ETL Project

Jihye Choi

My topic is Gun violation so I looked into how many minors died in 2019 and how many murders have been happened in 2016. (I couldn't find murder data for 2019 so I used data for 2016)

Extract Data

I used 4 different datasets from the public platform Kaggle which lead me to the Gun Violence Archive site, https://www.gunviolencearchive.org/reports. And I found 2 table on the Gun to Carry site, https://www.gunstocarry.com/concealed-carry-statistics/#numbers and https://woocommerce-262453-958953.cloudwaysapps.com/concealed-carry-statistics/concealed-carry-statistics-fbi/.

- Children Killed.csv
- Teens Killed.csv
- Accidental Deaths (Children Age 0-11).csv
- Accidental Deaths (Teens Ages 12-17).csv
- % of Population with Permit (table)
- 2016 Murder Statics - Weapon Used(table)

Transform

I wanted to know the total number of underages killed in 2019, so I used 4 CSV datasets from Gun Violence Archive site and dropped some columns that I didn't need and filtered only 2019. And merged them into one data frame since each file had a number that I needed (Children Killed, Teens Killed, Accidental Deaths) and added a total minor killed column to sum the numbers.
And I got some data from tables so I used panda to read HTML and got the tables. Also, I added columns that I needed and dropped that I didn't need and saved the data frame to CSV format.

Load

I created the following SQL tables:

- gun_violance table was for the total number of underages killed in 2019
- guntocarry table was for the % of population with permit to have gun
- weapon_used table was for the number of murders by gun in 2016.

And it could be also uploaded CSV format.