Relationship between Amount Spent on Wine and Other Aspects

2021/11/9

Group Members:

Jiapan Wang #29322674
Hanliang Liu #86009776
Catherine Cai #52204310
Jack Guo #57752750

Motivation:

Last month there was a tragic car accident near UBC campus late at night, which might have been caused by a possibly drunk driver. This dataset provides the amount spent on wine by customers and other aspects, for example, Birth Year and the Income of customers. It would give help to know the relationship between predictor variables and purchasing power in future careers.

Introduction:

In the last 70 years, the development of economics has recovered from the second World War. The dramatic increase in this period was never observed in history before due to the rapid development of industry and globalization. Meanwhile, the consumption power of citizens has increased with the development of economics. Therefore, we are interested in finding which set of factors has a great impact on the consumption power of customers, and in this study we will focus on the amount spent on wine, which represents the consumption power of customers. Our study is observational in nature. It is unfortunately not possible to randomly allocate experiment groups where some countries allocate more resources on certain factors while other countries do not.

We have chosen the "Customer Personality Analysis" hosted on Kaggle.com. This dataset is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers. The response variable chosen from this dataset is MntWines, amount spent on wine in the last 2 years, which is measured in dollars.

Data analysis:

The description of variables we may use in analysis

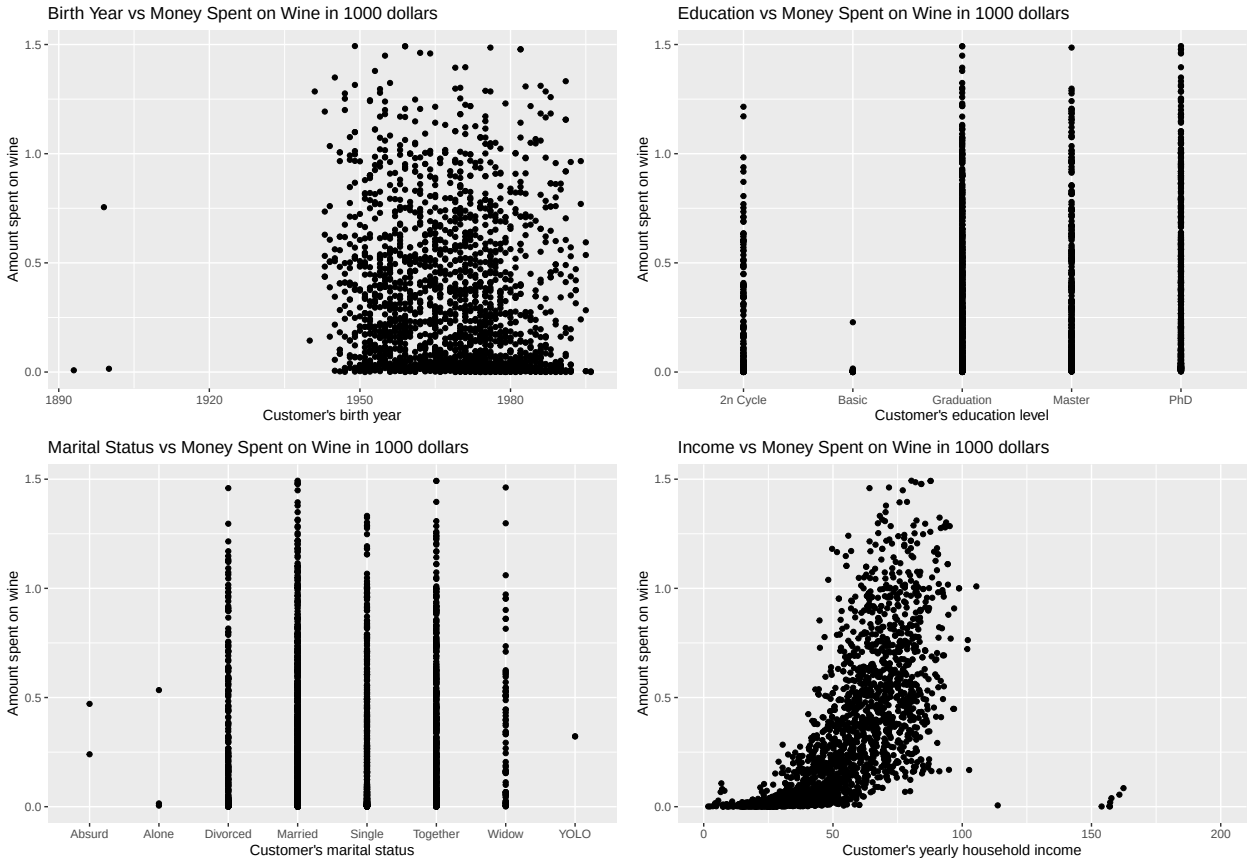
ID: Customer's unique identifier
Year_Birth: Customer's birth year
Education: Customer's education level
Marital_Status: Customer's marital status
Income: Customer's yearly household income
Kidhome: Number of children in customer's household

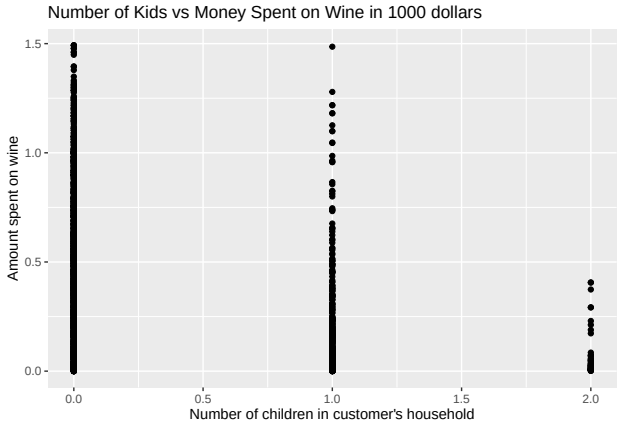
Teenhome: Number of teenagers in customer's household
Recency: Number of days since customer's last purchase
Complain: 1 if customer complained in the last 2 years, 0 otherwise
MntFruits: Amount spent on fruits in last 2 years
MntMeatProducts: Amount spent on meat in last 2 years
MntFishProducts: Amount spent on fish in last 2 years
MntSweetProducts: Amount spent on sweets in last 2 years
MntGoldProds: Amount spent on gold in last 2 years
MntWines: Amount spent on wine in last 2 years
NumDealsPurchases: Number of purchases made with a discount

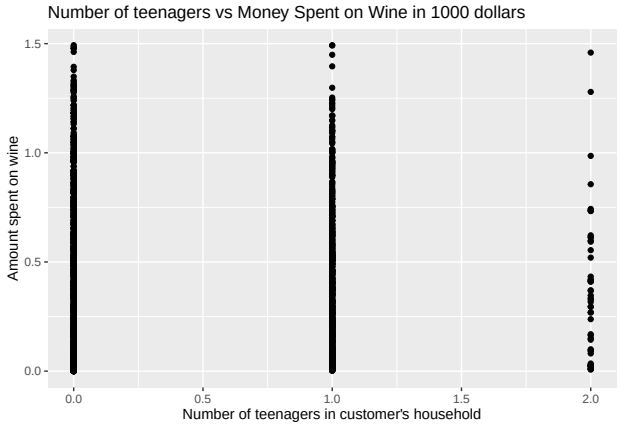
Notice: as mentioned in the requirement, we made far more plots than we will keep.

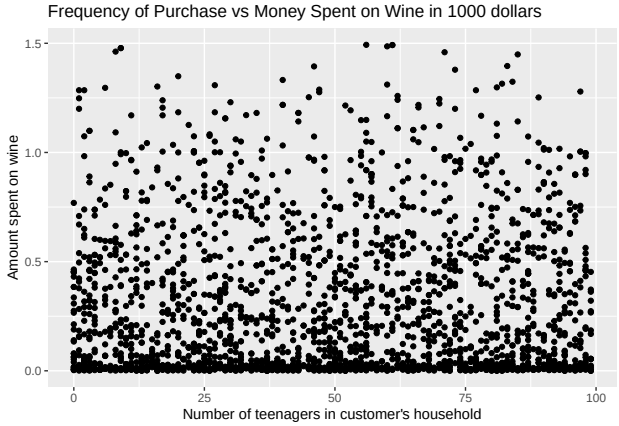
We explore the relationship between the response variable MntWines, which indicates amount spent on wine in last 2 years, and the other explanatory variables.

First, we plot MntWines against each plausible explanatory variable as follows:

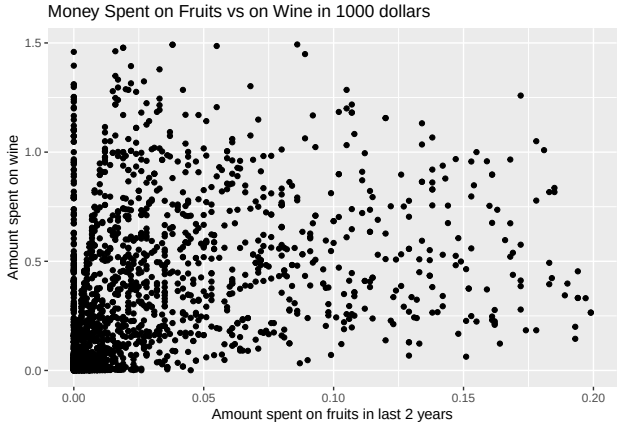



Number of Kids vs Money Spent on Wine in 1000 dollars
Amount spent on wine
Number of children in customer's household


Number of teenagers vs Money Spent on Wine in 1000 dollars
Amount spent on wine
Number of teenagers in customer's household

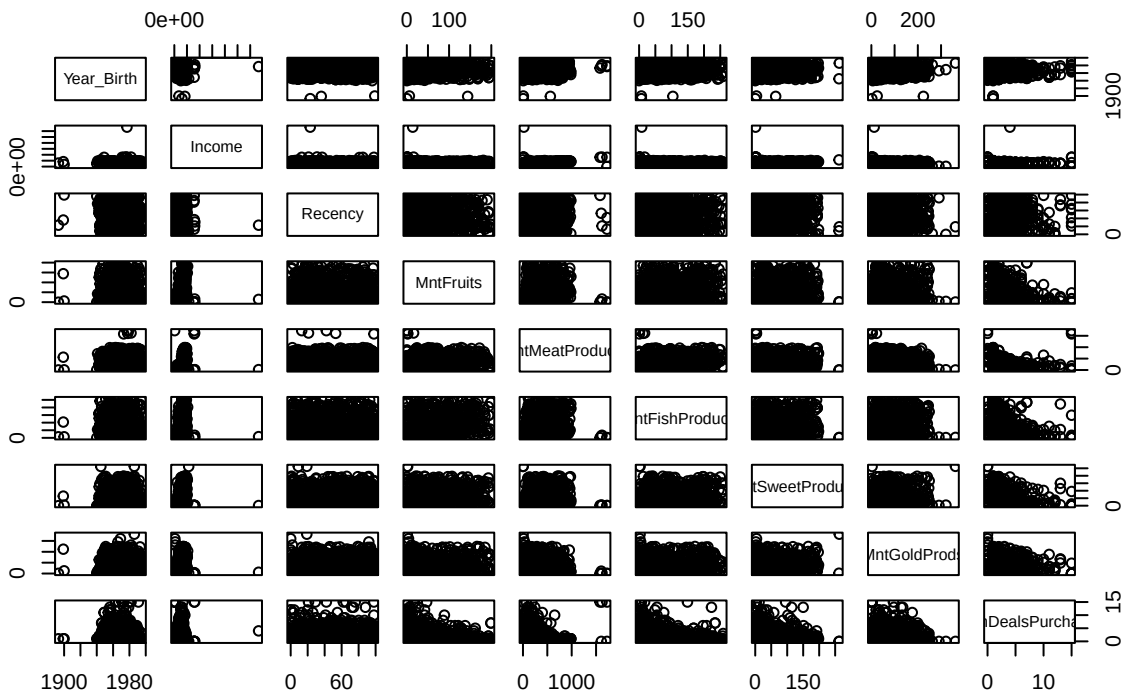

Frequency of Purchase vs Money Spent on Wine in 1000 dollars
Amount spent on wine
Number of teenagers in customer's household


Complain vs Money Spent on Wine in 1000 dollars
Amount spent on wine
1 if customer complained in the last 2 years, 0 otherwise


Money Spent on Fruits vs on Wine in 1000 dollars
Amount spent on wine
Amount spent on fruits in last 2 years


Money Spent on Meat Product vs on Wine in 1000 dollars
Amount spent on wine
Amount spent on meat in last 2 years

Then, we plot pairwise correlation to have a glance.



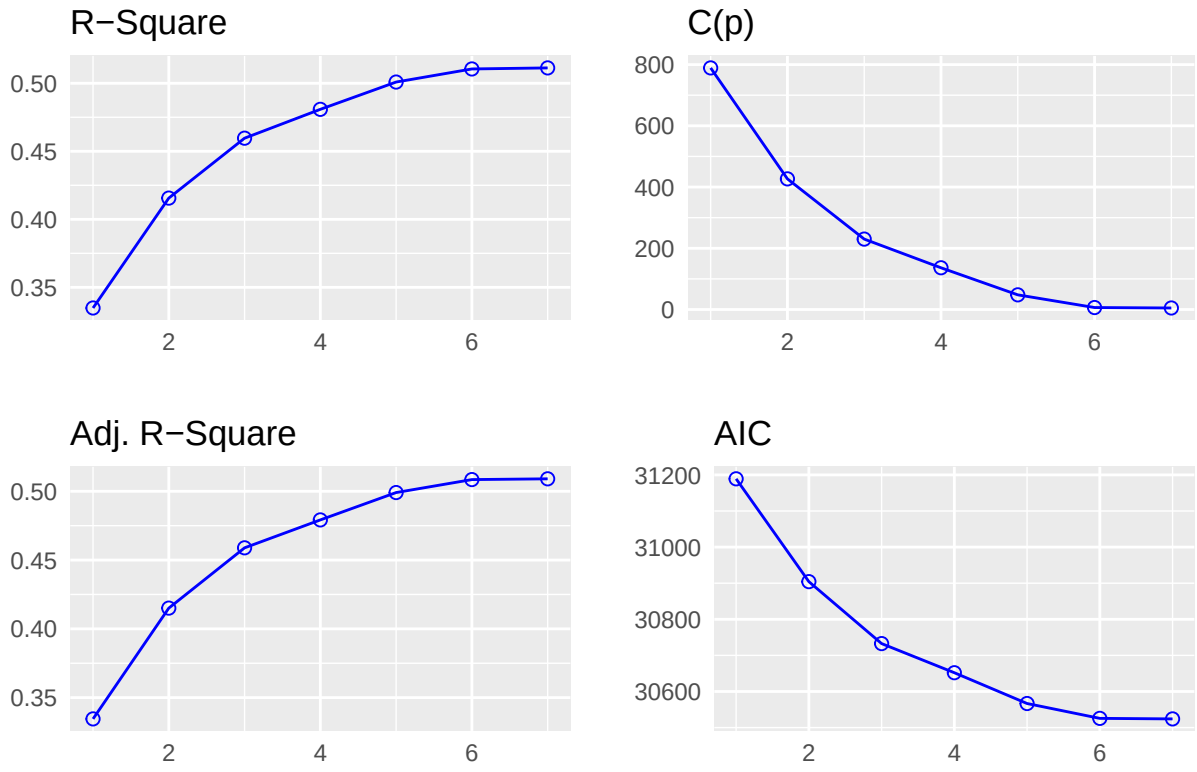
Then, we apply model selecting methods to explore the prediction risk (Cp, AIC):

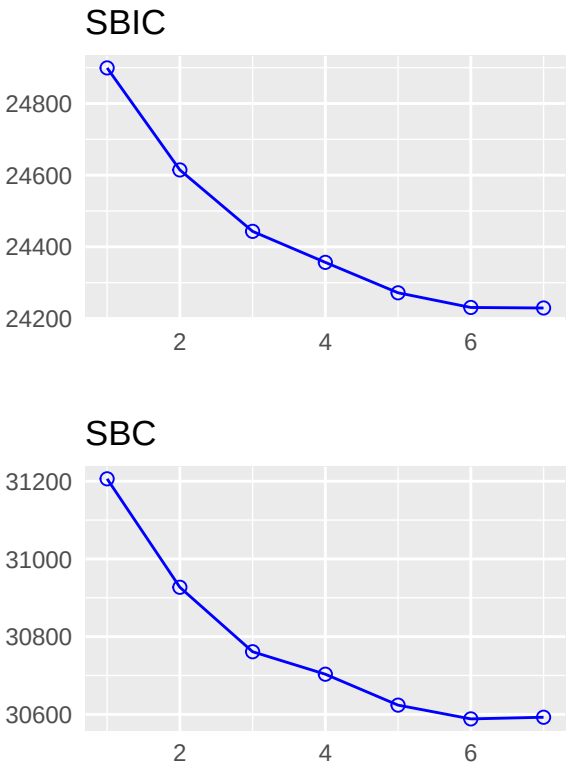
```
##
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------------
##                                Added/                   Adj.
## Step        Variable          Removed     R-Square    R-Square      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.

```
##                                         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
```

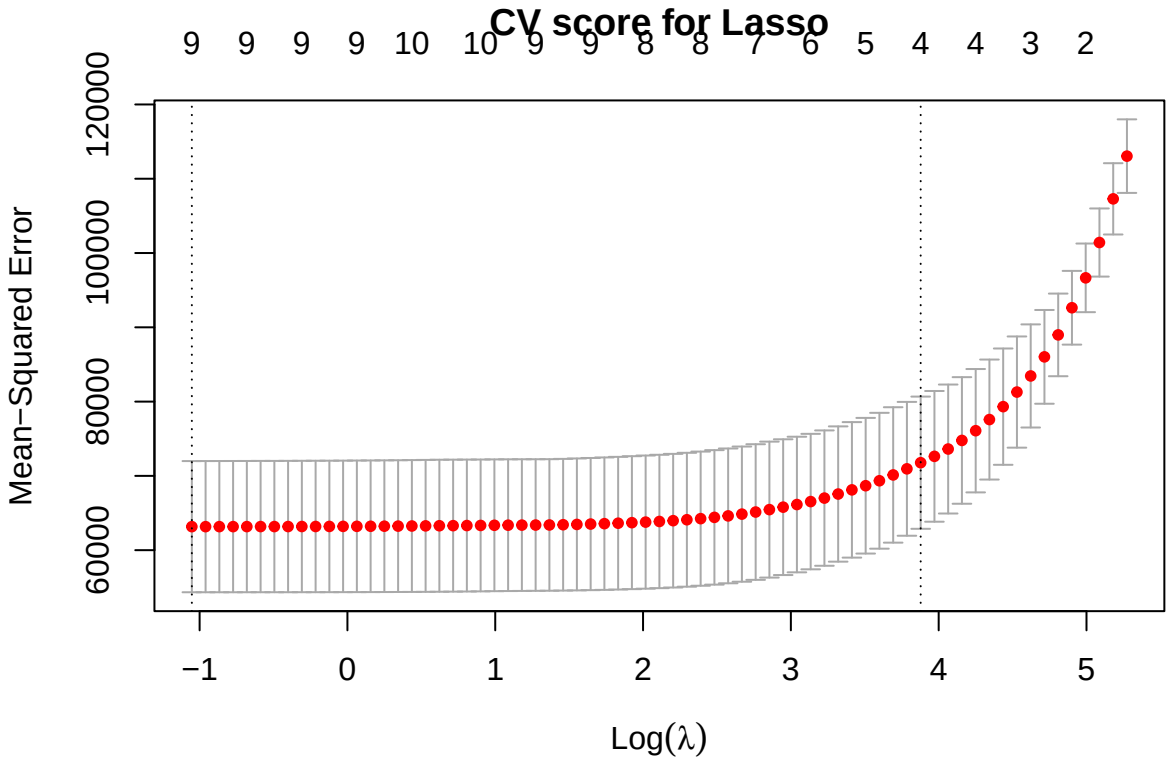
Next, we plot risk against different possible models (in the table above) as follows:


R−Square
0.50
0.45
0.40
0.35
2
4
6
C(p)
800
600
400
200
0
2
4
6
Adj. R−Square
0.50
0.45
0.40
0.35
2
4
6
AIC
31200
31000
30800
30600
2
4
6



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)