**Term Project Part II - Police shooting data (v2)**

**Ripa Shah & Shivranjan Jaganathan**

**Code Used To Make These Plots**

Below are the code chunks used to make these plots.

*# Use the police shooting data and complete the following: #*

*# 1.Describe the police shooting data set by filling out # the table below (follow the 'age' example)*

*#*

*# attribute definition data type % of missing data # age age of the victim at the*

*# time of the incident number 0%*

*#*

*# 2. In the demo code, the correlation between state # incident count and state population was examined. # What other correlation analyses could be done?*

*# For example, does age correlate with weapon used? # Perform a correlation analysis (Pearson or X-squared) # between a pair of variables of your choice.*

**rm**(list = **ls**())

**library**(ggplot2)

**library**(dplyr)

**library**(gridExtra)

**library**(tidyverse)

**library**(hrbrthemes)

**library**(ggthemes)

**library**(corrplot)

**library**(patchwork)

**library**(scales)

**library**(ggcorrplot)

*# Load the data*

*# shooting\_orig <- read\_csv("fatal-police-shootings-data1.csv")* shooting\_orig <-**read\_csv**("fatal-police-shootings-data1.csv", col\_names = TRUE, na="")

*#view the data*

**str**(shooting\_orig)

**sapply**(shooting\_orig, typeof)

*# #examine missing data*

**filter**(shooting\_orig, **!complete.cases**(shooting\_orig))

*# apply(shooting\_orig, 2, function(x) sum(is.na(x))) %>%*

*# sort(decreasing=TRUE)*

*# Get % of Missing data*

**apply**(shooting\_orig, 2, **function**(x) **percent**(**mean**(**is.na**(x)),accuracy=1)) **%>% sort**(decreasing=TRUE)

*# Split date & time*

shooting\_orig <- shooting\_orig **%>%**

**mutate**(date = **format**(**as.POSIXct**(date, format='%Y-%m-%d'))) **%>% mutate**(year = **year**(date)) **%>%**

**mutate**(month = **month**(date))

group <- **function** (string){

**if**(**is.na**(string)) **return** ("NA")

**if**(string **==** "unarmed") **return** ("Unarmed")

**else if** (string **%in% c**("undetermined", "unknown")) **return** ("Undetermined") **else if** (string **==** "vehicle") **return** ("Vehicle")

**else if** (string **%in% c**("gun", "gun;knife",

"other;gun", "replica",

"gun;vehicle", "vehicle;gun")) **return** ("Gun") **else if** (string **%in% c**("knife", "blunt\_object",

"blunt\_object;blunt\_object",

"sharp object", "knife;blunt\_object")) **return** ("Sharp Object")

**else return** ("Other")

}

*#apply group on each element of a vector*

shooting\_orig**$**armedType <- **sapply**(shooting\_orig**$**armed\_with, group)

*# create age category to consolidate different age groups*

shooting\_orig <- shooting\_orig **%>%**

**mutate**(ageCategory = **case\_when**(**between**(age, 0, 10) **~** "0 - 10", **between**(age, 11, 20) **~** "11 - 20", **between**(age, 21, 30) **~** "21 - 30", **between**(age, 31, 40) **~** "31 - 40", **between**(age, 41, 50) **~** "41 - 50", **between**(age, 51, 60) **~** "51 - 60", **between**(age, 61, 70) **~** "61 - 70", **between**(age, 71, 80) **~** "71 - 80", **between**(age, 81, 91) **~** "81 - 91")) **%>% select**(state, age, gender, race, ageCategory, armedType, year)

*# state region mapping*

state\_region\_mapping <- **data.frame**(

state = **c**(

"CA", "CO", "WA", "MT", "ID", "OR", "WY", "UT", "NV", "HI", "ND", "SD", "MN", "WI","MI", "NE","KS", "MO", "IA", "IL", "IN", "OH", "NY", "NJ", "VT", "PA", "MD", "NH", "CT", "RI", "DE", "DC", "MA", "AZ","NM", "OK", "TX",

"SC", "NC", "GA", "FL", "LA", "MS", "AL", "AR", "TN", "KY", "WV", "ME", "VA", "AK"),

region = **c**(

"West", "West", "West", "West", "West", "West", "West", "West","West", "West",

"Midwest", "Midwest", "Midwest", "Midwest","Midwest", "Midwest", "Midwest", "Midwest",

"Midwest", "Midwest", "Midwest", "Midwest",

"Northeast", "Northeast", "Northeast", "Northeast", "Northeast", "Northeast",

"Northeast", "Northeast","Northeast", "Northeast", "Northeast", "Southwest", "Southwest", "Southwest", "Southwest",

"Southeast", "Southeast","Southeast", "Southeast", "Southeast", "Southeast", "Southeast", "Southeast","Southeast", "Southeast", "Southeast", "Southeast","Southeast",

"Southeast"))

*# Join the Region column on ufo dataset*

shooting\_orig\_region <- shooting\_orig **%>%**

**filter**(**!is.na**(state))**%>%**

**filter**(**!is.na**(age))**%>%**

**filter**(**!is.na**(gender))**%>%**

**filter**(gender**!=**"non-binary")**%>%**

**filter**(**!is.na**(race))**%>%**

**filter**(**!is.na**(armedType))**%>%**

**filter**(armedType**!=**"NA")**%>%**

**left\_join**(state\_region\_mapping, by = "state") **%>%**

**select**(state, age, gender, race, ageCategory, armedType, year, region)

shooting\_orig\_reg<- shooting\_orig\_region **%>%**

**filter**(**!is.na**(region)) **%>%**

**group\_by**(region, age) **%>%**

**arrange**(region)**%>%**

**summarise**(wcount = **n**())

*#Q2 create joint data frame to run correlation*

(state\_count <- **arrange**(**count**(shooting\_orig\_region, state), **desc**(n))) (race\_count <- **arrange**(**count**(shooting\_orig\_region, race), **desc**(n))) (weapon\_count <- **arrange**(**count**(shooting\_orig\_region, armedType), **desc**(n)))

(age\_join\_arms <- **inner\_join**(weapon\_count, shooting\_orig\_region[, **c**("armedType", "age")],

by=**c**("armedType")))

(year\_join\_arms <- **inner\_join**(weapon\_count, shooting\_orig\_region[, **c**("armedType", "year")],

by=**c**("armedType")))

(race\_join\_age <- **inner\_join**(race\_count, shooting\_orig\_region[, **c**("race", "age")],

by=**c**("race")))

*# \*\* Correlation relationship between two variables\*\**

*# Perform the correlation test*

cr1 <- **cor.test**(shooting\_orig\_reg**$**age, shooting\_orig\_reg**$**wcount) **print**("Age vs Region")

**print**(cr1)

cr2 <- **cor.test**(race\_join\_age**$**age, race\_join\_age**$**n)

**print**("Age vs Race")

**print**(cr2)

cr3 <- **cor.test**(age\_join\_arms**$**age, age\_join\_arms**$**n)

**print**("Age vs Weapons")

**print**(cr3)

cr4 <- **cor.test**(year\_join\_arms**$**year, year\_join\_arms**$**n)

**print**("Year vs Weapons")

**print**(cr4)

*# Correlation Matrix*

**model.matrix**(**~**0**+**., data=race\_join\_age) **%>%**

**cor**(use="pairwise.complete.obs") **%>%**

**ggcorrplot**(show.diag=FALSE, type="lower", lab=TRUE, lab\_size=2)

**model.matrix**(**~**0**+**., data=shooting\_orig\_reg) **%>%**

**cor**(use="pairwise.complete.obs") **%>%**

**ggcorrplot**(show.diag=FALSE, type="lower", lab=TRUE, lab\_size=2)

**model.matrix**(**~**0**+**., data=year\_join\_arms) **%>%**

**cor**(use="pairwise.complete.obs") **%>%**

**ggcorrplot**(show.diag=FALSE, type="lower", lab=TRUE, lab\_size=2)

**model.matrix**(**~**0**+**., data=age\_join\_arms) **%>%**

**cor**(use="pairwise.complete.obs") **%>%**

**ggcorrplot**(show.diag=FALSE, type="lower", lab=TRUE, lab\_size=2) *# \*\*Correlation relationship Plots\*\**

c1 <- **ggplot**(shooting\_orig\_reg) **+**

**aes**(x = age, y = wcount) **+**

**geom\_point**(colour = "green") **+**

**theme\_minimal**()**+**

*# scale\_fill\_colorblind()+*

**ggtitle**("Age vs Region relationship")**+**

**labs**(y = "count", x = "Age" )**+**

**scale\_x\_continuous**(breaks = **seq**(0,95,5), limits = **c**(0, 95) )**+ scale\_y\_continuous**(breaks = **seq**(0,80,5), limits = **c**(0, 80) ) c1

c2 <- **ggplot**(race\_join\_age) **+**

**aes**(x = age, y = n) **+**

*# geom\_point(colour = "#0c4c8a") +*

**geom\_point**(colour = "red") **+**

**theme\_minimal**()**+**

*# scale\_fill\_colorblind()+*

**ggtitle**("Age vs Race relationship")**+**

**labs**(y = "Count", x = "Age" )**+**

**scale\_y\_continuous**(breaks = **seq**(0,4000,500), limits = **c**(0, 4000) )**+ scale\_x\_continuous**(breaks = **seq**(0,95,5), limits = **c**(0, 95) ) c2

c3 <- **ggplot**(age\_join\_arms) **+**

**aes**(x = age, y = n) **+**

*# geom\_point(colour = "#0c4c8a") +*

**geom\_point**(colour = "blue") **+**

**theme\_minimal**()**+**

*# scale\_fill\_colorblind()+*

**ggtitle**("Age vs Weapons relationship")**+**

**labs**(y = "Count", x = "Age" )**+**

**scale\_y\_continuous**(breaks = **seq**(0,1850,200), limits = **c**(0, 1850) )**+ scale\_x\_continuous**(breaks = **seq**(0,95,5), limits = **c**(0, 95) ) c3

c4 <- **ggplot**(year\_join\_arms) **+**

**aes**(x = year, y = n) **+**

**geom\_point**(colour = "#0c4c8a") **+**

**theme\_minimal**()**+**

*# scale\_fill\_colorblind()+*

**ggtitle**("Year vs Weapons relationship")**+**

**labs**(y = "Count", x = "Year" )**+**

**scale\_y\_continuous**(breaks = **seq**(0,5500,500), limits = **c**(0, 5500) )**+ scale\_x\_continuous**(breaks = **seq**(2015,2023,1), limits = **c**(2015, 2023) ) c4

Q1

Final\_Project 11-1-2023 Part II

## Including Plots

You can also embed plots, for example:

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

*#Police Shooting Data Data warehouse and OLAP* *#Use the police shooting data and complete the following:* *#Describe the police shooting data set by filling out the table below (follow the 'age' #example)* *#attribute definition data type %of missing data* *#age age of the victim at the time of incident number 0%* *#2. In the demo code, the correlation between state incident count and state population was examined. What other correlation analyses could be #done? For example, does age correlate with weapon used? Perform a correlation analysis (Pearson or X-squared) between a pair of variables of your #choice.  
  
 #Post your code and your findings.* setwd("C:/Users/yashs/OneDrive/Desktop/R/ArizonaPro")  
 *#load("PoliceShooting2-OLAP.RData")* getwd()

## [1] "C:/Users/yashs/OneDrive/Desktop/R/ArizonaPro"

setwd("C:/Users/yashs/OneDrive/Desktop/R/ArizonaPro")  
 rm(list = ls())  
 *#install.packages("readr")* *#install.packages("formattable")* *#install.packages("scales")* library(scales)  
 library(readr)

##  
 ## Attaching package: 'readr'

## The following object is masked from 'package:scales':  
 ##  
 ## col\_factor

*#install.packages("tibble")* library("tibble")  
 library("tidyverse")

library("sos")

library("dplyr")  
 library("ggplot2")  
 *#findFn("laply")* *#findFn("ggplot")* *#install.packages("ggplot2", dependencies=TRUE)* *#install.packages("pak")* *#pak::pak("tidyverse/ggplot2")*

RShah\_ShootingData<-read\_csv("fatal-police-shootings-data1.csv", col\_names = TRUE, na="")

## Rows: 8720 Columns: 19  
 ## ── Column specification ────────────────────────────────────────────────────────  
 ## Delimiter: ","  
 ## chr (12): threat\_type, flee\_status, armed\_with, city, county, state, locati...  
 ## dbl (4): id, latitude, longitude, age  
 ## lgl (2): was\_mental\_illness\_related, body\_camera  
 ## date (1): date  
 ##  
 ## ℹ Use `spec()` to retrieve the full column specification for this data.  
 ## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

save(RShah\_ShootingData, file="Example PartII.Rdata")  
  
 typeof(RShah\_ShootingData)

## [1] "list"

is.data.frame(RShah\_ShootingData)

## [1] TRUE

is\_tibble(RShah\_ShootingData)

## [1] TRUE

as\_tibble(RShah\_ShootingData)

## # A tibble: 8,720 × 19  
 ## id date threat\_type flee\_status armed\_with city county state  
 ## <dbl> <date> <chr> <chr> <chr> <chr> <chr> <chr>  
 ## 1 3 2015-01-02 point not gun Shelton Mason WA   
 ## 2 4 2015-01-02 point not gun Aloha Washi… OR   
 ## 3 5 2015-01-03 move not unarmed Wichita Sedgw… KS   
 ## 4 8 2015-01-04 point not replica San Francis… San F… CA   
 ## 5 9 2015-01-04 point not other Evans Weld CO   
 ## 6 11 2015-01-04 attack not gun Guthrie Logan OK   
 ## 7 13 2015-01-05 shoot car gun Chandler Maric… AZ   
 ## 8 15 2015-01-06 point not gun Assaria Saline KS   
 ## 9 16 2015-01-06 accident not unarmed Burlington Des M… IA   
 ## 10 17 2015-01-06 point not replica Knoxville Alleg… PA   
 ## # ℹ 8,710 more rows  
 ## # ℹ 11 more variables: latitude <dbl>, longitude <dbl>,  
 ## # location\_precision <chr>, name <chr>, age <dbl>, gender <chr>, race <chr>,  
 ## # race\_source <chr>, was\_mental\_illness\_related <lgl>, body\_camera <lgl>,  
 ## # agency\_ids <chr>

*#view the data* str(RShah\_ShootingData)

## spc\_tbl\_ [8,720 × 19] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
 ## $ id : num [1:8720] 3 4 5 8 9 11 13 15 16 17 ...  
 ## $ date : Date[1:8720], format: "2015-01-02" "2015-01-02" ...  
 ## $ threat\_type : chr [1:8720] "point" "point" "move" "point" ...  
 ## $ flee\_status : chr [1:8720] "not" "not" "not" "not" ...  
 ## $ armed\_with : chr [1:8720] "gun" "gun" "unarmed" "replica" ...  
 ## $ city : chr [1:8720] "Shelton" "Aloha" "Wichita" "San Francisco" ...  
 ## $ county : chr [1:8720] "Mason" "Washington" "Sedgwick" "San Francisco" ...  
 ## $ state : chr [1:8720] "WA" "OR" "KS" "CA" ...  
 ## $ latitude : num [1:8720] 47.2 45.5 37.7 37.8 40.4 ...  
 ## $ longitude : num [1:8720] -123.1 -122.9 -97.3 -122.4 -104.7 ...  
 ## $ location\_precision : chr [1:8720] "not\_available" "not\_available" "not\_available" "not\_available" ...  
 ## $ name : chr [1:8720] "Tim Elliot" "Lewis Lee Lembke" "John Paul Quintero" "Matthew Hoffman" ...  
 ## $ age : num [1:8720] 53 47 23 32 39 18 22 35 34 47 ...  
 ## $ gender : chr [1:8720] "male" "male" "male" "male" ...  
 ## $ race : chr [1:8720] "A" "W" "H" "W" ...  
 ## $ race\_source : chr [1:8720] "not\_available" "not\_available" "not\_available" "not\_available" ...  
 ## $ was\_mental\_illness\_related: logi [1:8720] TRUE FALSE FALSE TRUE FALSE FALSE ...  
 ## $ body\_camera : logi [1:8720] FALSE FALSE FALSE FALSE FALSE FALSE ...  
 ## $ agency\_ids : chr [1:8720] "73" "70" "238" "196" ...  
 ## - attr(\*, "spec")=  
 ## .. cols(  
 ## .. id = col\_double(),  
 ## .. date = col\_date(format = ""),  
 ## .. threat\_type = col\_character(),  
 ## .. flee\_status = col\_character(),  
 ## .. armed\_with = col\_character(),  
 ## .. city = col\_character(),  
 ## .. county = col\_character(),  
 ## .. state = col\_character(),  
 ## .. latitude = col\_double(),  
 ## .. longitude = col\_double(),  
 ## .. location\_precision = col\_character(),  
 ## .. name = col\_character(),  
 ## .. age = col\_double(),  
 ## .. gender = col\_character(),  
 ## .. race = col\_character(),  
 ## .. race\_source = col\_character(),  
 ## .. was\_mental\_illness\_related = col\_logical(),  
 ## .. body\_camera = col\_logical(),  
 ## .. agency\_ids = col\_character()  
 ## .. )  
 ## - attr(\*, "problems")=<externalptr>

*#sapply(RShah\_ShootingData,typeof)*

*# #examining missing data* *# #get a sense of the extend of missing data (lots of values for race is missing)* *# #815 out of 4478 rows has some missing value* filter(RShah\_ShootingData, !complete.cases(RShah\_ShootingData))

## # A tibble: 5,872 × 19  
 ## id date threat\_type flee\_status armed\_with city county state  
 ## <dbl> <date> <chr> <chr> <chr> <chr> <chr> <chr>  
 ## 1 29 2015-01-08 undetermined not <NA> Huntley Yello… MT   
 ## 2 49 2015-01-14 attack not <NA> St. Paul Ramsey MN   
 ## 3 50 2015-01-14 attack not <NA> Franklin C… Frank… MO   
 ## 4 85 2015-01-20 threat car <NA> Scottsdale Maric… AZ   
 ## 5 101 2015-01-23 threat not <NA> Fort Laude… Browa… FL   
 ## 6 110 2015-01-25 point not gun Winslow Camden NJ   
 ## 7 348 2015-01-27 threat not <NA> Phoenix Maric… AZ   
 ## 8 145 2015-02-13 attack car <NA> San Bernar… San B… CA   
 ## 9 162 2015-02-17 threat not <NA> Del Rio Val V… TX   
 ## 10 584 2015-02-20 point car gun Houston Harris TX   
 ## # ℹ 5,862 more rows  
 ## # ℹ 11 more variables: latitude <dbl>, longitude <dbl>,  
 ## # location\_precision <chr>, name <chr>, age <dbl>, gender <chr>, race <chr>,  
 ## # race\_source <chr>, was\_mental\_illness\_related <lgl>, body\_camera <lgl>,  
 ## # agency\_ids <chr>

*#  
 # #apply() returns a vector or array or list of values obtained by applying a function to margins of an array or matrix.* apply(RShah\_ShootingData, 2, function(x) sum(is.na(x))) %>% sort(decreasing=TRUE) *#number of NAs in each column ('2')*

## county race  
 ## 4856 1389  
 ## race\_source flee\_status  
 ## 1367 1175  
 ## latitude longitude  
 ## 980 980  
 ## location\_precision age  
 ## 980 600  
 ## name armed\_with  
 ## 572 210  
 ## city gender  
 ## 53 47  
 ## threat\_type agency\_ids  
 ## 39 2  
 ## id date  
 ## 0 0  
 ## state was\_mental\_illness\_related  
 ## 0 0  
 ## body\_camera  
 ## 0

*# #What shall we do with NAs? Not clear at this time, so keep NAs as is.* unique(RShah\_ShootingData$age)

## [1] 53 47 23 32 39 18 22 35 34 25 31 41 30 37 28 42 36 49 71 33 29 43 24 75 68  
 ## [26] 27 48 21 67 19 54 17 56 61 45 26 40 59 38 51 74 57 46 16 50 20 77 NA 58 64  
 ## [51] 52 63 44 60 66 83 72 76 62 55 69 86 15 65 6 12 70 80 14 82 13 73 91 79 78  
 ## [76] 84 81 89 88 8 92 2

*#percentage of missing data* apply(RShah\_ShootingData, 2, function(x) percent(mean(is.na(x)))) %>% sort(decreasing=TRUE)

## name age  
 ## "7%" "7%"  
 ## county armed\_with  
 ## "56%" "2%"  
 ## race race\_source  
 ## "16%" "16%"  
 ## flee\_status latitude  
 ## "13%" "11%"  
 ## longitude location\_precision  
 ## "11%" "11%"  
 ## city gender  
 ## "1%" "1%"  
 ## id date  
 ## "0%" "0%"  
 ## threat\_type state  
 ## "0%" "0%"  
 ## was\_mental\_illness\_related body\_camera  
 ## "0%" "0%"  
 ## agency\_ids  
 ## "0%"

*#Shooting data has all the dimension data, but not a measure.  
 #Let's add a measure "count" to shooting* RShah\_ShootingData$count <-rep(1, nrow(RShah\_ShootingData))  
 library(dplyr)  
  
 *#method1: binning with fixed equal-width cut* *#<25, 25-50, 50-75, >75* *#brains are not fully developed before 25.* RShah\_ShootingData$age <- as.numeric(as.character(RShah\_ShootingData$age))  
 (RShah\_AgeGroup\_dim <- cut(RShah\_ShootingData$age,  
 breaks = c(0, 25, 50, 75, 100),  
 labels=c('young', 'grown', 'mature', 'old')))

*# Create group armedTyPe for weapons* count(RShah\_ShootingData, armed\_with, sort=TRUE)

## # A tibble: 22 × 2  
 ## armed\_with n  
 ## <chr> <int>  
 ## 1 gun 5055  
 ## 2 knife 1466  
 ## 3 unarmed 514  
 ## 4 undetermined 339  
 ## 5 vehicle 306  
 ## 6 replica 288  
 ## 7 blunt\_object 213  
 ## 8 <NA> 210  
 ## 9 unknown 136  
 ## 10 other 88  
 ## # ℹ 12 more rows

RShah\_Armedgroup\_dim <- function (string){  
 if(is.na(string)) return ("NA")  
 if(string == "unarmed") return ("unarmed")  
 else if (string == "undetermined") return ("undetermined")  
 else if (string == "vehicle") return ("vehicle")  
 else if (string %in% c("gun", "toy weapon", "gun and knife",  
 "gun and car", "BB gun",  
 "guns and explosives", "gun and vehicle",  
 "hatchet and gun", "gun and sword",  
 "machete and gun", "vehicle and gun",  
 "pellet gun")) return ("gun")  
 else if (string %in% c("knife", "ax", "sword", "box cutter",  
 "hatchet", "sharp object",  
 "scissors", "meat cleaver", "pick-axe",  
 "straight edge razor",  
 "pitchfork", "chainsaw", "samurai sword",  
 "spear")) return ("sharpObject")  
 else return ("other")  
 }  
 age <- as.numeric(RShah\_ShootingData$age[!is.na(RShah\_ShootingData$age)])  
 RShah\_ShootingData$armed\_group <- sapply(RShah\_ShootingData$armed\_with, RShah\_Armedgroup\_dim)  
 *#RShah\_ShootingData$armed\_group*   
  
   
  
  
  
  
  
 *#apply group on each element of a vector* *#RShah\_ShootingData$armed\_with <- sapply(RShah\_ShootingData$armed\_with, RShah\_Armedgroup\_dim)  
  
 #RShah\_ShootingData$armed\_with <- sapply(RShah\_ShootingData$armed\_with, RShah\_Armedgroup\_dim\_n)  
 #RShah\_ShootingData$armed\_with*

*# create age category to consolidate different age groups* *# Identify different characteristics based on age, weapons, race and gender* year <- as.numeric(format(RShah\_ShootingData$date,'%Y'))  
 RShah\_AgeRange\_dim <- RShah\_ShootingData %>%  
 mutate(ageCategory = case\_when(between(age, 0, 10) ~ "0 - 10 Years",  
 between(age, 11, 20) ~ "11 - 20 Years",  
 between(age, 21, 30) ~ "21 - 30 Years",  
 between(age, 31, 40) ~ "31 - 40 Years",  
 between(age, 41, 50) ~ "41 - 50 Years",  
 between(age, 51, 60) ~ "51 - 60 Years",  
 between(age, 61, 70) ~ "61 - 70 Years",  
 between(age, 71, 80) ~ "71 - 80 Years",  
 between(age, 81, 91) ~ "81 - 91 Years")) %>%  
 select(age,race,gender,armed\_group,ageCategory)  
 *#RShah\_AgeRange\_dim*

*#Append result* *#RShah\_appendresult$AgeGroup <- RShah\_AgeGroup\_dim* *#RShah\_appendresulta$armed\_group <- sapply(RShah\_appendresult$armed\_with,RShah\_Armedgroup\_dim)* AgeRange <- RShah\_AgeRange\_dim$ageCategory  
 AgeGroup <- RShah\_AgeGroup\_dim  
 ArmedGroup <- sapply(RShah\_ShootingData$armed\_with,RShah\_Armedgroup\_dim)  
 year <- as.integer(format(RShah\_ShootingData$date,'%Y'))  
 df <- cbind(RShah\_ShootingData,AgeRange)  
 df1 <- cbind(df,AgeGroup)  
 str(df1)

## 'data.frame': 8720 obs. of 23 variables:  
 ## $ id : num 3 4 5 8 9 11 13 15 16 17 ...  
 ## $ date : Date, format: "2015-01-02" "2015-01-02" ...  
 ## $ threat\_type : chr "point" "point" "move" "point" ...  
 ## $ flee\_status : chr "not" "not" "not" "not" ...  
 ## $ armed\_with : chr "gun" "gun" "unarmed" "replica" ...  
 ## $ city : chr "Shelton" "Aloha" "Wichita" "San Francisco" ...  
 ## $ county : chr "Mason" "Washington" "Sedgwick" "San Francisco" ...  
 ## $ state : chr "WA" "OR" "KS" "CA" ...  
 ## $ latitude : num 47.2 45.5 37.7 37.8 40.4 ...  
 ## $ longitude : num -123.1 -122.9 -97.3 -122.4 -104.7 ...  
 ## $ location\_precision : chr "not\_available" "not\_available" "not\_available" "not\_available" ...  
 ## $ name : chr "Tim Elliot" "Lewis Lee Lembke" "John Paul Quintero" "Matthew Hoffman" ...  
 ## $ age : num 53 47 23 32 39 18 22 35 34 47 ...  
 ## $ gender : chr "male" "male" "male" "male" ...  
 ## $ race : chr "A" "W" "H" "W" ...  
 ## $ race\_source : chr "not\_available" "not\_available" "not\_available" "not\_available" ...  
 ## $ was\_mental\_illness\_related: logi TRUE FALSE FALSE TRUE FALSE FALSE ...  
 ## $ body\_camera : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
 ## $ agency\_ids : chr "73" "70" "238" "196" ...  
 ## $ count : num 1 1 1 1 1 1 1 1 1 1 ...  
 ## $ armed\_group : chr "gun" "gun" "unarmed" "other" ...  
 ## $ AgeRange : chr "51 - 60 Years" "41 - 50 Years" "21 - 30 Years" "31 - 40 Years" ...  
 ## $ AgeGroup : Factor w/ 4 levels "young","grown",..: 3 2 1 2 2 1 1 2 2 2 ...

year <- as.integer(format(RShah\_ShootingData$date,'%Y'))  
 RShah\_df <- cbind(df1,year)  
 *#RShah\_appendresult<-cbind(df2,ArmedGroup)* *#write.table(RShah\_appendresult, "fatal-police-shootings-data-new1.csv", append=F,sep=",")* *#df2 <- cbind(df1,year)* RShah\_appendresult<-RShah\_df  
  
  
  
  
  
 *#write.table(RShah\_appendresult, "fatal-police-shootings-data-new2.csv", append=F,sep=",")* *#write.csv(bind,file = "fatal-police-shootings-data.csv", append = TRUE)*

*#table field name and data types*

*#str(RShah\_ShootingData)* str(RShah\_df)

## 'data.frame': 8720 obs. of 24 variables:  
 ## $ id : num 3 4 5 8 9 11 13 15 16 17 ...  
 ## $ date : Date, format: "2015-01-02" "2015-01-02" ...  
 ## $ threat\_type : chr "point" "point" "move" "point" ...  
 ## $ flee\_status : chr "not" "not" "not" "not" ...  
 ## $ armed\_with : chr "gun" "gun" "unarmed" "replica" ...  
 ## $ city : chr "Shelton" "Aloha" "Wichita" "San Francisco" ...  
 ## $ county : chr "Mason" "Washington" "Sedgwick" "San Francisco" ...  
 ## $ state : chr "WA" "OR" "KS" "CA" ...  
 ## $ latitude : num 47.2 45.5 37.7 37.8 40.4 ...  
 ## $ longitude : num -123.1 -122.9 -97.3 -122.4 -104.7 ...  
 ## $ location\_precision : chr "not\_available" "not\_available" "not\_available" "not\_available" ...  
 ## $ name : chr "Tim Elliot" "Lewis Lee Lembke" "John Paul Quintero" "Matthew Hoffman" ...  
 ## $ age : num 53 47 23 32 39 18 22 35 34 47 ...  
 ## $ gender : chr "male" "male" "male" "male" ...  
 ## $ race : chr "A" "W" "H" "W" ...  
 ## $ race\_source : chr "not\_available" "not\_available" "not\_available" "not\_available" ...  
 ## $ was\_mental\_illness\_related: logi TRUE FALSE FALSE TRUE FALSE FALSE ...  
 ## $ body\_camera : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
 ## $ agency\_ids : chr "73" "70" "238" "196" ...  
 ## $ count : num 1 1 1 1 1 1 1 1 1 1 ...  
 ## $ armed\_group : chr "gun" "gun" "unarmed" "other" ...  
 ## $ AgeRange : chr "51 - 60 Years" "41 - 50 Years" "21 - 30 Years" "31 - 40 Years" ...  
 ## $ AgeGroup : Factor w/ 4 levels "young","grown",..: 3 2 1 2 2 1 1 2 2 2 ...  
 ## $ year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...

getwd()

## [1] "C:/Users/yashs/OneDrive/Desktop/R/ArizonaPro"

*#RShah\_appendresult<-read\_csv("fatal-police-shootings-data-new1.csv", col\_names = TRUE, na="")* save(RShah\_ShootingData, file="Example PartII.Rdata")  
 *#examining missing data* filter(RShah\_ShootingData,!complete.cases(RShah\_ShootingData))

## # A tibble: 5,872 × 21  
 ## id date threat\_type flee\_status armed\_with city county state  
 ## <dbl> <date> <chr> <chr> <chr> <chr> <chr> <chr>  
 ## 1 29 2015-01-08 undetermined not <NA> Huntley Yello… MT   
 ## 2 49 2015-01-14 attack not <NA> St. Paul Ramsey MN   
 ## 3 50 2015-01-14 attack not <NA> Franklin C… Frank… MO   
 ## 4 85 2015-01-20 threat car <NA> Scottsdale Maric… AZ   
 ## 5 101 2015-01-23 threat not <NA> Fort Laude… Browa… FL   
 ## 6 110 2015-01-25 point not gun Winslow Camden NJ   
 ## 7 348 2015-01-27 threat not <NA> Phoenix Maric… AZ   
 ## 8 145 2015-02-13 attack car <NA> San Bernar… San B… CA   
 ## 9 162 2015-02-17 threat not <NA> Del Rio Val V… TX   
 ## 10 584 2015-02-20 point car gun Houston Harris TX   
 ## # ℹ 5,862 more rows  
 ## # ℹ 13 more variables: latitude <dbl>, longitude <dbl>,  
 ## # location\_precision <chr>, name <chr>, age <dbl>, gender <chr>, race <chr>,  
 ## # race\_source <chr>, was\_mental\_illness\_related <lgl>, body\_camera <lgl>,  
 ## # agency\_ids <chr>, count <dbl>, armed\_group <chr>

*#apply() returns a vector or array or list of values obtained by applying a function to margins of an array or matrix.* apply(RShah\_ShootingData, 2, function(x) sum(is.na(x))) %>% sort(decreasing=TRUE)

## county race  
 ## 4856 1389  
 ## race\_source flee\_status  
 ## 1367 1175  
 ## latitude longitude  
 ## 980 980  
 ## location\_precision age  
 ## 980 600  
 ## name armed\_with  
 ## 572 210  
 ## city gender  
 ## 53 47  
 ## threat\_type agency\_ids  
 ## 39 2  
 ## id date  
 ## 0 0  
 ## state was\_mental\_illness\_related  
 ## 0 0  
 ## body\_camera count  
 ## 0 0  
 ## armed\_group  
 ## 0

*#number of NAs in each column ('2')* *# #What shall we do with NAs? Not clear at this time, so keep NAs as is.* apply(RShah\_ShootingData,2,function(x) percent(mean(is.na(x)))) %>% sort(decreasing=TRUE)

## name age  
 ## "7%" "7%"  
 ## county armed\_with  
 ## "56%" "2%"  
 ## race race\_source  
 ## "16%" "16%"  
 ## flee\_status latitude  
 ## "13%" "11%"  
 ## longitude location\_precision  
 ## "11%" "11%"  
 ## city gender  
 ## "1%" "1%"  
 ## id date  
 ## "0%" "0%"  
 ## threat\_type state  
 ## "0%" "0%"  
 ## was\_mental\_illness\_related body\_camera  
 ## "0%" "0%"  
 ## agency\_ids count  
 ## "0%" "0%"  
 ## armed\_group  
 ## "0%"

*#Next, convert shooting to an multi-dimensional array.  
 #This array is the base cuboid with measure=count.  
  
 #tapply() applys a function to each (non-empty) group of values  
 #given by a unique combination of the levels of factors.* *#To make is easier to understand, we will select 4 factors as the dimensions for the cube* str(RShah\_df)

## 'data.frame': 8720 obs. of 24 variables:  
 ## $ id : num 3 4 5 8 9 11 13 15 16 17 ...  
 ## $ date : Date, format: "2015-01-02" "2015-01-02" ...  
 ## $ threat\_type : chr "point" "point" "move" "point" ...  
 ## $ flee\_status : chr "not" "not" "not" "not" ...  
 ## $ armed\_with : chr "gun" "gun" "unarmed" "replica" ...  
 ## $ city : chr "Shelton" "Aloha" "Wichita" "San Francisco" ...  
 ## $ county : chr "Mason" "Washington" "Sedgwick" "San Francisco" ...  
 ## $ state : chr "WA" "OR" "KS" "CA" ...  
 ## $ latitude : num 47.2 45.5 37.7 37.8 40.4 ...  
 ## $ longitude : num -123.1 -122.9 -97.3 -122.4 -104.7 ...  
 ## $ location\_precision : chr "not\_available" "not\_available" "not\_available" "not\_available" ...  
 ## $ name : chr "Tim Elliot" "Lewis Lee Lembke" "John Paul Quintero" "Matthew Hoffman" ...  
 ## $ age : num 53 47 23 32 39 18 22 35 34 47 ...  
 ## $ gender : chr "male" "male" "male" "male" ...  
 ## $ race : chr "A" "W" "H" "W" ...  
 ## $ race\_source : chr "not\_available" "not\_available" "not\_available" "not\_available" ...  
 ## $ was\_mental\_illness\_related: logi TRUE FALSE FALSE TRUE FALSE FALSE ...  
 ## $ body\_camera : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
 ## $ agency\_ids : chr "73" "70" "238" "196" ...  
 ## $ count : num 1 1 1 1 1 1 1 1 1 1 ...  
 ## $ armed\_group : chr "gun" "gun" "unarmed" "other" ...  
 ## $ AgeRange : chr "51 - 60 Years" "41 - 50 Years" "21 - 30 Years" "31 - 40 Years" ...  
 ## $ AgeGroup : Factor w/ 4 levels "young","grown",..: 3 2 1 2 2 1 1 2 2 2 ...  
 ## $ year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...

**#Additional work**

shooting\_base\_cubiod <-  
 tapply(RShah\_df$count,  
 RShah\_df[,c("race", "state", "AgeRange","AgeGroup","armed\_group","year")],  
 FUN=function(x){return(sum(x, na.rm=TRUE))})  
  
 *#set all NAs to 0* shooting\_base\_cubiod[is.na(shooting\_base\_cubiod)] <- 0  
 shooting\_base\_cubiod