DATA 606 Data Project Proposal

Gabriel Campos

Data Preparation

```
west <- project_data %>%
  filter(State == "Arizona"| State == "California"| State =="Colorado"
       | State =="Idaho"| State =="Montana"| State =="Nevada"
       | State == "New Mexico"| State =="Oregon" | State == "Utah"
       | State =="Washington"| State =="Wyoming")
midwest <- project_data %>%
  filter(State == "Illinois"| State == "Indiana"| State =="Iowa"
       | State =="Kansas"| State =="Michigan"| State =="Minnesota"
       | State == "Missouri"| State =="Nebraska"| State == "North Dakota"
       | State =="Ohio"| State =="South Dakota"| State =="Wisconsin")
northeast <- project_data %>%
  filter(State == "Connecticut"| State == "Delaware"| State =="Maine"
       | State =="Massachusetts"| State =="New Hampshire"
       | State =="New Jersey"| State == "New York"| State =="Pennsylvania"
       | State == "Rhode Island"| State =="Vermont")
south <- project_data %>%
  filter(State == "Alabama"| State == "Arkansas"| State =="Florida"
       | State =="Georgia"| State =="Kentucky"| State =="Louisiana"
       | State == "Mississippi"| State =="North Carolina"| State == "Oklahoma"
       | State =="South Carolina"| State =="South Carolina"| State =="Tennessee"
       | State =="Texas"| State =="Virgina"| State =="West Virgina")
pacific<- project_data %>%
  filter(State == "Alaska"| State == "Hawaii")
```

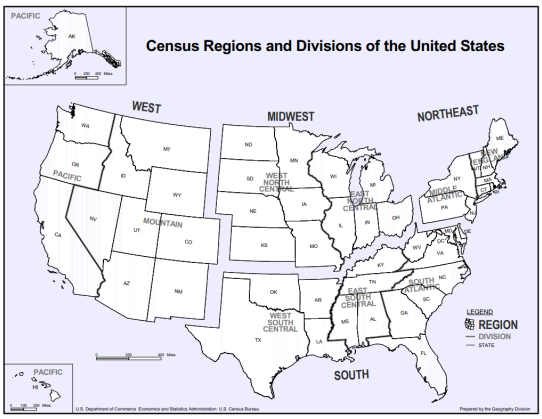


Figure 1: US regions based on census.gov

Research question

Question: Between 1980-2014, was the South the most dangerous US region for women to live in?

$H_0 = Region_{south.30yrs.OR.less} \leq Region_{(west...midwest...northeast...pacific).30yrs.OR.less}$
$H_A = Region_{south.30yrs.OR.less} > Region_{(west...midwest...northeast...pacific).30yrs.OR.less}$

Cases

What are the cases, and how many are there?

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 (**Derived by creating sub datasets NOTE CHUNK** $r\ regional - subsets$)

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

Outliers: 998

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

```
project_data %>% filter(`Victim Sex` == "Female" & `Victim Age`< 31) %>% glimpse()
```

Data collection

Describe the method of data collection.

Data source: kaggle.com

```
Content
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.

Acknowledgements
The data was compiled and made available by the Murder Accountability Project, founded by
Thomas Hargrove.
```

Homicide Reports, 1980-2014 Project https://www.kaggle.com/murderaccountability/homicide-reports?select=database.csv

Type of study

What type of study is this (observational/experiment)?

Observational

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

Homicide Reports, 1980-2014 Project https://www.kaggle.com/murderaccountability/homicide-reports?select=database.csv

Dependent Variable

```
##  [1] "Record ID"             "Agency Code"           "Agency Name"
##  [4] "Agency Type"           "City"                  "State"
##  [7] "Year"                  "Month"                 "Incident"
## [10] "Crime Type"            "Crime Solved"          "Victim Sex"
## [13] "Victim Age"            "Victim Race"           "Victim Ethnicity"
## [16] "Perpetrator Sex"       "Perpetrator Age"       "Perpetrator Race"
## [19] "Perpetrator Ethnicity" "Relationship"          "Weapon"
## [22] "Victim Count"          "Perpetrator Count"     "Record Source"
```

What is the response variable? Is it quantitative or qualitative?

Murder count, which is counted with Record ID and is quantitative.

Independent Variable

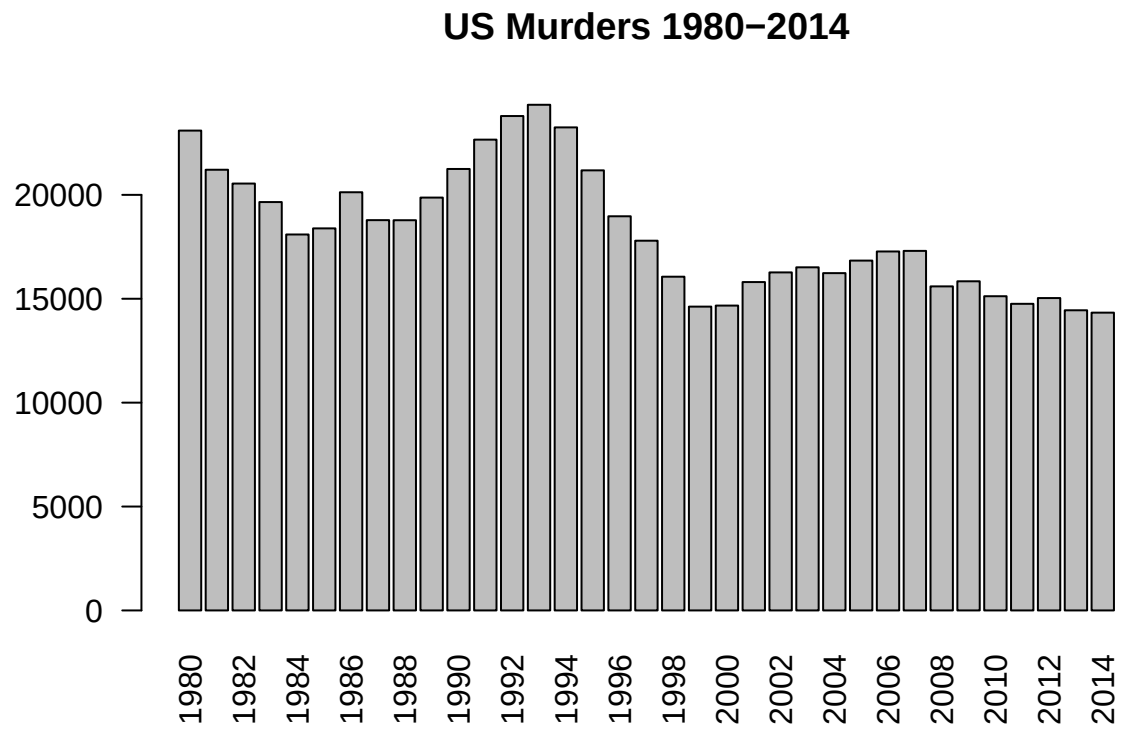
You should have two independent variables, one quantitative and one qualitative.
Victim Age (Quantitative) and Victim Sex (Qualitative)

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

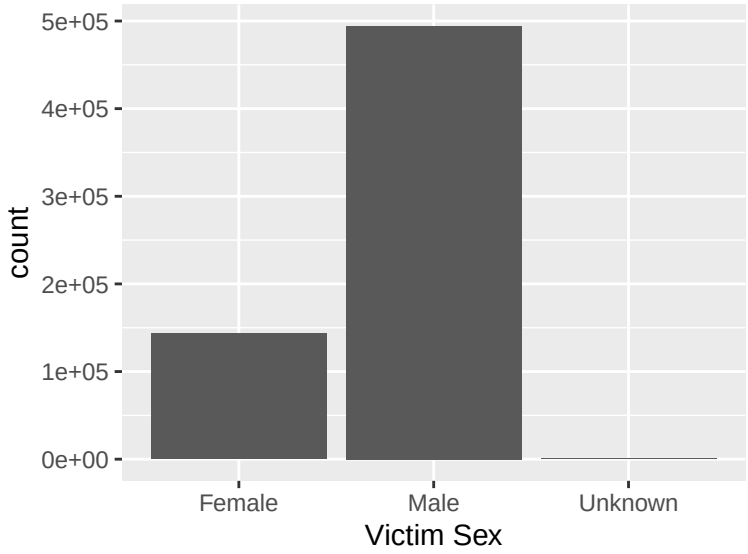
```
glimpse(project_data)
```

There were 638,454 total murders between 1980-2014



143,345 of the murders were specifically against Women

```
project_data %>% filter(`Victim Sex` == "Female") %>% glimpse()
```



A break down of murders by age are as follows *Note : age 998 was assigned to victims who's age could not be determined*

```
## 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
```

```
labs <- c("1-10","11-20","21-30","31-40","41-50","51-60","61-70","71-80","81-90","91-100")
#labs
project_data$`Victim AgeGroup`<- cut(project_data$`Victim Age`,
                                     breaks=c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100),
                                     right = FALSE)
ggplot(project_data)+geom_bar(aes(x=`Victim AgeGroup`))
```