Sales Classification and Prediction on Retail Data

Author

Prepared for UMBC Data Science Master Degree Capstone course under the guidence of Dr Chaojie (Jay) Wang

Link to GitHub Profile: https://github.com/SaiGangadharVeeramreddy

Link to LinkedIn Profile: https://www.linkedin.com/in/gangaveeramreddy/

Link to PowerPoint Presentation File:https://github.com/DATA-606-2023-FALL-MONDAY/Veeramreddy_Sai_Gangadhar/blob/main/docs/Sales%20Classification%20and%20Predicti

Link to YouTube Video: https://www.youtube.com/watch?v=eQfvI5n-f3I

Abstract

This project seeks to develop a sales forecasting model using the Retail Sales Data. The project explores various prediction models including Decision Tree, Lineear Regression, and K-Nearest Neighbors. Decision Tree is determined to be the most appropriate prediction algorithm with a R-Square score of 0.95.

Introduction

Data analytics is a crucial tool for businesses across all industries. For example, machine learning algorithms can be used to examine databases in search of patterns and correlations. Prediction, anomaly detection, data clustering, fraud detection, and automation are just few of the many applications of historical data. Machine learning can be applied on sales data over time to reveal hidden correlations between factors, such as customer buying behavior, store performance, and seasonal variations in sales, to assist businesses make informed decisions. Proper use of machine learning makes sales forecasting a powerful tool that may inform strategic planning.

Background Information

No matter the size of a company, accurate sales forecasting is essential. It's a method for businesses to anticipate sales by looking at past performance. Financial planning, resource allocation, strategy formulation in marketing, and projecting future demand are all aided by sales forecasting (Ma & Fildes, 2021). It's a method that can give a company an edge by allowing them to quickly adapt to new market conditions. Because it helps retail managers forecast sales volume at the Stock Keeping Unit (SKU) and store levels, it also contributes to improved consumer satisfaction through distribution and replenishment methods (Ma & Fildes, 2021). As a result, this helps alleviate the issue of unfilled orders from clients during times of high demand. Furthermore, being able to estimate demand benefits in reducing waste, improving sales revenue, and attaining effective distribution (Ma & Fildes, 2021). The ability to accurately predict future sales is a cornerstone of effective management.

What is the Project about?

This project is all about using data to help retail businesses make better decisions. The primary objective of this project is to offer a comprehensive understanding of the retail landscape, enabling businesses to make data-driven decisions that enhance their operations, boost profitability, and ultimately enrich the customer experience. By doing this, we can discover important trends and patterns that can help stores improve their operations, make more money, and give customers a better shopping experience. In the end, the goal is to give retailers the tools they need to succeed in a competitive market by using data in smart ways.

Why does it matter?

It matters because it can make stores more successful and customers happier. When stores use data to make decisions, they can sell things better and not waste money on stuff people don't want. This helps stores stay in business and keep prices reasonable. It also means shoppers can find what they need and maybe even get good deals. So, using data is like a win-win for stores and customers—it helps everyone.

What are your research questions?

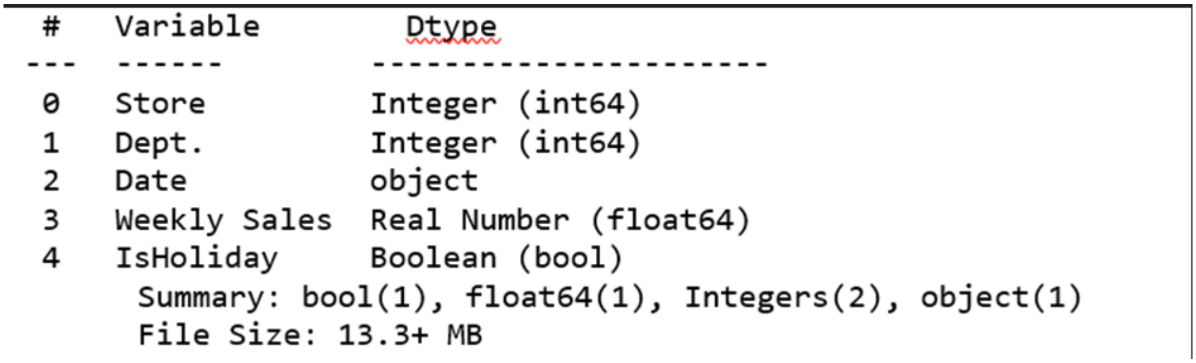
- **Sales Prediction:** Can we use the information about store location, temperature, fuel prices, and special holiday weeks to predict how much a store will sell in a given week? This could help stores plan better.
- **Markdown Impact Analysis:** Do discounts and promotional markdowns have a big effect on sales? We can investigate if offering discounts during certain weeks leads to increased sales.
- **Store Clustering:** Can we group similar stores together based on their sales history and the features like location and economic conditions? This might help stores understand their competition and market better.
- **Holiday Sales Assessment:** Does having a special holiday week significantly change sales patterns? We can examine if stores should prepare differently for these weeks in terms of stocking and staffing.

Data

The Retail Sales dataset comprises three CSV files: sales, features, and stores. The sales CSV file is the largest, containing more than 420,000 rows. It is made up of four columns. The first is "Store," which captures the store number for the 45 different stores. The second column (Dept.) provides information regarding the department to which the matching weekly sales values (column 4) are assigned. Column three details the dates in seven-day intervals (spread across three years between 2010 and 2013), while the last column, called IsHolidays, indicates whether the week was a holiday or not. Features CSV has slightly more than 8000 rows and comprises 12 columns. The first two columns are Store Number and Data, similar to the Stores CVS file. Column three (Temperature) gives the weekly average temperature in Fahrenheit for the region where the corresponding store is located. Column four (Fuel_Price) details the average fuel cost in the region where a store is located. Columns five to nine (labeled MarkDown1 to 5) detail data related to promotional markdowns, such as discounts, coupons, and endcaps, and are only available for a few months in a year. Column 10 of the Features CSV files is called CPI, detailing the consumer price index. Columns 11 (unemployment rate) and 12 (IsHoliday) detail the unemployment rate and whether the week is a special holiday week. The last file contains anonymized data about the 45 stores (hence 45 rows) and has three columns detailing the store number (column one) and the corresponding type of store (column two) and size (column three). The data types for all the columns in the three files are summarized in the tables below. Also featured in the table are file sizes for the three files.

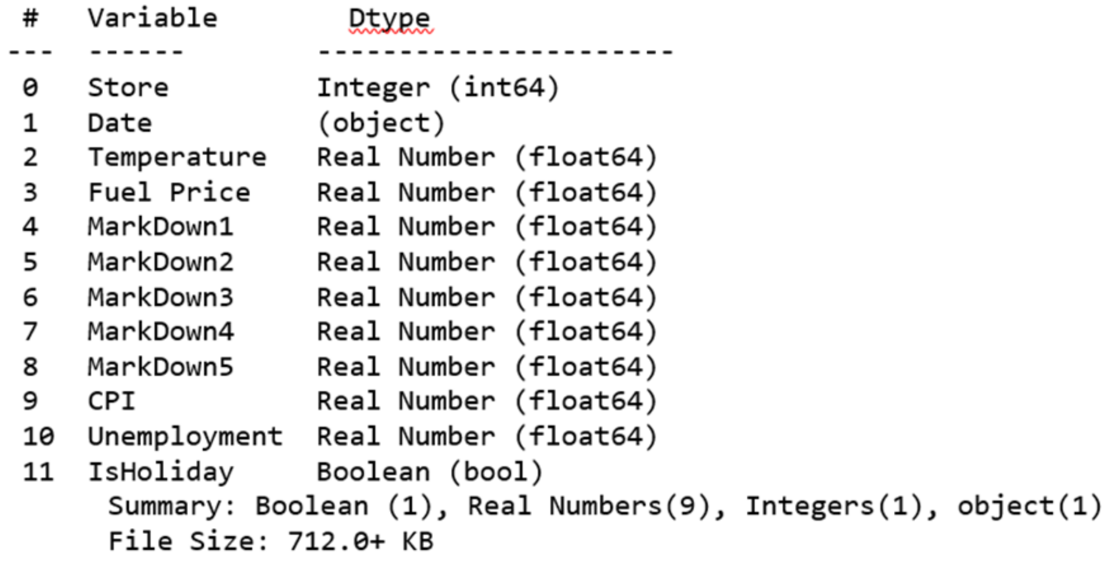
Data sources: The data set is available at https://www.kaggle.com/datasets/manjeetsingh/retaildataset?select=stores+data-set.csv

Table 1\ *Data Types for the Variables in the Sales Data File*



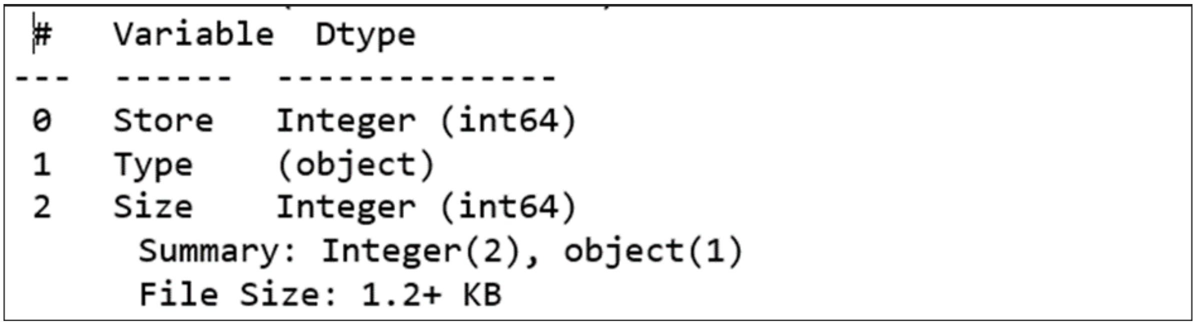
```
 #   Variable        Dtype
---  ------         ----------------------
 0   Store          Integer (int64)
 1   Dept.          Integer (int64)
 2   Date           object
 3   Weekly Sales   Real Number (float64)
 4   IsHoliday      Boolean (bool)
      Summary: bool(1), float64(1), Integers(2), object(1)
      File Size: 13.3+ MB
```

Table 2\ *Data Types for the Variables in the Features Data File*



```
  #   Variable       Dtype
 ---  ------        ----------------------
  0   Store         Integer (int64)
  1   Date          (object)
  2   Temperature   Real Number (float64)
  3   Fuel Price    Real Number (float64)
  4   MarkDown1     Real Number (float64)
  5   MarkDown2     Real Number (float64)
  6   MarkDown3     Real Number (float64)
  7   MarkDown4     Real Number (float64)
  8   MarkDown5     Real Number (float64)
  9   CPI           Real Number (float64)
  10  Unemployment  Real Number (float64)
  11  IsHoliday     Boolean (bool)
        Summary: Boolean (1), Real Numbers(9), Integers(1), object(1)
        File Size: 712.0+ KB
```

Table 3\ *Data Types for the Variables in the Stores Data File*



```
 #   Variable  Dtype
---  ------  --------------
 0   Store   Integer (int64)
 1   Type    (object)
 2   Size    Integer (int64)
      Summary: Integer(2), object(1)
      File Size: 1.2+ KB
```

Which variable/column will be your target/label in your ML model?

- Store Number
- Isholiday
- Temperature
- Fuel_Price
- CPI
- Unemployment rate
- type of store
- size of store
- Departmental_Sales

Which variables/columns may be selected as features/predictors for your ML models?

- Weekly_Sales

Exploratory Data Analysis (EDA)

Data Preprocessing

Data Integration

Data preprocessing is an important step in modeling. It improves performance by removing duplicates, irrelevant values, and outliers that might lead to misleading results. For this project, several data preprocessing techniques were employed. First was data integration, used to merge the three files (Sales, Features, and Stores) into a single comprehensive dataset. The first to be merged were the Sales and Features files, merged based on three identical columns ("Store," "Date" and "IsHoliday") present in both of them. The new combined dataset was subsequently integrated with the Stores file, with the "Store number" serving as the column for comparison.

Data Cleaning

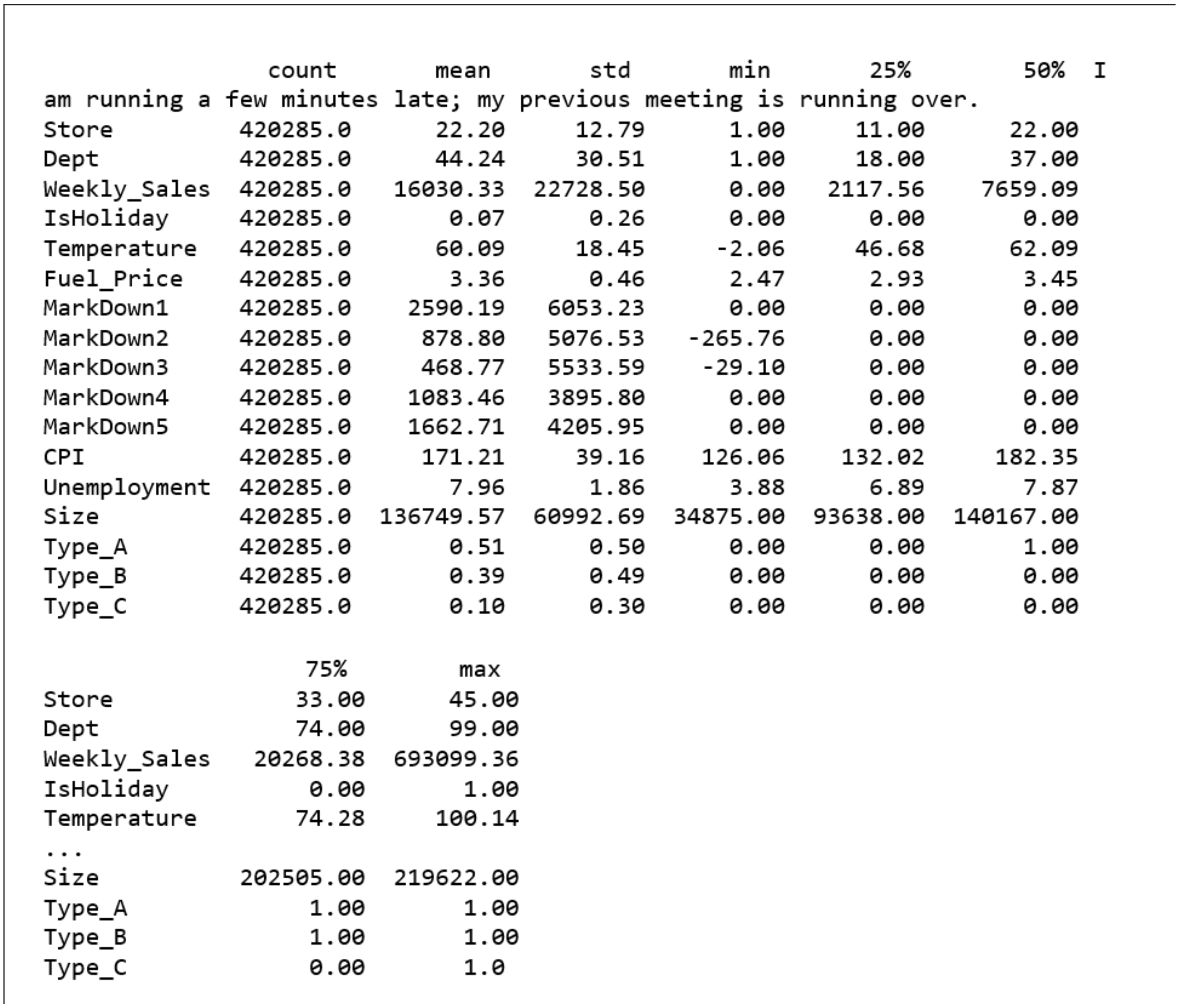
Data integration was followed by data cleaning. The merged data was examined for obvious flows or issues during this step. The initial analysis involved assessing the data for missing values. All of the columns were found to be complete, with the exception of the five Markdown columns. As anticipated, these columns only had values for particular months that are known for increased shopping behavior. Notably, for the weeks with no Markdown values, cells were populated with "NA," which were then replaced with zeros. Because of the absence of duplicates, the preprocessing focus was shifted to checking for negatives. Special focus was placed on the sales column, as sales are typically supposed to be either "0" or positive. A total of 1285 instances of negative values were observed, and the corresponding rows were eliminated. Other columns with negative values included Temperature, Markdown 2, and Markdown 3. Given the typical occurrence of negative values in such variables, negative values were deemed normal, hence obviating the necessity to erase them. DR_NO - Division of Records Number: Official file number made up of a 2-digit year, area ID, and 5 digits.

Data Transformation

The two columns with string data types were then transformed. First to be transformed was the Type column, which was transformed using One-hot encoding to create dummy variables for the three entry types in it contained. The resulting dummy variables, together with the IsHoliday column, were transformed further through Boolean encoding that changed the string data types to integers.

Data Summary

Table 4\ *Data Summaries*



```
                  count       mean       std       min       25%       50%  I
am running a few minutes late; my previous meeting is running over.
Store          420285.0      22.20     12.79      1.00     11.00     22.00
Dept           420285.0      44.24     30.51      1.00     18.00     37.00
Weekly_Sales   420285.0   16030.33  22728.50      0.00   2117.56   7659.09
IsHoliday      420285.0       0.07      0.26      0.00      0.00      0.00
Temperature    420285.0      60.09     18.45     -2.06     46.68     62.09
Fuel_Price     420285.0       3.36      0.46      2.47      2.93      3.45
MarkDown1      420285.0    2590.19   6053.23      0.00      0.00      0.00
MarkDown2      420285.0     878.80   5076.53   -265.76      0.00      0.00
MarkDown3      420285.0     468.77   5533.59    -29.10      0.00      0.00
MarkDown4      420285.0    1083.46   3895.80      0.00      0.00      0.00
MarkDown5      420285.0    1662.71   4205.95      0.00      0.00      0.00
CPI            420285.0     171.21     39.16    126.06    132.02    182.35
Unemployment   420285.0       7.96      1.86      3.88      6.89      7.87
Size           420285.0  136749.57  60992.69  34875.00  93638.00 140167.00
Type_A         420285.0       0.51      0.50      0.00      0.00      1.00
Type_B         420285.0       0.39      0.49      0.00      0.00      0.00
Type_C         420285.0       0.10      0.30      0.00      0.00      0.00

                    75%        max
Store             33.00      45.00
Dept              74.00      99.00
Weekly_Sales   20268.38  693099.36
IsHoliday          0.00       1.00
Temperature       74.28     100.14
...
Size          202505.00  219622.00
Type_A             1.00       1.00
Type_B             1.00       1.00
Type_C             0.00       1.0
```

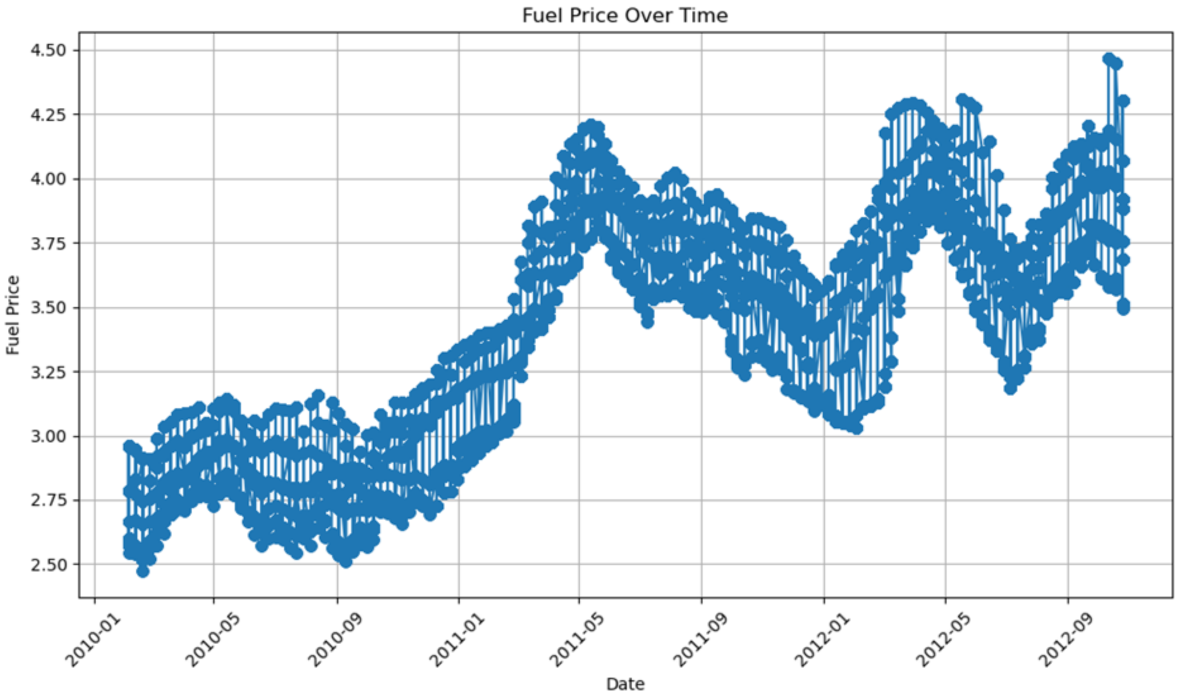
Alt text

The summary above shows that the target variable (Weekly Sales) has a mean value of approximately 16,030.33 and a relatively large standard deviation of about 22,728.50. The minimum weekly sales value is 0.00, while the 25th, 50th(median), and & 75th percentile values are 2117.56, 7659.09, and 20268.38, respectively.

Long Term Tilt Data for Fuel Price Changes

Fuel prices rose steadily between 2010 and 2012, as shown by the Long Term Tilt Data for Fuel Price Changes below. Therefore,for planning reason, it is necessary to expect increased fuel prices in future.

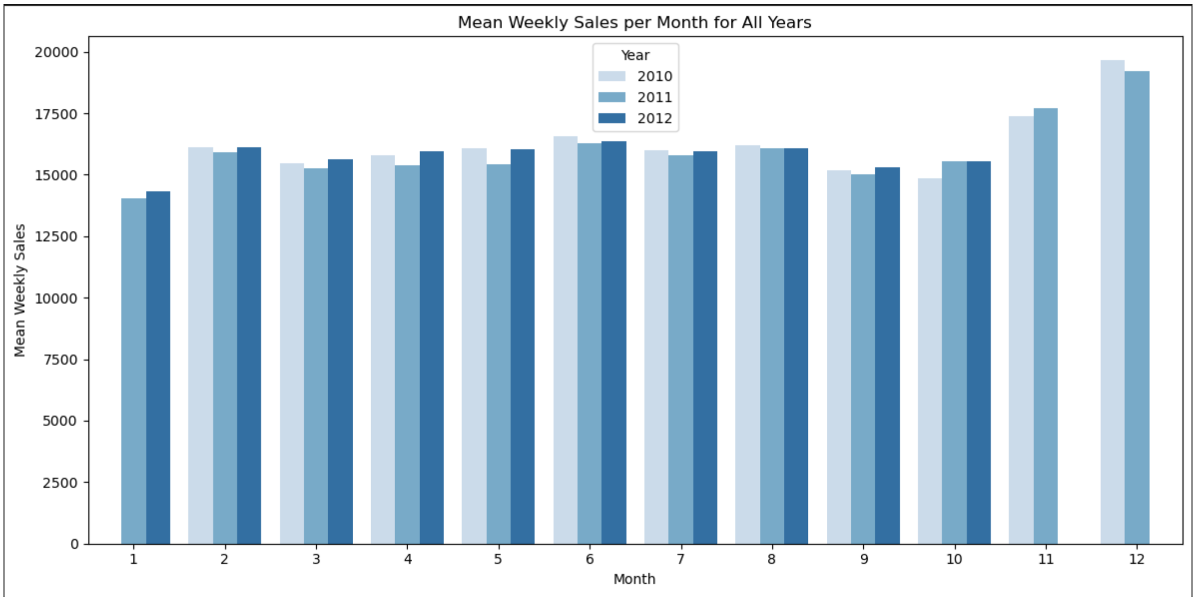
Figure 1\ *Long Term Tilt Data for Fuel Price Changes*



Mean Weekly Sales Per Monther for 2010, 2011, and 2012

The average monthly sales for 2010, 2011, and 2012 are all almost the same from the graph of Mean Weekly Sales Per Monther for those years. One possible explanation for the supermarket chain's relatively consistent monthly sales is the presence of a large number of loyal, repeat customers. Factors such as local events, holidays, and cultural influences may not have changed considerably during this time, and neither may the retail market or industry in which the store works.In addition, it can be seen that the months of January, September, and October have the lowest sales figures in every year. January is usually connected with the winter season, whereas September and October are likely to observe a fall in sales as a result of the conclusion of summer.

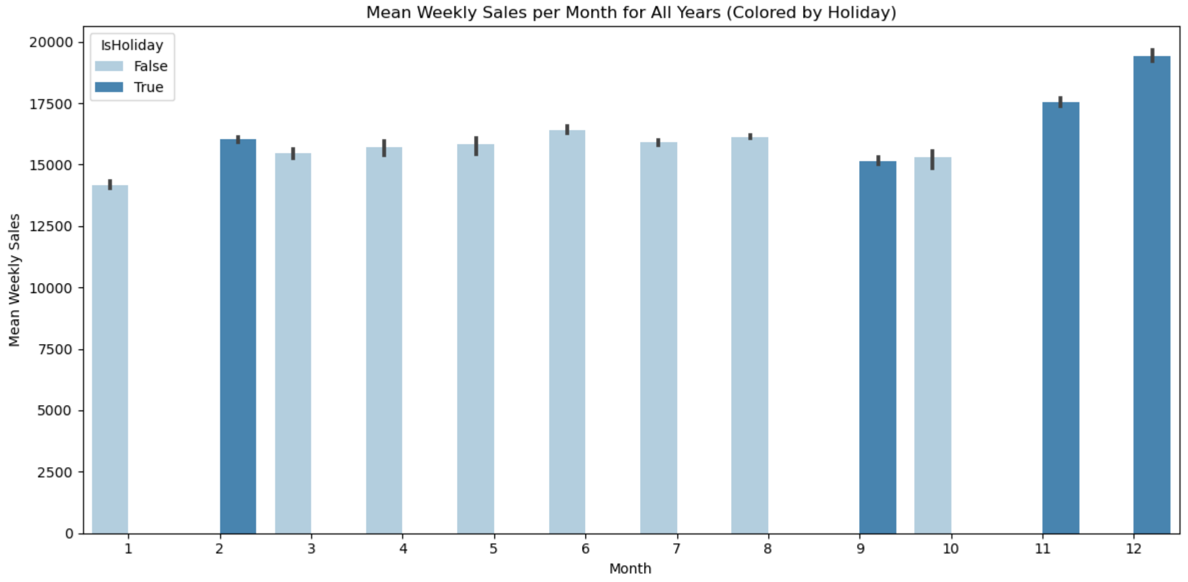
Figure 2\ *Mean Weekly Sales Per Monther for 2010, 2011, and 2012*



Mean Weekly Sales Per Month for 2010, 2011, and 2012 as Influenced by Holidays

Holidays often have a small effect on monthly sales. In actuality, September's monthly sales total is somewhat below the average for the preceding months. Nevertheless, it can be maintained that the reduction in sales during September can be linked to the closure of the summer season in certain places. After the summer is over, people often try to save money by cutting back on seasonal purchases and activities.

Figure 3\ *Mean Weekly Sales Per Monther for 2010, 2011, and 2012 as Influenced by Holidays*



As can be seen in Figure 2 and 3 below, which graphically illustrate annual sales and total weekly sales across different store types, Store Type A continuously displays the greatest sales statistics over the course of three consecutive years, followed by Store Type B. There is a substantially larger difference between Store Type A and Store Type B in terms of annual sales than there is between Store Type B locations over the course of all three years combined. There is no evidence provided to explain the discrepancy in sales results between the various types of stores.

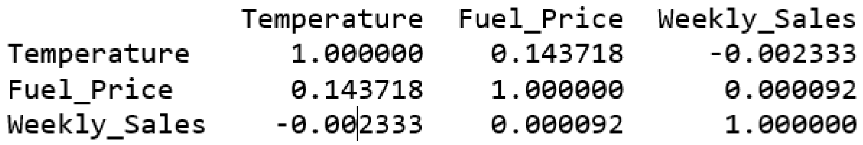
Relationship Between Temperature, Fuel Price and Weekly Sales

Increasing temperatures may lead to a little drop in weekly sales because of the inverse relationship between temperature and business activity. However, the correlation is so weak that it is almost inconsequential, showing that the impact of temperature on weekly sales is limited.

The price of gasoline has a very modest positive correlation with weekly revenue. This may indicate that there is some link between the recent rise in fuel prices and the marginal rise in weekly sales. This weak correlation, however, shows that the impact of gas prices on weekly revenue is quite small. Most likely, the correlation shown here results from chance rather than any real causality.

However, there is a slight positive correlation between fuel prices and temperatures. The data suggests that as temperatures rise, fuel prices tend to follow suit. It's likely that the minor positive association between temperature and fuel price can be attributed to an increase in demand for fuel during hot months, as more people tend to travel during the summer.

Figure 4\ *Relationship Between Temperature, Fuel Price, and MOnthly Sales*



```
              Temperature  Fuel_Price  Weekly_Sales
Temperature      1.000000    0.143718     -0.002333
Fuel_Price       0.143718    1.000000      0.000092
Weekly_Sales    -0.002333    0.000092      1.000000
```

Relationship Between Unemployment and Weekly Sales

There is a statistically significant negative link between unemployment and weekly sales (r=-0.03, p=0.0000). The minor negative link between the two variables, as indicated by the negative correlation coefficient, shows that weekly sales are likely to fall when unemployment levels rise. The little p-value indicates a high degree of confidence that the observed correlation is not due to chance. An increase in the unemployment rate is likely to lead to a decrease in sales because of the negative impact on customer confidence that it has.

Additional Visualization Summaries

Figure 5\ *Total Sales Per year against Store Type*

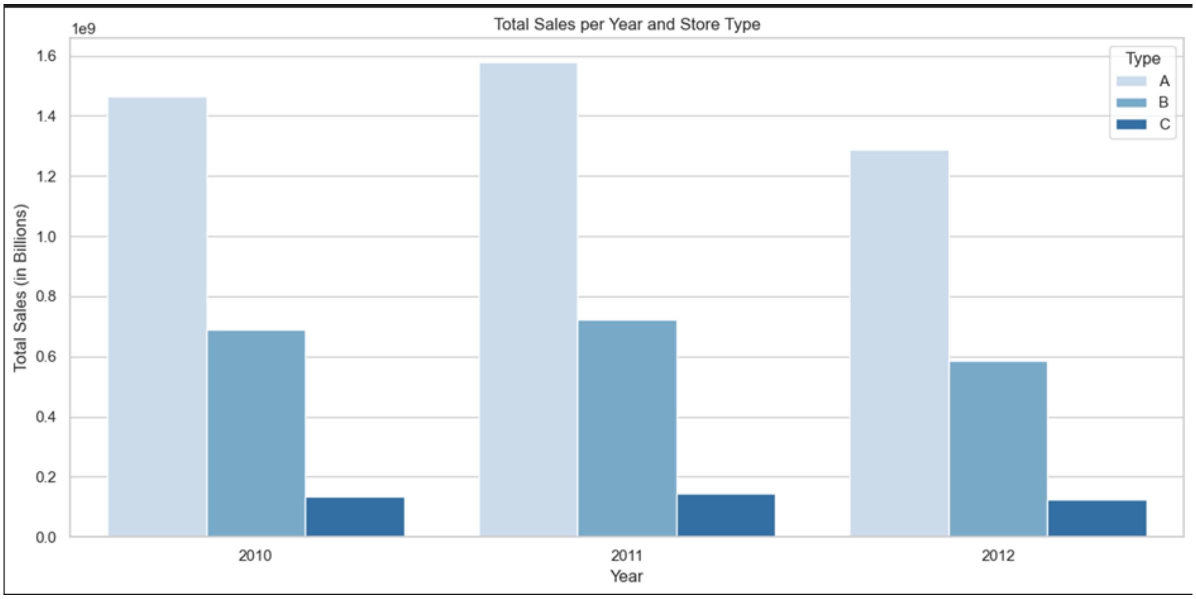


Figure 6\ *Weekly against Store Type*

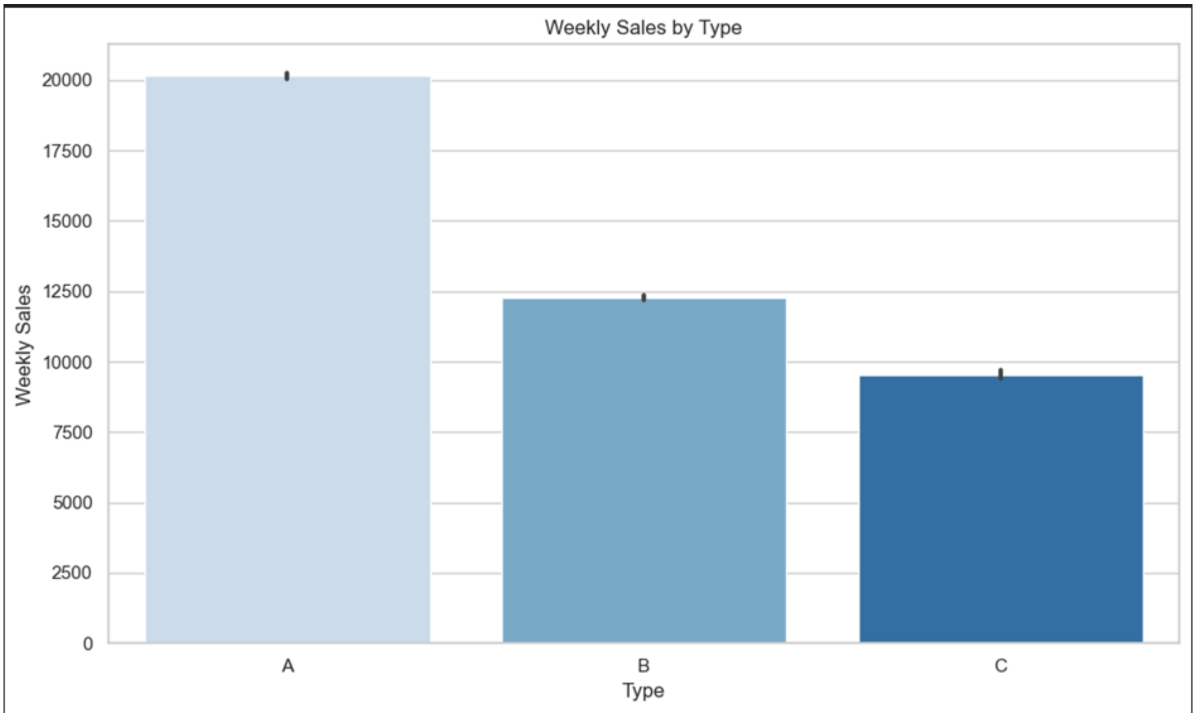
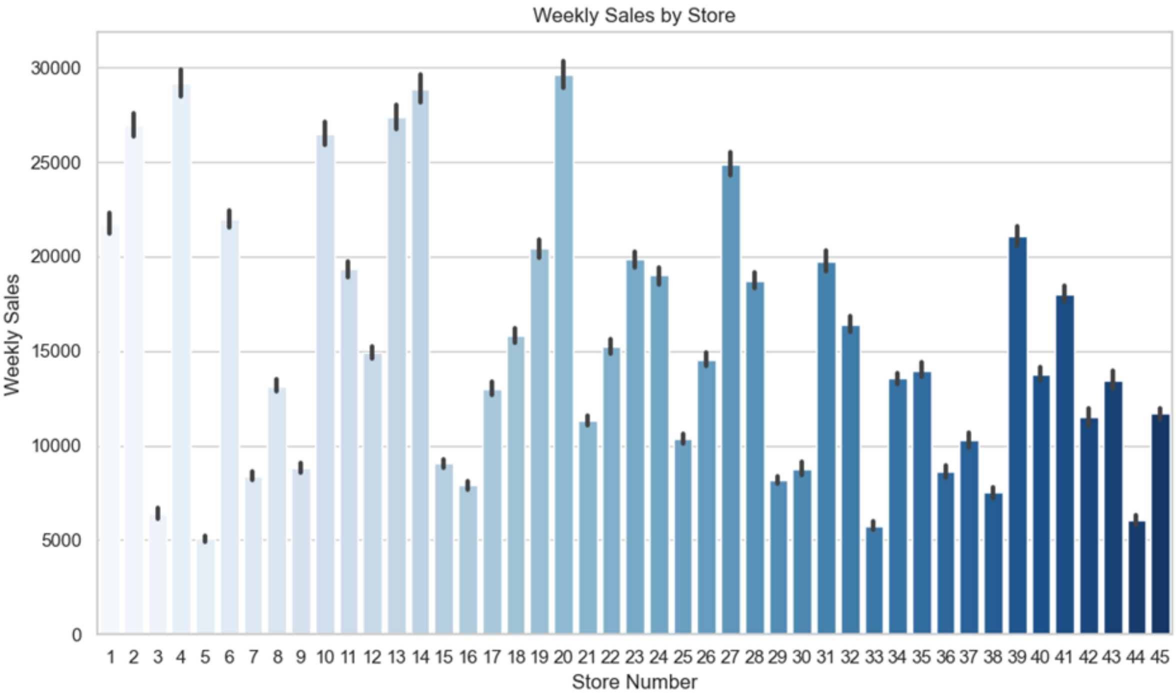


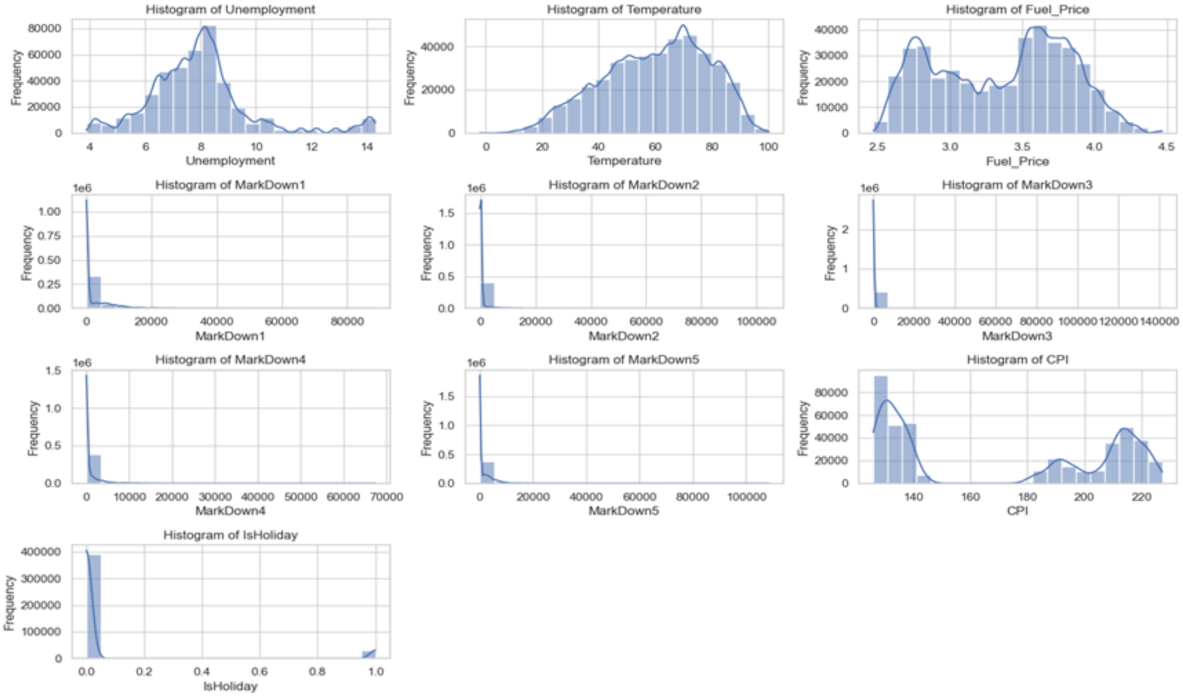
Figure 7\ *Weekly Sales by Type*



Univariate Analysis

From the univariate analysis photos below, it can be noted that Unemployment and Temperatures and the only near-normally distributed variables. Fuel_Price and CPI exhibit bimodal distribution while MarkDowns 1-5 have skewed distribution.

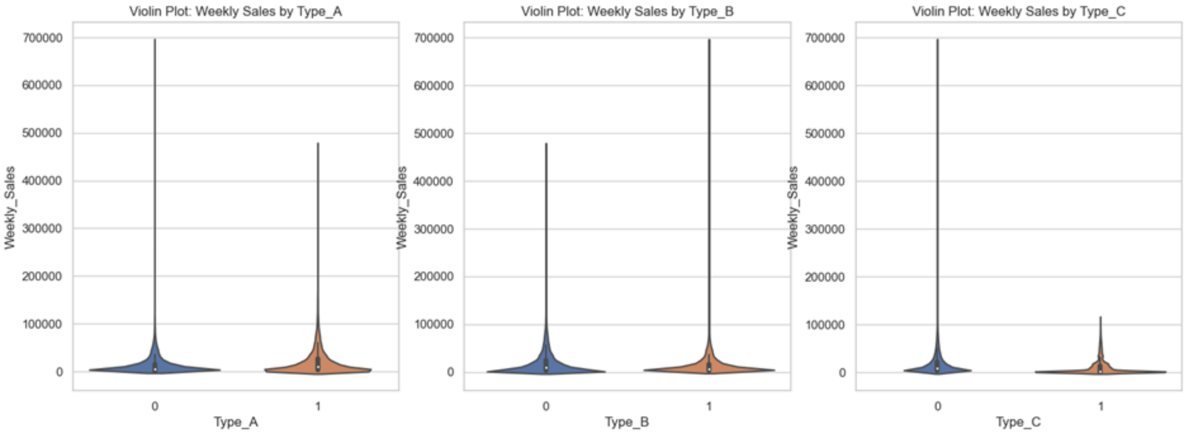
Figure 8\ *Univariate Analysis Tables for all Variables*



Bivariate Analysis

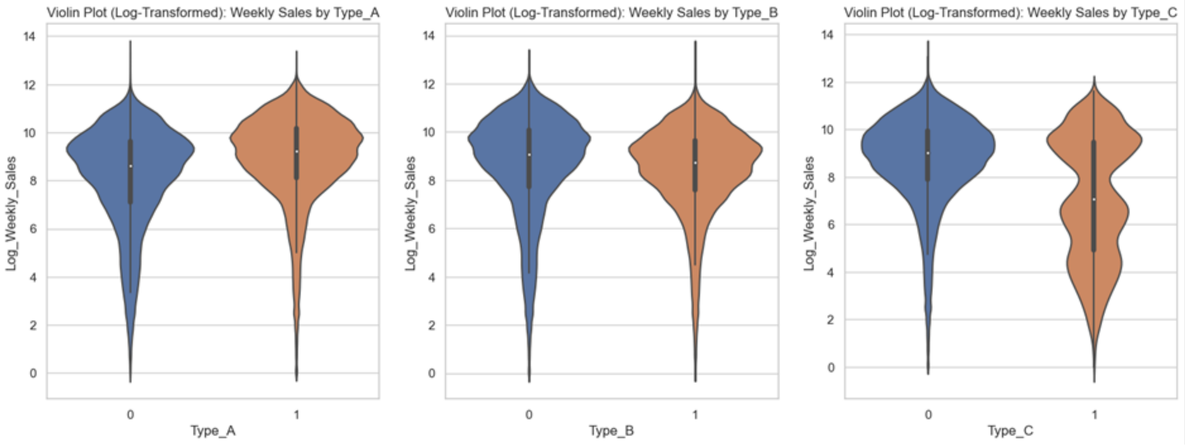
Bivariate analysis for Type variable revealed that it is highly skewed for all gummy variables.

Figure 9 *Bivariate Analysis Tables for Weekly Sales against Type*



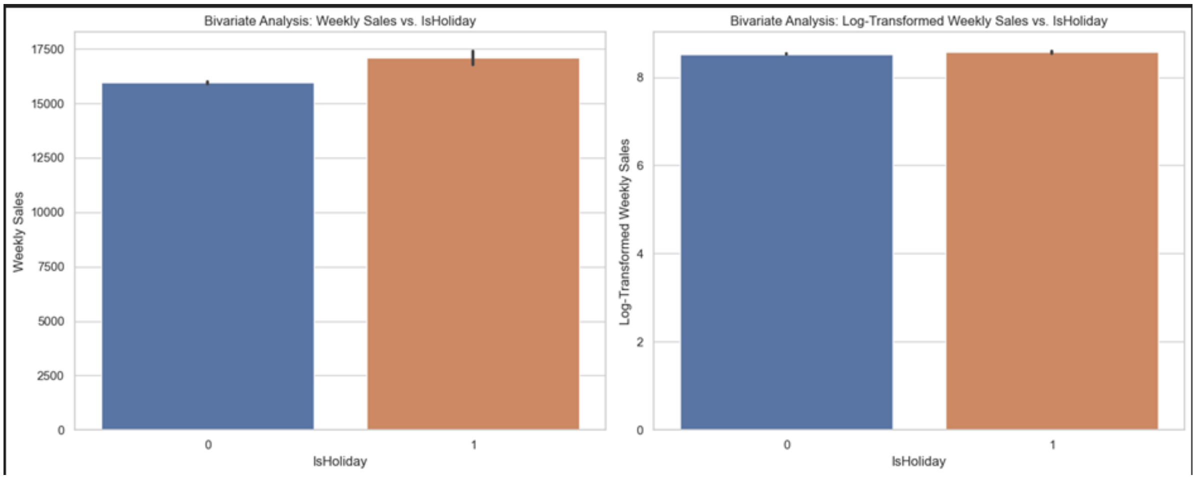
As a corrective measure, the variable Weekly Sales was subjected to log transformation to reduce the impact of outliers or extreme values.

Figure 10 *Bivariate Analysis Tables for Weekly Sales against Type after Log Transformation*



Notable changes in distribution can also be seen in pair plots of Weekly Sales and other variables after logarithmic trans formation

Figure 11 *Bivariate Analysis Tables for Weekly Sales against IsHoliday after Log Transformation*



Distribution has improved in the IsHoliday variable as be seen in figure 9 and 10 below.

Figure 12 *Pair Plots for Weekly Sales against IsHoliday Before Log Transformation*

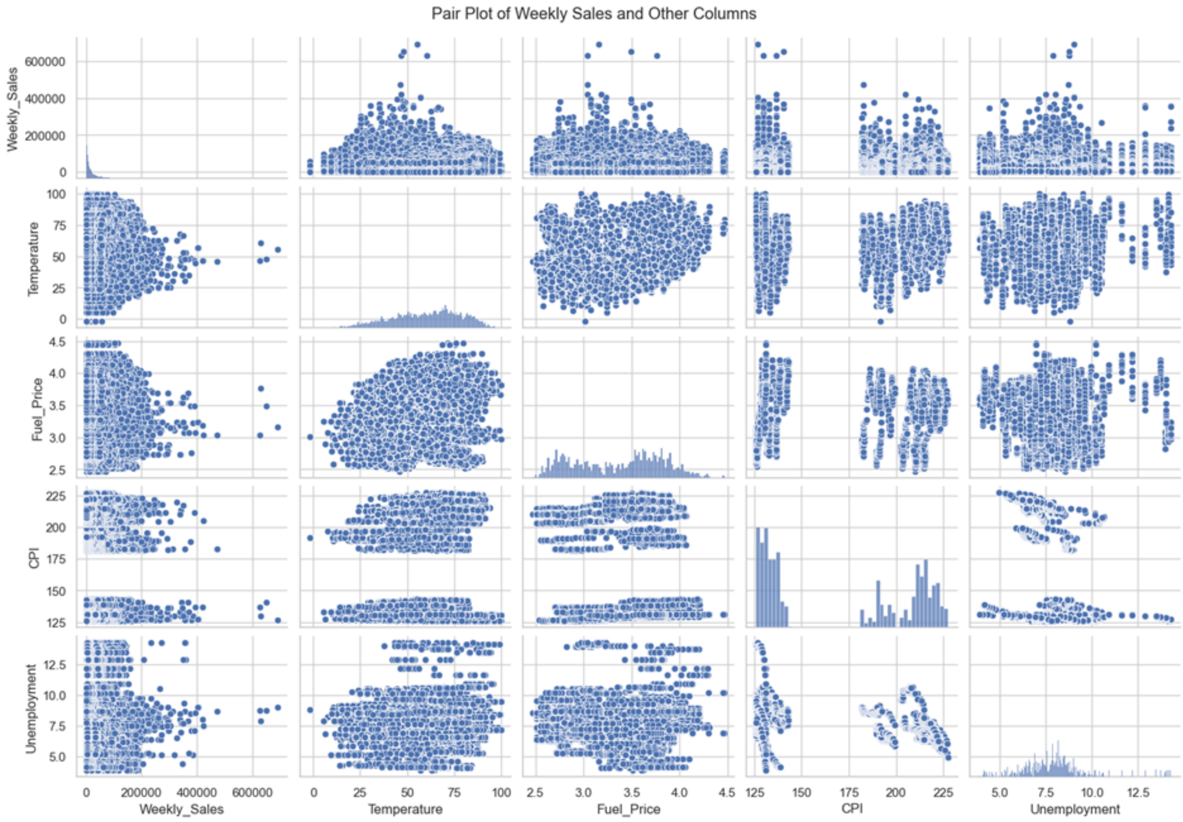
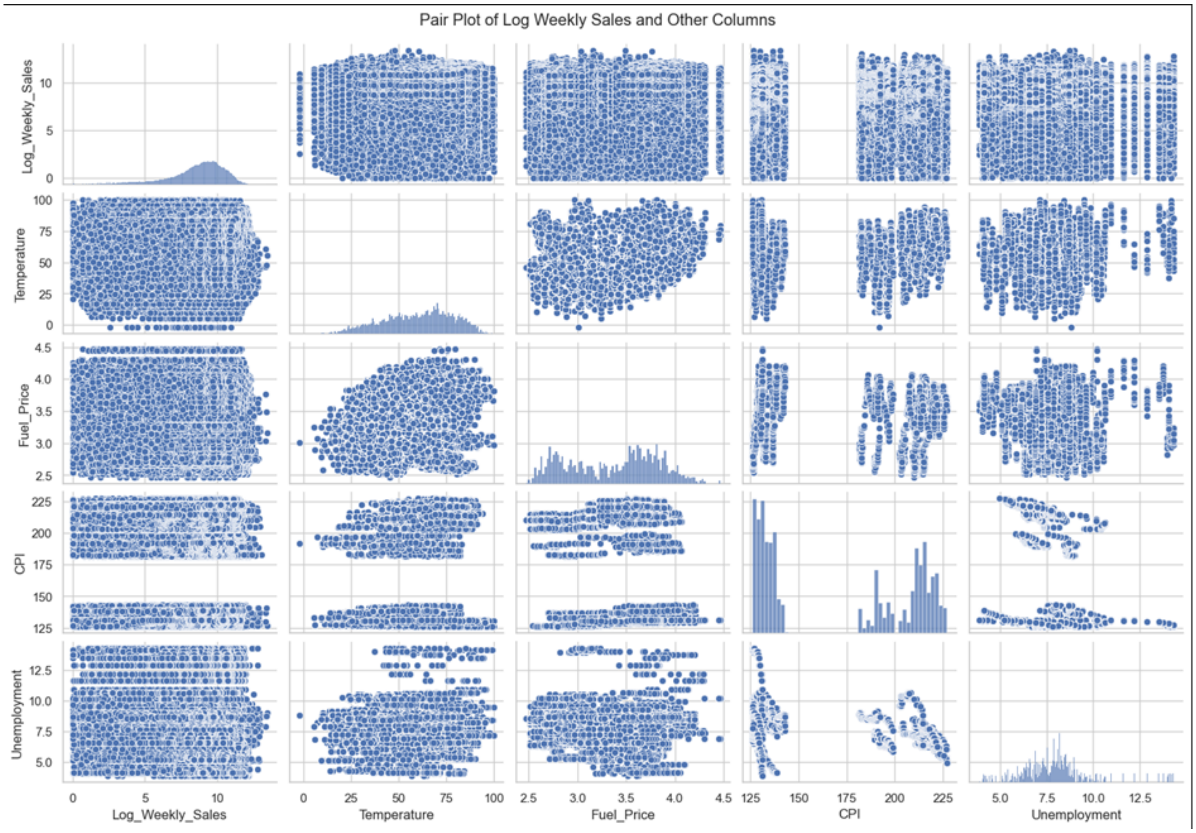


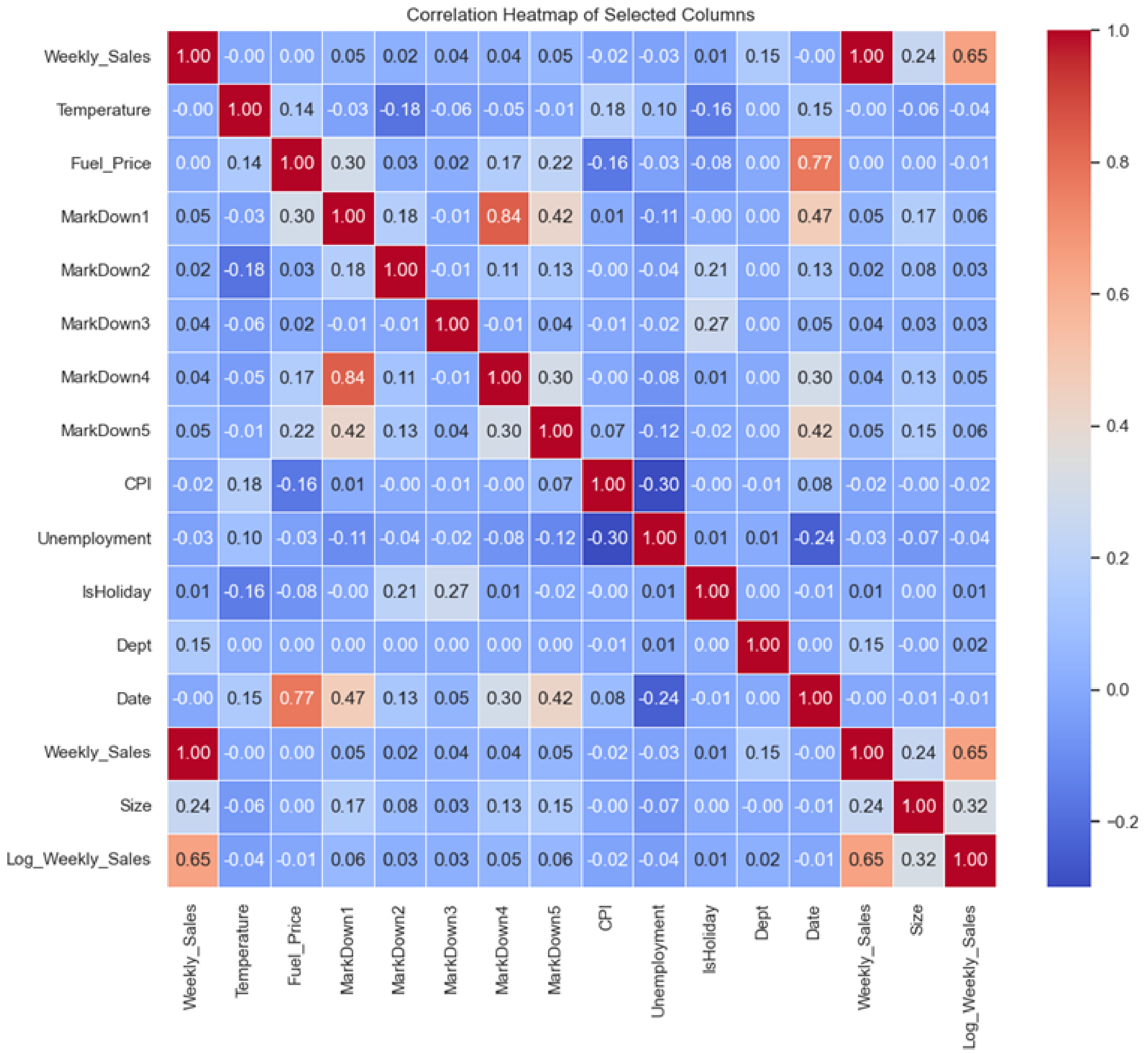
Figure 13 *Pair Plots for Weekly Sales against IsHoliday After Log Transformation*



Correlation Matrix

The was no strong relationships between variables, particularly with respect to the target feature values (Weekly Sales) for this task. It can be noted most of the various do not have strong correlations with the Target value. However, CPI, MarkDown1, MarkDown2, MarkDown3, MarkDown4, and MarkDown5 seem to have very week correlations with Log_Weekly_Sales, hence are dropped for the final data frame for modeling

Figure 14\ *Correlation Matrix*



	Weekly_Sales	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday	Dept	Date	Weekly_Sales	Size	Log_Weekly_Sales
Weekly_Sales	1.00	-0.00	0.00	0.05	0.02	0.04	0.04	0.05	-0.02	-0.03	0.01	0.15	-0.00	1.00	0.24	0.65
Temperature	-0.00	1.00	0.14	-0.03	-0.18	-0.06	-0.05	-0.01	0.18	0.10	-0.16	0.00	0.15	-0.00	-0.06	-0.04
Fuel_Price	0.00	0.14	1.00	0.30	0.03	0.02	0.17	0.22	-0.16	-0.03	-0.08	0.00	0.77	0.00	0.00	-0.01
MarkDown1	0.05	-0.03	0.30	1.00	0.18	-0.01	0.84	0.42	0.01	-0.11	-0.00	0.00	0.47	0.05	0.17	0.06
MarkDown2	0.02	-0.18	0.03	0.18	1.00	-0.01	0.11	0.13	-0.00	-0.04	0.21	0.00	0.13	0.02	0.08	0.03
MarkDown3	0.04	-0.06	0.02	-0.01	-0.01	1.00	-0.01	0.04	-0.01	-0.02	0.27	0.00	0.05	0.04	0.03	0.03
MarkDown4	0.04	-0.05	0.17	0.84	0.11	-0.01	1.00	0.30	-0.00	-0.08	0.01	0.00	0.30	0.04	0.13	0.05
MarkDown5	0.05	-0.01	0.22	0.42	0.13	0.04	0.30	1.00	0.07	-0.12	-0.02	0.00	0.42	0.05	0.15	0.06
CPI	-0.02	0.18	-0.16	0.01	-0.00	-0.01	-0.00	0.07	1.00	-0.30	-0.00	-0.01	0.08	-0.02	-0.00	-0.02
Unemployment	-0.03	0.10	-0.03	-0.11	-0.04	-0.02	-0.08	-0.12	-0.30	1.00	0.01	0.01	-0.24	-0.03	-0.07	-0.04
IsHoliday	0.01	-0.16	-0.08	-0.00	0.21	0.27	0.01	-0.02	-0.00	0.01	1.00	0.00	-0.01	0.01	0.00	0.01
Dept	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.01	0.00	1.00	0.00	0.15	-0.00	0.02
Date	-0.00	0.15	0.77	0.47	0.13	0.05	0.30	0.42	0.08	-0.24	-0.01	0.00	1.00	-0.00	-0.01	-0.01
Weekly_Sales	1.00	-0.00	0.00	0.05	0.02	0.04	0.04	0.05	-0.02	-0.03	0.01	0.15	-0.00	1.00	0.24	0.65
Size	0.24	-0.06	0.00	0.17	0.08	0.03	0.13	0.15	-0.00	-0.07	0.00	-0.00	-0.01	0.24	1.00	0.32
Log_Weekly_Sales	0.65	-0.04	-0.01	0.06	0.03	0.03	0.05	0.06	-0.02	-0.04	0.01	0.02	-0.01	0.65	0.32	1.00

Modeling

Because of their law scores in the correlation matrix, 'CPI', 'MarkDown1', 'MarkDown2', 'MarkDown3', 'MarkDown4', 'MarkDown5'. On its part, 'Weekly_Sales' was dropped because it had early be logarithmically transformed and replaced with ‘Log_Weekly_Sales’, which was selected as the target variable. The dataset was then split split into training and testing sets, using an 80/20 ratio. • Rpt Dist No - Code that represents a sub-area within a Geographic Area.

Four algorithms were chosen for training and prediction. They included Linear Regression, Decision Tree, Random Forest, and Support Vector Regression. Linear Regression was chosen for its simplicity and widespread usage in modelling linear data. However, since the sales data was not linear, it was expected to have least performance statistics. Random Forest Classifier: The Random Forest classifier is an ensemble learning technique that combines multiple decision trees to make vigorous predictions. It's well-suited for classification tasks and is skilled to handle complex datasets with high-dimensional features. We will use it to predict categorical outcomes, such as store types, based on numerous features from the dataset. Support Vector Machine (SVM) Classifier: The Support Vector Machine is a dominant classification algorithm that purposes to find the best hyperplane to separate different classes. SVMs are operative for both binary and multi-class classification tasks. In our analysis, we will leverage SVM to categorize stores into different types based on particular features.

The dataset was split into training, validation and testing sets, using an 80/20 ratio. This split was selected to maintain a balance between a satisfactorily large training dataset for vigorous model development and a reserved portion for model evaluation.The training set, encompassing 80% of the data, was used for training machine learning models to study from historical patterns and relationships within the dataset (Guu et al, 2020). In contrast, the testing set, signifying the remaining 20%, served as an unobserved dataset to assess model performance and overview to new, unseen data.

Linear Regresion

The R-Squared score, also known as the coefficient of determination, of 0.117 for Linear Regresion suggests that the model explains only 11.7% of the variance in the sales data. In other words, the model's predictions do not fit the data very well, as a higher R2 score is desired for better predictive performance.

Decision Tree

The algorithm has a high R-squared (R2): 0.9540453930332063, suggesting that it is able to predict 95% of the test set accurately. Efforts to fine tune the model using GridSearchCV were futile. GridSearchCV showed that the best parameters are:\ Max Depth: 30\ Max Features: 'sqrt'\ Min Samples Leaf: 10\ Min Samples Split: 2\ However, the R-Squared score has dropped to 0.5997, which is lower that the previous score of 0.9545

SVM

SVM did not perform well. The algorithm has a an extended run time hence was stopped before it completed modeling.

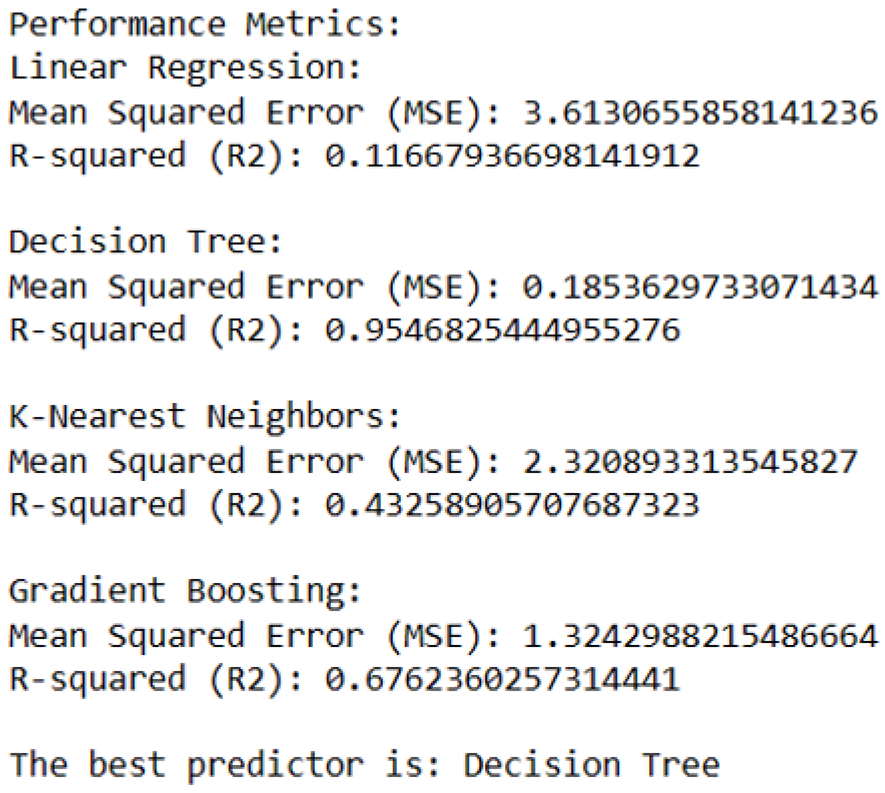
KNN

KNN had relatively large error values and a low R-squared value (0.42951) making it an suitable for this modeling

Best machnie learning model

The classification report which is generated by using the Machine Learning models says that the Logistic Regression, Decision Tree and KNN Model has almost the same accuracy value i.e around 70%. So all the three models are good for this data. Random Forest is also a good model but a bit less when we compared with the rest three models.

Figure 15\ *Model Comparision*



```
Performance Metrics:
Linear Regression:
Mean Squared Error (MSE): 3.6130655858141236
R-squared (R2): 0.11667936698141912

Decision Tree:
Mean Squared Error (MSE): 0.1853629733071434
R-squared (R2): 0.9546825444955276

K-Nearest Neighbors:
Mean Squared Error (MSE): 2.320893313545827
R-squared (R2): 0.43258905707687323

Gradient Boosting:
Mean Squared Error (MSE): 1.3242988215486664
R-squared (R2): 0.6762360257314441

The best predictor is: Decision Tree
```

Conclusion

There exists a negligible correlation between Weekly Sales, Fuel Prices, and Temperature. Nevertheless, there exists a statistically significant association between the factors such as Unemployment and Weekly Sales. At the same time, it is evident that variables such as markdowns and holidays have no substantial impacts on sales.

Future Work

It is important further investigations to be done to find out why obvious predictor of high sales like Markdowns and Holidays had very small correlation values with Weekly sales.

References:

[1] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons, Inc.

[2] Jadhav, P. V., & Patil, V. V. (2022). Application of decision tree for developing accurate prediction models. Lulu Publishers