US Census Dataset

LA - 2 Exploratory Data Analysis

Submitted By:- Shivee Jaiswal,1NT21IS152, 5 C

Introduction

The U.S. Census Dataset is a comprehensive collection of demographic, economic, and social data gathered by the United States Census Bureau. Conducted decennially, the census provides a detailed snapshot of the American population and serves as a crucial resource for policymakers, researchers, and businesses. The dataset covers a wide range of variables, including population counts, age distribution, racial and ethnic demographics, housing characteristics, income levels, educational attainment, and employment statistics. The census data plays a vital role in shaping public policies, resource allocation, and political representation. Additionally, it facilitates in-depth analyses of societal trends, disparities, and the overall well-being of the U.S. population. As a foundational tool for understanding the dynamics of American society, the U.S. Census Dataset contributes significantly to informed decision-making and informed social and economic research.

Loading Libraries

```
set.seed(123)

# Data manipulation

library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.3.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
library(DT)
```

```
## Warning: package 'DT' was built under R version 4.3.2
```

```
# Time manipulation

library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.3.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

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

```
# Visualization

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(plotrix)
```

```
## Warning: package 'plotrix' was built under R version 4.3.2
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```

```
library(ggdendro)
```

```
## Warning: package 'ggdendro' was built under R version 4.3.2
```

```
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.3.2
```

```
# Wordcloud

library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 4.3.2
```

```
## Loading required package: RColorBrewer
```

```
# Text manipulation

library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 4.3.2
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.3.2
```

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 4.3.2
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(sentimentr)
```

```
## Warning: package 'sentimentr' was built under R version 4.3.2
```

```
library(wordcloud)

library(RSentiment)
```

```
## Warning: package 'RSentiment' was built under R version 4.3.2
```

```
# Load the Dataset
df <- read.csv("C:\\Users\\Shivee Jaiswal\\Desktop\\acs2015_census_tract_data.csv")
# Data Summary
summary(df)
```

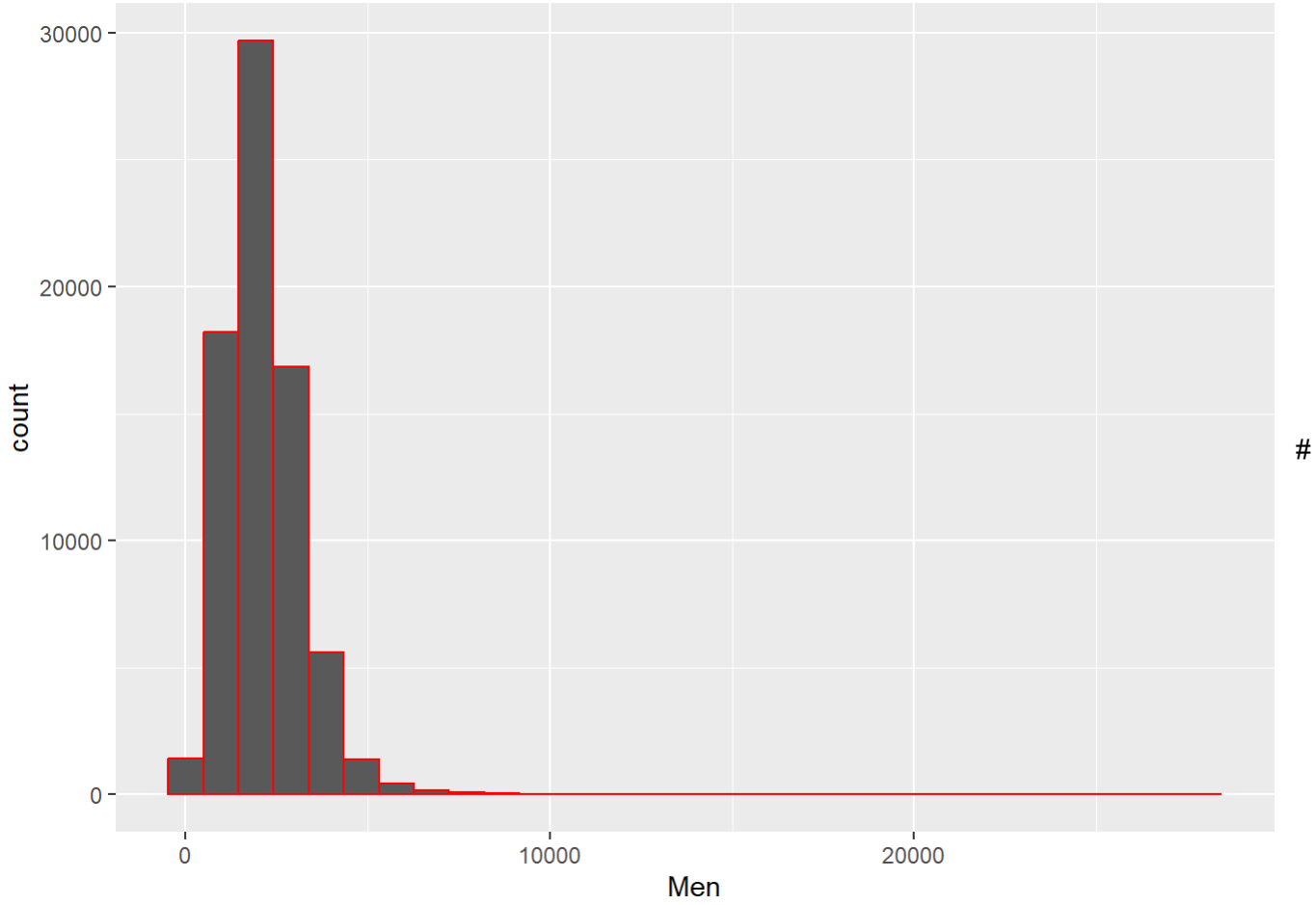
```
##   CensusTract           State              County             TotalPop
##  Min.   :1.001e+09   Length:74001       Length:74001       Min.   :    0
##  1st Qu.:1.304e+10   Class :character   Class :character   1st Qu.: 2891
##  Median :2.805e+10   Mode  :character   Mode  :character   Median : 4063
##  Mean   :2.839e+10                                         Mean   : 4326
##  3rd Qu.:4.200e+10                                         3rd Qu.: 5442
##  Max.   :7.215e+10                                         Max.   :53812
##
##       Men             Women           Hispanic          White
##  Min.   :    0   Min.   :    0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 1409   1st Qu.: 1461   1st Qu.:  2.40   1st Qu.: 39.40
##  Median : 1986   Median : 2066   Median :  7.00   Median : 71.40
##  Mean   : 2128   Mean   : 2198   Mean   : 16.86   Mean   : 62.03
##  3rd Qu.: 2674   3rd Qu.: 2774   3rd Qu.: 20.40   3rd Qu.: 88.30
##  Max.   :27962   Max.   :27250   Max.   :100.00   Max.   :100.00
##                                  NA's   :690      NA's   :690
##      Black            Native             Asian           Pacific
##  Min.   :  0.00   Min.   :  0.0000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:  0.70   1st Qu.:  0.0000   1st Qu.: 0.200   1st Qu.: 0.000
##  Median :  3.70   Median :  0.0000   Median : 1.400   Median : 0.000
##  Mean   : 13.27   Mean   :  0.7277   Mean   : 4.588   Mean   : 0.145
##  3rd Qu.: 14.40   3rd Qu.:  0.4000   3rd Qu.: 4.800   3rd Qu.: 0.000
##  Max.   :100.00   Max.   :100.0000   Max.   :91.300   Max.   :84.700
##  NA's   :690      NA's   :690        NA's   :690      NA's   :690
##     Citizen           Income         IncomeErr       IncomePerCap
##  Min.   :    0   Min.   :  2611   Min.   :   390   Min.   :   128
##  1st Qu.: 2037   1st Qu.: 37683   1st Qu.:  5317   1st Qu.: 19123
##  Median : 2863   Median : 51094   Median :  7732   Median : 25344
##  Mean   : 3043   Mean   : 57226   Mean   :  9134   Mean   : 28491
##  3rd Qu.: 3838   3rd Qu.: 70117   3rd Qu.: 11258   3rd Qu.: 33894
##  Max.   :37416   Max.   :248750   Max.   :123116   Max.   :254204
##                  NA's   :1100     NA's   :1100     NA's   :740
##  IncomePerCapErr     Poverty        ChildPoverty     Professional
##  Min.   :    85   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  2312   1st Qu.:  7.20   1st Qu.:  7.00   1st Qu.: 24.10
##  Median :  3127   Median : 13.40   Median : 17.80   Median : 32.60
##  Mean   :  3943   Mean   : 16.96   Mean   : 22.49   Mean   : 34.80
##  3rd Qu.:  4537   3rd Qu.: 23.10   3rd Qu.: 33.80   3rd Qu.: 43.88
##  Max.   :134380   Max.   :100.00   Max.   :100.00   Max.   :100.00
##  NA's   :740      NA's   :835      NA's   :1118     NA's   :807
##     Service           Office        Construction       Production
##  Min.   :  0.0   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
##  1st Qu.: 13.4   1st Qu.: 20.10   1st Qu.:  5.000   1st Qu.:  7.10
##  Median : 17.9   Median : 23.80   Median :  8.400   Median : 11.80
##  Mean   : 19.1   Mean   : 23.95   Mean   :  9.292   Mean   : 12.86
##  3rd Qu.: 23.6   3rd Qu.: 27.50   3rd Qu.: 12.500   3rd Qu.: 17.40
##  Max.   :100.0   Max.   :100.00   Max.   :100.000   Max.   :100.00
##  NA's   :807     NA's   :807      NA's   :807       NA's   :807
##      Drive            Carpool            Transit             Walk
##  Min.   :  0.00   Min.   :  0.000   Min.   :  0.000   Min.   :  0.000
##  1st Qu.: 72.00   1st Qu.:  6.000   1st Qu.:  0.000   1st Qu.:  0.400
##  Median : 79.70   Median :  8.800   Median :  1.100   Median :  1.400
##  Mean   : 75.53   Mean   :  9.627   Mean   :  5.456   Mean   :  3.123
##  3rd Qu.: 84.90   3rd Qu.: 12.300   3rd Qu.:  4.700   3rd Qu.:  3.500
##  Max.   :100.00   Max.   :100.000   Max.   :100.000   Max.   :100.000
```

```
##  NA's   :797      NA's   :797       NA's   :797       NA's   :797
##   OtherTransp        WorkAtHome       MeanCommute       Employed
##  Min.   :  0.000   Min.   :  0.000   Min.   : 1.20   Min.   :    0
##  1st Qu.:  0.400   1st Qu.:  1.800   1st Qu.:20.80   1st Qu.: 1249
##  Median :  1.100   Median :  3.500   Median :25.00   Median : 1846
##  Mean   :  1.892   Mean   :  4.368   Mean   :25.67   Mean   : 1984
##  3rd Qu.:  2.500   3rd Qu.:  5.900   3rd Qu.:29.80   3rd Qu.: 2553
##  Max.   :100.000   Max.   :100.000   Max.   :80.00   Max.   :24075
##  NA's   :797       NA's   :797       NA's   :949
##   PrivateWork       PublicWork      SelfEmployed       FamilyWork
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   : 0.0000
##  1st Qu.: 74.60   1st Qu.:  9.60   1st Qu.:  3.500   1st Qu.: 0.0000
##  Median : 80.10   Median : 13.40   Median :  5.500   Median : 0.0000
##  Mean   : 78.98   Mean   : 14.62   Mean   :  6.234   Mean   : 0.1698
##  3rd Qu.: 84.60   3rd Qu.: 18.20   3rd Qu.:  8.100   3rd Qu.: 0.0000
##  Max.   :100.00   Max.   :100.00   Max.   :100.000   Max.   :26.5000
##  NA's   :807      NA's   :807      NA's   :807       NA's   :807
##   Unemployment
##  Min.   :  0.000
##  1st Qu.:  5.100
##  Median :  7.700
##  Mean   :  9.029
##  3rd Qu.: 11.400
##  Max.   :100.000
##  NA's   :802
```

Histogram Graph

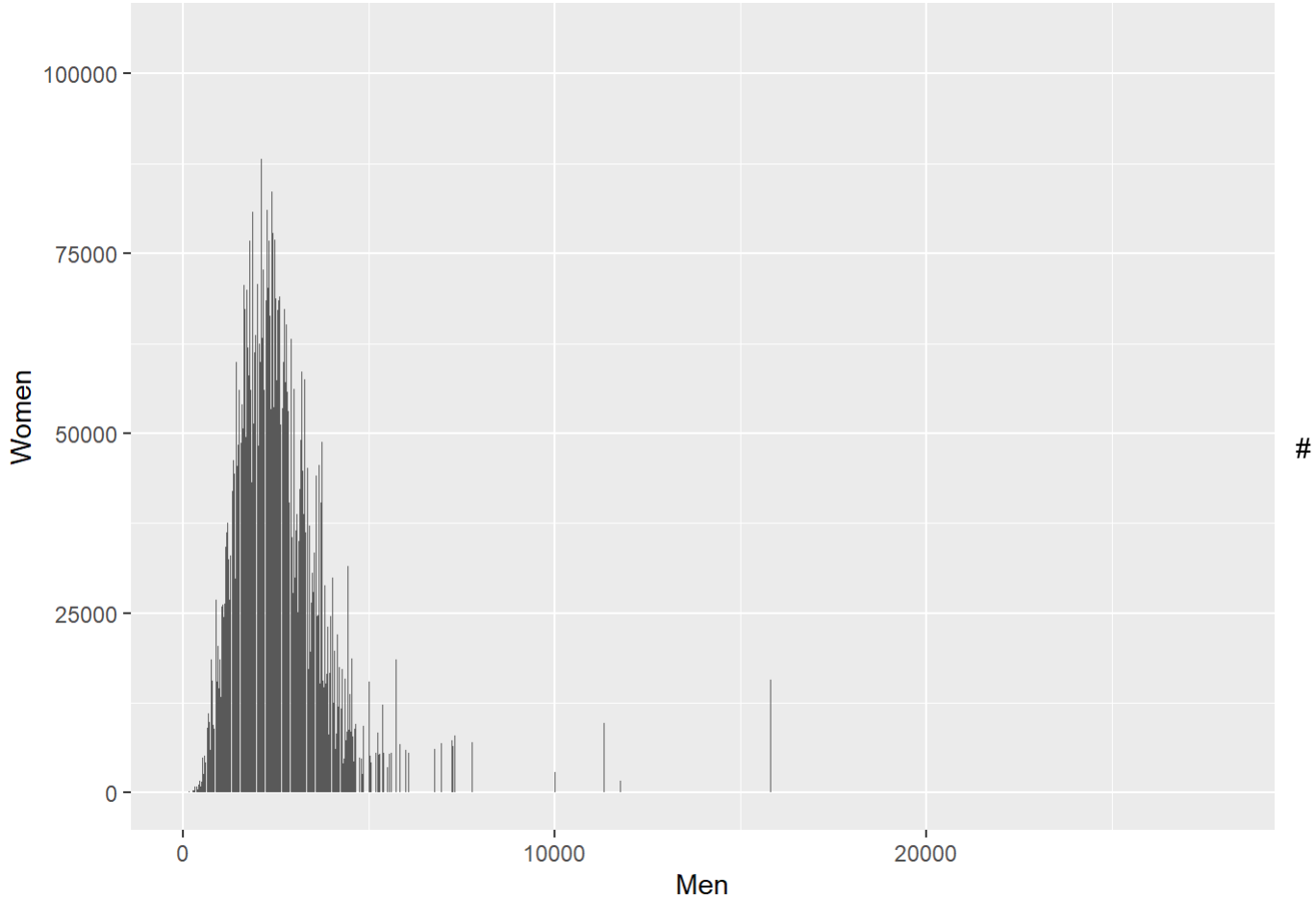
```
library(ggplot2)
ggplot(df, aes(x=Men)) +
  geom_histogram(colour="red")
```

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



Frequency graph

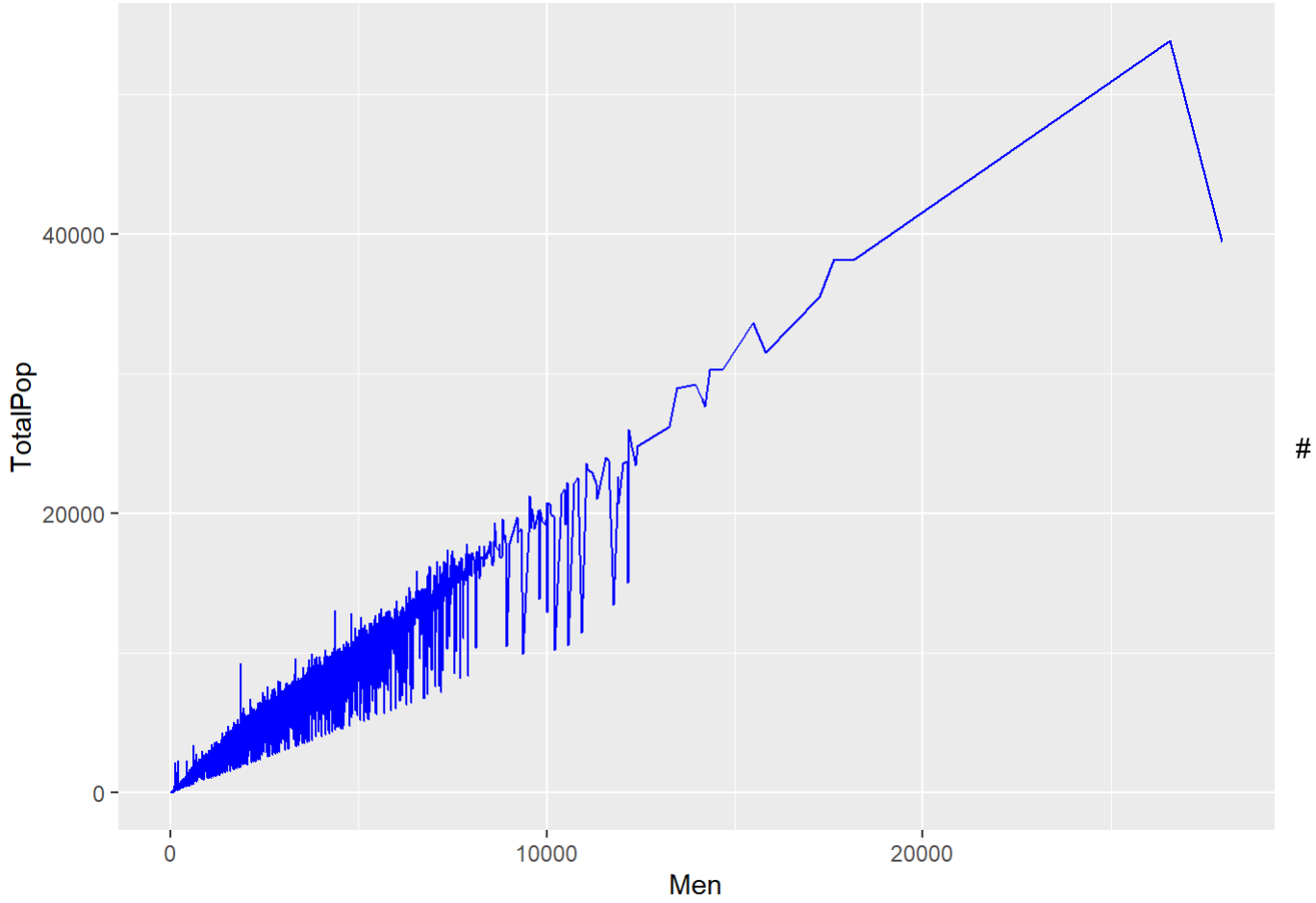
```
ggplot(df, aes(x=Men, y=Women)) +
  geom_col()
```



#

Line Graph

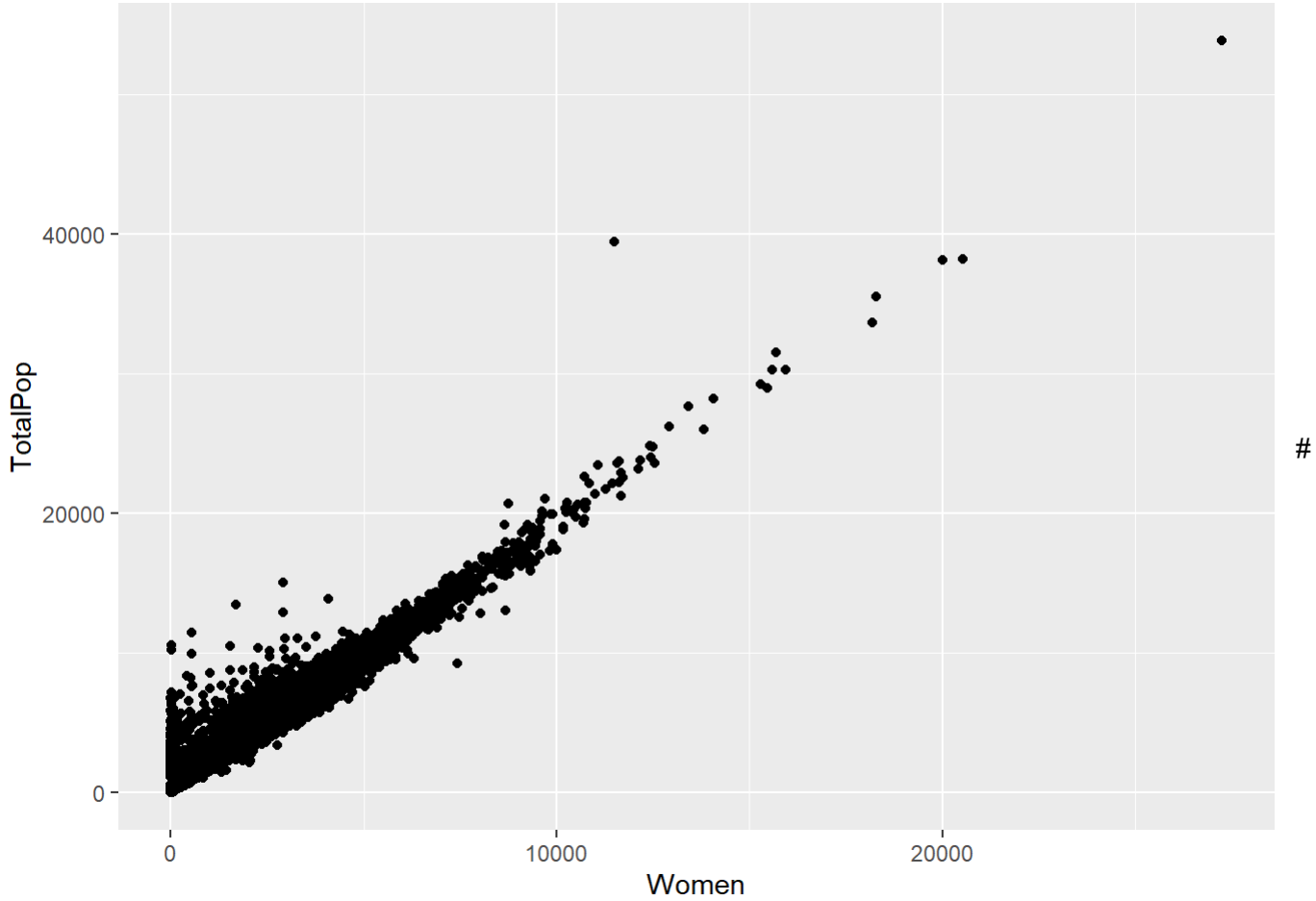
```
ggplot(df, aes(x=Men, y=TotalPop)) +
  geom_line(colour="blue")
```



#

Point graph

```
ggplot(df, aes(x=Women, y=TotalPop)) +
  geom_point()
```



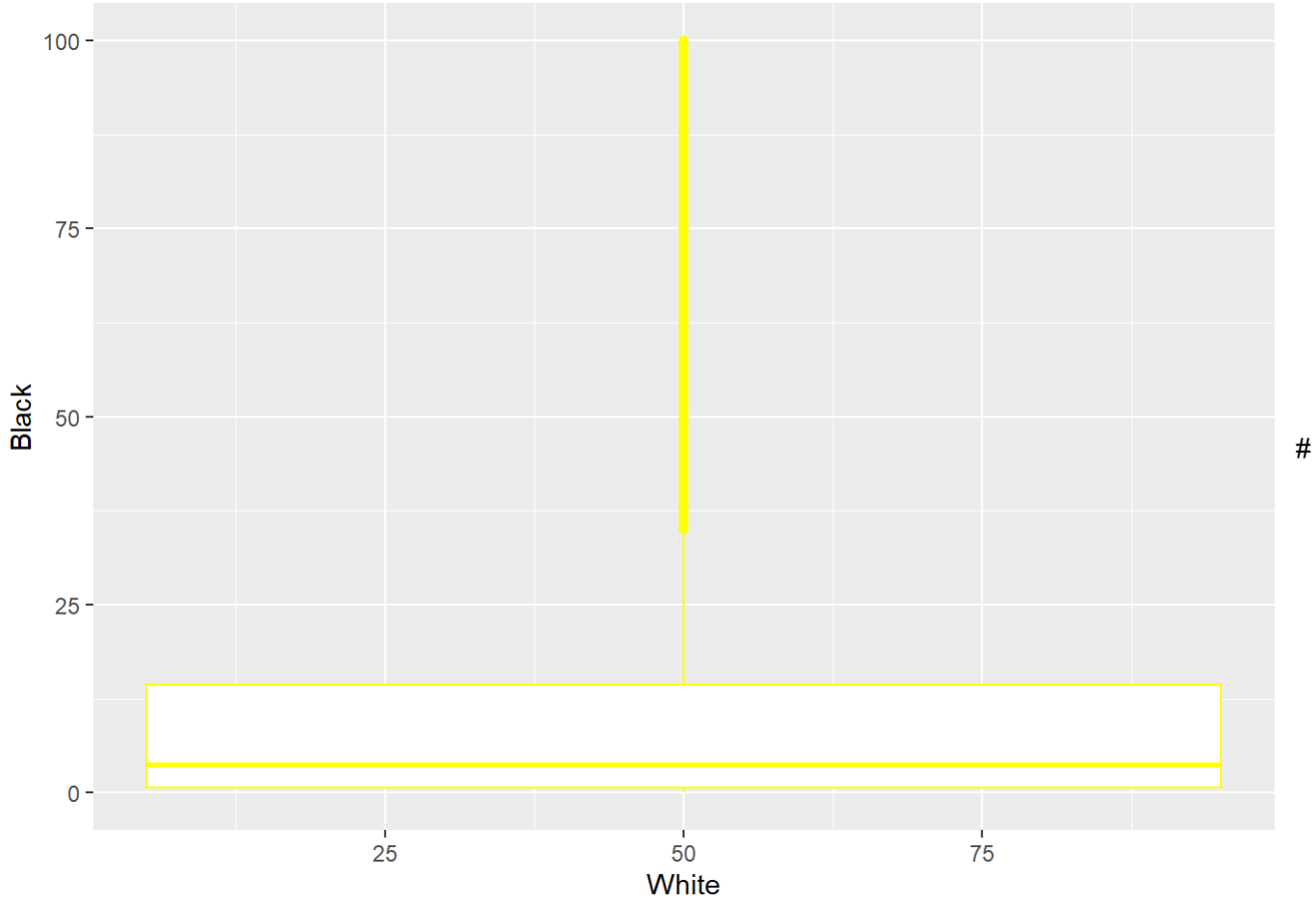
#

Boxplot

```
ggplot(df, aes(x=White, y=Black)) +
  geom_boxplot(colour="yellow")
```

```
## Warning: Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?
```

```
## Warning: Removed 690 rows containing missing values (`stat_boxplot()`).
```

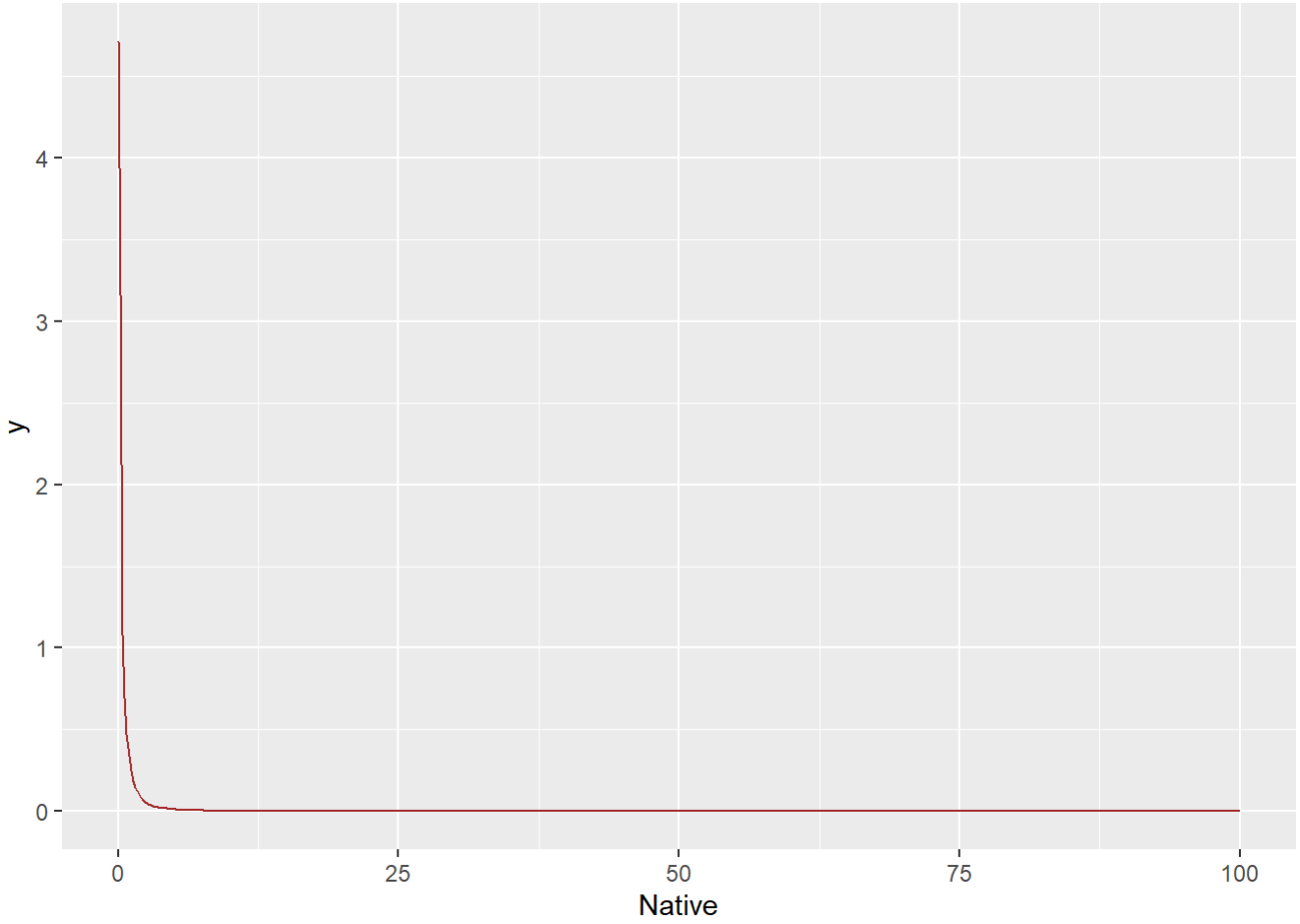


#

Line graph

```
ggplot(df, aes(x=Native)) +
  geom_line(stat="density", colour="brown") +
  expand_limits(y=0)
```

```
## Warning: Removed 690 rows containing non-finite values (`stat_density()`).
```



#

Density graph

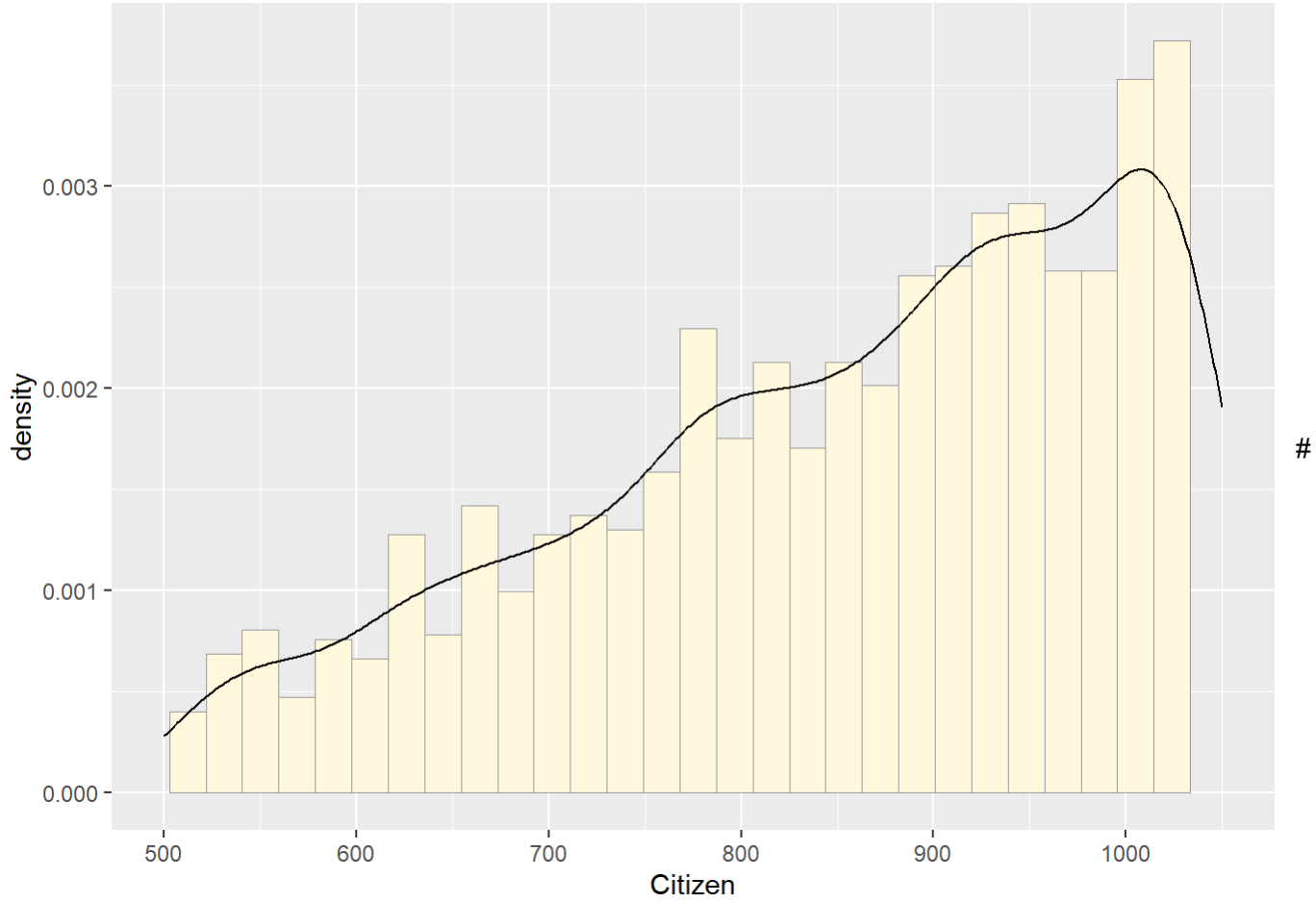
```
ggplot(df, aes(x=Citizen, y=after_stat(density))) +
  geom_histogram(fill="cornsilk", colour="grey60", linewidth=.2) +
  geom_density() +
  xlim(500, 1050)
```

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

```
## Warning: Removed 71774 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 71774 rows containing non-finite values (`stat_density()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



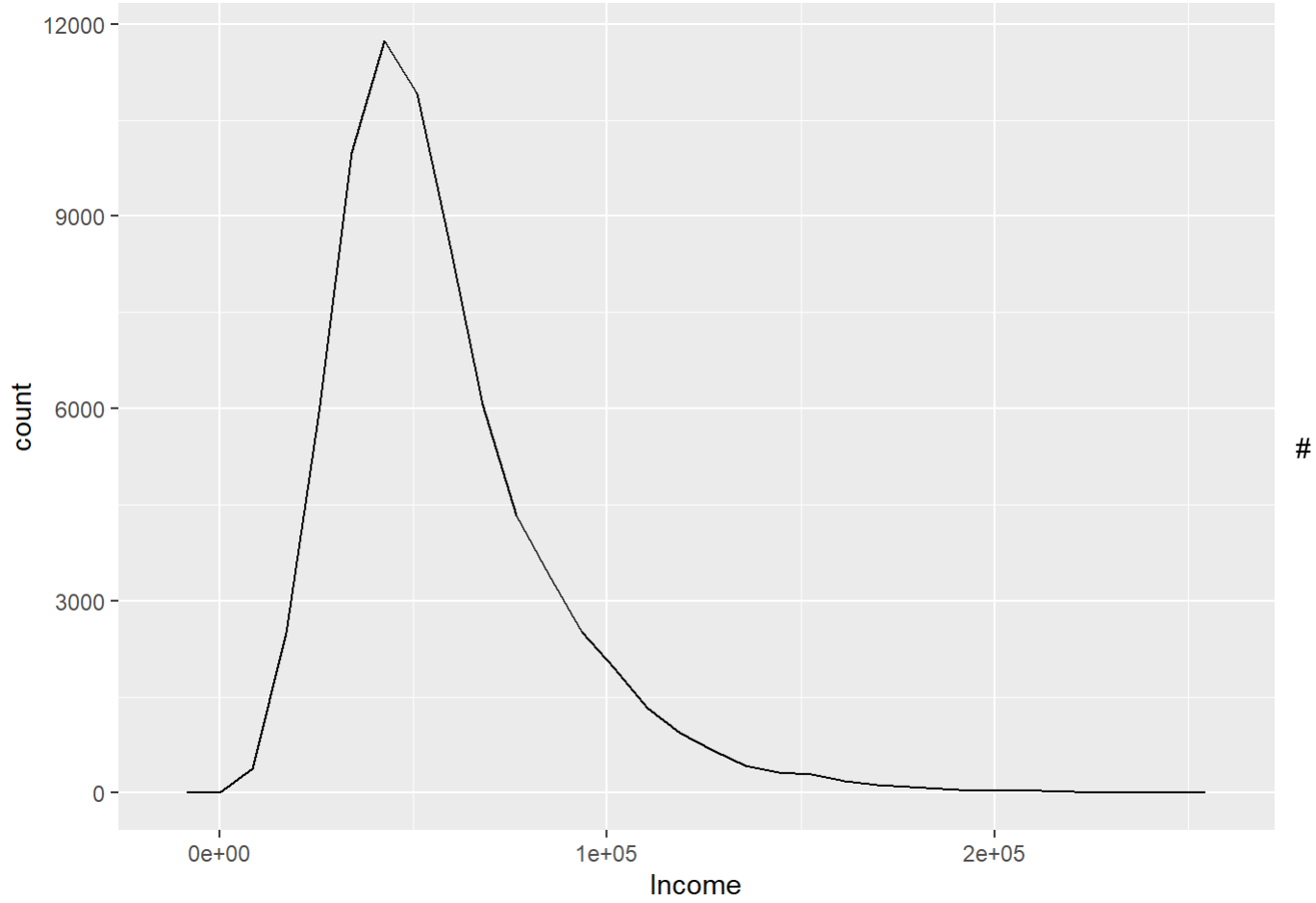
#

Frequency polygon graph

```
ggplot(df, aes(x=Income)) +
  geom_freqpoly()
```

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

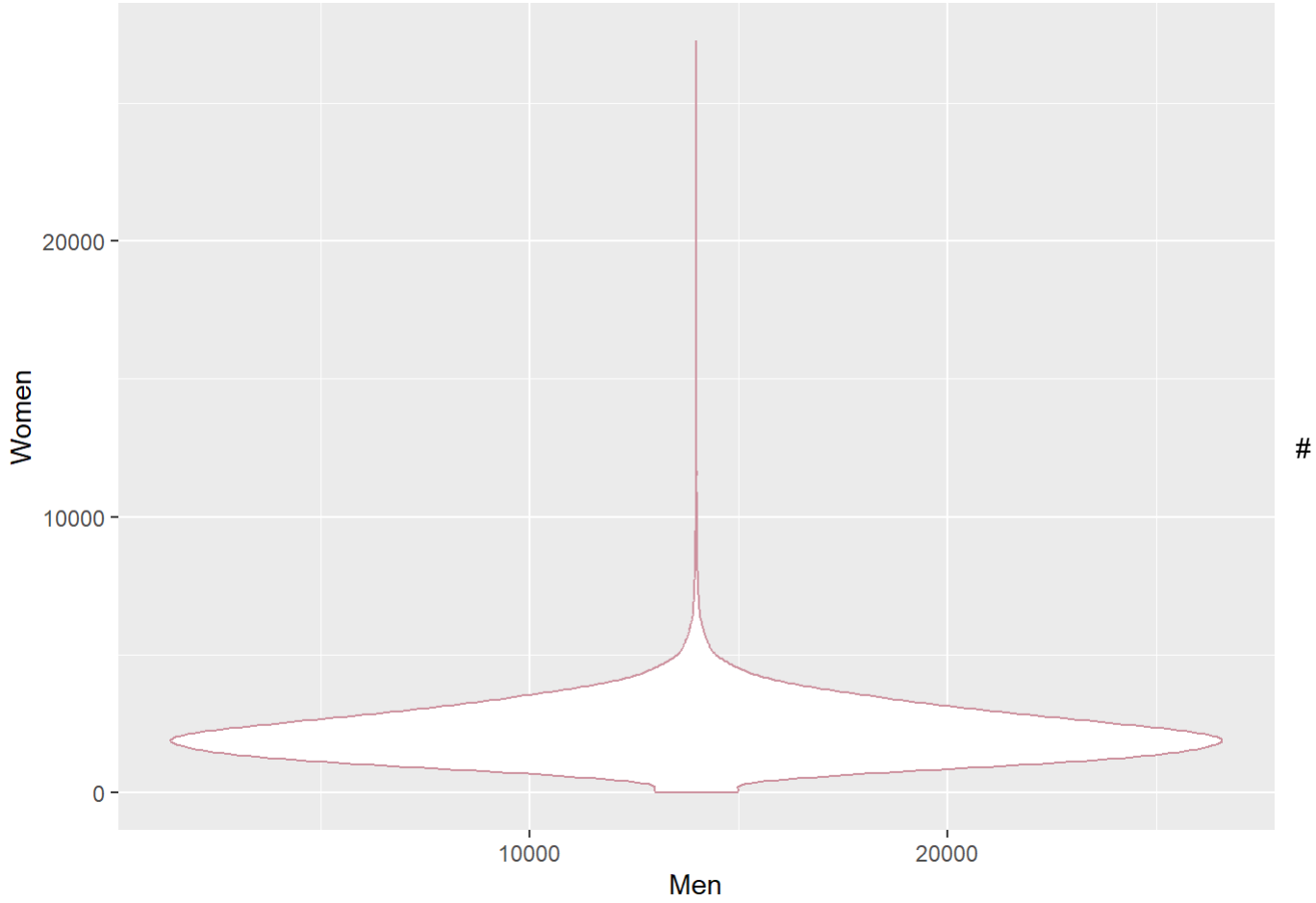
```
## Warning: Removed 1100 rows containing non-finite values (`stat_bin()`).
```



#

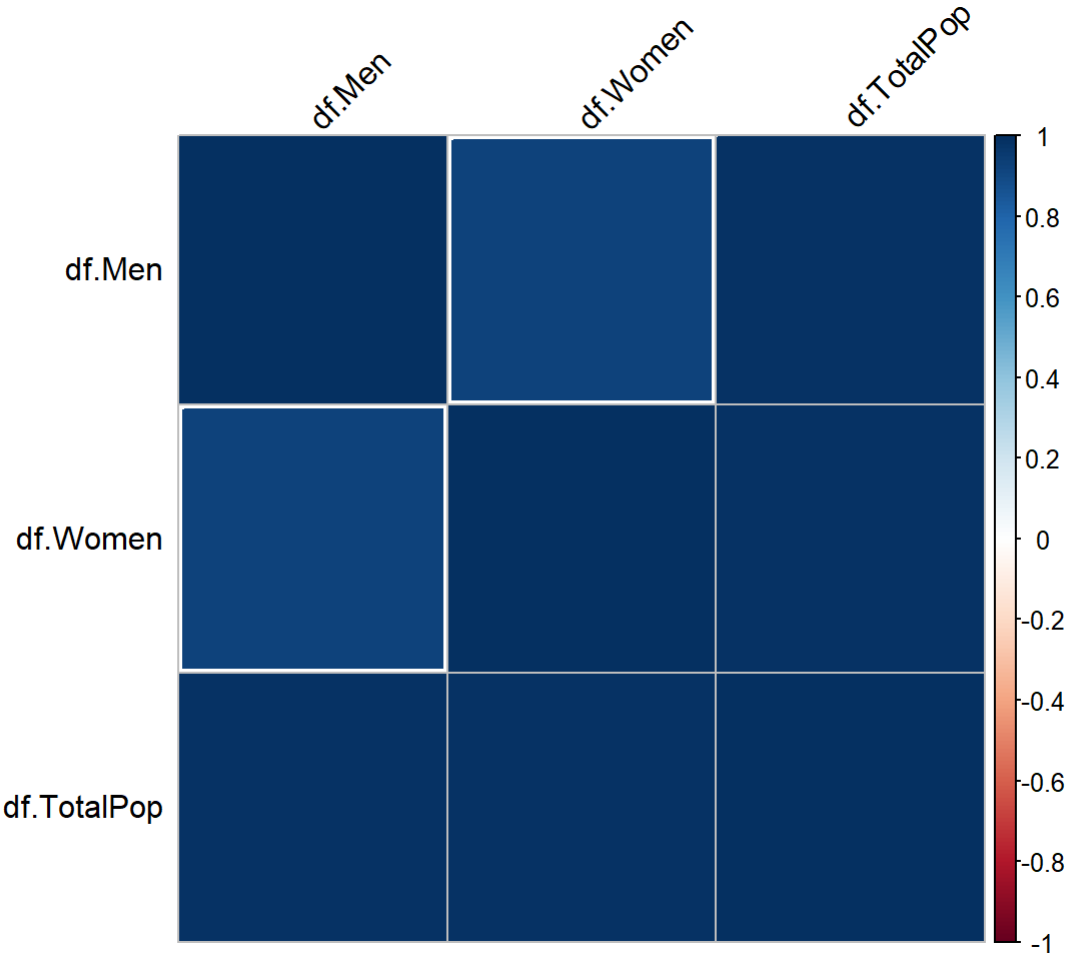
Violin graph

```
library(gcookbook)
df_p <- ggplot(df, aes(x= Men, y=Women))
df_p + geom_violin(adjust=2, colour="pink3")
```

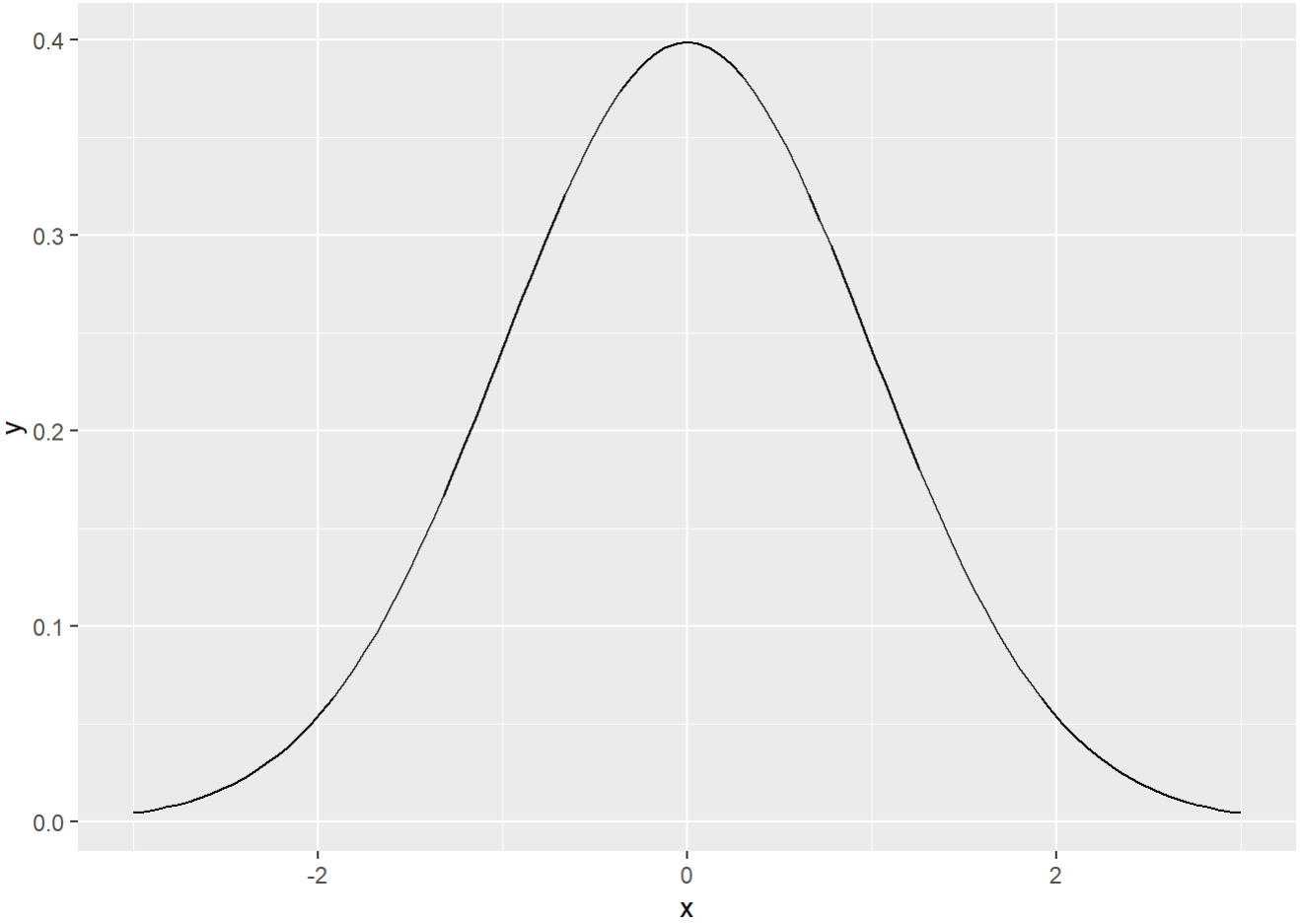


Correlation graph

```
library(corrplot)
dfcorr <- data.frame(df$Men, df$Women, df$TotalPop)
dfcorr <- cor(dfcorr)
corrplot(dfcorr, method="square", shade.col= NA, tl.col="black", tl.srt=45)
```




```
p <- ggplot(data.frame(x=c(-3,3)), aes(x=x))
p + stat_function(fun=dnorm)
```



#

Network graph

```
library(igraph)
```

```
## Warning: package 'igraph' was built under R version 4.3.2
```

```
##
## Attaching package: 'igraph'
```

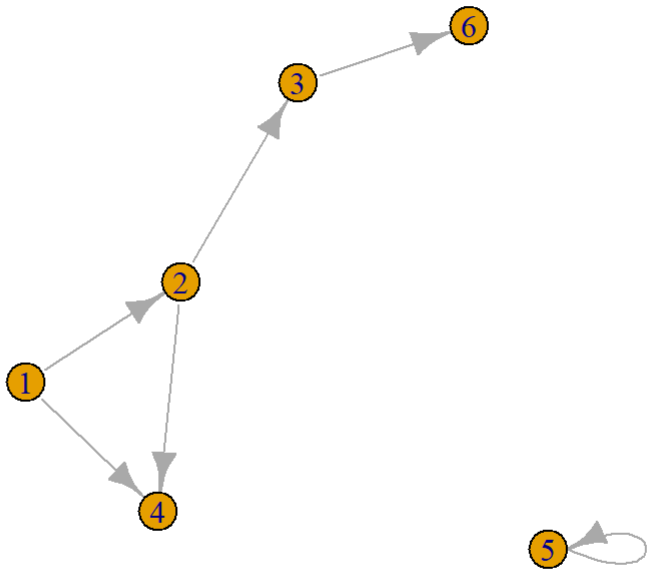
```
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

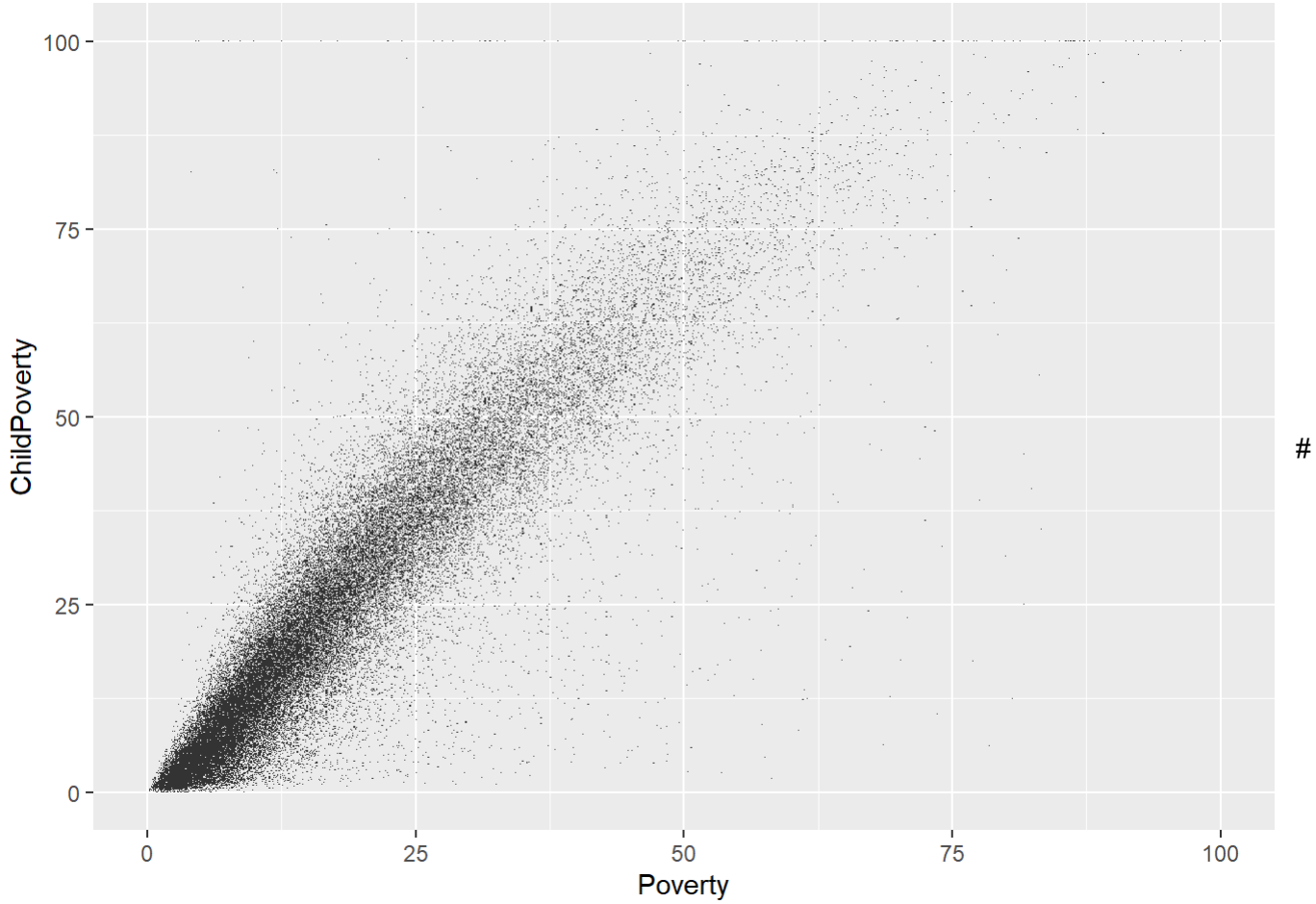
```
## The following object is masked from 'package:base':
##
##     union
```

```
gd <- graph(c(1,2,2,3,2,4,1,4,5,5,3,6))
plot(gd)
```




```
p <- ggplot(df, aes(x=Poverty, y=ChildPoverty))
p + geom_tile()
```

```
## Warning: Removed 1118 rows containing missing values (`geom_tile()`).
```

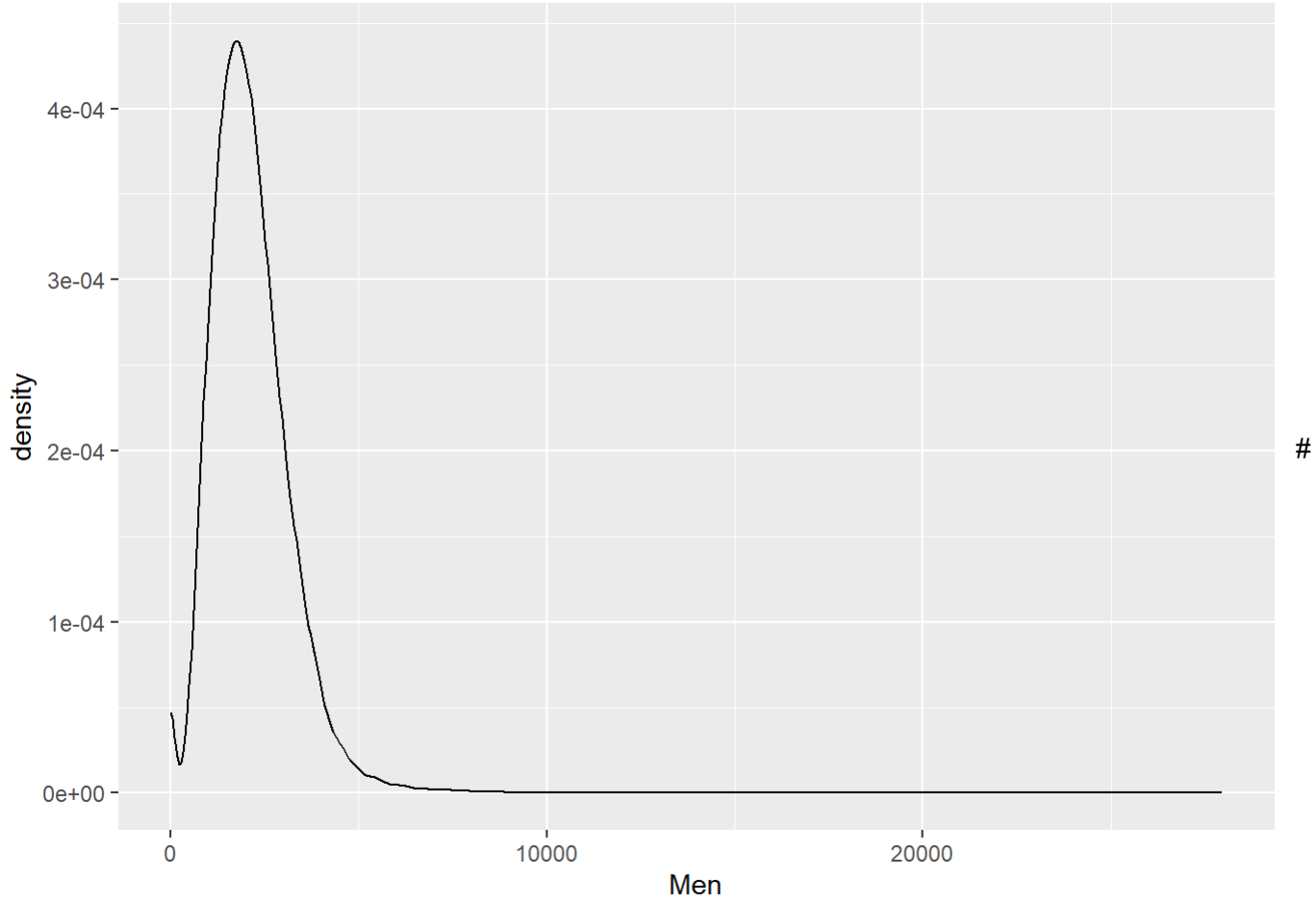


#

Density graph

```
library(ggplot2)

ggplot(df, aes(x = Men, fill = Women)) +
  geom_density(alpha = 0.5)
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



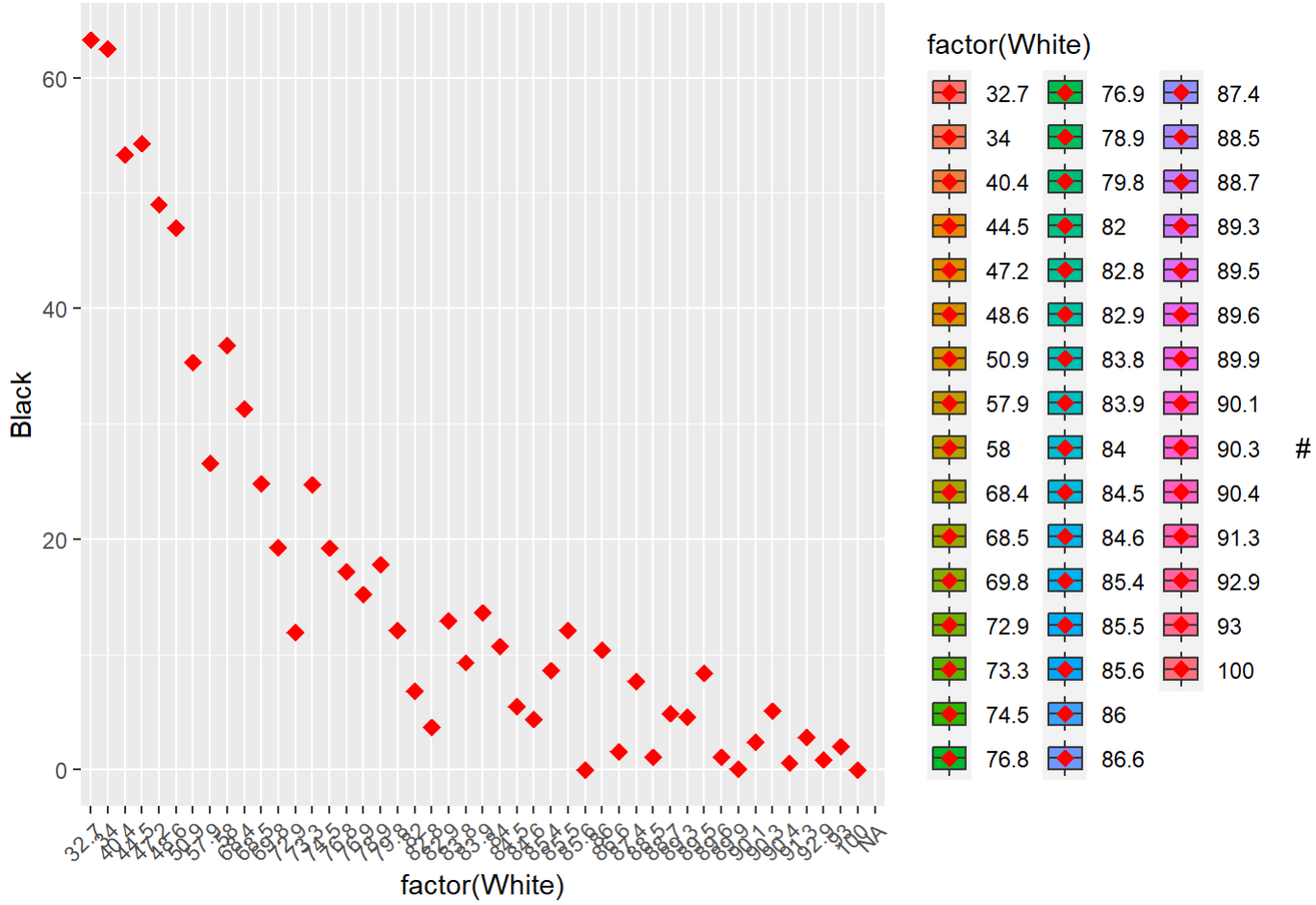
#

Boxplot graph

```
ggplot(head(df,50), aes(x = factor(White), y = Black, group = factor(White),fill=factor(Whit
e))) +
  geom_boxplot(width = 0.3) +  # Adjust the width as needed
  stat_summary(fun = "mean", geom = "point", shape = 18, size = 3, color = "red") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Adjust x-axis label rotation if
needed
```

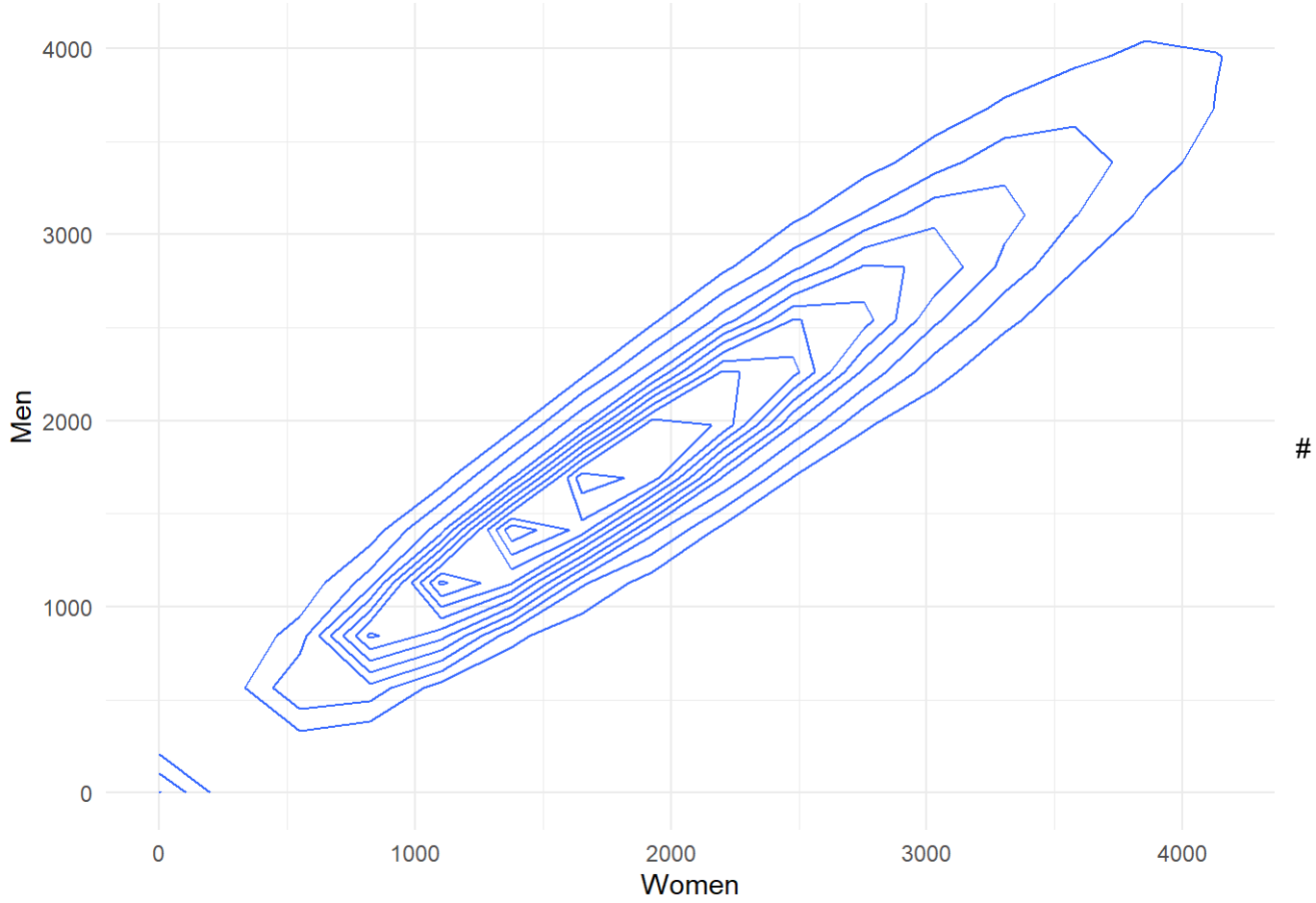
```
## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_summary()`).
```



2D Density graph

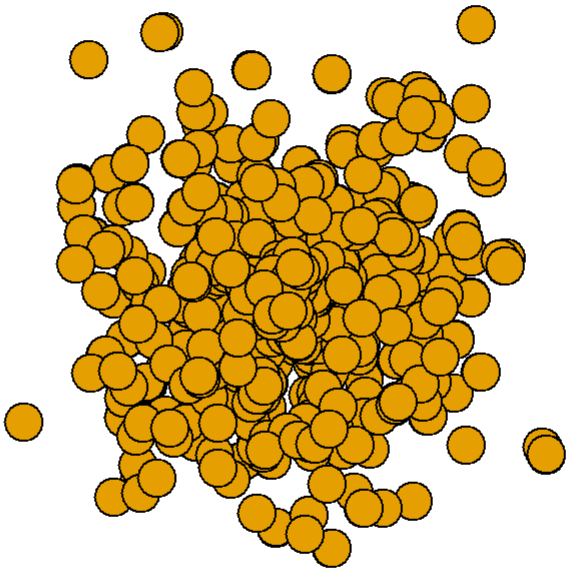
```
ggplot(df, aes(x = Women, y = Men)) +
  geom_density_2d() +
  labs(x = "Women", y = "Men") +
  theme_minimal()
```



#

Network graph

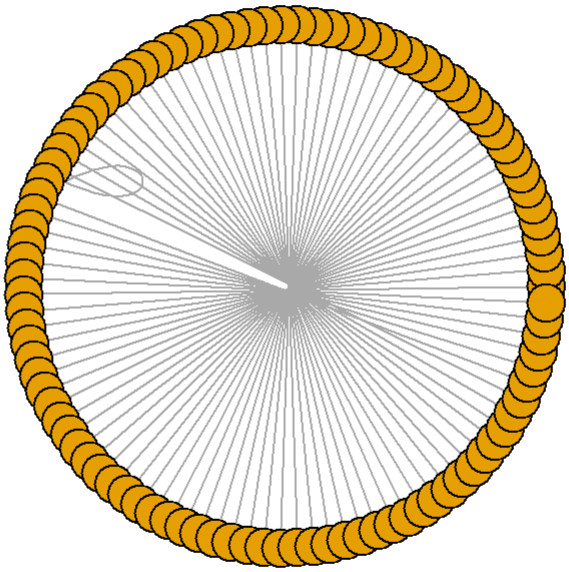
```
library(igraph)

set.seed(42)  # for reproducibility

# Create a graph
graph <- graph.data.frame(df[, c("Men", "Women", "TotalPop")], directed = FALSE)

# Plot the network graph
plot(graph,
     layout = layout_with_fr(graph),  # Use Fruchterman-Reingold layout
     vertex.label = NA,  # Display category_id as node labels
     edge.label = E(graph)$Women)  # Display views as edge labels
```



#

Network graph

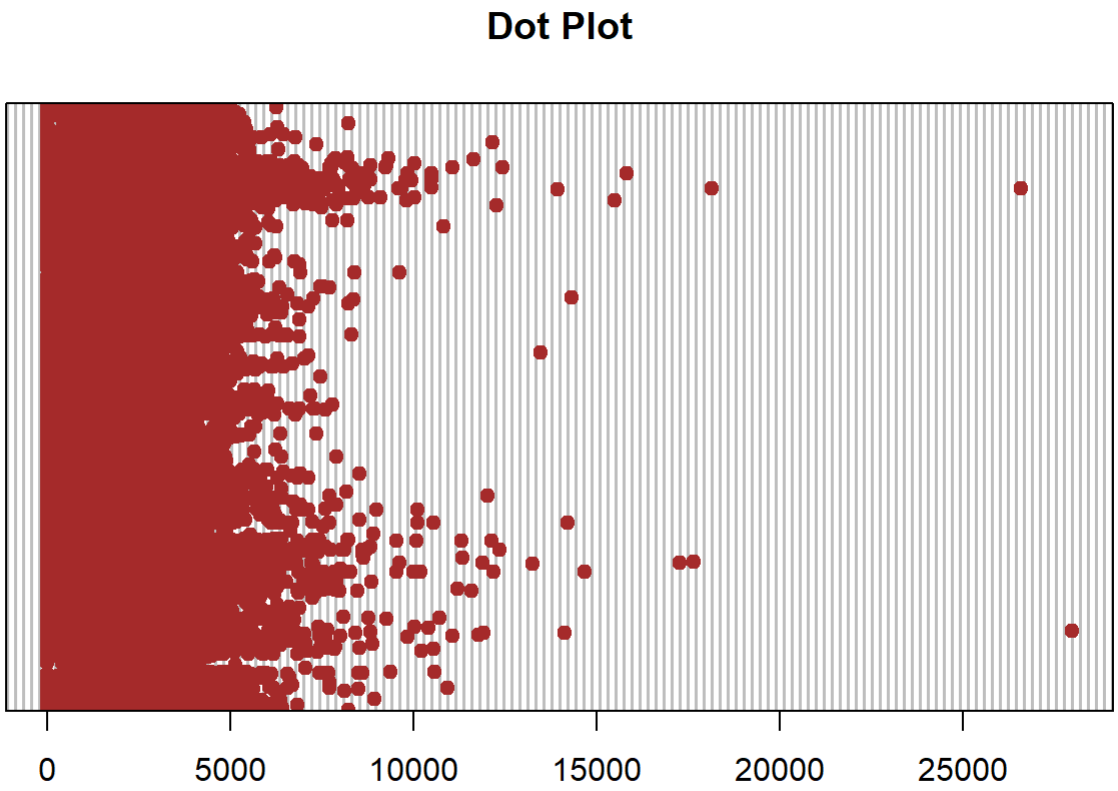
```
set.seed(42)  # for reproducibility

# Create a graph
graph <- graph.data.frame(head(df,50)[, c("Men", "Women", "TotalPop")], directed = FALSE)

# Plot the network graph
plot(graph,
     layout = layout.circle(graph),  # Use Fruchterman-Reingold layout
     vertex.label = NA,  # Display category_id as node labels
     edge.label = E(graph)$Men)  # Display views as edge labels
```



#

Dot plot

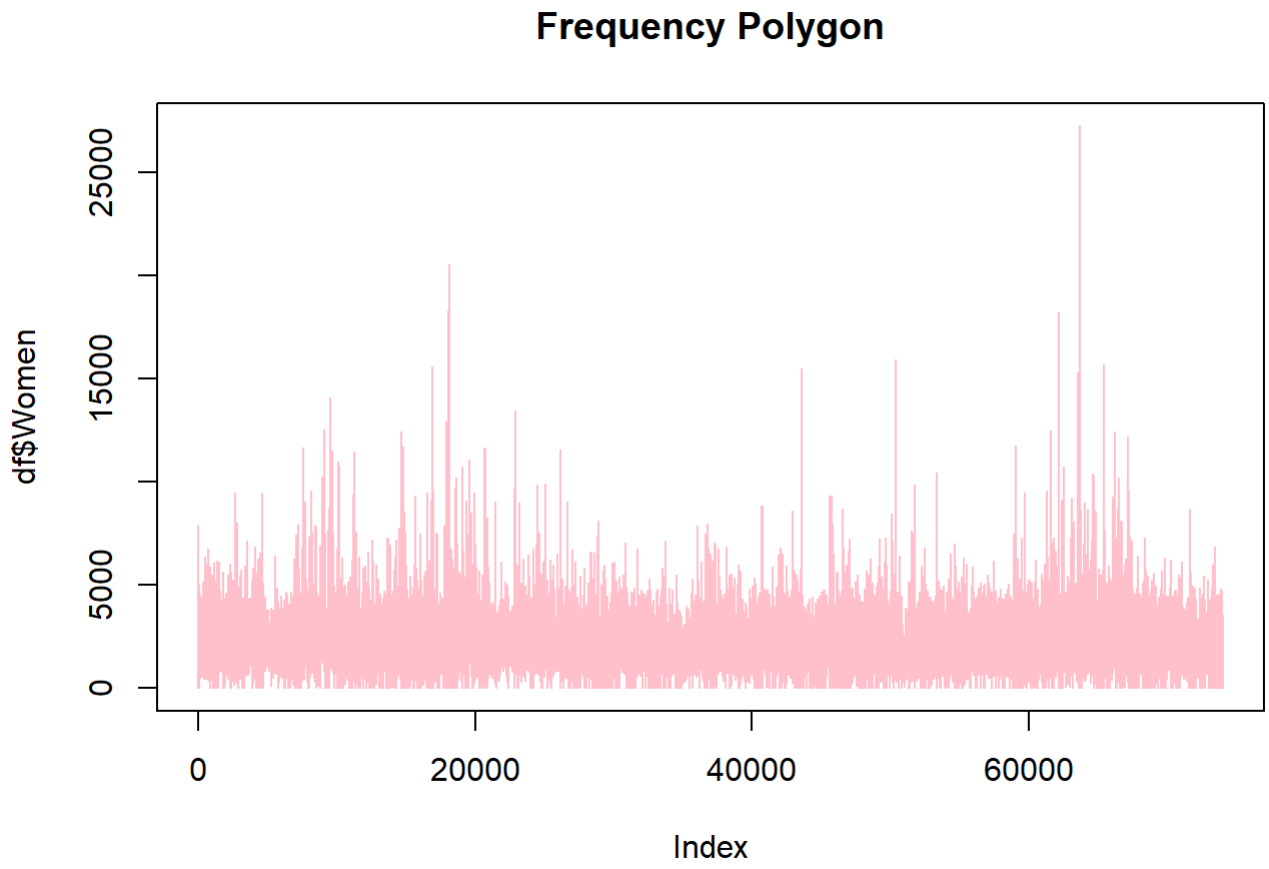
```
# Making a Dot Plot of US Census Dataset

dotchart(df$Men, pch = 19, col = "brown", main = "Dot Plot")
```



#

Frequency graph

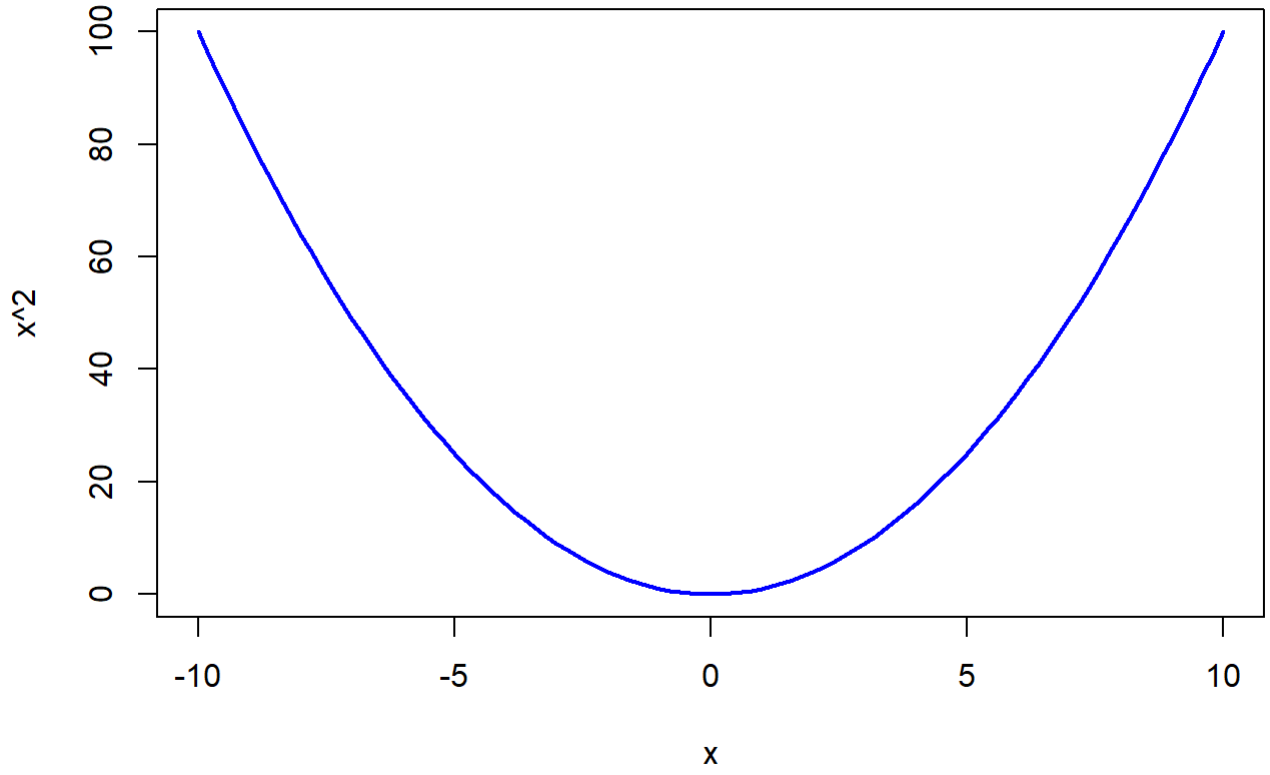
```
# Making a Frequency Polygon for the US Census Dataset
plot(df$Women, type = "l", lty = 1, col = "pink", main = "Frequency Polygon")
```



#

Curve function

```
# Plotting a Function Curve for the given dataset
curve(x^2, from = -10, to = 10, col = "blue", lwd = 2)
```



Github Link - https://github.com/shiveejaiswal/EDA (https://github.com/shiveejaiswal/EDA)