Group exercise 2: Analysing spatial data

DATA5207: Data Analysis in the Social Sciences

Emily and Sarah

Introduction

We will be examining the relationship between opiate drug prescriptions, poverty, education and legality of marijuana and mortality caused by drugs and alcohol using the methods covered in this lab, and previously in this unit. The spatial aspect of this is that we are going to use county-level data sourced from the *US Census Bureau* and the *Centers for Disease Control and Prevention (CDC)*.

We have included an additional variables from county.data:

- Education - Percent of adults with less than a high school diploma 2014-18: We believe this is a confounding factor of poverty and mortality from drugs since education will affect the jobs you are able to obtain (leading to poverty if unobtainable) and educated drug usage (which we are assuming gets taught in high school in the US). If they haven't been to high school, there is a more likely abuse of drugs since they are uneducated on the effects on their drug usage. Similarly, with poverty, if they are uneducated, they are less likely to have the skills to get a job and be able to obtain an income and then are more likely to enter poverty.

Data cleaning

Firstly, we shall load all the data files.

We will join datasets on their similar column - the FIPs code for the county:

Transformation of data

1. Standardisation of Geographical Identifiers: To ensure consistent merging across datasets, geographical identifiers such as county names were standardised. The gsub function was utilised to remove specific terms like 'County', 'Parish', 'City', 'Borough', and 'Census Area'. This step was crucial to prevent mismatches during the data merging process due to different naming conventions across datasets.
2. Calculation of Z-Scores: To facilitate direct comparisons across regions with different scales and units, critical variables such as the percentage of adults without a high school diploma, poverty rates, and opiate prescribing rates were transformed into z-scores. This standardisation adjusted each variable to have a mean of zero and a standard deviation of one, highlighting deviations from the national average.
3. Merging of Datasets: After cleaning and transforming the relevant variables, the datasets were merged based on FIPS codes. This integration created a unified dataset that combined socio-economic and health-related data, enabling a more robust analysis of the factors influencing opioid mortality rates.

Viewing the first six rows of our new data frame, we can see the structure of these data:

```
##    Id2      county mortality.rate below.poverty.rate county.state
## 1 1001 Autauga, AL           12.5                9.4           AL
## 2 1003 Baldwin, AL           22.6                9.3           AL
## 3 1005 Barbour, AL            9.0               20.0           AL
## 4 1007    Bibb, AL           14.1               11.7           AL
## 5 1009  Blount, AL           18.1               12.2           AL
## 6 1011 Bullock, AL           11.1               25.3           AL
##   education.highschool z.highschool     z.poverty  z.opiates
## 1                 11.3   -0.3612033 -0.4937185239  1.2662069
## 2                  9.7   -0.6019626 -0.5113760968  1.1283754
## 3                 27.0    2.0012473  1.3779842063  0.3893133
## 4                 16.8    0.4664068 -0.0875943466  0.4962515
## 5                 19.8    0.9178304  0.0006935181 -0.4614399
## 6                 24.8    1.6702032  2.3138355714 -1.3430863
##   X2016.Prescribing.Rate
## 1                  129.6
## 2                  123.8
## 3                   92.7
## 4                   97.2
## 5                   56.9
## 6                   19.8
```

The first row is the ID variable we used to merge our data frames. The below.poverty.rate is the percentage of each county's families with incomes below the poverty rate. county, the name of each country (and the state within which they sit) while as county.state is the state code. X2016.Prescribing.Rate is the rate of opiod prescriptions. education.highschool is the percentage that completed highschool only. Finally, the mortality rate associated with drug and alcohol use (per 100,000 people).

In this code, we also standardise the education, poverty and prescription rates in the same line of block of code that we use to merge the data.

There is one last edit to make to these data before we examine them. If you run the code

you will see that some of the observations for the variable `mortality.rate` are coded as 'Unreliable'. This creates two issues for us. Those observations are effectively missing for our purposes, and this converts the variable to a string when we want it to be a number.

We then want to link these shapefiles to the county results. This file, which we were required to modify a little above, can be matched to the shapefile data using similar syntax to above:

We edit the variable containing county names in shapefile using the `paste0()` function, which combines the county name with the state abbreviation (from the fip concordance file). We then modify county names with some different spelling in the shapefiles and our county level data:

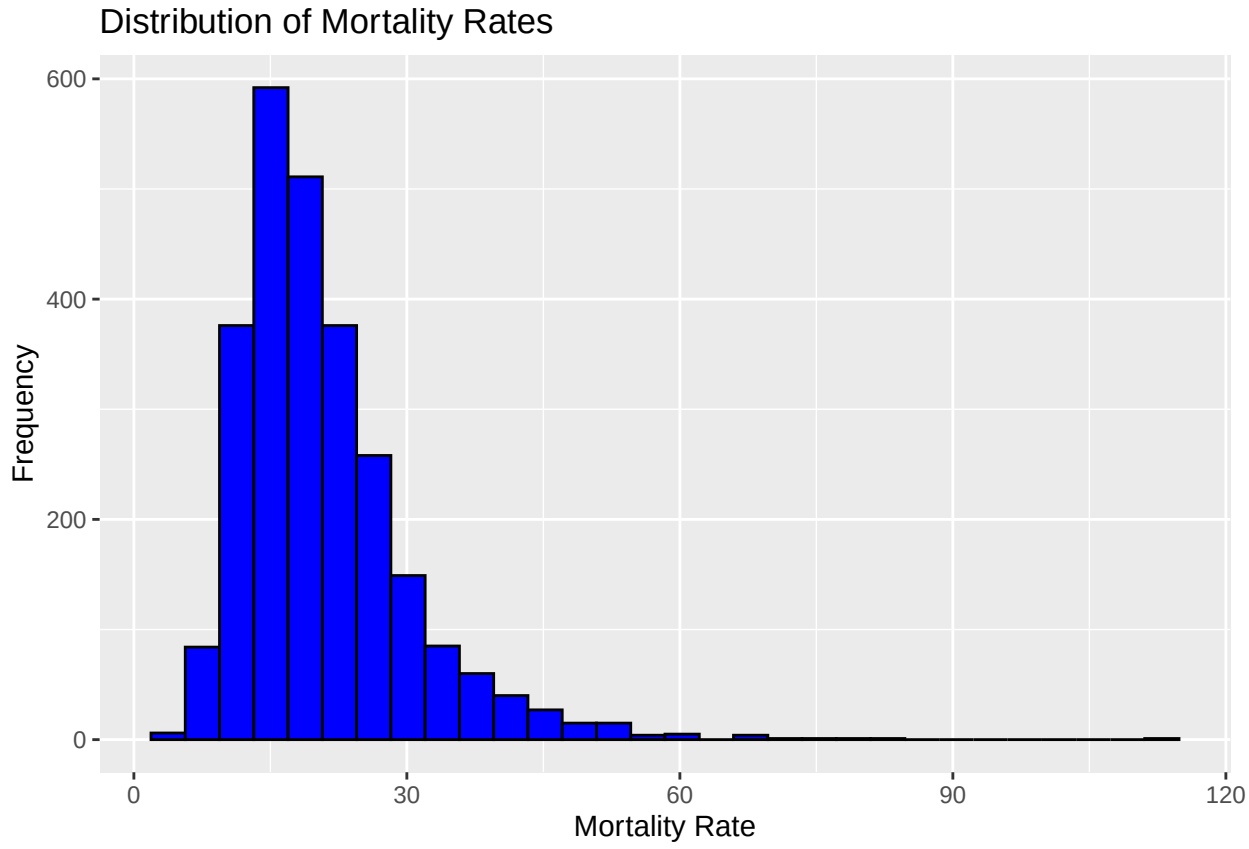
And then we merge the shapefile and the county-level data:

```
## Warning in sf_column %in% names(g): Detected an unexpected many-to-many relationship between `x` and
## i Row 1205 of `x` matches multiple rows in `y`.
## i Row 1138 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

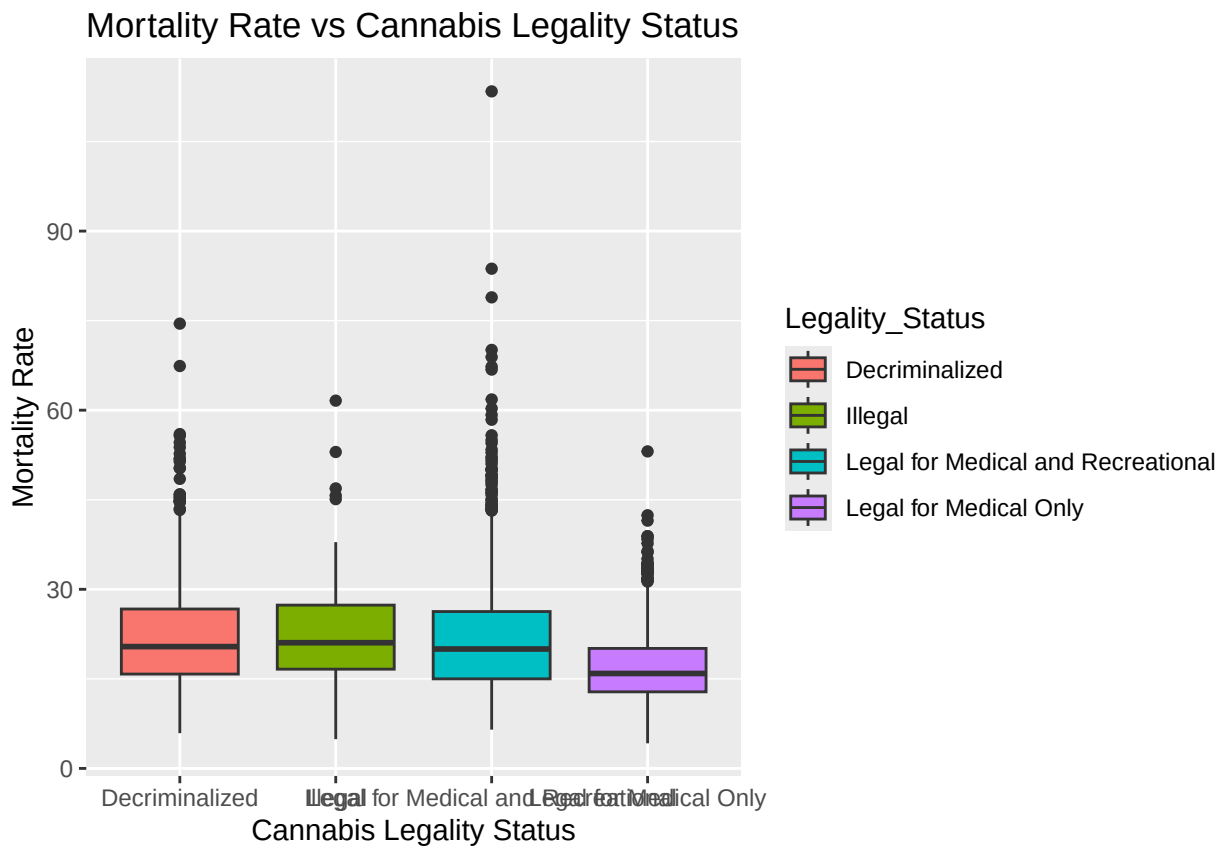
Extra Data set - Drug laws by state

In the context of opioid-related mortality rates, categorising states by marijuana laws into "Legal for Medical Only," "Legal for Medical and Recreational," "Illegal," and "Decriminalised" provides a nuanced framework for analysis. States where marijuana is legally available, either medically or recreationally, may experience a substitution effect, where individuals seeking pain relief opt for marijuana over opioids, potentially reducing opioid misuse and overdoses. Moreover, legalisation can reflect a broader, more progressive attitude towards healthcare and substance use treatment, increasing access to various addiction services. In states where marijuana remains illegal, strict law enforcement may perpetuate stigma around substance use, discouraging individuals from seeking treatment. Conversely, states that have decriminalised marijuana might allocate fewer resources to punitive measures and more towards harm reduction strategies, which could indirectly influence opioid mortality rates. This categorical analysis can reveal correlations between marijuana policies and opioid fatalities, offering crucial insights for future legislative decisions.

Descriptive Analysis

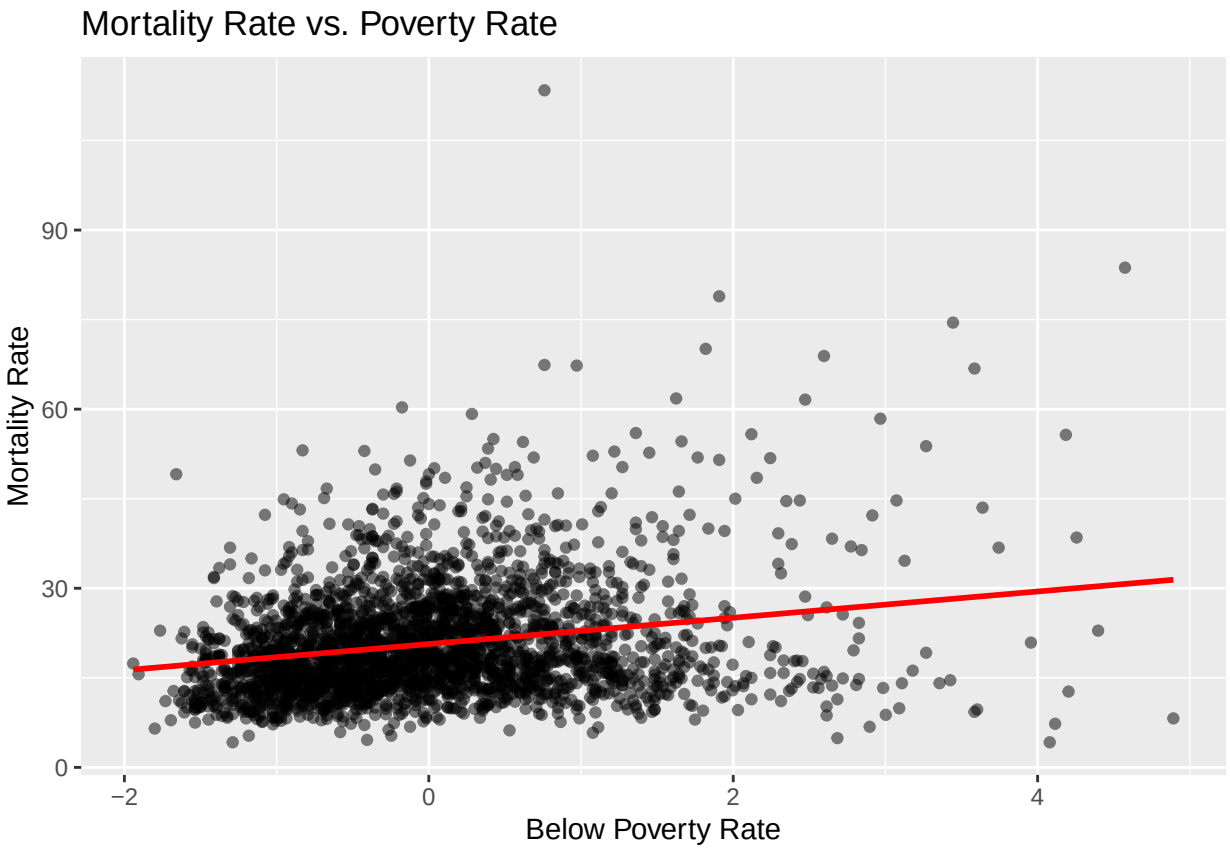


This histogram illustrates the distribution of opioid-related mortality rates, predominantly clustered between 0 and 30 deaths per 100,000 population. The data is positively skewed towards lower rates, peaking between 10-20 deaths, indicating this is the most common range. The frequency of occurrences drops sharply for higher rates, with a long tail extending up to 120 deaths, highlighting a few regions with significantly higher mortality rates. Overall, the distribution shows that high opioid mortality rates are relatively rare, with most areas experiencing lower rates.



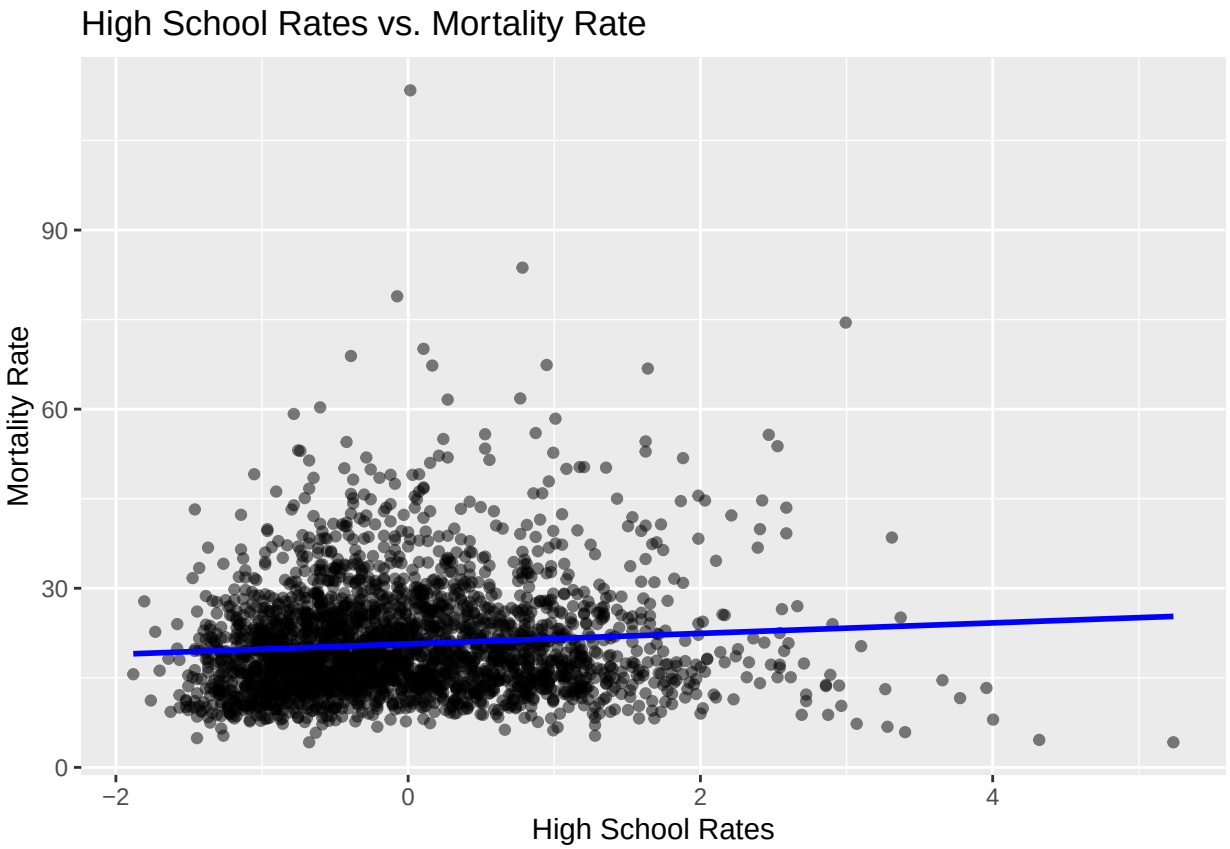
Decriminalised (Red): Shows lower and more consistent mortality rates with few outliers. Illegal (Green): Features a wider spread and higher median, indicating generally higher rates. Legal for Medical and Recreational (Blue): Displays a broad range but a lower median than illegal states, suggesting a moderating effect of more liberal cannabis policies. Legal for Medical Only (Purple): Similar range to the recreational category but with a slightly lower median, indicating moderately lower mortality rates. The plot highlights significant variability within each category, particularly in regions with higher rates, demonstrating the potential impact of cannabis legality on opioid mortality. Overall, this visualisation underscores the potential link between cannabis legality and opioid mortality, suggesting that where cannabis is more accessible, it might serve as a viable alternative to opioids, potentially leading to lower opioid death rates.

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



The red trend line shows a positive correlation, indicating that as poverty rates increase, there is a general rise in opioid mortality rates. This suggests that higher poverty levels are associated with increased opioid-related deaths. However, the spread of mortality rates at lower poverty rates indicates that poverty is a significant but not the only factor influencing opioid mortality, highlighting the complex interplay between socioeconomic factors and health outcomes.

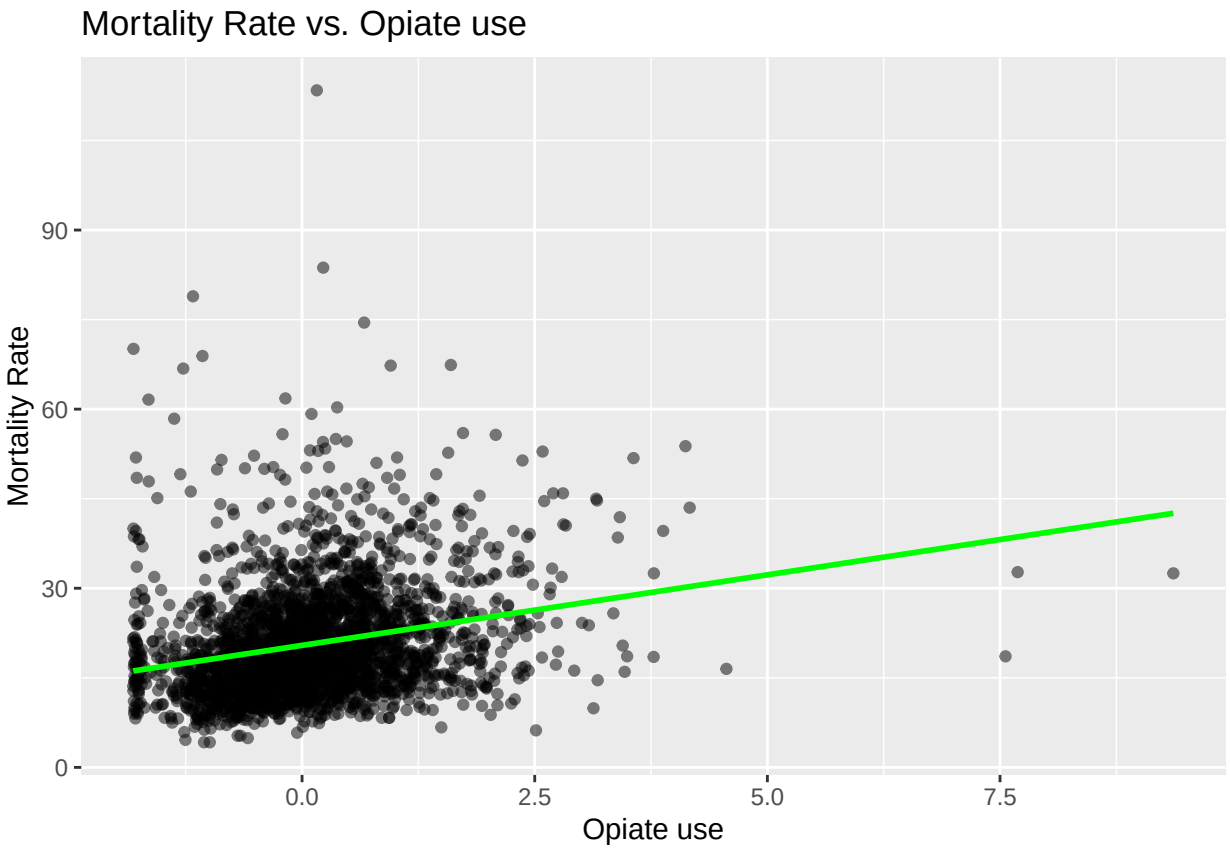
```
## `geom_smooth()` using formula = 'y ~ x'
```



The trend line suggests a weak positive correlation, indicating that regions with higher percentages of adults lacking a high school diploma tend to have slightly higher opioid mortality rates. However, the data points are widely dispersed, suggesting that while educational attainment may influence opioid mortality rates, it is one of several factors contributing to these outcomes.

The clustering of many data points at lower percentages of adults without high school diplomas with a broad range of mortality rates further highlights the complexity of this relationship. This visualisation suggests that while lower educational attainment may be associated with higher opioid mortality, the impact is not uniformly strong across all regions.

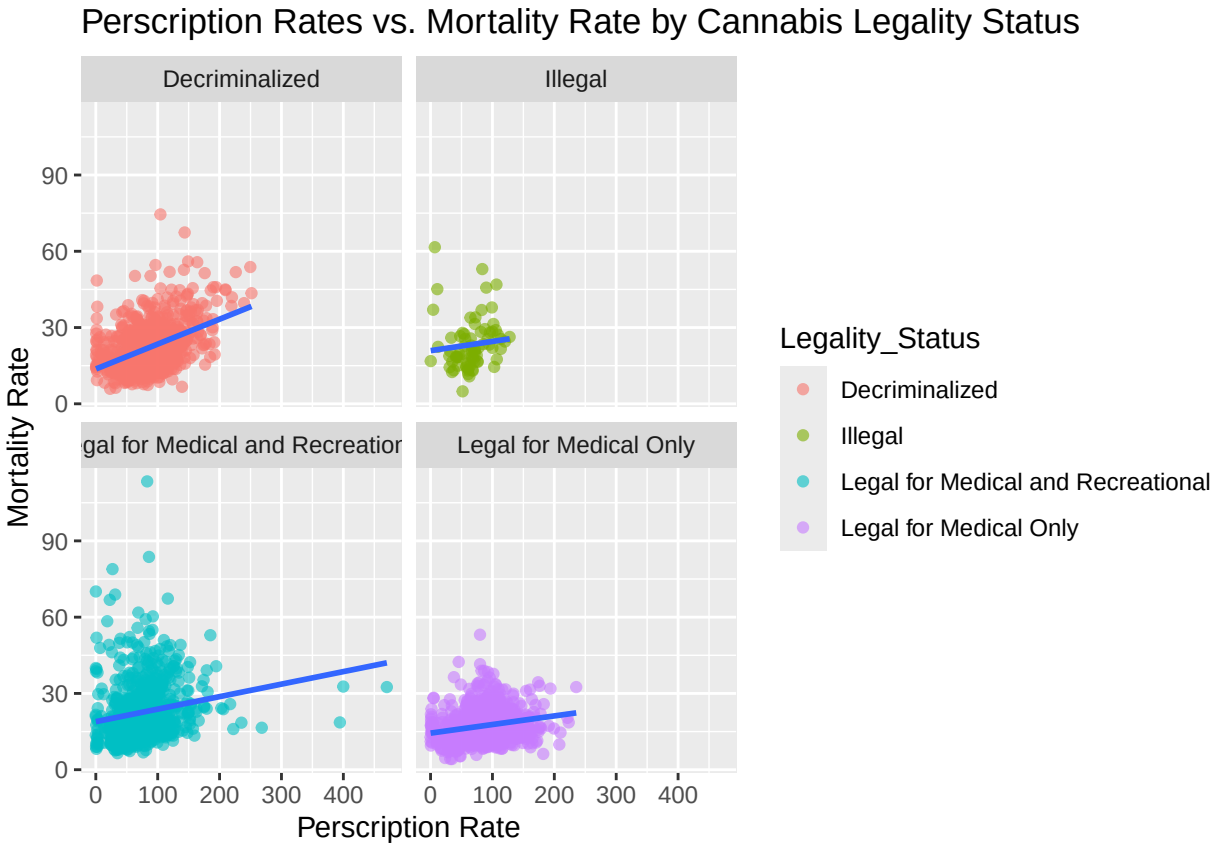
```
## `geom_smooth()` using formula = 'y ~ x'
```



The green trend line indicates a clear positive correlation, suggesting that as opiate use increases, there is a corresponding rise in opioid mortality rates. The dense clustering of data points around lower opiate use rates with varied mortality outcomes highlights that while higher opiate use is associated with higher mortality, there is also significant variability in mortality rates at lower levels of opiate use.

This plot demonstrates that higher opiate use is unsurprisingly strong predictor of opioid mortality, underscoring the need for targeted interventions in regions with higher opiate consumption to potentially mitigate the associated risks.

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



Decriminalised (Red): Shows a strong positive correlation, with higher prescription rates correlating significantly with higher mortality rates.

Illegal (Green): Displays fewer data points, with a moderate positive correlation indicating a slight increase in mortality rates as prescription rates rise.

Legal for Medical and Recreational (Cyan): Exhibits a wide spread of prescription rates with a very gentle upward trend in mortality rates, suggesting a minimal impact of increased prescriptions on mortality.

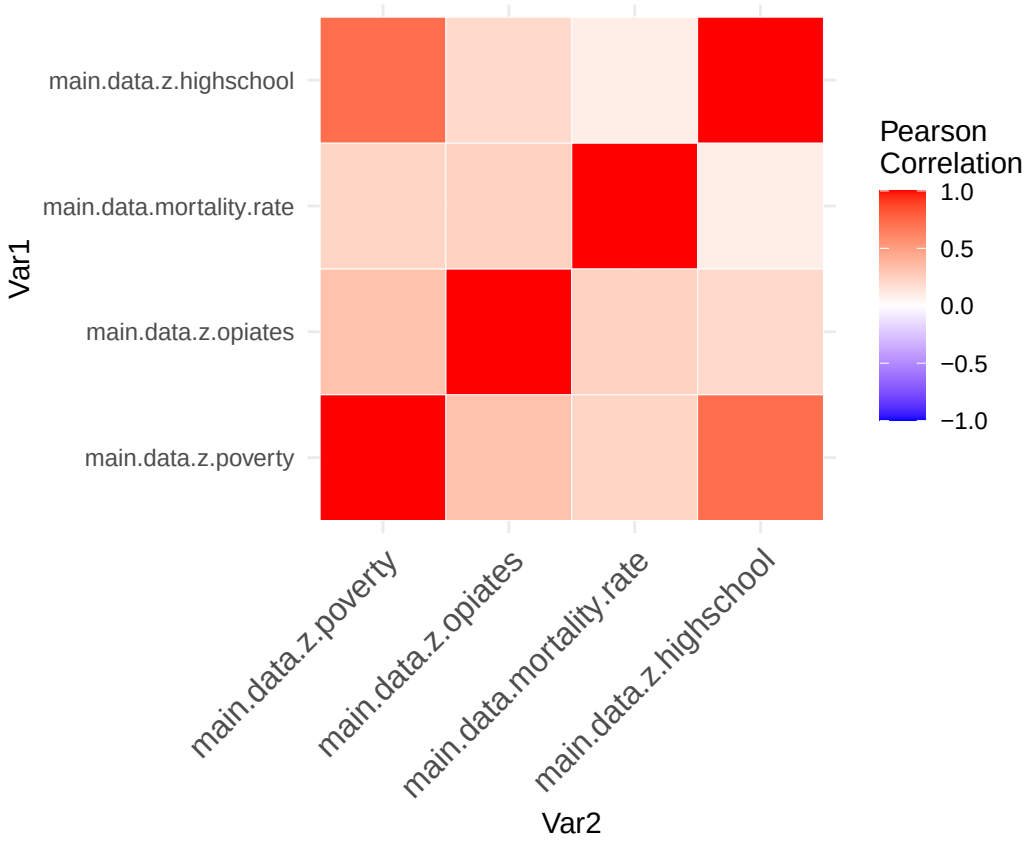
Legal for Medical Only (Purple): Features a low to moderate positive correlation, with a gradual increase in mortality rates as prescription rates increase.

The plot highlights varying relationships between opioid prescriptions and mortality rates, influenced by the cannabis legality in different regions. More liberal cannabis laws appear to be associated with less pronounced increases in opioid mortality rates against rising prescription rates, suggesting potential differences in prescribing practices.

Correlation heatmap

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```



Relationship of dependent and independent variables

To explore the question 'How does poverty and opiate prescriptions predict mortality rates associated with drugs and alcohol?' we have explored potential variables including poverty rates, opiate prescription rates, education levels and state laws for drug use. With a numerical dependent variable, mortality rate, a linear regression type may be suitable to model the relationship.

A few assumptions need to be made in order to use this model:

Linearity: The relationship between the predictor variables (poverty, opioid prescriptions) and the response variable (mortality rates associated with drugs and alcohol) should be approximately linear. This assumption is assessed through our descriptive analysis scatter plots.

Independence: The observations should be independent of each other. In other words, the mortality rates for one county should not influence the mortality rates for another county.

Homoscedasticity: The variance of the residuals should be constant across all levels of the predictors. This can be checked by examining a plot of residuals versus fitted values.

Normality of Residuals: The residuals should follow a normal distribution. This assumption can be assessed by examining a histogram or a Q-Q plot of the residuals.

No Multicollinearity: The predictor variables should not be highly correlated with each other. Multicollinearity can inflate standard errors and make interpretation of coefficients difficult. This is seen in the correlation matrix and it can be assumed this assumption is met.

Make sure to drop the NA values for all the other columns:

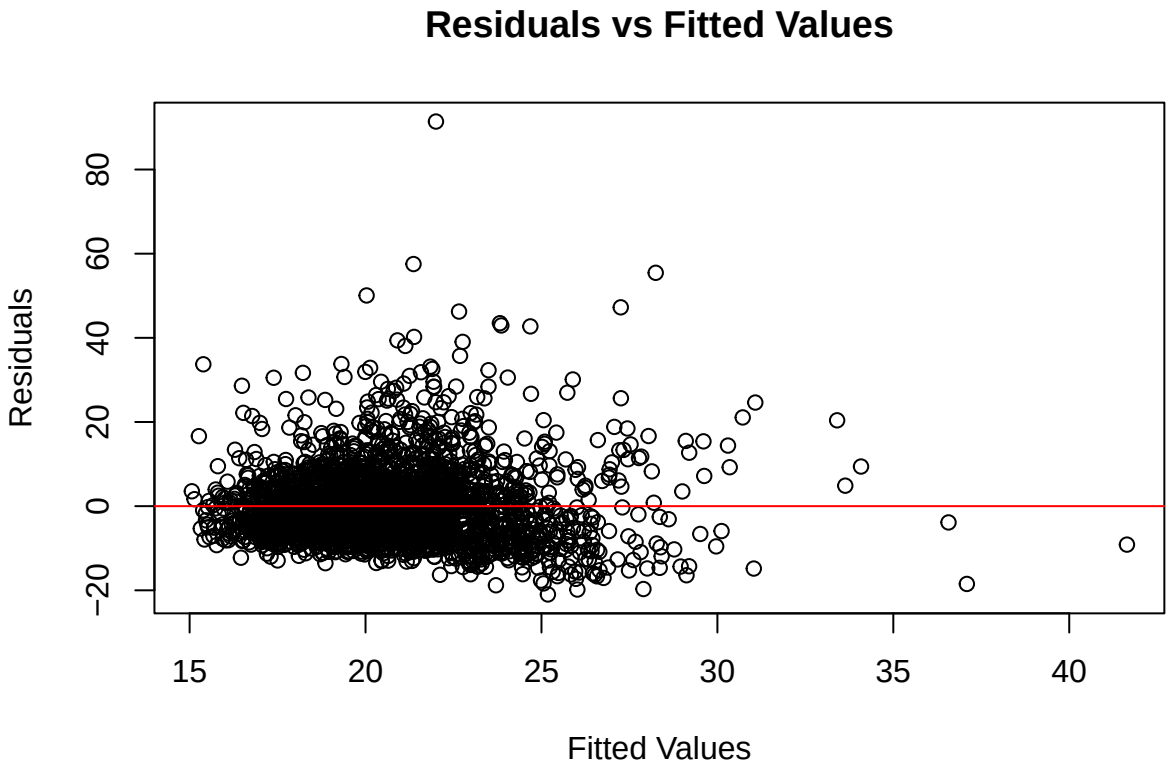
Model including just the standardised poverty and opiate variables.

```
##
## Call:
## lm(formula = mortality.rate ~ z.poverty + z.opiates, data = main.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.983  -5.824  -1.562   3.963  91.399
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.4864     0.1757 116.591   <2e-16 ***
## z.poverty     1.6040     0.1941   8.265   <2e-16 ***
## z.opiates     1.8630     0.1916   9.724   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.934 on 2609 degrees of freedom
## Multiple R-squared:  0.08339,    Adjusted R-squared:  0.08269
## F-statistic: 118.7 on 2 and 2609 DF,  p-value: < 2.2e-16
```

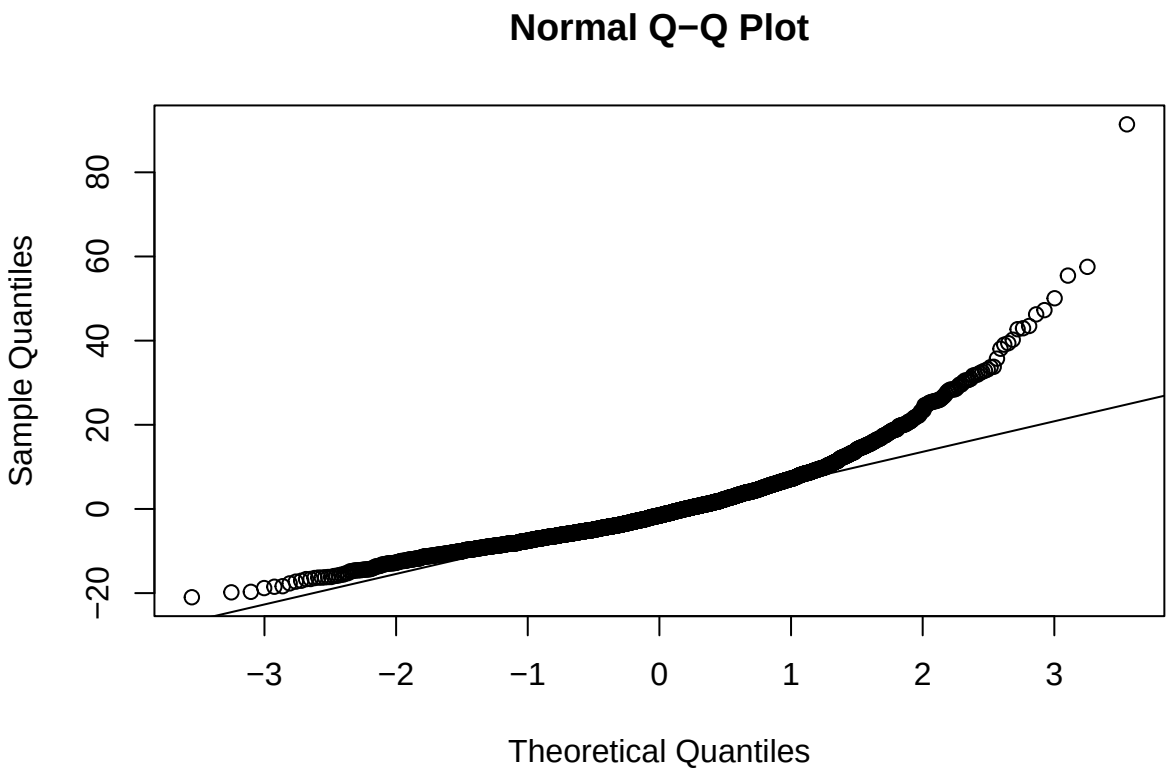
Model includes the standardised numerical variables and state law types.

```
##
## Call:
## lm(formula = mortality.rate ~ z.poverty + z.opiates + z.highschool +
##     Legality_Status, data = legalplus.main.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.350  -5.559  -1.214   3.978  87.919
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                       21.6153     0.2961  72.995
## z.poverty                                          2.8244     0.2602  10.856
## z.opiates                                          1.8133     0.1824   9.944
## z.highschool                                      -0.9000     0.2680  -3.359
## Legality_StatusIllegal                             2.8086     1.0328   2.719
## Legality_StatusLegal for Medical and Recreational  1.4446     0.4200   3.439
## Legality_StatusLegal for Medical Only             -5.1230     0.4128 -12.412
##                                                  Pr(>|t|)
## (Intercept)                                       < 2e-16 ***
## z.poverty                                         < 2e-16 ***
## z.opiates                                         < 2e-16 ***
## z.highschool                                     0.000795 ***
## Legality_StatusIllegal                           0.006583 **
## Legality_StatusLegal for Medical and Recreational 0.000593 ***
## Legality_StatusLegal for Medical Only             < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.439 on 2605 degrees of freedom
## Multiple R-squared:  0.1835, Adjusted R-squared:  0.1816
## F-statistic: 97.56 on 6 and 2605 DF,  p-value: < 2.2e-16
```

Testing Homoscedasticity: There is no real pattern in the plot which is good for homoscedasticity. The spread is quite random.



Testing for Normality of Residuals: the plotted points don't form a straight line and indicate that the residuals deviate slightly and the tail ends suggest a skewness.



Interpretations

We have made two models to identify if the inclusion of confounding factors better explains the mortality rate variable.

Model 1 identifies that the linear regression model can be suitable for predicting mortality rates with the assumptions holding. The coefficients in this model are all statistically significant. The intercept coefficient is the estimated value of the response variable when all predictor variables are set to zero. In this case, it represents the mortality rate is 20.49 (2dp) when both poverty rates and opiod prescriptions are 0.

Furthermore, the poverty coefficient and opiate coefficient are 1.6 and 1.9 respectively. Since the coefficients were standardised, the interpretation of this is one standard deviation increase in poverty is associated with an estimated increase of 1.6 standard deviation units in mortality rate, holding other variables constant. Similarly, one standard deivation increase in opiod prescription rates is associated with an estimated increase of 1.9 standard deviation units in mortality rate, holding other variables constant.

This positive relationship could be explained by our hypothesis on the social and economical aspects to mortality rate from drug and alcohol. Increases in poverty could relate to limited access to healthcare resources including mental health services and substance abuse treatment programs, increasing risk-taking behaviours associated with drugs and alcohol, increasing mortality rate. Furthermore, people with higher poverty rates could experience higher stress and social isolation due to financial instability, inadequate housing and could contribute to self-medication through substance use, leading to possibilities of addiction and again, higher mortality rates. With an increase in opiate prescriptions, there is an increase of availability and accessibility of opiates which can increase the risk of misuse, addiction and overdose deaths. The increase in prescriptions could also reflect counties that have less regulated prescribing practices.

Model 2 includes high school attainment and the laws associated with weed access to further explain the

relationship of mortality rates from drug and alcohol. Interestingly, all the variables are statistically significant except for legal status of 'legal for medical use only'. The positive relationships between mortality rate and poverty and opiate prescriptions hold true in this model as well. The educational status demonstrates a negative relationship with an increase of one standard deviation associated with high school rates causing approximately a 1.1 standard deviation decrease in mortality rates. Since the variable is associated with the rate of individuals in a county that have attained less than high school education, this relationship explains the social aspect of this problem with individuals that are more educated, making better decisions with drugs and alcohol leading to decreases in mortality rates.

The baseline of the legal variable is 'decriminalised' and the coefficients represent the estimated difference between the baseline and the categories in the predictor variable. Specifically, areas where weed is illegal is estimated to have a 7.4 unit higher mortality rate compared to where weed is decriminalised. Furthermore, where weed is legal for medical and recreational use leads to a 5.9 unit higher mortality rate in comparison. This relationship highlights political and social factors related to mortality rates associated with drugs and alcohol. Even though this only explains the access individuals have to one particular drug (marijuana), it can explain the stigma and criminalisation of drug use in general as well. The biggest increase in mortality rates were associated with legal status of 'illegal' there is two possible suggestions that can be made from this.

One being there may be a lack of access to proper state approved drugs. The illegality of marijuana can create economic incentives for illegal drug markets, these economic factors can contribute to higher mortality rates through factors such as drug-related violence, accidental overdose, and lack of quality control in illicit drug production.

A second response may be the fact that those who do not have access to medical marijuana would instead be prescribed and opiod. Therefore, increase the probability of misuse. Whilst states that allow recreational or medical marijuana way find pain relief through this avenue rather than opiod prescriptions.

Model Fit

Both models had particularly low adjusted r-squared values of 0.08 and 0.13 respectively. The second model was able to explain a greater proportion of variability in the mortality rate associated with drugs and alcohol, however only 13% of this variability is explained which could be from including more relevant predictors.

Map

Plotting the spatial patterns of poverty rates, prescription rates and mortality rates doesn't show distinct patterns of correlation between the two variables and the dependent variable. The map might show a high mortality rate in a particular county, but within that county, there may be significant variation in individual-level factors such as socioeconomic status, access to healthcare, and lifestyle behaviors. We also can see there is particular missing data for Alaska counties in the prescription map and the central counties mortality rates, leading to missing data bias.

We also need to remember ecological fallacy and that inferences about individuals because of a given demographic group may not represent how those individuals may behave. In this example, assuming that individuals who live in areas with high poverty rates, high mortality rates, or high opiate prescriptions also have similar characteristics can lead to incorrect conclusions.

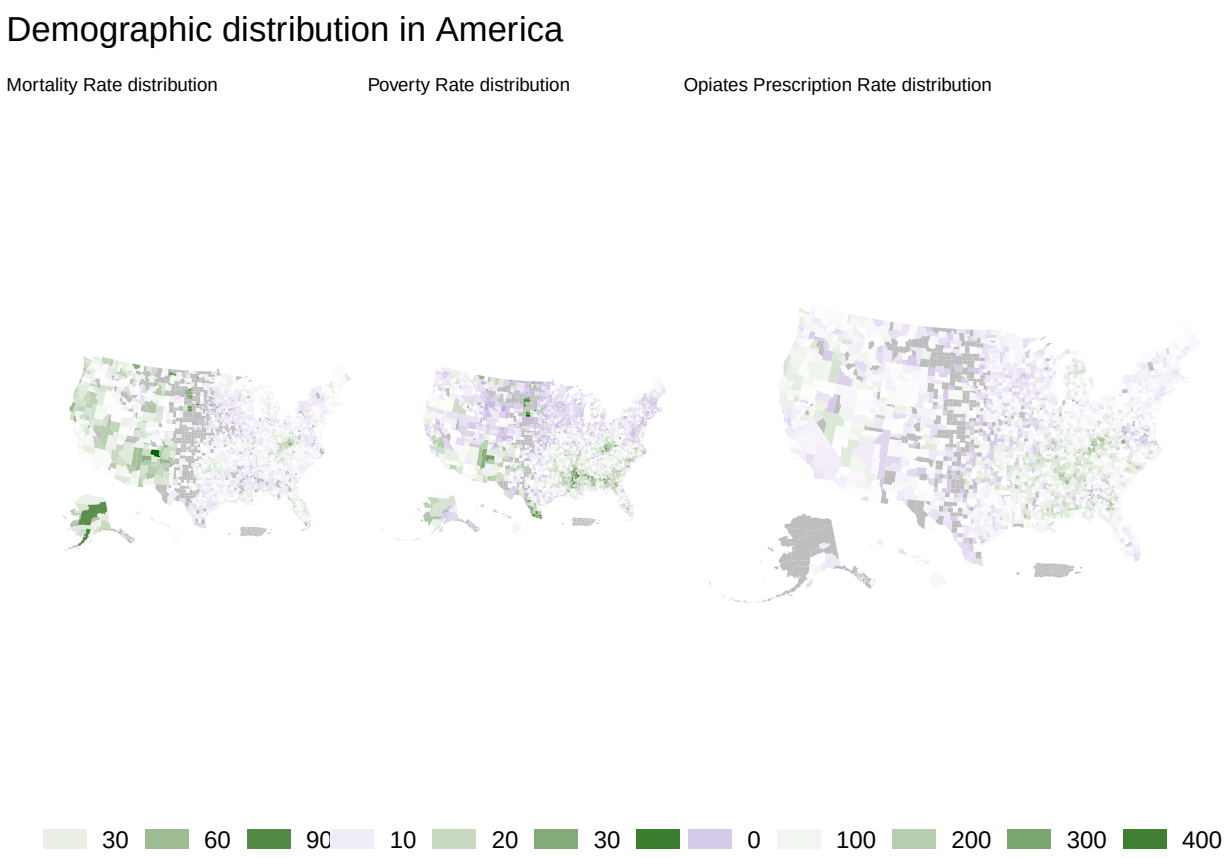


Figure 1: Demographic distribution in America