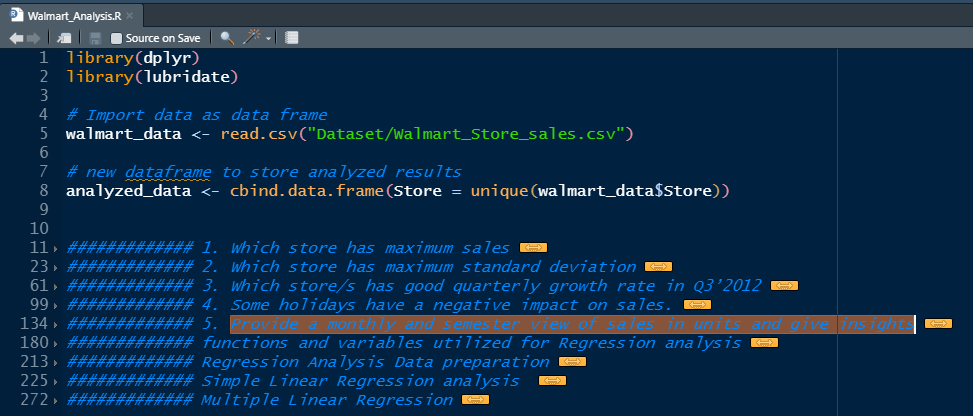
Code glimpse:



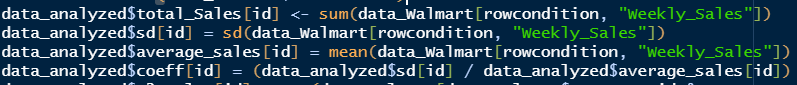
**Prerequisite libraries: dplyr , lubridate**

Imp Variables:

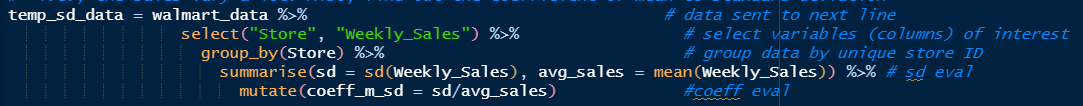
1. walmart\_data -> reads walmart csv data. Over the code this is populated with new variables (year, month…) as per questionnaire
2. analyzed\_data -> contains unique store ID’s as first column and relevant columns populated and utilized throughout code.

Code is primarily done using dplyr package since it has significant ease of code over base R ->

Base R Code:



dplyr code:



Document Design Template:

|  |  |
| --- | --- |
| Topic | Highlighter |
| Code and snapshot of code |  |
| Answer to question | |  |  | | --- | --- | |  |  | |  |  | |
| Insights | |  |  | | --- | --- | |  |  | |  |  | |

Note: Code mentioned in document may not run as is and so check source code to properly run. Only relevant code snippets are shown.

1. Which store has maximum sales

|  |  |
| --- | --- |
| Store ID | Max SD |
| 20 | **301397792.46** |

total\_sales\_data = walmart\_data %>% # data sent to next line

select("Store", "Weekly\_Sales") %>% # select variables (columns) of interest

group\_by(Store) %>% # group data by unique store ID

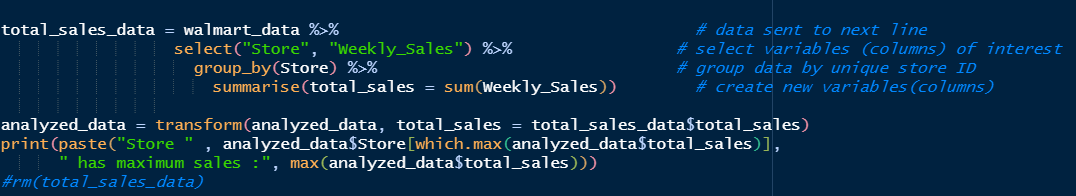
summarise(total\_sales = sum(Weekly\_Sales)) # create new variables(columns)

analyzed\_data = transform(analyzed\_data, total\_sales = total\_sales\_data$total\_sales)

print(paste("Store " , analyzed\_data$Store[which.max(analyzed\_data$total\_sales)],

" has maximum sales :", max(analyzed\_data$total\_sales)))

#rm(total\_sales\_data)



|  |  |
| --- | --- |
| Each store with total sales |  |

1. Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation

|  |  |
| --- | --- |
| Store ID | Max SD |
| 14 | 317569.949475508 |

temp\_sd\_data = walmart\_data %>% # data sent to next line

select("Store", "Weekly\_Sales") %>% # select variables (columns) of interest

group\_by(Store) %>% # group data by unique store ID

summarise(sd = sd(Weekly\_Sales), avg\_sales = mean(Weekly\_Sales)) %>% # sd eval

mutate(coeff\_m\_sd = sd/avg\_sales) #coeff eval

#populate our analyzed\_data

analyzed\_data = transform(analyzed\_data, sd = temp\_sd\_data$sd,

avg\_sales = temp\_sd\_data$avg\_sales,

coeff\_m\_sd = temp\_sd\_data$coeff\_m\_sd)

#analyze the coefficient against Ideal\_ratio = 10 % (0.1)

tmp\_sd\_analysis = analyzed\_data %>%

select(c("Store", "coeff\_m\_sd")) %>% # prepare data

mutate(ratio\_percentage = round((coeff\_m\_sd \* 100) , digits = 0), # in percentage

dev\_from\_IdelRatio = ifelse (between(ratio\_percentage, 5, 15), "Good",

ifelse (between(ratio\_percentage, 0, 5) | between(ratio\_percentage, 15, 20),

"Acceptable", "Beyond"))

) # deviations from Ideal ratio (10%)

analysis = tmp\_sd\_analysis %>% count(dev\_from\_IdelRatio)

print(analysis)

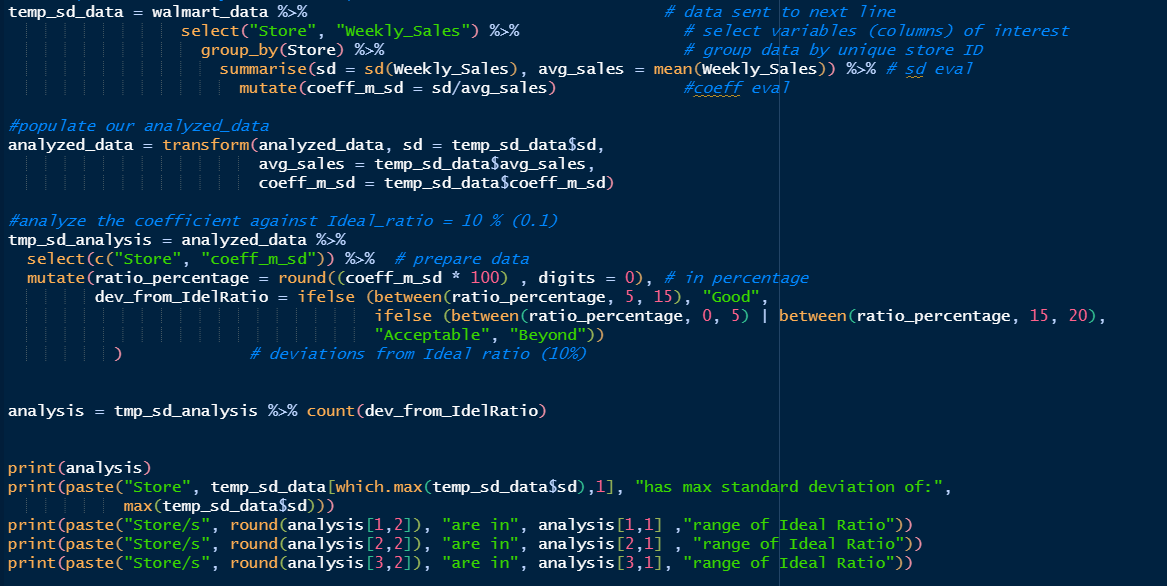
print(paste("Store", temp\_sd\_data[which.max(temp\_sd\_data$sd),1], "has max standard deviation of:",

max(temp\_sd\_data$sd)))

print(paste("Store/s", round(analysis[1,2]), "are in", analysis[1,1] ,"range of Ideal Ratio"))

print(paste("Store/s", round(analysis[2,2]), "are in", analysis[2,1] , "range of Ideal Ratio"))

print(paste("Store/s", round(analysis[3,2]), "are in", analysis[3,1], "range of Ideal Ratio"))



|  |  |
| --- | --- |
| sd, mean and coeff for all stores |  |
| Analysis table with coeff ratios as percentage.  Deviations in following logic -> |  |
| Insight:   1. Store/s 14 are in Acceptable range of Ideal Ratio 2. Store/s 1 are in Beyond range of Ideal Ratio 3. Store/s 30 are in Good range of Ideal Ratio |  |

1. Which store/s has good quarterly growth rate in Q3’2012

|  |  |
| --- | --- |
| Store ID | Q3-Growth |
| 7 | 13% |
| 16 | 8% |
| 26 | 4% |
| 35 | 4% |

#' Check whether date is in proper format... if not update in format

if (!is.Date(mode(walmart\_data$Date))) walmart\_data$Date = lubridate::dmy(walmart\_data[,"Date"])

#' add variables(columns) year, month, quarter in walmart\_data

walmart\_data = walmart\_data %>%

mutate(quarter = quarter(Date),

month = month(Date, label = TRUE, abbr = TRUE),

semester = semester(Date),

year = year(Date),

quarter = factor(quarter, levels = c(1:4), labels= c("Q1", "Q2", "Q3", "Q4")),

semester = factor(semester, levels = c(1:2), labels= c("S1", "S2")))

# compute (2 quarter data (Q2 2012,Q3 2012) for Weekly\_Sales) from walmart\_data

tmp\_quarter\_data = walmart\_data %>%

subset((quarter == "Q2" | quarter == "Q3") & year == 2012, select = c("Store", "quarter", "Weekly\_Sales")) %>%

group\_by(Store, quarter) %>%

summarise(sales = sum(Weekly\_Sales))

# append Q2,Q3 to analyzed\_data

analyzed\_data = analyzed\_data %>%

mutate(Q2\_2012 = pull(subset(tmp\_quarter\_data, tmp\_quarter\_data$quarter == "Q2", select = "sales")), #from tibble created by subset, pull vector

Q3\_2012 = pull(subset(tmp\_quarter\_data, tmp\_quarter\_data$quarter == "Q3", select = "sales")), #from tibble created by subset, pull vector

Q3\_2012\_growth = (( (Q3\_2012 - Q2\_2012) / Q2\_2012) \* 100) )

#analyze the positive growths

tmp\_stores\_good\_growth = analyzed\_data %>%

subset(Q3\_2012\_growth > 0, select = c("Store", "Q3\_2012\_growth")) %>%

arrange(desc(Q3\_2012\_growth)) %>%

slice(1:4) %>%

mutate(Q3\_2012\_grate = paste0(round(Q3\_2012\_growth, digits = 0), "%"))

print(tmp\_stores\_good\_growth)



|  |  |
| --- | --- |
| analyzed\_data table already has previous steps computed data |  |
| Take total sales Q2 2012 and Q3 2012 for all stores |  |
| Populate in analyzed\_data table, these quarter information store wise and compute Q3\_2012\_growth (Q3 – Q2) |  |
| Analyze top 4 store performers |  |

1. Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together.

|  |
| --- |
| Holidays with high sales |
| 2010-02-12 |
| 2010-11-26 |
| 2011-02-11 |
| 2011-11-25 |
| 2012-02-10 |
| 2012-09-07 |

#' # list of holidays

temp\_holiday = c("12-Feb-10", "11-Feb-11", "10-Feb-12", "8-Feb-13", "10-Sep-10", "9-Sep-11",

"7-Sep-12", "6-Sep-13", "26-Nov-10", "26-Nov-10", "23-Nov-12", "29-Nov-13",

"31-Dec-10", "30-Dec-11", "28-Dec-12", "27-Dec-13")

# holidays in Date format

temp\_holiday\_date = lubridate::dmy(temp\_holiday)

# avg of all sales non-holiday

temp\_s\_nh = walmart\_data %>%

subset(Holiday\_Flag == 0, select = c("Weekly\_Sales")) %>%

summarise(sales = mean(Weekly\_Sales))

# analyze avg of all sales on holiday\_date

temp\_s\_h = walmart\_data %>%

subset((Holiday\_Flag == 1 & any(Date %in% temp\_holiday\_date)),

select = c("Date", "Weekly\_Sales")) %>%

group\_by(Date) %>%

summarise(sales = mean(Weekly\_Sales)) %>%

mutate(sales\_high = sales > temp\_s\_nh$sale)

print(temp\_s\_h)

paste("Following holiday dates have higher sales than non-holiday: ")

print(temp\_s\_h %>% subset(sales\_high == TRUE, select= c("Date", "sales\_high") ))



|  |  |
| --- | --- |
| List holidays, convert them to Date format. |  |
| Average of all sales non-holiday |  |
| From Walmart\_data, take out all unique holidays and conform they are in the list of question holidays.  Once confirmed, Average of sales for each holiday across all stores.  Add column sales\_high to show whether High or not than non-holiday sales. |  |
| Following holiday dates have higher sales than non-holiday |  |

1. Provide a monthly and semester view of sales in units and give insights.

|  |
| --- |
| Plots for Semester and Month Sales |
|  |
|  |
|  |
| There is spike in Sales in December 2012 and 2012 which may be due to Christmas holidays. |
| Sales immediately dip in next month January (2011 and 2012) (people have already spent last month) |
| This explains sales rise (2010 S1 -> 2010 S2, 2011 S1, 2011 S2) and dip (2010 S2 -> 2011 S1, 2011 S2 -> 2012 S1 ) |
| Sales seem to fall from S1 to S2 in 2012 in contrast to 2010, 2011 but is not the case. We don’t have data month of december for 2012 which as per trend might push sales |
| December month has high influence on overall sales for a Year, Semester. |

temp\_sales\_sem = walmart\_data %>%

group\_by(year, semester) %>%

summarise(sales\_sem = sum(Weekly\_Sales))

temp\_sales\_sem$Sem\_Year = interaction(temp\_sales\_sem$semester, temp\_sales\_sem$year, sep=" ")

temp\_sales\_month = walmart\_data %>%

group\_by(year, month) %>%

summarise(sales\_month = sum(Weekly\_Sales))

temp\_sales\_month$month\_year = interaction(temp\_sales\_month$month, temp\_sales\_month$year, sep=" ")

# plot month

options(scipen=999999)

len = length(temp\_sales\_month$month\_year)

x1 = 1:len

y1 = temp\_sales\_month$sales\_month

plot(y1 ~ x1, frame.plot= TRUE, xaxt='n', xlab = "Months of 2011, 2012, 2013", ylab = "Sales")

axis(1, 1:len, temp\_sales\_month$month)

curve(splinefun(x1, y1, method = "monoH.FC")(x), add = TRUE, lty = 1, lwd = 1.5, col = "blue", n = 1001)

grid()

# plot semester

options(scipen=999999)

len = length(temp\_sales\_sem$Sem\_Year)

x1 = 1:len

y1 = temp\_sales\_sem$sales\_sem

plot.default(y1 ~ x1, frame.plot= TRUE, xaxt='n', xlab = "Semester Year", ylab="Sales",

panel.first = grid(length(len), length(len)))

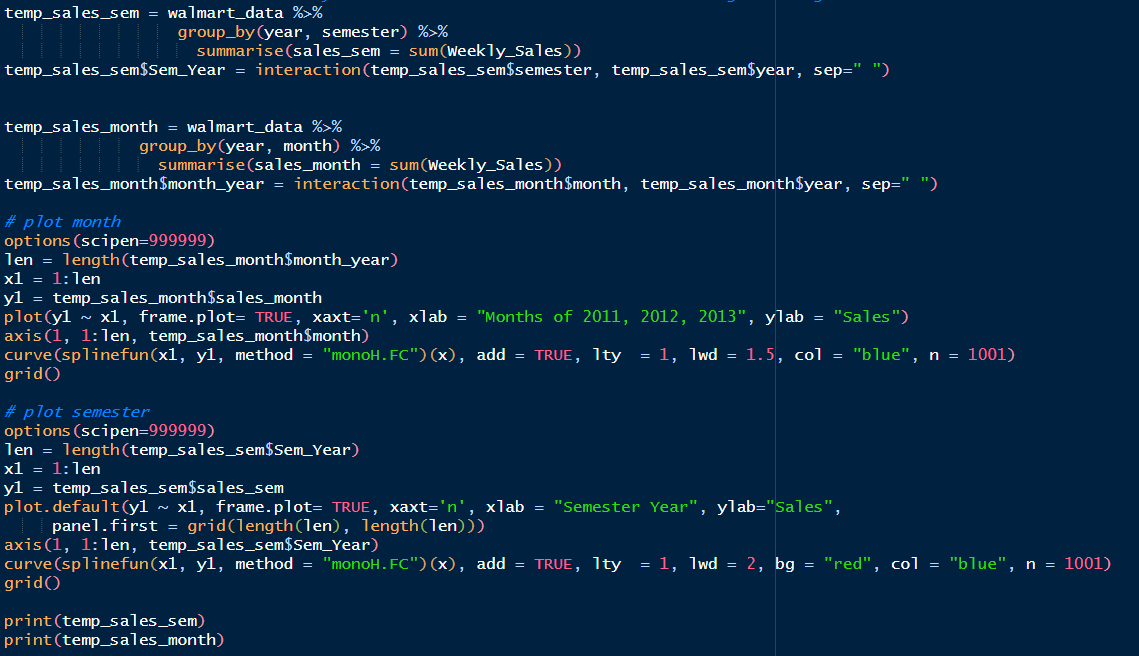
axis(1, 1:len, temp\_sales\_sem$Sem\_Year)

curve(splinefun(x1, y1, method = "monoH.FC")(x), add = TRUE, lty = 1, lwd = 2, bg = "red", col = "blue", n = 1001)

grid()

print(temp\_sales\_sem)

print(temp\_sales\_month)



|  |  |
| --- | --- |
| Grouped (year, semester) semester sales. |  |
| Grouped (year, month) monthly sales |  |

Statistical Model

For Store 1 – Build prediction models to forecast demand

Linear Regression – Utilize variables like date and restructure dates as 1 for 5 Feb 2010 (starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.

Change dates into days by creating new variable.

Select the model which gives best accuracy.

|  |  |
| --- | --- |
| Selected Model | Accuracy |
| lm (Weekly\_Sales ~CPI+Unemployment+Fuel\_Price) | 93.65 % |

1. Simple linear Regression on individual variables (summarized in a table for convenience) with respective plots(with correlations) ->

# run model

lm1 = lm(Weekly\_Sales ~Date)

lm2 = lm(Weekly\_Sales ~CPI)

lm3 = lm(Weekly\_Sales ~Unemployment)

lm4 = lm(Weekly\_Sales ~Fuel\_Price)

# summary

s\_lm1 = summary(lm1) # Date significance = "\*"

s\_lm2 = summary(lm2) # CPI significance = "\*\*"

s\_lm3 = summary(lm3) # Unemployment significance = NULL

s\_lm4 = summary(lm4) # Fuel\_Price significance = NULL

#bind coeficients

summary\_data = data.frame() %>%

bind\_coeff(s\_lm1) %>%

bind\_coeff(s\_lm2) %>%

bind\_coeff(s\_lm3) %>%

bind\_coeff(s\_lm4)

#bind significance, adjusted r-squared,

significances = pull(summary\_data %>% select(starts\_with("Pr")))

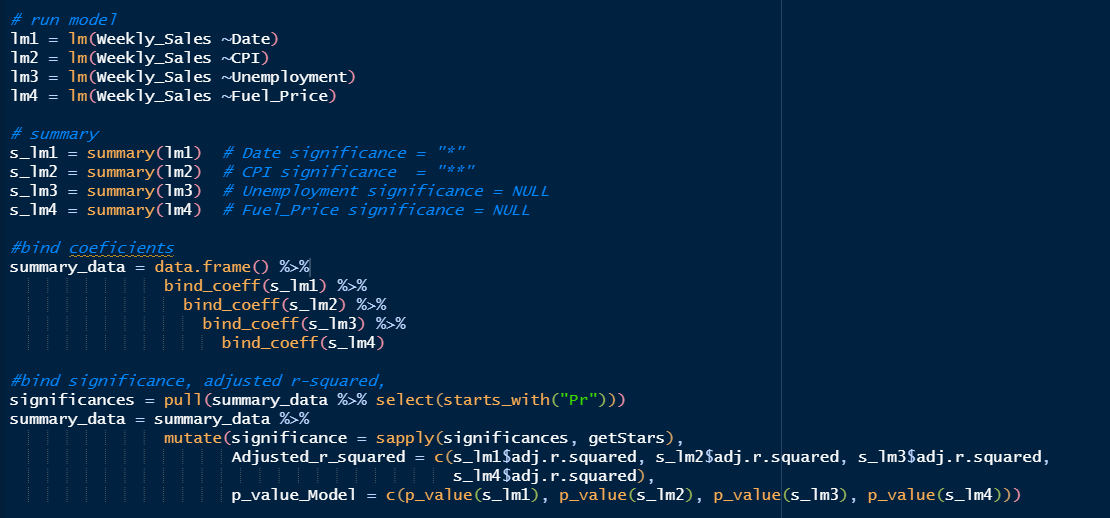
summary\_data = summary\_data %>%

mutate(significance = sapply(significances, getStars),

Adjusted\_r\_squared = c(s\_lm1$adj.r.squared, s\_lm2$adj.r.squared, s\_lm3$adj.r.squared,

s\_lm4$adj.r.squared),

p\_value\_Model = c(p\_value(s\_lm1), p\_value(s\_lm2), p\_value(s\_lm3), p\_value(s\_lm4)))





# plot graphs

op <- par(mfrow = c(2,2), mar = .1+ c(2,2,3,1))

plot(Weekly\_Sales ~ Date, main = "Weekly\_Sales(y), Date(x)")

mtext(paste("corr: ", cor(Weekly\_Sales, Date)), col = "red", cex = 0.8)

plot(Weekly\_Sales ~ CPI, main = "Weekly\_Saless(y), CPI(x)")

mtext(paste("corr: ", cor(Weekly\_Sales, CPI)), col = "red", cex = 0.8)

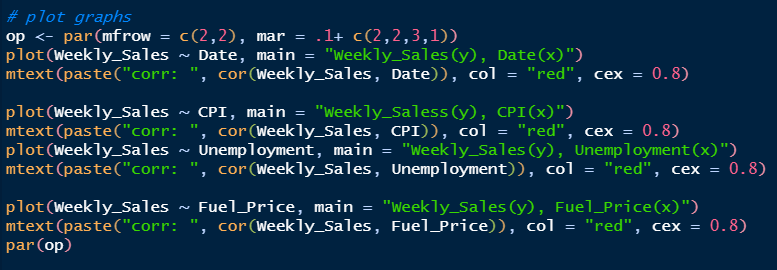
plot(Weekly\_Sales ~ Unemployment, main = "Weekly\_Sales(y), Unemployment(x)")

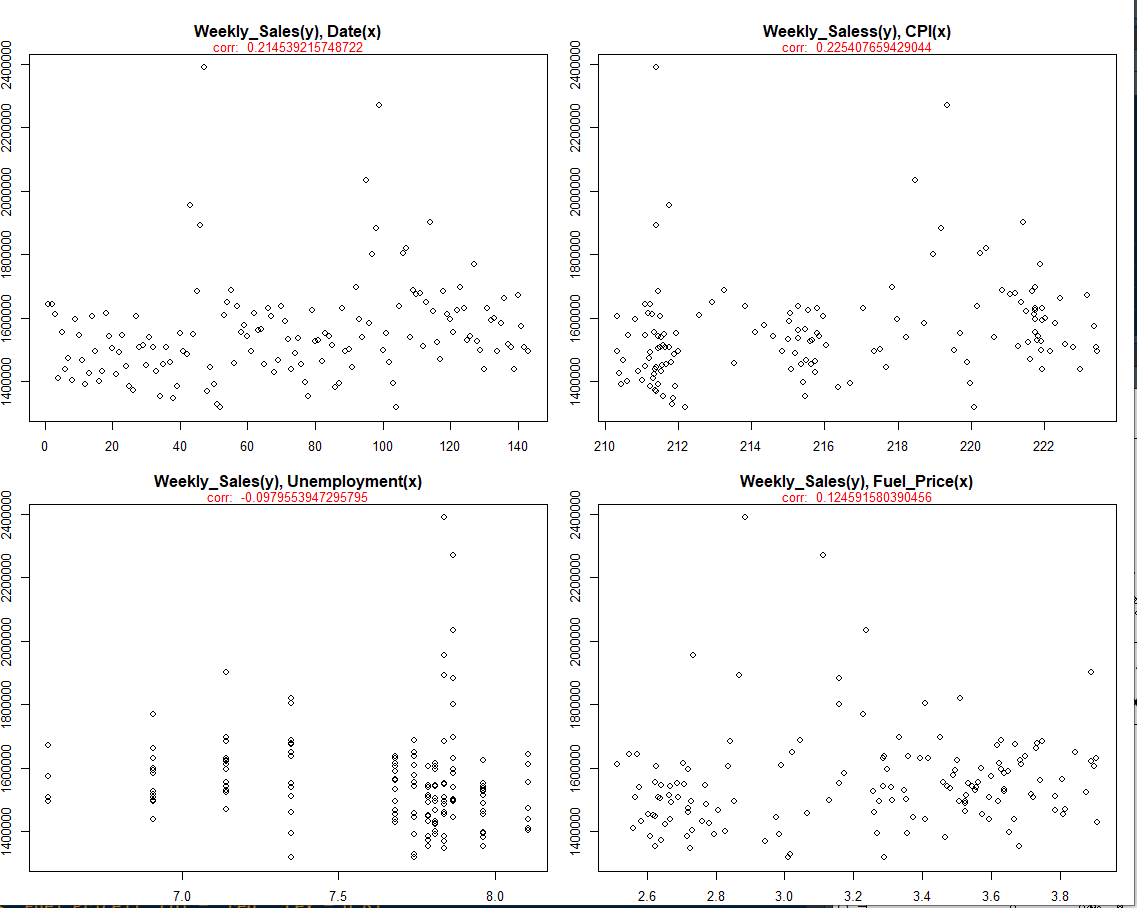
mtext(paste("corr: ", cor(Weekly\_Sales, Unemployment)), col = "red", cex = 0.8)

plot(Weekly\_Sales ~ Fuel\_Price, main = "Weekly\_Sales(y), Fuel\_Price(x)")

mtext(paste("corr: ", cor(Weekly\_Sales, Fuel\_Price)), col = "red", cex = 0.8)

par(op)





***Date and CPI have good significance and positive correlation with Sales.***

1. Multiple Linear Regression with (Date, CPI, Unemployment, Fuel\_Price)

# run regression across all variables

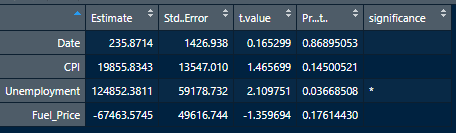
lm\_multi = lm(Weekly\_Sales ~Date+CPI+Unemployment+Fuel\_Price)

# summary

s\_lm\_multi = summary(lm\_multi)

Result summary ->

|  |  |
| --- | --- |
| Adjusted-r | 0.05865 |
| p-value | 0.01479 |
| F-statistic | 3.212 |



Date has an exceptionally high p-value = 0.86 which is way beyond 0.05.

With respect to simple liner regression in step (1), its in complete contrast -> 

Looks Date is bloating results. Possible candidate for rejection. Drop Date.

1. Check dropping Date and running model

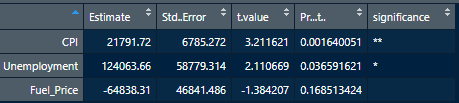
# run linear model dropping Date (with CPI, Unemployment, Fuel\_Price)

lm\_multi2 = lm(Weekly\_Sales ~CPI+Unemployment+Fuel\_Price)

# summary

s\_lm\_multi2 = summary(lm\_multi2)

|  |  |
| --- | --- |
| Adjusted-r | 0.06524 |
| p-value | 0.006162 |
| F-statistic | 4.303 |



Without Date variable,

1. Adjusted-r has increased indicating there was dilution in overall result by introducing Date.
2. Comparatively higher F-statistic moves model towards being significant model
3. Continuing with point (3) model ( lm(Weekly\_Sales ~CPI+Unemployment+Fuel\_Price)). Retrieve accuracy

# summary

s\_lm\_multi2 = summary(lm\_multi2)

# choosing the model with (Weekly\_Sales ~CPI+Unemployment+Fuel\_Price)

predicted\_sales = predict(lm\_multi2)

store1\_data = mutate(store1\_data, predicted\_sale = predicted\_sales)

store1\_data = transform(store1\_data, Err\_pct=(abs(store1\_data$Weekly\_Sales -store1\_data$predicted\_sale)/store1\_data$Weekly\_Sales))

View(store1\_data)

# Model accuracy

mean(store1\_data$Err\_pct)

1- mean(store1\_data$Err\_pct)

Model Accuracy = 93.65 %

1. An overshot since question is already answered. Why Unemployment in simple liner regression have no significance whereas in Multiple it has

Plotting correlation of Unemployment with other variables (Date, CPI)

# check correlations with other variables

op <- par(mfrow = c(2,2), mar = .1+ c(2,2,3,1))

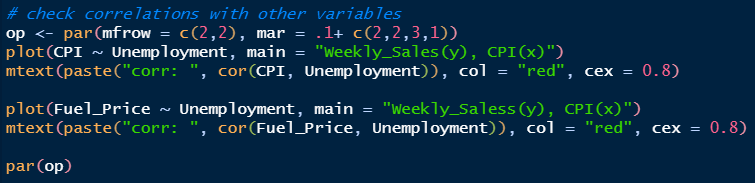
plot(CPI ~ Unemployment, main = "Weekly\_Sales(y), CPI(x)")

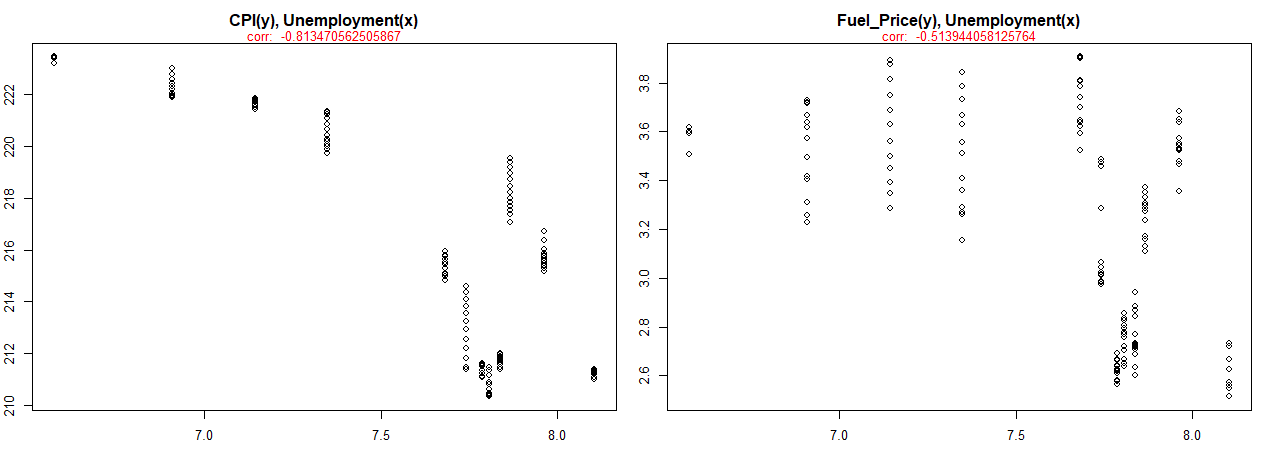
mtext(paste("corr: ", cor(CPI, Unemployment)), col = "red", cex = 0.8)

plot(Fuel\_Price ~ Unemployment, main = "Weekly\_Saless(y), CPI(x)")

mtext(paste("corr: ", cor(Fuel\_Price, Unemployment)), col = "red", cex = 0.8)

par(op)





It seems Unemployment has a strong negative correlation with CPI which directly influences sales (so indirect influence). Could be the reason model has seen indirect significance.