Walmart Sales Forcasting

Daniel Crofts

Libraries and Data

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(dplyr)
library(lubridate)
library(ggplot2)


dat_train <- read.csv("train.csv")
dat_test <- read.csv("test.csv")
glimpse(dat_train)
```

```
Rows: 421,570
Columns: 5
$ Store        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ Dept         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ Date         <chr> "2010-02-05", "2010-02-12", "2010-02-19", "2010-02-26", "~
$ Weekly_Sales <dbl> 24924.50, 46039.49, 41595.55, 19403.54, 21827.90, 21043.3~
$ IsHoliday    <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA~
```

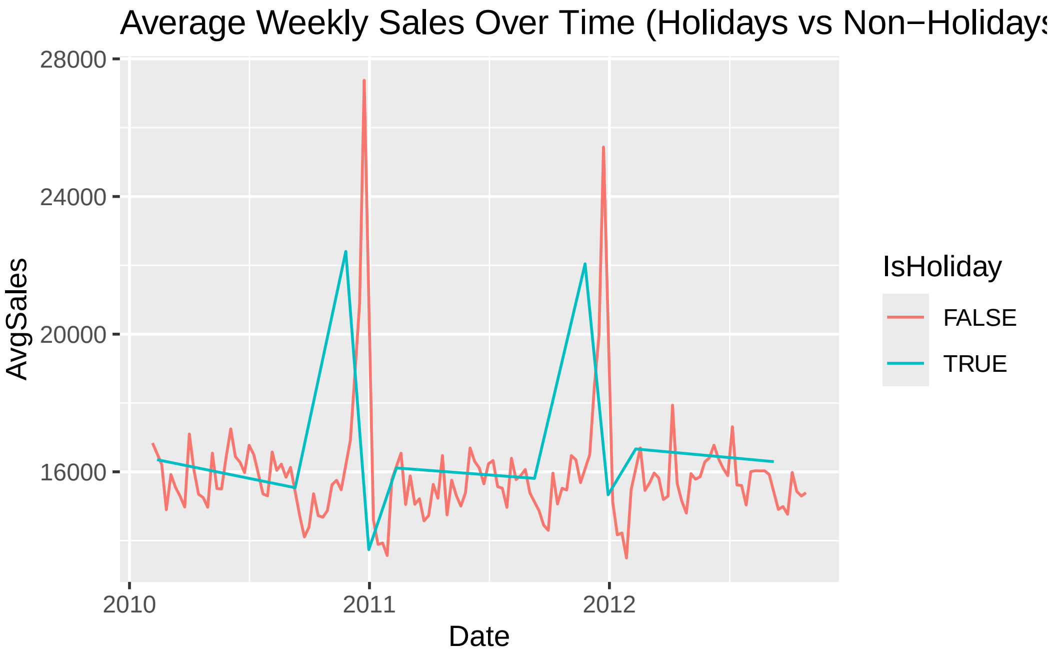
EDA

1

```
dat_train2 <- dat_train %>%
  mutate(Date = as.Date(Date),
         Week = isoweek(Date),
         Year = year(Date))

holiday_plot <- dat_train2 %>%
  group_by(Date, IsHoliday) %>%
  summarise(AvgSales = mean(Weekly_Sales), .groups = "drop")

ggplot(holiday_plot, aes(x = Date, y = AvgSales, color = IsHoliday)) +
  geom_line() +
  labs(title = "Average Weekly Sales Over Time (Holidays vs Non-Holidays)")
```

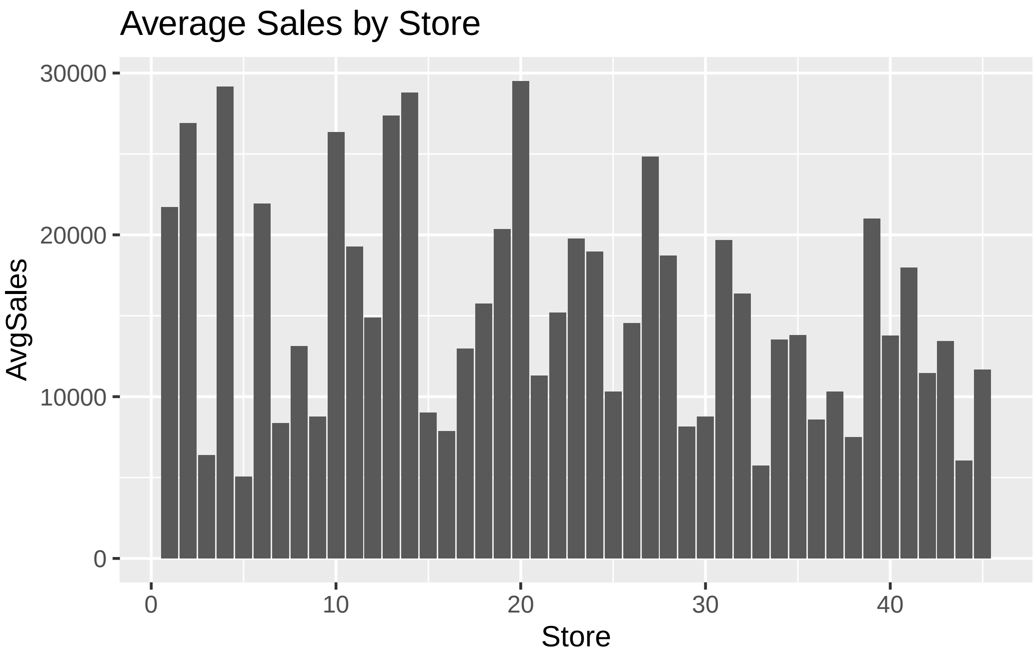


One issue we need to consider as shown in this graph is that holday weeks are weighted five times more heavily in this competition so that needs to be addressed in the model.

2

```r
store_var <- dat_train %>%
  group_by(Store) %>%
  summarise(AvgSales = mean(Weekly_Sales))


ggplot(store_var, aes(x = Store, y = AvgSales)) +
  geom_col() + labs(title = "Average Sales by Store")
```



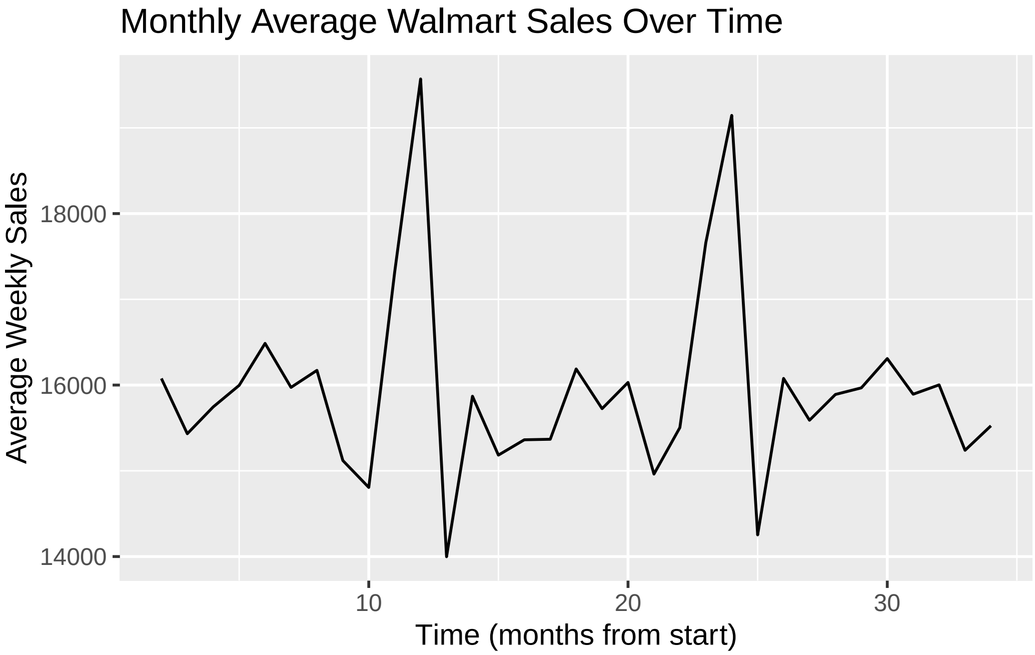
This graph shows that there is significant variation in sales accross store locations. For our predictions to be accurate we will have to account for these differences at the store level.

3

```r
dat_train <- dat_train %>%
  mutate(
    Date = as.Date(Date),
    Year  = year(Date),
    Month = month(Date),
    Week  = isoweek(Date)
  )

trend <- dat_train %>%
  group_by(Year, Month) %>%
  summarise(AvgSales = mean(Weekly_Sales), .groups = "drop") %>%
  mutate(TimeIndex = (Year - min(Year)) * 12 + Month)

ggplot(trend, aes(x = TimeIndex, y = AvgSales)) +
  geom_line() +
  labs(title = "Monthly Average Walmart Sales Over Time",
       x = "Time (months from start)",
       y = "Average Weekly Sales")
```



This EDA shows that there is also significant variation in sales over time. We must consider time related patterns when predicting the data.