Tutor - DPLYR

La Minh Hieu

2024-01-17

Data: File dplyr_data_training.csv

Bài 1

Import data từ file .csv vào trong R. Phân tích đánh giá biến trong dữ liệu thuộc loại gì, loại dữ liệu đó đã hợp lý chưa?

Bài 2

a) Kiểm tra xem dữ liệu có quan sát nào bị trùng không? Hai quan sát bị trùng là hai quan sát mà tất cả các giá trị trong hai quan sát đó đều giống hệt nhau.

b) Xoá 2 cột Naive_Bayes ở cuối data bằng hai cách: dùng package dplyr và không dùng package dplyr.

c) View các dòng từ 9409 đến 9430 của file data.

Bài 3

a) Sử dụng lệnh apply() để đổi các vector "Attrition_Flag", "Gender", "Education_Level", "Marital_Status", "Income_Category", "Card_Category" sang dạng factor.

b) Sử dụng lệnh apply() để đổi các vector Total_Relationship_Count và Months_Inactive_12_mon thành dạng character, sau đó đổi ngược lại thành numeric.

Bài 4

Liệt kê các giá trị khác nhau của các vector Customer_Age, Attrition_Flag, Gender, Education_Level, Martial_Status.

Bài 5

a) Thêm cột ID vào ngoài cùng bên trái của dữ liệu, đánh số từ 1 cho đến hết số dòng của data.

b) Thêm cột Credit_Limit_Book với giá trị bằng Credit_Limit/Book_on_value, đổi vị trí cột mới được tạo thành vào ngay bên phải cột Credit_Limit.

Bài 6

a) Xem các quan sát mà giá trị Credit_Limit_Book của nó lớn hơn giá trị trung bình của cột Credit_Limit_Book.

b) Câu hỏi tương tự, nhưng sử dụng lệnh filter() trong package dplyr để thực hiện.

Bài 7

a) Tạo dataframe dat_select với dữ liệu từ cột CLIENTNUM đến cột Avg_Utilization_Ratio

b) Trong dataframe dat_select, đổi chỗ các cột từ Education_Level đến Marital_Status ra trước cột Dependent_count

c) Sắp xếp lại dữ liệu sao cho biến Credit_Limit có xu hướng tăng dần/giảm dần.

d) Lưu dữ liệu dat_select vào trong thư mục bất kỳ trên máy tính của mình.

Bài 8

a) Viết function tính mode của một vector (Gợi ý: Sử dụng hàm which.max), sau đó sử dụng hàm summarise để tính mean, median, mode của các biến Credit_Limit và Customer_Age.

b) Tính mean, median, mode của các biến Credit_Limit, Customer_Age, Total_Revolving_Bal theo các nhóm trong cột Gender hoặc Card_Category, sau đó lưu kết quả vào biến dat_summarise_card_cate đối với bảng theo nhóm Card_Category, biến dat_summarise_gender đối với bảng theo nhóm Gender.

c) Tính Trung bình, Trung vị, Mốt, tứ phân vị thứ nhất, tứ phân vị thứ ba của biến Months_on_book dựa theo các nhóm của biến Education_Level, Marital_Status.

d) Tính mean, median, mode của hai vector Total_Trans_Amt, Credit_Limit, Avg_Open_To_Buy dựa trên hai nhóm Income_Category, Gender.

e) Tính mean, mode, median của vector Avg_Total_Trans_Month, được tính bằng cách lấy vector Total_Trans_Amt chia cho vector Total_Trans_Ct, phân nhóm theo vector Gender và Income_Category.