Dataset Analysis Report

Aleksandra Viktorova

March 1, 2020

The loading libraries

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(cluster)
library(factoextra)
library(maps)
```

Some information about the dataset

The “Homicide Reports” dataset resembles data about murders from 1980 to 2014 in US.
Number of Instances: 638454 murders.
Number of Attributes: 20.
Read more in the file “Description of dataset”.

Based on the available information, the following questions were formulated:

1. What is the average number of victims among men and women in each state of the United States?
2. Does the status of the solved crime depend on the type of agency?
3. Does the choice of weapons depend on the race of the perpetrator?
4. Does the age of the victim and the age of the perpetrator differ significantly?
5. Are there similar US states based on the values of variables such as average of “perpetrator age”, average of “victim age” and average number of crimes?

Load the data set into R from the file.

```
DataSet <- read.csv(file=file.choose(), header=TRUE, sep=";")
```

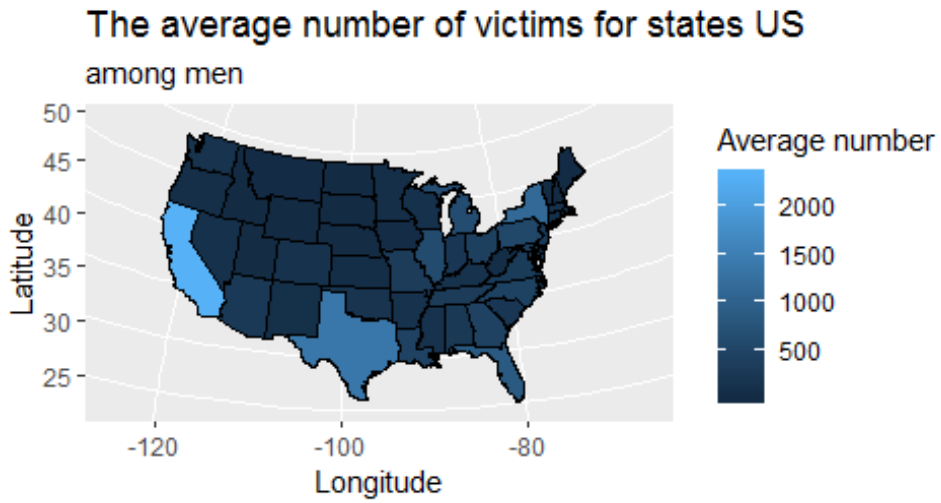
The answer to the first question

To answer this question, it is necessary to select data for men and women by state separately. Since the data set was given for the period from 1980 to 2014, we can calculate the total number of male and female victims for each year by state, and then take the average value for the results.

```
# Select and Group the data by Sate and Year,summarize the number of victims
among women and
# men and then group a data by state. Find the average number of victims among
women and men.
St_F <-as_tibble(select(DataSet[DataSet$Victim.Sex == "Female",], State, Year,
Victim.Sex)) %>%

          group_by(State, Year) %>% summarise(mCo_F = length(Victim.Sex)) %>%
                  group_by(State) %>% summarise(PFemale = mean(mCo_F))

St_M <- as_tibble(select(DataSet[DataSet$Victim.Sex == "Male",], State, Year,
Victim.Sex)) %>%
           group_by(State, Year) %>% summarise(mCo_M = length(Victim.Sex)) %>%
                       group_by(State) %>% summarise(PMale = mean(mCo_M))

# Create a shared data table
crimes <- data.frame(state=tolower(St_F$State), PFemale=St_F$PFemale, PMale =
St_M$PMale)
```

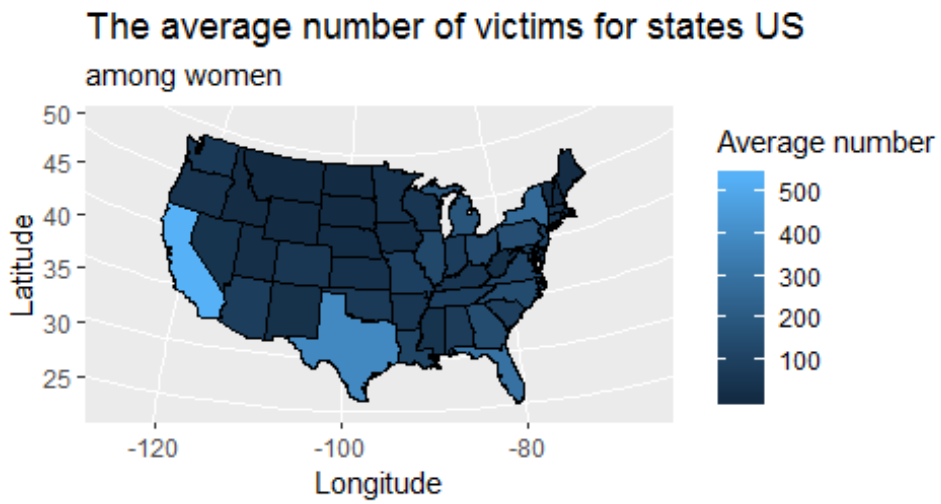
For clarity, display the results on a map of the United States.

```
# Turn data from the maps package in to a data frame
states_map <- map_data("state")

# Merge the data sets together and sort by group, then order
crime_map <- merge(states_map, crimes, by.x="region", by.y="state") %>%
arrange(group, order)

# Build two geo-maps for data for men and women
ggplot(data = crime_map, aes(x = long, y = lat, group = group, fill = PMale)) +
  geom_polygon(colour = "black") +
  coord_map("polyconic") +
  labs(x = "Longitude", y = "Latitude", fill = "Average number", title = "The
average number of victims for states US", subtitle = "among men")
```




```
ggplot(data = crime_map, aes(x = long, y = lat, group = group, fill = PFemale))
+
```

```
  geom_polygon(colour = "black") +
  coord_map("polyconic")+
  labs(x = "Longitude", y = "Latitude", fill = "Average number", title="The
average number of victims for states US", subtitle = "among women")
```



We can see that in both cases, Texas and California have the highest average number of victims among men and women.

The answer to the second question

To answer the question posed, it is necessary to select and group the data by the variables "Crime Solved" and "Agency Type", and then apply the Chi-square test.

```
# Select and Group the data by Agency.Type and Crime.Solved,summarize the number
of victims
TypeAgency <-as_tibble(select(DataSet, Agency.Type, Crime.Solved, Victim.Sex))
%>%
             group_by(Agency.Type, Crime.Solved) %>% summarise(mVi =
length(Victim.Sex))

#Create a data table
AgencyCrime <- cbind(TypeAgency[TypeAgency$Crime.Solved == "No",3],
                     TypeAgency[TypeAgency$Crime.Solved == "Yes",3])
colnames(AgencyCrime) <- unique(TypeAgency$Crime.Solved)
rownames(AgencyCrime) <- unique(TypeAgency$Agency.Type)
AgencyCrime

##                       No    Yes
## County Police       7533  15160
## Municipal Police 157017 336009
## Regional Police       49    186
## Sheriff            22328  82994
## Special Police       831   2058
## State Police        2520  11715
## Tribal Police          4     50
```

```
#Chi-Square Test
chisq.test(AgencyCrime)

##
##  Pearson's Chi-squared test
##
## data:  AgencyCrime
## X-squared = 5855.8, df = 6, p-value < 2.2e-16
```

We can see that p-value is < 0.05 and this means that the status of the solved crime depends on the type of agency.

The answer to the third question

To answer this question, it is necessary to select all the known values of the parameters of the perpetrator race and weapon, and then group the data.

```
# Select and Group the data by Perpetrator.Race and Weapon,summarize the number
of perpetrators
WePerpetrator <- as_tibble(select(DataSet[DataSet$Perpetrator.Race != "Unknown"
                                          & DataSet$Weapon != "Unknown",],
                                  Perpetrator.Race, Perpetrator.Sex, Weapon))
%>%
                  group_by(Perpetrator.Race, Weapon) %>%
                        summarise(mPer = length(Perpetrator.Sex))

# Look the result
summary(WePerpetrator)

##                         Perpetrator.Race          Weapon        mPer
##  Asian/Pacific Islander       :15      Blunt Object: 4   Min.   :     2
##  Black                        :15      Drowning    : 4   1st Qu.:    65
##  Native American/Alaska Native:14      Drugs       : 4   Median :   298
##  Unknown                      : 0      Fall        : 4   Mean   :  7192
##  White                        :15      Fire        : 4   3rd Qu.:  2132
##                                        Firearm     : 4   Max.   :116477
##                                        (Other)     :35
```

We can see that “Native Americans/Alaska Natives” has only 14 meanings compared to others, so this type of race will not be considered in a future analysis.Create the data table, and then apply the Chi-square test.

```
# Create a data table for the three races of perpetrators
PerWeapon <- cbind(WePerpetrator[WePerpetrator$Perpetrator.Race ==
"Asian/Pacific Islander",3],
                   WePerpetrator[WePerpetrator$Perpetrator.Race ==
"Black",3],
                   WePerpetrator[WePerpetrator$Perpetrator.Race ==
"White",3])
colnames(PerWeapon) <- c("Asian", "Black", "White")
rownames(PerWeapon) <- unique(WePerpetrator$Weapon)
PerWeapon
```

```
##                Asian  Black White
## Blunt Object     676  20400 29170
## Drowning          26    276   660
## Drugs             15    167  1190
## Explosives         9     75   298
## Fall               5     73    85
## Fire              80   1679  2281
## Firearm          258  14296  9336
## Gun                9    710   543
## Handgun         2976 116477 93662
## Knife           1136  35395 36322
## Poison            10     63   294
## Rifle            242   6135 13023
## Shotgun          219   9512 15890
## Strangulation     95   1466  2725
## Suffocation       67    855  1982
```

```
# Chi-Square Test for three races of perpetrators
chisq.test(PerWeapon)
```

```
## Warning in chisq.test(PerWeapon): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  PerWeapon
## X-squared = 11483, df = 28, p-value < 2.2e-16
```

We have a warning message: Chi-squared approximation may be incorrect.This is because the category “Asian” has small meanings. In this case, you can use “simulate p-value” or try to remove this criterion from consideration.

```
# Chi-Square Test for three races of perpetrators with simulate p-value
chisq.test(PerWeapon, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  PerWeapon
## X-squared = 11483, df = NA, p-value = 0.0004998
```

```
# Chi-Square Test for two races of perpetrators without Asians
chisq.test(select(PerWeapon, -Asian))
```

```
##
##  Pearson's Chi-squared test
##
## data:  select(PerWeapon, -Asian)
## X-squared = 11312, df = 14, p-value < 2.2e-16
```

In both cases, p-value < 0.05, which means that the choice of weapons depends on the race of the perpetrator.

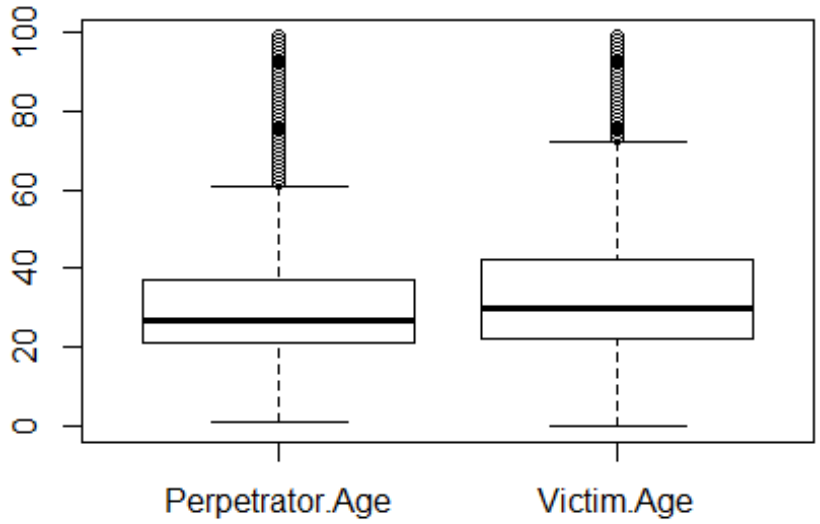
The answer to the fourth question

To answer this question, it is necessary to select data on the age of victims and perpetrator.

```
# Select of the necessary data
AgeTest <- as_tibble(select(DataSet[DataSet$Perpetrator.Age != 0 &
DataSet$Victim.Age != 998,],
                            Perpetrator.Age, Victim.Age))
```

Imagine groups of numerical data through quartiles.

```
boxplot(AgeTest)
```



From the graph obtained, it is clear that the medians of both indicators have different values, but for complete reliability it is necessary to conduct a t-test.

```
# T-test for two variables
t.test(AgeTest$Perpetrator.Age, AgeTest$Victim.Age)

##
##  Welch Two Sample t-test
##
## data:  AgeTest$Perpetrator.Age and AgeTest$Victim.Age
## t = -73.946, df = 769930, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.545234 -2.413793
## sample estimates:
## mean of x mean of y
##  30.73805  33.21756
```

Thus, the age of the victim and the age of the perpetrator differ significantly.

The answer to the fifth question

First, we select the necessary data for analysis.

```
# Select of the necessary data and their group by parameters
Age_Victim <- as_tibble(select(DataSet[DataSet$Victim.Age != 998,], State,
Victim.Age)) %>%
                  group_by(State) %>%  summarise(vAv =
round(mean(Victim.Age),2))

Age_Perpet <- as_tibble(select(DataSet[DataSet$Perpetrator.Age != 0,], State,
Perpetrator.Age))             %>%  group_by(State) %>% summarise(pAv =
round(mean(Perpetrator.Age), 2))

Crime_State <-as_tibble(select(DataSet, State, Year, Crime.Solved)) %>%
                group_by(State, Year) %>% summarise(Crn = length(Crime.Solved))
%>%
                      group_by(State) %>% summarise(cAv = round(mean(Crn), 2))
```

For cluster analysis, we need variables that do not strongly collate with each other, so we calculate the correlation matrix for the data obtained.

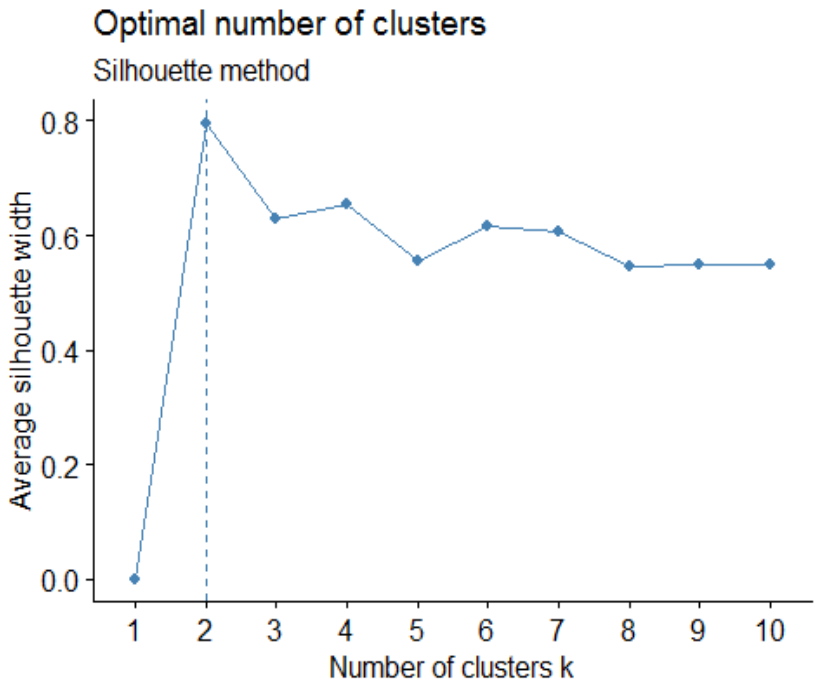
```
# Check the correlation between the variables
cor(data.frame(Age_Victim$vAv, Age_Perpet$pAv, Crime_State$cAv))

##                 Age_Victim.vAv Age_Perpet.pAv Crime_State.cAv
## Age_Victim.vAv       1.0000000      0.7878831      -0.2144571
## Age_Perpet.pAv       0.7878831      1.0000000      -0.3494318
## Crime_State.cAv     -0.2144571     -0.3494318       1.0000000
```

We can see that variables “Age_Victim.vAv” and “Age_Perpet.pAv” have the highest correlation value, therefore we cannot use both of these variables together. We will try to find similar states in the US by the “Age_Victim.vAv” and “Crime_State.cAv” variables since they have the smallest correlation value.

```
# Create a data set
Clust_State <- data.frame(Age_Victim$vAv, Crime_State$cAv)
row.names(Clust_State) <- unique(Crime_State$State)

# Compute all the pairwise distances between observations in the data set
Clust_State.dist <- daisy(Clust_State, metric="euclidean")

# Hierarchical cluster analysis on a set of dissimilarities
Clust_State.h <- hclust(Clust_State.dist, method="ward.D")
```

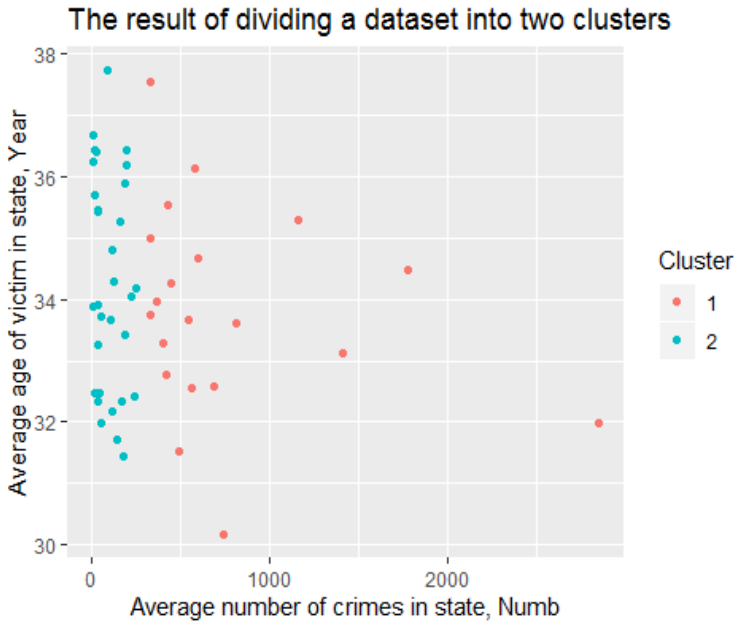
To find the optimal number of clusters, we use the following function:

```
# Find the optimal number of clusters
fviz_nbclust(Clust_State, kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette method")
```



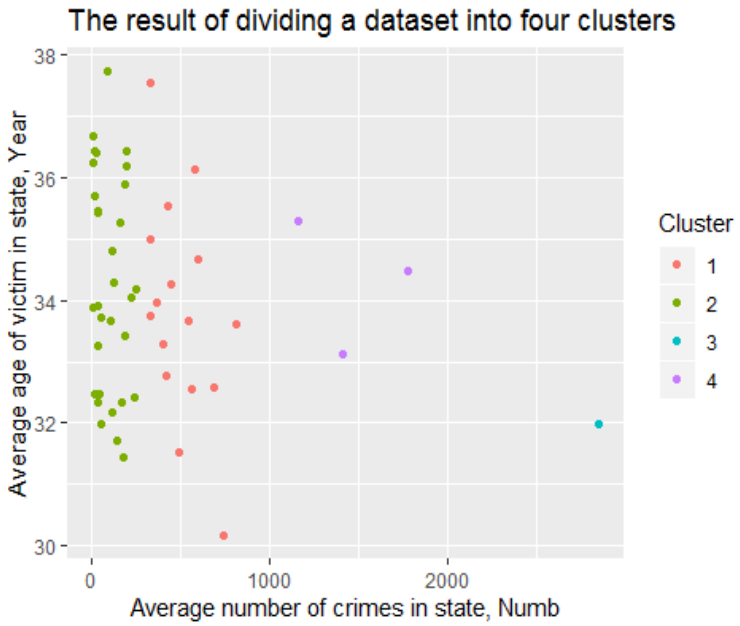
We can see that optimal number of clusters is 2. Divide the available data into two clusters.

```
# The dividing a dataset into two clusters
groups_cl <- cutree(Clust_State.h, k = 2)
Clust_State$groups_cl <- factor(groups_cl)

#Show the result on a graph
ggplot(data = Clust_State, aes(x = Crime_State.cAv , y = Age_Victim.vAv)) +
  geom_point(aes(color = groups_cl)) +
  labs(x = "Average number of crimes in state, Numb", y = "Average age of victim
in state, Year", color ="Cluster", title = "The result of dividing a dataset
into two clusters")
```



We can see that 4 points of 1 cluster have significant difference: 3 points between 1000 and 2000 of average number of crimes in state and 1 point after a value of 2000.Try to divide dataset into 4 clusters.

```
# The dividing a dataset into four clusters
groups_cl <- cutree(Clust_State.h, k = 4)
Clust_State$groups_cl <- factor(groups_cl)

#Show the result on a graph
ggplot(data = Clust_State, aes(x = Crime_State.cAv , y = Age_Victim.vAv)) +
  geom_point(aes(color = groups_cl))  +
  labs(x = "Average number of crimes in state, Numb", y = "Average age of victim
in state, Year", color ="Cluster", title = "The result of dividing a dataset
into four clusters")
```



We estimate the resulting clustering using the Kruskal-Wallis test.

```
#Kruskal test for two variables
kruskal.test(Clust_State$Age_Victim.vAv ~ Clust_State$groups_cl)

##
##  Kruskal-Wallis rank sum test
##
## data:  Clust_State$Age_Victim.vAv by Clust_State$groups_cl
## Kruskal-Wallis chi-squared = 2.5772, df = 3, p-value = 0.4615

kruskal.test(Clust_State$Crime_State.cAv ~ Clust_State$groups_cl)

##
##  Kruskal-Wallis rank sum test
##
## data:  Clust_State$Crime_State.cAv by Clust_State$groups_cl
## Kruskal-Wallis chi-squared = 37.231, df = 3, p-value = 4.112e-08
```

We can seee that in case of “Crime_State.cAv” p-value close to zero and this means that the distribution of the average number of crimes in each cluster is different. However, in case of “Age_Victim.vAv” the data in the samples are taken from one distribution. Thus, clustering based on the two variables taken was not the best. We construct a dendogram and select 4 obtained clusters on it.

```
#Cluster dendrogram
plot(Clust_State.h, labels = abbreviate(rownames(Clust_State), 1, method
="left.kept"),
     hang = 0.3, xlab = "States of USA", main = "Cluster Dendrogram")

#Select 4 clusters
rect.hclust(Clust_State.h, k = 4)
```