Assignment #1: German Credit

Luke Schwenke

2023-01-13

Regression Model

```
df <- GermanCredit
df <- df %>% dplyr::select(-Class) # Remove Class variable

full_model <- lm(Amount ~ ., data=df)

# Save and Print Coefficients
full_model_coeff <- full_model$coefficients
full_model_coeff
```

```
##                            (Intercept)                               Duration
##                             6473.24572                              127.22804
##              InstallmentRatePercentage                      ResidenceDuration
##                             -783.07567                              -52.07222
##                                    Age                  NumberExistingCredits
##                                5.75204                               75.35581
##                NumberPeopleMaintenance                              Telephone
##                             -198.28236                             -483.36894
##                          ForeignWorker             CheckingAccountStatus.lt.0
##                             -249.97139                             -104.08052
##         CheckingAccountStatus.0.to.200           CheckingAccountStatus.gt.200
##                              213.45397                             -614.78718
##             CheckingAccountStatus.none         CreditHistory.NoCredit.AllPaid
##                                     NA                              858.17909
##         CreditHistory.ThisBank.AllPaid                 CreditHistory.PaidDuly
##                              -50.32424                              -15.85221
##                    CreditHistory.Delay                 CreditHistory.Critical
##                              102.78461                                     NA
##                         Purpose.NewCar                        Purpose.UsedCar
##                            -1757.06185                            -1119.30058
##            Purpose.Furniture.Equipment               Purpose.Radio.Television
##                            -1836.09216                            -2071.60201
##              Purpose.DomesticAppliance                        Purpose.Repairs
##                            -2464.06380                            -1707.78133
##                      Purpose.Education                       Purpose.Vacation
##                            -1892.28951                                     NA
##                     Purpose.Retraining                       Purpose.Business
##                            -2209.35799                            -1989.46611
##                          Purpose.Other             SavingsAccountBonds.lt.100
##                                     NA                             -327.00664
##         SavingsAccountBonds.100.to.500        SavingsAccountBonds.500.to.1000
##                             -569.23487                             -643.52855
##            SavingsAccountBonds.gt.1000            SavingsAccountBonds.Unknown
##                             -385.98649                                     NA
##                EmploymentDuration.lt.1              EmploymentDuration.1.to.4
##                              115.35227                               51.82530
##              EmploymentDuration.4.to.7                EmploymentDuration.gt.7
##                              113.35764                             -163.18910
##          EmploymentDuration.Unemployed       Personal.Male.Divorced.Seperated
##                                     NA                              480.33678
##              Personal.Female.NotSingle                   Personal.Male.Single
##                              282.09482                              730.16723
##          Personal.Male.Married.Widowed                 Personal.Female.Single
##                                     NA                                     NA
##            OtherDebtorsGuarantors.None     OtherDebtorsGuarantors.CoApplicant
##                              138.71303                              752.18304
##       OtherDebtorsGuarantors.Guarantor                    Property.RealEstate
##                                     NA                             -840.99813
##                     Property.Insurance                      Property.CarOther
##                             -588.80108                             -570.75237
##                       Property.Unknown             OtherInstallmentPlans.Bank
##                                     NA                             -143.90714
##           OtherInstallmentPlans.Stores             OtherInstallmentPlans.None
##                              -62.89011                                     NA
##                           Housing.Rent                            Housing.Own
##                              243.68008                              131.43220
##                        Housing.ForFree                Job.UnemployedUnskilled
##                                     NA                            -1715.11138
##                  Job.UnskilledResident                    Job.SkilledEmployee
##                            -1198.70995                            -1253.90023
## Job.Management.SelfEmp.HighlyQualified
##                                     NA
```

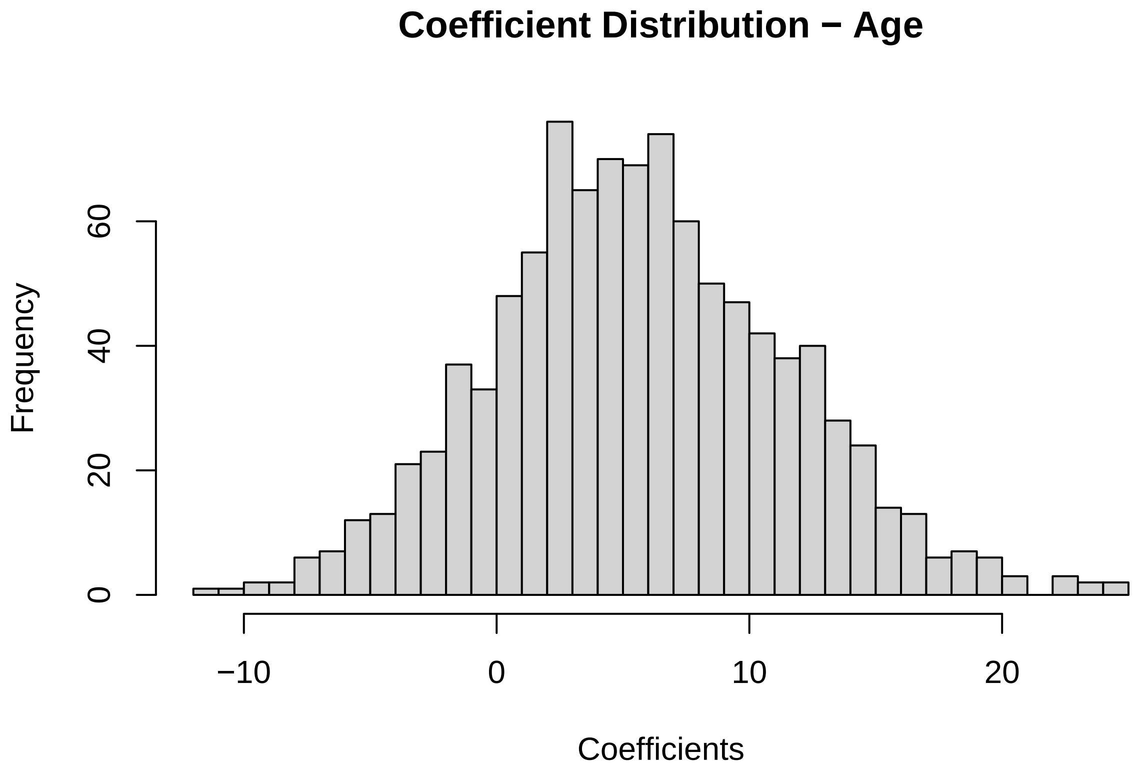
Split into Train/Test Sets and Run Model 1,000 times

```
set.seed(777)

# Create ID to help with splitting into Train/Test
df$id <- 1:nrow(df)

# Initialize empty Data Frame
my_df <- data.frame()

# Run 1,000 Linear models
for(n in 1:1000){

  # Split 63.20% of the data into the train set and the rest into the test set
  train <- df %>% dplyr::sample_frac(0.632)
  test  <- dplyr::anti_join(df, train, by = 'id')
  # Drop id column before running model since it does not have value
  train$id <- NULL
  test$id  <- NULL

  # Run the linear model on all the independent variables
  model <- lm(Amount ~., data=train)

  # Capture the predictions on the Test / Holdout set
  predictions <- predict(model, test)

  # Save the coefficients, and R-squared for the training and holdout
  save_coeff <- model$coefficients
  save_r2_training <- summary(model)$r.squared
  save_r2_holdout <- cor(test$Amount, predict(model, newdata = test))^2

  con <- c(save_coeff, save_r2_training, save_r2_holdout)
  my_df <- rbind(my_df, con)
}

# Remove ID variable and Amount (dependent) variable
df$id <- NULL
df$Amount <- NULL

# Update Column Names
colnames(my_df)[1] <- "(Intercept)"
colnames(my_df)[2:61] <- names(df)
colnames(my_df)[62] <- "training_r2"
colnames(my_df)[63] <- "holdout_r2"

head(my_df, 3)
```

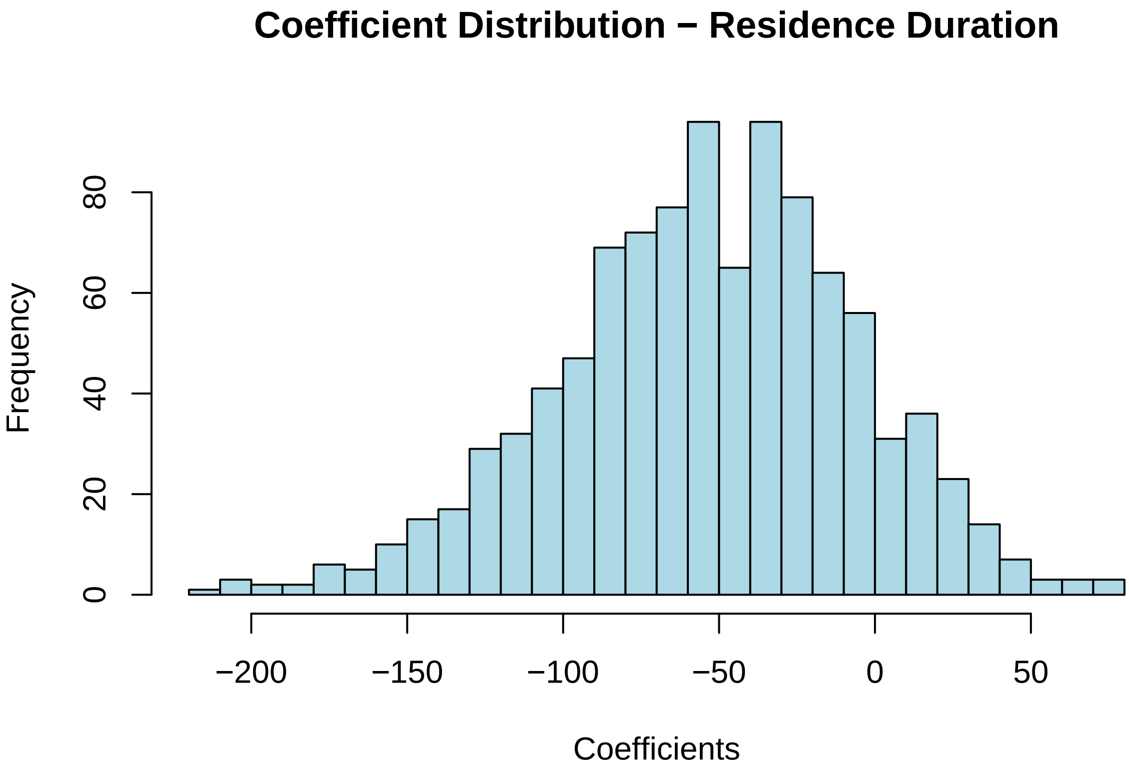
```
##   (Intercept) Duration InstallmentRatePercentage ResidenceDuration        Age
## 1    7053.006 132.5061                 -811.3683          4.879213 -3.3883285
## 2    8742.704 128.5282                 -808.0595       -125.022627  0.4969038
## 3    6213.055 128.2768                 -855.5463         -7.752832 -3.5933384
##   NumberExistingCredits NumberPeopleMaintenance Telephone ForeignWorker
## 1             -64.89621               -276.0670 -361.8853     -332.9213
## 2             117.92245               -111.0212 -513.6035     -139.2039
## 3             159.22881               -217.7662 -372.9430     -196.7809
##   CheckingAccountStatus.lt.0 CheckingAccountStatus.0.to.200
## 1                 -145.86101                       220.0945
## 2                  102.85244                       195.0846
## 3                   39.38775                       379.5955
##   CheckingAccountStatus.gt.200 CheckingAccountStatus.none
## 1                    -584.4721                         NA
## 2                    -677.3200                         NA
## 3                    -410.8699                         NA
##   CreditHistory.NoCredit.AllPaid CreditHistory.ThisBank.AllPaid
## 1                       962.4469                      -402.1201
## 2                       715.3007                       127.1241
## 3                       847.2132                      -264.0714
##   CreditHistory.PaidDuly CreditHistory.Delay CreditHistory.Critical
## 1             -325.39323          -248.09909                     NA
## 2              107.28357           248.44989                     NA
## 3               83.04295           -81.16513                     NA
##   Purpose.NewCar Purpose.UsedCar Purpose.Furniture.Equipment
## 1      -1190.621       -163.0444                   -1110.880
## 2      -3616.351      -2960.3367                   -3900.596
## 3      -2202.571      -1245.5712                   -2258.378
##   Purpose.Radio.Television Purpose.DomesticAppliance Purpose.Repairs
## 1                -1343.701                 -2101.464       -1402.828
## 2                -3987.831                 -4747.327       -3527.883
## 3                -2402.341                 -2819.007       -1916.458
##   Purpose.Education Purpose.Vacation Purpose.Retraining Purpose.Business
## 1         -809.9535               NA          -1437.532        -1149.776
## 2        -3660.1191               NA          -4200.430        -3827.898
## 3        -1971.9619               NA          -3509.607        -2588.997
##   Purpose.Other SavingsAccountBonds.lt.100 SavingsAccountBonds.100.to.500
## 1            NA                  -442.9623                      -636.0945
## 2            NA                  -432.0049                      -633.2180
## 3            NA                  -444.3698                      -320.4523
##   SavingsAccountBonds.500.to.1000 SavingsAccountBonds.gt.1000
## 1                       -813.7547                   -349.6858
## 2                       -499.8858                   -474.0402
## 3                       -798.5119                   -366.0600
##   SavingsAccountBonds.Unknown EmploymentDuration.lt.1 EmploymentDuration.1.to.4
## 1                          NA              -250.82327                -413.26081
## 2                          NA               -79.14711                 -39.77819
## 3                          NA               162.34220                 212.60865
##   EmploymentDuration.4.to.7 EmploymentDuration.gt.7
## 1                -318.16077              -618.95656
## 2                  51.27917               -77.74344
## 3                 281.68776               -37.57796
##   EmploymentDuration.Unemployed Personal.Male.Divorced.Seperated
## 1                            NA                         624.6313
## 2                            NA                         719.7154
## 3                            NA                         678.8079
##   Personal.Female.NotSingle Personal.Male.Single Personal.Male.Married.Widowed
## 1                  333.4951             812.9995                            NA
## 2                  308.9360             714.9203                            NA
## 3                  -31.7201             638.3747                            NA
##   Personal.Female.Single OtherDebtorsGuarantors.None
## 1                     NA                    -88.6265
## 2                     NA                    164.2454
## 3                     NA                    573.1364
##   OtherDebtorsGuarantors.CoApplicant OtherDebtorsGuarantors.Guarantor
## 1                           572.7070                               NA
## 2                           516.7521                               NA
## 3                          1048.1418                               NA
##   Property.RealEstate Property.Insurance Property.CarOther Property.Unknown
## 1           -765.9383          -717.8129         -736.9864               NA
## 2           -725.9376          -507.3184         -430.6413               NA
## 3          -1062.9615          -749.6071         -850.4425               NA
##   OtherInstallmentPlans.Bank OtherInstallmentPlans.Stores
## 1                 -220.58870                    -24.26371
## 2                 -253.11178                   -110.13515
## 3                  -64.04921                     23.08561
##   OtherInstallmentPlans.None Housing.Rent Housing.Own Housing.ForFree
## 1                         NA   232.368137    258.7847              NA
## 2                         NA     8.499965   -170.6788              NA
## 3                         NA   673.179510    764.3539              NA
##   Job.UnemployedUnskilled Job.UnskilledResident Job.SkilledEmployee
## 1               -1402.272             -1144.332           -1258.025
## 2               -1830.999             -1187.417           -1263.488
## 3               -1582.527             -1269.402           -1220.914
##   Job.Management.SelfEmp.HighlyQualified training_r2 holdout_r2
## 1                                     NA   0.6178389  0.5648368
## 2                                     NA   0.6122784  0.5816260
## 3                                     NA   0.6231289  0.5467266
```

Coefficient & R-Squared Distributions

```
hist(my_df$Age,
     breaks=30,
     xlab = "Coefficients", main = "Coefficient Distribution - Age")
```

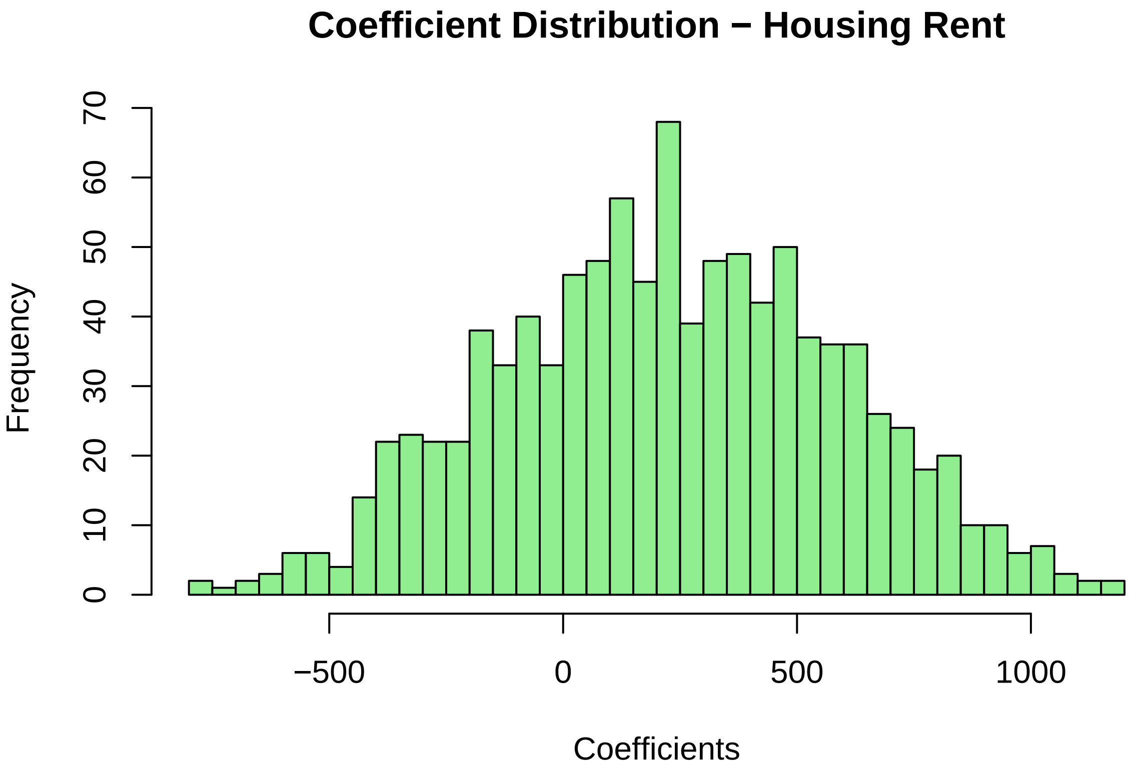


```
hist(my_df$ResidenceDuration,
     breaks=30,
     xlab = "Coefficients", main = "Coefficient Distribution - Residence Duration", col="lightblue")
```



Plot Interpretation:

```
hist(my_df$Housing.Rent,
     breaks=30,
     xlab = "Coefficients", main = "Coefficient Distribution - Housing Rent", col="lightgreen")
```



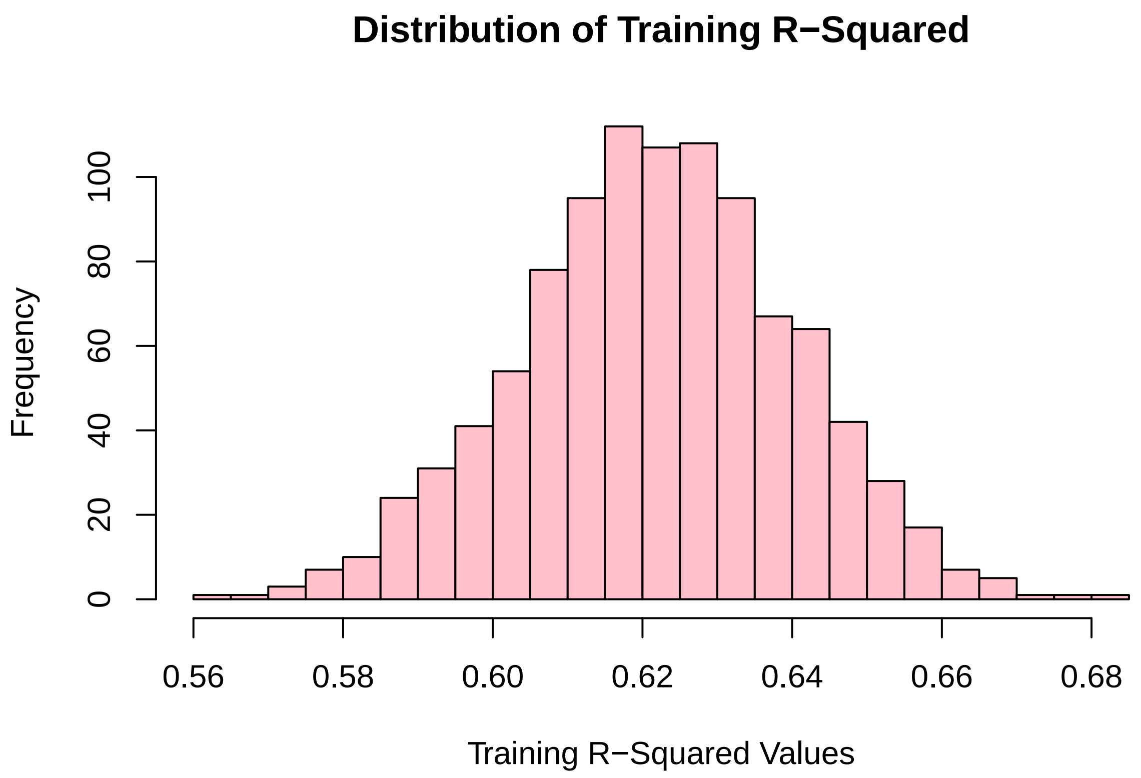
Summary - Interpretation of Above Plots ~ all plots printed above are fairly normally distributed. See individual interpretations here:

Age: the age coefficients are sometimes negatives which is the opposite of what I would normally think. Generally older people have higher credit, so seeing that the model sometimes says a higher age means a lesser loan is surprising. Overall the plot indicates age tends to have a positive impact on the loan amount, i.e., older people get higher loans.

Residence Duration: this coefficient value is mostly negative meaning the longer someone resides in a location, the lower their predicted loan Amount. This may be because wealthier people who can get higher loans tend to move around more and can afford to move to new places. This matches expectations.

Housing Rent: this coefficient value is mostly positive meaning the higher someone's rent the higher their predicted loan amount tends to be. This makes sense is people who can afford more expensive living places will generally be approved for higher loans due to high credit scores and proof of income.

```
hist(my_df$training_r2,
     breaks=30,
     xlab = "Training R-Squared Values", main = "Distribution of Training R-Squared", col="pink")
```



```
mean(my_df$training_r2)
```

```
## [1] 0.6219217
```

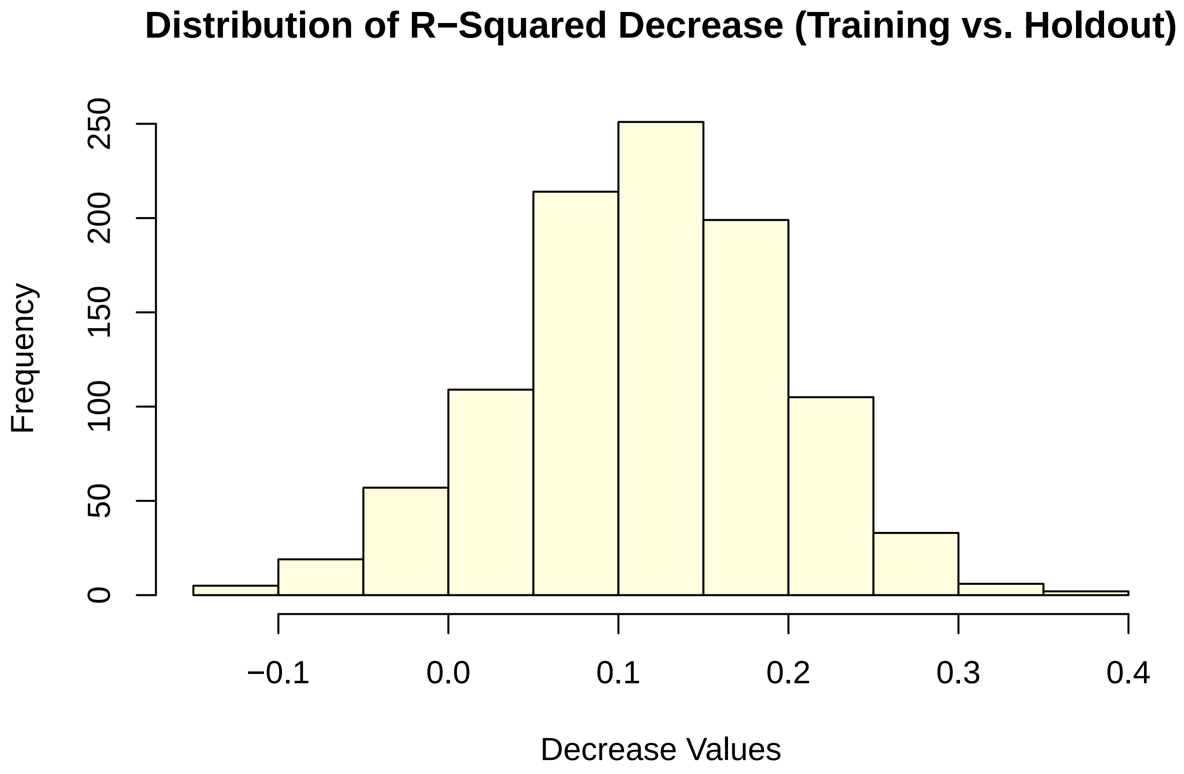
```
median(my_df$training_r2)
```

```
## [1] 0.6221699
```

```
my_df <- my_df %>% mutate(r2_decrease = (training_r2 - holdout_r2)/training_r2)
head(my_df$r2_decrease, 5)
```

```
## [1] 0.08578634 0.05006284 0.12261069 0.26808393 0.02836774
```

```
hist(my_df$r2_decrease,
     breaks=12,
     xlab = "Decrease Values", main = "Distribution of R-Squared Decrease (Training vs. Holdout)",
     col="lightyellow")
```



```
mean(my_df$r2_decrease)
```

```
## [1] 0.1170542
```

```
median(my_df$r2_decrease)
```

```
## [1] 0.1206984
```

Interpretation:

These plots are also normally distributed and indicate that on average we can expect around a 12% percentage decrease from our training dataset R-squared of 62% to our holdout dataset. This means the model performs slightly worse on the holdout data. This is to be expected because this is unseen data – we just want the model to be able to generalize well, not necessarily have an excellent R-squared value. This is a good result.

```
coefficient_means <- colMeans(my_df)
head(coefficient_means, 5)
```

```
##             (Intercept)                Duration InstallmentRatePercentage
##             6514.445005              126.992393               -782.770921
##       ResidenceDuration                     Age
##              -53.757369                5.662497
```

```r
coefficient_sds <- sapply(my_df, sd)
head(coefficient_sds, 5)
```

```
##               (Intercept)                  Duration InstallmentRatePercentage 
##               1288.098431                  5.387694                 47.668309 
##         ResidenceDuration                       Age 
##                 47.748450                  5.836145 
```

```r
# Difference between coefficient_means and actual model coefficients
bind <- data.frame(cbind(coefficient_means, save_coeff))
bind$abs_raw_diff <- abs(coefficient_means - save_coeff)
bind$percent_diff <- 100*abs(((coefficient_means - save_coeff) / save_coeff))

print(bind %>% na.omit())
```

```
##                                    coefficient_means    save_coeff abs_raw_diff
## (Intercept)                             6514.4450052  6787.3109335   272.865928
## Duration                                 126.9923926   121.7024620     5.289931
## InstallmentRatePercentage               -782.7709212  -803.3828621    20.611941
## ResidenceDuration                        -53.7573686    20.1154552    73.872824
## Age                                        5.6624968     1.4118155     4.250681
## NumberExistingCredits                     75.5933144     9.3048900    66.288424
## NumberPeopleMaintenance                 -207.9595145  -294.5123347    86.552820
## Telephone                               -478.1912394  -430.0143516    48.176888
## ForeignWorker                           -236.7257922  -174.4323728    62.293419
## CheckingAccountStatus.lt.0              -103.2913906    -0.3607945   102.930596
## CheckingAccountStatus.0.to.200           205.8951983   279.2494699    73.354272
## CheckingAccountStatus.gt.200            -616.6448860  -573.6950070    42.949879
## CreditHistory.NoCredit.AllPaid           849.4984033   508.5678641   340.930539
## CreditHistory.ThisBank.AllPaid           -47.9907468  -459.7257283   411.734981
## CreditHistory.PaidDuly                   -15.6668989   -61.1902427    45.523344
## CreditHistory.Delay                      113.9620829   138.9943540    25.032271
## Purpose.NewCar                         -1768.1136034 -2180.1076255   411.994022
## Purpose.UsedCar                        -1131.1941473 -1440.8553893   309.661242
## Purpose.Furniture.Equipment            -1851.4854148 -2295.6493157   444.163901
## Purpose.Radio.Television               -2085.7345008 -2454.9706733   369.236173
## Purpose.DomesticAppliance              -2481.7388581 -2559.0089901    77.270132
## Purpose.Repairs                        -1710.5724217 -1977.8988925   267.326471
## Purpose.Education                      -1907.1014641 -2429.6229301   522.521466
## Purpose.Business                       -1992.6396355 -2597.8420276   605.202392
## SavingsAccountBonds.lt.100              -329.4394441  -334.6707759     5.231332
## SavingsAccountBonds.100.to.500          -564.8601422  -201.3723119   363.487830
## SavingsAccountBonds.500.to.1000         -659.1950634  -769.0039648   109.808901
## SavingsAccountBonds.gt.1000             -386.4586074  -108.6142881   277.844319
## EmploymentDuration.lt.1                  114.2590421  -293.0847250   407.343767
## EmploymentDuration.1.to.4                 48.4814756   -89.4589353   137.940411
## EmploymentDuration.4.to.7                109.5993845  -200.9872948   310.586679
## EmploymentDuration.gt.7                 -159.3121487  -510.3288707   351.016722
## Personal.Male.Divorced.Seperated         480.5116417   503.3091425    22.797501
## Personal.Female.NotSingle                278.5919279   124.8996358   153.692292
## Personal.Male.Single                     735.2182940   892.2249404   157.006646
## OtherDebtorsGuarantors.None              129.5228933   133.9857097     4.462816
## OtherDebtorsGuarantors.CoApplicant       736.4230085   677.5142604    58.908748
## Property.RealEstate                     -837.7940304  -900.2620942    62.468064
## Property.Insurance                      -590.7038970  -511.6185864    79.085311
## Property.CarOther                       -566.1199297  -605.4682045    39.348275
## OtherInstallmentPlans.Bank              -150.7777877    31.0931649   181.870953
## OtherInstallmentPlans.Stores             -44.9924066   582.1274420   627.119849
## Housing.Rent                             227.0673423   822.5800230   595.512681
## Housing.Own                              108.5780856   583.4525437   474.874458
## Job.UnemployedUnskilled                -1688.5415978 -1857.0267610   168.485163
## Job.UnskilledResident                  -1192.5282121 -1174.7391001    17.789112
## Job.SkilledEmployee                    -1246.9380397 -1141.4595832   105.478456
## training_r2                                0.6219217  6787.3109335  6786.689012
## holdout_r2                                 0.5477035   121.7024620   121.154758
## r2_decrease                                0.1170542  -803.3828621   803.499916
##                                    percent_diff
## (Intercept)                            4.020236
## Duration                               4.346609
## InstallmentRatePercentage              2.565644
## ResidenceDuration                    367.244107
## Age                                  301.079107
## NumberExistingCredits                712.404169
## NumberPeopleMaintenance               29.388521
## Telephone                             11.203553
## ForeignWorker                         35.712075
## CheckingAccountStatus.lt.0         28528.869177
## CheckingAccountStatus.0.to.200        26.268366
## CheckingAccountStatus.gt.200           7.486535
## CreditHistory.NoCredit.AllPaid        67.037374
## CreditHistory.ThisBank.AllPaid        89.561005
## CreditHistory.PaidDuly                74.396410
## CreditHistory.Delay                   18.009560
## Purpose.NewCar                        18.897875
## Purpose.UsedCar                       21.491487
## Purpose.Furniture.Equipment           19.348073
## Purpose.Radio.Television              15.040350
## Purpose.DomesticAppliance              3.019533
## Purpose.Repairs                       13.515679
## Purpose.Education                     21.506278
## Purpose.Business                      23.296351
## SavingsAccountBonds.lt.100             1.563128
## SavingsAccountBonds.100.to.500       180.505367
## SavingsAccountBonds.500.to.1000       14.279367
## SavingsAccountBonds.gt.1000          255.808259
## EmploymentDuration.lt.1              138.984987
## EmploymentDuration.1.to.4            154.194112
## EmploymentDuration.4.to.7            154.530504
## EmploymentDuration.gt.7               68.782454
## Personal.Male.Divorced.Seperated       4.529522
## Personal.Female.NotSingle            123.052634
## Personal.Male.Single                  17.597204
## OtherDebtorsGuarantors.None            3.330815
## OtherDebtorsGuarantors.CoApplicant     8.694835
## Property.RealEstate                    6.938875
## Property.Insurance                    15.457865
## Property.CarOther                      6.498818
## OtherInstallmentPlans.Bank           584.922613
## OtherInstallmentPlans.Stores         107.728962
## Housing.Rent                          72.395714
## Housing.Own                           81.390417
## Job.UnemployedUnskilled                9.072845
## Job.UnskilledResident                  1.514303
## Job.SkilledEmployee                    9.240665
## training_r2                           99.990837
## holdout_r2                            99.549965
## r2_decrease                          100.014570
```

Calculate Confidence Intervals & Width

```
# Confidence Interval for Rep 1,000 Coefficients -------------------------------
# Transposed dataframe of confidence intervals for each variable
rep_conf <- data.frame(t(sapply(my_df[,1:63], function(x) Rmisc::CI(x, ci=0.975)))) #%>% na.omit()

# Calculate Width
# rep_conf$width <- (rep_conf$upper-rep_conf$lower)*sqrt(0.632)
rep_conf$width <- (rep_conf$upper-rep_conf$lower)*sqrt(0.632)

# MANUAL CHECK ==============================
# Calculate Means
#means <- data.frame(means=sapply(my_df[1:61], function(x) mean(x)))
#n <- 1000
# Calculate Standard Deviation
#std_dev <- data.frame(std_dev=sapply(my_df[1:61], function(x) sd(x)))

#std_error <- std_dev / sqrt(n)
#alpha = 0.025
#degrees_of_freedom = n-1
#t_score = qt(p=alpha/2, df=degrees_of_freedom,lower.tail=F)
#margin_error <- t_score * std_error

#lower_bound_new <- means - margin_error
#upper_bound_new <- means + margin_error

#x <- cbind(lower_bound_new, upper_bound_new) %>% dplyr::rename(lower_new = 1, upper_new = 2)

#margin <- qnorm(0.975)*std
#rep_CI_low_manual <- means - margin
#rep_CI_high_manual <- means + margin
#manual <- cbind(rep_CI_low_manual, rep_CI_high_manual) %>%
#                dplyr::rename(lower_manual = 1, upper_manual = 2)


# ==============================

# Reorder columns
rep_conf <- rep_conf %>% select(lower, upper, width)

# Keep only valid columns, remove R2 values
```

```
rep_conf <- rep_conf[1:61,]

# COMBINE for check
#z <- cbind(rep_conf, manual)
#z$width_manual <- z$upper_manual - z$lower_manual*sqrt(0.632)
#View(z)

# Check row count after ommitting NA
nrow(rep_conf)
```

```
## [1] 61
```

```
# Confidence Interval for Full Model ------------------------------------------
# Calculate confidence interval and rename the columns
full_model_conf <- data.frame(confint(full_model, level=0.975)) %>%
                   #na.omit() %>%
                   dplyr::rename(lower_full = 1, upper_full = 2)

# Calculate Width
full_model_conf <- full_model_conf %>%
                   #mutate(index=1:nrow(full_model_conf)) %>%
                   #filter(index %in% (2:48)) %>%
                   mutate(width_full=upper_full-lower_full)

# Check row count
nrow(full_model_conf)
```

```
## [1] 61
```

```
#t <- cbind(rep_conf, full_model_conf)

row.names(rep_conf) == row.names(full_model_conf)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [61] TRUE
```

```
#View(cbind(row.names(rep_conf), row.names(full_model_conf)))
```

Determine how many of the repeated sample CI’s are tighter

```
# Combine the 2 dataframes and remove NA values
calc <- cbind(rep_conf, full_model_conf) %>% na.omit()

#Calculate how many of the repeated sample CI’s are tighter or broader than the full model CI’s. If the

# 1 means the CI is tighter (width of repeated is smaller)
```

```
calc$tighter_CI_flag <- ifelse(calc$width < calc$width_full, 1, 0)

percent <- sum(calc$tighter_CI_flag)/nrow(calc)
print(percent*100)
```

```
## [1] 100
```

Conclusion: 100% of the simulated/repeated model's CI's are tighter/smaller. In other words, of the 47 columns left after removing NA's and the class column, 47 of the columns' confidence intervals from the repeated samples are smaller/tighter than those comapred to the CI's from the full model.