Robert K Burns

DSC680 Applied Data Science

Project 1: Retail Analytics – Draft Report

December 21, 2019

Instructor: Catie Williams

Introduction

Retail Analytics is the process by which analytical data is provided and analyzed. This data can be used to report on various aspects of the retail supply chain process including sales, returns, discounts and more [Alloy Client Solutions]. Many manufacturers, retailers, and marketing teams utilize this information to detect trends and understand more about their business. One of the key aspects of this type of analytics growth in the retail industry is that customers encounter a more personalized experience when they shop [Jones, 2018]. Not only do the retailers understand their customers' needs and interests better, but just as important, their customers' behavior.

Discussion

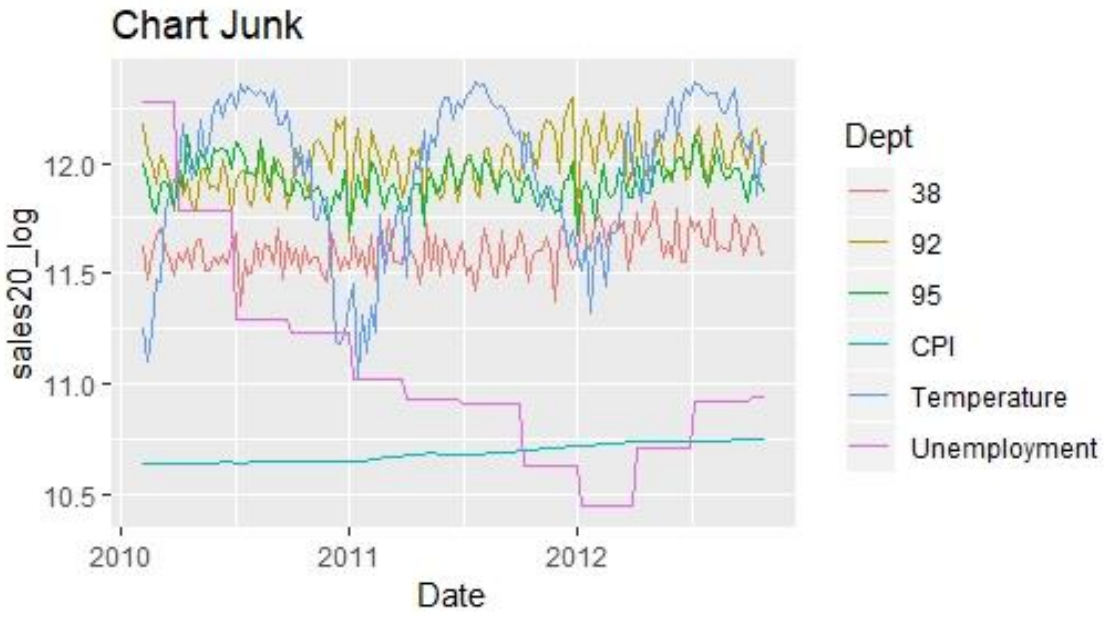
This report explores the correlation between external factors, such as the weather and the overall economy, and the retail sales experienced by the stores reported on in the dataset found on Kaggle.com/datasets/retaildataset. Understanding how customers behave can provide a great deal of insight to a retailer on staffing, resupplying of stocked goods, appropriate times to mark prices down and conduct sales and what to expect during the holiday season.

The data set that was utilized in creating this report was an anonymized report that detailed the weekly sales amounts of 45 retailers, with as many as 99 Individual departments within each of those locations. The primary focus of this report will revolve around the Top 3 retailers with the highest average weekly sales and drill down to the three departments in each of those retailers with the highest average sales, as well.

All measurements that were considered are against a timeline, but because they are all different units of measure, each feature has been log transformed so they may be appropriately visualized against the

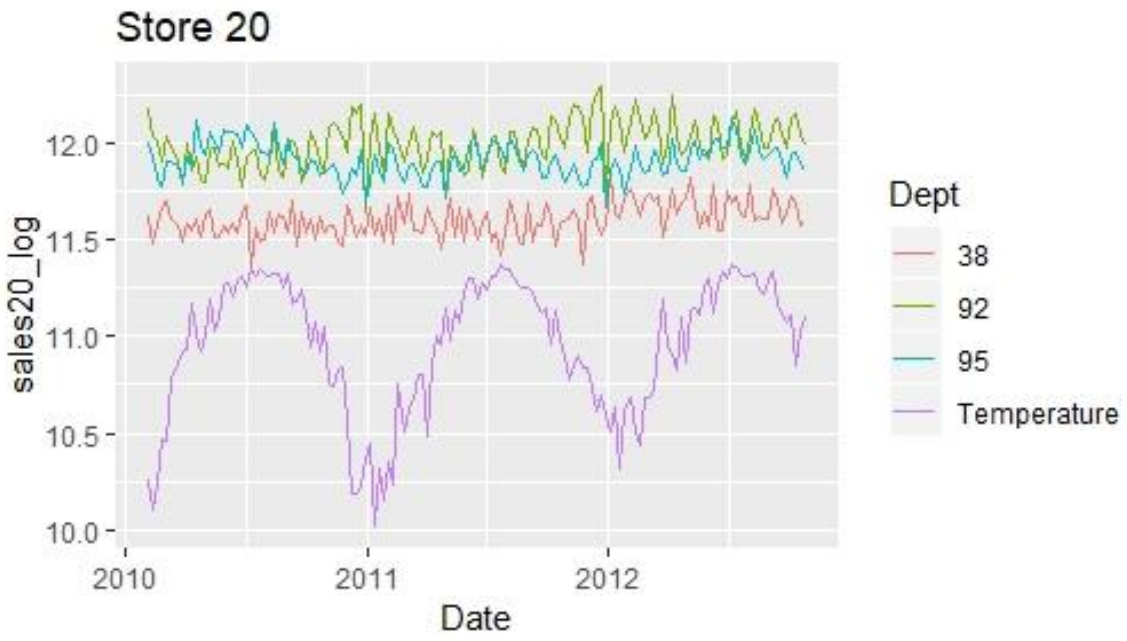
same scale. Some features had little to no change over the time period being analyzed, such as The Consumer Price Index (CPI) and proved to be of no real significance to the results.

This graph represents the overall performance of the top three departments of store #20, the store with the highest average weekly sales. To illustrate the WRONG thing to do, all three features being measured are also represented.

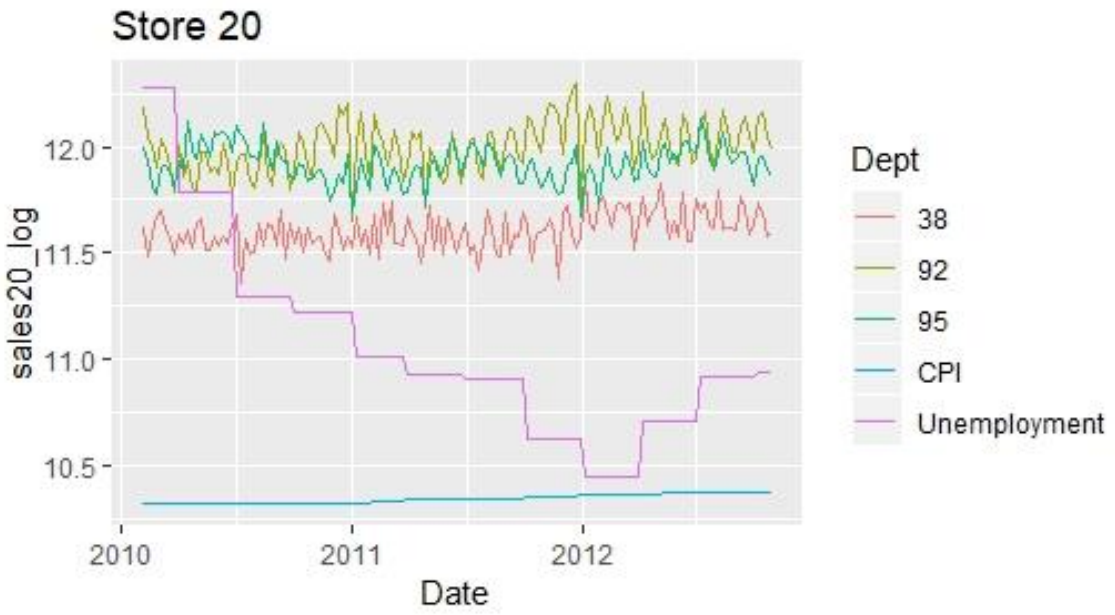


This is a lot to look at.

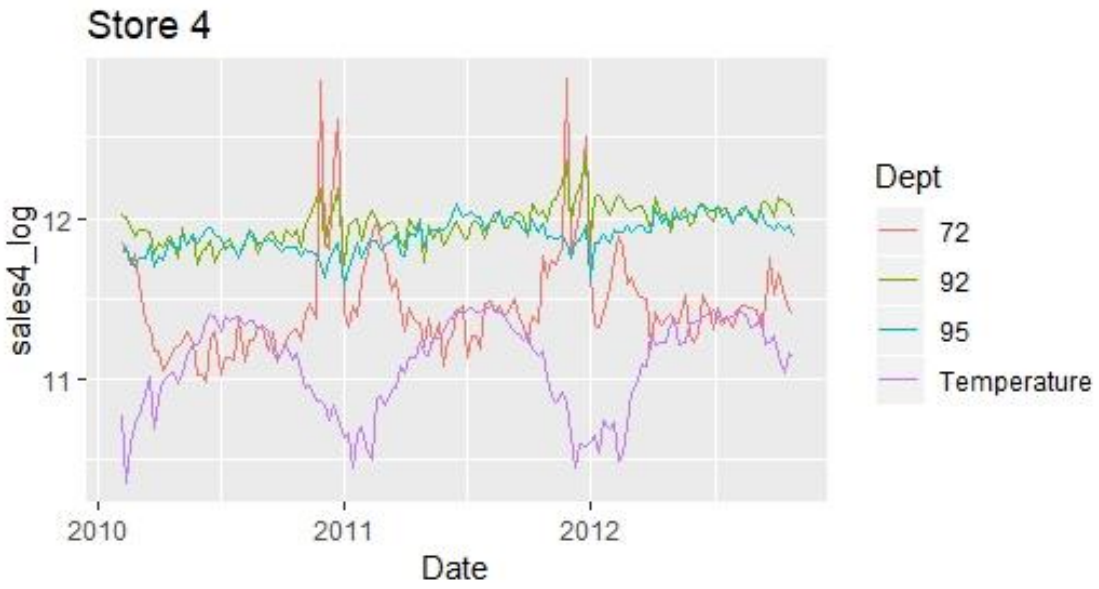
This next illustration shows how each of those departments has reacted differently with regard to the weather (average temperature). When we see how the performance varies from one store to another, even one department to another, it is important to understand that these stores are located in different areas. Based on the variation in temperature from week to week, it is safe to assume they are located in various parts of the country, so it is safe to derive that it is entirely possible that temperature will affect stores differently, but if each store understands the trends they experience as a result of these external factors, they can see how every individual department at every individual store has its own predictive analytics to develop.



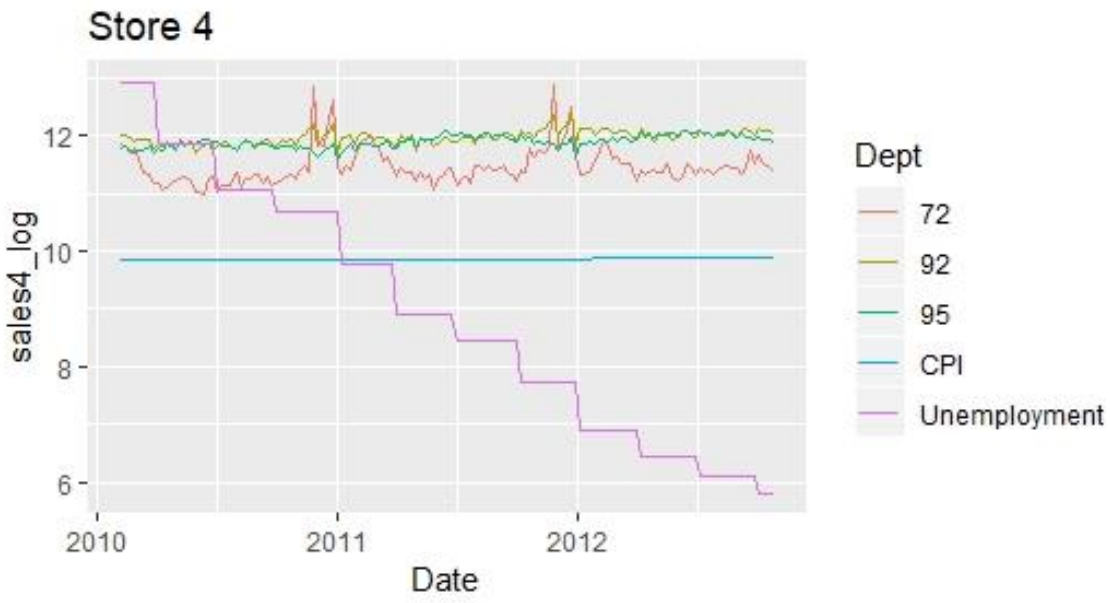
While this is still a lot, this graph only displays the top three departments at Store 20 and the temperature (in purple). For Department 92, we can see sales increase with cooler weather with slight spikes at the end of each year and more steady performance in the middle of the year, when temperatures are highest. Conversely, Department 95 shows increases in sales during warmer weather, while Department 38 shows steady performance throughout the year and an increase in overall sales in 2012.



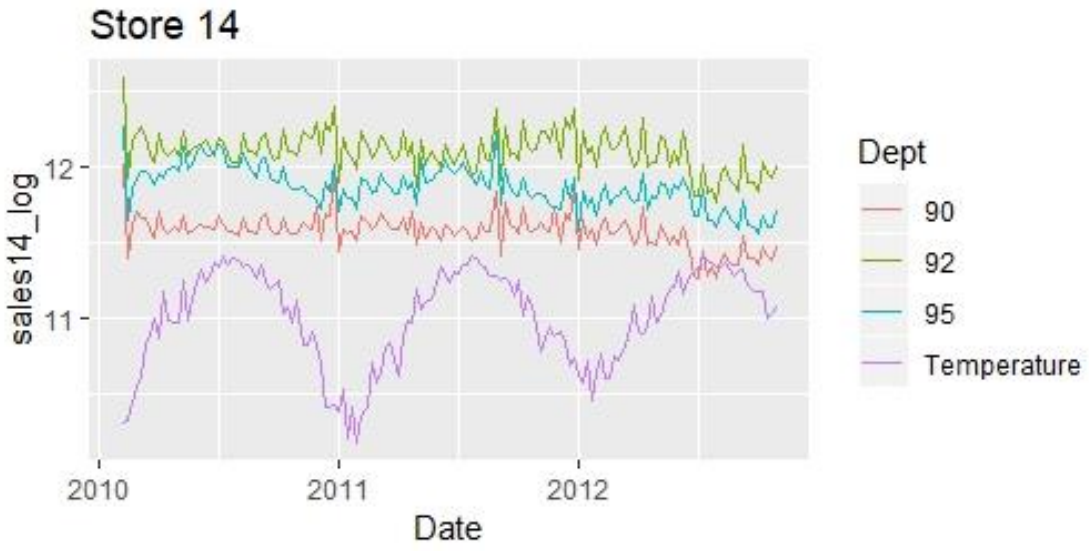
Departments 92 and 38 both show upward trends as unemployment starts to drop.



At Store 4, Department 72 showed dramatic peaks during cooler weather and a decline to steady performance when the temperatures warmed up. Department 92 also peaked in cooler temperatures but did not show as drastic changes as Department 72. Department 95 showed slight growth over the course of the three years but tended to have small dips in sales at the coldest temperatures.

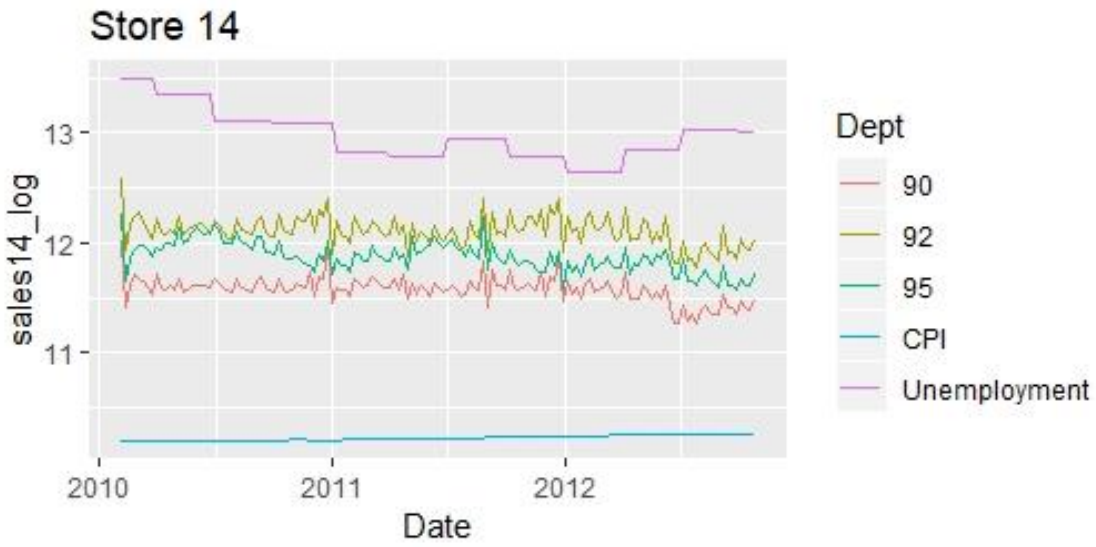


There does not appear to be any real change resulting from the CPI, which remained steady, or from the Unemployment rate dropping. Sales appear to fluctuate more so from the changes in weather.



Store 14 appears to reflect sales spikes that are consistent with the performance of Store 20. Department 92 has some slight increasing in cooler temperatures. Department 95 shows sales trends following the temperature – higher temperatures reflect higher sales and lower temperatures reflect

lower sales. Department 90 shows very steady performance with some slight peaking at years end, but all three of these departments had a decline in sales halfway through the year 2012.



There does not appear to be any clear correlation between the CPI and Unemployment rates with regard to how these features affect sales at Store 14. With all three of the top producing departments reflecting a decline in 2012, the initial observation would be that these changes are consistent with the increase in Unemployment, but by the end of 2012, Unemployment rates were not has high as they were in 2010, when the store reported better performance.

Conclusion

Should any one of these stores decide to pursue answers regarding sales performance, it is clear analytics can be broken down by specific department, allowing those retailers to understand how each department functions as a result of outside factors. Further analysis can be performed with data reflecting sale (markdown) periods and gain additional insight as to when the best time to conduct a sale would be on an individual department level.

As this report is not nearly long enough to be all inclusive, it should be known the code used to produce this information is completely reproduceable by selecting any store number and department numbers the user wishes to examine.

References:

Alloy Client Solutions. March 21, 2019. "Retail analytics trends in 2019 and beyond." https://medium.com/alloytech/retail-analytics-trends-in-2019-and-beyond-c671783c67b7

Jones, Joshua. March 2, 2018. "Fast Forward: How retailers will use data and analytics to succeed in 2018 and beyond." https://www.strategywise.com/fast-forward-how-retailers-will-use-data-and-analytics-to-succeed-in-2018-and-beyond/

Appendix

Using R to process the data frames:

```
sales_df <- read_csv("sales data-set.csv")
features_df <- read_csv('Features data set.csv')
```

I then made some adjustments to the features:

```
sales_df <-
  sales_df %>%
  mutate(Store = as.factor(Store),
         Dept = as.factor(Dept),
         Date = as.Date(Date, "%d/%m/%Y"))
```

```
features_df <-
  features_df %>%
  mutate(Store = as.factor(Store),
         Date = as.Date(Date, "%d/%m/%Y"))
```

I used Excel to find the highest grossing stores and the highest grossing departments within those top three stores. I then used R to create individual dataframes for each of those stores with those top departments.

```
sales_trimmed_20 <-
  sales_df %>%
  filter(Store == 20,
         Dept == 92 | Dept == 95 | Dept == 38) %>%
  dplyr::select(-IsHoliday)
```

```
sales_trimmed_4 <-
  sales_df %>%
  filter(Store == 4,
         Dept == 92 | Dept == 95 | Dept == 72) %>%
  dplyr::select(-IsHoliday)
```

```
sales_trimmed_14 <-
  sales_df %>%
  filter(Store == 14,
         Dept == 92 | Dept == 95 | Dept == 90) %>%
  dplyr::select(-IsHoliday)
```

```
features_trimmed_20 <-
  features_df %>%
  filter(Store == 20)


# df of store 4 features
features_trimmed_4 <-
  features_df %>%
  filter(Store == 4)


# df of store 14 features
features_trimmed_14 <-
  features_df %>%
  filter(Store == 14)
```

```
# cobine the two df's for store 20
dfcombo20 <- merge(sales_trimmed_20, features_trimmed_20, by = "Date")


# cobine the two df's for store 4
dfcombo4 <- merge(sales_trimmed_4, features_trimmed_4, by = "Date")

# cobine the two df's for store 14
dfcombo14 <- merge(sales_trimmed_14, features_trimmed_14, by = "Date")
```

I then performed log transformations so I could represent different measurements on the same scale.

```r
# Log transformations
# Log transformatios for store 20
sales20_log <- log(dfcombo20$Weekly_Sales)
temp20_log <- log(dfcombo20$Temperature)
cpi20_log <- log(dfcombo20$CPI)
unem20_log <- log(dfcombo20$Unemployment)


# Add log tranformed features for store 20 to df
dfcombo20["sales20_log"] <- sales20_log
dfcombo20["temp20_log"] <- temp20_log
dfcombo20["cpi20_log"] <- cpi20_log
dfcombo20["unem20_log"] <- unem20_log
```

```r
# Log transformatios for store 4
sales4_log <- log(dfcombo4$Weekly_Sales)
temp4_log <- log(dfcombo4$Temperature)
cpi4_log <- log(dfcombo4$CPI)
unem4_log <- log(dfcombo4$Unemployment)


# Add log tranformed features for store 4 to df
dfcombo4["sales4_log"] <- sales4_log
dfcombo4["temp4_log"] <- temp4_log
dfcombo4["cpi4_log"] <- cpi4_log
dfcombo4["unem4_log"] <- unem4_log
```

```
# Log transformatios for store 14
sales14_log <- log(dfcombo14$Weekly_Sales)
temp14_log <- log(dfcombo14$Temperature)
cpi14_log <- log(dfcombo14$CPI)
unem14_log <- log(dfcombo14$Unemployment)


# Add log tranformed features for store 14 to df
dfcombo14["sales14_log"] <- sales14_log
dfcombo14["temp14_log"] <- temp14_log
dfcombo14["cpi14_log"] <- cpi14_log
dfcombo14["unem14_log"] <- unem14_log
```

Then, I created charts to show the relationships between these features:

This first one was junk:

```
# plot 1 - chart junk
ggplot(dfcombo20, aes(x = Date)) +
  labs(title = "Chart Junk") +
  geom_line(aes(y = sales20_log, colour = Dept)) +
  geom_line(aes(y = temp20_log + 8, colour = 'Temperature')) +
  geom_line(aes(y = cpi20_log * 2, colour = 'CPI')) +
  geom_line(aes(y = Unemployment * 1.5, colour = 'Unemployment'))
```

The rest made better sense.

```
# effects of external factors on retail
#_________________Store 20_____________________________
# plot data of store 20 with temperature
ggplot(dfcombo20, aes(x = Date)) +
  labs(title = "Store 20") +
  geom_line(aes(y = sales20_log, colour = Dept)) +
  geom_line(aes(y = temp20_log + 7, colour = 'Temperature'))


# plot data of store 20 with unemployent and CPI
ggplot(dfcombo20, aes(x = Date)) +
  labs(title = "Store 20") +
  geom_line(aes(y = sales20_log, colour = Dept)) +
  geom_line(aes(y = cpi20_log + 5, colour = 'CPI')) +
  geom_line(aes(y = Unemployment * 1.5, colour = 'Unemployment'))
```

```r
#_________________Store 4_____________________________
# plot data of store 4 with temperature
ggplot(dfcombo4, aes(x = Date)) +
  labs(title = "Store 4") +
  geom_line(aes(y = sales4_log, colour = Dept)) +
  geom_line(aes(y = temp4_log + 7, colour = 'Temperature'))


# plot data of store 4 with unemployent and CPI
ggplot(dfcombo4, aes(x = Date)) +
  labs(title = "Store 4") +
  geom_line(aes(y = sales4_log, colour = Dept)) +
  geom_line(aes(y = cpi4_log + 5, colour = 'CPI')) +
  geom_line(aes(y = Unemployment * 1.5, colour = 'Unemployment'))
```

```r
#_________________Store 14_____________________________
# plot data of store 14 with temperature
ggplot(dfcombo14, aes(x = Date)) +
  labs(title = "Store 14") +
  geom_line(aes(y = sales14_log, colour = Dept)) +
  geom_line(aes(y = temp14_log + 7, colour = 'Temperature'))


# plot data of store 14 with unemployent and CPI
ggplot(dfcombo14, aes(x = Date)) +
  labs(title = "Store 14") +
  geom_line(aes(y = sales14_log, colour = Dept)) +
  geom_line(aes(y = cpi14_log + 5, colour = 'CPI')) +
  geom_line(aes(y = Unemployment * 1.5, colour = 'Unemployment'))
```