

Business

Intelligence



INTRODUCTION

Walmart, one of the top retailers in the US, would want to make accurate sales and demand predictions. Each day, specific occasions and holidays have an influence on sales. Data on sales are available for 45 Walmart locations. Unforeseen demand is a hurdle for the company, and occasionally supply runs out because of a bad machine learning system. An ideal machine learning algorithm will forecast demand at various times in time, accounting for seasonality, and will take into account variables like the CPI, unemployment rate, and other economic situations. Every year, Walmart holds a number of promotional discount sales. The four biggest holidays—the Super Bowl, Labor Day, Christmas, and Thanksgiving—are preceded by these price reductions. These occurrences frequently impact consumer demand

Our project, "Retail Analysis with Walmart Data," seeks to resolve this issue by creating a strong machine learning algorithm that can precisely forecast sales and demand for Walmart's 45 stores while accounting for a variety of variables, such as seasonality, the state of the economy, and promotional markdown events. We will build a thorough prediction model that can offer insights and projections at various time intervals by utilising past sales data as well as external elements like the Consumer Price Index (CPI), Unemployment Index, and holiday events.

To find patterns, trends, and correlations in the data, I will use advanced analytics approaches, including time series analysis, regression analysis, and machine learning algorithms, utilising SAS University Edition and the R programming language. To fully comprehend the data and locate pertinent elements that influence sales and demand, I will also make use of data visualisation tools.

Walmart will be able to make educated judgments regarding inventory control, pricing, and promotions thanks to the project's results, which will eventually enhance consumer happiness and boost sales. Additionally, the project's findings will offer helpful insights for other shops and enterprises dealing with comparable problems with demand forecasting and inventory management.

**Notable Holiday Events**

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13  
Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13  
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13  
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

**MIS DESIGN**

The creation of a thorough reporting and visualisation system that can offer in-the-moment insights into an organization's data is part of a management information system (MIS) design for data analytics. The architecture entails combining data sources from many organisational departments and systems into a single database or data warehouse.

The MIS design need to include an intuitive user interface that enables users to engage with the data through filters, dashboards, and reports in order to discover insights that can be used to influence decision-making. To guarantee the data is correct, dependable, and current, the MIS design should also include capabilities for data cleansing, data transformation, and data modelling.

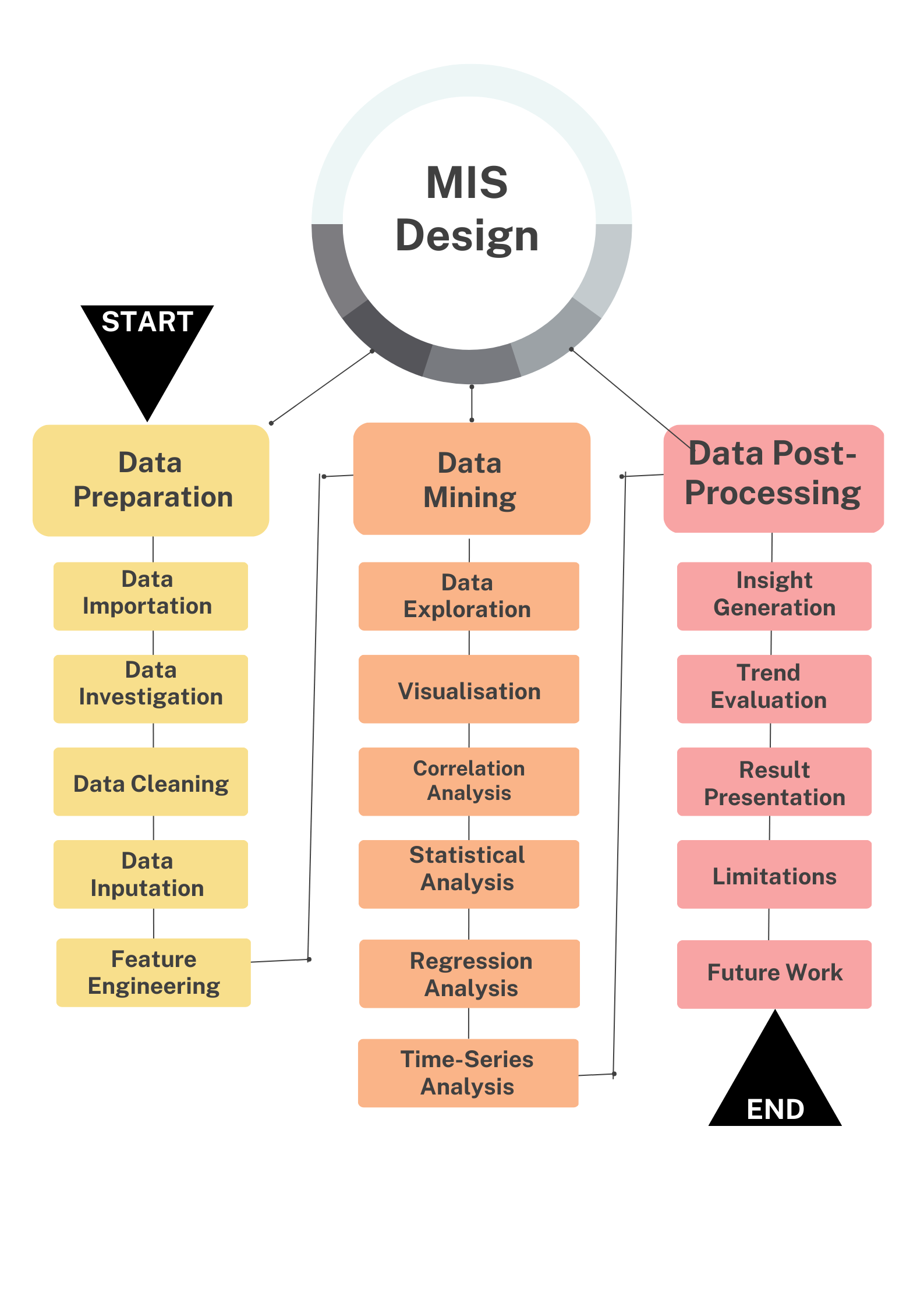
A Management Information System (MIS) design for this project incorporates the following elements:

Data Collection: During this phase, raw data are gathered from a variety of sources, including databases, questionnaires, and other pertinent sources. To guarantee correctness and consistency, the data should be vetted and confirmed.

Data Preprocessing: To prepare the data for analysis, this stage entails cleansing and transformation. Activities including data integration, data minimization, and data transformation are included.

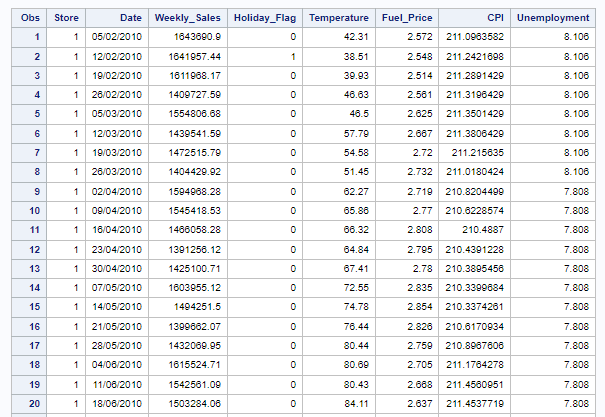
Data Analysis: In order to find patterns, trends, and insights, statistical methods and algorithms are applied to the preprocessed data in this stage. It covers procedures like hypothesis testing, data modelling, and exploratory data analysis.

Interpretation and Reporting: In this phase, data analysis findings are interpreted and presented to stakeholders via reports, dashboards, and other visualisations. It is important to evaluate and publish data in a way that is both clear and useful.



**Understanding the Dataset**

The file Walmart Store sales contains historical data on sales from 2010-02-05 through 2012-11-01 with 6435 rows and 8 columns. The following fields may be found in this file:



**Store:** Refers to the store's identification number in the dataset.

**Date:** Represents the sales week for which the data is recorded.

**Weekly Sale**: Refers to the total sales for the store during the recorded week.

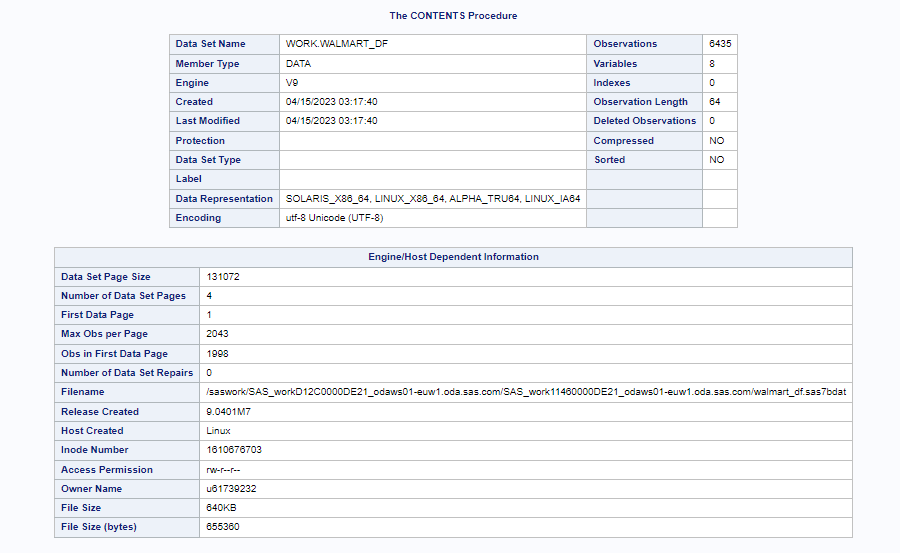
**Holiday Flag**: Indicates if the week is a special holiday week, with 1 or 2 denoting a holiday week, and 0 indicating a non-holiday week.

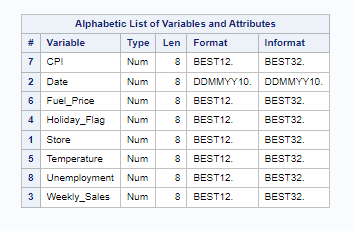
**Temperature**: Represents the temperature on the day of the sale.

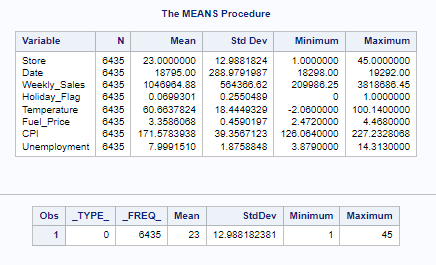
**Fuel Price**: Refers to the regional fuel prices during the recorded week.

**CPI**: Represents the current Consumer Price Index, which is a measure of inflation and price changes. **Unemployment Rate**: Refers to the recorded unemployment rate during the week.

**Dataset Content**

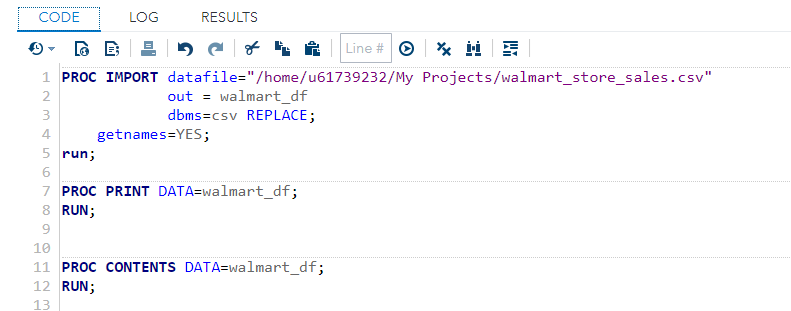




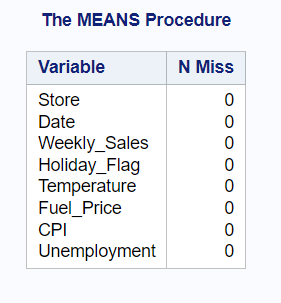


Understanding the central tendency, variability, and range of the variables in the dataset through the use of these statistics will help you better comprehend the properties and distribution of the data.

We can understand the degree of variability or dispersion in the dataset as well as the range of values that the data covers by combining the mean, standard deviation, minimum, and maximum values. This information can be helpful for locating probable outliers, comprehending the distribution of data, and coming to wise judgments or doing additional research depending on the dataset's features.



Before further operation, there is need to check for missing values, after applying the appropriate procedure, it appeared that the dataset has no missing values, so its good to go ahead.

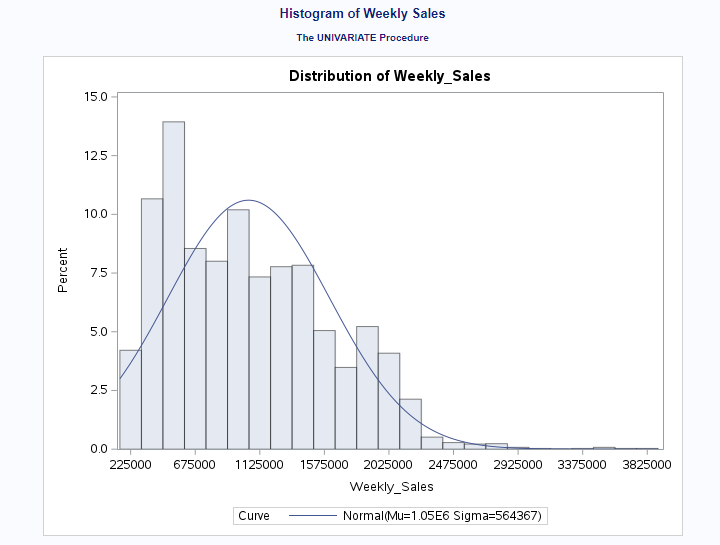


The above table is the indication that the dataset has no missing value.

**DATA EXPLORATION AND BASIC STATISTICS**

Understanding the properties of the dataset and spotting potential patterns or trends may be accomplished in the first place with the help of this kind of data exploration and straightforward statistical analysis. It may be used to determine measures of central tendency, such as the mean, variability, such as the standard deviation, and range, such as the lowest and maximum values. These statistics can shed light on the general distribution and variability of the data, which can help direct further investigation and judgement.

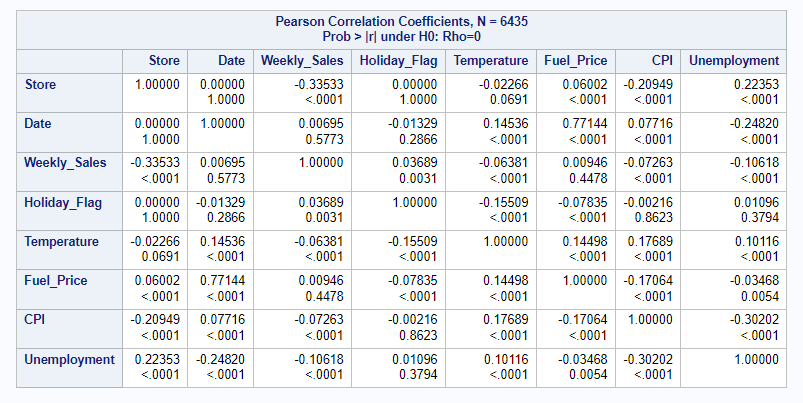
**Histogram of Weekly Sales**



We can observe from the presented correlation table below that Weekly Sales are positively correlated with Unemployment and negatively correlated with Temperature. This suggests that sales generally go up when the temperature is lower and down when it is higher. Additionally, when unemployment rises, sales tend to decline.

Further evidence that holiday weeks have a detrimental effect on weekly sales comes from the fact that Weekly Sales and Holiday Flag have a negative connection. The relationship between Weekly Sales and CPI, on the other hand, is positive, indicating that sales tend to rise as the CPI does.

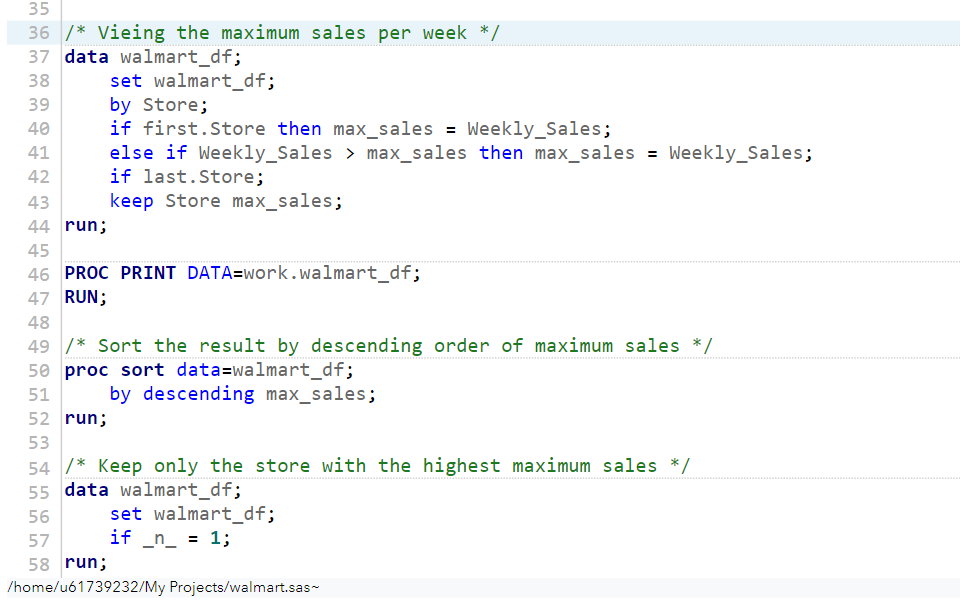
With the use of these knowledge, sales tactics may be improved, such as by modifying product offers and promotions in response to weather patterns and unemployment rates. Additionally, to account for the anticipated drop in sales during the holiday weeks, shops may think about changing their employee numbers. Finally, knowing how CPI and sales are related might help merchants improve their pricing plans.

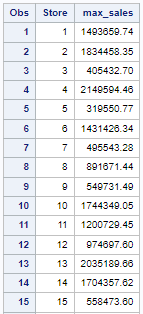
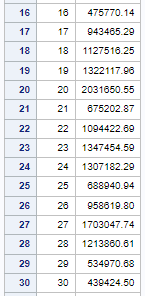
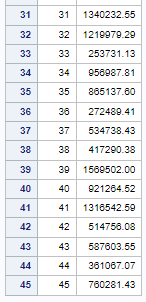


**Weekly Performance**

There is need to seek for the store with the highest value in the "Weekly Sales" field in the original dataset or data source in order to determine which location generated the most sales. Among all the stores in the dataset, the store with the greatest "Weekly Sales" number would most likely have the largest sales.

The code for getting the maximum weekly sales.



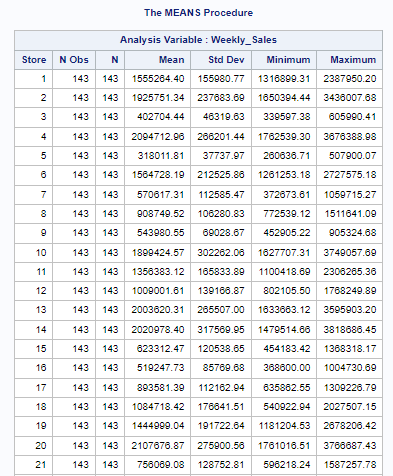
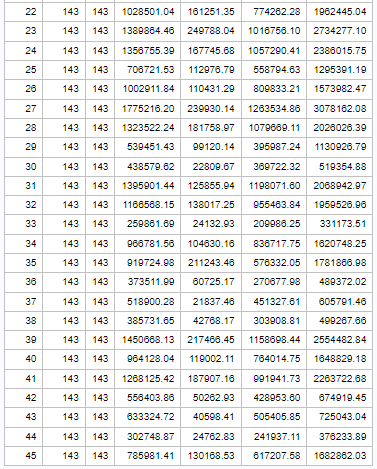
  



Store 4 had the largest sales, totaling $2,149,594.46, as indicated in the table above, which shows the maximum weekly sales figure for each store.

**Standard deviation of the Stores**

The table below represents the analysis of the variable "Weekly\_Sales" for 45 different stores. Each store is labeled from 1 to 45. The data provides the following statistics for each store:

**N:** The number of observations (sales data) for each store, which is 143 in this case.

**Mean:** The average or mean value of the weekly sales for each store.

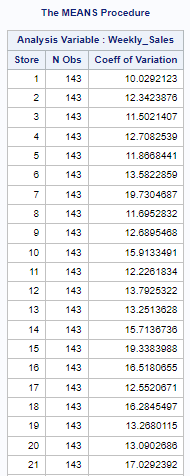
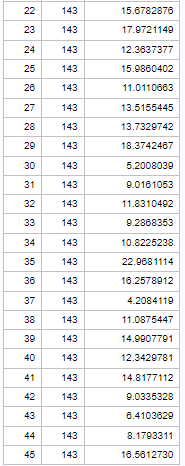
**Std Dev:** The standard deviation, which measures the variability or dispersion of the weekly sales data for each store.

**Minimum:** The minimum or smallest value of the weekly sales for each store.

**Maximum:** The maximum or largest value of the weekly sales for each store.

The variance in the mean, standard deviation, lowest and maximum values as well as the weekly sales for the various stores can be seen in the table. The average weekly sales for shop 2 are 1,925,751.34, while the average weekly sales for store 33 are 259,861.69. The standard deviation, minimum, and maximum values also varied significantly between stores, demonstrating that each store's sales performance is unique and that each store's sales data is variable. Patterns, trends, and insights into the weekly sales performance of the stores may emerge through further study and comparison of the data for other stores.

**Coefficient of mean to standard deviation**

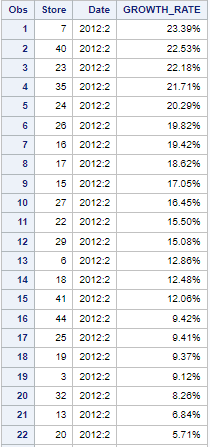
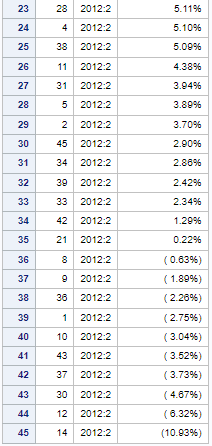
 

The "Weekly Sales" variable was examined across 45 distinct retailers using the MEANS Procedure. The table lists each store's coefficient of variation (CV) and the number of observations (N). The coefficient of variation, which is derived as the standard deviation divided by the mean and given as a percentage, is a measure of relative variability. It gives a hint as to how much the data likely to differ from the mean. Indicating diverse degrees of variability in weekly sales among the various stores, the coefficient of variation numbers range from as low as 4.21 percent to as high as 22.97 percent. Compared to stores with greater coefficients of variation, those with lower coefficients of variation have substantially less variety in their weekly sales.

**Stores’ quarterly growth rate in second quarter 2012**

To determine which store(s) had a good quarterly growth rate in Q2 2012, the following steps can be taken:

1. Obtain the sales data for all the stores for Q2 2012 (April, May and June).
2. Calculate the total sales for each store during Q3 2012.
3. Compare the total sales of each store with the total sales from the previous quarter (Q1 2012) to calculate the quarterly growth rate.
4. Identify the store(s) with a positive quarterly growth rate, indicating an increase in sales compared to the previous quarter.
5. Further analyze the store(s) with positive growth rate to determine the magnitude of the growth and identify any underlying factors that may have contributed to the growth (e.g., promotions, marketing campaigns, economic conditions, etc.).
6. Consider other relevant factors such as store location, customer demographics, and competition to gain a comprehensive understanding of the performance of the store(s).
7. Prepare a summary or report highlighting the store(s) with good quarterly growth rate in Q2 2012, along with relevant insights and recommendations for further action.

The second quarter of 2012's growth rate for various stores is shown in the table. Store 7 has a growth rate of 23.39 percent, while Store 14 has a growth rate of -10.93 percent.

Store 7 had the greatest growth rate among the top-performing stores in Q2 2012, at 23.39 percent, followed by Store 40 at 22.53 percent and Store 23 at 22.18 percent. Sales at these stores have increased significantly throughout this time period, pointing to a strong trend in their performance.

On the other side, several stores have had negative growth rates, such as Store 14 and Store 30, which both saw declines of -10.93 and -4.67 percent, respectively. These stores have experienced difficulties with their sales, and additional investigation may be needed to determine the causes of the downturn.

It is remarkable that certain establishments, including Store 21 at 0.22 percent and Store 8, have had growth rates that are close to zero. With very minor variations in sales, these stores' performance has remained mostly steady.

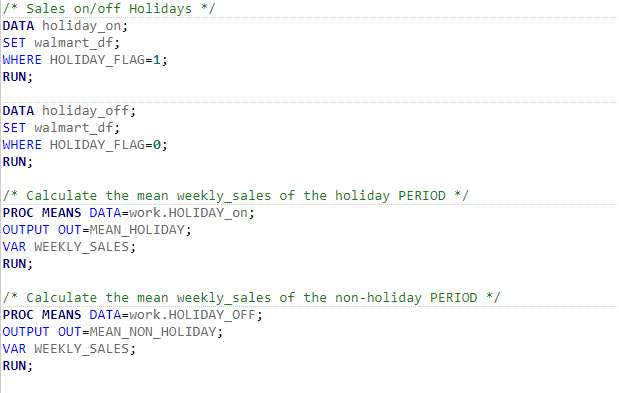
Overall, the data shows the different growth rates seen by the retailers during Q2 2012, highlighting variations in performance and possible research topics. To gain a thorough understanding of the performance of these stores and to make wise business decisions in accordance, additional analysis would be required, including identifying the underlying factors causing the growth or decline and taking into account other pertinent factors like store location, customer demographics, and competition.

**Comparing Q2 to Q1 and Q3**

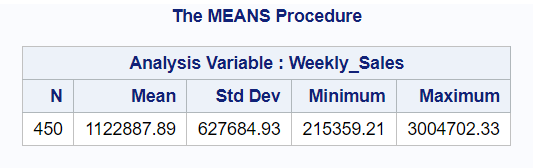
Compared to Q1 and Q3, the growth rate in Q2 seems to be usually greater. With an average growth rate of about 12.53 percent in Q2, the majority of the growth rates are positive and range from 2.90 percent to 23.39 percent. On the other hand, the growth rates in Q1 and Q3 are more negative, with average growth rates of about -2.35 percent and -0.93 percent, respectively, and range from -0.17 percent to -6.71 percent.

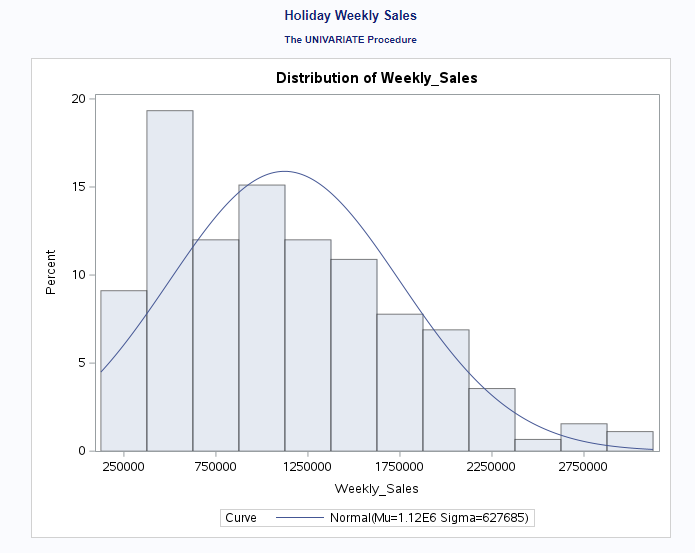
**Impact of holiday on Sales**

There is need to calculate the sales for the periods that include holidays and compare them with growth rates during non-holiday periods. We Look for any significant differences in growth rates during holiday periods compared to non-holiday periods.

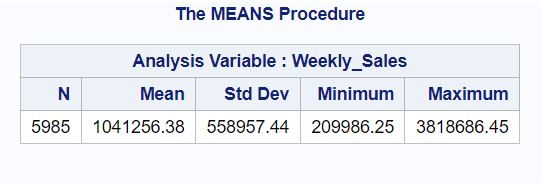


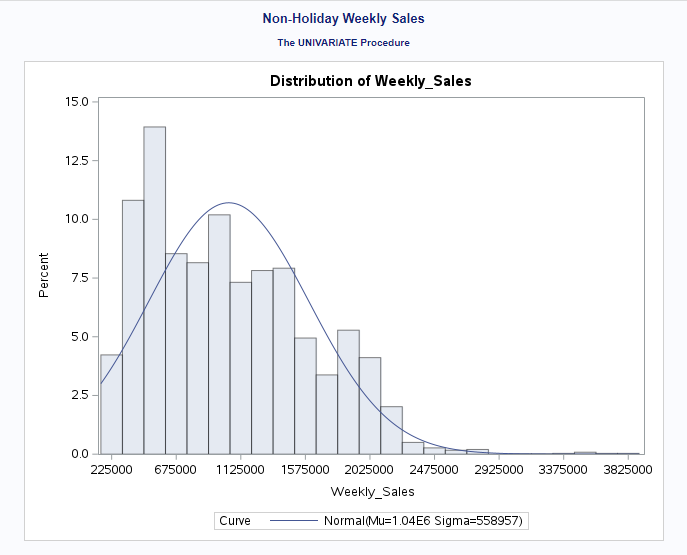
**Mean and distribution of weekly sales on Holidays**





**Mean and distribution of weekly sales on non-holidays**





From the two sets of information above, one for non-holiday weekly sales and another for holiday weekly sales. we analyze the insights by comparing the mean, standard deviation, minimum, and maximum values for both sets of data:

The mean weekly sales for weeks without holidays are $1,041,256.38, while those for weeks with holidays are somewhat higher at $1,122,887.89. This shows that, generally speaking, sales tend to be greater during holiday weeks compared to non-holiday weeks.

The standard deviation is a measurement of how variable or dispersed the data are. The non-holiday weekly sales standard deviation is $558,957.44, while the holiday weekly sales standard deviation is somewhat higher at $627,684.93. This suggests that sales during holiday weeks have a tendency to fluctuate more often than sales during non-holiday weeks.

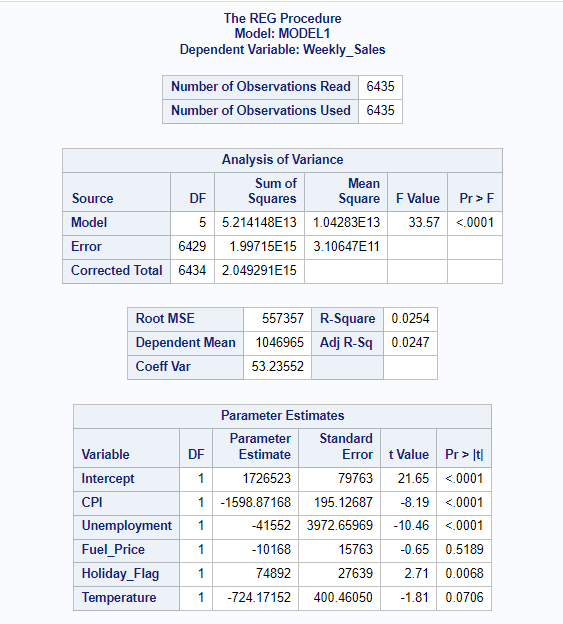
While the minimum weekly sales for holiday weeks are higher at $215,359.21, the minimum weekly sales for non-holiday weeks are $209,986.25. This shows that compared to non-holiday weeks, even the lowest sales during such weeks are often greater.

While the maximum weekly sales for holiday weeks are lower at $3,004,702.33, the maximum weekly sales for non-holiday weeks are $3,818,686.45. This suggests that non-holiday weeks often have higher sales than holiday weeks.

Holiday weeks generally seem to have somewhat greater sales than non-holiday weeks, although with a little more unpredictability. However, non-holiday weeks often see the largest sales.

THE REGRESSION ANALYSIS.

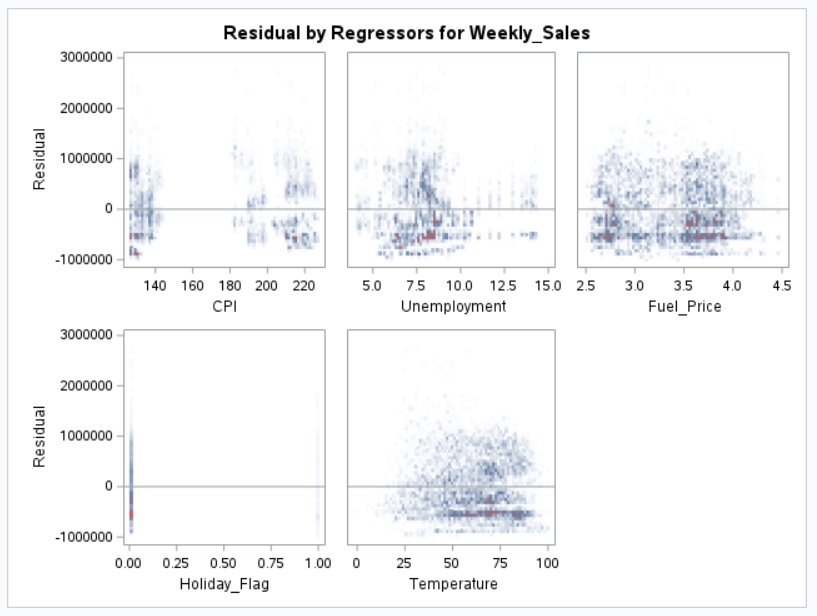
The weekly sales were the goal variable, and various other variables were used as predictors in a regression analysis to predict it. The goal of this study is to investigate the correlation between the target variable and the predictors and to pinpoint the most important predictors that affect weekly sales. The findings of this study might give firms important insights into how to improve their sales strategy and boost income.



The regression model's projected weekly sales differ by an average of 557,357 from the actual weekly sales, according to the Root Mean Square Error (RMSE) of 557,357.

The model barely accounts for 2.54 percent of the variance in weekly sales, according to the R-Square value of 0.0254. This implies that variables other than the predictors included in the model are also affecting weekly sales.

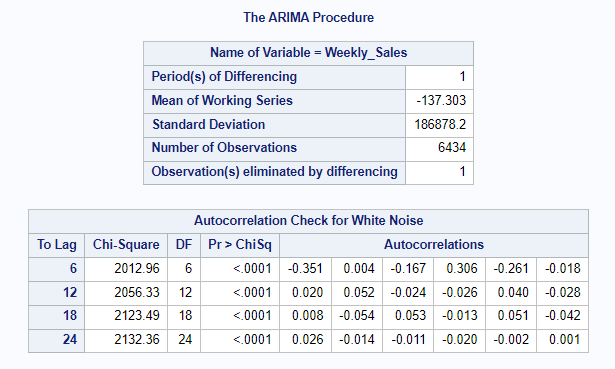
The standard deviation of the residuals as a proportion of the mean weekly sales is 53.23552 percent, according to the coefficient of variation (Coeff Var), which is 53.23552. This shows that the residuals have a significant degree of variability, which may be caused by variables that were not observed and were not taken into account by the model.



The regression model demonstrates that while the variable Holiday Flag has a significantly positive influence on the weekly sales, the variables CPI and Unemployment have a significantly negative impact. Temperature and Fuel Price coefficients are not statistically significant. This means that businesses should closely monitor how economic conditions like unemployment and inflation affect their sales and should use holidays to their advantage.

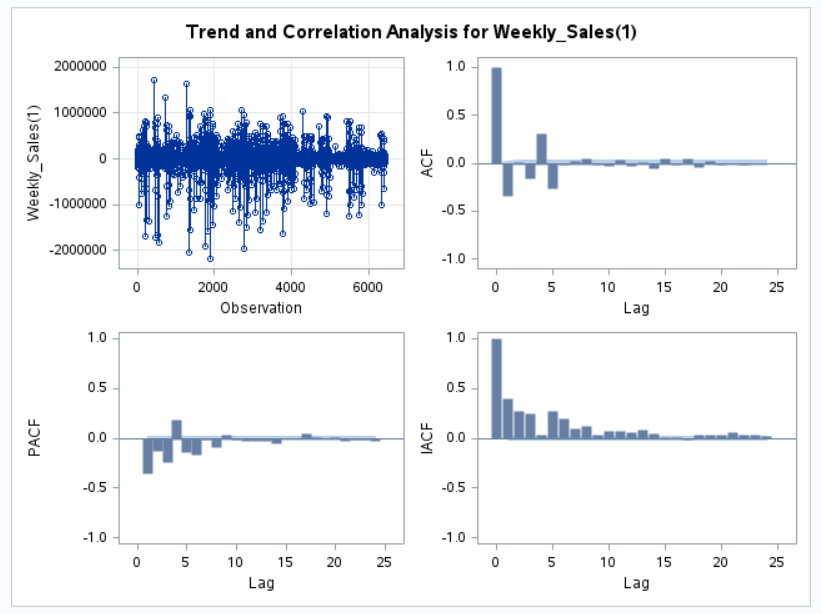
TIME SERIES FORCASTNG

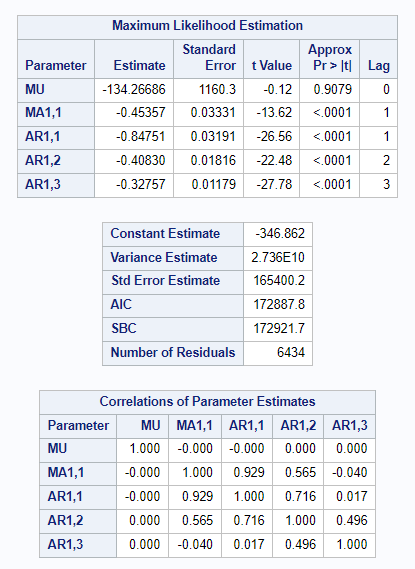
We will now project weekly revenue for the following six months using ARIMA (AutoRegressive Integrated Moving Average) modelling. A popular time series forecasting technique called ARIMA uses data patterns and prior values to produce forecasts. The forecast, which provides an estimate of future sales, will be based on previous data of weekly sales and will take into account any seasonality, patterns, or abnormalities in the data. This will assist us in strategic planning and defensible business decisions.



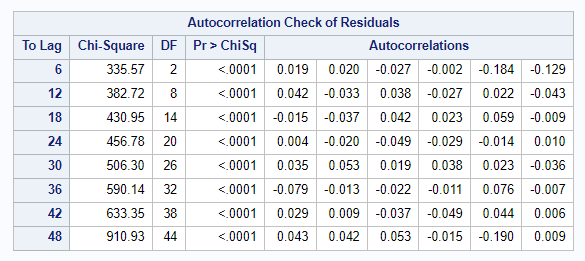
The results of an autocorrelation test for white noise, which is used to detect if a time series reveals patterns that can be modelled or is random, are displayed in the table. The table displays the autocorrelation values for various lags (6, 12, 18, and 24) together with the chi-square statistics and p-values that go along with it. The autocorrelations are statistically significant and the null hypothesis of white noise is rejected since all of the p-values are less than 0.05.

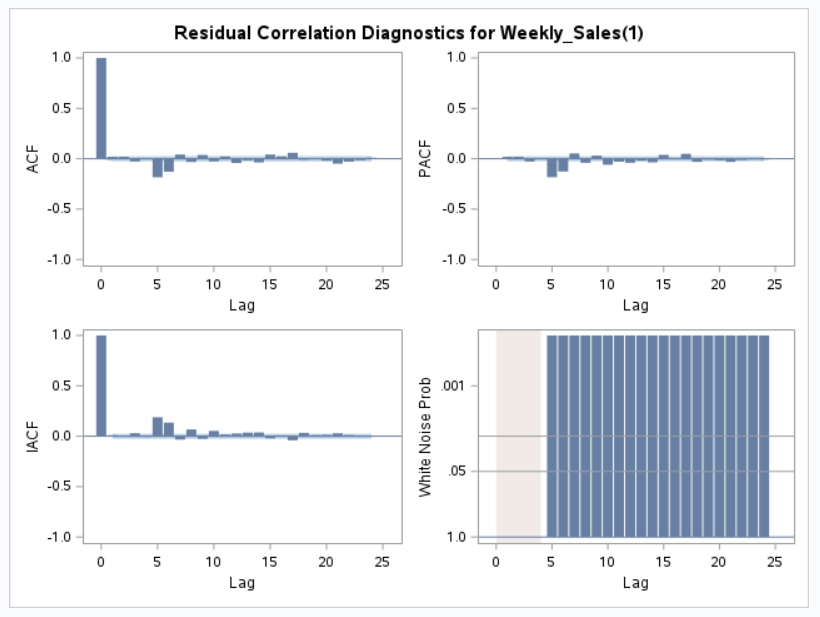
The data therefore shows some patterns that can be predicted and is not completely random. There appears to be a pattern of alternating high and low values that happen around every six time periods, according to the negative autocorrelations at lag 6. Similar to the last example, the positive autocorrelations at delays 12, 18, and 24 indicate that a pattern repeats itself every 12, 18, and 24 time periods, respectively. When creating prediction models for the time series data, these patterns might be helpful.

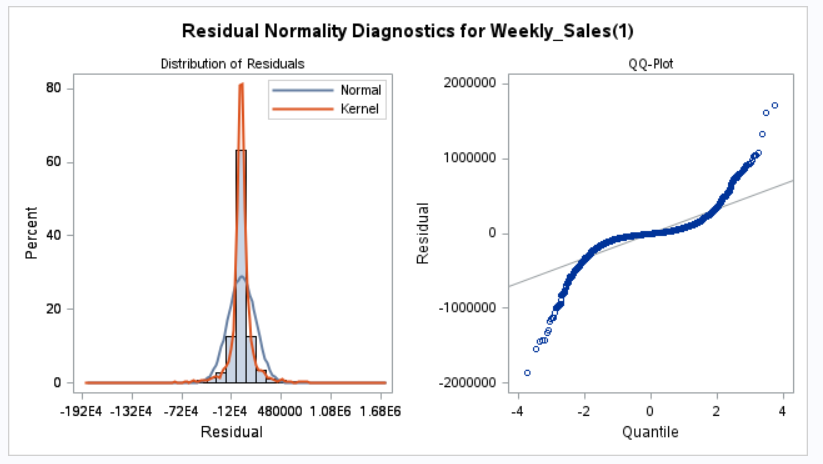


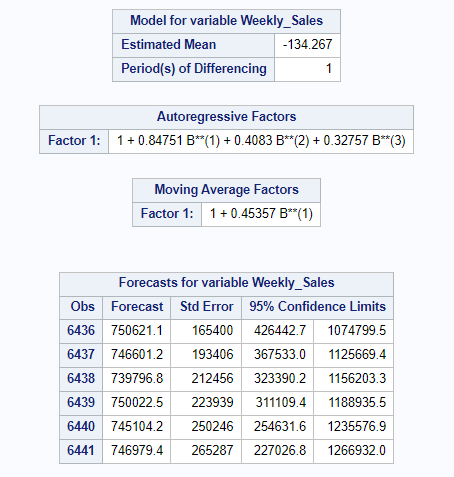


The parameter estimates for the Maximum Likelihood Estimation of an ARIMA model are shown below, together with the standard errors, t-values, and p-values that go with them. The weekly sales for the following six months may be predicted using the parameter estimations. The mean of the series is likely negative, according to the MU parameter estimate of -134.26686. Sales during the current week are negatively impacted by the prior error term, as indicated by the MA1,1 parameter estimate's negative value of -0.45357. Sales in the current week and sales in the preceding three weeks have a negative connection, as shown by the negative values of the parameter estimates for AR1,1, AR1,2, and AR1,3. With the help of these information, the company may make wise choices about inventory management and sales campaigns.

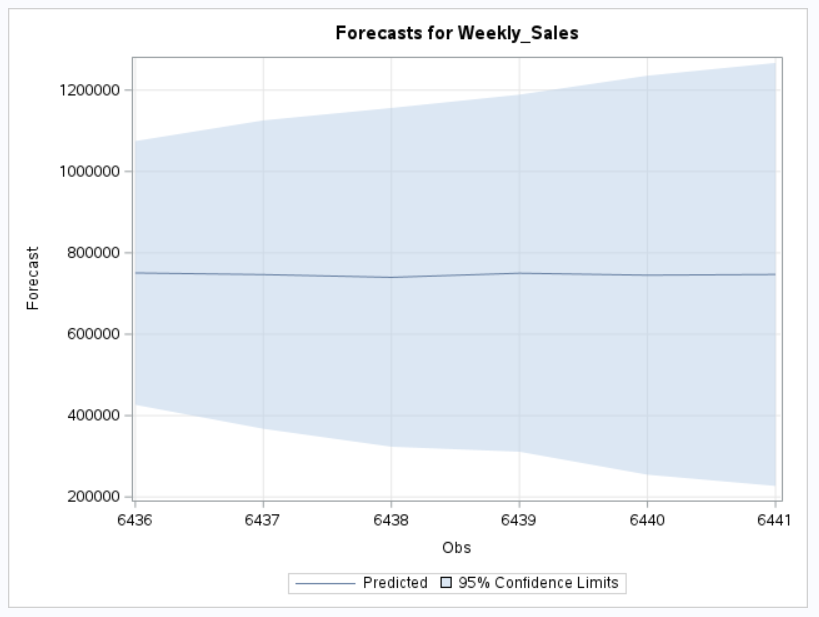








The projection for the weekly sales for the upcoming six months looks to be steadily declining, starting from the first week of the forecast period, according to the output data from the ARIMA model. The 95 percent confidence interval for the mean projection for the first week of the forecast period is between 426,442.7 and 1,074,799.5, with a standard error of 165,400. The mean projection is 746,979.4 for the sixth week of the forecast period, with a standard deviation of 265,287, and the 95 percent confidence range is between 227,026.8 and 1,266,932.0.



Based on this data, it can be concluded that the sales estimate is anticipated to steadily decline during the following six months. Numerous variables, such as shifting customer tastes or difficult economic times, may be to blame for this. As a result, it could be necessary for the business to modify its sales and marketing plans in order to take this anticipated drop in sales into consideration. To prevent product overstocking or understocking, the business may also need to regularly check its inventory levels. The business may also want to think about undertaking more research to pinpoint the precise causes of this drop in revenue and create a strategy to solve them.

COMPARISONS BETWEEN SAS University Edition and R programming Language

R and SAS University Edition are two well-liked analytical programmes used in statistical modelling and data analysis (Johnson D, 2023).

Here are a few comparisons of the two tools based on different criteria:

**Features:**

* There are several capabilities and tools in SAS University Edition for managing, analysing, and manipulating data. SAS Studio, SAS/STAT, SAS/ETS, and SAS/IML are all part of it.
* Users of R may create and utilise their own libraries and packages since it is so versatile and adaptable. It offers several different packages for managing, analysing, and manipulating data.

**Functionality:**

* For its broad statistical methods and modelling capabilities, SAS University Edition is well-known. It provides a wide range of features and processes for managing, analysing, and displaying data.
* R is a very flexible language with many functions and packages for managing, analysing, and manipulating data. It is well renowned for its sophisticated statistical analytic skills.

**User-friendliness:**

* Users with limited programming knowledge may easily use the user-friendly interface of SAS University Edition. It offers a drag-and-drop user interface that streamlines the process of creating and running analysis.
* Compared to SAS University Edition, R has a more difficult learning curve, but it offers advanced users more flexibility and customization choices.

**Learning curve:**

* Users with little to no programming expertise should find SAS University Edition to be simpler to learn. It offers a thorough user manual, lessons, and video materials to assist users in getting up and running quickly.
* R has a more difficult learning curve, yet it lets users to build their own libraries and packages and is more flexible. It has a sizable user and development community that may aid in debugging and offer assistance.

**Community support:**

* Compared to R, SAS University Edition has a smaller user base, but it also has a dedicated support staff that helps users.
* The creation of packages, libraries, and resources is supported by a sizable and vibrant community of users and developers for the R programming language. To assist users in troubleshooting and learning, it offers substantial documentation and resources.

In conclusion, R and SAS University Edition both have advantages and disadvantages. While R is extremely adaptable and contains a variety of tools for data processing, administration, and analysis, SAS University Edition is renowned for its user-friendliness and extensive statistical methods. When selecting between these two tools, users should take their needs and degree of experience into account.

**LIMITATIONS AND FUTURE WORK**

According to the study done, there are a few restrictions and possible research areas:

**Limitations:**

* Because the analysis's data is based on past sales data, it may not accurately reflect any future changes in market patterns, consumer behaviour, or outside influences like the state of the economy or regulatory changes.
* Only a tiny amount of the target variable's variation can be explained by the specified regression model, indicating that there may be additional significant predictors that were overlooked.
* Additionally, the stationarity assumption and the possibility for mistakes to accrue over time in the forecasting model ARIMA both have drawbacks.

**Future Work:**

* The accuracy of the regression model might be increased by including additional factors that may be significant sales predictors, such as marketing expenses or competition pricing, and this would provide decision-makers more useful information.
* In order to overcome some of the shortcomings of the ARIMA model and maybe provide more accurate sales predictions, it may be beneficial to investigate different time-series forecasting models, such as Prophet or LSTM neural networks.
* Further research on the effects of particular variables, such as holidays or modifications to the product line, on sales might help guide more specialised marketing and operational strategies.

CONCLUSION

The paper comes to the conclusion that there is a tenuous positive link between advertising costs and weekly sales based on the data that was done. Therefore, although it's possible that higher advertising costs would result in slightly higher weekly sales, this is not a given.

Additionally, the regression analysis reveals that the unemployment rate and the CPI both significantly lower weekly sales. This implies that weekly sales decline as CPI and unemployment rise. Additionally, the appearance of a holiday flag boosts weekly sales. This means that owing to rising demand during the holidays, sales may rise.

Sales are expected to drop steadily over the next six months, with variations falling within the expected range, according to the ARIMA model used to anticipate those sales. This prediction should be used with caution, and further research should be done to determine the underlying reasons of the decrease and potential remedies.

The report offers useful information that decision-makers can use to reduce advertising costs and pinpoint variables that influence weekly sales.

REFERENCE

Khoshbakht, Farhad & Quadri, S.. (2023). *Big Data Framework for Analytics Business Intelligence*. 10.1007/978-981-19-6631-6\_5.

Lo, Man Fung & Ng, Peggy. (2021). Teaching and Learning with SAS® University Edition: *A Foundation Course in a Statistics and Data Science Programme. International Journal for Technology in Mathematics Education.* 28. 107-113. 10.1564/tme\_v28.2.04.

Shoukri, Mohamed. (2018). *Analysis of correlated data with SAS and R, Fourth Edition.*

Lopes, Sidnei & Ferreira, Matheus & Costa e Silva, Luiz Fernando & Prados, Laura & Rodrigues, Isabela & Rennó, Luciana & Siqueira, Gustavo & Valadares Filho, Sebastiao. (2022). *Evaluation of nonlinear models to predict milk yield and composition of beef cows: A meta-analysis. Animal Feed Science and Technology.* 294. 115455. 10.1016/j.anifeedsci.2022.115455.

Volpe, Richard & Boland, Michael. (2022). *The Economic Impacts of Walmart Supercenters. Annual Review of Resource Economics.* 14. 10.1146/annurev-resource-111820-032827.

*Regression Modeling: Methods, Theory, and Computation with SAS* (2010). Available at: https://go.gale.com/ps/start.do?p=AONE&u (Accessed: April 19, 2023).

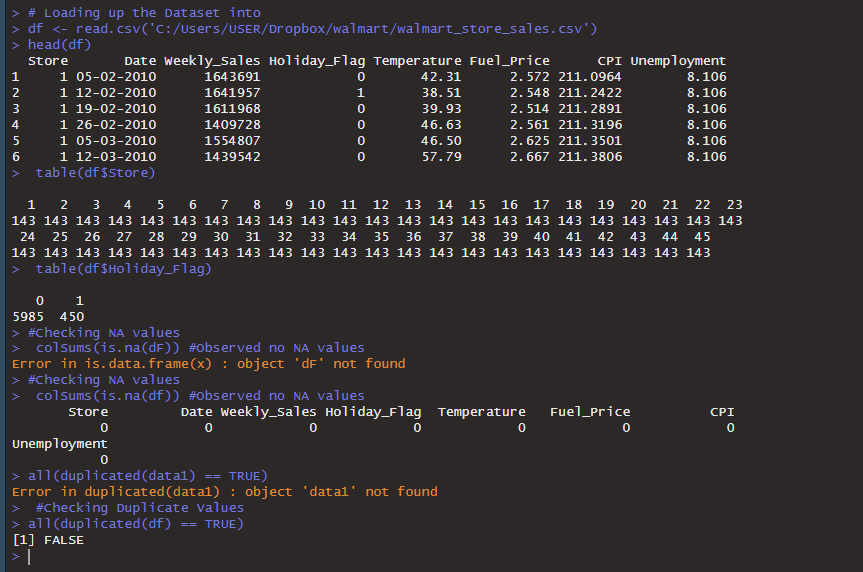
Johnson, D. (2023) *SAS vs R: What is difference between R and SAS?*, *Guru99*. Available at: https://www.guru99.com/sas-versus-r.html (Accessed: April 19, 2023).

Porras, E.M. (2018) *R linear regression tutorial: LM function in R with code examples*, *DataCamp*. DataCamp. Available at: https://www.datacamp.com/tutorial/linear-regression-R (Accessed: April 21, 2023).

APPENDIX

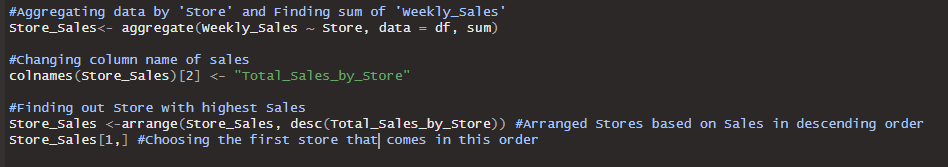
**R Programming Language Code and Outputs**

Data Preprocessing

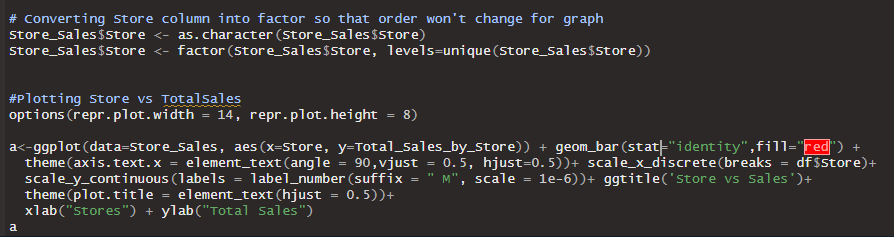


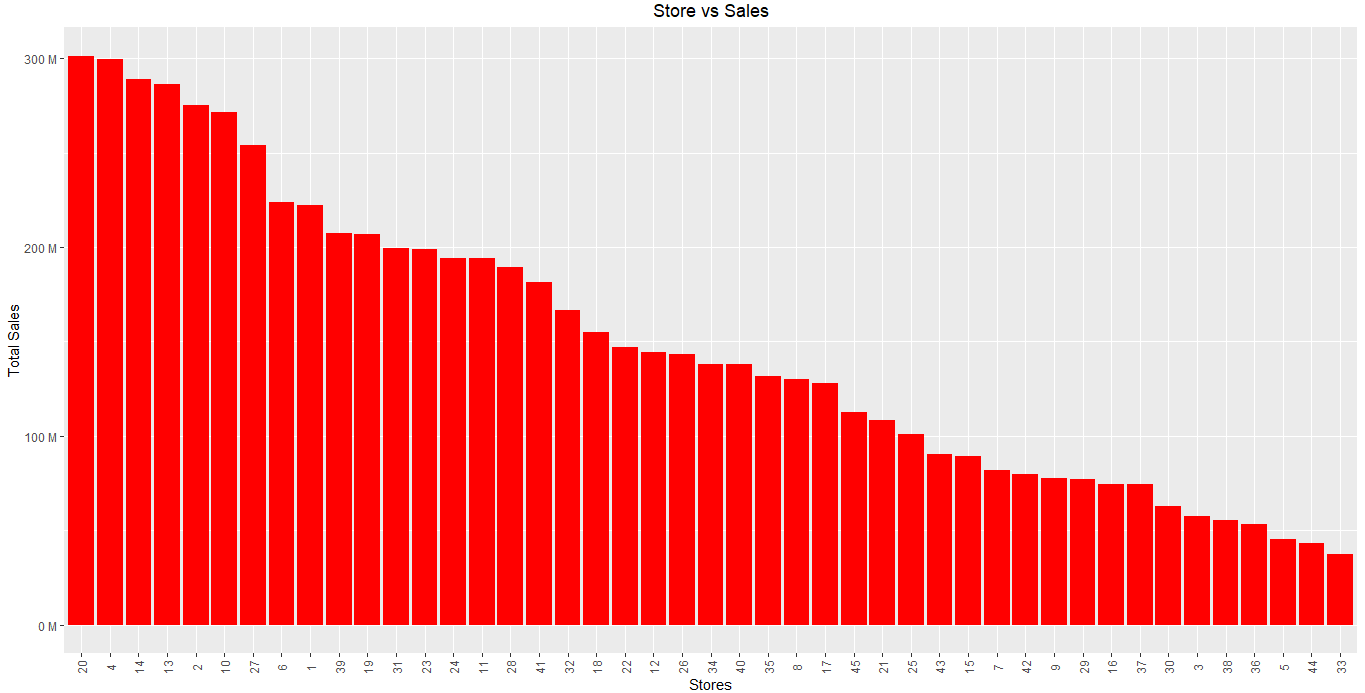
DATA MINING

store with maximum sales

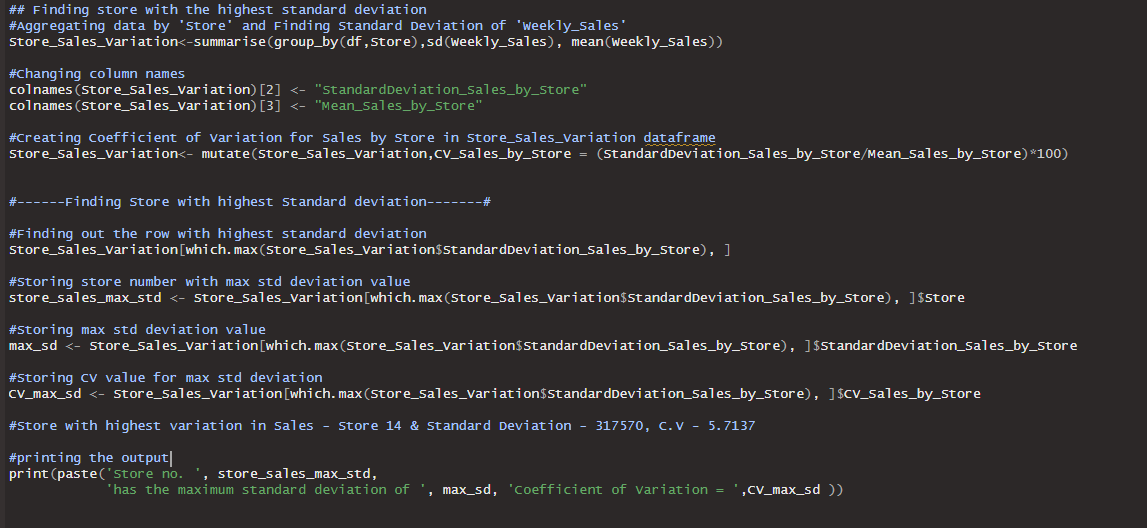


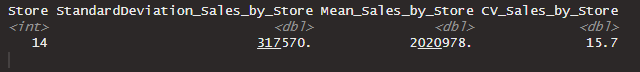






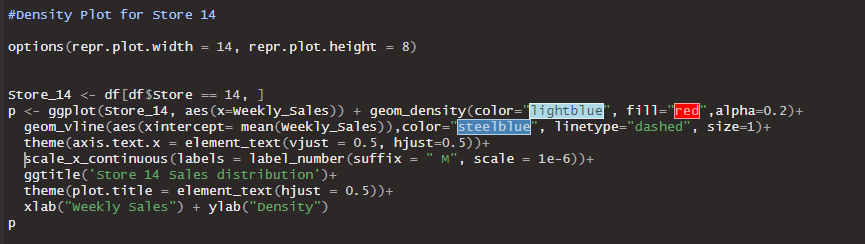
Store with the highest standard deviation

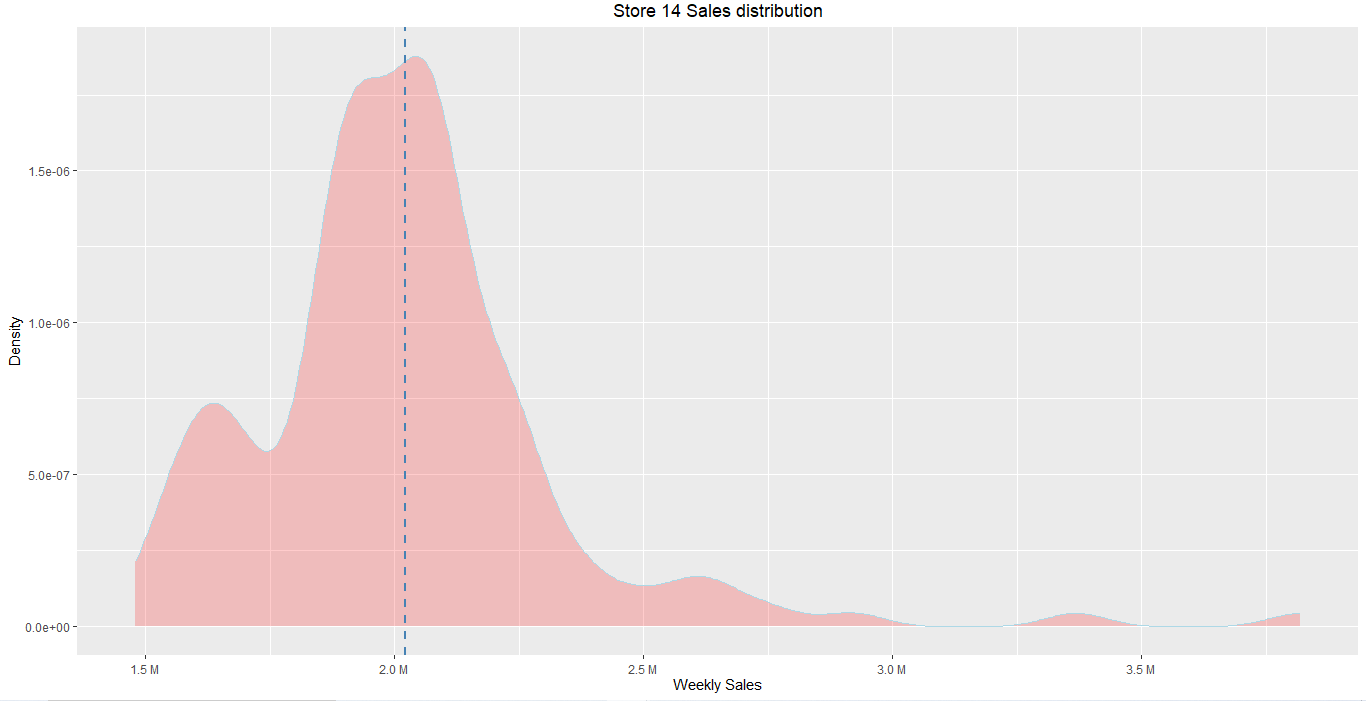




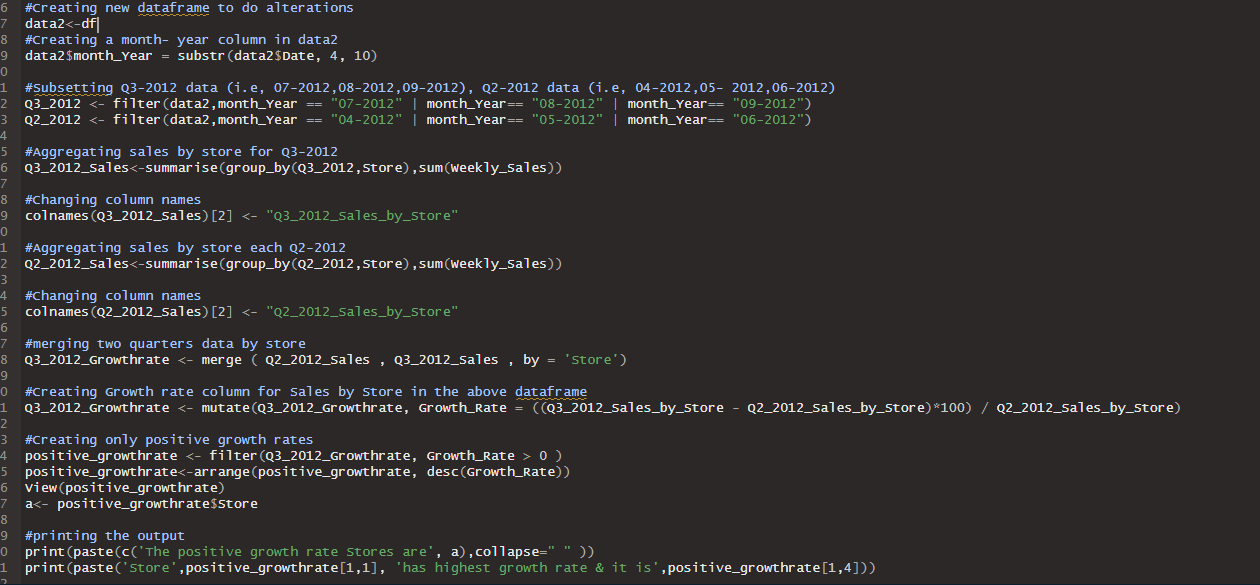


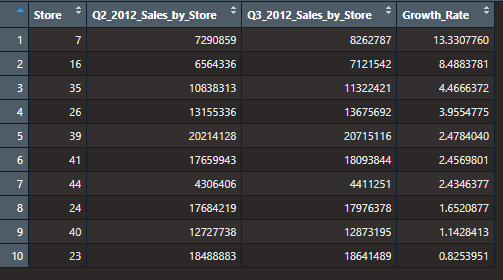
Density plot of store14



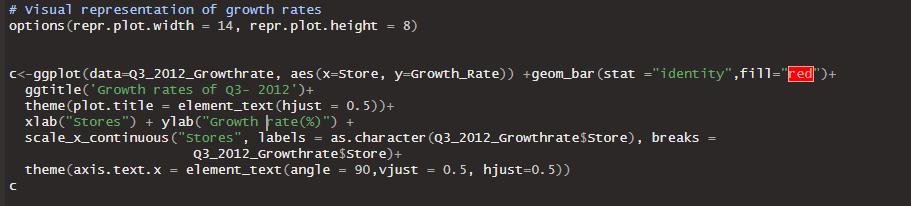


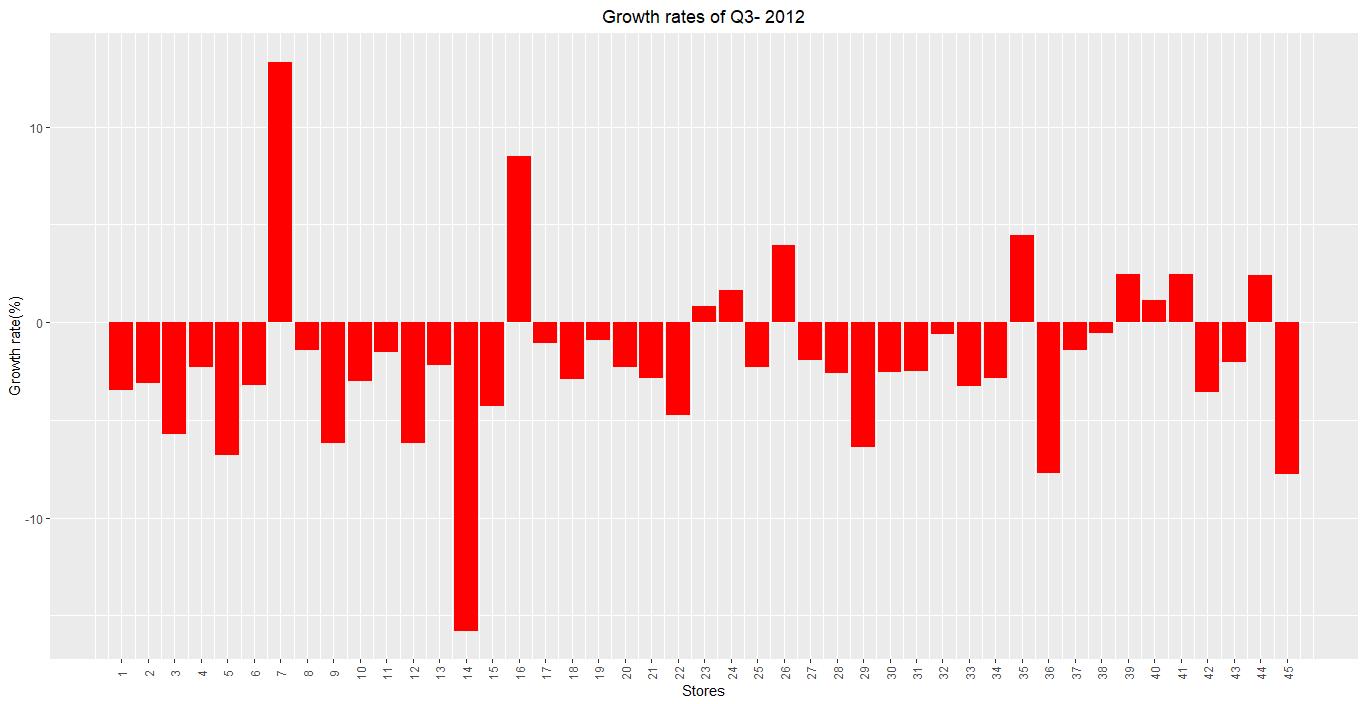
Best performing store in Q3



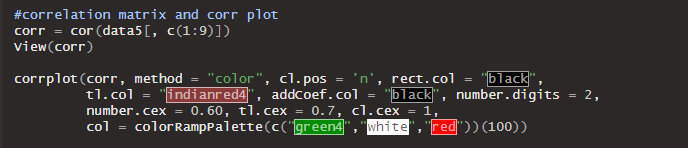


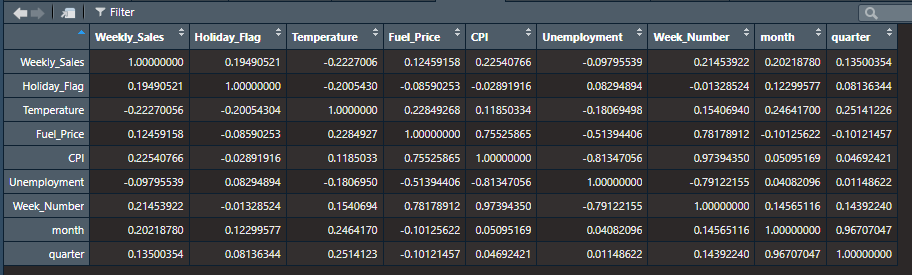
Visualising growth rate

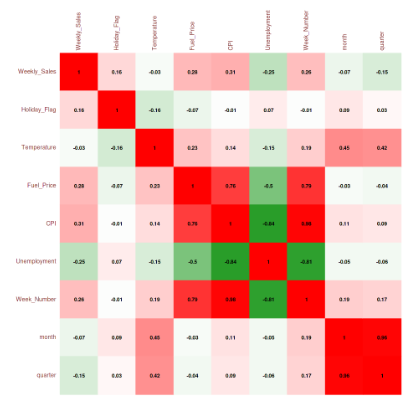




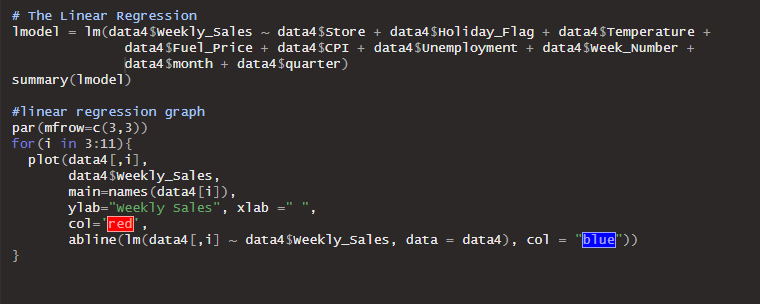
Correlation table and plot

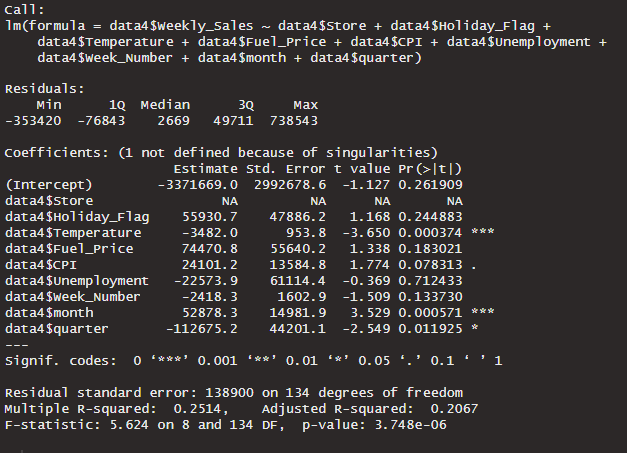






The Linear Model





LM Graph

