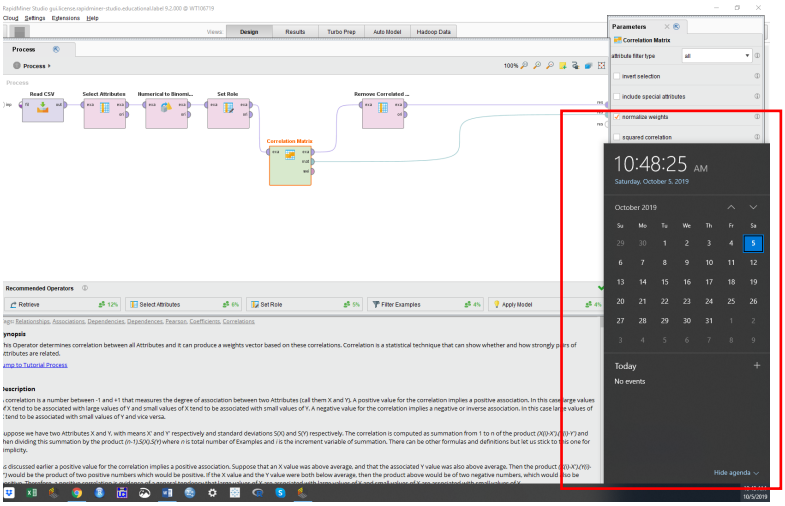
CIDM 6355 Data Mining Methods Exam 1 Part 2 Submission

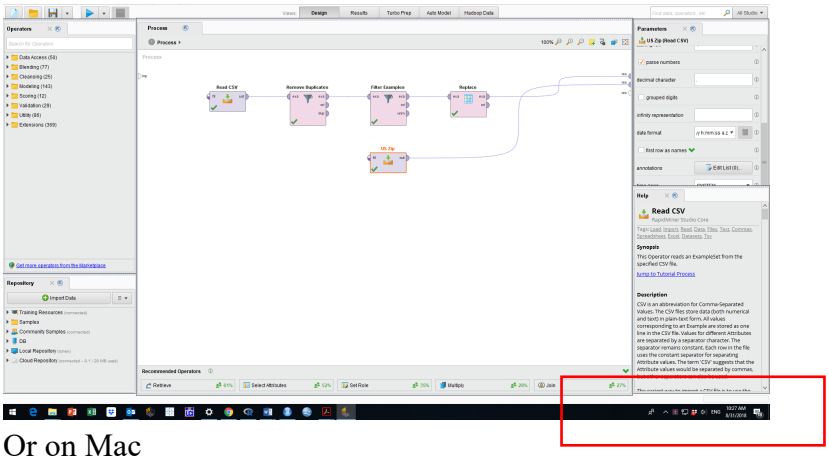
(50 points in total; due 11:59 pm CDT, October 13, 2024)

This exam is open book, open slides, and open notes, but you are not allowed to collaborate nor discuss with anyone else. Sharing your screenshots, RM processes, R script, or answers with other students are considered as cheating for this exam. Should you have any question or unclarity, please contact with the instructor. Please put all your deliverables in a word document and submit it to WT class before the deadline; Make sure that all your screenshots include dates and time [see the examples as below]; otherwise, a penalty of 50% of your grade will be applied. Please type your name below to indicate whether or not you have understood and complied with such requirements in this exam.

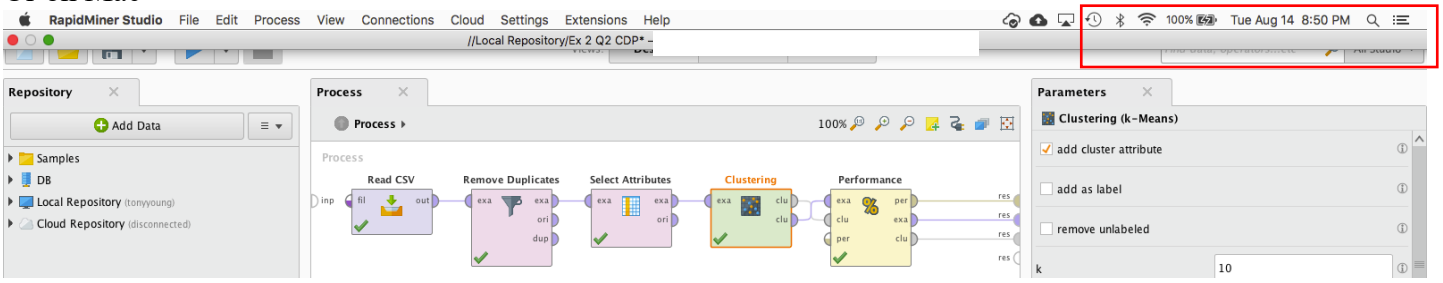
Your name (First Last): Sera Hill



Or



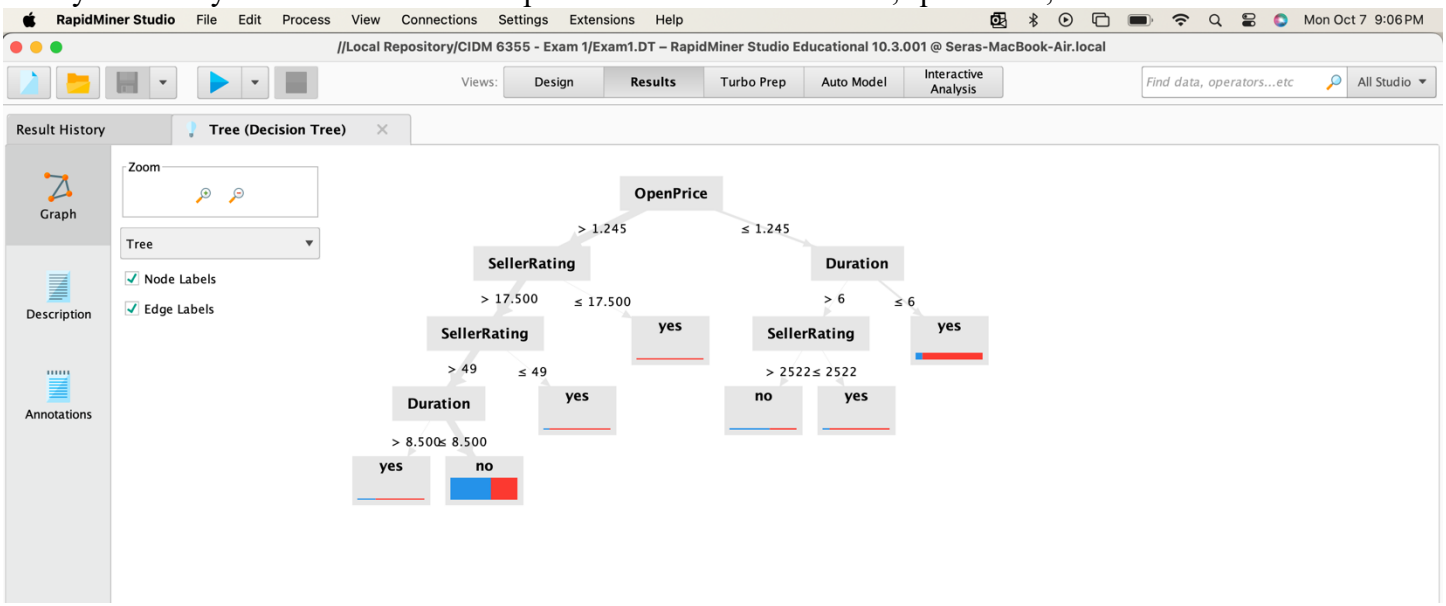
Or on Mac



PLEASE MAKE SURE THAT YOU TYPE YOUR NAME AT THE FIRST PAGE; OTHERWISE, YOUR SUBMISSIN WILL NOT BE GRADED AND A ZERO POINT WILL BE ASSIGNED.

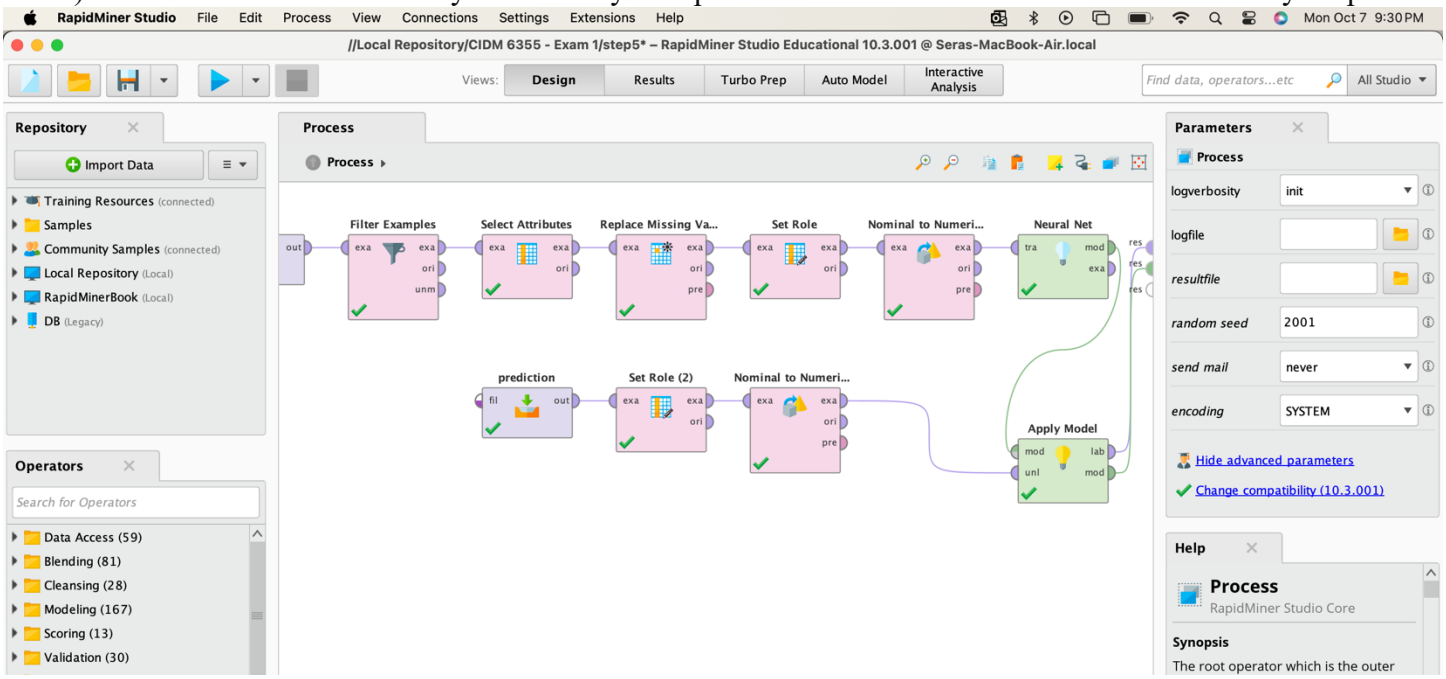
2 Screenshots in RM (6 points for each: 3 pts for your screenshot and 3 pts for your description/discussion).

Screenshot 1 with description (6 pts): A screenshot of your decision tree graph with date and time at Step 2.3 and briefly describe your model. Your description must include root node, split nodes, and leaf nodes.



The root node in this Decision Tree is "OpenPrice". It has 5 split nodes, and 7 leaf nodes.

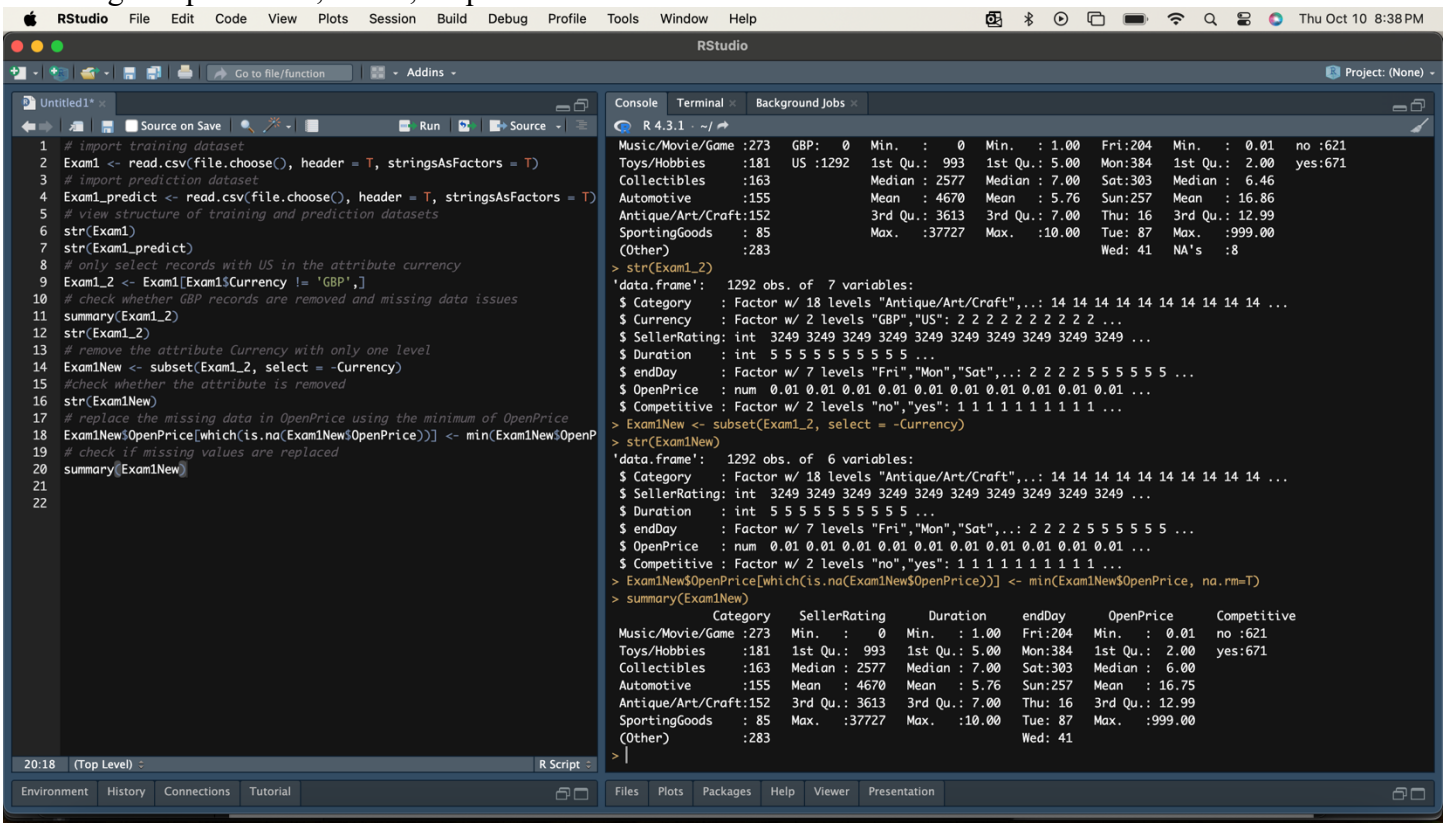
Screenshot 2 with your discussion (6 pts): A screenshot of your RapidMiner Process (the flow chart in your design mode) with date and time and briefly discuss why the operator Nominal to Numerical must be used in your process.



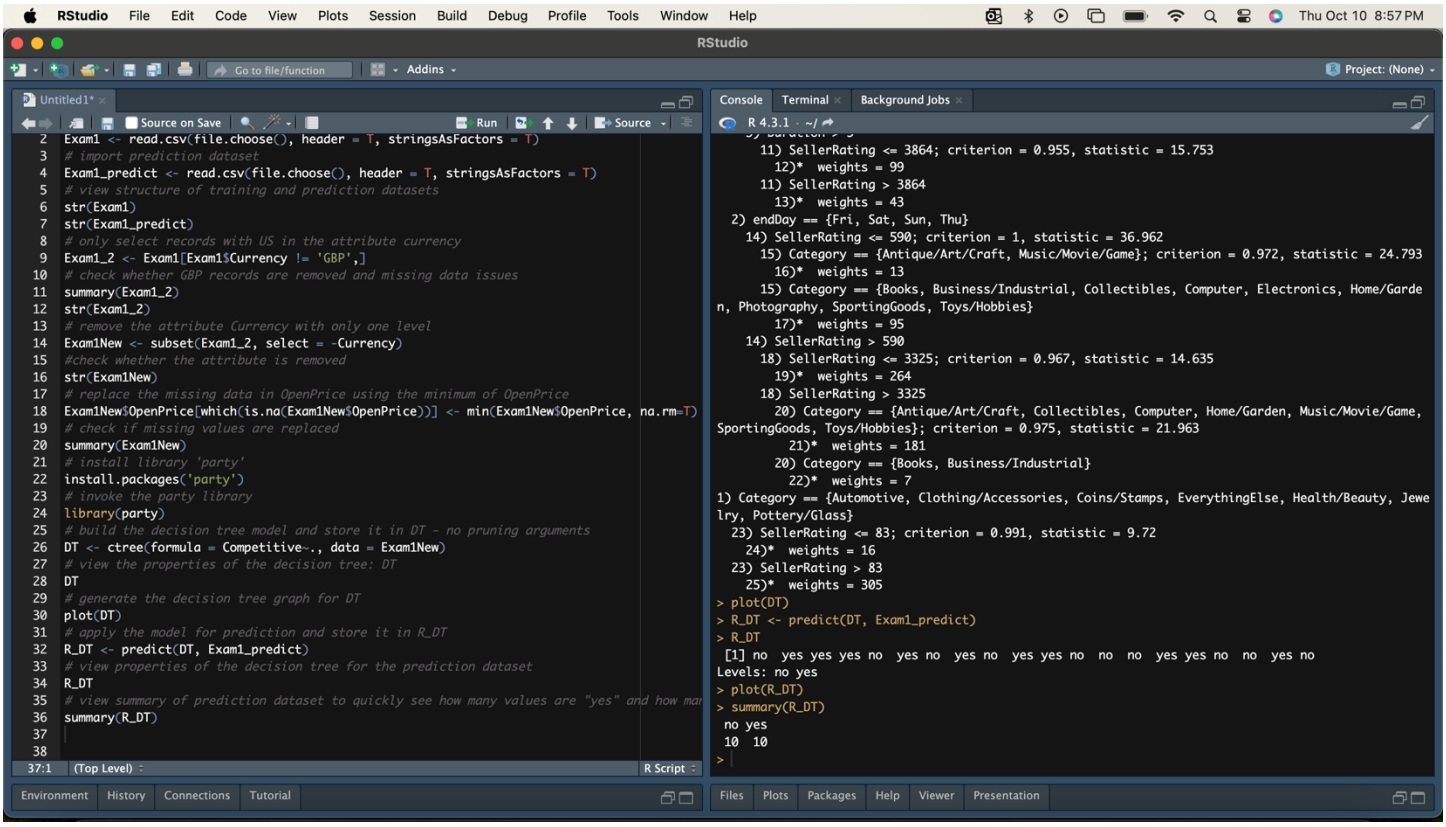
The Nominal to Numerical operator is necessary for this model to convert the categorical variables into numerical dummy variable values because Neural Networks are mathematical in nature and can't directly process non-numerical data.

5 Screenshots in R: Each screenshot must include complete and correct code description or comments (20 pts in total and 4 points for each).

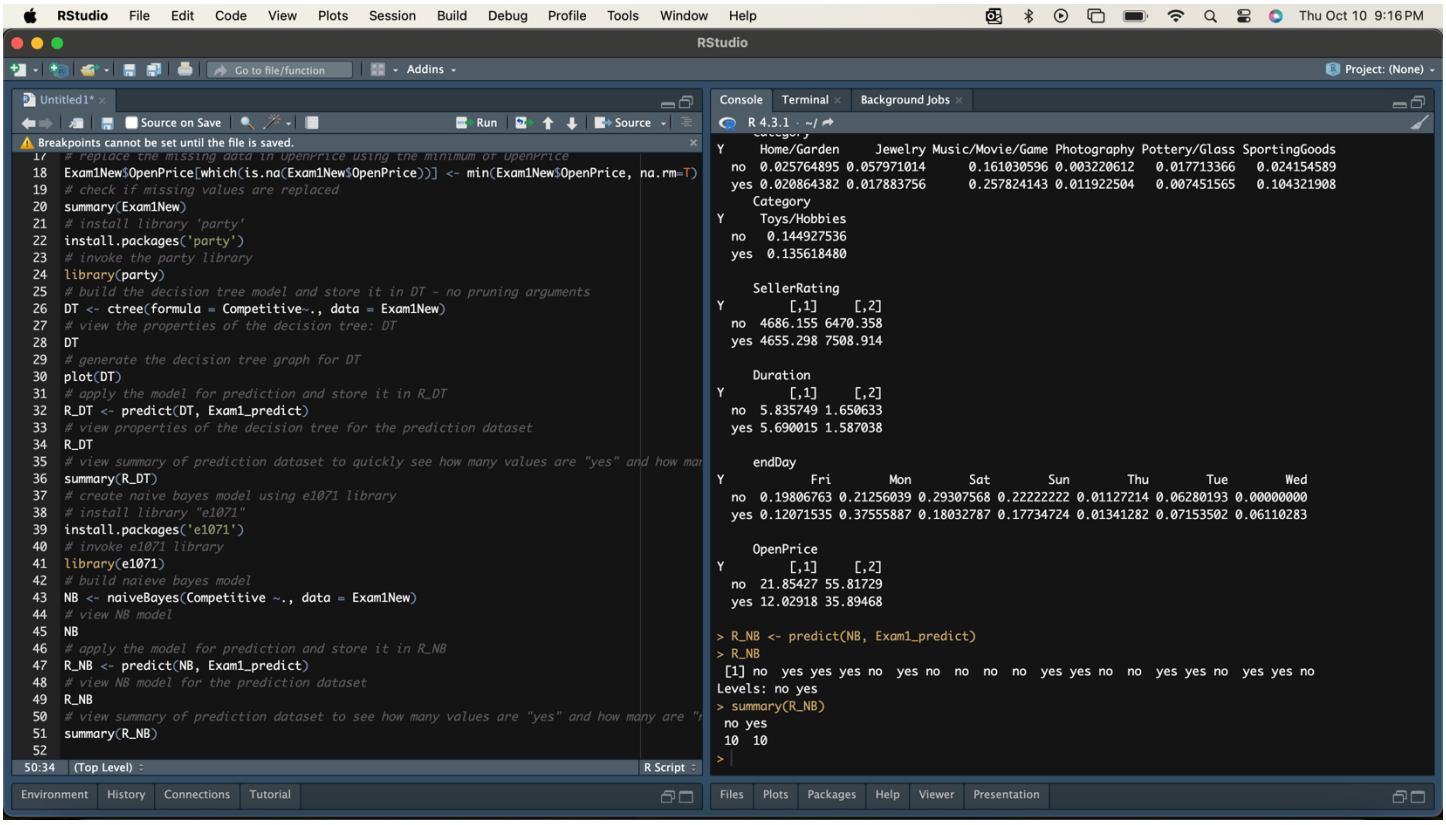
Screenshot 3: A screenshot of your R codes with date and time to show how you import and prepare the data for modeling and prediction, that is, Steps 6.1-6.3.



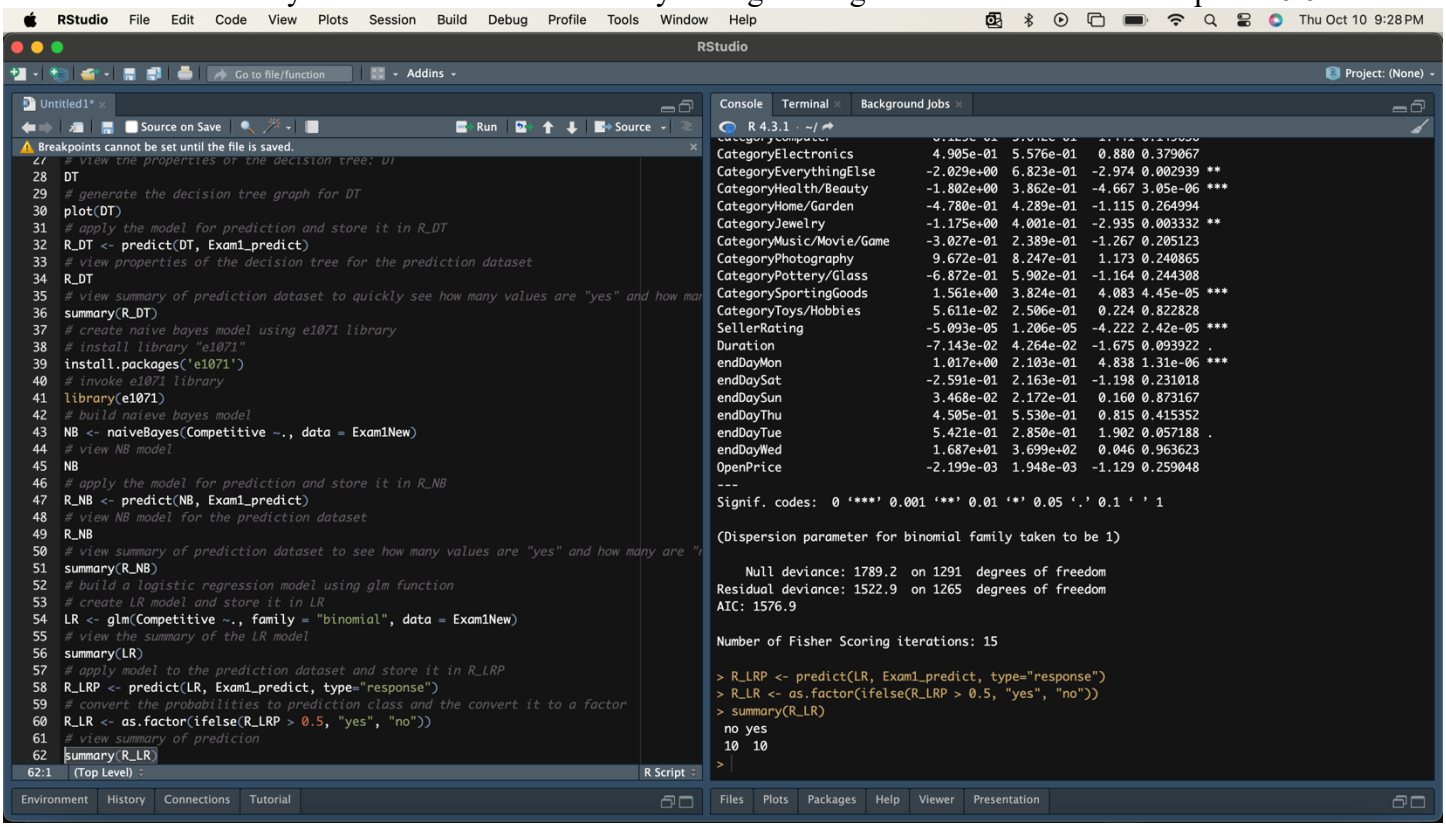
Screenshot 4: A screenshot of your R codes with date and time to show Step 6.4.1-6.4.3. Requirements: your screenshot must clearly include all the R codes for your decision tree model and the output of 6.4.3.



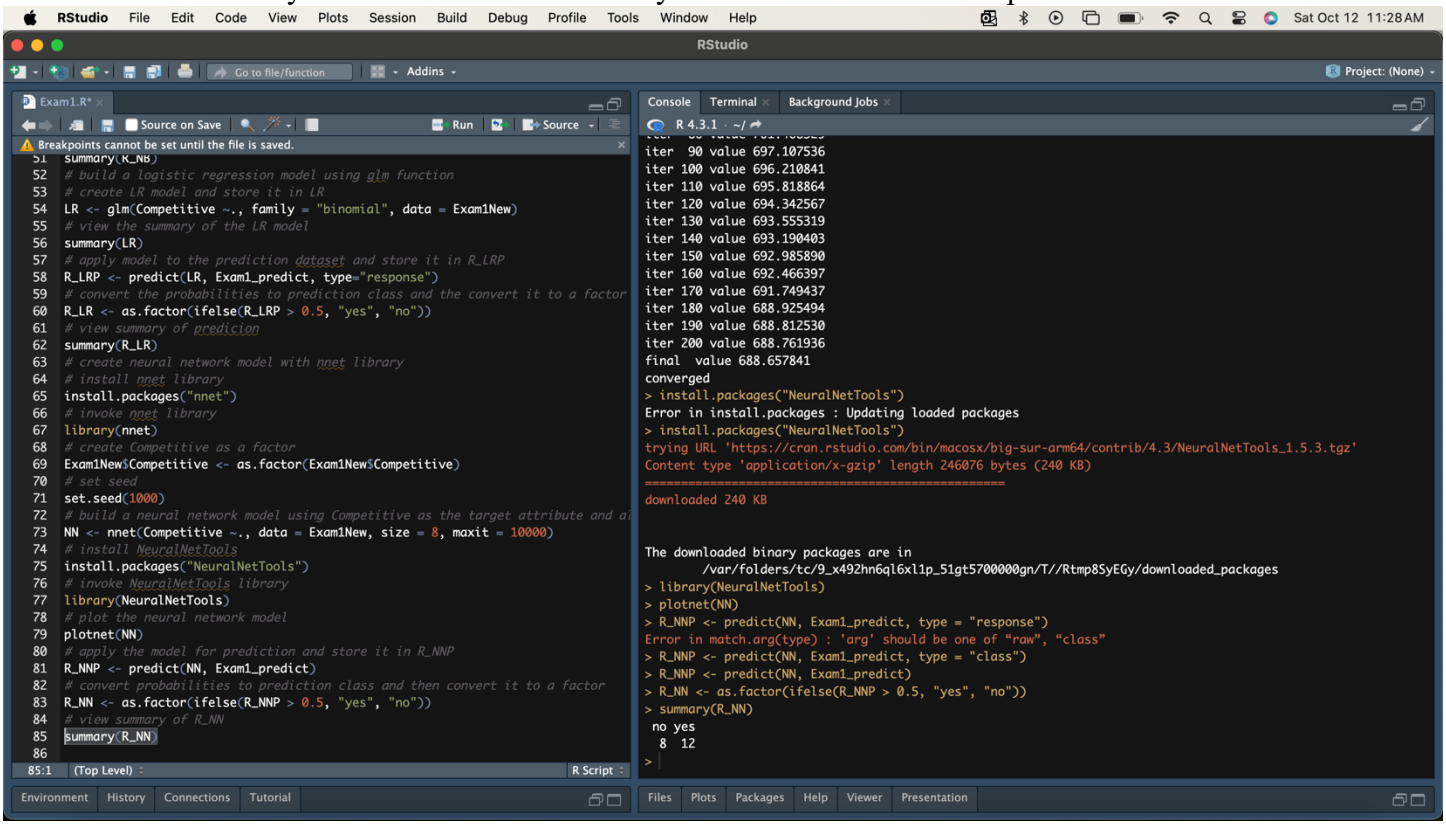
Screenshot 5: A screenshot of your R codes with date and time to show Step 6.5.1-6.5.3. Requirements: your screenshot must clearly include all the R codes for your NB model and the output of 6.5.3.



Screenshot 6: A screenshot of your R codes with date and time to show Step 6.6.1-6.6.4. Requirements: your screenshot must clearly include all the R codes for your logistic regression model and the output of 6.6.4.



Screenshot 7: A screenshot of your R codes with date and time to show Step 6.7.1-6.7.4. Requirements: your screenshot must clearly include all the R codes for your NN model and the output of 6.7.4.



Step 7: Comparative Analysis (18 points)

7.3. Please include the following deliverables in your submission:

7.3.1. Please copy and paste the provided table into your submission (your table must be accessible; screenshot or image is not accepted). Ensure that the table includes the predicted results of 20 records using 8 different methods (8 pts in total and each column is worth 1 point).

RM_DT	RM_NB	RM_LR	RM_NN	R_DT	R_NB	R_LR	R_NN
no	no	no	no	no	no	no	no
yes	yes	yes	yes	yes	yes	yes	no
no	no	yes	yes	yes	yes	yes	yes
yes	yes	yes	yes	yes	yes	yes	yes
yes	yes	no	yes	no	no	no	yes
no	no	yes	no	yes	yes	yes	no
no	no	no	no	no	no	no	no
yes	yes	yes	yes	yes	no	no	yes

no	no	no	yes	no	no	no	no
yes	yes	yes	yes	yes	no	no	yes
no	no	yes	yes	yes	yes	yes	yes
yes	yes	yes	yes	no	yes	yes	no
yes	yes	no	no	no	no	no	yes
no	no	no	yes	no	no	no	yes
yes	yes	yes	yes	yes	yes	yes	yes
yes	yes	yes	yes	yes	yes	yes	no
no	no	no	no	no	no	no	no
yes	yes	yes	yes	no	yes	yes	yes
no	no	yes	yes	yes	yes	yes	yes
no	no	no	yes	no	no	no	yes

7.3.2. Discuss the number of records predicted to be "yes" or “no” by each method in the RM and R datasets. For example, among the 20 records, RM_DT and R_DT jointly predict "yes" for 6 records (ID =2, …..) , and jointly predict "no" for 6 records (ID =1, …….). (8 pts: 2 pts for each pair of methods)

For the 20 records, RM_DT and R_DT jointly predicted “yes” for 6 records, IDs = 2, 4, 8, 10, 15, 16. They jointly predicted “no” for 6 records, IDs = 1, 7, 9, 14, 17, 20.

For the 20 records, RM_NB and R_NB jointly predicted “yes” for 6 records, IDs = 2, 4, 12, 15, 16, 18. They jointly predicted “no” for 6 records, IDs = 1, 7, 9, 14, 17, 20.

For the 20 records, RM_LR and R_LR jointly predicted “yes” for 10 records, IDs = 2, 3, 4, 6, 11, 12, 15, 16, 18, 19. They jointly predicted “no” for 8 records, IDs = 1, 5, 7, 9, 13, 14, 17, 20.

For the 20 records, RM_NN and R_NN jointly predicted “yes” for 11 records, IDs = 3, 4, 5, 8, 10, 11, 14, 15, 18, 19, 20. They jointly predicted “no” for 4 records, IDs = 1, 6, 7, 17.

7.3.3. Finally, provide an analysis of the number of records that all eight models predict as "yes" and the number of records that all eight models predict as "no." For example, all eight models jointly predict as "yes" for 2 records (ID =4…..), and "no" for 4 records (ID =1,……). (2 pts).

All eight models predicted “yes” for 2 records, IDs = 4, 15.

All eight models predicted “no” for 3 records, IDs = 1, 7, 17.