# Lesson 2 Assignment: SQL & Data Wrangling

## Case Study

**Background:** A retail firm (AdventureWorks Inc.) wants to move their Research and Development unit to a different location. The idea is to keep the unit small, agile and independent of "corporate pressures". You are given data files with employee ids and addresses. Your goal is to get a list of employees in research functions and the city where they live. This will help you suggest a location that is close to most of the employees.

## Assignment Tasks

### Datasources

For this assignment, assume that your datasources are the following (for just 1 plant in US):

1. Employee : CSV file named Employee . Contains list of employees (IDs, Login, Date of Birth, Gender etc.)
2. BusinessEntityAddress : CSV file named BusinessEntityAddress. Contains a list of employee IDs and their their AddressesID
3. Address : CSV file named Address. Contains a list of AddressesID and corresponding street, city, zip code etc.

If you are using your own installation of R studio download the CSV files from the assignment description in the course site. If you are using R studio Virtual lab use the location below to use the CVS files in the R Studio Virtual lab just like you did in the R lab exercise

[**https://library.startlearninglabs.uw.edu/DATAAVS210/L02\_HW/**](https://library.startlearninglabs.uw.edu/DATAAVS210/L02_HW/)

### Instructions

Perform the following using R, SQL and/or Excel as prompted. You will be required to submit your code and you Excel workbook for grading.

1. Load the two CSV files in R using fread function from data.table package. They can have any name in R (example, dt1, dt2) but for simplicity sake, call them Employee, BusinessEntityAddress and Address.
2. Assume the 3 files are database tables (with same names as their filename), write a SQL query that does all the following:
   * JOINS the three tables into 1 dataset (you can use either LEFT OUTER or INNER join or a combination of the two for this assignment)
   * Filters the data for Job Titles that start with the word Research (Hint: lookup like keyword and wildcard (%))
   * The resulting dataset should have these columns: BusinessEntityID, LoginID, JobTitle, City
3. Use the sqldf package to run the sql from the step above.
4. Submit your R script, the output, and the answer to the Question in a word document.

### Question

Based on the information you obtained after wrangling the data. What location would you recommend to move the Research and development unit?