hw4

September 23, 2025

1 Homework 4

Rex Wang 1. Create the opioid sqlite database from https://smart-stats.github.io/ds4bio_book/book/_build/html/sqlite.html. However, only go to the step where the csv files are read into the database. Then exit sqlite and you should have a file opioid.db that has the data. Next, read the three tables into pandas dataframes and do the remaining data wrangling from the sqlite chapter directly in pandas. Add the python code to your hw4.ipynb file.

[1]:
```
import sqlite3 as sq3
import pandas as pd
import numpy as np

# create the connection
con = sq3.connect("opioid.db")

# read csv
annual = pd.read_sql_query("SELECT * from annual", con)
land = pd.read_sql_query("SELECT * from land", con)
population = pd.read_sql_query("SELECT * from population", con)

# fill in fips for Montgomery, AR
annual.loc[
    (annual["BUYER_STATE"] == "AR") &
    (annual["BUYER_COUNTY"] == "MONTGOMERY"),
    "countyfips"
    ] = "05097"
```

[2]:
```
# land area in 2010 with fips code
land2010 = land[["Areaname", "STCOU", "LND110210D"]].rename(columns={"STCOU":␣
 ↪"countyfips"})

# assign land area to population
county_info = pd.merge(
    population,
    land2010,
    how="left",           # left join
    on="countyfips"       # key column
)
```

```
# # Convert columns to numeric (int or float)
# county_info["population"] = pd.to_numeric(county_info["population"],␣
 ↪errors="coerce")
# county_info["LND110210D"] = pd.to_numeric(county_info["LND110210D"],␣
 ↪errors="coerce")
# # calculate population density
# county_info["density"] = county_info["population"] / county_info["LND110210D"]

# # turn year into a category
# county_info["year"] = county_info["year"].astype("category")

# county_info["rural_urban"] = np.where(
#     county_info["density"] >= 500,
#     "Urban",
#     "Rural"
# )

# assign pop density to drug dosage
annual["year"] = annual["year"].astype("category")

all_data = pd.merge(
    annual,
    county_info,
    how="inner",           # inner join
    on=[
        "BUYER_COUNTY",
        "BUYER_STATE",
        "year",
        "countyfips"
        ]
)

# pills in millions
all_data["DOSAGE_UNIT"] = pd.to_numeric(all_data["DOSAGE_UNIT"])
all_data["Pills_in_millions"] = all_data["DOSAGE_UNIT"] / 1000000
```

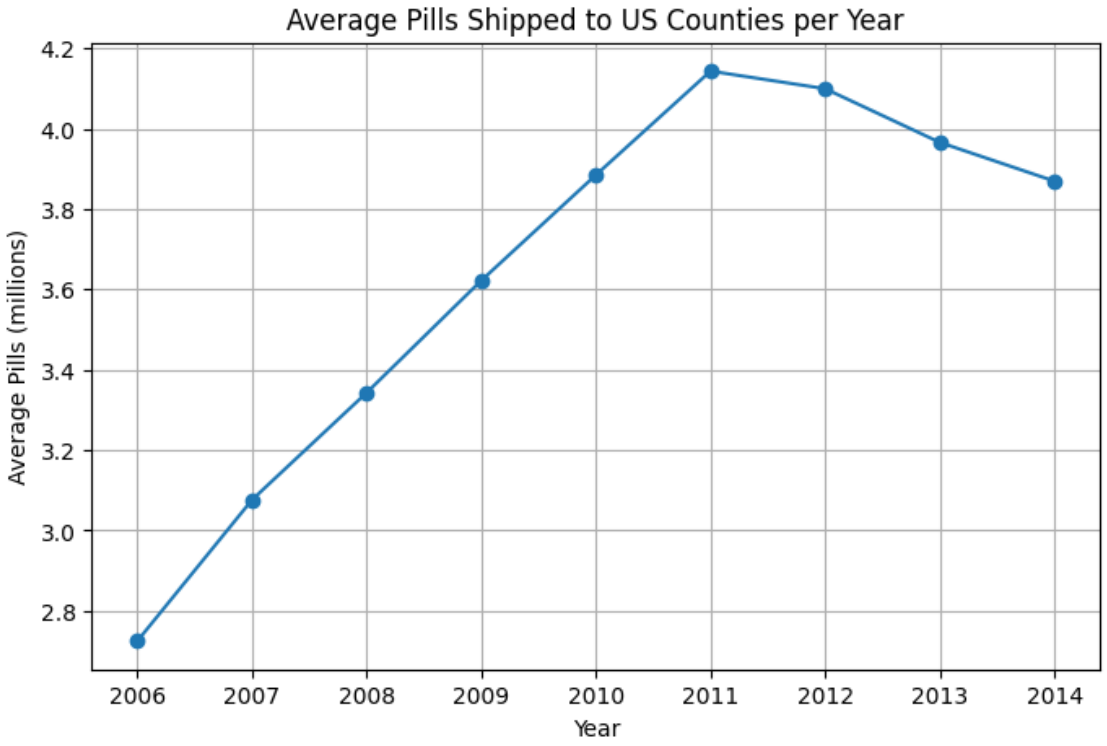
2. Create a scatterplot of the average number of opioid pills by year by loading the sql database in python. See the example here. Don't do the intervals (little vertical lines), only the points.

[3]:
```
import matplotlib.pyplot as plt

# Compute average pills per county per year
avg_pills = all_data.groupby("year")["Pills_in_millions"].mean().reset_index()

plt.figure(figsize=(8,5))
plt.plot(avg_pills["year"], avg_pills["Pills_in_millions"], marker='o')
```

```
plt.title("Average Pills Shipped to US Counties per Year")
plt.xlabel("Year")
plt.ylabel("Average Pills (millions)")
plt.grid(True)
plt.show()
```



3. Repeat the steps of loading and merging the opioid data files in R. That is, follow the steps of loading and merging the three csv files as well as the data cleaning described in the notes in R.

4. Take your R code from the previous step and call it from python instead of R. Convert the resulting dataset to a pandas dataframe.

[4]:
```
# import os
# os.environ['R_HOME'] = r"C:\PROGRA~1\R\R-43~1.0"
# os.environ['R_USER'] = r"C:\Users\rexsw\Documents"

import rpy2.robjects as ro

ro.r(''' # The code runs fine in RStudio but here python can't seem to locate␣
 ↪the libraries
library(dplyr)
library(tidyr)
```

```
library(tibble)
library(magrittr)
library(ggplot2)

# import data
annualDosage <- read.csv("county_annual.csv"); annualDosage <-␣
 ↪as_tibble(annualDosage);
county_pop <- read.csv("county_pop_arcos.csv"); county_pop <-␣
 ↪as_tibble(county_pop);
land <- read.csv("land_area.csv"); land <- as_tibble(land)

# select land area from 2010
land_area <-
  land %>%
  select(Areaname, STCOU, LND110210D)

# fill in fips code for Montgomery, AR
annualDosage %<>%
  mutate(countyfips = case_when(BUYER_STATE == "AR" &
                                BUYER_COUNTY == "MONTGOMERY" ~ 05097,
                                TRUE ~ countyfips))

# remove American terrirories with no fips code
annualDosage %<>%
  filter(!is.na(countyfips))

# join population with land area
land_area %<>%
  rename(countyfips = STCOU) # match column name
county_info <-
  left_join(x = county_pop, y = land_area, by = "countyfips")

# join county info with drug data
annualDosage %<>%
  mutate(countyfips = as.factor(countyfips),
         year = as.factor(year))
county_info %<>%
  mutate(countyfips = as.factor(countyfips),
         year = as.factor(year))
Annual <-
  left_join(annualDosage, county_info, by = c("BUYER_COUNTY",
                                              "BUYER_STATE",
                                              "year",
                                              "countyfips"))

# remove NA
Annual %<>%
```

```
  filter(!is.na(STATE))

# add column pills in millions
Annual %<>%
  mutate(Pills_in_millions = DOSAGE_UNIT/1000000)

# plot mean pills per county per year
raw_average <-
  Annual %>%
  ggplot(aes(x = year, y = Pills_in_millions, group = 1)) +
  stat_summary(fun = mean, geom = "point", size = 2) +
  labs(title = "Average Number of Opioid Pills Shipped to a US County",
       y = "Number of pills in millions") +
  theme_minimal()

raw_average
''')
```

```
Error importing in API mode: ImportError('On Windows, cffi mode "ANY" is only
"ABI".')
Trying to import in ABI mode.
R callback write-console: Error in library(dplyr) : there is no package called
'dplyr'
```

```
---------------------------------------------------------------------------
RRuntimeError                             Traceback (most recent call last)
Cell In[4], line 7
      1 # import os
      2 # os.environ['R_HOME'] = r"C:\PROGRA~1\R\R-43~1.0"
      3 # os.environ['R_USER'] = r"C:\Users\rexsw\Documents"
      5 import rpy2.robjects as ro
----> 7␣
  ↪ro.r(''' # The code runs fine in RStudio but here python can't seem to locate the librarie
      8 library(dplyr)
      9 library(tidyr)
     10 library(tibble)
     11 library(magrittr)
     12 library(ggplot2)
     13
     14 # import data
     15␣
  ↪annualDosage <- read.csv("county_annual.csv"); annualDosage <- as_tibble(annu lDosage);
     16␣
  ↪county_pop <- read.csv("county_pop_arcos.csv"); county_pop <- as_tibble(count _pop);
     17 land <- read.csv("land_area.csv"); land <- as_tibble(land)
     18
```

```r
# select land area from 2010
land_area <-
  land %>%
  select(Areaname, STCOU, LND110210D)

# fill in fips code for Montgomery, AR
annualDosage %<>%
  mutate(countyfips = case_when(BUYER_STATE == "AR" &
                                BUYER_COUNTY == "MONTGOMERY" ~ 05097,
                                TRUE ~ countyfips))

# remove American terrirories with no fips code
annualDosage %<>%
  filter(!is.na(countyfips))

# join population with land area
land_area %<>%
  rename(countyfips = STCOU) # match column name
county_info <-
  left_join(x = county_pop, y = land_area, by = "countyfips")

# join county info with drug data
annualDosage %<>%
  mutate(countyfips = as.factor(countyfips),
         year = as.factor(year))
county_info %<>%
  mutate(countyfips = as.factor(countyfips),
         year = as.factor(year))
Annual <-
  left_join(annualDosage, county_info, by = c("BUYER_COUNTY",
                                              "BUYER_STATE",
                                              "year",
                                              "countyfips"))

# remove NA
Annual %<>%
  filter(!is.na(STATE))

# add column pills in millions
Annual %<>%
  mutate(Pills_in_millions = DOSAGE_UNIT/1000000)

# plot mean pills per county per year
raw_average <-
  Annual %>%
  ggplot(aes(x = year, y = Pills_in_millions, group = 1)) +
  stat_summary(fun = mean, geom = "point", size = 2) +
  labs(title = "Average Number of Opioid Pills Shipped to a US County",
```

```
     67         y = "Number of pills in millions") +
     68   theme_minimal()
     69
     70 raw_average
     71 ''')

File c:
 ↪\Users\rexsw\AppData\Local\Programs\Python\Python313\Lib\site-packages\rpy2\robjects\__init
 ↪py:552, in R.__call__(self, string, invisible, print_r_warnings)
    550     invisible = self._invisible
    551 if invisible:
--> 552     res, visible = rinterface.evalr_expr_with_visible(   # type: ignore
    553         r_expr
    554     )
    555     if not visible[0]:  # type: ignore
    556         res = None

File c:
 ↪\Users\rexsw\AppData\Local\Programs\Python\Python313\Lib\site-packages\rpy2\rinterface\__in
 ↪py:205, in evalr_expr_with_visible(expr, envir)
    198 r_res = rmemory.protect(
    199         openrlib.rlib.R_tryEval(
    200             r_call,
    201             envir.__sexp__._cdata,  # call context.
    202             error_occured)
    203 )
    204 if error_occured[0]:
--> 205     raise embedded.RRuntimeError(_rinterface._geterrmessage())
    206 res = conversion._cdata_to_rinterface(r_res)
    207 assert isinstance(res, ListSexpVector)

RRuntimeError: Error in library(dplyr) : there is no package called 'dplyr'
```

[ ]:
```
from rpy2.robjects import pandas2ri
from rpy2.robjects.conversion import localconverter

# load r data
ro.r('df <- readRDS("data.rds")')  # or however you load it
r_df = ro.r['df']

# convert to pd dataframe
with localconverter(ro.default_converter + pandas2ri.converter):
    py_df = ro.conversion.rpy2py(r_df)

# Check
print(py_df.head())
```

```
   X.x BUYER_COUNTY BUYER_STATE  year  count  DOSAGE_UNIT countyfips    X.y  \
1    1    ABBEVILLE          SC  2006    877     363620.0      45001   2313
2    2    ABBEVILLE          SC  2007    908     402940.0      45001   5455
3    3    ABBEVILLE          SC  2008    871     424590.0      45001   8597
4    4    ABBEVILLE          SC  2009    930     467230.0      45001  11737
5    5    ABBEVILLE          SC  2010   1197     539280.0      45001  14877

   STATE  COUNTY county_name                              NAME    variable  \
1     45       1   Abbeville  Abbeville County, South Carolina  B01003_001
2     45       1   Abbeville  Abbeville County, South Carolina  B01003_001
3     45       1   Abbeville  Abbeville County, South Carolina  B01003_001
4     45       1   Abbeville  Abbeville County, South Carolina  B01003_001
5     45       1   Abbeville  Abbeville County, South Carolina  B01003_001

   population       Areaname  LND110210D  Pills_in_millions
1       25821  Abbeville, SC      490.48            0.36362
2       25745  Abbeville, SC      490.48            0.40294
3       25699  Abbeville, SC      490.48            0.42459
4       25347  Abbeville, SC      490.48            0.46723
5       25643  Abbeville, SC      490.48            0.53928
```