Final Project - Retail Store Analysis

Caden Goodwin, Bradley Goulart, Long Tran

Introduction

Our dataset gives us data from 45 department stores in 2012. Our goal is to find out trends in this data and provide our findings to the different stores on which promotions work best to optimize sales in different departments. The data was from Kaggle and sourced from walmart stores.

Variables:

Store: The store number. Date: The week the record is for. Temperature: The average temperature in the region. Fuel_Price: The cost of fuel in the region. MarkDown1-5: Anonymized data related to promotional markdowns. CPI: Consumer Price Index. Unemployment: The unemployment rate. IsHoliday: Whether the week is a special holiday week. Dept: The department number. Weekly_Sales: Sales for the given department in the given store. Type: Type of store (categorized into A, B, or C). Size: Size of the store.

Question 1: How can we predict weekly sales for each department using variables like store size, type, regional factors (temperature, fuel price, CPI, unemployment), and holiday information?

I used random forest regression analysis along with RMSE, R squared, and visualizations to show the factors that influence weekly sales in the retail dataset.

```
Root Mean Squared Error (RMSE): 4452.854948608919
R^2 Score: 0.962334722648165
```

Figure 1: RMSE

I got a RMSE score of 4452.85, which means the model's predictions on average are within approximately $4452.85 of the actual sales values. Considering the scale of weekly sales, which ranges into the hundreds of thousands, this level of error is relatively low. This is shown by a very high RSquared Score of 0.9623, meaning my model explains over 96% of the variance in weekly sales. Such a high R^2 value shows I have a strong model that fits the data well.

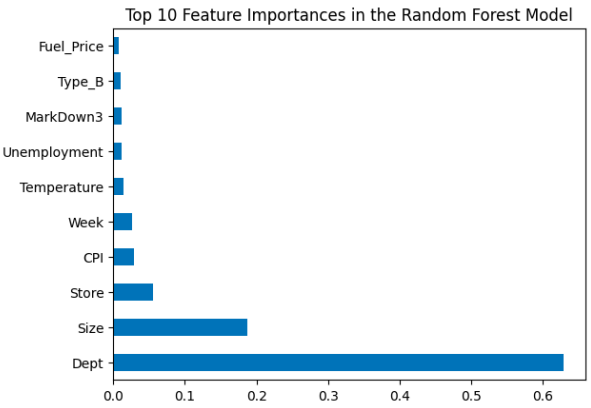


Figure 2: Feature Importance Plot

The first 3 most significant factors are internal store factors which were department, size of the store, and the store number.

Dept (Department Number): The plot showed that the department number is the largest factor for sales. This makes sense since sales can vary a lot between many different departments. Size (Store Size): The second most important factor was store size. This tells us that the size of the store is a major factor in sales, possibly due to larger stores having a wider product range. Store (Store Number): This feature's importance shows us that there are store specific factors affecting sales not captured by other features in the model. This could tell us that sales are influenced by a store's location.

CPI (Consumer Price Index): The CPI's significance tells us that general economic conditions, which can affect consumer purchasing power, are also important for predicting sales. Possibly telling us in times where the economy is good consumers are more likely to spend more. If we just take external factors into account, meaning we don't look at the departments, store size, or store number, and just look at the factors outside of the store itself, then this would be the most influential factor.

Week: The importance of the week indicates there are likely seasonal patterns or specific times of the year that are crucial for sales.

Other factors like 'Temperature', 'Unemployment', 'MarkDown3', 'Type_B' (indicating store type), and 'Fuel_Price' have lesser but still important influences on the model's predictions.

From this plot, the takeaway a store can make is that internal factors like the department sales are in, the store size, and what store location they are, are the main influences in sales. Stores can look at this and realize that some other stores have larger amounts of sales because they might have more departments, a larger store size, and a more prime location. Taking internal factors aside, they can see that CPI and Week are the leading external factors. So they should keep in mind the current CPI, and adjust prices in times where the economy isn't as good to offset this, or employ more in times where the CPI is high. Week is another important determinant of sales, so the company should employ more during high sale weeks. I'll analyze more season trends in my second question analysis.

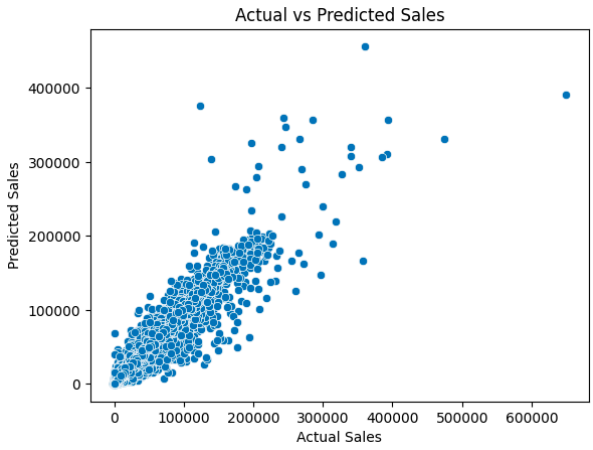


Figure 3: Scatter Plot Analysis

The scatter plot shows that there's a strong positive correlation between actual and predicted sales, which is a good indication that the model can predict sales with a high level of accuracy. There's a few outliers mostly in the higher values of sales where the model does not predict as accurately. This could be due to the complexity of sales that are not fully captured by the model. However overall this scatter plot is a good indication of a strong positive correlation and most of the time the predicted sales is in line with actual sales.

The stores should keep in mind the predicted sales, and employ and stock accordingly due to predicted sales being very accurate to how much they will actually make in sales. However they should take the higher predicted sales with a little grain of salt because as mentioned, there can be outliers as shown in the graph when it predicts a really high sale amount as those can be a bit more unpredictable due to our model not having enough information from these high sale numbers to give an accurate prediction.

Question 2: How significant is the impact on holiday weeks vs non holiday weeks, and what holidays contribute the most to high sales periods?

I compare the sales distributions across various holidays and between holiday and non-holiday periods as well as comparing different holiday periods to see which holidays are the most impactful. The goal of this question is first see how influential holiday seasons really are, and then see which holidays exactly bring the largest amount of sales so the stores can prepare by stocking shelves and hiring more employees around these times.

The sales data was merged with store information and prepared by creating indicators for each holiday based on the month and day that matched big sales holidays like thanksgiving, christmas, labor day, and the super bowl.

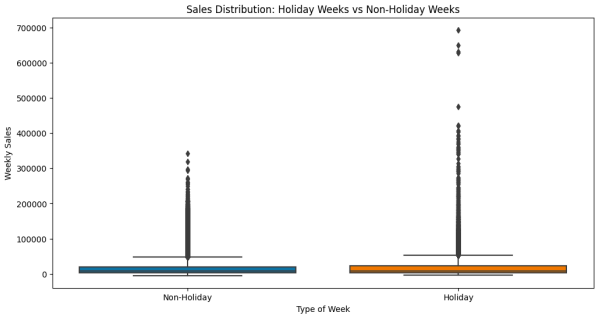


Figure 4: Holiday vs Non-Holiday Weeks

The first box plot shows the sales distribution for holidays compared to non-holiday periods. Some important findings were:

The median sales during holidays are higher than non-holiday weeks, showing holidays have a positive impact on sales.

There's some outliers which indicate that there are peak sales periods associated with holidays that significantly exceed average sales levels. This could be that certain holidays are significantly higher than the average boost sales get from a typical holiday period.

To show more specificity in what holidays are causing this boost and some holidays that might be more baseline to a typical non holiday sale time, I created another plot that shows the specific high sale holiday seasons shown below.

Some important findings from the second box plot that show the specific holiday periods were:

Variability in Sales: There's a noticeable variability in sales during the holidays. This is shown from the spread of the data points and the length of the whiskers in the box plot. Thanksgiving

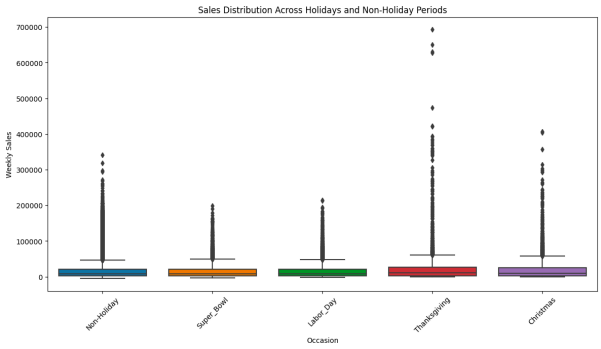


Figure 5: Sales Distribution Across Holidays and Non-Holidays

and Christmas have a wider distribution, which could mean an increased promotional activities or seasonal shopping behaviors around those times.

Median Sales Differences: The median sales (the line within the box) during holiday weeks appear to be higher compared to non holiday weeks. This shows us that on average holidays tend to boost sales.

Outliers: There are several outliers for each holiday, showing that there are weeks with even higher sales during the holidays, but also in the non holiday category there's also high outlier points. This could mean there's successful promotional campaigns or other events that are not influenced by just holidays.

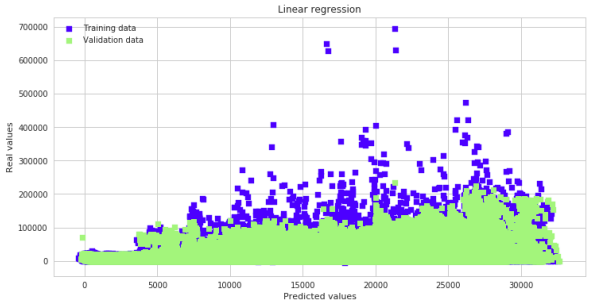
Retailers can use this information to stock up more on inventory and staff during holiday seasons, especially for the holidays that drive a much higher than average sales like thanksgiving or christmas.

Although holidays are a large driver of sales, The non holiday column showed us that there's also spikes for non holiday periods suggesting that there are other factors for certain non holiday times of the year that also drive sales like a non holiday related discount. My group members take a look at these other factors like markdowns in their analysis.

Question 3: How can we predict future weekly sales?

I used the variables Store, Dept, Temperature, Fuel_Price, CPI, Unemployment, IsHoliday and Size to help predict Weekly_Sales. First I used a logistic regression with these variables with a 20/80 split.

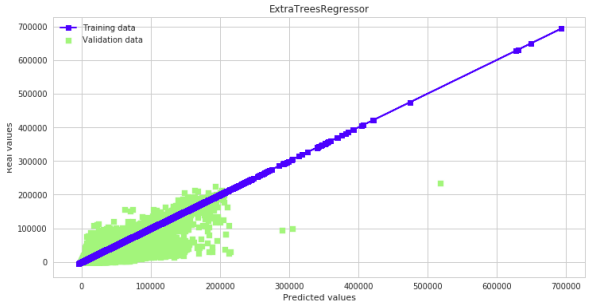
Results of Logistic Regression:




```
Coefficients:
 [ -9.08863111e+01   1.08346205e+02   1.48258168e+01  -7.80356148e+01
  -2.10052743e+01  -1.10786134e+02   1.32255758e+03   8.81126991e-02]
Mean squared error: 444775379.20
Variance score: 0.09
```

This model was only able to predict about 9% of the data when the test set was introduced, which is a very bad model. It averaged a $444 million error. Since this model performed very poorly I used an extra trees model with the same variables and a 20/80 split.

ExtraTrees model results:



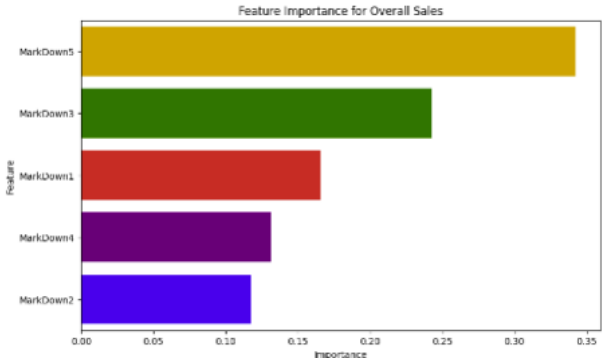

```
ExtraTreeRegressor Score: 0.89227072332
Mean squared error: 52713507.09
```

This model performed much better with 89% of the variance explained when the test set was introduced, having only a $52 million error on average, which is a huge improvement from a $444 million error.

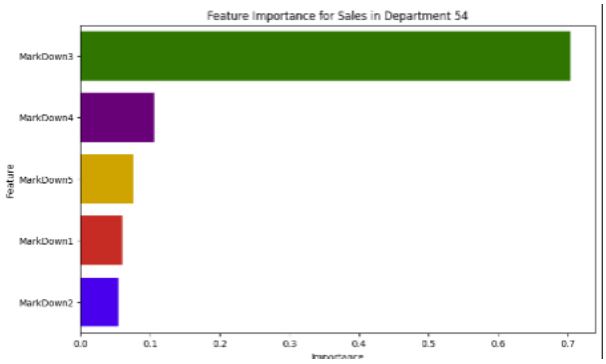
After comparing the models, we see that the extratrees one performs much better. Using this model, we will be able to predict sales in the future, which will help the stores to prepare inventory in a timely and accurate manner.

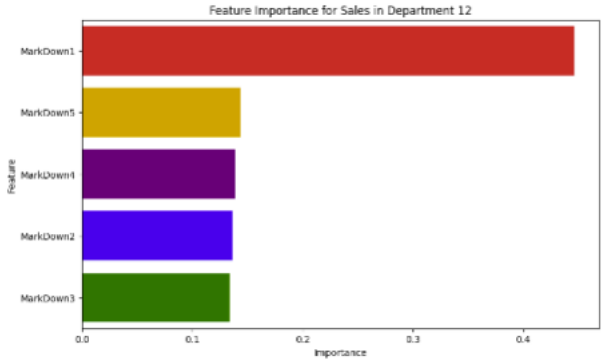
Question 4: Which markdown types 1-5 are most effective in increasing overall sales? Are there better markdown promotions for certain departments? What are the departments and markdowns that work the best to improve weekly sales?

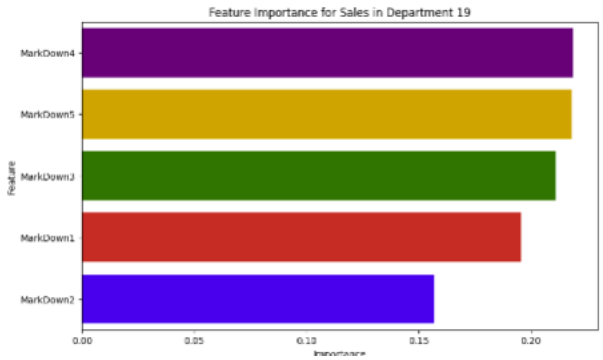
To address this question, I did feature importance on the different markdowns to see which ones affected Weekly_Sales the most. The variables used were Weekly_Sales, MarkDown1-5, and Dept. I found the best markdown was 5, 3, 1, 4, then 2. With 5 being used significantly more than the others and having a larger importance when affecting Weekly_Sales:



We were able to see the importance over all of the departments, but I also figured that each department uses different MarkDowns that are better than others. So I did feature importance on the markdowns for each separate departament. Here are 3 of 99 of the different departments:







As I thought, each department has a certain markdown that is best to help increase weekly sales. We can see that some of the departments use only one primary MarkDown and some of them use a combination of discounts.

Using the feature importance for each department, the stores will be able to know what they should use that have previously worked in the past. They can also experiment using different markdowns or combinations of markdowns and see how that will compare to the past data to increase their sales.

Question 5: What is the relationship between unemployment rates and sales performance in individual departments? Do certain departments thrive in regions with lower unemployment, while others perform better in higher unemployment areas?

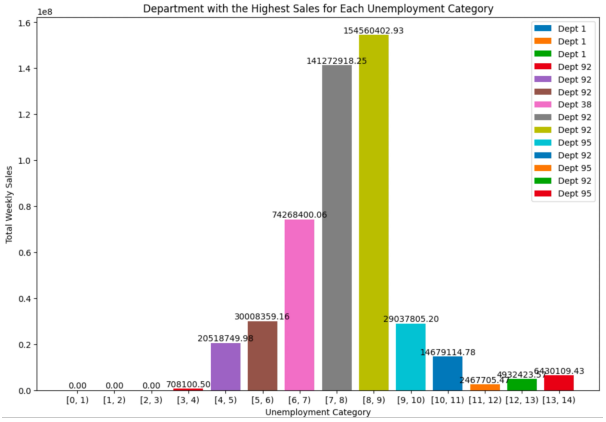
Initially I had to figure out what these store types are so I decided to analyze the data between these stores to find out the differences. I found out that Type A stores, has an average store number of 21.74, are larger and achieve weekly sales from $39,690 to an impressive $474,330.10 (mean: $182,231.29). Type B stores (average store number: 18.45) fall in the middle, with moderate sizes and weekly sales ranging from $34,875 to $693,099.36 (mean: $101,818.74). Type C stores (average store number: 38.94) have the smallest sizes, with weekly sales ranging from $39,690 to $112,152.35 (mean: $40,535.73). Temperature averages follow a similar order, with Type C being the highest (67.55), followed by Type A (60.53) and Type B (57.56). In essence, Type A stores excel in size and sales, while Type C stores, despite higher store numbers and smaller sizes, achieve lower sales. This detailed exploration provides a comprehensive understanding of each store type's unique characteristics and performance metrics.

For this analysis, I decided to look at clustering. After clustering, I found out that Cluster 2 seems to be doing the best in terms of sales, with an average of $65,951.43 per week. Cluster 1 is in the middle, with around $13,544.54 in weekly sales, showing decent performance. On the flip side, Cluster 0 has the lowest average sales, sitting at $9,281.12 per week, indicating a quieter market. Now, talking about unemployment rates, Cluster 1 has the highest at 13.83%,

hinting at some economic challenges. Cluster 0 has a lower unemployment rate of 7.85%, pointing to a more stable economic situation. Interestingly, Cluster 2, despite leading in sales, also maintains a low unemployment rate of 7.85%, which is kind of unique. So, it looks like there's a connection between how well things sell and what's happening in the local job market across these different clusters.

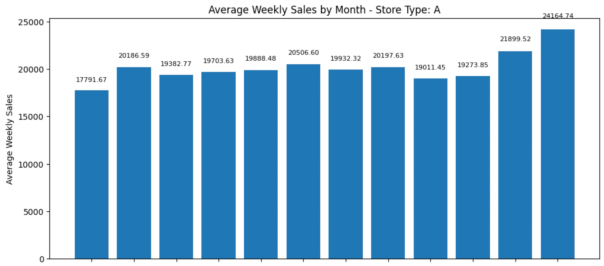
```
Average Sales by Cluster:
Cluster
0     9281.115625
1    13544.539548
2    65951.428028
Name: Weekly_Sales, dtype: float64

Average Unemployment Rate by Cluster:
Cluster
0     7.851773
1    13.831295
2     7.846535
```

After clustering, I decided to look at which departments have the most sales in each specific region. When the unemployment rate was 3-4%, Department 92 had the most sales of $708,100.50. As the unemployment rate increased to 4-9%, the sales from Department 92 skyrocketed into the multimillion-dollar range, hitting an impressive total weekly sales of $154,560,402.93 when the unemployment rate reached 8-9%. However, after this peak, there's a significant decline, which is pretty standard when unemployment rates drop, although they're still higher than areas with 3-4% unemployment. One reason for this could be shoppers from regions with unemployment rates of 3-4% seem to prefer store type A over other types. While these shoppers may buy pricier products, the sheer number of customers from regions with higher unemployment makes a notable impact on total sales.

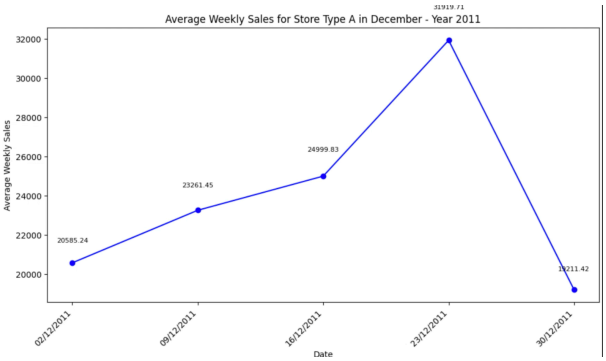


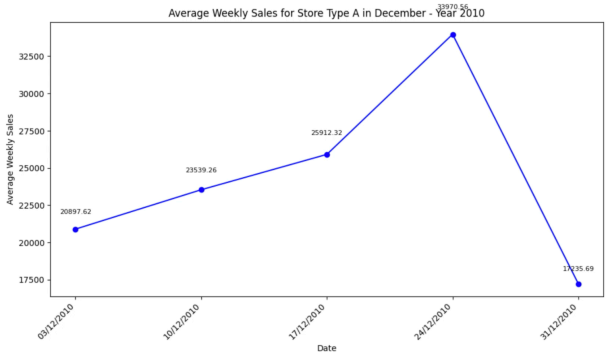
Question 6: How do different store types respond to special holiday weeks? Do certain store types experience more significant increases in sales during holidays, and are there specific departments that contribute more to this boost?

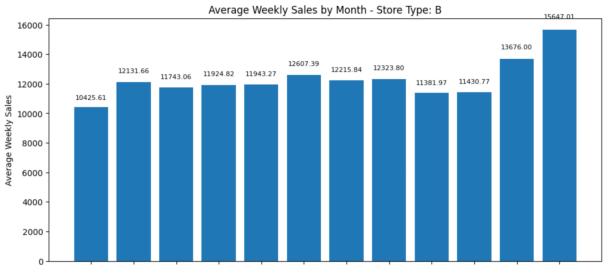
Store Type A sees fluctuations in average weekly sales throughout the year, with sales generally rising from January to December, peaking at $24,164.74 in December. Various factors, such as special promotions or holiday shopping may contribute to this increase. The peak in November and December aligns with the holiday season, suggesting a substantial boost in sales during festive times. Understanding these monthly patterns is crucial for inventory management, strategic marketing planning, and optimizing staff levels during peak sales periods.



In December of 2010 and 2011, Store Type A exhibited interesting sales patterns. Average weekly sales steadily increased, reaching a peak around Christmas day, indicating heightened spending during the holiday week. In 2010, sales almost doubled from the beginning of December to the week of Christmas, with a similar trend observed in 2011, featuring a notable spike in sales on December 23. Overall, department 92 has the highest sales within store type A.

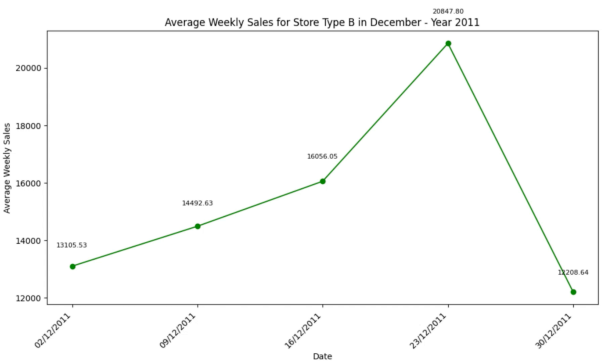


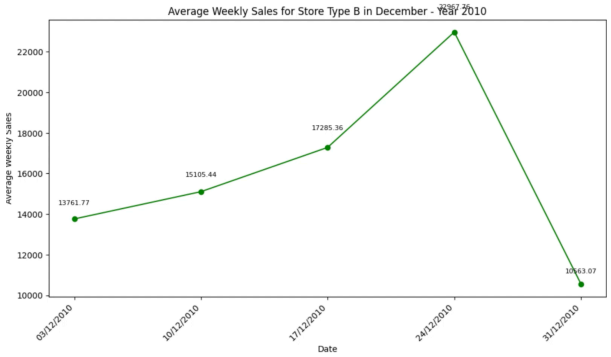


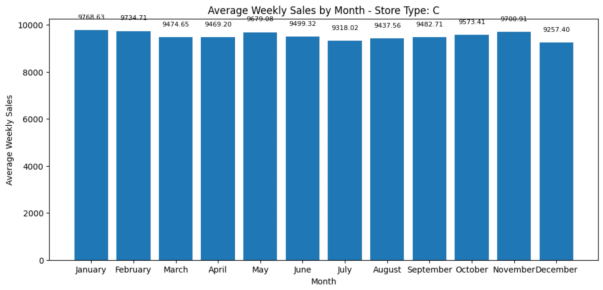


Store Type B shows a consistent upward trend in average weekly sales throughout the year, starting at $10,425.61 in January and reaching a peak of $15,647.01 in December. The significant increase in November and December points to strong performance during the holiday season, a time when consumers tend to spend more.

For Store Type B in 2010 and 2011, average weekly sales also exhibited a consistent upward trend in December, with a substantial increase leading up to the Christmas week and peaking on December 24. This suggests that Store Type B, similar to Store Type A, experiences a surge in consumer demand during the holiday season, particularly in the days just before Christmas. Overall, department 72 has the most sales within this store type.







Store Type C maintains relatively stable average weekly sales, ranging from approximately $9257.40 to $9768.63 throughout the months. This stability may indicate a more predictable consumer demand or a specific market niche. However, achieving growth and increasing sales could pose a challenge for Store Type C, given the steady pattern suggesting limited seasonality or promotional peaks. Comparing average weekly sales data for Store Type C in 2010 and

2011 with Store Types A and B reveals distinctive patterns. Store Type C exhibits a more moderate range of average weekly sales, differing from the higher averages observed in Store Types A and B during the same periods. Overall, department 72 has the highest sales within store type C.