SQL For Data Science with R Final Project

Myron Keith Gibert Jr

2022-10-25

Contents

Introduction

Imagine you have just been hired by a US Venture Capital firm as a data scientist.

The company is considering foreign grain markets to help meet its supply chain requirements for its recent investments in the microbrewery and microdistillery industry, which is involved with the production and distribution of craft beers and spirits.

Your first task is to provide a high level analysis of crop production in Canada. Your stakeholders want to understand the current and historical performance of certain crop types in terms of supply and price. For now they are mainly interested in a macro-view of Canada's crop farming industry, and how it relates to the relative value of the Canadian and US dollars.

You will be asked questions that will help you understand the data just like a data analyst or data scientist would. You will also be asked to create four tables in Db2, and load the tables using the provided datasets from R using the RODBC package. You will be assessed both on the correctness of your SQL queries and results, as well as the correctness of your table creation and data loading results.

An R based Jupyter notebook has been provided to help with completing this assignment. Follow the instructions to complete all the problems. Then share your solutions with your peers for reviewing.

Objectives

Understand four datasets Load the datasets into four separate tables in a Db2 database *Execute SQL queries using the RODBC R package to answer assignment questions* You have already encountered two of these datasets in the previous practice lab, and you will be able to reuse much of the work you did there to successfully prepare your database tables for executing SQL queries.

Setup

You can download the DB2 driver for you computer here.

```
if (!require("tidyverse")) install.packages("tidyverse")
library("tidyverse")

if (!require("ggplot2")) install.packages("ggplot2")
library("ggplot2")

if (!require("RODBC")) install.packages("RODBC")
library("RODBC")

### Formatting###

if (!require("formatR")) install.packages("formatR")
library("formatR")

knitr::opts_chunk$set(echo = TRUE, tidy = TRUE, tidy.opts = list(width.cutoff = 60))

### DB Connection###

driver.name <- "DB2"
db.name <- "BLUDB"
host.name <- "19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0t
  gtu0lqde00.databases.appdomain.cloud"
port <- "30699"  # 50000 if not using SSL or 50001 if using SSL
user.name <- "vkj26480"
user.pwd <- "PjtfHxJhreFiR28i"
```

```
con.text <- paste("CData DB2 Source;DRIVER=", driver.name, ";Database=",
    db.name, ";Hostname=", host.name, ";Port=", port, ";PROTOCOL=TCPIP",
    ";UID=", user.name, ";PWD=", user.pwd, sep = "")

conn <- odbcConnect(con.text)

knitr::opts_chunk$set(connection = "conn")
```

Exercise 1 : Understand the datasets

To complete the assignment problems in the notebook you will be using subsetted snapshots of two datasets from Statistics Canada, and two small datasets created from a third datasource from the Bank of Canada. The links to the prepared datasets are provided in the next section; interested students can explore the landing pages for the source datasets as follows:

1.Canadian Principal Crops (Data & Metadata)

2.Farm product prices (Data & Metadata)

3.Bank of Canada daily average exchange rates

1. Canadian Principal Crops Data *

This dataset contains agricultural production measures for the principle crops grown in Canada, including a breakdown by province and teritory, for each year from 1908 to 2020.

For this assignment you will use a preprocessed snapshot of this dataset (see next section for the link).

A detailed description of this dataset can be obtained from the StatsCan Data Portal at: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3210035901 Detailed information is included in the metadata file and as header text in the data file, which can be downloaded - look for the 'download options' link.

2. Farm product prices

This dataset contains monthly average farm product prices for Canadian crops and livestock by province and teritory, from 1980 to 2020.

For this assignment you will use a preprocessed snapshot of this dataset (see next section for the link).

A description of this dataset can be obtained from the StatsCan Data Portal at: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3210007701 The information is included in the metadata file, which can be downloaded - look for the 'download options' link.

3. Bank of Canada daily average exchange rates *

This dataset contains the daily average exchange rates for multiple foreign currencies. Exchange rates are expressed as 1 unit of the foreign currency converted into Canadian dollars. It includes only the latest four years of data, and the rates are published once each business day by 16:30 ET.

For this assignment you will use a snapshot of this dataset with only the USD-CAD exchange rates included (see next section). We have also prepared a monthly averaged version which you will be using below.

A brief description of this dataset and the original dataset can be obtained from the Bank of Canada Data Portal at: https://www.bankofcanada.ca/rates/exchange/daily-exchange-rates/

(* these datasets are the same as the ones you used in the practice lab)

Dataset URLs

Annual Crop Data: Annual_Crop_Data.csv

Daily FX Data: Daily_FX.csv

Monthly Farm Prices: Monthly_Farm_Prices.csv

Monthly FX Data: Monthly_FX.csv

IMPORTANT: You will be loading these datasets directly into R data frames from these URLs instead of from the StatsCan and Bank of Canada portals. The versions provided at these URLs are simplified and subsetted versions of the original datasets.

Exercise 2 : Load these datasets into four separate Db2 tables.

In this exercise, you will prepare the database so you can solve problems using SQL in the last portion of the assignemnt, Exercise 3. You will create four tables and load the datasets into them.

Problem 1: Create tables

Establish a connection to the Db2 database, and create the following four tables using the RODBC package in R.

1.CROP_DATA

2.FARM_PRICES

3.DAILY_FX

4.MONTHLY_FX

The previous practice lab will help you accomplish this.

```
sqlQuery(conn, "CREATE TABLE CROP_DATA (
    CD_ID int,
    YEAR DATE,
    CROP_TYPE varchar(255),
    GEO varchar(255),
    SEEDED_AREA int,
    HARVESTED_AREA int,
    PRODUCTION int,
    AVG_YIELD int
)")

sqlQuery(conn, "CREATE TABLE FARM_PRICES (
    CD_ID int,
    DATE DATE,
    CROP_TYPE varchar(255),
    GEO varchar(255),
    PRICE_PRERMT int
)")

sqlQuery(conn, "CREATE TABLE DAILY_FX (
    DFX_ID int,
    DATE DATE,
    FXUSDCAD DEC(1,4)
)")
```

```
sqlQuery(conn, "CREATE TABLE MONTHLY_FX (
    DFX_ID int,
    DATE DATE,
    FXUSDCAD DEC(1,6)
)")
```

```
##   TABLE_CAT  TABLE_NAME
## 1     BLUDB   CROP_DATA
## 2     BLUDB    DAILY_FX
## 3     BLUDB FARM_PRICES
## 4     BLUDB  MONTHLY_FX
```

Problem 2: Read Datasets and Load Tables

You will read the datasets directly into R dataframes using the urls provided above, and use these to load the tables you created.

```
CROP_DATA <- read.csv("https://cf-courses-data.s3.us.cloud-object-storage.
                      appdomain.cloud/IBM-RP0203EN-SkillsNetwork/labs/
                      Final%20Project/Annual_Crop_Data.csv")

sqlSave(conn, CROP_DATA, tablename = "CROP_DATA", append = FALSE,
    rownames = FALSE, colnames = FALSE, safer = FALSE, fast = FALSE)

FARM_PRICES <- read.csv("https://cf-courses-data.s3.us.cloud-object-storage.
                        appdomain.cloud/IBM-RP0203EN-SkillsNetwork/labs/
                        Final%20Project/Monthly_Farm_Prices.csv")

sqlSave(conn, FARM_PRICES, tablename = "FARM_PRICES", append = FALSE,
    rownames = FALSE, colnames = FALSE, safer = FALSE, fast = FALSE)

DAILY_FX <- read.csv("https://cf-courses-data.s3.us.cloud-object-storage.
                     appdomain.cloud/IBM-RP0203EN-SkillsNetwork/labs/
                     Final%20Project/Daily_FX.csv")

sqlSave(conn, DAILY_FX, tablename = "DAILY_FX", append = FALSE,
    rownames = FALSE, colnames = FALSE, safer = FALSE, fast = FALSE)

MONTHLY_FX <- read.csv("https://cf-courses-data.s3.us.cloud-object-storage.
                       appdomain.cloud/IBM-RP0203EN-SkillsNetwork/labs/
                       Final%20Project/Monthly_FX.csv")

sqlSave(conn, MONTHLY_FX, tablename = "MONTHLY_FX", append = FALSE,
    rownames = FALSE, colnames = FALSE, safer = FALSE, fast = FALSE)

## Show data frames

ls()[sapply(ls(), function(i) class(get(i))) == "data.frame"]

## Show Tables in Database

data <- sqlTables(conn) %>%
    filter(TABLE_SCHEM == "VKJ26480") %>%
    select(TABLE_CAT, TABLE_NAME)
```

```
data
```

```
## [1] "CROP_DATA"   "DAILY_FX"    "data"        "FARM_PRICES" "MONTHLY_FX"
```

```
##   TABLE_CAT  TABLE_NAME
## 1     BLUDB   CROP_DATA
## 2     BLUDB    DAILY_FX
## 3     BLUDB FARM_PRICES
## 4     BLUDB  MONTHLY_FX
```

Exercise 3: Execute SQL queries using the RODBC R package

Problem 3: How many records are in the farm prices dataset?

ANSWER: There are 2,678 records in the Farm Prices Dataset.

```
sqlQuery(conn, "SELECT COUNT(*) FROM FARM_PRICES")
```

```
##      1
## 1 2678
```

Problem 4: Which provinces are included in the farm prices dataset?

ANSWER: Alberta and Saskatchewan are included in the farm prices dataset.

```
sqlQuery(conn, "SELECT DISTINCT GEO FROM FARM_PRICES")
```

```
##            GEO
## 1      Alberta
## 2 Saskatchewan
```

Problem 5: How many hectares of Rye were harvested in Canada in 1968?

ANSWER: 274,100 hectares of Rye were harvested in Canada in 1968.

```
sqlQuery(conn, "SELECT HARVESTED_AREA FROM CROP_DATA
        WHERE YEAR='1968-12-31' and CROP_TYPE='Rye' and GEO='Canada'")
```

```
##   HARVESTED_AREA
## 1         274100
```

Problem 6: Query and display the first 6 rows of the farm prices table for Rye.

ANSWER: Below are the first six rows of the farm_prices table for Rye.

```
sqlQuery(conn, "SELECT * FROM FARM_PRICES WHERE CROP_TYPE='Rye' LIMIT 6")
```

```
##   CD_ID       DATE CROP_TYPE          GEO PRICE_PRERMT
## 1     4 1985-01-01       Rye      Alberta       100.77
## 2     5 1985-01-01       Rye Saskatchewan       109.75
## 3    10 1985-02-01       Rye      Alberta        95.05
## 4    11 1985-02-01       Rye Saskatchewan       103.46
## 5    16 1985-03-01       Rye      Alberta        96.77
## 6    17 1985-03-01       Rye Saskatchewan       106.38
```

Problem 7: Which provinces grew Barley?

ANSWER: Alberta and Saskatchewan grew Barley, and so did the Country of Canada as a whole.

```
sqlQuery(conn, "SELECT DISTINCT GEO FROM CROP_DATA WHERE CROP_TYPE='Barley'")
```

```
##            GEO
## 1      Alberta
## 2       Canada
## 3 Saskatchewan
```

Problem 8: Find the first and last dates for the farm prices data.

ANSWER: The table contains information on farm prices from 1965-2020.

```
sqlQuery(conn, "SELECT MAX(DATE),MIN(DATE) FROM FARM_PRICES")
```

```
##            1          2
## 1 2020-12-01 1985-01-01
```

Problem 9: Which crops have ever reached a farm price greater than or equal to $350 per metric tonne?

ANSWER: Canola is the only crop that has ever reached a farm price greater than or equal to $350 per metric tonne.

```
sqlQuery(conn, "SELECT DISTINCT CROP_TYPE FROM FARM_PRICES WHERE PRICE_PRERMT >= 350")
```

```
##   CROP_TYPE
## 1    Canola
```

Problem 10: Rank the crop types harvested in Saskatchewan in the year 2000 by their average yield. Which crop performed best?

ANSWER: Barley performed best, with an average yield of 2800.

```
sqlQuery(conn, "SELECT * FROM CROP_DATA
    WHERE GEO = 'Saskatchewan' and YEAR = '2000-12-31'
    ORDER BY AVG_YIELD DESC")
```

```
##   CD_ID       YEAR CROP_TYPE          GEO SEEDED_AREA HARVESTED_AREA PRODUCTION
## 1   422 2000-12-31    Barley Saskatchewan     2063900        1922300    5301600
## 2   431 2000-12-31     Wheat Saskatchewan     6145100        6080300   13411800
## 3   428 2000-12-31       Rye Saskatchewan       66800          46500      97800
## 4   425 2000-12-31    Canola Saskatchewan     2387600        2371500    3424600
##   AVG_YIELD
## 1      2800
## 2      2200
## 3      2100
## 4      1400
```

Problem 11: Rank the crops and geographies by their average yield (KG per hectare) since the year 2000. Which crop and province had the highest average yield since the year 2000?

ANSWER: Barley in Alberta performed best, with an average yield of 4100 in 2013 and 2016 and a average yield sum of 72,465 since 2000.

```
sqlQuery(conn, "
         SELECT CROP_TYPE,GEO,
         AVG(AVG_YIELD) AS TEN_YEAR_AVG_YIELD
         FROM CROP_DATA
         WHERE YEAR >= '2000-12-31'
         GROUP BY CROP_TYPE,GEO
         ORDER BY TEN_YEAR_AVG_YIELD DESC")
```

```
##    CROP_TYPE          GEO TEN_YEAR_AVG_YIELD
## 1     Barley      Alberta               3450
## 2     Barley       Canada               3253
## 3      Wheat      Alberta               3100
## 4     Barley Saskatchewan               2971
## 5      Wheat       Canada               2845
## 6        Rye      Alberta               2683
## 7        Rye       Canada               2543
## 8      Wheat Saskatchewan               2429
## 9        Rye Saskatchewan               2226
## 10    Canola      Alberta               1999
## 11    Canola       Canada               1873
## 12    Canola Saskatchewan               1754
```

Problem 12: Use a subquery to determine how much wheat was harvested in Canada in the most recent year of the data.

ANSWER: 35,183,000 metric tons of wheat were harvested in Canada in the most recent year of the data (2020) from over 10,017,800 hectares of harvested land.

```
sqlQuery(conn, "SELECT * FROM CROP_DATA WHERE GEO='Canada' and CROP_TYPE='Wheat' and YEAR = (SELECT MAX
```

```
##   CD_ID       YEAR CROP_TYPE    GEO SEEDED_AREA HARVESTED_AREA PRODUCTION
## 1   670 2020-12-31     Wheat Canada    10193600       10017800   35183000
##   AVG_YIELD
## 1      3512
```

Problem 13: Use an implicit inner join to calculate the monthly price per metric tonne of Canola grown in Saskatchewan in both Canadian and US dollars. Display the most recent 6 months of the data.

ANSWER: See results below. The price per metric ton (PRICE_PRERMT) is in CAD, and can be converted to USD by dividing by the exchange rate (FXUSDCAD). The exchange rate is the number of Canadian Dollars per US dollar.

```
sqlQuery(conn, "SELECT *,PRICE_PRERMT AS CAD_PRICE,PRICE_PRERMT/FXUSDCAD AS USD_PRICE
     FROM (SELECT FARM_PRICES.DATE,GEO,CROP_TYPE,PRICE_PRERMT,FXUSDCAD
           FROM FARM_PRICES
           INNER JOIN MONTHLY_FX ON FARM_PRICES.DATE = MONTHLY_FX.DATE
           WHERE CROP_TYPE = 'Canola' and GEO = 'Saskatchewan'
           ORDER BY MONTHLY_FX.DATE DESC
           LIMIT 6)")
```

```
##         DATE          GEO CROP_TYPE PRICE_PRERMT FXUSDCAD CAD_PRICE USD_PRICE
## 1 2020-12-01 Saskatchewan    Canola       507.33 1.280771    507.33  396.1130
## 2 2020-11-01 Saskatchewan    Canola       495.64 1.306820    495.64  379.2718
## 3 2020-10-01 Saskatchewan    Canola       474.80 1.321471    474.80  359.2966
## 4 2020-09-01 Saskatchewan    Canola       463.52 1.322810    463.52  350.4056
## 5 2020-08-01 Saskatchewan    Canola       464.60 1.322205    464.60  351.3827
## 6 2020-07-01 Saskatchewan    Canola       462.88 1.349850    462.88  342.9122
```