BUAN6356.006 - Group 14 project Report

Walmart Recruiting - Store Sales Forecasting

Members: Kai-Cheng,Lo - KXL180005

Kshitij Soni - KXS180022

Sakshi. - SXX180001

Sogoloba Coulibaly - SMC180005

Satyam Singh - SXS180097

Project Motivation/Background

There are lots of different dataset online, as students in Business analytic program, so we want to solve the business problem by using data analytic skills. After checking this dataset, we knew that we have to do sales forecasting for Walmart which is very challenging but also very interesting at the same time. It is very good to put what we have learned in class into a well-know company case.

Data Description

The data is about historical sales of 45 Walmart stores located across different regions with a total of 537k observations.

The file contains additional data related to the store, department, and regional activity for the given dates and also features from Walmart stores like markdowns which precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks.

The data contains the following features:

- Store - the store number
- Date - the week
- Temperature - average temperature in the region
- Fuel_Price - cost of fuel in the region
- MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
- CPI - the consumer price index
- Unemployment - the unemployment rate
- IsHoliday - whether the week is a special holiday week

Exploratory data analysis

1. We will observe the relationship between some independent variables and Weekly_sales
2. This analysis could help us to define the characteristics of the different type of stores that we don't know

Dept x WeeklySales

WALMART DEPARTMENT NUMBERS

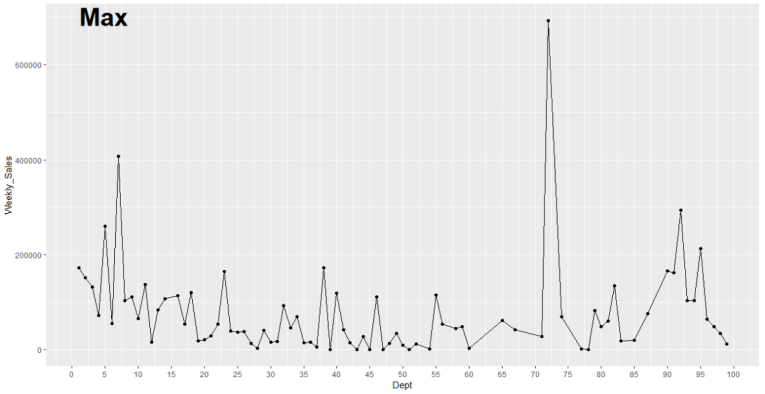
SORTED BY NUMBER

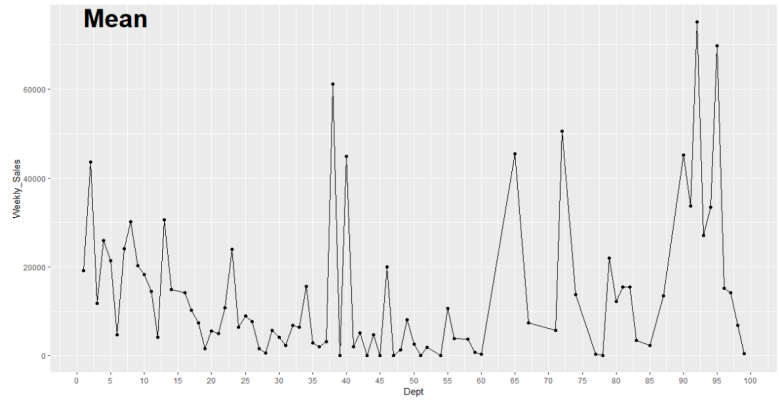
#	Department Name	#	Department Name
1	Candy & Tobacco	38	Pharmacy Rx
2	Heath & Beauty Aids	39	Consumer Service
3	Stationery	40	OTC Pharmacy
4	Household Paper	46	Cosmetics & Skincare
5	Media & Gaming	49	Optical
6	Cameras & Supplies	50	Optical
7	Toys	56	Horticulture
8	Pets & Supplies	58	Wireless Service, Inc.
9	Sporting Goods	60	Concept Stores
10	Automotive	65	Gasoline
11	Hardware	67	Celebration
12	Paint & Accessories	71	Furniture
13	Household Chemicals	72	Electronics
14	Cook & Dine	74	Home Management
15	Health and Wellness Clinics	75	Fixtures
16	Lawn & Garden	77	Large Household Goods
17	Home Decor	79	Infant Consumables Hardlines
18	Seasonal	80	Service Deli
19	Piece Goods & Crafts	81	Commercial Bread
20	Bath & Shower	82	Impulse Merchandise
21	Books & Magazines	83	Seafood
22	Bedding	84	Balloons & Flowers
23	Menswear	85	One-Hour Photo
24	Boyswear	86	Walmart Services
25	Shoes	87	Wireless
26	Infant Apparel	88	PMDC Signing
27	Family Socks	89	Travel
28	Hosiery	90	Dairy
29	Sleepwear	91	Frozen
30	Foundations	92	Grocery
31	Accessories	93	Meat
32	Jewelry	94	Produce
33	Girlswear	95	DSD Grocery
34	Ladies Apparel	96	Liquor
35	Plus Sizes & Maternity	97	Wall Deli
36	Ladies Outerwear	98	Bakery
37	Auto Service	99	Office & Store

According to the walmart website, different Dept numbers represent different products they are focus on.

Dept	count	mean(Weekly_Sales)
92	6435	75204.870531
95	6435	69824.423080
38	6435	61090.619568
72	6046	50566.515417
65	143	45441.706224
90	6435	45232.084488
40	6435	44900.702727
2	6435	43607.020113
91	6435	33687.910758
94	5685	33405.883963
13	6435	30663.802634
8	6435	30191.263517
93	5913	27008.060746
4	6435	25974.630238
7	6435	24161.237413
23	5774	23931.473983
79	6435	21973.044932
5	6347	21365.583515
9	6354	20206.681878
46	6435	19944.741483

Dept	Weekly_Sales	Type
72	693099.4	B
72	649770.2	B
72	630999.2	B
72	627962.9	B
72	474330.1	A
72	422306.2	A
72	420586.6	A
7	406988.6	B
72	404245.0	B
72	393705.2	B
72	392023.0	A
72	385051.0	A
72	381072.1	A
72	375948.3	A
72	369831.0	B
72	368484.2	A
72	360140.7	B
72	359995.6	B
7	356867.2	A
72	355356.4	A
72	353008.6	B
72	351763.7	A
72	351554.0	A
72	347680.1	B
72	345532.2	B



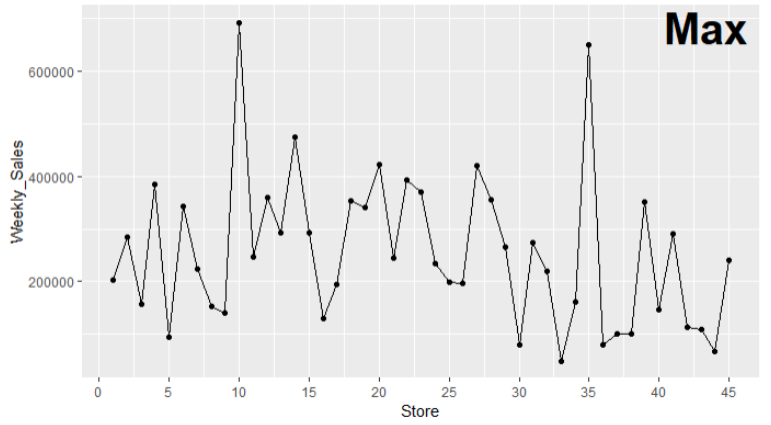


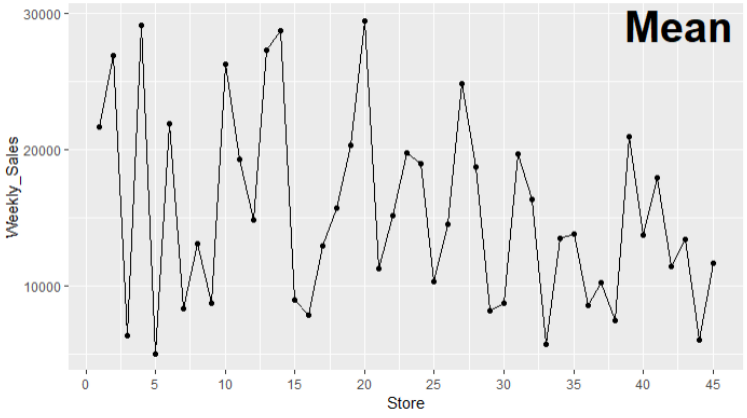
When we order the sales by department, we can find that the highest amount of weekly sales by department happened in Dept 72. This Department in Walmart represents Electronics, so TV's, Laptops, Speakers, Headphones, etc. which sell expensive stuffs.

Despite this fact, considering the average of weekly sales by department,the Department 92 had the highest average of weekly sales of 75204.87$. This phenomenon is explained by the fact that department 92 is the grocery's department which sells all the time throughout the year, so that it has the highest average sale.

Stores x WeeklySales

Store	count	mean(Weekly_Sales)
20	10214	29508.302
4	10272	29161.210
14	10040	28784.852
13	10474	27355.137
2	10238	26898.070
10	10315	26332.304
27	10225	24826.985
6	10211	21913.244
1	10244	21710.544
39	9878	21000.764
19	10148	20362.127
23	10050	19776.181
31	10142	19681.907
11	10062	19276.763
24	10228	18969.106
28	10113	18714.890
41	10088	17976.005
32	10202	16351.622
18	9859	15733.313
22	9688	15181.219
12	9705	14867.309





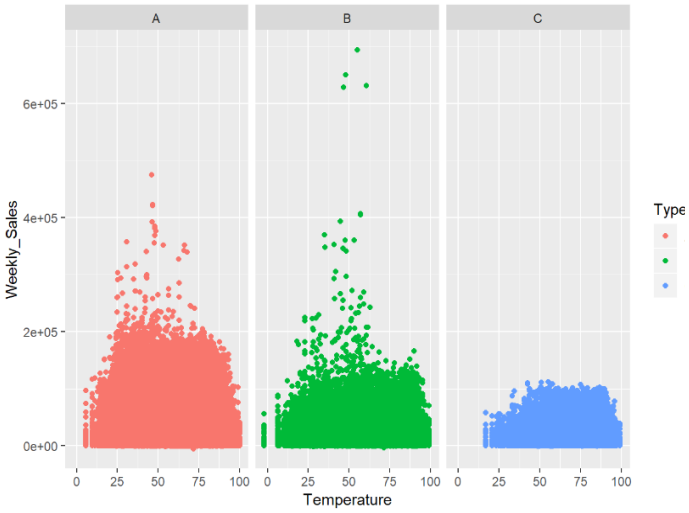
Considering the average of weekly sales by store, Store 20 has the highest sales of 29508.30 and is of Type A. if we focus on the top 10 high-sales store, we would find that nine of them are type A. only store 10 is type B.

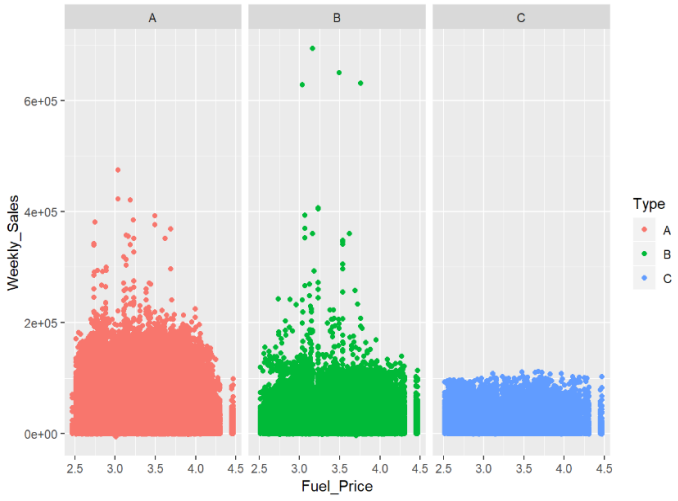
Type x WeeklySales

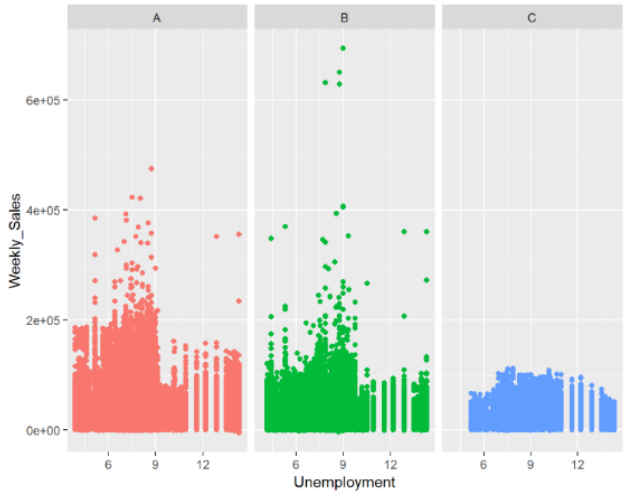
According to Walmart website, there is three types of stores in Walmart, like, the Supercenter which is the backbone of the Walmart empire. The average size of a Supercenter is 179,000 square feet, and the stores offer both merchandise and groceries. There is also, the 15,000 square-foot Walmart Express stores or Discount Store with an average size of 105,000 square feet. These stores typically don't offer groceries and are open only 14 or 15 hours a day and everyday.

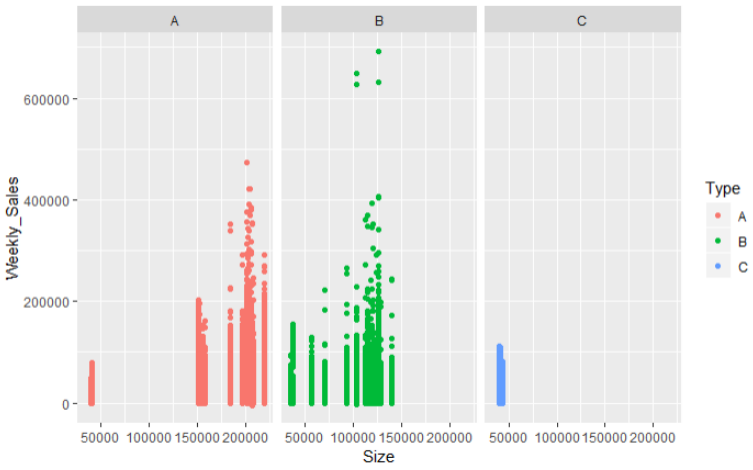
Type	count	mean(Weekly_Sales)	mean(Size)
A	215478	20099.568	182231.29
B	163495	12237.076	101818.74
C	42597	9519.533	40535.73

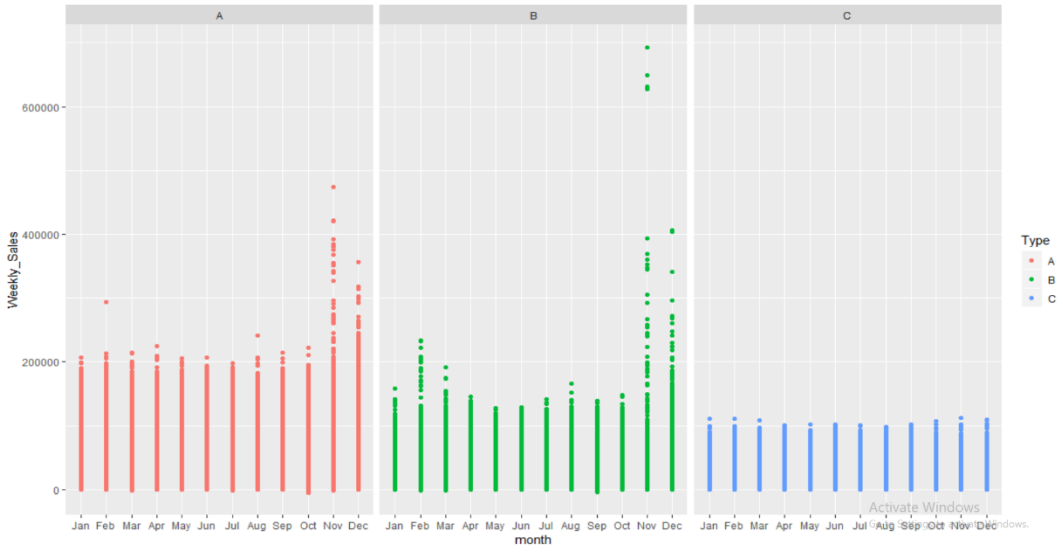
What are the characteristics of the differents types of variables











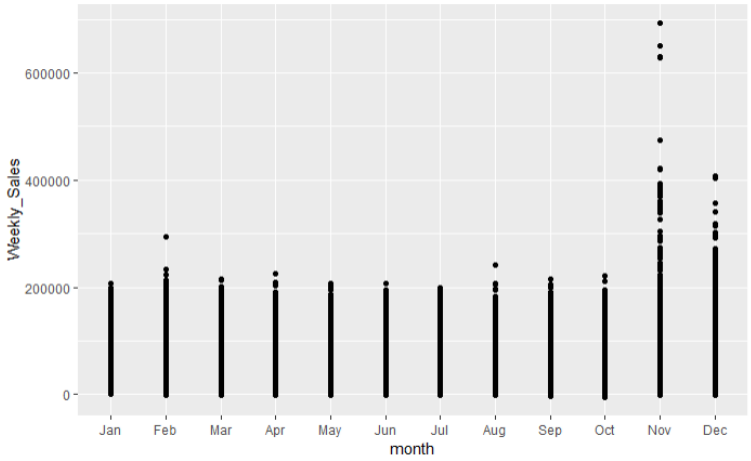
Obviously, in every variable, the store of type is A always has much sales and according to the mean size of each type, we can indicate that the store of type A is supercenter. What is more, we have already known that top 4 high sales happened in Dept 72 with Type B stores which sell electronic things and all of them happened in Nov, after checking the date, we can make sure

that someone bought a lot high-price electronic products during black-friday. So, we indicate that Type B store is Discount store. In the end, Type C is the express store.

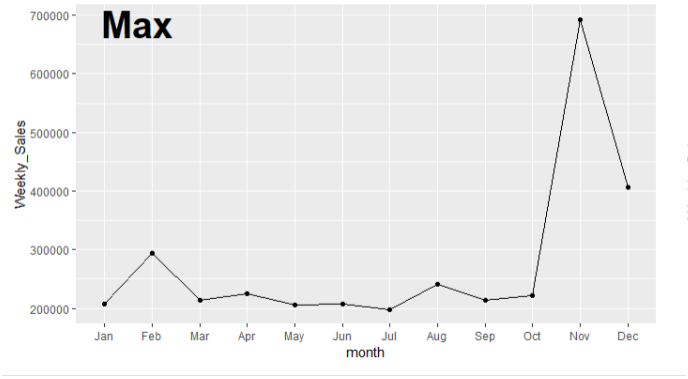
Month x Weeklysales

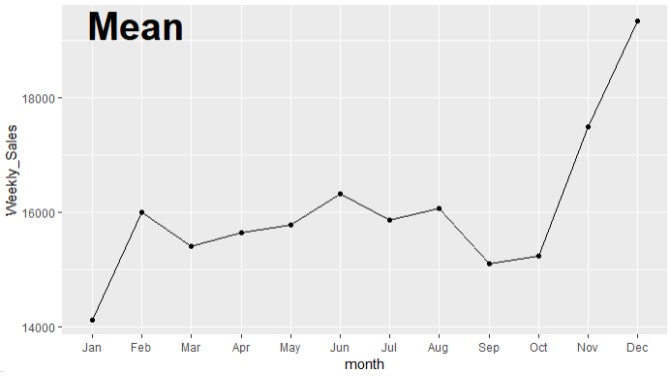
Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Date	IsHoliday	Dept	Weekly_Sales
2010-11-26	TRUE	72	693099.4
2011-11-25	TRUE	72	649770.2
2011-11-25	TRUE	72	630999.2
2010-11-26	TRUE	72	627962.9
2010-11-26	TRUE	72	474330.1
2010-11-26	TRUE	72	422306.2
2010-11-26	TRUE	72	420586.6
2010-12-24	FALSE	7	406988.6
2010-12-24	FALSE	72	404245.0
2010-11-26	TRUE	72	393705.2
2011-11-25	TRUE	72	392023.0
2011-11-25	TRUE	72	385051.0
2010-11-26	TRUE	72	381072.1
2011-11-25	TRUE	72	375948.3
2010-11-26	TRUE	72	369831.0
2011-11-25	TRUE	72	368484.2
2011-11-25	TRUE	72	360140.7
2010-11-26	TRUE	72	359995.6

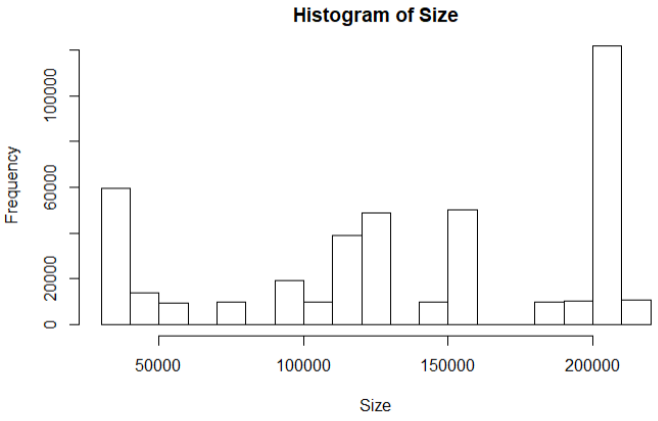


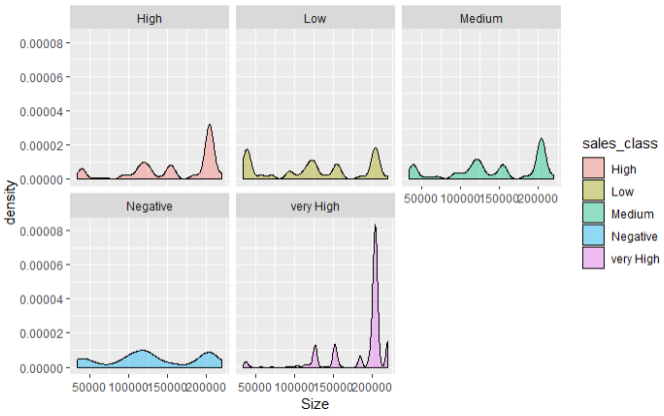
Obviously, during November and December has more sales because of Thanksgiving and Christmas and New year.





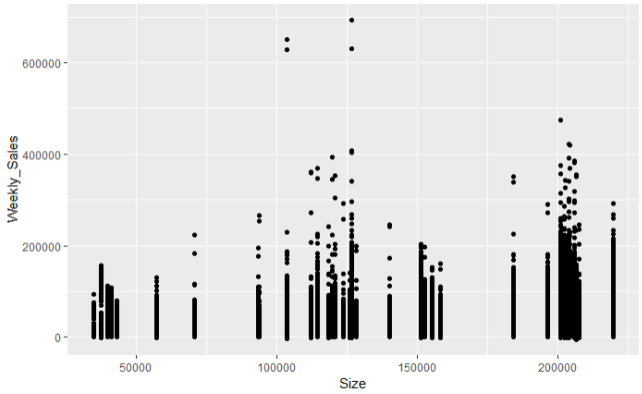
If we order the data by weekly_sales, we would find that all top 8 high sales happened in Thanksgiving, but when we calculate the average, we found that December has more sales. I think because the period of location in December is longer. (Christmas, New year and Winter vacation). Otherwise, Thanksgiving is the only one holiday in November.

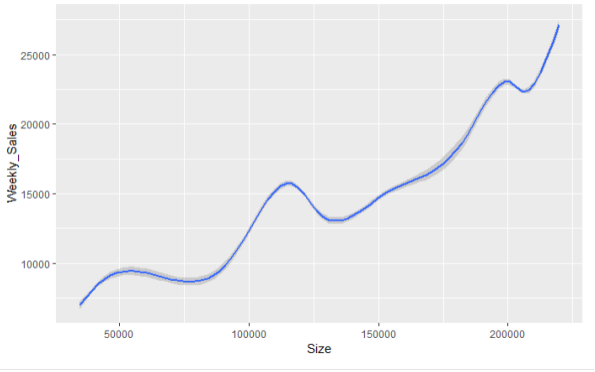




Size x Weeklysales

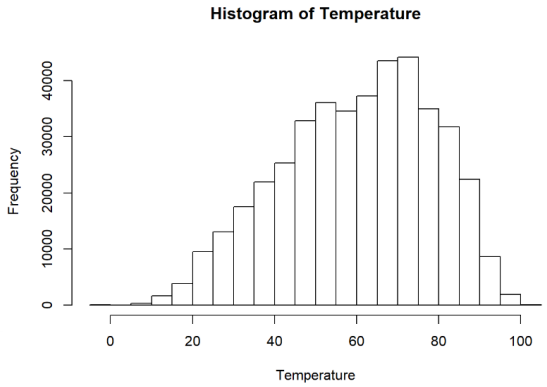
Looking for the right plot, and focus on 'Low Sales', lots of their size are small, But when we focus on medium, high and very high sales plot, we can find that the number of big size store are much. For stores which has very high sales, even most of their size around 20000. I can indicate that the size of store is a significant variable for Weekly Sales.

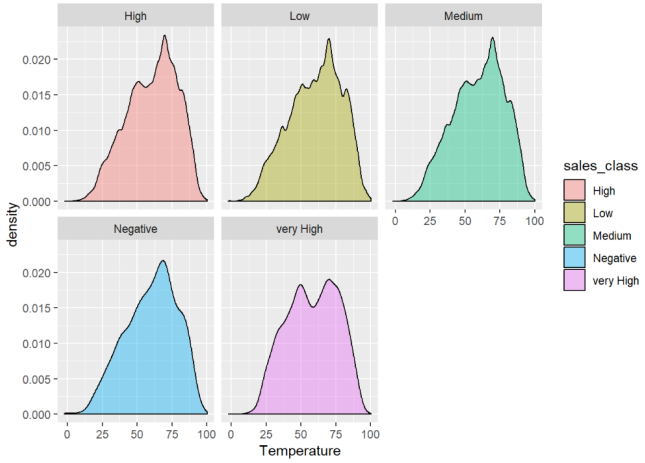




We have already known that the type of supercenter store in Walmart has much sales in general, and the average size of supercenter store is very big (179,000), but according to the scatter plot, it's not obvious to tell the store with huge size has much sales than medium size stores because there are some super high-sales points there. so after using geom_smooth to plot, it will be very clear to see that as size increase, the Weekly Sales increased.

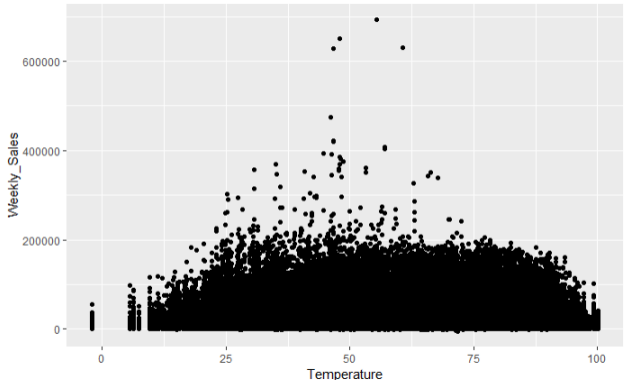
Temperature x Weekly_sales

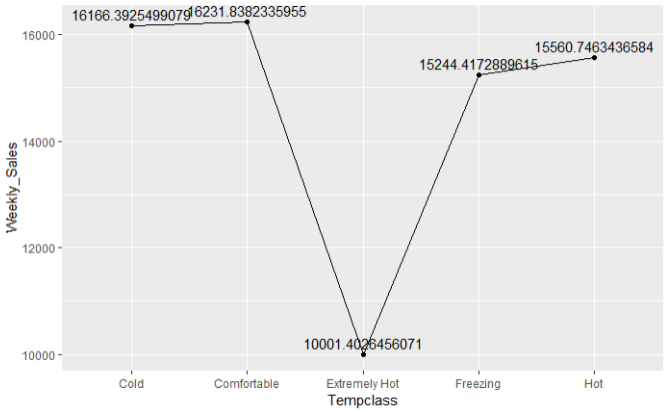




No matter which sales_class is, most of stores were built in the region which average temperature is in 40~80. It can indicate people who lives in this range of temperature tend to shopping.

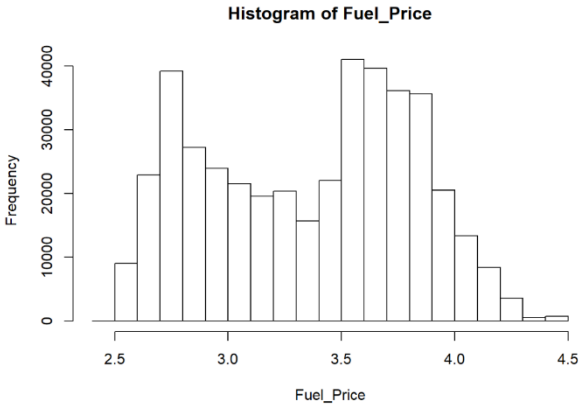
(Freezing = 32 < Cold = 32 – 64 Comfortable = 64-79 Hot = 79-95 Extremely hot = > 95)

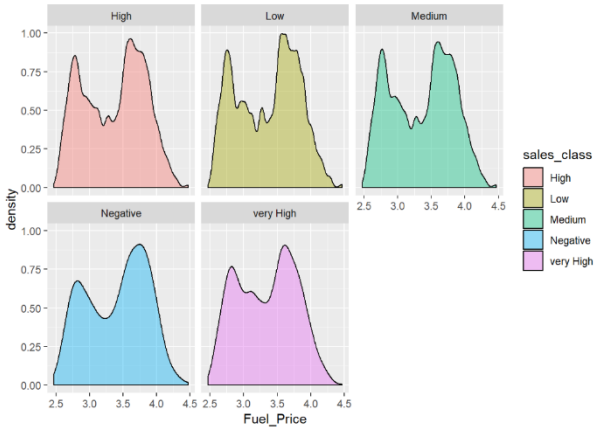




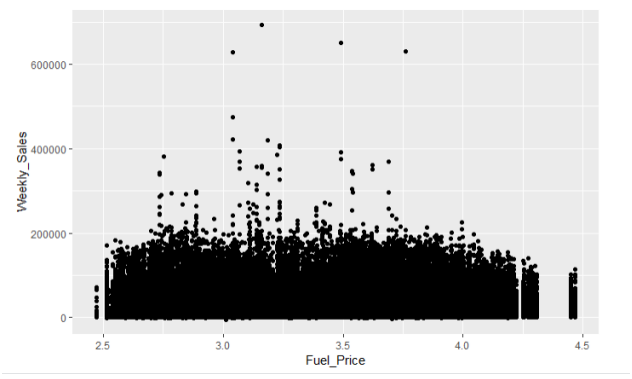
From the first plot, we can see that there is a convex shape in 30 ~ 70 temperature which means lots of high sales falls in this interval. If we divide temperature to different group, it's more intuitive to see the price in different temperature.

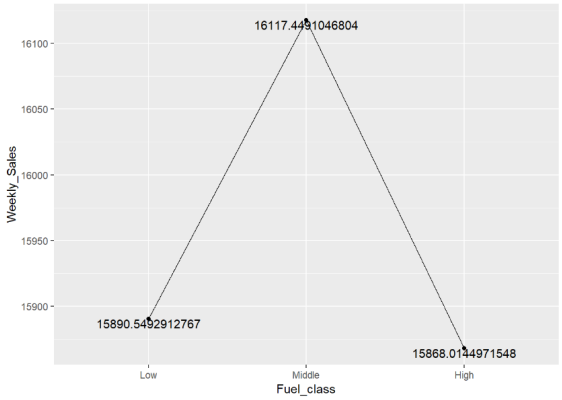
Fuel x WeeklySales





The data distribution under different sales_class is quite same, most of stores' fuel price falls in the interval between 2.6~3.0 and 3.5~3.8.



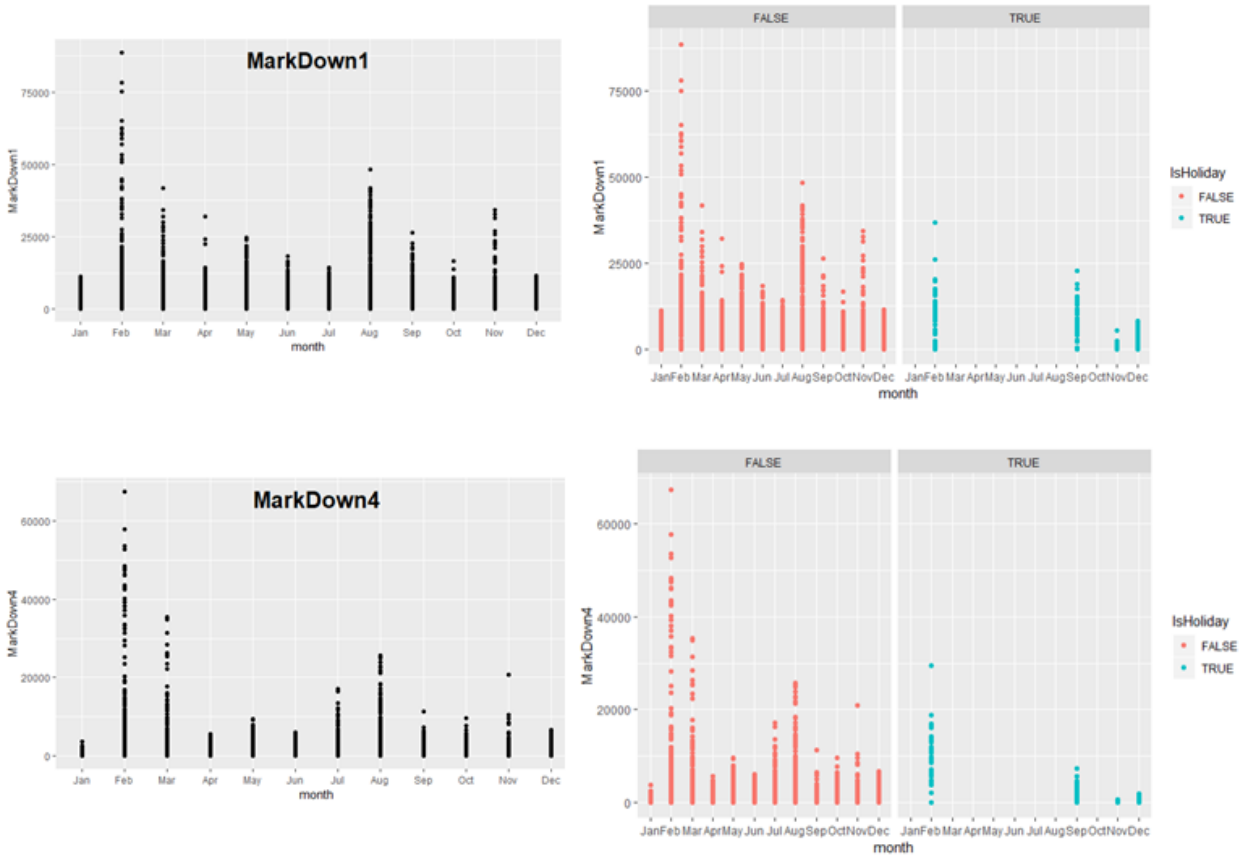


(Low price = 3 < Middle price = 3 – 3.7 High = >3.7)

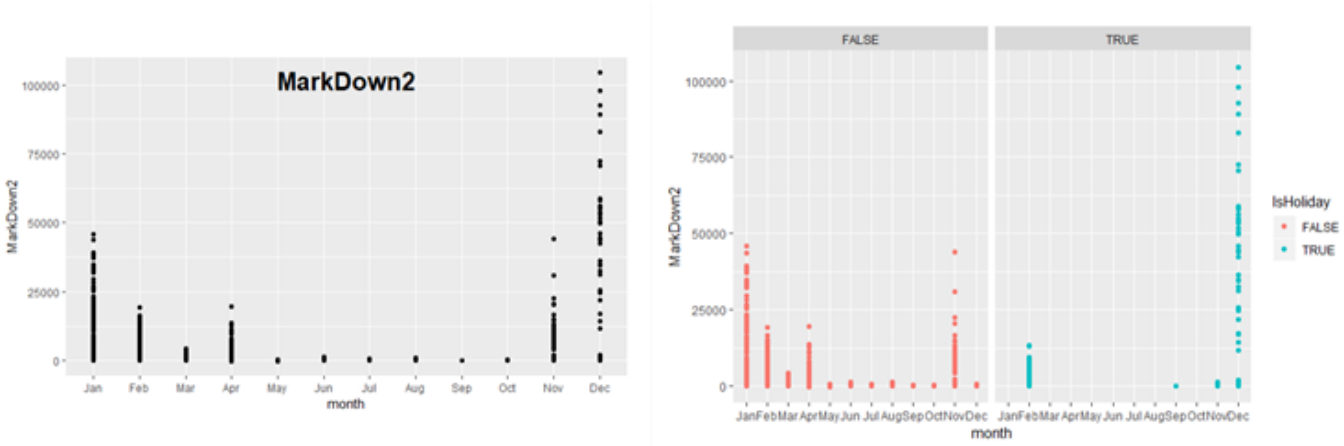
If we divide them to three classes, it's clearer to see the stores with middle fuel prices(3~3.7) have more sales. We don't think the Fuel_Price has a strong relationship with weekly_sales, because when we focus on the scatter plot, the distribution is fairly even. Because of Thanksgiving and Christmas, there are some transaction amounts that are so high and which exactly is the reason make the Line Graph look like.

Markdown x month x Holiday

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labor 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

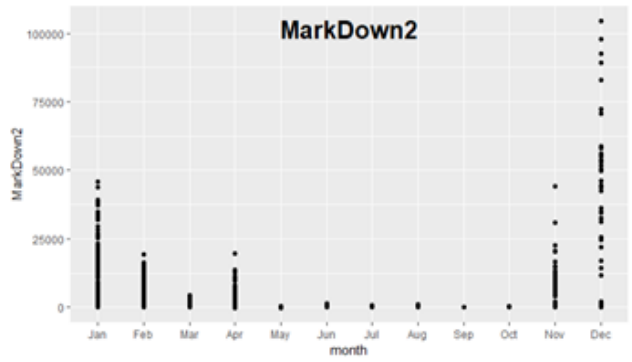


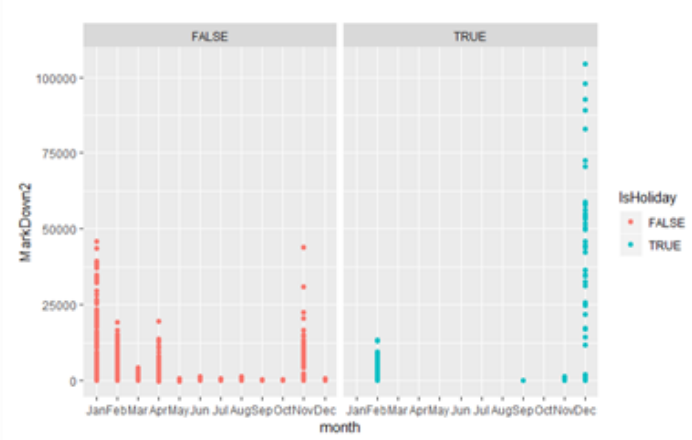
Markdown 1 and Markdown 4 are mainly offered on February, March and August and most of them offered during non-holiday. This discount is possibly related to Super bowl season, Mother's day and back-to-school season.



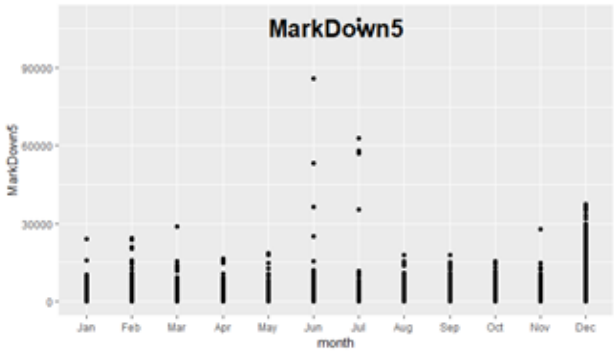
Markdown 2 is offered mainly during Dec holiday and Jan, but also has its peaks during

November. This discount is possibly related to Christmas and New year.



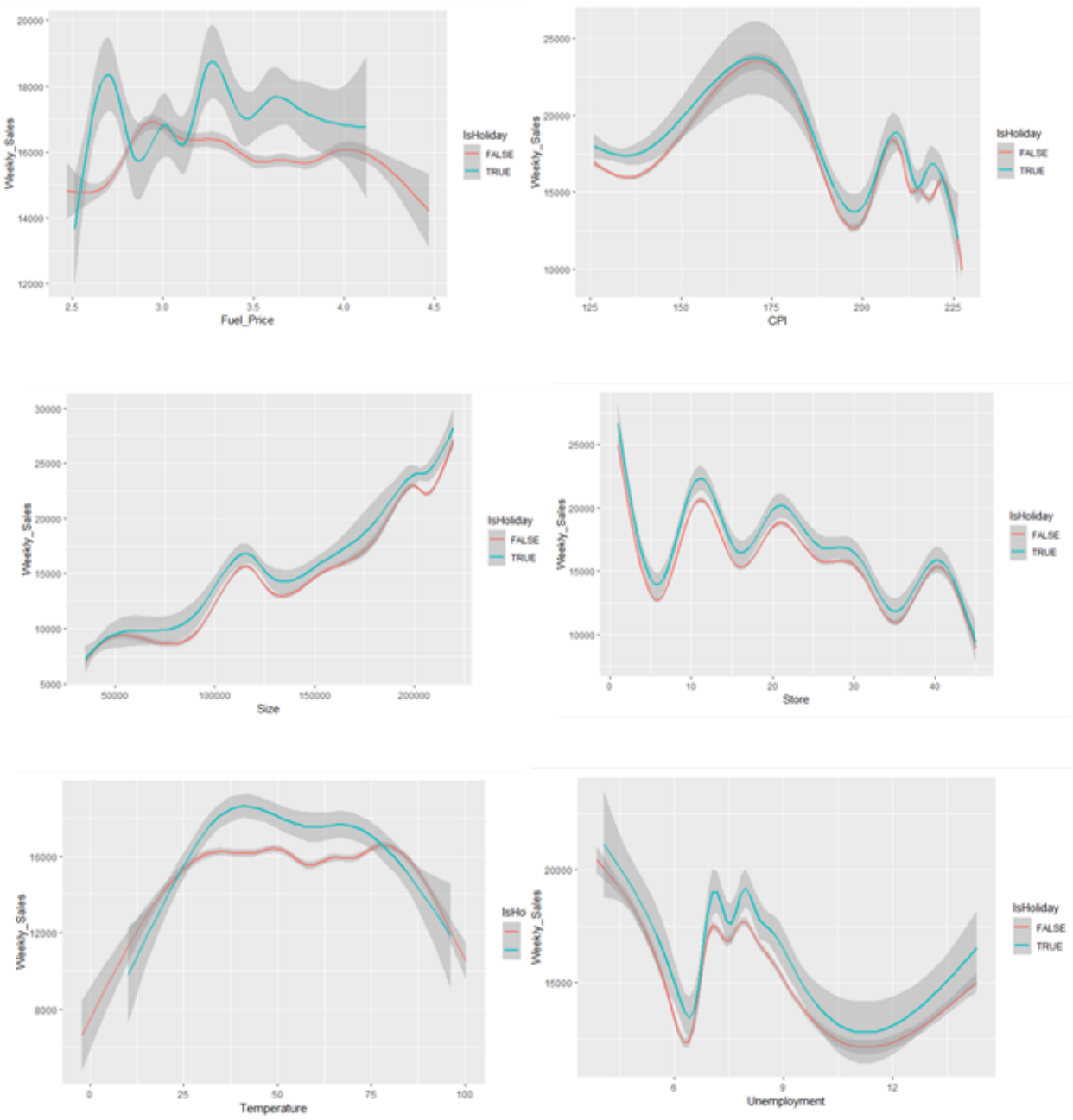


Markdown 3 is offered only during Nov holiday and Dec. This discount is possibly related to Thanksgiving.



Markdown 5 is mainly offered during non-holiday season

Holiday x Weekly_Sales

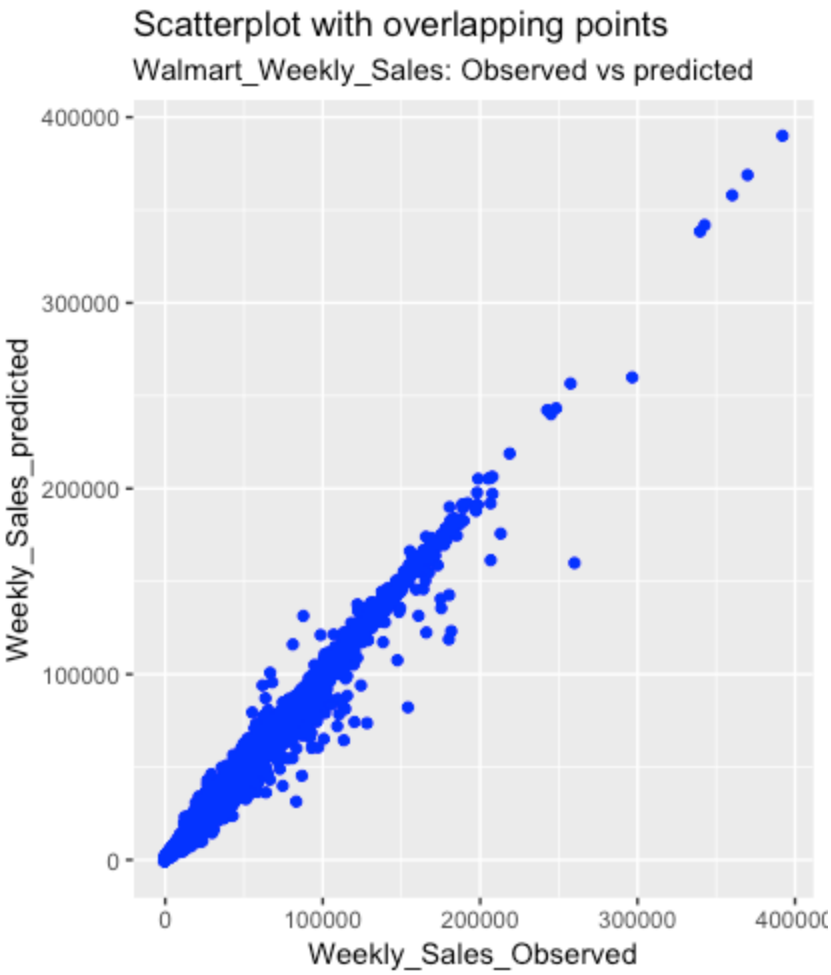


Obviously, from the different variable perspective, weekly sales is higher during holidays.

Models and analysis

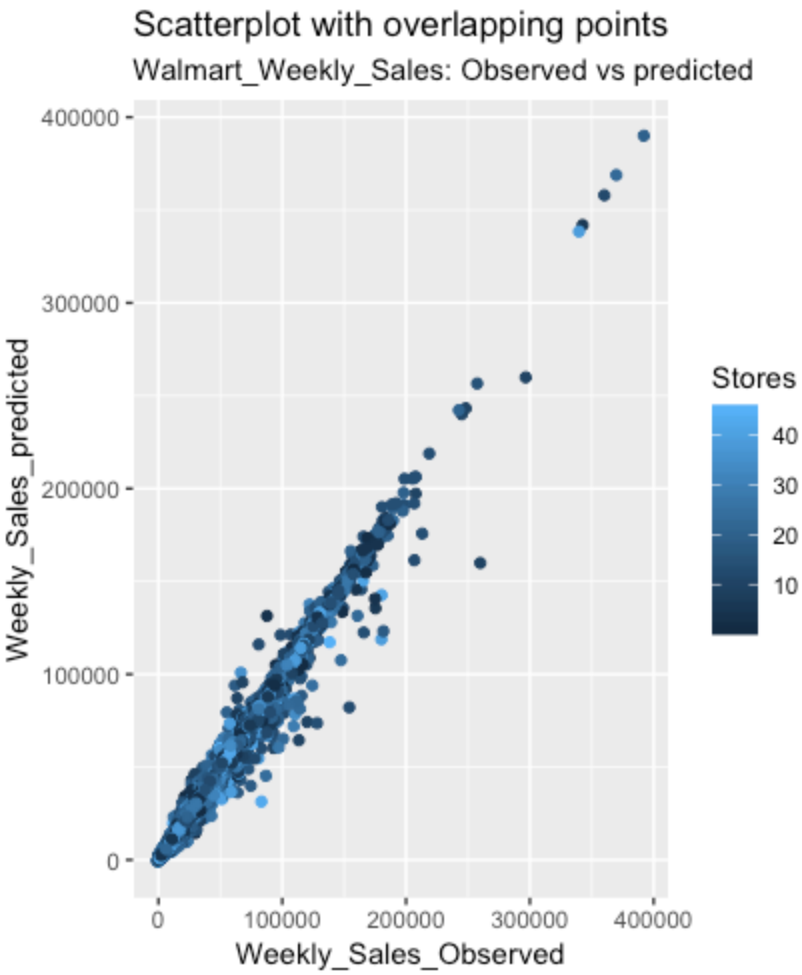
In model analysis we have divided the dataset into two parts: We have kept 80% of the our data as training and rest we have used for testing. As we have seen from our variable analysis that date is the most important variable in our data. We have included various dates elements like month, year, days, total days to include the pattern. We have increase the weights of certain dates for which there is a holiday by duplication, we could have used simple multiplication technique also, the result would have been the same. We have selected the important variables using lasso regression.

We have first used linear regression but were not getting the required accuracy. After that we have tried different algorithms such as decision trees etc. and finally we build our model using Random Forest. Algorithm is implemented using two conditions. First condition is if the sample size of filtered data(filtered on store and dept) is very less then we will use all the department of that store otherwise we will use data filtered on store and dept. We have build different random forest models for different department and stores. In the end we were able to predict department-wise weekly sales of stores with a MAPE (mean absolute % error) of 10%.

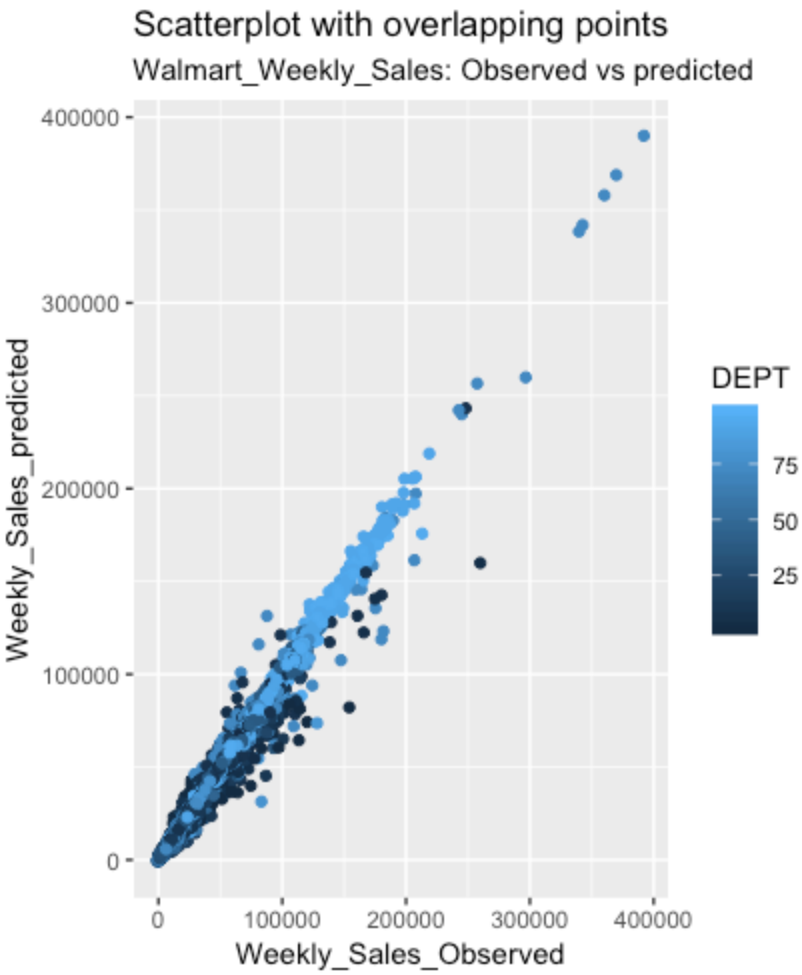


This graph represents total weekly sales prediction on the Y axis and observed Weekly Sales on X axis.

We observe points form a straight line which means, errors are less in the predicted values.



This graph represents store wise prediction of sales.



This graph represents department wise prediction of weekly sales.

Findings and managerial implications

These are some important findings:

- Dept 72 in Walmart is Electronics like TVs, Laptops, Speakers, Headphones, etc, the store with dept 72 has very high weekly sales in general.
- Type A store is the supercenter store which is the backbone of the Walmart empire. The average size of it is 179,000 square feet, and the stores offer both merchandise and groceries. Type A store has high weekly sales.
- Top 4 high sales happened in Dept 72 with Type B stores, after checking the date, we can make sure that someone bought a lot high-price electronic products during black-friday,so we indicate that Type B store is Discount store.
- Stores which are built in the region with the range of 30 ~ 70 F temperature, have many sales.
- Markdown 1 and Markdown 4 are mainly offered on Feb, Mar and Aug and most of them offered during non-holiday. These discounts is possibly related to Super bowl season, Mother's day and Back-to-school season.
- Markdown 2 is offered mainly during Dec holiday and Jan, but also has its peaks during November. This discount is possibly related to Christmas and New year.
- Markdown 3 is offered only during Nov holiday and Dec. This discount is possibly related to Thanksgiving.
- Holiday always bring in more sales than non-holiday.

Conclusions

- Electronics have highest sales overall so we can position the other department accordingly so that people can go through department with lower sales.
- We can predict department-wise sales of store within 10% of error
- Can hire more personnel as stores have huge spikes in sales during holidays.
- Incentives to stores with lower sales and hiring right sales executive.
- Optimal store size should be between 100,000-150,000 square feet.

References:

https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

https://github.com/mikeskim

https://www.8thandwalton.com/departments