Group 10 Final Report

I. INTRODUCTION AND PREPROCESSING

This analysis report aims to provide insights into the drug overdose dataset. The dataset was extracted, cleaned, and preprocessed, and initial data exploration and visualizations were conducted. The analysis was performed on various attributes such as gender, age, location, cause of deaths, city, county, race, and ethnicity.

The dataset was loaded as a Pandas DataFrame and preprocessed by converting the 'Date' attribute to a form that is processable by Pandas.

```
df['Date']=pd.to_datetime(df['Date'])
```

The Residence City Geo attribute was split into two sub-attributes, the state of the city and the coordinates.

```
df['RCG']=df['ResidenceCityGeo'].str.split(',').str[0
]

df[['Residence        City','RCG']].loc[(df['Residence
City']!=df['RCG'])  &  (df['Residence City'].notnull())
& (df['RCG'].notnull())].sum()
```

The same technique was employed with 'InjuryCity' and 'DeathCity' attributes, and the DataFrame was updated. Subsequently, irrelevant columns like RCG, ICG, and DCG were dropped. The 'Y' and 'Nan' values were replaced with 1s & 0s respectively for easier data manipulation and analysis.

```
for i in drugs:
   df[i] = df[i].replace(['Y',np.nan],[1,0])
```

The analysis was performed using bar graphs (both vertical and horizontal) and heatmaps. The number of deaths was grouped by gender, drug-wise analysis, age, location, city, county, race, and ethnicity.

II. EXPLORATORY DATA ANALYSIS (EDA)

The analysis showed that men are more susceptible to deaths by drug overdose than women as shown in Figure 2.1.

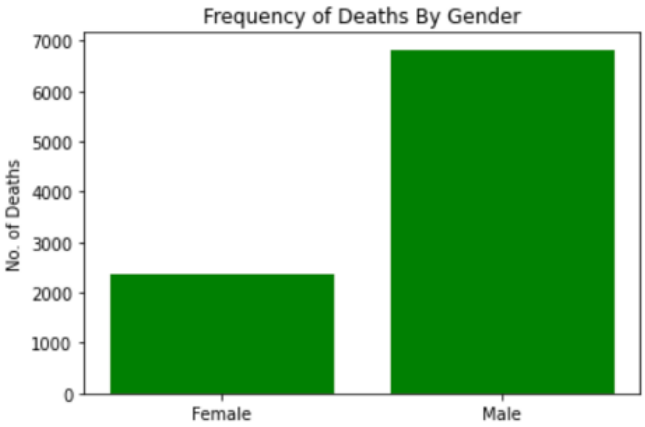


Figure 2.1: Frequency of Deaths by Gender

Our analysis showed that Fentanyl was found to be the highest cause of deaths in both genders, followed by opioids. Heroin was the third highest cause of death in males, while Benzodiazepine was the third highest cause of death in females. This is visualized in Figure 2.2.

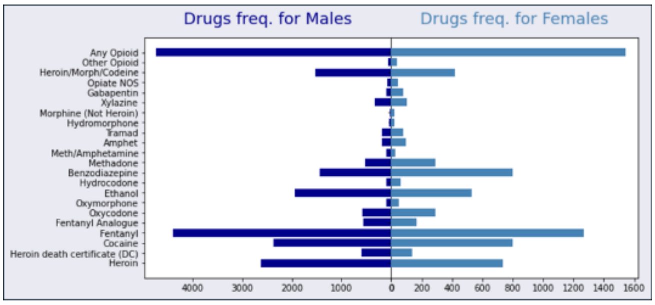


Figure 2.2: Drugs Frequency of Males and Females

Figure 2.3 shows that the highest number of deaths took place within the age group of 30-36, followed by the 50-54 group.

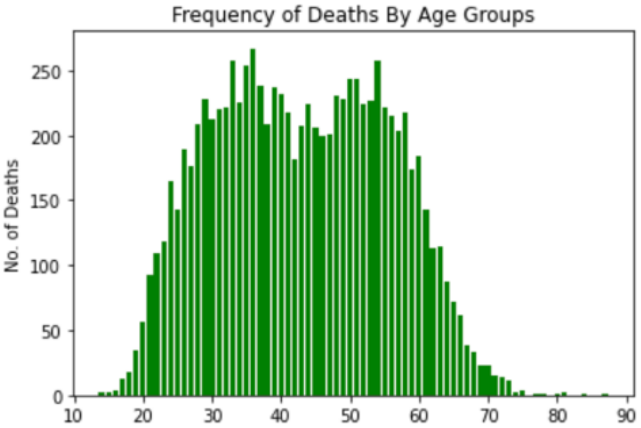


Figure 2.3: Frequency of Deaths by Age Groups

Next, the highest frequency of deaths took place in the victims' residences, followed by hospitals as visualized in Figure 2.4.

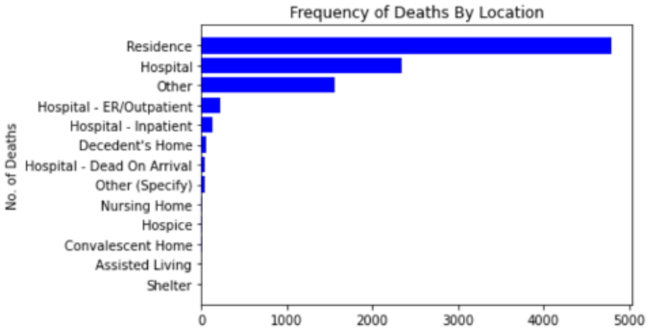


Figure 2.4: Frequency of Deaths by Location

In order to highlight the frequency of deaths by city, we first extracted a list of causes of death in the city, and then counted it for each city, forming a table for it. We then used this table to plot a 'Frequency of Deaths by City' bar graph.

```
my_list = list(df)

city_drug = [my_list[6]] + my_list[22:44]
```

```
city_drug_df = df.groupby('Residence
City').size().reset_index(name='Count')
city_drug_df =
city_drug_df.sort_values(by=['Count'],ascending=True)
```

Hartford was found to be the leading city of deaths, followed by Waterbury, New Haven, and others as shown in Figure 2.5.

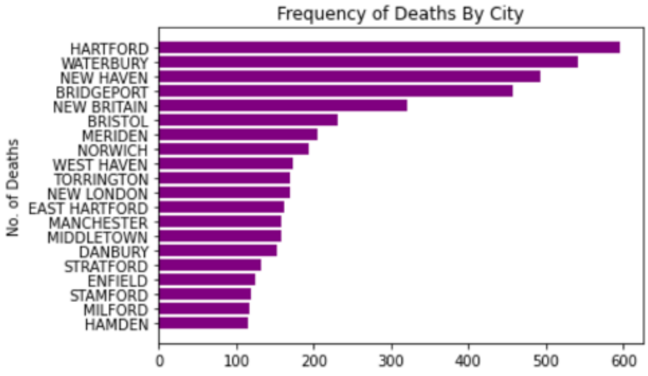


Figure 2.5: Frequency of Deaths by City

There was a high correlation association found between Residence City, Injury City, Death City, and Death County as visualized in the form of a heatmap in Figure 2.6.

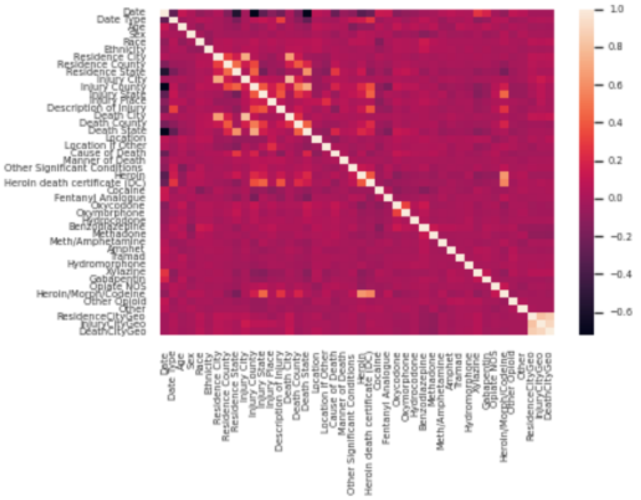


Figure 2.6: Heatmap of Correlation between Attributes

Among counties, Hartford had the highest frequency of deaths as shown in Figure 2.7.

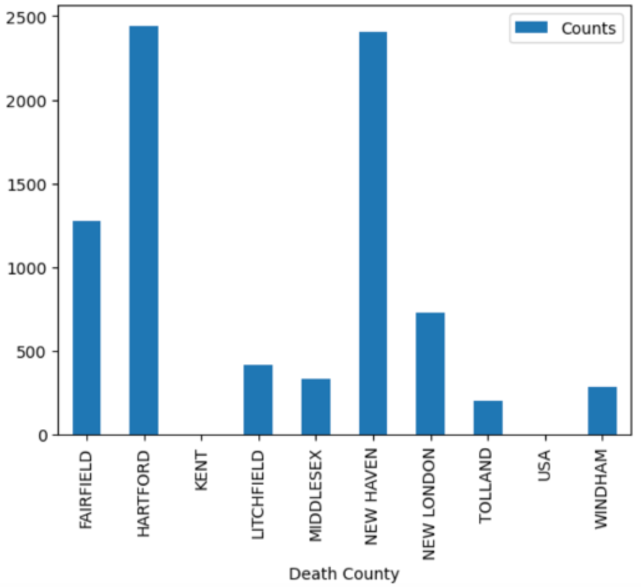


Figure 2.7: Deaths by County

Hispanics were the most vulnerable group ethnicity-wise (Figure 2.8), while White people had the highest chances of death by drug overdose race-wise (Figure 2.9).

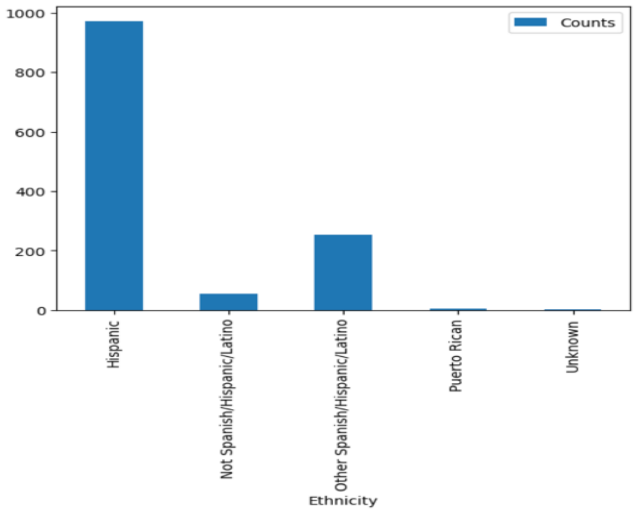


Figure 2.8: Deaths by Ethnicity

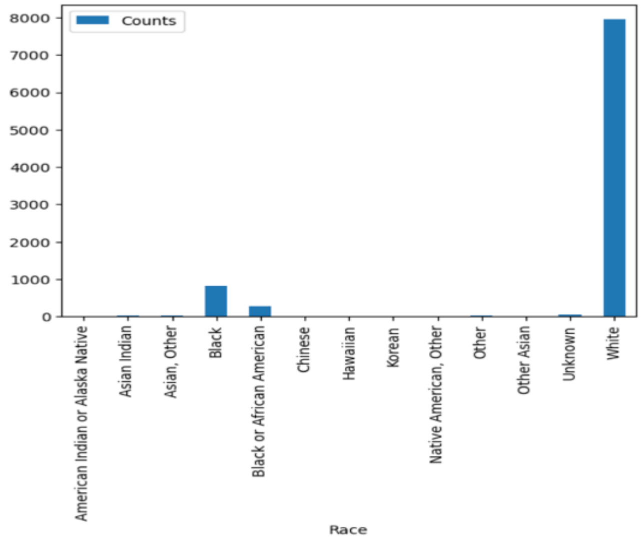


Figure 2.9: Deaths by Race

Lastly, Figure 2.10 shows our data points plotted over the Interactive Plotly map, exhibiting locations of deaths spread geographically.

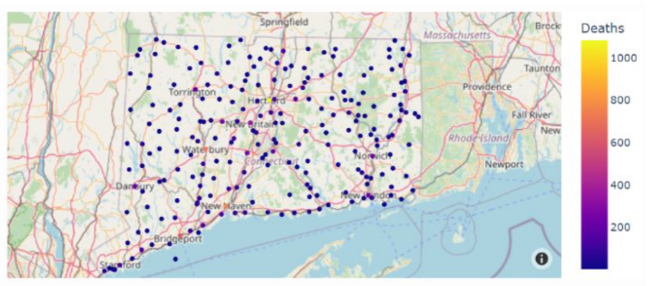


Figure 2.10: Plotly Map

III. CLUSTER ANALYSIS

Clustering is a technique used to find meaningful patterns and clusters in the data. In order to understand the dataset better, we need to gather insights in every form possible. There are many ways of doing clustering but the most popular and widely-used are K-Means and DBSCAN. K-means finds clusters based on the distance to the center of each cluster, while DBSCAN is a density based approach. DBSCAN discards data points that it considers noise and outliers and does not include them in our final clusters.

Although DBSCAN has more advantages, we have used k-means here due to the following reasons:

- DBSCAN can not efficiently handle large datasets.
- K-Means works better on high dimensional data.
- DBSCAN fails in determining neighborhood distances when the cluster densities are very different.
- It requires domain knowledge which is limited in our dataset.

As the k-means algorithm works on numerical data, we convert all of the entries of our dataset from string to numeric as shown in the following code snippet. We also disregard the features that are not relevant for our clustering task.

```
from sklearn.preprocessing import LabelEncoder

labels = LabelEncoder()

df_new=df_new[['Age','Sex','Race','Residence
City','Death City','Location','Cause of Death']]

col_list = list(df_new)

for col in col_list:

   try:

        df_new[col]=labels.fit_transform(df_new[col])

   except TypeError:

          df_new[col]=df_new[col].astype('string')
```

To run k-means on the above dataset, we need to find an appropriate value of k. We can get this value by using the elbow method. This method iterates over the dataset and finds clusters using many k values, and returns the most suitable value to use. Using the elbow method, we get a value of 3 for k.

Using this value, we run k-means and labels are assigned to all the data points based on the cluster they are part of. An important thing to note here is that k-means only works on numerical data and does not give desired clusters for categorical data. Let us separate numeric and categorical features and run k-means on them individually.

A. Categorical Features

```
cat_df = label_df[['Sex','Race','Location','labels']]

sns.pairplot(cat_df.sample(300),hue='labels')
```

It can be seen in figure 3.1 that categorical data clusters are not meaningful and do not give us any information regarding the relationship between two attributes or specific patterns.

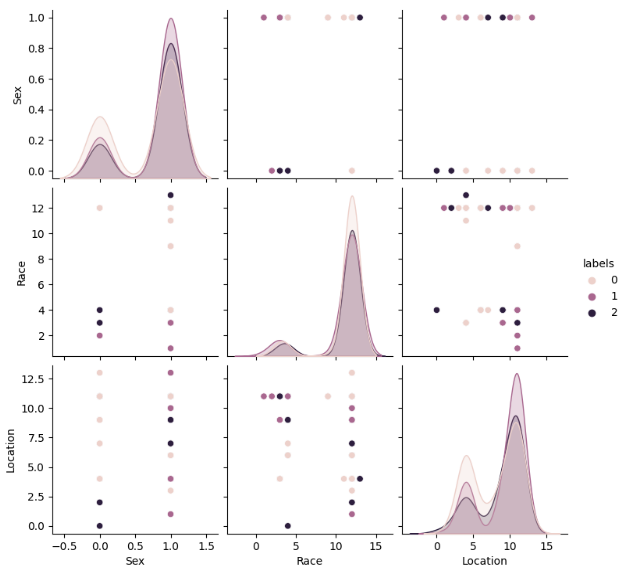
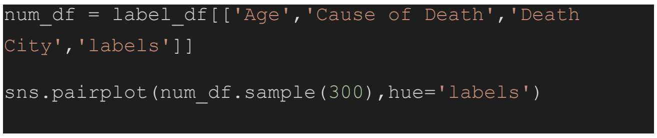


Figure 3.1: Clustering on Categorical Data

B. *Numerical Features*

Now, we look at numerical data clustering. We separate the columns which contain numerical values in a certain range.



```
num_df = label_df[['Age','Cause of Death','Death
City','labels']]

sns.pairplot(num_df.sample(300),hue='labels')
```

Let's visualize the clusters.

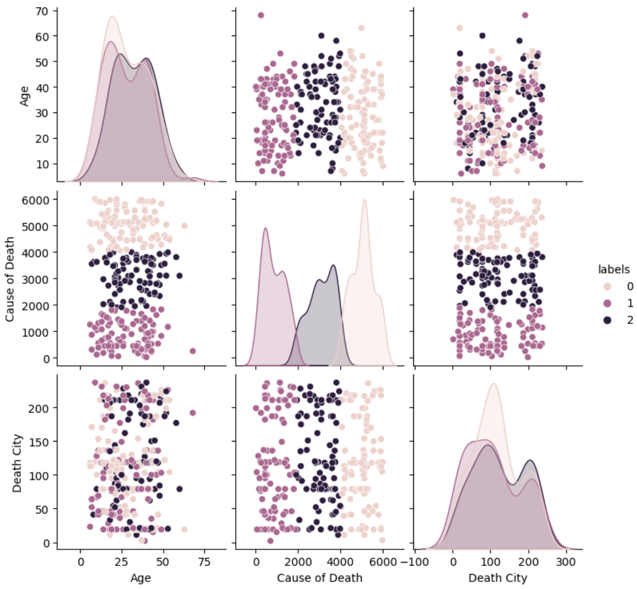


Figure 3.2: Clustering on Numerical Data

Figure 3.2 gives us clear and understandable patterns because the data were all numerical. We can see that age, cause of death and death city are highly correlated because there are distinct clusters formed which were not present in figure 3.1.

Hence, clustering is an important technique that provides us useful insights regarding which features can be used in future models and gain further understanding of the problem domain.

IV. TIME-SERIES ANALYSIS

After EDA and attaining the clean dataset, we proceed onto checking trends and stories in the data according to the changes over time.

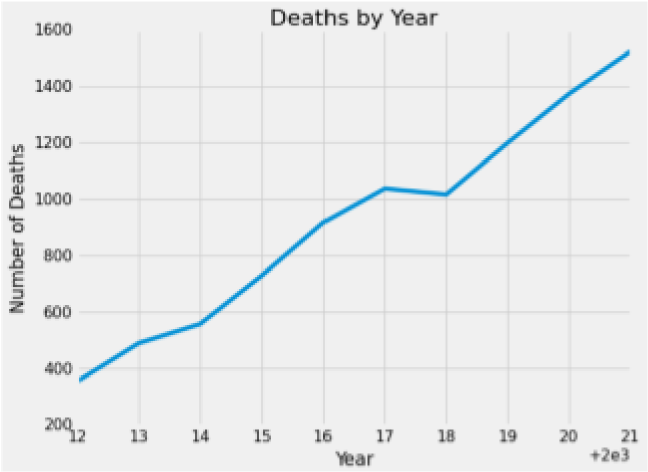


Figure 4.1 - Deaths By Year

Figure 4.1 shows us that the number of deaths have steadily grown over the years, almost linearly. The increase is almost 300%. The only year where deaths decreased was 2017. Otherwise, it has been growing.

We now proceeded to check the average age of those dying from drugs every year. The assumption is that more and more young people start doing drugs. However the trend showed us something different.

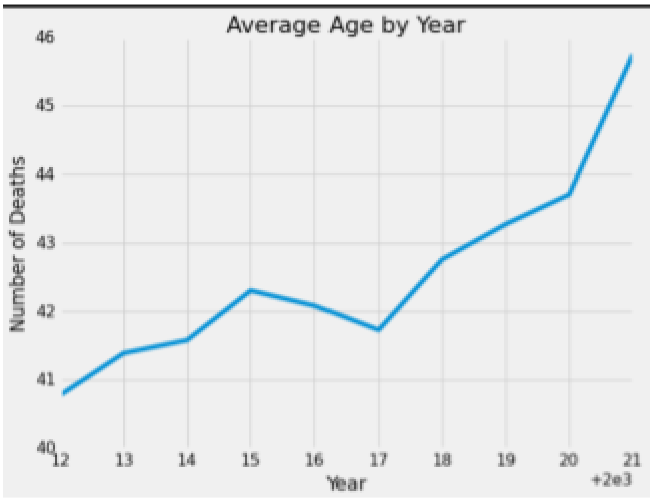
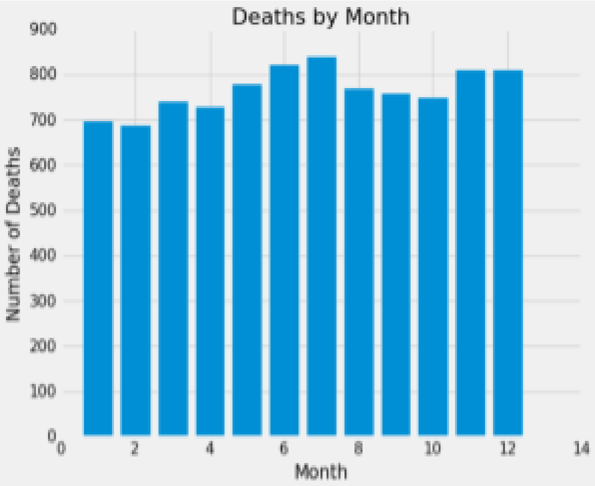


Figure 4.2 - Average Age with every passing year

The average age increased from around 40.77 to 42.30 from 2012 to 2015 according to figure 4.2. From 2015 until 2017, the average age fell 42 to less than 41, signaling new young users. From 2017 to 2020, an increase of almost 3 years in average age, simply meaning more older people started doing drugs and died. After that, it is consistent growth.

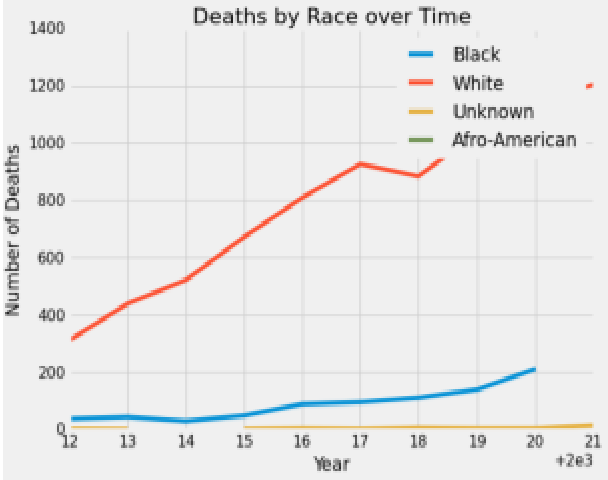
We also wanted to see the seasonal/over-the-year changes in the death rate.



4.3 - Average deaths each month in a 10-year period

In a period of ten years, we see that drug use increases in the summer months (May, June, July). This is the time where many people take holidays. The data we have is of places between New York and Boston, where winters are extremely cold, hence, in the summers people prefer to take holidays and students as well. Also, near Christmas, drug use also increases. Figure 4.3 depicts this extremely accurately.

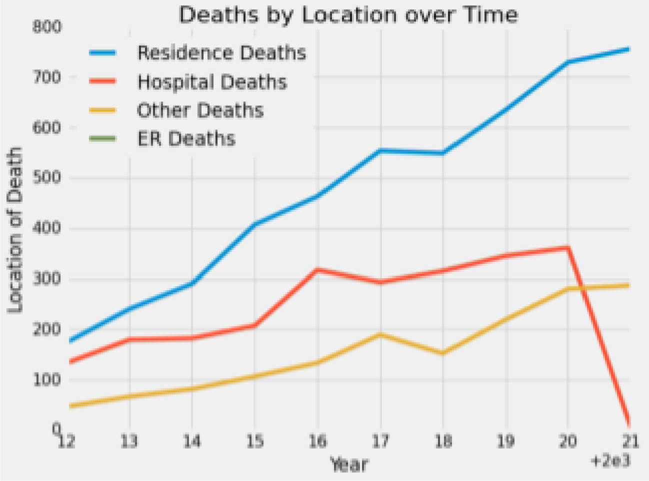
Checking deaths race-wise was incredibly interesting as seen in Figure 4.4.



4.4 - Race-wise Deaths over a 10-year period

Most drug-related deaths are of White people. Black people's death counts are much less. In fact, no black deaths were recorded during and after 2020.

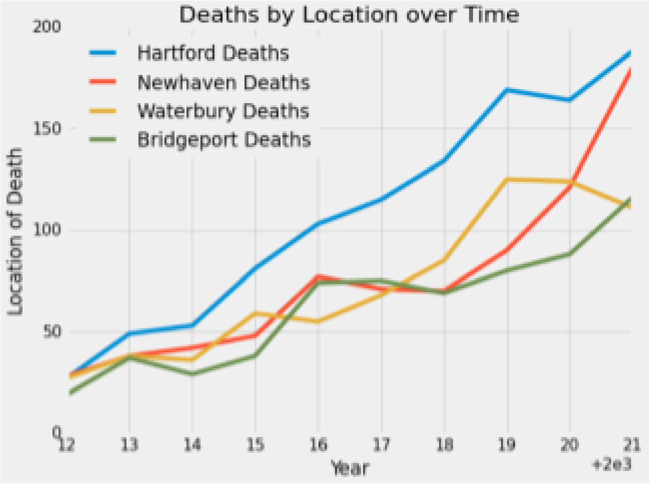
Next, we checked the various places where these dead bodies were found.



4.5 - Deaths by location of where the dead bodies were found

As seen in Figure 4.5, residential deaths continue to be the most common place for drug-related deaths and the number has increased. A sharp increase after 2018 can be seen. Hospital deaths followed a similar trend, albeit, with a lower magnitude. However, a sharp decline can be seen after 2020, owing to the COVID-19 pandemic and lockdown. ER deaths are very negligible.

As we only have data of a certain region of the United States, it is essential that we find out the city-wise pattern of deaths as well. Figure 4.6 comes into question.



4.6 - Deaths by City over a 10-year period

New Haven, Waterbury, and Bridgeport have almost identical trends. Increasing irregularly until 2017, and then a slight dip, followed by a sharp increase with New Haven seeing the biggest increase between 2020 and 2021. Hartford remains the highest in terms of magnitude and trends until a slight dip between 2019 and 2020.

V. ASSOCIATION RULE MINING

Results of Apriori(cumulative results on 3% and 5% minimun_support):

1) Support: Heroin, Cocaine, Fentanyl, Fentanyl Analogue, Oxycodone emerge as the individual drugs with highest confidence.

2) Confidence: Cocaine-Heroin, Fentanyl-Heroin, Fentanyl Analogue-Heroin, Ethanol-Heroin and Benzodiazepine-Heroin emerge as the double combinations that resulted in the most deaths. (Heroin, Any Opioid, Cocaine) , (Fentanyl Analogue, Heroin,Fentanyl) , (FentanylAnalogue, Fentanyl, Heroin) and (Ethanol, Heroin, Fentanyl) emerge as the most common triple combinations.

Results of FP-Growth(cumulative results on 3% and 5% minimun_support):

1) Support: Heroin, Cocaine, Any Opioid, Fentanyl, Oxycodone emerge as the individual drugs with highest support.

2) Cocaine-Any Opioid, Cocaine-Heroin, Fentanyl-Cocaine, emerge as the double combinations that resulted in the most Cocaine-Heroin, Fentanyl-Cocaine, emerge as the double combinations that resulted in the most deaths.(Cocaine, Heroin, Any Opioid) , (Cocaine, Any Opioid,Heroin) , (Heroin, Any Opioid,Cocaine) and (Cocaine, Heroin, Fentanyl) emerge as the most common triple combinations.

Conclusively, Heroin and Cocaine are common in both these algorithm results. However, in FP-Growth, Opioids show a greater impact on the death toll than those in Apriori. Fentanyl is also very common. We can conclude that Cocaine, Heroin and Fentanyl are deadly on their own. A combination of Cocaine and Heroin, Ethanol and Heroin, Benzodiazepine and Heroin are incredibly deadly.

VI. GEOSPATIAL ANALYSIS

In our mapping of deaths to their geographical locations, we filtered out cities with less than 50 deaths to find the cities that are most affected by drug related deaths and injuries. Figure 6.1 shows the color coded map of these deaths:

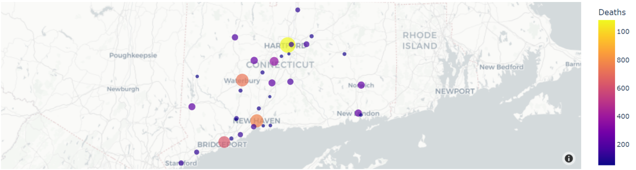


Figure 6.1

The map clearly shows that the most affected cities are Hartford with 1084 deaths, New Haven with 766 deaths, Waterbury with 728 deaths, and Bridgeport with 625 deaths. These cities are also the top cities by population in Connecticut with the exception of Stamford with 136 deaths. Stamford has a population of approximately 129,000 while that of New Haven is 135,000. The difference in deaths is 560% while the difference in population is only 5%. Hence preventative measures taken by Stamford city authorities can be analyzed further to determine if this large difference is a consequence of these measures or stems from something else entirely, this will help in replicating results in other cities.

VII. TEMPORAL ANALYSIS

According to our yearly analysis of drug-related deaths in Connecticut, the number of deaths due to heroin decreased steadily after 2016. However, deaths due to other drugs such as cocaine, fentanyl, oxycodone, oxymorphone, and ethanol continued to increase as shown by 7.1.

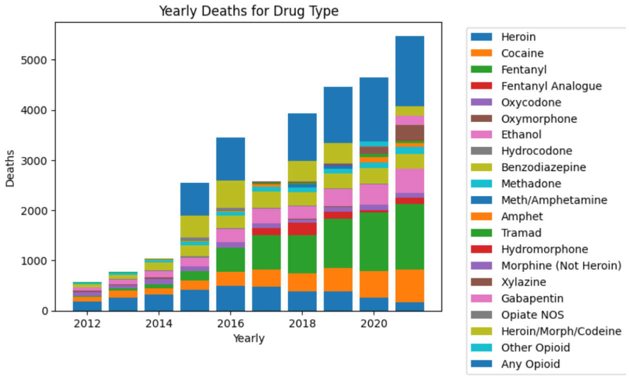
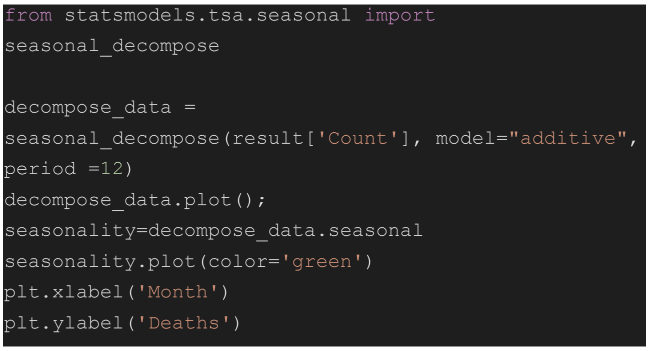


Figure 7.1

In our attempt to find patterns in our data that would enable us to produce accurate forecasts of drug related deaths in the state, we generated similar graphs like 7.1 for weekly and monthly timeframes. However no meaningful patterns could be extracted through these attempts aside from highlighting a slight uptick in deaths on Fridays on a weekly timeframe and in the month of July when analyzing monthly.

Hence we had to turn to seasonal fluctuations in deaths to provide us with patterns concrete enough for forecasting.

The following code demonstrates how the statsmodels libraries were used for decomposing the data into seasonal chunks:



```
from statsmodels.tsa.seasonal import
seasonal_decompose

decompose_data =
seasonal_decompose(result['Count'], model="additive",
period =12)
decompose_data.plot();
seasonality=decompose_data.seasonal
seasonality.plot(color='green')
plt.xlabel('Month')
plt.ylabel('Deaths')
```

Through this method a wave pattern emerged in seasonal deaths as shown by figure 7.2:

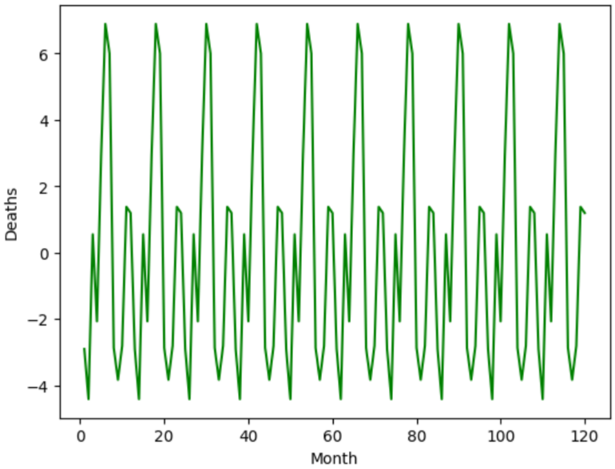
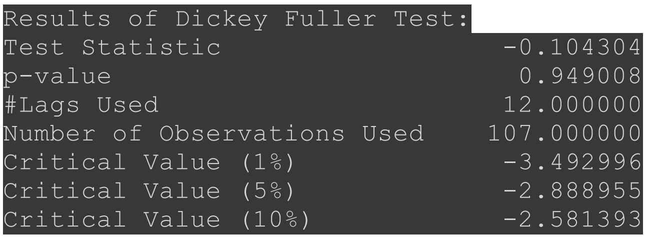


Figure 7.2

Now in order to proceed further with our forecasting we had to apply the Dickey Fuller Test on the seasonal data to ensure that it is fit for forecasting. The test results are as follows:

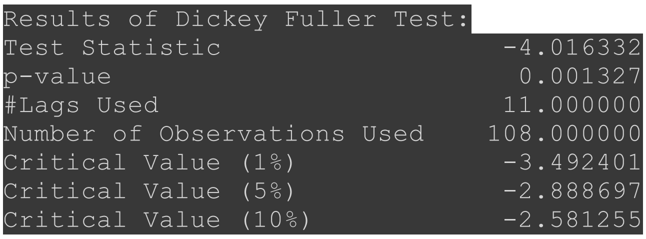


```
Results of Dickey Fuller Test:
Test Statistic                   -0.104304
p-value                           0.949008
#Lags Used                       12.000000
Number of Observations Used     107.000000
Critical Value (1%)              -3.492996
Critical Value (5%)              -2.888955
Critical Value (10%)             -2.581393
```

As the p-value in the test is a lot higher than the threshold of 0.05, it is unsuitable for forecasting. To make our data values pass the test we applied rolling mean differencing as demonstrated by the following code:

```
rolling_mean = result['Count'].rolling(window = 
12).mean()
result['rolling_mean_diff'] = rolling_mean - 
rolling_mean.shift()
```

Applying the Dickey Fuller Test after rolling mean differencing we get the following result:



```
Results of Dickey Fuller Test:
Test Statistic                   -4.016332
p-value                           0.001327
#Lags Used                       11.000000
Number of Observations Used     108.000000
Critical Value (1%)              -3.492401
Critical Value (5%)              -2.888697
Critical Value (10%)             -2.581255
```

Now that our data is ready for forecasting, we proceed to train a SARIMAX model with it:

```
import statsmodels.api as sm

model=sm.tsa.statespace.SARIMAX(result['Count'],order
=(1, 1, 1),seasonal_order=(1,1,1,12))
results=model.fit()
result['forecast']=results.predict(start=100,end=113,dy
namic=True)
```

```
result[['Count','forecast']].plot(figsize=(12,8))
```

Plotting the prediction from our model and comparing it with original values gives us the result shown by figure 7.3:

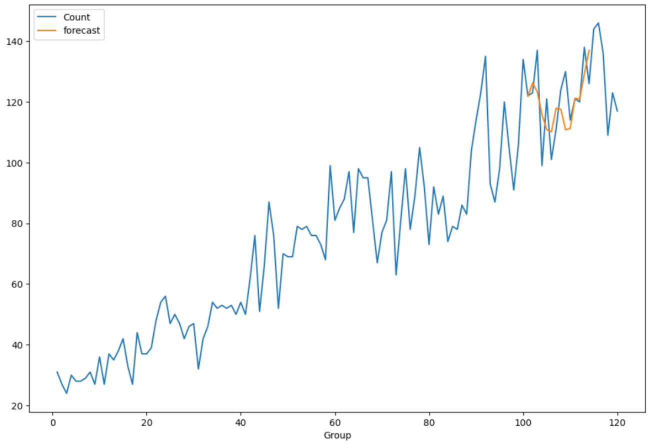


Figure 7.3

The graph in figure 7.3 shows that our model is underfitting so we need to apply further tests before confirming that there are no significant patterns in our data that can be used for forecasting. We chose the shapiro-wilk test for this purpose which divides our timeline into yearly timeframes and compares them to the first year.

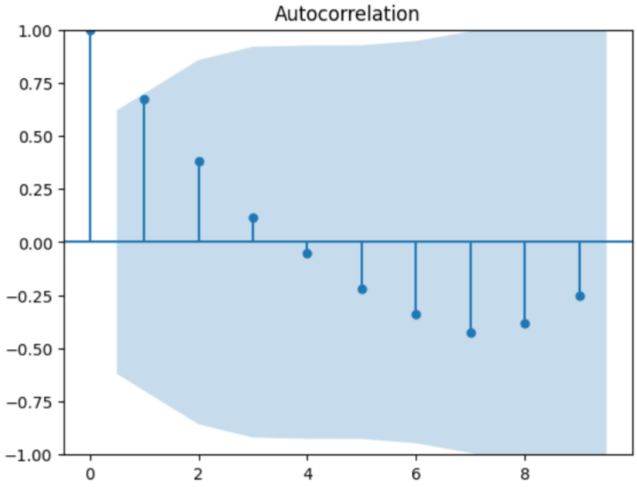


Figure 7.4

Figure 7.4 confirms that all subsequent years are not correlated with the first year as they lie within the blue highlighted area. The p-value for this test came out to be 0.88 hence the data is not normally distributed and cannot be used for forecasting. Therefore the conclusions that can be drawn from our temporal analysis are as follows: 1) Heroin related deaths have steadily decreased from 2016 onwards. 2) The most notorious drugs in terms of deaths caused are Cocaine, Fentanyl and Any Opioid(RX Morphine and Heroin-based morphine) 3) There is a slight increase in drug-related deaths on Fridays which may be explained by an increased number of individuals getting free from work for the week and indulging in intoxicants on a larger scale. 4) There are detectable seasonal patterns in the data however they are not significant enough to be used for forecasting.

VIII. CONCLUSION

The findings in this report provide crucial insights for authorities to develop more effective strategies to combat drug-related deaths. The disproportionate impact on certain ethnic and racial groups, such as Hispanics and Whites, respectively, underscores the need for culturally responsive interventions that address the unique needs and experiences of these groups. The sharp increase in deaths in recent years, particularly among older adults, suggests the need for interventions that address the complex factors contributing to drug-related deaths, such as chronic pain among others. Efforts to address the rise in deaths should focus on an approach that includes prevention, early intervention, and treatment, with a particular focus on addressing the root causes of addiction and providing access to effective treatment options. The decrease in heroin-related deaths in recent years suggests that certain interventions, such as increased access to medication-assisted treatment, may be effective in reducing drug-related deaths. In conclusion, the findings from this study provide important insights into the factors contributing to drug-related deaths in Connecticut between 2012 and 2021. These insights can inform the development of more targeted and effective interventions to combat the opioid epidemic and reduce the number of deaths related to drug use.

IX. FUTURE WORK

Additional insights from the concerned dataset may be drawn by analyzing the impact of COVID-19 on drug overdose deaths in Connecticut. This could involve comparing

overdose death rates before and after the pandemic and identifying any changes in the drugs being used or the demographic characteristics of those affected.