# Load the dataset  
credit\_data <- read.csv("C:/Users/me/Downloads/Credit.csv")  
# 1) Generate summary statistics  
summary(credit\_data)

## Income Limit Rating Age   
## Min. : 10354 Min. : 1160 Min. :126.0 Min. :23.00   
## 1st Qu.: 23150 1st Qu.: 3976 1st Qu.:304.0 1st Qu.:42.00   
## Median : 37141 Median : 5147 Median :380.0 Median :55.50   
## Mean : 49979 Mean : 5485 Mean :405.1 Mean :55.61   
## 3rd Qu.: 63740 3rd Qu.: 6453 3rd Qu.:469.0 3rd Qu.:69.00   
## Max. :186634 Max. :13913 Max. :982.0 Max. :98.00   
## Education Student Gender Married   
## Min. : 5.00 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:11.00 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :14.00 Median :0.0000 Median :1.0000 Median :1.0000   
## Mean :13.43 Mean :0.1258 Mean :0.5323 Mean :0.6194   
## 3rd Qu.:16.00 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :20.00 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Balance   
## Min. : 5.0   
## 1st Qu.: 338.0   
## Median : 637.5   
## Mean : 671.0   
## 3rd Qu.: 960.8   
## Max. :1999.0

# Quiz question #1: How many cardholders in the full dataset are students?  
students\_count <- sum(credit\_data$Student == 1)  
students\_count

## [1] 39

# 2) Partition the dataset into training and validation sets  
set.seed(42) # Set the seed for reproducibility  
n <- nrow(credit\_data)  
train\_indices <- sample(1:n, n/2)  
train\_data <- credit\_data[train\_indices, ]  
validation\_data <- credit\_data[-train\_indices, ]  
# 3) Create a correlation matrix with quantitative variables in the training dataframe  
quant\_vars <- c("Income", "Limit", "Rating", "Age", "Education", "Balance")  
cor\_matrix <- cor(train\_data[quant\_vars])  
cor\_matrix

## Income Limit Rating Age Education  
## Income 1.00000000 0.864047148 0.86065289 0.177950916 -0.028697641  
## Limit 0.86404715 1.000000000 0.99613733 0.137445826 -0.003477829  
## Rating 0.86065289 0.996137328 1.00000000 0.131264178 -0.020412443  
## Age 0.17795092 0.137445826 0.13126418 1.000000000 0.045031014  
## Education -0.02869764 -0.003477829 -0.02041244 0.045031014 1.000000000  
## Balance 0.44421732 0.783043777 0.78640576 0.003618942 0.051509856  
## Balance  
## Income 0.444217315  
## Limit 0.783043777  
## Rating 0.786405762  
## Age 0.003618942  
## Education 0.051509856  
## Balance 1.000000000

# 4) Multiple regression analysis  
model <- lm(Balance ~ ., data = train\_data)  
summary(model)

##   
## Call:  
## lm(formula = Balance ~ ., data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -78.081 -22.004 -2.532 20.292 104.743   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.715e+02 1.921e+01 -34.955 < 2e-16 \*\*\*  
## Income -1.003e-02 1.299e-04 -77.171 < 2e-16 \*\*\*  
## Limit 2.205e-01 1.438e-02 15.337 < 2e-16 \*\*\*  
## Rating 1.563e+00 2.110e-01 7.406 9.57e-12 \*\*\*  
## Age -9.563e-01 1.546e-01 -6.186 5.88e-09 \*\*\*  
## Education 1.627e-01 8.399e-01 0.194 0.8467   
## Student 4.897e+02 8.074e+00 60.659 < 2e-16 \*\*\*  
## Gender -4.894e+00 5.257e+00 -0.931 0.3534   
## Married -1.071e+01 5.477e+00 -1.956 0.0524 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.3 on 146 degrees of freedom  
## Multiple R-squared: 0.9938, Adjusted R-squared: 0.9935   
## F-statistic: 2925 on 8 and 146 DF, p-value: < 2.2e-16

# Quiz question #3: What is the slope coefficient for the Rating variable?  
slope\_rating <- coef(model)["Rating"]  
slope\_rating

## Rating   
## 1.562508

# 5) Calculate the Variance Inflation Factor (VIF  
library(car)

## Loading required package: carData

vif\_values <- vif(model)  
vif\_values

## Income Limit Rating Age Education Student Gender   
## 4.144032 141.542257 137.216277 1.047025 1.061899 1.088548 1.026109   
## Married   
## 1.020589

# Quiz question #4: What is the VIF for the Limit variable?  
vif\_limit <- vif\_values["Limit"]  
vif\_limit

## Limit   
## 141.5423

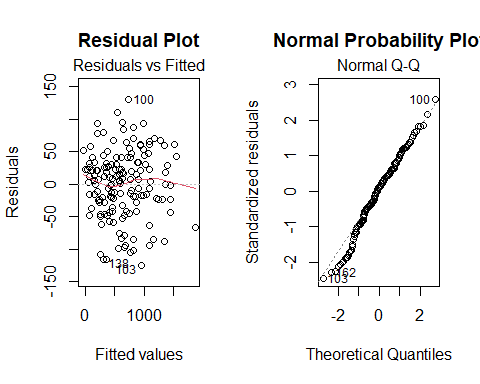
# Quiz question #5: What problem does the VIF for Limit suggest that we have with the analysis? (MC)  
# A high VIF for a predictor variable suggests multicollinearity, indicating that the predictor variable is highly correlated with other predictor variables in the model.  
# 6) New multiple regression analysis with selected predictor variables  
new\_model <- lm(Balance ~ Income + Rating + Age + Education + Student + Gender + Married, data = train\_data)  
summary(new\_model)

##   
## Call:  
## lm(formula = Balance ~ Income + Rating + Age + Education + Student +   
## Gender + Married, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -124.75 -28.24 5.45 32.09 129.95   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.027e+02 2.770e+01 -28.976 < 2e-16 \*\*\*  
## Income -9.696e-03 2.063e-04 -46.992 < 2e-16 \*\*\*  
## Rating 4.750e+00 5.838e-02 81.364 < 2e-16 \*\*\*  
## Age -8.614e-01 2.487e-01 -3.463 0.000699 \*\*\*  
## Education 2.863e+00 1.322e+00 2.165 0.031998 \*   
## Student 4.756e+02 1.292e+01 36.822 < 2e-16 \*\*\*  
## Gender -6.316e+00 8.464e+00 -0.746 0.456722   
## Married -1.873e+01 8.780e+00 -2.134 0.034530 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52.01 on 147 degrees of freedom  
## Multiple R-squared: 0.9838, Adjusted R-squared: 0.983   
## F-statistic: 1276 on 7 and 147 DF, p-value: < 2.2e-16

# Quiz question #6: What is the new slope coefficient for the Rating variable?  
new\_slope\_rating <- coef(new\_model)["Rating"]  
new\_slope\_rating

## Rating   
## 4.749937

# 7) Residual plot and normal probability plot  
par(mfrow = c(1, 2)) # Set up a 1x2 plot layout  
  
# Residual plot  
plot(new\_model, which = 1, main = "Residual Plot")  
# Normal probability plot  
plot(new\_model, which = 2, main = "Normal Probability Plot")



Quiz question #7: What pattern do you see in the residual plot? (MC)

Majority of data are along the residual line

Quiz question #8: What does this pattern tell you? (MC)

The pattern tells that the regression model is a good model for fitting the data.

Quiz question #9: What pattern do you see in the normal probability plot? (MC)

The data points are concentrated along the straight line

Quiz question #10: What does this pattern tell you? (MC)

The data follows a normal probability distribution

#9) New multiple regression analysis with significant predictor variables  
significant\_vars <- c("Rating", "Student")  
new\_model\_sig <- lm(Balance ~ Income + Limit + Rating + Age + Education + Student + Married, data = train\_data)  
summary(new\_model\_sig)

##   
## Call:  
## lm(formula = Balance ~ Income + Limit + Rating + Age + Education +   
## Student + Married, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -75.293 -21.617 -2.997 20.629 107.430   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.741e+02 1.900e+01 -35.480 < 2e-16 \*\*\*  
## Income -1.002e-02 1.295e-04 -77.358 < 2e-16 \*\*\*  
## Limit 2.207e-01 1.437e-02 15.363 < 2e-16 \*\*\*  
## Rating 1.558e+00 2.108e-01 7.389 1.03e-11 \*\*\*  
## Age -9.483e-01 1.543e-01 -6.147 7.07e-09 \*\*\*  
## Education 1.785e-01 8.393e-01 0.213 0.8319   
## Student 4.891e+02 8.042e+00 60.822 < 2e-16 \*\*\*  
## Married -1.100e+01 5.466e+00 -2.012 0.0461 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.28 on 147 degrees of freedom  
## Multiple R-squared: 0.9938, Adjusted R-squared: 0.9935   
## F-statistic: 3346 on 7 and 147 DF, p-value: < 2.2e-16

# Quiz question #12: What is the slope coefficient for the Age variable?

-9.483e-01   
# Quiz question #13: How would you interpret the slope coefficient for the Rating variable? (MC)

The slope coefficient for rating variable is 1.558e+00. Therefore, when rating increases by 1 unit, balance increases by 1.558e+00   
# Quiz question #14: How would you interpret the slope coefficient for the Student variable? (MC)

When student variable increases by 1 unit, balance increase by 4.891e+02  
# Quiz question #15: What is the adjusted R2 for this regression analysis?

0.9935  
# Quiz question #16: How can this adjusted R2 value be interpreted? (MC)

99.35% variation in the dependent variable is explained by the independent variables  
# Quiz question #17: What is the standardized slope coefficient for the Income variable? 16.33681

# Quiz question #18: Looking at the standardized slope coefficients, which variable makes the strongest unique contribution to predicting credit card balance? (MC)

Student  
# 10) Final multiple regression analysis on validation dataframe with significant predictor variables  
validation\_model <- lm(Balance ~ Income + Limit + Rating + Age + Education + Student + Married, data = validation\_data)  
summary(validation\_model)

##   
## Call:  
## lm(formula = Balance ~ Income + Limit + Rating + Age + Education +   
## Student + Married, data = validation\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -69.12 -21.43 -8.42 20.14 90.28   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.752e+02 1.862e+01 -36.260 < 2e-16 \*\*\*  
## Income -1.004e-02 1.338e-04 -75.010 < 2e-16 \*\*\*  
## Limit 2.332e-01 1.581e-02 14.749 < 2e-16 \*\*\*  
## Rating 1.416e+00 2.328e-01 6.084 9.69e-09 \*\*\*  
## Age -1.048e+00 1.618e-01 -6.474 1.34e-09 \*\*\*  
## Education -7.286e-01 8.368e-01 -0.871 0.385   
## Student 4.980e+02 8.482e+00 58.711 < 2e-16 \*\*\*  
## Married 2.195e-01 5.578e+00 0.039 0.969   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33.07 on 147 degrees of freedom  
## Multiple R-squared: 0.9942, Adjusted R-squared: 0.994   
## F-statistic: 3621 on 7 and 147 DF, p-value: < 2.2e-16

###11  
prediction\_model<- lm(Balance ~ Income + Rating + Age + Student, data = validation\_data)  
prediction\_data <- read.csv("C:/Users/me/Downloads/credit\_card\_prediction.csv")  
predicted\_balances <- predict(prediction\_model, newdata = prediction\_data, interval = "prediction", level = 0.95)  
# Quiz question #19: What is the predicted balance for new cardholder #1?  
predicted\_balance\_card1 <- predicted\_balances[1, "fit"]  
predicted\_balance\_card1

## [1] 390.5752

# Quiz question #20: What is the 95% prediction interval for the predicted balance for new cardholder #2?  
prediction\_interval\_card2 <- predicted\_balances[2, c("lwr", "upr")]  
prediction\_interval\_card2

## lwr upr   
## 1412.314 1625.545