Springboard Capstone Project

Bryon Kent

September 05, 2016

Introduction

One of the largest epidemics sweeping the veterans community is that of suicide, with roughly 22 people a day committing suicide accoridng to some estimates due to a number of often-times interrelated causes. I set-out to create a model that predicts the risk level for veteran suicide in a given community such that ones with a higher risk profile could then be honed-in on for specialized programs to increase awareness and to ensure that the proper resources are put in place to help eradicate this phenomena. Given that the Department of Defense (DoD) closeley protects Peronsally Identifiable Information (PII) and Personal Health Infomation (PHI), it was not possible to obtain the necessary data from the DoD to accomplish this directly. As such, it was necessary to develop a risk-profile using the population writ-large. This script explores the mortality data in the United States for *(the year 2014)* based on the data released by the Centers for Disease Control and Prevention. This script looks looks into the frequency of suicide cases for the different factors such as age, sex, race, education, and marital status

Pre-Processing

```
#calling libraries

library(dplyr)
library(ggplot2)
```

Data Loading

```
DeathRec<-read.csv("/Users/Athena/Documents/My-Github-
Repository/DeathRecords.csv",stringsAsFactors = F,header = T)
Race<-read.csv("/Users/Athena/Documents/My-Github-
Repository/Race.csv",stringsAsFactors = F,header = T)
AgeType<-read.csv("/Users/Athena/Documents/My-Github-
Repository/AgeType.csv",stringsAsFactors = F,header = T)
Edu2003<-read.csv("/Users/Athena/Documents/My-Github-
Repository/Education2003Revision.csv",stringsAsFactors = F,header = T)
Marital_table<-read.csv("/Users/Athena/Documents/My-Github-
Repository/MaritalStatus.csv",stringsAsFactors = F,header = T)
```

Data Exploration

In the following sections we will explore the data behind reported suicide statistics for 2014, such as age, sex, education, and marital status.

Extracting and filtering Suicide Cases

As we are only interested in the reported suicide cases, we will extract the entries from the total death record by filtering with MannerOfDeath==2 *(listed as the code for suicide)*.

```
Suicide<-DeathRec %>%
         filter(MannerOfDeath==2 )
```

The column AgeType specifies the units of Age column as per the following codes:

```
##   Code    Description
## 1    1          Years
## 2    2         Months
## 3    4           Days
## 4    5          Hours
## 5    6        Minutes
## 6    9 Age not stated
```

For the extracted suicide cases we have the following AgeTypes, which means all the listed ages are in years except for 9 entries with non-stated age.

```
table(Suicide$AgeType)

##
##     1     9
## 43132     7
```

We will exclude these 9 cases in order to avoid bias in the results.

```
Suicide<-Suicide %>%
         filter(AgeType!=9)
```

Distribution of Suicide Cases Over Age

For both males and females, we can see the number of suicide cases, mean age, and standard deviation are as follows:

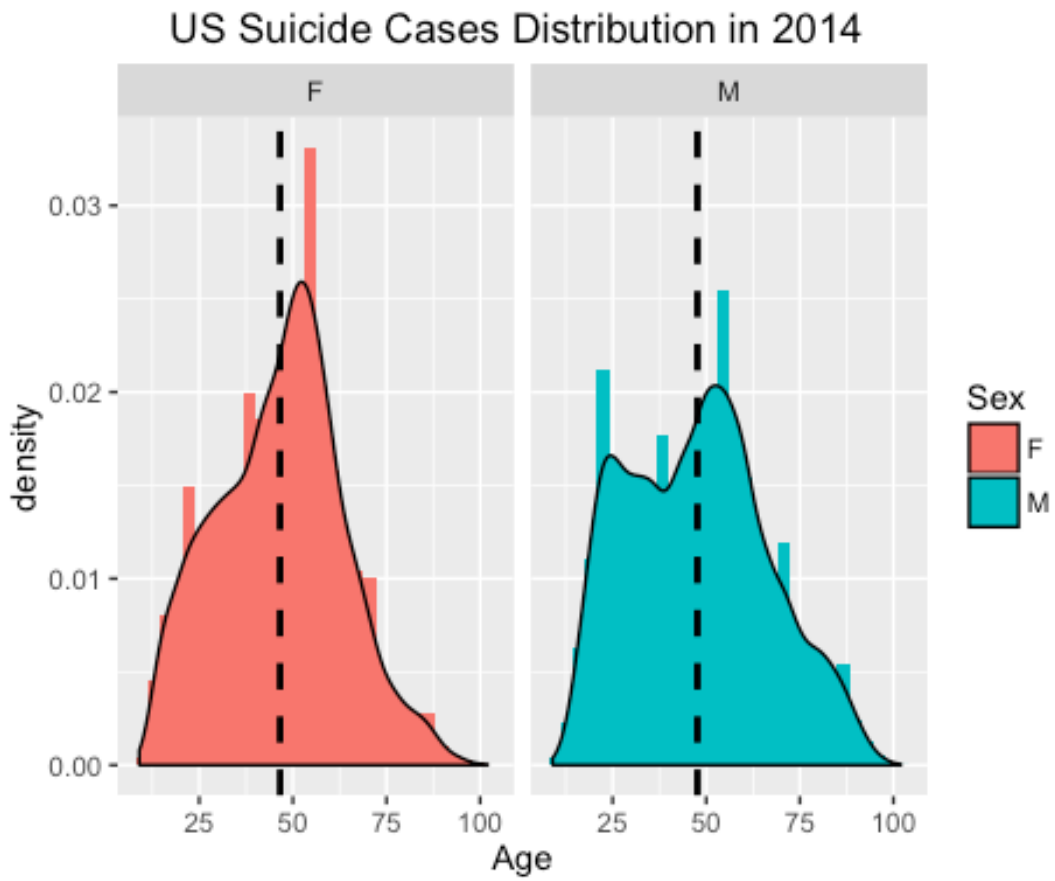
```
SuicideSex<-Suicide %>%
            group_by(Sex)%>%

summarise(Cases=n(),Percentage=100*n()/length(.$Id),Mean=mean(Age),Std=sd(Age
))

SuicideSex

## # A tibble: 2 x 5
##     Sex Cases Percentage     Mean      Std
##   <chr> <int>      <dbl>    <dbl>    <dbl>
## 1     F  9775   22.66299 46.57483 16.78307
## 2     M 33357   77.33701 47.65932 18.85484
```

From this, we can see that:

- The mean age for both groups is approximately the same.
- The overwhelming majority of the cases reported are male.

We would like to get a sense for the distribution and the mean for both males and females, so we'll group the data by sex and plot the distribution as follows:

```
#group data by Sex and caclculate mean

SuicideDist<-Suicide %>%
             group_by(Sex) %>%
             mutate(Mean=mean(Age))

ggplot(SuicideDist,aes(x=Age,y=..density.., fill=Sex))+
             geom_histogram()+
             geom_density()+
             geom_vline(aes(xintercept=Mean),color="black",
linetype="dashed", size=1)+
             facet_grid(.~Sex)+
             labs(title="US Suicide Cases Distribution in 2014 ")
```



Distribution of Suicide Cases Over Race

Across the reported races, we can see the number of suicide cases, mean age, and standard deviation are as follows:

```
SuicideRace<-Suicide %>%
             group_by(Race)%>%

summarise(Cases=n(),Percentage=100*n()/length(.$Id),Mean=mean(Age),Std=sd(Age
))

SuicideRace
```

```
## # A tibble: 14 x 5
##     Race Cases  Percentage     Mean      Std
##    <int> <int>       <dbl>    <dbl>    <dbl>
## 1      1 38983 90.38069183 48.27907 18.32820
## 2      2  2449  5.67791895 38.43895 16.70289
## 3      3   491  1.13836595 34.86558 15.10745
## 4      4   218  0.50542521 46.16514 19.29519
## 5      5    76  0.17620328 56.09211 19.31575
## 6      6     6  0.01391079 34.00000 24.45404
## 7      7   139  0.32226653 38.66187 16.47420
## 8     18   149  0.34545117 40.18792 16.73133
## 9     28   189  0.43818974 47.94180 18.42558
## 10    38    15  0.03477696 36.13333 16.85597
## 11    48   101  0.23416489 41.37624 16.90376
## 12    58     9  0.02086618 31.11111 13.75177
## 13    68   225  0.52165446 38.27556 16.69005
## 14    78    82  0.19011407 38.47561 14.39670
```

From this, we can see that:

- The mean age for all race groups is...
- The overwhelming majority of the cases reported are White.

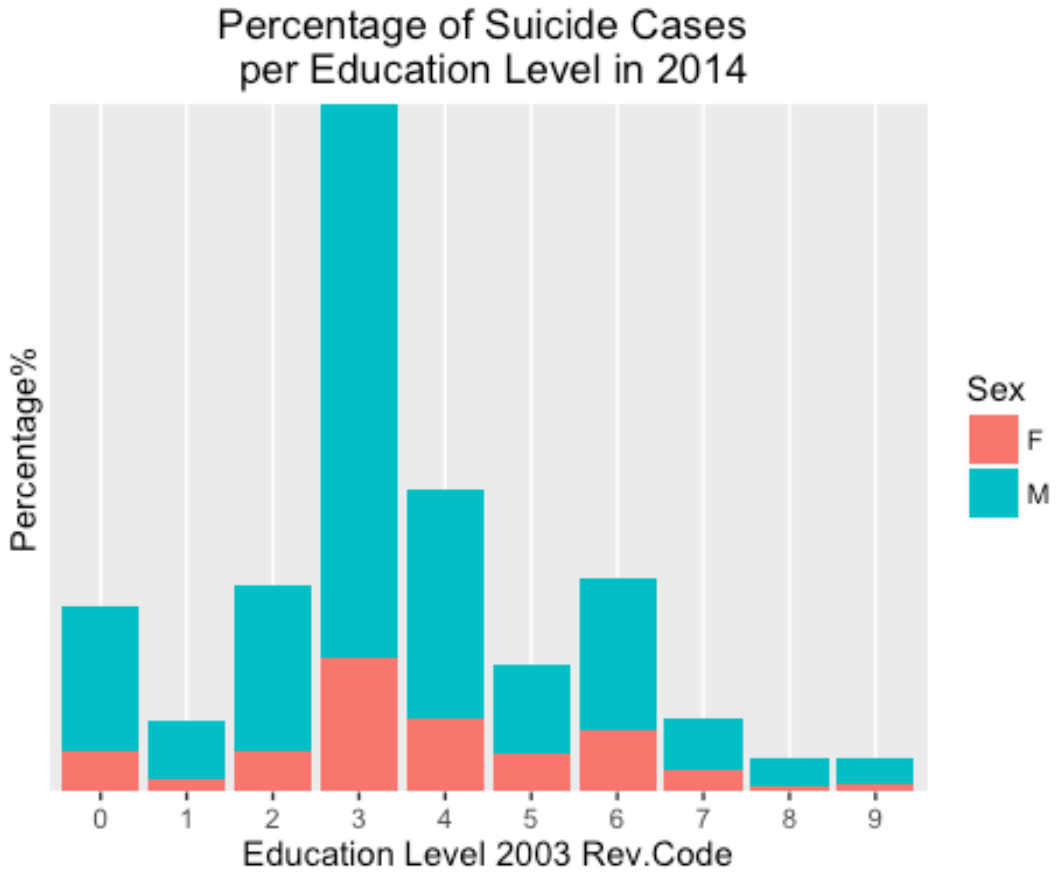
We would like to get a sense for the distribution and density for the reported race groups, so we'll group the data by race and plot the distribution as follows:

Percentage of Suicide Cases Per Education Level

It is stated in the *Education2003Revision.csv* data that we have 8 levels of education *(+ one for the unknown status)* . We would like to see if there are any patterns related to the level of education *(as per 2003 revision)*

```
ByEdu<-Suicide %>%
        group_by(Education2003Revision,Sex) %>%
        summarise(Sum=n(),Percentage=100*n()/length(.$Id))

ggplot(ByEdu,aes(x=factor(Education2003Revision),y=Percentage,fill=Sex))+
```

```
        geom_bar(stat="identity")+
        # facet_grid(.~Sex) +
        scale_x_discrete(name="Education Level 2003 Rev.Code")+
        scale_y_discrete(name="Percentage%",breaks=seq(0, 40, by = 2))+
        labs(title="Percentage of Suicide Cases \n per Education Level in
2014")
```



NOTE There are many cases with education level code=0, which is not listed in the lookup table!

```
##   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
```

From the graph, we can see that:

- a high percentage of the cases comes from an educational background of *(3:high school graduate or GED completed)* which should be highlighted for further reserach and comparisons with previous years whenever available.
- for both sexes, the lowest percentage comes from an educational background *(8:Doctorate or professional degree)*

But we cannot jump to conclusions without taking a deeper look into the percentage of each eduaction level in the original data set (All the death records), which justifies the results. Since we can see that the percentage of Doctorate holders or high-education levels is low compared to the high school graduates in the first place *(which is reasonable in any society)*.

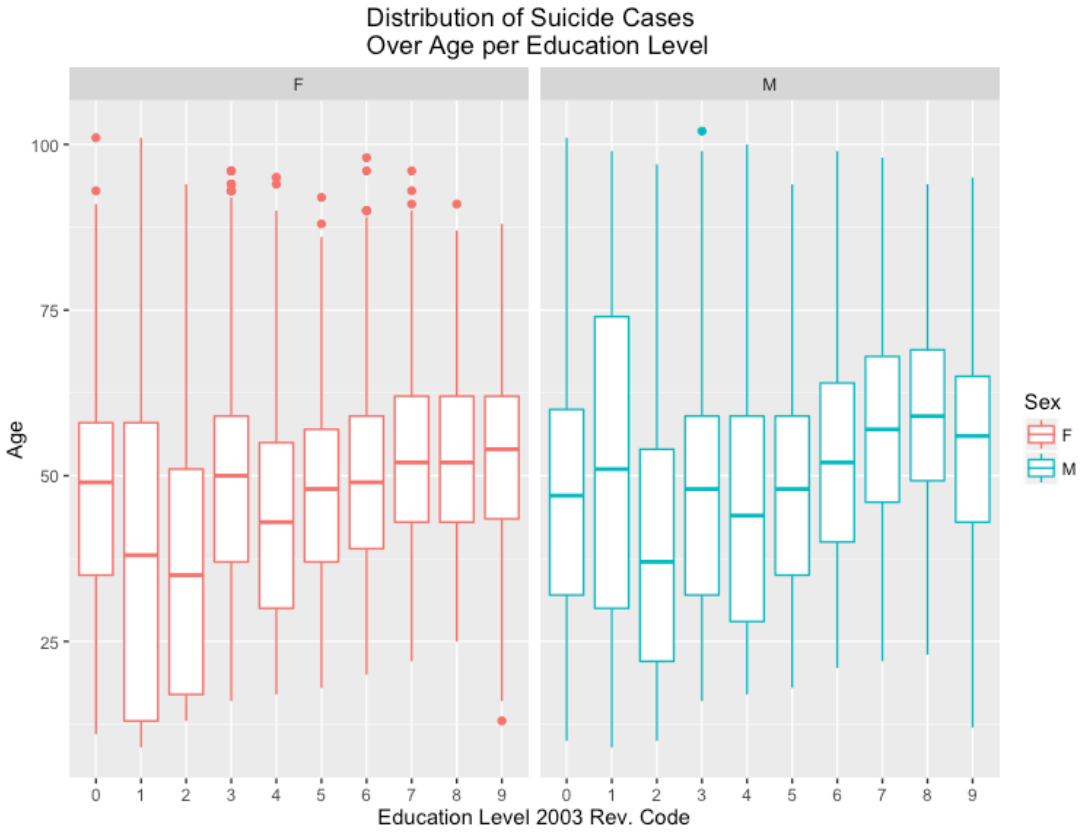
```
#calculate the percentage of each education level in the original data set
for all death manners
EduPercent <- DeathRec %>% group_by(Education2003Revision) %>%
summarise(Cases=n(),Percentage=100*n()/length(.$Id))

#rename the first column as Code to use it to merge with the lookup table
colnames(EduPercent)[1]="Code"

merge(Edu2003, EduPercent,all=T)
```

```
##    Code                            Description   Cases Percentage
## 1     0                                    <NA>  241607   9.182489
## 2     1                     8th grade or less  278271  10.575937
## 3     2            9 - 12th grade, no diploma  259335   9.856258
## 4     3 high school graduate or GED completed 1006055  38.236017
## 5     4    some college credit, but no degree  289850  11.016008
## 6     5                      Associate degree  140143   5.326260
## 7     6                     Bachelor's degree  234316   8.905389
## 8     7                       Master's degree   92593   3.519080
## 9     8      Doctorate or professional degree   38995   1.482040
## 10    9                               Unknown   50006   1.900523
```

Distribution of Suicide Cases Over Age per Education Level

We can also look at the mean and the range of ages within each educational level by plotting a box plot.

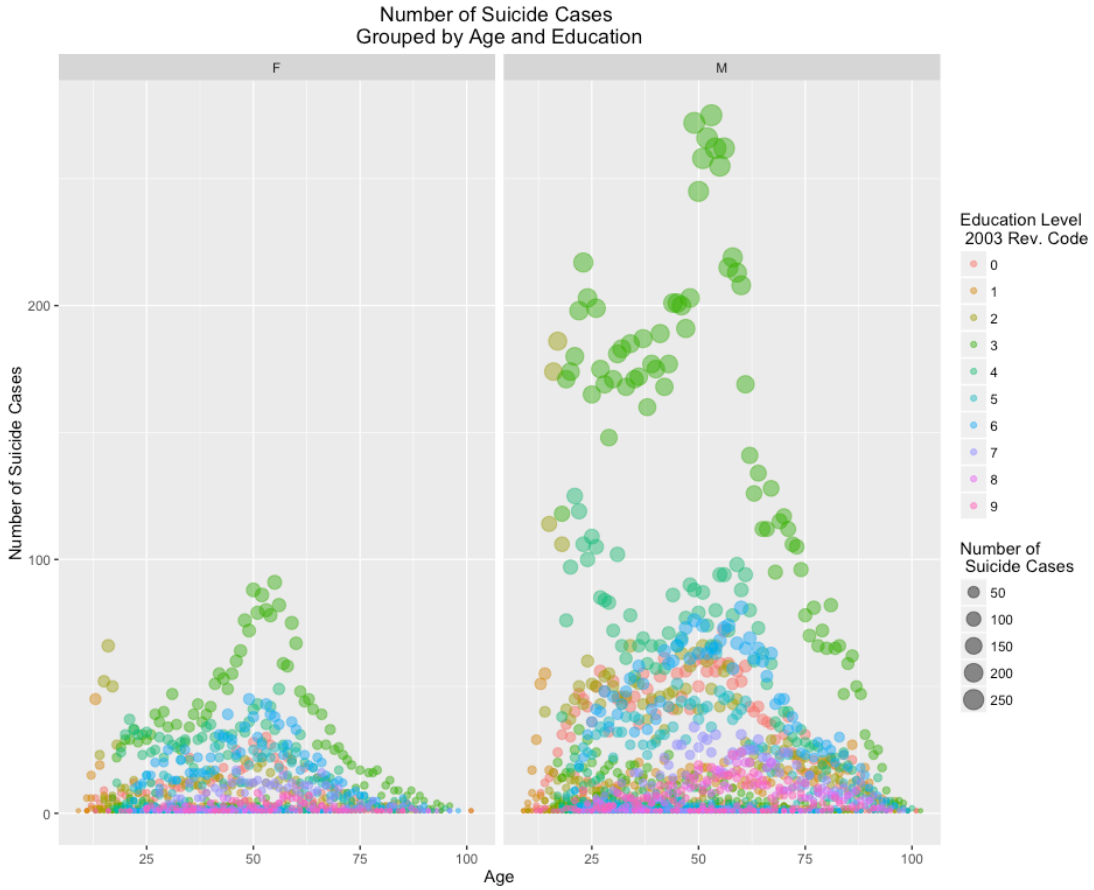
```
ggplot(Suicide,aes(y=Age,x=factor(Education2003Revision)))+
        geom_boxplot(aes(color=Sex))+
        facet_grid(.~Sex)+
        scale_x_discrete(name="Education Level 2003 Rev. Code")+
        labs(title="Distribution of Suicide Cases \n Over Age per Education
Level")
```



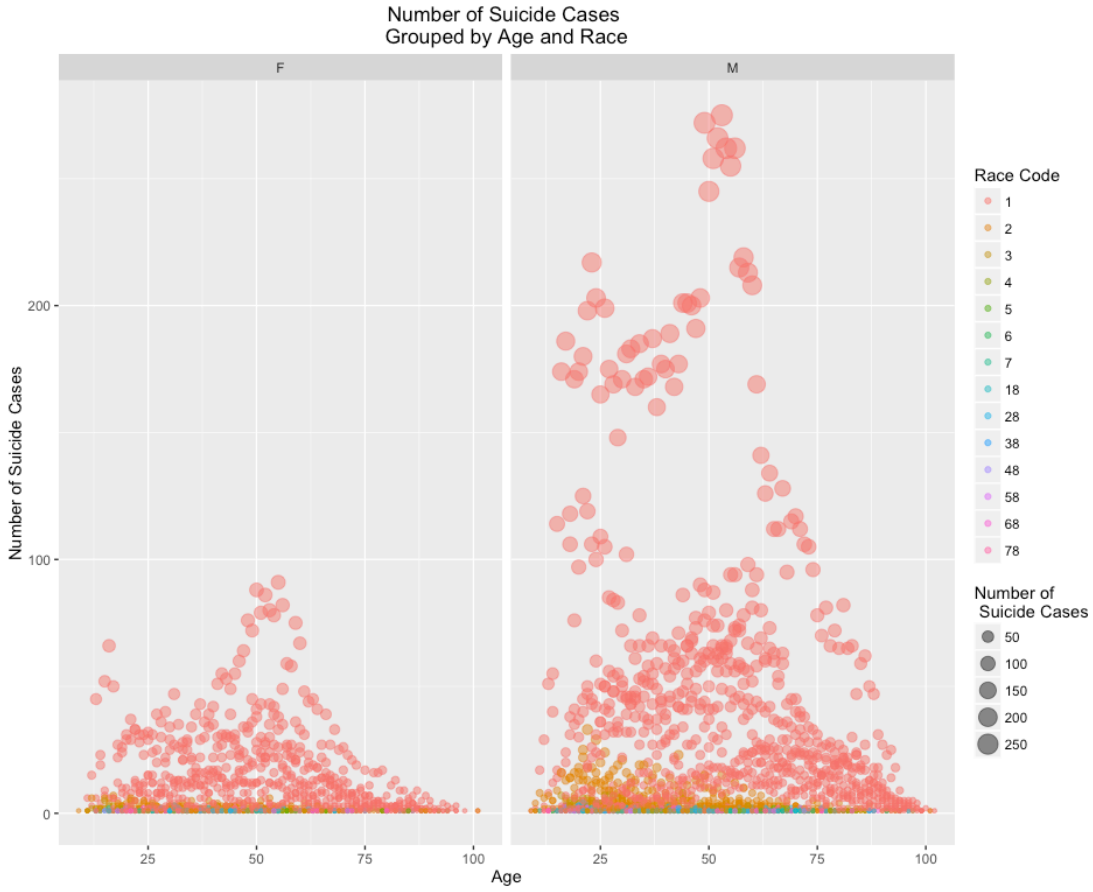
Number of Suicide Cases Grouped by Age, Race, and Education

We can also look further into the data and see the number of Suicide cases for each age, grouped by the education level and Race. As we expect from the previous results, most of the cases are around the mean and come from an educational background of high school or GED.

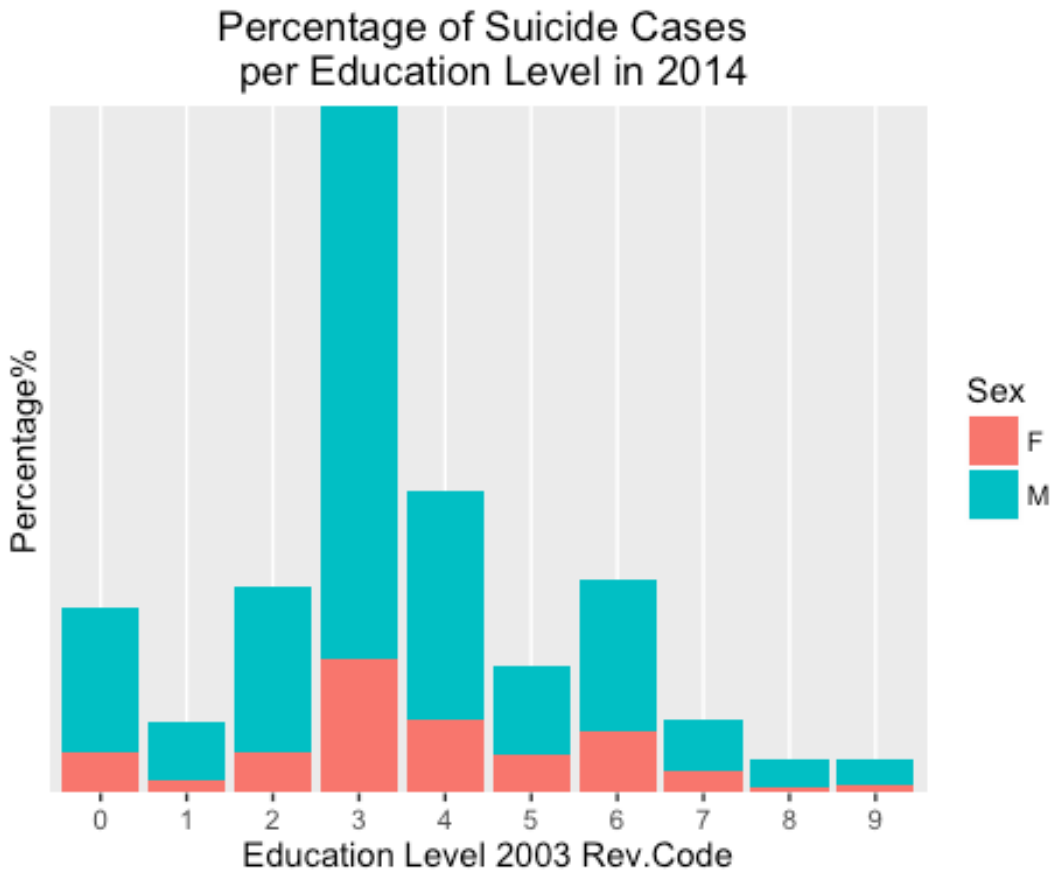
```
ByEdu2<-Suicide %>%
        group_by(Education2003Revision,Sex,Age,Race) %>%
        summarise(Sum=n())


ggplot(ByEdu2,aes(y=Sum,x=Age))+

geom_point(aes(color=factor(Education2003Revision),size=Sum),alpha=0.5)+
        facet_grid(.~Sex)+
        scale_y_continuous(name="Number of Suicide Cases")+
        scale_colour_discrete("Education Level \n 2003 Rev. Code")+
        scale_size_continuous("Number of \n Suicide Cases",
breaks=seq(0,300,by=50))+
        #theme(legend.position="top", legend.box = "horizontal")+
        labs(title="Number of Suicide Cases \n Grouped by Age and Education
")
```



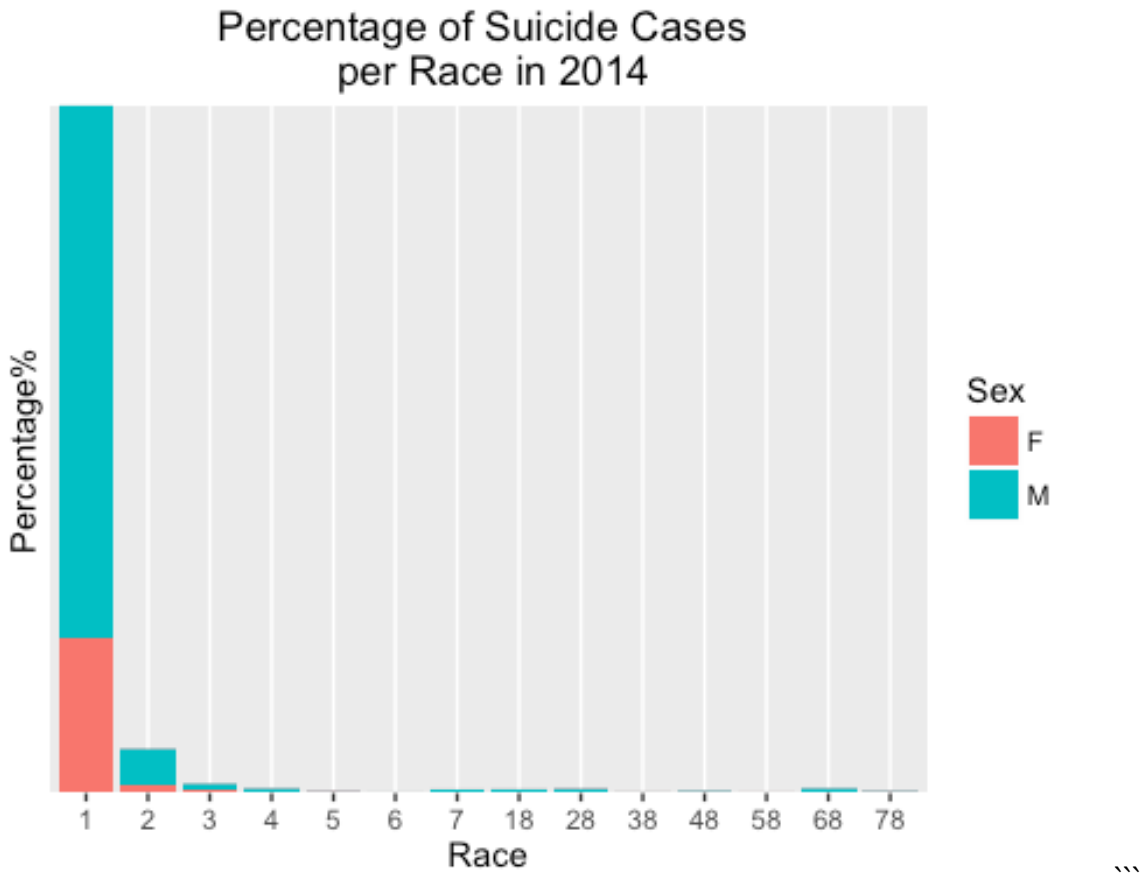

```
ggplot(ByEdu2,aes(y=Sum,x=Age))+
        geom_point(aes(color=factor(Race),size=Sum),alpha=0.5)+
        facet_grid(.~Sex)+
        scale_y_continuous(name="Number of Suicide Cases")+
        scale_colour_discrete("Race Code")+
        scale_size_continuous("Number of \n Suicide Cases",
breaks=seq(0,300,by=50))+
        #theme(legend.position="top", legend.box = "horizontal")+
        labs(title="Number of Suicide Cases \n Grouped by Age and Race ")
```




```
ByEdu<-Suicide %>%
        group_by(Education2003Revision,Sex) %>%
        summarise(Sum=n(),Percentage=100*n()/length(.$Id))

ggplot(ByEdu,aes(x=factor(Education2003Revision),y=Percentage,fill=Sex))+
        geom_bar(stat="identity")+
        # facet_grid(.~Sex) +
        scale_x_discrete(name="Education Level 2003 Rev.Code")+
        scale_y_discrete(name="Percentage%",breaks=seq(0, 40, by = 2))+
        labs(title="Percentage of Suicide Cases \n per Education Level in
2014")
```



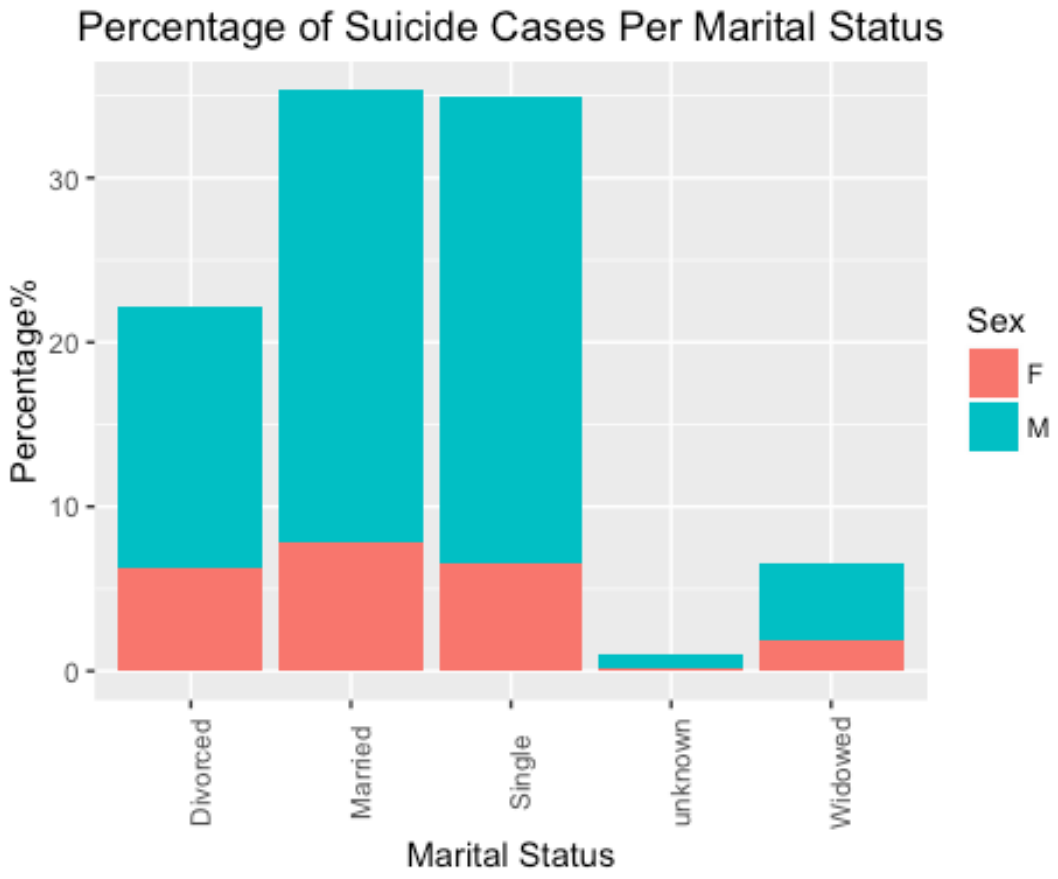

```
ByRace<-Suicide %>%
        group_by(Race, Sex, Age, Education2003Revision) %>%
        summarise(Sum=n(),Percentage=100*n()/length(.$Id))

ggplot(ByRace,aes(x=factor(Race),y=Percentage,fill=Sex))+
        geom_bar(stat="identity")+
        # facet_grid(.~Sex) +
        scale_x_discrete(name="Race")+
        scale_y_discrete(name="Percentage%",breaks=seq(0, 100, by = 2))+
        labs(title="Percentage of Suicide Cases \n per Race in 2014")
```




```

## Percentage of Suicide Cases Per Marital Status

One more factor to look at is the marital status of the people who committed suicide.

```
ByMarital<-Suicide %>%
        group_by(MaritalStatus,Sex) %>%
        summarise(Cases_Suicide=n(),Percentage_Suicide=100*n()/length(.$Id))

ggplot(ByMarital,aes(x=MaritalStatus,y=Percentage_Suicide,fill=Sex))+
        geom_bar(stat="identity")+
        # facet_grid(.~Sex)+
        scale_x_discrete(name="Marital Status",

labels=c("Divorced","Married","Single","unknown","Widowed"))+
        theme(axis.text.x=element_text(angle = 90,vjust=1))+
        scale_y_continuous(name="Percentage%")+
        labs(title="Percentage of Suicide Cases Per Marital Status")
```
```



We can see that:

- the singles represent the highest percentage in males, but not significantly higher than the married. And it is the other way around for the females.
- apart from the cases with unknown status, the widowed represent the lowest percentage for both sexes.

Again, we should be careful while interpreting this data, because we cannot just look at them without considering the percentage of each marital status in all the given death records. So we will check both as follows.

```
ByMarital2<-Suicide %>%
        group_by(MaritalStatus) %>%

summarise(Cases_Suicide=n(),Percentage_Suicide=round((Percentage_Suicide=100*
n()/length(.$Id)),2))

ByMarital_All<-DeathRec %>%
        group_by(MaritalStatus) %>%

summarise(Cases_All=n(),Percentage_All=round((Percentage_All=100*n()/length(.
$Id)),2))
```

```
MaritalData <- merge(ByMarital2,ByMarital_All) %>% merge(Marital_table,
by.x=("MaritalStatus"), by.y=("Code"))

MaritalData
```

```
##   MaritalStatus Cases_Suicide Percentage_Suicide Cases_All Percentage_All
## 1             D          9557               22.16    400959          15.24
## 2             M         15247               35.35    980016          37.25
## 3             S         15096               35.00    333043          12.66
## 4             U           434                1.01     18713           0.71
## 5             W          2798                6.49    898440          34.15
##              Description
## 1               Divorced
## 2                Married
## 3  Never married, single
## 4 Marital Status unknown
## 5                Widowed
```

The MaritalData data frame contains:

- **Cases_Suicide**:number of the people with certain marital status and sex in suicide cases
- **Percentage_Suicide**: percentage of the people with certain marital status and sex in suicide cases
- **Cases_All**: number of the people with certain marital status and sex in all death cases
- **Percentage_All**: percentage of the people with certain marital status and sex in all death cases

We can see that, the most significant value is for the singles, because the percentage of singles in all the death cases is around *12%*. On the other hand, around *35%* of the suicide cases come from singles. So this is worth highlighting.

Conclusion

Analyzing the data released by the Centers for Disease Control and Prevention about mortality rates in 2014, it is possible to investigate suicide cases and explore the underlying patterns. In this project we could see that:

- the most remarkable fact about the data is the high percentage of suicide cases between males *(around 77.3%)* compared to females *(around 22.7%)*.To be able to draw conclusions, it is recommended to compare the data over years to see whether this is a patterns to be studied. It is also worth consideration to Look into the data in the light of psychological studie and reports.
- the mean age for commiting suicide for both males and females is around 47.

- a high percentage of the cases came from a low/mid-level education, which is justified because it is proportional to the percentage of this level of education in the main data set and expectedly in the society.
- a significant percentage of cases are committed by singles *(around 35%)*, although in all the death cases singles represent around *(around 12%)*. So we can consider it a significant value to be investigated further.