Krish Patel
kbp98
GitHub Link

Final Report

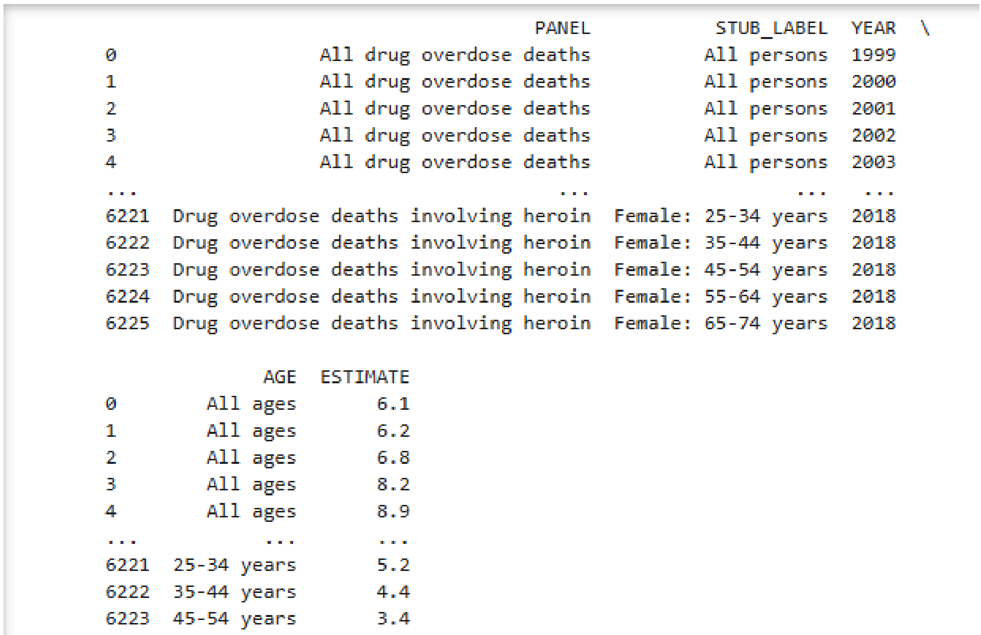
This project and dataset used are to address the critical issue of drug overdose deaths in the United States, with a specific focus on how these deaths vary across demographic lines — age, race, sex, and Hispanic origin — and seek to identify high-risk groups and forecast future overdose trends. The core question is twofold: How do overdose death rates vary across demographic groups, and can the model predict future trends for high-risk populations? By examining demographic disparities, the project aims to highlight significant trends in overdose deaths and predict future changes, providing data-driven insights to support possible public health initiatives. In the current landscape of pressing public health challenges, the issue of drug overdose deaths stands out as a profound and urgent crisis. This project's point lies in determining the limitations of standard epidemiological methods. The project relates to lectures and papers in data management by emphasizing data integration, transformation, and analysis. Lectures on SQL and database management provide techniques for structuring and storing the data to allow efficient querying for demographics and overdose types. Alongside data cleaning techniques taught in class, it will eliminate any outlier data or fill in any missing data in the file. Machine learning concepts covered in the course, such as regression and time-series forecasting, will be applied to predict future overdose trends, with data management principles ensuring data integrity and consistency for reliable analysis. This project involves descriptive analysis to uncover disparities and predictive modeling to project future outcomes.

This project is significant as drug overdose deaths continue to be a major public health crisis, with the opioid epidemic exacerbating this issue in recent years alongside a new killer that has taken over the youth: vaping. Public health authorities can use this project but need reliable, targeted data to design effective interventions. The novelty of this project lies in its combined approach of analyzing historical overdose disparities by demographic and using predictive analytics to forecast future trends. This dual approach identifies at-risk populations and predicts potential future spikes. One challenge in current data management is handling and cleaning complex datasets with various demographic markers, which often need more information. Additionally, demographic-specific data can be challenging to analyze due to slight differences between groups, which can sometimes go unnoticed without tailored analysis. While previous studies on overdose deaths might have identified general trends, few have combined demographic analysis with predictive modeling. Existing works report historical overdose data or offer predictive insights without describing demographic details. This project aims to bridge that gap by providing a comprehensive, demographic-specific analysis alongside predictive insights to inform public health efforts.

The project will use the “Drug Overdose Death Rates” dataset, which includes fields for drug type, sex, age, race, Hispanic origin, and more. This dataset is crucial as it contains the demographic details necessary for our analysis. The process begins with data cleaning and preparation, ensuring all demographic variables are consistent and handling missing or incomplete data entries. The data will be stored in an SQL database, organized by year, demographic categories, and drug type for easy querying. Once preprocessed, statistical methods and visualizations will be employed to identify overdose trends and disparities among demographic groups. Graphs can be used to visually display the data to show a more precise understanding and explain our big question.

Implementation steps include Data Integration, Data Cleaning and Transformation, Exploratory Data Analysis (EDA), Model Development, and Model Evaluation. A time-series approach or regression analysis will be used for predictive modeling to forecast future overdose rates. Using demographic variables as predictors, a supervised learning model will help predict trends, mainly focusing on groups identified as high-risk in the descriptive analysis. We’ll split the data into training and test sets to evaluate the model, using standard metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to assess accuracy. Success will be measured by the model’s ability to forecast overdose rates accurately and by the clarity of the demographic trends and disparities identified through data analysis.

The data used was from Data.gov, and it was a CSV file of Drug overdose death rates by drug type, sex, age, race, and Hispanic origin in the United States. I developed a multi-tiered approach to data analysis, implementing three distinct predictive modeling techniques: linear regression, polynomial regression, and an autoregressive model. Each approach offered a unique perspective, allowing me to capture the subtle and often nonlinear nature of overdose death trends. One of the existing issues in current data management practices is handling and cleaning complex datasets with various demographic markers. These datasets often need more complete or consistent information, making it challenging to conduct thorough analyses. Additionally, demographic-specific data can be complex to analyze due to subtle differences between groups, which can sometimes be overlooked without tailored analysis.



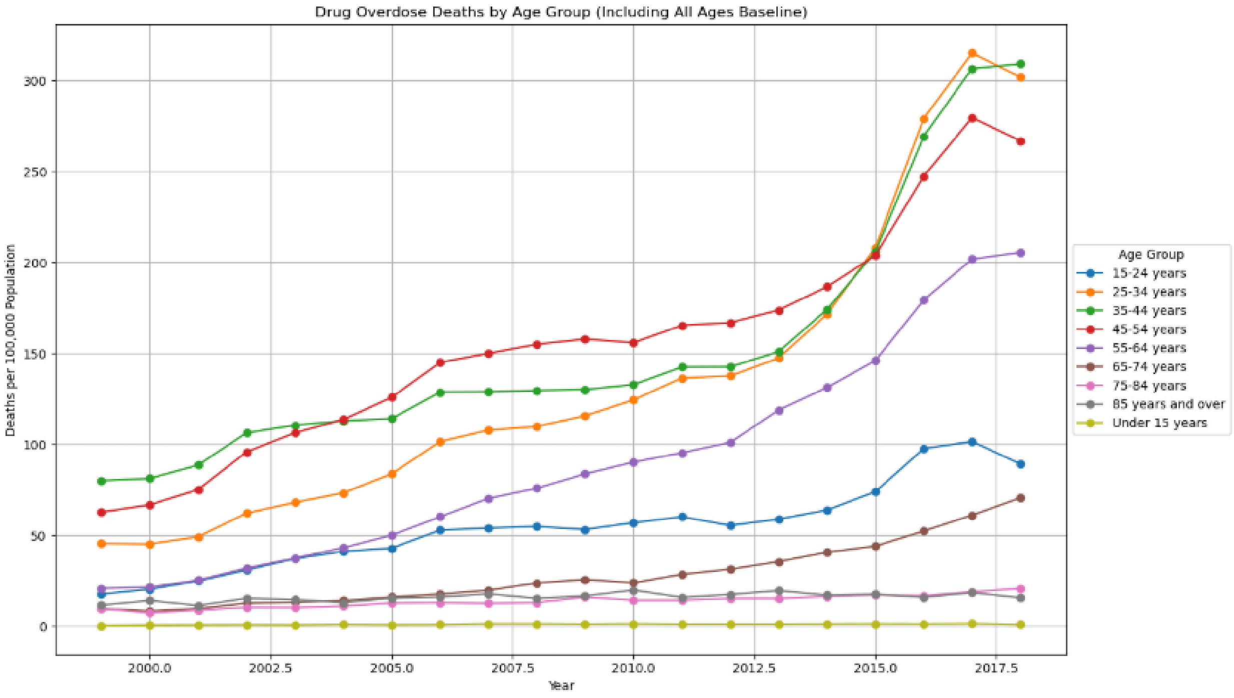
```
                                    PANEL          STUB_LABEL  YEAR  \
0                    All drug overdose deaths         All persons  1999
1                    All drug overdose deaths         All persons  2000
2                    All drug overdose deaths         All persons  2001
3                    All drug overdose deaths         All persons  2002
4                    All drug overdose deaths         All persons  2003
...                                       ...                 ...   ...
6221  Drug overdose deaths involving heroin  Female: 25-34 years  2018
6222  Drug overdose deaths involving heroin  Female: 35-44 years  2018
6223  Drug overdose deaths involving heroin  Female: 45-54 years  2018
6224  Drug overdose deaths involving heroin  Female: 55-64 years  2018
6225  Drug overdose deaths involving heroin  Female: 65-74 years  2018

              AGE  ESTIMATE
0         All ages       6.1
1         All ages       6.2
2         All ages       6.8
3         All ages       8.2
4         All ages       8.9
...            ...       ...
6221  25-34 years       5.2
6222  35-44 years       4.4
6223  45-54 years       3.4
```

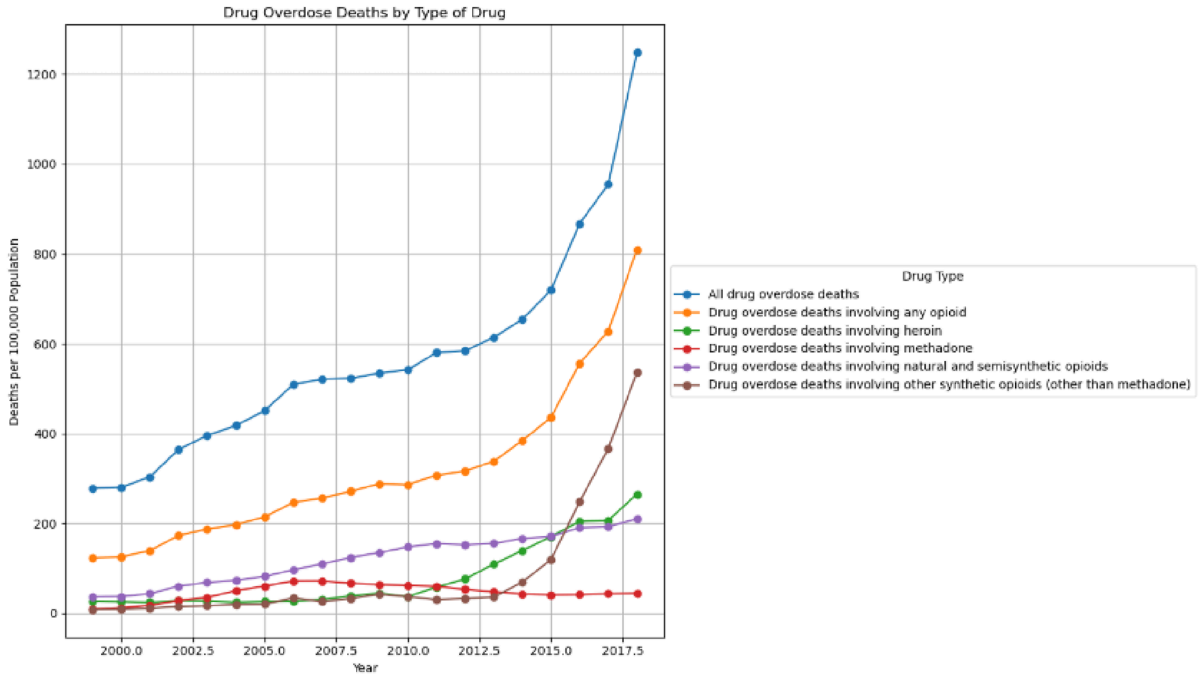
The data revealed stark and compelling insights into demographic vulnerabilities. The apparent risk was among males, particularly those not identifying as Hispanic or Latino, across multiple age groups. The 25-44 year demographic emerged as a critical zone of heightened mortality, with males in this age range experiencing the most amount of overdose deaths.

My analysis uncovered:

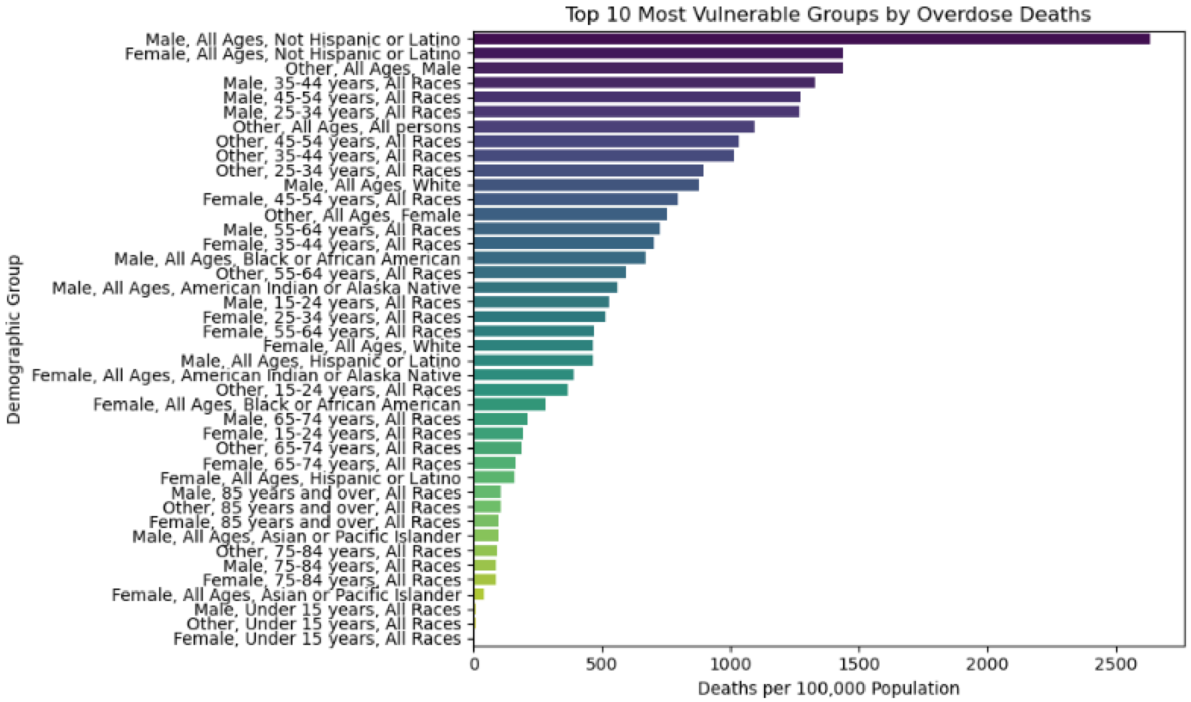
- Age Group: The 25-34 years age group experienced the most significant increase in drug overdose deaths, with a rise of 256.5 deaths per 100,000 population, highlighting this demographic as particularly vulnerable.
- Race/Ethnicity: The White population has the highest cumulative drug overdose deaths, with a total of 1,346.50 deaths per 100,000 population, indicating a significant vulnerability within this group.
- Gender: Males are the most vulnerable to drug overdoses, with a cumulative total of 6,721.40 deaths per 100,000 population, aligning with trends where males often exhibit higher fatality rates.
- Substance: Opioid-related deaths are the leading cause of overdose fatalities, with a total of 6,287.30 deaths per 100,000 population.

I designed experiments to test the accuracy and generalizability of each predictive model. The experiments involved splitting the data into training and test sets, fitting each model to the training data, and evaluating the model's performance on the test data. Standard metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were used to assess the accuracy of the models.

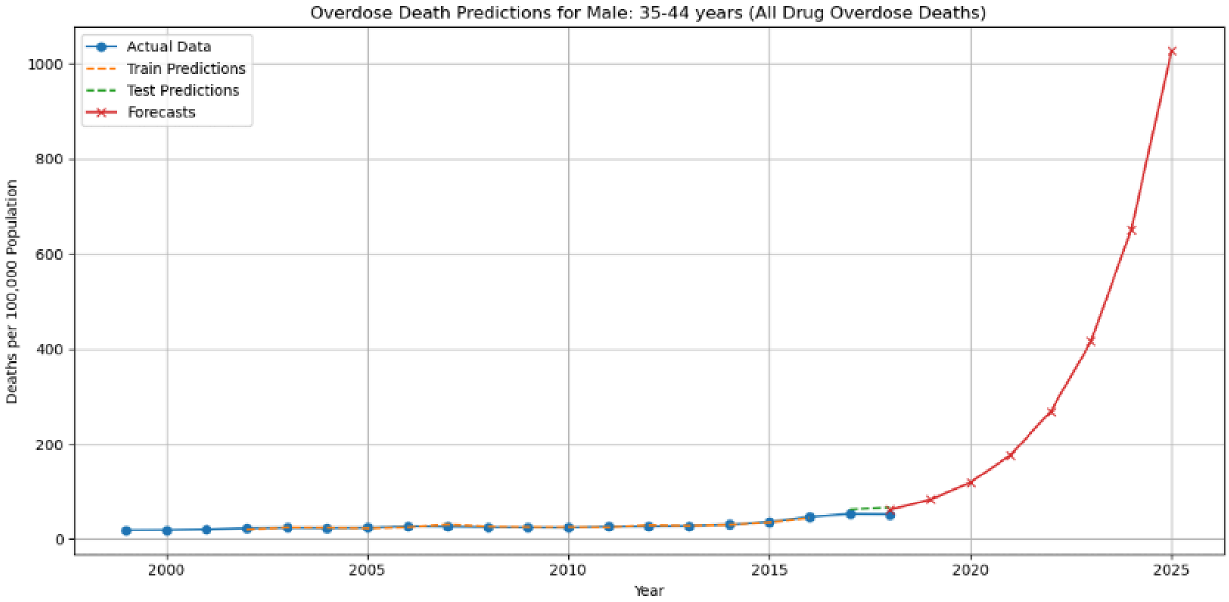




The predictive modeling results showed that the AR model had the lowest test MAE (7.09), indicating it provided the most accurate predictions among the models tested. These findings verified my hypothesis that demographic-specific analysis combined with predictive modeling could provide valuable insights for public health interventions.

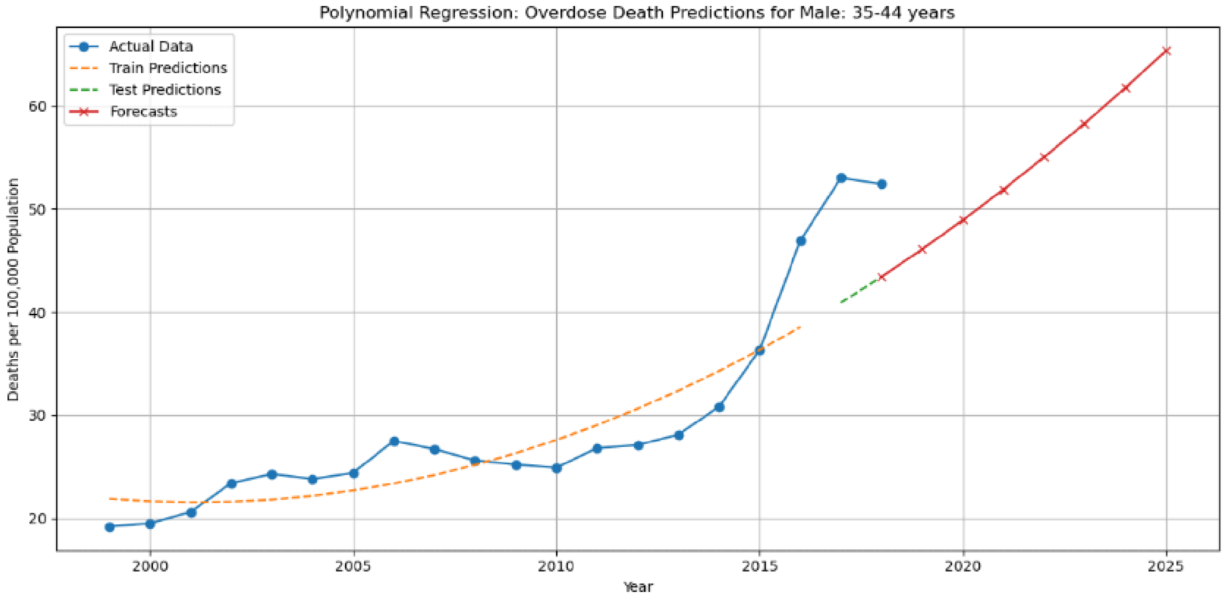


Here is my graph of Linear Regression Model:



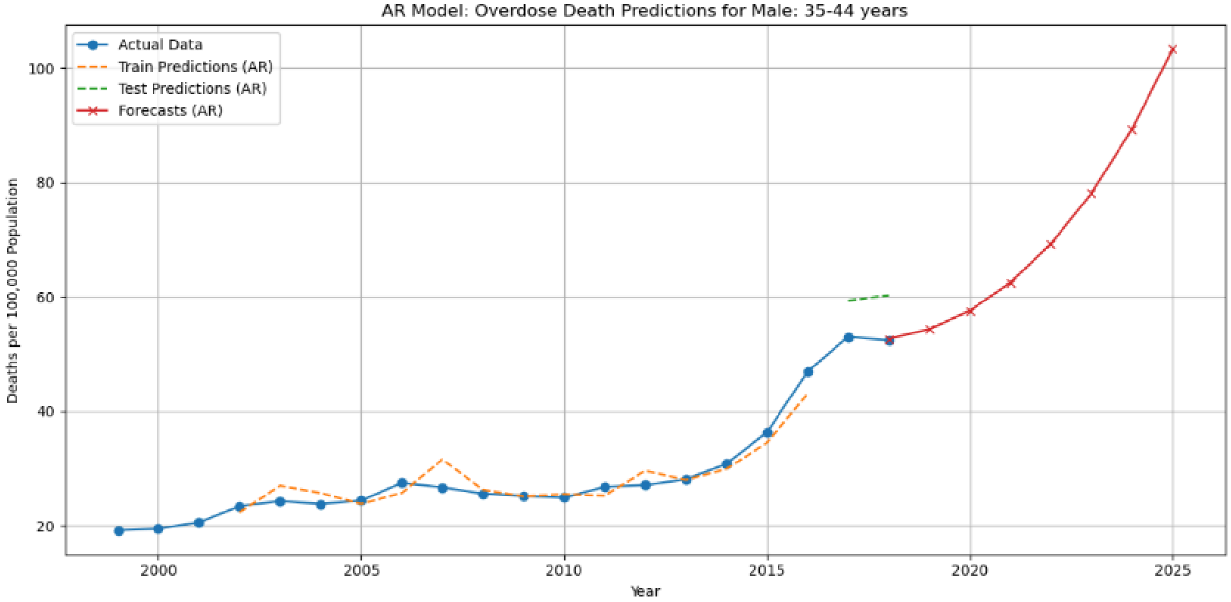
A simple linear regression model might not fully capture the underlying trends, especially if the relationship is nonlinear or influenced by multiple interacting factors. The linear regression model underperforms when comparing it to the other two models. While it tracks the actual data fairly well up to around 2010, it starts to deviate from the upward trend after that, especially after 2015. Its predictions for the test data are much less accurate, and the forecasts rise steeply toward 2025, predicting an unrealistic number of deaths per 100,000 population.

Here is my graph of Polynomial Regression Model:



This model shows a smoother curve for both train and test predictions, following the actual data up to 2015. The forecast after 2020 is also less extreme compared to the AR model, predicting a more gradual increase. However, the forecast still diverges from reality significantly towards 2025, potentially underestimating the impact of recent trends like the opioid crisis.

Here is my graph of Polynomial Regression Model:



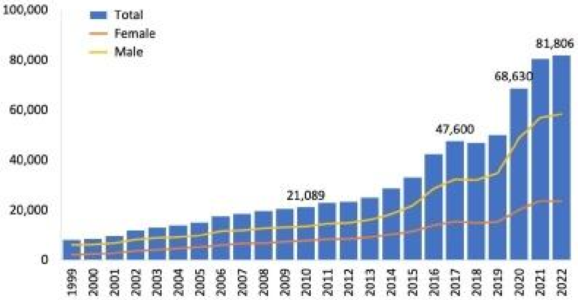
The AR model captures the upward trend in the actual data well, especially after 2015, with a noticeable increase in overdose deaths. However, its forecasts beyond 2020 tend to overshoot significantly, predicting far higher values than observed in the past. This suggests that the model may be overly sensitive to recent data and does not account for any external factors that could influence overdose rates.

The advantages of my approach are comprehensive analysis, which is the combined approach of historical analysis and predictive modeling that provides a holistic understanding of overdose trends, and targeted insights where identifying at-risk populations allows for more effective and targeted public health interventions, and data-driven decisions in which predictive analytics offer data-driven insights to inform future policies and strategies.

The limitation of my approach is model sensitivity because predictive models, especially the AR model, can be overly sensitive to recent data, leading to potential overestimation of future trends and generalizability since the models may struggle to generalize to unseen data, mainly if the relationship is nonlinear or influenced by multiple interacting factors.

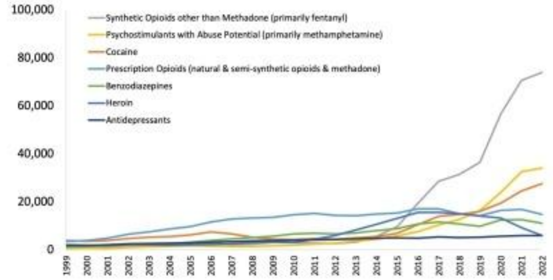
The final report broadly aligns with the proposed project, with a few refinements based on the data analysis and model performance. One significant change was the emphasis on the AR model, which provided the most accurate predictions among the models tested. I also included a more detailed demographic analysis, highlighting specific vulnerabilities within different age, race, and gender groups. These changes were made to enhance the comprehensiveness and accuracy of the findings. Common bottlenecks included handling missing or incomplete data entries, which required additional cleaning and preprocessing steps. Additionally, ensuring the models accurately captured the delicate nature of overdose trends was challenging.

Figure 3. U.S. Overdose Deaths Involving Any Opioid* by Sex, 1999-2022



*Among deaths with drug overdose as the underlying cause, the "any opioid" subcategory was determined by the following ICD-10 multiple cause-of-death codes: natural and semi-synthetic opioids (T40.2), methadone (T40.3), other synthetic opioids (other than methadone) (T40.4), or heroin (T40.1). Source: Centers for Disease Control and Prevention, National Center for Health Statistics. Multiple Cause of Death 1999-2022 on CDC WONDER Online Database, released 4/2024.

Figure 2. U.S. Overdose Deaths*, Select Drugs or Drug Categories, 1999-2022



*Includes deaths with underlying causes of unintentional drug poisoning (X40–X44), suicide drug poisoning (X60–X64), homicide drug poisoning (X85), or drug poisoning of undetermined intent (Y10–Y14), as coded in the International Classification of Diseases, 10th Revision. Source: Centers for Disease Control and Prevention, National Center for Health Statistics. Multiple Cause of Death 1999-2022 on CDC WONDER Online Database, released 4/2024.

In summary, the critical issue of drug overdose deaths in the United States can be sourced by analyzing demographic disparities and predicting future trends; this project provides valuable insights that can inform targeted public health interventions and policies. The use of advanced predictive models, such as the AR model, highlights the potential of data science in addressing complex social challenges. Recent reports indicate a significant decline in overdose deaths attributed to the widespread use of naloxone and improved addiction healthcare. According to a report by NPR, the Centers for Disease Control and Prevention (CDC) noted, "The latest data show that our efforts are working," with a 14.5% reduction in fatal overdoses over the past year, translating into more than 16,000 lives saved (Mann, 2024). This underscores the importance of continued focus and funding for addiction treatment and healthcare services, especially in marginalized and underserved communities.

Sources

https://catalog.data.gov/dataset/drug-overdose-death-rates-by-drug-type-sex-age-race-and-hispanic-origin-united-states-3f72f

https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates#Fig1

https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm

https://www.npr.org/2024/11/14/nx-s1-5191743/overdose-deaths-drop-fentanyl-opioid-crisis