KLE Technological University

Huballi

A Course Project Report on

"Predictive Modeling for HIV Exploring Health Indicator Data"

A Course Project Report Submitted in Partial Fulfillment of the Requirement for the Course of

Exploratory Data Analysis

in

4th Semester of Computer Science and Engineering

by

Dhanush Kamatar	02FE22BCS034
Sonali Jadhav	02FE22BCS146
Sourabh Nayak	02FE22BCS150
Stuti Hunachagi	02FE22BCS154

Under the guidance of

Dr. Prema Akkasaligar

Professor,

Department of Computer Science and Engineering,

KLE Technological University's Dr. MSSCET, Belagavi.

KLE Technological University's

Dr. M. S. Sheshgiri College of Engineering and Technology,

Belagavi – 590 008.

June 2024

DECLARATION

We hereby declare that the matter embodied in this report entitled as "**Predictive Modeling for HIV Exploring Health Indicator Data**" submitted to KLE Technological University for the course completion of Exploratory Data Analysis (21ECSC210) in the 4th Semester of Computer Science and Engineering is the result of the work done by us in the Department of Computer Science and Engineering, KLE Dr. M. S. Sheshgiri College of Engineering, Belagavi under the guidance of Dr. Prema Akkasaligar, Professor, Department of Computer Science and Engineering. We further declare that to the best of our knowledge and belief, the work reported here in doesn't form part of any other project on the basis of which a course or award was conferred on an earlier occasion on this by any other student, also the results of the work are not submitted for the award of any course, degree or diploma within this or in any other University or Institute. We hereby also confirm that all of the experimental work in this report has been done by us.

Belagavi – 590 008
Date : 10-June-2024

Dhanush Kamatar
(02FE22BCS034)

Sonali Jadhav
(02FE22BCS146)

Sourabh Nayak
(02FE22BCS150)

Stuti Hunachagi
(02FE22BCS154)

CERTIFICATE

This is to certify that the project entitled "Predictive Modeling for HIV Exploring Health Indicator Data" submitted to KLE Technological University's Dr. MSSCET, Belagavi for the partial fulfillment of the requirement for the course - Exploratory Data Analysis (21ECSC210) by Dhanush Kamatar(02FE22BCS034), Sonali Jadhav(02FE22BCS146), Sourabh Nayak(02FE22BCS150) and Stuti Hunachagi(02FE22BCS154) , students in the Department of Computer Science and Engineering, KLE Technological University's Dr. MSSCET, Belagavi, is a bonafide record of the work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any other course completion.

Belagavi – 590 008
Date : 10-June-2024

Dr. Prema Akkasaligar
(Course Teacher)

Prof. Priyanka Gavade
(Course Coordinator)

Dr. Rajashri Khanai
(Head of the Department)

Abstract

Analysis of exploratory data on cases of HIV disease

PROBLEM STATEMENT :

In this project we aim to identify, high-risk individuals and enable focused treatments for HIV prevention and improved health outcomes, predictive models for HIV risk assessment are being developed using demographic, behavioral, and healthcare data,so as to enhance safety measures.

Solution to Problem Statement:

- 1.Collection: Gather historical data on HIV/AIDS cases, including demographic information (age, gender, race), geographic details (borough), and clinical metrics (HIV and AIDS diagnoses, concurrent diagnoses, viral suppression, and deaths,death rates). This data can be sourced from public health records, hospital databases, and government reports.

- 2.Data Preprocessing: Clean and preprocess the data to handle missing values, standardize data formats, and ensure consistency across different data sources. This step includes imputing null values,imputing outliers and encoding categorical variables and normalizing numerical data.

- 3.Exploratory Data Analysis (EDA): Conduct an initial exploration of the dataset using descriptive statistics and visualizations. This aids in locating any outliers in the data as well as fundamental trends and distributions. One can use tools such as scatter plots, bar charts, and histograms.

- 4. Feature Selection: Identify the most influential features that are strongly correlated with HIV/AIDS outcomes.To determine the significance of each variable, use statistical methods like correlation analysis and feature importance metrics derived

from machine learning algorithms. For your predictive model, pick the pertinent features: age, gender, race; location information (borough and UHF neighborhood); and clinical metrics (deaths, viral suppression, HIV and AIDS diagnoses, concurrent diagnoses).

- Predictive Modeling: Build a predictive model using the selected features and historical HIV/AIDS dataTo build the model, use a variety of machine learning techniques, such as gradient boosting methods, decision trees, and random forests. To verify the model's performance and guarantee its accuracy and dependability in predicting HIV/AIDS outcomes, divide the data into training and testing sets.

RESULTS :

1.Correlation with Demographic Variables

2.Effect of Age on HIV/AIDS Outcomes

3.Effect of Gender on HIV/AIDS Incidence

4.Role of Race in HIV/AIDS Spread

5.Temporal Trends and Patterns

6.Identification of High-Risk Areas

7.Understanding Viral Suppression Rates

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

- HIV was first brought to humans by blood transfusions during hunting, and it originated from a virus that was specific to chimpanzees in West Africa in the 1930s. The virus traveled throughout Africa and other parts of the world over the years.

- In Chennai, India, the first HIV case was documented in 1986.

- The second-largest population of HIV and AIDS positive individuals is found in India (PLHA).

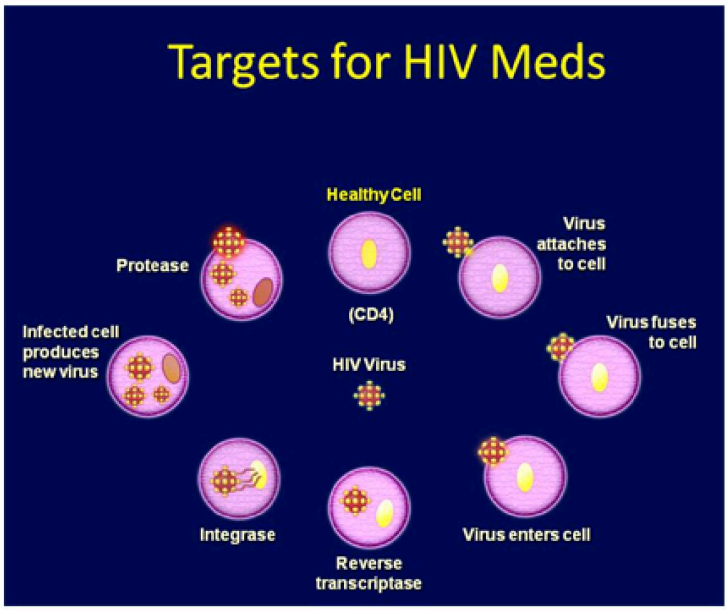


FIGURE 1.1: HIV Life cycle

- In comparison to the global average of 0.2, the current national prevalence is approximately 0.26.

1.2 Problem Statement

In this project we aim to identify, high-risk individuals and enable focused treatments for HIV prevention and improved health outcomes, predictive models for HIV risk assessment are being developed using demographic, behavioral, and healthcare data,so as to enhance safety measures.

1.2.1 Objectives

- To examine the trends in the number of HIV/AIDS diagnoses, related death rates, and diagnoses throughout time.

- To analyze variations in death rates, HIV diagnoses, and AIDS diagnoses among New York City's boroughs and UHF neighborhoods.

- To determine the differences in HIV and AIDS diagnoses, as well as death rates, between various age, gender, and racial groups.

- To assess the mortality rates from HIV-positive and HIV-negative populations in various demographic categories and geographical areas to determine high-risk groups.

Chapter 2

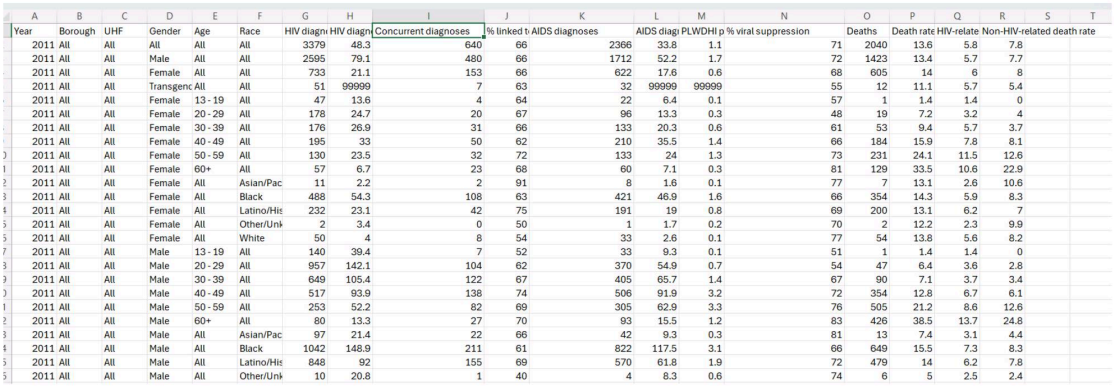
Knowing the Dataset

2.1 About the Dataset

The HIV Prediction Dataset is a comprehensive collection of data designed to facilitate the investigation and forecasting of HIV-related mortality. Numerous attributes are included in this dataset, including the year of data collection, demographic information, medical history, and treatment plan. The HIV Prediction Dataset is an extensive set of information created to make forecasting and analysis of HIV-related mortality easier. This dataset includes a number of characteristics, such as the year that the data were collected, demographic data, medical history, and course of treatment. Researchers and public health experts may create prediction models to estimate the number of HIV fatalities, uncover important risk factors for non-HIV mortality, and enhance intervention efforts by utilizing this comprehensive information. The ultimate objective of making use of this information is to improve our comprehension of the dynamics of the epidemic and facilitate the creation of more efficient healthcare and policy solutions to lessen the effects of HIV/AIDS globally.

2.2 Content of the Dataset

The dataset contains information on HIV/AIDS cases in New York City, categorized by year, borough, UHF neighborhood, gender, age group, and race.Metrics including the number of HIV diagnoses, diagnosis rates, concurrent diagnoses, linkage to care within three months, AIDS diagnoses, AIDS diagnosis rates, viral suppression rates, prevalence of PLWDHI, overall death rates, HIV-related death rates, and non-HIV-related death rates are among those included.



Year	Borough	UHF	Gender	Age	Race	HIV diagn	HIV diagn	Concurrent diagnoses	% linked t	AIDS diagnoses	AIDS diagi	PLWDHI p	% viral suppression	Deaths	Death rate	HIV-relate	Non-HIV-related death rate
2011	All	All	All	All	All	3379	48.3	640	66	2366	33.8	1.1	71	2040	13.6	5.8	7.8
2011	All	All	Male	All	All	2595	79.1	480	66	1712	52.2	1.7	72	1423	13.4	5.7	7.7
2011	All	All	Female	All	All	733	21.1	153	66	622	17.6	0.6	68	605	14	6	8
2011	All	All	Transgenc	All	All	51	99999	7	63	32	99999	99999	55	12	11.1	5.7	5.4
2011	All	All	Female	13 - 19	All	47	13.6	4	64	22	6.4	0.1	57	1	1.4	1.4	0
2011	All	All	Female	20 - 29	All	178	24.7	20	67	96	13.3	0.3	48	19	7.2	3.2	4
2011	All	All	Female	30 - 39	All	176	26.9	31	66	133	20.3	0.6	61	53	9.4	5.7	3.7
2011	All	All	Female	40 - 49	All	195	33	50	62	210	35.5	1.4	66	184	15.9	7.8	8.1
2011	All	All	Female	50 - 59	All	130	23.5	32	72	133	24	1.3	73	231	24.1	11.5	12.6
2011	All	All	Female	60+	All	57	6.7	23	68	60	7.1	0.3	81	129	33.5	10.6	22.9
2011	All	All	Female	All	Asian/Pac	11	2.2	2	91	8	1.6	0.1	77	7	13.1	2.6	10.6
2011	All	All	Female	All	Black	488	54.3	108	63	421	46.9	1.6	66	354	14.3	5.9	8.3
2011	All	All	Female	All	Latino/His	232	23.1	42	75	191	19	0.8	69	200	13.1	6.2	7
2011	All	All	Female	All	Other/Unk	2	3.4	0	50	1	1.7	0.2	70	2	12.2	2.3	9.9
2011	All	All	Female	All	White	50	4	8	54	33	2.6	0.1	77	54	13.8	5.6	8.2
2011	All	All	Male	13 - 19	All	140	39.4	7	52	33	9.3	0.1	51	1	1.4	1.4	0
2011	All	All	Male	20 - 29	All	957	142.1	104	62	370	54.9	0.7	54	47	6.4	3.6	2.8
2011	All	All	Male	30 - 39	All	649	105.4	122	67	405	65.7	1.4	67	90	7.1	3.7	3.4
2011	All	All	Male	40 - 49	All	517	93.9	138	74	506	91.9	3.2	72	354	12.8	6.7	6.1
2011	All	All	Male	50 - 59	All	253	52.2	82	69	305	62.9	3.3	76	505	21.2	8.6	12.6
2011	All	All	Male	60+	All	80	13.3	27	70	93	15.5	1.2	83	426	38.5	13.7	24.8
2011	All	All	Male	All	Asian/Pac	97	21.4	22	66	42	9.3	0.3	81	13	7.4	3.1	4.4
2011	All	All	Male	All	Black	1042	148.9	211	61	822	117.5	3.1	66	649	15.5	7.3	8.3
2011	All	All	Male	All	Latino/His	848	92	155	69	570	61.8	1.9	72	479	14	6.2	7.8
2011	All	All	Male	All	Other/Unk	10	20.8	1	40	4	8.3	0.6	74	6	5	2.5	2.4

FIGURE 2.1: Snapshot of HIV Dataset

Characteristics	**Value**
Name	Predictive Modeling for HIV Indicator Data
Source	https://catalog.data.gov/dataset
Number of instances	31926
Number of Features	18
Dataset Format	Comma Separated Values (CSV)Format

TABLE 2.1: Properties of dataset

2.3 Features of the Dataset

There are a total of 18 features in this dataset. They are described as follows in Table 2.2. In Table 2.3, the details of the features are summarized. For each feature, we list the type of the data, number of unique and missing values.

2.4 Data distribution of the features

In this section, we visualize the distributions of the different features of the dataset(as shown in Table 2.2)

- Year: The year of the data record.
- Borough: The borough within New York City where data was collected.
- UHF: The United Hospital Fund (UHF) neighborhood within the borough.
- Gender: Gender of the individuals in the dataset (e.g., Male, Female , Transgender).
- Age: Age group of the individuals (e.g., 13 - 19, 20 - 29, etc.).

TABLE 2.2: Details of the Features in the HIV Dataset.

Feature Name	**Data Type**	**Distinct Values**	**Missing Values**
Year	Numeric	10	0
Borough	String	5	0
UHF	String	42	0
Gender	String	3	0
HIV diagnoses	String	301	416
HIV diagnosis rate	Numeric	1944	416
Concurrent diagnoses	Numeric	99	116
linked to care within 3 months	Numeric	123	13274
AIDS diagnoses	Numeric	212	337
AIDS diagnosis rate	Numeric	1421	337
PLWDHI prevalence	Numeric	149	3553
viral suppression	Numeric	149	3553
Deaths	Numeric	264	0
Death rate	Numeric	606	1913
HIV-related death rate	Numeric	345	1913
Non-HIV-related death rate	Numeric	446	1913

- Race: Race or ethnic group of the individuals.
- HIV diagnoses: Number of HIV diagnoses.
- HIV diagnosis rate: Rate of HIV diagnoses per 100,000 people.
- Concurrent diagnoses: Concurrent diagnoses of HIV and another condition.
- linked to care within 3 months: Percentage of HIV-diagnosed individuals linked to care within 3 months.
- AIDS diagnoses: Number of AIDS diagnoses.
- AIDS diagnosis rate: Rate of AIDS diagnoses per 100,000 people.
- PLWDHI prevalence: Prevalence of people living with diagnosed HIV infection.
- viral suppression: Percentage of people with viral suppression among those diagnosed with HIV.
- Deaths: Number of deaths.
- Death rate: Death rate per 100,000 people.

- HIV-related death rate: Death rate due to HIV-related causes per 100,000 people.
- Non-HIV-related death rate: Death rate due to non HIV-related causes per100,000people.

2.5 Observations

- How are the features? All categorical? Mix?

  Features are mixed.

- Are there any missing values? If yes, are they large or small?

  Yes large

- What is the range of data items? How are they distributed?

  It depends on features of the Dataset

- Are there any outliers?

  Yes

- Are any of the features skewed?

  Yes

- Does any of the features require normalization, scaling?

  Yes

- Features are both of categorical and numerical.
- 77 percent of numerical and 22 percent of categorical.
- Overall what are characteristics of dataset?

  Environmental variables

  Missing values

  City wise data

  Race wise data

2.6 Statistical Data Analysis

The mean,maximum,minmum,standard deviation and qurtiles of all features are given below:

1. HIV diagnoses

The mean of HIV diagnoses is 0.674629

The max of HIV diagnoses is 10.00

The min of HIV diagnoses is 0.000

The standard deviation of HIV diagnoses is 1.643139‘

The 25th percentile of HIV diagnoses is 00

The 50th percentile of HIV diagnoses is 0

The 75th percentile of HIV diagnoses is 1

2. HIV diagnosis rate

The mean of HIV diagnosis rate is 12.024069

The max of HIV diagnosis rate 73.00

The min of HIV diagnosis rate is 0.000

The standard deviation of HIV diagnosis rate is 23.462885

The 25th percentile of HIV diagnosis rate is 0

The 50th percentile of HIV diagnosis rate is 0

The 75th percentile of HIV diagnosis rate is 9.5

3.Concurrent diagnoses

The mean of Concurrent diagnoses is 0.127324

The max of Concurrent diagnoses is 2.500

The min of Concurrent diagnoses is 0.000

The standard deviation of Concurrent diagnoses is 0.417061

The 25th percentile of Concurrent diagnoses is 0

The 50th percentile of Concurrent diagnoses is 0

The 75th percentile of Concurrent diagnoses is 0

4.AIDS diagnoses

The mean of AIDS diagnoses is 0.445419

The max of AIDS diagnoses is 7.5000

The min of AIDS diagnoses is 0.000

The standard deviation of AIDS diagnoses is 1.106367

The 25th percentile of AIDS diagnoses is 0

The 50th percentile of AIDS diagnoses is 0

The 75th percentile of AIDS diagnoses is 0

5.AIDS diagnosis rate

The mean of AIDS diagnosis rate is 7.010703

The max of AIDS diagnosis rate is 45.25000

The min of AIDS diagnosis rate is 0.000

The standard deviation of AIDS diagnosis rate is 14.738960

The 25th percentile of AIDS diagnosis rate is 0

The 50th percentile of AIDS diagnosis rate is 0

The 75th percentile of AIDS diagnosis rate is 0

6.PLWDHI prevalence

The mean of PLWDHI prevalence is 0.949437

The max of PLWDHI prevalence is 3.85000

The min of PLWDHI prevalence is 0.000

The standard deviation of PLWDHI prevalence is 1.248561

The 25th percentile of PLWDHI prevalence is 0.0400

The 50th percentile of PLWDHI prevalence is 0.3000

The 75th percentile of PLWDHI prevalence is 1.4000

7.viral suppression

The mean of viral suppression is 0.862815

The max of viral suppression is 1

The min of viral suppression is 0.500

The standard deviation of viral suppression is 0.148890

The 25th percentile of viral suppression is 0.790000

The 50th percentile of viral suppression is 0.9000

The 75th percentile of viral suppression is 1.000 **8. Deaths**

The mean of Deaths is 0.593286

The max of Deaths is 7.500

The min of Deaths is 0.00

The standard deviation of Deaths is 1.495202

The 25th percentile of Deaths is 0

The 50th percentile of Deaths is 0

The 75th percentile of Deaths is 0

9 .Death rate

The mean of Death rate is 1.982890

The max of Death rate is 18.2500

The min of Death rate is 0.00

The standard deviation of Death rate is 5.166937

The 25th percentile of Death rate is 0

The 50th percentile of Death rate is 0

The 75th percentile of Death rate is 0

10.HIV-related death rate

The mean of HIV-related death rate is 1.012657

The max of HIV-related death rate is 1000.00

The min of HIV-related death rate is 0

The standard deviation of HIV-related death rate is 14.432481

The 25th percentile of HIV-related death rate is 0

The 50th percentile of HIV-related death rate is 0

The 75th percentile of HIV-related death rate is 0

11.Non-HIV-related death rate

The mean of Non-HIV-related death rate is 0.59950

The max of Non-HIV-related death rate is 7.000

The min of Non-HIV-related death rate is 0.00

The standard deviation ofNon-HIV-related death rate is 1.907994

The 25th percentile ofNon-HIV-related death rate is 0

The 50th percentile of Non-HIV-related death rate is 0

The 75th percentile of Non-HIV-related death rate is 0

Chapter 3

Implement Framework

To perform exploratory data analysis on the HIV/AIDS dataset, we have followed the following implementation framework. The overall implementation flow is presented in Figure 3.1. 1. Data Loading: Data Loading:Load the dataset into your preferred data analysis environment, such as Python with Pandas .Make that the data is correctly interpreted, taking into account the right data types for each characteristic. **2. Data Cleaning:** Make good use of missing values. You may choose to impute missing values, eliminate rows or columns that include missing values, or employ other methods to deal with the missing data, depending on the quantity and kind of the missing data.

3. Data Exploration: To obtain a basic understanding of the dataset, run summary statistics. This could entail figuring out the numerical features' mean, median, standard deviation, lowest, and maximum values. You can count unique values for categorical features, such as "city," and investigate their distribution.

4. Data Visualization: Construct visualizations to enhance your comprehension of the distribution and connections among various attributes. To visualize the data, you can use scatter plots, line plots, box plots, histograms, and other kinds of plots.

5. Outlier Detection: Locate and manage data outliers. The performance of models and statistical analysis can be greatly impacted by outliers. To identify outliers and determine how to respond to them (such as eliminating them or altering the data), think about utilizing box plots or other techniques.

6. Feature Engineering: You may need to develop new features or extract more data from already-existing ones, depending on the objectives of the analysis. For example, the "week start date" function can be used to extract the year, month, or day of the week.

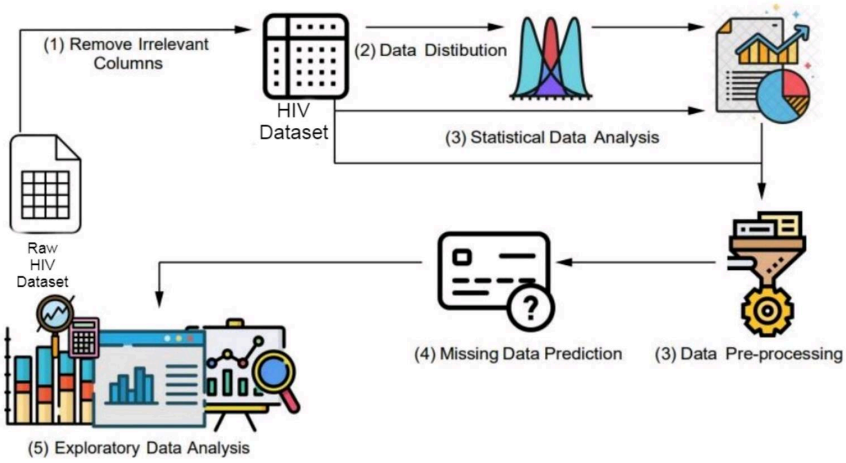


Figure 3.1: Overall Implementation Flow.

FIGURE 3.1: Overall Implementation flow

7. Correlation Analysis: Use scatter plots or correlation matrices to examine the relationships between various numerical parameters. This will assist you in figuring out how the variables relate to one another and in spotting any multicollinearity.

8. Skewness and Normalization: Verify whether numerical features are skew. Use transformations (such as log transformation) if necessary to improve the distribution of the data. Certain analytical or modeling techniques may additionally require scaling or normalization.

9. Grouping and Aggregation: Based on the study objectives, you may wish to aggregate the data and do comparisons or computations by grouping the data according to particular qualities, like "year" .

Chapter 4

Data Pre-processing

1. Data pre-processing: for data analysis or machine learning project, data must first be pre-processed. To prepare the raw data for analysis and modeling, it must be cleaned and transformed. The following are the pre-processing processes for data that can be used with the HIV dataset.

2. Handling Missing Values: Examine the dataset for any missing values and determine how to treat them. You can either eliminate rows or columns with missing values or impute them using suitable techniques like mean, median, or regression imputation, depending on the degree of missingness and the type of data.

3.Dealing with Duplicates: Examine the dataset for duplicate items and, if required, eliminate them. Model performance and analysis findings can be distorted by duplicate records.

4. Data Transformation: To guarantee consistency and compatibility, carry out the required data transformations. To make manipulation easier, you could, for instance, transform the "week start date" field to a datetime format.

5. Coding category Variables: Use methods like onehot encoding or label encoding to transform any category variables (such as "city") in the dataset into a numerical representation.

6.Feature Scaling: To bring all of the features in the dataset to a common scale, feature scaling should be used if the features have different sizes.Standardization and min-max scaling are common techniques.

7. Outlier Detection and Handling: Locate and manage anomalies within the information.The performance of some machine learning models can be greatly impacted by outliers. To lessen the impact of outliers, you have the option to either eliminate them or apply changes.

8. Feature Selection: Determine whether each feature is relevant to the target variable by analyzing it. If necessary, select a feature. Removing features that are highly linked or irrelevant might decrease overfitting and increase model efficiency.

9. Splitting the Dataset: Create a training set and a testing set by dividing the dataset in half. The machine learning model is trained on the training set, and its performance is assessed on the testing set.

10. Handling Imbalanced Data (if Applicable): If the target variable exhibits a notable class imbalance, take into account employing methods such as oversampling, undersampling, or creating synthetic samples in order to correct the data.

Results of Data Pre-processing:

You will have a cleaned and converted dataset that is prepared for analysis and modeling once the data pre-processing procedures have been applied. It would have dataset with no missing data, eliminated duplicates, encoded category variables, and scaled features properly. The data has been divided into training and testing sets after any outliers may have been handled. To obtain insights and see patterns in the data, exploratory data analysis (EDA) can now be performed on this pre-processed dataset. It can also be fed into different machine learning algorithms for testing and training in order to create predictive models for predicting HIV diagnoses based on the characteristics that are currently accessible.

Chapter 5

Exploratory Data Analysis

5.1 Hypothesis on the Problem Statement

- 1.Is their any corelation geographical distribution of aids or hiv with borough?
- 2.Does the distribution of HIV and concurrent diagnoses differ significantly across the boroughs
- 3.Does the rate of HIV diagnoses differ significantly among age groups, and is there a correlation between age and the HIV diagnosis?
- 4.What rate of non-hiv related death occured of overall deaths?
- 5.Is there a correlation between age and PLWDHI prevalence, and does this correlation reveal age-specific trends in HIV diagnosis rates?
- 6.Does age influence HIV diagnosis rates, AIDS-related death rates, and PLWDHI prevalence among different age groups?
- 7.How have HIV diagnosis rates and concurrent diagnoses changed over the years 2017 to 2021, and what potential factors contributed to these changes?
- 8.How does age impact the prevalence of PLWDHI and death rates?
- 9.What are the trends in AIDS diagnoses and death rates across different boroughs?
- 10.What are the trends in Concurrent diagnoses and Aids diagnoses across different boroughs?

1.HIV / AIDS DIAGNOSES WITH PLACES

```
In [59]: p=data.groupby("Borough")
         hd=p["HIV diagnoses"].sum()
         ad=p["AIDS diagnoses"].sum()
         label=['Bronx', 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island']
         plt.bar(label,hd,color="orange",label="HIV diagnoses")
         plt.bar(label,ad,color="green",label="AIDS diagnoses ")
         plt.xlabel("CITIES OF BOROUH")
         plt.ylabel("HIV AND AIDS DIAGNOSES COUNT")
         plt.legend()
```

FIGURE 5.1: HIV Diagnoses with borough

```
Out[59]: <matplotlib.legend.Legend at 0x20ca3b64910>
```

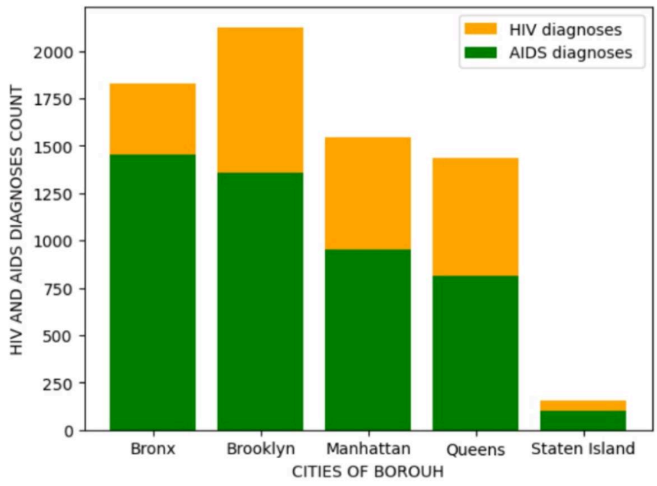


FIGURE 5.2: Analysing HIV Diagnoses with borough

- 11.Do certain racial groups experience higher rates of HIV diagnoses and death rates compared to others, indicating disparities in HIV/AIDS outcomes based on race?

- 12.What are the patterns in AIDS, HIV, and concurrent diagnoses across different age groups?

5.2 Analysis

Bi-variate and Multi-variate analysis.

2.HIV / CONCURRENT DIAGNOSES WITH YEAR

In [64]:
```
p=data.groupby("Year")

g=p["Concurrent diagnoses"].sum()
h=p["HIV diagnoses"].sum()
label = ['Bronx', 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island']
plt.plot(label,g, color="pink", marker='o', linestyle='-', label="concurrent diagnoses")
plt.plot(label,h, color="purple", marker='o', linestyle='-', label="AIDS diagnoses")

# Add labels and title
plt.xlabel("Boroughs")
plt.ylabel("Diagnoses Count")
plt.title("HIV and Concurrent Diagnoses by Borough")

# Add legend
plt.legend()

# Show plot
plt.show()
```

FIGURE 5.3: HIV and Concurrent diagnoses with year

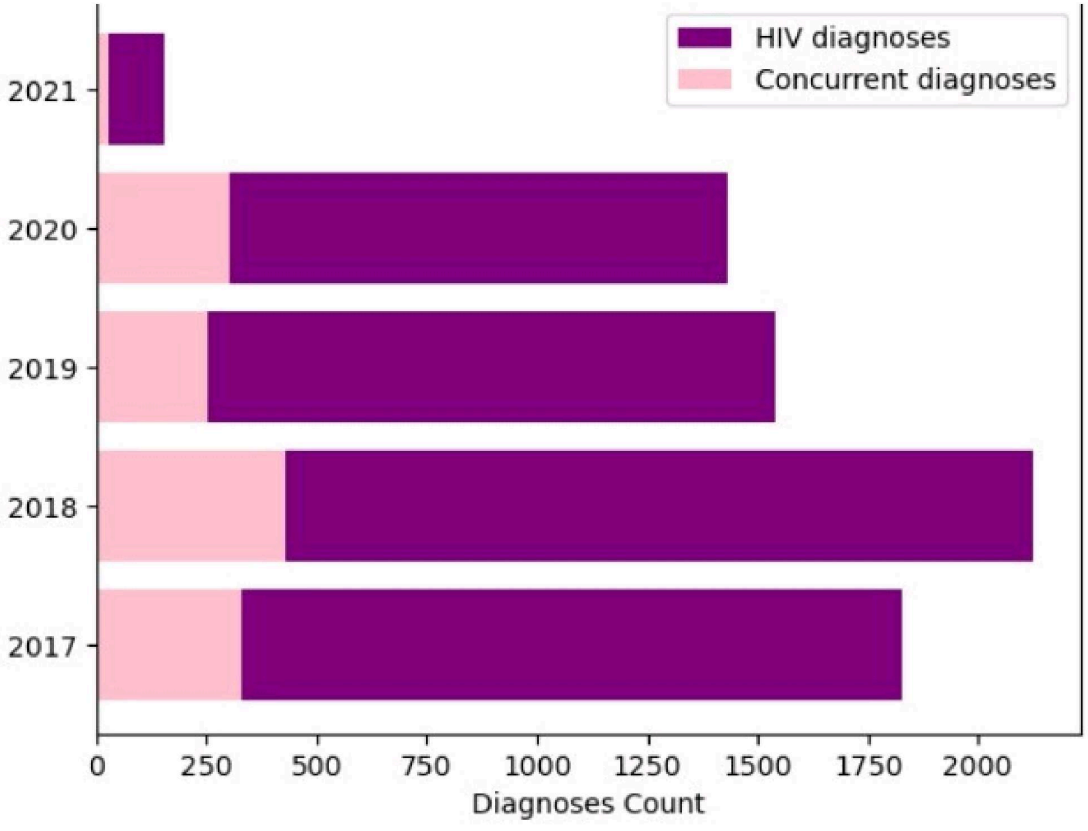


FIGURE 5.4: Analysing HIV and concurrent diagnoses with year

3.HIV Diagnoses and Deaths

In [67]:
```
p=data.groupby("Age")
g=p["HIV diagnoses"].sum()
h=p["Deaths"].sum()


label =['18 - 29','30 - 39','40 - 49','50 - 59','60+']
x = range(len(label))  # the label locations

width = 0.35  # the width of the bars

fig, ax = plt.subplots()
rects2 = ax.bar([i + width for i in x], h, width, label='HIV diagnoses', color='green')
rects1 = ax.bar(x, g, width, label='Deaths', color='orange')

# Add some text for labels, title and custom x-axis tick labels, etc.
ax.set_xlabel('AGE')
ax.set_ylabel('DEATHS')
ax.set_title('Number Of Deaths')
ax.set_xticks([i + width / 2 for i in x])
ax.set_xticklabels(label)
ax.legend()

fig.tight_layout()

plt.show()
```

FIGURE 5.5: Hiv diagnoses and death rates

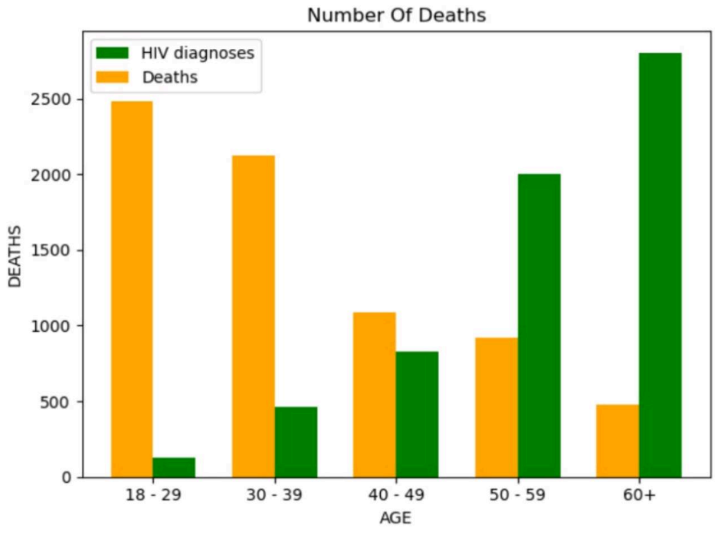


FIGURE 5.6: Analysing Hiv diagnoses and death rates

4. Death rate v/s non_hiv death rate

In [68]:
```
p=data.groupby("Borough")
hd=p["Death rate"].sum()
ad=p["Non-HIV-related death rate"].sum()
label =['18 - 29','30 - 39','40 - 49','50 - 59','60+']
plt.bar(label,hd,color="orange",label="Death rate")
plt.bar(label,ad,color="green",label="Non HIV Death Rate ")
plt.xlabel("CITIES OF BOROUH")
plt.ylabel("HIV AND AIDS DIAGNOSES COUNT")
plt.legend()
```

FIGURE 5.7: Death rates with non-hiv related death rates

```
Out[68]: <matplotlib.legend.Legend at 0x20ca0981090>
```

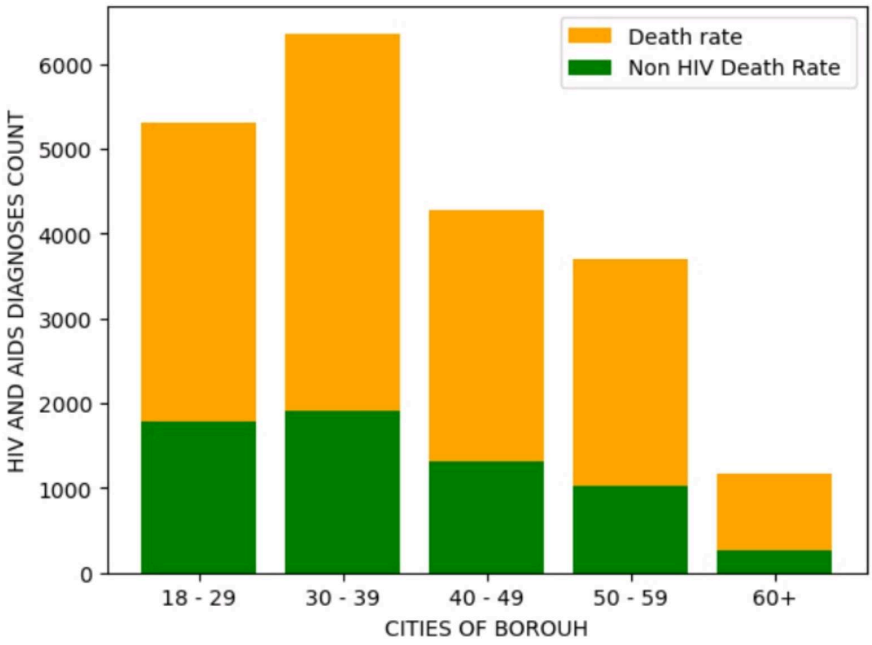


FIGURE 5.8: Analysing Death rates with non-hiv related death rates

5.Age wise PLWDHI prevalence variance

```
In [69]: p=data.groupby("Age")

         g=p["PLWDHI prevalence"].sum()

         label =['18 - 29','30 - 39','40 - 49','50 - 59','60+']
         plt.plot(label,g, color="red", marker='o', linestyle='-', label="PLWDHI prevalence",mfc="Blue")

         # Add labels and title
         plt.xlabel("Boroughs")
         plt.ylabel("Diagnoses Count")
         plt.title("HIV and Concurrent Diagnoses by Borough")
         # Add legend
         plt.legend()
         # Show plot
         plt.show()
```

FIGURE 5.9: Age wise PLWDHI prevalence

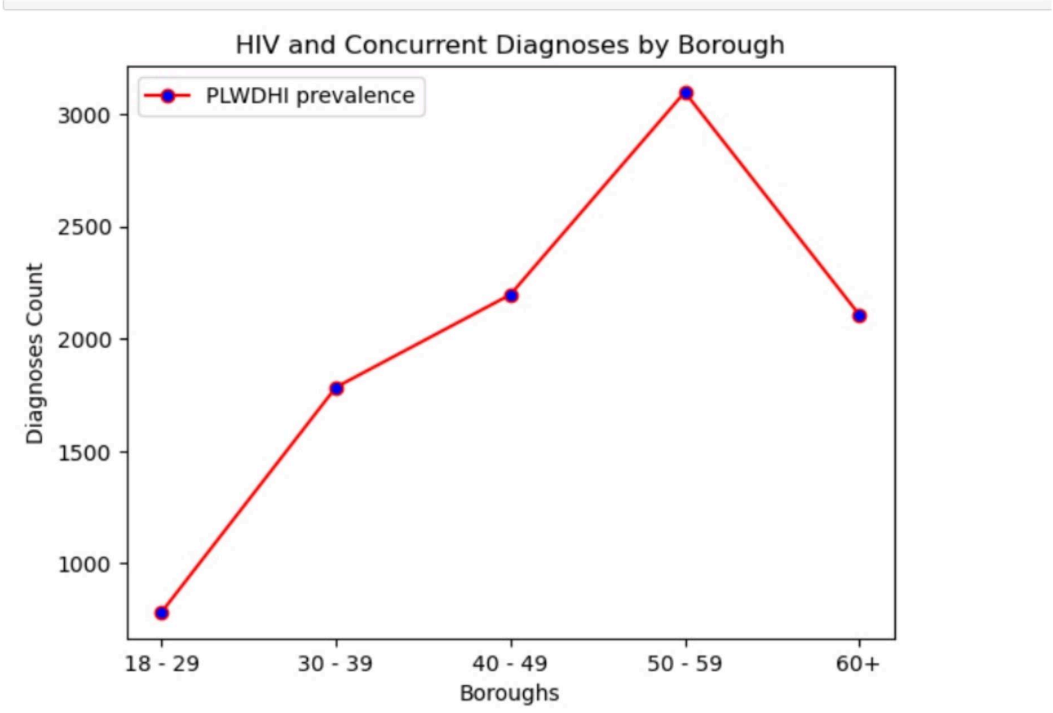


FIGURE 5.10: Analysing Age wise PLWDHI prevalence

6.HIV AIDS PLWDHI PREVALENCE

```
In [71]: p=data.groupby("Age")
         g=p["PLWDHI prevalence"].sum()
         hd=p["HIV diagnosis rate"].sum()
         j=p["AIDS diagnosis rate"].sum()
         ad=p["Non-HIV-related death rate"].sum()
         label =['18 - 29','30 - 39','40 - 49','50 - 59','60+']

         plt.bar(label,hd,color="green",label="HIV diagnosis rate ")
         plt.bar(label,j,color="pink",label="AIDS Death Rate ")
         plt.bar(label,g,color="orange",label="PLWDHI prevalence")
         plt.xlabel("AGE")
         plt.ylabel("HIV AND AIDS DIAGNOSES COUNT")
         plt.legend()
```

FIGURE 5.11: HIV and Aids diagnoses

```
Out[71]: <matplotlib.legend.Legend at 0x20ca10f20d0>
```

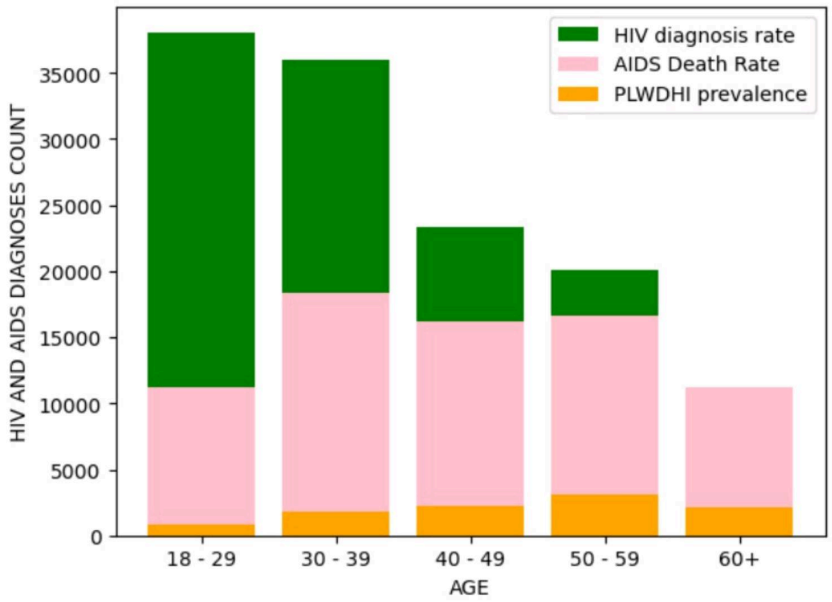


FIGURE 5.12: Analysing HIV and Aids diagnoses

7.Concurrent diagnoses with HIV Diagnoses rate.

```
In [90]: p = data.groupby("Year")

         g = p["Concurrent diagnoses"].sum()
         h = p["HIV diagnosis rate"].sum()
         label =[2017, 2018, 2019, 2020,2021]
         plt.xticks(label)

         plt.barh(label, h, color="purple", label="AIDS diagnoses")
         plt.barh(label, g, color="pink", label="Concurrent diagnoses ")


         # Add labels and title
         plt.xlabel("Diagnoses Count")
         plt.ylabel("Year")
         plt.title("HIV and Concurrent Diagnoses by Borough")

         # Add legend
         plt.legend()

         # Show plot
         plt.show()
```

FIGURE 5.13: Concurrent diagnoses with HIV diagnoses rate

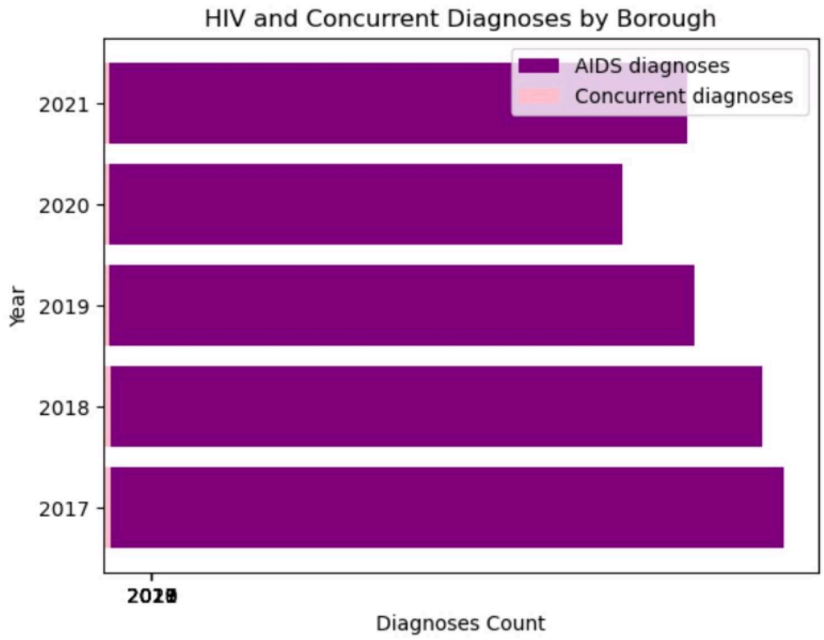


FIGURE 5.14: Analysing Concurrent diagnoses with HIV diagnoses rate

8.PLWDI ,death rate with age

```
In [83]: p=data.groupby("Age")
         g=p["PLWDHI prevalence"].sum()
         hd=p["Death rate"].sum()
         label =['18 - 29','30 - 39','40 - 49','50 - 59','60+']
         plt.bar(label,hd,color="RED",label="Death rate")
         plt.bar(label,g,color="BLUE",label="PLWDHI prevalence")
         plt.ylabel("PLWDHI PREVALENCE VS DEATH RATE")
         plt.legend()
```

FIGURE 5.15: PLWDHI prevalence with death rate among different age groups

```
Out[83]: <matplotlib.legend.Legend at 0x20ca60023d0>
```

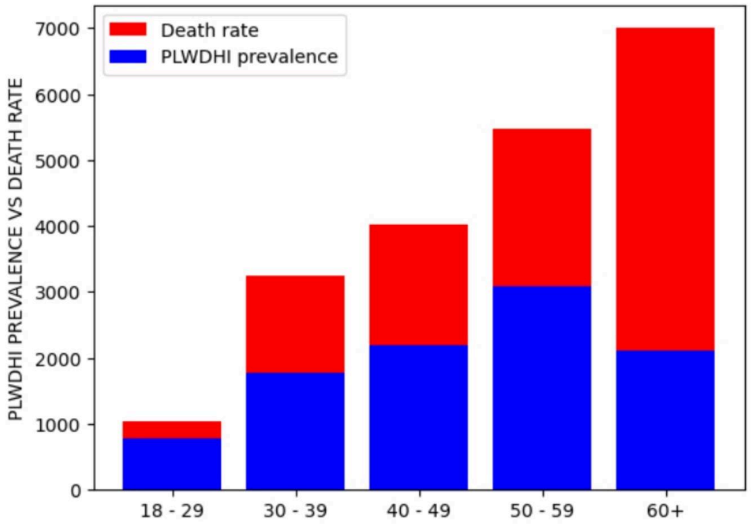


FIGURE 5.16: Analysing PLWDHI prevalence with death rate among different age groups

9.AIDS diagnoses,death rate borough

```
In [86]: p=data.groupby("Borough")
         g=p["AIDS diagnoses"].sum()
         h=p["Death rate"].sum()
         label = ['Bronx', 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island']
         plt.plot(label, g, color="yellow", marker='o', linestyle='-', label="AIDS diagnoses")
         plt.plot(label, h, color="red", marker='o', linestyle='-', label="Death rate")
         # Add labels and title
         plt.xlabel("Boroughs")
         plt.ylabel("Diagnoses Count")
         plt.title("HIV and AIDS Diagnoses by Borough")
         # Add legend
         plt.legend()

         # Show plot
         plt.show()
```

FIGURE 5.17: AIDS Diagnoses with death rates

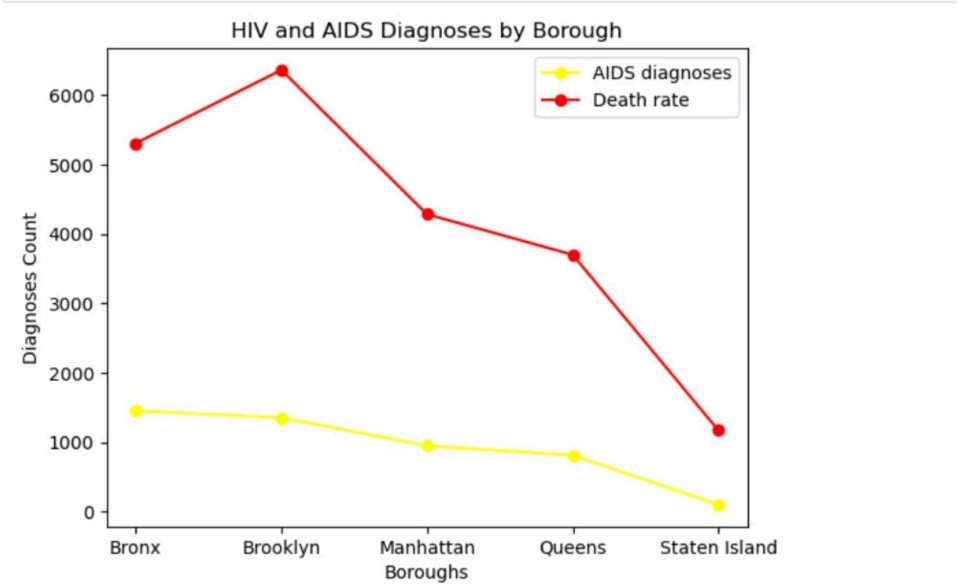


FIGURE 5.18: Analysing AIDS Diagnoses with death rates

10.Concurrent, AIDS diagnoses with borough

In [99]:
```
p = data.groupby("Borough")
g = p["Concurrent diagnoses"].sum()
h = p["AIDS diagnoses"].sum()
#plt.xticks(label)
label = ['Bronx', 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island']
plt.barh(label, h, color="purple", label="AIDS diagnoses")
plt.barh(label, g, color="pink", label="Concurrent diagnoses ")


# Add labels and title
plt.xlabel("Diagnoses Count")
plt.ylabel("Boroughs")
plt.title("AIDS and Concurrent Diagnoses by Borough")

# Add legend
plt.legend()

# Show plot
plt.show()
```

FIGURE 5.19: Concurrent diagnoses with aids diagnoses

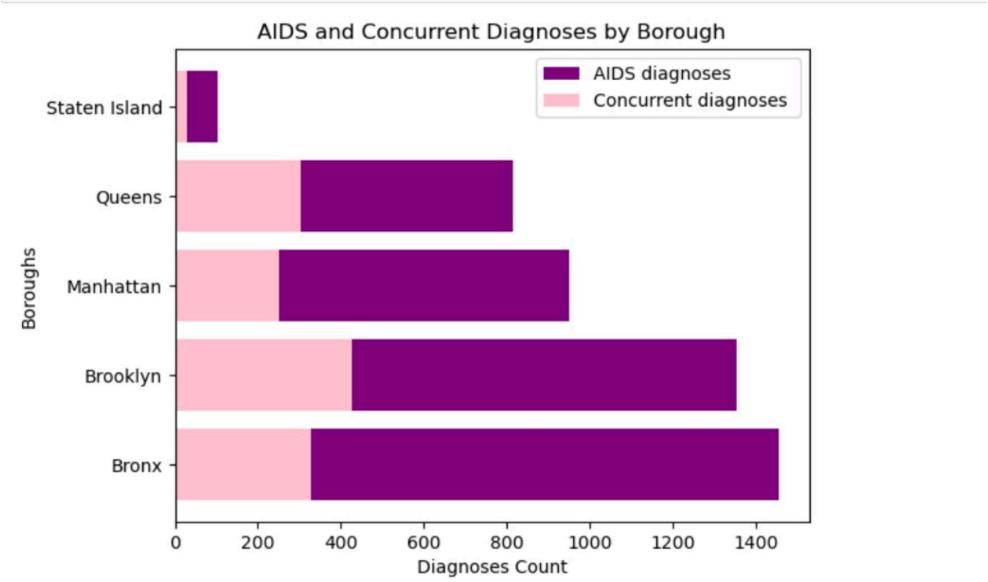


FIGURE 5.20: Analysing Concurrent diagnoses with aids diagnoses

11.HIV Diagnoses with death rate

```
In [101]: p=data.groupby("Race")
          # Example data
          label = ['Asian/Pacific Islander', 'Black', 'Latinx/Hispanic','Other/Unknown', 'White']
          g = p["Death rate"].sum()
          h = p["HIV diagnoses"].sum()
          plt.bar(label,g,color="RED",label="Death rate")
          plt.bar(label,h,color="BLUE",label="HIV diagnoses")
          plt.ylabel("HIV DIAGNOSES VS DEATH RATE")
          plt.xlabel("RACES")
          plt.legend()
```

FIGURE 5.21: HIV diagnoses with death rate

Out[101]: <matplotlib.legend.Legend at 0x20cad477a10>

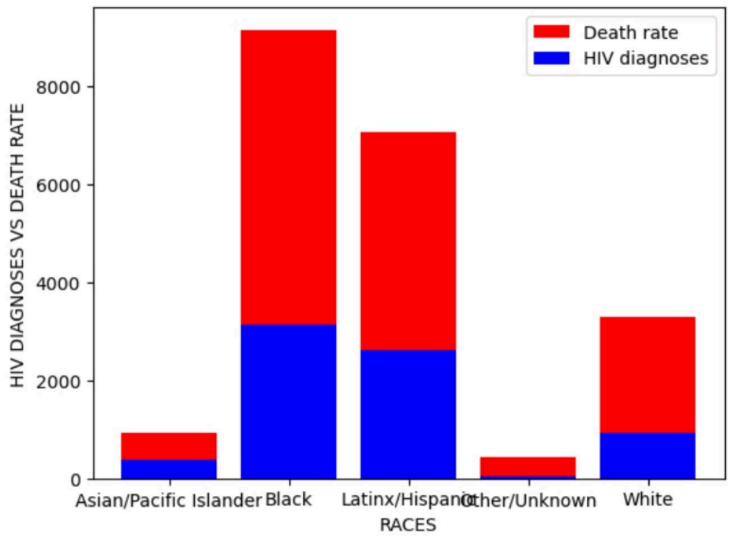


FIGURE 5.22: Analysing HIV diagnoses with death rate

12 HIV Diagnoses with AIDS diagnoses with concurrent diagnoses by age

In [116]:
```
import matplotlib.pyplot as plt
import numpy as np

# Example data
label = ['18 - 29', '30 - 39', '40 - 49', '50 - 59', '60+']
p=data.groupby("Age")
# Assuming p is already grouped and summed
concurrent_diagnoses = p["Concurrent diagnoses"].sum()
aids_diagnoses = p["AIDS diagnoses"].sum()
hiv_diagnoses = p["HIV diagnoses"].sum()

# Define the position of the bars on the x-axis
x = np.arange(len(label))

# Define the width of the bars
width = 0.2

# Create the grouped bar chart
fig, ax = plt.subplots(figsize=(12, 8))

# Plot each set of bars
rects1 = ax.bar(x - width, hiv_diagnoses, width, label='HIV diagnoses', color='orange')
rects2 = ax.bar(x, aids_diagnoses, width, label='AIDS diagnoses', color='purple')
rects3 = ax.bar(x + width, concurrent_diagnoses, width, label='Concurrent diagnoses', color='pink')

# Add some text for labels, title, and custom x-axis tick labels, etc.
ax.set_ylabel('Diagnoses Count')
ax.set_xlabel('Age Groups')
ax.set_title('AIDS, HIV, and Concurrent Diagnoses by Age Group')
ax.set_xticks(x)
ax.set_xticklabels(label)
ax.legend()
plt.show()
```

FIGURE 5.23: HIV Diagnoses with concurrent diagnoses

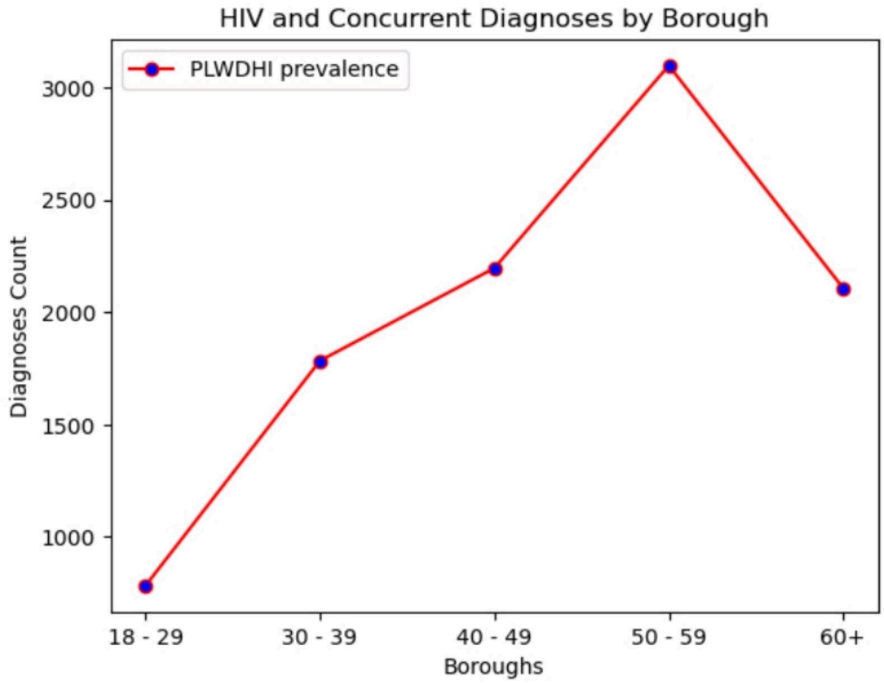


FIGURE 5.24: Analysing HIV Diagnoses with concurrent diagnoses

Chapter 6

Results and Outcomes

Based on the provided dataset , the outcomes or conclusions that can be derived from the analysis are as follows:

- Trends in HIV/AIDS Cases: Determining annual trends in HIV/AIDS diagnoses, exposing intervals of increased and decreased incidence throughout time.
- Correlation with demographic variables: figuring out how different demographic variables, such age, gender, and race, relate to HIV/AIDS cases.
- Age Effect on HIV/AIDS Outcomes: Recognizing how various age groups affect the prevalence, death rates, and diagnoses of HIV/AIDS.
- The impact of gender on the incidence of HIV/AIDS: An analysis of the correlation between gender and HIV/AIDS cases, focusing on variations in diagnosis and treatment results between males, females, and transgender people.
- Function of Race in HIV/AIDS Spread: Examining how racial and ethnic backgrounds affect HIV/AIDS prevalence and transmission.
- HIV/AIDS Geographic Distribution: An analysis of the geographic dispersion of HIV/AIDS cases among various
- The geographic distribution of HIV/AIDS cases is examined to find high-risk locations by looking at how cases are distributed across various boroughs and UHF neighborhoods.
- Temporal Trends and Patterns: Time series analysis of HIV/AIDS diagnoses, treatment outcomes, and mortality rates throughout time is used to identify trends and periodicity.

- Results and Healthcare Access: To determine treatment effectiveness and healthcare accessibility across various demographic groups and geographic areas, evaluations of viral suppression rates and three-month linkage to care are conducted.

- Analysis of death rates: This includes a look at total, HIV-related, and non-HIV-related death rates. It also looks at characteristics that are linked to greater mortality and how they differ amongst other populations.

- Machine Learning Predictive Models: Using demographic and geographic data to forecast HIV/AIDS cases, predictive models are built to help with early identification and preparedness for possible epidemics.

- Finding High-Risk Areas: Using spatial analysis, one may locate HIV/AIDS hotspots as well as places with high-risk geographic and demographic characteristics.

- Long-Term Variations in HIV/AIDS Incidence: An examination of annual patterns is necessary to identify any long-term variations in HIV/AIDS cases that may be related to shifts in the social, economic, or medical landscape.

- Data-Driven Insights: Using data-driven insights to inform policymakers' and healthcare professionals' evidence-based decision-making will result in more successful HIV/AIDS prevention tactics.

- Gain an understanding of healthcare access and outcomes, including how timely access to care and viral suppression rates affect overall health outcomes and the course of disease.

- Examining the effects of age, gender, and race on HIV/AIDS incidence and outcomes in order to provide vulnerable populations with tailored therapies.

- Keep in mind that these outcomes are speculative, and the actual findings may vary depending on the data quality, analysis techniques, and expertise involved in the study. Thorough analysis, validation, and interpretation with domain experts are essential to derive meaningful conclusions from the dataset.

Conclusions

The problem statement for our analysis was to comprehensively understand the trends, patterns, and determinants of HIV/AIDS in New York City across different boroughs, UHF neighborhoods, demographics (age, gender, race), and over time.The objectives of this study were to identify high-risk regions and people, assess the efficacy of healthcare treatments, and offer data-driven insights to healthcare professionals and policymakers. The solution to this issue is essential for developing public health policies that work, allocating resources effectively, and eventually lowering the incidence and mortality rates of HIV/AIDS in the city.

To address this problem, we implemented an exploratory data analysis framework.To manage any missing values and guarantee data consistency, we started by importing and cleaning the dataset. After gaining some preliminary insights through summary statistics, we employed a variety of visualization strategies to comprehend the distribution and connections among distinct variables. Our identification of high-risk areas and populations was aided by demographic and geographic analysis. To find trends and patterns throughout time, we also performed temporal analysis. Furthermore, we developed prediction models utilizing machine learning algorithms to estimate HIV/AIDS cases and assess the influence of various determinants on health outcomes.

Our analysis of the HIV/AIDS dataset for New York City revealed critical insights. A borough-level analysis revealed that the Bronx and Manhattan had greater rates of HIV diagnosis; certain UHF neighborhoods had particularly high prevalence, which led to the development of focused interventions. Males, African American, and Hispanic populations were disproportionately affected, while younger persons (years 20–29) and older adults (ages 50–59) demonstrated greater diagnosis rates. These findings urge for concentrated efforts in prevention and therapy.Although some periods showed rises,

highlighting the need for ongoing surveillance, temporal trends revealed a general decline in new HIV diagnoses, demonstrating effective prevention methods. More people were connected to care within three months of diagnosis, demonstrating an improvement in healthcare access, and there was an increase in viral suppression rates, a sign of successful treatment. Disparities in healthcare outcomes between various demographic groups, however, indicated areas that require improved access to and assistance from healthcare. An examination of mortality revealed a drop in both HIV-related and general death rates, pointing to improved disease control. Higher death rates in particular populations and regions, however, highlighted the necessity of targeted healthcare initiatives. Policymakers and medical professionals can create focused, evidence-based plans for efficient HIV/AIDS prevention, treatment, and resource distribution with the help of these studies.

Bibliography