Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

Homework1

Sai Charan Talipineni

17 January 2017

1-

Response variable: AmountSpent

Predictor variables: Age, Gender, OwnHome, Married, Location, Salary, Children, History, Catalogs.

2-

a-

There are no explicit missing values in the dataset. The values in the "History" coulmn were set to 'NA' means 'not yet purchased'(according to data description).so, I changed 'NA' to 'not_yet_purchased' to make work easier.

```
orgdata<-read.csv("D:/semester/2nd sem/DATA_MINING/hw1/DirectMarketing.csv")
orgdata$History<-as.character(orgdata$History)
orgdata$History[is.na(orgdata$History)]<-"not_yet_purchased"
orgdata$History<-as.factor(orgdata$History)
```

2b-

```
summary(orgdata)

##      Age         Gender     OwnHome       Married      Location
##  Middle:508   Female:506   Own :516   Married:502   Close:710
##  Old   :205   Male  :494   Rent:484   Single :498   Far  :290
##  Young :287
##
##
##
##      Salary          Children                      History       Catalogs
##  Min.   : 10100   Min.   :0.000   High             :255   Min.   : 6.00
##  1st Qu.: 29975   1st Qu.:0.000   Low              :230   1st Qu.: 6.00
##  Median : 53700   Median :1.000   Medium           :212   Median :12.00
##  Mean   : 56104   Mean   :0.934   not_yet_purchased:303   Mean   :14.68
##  3rd Qu.: 77025   3rd Qu.:2.000                           3rd Qu.:18.00
##  Max.   :168800   Max.   :3.000                           Max.   :24.00
##   AmountSpent
##  Min.   :   38.0
##  1st Qu.: 488.2
##  Median : 962.0
##  Mean   :1216.8
```

Name: Sai Charan Talipineni Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
##  3rd Qu.:1688.5
##  Max.   :6217.0
```

```r
print("Salary")
```

```
## [1] "Salary"
```

```r
summary(orgdata$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10100   29980   53700   56100   77020  168800
```

```r
sd(orgdata$Salary)
```

```
## [1] 30616.31
```

```r
print("Children")
```

```
## [1] "Children"
```

```r
summary(orgdata$Children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   0.934   2.000   3.000
```

```r
sd(orgdata$Children)
```

```
## [1] 1.05107
```

```r
print("Catalogs")
```

```
## [1] "Catalogs"
```

```r
summary(orgdata$Catalogs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00    6.00   12.00   14.68   18.00   24.00
```

```r
sd(orgdata$Catalogs)
```

```
## [1] 6.622895
```

```r
print("AmountSpent")
```

```
## [1] "AmountSpent"
```

```r
summary(orgdata$AmountSpent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    38.0   488.2   962.0  1217.0  1688.0  6217.0
```

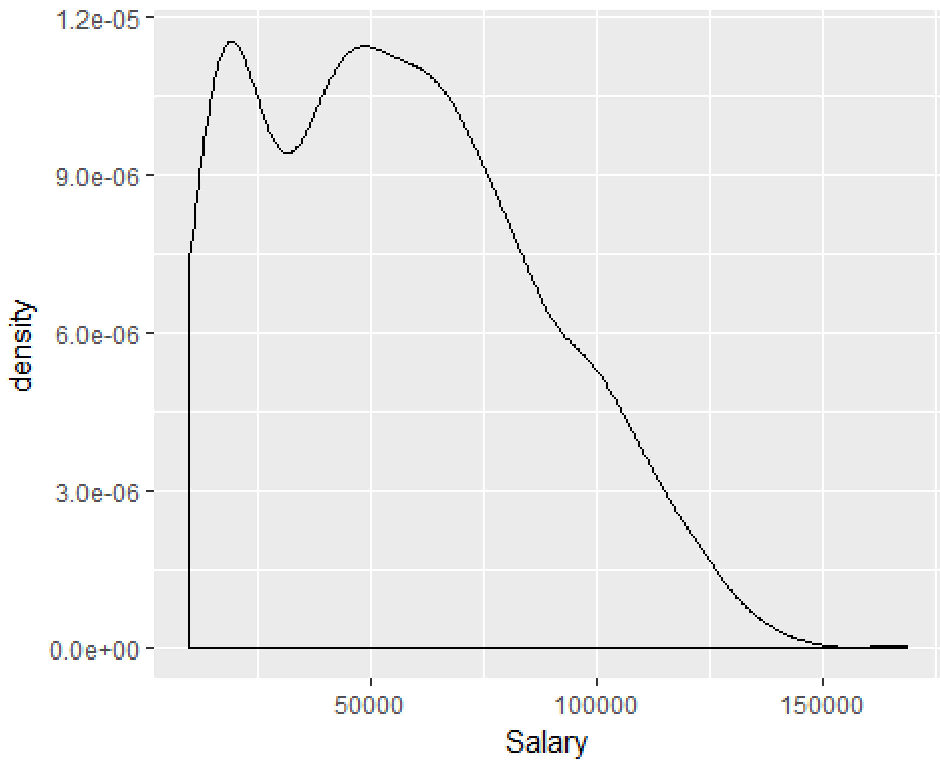
```r
sd(orgdata$AmountSpent)
```

```
## [1] 961.0686
```

Name: Sai Charan Talipineni Email-ID: sat122@pitt.edu
Pitt ID: 4105769

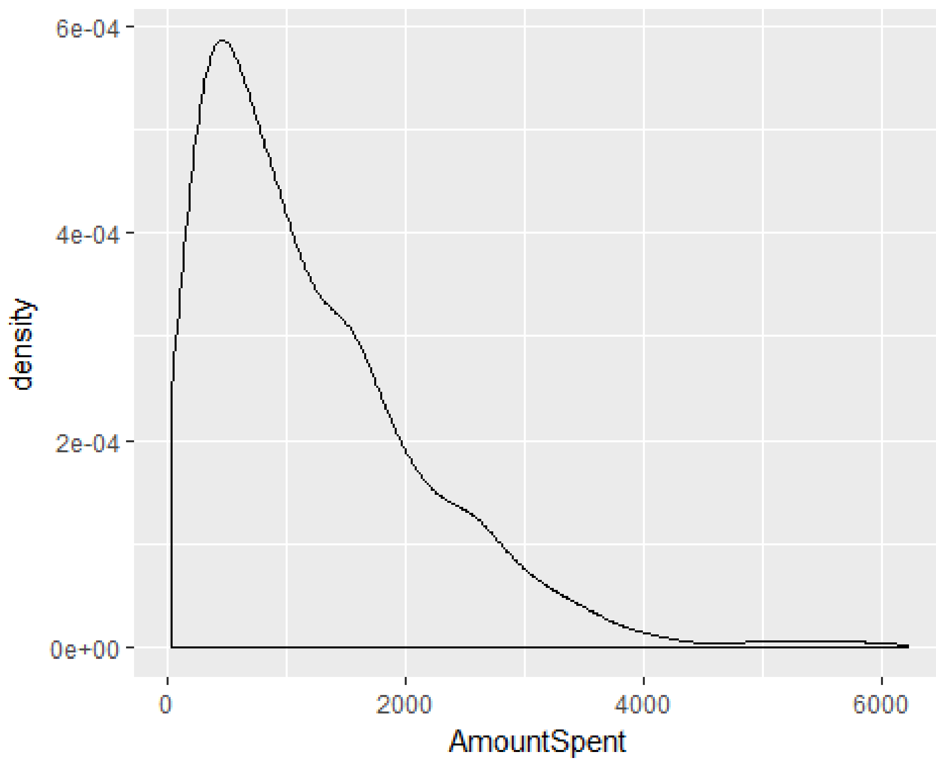
2c-

```
require(ggplot2)

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.3.2

ggplot(orgdata, aes(x=Salary)) + geom_density() #right skewed/positive skewed distribution
```



Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
ggplot(orgdata, aes(x=AmountSpent)) + geom_density() #right skewed/positive
skewed distribution
```



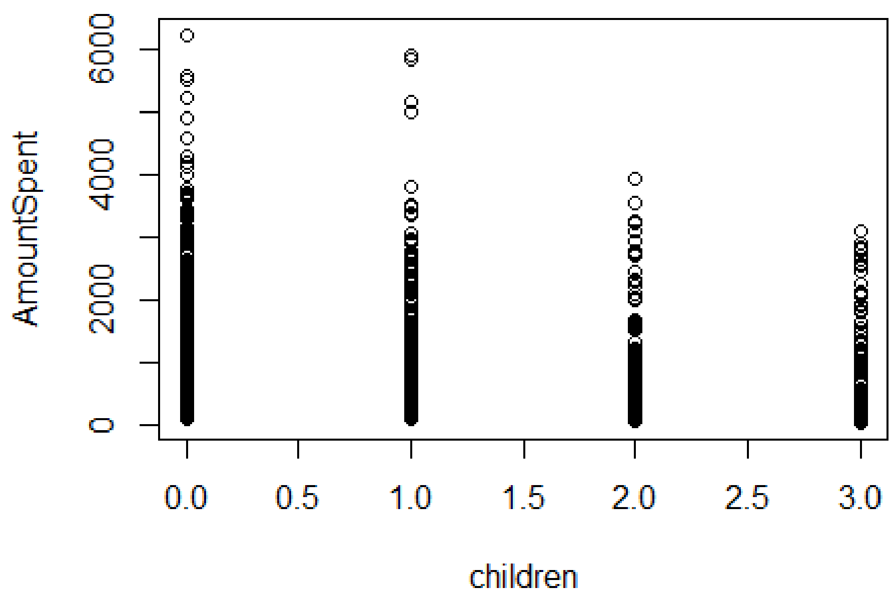
Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

2d-

```
cor(orgdata$Salary, orgdata$AmountSpent)

## [1] 0.6995957

#Positive correlation

plot(orgdata$Salary, orgdata$AmountSpent, xlab="Salary", ylab="AmountSpent")
```

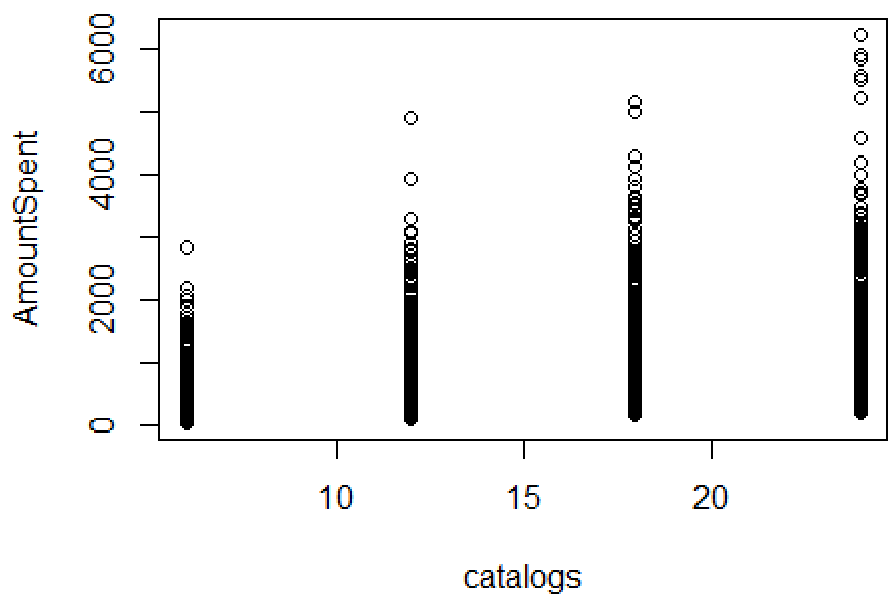
Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
cor(orgdata$Children, orgdata$AmountSpent)

## [1] -0.2223082

#Negative correlation

plot(orgdata$Children, orgdata$AmountSpent, xlab="children",
ylab="AmountSpent")
```



Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

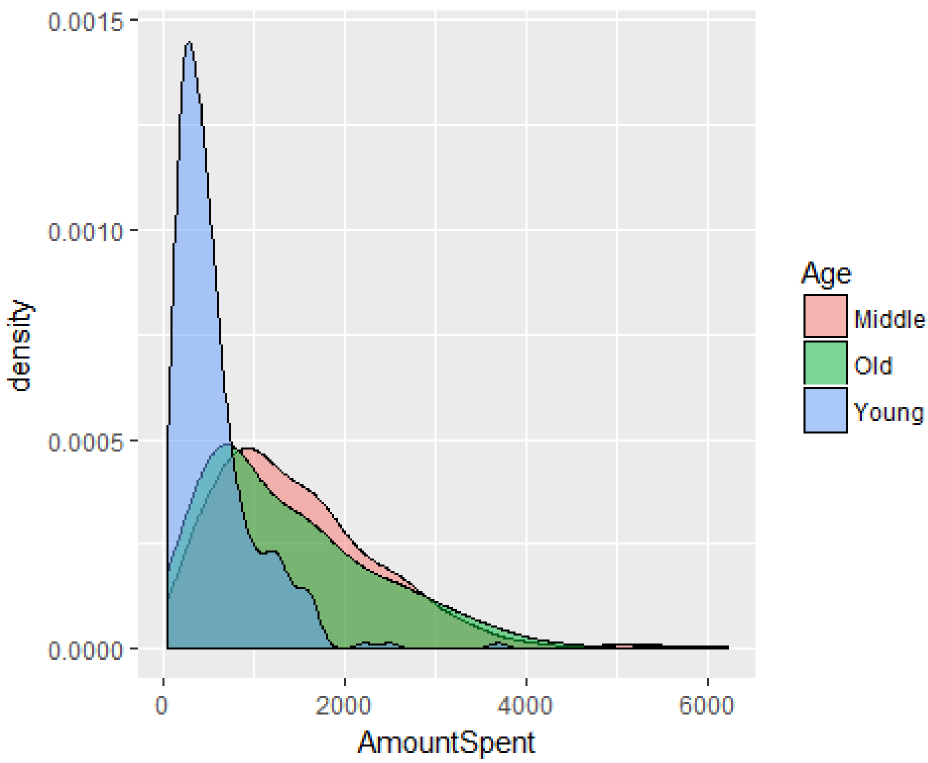
```
cor(orgdata$Catalogs, orgdata$AmountSpent)

## [1] 0.4726499

#Positive correlation but less than Salary

plot(orgdata$Catalogs, orgdata$AmountSpent, xlab="catalogs",
ylab="AmountSpent")
```



Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

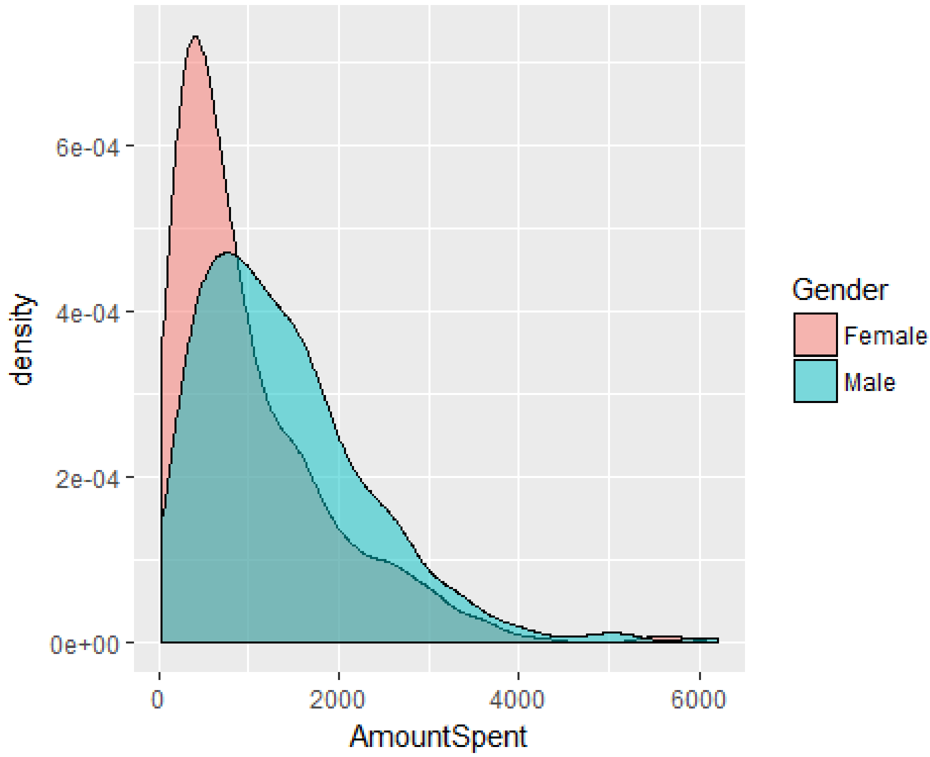
2e-

```
ggplot(orgdata, aes(x = AmountSpent, fill = Age)) + geom_density(alpha = 0.5)
```

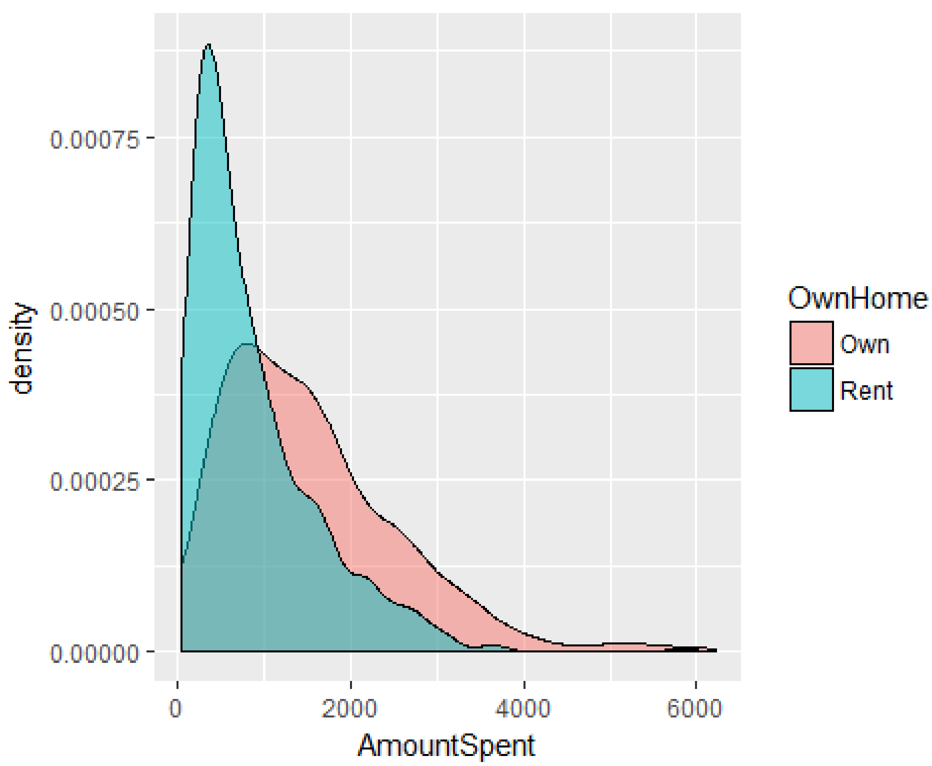



```
ggplot(orgdata, aes(x = AmountSpent, fill = Gender)) + geom_density(alpha = 
0.5)
```

Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

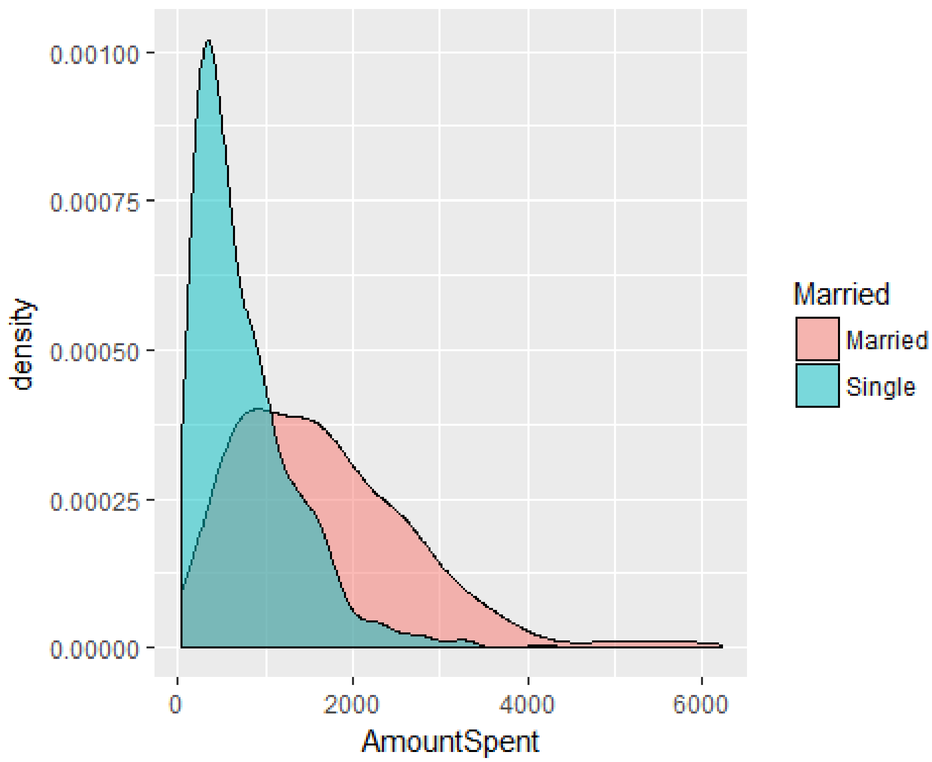



```
ggplot(orgdata, aes(x = AmountSpent, fill = OwnHome)) +  geom_density(alpha = 0.5)
```



Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

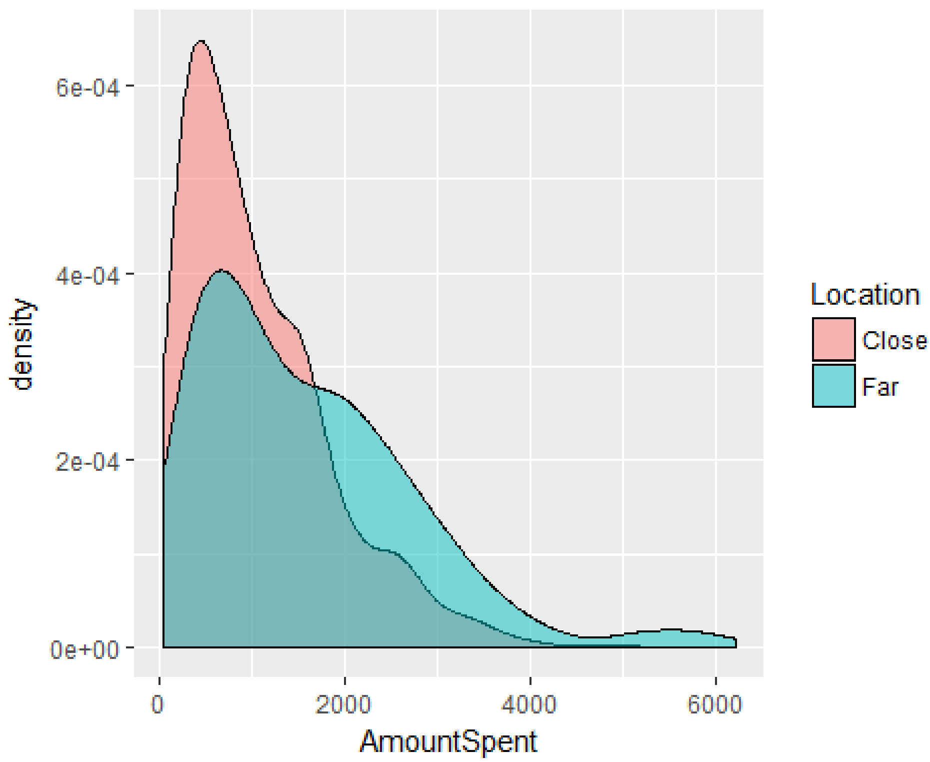
```
ggplot(orgdata, aes(x = AmountSpent, fill = Married)) +  geom_density(alpha =
0.5)
```



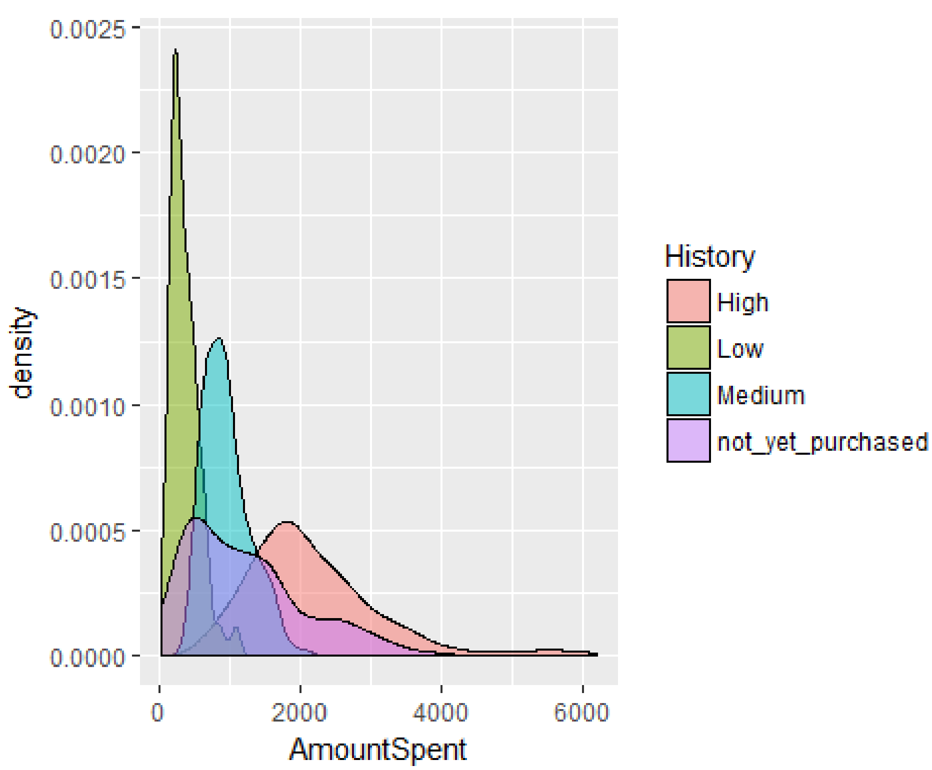

```
ggplot(orgdata, aes(x = AmountSpent, fill = Location)) +  geom_density(alpha
= 0.5)
```

Name: Sai Charan Talipineni

Email-ID: sat122@pitt.edu
Pitt ID: 4105769




```
ggplot(orgdata, aes(x = AmountSpent, fill = History)) +  geom_density(alpha = 0.5)
```



Name: Sai Charan Talipineni Email-ID: sat122@pitt.edu
Pitt ID: 4105769

2f-

```
#Amountspent and Age
Middle_sum<-0
Middle_count<-0
Young_sum<-0
Young_count<-0
Old_sum<-0
Old_count<-0
for(i in 1:1000){
  if(orgdata$Age[i]=="Young"){
    Young_sum<-Young_sum+(orgdata$AmountSpent)[i]
    Young_count<-Young_count+1
  }
  else if(orgdata$Age[i]=="Middle"){
    Middle_sum<-Middle_sum+(orgdata$AmountSpent)[i]
    Middle_count<-Middle_count+1
  }
 else if(orgdata$Age[i]=="Old"){
    Old_sum<-Old_sum+(orgdata$AmountSpent)[i]
    Old_count<-Old_count+1
  }
}

Age_Young_Mean<-(Young_sum/Young_count)
Age_Middle_Mean<-(Middle_sum/Middle_count)
Age_Old_Mean<-(Old_sum/Old_count)
Age_Young_Mean

## [1] 558.6237

Age_Middle_Mean

## [1] 1501.691

Age_Old_Mean

## [1] 1432.127

#From the above observations, we can say that Middle>Old>Young (Means)

#Amountspent and Gender
Male_sum<-0
Male_count<-0
Female_sum<-0
Female_count<-0
for(i in 1:1000){
  if(orgdata$Gender[i]=="Female"){
    Female_sum<-Female_sum+(orgdata$AmountSpent)[i]
    Female_count<-Female_count+1
  }
```

```
  else if(orgdata$Gender[i]=="Male"){
    Male_sum<-Male_sum+(orgdata$AmountSpent)[i]
    Male_count<-Male_count+1
  }
}
Gender_Female_Mean<-(Female_sum/Female_count)
Gender_Male_Mean<-(Male_sum/Male_count)
Gender_Female_Mean

## [1] 1025.34

Gender_Male_Mean

## [1] 1412.85

#From the above observations, we can say that Male>Female (Means)

#Amountspent and Location
Far_sum<-0
Far_count<-0
Close_sum<-0
Close_count<-0
for(i in 1:1000){
  if(orgdata$Location[i]=="Close"){
    Close_sum<-Close_sum+(orgdata$AmountSpent)[i]
    Close_count<-Close_count+1
  }
  else if(orgdata$Location[i]=="Far"){
    Far_sum<-Far_sum+(orgdata$AmountSpent)[i]
    Far_count<-Far_count+1
  }
}
Location_Close_Mean<-(Close_sum/Close_count)
Location_Far_Mean<-(Far_sum/Far_count)
Location_Close_Mean

## [1] 1061.686

Location_Far_Mean

## [1] 1596.459

#From the above observations, we can say that Far>Close (Means)

#Amountspent and History
High_sum<-0
High_count<-0
Medium_sum<-0
Medium_count<-0
Low_sum<-0
Low_count<-0
not_yet_purchased_sum<-0
```

Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
not_yet_purchased_count<-0
for(i in 1:1000){
  if(orgdata$History[i]=="Medium"){
    Medium_sum<-Medium_sum+(orgdata$AmountSpent)[i]
    Medium_count<-Medium_count+1
  }
  else if(orgdata$History[i]=="High"){
    High_sum<-High_sum+(orgdata$AmountSpent)[i]
    High_count<-High_count+1
  }
 else if(orgdata$History[i]=="Low"){
    Low_sum<-Low_sum+(orgdata$AmountSpent)[i]
    Low_count<-Low_count+1
  }
  else if(orgdata$History[i]=="not_yet_purchased"){
    not_yet_purchased_sum<-not_yet_purchased_sum+(orgdata$AmountSpent)[i]
    not_yet_purchased_count<-not_yet_purchased_count+1
  }
}

History_Medium_Mean<-(Medium_sum/Medium_count)
History_High_Mean<-(High_sum/High_count)
History_Low_Mean<-(Low_sum/Low_count)
History_not_yet_purchase_Mean<-
(not_yet_purchased_sum/not_yet_purchased_count)
History_Medium_Mean

## [1] 950.4009

History_High_Mean

## [1] 2186.137

History_Low_Mean

## [1] 357.087

History_not_yet_purchase_Mean

## [1] 1239.901

#From the above observations, we can say that
High>not_yet_purchased>Medium>Low (Means)
```

For the remaining categorical attributes, with the above numerical data and conditional density plots, we can conclude that, (Means)

OwnHome:

Own>Rent

```
Marriage:

Married>Single
```

3a-

```
as_lr<-lm(AmountSpent~(Age)+ (Gender)+ (OwnHome)+ (Married)+ (Location)+
Salary+ Children+ (History)+ Catalogs, data=orgdata)
summary(as_lr)

##
## Call:
## lm(formula = AmountSpent ~ (Age) + (Gender) + (OwnHome) + (Married) +
##     (Location) + Salary + Children + (History) + Catalogs, data = orgdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1711.44  -292.41   -17.56   237.87  2876.91
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -285.74892  116.39444  -2.455   0.0143 *
## AgeOld                     63.36828   47.79586   1.326   0.1852
## AgeYoung                    8.90120   49.70059   0.179   0.8579
## GenderMale                -46.99837   32.85192  -1.431   0.1529
## OwnHomeRent               -16.63382   36.64327  -0.454   0.6500
## MarriedSingle              32.74314   44.54067   0.735   0.4624
## LocationFar               436.50575   35.92138  12.152  < 2e-16 ***
## Salary                      0.01920    0.00103  18.652  < 2e-16 ***
## Children                 -162.73555   18.00348  -9.039  < 2e-16 ***
## HistoryLow               -352.89534   65.57529  -5.382 9.23e-08 ***
## HistoryMedium            -404.41014   52.94420  -7.638 5.19e-14 ***
## Historynot_yet_purchased    6.99218   51.32915   0.136   0.8917
## Catalogs                   41.86880    2.45796  17.034  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 485.7 on 987 degrees of freedom
## Multiple R-squared:  0.7476, Adjusted R-squared:  0.7446
## F-statistic: 243.7 on 12 and 987 DF,  p-value: < 2.2e-16

#RMSE
n = length(orgdata$AmountSpent)
error = dim(n)
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent~(Age)+ (Gender)+ (OwnHome)+ (Married)+ (Location)+
Salary+ Children+ (History)+ Catalogs, data=orgdata[train, ])
```

Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
  pred = predict(m2, newdat=orgdata[-train ,])
  obs = orgdata$AmountSpent[-train]
  error[k] = obs-pred
}
me=mean(error)
rmse=sqrt(mean(error^2))
rmse

## [1] 489.3011
```

3b-

```
#linear
Y<-orgdata$AmountSpent
as_lr1<-lm(Y~Salary+Children+Catalogs, data=orgdata)
summary(as_lr1)

##
## Call:
## lm(formula = Y ~ Salary + Children + Catalogs, data = orgdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1775.9  -348.7   -38.7   255.5  3211.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.428e+02  5.372e+01  -8.242 5.29e-16 ***
## Salary       2.041e-02  5.929e-04  34.417  < 2e-16 ***
## Children    -1.987e+02  1.709e+01 -11.628  < 2e-16 ***
## Catalogs     4.770e+01  2.755e+00  17.310  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 562.5 on 996 degrees of freedom
## Multiple R-squared:  0.6584, Adjusted R-squared:  0.6574
## F-statistic:   640 on 3 and 996 DF,  p-value: < 2.2e-16

#evaluation_linear_rmse
n = length(Y)
error = dim(n)
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ (Salary+Children+Catalogs), data=orgdata[train, ])
  pred = predict(m2, newdat=orgdata[-train ,])
  obs = orgdata$AmountSpent[-train]
  error[k] = obs-pred
```

```
}
me=mean(error)
rmse=sqrt(mean(error^2))
rmse

## [1] 564.2606

#linear
as_lr2<-lm(Y~Salary+Catalogs+Location+History, data=orgdata)
summary(as_lr2)

##
## Call:
## lm(formula = Y ~ Salary + Catalogs + Location + History, data = orgdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1388.30  -317.28   -18.88   260.36  3085.80
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.025e+02  8.210e+01  -1.249 0.212041
## Salary                   1.629e-02  6.682e-04  24.379  < 2e-16 ***
## Catalogs                 4.218e+01  2.576e+00  16.373  < 2e-16 ***
## LocationFar              3.822e+02  3.729e+01  10.252  < 2e-16 ***
## HistoryLow              -6.573e+02  6.104e+01 -10.770  < 2e-16 ***
## HistoryMedium           -5.582e+02  5.288e+01 -10.555  < 2e-16 ***
## Historynot_yet_purchased -1.824e+02  5.034e+01  -3.623 0.000306 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 510.4 on 993 degrees of freedom
## Multiple R-squared:  0.7197, Adjusted R-squared:  0.718
## F-statistic: 424.9 on 6 and 993 DF,  p-value: < 2.2e-16

#evaluation_linear_rmse
n = length(Y)
error = dim(n)
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ (Salary+Catalogs+Location+History),
data=orgdata[train, ])
  pred = predict(m2, newdat=orgdata[-train ,])
  obs = orgdata$AmountSpent[-train]
  error[k] = obs-pred
}
me=mean(error)
rmse=sqrt(mean(error^2))
rmse
```

Name: Sai Charan Talipineni Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
## [1] 512.6858
```

```
#non_linear
summary(lm(Y ~ (poly(Salary,degree = 3)+poly(Children,
degree=3)+poly(Catalogs,degree = 3)), data = orgdata))
```

```
##
## Call:
## lm(formula = Y ~ (poly(Salary, degree = 3) + poly(Children, degree = 3) +
##     poly(Catalogs, degree = 3)), data = orgdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1446.17  -341.91   -40.52   245.91  3154.38
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1216.770     17.732  68.620   <2e-16 ***
## poly(Salary, degree = 3)1    19783.922    573.256  34.512   <2e-16 ***
## poly(Salary, degree = 3)2      627.740    564.356   1.112   0.2663
## poly(Salary, degree = 3)3     -952.588    562.780  -1.693   0.0908 .
## poly(Children, degree = 3)1 -6725.880    568.229 -11.837   <2e-16 ***
## poly(Children, degree = 3)2    -9.906    563.149  -0.018   0.9860
## poly(Children, degree = 3)3   971.219    562.656   1.726   0.0846 .
## poly(Catalogs, degree = 3)1 10035.034    577.331  17.382   <2e-16 ***
## poly(Catalogs, degree = 3)2   843.720    563.770   1.497   0.1348
## poly(Catalogs, degree = 3)3  -941.450    561.343  -1.677   0.0938 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 560.7 on 990 degrees of freedom
## Multiple R-squared:  0.6627, Adjusted R-squared:  0.6596
## F-statistic: 216.1 on 9 and 990 DF,  p-value: < 2.2e-16
```

```
#evaluation_nonlinear_rmse
n = length(Y)
error = dim(n)
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ (poly(Salary,degree = 3)+poly(Children,
degree=3)+poly(Catalogs,degree = 3)), data=orgdata[train, ])
  pred = predict(m2, newdat=orgdata[-train ,])
  obs = orgdata$AmountSpent[-train]
  error[k] = obs-pred
}
me=mean(error)
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 569.2569
```

Name: Sai Charan Talipineni Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
#non_linear
summary(lm(Y ~ (poly(Salary,degree = 3)+poly(Catalogs,
degree=3)+Location+History), data = orgdata))

##
## Call:
## lm(formula = Y ~ (poly(Salary, degree = 3) + poly(Catalogs, degree = 3) +
##     Location + History), data = orgdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1719.57  -319.57   -17.76   260.99  3062.65
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1434.00      42.93  33.401  < 2e-16 ***
## poly(Salary, degree = 3)1   15705.85     653.97  24.016  < 2e-16 ***
## poly(Salary, degree = 3)2     517.74     547.25   0.946 0.344340
## poly(Salary, degree = 3)3     267.42     528.00   0.506 0.612632
## poly(Catalogs, degree = 3)1  8853.43     540.22  16.389  < 2e-16 ***
## poly(Catalogs, degree = 3)2   325.67     513.77   0.634 0.526304
## poly(Catalogs, degree = 3)3  -424.28     514.36  -0.825 0.409648
## LocationFar                   379.49      37.39  10.148  < 2e-16 ***
## HistoryLow                   -669.63      62.66 -10.687  < 2e-16 ***
## HistoryMedium                -553.04      54.67 -10.115  < 2e-16 ***
## Historynot_yet_purchased     -184.90      50.75  -3.644 0.000283 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 510.8 on 989 degrees of freedom
## Multiple R-squared:  0.7203, Adjusted R-squared:  0.7175
## F-statistic: 254.7 on 10 and 989 DF,  p-value: < 2.2e-16

#evaluation_nonlinear_rmse
n = length(Y)
error = dim(n)
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ (poly(Salary,degree = 3)+poly(Catalogs,
degree=3)+Location+History), data=orgdata[train, ])
  pred = predict(m2, newdat=orgdata[-train ,])
  obs = orgdata$AmountSpent[-train]
  error[k] = obs-pred
}
me=mean(error)
rmse=sqrt(mean(error^2))
rmse

## [1] 522.0549
```

Name: Sai Charan Talipineni Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
#non_linear
summary(lm(Y ~ (poly(Salary,degree = 3)+poly(Catalogs,
degree=3)+Location+History+Age+Children+Gender), data = orgdata))

##
## Call:
## lm(formula = Y ~ (poly(Salary, degree = 3) + poly(Catalogs, degree = 3) +
##     Location + History + Age + Children + Gender), data = orgdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1726.91  -293.42   -17.49   245.35  2871.56
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1420.27      52.49  27.058  < 2e-16 ***
## poly(Salary, degree = 3)1   18172.66     816.81  22.248  < 2e-16 ***
## poly(Salary, degree = 3)2     242.65     569.36   0.426    0.670
## poly(Salary, degree = 3)3     -80.06     507.75  -0.158    0.875
## poly(Catalogs, degree = 3)1  8751.93     514.60  17.007  < 2e-16 ***
## poly(Catalogs, degree = 3)2   520.24     489.21   1.063    0.288
## poly(Catalogs, degree = 3)3  -519.18     489.77  -1.060    0.289
## LocationFar                   434.20      35.98  12.067  < 2e-16 ***
## HistoryLow                   -357.41      67.32  -5.309 1.36e-07 ***
## HistoryMedium                -396.82      54.31  -7.307 5.66e-13 ***
## Historynot_yet_purchased       10.11      51.84   0.195    0.845
## AgeOld                         56.57      46.19   1.225    0.221
## AgeYoung                      -12.67      51.01  -0.248    0.804
## Children                     -163.47      18.02  -9.070  < 2e-16 ***
## GenderMale                    -43.39      32.92  -1.318    0.188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 485.8 on 985 degrees of freedom
## Multiple R-squared:  0.748,  Adjusted R-squared:  0.7445
## F-statistic: 208.9 on 14 and 985 DF,  p-value: < 2.2e-16

#evaluation_nonlinear_rmse
n = length(Y)
error = dim(n)
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ (poly(Salary,degree = 3)+poly(Catalogs,
degree=3)+Location+History+Age+Children+Gender), data=orgdata[train, ])
  pred = predict(m2, newdat=orgdata[-train ,])
  obs = orgdata$AmountSpent[-train]
  error[k] = obs-pred
}
me=mean(error)
```

```
rmse=sqrt(mean(error^2))
rmse

## [1] 498.0905
```

3c-

I considered the best model as the one (3a) with highest R^2 and low RMSE as the linear regression model with all numerical and categorical variables.

```
library(MASS)
fit = lm(AmountSpent~(Age)+ (Gender)+ (OwnHome)+ (Married)+ (Location)+
Salary+ Children+ (History)+ Catalogs, data=orgdata)
stepAIC(fit, direction="backward")

## Start:  AIC=12384.2
## AmountSpent ~ (Age) + (Gender) + (OwnHome) + (Married) + (Location) +
##     Salary + Children + (History) + Catalogs
##
##             Df Sum of Sq       RSS   AIC
## - Age        2    443097 233304046 12382
## - OwnHome    1     48616 232909565 12382
## - Married    1    127499 232988448 12383
## <none>                   232860949 12384
## - Gender     1    482863 233343812 12384
## - Children   1  19276638 252137587 12462
## - History    3  28426404 261287353 12493
## - Location   1  34838025 267698974 12522
## - Catalogs   1  68455782 301316731 12640
## - Salary     1  82083034 314943983 12684
##
## Step:  AIC=12382.1
## AmountSpent ~ Gender + OwnHome + Married + Location + Salary +
##     Children + History + Catalogs
##
##             Df Sum of Sq       RSS   AIC
## - Married    1     55318 233359364 12380
## - OwnHome    1    147202 233451248 12381
## <none>                   233304046 12382
## - Gender     1    664803 233968849 12383
## - Children   1  24879626 258183672 12481
## - History    3  28889456 262193501 12493
## - Location   1  35011045 268315091 12520
## - Catalogs   1  68392954 301697000 12637
## - Salary     1 107651722 340955767 12760
##
## Step:  AIC=12380.33
## AmountSpent ~ Gender + OwnHome + Location + Salary + Children +
##     History + Catalogs
##
```

Name: Sai Charan Talipineni
Email-ID: sat122@pitt.edu
Pitt ID: 4105769

```
##             Df Sum of Sq       RSS   AIC
## - OwnHome   1    162809 233522173 12379
## <none>                  233359364 12380
## - Gender    1    634446 233993810 12381
## - Children  1  24825054 258184418 12479
## - History   3  29027254 262386618 12492
## - Location  1  34961973 268321337 12518
## - Catalogs  1  68354217 301713581 12635
## - Salary    1 153879921 387239285 12885
##
## Step:  AIC=12379.03
## AmountSpent ~ Gender + Location + Salary + Children + History +
##     Catalogs
##
##             Df Sum of Sq       RSS   AIC
## <none>                  233522173 12379
## - Gender    1    670888 234193061 12380
## - Children  1  24994947 258517120 12479
## - History   3  29194376 262716549 12491
## - Location  1  34842146 268364319 12516
## - Catalogs  1  68330846 301853019 12634
## - Salary    1 177237435 410759607 12942

##
## Call:
## lm(formula = AmountSpent ~ Gender + Location + Salary + Children +
##     History + Catalogs, data = orgdata)
##
## Coefficients:
##               (Intercept)                 GenderMale
##                -228.38437                  -54.28354
##               LocationFar                     Salary
##                 436.04608                    0.01892
##                  Children                 HistoryLow
##                -171.98225                 -355.05647
##             HistoryMedium  Historynot_yet_purchased
##                -408.81287                   -0.03510
##                  Catalogs
##                  41.74594
```

The most important predictors are seem to be above all from the observations and the probabilities of corresponding values in the summary of the above best model. In all these, Salary predictor is seems to be the best one with low probability and correlation with AmountSpent and an important predictor from StepAIC.