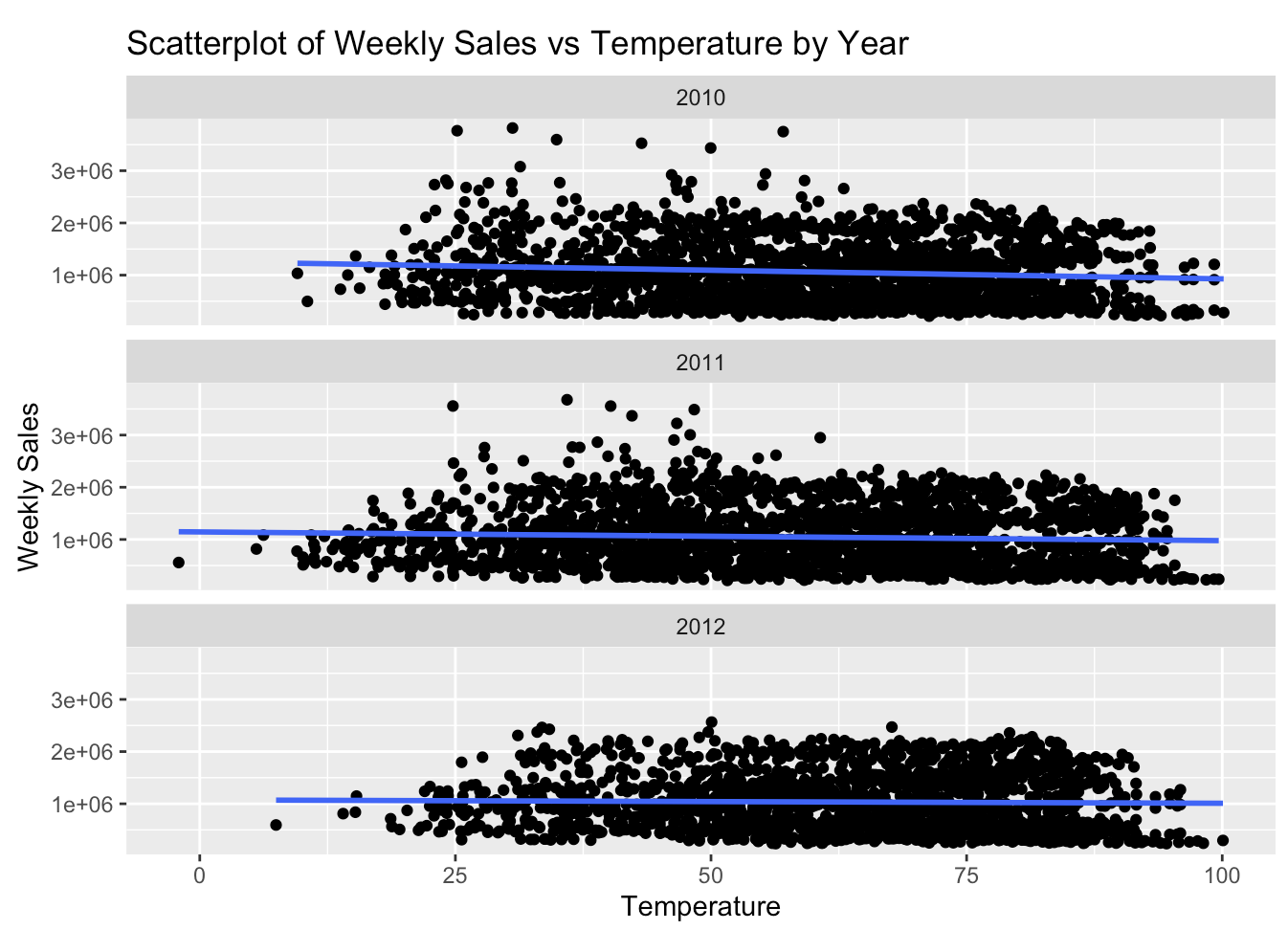




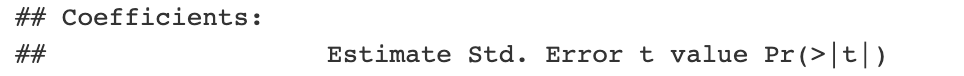
*Code Figure 13*



*Code Figure 14*



*Code Figure 2*





1. Interpretation:

The initial exploratory data analyses and the Gamma regression model both confirmed that increasing unemployment rates and temperature showed a slight decrease in weekly sales. Increasing unemployment rates may cause customers to be more careful with their money due to job insecurity. Higher temperatures also show a slight decrease in weekly sales which may be due to the possibility that customers do not want to go outside during a heatwave. Nevertheless, customers still have to buy necessities like groceries to survive which may account for there being only a slight decrease in weekly sales.

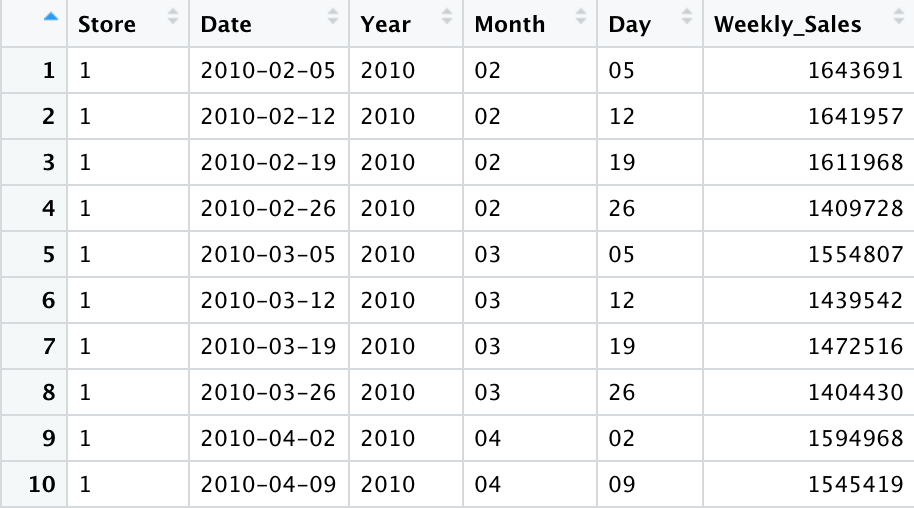
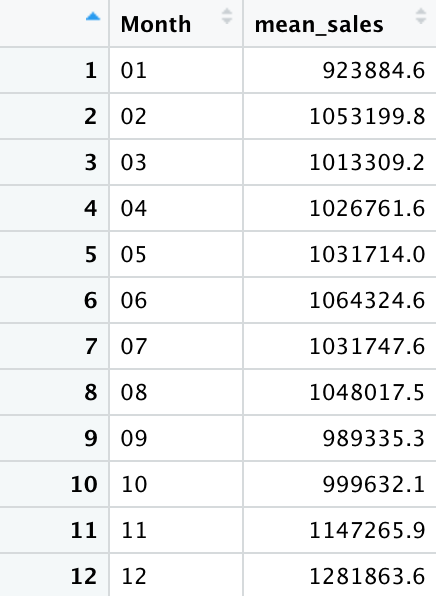
1. **What range can we expect average weekly sales to fall in?**
2. Important details of the analysis:

Since the “Walmart” data set had the date in the format YYYY/MM/DD, we separated the year, month, and day into 3 separate variables saving it into the tibble “Walmart\_year”. This tibble contained the same number of observations and variables as “Walmart” in addition to the new variables “Year”, “Month”, and “Day”. Then, we wanted to get the average sales of each month from all three years combined together. To do so, we created a new tibble called “Walmart\_month” and grouped by the month to get the mean sales using the summarize() function. This tibble only contained 12 observations and the 2 variables month and mean sales. “Walmart\_month” will be used as the initial sample of our bootstrap distribution of the sample mean of weekly sales.

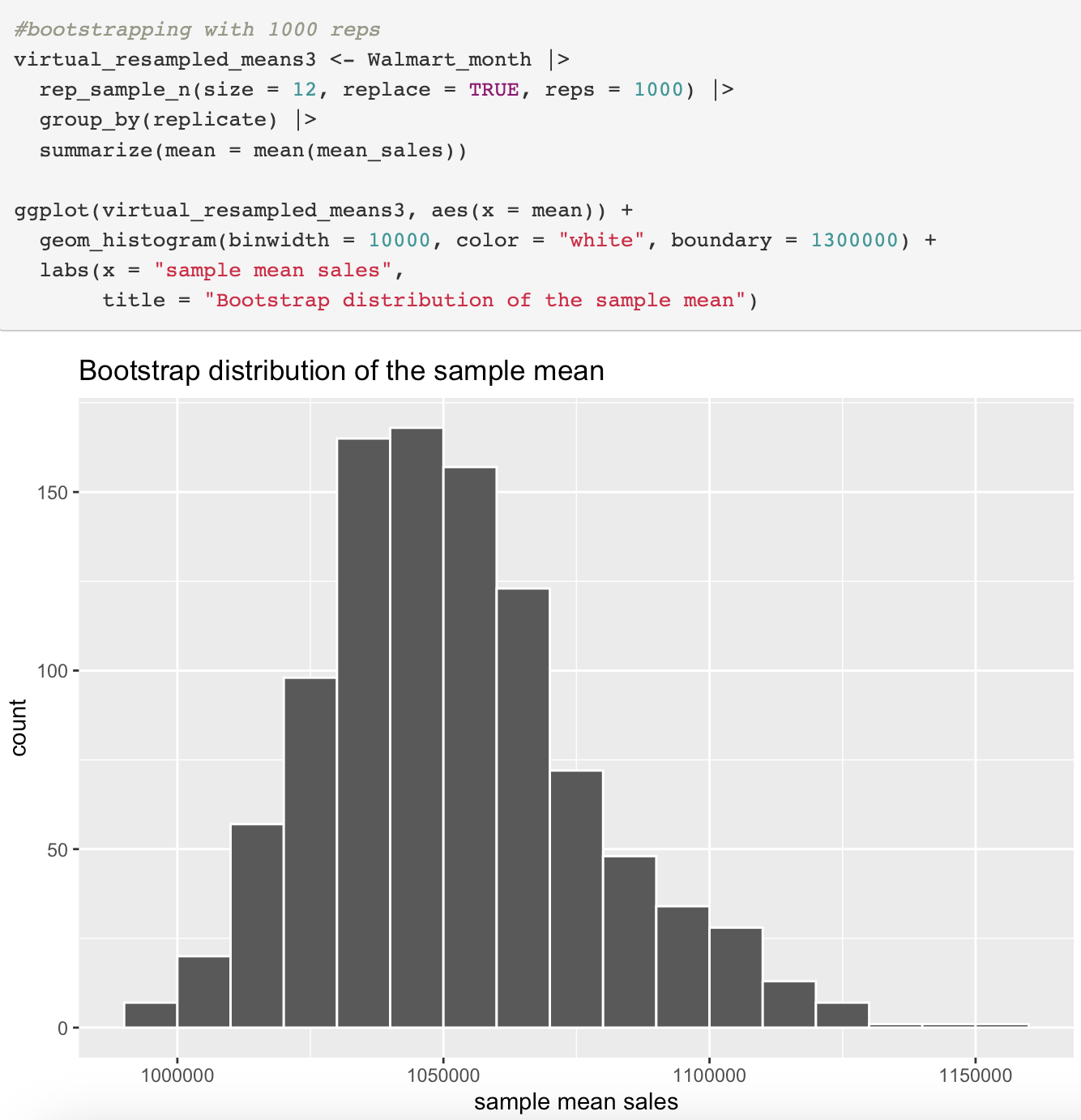
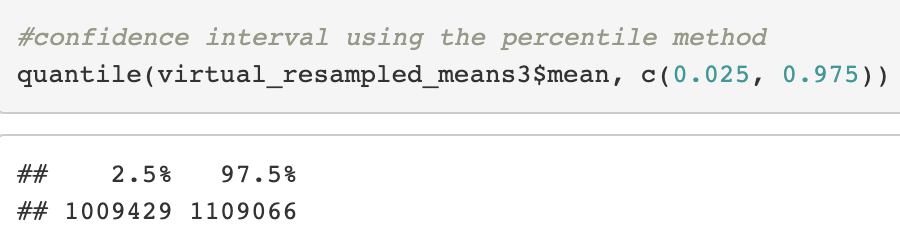
After bootstrapping with 50, 100, and 1000 repetitions, we decided that using the bootstrap distribution of 1000 repetitions would be the most appropriate in conducting our confidence intervals since it had the most normal distribution. However, there was still a very slight right skew in the bootstrap distribution which would make the percentile method in constructing the confidence interval the most appropriate method. The 95% confidence interval ended up being ($1,006,330, $1,102,230).

1. Coding:

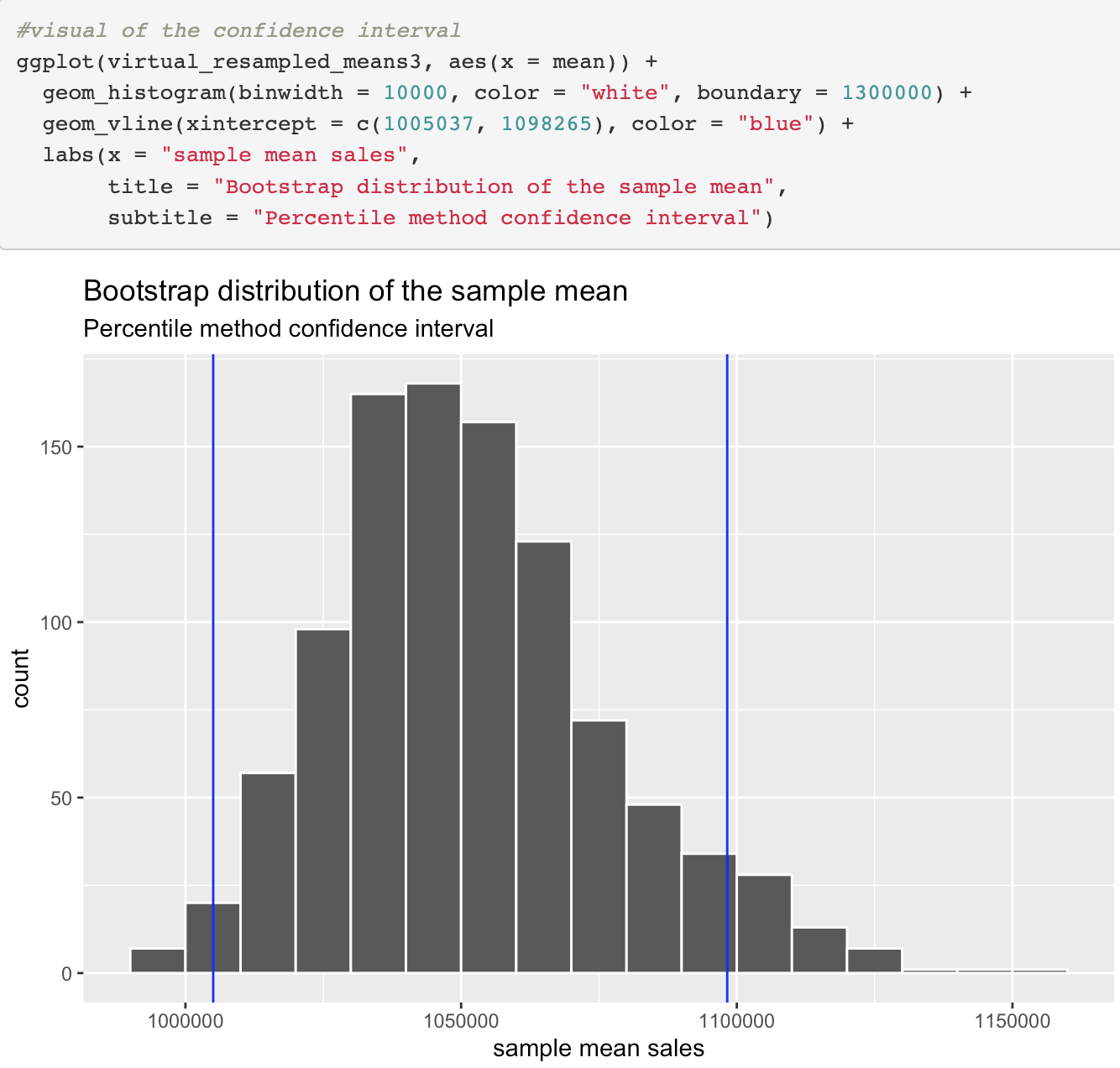
*Code Figure 15: Walmart\_year Code Figure 16: Walmart\_month*

* *

*Code Figure 17 Code Figure 18*

* *

*Code Figure 19*

**

1. Interpretation:

Comparing the actual mean of weekly sales and the sample mean of the bootstrap distribution, we see that the 2 values are very close with the actual mean being $1,046,965 and the sample mean being $1,050,444. With 95% confidence, we can say that the population mean of weekly sales at Walmart will be in the interval ($1,006,330, $1,102,230).

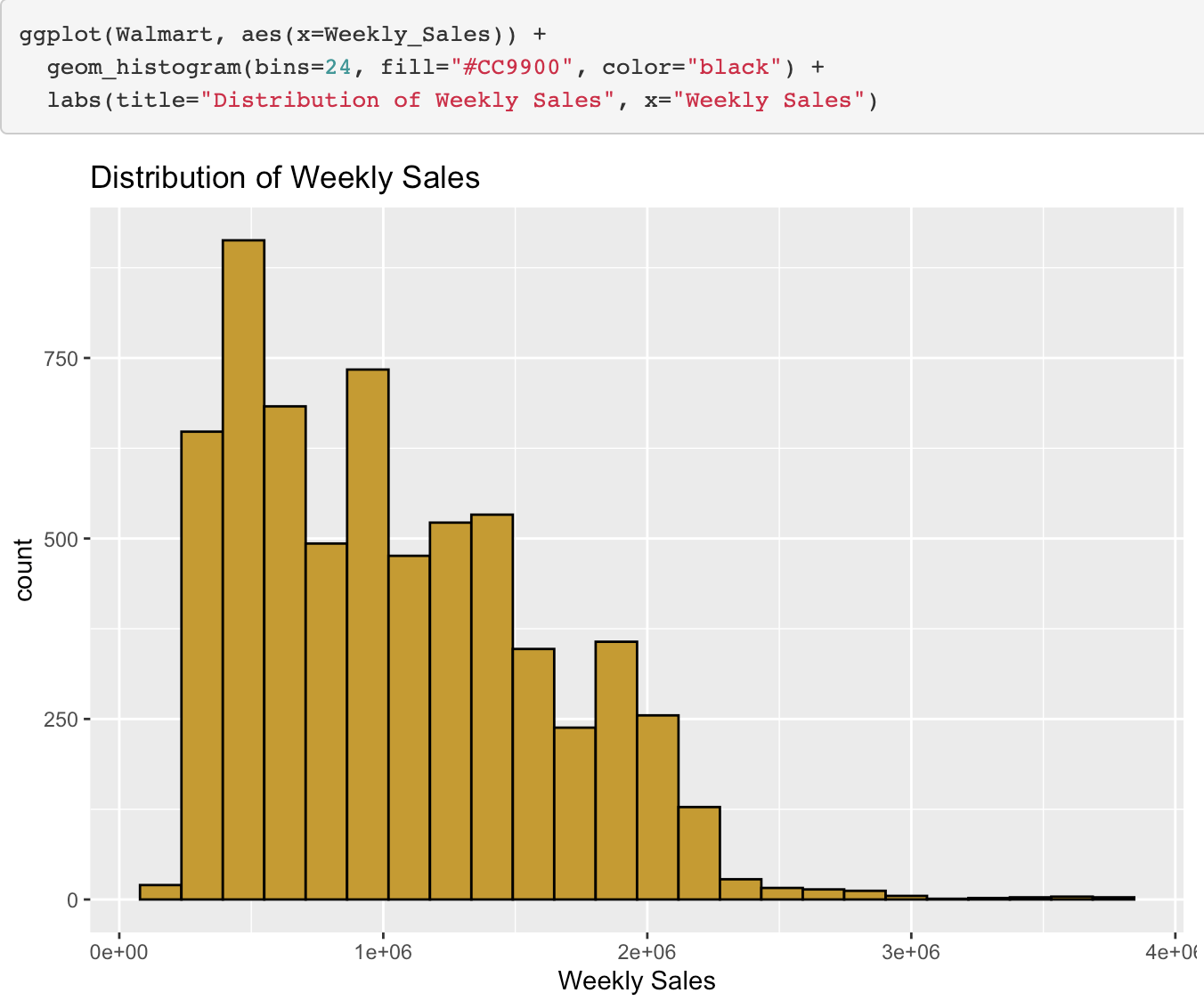
**Conclusion**

Due to the right skew in the distribution of Walmart’s weekly sales, we fitted the Gamma regression model to account for it. All variables were found to be significant predictors of weekly sales. However, after doing some predictions using the fitted model, we found that it overestimated weekly sales by over 30%. After doing some exploratory analyses, we found that many outliers of high weekly sales were found in the weeks from Thanksgiving to Christmas. This seemed to explain why the fitted model was overestimating weekly sales and why the bootstrap distribution had a very slight right skew even after 1000 resamples.

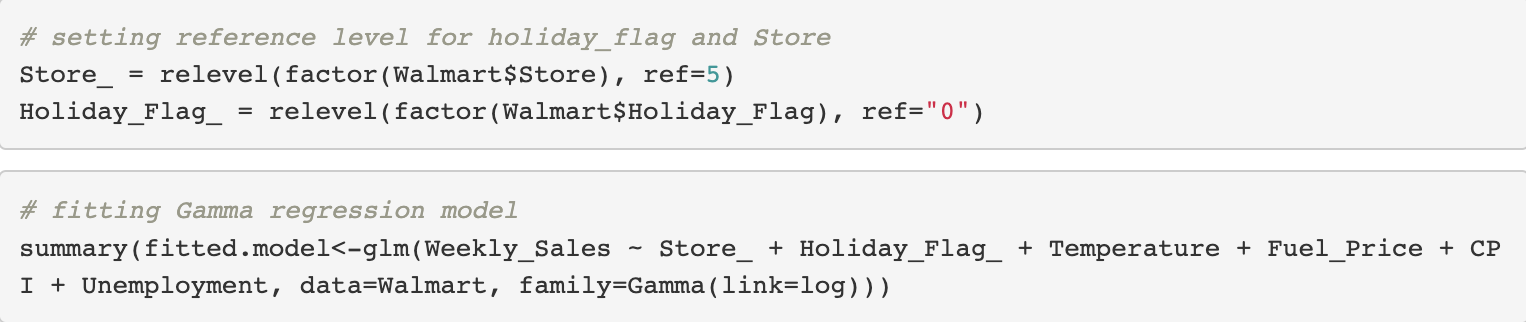
To put simply, the weeks from Thanksgiving to Christmas garnered the most sales at Walmart. This could be due to customers buying gifts and preparing dinners for the holiday season. Around this time, there were also many sales for Black Friday and Christmas which could also be a driving factor for the high sales. Variables such as CPI, unemployment rate, and temperature do not appear to affect Walmart sales as much as the aforementioned weeks since customers still need to spend money on basic necessities. Walmart should increase their inventory and sales during Thanksgiving to Christmas to drive their sales up even more.

**Appendix**

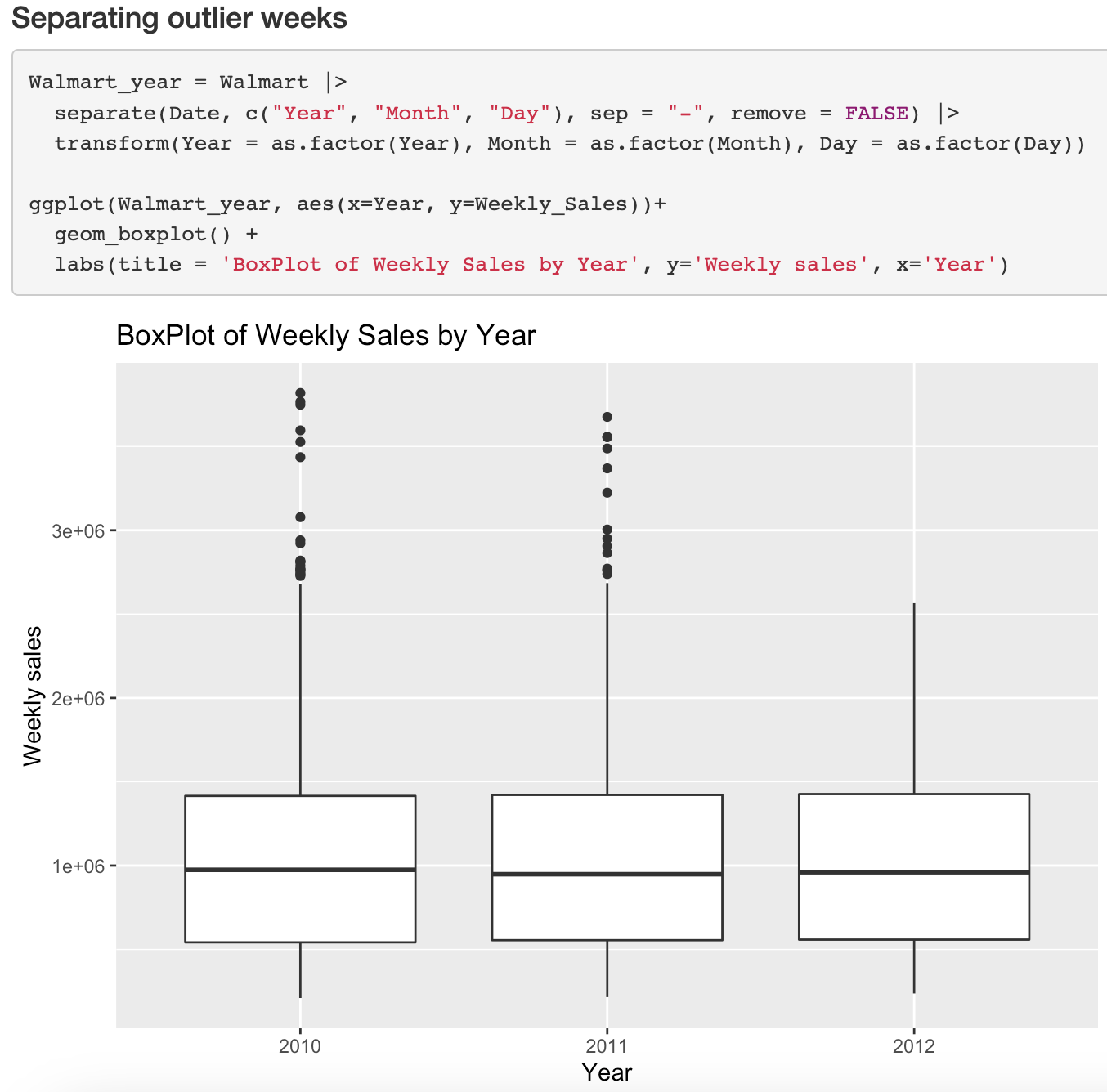
*Figure 1*



*Figure 2*

**

*Figure 3*

**

*Figure 4*



*Figure 5*

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

*Figure 6*

Chart

Description automatically generated

*Figure 7*

Table

Description automatically generated with medium confidence

*Figure 8*

Graphical user interface

Description automatically generated with medium confidence

*Figure 9*

Text

Description automatically generated

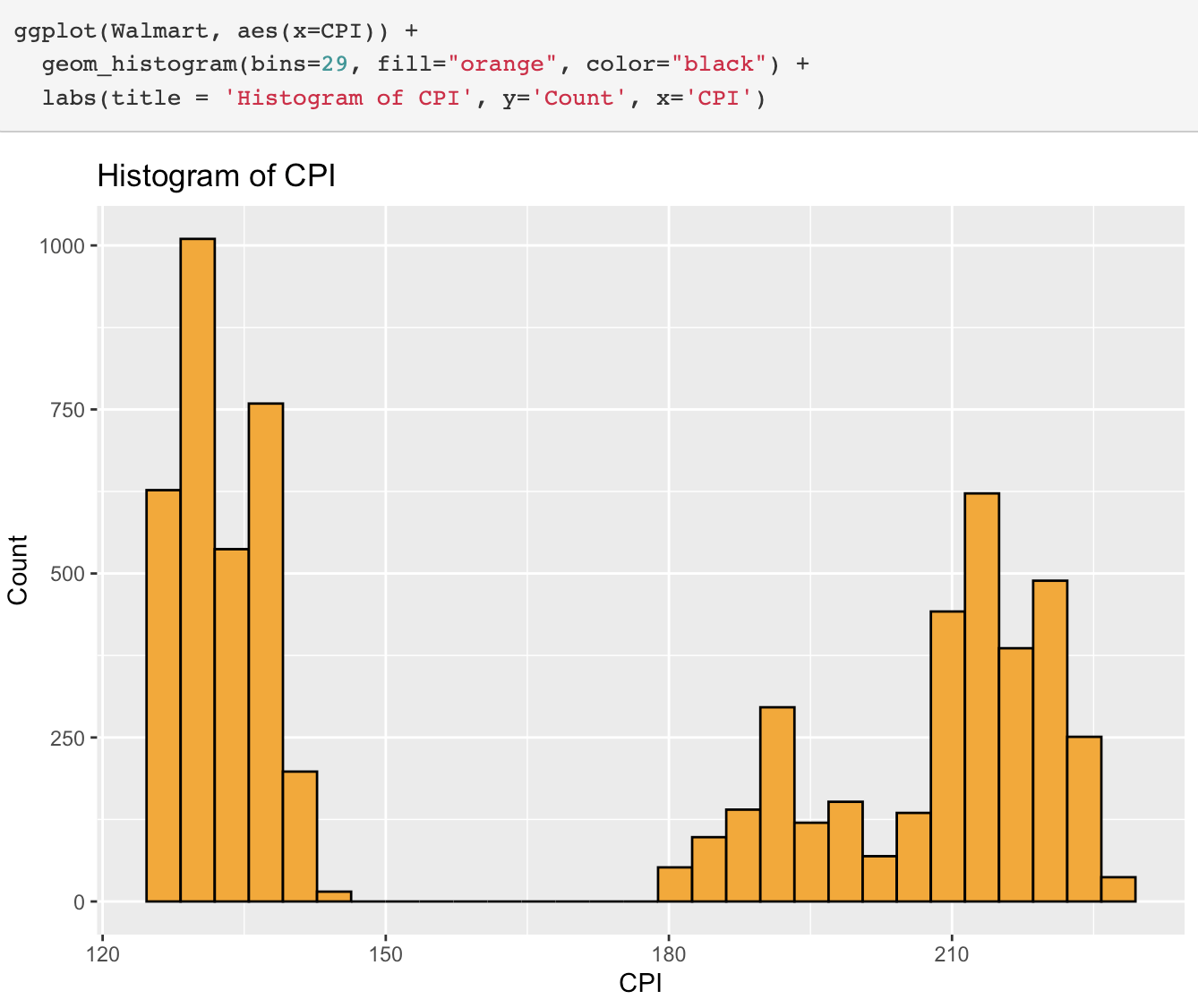
Graphical user interface, text, application

Description automatically generated

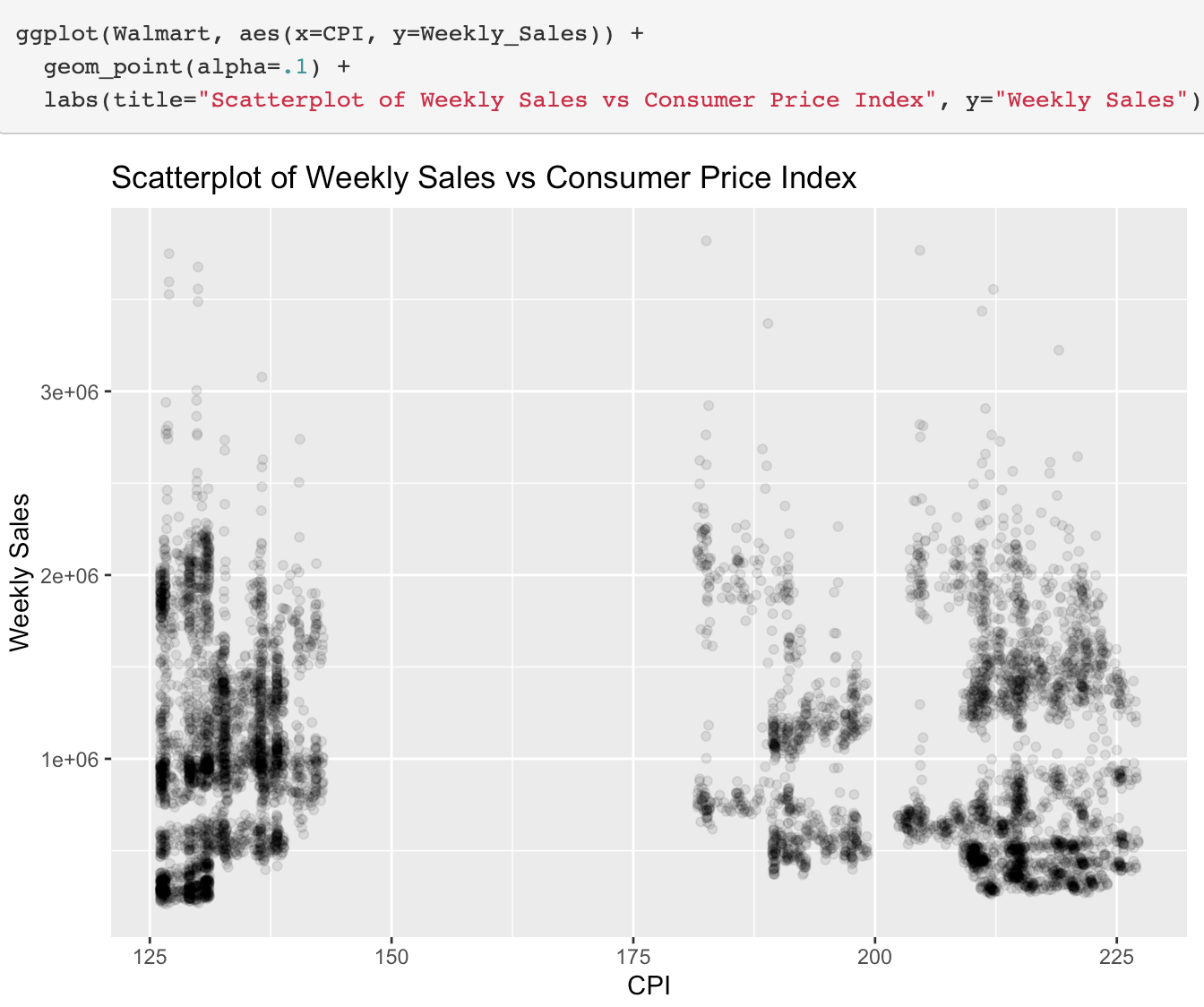
Text, letter

Description automatically generated

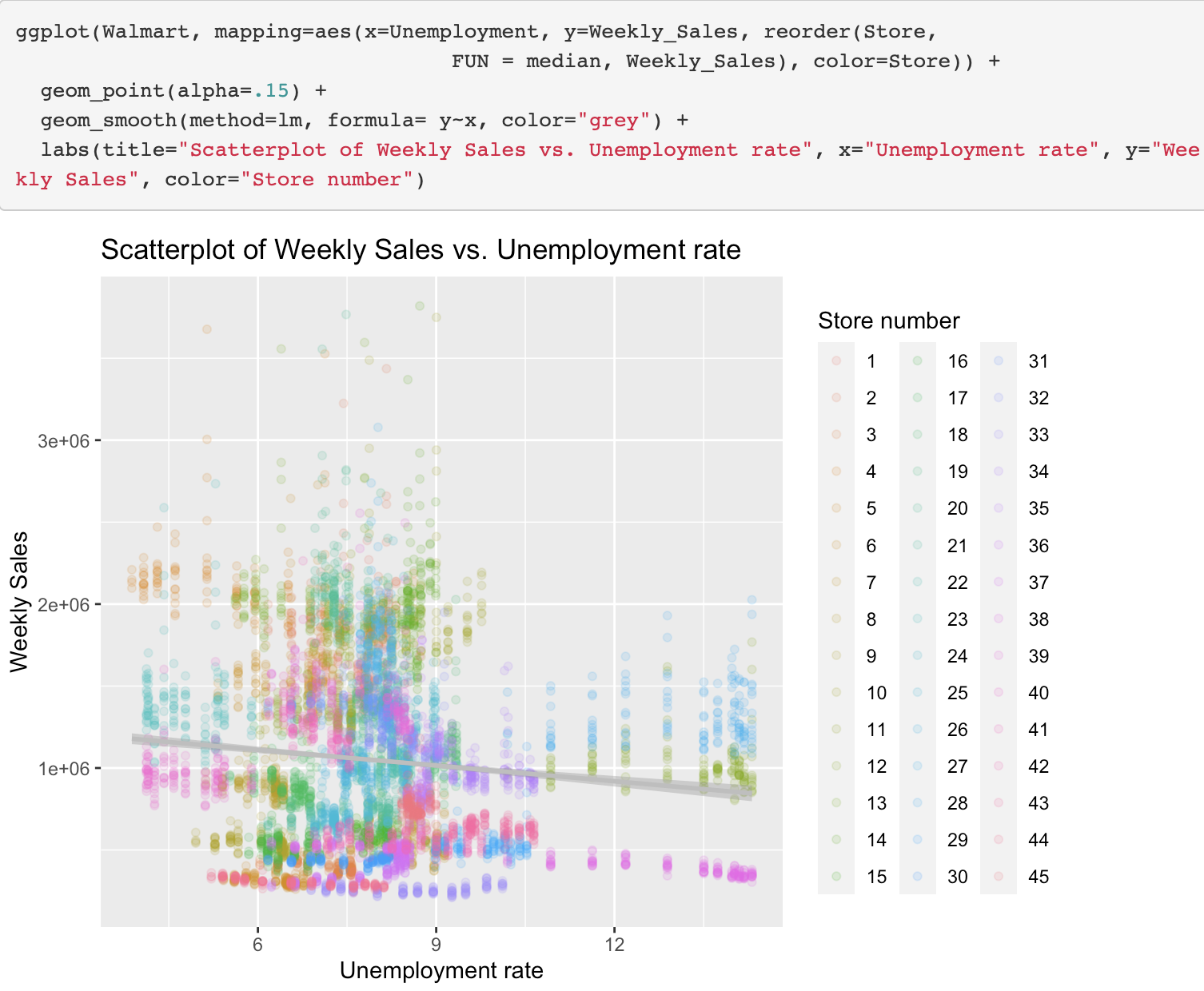
*Figure 10*

**

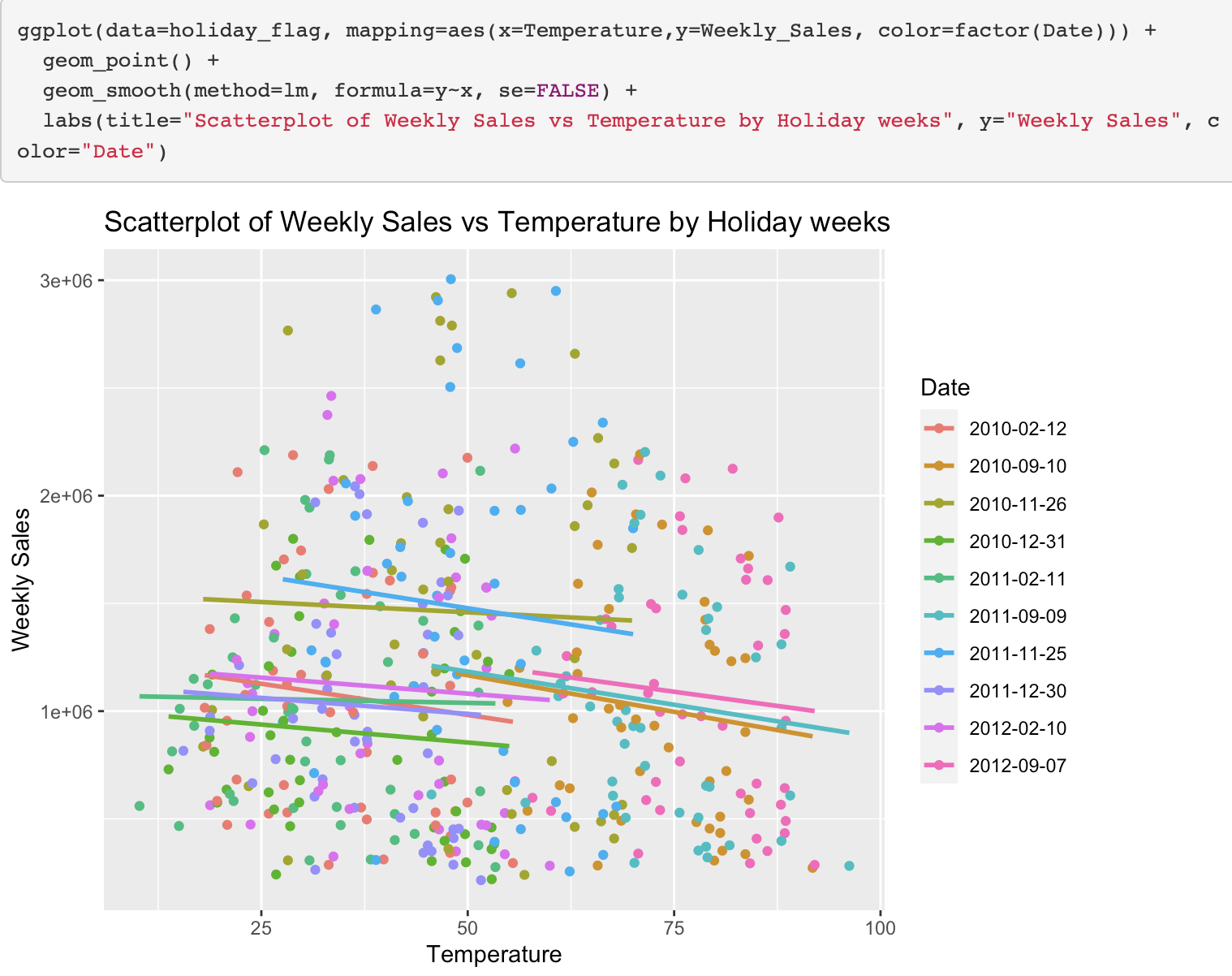
*Figure 11*



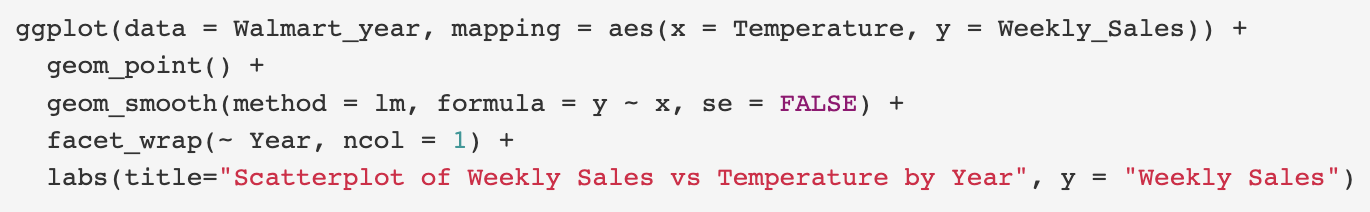
*Figure 12*



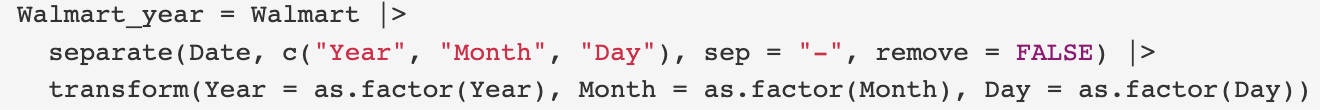
*Figure 13*



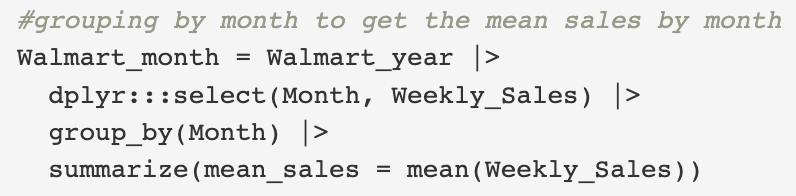
*Figure 14*

**

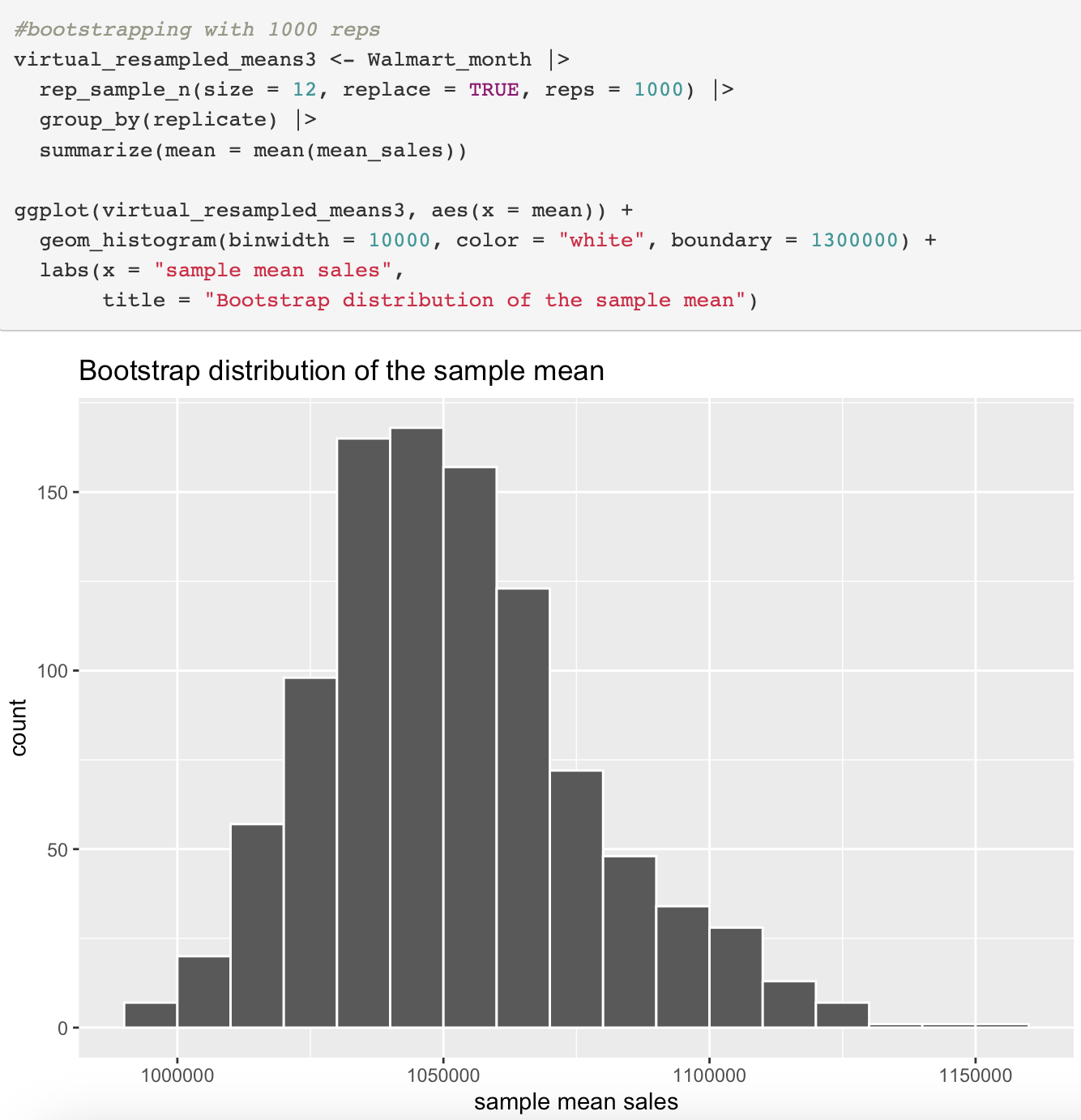
*Figure 15*



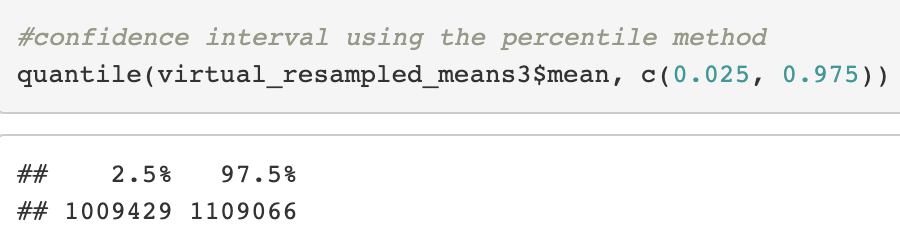
*Figure 16*

**

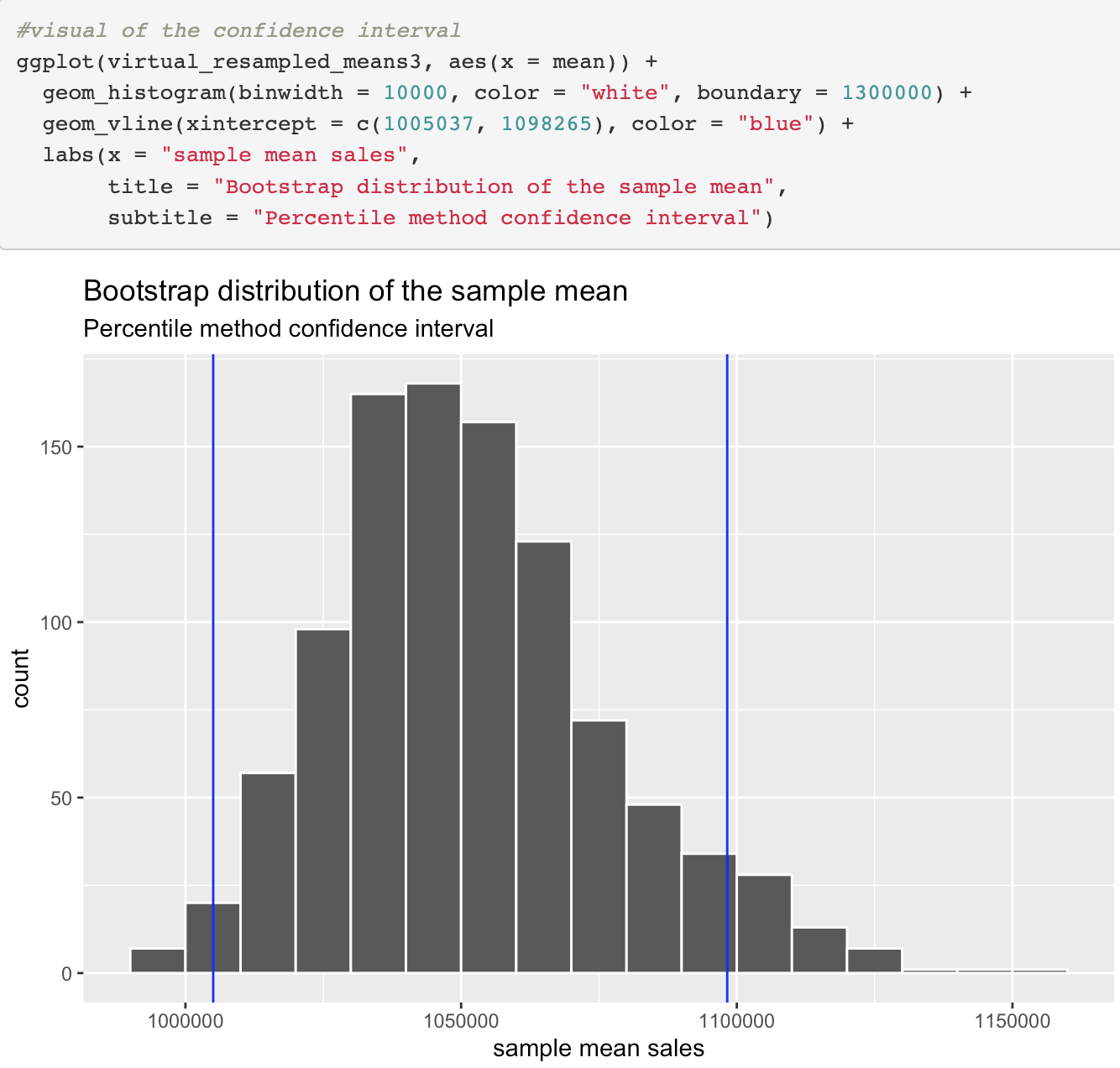
*Figure 17*

**

*Figure 18*

**

*Figure 19*

**