

# [90 min] DO: Assignment 5 / Query Databases and Dataframes in R with SQL

---

**Due** Feb 26 by 11:59pm    **Points** 100    **Submitting** a file upload  
**File Types** rmd, pdf, and zip    **Available** until Feb 28 at 5:59pm

---

This assignment was locked Feb 28 at 5:59pm.

SQL has emerged as a generic query language on virtually any tabular data, including data frame in R, pandas in Python, and even Excel worksheets. This assignment provides you some exposure to querying databases from R and also using SQL to query non-database tabular objects using the **sqldf** package; it is quite simple and explained in [First Steps in R](#). Another objective for this assignment is to require that you learn some R as we will need that for the first practicum that is coming up.

## Format


- Group or individual. Group collaboration is encouraged and permitted, although individual submission is required.

## Resources

- [First Steps in R](#)
- [Lecture Notes on R and Examples](#)

## Instructions

Create an R Notebook and do the following tasks in separate code chunks:

1. (20 pts) In the R Notebook, connect to the SQLite [MediaDB.db](#)  ([https://northeastern.instructure.com/courses/63394/files/6999739/download?download\\_frd=1](https://northeastern.instructure.com/courses/63394/files/6999739/download?download_frd=1)) database and then load, using SQL SELECT, the "invoice\_items" table into a data frame called *rs*. Add a new column to *rs* for the extended price called *ExtPrice* that is *Quantity* times *Price*. Using R, what is the average extended price (rounded to 2 decimals)? Do not use {sql} chunks for this.
2. (30 pts) Using **sqldf**, write a SQL query against the data frame *rs* from the question above that finds the total amount for each invoice (i.e., the sum of the extended prices for the *invoice\_items* in each invoice) and the number of items in the invoice. So, the result set contains rows that each have the invoice ID, the total, and the number of items.
3. (30 pts) Using R and the result from the prior question, create a scatter plot of the total number of items in an invoice (x axis) versus the total (y axis). Add proper axis labels.
4. (20 pts) Write and execute a SQL UPDATE statement that applies a 10% discount to the total amount for each invoice if it has more than 5 items and stores that discounted amount in a new column called *DiscPrice*. Using a separate `{r}` chunk show that the query executed properly by

displaying a part of the table. If you cannot use **sqldf**, then apply to the database directly instead; either are acceptable as long as you use a SQL UPDATE.

## Submission Details

- Submit the R Notebook (*.Rmd* file). Your code must execute to receive any points.