COMP-3150 COVID-19 Database Project

Group Members:
Ben Chittle
Danielle Nguyen
Najia Shinneeb

CONTENTS

01

Overview

Database Goals

Our goal was to construct a database that would manage data based on the following principles (from project outline):

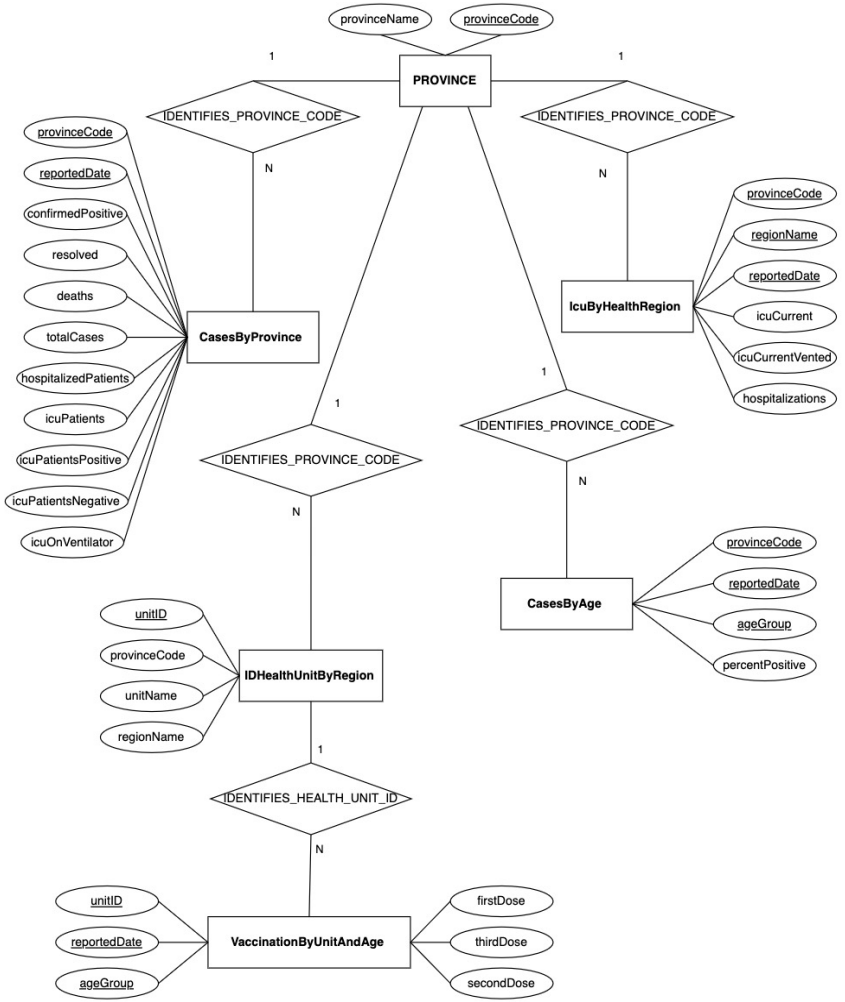
- COVID-19 Hospitalization and ICU information (both provincial and according to health regions)
- COVID-19 testing (including positivity rate)
- COVID-19 age-wise breakdown of the positive cases
- COVID-19 vaccine data (ONLY adults but age-wise)

02

ER Diagram

ER Diagram

- Modified accordingly to our database
- Six entities
- Five relationships



03

Relational Schema

Table Province

This Table Contains:

- All province names and province codes in Canada

Primary Key: Province Code

- Also used as a Foreign key in the other tables (to distinguish data by province)

Normalization note:
✓ INF: cells contain single value
✓ 2NF: single-attribute primary key
✓ 3NF: no transitive dependency

```
CREATE TABLE Province (
provinceName        VARCHAR(50)    NOT NULL,
provinceCode        CHAR(2)        NOT NULL,
PRIMARY KEY (provinceCode),
UNIQUE (provinceName)
);
```

provinceName	provinceCode
Alberta	AB
British Columbia	BC
Newfoundland and Labrador	NL
Prince Edward Island	PE
Nova Scotia	NS
New Brunswick	NB
Quebec	QC
Ontario	ON
Manitoba	MB
Saskatchewan	SK
Yukon	YT
Northwest Territories	NT
Nunavut	NU

Table CasesByProvince

This Table Contains:

- Information about COVID-19 case numbers and ICU for the whole province

Primary Key:

- Province Code (from Province)
- Reported Date

```
CREATE TABLE CasesByProvince (
provinceCode            CHAR(2)     NOT NULL,
reportedDate            DATE        NOT NULL,
confirmedPositive       INT         CHECK (confirmedPositive >= 0),
resolved                INT         CHECK (resolved >= 0),
deaths                  INT         CHECK (deaths >= 0),
totalCases              INT         CHECK (totalCases >=0),
hospitalizedPatients    INT         CHECK (hospitalizedPatients >= 0),
icuPatients             INT         CHECK (icuPatients>= 0),
icuPatientsPositive     INT         CHECK (icuPatientsPositive>= 0),
icuPatientsNegative     INT         CHECK (icuPatientsNegative>= 0),
icuOnVentilator         INT         CHECK (icuOnVentilator>= 0),
PRIMARY KEY (provinceCode, reportedDate),
FOREIGN KEY (provinceCode) REFERENCES Province(provinceCode)
);
```

provinceCode	reportedDate	confirmedPositive	resolved	deaths	totalCases	hospitalizedPatients	icuPatients	icuPatientsPositive	icuPatientsNegative	icuOnVentilator
ON	1-Jan-2021	21582	158472	4581	184635	1260	355	336	19	232
ON	2-Jan-2021	22847	160526	4625	187998	1003	344	322	22	230
ON	3-Jan-2021	23611	162701	4650	190962	998	352	329	23	240
ON	4-Jan-2021	24778	164775	4679	194232	1190	355	333	22	205
ON	5-Jan-2021	25840	166790	4730	197360	1347	375	352	23	257
ON	6-Jan-2021	26064	169795	4767	200626	1463	381	361	20	256
ON	7-Jan-2021	26718	172571	4856	204145	1472	382	363	19	251
ON	8-Jan-2021	28203	175309	4882	208394	1446	394	369	25	263

Normalization note:

✓ INF: cells contain single value

✓ 2NF: non-key attributes have no partial dependency

✓ 3NF: no transitive dependency

Table IcuByHealthRegion

```
CREATE TABLE IcuByHealthRegion (
provinceCode          CHAR(2)          NOT NULL,
regionName            VARCHAR(20)      NOT NULL,
reportedDate          DATE             NOT NULL,
icuCurrent            INT              CHECK (icuCurrent >= 0),
icuCurrentVented      INT              CHECK (icuCurrentVented >= 0),
hospitalizations      INT              CHECK (hospitalizations >= 0),
PRIMARY KEY (provinceCode, regionName, reportedDate),
FOREIGN KEY (provinceCode) REFERENCES Province(provinceCode)
);
```

This Table Contains:

- Information on COVID-19 ICU and hospitalization for each provincial region

Primary Key:

- Province Code (from Province)
- Provincial Regions
- Reported Date

provinceCode	regionName	reportedDate	icuCurrent	icuCurrentVented	hospitalizations
ON	Central	1-Jan-2021	127	94	517
ON	Central	2-Jan-2021	117	88	503
ON	Central	3-Jan-2021	120	93	534
ON	Central	4-Jan-2021	122	77	542
ON	Central	5-Jan-2021	125	91	562
ON	Central	6-Jan-2021	133	98	598
ON	Central	7-Jan-2021	134	101	595
ON	Central	8-Jan-2021	140	104	592
ON	Central	9-Jan-2021	141	98	603

Normalization note:

✓ INF: cells contain single value

✓ 2NF: non-key attributes have no partial dependency

✓ 3NF: no transitive dependency

Table IDHealthUnitByRegion

This Table Contains:

- Health Unit Names and their corresponding Region Names

Primary Key: Unit ID

Foreign Key: Province Code (from Province)

```
CREATE TABLE IDHealthUnitByRegion (
unitID          INT          NOT NULL,
provinceCode    CHAR(2)      NOT NULL,
unitName        VARCHAR(50)  NOT NULL,
regionName      VARCHAR(10)  NOT NULL,
PRIMARY KEY (unitID),
FOREIGN KEY (provinceCode) REFERENCES Province(provinceCode)
);
```

unitId	provinceCode	unitName	regionName
1	ON	Algoma District	Northern
2	ON	Brant County	Western
3	ON	Durham Region	Central
4	ON	Grey Bruce	Western
5	ON	Haldimand-Norfolk	Western
6	ON	Haliburton, Muskoka and Kawartha Lakes	Eastern
7	ON	Halton Region	Western
8	ON	City of Hamilton	Western
9	ON	Hastings and Prince Edward	Eastern
10	ON	Chatham-Kent	Western
11	ON	Kingston, Frontenac, Lennox and Addington	Eastern
12	ON	Lambton County	Western
13	ON	Leeds, Greenville and Lanark District	Eastern
14	ON	Middlesex-London	Western

Normalization note:

✓ INF: cells contain single value

✓ 2NF: single-attribute primary key

Table VaccinationByUnitAndAge

This Table Contains:

- Daily information on COVID-19 vaccination rates for each age group in every provincial health unit

Primary Key:

- Unit ID (from IDHealthUnitByRegion)
- Reported Date
- Age Group

Normalization note:

✓ INF: cells contain single value

✓ 2NF: non-key attributes have no partial dependency

✓ 3NF: no transitive dependency

```
CREATE TABLE VaccinationByUnitAndAge (
unitID          INT     NOT NULL,
reportedDate    DATE    NOT NULL,
ageGroup        VARCHAR(10)  NOT NULL,
firstDose       INT     CHECK (firstDose >= 0),
secondDose      INT     CHECK (secondDose >= 0),
thirdDose       INT     CHECK (thirdDose >= 0),
PRIMARY KEY (unitID, reportedDate, ageGroup),
FOREIGN KEY (unitID) REFERENCES IDHealthUnitByRegion (unitID)
);
```

unitId	reportedDate	ageGroup	firstDose	secondDose	thirdDose
1	26-Jul-2021	18-29	9811	6763	0
2	26-Jul-2021	18-29	15753	11325	0
3	26-Jul-2021	18-29	79765	59785	0
4	26-Jul-2021	18-29	13018	9833	0
5	26-Jul-2021	18-29	8171	5819	0
6	26-Jul-2021	18-29	14374	9824	0
7	26-Jul-2021	18-29	66917	46763	0
8	26-Jul-2021	18-29	64059	47055	0
9	26-Jul-2021	18-29	12858	8086	0
10	26-Jul-2021	18-29	8137	5775	0
11	26-Jul-2021	18-29	23996	17802	0
12	26-Jul-2021	18-29	10937	7432	0

Table CasesByAge

This Table Contains:

- Daily information on percent positive cases for adults per province

Primary Key:

- Province Code (from Province)
- Reported Date
- Age

Normalization note:

✓ INF: cells contain single value

✓ 2NF: non-key attributes have no partial dependency

✓ 3NF: no transitive dependency

```
CREATE TABLE CasesByAge (
provinceCode        CHAR(2)          NOT NULL,
reportedDate        DATE             NOT NULL,
ageGroup            VARCHAR(20)      NOT NULL,
percentPositive     DECIMAL(5, 4),
PRIMARY KEY (provinceCode, reportedDate, ageGroup),
FOREIGN KEY (provinceCode) REFERENCES Province(provinceCode)
);
```

provinceCode	reportedDate	ageGroup	percentPositive
ON	1-Jan-2021	18-24	0.0873
ON	2-Jan-2021	18-24	0.0908
ON	3-Jan-2021	18-24	0.0945
ON	4-Jan-2021	18-24	0.0932
ON	5-Jan-2021	18-24	0.0967
ON	6-Jan-2021	18-24	0.0969
ON	7-Jan-2021	18-24	0.0967
ON	8-Jan-2021	18-24	0.0962
ON	9-Jan-2021	18-24	0.0944
ON	10-Jan-2021	18-24	0.0911
ON	11-Jan-2021	18-24	0.0886
ON	12-Jan-2021	18-24	0.0855
ON	13-Jan-2021	18-24	0.0818
ON	14-Jan-2021	18-24	0.0787
ON	15-Jan-2021	18-24	0.0755

04

Database Design

Excel to SQL Queries (ExcelToSQL.py)

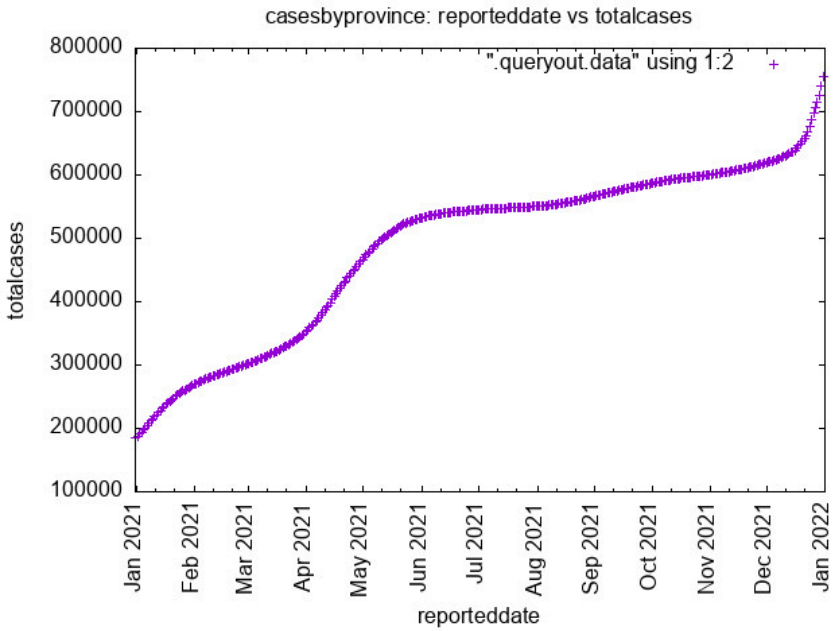
- Excel data populated using data from Ontario governmental website
- Custom Python script that transforms data in an Excel sheet into INSERT statements
- Output is a .sql file that can be imported in SQLPLUS Server

ExcelToSQL.py

```
"""

# ------------------------------------------------------------------------------

import pandas as pd

# (Unix) Path to the input Excel workbook.
INPUT_PATH = r"~/Downloads/Database.xlsx"
# (Unix) Path to output file (can be a relative path; will be created or overwritten).
OUTPUT_PATH = r"inserts.sql"

# Read in data from Excel workbook.
data = pd.read_excel(INPUT_PATH, None)

# List for all INSERT statements.
all_statements = []

# Iterate over each sheet.
for table, df in data.items():
    # List of statements for this sheet.
    statements = []
    # We need to format any date columns correctly for SQL to read them.
    if ("reportedDate" in df.columns):
        df["reportedDate"] = df["reportedDate"].apply(lambda x: x.strftime("%d-%b-%Y"))

    # Iterate over the rows of the sheet.
    for row in df.iterrows():
        # Create an INSERT statement for each row of data.
        statements.append(f"INSERT INTO {table} VALUES {tuple(row[1])};")

    # Join the individual INSERT statements into a single string separated by
    # newlines.
    all_statements.append("\n".join(statements))

# Open / create a file for the output statements.
with open(OUTPUT_PATH, "w") as file:
    # Join each string of statements (we have a string for each sheet) into one
    # long string separated by newlines and write it to a file.
    file.write("\n".join(all_statements))
```

Plotting Script (plot.sh)

- Another custom script
- Script asks for user input for x and y data columns
- Generates .jpg file using gnuplot GUI program program on the CS server
- Example output on right



Limitations

- Actual Canadian database used various field definitions causing inconsistencies in:
  - Age Groupings
  - Provincial Regions
  - Health Unit Regions
  - ICU by Age Group or ICU by Health Region was not available

SOLUTION:

- The Province and the IDHealthUnitByRegion tables were created in order to coordinate all the various data together

- Our database is limited to 2021 COVID-19 data in Ontario as the school CS server can only support so much

SOLUTION:

- The Province table provides the option to expand the data into other province data.
- On local servers, there is the option to extract older (2020) and newer (2022) data

05

Queries

Scenarios

- **QUERY 1:** *Get the average number of hospitalizations in each province for each month and order the results by province and date.*
- **QUERY 2:** *Get the average ratio of ICU ventilators in use to total hospitalizations for each month in each health region. Order the results by province, date, and health region.*
- **QUERY 3:** *Prompt the user to enter an age group. Then, display the total number of people in that age group who have received their first and second dose of the vaccine, grouped and sorted by region and date.*
- **QUERY 4:** *Assume a COVID outbreak occurred due to holiday gatherings. Find how many ICU cases each provincial region saw during the last 2 weeks of December of 2021 (December 18 – December 31). Order them from most cases to least cases.*
- **QUERY 5:** *The Omicron COVID variant was first reported in Canada November 28, 2021. Find the percent increase of Ontario COVID cases between November and December of 2021. (or the percent increase over the months).*
- **QUERY 6***: Prompt the user to enter an age group. Get the percentage of people who took the first dose and did not take the second dose for each month by age group.*
- **QUERY 7:** *Prompt the user to enter a column name from CasesByProvince to get the smallest value for each month of the reported column.*

Live Demo...

Scenarios

- **QUERY 1:** *Get the average number of hospitalizations in each province for each month and order the results by province and date.*

```
SELECT P.provinceName,
       EXTRACT(YEAR FROM C.reportedDate) AS "year",
       EXTRACT(MONTH FROM C.reportedDate) AS "month",
       ROUND(AVG(C.hospitalizedPatients), 0) AS "avgHospitalized"
FROM CasesByProvince C
JOIN Province P ON P.provinceCode = C.ProvinceCode
GROUP BY P.provinceName,
         EXTRACT(YEAR FROM C.reportedDate),
         EXTRACT(MONTH FROM C.reportedDate)
ORDER BY P.provinceName,
         EXTRACT(YEAR FROM C.reportedDate),
         EXTRACT(MONTH FROM C.reportedDate);
```

PROVINCENAME	year	month	avgHospitalized
Ontario	2021	1	1440
Ontario	2021	2	826
Ontario	2021	3	765
Ontario	2021	4	1800
Ontario	2021	5	1465
Ontario	2021	6	432
Ontario	2021	7	149
Ontario	2021	8	171
Ontario	2021	9	310
Ontario	2021	10	219
Ontario	2021	11	215

Scenarios

- **QUERY 2:** *Get the average ratio of ICU ventilators in use to total hospitalizations for each month in each health region. Order the results by province, date, and health region.*

```
SELECT P.provinceName,
       EXTRACT(YEAR FROM H.reportedDate) AS "year",
       EXTRACT(MONTH FROM H.reportedDate) AS "month",
       H.regionName,
       ROUND(AVG(H.icuCurrentVented), 0) AS "vented",
       ROUND(AVG(H.hospitalizations), 0) AS "hospitalizations",
       ROUND(AVG(H.icuCurrentVented) / AVG(H.hospitalizations) * 100, 2) AS "ratio (%)"
FROM IcuByHealthRegion H
JOIN Province P ON P.provinceCode = H.ProvinceCode
GROUP BY P.provinceName,
         EXTRACT(YEAR FROM H.reportedDate),
         EXTRACT(MONTH FROM H.reportedDate),
         H.regionName
ORDER BY P.provinceName,
         EXTRACT(YEAR FROM H.reportedDate),
         EXTRACT(MONTH FROM H.reportedDate),
         H.regionName;
```

PROVINCENAME	year	month	REGIONNAME	vented	hospitalizations	ratio (%)
Ontario	2021	1	Central	95	578	16.42
Ontario	2021	1	Eastern	46	197	23.42
Ontario	2021	1	Northern	2	13	11.82
Ontario	2021	1	Toronto	59	260	22.84
Ontario	2021	1	Western	64	439	14.66
Ontario	2021	2	Central	76	338	22.37
Ontario	2021	2	Eastern	45	125	35.78
Ontario	2021	2	Northern	3	20	15.54
Ontario	2021	2	Toronto	40	147	27.11
Ontario	2021	2	Western	45	218	20.81
Ontario	2021	3	Central	73	296	24.79

Scenarios

- **QUERY 3:** *Prompt the user to enter an age group. Then, display the total number of people in that age group who have received their first and second dose of the vaccine, grouped and sorted by region and date.*

```
PROMPT Enter one of the following age groups to display vaccination data for: 18-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80
ACCEPT ageGroup PROMPT "ageGroup: "
SET echo ON
SELECT I.regionName,
       EXTRACT(YEAR FROM V.reportedDate) AS "year",
       EXTRACT(MONTH FROM V.reportedDate) AS "month",
       SUM(V.firstDose) AS "total first doses (age=&ageGroup)",
       SUM(V.secondDose) AS "total second doses (age=&ageGroup)"
FROM VaccinationByUnitAndAge V
JOIN IdHealthUnitByRegion I
ON V.unitId = I.unitId
WHERE V.ageGroup = '&ageGroup'
GROUP BY EXTRACT(YEAR FROM V.reportedDate),
         EXTRACT(MONTH FROM V.reportedDate),
         I.regionName
ORDER BY EXTRACT(YEAR FROM V.reportedDate),
         EXTRACT(MONTH FROM V.reportedDate),
         I.regionName;
```

REGIONNAME	year	month	total first doses (age=18-29)	total second doses (age=18-29)
Central	2021	7	2745934	2064142
Eastern	2021	7	1443170	1063428
Northern	2021	7	806257	570250
Toronto	2021	7	2448297	1965622
Western	2021	7	2989376	2175827
Central	2021	8	14554055	11940183
Eastern	2021	8	7677165	6275662
Northern	2021	8	4331577	3419761
Toronto	2021	8	12978197	11063817
Western	2021	8	15970564	13097184
Central	2021	9	14863828	12977090

Scenarios

- **QUERY 4:** *Assume a COVID outbreak occurred due to holiday gatherings. Find how many ICU cases each provincial region saw during the last 2 weeks of December of 2021 (December 18 – December 31). Order them from most cases to least cases.*

```
SELECT regionName,
SUM(icuCurrent)
FROM IcuByHealthRegion
WHERE reportedDate BETWEEN '18-Dec-2021' AND '31-Dec-2021'
GROUP BY regionName
ORDER BY SUM(icuCurrent) DESC;
```

REGIONNAME	SUM(ICUCURRENT)
Western	1073
Central	438
Eastern	427
Toronto	188
Northern	138

Scenarios

- **QUERY 5:** *The Omicron COVID variant was first reported in Canada November 28, 2021. Find the percent increase of Ontario COVID cases between November and December of 2021. (or the percent increase over the months).*

```
SELECT
    P.provinceName,
    EXTRACT(YEAR FROM C.reportedDate) AS "year",
    EXTRACT(MONTH FROM C.reportedDate) AS "month",
    SUM(C.confirmedPositive) AS "total positive cases",
    LAG(SUM(C.confirmedPositive), 1, NULL) OVER (ORDER BY EXTRACT(YEAR FROM C.reportedDate), EXTRACT(MONTH FROM C.reportedDate)) AS "last month total",
    (SUM(C.confirmedPositive) - LAG(SUM(C.confirmedPositive), 1, NULL) OVER (ORDER BY EXTRACT(YEAR FROM C.reportedDate), EXTRACT(MONTH FROM C.reportedDate))) AS "difference",
    (SUM(C.confirmedPositive) - LAG(SUM(C.confirmedPositive), 1, NULL) OVER (ORDER BY EXTRACT(YEAR FROM C.reportedDate), EXTRACT(MONTH FROM C.reportedDate))) / LAG(SUM
(C.confirmedPositive), 1, NULL) OVER (ORDER BY EXTRACT(YEAR FROM C.reportedDate), EXTRACT(MONTH FROM C.reportedDate)) * 100 AS "% increase"
FROM CasesByProvince C
JOIN Province P
ON C.provinceCode = P.provinceCode
GROUP BY provinceName, EXTRACT(YEAR FROM C.reportedDate), EXTRACT(MONTH FROM C.reportedDate)
ORDER BY provinceName, EXTRACT(YEAR FROM C.reportedDate), EXTRACT(MONTH FROM C.reportedDate);
```

PROVINCENAME	year	month	total positive cases	last month total	difference	% increase
Ontario	2021	1	796555			
Ontario	2021	2	355178	796555	-441377	-55.410737
Ontario	2021	3	417311	355178	62133	17.4934821
Ontario	2021	4	1050224	417311	632913	151.664586
Ontario	2021	5	811729	1050224	-238495	-22.708965
Ontario	2021	6	164496	811729	-647233	-79.735109
Ontario	2021	7	49793	164496	-114703	-69.729963
Ontario	2021	8	114957	49793	65164	130.869801
Ontario	2021	9	179580	114957	64623	56.2149325
Ontario	2021	10	122110	179580	-57470	-32.00245
Ontario	2021	11	142874	122110	20764	17.0043403

Scenarios

- **QUERY 6***: Prompt the user to enter an age group. Get the percentage of people who took the first dose and did not take the second dose for each month by age group.*

```
PROMPT Enter one of the following age groups to display vaccination data for: 18-29,
30-39, 40-49, 50-59, 60-69, 70-79, 80
ACCEPT ageGroup PROMPT "ageGroup: "
SET echo ON
SELECT
    I.regionName AS "regionName",
    EXTRACT(YEAR FROM V.reportedDate) AS "year",
    EXTRACT(MONTH FROM V.reportedDate) AS "month",
    ROUND((SUM(V.firstDose) - SUM(V. secondDose)) /SUM(V.firstDose) * 100, 2) AS
"percentage (&ageGroup)"
FROM VaccinationByUnitAndAge V
JOIN IdHealthUnitByRegion I
ON V.unitId = I.unitId
WHERE V.ageGroup = '&ageGroup'
GROUP BY
    EXTRACT(YEAR FROM V.reportedDate),
    EXTRACT(MONTH FROM V.reportedDate),
    V.ageGroup,
    I.regionName
ORDER BY
    EXTRACT(YEAR FROM V.reportedDate),
    EXTRACT(MONTH FROM V.reportedDate),
    I.regionName;
```

regionName	year	month	percentage (60–69)
Central	2021	7	8.73
Eastern	2021	7	8.72
Northern	2021	7	9.76
Toronto	2021	7	8.52
Western	2021	7	8.98
Central	2021	8	5.05
Eastern	2021	8	4.9
Northern	2021	8	5.47
Toronto	2021	8	5.54
Western	2021	8	5.1
Central	2021	9	3.05

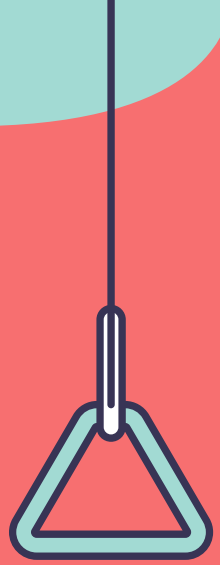
Scenarios

- **QUERY 7***: Prompt the user to enter a column name from CasesByProvince to get the smallest value for each month of the reported column.*

```
PROMPT Enter one of the following columns: confirmedPositive, resolved, deaths, totalCases, hospitalizedPatients, icuPatients to find the minimum value for each month
ACCEPT ans CHAR PROMPT "ans: "
SET echo ON
SELECT
    EXTRACT(YEAR FROM C.reportedDate) AS "year",
    EXTRACT(MONTH FROM C.reportedDate) AS "month",
    MIN(C.&ans) AS "min(&ans)"
FROM CasesByProvince  C
GROUP BY
    EXTRACT(YEAR FROM C.reportedDate),
    EXTRACT(MONTH FROM C.reportedDate)
ORDER BY
    EXTRACT(YEAR FROM C.reportedDate),
    EXTRACT(MONTH FROM C.reportedDate);
```

year	month	min(confirmedPositive)
2021	1	19216
2021	2	10050
2021	3	10210
2021	4	20875
2021	5	12567
2021	6	2257
2021	7	1334
2021	8	1667
2021	9	4947
2021	10	2978
2021	11	3093

THANKS!



CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik.**