BST210 Project appendix

Exploratory Data Analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(viridisLite) # nice colours
```

What's the structure of our data?

```
dat <- read_csv("data/cancer_reg.csv")
```

```
## Rows: 3047 Columns: 34
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): binnedInc, Geography
## dbl (32): avgAnnCount, avgDeathsPerYear, TARGET_deathRate, incidenceRate, me...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(dat)
```

```
## spec_tbl_df [3,047 x 34] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ avgAnnCount            : num [1:3047] 1397 173 102 427 57 ...
##  $ avgDeathsPerYear       : num [1:3047] 469 70 50 202 26 152 97 71 36 1380 ...
##  $ TARGET_deathRate       : num [1:3047] 165 161 175 195 144 ...
##  $ incidenceRate          : num [1:3047] 490 412 350 430 350 ...
##  $ medIncome              : num [1:3047] 61898 48127 49348 44243 49955 ...
##  $ popEst2015             : num [1:3047] 260131 43269 21026 75882 10321 ...
##  $ povertyPercent         : num [1:3047] 11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
```

```
##  $ studyPerCap           : num [1:3047] 499.7 23.1 47.6 342.6 0 ...
##  $ binnedInc             : chr [1:3047] "(61494.5, 125635]" "(48021.6, 51046.4]" "(48021.6, 51046.4]
##  $ MedianAge             : num [1:3047] 39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
##  $ MedianAgeMale         : num [1:3047] 36.9 32.2 44 42.2 47.8 43.5 42.2 50.8 48.4 34.7 ...
##  $ MedianAgeFemale       : num [1:3047] 41.7 33.7 45.8 43.4 48.9 48 43.5 52.5 49.8 37 ...
##  $ Geography             : chr [1:3047] "Kitsap County, Washington" "Kittitas County, Washington" "M
##  $ AvgHouseholdSize      : num [1:3047] 2.54 2.34 2.62 2.52 2.34 2.58 2.42 2.24 2.38 2.65 ...
##  $ PercentMarried        : num [1:3047] 52.5 44.5 54.2 52.7 57.8 50.4 54.1 52.7 55.9 50 ...
##  $ PctNoHS18_24          : num [1:3047] 11.5 6.1 24 20.2 14.9 29.9 26.1 27.3 34.7 15.6 ...
##  $ PctHS18_24            : num [1:3047] 39.5 22.4 36.6 41.2 43 35.1 41.4 33.9 39.4 36.3 ...
##  $ PctSomeCol18_24       : num [1:3047] 42.1 64 NA 36.1 40 NA NA 36.5 NA NA ...
##  $ PctBachDeg18_24       : num [1:3047] 6.9 7.5 9.5 2.5 2 4.5 5.8 2.2 1.4 7.1 ...
##  $ PctHS25_Over          : num [1:3047] 23.2 26 29 31.6 33.4 30.4 29.8 31.6 32.2 28.8 ...
##  $ PctBachDeg25_Over     : num [1:3047] 19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
##  $ PctEmployed16_Over    : num [1:3047] 51.9 55.9 45.9 48.3 48.2 44.1 51.8 40.9 39.5 56.6 ...
##  $ PctUnemployed16_Over  : num [1:3047] 8 7.8 7 12.1 4.8 12.9 8.9 8.9 10.3 9.2 ...
##  $ PctPrivateCoverage    : num [1:3047] 75.1 70.2 63.7 58.4 61.6 60 49.5 55.8 55.5 69.9 ...
##  $ PctPrivateCoverageAlone: num [1:3047] NA 53.8 43.5 40.3 43.9 38.8 35 33.1 37.8 NA ...
##  $ PctEmpPrivCoverage    : num [1:3047] 41.6 43.6 34.9 35 35.1 32.6 28.3 25.9 29.9 44.4 ...
##  $ PctPublicCoverage     : num [1:3047] 32.9 31.1 42.1 45.3 44 43.2 46.4 50.9 48.1 31.4 ...
##  $ PctPublicCoverageAlone : num [1:3047] 14 15.3 21.1 25 22.7 20.2 28.7 24.1 26.6 16.5 ...
##  $ PctWhite              : num [1:3047] 81.8 89.2 90.9 91.7 94.1 ...
##  $ PctBlack              : num [1:3047] 2.595 0.969 0.74 0.783 0.27 ...
##  $ PctAsian              : num [1:3047] 4.822 2.246 0.466 1.161 0.666 ...
##  $ PctOtherRace          : num [1:3047] 1.843 3.741 2.747 1.363 0.492 ...
##  $ PctMarriedHouseholds  : num [1:3047] 52.9 45.4 54.4 51 54 ...
##  $ BirthRate             : num [1:3047] 6.12 4.33 3.73 4.6 6.8 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   avgAnnCount = col_double(),
##   ..   avgDeathsPerYear = col_double(),
##   ..   TARGET_deathRate = col_double(),
##   ..   incidenceRate = col_double(),
##   ..   medIncome = col_double(),
##   ..   popEst2015 = col_double(),
##   ..   povertyPercent = col_double(),
##   ..   studyPerCap = col_double(),
##   ..   binnedInc = col_character(),
##   ..   MedianAge = col_double(),
##   ..   MedianAgeMale = col_double(),
##   ..   MedianAgeFemale = col_double(),
##   ..   Geography = col_character(),
##   ..   AvgHouseholdSize = col_double(),
##   ..   PercentMarried = col_double(),
##   ..   PctNoHS18_24 = col_double(),
##   ..   PctHS18_24 = col_double(),
##   ..   PctSomeCol18_24 = col_double(),
##   ..   PctBachDeg18_24 = col_double(),
##   ..   PctHS25_Over = col_double(),
##   ..   PctBachDeg25_Over = col_double(),
##   ..   PctEmployed16_Over = col_double(),
##   ..   PctUnemployed16_Over = col_double(),
##   ..   PctPrivateCoverage = col_double(),
##   ..   PctPrivateCoverageAlone = col_double(),
```

```
##   ..   PctEmpPrivCoverage = col_double(),
##   ..   PctPublicCoverage = col_double(),
##   ..   PctPublicCoverageAlone = col_double(),
##   ..   PctWhite = col_double(),
##   ..   PctBlack = col_double(),
##   ..   PctAsian = col_double(),
##   ..   PctOtherRace = col_double(),
##   ..   PctMarriedHouseholds = col_double(),
##   ..   BirthRate = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(dat)
```

```
##   avgAnnCount      avgDeathsPerYear TARGET_deathRate incidenceRate
##  Min.   :    6.0   Min.   :    3    Min.   : 59.7    Min.   : 201.3
##  1st Qu.:   76.0   1st Qu.:   28    1st Qu.:161.2    1st Qu.: 420.3
##  Median :  171.0   Median :   61    Median :178.1    Median : 453.5
##  Mean   :  606.3   Mean   :  186    Mean   :178.7    Mean   : 448.3
##  3rd Qu.:  518.0   3rd Qu.:  149    3rd Qu.:195.2    3rd Qu.: 480.9
##  Max.   :38150.0   Max.   :14010    Max.   :362.8    Max.   :1206.9
##
##    medIncome        popEst2015       povertyPercent   studyPerCap
##  Min.   : 22640   Min.   :     827   Min.   : 3.20   Min.   :   0.00
##  1st Qu.: 38883   1st Qu.:   11684   1st Qu.:12.15   1st Qu.:   0.00
##  Median : 45207   Median :   26643   Median :15.90   Median :   0.00
##  Mean   : 47063   Mean   :  102637   Mean   :16.88   Mean   : 155.40
##  3rd Qu.: 52492   3rd Qu.:   68671   3rd Qu.:20.40   3rd Qu.:  83.65
##  Max.   :125635   Max.   :10170292   Max.   :47.40   Max.   :9762.31
##
##   binnedInc           MedianAge      MedianAgeMale   MedianAgeFemale
##  Length:3047        Min.   : 22.30   Min.   :22.40   Min.   :22.30
##  Class :character   1st Qu.: 37.70   1st Qu.:36.35   1st Qu.:39.10
##  Mode  :character   Median : 41.00   Median :39.60   Median :42.40
##                     Mean   : 45.27   Mean   :39.57   Mean   :42.15
##                     3rd Qu.: 44.00   3rd Qu.:42.50   3rd Qu.:45.30
##                     Max.   :624.00   Max.   :64.70   Max.   :65.70
##
##   Geography          AvgHouseholdSize PercentMarried   PctNoHS18_24
##  Length:3047        Min.   :0.0221   Min.   :23.10   Min.   : 0.00
##  Class :character   1st Qu.:2.3700   1st Qu.:47.75   1st Qu.:12.80
##  Mode  :character   Median :2.5000   Median :52.40   Median :17.10
##                     Mean   :2.4797   Mean   :51.77   Mean   :18.22
##                     3rd Qu.:2.6300   3rd Qu.:56.40   3rd Qu.:22.70
##                     Max.   :3.9700   Max.   :72.50   Max.   :64.10
##
##    PctHS18_24   PctSomeCol18_24 PctBachDeg18_24   PctHS25_Over
##  Min.   : 0.0   Min.   : 7.10   Min.   : 0.000   Min.   : 7.50
##  1st Qu.:29.2   1st Qu.:34.00   1st Qu.: 3.100   1st Qu.:30.40
##  Median :34.7   Median :40.40   Median : 5.400   Median :35.30
##  Mean   :35.0   Mean   :40.98   Mean   : 6.158   Mean   :34.80
##  3rd Qu.:40.7   3rd Qu.:46.40   3rd Qu.: 8.200   3rd Qu.:39.65
##  Max.   :72.5   Max.   :79.00   Max.   :51.800   Max.   :54.80
##                 NA's   :2285
```

```
##  PctBachDeg25_Over PctEmployed16_Over PctUnemployed16_Over PctPrivateCoverage
##  Min.   : 2.50     Min.   :17.60      Min.   : 0.400       Min.   :22.30
##  1st Qu.: 9.40     1st Qu.:48.60      1st Qu.: 5.500       1st Qu.:57.20
##  Median :12.30     Median :54.50      Median : 7.600       Median :65.10
##  Mean   :13.28     Mean   :54.15      Mean   : 7.852       Mean   :64.35
##  3rd Qu.:16.10     3rd Qu.:60.30      3rd Qu.: 9.700       3rd Qu.:72.10
##  Max.   :42.20     Max.   :80.10      Max.   :29.400       Max.   :92.30
##                    NA's   :152
##  PctPrivateCoverageAlone PctEmpPrivCoverage PctPublicCoverage
##  Min.   :15.70           Min.   :13.5       Min.   :11.20
##  1st Qu.:41.00           1st Qu.:34.5       1st Qu.:30.90
##  Median :48.70           Median :41.1       Median :36.30
##  Mean   :48.45           Mean   :41.2       Mean   :36.25
##  3rd Qu.:55.60           3rd Qu.:47.7       3rd Qu.:41.55
##  Max.   :78.90           Max.   :70.7       Max.   :65.10
##  NA's   :609
##  PctPublicCoverageAlone    PctWhite         PctBlack          PctAsian
##  Min.   : 2.60          Min.   : 10.20   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.:14.85          1st Qu.: 77.30   1st Qu.: 0.6207   1st Qu.: 0.2542
##  Median :18.80          Median : 90.06   Median : 2.2476   Median : 0.5498
##  Mean   :19.24          Mean   : 83.65   Mean   : 9.1080   Mean   : 1.2540
##  3rd Qu.:23.10          3rd Qu.: 95.45   3rd Qu.:10.5097   3rd Qu.: 1.2210
##  Max.   :46.60          Max.   :100.00   Max.   :85.9478   Max.   :42.6194
##
##   PctOtherRace     PctMarriedHouseholds   BirthRate
##  Min.   : 0.0000   Min.   :22.99        Min.   : 0.000
##  1st Qu.: 0.2952   1st Qu.:47.76        1st Qu.: 4.521
##  Median : 0.8262   Median :51.67        Median : 5.381
##  Mean   : 1.9835   Mean   :51.24        Mean   : 5.640
##  3rd Qu.: 2.1780   3rd Qu.:55.40        3rd Qu.: 6.494
##  Max.   :41.9303   Max.   :78.08        Max.   :21.326
##
```

```
ed <- read_csv("data/Education.csv")
```

```
## Rows: 169235 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): State, Area name, Attribute
## dbl (2): FIPS Code, Value
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

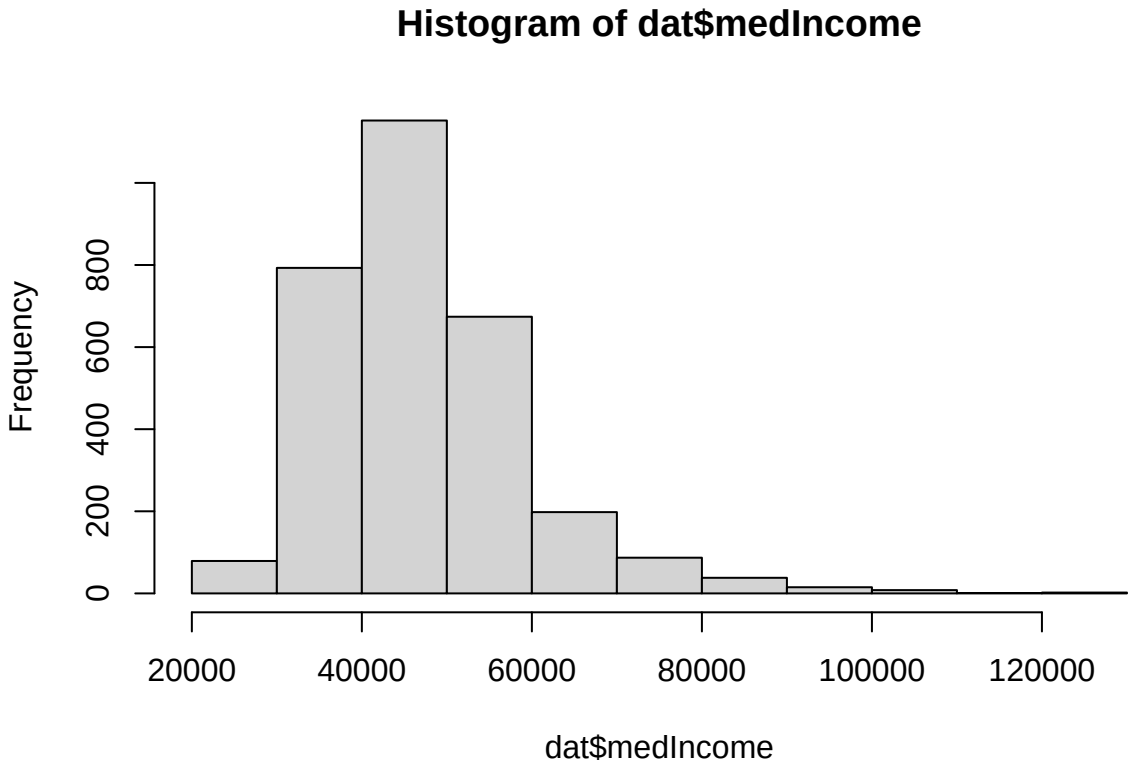
```
states <- read_csv("data/50_states.csv") %>%
    add_row(State="District of Columbia", Abbr="DC", `State Capital`="Washington", Region="East")
```

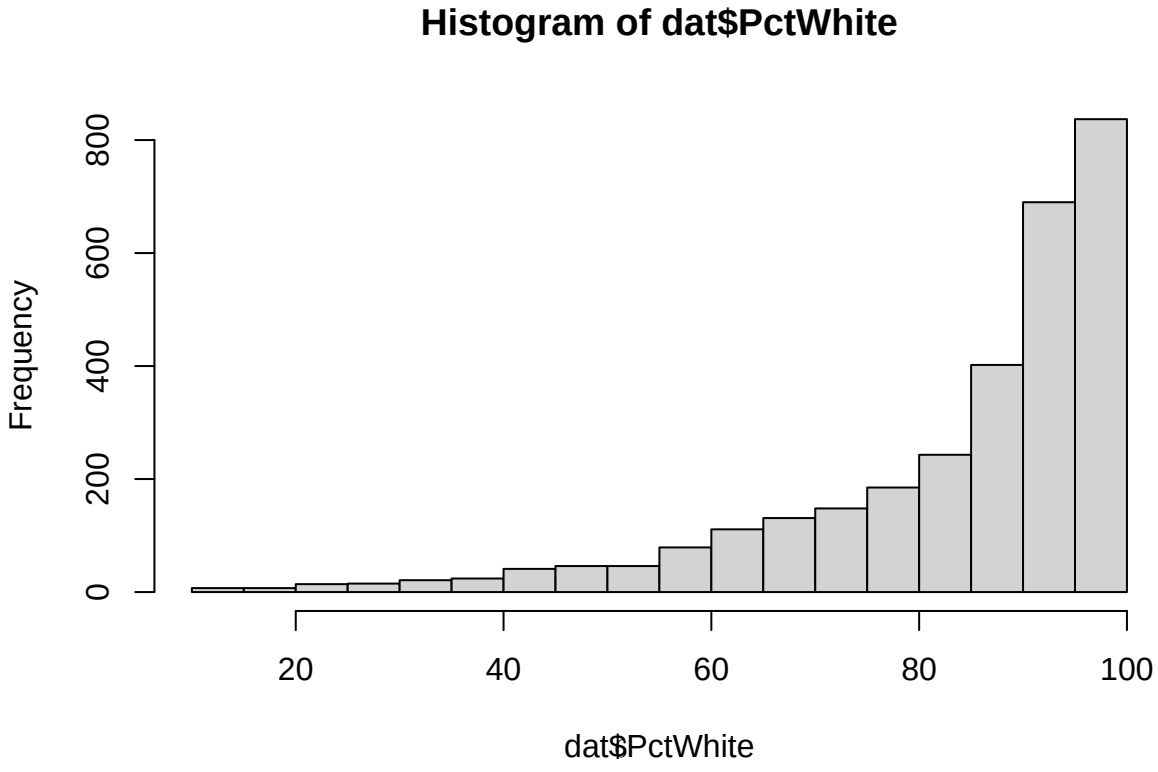
```
## Rows: 50 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): State, Abbr, State Capital, Region
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

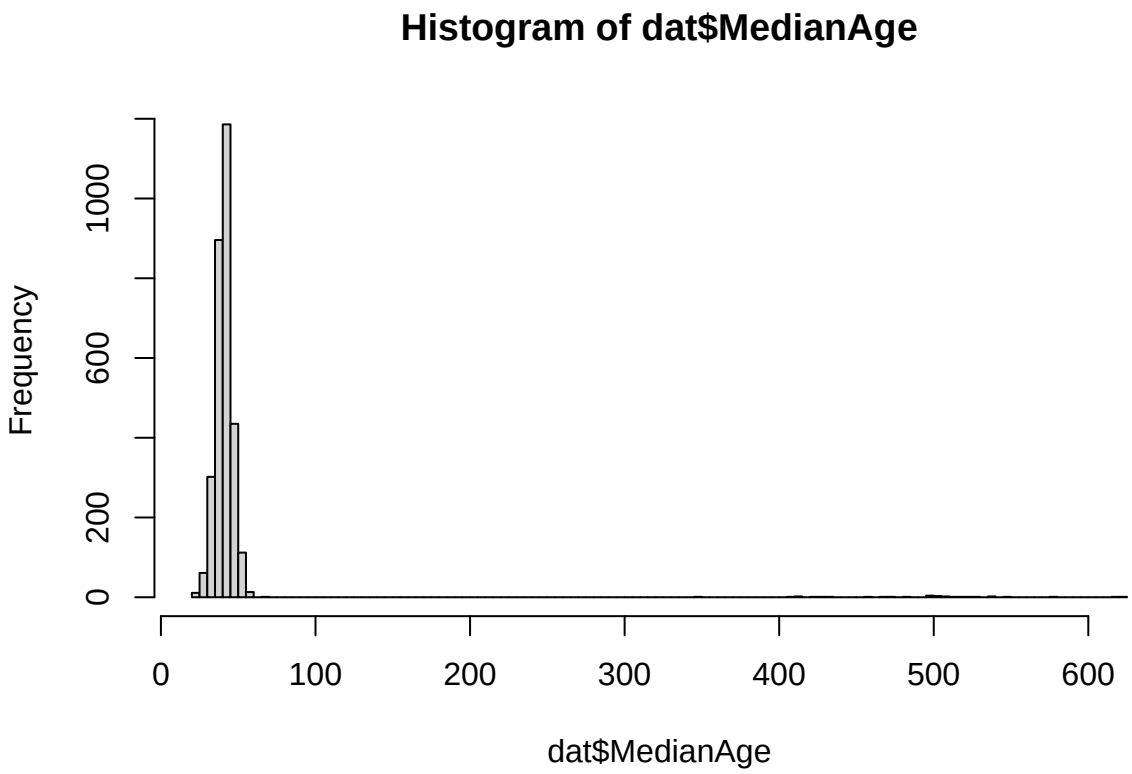
```r
# get county names
dat <- dat %>%
    mutate(county_name = str_split(Geography, ",") %>% map_chr(., 1))%>%
    mutate(state = str_extract(Geography, "[^,]+$")) %>% # regex to select everything after ','
    mutate(state = str_trim(state)) %>%
    left_join(states, by=c("state" = "State")) %>%
    filter(county_name != "Valdez-Cordova Census Area")

Encoding(dat$county_name) <- "UTF-8"
dat$county_name <- iconv(dat$county_name, "UTF-8", "UTF-8", sub='')
dat <- dat %>%
    mutate(county_name = ifelse(county_name == "Doa Ana County", "Dona Ana County", county_name))

# get education stats from 2016-2020 https://www.ers.usda.gov/data-products/county-level-data-sets/down
ed <- ed %>%
    filter(grepl("2016", Attribute)) %>%
    rename(county_name = `Area name`) %>%
    pivot_wider(names_from = Attribute, values_from = Value)  %>%
    mutate(county_name = case_when(county_name == "La Salle County" & State != "TX" ~ "LaSalle County",
                                   county_name == "La Salle Parish" ~ "LaSalle Parish",
                                   TRUE ~ county_name))

dat <- ed %>%
    right_join(dat, by=c("county_name", "State" = "Abbr")) %>%
    distinct()
```

5. Modelling Approches

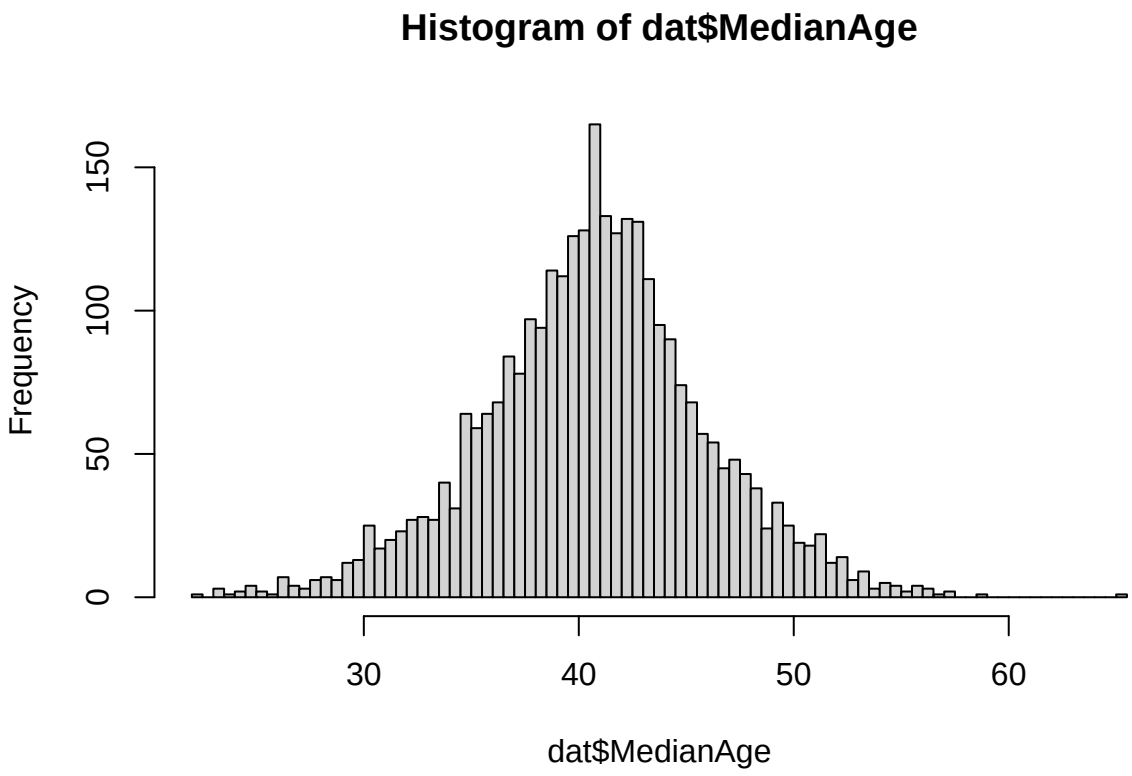
a. Fitting an linear model








```
dat = dat %>% filter(MedianAge <= 100)

hist(dat$MedianAge, breaks = 100)
```




```
mod1 = lm(data = dat, TARGET_deathRate ~ medIncome + MedianAge + PctWhite)
summary(mod1)
```

model fitting:

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + MedianAge + PctWhite,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.117  -14.061    0.904   15.057  175.883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.411e+02  4.250e+00  56.735  < 2e-16 ***
## medIncome   -9.492e-04  3.889e-05 -24.409  < 2e-16 ***
## MedianAge   -7.100e-02  9.591e-02  -0.740    0.459
## PctWhite    -1.785e-01  3.065e-02  -5.823  6.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 24.89 on 3013 degrees of freedom
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.1946
## F-statistic:   244 on 3 and 3013 DF,  p-value: < 2.2e-16
```

```
mod1.1 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite)
summary(mod1.1)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.987  -14.167    0.874   15.145  175.870
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.387e+02  2.748e+00  86.894  < 2e-16 ***
## medIncome   -9.436e-04  3.813e-05 -24.746  < 2e-16 ***
## PctWhite    -1.876e-01  2.806e-02  -6.686 2.73e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.89 on 3014 degrees of freedom
## Multiple R-squared:  0.1953, Adjusted R-squared:  0.1947
## F-statistic: 365.7 on 2 and 3014 DF,  p-value: < 2.2e-16
```

```
anova(mod1.1, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   3014 1867102
## 2   3013 1866763  1    339.57 0.5481 0.4592
```

```
mod1.2 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite + MedianAge + medIncome*MedianAge)
summary(mod1.2)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + MedianAge +
##     medIncome * MedianAge, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.331  -13.843    0.929   14.955  175.902
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.012e+02  1.625e+01  12.379  < 2e-16 ***
```

```
## medIncome           -7.301e-05  3.464e-04  -0.211   0.8331
## PctWhite            -1.806e-01  3.064e-02  -5.894  4.2e-09 ***
## MedianAge            9.390e-01  4.082e-01   2.301   0.0215 *
## medIncome:MedianAge -2.213e-05  8.693e-06  -2.546   0.0110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.87 on 3012 degrees of freedom
## Multiple R-squared:  0.1972, Adjusted R-squared:  0.1961
## F-statistic: 184.9 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
anova(mod1.2, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + MedianAge + medIncome *
##     MedianAge
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1   3012 1862754
## 2   3013 1866763 -1   -4008.3 6.4812 0.01095 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod1.3 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite + MedianAge + PctWhite*MedianAge)
summary(mod1.3)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + MedianAge +
##     PctWhite * MedianAge, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.122  -14.063    0.896   15.064  175.865
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.400e+02  1.824e+01  13.160   <2e-16 ***
## medIncome          -9.493e-04  3.890e-05 -24.401   <2e-16 ***
## PctWhite           -1.653e-01  2.123e-01  -0.779    0.436
## MedianAge          -4.166e-02  4.759e-01  -0.088    0.930
## PctWhite:MedianAge -3.438e-04  5.461e-03  -0.063    0.950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.9 on 3012 degrees of freedom
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.1944
## F-statistic: 182.9 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
anova(mod1.3, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + MedianAge + PctWhite *
##     MedianAge
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1   3012 1866760
## 2   3013 1866763 -1   -2.4565 0.004 0.9498
```

```
mod1.4 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite + PctBlack + PctAsian + PctOtherRace)
summary(mod1.4)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + PctBlack +
##     PctAsian + PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.383  -13.918    0.866   14.279  174.216
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.364e+02  5.931e+00  39.869  < 2e-16 ***
## medIncome    -8.450e-04  4.302e-05 -19.640  < 2e-16 ***
## PctWhite     -1.917e-01  6.064e-02  -3.161  0.00159 **
## PctBlack      1.204e-01  6.406e-02   1.879  0.06033 .
## PctAsian     -2.622e-01  2.130e-01  -1.231  0.21854
## PctOtherRace -1.395e+00  1.414e-01  -9.865  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.33 on 3011 degrees of freedom
## Multiple R-squared:  0.2317, Adjusted R-squared:  0.2305
## F-statistic: 181.7 on 5 and 3011 DF,  p-value: < 2.2e-16
```

```
cor(dat$PctAsian, dat$PctWhite)
```

```
## [1] -0.2658648
```

```
cor(dat$PctBlack, dat$PctWhite)
```

```
## [1] -0.8312116
```

```
cor(dat$PctOtherRace, dat$PctWhite)
```

```
## [1] -0.2331931
```

```
mod1.5 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite + PctAsian + PctOtherRace)
summary(mod1.5)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + PctAsian +
##     PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.382  -13.873    0.839   14.226  174.620
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.463e+02  2.763e+00  89.160   <2e-16 ***
## medIncome    -8.482e-04  4.301e-05 -19.723   <2e-16 ***
## PctWhite     -2.905e-01  3.025e-02  -9.604   <2e-16 ***
## PctAsian     -3.827e-01  2.032e-01  -1.883   0.0598 .
## PctOtherRace -1.495e+00  1.310e-01 -11.413   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.34 on 3012 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.2298
## F-statistic:   226 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
anova(mod1.5, mod1.1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + PctAsian + PctOtherRace
## Model 2: TARGET_deathRate ~ medIncome + PctWhite
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1   3012 1784592
## 2   3014 1867102 -2    -82510 69.63 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod1.6 = lm(data = dat, TARGET_deathRate ~ medIncome)
summary(mod1.6)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.962  -14.433    0.937   15.098  177.402
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.250e+02  1.840e+00  122.29   <2e-16 ***
## medIncome   -9.856e-04  3.788e-05  -26.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 25.07 on 3015 degrees of freedom
## Multiple R-squared:  0.1833, Adjusted R-squared:  0.1831
## F-statistic: 676.9 on 1 and 3015 DF,  p-value: < 2.2e-16
```

```
mod1.6.1 = lm(data = dat, TARGET_deathRate ~ medIncome + I(medIncome ^2))
summary(mod1.6.1)
```

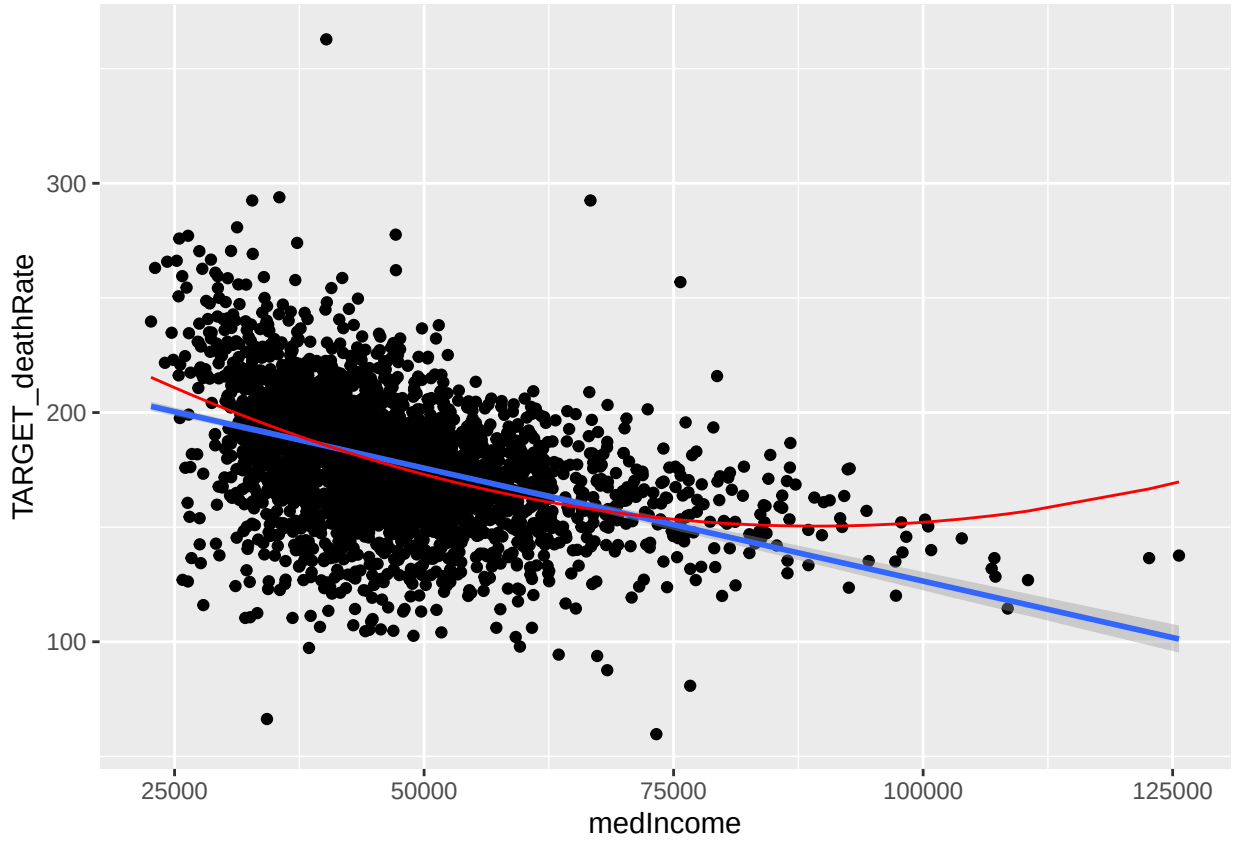
```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + I(medIncome^2), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -128.419  -13.923    1.128   14.799  177.132
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.670e+02  4.952e+00  53.913   <2e-16 ***
## medIncome      -2.609e-03  1.822e-04 -14.318   <2e-16 ***
## I(medIncome^2)  1.461e-08  1.605e-09   9.104   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.74 on 3014 degrees of freedom
## Multiple R-squared:  0.2052, Adjusted R-squared:  0.2047
## F-statistic: 389.1 on 2 and 3014 DF,  p-value: < 2.2e-16
```

```
anova(mod1.6, mod1.6.1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome
## Model 2: TARGET_deathRate ~ medIncome + I(medIncome^2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   3015 1894793
## 2   3014 1844082  1     50711 82.883 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
predict = data.frame(TARGET_deathRate = predict(mod1.6.1, dat), medIncome = dat$medIncome)

dat %>% ggplot(aes(medIncome, TARGET_deathRate)) + geom_point() + geom_smooth(method = "lm") + geom_lin
```

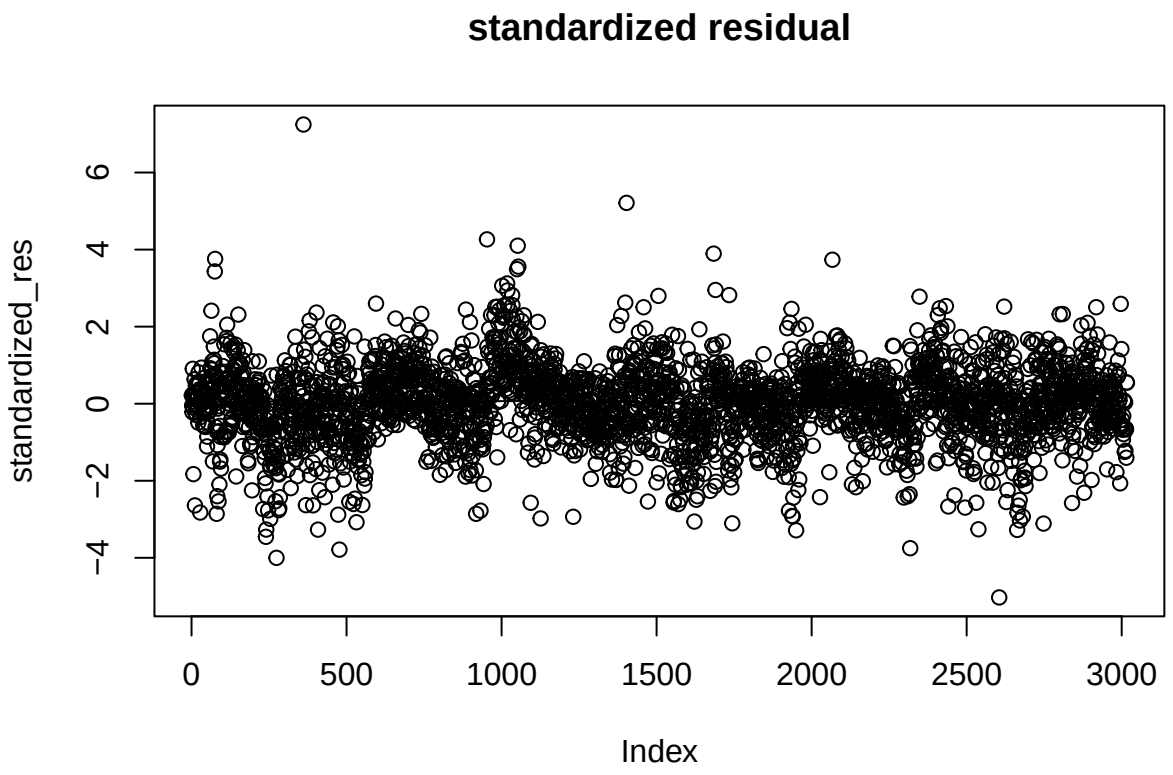
```
## `geom_smooth()` using formula 'y ~ x'
```



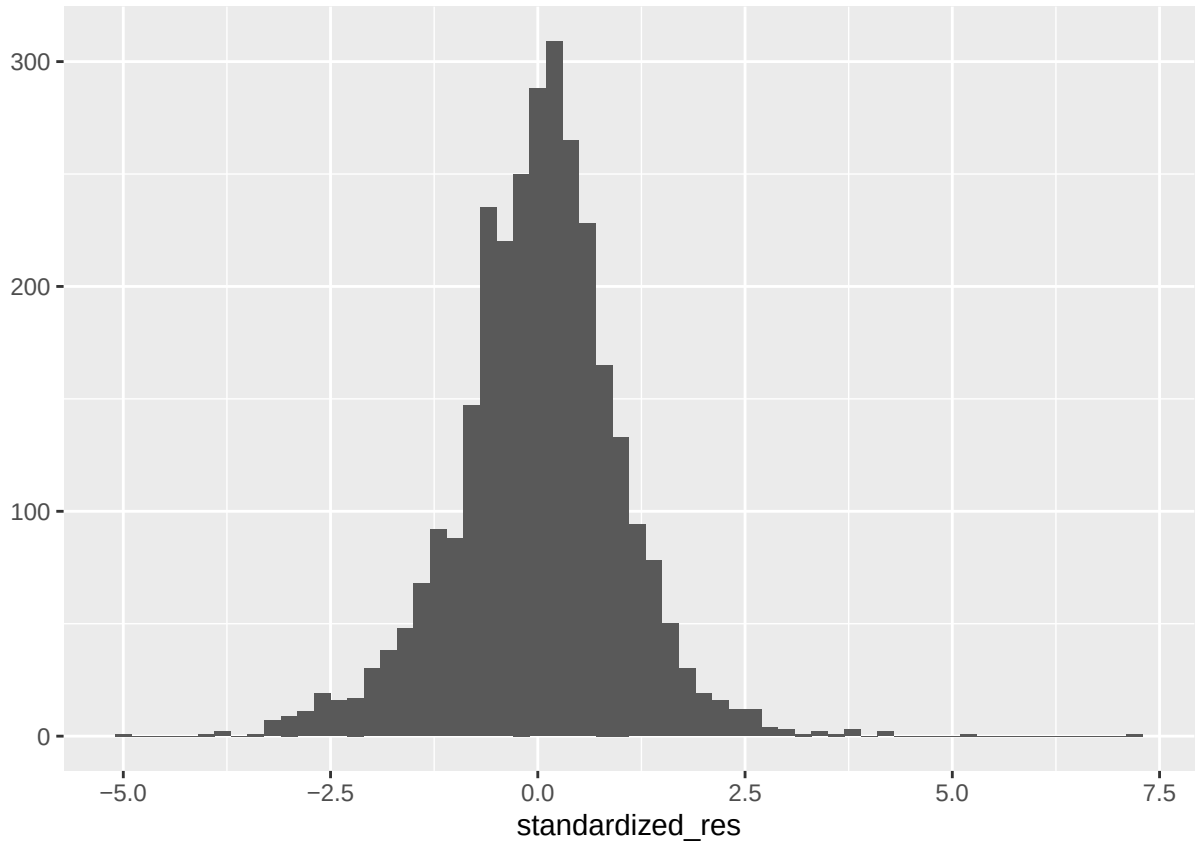

```
mod1_core = lm(data = dat, TARGET_deathRate ~ medIncome + I(medIncome ^2)+ PctWhite + PctAsian + PctOth
summary(mod1_core)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + I(medIncome^2) +
##     PctWhite + PctAsian + PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.247  -13.750    1.302   14.195  174.871
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.756e+02  4.938e+00  55.821  < 2e-16 ***
## medIncome      -2.159e-03  1.885e-04 -11.456  < 2e-16 ***
## I(medIncome^2)  1.182e-08  1.656e-09   7.141 1.16e-12 ***
## PctWhite       -2.362e-01  3.095e-02  -7.632 3.08e-14 ***
## PctAsian       -5.664e-01  2.032e-01  -2.787  0.00535 **
## PctOtherRace   -1.419e+00  1.303e-01 -10.884  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.14 on 3011 degrees of freedom
## Multiple R-squared:  0.2437, Adjusted R-squared:  0.2424
## F-statistic:   194 on 5 and 3011 DF,  p-value: < 2.2e-16
```

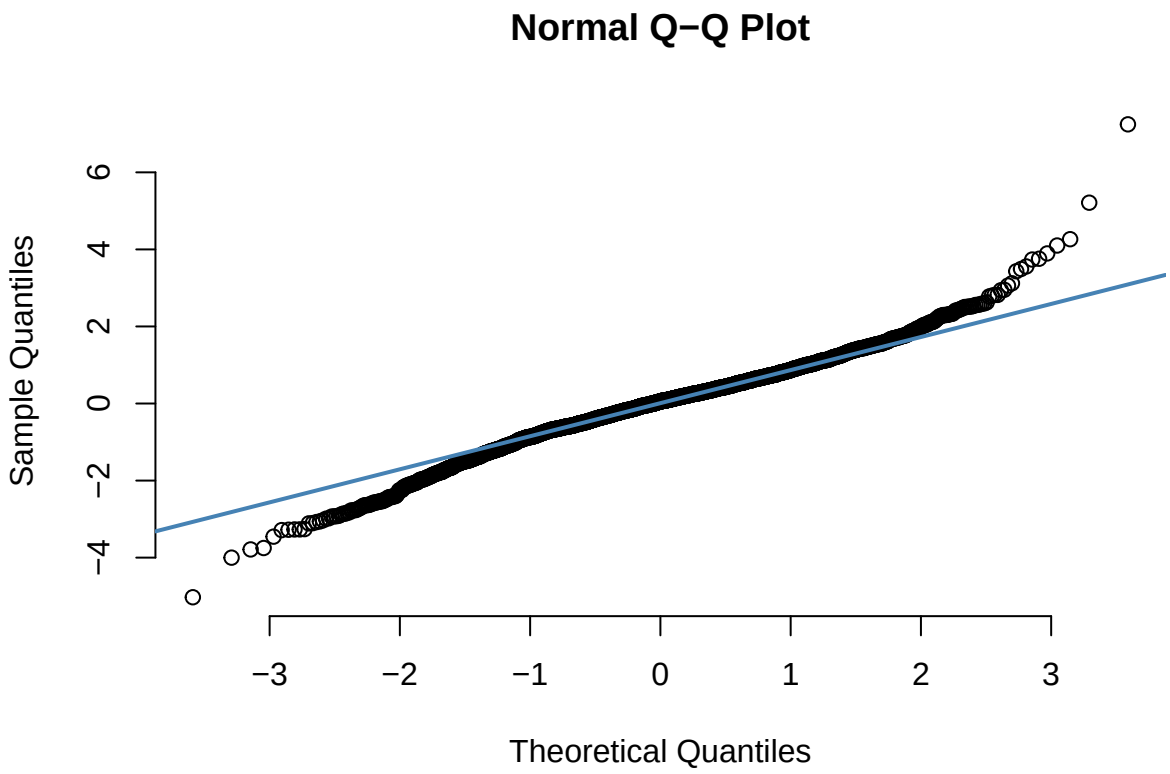
```
standardized_res = rstandard(mod1_core)
scatter.smooth(standardized_res, main = "standardized residual")
```




```
qplot(standardized_res, binwidth = 0.2)
```

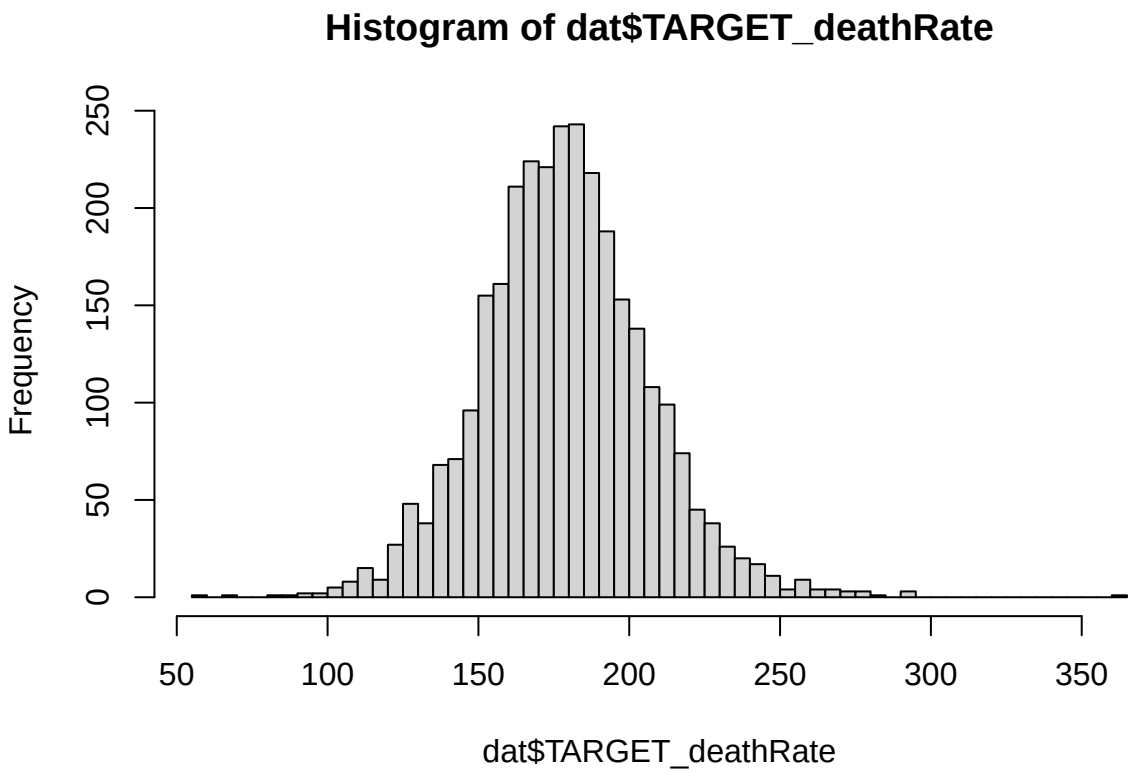



```
qqnorm(standardized_res, pch = 1, frame = FALSE)
qqline(standardized_res, col = "steelblue", lwd = 2)
```



b. Logistic/multinomial/ordinal regression

```
hist(dat$TARGET_deathRate, breaks = 100)
```




```
#create the three bins
dat = dat %>% mutate(multi = case_when(TARGET_deathRate < 150 ~ 1, TARGET_deathRate < 200 ~ 2, T ~ 3))
# 3 is bad quality lung cancer prevention, 2 is medium, 1 is good quality.
```

```
library(nnet)
mod2.1 <- multinom(multi ~ medIncome + I(medIncome ^2), data = dat)
```

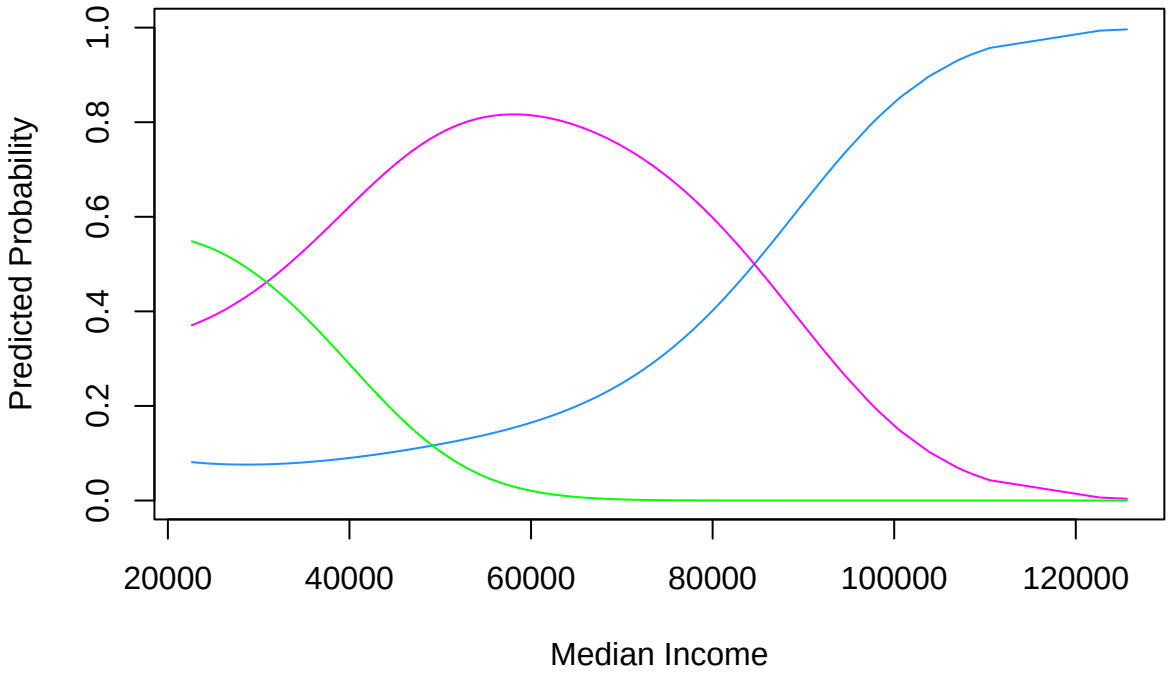
```
## # weights:  12 (6 variable)
## initial  value 3314.513275
## iter  10 value 2320.489209
## final  value 2316.095848
## converged
```

```
summary(mod2.1)
```

```
## Call:
## multinom(formula = multi ~ medIncome + I(medIncome^2), data = dat)
##
## Coefficients:
##    (Intercept)    medIncome I(medIncome^2)
## 2 2.417464e-09 9.165115e-05  -1.083448e-09
## 3 7.931352e-09 1.565121e-04  -3.185035e-09
##
## Std. Errors:
```

```
##    (Intercept)    medIncome I(medIncome^2)
## 2 6.249523e-21 2.523511e-16   1.799483e-11
## 3 1.659987e-20 7.128037e-16   3.143898e-11
##
## Residual Deviance: 4632.192
## AIC: 4644.192
```

```
plot(mod2.1$fitted.values[,1][order(dat$medIncome)] ~ sort(dat$medIncome), type="l", col="dodgerblue", 
points(mod2.1$fitted.values[,2][order(dat$medIncome)] ~ sort(dat$medIncome), type="l", col="magenta")
points(mod2.1$fitted.values[,3][order(dat$medIncome)]~sort(dat$medIncome), type="l", col="green")
```




```
mod2.2 <- multinom(multi ~ PctWhite , data = dat)
```

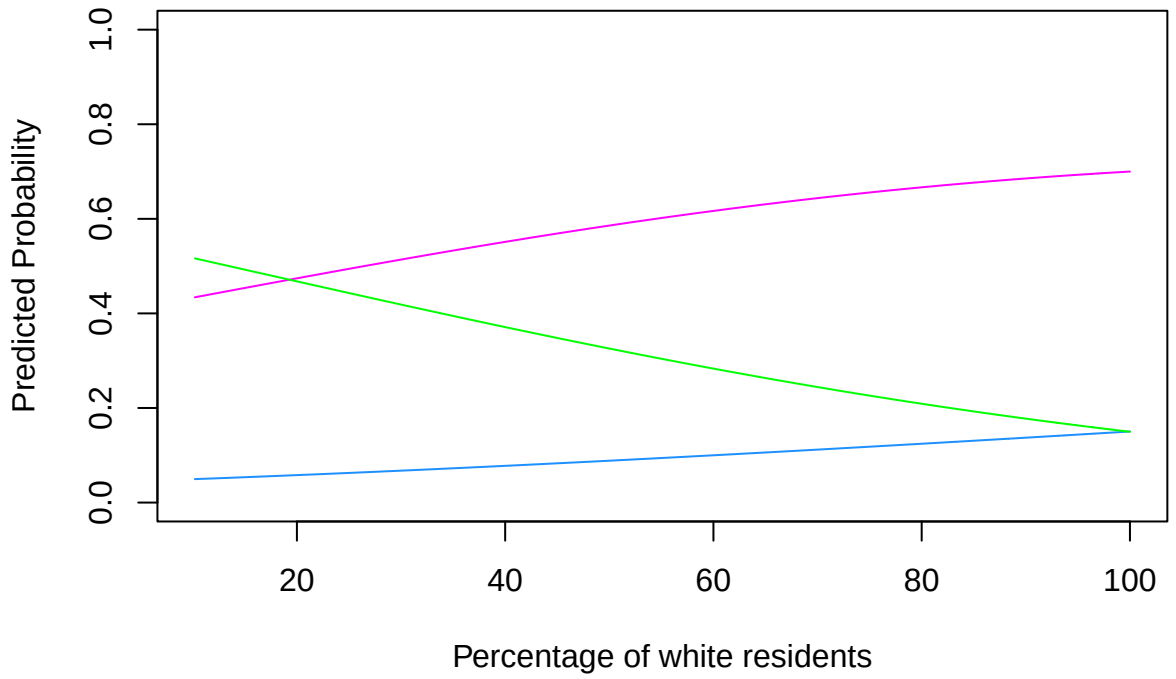
```
## # weights:  9 (4 variable)
## initial  value 3314.513275
## iter  10 value 2552.935350
## iter  10 value 2552.935350
## final  value 2552.935350
## converged
```

```
summary(mod2.2)
```

```
## Call:
## multinom(formula = multi ~ PctWhite, data = dat)
```

```
##
## Coefficients:
##   (Intercept)     PctWhite
## 2    2.241079 -0.007027921
## 3    2.609593 -0.026135639
##
## Std. Errors:
##   (Intercept)    PctWhite
## 2   0.3398884 0.003901638
## 3   0.3612030 0.004210873
##
## Residual Deviance: 5105.871
## AIC: 5113.871
```

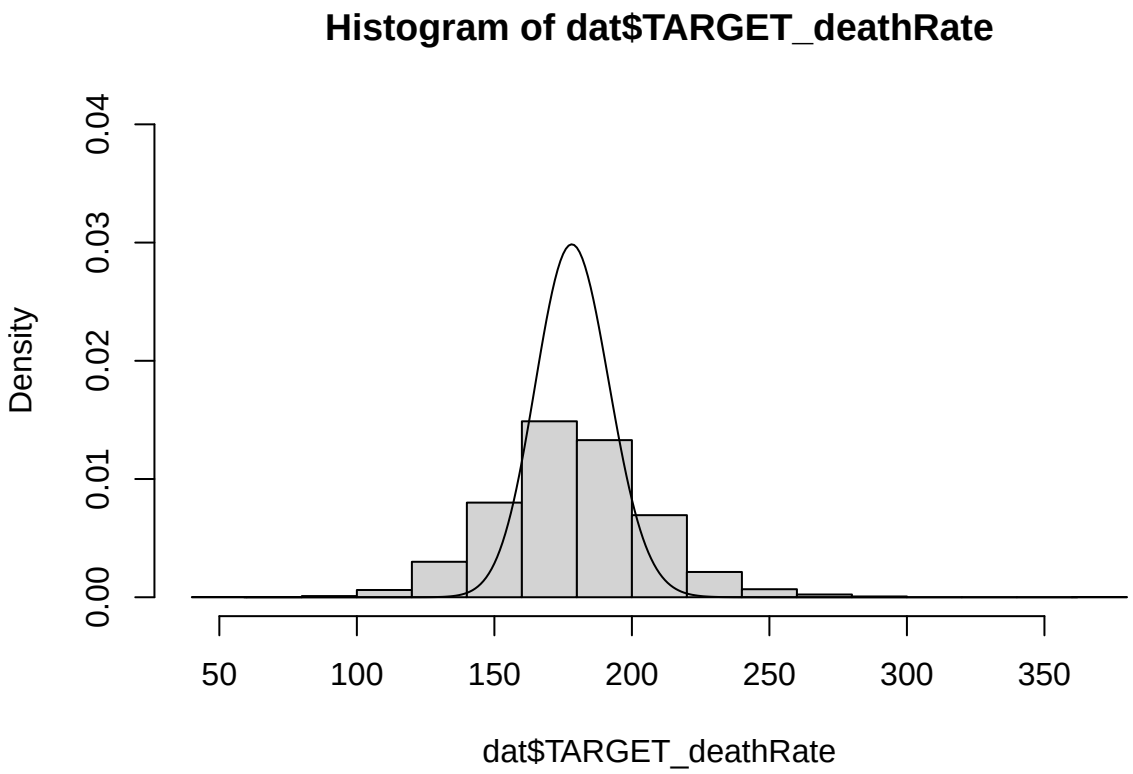
```
plot(mod2.2$fitted.values[,1][order(dat$PctWhite)] ~ sort(dat$PctWhite), type="l", col="dodgerblue", xla
points(mod2.2$fitted.values[,2][order(dat$PctWhite)] ~ sort(dat$PctWhite), type="l", col="magenta")
points(mod2.2$fitted.values[,3][order(dat$PctWhite)]~sort(dat$PctWhite), type="l", col="green")
```



c. Poisson Regression

Over-dispersion

```
hist(dat$TARGET_deathRate, freq = F, ylim = c(0, 0.04))
lines(as.integer(min(dat$TARGET_deathRate)):as.integer(max(dat$TARGET_deathRate)), dpois(as.integer(min
```



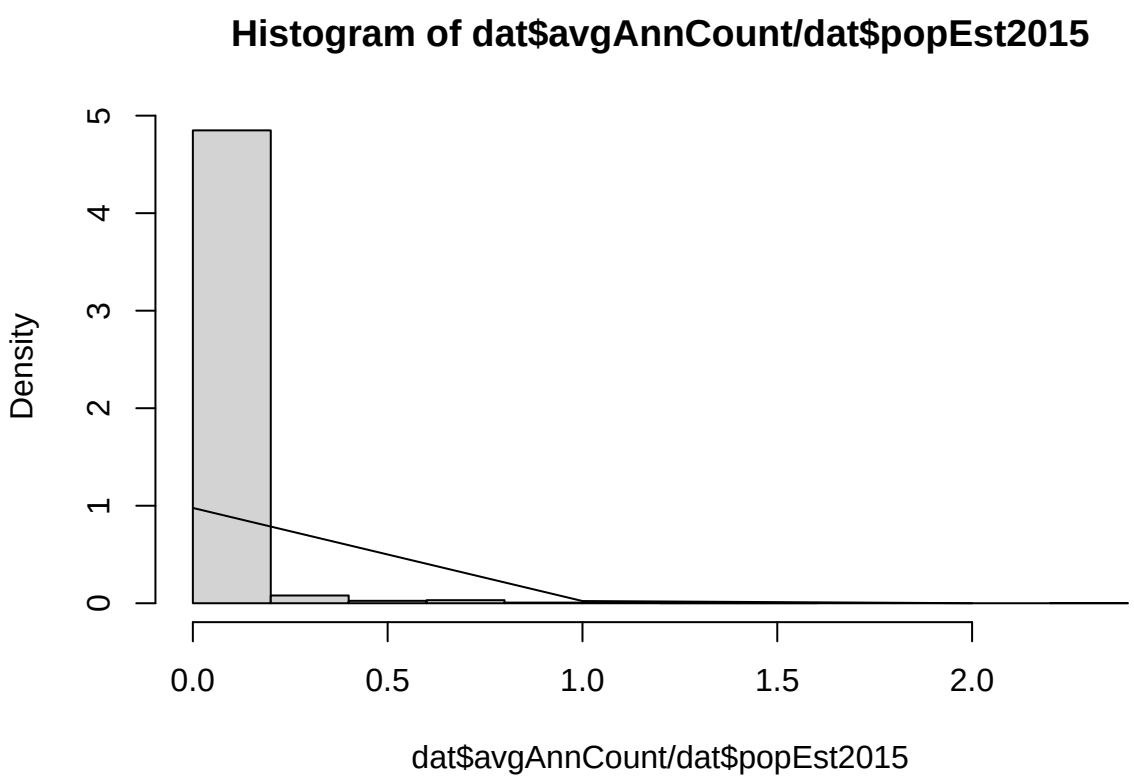

```
print(mean(dat$TARGET_deathRate))
```

```
## [1] 178.6452
```

```
print(var(dat$TARGET_deathRate))
```

```
## [1] 769.2961
```

```
hist(dat$avgAnnCount/dat$popEst2015, freq = F)#, ylim = c(0, 0.04))
lines(as.integer(min(dat$avgAnnCount/dat$popEst2015)):as.integer(max(dat$avgAnnCount/dat$popEst2015)), d
```



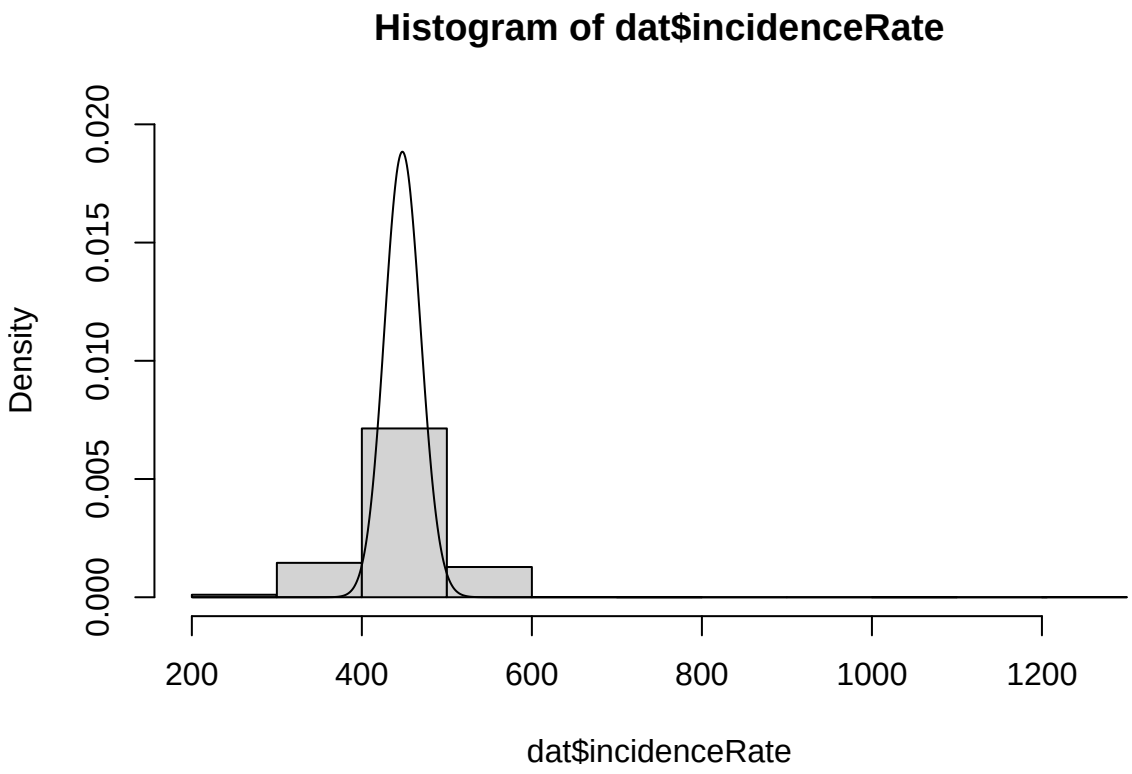

```
mean(dat$avgDeathsPerYear/dat$popEst2015)
```

```
## [1] 0.002287129
```

```
var(dat$avgDeathsPerYear/dat$popEst2015)
```

```
## [1] 3.729806e-07
```

```
hist(dat$incidenceRate, freq = F, ylim = c(0, 0.02))
lines(as.integer(min(dat$incidenceRate)):as.integer(max(dat$incidenceRate)), dpois(as.integer(min(dat$i
```




```
print(mean(dat$incidenceRate))
```

```
## [1] 448.1764
```

```
print(var(dat$incidenceRate))
```

```
## [1] 2982.145
```

Model fits

```
# poisson fit
state_inc_pop_pois <- dat %>% glm(formula = TARGET_deathRate ~ medIncome + State + popEst2015, family=po
summary(state_inc_pop_pois)
```

```
##
## Call:
## glm(formula = TARGET_deathRate ~ medIncome + State + popEst2015,
##     family = poisson(), data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.7899  -0.9173   0.0055   0.9062  11.6936
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.548e+00  1.957e-02 283.437  < 2e-16 ***
## medIncome   -4.502e-06  1.435e-07 -31.378  < 2e-16 ***
## StateAL     -1.111e-01  1.995e-02  -5.571 2.54e-08 ***
## StateAR     -7.953e-02  1.958e-02  -4.062 4.86e-05 ***
## StateAZ     -3.451e-01  2.765e-02 -12.482  < 2e-16 ***
## StateCA     -2.272e-01  2.063e-02 -11.015  < 2e-16 ***
## StateCO     -3.604e-01  2.065e-02 -17.454  < 2e-16 ***
## StateCT     -1.658e-01  3.318e-02  -4.998 5.78e-07 ***
## StateDC     -1.967e-02  5.528e-02  -0.356 0.722031
## StateDE     -9.943e-02  4.658e-02  -2.134 0.032813 *
## StateFL     -1.562e-01  1.995e-02  -7.829 4.90e-15 ***
## StateGA     -1.527e-01  1.871e-02  -8.160 3.35e-16 ***
## StateHI     -3.009e-01  4.533e-02  -6.637 3.19e-11 ***
## StateIA     -2.000e-01  1.917e-02 -10.433  < 2e-16 ***
## StateID     -3.036e-01  2.154e-02 -14.092  < 2e-16 ***
## StateIL     -1.067e-01  1.901e-02  -5.612 2.00e-08 ***
## StateIN     -8.772e-02  1.911e-02  -4.589 4.46e-06 ***
## StateKS     -2.028e-01  1.914e-02 -10.591  < 2e-16 ***
## StateKY      6.348e-04  1.881e-02   0.034 0.973082
## StateLA     -7.386e-02  1.984e-02  -3.723 0.000197 ***
## StateMA     -1.453e-01  2.723e-02  -5.335 9.54e-08 ***
## StateMD     -7.295e-02  2.326e-02  -3.136 0.001710 **
## StateME     -1.317e-01  2.552e-02  -5.161 2.45e-07 ***
## StateMI     -1.645e-01  1.948e-02  -8.443  < 2e-16 ***
## StateMN     -2.181e-01  1.941e-02 -11.237  < 2e-16 ***
## StateMO     -1.165e-01  1.896e-02  -6.144 8.04e-10 ***
## StateMS     -7.530e-02  1.948e-02  -3.864 0.000111 ***
## StateMT     -2.559e-01  2.101e-02 -12.182  < 2e-16 ***
## StateNC     -1.783e-01  1.928e-02  -9.248  < 2e-16 ***
## StateND     -2.243e-01  2.068e-02 -10.847  < 2e-16 ***
## StateNE     -2.474e-01  1.968e-02 -12.571  < 2e-16 ***
## StateNH     -1.354e-01  2.982e-02  -4.540 5.63e-06 ***
## StateNJ     -9.329e-02  2.430e-02  -3.839 0.000123 ***
## StateNM     -3.131e-01  2.282e-02 -13.720  < 2e-16 ***
## StateNV     -1.216e-01  2.525e-02  -4.815 1.47e-06 ***
## StateNY     -1.560e-01  2.004e-02  -7.781 7.17e-15 ***
## StateOH     -9.885e-02  1.924e-02  -5.138 2.77e-07 ***
## StateOK     -8.282e-02  1.945e-02  -4.259 2.05e-05 ***
## StateOR     -1.967e-01  2.184e-02  -9.009  < 2e-16 ***
## StatePA     -1.547e-01  1.993e-02  -7.761 8.40e-15 ***
## StateRI     -1.563e-01  4.256e-02  -3.673 0.000239 ***
## StateSC     -1.265e-01  2.073e-02  -6.100 1.06e-09 ***
## StateSD     -2.384e-01  2.035e-02 -11.720  < 2e-16 ***
## StateTN     -6.076e-02  1.916e-02  -3.172 0.001515 **
## StateTX     -1.964e-01  1.830e-02 -10.731  < 2e-16 ***
## StateUT     -3.846e-01  2.404e-02 -15.999  < 2e-16 ***
## StateVA     -1.029e-01  1.873e-02  -5.491 3.99e-08 ***
## StateVT     -1.459e-01  2.668e-02  -5.468 4.54e-08 ***
## StateWA     -2.024e-01  2.149e-02  -9.420  < 2e-16 ***
## StateWI     -1.687e-01  1.968e-02  -8.569  < 2e-16 ***
## StateWV     -8.860e-02  2.018e-02  -4.390 1.13e-05 ***
```

```
## StateWY     -2.220e-01  2.404e-02  -9.235  < 2e-16 ***
## popEst2015  -1.060e-08  4.808e-09  -2.205 0.027479 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13026.8  on 3016  degrees of freedom
## Residual deviance:  7555.3  on 2964  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

```
# neg bin fit
state_inc_pop_nb <- dat %>% MASS::glm.nb(formula = TARGET_deathRate ~ medIncome + State + popEst2015, da
summary(state_inc_pop_nb)
```

```
##
## Call:
## MASS::glm.nb(formula = TARGET_deathRate ~ medIncome + State +
##     popEst2015, data = ., init.theta = 118.3508188, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7153  -0.5830   0.0045   0.5662   6.8103
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.542e+00  3.153e-02 175.761  < 2e-16 ***
## medIncome   -4.428e-06  2.224e-07 -19.913  < 2e-16 ***
## StateAL     -1.088e-01  3.236e-02  -3.361 0.000776 ***
## StateAR     -7.718e-02  3.182e-02  -2.426 0.015280 *
## StateAZ     -3.421e-01  4.296e-02  -7.963 1.68e-15 ***
## StateCA     -2.255e-01  3.298e-02  -6.837 8.11e-12 ***
## StateCO     -3.588e-01  3.273e-02 -10.964  < 2e-16 ***
## StateCT     -1.662e-01  5.159e-02  -3.221 0.001277 **
## StateDC     -1.972e-02  8.826e-02  -0.223 0.823198
## StateDE     -9.806e-02  7.408e-02  -1.324 0.185601
## StateFL     -1.549e-01  3.222e-02  -4.808 1.52e-06 ***
## StateGA     -1.501e-01  3.030e-02  -4.954 7.28e-07 ***
## StateHI     -3.001e-01  6.833e-02  -4.392 1.12e-05 ***
## StateIA     -1.985e-01  3.091e-02  -6.422 1.34e-10 ***
## StateID     -3.020e-01  3.423e-02  -8.822  < 2e-16 ***
## StateIL     -1.046e-01  3.078e-02  -3.399 0.000676 ***
## StateIN     -8.608e-02  3.099e-02  -2.778 0.005476 **
## StateKS     -2.014e-01  3.088e-02  -6.521 6.97e-11 ***
## StateKY      2.029e-03  3.063e-02   0.066 0.947177
## StateLA     -7.152e-02  3.223e-02  -2.219 0.026472 *
## StateMA     -1.436e-01  4.300e-02  -3.339 0.000842 ***
## StateMD     -7.193e-02  3.733e-02  -1.927 0.053977 .
## StateME     -1.299e-01  4.107e-02  -3.163 0.001559 **
## StateMI     -1.627e-01  3.147e-02  -5.169 2.36e-07 ***
## StateMN     -2.161e-01  3.120e-02  -6.925 4.35e-12 ***
## StateMO     -1.144e-01  3.073e-02  -3.723 0.000197 ***
```

```
## StateMS     -7.231e-02  3.167e-02  -2.283 0.022432 *
## StateMT     -2.536e-01  3.359e-02  -7.548 4.43e-14 ***
## StateNC     -1.760e-01  3.116e-02  -5.650 1.61e-08 ***
## StateND     -2.228e-01  3.306e-02  -6.739 1.60e-11 ***
## StateNE     -2.452e-01  3.160e-02  -7.760 8.50e-15 ***
## StateNH     -1.336e-01  4.725e-02  -2.828 0.004683 **
## StateNJ     -9.247e-02  3.869e-02  -2.390 0.016846 *
## StateNM     -3.105e-01  3.615e-02  -8.589  < 2e-16 ***
## StateNV     -1.222e-01  4.045e-02  -3.022 0.002511 **
## StateNY     -1.541e-01  3.227e-02  -4.774 1.81e-06 ***
## StateOH     -9.731e-02  3.117e-02  -3.122 0.001799 **
## StateOK     -8.144e-02  3.157e-02  -2.580 0.009875 **
## StateOR     -1.942e-01  3.500e-02  -5.549 2.88e-08 ***
## StatePA     -1.528e-01  3.214e-02  -4.752 2.01e-06 ***
## StateRI     -1.558e-01  6.651e-02  -2.342 0.019186 *
## StateSC     -1.248e-01  3.357e-02  -3.717 0.000202 ***
## StateSD     -2.368e-01  3.261e-02  -7.263 3.78e-13 ***
## StateTN     -5.883e-02  3.114e-02  -1.890 0.058816 .
## StateTX     -1.936e-01  2.963e-02  -6.535 6.36e-11 ***
## StateUT     -3.825e-01  3.729e-02 -10.257  < 2e-16 ***
## StateVA     -1.011e-01  3.034e-02  -3.331 0.000865 ***
## StateVT     -1.434e-01  4.263e-02  -3.364 0.000768 ***
## StateWA     -2.004e-01  3.437e-02  -5.830 5.53e-09 ***
## StateWI     -1.674e-01  3.174e-02  -5.276 1.32e-07 ***
## StateWV     -8.637e-02  3.278e-02  -2.635 0.008423 **
## StateWY     -2.196e-01  3.800e-02  -5.778 7.58e-09 ***
## popEst2015  -1.010e-08  7.324e-09  -1.379 0.168041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(118.3508) family taken to be 1)
##
##     Null deviance: 5232.5  on 3016  degrees of freedom
## Residual deviance: 3051.8  on 2964  degrees of freedom
## AIC: 27080
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  118.35
##           Std. Err.:  5.12
##
##  2 x log-likelihood:  -26971.87
```