Exploratory Analysis on Death In U.S.

Anushree Shivarudrappa

June 1, 2016

1. Introduction

I am using Death in United States dataset. This mortality dataset is a record of every death in the country for the year 2014, which includes detailed information about causes of death and the demographic background of the deceased. These data provide information on mortality patterns among U.S. residents by such variables as sex, race and ethnicity, and cause of death. Information on mortality patterns is key to understanding changes in the health and well-being of the U.S. population.

2. Pre-Processing

```
library(data.table)
library(ggplot2)
library(dplyr)
library(scales)
library(RColorBrewer)
library(tidyr)
```

3. Data Loading

```
Death_US <- fread("DeathRecords.csv", header = T)
MannerOfDeath <- read.csv("MannerOfDeath.csv", header = T)
Edu_1989 <- read.csv("Education1989Revision.csv", header = T)
Edu_2003 <- read.csv("Education2003Revision.csv", header = T)
MaritalStatus <- read.csv("MaritalStatus.csv", header = T)
Race3 <- read.csv("RaceRecode3.csv", header = T)
RaceAll <- read.csv("Race.csv", header = T)
icd10_code <- fread("Icd10Code.csv")
```

4. Prepare Tidy DataSet

Add labels to "MannerOfDeath" variable

```
Death_US$MannerOfDeath <- factor(Death_US$MannerOfDeath)
Death_US$MannerOfDeath <- factor(Death_US$MannerOfDeath,
                           levels = c("0","1","2","3","4","5","7"),
                           labels =c("Not specified", "Accident", "Suicide",
                                     "Homicide", "Pending investigation",
                                     "Could not determine", "Natural") )
```

Create a new subgroup for "Age" variable

```
# teenager ( < 19 years) , young adult (between 19 to 25 years), adult (26-39),
# middle aged ( 40-60), senior citizens ( > 60 )

Death_US$AgeGroup <- cut(Death_US$Age, c(0,19,25,39,60,110))
levels(Death_US$AgeGroup) <- c("Teenager", "Young_Adult", "Adult", "Middle_Age",
                               "Senior_Citizens")
```

Add labels to "MaritalStatus" variable

```
Death_US$MaritalStatus <- factor(Death_US$MaritalStatus)
levels(Death_US$MaritalStatus) <- c("Divorced", "Married", "Single(NM)",
                                    "Unknown", "Widowed")
```

Add labels to "RaceRecode3" variable

```
Death_US$RaceRecode3 <- factor(Death_US$RaceRecode3)
levels(Death_US$RaceRecode3) <- c("White", "Races_Other_Than_White/Black", "Black")
```

5. Exploratory Analysis On Dataset

```
str(Death_US)
```

```
## Classes 'data.table' and 'data.frame':   2631171 obs. of  39 variables:
##  $ Id                           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ResidentStatus               : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Education1989Revision        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Education2003Revision        : int  2 2 7 6 3 5 4 4 3 3 ...
##  $ EducationReportingFlag       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ MonthOfDeath                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex                          : chr  "M" "M" "F" "M" ...
##  $ AgeType                      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Age                          : int  87 58 75 74 64 93 82 55 86 23 ...
##  $ AgeSubstitutionFlag          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AgeRecode52                  : int  43 37 41 40 38 44 42 37 43 30 ...
##  $ AgeRecode27                  : int  23 17 21 20 18 24 22 17 23 10 ...
##  $ AgeRecode12                  : int  11 8 10 9 8 11 10 8 11 4 ...
##  $ InfantAgeRecode22            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PlaceOfDeathAndDecedentsStatus: int  4 4 4 6 4 6 4 7 4 7 ...
##  $ MaritalStatus                : Factor w/ 5 levels "Divorced","Married",..: 2 1 5 1 1 5 2 3 5 3 .
##  $ DayOfWeekOfDeath             : int  4 3 2 1 2 4 6 6 6 4 ...
##  $ CurrentDataYear              : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
##  $ InjuryAtWork                 : chr  "U" "U" "U" "U" ...
##  $ MannerOfDeath                : Factor w/ 7 levels "Not specified",..: 7 7 7 7 7 7 7 7 7 3 ...
##  $ MethodOfDisposition          : chr  "C" "C" "C" "C" ...
##  $ Autopsy                      : chr  "N" "N" "N" "N" ...
##  $ ActivityCode                 : int  99 99 99 99 99 99 99 99 99 9 ...
##  $ PlaceOfInjury                : int  99 99 99 99 99 99 99 99 99 5 ...
##  $ Icd10Code                    : chr  "I64" "I250" "J449" "I48" ...
```

```
##  $ CauseRecode358                : int  238 214 267 228 214 280 215 214 175 429 ...
##  $ CauseRecode113                : int  70 62 86 68 62 111 63 62 111 125 ...
##  $ InfantCauseRecode130          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CauseRecode39                 : int  24 21 28 22 21 37 21 21 37 40 ...
##  $ NumberOfEntityAxisConditions  : int  1 3 2 3 1 5 4 2 1 3 ...
##  $ NumberOfRecordAxisConditions  : int  1 3 2 3 1 5 3 2 1 3 ...
##  $ Race                          : int  1 1 1 1 1 1 1 2 1 1 ...
##  $ BridgedRaceFlag               : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ RaceImputationFlag            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RaceRecode3                   : Factor w/ 3 levels "White","Races_Other_Than_White/Black",..: 1 1
##  $ RaceRecode5                   : int  1 1 1 1 1 1 1 2 1 1 ...
##  $ HispanicOrigin                : int  100 100 100 100 100 100 100 100 100 100 ...
##  $ HispanicOriginRaceRecode      : int  6 6 6 6 6 6 6 7 6 6 ...
##  $ AgeGroup                      : Factor w/ 5 levels "Teenager","Young_Adult",..: 5 4 5 5 5 5 5 4 5
##  - attr(*, ".internal.selfref")=<externalptr>
```

Total Number Of Death

```
nrow(Death_US)
```

```
## [1] 2631171
```

How long can we expect to live?

```
mean(Death_US$Age)
```

```
## [1] 73.41336
```

```
#  Life expectancy of Men
mean(Death_US$Age[Death_US$Sex == 'M'])
```

```
## [1] 70.22786
```

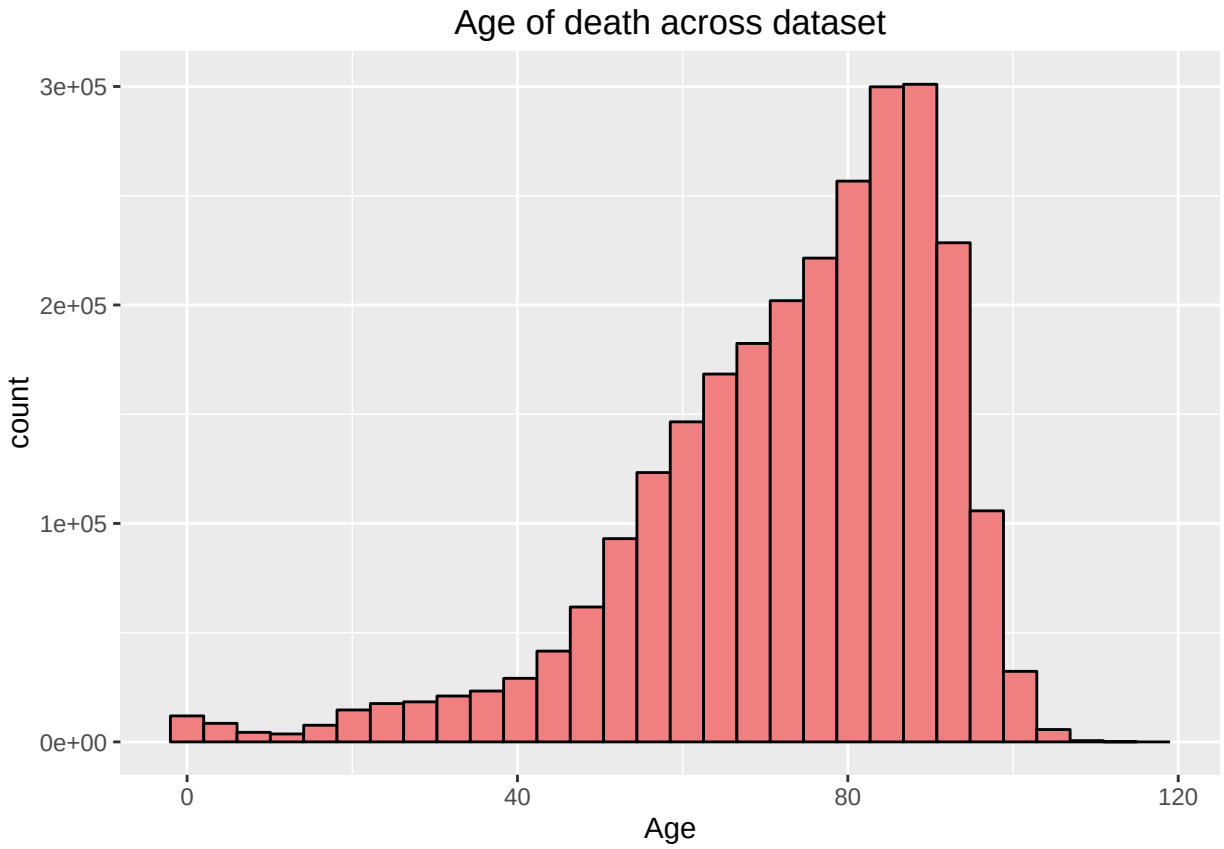
```
#  Life expectancy of Women
mean(Death_US$Age[Death_US$Sex == 'F'])
```

```
## [1] 76.67668
```

Life expectancy at birth represents the average number of years that a group of infants would live which is 73.4 years for he total U.S. population.
Life expectancy for females is 76.6 years.
Life expectancy for males is 70.2 years.
That difference of 6.2 years.

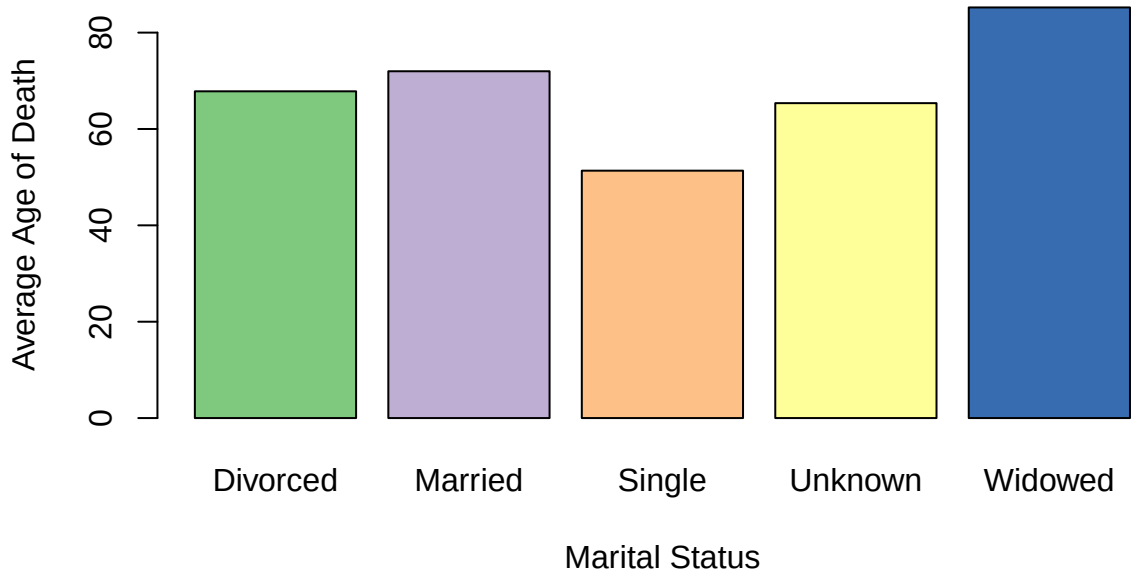
Histogram of Age of death across dataset

```r
# There are 571 observations which are having Age = 999, so removing these observations
# count(Death_US[Death_US$Age > 125, ])
Death_US <- Death_US[Death_US$Age <= 125, ]
Death_US %>% ggplot(aes(x=Age)) + geom_histogram(color="Black",fill="lightcoral") +
             ggtitle("Age of death across dataset")
```



Marital Status Analysis

```r
#gets mean age of death for each marital status category
marMeans <- data.frame(aggregate(Death_US$Age ~ Death_US$MaritalStatus, FUN = mean))

barplot(marMeans$Death_US.Age,
        col = brewer.pal(5, "Accent"),
        ylab = "Average Age of Death",
        xlab = "Marital Status",
        main = "Average Age of Death vs Marital Status",
        names.arg = c("Divorced", "Married", "Single", "Unknown", "Widowed")
)
```



As per graph windows have the highest average age of death, this is because many older married people who die (at a mean age of around 80) leave behind a partner in the marriage who will not remarry. The remaining partners make up a significant portion of the widow demographic who have an average age of death of over 80 years.
People who are reported as single have died at a substantially younger average age than other marital demographics. While single people may die at a lower age for certain reasons, this mean is surely brought down by all the children and young adults who die early in life while still single.

Analyze the manner of death across different Race

Before beginning the race analysis it's important to know that about 85% of the deaths in this dataset are white, about 12% are black, and the other races each account for under %1 of the total deaths recorded.

Below we can check the table of total death proportion by race.

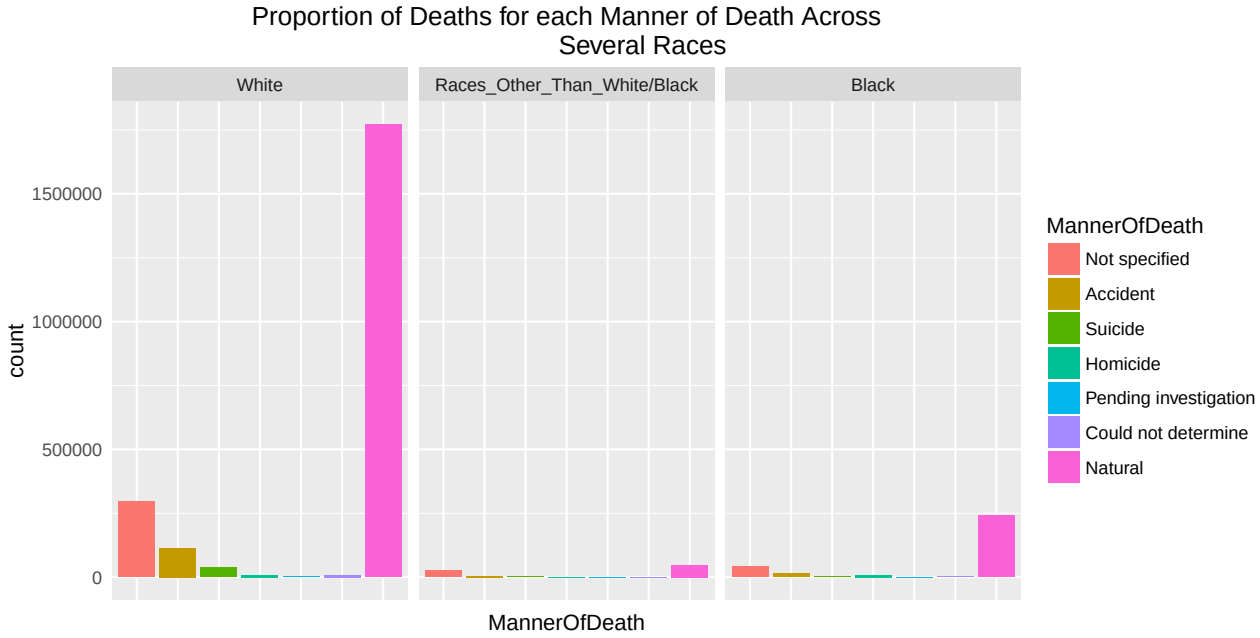
```
# Race labels
RaceName <- c("White", "Black", "American Indian", "Chinese", "Japanese", "Hawaiian",
              "Filipino", "Asian Indian", "Korean", "Samoan", "Vietnamese","Guamanian",
              "Other Asian or Pacific Islander codes18-58",
              "Combined Other Asian or Pacific Islander codes18-68")
#Race death proportions
proportion <- Death_US %>% group_by(Race) %>% summarise(count=n()) %>%
              mutate(pct = count/sum(count))
Race_Proportion <- data.frame(RaceName, round(proportion$pct,3)*100)
colnames(Race_Proportion) <- c("Race", "% of Deaths")
Race_Proportion
```

```
##                                          Race % of Deaths
## 1                                       White        85.2
## 2                                       Black        11.8
## 3                            American Indian         0.7
## 4                                    Chinese         0.2
```

```
## 5                                          Japanese        0.4
## 6                                          Hawaiian        0.0
## 7                                          Filipino        0.3
## 8                                       Asian Indian        0.3
## 9                                            Korean        0.5
## 10                                           Samoan        0.1
## 11                                        Vietnamese        0.0
## 12                                        Guamanian        0.0
## 13          Other Asian or Pacific Islander codes18-58        0.3
## 14 Combined Other Asian or Pacific Islander codes18-68        0.2
```

Lets check the Proportion of Reported Deaths for each Manner of Death Across Several Races

```
Death_US %>% ggplot(aes(x = MannerOfDeath, fill=MannerOfDeath)) + geom_bar() +
             facet_grid(. ~ RaceRecode3) + theme(axis.ticks = element_blank(),
             axis.text.x = element_blank()) +
             ggtitle("Proportion of Deaths for each Manner of Death Across
                     Several Races")
```



Check the Homicide cases across race

```
Death_US_Homicide <- Death_US[Death_US$MannerOfDeath=="Homicide",]

count_homicide <- as.data.frame(Death_US_Homicide %>% group_by(RaceRecode3) %>%
                                summarise(count=n()))
count_homicide
```

```
##                   RaceRecode3 count
```

```
## 1                         White  8076
## 2                         Black  8130
## 3 Races_Other_Than_White/Black   623
```

As discussed above, 85% of the deaths in dataset are belonging to white race, about 12% to black race. Look at the Homicide count across White and Black Race. Black race has a more count of Homicide cases even though the black race death count is only 12% of total count in US.

Lets calculate the percentage of homicide cases across race

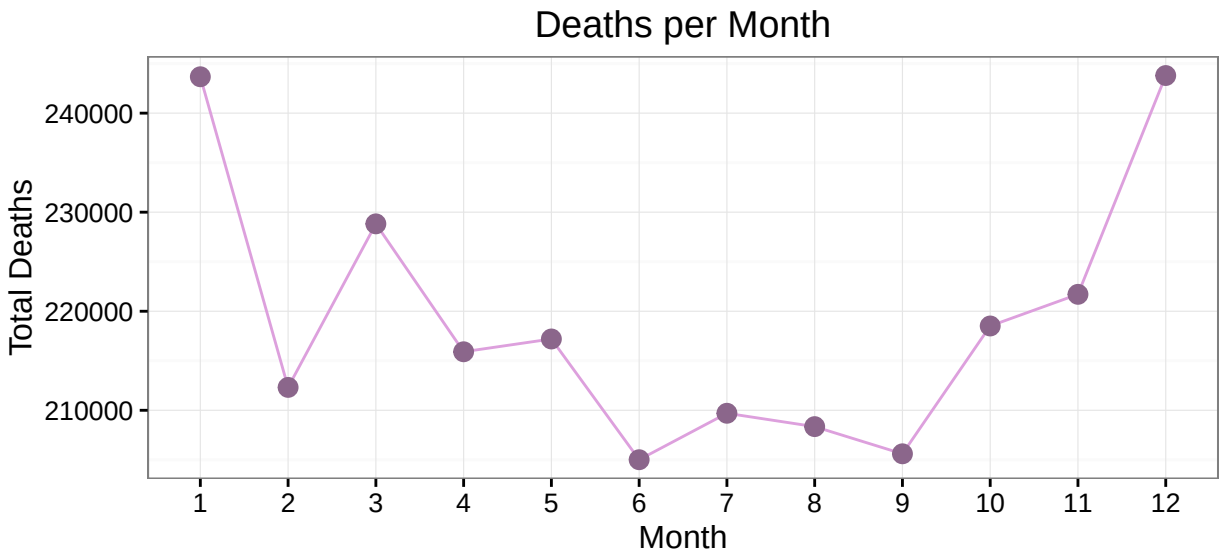
```
count_all <- as.data.frame(Death_US %>% group_by(RaceRecode3) %>% summarise(count=n()))
ratio <- merge(count_homicide, count_all, by="RaceRecode3")
names(ratio) <- c("RaceRecode3", "HomicideCount", "AllCount")
ratio$homicdeRation <- round((ratio$HomicideCount / ratio$AllCount),3)*100
ratio
```

```
##                    RaceRecode3 HomicideCount AllCount homicdeRation
## 1                        Black          8130   309372           2.6
## 2 Races_Other_Than_White/Black           623    80133           0.8
## 3                        White          8076  2241095           0.4
```

From the result we can say that homicide death count in Black race is 6 times more than the homicide death count in White race.

Analyze Death Per Month

```
MonthCount <- Death_US %>% group_by(MonthOfDeath) %>% summarise(count=n())

MonthCount %>% ggplot(aes(x = factor(MonthOfDeath), y = count, group = 1)) +
               geom_line(colour = "plum") + geom_point(colour="plum4", size=3) +
               theme_bw() + xlab("Month") + ylab("Total Deaths") +
               ggtitle("Deaths per Month")
```

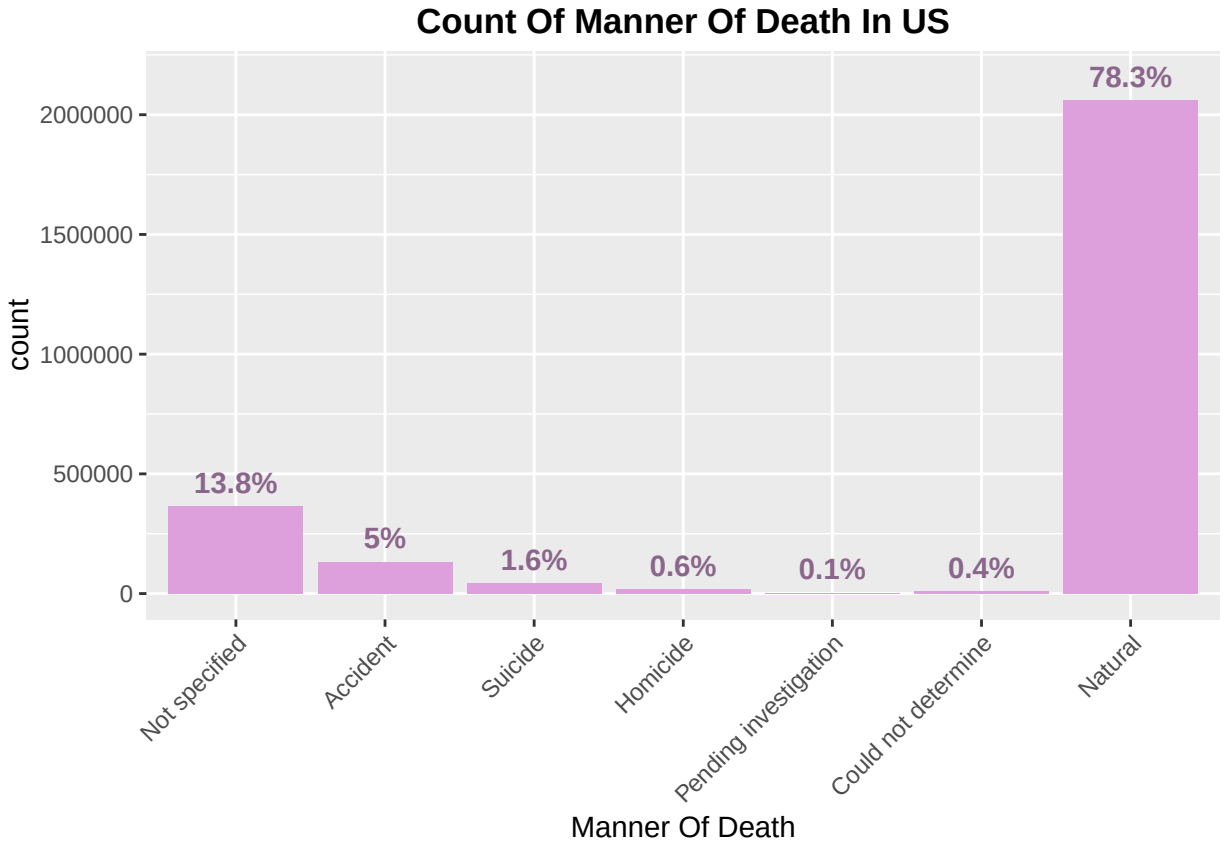


Mr. Trollope was an Englishman writing in the 19th century, but his quote is relevant to modern day U.S.

Indeed, there are seasonal fluctuations in U.S. deaths. One's chances of dying in the winter months are significantly greater than in the summer. This is a statistical fact. It is generally true regardless of where one lives in the U.S.

6. Plot the bar chart to find "Main leading cause of death in US"

```
## calculate the percentage of MannerOfDeath
test.pct <- Death_US %>% group_by(MannerOfDeath) %>% summarise(count=n()) %>%
            mutate(pct = count/sum(count))
# plot bar chart
Death_US %>% ggplot(aes(x = factor(MannerOfDeath))) +
             geom_bar(fill = "plum") +
             xlab("Manner Of Death") + ggtitle("Count Of Manner Of Death In US") +
             theme(plot.title = element_text(lineheight=.8, face="bold")) +
             theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1)) +
             geom_text(data=test.pct, aes(label=paste0(round(pct*100,1),"%"),
             y=count+100000), size=4, color = "plum4", fontface = "bold")
```



"Natural Death" is major factor of Manner Of Death. "Not Specified", "Accident" and "Suicide" factors are also showing considerable effect on US death count.
Accidents and suicides are the manners of death chosen to explore further in the context.

7. Suicide rate in the United States

U.S population in 2014 was 318.9 million, lets calculate the suicide rate . **Extracting and filtering Suicidal death cases**

```
Death_US_Suicide <- Death_US[Death_US$MannerOfDeath == "Suicide", ]

(nrow(Death_US_Suicide) * 100000 ) / 318900000
```

```
## [1] 13.52524
```

Suicidal rate in 2014 is 13.5 per 100,000 population.
As per research from 1999 through 2014, the age-adjusted suicide rate in the United States increased 25%, from 10.5 to 13.5 per 100,000 population, with the pace of increase greater after 2006.

8. Suicide cases by discharge of firearms

To find the Suicide cacses by discharching firearms refereing CauseRecode113 variable. 125 is the list code 125 for firearms sicidal cases.
125 : Intentional self-harm (suicide) by discharge of firearms (X72-X74). The deail can be found here.

```
firearms <- Death_US_Suicide[Death_US_Suicide$CauseRecode113 == "125",]
nrow(firearms) / nrow(Death_US_Suicide)
```

```
## [1] 0.4947603
```

50% of suicidal cases use firearms.

9. Male-to-female suicide death ratio? Who likely take their own lives.

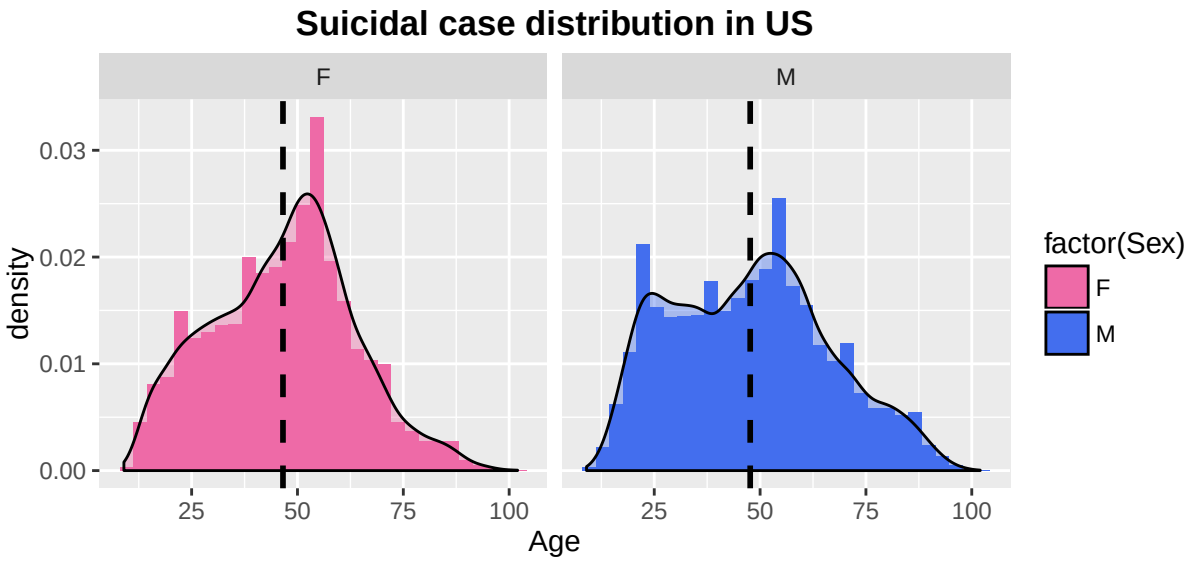
```
# calculate the percentage of Male & Female suicidal death cases
suicide.pct <- as.data.frame(Death_US_Suicide %>% group_by(Sex) %>% summarise(Cases=n(),
                      Mean=mean(Age), Std=sd(Age))) %>% mutate(pct = Cases/sum(Cases))
suicide.pct
```

```
##   Sex Cases     Mean      Std       pct
## 1   M 33357 47.65932 18.85484 0.7733701
## 2   F  9775 46.57483 16.78307 0.2266299
```

Notice that:
* Male and Female groups has the same “mean” for Age of death.
* Majority of suicidal cases belong to male, they represent the 77.34% of the total suicidal cases.

```
# colour palette
cbbPalette <- c("hotpink2", "royalblue2")
Death_US_Suicide_By_Sex <- Death_US_Suicide %>% group_by(Sex) %>%
                           mutate(MeanCalculated=mean(Age))

# plot the bar chart
Death_US_Suicide_By_Sex %>% ggplot(aes(x=Age, y=..density.., fill=factor(Sex))) +
                            geom_histogram() + geom_density(alpha=0.4) +
                            scale_fill_manual(values=cbbPalette) + facet_grid(.~Sex) +
                            geom_vline(aes(xintercept=MeanCalculated),color="black",
                                       linetype="dashed", size=1)+
                            ggtitle("Suicidal case distribution in US") +
                            theme(plot.title = element_text(lineheight=.8, face="bold"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



10. Distribution of suicidal cases over Age & Education level.

```
# Continue analysis on "Death_US_Suicide" data frame
# Check how many suicidal cases belong to 2003 Education levels
table(factor(Death_US_Suicide$EducationReportingFlag))
```

```
##
##     0     1
##  4116 39016
```

"0" :: 1989 revision of education item on certificate
"1" :: 2003 revision of education item on certificate
From the table we can calculate and say 9.5% of data belongs to 1989 education attainment and 90.5% of data belongs to 2003 education attainment. So next we will be looking into the data of 2003 education attainment.

Extract 2003 Education levels

```
Death_US_Suicide_Edu_2003 <-
  Death_US_Suicide[Death_US_Suicide$Education2003Revision != 0, ]
Edu_2003
```

```
##   Code                          Description
## 1    1                    8th grade or less
## 2    2           9 - 12th grade, no diploma
## 3    3 high school graduate or GED completed
## 4    4    some college credit, but no degree
## 5    5                     Associate degree
## 6    6                    Bachelor's degree
## 7    7                      Master's degree
## 8    8      Doctorate or professional degree
## 9    9                               Unknown
```

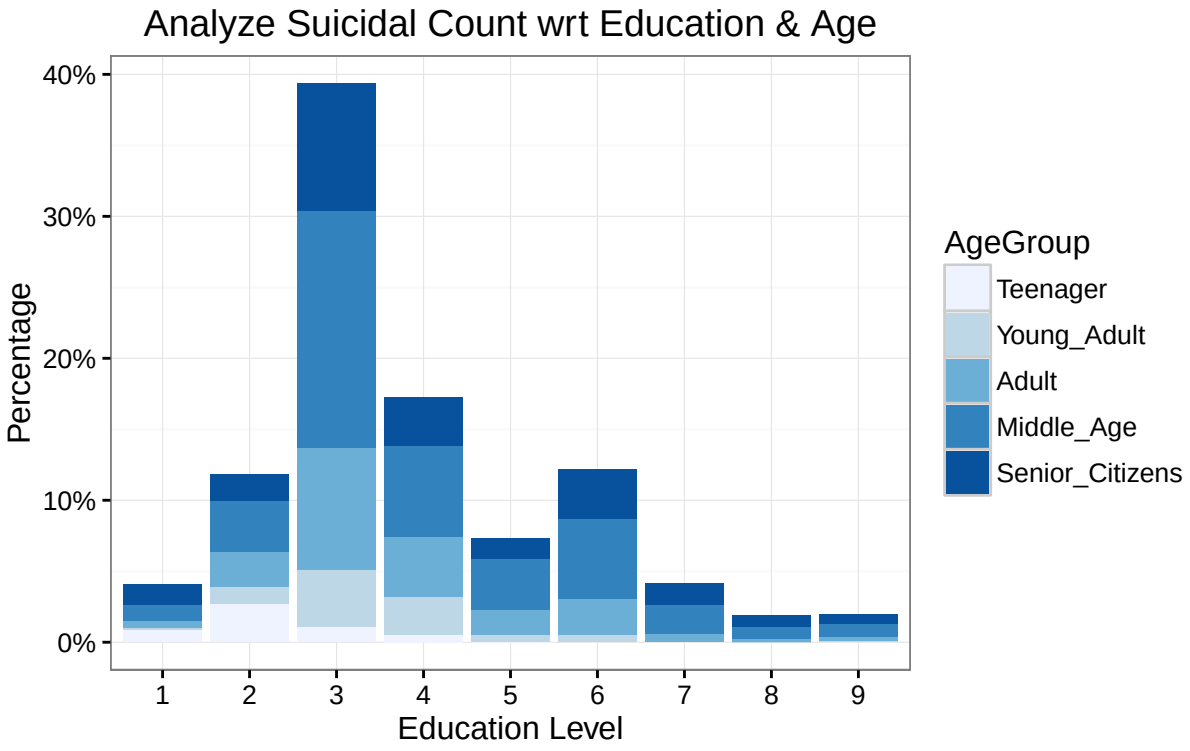
```
table(factor(Death_US_Suicide_Edu_2003$Education2003Revision))
```

```
##
##     1     2     3     4     5     6     7     8     9
##  1577  4622 15356  6726  2857  4754  1618   735   771
```

We have some unknown education entries. 771 / 39016 = 1.9%. Its a less entries, so lets remove the unknown education entries.

Remove unknown education values

```
Death_US_Suicide_Edu_2003 <-
  Death_US_Suicide_Edu_2003[Death_US_Suicide_Edu_2003$Education2003Revision != "unknown", ]

# Plot BarChart
Death_US_Suicide_Edu_2003 %>% ggplot(aes(x=factor(Education2003Revision),fill=AgeGroup))+
                          geom_bar(aes(y = (..count..)/sum(..count..))) +
                          scale_fill_brewer() +
                          scale_y_continuous(labels=percent) +
                          ggtitle("Analyze Suicidal Count wrt Education & Age") +
                          ylab("Percentage") + xlab("Education Level") +
                          theme(plot.title = element_text(lineheight=.8,
                                                          face="bold")) +
                          theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1)) +
                          theme_bw()
```



- From the bar chart we can notice that education level 3, i.e. "high school graduate or GED completed" is displaying high percentage of suicidal cases. This level should be highlighted for further research to compare this data with the earlier years data.
- Also, notice that "Middle_Aged_Person" (41 %) & "Senior_Citizens" (23 %) are the one who committed more suicide. i.e Person who falls to the "AgeGroup > 40" has more tendency to commit the suicide.

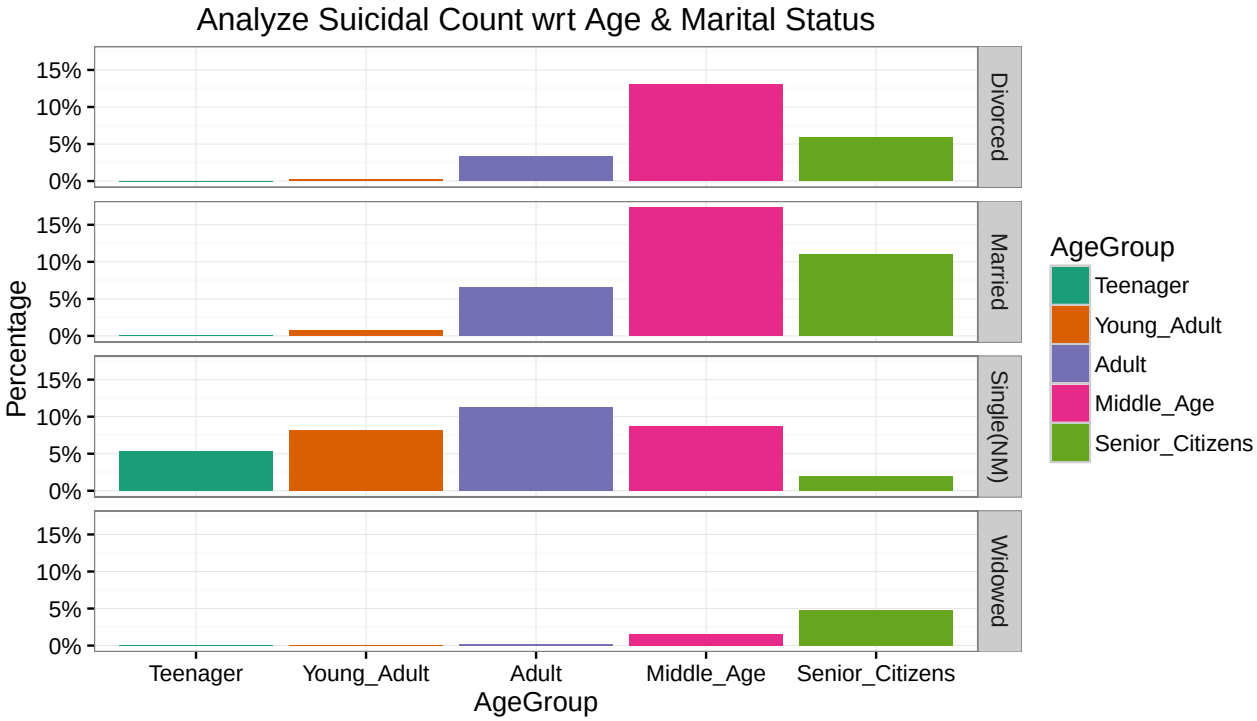
11. Is Marital Status has effect on suicidal cases?

```
table(factor(Death_US_Suicide_Edu_2003$MaritalStatus))
```

```
##
##   Divorced    Married Single(NM)    Unknown    Widowed
##       8663      13725      13692        408       2528
```

```
# There are few unknown MaritalStatus observations which is very less.
# So remove the unknown MaritalStatus observations
Death_US_Suicide_Edu_2003_MariStat <-
  Death_US_Suicide_Edu_2003[Death_US_Suicide_Edu_2003$MaritalStatus!="Unknown", ]

Death_US_Suicide_Edu_2003_MariStat %>% ggplot(aes(x=AgeGroup, fill=AgeGroup)) +
                      geom_bar(aes(y = (..count..)/sum(..count..))) +
                      scale_fill_brewer(palette = "Dark2") +
                      scale_y_continuous(labels=percent) +
                      ggtitle("Analyze Suicidal Count wrt Age & Marital Status") +
                      ylab("Percentage") +
                      facet_grid(MaritalStatus ~ . ) +
```

```
                      theme(plot.title = element_text(lineheight=.8, face="bold")) +
                      theme(axis.text.x=element_text(angle=45,hjust=0.5,vjust=1)) +
                      theme_bw()
```



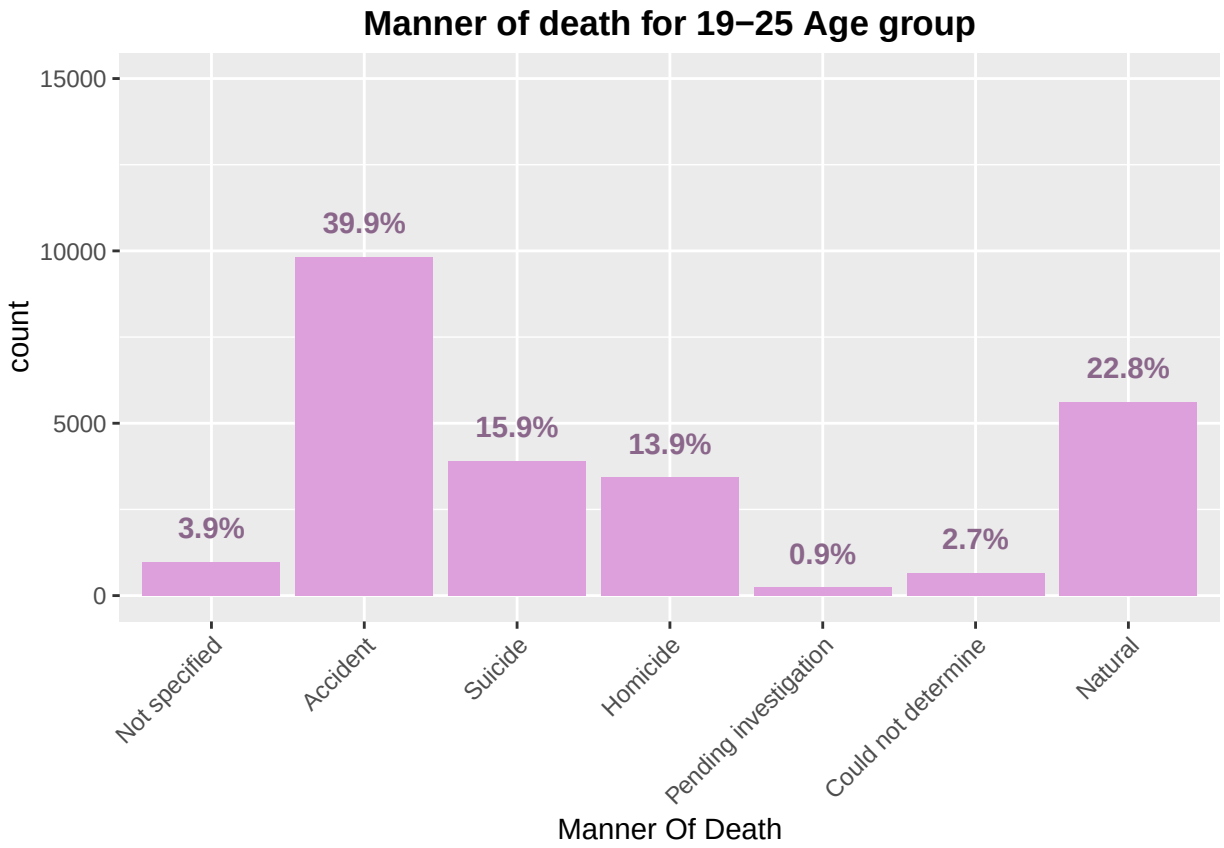
From the graph we can say that Suicidal percentage with MaritalStatus as "Widowed" is less, this again add point to our above discussion saying Women are less likely to commit suicide. Also we can notice that person belongs to "Middle_Age & Married"" group are more likely to commit suicide. But, there is no considerable link between Age & Marital Status.

12. Distribution of cause death in Young Adult (19-25) Age Group

```
Death_US_YoungAdult <- Death_US[Death_US$AgeGroup == "Young_Adult"]
nrow(Death_US_YoungAdult)
```

```
## [1] 24635
```

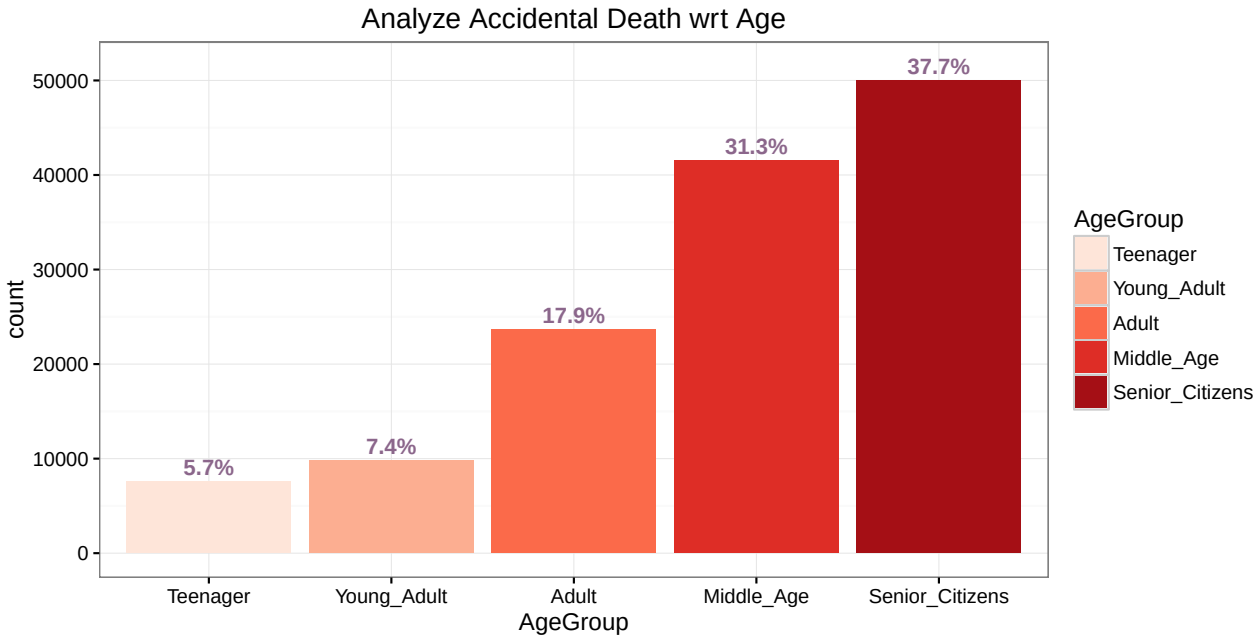
```
test.pct2 <- Death_US_YoungAdult %>% group_by(MannerOfDeath) %>% summarise(count=n()) %>%
  mutate(pct = count/sum(count))

Death_US_YoungAdult %>% ggplot(aes(x = factor(MannerOfDeath))) +
  geom_bar(fill = "plum") + ylim(0, 15000) +
  xlab("Manner Of Death") + ggtitle("Manner of death for 19-25 Age group") +
  theme(plot.title = element_text(lineheight=.8, face="bold")) +
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1)) +
  geom_text(data=test.pct2, aes(label=paste0(round(pct*100,1),"%"),
                                y=count+1000), size=4, color = "plum4", fontface = "bold")
```



- From the barplot we can notice that the Natural death of people belong to 19-25 age group is less 23%.
- 40% of young adult grouped people die in Accident.
- 30% of young adult grouped people die in Suicide & Homicide cases.

13. Which age group has more accidental death? Is younger group shows more rate?

```
Death_US_Accident <- Death_US[Death_US$MannerOfDeath=="Accident",]
# to remove NA values
Death_US_Accident <- Death_US_Accident[!is.na(Death_US_Accident$AgeGroup),]

# calculate the percentage of Accidental death over age
Acci.pct <- Death_US_Accident %>% group_by(AgeGroup) %>% summarise(count=n()) %>%
  mutate(pct = count/sum(count))

Death_US_Accident %>% ggplot(aes(x=AgeGroup, fill=AgeGroup)) +
                geom_bar() +
                scale_fill_brewer(palette = "Reds") +
                ggtitle("Analyze Accidental Death wrt Age") +
                theme(plot.title = element_text(lineheight=.8, face="bold")) +
                theme(axis.text.x=element_text(angle=45,hjust=0.5,vjust=1)) +
                theme_bw() +
                geom_text(data=Acci.pct, aes(label=paste0(round(pct*100,1),"%"),
                        y=count+1500), size=4, color = "plum4", fontface = "bold")
```



From the graph we got to know that 69% of total accidental death happens in "AgeGroup > 40". But "AgeGroup > 40" has a very large population. Also, as observed in "Distribution of cause death in Young Adult (19-25) Age Group" recalls that 40% of young adult grouped people die in Accident. So lets perform analysis with accidental death percentage for each age group.

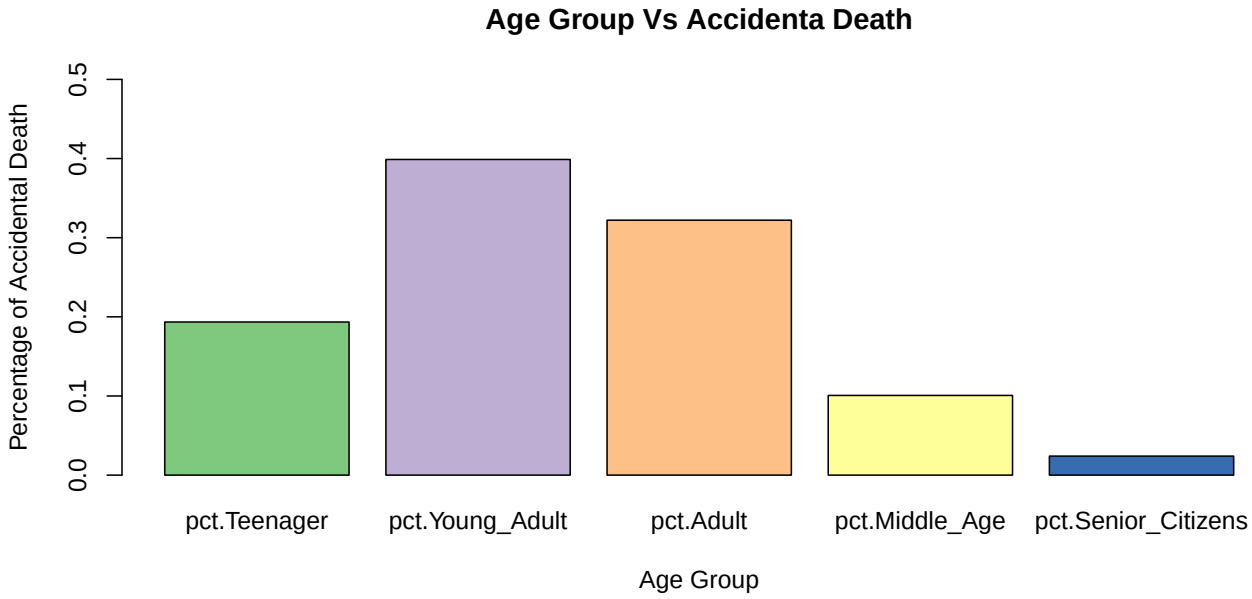
Accidental Death Percentage for each age group

```
# calculate percentage for "Teenager" age group
test.pct1 <-as.data.frame(Death_US[Death_US$AgeGroup == "Teenager"] %>%
                            group_by(MannerOfDeath) %>% summarise(count = n()) %>%
                            mutate(pct = count/sum(count)))

# calculate percentage for "Adult" age group
test.pct3 <- as.data.frame(Death_US[Death_US$AgeGroup == "Adult"] %>%
                             group_by(MannerOfDeath) %>% summarise(count = n()) %>%
                             mutate(pct = count/sum(count)))

# calculate percentage for "Middle_Age" age group
test.pct4 <- as.data.frame(Death_US[Death_US$AgeGroup == "Middle_Age"] %>%
                             group_by(MannerOfDeath) %>% summarise(count = n()) %>%
                             mutate(pct = count/sum(count)))

# calculate percentage for "Senior_Citizens" age group
test.pct5 <- as.data.frame(Death_US[Death_US$AgeGroup == "Senior_Citizens"] %>%
                             group_by(MannerOfDeath) %>% summarise(count = n()) %>%
                             mutate(pct = count/sum(count)))

# Merge all the records
set.seed(123)
merge1 <- merge(test.pct1, test.pct2, by = "MannerOfDeath")
```

```r
names(merge1) <- c("MannerOfDeath", "count.Teenager", "pct.Teenager", "count.Young_Adult",
                   "pct.Young_Adult")

merge2 <- merge(merge1, test.pct3, by = "MannerOfDeath")
colnames(merge2)[c(6,7)] <- c("count.Adult", "pct.Adult")

merge3 <- merge(merge2, test.pct4, by = "MannerOfDeath")
colnames(merge3)[c(8,9)] <- c("count.Middle_Age", "pct.Middle_Age")

merge4 <- merge(merge3, test.pct5, by = "MannerOfDeath")
colnames(merge4)[c(10,11)] <- c("count.Senior_Citizens", "pct.Senior_Citizens")

merge4 <- as.data.frame(merge4)

# select related columns and filter the dataframe
data_accident_per <- merge4[1, c(1,3,5,7,9,11)]
data_accident_per <-as.data.frame( t(data_accident_per) )
data_accident_per <- as.data.frame(data_accident_per[-1, ])
names(data_accident_per) <- "Accident_Per"
data_accident_per
```

```
##                     Accident_Per
## pct.Teenager            0.193456
## pct.Young_Adult         0.3988228
## pct.Adult               0.3220788
## pct.Middle_Age          0.1006917
## pct.Senior_Citizens    0.02402903
```

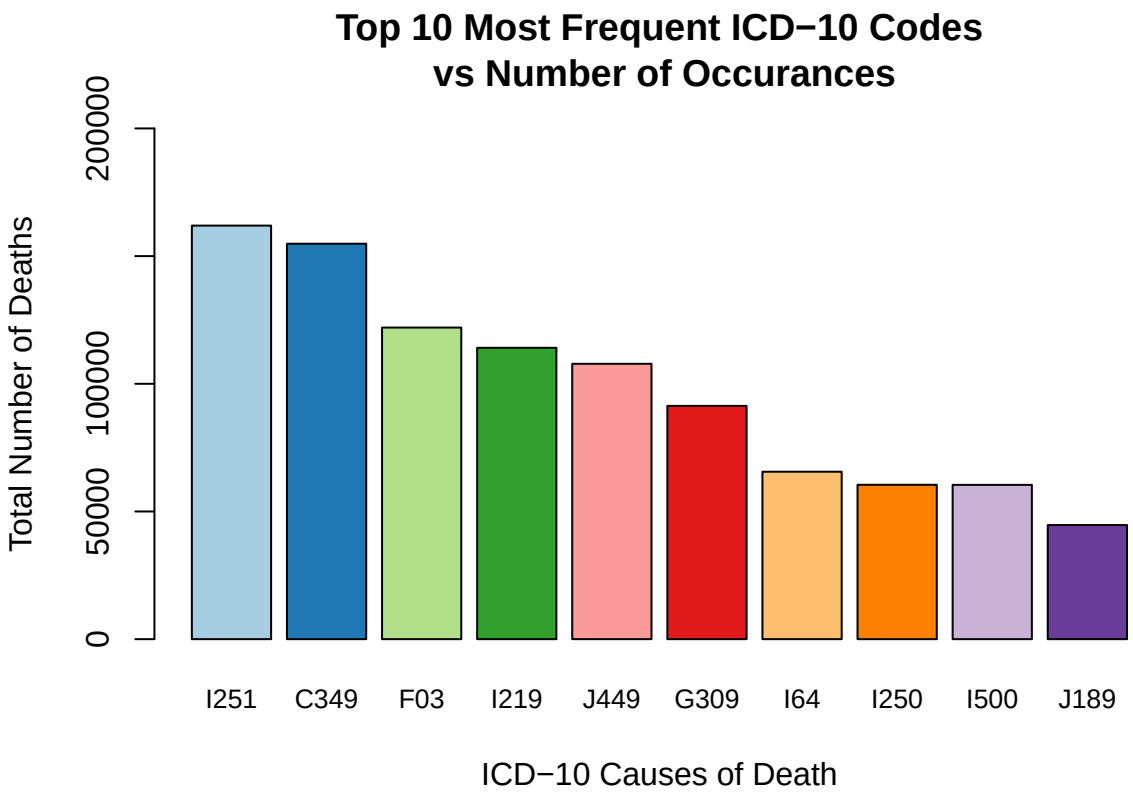
```r
barplot(as.numeric(as.character(data_accident_per$Accident_Per)),
        col = brewer.pal(5, "Accent"), ylim = c(0.0, 0.5),
        ylab = "Percentage of Accidental Death",
        xlab = "Age Group",
        main = "Age Group Vs Accidenta Death",
        names.arg = c("pct.Teenager", "pct.Young_Adult", "pct.Adult", "pct.Middle_Age",
                      "pct.Senior_Citizens")
)
```



Finally from the plot we can say that younger group show the more accidental death rate.

14. Top 10 ICD10 Codes Associated with Deaths

```
# group by Icd10Code variable and count the deaths
icd <-as.data.frame( Death_US %>% group_by(Icd10Code) %>% summarise(count=n()))
# sort the Icd10Code
icd <- icd[order(icd$count,decreasing = T),]
# extract Top 10 Icd10Code and add discription
icd_top10 <- head(icd,10)
ICDcode_names <- icd10_code[icd10_code$Code %in% icd_top10$Icd10Code]
icd_top10 <- merge(icd_top10,ICDcode_names, by.x= "Icd10Code", by.y="Code")
icd_top10 <- icd_top10[order(icd_top10$count,decreasing = T),]

barplot(icd_top10$count, col = brewer.pal(10, "Paired"),
        names.arg = icd_top10$`Icd10Code`,
        main = "Top 10 Most Frequent ICD-10 Codes\n vs Number of Occurances",
        xlab = "ICD-10 Causes of Death",
        ylab = "Total Number of Deaths",
        ylim = c(0, 200000),
        cex.names=0.85)
```




```
icd_top10[-2]
```

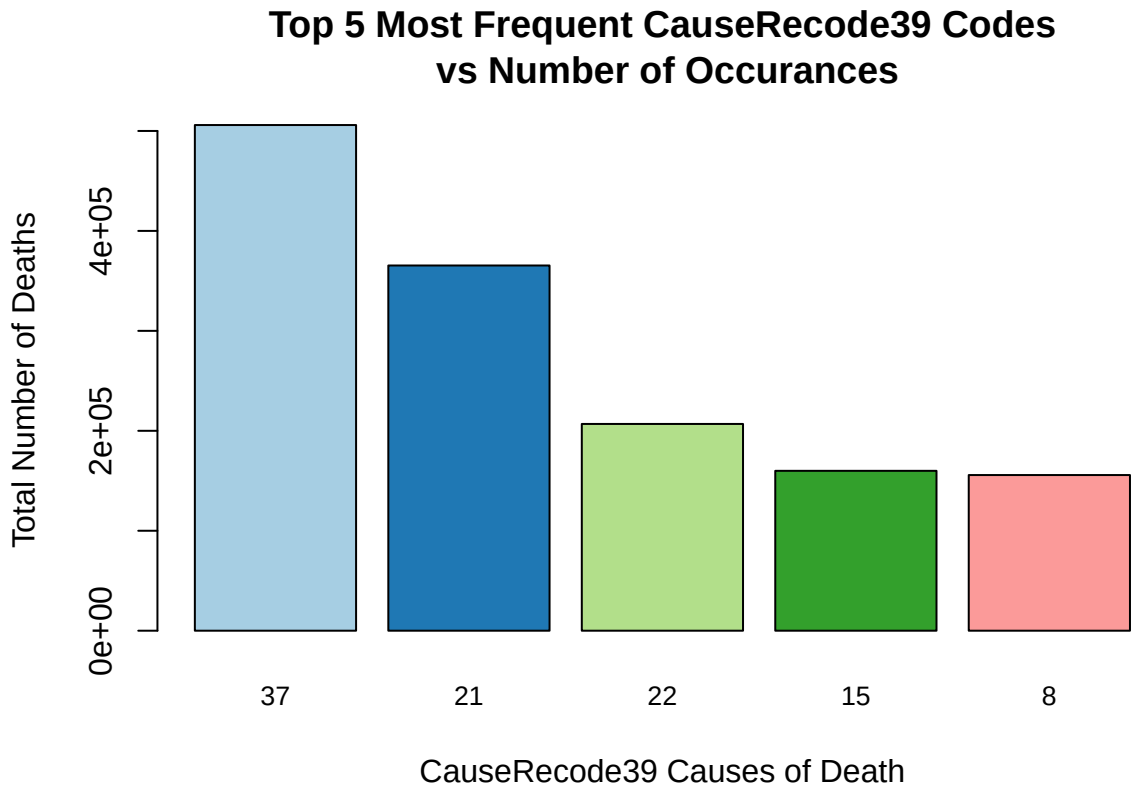
```
##    Icd10Code                                        Description
## 6       I251                          Atherosclerotic heart disease
## 1       C349    Malignant neoplasm: Bronchus or lung, unspecified
## 2        F03                                 Unspecified dementia
## 4       I219             Acute myocardial infarction, unspecified
## 10      J449   Chronic obstructive pulmonary disease, unspecified
## 3       G309                         Alzheimer disease, unspecified
## 8        I64   Stroke, not specified as haemorrhage or infarction
## 5       I250 Atherosclerotic cardiovascular disease, so described
## 7       I500                              Congestive heart failure
## 9       J189                                Pneumonia, unspecified
```

15. Top 5 CauseRecode39 Codes Associated with Deaths

CauseRecode39 is a Cause of death recoded into 39 bins

```
# group by CauseRecode39 variable and count the deaths
icd_39 <-as.data.frame( Death_US %>% group_by(CauseRecode39) %>% summarise(count=n()))
# sort the CauseRecode39
icd_39 <- icd_39[order(icd_39$count,decreasing = T),]
# extract Top 10 CauseRecode39
icd39_top10 <- head(icd_39,5)

barplot(icd39_top10$count, col = brewer.pal(5, "Paired"),
```

```
      names.arg = icd39_top10$`CauseRecode39`,
       main = "Top 5 Most Frequent CauseRecode39 Codes\n vs Number of Occurances",
       xlab = "CauseRecode39 Causes of Death",
       ylab = "Total Number of Deaths",
       cex.names=0.85)
```



CODES

37 :: All other diseases (Residual) (A00-A09,A20-A49,A54-B19,B25-B99,D00-E07, E15-G25,G31-H93,I80-J06,J20-J39,J60-K22,K29-K66,K71-K72, K75-M99,N10-N15,N20-N23,N28-N98,U04)

21 :: Ischemic heart diseases (I20-I25)

22 :: Other diseases of heart (I00-I09,I26-I51)

25 :: Other malignant neoplasms (C00-C15,C17,C22-C24,C26-C32,C37-C49,C51-C52,C57-C60,C62-C63,C69-C81,C88,C90,C96-C97)

8 :: Malignant neoplasms of trachea, bronchus and lung (C33-C34)