group_project_KG.R

danny

2020-04-18

```r
suppressWarnings(suppressPackageStartupMessages({
library(dplyr)
library(ggplot2)
library(stargazer)
library(plm)
library(scales)
}))

## Set working dir
setwd("C:/Users/danny/Downloads")

#### Load the data ####
stores = read.csv("stores data-set.csv")
sales = read.csv("sales data-set.csv")
features = read.csv("Features data set.csv")

### Data Preparation and Transformation ###

# Convert Date field in sales from dd/mm/yyyy to yyyy/mm/dd
sales$Date_new <- strptime(as.character(sales$Date), "%d/%m/%Y")
sales$Date <- format(sales$Date_new, "%Y-%m-%d")
sales$Date <- as.Date(sales$Date, format = "%Y-%m-%d")
sales <- select(sales, -c(Date_new))

# Convert Date field in features from dd/mm/yyyy to yyyy/mm/dd
features$Date_new <- strptime(as.character(features$Date), "%d/%m/%Y")
features$Date <- format(features$Date_new, "%Y-%m-%d")
features$Date <- as.Date(features$Date, format = "%Y-%m-%d")
features <- select(features, -c(Date_new))

# Make Type field in sales a factor
stores$Type <- as.factor(stores$Type)

# convert IsHoliday True/False indicator --> binary (1 = True, 0 = False), make a factor
sales$IsHoliday <- ifelse(sales$IsHoliday=="TRUE",1,0)
sales$IsHoliday <- as.factor(sales$IsHoliday)

# Sum daily total sales by Store (summing daily sales for the 72 different departments)
sales_daily <- sales %>% group_by(Store, Date, IsHoliday) %>%
  summarise(Total_Weekly_Sales = sum(Weekly_Sales))

# Add Week Number field to Sales
store_date_grouped <- sales_daily %>% group_by(Store, Date) %>% summarise() %>% ungroup()
store_date_grouped <- store_date_grouped %>% mutate(Week_No = rep(1:length(unique(sales_daily$Date)),
                                                                   length(unique(sales_daily$Store))))
sales_date <- inner_join(sales_daily, store_date_grouped, by=c('Store','Date'))
```

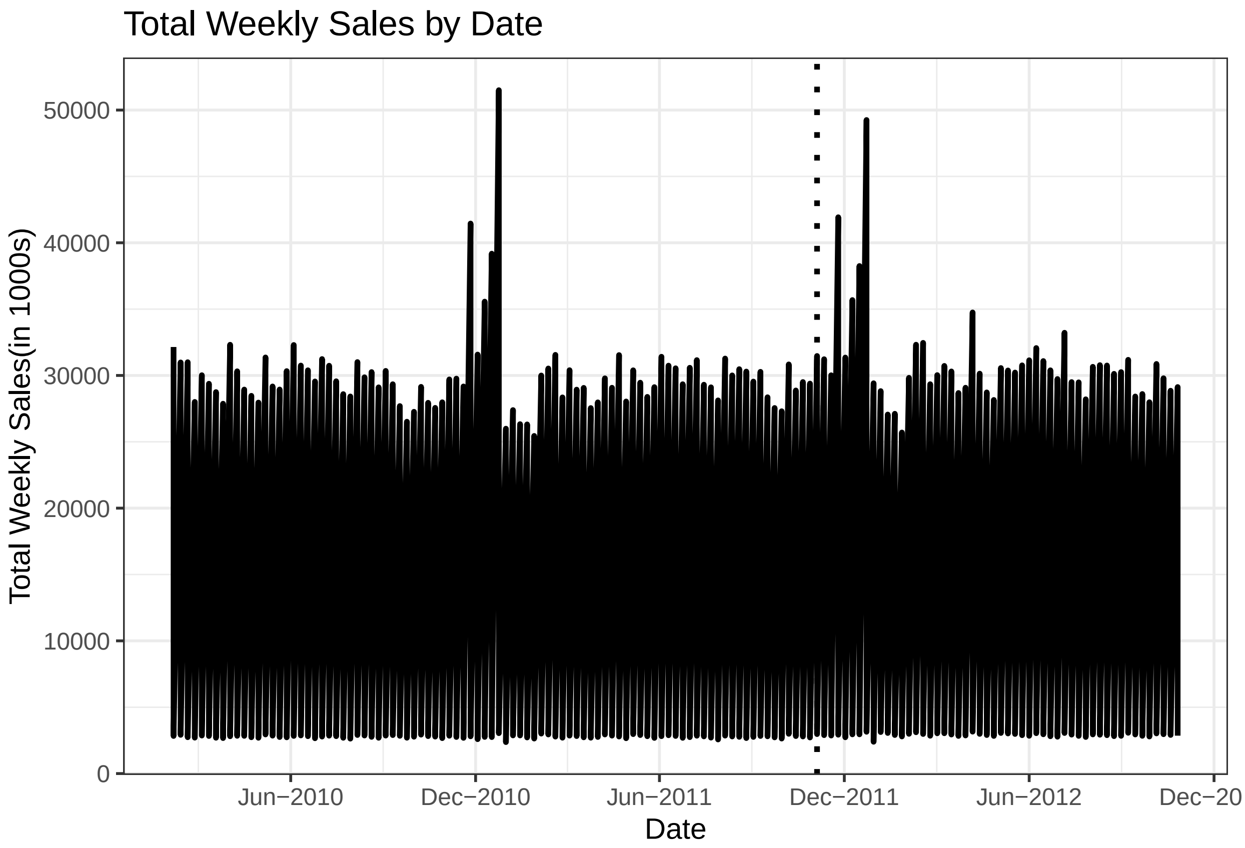
```r
# Features: Add MarkDown Total column (MarkDown's 1-5, summed)
features <- features %>% mutate(MarkDown_Total = MarkDown1 + MarkDown2 +
                                  MarkDown3 + MarkDown4 + MarkDown5)

# Join sales_date and stores tables
sales_stores <- inner_join(sales_date, stores, by='Store')

# Join sales_stores and features tables
sales_full <- inner_join(sales_stores, select(features, c(-IsHoliday)), by=c('Store','Date'))

# Identify which week treatment period begins (after >= 11/1/2011)
# sales_full %>% filter(sales_full$Date >='2011-11-01') %>% head()
# Treatment period: Week_Number >= 92

# add week identifier ("after')
  # whether or not a week number was part of the treatment period
# Weeks >=92: marked as 1; Weeks < 92: marked as 0
sales_full <- mutate(sales_full, after = ifelse(Week_No >= 92, 1, 0))

# Convert NA's to 0
sales_full <- sales_full %>% replace(is.na(.), 0)

# convert Week_No to factor for regression
sales_full$Week_No <- as.factor(sales_date$Week_No)

# Add HasMarkDown field: whether that store for that week had any MarkDown (MarkDown1 - MarkDown5)
sales_full <- mutate(sales_full,
                     HasMarkDown = ifelse((MarkDown1 > 0 | MarkDown2 > 0 |
                                  MarkDown3 > 0 | MarkDown4 > 0 | MarkDown5 > 0 ),1,0))
head(sales_full)
```

```
## # A tibble: 6 x 19
## # Groups:   Store, Date [6]
##   Store Date       IsHoliday Total_Weekly_Sa~ Week_No Type    Size
##   <int> <date>     <fct>                <dbl> <fct>   <fct>  <int>
## 1     1 2010-02-05 0                 1643691. 1       A     151315
## 2     1 2010-02-12 1                 1641957. 2       A     151315
## 3     1 2010-02-19 0                 1611968. 3       A     151315
## 4     1 2010-02-26 0                 1409728. 4       A     151315
## 5     1 2010-03-05 0                 1554807. 5       A     151315
## 6     1 2010-03-12 0                 1439542. 6       A     151315
## # ... with 12 more variables: Temperature <dbl>, Fuel_Price <dbl>,
## #   MarkDown1 <dbl>, MarkDown2 <dbl>, MarkDown3 <dbl>, MarkDown4 <dbl>,
## #   MarkDown5 <dbl>, CPI <dbl>, Unemployment <dbl>, MarkDown_Total <dbl>,
## #   after <dbl>, HasMarkDown <dbl>
```

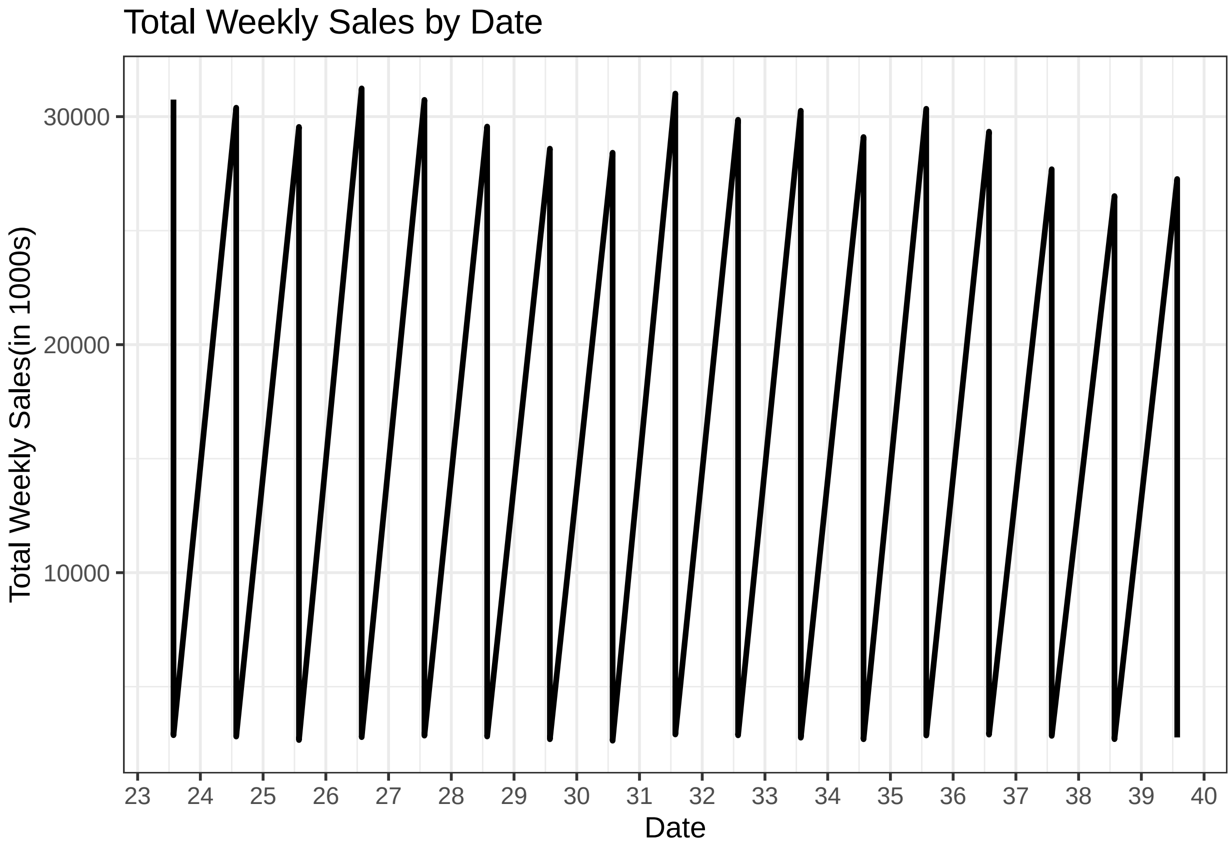
```r
### EDA ###

# Aggregate Total Weekly Sales by Date by Store Type --> new table sales_weekly_type
sales_weekly_type <- sales_full %>% select(Date, Total_Weekly_Sales, Type, IsHoliday) %>%
  group_by(Date, Type, IsHoliday) %>%
  summarise(Total_Weekly_Sales = sum(Total_Weekly_Sales))
```

```
## Adding missing grouping variables: `Store`
```

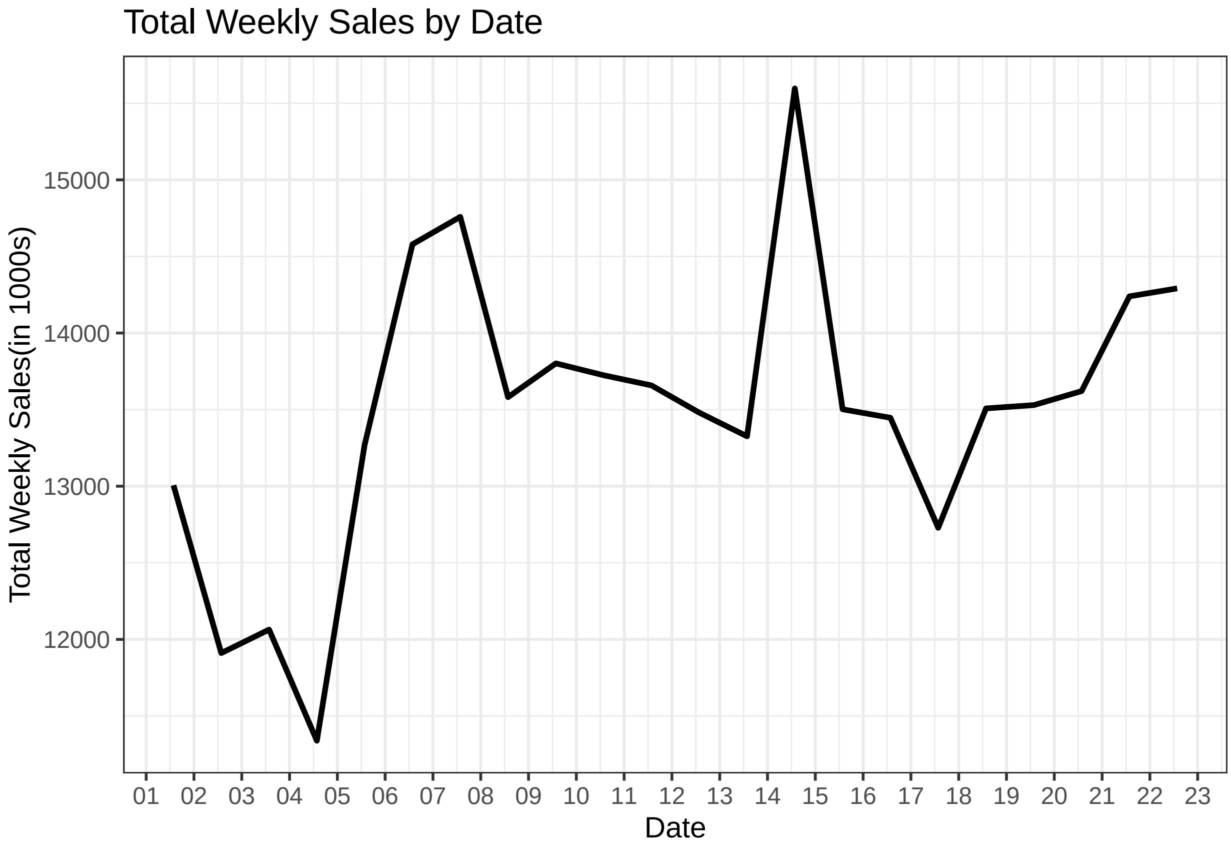
```
# Unfiltered - Plot of Weekly Sales by Date;
# vertical line indicates start of MarkDown Treatment period after 11/1/2011
  # interpretation: obvious patterns in weekly sales per month; need to breakdown further
ggplot(sales_weekly_type, aes(x = Date, y = Total_Weekly_Sales/1000)) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Total Weekly Sales by Date", y="Total Weekly Sales(in 1000s)") +
  scale_x_date(breaks = date_breaks("6 months"),
               labels = date_format("%b-%Y"))
```



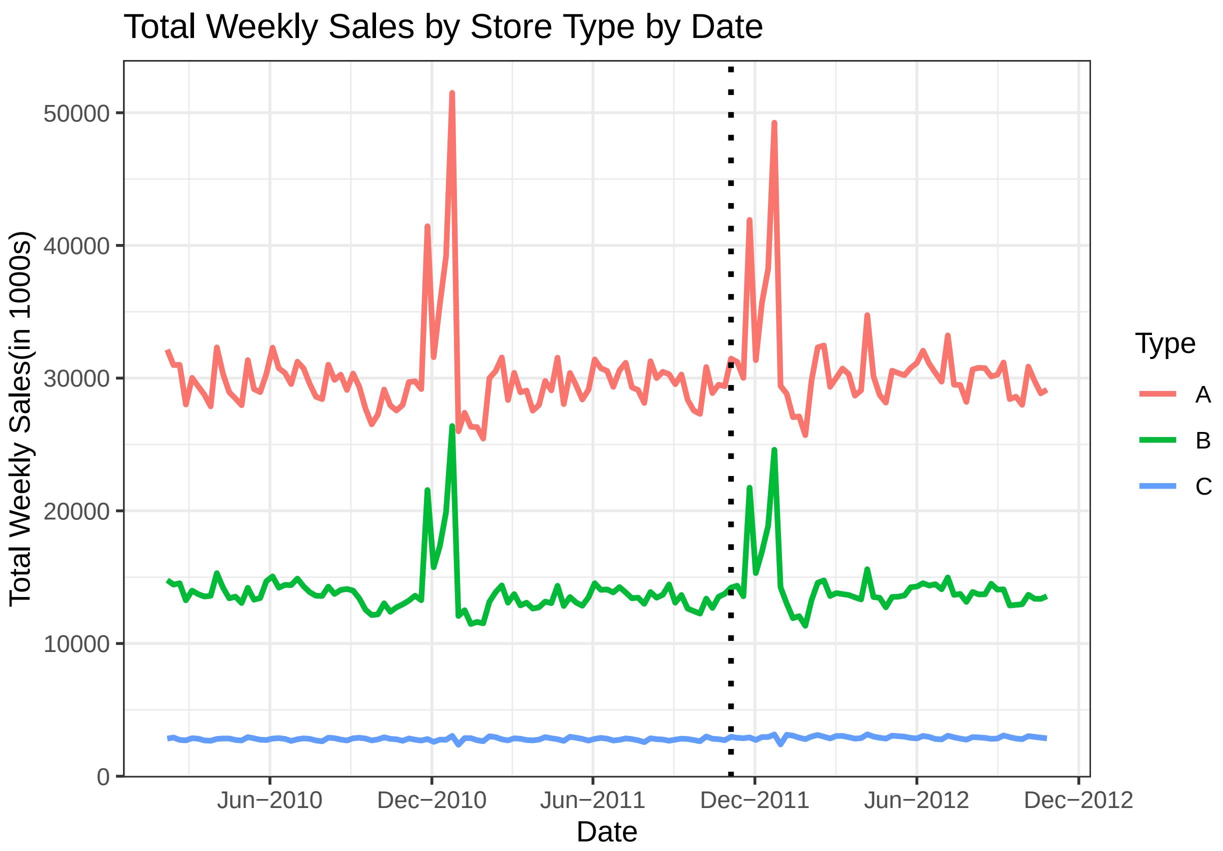
```
# Plot of Weekly Sales over a 3 month period;
# interpretation: high and low sales volumes alternate every other week
  # how does this change during the treatment period?
sales_weekly_type %>% filter(Date >='2010-06-10' & Date <='2010-10-01') %>%
  ggplot(aes(x = Date, y = Total_Weekly_Sales/1000)) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Total Weekly Sales by Date", y="Total Weekly Sales(in 1000s)") +
  scale_x_date(breaks = date_breaks("1 week"),
               labels = date_format("%U"))
```



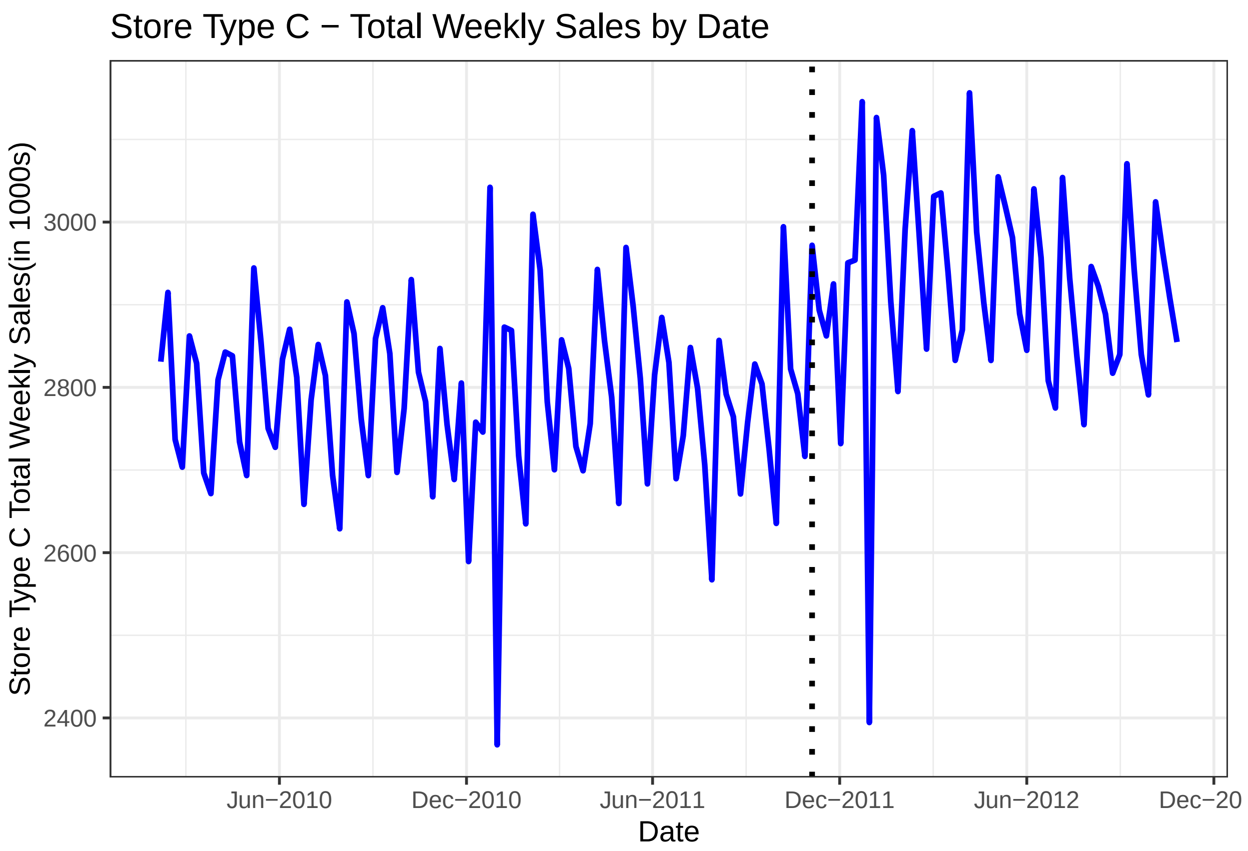
```
# Plot of Weekly Sales over a 3 month period;
# interpretation: high and low sales volumes alternate every other week
# how does this change during the treatment period?
sales_weekly_type %>% filter(Date >='2012-01-01' & Date <='2012-06-01' & Type=='B') %>%
  ggplot(aes(x = Date, y = Total_Weekly_Sales/1000)) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Total Weekly Sales by Date", y="Total Weekly Sales(in 1000s)") +
  scale_x_date(breaks = date_breaks("1 week"),
               labels = date_format("%U"))
```



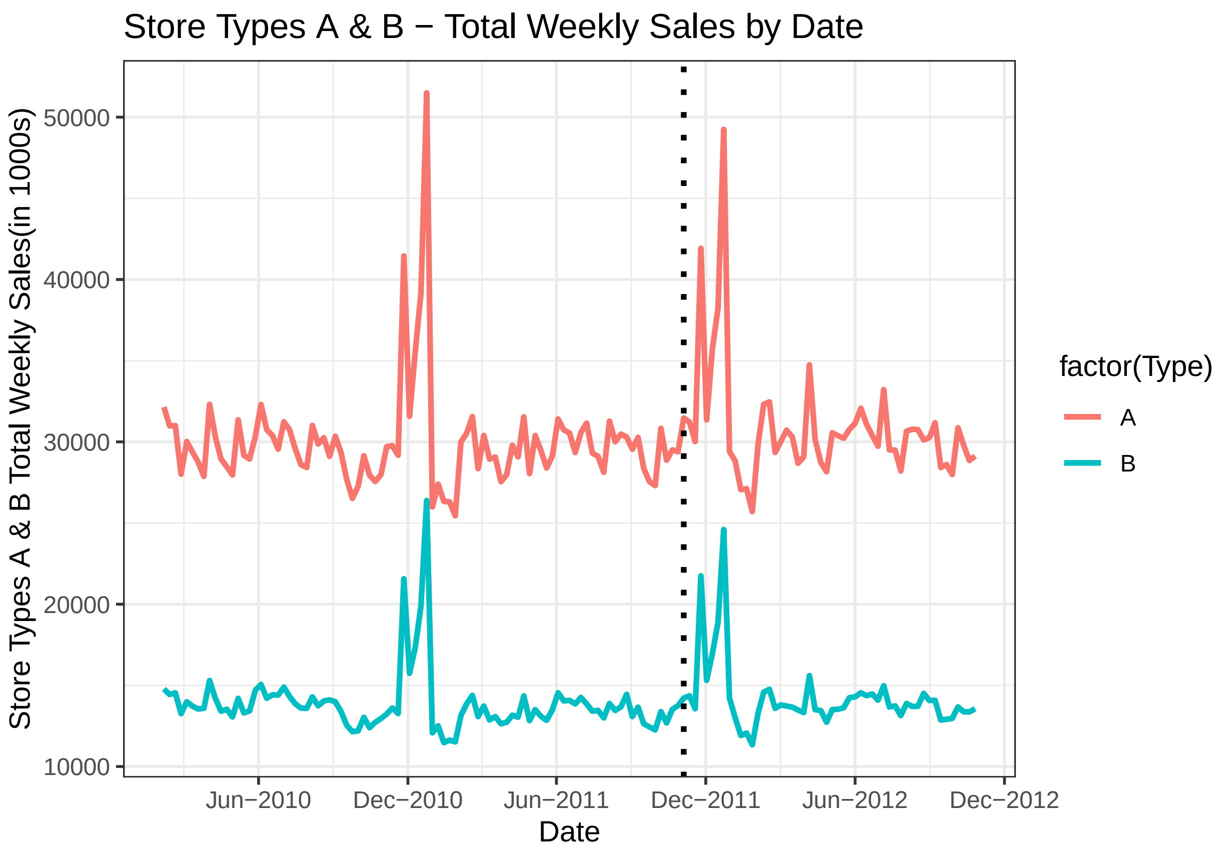
```
# All Store Types - Plot of Weekly Sales by Date;
ggplot(sales_weekly_type, aes(x = Date, y = Total_Weekly_Sales/1000, color = Type)) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Total Weekly Sales by Store Type by Date", y="Total Weekly Sales(in 1000s)") +
  scale_x_date(breaks = date_breaks("6 months"),
               labels = date_format("%b-%Y"))
```



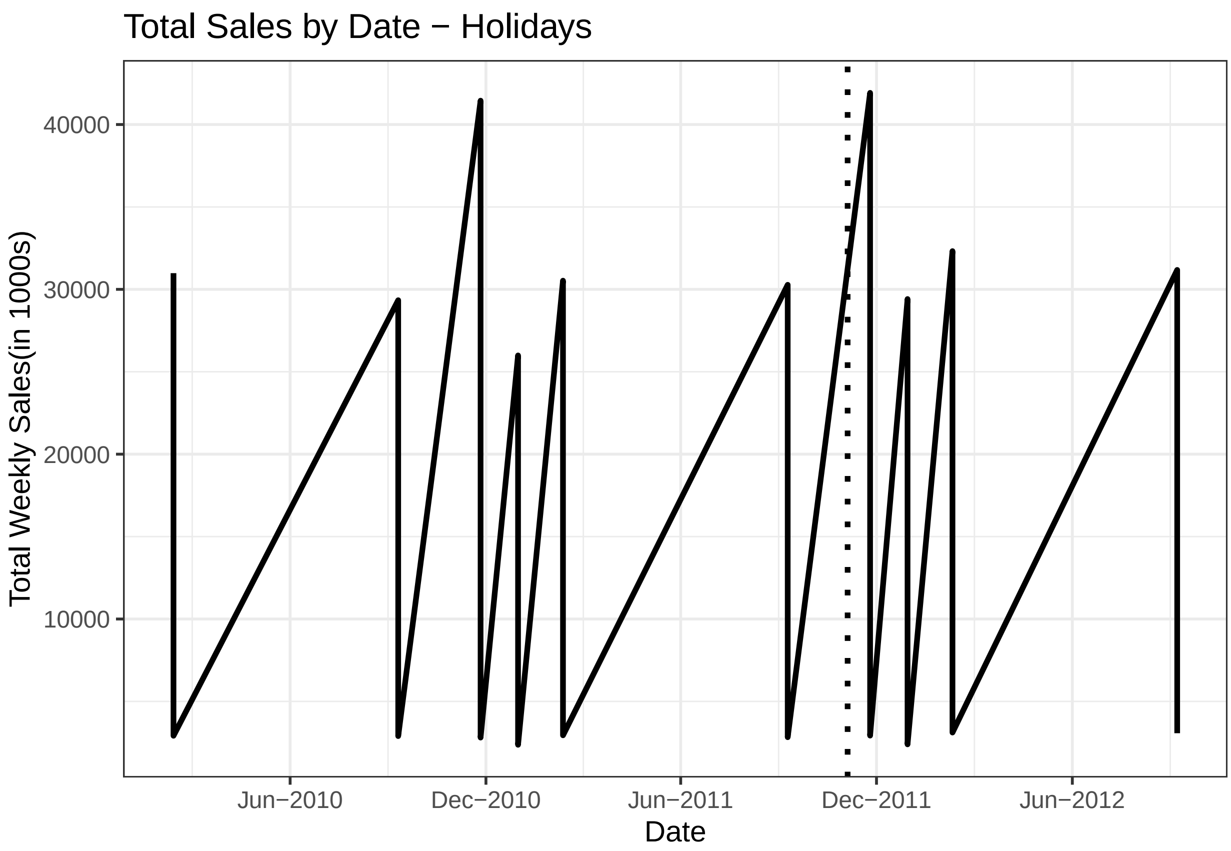
```
# hard to tell visually sales have increased during the treatment period
# Type C stores do not make up much business. May have to treat them differently,
    # or plot Type C by itself to be visually see any trending once the treatment period begun.

    # Further exploration showed that while Type C stores represent 33% of all store locations,
    # they only comprise 6% of total sales.

# Store Type C - Plot of Sales by Date
  # interpretation: different trend then Store Types A & B
    # appears visually to be slightly increase since treatment period, outside of annual sales
    # drop in December
sales_weekly_type %>% filter(Type=='C') %>%
  ggplot(aes(x = Date, y = Total_Weekly_Sales/1000)) +
  geom_line(size=1, color='blue') +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Store Type C - Total Weekly Sales by Date",
       y="Store Type C Total Weekly Sales(in 1000s)") +
  scale_x_date(breaks = date_breaks("6 months"),
               labels = date_format("%b-%Y"))
```



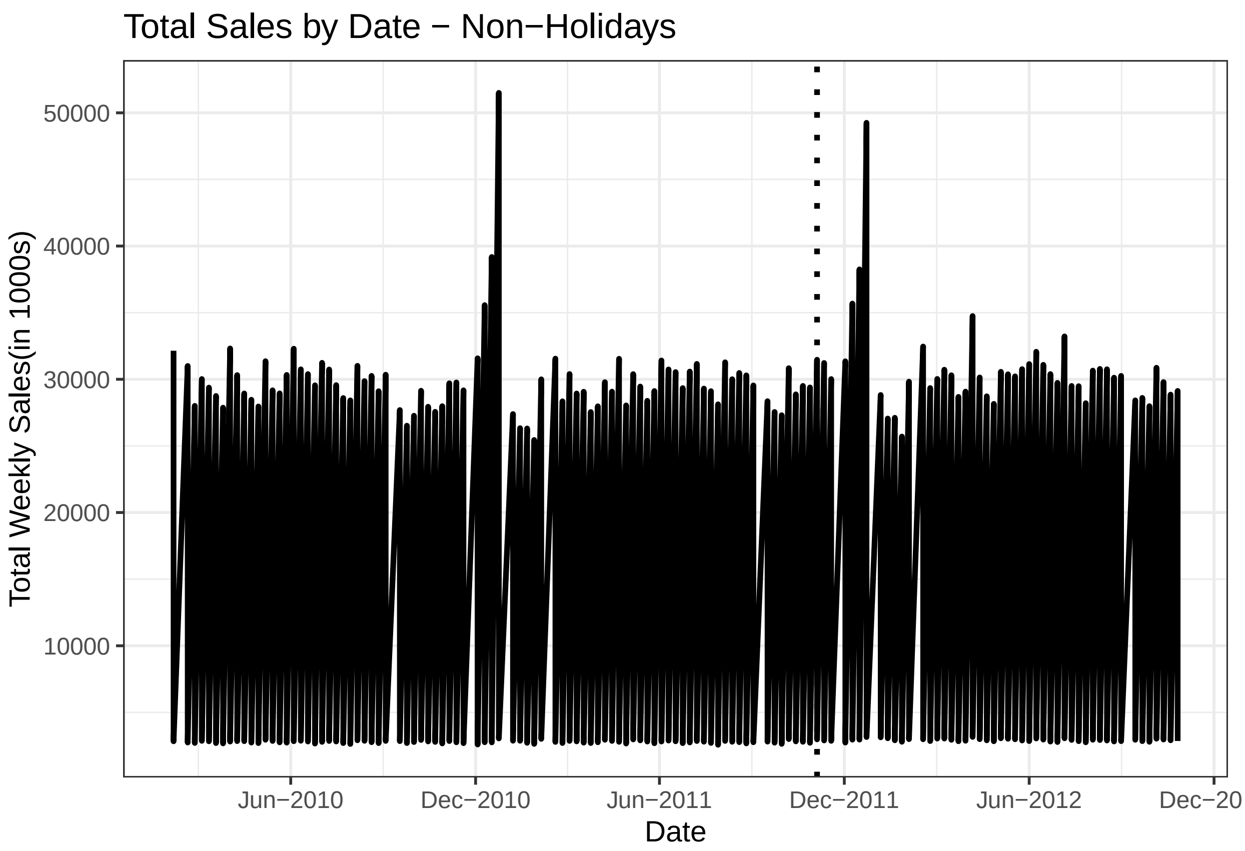
```
# Store Types A & B - Plot of Sales by Date
  # interpretation: similar sales trends over time - just different sales volumes
sales_weekly_type %>% filter(Type=='A' | Type=='B') %>%
  ggplot(aes(x = Date, y = Total_Weekly_Sales/1000, color=factor(Type))) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Store Types A & B - Total Weekly Sales by Date",
       y="Store Types A & B Total Weekly Sales(in 1000s)") +
  scale_x_date(breaks = date_breaks("6 months"),
               labels = date_format("%b-%Y"))
```



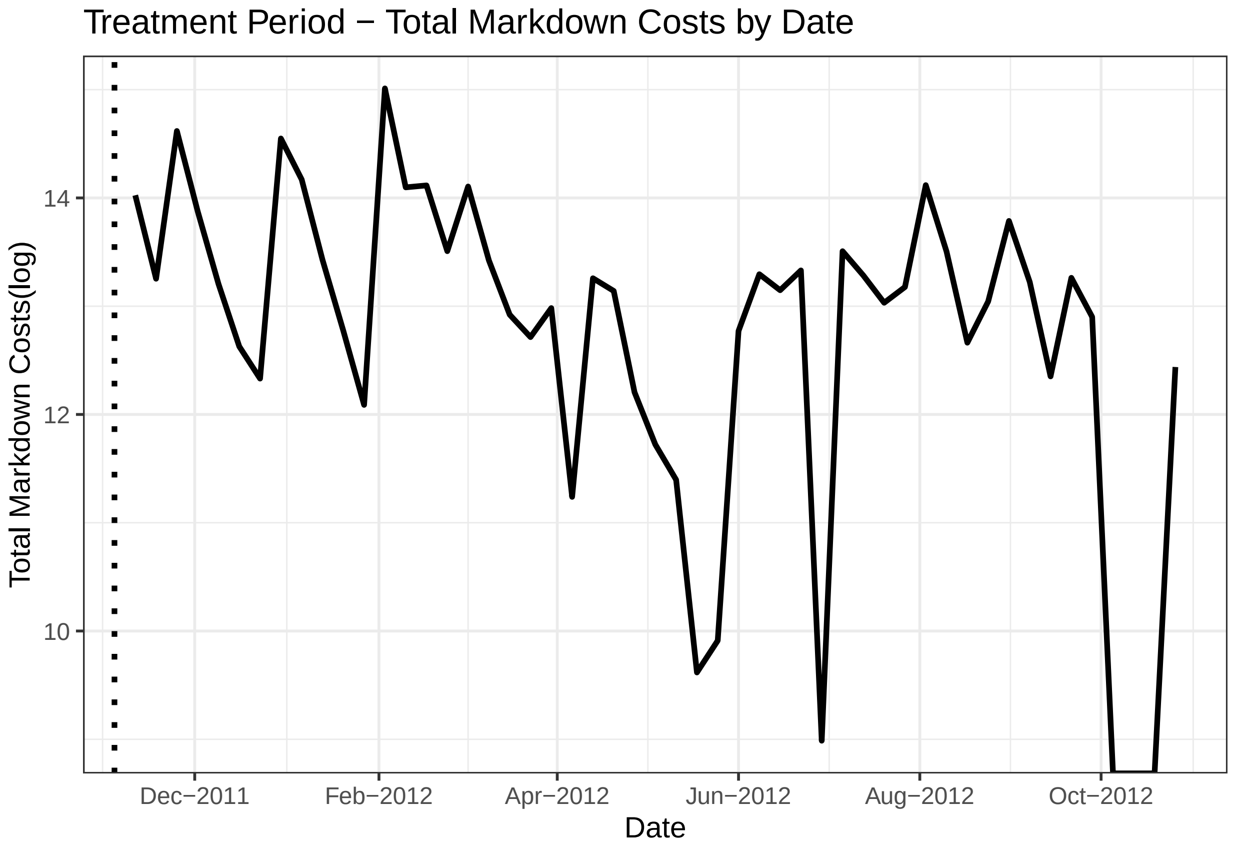
```
# Holidays - Plot of Sales by Date
  # interpretation: holiday status not consistently correlated with higher sales; expected
  # spikes in Xmas period
sales_weekly_type %>% filter(IsHoliday==1) %>%
  ggplot(aes(x = Date, y = Total_Weekly_Sales/1000)) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Total Sales by Date - Holidays", y="Total Weekly Sales(in 1000s)") +
  scale_x_date(breaks = date_breaks("6 months"),
               labels = date_format("%b-%Y"))
```



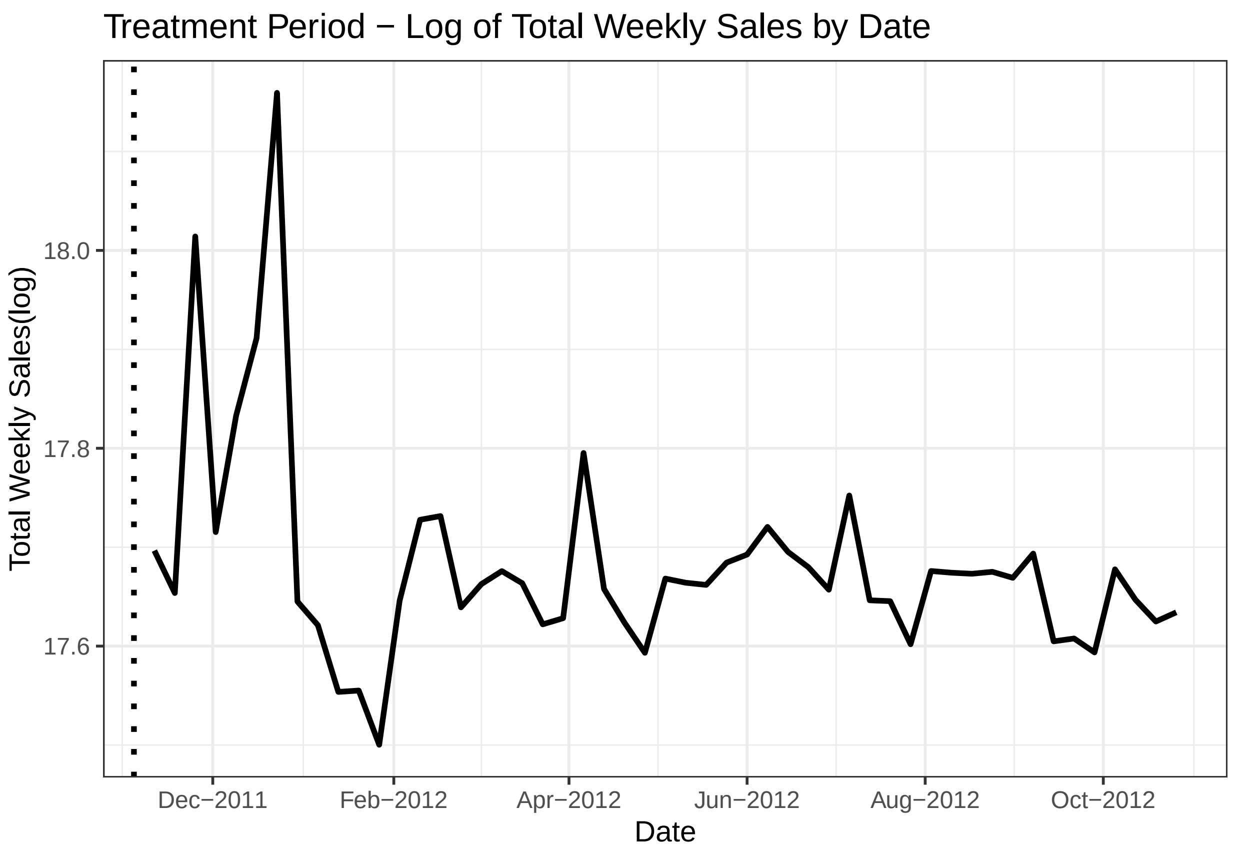
```
# Non-Holidays - Plot of Sales by Date
  # shows general bi-weekly fluctating sales pattern
sales_weekly_type %>% filter(IsHoliday==0) %>%
  ggplot(aes(x = Date, y = Total_Weekly_Sales/1000)) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Total Sales by Date - Non-Holidays", y="Total Weekly Sales(in 1000s)") +
  scale_x_date(breaks = date_breaks("6 months"),
               labels = date_format("%b-%Y"))
```



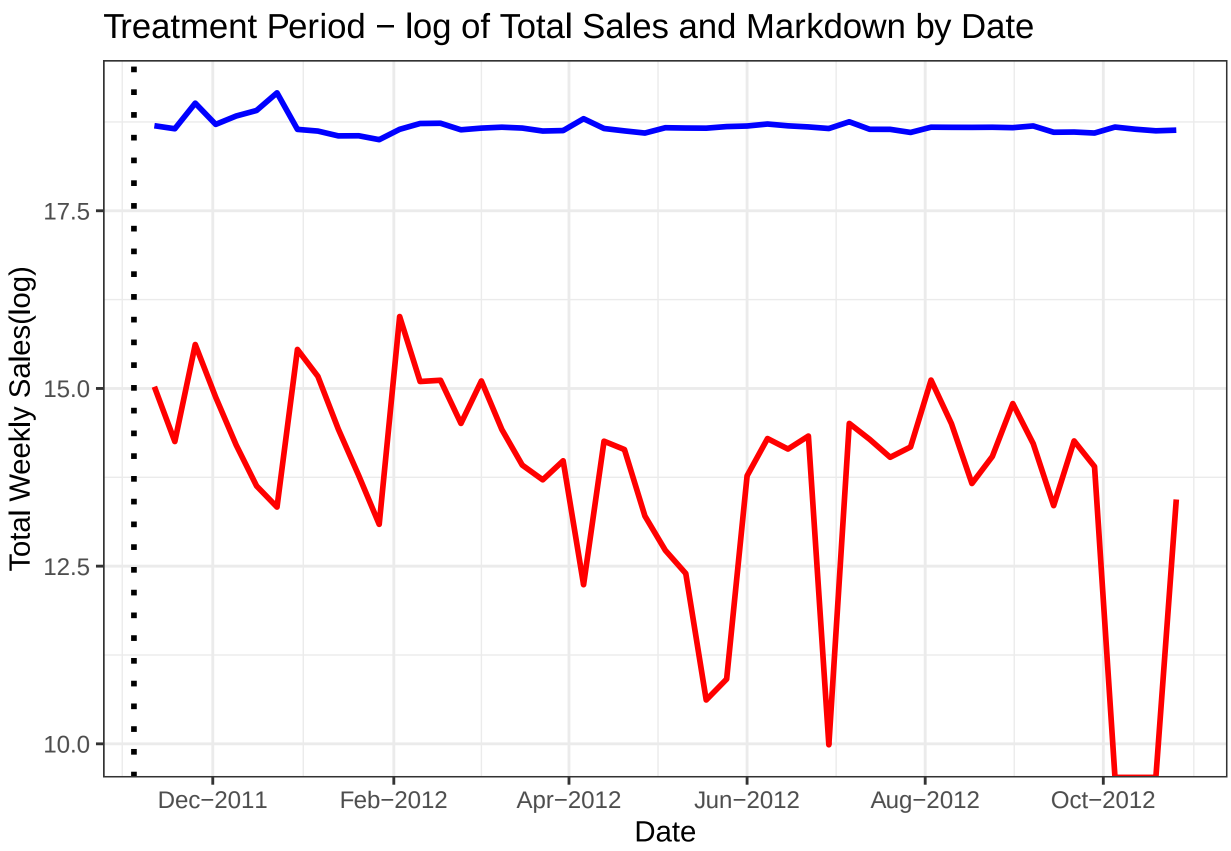
```
# Total Sales by Date by Store with MarkDown Total and MarkDown % of Sales
sales_markdown_weekly_date_store <- sales_full %>% filter(HasMarkDown==1) %>%
  select(Store, Date, Total_Weekly_Sales, MarkDown_Total, Store, IsHoliday) %>%
  mutate(MarkDown_Perc_Sales = MarkDown_Total/ Total_Weekly_Sales*100)

# Total Sales by Date with MarkDown Total and MarkDown % of Sales
sales_markdown_weekly_date <- sales_full %>% filter(HasMarkDown==1) %>%
  group_by(Date, IsHoliday) %>%
  mutate(Total_Weekly_Sales = sum(Total_Weekly_Sales),
         MarkDown_Total = sum(MarkDown_Total),
         MarkDown_Perc_Sales = (sum(MarkDown_Total)/ sum(Total_Weekly_Sales)*100)) %>%
  select(Date, Total_Weekly_Sales, MarkDown_Total, MarkDown_Perc_Sales, IsHoliday)

# Unfiltered - Plot of Weekly Total Markdown Costs by Date;
# interpretation: after costs spikes in Dec 2011, Jan 2012 and Feb 2012,
  # weekly markdown costs have declined on the whole, with occasional cost spikes not showing
  # an obvious pattern
ggplot(sales_markdown_weekly_date, aes(x = Date, y = log(MarkDown_Total))) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Treatment Period - Total Markdown Costs by Date", y="Total Markdown Costs(log)") +
  scale_x_date(breaks = date_breaks("2 months"),
               labels = date_format("%b-%Y"))
```



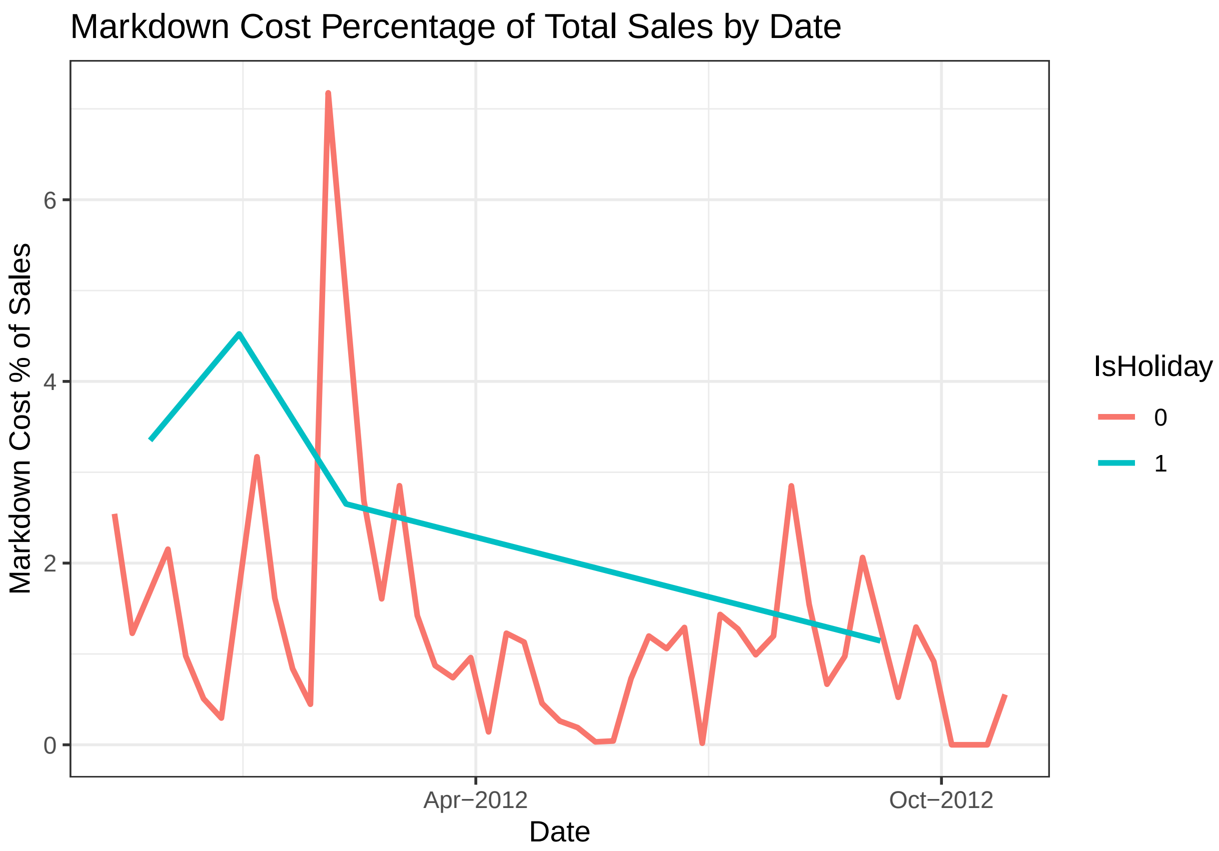
```r
# Treatment Period Sales - Plot of Weekly Total Sales by Date during the treatment period;
# interpretation: after sales peaks in December 2011/January 2012, sales fell in February.
  # While sales grew in March, no apparent upward trend in sales since March 2012
ggplot(sales_markdown_weekly_date, aes(x = Date, y = log(Total_Weekly_Sales))) +
  geom_line(size=1) +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Treatment Period - Log of Total Weekly Sales by Date", y="Total Weekly Sales(log)") +
  scale_x_date(breaks = date_breaks("2 months"),
               labels = date_format("%b-%Y"))
```



```r
# Treatment Period Sales and Markdown together - Plot of Sales & Markdown by Date during
# treatment period;
# log values interpretation: reinforces that sales have remained flat during treatment period,
# while Markdown costs have decreased
  # We have incomplete Markdown data, only for about the last 1 year in the dataset.
    # Without the inclusion of the Markdown data for the first ~365 days of data, we can't
    # tell whether the cost trend in the last 12 month is a new trend, or whether the previous 12-24
    # month period followed a similar pattern.
ggplot(sales_markdown_weekly_date) +
  geom_line(aes(x = Date, y = log(Total_Weekly_Sales)+1), size=1, color='blue') +
  geom_line(aes(x = Date, y = log(MarkDown_Total)+1), size=1, color='red') +
  geom_vline(xintercept=as.numeric(sales_weekly_type$Date[274]), linetype='dotted', size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Treatment Period - log of Total Sales and Markdown by Date",
       y="Total Weekly Sales(log)") +
  scale_x_date(breaks = date_breaks("2 months"),
               labels = date_format("%b-%Y"))
```



```
# Plot change in Markdown percentage of sales over time, with Holiday indicator
  # interpretation: markdown costs as a percentage of sales have decreased during the
  # treatment period,whether a holiday week or not.

ggplot(sales_markdown_weekly_date, aes(x = Date, y = MarkDown_Perc_Sales, color=IsHoliday)) +
  geom_line(size=1) +
  # ylim(0, 6) + xlim(2220,2233) +
  theme_bw() +
  labs(title="Markdown Cost Percentage of Total Sales by Date", y="Markdown Cost % of Sales") +
  scale_x_date(breaks = date_breaks("6 months"),
               labels = date_format("%b-%Y"))
```



```
### Regressions ###
  # haven't done a lot yet...
  # haven't included a Premium variable designation - need to determine what the Treatment Group is

# Linear Regression of Weekly Sales by the After date
did_after = lm(log(Total_Weekly_Sales+1) ~ after, data=sales_full)
summary(did_after)
```

```
##
## Call:
## lm(formula = log(Total_Weekly_Sales + 1) ~ after, data = sales_full)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44162 -0.47911  0.07347  0.46620  1.46481
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 13.690609   0.009191 1489.556   <2e-16 ***
## after        0.031078   0.015242    2.039   0.0415 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5882 on 6433 degrees of freedom
## Multiple R-squared:  0.0006459,  Adjusted R-squared:  0.0004905
## F-statistic: 4.157 on 1 and 6433 DF,  p-value: 0.04149
```

```
#interpretation:
# reinforces how sales haven't been increasing during the treatment period; only slight increase (3%)

did_after_type = lm(log(Total_Weekly_Sales+1) ~ after + Type, data=sales_full)
summary(did_after_type)
```

```
##
## Call:
## lm(formula = log(Total_Weekly_Sales + 1) ~ after + Type, data = sales_full)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76924 -0.24747  0.06587  0.29409  1.63470
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 14.018232   0.009472 1479.939   <2e-16 ***
## after        0.031078   0.012172    2.553   0.0107 *
## TypeB       -0.515915   0.012684  -40.674   <2e-16 ***
## TypeC       -0.995412   0.018091  -55.023   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4697 on 6431 degrees of freedom
## Multiple R-squared:  0.3628, Adjusted R-squared:  0.3625
## F-statistic:  1221 on 3 and 6431 DF,  p-value: < 2.2e-16
```

```
#interpretation:
# flat sales for Store Type A (3%), but large sales drops for Type B and Type C stores
  # need to validate
```