SURV 727 Assignment 4

Mathew Hill

2024-10-31

GitHub link: https://github.com/mathewhill/surv727hw4

After you have initialized a project, paste your project ID into the following chunk.

```
project <- "surv-727-project-4"
```

We will connect to a public database, the Chicago crime database, which has data on crime in Chicago.

```
con <- dbConnect(
  bigrquery::bigquery(),
  project = "bigquery-public-data",
  dataset = "chicago_crime",
  billing = project
)
con
```

```
## <BigQueryConnection>
##   Dataset: bigquery-public-data.chicago_crime
##   Billing: surv-727-project-4
```

```
dbListTables(con)
```

```
## ! Using an auto-discovered, cached token.

##   To suppress this message, modify your code or options to clearly consent to
##   the use of a cached token.

##   See gargle's "Non-interactive auth" vignette for more details:

##   <https://gargle.r-lib.org/articles/non-interactive-auth.html>

## i The bigrquery package is using a cached token for 'mathewfhill@gmail.com'.

## [1] "crime"
```

Write a first query that counts the number of rows of the 'crime' table in the year 2016. Use code chunks with {sql connection = con} in order to write SQL code within the document

```
SELECT count(primary_type), count(*)
FROM crime
WHERE year = 2016
LIMIT 10
```

Table 1: 1 records

f0_	f1_
269921	269921

Next, count the number of arrests grouped by primary_type in 2016. Note that is a somewhat similar task as above, with some adjustments on which rows should be considered. Sort the results, i.e. list the number of arrests in a descending order.

```
SELECT primary_type, COUNT(*) AS arrest_count
FROM crime
WHERE year = 2016 AND arrest = TRUE
GROUP BY primary_type
ORDER BY arrest_count DESC
```

Table 2: Displaying records 1 - 10

primary_type	arrest_count
NARCOTICS	13327
BATTERY	10333
THEFT	6522
CRIMINAL TRESPASS	3724
ASSAULT	3492
OTHER OFFENSE	3415
WEAPONS VIOLATION	2511
CRIMINAL DAMAGE	1669
PUBLIC PEACE VIOLATION	1116
MOTOR VEHICLE THEFT	1098

We can also use the date for grouping. Count the number of arrests grouped by hour of the day in 2016. You can extract the latter information from date via EXTRACT(HOUR FROM date). Which time of the day is associated with the most arrests?

```
SELECT EXTRACT(HOUR FROM date) AS arrest_hour, COUNT(*) AS arrest_count
FROM crime
WHERE year = 2016 AND arrest = TRUE
GROUP BY arrest_hour
ORDER BY arrest_count DESC;
```

Table 3: Displaying records 1 - 10

arrest_hour	arrest_count
19	3843
18	3481
20	3302
21	2961
16	2933
22	2896
11	2895
17	2820
12	2787
14	2774

Hour 19 appears to be associated with the most arrests.

Focus only on HOMICIDE and count the number of arrests for this incident type, grouped by year. List the results in descending order.

```
SELECT year, COUNT(*) AS homicide_arrests
FROM crime
WHERE primary_type = 'HOMICIDE' AND arrest = TRUE
GROUP BY year
ORDER BY homicide_arrests DESC;
```

Table 4: Displaying records 1 - 10

year	homicide_arrests
2001	430
2002	427
2003	382
2020	349
2022	306
2004	294
2021	291
2016	289
2008	287
2006	284

Find out which districts have the highest numbers of arrests in 2015 and 2016. That is, count the number of arrests in 2015 and 2016, grouped by year and district. List the results in descending order.

```
SELECT year, district, COUNT(*) AS arrest_count
FROM crime
WHERE year IN (2015, 2016) AND arrest = TRUE
GROUP BY year, district
ORDER BY arrest_count DESC;
```

Table 5: Displaying records 1 - 10

year	district	arrest_count
2015	11	8974
2016	11	6575
2015	7	5549
2015	15	4514
2015	6	4474
2015	25	4450
2015	4	4325
2015	8	4113
2016	7	3655
2015	10	3622

Lets switch to writing queries from within R via the DBI package. Create a query object that counts the number of arrests grouped by primary_type of district 11 in year 2016. The results should be displayed in descending order. Execute the query.

```
DBI_query <- dbSendQuery(con, "
  SELECT primary_type, COUNT(*) AS arrest_count
  FROM crime
  WHERE year = 2016 AND district = 11 AND arrest = TRUE
  GROUP BY primary_type
  ORDER BY arrest_count DESC
")
DBI_result <- dbFetch(DBI_query)
DBI_result
```

```
## # A tibble: 27 x 2
##    primary_type                      arrest_count
##    <chr>                                    <int>
##  1 NARCOTICS                                 3634
##  2 BATTERY                                    635
##  3 PROSTITUTION                               511
##  4 WEAPONS VIOLATION                          303
##  5 OTHER OFFENSE                              255
##  6 ASSAULT                                    206
##  7 CRIMINAL TRESPASS                          205
##  8 PUBLIC PEACE VIOLATION                     135
##  9 INTERFERENCE WITH PUBLIC OFFICER           119
## 10 CRIMINAL DAMAGE                            106
## # i 17 more rows
```

Try to write the very same query, now using the dbplyr package. For this, you need to first map the crime table to a tibble object in R.

```
crime_tibble <- tbl(con, "crime")

tibble_result <- crime_tibble %>%
```

```
  filter(year == 2016, district == 11, arrest == TRUE) %>%
  group_by(primary_type) %>%
  summarise(arrest_count = n()) %>%
  arrange(desc(arrest_count))

local_results <- collect(tibble_result)
local_results
```

```
## # A tibble: 27 x 2
##    primary_type                      arrest_count
##    <chr>                                    <int>
##  1 NARCOTICS                                 3634
##  2 BATTERY                                    635
##  3 PROSTITUTION                               511
##  4 WEAPONS VIOLATION                          303
##  5 OTHER OFFENSE                              255
##  6 ASSAULT                                    206
##  7 CRIMINAL TRESPASS                          205
##  8 PUBLIC PEACE VIOLATION                     135
##  9 INTERFERENCE WITH PUBLIC OFFICER           119
## 10 CRIMINAL DAMAGE                            106
## # i 17 more rows
```

###Again, count the number of arrests grouped by primary_type of district 11 in year 2016, now using dplyr syntax.

```
district_11_query <- dbSendQuery(con, "
  SELECT primary_type, COUNT(*) AS arrest_count
  FROM crime
  WHERE year = 2016 AND district = 11 AND arrest = TRUE
  GROUP BY primary_type
  ORDER BY arrest_count DESC
")
district_11_result <- dbFetch(district_11_query)
district_11_result
```

```
## # A tibble: 27 x 2
##    primary_type                      arrest_count
##    <chr>                                    <int>
##  1 NARCOTICS                                 3634
##  2 BATTERY                                    635
##  3 PROSTITUTION                               511
##  4 WEAPONS VIOLATION                          303
##  5 OTHER OFFENSE                              255
##  6 ASSAULT                                    206
##  7 CRIMINAL TRESPASS                          205
##  8 PUBLIC PEACE VIOLATION                     135
##  9 INTERFERENCE WITH PUBLIC OFFICER           119
## 10 CRIMINAL DAMAGE                            106
## # i 17 more rows
```

Count the number of arrests grouped by primary_type and year, still only for district 11. Arrange the result by year.

```
crime_tibble %>%
  filter(district == 11, arrest == TRUE) %>%
  group_by(primary_type, year) %>%
  summarise(arrest_count = n()) %>%
  arrange(year) %>%
  collect()
```

```
## `summarise()` has grouped output by "primary_type". You can override using the
## `.groups` argument.
```

```
## # A tibble: 613 x 3
## # Groups:   primary_type [32]
##    primary_type                        year arrest_count
##    <chr>                              <int>        <int>
##  1 ASSAULT                             2001          322
##  2 HOMICIDE                            2001           48
##  3 LIQUOR LAW VIOLATION                2001           49
##  4 INTERFERENCE WITH PUBLIC OFFICER    2001           14
##  5 KIDNAPPING                          2001            4
##  6 NARCOTICS                           2001         7979
##  7 CRIMINAL TRESPASS                   2001          389
##  8 MOTOR VEHICLE THEFT                 2001          179
##  9 PUBLIC PEACE VIOLATION              2001           34
## 10 WEAPONS VIOLATION                   2001          236
## # i 603 more rows
```

Assign the results of the query above to a local R object.

```
yearly_result <- crime_tibble %>%
  filter(district == 11, arrest == TRUE) %>%
  group_by(primary_type, year) %>%
  summarise(arrest_count = n()) %>%
  arrange(year) %>%
  collect()
```

```
## `summarise()` has grouped output by "primary_type". You can override using the
## `.groups` argument.
```

Confirm that you pulled the data to the local environment by displaying the first ten rows of the saved data set.

```
head(yearly_result, 10)
```

```
## # A tibble: 10 x 3
## # Groups:   primary_type [10]
```

```
##    primary_type                      year arrest_count
##    <chr>                            <int>        <int>
##  1 ASSAULT                           2001          322
##  2 HOMICIDE                          2001           48
##  3 LIQUOR LAW VIOLATION              2001           49
##  4 INTERFERENCE WITH PUBLIC OFFICER  2001           14
##  5 KIDNAPPING                        2001            4
##  6 NARCOTICS                         2001         7979
##  7 CRIMINAL TRESPASS                 2001          389
##  8 MOTOR VEHICLE THEFT               2001          179
##  9 PUBLIC PEACE VIOLATION            2001           34
## 10 WEAPONS VIOLATION                 2001          236
```

Close the connection.

```
dbDisconnect(con)
```