Analysis of mass shootings in the USA between 1966 and 2017

Introduction

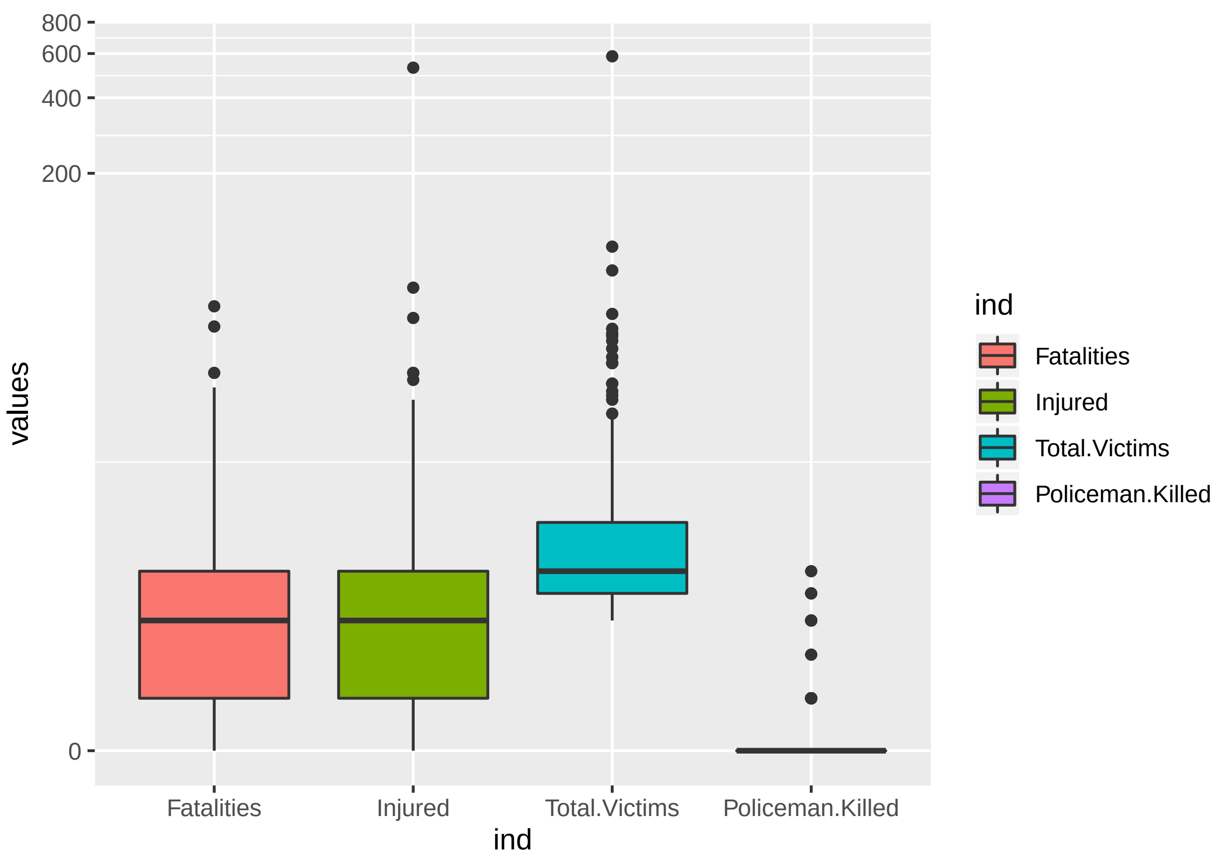
Ce document est une présentation de plusieurs analyses faites sur un jeu de données présentant les fusillades de masses de 1966 à 2017. Aucune conclusion définitive ne pourra être tirer de ces analyses car le jeu de données ne compte qu'un nombre limité de variables et que d'autres variables non présentes peuvent influer sur les corrélations relevées.

Cependant ce document exposent des pistes d'études relevantes à explorer pour mieux comprendre les causes des fusillades.

Graphes

```r
ggplot(stack(shootings.quantitative[,-c(5,6)]), aes(x = ind, y = values, fill = ind)) +
  geom_boxplot() + scale_y_continuous(trans='pseudo_log')
```

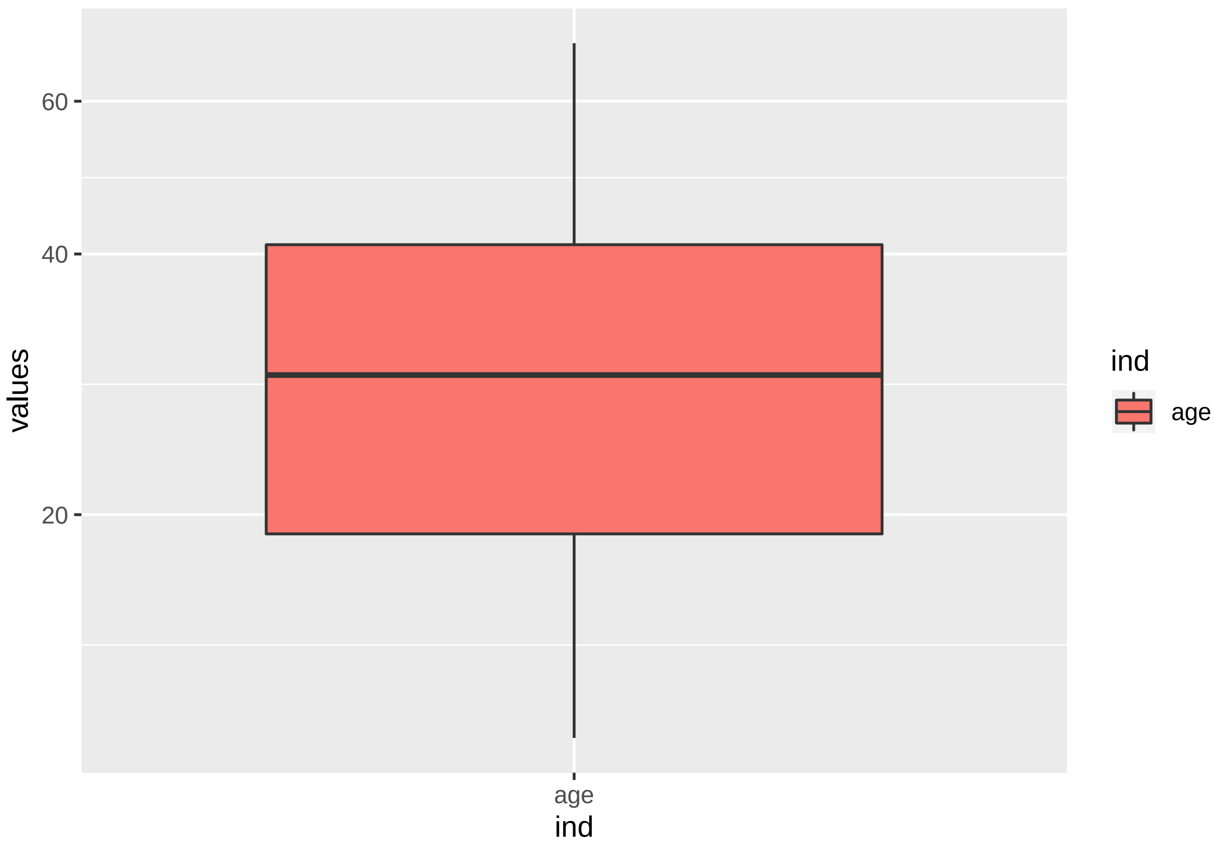
```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```



On fait le graphique pour Age/Age2 séparéments car il y a plus de valeurs

```
ggplot(stack(shootings.ages), aes(x = ind, y = values, fill = ind)) +
  geom_boxplot() + scale_y_continuous(trans='pseudo_log')
```

```
## Warning: Removed 449 rows containing non-finite values (stat_boxplot).
```



Statistiques sur le nombre de mort

```
summary(shootings.quantitative$Fatalities)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   4.379   5.000  59.000
```

```
CI(shootings.quantitative$Fatalities, ci=0.95)
```

```
##    upper     mean    lower
## 5.025635 4.378981 3.732327
```

```
fatalities_outliers <- boxplot.stats(shootings.quantitative$Fatalities)$out
fatalities_outliers_row <- which(shootings.quantitative$Fatalities %in% c(fatalities_outliers))

shootings[fatalities_outliers_row,]
```

```
## # A tibble: 17 x 26
##       S. Title Location State Incident.Area Open.Close.Loca~ Target Cause
##    <dbl> <chr> <chr>    <chr> <chr>         <chr>            <chr>  <chr>
##  1     1 Texa~ Sutherl~ Texas place of wor~ close            random <NA>
##  2     4 Las ~ Las Veg~ Neva~ event         open             random <NA>
##  3    14 Orla~ Orlando  Flor~ place of ent~ close            random <NA>
```

```
##  4    81 San ~ San Ber~ Cali~ event         close            random terr~
##  5   164 Wash~ Washing~ Dist~ NA            close            random terr~
##  6   177 Sand~ Newtown  Conn~ NA            <NA>             famil~ terr~
##  7   183 Auro~ Aurora   Colo~ place of ent~ close            random terr~
##  8   202 Fort~ Fort Ho~ Texas millitary fa~ close            random terr~
##  9   203 Bing~ Bingham~ New ~ association   close            random terr~
## 10   221 Virg~ Blacksb~ Virg~ university    close            random terr~
## 11   250 Colu~ Littlet~ Colo~ high school   close            stude~ terr~
## 12   288 Luby~ Killeen  Texas restaurant    open             random unem~
## 13   291 GMAC~ Jackson~ Flor~ NA            close            random psyc~
## 14   303 Post~ Edmond   Okla~ administrati~ close            cowor~ <NA>
## 15   307 McDo~ San Ysi~ Cali~ restaurant    close            random psyc~
## 16   311 Wah ~ Seattle  Wash~ place of ent~ close            random terr~
## 17   323 Univ~ Austin   Texas university    close            random terr~
## # ... with 18 more variables: Summary <chr>, Fatalities <dbl>, Injured <dbl>,
## #   Total.victims <dbl>, Policeman.Killed <dbl>, Age <dbl>, Weapon.Type <chr>,
## #   Mental.Health.Issues <chr>, Race <chr>, Gender <chr>, Latitude <dbl>,
## #   Longitude <dbl>, Age2 <dbl>, AverageAge <dbl>, Day <chr>, Month <chr>,
## #   Year <dbl>, Ten.Casualities.Min <dbl>
```

Distribution

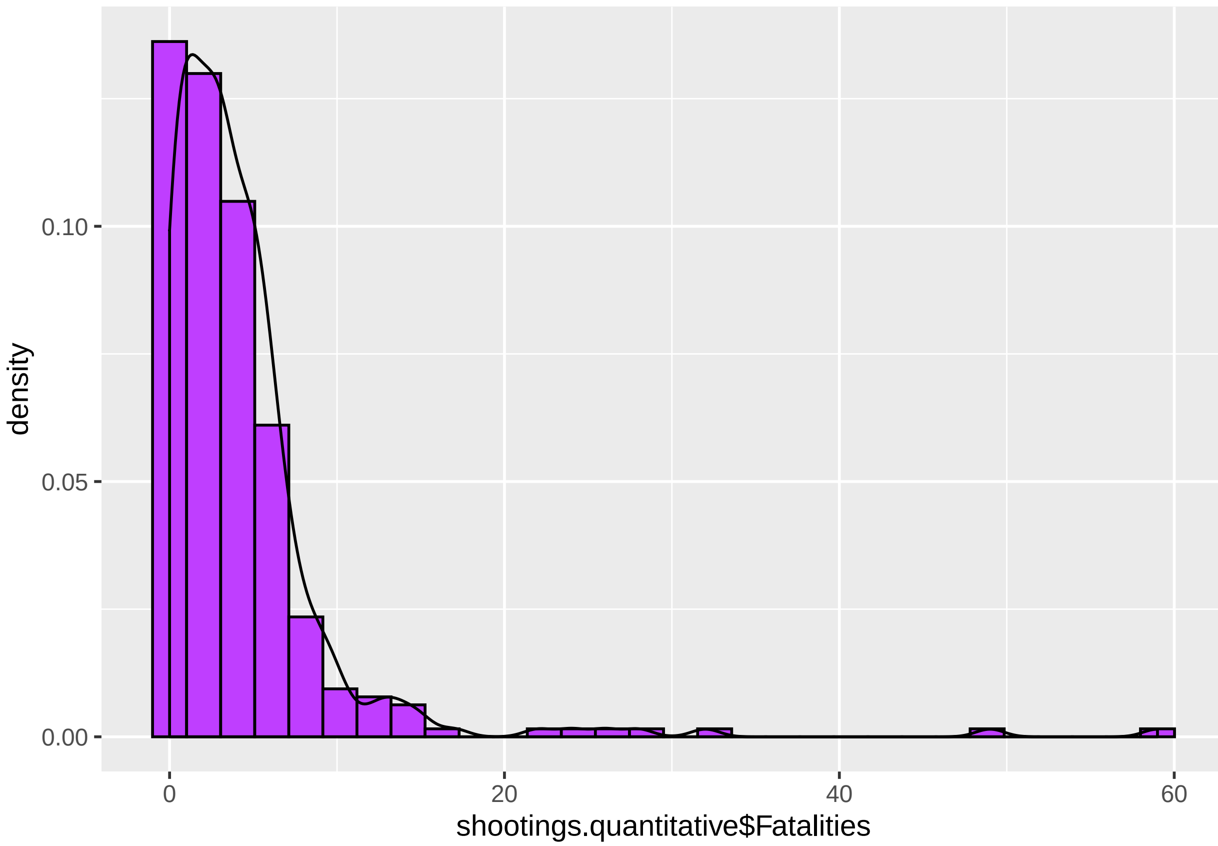
```r
shapiro.test(shootings.quantitative$Fatalities)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shootings.quantitative$Fatalities
## W = 0.57015, p-value < 2.2e-16
```

On peux assumer que la loi n'est pas normale car p-value < 0.05

```r
ggplot(shootings.quantitative, aes(x=shootings.quantitative$Fatalities)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill="darkorchid1") +
  geom_density()
```

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



On remarque que la distribution ressemble à une loi exponentielle, pour le tester on va utiliser un test "goodness of fit"

```
## Warning in ks.test(shootings.quantitative$Fatalities, "pexp", fit$estimate):
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  shootings.quantitative$Fatalities
## D = 0.13057, p-value = 4.478e-05
## alternative hypothesis: two-sided
```

Statistiques sur le nombre de blessés

```
summary(shootings.quantitative$Injured)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    3.00    6.15    5.00  527.00
```

```
CI(shootings.quantitative$Injured, ci=0.95)
```

```
##    upper     mean    lower
## 9.512986 6.149682 2.786377
```

```
injured_outliers <- boxplot.stats(shootings.quantitative$Injured)$out
injured_outliers_row <- which(shootings.quantitative$Injured %in% c(injured_outliers))

shootings[injured_outliers_row,]
```

```
## # A tibble: 27 x 26
##        S. Title Location State Incident.Area Open.Close.Loca~ Target Cause
##     <dbl> <chr> <chr>    <chr> <chr>         <chr>            <chr>  <chr>
##  1      1 Texa~ Sutherl~ Texas place of wor~ close            random <NA>
##  2      4 Las ~ Las Veg~ Neva~ event         open             random <NA>
##  3     14 Orla~ Orlando  Flor~ place of ent~ close            random <NA>
##  4     52 Exce~ Hesston  Kans~ company       close            random <NA>
##  5     81 San ~ San Ber~ Cali~ event         close            random terr~
##  6    155 Isla~ Santa B~ Cali~ NA            <NA>             random psyc~
##  7    157 Fort~ Fort Ho~ Texas millitary fa~ open             polic~ psyc~
##  8    180 The ~ Miami    Flor~ place of ent~ close            random terr~
##  9    183 Auro~ Aurora   Colo~ place of ent~ close            random terr~
## 10    197 Tucs~ Tucson   Ariz~ NA            open             congr~ terr~
## # ... with 17 more rows, and 18 more variables: Summary <chr>,
## #   Fatalities <dbl>, Injured <dbl>, Total.victims <dbl>,
## #   Policeman.Killed <dbl>, Age <dbl>, Weapon.Type <chr>,
## #   Mental.Health.Issues <chr>, Race <chr>, Gender <chr>, Latitude <dbl>,
## #   Longitude <dbl>, Age2 <dbl>, AverageAge <dbl>, Day <chr>, Month <chr>,
## #   Year <dbl>, Ten.Casualities.Min <dbl>
```

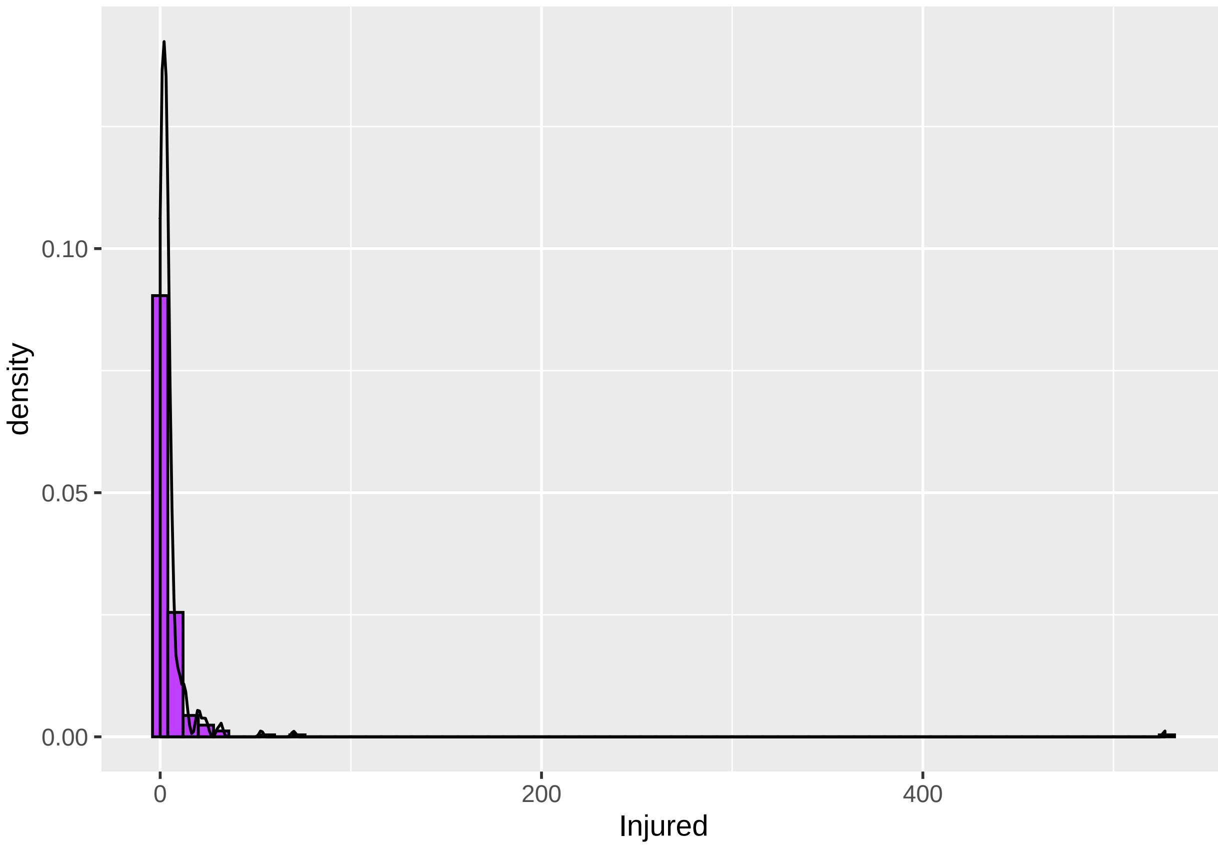
Distribution

```r
shapiro.test(shootings.quantitative$Injured)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shootings.quantitative$Injured
## W = 0.11136, p-value < 2.2e-16
```

On peux assumer que la loi n'est pas normale car p-value < 0.05

```r
ggplot(shootings.quantitative, aes(x=Injured)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill="darkorchid1", binwidth = 8) +
  geom_density()
```



On remarque que la distribution ressemble à une loi exponentielle, pour le tester on va utiliser un test "goodness of fit"

```
## Warning in ks.test(shootings.quantitative$Injured, "pexp", fit$estimate): ties
## should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  shootings.quantitative$Injured
## D = 0.24475, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Statistiques sur le nombre total de victimes

```r
summary(shootings.quantitative$Total.Victims)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    4.00    5.00   10.18    8.00  585.00
```

```r
CI(shootings.quantitative$Total.Victims, ci=0.95)
```

```
##     upper      mean     lower
## 13.971588 10.184713  6.397839
```

```r
victims_outliers <- boxplot.stats(shootings.quantitative$Total.Victims)$out
victims_outliers_row <- which(shootings.quantitative$Total.Victims %in% c(victims_outliers))
shootings[victims_outliers_row,]
```

```
## # A tibble: 36 x 26
##       S. Title Location State Incident.Area Open.Close.Loca~ Target Cause
##    <dbl> <chr> <chr>    <chr> <chr>         <chr>            <chr>  <chr>
##  1     1 Texa~ Sutherl~ Texas place of wor~ close            random <NA>
##  2     4 Las ~ Las Veg~ Neva~ event         open             random <NA>
##  3    13 Dall~ Dallas   Texas protest       open             police raci~
##  4    14 Orla~ Orlando  Flor~ place of ent~ close            random <NA>
##  5    52 Exce~ Hesston  Kans~ company       close            random <NA>
##  6    81 San ~ San Ber~ Cali~ event         close            random terr~
##  7    93 Umpq~ Roseburg Oreg~ university    close            stude~ terr~
##  8   155 Isla~ Santa B~ Cali~ NA            <NA>             random psyc~
##  9   157 Fort~ Fort Ho~ Texas millitary fa~ open             polic~ psyc~
## 10   164 Wash~ Washing~ Dist~ NA            close            random terr~
## # ... with 26 more rows, and 18 more variables: Summary <chr>,
## #   Fatalities <dbl>, Injured <dbl>, Total.victims <dbl>,
## #   Policeman.Killed <dbl>, Age <dbl>, Weapon.Type <chr>,
## #   Mental.Health.Issues <chr>, Race <chr>, Gender <chr>, Latitude <dbl>,
## #   Longitude <dbl>, Age2 <dbl>, AverageAge <dbl>, Day <chr>, Month <chr>,
## #   Year <dbl>, Ten.Casualities.Min <dbl>
```

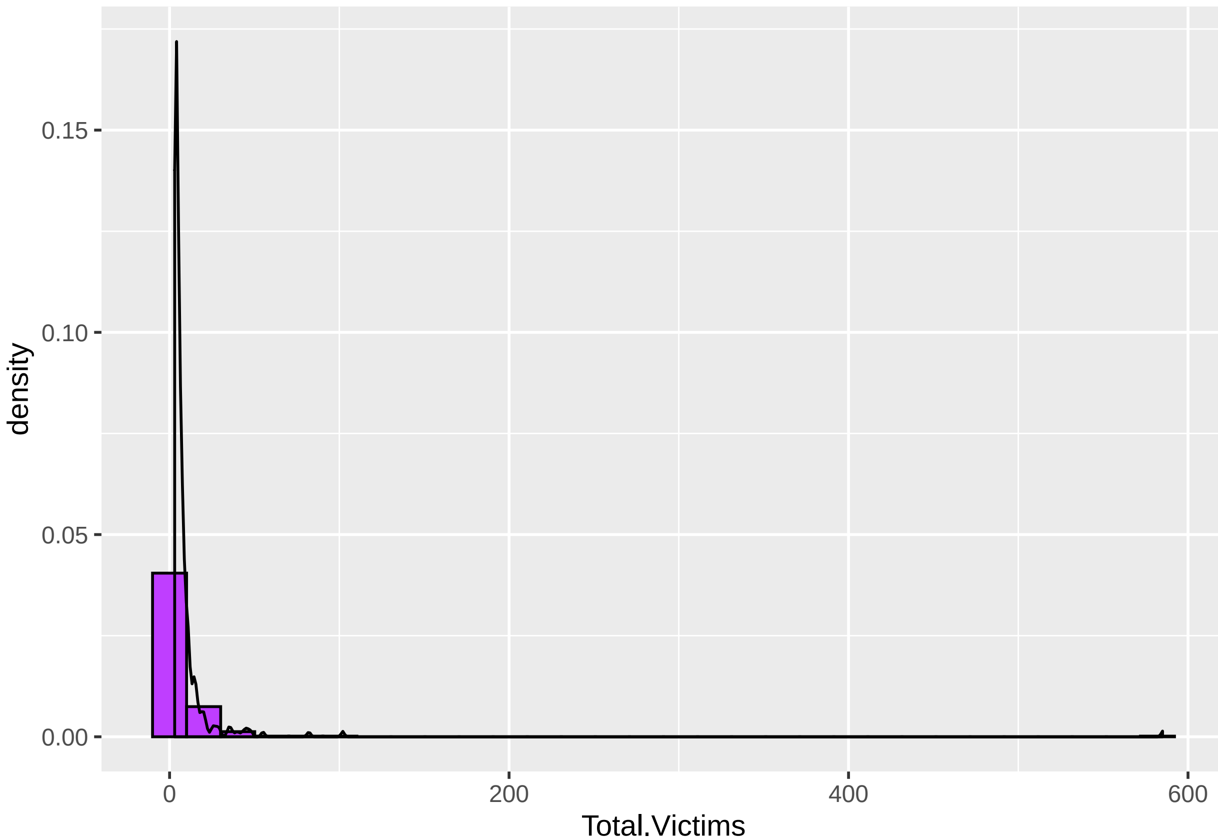
```r
shapiro.test(shootings.quantitative$Total.Victims)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shootings.quantitative$Total.Victims
## W = 0.13555, p-value < 2.2e-16
```

On peux assumer que la loi n'est pas normale car p-value <= 0.05

```r
ggplot(shootings.quantitative, aes(x=Total.Victims)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill="darkorchid1") +
  geom_density()
```

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



On remarque que la distribution ressemble à une loi exponentielle, pour le tester on va utiliser un test “goodness of fit”

```
## Warning in ks.test(shootings.quantitative$Total.Victims, "pexp", fit$estimate):
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  shootings.quantitative$Total.Victims
## D = 0.25514, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Statistiques sur le nombre de policier tués

```
summary(shootings.quantitative$Policeman.Killed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.1104  0.0000  5.0000       6
```

```
CI(shootings.quantitative$Policeman.Killed, ci=0.95)
```

```
## upper  mean lower
##    NA    NA    NA
```

```
policemans_outliers <- boxplot.stats(shootings.quantitative$Policeman.Killed)$out
policemans_outliers_row <- which(shootings.quantitative$Policeman.Killed %in% c(policemans_outliers))
shootings[policemans_outliers_row,]
```

```
## # A tibble: 17 x 26
```

```
##       S. Title Location State Incident.Area Open.Close.Loca~ Target Cause
##    <dbl> <chr> <chr>    <chr> <chr>         <chr>            <chr>  <chr>
##  1     4 Las ~ Las Veg~ Neva~ event         open             random <NA>
##  2     8 Rura~ Kirkers~ Ohio  hospital      close            cowor~ <NA>
##  3    12 Bato~ Baton R~ Loui~ NA            open             police <NA>
##  4    13 Dall~ Dallas   Texas protest       open             police raci~
##  5    50 Wood~ Woodbri~ Virg~ home          open             random <NA>
##  6    59 Iuka~ Iuka     Miss~ home          open             polic~ dome~
##  7    83 Plan~ Colorad~ Colo~ street        close            random <NA>
##  8   126 Litt~ Little ~ New ~ NA            close            polic~ psyc~
##  9   153 Nell~ Las Veg~ Neva~ restaurant;s~ close            polic~ psyc~
## 10   157 Fort~ Fort Ho~ Texas millitary fa~ open             polic~ psyc~
## 11   162 Los ~ Los Ang~ Cali~ airport       open             tsa o~ anger
## 12   174 Los ~ Irvine   Cali~ NA            <NA>             cowor~ anger
## 13   201 Park~ Lakewood Wash~ restaurant    close            polic~ reve~
## 14   257 Calt~ Orange   Cali~ company       close            ex-co~ unem~
## 15   260 R.E.~ Aiken    Sout~ company       <NA>             ex-co~ unem~
## 16   297 Come~ Chicago  Illi~ NA            <NA>             random terr~
## 17   320 New ~ New Orl~ Loui~ NA            close            random psyc~
## # ... with 18 more variables: Summary <chr>, Fatalities <dbl>, Injured <dbl>,
## #   Total.victims <dbl>, Policeman.Killed <dbl>, Age <dbl>, Weapon.Type <chr>,
## #   Mental.Health.Issues <chr>, Race <chr>, Gender <chr>, Latitude <dbl>,
## #   Longitude <dbl>, Age2 <dbl>, AverageAge <dbl>, Day <chr>, Month <chr>,
## #   Year <dbl>, Ten.Casualities.Min <dbl>
```

```r
shapiro.test(shootings.quantitative$Policeman.Killed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shootings.quantitative$Policeman.Killed
## W = 0.20313, p-value < 2.2e-16
```

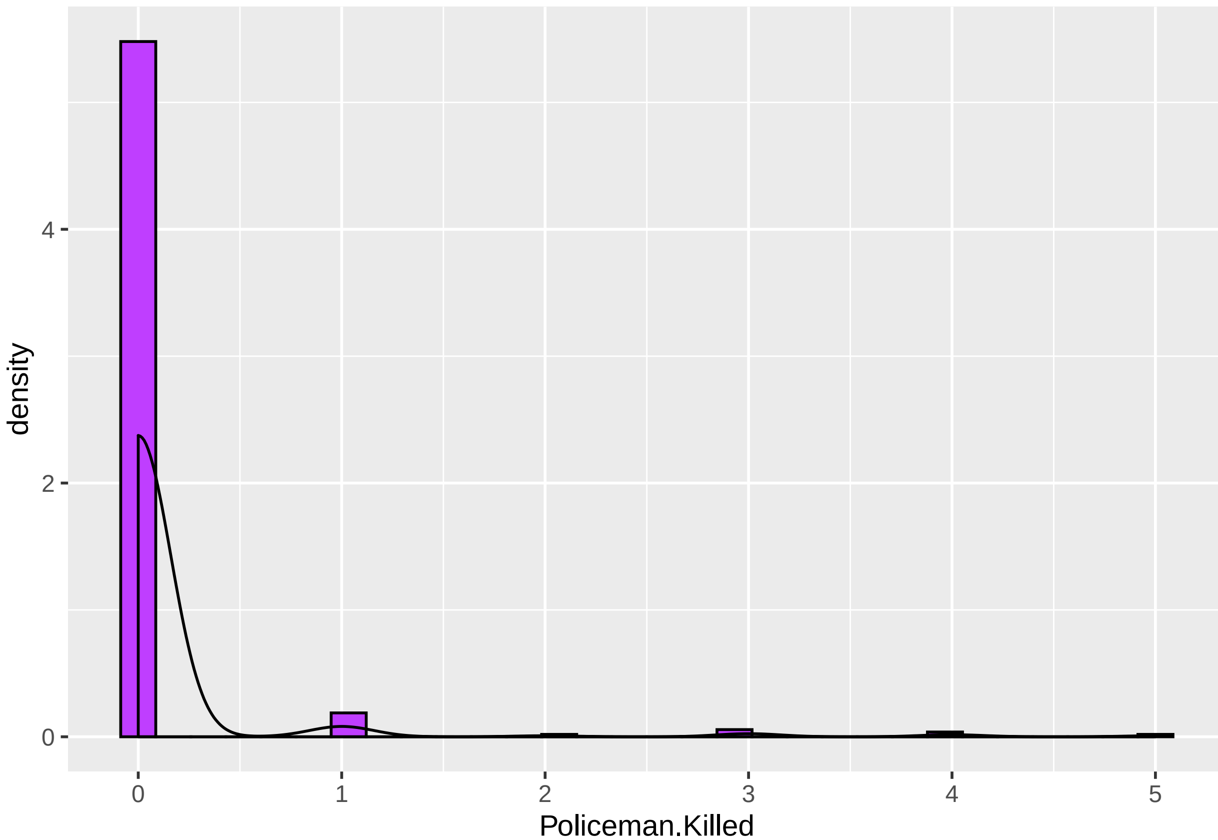
On peux assumer que la loi n'est pas normale car p-value <= 0.05

```r
ggplot(shootings.quantitative, aes(x=Policeman.Killed)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill="darkorchid1") +
  geom_density()
```

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

```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 6 rows containing non-finite values (stat_density).
```



Statistiques sur l'age

```
summary(shootings.ages$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   11.00   19.00   29.00   31.41   41.00   70.00     449
```

```
CI(shootings.ages$age, ci=0.95)
```

```
## upper  mean lower
##    NA    NA    NA
```

```
ages_outliers <- boxplot.stats(shootings.ages$age)$out
ages_outliers_row <- which(shootings.ages %in% c(ages_outliers))
shootings[ages_outliers_row,]
```

```
## # A tibble: 0 x 26
## # ... with 26 variables: S. <dbl>, Title <chr>, Location <chr>, State <chr>,
## #   Incident.Area <chr>, Open.Close.Location <chr>, Target <chr>, Cause <chr>,
## #   Summary <chr>, Fatalities <dbl>, Injured <dbl>, Total.victims <dbl>,
## #   Policeman.Killed <dbl>, Age <dbl>, Weapon.Type <chr>,
## #   Mental.Health.Issues <chr>, Race <chr>, Gender <chr>, Latitude <dbl>,
## #   Longitude <dbl>, Age2 <dbl>, AverageAge <dbl>, Day <chr>, Month <chr>,
## #   Year <dbl>, Ten.Casualities.Min <dbl>
```

Age n'a pas d'outliers

```
shapiro.test(shootings.ages$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shootings.ages$age
## W = 0.94885, p-value = 4.726e-06
```

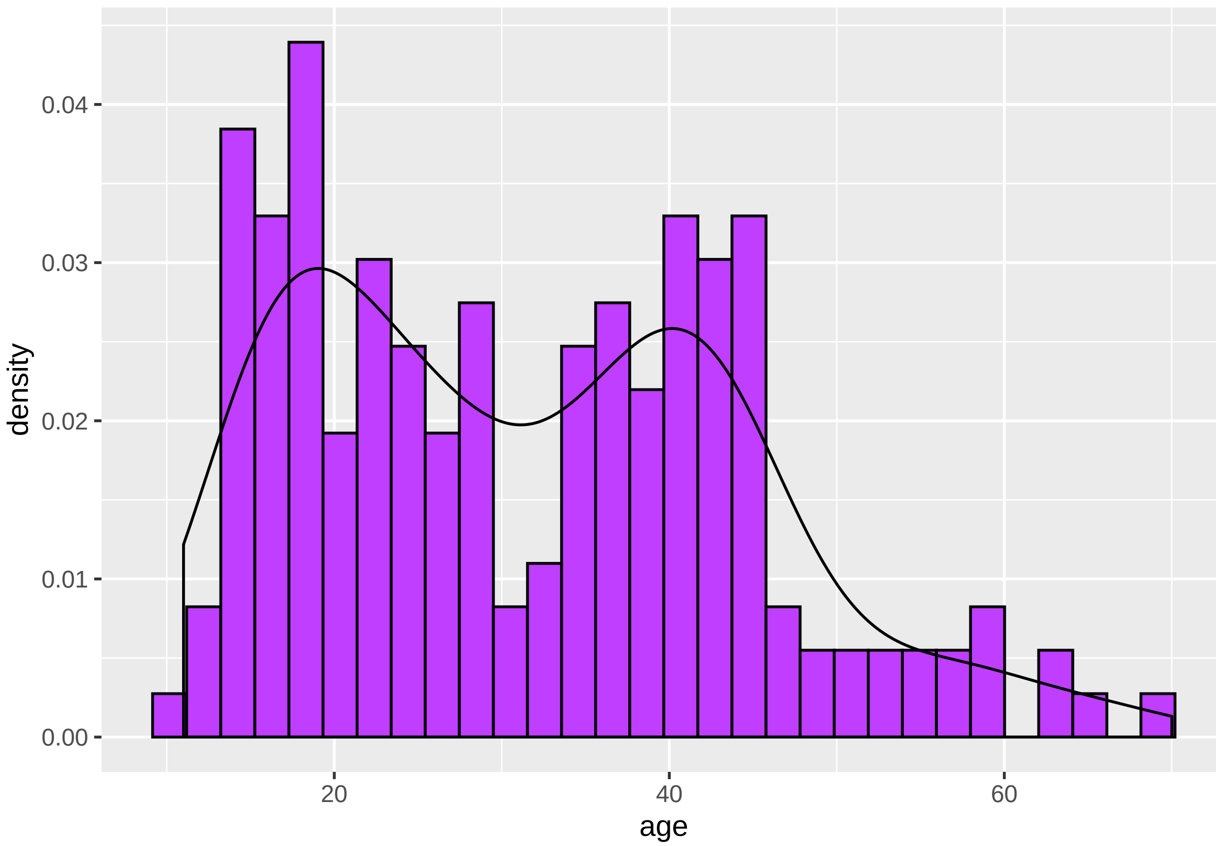
On peux assumer que la loi n'est pas normale car p-value <= 0.05

```r
ggplot(shootings.ages, aes(x=age)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill="darkorchid1") +
  geom_density()
```

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

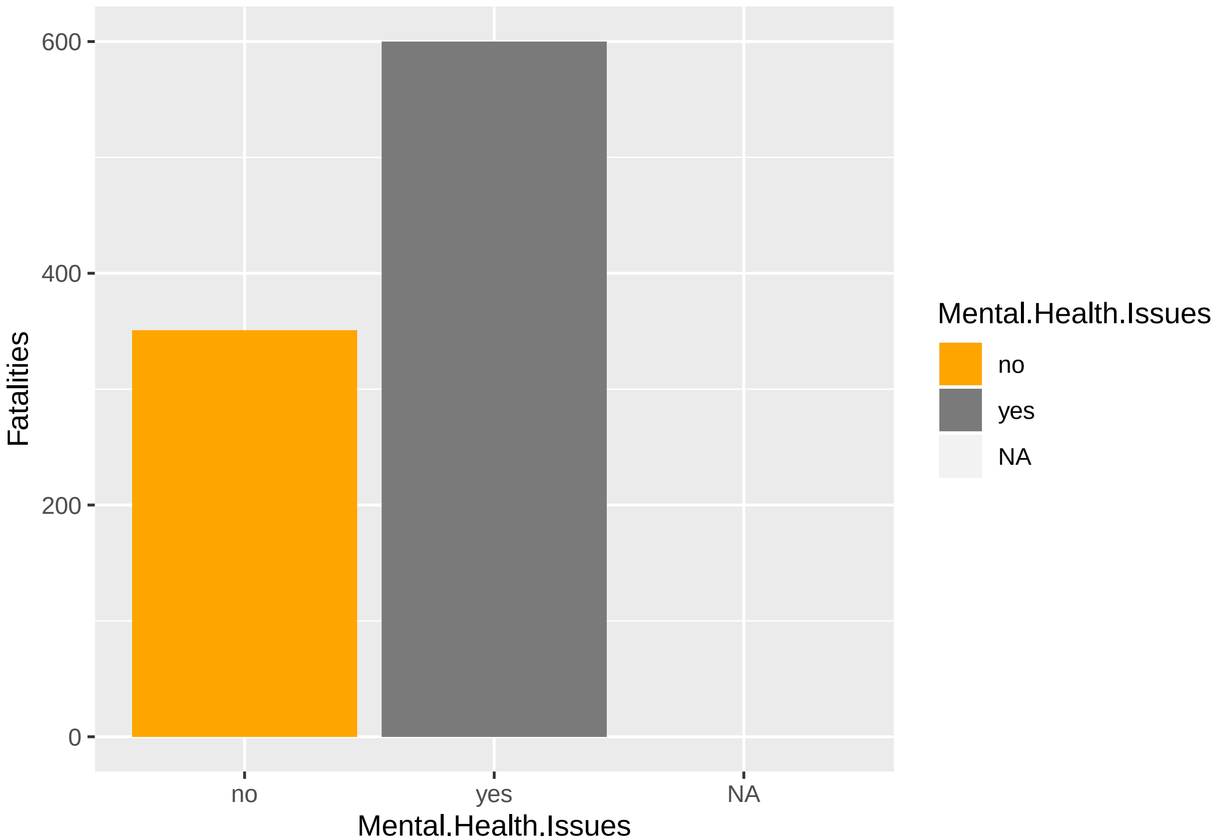
```
## Warning: Removed 449 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 449 rows containing non-finite values (stat_density).
```

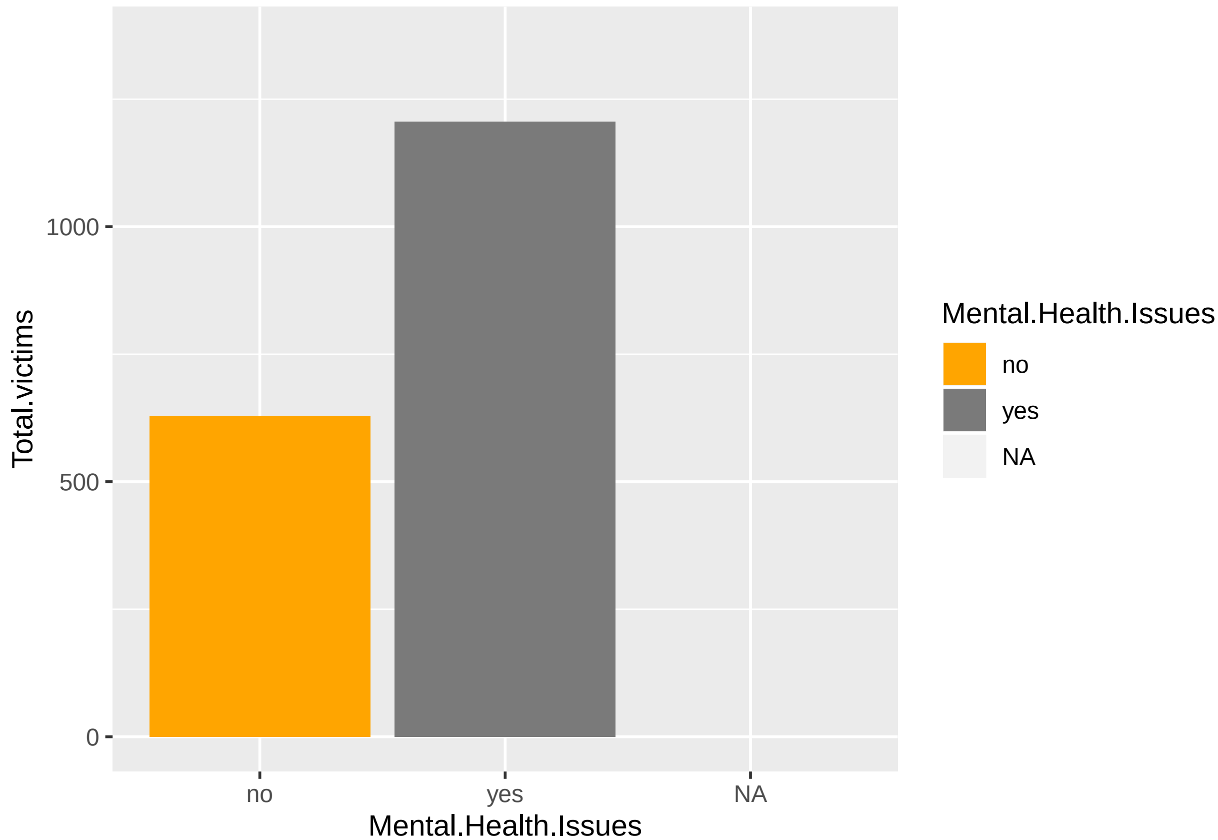


Associations with caterorical variables

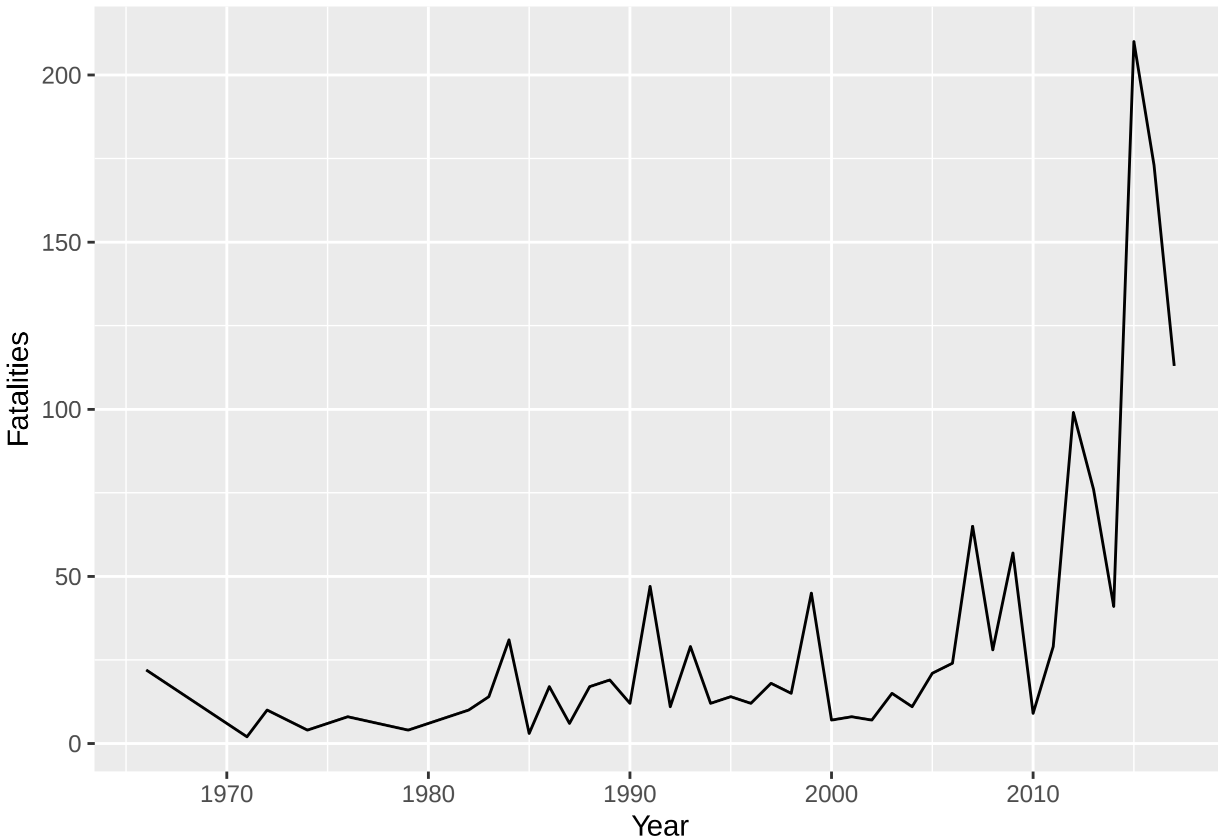
```r
ggplot(data=shootings, aes(x=Mental.Health.Issues, y=Fatalities, fill=Mental.Health.Issues)) +
  scale_fill_manual(values=c("orange1", "grey48", "firebrick1")) +
  geom_bar(stat="identity")
```



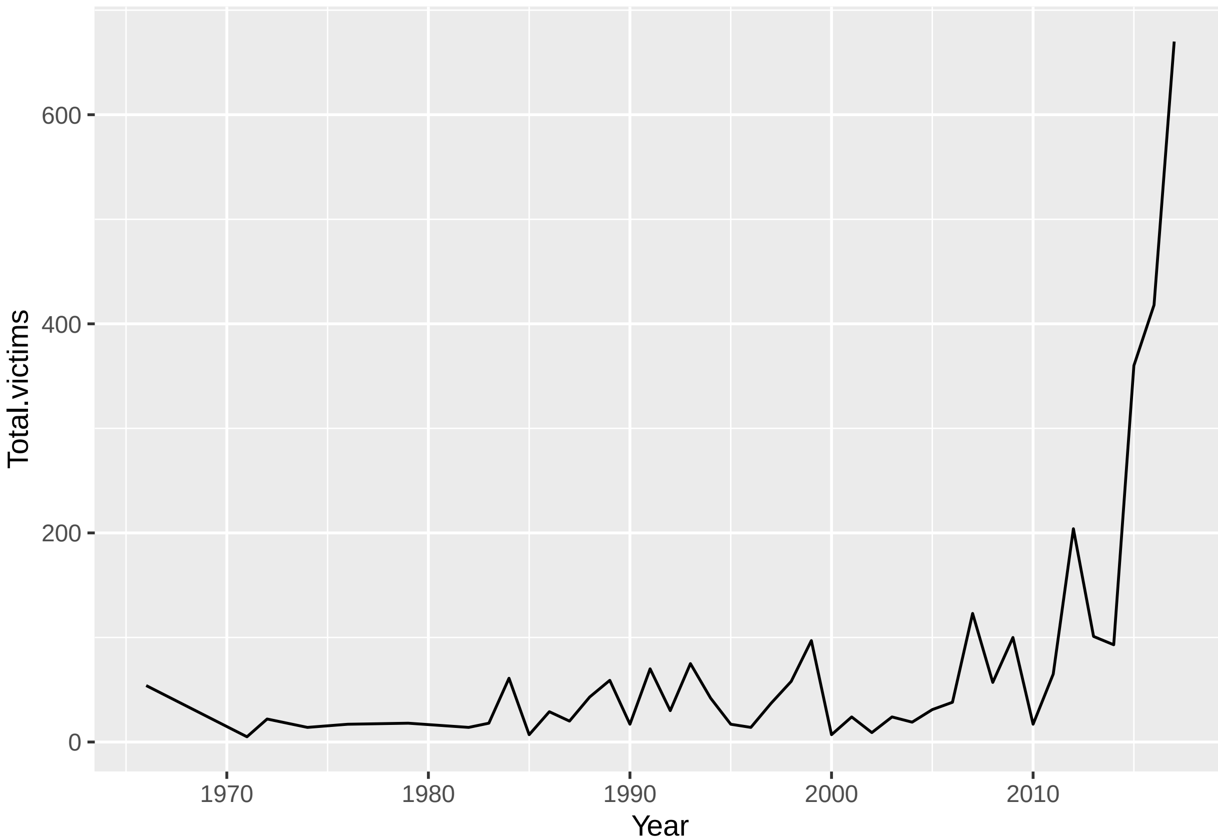

```r
ggplot(data=shootings, aes(x=Mental.Health.Issues, y=Total.victims, fill=Mental.Health.Issues)) +
  scale_fill_manual(values=c("orange1", "grey48", "firebrick1")) +
  geom_bar(stat="identity")
```



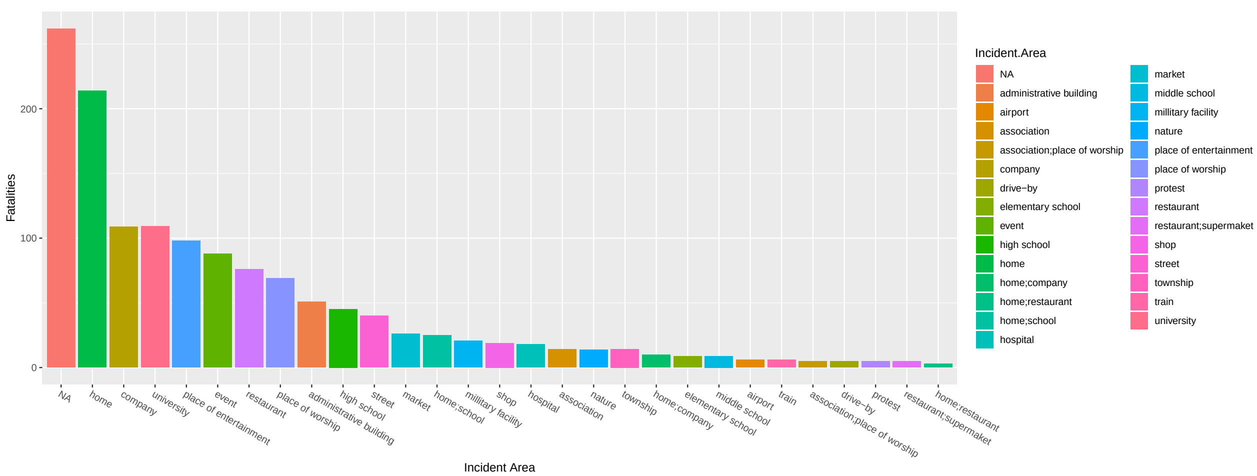

```
ggplot(shootings, aes(x=Year, y=Fatalities)) +
  stat_summary(fun.y = sum, geom="line")
```



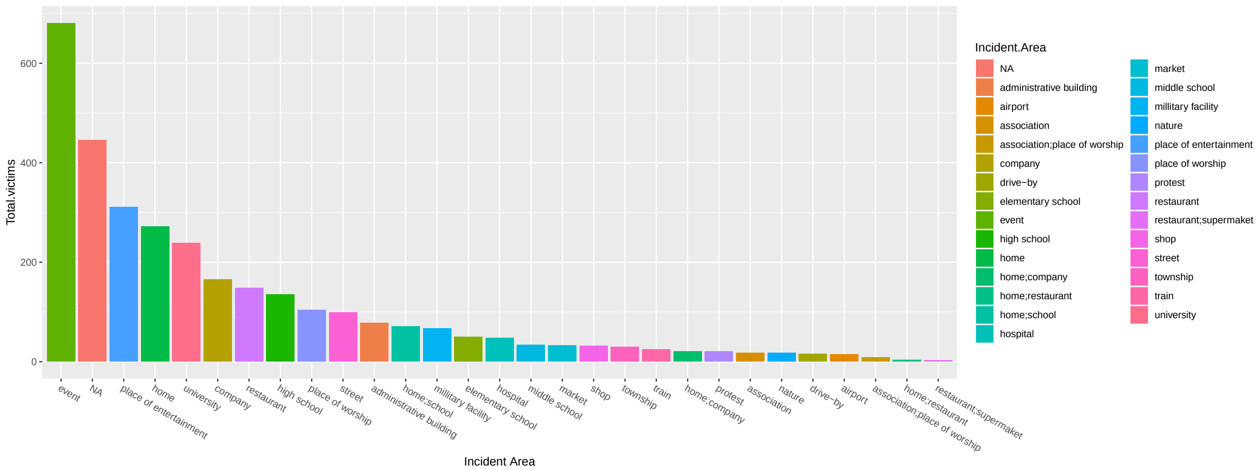

```r
ggplot(shootings, aes(x=Year, y=Total.victims)) +
  stat_summary(fun.y = sum, geom="line")
```




```
ggplot(data=shootings, aes(x=reorder(Incident.Area, -Fatalities, function(x){ sum(x) }), y=Fatalities, f
  theme(axis.text.x=element_text(angle=-30,hjust=0)) +
  xlab("Incident Area") +
  geom_bar(stat="identity")
```



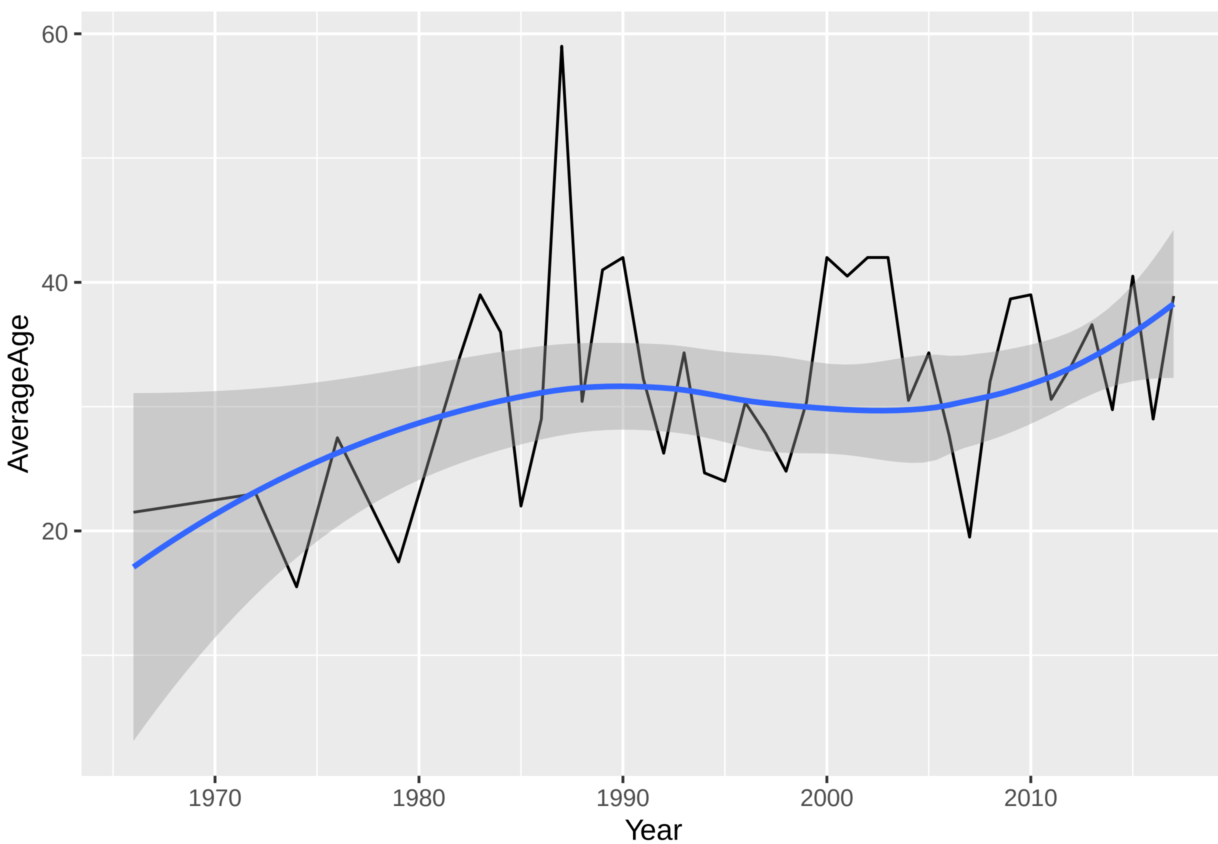

```
ggplot(data=shootings, aes(x=reorder(Incident.Area, -Total.victims, function(x){ sum(x) }), y=Total.vict
  theme(axis.text.x=element_text(angle=-30,hjust=0)) +
  xlab("Incident Area") +
  geom_bar(stat="identity")
```



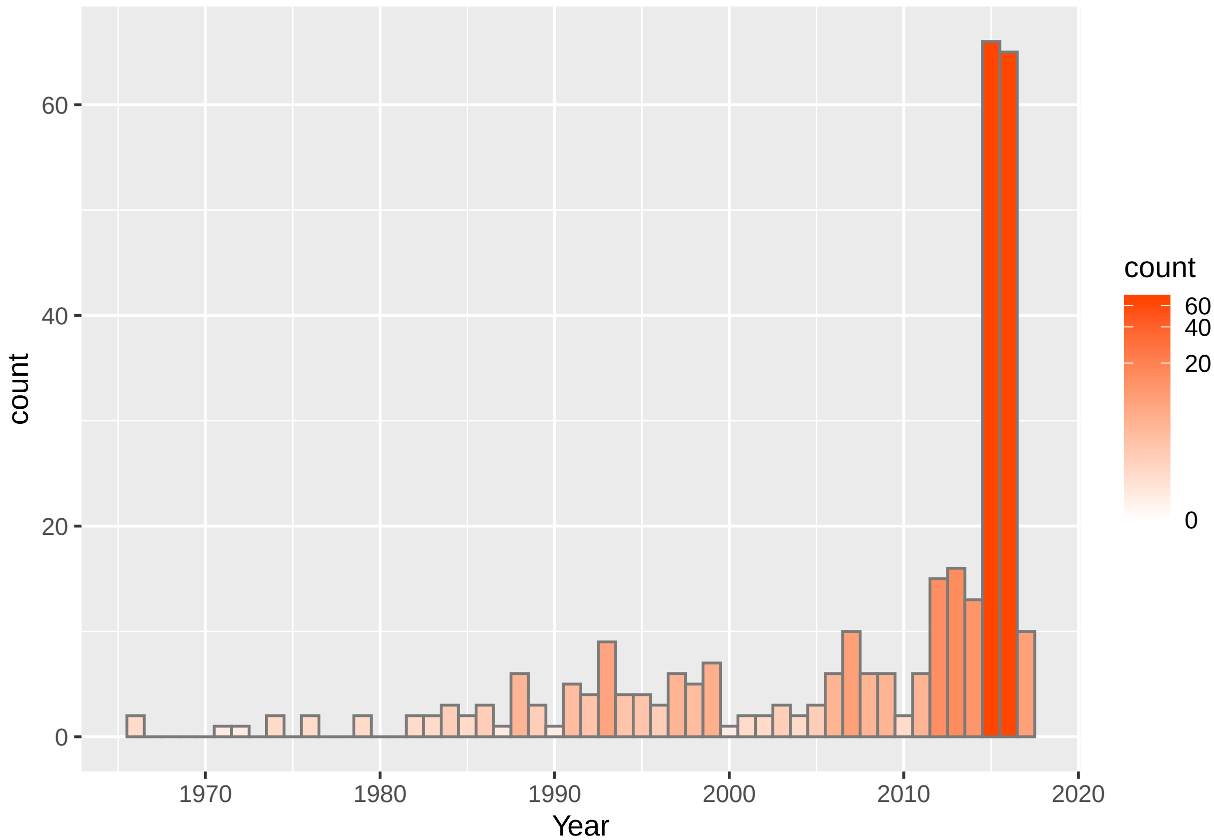
Age Year

```r
ggplot(shootings, aes(x=Year, y=AverageAge)) +
  stat_summary(fun.y = mean, geom="line", na.rm = TRUE) +
  geom_smooth(na.rm = TRUE)
```

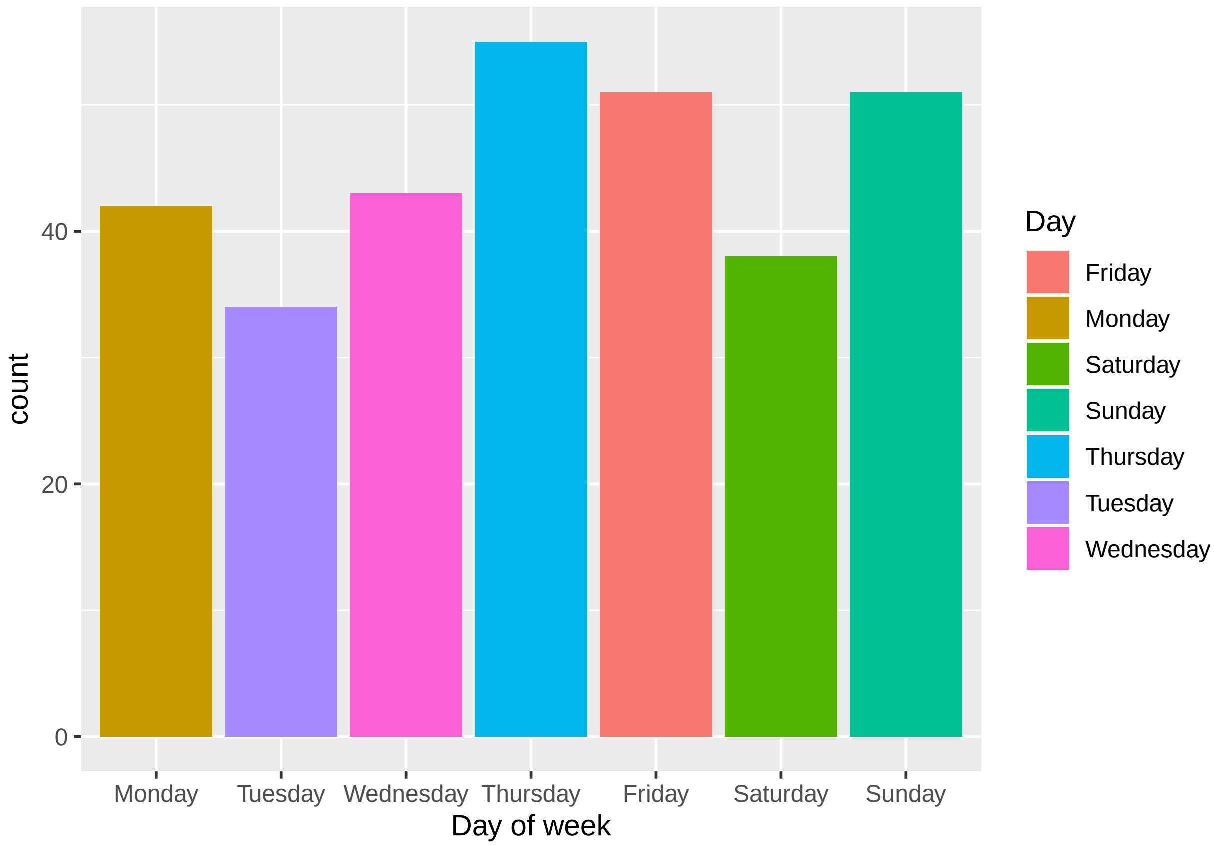
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



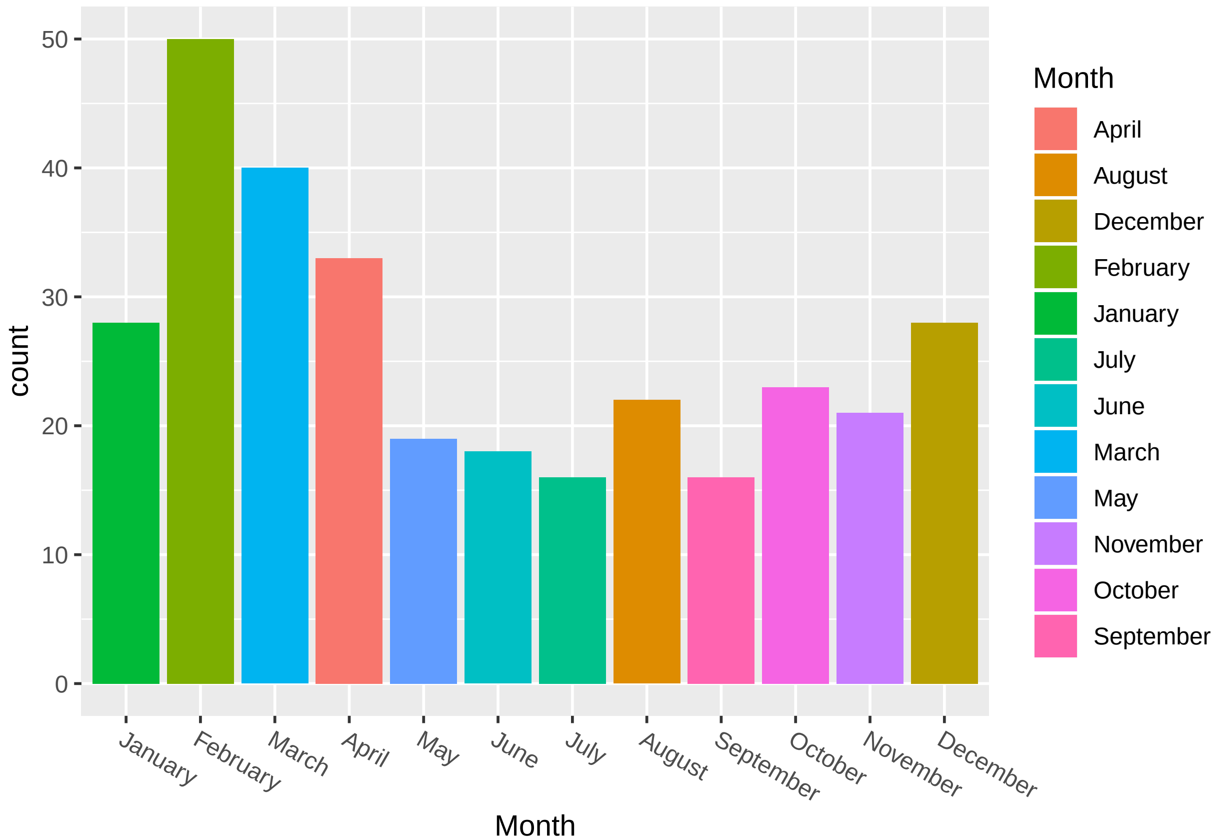

```r
ggplot(shootings, aes(x=Year, fill=..count..)) +
  geom_histogram(aes(y = stat(count)), colour="grey48", binwidth = 1) +
  scale_fill_gradient(low='white', high='orangered', trans = "pseudo_log")
```




```r
day_order <- c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday')
ggplot(data=shootings, aes(x=factor(Day, level = day_order), fill=Day)) +
  xlab("Day of week") +
  geom_bar()
```




```
month_order <- c('January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September',
ggplot(data=shootings, aes(x=factor(Month, level = month_order), fill=Month)) +
  theme(axis.text.x=element_text(angle=-30,hjust=0)) +
  xlab("Month") +
  geom_bar()
```



Test des correlations

```r
summary(aov(Fatalities ~ Race, data = shootings))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Race          6    572   95.39   2.631 0.0171 *
## Residuals   264   9572   36.26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 43 observations deleted due to missingness
```

```r
summary(aov(Total.victims ~ Race, data = shootings))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Race          6   3103   517.2    0.38  0.891
## Residuals   264 359283  1360.9
## 43 observations deleted due to missingness
```

```r
summary(aov(Fatalities ~ Weapon.Type, data = shootings))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Weapon.Type 23   1239   53.86   0.773  0.755
## Residuals   91   6344   69.71
## 199 observations deleted due to missingness
```

```r
summary(aov(Total.victims ~ Weapon.Type, data = shootings))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
```

```
## Weapon.Type 23  33838    1471   0.424  0.989
## Residuals   91 315405    3466
## 199 observations deleted due to missingness
```

```
summary(aov(Fatalities ~ Cause, data = filter(shootings, Cause != "unknown")))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Cause        13    465   35.78   1.961 0.0252 *
## Residuals   219   3995   18.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(Total.victims ~ Cause, data = filter(shootings, Cause != "unknown")))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Cause        13   2167  166.67   1.946 0.0265 *
## Residuals   219  18752   85.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(AverageAge ~ Cause, data = filter(shootings, Cause != "unknown")))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Cause        12   2930   244.1   1.519  0.124
## Residuals   137  22025   160.8
## 83 observations deleted due to missingness
```

3.3

Total Victims

```
t.test(shootings$Total.victims ~ shootings$Mental.Health.Issues)
```

```
##
##  Welch Two Sample t-test
##
## data:  shootings$Total.victims by shootings$Mental.Health.Issues
## t = -3.4376, df = 142.57, p-value = 0.0007696
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.554907 -2.038393
## sample estimates:
##  mean in group no mean in group yes
##          6.912088         11.708738
```

```
t.test(shootings$Total.victims ~ shootings$Open.Close.Location)
```

```
##
##  Welch Two Sample t-test
##
## data:  shootings$Total.victims by shootings$Open.Close.Location
## t = -0.58158, df = 79.897, p-value = 0.5625
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.02473  10.41989
## sample estimates:
## mean in group close  mean in group open
##            9.127962           13.430380
```

```r
t.test(shootings$Total.victims ~ shootings$Gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  shootings$Total.victims by shootings$Gender
## t = -1.2999, df = 112.34, p-value = 0.1963
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.604488  1.579135
## sample estimates:
## mean in group female   mean in group male
##              7.60000             10.61268
```

Age

```r
t.test(shootings$AverageAge ~ shootings$Mental.Health.Issues)
```

```
##
##  Welch Two Sample t-test
##
## data:  shootings$AverageAge by shootings$Mental.Health.Issues
## t = 1.4516, df = 110.65, p-value = 0.1495
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.164305  7.540808
## sample estimates:
##  mean in group no mean in group yes
##          33.31034          30.12209
```

```r
t.test(shootings$AverageAge ~ shootings$Open.Close.Location)
```

```
##
##  Welch Two Sample t-test
##
## data:  shootings$AverageAge by shootings$Open.Close.Location
## t = -0.77815, df = 27.136, p-value = 0.4432
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.958124  4.480866
## sample estimates:
## mean in group close  mean in group open
##            31.19615            33.93478
```

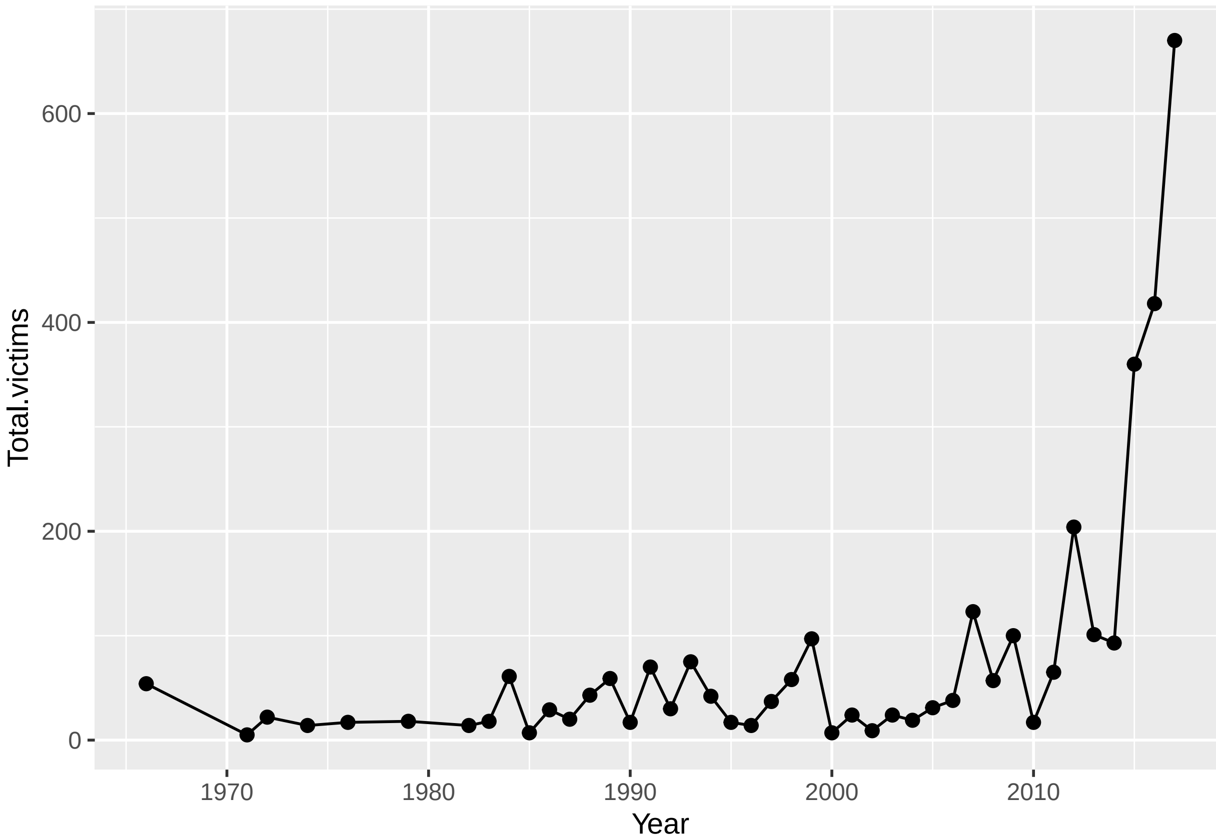
```r
t.test(shootings$Age ~ shootings$Gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  shootings$Age by shootings$Gender
## t = 0.69416, df = 4.2694, p-value = 0.5235
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.46681  19.36919
## sample estimates:
## mean in group female   mean in group male
```

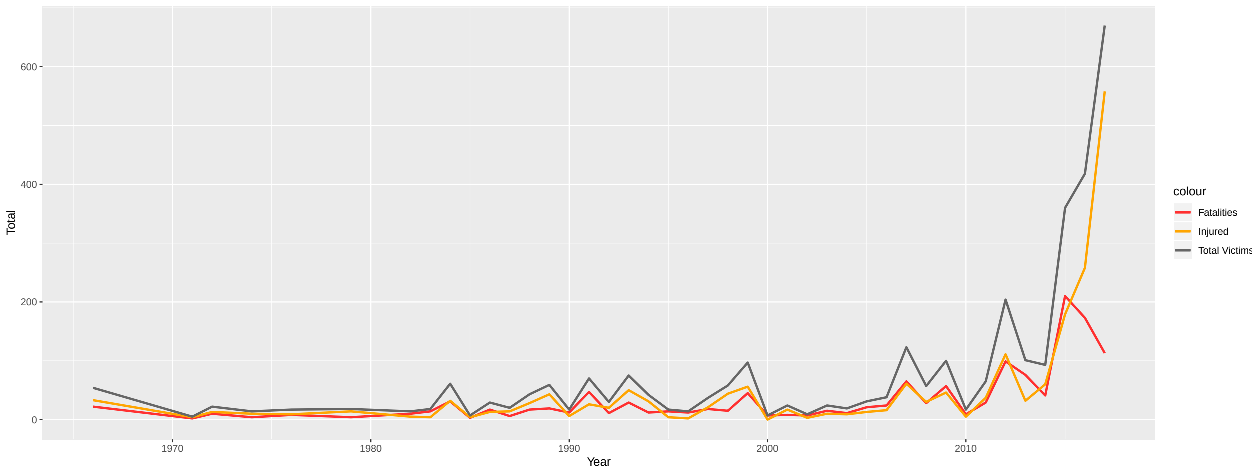
```
##             35.60000             31.64881
```

3.5

```r
ggplot(shootings.by.year, aes(x=Year, y=Total.victims)) +
  geom_point(size=2, aes(size=20)) +
  stat_summary(fun.y = sum, geom="line")
```




```r
ggplot(shootings.by.year, aes(x=Year)) +
  geom_line(aes(Year, Fatalities, color="Fatalities"), size=1) +
  geom_line(aes(Year, Injured, color="Injured"), size=1) +
  geom_line(aes(Year, Total.victims, color="Total Victims"), size=1) +
  scale_colour_manual(values = c("Fatalities" = "firebrick1", "Injured" = "orange", "Total Victims" = "
  ylab("Total")
```




```r
shootings$Total.victims <- as.numeric(shootings$Total.victims)
shootings$Injured <- as.numeric(shootings$Injured)
shootings$Fatalities <- as.numeric(shootings$Fatalities)

data <- shootings %>%
  gather(key = "variable", value = "value", Injured, Fatalities)
data <- data[,c("variable", "value", "Year")]

data <- data  %>%
  dplyr::group_by(Year, variable) %>%
  dplyr::summarise(n = sum(value)) %>%
  dplyr::mutate(percentage = n / sum(n))

ggplot(data, aes(x=Year, y = percentage, fill = variable)) +
  geom_area(color = "black")
```


```
victims.regression.linear <- lm(Total.victims ~ Year, data=shootings.by.year)

summary(victims.regression.linear)
```

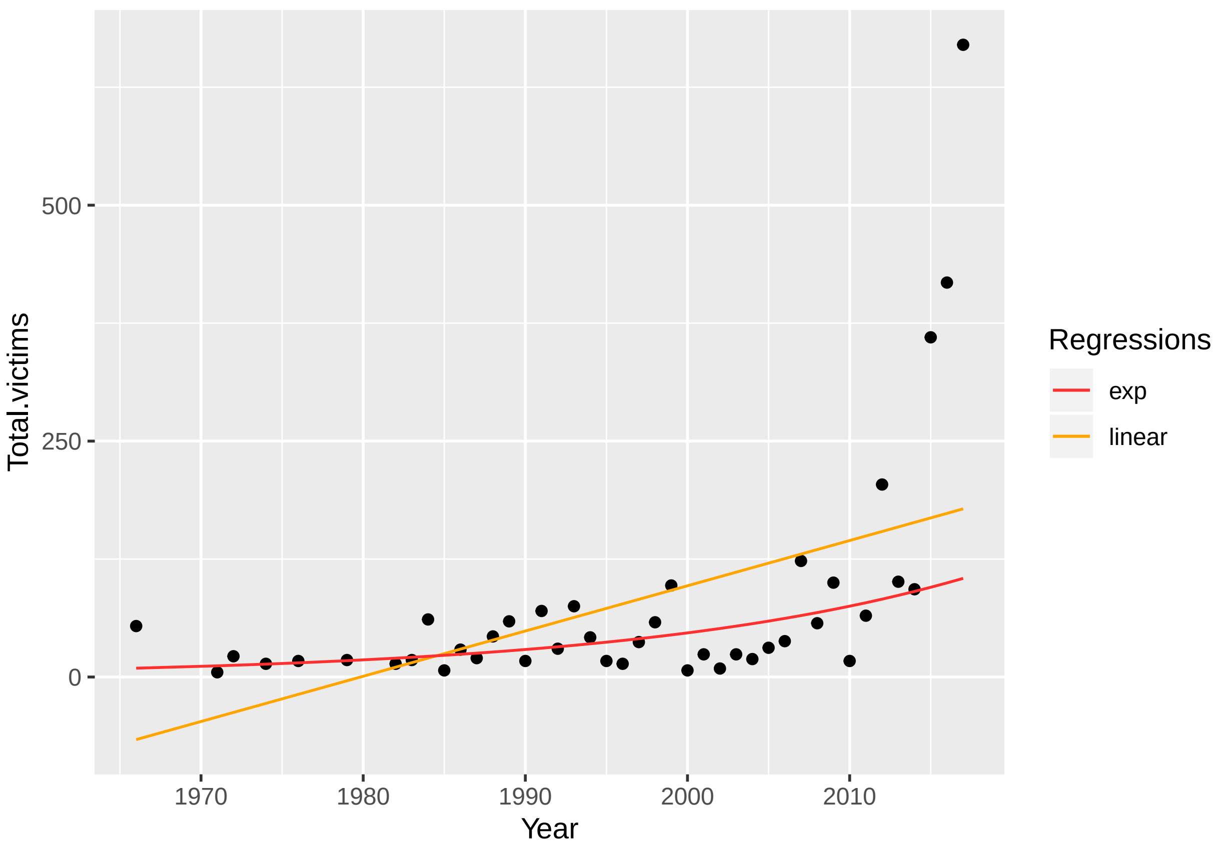
```
##
## Call:
## lm(formula = Total.victims ~ Year, data = shootings.by.year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -127.64  -68.99  -16.07   20.61  491.80
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9492.560   2497.079  -3.801 0.000481 ***
## Year            4.795      1.251   3.832 0.000439 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109.1 on 40 degrees of freedom
## Multiple R-squared:  0.2685, Adjusted R-squared:  0.2502
## F-statistic: 14.68 on 1 and 40 DF,  p-value: 0.0004394
```

Une régression linéaire n’est pas satisfaisante

```
victims.regression.exp <- lm(log(Total.victims) ~ Year, data=shootings.by.year)

summary(victims.regression.exp)
```

```
##
## Call:
## lm(formula = log(Total.victims) ~ Year, data = shootings.by.year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8991 -0.6204  0.0247  0.6197  1.8578
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -90.79047   20.55466  -4.417 7.42e-05 ***
## Year          0.04732    0.01030   4.594 4.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8977 on 40 degrees of freedom
## Multiple R-squared:  0.3454, Adjusted R-squared:  0.3291
## F-statistic: 21.11 on 1 and 40 DF,  p-value: 4.27e-05
```

```r
shootings.by.year %>%
  mutate( model = predict(victims.regression.linear)) %>%
  mutate( model.exp = exp(predict(victims.regression.exp))) %>%
  ggplot(x = Year) +
  geom_point( aes(Year, Total.victims)) +
  geom_line( aes(Year, model, colour="linear")) +
  geom_line( aes(Year, model.exp, colour="exp")) +
  scale_colour_manual(name = 'Regressions', values =c('linear'='orange','exp'='firebrick1'))
```



```r
shootings.by.year.2010 <- filter(shootings.by.year, Year >= 2010)

victims.regression.linear.2010 <- lm(Total.victims ~ Year, data=shootings.by.year.2010)

summary(victims.regression.linear.2010)
```

```
##
## Call:
## lm(formula = Total.victims ~ Year, data = shootings.by.year.2010)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -188.45  -43.83   11.95   65.46  145.83
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -162660.74   35569.51  -4.573  0.00380 **
## Year            80.90      17.67   4.580  0.00377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114.5 on 6 degrees of freedom
## Multiple R-squared:  0.7776, Adjusted R-squared:  0.7405
## F-statistic: 20.97 on 1 and 6 DF,  p-value: 0.003771
```

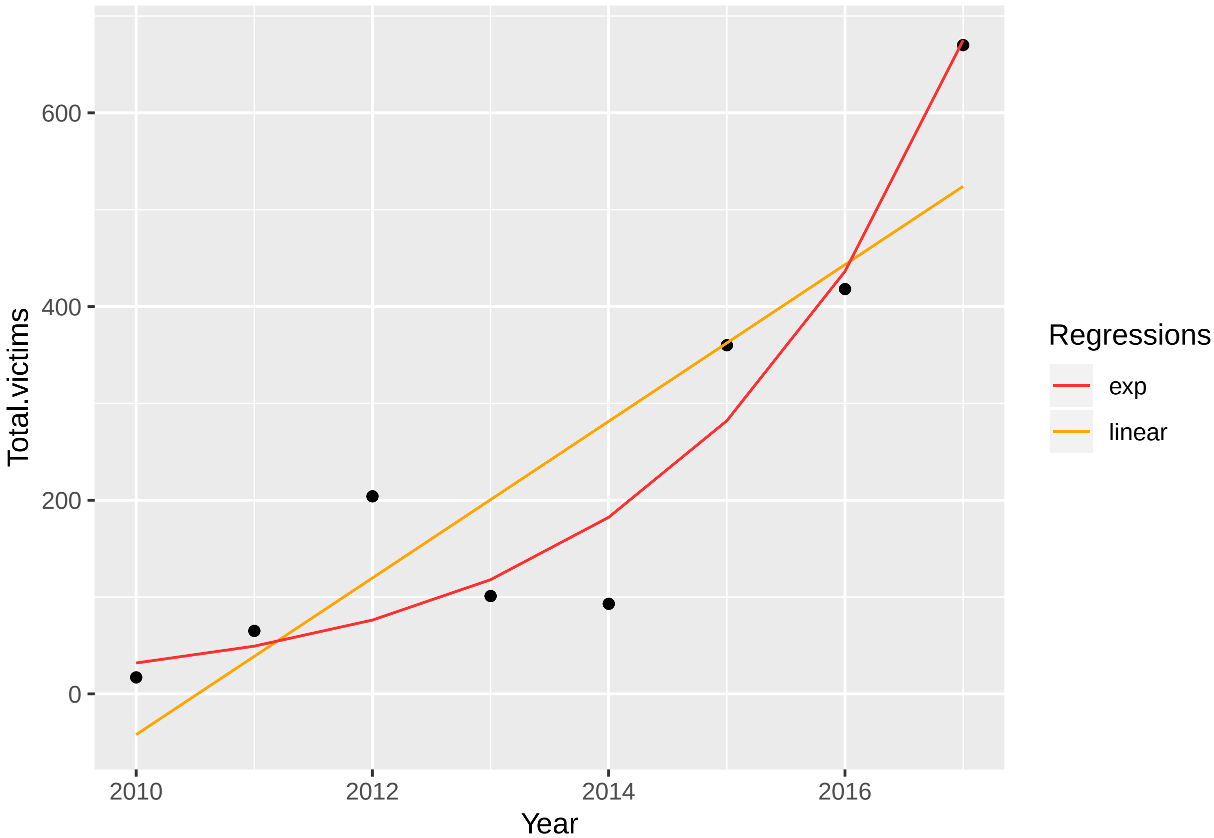
Une régression linéaire n'est pas satisfaisante

```r
victims.regression.exp.2010 <- lm(log(Total.victims) ~ Year, data=shootings.by.year.2010)

summary(victims.regression.exp.2010)
```

```
##
## Call:
## lm(formula = log(Total.victims) ~ Year, data = shootings.by.year.2010)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67332 -0.27283 -0.02517  0.25226  0.98471
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -873.41004  178.43938  -4.895  0.00273 **
## Year           0.43625    0.08862   4.923  0.00265 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5743 on 6 degrees of freedom
## Multiple R-squared:  0.8015, Adjusted R-squared:  0.7685
## F-statistic: 24.23 on 1 and 6 DF,  p-value: 0.00265
```

```r
shootings.by.year.2010 %>%
  mutate( model = predict(victims.regression.linear.2010)) %>%
  mutate( model.exp = exp(predict(victims.regression.exp.2010))) %>%
  ggplot(x = Year) +
  geom_point( aes(Year, Total.victims) ) +
```

```r
  geom_line( aes(Year, model, colour="linear")) +
  geom_line( aes(Year, model.exp, colour="exp")) +
  scale_colour_manual(name = 'Regressions', values =c('linear'='orange','exp'='firebrick1'))
```




```r
dates <- data.frame(Year=c(2018, 2020, 2025, 2030, 2050, 2075, 2100))
predict(victims.regression.linear, dates)
```

```
##        1        2        3        4        5        6        7
## 182.9945 192.5838 216.5569 240.5300 336.4225 456.2882 576.1538
```

```r
dates <- data.frame(Year=c(2018, 2020, 2025, 2030, 2050, 2075, 2100))
exp(predict(victims.regression.exp, dates))
```

```
##         1         2         3         4         5         6         7
##  109.5914  120.4692  152.6249  193.3635  498.1629 1625.9883 5307.1755
```

```r
dates <- data.frame(Year=c(2018, 2020, 2025, 2030, 2050, 2075, 2100))
predict(victims.regression.linear.2010, dates)
```

```
##         1         2         3         4         5         6         7
##  605.0714  766.8810 1171.4048 1575.9286 3194.0238 5216.6429 7239.2619
```

```r
dates <- data.frame(Year=c(2018, 2020, 2025, 2030, 2050, 2075, 2100))
exp(predict(victims.regression.exp.2010, dates))
```

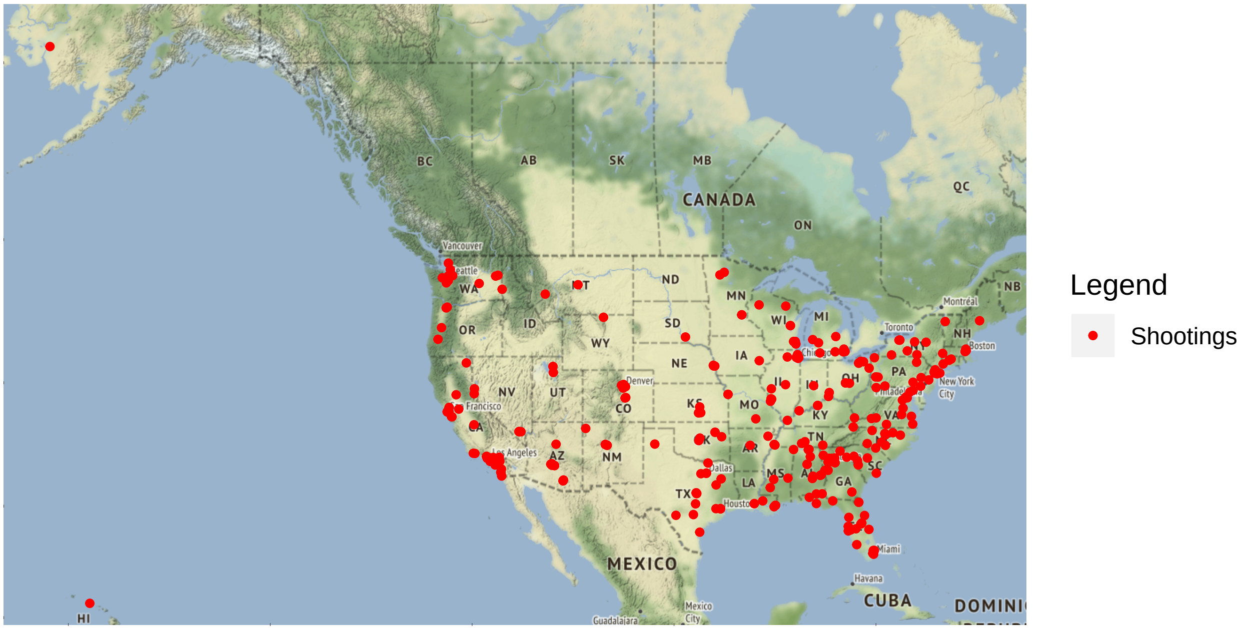
```
##            1            2            3            4            5            6
## 1.044122e+03 2.498486e+03 2.213048e+04 1.960220e+05 1.206593e+09 6.578553e+13
##            7
## 3.586742e+18
```

4

```r
qmplot(Longitude, Latitude, data = shootings, maptype = "toner-lite", color = "red", size = I(1)) +
  labs(title = "Shootings' Location\n", x = "", y = "", color = "Legend") +
  scale_color_manual(labels = c("Shootings"), values = c("red"))
```

```
## Using zoom = 4...
## Source : http://tile.stamen.com/terrain/4/0/4.png
## Source : http://tile.stamen.com/terrain/4/1/4.png
## Source : http://tile.stamen.com/terrain/4/2/4.png
## Source : http://tile.stamen.com/terrain/4/3/4.png
## Source : http://tile.stamen.com/terrain/4/4/4.png
## Source : http://tile.stamen.com/terrain/4/5/4.png
## Source : http://tile.stamen.com/terrain/4/0/5.png
## Source : http://tile.stamen.com/terrain/4/1/5.png
## Source : http://tile.stamen.com/terrain/4/2/5.png
## Source : http://tile.stamen.com/terrain/4/3/5.png
## Source : http://tile.stamen.com/terrain/4/4/5.png
## Source : http://tile.stamen.com/terrain/4/5/5.png
## Source : http://tile.stamen.com/terrain/4/0/6.png
## Source : http://tile.stamen.com/terrain/4/1/6.png
## Source : http://tile.stamen.com/terrain/4/2/6.png
## Source : http://tile.stamen.com/terrain/4/3/6.png
## Source : http://tile.stamen.com/terrain/4/4/6.png
## Source : http://tile.stamen.com/terrain/4/5/6.png
## Source : http://tile.stamen.com/terrain/4/0/7.png
## Source : http://tile.stamen.com/terrain/4/1/7.png
## Source : http://tile.stamen.com/terrain/4/2/7.png
## Source : http://tile.stamen.com/terrain/4/3/7.png
## Source : http://tile.stamen.com/terrain/4/4/7.png
## Source : http://tile.stamen.com/terrain/4/5/7.png
```

Shootings' Location

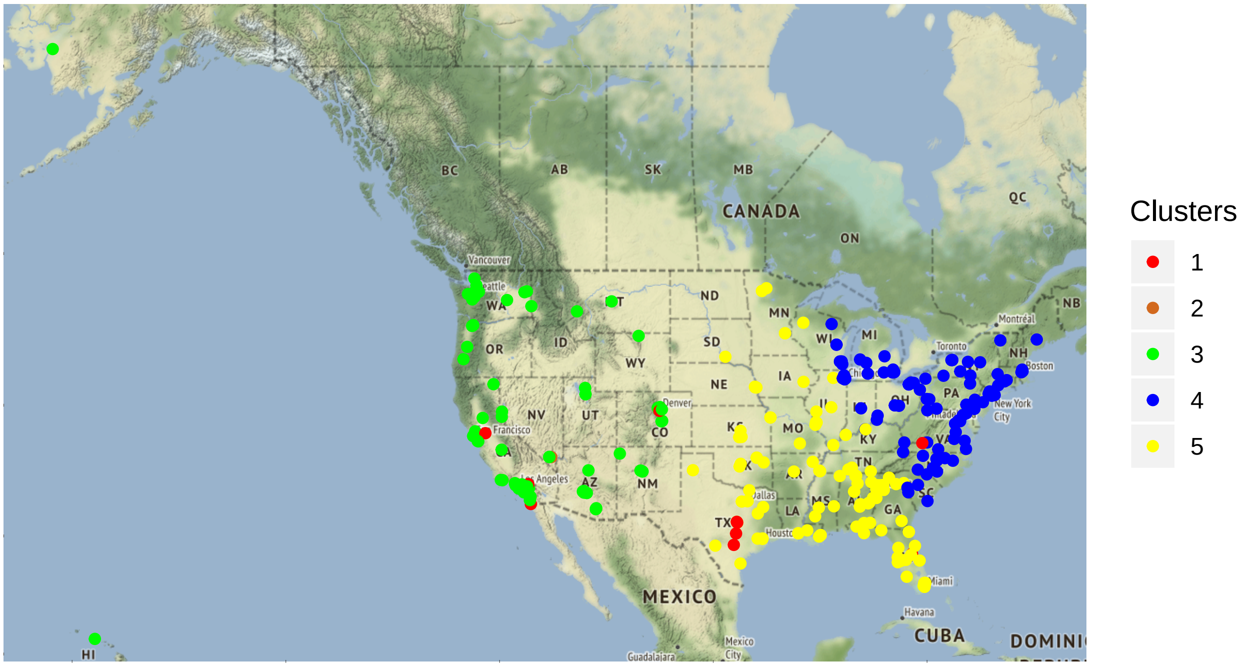


5

```
shootings_bis <- shootings

clusters <- kmeans(shootings_bis[c('Latitude', 'Longitude', 'Total.victims')], 5)

# Save the cluster number in the dataset as column
shootings_bis$Clusters <- as.factor(clusters$cluster)

qmplot(Longitude, Latitude, data = shootings_bis, maptype = "terrain-background", color = Clusters) +
  labs(title = "K Means clustering Visualition of mass shootings in the US\n", x = "", y = "", color =
  scale_color_manual(labels = c("1", "2", "3", "4", "5", "6"),
                     values =  c("red", "chocolate", "green", "blue", "yellow", "deeppink"))
```

```
## Using zoom = 4...
```

K Means clustering Visualition of mass shootings in the US

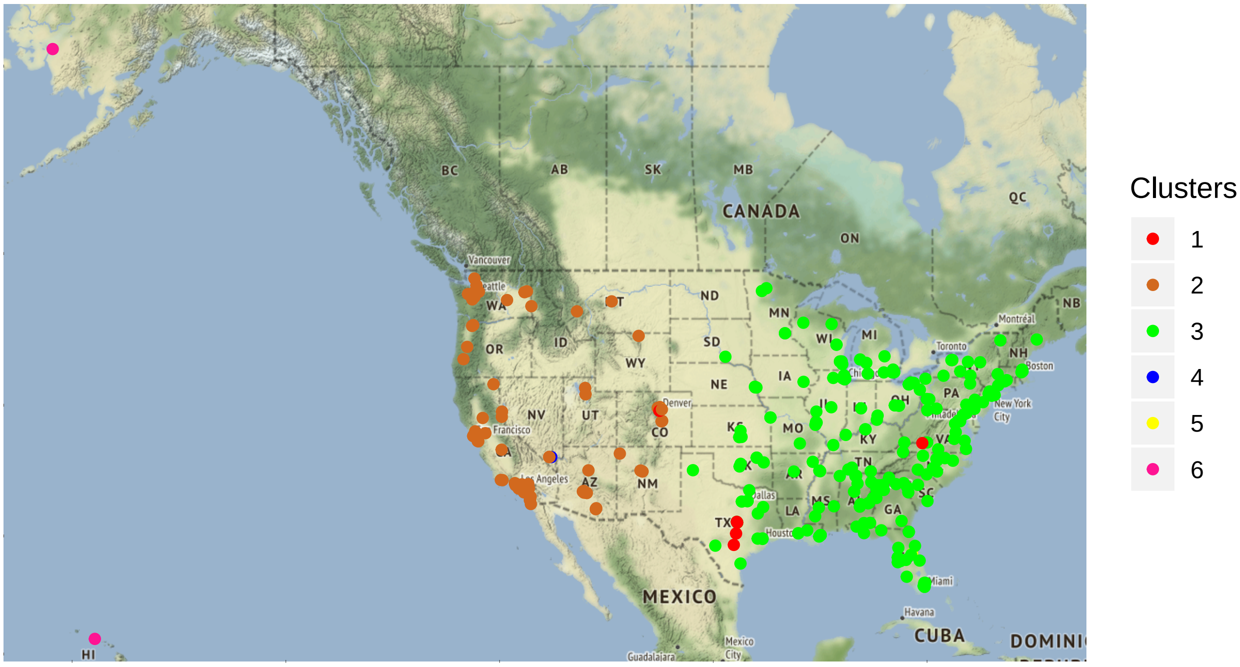


6

```r
hierachical.cluster <- hclust(dist(shootings[c('Latitude', 'Longitude', 'Total.victims')]))
cut_cluster <- cutree(hierachical.cluster, k = 6)

shootings_bis$HierarchicalClusters <- as.factor(cut_cluster)

qmplot(Longitude, Latitude, data = shootings_bis, maptype = "terrain-background", color = HierarchicalCl
  labs(title = "Hierarchical clustering Visualition of mass shootings in the US\n", x = "", y = "", col
  scale_color_manual(labels = c("1", "2", "3", "4", "5", "6"),
                     values =  c("red", "chocolate", "green", "blue", "yellow", "deeppink"))
```

```
## Using zoom = 4...
```

Hierarchical clustering Visualition of mass shootings in the US



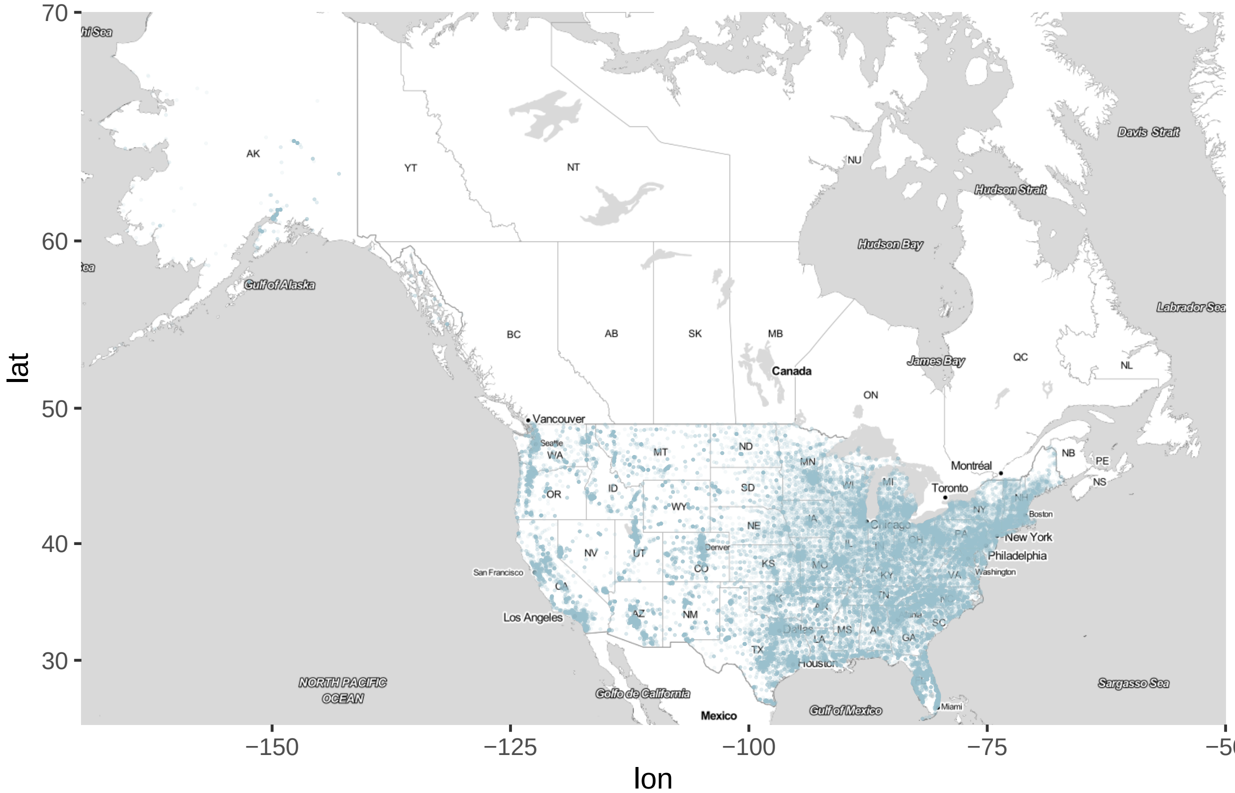
7

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `Lic Regn` = col_double(),
##   `Lic Dist` = col_double(),
##   `Lic Cnty` = col_double(),
##   `Lic Type` = col_double(),
##   `Lic Seqn` = col_double(),
##   `Premise Zip Code` = col_double(),
##   `Mail Zip Code` = col_double(),
##   `Voice Phone` = col_double(),
##   Longitude = col_double(),
##   Latitude = col_double()
## )

## See spec(...) for full column specifications.
```

```
ggmap(get_stamenmap(bbox = c(left = -170, bottom = 24, right = -50, top = 70), zoom = 4,
        color = "color",
        maptype = "toner-lite", force = FALSE
)) + geom_point(aes(x = Longitude, y = Latitude), color = "lightblue3", data = gun_license, size = 0.01
  theme(legend.position="bottom")
```

```
## Source : http://tile.stamen.com/toner-lite/4/0/3.png

## Source : http://tile.stamen.com/toner-lite/4/1/3.png

## Source : http://tile.stamen.com/toner-lite/4/2/3.png

## Source : http://tile.stamen.com/toner-lite/4/3/3.png

## Source : http://tile.stamen.com/toner-lite/4/4/3.png
```

```
## Source : http://tile.stamen.com/toner-lite/4/5/3.png
## Source : http://tile.stamen.com/toner-lite/4/0/4.png
## Source : http://tile.stamen.com/toner-lite/4/1/4.png
## Source : http://tile.stamen.com/toner-lite/4/2/4.png
## Source : http://tile.stamen.com/toner-lite/4/3/4.png
## Source : http://tile.stamen.com/toner-lite/4/4/4.png
## Source : http://tile.stamen.com/toner-lite/4/5/4.png
## Source : http://tile.stamen.com/toner-lite/4/0/5.png
## Source : http://tile.stamen.com/toner-lite/4/1/5.png
## Source : http://tile.stamen.com/toner-lite/4/2/5.png
## Source : http://tile.stamen.com/toner-lite/4/3/5.png
## Source : http://tile.stamen.com/toner-lite/4/4/5.png
## Source : http://tile.stamen.com/toner-lite/4/5/5.png
## Source : http://tile.stamen.com/toner-lite/4/0/6.png
## Source : http://tile.stamen.com/toner-lite/4/1/6.png
## Source : http://tile.stamen.com/toner-lite/4/2/6.png
## Source : http://tile.stamen.com/toner-lite/4/3/6.png
## Source : http://tile.stamen.com/toner-lite/4/4/6.png
## Source : http://tile.stamen.com/toner-lite/4/5/6.png
## Warning: Removed 1806 rows containing missing values (geom_point).
```




```r
cor(Data$ShootingsPct, Data$ShopsPct)
```

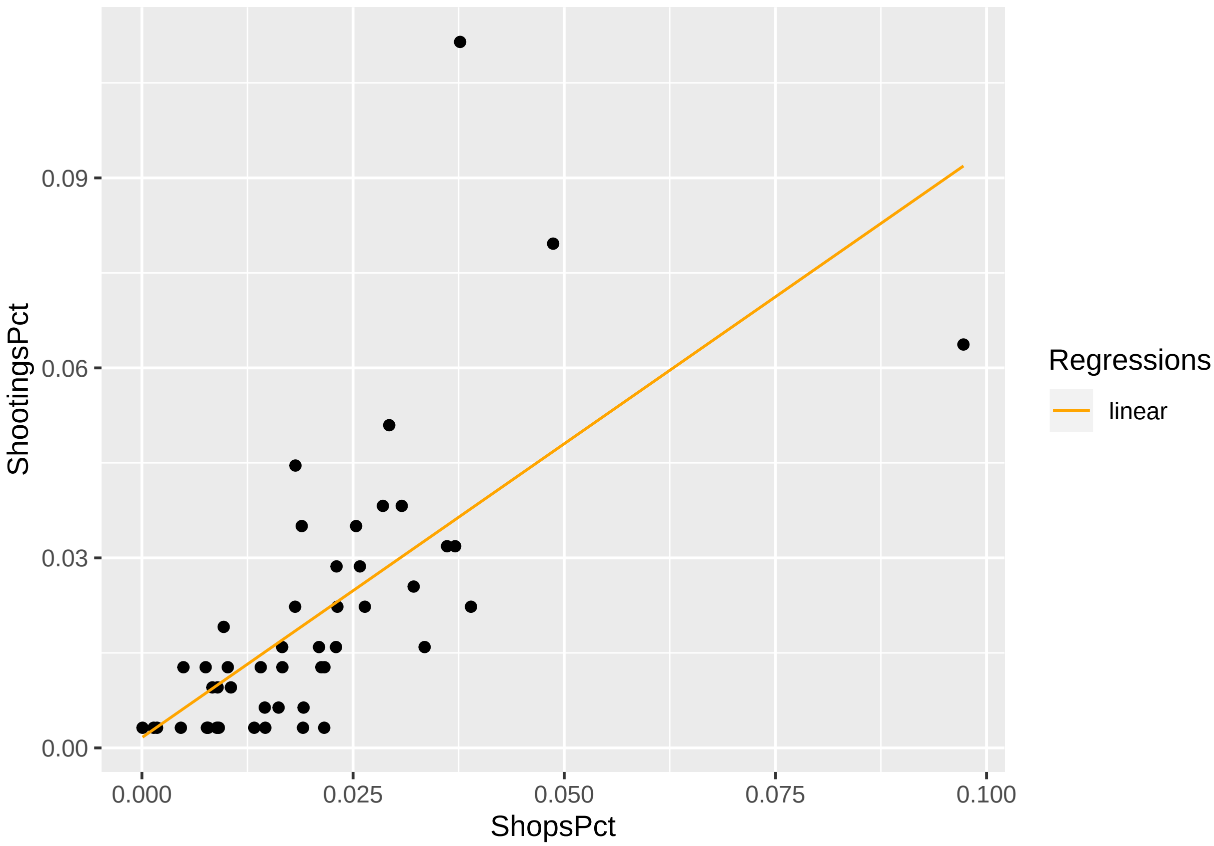
```
## [1] 0.6821753
```

```r
rel <- lm(ShootingsPct ~ ShopsPct, data = Data)
summary(rel)
```

```
##
## Call:
## lm(formula = ShootingsPct ~ ShopsPct, data = Data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028186 -0.007422 -0.002339  0.003900  0.074878
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.001631   0.003778   0.432    0.668
## ShopsPct    0.927807   0.146626   6.328 9.33e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01588 on 46 degrees of freedom
## Multiple R-squared:  0.4654, Adjusted R-squared:  0.4537
## F-statistic: 40.04 on 1 and 46 DF,  p-value: 9.328e-08
```

```r
Data %>%
  mutate( model = predict(rel)) %>%
  ggplot(x = ShopsPct) +
  geom_point( aes(ShopsPct, ShootingsPct) ) +
```

```
  geom_line( aes(ShopsPct, model, colour="linear")) +
  scale_colour_manual(name = 'Regressions', values =c('linear'='orange'))
```



8

```
shootings.State.data <- shootings %>%
  dplyr::group_by(State) %>%
  dplyr::count() %>%
  dplyr::ungroup() %>%
  dplyr::mutate(ShootingsPct=`n`/sum(`n`)) %>%
  dplyr::mutate(Shootings=`n`) %>%
 dplyr:: arrange(desc(State))

shootings.State.data$state <- shootings.State.data$State

plot_usmap(data = shootings.State.data, values = "n", color = "black") +
  scale_fill_continuous(low = "lightgoldenrodyellow", high = "red", name = "Shootings", label = scales:
  theme(legend.position = "right") +
  ggtitle("US Shootings By State")
```

US Shootings By State