

ROSSMANN STORE SALES

Forecast sales using store, promotion, and competitor data



CIS 5270: Business Intelligence (Spring 2017)

By: Niklas Melcher & Yashaswi Ananth



SUBMITTED TO – DR. SHILPA BALAN

CALIFORNIA STATE UNIVERSITY, LOS ANGELES.

# **DATASET**

This dataset that was collected from Kaggle, holds the daily sales report of Rossmann drug store. Rossmann is a nationwide chain of German chemists operating over 3,000 drug stores over 7 European countries. This dataset was originally created for the managers to predict daily sales statistics upto 6 weeks in advance. The goal was to analyze data and provide stores with information to make confident inventories and merchandise decisions that will lead to improved sales, profit, efficiency, and customer satisfaction. Their store sales are influenced by many factors and predictions were to be based on several criterias such as promotions, competition, school and state holidays, seasonality, and locality. The data for analysis is collected for the year 2013, 2014, and 2015.

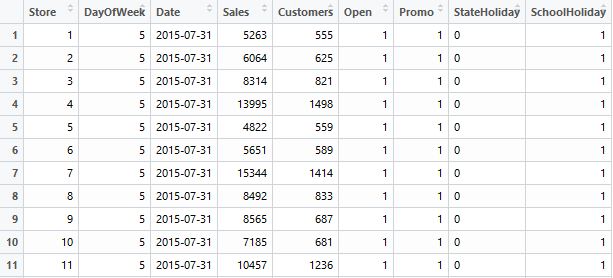
The dataset contains colums like ‘DayOfWeek’ that has values from 1 to 7 representing the 5 days of the week starting Monday to Sunday respectively. The column ‘Sales’ gives the turnover for that particular day. ‘Open’ indicates if the store was open or closed for the day. Value 1 represents open and 0 represents closed. Column ‘Customers’ gives the number of customers for the day. The ‘StateHoiday’ column indicates a state holiday (Eg. New years, Good Friday, Easter Monday etc.). Normally all stores, except for few were found to be closed on state holidays. Whereas, all schools were closed on public holidays and weekends. The calues for the StateHoliday column are a = public holiday, b = Easter holiday, c = Christmas and 0 = None. The promo column describes of the store was running a promo on that particular day or not. 1 suggests that it was a promo day at the store and 0 indicates a no-promo day. The ‘Store’ column is the primary key of the dataset which gives a unique ID to every store. This dataset is of the size 37.16 KB and is in a .CSV format.

# **DATA CLEANING**

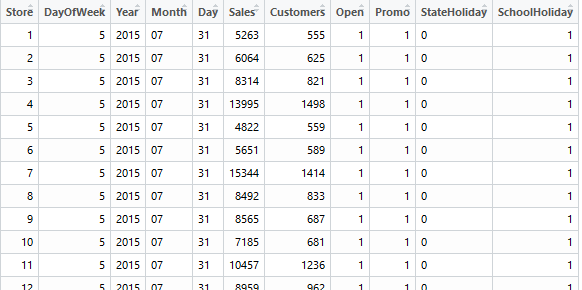
Below were the modifications done to the data during the process of data cleaning.

1. **Splitting the columns:**

**Code -** *cl1\_df<-separate(data,Date, c("Year", "Month", "Day"), sep = "-")*

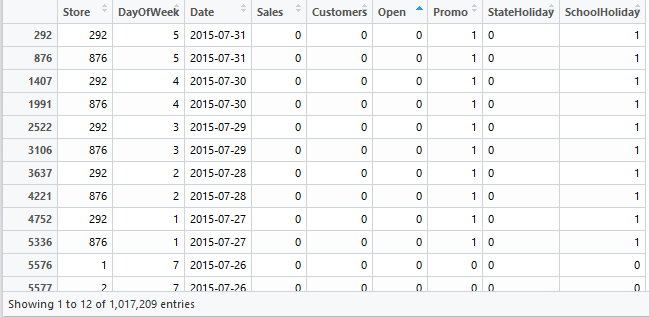


The Date column originally contained the year, month and day separated by a ‘ - ‘. This column was split into three different columns names ‘Year’, ‘Month’, and ‘Day’ for better understanding and data analysis

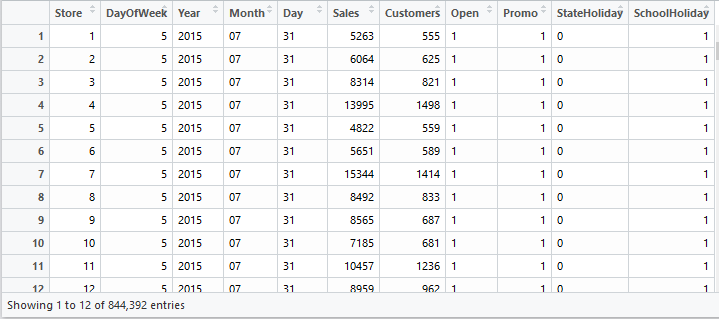


1. **Deleting unnecessary rows:**

**Code** *-*



The dataset contained reports of days when the store was closed (Open = 0). Other column values in corresponding to these rows (Eg. Sales and number of customers for the day) were insignificant and did not add any value to our analysis. Hence, these rows were eliminated and analysis was targeted only towards the days the store was open and functioning.



1. **Replacing string with numerical values:**

**Code** *–*

*just\_open<-subset(cl1\_df, Open == 1)*

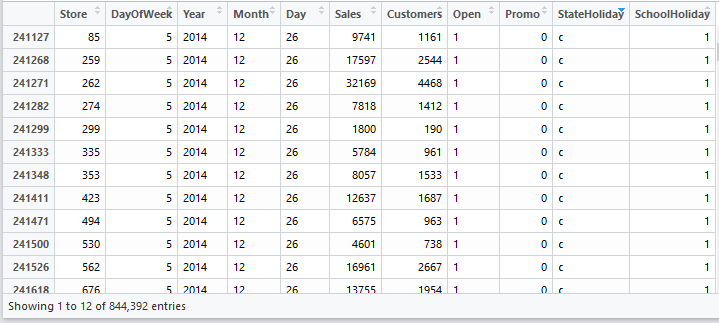
*df<-just\_open*

*df[df$StateHoliday == "a",]$StateHoliday = 1*

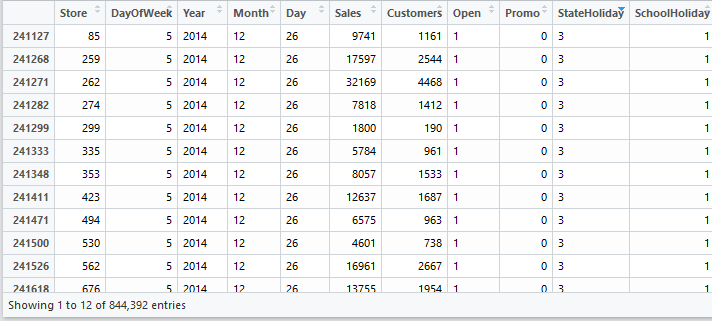
*df[df$StateHoliday == "b",]$StateHoliday = 2*

*df[df$StateHoliday == "c",]$StateHoliday = 3*

*df[df$StateHoliday == "0",]$StateHoliday = 0*



The column ‘StateHoliday’ consisted of values a = public holiday, b = Easter holiday, c = Christmas, and 0 = None. These string values were changed to 1, 2, 3, and 4 respectively for better and faster calculations.



**ANALYSIS AND VISUALIZATIONS**

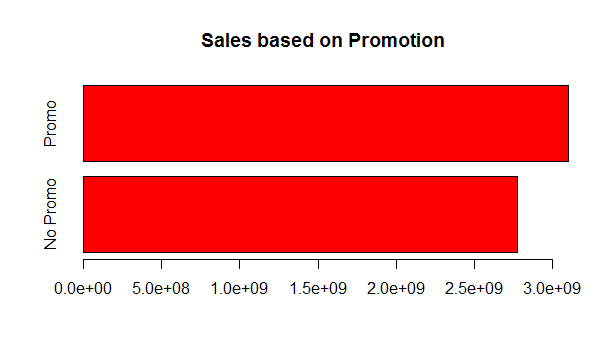
Following were the insights derived from the dataset.

1. **What is the distribution of the number of customers on a daily basis?**



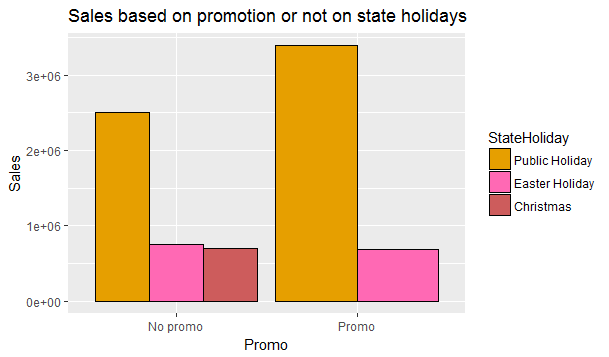
This is a histogram that depicts the daily inflow of customers at the Rossmann stores on a daily basis. It is evident from the graph that the store mostly gets around 500 – 1000 customers in a day. The second most common distribution of the number of customers is between 0 and 500. Furthermore, it is very rare that the number of customers crosses more than 2000 in a single day.

1. **What is the difference in sales between a promo and a non-promo day at the store?**



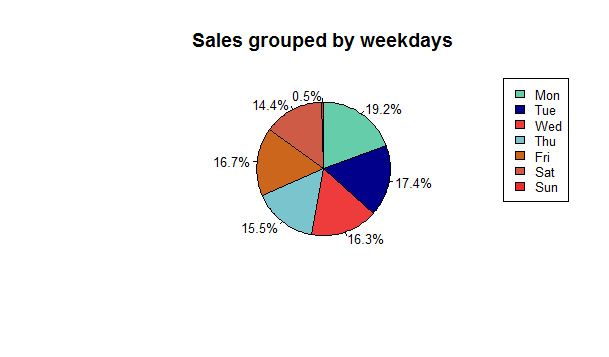
This vertical bar chart shows us the sales rate based on promotion. We could see that the sales were higher when there were promotions as compared to a non-promo day at the Rossmann stores. After calculations, we found that the difference between sales with and without promotion was nearly 329,231,949 € per day although it is not obvious from the chart. Using this figure, we could conclude that the sales were significantly higher when products were promoted.

1. **What is the effect of promotions on store sales on a state holiday?**



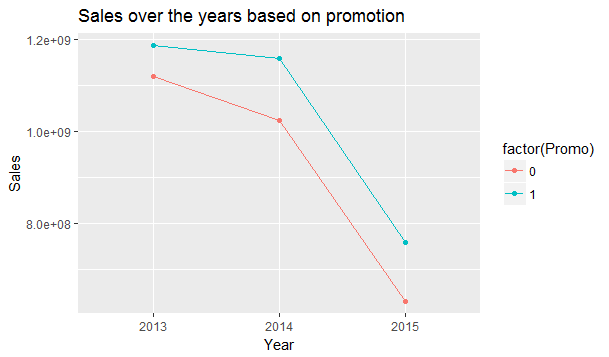
This horizontal bar chart shows the sales based on promotion and on state holidays. The sales on a non-state holiday has been excluded for this visualization. It is visible from the graph that sales were comparatively more when the products were promoted on a public holiday. However, there is not much difference between the sales on Easter holiday with or without sales which leads to the conclusion that promotions were not needed that day. Furthermore, there were no sales on Christmas even with promotions. This could be a point for further investigation as to why there are no sales on Christmas on the promoted products or is Rossmann just not having any promotions on Christmas.

1. **Which are the days of the week when the sales are maximum and minimum?**



This pie chart shows us the sales rate for each weekday. Sales were maximum on Mondays around 19.2% of the week’s total sales and that is probably because of the fact that people shop on Mondays for the rest of the week. The least sales were made on the weekend, especially on Sundays merely 0.5% of the total sales for the week. This could be because the stores are closed on weekends in most European countries.

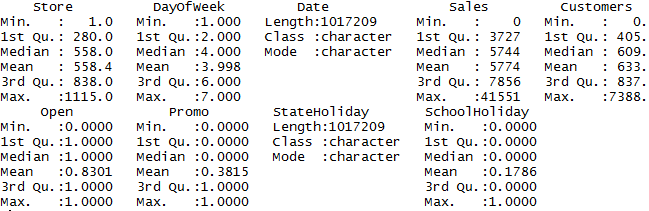
1. **What is the growth or decline in sales over the years based on promotions?**



The above line graph shows us the sales based on promotion over the years 2013, 2014, and 2015. 0 indicates sales with no promotion and 1 indicates sales with promotion. It is clear that the sales have always been higher with promotion every year. However, we see a huge drop in sales from 2014 to 2015 regardless of whether there was promotion or not. This is could be because the latest date for a sale in the data is the 31/7/2015. The dataset lacks data from August to December 2015 which makes the drop comprehensible.

**STATISTICAL AVERAGES OF THE DATA**

Below is the compilation of all the statistical averages for the dataset.



**REFERENCES**

In order to do maximum justice in understanding the impact of product promotions in store sales, below mentioned are few of the research papers that we went through during the course of carrying out this project: -

* Boddewyn, Jean J., and Monica Leardi. "Sales promotions: Practice, regulation and self-regulation around the world." International Journal of Advertising 8.4 (1989), p. 365.
* Chaharsoughi, Shahriar Ansari. "Effect of sales promotion on consumer behavior based on culture." African Journal Of Business Management 6.1 (2012), p. 1.

**PROJECT CODE**

**R script**

*data<-read.csv('train.csv', header = TRUE, sep=',', stringsAsFactors=FALSE)*

*View(data)*

#User defined function: Check for every column if there are missing values

*co\_li<-c(data['Store'], data['DayOfWeek'], data['Date'], data['Sales'], data['Customers'], data['Open'], data['Promo'], data['StateHoliday'], data['SchoolHoliday'] )*

*checkna<-function(col) {*

*i=1*

*for (co in col){*

*cat('Column: ', names(col[i]),', NA Values: ',sum(is.na(co)), '\n')*

*i=i+1*

*}*

*}*

*func<-checkna(co\_li)*

*summary(data)*

#Data cleaning 1: Split the column 'Date'

*install.packages("tidyr")*

*library(tidyr)*

*cl1\_df<-separate(data,Date, c("Year", "Month", "Day"), sep = "-")*

*View(cl1\_df)*

#How are the sales when store is open or not?

*just\_open<-subset(cl1\_df, Open == 1)*

*cat("Sum sales for open stores", sum(as.numeric(just\_open$Sales)))*

*just\_closed<-subset(cl1\_df, Open == 0)*

*cat("Sum sales for closed stores", sum(just\_closed$Sales))*

#Data cleaning 2: The sum of sales for closed stores is zero, so wie drop these rows (meanswe keep the just\_open dataframe as our dataframe to proceed)

#Data cleaning 3: Replace the string values in the column 'state holiday' by numeric valuesto calculate faster

*df<-just\_open*

*df[df$StateHoliday == "a",]$StateHoliday = 1*

*df[df$StateHoliday == "b",]$StateHoliday = 2*

*df[df$StateHoliday == "c",]$StateHoliday = 3*

*df[df$StateHoliday == "0",]$StateHoliday = 0*

*View(df)*

#VISUALIZATION

#1. What is the distribution of customers coming into the stores?

*hist(df$Customers, main='Distribution of Customers coming into stores on a daily basis', xlab='Customers', col='seagreen', xlim=c(0, 3500))*

#2. Sales based on Promotion

*install.packages('ggplot2')*

*library(ggplot2)*

*d <- aggregate(df$Sales, by=list(Promo=df$Promo), FUN = function(x){sum(as.numeric(x))})*

*barplot(d$x, main='Sales based on Promotion', names.arg = c('No Promo', 'Promo'), horiz = TRUE, col='red')*

#3. Sales based on promotion or not on state holidays

*st <- aggregate(df$Sales, by=list(Promo=df$Promo, StateHoliday=df$StateHoliday), FUN = function(x){sum(as.numeric(x))})*

*st\_cl <- subset(st, StateHoliday != 0)*

*st\_cl[st\_cl$Promo == 0,]$Promo = "No promo"*

*st\_cl[st\_cl$Promo == 1,]$Promo = "Promo"*

*ggplot(data=st\_cl, aes(x=Promo, y=x, fill=StateHoliday)) +*

*geom\_bar(stat="identity", position=position\_dodge(), colour="black") +*

*labs(title="Sales based on promotion or not on state holidays", y="Sales") +*

*scale\_fill\_manual(labels = c("Public Holiday", "Easter Holiday", "Christmas"), values=c("#E69F00", "#FF69B4", "#CD5C5C"))*

#4. Sales based on weekdays

*we <- aggregate(df$Sales, by=list(DayOfWeek=df$DayOfWeek), FUN = function(y){sum(as.numeric(y))})*

*cols <- c("aquamarine3","blue4","brown2","cadetblue3","chocolate3","coral3","firebrick2")*

*percentlabels<- round(100\*we$x/sum(we$x), 1)*

*pielabels<- paste(percentlabels, "%", sep="")*

*pie(we$x, main="Sales grouped by weekdays", col=cols, labels=pielabels, cex=0.8, clockwise=TRUE)*

*legend("topright", c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"), cex=0.8, fill=cols)*

#5. Sales over the years based on promotion

*pr <-aggregate(Sales~Year+Promo, df, FUN = function(y){sum(as.numeric(y))})*

*ggplot(data=pr, aes(x=Year, y=Sales, colour=factor(Promo), group=factor(Promo)))+*

*labs(title="Sales over the years based on promotion") +*

*geom\_line() +*

*geom\_point()*

**User defined R function (To check columns for missing values)**

*co\_li<-c(data['Store'], data['DayOfWeek'], data['Date'], data['Sales'], data['Customers'], data['Open'], data['Promo'], data['StateHoliday'], data['SchoolHoliday'] )*

*checkna<-function(col) {*

*i=1*

*for (co in col){*

*cat('Column: ', names(col[i]),', NA Values: ',sum(is.na(co)), '\n')*

*i=i+1*

*}*

*}*

*func<-checkna(co\_li)*

**R Code for the Visualizations**

#VISUALIZATION

#1. What is the distribution of customers coming into the stores?

*hist(df$Customers, main='Distribution of Customers coming into stores on a daily basis', xlab='Customers', col='seagreen', xlim=c(0, 3500))*

*#2. Sales based on Promotion*

*install.packages('ggplot2')*

*library(ggplot2)*

*d <- aggregate(df$Sales, by=list(Promo=df$Promo), FUN = function(x){sum(as.numeric(x))})*

*barplot(d$x, main='Sales based on Promotion', names.arg = c('No Promo', 'Promo'), horiz = TRUE, col='red')*

#3. Sales based on promotion or not on state holidays

*st <- aggregate(df$Sales, by=list(Promo=df$Promo, StateHoliday=df$StateHoliday), FUN = function(x){sum(as.numeric(x))})*

*st\_cl <- subset(st, StateHoliday != 0)*

*st\_cl[st\_cl$Promo == 0,]$Promo = "No promo"*

*st\_cl[st\_cl$Promo == 1,]$Promo = "Promo"*

*ggplot(data=st\_cl, aes(x=Promo, y=x, fill=StateHoliday)) +*

*geom\_bar(stat="identity", position=position\_dodge(), colour="black") +*

*labs(title="Sales based on promotion or not on state holidays", y="Sales") +*

*scale\_fill\_manual(labels = c("Public Holiday", "Easter Holiday", "Christmas"), values=c("#E69F00", "#FF69B4", "#CD5C5C"))*

#4. Sales based on weekdays

*we <- aggregate(df$Sales, by=list(DayOfWeek=df$DayOfWeek), FUN = function(y){sum(as.numeric(y))})*

*cols <- c("aquamarine3","blue4","brown2","cadetblue3","chocolate3","coral3","firebrick2")*

*percentlabels<- round(100\*we$x/sum(we$x), 1)*

*pielabels<- paste(percentlabels, "%", sep="")*

*pie(we$x, main="Sales grouped by weekdays", col=cols, labels=pielabels, cex=0.8, clockwise=TRUE)*

*legend("topright", c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"), cex=0.8, fill=cols)*

#5. Sales over the years based on promotion

*pr <-aggregate(Sales~Year+Promo, df, FUN = function(y){sum(as.numeric(y))})*

*ggplot(data=pr, aes(x=Year, y=Sales, colour=factor(Promo), group=factor(Promo)))+*

*labs(title="Sales over the years based on promotion") +*

*geom\_line() +*

*geom\_point()*