Task 3 - Feature Engineering

Goal:

1. Avoid Geo data missing.
2. Aggregate data

Data in:

1. Task2
2. MedianHouseholdIncome2015.csv
3. PercentagePeopleBelowPovertyLevel.csv
4. PercentOver25CompletedHighSchool.csv
5. ShareRaceByCity.csv
6. State.xlsx
7. US Cities.xlsx

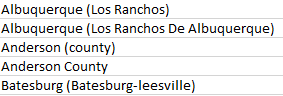
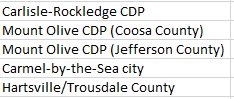
Data out:

Task3-aggregate to Weekly/Monthly /Cities

**Challenge:**

1. **How to match the cities**

**e.g.**



* Need to go back to the metadata to find why the cities cannot match.
* Check specific symbol/ stop words/ uppercase/ space
* The length different (e.g. last string is space)

1. aggregate the data in the meaningful way.

* We try to aggregate Weekly and monthly data, but finally we do the city level. That is most fit out goal

**Performance:**

1. Before aggregate, missing of geo data(same dataframe) reduce **from 18% to 8%**
2. But we still has 30% missing by cities geo data 🙁
3. Missing reason:
   1. Original missing
   2. city cannot exact match,
   3. state, region and city typo
4. Number of incident by cities can be considered as our objective variables. then using linear regression to build model
5. We assume our model results seems like:
   1. “XXX(city or area) occur more gun shooting incidents”
   2. “XX(race/income/education%..) might related more shooting
   3. Participant type or gun type might related to the number of incidents

**Explanation :**

1. how to we explain the 30% missing value -- **percentage missing data acceptable**

* “Missing data are a rule rather than an exception in quantitative research. Enders ( 2003 ) stated that a missing rate of 15% to 20% was common in educational and psychological studies.”

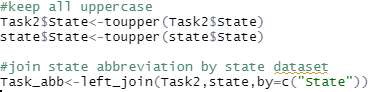
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/#:~:text=Missing%20data%20are%20a%20rule,in%20educational%20and%20psychological%20studies>.

* In the real-life data, theoretically, 25%-30% of missing values are allowed.
* And we still have over 10000 non-missing data.

**Technical Step:**

**PART ONE – JOIN IN AND REDUCE MISSING**

1. Create table Task\_abb to Add the state abbreviation to Task2 (keep all to uppercase)



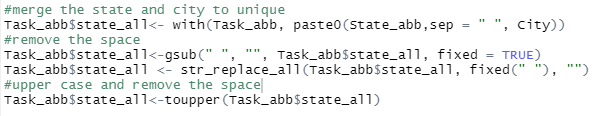
\*Note: state.abb[match(Task2$State,state.name)] cause missing. So we try to import the full state and state abbreviation.

* After join in, State\_abb is created and no missing to this column.

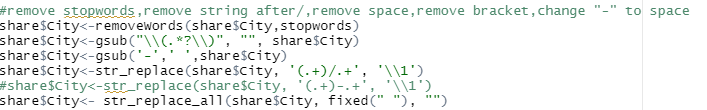
1. clean column “City” in both joined table and Task2
   1. Remove string in Bracket
   2. Replace “Saint” to “st.” -- this step need to check specific city from dataset.
   3. Remove Words like “city,CDP,country,town...”



1. Merge state\_abb and city because there are same cities in different state. Then clean
   1. Merge
   2. Remove all space (double check the space at the end
   3. To uppercase

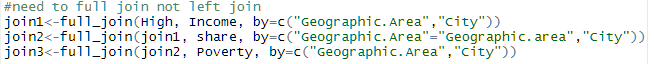


1. Import geographic data and clean the city column
   1. Remove stop-word
   2. Remove string in bracket
   3. Remover strings after “/”
   4. Change “-” to space
   5. Remove all space



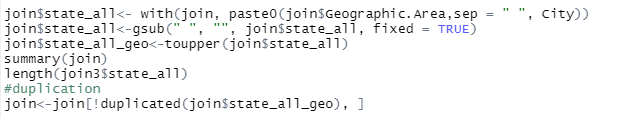
* Need to repeat same way to clean for the 4 datasets

1. join geographic dataset(Income, poverty level, high school completed and share race) by ‘Geographic area’ and ‘city’



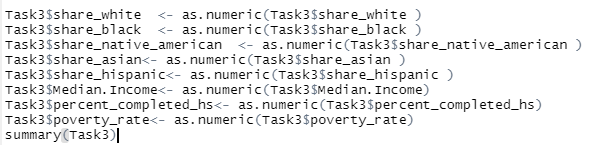
* We want to keep all information, so we need use full join

1. Merge state and city in joined geo table named “join” to keep city unique in different state for task2 and joined geo table
   1. Remove all blank (check length of strings)
   2. To all upper case
   3. Remove the duplicated.



1. Left join in Task2 and join3 and replace the space and change the type of variables.





* The missing from geographic data is under 8%.
* The reason of missing because:
  + original missing from geo dataset
  + Cannot 100% matching about the city’s name

**PART TWO -- AGGREGATE**

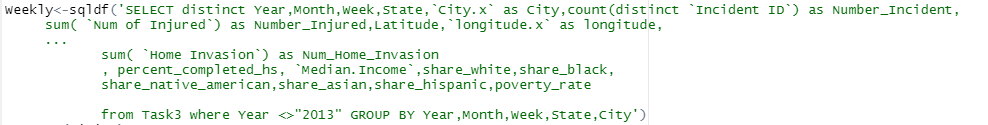
1. **Aggregate to weekly /monthly level** – create column week and year based on date

Task3$Week <- week(Task3$date)

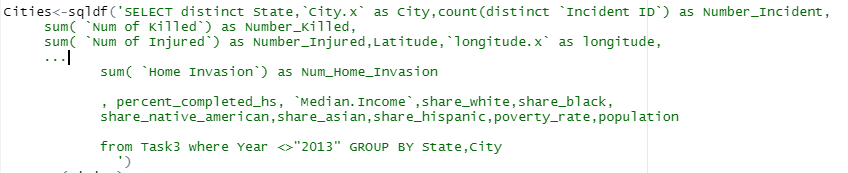
Task3$Month <- month(Task3$date)

Task3$Year <- substring(Task3$date,1,4)

1. Count Incident ID and sum rest of column(without geo data) group by week, year, state,city using SQL in r
   1. Weekly or monthly

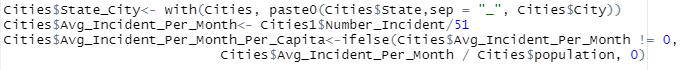


* 1. Cities

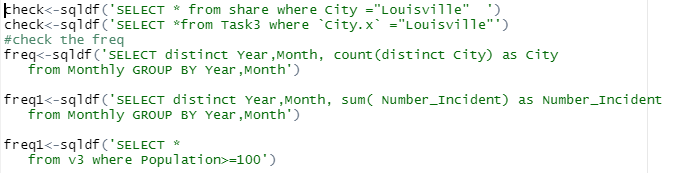


Note: After check the frequency, we found the 2013 data is weird(too small). So we decided not use data from 2013

* 1. Merge state and city, Remove all the space
  2. Create the new columns - Avg\_Incident\_Per\_Month & Avg\_Incident\_Per\_Month\_Per\_Capita



1. Some check about missing and frequency.



Note: need to check specific cities that is missing geo data

1. Export data to excel

