Week_6_NSDC_project

Lea Choe, Elias Grant

2025-05-14

```r
# Load necessary libraries
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2

## Warning: package 'ggplot2' was built under R version 4.4.2

## Warning: package 'tibble' was built under R version 4.4.2

## Warning: package 'tidyr' was built under R version 4.4.2

## Warning: package 'dplyr' was built under R version 4.4.2

## Warning: package 'lubridate' was built under R version 4.4.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# install.packages("car")
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.4.2
```

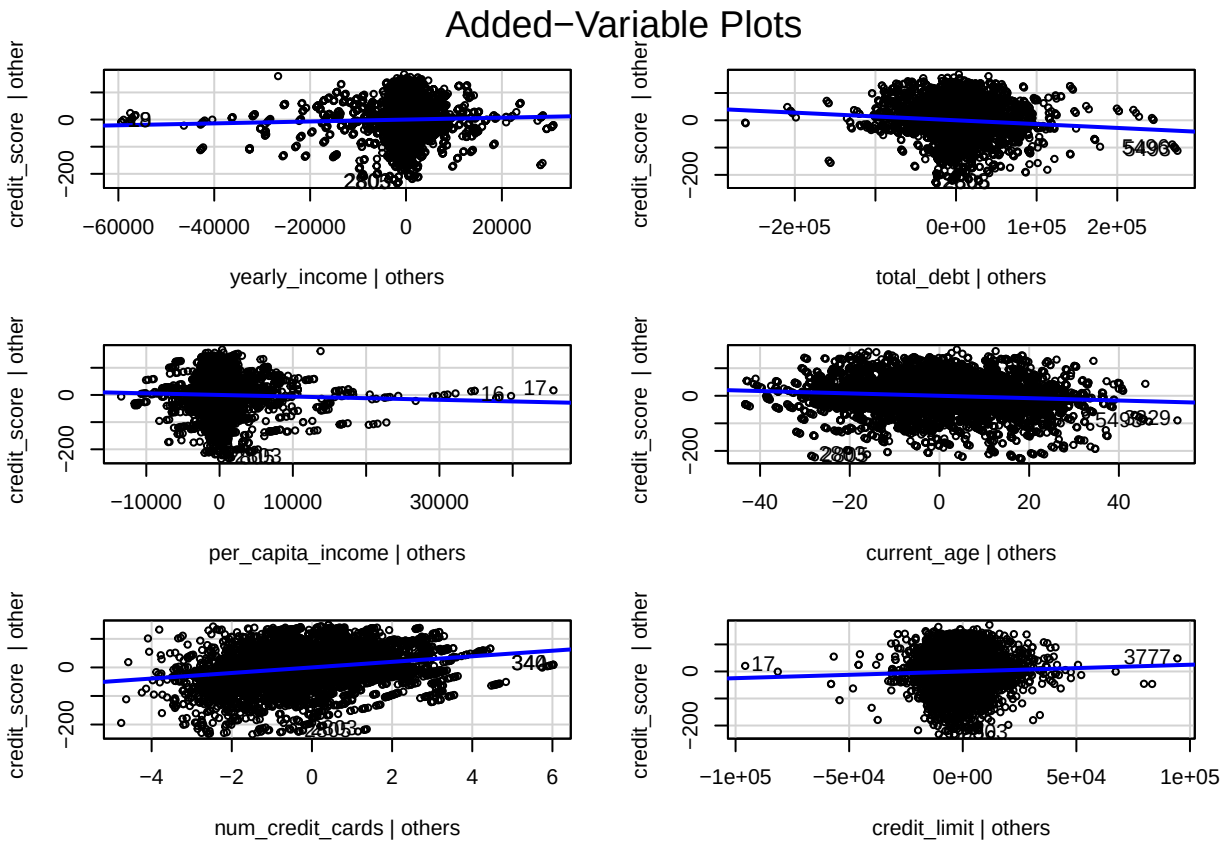
```
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
# Read the data
cards_df <- read.csv("C:/Users/elias/Downloads/NSDC/cards_data.csv")
users_df <- read.csv("C:/Users/elias/Downloads/NSDC/users_data.csv")

# Merge datasets on client_id (cards) and id (users)
merged_df <- cards_df %>%
  rename(user_id = client_id) %>%
  left_join(users_df, by = c("user_id" = "id"))

# Clean income and debt columns (remove $ and commas, convert to numeric)
merged_df <- merged_df %>%
  mutate(
    credit_limit = as.numeric(gsub("[$,]", "", credit_limit)),
    yearly_income = as.numeric(gsub("[$,]", "", yearly_income)),
    total_debt = as.numeric(gsub("[$,]", "", total_debt)),
    per_capita_income = as.numeric(gsub("[$,]", "", per_capita_income))
  )

# Remove rows with missing or NA credit_score
merged_df <- merged_df %>% filter(!is.na(credit_score))

# Subset relevant variables for modeling
model_data <- merged_df %>%
  select(credit_score, yearly_income, total_debt, per_capita_income, current_age, num_credit_cards, cre

# Fit the linear model
model <- lm(credit_score ~ ., data = model_data)
summary(model)
```

```
##
## Call:
## lm(formula = credit_score ~ ., data = model_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -235.438  -34.050   -0.935   39.254  169.442
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.039e+02  3.189e+00 220.705  < 2e-16 ***
## yearly_income      3.508e-04  1.201e-04   2.921   0.0035 **
## total_debt        -1.398e-04  1.881e-05  -7.432 1.21e-13 ***
## per_capita_income -5.973e-04  2.404e-04  -2.485   0.0130 *
```

```
## current_age       -4.302e-01  5.263e-02  -8.175 3.57e-16 ***
## num_credit_cards   9.744e+00  5.416e-01  17.992  < 2e-16 ***
## credit_limit       2.482e-04  7.949e-05   3.122   0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.2 on 6139 degrees of freedom
## Multiple R-squared:  0.06401,    Adjusted R-squared:  0.06309
## F-statistic: 69.97 on 6 and 6139 DF,  p-value: < 2.2e-16
```

```
# Added variable plots
avPlots(model)
```




```
# Variance Inflation Factors
vif(model)
```

```
##     yearly_income       total_debt per_capita_income       current_age
##         12.855461          1.547472         12.564242          1.636418
##  num_credit_cards      credit_limit
##          1.338881          1.496274
```