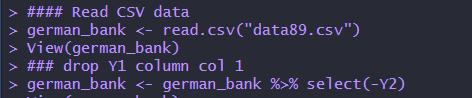
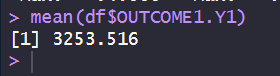
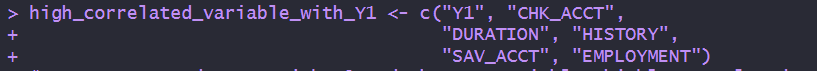
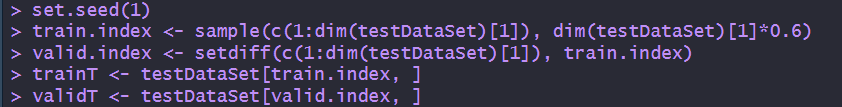
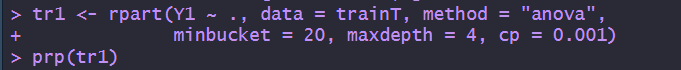
**Regression Tree Model (CART) for Y1. Y2 is excluded from this analysis.**

**1. Explain what data you are using, number of observations, number of variables**  
  
  
I have loaded the German Credit dataset into a data frame called '**german\_bank**'. This data frame includes **912** observations, with each observation corresponding to a row, and **30** variables, with each variable corresponding to a column.  
  
**What is Y1**  
The variable Y1 represents the Credit Amount in Dollars  
  
**What is the mean for Y1.**   


**2. (From HW#5). Create a new dataset with Y1 and the top variables with high correlation with Y1 (positive or negative).**   
  
  
  
  
  


**3. Partition the data in 60% training set and 40% validation set.**  
  

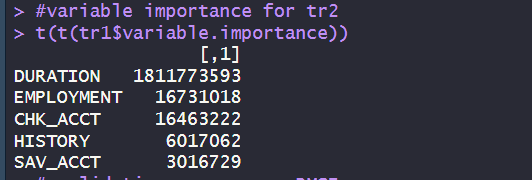

**4. Use the training data to build regression tree model for Y1 (CART). Use 3 to 5 tree levels.**  
  
   
**Comment on the results:** This code constructs a regression tree with a minimum bucket size of 20, a maximum depth of 4, and uses a complexity parameter of 0.001 to regulate the tree's size and prevent overfitting.

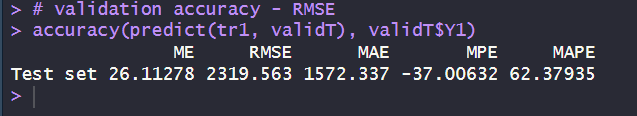
**5. Plot the CART tree (must be readable).**   
A diagram of a company structure

Description automatically generated  
  
**Comment on the tree:** This regression tree has four (4) levels, indicating the depth of questions from top to bottom. It branches out based on conditions starting from the top variable, **Duration of Credit in Month**, which is the root node of the tree. The tree splits into **9 end nodes**, or 'leaves,' where each leaf represents a final decision point or outcome.

**6. Pick the end-node with the lowest Y1 (lowest Y1 average for the node) and write down the path to this node in plain English.**  
  
To reach End-node 1, which has the lowest average Y1 of $1660, follow this path: Start by identifying loans with a duration longer than 25 months. Next, within that group, focus on loans with durations greater than 17 months. Finally, among these, select loans with durations exceeding 15 months. This specific pathway leads to End-node 1, marking the group with the lowest average loan amount.

**7. Pick one end-node with the highest Y1 (highest Y1 average for the node) and write down the path to this node in plain English.**  
  
To reach End-node 9, which has the highest average Y1 of $7433, follow this path: Start by identifying loans with a duration less than 25 months. Next, within that group, focus on loans with a duration less than 39 months. This specific pathway leads to End-node 9, marking the group with the highest average loan amount.

**8. Show the variable importance measure**   
  
  
**Comment on the results:** The results show the variable importance scores from this regression tree model. The " **DURATION** " variable has the highest score, indicating it is the most important predictor in the model. " **EMPLOYMENT** " follows as the second most significant predictor, with " **CHK\_ACCT** " third. These scores help determine which features most influence the model's decisions.

**9. Use the validation data to estimate the accuracy of the model (e.g. RMSE)**   
  
  
  
**Comment on the result:** This model demonstrates good predictability because its RMSE (Root Mean Square Error) of 2320 is lower than the average Y1 value (the credit amount in dollars), which is 3254.