project

Keerthi Sreenivas Konjety, Vishnu Elangonvan

12/15/2021

```
mkt = read.csv("marketing_campaign.csv")
```

Few Exploratory Analysis: ## R Markdown

```
#Adding Age Column
mkt$Age = 2021-mkt$Year_Birth
```

Number of day customer enroll with the company (until 07/31/2021)

Calculate days difference between the day customer enrolled with the company and the day the author uploaded the data on Kaggle, 07/31/2021.

```
mkt$Dt_CustomerCovert1 = as.Date(mkt$Dt_Customer)
mkt$Dt_CustomerCovert2 = as.Date("2021-07-31") - as.Date(mkt$Dt_CustomerCovert1)
mkt$NumberofDayEnrolled = as.numeric(mkt$Dt_CustomerCovert2, units="days")
```

In our case, the most important information about a customer is only required to predict the amount spent on different product categories. So therefore all irrelevant variables have been excluded from our model.

```
#dropping columns that are not of our interest
mkt = subset(mkt, select = -c(Year_Birth,ID, Dt_Customer,NumDealsPurchases,Recency, AcceptedCmp3,Accept
#displaying the column names
colnames(mkt)
```

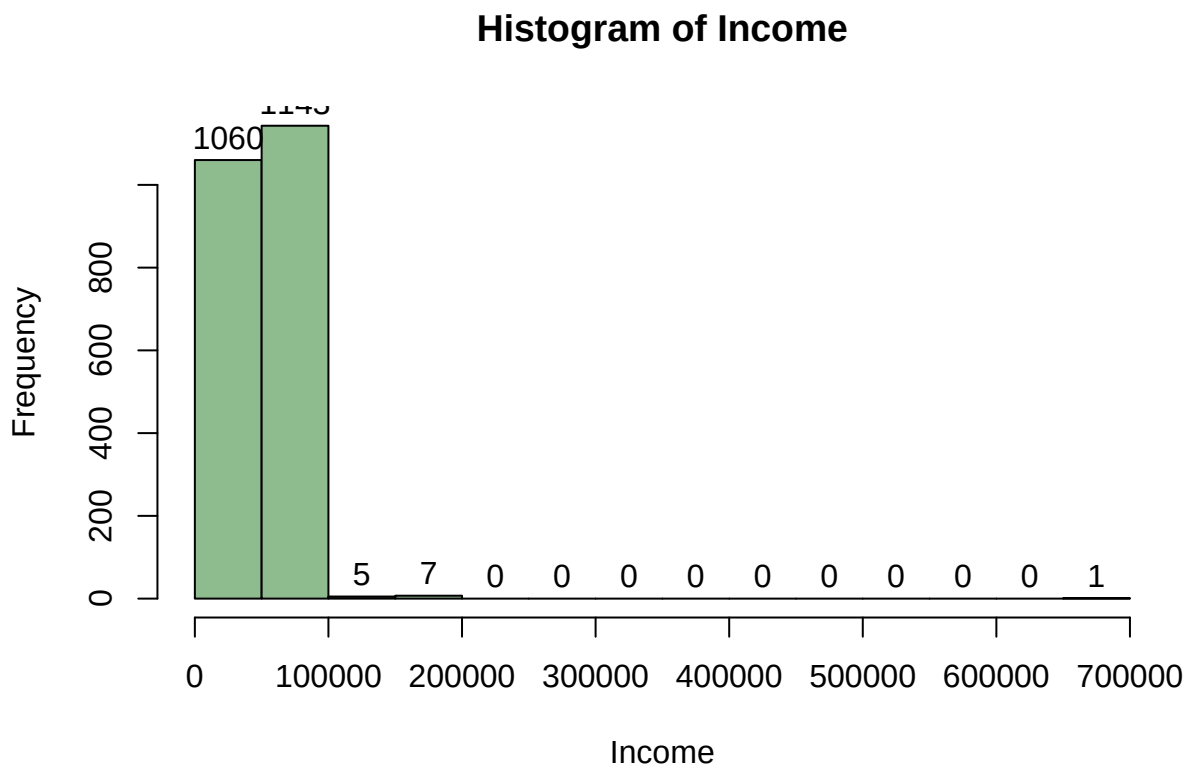
```
##  [1] "Education"           "Marital_Status"      "Income"
##  [4] "Kidhome"             "Teenhome"            "MntWines"
##  [7] "MntFruits"           "MntMeatProducts"     "MntFishProducts"
## [10] "MntSweetProducts"    "MntGoldProds"        "NumWebPurchases"
## [13] "NumCatalogPurchases" "NumStorePurchases"   "NumWebVisitsMonth"
## [16] "Age"                 "NumberofDayEnrolled"
```

```
## 1 - ID : 2 - Year_Birth
## 8 - Dt_Customer : 9 - Recency
## 16 - NumDealsPurchases : 25 - AcceptedCmp2
## 27 - Z_CostContact : 29 - Response
## 31 - Dt_CustomerCovert1 : 32 - Dt_CustomerCovert2
#mkt <- mkt[-c( 1:2 , 8:9 , 16:25 , 27:29 , 31:32 )]
#mkt
```

Next step is investigating the outliers , based on my dataset exploration, I can see extreme outliers in some variables , to better understand this , I am building histograms to see the distribution of each variable.

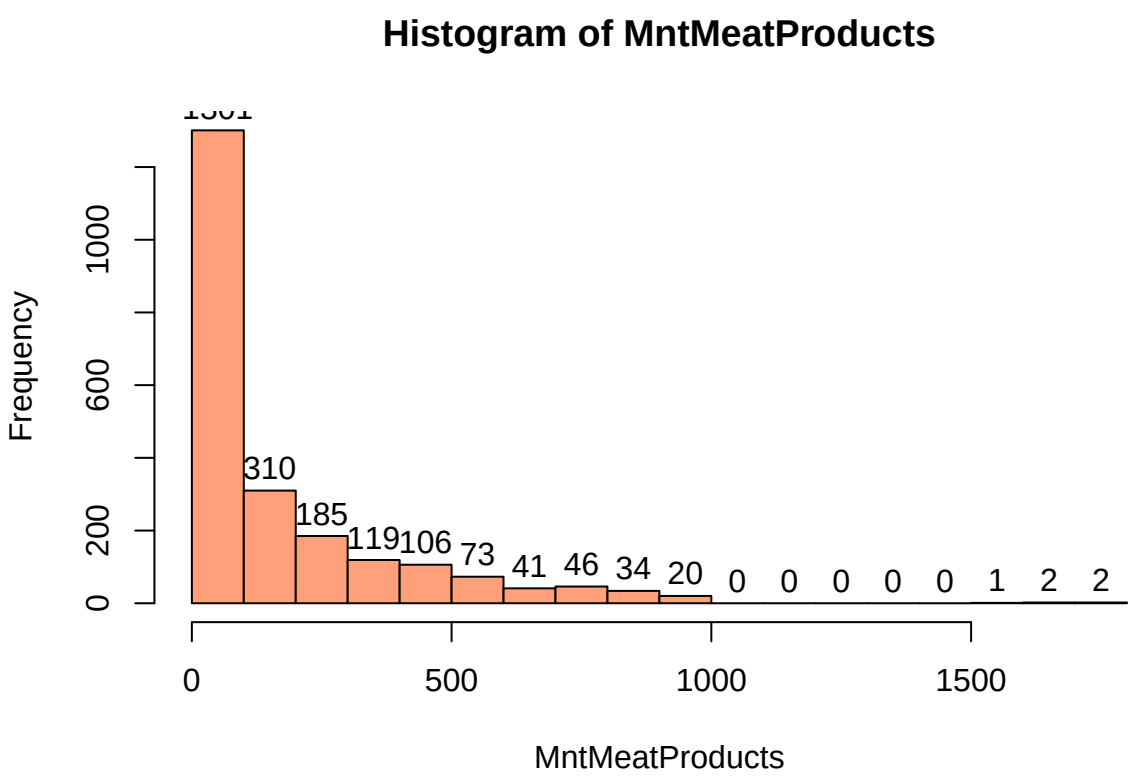
Income variable

```r
options(scipen = 100)
hist(mkt$Income,
     xlab = "Income",
     main = "Histogram of Income",
     col = "darkseagreen",
     breaks = 20,
     labels = TRUE)
```



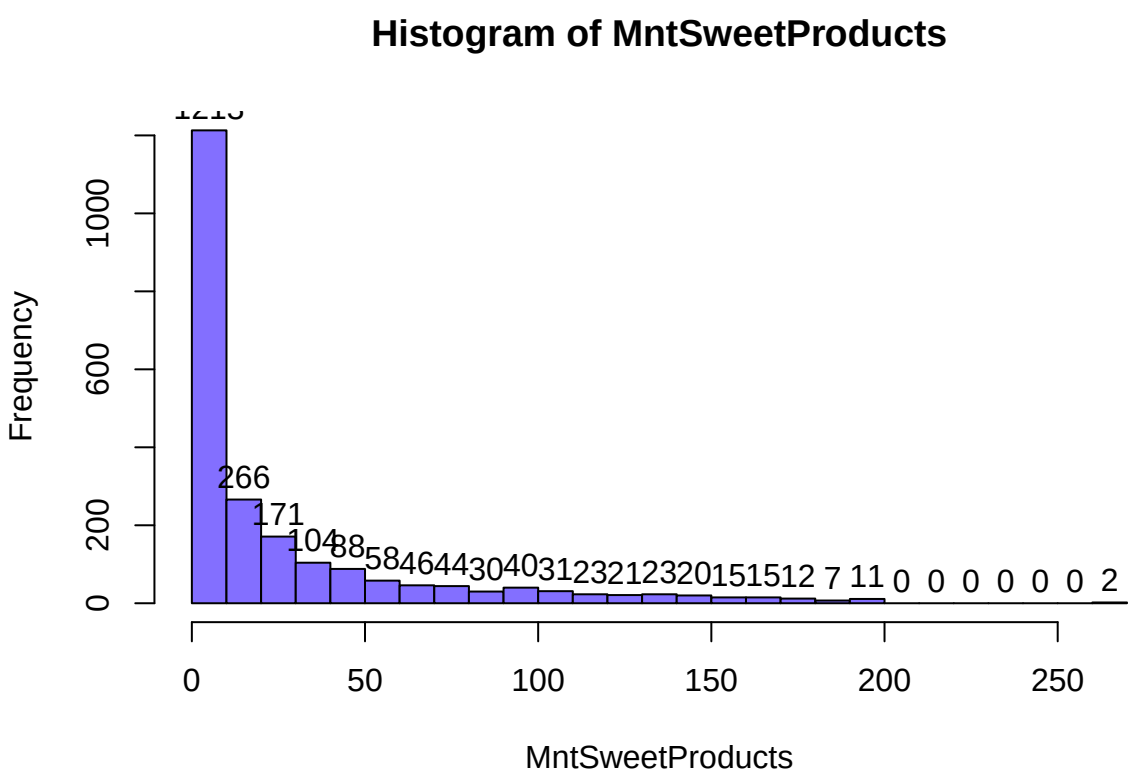
Amount spent on Meat Product

```r
options(scipen = 100)
hist(mkt$MntMeatProducts,
     xlab = "MntMeatProducts",
     main = "Histogram of MntMeatProducts",
     col = "lightsalmon",
     breaks = 20,
     labels = TRUE)
```

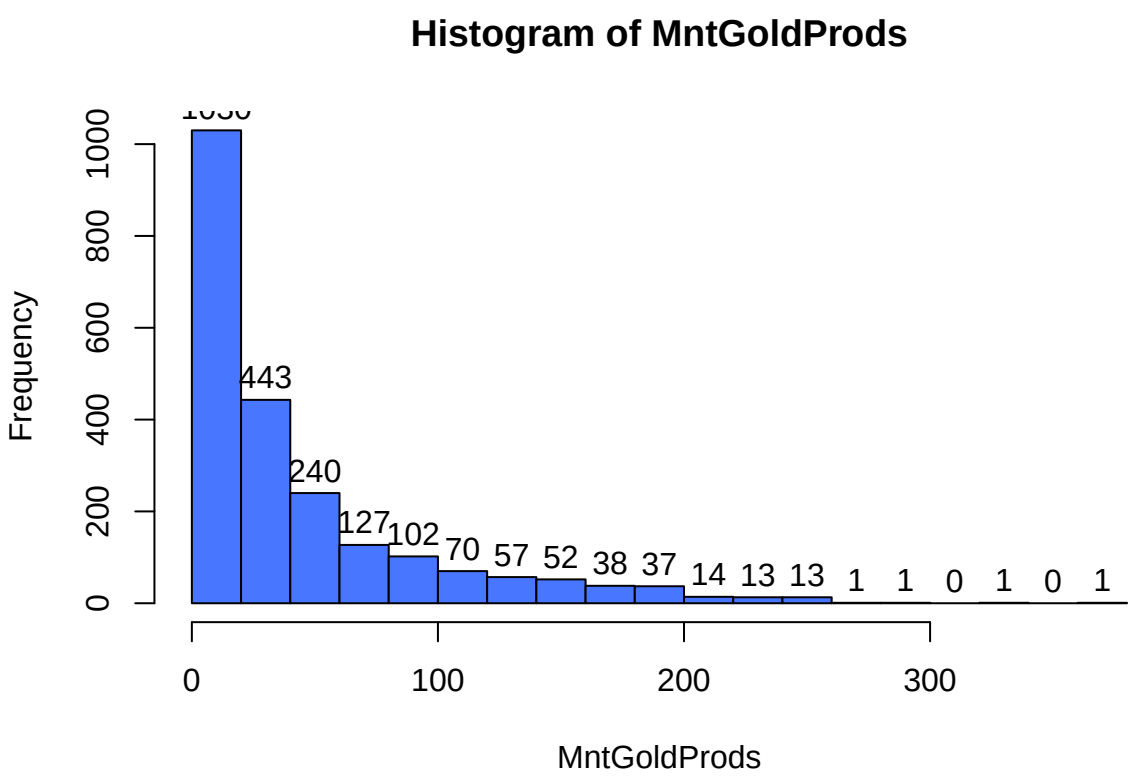


Amount spent on Sweet Product

```
options(scipen = 100)
hist(mkt$MntSweetProducts,
     xlab = "MntSweetProducts",
     main = "Histogram of MntSweetProducts",
     col = "slateblue1",
     breaks = 20,
     labels = TRUE)
```




```
options(scipen = 100)
hist(mkt$MntGoldProds,
     xlab = "MntGoldProds",
     main = "Histogram of MntGoldProds",
     col = "royalblue1",
     breaks = 20,
     labels = TRUE)
```



I am removing Outliers!!

```
mkt <- mkt[!(mkt$Income>150000 |
                              mkt$MntMeatProducts>1000 |
                              mkt$MntSweetProducts>200 |
                              mkt$MntGoldProds>260 ) , ]
```

Removing the missing data Since the number of rows with missing data In Income variable is 24, which accounted for only 1% of the dataset, it is safe to remove them.

```
mkt <- mkt[!(is.na(mkt$Income)),]
```

I want to treat Education,Marital_Status, and Complain as categorical variables and thus I use convert them to factors.

```
mkt$Education <- as.factor(mkt$Education)
mkt$Marital_Status <- as.factor(mkt$Marital_Status)
#mkt$Complain <- as.factor(mkt$Complain)
```

I want to see the frequency distribution of my Marital_Status column , so I do this below:

```
MarritalStatfreq <- data.frame(table(mkt$Marital_Status))
MarritalStatfreq[order(MarritalStatfreq$Freq, decreasing = TRUE),]
```

```
##       Var1 Freq
```

```
## 4  Married  852
## 6 Together  569
## 5   Single  468
## 3 Divorced  231
## 7    Widow   76
## 2    Alone    3
## 1   Absurd    2
## 8     YOLO    2
```

I am interested in seeing how many types I have in my Marital_Status column and I observe that there are some types with very less frequency. So I am checking if my marital_status lies in atleast 1% of population.

```
MarritalStatfreq[MarritalStatfreq$Freq / nrow(mkt) > .01, ]
```

```
##       Var1 Freq
## 3 Divorced  231
## 4  Married  852
## 5   Single  468
## 6 Together  569
## 7    Widow   76
```

There are eight statuses in the Marital Status variable. However, only five appear in at least 1% of the records. Therefore, we will combine the other three statuses into a group called "Others."

```
mkt$Marital_Status <- as.factor(ifelse(mkt$Marital_Status %in%
                                         c("Divorced", "Married", "Single","Together","Widow"),
                                       as.character(mkt$Marital_Status),
                                       "Other"))
MarritalStatfreq <- data.frame(table(mkt$Marital_Status))
MarritalStatfreq[order(MarritalStatfreq$Freq, decreasing = TRUE),]
```

```
##       Var1 Freq
## 2  Married  852
## 5 Together  569
## 4   Single  468
## 1 Divorced  231
## 6    Widow   76
## 3    Other    7
```

My idea is to create dummy vaariables for mt 3 categorical variables - Education, Marital_Status, and Complain.

```
library(fastDummies)
mkt <- dummy_cols( mkt,select_columns=c("Education", "Marital_Status","Complain_1"),remove_first_dummy='
```

```
## Warning in dummy_cols(mkt, select_columns = c("Education", "Marital_Status", : NOTE: The following s
##
```

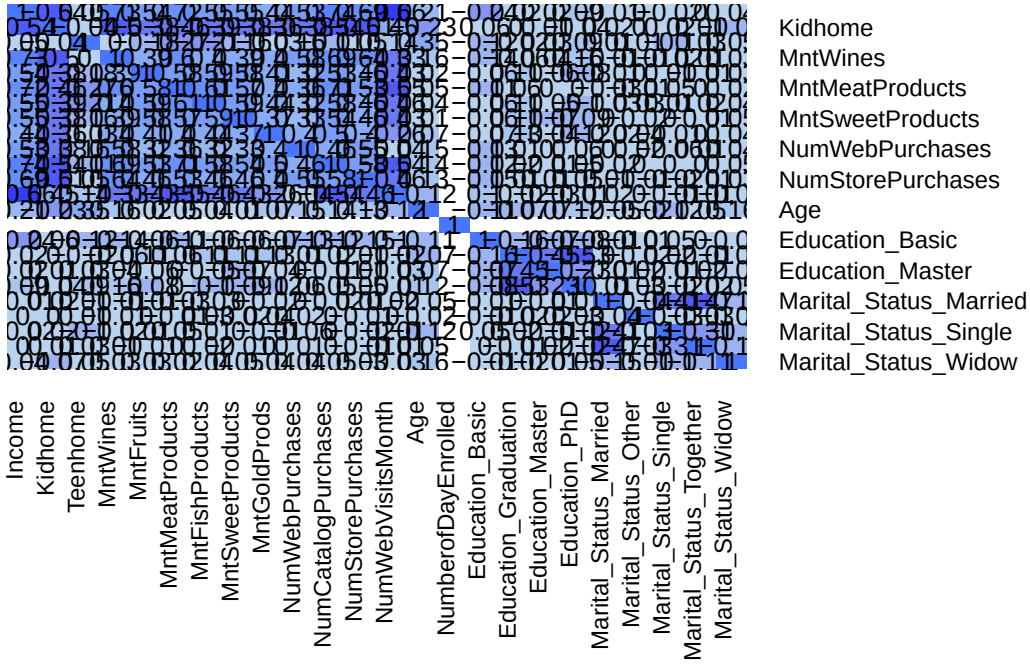
To examine the correlation between variables - A heat map is used to illustrate the correlation between variables. There are no pairs with correlation > 0.95, so we won't remove any variables.

```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
colfunc <- colorRampPalette(c("blue", "slategray2", "royalblue1"))
heatmap.2(cor(mkt),
          Rowv = FALSE, Colv = FALSE,
          dendrogram = "none",
          lwid=c(0.1,4), lhei=c(0.1,4), col = colfunc(15),
          cellnote = round(cor(mkt),2),
          notecol = "black",
          key = FALSE, trace = 'none', margins = c(15,15))
```



Multiple Regression - I am using 70% of my data for training, and 30% of data to evaluate my model.

```
mkt <- mkt[ ,-c(5:9)]

set.seed(14)
train.rows <- sample(rownames(mkt), nrow(mkt)*0.7)
```

```
train.data <- mkt[train.rows , ]
valid.rows <- setdiff(rownames(mkt), train.rows)
valid.data <- mkt[valid.rows , ]
```

Full Model - This model will include all the variables. We will start with a model to predict the amount spent on Wines.

```
customer.full.lm <- lm( MntWines ~ . ,
                        data = train.data )
options( scipen = 999 )
sum.full <- summary(customer.full.lm)
sum.full
```

```
##
## Call:
## lm(formula = MntWines ~ ., data = train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -782.98 -104.83  -17.39   92.14  896.50
##
## Coefficients:
##                            Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)            -3421.011070  4369.007049  -0.783              0.43392
## Income                     0.008799     0.000797  11.040 < 0.0000000000000002
## Kidhome                  -50.047602    18.831889  -2.658              0.00808
## Teenhome                 -45.934336    16.390031  -2.803              0.00523
## NumWebPurchases           14.422398     4.663535   3.093              0.00208
## NumCatalogPurchases       32.233199     4.422665   7.288    0.000000000000986
## NumStorePurchases          8.672165     3.550423   2.443              0.01487
## NumWebVisitsMonth         33.002711     6.223828   5.303    0.000000160251767
## Age                        0.296704     0.761067   0.390              0.69678
## NumberofDayEnrolled        0.003952     0.005938   0.666              0.50590
## Education_Basic          109.735835    59.596192   1.841              0.06606
## Education_Graduation      38.957065    25.897570   1.504              0.13303
## Education_Master          76.838545    30.517645   2.518              0.01206
## Education_PhD            120.101383    29.142781   4.121    0.000042959331208
## Marital_Status_Married   -40.747475    31.090816  -1.311              0.19049
## Marital_Status_Other     -93.984587   115.518847  -0.814              0.41620
## Marital_Status_Single    -20.690090    33.387673  -0.620              0.53569
## Marital_Status_Together  -22.811104    32.264662  -0.707              0.47984
## Marital_Status_Widow    -102.764229    46.179907  -2.225              0.02643
##
## (Intercept)
## Income                  ***
## Kidhome                 **
## Teenhome                **
## NumWebPurchases         **
## NumCatalogPurchases     ***
## NumStorePurchases       *
## NumWebVisitsMonth       ***
## Age
## NumberofDayEnrolled
```

```
## Education_Basic         .
## Education_Graduation
## Education_Master        *
## Education_PhD           ***
## Marital_Status_Married
## Marital_Status_Other
## Marital_Status_Single
## Marital_Status_Together
## Marital_Status_Widow    *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 191.8 on 605 degrees of freedom
##   (918 observations deleted due to missingness)
## Multiple R-squared:  0.6776, Adjusted R-squared:  0.668
## F-statistic: 70.65 on 18 and 605 DF,  p-value: < 0.00000000000000022
```

To see how well the model performs, we assess its performance on the validation data. By comparing the predicted value with the actual value, we can calculate the residual and the error (ME, RMSE, MAE). We can also calculate R squared, adjusted R squared, AIC, and BIC value of the model.

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
valid.full.lm.pred <- predict(customer.full.lm, valid.data)
options(scipen = 999, digits = 1)
valid.resid <- valid.data$MntWines - valid.full.lm.pred

data.frame( "Predicted" = valid.full.lm.pred[1:15],
            "Actual" = valid.data$MntWines[1:15],
            "Residual" = valid.resid[1:15])
```

```
##    Predicted Actual Residual
## 7         NA    235       NA
## 14       -53      3       56
## 16        NA     53       NA
## 18        NA   1012       NA
## 20       250     86     -164
## 24        10      6       -4
## 32        NA    482       NA
## 33        NA     40       NA
## 34        NA    702       NA
## 37       734    437     -297
## 42        NA    123       NA
## 45       631    826      195
## 49       707    510     -197
## 55       590    712      122
## 57        NA    523       NA
```

```
options(digits = 6)
accuracy(valid.full.lm.pred, valid.data$MntWines)
```

```
##                 ME   RMSE     MAE MPE MAPE
## Test set -16.3783 207.49 144.294 Inf  Inf
```

```
sum.full$r.squared
```

```
## [1] 0.677622
```

```
sum.full$adj.r.squared
```

```
## [1] 0.668031
```

```
AIC(customer.full.lm)
```

```
## [1] 8351.73
```

```
BIC(customer.full.lm)
```

```
## [1] 8440.46
```

The Forward Selection model begins with no variables and adds one predictors at a time. Each predictor added is the one (among all remaining predictors) contributes the most to R squared on top of the predictors that have already been added to the model. When the contribution of additional predictors is no longer statistically significant, the model will stop adding predictors.

Initial Baseline Model

```
customer.lm<- lm(MntWines ~ .,data = train.data)
customer.lm.null<- lm(MntWines ~ 1,data = train.data)
```

b. Build model

```
customer.lm.fwd <- step(customer.lm.null,scope =list(customer.lm.null,upper =customer.lm),direction ="f
```

```
## Start:  AIC=17969.3
## MntWines ~ 1
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit
```

```
##                          Df Sum of Sq      RSS  AIC
## + Income                  1  38184921 30868353 6749
## + NumCatalogPurchases     1  32323158 36730115 6857
## + NumStorePurchases       1  23196186 45857088 6996
## + NumWebPurchases         1  23054613 45998660 6998
## + Kidhome                 1  17890818 51162456 7064
```

```
## + NumWebVisitsMonth        1  10515959 58537314 7148
## + Education_PhD            1   1930938 67122336 7234
## + Age                      1   1428799 67624475 7238
## + Education_Basic          1   1216243 67837030 7240
## + Marital_Status_Married   1    782670 68270603 7244
## <none>                                69053274 7249
## + Marital_Status_Single    1    205989 68847284 7249
## + Marital_Status_Widow     1    149072 68904202 7250
## + Education_Graduation     1     67853 68985420 7251
## + NumberofDayEnrolled      1     49996 69003278 7251
## + Marital_Status_Together  1     39861 69013412 7251
## + Teenhome                 1     32013 69021260 7251
## + Education_Master         1     17799 69035474 7251
## + Marital_Status_Other     1      6594 69046679 7251
##
## Step:  AIC=16710.9
## MntWines ~ Income
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit
```

```
##                           Df Sum of Sq      RSS  AIC
## + NumWebPurchases          1   3585978 27282375 6674
## + NumCatalogPurchases      1   3273738 27594615 6681
## + NumWebVisitsMonth        1   1900685 28967668 6711
## + NumStorePurchases        1    939802 29928551 6732
## + Kidhome                  1    938970 29929383 6732
## + Education_PhD            1    497907 30370446 6741
## + Education_Graduation     1    251653 30616700 6746
## + Teenhome                 1    185127 30683226 6747
## + Education_Basic          1    100503 30767850 6749
## <none>                                30868353 6749
## + Marital_Status_Widow     1     90108 30778245 6749
## + Marital_Status_Married   1     84988 30783365 6749
## + Marital_Status_Together  1     54390 30813963 6750
## + Marital_Status_Single    1     43256 30825097 6750
## + Education_Master         1     26056 30842297 6750
## + NumberofDayEnrolled      1     21378 30846975 6750
## + Marital_Status_Other     1     13700 30854653 6751
## + Age                      1      5742 30862611 6751
##
## Step:  AIC=16523.2
## MntWines ~ Income + NumWebPurchases
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit
```

```
##                           Df Sum of Sq      RSS  AIC
## + NumCatalogPurchases      1   2355670 24926705 6619
## + Teenhome                 1    542767 26739608 6663
## + Kidhome                  1    472000 26810376 6665
## + Education_PhD            1    421558 26860817 6666
## + Education_Graduation     1    232920 27049455 6670
```

```
## + NumStorePurchases        1    219309 27063067 6671
## + NumWebVisitsMonth        1    168544 27113832 6672
## + Education_Basic          1    128487 27153888 6673
## + Marital_Status_Married   1    126428 27155947 6673
## + Marital_Status_Widow     1    122356 27160019 6673
## + Marital_Status_Single    1    108006 27174369 6673
## <none>                                 27282375 6674
## + Education_Master         1     55795 27226580 6675
## + Age                      1     44539 27237836 6675
## + Marital_Status_Other     1     41910 27240465 6675
## + Marital_Status_Together  1     25019 27257357 6675
## + NumberofDayEnrolled      1     22019 27260357 6675
##
## Step:  AIC=16395.2
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit

##                           Df Sum of Sq      RSS  AIC
## + NumWebVisitsMonth        1    629047 24297658 6606
## + Education_PhD            1    608287 24318418 6606
## + Education_Graduation     1    238646 24688059 6615
## + Marital_Status_Widow     1    157886 24768819 6618
## + Teenhome                 1    133452 24793253 6618
## + Kidhome                  1    104755 24821950 6619
## + NumStorePurchases        1     79834 24846871 6619
## <none>                                 24926705 6619
## + Marital_Status_Single    1     78979 24847726 6619
## + Marital_Status_Married   1     71023 24855682 6620
## + Education_Basic          1     57931 24868774 6620
## + Education_Master         1     42679 24884025 6620
## + NumberofDayEnrolled      1     31416 24895288 6621
## + Marital_Status_Other     1     21576 24905129 6621
## + Marital_Status_Together  1     13976 24912729 6621
## + Age                      1      9262 24917443 6621
##
## Step:  AIC=16335.7
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumWebVisitsMonth

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit

##                           Df Sum of Sq      RSS  AIC
## + Education_PhD            1    554902 23742756 6593
## + Kidhome                  1    329765 23967893 6599
## + Teenhome                 1    268210 24029448 6601
## + Education_Graduation     1    261510 24036148 6601
## + NumStorePurchases        1    248665 24048993 6601
## + Marital_Status_Widow     1    173645 24124013 6603
## + Education_Basic          1     86130 24211528 6605
## <none>                                 24297658 6606
## + Marital_Status_Single    1     68640 24229018 6606
```

```
## + Marital_Status_Married   1     60769 24236889 6606
## + Education_Master         1     35052 24262606 6607
## + NumberofDayEnrolled      1     29929 24267729 6607
## + Marital_Status_Together  1     18764 24278894 6607
## + Marital_Status_Other     1     15835 24281823 6607
## + Age                      1     10922 24286736 6607
##
## Step:  AIC=16298.1
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumWebVisitsMonth +
##     Education_PhD
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit
```

```
##                           Df Sum of Sq      RSS  AIC
## + Kidhome                  1    305408 23437347 6587
## + Teenhome                 1    298034 23444721 6587
## + NumStorePurchases        1    283885 23458871 6588
## + Marital_Status_Widow     1    201427 23541329 6590
## + Education_Master         1    135507 23607249 6592
## + Education_Basic          1    105757 23636998 6592
## <none>                                 23742756 6593
## + Marital_Status_Married   1     52236 23690520 6594
## + Marital_Status_Single    1     48241 23694515 6594
## + NumberofDayEnrolled      1     33922 23708834 6594
## + Marital_Status_Together  1     25568 23717187 6594
## + Education_Graduation     1     21191 23721565 6595
## + Age                      1     17726 23725030 6595
## + Marital_Status_Other     1      9168 23733587 6595
##
## Step:  AIC=16271.4
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumWebVisitsMonth +
##     Education_PhD + Kidhome
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit
```

```
##                           Df Sum of Sq      RSS  AIC
## + Teenhome                 1    360565 23076782 6579
## + NumStorePurchases        1    216503 23220844 6583
## + Marital_Status_Widow     1    215746 23221601 6583
## + Education_Master         1    141818 23295529 6585
## + Education_Basic          1     88234 23349113 6587
## <none>                                 23437347 6587
## + Marital_Status_Married   1     44131 23393216 6588
## + Age                      1     42472 23394875 6588
## + Marital_Status_Single    1     34573 23402774 6588
## + Marital_Status_Together  1     31499 23405849 6588
## + NumberofDayEnrolled      1     27510 23409838 6588
## + Education_Graduation     1     19429 23417919 6589
## + Marital_Status_Other     1      5151 23432196 6589
##
## Step:  AIC=16259.4
```

```
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumWebVisitsMonth +
##     Education_PhD + Kidhome + Teenhome

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit

##                           Df Sum of Sq      RSS  AIC
## + NumStorePurchases        1    207156 22869627 6576
## + Marital_Status_Widow     1    174211 22902571 6577
## + Education_Master         1    136374 22940408 6578
## <none>                                 23076782 6579
## + Education_Basic          1     61049 23015733 6580
## + Marital_Status_Married   1     48991 23027791 6580
## + Marital_Status_Together  1     36039 23040743 6580
## + Marital_Status_Single    1     26148 23050634 6581
## + NumberofDayEnrolled      1     19667 23057115 6581
## + Marital_Status_Other     1      8280 23068502 6581
## + Education_Graduation     1      7491 23069291 6581
## + Age                      1      1521 23075261 6581
##
## Step:  AIC=16233.2
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumWebVisitsMonth +
##     Education_PhD + Kidhome + Teenhome + NumStorePurchases

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit

##                           Df Sum of Sq      RSS  AIC
## + Education_Master         1    163348 22706279 6573
## + Marital_Status_Widow     1    156319 22713308 6573
## <none>                                 22869627 6576
## + Education_Basic          1     63977 22805650 6576
## + Marital_Status_Married   1     58589 22811038 6576
## + Marital_Status_Together  1     39801 22829826 6577
## + Marital_Status_Single    1     19033 22850594 6577
## + Education_Graduation     1     14376 22855251 6577
## + NumberofDayEnrolled      1     12432 22857195 6577
## + Marital_Status_Other     1      5727 22863900 6578
## + Age                      1      5093 22864534 6578
##
## Step:  AIC=16219.3
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumWebVisitsMonth +
##     Education_PhD + Kidhome + Teenhome + NumStorePurchases +
##     Education_Master

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit

##                           Df Sum of Sq      RSS  AIC
## + Marital_Status_Widow     1    147243 22559035 6571
## + Education_Basic          1     82455 22623823 6573
## <none>                                 22706279 6573
```

```
## + Marital_Status_Married   1     52415 22653863 6574
## + Education_Graduation     1     40504 22665774 6574
## + Marital_Status_Together  1     32870 22673408 6574
## + Marital_Status_Single    1     23877 22682402 6575
## + NumberofDayEnrolled      1     13378 22692901 6575
## + Marital_Status_Other     1     12494 22693785 6575
## + Age                      1       528 22705750 6575
##
## Step:  AIC=16219
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumWebVisitsMonth +
##     Education_PhD + Kidhome + Teenhome + NumStorePurchases +
##     Education_Master + Marital_Status_Widow
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 624/1542 rows from a combined fit
```

```
##                            Df Sum of Sq      RSS  AIC
## + Marital_Status_Married    1     89115 22469920 6571
## + Education_Basic          1     81731 22477305 6571
## <none>                                 22559035 6571
## + Education_Graduation     1     39178 22519857 6572
## + Marital_Status_Together  1     17181 22541855 6573
## + Marital_Status_Other     1     13301 22545735 6573
## + Marital_Status_Single    1     12042 22546993 6573
## + NumberofDayEnrolled      1     11414 22547622 6573
## + Age                      1      7623 22551413 6573
##
## Step:  AIC=16220.5
## MntWines ~ Income + NumWebPurchases + NumCatalogPurchases + NumWebVisitsMonth +
##     Education_PhD + Kidhome + Teenhome + NumStorePurchases +
##     Education_Master + Marital_Status_Widow + Marital_Status_Married
```

```
sum.forward <-summary(customer.lm.fwd )
sum.forward
```

```
##
## Call:
## lm(formula = MntWines ~ Income + NumWebPurchases + NumCatalogPurchases +
##     NumWebVisitsMonth + Education_PhD + Kidhome + Teenhome +
##     NumStorePurchases + Education_Master + Marital_Status_Widow +
##     Marital_Status_Married, data = train.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -786.2 -108.4  -12.8   80.8  966.5
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           -506.78380   33.31847  -15.21 < 0.0000000000000002 ***
## Income                   0.00853    0.00049   17.43 < 0.0000000000000002 ***
## NumWebPurchases         13.73677    2.80597    4.90      0.00000108399816 ***
## NumCatalogPurchases     31.48544    2.86387   10.99 < 0.0000000000000002 ***
## NumWebVisitsMonth       34.68396    3.44871   10.06 < 0.0000000000000002 ***
```

```
## Education_PhD            93.00028   12.46438    7.46     0.00000000000014 ***
## Kidhome                 -60.37486   11.92643   -5.06     0.00000046445176 ***
## Teenhome                -36.72654    9.60504   -3.82              0.00014 ***
## NumStorePurchases        12.34059    2.28580    5.40     0.00000007766427 ***
## Education_Master         53.85718   13.60665    3.96     0.00007899419360 ***
## Marital_Status_Widow    -45.44271   28.05389   -1.62              0.10547
## Marital_Status_Married   -6.78782   10.13726   -0.67              0.50322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 192 on 1530 degrees of freedom
## Multiple R-squared:  0.683,  Adjusted R-squared:  0.681
## F-statistic:  299 on 11 and 1530 DF,  p-value: <0.0000000000000002
```

Assess performance on validation data

```
library(forecast)
valid.fwd.pred <- predict(customer.lm.fwd, valid.data)
options(digits = 6)
accuracy(valid.fwd.pred, valid.data$MntWines)  # performance of variable selection
```

```
##                ME    RMSE     MAE MPE MAPE
## Test set -10.376 190.674 137.456 Inf  Inf
```

```
sum.forward$r.squared
```

```
## [1] 0.682847
```

```
sum.forward$adj.r.squared
```

```
## [1] 0.680567
```

```
AIC(customer.lm.fwd)
```

```
## [1] 20598.5
```

```
BIC(customer.lm.fwd)
```

```
## [1] 20667.9
```