Midterm Q2

Yufei Yin

Question 2: Database

Consider the database in the file `stat240.sqlite` provided in this midterm archive. This database contains a table named **citiesA** containing the area of cities and a table named **citiesP** containing the population of cities. This question has 4 parts, which must all be completed. For each part, provide code and output in a single *pdf* file through *crowdmark*. Provide axis-lables and titles for all of your plots.

Question 2, Part I

Connect to the database in the file `stat240.sqlite` and output the names of the tables in the database. For each table, output the names of the columns of the table and the data types of the columns and the number of entries in the table.

```
library(RSQLite)
library(DBI)
dbcon = dbConnect(SQLite(), dbname="stat240.sqlite")
```

The names of the tables

```
dbListTables(dbcon)
```

```
## [1] "Locations" "citiesA"   "citiesP"
```

```
Locations = dbReadTable(dbcon, "Locations")
citiesA = dbReadTable(dbcon, "citiesA")
citiesP = dbReadTable(dbcon, "citiesP")
```

The names of the columns of the table

```
names(Locations)
```

```
## [1] "ID"              "Country"         "Geographic_name" "Region"
## [5] "Province"        "Prov_acr"        "Latitude"        "Longitude"
## [9] "Region_Index"
```

```
names(citiesA)
```

```
## [1] "rank"     "name"     "province" "status"   "area"
```

```
names(citiesP)
```

```
## [1] "rank2016"       "rank2011"       "name"           "province"
## [5] "type"           "population2016" "population2011" "noise"
```

The data types of the columns

```
str(Locations)
```

```
## 'data.frame':    1640 obs. of  9 variables:
##  $ ID             : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country        : chr  "CA" "CA" "CA" "CA" ...
##  $ Geographic_name: chr  "T0A" "T0B" "T0C" "T0E" ...
##  $ Region         : chr  "Eastern Alberta (St. Paul)" "Wainwright Region (Tofield)" "Central Alberta
##  $ Province       : chr  "Alberta" "Alberta" "Alberta" "Alberta" ...
##  $ Prov_acr       : chr  "AB" "AB" "AB" "AB" ...
##  $ Latitude       : num  54.8 53.1 52.5 53.4 55.7 ...
##  $ Longitude      : num  -112 -112 -113 -117 -114 ...
##  $ Region_Index   : int  NA NA NA NA NA NA NA NA NA NA ...
```

```
str(citiesA)
```

```
## 'data.frame':    100 obs. of  5 variables:
##  $ rank    : chr  "1" "2" "  3" "  4" ...
##  $ name    : chr  "La Tuque" "Senneterre" "Rouyn-Noranda" "Val-d'Or" ...
##  $ province: chr  "Quebec" "Quebec" "Quebec" "Quebec" ...
##  $ status  : chr  "Ville" "Ville" "Ville" "Ville" ...
##  $ area    : num  22 12 86 35 34 33 14 13 11 10 ...
```

```
str(citiesP)
```

```
## 'data.frame':    152 obs. of  8 variables:
##  $ rank2016      : chr  "1" "                2" "                3" "                4" ...
##  $ rank2011      : chr  "1" "2" "3" "5" ...
##  $ name          : chr  "Toronto" "Montreal " "Vancouver (Surrey)" "Calgary" ...
##  $ province      : chr  "Ontario" "Quebec" "British Columbia" "Alberta" ...
##  $ type          : chr  "CMA" "CMA" "CMA" "CMA" ...
##  $ population2016: chr  "5928040" "4098927" "2463431" "1392609" ...
##  $ population2011: chr  "5583064" "3934078" "2313328" "1214839" ...
##  $ noise         : chr  "b6.8" "d4.2" "b8.2" "b14.1" ...
```

"num": numeric

"chr": character

"int": integer

The number of entries in the table

```
dim(Locations)[1]*dim(Locations)[2]
```

```
## [1] 14760
```

```
dim(citiesA)[1]*dim(citiesP)[2]
```

```
## [1] 800
```

```
dim(citiesP)[1]*dim(citiesP)[2]
```

```
## [1] 1216
```

Question 2, Part II

Use *SQL* to extract the unique combinations of **province** and **type** from the **citiesP** table, and provide the number of such unique combinations.

The unique combinations of province and type from the citiesP table

```
sql = "SELECT DISTINCT province, type FROM citiesP"
dbGetQuery(dbcon, sql)
```

```
##                     province type
## 1                    Ontario  CMA
## 2                     Quebec  CMA
## 3           British Columbia  CMA
## 4                    Alberta  CMA
## 5             Ontario/Quebec  CMA
## 6                   Manitoba  CMA
## 7                Nova Scotia  CMA
## 8               Saskatchewan  CMA
## 9  Newfoundland and Labrador  CMA
## 10             New Brunswick  CMA
## 11          British Columbia   CA
## 12                   Ontario   CA
## 13             New Brunswick   CA
## 14                   Alberta   CA
## 15               Nova Scotia   CA
## 16                    Quebec   CA
## 17      Prince Edward Island   CA
## 18                  Manitoba   CA
## 19              Saskatchewan   CA
## 20      Alberta/Saskatchewan   CA
## 21 Newfoundland and Labrador   CA
## 22                     Yukon   CA
## 23     Northwest Territories   CA
## 24            Ontario/Quebec   CA
```

The number of such unique combinations

```
dim(dbGetQuery(dbcon, sql))[1]
```

```
## [1] 24
```

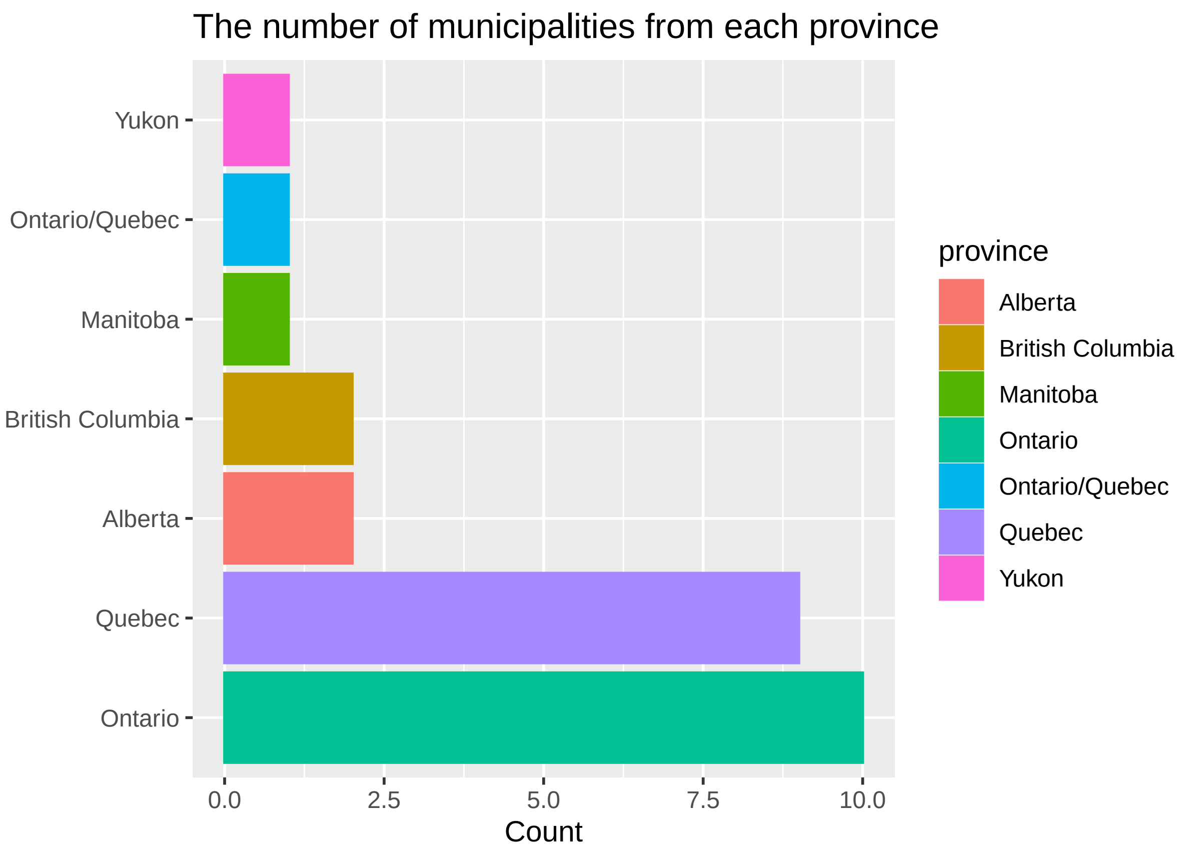
Question 2, Part III

Use *SQL* to obtain the number of municipalities within each province in the database. Restrict to location names that are present in both **citiesA** and **citiesP** tables. Provide a plot of the number of resulting municipalities from each province.

```
sql = "SELECT name,
              province,
              COUNT(province) AS count
       FROM(SELECT * FROM citiesP
            INNER JOIN citiesA
            ON
             citiesP.name = citiesA.name)
       GROUP BY province"
dbGetQuery(dbcon, sql)
```

```
##         name         province count
## 1    Calgary          Alberta     2
## 2   Kamloops British Columbia     2
## 3   Winnipeg         Manitoba     1
## 4    Toronto          Ontario    10
## 5     Ottawa   Ontario/Quebec     1
## 6   Saguenay           Quebec     9
## 7 Whitehorse            Yukon     1
```

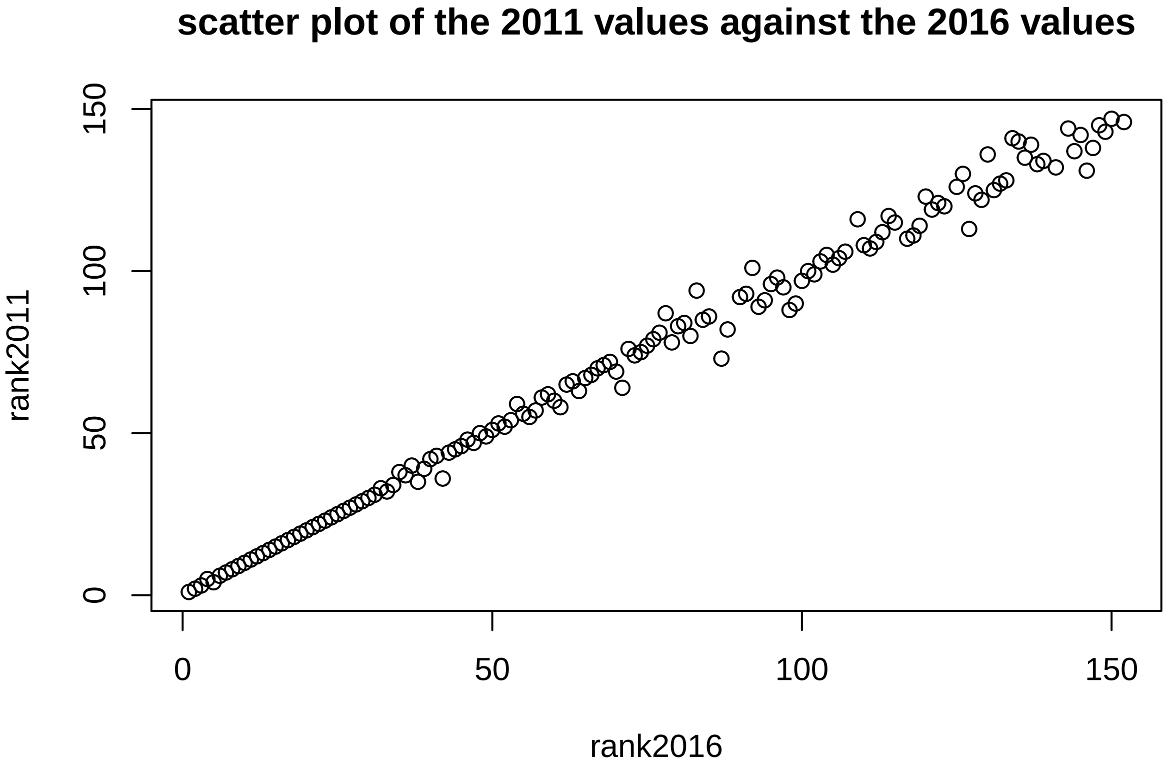
```
# plot
library(tidyverse)
sql = "SELECT province
       FROM(SELECT * FROM citiesP
            INNER JOIN citiesA
            ON
             citiesP.name = citiesA.name)"
province = dbGetQuery(dbcon, sql)

ggplot(data = province, aes(y = fct_infreq(province), color = province, fill = province)) +
  geom_bar() +
  labs(x = "Count", y = "",title = "The number of municipalities from each province")
```


The number of municipalities from each province
Yukon
Ontario/Quebec
Manitoba
British Columbia
Alberta
Quebec
Ontario
0.0
2.5
5.0
7.5
10.0
Count
province
Alberta
British Columbia
Manitoba
Ontario
Ontario/Quebec
Quebec
Yukon

Question 2, Part IV

For each location in the **citiesP** table, the columns **rank2011** and **rank2016** represent the popularity of the destination with tourists in the years 2011 and 2016 respectively (the tourist rank orders). Extract the tourist rank order for 2011 (**rank2011**) and for 2016 (**rank2016**) for each location in the **citiesP** table. Provide a scatter plot of the 2011 values against the 2016 values (i.e., a plot with one point per location and the 2011 values on the y-axis and the 2016 values on the x-axis).

```
sql = "SELECT rank2011, rank2016, name, province FROM citiesP WHERE rank2011 != ’NR’"
rank <- dbGetQuery(dbcon, sql)
plot(x = rank$rank2016,
     y = rank$rank2011,
     xlab = "rank2016",
     ylab = "rank2011",
     main = "scatter plot of the 2011 values against the 2016 values")
```


scatter plot of the 2011 values against the 2016 values
rank2011
rank2016