Assignment2

Liju Robin George

January 27, 2017

1. Data Description and Identifying Target and Predictor variables

The dataset consists of n=2000 individual tax returns.Given below are the variables,their description, their type and whether they are predictors, non-predictors or target variables

1. ID - unique identifier - Numerical - Non-Predictor
2. Age - Age of individual - Numerical - Predictor
3. Employment - Employment Type - Categorical - Predictor
4. Education - Highest level of education - Categorical - Predictor
5. Marital - Marital status - Categorical - Predictor
6. Occupation - Occupation Type - Categorical - Predictor
7. Income - Income declared - Numerical - Predictor
8. Gender - Gender of individual - Categorical - Predictor
9. Deductions - Financial claim - Numerical - Predictor
10. Hours - Avg working hours per week - Numerical - Predictor
11. RISK_Adjustment - Amount of monetary adjustment on claim. Records the risk associated with an individual - Numerical - Target variable
12. TARGET_Adjusted - Variable deciding a productive or non-productive audit. Productive being adjustment being made - Categorical - Target variable

The goal is to predict the binary TARGET_Adjusted and the continuous RISK_adjustment variables.

2.Descriptive and Graphical Statistics

2.a. Summary of data and of Numerical variables

Below is the summary of the data

```
variable.names(audit)
```

```
##  [1] "ID"              "Age"             "Employment"
##  [4] "Education"       "Marital"         "Occupation"
##  [7] "Income"          "Gender"          "Deductions"
## [10] "Hours"           "RISK_Adjustment" "TARGET_Adjusted"
```

```
summary(audit)
```

```
##        ID                Age              Employment        Education
##  Min.   :1004641   Min.   :17.00   Private    :1411   HSgrad    :660
##  1st Qu.:3437052   1st Qu.:28.00   Consultant: 148   College   :442
##  Median :5638451   Median :37.00   PSLocal    : 119   Bachelor  :345
##  Mean   :5624348   Mean   :38.62   SelfEmp    :  79   Master    :102
##  3rd Qu.:7876535   3rd Qu.:48.00   PSState    :  72   Vocational: 86
##  Max.   :9996101   Max.   :90.00   (Other)    :  71   Yr11      : 74
##                                    NA's       : 100   (Other)   :291
##                  Marital           Occupation      Income
##  Absent               :669   Executive    :289   Min.   :   609.7
##  Divorced             :266   Professional:247   1st Qu.: 34433.1
##  Married              :917   Clerical     :232   Median : 59768.9
##  Married-spouse-absent: 22   Repair       :225   Mean   : 84688.5
##  Unmarried            : 67   Service      :210   3rd Qu.:113842.9
##  Widowed              : 59   (Other)      :696   Max.   :481259.5
##                              NA's         :101
##     Gender       Deductions          Hours       RISK_Adjustment
##  Female: 632   Min.   :   0.00   Min.   : 1.00   Min.   : -1453
##  Male  :1368   1st Qu.:   0.00   1st Qu.:38.00   1st Qu.:     0
##                Median :   0.00   Median :40.00   Median :     0
##                Mean   :  67.57   Mean   :40.07   Mean   :  2021
##                3rd Qu.:   0.00   3rd Qu.:45.00   3rd Qu.:     0
##                Max.   :2904.00   Max.   :99.00   Max.   :112243
##
##  TARGET_Adjusted
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2315
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```r
dim(audit)
```

```
## [1] 2000   12
```

We can observe that there are some missing values in the data for columns Employment and Occupation. Let us try to handle them. We can also see that TARGET_Adjusted is numerical. I do not change it into a factor here, rather, I handle it in the sections below while predicting for it.

```r
audit.clean = audit

audit.clean$Occupation<- as.character(audit.clean$Occupation)
audit.clean$Occupation[is.na(audit.clean$Occupation)] <- "Unknown"
audit.clean$Occupation<- factor (audit.clean$Occupation)

audit.clean$Employment<- as.character(audit.clean$Employment)
audit.clean$Employment[is.na(audit.clean$Employment)] <- "Unknown"
audit.clean$Employment<- factor(audit.clean$Employment)

summary(audit.clean)
```

```
##        ID                Age             Employment         Education
##  Min.   :1004641   Min.   :17.00   Private    :1411   HSgrad    :660
##  1st Qu.:3437052   1st Qu.:28.00   Consultant: 148   College   :442
##  Median :5638451   Median :37.00   PSLocal    : 119   Bachelor  :345
##  Mean   :5624348   Mean   :38.62   Unknown    : 100   Master    :102
##  3rd Qu.:7876535   3rd Qu.:48.00   SelfEmp    :  79   Vocational: 86
##  Max.   :9996101   Max.   :90.00   PSState    :  72   Yr11      : 74
##                                    (Other)    :  71   (Other)   :291
##                   Marital          Occupation      Income
##  Absent               :669   Executive    :289   Min.   :   609.7
##  Divorced             :266   Professional:247   1st Qu.: 34433.1
##  Married              :917   Clerical     :232   Median : 59768.9
##  Married-spouse-absent: 22   Repair       :225   Mean   : 84688.5
##  Unmarried            : 67   Service      :210   3rd Qu.:113842.9
##  Widowed              : 59   Sales        :206   Max.   :481259.5
##                              (Other)      :591
##     Gender      Deductions          Hours       RISK_Adjustment
##  Female: 632   Min.   :   0.00   Min.   : 1.00   Min.   : -1453
##  Male  :1368   1st Qu.:   0.00   1st Qu.:38.00   1st Qu.:     0
##                Median :   0.00   Median :40.00   Median :     0
##                Mean   :  67.57   Mean   :40.07   Mean   :  2021
##                3rd Qu.:   0.00   3rd Qu.:45.00   3rd Qu.:     0
##                Max.   :2904.00   Max.   :99.00   Max.   :112243
##
##  TARGET_Adjusted
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2315
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

Now let us look at the summary and statistics for each of the Numerical

```r
#Standard Deviation
standardDevs = c(Age = sd(audit.clean$Age), Income = sd(audit.clean$Income),
                 Deductions = sd(audit.clean$Deductions), Hours = sd(audit.clean$Hours),
                 RISK_Adjustment=sd(audit.clean$RISK_Adjustment))
standardDevs
```

```
##             Age          Income      Deductions           Hours
##        13.58475     69621.64450       340.70470        12.15372
## RISK_Adjustment
##      8341.87229
```

```r
#Summary
totSummary = c(Age = summary(audit.clean$Age), Income = summary(audit.clean$Income),
               Deductions = summary(audit.clean$Deductions), Hours = summary(audit.clean$Hours),
               HoursRISK_Adjustment=summary(audit.clean$RISK_Adjustment))
totSummary
```

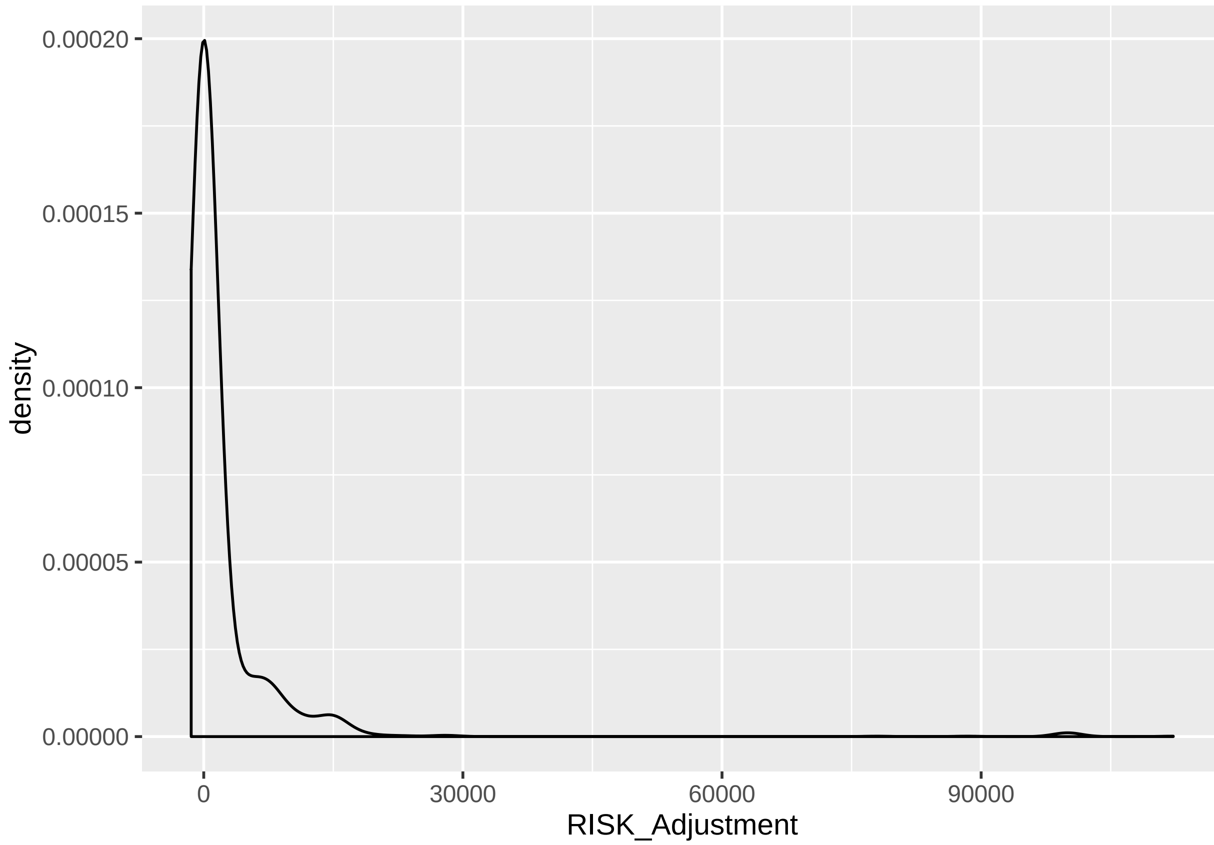
```
##                       Age.Min.                    Age.1st Qu.
```

```
##                           17.00                          28.00
##                      Age.Median                       Age.Mean
##                           37.00                          38.62
##                    Age.3rd Qu.                       Age.Max.
##                           48.00                          90.00
##                    Income.Min.                 Income.1st Qu.
##                          609.70                       34430.00
##                   Income.Median                    Income.Mean
##                        59770.00                       84690.00
##                 Income.3rd Qu.                    Income.Max.
##                       113800.00                      481300.00
##                 Deductions.Min.             Deductions.1st Qu.
##                            0.00                           0.00
##               Deductions.Median                Deductions.Mean
##                            0.00                          67.57
##             Deductions.3rd Qu.                Deductions.Max.
##                            0.00                        2904.00
##                     Hours.Min.                 Hours.1st Qu.
##                            1.00                          38.00
##                    Hours.Median                     Hours.Mean
##                           40.00                          40.07
##                   Hours.3rd Qu.                     Hours.Max.
##                           45.00                          99.00
##    HoursRISK_Adjustment.Min. HoursRISK_Adjustment.1st Qu.
##                        -1453.00                           0.00
##  HoursRISK_Adjustment.Median    HoursRISK_Adjustment.Mean
##                            0.00                        2021.00
## HoursRISK_Adjustment.3rd Qu.    HoursRISK_Adjustment.Max.
##                            0.00                      112200.00
```

2.b. Density plots for Numerical variables

RISK_Adjustment

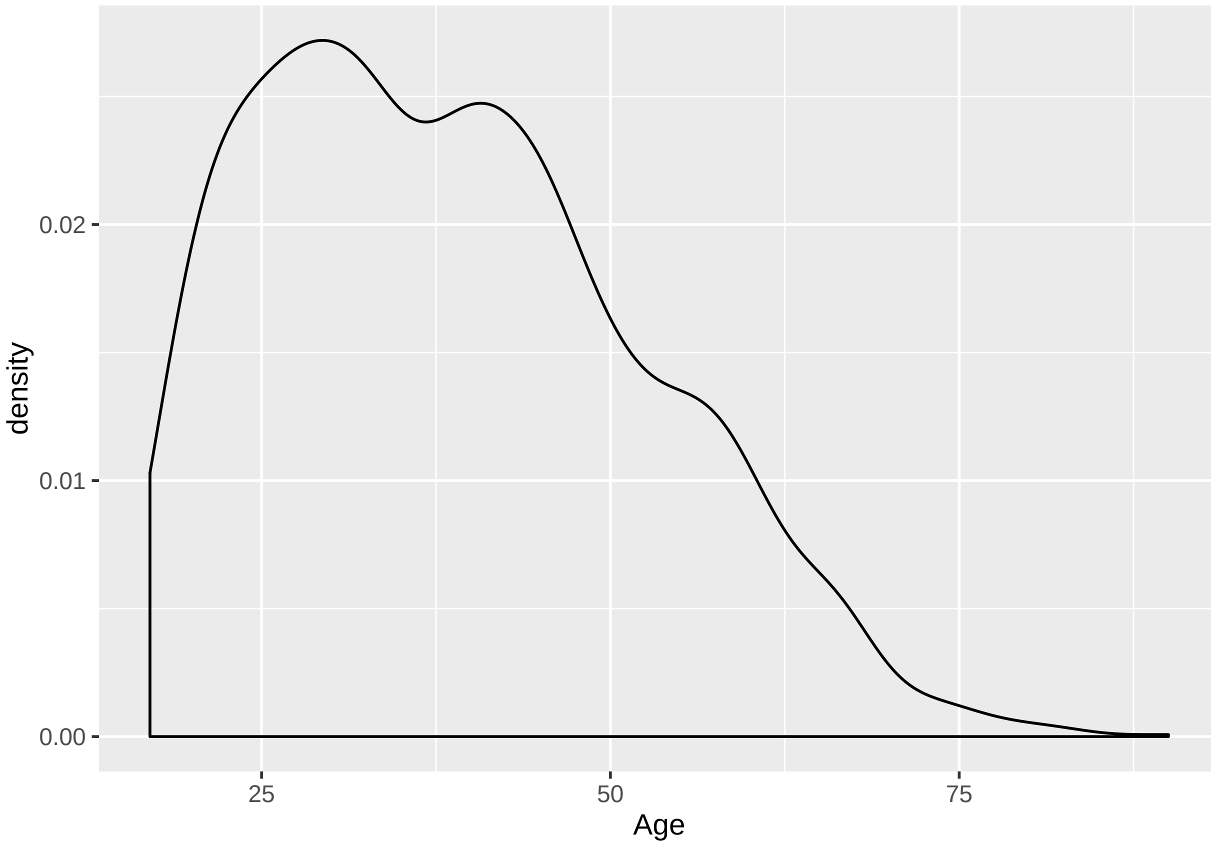
```r
suppressWarnings(library("ggplot2"))

ggplot(audit.clean, aes(x = RISK_Adjustment)) + geom_density()
```



We can see that there is a good amount of right skewness for RISK_Adjustment suggesting inconsistencies.

Age

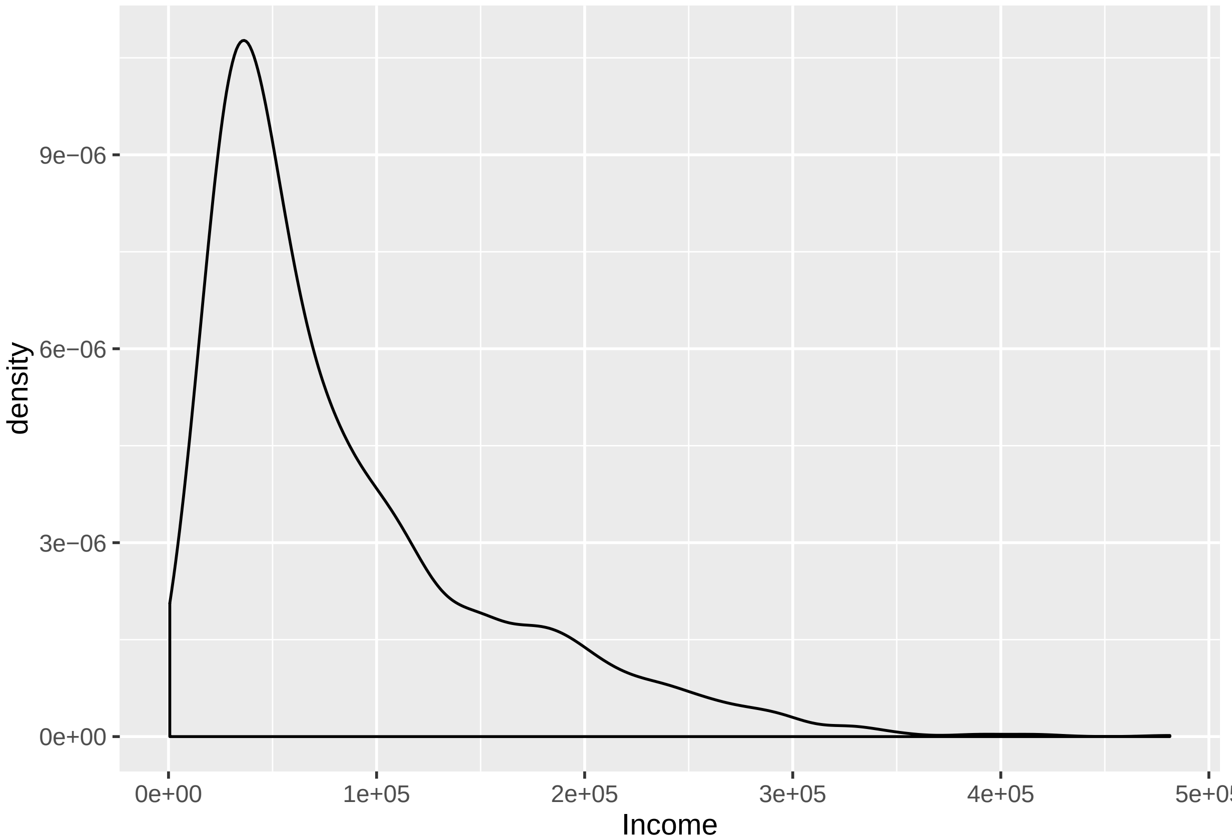
```
ggplot(audit.clean, aes(x = Age)) + geom_density()
```



There is a slight skewness to the right and also multiple bumps in the distribution.

Income

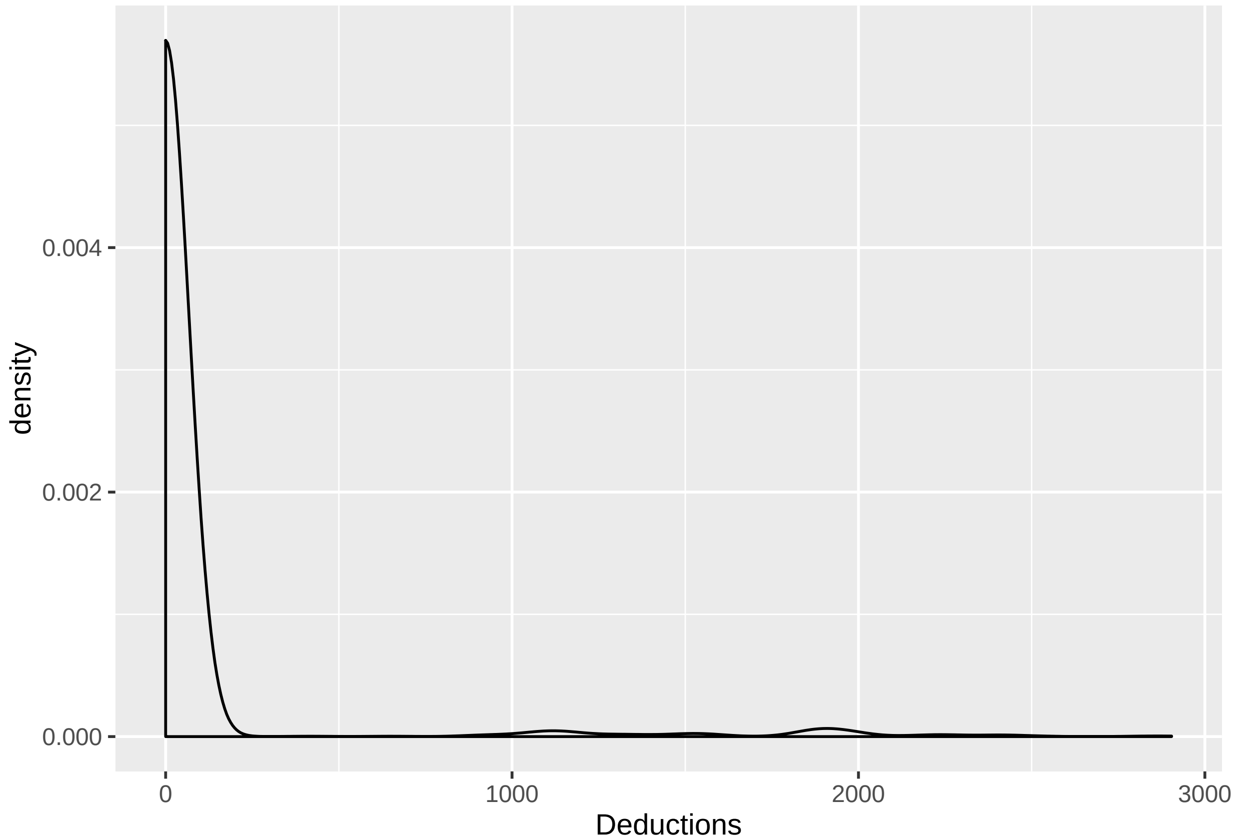
```
ggplot(audit.clean, aes(x = Income)) + geom_density()
```



Quite right skewed again for the Income variable.

Deductions

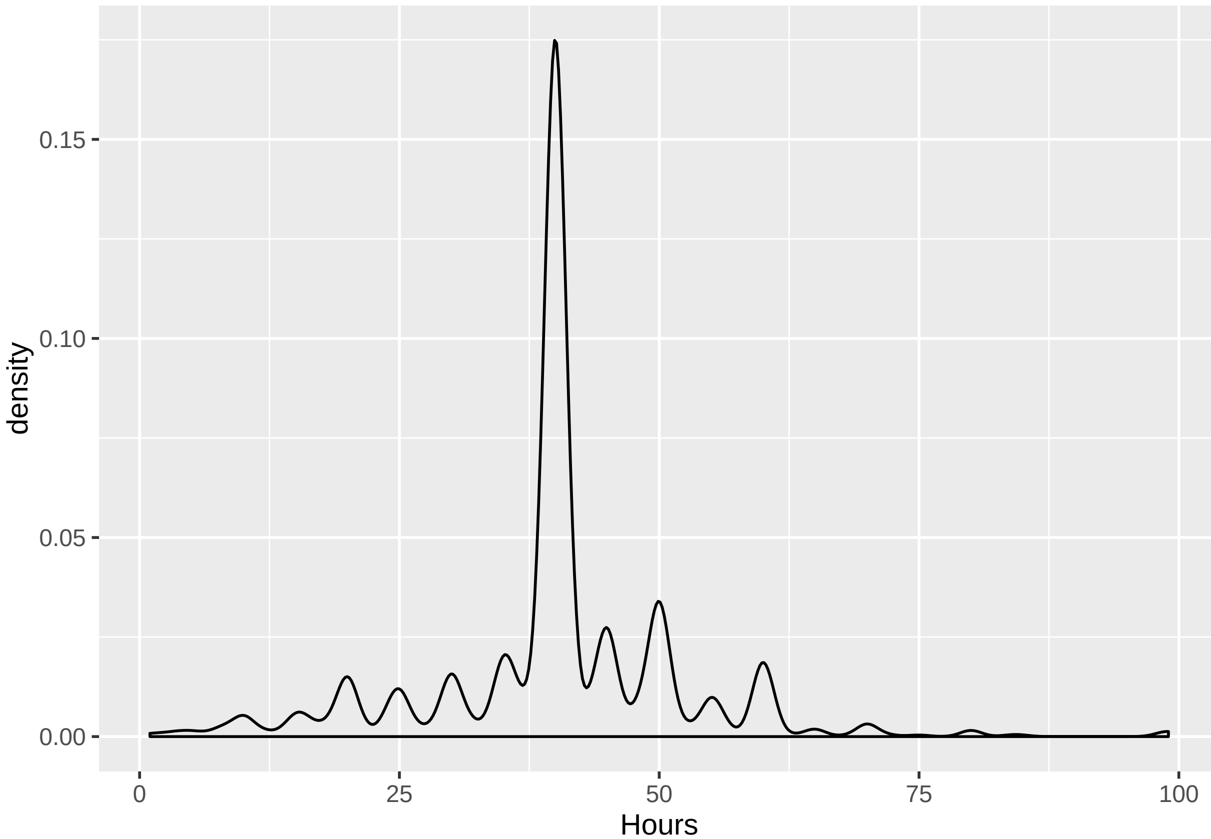
```
ggplot(audit.clean, aes(x = Deductions)) + geom_density()
```



Very skewed and there is a long tail to the right.

Hours

```
ggplot(audit.clean, aes(x = Hours)) + geom_density()
```



Hours seems much better distributed thatn the others. There are however multilple peaks in the distribution suggesting various classes of working hours.

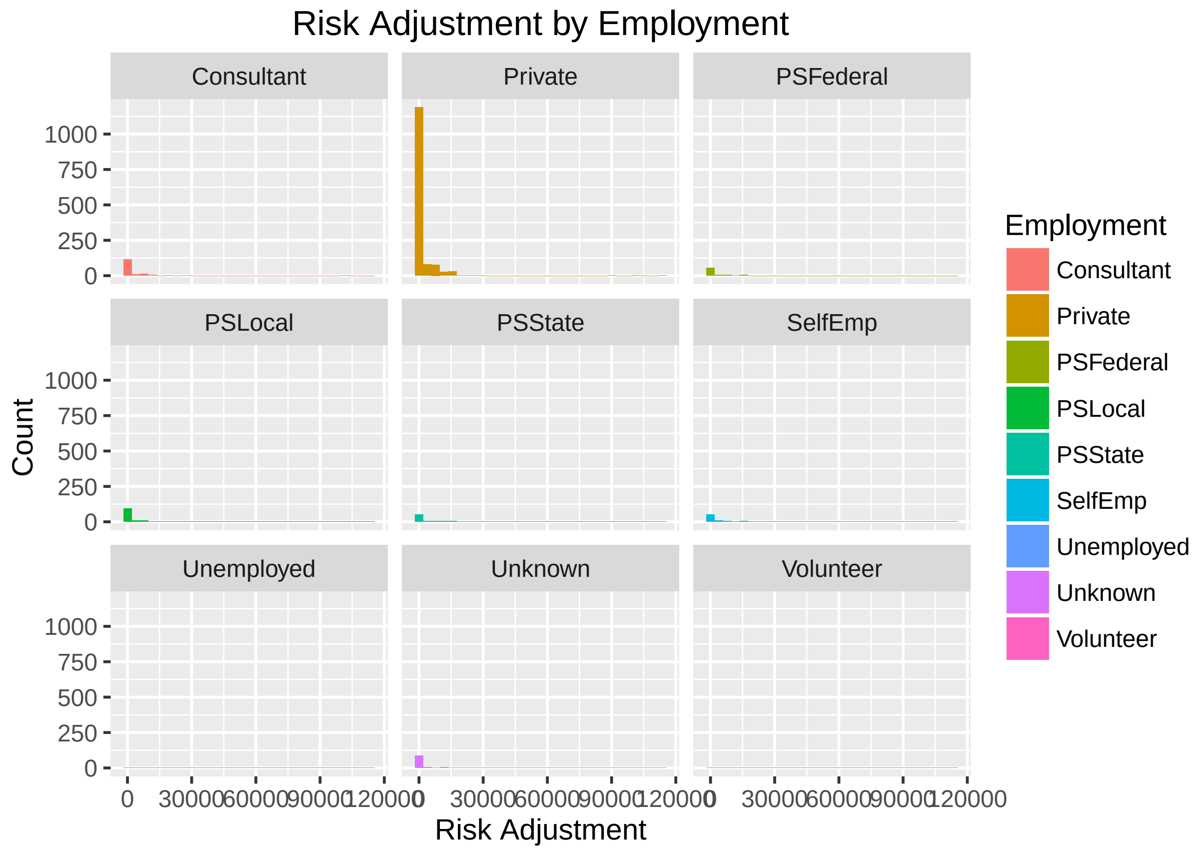
We see the data being skewed for most cases. There are appropriate transformations that we can apply to better normalize these points. We shal first look at the correlation and scatterplots and take a call.

2.c. Conditional Histogram Plot of Categorical Variables

i). RISK_Adjustment and Categorical Variables

RISK_Adjustment by Employment

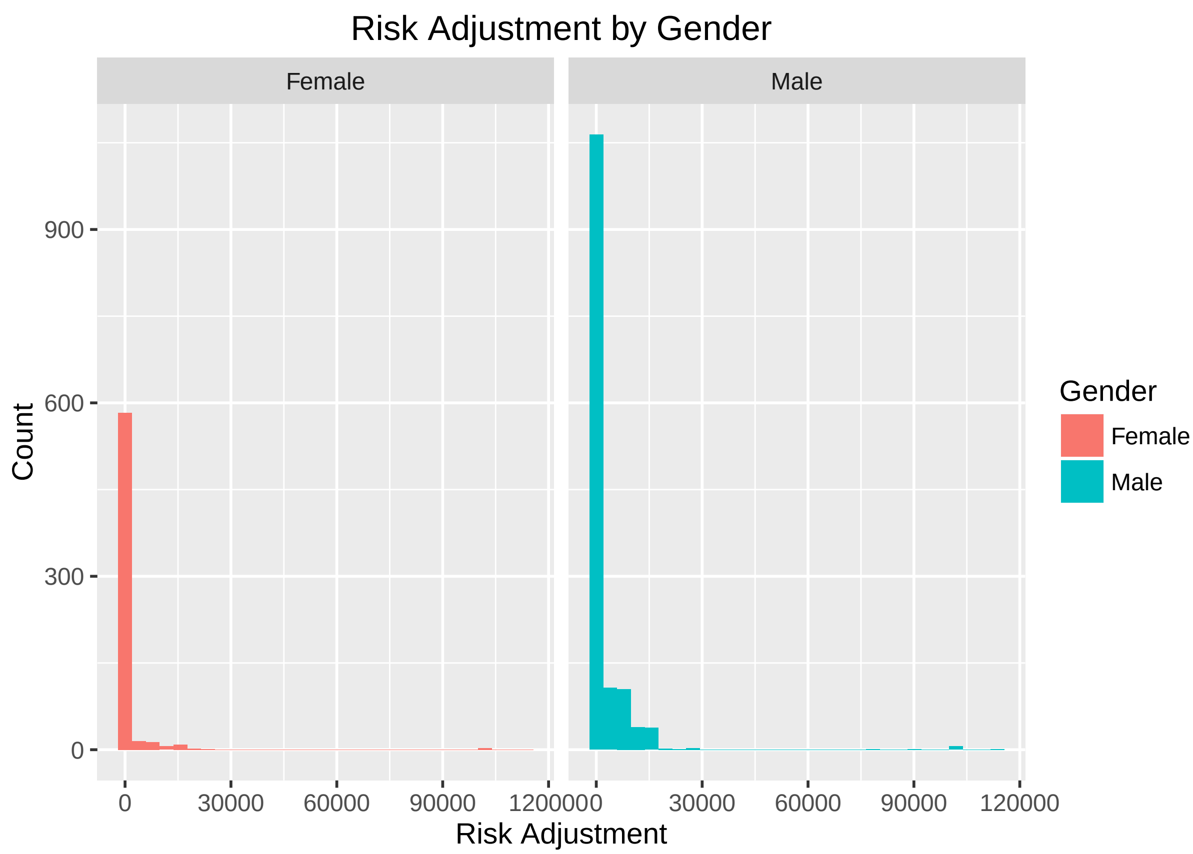
```
hist_Employment <- ggplot(audit.clean, aes(x=RISK_Adjustment, fill=Employment) )+
  geom_histogram(bins=30)+facet_wrap(~Employment)

hist_Employment+xlab("Risk Adjustment")+ylab("Count")+ggtitle("Risk Adjustment by Employment")
```



We can observe the skewness here. A vast majority if Employment types tend to have low Risk_Adjustment. Privates consists of a good number. But we can also observe that as we look to the right we see private employment have higher Risk_Adjustment too. Others also have but are very few in number.

RISK_Adjustment by Gender

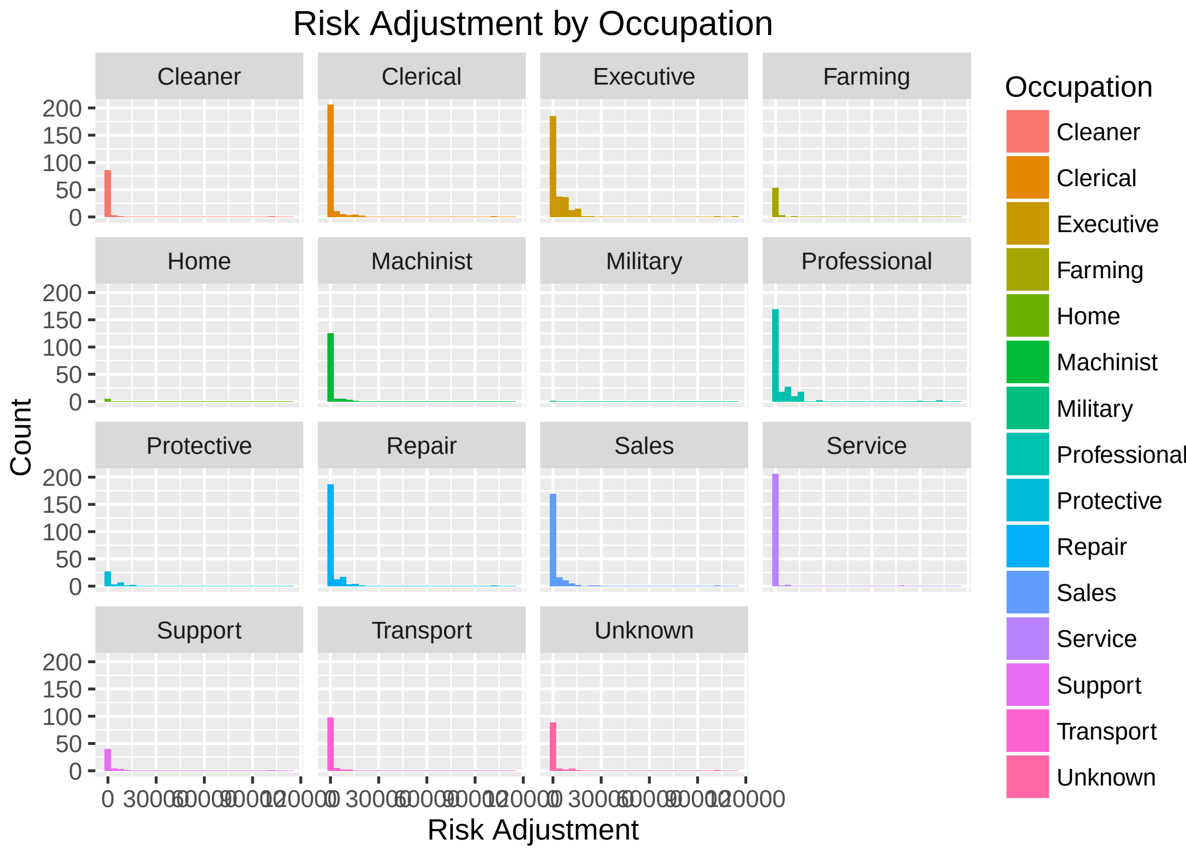
```
hist_Gender <- ggplot(audit.clean, aes(x=RISK_Adjustment, fill=Gender) )+
  geom_histogram(bins=30)+facet_wrap(~Gender)

hist_Gender+xlab("Risk Adjustment")+ylab("Count")+ggtitle("Risk Adjustment by Gender")
```



As can be seen above the Male tends to constitute more in low risk adjustment also for higher risk adjustment. This may be due to higher male working population.

RISK_Adjustment by Occupation

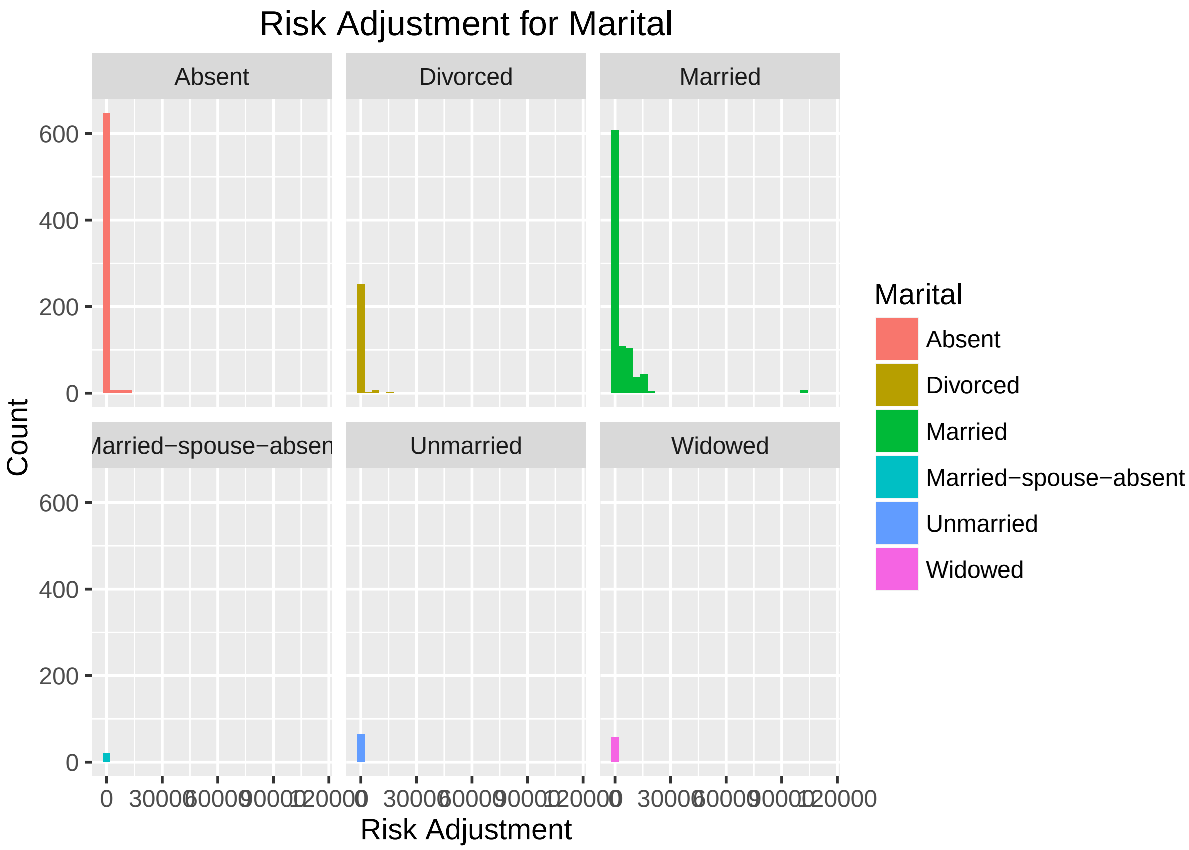
```
hist_Occupation<- ggplot(audit.clean, aes(x=RISK_Adjustment, fill=Occupation) )+
  geom_histogram(bins=30)+facet_wrap(~Occupation)

hist_Occupation+xlab("Risk Adjustment")+ylab("Count")+ggtitle("Risk Adjustment by Occupation")
```



Occupation shows considerable variation which is interesting. Clerical, Executive, Machinist, Professional, Repair, Sales have larger numbers in the high RISK_Adjustment values. A trend to observe here is that these people are more liekely to maintain tax returns.

RISK_Adjustment by Marital

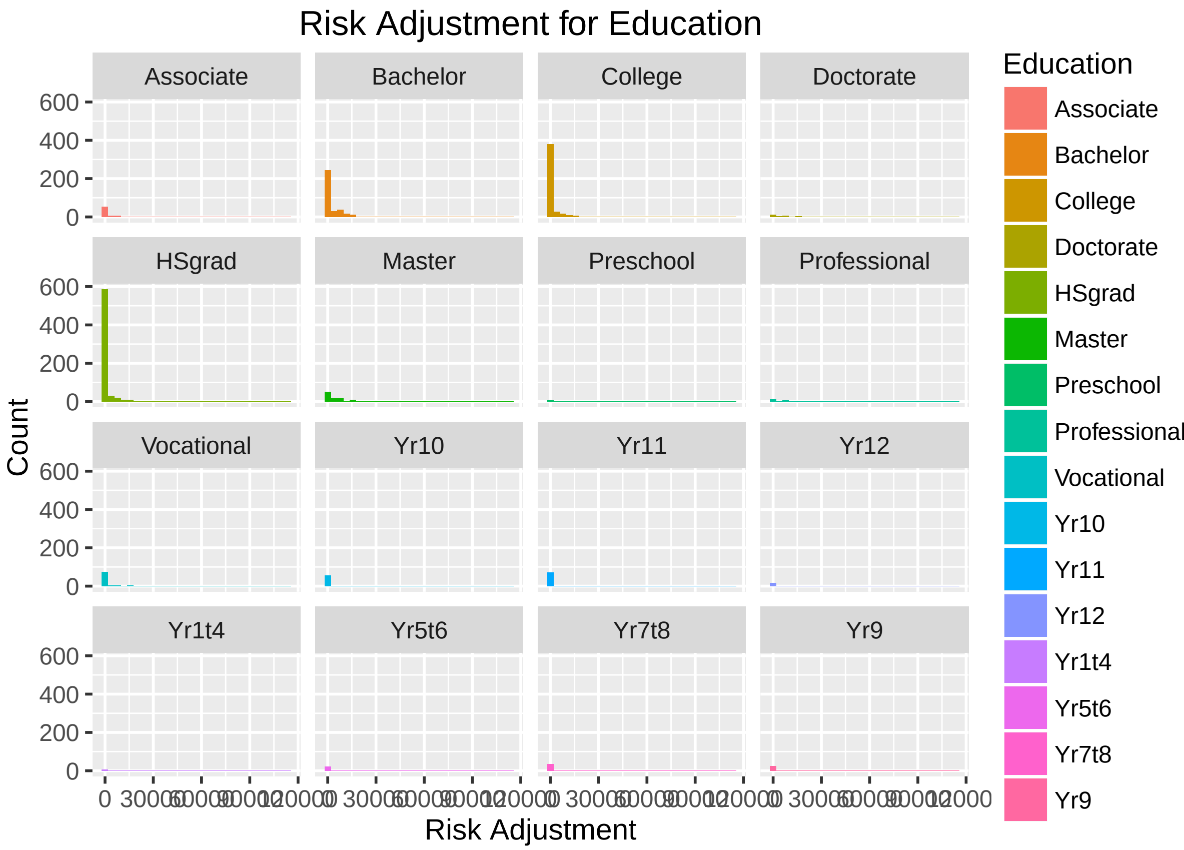
```
hist_Marital<- ggplot(audit.clean, aes(x=RISK_Adjustment, fill=Marital) )+
  geom_histogram(bins=30)+facet_wrap(~Marital)

hist_Marital+xlab("Risk Adjustment")+ylab("Count")+ggtitle("Risk Adjustment for Marital")
```



Here we see that Married shows a higher count for higher Risk adjusments. There are traces of other statuses but Married seems to be the prominent one.

RISK_Adjustment by Education

```
hist_Education<- ggplot(audit.clean, aes(x=RISK_Adjustment, fill=Education) )+
  geom_histogram(bins=30)+facet_wrap(~Education)

hist_Education+xlab("Risk Adjustment")+ylab("Count")+ggtitle("Risk Adjustment for Education")
```

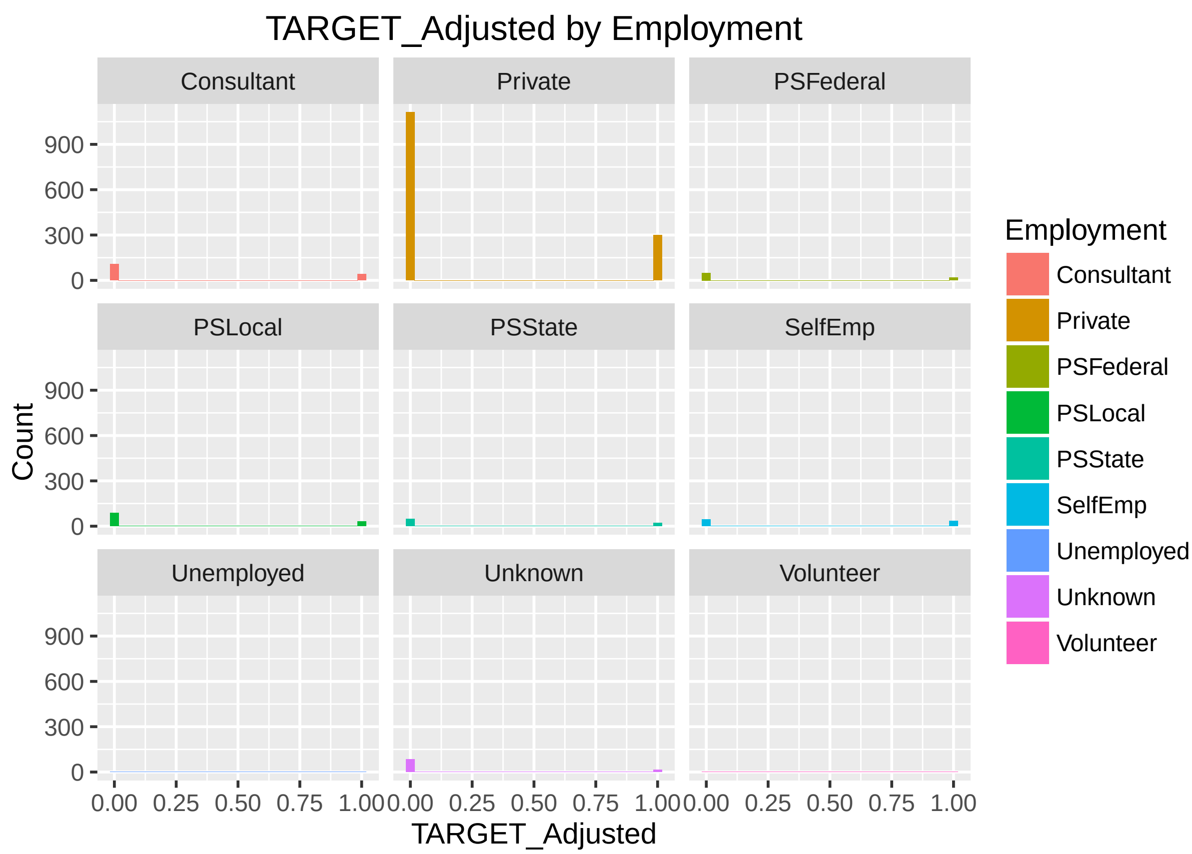


In the above distributions for Education, we can observe that Bachelor, College, HSgrad, Master tends to have higher Risk adjustment. This may attribute to the general financial instability of individuals at that range.

ii). TARGET_Adjusted and categorical variables

TARGET_Adjusted by Employment

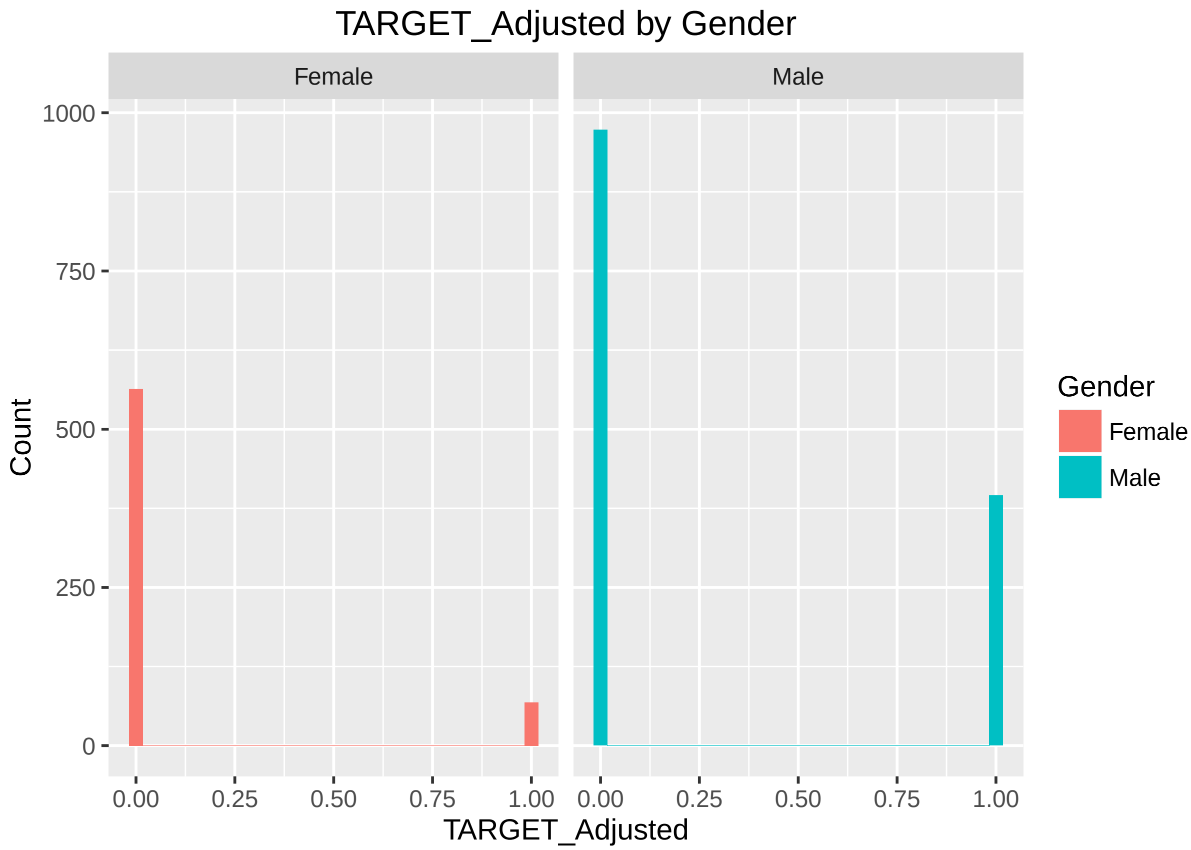
```
hist_Employment <- ggplot(audit.clean, aes(x=TARGET_Adjusted, fill=Employment) )+
  geom_histogram(bins=30)+facet_wrap(~Employment)

hist_Employment+xlab("TARGET_Adjusted")+ylab("Count")+ggtitle("TARGET_Adjusted by Employment")
```



We can see that Private has counts in both productive and non-productive audits. The count in non-productive audits being maximum. We see Consultant, PSLocal, PSState, PSFederal, SelfEmp all having counts in both sections but Private is the major category.

TARGET_Adjusted by Gender

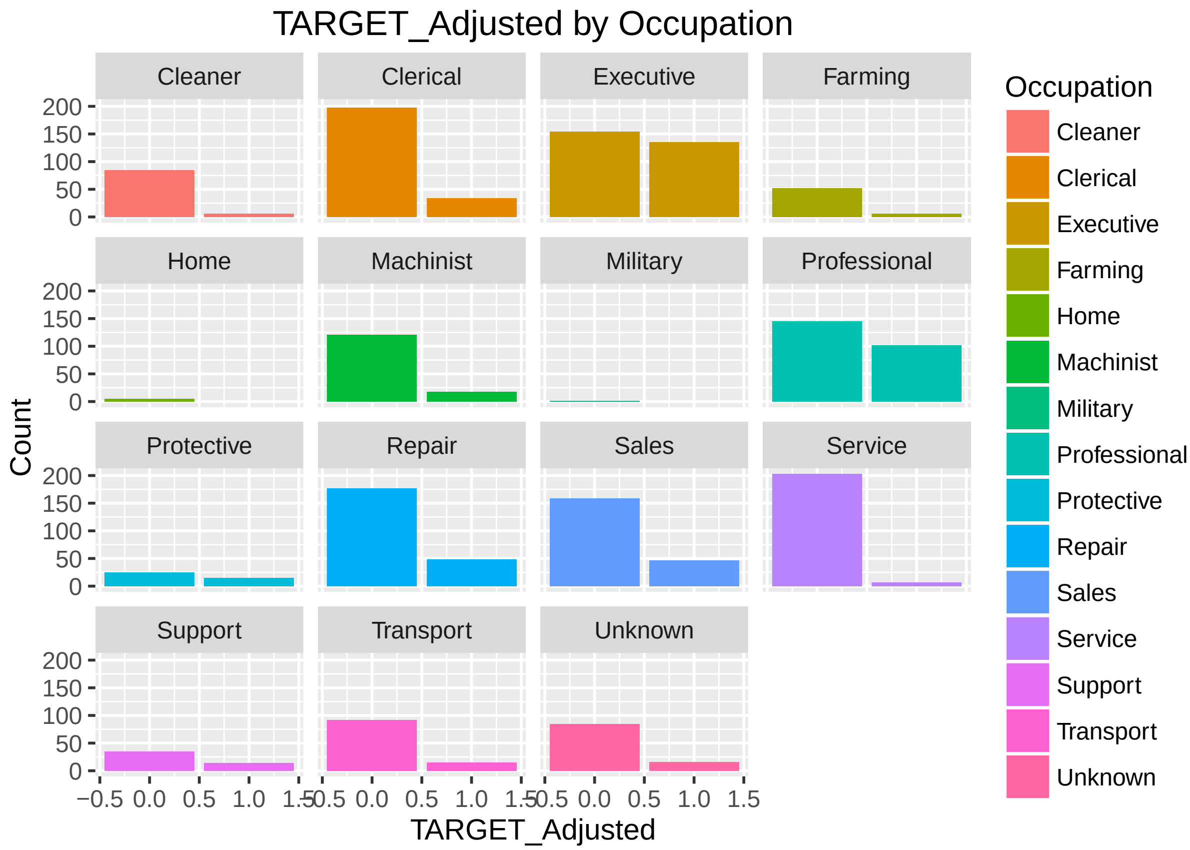
```
hist_Gender <- ggplot(audit.clean, aes(x=TARGET_Adjusted, fill=Gender) )+
  geom_histogram(bins=30)+facet_wrap(~Gender)

hist_Gender+xlab("TARGET_Adjusted")+ylab("Count")+ggtitle("TARGET_Adjusted by Gender")
```



Gender shows that Male are larger in number in both productive and non-productive audits. Also the ration of between the audits is higher for Male.

TARGET_Adjusted by Occupation

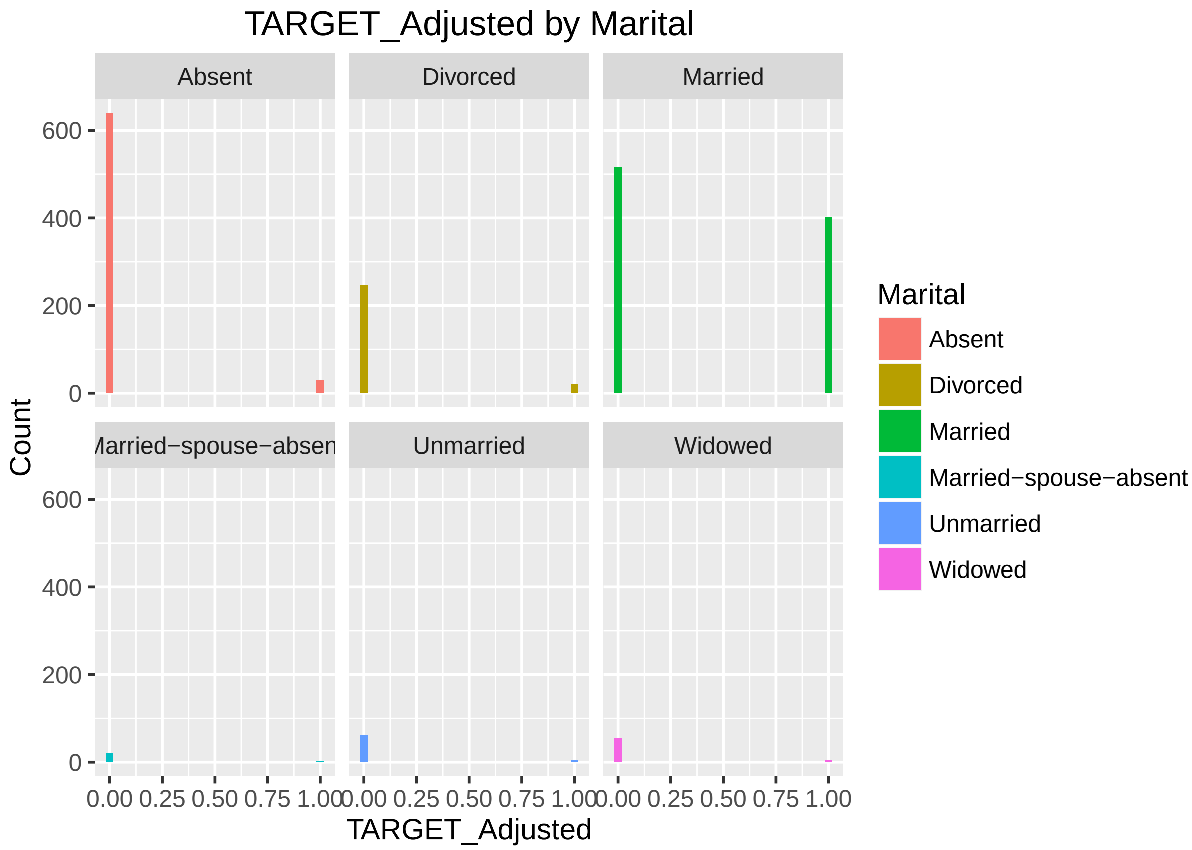
```
hist_Occupation <- ggplot(audit.clean, aes(x=TARGET_Adjusted, fill=Occupation) )+
  geom_bar()+facet_wrap(~Occupation)

hist_Occupation+xlab("TARGET_Adjusted")+ylab("Count")+ggtitle("TARGET_Adjusted by Occupation")
```



Clerical. Executive, Professional, Repair, Sales are the categories in Occupation that show high values for both productive and non-productive audits. Executive and Professional showing a greater value in productive audits than others.

TARGET_Adjusted by Marital

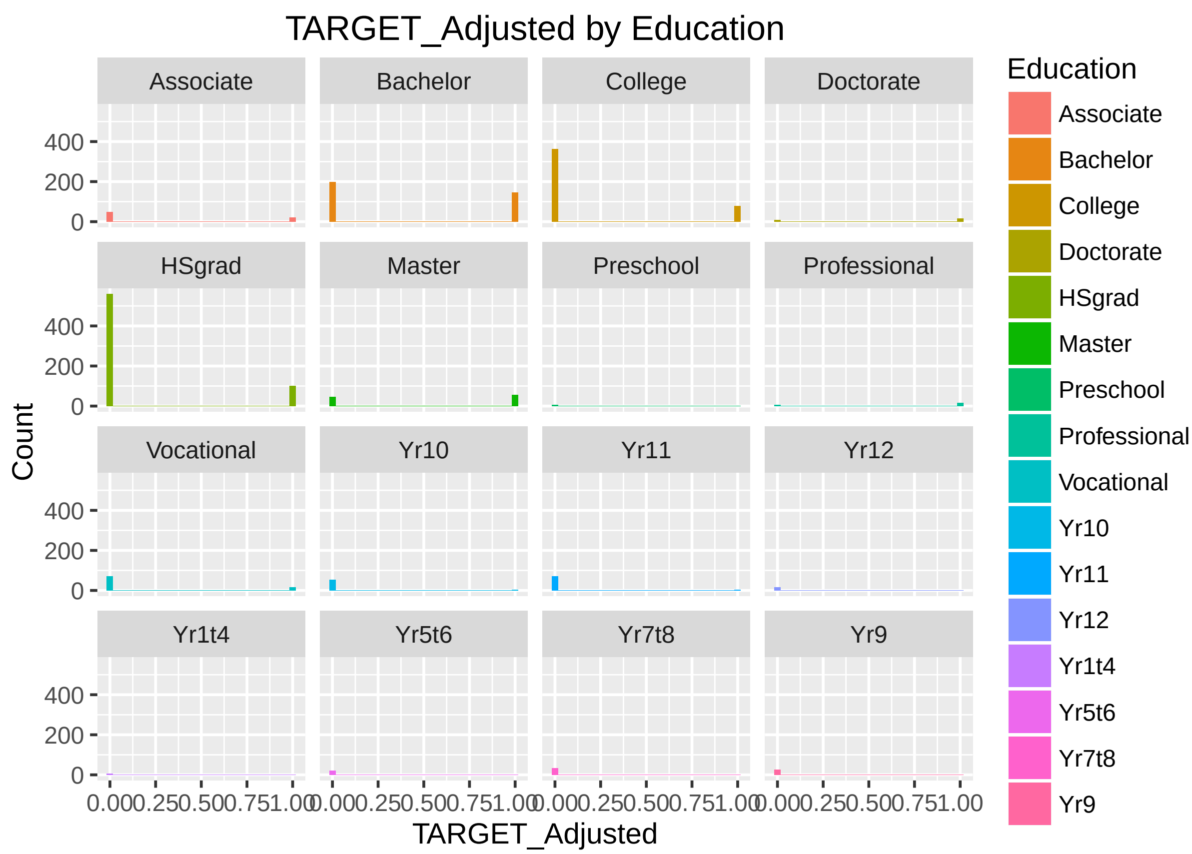
```
hist_Marital <- ggplot(audit.clean, aes(x=TARGET_Adjusted, fill=Marital) )+
  geom_histogram(bins=30)+facet_wrap(~Marital)

hist_Marital+xlab("TARGET_Adjusted")+ylab("Count")+ggtitle("TARGET_Adjusted by Marital")
```



We see that Married is also the category with considerable count in both the productive and non-productive audits.

TARGET_Adjusted by Education

```
hist_Education <- ggplot(audit.clean, aes(x=TARGET_Adjusted, fill=Education) )+
  geom_histogram(bins=30)+facet_wrap(~Education)

hist_Education+xlab("TARGET_Adjusted")+ylab("Count")+ggtitle("TARGET_Adjusted by Education")
```



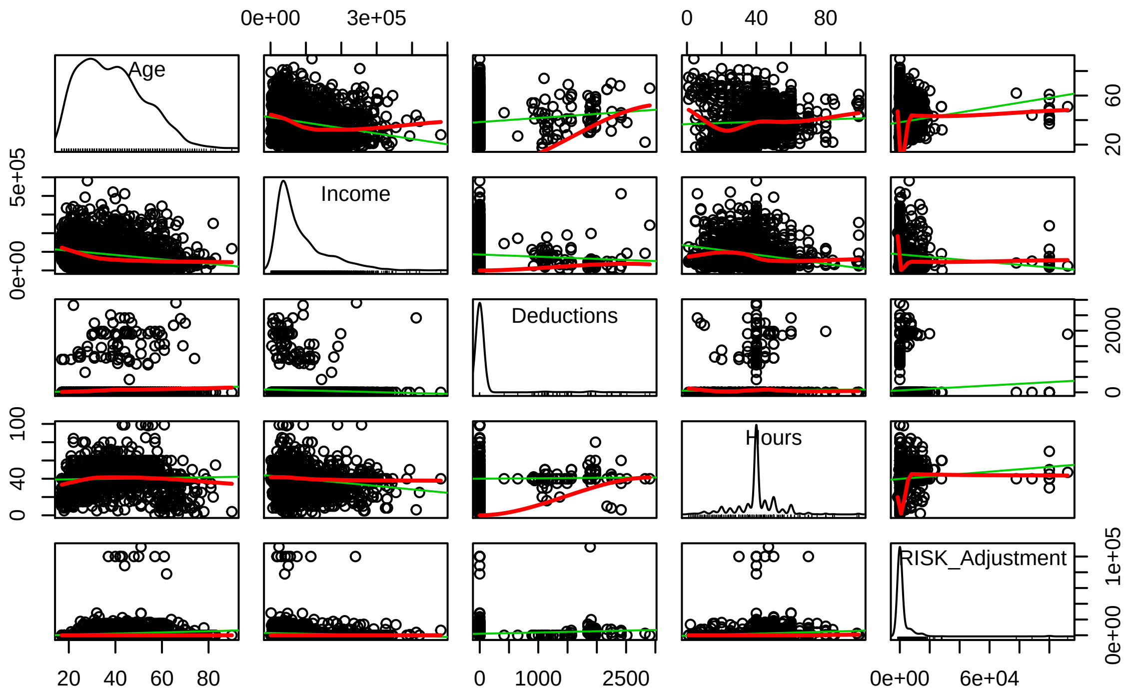
As in RISK_Adjustment, here also Bacheor, College, HSGrad and Master have good numbers for both the audits. There are traces of others too.

2.d. Correlation and Scatterplots of Numerical Variables with Response variables

Let us now observe the correlations of the numerical predictors with RISK_Adjustment and also the scatterplots between them

```
##                        Age      Income  Deductions       Hours
## Age              1.00000000 -0.22686777  0.08399899  0.04236487
## Income          -0.22686777  1.00000000 -0.05734147 -0.21269065
## Deductions       0.08399899 -0.05734147  1.00000000  0.01365124
## Hours            0.04236487 -0.21269065  0.01365124  1.00000000
## RISK_Adjustment  0.12274079 -0.08339021  0.06559720  0.09060735
##                 RISK_Adjustment
## Age                  0.12274079
## Income              -0.08339021
## Deductions           0.06559720
## Hours                0.09060735
## RISK_Adjustment      1.00000000
```

Scatter Plot Matrix



As we can see from the plots and the correlation matrix, There is very less correlation between the numerical predictors and RISK_Adjustment. We see a slight positive correlation of .12 between Age and RISK_Adjustment. A negative coorelation of -0.08 with Income and very minor correlation of RISK_Adjustment with Deductions 0.06 and Hours 0.09.

3. Applying logistic regression analysis to predict TARGET_Adjusted.

Let us now try to predict the TARGET_Adjusted boolean variable using logistic regression. Please note here we take 1 as Positive and 0 as Negative. Appropriate code adjustments are made to accomodate this in the sections below.

3.a. Predicting the best model via 10-fold CV, Accuracy, Precision, Recall, Lift Chart, ROC chart, AUC

Let us look into a baseline model. Note we do not consider RISK_Adjustment as a predictor here

```
basemodel =  glm(TARGET_Adjusted ~ Age+Employment+Education+Marital+Occupation+
                   Income+Gender+Deductions+Hours,data = audit.clean,family="binomial")
summary(basemodel)
```

```
##
## Call:
## glm(formula = TARGET_Adjusted ~ Age + Employment + Education +
##     Marital + Occupation + Income + Gender + Deductions + Hours,
```

```
##     family = "binomial", data = audit.clean)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.58462  -0.53212  -0.21898  -0.00025   2.88847
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.591e+00  8.329e-01  -7.913 2.51e-15 ***
## Age                           2.987e-02  6.515e-03   4.585 4.54e-06 ***
## EmploymentPrivate             3.387e-01  2.547e-01   1.330  0.18352
## EmploymentPSFederal           2.900e-01  4.220e-01   0.687  0.49204
## EmploymentPSLocal             9.842e-02  3.891e-01   0.253  0.80029
## EmploymentPSState             3.005e-01  4.336e-01   0.693  0.48818
## EmploymentSelfEmp             1.388e-01  3.684e-01   0.377  0.70643
## EmploymentUnemployed         -1.062e+01  3.956e+03  -0.003  0.99786
## EmploymentUnknown             7.210e-01  6.346e-01   1.136  0.25589
## EmploymentVolunteer          -1.758e+01  3.956e+03  -0.004  0.99645
## EducationBachelor             9.887e-02  3.654e-01   0.271  0.78672
## EducationCollege             -8.552e-01  3.678e-01  -2.325  0.02006 *
## EducationDoctorate            1.011e+00  6.218e-01   1.627  0.10380
## EducationHSgrad              -1.155e+00  3.618e-01  -3.192  0.00141 **
## EducationMaster               4.820e-01  4.449e-01   1.083  0.27867
## EducationPreschool           -1.561e+01  1.410e+03  -0.011  0.99117
## EducationProfessional         1.733e+00  6.560e-01   2.641  0.00827 **
## EducationVocational          -9.833e-01  4.805e-01  -2.046  0.04071 *
## EducationYr10                -1.546e+00  6.585e-01  -2.348  0.01885 *
## EducationYr11                -1.601e+00  7.297e-01  -2.194  0.02820 *
## EducationYr12                -1.739e+00  1.190e+00  -1.461  0.14403
## EducationYr1t4               -1.706e+01  1.484e+03  -0.012  0.99082
## EducationYr5t6               -2.224e+00  9.133e-01  -2.435  0.01488 *
## EducationYr7t8               -1.666e+01  5.923e+02  -0.028  0.97756
## EducationYr9                 -2.930e+00  1.131e+00  -2.590  0.00960 **
## MaritalDivorced              -6.347e-02  3.367e-01  -0.189  0.85045
## MaritalMarried                2.681e+00  2.383e-01  11.254  < 2e-16 ***
## MaritalMarried-spouse-absent  3.562e-01  8.146e-01   0.437  0.66190
## MaritalUnmarried              5.921e-01  5.378e-01   1.101  0.27091
## MaritalWidowed               -1.340e-01  6.560e-01  -0.204  0.83819
## OccupationClerical            1.181e+00  5.304e-01   2.226  0.02604 *
## OccupationExecutive           1.587e+00  4.964e-01   3.197  0.00139 **
## OccupationFarming             2.495e-02  6.929e-01   0.036  0.97128
## OccupationHome               -1.250e+01  1.727e+03  -0.007  0.99422
## OccupationMachinist           4.818e-01  5.427e-01   0.888  0.37470
## OccupationMilitary           -1.293e+01  3.956e+03  -0.003  0.99739
## OccupationProfessional        1.233e+00  5.188e-01   2.377  0.01746 *
## OccupationProtective          1.866e+00  6.567e-01   2.841  0.00449 **
## OccupationRepair              6.786e-01  5.010e-01   1.355  0.17557
## OccupationSales               9.625e-01  5.145e-01   1.871  0.06141 .
## OccupationService            -3.747e-01  6.297e-01  -0.595  0.55187
## OccupationSupport             1.279e+00  6.104e-01   2.095  0.03616 *
## OccupationTransport           2.472e-01  5.562e-01   0.445  0.65668
## OccupationUnknown                    NA         NA      NA       NA
## Income                        2.405e-06  1.437e-06   1.674  0.09415 .
## GenderMale                    1.911e-01  2.469e-01   0.774  0.43905
```

```
## Deductions                        1.053e-03  1.950e-04   5.399 6.71e-08 ***
## Hours                             3.465e-02  6.397e-03   5.416 6.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2164.3  on 1999  degrees of freedom
## Residual deviance: 1329.9  on 1953  degrees of freedom
## AIC: 1423.9
##
## Number of Fisher Scoring iterations: 16
```

We can see some significnce of Age, MaritalMarried, Deductions, Hours. Let us try and see what logits we get for them.

```
#Age
exp(basemodel$coef[2])
```

```
##     Age
## 1.03032
```

```
#MaritalMarried
exp(basemodel$coef[27])
```

```
## MaritalMarried
##       14.60591
```

```
#Deductions
exp(basemodel$coef[47])
```

```
## Deductions
##   1.001053
```

```
#Hours
exp(basemodel$coef[48])
```

```
##    Hours
## 1.035258
```

Conditioning on other variables we can see that all have an effect on TARGET_Adjusted to be postive. MaritalMarried has the highest effect among them.

Let us now create a model by splitting the data into training and test sets but without cross-validation. We create a model matrix first to account for undersampling of some data in the Employment and Occupation variables.

```
Xdel = model.matrix(TARGET_Adjusted~Age+Employment+Education+Marital
                    +Occupation+Income+Gender+Deductions+Hours,data=audit.clean)[,-1]

n.total=length(audit.clean$TARGET_Adjusted)
n.total
```

```
## [1] 2000
```

```
n.train=floor(n.total*(0.6))
n.train
```

```
## [1] 1200
```

```
n.test=n.total-n.train
n.test
```

```
## [1] 800
```

```
train=sample(1:n.total,n.train)
xtrain = Xdel[train,]
xtest = Xdel[-train,]

ytrain = audit.clean$TARGET_Adjusted[train]
ytest = audit.clean$TARGET_Adjusted[-train]

m1 = glm(TARGET_Adjusted~.,family=binomial,data=data.frame(TARGET_Adjusted=ytrain,xtrain))
summary(m1)
```

```
##
## Call:
## glm(formula = TARGET_Adjusted ~ ., family = binomial, data = data.frame(TARGET_Adjusted = ytrain,
##     xtrain))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.40135  -0.49288  -0.18809  -0.00007   3.08113
##
## Coefficients: (2 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -7.333e+00  1.100e+00  -6.668 2.59e-11 ***
## Age                          3.560e-02  8.440e-03   4.218 2.47e-05 ***
## EmploymentPrivate            6.125e-01  3.434e-01   1.783  0.07451 .
## EmploymentPSFederal          7.134e-01  5.302e-01   1.346  0.17846
## EmploymentPSLocal            3.387e-01  5.410e-01   0.626  0.53128
## EmploymentPSState            7.700e-01  5.665e-01   1.359  0.17408
## EmploymentSelfEmp            7.252e-01  4.745e-01   1.528  0.12647
## EmploymentUnemployed                NA         NA      NA       NA
## EmploymentUnknown            1.523e+00  8.112e-01   1.878  0.06041 .
## EmploymentVolunteer         -1.847e+01  3.956e+03  -0.005  0.99628
## EducationBachelor            3.074e-02  4.651e-01   0.066  0.94730
## EducationCollege            -9.642e-01  4.672e-01  -2.064  0.03905 *
## EducationDoctorate           8.365e-01  8.815e-01   0.949  0.34266
## EducationHSgrad             -1.390e+00  4.629e-01  -3.004  0.00267 **
## EducationMaster             -6.563e-02  5.644e-01  -0.116  0.90742
## EducationPreschool          -1.546e+01  1.757e+03  -0.009  0.99298
## EducationProfessional        8.731e-01  8.498e-01   1.027  0.30421
## EducationVocational         -1.169e+00  6.279e-01  -1.862  0.06262 .
## EducationYr10               -1.938e+00  7.969e-01  -2.432  0.01500 *
```

```
## EducationYr11                 -2.403e+00  1.186e+00  -2.026  0.04282 *
## EducationYr12                 -2.925e-01  1.530e+00  -0.191  0.84838
## EducationYr1t4                -1.766e+01  1.912e+03  -0.009  0.99263
## EducationYr5t6                -1.988e+00  1.002e+00  -1.984  0.04727 *
## EducationYr7t8                -1.704e+01  7.688e+02  -0.022  0.98232
## EducationYr9                  -2.312e+00  1.201e+00  -1.925  0.05421 .
## MaritalDivorced                4.574e-01  4.484e-01   1.020  0.30769
## MaritalMarried                 3.047e+00  3.375e-01   9.027  < 2e-16 ***
## MaritalMarried.spouse.absent   1.163e+00  9.293e-01   1.251  0.21080
## MaritalUnmarried               1.672e+00  6.105e-01   2.739  0.00617 **
## MaritalWidowed                 1.153e-01  7.996e-01   0.144  0.88535
## OccupationClerical             1.246e+00  6.696e-01   1.861  0.06281 .
## OccupationExecutive            1.942e+00  6.289e-01   3.089  0.00201 **
## OccupationFarming              9.937e-01  8.112e-01   1.225  0.22062
## OccupationHome                -1.199e+01  3.956e+03  -0.003  0.99758
## OccupationMachinist           -5.872e-02  7.436e-01  -0.079  0.93706
## OccupationMilitary            -1.262e+01  3.956e+03  -0.003  0.99745
## OccupationProfessional         1.447e+00  6.605e-01   2.191  0.02843 *
## OccupationProtective           1.823e+00  8.540e-01   2.134  0.03283 *
## OccupationRepair               9.659e-01  6.330e-01   1.526  0.12706
## OccupationSales                9.152e-01  6.568e-01   1.393  0.16350
## OccupationService             -2.138e-01  7.763e-01  -0.275  0.78306
## OccupationSupport              1.704e+00  7.893e-01   2.159  0.03085 *
## OccupationTransport            8.408e-01  6.763e-01   1.243  0.21378
## OccupationUnknown                     NA          NA      NA       NA
## Income                         2.450e-06  1.900e-06   1.289  0.19726
## GenderMale                     3.580e-01  3.320e-01   1.078  0.28097
## Deductions                     7.398e-04  2.743e-04   2.697  0.00699 **
## Hours                          3.011e-02  8.173e-03   3.683  0.00023 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1334.00  on 1199  degrees of freedom
## Residual deviance:  779.93  on 1154  degrees of freedom
## AIC: 871.93
##
## Number of Fisher Scoring iterations: 16
```

```r
ptest = suppressWarnings(predict(m1,newdata=data.frame(xtest),type="response"))

data.frame(ytest,ptest)[1:10,]
```

```
##    ytest      ptest
## 2      0 0.03933308
## 3      0 0.03466078
## 4      1 0.73238851
## 8      0 0.25544142
## 11     0 0.01863075
## 15     1 0.83628895
## 18     0 0.08387798
## 19     0 0.58590387
## 22     0 0.28144782
```

```
## 27     0 0.03036353
```

```
btest=floor(ptest+0.5)

conf.matrix = table(ytest,btest)
conf.matrix
```

```
##      btest
## ytest   0   1
##     0 564  66
##     1  74  96
```

```
error=(conf.matrix[1,2]+conf.matrix[2,1])/n.test
error
```

```
## [1] 0.175
```

```
acc = 1- error
acc
```

```
## [1] 0.825
```

We can observe that the accuracy ranges from 80 - 85%. This range exists because of the fact that we use a random shuffling of data. Below I give the models based on accuracy, precision and recall.

Before the models. I would like to highlight the steps and approaches that I took for the model for better clarity.

1. I removed ID and RISK_Adjustment columns in the first step.
2. Created a model matrix and chose which variables to incude (variable selection) here rather than the model below.
3. To find the best model I did not use random shuffling at first. If we do that we get varying results, so to get the best model I commented the line (df = df[sample(nrow(df)),]). Once I found the best model then I computed further scores with the shuffling of data.
4. The major difference is that I did not go for the regular 10-fold cross validation on the whole data. Instead, I first split the data into Training 80% and testing 20% and then performed 10-fold CV on the 80% Training set and finally predicted the 20% held out test data. So, please observe that my confusion matrix, lift chart and roc chart are on the 20% (400 data points). I chose this approach due to a personal preference and also after reading through the piaza post on this discussion.
5. I also reversed the TP to be on 1 and TN to be on 0 respectively for TARGET_Adjusted in the confusion matirix while predicting precision and recall.

```
ten.fold.cv <- function(data, query) {
audit.del=audit.clean[,c(-1, -11)]
#audit.del[1:3,]
audit.train<-audit.del
#audit.train[1:3,]
Xdel = model.matrix(query,data=audit.train)[,-1]
#Xdel[1:3,]
```

```
n.total=length(audit.train$TARGET_Adjusted)
#n.total

n.train=floor(n.total*(0.8))
#n.train

n.test=n.total-n.train
#n.test

xtrain = Xdel[1:n.train,]
xtest = Xdel[(n.train+1):n.total,]

ytrain = audit.train$TARGET_Adjusted[1:n.train]
ytest = audit.train$TARGET_Adjusted[(n.train+1):n.total]

#Create 10 equally size folds
folds <- cut(seq(1,nrow(xtrain)),breaks=10,labels=FALSE)

df = data.frame(TARGET_Adjusted=ytrain,xtrain)
#df = df[sample(nrow(df)),]

#Perform 10 fold cross validation on the 80% training data
for(i in 1:10){
  #Segement data by fold using the which() function
  testIndexes <- which(folds==i,arr.ind=TRUE)

  df.cv = df[testIndexes,]

  model.train = glm(formula=TARGET_Adjusted~.,family=binomial,data=df.cv)
}

#Use the model to predict the 20% test data
ptest = suppressWarnings(predict(model.train,newdata=data.frame(xtest),type="response"))
btest=floor(ptest+0.5)
conf.matrix = table(ytest,btest)
return (conf.matrix)
}
```

Below is the first model with 10-fold cross validation on all the predictors.

```
formulaA = TARGET_Adjusted~Age+Employment+Education+Marital+Occupation+Income+Hours+Gender+Deductions
suppressWarnings(conf.matrix <- ten.fold.cv(audit.clean, formulaA ))
conf.matrix
```

```
##      btest
## ytest   0   1
##     0 262  52
##     1  31  55
```

```
#Accuracy
error=(conf.matrix[1,2]+conf.matrix[2,1])/n.test
error
```

```
## [1] 0.10375
```

```r
acc = 1 - error
acc
```

```
## [1] 0.89625
```

```r
#treating the given binary values "1/0" as "positive/negative"
#so switching TN with TP and FN with FP in conf.matrix
TN = conf.matrix[1,1]
TP = conf.matrix[2,2]
FN = conf.matrix[2,1]
FP = conf.matrix[1,2]

Precision = TP/(TP+FP)
Precision
```

```
## [1] 0.5140187
```

```r
Recall = TP/(TP+FN)
Recall
```

```
## [1] 0.6395349
```

```r
F1 = 2*(Precision*Recall)/(Precision+Recall)
F1
```

```
## [1] 0.5699482
```

As we can see above using all the predictors we get a good accuracy rate of 89.625%. However if we go down and see the precision and recall and F1 scores aren't that great. Let us try to reduce the predictors and see.

```r
formulaB = TARGET_Adjusted~Age+Employment+Education+Marital+Occupation+Income+Hours
suppressWarnings(conf.matrix <- ten.fold.cv(audit.clean, formulaB ))
conf.matrix
```

```
##      btest
## ytest   0   1
##     0 270  44
##     1  34  52
```

```r
#Accuracy
error=(conf.matrix[1,2]+conf.matrix[2,1])/n.test
error
```

```
## [1] 0.0975
```

```r
acc = 1 - error
acc
```

```
## [1] 0.9025
```

```r
#treating the given binary values "1/0" as "positive/negative"
#so switching TN with TP and FN with FP in conf.matrix
TN = conf.matrix[1,1]
TP = conf.matrix[2,2]
FN = conf.matrix[2,1]
FP = conf.matrix[1,2]

Precision = TP/(TP+FP)
Precision
```

```
## [1] 0.5416667
```

```r
Recall = TP/(TP+FN)
Recall
```

```
## [1] 0.6046512
```

```r
F1 = 2*(Precision*Recall)/(Precision+Recall)
F1
```

```
## [1] 0.5714286
```

As seen above we do have an increase in accuracy to 90.25%, but we can also see that the precision has increased and recall has reduced but overall F1 score is better.This seems to be a good model.

Let us try to remove Occupation and Hours as intuitively we can think of them of having a great impact as working hours should not matter as long as income is present as a separate predictor.

```r
formulaC = TARGET_Adjusted~Age+Employment+Education+Marital+Income
suppressWarnings(conf.matrix <- ten.fold.cv(audit.clean, formulaC ))
conf.matrix
```

```
##      btest
## ytest   0   1
##     0 260  54
##     1  35  51
```

```r
#Accuracy
error=(conf.matrix[1,2]+conf.matrix[2,1])/n.test
error
```

```
## [1] 0.11125
```

```r
acc = 1 - error
acc
```

```
## [1] 0.88875
```

```r
#treating the given binary values "1/0" as "positive/negative"
#so switching TN with TP and FN with FP in conf.matrix
TN = conf.matrix[1,1]
TP = conf.matrix[2,2]
FN = conf.matrix[2,1]
FP = conf.matrix[1,2]

Precision = TP/(TP+FP)
Precision
```

```
## [1] 0.4857143
```

```r
Recall = TP/(TP+FN)
Recall
```

```
## [1] 0.5930233
```

```r
F1 = 2*(Precision*Recall)/(Precision+Recall)
F1
```

```
## [1] 0.5340314
```

We have seen that among all the combinations we have the second model has good accuracy as well as good F1 scores. The formula being:

formulaB = TARGET_Adjusted~Age+Employment+Education+Marital+Occupation+Income+Hours

Having decided the best model let us now further solidify our findings by including random suffling on cross-validation and then use it to construct the Lift Chart, ROC chart and find the AUC.

Final Model:

```r
audit.train<-audit.clean[,c(-1, -11)]

Xdel = model.matrix(TARGET_Adjusted~Age+Employment+
                      Education+Marital+Occupation+Income+Hours,data=audit.train)[,-1]

n.total=length(audit.train$TARGET_Adjusted)

n.train=floor(n.total*(0.8))

n.test=n.total-n.train

xtrain = Xdel[1:n.train,]
xtest = Xdel[(n.train+1):n.total,]

ytrain = audit.train$TARGET_Adjusted[1:n.train]
ytest = audit.train$TARGET_Adjusted[(n.train+1):n.total]

summary(ytest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.215   0.000   1.000
```

```r
#Create 10 equally size folds
folds <- cut(seq(1,nrow(xtrain)),breaks=10,labels=FALSE)

#Perform 10 fold cross validation

df = data.frame(TARGET_Adjusted=ytrain,xtrain)
df = df[sample(nrow(df)),]

for(i in 1:10){
  #Segement your data by fold using the which() function
  testIndexes <- which(folds==i,arr.ind=TRUE)

  df.cv = df[testIndexes,]

  model.train = suppressWarnings(glm(formula=TARGET_Adjusted~.,family=binomial,data=df.cv))
}

ptest = suppressWarnings(predict(model.train,newdata=data.frame(xtest),type="response"))
btest=floor(ptest+0.5)
conf.matrix = table(ytest,btest)
conf.matrix
```

```
##      btest
## ytest   0   1
##     0 263  51
##     1  48  38
```

```r
#Accuracy
error=(conf.matrix[1,2]+conf.matrix[2,1])/n.test
error
```

```
## [1] 0.2475
```

```r
acc = 1 - error
acc
```

```
## [1] 0.7525
```

```r
#treating the given binary values "1/0" as "positive/negative" so switching TN with TP and FN with FP i
TN = conf.matrix[1,1]
TP = conf.matrix[2,2]
FN = conf.matrix[2,1]
FP = conf.matrix[1,2]

Precision = TP/(TP+FP)
Precision
```

```
## [1] 0.4269663
```

```r
Recall = TP/(TP+FN)
Recall
```

```
## [1] 0.4418605
```

```r
F1 = 2*(Precision*Recall)/(Precision+Recall)
F1
```

```
## [1] 0.4342857
```

With random shuffling added we can observe the accuracy to be between 75 -80%. We also have decent precision and recall ranges.

Let us build the Lift Chart for this model:

```r
#Plotting LIFT
df=cbind(ptest,ytest)
df[1:20,]
```

```
##             ptest ytest
## 1601 1.052897e-01     0
## 1602 2.945523e-01     1
## 1603 6.347918e-09     0
## 1604 6.210373e-09     0
## 1605 2.812638e-01     0
## 1606 7.061886e-01     1
## 1607 3.275931e-03     0
## 1608 8.147171e-02     0
## 1609 4.379567e-09     0
## 1610 1.410419e-01     0
## 1611 6.226162e-02     0
## 1612 3.362736e-01     0
## 1613 4.058294e-01     1
## 1614 2.097948e-10     0
## 1615 4.036424e-01     1
## 1616 7.693797e-01     0
## 1617 1.044139e-01     0
## 1618 5.686884e-03     0
## 1619 4.592545e-01     0
## 1620 1.577288e-08     0
```

```r
summary(df)
```

```
##      ptest              ytest
##  Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:0.000
##  Median :0.05147   Median :0.000
##  Mean   :0.22752   Mean   :0.215
##  3rd Qu.:0.44130   3rd Qu.:0.000
##  Max.   :1.00000   Max.   :1.000
```

```r
rank.df=as.data.frame(df[order(ptest,decreasing=TRUE),])
colnames(rank.df) = c('predicted','actual')
rank.df[1:20,]
```

```
##      predicted actual
## 1627 1.0000000      1
## 1632 1.0000000      0
## 1710 1.0000000      0
## 1981 1.0000000      0
## 1638 1.0000000      0
## 1972 1.0000000      1
## 1874 1.0000000      1
## 1700 1.0000000      1
## 1868 1.0000000      1
## 1810 0.9999999      0
## 1656 0.9999997      0
## 1682 0.9519467      0
## 1947 0.9377505      0
## 1871 0.9374046      0
## 1702 0.9289868      0
## 1817 0.9194620      0
## 1781 0.9148226      0
## 1887 0.8964318      1
## 1663 0.8694135      1
## 1973 0.8546807      1
```

```r
baserate=mean(ytest)
baserate
```

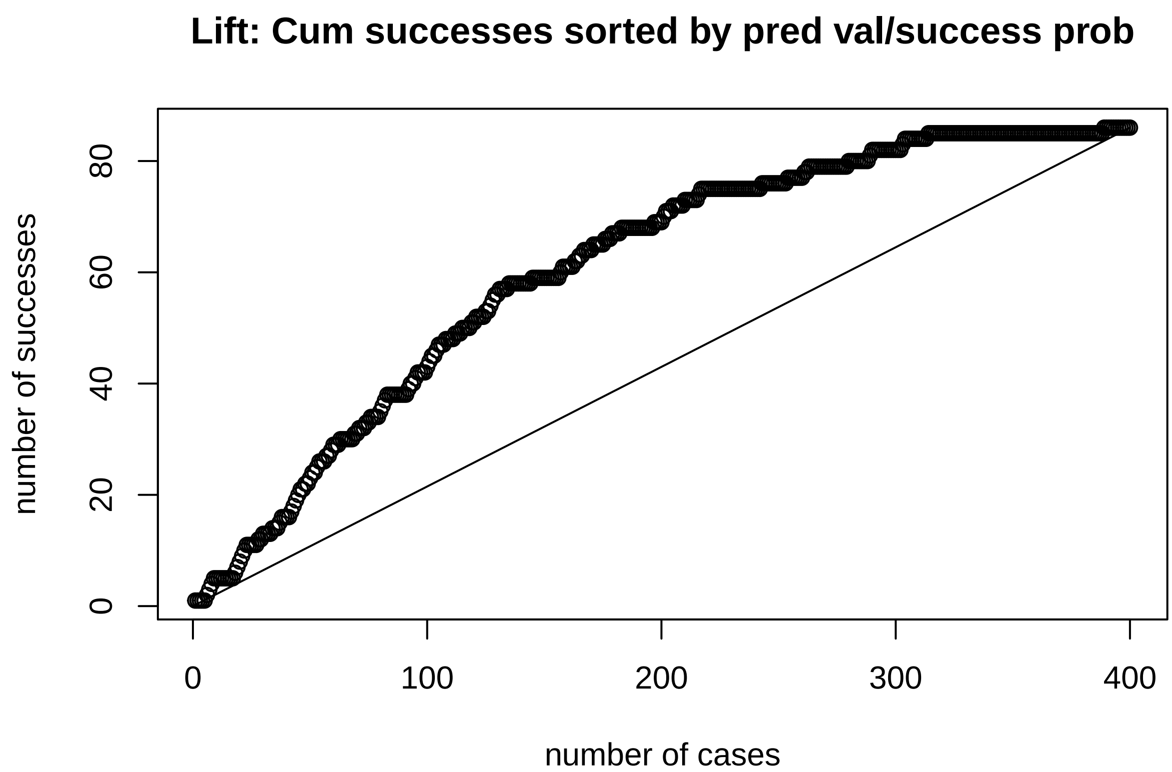
```
## [1] 0.215
```

```r
ax=dim(n.test)
ay.base=dim(n.test)
ay.pred=dim(n.test)
ax[1]=1
ay.base[1]=baserate
ay.pred[1]=rank.df$actual[1]
for (i in 2:n.test) {
  ax[i]=i
  ay.base[i]=baserate*i ## uniformly increase with rate xbar
  ay.pred[i]=ay.pred[i-1]+rank.df$actual[i]
}

df=cbind(rank.df,ay.pred,ay.base)
df[1:20,]
```

```
##      predicted actual ay.pred ay.base
## 1627 1.0000000      1       1   0.215
## 1632 1.0000000      0       1   0.430
## 1710 1.0000000      0       1   0.645
## 1981 1.0000000      0       1   0.860
## 1638 1.0000000      0       1   1.075
## 1972 1.0000000      1       2   1.290
## 1874 1.0000000      1       3   1.505
## 1700 1.0000000      1       4   1.720
## 1868 1.0000000      1       5   1.935
## 1810 0.9999999      0       5   2.150
```

```
## 1656 0.9999997      0       5   2.365
## 1682 0.9519467      0       5   2.580
## 1947 0.9377505      0       5   2.795
## 1871 0.9374046      0       5   3.010
## 1702 0.9289868      0       5   3.225
## 1817 0.9194620      0       5   3.440
## 1781 0.9148226      0       5   3.655
## 1887 0.8964318      1       6   3.870
## 1663 0.8694135      1       7   4.085
## 1973 0.8546807      1       8   4.300
```

```r
plot(ax,ay.pred,xlab="number of cases",ylab="number of successes",main="Lift: Cum successes sorted by pr
points(ax,ay.base,type="l")
```



As we can observe above we can see a good lift from the baseline. Let us look at the ROC chart and find the AUC.

```r
#Plotting ROC
suppressWarnings(library(ROCR))
```

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

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

```
##
##     lowess
```

```r
data=data.frame(predictions=ptest,labels=ytest)
data[1:10,]
```

```
##       predictions labels
## 1601 1.052897e-01      0
## 1602 2.945523e-01      1
## 1603 6.347918e-09      0
## 1604 6.210373e-09      0
## 1605 2.812638e-01      0
## 1606 7.061886e-01      1
## 1607 3.275931e-03      0
## 1608 8.147171e-02      0
## 1609 4.379567e-09      0
## 1610 1.410419e-01      0
```

```r
pred <- prediction(data$predictions,data$labels)
str(pred)
```

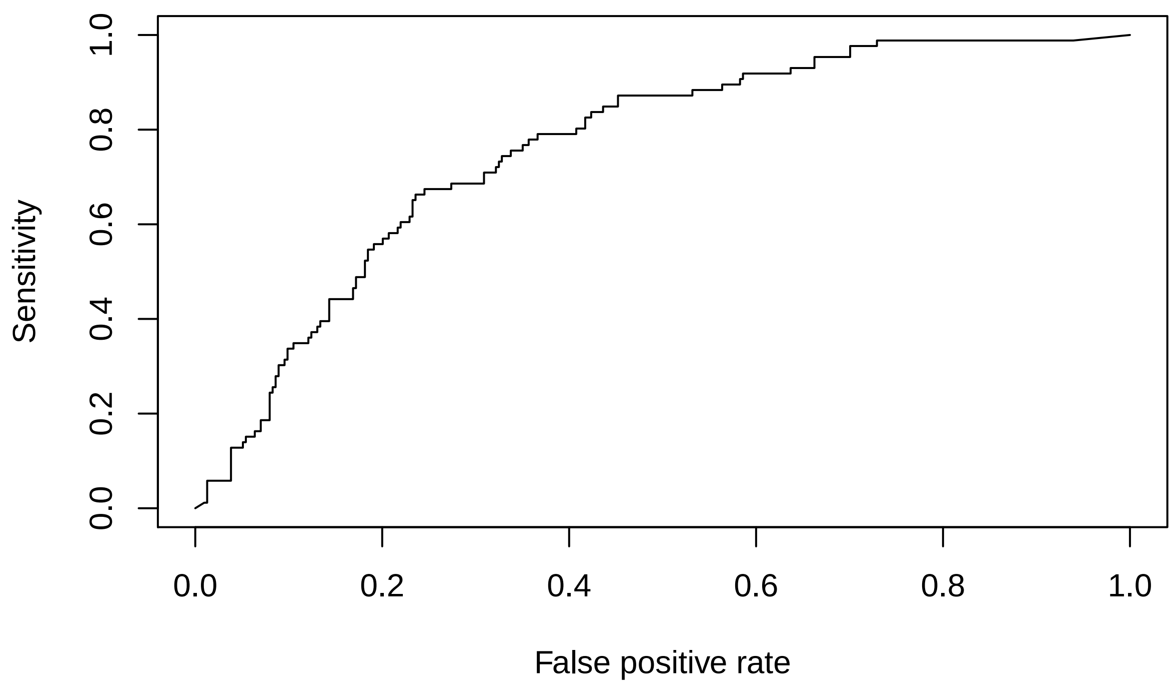
```
## Formal class 'prediction' [package "ROCR"] with 11 slots
##   ..@ predictions:List of 1
##   .. ..$ : num [1:400] 1.05e-01 2.95e-01 6.35e-09 6.21e-09 2.81e-01 ...
##   ..@ labels     :List of 1
##   .. ..$ : Ord.factor w/ 2 levels "0"<"1": 1 2 1 1 1 2 1 1 1 1 ...
##   ..@ cutoffs    :List of 1
##   .. ..$ : num [1:379] Inf 1 1 1 1 ...
##   ..@ fp         :List of 1
##   .. ..$ : num [1:379] 0 3 4 4 4 4 4 5 6 7 ...
##   ..@ tp         :List of 1
##   .. ..$ : num [1:379] 0 1 1 2 3 4 5 5 5 5 ...
##   ..@ tn         :List of 1
##   .. ..$ : num [1:379] 314 311 310 310 310 310 310 309 308 307 ...
##   ..@ fn         :List of 1
##   .. ..$ : num [1:379] 86 85 85 84 83 82 81 81 81 81 ...
##   ..@ n.pos      :List of 1
##   .. ..$ : int 86
##   ..@ n.neg      :List of 1
##   .. ..$ : int 314
##   ..@ n.pos.pred :List of 1
##   .. ..$ : num [1:379] 0 4 5 6 7 8 9 10 11 12 ...
##   ..@ n.neg.pred :List of 1
##   .. ..$ : num [1:379] 400 396 395 394 393 392 391 390 389 388 ...
```

```r
perf <- performance(pred, "sens", "fpr")
str(perf)
```

```
## Formal class 'performance' [package "ROCR"] with 6 slots
##   ..@ x.name      : chr "False positive rate"
##   ..@ y.name      : chr "Sensitivity"
##   ..@ alpha.name  : chr "Cutoff"
```

```
##   ..@ x.values    :List of 1
##   .. ..$ : num [1:379] 0 0.00955 0.01274 0.01274 0.01274 ...
##   ..@ y.values    :List of 1
##   .. ..$ : num [1:379] 0 0.0116 0.0116 0.0233 0.0349 ...
##   ..@ alpha.values:List of 1
##   .. ..$ : num [1:379] Inf 1 1 1 1 ...
```

```r
plot(perf)
```




```r
#Finding AUC
suppressWarnings(library(pROC))
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
auc <-  auc(data$labels, data$predictions)
auc
```

```
## Area under the curve: 0.761
```

We can see a good range of sensitivity above the linear baseline of 50-50 chance. As sensitivity increases we get higher FP rate and that is a tradeoff but the increase in not linear, if it were linear it would have been a bad model.

The high AUC value further confirms our findings above.

3.b. Computing Odds ratio for best model.

For computing the odds ratio we can consider the variables showing significance. We can observe that from the model summary as below:

```
model.final = glm(formula=TARGET_Adjusted~Age+Employment+
                    Education+Marital+Occupation+Income+Hours,family=binomial,data=audit.clean)

exp(model.final$coefficients)
```

```
##                 (Intercept)                          Age
##                1.823785e-03                 1.032015e+00
##           EmploymentPrivate            EmploymentPSFederal
##                1.369298e+00                 1.317955e+00
##           EmploymentPSLocal              EmploymentPSState
##                1.072531e+00                 1.274635e+00
##           EmploymentSelfEmp          EmploymentUnemployed
##                1.075925e+00                 2.167621e-05
##           EmploymentUnknown            EmploymentVolunteer
##                2.021543e+00                 2.011959e-08
##           EducationBachelor              EducationCollege
##                1.105429e+00                 4.327473e-01
##          EducationDoctorate               EducationHSgrad
##                2.591162e+00                 3.207101e-01
##             EducationMaster            EducationPreschool
##                1.585154e+00                 1.599432e-07
##       EducationProfessional           EducationVocational
##                5.474589e+00                 3.468145e-01
##              EducationYr10                 EducationYr11
##                2.214366e-01                 1.875861e-01
##              EducationYr12                EducationYr1t4
##                1.585028e-01                 3.554358e-08
##             EducationYr5t6                EducationYr7t8
##                1.260861e-01                 5.481506e-08
##               EducationYr9               MaritalDivorced
##                5.010088e-02                 9.190138e-01
##             MaritalMarried MaritalMarried-spouse-absent
##                1.467247e+01                 1.276927e+00
##           MaritalUnmarried                MaritalWidowed
##                1.917355e+00                 9.064054e-01
##          OccupationClerical           OccupationExecutive
##                3.297438e+00                 5.006916e+00
##           OccupationFarming                OccupationHome
##                1.070590e+00                 2.915048e-06
##        OccupationMachinist             OccupationMilitary
```

```
##                1.602257e+00                2.413422e-06
##      OccupationProfessional        OccupationProtective
##                3.379351e+00                5.960450e+00
##            OccupationRepair             OccupationSales
##                2.119397e+00                2.587714e+00
##           OccupationService           OccupationSupport
##                6.265928e-01                3.424210e+00
##         OccupationTransport           OccupationUnknown
##                1.289794e+00                          NA
##                      Income                       Hours
##                1.000002e+00                1.033908e+00
```

We can observe the following effect for the predictors with highest odds rato:

1. MaritalMarried - 14.67: Which means those individuals have the 14.67 times the odds of being a productive audit than the
   other Individuals.
2. EducationProfessional - 5.47: Which means those individuals have the 5.47 times the odds of being a productive audit than the other Individuals.
3. OccupationExecutive - 5: Which means those individuals have the 5 times the odds of being a productive audit than the other Individuals
4. OccupationProtective - 5.9: Which means those individuals have the 5.9 times the odds of being a productive audit than the other Individuals

We can observe that MaritalMarried has the odds of being a productive audit by 128.31 times than the other variables. Age and Hours show odds of 1.031 and 1.043 repectively.

4 Applying linear and non-linear regression analysis to predict RISK_Adjustment

4.a. Simple regression using all predictors.

Our first model is one that uses all the predictors. Please note that we do not use TARGET_Adjusted as a predictor now as per discussions on Piazza. We see the significance of adding that in below sections

```
audit.lm=audit.clean

audit.lm = model.matrix(RISK_Adjustment~ Age+Employment+Gender
                        +Education+Marital+Occupation+Income+Hours+Deductions,data=audit.lm)[,-1]

audit.df = data.frame(RISK_Adjustment=audit.clean$RISK_Adjustment,audit.lm)

fitFull = lm(RISK_Adjustment~., data=audit.df)

summary(fitFull)
```

```
##
## Call:
## lm(formula = RISK_Adjustment ~ ., data = audit.df)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -14702  -2576   -590    609 104027
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.101e+02  1.856e+03   0.329   0.7425
## Age                         3.448e+01  1.729e+01   1.994   0.0463 *
## EmploymentPrivate          -1.426e+03  7.251e+02  -1.966   0.0494 *
## EmploymentPSFederal        -1.898e+03  1.215e+03  -1.562   0.1184
## EmploymentPSLocal           9.483e+02  1.057e+03   0.897   0.3699
## EmploymentPSState          -1.700e+03  1.199e+03  -1.418   0.1562
## EmploymentSelfEmp          -1.552e+02  1.141e+03  -0.136   0.8919
## EmploymentUnemployed       -4.297e+02  8.175e+03  -0.053   0.9581
## EmploymentUnknown          -1.184e+03  1.380e+03  -0.858   0.3908
## EmploymentVolunteer        -5.082e+03  8.190e+03  -0.621   0.5349
## GenderMale                 -1.108e+02  4.981e+02  -0.223   0.8239
## EducationBachelor          -6.205e+02  1.082e+03  -0.574   0.5664
## EducationCollege           -1.017e+03  1.056e+03  -0.963   0.3354
## EducationDoctorate          9.414e+02  1.899e+03   0.496   0.6202
## EducationHSgrad            -1.613e+03  1.040e+03  -1.551   0.1212
## EducationMaster             9.971e+02  1.311e+03   0.760   0.4471
## EducationPreschool         -1.974e+03  3.494e+03  -0.565   0.5720
## EducationProfessional       8.213e+03  1.966e+03   4.179 3.06e-05 ***
## EducationVocational        -1.702e+03  1.316e+03  -1.294   0.1959
## EducationYr10               5.945e+01  1.467e+03   0.041   0.9677
## EducationYr11              -1.530e+03  1.390e+03  -1.101   0.2711
## EducationYr12              -1.480e+03  2.205e+03  -0.671   0.5021
## EducationYr1t4             -3.874e+03  3.453e+03  -1.122   0.2621
## EducationYr5t6             -2.717e+03  1.983e+03  -1.371   0.1706
## EducationYr7t8             -2.619e+03  1.716e+03  -1.526   0.1272
## EducationYr9               -2.592e+03  1.854e+03  -1.398   0.1622
## MaritalDivorced            -7.208e+02  6.498e+02  -1.109   0.2675
## MaritalMarried              2.724e+03  5.177e+02   5.260 1.59e-07 ***
## MaritalMarried.spouse.absent -1.029e+03  1.786e+03  -0.576   0.5645
## MaritalUnmarried           -2.250e+02  1.055e+03  -0.213   0.8311
## MaritalWidowed             -1.011e+03  1.231e+03  -0.821   0.4115
## OccupationClerical          5.628e+01  1.044e+03   0.054   0.9570
## OccupationExecutive         2.740e+02  1.029e+03   0.266   0.7901
## OccupationFarming          -1.941e+03  1.405e+03  -1.381   0.1675
## OccupationHome              2.800e+02  3.732e+03   0.075   0.9402
## OccupationMachinist        -1.090e+03  1.097e+03  -0.994   0.3202
## OccupationMilitary          1.100e+02  8.148e+03   0.014   0.9892
## OccupationProfessional      5.630e+01  1.101e+03   0.051   0.9592
## OccupationProtective       -9.032e+02  1.610e+03  -0.561   0.5749
## OccupationRepair           -8.450e+02  1.018e+03  -0.830   0.4068
## OccupationSales            -1.897e+02  1.042e+03  -0.182   0.8555
## OccupationService          -5.454e+02  1.027e+03  -0.531   0.5953
## OccupationSupport           9.031e+02  1.465e+03   0.617   0.5375
## OccupationTransport        -2.120e+03  1.166e+03  -1.818   0.0691 .
## OccupationUnknown                  NA         NA      NA       NA
## Income                      2.860e-03  3.078e-03   0.929   0.3528
## Hours                       3.033e+01  1.641e+01   1.848   0.0648 .
## Deductions                  1.005e+00  5.359e-01   1.875   0.0609 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8032 on 1953 degrees of freedom
## Multiple R-squared:  0.09416,    Adjusted R-squared:  0.07283
## F-statistic: 4.413 on 46 and 1953 DF,  p-value: < 2.2e-16
```

```
model.mse = mean(residuals(fitFull)^2)
model.mse
```

```
## [1] 63002753
```

```
rmse = sqrt(model.mse)
rmse
```

```
## [1] 7937.427
```

We see that we have a poor model with R squared of 0.094 and Adjusted R-squared of 0.072 . We see significance of MaritalMarried, EducationProfessional, and slight significance for Age and EmploymentPrivate. However, there is high residual error 8032, and high RMSE of 7937 (RMSE taken without cross-validation). These values show that the model is not useful as well as the performance is low(High RMSE) and the data does not fit the model well(high residual error).

Below is the effect of the significant predictors:

1. A unit increase in MaritalMarried influences a 2.730e+03 increase in RISK_Adjustment, controlling for all other predictors.
2. A unit increase in EducationProfessional influences an 8.212e+03 increase in RISK_Adjustment, controlling for all other predictors.
3. A unit increase in Age may influence a 3.612e+01 increase in RISK_Adjustment, controlling for all other predictors.
4. A unit increase in EmploymentPrivate may influence a -1.426e+03 decrease in RISK_Adjustment, controlling for all other predictors.

The high p-values of all the other variables suggests that none of them are linearly related to RISK_Adjustment controlling for all other variables.

Together all these predictors account for only 9.4% of the variance in RISK_Adjustment across individuals.

4.b. Model Selection

Let us now try to combine various predictors and build linear as well as non linear models and use out-of-sample evaluation to select the best model.

We will describe the function first: 1. We use leave one out cross-validation here 2. A model matrix is used to account for Undersampling

```
lm.loov <- function(audit.clean, query) {

audit.lm=audit.clean
audit.lm[1:3,]
```

```
audit.lm = model.matrix(query,data=audit.lm)[,-1]
audit.lm[1:3,]

audit.df = data.frame(RISK_Adjustment=audit.clean$RISK_Adjustment,audit.lm)

## leave-one-out cross validation
n = length(audit.clean$RISK_Adjustment)
n
error = dim(n)
for (k in 1:n) {
  train1 = c(1:n)
  train2 = train1[train1!=k] ## pick elements that are different from k
  m2 = lm(RISK_Adjustment ~., data=audit.df[train2,])
  pred = predict(m2, newdat=audit.df[-train2 ,])
  obs = audit.clean$RISK_Adjustment[-train2]
  error[k] = obs-pred
}

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

Our first model is one that uses all the predictors but now with cross validation:

```
formulaA = RISK_Adjustment~ Age+Employment+Education+Marital+Occupation+Income+Gender+Deductions+Hours
suppressWarnings(rmseA <- lm.loov(audit.clean, formulaA))
rmseA
```

```
## [1] 8149.902
```

We see an increase in the RMSE when we use cross-validation showing that the previous simple model was over-fitting the data.

Let us perform a test by using only the significant variables of Education, Marital and Age.

```
formulaB = RISK_Adjustment~ Age+Education+Marital
suppressWarnings(rmseB <- lm.loov(audit.clean, formulaB))
rmseB
```

```
## [1] 8104.054
```

Yes we have a better model than before.

Let us now use STEPAIC to see what variables are suggested to be present. I ran the following command and present here only the final formula and scores after variable selection as the process ran for close to 27 pages and including that in the report is an overkill:

library(MASS)

fitSelect = lm(RISK_Adjustment ~ Age + Employment
+ Education + Marital + Occupation+ Income+ Gender+ Deductions+ Hours, data=audit.clean)
stepAIC(fitFull, direction="backward")

Call: lm(formula = RISK_Adjustment ~ EmploymentPrivate + EmploymentPSLocal + EducationProfessional + EducationYr10 + OccupationProtective + TARGET_Adjusted, data = audit.df)

Coefficients: (Intercept) EmploymentPrivate EmploymentPSLocal EducationProfessional EducationYr10
445.4 -846.5 2001.8 6313.5 1586.9
OccupationProtective TARGET_Adjusted
-1973.6 8515.5

So we see that taking in Age, Employment, Education, Marital, Occupation, Hours and Deductions might give a better model. Let us do that

```
formulaC = RISK_Adjustment~ Age+Education+Marital+Employment+Occupation+Hours+Deductions
suppressWarnings(rmseC <- lm.loov(audit.clean, formulaC))
rmseC
```

```
## [1] 8145.022
```

Not the value we might have hoped for. Let us head for non-linear regression using polyfit. Let us try it out for the best score we got before for Age, Education and Marital. Note: I chose 4 as the degree for Age after trials not shown here.

```
formulaD = RISK_Adjustment~ poly(Age, degree=4) + Education + Marital
suppressWarnings(rmseD <- lm.loov(audit.clean, formulaD))
rmseD
```

```
## [1] 8084.639
```

We have the best rmse so far. Can we improve further? Let us try

```
formulaE = RISK_Adjustment~poly(Age, degree=4) +Education + Marital + poly(Deductions, degree=5)+ Hours
suppressWarnings(rmseE <- lm.loov(audit.clean, formulaE))
rmseE
```

```
## [1] 8087.073
```

The rmse has increased so our best model so far can be with formulaD = RISK_Adjustment~ poly(Age, degree=4) + Education + Marital with RMSE of 8084.639

In addition, I just tried to see how much TARGET_Adjusted is related to RISK_Adjustment by including that in one of the models. Below is the model that I got:

```
formulaTAdjusted = RISK_Adjustment~ Employment + TARGET_Adjusted
suppressWarnings(rmseTA <- lm.loov(audit.clean, formulaTAdjusted))
rmseTA
```

```
## [1] 7506.031
```

The above model shows a drastic decrease in RMSE but since TARGET_Adjusted is also a response variable we do not include it.

4.c. Determining Most important predictor

Since we are choosing the following model as the best(Not considering TARGET_Adjusted). Let us try and narrow down the best predictor from it:

formulaD = RISK_Adjustment~ poly(Age, degree=4) + Education + Marital

Removing Age

```
formulaD = RISK_Adjustment~  Education + Marital
suppressWarnings(rmseD <- lm.loov(audit.clean, formulaD))
rmseD
```

```
## [1] 8111.894
```

Removing Age increases the RMSE of the model by 8111.894 - 8084.639 = 27.255. Let us see the impact of others.

Removing Education

```
formulaD = RISK_Adjustment~ poly(Age, degree=4) + Marital
suppressWarnings(rmseD <- lm.loov(audit.clean, formulaD))
rmseD
```

```
## [1] 8137.712
```

Removing the Education definitely has a greater impact. The difference in RMSE is 8137.712 - 8084.639 = 53.073.

Removing Marital

```
formulaD = RISK_Adjustment~ poly(Age, degree=4) + Education
suppressWarnings(rmseD <- lm.loov(audit.clean, formulaD))
rmseD
```

```
## [1] 8189.021
```

Removing Marital definitely has the greatest impact we did see the influence in the Conditional Histogram too. We also see this influence for TARGET_Adjusted which can suggest a link. The difference in RMSE is 8189.021 - 8084.639 = 104.382

From the above analysis we can see that Marital is the most important predictor.

Conclusion

This assignment showed how to perfrom Logistic regression and what are the performance measures and how it is different from Linear Regression. Summarising:

1. Initial anaysis involves understanding various variables and the values they hold. Highlighting missing values.

2. Using conditional histograms and facets for various categories w.r.t. the response variables

3. Analysing the descriptive statistics of the variables

4. Building a baseine model and checking model fit, accuracy and performance.
5. Using characteristics like accuracy, precision, recall, AUC, ROC, Lift Chart etc to analyse model fit and performance for logistic
   models
6. Continuously imrpoving the model using apporaches like variable selection, 10-fold cross-validation etc
7. Identifying importance of each variable.

References:

1. Cross-validation - https://en.wikipedia.org/wiki/Cross-validation_(statistics)
2. ROC - http://gim.unmc.edu/dxtests/roc2.htm