Capestone Project

James Strayer

02/01/2020

Introduction

I was asked by a manager of Restaurant that does retail, to have a model to predict his sales for a given week, he gave me a file with the sales since end of March 2015. He started as the manager on 01-10-2018 ##Data import and cleaning

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --
```

```
## <U+2713> ggplot2 3.2.1     <U+2713> purrr   0.3.2
## <U+2713> tibble  2.1.3     <U+2713> dplyr   0.8.3
## <U+2713> tidyr   0.8.3     <U+2713> stringr 1.4.0
## <U+2713> readr   1.3.1     <U+2713> forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Parsed with column specification:
## cols(
##   Date = col_date(format = ""),
##   Weekday = col_character(),
##   Year = col_double(),
##   Month = col_character(),
##   Day = col_double(),
##   `annee fiscale` = col_character(),
##   Sales = col_number(),
##   Retail = col_number(),
##   TakeOutSales = col_number(),
##   Bar_Sales = col_number(),
##   Sales_Restaurant = col_number()
## )
```

```
## Observations: 1,715
## Variables: 11
## $ Date             <date> 2015-03-29, 2015-03-30, 2015-03-31, 2015-04-01, 201...
## $ Weekday          <chr> "SUNDAY", "MONDAY", "TUESDAY", "WEDNESDAY", "THURSDA...
## $ Year             <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015...
## $ Month            <chr> "March'15", "March'15", "March'15", "April'15", "Apr...
## $ Day              <dbl> 29, 30, 31, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1...
## $ `annee fiscale`  <chr> "2014-15", "2014-15", "2014-15", "2014-15", "2014-15...
## $ Sales            <dbl> 1792.00, NA, 1526.30, 2250.26, 2077.57, 2357.48, 150...
## $ Retail           <dbl> 373.00, NA, 380.95, 363.72, 268.80, 394.23, 476.17, ...
## $ TakeOutSales     <dbl> 0.00, 0.00, 99.25, 402.75, 0.00, 32.50, 93.73, 0.00,...
## $ Bar_Sales        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 227....
## $ Sales_Restaurant <dbl> 1419.00, 0.00, 1046.10, 1483.79, 1808.77, 1930.75, 9...
```

The manager is using the data to have a day to day idea of the health of his buisness and therefore he added an accounting variable, annee fiscale (fiscal year) ,that we don't need and that we can remove.The Month variable is in a format not pratical for analysis (Month, YY) we want to have just the Full moth written without the year. So for simplicity and because we have the full date in the Date column, we are going to remove the month column as it is and create a new one based on the date column, we will do the same for the weekdays column. Furthermore we don't need the day number in our analysis so we will remove it too:

```
#Remove unncessary column for the analysis
DataForAnalysis<-DailySales%>%select(-Weekday,-Month,-Day,-`annee fiscale`)
#Add a column for weekday and month
DataForAnalysis<-DataForAnalysis%>%mutate(Weekday=weekdays(Date),Month=months(Date))
```

If we look again at the data we see that we have a Sales column, representing the Total Sales for the day and then each column after it, is the total sales for each day for each component of the restaurant possible sales revenu, so Retail, Take-out, Bar and Restaurant. We can see that there is NAs in all of those data

```
#looking for NAs in the sales data
sum(is.na(DataForAnalysis$Sales))
```

```
## [1] 175
```

```
sum(is.na(DataForAnalysis$Bar_Sales))
```

```
## [1] 208
```

```
sum(is.na(DataForAnalysis$Retail))
```

```
## [1] 188
```

```
sum(is.na(DataForAnalysis$TakeOutSales))
```

```
## [1] 1
```

```
sum(is.na(DataForAnalysis$Sales_Restaurant))
```

```
## [1] 1
```

we will change those NAs to 0, considering a $0 CAD sales for that day and variable

```
#Chaning NAs to 0 in the sales data
DataForAnalysis[is.na(DataForAnalysis$Sales),]$Sales<-0
DataForAnalysis[is.na(DataForAnalysis$Bar_Sales),]$Bar_Sales<-0
DataForAnalysis[is.na(DataForAnalysis$Retail),]$Retail<-0
DataForAnalysis[is.na(DataForAnalysis$TakeOutSales),]$TakeOutSales<-0
DataForAnalysis[is.na(DataForAnalysis$Sales_Restaurant),]$Sales_Restaurant<-0
```

Once we Cleaned the data. We are interested to add some classification for the days, specialy considering that holidays and special event should have an impact on the sales of a restaurant, to test this hypothesis we create a data frame event for all the bank holidays and events in the province of Quebec during the year:

here is an example of the data used to create that data frame: https://www.statutoryholidays.com/2017.php, all dates with observance National, QC and event such as Mother's day and Valentine's day. Once this vector is created we can create a new variable called EventDay which is true if the date equals one of the date in the vector

```
#Add a column for EventDay
DataForAnalysis<-DataForAnalysis%>%mutate(EventDay=ifelse(Date %in% Event,TRUE,FALSE))
```

Data exploration

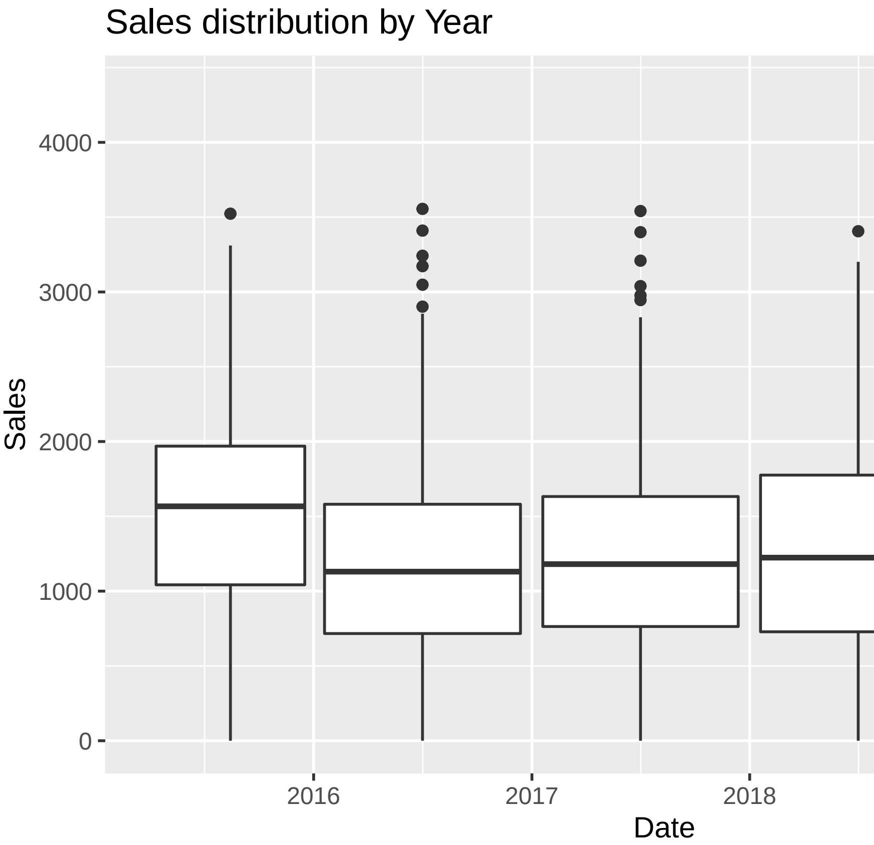
```
#look at the structure of the data
glimpse(DataForAnalysis)
```

```
## Observations: 1,715
## Variables: 10
## $ Date             <date> 2015-03-29, 2015-03-30, 2015-03-31, 2015-04-01, 201...
## $ Year             <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015...
## $ Sales            <dbl> 1792.00, 0.00, 1526.30, 2250.26, 2077.57, 2357.48, 1...
## $ Retail           <dbl> 373.00, 0.00, 380.95, 363.72, 268.80, 394.23, 476.17...
## $ TakeOutSales     <dbl> 0.00, 0.00, 99.25, 402.75, 0.00, 32.50, 93.73, 0.00,...
## $ Bar_Sales        <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00...
## $ Sales_Restaurant <dbl> 1419.00, 0.00, 1046.10, 1483.79, 1808.77, 1930.75, 9...
## $ Weekday          <chr> "Sunday", "Monday", "Tuesday", "Wednesday", "Thursda...
## $ Month            <chr> "March", "March", "March", "April", "April", "April"...
## $ EventDay         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
```

```
summary(DataForAnalysis)
```

```
##       Date                 Year          Sales            Retail
##  Min.   :2015-03-29   Min.   :2015   Min.   :   0.0   Min.   :   0.0
##  1st Qu.:2016-05-30   1st Qu.:2016   1st Qu.: 779.9   1st Qu.: 168.2
##  Median :2017-08-02   Median :2017   Median :1266.8   Median : 263.2
##  Mean   :2017-08-02   Mean   :2017   Mean   :1275.9   Mean   : 271.1
##  3rd Qu.:2018-10-04   3rd Qu.:2018   3rd Qu.:1764.8   3rd Qu.: 372.3
##  Max.   :2019-12-29   Max.   :2019   Max.   :4360.8   Max.   :1436.9
##   TakeOutSales       Bar_Sales      Sales_Restaurant    Weekday
##  Min.   :   0.00   Min.   :   0.0   Min.   :-2299.8   Length:1715
##  1st Qu.:   3.00   1st Qu.:  72.5   1st Qu.:  377.6   Class :character
##  Median :  65.49   Median : 160.8   Median :  709.8   Mode  :character
##  Mean   :  86.86   Mean   : 181.9   Mean   :  736.0
##  3rd Qu.: 125.37   3rd Qu.: 265.0   3rd Qu.: 1037.7
##  Max.   :2965.72   Max.   :1045.0   Max.   : 4284.2
##     Month             EventDay
##  Length:1715        Mode :logical
##  Class :character   FALSE:1715
##  Mode  :character
##
##
##
```

In our data set we have 10 variables for 1715 observations corresponding to the number of days the restaurant



was open since the 29-03-2015. ## Sales variable
We see In the box plot above that 2019 was similar to 2018 in term of sales but with 2 outliers ##Preparing Data for Modeling

Modeling Data

Discussion