Stat 222- EDA

A.J. Torre

February 19, 2019

Data Cleaning/ Working with Initial Data Set

This data is pretty clean to begin with so there's only a few steps to clean it.

```
# read data into R
death_rate_data <- read.csv("C:/Users/AJ/OneDrive/Documents/Stat 222/deathrate.csv")

# do some cleaning, data is relatively clean so we don't have
# too much to do

# removing missing values
death_rate_data <- na.omit(death_rate_data)
# checking to make sure there are no missing values
nrow(death_rate_data[!complete.cases(death_rate_data), ])
```

```
## [1] 0
```

```
# removing any duplicated values
death_rate_data <- distinct(death_rate_data)
```

Next, after cleaning the data, we can add a column for poverty rate in each county by dividing the number of people in poverty by the total population.

```
# getting poverty rate by divding poverty by population, see
# what data says about what exactly poverty means?
death_rate_data <- transform(death_rate_data, new = Poverty/Population)
colnames(death_rate_data)[colnames(death_rate_data) == "new"] <- "PovertyRate"

# check to see that this happened
head(death_rate_data)
```

```
##   X Year               County  FIPS Deathrate Population Poverty
## 1 1 1999 Abbeville County, SC 45001         1      25921    3257
## 2 2 1999    Acadia Parish, LA 22001         7      58762   12461
## 3 3 1999  Accomack County, VA 51001         5      37614    6107
## 4 4 1999       Ada County, ID 16001         7     294292   24964
## 5 5 1999     Adair County, IA 19001         1       8298     697
## 6 6 1999     Adair County, KY 21001         5      17054    3656
##   PovertyRate
## 1  0.12565102
## 2  0.21205881
## 3  0.16235976
## 4  0.08482731
## 5  0.08399614
## 6  0.21437786
```

Next, we will subset the data from 2014 to plot it as 2014 is the most recent year in our provided data set. Then, we plot the OD deathrate against the poverty rate and see if there are any clear trends.

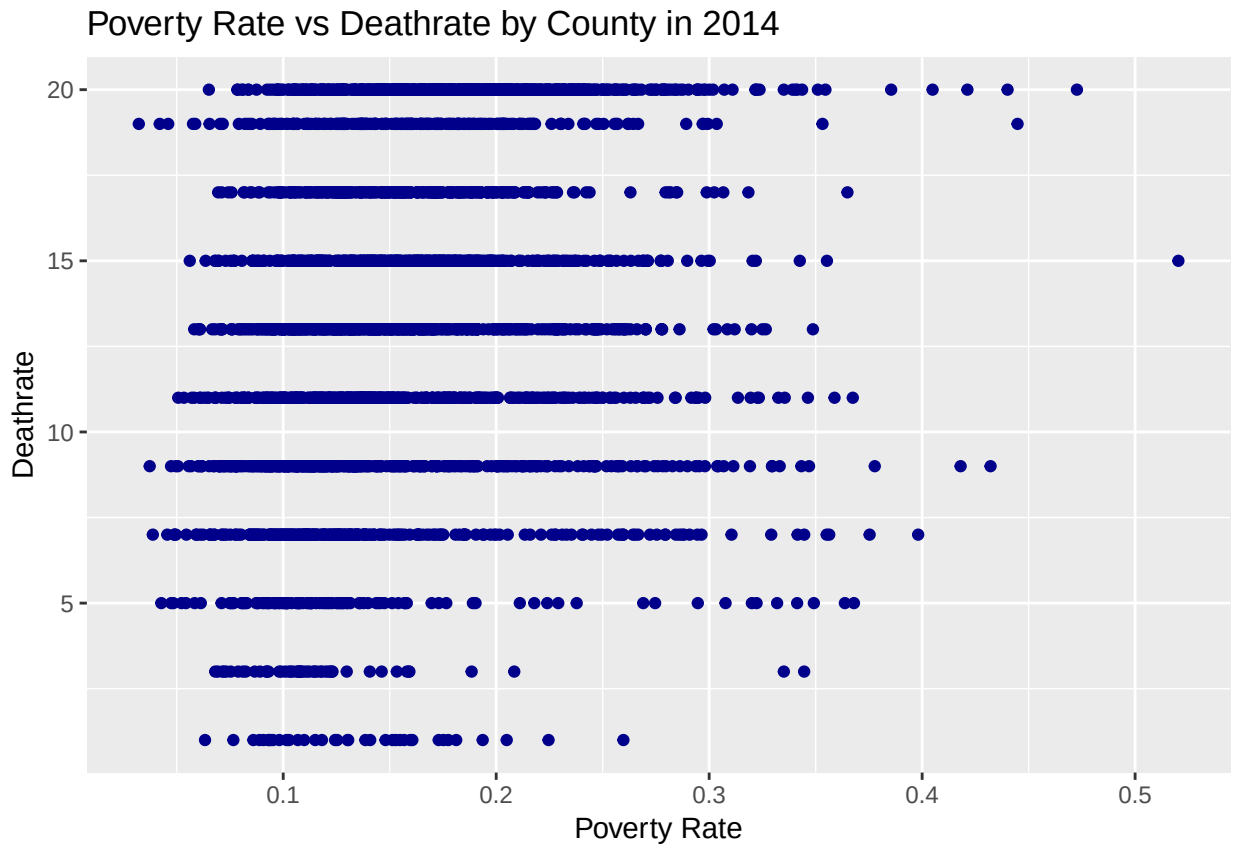
```
# get only 2014 data
death_rate_14 <- subset(death_rate_data, Year == 2014)

head(death_rate_14)
```

```
##           X Year               County  FIPS Deathrate Population Poverty
## 47020 47030 2014 Abbeville County, SC 45001         9      24965    5178
## 47021 47031 2014    Acadia Parish, LA 22001        20      62486   13527
## 47022 47032 2014  Accomack County, VA 51001        13      33021    6319
## 47023 47033 2014       Ada County, ID 16001        17     426236   48083
## 47024 47034 2014     Adair County, IA 19001        11       7454     751
## 47025 47035 2014     Adair County, KY 21001        19      19204    4644
##       PovertyRate
## 47020   0.2074104
## 47021   0.2164805
## 47022   0.1913631
## 47023   0.1128084
## 47024   0.1007513
## 47025   0.2418246
```

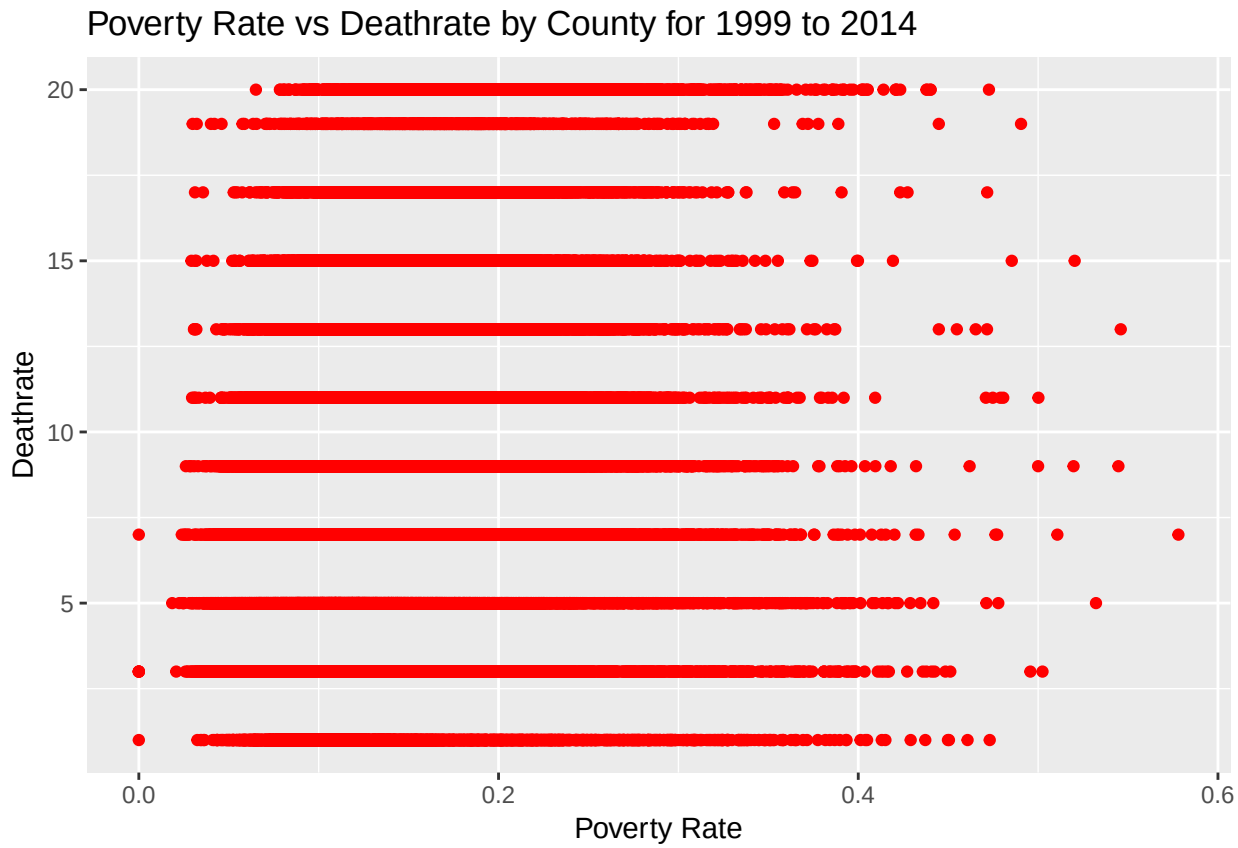
```
# just for 2014, plot poverty rate against death
plot_poverty_death_14 <- ggplot(death_rate_14, aes(x = PovertyRate,
    y = Deathrate)) + geom_point(color = "darkblue") + ggtitle("Poverty Rate vs Deathrate by County in 
    xlab("Poverty Rate") + ylab("Deathrate")

# don't see any very clear correlation?!
plot_poverty_death_14
```



Actually, if we plot all the deathrates against all the poverty rates, there is still no clear correlation.

```
# for all years,plot poverty rate against death rate
points_poverty_death <- ggplot(death_rate_data, aes(x = PovertyRate,
    y = Deathrate)) + geom_point(color = "red") + ggtitle("Poverty Rate vs Deathrate by County for 1999
    xlab("Poverty Rate") + ylab("Deathrate")
points_poverty_death
```



Next, we try to look at three counties that had high deathrates due to overdose in 2014, and we examine their trends overtime.

```
# look at which counties have high deathrates
head(death_rate_14[order(-death_rate_14$Deathrate), ])
```

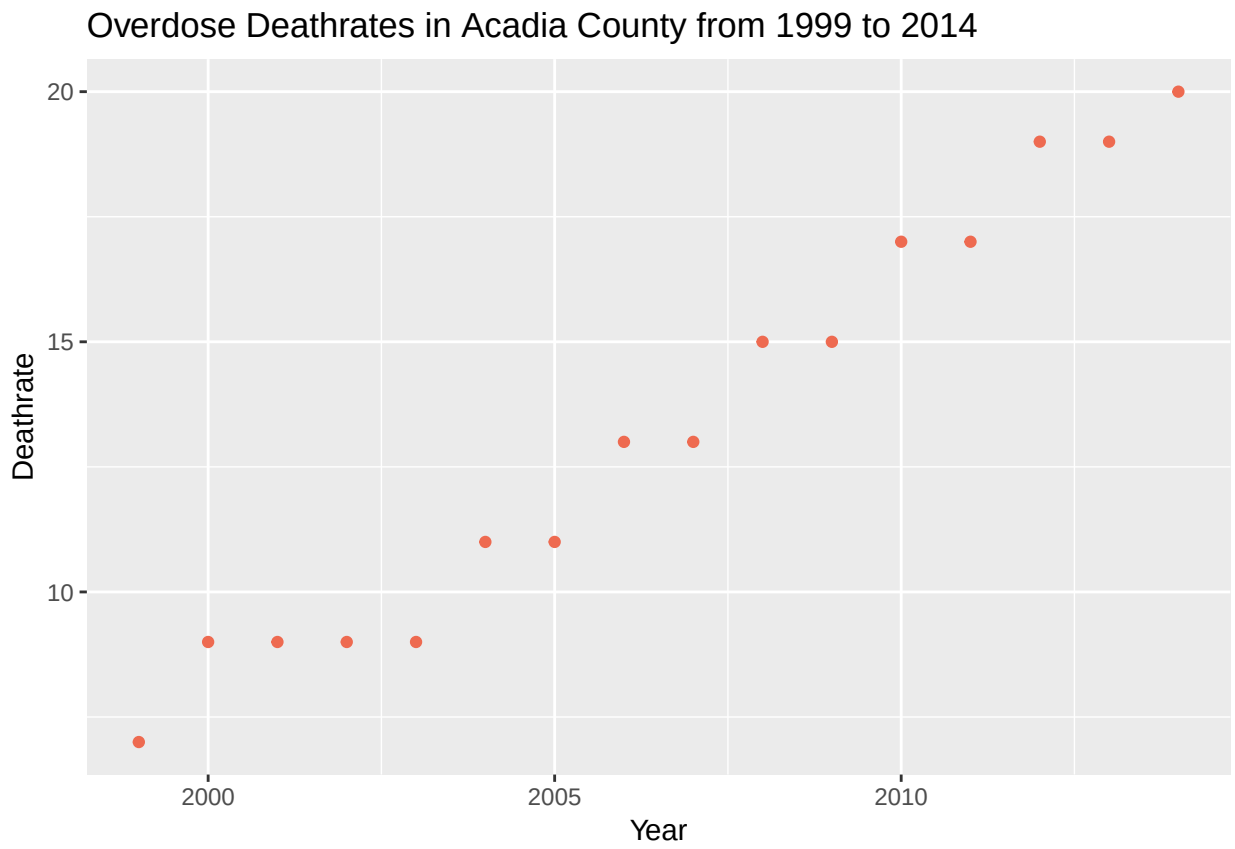
```
##           X Year             County  FIPS Deathrate Population Poverty
## 47021 47031 2014  Acadia Parish, LA 22001        20      62486   13527
## 47028 47038 2014   Adams County, CO  8001        20     480718   61384
## 47036 47046 2014   Adams County, OH 39001        20      28129    6864
## 47046 47056 2014 Alamosa County, CO  8003        20      16177    3385
## 47050 47060 2014  Alcona County, MI 26001        20      10454    1803
## 47057 47067 2014 Alfalfa County, OK 40003        20       5790     783
##       PovertyRate
## 47021   0.2164805
## 47028   0.1276923
## 47036   0.2440186
## 47046   0.2092477
## 47050   0.1724699
## 47057   0.1352332
```

```
# see increase over time for Acadia county in LA
acadia_la_data <- subset(death_rate_data, County == "Acadia Parish, LA")
acadia_la_data
```

```
##           X Year            County  FIPS Deathrate Population Poverty
```

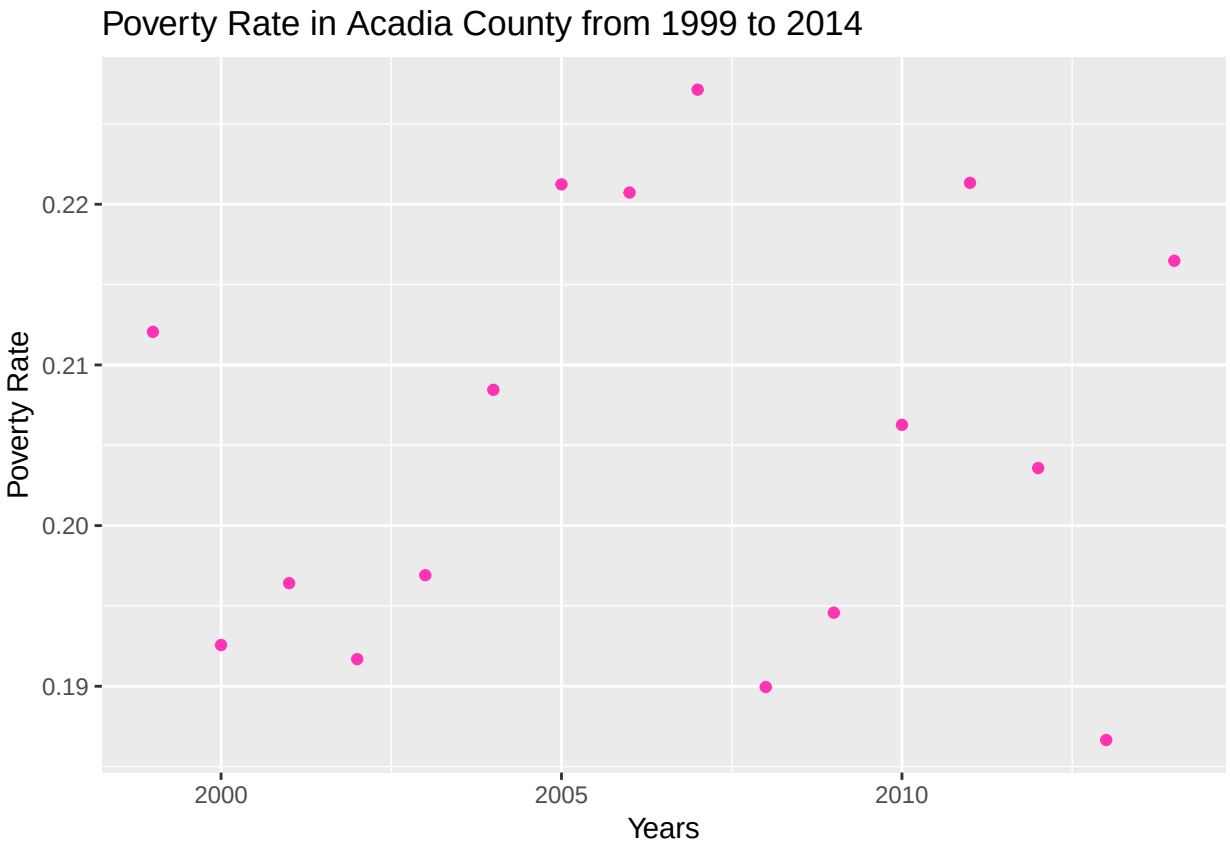
```
## 2         2 1999 Acadia Parish, LA 22001         7      58762   12461
## 3137   3138 2000 Acadia Parish, LA 22001         9      58795   11322
## 6273   6274 2001 Acadia Parish, LA 22001         9      58844   11558
## 9409   9410 2002 Acadia Parish, LA 22001         9      59065   11322
## 12545 12546 2003 Acadia Parish, LA 22001         9      59194   11656
## 15680 15681 2004 Acadia Parish, LA 22001        11      59223   12345
## 18814 18816 2005 Acadia Parish, LA 22001        11      59524   13169
## 21948 21951 2006 Acadia Parish, LA 22001        13      60522   13359
## 25083 25086 2007 Acadia Parish, LA 22001        13      60762   13801
## 28217 28221 2008 Acadia Parish, LA 22001        15      61115   11609
## 31351 31356 2009 Acadia Parish, LA 22001        15      61451   11957
## 34485 34491 2010 Acadia Parish, LA 22001        17      61861   12760
## 37619 37626 2011 Acadia Parish, LA 22001        17      61766   13671
## 40753 40761 2012 Acadia Parish, LA 22001        19      61873   12596
## 43887 43896 2013 Acadia Parish, LA 22001        19      62169   11604
## 47021 47031 2014 Acadia Parish, LA 22001        20      62486   13527
##       PovertyRate
## 2       0.2120588
## 3137    0.1925674
## 6273    0.1964176
## 9409    0.1916871
## 12545   0.1969118
## 15680   0.2084494
## 18814   0.2212385
## 21948   0.2207297
## 25083   0.2271321
## 28217   0.1899534
## 31351   0.1945778
## 34485   0.2062689
## 37619   0.2213354
## 40753   0.2035783
## 43887   0.1866525
## 47021   0.2164805
```

```
acadia__death_plot <- ggplot(acadia_la_data, aes(x = Year, y = Deathrate)) +
    geom_point(color = "coral2") + ggtitle("Overdose Deathrates in Acadia County from 1999 to 2014") +
    xlab("Year") + ylab("Deathrate")
acadia__death_plot
```



Out of curiosity, we can look at the rate of poverty over time in Acadia, LA to see if there is any clear increase in the poverty rate, just as there is a clear increase in deathrate. For Acadia county, it doesn't look like the deathrates and poverty rates are growing together.

```
# plot poverty in Acadia county over time
acadia_poverty_plot <- ggplot(acadia_la_data, aes(x = Year, y = PovertyRate)) +
    geom_point(color = "maroon1") + ggtitle("Poverty Rate in Acadia County from 1999 to 2014") +
    xlab("Years") + ylab("Poverty Rate")
acadia_poverty_plot
```



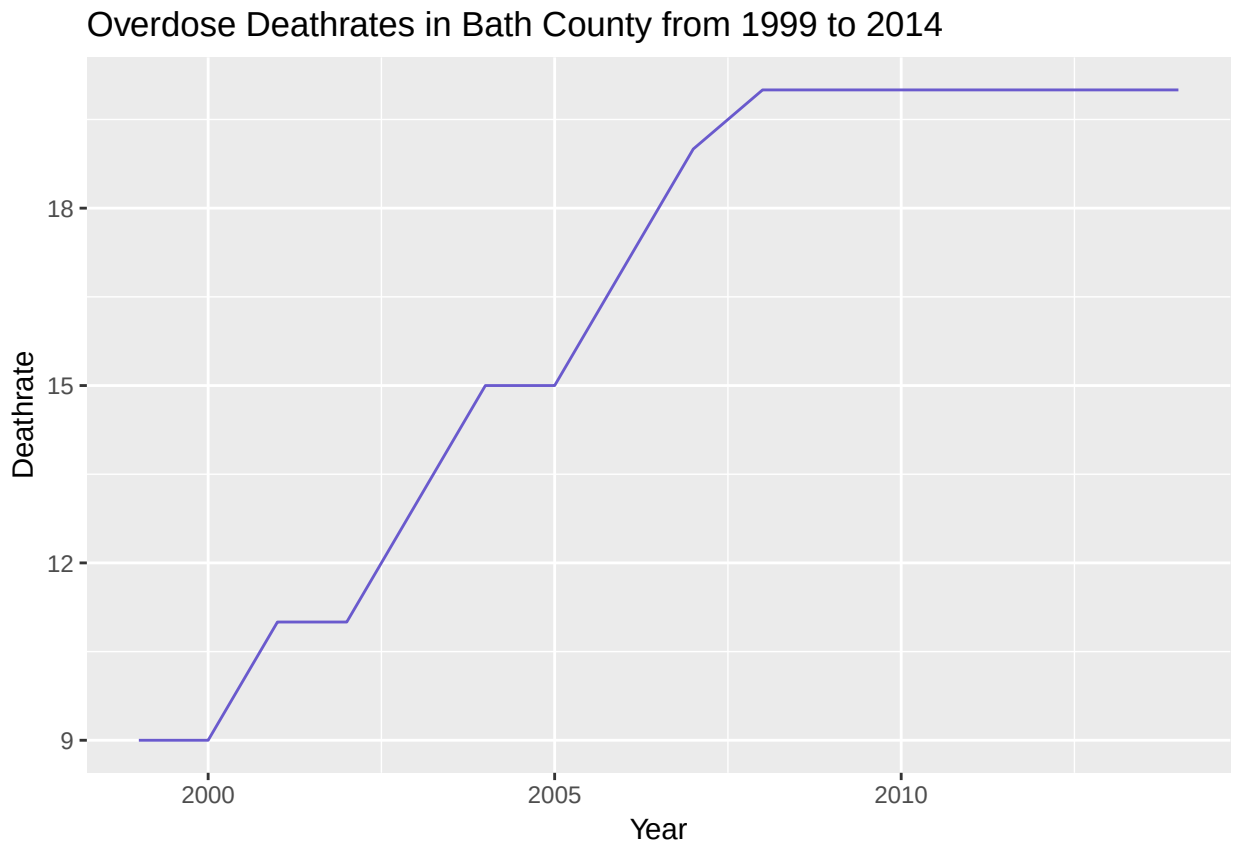
We will do the same for 2 other counties that have high death rates in 2014: Bath, KY and Abbeville, SC.

```
# see increase in OD deaths over time
bath_ky_data <- subset(death_rate_data, County == "Bath County, KY")
bath_ky_data
```

```
##           X Year          County  FIPS Deathrate Population Poverty
## 148     148 1999 Bath County, KY 21011         9      10911    2209
## 3283   3284 2000 Bath County, KY 21011         9      11128    2104
## 6419   6420 2001 Bath County, KY 21011        11      11280    2196
## 9555   9556 2002 Bath County, KY 21011        11      11403    2140
## 12691 12692 2003 Bath County, KY 21011        13      11428    2115
## 15826 15827 2004 Bath County, KY 21011        15      11447    2324
## 18960 18962 2005 Bath County, KY 21011        15      11538    2519
## 22094 22097 2006 Bath County, KY 21011        17      11558    2495
## 25229 25232 2007 Bath County, KY 21011        19      11470    2724
## 28363 28367 2008 Bath County, KY 21011        20      11647    3094
## 31497 31502 2009 Bath County, KY 21011        20      11544    2872
## 34631 34637 2010 Bath County, KY 21011        20      11623    3191
## 37765 37772 2011 Bath County, KY 21011        20      11715    2974
## 40899 40907 2012 Bath County, KY 21011        20      11799    2971
## 44033 44042 2013 Bath County, KY 21011        20      12008    3135
## 47167 47177 2014 Bath County, KY 21011        20      12206    2733
##       PovertyRate
## 148     0.2024562
## 3283    0.1890726
## 6419    0.1946809
```

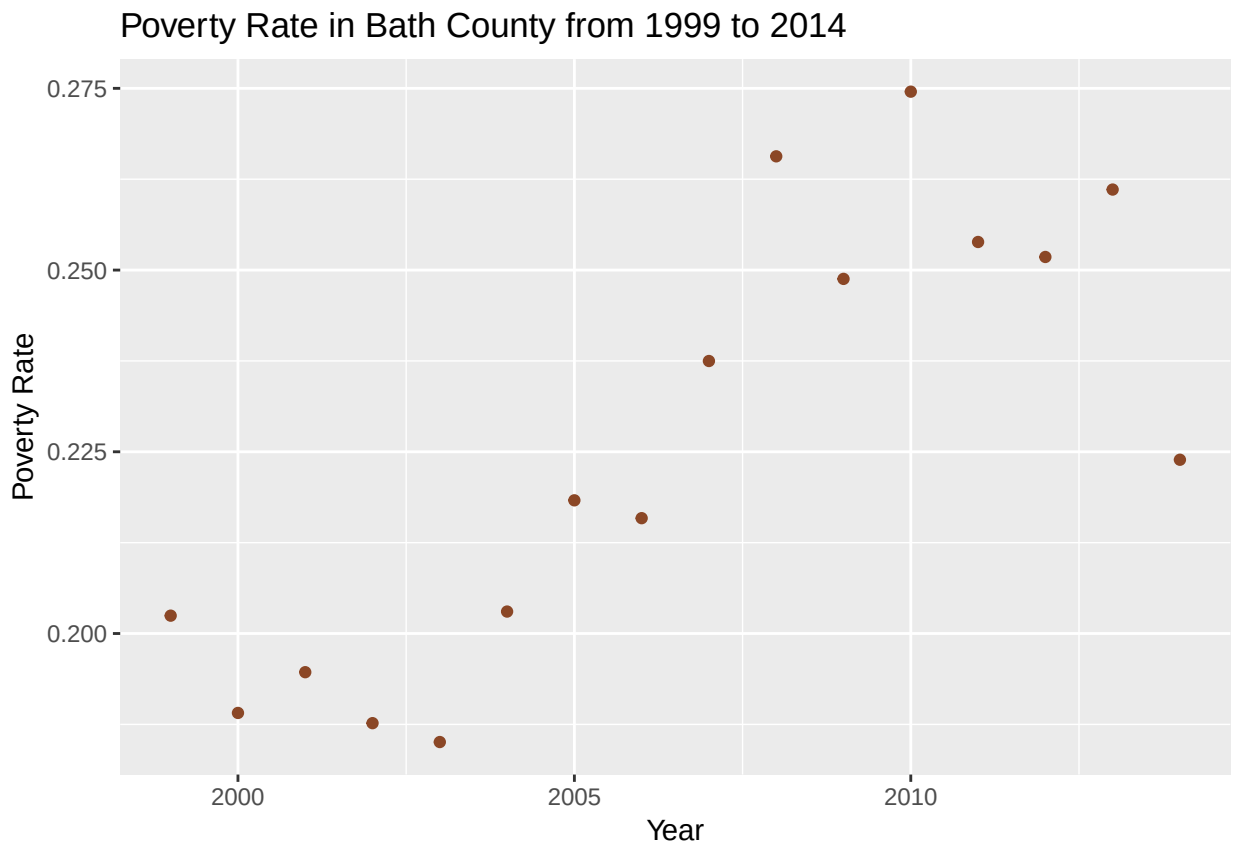
```
## 9555    0.1876699
## 12691   0.1850718
## 15826   0.2030226
## 18960   0.2183221
## 22094   0.2158678
## 25229   0.2374891
## 28363   0.2656478
## 31497   0.2487872
## 34631   0.2745419
## 37765   0.2538626
## 40899   0.2518010
## 44033   0.2610759
## 47167   0.2239063
```

```
bath_death_plot <- ggplot(bath_ky_data, aes(x = Year, y = Deathrate)) +
    geom_line(color = "slateblue") + ggtitle("Overdose Deathrates in Bath County from 1999 to 2014") +
    xlab("Year") + ylab("Deathrate")
bath_death_plot
```



Below, we plot the poverty rate in Bath County. It does seem like an overall trend that poverty is increasing overtime, but also from the graph it also seems like poverty drastically decreased in 2014.

```
# looking at poverty overtime in Bath county
bath_poverty_plot <- ggplot(bath_ky_data, aes(x = Year, y = PovertyRate)) +
    geom_point(color = "sienna4") + ggtitle("Poverty Rate in Bath County from 1999 to 2014") +
    xlab("Year") + ylab("Poverty Rate")
bath_poverty_plot
```



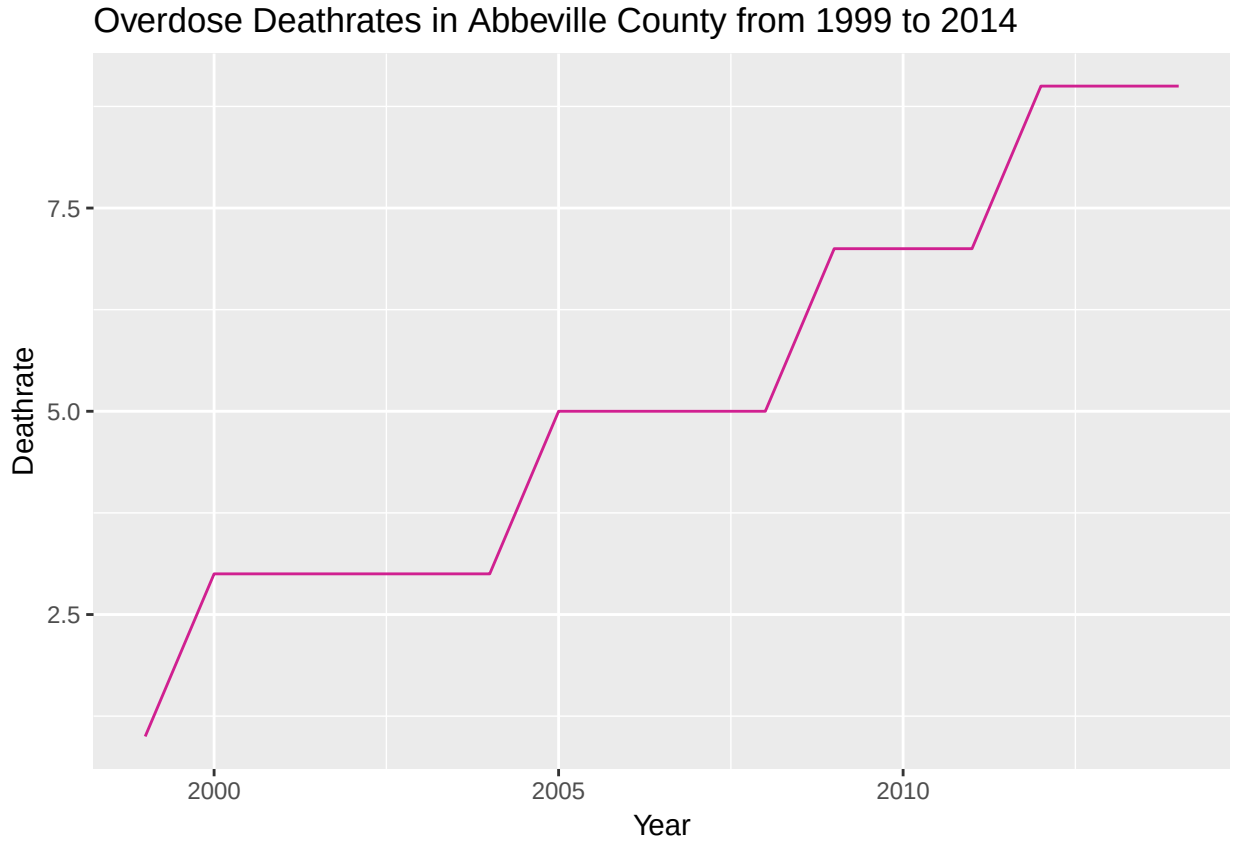
Next, we plot the death and poverty rates for Abbeville County.

```
# see increase in OD deaths
abbeville_sc_data <- subset(death_rate_data, County == "Abbeville County, SC")
abbeville_sc_data
```

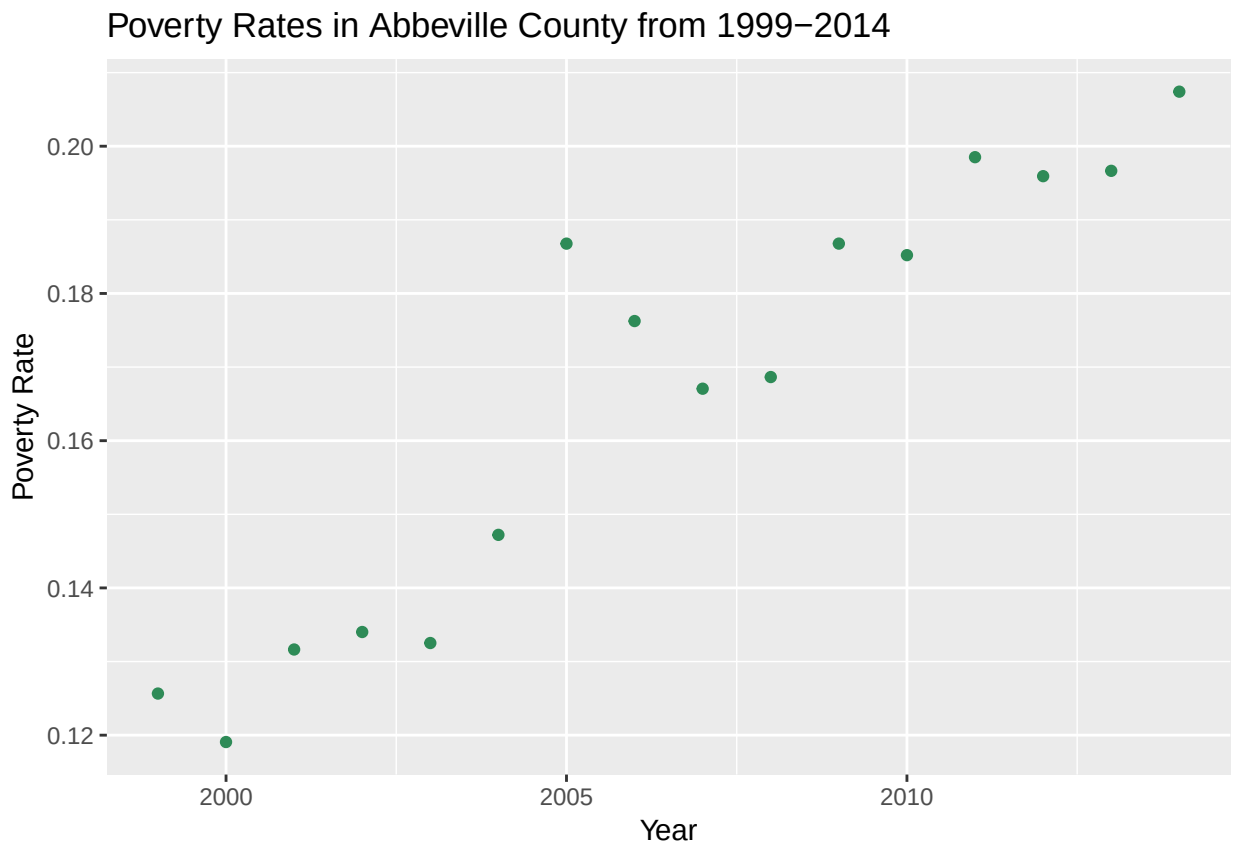
```
##           X Year               County  FIPS Deathrate Population Poverty
## 1         1 1999 Abbeville County, SC 45001         1      25921    3257
## 3136   3137 2000 Abbeville County, SC 45001         3      26229    3123
## 6272   6273 2001 Abbeville County, SC 45001         3      26330    3466
## 9408   9409 2002 Abbeville County, SC 45001         3      26311    3526
## 12544 12545 2003 Abbeville County, SC 45001         3      26306    3486
## 15679 15680 2004 Abbeville County, SC 45001         3      26235    3862
## 18813 18815 2005 Abbeville County, SC 45001         5      25995    4855
## 21947 21950 2006 Abbeville County, SC 45001         5      25821    4551
## 25082 25085 2007 Abbeville County, SC 45001         5      25745    4301
## 28216 28220 2008 Abbeville County, SC 45001         5      25699    4334
## 31350 31355 2009 Abbeville County, SC 45001         7      25614    4784
## 34484 34490 2010 Abbeville County, SC 45001         7      25345    4694
## 37618 37625 2011 Abbeville County, SC 45001         7      25117    4986
## 40752 40760 2012 Abbeville County, SC 45001         9      25065    4911
## 43886 43895 2013 Abbeville County, SC 45001         9      25008    4918
## 47020 47030 2014 Abbeville County, SC 45001         9      24965    5178
##       PovertyRate
## 1       0.1256510
## 3136    0.1190667
## 6272    0.1316369
```

```
## 9408    0.1340124
## 12544   0.1325173
## 15679   0.1472079
## 18813   0.1867667
## 21947   0.1762519
## 25082   0.1670616
## 28216   0.1686447
## 31350   0.1867729
## 34484   0.1852042
## 37618   0.1985110
## 40752   0.1959306
## 43886   0.1966571
## 47020   0.2074104
```

```
abbeville_death_plot <- ggplot(abbeville_sc_data, aes(x = Year,
    y = Deathrate)) + geom_line(color = "violetred") + ggtitle("Overdose Deathrates in Abbeville County
    xlab("Year") + ylab("Deathrate")
abbeville_death_plot
```




```
# looking at poverty overtime in Abbeville county
abbeville_poverty_plot <- ggplot(abbeville_sc_data, aes(x = Year,
    y = PovertyRate)) + geom_point(color = "seagreen") + ggtitle("Poverty Rates in Abbeville County fro
    xlab("Year") + ylab("Poverty Rate")
abbeville_poverty_plot
```

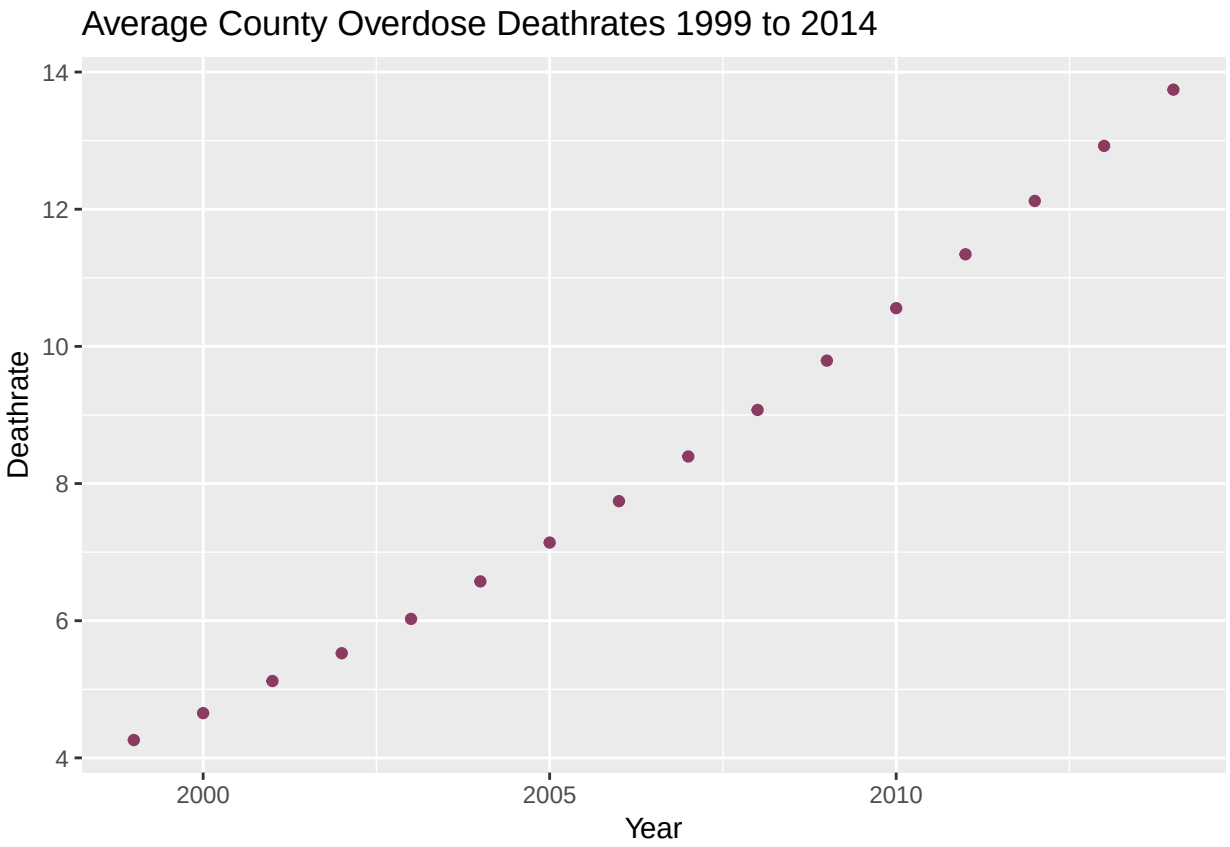


Next, we plot the average county deathrate over the years. We use the `dplyr` package to do some subsetting and averaging of data. We can see a clear increase in deathrates due to overdose over this 15 year period.

```
# getting yearly increase in overall death rates data
yearly_death <- death_rate_data %>% group_by(Year) %>% summarize(mean_death_rate = mean(Deathrate,
    na.rm = TRUE))
typeof(yearly_death)
```

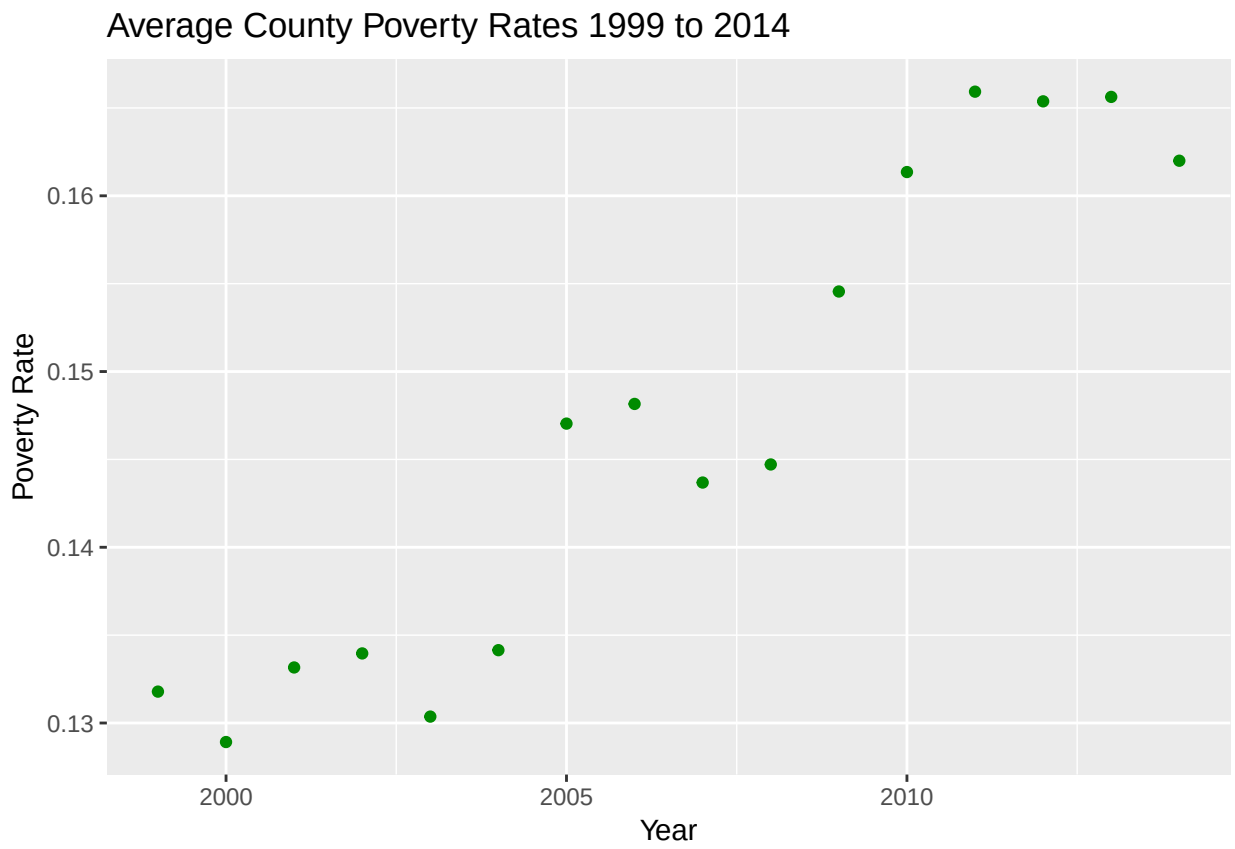
```
## [1] "list"
```

```
# plot average death due to OD overtime
yearly_death_plot <- ggplot(yearly_death, aes(x = Year, y = mean_death_rate)) +
    geom_point(color = "hotpink4") + ggtitle("Average County Overdose Deathrates 1999 to 2014") +
    xlab("Year") + ylab("Deathrate")
yearly_death_plot
```



Next, we can look at the average poverty rate per county and see if there are any trends with the poverty rate. We also see a general trend of poverty increasing over time, which (to me) is surprising.

```
# want to see how poverty is over time
yearly_poverty <- death_rate_data %>% group_by(Year) %>% summarize(mean_poverty_rate = mean(PovertyRate
    na.rm = TRUE))
yearly_poverty_plot <- ggplot(yearly_poverty, aes(x = Year, y = mean_poverty_rate)) +
    geom_point(color = "green4") + ggtitle("Average County Poverty Rates 1999 to 2014") +
    xlab("Year") + ylab("Poverty Rate")
yearly_poverty_plot
```

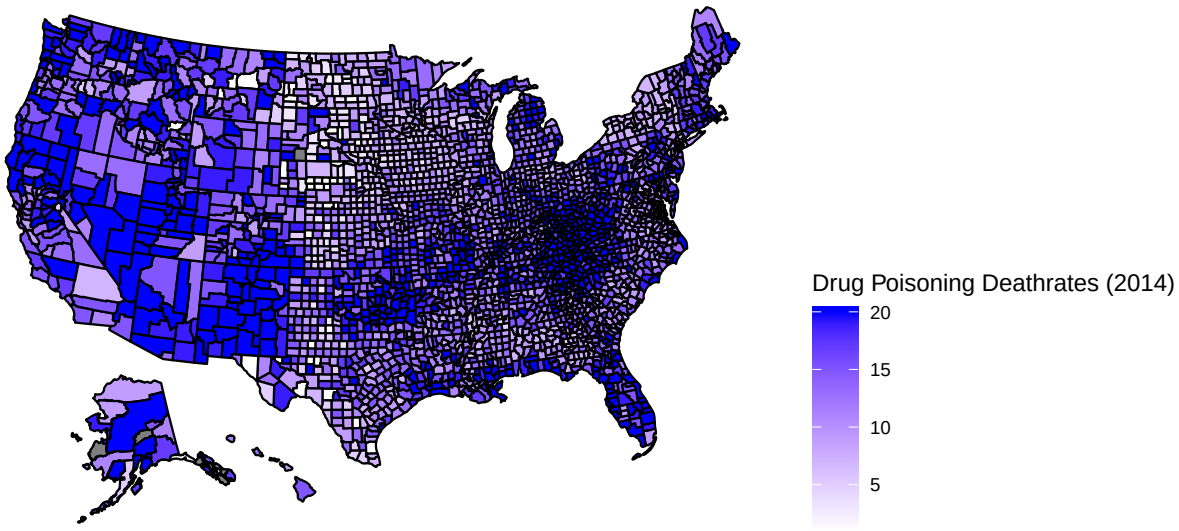



```
# see that poverty has been increasing overtime
```

Next, using the `usmap` package in R, we can create our own heatmaps of the US with the deathrate data.

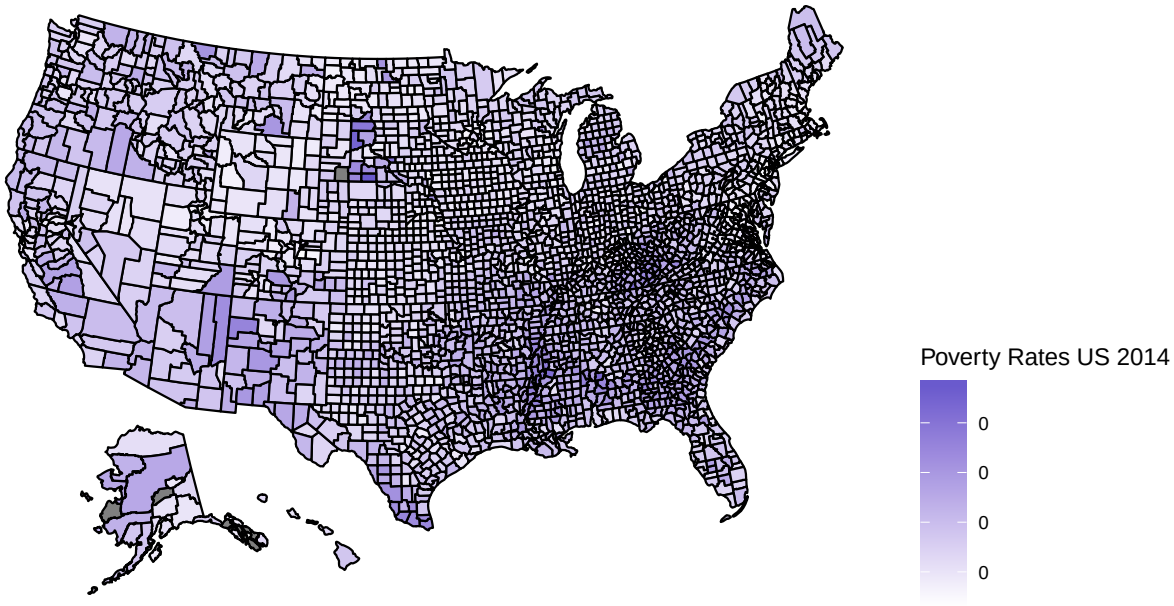
```
# in order to use the usmap package, need column to be
# exactly named 'fips'
colnames(death_rate_14)[colnames(death_rate_14) == "FIPS"] <- "fips"

# creating heatmap for deathrate data 2014
plot_usmap(data = death_rate_14, values = "Deathrate", lines = "black") +
    scale_fill_continuous(low = "white", high = "blue", name = "Drug Poisoning Deathrates (2014)",
        label = scales::comma) + theme(legend.position = "right")
```



We can also make a heatmap with the poverty rates to again see if there is any overlap between the poverty and overdose deathrates.

```
# plot poverty in US by county in 2014
plot_usmap(data = death_rate_14, values = "PovertyRate", lines = "black") +
    scale_fill_continuous(low = "white", high = "slateblue",
        name = "Poverty Rates US 2014", label = scales::comma) +
    theme(legend.position = "right")
```



Working with Outside Data

For the below two graphs, I downloaded the drug overdose deathrates for 2016 and 2017 from the CDC website. I also downloaded data on the percent change of deathrates due to drug OD from 2016 to 2017. The heat maps are made using the `usmap` package in R.

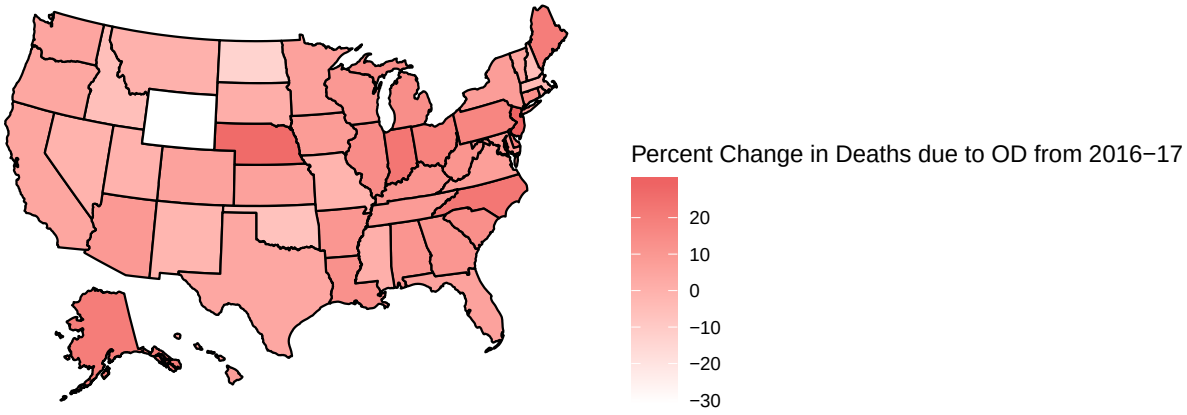
This map is the percent change in deaths due OD from 2016 to 2017 at the state level.

```
# downloaded data from CDC website on states' percent change
# in deaths due to OD from 2016 to 2017
change_16_17_deaths <- read.csv("C:/Users/AJ/OneDrive/Documents/Stat 222/DrugODDeathRateIncreaseFrom201
head(change_16_17_deaths)
```

```
##   state significant change
## 1    AL         Yes   11.1
## 2    AK          No   20.2
## 3    AZ         Yes    9.4
## 4    AR          No   10.7
## 5    CA         Yes    4.5
## 6    CO          No    6.0
```

```
plot_usmap(data = change_16_17_deaths, values = "change", lines = "black") +
    scale_fill_continuous(low = "white", high = "indianred2",
```

```
        name = "Percent Change in Deaths due to OD from 2016-17",
        label = scales::comma) + theme(legend.position = "right")
```

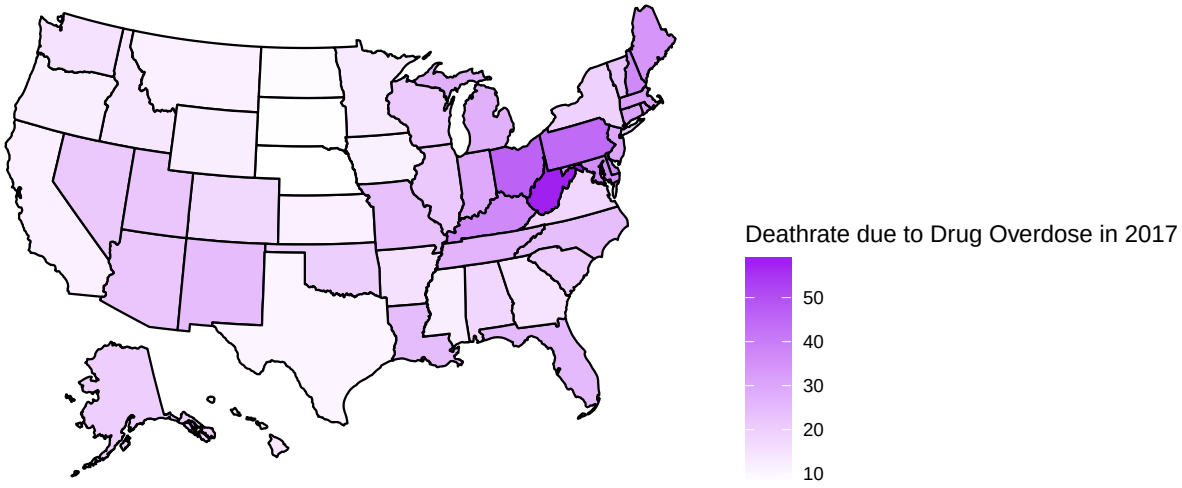


This heatmap is the deathrate due to OD in 2017 at the state level

```
# make a heat map of deaths due to Overdose in 2017, plotted
# deathrates so deaths per 100,000 people
death_17_rates <- read.csv("C:/Users/AJ/OneDrive/Documents/Stat 222/drug-overdose-deaths-state-2017.csv
head(death_17_rates)
```

```
##   State        range rate  number
## 1    AL 16.1 to 18.5 18.0    835
## 2    AK 18.6 to 21.0 20.2    147
## 3    AZ 21.1 to 57.0 22.2  1,532
## 4    AR 13.6 to 16.0 15.5    446
## 5    CA 11.1 to 13.5 11.7  4,868
## 6    CO 16.1 to 18.5 17.6  1,015
```

```
colnames(death_17_rates)[colnames(death_17_rates) == "State"] <- "state"
plot_usmap(data = death_17_rates, values = "rate", lines = "black") +
    scale_fill_continuous(low = "white", high = "purple", name = "Deathrate due to Drug Overdose in 201
        label = scales::comma) + theme(legend.position = "right")
```

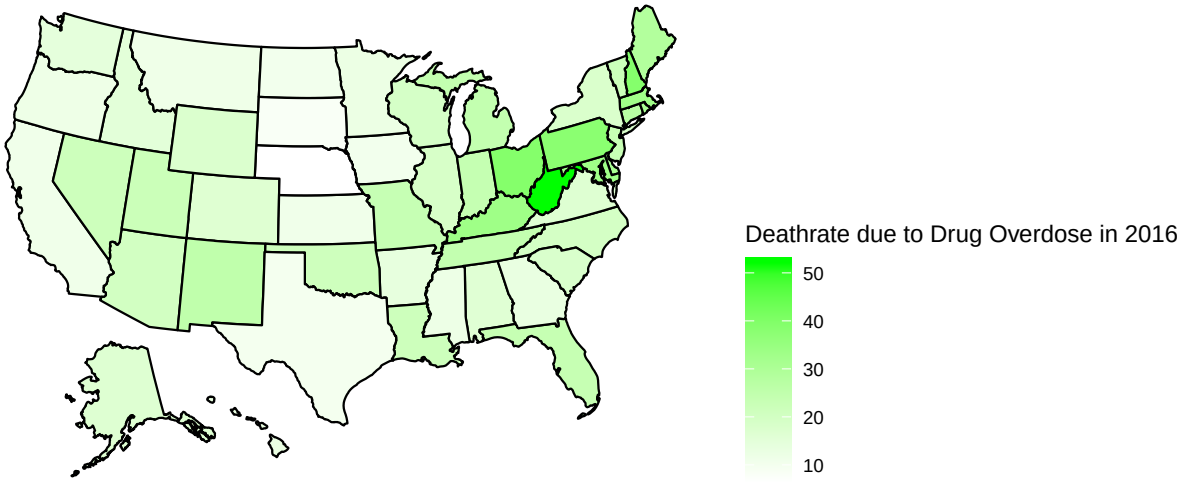


Lastly, this heatmap is the deathrate due to OD in 2016 at the state level.

```
# make a heat map of deaths due to overdose in 2016, plotted
# deathrates so deaths per 100,000 people
death_16_rates <- read.csv("C:/Users/AJ/OneDrive/Documents/Stat 222/death-rate-overdose-2016.csv")
head(death_16_rates)
```

```
##   State        range rate number
## 1    AL 16.1 to 18.5 16.2    756
## 2    AK 16.1 to 18.5 16.8    128
## 3    AZ 18.6 to 21.0 20.3  1,382
## 4    AR 13.6 to 16.0 14.0    401
## 5    CA 11.1 to 13.5 11.2  4,654
## 6    CO 16.1 to 18.5 16.6    942
```

```
colnames(death_16_rates)[colnames(death_16_rates) == "State"] <- "state"
plot_usmap(data = death_16_rates, values = "rate", lines = "black") +
    scale_fill_continuous(low = "white", high = "green", name = "Deathrate due to Drug Overdose in 2016
        label = scales::comma) + theme(legend.position = "right")
```


Deathrate due to Drug Overdose in 2016
50
40
30
20
10