Walmart weekly sales mid-term project

Yingmai Chen

2023-12-10

I.Abstract

The project aims to find the relationship between weekly sales of walmart with other variables,and the project will include six parts:“Abstract”,“Introduction”,“method”,“results”,“discussion”,“appendix”. The first two parts are the descriptions of this project,and the method part would have two parts:EDA(Exploratory Data Analysis)which include some visualization of my data and description of it,and for the model part,it will include the formula of my model and some analysis of the model,I plan to make six model.The results would be the comparison of the model,like among these six model,which one I think is best.Discussion part would discuss what should be improved about the final model I choose,and some future questions about my project,appendix would include some poorly visualized graphs when I do eda,so I don’t want to put it in eda part.

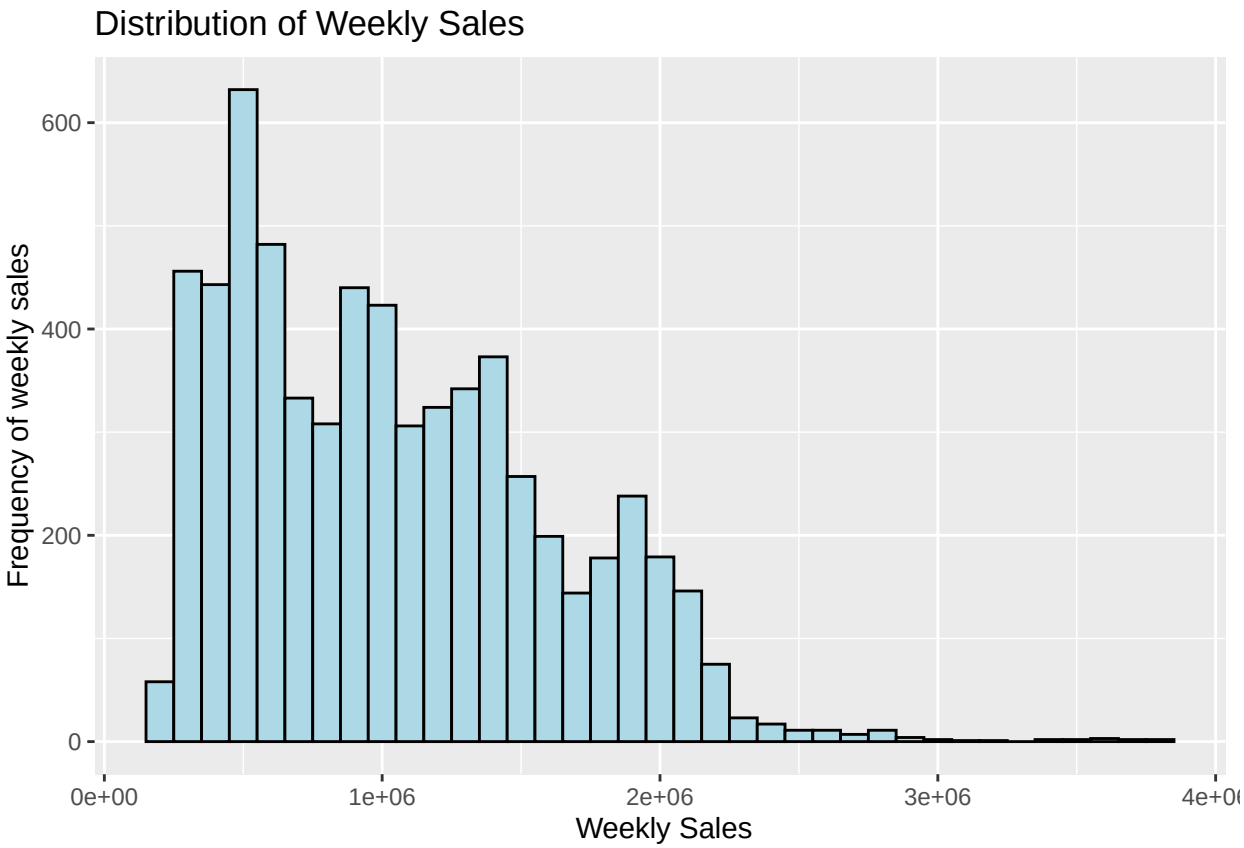
II.Introduction

In this study, we leveraged a dataset from Kaggle to identify and quantify various factors impacting weekly sales. The dataset includes key variables such as holidays, oil prices, temperature, unemployment rates, and the Consumer Price Index (CPI), we utilized regression analysis to assess the relative impact of these variables on sales volumes.Through this study, we hope to offer a more comprehensive and accurate sales forecasting model, helping businesses better understand market dynamics and make more effective business decisions.

III.Method

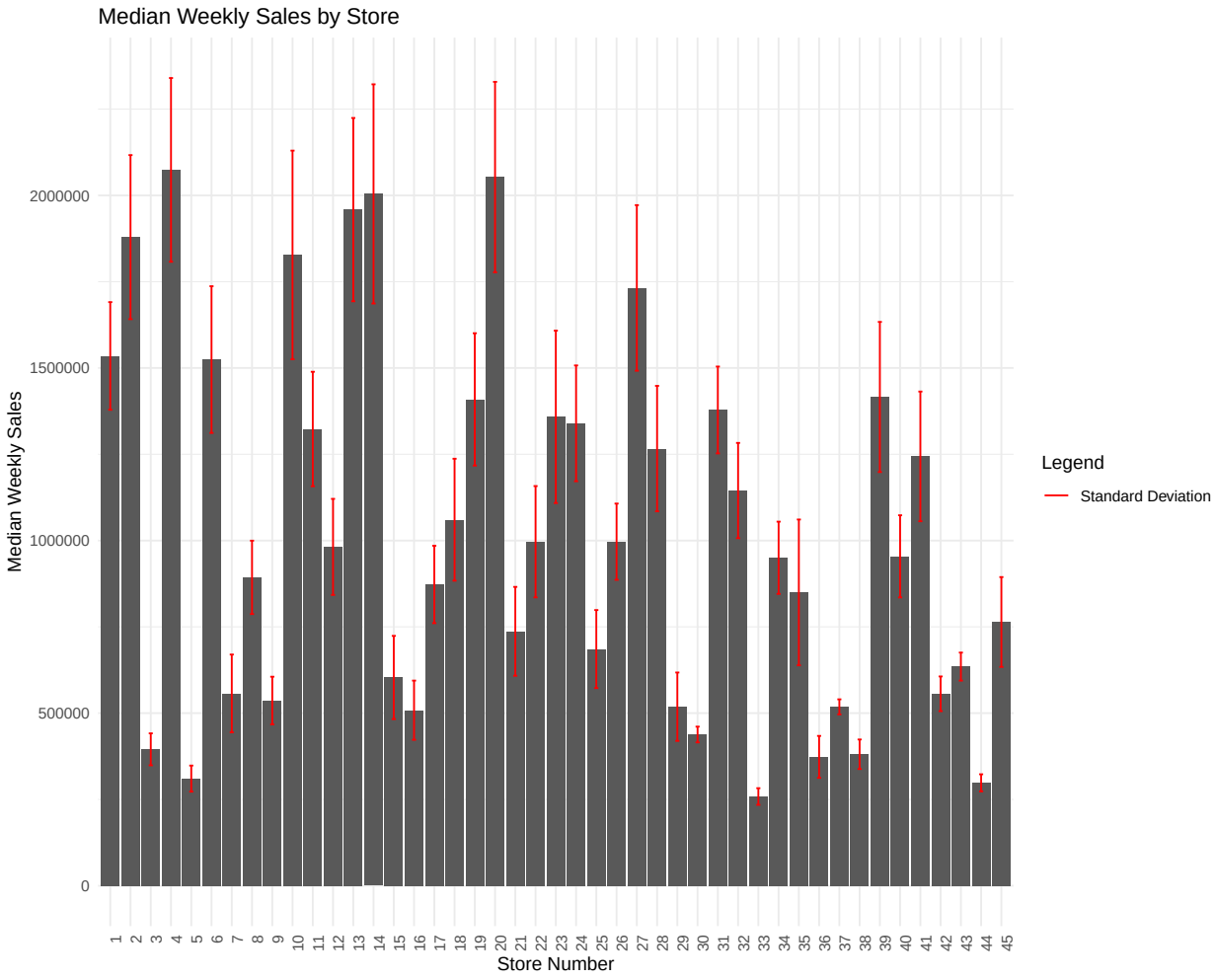
I check the summary of the data first and see there don’t have NA’s in the data,so we don’t need to deal with the missing value of data.

EDA



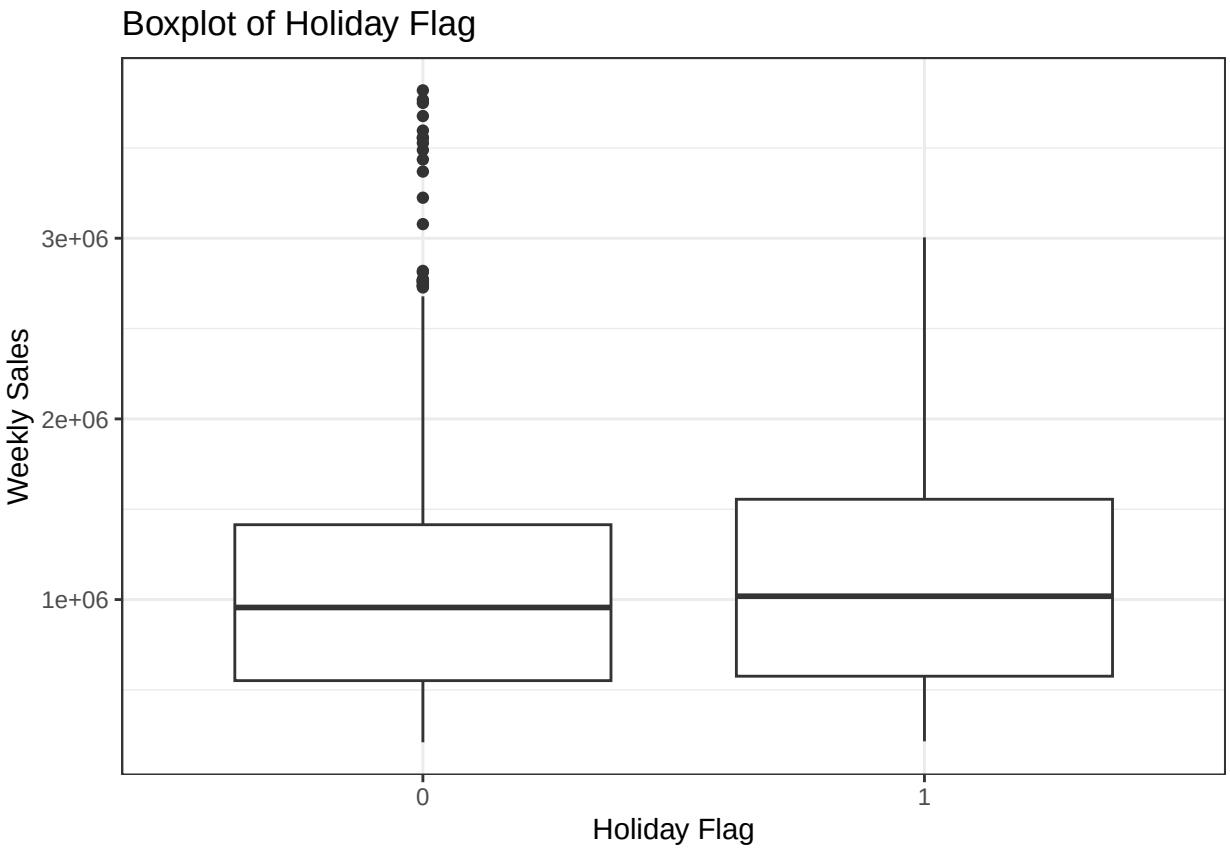
description

From this plot,we can see that there are fewer weeks with very high sales compared to weeks with low sales.This is typical for sales data where a small number of periods (like holiday seasons) might have exceptionally high sales.



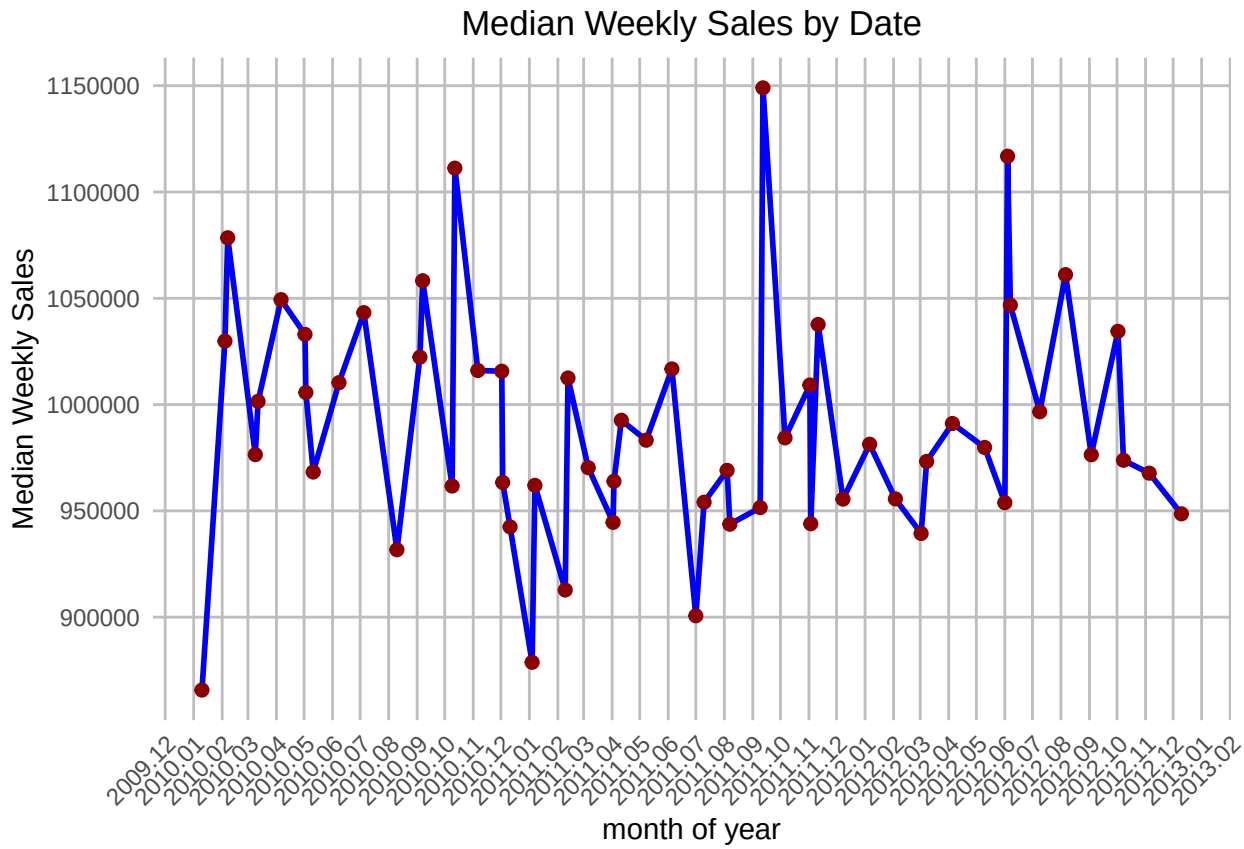
description

From this plot,we can see there exist obvious differences between each store,some has higher median of weekly sales, and some would have bigger error bar which means sales are more volatile.this suggests that we should trying multilevel model in the model part.



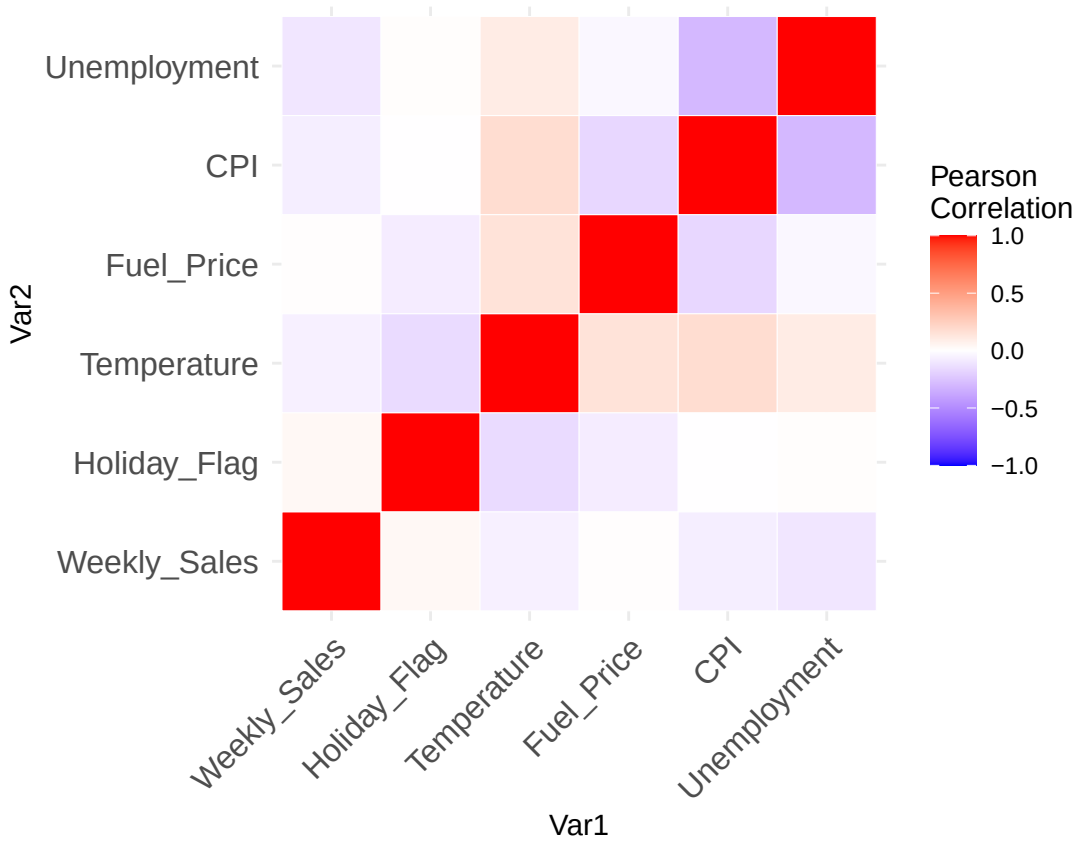
description

From this plot, we can see that the median of holiday weekly sales is a little bit higher than non-holiday weekly sale, and the numbers of outliers in non-holiday weekly sales are more than holiday-weekly sales,based on this,I think there would have some relationships between holidays and weekly sales.



description

From this plot,we can see that there is a recurring trend of lower sales at the beginning of each year. This dip in sales could be attributed to post-holiday season effects, where consumer spending typically drops following the end-of-year holidays,which means the date would have impact on weekly sales.



description

As we can see, from this plot,It is not obvious that these variables are correlated to weekly sales.

Model

Because we can not easily draw a conclusion by doing EDA,so the next step for us shoud be modeling.

model1:null model

The model equation is given by:

$$\text{WeeklySales} = \beta_0 + \varepsilon_i$$

- β_0 is the intercept, which represents the average sales over all observations in the data.
- ε_i is the error term for the i-th observation.

analysis of model1 Intercept: The estimated intercept is 1,046,965 with a standard error of 7,035. This intercept represents the average weekly sales across all stores and dates included in the dataset.

Residuals: The residuals' range from -836,979 to 277,1722, with the interquartile range (IQR) from -493,615 (1st quartile) to 37,194 (3rd quartile). The large range of residuals indicates there is considerable variability

in weekly sales that the null model (using only the mean) does not capture. The median of the residuals is -86,219, suggesting that the model may systematically overestimate the weekly sales.

Mean Squared Error (MSE):The provided image shows the Mean Squared Error (MSE) of the null model, which is 318460187634.It illustrated that The null model's high MSE indicates it fails to accurately predict weekly sales, lacking explanatory variables. To improve, incorporating factors like promotions, store attributes, and seasonality into a more complex model is essential.

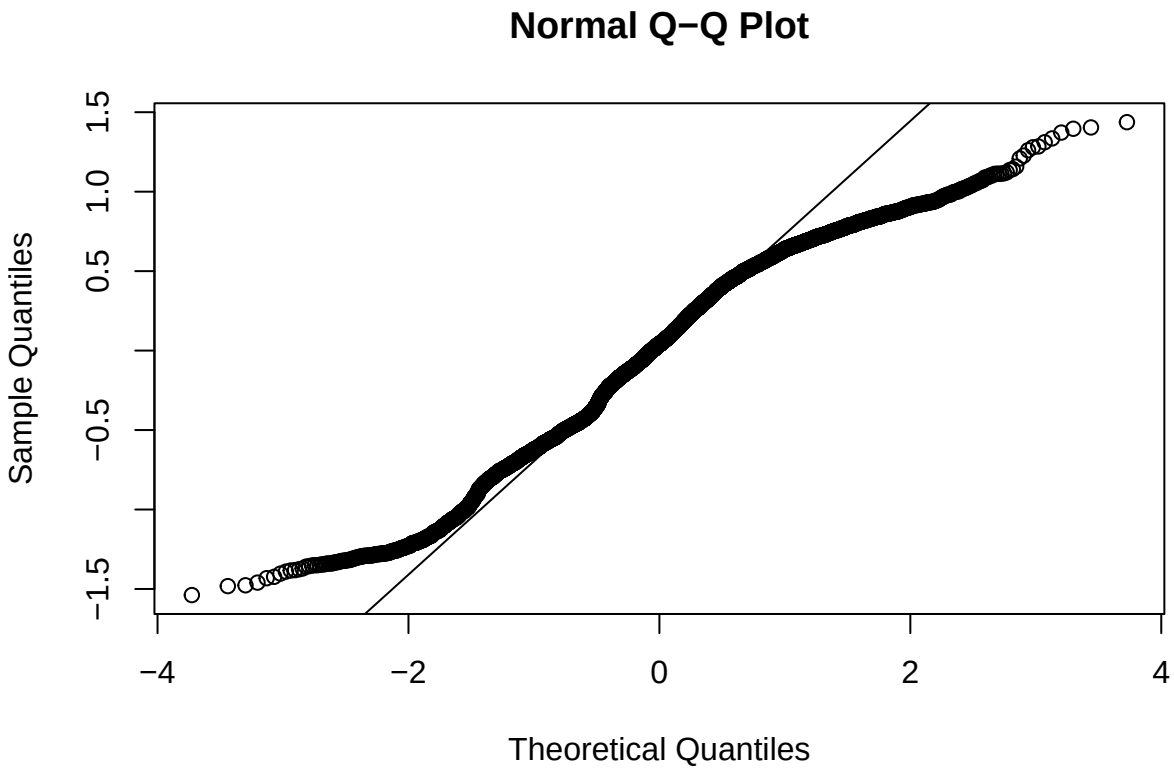
model2:simple linear model

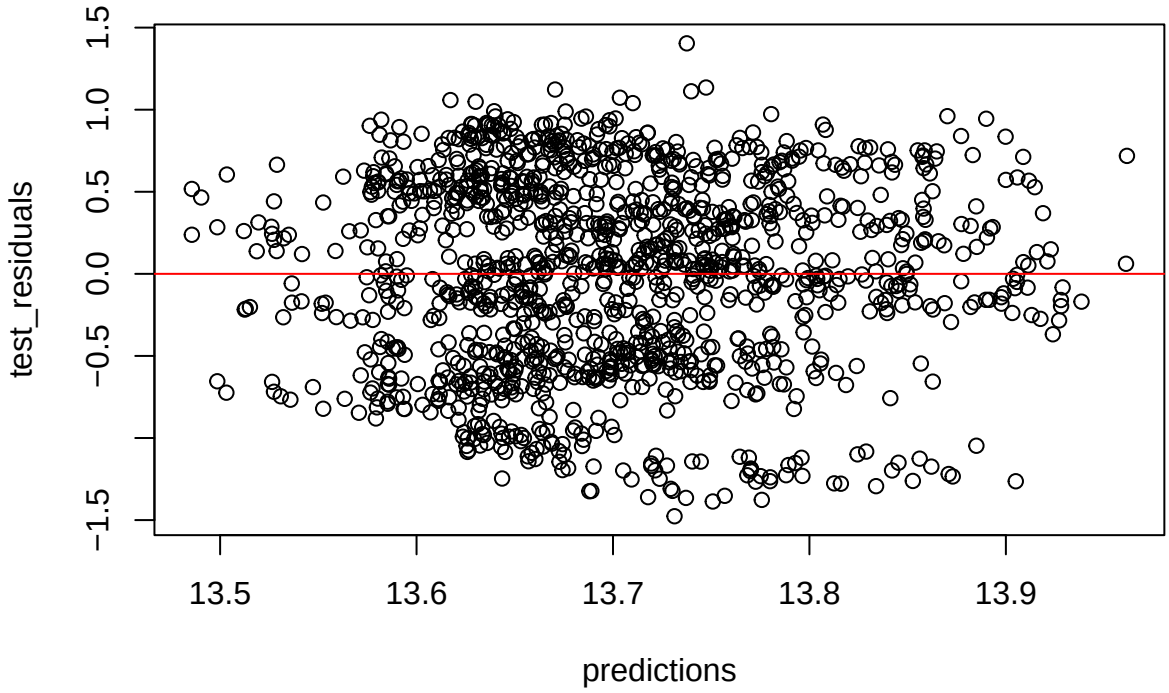
The model equation is given by:

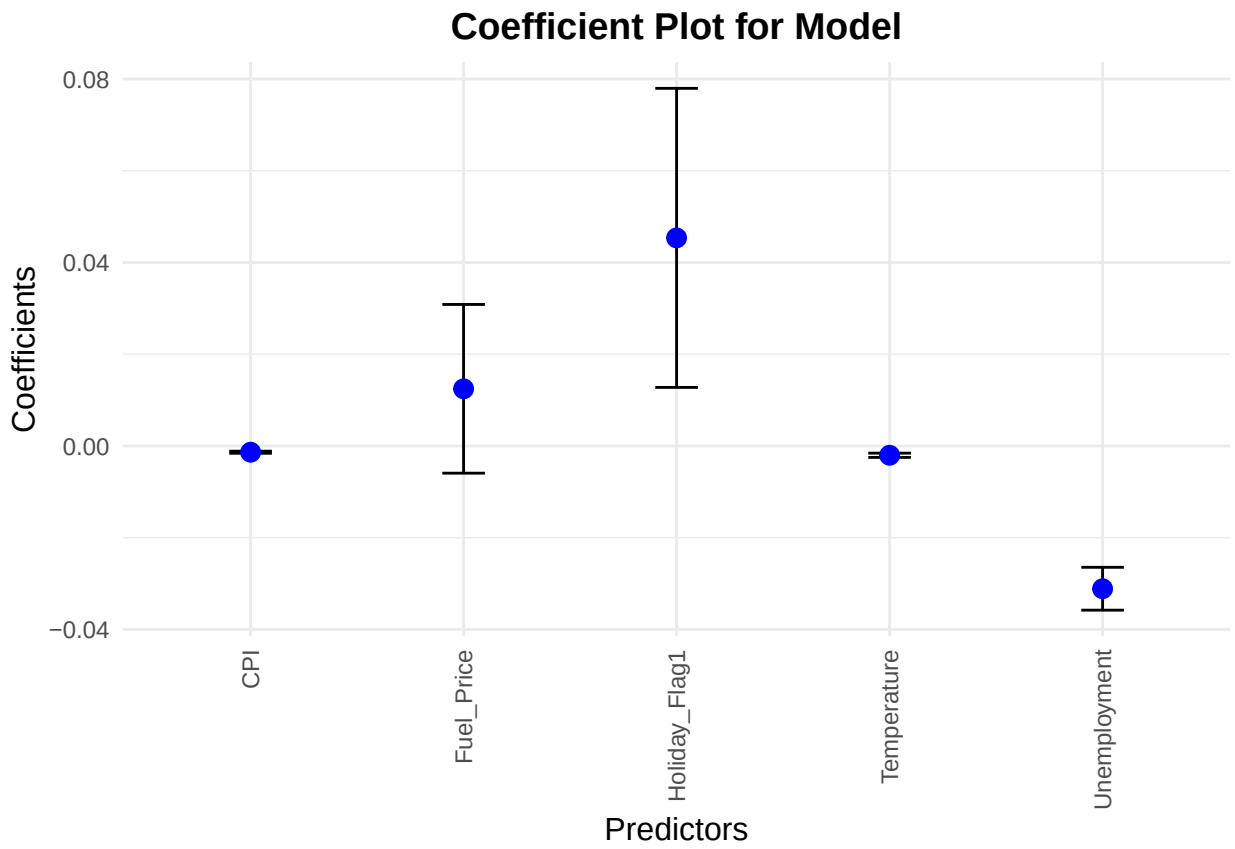
$$\log(\text{Weekly\_Sales}) = \beta_0 + \beta_1 \text{Holiday\_Flag} + \beta_2 \text{Temperature} + \beta_3 \text{Fuel\_Price} + \beta_4 \text{CPI} + \beta_5 \text{Unemployment} + \varepsilon$$

- β_0 is the intercept.
- β_1 is the coefficient for the variable `Holiday_Flag`.
- β_2 is the coefficient for the variable `Temperature`.
- β_3 is the coefficient for the variable `Fuel_Price`.
- β_4 is the coefficient for the variable `CPI`.
- β_5 is the coefficient for the variable `Unemployment`.
- ε represents the error term.

analysis of model2




test_residuals
1.5
1.0
0.5
0.0
−0.5
−1.5
13.5
13.6
13.7
13.8
13.9
predictions



The model has a very low R-squared value, indicating it does not explain much of the variation in the data.Some predictors are statistically significant, but given the low R-squared, their practical impact might be limited. The assumptions of normality and homoscedasticity may be violated based on the Q-Q plot and the Residuals vs. Predictions plot, which can affect the reliability of the model's standard errors and p-values.There are outliers or extreme values that might be influencing the results significantly.So all in all,simple linear model seems not to be a good model to fit the data.Besides,we can see that fuel price,holiday flag and unemployment are statistical significant.But temperatures and CPI are not

model3:multilevel linear model with random intercept(partial pooling)

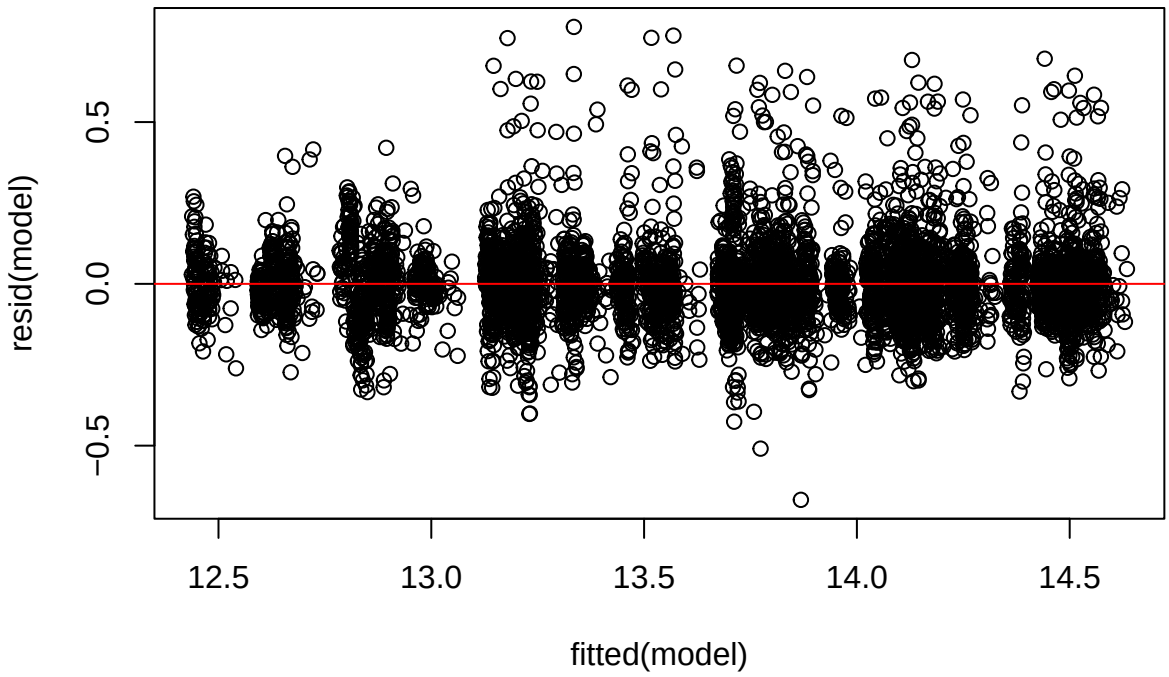
The model equation is given by:

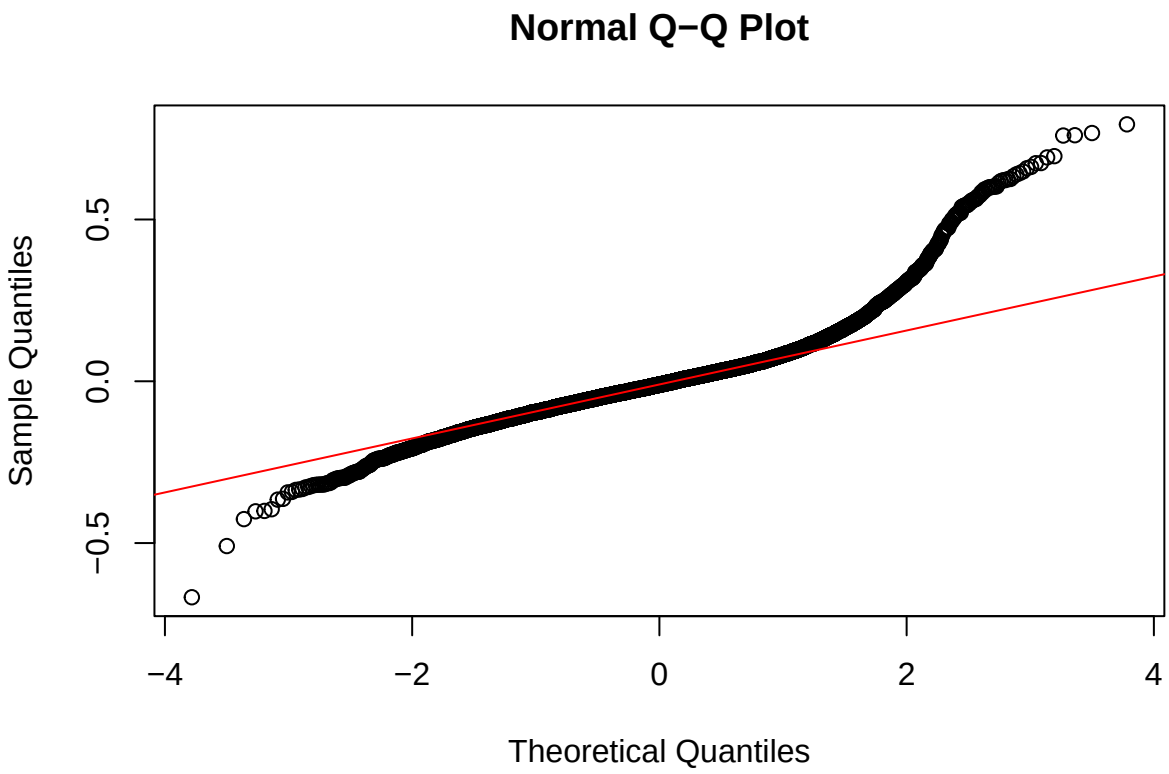
$$\log(\text{Weekly\_Sales}) = \beta_0+\beta_1\text{Temperature}+\beta_2\text{Fuel\_Price}+\beta_3\text{CPI}+\beta_4\text{Unemployment}+\beta_5\text{Holiday\_Flag}+u_j+\epsilon_i$$

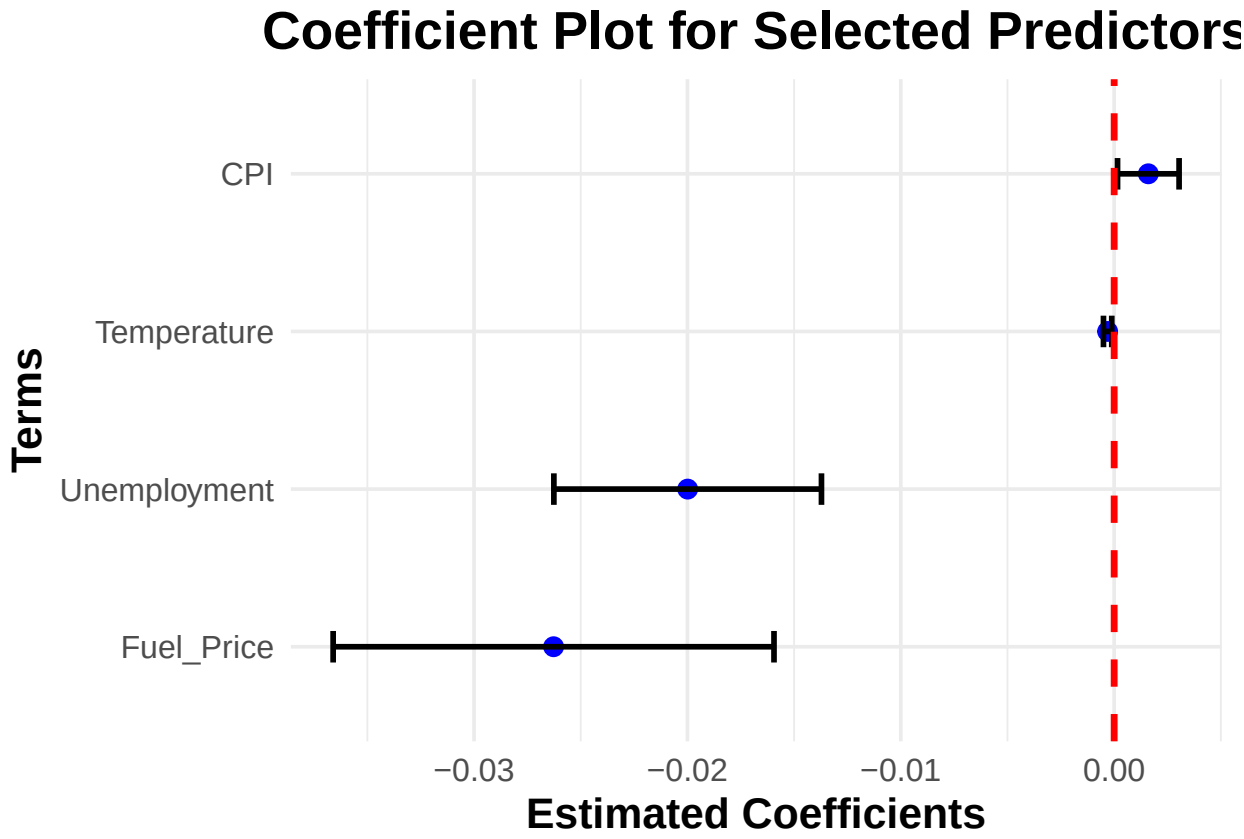
- β_0 is the intercept term, representing the expected value of the log sales when all other explanatory variables are zero.
- u_j is the random effects term, capturing the effects introduced by different stores (store j). This means each store has its own specific impact (like location, size, customer base, etc.) that is not shared with other stores.
- ϵ_i is the error term.

So in this model,I will use the random value of intercept of store to distinguish the different store.

analysis of model3




Normal Q−Q Plot
Sample Quantiles
0.5
0.0
−0.5
−4
−2
0
2
4
Theoretical Quantiles



Model Fit:

The conditional R-squared is extremely high, suggesting that when accounting for both fixed and random effects, the model explains a large proportion of the variance in sales.

The marginal R-squared is quite low, indicating that without the random effects, the fixed effects do not explain much of the variance in sales.

Residual Diagnostics:

The residual plot does not show obvious patterns, which suggests that the model does not suffer from non-linearity or heteroscedasticity issues.

The QQ plot indicates deviations from normality, particularly in the tails. This could point to issues with outlier responses or indicate that the normal distribution may not be the best fit for the residuals.

From coefficient plot,we can see except for CPI and temperature,other variables are statistical significant.

In conclusion, while appears to be a good fit for capturing store-to-store variation in sales, the fixed effects alone have limited explanatory power.

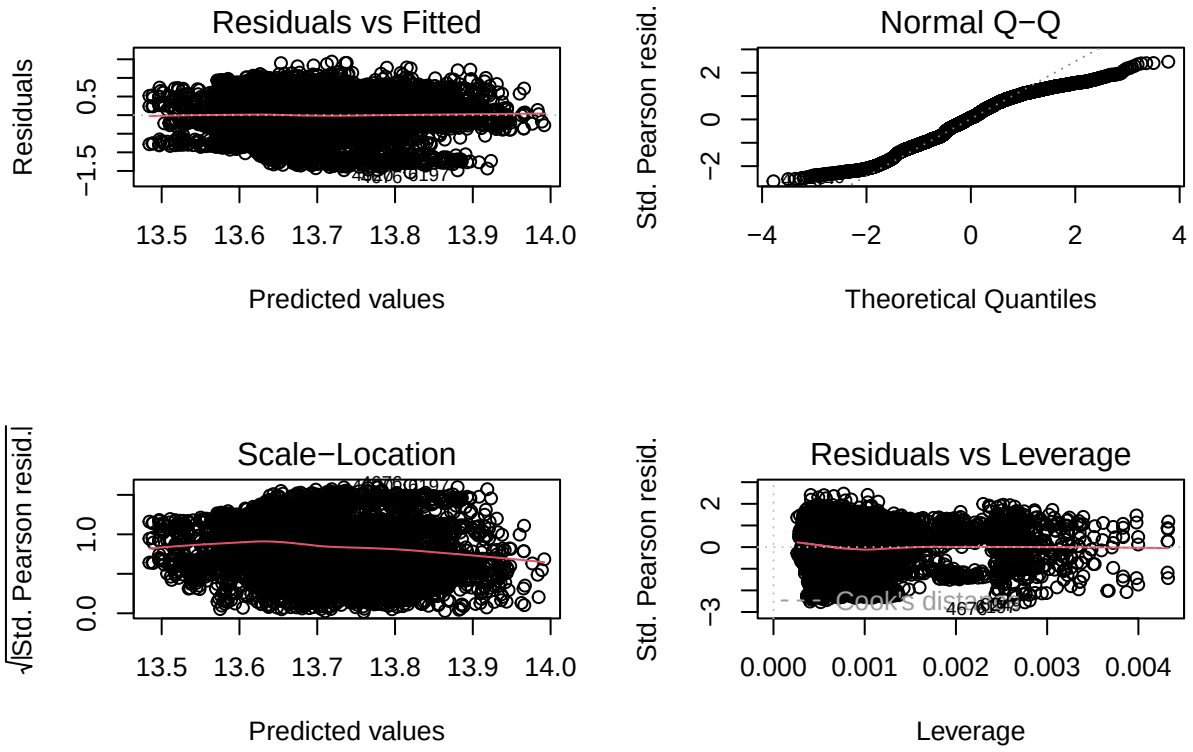
model4:generalized linear Model

The model equation is given by:

$$\log(\text{Weekly\_Sales}) = \beta_0 + \beta_1 \text{Holiday\_Flag} + \beta_2 \text{Temperature} + \beta_3 \text{Fuel\_Price} + \beta_4 \text{CPI} + \beta_5 \text{Unemployment}$$

- β_0 is the intercept term.
- β_1 is the coefficient for the variable `Holiday_Flag`.
- β_2 is the coefficient for the variable `Temperature`.
- β_3 is the coefficient for the variable `Fuel_Price`.
- β_4 is the coefficient for the variable `CPI`.
- β_5 is the coefficient for the variable `Unemployment`.

analysis of model4



After doing many analysis of this data,we find that it also shows a poor fit with the data,Neither residual plot nor qq plot all shows it would not fit to the data,so I decide not to do the no pooling anymore,I want to try to improve model3 to make it fit better to the data.Besides,for the summary of the model,we can see temperature,CPI and unemployment are statistical significant,it is different with other models.

model5:multilevel generalized linear Model with random intercept(partial pooling)

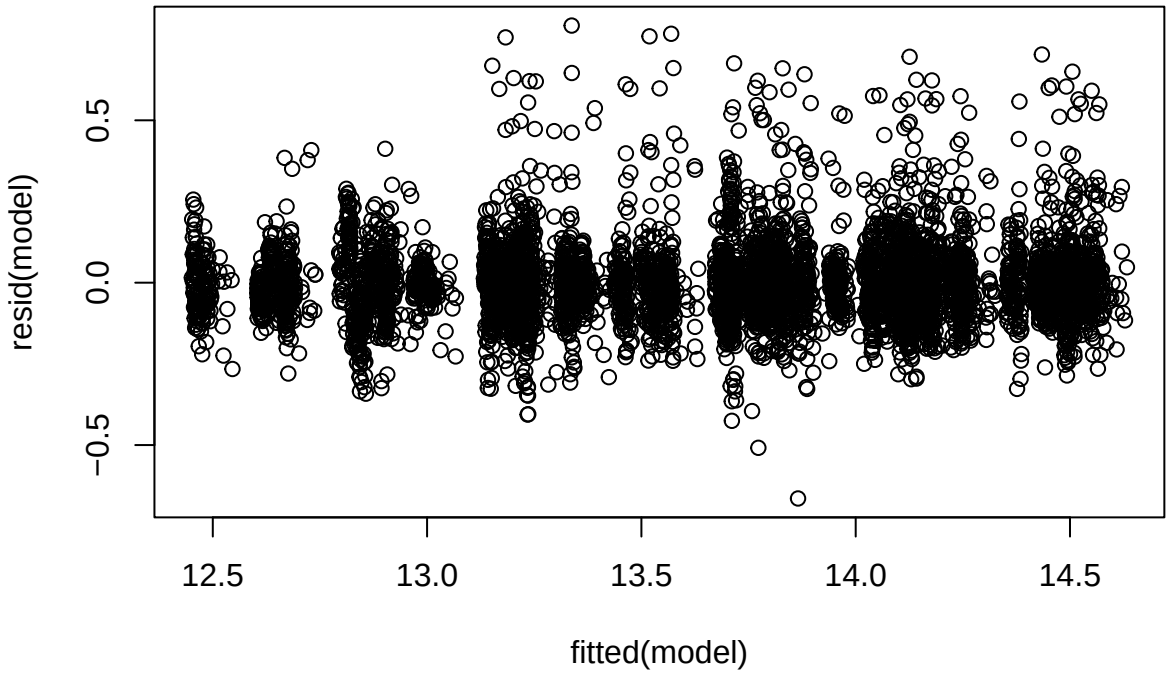
The model equation is given by:

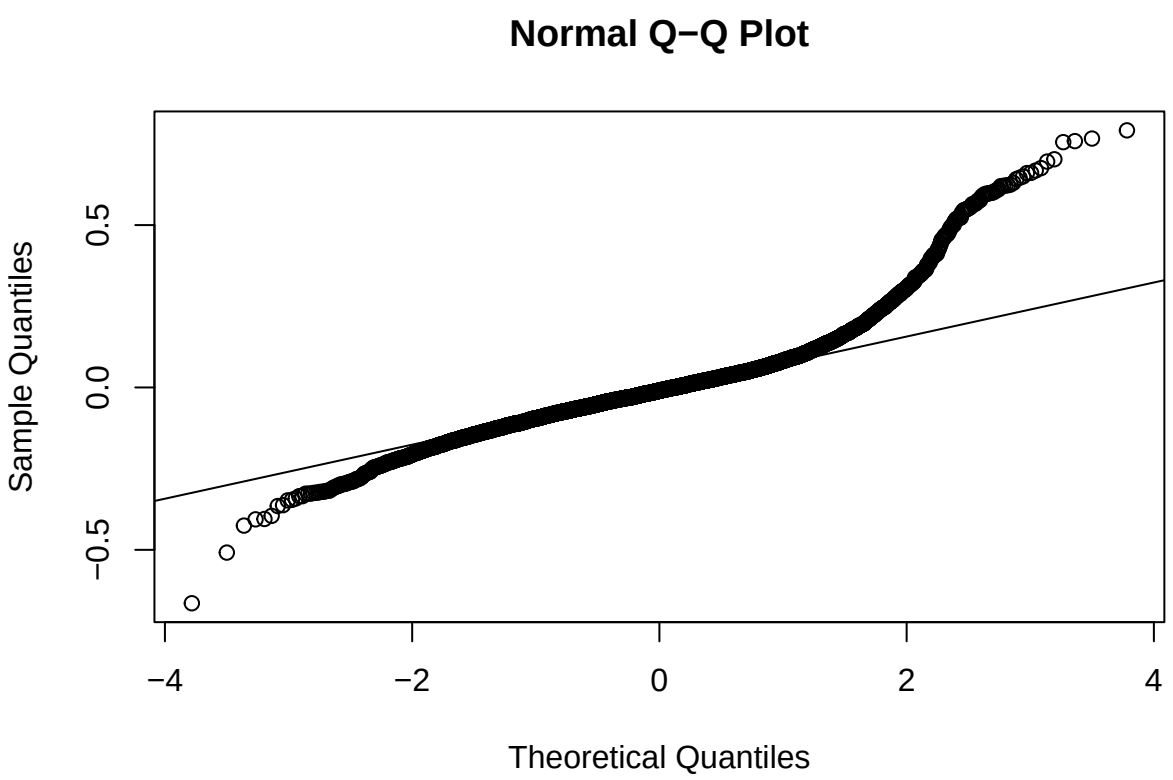
$$\log(\text{Weekly\_Sales}) = \beta_0 + \beta_1 \text{Temperature} + \beta_2 \text{Fuel\_Price} + \beta_3 \text{CPI} + \beta_4 \text{Unemployment} + \beta_5 \text{Holiday\_Flag} + u_j + \epsilon_i$$

- β_0 is the intercept term, representing the expected value of the log sales when all other explanatory variables are zero.

- u_j is the random effects term, capturing the effects introduced by different stores (store j). This means each store has its own specific impact (like location, size, customer base, etc.) that is not shared with other stores.
- ϵ_i is the error term.

analysis of model5





From the qqplot ,we can see a deviation from the line in the tails, suggesting that the residuals may not be normally distributed, especially in the extremes. This could affect the validity of statistical tests and confidence intervals.From the residual plot,we can see a clear pattern of residuals, which is not ideal. We would prefer a random scatter. The banding pattern could be due to discrete features in the data or overdispersion.Besides,from summary of model,we can see except for CPI and temperature,other variables are statistical significant.So it seems this model may not be a good choice.

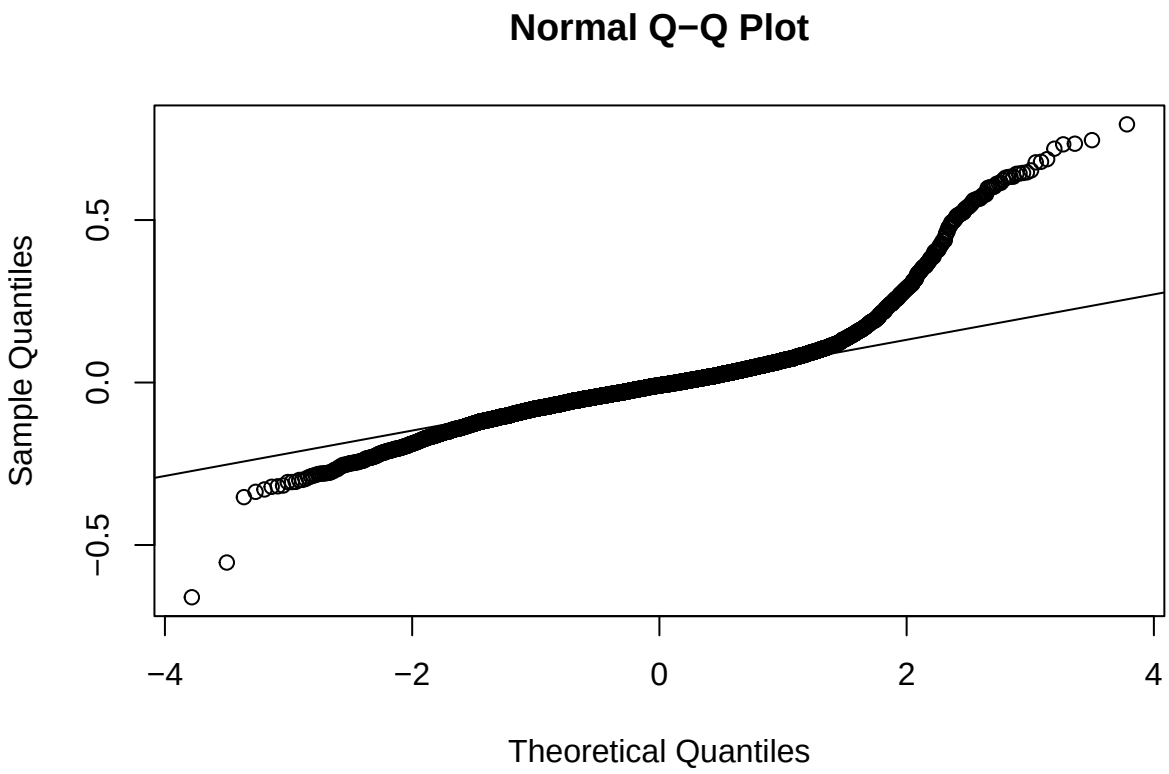
model6:multilevel linear Model with random coefficient and random intercept (partial pooling)

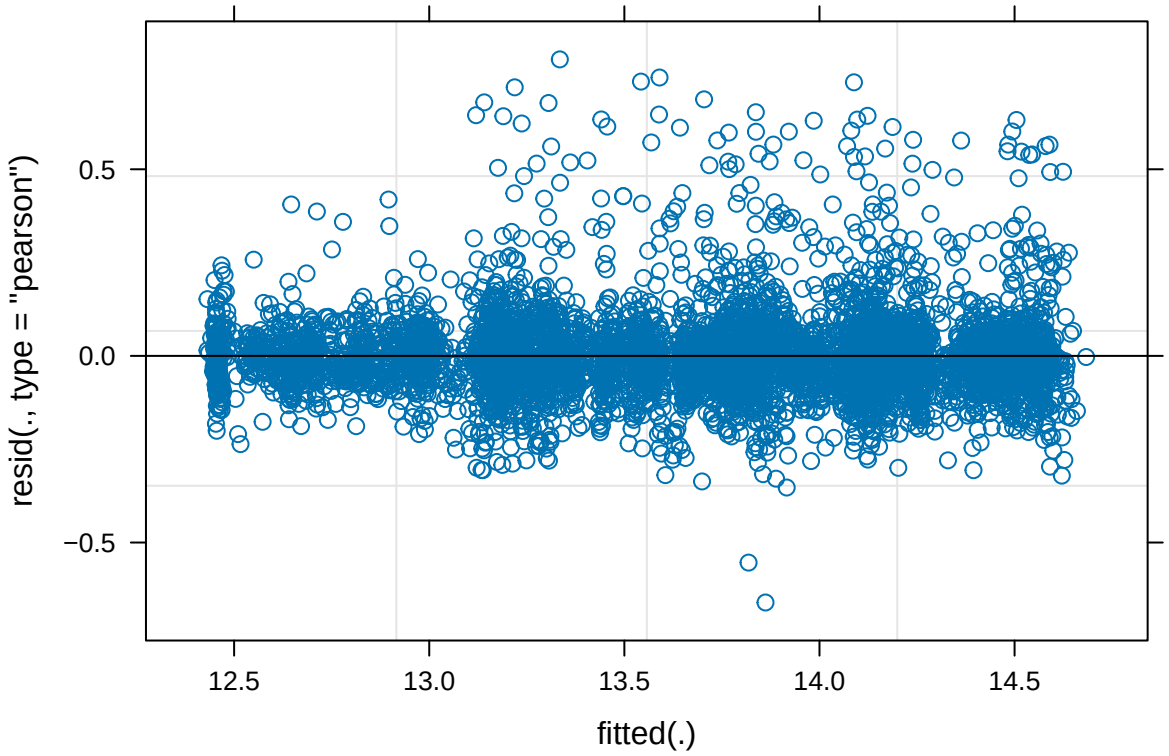
The model equation is given by:

$$\log(\text{Weekly\_Sales}) = \alpha_i + \beta_{1[i]}\text{Temperature} + \beta_{2[i]}\text{Fuel\_Price} + \beta_{3[i]}\text{CPI} + \beta_{4[i]}\text{Unemployment} + \beta_{5[i]}\text{Holiday\_flag} + \epsilon_i$$

- α_i is the random intercept for each store.
- $\beta_{1[i]}$ represents the random coefficient for the variable `Temperature` for store i.
- $\beta_{2[i]}$ represents the random coefficient for the variable `Fuel_Price` for store i.
- $\beta_{3[i]}$ represents the random coefficient for the variable `CPI` for store i.
- $\beta_{4[i]}$ represents the random coefficient for the variable `Unemployment` for store i.
- $\beta_{5[i]}$ represents the random coefficient for the variable `Holiday_Flag` for store i.
- ϵ_i is the error term.

analysis of model6


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
0.5
0.0
−0.5
−4
−2
0
2
4



From this qqplot,we can see some deviation from normality, especially in the tails. This could affect the reliability of the p-values associated with the fixed effects.From the residual plot, we can see it shows a fairly random dispersion,besides,we also see the summary of the model,it shows that the intercept and all predictors except for Temperature have significant p-values, suggesting they contribute meaningfully to the model.so based on these outputs, the model appears to be decent.

IV.result

I have mentioned in the abstract part,for the result part,what I want to do is to choose from the six model I used to judge which one is the best model,So what I do are as follows:

comparison

Mean square error

First,I exclude model5,it seems that the mse of model5 is much bigger than others.

From mean square error part,we can see clearly that model3 and model6 are better,in mse part,it seems model6 is better than model3,but to be honest,model6 is more complex than model3,so to compare those two models,I also use anova test to compare those two models.

Table 1: MSE of Various Models

Model	MSE
Model 1	0.3460442
Model 2	0.3383922
Model 3	0.0144211
Model 4	0.3383922
Model 6	0.0119610

AIC and BIC

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
model3	8	-8590.185	-8536.029	4303.093	-8606.185	NA	NA	NA
model6	28	-9356.132	-9166.585	4706.066	-9412.132	805.9461	20	0

model3 is multilevel linear Model with random intercept,model6 is multilevel linear Model with random coefficient and random intercept.

AIC and BIC: model6 has the lowest AIC and BIC values, suggesting that it fits the data best while considering model complexity. AIC and BIC are fit indices that penalize for complexity, with lower values typically indicating a model that fits well with fewer parameters.

Log-Likelihood: model6 has the highest log-likelihood value, which indicates the most accurate fit to the data.

Significance Testing: The p-value for the Chisq statistic is well below 0.001, indicating that model6 represents a statistically significant improvement over the other models.

Combining these insights, model6 appears to be better than model3. It not only provides the best fit to the data (based on AIC, BIC, and log-likelihood) but also accounts for model complexity in a penalized manner. Although this model is more complex (as it includes random slopes), its ability to improve the fit seems to outweigh the costs of increased complexity. Therefore, if the data supports a more complex model structure and predictive ability is an important consideration, model16 would be the better choice.

V.discussion

For the discussion part,I have mentioned in the abstract part,I will point out some limitation of multilevel linear Model with random coefficient and random intercept(model I choose) and future questions.

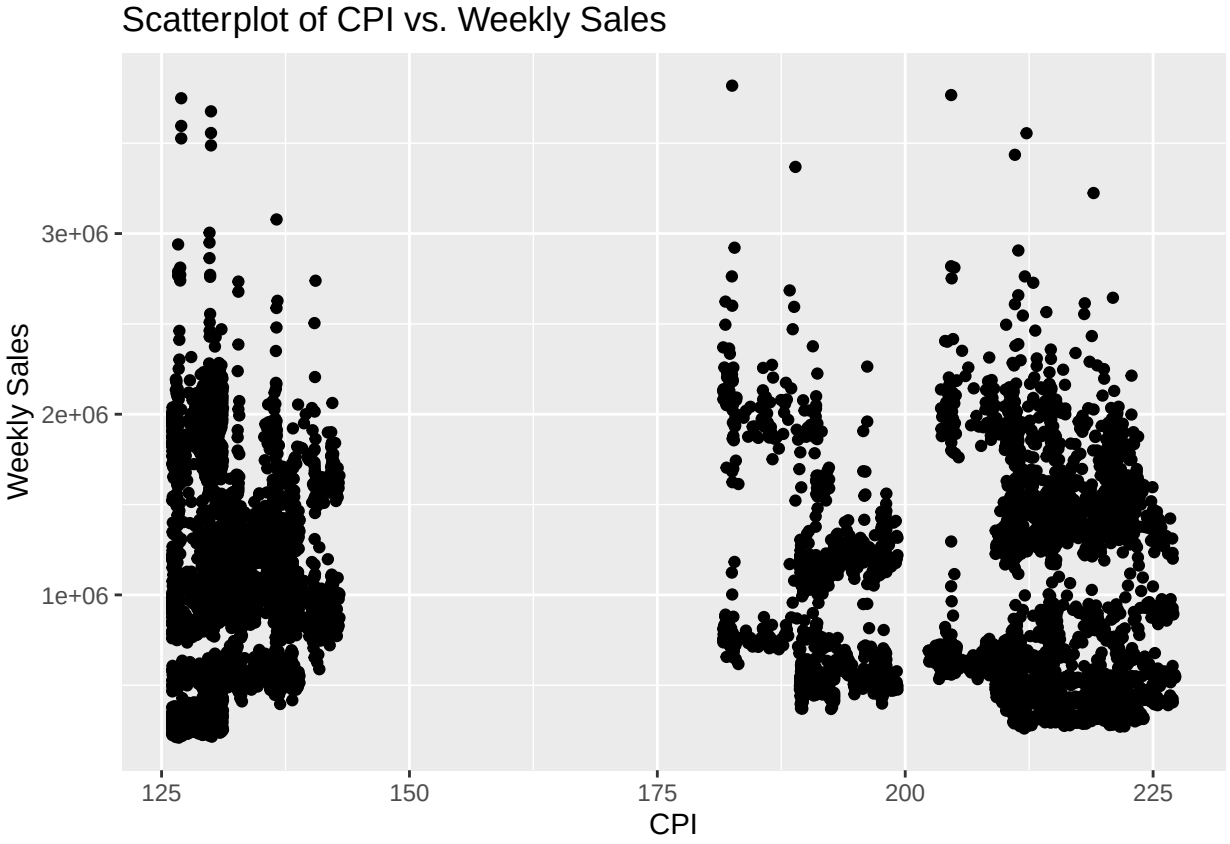
limitation

For the predictor Significance temperature is insignificance ($p > 0.05$),which suggests it may not be an important predictor for `Log_Weekly_Sales` within this model, potentially limiting the model's explanatory power regarding the influence of temperature on sales. For the qqplot part, deviations from normality, particularly in the tails as shown by the Q-Q plot, question the reliability of the model's p-values and confidence intervals, impacting the validity of statistical inferences.Foe r the residual plot,the funnel pattern observed in the residuals versus fitted values plot indicates the presence of heteroscedasticity, which means that the error variance is not constant. This can affect the efficiency of the estimators and the predictive performance of the model.

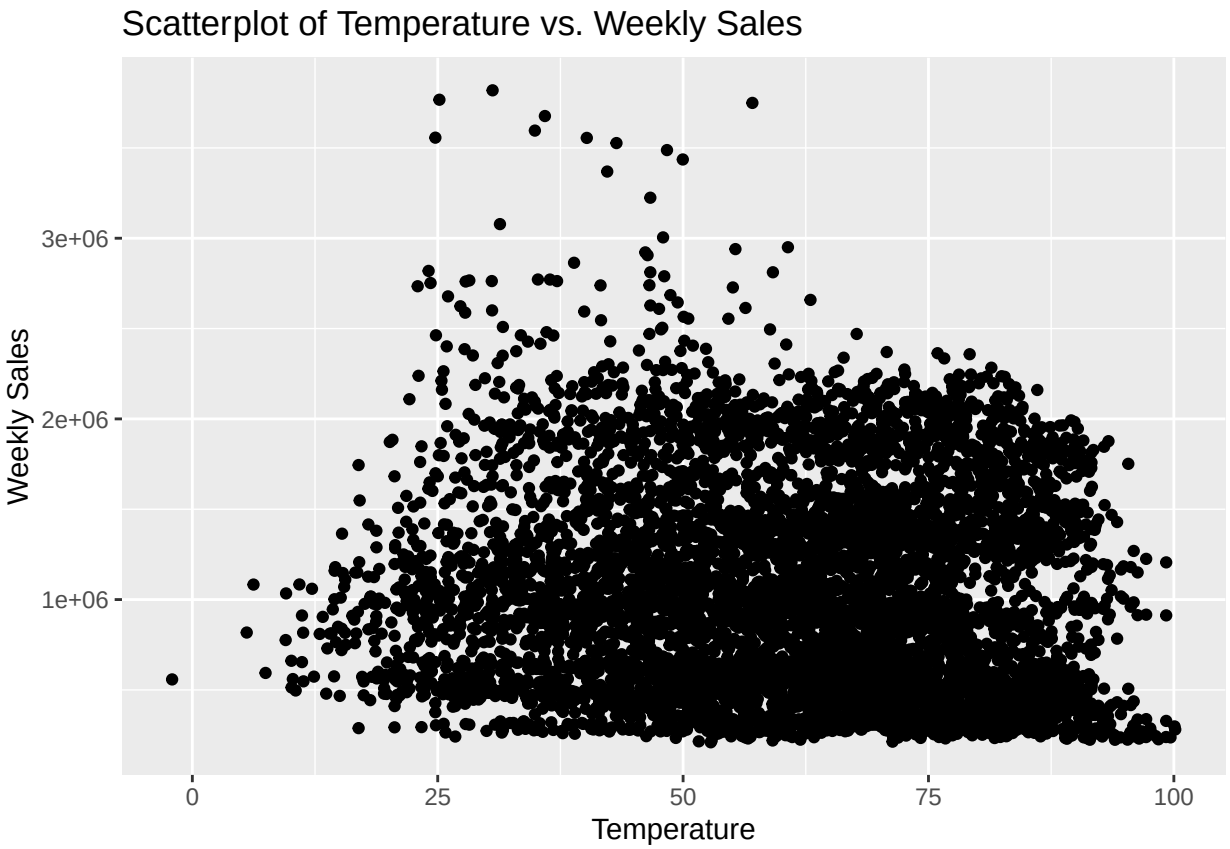
future questions

For the future question,I think first we should further investigate the Temperature variable to see if there is a nonlinear relationship or interactions with other variables that are not captured in the current model. And we should try to explore more complex random effects structures to better capture the variability in the data, such as including random slopes for certain predictors.

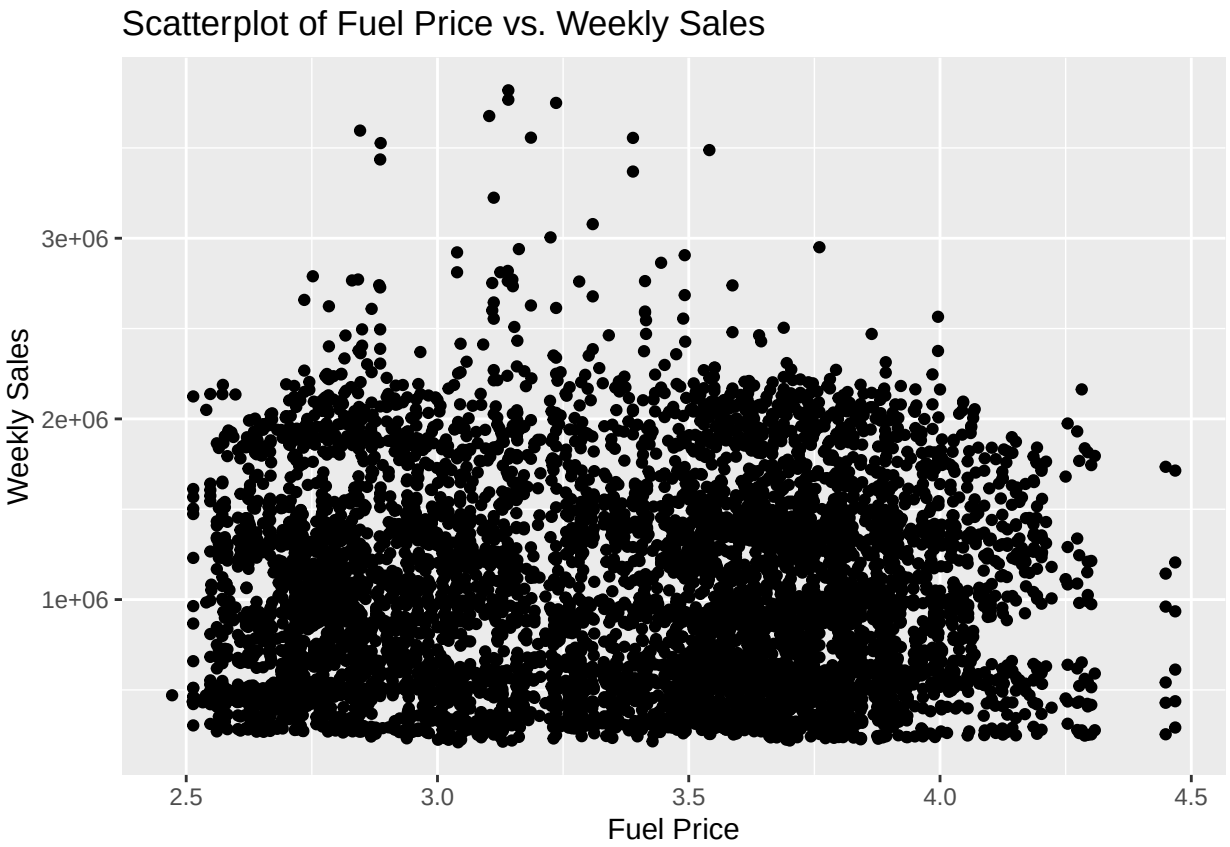
VI.appendix



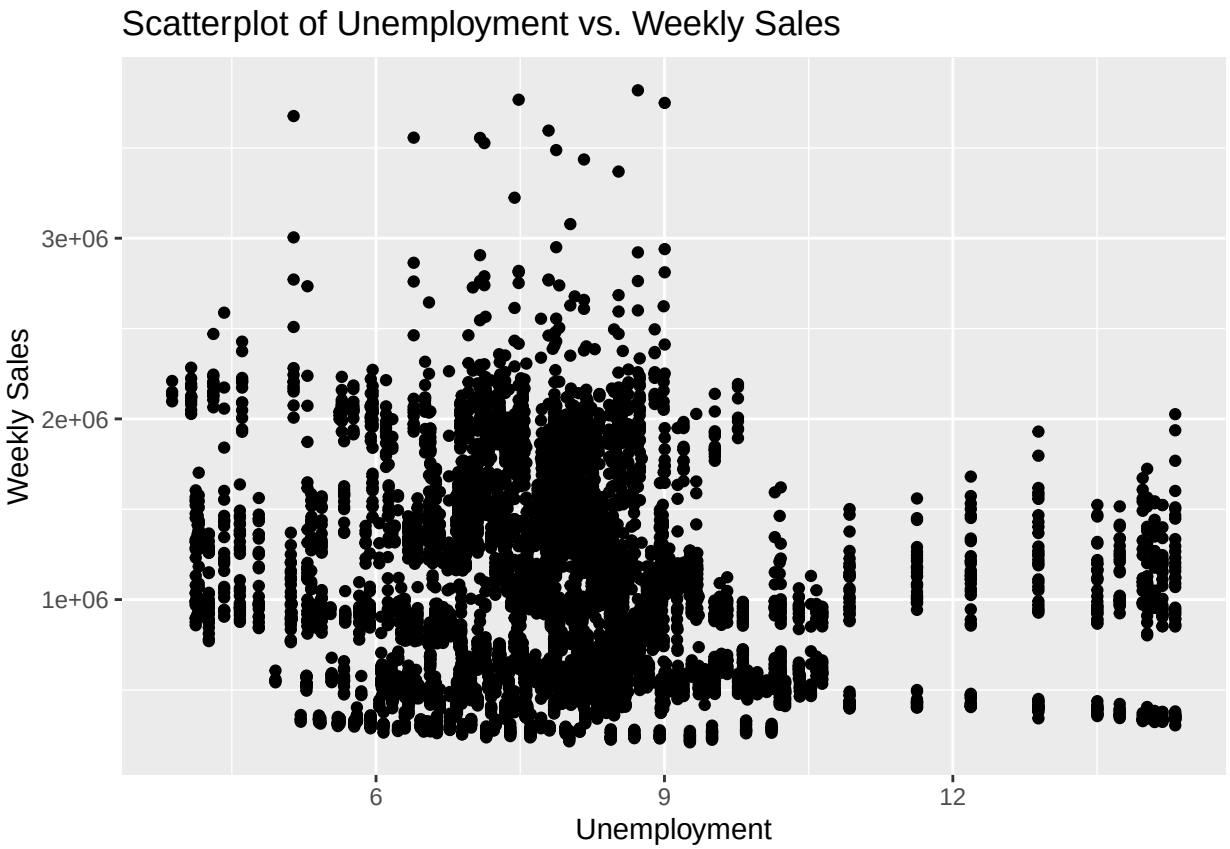
From this plot, we can hardly see if there exist relationships between CPI and weekly sales.



From this plot,we can see temperatures from 25 to 75 of the country tend to have much more numbers of higher weekly sales.



From this plot, we can hardly see if there exist relationships between fuel price and weekly sales.



From this plot,we can see the less of proportions of unemployment tends to have higher weekly sales.