Introduction :

Sales are the most important variable in a company income its analysis is very important for the company expansion and future strategy. The variable sales is affected by many external and internal variables the relationship between these features can determine the project's performance .
The data is collected from 45 stores in 99 departments weekly between the year 2010 and 2012 with all the different internal data that may affect it such as markdowns (promotions) ,store size and type and external data like unemployment , CPI, fuel price , holiday and temperature .
Our main purpose , is to explain the variable Sales based on the given features, and to find the determinants of performance of stores?

I. Data explanation :

The data set gives details about features that affects the sales variable . Each of the features has its specifications that we are going to be describing in details in the coming sections.

- Store: the ID of the store, contains the number of each store going from one to 45 .
- Store size : the size of the store, contains the number of products in store ranging from 34875-219622 .
- Store type : the type of the store, three types are observed A,B and C.
- Date : the date of the weekly purchase
- Dept : the department where the store is located .
- Holiday : contains information about the week of the purchase is a holiday or not
- Temperature: the average temperature in the region for the corresponding week.
- Fuel price : the price of fuel in the week of purchase in the region .
- CPI : the consumer price index of the corresponding region and week.
- Unemployment: the unemployment rate in the region for the corresponding week

II. Data cleaning and preprocessing :

After importing the data I had a hard time processing some of its features so I had to make many conversions since all the features came in factor type and many other processes were needed in order to get a data set I can work on with R .

1. Data overview :

I started my analysis by trying to get a general idea about this data set and its different components so I used different R functions :

```
       V1               V2                 V3               V4               V5               V6
 Min.   :     1   13     : 10474   23/12/2011:  3027   FALSE  :391909   A    :215478   39690  : 20802
 1st Qu.:105393   10     : 10315   25/11/2011:  3021   Holiday:     1   B    :163495   39910  : 20597
 Median :210786   4      : 10272   16/12/2011:  3013   TRUE   : 29661   C    : 42597   203819 : 20376
 Mean   :210786   1      : 10244   09/12/2011:  3010                    Type:     1   219622 : 10474
 3rd Qu.:316178   2      : 10238   17/02/2012:  3007                                   126512 : 10315
 Max.   :421570   24     : 10228   30/12/2011:  3003                                   205863 : 10272
 NA's   :1        (Other):359800   (Other)   :403490                                   (Other):328735
       V7               V8               V9              V10              V11              V12
 1      :  6435   10     :   353   50.43  :   709   3.638  :  2548   1.5    :   102   1.91   :   539
 10     :  6435   5      :   289   67.87  :   646   3.63   :  2164   460.73 :   102   3      :   493
 13     :  6435   20     :   232   72.62  :   594   2.771  :  1917   175.64 :    93   0.5    :   485
 14     :  6435   15     :   215   76.67  :   583   3.891  :  1856   1282.42:    75   1.5    :   471
 16     :  6435   12     :   175   70.28  :   563   3.594  :  1796   1483.17:    75   4      :   367
 2      :  6435   1      :   169   76.03  :   555   3.524  :  1793   (Other):150235   (Other):108894
 (Other):382961   (Other):420138   (Other):417921   (Other):409497   NA's   :270889   NA's   :310322
       V13              V14              V15              V16              V17
 3      :   754   9      :   280   2743.18 :   136   130.7196:   833   8.099  :  5152
 6      :   710   4      :   200   1064.56 :   120   130.683 :   706   8.163  :  3636
 2      :   660   2      :   197   20371.02:    75   129.8364:   704   7.852  :  3614
 1      :   611   3      :   146   3557.67 :    75   130.244 :   703   7.343  :  3416
 0.22   :   487   47     :   143   3567.03 :    75   130.7929:   703   7.057  :  3414
 (Other):133870   (Other):134002   (Other) :150952   131.043 :   703   7.931  :  3400
 NA's   :284479   NA's   :286603   NA's    :270138   (Other) :417219   (Other):398939
[1] 421571     17
```

Head

	V1 <int>	V2 <fctr>	V3 <fctr>	V4 <fctr>	V5 <fctr>	V6 <fctr>	V7 <fctr>	V8 <fctr>	V9 <fctr>	V10 <fctr>	V11 <fctr>	V12 <fctr>	V13 <fctr>	V14 <fctr>
1	*NA*	Store	Date	Holiday	Type	Size	Dept	Sales	Temperature	Fuel_Price	Promotion1	Promotion2	Promotion3	Promotion4
2	1	1	01/04/2011	FALSE	A	151315	49	13167.85	59.17	3.524	*NA*	*NA*	*NA*	*NA*
3	2	1	01/04/2011	FALSE	A	151315	26	5946.53	59.17	3.524	*NA*	*NA*	*NA*	*NA*
4	3	1	01/04/2011	FALSE	A	151315	81	28545.23	59.17	3.524	*NA*	*NA*	*NA*	*NA*
5	4	1	01/04/2011	FALSE	A	151315	34	9949.54	59.17	3.524	*NA*	*NA*	*NA*	*NA*
6	5	1	01/04/2011	FALSE	A	151315	59	316.86	59.17	3.524	*NA*	*NA*	*NA*	*NA*

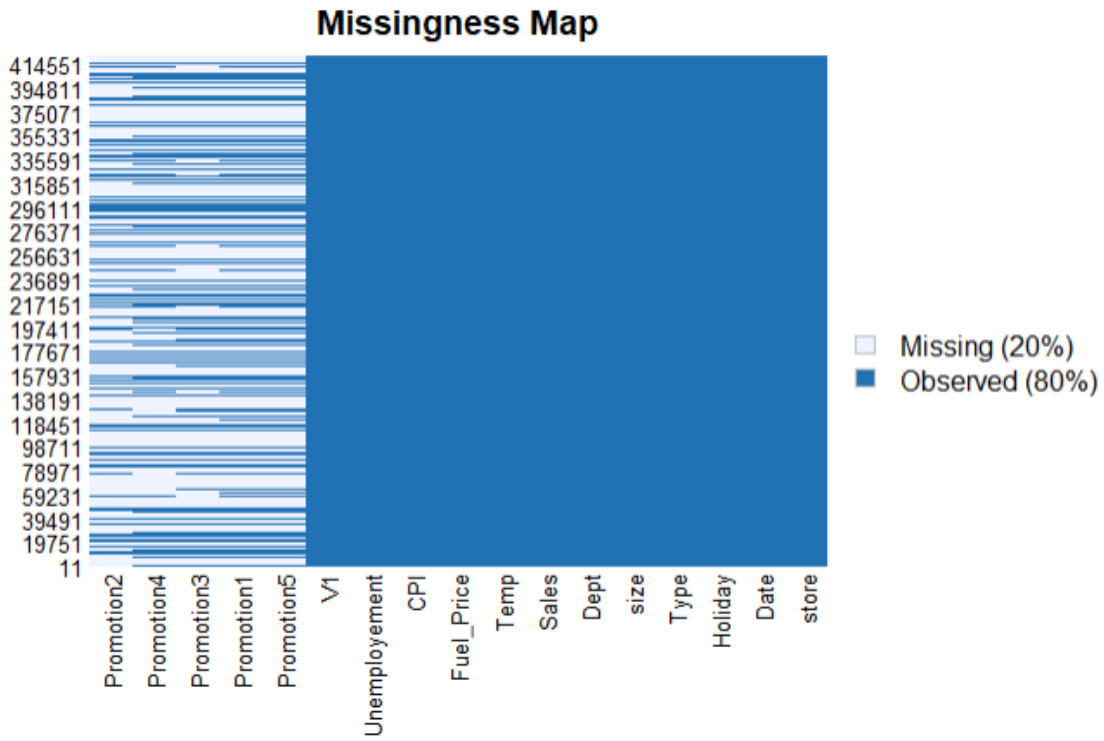
6 rows | 1-15 of 17 columns

Tail

	V1 <int>	V2 <fctr>	V3 <fctr>	V4 <fctr>	V5 <fctr>	V6 <fctr>	V7 <fctr>	V8 <fctr>	V9 <fctr>	V10 <fctr>	V11 <fctr>	V12 <fctr>	V13 <fctr>	V14 <fctr>	V15 <fctr>	V16 <fctr>	V17 <fctr>
421566	421565	9	31/12/2010	TRUE	B	125833	34	3415	45.92	2.943	*NA*	*NA*	*NA*	*NA*	*NA*	214.9268	6.56
421567	421566	9	31/12/2010	TRUE	B	125833	74	9559.24	45.92	2.943	*NA*	*NA*	*NA*	*NA*	*NA*	214.9268	6.56
421568	421567	9	31/12/2010	TRUE	B	125833	31	883.7	45.92	2.943	*NA*	*NA*	*NA*	*NA*	*NA*	214.9268	6.56
421569	421568	9	31/12/2010	TRUE	B	125833	24	762.06	45.92	2.943	*NA*	*NA*	*NA*	*NA*	*NA*	214.9268	6.56
421570	421569	9	31/12/2010	TRUE	B	125833	23	7147.8	45.92	2.943	*NA*	*NA*	*NA*	*NA*	*NA*	214.9268	6.56
421571	421570	9	31/12/2010	TRUE	B	125833	10	8987.31	45.92	2.943	*NA*	*NA*	*NA*	*NA*	*NA*	214.9268	6.56

6 rows

I also wanted to check if the data was complete or I had some missing data for that I used the Amelia library and the missmap function



The missmap function output shows a missing data percentage of 20 percent only present in the promotion variables and a complete data for the other variables . I estimate that we are not really in the case of a missing data but promotion don't take place at the same time neither they are always present so many of them came with the NA as a default value because no entry was affected to their value.

2.Cleaning Data :

we noticed that the data we imported came with features that all had the same type which was factor as observed in the previous analysis (head, tail, summary) that we after verified using the function type of .

so we had to convert all the features to the right type, most of them had to be converted to integers using the as.numeric function applied on the as.character function .

we also had to replace the NA values with null values so we can do different plots

As for the date, we use the library lubridate to convert it to a date type that can be understoodby the R software.we also had to rename all the columns with their right name.

We also had to delete the first line and row cause they had no added value to the data set .

```r
df$store<-as.numeric(as.character(df$store))
df$size<-as.numeric(as.character(df$size))
df$Temp<-as.numeric(as.character(df$Temp))
df$Sales<-as.numeric(as.character(df$Sales))
df$Fuel_Price<-as.numeric(as.character(df$Fuel_Price))
df$Promotion1<-as.numeric(as.character(df$Promotion1))
df$Promotion2<-as.numeric(as.character(df$Promotion2))
df$Promotion3<-as.numeric(as.character(df$Promotion3))
df$Promotion4<-as.numeric(as.character(df$Promotion4))
df$Promotion5<-as.numeric(as.character(df$Promotion5))
head(df)
```

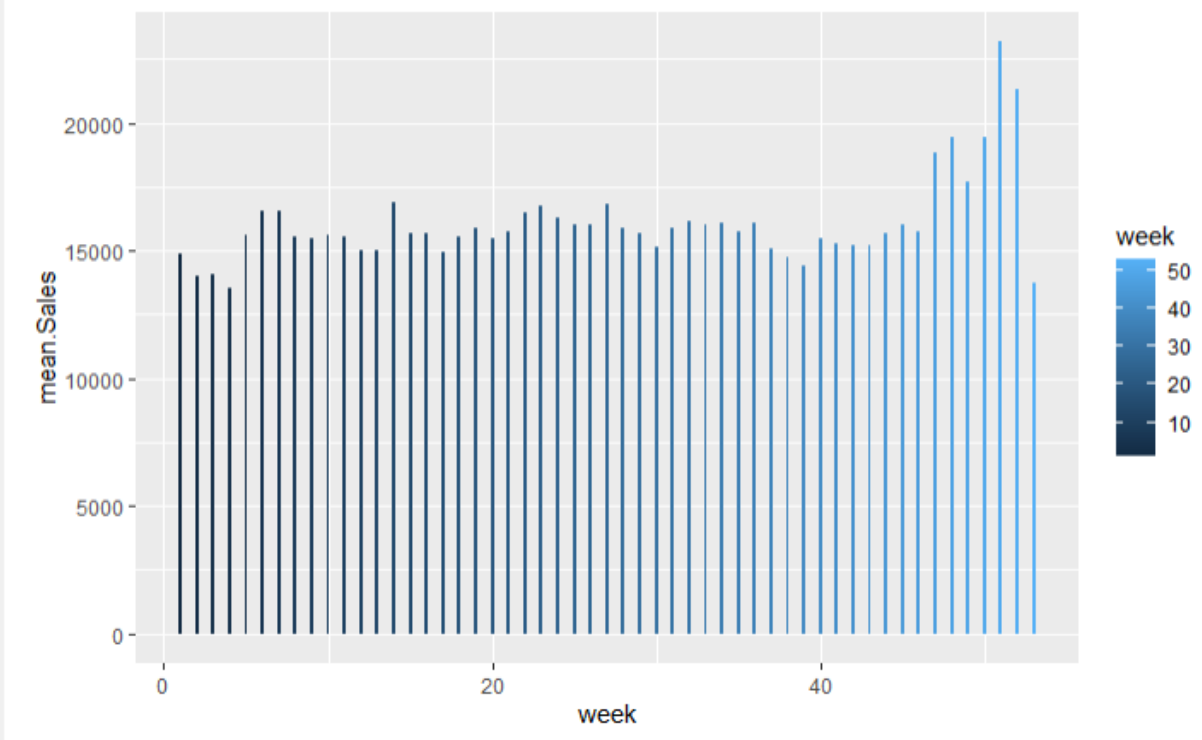
```r
names(df)[2] <- "store"
names(df)[3] <- "Date"
names(df)[4] <- "Holiday"
names(df)[5] <- "Type"
names(df)[6] <- "size"
names(df)[7] <- "Dept"
names(df)[8] <- "Sales"
names(df)[9] <- "Temp"
names(df)[10] <- "Fuel_Price"
names(df)[11] <- "Promotion1"
names(df)[12] <- "Promotion2"
names(df)[13] <- "Promotion3"
names(df)[14] <- "Promotion4"
names(df)[15] <- "Promotion5"
names(df)[16] <- "CPI"
names(df)[17] <- "Unemployement"
head(df)
```

Data visualization :

I. variation with time :

1. weeks :

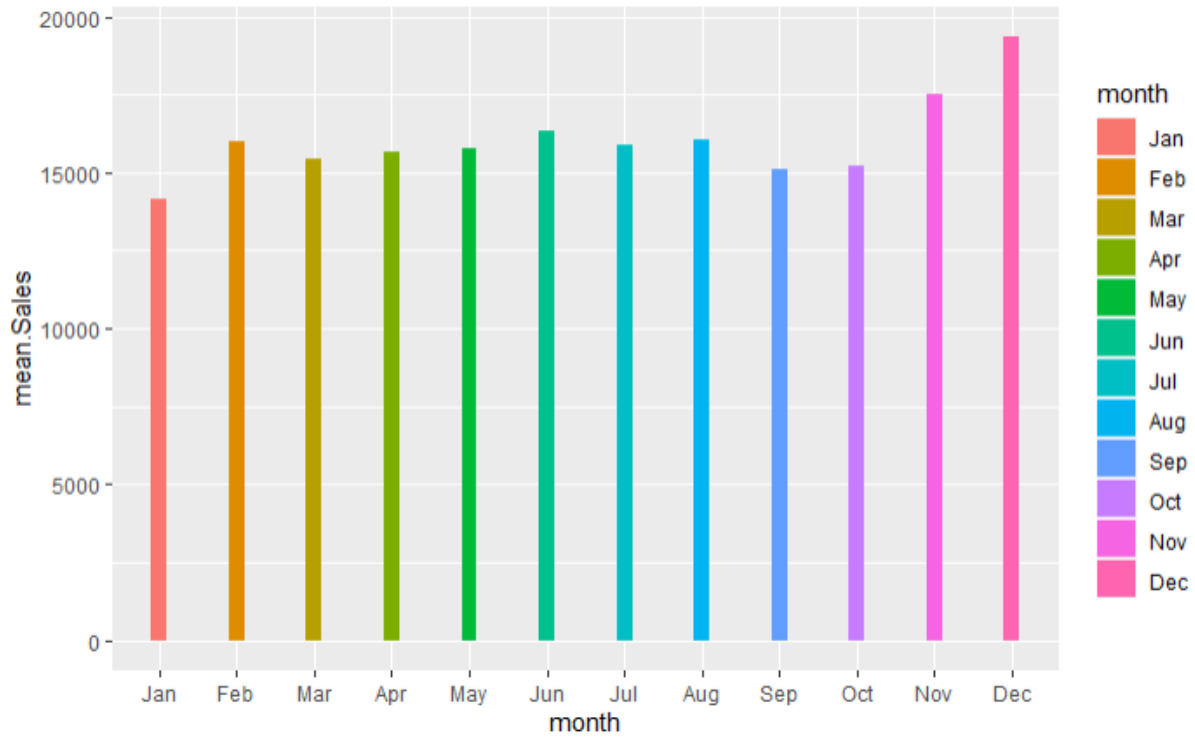
Since data is collected by weeks we wanted to see what effect the week variable had on the sales factors so we ploted the mean of sales by week for that we had to add the column date to the date sate using the mutate function and the week function from the lubridate library. We remarked that the sales are quite equal all along the year but tend to rise in the last six weeks of the year reaching a maximum by the 50^{th} year and in the minimum in the last week of the year this can be explained by the end of the year holidays such as thanks giving.



2. Month :

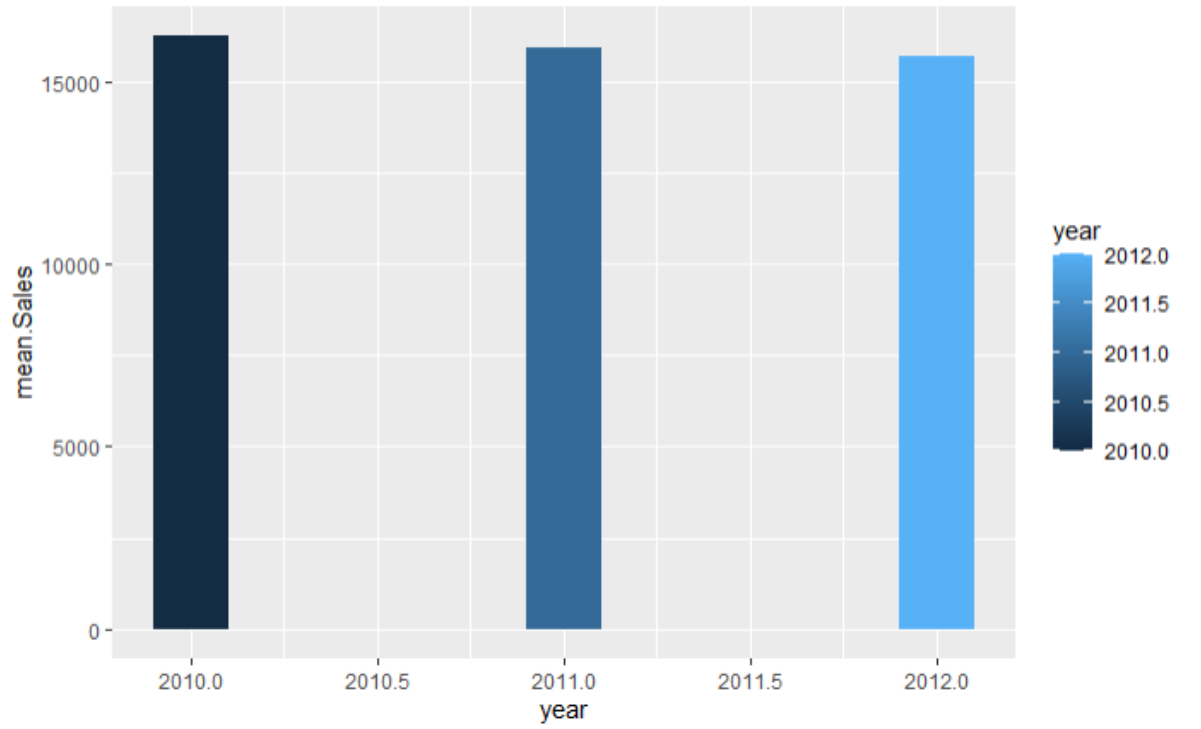
As seen in the week section , we wanted to check if the sales are affected by the month variable so we added a month column using the mutate and lubridate month function.

We notice that the sales do vary with the month with the lowest value on January and the highest sales on the month of December and November which confirms our deductions from the week analysis.



3. Year :

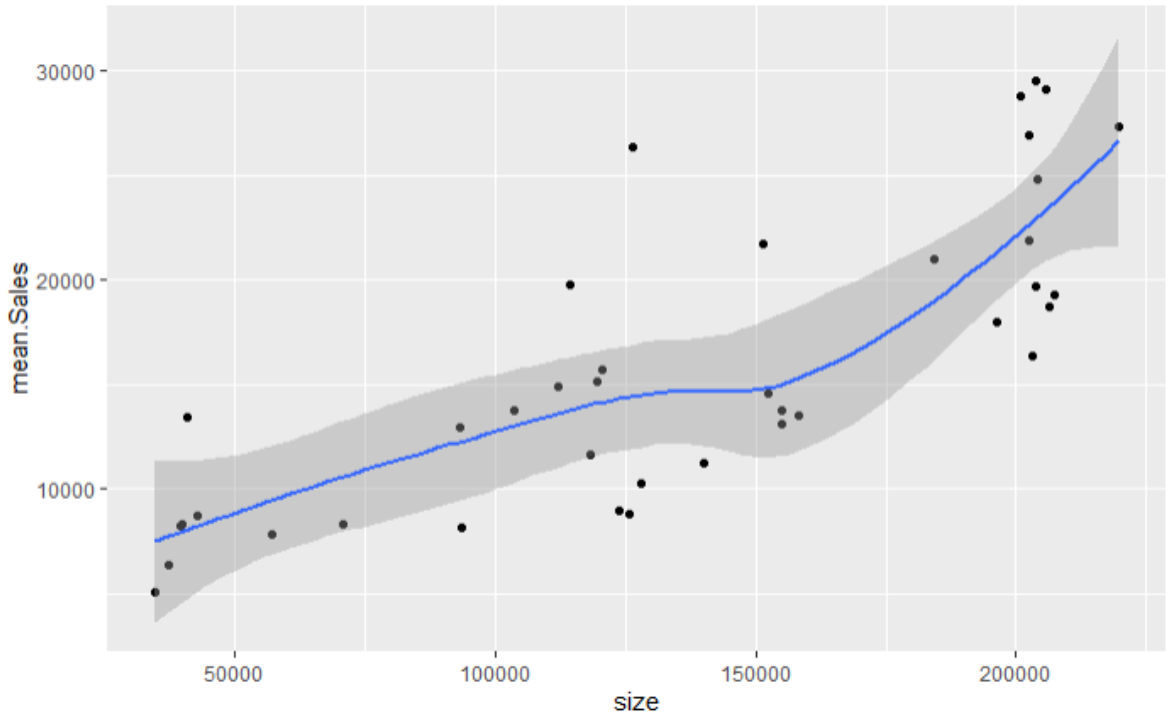
The year is not really shown in the data set but we thought that the sales may change depending on the year so we had to add a year column using the mutate and year function from the lubridate library . We can conclude that the sales are equal for the three years of the study but tend to be slightly lower every year with 2010 having the maximum sales and 2012 the lowest but the margin is not realy that important



II. Variation with the store :

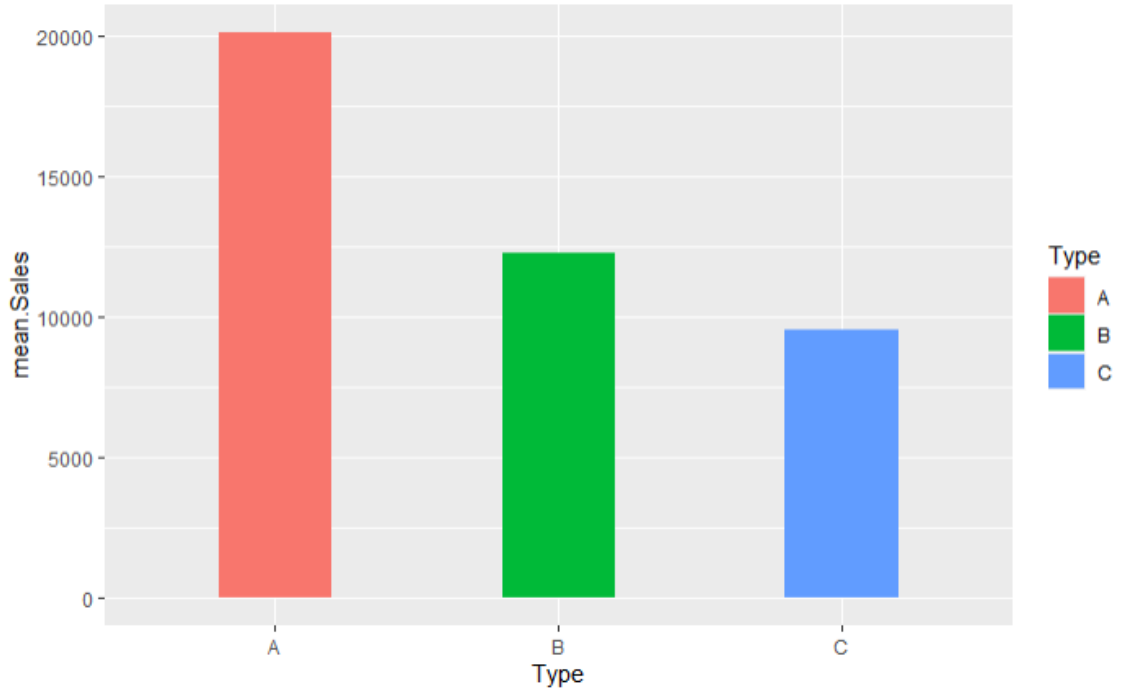
1. Store size :

As we defined earlier the store size means the number of articles present on that store so we wanted to check if the store size has an effect on the sales variable. Therefore, we tried to plot the variation of the sales depending on the store size as shown in the next figure .As we expected sales tend to get higher when the store size is more important which means people tend to buy from bigger store that have more articles and more choice eventually .



2. Store type :

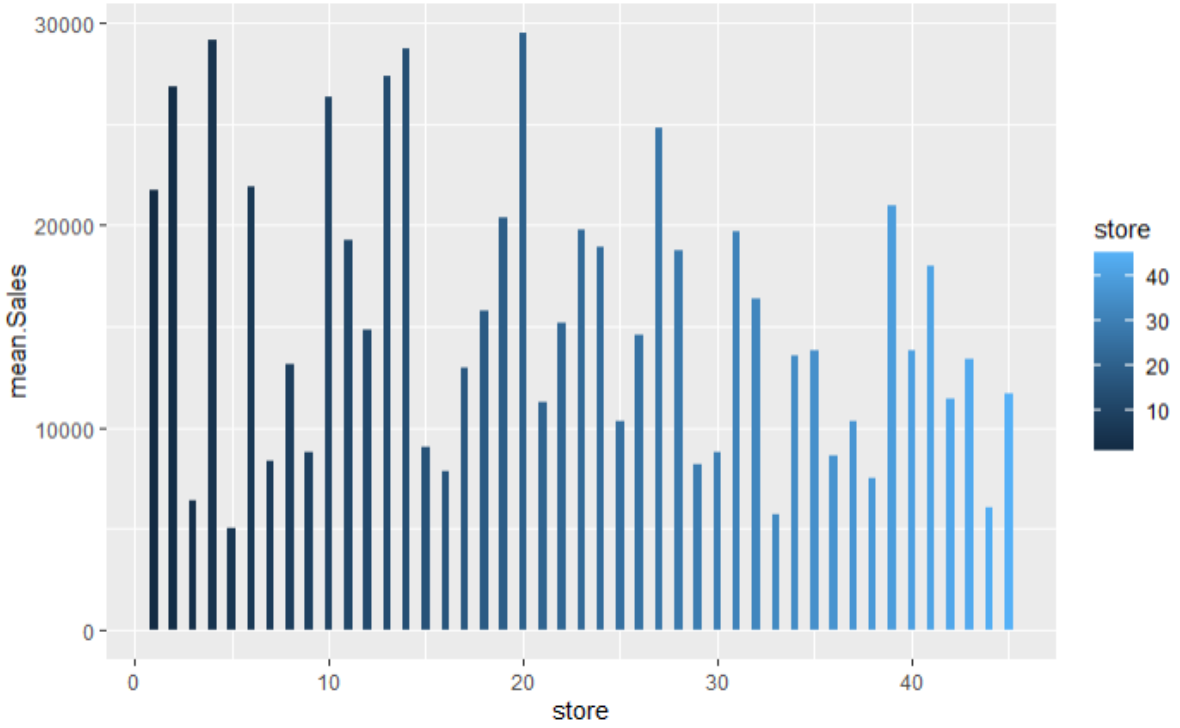
The type of the store is certainly a very important variable that might obviously affect the sales variable so we wanted to show that effect on the studied data set by ploting the sales as a function of the store type using the bar plot . As we can see in the figure the type A store has the most important sales among the stores and the type see has the lowest .



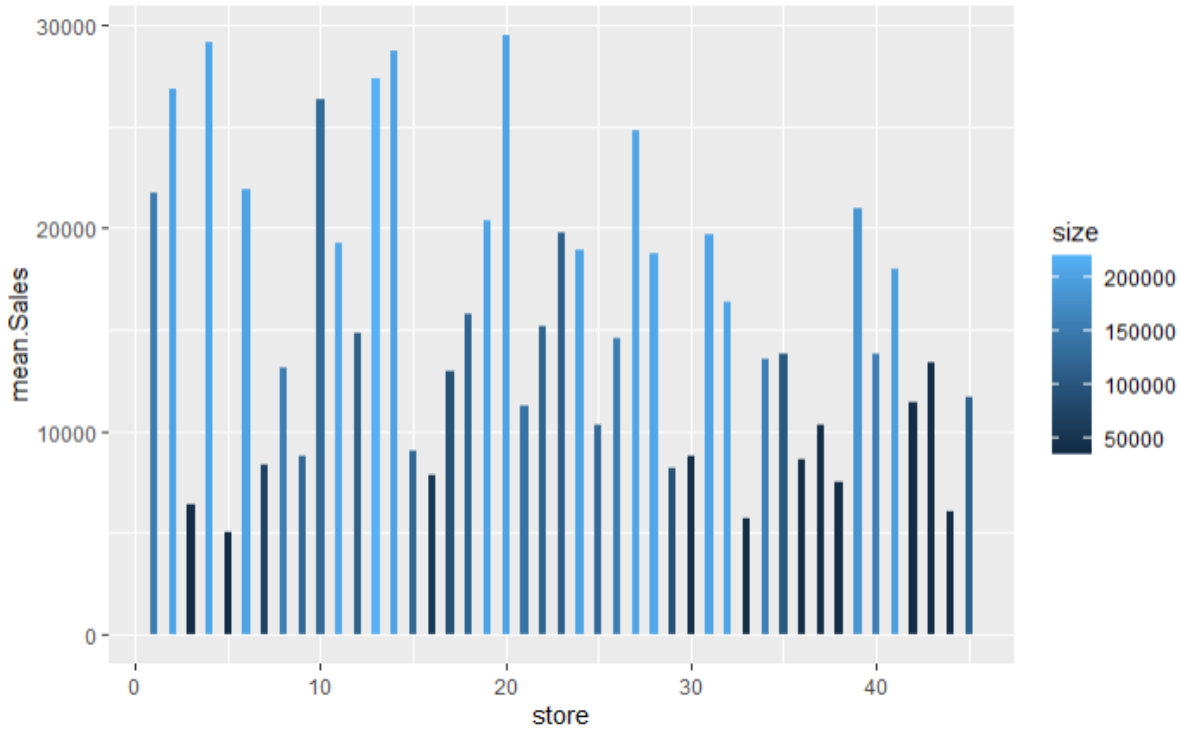
3. Store ID :

As we explained earlier that study was conducted on 45 stores we wanted to verify if the store ID varying from 1 to 45 can have a significant effect on the sales variation. The figure shows the variation of the sales depending on the store ID . as expected sales do vary from one store to an other the store N°20 has the best sales and the store N°5 has the

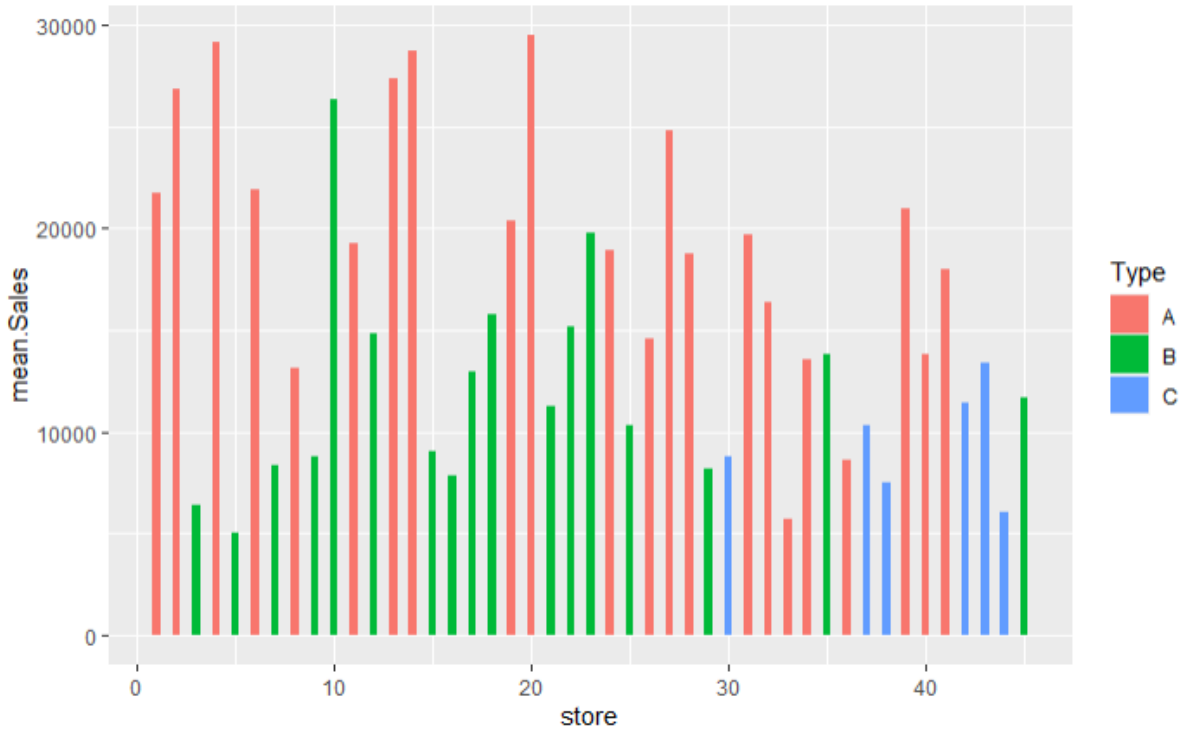
lowest of them all .



But these results don’t seem to be enough on their own we want to explain futher so we changed the filling to see what really affected the sales in these stores we started by filling by store size as shown in the figure and we see that our first hypothesis is confirmed about the importance of the size



To confirm our second hypothesis we went further to see the store ID and type effect on the sales by using the type variable as a filling. We ended up confirming that our deduction. The stores that made the most important sales are generally type A stores and have bigger size compared to others



III. Variation with other features :

1. Holiday :

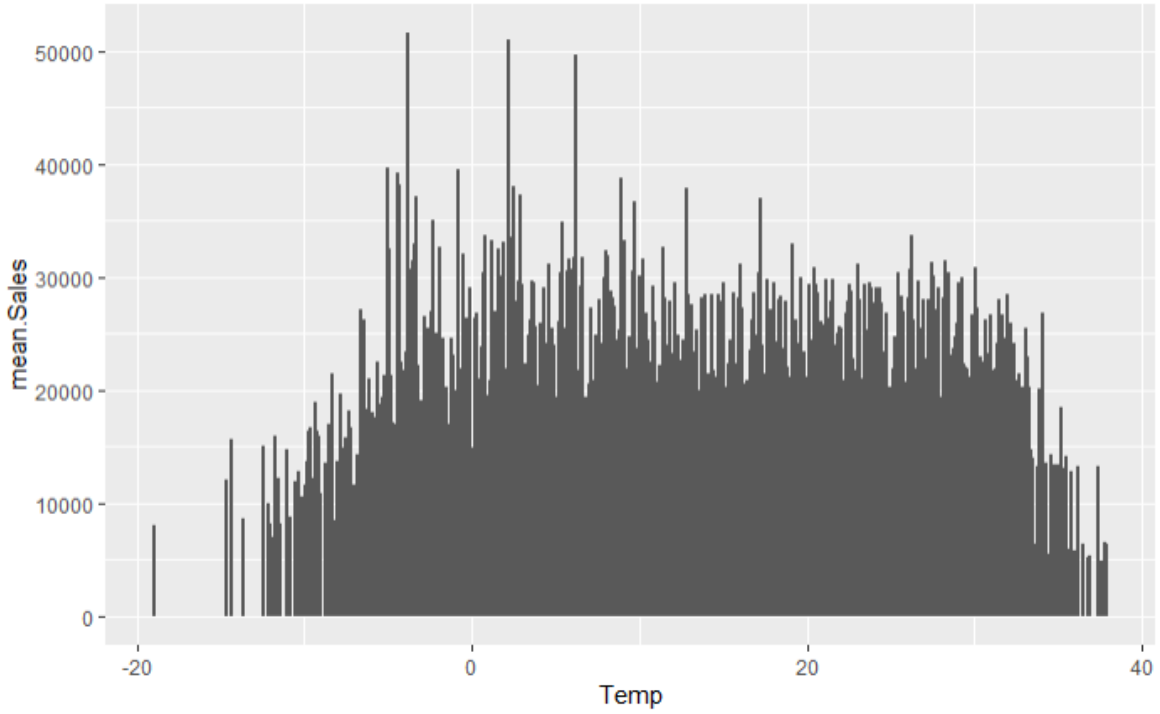
Holiday in this data is a Boolean variable telling if that date of was a holiday which can be bringing a very import change to the sales variable. We plotted the sales variable as a factor of the holiday variable. Results came as we expected the sales tend to augment if the date of the purchase is a holiday. People tend to buy more on holidays then on normal days.

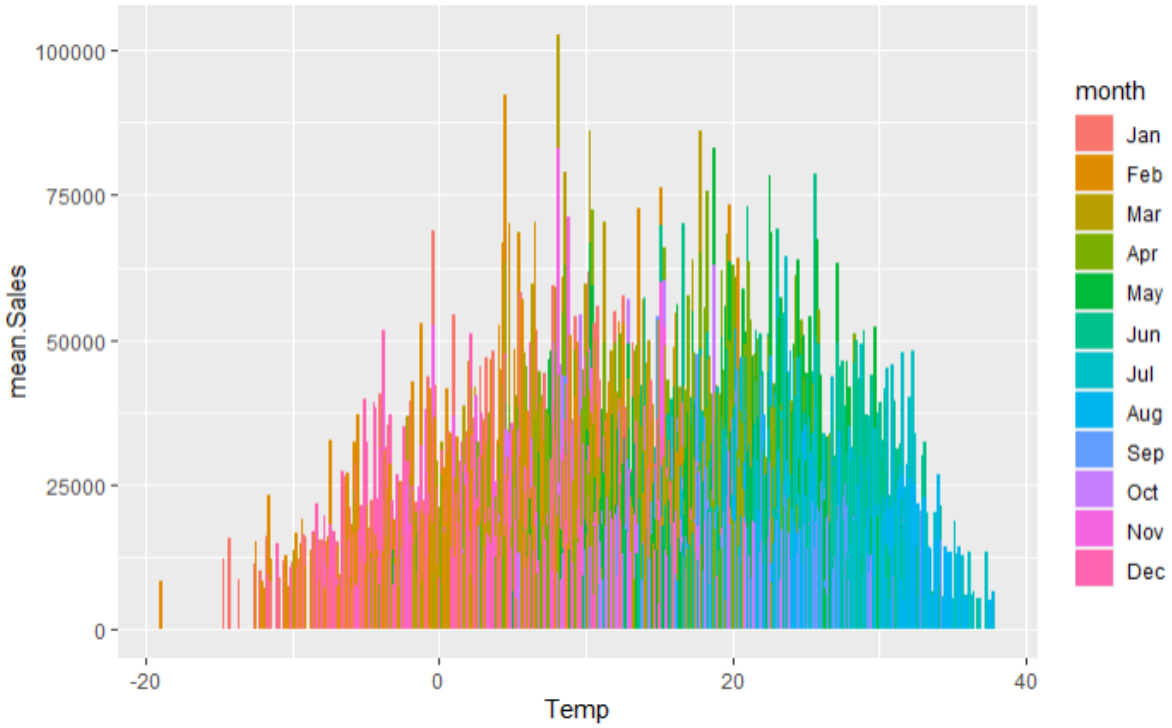
2. Temperature :

Temperature is an external variable and we don't think it would have a significant effect on sales so we wanted to check if this hypothesis was right buy creating the sales by temperature plot. We noticed that the temperature values were in Fahrenheit so we converted them to Celsius.

For this result we don't really think that temperature do affect the sales variable yet we can estimate that the temperature variable itself is affected by more significant variable such as the date we can clearly see that higher sales come with low temperatures that are observed in the Month of November and December as we can see in the figure using month variable as a filling.

We can also confirm that extreme temperatures very high or very low tend to lower sales but the average temperature don't have that much of an effect .

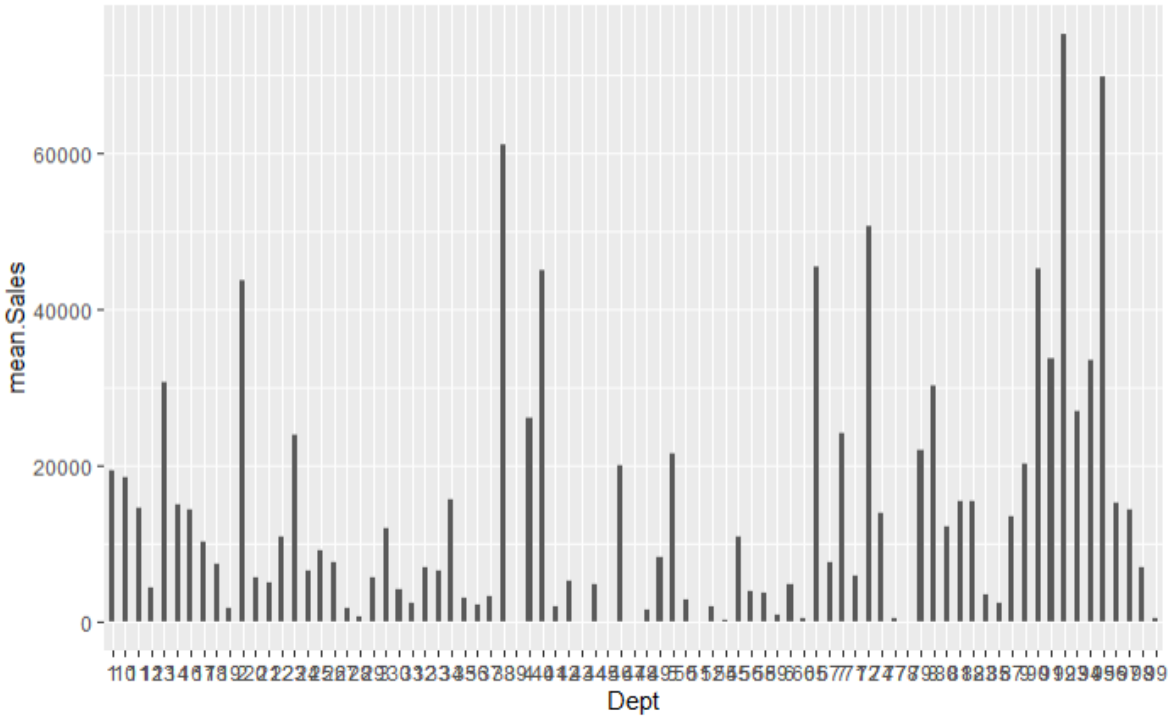


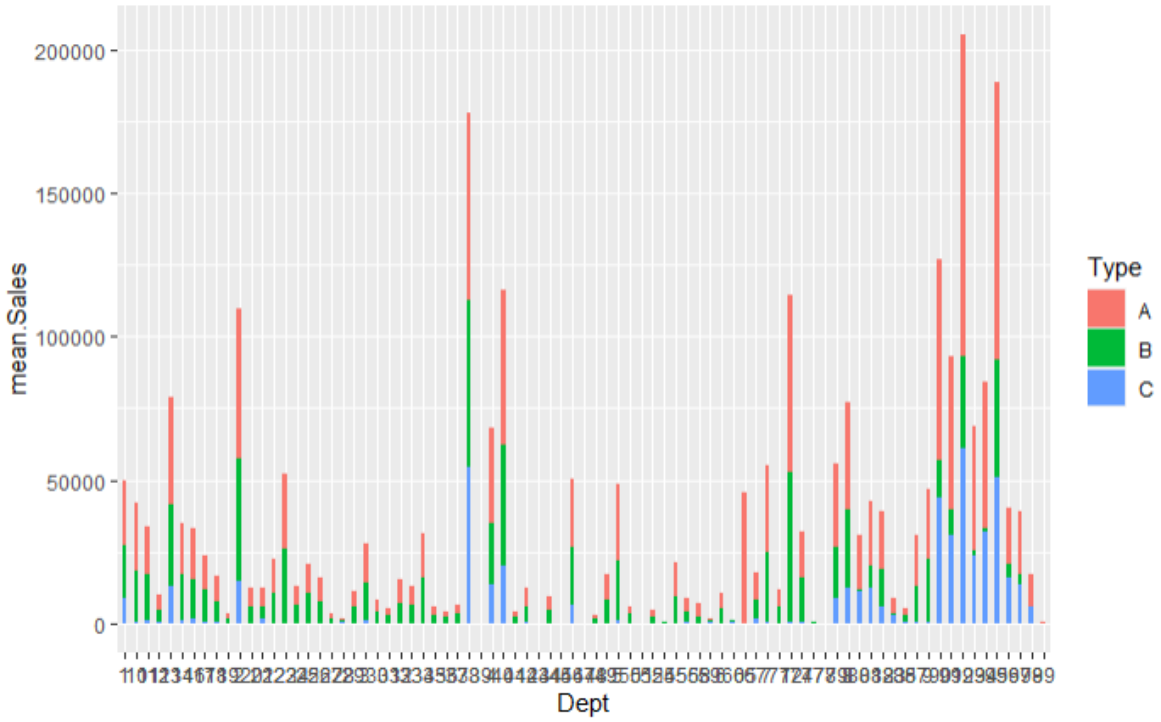


3. Department :

The department variable gives an idea about the location of the store and we think it would be really affecting the sales variable so we wanted to check by creating the sales by dept representation. We can clearly see that the sales vary obviously by department seeing the highest values in the department N°92 if this data is taking from French sales it is quite clear that department from the Paris region Ile de France have the highest sales . To make sure this

variation is totally due to dept alone we wanted to recreate the figure using the type as a filling and results came proving that department has a very significant effect no matter what the store type is.

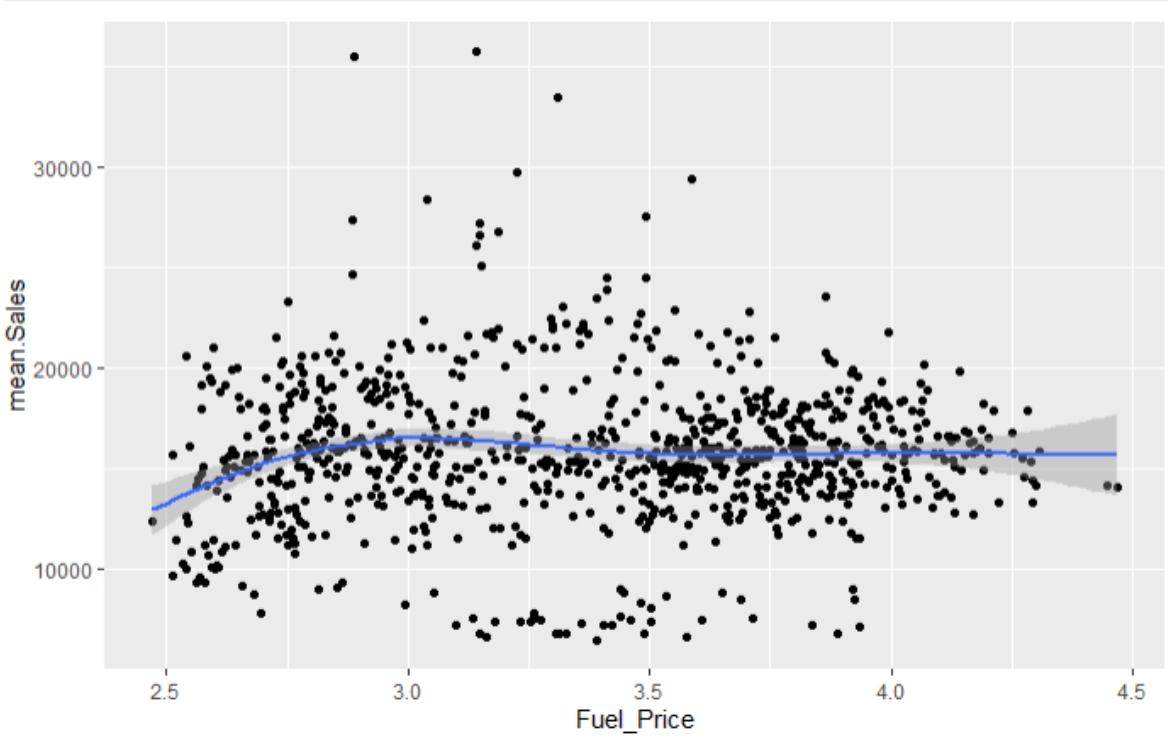




4. Fuel Price :

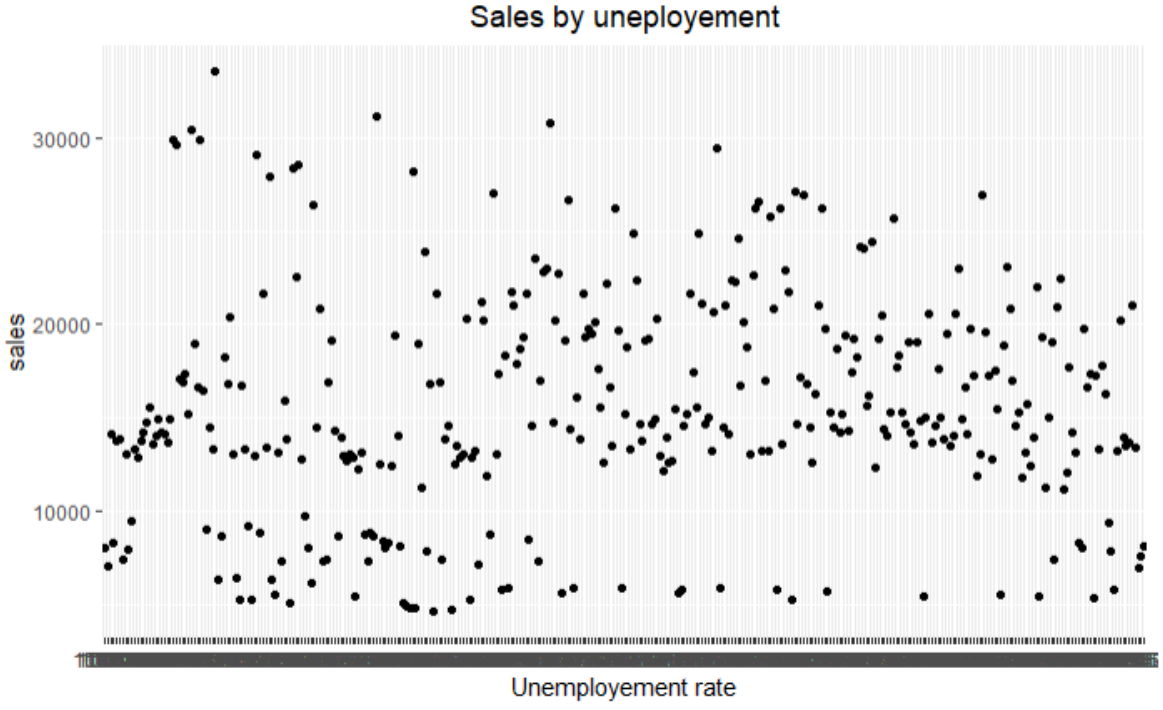
Fuel price is a very important variable and might have an effect on sales variable. Trying to prove that we represented the sales variation by fuel price. By analyzing the output results

came proving a quite stable variation but also an optimum with the highest sales observed in range between 2.75 and 3.25 per liter of fuel .



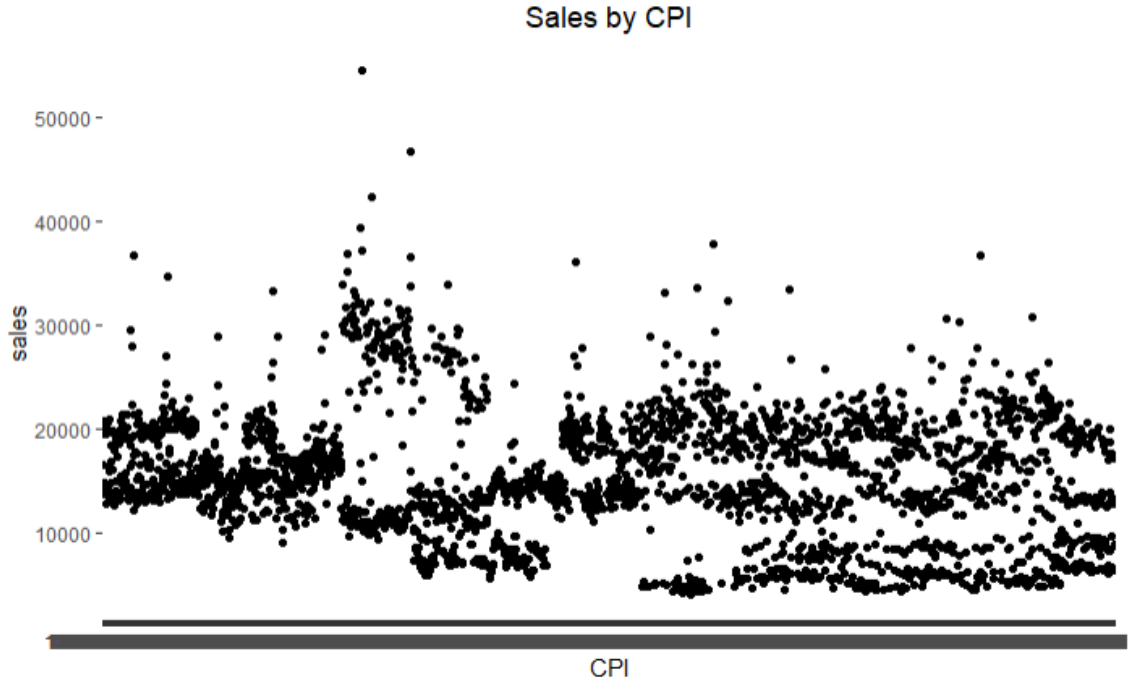
5. Unemployment :

Unemployment rate by region and week clearly affects the variable sales as it is shown in the plot . we can observe that sales tend to get lower when the unemployment index gets higher. The highest sales variable is seen in the low unemployment range .



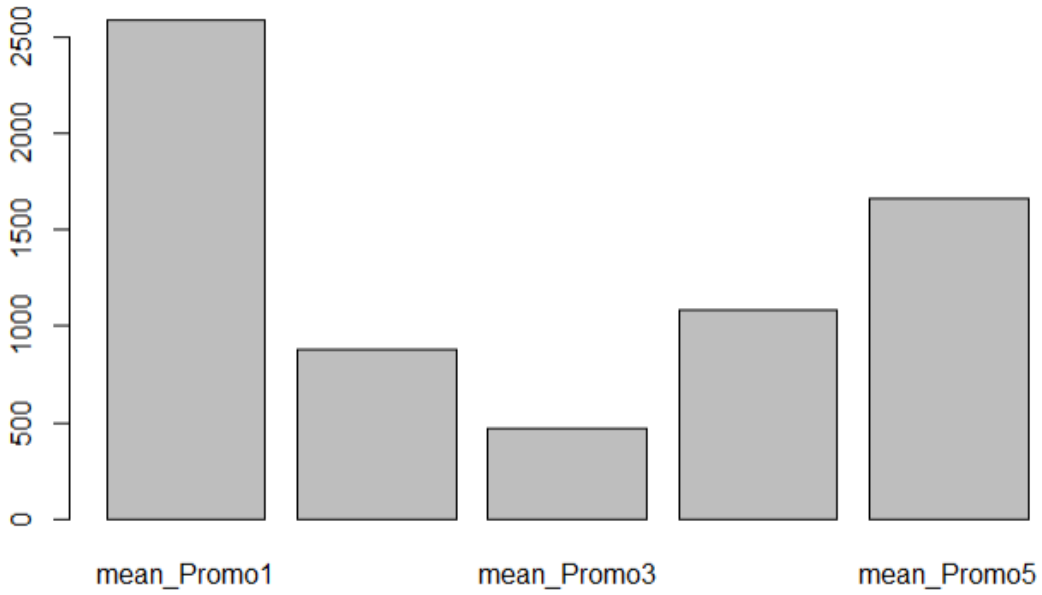
6. CPI: consumer price index :

The consumer price index of the corresponding region and week is a variable that might affect the sales variable. When the CPI augment sales tend to get lower .



7. Promotion :

Promotions really affects sales in a various way as seen in the next figure we see that promotion 1 really affects the sales making them the highest and promotion 3 gives the least sales



.