Module 4: Data Manipulation

Case Study

edureka!

Things you will learn in this case-study: You will learn the basics of verbs in dplyr

1. select(), which returns a subset of the columns,
2. filter(), that is able to return a subset of the rows,
3. arrange(), that reorders the rows according to single or multiple variables,
4. mutate(), used to add columns from existing data,
5. summarise(), which reduces each group to a single row by calculating aggregate measures.

Back Ground:

Real Estate Companies like Godrej Reality spend billions on customer acquisition and retention. This spending is in the areas of marketing, advertisement, sales staff etc. They are facing a lot of competition in the sector and with all the regulatory compliances coming into place, the task has become even more daunting.
As a result, these companies are using analytics to understand their data about customers and transactions to have a focused approach towards the entire process. This is where they require your support.

Overview of the problem:

In this project, you will play the role of Data Scientist who would be carrying out Descriptive Analytics for the company and come up with insights about What the Data is talking about?
The Data you will be dealing is with Transactions Data captured for the Company. The data is in data frame format which has been stored in the file named 'housingdata.csv'

Data set description:

Record: Record ID
Gender: Gender of the applicant.
No_kids: Number of kids.
Education: Education level of the individual.
HasCar: whether the individual has car or not.
Income: income level of the individual.
PropertyValue: Property Value of the Flat / House (in USD)
Loan_Period: period of the loan (in months)
Credit_Record: value of 1, if the Credit Record is good and 0, if it is not.
Housing_type: category of housing property – (Affordable, Mid-Range and Premium)

Property_Purchased: takes the value 'Y', if the inquiry led to Purchase of the property, and 'N' if the property was not Purchased.

Objective:

Based on the knowledge you acquired in Module 4, you are expected to complete the below mentioned activities by using the dplyr function, which is the most widely used library in R for data manipulation.

You should do the following:

1. Load the required libraries and the data.

2. Understand the data structure and provide concise summary on the following –
   - no of observations
   - total number of variables
   - number of continuous variables
   - number of categorical variables

3. Select and Mutate : use the select() and mutate() functions in R to answer the following
   - Select the columns Gender, Education, and Income and print the first five rows
   - Select the columns from Gender to Loan Period and print the first five rows
   - Be concise! - select columns by removing Record Column and Gender and print the first five rows
   - Use mutate() function to add the new variables var1 which calculates the ratio of property value to total income and save the result as g1. Print the first five rows.
   - Add the new variable var2 which returns the ratio of property value to loan period and save the result as g2. Print the first five rows.

4. Filter and Arrange:
   - Filter all the observations that have Property Value lower than 80000 or higher than 150000 and store it in df g3. Print the first five rows. How many observations are there.
   - Filter all the observations that have Property Value > 1000000 and Income < 3185 and store it in df g4. Print the first five rows. How many observations are there.
   - Filter all observations where Income < 3185 and still Property was purchased. How many such records are there in the data set. Print the first five rows.

- Use the arrange() function in dplyr to -:
- Create a data frame by the name 'bought' – which includes observations when the Property was purchased. How many observations are there.
- Arrange the data frame bought by Income and print the first five rows.
- Arrange the data frame bought by Gender and print the first five rows.
- Arrange the data frame bought so that Gender and Education is grouped and print the first five rows.
- Create a data frame by the name 'notbought' – which includes observations when the Property was not purchased. How many observations are there.
- Arrange the data frame notbought by Income and print the first five rows.
- Arrange the data frame notbought by Gender and print the first five rows.
- Arrange the data frame notbought so that Gender and Education is grouped and print the first five rows.
- Reverse the order of arranging - Arrange the housing data according to State and decreasing Income. Print the first five rows.

5. Summarise function:
    - Print out a summary with variables min_income and max_income.
    - Generate summary statistics about Income column of housing dataframe. The summary should print minimum, maximum, average, standard deviation, and IQR of the variable.
    - Generate summary about PropertyValue column of housing. The output should print minimum, maximum, average, standard deviation, and IQR of the variable.
    - Generate summary about Loan_Period column of housing. The output should print minimum, maximum, average, standard deviation, and IQR of the variable.

6. the pipe operator of dplyr: reproduce the below steps using dplyr and pipe operator
    - Start with the housing data set and then
    - Add the new variable var1 which calculates the ratio of property value to total income
    - Pick all-of the rows whose var1 value exceeds 50, and then
    - Summarize the data set with a value named avg. that is the mean value of var1.
    - Finally report the output of the above steps.

7. using group_by function of dplyr: reproduce the below steps
    - Start with the housing data set and then

- Use group_by() to group housing by Education.
- summarise() the grouped df with two summary variables: avg_income, the average of Income, and avg_Value, the average value of purchased property.
- Finally, order the summary from low to high by these two summarized variables
- Finally report the output of the above steps.

Submission should include the following:

1. Answers to the above questions. Print the first five rows as output wherever applicable.
2. Summary on approach should be documented and submitted for each question.
3. R Code File.