San José State University
Department of Applied Data Science

DATA 225
Database Systems for Analytics

Fall 2023
Section 21
Instructor: Ron Mak

Assignment #8

Assigned:	Monday, October 23
Due:	Monday, October 30 at 5:30 pm
	Team assignment, 170 points max

Missing values

The purpose of this assignment is to give you practice replacing missing values in your source data using SQL operations on the database server. Choose a multidimensional dataset (i.e., one with multiple variables) that you find interesting. There should be at least four variables — if your chosen dataset has more than four variables, pick four of them for this assignment.

Tip: Google “datasets for regression analysis”.

Unless the dataset already has missing values, artificially create missing values. Choose one of variables as your target variable (such as age in the Titanic data), and randomly remove some of its values, up to about 15%. Load your dataset including the missing values into database table(s).

Part A: Replace with average values

Use the Titanic example from class as an example to perform these steps:

1. On the client side, without the records containing the missing values, calculate and print a pairwise correlation matrix of your four variables.
2. On the server side, without the missing values, use SQL to calculate the overall average of your target variable and its averages within the major subgroups (such as the Titanic passenger classes). Query for and print the averages on the client side.
3. On the server side, use SQL to calculate the percentages of missing target variable values in ever smaller subgroups. Query for and print the percentages on the client side.
4. On the client side, determine the smallest subgroup whose averages you can use to replace the missing values.
5. On the server side, use SQL code to replace each missing value with the appropriate average.

6. Download the cleaned data and redo steps 1 and 2. Note any changes in the results.

Part B: Replace using multiple regression

Use the SQL code from class to perform multiple regression. Use your target variable with missing values as the dependent variable and the remaining three variables as the independent variables to perform these steps:

1. On the client side, without the records containing the missing values, calculate and print a pairwise correlation matrix.
2. On the server side, use SQL to calculate the linear regression coefficients for the data without the missing values. Query for and print the coefficients on the client side.
3. On the server side, use SQL code to replace each missing value with an estimate calculated from the regression equation.
4. Download the cleaned data and redo step 1. Note any changes in the results.

What to submit

Two Jupyter notebooks, one for Part A and one for Part B.

Rubric

Criteria	Max points
• **Part A**	• **100**
○ Step 1	○ 10
○ Step 2	○ 20
○ Step 3	○ 20
○ Step 4	○ 10
○ Step 5	○ 20
○ Step 6	○ 20
• **Part B**	• **70**
○ Step 1	○ 10
○ Step 2	○ 25
○ Step 3	○ 25
○ Step 4	○ 10