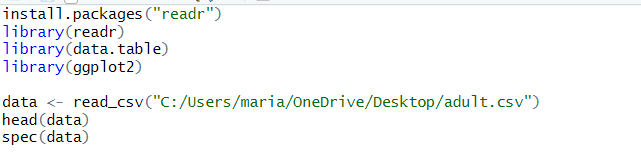
**INFS-417**

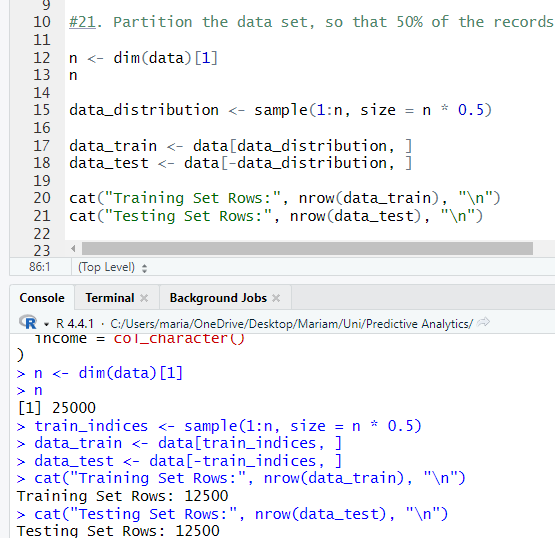
**CHAPTER 5-ASSIGNMENT 2**

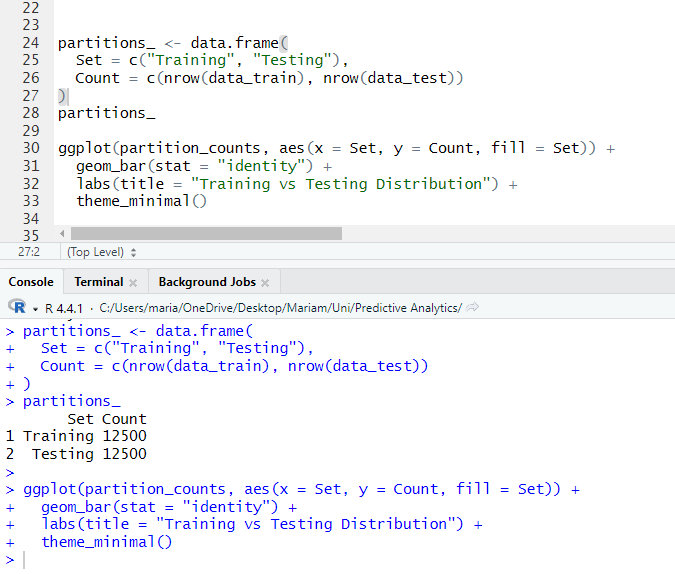
**STEP 1: INSTALLING LIBRARIES & IMPORTING DATA**

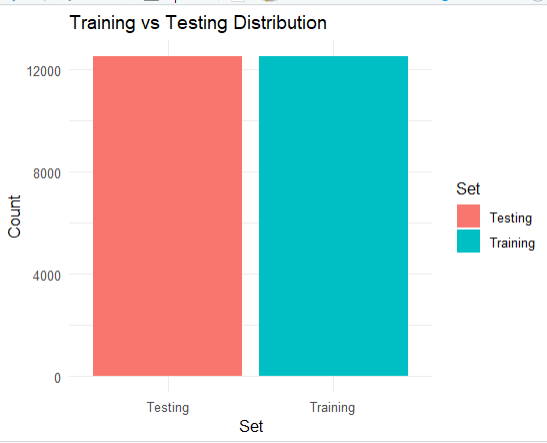


**STEP 2: CODE FOR QUESTIONS**

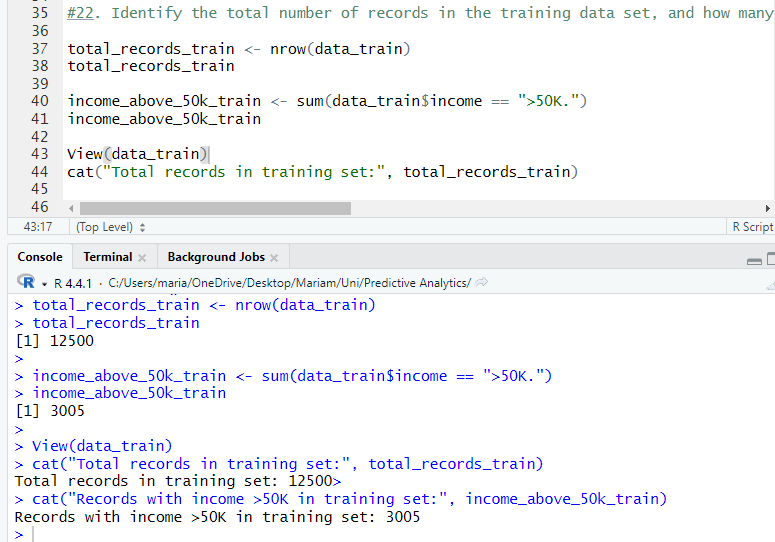
**21. Partition the data set, so that 50% of the records are included in the training data set and 50% are included in the test data set. Use a bar graph to confirm your proportions.**





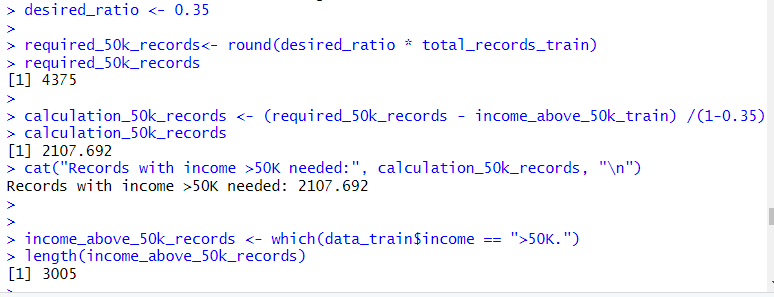


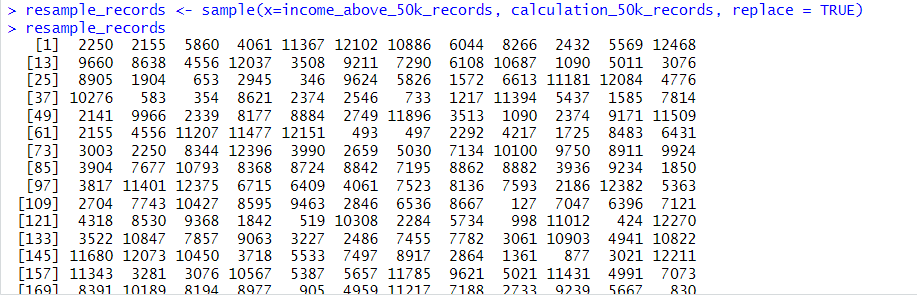
**22. Identify the total number of records in the training data set, and how many records in the training data set have an income value of >50 K.**



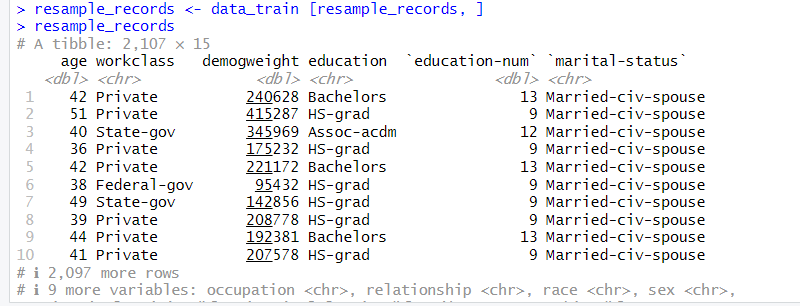
The total number of records in the training dataset is 12500 and 3005 are the data points where income is greater than 50k.

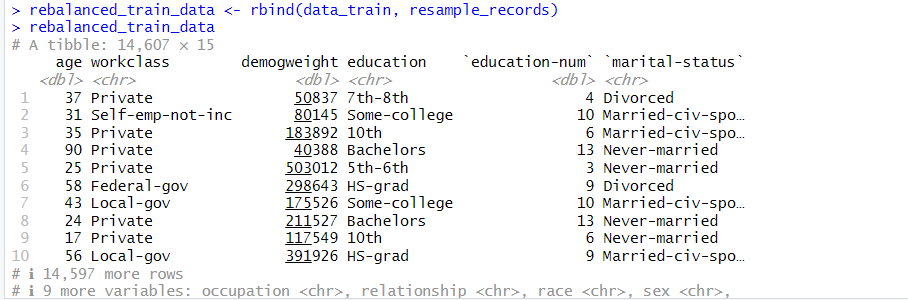
**23. Use your answers from the previous exercise to calculate how many records with income >50 K you need to resample in order to have 35% of the rebalanced data set have incomes of >50 K.**



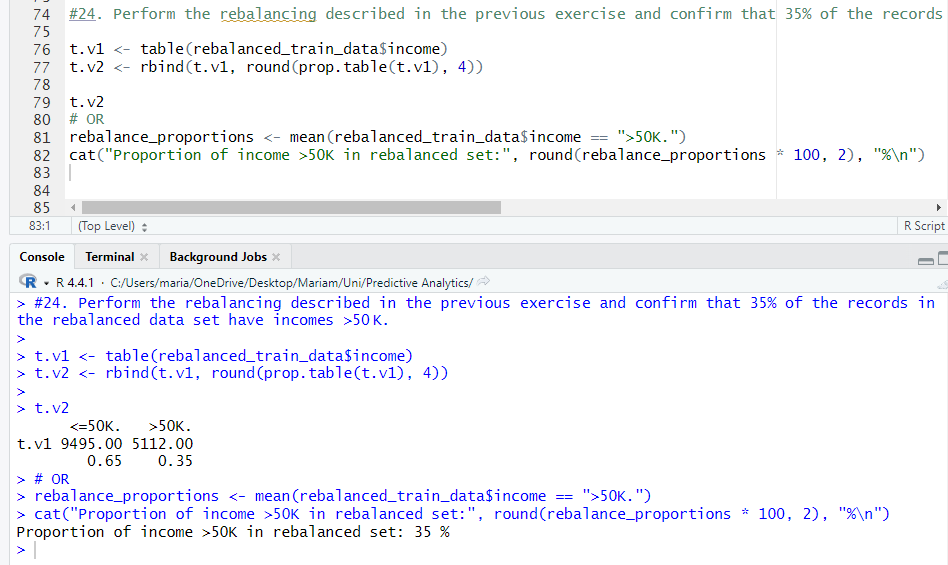


The number of records seems to be 2108 as per requirement with income > 50k.

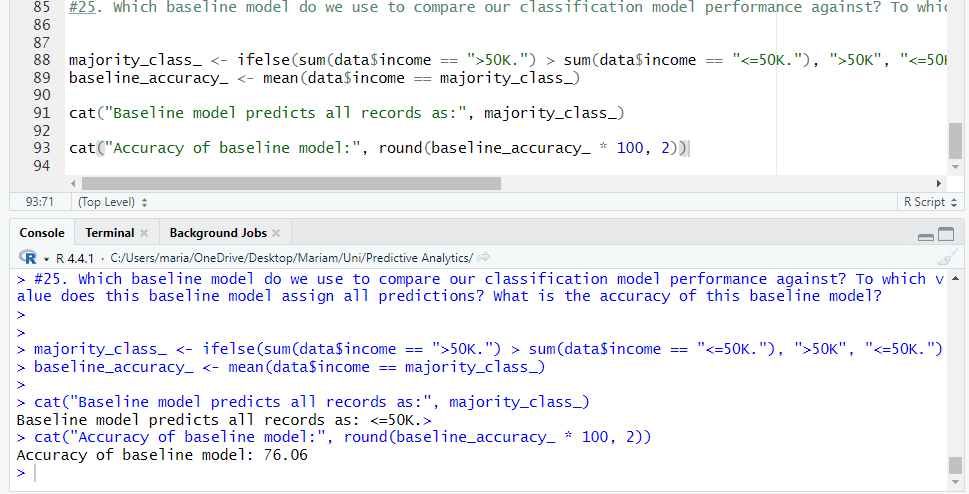




**24. Perform the rebalancing described in the previous exercise and confirm that 35% of the records in the rebalanced data set have incomes >50 K.**



**25. Which baseline model do we use to compare our classification model performance against? To which value does this baseline model assign all predictions? What is the accuracy of this baseline model?**



In classification tasks, the baseline model commonly used for performance comparison is the majority class model. This model predicts all outcomes based on the class that occurs most frequently in the training data.

**Baseline Model Used:** Majority class model (which predicts the most prevalent class in the dataset).

**Prediction Value Assigned:** This model assigns every prediction to the majority class (in our case “income ≤ 50K” is considered, it would categorize all records under that class).