FINAL REPORT

Project by
Sai Ram Ajay Krishna Gabbula Venkata Vijay Krishna Gabbula
Snehith Reddy Kallem Manichandana Kulluri

Data Science CSC-605

THE UNIVERSITY OF NORTH CAROLINA AT GREENSBORO
Greensboro, North Carolina

TABLE OF CONTENTS

LIST OF FIGURS

Chapter 1

STAGE - 1 DATA AND PROJECT UNDERSTANDING

1.1 Introduction

The United States grappling with its worst-ever opioid drug crisis. This had started over-prescription of the pain-relieving drug. In 1990, the pharmaceutical companies reassured that people or patients would not become addicted to pain relief drugs. Increased prescription of opioid medications led to widespread misuse of both prescription and non-prescription opioids before it became clear that these medications could indeed be highly addictive. Opioid overdoses and addiction present significant challenges to communities across the country.

The deaths related to opioid overdose are 78 each day and the prescriptions on opioid are more than 650,000 per day in the USA. Out of all the opioid related deaths one third of the deaths are due to the commonly prescribed opioid drugs. The control in the prescriptions of opioid can reduce the deaths related to them as the past data shows that the increase in the prescriptions lead to more opioid related deaths. However many other factors other than the prescription of opioids needs to be considered to make significant impact in decreasing the mortality related to opioid.

Main aim of this project is to collect the data from different datasets such as Drug Overdose Dataset, County Health Rankings, County Opioid Dispensing Rates and understand and visualize them with the help of Data Science techniques to estimate the impact of Opioid Epidemic across different socio-economic, demographic, geographic variables that are present for the counties in US and to give future predictions.

By applying data analysis techniques to data, we can get to know the effect of drug usage and prescription rate county-wise, what factors affecting the over-drug usage, and how many people were exposed to disorders and death. The analysis results give a better understanding of the problem based on the results which can help to find a solution to prevent or reduce the effects of opioid related issues.

1.2 Analysing the datasets

In this stage we acquire different datasets to analyse the opioid pandemic. The following are the different datasets which are used in this project.

- Drug Overdose Dataset[1]
- County Health Rankings[2]
- County Opioid Dispensing Rates[3]

The opioid drug overdose by category dataset contains data from 1999 to 2000. It includes the cause of death and the number of deaths over the population at the county level. The opioid drug dispensing rate data set contains the drug dispensing rate at the county level. County health ranking includes socio-economic characteristics.

1.3 Linking Data

In stage 1, we link different datasets available to obtain a super dataframe which can be used for further analysis. All the datasets are in CSV format, with the help of Pandas library in python pre-processing the data to make the merge between all the datasets possible.

Merging two datasets is the process of bringing two datasets together into one and aligning the rows from each based on common attributes or columns. Pandas is a library in python which helps to handle the dataframes. So, with the help of pandas merging of the datasets are performed. This can be done by using merge, join and concat functions. Join in pandas is similar to join operation in SQL. Here we are merging the dataframes based on the common columns between them. 'merge' is the keyword used to combine the different types of datasets. The function merge performs the inner join. Inner merge gives the intersection of datasets. Outer merge performs union operation on datasets. Left merge performs the merge operation from left side of the datasets. Right merge performs the merge operation from the right side of the datasets. Merge operation can be performed on two datasets at time like pd.merge(dataframe1,dataframe2)

The super dataframe obtained has 2527 rows and 542 columns. The mortality data is normalized per 100,000 population. The obtained dataframe is saved into a CSV file so that it can be used in the later stages for analysis.

1.4 Issues with the data

The datasets which are used in this project have some issues in it. So these issues needs to be handled before the merge.

Underlying Causes of Deaths Dataset

The Notes column have all values as NaN. This means that the column is of no use in any analysis hence this column is removed from the dataset. Crude rate have some entries as unreliable and we don't know about this data. So to avoid errors these values are made zeros.

Opioid Overdose dataset

In this dataset the data regarding Drug/Alcohol Induced Cause Code is not needed to merge the data and it cannot be used to group the data to get the state level data. This issue is rectified by drooping the data from the dataset. To make the merge possible we have renamed the FIPS column name to county code to match the column names.

Drug Overdose Dataset and County Health Rankings dataset have the county as a column and dropping the column of county from opioid rate dataset to avoid duplicates in the data.

1.5 Data Dictionary

All the variables from the data are analysed and a data dictionary is formulated.

The data dictionary is shown below where 10 variables are selected from the provided data sets. These variables are important for any kind of data analysis. The variables county, state , 5 digit FIPS code, county FIPS code are used to identify the data with respect to a state and it's county. The variables population and deaths gives the information about the total population in that county and the total number of deaths that have been reported related to opioid. The variable release year helps us to identify in which year the data is released. The variable opioid dispensing rate gives us the data on the percentage of opioid that is prescribed to the population in that county. The variable county ranked tells us whether the specific county have violated the dispensing of opioid drug to the patients.

Data Dictionary			
S.No	Variable	Measure	Data Type
1	County	Name of the county	object
2	State	State abbreviation	object
3	5-digit fips code	Fips code	Int64
4	Population	Population in the country	Int64
5	Deaths	Deaths in the country	Int64
6	Release Year	Year	Int64
7	Norm_Deaths	Normalized deaths per 100K Population	Float64
8	County Ranked (Yes=1/No=0)	Ranking of county	Float64
9	County FIPS Code	County FIPS Code	Int64
10	Opioid_Dispensing_ Rate	Dispensing rate of opioid	Float64

1.6 Identifying the top counties and states with highest opioid mortality rates

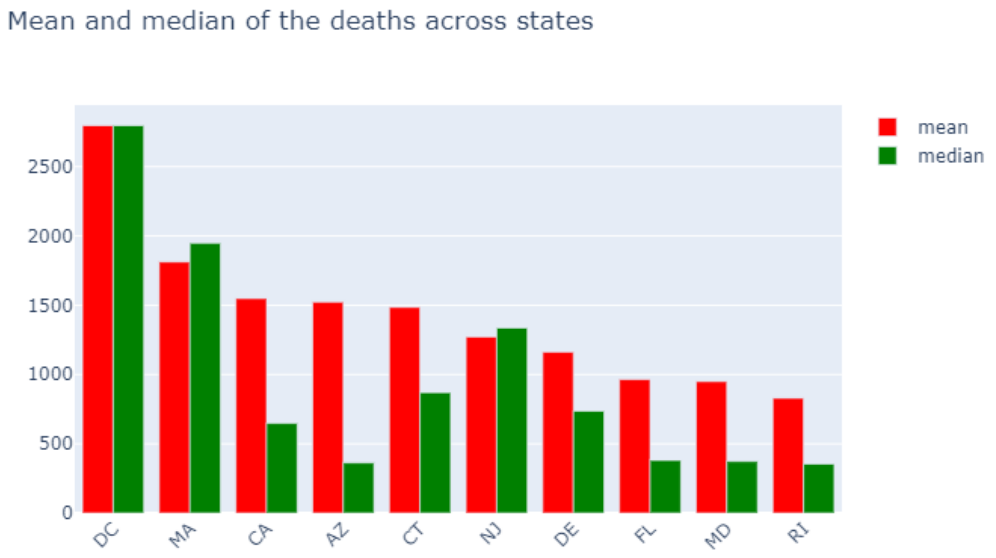


Figure 1.1: Mean and Median of the deaths across different states

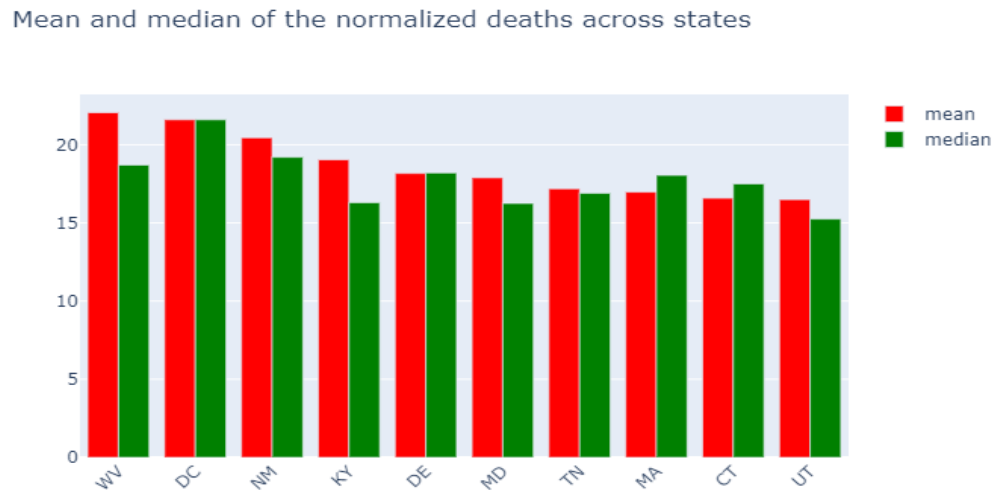


Figure 1.2: Mean and Median of the normalized deaths across different states

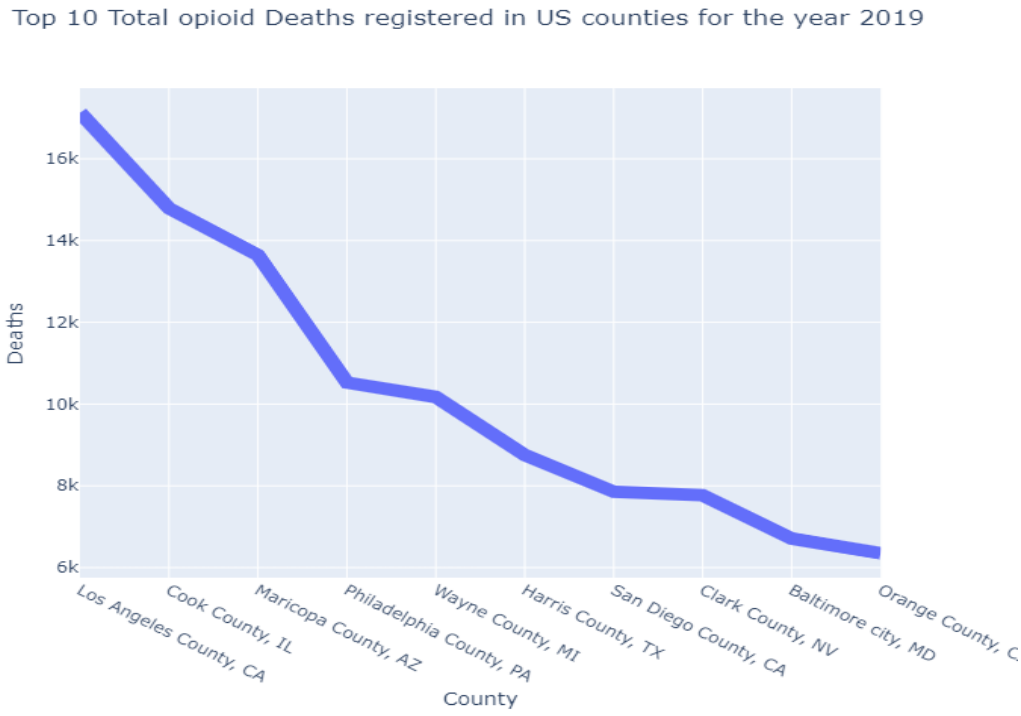


Figure 1.3: Top 10 Total opiopid deaths registered in US counties in 2019

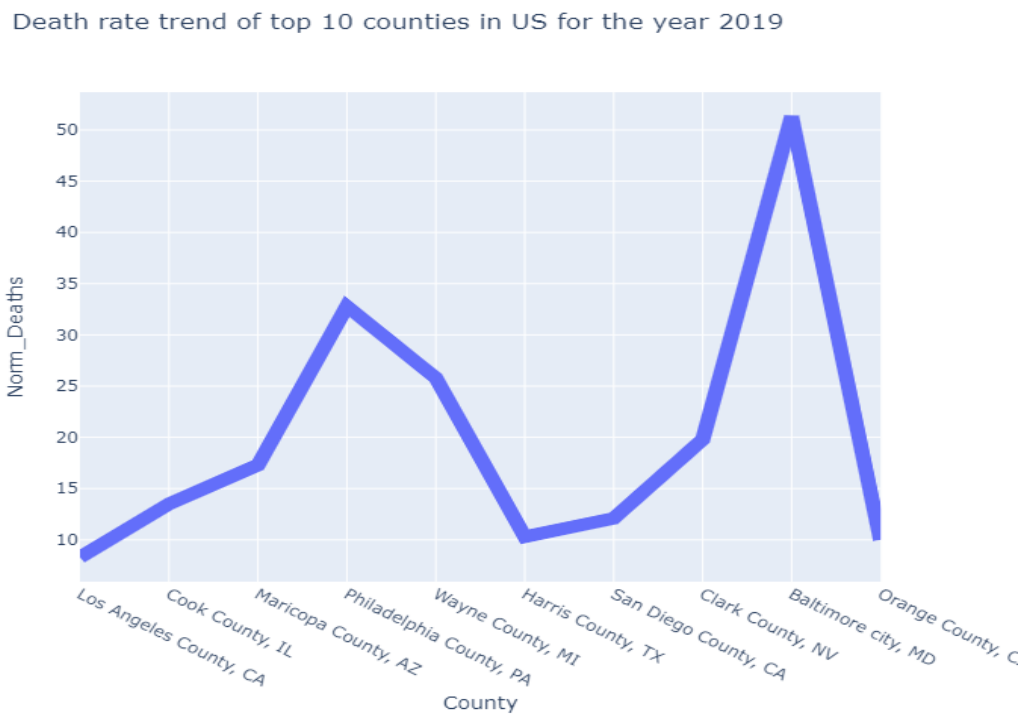


Figure 1.4: Death rate trend in Top 10 counties in US in 2019

1.7 Intution on high rates in these states and counties

- The deaths related to opioid depends on the prescription rate but it is not directly proportional.
- The factors such as socio economic status, per capita income, race and ethinicity, health care systems in the area, prescription rates in the area, addiction rates in the area, demographic location, history of drug usage in the area, etc. may lead to the opioid deaths in US.
- Los Angeles county in california is the county with highest opioid deaths in the country, this can be because of many factors as above.
- Conducting detailed analysis on many other factors can give better results and relation on the death rate.

Chapter 2

STAGE-2 DATA MODELING

2.1 Introduction

In this stage we are developing the data for modeling and comparative analysis. In this we are graphically comparing how different states are doing with respect to opioid mortality rate. County based information for different states in the US is also analysed. We are using a historical dataset here for opioid related mortality from 1999 -2019 for different causes of death[4].

2.2 Identifing peaks in US Opioid Mortality

The deaths rate increased from 21k in 2010 to 48k in 2017, The death rate remained Steady in the following years and a significant increase in the deaths to 68k in the year 2020. In the years from 2018 to 2020 we can see that more number of deaths in states like west virginia, new mexico, kentuky, california and maryland.

If we compare the mean of Normalized Deaths from the Drug Overdose Dataset we can say that states like West Virginia, Arkansas, Mississippi, Alabama, Iowa, Pennsylvania top among all other states in US. If we compare the factors of health care, education and economy the above said states (West Virginia, Arkansas, Mississippi) perform really poor and stand among the last five rankings of all the states in US. With highly poor Health care and education along with highly poor quality of life we can see more people are addicted to opioid drugs as a result more number of deaths in these states.

As already discussed in these states we have poor economic conditions and poor health care more people are under the drug addiction and in the recent years like 2020 more number of drug overdose deaths are seen because of drug addictions leading to unprescribed drug usage especially in the above said states.

We can see that there is definite peak in the deaths in 2020 in comparision with 2019. If we compare the percentage increase of deaths in 2020, there is increase of almost 50% of deaths in the states like west Virginia, Mississippi, Arkansas and california. If we consider the data from 2013-2017 for the deaths, States like West Virginia, Ohio, pensylvania rank among top for percentage increase in deaths when compared to previous years.

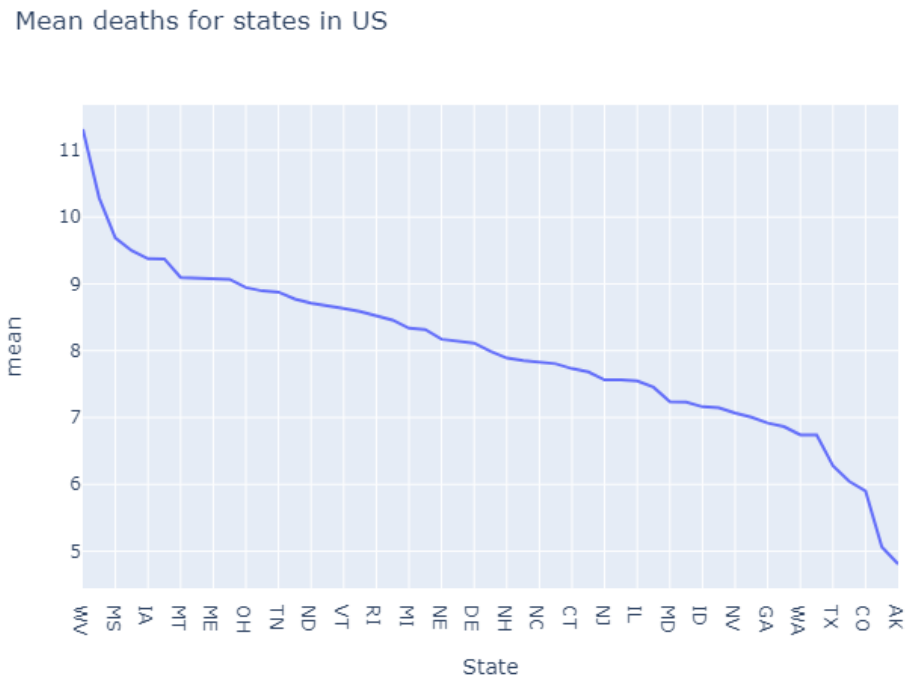


Figure 2.1: Mean Deaths for states in US

2.3 Identifying trends in states

We are considering the data from 2016 to 2020 i.e., 5 years.

When considered the 2019 data with percentage change we can say that AK, SD, MT, ND, NM have the highest percentage change for the year 2019 and MA, VT, MO, DE, DC states have the least perentage change in Norm_Deaths. In the same way if we consider the data from 2016-2020 then states like NY, NJ, LA, NM, MS have the most percentage change in the Norm_Deaths. Similarly, VT, NH, WA, ME, OR have the least percentage change. In the below graph for the up-trend states plot we have a definite rise in Norm_Deaths for these 5 states in the year 2020 when compared to 2016-2019 it is almost same. Similar type of graph for the least percentage change states same trend can be seen as more norm_deaths for the year 2020 in compared to 2016 to 2019.

2.4 Normalized Deaths Vs Log Population

- The Mean opioid normalized deaths for the US is 12.7192 and the Median opioid normalized deaths for the US is 11.7.

- The Normalized deaths V/S Population scatter plot tells us that the normalized

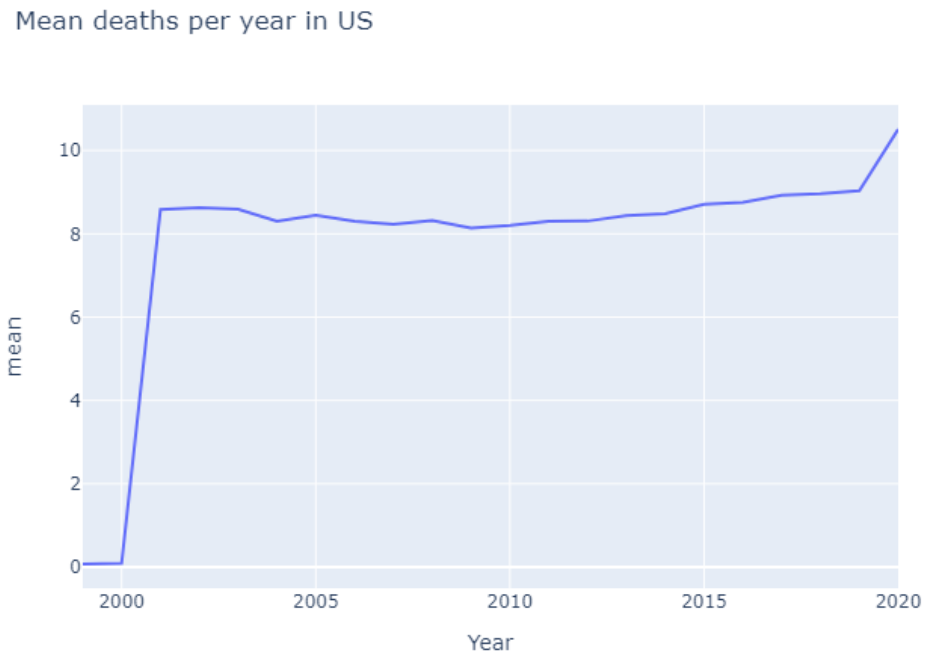


Figure 2.2: Mean Deaths per year in US

deaths are similar for a major portion of population i.e., most of the points are in between 0-30. A few outliers present in the data which might influence the distribution of the data.

- The Normalized deaths V/S Opioid dispensing rate scatter plot tells us that most of the normalized deaths are in between 5-30 for the opioid dispensing rate of 0-100. There are a few outliers in the data and the data is weighted in the bottom left corner of the plot.

- The Normalized deaths V/S County scatter plot shows how the deaths are for a respective county. The plot is scattered a lot because county is a variable which is different every time and we have more than 3000 counties so the Y-axis of the scatter plot has more than 3000 data points to display.

- The Normalized deaths V/S State scatter plot shows how the deaths are for a respective State. Here there are 51 states so there are 51 different data points on the y-axis and the points in the scatter plot is independent with each other as each state is a seperate variable.

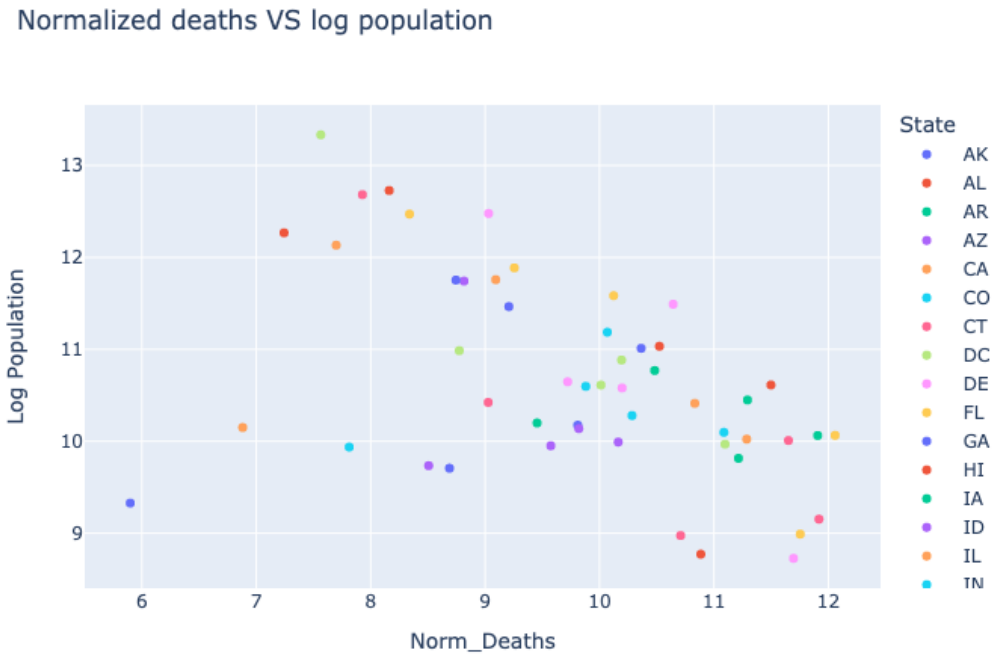


Figure 2.3: Normalized Deaths Vs Log Population

- The Normalized deaths V/S Release Year scatter plot shows how the deaths are for a respective year. Here the year taken for analysis is 2019 so the plot has all the points for year 2019 and so is a horizontal dotted line plot is obtained.
- The Normalized deaths V/S 5-digit fips code scatter plot shows how the deaths are for a respective fips. This is helpfull in estimating which fips codes are having more deaths and can extract insights on the deaths for a group of fips codes from this plot. The plot have a straight lined plots because the y-axis is scaled to show all the fips codes in a small plot.
- The Normalized deaths V/S Normalized deaths scatter plot shows the npormalized deaths for normalized deaths. The plot is between the same variable and hence the data points are placed as a linear dotted line increasing evenly.

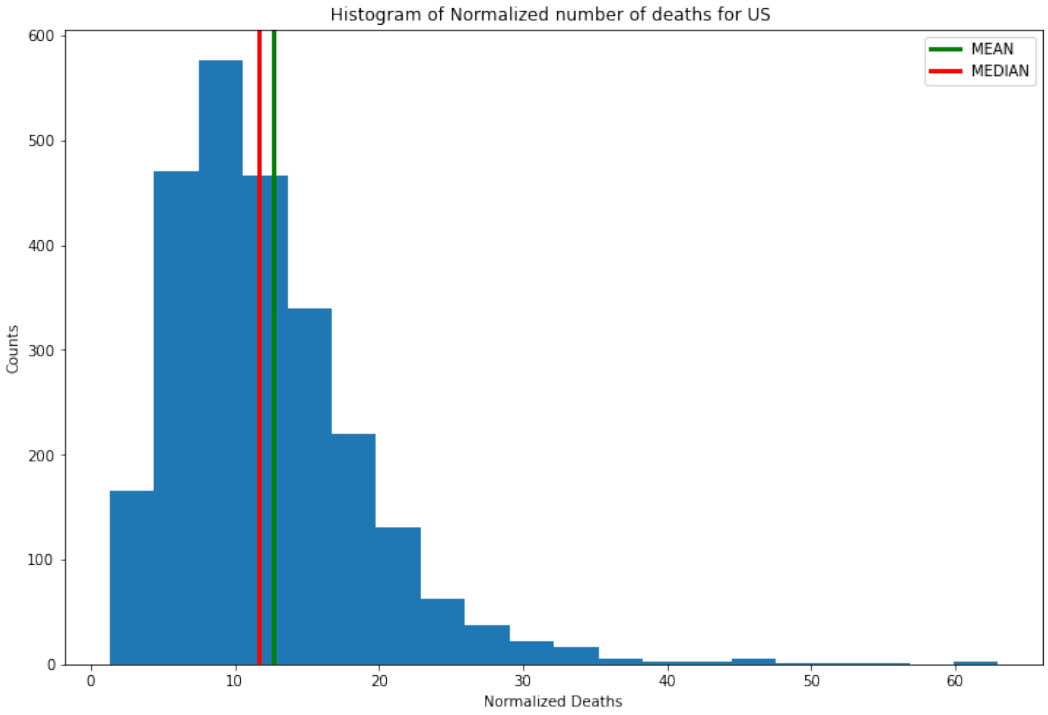


Figure 2.4: Histogram Of Normalized Deaths Variable for entire US

Chapter 3

STAGE-3 DISTRIBUTIONS AND HYPOTHESIS TESTING

3.1 Distribution Analysis

In this we compare NC and KY states opioid mortality rate on 2019 data and created Histogram for the same. Both the histograms were merged on to a single graph along with mean lines for the histograms as given in the figure 3.1.

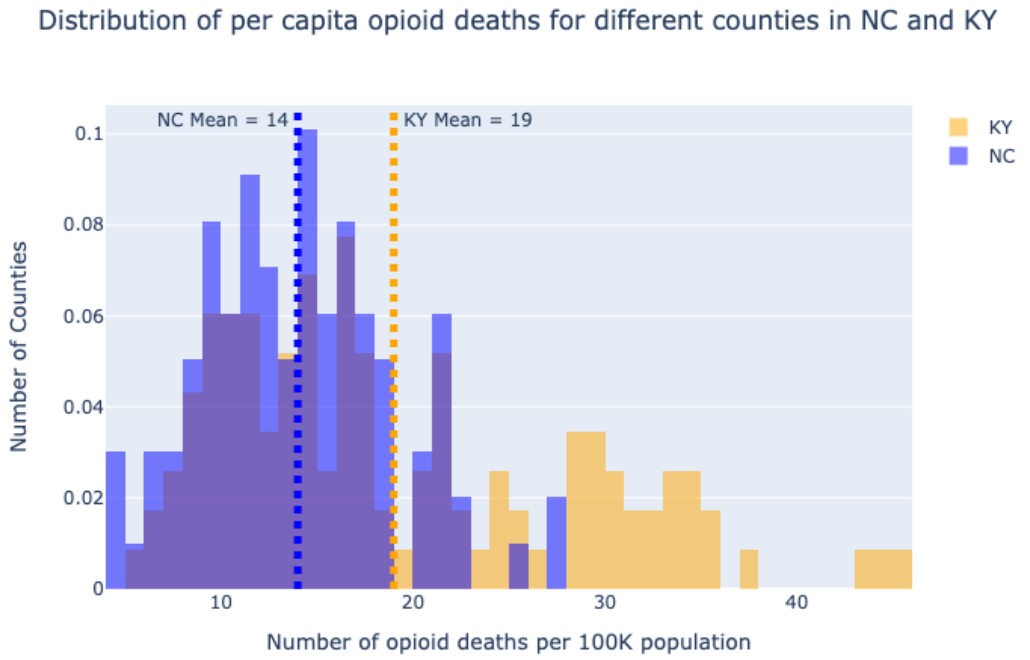


Figure 3.1: Histogram Of Normalized Deaths Variable for KY and NC

Evaluation of distribution for the opioid mortality rate for 100K population. We choose the normal distribution for the analysis as the mean values of the distribution are floating values. We developed distribution estimators with MOM, MLE and KDE plots for the normal distribution which we consider the best fit curve will be the mean of distribution. So MLE and MOM will be the same for the distribution. The plot for the most likelihood estimator probability density function is shown in the figure 3.2.

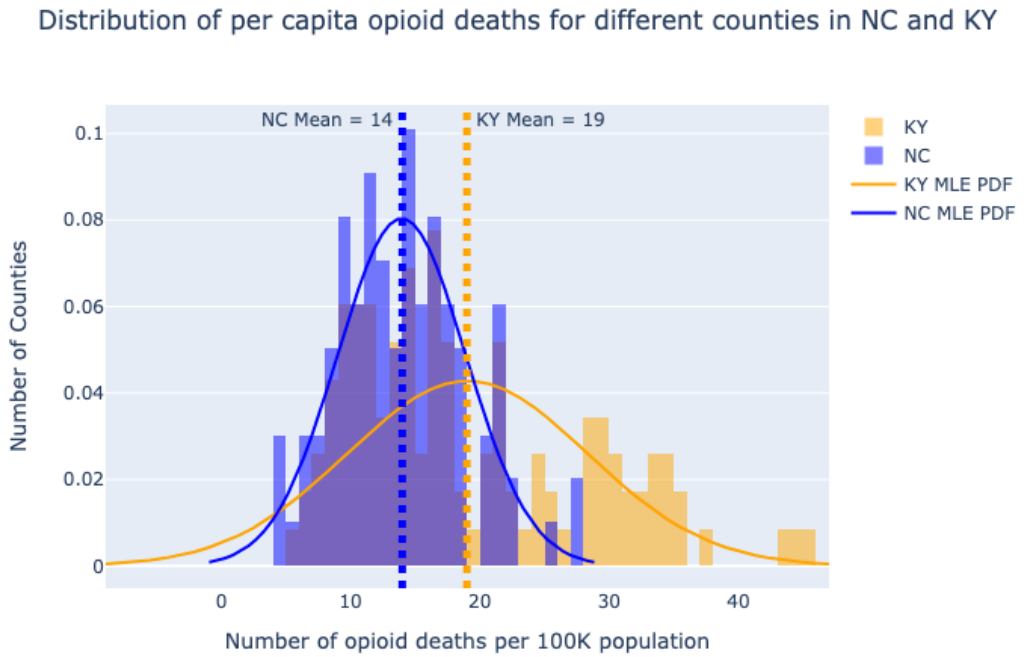


Figure 3.2: Histogram Of Normalized Deaths Variable with MLE for KY and NC states

KDE plot for the states NC and KY are shown over the histogram plot as of figure 3.3.

All the above said process was replicated for the top two states i.e., WE and NM for highest mortality rate.

3.2 Hypothesis Testing

From stage 1 we have identified 5 variables to formulate and test the hypothesis of these variables. The variables are opioid dispensing rates, Life Expectancy Raw Value, Drug Overdose Raw Value, Primary Care Physicians Raw Value and Uninsured raw value.

For each variable, dividing the observations into two categories, high and low by considering the mean value of the variable. In this low is labelled as '0' and high is labelled as '1'.

So for the observed categories we formulate a hypothesis test.

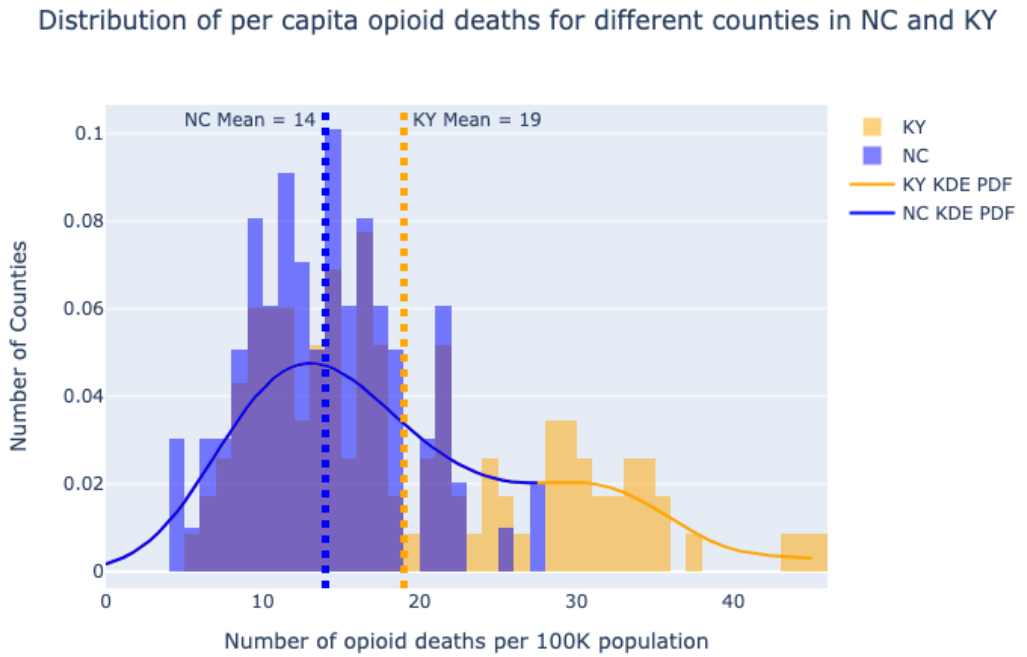


Figure 3.3: Histogram Of Normalized Deaths Variable with KDE for KY and NC states

3.3 Two Sample T test

A two sample T-test is selected for the hypothesis testing. A two-sample t-test investigates whether the means of two independent data samples differ from one another. Null hypothesis selected as, The mean values of both the considered groups (Divided each observation into low and high) are same. Alternate hypothesis can be considered as the mean values for both the groups are not similar.

The confidence interval is taken as 95%. If 'p' value greater than 0.05 then null hypothesis is failed to reject and we can say their is a relation between the mean values of the both mean values of both the considered groups are almost similar when the 'p' value greater than 0.05 then the test will fail to reject.

For variable 1 (Opioid dispensing rate), obtained 'p' value is almost equal to zero, so we reject the null hypothesis. For variable 2 (life expectancy raw value) 'p' value is almost zero so we reject the null hypothesis for variable 2. For variable 3 (Drug overdose deaths raw value) 'p' value is 0.27 which is greater than 0.05 so we fail to reject the null hypothesis for variable 3. For variable 4 (Primary Care Physicians Raw Value) 'p' value is greater than 0.05 which is 0.07 so we fail to reject the null

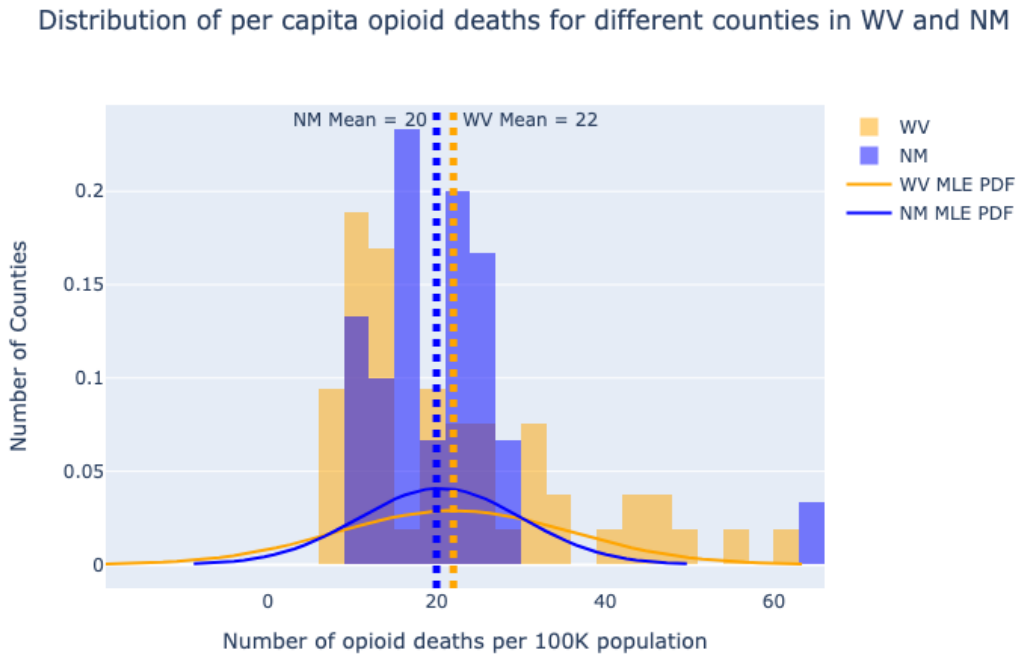


Figure 3.4: Histogram Of Normalized Deaths Variable with MLE for WV and NM states

hypothesis for variable 4. For variable 5 (Uninsured raw value) 'p' value is greater than 0.05 which is 0.97 so we fail to reject the null hypothesis for variable 5.

3.4 Regression

Linear Regression

Linear regression is performed between opioid mortality rate for 100K population and opioid dispensing rate shown in the figure 3.5. Trace 0 is the data points of the variables and trace 1 is the trend line.

Multi Linear regression

A multi linear regression is performed with the 5 variables considered above and the opioid dispensing rate. Following are the model parameters obtained for the multi linear regression model.

- Intercept - 26.982816
- Life expectancy raw value - -0.003363

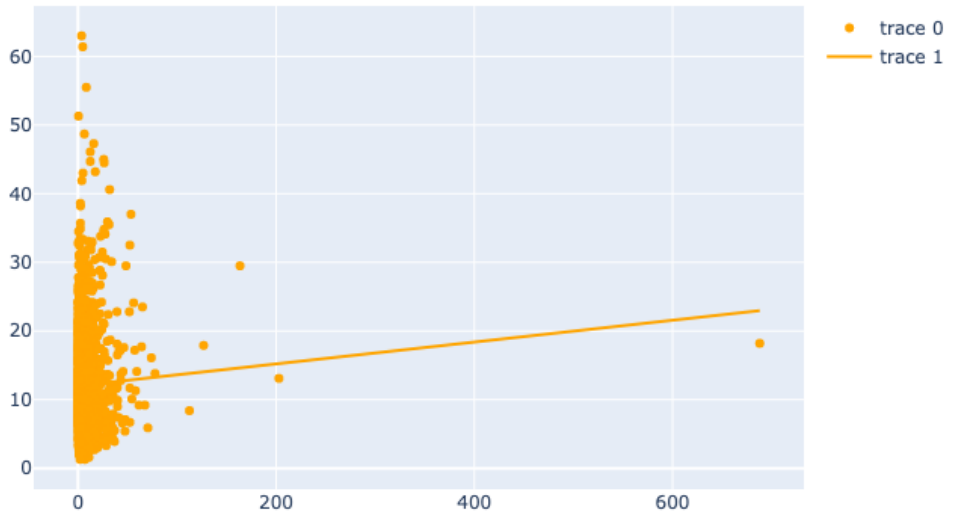


Figure 3.5: Linear regression between opioid mortality rate for 100K population and opioid dispensing rate

- Primary care physicians raw value - 187.839972
- Drug overdose deaths raw value - 0.095825
- Uninsured raw value - 17.174813
- Mental health providers raw value - -653.297929
- Norm Deaths - 1.269584

Non - Linear regression

A polynomial regression was performed for the opioid dispensing rate and the normalized deaths for n=1,2,3,4.

Performed a non linear regression model with 5+1 variables. The results were evaluated and discussed.

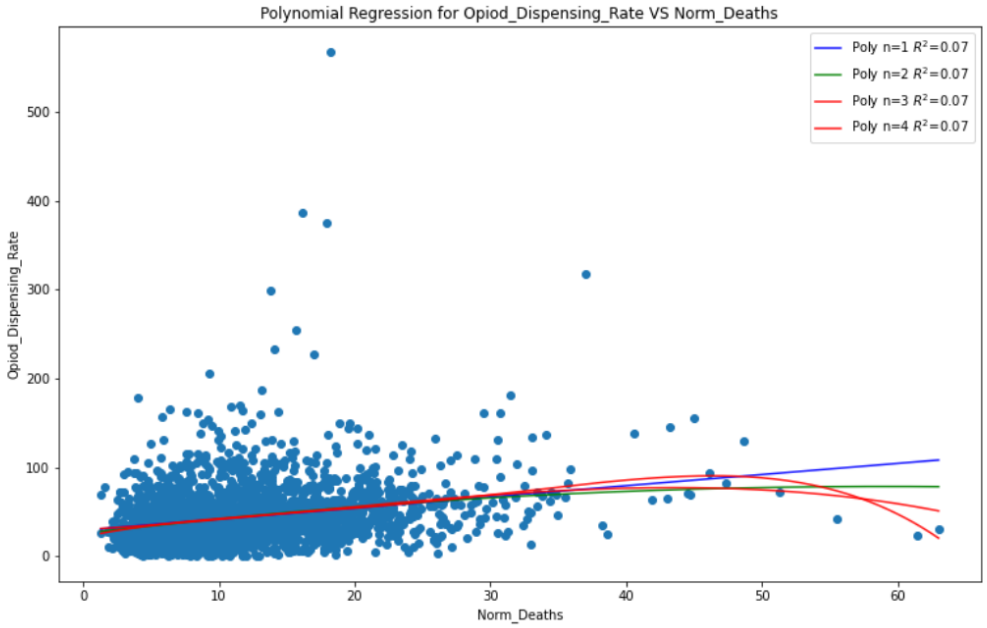


Figure 3.6: Non linear regression for opioid dispensing rate and Normalized deaths (n=1,2,3,4)

Chapter 4

STAGE-4 DASHBOARD

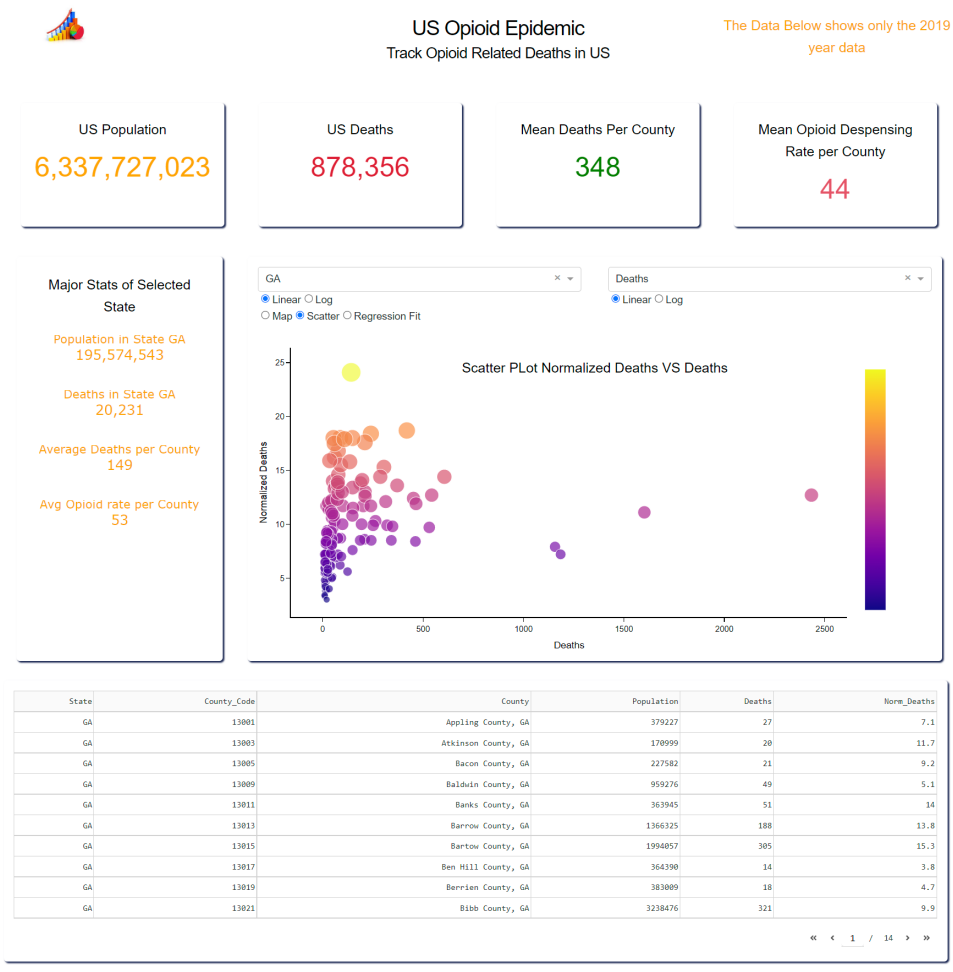


Figure 4.1: Dashboard

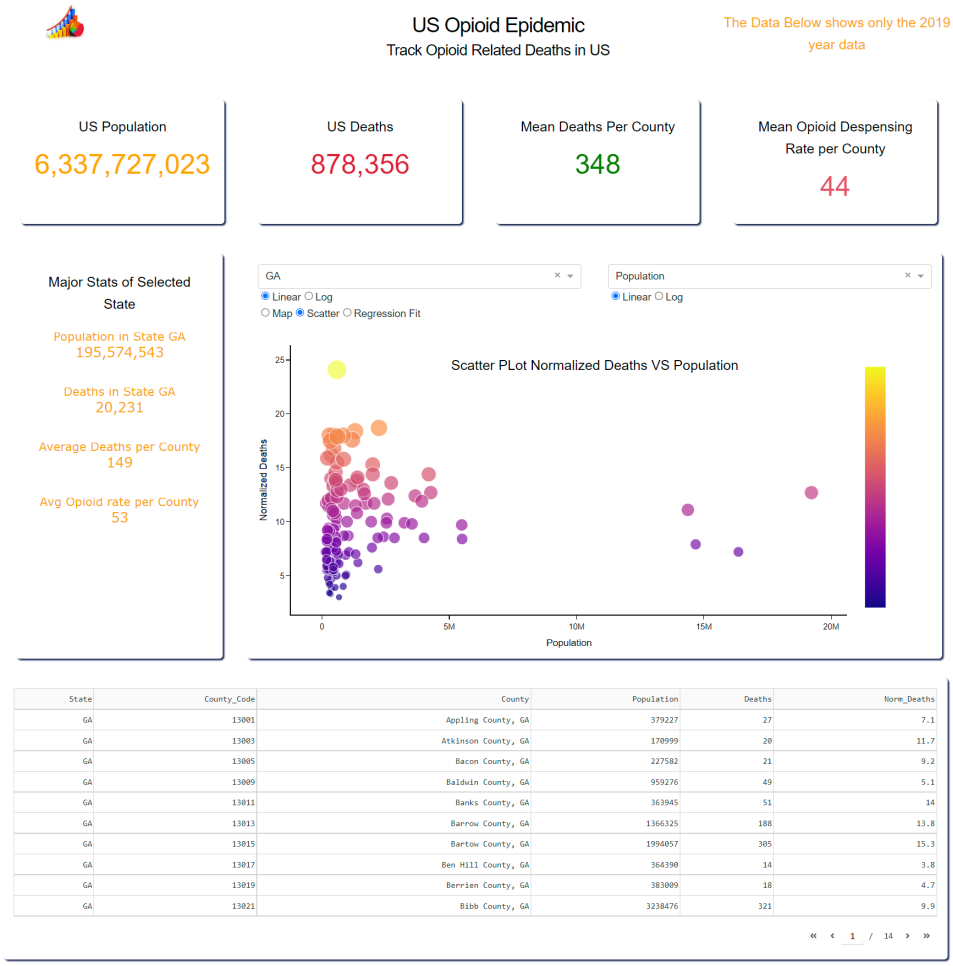
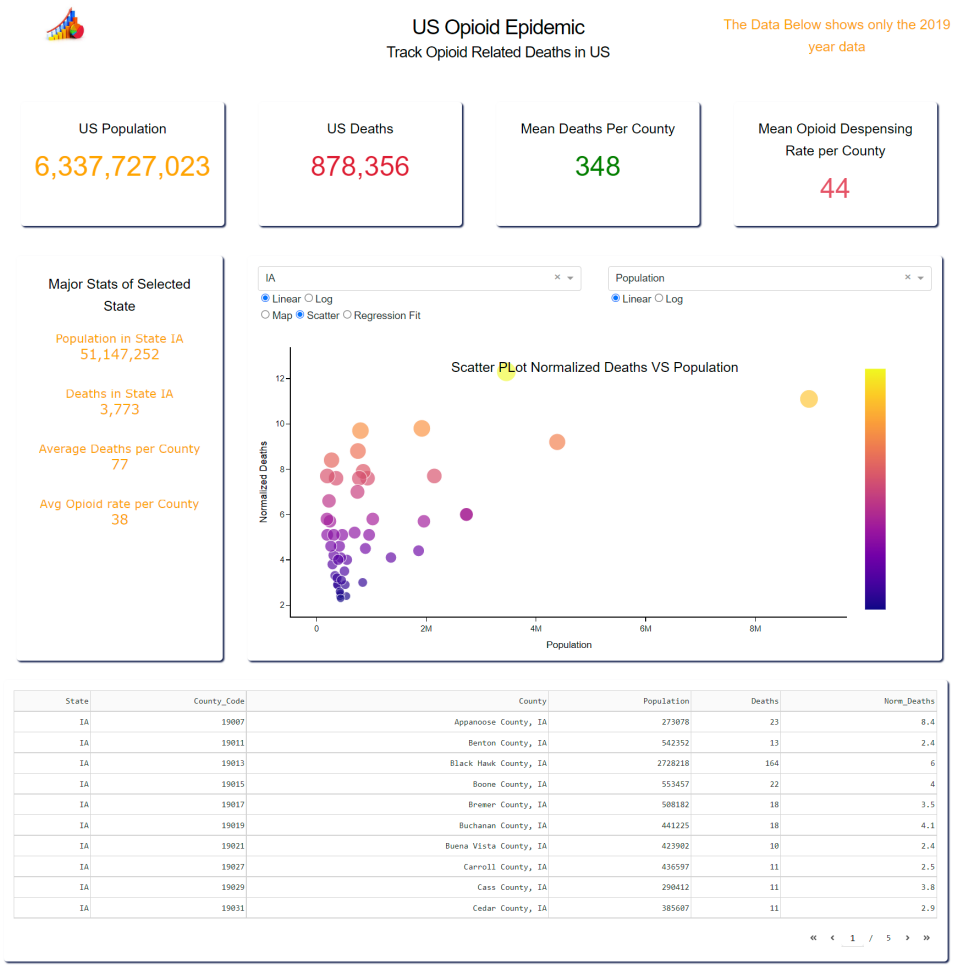


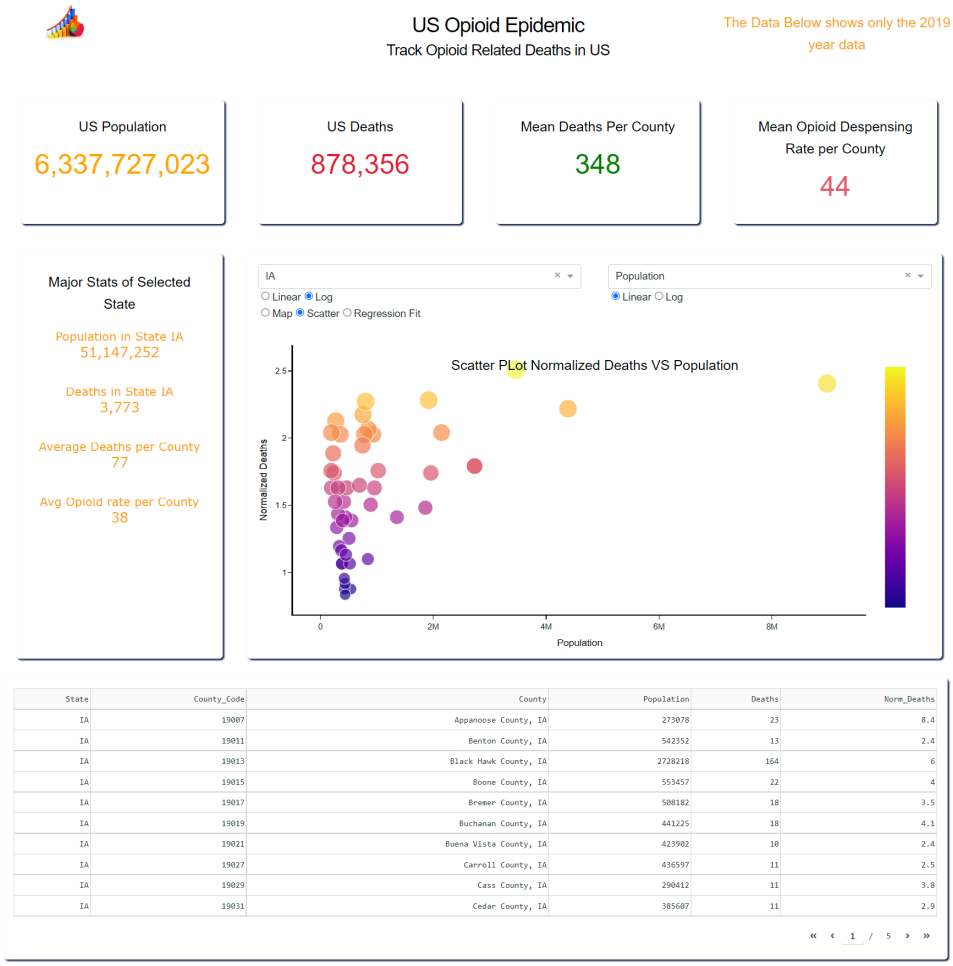
Figure 4.2: Scatter plot updation for the selected variable



State	County_Code	County	Population	Deaths	Norm_Deaths
IA	19007	Appanoose County, IA	273078	23	8.4
IA	19011	Benton County, IA	542352	13	2.4
IA	19013	Black Hawk County, IA	2728218	164	6
IA	19015	Boone County, IA	553457	22	4
IA	19017	Bremer County, IA	508182	18	3.5
IA	19019	Buchanan County, IA	441225	18	4.1
IA	19021	Buena Vista County, IA	423902	10	2.4
IA	19027	Carroll County, IA	436597	11	2.5
IA	19029	Cass County, IA	290412	11	3.8
IA	19031	Cedar County, IA	385607	11	2.9

« ‹ 1 / 5 › »

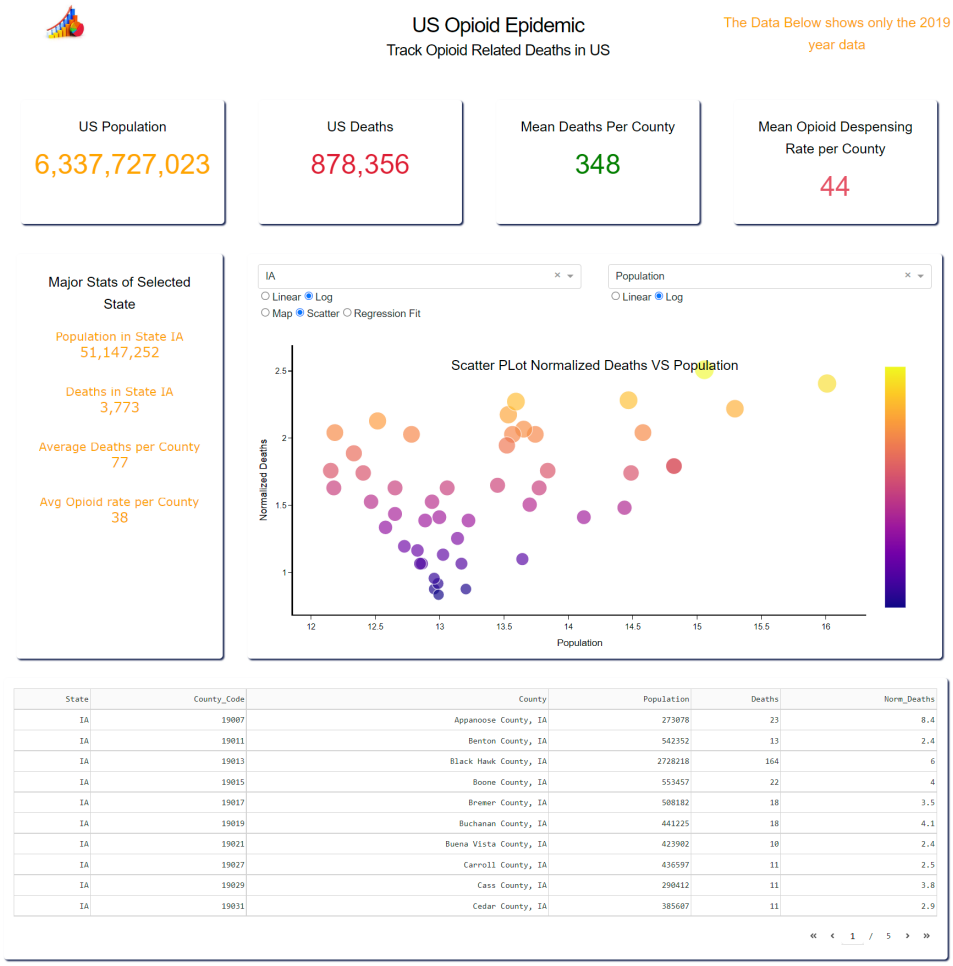
Figure 4.3: Showing the Scatter plot and the data table for the selection



State	County_Code	County	Population	Deaths	Norm_Deaths
IA	19007	Appanoose County, IA	273078	23	8.4
IA	19011	Benton County, IA	542352	13	2.4
IA	19013	Black Hawk County, IA	2728218	164	6
IA	19015	Boone County, IA	553457	22	4
IA	19017	Bremer County, IA	508182	18	3.5
IA	19019	Buchanan County, IA	441225	18	4.1
IA	19021	Buena Vista County, IA	423902	10	2.4
IA	19027	Carroll County, IA	436597	11	2.5
IA	19029	Cass County, IA	290412	11	3.8
IA	19031	Cedar County, IA	385607	11	2.9

« ‹ 1 / 5 › »

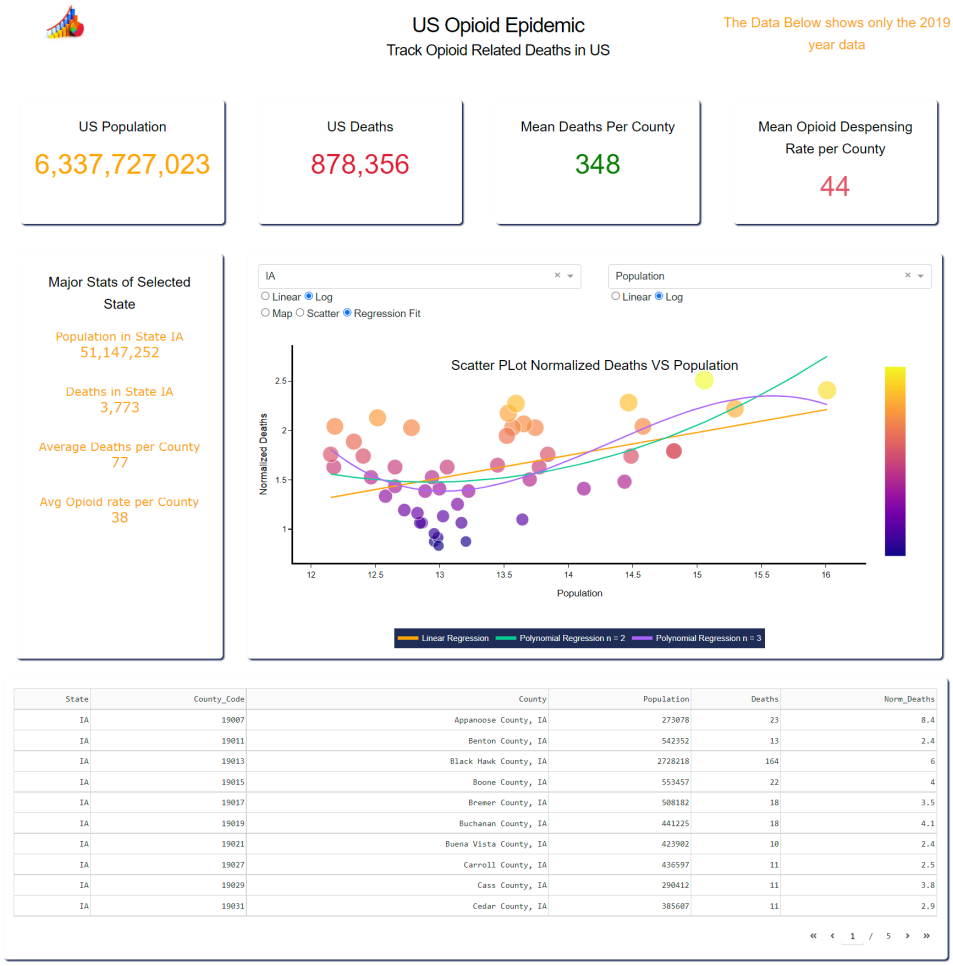
Figure 4.4: Showing the Scatter plot for the log of Y-axis



State	County_Code	County	Population	Deaths	Norm_Deaths
IA	19007	Appanoose County, IA	273078	23	8.4
IA	19011	Benton County, IA	542352	13	2.4
IA	19013	Black Hawk County, IA	2728218	164	6
IA	19015	Boone County, IA	553457	22	4
IA	19017	Bremer County, IA	508182	18	3.5
IA	19019	Buchanan County, IA	441225	18	4.1
IA	19021	Buena Vista County, IA	423902	10	2.4
IA	19027	Carroll County, IA	436597	11	2.5
IA	19029	Cass County, IA	290412	11	3.8
IA	19031	Cedar County, IA	385607	11	2.9

« ‹ 1 / 5 › »

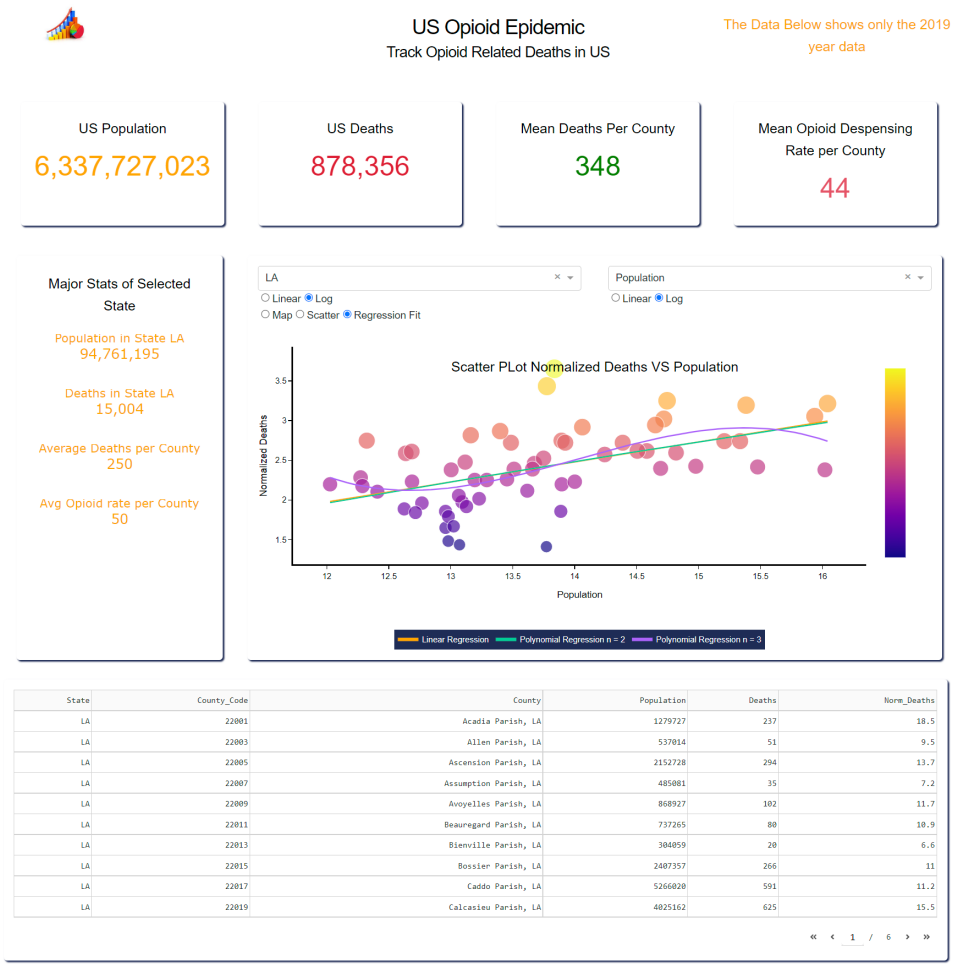
Figure 4.5: Showing the Scatter plot for the log of X-axis



State	County_Code	County	Population	Deaths	Norm_Deaths
IA	19007	Appanoose County, IA	273078	23	8.4
IA	19011	Benton County, IA	542352	13	2.4
IA	19013	Black Hawk County, IA	2728218	164	6
IA	19015	Boone County, IA	553457	22	4
IA	19017	Bremer County, IA	508182	18	3.5
IA	19019	Buchanan County, IA	441225	18	4.1
IA	19021	Buena Vista County, IA	423902	10	2.4
IA	19027	Carroll County, IA	436597	11	2.5
IA	19029	Cass County, IA	290412	11	3.8
IA	19031	Cedar County, IA	385607	11	2.9

« ‹ 1 / 5 › »

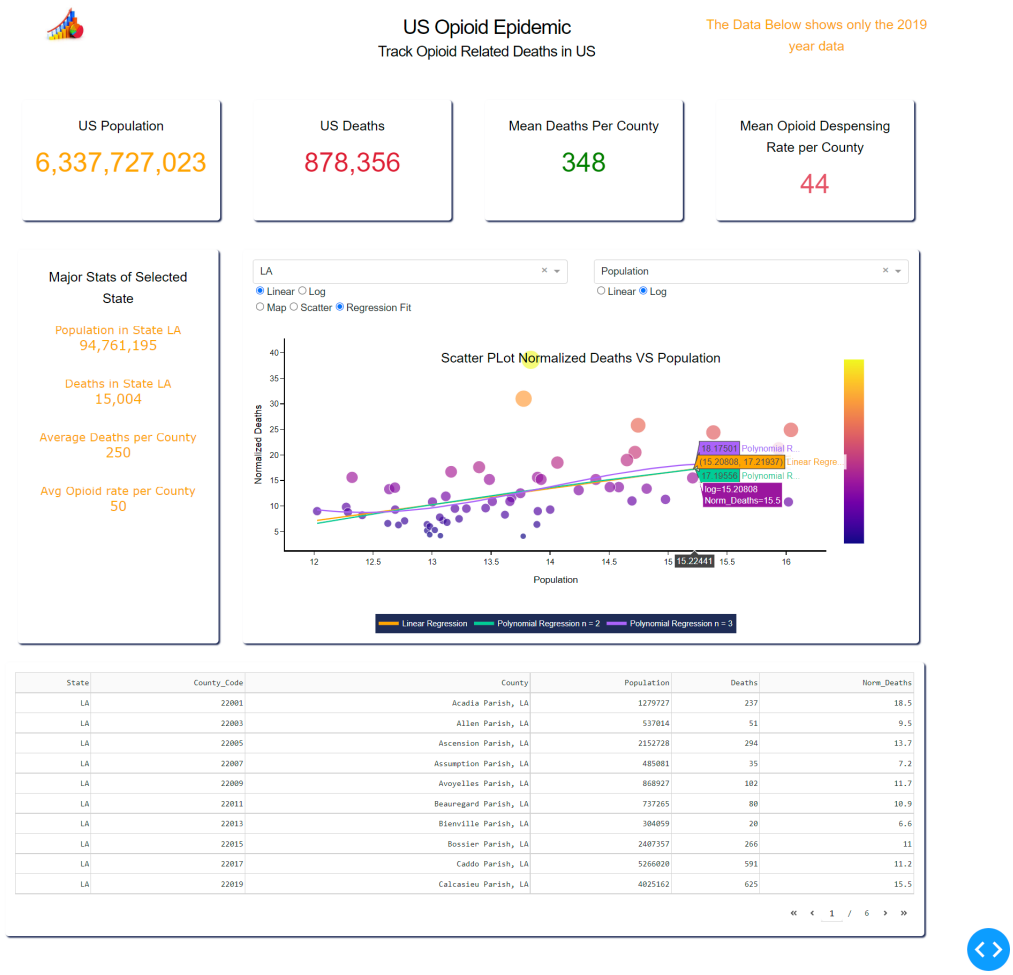
Figure 4.6: Showing the Scatter plot with linear and polynomial regression (n=2,3)



State	County_Code	County	Population	Deaths	Norm_Deaths
LA	22001	Acadia Parish, LA	1279727	237	18.5
LA	22003	Allen Parish, LA	537014	51	9.5
LA	22005	Ascension Parish, LA	2152728	294	13.7
LA	22007	Assumption Parish, LA	485081	35	7.2
LA	22009	Avoyelles Parish, LA	868927	102	11.7
LA	22011	Beauregard Parish, LA	737265	80	10.9
LA	22013	Bienville Parish, LA	304059	20	6.6
LA	22015	Bossier Parish, LA	2407357	266	11
LA	22017	Caddo Parish, LA	5266020	591	11.2
LA	22019	Calcasieu Parish, LA	4025162	625	15.5

« ‹ 1 / 6 › »

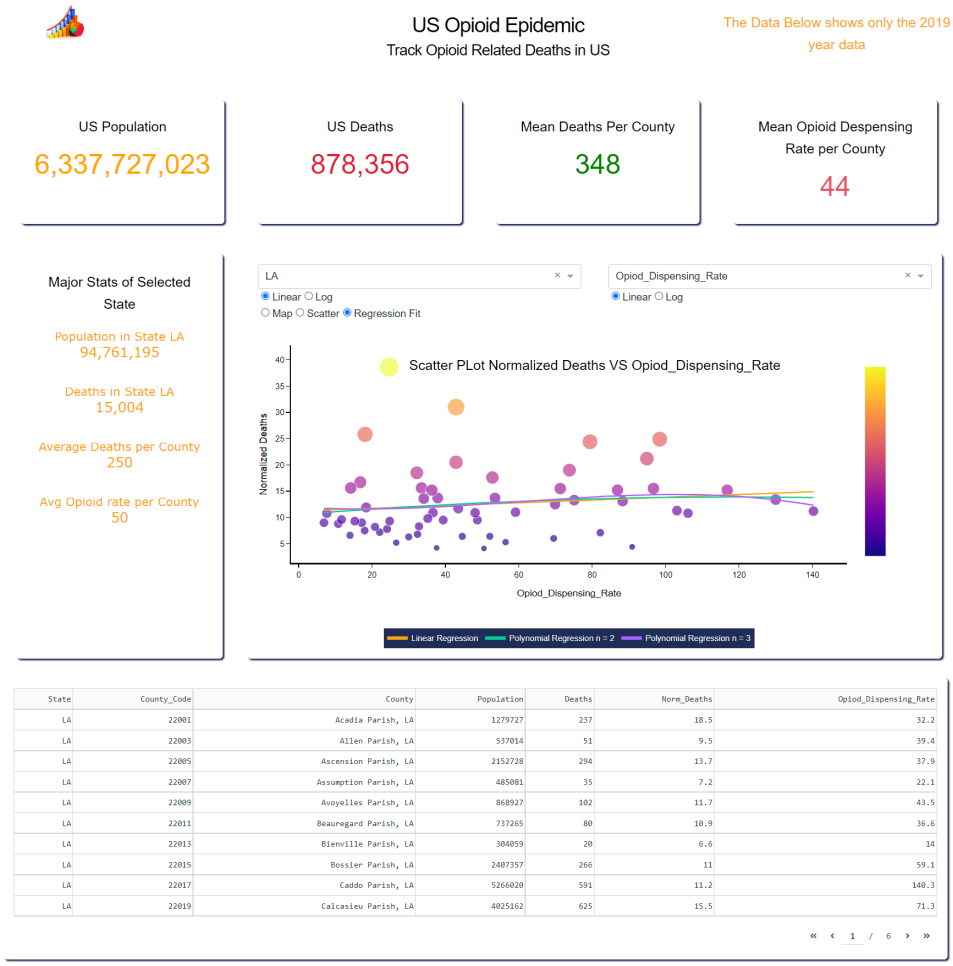
Figure 4.7: Showing the Scatter plot with linear and polynomial regression (n=2,3) for the selected state



State	County_Code	County	Population	Deaths	Norm_Deaths
LA	22001	Acadia Parish, LA	1279727	237	18.5
LA	22003	Allen Parish, LA	537014	51	9.5
LA	22005	Ascension Parish, LA	2152728	294	13.7
LA	22007	Assumption Parish, LA	485081	35	7.2
LA	22009	Avoyelles Parish, LA	868927	102	11.7
LA	22011	Beauregard Parish, LA	737265	80	10.9
LA	22013	Bienville Parish, LA	304059	20	6.6
LA	22015	Bossier Parish, LA	2407357	266	11
LA	22017	Caddo Parish, LA	5266020	591	11.2
LA	22019	Calcasieu Parish, LA	4025162	625	15.5

« ‹ 1 / 6 › »

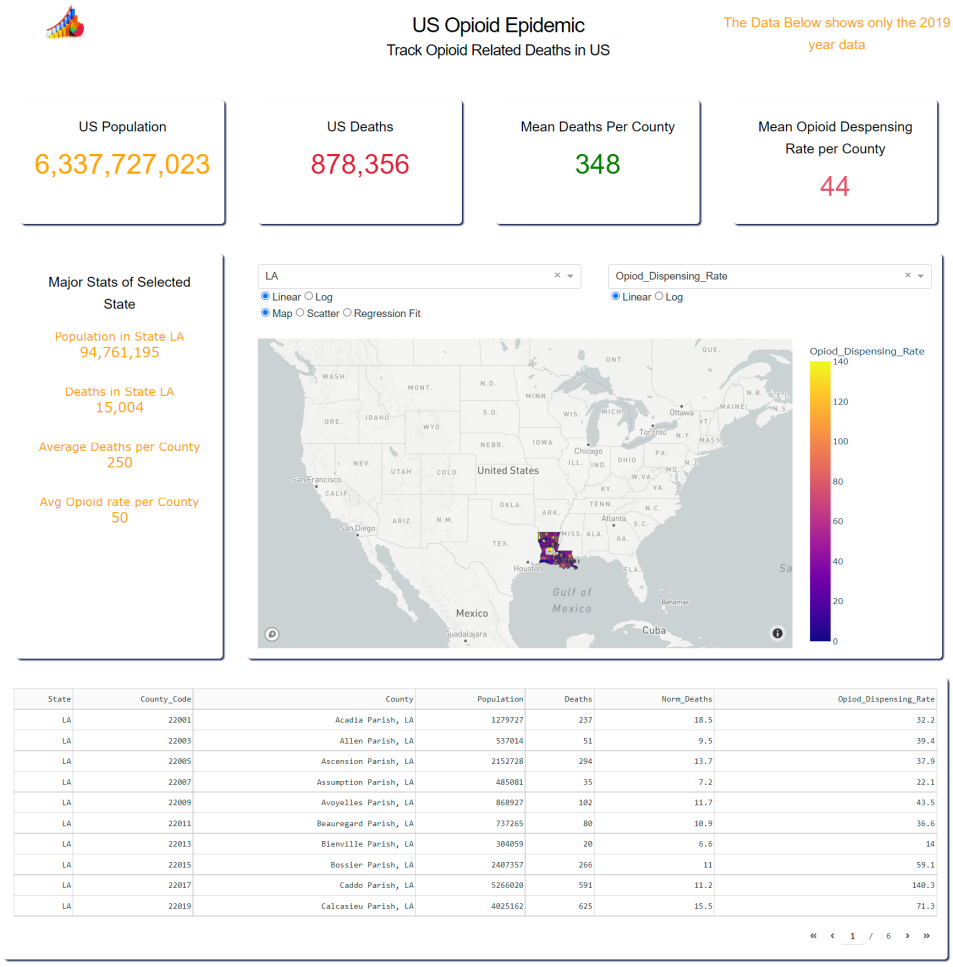
Figure 4.8: Showing the Scatter plot with linear and polynomial regression (n=2,3) for the selected state for linear scale



State	County_Code	County	Population	Deaths	Norm_Deaths	Opiod_Dispensing_Rate
LA	22001	Acadia Parish, LA	1279727	237	18.5	32.2
LA	22003	Allen Parish, LA	537014	51	9.5	39.4
LA	22005	Ascension Parish, LA	2152728	294	13.7	37.9
LA	22007	Assumption Parish, LA	485081	35	7.2	22.1
LA	22009	Avoyelles Parish, LA	868927	102	11.7	43.5
LA	22011	Beauregard Parish, LA	737265	80	10.9	36.6
LA	22013	Bienville Parish, LA	304059	20	6.6	14
LA	22015	Bossier Parish, LA	2407357	266	11	59.1
LA	22017	Caddo Parish, LA	5266020	591	11.2	140.3
LA	22019	Calcasieu Parish, LA	4025162	625	15.5	71.3

« ‹ 1 / 6 › »

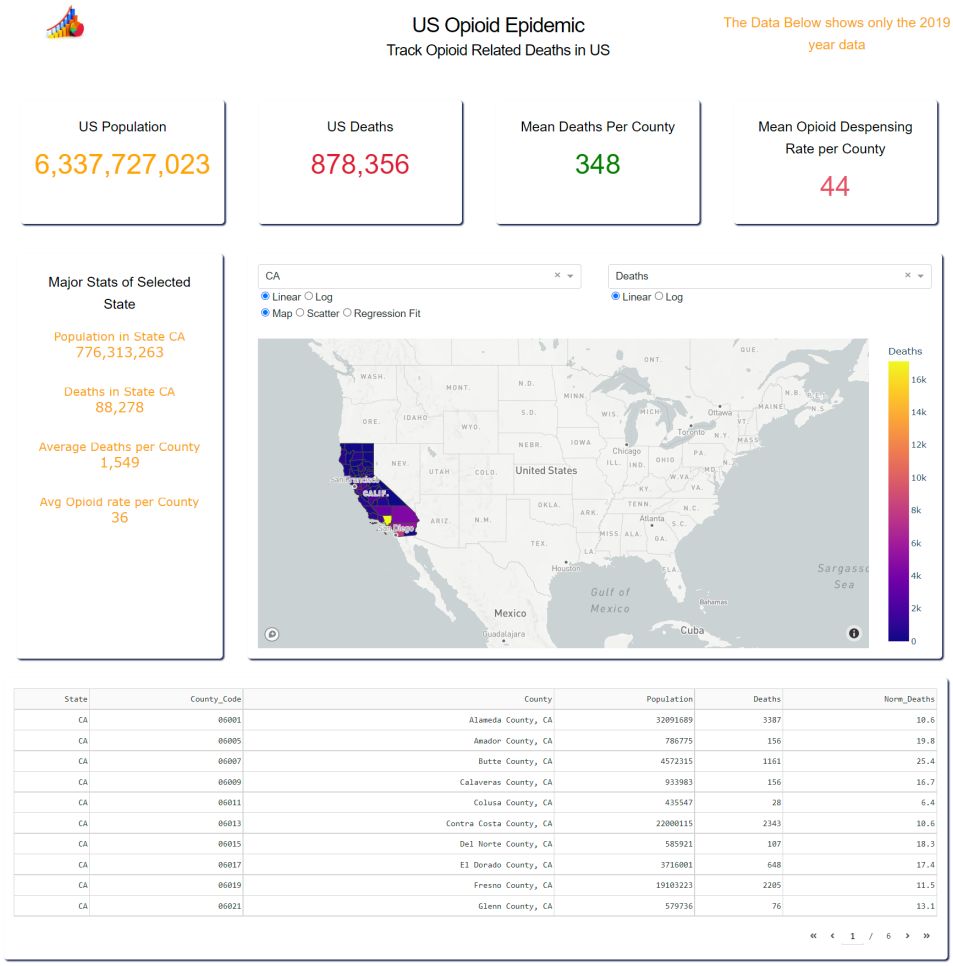
Figure 4.9: Another instance of dashboard



State	County_Code	County	Population	Deaths	Norm_Deaths	Opiod_Dispensing_Rate
LA	22001	Acadia Parish, LA	1279727	237	18.5	32.2
LA	22003	Allen Parish, LA	537014	51	9.5	39.4
LA	22005	Ascension Parish, LA	2152728	294	13.7	37.9
LA	22007	Assumption Parish, LA	485081	35	7.2	22.1
LA	22009	Avoyelles Parish, LA	868927	102	11.7	43.5
LA	22011	Beauregard Parish, LA	737265	80	10.9	36.6
LA	22013	Bienville Parish, LA	304059	20	6.6	14
LA	22015	Bossier Parish, LA	2407357	266	11	59.1
LA	22017	Caddo Parish, LA	5266020	591	11.2	140.3
LA	22019	Calcasieu Parish, LA	4025162	625	15.5	71.3

« ‹ 1 / 6 › »

Figure 4.10: Showing the Choropleth map of USA for the selected variables



State	County_Code	County	Population	Deaths	Norm_Deaths
CA	06001	Alameda County, CA	32091689	3387	10.6
CA	06005	Amador County, CA	786775	156	19.8
CA	06007	Butte County, CA	4572315	1161	25.4
CA	06009	Calaveras County, CA	933983	156	16.7
CA	06011	Colusa County, CA	435547	28	6.4
CA	06013	Contra Costa County, CA	22000115	2343	10.6
CA	06015	Del Norte County, CA	585921	107	18.3
CA	06017	El Dorado County, CA	3716001	648	17.4
CA	06019	Fresno County, CA	19103223	2205	11.5
CA	06021	Glenn County, CA	579736	76	13.1

« ‹ 1 / 6 › »

Figure 4.11: Showing the Choropleth map of USA for California state

BIBLIOGRAPHY

[1] Drug Overdose Dataset, https://wonder.cdc.gov/ucd-icd10.html

[2] County Health Rankings, https://www.countyhealthrankings.org/

[3] County Opioid Dispensing Rate, https://www.cdc.gov/drugoverdose/maps/rxcounty2019.html

[4] opioid related mortality from 1999 -2019 for different causes of death, https://wonder.cdc.gov/wonder/help/ucd.html#Drug/Alcohol%20Induced%20Causes