**Walmart sales**

**EDA**

> # Importing csv file

> T=read.csv(file.choose())

> View(T)

> summary(T)

Store Date

Min. : 1 Length:6435

1st Qu.:12 Class :character

Median :23 Mode :character

Mean :23

3rd Qu.:34

Max. :45

Weekly\_Sales Holiday\_Flag

Min. : 209986 Min. :0.00000

1st Qu.: 553350 1st Qu.:0.00000

Median : 960746 Median :0.00000

Mean :1046965 Mean :0.06993

3rd Qu.:1420159 3rd Qu.:0.00000

Max. :3818686 Max. :1.00000

Temperature Fuel\_Price

Min. : -2.06 Min. :2.472

1st Qu.: 47.46 1st Qu.:2.933

Median : 62.67 Median :3.445

Mean : 60.66 Mean :3.359

3rd Qu.: 74.94 3rd Qu.:3.735

Max. :100.14 Max. :4.468

CPI Unemployment

Min. :126.1 Min. : 3.879

1st Qu.:131.7 1st Qu.: 6.891

Median :182.6 Median : 7.874

Mean :171.6 Mean : 7.999

3rd Qu.:212.7 3rd Qu.: 8.622

Max. :227.2 Max. :14.313

> length(T)

[1] 8

>

> library(ggplot2)

> library(dplyr)

> library(tidyr)

> library(lubridate)

>

> str(T) # Structure of data

'data.frame': 6435 obs. of 8 variables:

$ Store : int 1 1 1 1 1 1 1 1 1 1 ...

$ Date : chr "05-02-2010" "12-02-2010" "19-02-2010" "26-02-2010" ...

$ Weekly\_Sales: num 1643691 1641957 1611968 1409728 1554807 ...

$ Holiday\_Flag: int 0 1 0 0 0 0 0 0 0 0 ...

$ Temperature : num 42.3 38.5 39.9 46.6 46.5 ...

$ Fuel\_Price : num 2.57 2.55 2.51 2.56 2.62 ...

$ CPI : num 211 211 211 211 211 ...

$ Unemployment: num 8.11 8.11 8.11 8.11 8.11 ...

>

> colSums(is.na(T)) # For checking missing values

Store Date Weekly\_Sales

0 0 0

Holiday\_Flag Temperature Fuel\_Price

0 0 0

CPI Unemployment

0 0

> # No missing values found.

> #boxplot and histograms for visualizing data

>

> str(T[,c(1,5,6,7)])

'data.frame': 6435 obs. of 4 variables:

$ Store : int 1 1 1 1 1 1 1 1 1 1 ...

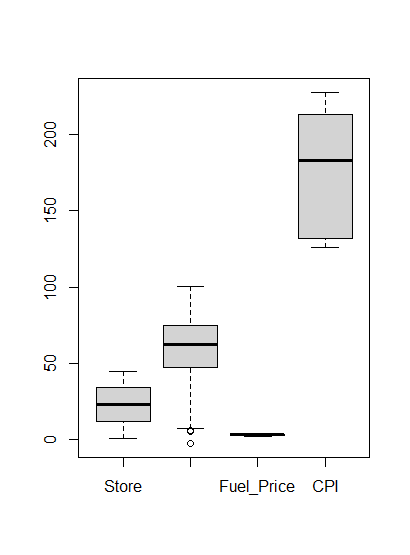
$ Temperature: num 42.3 38.5 39.9 46.6 46.5 ...

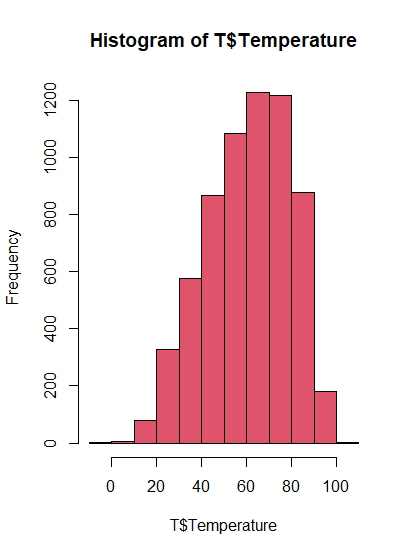
$ Fuel\_Price : num 2.57 2.55 2.51 2.56 2.62 ...

$ CPI : num 211 211 211 211 211 ...

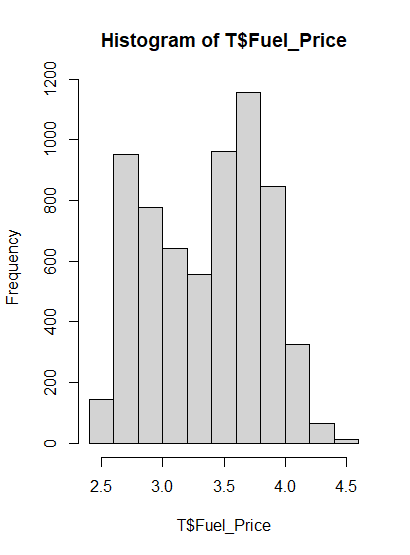
>

> boxplot(T[,c(1,5,6,7)])

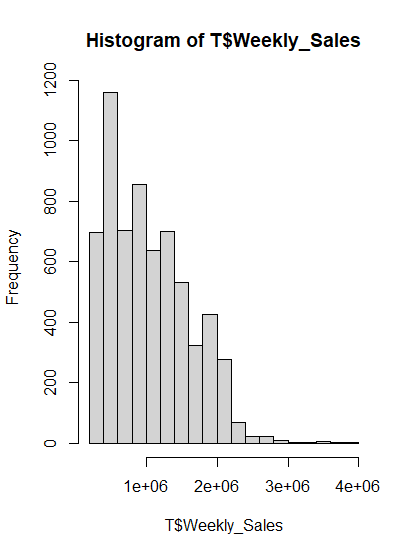


> hist(T$Temperature,col=2) 

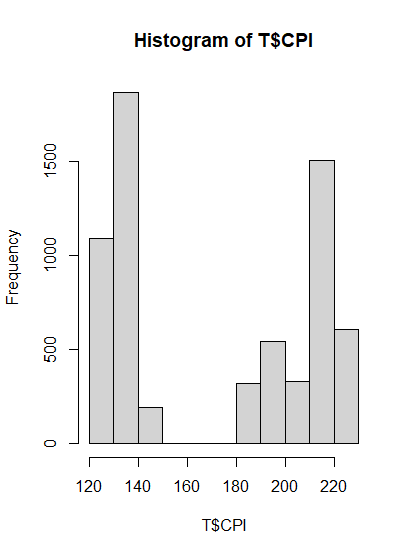
> hist(T$Fuel\_Price)



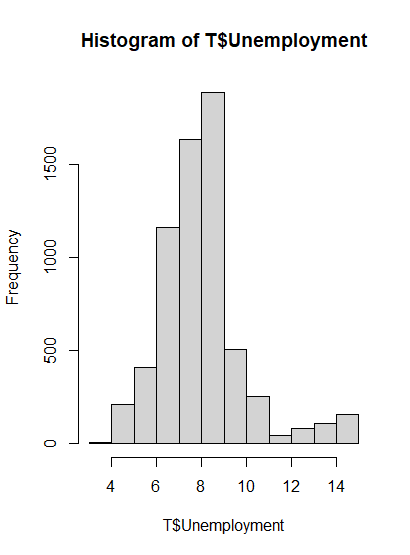
> hist(T$Weekly\_Sales)



> hist(T$CPI )



> hist(T$Unemployment)



>

# Finding correlation

> t=T[,c(3,5,6,7,8)] # taking only numeric variables

> cor\_matrix=cor(t) # Correlation matrix

> cor\_matrix

Weekly\_Sales Temperature

Weekly\_Sales 1.000000000 -0.06381001

Temperature -0.063810013 1.00000000

Fuel\_Price 0.009463786 0.14498181

CPI -0.072634162 0.17688768

Unemployment -0.106176090 0.10115786

Fuel\_Price CPI

Weekly\_Sales 0.009463786 -0.07263416

Temperature 0.144981806 0.17688768

Fuel\_Price 1.000000000 -0.17064180

CPI -0.170641795 1.00000000

Unemployment -0.034683745 -0.30202006

Unemployment

Weekly\_Sales -0.10617609

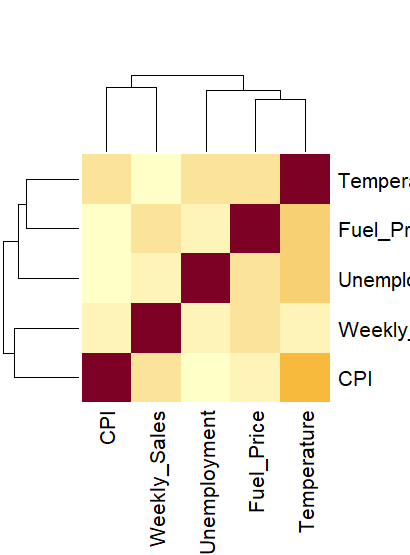
Temperature 0.10115786

Fuel\_Price -0.03468374

CPI -0.30202006

Unemployment 1.00000000

> heatmap(cor\_matrix)



> # As the correlation between the variables is almost zero, there is no relation between Weekly sales and other variables.

# Finding which store has the maximum sales...

> T=read.csv(file.choose())

Error in file.choose() : file choice cancelled

> T<- T %>%

+ mutate(Date = dmy(Date))

> sum(is.na(T))

[1] 0

>

> maximum\_sales <- T %>%

+ group\_by(Store) %>%

+ summarize(totalsales=sum(Weekly\_Sales))

> head(arrange(maximum\_sales,desc(totalsales)))

# A tibble: 6 × 2

Store totalsales

<int> <dbl>

1 20 301397792.

2 4 299543953.

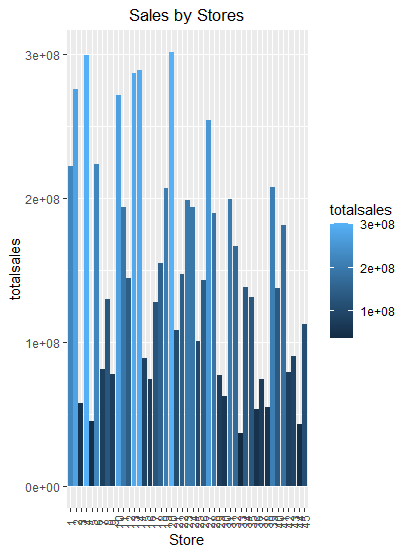
3 14 288999911.

4 13 286517704.

5 2 275382441.

6 10 271617714.

> ggplot(maximum\_sales,aes(x=factor(Store),y=totalsales,fill=totalsales)) + geom\_col()+theme(axis.text.x = element\_text(angle = 90, vjust = 0.5),plot.title = element\_text(hjust = 0.5))+xlab("Store")+ggtitle("Sales by Stores")

> 

> # store 20 has the maximum sales.

>

>

> T %>%

+ group\_by(Store) %>%

+ summarize(sd= sd(Weekly\_Sales),mean= mean(Weekly\_Sales),CV= sd/mean) %>%

+ arrange(desc(sd))

# A tibble: 45 × 4

Store sd mean CV

<int> <dbl> <dbl> <dbl>

1 14 317570. 2020978. 0.157

2 10 302262. 1899425. 0.159

3 20 275901. 2107677. 0.131

4 4 266201. 2094713. 0.127

5 13 265507. 2003620. 0.133

6 23 249788. 1389864. 0.180

7 27 239930. 1775216. 0.135

8 2 237684. 1925751. 0.123

9 39 217466. 1450668. 0.150

10 6 212526. 1564728. 0.136

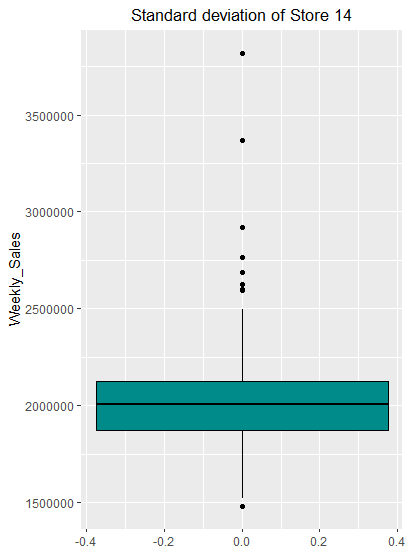
# ℹ 35 more rows

# ℹ Use `print(n = ...)` to see more rows

> sales\_data %>%

+ filter(Store == 14) %>%

+ ggplot(aes(y=Weekly\_Sales)) + geom\_boxplot(colour="black",fill="darkcyan")+ ggtitle("Standard deviation of Store 14") +theme(plot.title = element\_text(hjust = 0.5))



> # sales of store 14 varies a lot than other stores.

> monthly\_data <- T %>%

+ mutate(Month = month(Date,label = TRUE),Year = year(Date),Semester = as.character(semester(Date,with\_year = TRUE)))

>

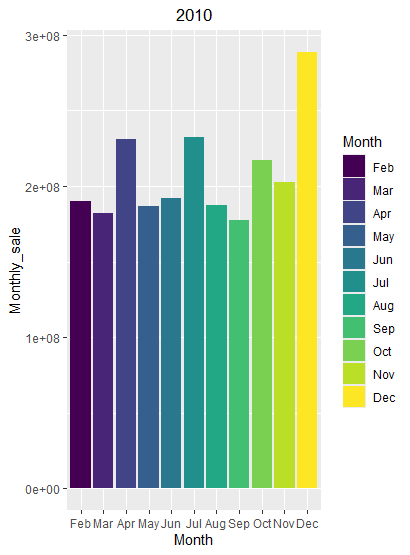
# Monthly Graph of 2010

> monthly\_data %>%

+ filter(Year == 2010) %>%

+ group\_by(Month) %>%

+ summarise(Monthly\_sale = sum(Weekly\_Sales)) %>%

+ ggplot(aes(x=Month,y=Monthly\_sale,fill=Month)) + geom\_col() + ggtitle("2010") +theme(plot.title = element\_text(hjust = 0.5)) 

>

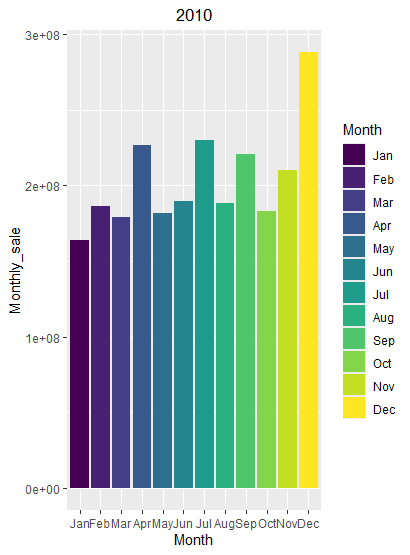
> # Monthly Graph of 2011

> monthly\_data %>%

+ filter(Year == 2011) %>%

+ group\_by(Month) %>%

+ summarise(Monthly\_sale = sum(Weekly\_Sales)) %>%

+ ggplot(aes(x=Month,y=Monthly\_sale,fill=Month)) + geom\_col() + ggtitle("2010") +theme(plot.title = element\_text(hjust = 0.5)) 

# Monthly Graph of 2012

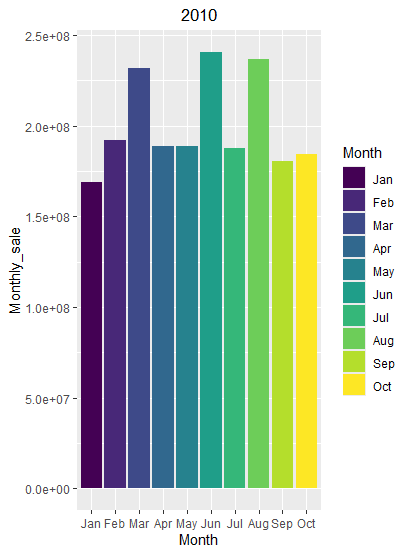
> monthly\_data %>%

+ filter(Year == 2012) %>%

+ group\_by(Month) %>%

+ summarise(Monthly\_sale = sum(Weekly\_Sales)) %>%

+ ggplot(aes(x=Month,y=Monthly\_sale,fill=Month)) + geom\_col() + ggtitle("2010") +theme(plot.title = element\_text(hjust = 0.5))



> # Conclusion: Trends are changed. Sales is increased in April, July and December.

> #########################################################