Lab for Final Project - Data Analytics for Canadian Crop Production Data Set

Estimated time needed: 45 minutes

Assignment Scenario

Congratulations! You have just been hired by a US Venture Capital firm as a data analyst.

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 volatility. 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.

Introduction

Using this R notebook you will:

1. Understand four datasets
2. Load the datasets into four separate tables in a database
3. Execute SQL queries to answer assignment questions

You have already encountered two of these datasets in the previous practice lab. You will be able to reuse much of the work you did there to prepare your database tables for executing SQL queries.

Understand the datasets

To complete the assignment problems in this notebook you will be using subsetted snapshots of two datasets from Statistics Canada, and one from the Bank of Canada. The links to the prepared datasets are provided in the next section; the interested student 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 below).

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 (or 'last year', whichever is greatest).

For this assignment you will use a preprocessed snapshot of this dataset (see below).

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

1. Annual Crop Data: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-RP0203EN-SkillsNetwork/labs/Final%20Project/Annual_Crop_Data.csv
2. Farm product prices: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-RP0203EN-SkillsNetwork/labs/Final%20Project/Monthly_Farm_Prices.csv
3. Daily FX Data: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-RP0203EN-SkillsNetwork/labs/Final%20Project/Daily_FX.csv
4. Monthly FX Data: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-RP0203EN-SkillsNetwork/labs/Final%20Project/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.

Now let's load these datasets into four separate tables.

Let's first load the RSQLite package.

Note: If you encounter a `non-zero exist status`, don't worry as it doesn't affect the functionality of the lab.

In [ ]:
```
install.packages("https://cran.r-project.org/src/contrib/Archive/RSQLite/RSQLite_0.
```

Restart Kernel

After installing the RSQLite package, it is necessary to restart R Kernel. Click **Kernel** > **Restart Kernel** from the main menu. This will register the newly installed packages. Now proceed to load the RSQLite package.

In [5]:
```
library("RSQLite")
```

```
Warning message:
"package 'RSQLite' was built under R version 4.5.2"
```

Problem 1

Create tables

Establish a connection **conn** to the RSQLite database **FinalDB.sqlite**, and create the following four tables.

1. **CROP_DATA**
2. **FARM_PRICES**
3. **DAILY_FX**
4. **MONTHLY_FX**

The previous practice lab will help you accomplish this.

In [6]:
```
# Write your query here



conn<-dbConnect(RSQLite::SQLite(),"FinalDB.sqlite")

dbExecute(conn, "DROP TABLE IF EXISTS CROP_DATA")
dbExecute(conn, "DROP TABLE IF EXISTS FARM_PRICES")
dbExecute(conn, "DROP TABLE IF EXISTS DAILY_FX")
dbExecute(conn, "DROP TABLE IF EXISTS MONTHLY_FX")

df1 <- dbExecute(conn,
                    "CREATE TABLE CROP_DATA (
                                        CD_ID INTEGER NOT NULL,
                                        YEAR DATE NOT NULL,
                                        CROP_TYPE VARCHAR(20) NOT NULL,
                                        GEO VARCHAR(20) NOT NULL,
                                        SEEDED_AREA INTEGER NOT NULL,
                                        HARVESTED_AREA INTEGER NOT NULL,
                                        PRODUCTION INTEGER NOT NULL,
                                        AVG_YIELD INTEGER NOT NULL,
                                        PRIMARY KEY (CD_ID)
                                        )"
                    )
df2 <- dbExecute(conn,"CREATE TABLE FARM_PRICES (
                                    CD_ID INTEGER NOT NULL,
                                    DATE DATE NOT NULL,
                                    CROP_TYPE VARCHAR(20) NOT NULL,
                                    GEO VARCHAR(20) NOT NULL,
                                    PRICE_PRERMT FLOAT(6) NOT NULL,
                                    PRIMARY KEY (CD_ID)
                 )"
                    )
df3 <- dbExecute(conn, "CREATE TABLE DAILY_FX (
                DFX_ID INTEGER NOT NULL,
                DATE DATE NOT NULL,
                FXUSDCAD FLOAT(6),
                PRIMARY KEY (DFX_ID)
                  )"
                  )
df4 <- dbExecute(conn,"CREATE TABLE MONTHLY_FX (
                    DFX_ID INTEGER NOT NULL,
                    DATE DATE NOT NULL,
                    FXUSDCAD FLOAT(9) NOT NULL,
```

```
                    PRIMARY KEY (DFX_ID)
                    )"
                        )
```

0

0

0

0

In [7]:
```
#check list of tables in the present db.
dbListTables(conn)
```

'CROP_DATA' · 'DAILY_FX' · 'FARM_PRICES' · 'MONTHLY_FX'

Problem 2

Read Datasets and load your tables in database

Read the datasets into R dataframes using the urls provided above. Then load your tables in database.

In [8]:
```
# Write your query here
crop_df <- read.csv('https://cf-courses-data.s3.us.cloud-object-storage.appdomain.c
farm_df <- read.csv('https://cf-courses-data.s3.us.cloud-object-storage.appdomain.c
daily_df <- read.csv('https://cf-courses-data.s3.us.cloud-object-storage.appdomain.
monthly_df <- read.csv('https://cf-courses-data.s3.us.cloud-object-storage.appdomai

dbWriteTable(conn, "CROP_DATA", crop_df, overwrite=TRUE, header = TRUE)
dbWriteTable(conn, "FARM_PRICES", farm_df, overwrite=TRUE, header=TRUE)
dbWriteTable(conn, "DAILY_FX", daily_df, overwrite=TRUE, header=TRUE)
dbWriteTable(conn, "MONTHLY_FX", monthly_df, overwrite=TRUE, header=TRUE)
```

Now execute SQL queries using the RSQLite R package to solve the assignment problems.

Problem 3

How many records are in the farm prices dataset?

In [9]:
```
# Write your query here
dbGetQuery(conn, "SELECT COUNT(CD_ID) FROM FARM_PRICES")
```

A data.frame: 1 × 1

COUNT(CD_ID)
<int>
2678

Problem 4

Which geographies are included in the farm prices dataset?

In [10]:
```
# Write your query here
dbGetQuery(conn, "SELECT DISTINCT(GEO)FROM FARM_PRICES")
```

A data.frame: 2 × 1

GEO <chr>
Alberta
Saskatchewan

Problem 5

How many hectares of Rye were harvested in Canada in 1968?

In [11]:
```
# Write your query here
dbGetQuery(conn, "SELECT YEAR, CROP_TYPE, GEO, HARVESTED_AREA
          FROM CROP_DATA
          WHERE YEAR='1968-12-31' AND GEO='Canada' AND CROP_TYPE='Rye'")
```

A data.frame: 1 × 4

YEAR <chr>	CROP_TYPE <chr>	GEO <chr>	HARVESTED_AREA <int>
1968-12-31	Rye	Canada	274100

Problem 6

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

In [20]:
```
# Wriet your query here
result_6 <- dbGetQuery(conn,
  "SELECT *
   FROM FARM_PRICES
   WHERE CROP_TYPE='Rye'
   LIMIT 6"
)

print(result_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?

In [13]:
```
# Write your query here
dbGetQuery(conn, "SELECT DISTINCT(GEO)FROM FARM_PRICES WHERE CROP_TYPE='Barley'")
```

A data.frame: 2 × 1

GEO <chr>
Alberta
Saskatchewan

Problem 8

Find the first and last dates for the farm prices data.

In [14]:
```
# Write your query here
dbGetQuery(conn,"SELECT min(DATE), max(DATE) FROM FARM_PRICES")
```

A data.frame: 1 × 2

min(DATE) <chr>	max(DATE) <chr>
1985-01-01	2020-12-01

Problem 9

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

In [15]:
```
# Write your query here
dbGetQuery(conn, "SELECT DISTINCT(CROP_TYPE) FROM FARM_PRICES WHERE PRICE_PRERMT>=3
```

A data.frame:

1 × 1

CROP_TYPE
<chr>
Canola

Problem 10

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

In [16]:
```
# Write your query here
dbGetQuery(conn,"SELECT YEAR, CROP_TYPE, AVG_YIELD FROM CROP_DATA WHERE YEAR='2000-
```

A data.frame: 4 × 3

YEAR	CROP_TYPE	AVG_YIELD
<chr>	<chr>	<int>
2000-12-31	Barley	2800
2000-12-31	Wheat	2200
2000-12-31	Rye	2100
2000-12-31	Canola	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?

In [17]:
```
# Write your query here
dbGetQuery(conn,"SELECT CROP_TYPE, GEO, AVG(AVG_YIELD) AS avg_yield_since_2000
           FROM CROP_DATA WHERE YEAR>='2000-12-31' GROUP BY CROP_TYPE, GEO ORDER BY
#Barley in Alberta had the highest average yield since the year 2000.
```

A data.frame: 12 × 3

CROP_TYPE	GEO	avg_yield_since_2000
<chr>	<chr>	<dbl>
Barley	Alberta	3450.714
Barley	Canada	3253.762
Wheat	Alberta	3100.619
Barley	Saskatchewan	2971.048
Wheat	Canada	2845.333
Rye	Alberta	2683.810
Rye	Canada	2543.905
Wheat	Saskatchewan	2429.381
Rye	Saskatchewan	2226.714
Canola	Alberta	1999.238
Canola	Canada	1873.381
Canola	Saskatchewan	1754.857

Problem 12

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

In [18]:
```
# Write your query here

dbGetQuery(conn, "SELECT HARVESTED_AREA
FROM CROP_DATA
WHERE
  CROP_TYPE = 'Wheat'
  AND GEO = 'Canada'
  AND YEAR = (
      SELECT MAX(YEAR)
      FROM CROP_DATA
           )
  ")
# 10,017,800 hectares of wheat were harvested in Canada in the most recent year.
```

A data.frame: 1 × 1

HARVESTED_AREA
<int>
10017800

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.

In [19]:
```
dbGetQuery(conn, "SELECT PRICE_PRERMT AS CAD, PRICE_PRERMT/FXUSDCAD AS USD

FROM FARM_PRICES fp, MONTHLY_FX fx
WHERE
  fp.DATE = fx.DATE
  AND fp.CROP_TYPE = 'Canola'
  AND fp.GEO = 'Saskatchewan'
ORDER BY fp.DATE DESC
LIMIT 6")
```

A data.frame: 6 × 2

CAD	USD
<dbl>	<dbl>
507.33	396.1128
495.64	379.2718
474.80	359.2965
463.52	350.4057
464.60	351.3827
462.88	342.9122

Author(s)

Jeff Grossman

D.M. Naidu

Contributor(s)

Rav Ahuja

Lakshmi Holla

Skills Network

IBM

In [ ]: