Regression Analyis of HMDA Loan Data

Loading the Packages for use

```
library(conflicted)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
```

```
library(magrittr)
library(lubridate)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(dplyr)
```

Reading the data

```
# Read data
df <- readRDS('hmdaInterestRate.rds')

# Display the structure of data
str(df)
```

```
## 'data.frame':    6509 obs. of  14 variables:
##  $ activity_year            : Factor w/ 2 levels "2019","2018": 2 2 2 2 2 2 2 2 2 2 ...
##  $ state_code               : Factor w/ 1 level "IL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ county_code              : Factor w/ 3 levels "Missing","Coles",..: 1 1 1 1 2 1 1 1 1 1 ...
##  $ aus_1                    : Factor w/ 6 levels "Desktop Underwriter (DU)",..: 2 1 3 4 1 4 1 1 1 1
##  $ loan_purpose             : Factor w/ 6 levels "Home purchase",..: 2 3 1 1 3 3 1 1 1 1 ...
##  $ applicant_ethnicity_1    : Factor w/ 8 levels "Not Hispanic or Latino",..: 2 1 1 1 2 1 1 1 1 1 ..
##  $ applicant_sex            : Factor w/ 4 levels "Male","Female",..: 1 2 1 1 3 1 1 1 2 1 ...
##  $ derived_loan_product_type: Factor w/ 6 levels "Conventional:First Lien",..: 1 1 1 2 4 2 1 1 1 1 .
##  $ interest_rate            : num  3.62 4.99 4.12 4.25 3.99 ...
##  $ loan_amount              : num  185000 105000 255000 255000 95000 205000 235000 105000 275000 750
##  $ loan_term                : num  180 360 360 360 240 360 360 360 360 360 ...
##  $ property_value           : num  235000 215000 265000 255000 105000 265000 335000 265000 285000 85
##  $ income                   : num  154000 88000 66000 89000 81000 61000 84000 76000 160000 35000 ...
##  $ applicant_age            : num  50 50 30 50 60 40 30 50 30 30 ...
```

Data Preparation

```r
# Replace the values in the following columns with the same value divided by 1,000: loan_amount, proper
df <- df %>%
  mutate(
    loan_amount = loan_amount / 1000,
    property_value = property_value / 1000,
    income = income / 1000
  )


# Create a new column, ltp, that is equal to the values in the loan_amount column divided by the values
df <- df %>%
  mutate(ltp = loan_amount / property_value)


# Filter the data to keep observations for which income is less than 300 (i.e., $300,000).
df <- df %>%
  dplyr::filter(income < 300)


# Display a summary of all columns
summary(df)
```

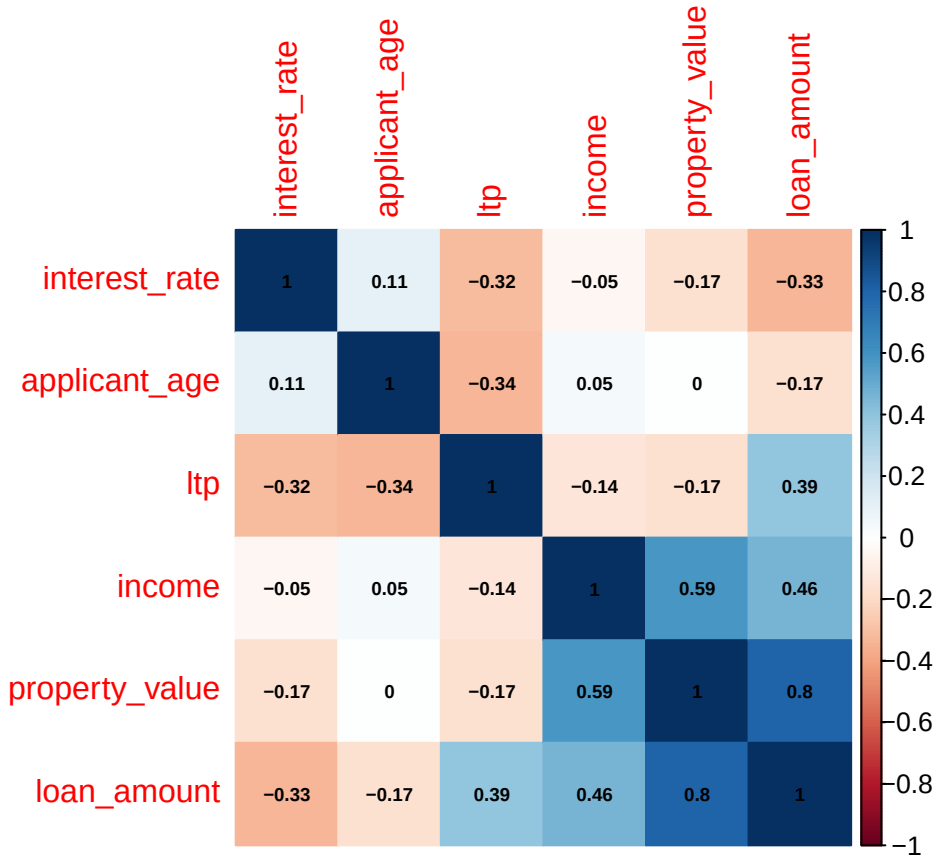
```
##  activity_year state_code     county_code
##  2019:3151     IL:6267    Missing   :5389
##  2018:3116                Coles     : 733
##                           Cumberland: 145
##
##
##
##
##                                                    aus_1
##  Desktop Underwriter (DU)                             :3091
##  Not applicable                                       : 941
##  Loan Propspector (LP) or Loan Product Advisor        :1548
##  Technology Open to Approved Lenders (TOTAL) Scorecard: 470
##  Guaranteed Underwriting System                       : 158
##  Other                                                :  59
##
##                loan_purpose                              applicant_ethnicity_1
##  Home purchase       :3945   Not Hispanic or Latino               :5558
##  Refinancing         :1002   Information not provided by applicant: 475
##  Cash-out refinancing: 824   Hispanic or Latino                   : 204
##  Other purpose       : 247   Other Hispanic or Latino             :  14
##  Home improvement    : 248   Mexican                              :  11
##  Not applicable      :   1   Not applicable                       :   2
##                              (Other)                              :   3
##                                  applicant_sex
##  Male                                   :3869
##  Female                                 :2063
##  information not provided by applicant  : 332
##  Applicant selected both male and female:   3
##
```

```
##
##
##                  derived_loan_product_type interest_rate    loan_amount
##  Conventional:First Lien      :4678        Min.   :2.400   Min.   :  5.0
##  FHA:First Lien               : 605        1st Qu.:3.990   1st Qu.: 85.0
##  Conventional:Subordinate Lien: 485        Median :4.500   Median :125.0
##  VA:First Lien                : 319        Mean   :4.516   Mean   :143.3
##  FSA/RHS:First Lien           : 179        3rd Qu.:4.990   3rd Qu.:185.0
##  FHA:Subordinate Lien         :   1        Max.   :6.980   Max.   :785.0
##
##    loan_term     property_value      income       applicant_age
##  Min.   :  6.0   Min.   : 15.0   Min.   :  1.00   Min.   :30.00
##  1st Qu.:240.0   1st Qu.:115.0   1st Qu.: 51.00   1st Qu.:30.00
##  Median :360.0   Median :165.0   Median : 77.00   Median :40.00
##  Mean   :307.2   Mean   :193.2   Mean   : 89.31   Mean   :45.53
##  3rd Qu.:360.0   3rd Qu.:245.0   3rd Qu.:114.00   3rd Qu.:60.00
##  Max.   :480.0   Max.   :925.0   Max.   :299.00   Max.   :70.00
##
##       ltp
##  Min.   :0.01408
##  1st Qu.:0.69697
##  Median :0.80822
##  Mean   :0.76296
##  3rd Qu.:0.93103
##  Max.   :1.26667
##
```

Creating a correlation matrix and correlation plot

```
# Create correlation matrix
ctrd = cor(df[,c('interest_rate', 'ltp', 'income', 'applicant_age', 'property_value', 'loan_amount')])

# Create the correlation plot
corrplot(ctrd
         , method = 'color'
         , order = 'hclust'
         , addCoef.col = 'black'
         , number.cex = .6
         )
```



The variable loan_amount exhibits the most pronounced negative correlation with interest_rate, standing at -0.33. This negative correlation implies that, as the loan amount increases, there is a tendency for the interest rate to decrease. This phenomenon can be attributed to factors such as risk assessment by lenders and market dynamics, where larger loans may be subject to lower interest rates to attract borrowers or manage risk.

Regression of interest_rate (dependent variable) on ltp (independent variable)

```
# Fit the linear regression model
model <- lm(interest_rate ~ ltp, data = df)

# Display a summary of the fitted model
summary(model)
```

```
##
## Call:
## lm(formula = interest_rate ~ ltp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8933 -0.5055  0.0285  0.4834  2.5221
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.31852    0.03147  168.98   <2e-16 ***
## ltp         -1.05188    0.03950  -26.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.7182 on 6265 degrees of freedom
## Multiple R-squared:  0.1017, Adjusted R-squared:  0.1015
## F-statistic: 709.1 on 1 and 6265 DF,  p-value: < 2.2e-16
```

The coefficient estimate for ltp is -1.05188. This implies that, on average, for every one-unit increase in the loan-to-property (ltp), the interest rate is expected to decrease by 1.05188 units.This negative relationship aligns with common lending practices and makes sense. A higher ltp is often associated with a lower level of risk for the lender, as it suggests the borrower has a larger equity stake in the property. Lower risk may lead to lenders offering lower interest rates to borrowers with higher equity.

Regression of interest_rate (dependent variable) on loan_amount (independent variable)

```r
# Fit the linear regression model
multi_model <- lm(interest_rate ~ ltp + loan_amount, data = df)

# Display a summary of the fitted model
summary(multi_model)
```

```
##
## Call:
## lm(formula = interest_rate ~ ltp + loan_amount, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.93510 -0.47999  0.03926  0.46210  2.29461
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.3848432  0.0307701  175.00   <2e-16 ***
## ltp         -0.7373702  0.0416838  -17.69   <2e-16 ***
## loan_amount -0.0021367  0.0001105  -19.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6978 on 6264 degrees of freedom
## Multiple R-squared:  0.1522, Adjusted R-squared:  0.152
## F-statistic: 562.5 on 2 and 6264 DF,  p-value: < 2.2e-16
```

- **Change in Adjusted R-squared:**
  - The adjusted R-squared for the multiple predictor model (0.152) is higher than that of the single predictor model (0.1015).
  - The increase in adjusted R-squared suggests that the additional predictor (loan_amount) contributes to explaining more variability in interest_rate. In other words, the model with both predictors provides a better fit to the data compared to the model with just ltp.
- **Change in Coefficient Estimate on ltp:**
  - The coefficient estimate on ltp decreased from -1.05188 in the single predictor model to -0.7373702 in the multiple predictor model.
  - This change suggests that when loan_amount is included in the model, the effect of ltp on interest_rate is attenuated. In other words, the relationship between ltp and interest_rate is influenced by the presence of the additional predictor.

Regression of interest_rate (dependent variable) on interest_rate on ltp, loan_amount, and aus_1 (independent variables)

```
# Fit the linear regression model
multi_model2 <- lm(interest_rate ~ ltp + loan_amount + aus_1, data = df)

# Display a summary of the fitted model
summary(multi_model2)
```

```
##
## Call:
## lm(formula = interest_rate ~ ltp + loan_amount + aus_1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.06937 -0.44178  0.04104  0.44611  2.33427
##
## Coefficients:
##                                                                Estimate Std. Error
## (Intercept)                                                    4.589412   0.039711
## ltp                                                            0.025007   0.048639
## loan_amount                                                   -0.001714   0.000106
## aus_1Not applicable                                            0.905072   0.030109
## aus_1Loan Propspector (LP) or Loan Product Advisor             0.007257   0.020211
## aus_1Technology Open to Approved Lenders (TOTAL) Scorecard     0.166965   0.032965
## aus_1Guaranteed Underwriting System                           -0.032727   0.054296
## aus_1Other                                                     0.403022   0.085503
##                                                               t value Pr(>|t|)
## (Intercept)                                                   115.571  < 2e-16 ***
## ltp                                                             0.514    0.607
## loan_amount                                                   -16.174  < 2e-16 ***
## aus_1Not applicable                                            30.060  < 2e-16 ***
## aus_1Loan Propspector (LP) or Loan Product Advisor              0.359    0.720
## aus_1Technology Open to Approved Lenders (TOTAL) Scorecard      5.065 4.20e-07 ***
## aus_1Guaranteed Underwriting System                            -0.603    0.547
## aus_1Other                                                      4.714 2.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6489 on 6259 degrees of freedom
## Multiple R-squared:  0.2675, Adjusted R-squared:  0.2667
## F-statistic: 326.6 on 7 and 6259 DF,  p-value: < 2.2e-16
```

- **Change in Adjusted R-squared:**
  - The adjusted R-squared for the multiple predictor model (0.2667) is higher than that of the previous model (0.152).
  - The increase in adjusted R-squared suggests that the addition of the aus_1 variable contributes to explaining more variability in interest_rate. Including more predictors has improved the overall model fit.
- **For the ltp Variable:**
  - Coefficient Estimate: The coefficient estimate for ltp changed from -0.7373702 to 0.025007. The change in the coefficient estimate suggests a reversal in the relationship between ltp and interest_rate. Previously, ltp had a negative coefficient, indicating a negative relationship. In the current model, the positive coefficient suggests a positive relationship.
  - P-value: The p-value for ltp increased substantially from <2e-16 (is statistically significant) to

0.607 (is not statistically significant). The increase in the p-value indicates that the relationship is no longer statistically significant.

- **For the loan_amount Variable:**
  - Coefficient Estimate: The coefficient estimate for loan_amount increased from -0.0021367 to -0.001714. The change in the coefficient estimate suggests a slight modification in the impact of loan_amount on interest_rate.
  - P-value: The p-value for loan_amount remains the same with (< 2e-16), indicating that the relationship remains statistically significant.
- **For the aus_1 Variable:**
  - Significant Levels: The aus_1 variable has multiple levels, and each level is associated with a different coefficient estimate.
  - The levels of the aus_1 variable provide information about the impact of different aus on interest_rate. The level "aus_1Not applicable" has a coefficient of 0.9051, indicating that loans with this aus tend to have highest interest rates compared to others. Also, the p-value is statistically significant (< 2e-16) for "aus_1Not applicable".