MVA project

Lokesh Arora, Ankita Shinde

10/1/2020

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.6.3
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library("plyr")
```

```
## Warning: package 'plyr' was built under R version 3.6.3
```

```
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
library(RColorBrewer)
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dataset = read.csv("data.csv", header= T)
head(dataset)
```

```
##   Store Dept       Date weeklySales isHoliday Type   Size Temperature
## 1     1    1 2010-02-05    24924.50     False    A 151315       42.31
## 2     1    1 2010-02-12    46039.49      True    A 151315       38.51
## 3     1    1 2010-02-19    41595.55     False    A 151315       39.93
## 4     1    1 2010-02-26    19403.54     False    A 151315       46.63
## 5     1    1 2010-03-05    21827.90     False    A 151315       46.50
## 6     1    1 2010-03-12    21043.39     False    A 151315       57.79
##   Fuel_Price MarkDown1 MarkDown2 MarkDown3 MarkDown4 MarkDown5      CPI
## 1      2.572        NA        NA        NA        NA        NA 211.0964
## 2      2.548        NA        NA        NA        NA        NA 211.2422
## 3      2.514        NA        NA        NA        NA        NA 211.2891
## 4      2.561        NA        NA        NA        NA        NA 211.3196
## 5      2.625        NA        NA        NA        NA        NA 211.3501
## 6      2.667        NA        NA        NA        NA        NA 211.3806
##   Unemployment
## 1        8.106
## 2        8.106
## 3        8.106
## 4        8.106
## 5        8.106
## 6        8.106
```

We can see that there are few null values in the data set for column Markdown 1 - 5. We will also split the data column in 3 as Day, Month and Year.

```
dataset$Year <- year(ymd(dataset$Date))
dataset$Month <- month(ymd(dataset$Date))
dataset$Day <- day(ymd(dataset$Date))
dataset$Dept = as.factor(dataset$Dept)
dataset$Store = as.factor(dataset$Store)
dataset$MarkDown1[is.na(dataset$MarkDown1)] = 0
dataset$MarkDown2[is.na(dataset$MarkDown2)] = 0
dataset$MarkDown3[is.na(dataset$MarkDown3)] = 0
dataset$MarkDown4[is.na(dataset$MarkDown4)] = 0
dataset$MarkDown5[is.na(dataset$MarkDown5)] = 0
head(dataset)
```

```
##   Store Dept       Date weeklySales isHoliday Type   Size Temperature
## 1     1    1 2010-02-05    24924.50     False    A 151315       42.31
## 2     1    1 2010-02-12    46039.49      True    A 151315       38.51
## 3     1    1 2010-02-19    41595.55     False    A 151315       39.93
## 4     1    1 2010-02-26    19403.54     False    A 151315       46.63
## 5     1    1 2010-03-05    21827.90     False    A 151315       46.50
## 6     1    1 2010-03-12    21043.39     False    A 151315       57.79
##   Fuel_Price MarkDown1 MarkDown2 MarkDown3 MarkDown4 MarkDown5      CPI
## 1      2.572         0         0         0         0         0 211.0964
## 2      2.548         0         0         0         0         0 211.2422
## 3      2.514         0         0         0         0         0 211.2891
## 4      2.561         0         0         0         0         0 211.3196
```

```
## 5      2.625         0         0         0         0 211.3501
## 6      2.667         0         0         0         0 211.3806
##   Unemployment Year Month Day
## 1        8.106 2010     2   5
## 2        8.106 2010     2  12
## 3        8.106 2010     2  19
## 4        8.106 2010     2  26
## 5        8.106 2010     3   5
## 6        8.106 2010     3  12
```

```
dim(dataset)
```

```
## [1] 421570     19
```

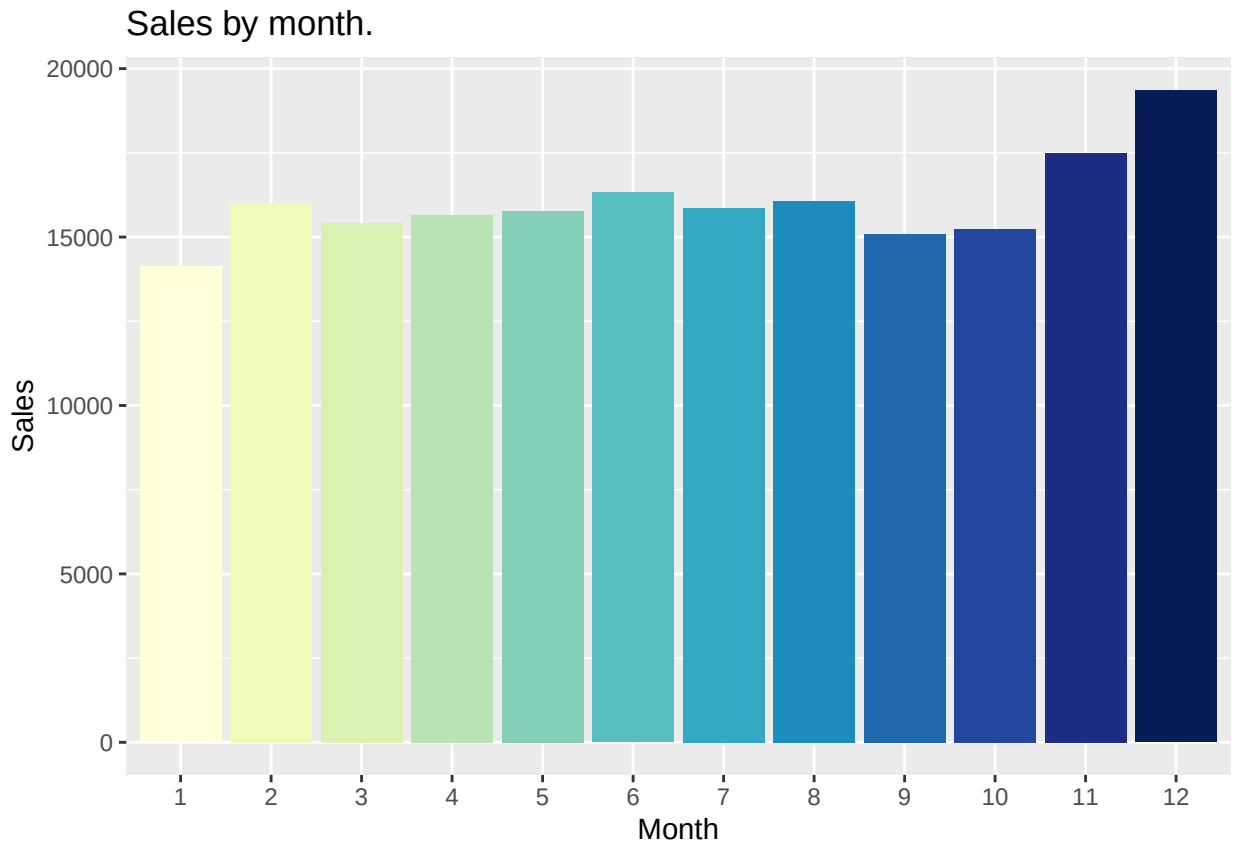
As we can see from the result now there are 19 columns and 421570 rows.

```
names(dataset)
```

```
##  [1] "Store"        "Dept"         "Date"         "weeklySales"  "isHoliday"
##  [6] "Type"         "Size"         "Temperature"  "Fuel_Price"   "MarkDown1"
## [11] "MarkDown2"    "MarkDown3"    "MarkDown4"    "MarkDown5"    "CPI"
## [16] "Unemployment" "Year"         "Month"        "Day"
```

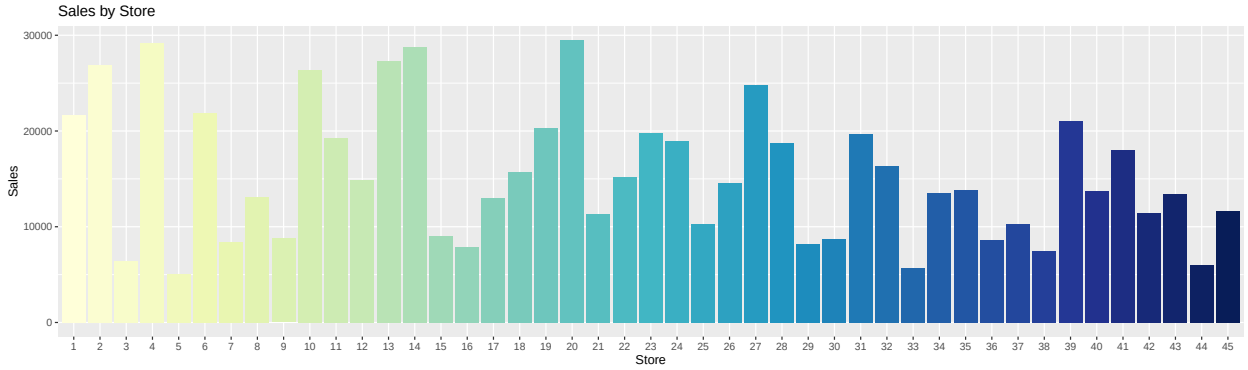
Here are all the columns.

```
month_wise =ddply(dataset, .(Month), summarize,  Sales=mean(weeklySales))
month_wise$Month = as.factor(month_wise$Month)
ggplot(month_wise,  aes(fill=Month, y=Sales, x=Month)) +
    geom_bar(position="dodge", stat="identity") + ggtitle("Sales by month.") +
  scale_fill_manual(values=colorRampPalette(brewer.pal(9, "YlGnBu"))(12)) +
  theme(legend.position = "none")
```

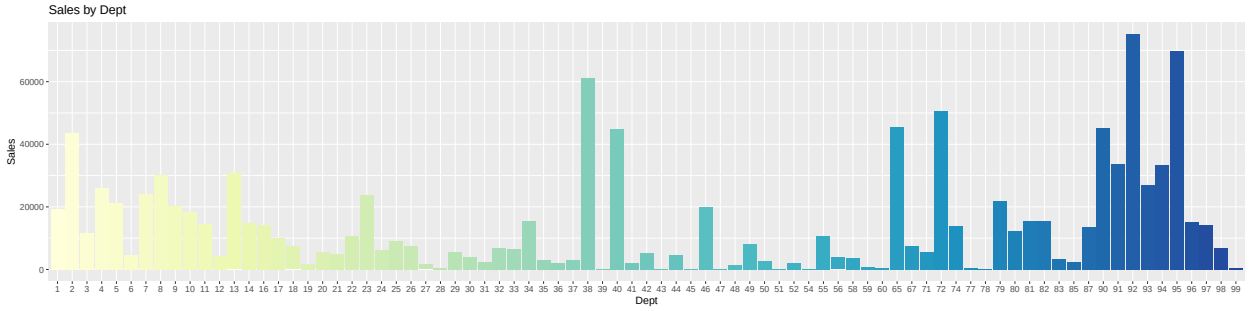


The above graph shows the average sales by walmart every month. We can see that the Sales is always most during December.

```
store_wise =ddply(dataset, .(Store), summarize,  Sales=mean(weeklySales))
ggplot(store_wise,  aes(fill=Store, y=Sales, x=Store)) +
  geom_bar(position="dodge", stat="identity") + ggtitle("Sales by Store") +
  scale_fill_manual(values=colorRampPalette(brewer.pal(9, "YlGnBu"))(45)) +
  theme(legend.position = "none")
```



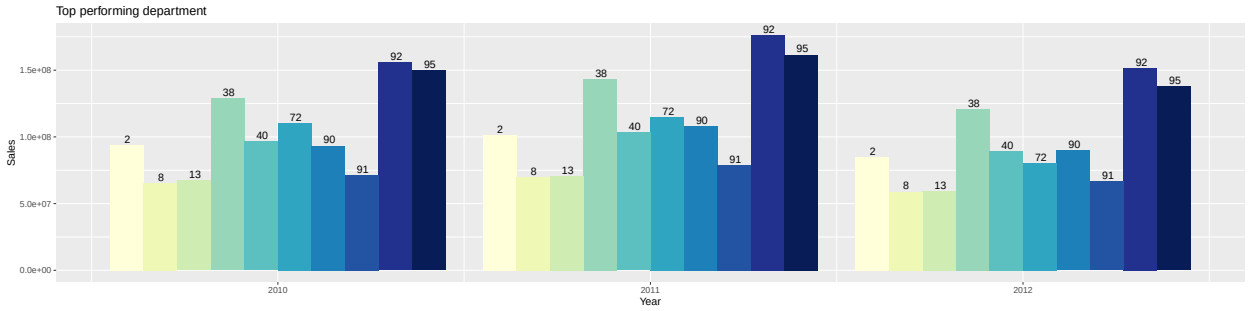

```
dept_wise =ddply(dataset, .(Dept), summarize,  Sales=mean(weeklySales))
ggplot(dept_wise,  aes(fill=Dept, y=Sales, x=Dept)) +
  geom_bar(position="dodge", stat="identity") + ggtitle("Sales by Dept") +
  theme(legend.position = "none") +
  scale_fill_manual(values=colorRampPalette(brewer.pal(9, "YlGnBu"))(99))
```




```
dept_sales =ddply(dataset, c("Year","Dept"), summarize,  Sales=sum(weeklySales))
top_depts = arrange(dept_sales,Year, desc(Sales)) %>% group_by(Year) %>% top_n(n = 10)
```

```
## Selecting by Sales
```

```
top_depts$Dept = as.factor(top_depts$Dept)
ggplot(data=top_depts, aes(x=Year, y=Sales, fill=Dept)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_text(aes(label=Dept), position = position_dodge2(width = 0.9, preserve = "single"), angle = 0, v
  scale_fill_manual(values=colorRampPalette(brewer.pal(9, "YlGnBu"))(10)) +
  ggtitle("Top performing department") + theme(legend.position = "none")
```

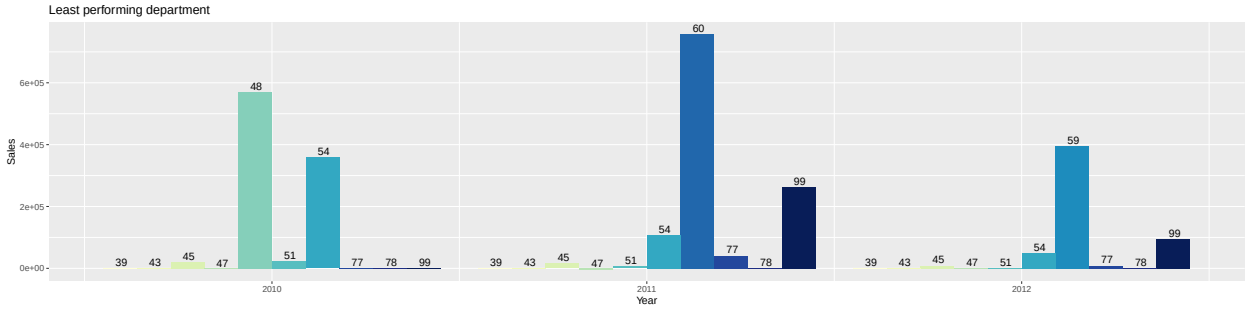


The above graph shows the depatment where the sales is maximum.

```
bottom_depts = arrange(dept_sales,Year, desc(Sales)) %>% group_by(Year) %>% top_n(n = -10)
```

```
## Selecting by Sales
```

```
bottom_depts$Dept = as.factor(bottom_depts$Dept)
ggplot(data=bottom_depts, aes(x=Year, y=Sales, fill=Dept)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_text(aes(label=Dept), position = position_dodge2(width = 0.9, preserve = "single"), angle = 0, v
  scale_fill_manual(values=colorRampPalette(brewer.pal(9, "YlGnBu"))(12)) +
  ggtitle("Least performing department") + theme(legend.position = "none")
```



The above graph shows the departments where sales is least. We can also see that the departmnets 47 in is loss.