XYZ Store Sales Analysis

Jayadev KP | Lydia Savatsky |Raj Vardhan | Zhaoyan Zhi
11-13-2021

Table of Contents

Executive Summary

XYZ is a nationwide retail company stretching around 45 stores throughout the United States. The company wants to determine how its weekly sales value relates to various attributes such as store-level attributes, macro-economic variables, and external factors. The company is also interested in forecasting weekly sales based on the given attributes. To meet this requirement, a sales model that explains almost 70% of our historical sales values has been created. The model uses the most significant set of attributes, which consisted of store size, unemployment rate, and CPI, to estimate a weekly sales amount. Estimating sales with our model will enable XYZ to make informed decisions and strategize actions towards optimizing inventory management, allocating staffs, opening stores, and setting sales benchmarks.

I. Introduction

XYZ is a retail chain with stores across the U.S. looking to estimate store sales. The advantages of being able to estimate the store sales include the following:

- Optimization of inventory management.
- Improving the efficiency of staff allocation.
- Planning promotional activities.

The key questions that we wish to address are:

1. What are the most important factors that impact weekly sales?
2. How useful are these factors in predicting weekly sales?

Our data consists of weekly sales per store for forty-five different stores across 143 weeks, amounting to about 6K data points. The data was collected from February 2nd, 2010, until November 1st, 2012. To better understand which factors have a relationship with weekly sales, we explored three areas that impact weekly sales: store level attributes, macro-economic variables, and external variables. Store level attributes include store type and store size. Macro-economic variables consist of CPI, unemployment rate, and fuel price. External variables include the average temperature per week and whether there was a holiday. The table below describes each of these variables and their ranges/levels in detail.

Variable Name	Description	Range/Levels
Weekly Sales	Total sales per store across one week of operations	$300K - $4M
Store Size	Size of store	35K sq ft - 220K sq ft
Store Type	Describes the three levels of stores on the basis of the products the store offers and the way the store operates	Store Type A, Store Type B and Store Type C
CPI	Consumer Price Index - Measure of average change over time in the prices paid for consumer goods and services	$126.1 - $227.2
Unemployment	The proportion of people who are unemployed as a percentage of the labor force	3.88% - 14.31%
Fuel Price	Price of gas per gallon in dollars for a particular week	$2.47 - $4.47
Temperature	Average temperature of the week	-2.06°F - 100.14°F
IsHoliday	Whether the week is a special holiday week	True, False

The variables in the above table are of interest because they can be collected with ease and updated with little cost. Additionally, these metrics are intuitively related to consumer demand, and thus they are objective measures that are unambiguous to a business.

II. Analysis

In order to develop an understanding of the important factors that relate to weekly sales, we explored each variable and their relationship with weekly sales. We then generated a model based on the most important factors to evaluate such relationships.

A. Objective

The objective of our analysis was to build a model to identify factors that significantly impact weekly sales and use these factors to estimate sales. We explored factors such as store size, store type, CPI, unemployment rate, fuel price, temperature, and whether the week lands on a holiday.

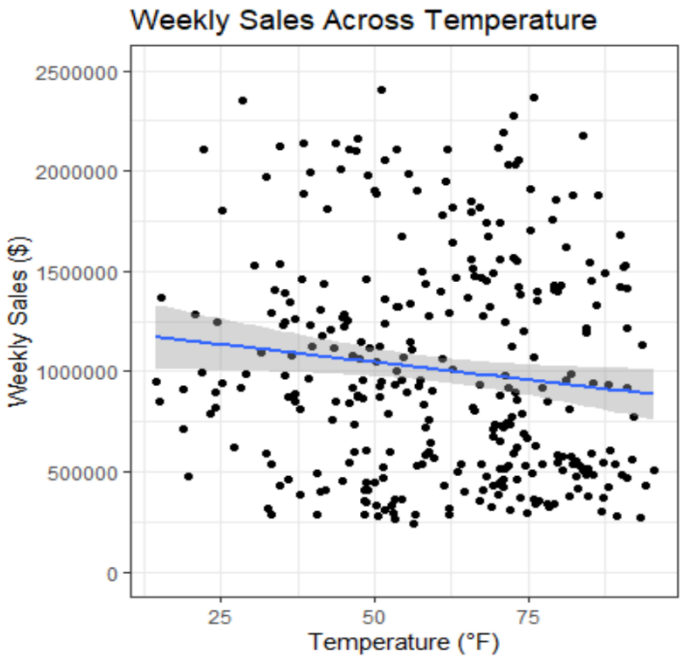
B. Exploratory Data Analysis

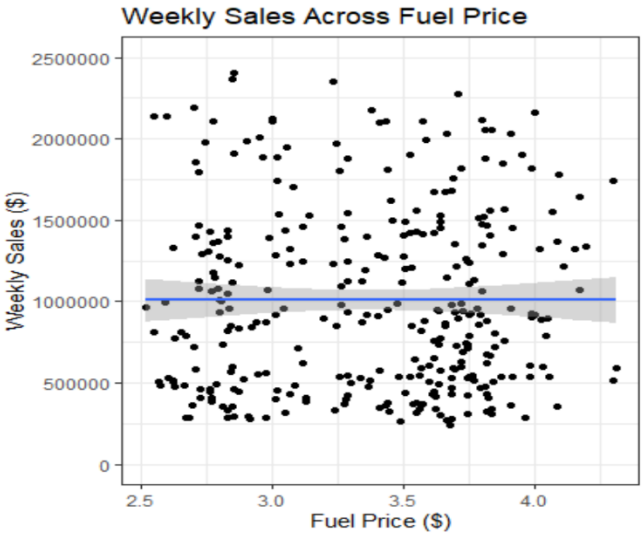
Before building our model, we first analyzed the relationship between sales and each variable to better understand whether there was a positive or negative relationship between the

variable and sales. Below we have included graphs to better understand how weekly sales change as each variable changes.

Temperature:

The figure to the right shows that as temperature increases, we see a drop in weekly sales. This can be understood by the blue line, which has a downward trend.



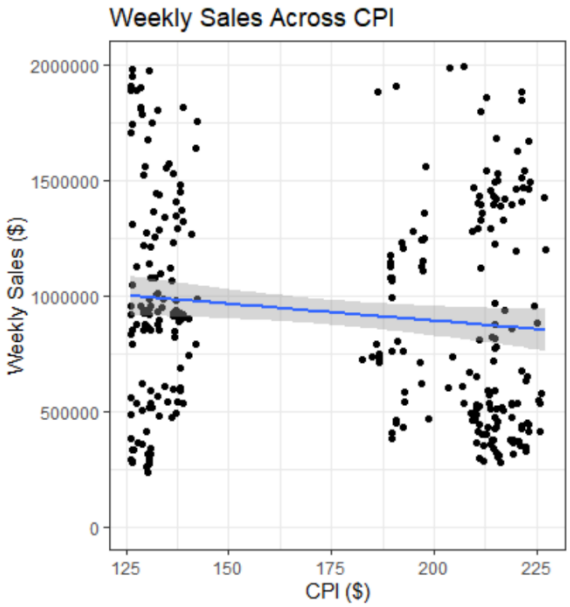


Fuel Price:

As highlighted by the blue line in the graph to the left, fuel price is not related to weekly sales as there is no downward/upward trend in sales over fuel price.

Unemployment

The figure to the right shows a blue line which denotes the relationship between the unemployment rate and weekly sales. There exists a negative trend, so we can conclude that weekly sales decreases with an increase in the unemployment rate.

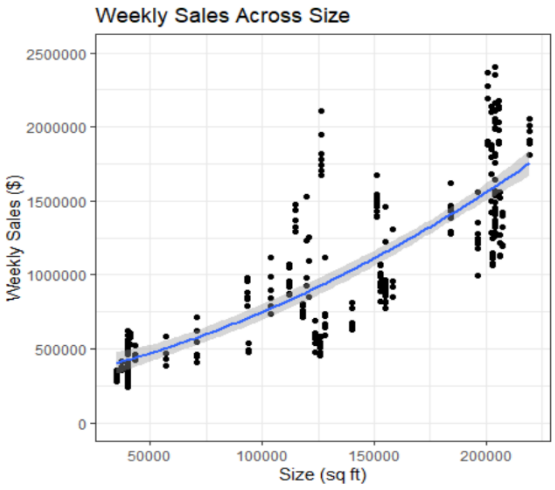


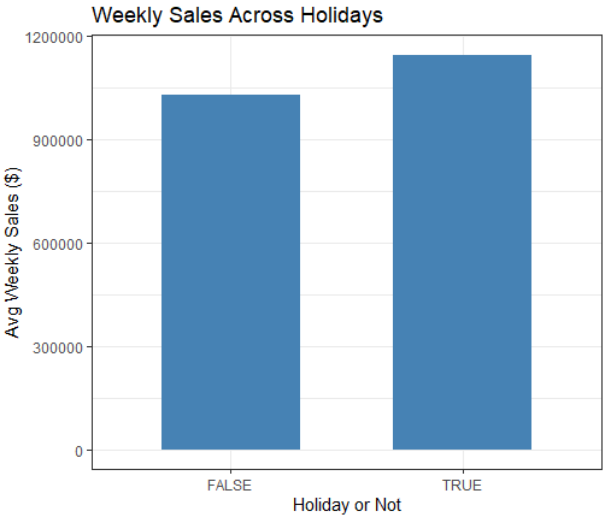
CPI:

CPI appears to be negatively related to Sales as indicated by the blue trendline in the graph to the left.

Store Size:

Size of the store has a curvilinear relationship with sales. When size increases for the same unit, the sales of stores increase disproportionately.



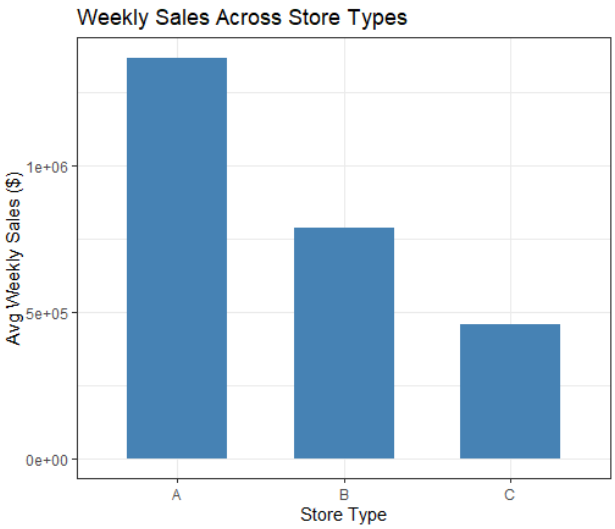


Is Holiday:

Weekly sales on holiday weeks are, on an average, 11.2% higher than non-holiday weeks.

Store Type:

We have three levels of store type for a retail store. Store type A represents supercenters, which offer a one-stop shopping experience by supplying groceries, electronics, apparel, toys, and home furnishings. Store Type B represents midsize discount stores that supply electronics, apparel, toys, home furnishings, health and beauty aids, hardware, and more. Store type C represents neighborhood markets, which include small stores that sell affordable groceries and merchandise. Store type C has average weekly sales 66% lower than store type A while store type B has average weekly sales around 42% lower than store type A.



C. Model Building

We generated a linear model to evaluate the relationship between the attributes described above and weekly sales per store for company XYZ. Multiple combinations of attributes were tested in the model, and the most efficient model was selected (kindly refer to the appendix for the model selection process). The model selection was based on the following criteria:

- Variation of the weekly sales explained by the model
- Degree of confidence of the relationship between different attributes and sales
- Simplicity of the model

- Interpretability of the model in business context

The final model selected included a relationship between store size, CPI, and unemployment rate. For technical details about this model, please refer to the appendix.

D. Model Evaluation

The model explains 68% of the weekly sales. Our degree of confidence of the relationship between Size, CPI and Unemployment with sales in the model is high (for details in evaluation see appendix).

Based on our model, we know that there is a positive curvilinear relationship between size and weekly sales. There is a negative relationship between CPI and sales: the weekly sales decreases by $19,930 for one unit increase in unemployment index when size and CPI stay unchanged. Finally, there is a negative relationship between unemployment and sales: the weekly sales decreases by $1,409 for one unit increase in CPI when size and unemployment stay unchanged. For more information kindly refer to the appendix.

Although this model explains a significant portion of weekly sales, there are a few limitations. There is an issue with the modeling assumption and additional analyses should be done to follow up with this limitation (kindly refer to the appendix for assumption checking). Another limitation of our model is that it does not include business levers related to the store such as store discounts, promotions, and store layout. The inclusion of these factors could help us further understand what attributes have a relationship with weekly sales.

III. Conclusion

In generating this linear model, our goal was to discover which factors impact the weekly sales of a store. We have found that the size of a store, unemployment rate, and the CPI value are the most important factors that impact the weekly sales. Furthermore, the model is able to explain 68% of weekly sales in our data.

This model can be used to benchmark a store's weekly sales. For example, the company can examine a particular store's size, the unemployment rate in that store's region, and the CPI value to understand how a store should be performing in terms of weekly sales. Additionally, the model can be utilized to scope out new regions to open stores in. By looking into a particular region's unemployment rate and CPI, a manager can understand what a store's weekly sales might look like, depending on the size of the store, if XYZ were to open a store in that region.

IV. References

Our data was collected from a Kaggle data set. For information concerning the data set and the problem statement please refer to https://www.kaggle.com/manjeetsingh/retaildataset. We used the variables weekly sales, temperature, fuel price, CPI, unemployment, size, holiday, and store type from this data set for our analysis. For information about data cleaning, please refer to the appendix.

V. Appendix

A. Data Cleaning

The dataset we collected from Kaggle was quite clean to begin with, but we did have to perform some pre-processing to reorganize the data in a manner that we could analyze. After reading in each data set provided on Kaggle (features, sales, and stores data sets), we first grouped the sales data set by store and date to get the total sales for each store on a weekly basis. We then merged the features and stores data sets to our grouped by sales, creating a data frame that consisted of weekly sales, temperature, fuel price, CPI, unemployment, size, IsHoliday, and Type for all 45 across 143 weeks. Throughout our pre-processing of the data, we also checked the distribution of the variables to detect for outliers. We found that no variables had explicit differences in their distributions, meaning there was no need to remove any outliers.

B. Model Selection

Below we have outlined the hypothesis model for each model that we tested along with a table of the beta and p-values from each model. We will then compare these eight models to express our selection of the best model.

Model 1: Weekly Sales Across all Factors

$$Weekly\ Sales = B0 +$$
$$B1 * Temperature +$$
$$B2 * Fuel\ Price +$$
$$B3 * CPI +$$
$$B4 * Unemployment +$$
$$B5 * Size$$
$$B6 * IsHoliday_{True} +$$
$$B7 * TypeB +$$
$$B8 * TypeC +$$
$$Error\ Term$$

Variable Name	Beta Values	P-values
Intercept	4.127e+05	< 2e-16
Temperature	1.262e+03	8.54e-08
Fuel Price	-2.643e+04	0.00387
CPI	-1.379e+03	< 2e-16
Unemployment	-2.505e+04	< 2e-16
Size	7.926e+00	< 2e-16
IsHolidayTrue	9.362e+04	5.24e-09
TypeB	4.824e+04	5.27e-05
TypeC	1.948e+05	< 2e-16

Model 2: Weekly Sales with Size^2 and all Factors

$$Weekly\ Sales = B0 +$$
$$B1 * Size +$$
$$B2 * Size\text{^}2 +$$
$$B3 * Temperature +$$
$$B4 * Fuel\ Price +$$
$$B5 * CPI +$$
$$B6 * Unemployment$$
$$B7 * IsHoliday_{True} +$$
$$B8 * TypeB +$$
$$B9 * TypeC +$$
$$Error\ Term$$

Variable Name	Beta Values	P-values
Intercept	6.518e+05	< 2e-16
Size	2.510e+00	6.11e-09
$Size^2$	2.277e-05	< 2e-16
Temperature	1.067e+03	4.67e-06
Fuel_Price	-2.461e+04	0.00644
CPI	-1.590e+03	< 2e-16
Unemployment	-2.566e+04	< 2e-16
IsHolidayTrue	9.167e+04	6.98e-09
TypeB	1.472e+05	< 2e-16
TypeC	1.865e+05	< 2e-16

Model 3: Removing Fuel Price from Model 2

$$Weekly\ Sales = B0 +$$
$$B1 * Size +$$
$$B2 * Size\text{^}2 +$$
$$B3 * Temperature +$$
$$B4 * CPI +$$
$$B5 * Unemployment$$
$$B6 * IsHoliday_{True} +$$
$$B7 * TypeB +$$
$$B8 * TypeC +$$
$$Error\ Term$$

Variable Name	Beta Values	P-values
Intercept	5.621e+05	< 2e-16
Size	2.479e+00	9.43e-09
$Size^2$	2.285e-05	< 2e-16
Temperature	9.432e+02	3.70e-05
CPI	-1.522e+03	< 2e-16
Unemployment	-2.489e+04	< 2e-16
IsHolidayTrue	9.371e+04	3.15e-09
TypeB	1.456e+05	< 2e-16
TypeC	1.846e+05	< 2e-16

Model 4: Removing Temperature from Model 3

$$Weekly\ Sales = B0 +$$
$$B1 * Size +$$
$$B2 * Size\text{^}2 +$$
$$B3 * CPI +$$
$$B4 * Unemployment$$
$$B5 * IsHoliday_{True} +$$
$$B6 * TypeB +$$
$$B7 * TypeC +$$
$$Error\ Term$$

Variable Name	Beta Values	P-values
Intercept	6.046e+05	< 2e-16
Size	2.334e+00	6.05e-08
$Size^2$	2.330e-05	< 2e-16
CPI	-1.434e+03	< 2e-16
Unemployment	-2.352e+04	< 2e-16
IsHolidayTrue	8.305e+04	
TypeB	1.422e+05	< 2e-16
TypeC	1.843e+05	< 2e-16

Model 5: Removing IsHoliday from Model 4

$$Weekly\ Sales = B0 +$$
$$B1 * Size +$$
$$B2 * Size\text{^}2 +$$
$$B3 * CPI +$$
$$B4 * Unemployment$$
$$B5 * TypeB +$$
$$B6 * TypeC +$$
$$Error\ Term$$

Variable Name	Beta Values	P-values
Intercept	6.093e+05	< 2e-16
Size	2.334e+00	6.48e-08
$Size^2$	2.330e-05	< 2e-16
CPI	-1.434e+03	< 2e-16
Unemployment	-2.339e+04	< 2e-16
TypeB	1.421e+05	< 2e-16
TypeC	1.841e+05	< 2e-16

Model 6: Removing Type from Model 5

$$Weekly\ Sales = B0 +$$
$$B1 * Size +$$
$$B2 * Size\text{^}2 +$$
$$B3 * CPI +$$
$$B4 * Unemployment$$
$$Error\ Term$$

Variable Name	Beta Values	P-values
Intercept	7.187e+05	<2e-16
Size	2.981e+00	<2e-16
$Size^2$	1.692e-05	<2e-16
CPI	-2.063e+04	<2e-16
Unemployment	-1.455e+03	<2e-16

Model 7: Removing Unemployment from Model 6

$$Weekly\ Sales = B0 +$$
$$B1 * Size +$$
$$B2 * Size\text{^}2 +$$
$$B3 * CPI +$$
$$Error\ Term$$

Variable Name	Beta Values	P-values
Intercept	4.797e+05	<2e-16
Size	3.319e+00	<2e-16
$Size^2$	1.579e-05	<2e-16
CPI	-1.142e+03	<2e-16

Model 8: Removing CPI from Model 6

$$Weekly\ Sales = B0 +$$
$$B1 * Size +$$
$$B2 * Size\text{^}2 +$$
$$B3 * Unemployment +$$
$$Error\ Term$$

Variable Name	Beta Values	P-values
Intercept	3.499e+05	< 2e-16
Size	3.853e+00	< 2e-16
$Size^2$	1.355e-05	< 2e-16
Unemployment	-1.110e+04	< 2e-16

Comparison of Models

Model	k	R^2	s	Highest P-Value
Model 1	8	0.6733	322800	Fuel Price: 0.00387
Model 2	9	0.6816	318700	Fuel Price: 4.67e-06
Model 3	8	0.6813	318800	Temperature: 3.70e-05
Model 4	7	0.6804	319200	IsHoliday_True: 1.06e-07
Model 5	6	0.679	319900	
Model 6	5	0.6726	323000	
Model 7	4	0.6684	325100	
Model 8	4	0.6636	327400	

We have chosen our final model as Model 6, which includes store size, store size squared, CPI, and the unemployment rate. We have chosen this model because 1) it is simple, 2) it explains a majority of the variation in sales compared to other models, 3) it has a low amount of unexplained variance in sales compared to other models, and 4) it is interpretable in the business context.

C. Assumption Checking and Other Limitations

Assumption Checking

The linear regression model has three assumptions:

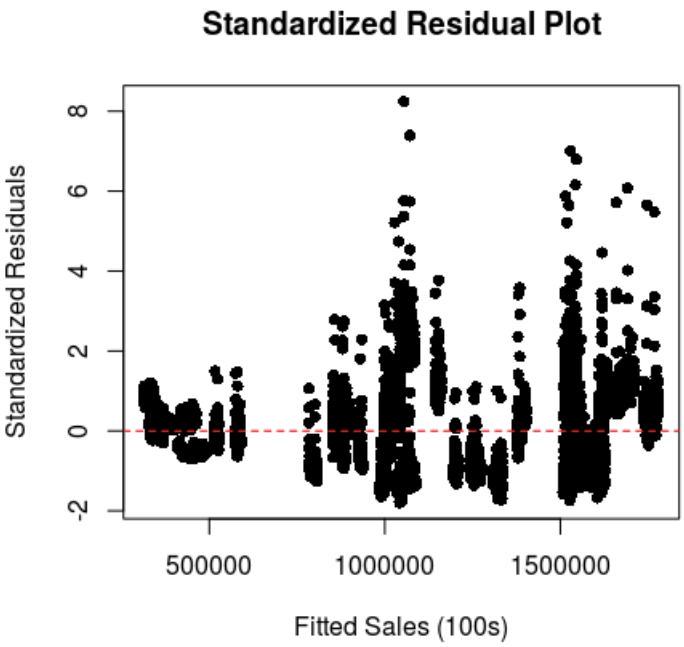
- Random sampling
- Stability over time
- Error terms are normally distributed with a mean at 0 and a constant standard deviation across all possible values of the predictors

While the first two assumptions may not be checked at this point, we check the third assumption by checking three things:

- Check the mean and constant variance across fitted values
- Check of Mean and constant variance across all attributes
- Check the normality of the residuals with histogram distribution and Normal QQ plot. The shapiro-wilk test was not conducted because the size of data set is too large.

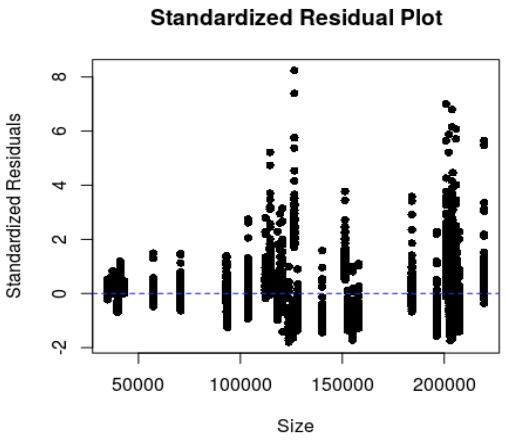
With detailed analysis below, we conclude that error terms are not normally distributed with a constant standard deviation across all possible values of the predictors. This is one limitation of our model and additional analyses should be done in the future to follow up with this issue.

Checking the mean and constant variance across fitted values

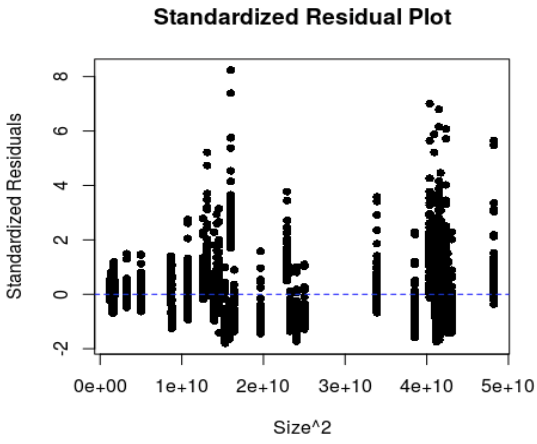


Standardized residuals over fitted values seems to have a mean at zero, but the standard deviation of the residuals seems to be inconsistent.

Checking of Mean and constant variance across all Predictors

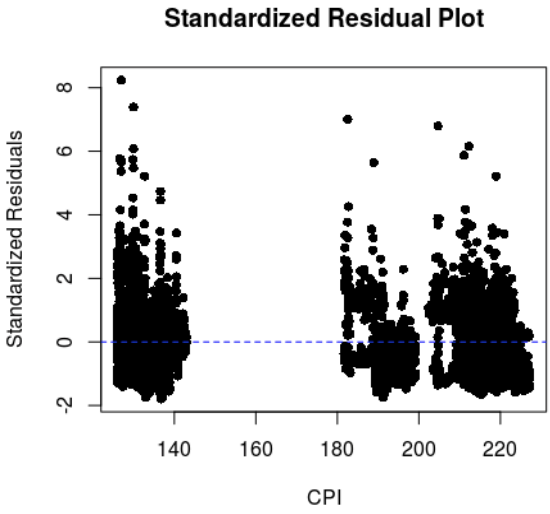


Standardized residuals over Size seems to have a mean at zero, but the standard deviation of the residuals seems to be inconsistent.



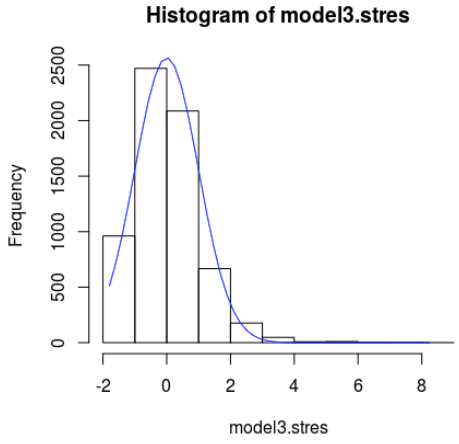
Standardized residuals over Size^2 seems to have a mean at zero, but its standard deviation seems to be inconsistent.

Standardized residuals over Unemployment seems to have a mean at zero and a constant standard deviation across all unemployment values.

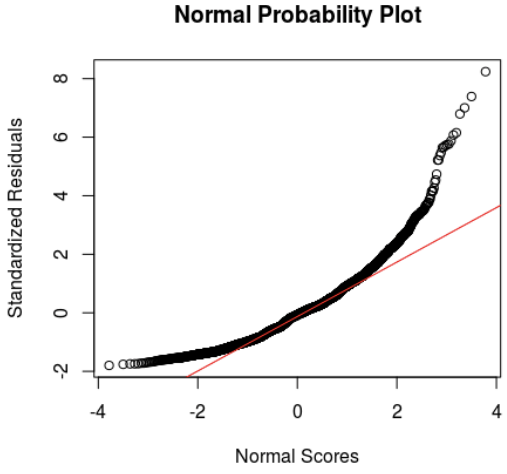


Standardized residuals over CPI seems to have a mean at zero, and a constant standard deviation across all Unemployment values.

Checking normality of the residuals



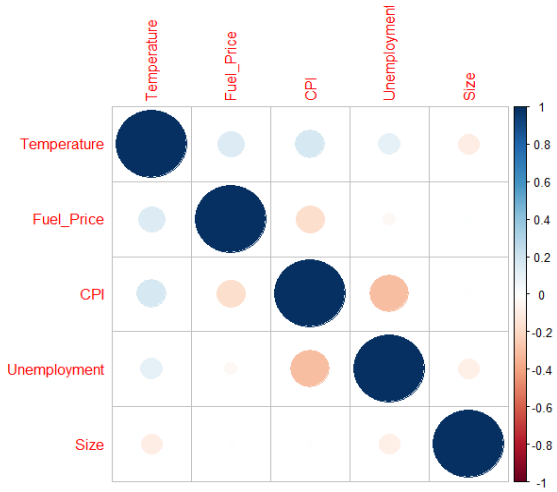
The above graph shows that the residuals have a smaller variance than a typical normal distribution.

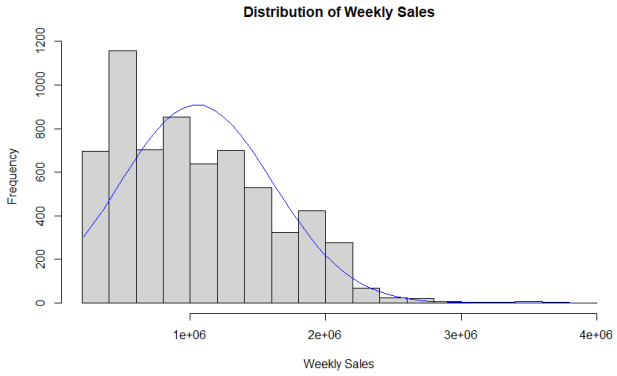


The graph above shows that the residuals have a different distribution from a typical normal distribution.

Other Limitations

A correlation plot of the predictors is plotted to the right. We see that there are no highly correlated variables amongst the predictors as the absolute value of the correlation coefficients is less than 0.2. This caveat is mentioned to express that this is not a limitation to our model, so we are allowed to interpret beta values as shown in the conclusions section.





Another limitation of the model is due to the high right skew in weekly sales as shown in the graph below. While having a highly skewed dependent variable does not violate an assumption, it may make OLS regression inappropriate. OLS regression models the mean weekly sales and the mean is not a good measure of central tendency in a skewed distribution.

Through our exploration of the data, we discovered that store size and store type seem to be related. To test this, we ran an ANOVA between Store Size (as a response) and Store type (as predictor). The results shown below do suggest that there is a relationship between them This would mean that if our model included store type and store size, we would not be able to interpret the beta values.

	Degrees of freedom	Sum Sq	Mean Sq	F value	Pr(>F)
Type	2	1.591e+13	7.953e+12	5260	<2e-16
Residuals	6432	9.725e+12	1.512e+09		