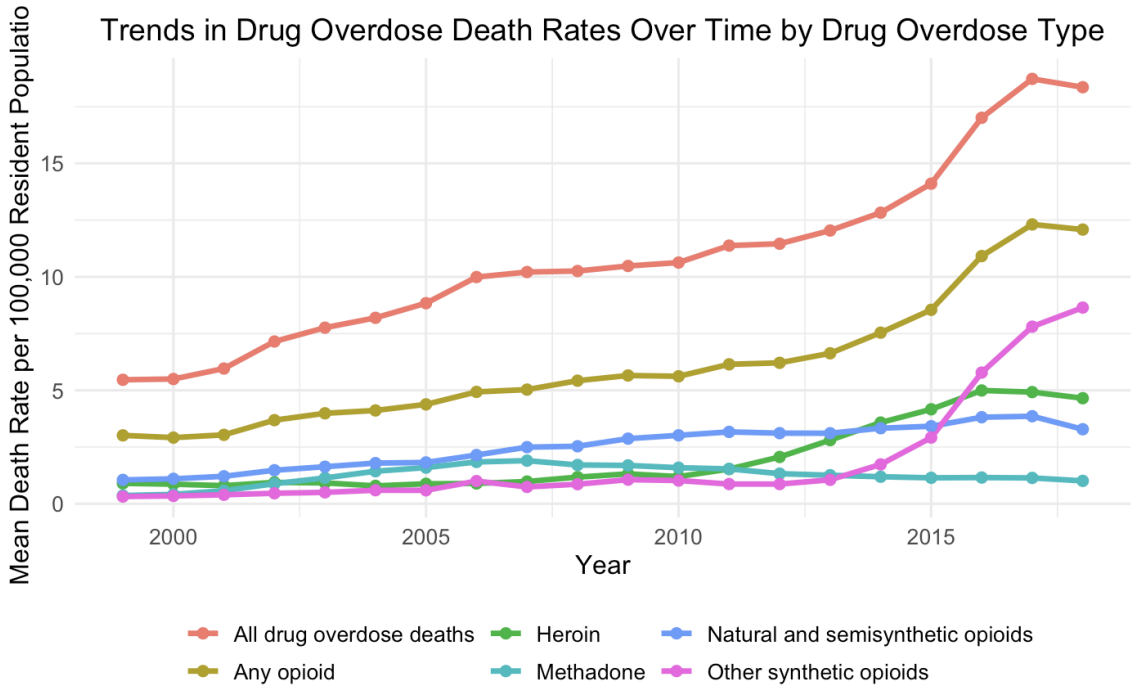
Intro

At the heart of this crisis lies the specter of opioids, which have emerged as the primary culprit behind the rising tide of overdose deaths. In 2019, opioids accounted for a staggering 70.6% of all drug overdose deaths—a harrowing statistic that underscores the urgent need for action.

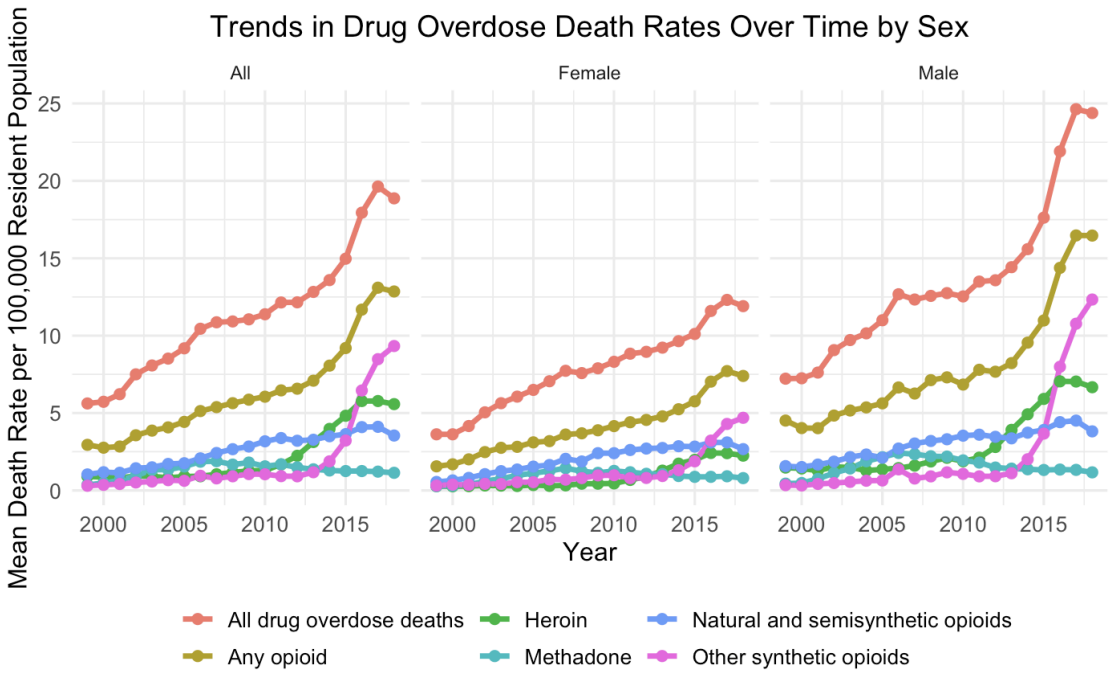
Data Visualization

1. What are the trends in drug overdose death rates over time?

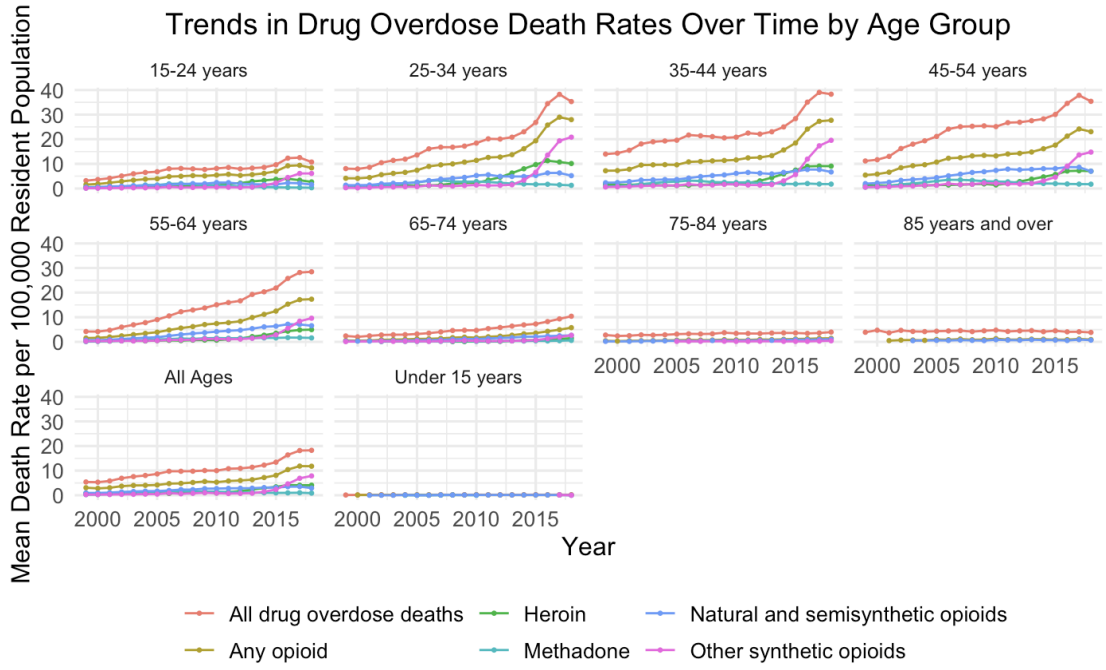
- Drug overdose type



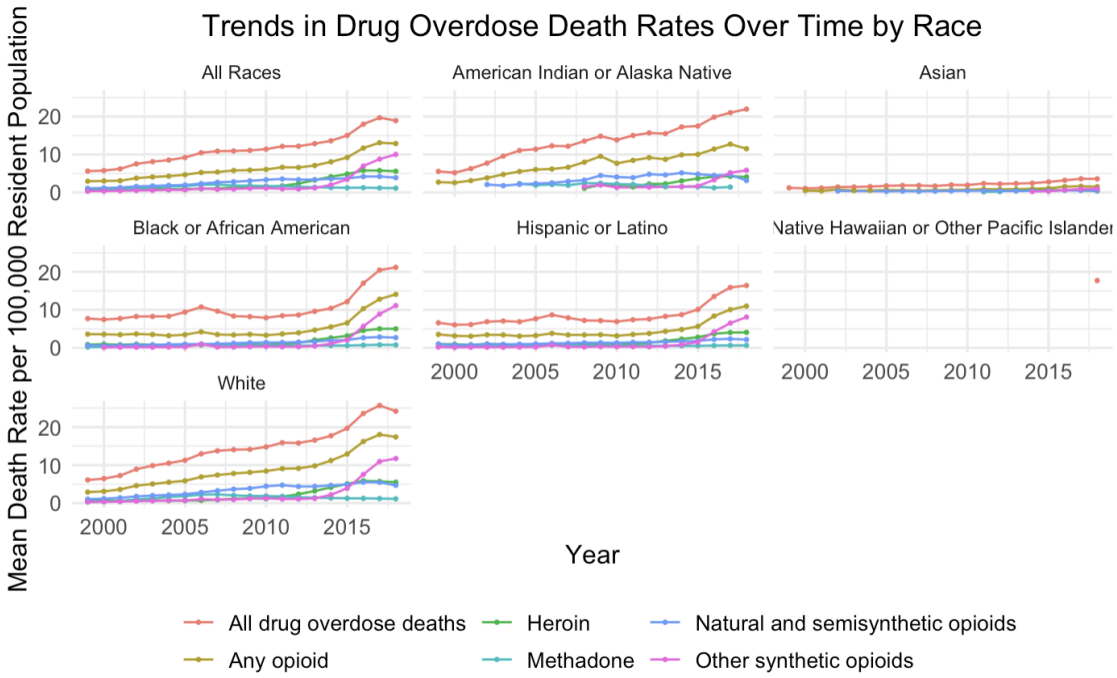
- Sex


Trends in Drug Overdose Death Rates Over Time by Sex
All
Female
Male
Mean Death Rate per 100,000 Resident Population
0
5
10
15
20
25
2000
2005
2010
2015
Year
All drug overdose deaths
Heroin
Natural and semisynthetic opioids
Any opioid
Methadone
Other synthetic opioids

- Age


Trends in Drug Overdose Death Rates Over Time by Age Group
15-24 years
25-34 years
35-44 years
45-54 years
55-64 years
65-74 years
75-84 years
85 years and over
All Ages
Under 15 years
Mean Death Rate per 100,000 Resident Population
0
10
20
30
40
2000
2005
2010
2015
Year
All drug overdose deaths
Heroin
Natural and semisynthetic opioids
Any opioid
Methadone
Other synthetic opioids

- Race



2、Are there any specific demographic groups (gender/sex/age) that are most affected by specific types of drug overdoses?

Conclusion

- Synthetic opioids (other than methadone) are the primary driver of the increase in drug overdose deaths across all demographics.
- Males are more affected by drug overdoses compared to females.
- Younger age groups (25-44 years) show the most significant increase in drug overdose death rates.
- Whites and American Indians or Alaska Natives are among the most affected racial groups.

3、Which factors are most strongly associated with higher death rates?

```
Call:
lm(formula = log_estimate ~ YEAR + PANEL_COMBINED + Sex + Age +
    Race, data = drug)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76711 -0.29097  0.00154  0.25708  2.28857

Coefficients:
                                                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                    -1.728e+02  2.424e+00  -71.274  < 2e-16 ***
YEAR                                            8.705e-02  1.207e-03   72.115  < 2e-16 ***
PANEL_COMBINEDAny opioid                       -7.834e-01  2.185e-02  -35.844  < 2e-16 ***
PANEL_COMBINEDHeroin                           -2.274e+00  2.465e-02  -92.266  < 2e-16 ***
PANEL_COMBINEDMethadone                        -2.525e+00  2.451e-02 -103.036  < 2e-16 ***
PANEL_COMBINEDNatural and semisynthetic opioids -1.614e+00  2.245e-02  -71.890  < 2e-16 ***
PANEL_COMBINEDOther synthetic opioids          -2.481e+00  2.425e-02 -102.333  < 2e-16 ***
SexFemale                                      -3.631e-01  1.978e-02  -18.355  < 2e-16 ***
SexMale                                         2.653e-01  1.979e-02   13.403  < 2e-16 ***
Age25-34 years                                  8.613e-01  3.682e-02   23.394  < 2e-16 ***
Age35-44 years                                  1.090e+00  3.682e-02   29.615  < 2e-16 ***
Age45-54 years                                  1.103e+00  3.682e-02   29.957  < 2e-16 ***
Age55-64 years                                  4.055e-01  3.708e-02   10.936  < 2e-16 ***
Age65-74 years                                 -6.819e-01  3.908e-02  -17.449  < 2e-16 ***
Age75-84 years                                 -1.203e+00  4.523e-02  -26.598  < 2e-16 ***
Age85 years and over                           -9.876e-01  5.083e-02  -19.429  < 2e-16 ***
AgeAll Ages                                     4.425e-01  3.192e-02   13.863  < 2e-16 ***
AgeUnder 15 years                              -3.550e+00  4.933e-02  -71.961  < 2e-16 ***
RaceAmerican Indian or Alaska Native            1.894e-01  3.263e-02    5.806 6.81e-09 ***
RaceAsian                                      -1.965e+00  3.729e-02  -52.697  < 2e-16 ***
RaceBlack or African American                  -4.334e-01  3.670e-02  -11.809  < 2e-16 ***
RaceHispanic or Latino                         -5.619e-01  2.938e-02  -19.126  < 2e-16 ***
RaceNative Hawaiian or Other Pacific Islander  -7.490e-01  4.930e-01   -1.519    0.129
RaceWhite                                       1.635e-01  2.923e-02    5.594 2.34e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4922 on 5081 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8684
F-statistic:  1466 on 23 and 5081 DF,  p-value: < 2.2e-16
```

- YEAR has a positive and significant association with higher death rates.
- PANEL_COMBINED categories (Any opioid, Heroin, Methadone, Natural and semisynthetic opioids, and Other synthetic opioids) have negative and significant associations with death rates.
- Sex: Being male is associated with higher death rates compared to being female.
- Age: Age groups 25-34, 35-44, 45-54, 55-64 have higher death rates, while age groups 65-74, 75-84, and 85+ have lower death rates compared to the reference age group.
- Race: American Indian or Alaska Native and White are associated with higher death rates compared to the reference race group, while Asian, Black or African American, and Hispanic or Latino are associated with lower death rates.

The factors most strongly associated with higher death rates, based on the magnitude and significance of the coefficients, are being in the age groups 35-44 years, 45-54 years, and 25-34 years, as well as being male.

4. Investigate whether certain combinations of demographics (e.g., young males, and elderly females) have particularly high or low death rates for specific drug overdose types.

Summary:

- Elderly Females: Higher death rates for older females in general.
- Elderly Males: Lower death rates for certain drug types, especially heroin.
- Young Males: Generally higher death rates, but not specifically tied to any drug type in this analysis.
- Specific Drug Types: Heroin shows a consistent trend of lower death rates across various age groups, particularly in older adults.

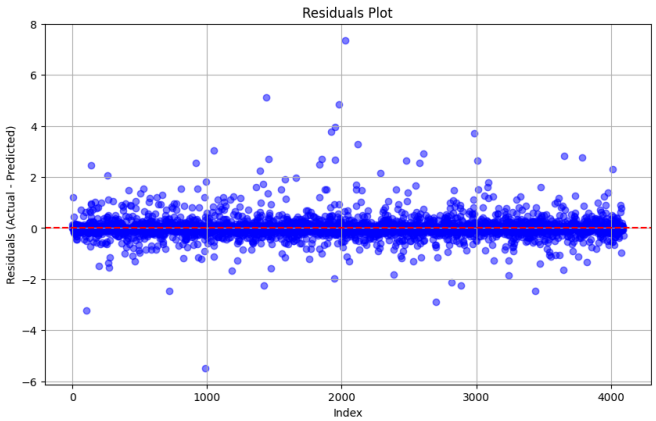
These findings indicate that demographic factors like age and sex significantly influence death rates from specific drug overdose types, with notable differences observed in interactions between drug types and demographics.

Machine Learning

Random Forest Regressor

Link to notebook:
https://colab.research.google.com/drive/1UejM94enGAzbYB71acZSMWiEIKYsWs5R?usp=sharing

- Data cleaning/processing
  - After reading the documentation, I decided that 'flag' is not very helpful without the additional notes, which requires word for word analysis, so I will just drop it for now.
  - And for 'estimate', I decided to make the rows with NaN values in this column act as representation of future data.
  - I will then train and test with the rest of the rows.
  - 'INDICATOR' has all the same value, so there is no meaning to predict with this column, it will be dropped.
  - As for 'PANEL', looking at the data values, it seems informative. But I believe there shouldn't be a rank in the values, so I will be dropping this column and 'PANEL_NUM', and use one-hot encoding.
  - For the rest of the columns, they all provide some information for predicting 'estimate'. Similarly, I will use some one-hot encoding since they are all not ordinal.
  - I will perform further feature selection if necessary.
- Modeling
  - The first model I will try is random forest, and I will split the data with value in 'estimate' into train and test data, so I can use the test data to evaluate future performance.
  - From the residual plot, it seems that the model produced good predictions. The residuals are scattered randomly, and close towards the horizontal line y = 0, which means the errors are mostly small and close to 0



Let's check the MSE and R^2

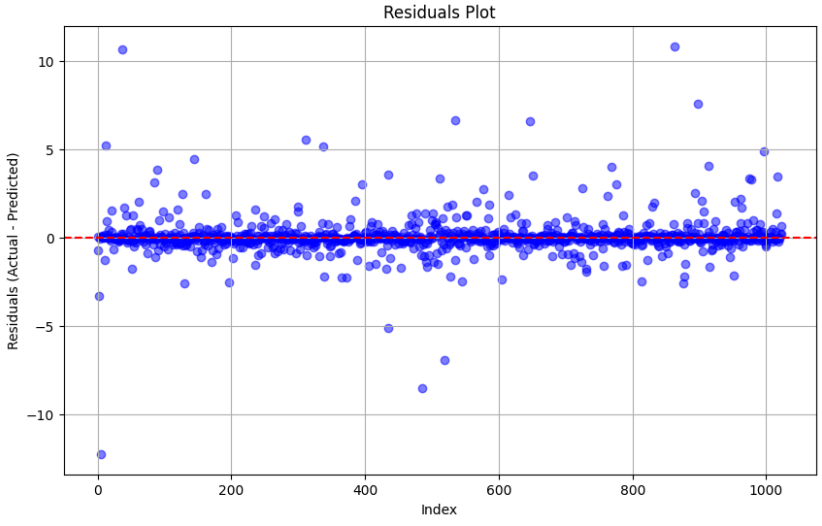
```
[22] mean_squared_error(pred_train_rf, y_train, squared=False)
```

```
0.42389758056365495
```

```
[23] r2_score(y_train, pred_train_rf)
```

```
0.9955779447755507
```

- The MSE is very close to 0, and the $R2$ very close to perfect. Since this is the training data, there are chances that the model overfitted. However, since it is a random forest model, it probably didn't overfit too much. We can check with our test data.



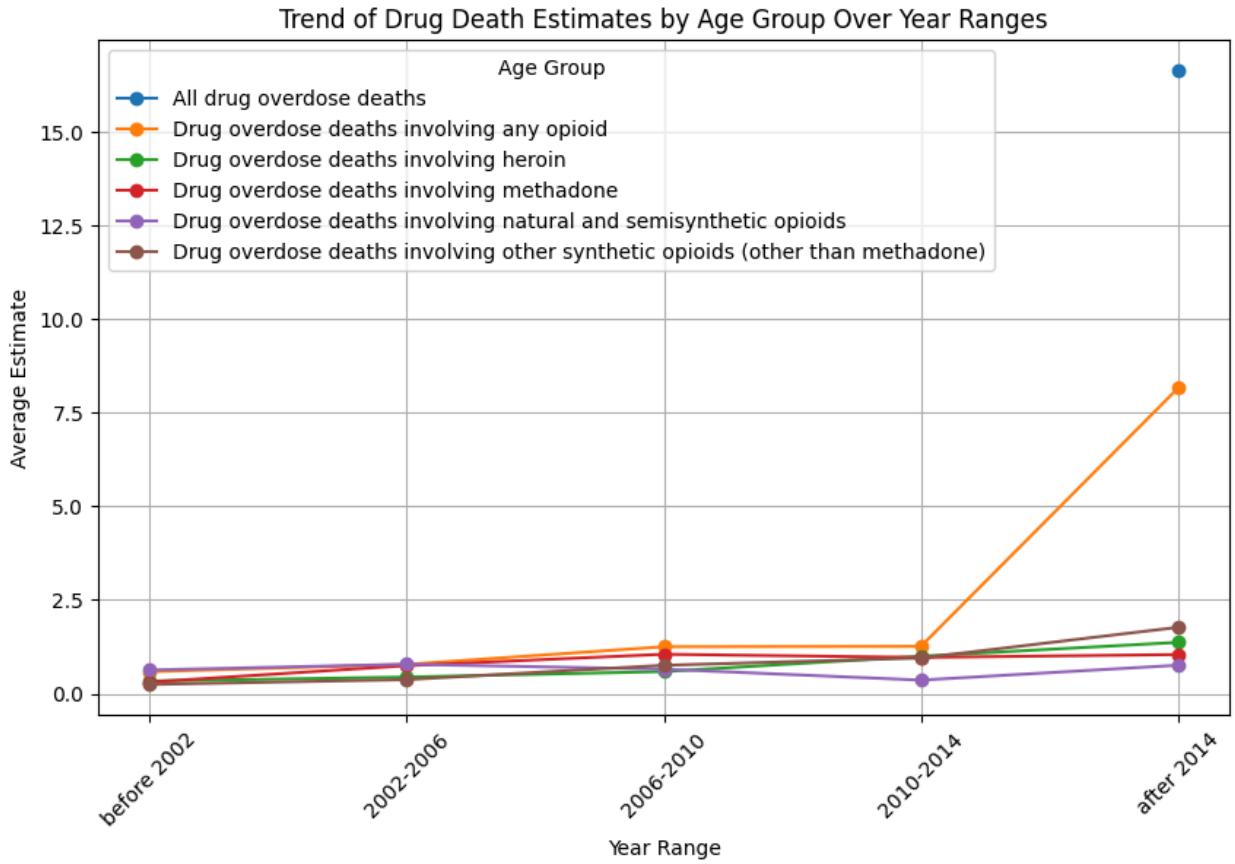

```
[26] mean_squared_error(pred_test_rf, y_test, squared=False)
```

```
1.1380673057349515
```

```
[27] r2_score(y_test, pred_test_rf)
```

```
0.9704193067859317
```

- Grid search
    - I then performed some grid search for the best hyperparameters. However, even with the best hyperparameters, the model performance did not change much. We might as well just stick with the default model.
- Estimate for missing data



    - I think the 'All drug overdose deaths' didn't have too many missing data, so it only have one point for 'after 2014', but we can interpret it as the sum of the rest of the categories.
    - So from the plotS, we can see the estimated deaths follow a similar trend as the trend given by the rest of the dataset. This gives some evidence that the model is producing valid estimates.
    - Besides all drug overdose, drug overdose involving any opioid continues to be the most influential death cause.