Gender Gap among Nation Origins and Enterprises

C. Francesco, O Matteo, W. Anna

31/07/2022

Introduction

In our seminar we analyzed a data set from the U.S. Federal Housing Finance Agency. The Enterprise PUDB contains data on single-family and multifamily mortgages purchased by the Federal National Mortgage Association (**Fannie Mae**) and the Federal Home Loan Mortgage Corporation (**Freddie Mac**) for calendar year 2019.

Our goal is to investigate the data set to test our research question, which is: does the data set have a gender and ethnicity gap between different Enterprises?

EDA

The data set consists of sixty-four variables for sixty-four millions of observations: thus, it is a high dimensional data set. Because it is composed by raw data, we have to deal with many issues and clean the data first. For this purpose, the variables were examined by the original description of them. We decided to **identify the missing observations** according to the description contained into the data set, then we continued by renaming these observations in NA's in order to count them and to investigate about the completeness of these variables in R program. By the second step, we **controlled the scale of all variables in data set**. Collected all this information we have been able to continue the analysis and to formulate our research question. As soon as we knew that the focus of our research was to focus on the comparison between gender and enterprises, we started the analysis.

We started with selecting the list of variables which are relevant for our research question. The following have been selected: Enterprise Flag, Borrower's Annual Income, Borrower Income Ratio, Acquisition Unpaid Principal Balance, Purpose of Loan, Federal Guarantee, First-Time Home Buyer, National Origin, Borrower Gender, Co-Borrower Gender, Age of Borrower, Age of Co-Borrower, Rate Spread, Loan-to-Value Ratio, Interest Rate at Origination, Credit Score Model - Borrower, Automated Underwriting System (AUS) Name, High Opportunity Area, Debt-to-Income Ratio, Property Value.

A subset containing all the observations for the selected variables has been created by the principal one, then it has split again per gender and enterprise. Finally, both of these additional subsets have been split per nation origin one more time. After that, **we have been able to analyze our research question.**

The comparison of mean values between the groups and in the groups and the visualization by plotting our results, lead to the small correction of our research question. Due to we find that the changes are not big enough for to make more research, we started to try another combinations of variables and find some interesting outcomes, which will be presented in the part of results.

Challenge 1

The first challenge of this particular dataset were the not available values. The problem was that all of them were define for each single variable in the different way. Therefore, we invest a big amount of time to find the definition of the not available values and then to rename it in an actual **“NA” value**.

```
frame$Loan.to.Value.Ratio..LTV.[frame$Loan.to.Value.Ratio..LTV.
frame[frame$Purpose.of.Loan == 9] <- NA
frame$Borrower.Gender[frame$Borrower.Gender == 9] <- NA
frame$Borrower.Gender[frame$Borrower.Gender == 4] <- NA
frame$Borrower.Gender[frame$Borrower.Gender == 3] <- NA
frame$Co.Borrower.Gender[frame$Co.Borrower.Gender == 9] <- NA
frame$Co.Borrower.Gender[frame$Co.Borrower.Gender == 4] <- NA
frame$Age.of.Borrower[frame$Age.of.Borrower == 9] <- NA
frame$Age.of.Co.Borrower[frame$Age.of.Co.Borrower == 9] <- NA
frame$Rate.Spread[frame$Rate.Spread == 0] <- NA
frame$Interest.Rate.at.Origination[frame$Interest.Rate.at.Origi
frame$Automated.Underwriting.System..AUS.[frame$Automated.Under
frame$Credit.Score.Model...Borrower[frame$Credit.Score.Model...
frame$Credit.Score.Model...Borrower[frame$Credit.Score.Model...
frame$Debt.to.Income..DTI..Ratio[frame$Debt.to.Income..DTI..Rat
frame$Property.Value[frame$Property.Value == 999999999] <- NA
frame$First.Time.Home.Buyer[frame$First.Time.Home.Buyer == 9] <
frame$Acquisition.Unpaid.Principal.Balance..UPB.[frame$Acquisit
frame$Borrower.Income.Ratio[frame$Borrower.Income.Ratio == 999
frame$Borrower.Race1[frame$Borrower.Race1 == 9] <- NA
frame$Borrower.Race1[frame$Borrower.Race1 == 7] <- NA
frame$Borrower.Race1[frame$Borrower.Race1 == 6] <- NA
frame$Borrower.s..or.Borrowers...Annual.Income[frame$Borrower.s
frame$Rate.Spread[frame$Borrower.Race1 == 0] <- NA
```

Challenge 2

The second challenge was to understand and define the structure of the analysis. For that reason, we need the **identify different groups for Enterprises and Gender** . The solution was found in splitting of the original, but already framed dataset into two groups for both. Then we decided to split the dataset per Nation Origin. After of that, this sub-group have been divided again per Gender and then per Enterprise.

```
Bg_1 <- frame[c(frame$Borrower.Gender ==1), ]
Bg_2 <- frame[c(frame$Borrower.Gender ==2), ]

Ent_1 <- frame[c(frame$Enterprise.Flag ==1), ]
Ent_2 <- frame[c(frame$Enterprise.Flag ==2), ]

Indian  <- frame[c(frame$Borrower.Race1 ==1), ]
Asian <- frame[c(frame$Borrower.Race1 ==2), ]
AfroAmerican <- frame[c(frame$Borrower.Race1 ==3), ]
White  <- frame[c(frame$Borrower.Race1 ==5), ]

IndianM  <- Indian[c(Indian$Borrower.Gender==1), ]
AsianM <- Asian[c(Asian$Borrower.Gender==1), ]
AfroAmericanM <- AfroAmerican[c(AfroAmerican$Borrower.Gender==1
WhiteM  <- White[c(White$Borrower.Gender==1), ]

IndianF  <- Indian[c(Indian$Borrower.Gender==2), ]
AsianF <- Asian[c(Asian$Borrower.Gender==2), ]
AfroAmericanF <- AfroAmerican[c(AfroAmerican$Borrower.Gender==2
WhiteF  <- White[c(White$Borrower.Gender==2), ]

IndianM_A <- IndianM[c(IndianM$Enterprise.Flag ==1), ]
IndianM_B <- IndianM[c(IndianM$Enterprise.Flag ==2), ]

IndianF_A <- IndianF[c(IndianF$Enterprise.Flag ==1), ]
IndianF_B <- IndianF[c(IndianF$Enterprise.Flag ==2), ]

AsianM_A <- AsianM[c(AsianM$Enterprise.Flag ==1), ]
AsianM_B <- AsianM[c(AsianM$Enterprise.Flag ==2), ]

AsianF_A <- AsianF[c(AsianF$Enterprise.Flag ==1), ]
AsianF_B <- AsianF[c(AsianF$Enterprise.Flag ==2), ]
```

```
WhiteM_A <- WhiteM[c(WhiteM$Enterprise.Flag == 1), ]
WhiteM_B <- WhiteM[c(WhiteM$Enterprise.Flag == 2), ]

WhiteF_A <- WhiteF[c(WhiteF$Enterprise.Flag == 1), ]
WhiteF_B <- WhiteF[c(WhiteF$Enterprise.Flag == 2), ]

AfroAmericanM_A <- AfroAmericanM[c(AfroAmericanM$Enterprise.Fla
AfroAmericanM_B <- AfroAmericanM[c(AfroAmericanM$Enterprise.Fla

AfroAmericanF_A <- AfroAmericanF[c(AfroAmericanF$Enterprise.Fla
AfroAmericanF_B <- AfroAmericanF[c(AfroAmericanF$Enterprise.Fla
```

Data Storytelling

In general, can be highlighted that the outcomes of our research question show only few differences over all data. **Changes between male and female borrowers and changes between Fannie MAE and Freddie MAC per male and female borrowers are slightly different over all**. Nevertheless, by controlling the outcomes per National Origin, we found remarkable differences. This is the reason why this last variable found it place in defining our research question. By National Origin we could observe the existence of differences between male and female borrowers and between the two enterprises.

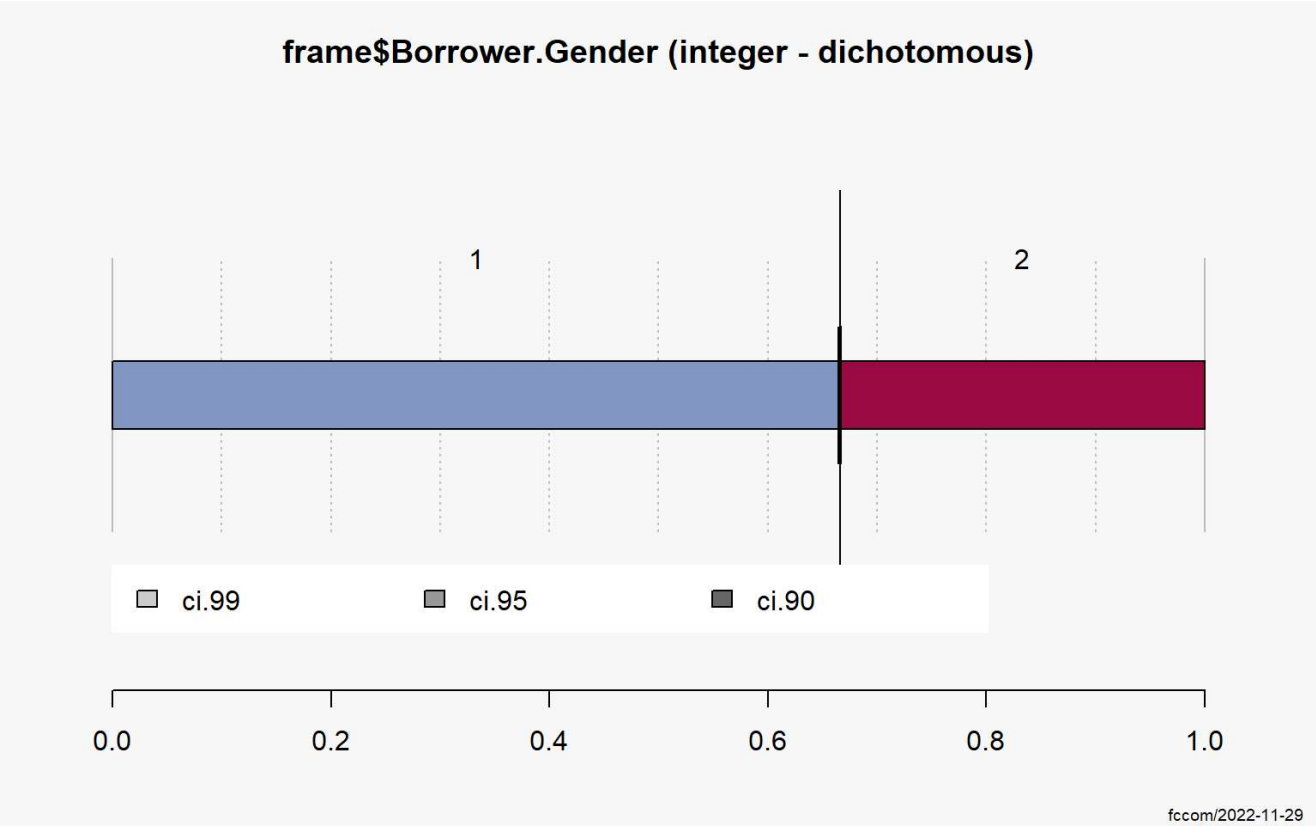
Another **finding is that the interest rate is higher for the female borrowers, rather than for the male ones.** Therefore, women pay more for mortgage than men in both enterprises. Nevertheless, this outcome is even more true for Freddie MAC enterprise than for Fanny MAE.

Finding 1

Gender distribution over the data set:

```
par(bg="#f7f7f7")
Desc(frame$Borrower.Gender)
```

```
## ------------------------------------------------------------
## frame$Borrower.Gender (integer - dichotomous)
##
##      length         n       NAs    unique
##   1'000'000   920'390    79'610         2
##                 92.0%      8.0%
##
##       freq   perc  lci.95  uci.95'
## 1  612'838  66.6%   66.5%   66.7%
## 2  307'552  33.4%   33.3%   33.5%
##
## ' 95%-CI (Wilson)
```

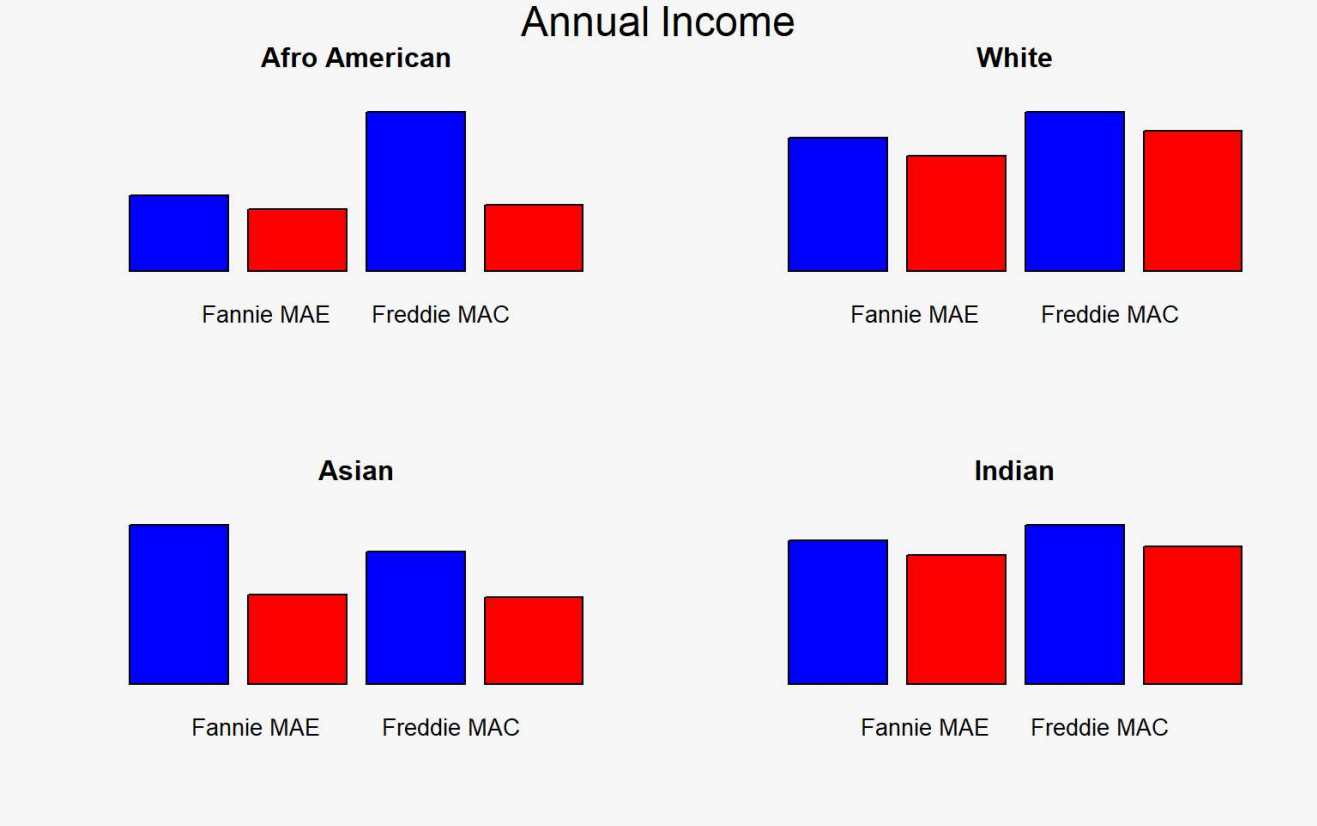


Finding 2

By **controlling per Annual Income and the Level of Default (UPS)**, this proportion applies in favor of men. This means that if men's earnings are high, their mortgages' payment will be high as well.

Annual Income per male and female borrowers between Gender, National Origin and Enterprise:

```
par(bg="#f7f7f7")
AfroAmerican_income <- c(108308.45,88793.86,226317.678359,93758
White_income <- c(122400.52,105985.232129, 145887.555095,128585
Asian_income <- c(214325.907059,120930.78,179053.061316,117527.
Indian_income <- c(103867.20,93309.86,114645.47,99181.91)
par(mfrow=c(2,2))
barplot(AfroAmerican_income, axes=FALSE, col= c("blue", "red",'
barplot(White_income, axes=FALSE, col= c("blue", "red","blue",
barplot(Asian_income, axes=FALSE, col= c("blue", "red","blue",
barplot(Indian_income, axes=FALSE, col= c("blue", "red","blue",
mtext("Annual Income", side = 3, line = -1.5, outer = TRUE, cex
```



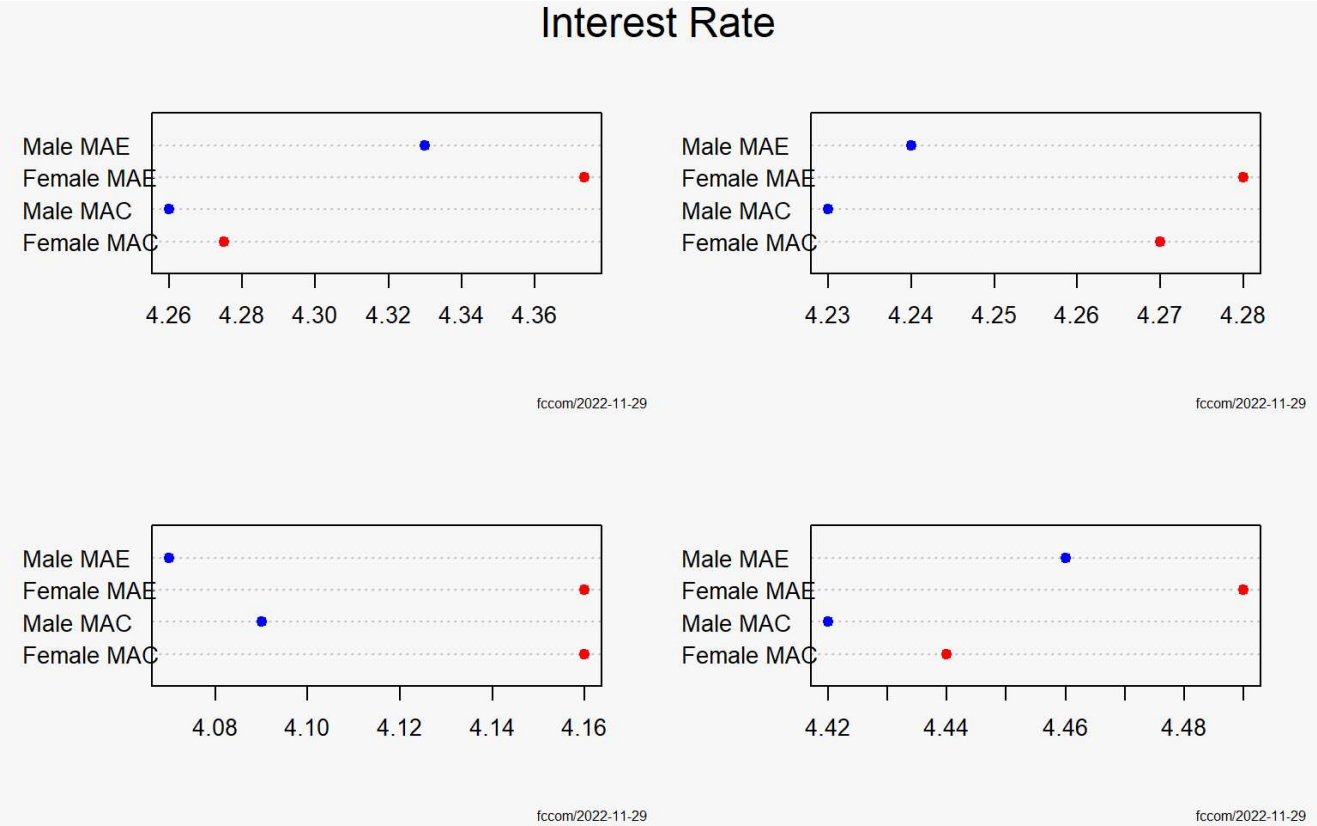
Finding 3

Similar results have been established by taking in consideration Interest Rate and Rate Spread. As far as Interest Rate and Rate of Spread increase, the gender gap declines.

In accordance with the data set we can see that **Fannie MAE presents similar gender proportions for the Spread Rate** in the white group National Origin. It could be explained by the composition of the data set, but if we look at the other groups National Origin the difference becomes bigger, for the Indian and Asian groups. **The second bank shows less gender gap** for Indian, afro American and Asian groups. However, for Freddie MAC the gender gap becomes more evident.

Interest Rates per male and female borrowers between Gender, National Origin Enterprises:

```
par(bg="#f7f7f7")
AfroAmerican_rate <- c(4.46, 4.49, 4.42, 4.44)
white_rate <- c(4.24, 4.28, 4.23, 4.27)
asian_rate <- c(4.07,4.16, 4.09, 4.16)
indian_rate <- c(4.33, 4.3736, 4.26, 4.275)
par(mfrow=c(2,2))
PlotDot(indian_rate, pch=19, col= c("blue", "red","blue", "red'
PlotDot(white_rate, pch=19, col= c("blue", "red","blue", "red")
PlotDot(asian_rate, pch=19, col= c("blue", "red","blue", "red")
PlotDot(AfroAmerican_rate, pch=19, col= c("blue", "red","blue",
mtext("Interest Rate", side = 3, line = -1.5, outer = TRUE, cex
```

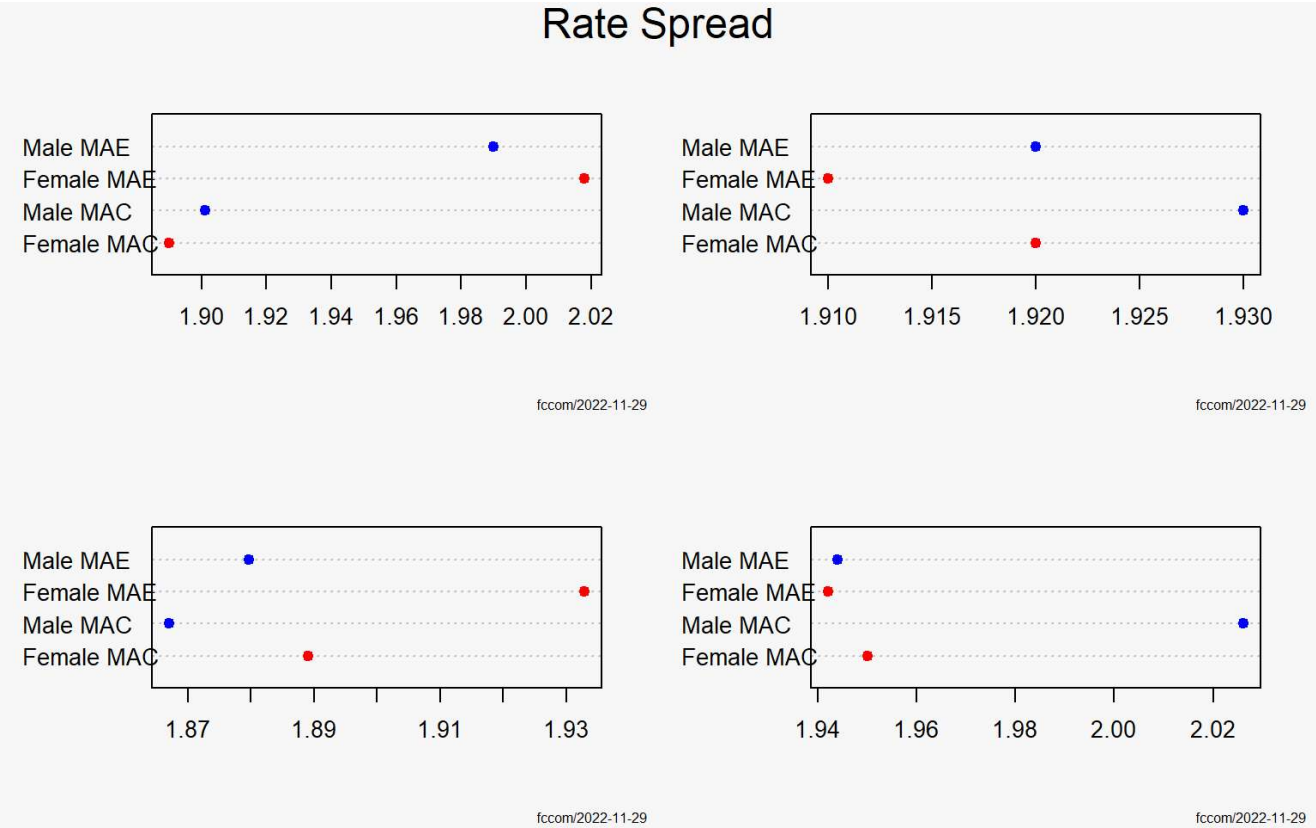


Finding 4

The same changes can be observed by observing the cost of mortgages for two extreme groups: in our case the group of Asian people (with the lower cost of mortgage) and the group of Afro American people (with the highest cost of mortgage). The higher are the earnings, the smaller is the gender gap. **Consequently it seems that the gap between female and male borrowers is relevant by controlling for national origin.**

Rate Spread per male and female borrowers between Gender, National Origin Enterprises:

```
par(bg="#f7f7f7")
indian_spread <- c(1.990, 2.018,1.9011,1.89)
white_spread <- c(1.92, 1.91, 1.93, 1.92)
asian_spread <- c(1.8797, 1.9329, 1.867, 1.889)
AfroAmerican_spread <- c(1.944, 1.942, 2.0262, 1.95)
par(mfrow=c(2,2))
PlotDot(indian_spread, pch=19, col= c("blue", "red","blue", "re
PlotDot(white_spread, pch=19, col= c("blue", "red","blue", "rec
PlotDot(asian_spread, pch=19, col= c("blue", "red","blue", "rec
PlotDot(AfroAmerican_spread, pch=19, col= c("blue", "red","blue
mtext("Rate Spread", side = 3, line = -1.5, outer = TRUE, cex =
```



Conclusion

In conclusion can be highlighted that in our data set **there is no big gender gap between male and female borrowers** in both Enterprises. This result could be explained by using other variables like the mortgages cost and the annual incomes, which play an important role.

The variable National Origin conduced so to our research question analysis. We could find more differences for male, female and Nationality borrowers for both Enterprises. To sum up, we recommend to analyze deeper the research question.