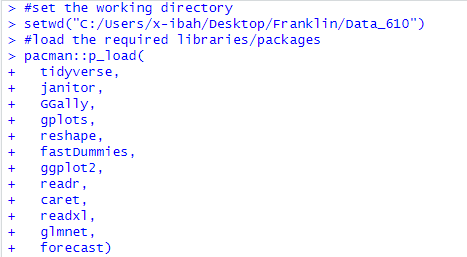
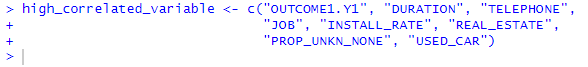
**Set the working Directory and installed and loaded packages**

**1. Explain what data are you using, number of observations, what is Y1, and how much is the mean for Y1.**  
  

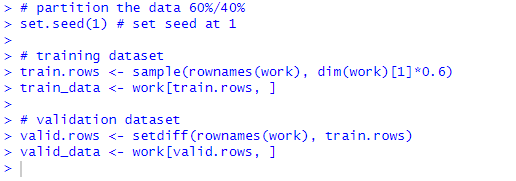

- I have loaded the German Credit dataset into a data frame called 'df'. This data frame includes **912** observations, with each observation corresponding to a row, and **30** variables, with each variable corresponding to a column. **- Y1** is the amount of the credit per dollars  
  
**- Mean of Y1**  

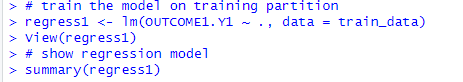
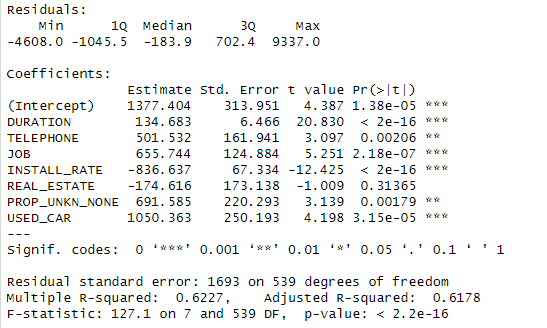

**2. Present correlation matrix with Y1 and the rest of the variables (continuous or dummy).   
List the top variables with high correlation with Y1 (positive or negative).**   
  
  
Variables with high correlation coefficients with OUTCOME1.Y1

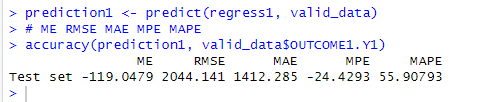
1. **Duration** 0.64 (strong positive relation)
2. **Telephone** 0.29 (moderate positive correlation)
3. **Job** 0.28 (moderate positive correlation)
4. **Install\_Rate** -0.27 (moderate negative relation)
5. **Used\_Car** 0.26 ((moderate positive correlation)
6. **Prop\_Unkn\_None** 0.25 (moderate positive correlation)
7. **Real\_Estate** -0.24 (moderate negative correlation)

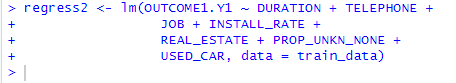
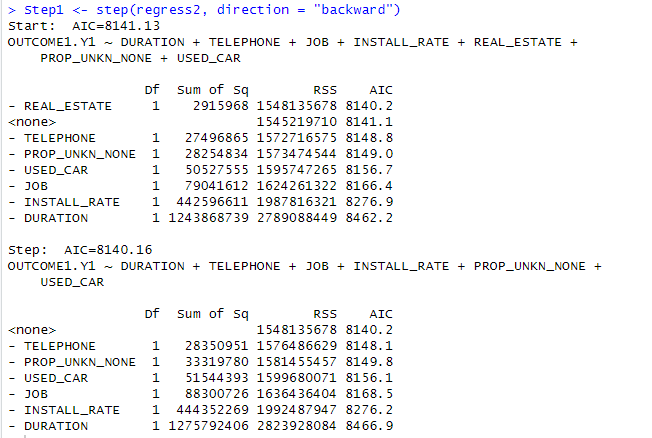
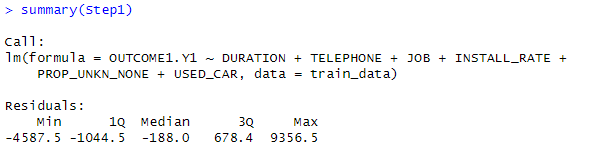
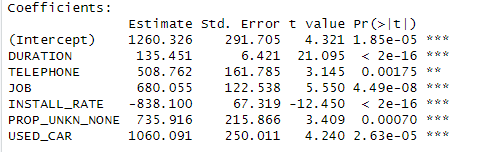
  
The variable "**high\_correlated\_variable**" denotes the top 7 variables highly correlated with the variable Y1.

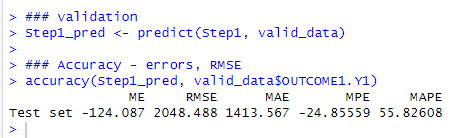
**3. Create a new dataset called "work" with Y1 and the top variables with high correlation with Y1 (positive or negative).**   
  


4. Partition the **work** dataset in 60% training set and 40% validation set.  
  


**5. Estimate regression model with Y1 as dependent variable and all variables with training data. Comment on the results.**  
  
  
  
  
**Comment:**  
In the regression model output, **DURATION** (p < 2e-16), **TELEPHONE** (p = 0.00206), **JOB** (p = 0.0002079), **INSTALL\_RATE** (p < 2e-16), **PROP\_UNKN\_NONE** (p = 0.00179), and **USED\_CAR** (p = 3.15e-05) are statistically significant predictors of the response variable, each with a p-value well below the 0.05 threshold. **REAL\_ESTATE,** with a p-value of 0.31365, is not a significant predictor. The significance levels are denoted by asterisks, with three asterisks marking the most significant predictors.

**6. Evaluate the accuracy of the above model (e.g. RMSE) prediction with validation data. Comment on the results.**  
  
  
**Comment** 🡪 The Root Mean Squared Error (RMSE) value is 2044.141, which measures the standard deviation of the prediction errors, seems quite high; this indicates that the model has a large average prediction error. The higher the RMSE, the worse the model is at predicting accurately.

**7. Run regression model with Y1 as dependent variable**   
  
  
  
  
**stepwise** selected variables with **training** data.   
  
  
  
  
Comment on the results.  
  
  
In the regression model results provided, all predictors are statistically significant. The DURATION variable has a p-value of less than 2e-16, indicating extremely high significance. TELEPHONE has a p-value of 0.00175, JOB has a p-value of 4.49e-08, INSTALL\_RATE has a p-value of less than 2e-16, PROP\_UNKN\_NONE has a p-value of 0.00070, and USED\_CAR has a p-value of 2.63e-05, each indicating strong evidence against the null hypothesis, hence they have significant associations with the response variable.

**8. Evaluate the accuracy of the above model (e.g. RMSE) prediction with validation data.**   
  
  
Comment 🡪 The Root Mean Squared Error (RMSE) value is **2048.488**, which measures the standard deviation of the prediction errors, seems quite high; this indicates that the model has a large average prediction error. The higher the RMSE, the worse the model is at predicting accurately.

**9. Compare the two regression models by RMSE.**Regression 1 has a Root Mean Square Error (RMSE) of 2044.141, while Regression 2 has a slightly higher RMSE of 2048.488. The lower RMSE of Regression 1 indicates it has a marginally better fit to the observed data compared to Regression 2. The numerical difference in RMSE between the two regressions is 4.347, which is relatively small.