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

DATA 606 Data Project Submission

Gabriel Campos

14 November 2020

```
## [1] " Attaching packages: tidyverse,openintro,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 focus specifically on murder rates for women under 30 against the entire data set.

Overview

The goals for this project are to:

- Think about the independent and dependent variables in correlation to murder rates including region, gender and age
- Compare murder rates by region, race, gender and age based on the data set
- Draw conclusions of the most at risk population by region for murder in the United States based on 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 qualifying toddlers 11 months old or younger 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

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   Year = col_double(),
##   Incident = col_double(),
##   `Victim Age` = col_double(),
##   `Perpetrator Age` = col_double(),
##   `Victim Count` = col_double(),
##   `Perpetrator Count` = col_double()
## )
```

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

```
## [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] "ifelse(State,Region...Default)) "
```

```
## [1] "Regions based on census.gov site : West, Midwest, Northeast, South, Pacific "
```

Census Map

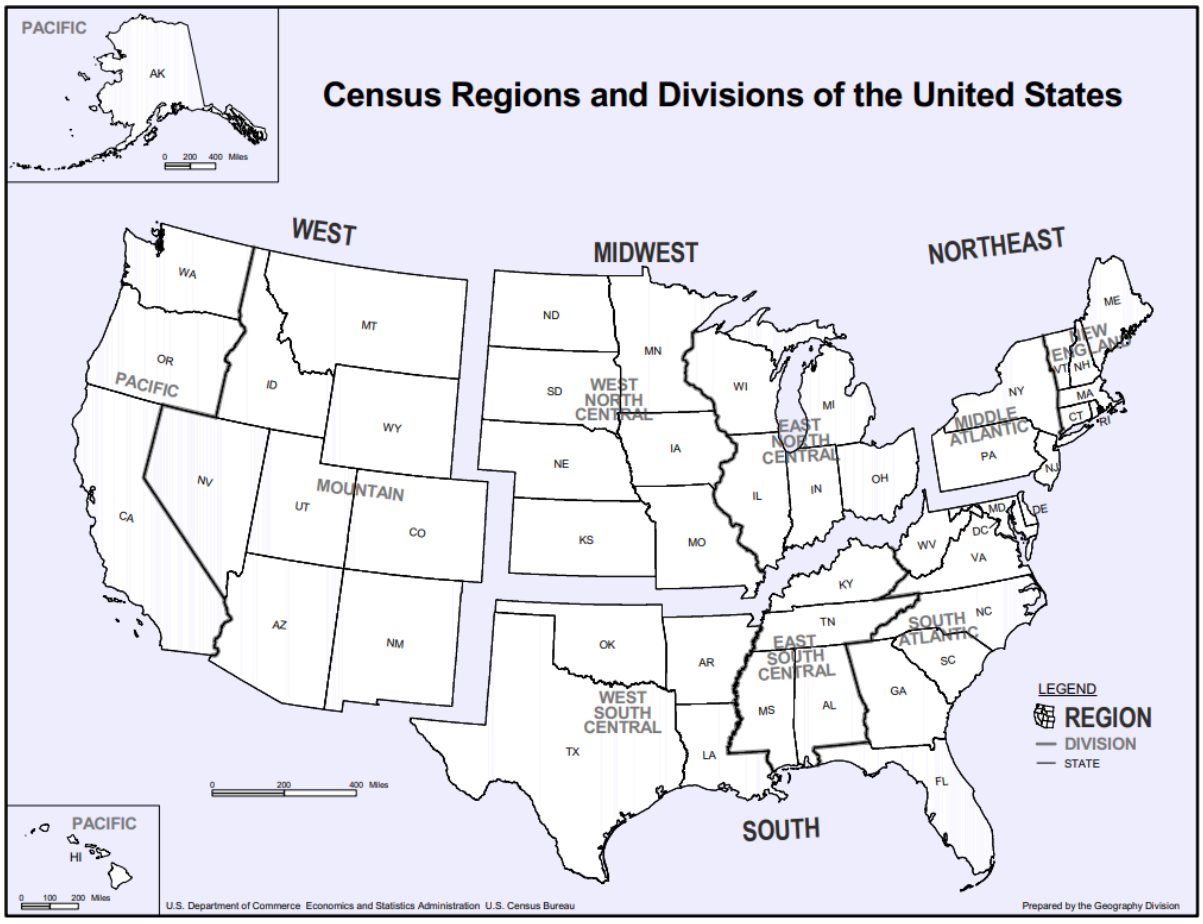


Figure 1: US regions based on census.gov

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

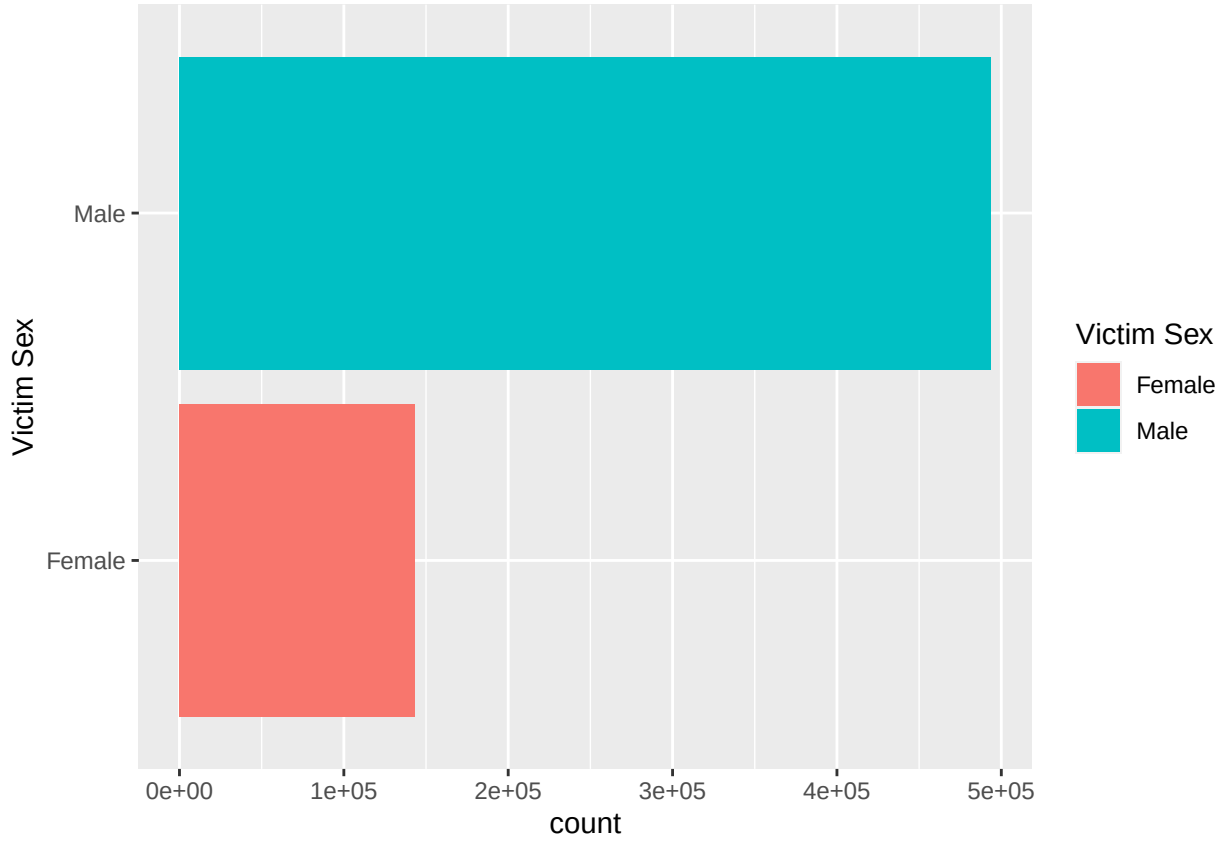
```
## count
##     0     1     2     3     4     5     6     7     8     9    10    11    12
##  8444  5525  3805  2378  1659  1194   999   915   852   834   854   911  1239
##    13    14    15    16    17    18    19    20    21    22    23    24    25
##  1897  3342  5905  9402 14030 18469 21939 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),]"
```

```
## [1] "Summary of age"
```

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



Part 3 - Exploratory data analysis

Part 4 - Inference

Part 5 - Conclusion

References:

Homicide Reports, 1980-2014

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

Appendix (optional):

Remove this section if you don't have an appendix