Relationship between US Regions and Homocide Rates of Young Women between 1980-2014

DATA 606 Data Project Submission

Gabriel Campos

20 November 2020

```
## [1] " Attaching packages: DATA606,infer,dplyr,VennDiagram,scales,data.table,readr "
```

Part 1 - Introduction

A primary concern regarding public safety that, impacts all of society is murder. The threat of murder negatively impacts the widely accepted fundamental human need for safety. Murder by region can impact the population psychologically, economically and can stifle a community's continued progress for a better standard of living. Gaining insight on those who are the most vulnerable and impacted by murder is necessary to forge initiatives in assisting those victimized by its threat. Like many approaches in troubleshooting or diagnosing an issue, understanding where to begin is a key first step. In order to focus this analysis, we will look further into what attributes impact the murder counts in the United States based on the data set.

Overview

The goals for this project are to:

- Think about the independent and dependent variables in correlation to murder counts including region, gender and age.
- Compare murder rates by region, gender and age based on the data set
- Draw conclusions of the factors impacting the counts and question if these factors are localized based on our findings.

Part 2 - Data

Collection

The Murder Accountability Project is the most complete database of homicides in the United States currently available. This dataset includes murders from the FBI's Supplementary Homicide Report from 1976 to the present and Freedom of Information Act data on more than 22,000 homicides that were not reported to the Justice Department. This dataset includes the age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used. A victims age is rounded down by year (e.g. toddlers 11 months old or younger qualify as 0 years old). If a victim's age cannot be determined, they will be categorized as 998 years old respectively.

Data source: kaggle.com

Load Data

```
## [1] "Loading Data: database.csv "
```

```
## [1] "Subset Data: Region <- project_data %>% filter() "
```

```
## [1] "Vector Region: project_data$Region<-ifelse(State,Region1,ifelse(State,Region2,"
```

```
## [1] "Vector AgeGroup: project_data$`Victim AgeGroup`<- cut(project_data$`Victim Age`,... "
```

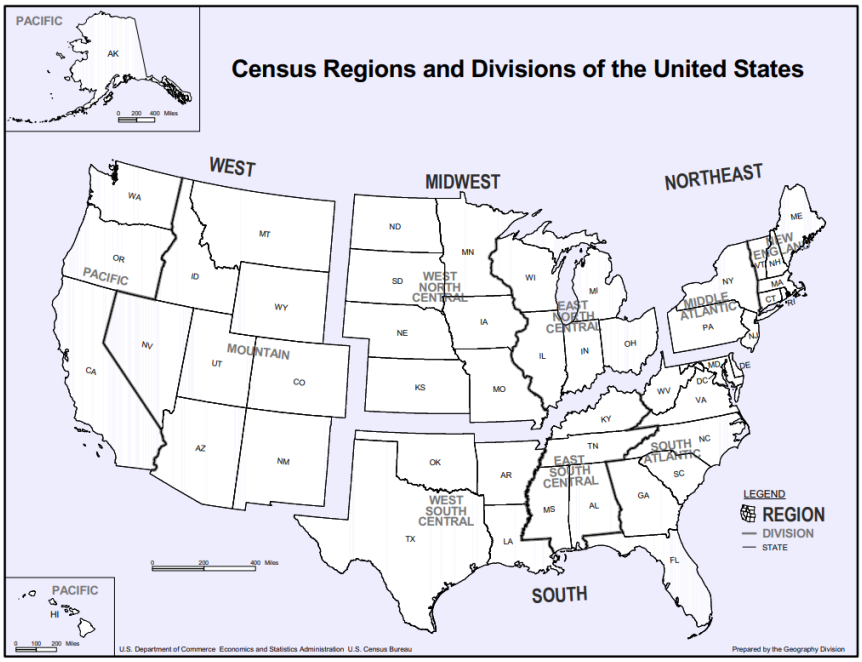


Figure 1: US regions based on census.gov

Cases

Categorical Data: Record ID, Agency Code, Agency Name, Agency Type, City, State,Incident, Crime Type,Crime Solved,Victim Sex,Victim Race,Victim Ethnicity,Perpetrator Sex, Perpetrator Race,Perpetrator Ethnicity,Relationship,Weapon,Record Source, Region

Numerical: Year,Month,Victim Age,Perpetrator Age,Victim Count,Perpetrator Count

Outliers: 974

There are **638,454** total cases in our data set, with **66,301** representing **murders** committed against **Female's under** the age of **30** throughout the United States from 1980-2014

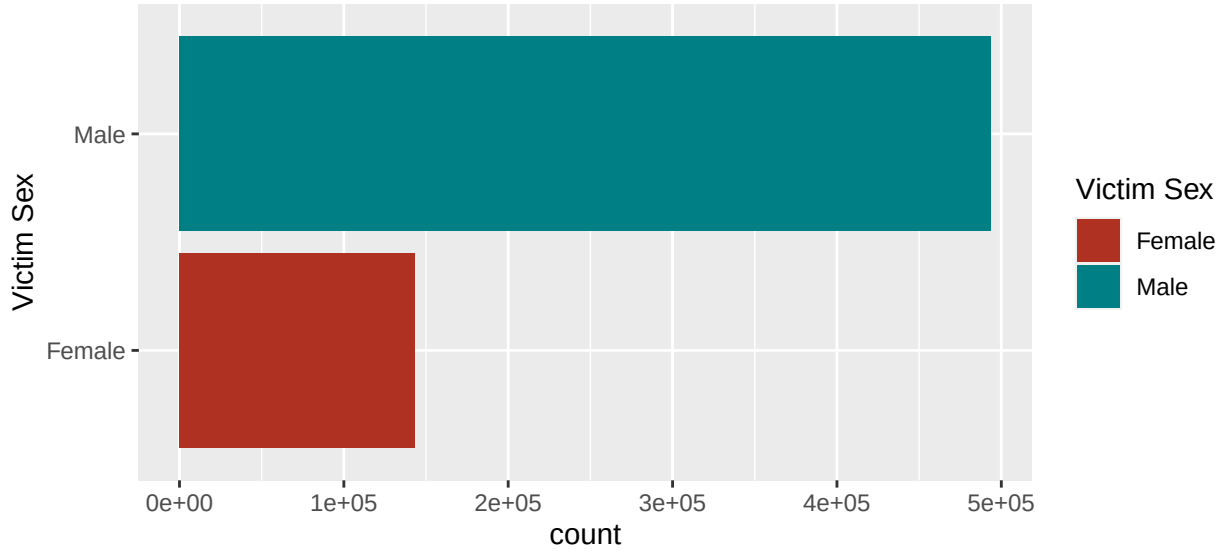
A break down of murders by age are as follows
Note : *Age* 998 *was assigned to victims who*′*s age could not be determined.*
To avoid the impact it will have on our results I will remove it from our existing dataframes

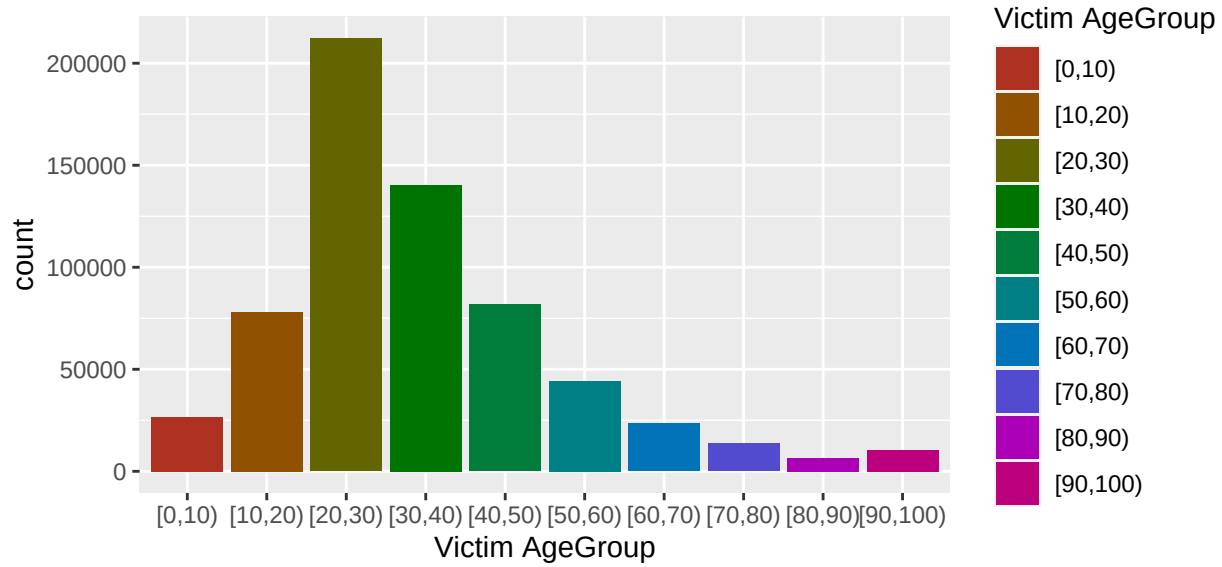
```
## count
##     0     1     2     3     4     5     6     7     8     9    10    11    12
##  8444  5525  3805  2378  1659  1194   999   915   852   834   854   911  1239
##    13    14    15    16    17    18    19    20    21    22    23    24    25
##  1897  3342  5905  9402 14030 18469 21939 23031 22796 23049 22438 21830 22939
##    26    27    28    29    30    31    32    33    34    35    36    37    38
## 20469 19465 18199 18037 18966 15762 15812 14463 14296 14314 12502 11829 11411
##    39    40    41    42    43    44    45    46    47    48    49    50    51
## 10921 11163  9594  9613  8629  7921  8157  7336  6902  6365  6149  6325  5270
##    52    53    54    55    56    57    58    59    60    61    62    63    64
##  5203  4788  4466  4246  3939  3721  3272  3184  3171  2797  2862  2519  2271
##    65    66    67    68    69    70    71    72    73    74    75    76    77
##  2418  1861  2013  1840  1663  1783  1566  1596  1390  1367  1411  1213  1135
##    78    79    80    81    82    83    84    85    86    87    88    89    90
##  1102  1098  1067   930   835   765   686   627   574   460   408   313   281
##    91    92    93    94    95    96    97    98    99   998
##   215   156   134   116    82    37    39    33  9281   974
```

```
## [1] "Non determined variables removed with the following command:"
```

```
## [1] "data.frame<-data.frame[!(dataframe$`Victim Age`==998),]"
```

The charts below show that of the murders reported, the majority of these are against men. Additionally, regardless of gender, nationally the highest murder victims age falls with Victims age 20-30.





Part 3 - Exploratory data analysis

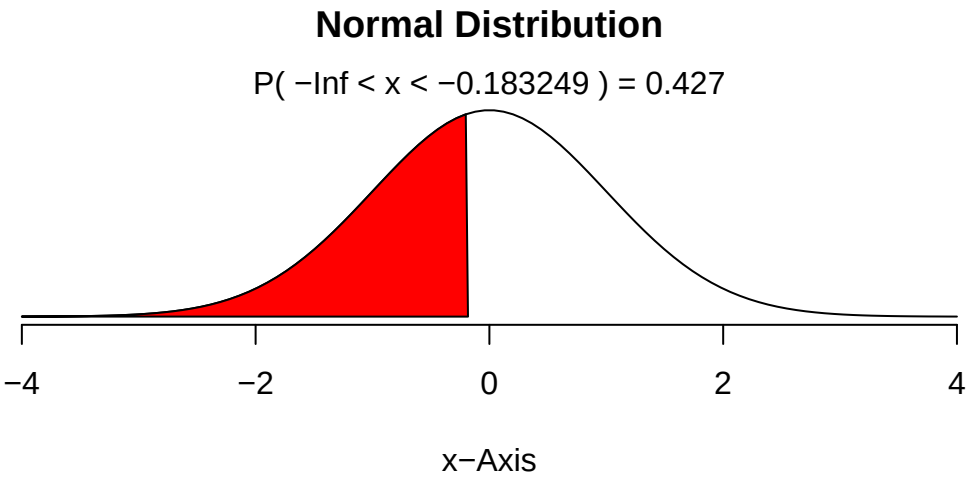
National Age Average

```
## [1] "Summary of age for entire dataset"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   22.00   30.00   33.56   42.00   99.00
```

```
## [1] 17.78703
```

- **Murders 1980-2014: N(μ = 33.26, $\sigma \approx$ 17.79)**
- I would like to see the probability of a murder taking place on `Victim Sex` ≤ 30
- In order to do so I can calculate $\mathbf{Z = \frac{x-\mu}{\sigma}}$ **Where** $x \leq \mathbf{30} = -0.183249$



What's interesting here, $\approx 42.7\%$ of all the reported murders in our data set, of victims whose age we can determine, occurs with the first 30 years of there life. Next I will compare the average age of murdered women in our dataset, by region.

Regional Age Averages

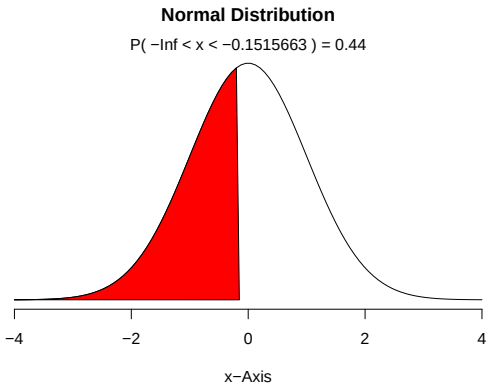
```
## [1] "In order to compare variabilities I created vectors in our master data set using"
```

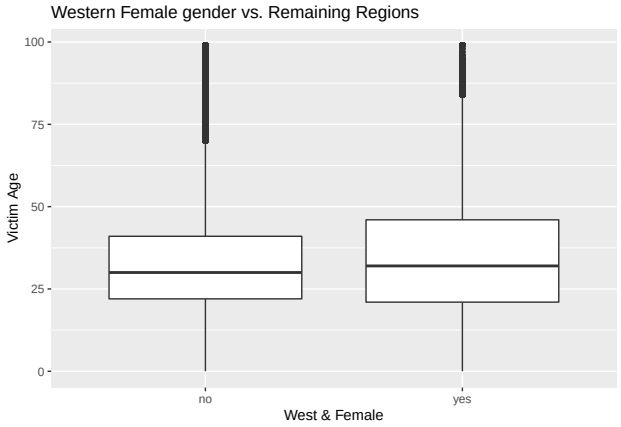
```
## [1] "project_data <- project_data %>% mutate(femW = ifelse(project_data$Region...."
```

```
## [1] "allowing for comparison of one regions variability with the remaining 3 combined"
```

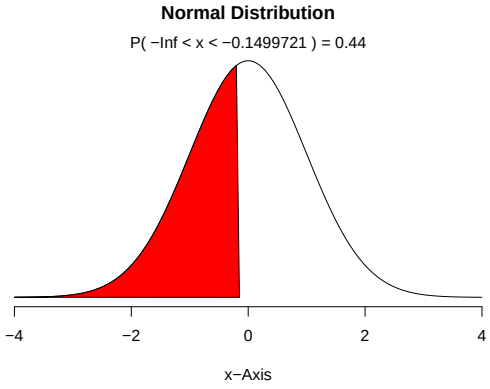
The boxplots below show that with respect to age, our data is right skewed, indicating murders occurred at a higher frequency within the first three decades of our victims lives. This is supported with consistent medians, relatively close means and similar number of outliers shown. The only standouts would be the Pacific, which has slightly more scattered outliers, and the South which has the lowest percentage of murder victims below 30 (40%). Based on the data, I suspect age brackets hardly impact murders among women below 30.

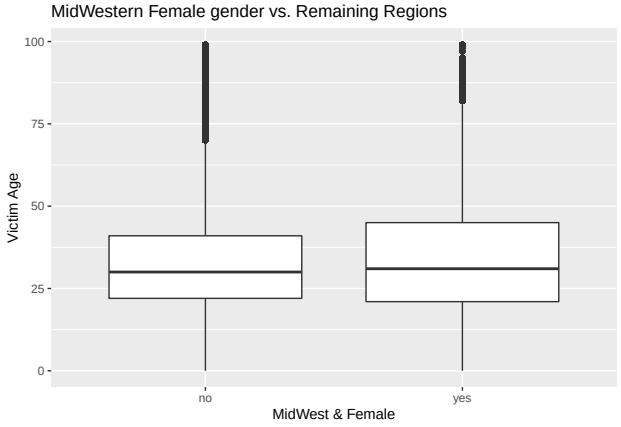
Western Murders 1980-2014: N(μ = 32.64, σ = 17.42)
Z = $\frac{x-\mu}{\sigma}$ Where $x \leq$ 30 $= -0.1515663$ making $\approx 44\%$ of murder victims years$\leq$ **30**



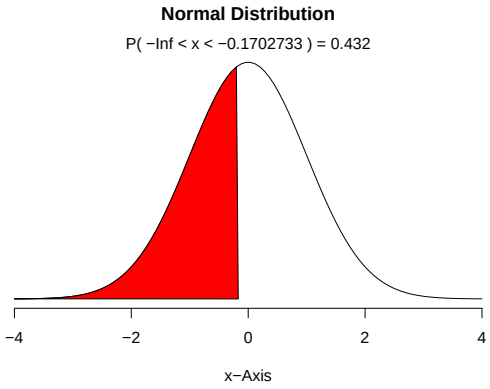


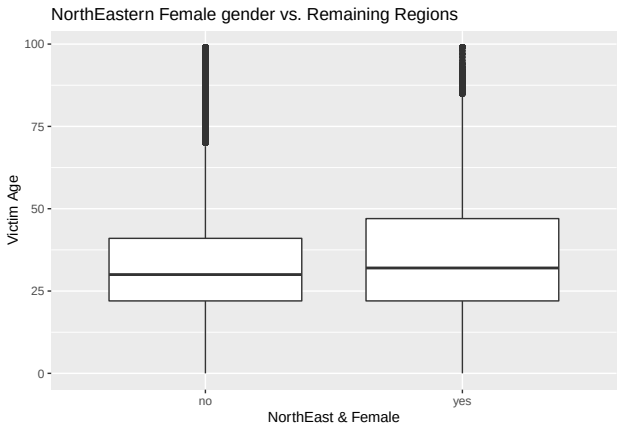
Midwestern Murders 1980-2014: N(μ = 32.6, $\sigma \approx$ 17.34)
Z = $\frac{x-\mu}{\sigma}$ Where $x \leq$ 30 $= -0.1499721$ making $\approx 44\%$ of murder victims years$\leq$ **30**



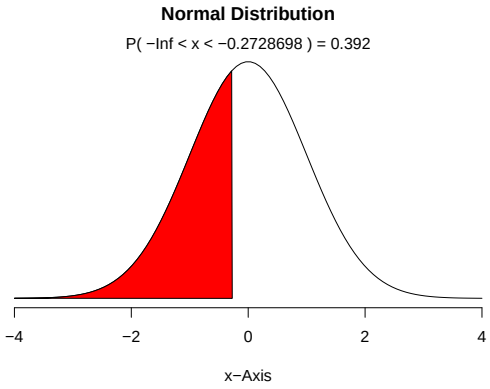


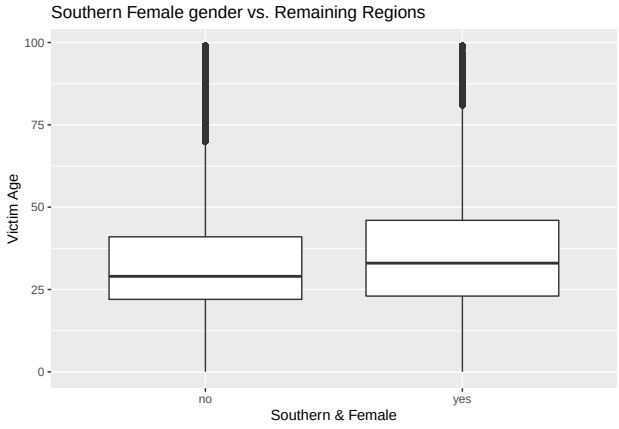
Northeastern Murders 1980-2014: N(μ = 32.99, $\sigma \approx$ 17.56)
Z = $\frac{x-\mu}{\sigma}$ **Where** $x \leq$ **30** $= -0.1702733$ making $\approx 43\%$ of murder victims years$\leq$ **30**



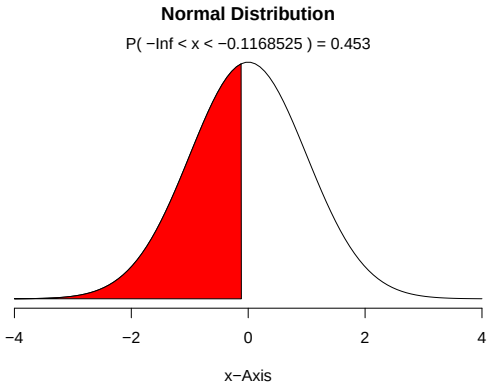


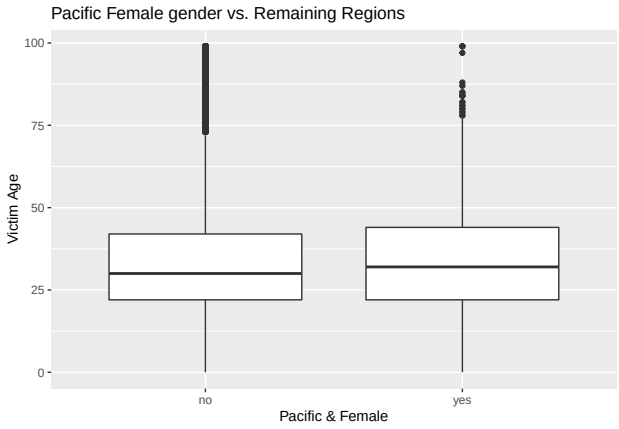
Southern Murders 1980-2014: N(μ = 35, $\sigma \approx$ 18.32)
Z = $\frac{x-\mu}{\sigma}$ **Where** $x \leq$ **30** $= -0.2728698$ making $\approx 40\%$ of murder victims years$\leq$ **30**





Pacific Murders 1980-2014: N(μ = 33.81, $\sigma \approx$ 17.12)
Z = $\frac{x-\mu}{\sigma}$ **Where** $x \leq$ **30** $= -0.1168525$ making $\approx 45\%$ of murder victims years$\leq$ **30**



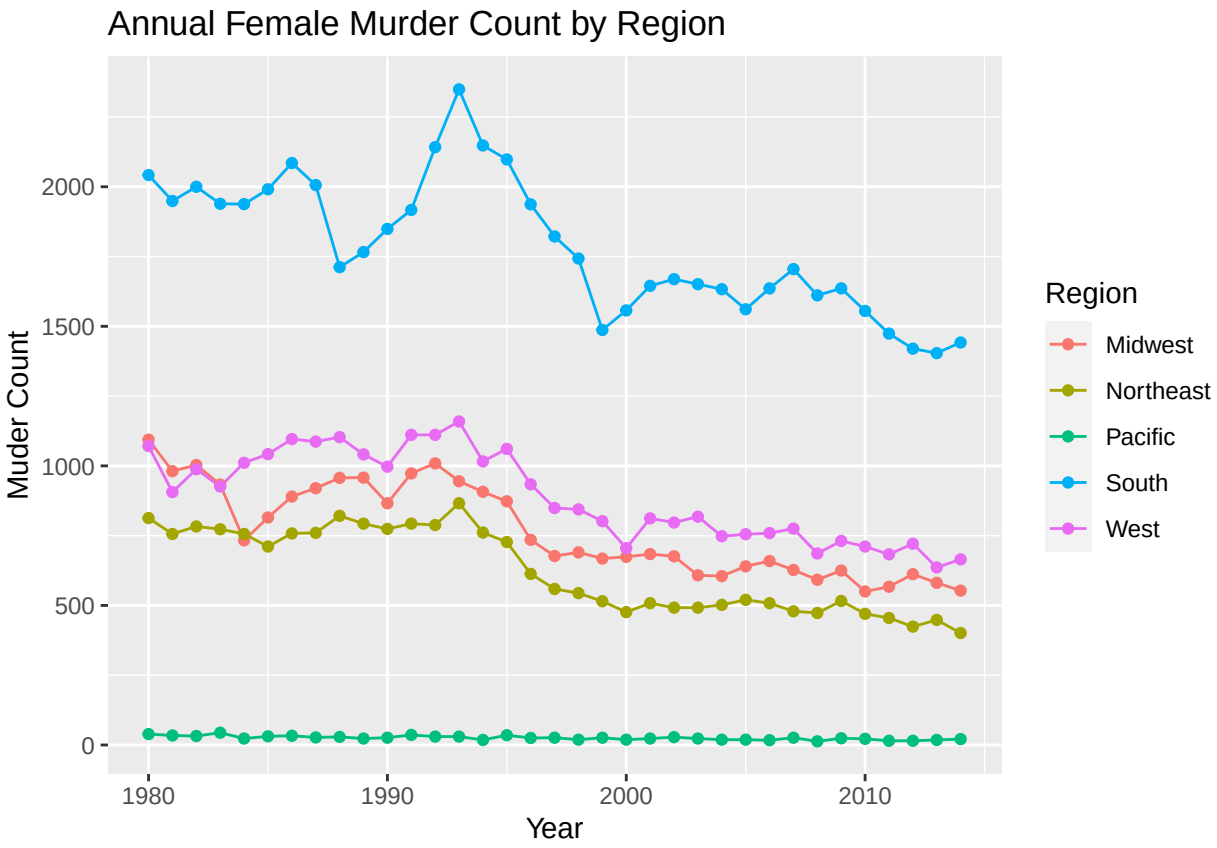


Average Female Murder Count by Region

Next I created a dataframe, categorizing the annual murder count of women by region for the years between 1980-2014. What I found appeared alarming at first. I seems the South has an overwhelming higher total female murder count:

Female Murder Count by Region

	West	Midwest	Northeast	South	Pacific	Total
Mean	890.2	768	623.65	1,786.25	25.37	818.6971
SD	160.84	163.9	149.09	241.81	7.23	591.7687
n	31,157	26,880	21,82	62,519	888	143,272




```
## [1] "Data Frame: female_only<-data.frame(Year = project_data$Year,R..."
```

```
## [1] " [female_only$Sex!='Male',] ...count(Sex,Region,Year)  "
```

Part 4 - Inference

The data above shows similar behavior with respect to mean age of murder count. It is also apparent that gender, impacts murder counts when comparing Males to Females. Now we will consider the impact Regions have on murder counts of cases of the same gender. I will infer based on age, by categorizing Females into 2 groups, those above 30 years of age and those below. The decision to split at 30 years, comes from looking at the National average 33.56, and my largest group set for age, the 20-30 bracket. The 2 regions I will use are the highest and lowest for my Annual "Female Murder Count by Region" line graph, the South and the Pacific.

South Inference

Inference Conditions

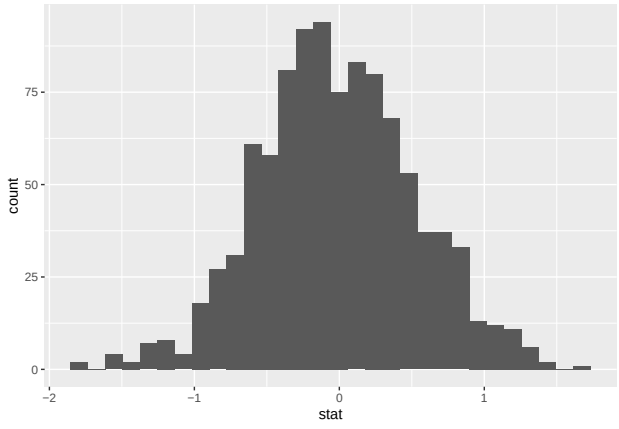
- Independence - Murder counts are assumed independent with 53072 observations✓
- Success-failure Conditions - $Count_{under\ 30}$=1079 & $Count_{over\ 30}$=2212 $\therefore n \geq 30$ ✓

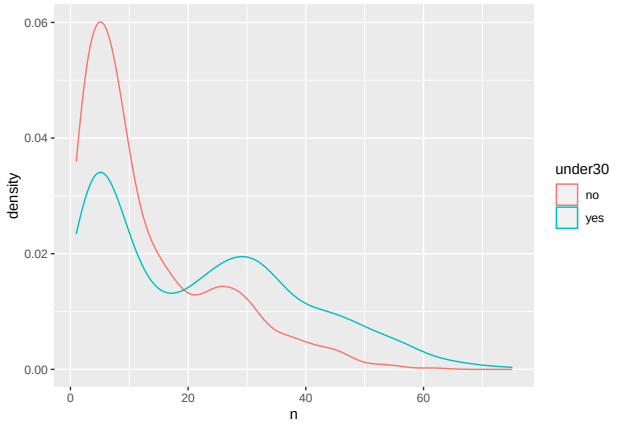
Hypothesis statement

$H_0 = South_{30yrs\ OR\ less} = 0$. *No Difference in average below or above* 30 *years for women*
$H_A = South_{30yrs\ OR\ less} \neq 0$. *There is a difference*

```
#New Variable Created 'under30'
south <- south %>%
  mutate(under30 = ifelse(south$`Victim Age` < 31, "yes", "no"))
```

Observation stat of 8.593917 generated & `hypothesize` function used to set null hypothesis as a test for independence. We have 0 null permutations.

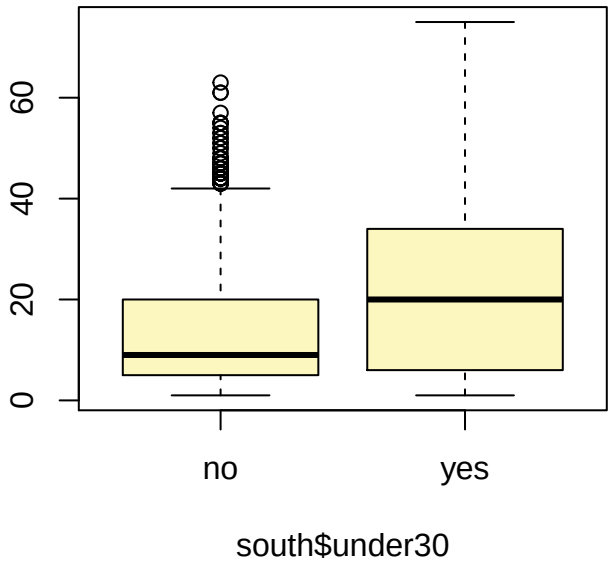





```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```r
inference(y = south$n, x = south$under30,est = "mean", type = "ci", method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_no = 2212, mean_no = 13.3088, sd_no = 11.7542
## n_yes = 1079, mean_yes = 21.9027, sd_yes = 16.8873
```




```
## Observed difference between means (no-yes) = -8.5939
##
## Standard error = 0.5716
## 95 % Confidence interval = ( -9.7143 , -7.4735 )
```

Pacific Inference

Inference Conditions

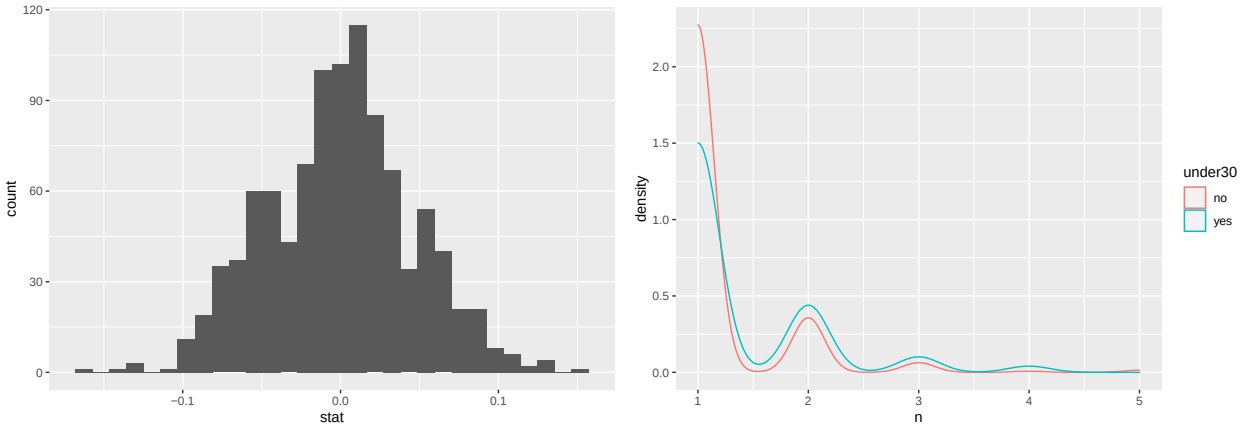
- Independence - Murder counts are assumed independent with 888 observations✓
- Success-failure Conditions - $Count_{under\ 30}$=308 & $Count_{over\ 30}$=387 $\therefore n \geq 30$ ✓

Hypothesis statement

$H_0 = Pacific_{30yrs\ OR\ less} = 0$. *No Difference in average below or above* 30 *years for women*
$H_A = Pacific_{30yrs\ OR\ less} \neq 0$. *There is a difference*

```
#New Variable Created 'under30'
pacific <- pacific %>%
  mutate(under30 = ifelse(pacific$`Victim Age` < 31, "yes", "no"))
```

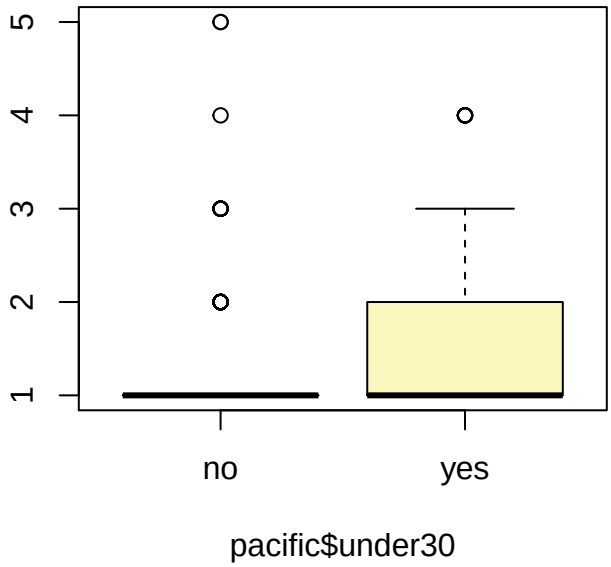
Observation stat of 0.1601648 generated & `hypothesize` function used to set null hypothesis as a test for independence. We have 0 null permutations.




```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```
inference(y = pacific$n, x = pacific$under30,est = "mean", type = "ci", method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_no = 387, mean_no = 1.2067, sd_no = 0.5374
## n_yes = 308, mean_yes = 1.3669, sd_yes = 0.6693
```




```
## Observed difference between means (no-yes) = -0.1602
##
## Standard error = 0.0469
## 95 % Confidence interval = ( -0.2521 , -0.0682 )
```

Part 5 - Conclusion

My conclusion is based on the following facts:

- Mean age on a national and a regional level hardly differs
- Region does not impact murder count of a gender by age
- Gender does impact murder count, but only when referencing male vs. female

Noting the higher murder count for the South vs. the Pacific, considering the Pacific is easily much smaller being composed of only 2 states, and the fact that both follow national trends indicating peak numbers occur for victims 20-30, I would say the difference is most likely because of population size. If I were to do a comparison of population of these regions during these years, I'd imagine the murder counts would be comparable. Therefore, my limitation with this data is not having the population size for each states, per year, between 1980-2014. Regardless the data does not highlight any specific region as having an abnormal murder count in my opinion. I can support that age is the biggest factor among those mentions.

References:

Homicide Reports, 1980-2014

Project https://www.kaggle.com/murderaccountability/homicide-reports?select=database.csv