Opiod Mortality Data Analysis Report: Fall 2022

Priyanka Budavi , Manish Shah , Aditi Darandale , Raveena Arasikere Rakesh
University of North Carolina Greenboro

1

Abstract. *The goal of this project is to identify the factors that affected the mortality during the OPIOD Epidemic. In the United States drug overdose was significant in young adults and was mainly because of opioids. These drugs are used as powerful pain relievers however, they also produce feelings of euphoria. This makes them highly addictive and prone to abuse. With the intake of more percentage of drugs it led to a drug overdose. Statistics say the problem started with the over prescription of legal pain medications, but this has recently intensified as people started mixing these drugs with heroin, fentanyl etc. Most of the pharmaceutical companies also stated that the pain killers are no harm to the human and made people believe that these are the best for quick relief. Thus, people purchased these medicines them and started getting addicted to these medicines. Our goal is to derive the insights from the opiod dataset and understand the patterns using data science techniques.*
The project was divided in four stages and each stage goal is as follows:

- *Data Understanding and Linking*
- *Data Modelling*
- *Distributions and Hypothesis Testing*
- *Dashboard*

1 Introduction

The goal of this project is to identify the factors that affected the mortality during the OPIOD Epidemic. In the United States drug overdose was significant in young adults and was mainly because of opioids. These drugs are used as powerful pain relievers however, they also produce feelings of euphoria. This makes them highly addictive and prone to abuse. With the intake of more percentage of drugs it led to a drug overdose. Statistics say the problem started with the over prescription of legal pain medications, but this has recently intensified as people started mixing these drugs with heroin, fentanyl etc. Most of the pharmaceutical companies also stated that the pain killers are no harm to the human and made people believe that these are the best for quick relief. Thus, people purchased these medicines them and started getting addicted to these medicines. The statistics of demographics shows that the drug overdose was common amongst Americans which was 70% of the data and 12-17% was among the Black Americans or Hispanic race. It also came to light that Veterans, who suffered from severe chronic pain because of their service, account for high number of opioid-related deaths. Their ratio is high compared to the rest of the population of the country. The study also showed that drug consumption was very common among the people with low income or below poverty line as compared to the other class of the society. Also, the numbers talk about the people with no education, consumed the drugs more often and were involved in the illegal activities which led to high fatalities. The consumption of Opioid also led to other health diseases like high rates of hepatitis C, HIV, etc., HIV was mainly due to shared syringes. It was also observed that pregnant mothers who consumed drugs also passed the drug to the infants even before the birth and this led to serious issues like incidences of neonatal abstinence syndrome, or withdrawal symptoms experienced by newborns exposed to drugs while in the womb. The percentage recorded was highest from 1999 – 2018. Thus, this led to the increase of the mortality rates in the country. The analysis also correlated endemic to the minority physicians who is likely to serve the undeserved communities. There were many prescriptions from these physicians which led to the high mortality rate. The age also played an important role during the analysis. The consumption was more common among young adults and the overdose led to increased violence, injuries as they had no control on what they were doing. To reduce the endemic there were many programs started like educating patients on how to safely use and store the opioids, enforcing state laws on drugs, prescription monitoring programs, creating awareness about the risks of prescription opioids, and the cost of overdose on patients and families. The variables linked to opioid endemic are racism, increased deaths in car accidents, physical abuse, mental issues, Deaths due to overdose of drugs, also neonatal deaths as pregnant women started consuming drugs which led to the complication during delivery. There were many car accidents due to excessive drinking and drug overdose. Age group was another factor from which we get to know that it was very common among the youngsters, unemployment due to crime rates committed by people, median household income where the individuals lost their jobs due to overdose and involving in violent activities, analysis of poverty as it was observed that people who were poor consumed more drugs compared to other class of the society. It was also found that education was inversely correlated to drug overdose. Less educated people were involved more in the illegal activities which led to the high rates of opioid deaths. The mentioned variables have affected people due to the addiction of the drug. Also, the pharmaceutical companies who claimed that the pain killers would not affect people was one of the major problems that led to the endemic.

2 Stage - I Downloading the data-sets and analyzing the effect of opioid deaths

In stage-1 we are trying analyze the data and try understand what could be the variables that are associated with

opiod deaths. There are several factors where the increase in deaths could be associated with the variables. Below are the reasons why I think the variables have affected the increase in in growth rates for deaths.

- Premature death raw value - help us understand how many pregnant women consumed the drug and what was the percent of neonatal deaths due to the drug consumption/overdose.
- Violent crime raw value - Due to the overdose people get violent and this helps in the analysis of the crime rate due to overdose.
- Drug overdose deaths raw value – This variable helps finding deaths due to overdose from the total population.
- Motor vehicle crash occupancy rate raw value – Whenever there is drug intake people are not in right state of mind and not alert while driving so with this data we could analyze the accidents due to overdose of drugs.
- Alcohol-impaired driving deaths raw value - would give information on the percentage of deaths due to excessive alcohol consumption
- Firearm fatalities raw value - During this endemic, firearm fatalities increased as people who had consumed drugs were not in the state of mind to keep the house safe and might have used the stove and forgot to turn it off.
- HIV prevalence - Many people with opioid use disorder, who initially were prescribed oral drugs to treat pain, now inject prescribed or illegal opioids. High-risk injection practices such as needle-sharing are causing a surge in infectious diseases. Additionally, risky sexual behaviors associated with injection drug use have contributed to the spread of sexually transmitted infections.
- Population – From this variable we can calculate the percentage of deaths from the total population
- Sexually transmitted infections raw value – Due to the increase in the drug cases there were many syringes that were used without sterilizing, and this led to the increase in the mortality rate.
- Injury deaths raw value – Drug overdose led to injury of people as they were not in the right state of mind and thus mortality rate increased.

After identifying the variables that could affect the increase in mortality rate, merge all the data-sets and create a super data-frame using the merge function. While merging all the three data-sets, data cleaning was important as the data had columns that had no values. In this task we also normalized the deaths column to get the values in a particular scale. The data-set obtained after merging all the three file is of shape (2527 , 542). The next task was to find the counties and states with high mortality rate. Here we used the sort function to sort the Norm deaths values from the super data-set which gives us information county wise. Thus we tried to display County and Norm deaths in the descending order to get the top ten counties with high mortality and then find the mean and median for the counties and states. Below are the inference drawn from the data. There were few issues with the data like inconsistent values or missing values. Before performing any type of analysis its necessary that these values are taken care. For example, the health ranking dataset has many missing values and this can have a significant effect on the conclusions that can be drawn from the data. Population growth raw value', Ratio of population to primary care providers other than physicians have negative values. There are few solution that we used in the project while dealing with missing and inconsistent values and they are as follows:

If whole column is empty or NULL, we can discard/remove/drop it. If there are few missing values in a column, then it can replaced by the mean, rolling average values. It will benefit by not distorting different parameters for analysis. We can use fillna() to fill it with any relevant values for the column.

Inference for Stage 1

- From the data we observe that West Virginia had the highest mean and median values. Mean: Since each state has different number of counties the mean value is dependant on that and from the statistics point of view the mean value in all counties in the state of west virginia is 22.094340.
- The median value gives an information of where the 50% of the data is lying. West Virginia has 19.0 as the median which means the death value in around 19 in each of the counties of a west virginia.
- McDowell County in West Virginia has the highest mean and median values. With mean and median being 61.0. As explained above the death rates in this county was at least 61. Kentucky State is the second highest in mortality death rates.

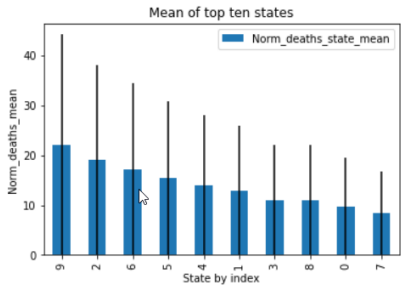


Figure 1: Mean Top 10 States

3 Stage - II Data Modelling

The goal of Stage II is to develop the data for modeling and comparative analysis. Here you will be graphically comparing how different states are doing with respect to opioid mortality rate. And you will be also analysing county based information for different states in the US.In this stage the task was to understand the distribution of normalized deaths in US. For viewing the graphical view we have used histogram plots for the entire US data with the normalised deaths in US column. This histogram provides

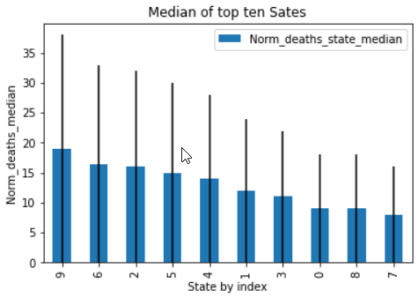


Figure 2: Median Top 10 States

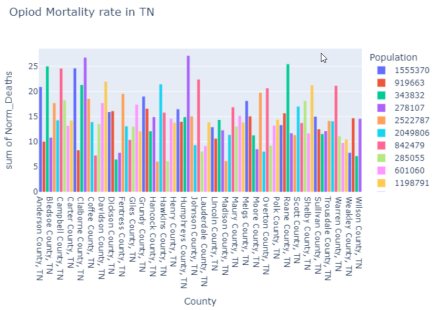


Figure 3: TN Opiod deaths rate

the two lines one for Histogram and other for mean and median for number of deaths nationwide. * Importing historical data and identifying peaks in Opioid mortality.

3.1 Dataset

The dataset used is download is downloaded from the "./data/1999-2020_Drug_Overdose_By_ Category.xls" . The variables in the dataset are described below.

County Code	Int64
Population	int64
Deaths	Float64
Female population 65+ raw value	Float64
Total female population raw value	Float64
Population growth raw value	Float64
FIPS	Float64
Opiod_Dispensing_Rate	Float64

Figure 4: Histogram US Norm Deaths

We have checked if there are any infinite values. There was no infinite values in the opiod mortality rate dataset. We have also plotted the histogram as shown below.

We can infer from the above histogram the data is discrete. We can also infer that data is uniformly distributed so we are using median to calculate the measure of centre.

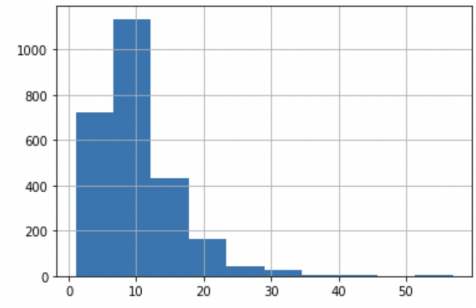


Figure 5: Histogram for Opioid mortality rate dataset

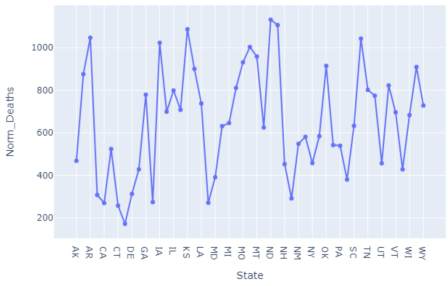


Figure 6: Opioid mortality across different states

We can infer from the above graph the highest mortality rate is across state ND and the mean death rate is 1131.521. So, we can hypothesise that the mortality rates across each of the states maybe influenced by the population. So, as the population is increasing the mortality increases. The peaks can be the result of the high opiod dispensing rate.

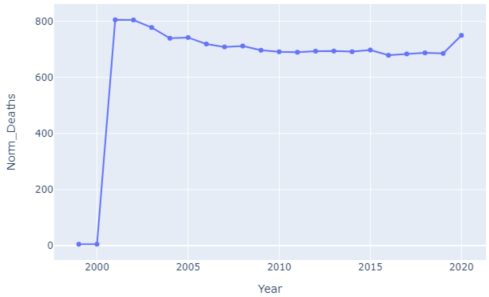


Figure 7: Opioid mortality across different states

We can infer the highest mortality rate in the year 2001 and mean death rate 734.8614. We can infer from the graph opiod mortality rate are changing.

Discuss why there was a peak the highest mortality rate was in the year 2001 with a mean death rate of 734.8614. Also,we can observe form the graph that the opiod mortality rate is significantly fluctuating, there was a sudden spike in early 2000's and then the rates have been constatnt since then, this maybe because there is a lot of missing data for the years 1999 and 2000. The peaks could have also been due to opiod dispensing rate.

To identify the trends in state the dataset was having county and state abbreviation together so we need to split that data for further operations. To find the increasing trend we have use 5 years information to see the increase in mortality rates. To compare the mortality with the previous years we find a difference and we have added difference coloumn. We are grouping state wise and calculating the differences of deaths values. So, this provided us with the total deaths in each state in year by sorting them we get the top five states with increasing values as shown in Figure below. We are identifying 5 states which are increasing in opiod numbers and also identified 5 states which are reducing their opiod numbers. Washington DC have the highest decrease in death rate over the period of 5 years and then the states i.e DE, HI , RI and then NH were the states in the decreasing number of Norm_Deaths.The above states have special programs granted by US government such as Prescription Drug Monitoring Programs (PDMPs) which helps in reducing the rate of opiod precription.

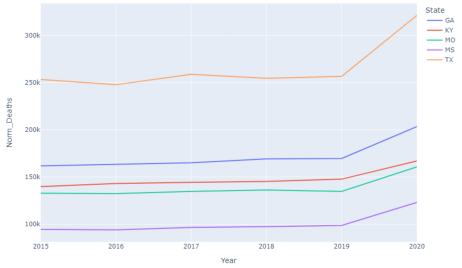


Figure 8: Five states with increasing trend

We can infer from the above figure Texas state has the highest population where as MS has the lowest population so the death rates are high in Texas as compared to other states.Since 2000 the use of opiod increased in the US , but the reports say that in texas, from 2015 there was increase in female deaths compared to male deaths.So from 2015 to 2019 the deaths were not increasing they were almost consistent but in 2019 there was a slight increase in the deaths.

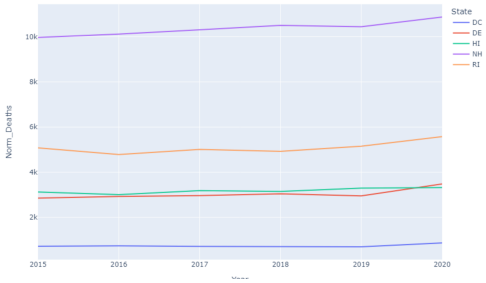


Figure 9: Five states with decreasing trend

Washington DC had the highest decrease in death rate over the period of 5 years and then the states i.e DE, HI , RI and then NH were the states in the decreasing number of Norm_Deaths. The above states had special programs granted by the US government such as Prescription Drug Monitoring Programs (PDMPs) which helps in reducing the rate of opioid prescription.DC has the lowest population compared to the other 4 states which results in the lowest number of opioid deaths.

Plotting a scatter plot graph of normalized mortality by state with respect to the log of the population. For this task, we have taken the opioid data from 1990's to 2020. Initially, determining the mortality rate by dividing the death value by population per 100000. Taking the natural log of population value reduces the large values of population typically in millions to 100, 000. By grouping the data by county and plotting the total sum of Norm_Deaths andLog of population in a scatter. we are plotting a scatter plot graph of normalized mortality by state with respect to the log of the population.

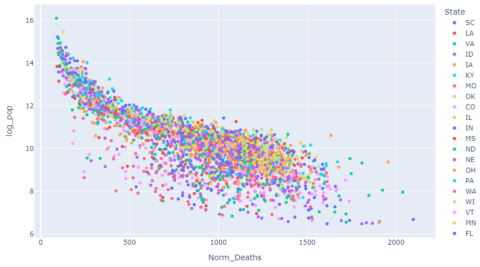


Figure 10: Scatter plot for Normalized deaths and log populations

We can infer from the above graph inverse relationship between log_pop and Norm_Deaths.It shows exponentially decreasing relation for all counties as norm death value increases. The dots are color coded by States and most of the cases and in the range of norm death value between 700 to 1400 with log population value between 8 to 10.

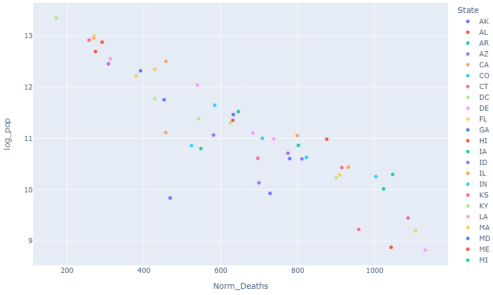


Figure 11: Scatter plot for Normalized deaths and log populations

We can infer from the scatter plot the values of norm_deaths values of 50 states given the log of population. we can infer that there exists a negative relationship between the state mortality rate and log of population. It is in sync with the above county based plot. The scatter plot shows there is an linear relationship between norm_deaths and log_pop.

In member task the main idea behind this task is to analyze the peaks generated by the mortality rates, when we analyze them for each state and year. Analyzing the mortality rate of each state

- Reading data of the year 1999-2000 dataset. Filling the nan values.
- Normalizing death rate by population per 100,000, So the scale for measuring the rates across different states is same
- Groupby using state and calculate mean of normalized deaths. Further plotting those values of each state.

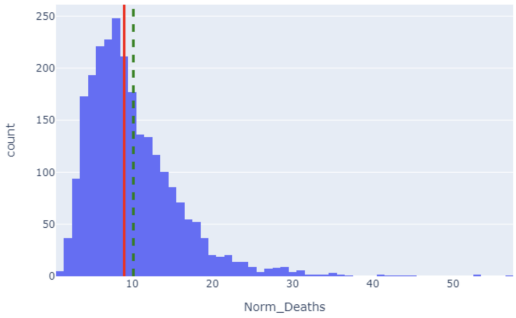


Figure 12: Histogram US Norm Deaths with mean and median line

Further, we are understanding variables to mortality relationships of 2019 data and we are codifying them number of Deaths per 100k Population - Norm_Deaths and store it in label coloumn. The codification is based on quantile distribution of the Normalized deaths. that are Very Low (v_low) from [1.999 - 8.0], Low (low) from [8.0 - 11.0], High (high) from [11.0 - 16.0] and Very High (v_high) from [16.0 - 64.0] as shown in below figure.

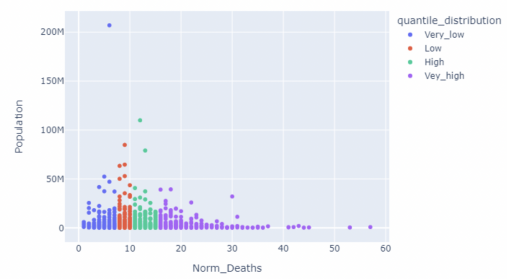


Figure 13: Displaying plot for Normalised deaths compared to population and displaying codification variables

Now try the variables you have identified in Stage I and plot them as a second variable to Normalized Mortality in a scatter plot to observe any trends. We have identified total female population from stage 1 plotted them as a second variable to Normalized Mortality in a scatter plot to observe the trends.

We can infer from above graph that when we compare the opiod mortality rate against the total male population, we can that the highest male population goes under the very_high quantile,This means that the opiod mortality rate is higher where the male population is higher.

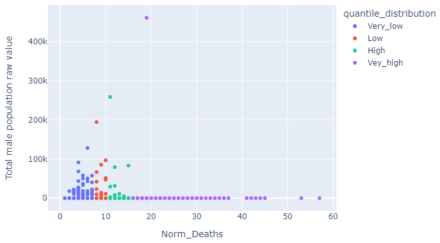


Figure 14: Displaying plot for total male population raw value wit normalized deaths

4 Stage - III Distributions and Hypothesis Testing

The goal of Stage III is to develop advanced data for modeling for comparative analysis and hypothesis testing.

4.0.1 Distribution Analysis

:In this section, we analysed the distribution across two states NC and KY. Then we first calculated the means across these states,

- The mean across NC is 12.10 and the median is 18.11.
- The mean across KY is 12.10 and the median is 16.0.

Next, we plotted a histogram of the Normalised opioid mortality rate across NC and KY in a single graph and indicated there respective means with a vertical line. Next, we analysed the distribution across the NC and KY

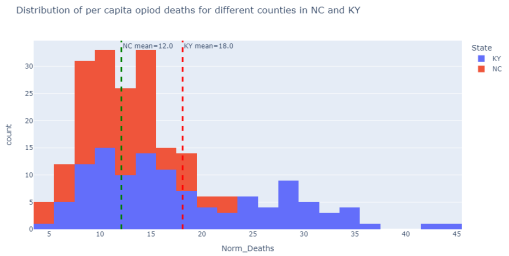


Figure 15: Opioid Mortality rate across NC and KY

dataset.Since the mortality rate across the entire dataset is rounded we can considered it as a **discrete distribution**.We assume it to be a Poisson distribution and try to fit the data using the mean and standard deviation which are the estimation parameters for the **poisson distribution**. Now, that we assume our distribution to be a poisson distribution, we will continue to develop the distribution estimators i.e Method of Moments(MOM),Maximum Likelihood estimate(MLE) and Kernel Density Function(KDE) by calculating the probability mass function(pmf) across this Poisson distribution. In a Poisson distribution , the calculated mean and the standard mean are going to be the same.Hence both the MOM and MLE are going to be the same, and this was verified by plotting a graph.

When we plot the graph both the MOM,MLE and KDE together in a single graph we can see that curves overlap. Hence, we regarded them both as the same. Also from

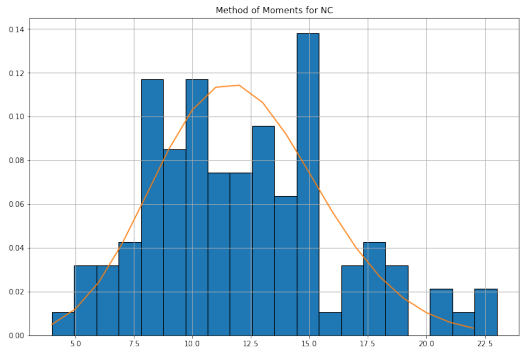


Figure 16: Method of Moments(MOM) across NC

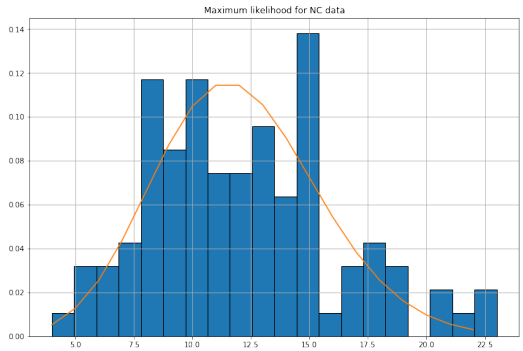


Figure 17: Maximum Likelihood estimate(MLE) across NC

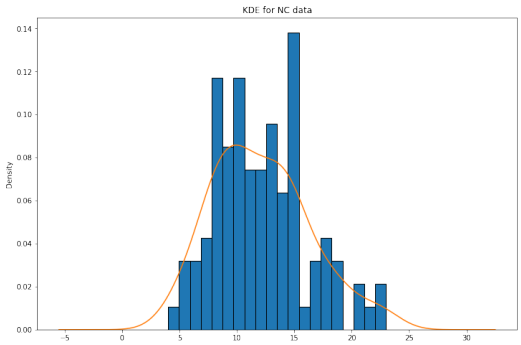


Figure 18: Kernel Density estimate(KDE) across NC

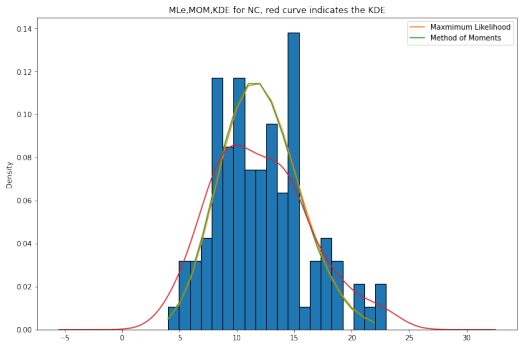


Figure 19: MOM,MLE and KDE across NC in a single graph

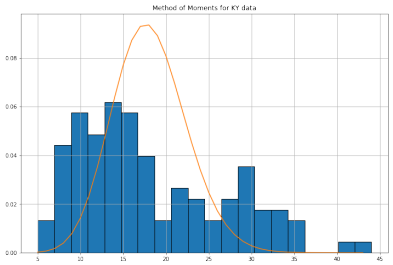
the Fig.15 we can see that the MOm/MLE curve fits the distribution better than the KDE curve, so we regard it as the best-fir curve. We then performed the same analysis on the KY data to analyse which distribution estimator fits the data the best. From the final curve generated, we can see that KDE generates a best fit curve and hence we choose KDE as the best fit curve for the KY data.Next, we plotted both the NC and KY data in a single graph along with their mean values and the best-fit curves (Here MOM/MLE is the best fit curve for the NC data and KDE is the best-fit curve for KY data-which is indicated by the legend) Similarly, we performed the same distribution estimation on the 2 states with the highest moratlity rates that was analysed in stage-1(which we WV and NM). From the graph generated we can see that MOM/MLE can be considered as the right fit curve for wv data. Next, we plot both WV and NM data into a single graph, along with their mean values and their best-fit curves.

4.0.2 Hypothesis Testing and Regression

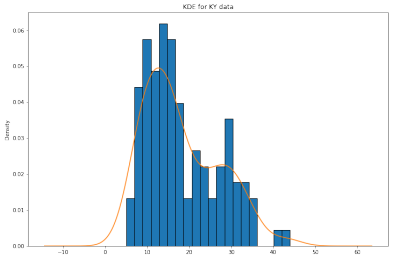
1. First we ,formulated Hypothesis for 5 identified variables in Stage 1 and test the hypothesis.

   The 5 variables selected in the stage-1 of the project were-

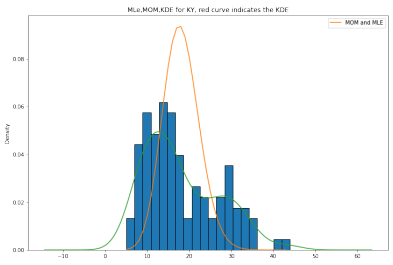
   - Excessive drinking raw value
   - Sexually transmitted infections raw value
   - Preventable hospital stays raw value
   - Mental health providers raw value
   - Premature death raw value



(a) MOM/MLE across KY



(b) KDE across KY



(c) MOM/MLE and KDE for KY data

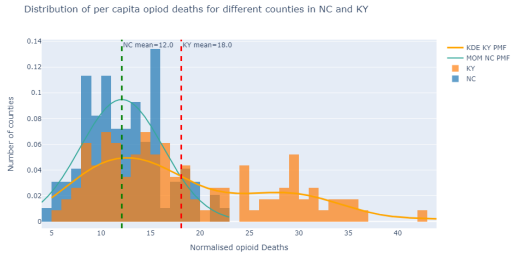
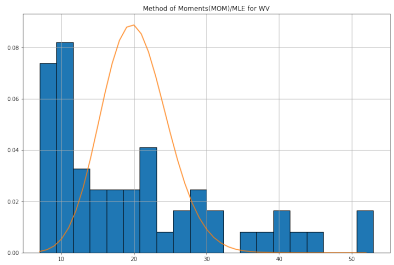
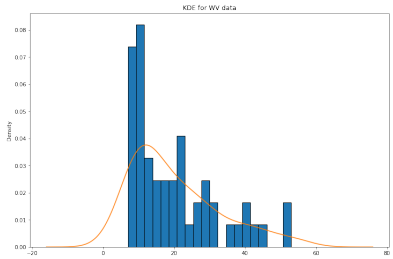


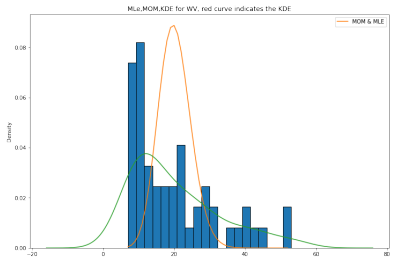
Figure 21: Opioid Mortality rate across NC and KY along with their best-fit curves



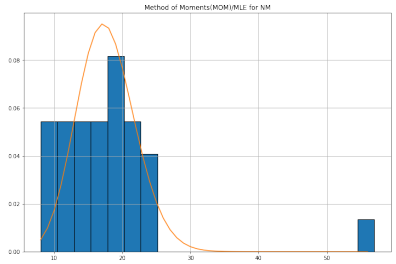
(a) MOM/MLE across WV



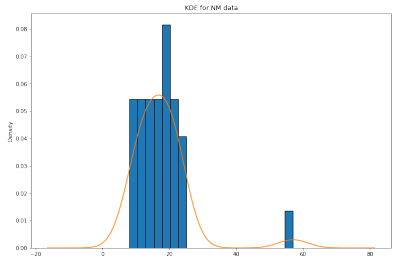
(b) KDE across WV



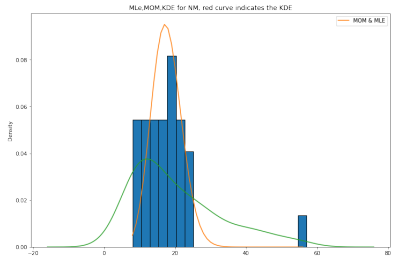
(c) MOM/MLE and KDE for WV data



(a) MOM/MLE across NM



(b) KDE across NM



(c) MOM/MLE and KDE for NM data

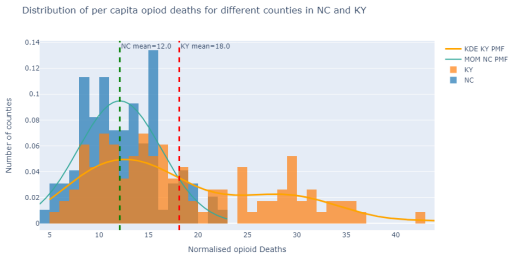


Figure 24: Opioid Mortality rate across WV and NM along with their best-fit curves

2. After we identified the 5 variables, we then need to normalize each of the column with population per 100,000 so that the scale of comparison is the same.
3. Next, we need formulate the hypothesis and then perform a two-sample t-test because we will be analyzing two independent variables against a single target variable.
4. We then calculate a threshold for each of the selected 5 variables, with the same formula-(min+max)/2. The Values that are less the threshold are set to **LOW** and the values greater than the threshold are set to **HIGH**.Once a threshold is set, we then perform a **two-tail ttest** to see check if we can accept or reject our NULL hypothesis.
5. We compare the HIGH values across Normalized Mortality rate with the HIGH values across the selected 5 variables.
6. Next, We compare the Low values across Normalized Mortality rate with the Low values across the selected 5 variables.

Once, all the pre-processing on the data is done, we then define a hypothesis across each variable and the result across each are as follows:

1. For excessive drinking raw value:
    - Hypothesis:
        - *H0: Across any state in the US, mean of excessive drinking rate similar to mean of the opiod mortality rate*
        - *H1: Across any state in the US, there is significant difference between the means of ex-*

cessive drinking rate and opiod mortality rate

- The threshold set is (max+min)/2=0.13396205075342518
- The hypothesis test performed is two-sample T-test
- **Result:** Reject the Null hypothesis.

2. For Sexually transmitted infections raw value:
   - Hypothesis:
     - *H0: Across any state in the US, mean of Sexually transmitted infections raw value is similar to mean of the opiod mortality rate*
     - *H1: Across any state in the US, there is significant difference between the means of Sexually transmitted infections raw value rate and opiod mortality rate*
   - The threshold set is (max+min)/2=665.0270501638447
   - The hypothesis test performed is two-sample T-test
   - **Result:** Reject the Null hypothesis.
3. For Preventable hospital stays raw value:
   - Hypothesis:
     - *H0: Across any state in the US, mean of Preventable hospital stays raw value is similar to mean of the opioid mortality rate*
     - *H1: Across any state in the US, there is significant difference between the means of Preventable hospital stays raw value rate and opioid mortality rate*
   - The threshold set is (max+min)/2=6557.256221900106
   - The hypothesis test performed is two-sample T-test
   - **Result:** Reject the Null hypothesis.
4. For Mental health providers raw value:
   - Hypothesis:
     - *H0: Across any state in the US, mean of Mental health providers raw value is similar to mean of the opiod mortality rate*
     - *H1: Across any state in the US, there is significant difference between the means of Mental health providers raw value rate and opiod mortality rate*
   - The threshold set is (max+min)/2=0.008067995525588043
   - The hypothesis test performed is two-sample T-test
   - **Result:** Reject the Null hypothesis.
5. For Premature death raw value:
   - Hypothesis:
     - *H0: Across any state in the US, mean of Premature death raw value is like mean of the opioid mortality rate*
     - *H1: Across any state in the US, there is significant difference between the means of Premature death raw value rate and opioid mortality rate*
   - The threshold set is (max+min)/2=11141.092263818999
   - The hypothesis test performed is two-sample T-test
   - **Result:** Reject the Null hypothesis.

Next, we Perform linear regression to discover patterns:

1. We first performed a linear regression between Normalized Mortality and Opiod_Dispensing_Rate. From, the generated graph we can see that there a linear relationship between opioid morality rate and the opioid dispensing rate.

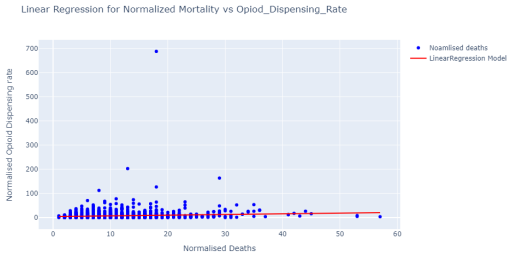


Figure 25: Linear Regression model for Normalised Opioid Mortality vs Normalised Dispensing rate

2. Next we performed a multiple linear regression model with your 5 variables and Opiod Dispensing Rate.
   - To perform a multiple linear regression model, we fit the data with the 5 variables and use the opioid dispensing rate as the target variable.
   - The results were as follows,
   - RMSE score for multiple linear regression: 14.760761738406782
   - R-squared value: 0.24644317140642225
   - The square root of the variance of the residuals is represented by the RMSE error- which is approximately 15, we can say that the model performed well but not did a great job in predicting the opioid mortality rate when the 5 variables are given.
   - Since, R2 value is less than half i.e., less than 0.5 we can say that the model does

not do a good job in explaining the proportion of the variance between the dependent and independent variable.

3. Next we tested for non-linear model with the 5 + 1 variables (n=2,3,4).
4. We then performed a non-linear(polynomial) regression across all 5 variables in comparission with the Normalised opoiod mortality rate,
5. For degree 1:
    - RMSE score for multiple linear regression for degree:1 14.760761738406782
    - R-squared value for degree:1 0.24644317140642213
6. For degree-2:
    - RMSE score for multiple linear regression for degree:2 12.485227815405297
    - R-squared value for degree:2 0.4608725907118788
7. For degree-3:
    - RMSE score for multiple linear regression for degree:3 7.85071603644558
    - R-squared value for degree:3 0.7868344556440677
8. For degree-4:
    - RMSE score for multiple linear regression for degree:4 6.765188282082652
    - R-squared value for degree:4 0.8417082627869036

    From the above generated results, we can see that the RMSE error is the lowest for degree-4 and the r-squared value is also the highest for degree-4 polynomial regression. Hence, we can say that the model performs the best for the degree-4 polynomial regression.

5 Stage - IV Dashboard

This is the final stage of the project. This stage aims at developing an interactive dashboard based on the analysis done so far. To design a dashboard in python, we'll be utilizing Plotly along with Dash as the main framework. Dynamic plots and graphical representation are some of the best methods to visualize data. Scatter plots in Plotly are some of the methods to plot the data, we can also plot maps on the dashboard and visualize various parameters on a map too. Data science is a broad field and the methods are not limited to the above-mentioned ones.

5.1 Scatter plot comparison

This task involves creation of a dynamic scatter plot in plotly. We can implement it either using plotly.express or graph.objects packages. The Y-axis of the scatter plot is fixed and represents the normalized mortality rate. While the X-axis contains different variables. We can select various different parameters to compared with the normalized mortality rate. The parameters are the ones that we have filtered in stage-I which we felt affect the mortality rate. For this stage, we are using the Opiod dispensing value, Unemployment value, Drug Overdose value, Insufficient sleep value, and Excessive drinking value.

Figure 10 shows the dashboard created using Dash with the scatter plot. Scatter plot have variables selected from the drop downs and radio buttons.

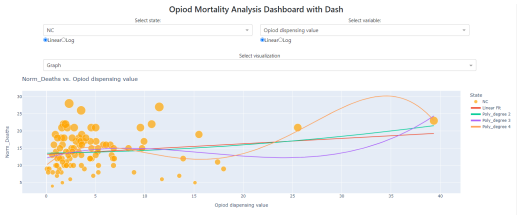


Figure 26: Dashboard with scatter plot

5.2 Incorporate best model trendline

Next step involves generation of linear and non-linear model prediction trendline. This trendline follows the data distribution for the given variables. A linear trendline is the one which shows the trend of data using a straight line. It's a good approach to describe the data distribution when the data is linear. But there are much better options i.e. non-linear trend as generally the data is not always linear. Linear regression is used to plot the linear trendline while Polynomial regression is used to plot non-linear trendline. In this task, we will show polynomial regression for the degree of 2, 3 and 4. As degree increases, the model tries to fit the data in a better manner which may result in overfitting. We have to use our discretion in order to analyse the best fit. Figure 11 shows linear trendline and Figure 12 shows non-linear trendline.

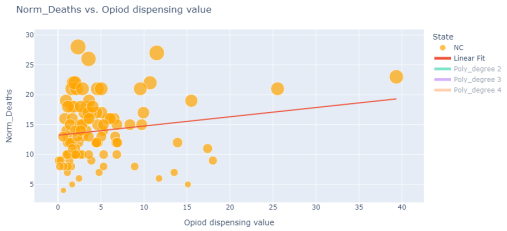


Figure 27: Linear Trendline on scatter plot

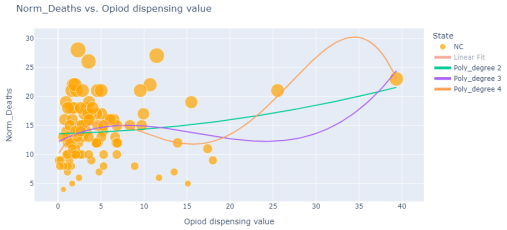


Figure 28: Non-linear Trendline on scatter plot

From Figure 12, we can see that non-linear model with the degree = 3 is the best fit given both linear/non-linear model.

5.3 Incorporate the Data table

A data table is similar to a dataframe. If we wish to show the data in the dashboard, it can done using a data table. Our main dataframe consists of 2527 rows and 542

columns. But we are not showing all these columns but selecting only few required columns and fixing the width and height of data table. Fixing the width and height provides the liberty to include as many data rows and columns we want to add. We are showing only 5 rows per data table entry. For viewing more data, a next button is provided and a scroll bar to view other columns. Figure 13 represents the Data Table created.

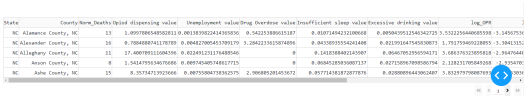


Figure 29: Data Table

5.4 Contains a map displaying values of either variables

We create a map of the USA at county level. It represents all the data for a selected variable and displays a heat-map intensity based on values of variables in the counties. The map changes as the values from the "Select Variable" dropdown changes. We are also creating a dropdown for map and graph selection. If we want to display the graph/map, change the selection from dropdown and vice versa.

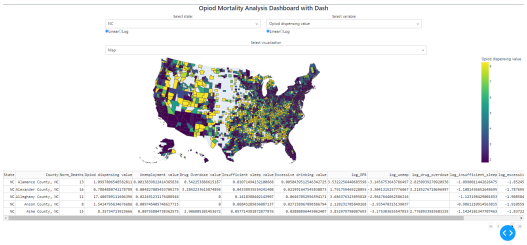


Figure 30: Linear selection in Map

Figure 16 shows the Map output selected a variable from the dropdown. The map also changes the values as we change the values from the radio buttons. From Figure 17, it can be seen that as we change the selection in the radio button from linear to log, the output of map changes too.

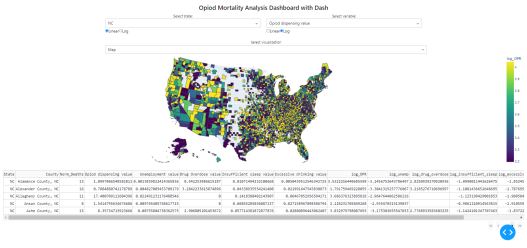


Figure 31: Log selection in Map

5.5 Allow selectors

In this step, we will show in detail about various selectors i.e. dropdowns and radio buttons. There are 2 major dropdowns: one for selection of state and other for selection of variables. The state dropdown contains values of all the state in the data. By selecting the value of state from the dropdown, the scatter plot displays the plot corresponding to that state. Data table contains the data corresponding to selection from the state dropdown.

The variable selection dropdown contains the variables that we want to plot against the mortality rate in the graph and the map.

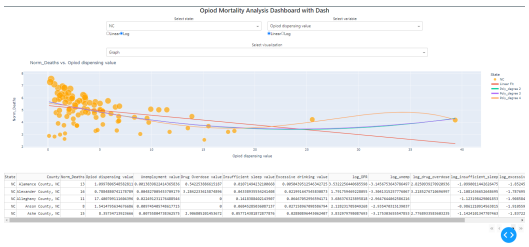


Figure 32: Dropdown and radio button selectors

There are 2 major radio buttons that are incorporated just below the 2 dropdowns. These radio buttons contains the linear/log selection for both x and y axis of the scatter plot. It also affects the selection in the map. Figure 18 shows the entire output of the dashboard with the dropdown selection.

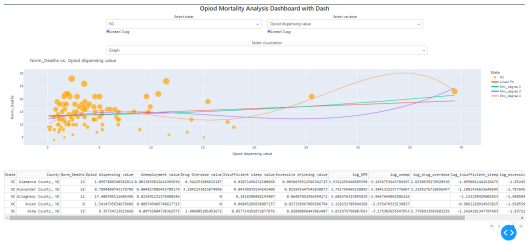


Figure 33: Data table displaying NC state

5.6 Selection of on the graph or Data Table highlights the other ones

For this task, we are altering the Data table based on the selection from the state dropdown. Figure 23, 24 shows the change in the data table based on selection from the dropdown.

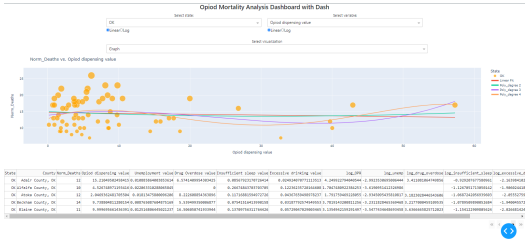


Figure 34: Data table changes the data based on state selection from dropdown

Fig 24 shows the complete dashboard. Here, we will brief all the parameters of the dash table. First parameter is the selector: radio buttons and the dropdown. We have 2 dropdowns in the top and one dropdown for Map/Graph selection. Secondly, the radio buttons are incorporated with the dropdowns and changes the selection criteria to log/linear.

An important parameter of the dashboard is Graph/Map. This shows the Graph and Map based on selection from the dropdown. Last, we have a data table that changes the data based on selection from the state dropdown. All these are incorporated in a single dashboard and create a multi-purpose dashboard.

6 Conclusion

Through all the stages we were able to perform data analysis on opiod dataset. Starting from data cleaning as its important aspect of the analysis to derive meaningful insights from the data. We were also able to analyze the distribution of death rate in the entire US using histograms. The statistical analysis of fitting the right distribution for the data set and then determining the method of moments, maximum likelihood and Kernel density estimation helped us understand the best estimation parameter for a given dataset. Finally, the dashboard was built to see the linear and log trends for deaths and the variables of interest.

References

[1] https://github.com/manishshah1698/CSC605-Fall2022-Team1/

[2] https://link.springer.com/article/10.1007/s40265-017-0846-6

[3] https://github.com/UNCG-CSE/CSC-605_Fall_2022/blob/main/Class_Resources/Lecture_06/Statistics/07_linear_regression.ipynb

[4] https://stackoverflow.com/questions/70328489/create-plotly-distplot-charts-in-plotly-express

[5] https://stackoverflow.com/questions/63865209/plotly-how-to-show-both-a-normal-distribution-and-a-kernel-density-estimation-i

[6] https://pyshark.com/poisson-distribution-and-poisson-process-in-python/

[7] https://plotly.com/python/distplot/

[8] https://www.youtube.com/watch?v=kXBQrPcDYDk