JIP45_HW2_INFSCI2160

Jing Pang

1/18/2019

Goal: Explain AmountSpent in terms of the provided customer characteristics.

Data Preparation

```
direct_marketing <- read.csv("~/Dropbox/19 Spring/INFSCI_2160_DM/HWs/HW2/HOMEWORK_2_DATASET_DIRECT_MARK
```

Question 1: Identifying variables

- Identify the response variable and the predictor variables.

```
head(direct_marketing)
```

```
##      Age Gender OwnHome Married Location Salary Children History Catalogs
## 1    Old Female     Own  Single      Far  47500        0    High        6
## 2 Middle   Male    Rent  Single    Close  63600        0    High        6
## 3  Young Female    Rent  Single    Close  13500        0     Low       18
## 4 Middle   Male     Own Married    Close  85600        1    High       18
## 5 Middle Female     Own  Single    Close  68400        0    High       12
## 6  Young   Male     Own Married    Close  30400        0     Low        6
##   AmountSpent
## 1         755
## 2        1318
## 3         296
## 4        2436
## 5        1304
## 6         495
```

In this case, the response variable is AmountSpent; the predictor variables are Age, Gender, OwnHome, Married, Location, Salary, Children, History, and Catalogs

Question 2: Exploring the dataset

Explore the dataset and generate a statistical and graphical summary:

- There are missing values in the dataset. Describe how you deal with them. (Hint: Check the data description to see what the missingness means.)

There are 303 missing values in the History column, and around 30 percentage of History values are missing compared with total. The missing data represents a pretty huge amount fraction of the dataset. So it's probably safe just not to drop these data from my analysis. One of solutions is just to create a new category for the variable, called missing.

```
# find out which column having missing values
colSums(is.na(direct_marketing))
```

```
##         Age      Gender     OwnHome     Married    Location      Salary
##           0           0           0           0           0           0
##    Children     History    Catalogs AmountSpent
##           0         303           0           0
```

```r
# find summary of History column
summary(direct_marketing$History)
```

```
##   High    Low Medium   NA's
##    255    230    212    303
```

```r
# create a new variable
direct_marketing$History.fix <- as.factor(ifelse(is.na(direct_marketing$History), "Missing",
                         ifelse(direct_marketing$History=="High","High",
                                ifelse(direct_marketing$History=="Low", "Low","Medium"))))

summary(direct_marketing$History.fix)
```

```
##    High     Low  Medium Missing
##     255     230     212     303
```

- Generate a summary table for the data. For each numerical variable, list the name of it, mean, median, 1st quartile, 3rd quartile, and standard deviation.

```r
nums <- unlist(lapply(direct_marketing, is.numeric))
num_direct_marketing <- direct_marketing[,nums]
sapply(num_direct_marketing, function(x) c("Mean"= mean(x,na.rm=TRUE),
                             "Median" = median(x),
                             "1st Quantile" = quantile(x,0.25),
                             "3rd Quartile" = quantile(x,0.75),
                             "Std" = sd(x)
                             )
       )
```

```
##                      Salary Children  Catalogs AmountSpent
## Mean               56103.90  0.93400 14.682000   1216.7700
## Median             53700.00  1.00000 12.000000    962.0000
## 1st Quantile.25% 29975.00  0.00000  6.000000    488.2500
## 3rd Quartile.75% 77025.00  2.00000 18.000000   1688.5000
## Std                30616.31  1.05107  6.622895    961.0686
```

- Plot the density distribution of the AmountSpent and Salary variables. What type of shape do they have?

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

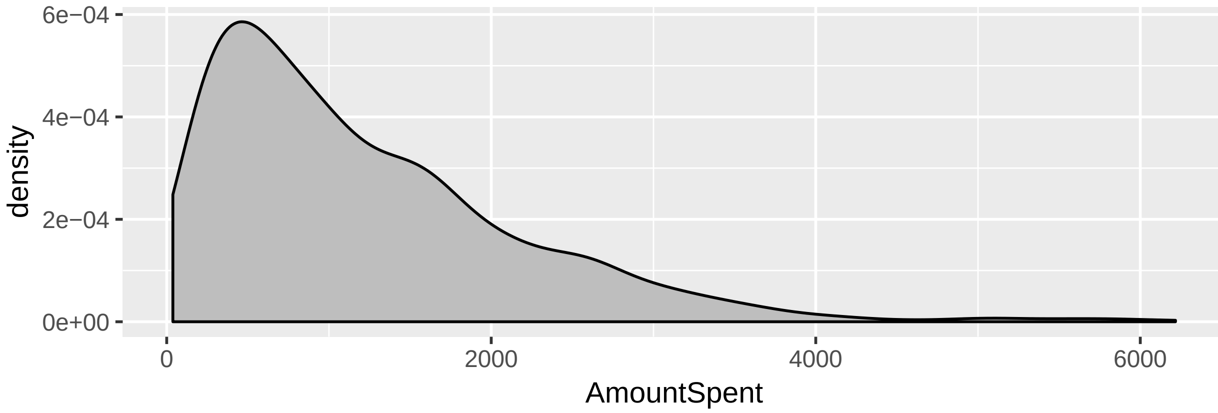
```r
p1 <- ggplot(direct_marketing) + geom_density(aes(x = AmountSpent),  binwidth = 100, fill = "grey", col
```

```
## Warning: Ignoring unknown parameters: binwidth
```

```r
p2 <- ggplot(direct_marketing) + geom_density(aes(x = Salary),  binwidth = 1000, fill = "grey", color =
```

```
## Warning: Ignoring unknown parameters: binwidth
```

```r
library("gridExtra")
grid.arrange(p1, p2, ncol = 1)
```



From the graphs above, the density graph of AmountSpent feature is clearly a right-skewed shape. But the shape of Salary feature is also a kind of right-skewed shape. However, this shape spread wider than AmountSpent feature, and it has two peaks in the graph.

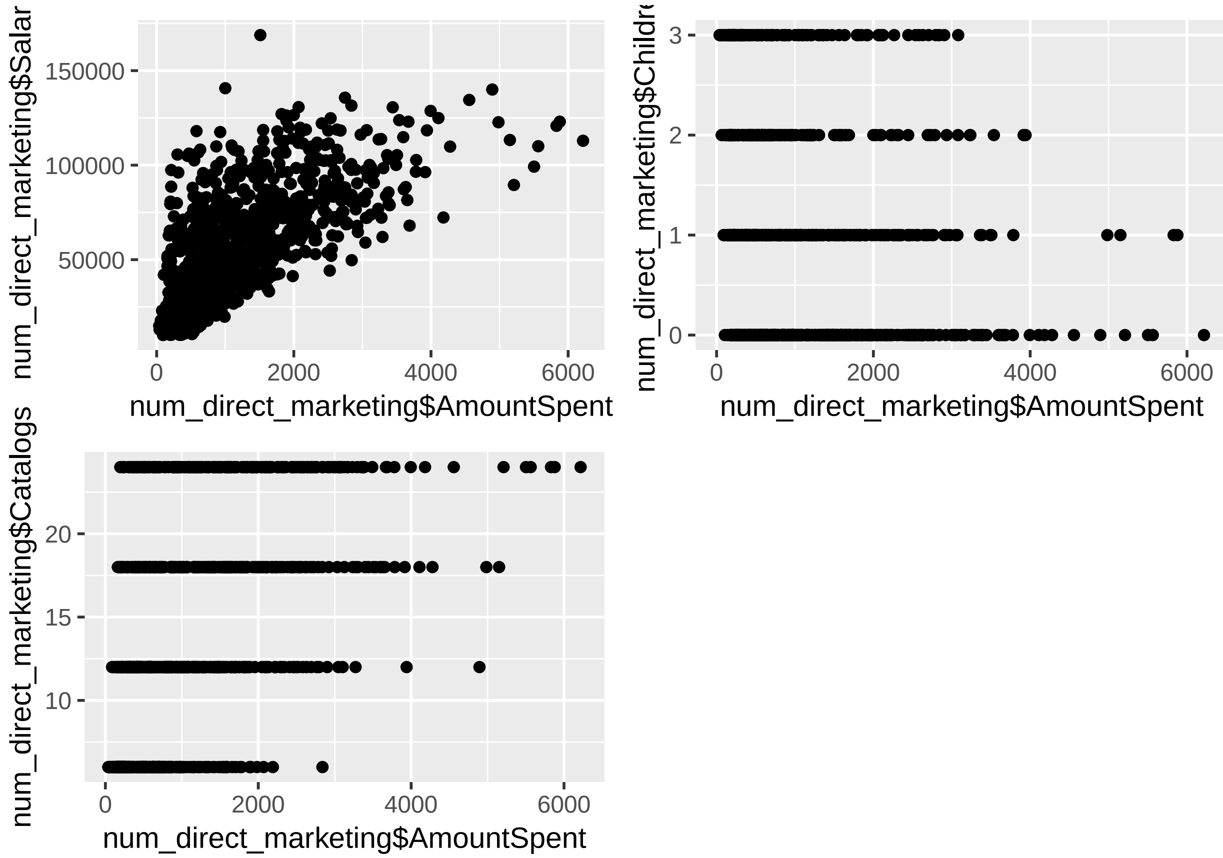
- Describe the relationship between all the continuous variables and the response variable in terms of correlation and a scatter plot.

In this case, because the description of dataset shows that features of Age, Gender, OwnHome, Married, Location and History are not continuous variables; we will only consider two variables Salary and Catalogs in this part.

```
# correlation
cor(num_direct_marketing)
```

```
##                 Salary    Children   Catalogs AmountSpent
## Salary      1.00000000  0.04966316  0.1835509   0.6995957
## Children    0.04966316  1.00000000 -0.1134554  -0.2223082
## Catalogs    0.18355086 -0.11345543  1.0000000   0.4726499
## AmountSpent 0.69959571 -0.22230817  0.4726499   1.0000000
```

```
# scatter plot
library(ggplot2)
p1 <- ggplot(num_direct_marketing) + geom_point(aes(num_direct_marketing$AmountSpent, num_direct_market
p2 <- ggplot(num_direct_marketing) + geom_point(aes(num_direct_marketing$AmountSpent, num_direct_market
p3 <- ggplot(num_direct_marketing) + geom_point(aes(num_direct_marketing$AmountSpent, num_direct_market
library("gridExtra")
grid.arrange(p1, p2, p3, nrow = 2, ncol = 2)
```

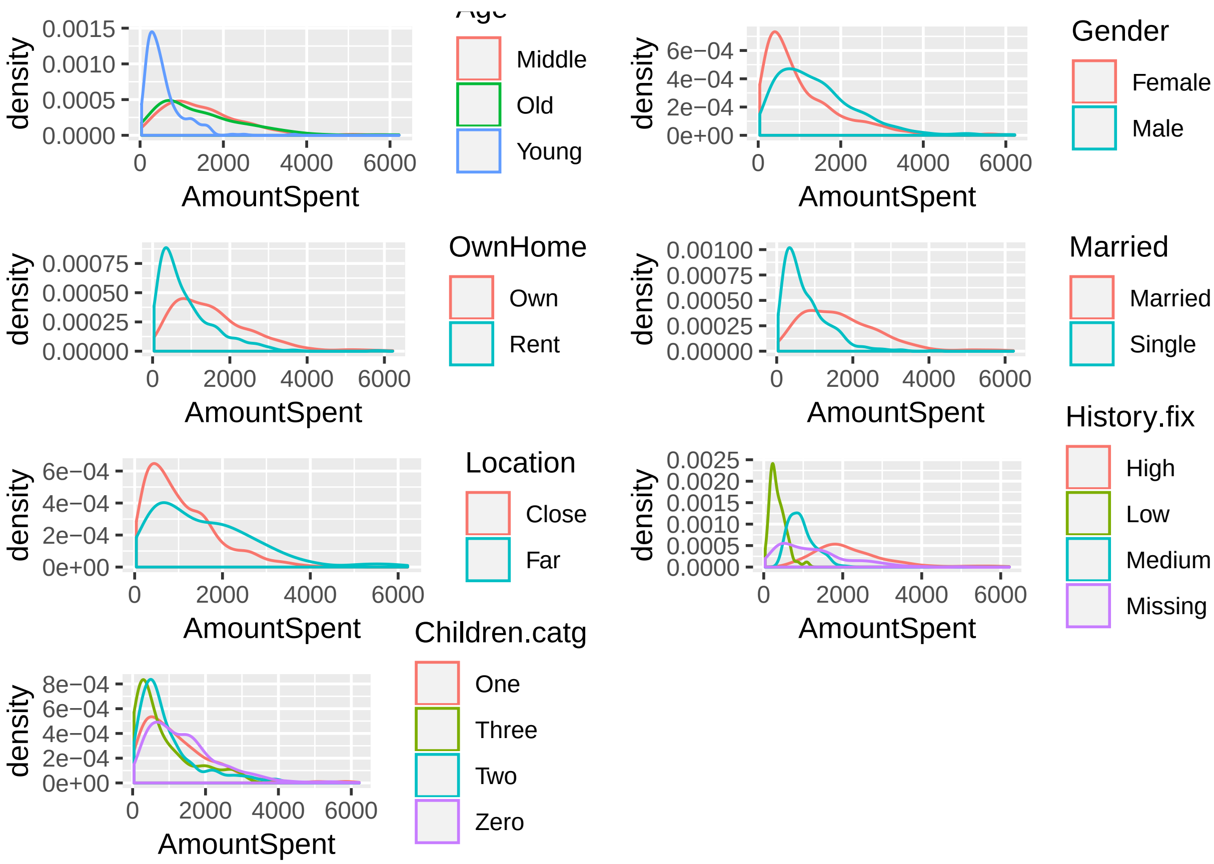


- For each categorical variable, generate a conditional density plot of the response variable. (Hint: Plot the density of the response variable into multiple distributions separated by the predictor's categories on the same figure. Use different colors or line shapes to differentiate the categories.)

```
sapply(direct_marketing,class)
```

```
##          Age       Gender      OwnHome      Married     Location       Salary
##     "factor"     "factor"     "factor"     "factor"     "factor"    "integer"
##     Children      History     Catalogs  AmountSpent  History.fix
##    "integer"     "factor"    "integer"    "integer"     "factor"
```

```
p1<-ggplot(direct_marketing) + geom_density(aes(x = AmountSpent, color = Age))
p2<-ggplot(direct_marketing) + geom_density(aes(x = AmountSpent, color = Gender))
p3<-ggplot(direct_marketing) + geom_density(aes(x = AmountSpent, color = OwnHome))
p4<-ggplot(direct_marketing) + geom_density(aes(x = AmountSpent, color = Married))
p5<-ggplot(direct_marketing) + geom_density(aes(x = AmountSpent, color = Location))
p6<-ggplot(direct_marketing) + geom_density(aes(x = AmountSpent, color = History.fix))

# children
direct_marketing$Children.catg <- as.factor(ifelse(direct_marketing$Children==0, "Zero",
                         ifelse(direct_marketing$Children==1,"One",
                                ifelse(direct_marketing$Children==2, "Two","Three"))))

p7<-ggplot(direct_marketing) + geom_density(aes(x = AmountSpent, color = Children.catg))

library("gridExtra")
grid.arrange(p1, p2, p3, p4, p5, p6, p7, nrow = 4, ncol = 2)
```



Question 3: Apply regression analysis to the dataset to predict AmountSpent

Run a multiple linear regression model. How does it perform?

Data transformation

We find out the features of Salary and AmountSpent are extremely huge compared to others. This will affect our model's precision. Therefore, we have to use normalization to downsize the variable into a reasonable scaling.

```
direct_marketing$AmountSpent.norm <- scale(direct_marketing$AmountSpent)
direct_marketing$Salary.norm <- scale(direct_marketing$Salary)
direct_marketing$Catalogs.norm <- scale(direct_marketing$Catalogs)
```

```
direct_marketing_lm <- lm(AmountSpent.norm ~ Age + Gender + OwnHome + Married + Location + Salary.norm +
summary(direct_marketing_lm)
```

```
##
## Call:
## lm(formula = AmountSpent.norm ~ Age + Gender + OwnHome + Married +
##     Location + Salary.norm + Children.catg + History.fix + Catalogs.norm,
##     data = direct_marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78069 -0.30174 -0.01886  0.24635  2.99265
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.050049   0.068029   0.736  0.46208
## AgeOld              0.079103   0.051638   1.532  0.12587
## AgeYoung            0.010676   0.051784   0.206  0.83671
## GenderMale         -0.044366   0.034518  -1.285  0.19898
## OwnHomeRent        -0.018489   0.038190  -0.484  0.62839
## MarriedSingle       0.035664   0.046406   0.769  0.44236
## LocationFar         0.453192   0.037411  12.114  < 2e-16 ***
## Salary.norm         0.611614   0.032837  18.626  < 2e-16 ***
## Children.catgThree -0.377267   0.057438  -6.568 8.23e-11 ***
## Children.catgTwo   -0.196482   0.053150  -3.697  0.00023 ***
## Children.catgZero   0.130788   0.043989   2.973  0.00302 **
## History.fixLow     -0.366875   0.068274  -5.374 9.64e-08 ***
## History.fixMedium  -0.422413   0.055462  -7.616 6.11e-14 ***
## History.fixMissing  0.005624   0.053465   0.105  0.91625
## Catalogs.norm       0.289111   0.016967  17.040  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5057 on 985 degrees of freedom
## Multiple R-squared:  0.7479, Adjusted R-squared:  0.7443
## F-statistic: 208.7 on 14 and 985 DF,  p-value: < 2.2e-16
```

Use ridge, lasso, and AIC for feature selection. For each model:

```r
library(rsample)
```

```
## Loading required package: tidyr
```

```r
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

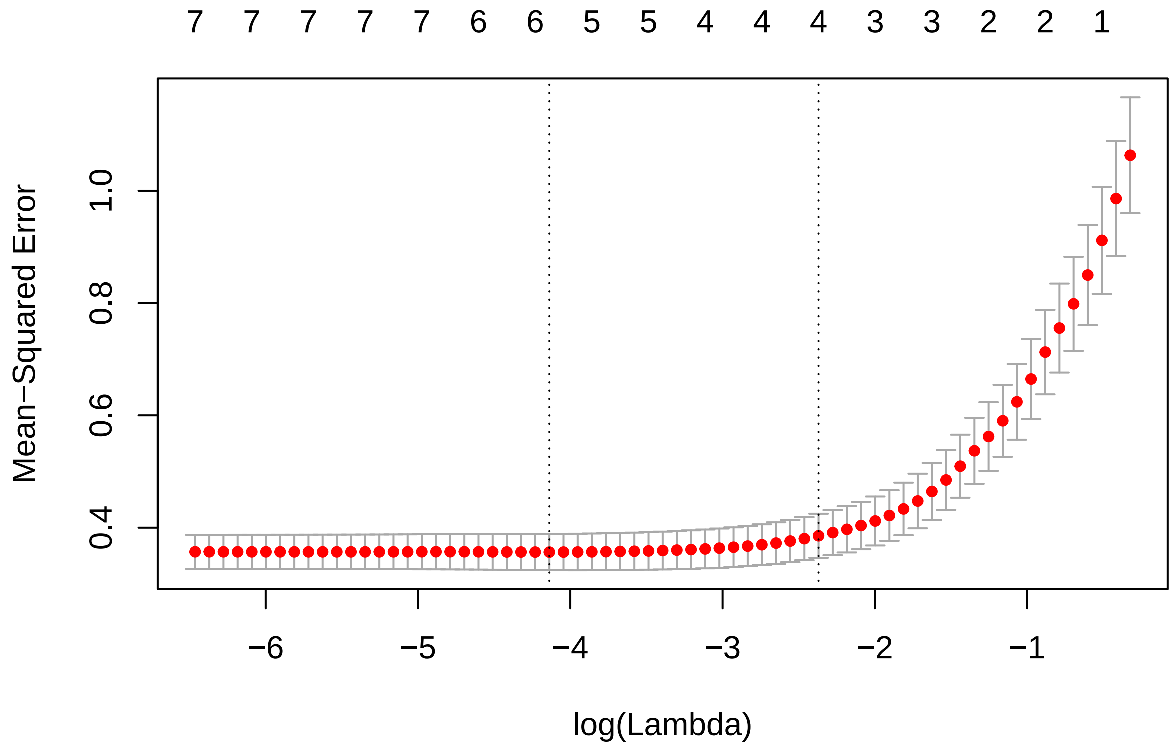
```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
## Loading required package: foreach
```

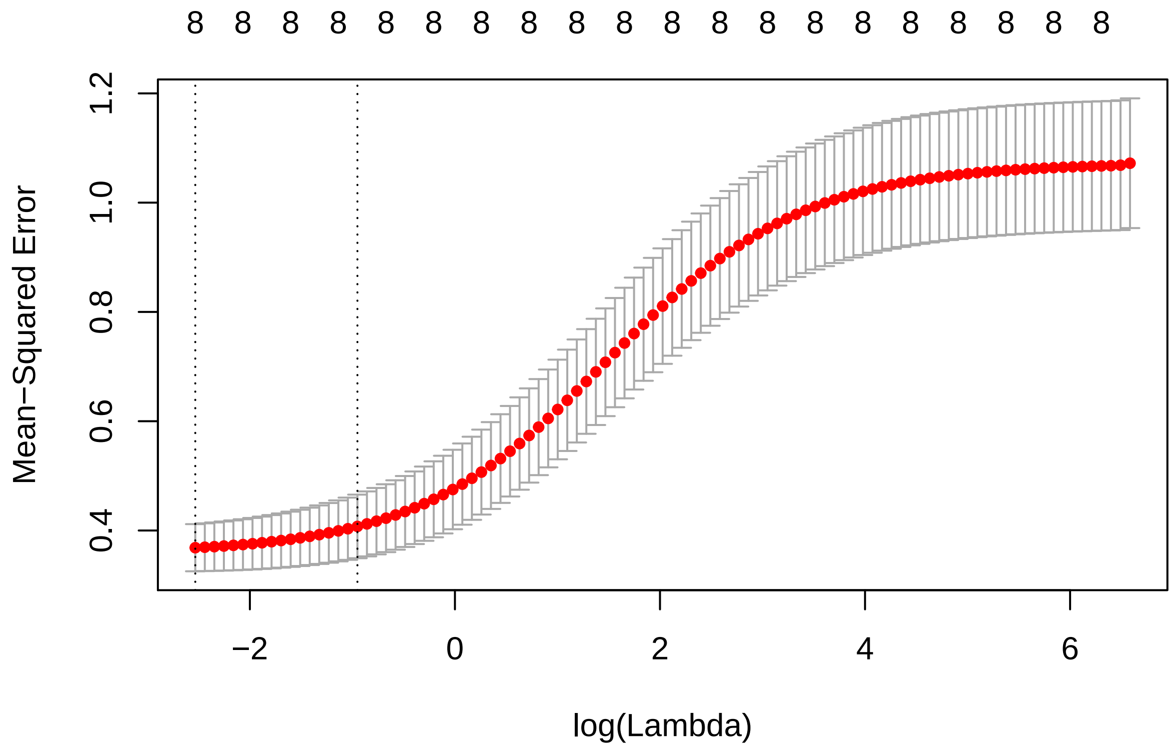
```
## Loaded glmnet 2.0-16
```

```r
# LASSO

## prepare train and test sets
direct_marketing.prep <- direct_marketing[,c("Age", "Gender", "OwnHome", "Married", "Location", "Childre
direct_marketing.prep[, c("Age", "Gender", "OwnHome", "Married", "Location", "Children.catg")] <- lappl
direct_marketing.prep[, c("Age", "Gender", "OwnHome", "Married", "Location", "Children.catg")] <- lappl

direct_marketing_train_test_split <- initial_split(direct_marketing.prep, prop = 0.8)
direct_marketing_train_tbl <- training(direct_marketing_train_test_split)
direct_marketing_test_tbl  <- testing(direct_marketing_train_test_split)
## separate variables and target
```

```
direct_marketing_vars_train<- direct_marketing_train_tbl[, c("Age", "Gender", "OwnHome", "Married", "Loc

direct_marketing_vars_test <- direct_marketing_test_tbl[, c("Age", "Gender", "OwnHome", "Married", "Loca

direct_marketing_target_train <- direct_marketing_train_tbl[, c("AmountSpent.norm")]
direct_marketing_target_test <- direct_marketing_test_tbl[, c("AmountSpent.norm")]

## cross-validation model
set.seed(1)
direct_marketing_cv_lasso = cv.glmnet(x = as.matrix(direct_marketing_vars_train), y = direct_marketing_t
plot(direct_marketing_cv_lasso, label=TRUE)
```




```
direct_marketing.bestlasso = direct_marketing_cv_lasso$lambda.min

# RIDGE

direct_marketing_cv_ridge = cv.glmnet(x = as.matrix(direct_marketing_vars_train), y = direct_marketing_t
plot(direct_marketing_cv_ridge, label=TRUE)
```




```r
direct_marketing.bestridge <- direct_marketing_cv_ridge$lambda.min

#Test MSE of LASSO and ridge
direct_marketing.pred_lasso = predict(direct_marketing_cv_lasso, s = direct_marketing.bestlasso, newx =
direct_marketing.pred_lasso_mse <- mean((direct_marketing.pred_lasso-direct_marketing_test_tbl$AmountSpe

direct_marketing.pred_ridge = predict(direct_marketing_cv_ridge, s = direct_marketing.bestridge, newx =
direct_marketing.pred_ridge_mse <- mean((direct_marketing.pred_ridge-direct_marketing_test_tbl$AmountSpe

t_mse <- matrix(c(direct_marketing.bestlasso,direct_marketing.bestridge,direct_marketing.pred_lasso_mse
colnames(t_mse) <- c("LASSO","RIDGE")
rownames(t_mse) <- c("lambda", "MSE")
t_mse<-as.table(t_mse)

# AIC
library(MASS)
# backward stepwise
direct_marketing_lm_backward <- stepAIC(direct_marketing_lm, direction = "backward")
```

```
## Start:  AIC=-1348.85
## AmountSpent.norm ~ Age + Gender + OwnHome + Married + Location +
##     Salary.norm + Children.catg + History.fix + Catalogs.norm
##
##                  Df Sum of Sq    RSS     AIC
## - OwnHome         1     0.060 251.93 -1350.6
## - Age             2     0.639 252.51 -1350.3
## - Married         1     0.151 252.02 -1350.3
## - Gender          1     0.422 252.29 -1349.2
## <none>                        251.87 -1348.8
## - Children.catg   3    21.112 272.98 -1274.4
```

```
## - History.fix    3    30.310 282.18 -1241.2
## - Location       1    37.524 289.39 -1212.0
## - Catalogs.norm  1    74.243 326.11 -1092.5
## - Salary.norm    1    88.710 340.58 -1049.1
##
## Step:  AIC=-1350.62
## AmountSpent.norm ~ Age + Gender + Married + Location + Salary.norm +
##     Children.catg + History.fix + Catalogs.norm
##
##                 Df Sum of Sq    RSS     AIC
## - Married        1     0.163 252.09 -1352.0
## - Age            2     0.761 252.69 -1351.6
## - Gender         1     0.428 252.35 -1350.9
## <none>                       251.93 -1350.6
## - Children.catg  3    21.052 272.98 -1276.4
## - History.fix    3    30.728 282.65 -1241.5
## - Location       1    37.466 289.39 -1214.0
## - Catalogs.norm  1    74.208 326.14 -1094.4
## - Salary.norm    1    93.437 345.36 -1037.2
##
## Step:  AIC=-1351.97
## AmountSpent.norm ~ Age + Gender + Location + Salary.norm + Children.catg +
##     History.fix + Catalogs.norm
##
##                 Df Sum of Sq    RSS      AIC
## - Age            2     0.673 252.76 -1353.30
## - Gender         1     0.398 252.49 -1352.39
## <none>                       252.09 -1351.97
## - Children.catg  3    21.187 273.28 -1277.27
## - History.fix    3    31.164 283.25 -1241.41
## - Location       1    37.380 289.47 -1215.70
## - Catalogs.norm  1    74.058 326.15 -1096.40
## - Salary.norm    1   143.707 395.80  -902.85
##
## Step:  AIC=-1353.3
## AmountSpent.norm ~ Gender + Location + Salary.norm + Children.catg +
##     History.fix + Catalogs.norm
##
##                 Df Sum of Sq    RSS      AIC
## <none>                       252.76 -1353.30
## - Gender         1     0.688 253.45 -1352.59
## - Children.catg  3    27.123 279.89 -1257.37
## - History.fix    3    31.032 283.79 -1243.50
## - Location       1    37.627 290.39 -1216.53
## - Catalogs.norm  1    73.962 326.72 -1098.64
## - Salary.norm    1   191.600 444.36  -791.12
```

```r
# forward stepwise
direct_marketing_lm_forward <- stepAIC(direct_marketing_lm, direction = "forward")
```

```
## Start:  AIC=-1348.85
## AmountSpent.norm ~ Age + Gender + OwnHome + Married + Location +
##     Salary.norm + Children.catg + History.fix + Catalogs.norm
```

```r
# mix
direct_marketing_lm_both <- stepAIC(direct_marketing_lm, direction = "both")
```

```
## Start:  AIC=-1348.85
## AmountSpent.norm ~ Age + Gender + OwnHome + Married + Location +
##     Salary.norm + Children.catg + History.fix + Catalogs.norm
##
##                 Df Sum of Sq    RSS     AIC
## - OwnHome        1     0.060 251.93 -1350.6
## - Age            2     0.639 252.51 -1350.3
## - Married        1     0.151 252.02 -1350.3
## - Gender         1     0.422 252.29 -1349.2
## <none>                       251.87 -1348.8
## - Children.catg  3    21.112 272.98 -1274.4
## - History.fix    3    30.310 282.18 -1241.2
## - Location       1    37.524 289.39 -1212.0
## - Catalogs.norm  1    74.243 326.11 -1092.5
## - Salary.norm    1    88.710 340.58 -1049.1
##
## Step:  AIC=-1350.62
## AmountSpent.norm ~ Age + Gender + Married + Location + Salary.norm +
##     Children.catg + History.fix + Catalogs.norm
##
##                 Df Sum of Sq    RSS     AIC
## - Married        1     0.163 252.09 -1352.0
## - Age            2     0.761 252.69 -1351.6
## - Gender         1     0.428 252.35 -1350.9
## <none>                       251.93 -1350.6
## + OwnHome        1     0.060 251.87 -1348.8
## - Children.catg  3    21.052 272.98 -1276.4
## - History.fix    3    30.728 282.65 -1241.5
## - Location       1    37.466 289.39 -1214.0
## - Catalogs.norm  1    74.208 326.14 -1094.4
## - Salary.norm    1    93.437 345.36 -1037.2
##
## Step:  AIC=-1351.97
## AmountSpent.norm ~ Age + Gender + Location + Salary.norm + Children.catg +
##     History.fix + Catalogs.norm
##
##                 Df Sum of Sq    RSS      AIC
## - Age            2     0.673 252.76 -1353.30
## - Gender         1     0.398 252.49 -1352.39
## <none>                       252.09 -1351.97
## + Married        1     0.163 251.93 -1350.62
## + OwnHome        1     0.072 252.02 -1350.26
## - Children.catg  3    21.187 273.28 -1277.27
## - History.fix    3    31.164 283.25 -1241.41
## - Location       1    37.380 289.47 -1215.70
## - Catalogs.norm  1    74.058 326.15 -1096.40
## - Salary.norm    1   143.707 395.80  -902.85
##
## Step:  AIC=-1353.3
## AmountSpent.norm ~ Gender + Location + Salary.norm + Children.catg +
##     History.fix + Catalogs.norm
```

```
##
##                 Df Sum of Sq    RSS      AIC
## <none>                       252.76 -1353.30
## - Gender         1     0.688 253.45 -1352.59
## + OwnHome        1     0.199 252.56 -1352.09
## + Age            2     0.673 252.09 -1351.97
## + Married        1     0.075 252.69 -1351.60
## - Children.catg  3    27.123 279.89 -1257.37
## - History.fix    3    31.032 283.79 -1243.50
## - Location       1    37.627 290.39 -1216.53
## - Catalogs.norm  1    73.962 326.72 -1098.64
## - Salary.norm    1   191.600 444.36  -791.12
```

- Identify which variables are statistically significant.

From the previous linear regression model, we can see that the variables Location, Salary, Children, History.fix, and Catalogs have statistically significant, while Age, Gender, OwnHome, and Married are not (all of them have p-values > 0.05)

- Evaluate the performance of your model.

```r
t_mse
```

```
##             LASSO      RIDGE
## lambda 0.01596327 0.07944960
## MSE    0.21586370 0.21454322
```

The table above shows that the difference between LASSO and RIDGE is not too huge. If compared both of models, we would consider the LASSO model is performing the better result than RIDGE because of the MSE number.

```r
summary(direct_marketing_lm_backward)
```

```
##
## Call:
## lm(formula = AmountSpent.norm ~ Gender + Location + Salary.norm +
##     Children.catg + History.fix + Catalogs.norm, data = direct_marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71191 -0.29992 -0.02124  0.24426  3.04797
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.072769   0.055759   1.305 0.192176
## GenderMale         -0.055148   0.033614  -1.641 0.101191
## LocationFar         0.453270   0.037356  12.134  < 2e-16 ***
## Salary.norm         0.602522   0.022006  27.380  < 2e-16 ***
## Children.catgThree -0.377561   0.057355  -6.583 7.47e-11 ***
## Children.catgTwo   -0.195027   0.053005  -3.679 0.000246 ***
## Children.catgZero   0.161092   0.040201   4.007 6.61e-05 ***
## History.fixLow     -0.369333   0.068141  -5.420 7.48e-08 ***
## History.fixMedium  -0.426087   0.054916  -7.759 2.13e-14 ***
## History.fixMissing -0.001494   0.053262  -0.028 0.977623
## Catalogs.norm       0.287927   0.016925  17.012  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5055 on 989 degrees of freedom
## Multiple R-squared:  0.747,  Adjusted R-squared:  0.7444
## F-statistic:   292 on 10 and 989 DF,  p-value: < 2.2e-16
```

```r
summary(direct_marketing_lm_forward)
```

```
##
## Call:
## lm(formula = AmountSpent.norm ~ Age + Gender + OwnHome + Married +
##     Location + Salary.norm + Children.catg + History.fix + Catalogs.norm,
##     data = direct_marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78069 -0.30174 -0.01886  0.24635  2.99265
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.050049   0.068029   0.736  0.46208
## AgeOld               0.079103   0.051638   1.532  0.12587
## AgeYoung             0.010676   0.051784   0.206  0.83671
## GenderMale          -0.044366   0.034518  -1.285  0.19898
## OwnHomeRent         -0.018489   0.038190  -0.484  0.62839
## MarriedSingle        0.035664   0.046406   0.769  0.44236
## LocationFar          0.453192   0.037411  12.114  < 2e-16 ***
## Salary.norm          0.611614   0.032837  18.626  < 2e-16 ***
## Children.catgThree -0.377267   0.057438  -6.568 8.23e-11 ***
## Children.catgTwo   -0.196482   0.053150  -3.697  0.00023 ***
## Children.catgZero   0.130788   0.043989   2.973  0.00302 **
## History.fixLow     -0.366875   0.068274  -5.374 9.64e-08 ***
## History.fixMedium  -0.422413   0.055462  -7.616 6.11e-14 ***
## History.fixMissing  0.005624   0.053465   0.105  0.91625
## Catalogs.norm       0.289111   0.016967  17.040  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5057 on 985 degrees of freedom
## Multiple R-squared:  0.7479, Adjusted R-squared:  0.7443
## F-statistic: 208.7 on 14 and 985 DF,  p-value: < 2.2e-16
```

```r
summary(direct_marketing_lm_both)
```

```
##
## Call:
## lm(formula = AmountSpent.norm ~ Gender + Location + Salary.norm +
##     Children.catg + History.fix + Catalogs.norm, data = direct_marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71191 -0.29992 -0.02124  0.24426  3.04797
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.072769   0.055759   1.305 0.192176
```

```
## GenderMale         -0.055148   0.033614  -1.641 0.101191    
## LocationFar         0.453270   0.037356  12.134  < 2e-16 ***
## Salary.norm         0.602522   0.022006  27.380  < 2e-16 ***
## Children.catgThree -0.377561   0.057355  -6.583 7.47e-11 ***
## Children.catgTwo   -0.195027   0.053005  -3.679 0.000246 ***
## Children.catgZero   0.161092   0.040201   4.007 6.61e-05 ***
## History.fixLow     -0.369333   0.068141  -5.420 7.48e-08 ***
## History.fixMedium  -0.426087   0.054916  -7.759 2.13e-14 ***
## History.fixMissing -0.001494   0.053262  -0.028 0.977623    
## Catalogs.norm       0.287927   0.016925  17.012  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5055 on 989 degrees of freedom
## Multiple R-squared:  0.747,  Adjusted R-squared:  0.7444 
## F-statistic:   292 on 10 and 989 DF,  p-value: < 2.2e-16
```

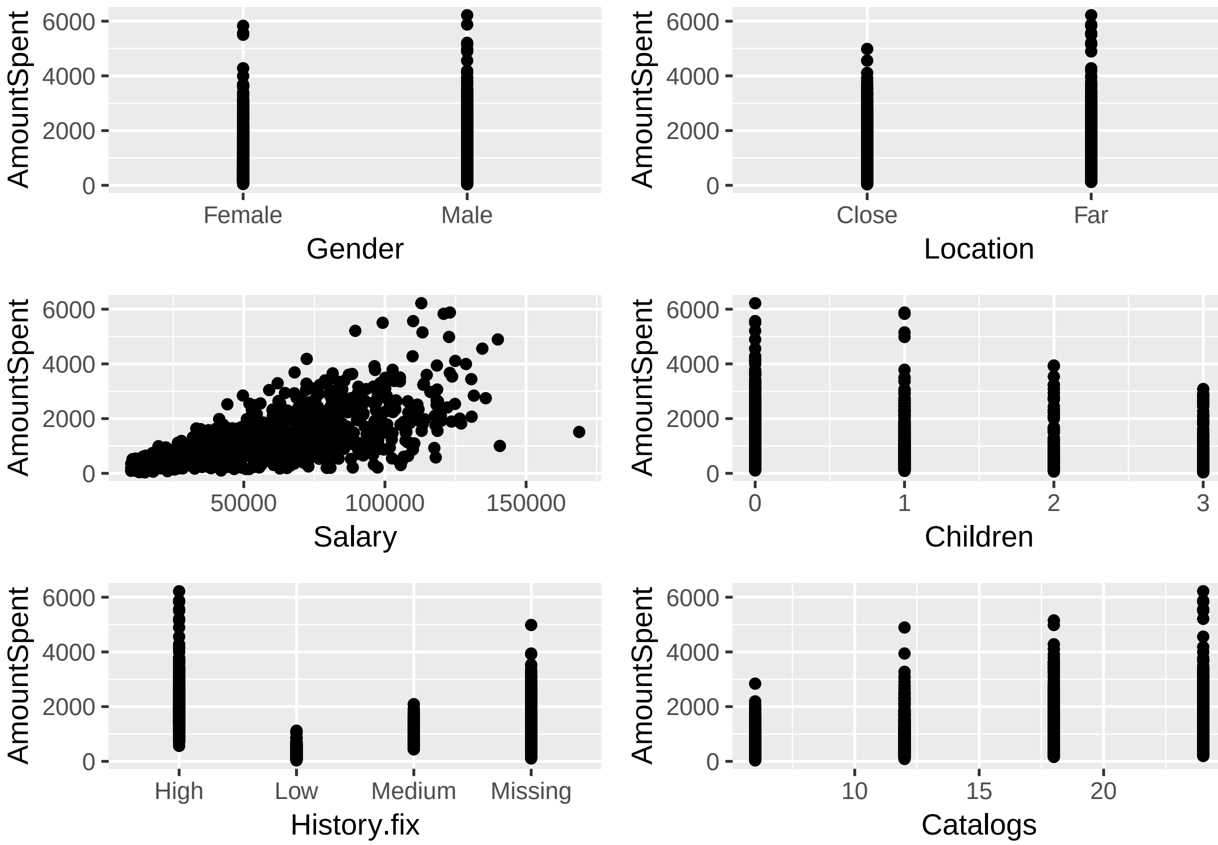
And from previous lists of summary, the forward AIC performs a better model than other two models because its F-statistic score is the best.

- Which model performed best? How did you decide this?

From the table above, we can see that the mean square error of LASSO and RIDGE feature selection model do not have too large difference. Therefore, both feature selections will work good enough. From AIC statistics, we can see that these six variables (Gender, Location, Salary, Children, History.fix and Catalogs) present a better model output.

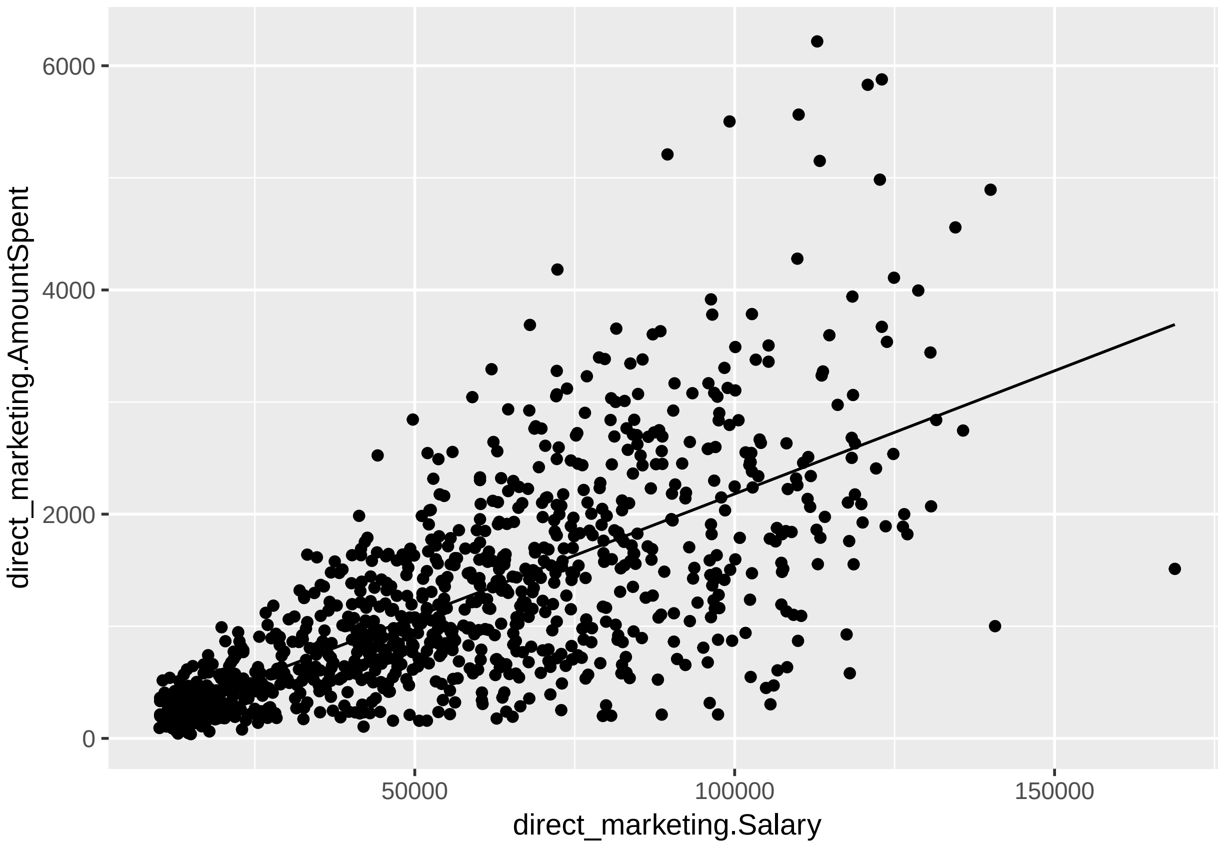
Question 4: Apply polynomial and locfit to the analysis

```r
p1<-ggplot(direct_marketing, aes(x = Gender, y = AmountSpent)) + geom_point()
p2<-ggplot(direct_marketing, aes(x = Location, y = AmountSpent)) + geom_point()
p3<-ggplot(direct_marketing, aes(x = Salary, y = AmountSpent)) + geom_point()
p4<-ggplot(direct_marketing, aes(x = Children, y = AmountSpent)) + geom_point()
p5<-ggplot(direct_marketing, aes(x = History.fix, y = AmountSpent)) + geom_point()
p6<-ggplot(direct_marketing, aes(x = Catalogs, y = AmountSpent)) + geom_point()
library("gridExtra")
grid.arrange(p1, p2, p3, p4, p5,p6, nrow = 3, ncol = 2)
```

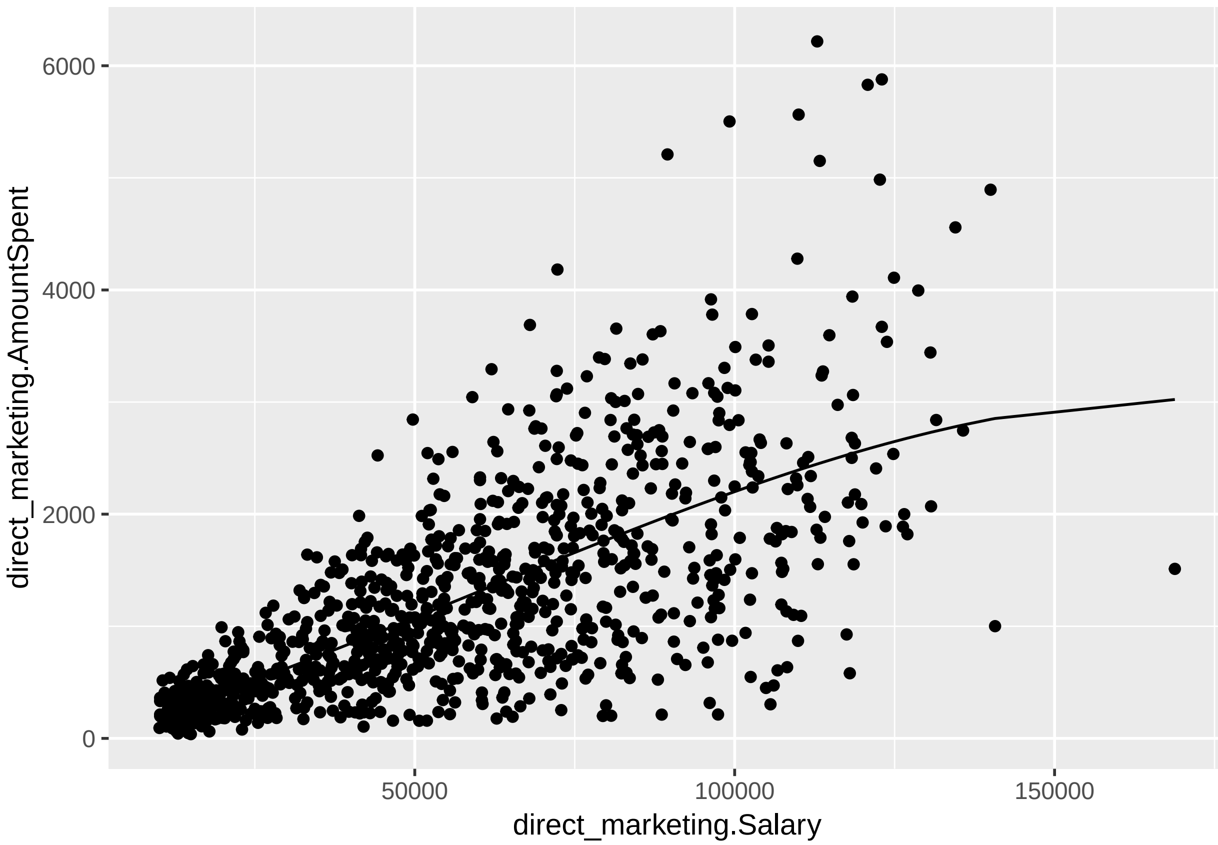


- Apply a polynomial model to the dataset to predict AmountSpent. How does it perform? How did you choose your parameters?

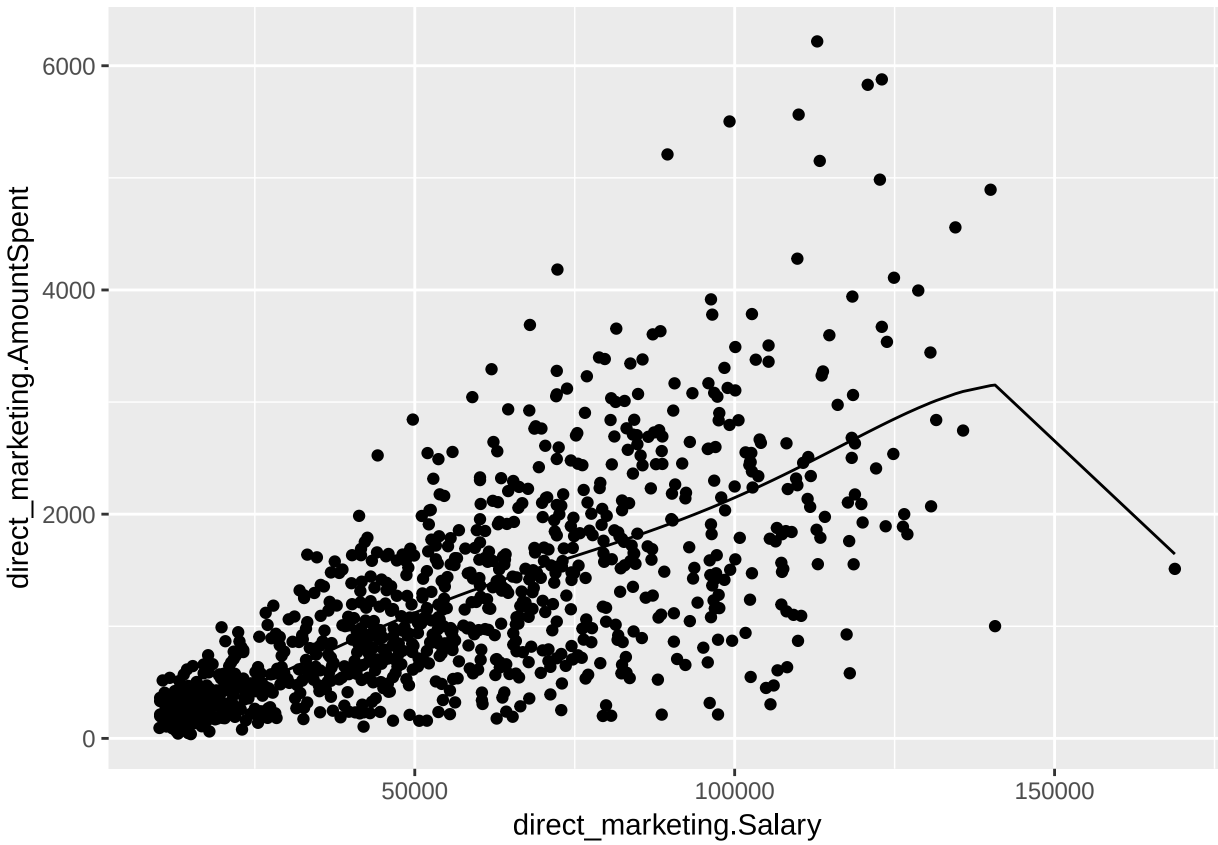
```
poly_direct_marketing <- data.frame(direct_marketing$Salary,direct_marketing$AmountSpent)
# Create polynomial regression models with degrees of 1, 3, 5
#Create polynomial regression with degree of 1
poly.fit <- lm(direct_marketing.AmountSpent ~ poly(direct_marketing.Salary, degree = 1), data = poly_di
poly_direct_marketing <- transform(poly_direct_marketing, PredictedY = predict(poly.fit))
ggplot(poly_direct_marketing , aes(x = direct_marketing.Salary, y = direct_marketing.AmountSpent)) +
  geom_point() + geom_line(data=poly_direct_marketing, aes(x = direct_marketing.Salary, y = PredictedY)
```




```
# #Create polynomial regression with degree of 3
poly.fit <- lm(direct_marketing.AmountSpent ~ poly(direct_marketing.Salary, degree = 3), data = poly_di
poly_direct_marketing <- transform(poly_direct_marketing, PredictedY = predict(poly.fit))
ggplot(poly_direct_marketing , aes(x = direct_marketing.Salary, y = direct_marketing.AmountSpent)) +
  geom_point() + geom_line(data=poly_direct_marketing, aes(x = direct_marketing.Salary, y = PredictedY)
```




```r
# #Create polynomial regression with degree of 5
poly.fit <- lm(direct_marketing.AmountSpent ~ poly(direct_marketing.Salary, degree = 5), data = poly_di
poly_direct_marketing <- transform(poly_direct_marketing, PredictedY = predict(poly.fit))
ggplot(poly_direct_marketing , aes(x = direct_marketing.Salary, y = direct_marketing.AmountSpent)) +
  geom_point() + geom_line(data=poly_direct_marketing, aes(x = direct_marketing.Salary, y = PredictedY)
```



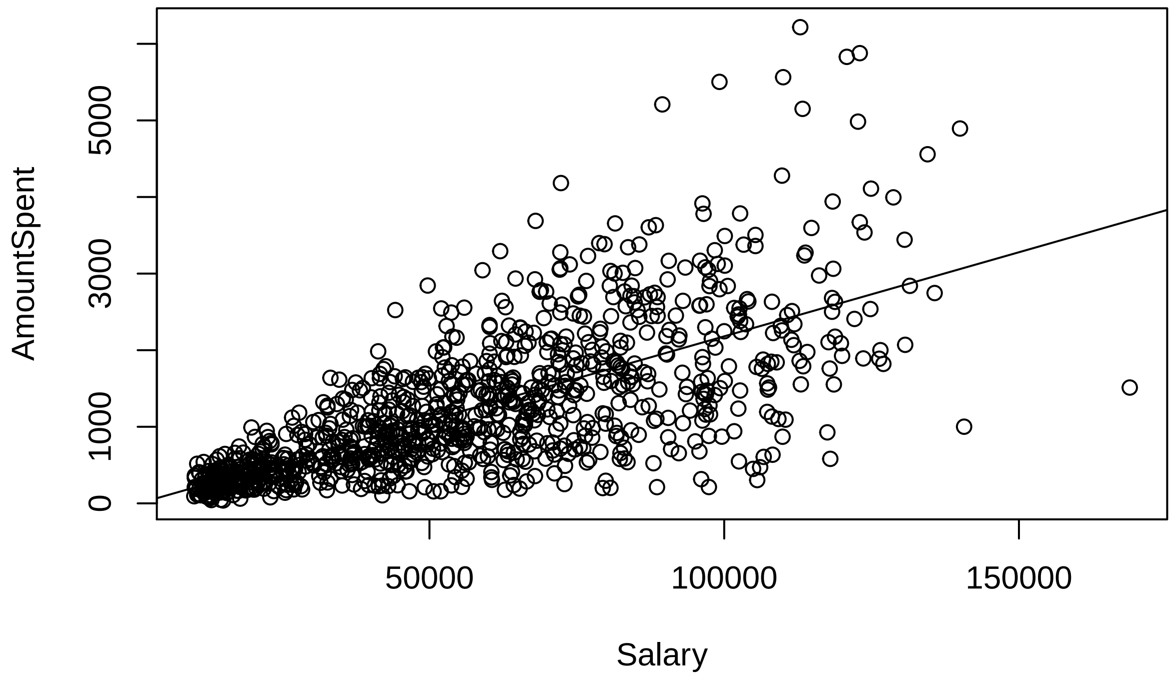
- Apply a locfit model to the dataset to predict AmountSpent. How does it perform?

The locfit model perform better and more reasonable graph than polynomial model.

```r
library(locfit)
```

```
## locfit 1.5-9.1    2013-03-22
```

```r
# Standard regression
regular_lm <- lm(AmountSpent ~Salary, data = direct_marketing)
plot(AmountSpent~Salary, data = direct_marketing)
abline(regular_lm)
```




```r
# fit a local polynomial regression model.  We will use the nearest-neighbor threshold of 0.5, or 50%
poly_fit <- locfit(AmountSpent ~ lp(Salary, nn = 0.5), data = direct_marketing)
plot(poly_fit)
```