# Task 2

# Quantium Virtual Internship

# Experiment and uplift testing

# 

# Parisa Rahbarian

# 12 Oct 2021

Load required libraries:

**library(data.table)**

**library(ggplot2)**

**library(ggmosaic)**

**library(readr)**

**library(tidyr)**

Fill in the path to your working directory. If you are on a Windows machine, you will need to #use forward slashes (/) instead of backshashes (\)

**filePath <- "D:/Parisa/Internship/Quantium/"**

**data <- fread(paste0(filePath,"QVI\_s=data.csv"))**

***Set themes for plots***

**theme\_set(theme\_bw())**

**theme\_update(plot.title = element\_text(hjust = 0.5))**

**Select control stores**

The client has selected store numbers 77, 86 and 88 as trial stores and want control stores to be established stores that are operational for the entire observation period. We would want to match trial stores to control stores that are similar to the trial store prior to the trial period of Feb 2019 in terms of:

- Monthly overall sales revenue

- Monthly number of customers

- Monthly number of transactions per customer

***Calculate these measures over time for each store***

**library(lubridate)**

**library(tidyverse)**

**library(dplyr)**

**monthYear <- format(as.Date(data$DATE),"%Y%m")**

**data[, YEARMONTH := monthYear]**

**data$YEARMONTH <- as.numeric(as.character(data$YEARMONTH))**

Next, we define the measure calculations to use during the analysis. For each store and month calculate total sales, number of customers, transactions per customer, chips per customer and the average price per unit.

**measureOverTime <- data %>% group\_by(STORE\_NBR,YEARMONTH) %>% summarise(totSales = sum(TOT\_SALES),nCustomers = uniqueN(LYLTY\_CARD\_NBR),nTxnPerCust = uniqueN(TXN\_ID)/uniqueN(LYLTY\_CARD\_NBR), nChipsPerTxn = sum(PROD\_QTY)/uniqueN(TXN\_ID), avgPricePerUnit = (sum(TOT\_SALES)/sum(PROD\_QTY)))**

**STORE\_NBR YEARMONTH totSales nCustomers nTxnPerCust nChipsPerTxn avgPricePerUnit**

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

**1 1 201807 189. 47 1.04 1.18 3.26**

**2 1 201808 168. 41 1 1.27 3.24**

**3 1 201809 268. 57 1.04 1.20 3.78**

**4 1 201810 175. 39 1.03 1.27 3.44**

**5 1 201811 185. 44 1.02 1.22 3.36**

**6 1 201812 161. 37 1.08 1.2 3.35**

**7 1 201901 150. 35 1 1.17 3.65**

**8 1 201902 195. 49 1.04 1.14 3.36**

**9 1 201903 185. 43 1.09 1.19 3.31**

**10 1 201904 177. 39 1.03 1.3 3.41**

***Filter to the pre-trial period and stores with full observation periods***

**storesWithFullObs <- as.data.table(table(measureOverTime$STORE\_NBR))**

**storesWithFullObs <- storesWithFullObs %>% filter(N==12)**

**storesWithFullObs<-setNames(storesWithFullObs,c("STORE\_NBR","N"))**

Now we need to work out a way of ranking how similar each potential control store is to the trial store. We can calculate how correlated the performance of each store is to the trial store.

*Create a function to calculate correlation for a measure, looping through each control store.*

**trialStore\_sales <- preTrialMeasures %>% filter(STORE\_NBR ==77)**

**trialStore\_sales <- trialStore\_sales %>% select(STORE\_NBR,YEARMONTH,totSales,nCustomers)**

**calCorr <- function(preTrialMeasures,trialStore\_sales,trialStoreN){ calTable = data.table(Store1 = numeric(), Store2 = numeric(), corr\_measure = numeric())stN <- preTrialMeasures %>% select(STORE\_NBR) for(i in stN$STORE\_NBR){ contSt <- preTrialMeasures %>% filter(STORE\_NBR==i) contSt <- contSt %>% select(totSales)**

**calMeasure = data.table("Store1" = trialStoreN, "Store2" = i, "corr\_measure" = cor(trialStore\_sales$totSales,contSt$totSales)) calTable <- rbind(calTable, calMeasure) } return(calTable) }**

***Customers***

**calculateCorrelation <- function(preTrialMeasures,trialStore\_sales,trialStoreN){ calTable = data.table(Store1 = numeric(), Store2 = numeric(), corr\_measure = numeric())**

**stN <- preTrialMeasures %>% select(STORE\_NBR) for(i in stN$STORE\_NBR){ contSt <- preTrialMeasures %>% filter(STORE\_NBR==i) contSt <- contSt %>% select(nCustomers) contSt <- contSt %>% select(nCustomers) calMeasure = data.table("Store1" = trialStoreN, "Store2" = i, "corr\_measure" = cor(trialStore\_sales$nCustomers,contSt$nCustomers)) calTable <- rbind(calTable, calMeasure) } return(calTable) }**

|  |  |
| --- | --- |
|  |  |
|  |  |

Apart from correlation, we can also calculate a standardised metric based on the absolute difference between the trial store's performance and each control store's performance.

#### ***Create a function to calculate a standardised magnitude distance for a measure,***

#### ***looping through each control store***

**calculateMagnitudeDistance1 <- function(preTrialMeasures,trialStore\_sales,trial\_storeN){**

**calTable = data.table(Store1 = numeric(), Store2 = numeric(), YEARMONTH = numeric(),mag\_measure = numeric())**

**stN <- preTrialMeasures %>% select(STORE\_NBR)**

**for(i in stN$STORE\_NBR){**

**contSt <- preTrialMeasures %>% filter(STORE\_NBR==i)**

**contSt <- contSt %>% select(totSales)**

**calMeasure = data.table("Store1" = trial\_storeN, "Store2" = i, "YEARMONTH" = preTrialMeasures$YEARMONTH ,"mag\_measure" = abs(trialStore\_sales$totSales - contSt$totSales))**

**calTable <- rbind(calTable,calMeasure)**

**calTable <- unique(calTable)**

**}**

**return(calTable)**

**}**

***Standardize***

**standMag1 <- function(magnitude\_nSales) {**

**minMaxDist <- magnitude\_nSales[, .(minDist = min( magnitude\_nSales$mag\_measure), maxDist = max(magnitude\_nSales$mag\_measure)), by = c("Store1", "YEARMONTH")]**

**distTable <- merge(magnitude\_nSales, minMaxDist, by = c("Store1", "YEARMONTH"))**

**distTable[, magnitudeMeasure := 1 - (mag\_measure - minDist)/(maxDist - minDist)]**

**finalDistTable <- distTable[, .(magN\_measure = mean(magnitudeMeasure)), by = .(Store1, Store2)]**

**return(finalDistTable)**

**}**

***Customers***

**calculateMagnitudeDistance2 <- function(preTrialMeasures,trialStore\_sales,trial\_storeN){**

**calTable = data.table(Store1 = numeric(), Store2 = numeric(), YEARMONTH = numeric(),mag\_measure = numeric())**

**stN <- preTrialMeasures %>% select(STORE\_NBR)**

**for(i in stN$STORE\_NBR){**

**contSt <- preTrialMeasures %>% filter(STORE\_NBR==i)**

**contSt <- contSt %>% select(nCustomers)**

**calMeasure = data.table("Store1" = trial\_storeN, "Store2" = i, "YEARMONTH" = preTrialMeasures$YEARMONTH ,"mag\_measure" = abs(trialStore\_sales$nCustomers - contSt$nCustomers))**

**calTable <- rbind(calTable,calMeasure)**

**calTable <- unique(calTable)**

**}**

**return(calTable)**

**}**

***Standardize***

**standMag2 <- function(magnitude\_nCustomers) {**

**minMaxDist <- magnitude\_nCustomers[, .(minDist = min( magnitude\_nCustomers$mag\_measure), maxDist = max(magnitude\_nCustomers$mag\_measure)), by = c("Store1", "YEARMONTH")]**

**distTable <- merge(magnitude\_nCustomers, minMaxDist, by = c("Store1", "YEARMONTH"))**

**distTable[, magnitudeMeasure := 1 - (mag\_measure - minDist)/(maxDist - minDist)]**

**finalDistTable <- distTable[, .(magN\_measure = mean(magnitudeMeasure)), by = .(Store1, Store2)]**

**return(finalDistTable)**

**}**

Now let's use the functions to find the control stores! We'll select control stores based on how similar monthly total sales in dollar amounts and monthly number of customers are to the trial stores. So we will need to use our functions to get four scores, two for each of total sales and total customers.

Use the function you created to calculate correlations against store 77 using total sales and number of customers.

***Use the function you created to calculate correlations against store 77 using total sales and number of customers.***

**trial\_store <- 77**

**corr\_nSales <- calCorr(preTrialMeasures,trialStore\_sales,trial\_store)**

**corr\_nSales <- unique(corr\_nSales)**

**corr\_nCustomers <- calculateCorrelation(preTrialMeasures, trialStore\_sales, trial\_store )**

**corr\_nCustomers <- unique(corr\_nCustomers)**

***Use the functions for calculating magnitude***

**magnitude\_nSales <- calculateMagnitudeDistance1(preTrialMeasures, trialStore\_sales, trial\_store)**

**magnitude\_nSales <- standMag1(magnitude\_nSales)**

**magnitude\_nCustomers <‐ calculateMagnitudeDistance2(preTrialMeasures,trialStore\_sales, trial\_store)**

**magnitude\_nCustomers <- standMag2(magnitude\_nCustomers)**

We'll need to combine the all the scores calculated using our function to create a composite score to rank on. Let's take a simple average of the correlation and magnitude scores for each driver. Note that if we consider it more important for the trend of the drivers to be similar, we can increase the weight of the correlation score (a simple average gives a weight of 0.5 to the corr\_weight) or if we consider the absolute size of the drivers to be more important, we can lower the weight of the correlation score. Create a combined score composed of correlation and magnitude, by first merging the correlations table with the magnitude table.

**corr\_weight <- 0.5**

**score\_nSales <- merge(corr\_nSales,magnitude\_nSales, by = c("Store1", "Store2"))**

**score\_nSales <- score\_nSales %>% mutate(scoreNSales = (score\_nSales$corr\_measure \* corr\_weight)+(score\_nSales$magN\_measure \* (1 - corr\_weight)))**

**score\_nCustomers <- merge(corr\_nCustomers,magnitude\_nCustomers, by = c("Store1", "Store2"))**

**score\_nCustomers <- score\_nCustomers %>% mutate(scoreNCust = (score\_nCustomers$corr\_measure \* corr\_weight)+(score\_nCustomers$magN\_measure \* (1 - corr\_weight)))**

Now we have a score for each of total number of sales and number of customers. Let's combine the two via a simple average.

**score\_Control <- merge(score\_nSales,score\_nCustomers, by = c("Store1", "Store2"))**

**score\_Control <- score\_Control %>% mutate(finalControlScore = (scoreNSales \* 0.5) + (scoreNCust \* 0.5))**

The store with the highest score is then selected as the control store since it is most similar to the trial store.

***Select control stores based on the highest matching store (closest to 1 but not the store itself, i.e. the second ranked highest store)***

**control\_store <- score\_Control[order(-finalControlScore),]**

**control\_store <- control\_store$Store2**

**control\_store <- control\_store[2]**

Now that we have found a control store, let's check visually if the drivers are indeed similar in the period before the trial. We'll look at total sales first.

***Visual checks on trends based on the drivers***

**measureOverTimeSales <- as.data.table(measureOverTime)**

**pastSales <- measureOverTimeSales[, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",ifelse(STORE\_NBR == control\_store,"Control", "Other stores"))][, totSales := mean(totSales), by = c("YEARMONTH","Store\_type")][, TransactionMonth := as.Date(paste(YEARMONTH %/%100, YEARMONTH %% 100, 1, sep = "‐"), "%Y‐%m‐%d")][YEARMONTH < 201903 , ]**

***Visualize***

**ggplot(pastSales, aes(TransactionMonth, totSales, color = Store\_type)) + geom\_line() + labs(x = "Month of Operation", y = "Total Sales", title = "Total Sales by Month")**

Chart, line chart

Description automatically generated

***Conduct visual checks on customer count trends by comparing the trial store to the control store and other stores.***

**measureOverTimeCusts <- as.data.table(measureOverTime)**

**pastCustomers <- measureOverTimeCusts[, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",ifelse(STORE\_NBR == control\_store,"Control", "Other stores"))][, numberCustomers := mean(nCustomers), by = c("YEARMONTH","Store\_type")][, TransactionMonth := as.Date(paste(YEARMONTH %/%100, YEARMONTH %% 100, 1, sep = "‐"), "%Y‐%m‐%d")][YEARMONTH < 201903 ]**

***Visualize***

**ggplot(pastCustomers, aes(TransactionMonth, numberCustomers, color = Store\_type)) + geom\_line() + labs(x = "Month of Operation", y = "Total Number of Customers", title = "Total Number of Customers by Month")**

Chart, line chart

Description automatically generated

The trial period goes from the start of February 2019 to April 2019. We now want to see if there has been an uplift in overall chip sales. We'll start with scaling the control store's sales to a level similar to control for any differences between the two stores outside of the trial period.

Scale pre-trial control sales to match pre-trial trial store sales

**preTrialMeasures <- as.data.table(preTrialMeasures)**

**scalingFactorForControlSales <- preTrialMeasures[STORE\_NBR == trial\_store & YEARMONTH < 201902, sum(totSales)]/preTrialMeasures[STORE\_NBR == control\_store & YEARMONTH < 201902, sum(totSales)]**

***Applying the Scaling Factor***

**measureOverTimeSales <- as.data.table(measureOverTime)**

**scaledControlSales <- measureOverTimeSales[STORE\_NBR == control\_store, ][ , controlSales := totSales \* scalingFactorForControlSales ]**

Now that we have comparable sales figures for the control store, we can calculate the percentage difference between the scaled control sales and the trial store's sales during the trial period.

***Calculate the percentage difference between scaled control sales and trial sales***

**measureOverTime <- as.data.table(measureOverTime)**

**percentageDiff <- merge(scaledControlSales[, c("YEARMONTH", "controlSales")], measureOverTime[STORE\_NBR == trial\_store, c("totSales", "YEARMONTH")], by = "YEARMONTH")[ , percentageDiff := abs(controlSales - totSales)/controlSales]**

Let's see if the difference is significant!

As our null hypothesis is that the trial period is the same as the pre-trial period, let's take the standard deviation based on the scaled percentage difference in the pre-trial period

**stdDev <- sd(percentageDiff[YEARMONTH < 201902, percentageDiff])**

**Note that there are 8 months in the pre-trial period**

**hence 8 - 1 = 7 degrees of freedom**

**degreesOfFreedom <- 7**

We will test with a null hypothesis of there being 0 difference between trial and control stores.

**Calculate the t-values for the trial months.**

**percentageDiff[ , tvalue := (percentageDiff - 0)/stdDev][ , TransactionMonth := as.Date(paste(YEARMONTH %/% 100, YEARMONTH %% 100, 1, sep = "-"),"%Y-%m-%d")][YEARMONTH < 201905 & YEARMONTH > 201901, .(TransactionMonth, tvalue)]**

TransactionMonth tvalue

1: 2019-02-01 1.223912

2: 2019-03-01 5.633494

3: 2019-04-01 11.336505

***Also,find the 95th percentile of the t distribution with the appropriate degrees of freedom to check whether the hypothesis is statistically significant.***

**qt(0.95, df = degreesOfFreedom)**

1.894579

We can observe that the t-value is much larger than the 95th percentile value of the t-distribution for March and April - i.e. the increase in sales in the trial store in March and April is statistically greater than in the control store. Let's create a more visual version of this by plotting the sales of the control store, the sales of the trial stores and the 95th percentile value of sales of the control store.

**measureOverTimeSales <- measureOverTime**

**#### Trial and control store total sales Create new variables Store\_type, totSales and TransactionMonth in the data table.**

**pastSales <- measureOverTimeSales[, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store, "Control", "Other stores"))**

**][, totSales := mean(totSales), by = c("YEARMONTH", "Store\_type")**

**][, TransactionMonth := as.Date(paste(YEARMONTH %/% 100, YEARMONTH %% 100, 1, sep = "-"), "%Y-%m-%d")**

**][Store\_type %in% c("Trial", "Control"), ]**

**Control store 95th percentile**

**pastSales\_Controls95 <- pastSales[Store\_type == "Control",**

**][, totSales := totSales \* (1 + stdDev \* 2)**

**][, Store\_type := "Control 95th % confidence**

**interval"]**

**Control store 5th percentile**

**pastSales\_Controls5 <- pastSales[Store\_type == "Control",**

**][, totSales := totSales \* (1 - stdDev \* 2)**

**][, Store\_type := "Control 5th % confidence**

**interval"]**

**trialAssessment <- rbind(pastSales, pastSales\_Controls95, pastSales\_Controls5)**

**#### Plotting these in one nice graph**

**ggplot(trialAssessment, aes(TransactionMonth, totSales, color = Store\_type)) +**

**geom\_rect(data = trialAssessment[ YEARMONTH < 201905 & YEARMONTH > 201901 ,],**

**aes(xmin = min(TransactionMonth), xmax = max(TransactionMonth), ymin = 0 , ymax =**

**Inf, color = NULL), show.legend = FALSE) +**

**geom\_line() +**

**labs(x = "Month of operation", y = "Total sales", title = "Total sales by month")**

**```**

The results show that the trial in store 77 is significantly different to its

control store in the trial period as the trial store performance lies outside the

5% to 95% confidence interval of the control store in two of the three trial

months.Let's have a look at assessing this for number of customers as well.

This would be a repeat of the steps before for total sales Scale pre-trial control customers to match pre-trial trial store customers Compute a scaling factor to align control store customer counts to our trial store**.**

Then, apply the scaling factor to control store customer counts.

Finally, calculate the percentage difference between scaled control store customers and trial customers.

**scalingFactorForControlCust <- preTrialMeasures[STORE\_NBR == trial\_store &**

**YEARMONTH < 201902, sum(nCustomers)] / preTrialMeasures[STORE\_NBR ==**

**control\_store & YEARMONTH < 201902, sum(nCustomers)]**

**measureOverTimeCusts <- measureOverTime**

**scaledControlCustomers <- measureOverTimeCusts[STORE\_NBR == control\_store,**

**][ , controlCustomers := nCustomers \* scalingFactorForControlCust**

**][, Store\_type := ifelse(STORE\_NBR ==trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store,"Control", "Other stores"))]**

**percentageDiff <- merge(scaledControlCustomers[, c("YEARMONTH", "controlCustomers")],**

**measureOverTimeCusts[STORE\_NBR == trial\_store,c("nCustomers", "YEARMONTH")],**

**by = "YEARMONTH"**

**)[, percentageDiff := abs(controlCustomers-nCustomers)/controlCustomers]**

Let's again see if the difference is significant visually!

**#### As our null hypothesis is that the trial period is the same as the pre-trial**

**#### period, let's take the standard deviation based on the scaled percentage difference**

**#### in the pre-trial period**

**stdDev <- sd(percentageDiff[YEARMONTH < 201902 , percentageDiff])**

**degreesOfFreedom <- 7**

**#### Trial and control store number of customers**

**pastCustomers <- measureOverTimeCusts[, nCusts := mean(nCustomers), by =**

**c("YEARMONTH", "Store\_type")**

**][Store\_type %in% c("Trial", "Control"), ]**

**Control store 95th percentile**

**pastCustomers\_Controls95 <- pastCustomers[Store\_type == "Control",**

**][, nCusts := nCusts \* (1 + stdDev \* 2)**

**][, Store\_type := "Control 95th % confidence**

**interval"]**

**Control store 5th percentile**

**pastCustomers\_Controls5 <- pastCustomers[Store\_type == "Control",**

**][, nCusts := nCusts \* (1 - stdDev \* 2)**

**][, Store\_type := "Control 5th % confidence**

**interval"]**

**trialAssessment <- rbind(pastCustomers, pastCustomers\_Controls95,**

**pastCustomers\_Controls5)**

Plot everything into one nice graph.

geom\_rect creates a rectangle in the plot. Use this to highlight the trial period in our graph.

**ggplot(trialAssessment, aes(TransactionMonth, nCusts, color = Store\_type)) +**

**geom\_rect(data = trialAssessment[ YEARMONTH < 201905 & YEARMONTH > 201901 ,],**

**aes(xmin = min(TransactionMonth), xmax = max(TransactionMonth), ymin = 0 ,**

**ymax = Inf, color = NULL), show.legend = FALSE) +**

**geom\_line() + labs(x = "Month of operation", y = "Total number of customers", title = "Total number of customers by month")**

Chart, line chart

Description automatically generated

Let's repeat finding the control store and assessing the impact of the trial for

each of the other two trial stores.

**Trial store 86**

**Calculate the metrics below as we did for the first trial store.**

**measureOverTime <- data[, .(totSales = sum(TOT\_SALES),**

**nCustomers = uniqueN(LYLTY\_CARD\_NBR),**

**nTxnPerCust = (uniqueN(TXN\_ID))/(uniqueN(LYLTY\_CARD\_NBR)),**

**nChipsPerTxn = (sum(PROD\_QTY))/(uniqueN(TXN\_ID)) ,**

**avgPricePerUnit = sum(TOT\_SALES)/sum(PROD\_QTY) ) , by = c("STORE\_NBR", "YEARMONTH")][order(STORE\_NBR, YEARMONTH)]**

**Use the functions we created earlier to calculate correlations and magnitude for each potential control store**

**trial\_store <- 86**

**corr\_nSales <- calculateCorrelation(preTrialMeasures, trialStore\_sales trial\_store)**

**corr\_nCustomers <- calculateCorrelation(preTrialMeasures, trialStore\_sales trial\_store)**

**magnitude\_nSales <- calculateMagnitudeDistance1(preTrialMeasures, trialStore\_sales, trial\_store)**

**magnitude\_nCustomers <- calculateMagnitudeDistance2(preTrialMeasures, trialStore\_sales, trial\_store)**

**Now, create a combined score composed of correlation and magnitude**

**corr\_weight <- 0.5**

**score\_nSales <- merge(corr\_nSales, magnitude\_nSales, by = c("Store1", "Store2"))[ , scoreNSales := (corr\_measure + mag\_measure)/2]**

**score\_nCustomers <- merge(corr\_nCustomers, magnitude\_nCustomers, by = c("Store1", "Store2"))[ , scoreNCust := (corr\_measure + mag\_measure)/2]**

**Finally, combine scores across the drivers using a simple average.**

**score\_Control <- merge(score\_nSales, score\_nCustomers, by = c("Store1","Store2"))**

**score\_Control[, finalControlScore := scoreNSales \* 0.5 + scoreNCust \* 0.5]**

**Select control stores based on the highest matching store (closest to 1 but not the store itself, i.e. the second ranked highest store)**

**Select control store for trial store 86**

**control\_store <- score\_Control[Store1 == trial\_store,**

**][order(-finalControlScore)][2, Store2]**

**control\_store**

Looks like store 155 will be a control store for trial store 86.

Again, let's check visually if the drivers are indeed similar in the period before

the trial.

We'll look at total sales first.

**#### Conduct visual checks on trends based on the drivers**

**measureOverTimeSales <- measureOverTime**

**pastSales <- measureOverTimeSales[, Store\_type:= ifelse(STORE\_NBR == trial\_store, "Trial", ifelse(STORE\_NBR== control\_store, "Control", "Other stores"))][, totSales := mean(totSales), by = c("YEARMONTH", "Store\_type")][, TransactionMonth:= as.Date(paste(YEARMONTH%/%100, YEARMONTH%% 100, 1, sep = "-"), "%Y-%m-%d")][YEARMONTH <210903]**

**ggplot(pastSales, aes(TransactionMonth, totSales, color = Store\_type)) +**

**geom\_line(aes(linetype = Store\_type)) +**

**labs(x = "Month of operation", y = "Total sales", title = "Total sales by month")**

Chart, line chart

Description automatically generated

Great, sales are trending in a similar way.

Next, number of customers.

**#### Conduct visual checks on trends based on the drivers**

**measureOverTimeCusts <- measureOverTime**

**pastCustomers <- measureOverTimeCusts[, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store, "Control", "Other stores"))**

**][, numberCustomers := mean(nCustomers), by = c("YEARMONTH", "Store\_type")**

**][, TransactionMonth := as.Date(paste(YEARMONTH %/%**

**100, YEARMONTH %% 100, 1, sep = "-"), "%Y-%m-%d")**

**][YEARMONTH < 201903 , ]**

**ggplot(pastCustomers, aes(TransactionMonth, numberCustomers, color = Store\_type)) +**

**geom\_line() +**

**labs(x = "Month of operation", y = "Total number of customers", title = "Total number of customers by month")**

Chart, line chart

Description automatically generated

Good, the trend in number of customers is also similar.

Let's now assess the impact of the trial on sales.

**Scale pre-trial control sales to match pre-trial trial store sales**

**scalingFactorForControlSales <- preTrialMeasures[STORE\_NBR == trial\_store &**

**YEARMONTH < 201902, sum(totSales)]/preTrialMeasures[STORE\_NBR == control\_store &**

**YEARMONTH < 201902, sum(totSales)]**

**Apply the scaling factor**

**measureOverTimeSales <- measureOverTime**

**scaledControlSales <- measureOverTimeSales[STORE\_NBR == control\_store, ][ ,**

**controlSales := totSales \* scalingFactorForControlSales]**

**Calculate the percentage difference between scaled control sales and trial sales**

**When calculating percentage difference, remember to use absolute difference**

**percentageDiff <- merge(scaledControlSales[, c("YEARMONTH", "controlSales")],**

**measureOverTime[STORE\_NBR == trial\_store, c("totSales", "YEARMONTH")],**

**by = "YEARMONTH"**

**)[, percentageDiff := abs(controlSales-totSales)/controlSales]**

**As our null hypothesis is that the trial period is the same as the pre-trial**

**period, let's take the standard deviation based on the scaled percentage difference**

**in the pre-trial period Calculate the standard deviation of percentage differences during the pre-trial period**

**stdDev <- sd(percentageDiff[YEARMONTH < 201902 , percentageDiff])**

**degreesOfFreedom <- 7**

**Trial and control store total sales**

**Create a table with sales by store type and month.**

**#### We only need data for the trial and control store.**

**measureOverTimeSales <- measureOverTime**

**pastSales <- measureOverTimeSales[, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store, "Control", "Other stores"))**

**][, totSales := mean(totSales), by = c("YEARMONTH", "Store\_type")**

**][, TransactionMonth := as.Date(paste(YEARMONTH %/%100, YEARMONTH %% 100, 1, sep = "-"), "%Y-%m-%d")**

**][Store\_type %in% c("Trial", "Control"), ]**

**Calculate the 5th and 95th percentile for control store sales.**

**The 5th and 95th percentiles can be approximated by using two standard deviations away from the mean.**

**Recall that the variable stdDev earlier calculates standard deviation in percentages, and not dollar sales.**

**Control store 95th percentile**

**pastSales\_Controls95 <- pastSales[Store\_type == "Control",**

**][, totSales := totSales \* (1 + stdDev \* 2)**

**][, Store\_type := "Control 95th % confidence interval"]**

**#### Control store 5th percentile**

**pastSales\_Controls5 <- pastSales[Store\_type == "Control",**

**][, totSales := totSales \* (1 - stdDev \* 2)**

**][, Store\_type := "Control 5th % confidence interval"]**

**Then, create a combined table with columns from pastSales, pastSales\_Controls95 and pastSales\_Controls5**

**trialAssessment <- rbind(pastSales, pastSales\_Controls95, pastSales\_Controls5)**

**Plotting these in one nice graph**

**ggplot(trialAssessment, aes(TransactionMonth, totSales, color = Store\_type)) +**

**geom\_rect(data = trialAssessment[ YEARMONTH < 201905 & YEARMONTH > 201901 ,],**

**aes(xmin = min(TransactionMonth), xmax = max(TransactionMonth), ymin = 0 , ymax =**

**Inf, color = NULL), show.legend = FALSE) +**

**geom\_line(aes(linetype = Store\_type)) +**

**labs(x = "Month of operation", y = "Total sales", title = "Total sales by month")**

Chart, line chart

Description automatically generated

The results show that the trial in store 86 is not significantly different to its

control store in the trial period as the trial store performance lies inside the 5%

to 95% confidence interval of the control store in two of the three trial months.

Let's have a look at assessing this for the number of customers as well.

**#This would be a repeat of the steps before for total sales**

**# Scale pre-trial control customers to match pre-trial trial store customers**

**scalingFactorForControlCust <- preTrialMeasures[STORE\_NBR == trial\_store &**

**YEARMONTH < 201902, sum(nCustomers)]/preTrialMeasures[STORE\_NBR == control\_store &**

**YEARMONTH < 201902, sum(nCustomers)]**

**Apply the scaling factor**

**measureOverTimeCusts <- measureOverTime**

**scaledControlCustomers <- measureOverTimeCusts[STORE\_NBR == control\_store,**

**][ , controlCustomers := nCustomers**

**\* scalingFactorForControlCust**

**][, Store\_type := ifelse(STORE\_NBR**

**== trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store,**

**"Control", "Other stores"))**

**]**

**Calculate the percentage difference between scaled control sales and trial sales**

**percentageDiff <- merge(scaledControlCustomers[, c("YEARMONTH",**

**"controlCustomers")],**

**measureOverTime[STORE\_NBR == trial\_store, c("nCustomers",**

**"YEARMONTH")],**

**by = "YEARMONTH"**

**)[, percentageDiff :=**

**abs(controlCustomers-nCustomers)/controlCustomers]**

**As our null hypothesis is that the trial period is the same as the pre-trial**

**period, let's take the standard deviation based on the scaled percentage difference**

**in the pre-trial period**

**stdDev <- sd(percentageDiff[YEARMONTH < 201902 , percentageDiff])**

**degreesOfFreedom <- 7**

**#### Trial and control store number of customers**

**pastCustomers <- measureOverTimeCusts[, nCusts := mean(nCustomers), by =**

**c("YEARMONTH", "Store\_type")**

**][Store\_type %in% c("Trial", "Control"), ]**

**#### Control store 95th percentile**

**pastCustomers\_Controls95 <- pastCustomers[Store\_type == "Control",**

**][, nCusts := nCusts \* (1 + stdDev \* 2)**

**][, Store\_type := "Control 95th % confidence**

**interval"]**

**Control store 5th percentile**

**pastCustomers\_Controls5 <- pastCustomers[Store\_type == "Control",**

**][, nCusts := nCusts \* (1 - stdDev \* 2)**

**][, Store\_type := "Control 5th % confidence**

**interval"]**

**trialAssessment <- rbind(pastCustomers, pastCustomers\_Controls95,**

**pastCustomers\_Controls5)**

**Plotting these in one nice graph**

**ggplot(trialAssessment, aes(TransactionMonth, nCusts, color = Store\_type)) +**

**geom\_rect(data = trialAssessment[ YEARMONTH < 201905 & YEARMONTH > 201901 ,],**

**aes(xmin = min(TransactionMonth), xmax = max(TransactionMonth), ymin = 0 , ymax =**

**Inf, color = NULL), show.legend = FALSE) +**

**geom\_line() +**

**labs(x = "Month of operation", y = "Total number of customers", title = "Total**

**number of customers by month")**

It looks like the number of customers is significantly higher in all of the three

months. This seems to suggest that the trial had a significant impact on increasing

the number of customers in trial store 86 but as we saw, sales were not

significantly higher. We should check with the Category Manager if there were

special deals in the trial store that were may have resulted in lower prices,

impacting the results.

**## Trial store 88**

**Conduct the analysis on trial store 88.**

**measureOverTime <- data[, .(totSales = sum(TOT\_SALES),**

**nCustomers = uniqueN(LYLTY\_CARD\_NBR),**

**nTxnPerCust = uniqueN(TXN\_ID)/uniqueN(LYLTY\_CARD\_NBR),**

**nChipsPerTxn = sum(PROD\_QTY)/uniqueN(TXN\_ID),**

**avgPricePerUnit = sum(TOT\_SALES)/sum(PROD\_QTY))**

**, by = c("STORE\_NBR", "YEARMONTH")][order(STORE\_NBR, YEARMONTH)]**

**Use the functions from earlier to calculate the correlation of the sales and number of customers of each potential control store to the trial store**

**trial\_store <- 88**

**corr\_nSales <- calculateCorrelation(preTrialMeasures, quote(totSales),trial\_store)**

**corr\_nCustomers <- calculateCorrelation(preTrialMeasures, quote(nCustomers), trial\_store)**

**Use the functions from earlier to calculate the magnitude distance of the sales and number of customers of each potential control store to the trial store**

**magnitude\_nSales <- calculateMagnitudeDistance(preTrialMeasures, quote(totSales), trial\_store)**

**magnitude\_nCustomers <- calculateMagnitudeDistance(preTrialMeasures, quote(nCustomers), trial\_store)**

**Create a combined score composed of correlation and magnitude by merging the correlations table and the magnitudes table, for each driver.**

**corr\_weight <- 0.5**

**score\_nSales <- merge(corr\_nSales, magnitude\_nSales, by = c("Store1", "Store2"))[ , scoreNSales := (corr\_measure + mag\_measure)/2]**

**score\_nCustomers <- merge(corr\_nCustomers, magnitude\_nCustomers, by = c("Store1", "Store2"))[ , scoreNCust := (corr\_measure + mag\_measure)/2]**

**Combine scores across the drivers by merging sales scores and customer scores, and compute a final combined score.**

**score\_Control <- merge(score\_nSales, score\_nCustomers, by = c("Store1","Store2"))**

**score\_Control[, finalControlScore := scoreNSales \* 0.5 + scoreNCust \* 0.5]**

**Select control stores based on the highest matching store**

**(closest to 1 but not the store itself, i.e. the second ranked highest store)**

**Select control store for trial store 88**

**control\_store <- score\_Control[Store1 == trial\_store, ][order(-finalControlScore)][2, Store2]**

**control\_store**

We've now found store 237 to be a suitable control store for trial store 88.

Again, let's check visually if the drivers are indeed similar in the period before

the trial.

We'll look at total sales first.

**Visual checks on trends based on the drivers**

**For the period before the trial, create a graph with total sales of the trial store for each month, compared to the control store and other stores.**

**measureOverTimeSales <- measureOverTime**

**pastSales <- measureOverTimeSales[, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store, "Control", "Other stores"))**

**][, totSales := mean(totSales), by = c("YEARMONTH","Store\_type")**

**][, TransactionMonth := as.Date(paste(YEARMONTH %/% 100, YEARMONTH %% 100, 1, sep = "-"), "%Y-%m-%d")**

**][YEARMONTH < 201903 , ]**

**ggplot(pastSales, aes(TransactionMonth, totSales, color = Store\_type)) +**

**geom\_line(aes(linetype = Store\_type)) +**

**labs(x = "Month of operation", y = "Total sales", title = "Total sales by month")**

**Great, the trial and control stores have similar total sales.**

**Next, number of customers.**

**Visual checks on trends based on the drivers**

**For the period before the trial, create a graph with customer counts of the trial store for each month, compared to the control store and other stores.**

**measureOverTimeCusts <- measureOverTime**

**pastCustomers <- measureOverTimeCusts[, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store, "Control", "Other stores"))**

**][, numberCustomers := mean(nCustomers), by = c("YEARMONTH", "Store\_type")**

**][, TransactionMonth := as.Date(paste(YEARMONTH %/%**

**100, YEARMONTH %% 100, 1, sep = "-"), "%Y-%m-%d")**

**][YEARMONTH < 201903 , ]**

**ggplot(pastCustomers, aes(TransactionMonth, numberCustomers, color = Store\_type)) +**

**geom\_line() + labs(x = "Month of operation", y = "Total number of customers", title = "Total number of customers by month")**

**Total number of customers of the control and trial stores are also similar.**

**Let's now assess the impact of the trial on sales.**

**Scale pre-trial control store sales to match pre-trial trial store sales**

**scalingFactorForControlSales <- preTrialMeasures[STORE\_NBR == trial\_store &**

**YEARMONTH < 201902, sum(totSales)]/preTrialMeasures[STORE\_NBR ==**

**control\_store & YEARMONTH < 201902, sum(totSales)]**

**Apply the scaling factor**

**measureOverTimeSales <- measureOverTime**

**scaledControlSales <- measureOverTimeSales[STORE\_NBR == control\_store, ][ ,controlSales := totSales \* scalingFactorForControlSales]**

**Calculate the absolute percentage difference between scaled control sales and trial sales**

**percentageDiff <- merge(scaledControlSales[, c("YEARMONTH", "controlSales")],measureOverTime[STORE\_NBR == trial\_store, c("totSales", "YEARMONTH")],by = "YEARMONTH")[, percentageDiff := abs(controlSales-totSales)/controlSales]**

**As our null hypothesis is that the trial period is the same as the pre-trial period,**

**let's take the standard deviation based on the scaled percentage difference in the pre-trial period**

**stdDev <- sd(percentageDiff[YEARMONTH < 201902 , percentageDiff])**

**degreesOfFreedom <- 7**

**Trial and control store total sales**

**measureOverTimeSales <- measureOverTime**

**pastSales <- measureOverTimeSales[, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store, "Control", "Other stores"))**

**][, totSales := mean(totSales), by = c("YEARMONTH", "Store\_type")**

**][, TransactionMonth := as.Date(paste(YEARMONTH %/%100, YEARMONTH %% 100, 1, sep = "-"), "%Y-%m-%d")**

**][Store\_type %in% c("Trial", "Control"), ]**

**Control store 95th percentile**

**pastSales\_Controls95 <- pastSales[Store\_type == "Control",**

**][, totSales := totSales \* (1 + stdDev \* 2)**

**][, Store\_type := "Control 95th % confidence interval"]**

**Control store 5th percentile**

**pastSales\_Controls5 <- pastSales[Store\_type == "Control",**

**][, totSales := totSales \* (1 - stdDev \* 2)**

**][, Store\_type := "Control 5th % confidence interval"]**

**trialAssessment <- rbind(pastSales, pastSales\_Controls95, pastSales\_Controls5)**

**Plotting these in one nice graph**

**ggplot(trialAssessment, aes(TransactionMonth, totSales, color = Store\_type)) +**

**geom\_rect(data = trialAssessment[ YEARMONTH < 201905 & YEARMONTH > 201901 ,],**

**aes(xmin = min(TransactionMonth), xmax = max(TransactionMonth), ymin = 0 ,**

**ymax = Inf, color = NULL), show.legend = FALSE) +**

**geom\_line(aes(linetype = Store\_type)) +**

**labs(x = "Month of operation", y = "Total sales", title = "Total sales by month")**

Chart, line chart

Description automatically generatedThe results show that the trial in store 88 is significantly different to its

control store in the trial period as the trial store performance lies outside of

the 5% to 95% confidence interval of the control store in two of the three trial

months.

Let's have a look at assessing this for number of customers as well.

**This would be a repeat of the steps before for total sales**

**Scale pre-trial control store customers to match pre-trial trial store customers**

**scalingFactorForControlCust <- preTrialMeasures[STORE\_NBR == trial\_store &**

**YEARMONTH < 201902, sum(nCustomers)]/preTrialMeasures[STORE\_NBR ==**

**control\_store & YEARMONTH < 201902, sum(nCustomers)]**

**Apply the scaling factor**

**measureOverTimeCusts <- measureOverTime**

**scaledControlCustomers <- measureOverTimeCusts[STORE\_NBR == control\_store,**

**][ , controlCustomers := nCustomers \* scalingFactorForControlCust**

**][, Store\_type := ifelse(STORE\_NBR == trial\_store, "Trial",**

**ifelse(STORE\_NBR == control\_store,"Control", "Other stores"))**

**]**

**Calculate the absolute percentage difference between scaled control sales and trial sales**

**percentageDiff <- merge(scaledControlCustomers[, c("YEARMONTH","controlCustomers")],measureOverTime[STORE\_NBR == trial\_store, c("nCustomers", "YEARMONTH")],**

**by = "YEARMONTH")[, percentageDiff := abs(controlCustomers-nCustomers)/controlCustomers]**

**As our null hypothesis is that the trial period is the same as the pre-trial**

**period, let's take the standard deviation based on the scaled percentage difference in the pre-trial period**

**stdDev <- sd(percentageDiff[YEARMONTH < 201902 , percentageDiff])**

**degreesOfFreedom <- 7**

**note that there are 8 months in the pre-trial period hence 8 - 1 = 7 degrees of freedom**

**Trial and control store number of customers**

**pastCustomers <- measureOverTimeCusts[, nCusts := mean(nCustomers), by = c("YEARMONTH", "Store\_type")**

**][Store\_type %in% c("Trial", "Control"), ]**

**Control store 95th percentile**

**pastCustomers\_Controls95 <- pastCustomers[Store\_type == "Control",**

**][, nCusts := nCusts \* (1 + stdDev \* 2)**

**][, Store\_type := "Control 95th % confidence interval"]**

**Control store 5th percentile**

**pastCustomers\_Controls5 <- pastCustomers[Store\_type == "Control",**

**][, nCusts := nCusts \* (1 - stdDev \* 2)**

**][, Store\_type := "Control 5th % confidence interval"]**

**Combine the tables pastSales, pastSales\_Controls95, pastSales\_Controls5**

**trialAssessment <- rbind(pastCustomers, pastCustomers\_Controls95,pastCustomers\_Controls5)**

**Plotting these in one nice graph**

**ggplot(trialAssessment, aes(TransactionMonth, nCusts, color = Store\_type)) +**

**geom\_rect(data = trialAssessment[ YEARMONTH < 201905 & YEARMONTH > 201901 ,],**

**aes(xmin = min(TransactionMonth), xmax = max(TransactionMonth), ymin = 0 ,**

**ymax = Inf, color = NULL), show.legend = FALSE) + geom\_line() +**

**labs(x = "Month of operation", y = "Total number of customers", title = "Total number of customers by month")**

Chart, line chart

Description automatically generated

Total number of customers in the trial period for the trial store is significantly

higher than the control store for two out of three months, which indicates a

positive trial effect.

**Conclusion**

Good work! We've found control stores 233, 155, 237 for trial stores 77, 86 and 88

respectively.

The results for trial stores 77 and 88 during the trial period show a significant

difference in at least two of the three trial months but this is not the case for

trial store 86. We can check with the client if the implementation of the trial was

different in trial store 86 but overall, the trial shows a significant increase in

sales. Now that we have finished our analysis, we can prepare our presentation to

the Category Manager.