MSDS660_Week6_Discussion_Apeetz

2022-11-13

Adam Peetz

MSDS660 Week 6 Discussion

Regis University

Dr. Siripun Sanguansintukul

November 24th 2022

Linear Regression

Discussion Activity: For Discussion, continue working on the loans file and respond to the following questions. 1.Does annual income matter for the loan amount or its interest rate? 2. Pick another DV (not loan_amount) and make a hypothesis of variables that maybe related. You need to include at least 3 IVs in the analysis. 3. Run several MLR models. Be sure to consider if you need to add/remove or transform variables. 4. Perform tests of diagnostics, i.e. with plot(), correlation, and vif. 5. Are there variables currently not in the data set that may be beneficial to your analysis? Does your initial hypothesis hold? 5. Post your rfile and responses to the questions to the Week 6 discussion.

```
# load libraries
library(tidyverse)
library(data.table)
library(dplyr)
library(car)
library(corrplot)
library(MASS)
library('fastDummies')

#heatmap and custom colors
#install.packages("reshape2")
library(reshape2)
#install.packages("viridis")
library("viridis")

# load data
data <- read_csv("loans_full_schema.csv",show_col_types = FALSE)
# convert data to table
df<-as.data.table(data)
```

1.Does annual income matter for the loan amount or its interest rate?

F-Statistic Tests

Ho: Where p > 0.05 There is no relationship between ANY of the independent variables and Y

Ha: Where p < 0.05 AT LEAST 1 independent variable is related to Y

Annual Income and Loan Amount

A F-Statistic of < 0.05 indicates that the null hypothesis is true. Loan amount is related to annual income.

The slope of the line is 0.05 with an intercept of 12,268.

Every 1,000 increase in income increases loan amount by 51 dollars.

Annual income and interest rate

A F-Statistic of < 0.05 indicates that the null hypothesis is true. Loan amount is related to interest rate.

The slope of the line is -0.000007 with an intercept of 13.

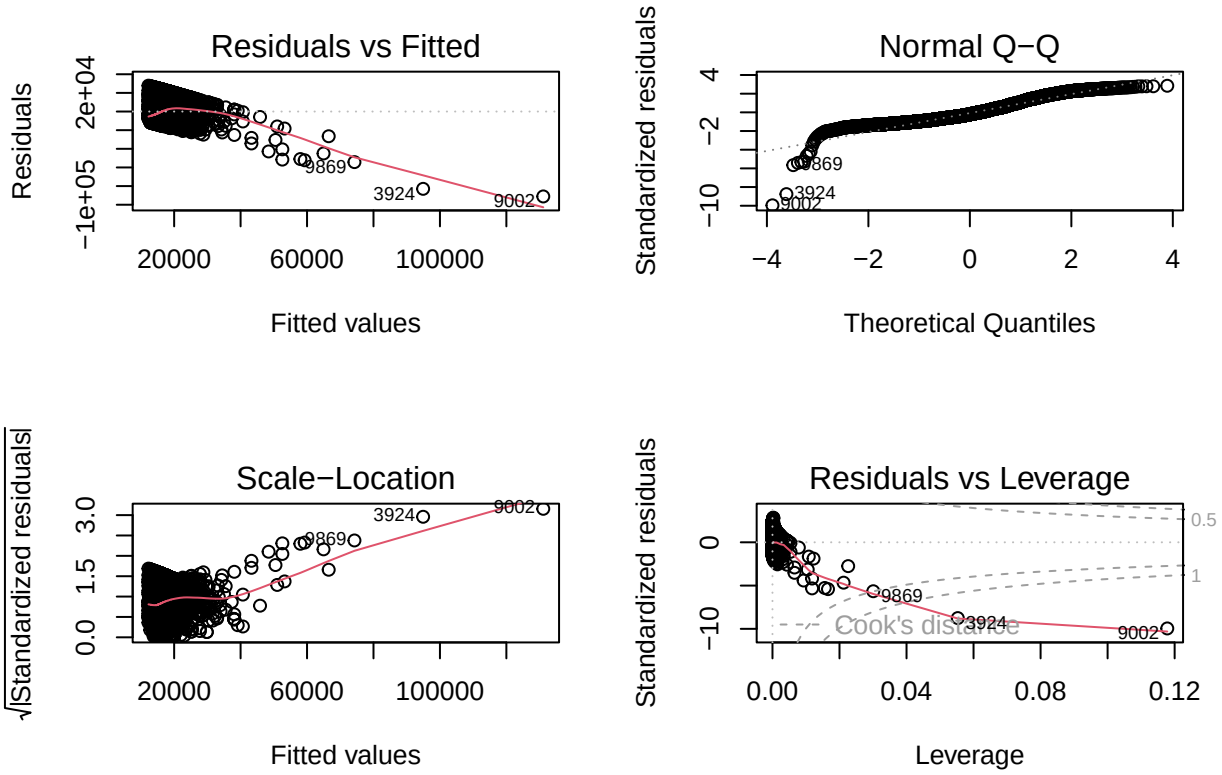
Every 1,000 increase in income decreases interest rate by 0.007 percent.

```
# fit model
slm <- lm(loan_amount ~ df$annual_income, data = df)

# show model summary
summary(slm)
```

```
##
## Call:
## lm(formula = loan_amount ~ df$annual_income, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91105  -7418  -1970   5911  27731
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.227e+04  1.540e+02   79.66   <2e-16 ***
## df$annual_income 5.167e-02  1.505e-03   34.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9744 on 9998 degrees of freedom
## Multiple R-squared:  0.1054, Adjusted R-squared:  0.1053
## F-statistic:  1178 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(slm)
```

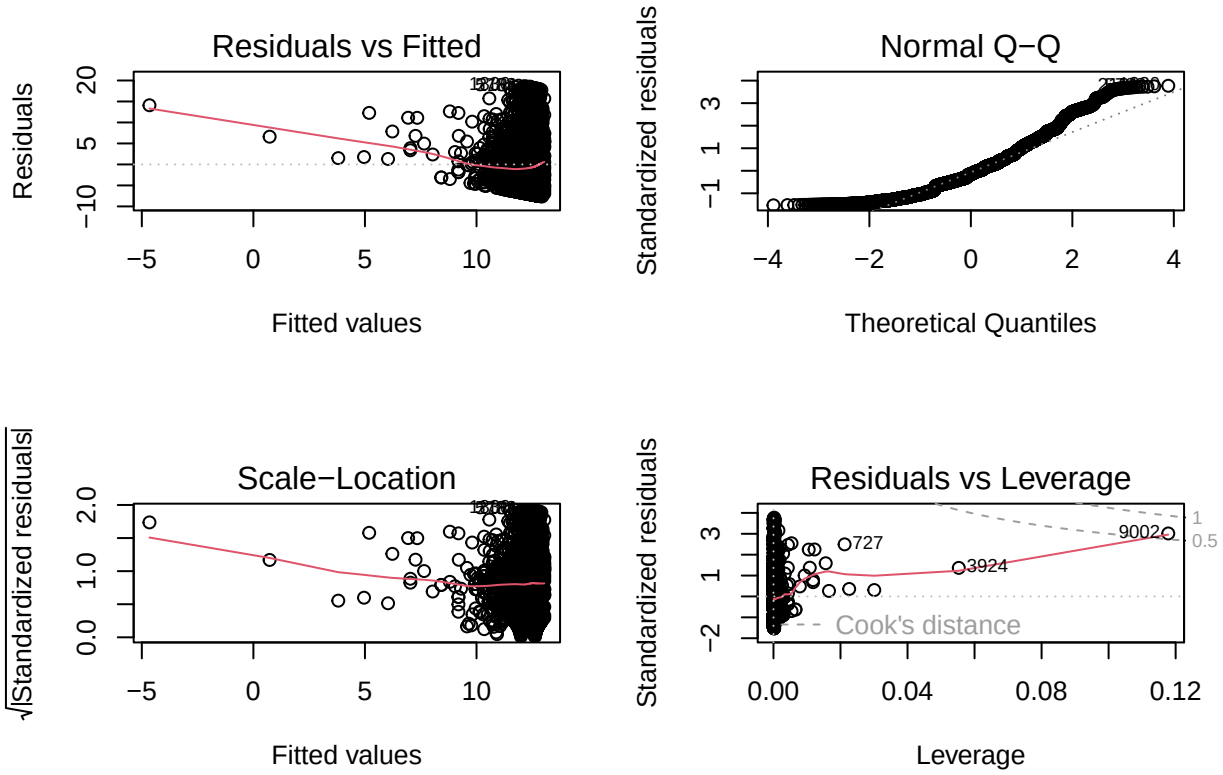



```
# Fit model
slm <- lm(interest_rate ~ df$annual_income, data = df)

#show model summary
summary(slm)
```

```
##
## Call:
## lm(formula = interest_rate ~ df$annual_income, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6247 -3.3277 -0.7034  2.6595 18.7670
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.304e+01  7.865e-02  165.76   <2e-16 ***
## df$annual_income -7.694e-06  7.688e-07  -10.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.976 on 9998 degrees of freedom
## Multiple R-squared:  0.009917,   Adjusted R-squared:  0.009818
## F-statistic: 100.1 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(slm)
```



2. Pick another DV (not loan_amount) and make a hypothesis of variables that maybe related. You need to include at least 3 IVs in the analysis.

Hypothesis Total_credit_limit is related to total_credit_lines, accounts_opened_24m, and num_mort_accounts. A prerson with more credit lines, recently opened accounts, and mortgages will have higher total credit.

Results

A F-Statistic of < 0.05 indicates that the null hypothesis is true. These variables are related to Total_credit_limit.

Number of mortgage accounts has a big impact on total_credit_limit, every additional mortgage account increases credit limit by 54804.

```
# select data for model
df_1 <- df %>% dplyr::select(total_credit_limit, total_credit_lines, accounts_opened_24m, num_mort_accou

# fit model
fit_1 <- lm(total_credit_limit ~ ., data = df_1)

# show model summary
summary(fit_1)
```

```
##
```

```
## Call:
## lm(formula = total_credit_limit ~ ., data = df_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -657874  -64652  -27514   36623 3220904
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           31242.0     3345.3   9.339  < 2e-16 ***
## total_credit_lines     3038.8      150.0  20.252  < 2e-16 ***
## accounts_opened_24m    1747.0      529.9   3.297 0.000981 ***
## num_mort_accounts     54804.3      933.6  58.700  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149300 on 9996 degrees of freedom
## Multiple R-squared:  0.3671, Adjusted R-squared:  0.3669
## F-statistic:  1933 on 3 and 9996 DF,  p-value: < 2.2e-16
```

```
AIC(fit_1)
```

```
## [1] 266657.7
```

3. Run several MLR models. Be sure to consider if you need to add/remove or transform variables.

Adding Variables

Additional variables may be related to total_credit_limit. In addition to the original variables, application_type, interest_rate, open_credit_lines, total_debit_limit, number_open_cc_accounts, and grade will be added to the model. Application_type and grade both contain category labels instead of numerical values. These two features are transformed into sparse numerical matrices using one-hot-encoding.

These additional features improve the AIC of the model from 266657 to 264793.

```
# Select additional variables
df_2 <- df %>% dplyr::select(total_credit_limit, total_credit_lines, accounts_opened_24m, num_mort_accou

# One hot encoding categorical variables
dum_df_2 <- dummy_cols(df_2,
                       select_columns=c('grade','application_type'),
                       remove_selected_columns = TRUE)

# fit model
fit_2 <- lm(total_credit_limit ~ ., data = dum_df_2)

# show model statistics
summary(fit_2)
```

```
##
## Call:
## lm(formula = total_credit_limit ~ ., data = dum_df_2)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -593806  -60736  -17724   29705 3166826
##
## Coefficients: (2 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.842e+04  5.494e+04  -0.335    0.737
## total_credit_lines            4.182e+01  1.902e+02   0.220    0.826
## accounts_opened_24m           3.793e+03  5.224e+02   7.260 4.14e-13 ***
## num_mort_accounts             5.037e+04  8.910e+02  56.531  < 2e-16 ***
## interest_rate                 4.138e+02  1.237e+03   0.335    0.738
## open_credit_lines             1.394e+04  5.443e+02  25.611  < 2e-16 ***
## total_debit_limit             2.199e+00  6.346e-02  34.646  < 2e-16 ***
## num_open_cc_accounts         -1.579e+04  5.669e+02 -27.859  < 2e-16 ***
## grade_A                       3.926e+04  4.927e+04   0.797    0.426
## grade_B                       4.778e+04  4.664e+04   1.025    0.306
## grade_C                       3.568e+04  4.440e+04   0.804    0.422
## grade_D                       3.786e+04  4.200e+04   0.901    0.367
## grade_E                       2.666e+04  4.058e+04   0.657    0.511
## grade_F                       3.508e+04  4.316e+04   0.813    0.416
## grade_G                              NA         NA      NA       NA
## application_type_individual -2.565e+04  3.856e+03  -6.652 3.05e-11 ***
## application_type_joint               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 135900 on 9985 degrees of freedom
## Multiple R-squared:  0.4759, Adjusted R-squared:  0.4752
## F-statistic: 647.7 on 14 and 9985 DF,  p-value: < 2.2e-16
```

```
AIC(fit_2)
```

```
## [1] 264793.2
```

Removing Outliers

There are several outliers in the total_credit_limit column. These will be removed to help normalize the distribution of the feature which will increase model performance.

Removing outliers further improves model performance from 264793 to 249100.

```
# remove outliers
Q <- quantile(dum_df_2$total_credit_limit, probs=c(.25, .75), na.rm = TRUE)
iqr <- IQR(dum_df_2$total_credit_limit, na.rm = TRUE)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
clean_dum_df_2<- subset(dum_df_2, dum_df_2$total_credit_limit > (Q[1] - 1.5*iqr) & dum_df_2$total_credi

# fit model
fit_3 <- lm(total_credit_limit ~ ., data = clean_dum_df_2)

# show model charts and results
summary(fit_3)
```

```
##
## Call:
## lm(formula = total_credit_limit ~ ., data = clean_dum_df_2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -485998  -49583  -19558   35331  447578
##
## Coefficients: (2 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.815e+04  3.979e+04   0.959    0.338
## total_credit_lines            1.340e+02  1.398e+02   0.958    0.338
## accounts_opened_24m           2.392e+03  3.833e+02   6.241 4.53e-10 ***
## num_mort_accounts             4.308e+04  6.744e+02  63.886  < 2e-16 ***
## interest_rate                -3.802e+02  9.029e+02  -0.421    0.674
## open_credit_lines             1.150e+04  3.993e+02  28.800  < 2e-16 ***
## total_debit_limit             1.454e+00  5.089e-02  28.562  < 2e-16 ***
## num_open_cc_accounts         -1.152e+04  4.180e+02 -27.554  < 2e-16 ***
## grade_A                       6.008e+03  3.559e+04   0.169    0.866
## grade_B                       9.692e+03  3.365e+04   0.288    0.773
## grade_C                       5.773e+03  3.200e+04   0.180    0.857
## grade_D                       1.032e+04  3.022e+04   0.341    0.733
## grade_E                       6.430e+03  2.916e+04   0.221    0.825
## grade_F                       2.137e+04  3.108e+04   0.688    0.492
## grade_G                              NA         NA      NA       NA
## application_type_individual -2.814e+04  2.824e+03  -9.967  < 2e-16 ***
## application_type_joint               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97570 on 9634 degrees of freedom
## Multiple R-squared:  0.5099, Adjusted R-squared:  0.5092
## F-statistic:   716 on 14 and 9634 DF,  p-value: < 2.2e-16
```

```
AIC(fit_3)
```

```
## [1] 249100.8
```

Stepwise AIC

A few of the variables are considered statistically insignificant for the prediction of total_credit_limit and could be removed from the model. The best combination of features can be programmatically generated using a stepwise selection process. The stepAIC function recommends a combination of accounts_opened_24m + num_mort_accounts + open_credit_lines + total_debit_limit + num_open_cc_accounts + grade_B + application_type_individual for prediction of total_credit_limit.

The model with features selected by stepwise AIC improves from 249100 to 249090.

```
# Show step AIC
stepAIC(fit_3, direction="both")
```

```
## Start:  AIC=221716.1
## total_credit_limit ~ total_credit_lines + accounts_opened_24m +
##     num_mort_accounts + interest_rate + open_credit_lines + total_debit_limit +
##     num_open_cc_accounts + grade_A + grade_B + grade_C + grade_D +
##     grade_E + grade_F + grade_G + application_type_individual +
##     application_type_joint
##
##
```

```
## Step:  AIC=221716.1
## total_credit_limit ~ total_credit_lines + accounts_opened_24m +
##     num_mort_accounts + interest_rate + open_credit_lines + total_debit_limit +
##     num_open_cc_accounts + grade_A + grade_B + grade_C + grade_D +
##     grade_E + grade_F + grade_G + application_type_individual
##
##
## Step:  AIC=221716.1
## total_credit_limit ~ total_credit_lines + accounts_opened_24m +
##     num_mort_accounts + interest_rate + open_credit_lines + total_debit_limit +
##     num_open_cc_accounts + grade_A + grade_B + grade_C + grade_D +
##     grade_E + grade_F + application_type_individual
##
##                               Df  Sum of Sq        RSS    AIC
## - grade_A                      1 2.7132e+08 9.1709e+13 221714
## - grade_C                      1 3.0993e+08 9.1710e+13 221714
## - grade_E                      1 4.6289e+08 9.1710e+13 221714
## - grade_B                      1 7.8980e+08 9.1710e+13 221714
## - grade_D                      1 1.1097e+09 9.1710e+13 221714
## - interest_rate                1 1.6882e+09 9.1711e+13 221714
## - grade_F                      1 4.5026e+09 9.1714e+13 221715
## - total_credit_lines           1 8.7435e+09 9.1718e+13 221715
## <none>                                      9.1709e+13 221716
## - accounts_opened_24m          1 3.7078e+11 9.2080e+13 221753
## - application_type_individual  1 9.4566e+11 9.2655e+13 221813
## - num_open_cc_accounts         1 7.2275e+12 9.8937e+13 222446
## - total_debit_limit            1 7.7658e+12 9.9475e+13 222498
## - open_credit_lines            1 7.8958e+12 9.9605e+13 222511
## - num_mort_accounts            1 3.8853e+13 1.3056e+14 225122
##
## Step:  AIC=221714.1
## total_credit_limit ~ total_credit_lines + accounts_opened_24m +
##     num_mort_accounts + interest_rate + open_credit_lines + total_debit_limit +
##     num_open_cc_accounts + grade_B + grade_C + grade_D + grade_E +
##     grade_F + application_type_individual
##
##                               Df  Sum of Sq        RSS    AIC
## - grade_C                      1 5.8855e+07 9.1710e+13 221712
## - grade_E                      1 2.1431e+08 9.1710e+13 221712
## - grade_D                      1 3.2790e+09 9.1713e+13 221712
## - interest_rate                1 4.1439e+09 9.1714e+13 221713
## - grade_F                      1 6.7327e+09 9.1716e+13 221713
## - total_credit_lines           1 8.7341e+09 9.1718e+13 221713
## - grade_B                      1 1.0970e+10 9.1720e+13 221713
## <none>                                      9.1709e+13 221714
## + grade_A                      1 2.7132e+08 9.1709e+13 221716
## + grade_G                      1 2.7132e+08 9.1709e+13 221716
## - accounts_opened_24m          1 3.7159e+11 9.2081e+13 221751
## - application_type_individual  1 9.4576e+11 9.2655e+13 221811
## - num_open_cc_accounts         1 7.2281e+12 9.8938e+13 222444
## - total_debit_limit            1 7.7764e+12 9.9486e+13 222497
## - open_credit_lines            1 7.8960e+12 9.9605e+13 222509
## - num_mort_accounts            1 3.8864e+13 1.3057e+14 225121
##
```

```
## Step:  AIC=221712.1
## total_credit_limit ~ total_credit_lines + accounts_opened_24m +
##     num_mort_accounts + interest_rate + open_credit_lines + total_debit_limit +
##     num_open_cc_accounts + grade_B + grade_D + grade_E + grade_F +
##     application_type_individual
##
##                               Df  Sum of Sq        RSS    AIC
## - grade_E                      1 2.4027e+08 9.1710e+13 221710
## - total_credit_lines           1 8.7369e+09 9.1718e+13 221711
## - grade_F                      1 1.1863e+10 9.1721e+13 221711
## - grade_D                      1 1.3175e+10 9.1723e+13 221712
## - interest_rate                1 1.3810e+10 9.1723e+13 221712
## <none>                                      9.1710e+13 221712
## - grade_B                      1 2.6384e+10 9.1736e+13 221713
## + grade_G                      1 3.2011e+08 9.1709e+13 221714
## + grade_C                      1 5.8855e+07 9.1709e+13 221714
## + grade_A                      1 2.0238e+07 9.1710e+13 221714
## - accounts_opened_24m          1 3.7156e+11 9.2081e+13 221749
## - application_type_individual  1 9.4584e+11 9.2655e+13 221809
## - num_open_cc_accounts         1 7.2307e+12 9.8940e+13 222442
## - total_debit_limit            1 7.7985e+12 9.9508e+13 222498
## - open_credit_lines            1 7.8966e+12 9.9606e+13 222507
## - num_mort_accounts            1 3.8881e+13 1.3059e+14 225121
##
## Step:  AIC=221710.2
## total_credit_limit ~ total_credit_lines + accounts_opened_24m +
##     num_mort_accounts + interest_rate + open_credit_lines + total_debit_limit +
##     num_open_cc_accounts + grade_B + grade_D + grade_F + application_type_individual
##
##                               Df  Sum of Sq        RSS    AIC
## - total_credit_lines           1 8.8027e+09 9.1719e+13 221709
## - grade_F                      1 1.1992e+10 9.1722e+13 221709
## - grade_D                      1 1.5679e+10 9.1725e+13 221710
## <none>                                      9.1710e+13 221710
## - interest_rate                1 2.0299e+10 9.1730e+13 221710
## - grade_B                      1 2.6145e+10 9.1736e+13 221711
## + grade_G                      1 4.1450e+08 9.1709e+13 221712
## + grade_E                      1 2.4027e+08 9.1710e+13 221712
## + grade_C                      1 8.4816e+07 9.1710e+13 221712
## + grade_A                      1 5.7208e+07 9.1710e+13 221712
## - accounts_opened_24m          1 3.7133e+11 9.2081e+13 221747
## - application_type_individual  1 9.4648e+11 9.2656e+13 221807
## - num_open_cc_accounts         1 7.2336e+12 9.8943e+13 222441
## - total_debit_limit            1 7.8936e+12 9.9603e+13 222505
## - open_credit_lines            1 7.8970e+12 9.9607e+13 222505
## - num_mort_accounts            1 3.8898e+13 1.3061e+14 225120
##
## Step:  AIC=221709.1
## total_credit_limit ~ accounts_opened_24m + num_mort_accounts +
##     interest_rate + open_credit_lines + total_debit_limit + num_open_cc_accounts +
##     grade_B + grade_D + grade_F + application_type_individual
##
##                               Df  Sum of Sq        RSS    AIC
## - grade_F                      1 1.2098e+10 9.1731e+13 221708
```

```
## - grade_D                       1 1.5804e+10 9.1734e+13 221709
## <none>                                       9.1719e+13 221709
## - interest_rate                 1 2.0632e+10 9.1739e+13 221709
## - grade_B                       1 2.6611e+10 9.1745e+13 221710
## + total_credit_lines            1 8.8027e+09 9.1710e+13 221710
## + grade_G                       1 4.2375e+08 9.1718e+13 221711
## + grade_E                       1 3.0608e+08 9.1718e+13 221711
## + grade_C                       1 1.1700e+08 9.1718e+13 221711
## + grade_A                       1 7.7218e+07 9.1718e+13 221711
## - accounts_opened_24m           1 3.8961e+11 9.2108e+13 221748
## - application_type_individual   1 9.4373e+11 9.2662e+13 221806
## - num_open_cc_accounts          1 7.3182e+12 9.9037e+13 222448
## - total_debit_limit             1 7.9128e+12 9.9631e+13 222506
## - open_credit_lines             1 1.1395e+13 1.0311e+14 222837
## - num_mort_accounts             1 4.4818e+13 1.3654e+14 225546
##
## Step:  AIC=221708.4
## total_credit_limit ~ accounts_opened_24m + num_mort_accounts +
##     interest_rate + open_credit_lines + total_debit_limit + num_open_cc_accounts +
##     grade_B + grade_D + application_type_individual
##
##                                Df  Sum of Sq        RSS    AIC
## - grade_D                       1 1.0926e+10 9.1742e+13 221708
## - interest_rate                 1 1.2972e+10 9.1744e+13 221708
## <none>                                       9.1731e+13 221708
## + grade_F                       1 1.2098e+10 9.1719e+13 221709
## - grade_B                       1 2.6168e+10 9.1757e+13 221709
## + total_credit_lines            1 8.9092e+09 9.1722e+13 221709
## + grade_A                       1 1.6209e+09 9.1729e+13 221710
## + grade_C                       1 1.5862e+09 9.1729e+13 221710
## + grade_G                       1 7.6669e+08 9.1730e+13 221710
## + grade_E                       1 3.0449e+08 9.1730e+13 221710
## - accounts_opened_24m           1 3.8898e+11 9.2120e+13 221747
## - application_type_individual   1 9.4473e+11 9.2675e+13 221805
## - num_open_cc_accounts          1 7.3268e+12 9.9057e+13 222448
## - total_debit_limit             1 7.9715e+12 9.9702e+13 222510
## - open_credit_lines             1 1.1397e+13 1.0313e+14 222836
## - num_mort_accounts             1 4.4889e+13 1.3662e+14 225550
##
## Step:  AIC=221707.5
## total_credit_limit ~ accounts_opened_24m + num_mort_accounts +
##     interest_rate + open_credit_lines + total_debit_limit + num_open_cc_accounts +
##     grade_B + application_type_individual
##
##                                Df  Sum of Sq        RSS    AIC
## - interest_rate                 1 5.1353e+09 9.1747e+13 221706
## <none>                                       9.1742e+13 221708
## - grade_B                       1 2.1530e+10 9.1763e+13 221708
## + grade_D                       1 1.0926e+10 9.1731e+13 221708
## + total_credit_lines            1 8.9941e+09 9.1733e+13 221709
## + grade_C                       1 8.2723e+09 9.1733e+13 221709
## + grade_F                       1 7.2203e+09 9.1734e+13 221709
## + grade_E                       1 4.1448e+09 9.1737e+13 221709
## + grade_A                       1 2.6050e+09 9.1739e+13 221709
```

```
## + grade_G                      1 1.4705e+09 9.1740e+13 221709
## - accounts_opened_24m          1 3.8897e+11 9.2131e+13 221746
## - application_type_individual  1 9.4443e+11 9.2686e+13 221804
## - num_open_cc_accounts         1 7.3249e+12 9.9067e+13 222447
## - total_debit_limit            1 7.9764e+12 9.9718e+13 222510
## - open_credit_lines            1 1.1393e+13 1.0313e+14 222835
## - num_mort_accounts            1 4.4883e+13 1.3662e+14 225548
##
## Step:  AIC=221706
## total_credit_limit ~ accounts_opened_24m + num_mort_accounts +
##     open_credit_lines + total_debit_limit + num_open_cc_accounts +
##     grade_B + application_type_individual
##
##                               Df  Sum of Sq        RSS    AIC
## <none>                                     9.1747e+13 221706
## + grade_C                      1 9.2048e+09 9.1738e+13 221707
## + total_credit_lines           1 9.1353e+09 9.1738e+13 221707
## - grade_B                      1 2.9733e+10 9.1776e+13 221707
## + grade_E                      1 8.1196e+09 9.1739e+13 221707
## + grade_A                      1 7.7388e+09 9.1739e+13 221707
## + interest_rate                1 5.1353e+09 9.1742e+13 221708
## + grade_F                      1 4.0425e+09 9.1743e+13 221708
## + grade_D                      1 3.0890e+09 9.1744e+13 221708
## + grade_G                      1 2.2789e+09 9.1744e+13 221708
## - accounts_opened_24m          1 3.8385e+11 9.2131e+13 221744
## - application_type_individual  1 9.3931e+11 9.2686e+13 221802
## - num_open_cc_accounts         1 7.3305e+12 9.9077e+13 222446
## - total_debit_limit            1 8.6345e+12 1.0038e+14 222572
## - open_credit_lines            1 1.1388e+13 1.0313e+14 222833
## - num_mort_accounts            1 4.5481e+13 1.3723e+14 225589
##
## Call:
## lm(formula = total_credit_limit ~ accounts_opened_24m + num_mort_accounts +
##     open_credit_lines + total_debit_limit + num_open_cc_accounts +
##     grade_B + application_type_individual, data = clean_dum_df_2)
##
## Coefficients:
##                  (Intercept)          accounts_opened_24m
##                    40223.478                     2398.269
##            num_mort_accounts            open_credit_lines
##                    43359.765                    11696.508
##            total_debit_limit         num_open_cc_accounts
##                        1.468                   -11556.767
##                      grade_B  application_type_individual
##                     3828.900                   -27958.094
```

```
# fit model
fit_4 <- lm(formula = total_credit_limit ~ accounts_opened_24m + num_mort_accounts +
            open_credit_lines + total_debit_limit + num_open_cc_accounts +
            grade_B + application_type_individual, data = clean_dum_df_2)

# model summary
summary(fit_4)
```

```
##
## Call:
## lm(formula = total_credit_limit ~ accounts_opened_24m + num_mort_accounts +
##     open_credit_lines + total_debit_limit + num_open_cc_accounts +
##     grade_B + application_type_individual, data = clean_dum_df_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -487639  -49605  -19610   35290  442119
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.022e+04  3.371e+03  11.932  < 2e-16 ***
## accounts_opened_24m          2.398e+03  3.776e+02   6.351 2.23e-10 ***
## num_mort_accounts            4.336e+04  6.272e+02  69.133  < 2e-16 ***
## open_credit_lines            1.170e+04  3.381e+02  34.593  < 2e-16 ***
## total_debit_limit            1.468e+00  4.874e-02  30.122  < 2e-16 ***
## num_open_cc_accounts        -1.156e+04  4.164e+02 -27.754  < 2e-16 ***
## grade_B                      3.829e+03  2.166e+03   1.768   0.0772 .
## application_type_individual -2.796e+04  2.814e+03  -9.935  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97550 on 9641 degrees of freedom
## Multiple R-squared:  0.5097, Adjusted R-squared:  0.5094
## F-statistic:  1432 on 7 and 9641 DF,  p-value: < 2.2e-16
```

```
AIC(fit_4)
```

```
## [1] 249090.7
```

Best model by AIC

Fit_4, the model with outliers removed and additional features selected by stepwise AIC, is the best performing model.

```
# score all models
# mode
AIC(fit_1)
```

```
## [1] 266657.7
```

```
AIC(fit_2)
```

```
## [1] 264793.2
```

```
AIC(fit_3)
```

```
## [1] 249100.8
```

```
AIC(fit_4)
```

```
## [1] 249090.7
```

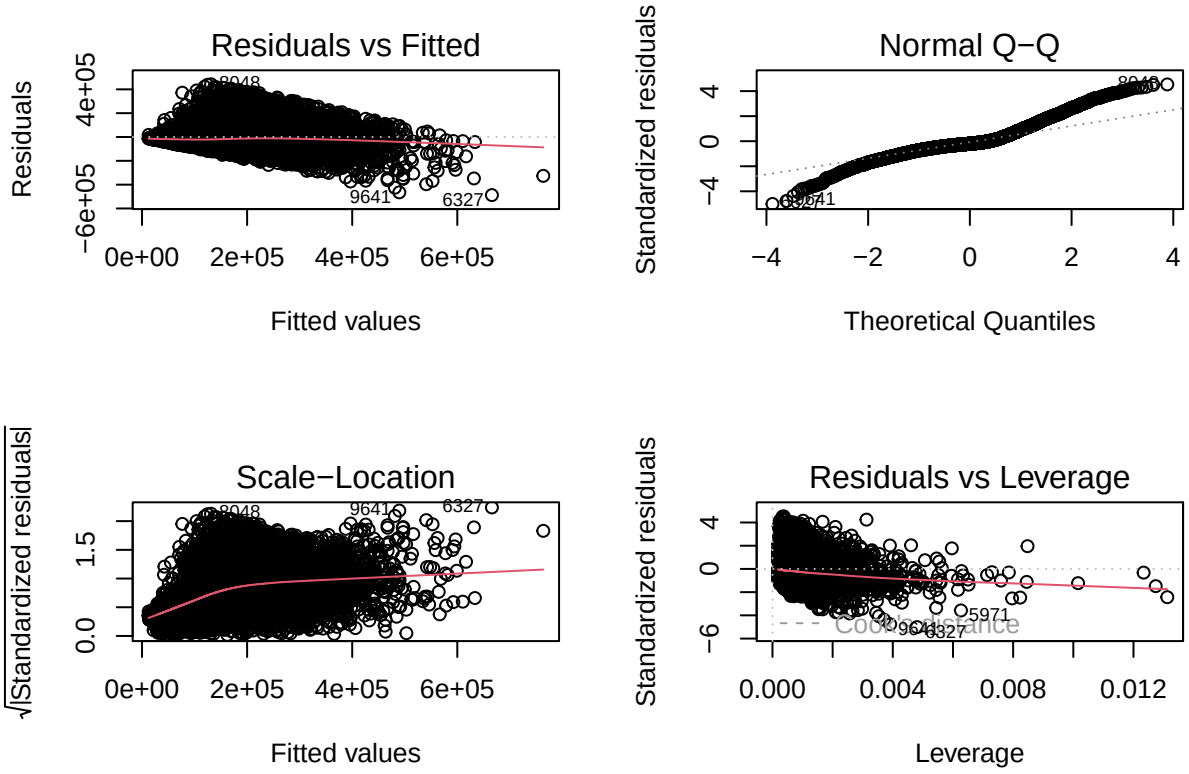
4. Perform tests of diagnostics, i.e. with plot(), correlation, and vif.

plot

Residual plots indicate some departures from homogeneous variance in the dataset.

A Q-Q plot shows the data set is not normally distributed. Additional corrections could be made to features with outliers to restore normality to the residual and Q-Q plots.

```
par(mfrow=c(2,2))
plot(fit_4)
```



Correlation

Checking for multicollinearity issues with a heatmap shows open_credit_lines is correlated to num_open_cc_accounts. A correlation score that exceed 0.8 indicates that these two features are essentially redundant. Open_credit_lines shows higher correlation to the prediction variable so it should be more impactful for the prediction of total_credit_limit than num_open_cc_accounts. Num_open_cc_accounts should be dropped from the model.

```
#define lower triangle function
get_lower_tri<-function(cormat){
    cormat[lower.tri(cormat)] <- NA
    return(cormat)}

#define upper triangle function
get_upper_tri <- function(cormat){
    cormat[upper.tri(cormat)]<- NA
```

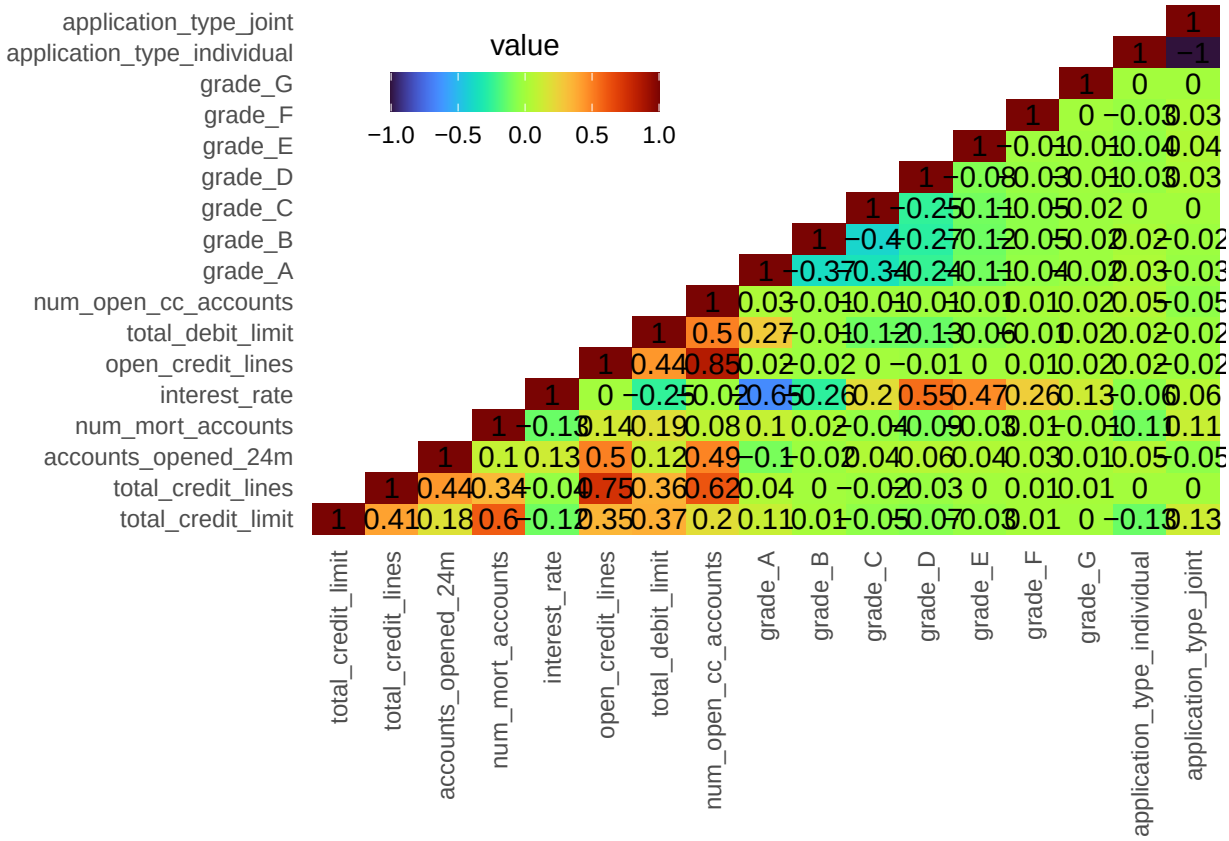
```
    return(cormat)}

#translate dataframe to correlation dataframe
cormap <- round(cor(clean_dum_df_2),2)

#get lower triangle
tri <- get_lower_tri(cormap)

#melt the corrleation dataframe
melted_cormap <- melt(tri, na.rm=TRUE)

#apply gg plotting function
ggplot(data = melted_cormap, aes(x=Var2, y=Var1, fill=value)) +
        geom_tile() +
        geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
        scale_fill_viridis(discrete = FALSE, option="H") +
        theme(
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank(),
        legend.justification = c(1, 0),
        legend.position = c(0.4, 0.7),
        legend.direction = "horizontal")+
        guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                      title.position = "top", title.hjust = 0.5))
```



VIF

Variance-Inflation-Factor (VIF) is another check for multicollinearity issues. The best possible VIF score is 1, which indicates the absence of multicollinearity. VIF values that exceed 5 are problematic and should be addressed in the model. All features of the fit_4 model have VIF scores below 5. Open_credit_lines and num_open_cc_accounts do have higher VIF scores which matches the results of the correlation plot that suggested these variables have multicollinearity issues.

```
vif(fit_4)
```

```
##          accounts_opened_24m                  num_mort_accounts
##                     1.424386                           1.072515
##            open_credit_lines                  total_debit_limit
##                     3.887283                           1.414333
##          num_open_cc_accounts                            grade_B
##                     4.155838                           1.001921
## application_type_individual
##                     1.018306
```

5. Are there variables currently not in the data set that may be beneficial to your analysis? Does your initial hypothesis hold?

There are a significant number of variables that impact total_credit_limit in the dataset. Additional variables are not needed to analyze total_credit_limit.

The three variables initially chosen remained in the model after reduction by stepwise AIC and through

collinearity checks. The initial hypothesis holds. Total_credit_limit is related to total_credit_lines, accounts_opened_24m, and num_mort_accounts. A person with more credit lines, recently opened accounts, and mortgages will have higher total credit.