Assignment7

Lokesh Arora, Ankita Shinde

10/29/2020

Github Link: https://github.com/ankita1598/Walmart

```
#Loading Packages
library(mvtnorm)
```

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

```
library(dplyr)
```

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

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

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

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

```
library(psych)
```

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

```
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
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
## 
## Attaching package: 'plyr'
```

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

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

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

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

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

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

```
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
#Loading Dataset
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
dataset = fastDummies::dummy_cols(dataset, select_columns = "Type")
dataset$IsHoliday[dataset$isHoliday == "False"] = 0
dataset$IsHoliday[dataset$isHoliday == "True"] = 1
dataset$Dept = as.numeric(as.factor(dataset$Dept))
dataset$Store = as.numeric(as.factor(dataset$Store))
features = c("weeklySales","Store","Dept","IsHoliday","Type_A","Type_B","Type_C","Size","Temperature","Fuel_Price","MarkDown
1","MarkDown2","MarkDown3","MarkDown4","MarkDown5","CPI","Unemployment","Year","Month","Day")
dataset = select(dataset,features)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(features)` instead of `features` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
head(dataset)
```

```
##   weeklySales Store Dept IsHoliday Type_A Type_B Type_C   Size Temperature
## 1    24924.50     1    1         0      1      0      0 151315       42.31
## 2    46039.49     1    1         1      1      0      0 151315       38.51
## 3    41595.55     1    1         0      1      0      0 151315       39.93
## 4    19403.54     1    1         0      1      0      0 151315       46.63
## 5    21827.90     1    1         0      1      0      0 151315       46.50
## 6    21043.39     1    1         0      1      0      0 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         0 211.3501
## 6      2.667         0         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     20
```

```
names(dataset)
```

```
##  [1] "weeklySales"  "Store"        "Dept"         "IsHoliday"    "Type_A"      
##  [6] "Type_B"       "Type_C"       "Size"         "Temperature"  "Fuel_Price"  
## [11] "MarkDown1"    "MarkDown2"    "MarkDown3"    "MarkDown4"    "MarkDown5"   
## [16] "CPI"          "Unemployment" "Year"         "Month"        "Day"
```

Since our depedent variable is continous so we check for min, max and distribution of data points to figure out if can form two catagories to run logistic regression on our dataset.

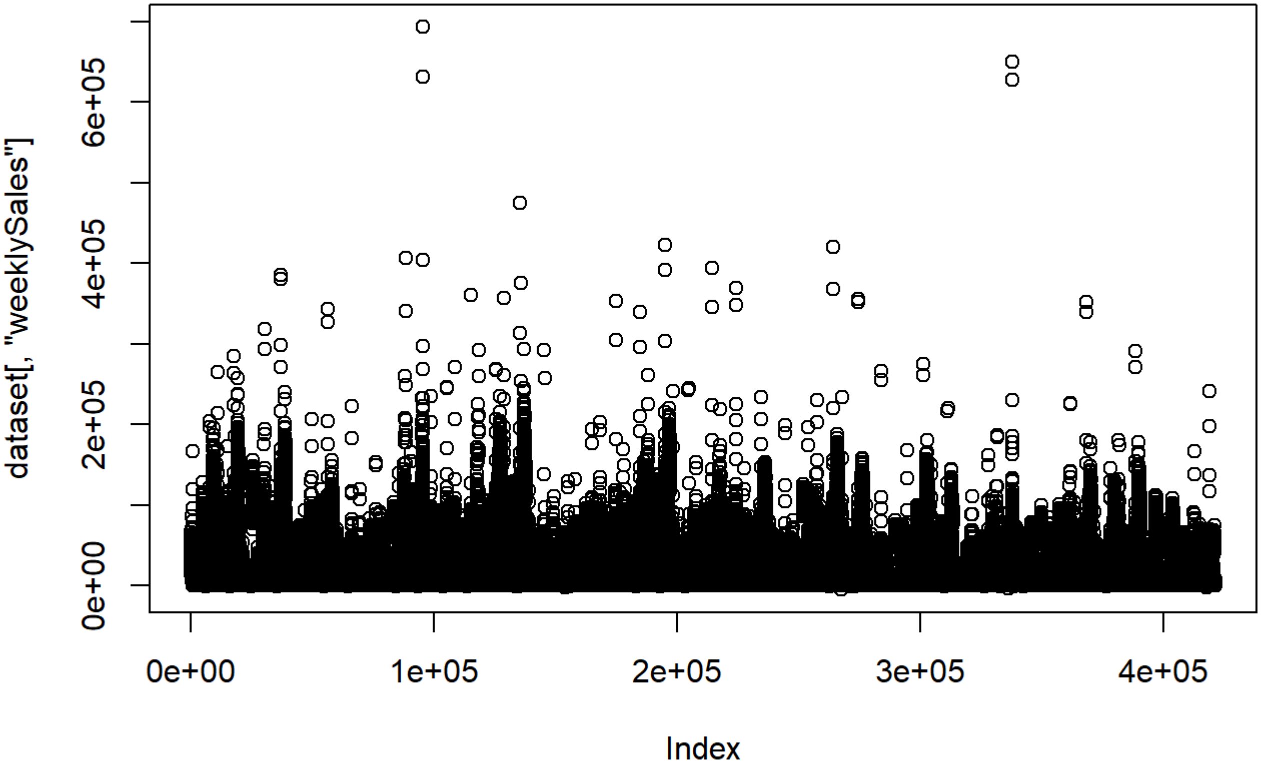
```
min(dataset[,"weeklySales"])
```

```
## [1] -4988.94
```

```
max(dataset[,"weeklySales"])
```

```
## [1] 693099.4
```

```
plot(dataset[,"weeklySales"])
```



If we look at the distriburtion of data it doesn’t seem to have two clusters so it is not possible to divide the dependent variable into two categories and run logistic regression.