What Does Walmart's Historical Store Sales Data Tell Us?

Use historical markdown data to predict store sales

Yougui Xiang
May 10 · 5 min read

Introduction

Given you have been provided with the past 3 years of historical weekly sales data for 45 Walmart stores located in different regions and each store contains many departments.

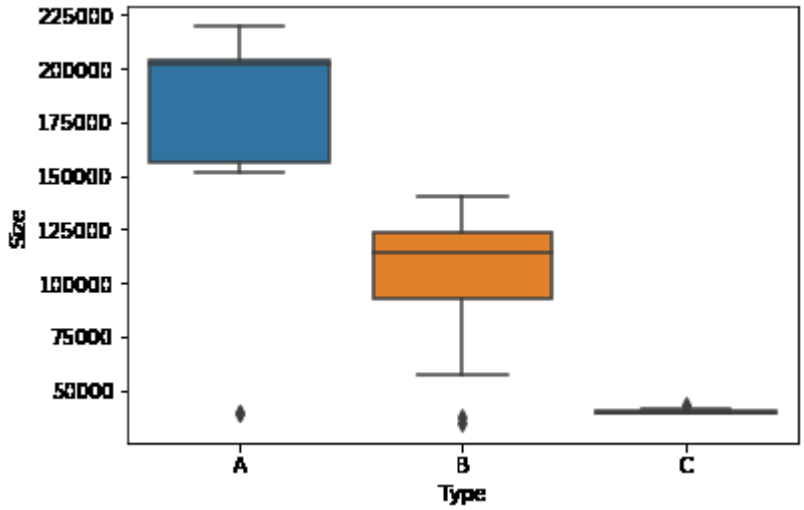
1. How can you utilize these data to predict the next year sales for each department in each store ?
2. In the historical weekly sales data, selected holiday markdown events are included. These markdowns are known to affect sales, but how to predict which departments are affected by these markdowns?
3. And to what extent do these markdowns impact weekly sales?

These challenge questions are coming from Walmart Recruiting — Store Sales Forecasting competition, answers to these questions are valuable for business decision making for big companies such as Walmart. In this blog, I will demonstrate how I approach these questions.

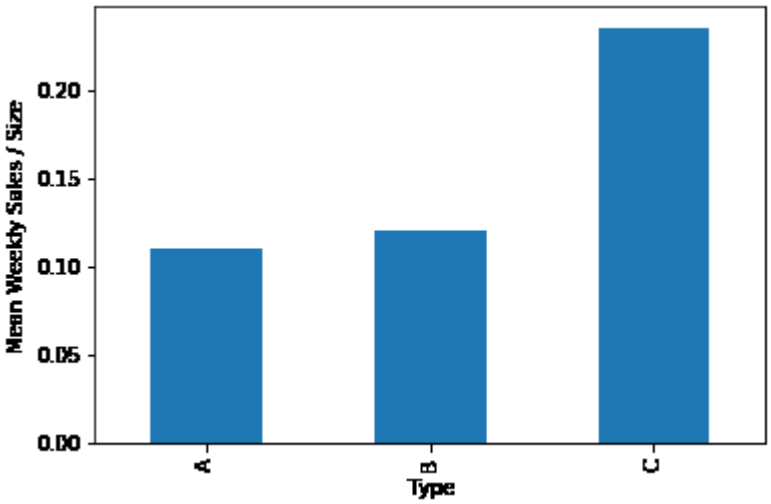
Part I: Data visualization and exploration

In this part, let us explore and visualize this Walmart Store Sales Forecasting data first. Data visualization and exploration will give us some hints on how to solve the above questions. The time frame of train data is from 2010–02–05 to 2012–10–26, and test data is from 2012–11–02 to 2013–07–26. Markdown events only occur during holidays. Markdowns only occur about 30 percent of the time.

There are three types of stores in the data set: 22 Type A stores,17 Type B stores, and 6 Type C stores. From bellow figure we can see that Type A is large Walmart store, Type B is medium size Walmart store, and Type C is small Walmart store.



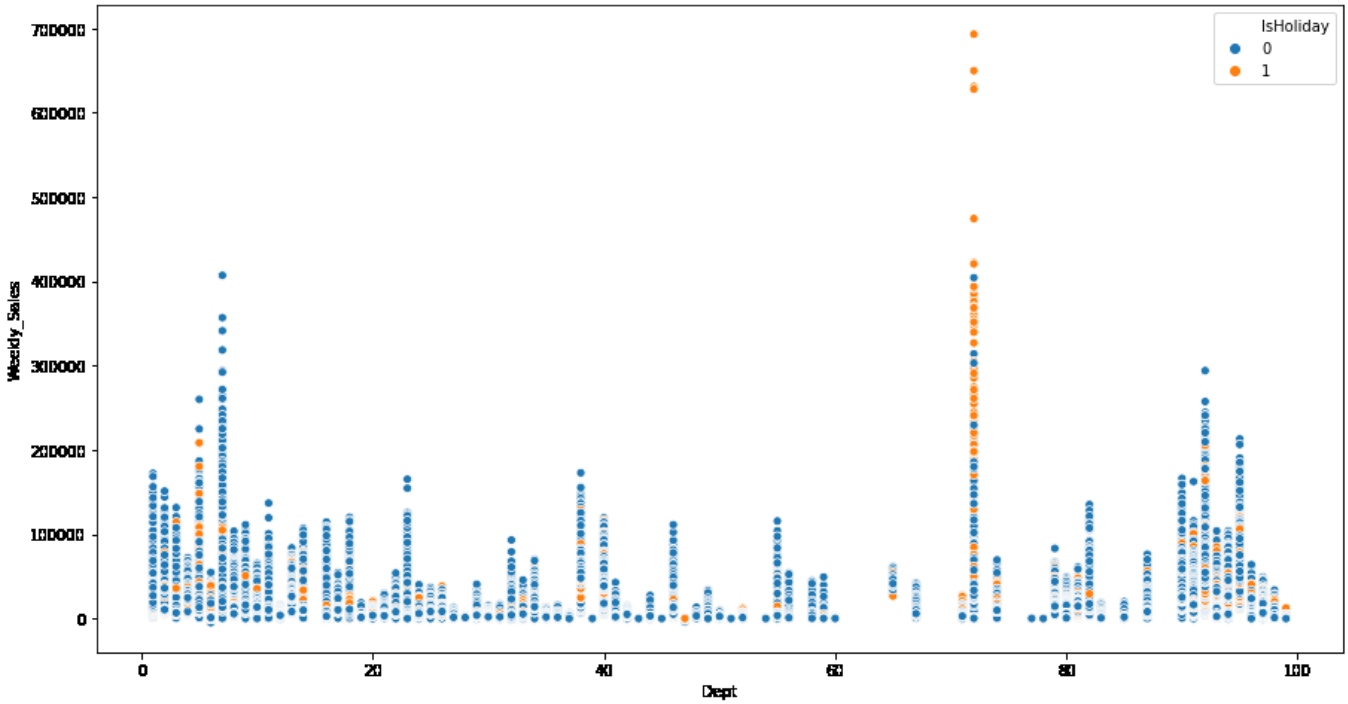
Type A stores have higher weekly sales, however, the average weekly sales are lower for Type A store compared with the smaller Type C store. The Store variable contains much intrinsic information such as Type, Size, Markdown, and Department, therefore it provides valuable information on weekly sales.



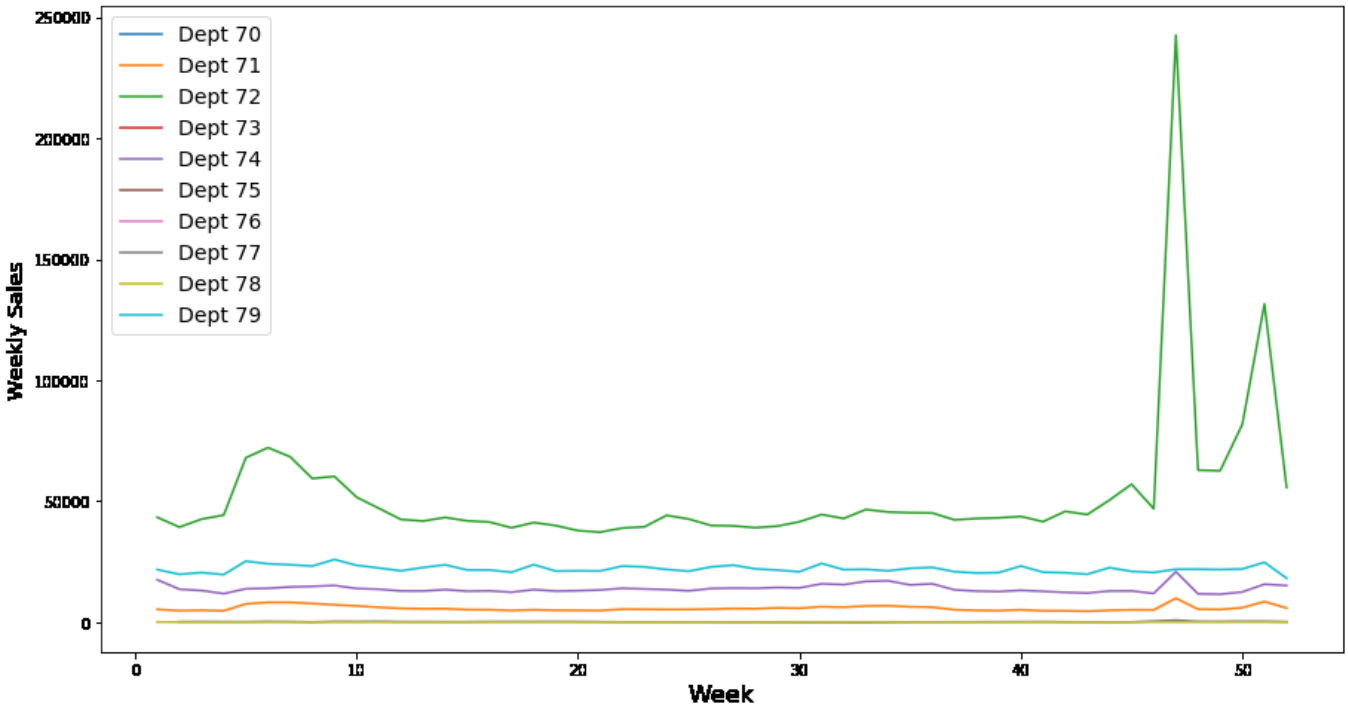
Part II: Which departments are most affected by holiday markdowns?

To analyze what departments are most affected by holiday markdowns, each department's weekly sales data were visualized by a scatter plot. Weekly sales during a

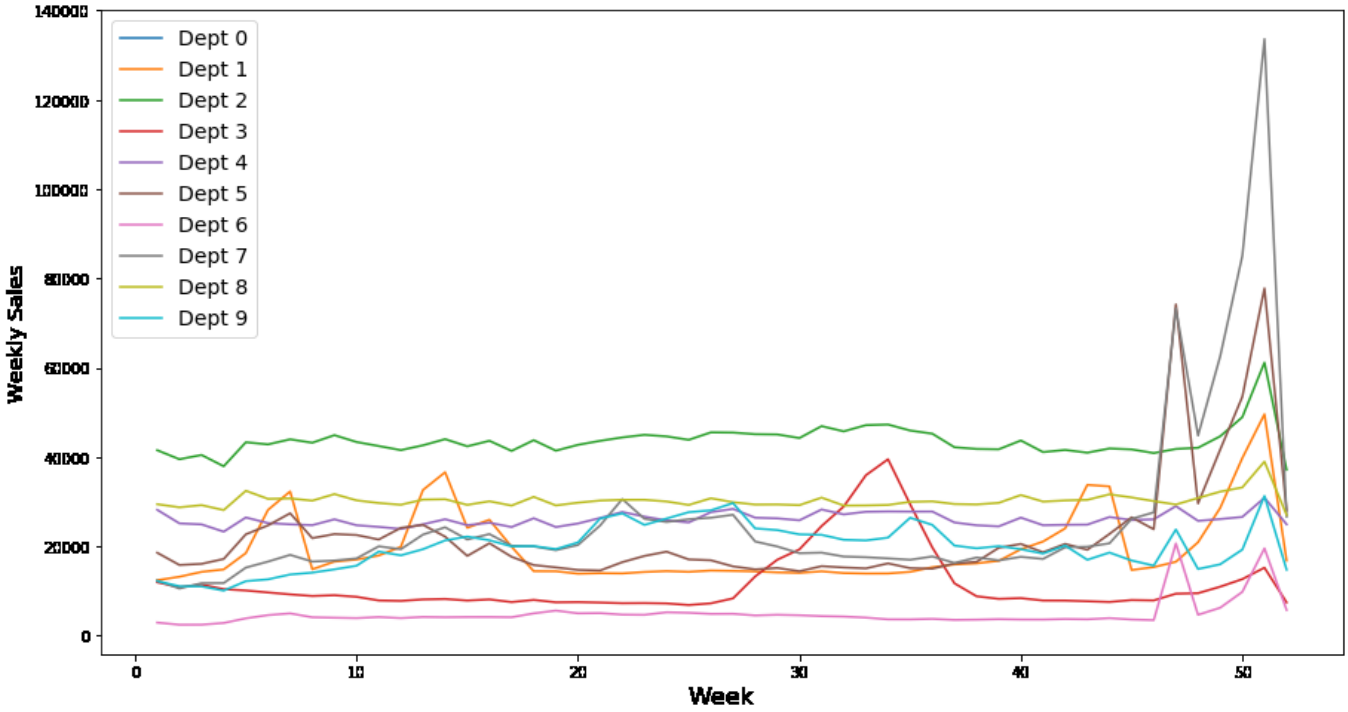
holiday week were marked with brown. From the figure we can see departments between the numbers 70 to 80 are affected dramatically by holiday markdowns.



Further analysis has shown that Department-72's weekly sales are affected by Thanksgiving and Christmas dramatically. Weekly sales during the Thanksgiving holiday week are much higher than the Christmas holiday week.Thanksgiving week has an about 5 folds sales surge than other weeks.

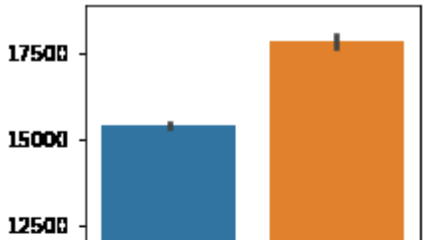


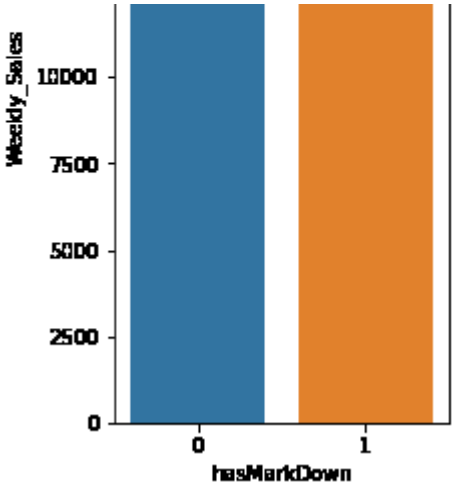
For Department-7, however, weekly sales during Christmas week are much higher than Thanksgiving week. Department-7's Christmas weekly sale is about 5 folds higher than non-holiday weeks. Interestingly, Department-3 has a weekly sale spike near the Labor Day holiday (near September 10th every year). These analyses suggest that some department's weekly sales are strongly correlated with holiday markdowns. Department should be a good indicator to predict weekly sales.



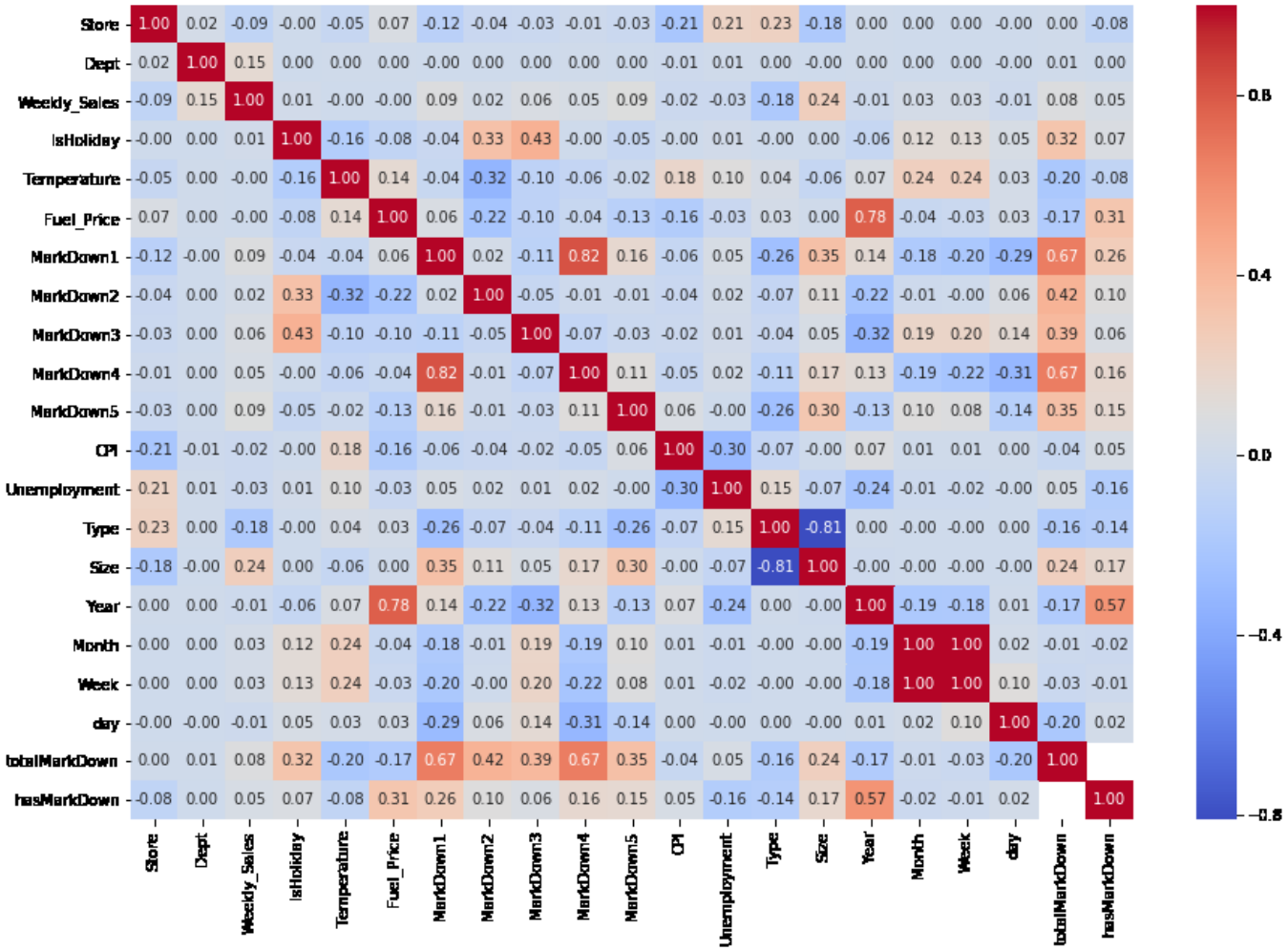
Part III: To what extent do markdowns impact weekly sales?

To further utilize Markdowns data, I engineered totalMarkDown feature, which is the sum of MarkDown1 to MarkDown5 and hasMarkDown feature. totalMarkDown returns yes whenever one kind of MarkDown1 to MarkDown5 event is happening on that day. Data shows that 23% of time Walmart stores have some kind of markdown event. From below figure we can see that Markdown events have significant impact on a store's weekly sales.





From the correlation coefficient heatmap, we can see that Weekly_Sales the variable strongly correlates with store Size, Type, and Department. MarkDown1 to MarkDown5 has positive correlations with store weekly sales. totalMarkDown and hasMarkDown are also correlated with store weekly sales.



Part IV: Using Machine Learning to predict department weekly sales.

First, I tried to use LinearRegression, DecisionTreeRegressor and RandomForestRegressor to predict weekly sales. DecisionTreeRegressor and RandomForestRegressor have a much better coefficient of determination scores (r2 score) compared to the LinearRegressor model. To better evaluate different machine learning algorithm, I also defined a function for calculating weighted mean absolute error (WMAE) since this Walmart Recruiting — Store Sales Forecasting competition was evaluated on the weighted holiday week sales.

> This competition is evaluated on the weighted mean absolute error (WMAE):
>
> $$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$
>
> where
>
> - n is the number of rows
> - $\hat{y}_i$ is the predicted sales
> - y_i is the actual sales
> - w_i are weights. w = 5 if the week is a holiday week, 1 otherwise

I then applied a grid search to find the best n_estimators and max_depth for RandomForestRegressor. The best n_estimators is 50 and max_depth is 20 for RandomForestRegressor and the WMAE score with these parameters is 1586.0. Finally, I tried AdaBoostRegressor with DecisionTreeRegressor as base estimator, this model gave me a WMAE score 1476.9, which is slightly better than best score from RandomForestRegressor. I opted for AdaBoostRegressor as my estimator to predict test data. I successfully submitted my predicted result to Walmart Recruiting — Store Sales Forecasting, my Private Score is 4117.399 and my Public Score is 4011.110.

Conclusion

In this post, I briefly described how I used data science skills and machine learning methods to perform Walmart Store Sales Forecasting .

Data Science Machine Learning Sales Forecasting Data Mining Data Visualization

About Help Legal

Get the Medium app