Exploratory Data Project

##Introduction: The data set I chose was the Police Shootings data set. This data set gives us information on people that have been killed in Police shootings in 2015. There are multiple variables in this data set, but I will be focusing on the signs of mental illness and race of those killed and if there is a correlation between the two variables.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readxl)
police_shootings<-read_excel("/Users/layladreyer/Math 130/data/fatal-police-shootings-data.xlsx", sheet=1, col_na
mes=TRUE)
```

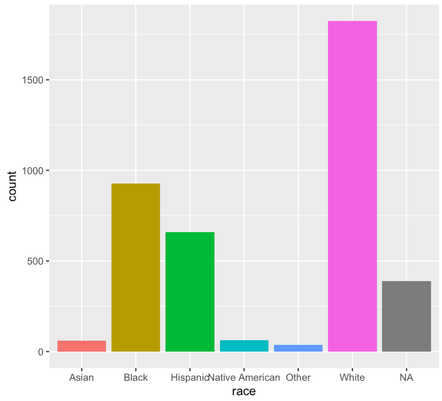
Univariate Analysis of variables

Race & Signs of mental illness

```
police_shootings$race[police_shootings$race%in%c('A')]<-"Asian"
police_shootings$race[police_shootings$race%in%c('B')]<-"Black"
police_shootings$race[police_shootings$race%in%c('H')]<-"Hispanic"
police_shootings$race[police_shootings$race%in%c('W')]<-"White"
police_shootings$race[police_shootings$race%in%c('N')]<-"Native American"
police_shootings$race[police_shootings$race%in%c('O')]<-"Other"
table(police_shootings$race)
```

```
##
##          Asian          Black       Hispanic Native American          Other
##             61            927            659             62             37
##          White
##           1825
```

```
ggplot(police_shootings,aes(x=race,fill=race))+geom_bar()
```

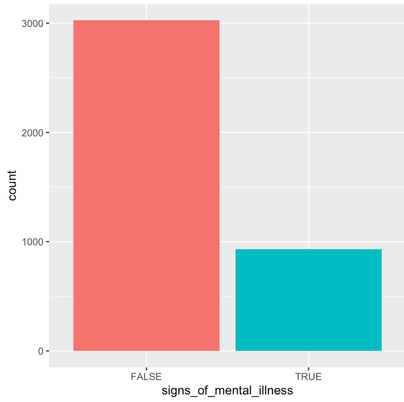


The graph and table shows the count of deaths per race. It clearly shows that people that identify as White were killed the most, followed by Black, and then Hispanic.

```
police_shootings$signs_of_mental_illness[police_shootings$signs_of_mental_illness%in%c('T')]<-"TRUE"
police_shootings$signs_of_mental_illness[police_shootings$signs_of_mental_illness%in%c('F')]<-"FALSE"
table(police_shootings$signs_of_mental_illness)
```

```
##
## FALSE  TRUE
##  3028   932
```

```
ggplot(police_shootings,aes(x=signs_of_mental_illness,fill=signs_of_mental_illness))+geom_bar()
```



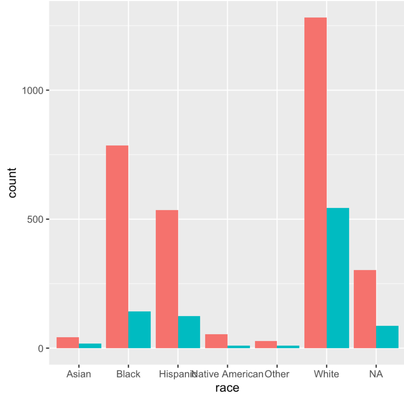
I expected that the data set would show that the deaths did have signs of mental illness but as you can see in the table and the graph that was not true. There was a difference of over 2,000 people that did not show signs of mental illness compared to those that did.

##Bivariate Analysis

```
table(police_shootings$race,police_shootings$signs_of_mental_illness)
```

```
##
##                   FALSE TRUE
##   Asian              43   18
##   Black             785  142
##   Hispanic          535  124
##   Native American    53    9
##   Other              28    9
##   White            1282  543
```

```
ggplot(police_shootings, aes(x=race,fill=signs_of_mental_illness))+geom_bar(position="dodge")
```



Based on the Graph all races show that the no signs of mental illness is still larger than those with signs of mental illness. Now that our two variables are shown on one table and graph we can still see that White, Black, and Hispanic lead the amount of deaths. It doesn't look like there is a clear correlation between the two variables based on the graph and table.