Retail Data Assignment

Appanna

11/28/2020

Introduction

We have received a dataset from our customer containing the retail data for 45 stores of his company. Using the columns in the dataset, our goal is to explain the target variable 'Sales'. In the course of our analysis, it is useful to first understand the data better. Let us start by describing the columns of the data given:

- Information shared by client about data:
  - Store: the ID of the store.
  - Data: the weekly balance sheet date.
  - Holiday: a binary variable informing if the considered week is Holyday or not. Type: the type of the store.
  - Size: the size of the store.
  - Dept: the department number of the store.
  - Sales: the volume of sales for the corresponding week.
  - Temperature: the average temperature in the region for the corresponding week.
  - Fuel_Price: the price of the fuel in the region for the corresponding week.
  - Promotion: price reduction for the corresponding store and week. The are 5 promotion categories.
  - CPI: the consumer price index of the corresponding region and week.
  - Unemployment: the unemployment rate in the region for the corresponding week.

Body

- First we will import the data into RStudio and explore the data for a meta-analysis.
- Next we will plot some visualizations to get an intuitive feel for the dataset.
- After this, we can plot some visualizations to give us the behavior of the Sales column.
- Finally we can fit the data into a linear model and check various combinations to identify to the best extent that the 'Sales' column can be explained.

Data Exploration

- Import into RStudio and view the data.

```
## X Store  Date       Holiday Type Size  Dept  Sales   Temperature Fuel_Price
1  1   1  01/04/2011   FALSE    A 151315   49 13167.85       59.17      3.524
2  2   1  01/04/2011   FALSE    A 151315   26  5946.53       59.17      3.524
```

```
## Promotion1 Promotion2 Promotion3 Promotion4 Promotion5 CPI Unemployment
 1   NA          NA          NA          NA          NA     214.8372  7.682
 2   NA          NA          NA          NA          NA     214.8372  7.682
```

- Just by looking at the above table, we can infer the following:
  - Only the 'Dept' and 'Sales' column have unique elements in each row. This means the Sales volume value refers to each unique 'Dept'. The 'X' column likely refers to row number which is insignificant as a feature for analysis.
  - The 'Promotion' columns have missing values.
  - All the other columns have the same values in both rows. This implies that they remain common across some grouping of the elements.
- Let us import the libraries that we might use:

```
library(tidyverse) #bundle of packages for data grouping, cleaning and
visualizing
library(reshape2) #if we choose to 'melt' the data for different plotting
options.
library(Amelia) #to check the missing values.
library(naniar) #to check the missing values.
library(ggplot2) #for beautiful visualizations.
library(GGally) #for a nice correlation plot with multiple columns.
library(lubridate) #to load the 'Date' column into a format for easy
manipulation.
```

- How many rows and columns does the dataset contain:

```
dim(retail_data)

## [1] 421570     17
```

Dataset contains 421.570 rows and 17 columns.

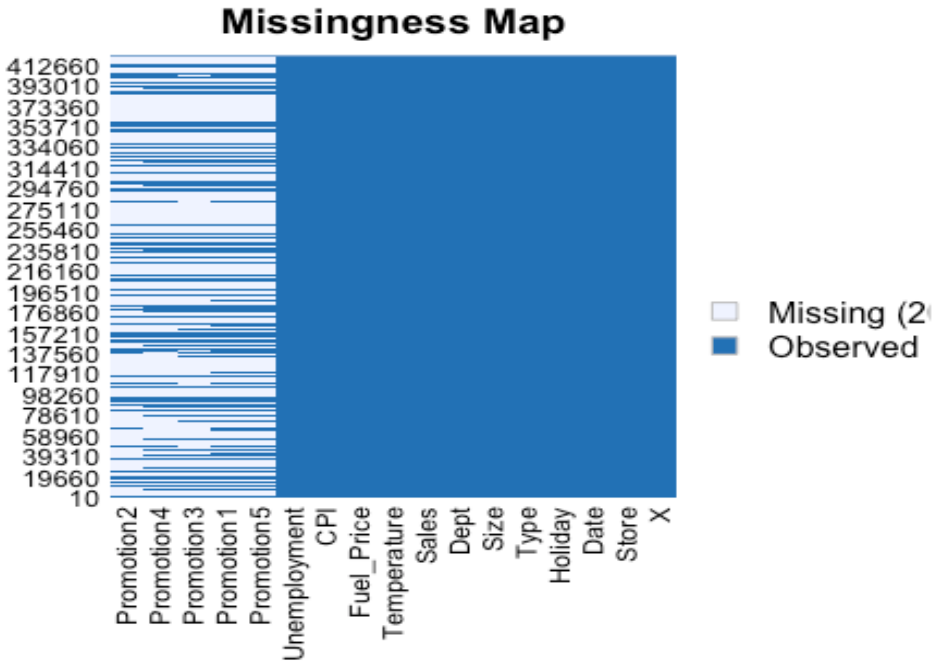
- Numeric summary of the data:
  - We can notice that the 'Dept' and 'Store' have been considered as continuous variables. For the plots and analysis to work well, we can convert these into factors.

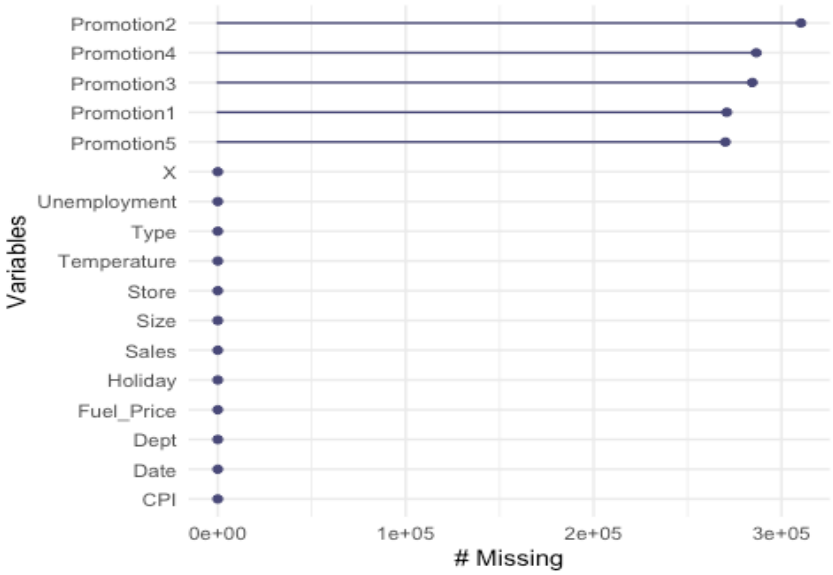
```
summary(retail_data)

##        X              Store              Date         Holiday        Type
##  Min.   :     1   Min.   : 1.0   23/12/2011:3027   Mode :logical  A:215478
##  1st Qu.:105393   1st Qu.:11.0   25/11/2011:3021   FALSE:391909   B:163495
##  Median :210786   Median :22.0   16/12/2011:3013   TRUE :29661    C:42597
##  Mean   :210786   Mean   :22.2   09/12/2011:3010
##  3rd Qu.:316178   3rd Qu.:33.0   17/02/2012:3007
##  Max.   :421570   Max.   :45.0   30/12/2011:3003
##                                  (Other)   :403489
##       Size             Dept           Sales          Temperature
##  Min.   : 34875   Min.   : 1.00   Min.   : -4989   Min.   : -2.06
##  1st Qu.: 93638   1st Qu.:18.00   1st Qu.:  2080   1st Qu.: 46.68
##  Median :140167   Median :37.00   Median :  7612   Median : 62.09
##  Mean   :136728   Mean   :44.26   Mean   : 15981   Mean   : 60.09
```

```
##  3rd Qu.:202505   3rd Qu.:74.00   3rd Qu.: 20206   3rd Qu.: 74.28
##  Max.   :219622   Max.   :99.00   Max.   :693099   Max.   :100.14
##
##    Fuel_Price      Promotion1         Promotion2         Promotion3
##  Min.   :2.472   Min.   :    0.27   Min.   :  -265.8   Min.   :   -29.10
##  1st Qu.:2.933   1st Qu.: 2240.27   1st Qu.:    41.6   1st Qu.:     5.08
##  Median :3.452   Median : 5347.45   Median :   192.0   Median :    24.60
##  Mean   :3.361   Mean   : 7246.42   Mean   :  3334.6   Mean   :  1439.42
##  3rd Qu.:3.738   3rd Qu.: 9210.90   3rd Qu.:  1926.9   3rd Qu.:   103.99
##  Max.   :4.468   Max.   :88646.76   Max.   :104519.5   Max.   :141630.61
##                  NA's   :270889     NA's   :310322     NA's   :284479
##    Promotion4         Promotion5          CPI         Unemployment
##  Min.   :    0.22   Min.   :   135.2   Min.   :126.1   Min.   : 3.879
##  1st Qu.:  504.22   1st Qu.:  1878.4   1st Qu.:132.0   1st Qu.: 6.891
##  Median : 1481.31   Median :  3359.4   Median :182.3   Median : 7.866
##  Mean   : 3383.17   Mean   :  4629.0   Mean   :171.2   Mean   : 7.960
##  3rd Qu.: 3595.04   3rd Qu.:  5563.8   3rd Qu.:212.4   3rd Qu.: 8.572
##  Max.   :67474.85   Max.   :108519.3   Max.   :227.2   Max.   :14.313
##  NA's   :286603     NA's   :270138
```

- Check for missing values:






```
##      X Store Date Holiday Type Size Dept Sales Temperature Fuel_Price
## [1,] 0 0     0    0       0    0    0    0     0           0
##      Promotion1 Promotion2 Promotion3 Promotion4 Promotion5 CPI
Unemployment
## [1,] 270889     310322     284479     286603     270138     0   0
```

We can see that only the 'Promotion' columns have missing values. 80% of the values are available while 20% of the values are missing from the total dataset.

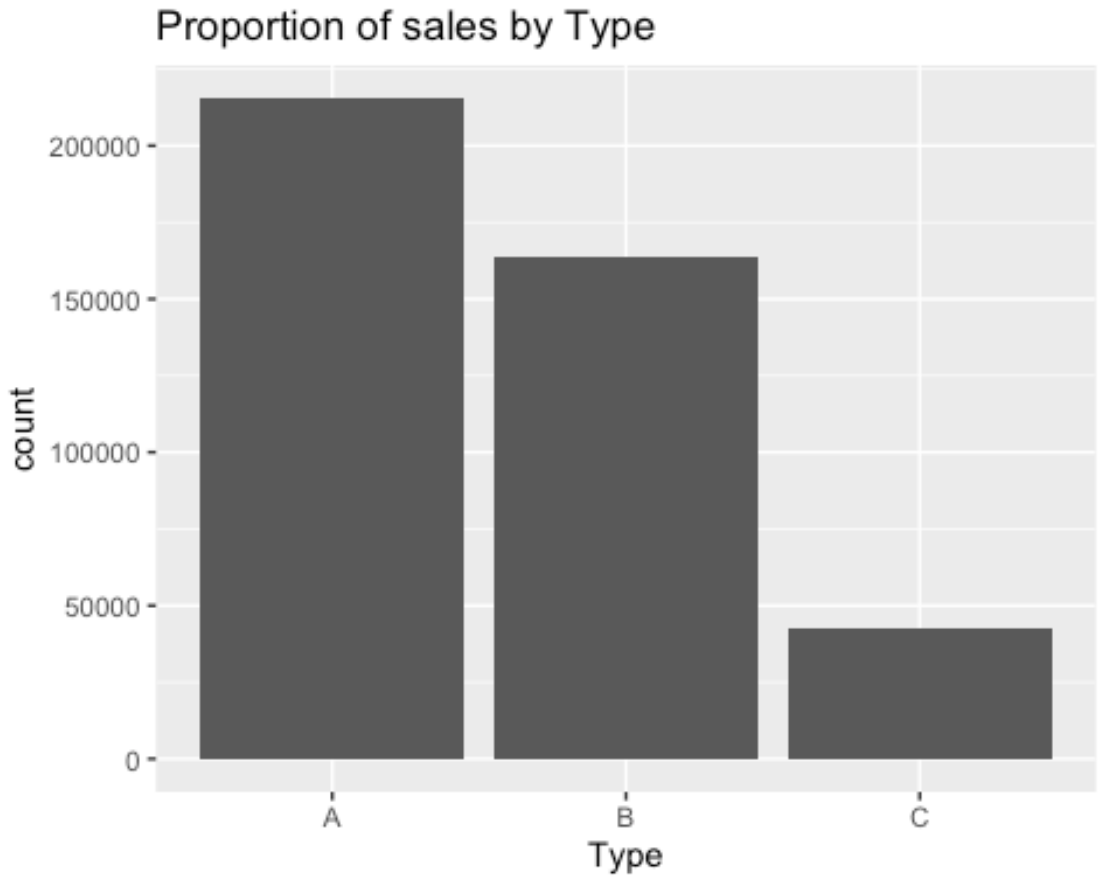
- What defines the uniqueness of each column in the dataset:
  - First we can check what unique combination of columns results in the total rows of 421570.
  - It appears that the rows are unique by Store, Date and Dept.
  - This means that 'Sales' is the value for sales volume of a 'Dept' in a 'Store' on a particular 'Date'. Each date refers to a week of sales.
  - 1 'Date' has many 'Stores'. Each 'Store' has many 'Depts'. Each 'Dept' represents a row in the table with 'Sales' corresponding to that 'Store' and 'Date'.
  - A particular 'Store' has only 1 value of 'Temperature', 'Fuel_Price', 'CPI', 'Unemployment' on a particular 'Date'. All 'Depts' in this 'Date', 'Store' combination have same values for these fields.
  - Many 'Stores' are affected by same 'Temp'/'Fuel_Price'/'CPI'/'Unemployement' on the same 'Date'. This could be because the stores are close by to each other.

Intuitive Visualizations

- These visualizations will mainly describe representation of different columns based on the count of rows in the dataset. This can give us an intuitive feel for the distibutions within the dataset.

- Proportion of sales data by Type:
  - There are 3 Types into which the stores are categorized.

- As we can see the Type A represents most of the dataset followed closely by Type B.
- Type C has much lesser representation.



- Number of stores within each 'Type'.

```
## # A tibble: 3 x 2
##   Type  nStores
##   <fct>   <int>
## 1 A          22
## 2 B          17
## 3 C           6
```

- Store column
  - The Store number ranges from 1 to 45 and the total number of stores are:

```
## [1] 45
```

- Dept column
  - 'Dept' numbers range from 1 to 99(inclusive) as observed in the 'summary' table and the number of unique 'Dept' are:

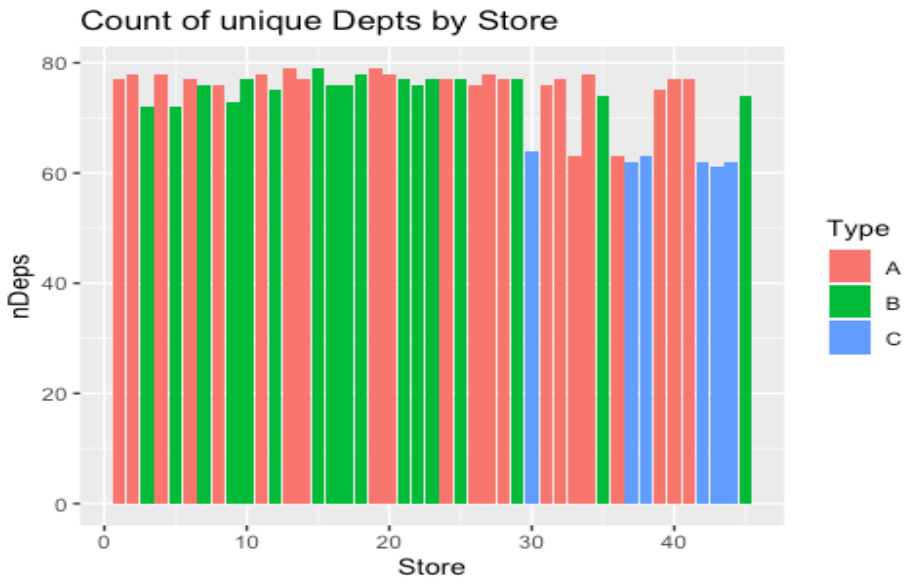
```
## [1] 81
```

- There are between 60 to 80 departments per store.

- No store has all 81 departments.

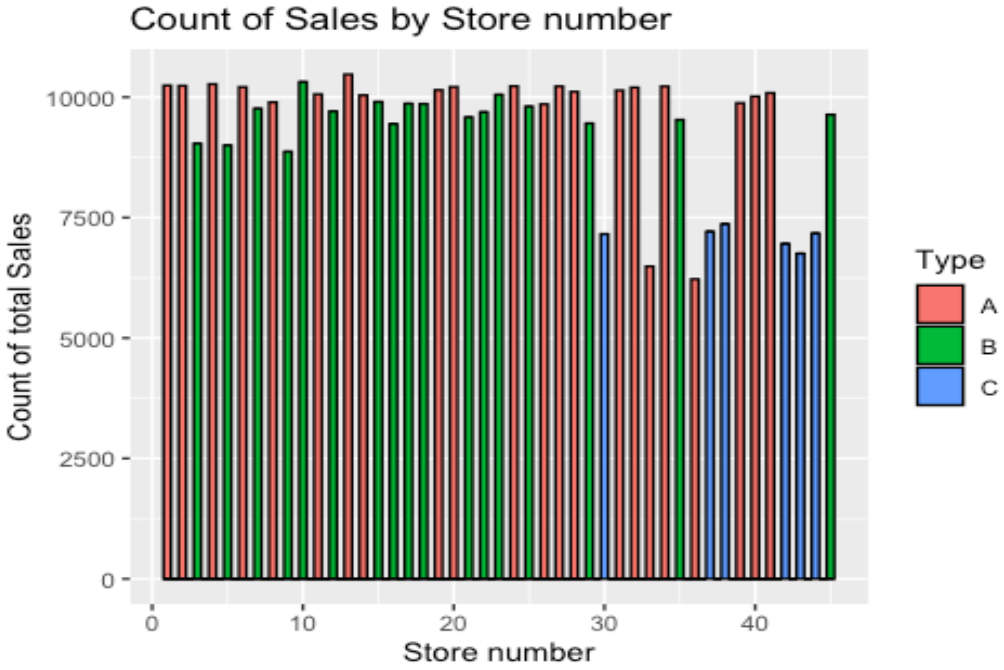
- Most of the stores have more than 70 departments

- If there are less than 70 Depts then it is likely to be a Type C store.

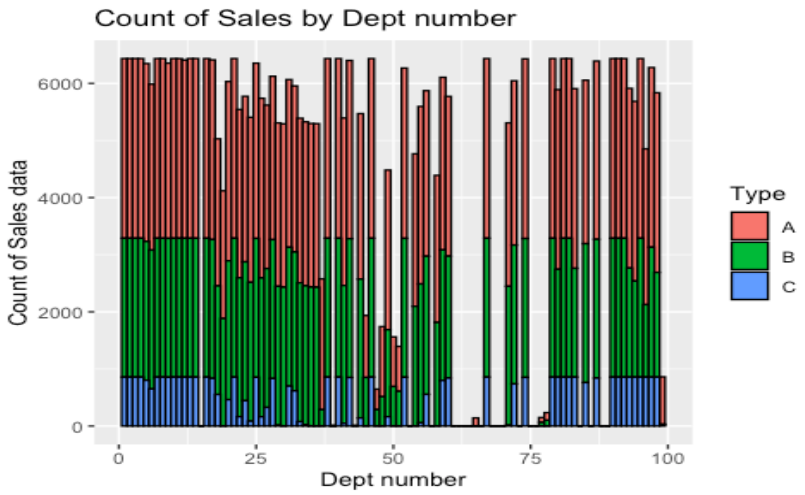


- Histogram of 'Store'

- There seems to be a correlation between number of departments in a store and count of sales. This is because if there are less 'Depts' then there is fewer data under a 'Store'.



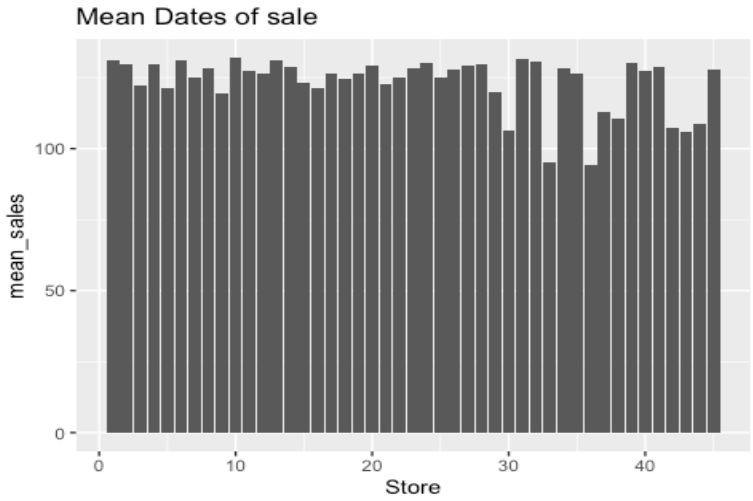
- Histogram of 'Dept'



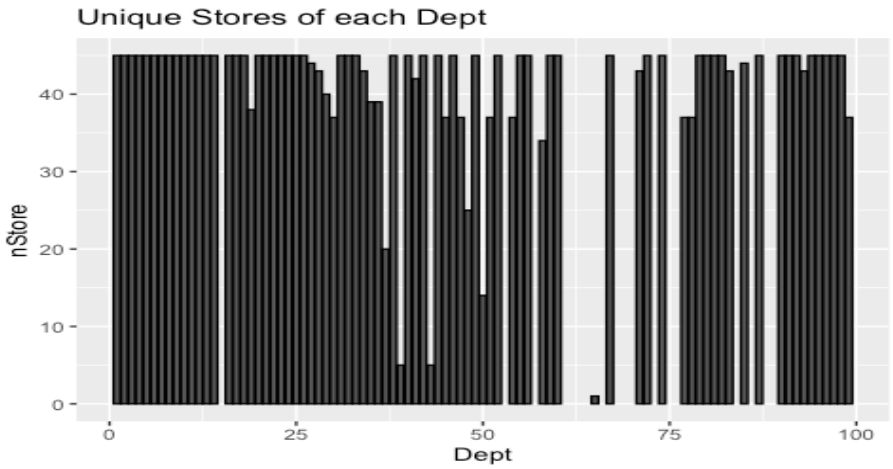
- Number of unique Dates is:

```
## [1] 143
```

- Mean number of Dates of Sale for which there is availability of sales data for different stores:
  - A Dept appears for a date only in there is Sales available at that 'Store' on that 'Date'.
  - Most of the Depts have stores on all 143 dates but some Depts of a store may have had sales on only a few days. This brings down the average 'Dates' of Sale for a 'Store'.
  - As we can see in the plot below, most 'Stores' have a mean number of around 125 Dates of sale.

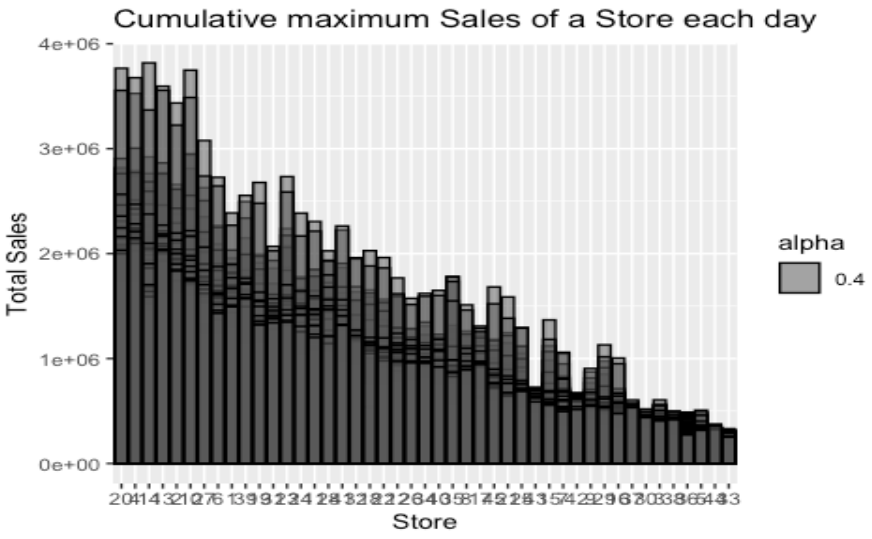


- Number of unique Stores for each Dept
  - Most of the departments appear at least on 1 date in all 45 stores.
  - In the number range of 1 to 100 for Depts, some numbers do not represent any Dept.

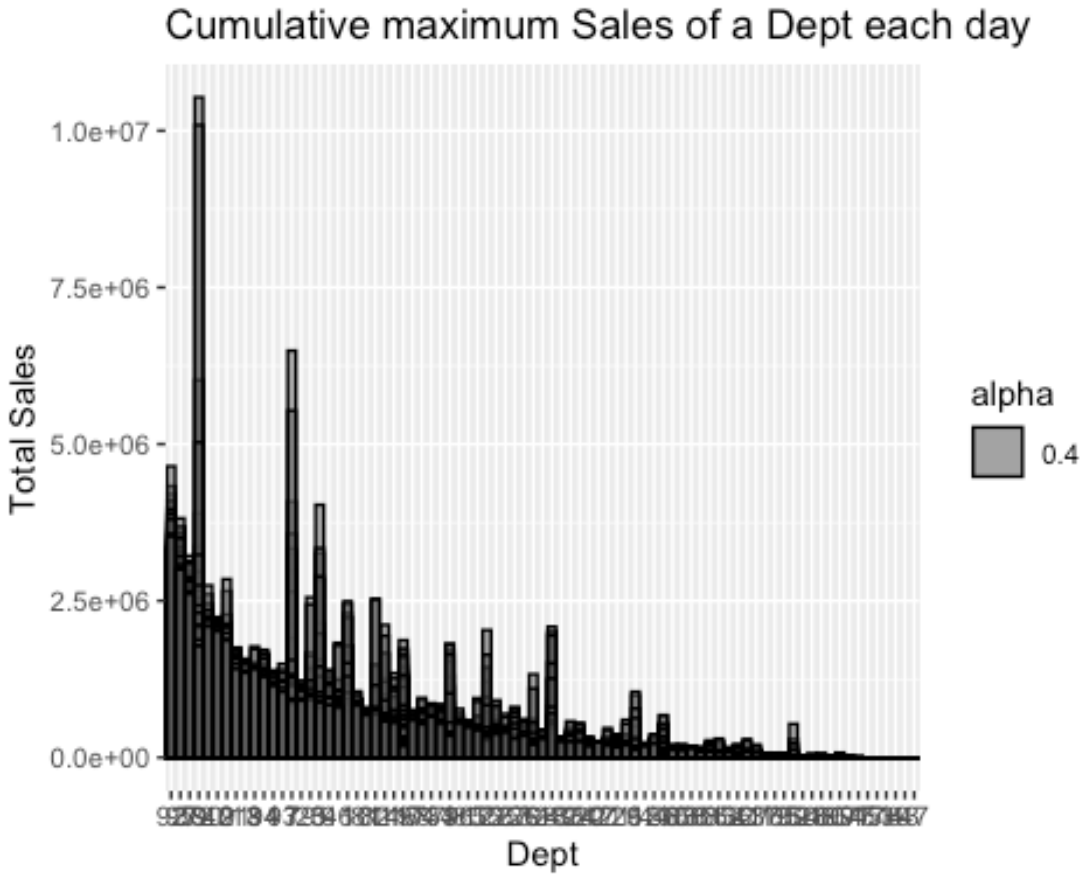


Patterns in Sales volume

- Max Sales on a day by Store
  - The different bars for each Store (which are stacked behind each other) refer to the different days.
  - We can observe that the top performing stores, are performing well for most days. This fairly uniform pattern of performance suggests that this might be a good variable for our regression model.
  - The bar chart is ordered by cumulative maximum sales per location(all days)
  - The top 6 Stores are: 20,4,14,13,2 and 10. After this there is a dip in performance by the other Stores.



- Max Sales by Dept
  - The different bars for each Dept (which are stacked behind each other) refer to the different days.
  - Observe that even for the abnormal peaks, there are not many 'Dates' (indicated by transparency of the bar segments). From a total of 145 Dates, only few of these are abnormal.
  - We can see that some 'Depts' have a much higher volume of Sales on some days. These non uniform spikes indicate that this might not be a good variable for our regression model.



- Top 10 Depts for maximum overall sales:
  - These correspond to the 10 Depts representing bars from the left of the above plot.

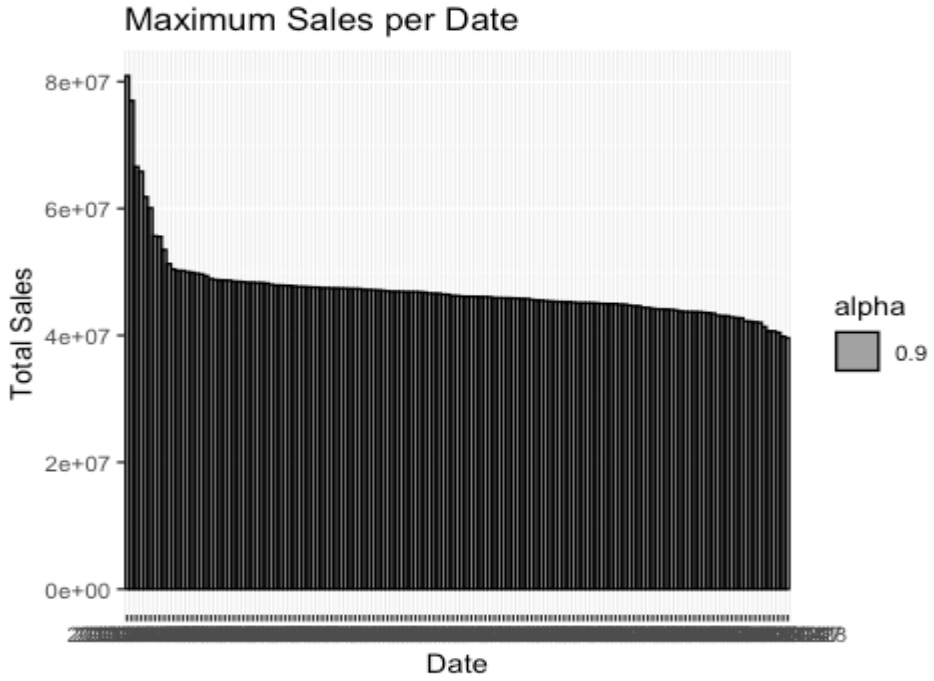
```
## # A tibble: 10 x 2
##     Dept    s_Sales
##    <int>      <dbl>
##  1    92 483943342.
##  2    95 449320163.
##  3    38 393118137.
##  4    72 305725152.
##  5    90 291068464.
##  6    40 288936022.
##  7     2 280611175.
```

```
##  8    91 216781706.
##  9    13 197321570.
## 10     8 194280781.
```

- Top 10 Depts which have the highest Sales for a day:
  - Each bar in the previous graph represents the total sales of a particular Dept across all Stores for a particular Date. So there are as many bar segments as there are Dates for that Dept.
  - The table below corresponds to the Depts creating the 10 highest spikes in the previous chart which is making the distribution abnormal:

```
## # A tibble: 10 x 3
## # Groups:   Dept [10]
##    Date        Dept   s_Sales
##    <date>     <int>     <dbl>
##  1 2010-11-26    72 10533641.
##  2 2010-12-24     7  6490778.
##  3 2011-12-23    92  4647998.
##  4 2010-12-24     5  4032443.
##  5 2012-07-06    95  3816877.
##  6 2011-01-07    38  3219846.
##  7 2011-12-23     2  2846583.
##  8 2011-12-23    90  2747983.
##  9 2010-12-24    23  2557656.
## 10 2011-12-23    82  2533132.
```

- Max Sales by Date
  - It appears that the top 10 Dates are higher than the remaining Dates which all have around the same volume of Sales.



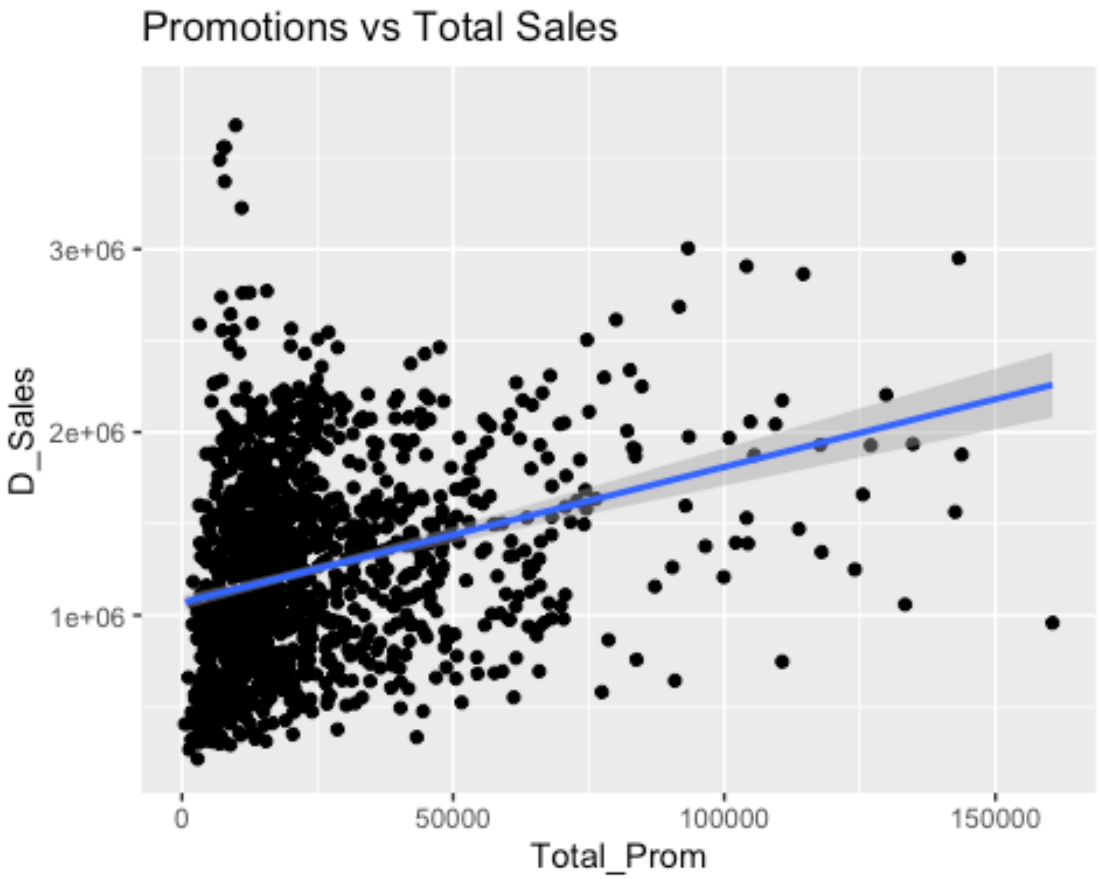
- Top 10 days of Sales
  - We notice that December is the most common month in the top 10 Dates with 6 days.
  - This is followed by the month of November with 2 days.
  - Lastly we have a single date each in April and July with large sales volume.

```
## # A tibble: 10 x 4
##    Year  Month Day     D_Sales
##    <fct> <fct> <fct>     <dbl>
##  1 2010  12    24    80931415.
##  2 2011  12    23    76998241.
##  3 2011  11    25    66593605.
##  4 2010  11    26    65821003.
##  5 2010  12    17    61820800.
##  6 2011  12    16    60085696.
##  7 2010  12    10    55666771.
##  8 2011  12    9     55561148.
##  9 2012  4     6     53502316.
## 10 2012  7     6     51253022.
```

- Top 10 peak days of Sales by Store:
  - We can see that Dec 23rd and 24th cause the highest sales across different Stores.

```
## # A tibble: 10 x 5
##    Year  Month Day   Store  D_Sales
##    <fct> <fct> <fct> <fct>    <dbl>
##  1 2010  12    24    14    3818686.
##  2 2010  12    24    20    3766687.
##  3 2010  12    24    10    3749058.
##  4 2011  12    23    4     3676389.
##  5 2010  12    24    13    3595903.
##  6 2011  12    23    13    3556766.
##  7 2011  12    23    20    3555371.
##  8 2010  12    24    4     3526713.
##  9 2011  12    23    10    3487987.
## 10 2010  12    24    2     3436008.
```

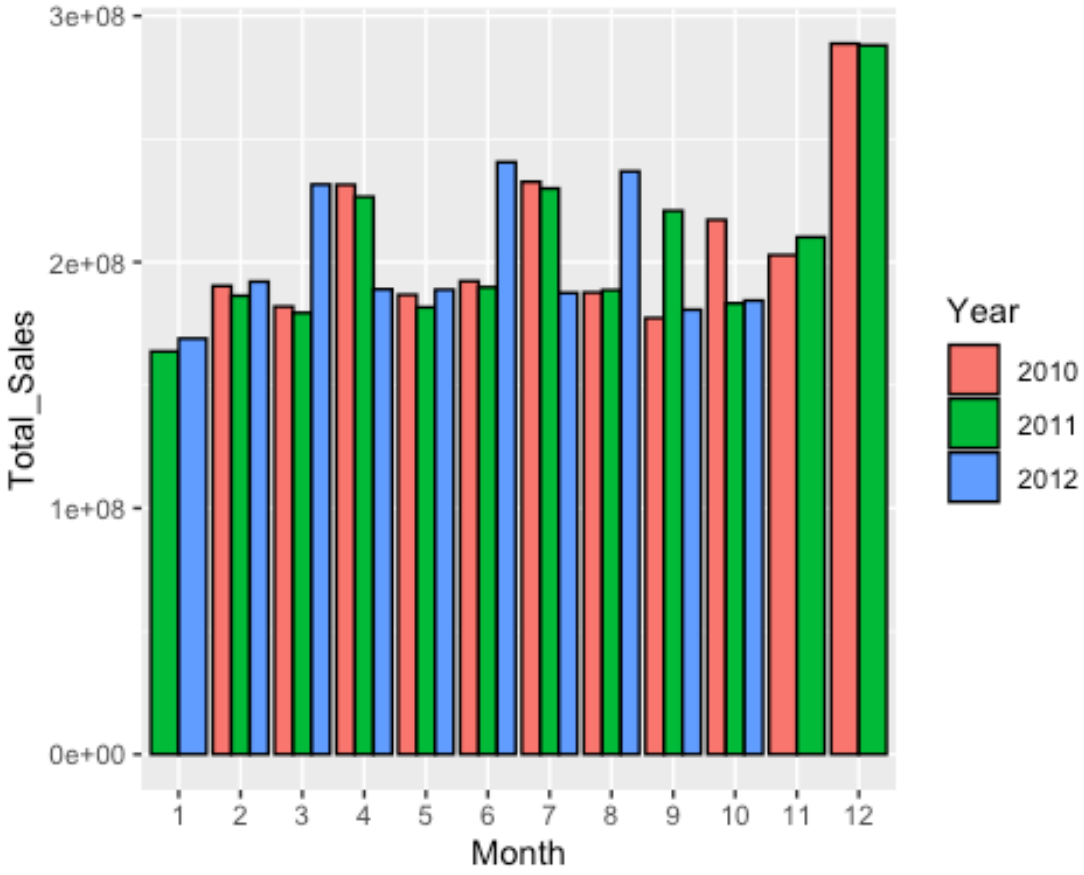
- Promotions by Date and Dept:
  - From the dataset we know that promotions apply to all Depts of a Store on a particular day.
  - Total_promotions is the sum of all the different types of promotions (Promotion1, Promotion2 etc).
  - We can see that there is a slight positive trend between total Sales of a Store on a Date and Total_promotions.
  - However, the highest sales from a Store on a given Date still occurs on a day with less Total_promotions.



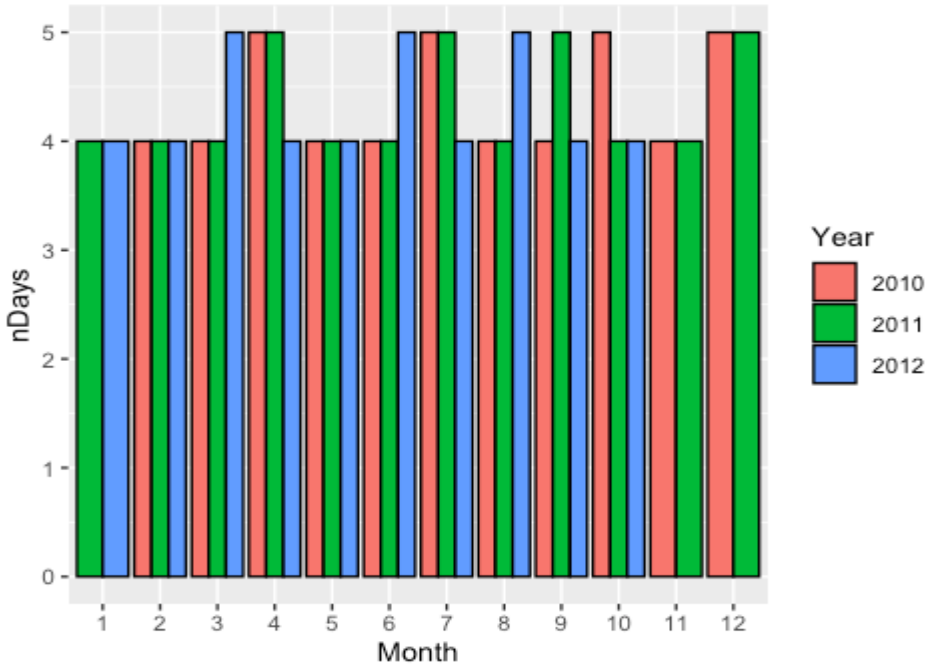
- Top 10 peak days of Sales by Dept and Store:
  - Store number 10 has 4 days of the highest volume of sales. Store 35 has 2 days and Stores 14,20,27, and 22 have 1 day each.
  - We can see that for most Stores, Dept 72 accounts for the maximum volume of Sales on a given day.
  - There are 3 main days where this happens: Nov 25th and 26th; Dec 24th. Most of these days are in the year 2010.

```
## # A tibble: 10 x 6
##    Year  Month Day   Store  Dept D_Sales
##    <fct> <fct> <fct> <fct> <int>   <dbl>
##  1 2010  11    26    10       72 693099.
##  2 2011  11    25    35       72 649770.
##  3 2011  11    25    10       72 630999.
##  4 2010  11    26    35       72 627963.
##  5 2010  11    26    14       72 474330.
##  6 2010  11    26    20       72 422306.
##  7 2010  11    26    27       72 420587.
##  8 2010  12    24    10        7 406989.
##  9 2010  12    24    10       72 404245.
## 10 2010  11    26    22       72 393705.
```

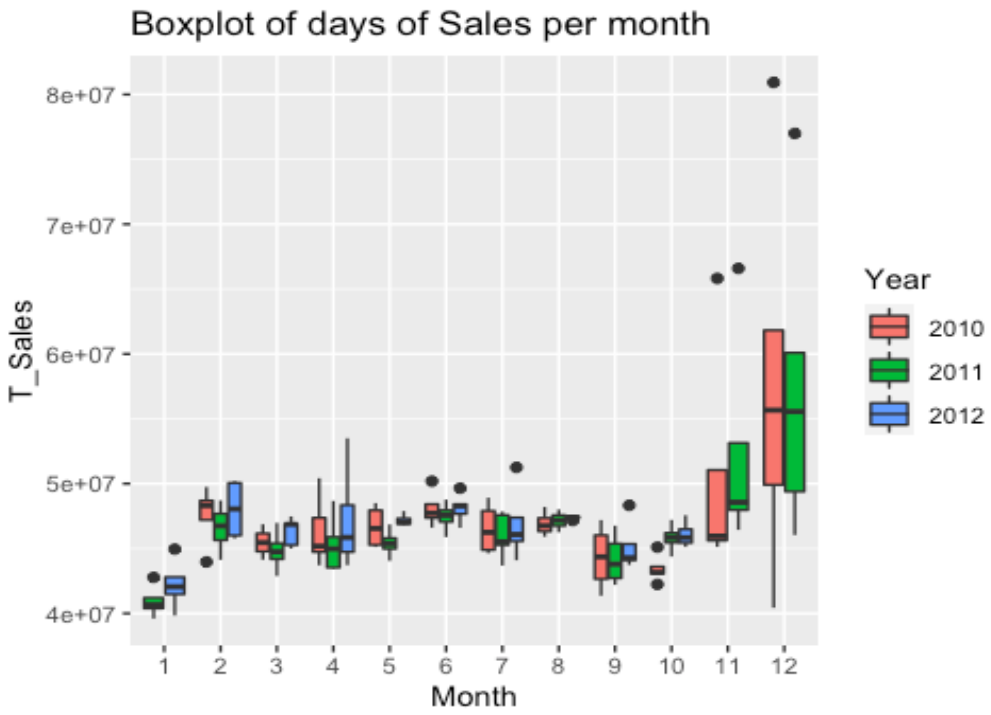
- Total Sales by Month
  - There are 3 years of data starting from Feb 2010 to Oct 2012.
  - We can see that December has the highest sales for any month.
  - There appears to be 4 major spikes in the Sales for any year:
    - 2010->Apr,July,Oct,Dec
    - 2011->Apr,July,Sept,Dec
    - 2012->Mar,June,Aug
  - These spikes seem to correspond to the extra date available for that month. Remember that each date mostly refers to the week of sales. It may be the case that, based on which day of week is considered for publishing the data for that week, some months may have 5 data points some may have 4.



- Check the number of Days per month in the dataframe
  - There are 5 days of data for the months corresponding to the Total Sales per month peaks.
  - On all other months only 4 days of data are available. These days correspond to a week of data.

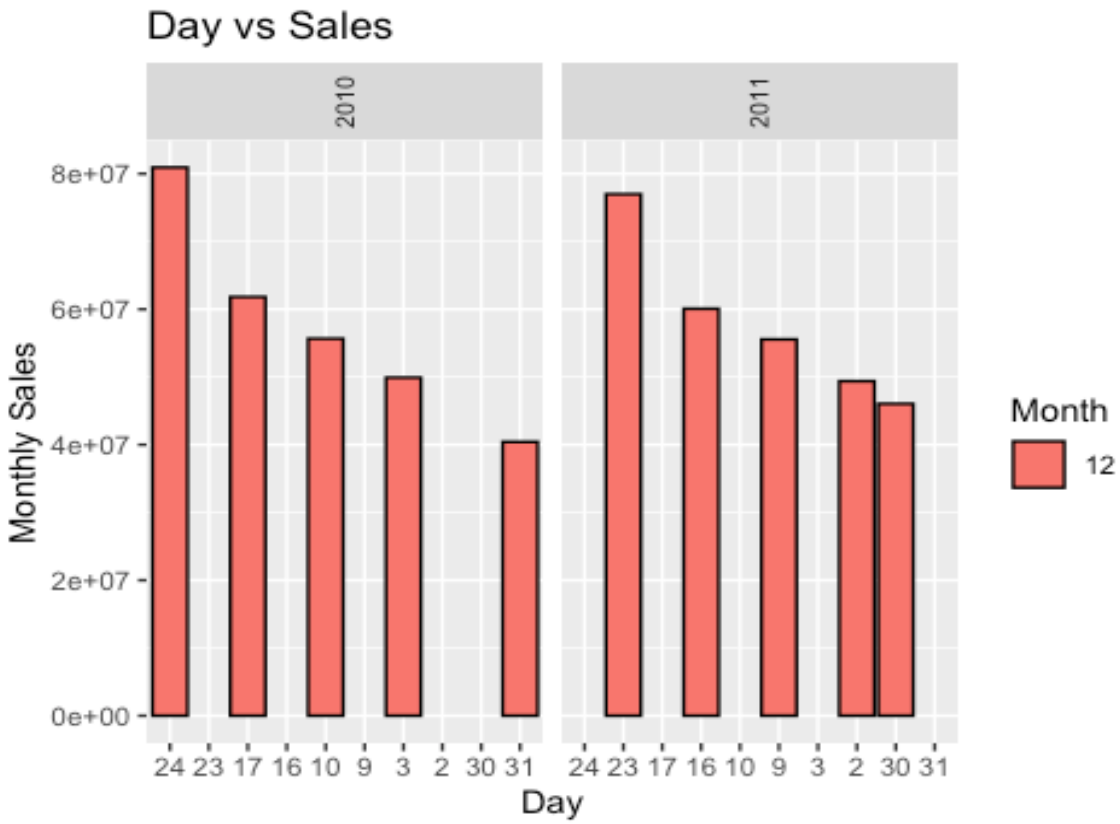


- Total Sales per day distribution as a box plot:
  - December 2010 has a day with the largest volume of sales.
  - The top outliers are explained by top 10 days list mentioned before.



- Deep dive on what is causing the spike in Dec:
  - Are specific Stores causing the outliers for Dec?
  - Check top 5 stores.
  - There is no significant spike in any 1 Store.

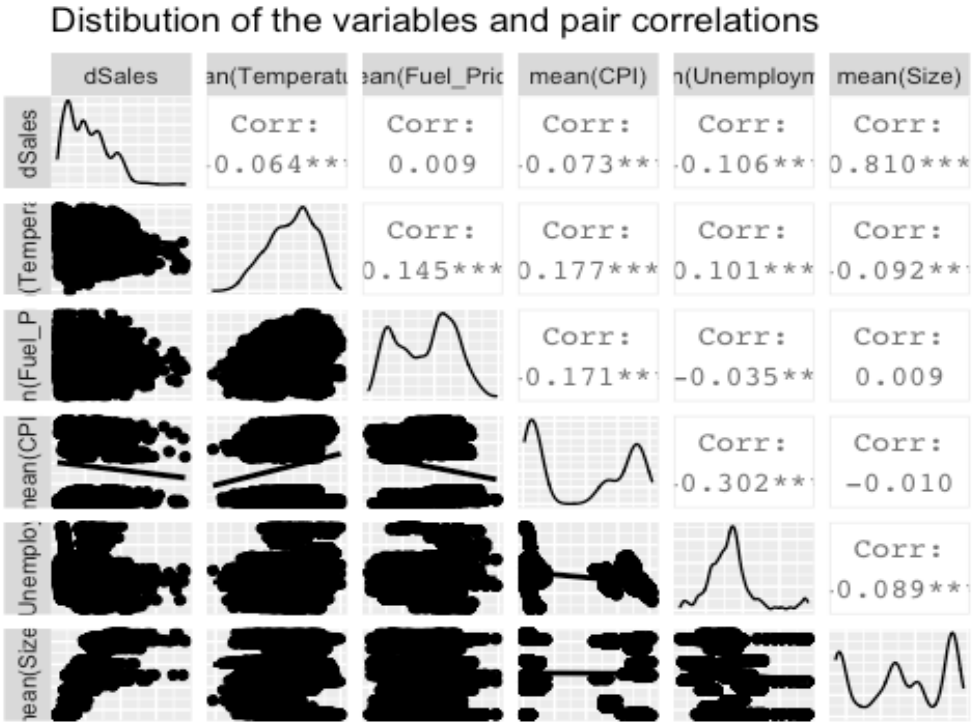
- As we can see below, there is a clear spike on 24th 2010 and 23rd 2011 in the month of December. And we can conclude that this was not due to a spike in any 1 Store in particular.



Correlation and linear regression

Correlation

- Approach:
  - We need to explain 'Sales'. Each line item of the dataset is for a certain 'Dept' which is part of a particular 'Store' on a particular 'Date'.
  - Let us define 'Continuous_variables' as those which vary continuously: Temp, Fuel_price, CPI, Unemployment, Promotions.
  - Let us define the other columns as 'Categorical_variables' such as 'Store', 'Holiday', 'Month' and 'Type', which behave more as categories or factors.
  - Lastly let us look at 'Promotions' separately.
- Let us group Sales by Store and Date since these Continuous_variables refer to a Store for the week. This means we will lose the 'Dept' level variation in Sales. We will consider only the total of all Depts for a Day which is the total Sales for that Store on that Date(week).
  - Temperature: Unique Temperature value affects at least 1 store and up to 5 stores on a given week
  - Fuel_Price: Fuel Price affects at least 1 Store and up to 16 Stores on a given day.
  - CPI: CPI affects at least 1 Store and up to 11 Stores on a given day.
  - Unemployment: Unemployment affects at least 1 Store and up to 6 Stores on a given day.
- Check pair correlations of the grouped data. 'Mean' is just for grouping and not actual.



- The correlation between Sales and the other Continuous_variables gets better after grouping:
  - Sales-Temp Ungrouped -0.002 Grouped -0.064
  - Sales-FP Ungrouped -0.000 Grouped 0.009
  - Sales-CPI Ungrouped -0.021 Grouped-0.073
  - Sales-UE Ungrouped -0.026 Grouped-0.106
  - Sales-Size Ungrouped 0.244 Grouped 0.810
- There is a strong positive correlation of 'Sales' with the 'Size' of the stores.This means that the greater the size of the stores, the greater is the sales. For the other variables however, the correlation is poor.

Linear Regression-Categorical_variables

- Let us check 'Categorical_variables such as 'Store', 'Holiday', 'Month' and 'Type' after grouping by the Date and Store.
  - Correlation plot for categorical_variables are difficult to interpret. So let us plug into the linear model to judge which of these are relevant.
- Use linear model to check if the categorical variables are salient with respect to Sales.
  - In linear model remember that with more independant variables R square 'R2' will increase but actually may not be significant. So pay attention to adjusted R2 and check when it begins to drop.
- Adjusted R squares for 'Categorical_variables':
  - Holiday 0.12%
  - Type 36.43%
  - Type*Holiday 36.56%
  - Store 91.68%
  - Store+Holiday+Type 91.82%
  - Holiday*Store 91.83%
  - Store + Month 93.57%
  - Store+Holiday+Type+Month 93.58%
  - Store+Month+Type+(Holiday*Store) 93.6%
  - Store+Month+(Holiday*Store) 93.6%
- Certain stores may have more Sales influenced by the fact that it is a holiday.
- We cannot reject null hypothesis for Store 6. But other Stores have statistically significant p-values.
- The p-values for 'Month' is also statistically significant. But for 'Holiday' the p-values are insignificant (large). Still, the adjusted R2 increases to 93.6% by including 'Month' and 'Holiday' along with 'Store'.

Linear Regression-Continuous_variables

- Let us check 'Categorical_variables' such as 'Temp', 'FP', 'CPI', 'UE', and 'Size' after grouping by the Date and Store.
- Adjusted R squares for 'Continuous_variables':

  - Temp+FP+CPI+UE 2.37%
  - Size 65.68%
  - Size+FP 65.68%
  - Size+Temp 65.69%
  - Size+UE 65.79%
  - Size+CPI 66.09%
  - Size+CPI+Temp 66.14%
  - Size+CPI+UE 66.41%
  - Size+CPI+UE+FP 66.42%
  - Size+CPI+UE+Temp 66.51%
  - Size+CPI+UE+Temp+FP 66.55%
- The p-values are statistically significant (<0.05) for all the variables: Size,CPI,UE,Temp,FP.

Linear Regression-Promotions and other combinations

- These Promotion columns are unique because they contain a numeric magnitude (amount of discount), so they behave as a continuous variable. But since they also contain missing values, they can be used as a binary factors: 'with promotion' or 'without promotion'. For this analysis we are only going to use the 'Promotions' columns as Continuous (numeric).

- Regression using the 'Promotions' values only.

  - With only 'Promotions' columns individually the adjusted R sq is 11.3%. The columns behave as continuous variables here.
  - Considered as total promotions,which is the sum of the individual promotions for a row, the adjusted R sq is 8.3%
  - The p-values for all the individual Promotions are statistically significant except for Promotion4. For this there is a 42.7% chance that the values are due to randomness.
- Use Promotion values in combination with some other variable to see if model performance improves:

- Adjusted R2 ('BCat' is best of categoricals, 'BCon' is best of continuous, and 'Promos' is all promotions):

  - Promos 11.3%
  - Promos*Holiday :13.96% ; Promos + Holiday 11.68%
  - Promos*Type 42.07% ; Promos + Type 42.18%
  - Promos*Size 56.09% ; Promos + Size 56.19%
  - Promos*Store 88.76% ; Promos + Store 87.74%
  - Store*Promos + (best combination of categorical variables) 93.14%
    - Best categoricals: Store+Month+(Holiday*Store) 93.6%
    - BCat+Prom1 93.68

        - BCat+Prom3 94.07
        - BCat+Prom5 93.67
        - BCat +Promotion1*Promotion3 94.12%
    - Store*Promos + (best combination of continuous variables) 89.37%
        - Best continuous: Size+CPI+UE+Temp+FP 66.55%
        - BCon + BCat 93.7%
        - BCon + BCat + Prom3 94.09%
        - BCon + BCat + Prom1*Prom3 94.14%
        - BCon + BCat + Prom3*Prom5 94.15%

Conclusion

- As we have seen above, there are various models which behave differently. The models which have the best adjusted R square values explain most of the variance. But this sometimes comes at the cost of low p-value for the variables.
    - The model with the highest adjusted R square contains: Store + Month + (Holiday * Store) + Size + CPI + UE + Temp + FP + (Promotion3 * Promotion5) which explains 94.15% of the variance in Sales. However the p-values for many of the variables are statistically insignificant. So this model may not be the best option.
    - The model with 'Store' number and 'Month' predictors, offers the best p-values and has an adjusted R square of 93.57%. This means that the 'Month' and which 'Store' explains the Sales of each Store the best.
    - Size appears to be the best continuous_variable predictor for success at a particular Store and explains 65.68% of the variance in Store Sales data. If we add the other continuous variables: Size + CPI + UE + Temp + FP then we add an additional 1% of explained variance for 66.55%. All these are parameters are statistically significant with low p-values.
        - So we can infer that Size influences the Sales of a individual Store the most. And based on which Store and month, we can predict the volume of Sales.
- Top performances from entire data set:
    - The top 6 Stores are: 20,4,14,13,2 and 10
    - The top 5 Depts are: 92,95,38,72,90
    - The top Dates are Nov 25th,26th and Dec 23rd,24th