**#CORRELATION BETWEEN SALES AND HOLIDAY**

**CODE:**

cor.test(my\_data$Weekly\_Sales,my\_data$IsHoliday,method="pearson")

**OUTPUT:**

Pearson's product-moment correlation

data: my\_data$Weekly\_Sales and my\_data$IsHoliday

t = 8.2948, df = 421570, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.009755942 0.015792266

sample estimates:

cor

0.01277422

The correlation between sales and holiday is significant because the p-value Is less than 5% (2.2e-16)

**## BINNING OF DATA (categorizing continuous values into smaller bins)**

**CODE:**

install.packages("classInt")

library(classInt)

bin\_data <- my\_data$Weekly\_Sales

classIntervals(bin\_data,5,style="equal") ##equal width binning

classIntervals(bin\_data,5,style="quantile") ##equal frequency binning

**OUTPUT:**

style: equal

[-4988.94,134628.7) [134628.7,274246.4) [274246.4,413864) [413864,553481.7) [553481.7,693099.4]

419694 1832 37 3 4

style: quantile

[-4988.94,1340.98) [1340.98,4887.96) [4887.96,11274.63) [11274.63,25217.61) [25217.61,693099.4]

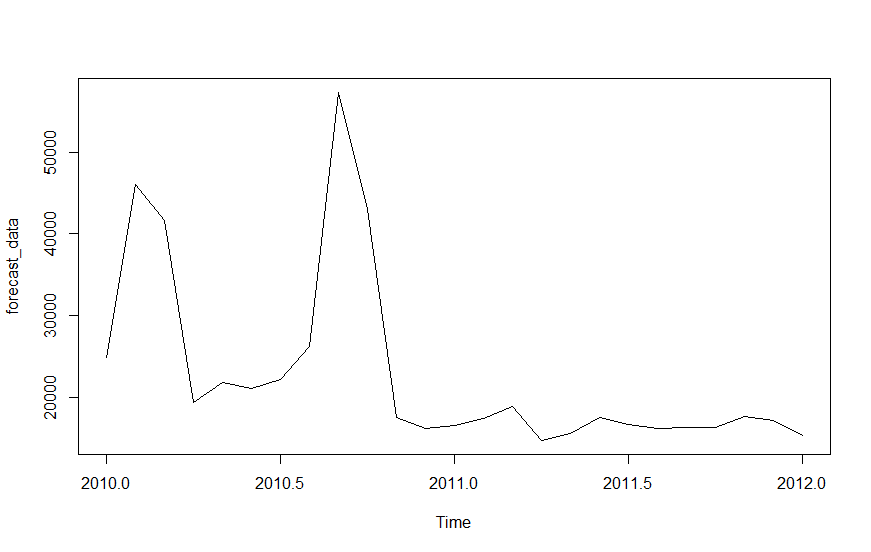
84314 84313 84315 84314 84314

**# TIME SERIES**

forecast\_data <- ts(my\_data$Weekly\_Sales, start=2010, end=2012,frequency=12)

plot(forecast\_data)

**OUTPUT:**



**# Best Performing Stores extraction using DPLYR Grouping**

**Code:**

library(dplyr)

aggregate(my\_data$Weekly\_Sales, by=list(Store=my\_data$Store), FUN=sum)

**Result:**

Store x

1 1 222402809

2 2 275382441

3 3 57586735

4 4 299543953

5 5 45475689

6 6 223756131

7 7 81598275

8 8 129951181

9 9 77789219

10 10 271617714

11 11 193962787

12 12 144287230

13 13 286517704

14 14 288999911

15 15 89133684

16 16 74252425

17 17 127782139

18 18 155114734

19 19 206634862

20 20 301397792 ##Store 20 is the best performing store

21 21 108117879

22 22 147075649

23 23 198750618

24 24 194016021

25 25 101061179

26 26 143416394

27 27 253855917

28 28 189263681

29 29 77141554

30 30 62716885

31 31 199613906

32 32 166819246

33 33 37160222

34 34 138249763

35 35 131520672

36 36 53412215

37 37 74202740

38 38 55159626

39 39 207445542

40 40 137870310

41 41 181341935

42 42 79565752

43 43 90565435

44 44 43293088

45 45 112395341

**# Correlation Matrix**

subset2 <- subset(my\_data, select= c("Store","Dept","Weekly\_Sales","IsHoliday"))

res <- cor(subset2)

head(res)

**Result:**

Store Dept Weekly\_Sales IsHoliday

Store 1.000000000 0.0240036675 -0.08519499 -0.0005479880

Dept 0.024003668 1.0000000000 0.14803206 0.0009161019

Weekly\_Sales -0.085194992 0.1480320600 1.00000000 0.0127742204

IsHoliday -0.000547988 0.0009161019 0.01277422 1.0000000000

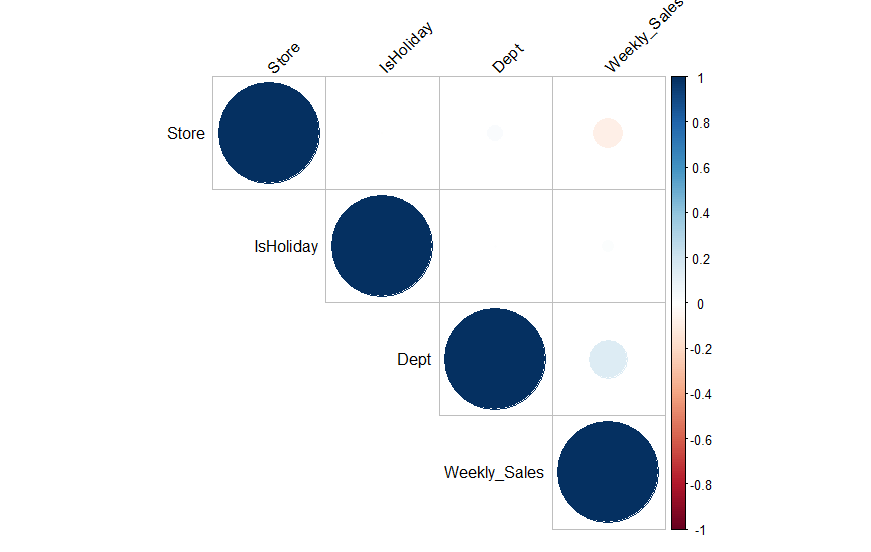
**There is positive correlation between department and weekly sales**

**# Correlation Plot**

library(corrplot)

corrplot(res, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)

**Result:**

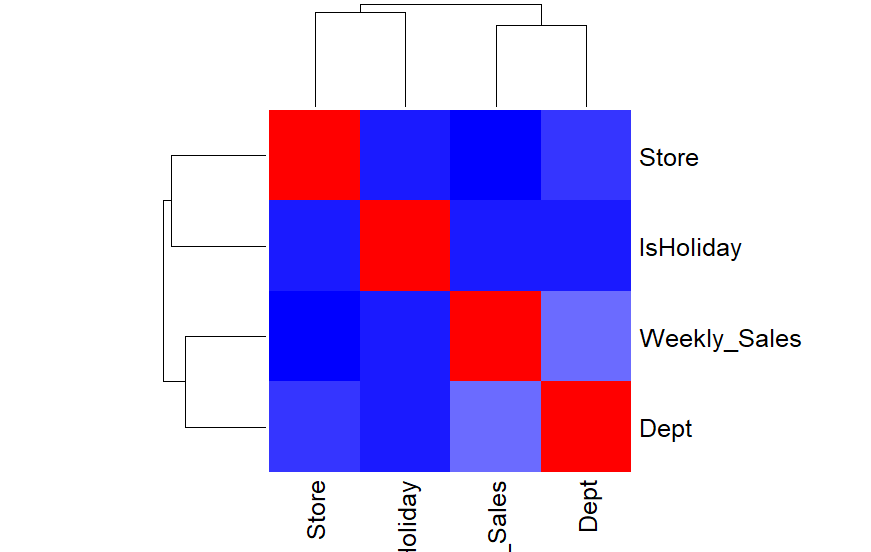


**Positive correlation between department and weekly sales**

**#Heatmap**

col<- colorRampPalette(c("blue", "white", "red"))(20)

heatmap(x = res, col = col, symm = TRUE )



**#Sales Prediction using: Multiple Linear Regression**

input<-my\_data[,c("Weekly\_Sales","Store","Dept","IsHoliday")]

head(input)

model1 <- lm(Weekly\_Sales~Store+Dept+IsHoliday, data = input)

print(model1)

summary(model1)

**Result:**

> print(model1)

Call:

lm(formula = Weekly\_Sales ~ Store + Dept + IsHoliday, data = input)

Coefficients:

(Intercept) Store Dept IsHoliday

14454.3 -157.7 111.8 1117.8

> summary(model1)

Call:

lm(formula = Weekly\_Sales ~ Store + Dept + IsHoliday, data = input)

Residuals:

Min 1Q Median 3Q Max

-26486 -13184 -7721 5148 671052

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 14454.297 84.945 170.162 <2e-16 \*\*\*

Store -157.726 2.695 -58.518 <2e-16 \*\*\*

Dept 111.837 1.130 98.956 <2e-16 \*\*\*

IsHoliday 1117.842 134.706 8.298 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22370 on 421566 degrees of freedom

Multiple R-squared: 0.02995, Adjusted R-squared: 0.02995

F-statistic: 4339 on 3 and 421566 DF, p-value: < 2.2e-16

**#Plot to check weekly\_sales VS store, Dept, Date and Is\_Holiday**

**CODE**

plot(my\_data$Weekly\_Sales, my\_data$Store)

plot(my\_data$Weekly\_Sales, my\_data$Dept)

