capstone_project_part0

July 24, 2021

Modeling and Forecasting Crime Rate in Colorado

Data Science Capstone Project, part 0,creating SQLite database * Student name: Elena Kazakova * Student pace: Full-time * Cohort: DS02222021 * Scheduled project review date: 07/26/2021 * Instructor name: James Irving * Application url: TBD

0.1 TABLE OF CONTENTS

This notebook is part ZERO of the project. Its' goal is to create an SQLite database from the csv files downloaded from the CDE FBI public data depository.

This notebook takes 3 minutes to run

1 Imports

[1]:

```
# Importing packages
import pandas as pd
from pandasql import sqldf
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import gzip
import shutil
import os
import sqlite3
import db_to_sqlite
from sqlite3 import Error
import csv
from pathlib import Path
import subprocess
import io
import warnings
warnings.filterwarnings(action='ignore', category=FutureWarning)
from functions_all import *


%reload_ext autoreload
%autoreload 2
%matplotlib inline
```

2 DataBase

2.1 Data

2.1.1 Data source

Data is from FBI Crime Data Explorer NIBRS data for Colorado from 2009-2019

The data dictionary is and a record descriptiopn are available.

The description of the main and reference tables is in data/README.md file. The agency implemented some changes to the files structure in 2016 and removed the sqlite create and load scripts from the zip directories. Another fact worth mentioning is that files 'nibrs_property_desc.csv' from 2014 and 2015 have duplicated nibrs_property_desc_ids (unique identifier in the nibrs_property_desc table) which complicated the loading of the data.

Loading it all up into SQLite database for easy access

2.1.2 Pre-processing original csv files

Before re-running the code if needed, remove the existing database

[2]:

```
# cur.close()
# conn.close()
```

[3]:

```
!rm data/sqlite/db/production1.db
```

[4]:

```
#I created a separate directory with only incident data files as a template for␣
 ↪lists of data (not reference tables)
# from 2009-2015

list_template_early=os.listdir('data/incidents/template_data/')
list_template_early
```

[4]:

```
['agency_participation.csv',
 'cde_agencies.csv',
 'nibrs_arrestee.csv',
 'nibrs_arrestee_weapon.csv',
 'nibrs_bias_motivation.csv',
 'nibrs_criminal_act.csv',
 'nibrs_incident.csv',
 'nibrs_month.csv',
 'nibrs_offender.csv',
 'nibrs_offense.csv',
 'nibrs_property.csv',
 'nibrs_property_desc.csv',
 'nibrs_suspected_drug.csv',
 'nibrs_suspect_using.csv',
 'nibrs_victim.csv',
 'nibrs_victim_circumstances.csv',
 'nibrs_victim_injury.csv',
 'nibrs_victim_offender_rel.csv',
 'nibrs_victim_offense.csv',
 'nibrs_weapon.csv']
```

[5]:

```
# List of incident data fiels from 2016-2019

list_template_late=list_template_early[2:]
list_template_late.append('agencies.csv')
list_template_late
```

[5]:

```
['nibrs_arrestee.csv',
 'nibrs_arrestee_weapon.csv',
 'nibrs_bias_motivation.csv',
 'nibrs_criminal_act.csv',
 'nibrs_incident.csv',
 'nibrs_month.csv',
 'nibrs_offender.csv',
 'nibrs_offense.csv',
 'nibrs_property.csv',
 'nibrs_property_desc.csv',
 'nibrs_suspected_drug.csv',
 'nibrs_suspect_using.csv',
 'nibrs_victim.csv',
 'nibrs_victim_circumstances.csv',
 'nibrs_victim_injury.csv',
 'nibrs_victim_offender_rel.csv',
 'nibrs_victim_offense.csv',
 'nibrs_weapon.csv',
 'agencies.csv']
```

I commented out the following cell to avoid overwriting changes to the directories

[6]:

```
# copy_files(list_template_early, '/Users/elena/Desktop/FBI_crime_files/CO-2009/
 ↪', 'data/incidents/2009/')
# copy_files(list_template_early, '/Users/elena/Desktop/FBI_crime_files/CO-2010/
 ↪', 'data/incidents/2010/')
# copy_files(list_template_early, '/Users/elena/Desktop/FBI_crime_files/CO-2011/
 ↪', 'data/incidents/2011/')
# copy_files(list_template_early, '/Users/elena/Desktop/FBI_crime_files/CO-2012/
 ↪', 'data/incidents/2012/')
# copy_files(list_template_early, '/Users/elena/Desktop/FBI_crime_files/CO-2013/
 ↪', 'data/incidents/2013/')
# copy_files(list_template_early, '/Users/elena/Desktop/FBI_crime_files/CO-2014/
 ↪', 'data/incidents/2014/')
# copy_files(list_template_early, '/Users/elena/Desktop/FBI_crime_files/CO-2015/
 ↪', 'data/incidents/2015/')
```

[7]:

```
# copy_files(list_template_late, '/Users/elena/Desktop/FBI_crime_files/CO-2016/
 ↪', 'data/incidents/2016/')
# copy_files(list_template_late, '/Users/elena/Desktop/FBI_crime_files/CO-2017/
 ↪', 'data/incidents/2017/')
# copy_files(list_template_late, '/Users/elena/Desktop/FBI_crime_files/CO-2018/
 ↪', 'data/incidents/2018/')
# copy_files(list_template_late, '/Users/elena/Desktop/FBI_crime_files/CO-2019/
 ↪', 'data/incidents/2019/')
```

[8]:

```
# Initiating a cursor
conn = sqlite3.connect('data/sqlite/db/production1.db')
cur = conn.cursor()
```

[9]:

```
q="SELECT name FROM sqlite_master WHERE type='table'"
df=table_query(q, cur=cur)
df
```

```
Nothing was found
```

[10]:

```
# Creating tables in the database
sql_file = open('script_to_create_tables.sql')
sql_as_string = sql_file.read()
cur.executescript(sql_as_string)
```

[10]:

```
<sqlite3.Cursor at 0x27e76dd9b90>
```

[11]:

```
q="SELECT name FROM sqlite_master WHERE type='table'"
df=table_query(q, cur=cur)
df
```

[11]:

```
                          name
0                     agencies
1         agency_participation
2                 cde_agencies
3          nibrs_activity_type
4                    nibrs_age
5            nibrs_arrest_type
6        nibrs_assignment_type
7              nibrs_bias_list
8          nibrs_location_type
9           nibrs_offense_type
10        nibrs_prop_desc_type
11           nibrs_victim_type
12         nibrs_circumstances
13        nibrs_cleared_except
14          nibrs_criminal_act
15     nibrs_criminal_act_type
16     nibrs_drug_measure_type
17             nibrs_ethnicity
18                nibrs_injury
19     nibrs_justifiable_force
20        nibrs_prop_loss_type
21          nibrs_relationship
22   nibrs_suspected_drug_type
23            nibrs_using_list
24           nibrs_weapon_type
25                    ref_race
26                   ref_state
27              nibrs_arrestee
28       nibrs_arrestee_weapon
29       nibrs_bias_motivation
30                 nibrs_month
31              nibrs_incident
32              nibrs_offender
33               nibrs_offense
34              nibrs_property
```

```
35          nibrs_property_desc
36          nibrs_suspect_using
37         nibrs_suspected_drug
38                 nibrs_victim
39  nibrs_victim_circumstances
40          nibrs_victim_injury
41    nibrs_victim_offender_rel
42         nibrs_victim_offense
43                 nibrs_weapon
```

[12]:

```
display_csvfileDF('nibrs_age.csv', 'Ref_tables/')
```

[12]:

```
   age_id age_code           age_name
0       1       NN     Under 24 Hours
1       2       NB       1-6 Days Old
2       3       BB     7-364 Days Old
3       4       00            Unknown
4       5       AG       Age in Years
5       6       99  Over 98 Years Old
```

[13]:

```
q="""SELECT * FROM nibrs_age"""
df=table_query(q, cur=cur)
df
```

```
Nothing was found
```

[14]:

```
#All reference table files are in this directory, the actual incident data␣
↪files are in all data/incidents, split by years
!ls -al data/Ref_tables/
```

```
total 56
drwxr-xr-x 1 elena 197121    0 Jun 30 14:49 .
drwxr-xr-x 1 elena 197121    0 Jul 14 00:47 ..
-rw-r--r-- 1 elena 197121  477 Jun 30 14:44 nibrs_activity_type.csv
-rw-r--r-- 1 elena 197121  137 Jun 30 14:44 nibrs_age.csv
-rw-r--r-- 1 elena 197121  105 Jun 30 14:44 nibrs_arrest_type.csv
-rw-r--r-- 1 elena 197121  266 Jun 30 14:44 nibrs_assignment_type.csv
-rw-r--r-- 1 elena 197121  993 Jun 30 14:44 nibrs_bias_list.csv
-rw-r--r-- 1 elena 197121  556 Jun 30 14:44 nibrs_circumstances.csv
-rw-r--r-- 1 elena 197121  217 Jun 30 14:44 nibrs_cleared_except.csv
-rw-r--r-- 1 elena 197121  442 Jun 30 14:44 nibrs_criminal_act_type.csv
-rw-r--r-- 1 elena 197121  218 Jun 30 14:44 nibrs_drug_measure_type.csv
-rw-r--r-- 1 elena 197121  134 Jun 30 14:44 nibrs_ethnicity.csv
-rw-r--r-- 1 elena 197121  194 Jun 30 14:44 nibrs_injury.csv
-rw-r--r-- 1 elena 197121  436 Jun 30 14:44 nibrs_justifiable_force.csv
-rw-r--r-- 1 elena 197121 1238 Jun 30 14:44 nibrs_location_type.csv
-rw-r--r-- 1 elena 197121 3811 Jun 30 14:44 nibrs_offense_type.csv
-rw-r--r-- 1 elena 197121 1696 Jun 30 14:44 nibrs_prop_desc_type.csv
```

```
-rw-r--r-- 1 elena 197121  142 Jun 30 14:44 nibrs_prop_loss_type.csv
-rw-r--r-- 1 elena 197121  793 Jun 30 14:44 nibrs_relationship.csv
-rw-r--r-- 1 elena 197121  354 Jun 30 14:44 nibrs_suspected_drug_type.csv
-rw-r--r-- 1 elena 197121  129 Jun 30 14:44 nibrs_using_list.csv
-rw-r--r-- 1 elena 197121  214 Jun 30 14:44 nibrs_victim_type.csv
-rw-r--r-- 1 elena 197121  717 Jun 30 14:44 nibrs_weapon_type.csv
-rw-r--r-- 1 elena 197121  639 Jun 30 14:49 ref_race.csv
-rw-r--r-- 1 elena 197121 1883 Jun 30 14:49 ref_state.csv
```

2.1.3 Importing files into the database

[15]:

```
# Creating a list of ref table files to import them into tables
files_ref=create_filelist('data/Ref_tables/',n=0)
files_ref
```

[15]:

```
['data/Ref_tables/nibrs_activity_type.csv',
 'data/Ref_tables/nibrs_age.csv',
 'data/Ref_tables/nibrs_arrest_type.csv',
 'data/Ref_tables/nibrs_assignment_type.csv',
 'data/Ref_tables/nibrs_bias_list.csv',
 'data/Ref_tables/nibrs_circumstances.csv',
 'data/Ref_tables/nibrs_cleared_except.csv',
 'data/Ref_tables/nibrs_criminal_act_type.csv',
 'data/Ref_tables/nibrs_drug_measure_type.csv',
 'data/Ref_tables/nibrs_ethnicity.csv',
 'data/Ref_tables/nibrs_injury.csv',
 'data/Ref_tables/nibrs_justifiable_force.csv',
 'data/Ref_tables/nibrs_location_type.csv',
 'data/Ref_tables/nibrs_offense_type.csv',
 'data/Ref_tables/nibrs_prop_desc_type.csv',
 'data/Ref_tables/nibrs_prop_loss_type.csv',
 'data/Ref_tables/nibrs_relationship.csv',
 'data/Ref_tables/nibrs_suspected_drug_type.csv',
 'data/Ref_tables/nibrs_using_list.csv',
 'data/Ref_tables/nibrs_victim_type.csv',
 'data/Ref_tables/nibrs_weapon_type.csv',
 'data/Ref_tables/ref_race.csv',
 'data/Ref_tables/ref_state.csv']
```

[16]:

```
#importing data into reference tables
import_data_to_tables('data/sqlite/db/production1.db', files_ref, 'data/
 ↪Ref_tables/')
```

[17]:

```
q='SELECT * FROM nibrs_using_list'
df=table_query(q, cur)
df
```

```
[17]:    suspect_using_id suspect_using_code  suspect_using_name
      0                 1                  A             Alcohol
      1                 2                  C  Computer Equipment
      2                 3                  D     Drugs/Narcotics
      3                 4                  N      Not Applicable
```

```
[18]: # Importing incidents data from 2009-2015 to the database

      list_inc_2009=create_filelist('data/incidents/2009/', n=0)
      import_data_to_tables('data/sqlite/db/production1.db', list_inc_2009, 'data/
       ↪incidents/2009/')

      list_inc_2010=create_filelist('data/incidents/2010/', n=0)
      import_data_to_tables('data/sqlite/db/production1.db', list_inc_2010, 'data/
       ↪incidents/2010/')

      list_inc_2011=create_filelist('data/incidents/2011/', n=0)
      import_data_to_tables('data/sqlite/db/production1.db', list_inc_2011, 'data/
       ↪incidents/2011/')

      list_inc_2012=create_filelist('data/incidents/2012/', n=0)
      import_data_to_tables('data/sqlite/db/production1.db', list_inc_2012, 'data/
       ↪incidents/2012/')

      list_inc_2013=create_filelist('data/incidents/2013/', n=0)
      import_data_to_tables('data/sqlite/db/production1.db', list_inc_2013, 'data/
       ↪incidents/2013/')

      list_inc_2014=create_filelist('data/incidents/2014/', n=0)
      import_data_to_tables('data/sqlite/db/production1.db', list_inc_2014, 'data/
       ↪incidents/2014/')

      list_inc_2015=create_filelist('data/incidents/2015/', n=0)
      import_data_to_tables('data/sqlite/db/production1.db', list_inc_2015, 'data/
       ↪incidents/2015/')
```

```
[19]: q='SELECT * FROM nibrs_incident'
      df=table_query(q,cur)
      len(df)
```

```
[19]: 1701394
```

2.1.4 Starting clean-up

All 2016-2019 files need to be cleaned up because FBI changed the file format. There is a YEAR column that needs to be removed as well as the legacy columns from the previous years need to be added up. It's a tedious job and it needs to be done once

and the files need to be backed up.

In order to clean the tables up the following needs to be done

1. Remove all **DATA_YEAR** columns from each file, it's the first column
2. Files that do not need any changes beyond **DATA_YEAR** column removal

   nibrs_arrestee_weapon.csv nibrs_bias_motivation.csv nibrs_criminal_act.csv nibrs_property_desc.csv nibrs_suspect_using.csv nibrs_suspected_drug.csv nibrs_victim_circumstances.csv nibrs_victim_injury.csv nibrs_victim_offender_rel.csv nibrs_victim_offense.csv nibrs_weapon.csv

3. in **nibrs_arestee.csv file**:

a. between **ARRESTEE_SEQ_NUM** and **ARREST_DATE** there should be an **arrest_num column**
b. Between **CLEARANCE_IND** and **AGE_RANGE_LOW_NUM** should be a **ff_line_number** column.

4. in **nibrs_incident** file: a.between **NIBRS_MONTH_ID** and **CARGO_THEFT_FLAG** column **incident_number** b.between **DATA_HOME** and **ORIG_FORMAT** column **ddocname** c.between **ORIG_FORMAT** and **DID** column **ff_line_number**
5. in **nibrs_month.csv** file: a.between **REPORT_DATE** and **UPDATE_FLAG** add **prepared_date** column b.between **ORIG_FORMAT** and **DATA_HOME** column **ff_line_number** c.column **MONTH_PUB_STATUS** removed
6. in **nibrs_offender.csv** file: a.between **ETHNICITY_ID** and **AGE_RANGE_LOW_NUM** column **ff_line_number**
7. in **nibrs_offense.csv** file:

   a. the last column **ff_line_number** should be added

8. in **nibrs_property.csv** file:

   a. the last column **ff_line_number** should be added

9. in **nibrs_victim.csv** file:

   a. between **RESIDENT_STATUS_CODE** and **AGE_RANGE_LOW_NUM** two columns **agency_data_year** and **ff_line_number** (in that order) should be added

[20]:
```
# Importing cleaned-up files to the tables and checking numbers along the way

list_inc_2016=create_filelist('data/incidents/2016/', n=0)
import_data_to_tables('data/sqlite/db/production1.db', list_inc_2016, 'data/
 ↪incidents/2016/')
```

[21]:
```
q='SELECT * FROM nibrs_incident'
df=table_query(q,cur)
len(df)
```

[21]: 1983733

[22]:
```
list_inc_2017=create_filelist('data/incidents/2017/', n=0)
import_data_to_tables('data/sqlite/db/production1.db', list_inc_2017, 'data/
 ↪incidents/2017/')
```

[23]:
```
q='SELECT * FROM nibrs_incident'
df=table_query(q, cur)
len(df)
```

[23]: 2269247

[24]:
```
list_inc_2018=create_filelist('data/incidents/2018/', n=0)
import_data_to_tables('data/sqlite/db/production1.db', list_inc_2018, 'data/
 ↪incidents/2018/')
```

[25]:
```
q='SELECT * FROM nibrs_incident'
df=table_query(q, cur)
len(df)
```

[25]: 2556043

[26]:
```
list_inc_2019=create_filelist('data/incidents/2019/', n=0)
import_data_to_tables('data/sqlite/db/production1.db', list_inc_2019, 'data/
 ↪incidents/2019/')
```

[27]:
```
q='SELECT * FROM nibrs_incident'
df=table_query(q, cur)
len(df)
```

[27]: 2819463

[28]:
```
df.head()
```

[28]:
```
   agency_id  incident_id  nibrs_month_id incident_number cargo_theft_flag  \
0       1971     51264520         4814762        09000019
1       1971     51264521         4814762        09000053
2       1971     51264523         4814762        09000082
3       1971     51264524         4814762        09000092
4       1971     51264525         4814762        09000097

  submission_date        incident_date report_date_flag incident_hour  \
0                  2009-01-05 00:00:00                             22
1                  2009-01-13 00:00:00
2                  2009-01-17 00:00:00                             19
3                  2009-01-20 00:00:00                R
4                  2009-01-21 00:00:00
```

```
   cleared_except_id cleared_except_date  incident_status data_home  \
0                  6                                    0         C
1                  6                                    0         C
2                  6                                    0         C
3                  6                                    0         C
4                  6                                    0         C

                               ddocname orig_format ff_line_number did
0  2009_01_CO0320000_09000019_INC_NIBRS
1  2009_01_CO0320000_09000053_INC_NIBRS
2  2009_01_CO0320000_09000082_INC_NIBRS
3  2009_01_CO0320000_09000092_INC_NIBRS
4  2009_01_CO0320000_09000097_INC_NIBRS
```

[29]:

```
cur.close()
conn.commit()
conn.close()
```

This notebook is the pre-work for the notebook, part I. The link to the next notebook here.

All the tables from 2009-2019 incidents in Colorado is in production1 database in data/sqlite/db folder. The database can be used moving forward.