Wine Retailer
Case Assignment

MSBA Cohort 2 Group 22
Pin Li, Jiawen Liang, Ruiling Shen, Chenxi Tao, Khanh Tran

EXECUTIVE SUMMARY

Is sending email effective?

YES !

Purchase value increases by $1.35 on average

Average Casual Effect

What group should we send email to?

Recent Buyers

- Sending email has **insignificant** impact on non-recent buyers
- The impact is **greater** on recent buyers

Past Purchase Value

- The impact of sending email to loyal customers is **5 times** higher than to non-loyal customers

Slicing & Dicing

Who should we send email to?

43,325 customers with expected profits greater than email cost

Targeted customers have:

Past purchase:
$55 higher

Days since last purchase:
38 days shorter

Causal Forest

METHODOLOGY

Is sending email effective?

Randomization Check

Passed !

Average Casual Effect

- Run regression on main effect
- `groupemail` is statistically significant
- Sending email on average increases purchase amount by **$1.35**

What group should we send email to?

Slicing & Dicing

- Plot **histograms** of `last_purch`, `past_purch` and `visits` to find threshold to split into groups
- Plot groups' difference
- Run regression with **interaction terms** between main effect and group dummy

Who should we send email to?

Individual-level Effect

- Train **causal forest** model on the entire dataset
- Predict causal effect estimates for each customer

Scoring

Score = 30%*Beta - 0.1

Profit Margin: 30%

Email Cost: $0.1

Targeting

Send email to individuals with **score > 0**

Average Casual Effect

- ➢ Sending email is **statistically significant** and can increase customers' purchase by **$1.34**

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.7727     0.2260  56.528  < 2e-16 ***
groupemail    1.3465     0.3195   4.214 2.52e-05 ***
```

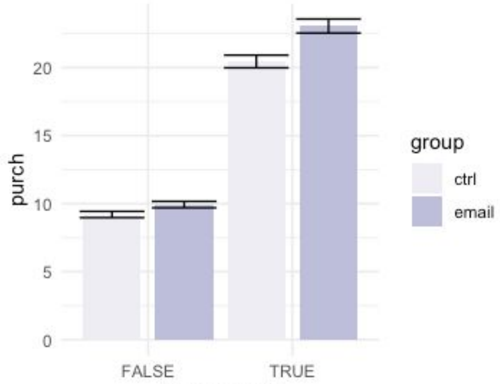
Average Purchase Value


```
==========================================================================
                                  Dependent variable:
                    ------------------------------------------------------
                                         purch
                               (1)                         (2)
--------------------------------------------------------------------------
groupemail                  1.346***                    1.260***
                            (0.320)                     (0.310)

--------------------------------------------------------------------------
Model                     No Controls                 With Controls
Observations                78,312                      78,312
R2                          0.0002                      0.058
Adjusted R2                 0.0002                      0.058
Residual Std. Error   44.712 (df = 78310)          43.394 (df = 78304)
F Statistic        17.755*** (df = 1; 78310) 693.252*** (df = 7; 78304)
==========================================================================
Note:                                         *p<0.1; **p<0.05; ***p<0.01
```

- ➢ Compared to controlling all Xs, the coefficients of `groupemail` is still statistically significant.
- ➢ The expected value is slightly lower than the previous results.

Slicing and Dicing

Recent Buyers

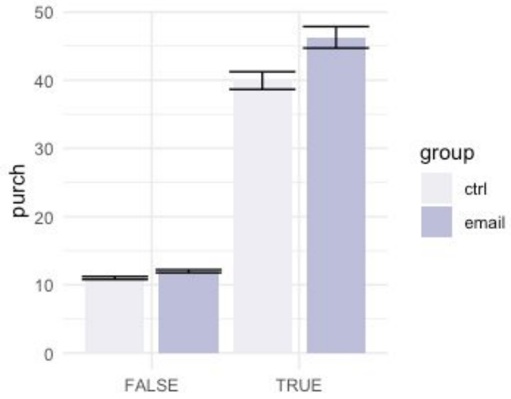


Recent Buyers (last_purch < 35)

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 9.1999     0.2713  33.912  < 2e-16 ***
groupemail                  0.7312     0.3839   1.905  0.05680 .
recentPurchTRUE            11.2520     0.4814  23.372  < 2e-16 ***
groupemail:recentPurchTRUE  1.8753     0.6804   2.756  0.00585 **
---
```

- ➢ Recent buyers purchase **$11.25** more than non-recent buyers on average
- ➢ Sending email has **insignificant** impact on non-recent buyers (p-value > 0.05)
- ➢ Sending email has **greater impact** on recent buyers.

Loyal Customers



Loyal Customers (past_purch > 450)

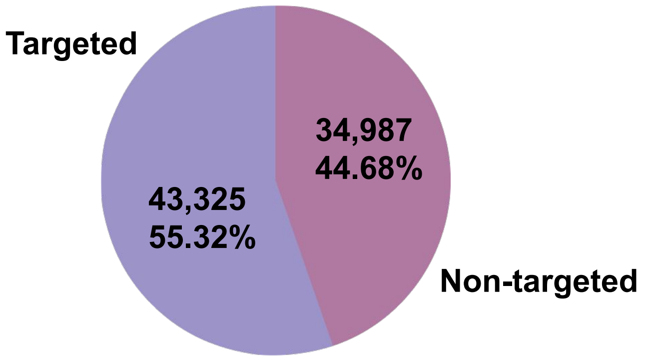
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          10.9973     0.2298  47.855  < 2e-16 ***
groupemail            1.0060     0.3250   3.095  0.00197 **
loyalTRUE            28.9529     0.9280  31.198  < 2e-16 ***
groupemail:loyalTRUE  5.3246     1.3104   4.063 4.84e-05 ***
---
```

- ➢ Loyal customers purchase **$28.95** more than non-loyal customers on average
- ➢ Sending email is **statistically significant** and can increase non-loyal customers' purchase by **$1.01**
- ➢ The impact of sending email is **$5.32 higher** for loyal customers.

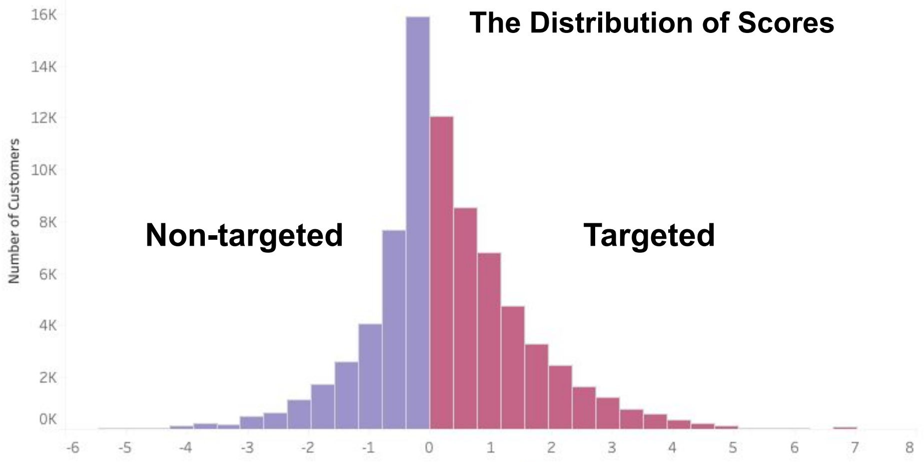
Scoring and Targeting

Findings:

- ➢ Scores concentrate between **-4** and **4**.
- ➢ Send e-mails to **43,325** customers.
- ➢ Our targeted customers have the following features on average:
  Past Purchase: 55 units higher
  Last Purchase: 38 days shorter



Number of Targeted and Non-targeted Customers



Summary of Baseline Variables for Targeted/Non-targeted Customers

GBA424: Analytics Design - Assignment 4

Pin Li, Jiawen Liang, Ruiling Shen, Chenxi Tao, Khanh Tran

2/15/2020

Table of Contents

Setup

The Wine Retailer's experiment data we will use has 78,312 observations and 13 variables.

```
dir = "/Users/srl/Desktop/UR/MSBA Class of 2021/Class/Spring A/GBA424 Analyti
cs Design:Application/Assignment /Assignment 4"
setwd(dir)
d = read.csv("test_data_1904.csv")
```

Descriptions of the variables:

- **userid** id number of users
- **cpgn_id** id number of campaigns
- **group** factor. Does the user receive an email? (`treatment`)
- **open** factor. Does the user open the email?
- **click** factor. Does the user click on the email?
- **purch** user's purchase amount (`target variable`)
- **chard** past purchased amount on `chard` (a wine type)
- **sav_blanc** past purchased amount on `sav_blance` (a wine type)
- **syrah** past purchased amount on `syrah` (a wine type)
- **cab** past purchased amount on `cab` (a wine type)
- **past_purch** total past purchased amount (`= chard + sav_blance + syrah + cab`)
- **last_purch** days since last purchase

- **visits** number of website visits

Summary of the variables:

```
summary(d)
```

```
##     user_id              cpgn_id          group             open       
##  Min.   :2000001   1904Email:78312   ctrl :39156   Min.   :0.0000  
##  1st Qu.:2019579                     email:39156   1st Qu.:0.0000  
##  Median :2039156                                   Median :0.0000  
##  Mean   :2039156                                   Mean   :0.3979  
##  3rd Qu.:2058734                                   3rd Qu.:1.0000  
##  Max.   :2078312                                   Max.   :1.0000  
##      click             purch             chard             sav_blanc      
##  Min.   :0.00000   Min.   :   0.00   Min.   :    0.00   Min.   :   0.00  
##  1st Qu.:0.00000   1st Qu.:   0.00   1st Qu.:    0.00   1st Qu.:   0.00  
##  Median :0.00000   Median :   0.00   Median :    0.00   Median :   0.00  
##  Mean   :0.06729   Mean   :  13.45   Mean   :   74.01   Mean   :  26.72  
##  3rd Qu.:0.00000   3rd Qu.:   0.00   3rd Qu.:   56.62   3rd Qu.:  21.03  
##  Max.   :1.00000   Max.   :1812.50   Max.   :13379.44   Max.   :3843.24  
##      syrah             cab             past_purch          last_purch    
##  Min.   :  0.00   Min.   :   0.00   Min.   :    0.00   Min.   :   0.00  
##  1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:    0.00   1st Qu.:  26.00  
##  Median :  0.00   Median :   0.00   Median :   52.95   Median :  63.00  
##  Mean   :  2.84   Mean   :  27.03   Mean   :  130.60   Mean   :  90.06  
##  3rd Qu.:  0.00   3rd Qu.:  21.10   3rd Qu.:  169.00   3rd Qu.: 125.00  
##  Max.   :360.32   Max.   :2649.78   Max.   :13379.44   Max.   :1225.00  
##      visits      
##  Min.   : 0.000  
##  1st Qu.: 4.000  
##  Median : 5.000  
##  Mean   : 5.647  
##  3rd Qu.: 7.000  
##  Max.   :64.000
```

Part A: Average Causal Effect

In this section, we will examine the impact of sending an email (variable `group`) on the purchase value (variable `purch`) that consumers make.

We will exclude `open` and `click` from this analysis and combine the remaining variables as `X`. Because `past_purch` is perfectly collinear with other variables, it is also excluded. The function `model.matrix` will expand factors to a set of dummy variables and expand interactions similarly.

```
# Group variables
X = model.matrix(~chard+sav_blanc+syrah+cab+last_purch+visits,
                 data=d)
X = X[,2:ncol(X)]
```

1. Randomization Check

Before analyzing the causal effect of sending an `email` on the target variable `purch`, we will do a randomization check to see whether the experiment is conducted correctly.

```
randomizationCheck = function(w, X){
  ##Assumes w is binary assignment variable (0,1) and X has columns with vari
ables for randomization check
  pvals = numeric(ncol(X))
  for(i in 1:ncol(X)){
    slm = summary(lm(X[,i]~w)) #save summary information
    pvals[i] = slm[[4]][2,4] #pull off the summary table ([[4]]) and 2nd coef
ficient's p-value (4th column), which is [2,4]
  }
  data.frame(variable=colnames(X),"p-value"=pvals,"Passed"=ifelse(pvals<.05,"
FAILED","passed"))
}

rC = randomizationCheck(d$group, X)
format(rC,digits=2)

##     variable p.value Passed
## 1       chard    0.25 passed
## 2   sav_blanc    0.97 passed
## 3       syrah    0.21 passed
## 4         cab    0.73 passed
## 5 last_purch    0.76 passed
## 6      visits    0.86 passed
```

Based on the results, we can conclude that the randomization check is passed and the experiment is conducted correctly. There is no difference in purchase history between people receiving an email and those who don't.

2. Average Casual Effect Analysis

In an experiment, we only need to run the regression on the target variable with the treatment variable. We don't have to control for other variables.

```
lm0 = lm(purch~group,data=d)
summary(lm0)

##
## Call:
## lm(formula = purch ~ group, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -14.12  -14.12  -12.77  -12.77 1798.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  12.7727     0.2260  56.528  < 2e-16 ***
## groupemail    1.3465     0.3195   4.214 2.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.71 on 78310 degrees of freedom
## Multiple R-squared:  0.0002267,  Adjusted R-squared:  0.0002139
## F-statistic: 17.76 on 1 and 78310 DF,  p-value: 2.515e-05
```

The coefficient of group is statiscally significant, indicating that people receiving an email have a different purchase amount from people not receiving one. The difference is the value of the coefficient.

For people not receiving an email, the expected purchased amount is $`12.7727`, and for people receiving an email, the expected purchased amount is $`1.3465` higher, or $`14.1192`. The standard error is `0.3195`.

Below we run the regression controlling for all `X`. With successful randomization, it should not affect the results we have above.

```
lm1 = lm(purch~group+X,data=d)
summary(lm1)
```

```
##
## Call:
## lm(formula = purch ~ group + X, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -420.37  -14.57  -10.31   -1.72 1798.77
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.5269957  0.4363336  33.293  < 2e-16 ***
## groupemail   1.2603997  0.3101382   4.064 4.83e-05 ***
## Xchard       0.0346117  0.0007959  43.489  < 2e-16 ***
## Xsav_blanc   0.0433309  0.0020630  21.004  < 2e-16 ***
## Xsyrah       0.0240070  0.0149648   1.604    0.109
## Xcab         0.0489413  0.0020948  23.363  < 2e-16 ***
## Xlast_purch -0.0718125  0.0017235 -41.667  < 2e-16 ***
## Xvisits     -0.0627548  0.0655217  -0.958    0.338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.39 on 78304 degrees of freedom
## Multiple R-squared:  0.05836,    Adjusted R-squared:  0.05827
## F-statistic: 693.3 on 7 and 78304 DF,  p-value: < 2.2e-16
```

The coefficients of `groupemail` is still statiscally significant. The expected value is slightly lower than the previous results. By controlling variables, we can absord some of the errors and reduce the standard errors.

```
stargazer(lm0, lm1, type="text", keep=c("groupemail"),
          add.lines=list(c("Model","No Controls","With Controls")))

##
## ========================================================================
##                                     Dependent variable:
##                      ---------------------------------------------------
##                                            purch
##                                 (1)                        (2)
## ------------------------------------------------------------------------
## groupemail                   1.346***                   1.260***
##                              (0.320)                    (0.310)
##
## ------------------------------------------------------------------------
## Model                       No Controls              With Controls
## Observations                  78,312                     78,312
## R2                            0.0002                     0.058
## Adjusted R2                   0.0002                     0.058
## Residual Std. Error    44.712 (df = 78310)       43.394 (df = 78304)
## F Statistic         17.755*** (df = 1; 78310) 693.252*** (df = 7; 78304)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

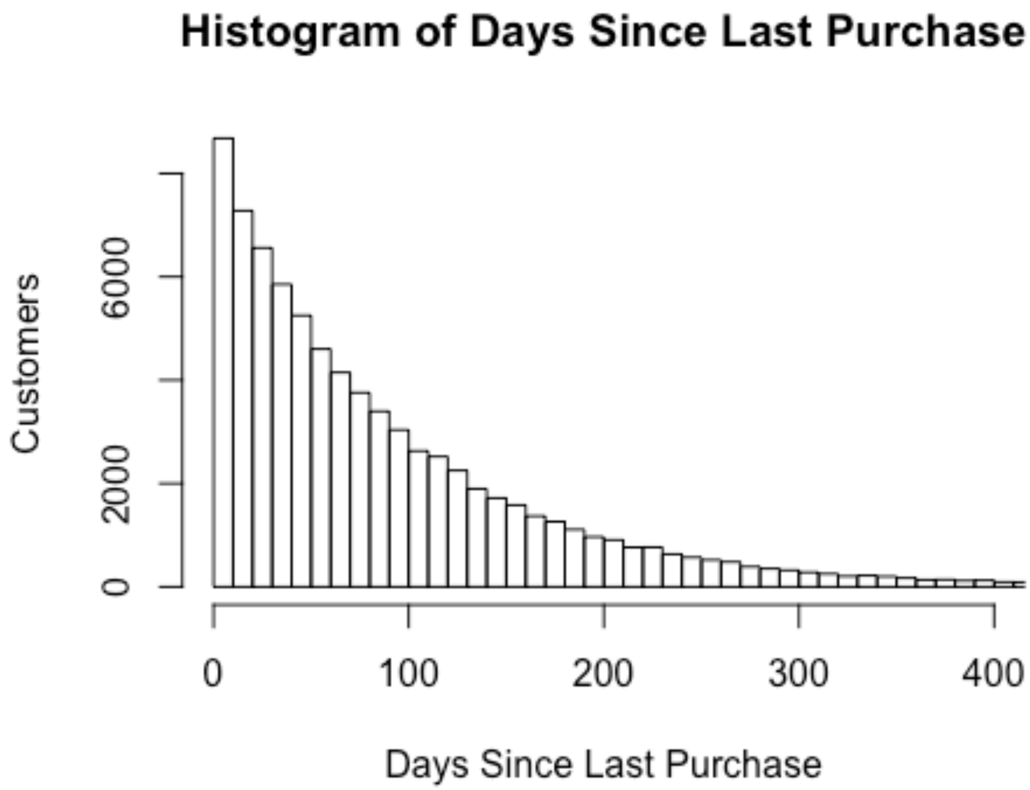
Part B: Slicing and Dicing

We then use slice and dice analysis to illustrate the potential for targeting on responses for this email campaign.

1. Recent Purchase

Firstly, we plot the histogram for last_purch.

```
# plot purchase rates for 'last_purch'
hist(d$last_purch,
     breaks = 100,
     xlab="Days Since Last Purchase", ylab="Customers",
     main="Histogram of Days Since Last Purchase",
     xlim = range(0,400))
```



We consider the customers who have made a purchase within the last 35 days as **Recent buyers.**

```
# differentiate new versus older customers
d$recentPurch = (d$last_purch < 35)
nrow(d[d$recentPurch==TRUE,])

## [1] 24925
```

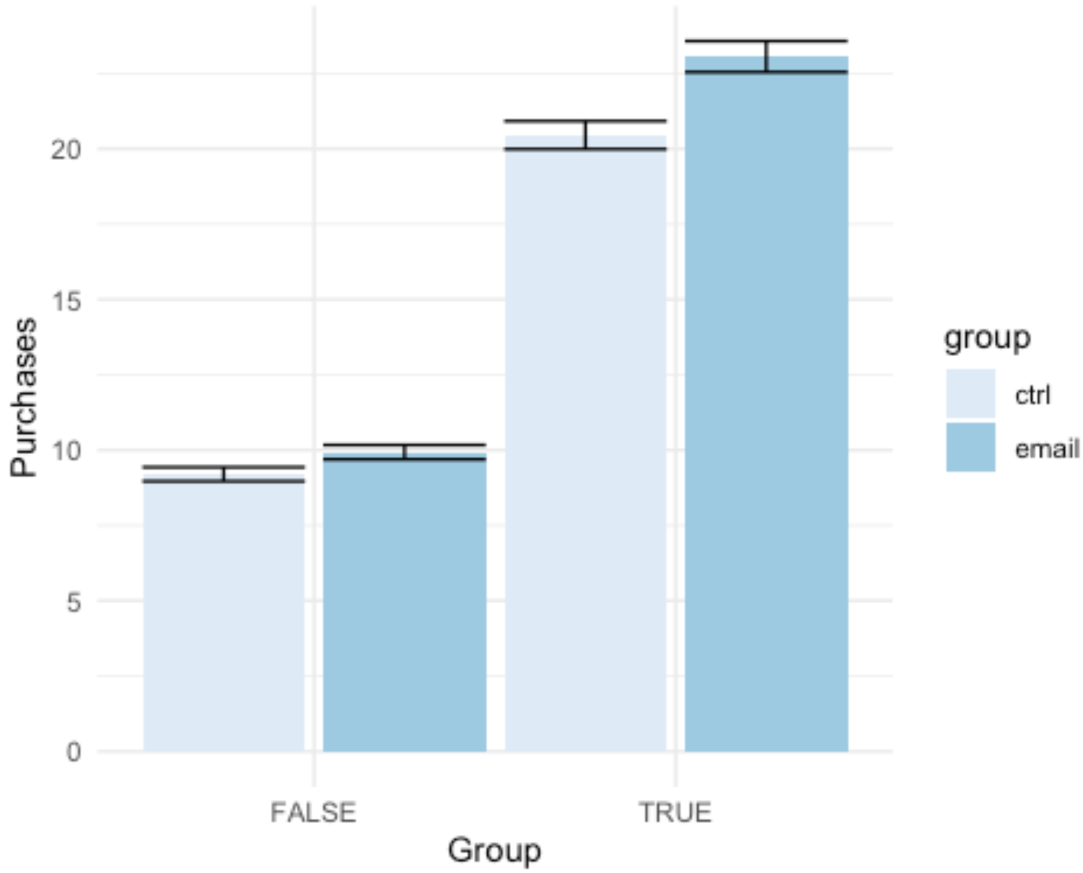
Recent buyers vs. Non-recent buyers

```
dt = data.table(d)
dagg_rec = dt[,.(open = mean(open), click=mean(click), purch = mean(purch),
                 seOpen = sd(open)/sqrt(.N), seClick=sd(click)/sqrt(.N),
                 sePurch = sd(purch)/sqrt(.N),.N), #standard error
              by = .(group,recentPurch)] #condition
dagg_rec = setorder(dagg_rec,group,-recentPurch) #display the data table via
group name by order
dagg_rec

##    group recentPurch      open     click     purch      seOpen     seClick
## 1:  ctrl        TRUE 0.0000000 0.0000000 20.451857 0.000000000 0.000000000
## 2:  ctrl       FALSE 0.0000000 0.0000000  9.199882 0.000000000 0.000000000
## 3: email        TRUE 0.9161063 0.1492955 23.058409 0.002480499 0.003188704
```

```
## 4: email       FALSE 0.7394239 0.1277003  9.931110 0.002688187 0.002043969
##      sePurch     N
## 1: 0.4650417 12433
## 2: 0.2330076 26723
## 3: 0.5156884 12492
## 4: 0.2381407 26664
```

- Recent buyers buy more on average
- The email seems to produce a stronger effect on purchases for more recent buyers (~$2.65 versus $0.74)

Is email more effective for recent buyers?



We can see that email is more effective for recent buyers.

Measuring causal effects with regression: Conditional causal effects

```
summary(lm(purch~group*recentPurch,data=d)) #compares each email to control g
roup

##
## Call:
## lm(formula = purch ~ group * recentPurch, data = d)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
##   -23.06    -9.93    -9.93    -9.20 1802.57
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   9.1999     0.2713  33.912  < 2e-16 ***
## groupemail                    0.7312     0.3839   1.905  0.05680 .
## recentPurchTRUE              11.2520     0.4814  23.372  < 2e-16 ***
## groupemail:recentPurchTRUE    1.8753     0.6804   2.756  0.00585 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.35 on 78308 degrees of freedom
## Multiple R-squared:  0.01645,    Adjusted R-squared:  0.01641
## F-statistic: 436.5 on 3 and 78308 DF,  p-value: < 2.2e-16
```

```r
p = summary(lm(purch~group*recentPurch,data=d))$coefficient[,4]
p.adjust(p, "bonferroni")
```

```
##                (Intercept)                 groupemail
##               1.174165e-249               2.271972e-01
##            recentPurchTRUE groupemail:recentPurchTRUE
##               8.626014e-120               2.340425e-02
```

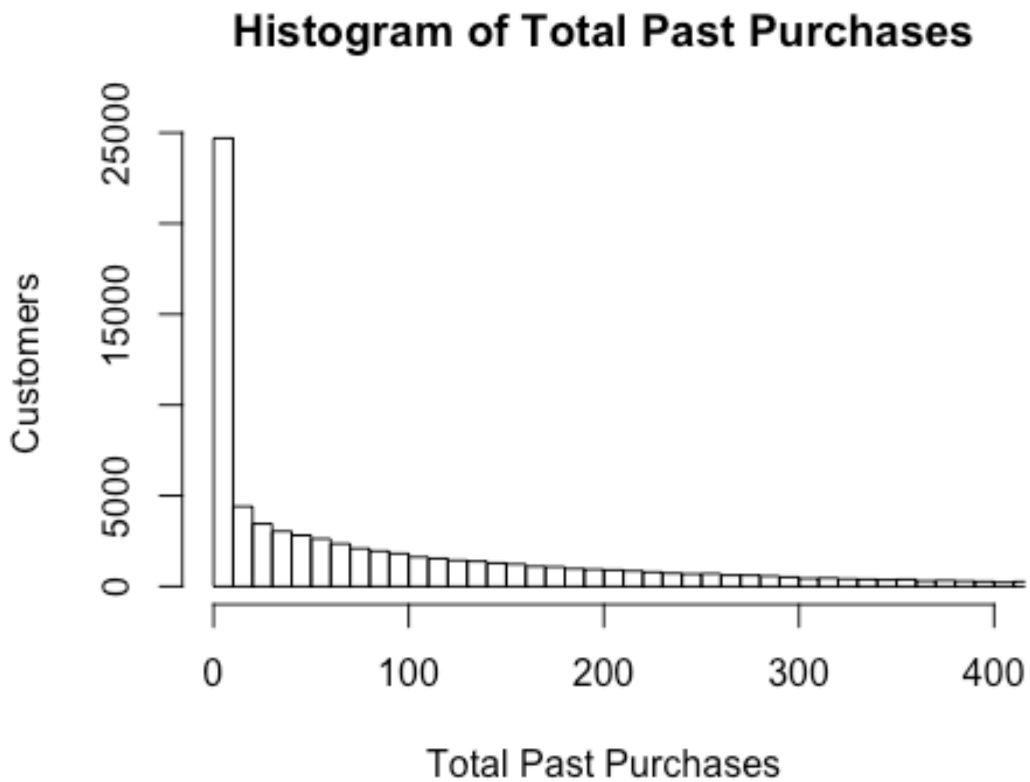
1. The main effect of the email variable is not significant (p-value = 0.23), indicating people who didn't purchase within the last 35 days are not significantly affected by the email.

2. Subgroups will vary in **how much they engage in behaviors** (*main effect of baseline variables*)
   - Recent buyers tend to have $11.25 higher average purchases in the future
3. Subgroups vary in **how much they respond to treatments** (*interaction effects*)
   - Recent buyers are more affected by the email, leading to addition $1.88 in spending

2. Past Purchase Amount

Firstly, we plot the histogram for past_purch.

```r
# plot purchase rates for 'last_purch'
hist(d$past_purch,
     breaks = 1000,
     xlab="Total Past Purchases", ylab="Customers",
     main="Histogram of Total Past Purchases",
     xlim = range(0,400))
```



We consider the customers who have made past purchase over $450 as **loyal buyers**.

```
# differentiate new versus older customers
d$pastPurch = (d$past_purch > 450)
nrow(d[d$pastPurch==TRUE,])

## [1] 4818
```

Because our test is big enough, we will have enough sample in the subgroup.

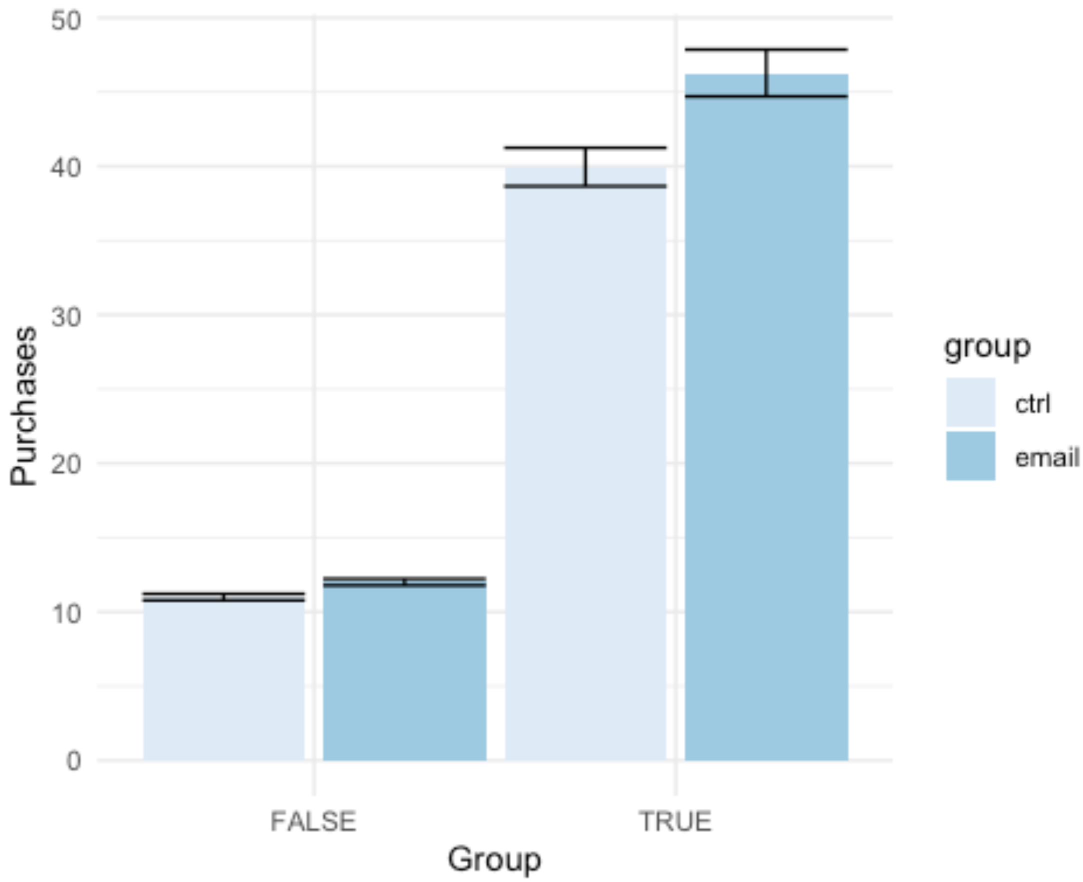
Loyal buyers vs. Non-loyal customers

```
dt = data.table(d)
dagg_past = dt[,.(open = mean(open), click=mean(click), purch = mean(purch),
                 seOpen = sd(open)/sqrt(.N), seClick=sd(click)/sqrt(.N),
                 sePurch = sd(purch)/sqrt(.N),.N), #standard error
              by = .(group,pastPurch)] #condition
dagg_past = setorder(dagg_past,group,-pastPurch) #display the data table via
group name by order
dagg_past

##    group pastPurch      open     click    purch       seOpen      seClick
## 1:  ctrl      TRUE 0.0000000 0.0000000 39.95017 0.000000000 0.000000000
## 2:  ctrl     FALSE 0.0000000 0.0000000 10.99731 0.000000000 0.000000000
```

```
## 3: email      TRUE 1.0000000 0.1832851 46.28077 0.000000000 0.007871370
## 4: email     FALSE 0.7823566 0.1313863 12.00327 0.002152868 0.001762506
##      sePurch     N
## 1: 1.2924694  2401
## 2: 0.2138109 36755
## 3: 1.5774902  2417
## 4: 0.2212722 36739
```

- Loyal buyers buy more on average
- The email seems to produce a stronger effect on purchases for loyal buyers (~$6.33 versus $1.01)

Is email more effective for loyal buyers?



We can see that email is much more effective for loyal buyers.

Measuring causal effects with regression: Conditional causal effects

```
summary(lm(purch~group*pastPurch, data=d)) #compares each email to control gr
oup

##
## Call:
## lm(formula = purch ~ group * pastPurch, data = d)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
##  -46.28  -12.00  -11.00  -11.00 1800.50 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 10.9973     0.2298  47.855  < 2e-16 ***
## groupemail                   1.0060     0.3250   3.095  0.00197 ** 
## pastPurchTRUE               28.9529     0.9280  31.198  < 2e-16 ***
## groupemail:pastPurchTRUE     5.3246     1.3104   4.063 4.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 44.06 on 78308 degrees of freedom
## Multiple R-squared:  0.02931,    Adjusted R-squared:  0.02927 
## F-statistic: 788.1 on 3 and 78308 DF,  p-value: < 2.2e-16
```

```
p_past = summary(lm(purch~group*pastPurch, data=d))$coefficient[,4]
p.adjust(p_past, "bonferroni")
```

```
##              (Intercept)               groupemail            pastPurchTRUE 
##             0.000000e+00             7.874855e-03            9.070223e-212 
## groupemail:pastPurchTRUE 
##             1.936590e-04
```

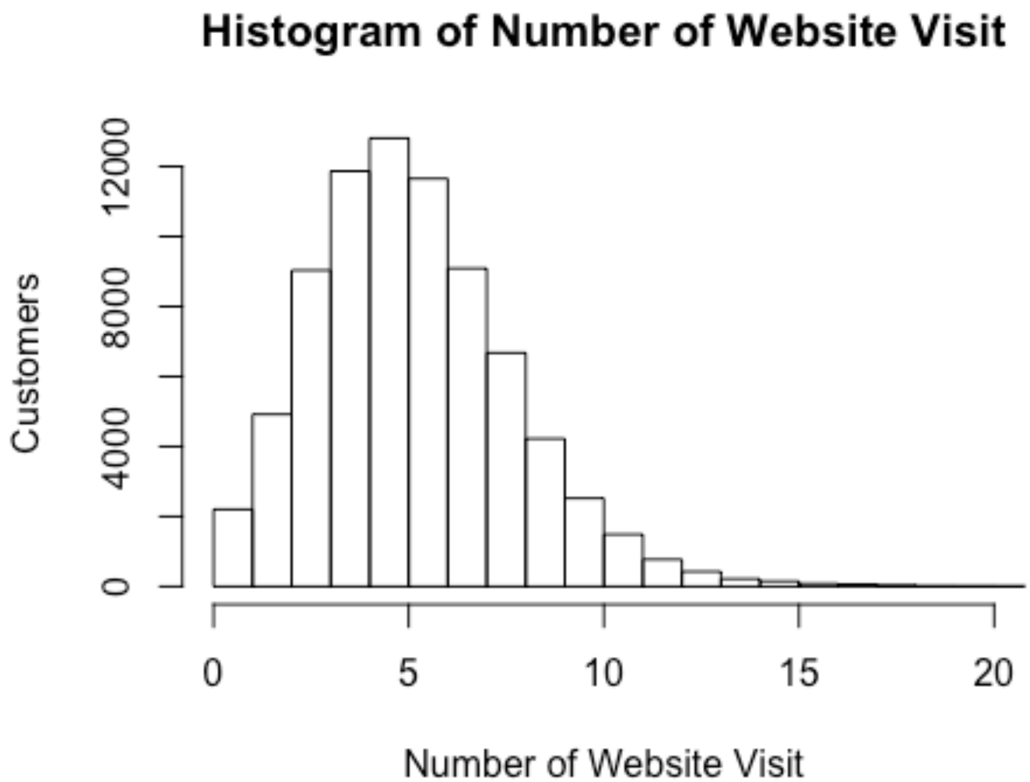
1. The main effect of the email variable is significant (p-value=0.008), leading to $1.01 more sales for those who have not bought much in the past, indicating this group of customers are significantly affected by the email.

2. Subgroups will vary in **how much they engage in behaviors** (*main effect of baseline variables*)
   - Loyal buyers tend to have $28.95 higher average purchases in the future
3. Subgroups vary in **how much they respond to treatments** (*interaction effects*)
   - Loyal buyers are more affected by the email, leading to addition $5.32 in spending

3. Frequent Visitors

Firstly, we plot the histogram for visits.

```
# plot purchase rates for 'visits'
hist(d$visits,
     breaks=50,
     xlab="Number of Website Visit", ylab="Customers",
     main="Histogram of Number of Website Visit",
     xlim = range(0,20))
```



We consider the customers who visit the website more than 5 times as **Frequent website visitors**.

```
# differentiate new versus older customers
d$Freq = (d$visits > 5)
sum(d$visits>5)

## [1] 37480
```

Because our test is big enough, we will have enough sample in the subgroup.

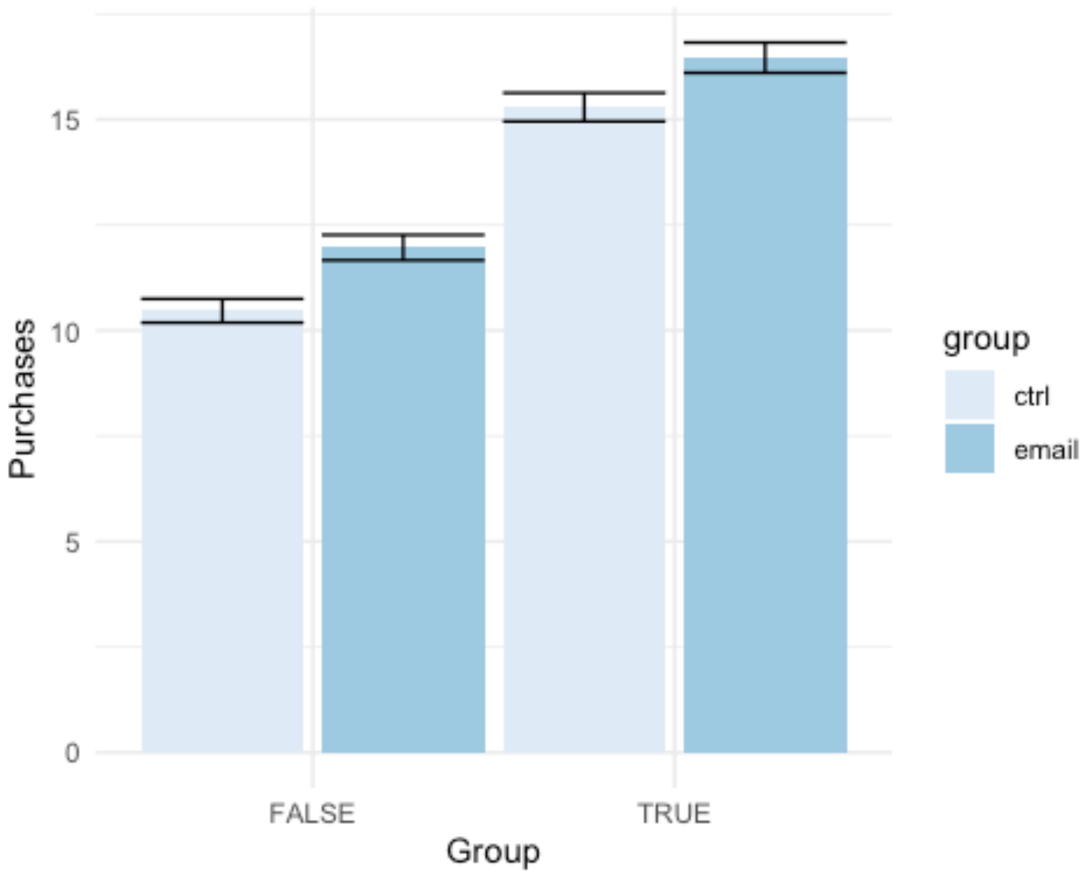
Frequent visitors vs. Infrequent visitors

```
dt = data.table(d)
dagg_freq = dt[,.(open = mean(open), click=mean(click), purch = mean(purch),
                  seOpen = sd(open)/sqrt(.N), seClick=sd(click)/sqrt(.N),
                  sePurch = sd(purch)/sqrt(.N),.N), #standard error
               by = .(group,Freq)] #condition
dagg_freq = setorder(dagg_freq,group,-Freq) #display the data table via group
name by order
dagg_freq

##    group  Freq      open     click    purch      seOpen     seClick
## 1:  ctrl  TRUE 0.0000000 0.0000000 15.29180 0.000000000 0.000000000
```

```
## 2:  ctrl FALSE 0.0000000 0.0000000 10.46647 0.000000000 0.000000000
## 3: email  TRUE 0.8272408 0.1442502 16.46249 0.002759703 0.002564823
## 4: email FALSE 0.7668465 0.1256989 11.96242 0.002961265 0.002321658
##      sePurch     N
## 1: 0.3371940 18714
## 2: 0.2819887 20442
## 3: 0.3592396 18766
## 4: 0.3008845 20390
```

- Frequent website visitors buy more on average
- The email seems to produce a stronger effect on purchases for infrequent buyers (~$1.5 versus $1.17)

Is email more effective for frequent visitors?



We can see that email is not more effective for frequent visitors.

Measuring causal effects with regression: Conditional causal effects

```
summary(lm(purch~group*Freq, data=d))

##
## Call:
## lm(formula = purch ~ group * Freq, data = d)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
##  -16.46  -15.29  -11.96  -10.47 1796.04 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           10.4665     0.3123  33.514  < 2e-16 ***
## groupemail             1.4960     0.4419   3.385 0.000712 ***
## FreqTRUE               4.8253     0.4517  10.682  < 2e-16 ***
## groupemail:FreqTRUE   -0.3253     0.6388  -0.509 0.610636    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 44.65 on 78308 degrees of freedom
## Multiple R-squared:  0.002943,   Adjusted R-squared:  0.002905 
## F-statistic: 77.05 on 3 and 78308 DF,  p-value: < 2.2e-16
```

```
p_freq = summary(lm(purch~group*Freq, data=d))$coefficient[,4]
p.adjust(p_freq, "bonferroni")
```

```
##         (Intercept)          groupemail            FreqTRUE 
##        6.541793e-244        2.849020e-03        5.175305e-26 
## groupemail:FreqTRUE 
##         1.000000e+00
```

The main effect of the email variable is significant (p-value=0.002), leading to $1.49 more sales for those who hasn't visited our website for over 5 times, indicating this group of customers are significantly affected by the email.

However, the difference of effect from email compaign between frequent and infrequent visitors are not significant at all (p-value=1). Therefore, `visits` may not be a good example for slicing and dicing.

Part C: Causal Forest

Now we will use machine learning to estimate the causal effect at the individual level. The method we apply is **causal forest**. Because Causal forests are an *alternative to regression* for identifying heterogeneous treatment effects and scoring customers based on predicted treatment effect uplift. Moreover, **causal forest** has the following advantages:

- Works well with a large number of baseline variables
- Doesn't require the analyst to define cut-offs for continuous baseline variables
- Will fit non-linear relationships between baseline variables and uplift

```
set.seed(22)
treatment <- (d$group == "email")*1
target <- d$purch
baseline <- d[c("last_purch", "visits", "chard", "sav_blanc", "syrah", "cab")
]
```

```r
# Time the training process
start = proc.time()
cf <- causal_forest(X=baseline, Y=target, W=treatment)
proc.time() - start
```

```
##    user  system elapsed
## 633.969  19.059 216.494
```

```r
print(cf)
```

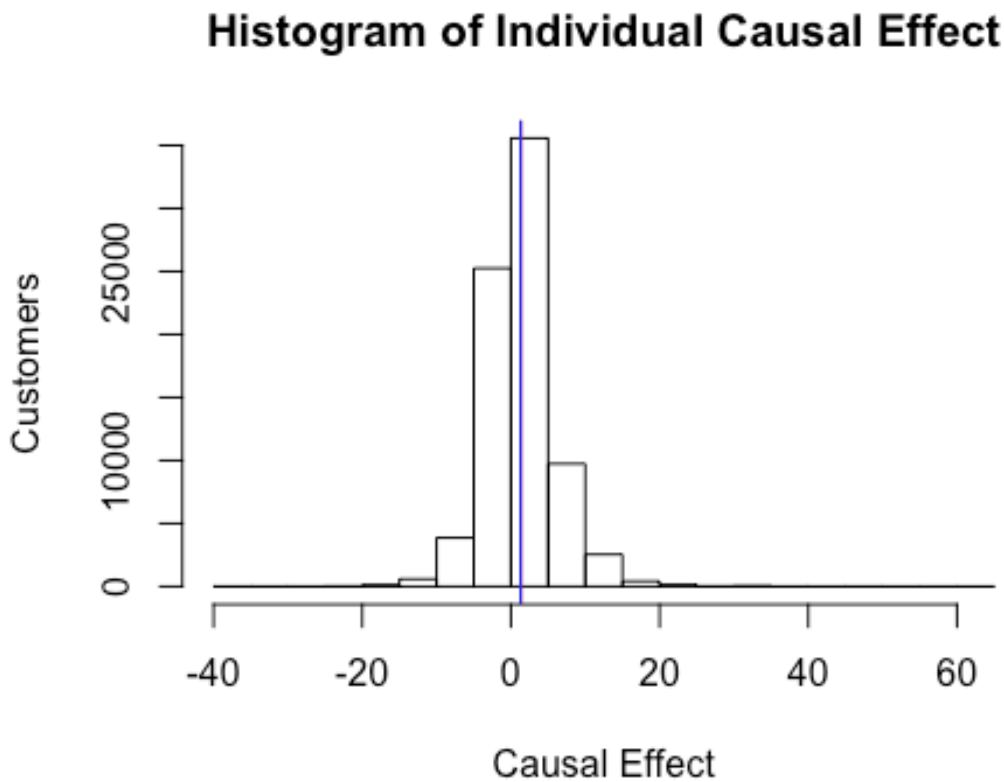
```
## GRF forest object of type causal_forest
## Number of trees: 2000
## Number of training samples: 78312
## Variable importance:
##     1     2     3     4     5     6
## 0.232 0.054 0.278 0.243 0.062 0.132
```

With the trained model, we can make predictions for causal effects on all consumers in the dataset.

```r
d.sub = d[, c("last_purch", "visits", "chard", "sav_blanc", "syrah", "cab")]
start = proc.time() # Start timing
preds = predict(cf, d.sub, estimate.variance=TRUE)
proc.time() - start # End timing
```

```
##    user  system elapsed
## 157.509   7.294  57.235
```

```r
# Plot histogram
hist(preds$predictions,
     xlab="Causal Effect", ylab="Customers",
     main="Histogram of Individual Causal Effect",)
abline(v=coef(lm0)[2], col=2); abline(v=mean(preds$predictions), col=4)
```



The causal forest method predictes causal effect estimates for each individual in the dataset. The individual estimates vary widely as shown in the histogram.

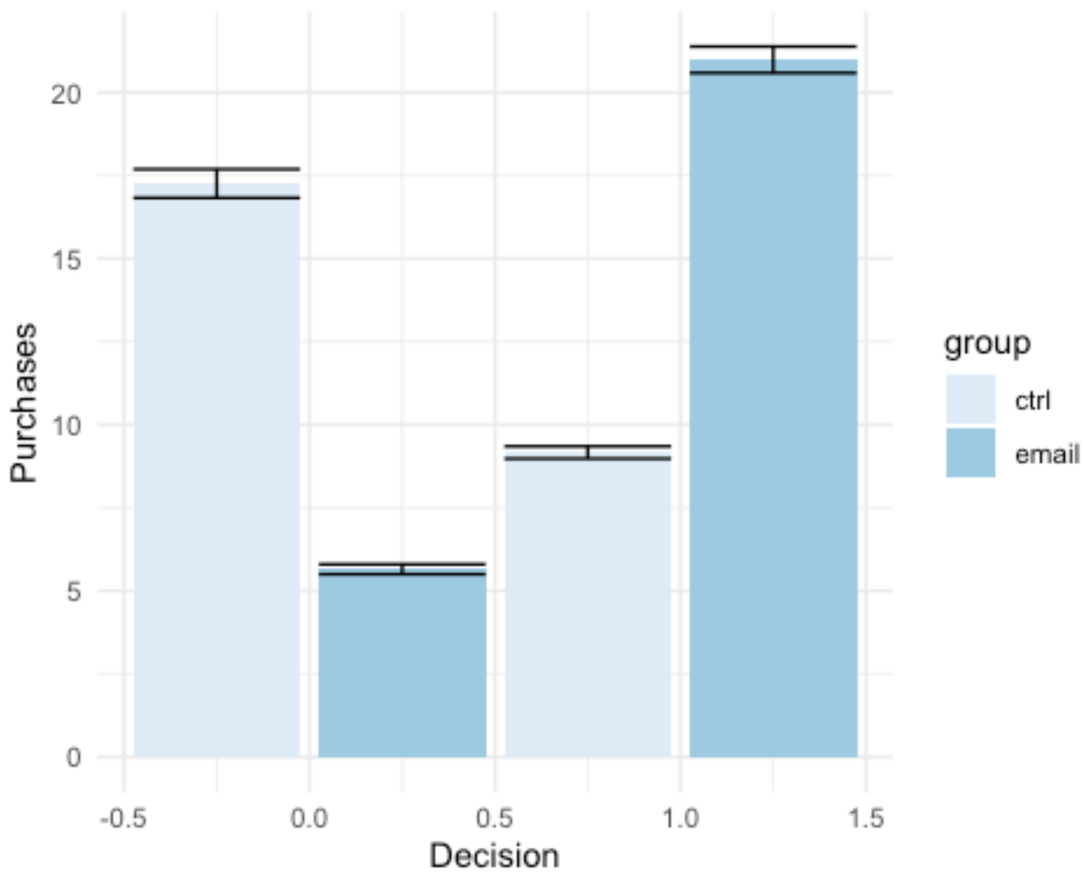
Now we will compute the score for each consumer. It is the profit we can gain by sending an email to a consumer subtracts the cost for sending an email. After computing the score, we can send emails to ones with a positive score. Because the causal effect estimates are the increases in purchased amount of consumers receiving an email, the gain is that increase multiply with the margin, which is 30% in this case. So, the formula to calculate the score for each customer is:

$$Score = \beta_1 \times 30\% - 0.1$$

```
preds$score = preds$predictions*0.3 - 0.1
preds$decision = (preds$score > 0)*1
table(preds$decision)

##
##     0     1
## 34987 43325
```

We will send emails to 43,325 consumers in our database. We can see that the causal effect is very clear on people that we decide to send an email to.

```r
d$decision = preds$decision
dt = data.table(d)

# Compare purchased amount between recent consumers and others
dagg = dt[, .(open = mean(open), click=mean(click), purch=mean(purch),
              seOpen=sd(open)/sqrt(.N), seClick=sd(click)/sqrt(.N),
              sePurch = sd(purch)/sqrt(.N),.N),
          by = .(group, decision)]

# Plot the difference
dodge = position_dodge(width=1); ##to form constant dimensions
ggplot(aes(fill=group, y=purch, x=decision,
           ymax=purch+sePurch, ymin=purch-sePurch),
       data=dagg) +
  geom_bar(position=dodge,stat="identity") +
  scale_fill_brewer(palette="Blues") +
  geom_errorbar(position=dodge) +
  labs(x="Decision", y="Purchases") +
  theme_minimal()
```



Below is the code to score new customers and making respective decision.

```
#### Code that generate score and targeting decisions for new data
# newdata <- data.frame(last_purch=xxx,visits=xxx,chard=xxx,sav_blanc=xxx,syr
ah=xxx,cab=xxx)
# pred <- predict(cf,newdata,estimate.variance=True)
# score <- pred[,1]*0.3 - 0.1
# desicion <- (score>0)
```

Finally, let's save our predictions for further exploratory analysis in Tableau.

```
write.csv(d, "full_data.csv")
write.csv(preds, "predictions.csv")
```