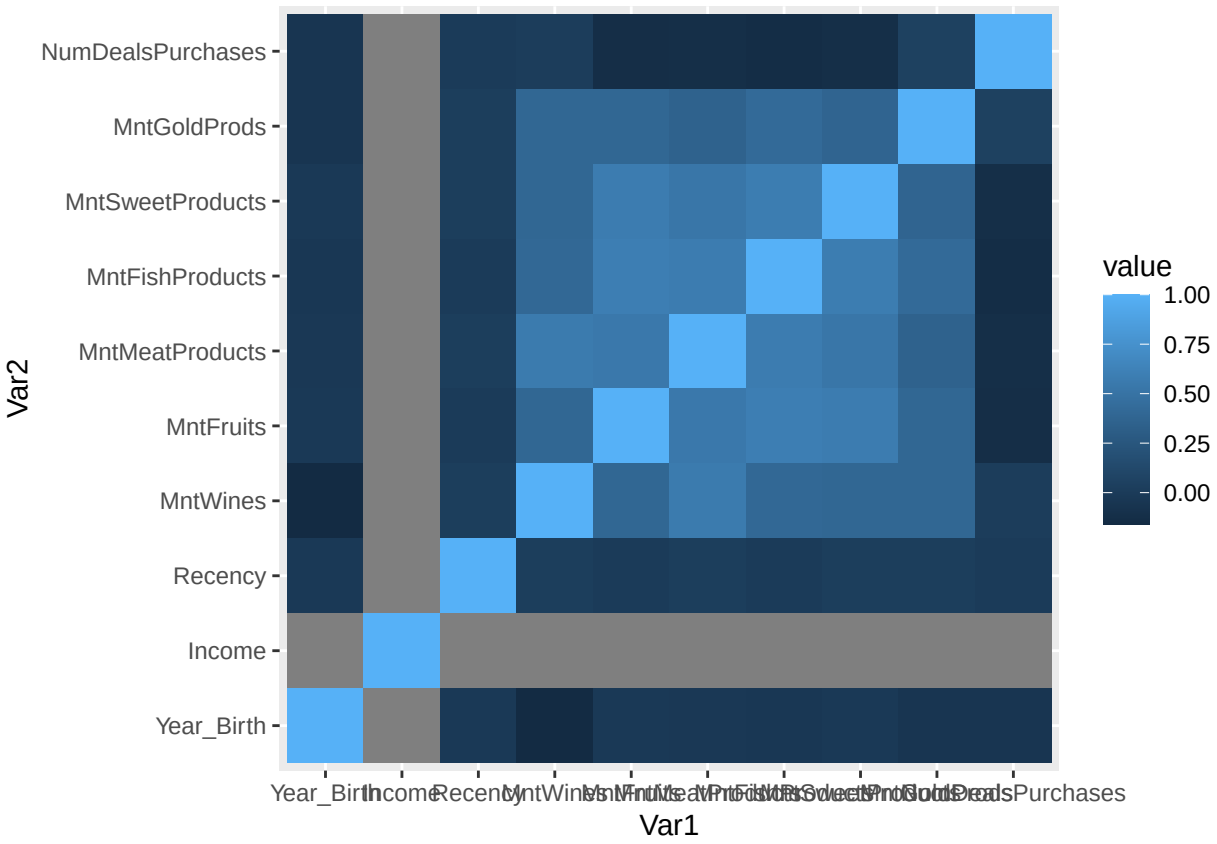
# Relationship between Amount Spent on Wine and Other Aspects - Checkpoint 3

2021/12/7

**This document is made for displaying the code we used, and the formal final report is the other file.



```
##              Var1       Var2 value
## 1      Year_Birth Year_Birth  1.00
## 2          Income Year_Birth    NA
## 3         Recency Year_Birth -0.02
## 4        MntWines Year_Birth -0.16
## 5       MntFruits Year_Birth -0.02
## 6 MntMeatProducts Year_Birth -0.03
```

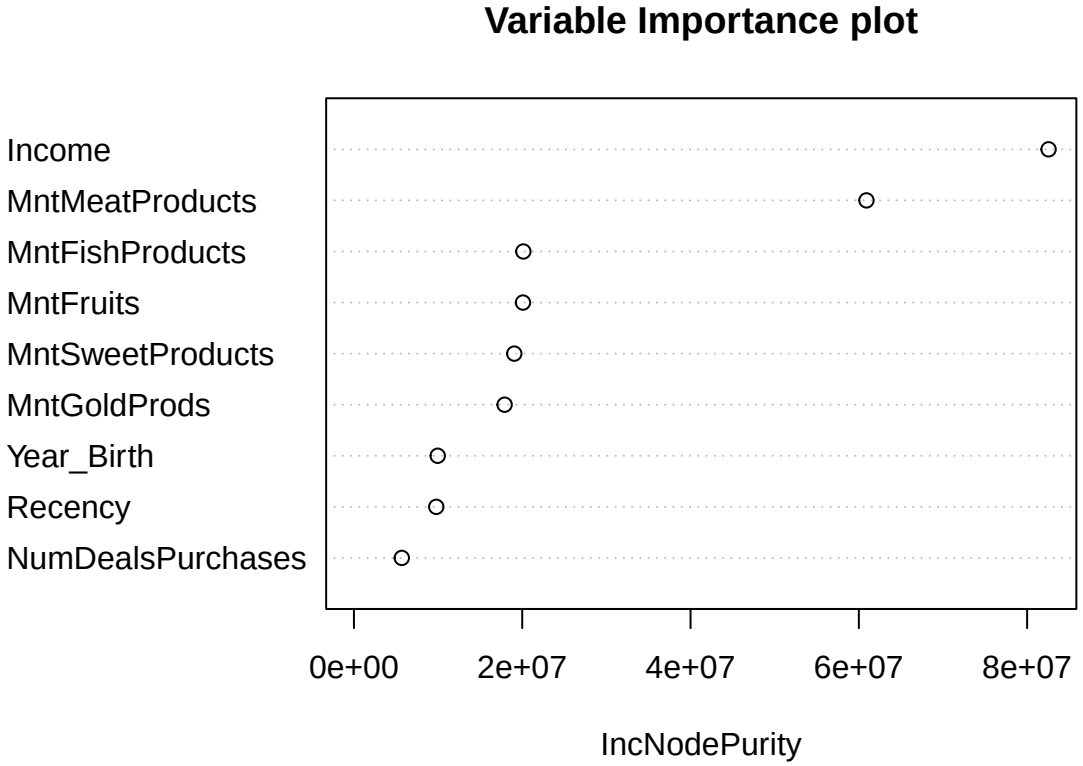



```
##
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------------
##                                 Added/                   Adj.
## Step        Variable            Removed     R-Square     R-Square      C(p)         AIC          RMSE
## ----------------------------------------------------------------------------------------------------
##    1         Income             addition       0.335        0.335    782.3200    31189.1438    275.1787
##    2     MntMeatProducts        addition       0.416        0.415    420.7530    30904.2891    257.9906
##    3        Kidhome             addition       0.460        0.459    224.5120    30732.6756    248.1359
##    4       Education            addition       0.481        0.479    131.0650    30651.9723    243.4392
##    5      MntGoldProds          addition       0.501        0.499     42.8000    30566.6858    238.7457
##    6    NumDealsPurchases       addition       0.511        0.509      1.4850    30525.5484    236.4868
##    7    MntSweetProducts        addition       0.511        0.509      0.0260    30524.0667    236.3547
## ----------------------------------------------------------------------------------------------------
```

It seems like "MntSweetProducts" has the smallest risk.
We apply best subsets regression method, and focus on $Cp$ and $AIC$, and temporarily ignore the other columns in the output , also the off-screen part.

```
##
## Call:
##  randomForest(formula = MntWines ~ ., data = newwine, ntree = ntrees,      type = classification, na
##                Type of random forest: regression
##                      Number of trees: 200
## No. of variables tried at each split: 3
```

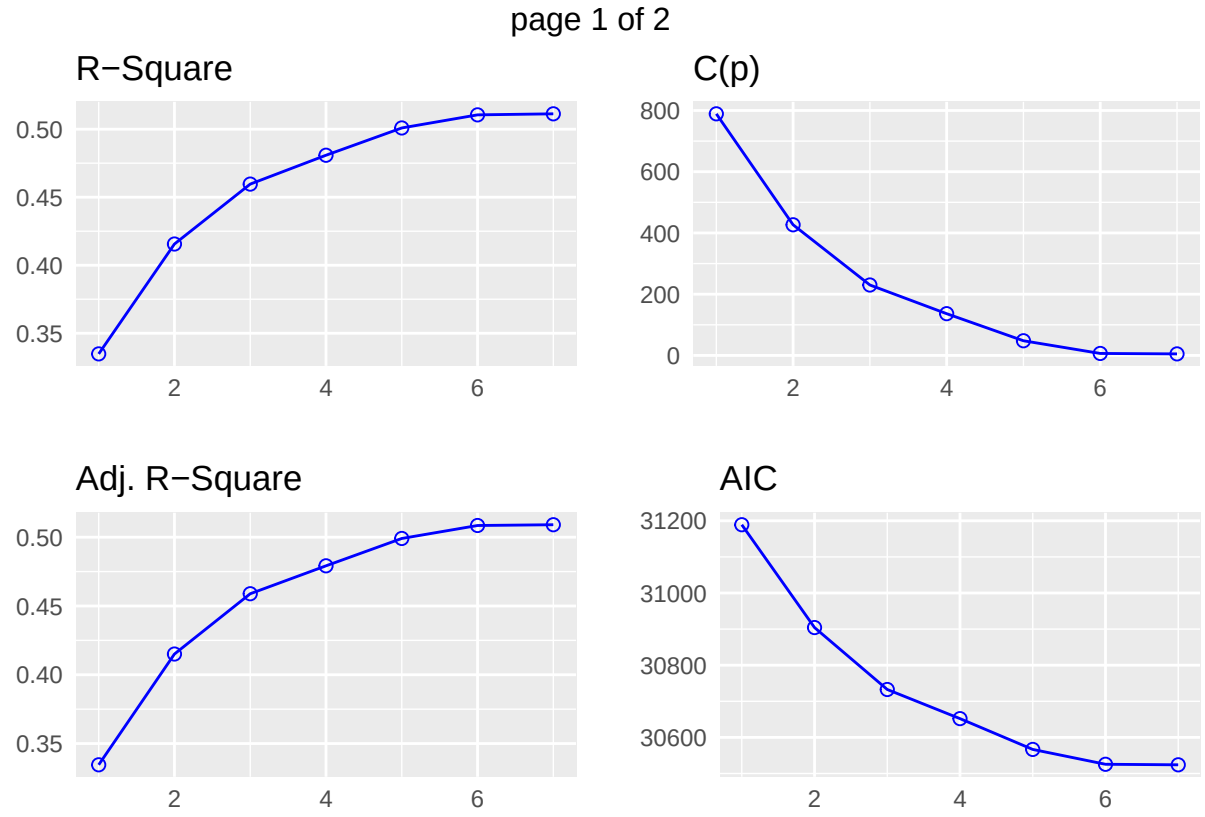
```
##
##           Mean of squared residuals: 35427.01
##                     % Var explained: 68.85
```

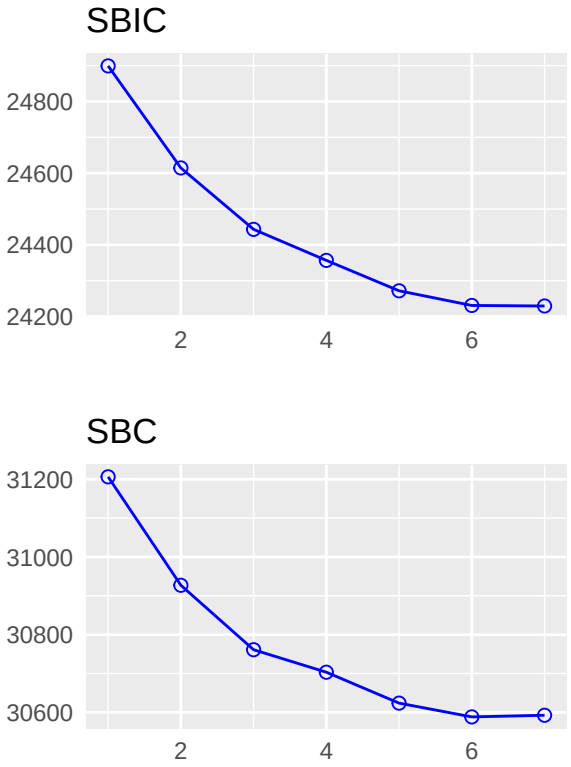



```
##                                     Best Subsets Regression
## ---------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------------------------------------------------------
##      1         Income
##      2         Income MntMeatProducts
##      3         Income MntMeatProducts Kidhome
##      4         Income MntMeatProducts Kidhome Education
##      5         Income MntMeatProducts Kidhome Education MntGoldProds
##      6         Income MntMeatProducts Kidhome Education MntGoldProds NumDealsPurchases
##      7         Income MntMeatProducts Kidhome Education MntGoldProds NumDealsPurchases MntSweetProdu
## ---------------------------------------------------------------------------------------------------------
##
##                                                  Subsets Regression Summary
## ---------------------------------------------------------------------------------------------------------
##                       Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)         AIC           SBIC          SBC
## ---------------------------------------------------------------------------------------------------------
##   1        0.3348      0.3345      0.2447    789.0891    31189.1438    24899.2249    31206.2541    1
##   2        0.4156      0.4151      0.3264    426.7002    30904.2891    24614.5453    30927.1030    1
##   3        0.4596      0.4589      0.3881    230.0115    30732.6756    24443.1949    30761.1929    1
##   4        0.4808      0.4792       0.413    136.3489    30651.9723    24356.6915    30703.3035    1
##   5        0.5009      0.4991      0.4399     47.8790    30566.6858    24271.7585    30623.7204    1
##   6        0.5105      0.5085      0.4462      6.4671    30525.5484    24230.8605    30588.2864    1
##   7        0.5113      0.5091      0.4447      5.0000    30524.0667    24229.4109    30592.5083    1
## ---------------------------------------------------------------------------------------------------------
```

```
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

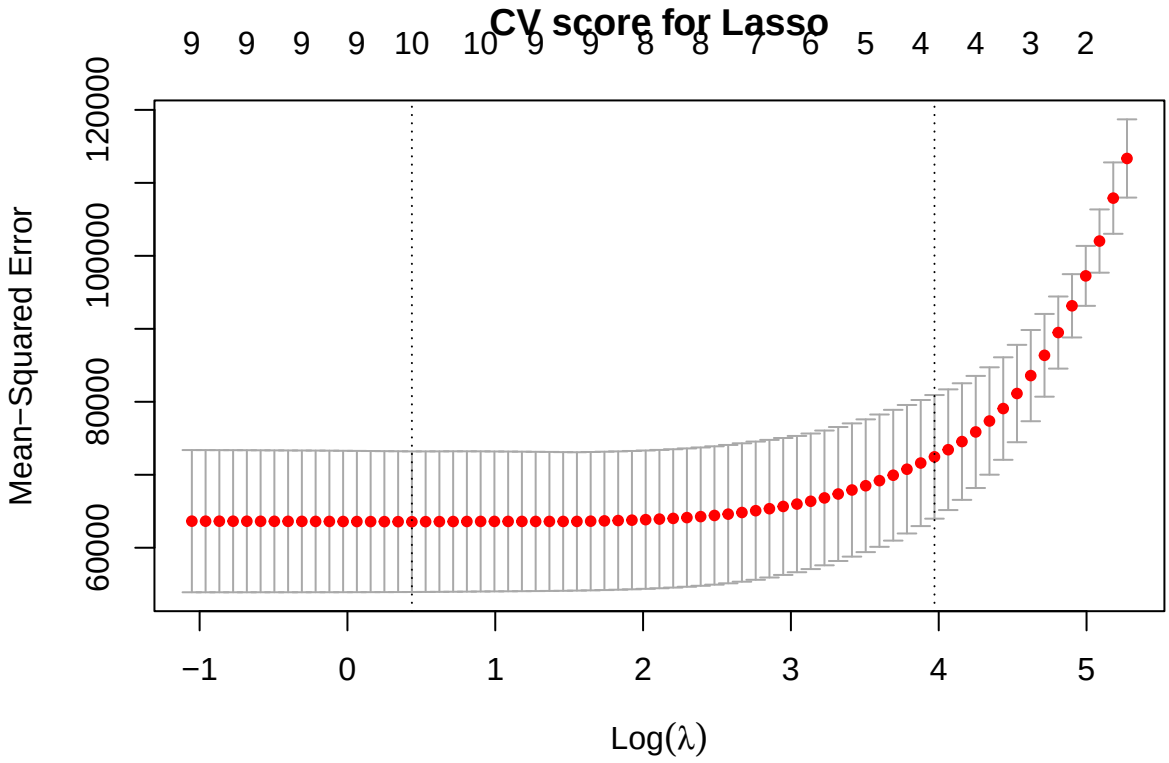
Next, we plot risk against different possible models (in the table above) as follows:






We conclude that a combination of "Income", "MntMeatProducts", "Kidhome", "Education", "MntGoldProds", "NumDealsPurchases", and "MntSweetProducts", is the best choice due to its lowest $Cp$ and $AIC$.

Below, we use the selected variables above to fit a ordinary linear model, and analyze it by shrinkage methods Ridge and Lasso.


CV score for Lasso
9 9 9 9 10 10 9 9 8 8 7 6 5 4 4 3 2
Mean−Squared Error
60000
80000
100000
120000
−1
0
1
2
3
4
5
Log(λ)