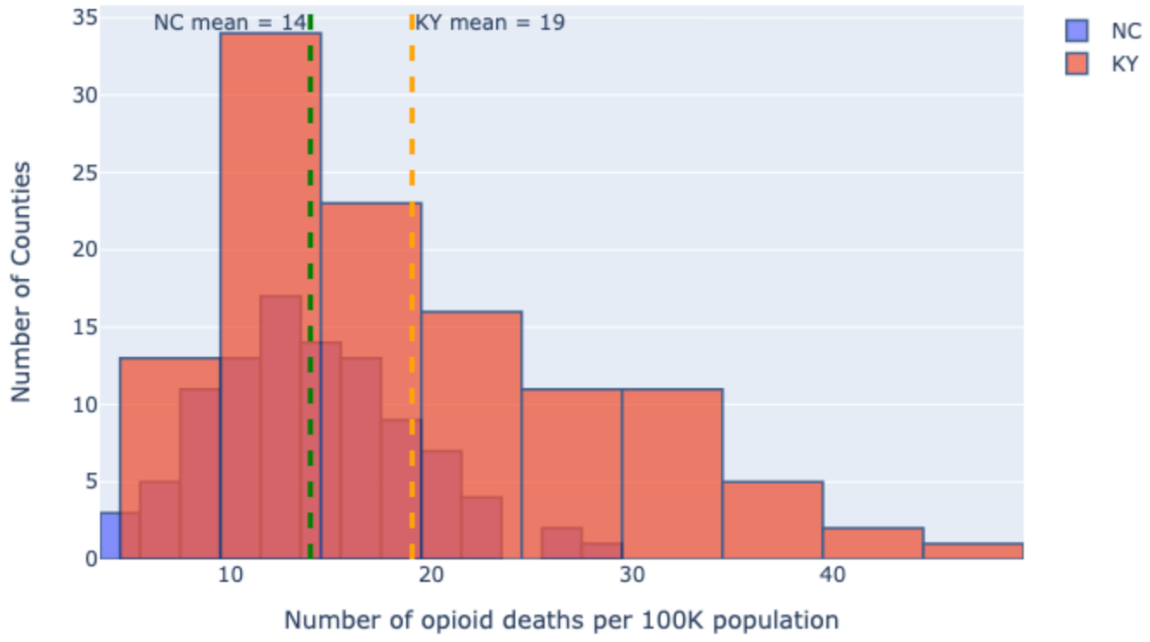
Stage 3

Task 1: Distribution Analysis

M1.1 Compare NC and KY on Opioid Mortality - 2019 Data. Create histograms for NC and KY for Opioid Mortality (Normalized Mortality Rate), Merge them into a single graph, Plot mean lines for both the histograms.

We have created data frames for NC and KY to plot the histograms.



M1.2 Evaluate a distribution for the Normalized Mortality Rate

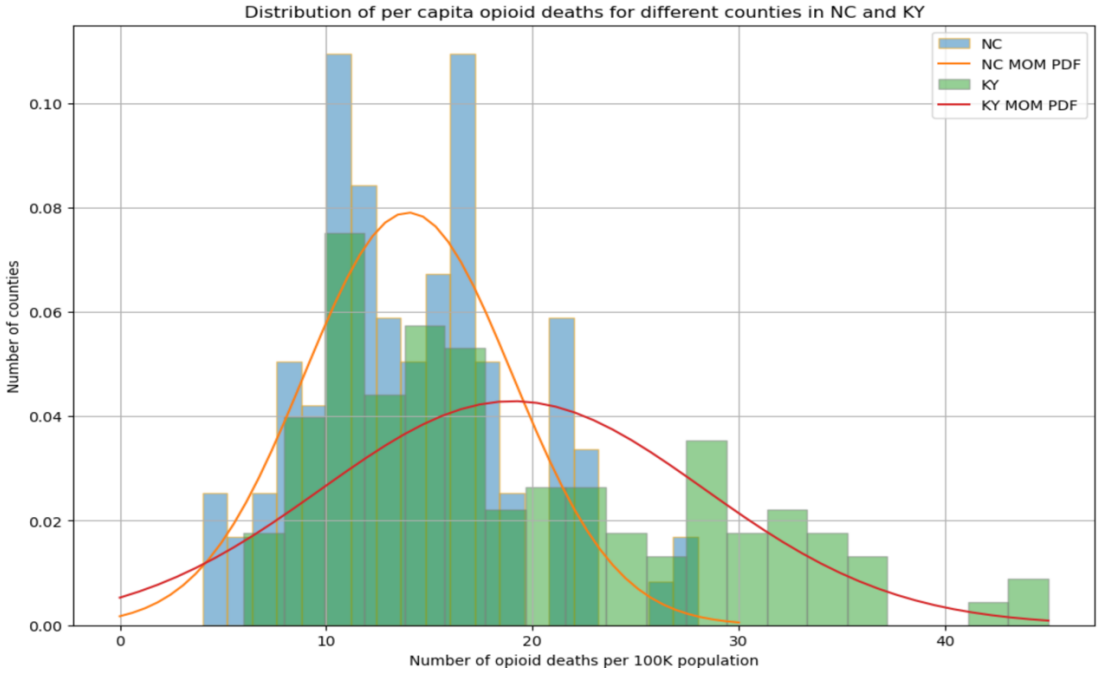
Evaluate a distribution for the Normalized Mortality Rate

- Choose a distribution for Normalized Mortality Rate

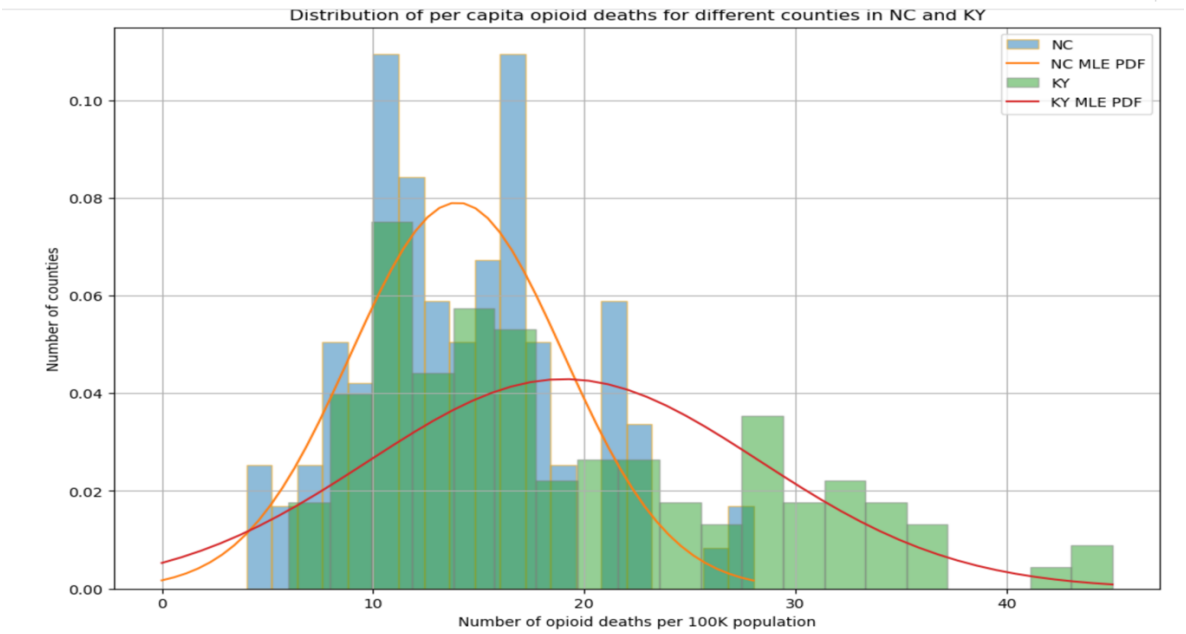
We have used KY and NC state for distribution for the normalized mortality rate.

- Develop distribution estimator with - Method of Moments (MoM), Maximum Likelihood (MLE), and Kernel Density Estimation (KDE)

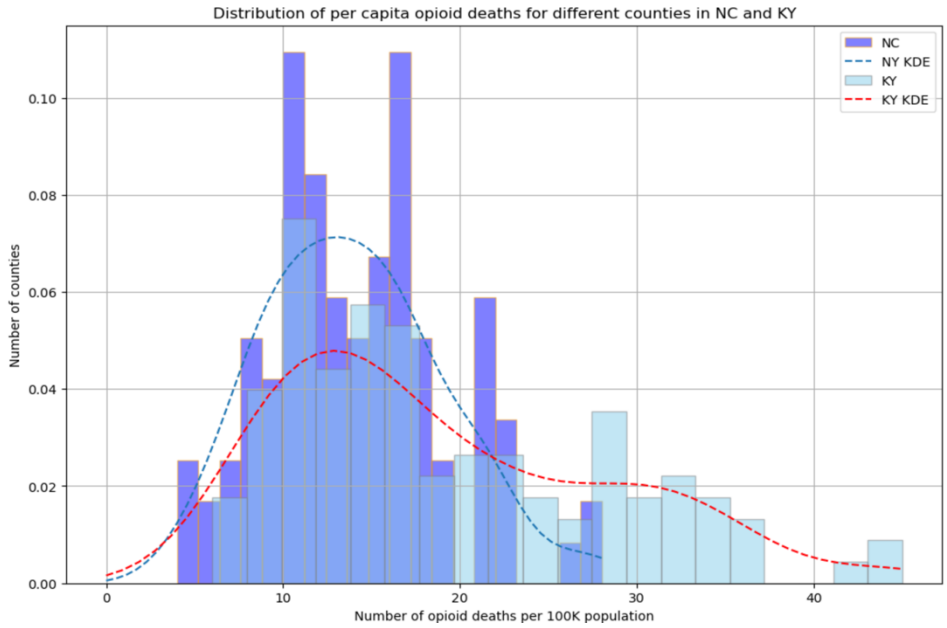
Method of Moments (MoM)



Maximum Likelihood (MLE)



Kernel Density Estimation (KDE)



Discuss which estimator works the best and why

- KDE estimators provide better distribution of data than MOM and MLE estimators

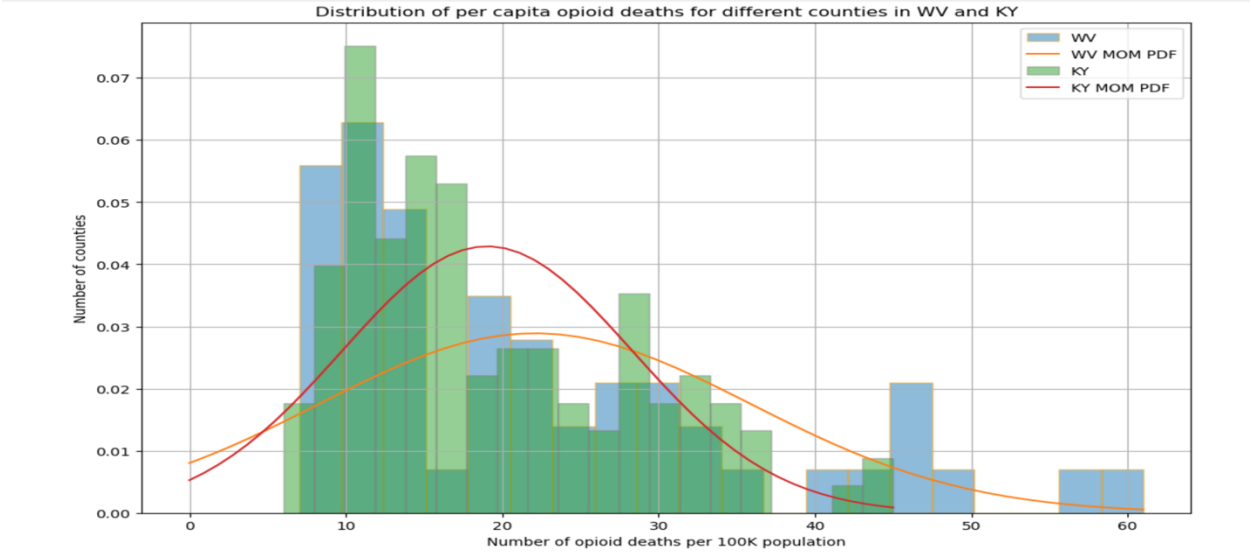
- The graph lines provide better information about the increase and decrease denth rate.

Select the top two states identified in Stage I and recreate the M1.2 task

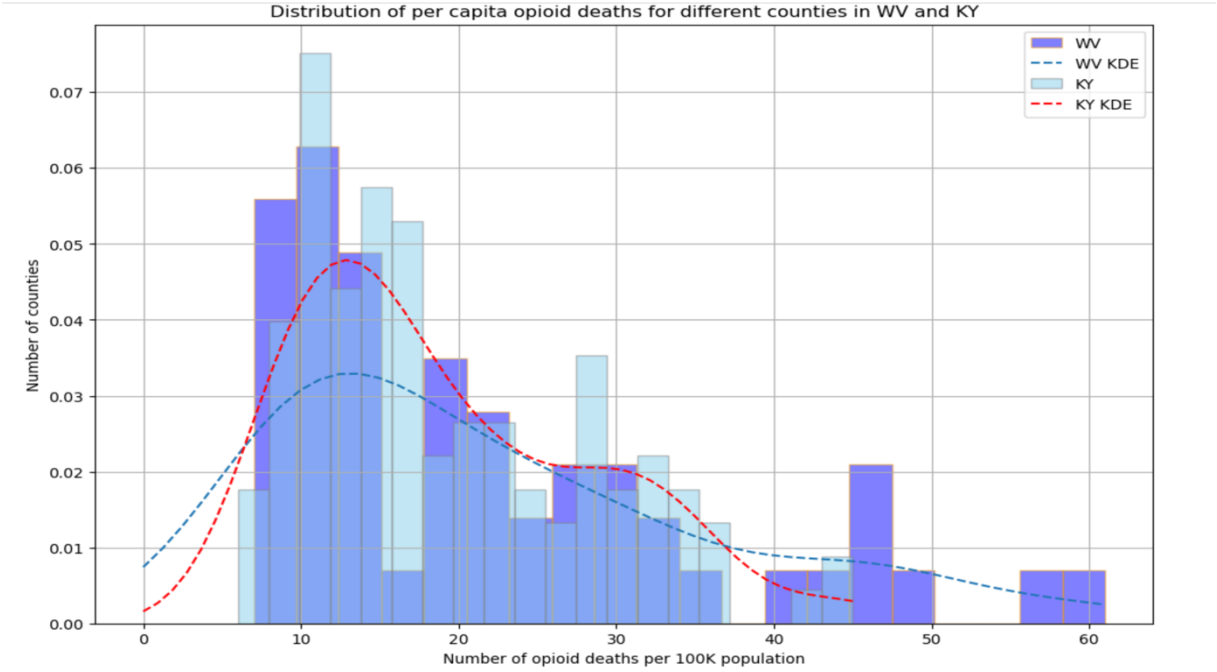
- Kentucky(KY) and West Virginia (WV) were identified top two states in Stage I
- Create dataframe for West Virginia for Opioid Mortality (Normalized Mortality Rate) 2019 Data

Method of moments

We will be calculating method of moments using norm.pdf function because we are using normal distribution



Kernel Density Estimation (KDE)



Discuss the results

- MOM, MLE and KDE estimatos show similar trends for Kentucky(KY) and West Virginia (WV) states.
- West Virginia (WV) has higher death rates compared to NC state analyzed earlier.

Task 2: Hypothesis Testing and Regression

M2.2.1 Formulate Hypothesis for 5 identified variables in Stage 1 and test the hypothesis

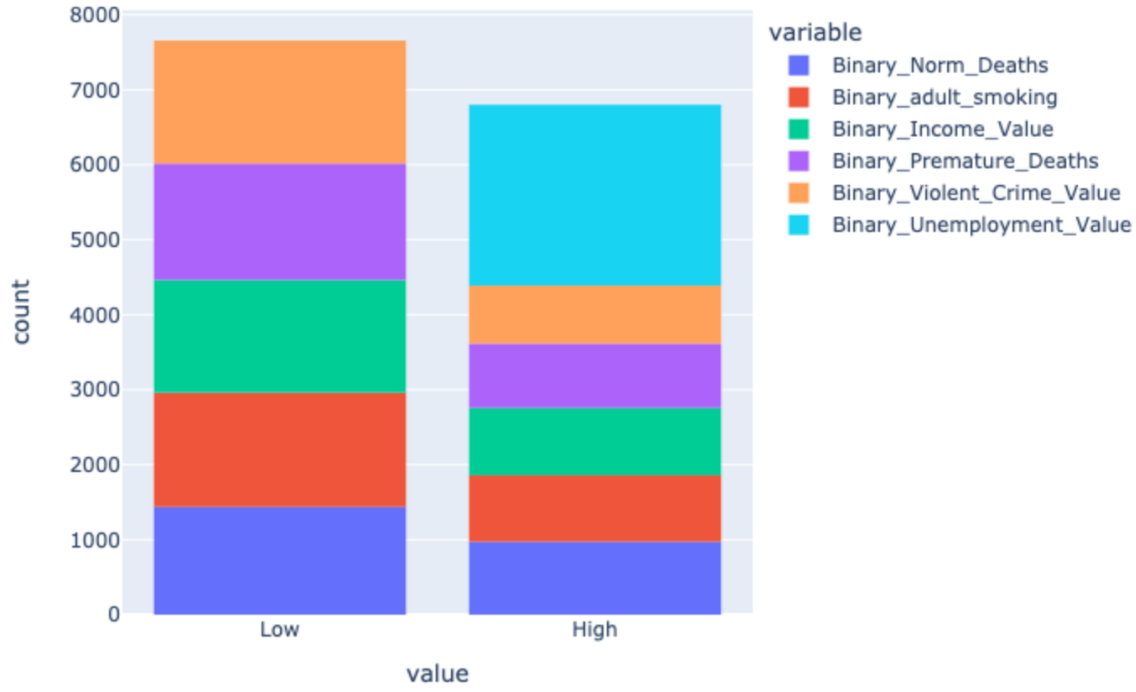
Variables identified in Stage I:

- Income inequality raw value
- Adult smoking raw value

- Premature death raw value
- Violent crime raw value
- Adult smoking raw value

We are normalizing variables determine with 100 k population

We are using mean function on the categorical variables to divide them into high and low values.



Formally state the Null and Alternative Hypothesis

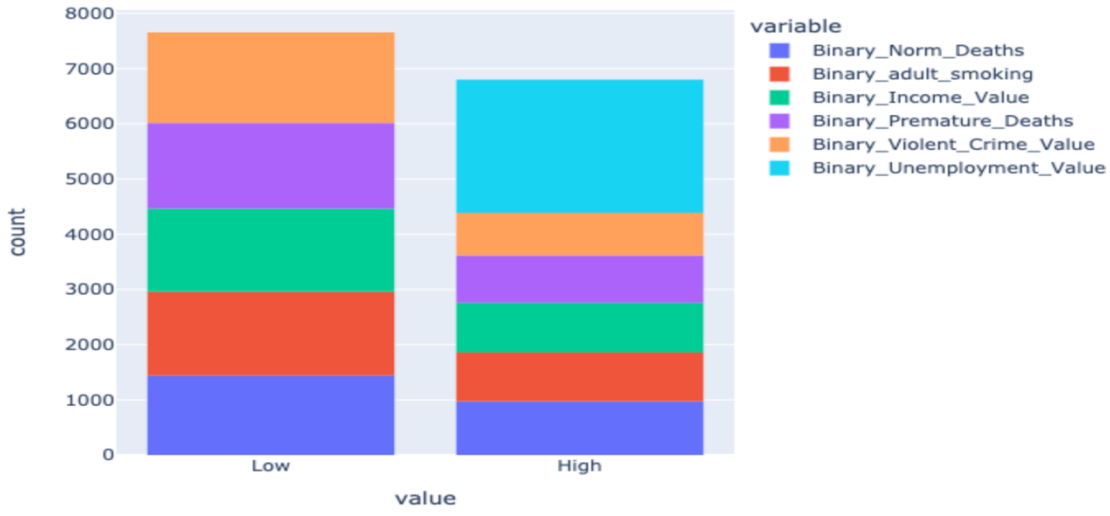
Task 2: Hypothesis Testing and Regression

M2.2.1 Formulate Hypothesis for 5 identified variables in Stage 1 and test the hypothesis

Variables identified in Stage I:

- Income inequality raw value
- Adult smoking raw value
- Premature death raw value
- Violent crime raw value
- Adult smoking raw value

Compare the distribution of the variables



Prove or disprove the following Null Hypothesis:

Variable 1 - Normalized_Adult_smoking_Raw_Value

- Null Hypothesis H0: The mortality rate due to Opioid is dependent on adult smoking rate in US.

- Alternate Hypothesis H1: The mortality rate due to Opioid is not dependent on adult smoking rate in US.

Variable 2 - Income inequality_Raw_Value

- Null Hypothesis H0: The mortality rate due to Opioid is dependent on Income inequality in US.
- Alternate Hypothesis H1: The mortality rate due to Opioid is not dependent on Income inequality in US.

Variable 3 - Premature death raw value

- Null Hypothesis H0: The mortality rate due to Opioid is dependent on premature death rate in US.
- Alternate Hypothesis H1: The mortality rate due to Opioid is not dependent on premature death rate in US.

Variable 4 - Violent crime raw value

- Null Hypothesis H0: The mortality rate due to Opioid is dependent on Violent crime rate value.
- Alternate Hypothesis H1: The mortality rate due to Opioid is not dependent on Violent crime rate value.

Variable 5 - Unemployment raw value

- Null Hypothesis H0: The mortality rate due to Opioid is dependent on Unemployment rate value.
- Alternate Hypothesis H1: The mortality rate due to Opioid is not dependent Unemployment rate value.

Define the type of hypothesis and the thresholds

Two-sample t-test is used to investigate whether the means of two independent data samples are same

We have seen the p-value is greated than 0.05 threshold for all the variables so we have to reject the Null Hypothesis and Accpet the alternate Hypothesis.

The p-values for variables with high mean values and low mean values are very different and have consideribly large values.

M2.2 Perform linear regression to discover patterns

Inferences:

- The slopes for Income Raw Value and Unemployment Raw values are negative which states that these variables are inversely proprtional to normalized dispensing rates
- The slopes for adult smoking raw value, violent crime value and premature death value are positive which states they are proprtional to normalized dispensing rate.

Test non-linear model with the 5 + 1 variables (n=2,3,4)

The R squared value is 0.267 for 2nd order polynomial regression

The R squared value is 0.273 for 3rd order polynomial regression

The R squared value is 0.294 for 4th order polynomial regression