Task 2 Process and summary Report

**The objective:** To do the essential data cleansing and explore the task 1 joined dataset, including understand the variable identification, dataset structure, univariate & bivariate analysis, deal with partial missing value and transformation.

**The input dataset description:** Task one final dataset

**The output dataset description:**

* Total 50 columns and 240092 rows
* There are two char variables are ‘State’ and ‘City’
* There is one data type variable called ‘date’ from 2013 to 2018
* There is the summary of all columns

Table

Description automatically generated

**The process:**

1. Load all necessary libraries and task one file to the R studio
2. Remove some useless or redounded columns:

Address: our model will only focus on the city scop, not on the street scop

Participant\_age: we already have the participant\_agegroup column and will only focus on age group instead of the specific ages

Participabnt\_name: we do not care about the individual participant’s name

City\_and\_state: we have city and state columns separately

State\_id: we already have a state column

Latitude.y: we already have the latitude. This is the redundant column

longitude.y: we already have the longitude. This is the redundant column

1. Change median\_income, poverty\_rate and high\_school\_completion\_rate into numeric value
2. Split ‘participant\_gender’ into ‘Number of Male(Participant)’ and ‘Number of Female(Participant)’, then change NA to 0

Graphical user interface, text, application

Description automatically generated

Before and after:

Table

Description automatically generated Table

Description automatically generated

1. Split ‘participant\_age\_group’ into ‘Number of Adult’, ‘Number of Teen’ and ‘Total # Child’, then change NA to 0

Before and after:

Table

Description automatically generated Table

Description automatically generated

1. Create the total # Participant which by the max of total genders and the total number of age groups
2. Repeat the steps 4 and 5 methods to dealt with ‘Participant Type’ and ‘gun types’, ‘gun\_stolen’ and ‘participant\_relationship’
3. Consolidate all gun types into seven main types are ‘Unknown(gun type)’,’ SG(gun type)’,’ Other(gun type)’,’ H(gun type)’, HR(gun type),’ Hunting Rifle(gun type)’ and ‘AR(gun type)’

* In this step we used the ‘gun legend’ datafile as a referenceCreate a flag called ‘massive shooting’, 1 means yes and 0 means no

1. Final clean up some columns and remove them
2. Rename the columns into the meaningful namesText, letter

   Description automatically generated
3. Check the missing value. They are majority came for demographic columns, will do imputation in the next step.

A picture containing text

Description automatically generated

1. Summary statistical analysis

Text

Description automatically generated

1. Bivariable analysis Text

   Description automatically generatedText

   Description automatically generated with medium confidence
2. Some basic visualization in R and tableau:

Chart, bar chart

Description automatically generatedChart

Description automatically generated

Graphical user interface

Description automatically generated with low confidenceGraphical user interface

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

1. Export file