NSDC Week 5 Analysis & Visualization

2025-05-07

Data

```r
#Upload 2 datasets
users_data <- read.csv("C:/Users/elias/Downloads/NSDC/users_data.csv")
cards_data <- read.csv("C:/Users/elias/Downloads/NSDC/clean_cards_data.csv")

#Clean users dataset (dates left as is)
users_data$per_capita_income <- as.numeric(sub("^\\$", "", users_data$per_capita_income))

users_data$yearly_income <- as.numeric(sub("^\\$", "", users_data$yearly_income))

users_data$total_debt <- as.numeric(sub("^\\$", "", users_data$total_debt))

users_data$gender <- as.factor(users_data$gender)
```

Multiple Regression Model

```r
#Add credit score to cards data set
cards2 <- cards_data[cards_data$id %in% users_data$id,]
cards2$credit_score <- numeric(length(cards2$id))

for (i in seq_along(cards2$id)) {
  score <- users_data$credit_score[cards2$id[i] == users_data$id]
  cards2$credit_score[cards2$id[i] == cards2$id] <- score
}


#Generate models based off of both data sets
model_users <- lm(credit_score ~ id + current_age + retirement_age + birth_year + birth_month + gender +

model_cards <- lm(credit_score ~ id + client_id + cvv + card_number + num_cards_issued + year_pin_last_

#Summarize the models
summary(model_users)
```

```
##
## Call:
## lm(formula = credit_score ~ id + current_age + retirement_age +
##     birth_year + birth_month + gender + latitude + longitude +
##     per_capita_income + yearly_income + total_debt + num_credit_cards,
```

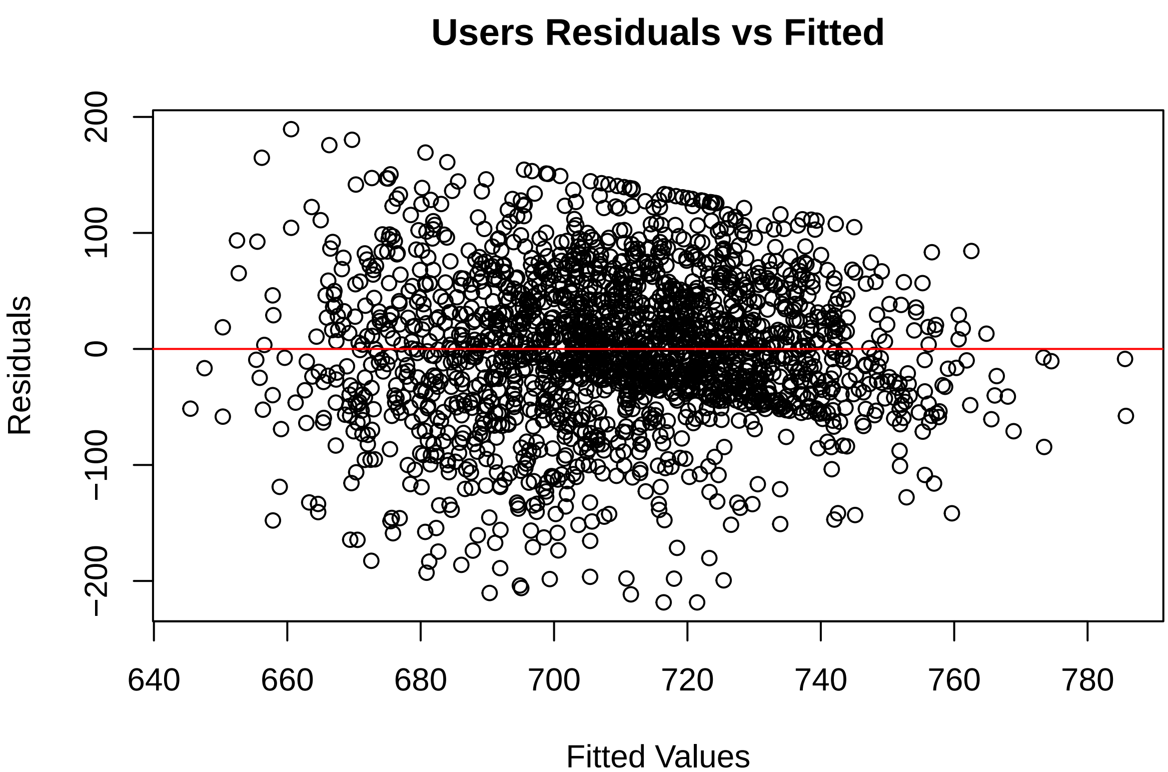
```
##     data = users_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -218.452  -35.775   -0.596   41.172  189.428
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.661e+04  9.965e+03  -1.667   0.0957 .
## id                 1.163e-04  2.480e-03   0.047   0.9626
## current_age        7.891e+00  4.936e+00   1.598   0.1101
## retirement_age     2.284e+00  4.017e-01   5.687 1.49e-08 ***
## birth_year         8.494e+00  4.934e+00   1.721   0.0853 .
## birth_month        1.811e-01  5.468e-01   0.331   0.7405
## genderMale         8.220e-01  2.859e+00   0.288   0.7737
## latitude           4.955e-01  2.834e-01   1.748   0.0806 .
## longitude          2.995e-02  8.861e-02   0.338   0.7354
## per_capita_income -4.236e-04  5.167e-04  -0.820   0.4124
## yearly_income      3.536e-04  2.615e-04   1.352   0.1765
## total_debt        -1.590e-04  3.469e-05  -4.584 4.84e-06 ***
## num_credit_cards   1.145e+01  1.019e+00  11.239  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.8 on 1987 degrees of freedom
## Multiple R-squared:  0.1047, Adjusted R-squared:  0.09928
## F-statistic: 19.36 on 12 and 1987 DF,  p-value: < 2.2e-16
```

```r
summary(model_cards)
```

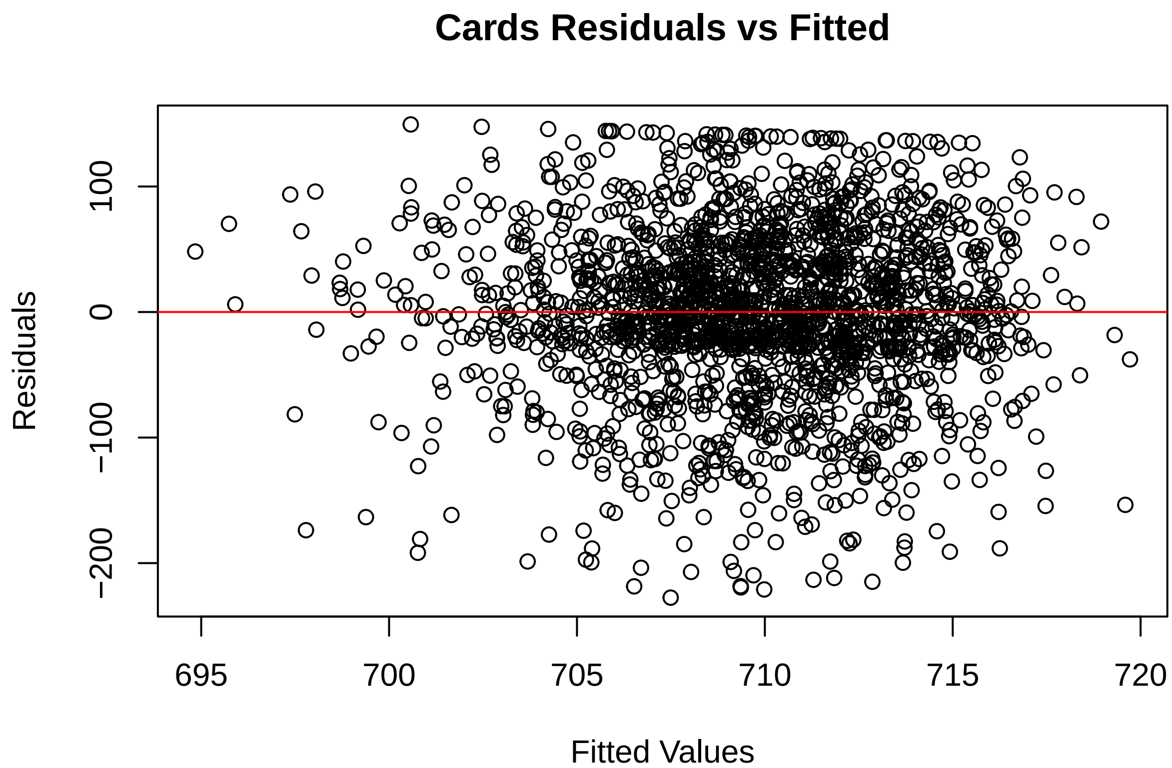
```
##
## Call:
## lm(formula = credit_score ~ id + client_id + cvv + card_number +
##     num_cards_issued + year_pin_last_changed + card_brand_num +
##     has_chip_num + credit_limit_num, data = cards2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -227.494  -29.978    1.341   43.264  149.423
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.915e+02  7.430e+02   0.796    0.426
## id                    -1.811e-03  2.878e-03  -0.629    0.529
## client_id              3.025e-03  2.638e-03   1.147    0.252
## cvv                   -4.165e-03  5.206e-03  -0.800    0.424
## card_number            2.918e-16  1.223e-15   0.239    0.811
## num_cards_issued       3.074e+00  2.880e+00   1.067    0.286
## year_pin_last_changed  5.517e-02  3.687e-01   0.150    0.881
## card_brand_num        -1.378e-01  2.083e+00  -0.066    0.947
## has_chip_num           4.965e+00  4.774e+00   1.040    0.298
## credit_limit_num      -1.434e-04  1.295e-04  -1.107    0.269
##
## Residual standard error: 67.27 on 1990 degrees of freedom
```

```
## Multiple R-squared:  0.003029,	Adjusted R-squared:  -0.00148
## F-statistic: 0.6717 on 9 and 1990 DF,  p-value: 0.7352
```

```r
#Plot Fitted Values vs residuals graphs for both.
plot(model_users$fitted.values, resid(model_users),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Users Residuals vs Fitted")
abline(h = 0, col = "red")
```



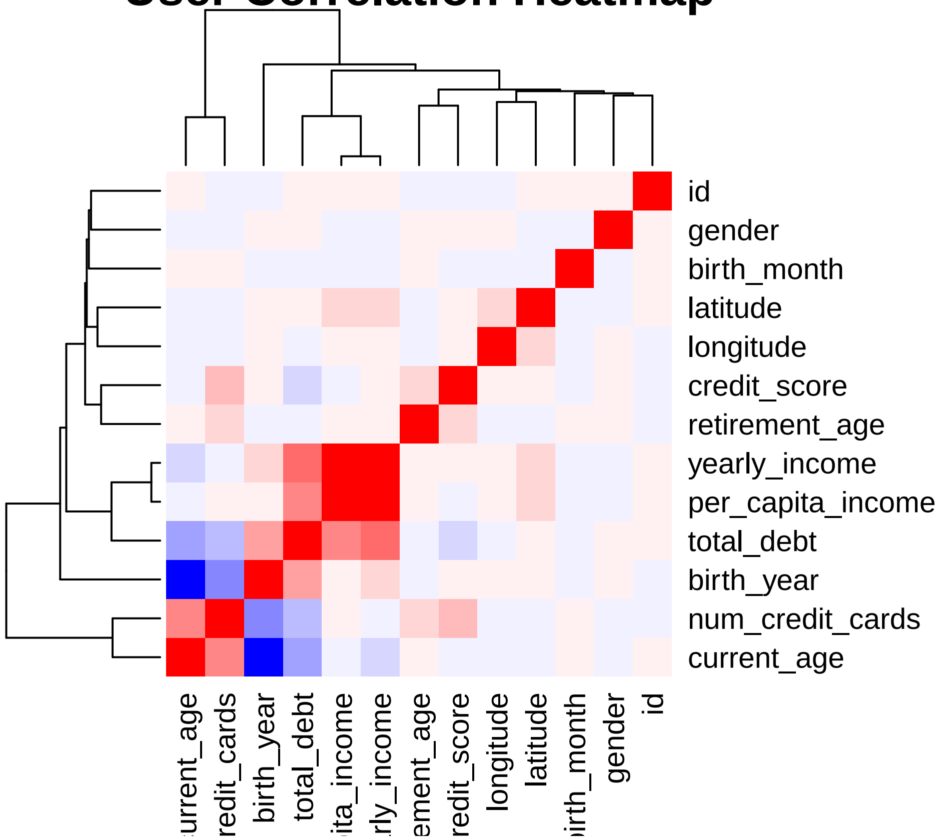

```r
plot(model_cards$fitted.values, resid(model_cards),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Cards Residuals vs Fitted")
abline(h = 0, col = "red")
```



Heatmaps

```
#Users Heatmap
subset_data <- users_data[sapply(users_data, function(x) is.numeric(x) || is.factor(x))]
subset_data$gender <- as.numeric(subset_data$gender)
cor_matrix <- cor(subset_data)

heatmap(cor_matrix,
        main = "User Correlation Heatmap",
        col = colorRampPalette(c("blue", "white", "red"))(20),
        scale = "none")
```

User Correlation Heatmap




```
subset2 <- cards2[sapply(cards2, function(x) is.numeric(x))]
subset2$card_on_dark_web_num <- NULL
cor_matrix <- cor(subset2)


#Cards Heatmap
library(ggplot2)
```

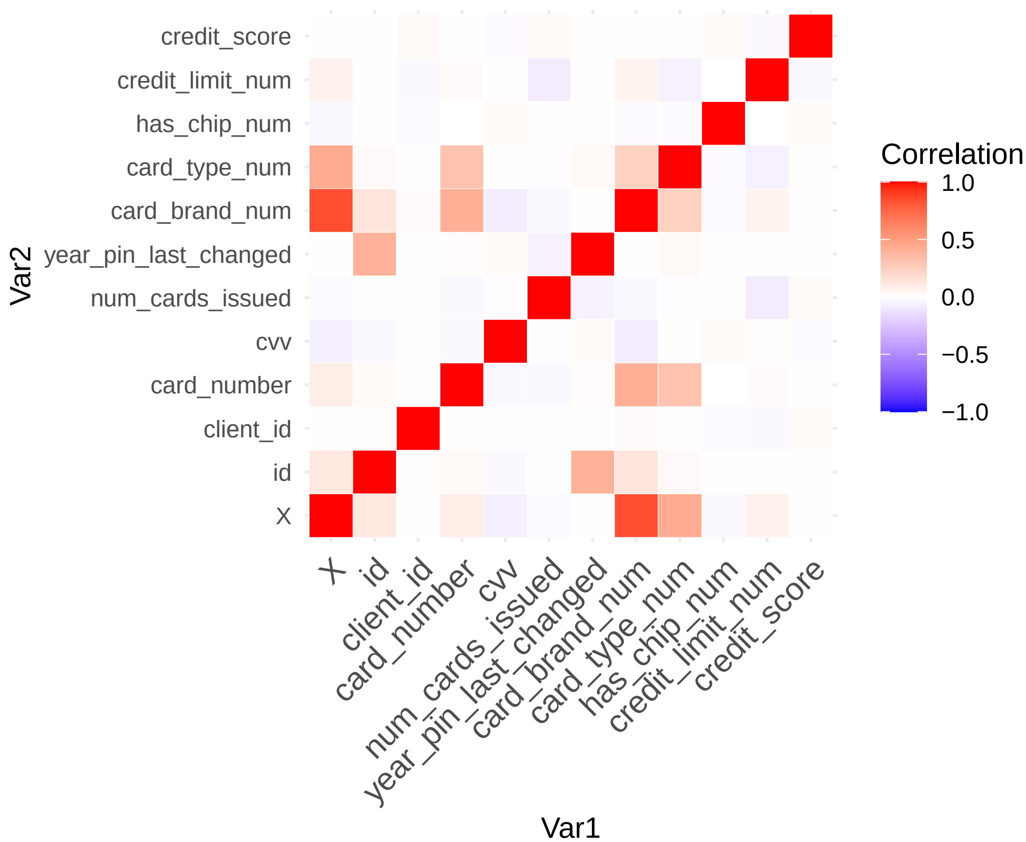
```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.4.3
```

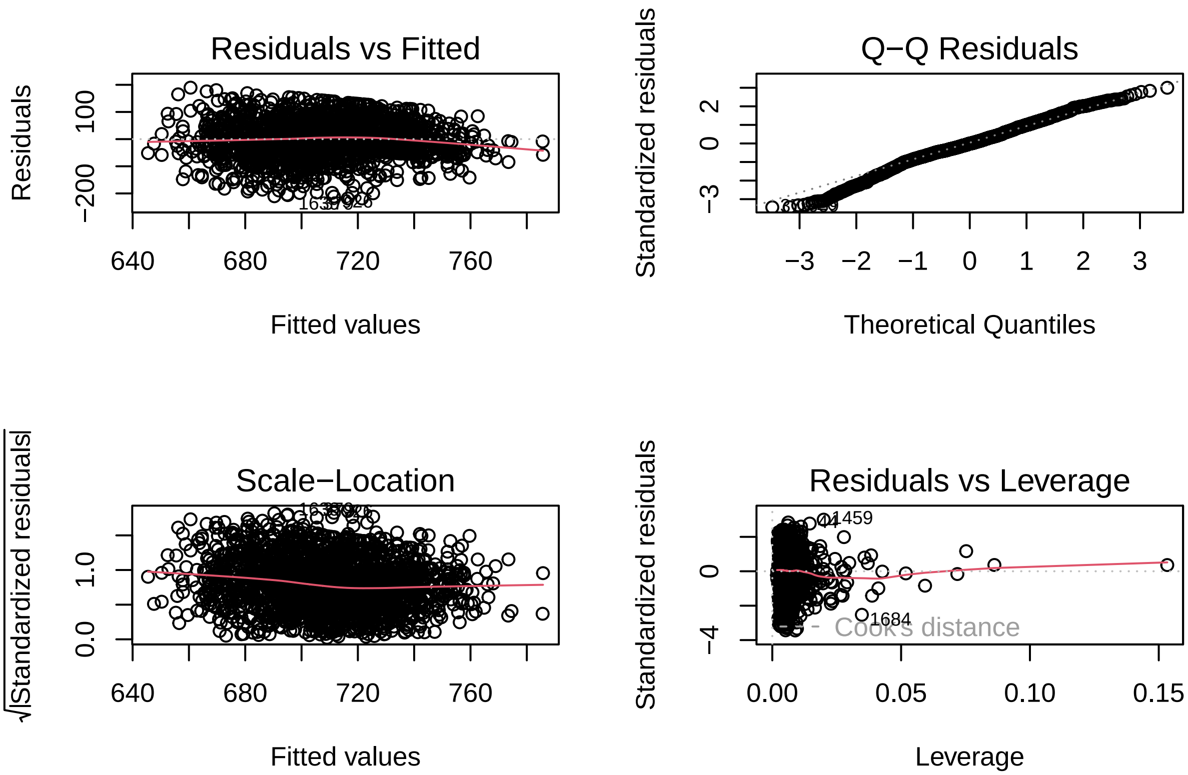
```
# Melt the matrix into long format
melted_cor <- melt(cor_matrix)

# Plot heatmap
ggplot(data = melted_cor, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlation") +
  theme_minimal() +
```

```r
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1)) +
  coord_fixed()
```

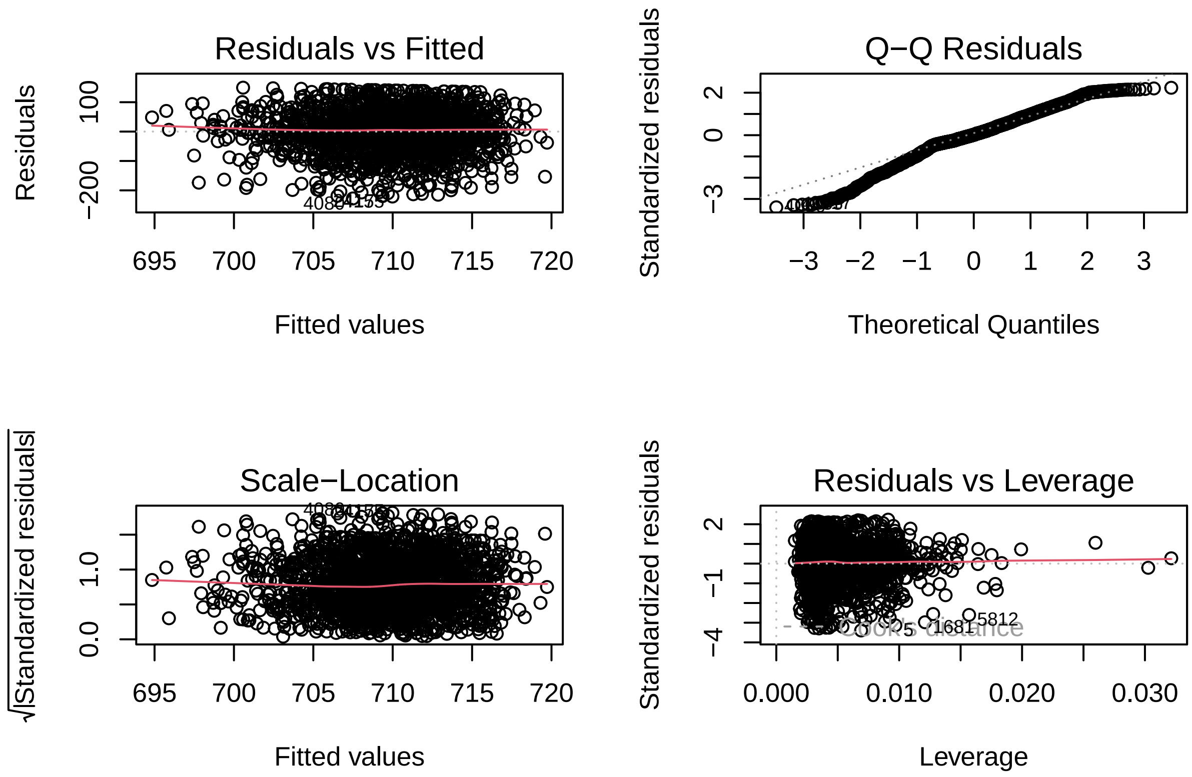


Diagnostic Plots

```r
#Users Data
par(mfrow = c(2, 2))
plot(model_users)
```




```
#Cards Data
par(mfrow = c(2, 2))
plot(model_cards)
```



R^2 and P-Values

```r
#Info from users data set

# R-squared and Adjusted R-squared
r_squared <- summary(model_users)$r.squared
adj_r_squared <- summary(model_users)$adj.r.squared

# p-values for each predictor
p_values <- summary(model_users)$coefficients[, 4]

print(paste("R-squared:", r_squared))
```

```
## [1] "R-squared: 0.104689660012925"
```

```r
print(paste("Adjusted R-squared:", adj_r_squared))
```

```
## [1] "Adjusted R-squared: 0.099282652423672"
```

```r
print("P-values for each predictor:")
```

```
## [1] "P-values for each predictor:"
```

```r
print(p_values)
```

```
##       (Intercept)                id       current_age    retirement_age 
##      9.565014e-02      9.625874e-01      1.100945e-01      1.487319e-08 
##        birth_year       birth_month        genderMale          latitude 
##      8.531736e-02      7.405347e-01      7.737494e-01      8.062354e-02 
##         longitude per_capita_income     yearly_income        total_debt 
##      7.353877e-01      4.124181e-01      1.764805e-01      4.838252e-06 
##  num_credit_cards 
##      1.864227e-28
```

```r
#Info from cards data set

# R-squared and Adjusted R-squared
r_squared <- summary(model_cards)$r.squared
adj_r_squared <- summary(model_cards)$adj.r.squared

# p-values for each predictor
p_values <- summary(model_cards)$coefficients[, 4]

print(paste("R-squared:", r_squared))
```

```
## [1] "R-squared: 0.00302872541951428"
```

```r
print(paste("Adjusted R-squared:", adj_r_squared))
```

```
## [1] "Adjusted R-squared: -0.00148018989265886"
```

```r
print("P-values for each predictor:")
```

```
## [1] "P-values for each predictor:"
```

```r
print(p_values)
```

```
##         (Intercept)                  id           client_id 
##           0.4260912           0.5292073           0.2516803 
##                 cvv         card_number     num_cards_issued 
##           0.4237663           0.8114309           0.2859292 
## year_pin_last_changed      card_brand_num         has_chip_num 
##           0.8810829           0.9472699           0.2984342 
##    credit_limit_num 
##           0.2685389
```