UNDERSTANDING DATA AND ITS ENVIRONMENT: REPORT ASSESSMENT

## 1. Introduction

Forecasting or making predictions of sales are fundamental and very challenging for retailers around the world (reference needed).

This report explains the approach taken to forecast sales for a nationwide retailer in the U.S based on historical sales data for the departments of 45 stores.

The consideration of the effects of promotional activities is an additional difficulty in the analysis given the fact that part of the promotion related data is absent from historical records.

All task has been carried out using R. The report is reproducible based on code available at <https://github.com/eugenividal/Understanding-data-report>.

The whole process is described in the sections below.

## 2. Data description

The first stage before describing the data is to load it into the R environment. To to that, we will use the tidyverse package and load it with the library()function:

library(tidyverse)  
# load data  
source("code/load-data.R")

We can check that these files have been loaded with the following command:

ls()

## [1] "features" "stores" "test" "train"

Four data sets have been provided about the company. These data sets have information about the stores, features… Each of the data sets and its features are explained below:

• stores.csv

* Store: the anonymised store number
* Type: store type, A: supercentre, B: superstore, C: supermarket
* Size : store size (in square feet)

• features.csv

* Store: the anonymised store number
* Date: the week with the dated Friday
* Temperature: average temperature in the region
* Fuel\_Price: cost of fuel in the region
* Promotions: anonymised data related to promotions, mainly price reductions that the retailer is running. Promotion 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

• train.csv

* Store: the anonymised store number
* Department: the anonymised department number
* Date: the week with the dated Friday
* Weekly\_Sales: sales for the given department in the given store
* IsHoliday: whether the week is a special holiday week

• test.csv

The validation dataset have the same fields as the train.csv, except we need to predict the weekly sales for each triplet of store, department, and date from 02/11/2012 to 26/07/2013.

## 3. Data preparation

In this step we will carry out some “techniques”" to arrange the data in a way that makes the analysis easier and to produce a better model.

First of all, we will join the three data sets (stores, train and features) by Store which is the common column.No difficulties of Data Integration were found.

# join store-level data onto training dataset (so we know size)  
train\_joined = inner\_join(train, y = stores)

## Joining, by = "Store"

# would use rename() function to rename columns if needed (not needed)  
train\_joined = inner\_join(train\_joined, y = features)

## Joining, by = c("Store", "Date", "IsHoliday")

The result is a data frame with all the features that contains 421570 observations.

Secondly, we will clean the data.

Thirdly, we will generate new features: 1) Include a Week Number of the year (code needed); 2) Add a return column (code needed).

Finally, we will reduce the data set. After performing this procedure we have x observations which makes our data more manageable for further analysis.

And this is the way the data set frame looks once pre-processed (code is needed).

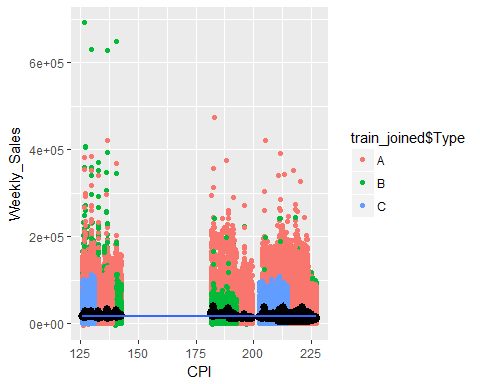
## 4. Identifying the key factors

Graphs to see the correlation between the different variables

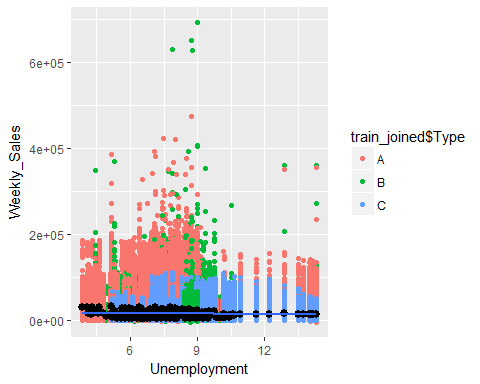
There are many packages and approach for forecasting. We could use the lm() function to do a linear regression, for example. Here we use the xgboost package

install.packages(“Hmisc”)

# Plot weekly sales vs CPI  
ggplot(train\_joined,aes(x= CPI, y= Weekly\_Sales)) + geom\_point(aes(color=train\_joined$Type)) +stat\_summary(fun.data=mean\_cl\_normal) +   
 geom\_smooth(method='lm')

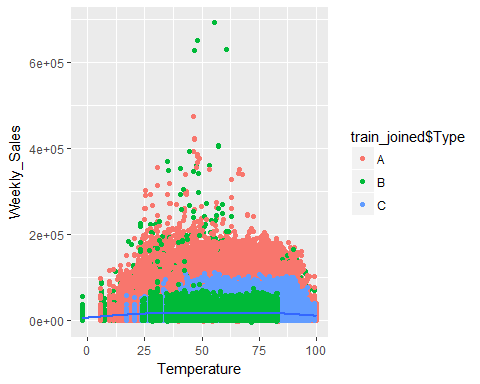


# Plot weekly sales vs Unemployment  
ggplot(train\_joined,aes(x= Unemployment, y= Weekly\_Sales)) + geom\_point(aes(color=train\_joined$Type))+stat\_summary(fun.data=mean\_cl\_normal) +   
 geom\_smooth(method='lm')



# Plot weekly sales vs Temperature  
ggplot(train\_joined,aes(x= Temperature, y= Weekly\_Sales)) + geom\_point(aes(color=train\_joined$Type)) + geom\_smooth()

## `geom\_smooth()` using method = 'gam'



There is no clear correlation between the varibles previously graphed. BUt there is correlation between…

Correlation matrix between all of our numerical features?

## 5. Creating the predictive model

m3 = lm(Weekly\_Sales ~ Dept + Store + Type + Promotion1 + Promotion2 + Promotion3 + Promotion4 + Promotion5 + CPI + Unemployment, data = train\_joined)  
summary (m3)

##   
## Call:  
## lm(formula = Weekly\_Sales ~ Dept + Store + Type + Promotion1 +   
## Promotion2 + Promotion3 + Promotion4 + Promotion5 + CPI +   
## Unemployment, data = train\_joined)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41046 -14496 -7194 6190 593823   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.715e+04 6.143e+02 44.205 < 2e-16 \*\*\*  
## Dept 1.097e+02 2.555e+00 42.945 < 2e-16 \*\*\*  
## Store -1.490e+02 6.753e+00 -22.064 < 2e-16 \*\*\*  
## TypeB -8.599e+03 1.636e+02 -52.575 < 2e-16 \*\*\*  
## TypeC -7.740e+03 8.340e+02 -9.281 < 2e-16 \*\*\*  
## Promotion1 7.507e-02 1.513e-02 4.961 7.01e-07 \*\*\*  
## Promotion2 2.446e-02 7.686e-03 3.183 0.00146 \*\*   
## Promotion3 1.406e-01 7.101e-03 19.803 < 2e-16 \*\*\*  
## Promotion4 -2.093e-02 1.928e-02 -1.086 0.27763   
## Promotion5 1.355e-01 1.212e-02 11.180 < 2e-16 \*\*\*  
## CPI -3.629e+01 2.080e+00 -17.447 < 2e-16 \*\*\*  
## Unemployment -3.997e+02 4.828e+01 -8.277 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23970 on 97044 degrees of freedom  
## (324514 observations deleted due to missingness)  
## Multiple R-squared: 0.06112, Adjusted R-squared: 0.06102   
## F-statistic: 574.3 on 11 and 97044 DF, p-value: < 2.2e-16

Linear model to find a specific value for Weekly Sales that we want to predict?

## 6. Evaluation of forecasting accuracy

## 7. Conclusions