SQL exploration

Mark Gallivan

SQL exploration of three databases on providers, patients, and healthcare plans.

Set-up

```
## Load packages
library(RSQLite) # used create connection to db and get sqlite query instead of bash call
library(knitr) # used to make nice tables
opts_chunk$set(tidy.opts=list(width.cutoff=80),tidy=TRUE)

#tinytex::install_tinytex(force = TRUE) #for latex
## set up connection to  db
SqliteDriver <- dbDriver("SQLite")
DB <- dbConnect(SqliteDriver,
             dbname = "/Volumes/GoogleDrive/My Drive/Admin/data/sqlite.db")
## Through terminal:
#system("cd /Volumes/GoogleDrive/My Drive/Admin/db/")
#system("sqlite3 $sqlite.db)
```

Part I.

Question 1. How many rows are in the provider database?

There are a total of 100 entries in the provider database.

```
kable(dbGetQuery(DB, "SELECT COUNT(*) as TotalEntries FROM providers"))
```

TotalEntries
100

Question 2. How many specialists are in the provier database?

There are a total of 33 specialists (or non-primary care physcians) in the database.

```
kable(dbGetQuery(DB, "SELECT Count(id) as CountNonICPs
                            FROM providers
                            WHERE is_pcp = 0"))
```

CountNonICPs
33

Question 3. What states are the providers in?

According to the provider directory, there are providers in five states: Massachusetts, Idaho, Florida, Alaska, and Arkansas.

```
kable(dbGetQuery(DB, "SELECT DISTINCT
                            substr(city, INSTR(city, ',')+1, length(city)) as ProviderStates
                            FROM providers"))
```

ProviderStates
MA
ID
FL
AK
AR

Question 3a. How many physicians are in each state?

There are 14 physicians in Alaska, 19 in Arkansas, 29 in Florida, 24 in Idaho, and 14 in Massachusetts.

```
kable(dbGetQuery(DB, "SELECT substr(city, INSTR(city, ',')+1, length(city)) as States, COUNT(id) as Cou
                      FROM providers
                      GROUP BY States"))
```

States	Count
AK	14
AR	19
FL	29
ID	24
MA	14

Question 4. How many rows are in the patient database?

There are 10000 total entries in the patient database.

```
kable(dbGetQuery(DB, "SELECT COUNT(*) as CountMembers FROM members"))
```

CountMembers
10000

Question 5. What states do members live in?

Members live in Arkansas, Alaska, Florida, Idaho, and Massachusetts.

```
kable(dbGetQuery(DB, "SELECT DISTINCT
                      substr(city, INSTR(city, ',')+1, length(city)) as MemberStates
                      FROM members"))
```

MemberStates
AK
MA
FL
ID
AR

Question 5a. How many patients are in each state?

As seen in the table below, there are 2036 patients in Alaska, 1974 in Arkansas, 1967 in Florida, 2007 in Idaho, and 2016 in Massachusetts.

```
kable(dbGetQuery(DB, "SELECT substr(city, INSTR(city, ',')+1, length(city)) as States,
                              COUNT(DISTINCT id) as Count
                              FROM members
                              GROUP BY States"))
```

States	Count
AK	2036
AR	1974
FL	1967
ID	2007
MA	2016

Question 6. How many members have a primary care provider in each month?

The number of members that have a PCP each month can be found in the table below. However, it is important to note there are 1437 rows with a missing provider_id, however we will assume this is a data entry error since there is a record in the Member_PCP_Spans table.

```
kable(dbGetQuery(DB, "WITH NewTable AS
  (SELECT member_id, CAST(MIN(substr(start_date, 3, 2)) AS INT)
  as MinStart_Month, CAST(MAX(substr(end_date, 3,2)) AS INT) as MaxEnd_Month
  FROM Member_PCP_Spans GROUP BY member_id)
  SELECT SUM(case when MinStart_Month = 1 then 1 else 0 end) as Jan,
  SUM(case when MinStart_Month <= 2 AND MaxEnd_Month >= 2 then 1 else 0 end) as Feb,
  SUM(case when MinStart_Month <= 3 AND MaxEnd_Month >= 3 then 1 else 0 end) as Mar,
  SUM(case when MinStart_Month <= 4 AND MaxEnd_Month >= 4 then 1 else 0 end) as Apr,
  SUM(case when MinStart_Month <= 5 AND MaxEnd_Month >= 5 then 1 else 0 end) as May,
  SUM(case when MinStart_Month <= 6 AND MaxEnd_Month >= 6 then 1 else 0 end) as June,
  SUM(case when MinStart_Month <= 7 AND MaxEnd_Month >= 7 then 1 else 0 end) as Jul,
  SUM(case when MinStart_Month <= 8 AND MaxEnd_Month >= 8 then 1 else 0 end) as Aug,
  SUM(case when MinStart_Month <= 9 AND MaxEnd_Month >= 9 then 1 else 0 end) as Sept,
  SUM(case when MinStart_Month <= 10 AND MaxEnd_Month >= 10 then 1 else 0 end) as Oct,
  SUM(case when MinStart_Month <= 11 AND MaxEnd_Month >= 11 then 1 else 0 end) as Nov,
  SUM(case when MinStart_Month <= 12 AND MaxEnd_Month >= 12 then 1 else 0 end) as Dec
  FROM NewTable"))
```

Jan	Feb	Mar	Apr	May	June	Jul	Aug	Sept	Oct	Nov	Dec
3354	3929	4477	5079	5632	6212	6793	7356	7949	8508	9103	9663

Question 7. How many members changed their primary care provider just once during the last 6 months of 2018?

There were 4826 members who changed their PCP exactly once during the last months of 2018.

```
kable(dbGetQuery(DB, "WITH NewTable AS (SELECT member_id, start_date, COUNT(*) as Count_Num
          FROM Member_PCP_Spans WHERE (substr(start_date,3,2) IN ('07','08','09','10','11','12'))
          GROUP BY member_id HAVING Count_Num = 1) SELECT COUNT(*) as Count FROM NewTable"))
```

Count
4826

Question 8. In November, how many members are assigned to a Florida provider?

There were 1622 members assigned to a provider in Florida in November. However, it's important to note this number only includes PCPs.

```
kable(dbGetQuery(DB, "WITH NewTable AS (SELECT provider_id, member_id,
                CAST(substr(start_date, 3, 2) AS INT) as Start_Month,
                CAST(substr(end_date, 3,2) AS INT) as End_Month FROM Member_PCP_Spans)
                SELECT COUNT(*) as COUNT FROM NewTable
                INNER JOIN Providers ON Providers.id = NewTable.provider_id
                WHERE NewTable.Start_Month <= 11 AND NewTable.End_Month >= 11
                AND substr(Providers.city, INSTR(Providers.city, ',')+2, length(Providers.city)) = 'FL
```

COUNT
1622

Question 9. How many patients are currently seeing a provider practicing outside of the patient's city?

There are currently (as of 12/31/18), 8477 members seeing a provider practicing outside of the member's city.

```
kable(dbGetQuery(DB, "SELECT COUNT(Member_PCP_Spans.member_id) as CountMembers
                            FROM Member_PCP_Spans
                            LEFT JOIN Providers ON Member_PCP_Spans.provider_id = Providers.id
                            LEFT JOIN Members ON Member_PCP_Spans.member_id = Members.id
                            WHERE Member_PCP_Spans.end_date = '999999' AND
                            Providers.city != Members.city"))
```

CountMembers
8477

Part II. Based on the available information, which providers provide the best healthcare? What are limitations with the available data? What other information would you like to know?

We are looking for the top five PCPs based on a metric that captures "the best customer service". Based on the available data, the pcp_rating variable from the Member_PCP_Spans table is the metric most aligned with our question. From the table below, we select the top five PCPs based on the highest mean of this metric and evaluating that observing that at all the selected providers had at least 30 ratings (range of 35 to 76).

```
kable(dbGetQuery(DB, "WITH NewTable AS
                (SELECT provider_id, COUNT(pcp_rating) as NumObs, avg(pcp_rating) as MeanRating
                FROM Member_PCP_Spans WHERE pcp_rating IS NOT NULL
                GROUP BY provider_id ORDER BY MeanRating DESC LIMIT 5)
                SELECT provider_id, NumObs, MeanRating, name, city
                FROM NewTable
                INNER JOIN Providers ON Providers.id = NewTable.provider_id"))
```

provider_id	NumObs	MeanRating	name	city
1730452909	43	4.186046	Daniel Robinson	New Scott, ID
3460630898	76	4.039474	Daniel Frederick	South Erikabury, AR
1464837296	44	4.000000	Andrew Shaw	Jacksonhaven, ID

provider_id	NumObs	MeanRating	name	city
2914760740	35	3.971429	Virginia Adán Carbajal Mondragón	Morrisonville, FL
938439286	36	3.944444	John Thompson	Johnsonville, FL

Unfortunately, there are a number of missing values for several of the variables including the pcp_rating. In fact, 10485 out of 14372 total entries in the table do not have a pcp_rating.

A major limitation of this approach comes from the fact the PCP ratings are voluntary and given at a time when a member changes providers or plans. Therefore, there may a selection bias meaning that the sample with pcp_rating's provided in the Member_PCP_Spans table are not representitative to the patient population as a whole. For example, there may be a bias in the sample towards patients that found a cheaper plan or changed plan due to unsatisfactory relationship with the PCP. You could imagine that an ideal provider would have many satsified patients who would not change provider or plan, which would not be captured by the pcp_variable.

To make a more informed decision, additional data collection should be performed. Additional data could include a survey (preferably one that is mandatory or incentivized to minimize selection bias and be more population-based) given immediately (electronically or paper form) after a patient visit (to minize recall bias) that asks several questions preferably on a Likert scale including "How satisified are you with your current provider" (primary metric), "How likely are you to switch providers?", "Would you recommend your provider to a friend?", "Would you see your primary care provider again?". Similar questions are described in the book "The Innovator's Prescription: A Disruptive Solution for Health Care" by Clayton Christensen.

Additionally, the provided data covers a year but there could be additional analyses if more longitudinal data is present. I would recommend considering secondary and more indirect measures as well. This could include identifying providers with the highest duration of patient-provider relationship or conversely, the providers with the smallest rate of member switching after adjusting for patient risk score, average co-pay, length of employment, and other variable that might be causally related to the provider and outcome of interest. For example, a mixed-effects model (clustered by hospital or clinic) may be helpful to disentengle the effect of hospital or clinic and provider-patient satistification.