Fahim Ahamed | Anthony Sinopoli | Fatima Nasyr

Summary

The project examines how HIV and AIDS diagnoses have changed in New York City from 2016 to 2021. Using data from NYC Open Data, it explores patterns across time, geography, and demographic groups to identify disparities in diagnosis rates and disease progression. The analysis focuses on variations among the five boroughs and across gender and racial or ethnic groups.

By highlighting which populations or areas experience higher rates, the study aims to provide insights that can guide public health strategies such as awareness programs, prevention efforts, and improved access to care and testing.

The exploratory data analysis (EDA) focuses on identifying key trends, distributions, and patterns in the data. Using descriptive statistics and visualizations, it highlights how HIV and AIDS diagnoses have shifted across New York City over time and across populations. The goal was to develop a clear, data-driven understanding of the overall landscape before moving into more detailed analyses.

About the Data

The dataset used in this project was obtained from **NYC Open Data**, which provides publicly available health statistics compiled by the New York City Department of Health and Mental Hygiene (DOHMH). It contains yearly records of HIV and AIDS diagnoses among New York City residents from **2010 to 2021**, excluding **2014 and 2015**, for which data were not reported.

Each record represents a specific combination of **year, borough, neighborhood, sex, and race/ethnicity**, along with multiple indicators describing both counts and rates. These include:

- Total number of **HIV diagnoses** and corresponding rates per 100,000 population
- Number and proportion of **concurrent HIV/AIDS diagnoses** (individuals diagnosed with both at the same time)
- Total number of **AIDS diagnoses** and corresponding rates per 100,000 population

More details about the dataset's features and definitions are available on the official **NYC Open Data page**.

Data Preparation

The raw dataset from NYC Open Data required several cleaning and transformation steps before exploratory analysis. First, the data were imported and inspected for duplicate entries, missing values, and inconsistent formatting across variables. All numeric columns originally stored as text were converted to numeric types for accurate computation.

To make the data consistent and analysis-ready:

- Duplicated rows were identified and removed, keeping only unique combinations of **year**, **borough**, **neighborhood**, **sex**, and **race/ethnicity**.

- Line breaks and special characters were removed from categorical fields such as **RACE/ETHNICITY** and **Neighborhood (U.H.F)**.

- Rows with missing or suppressed values (shown as “NA”) were removed to ensure the accuracy of summaries and visualizations.

- Data were filtered to include relevant combinations of attributes, such as cases where **borough**, **neighborhood**, **sex**, and **race/ethnicity** were reported as “All,” depending on the analysis focus.

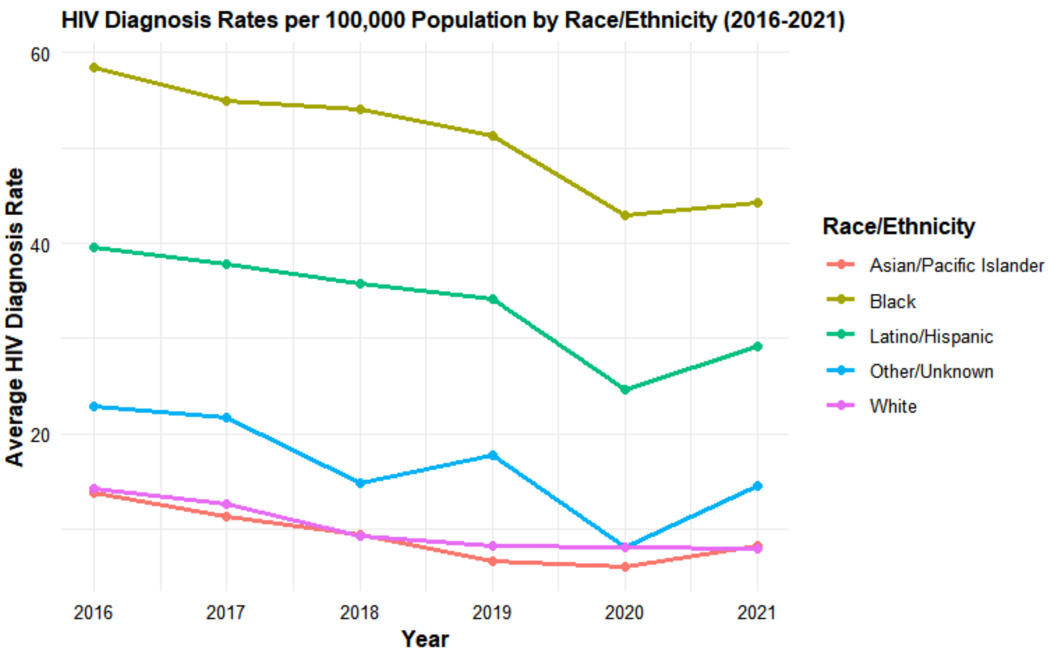
Exploratory Data Analysis

At the start of the analysis, we asked ourselves a key question: ***“What do we want to learn from the data?”***

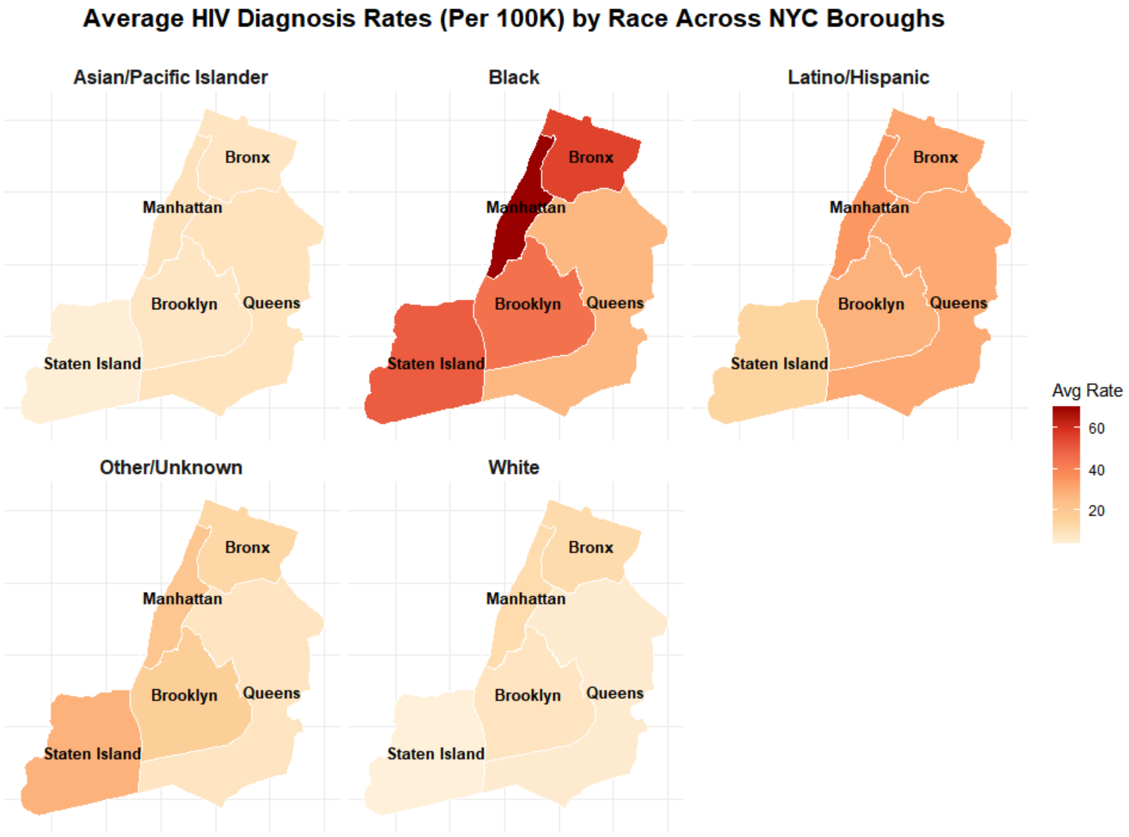
The goal of the EDA was to explore how HIV and AIDS diagnoses vary across New York City by examining the data from three different angles: yearly trends, geographic patterns (borough and neighborhood), and demographic factors (sex and race/ethnicity).

Although the dataset spans 2010-2021, most visual analyses focus on 2016-2021 due to missing or incomplete data in earlier years. This ensures consistent comparisons across boroughs and demographic groups.

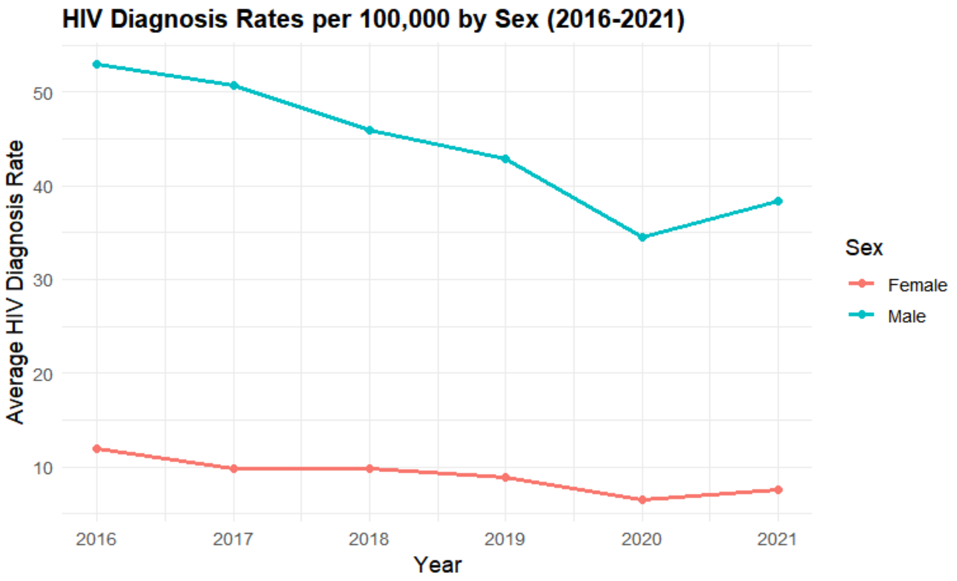
We wanted to see how these three views of the data relate to each other and whether changes over the years also appear across different areas or groups. Looking at the data this way helped us notice both the overall trends and the differences among specific populations.



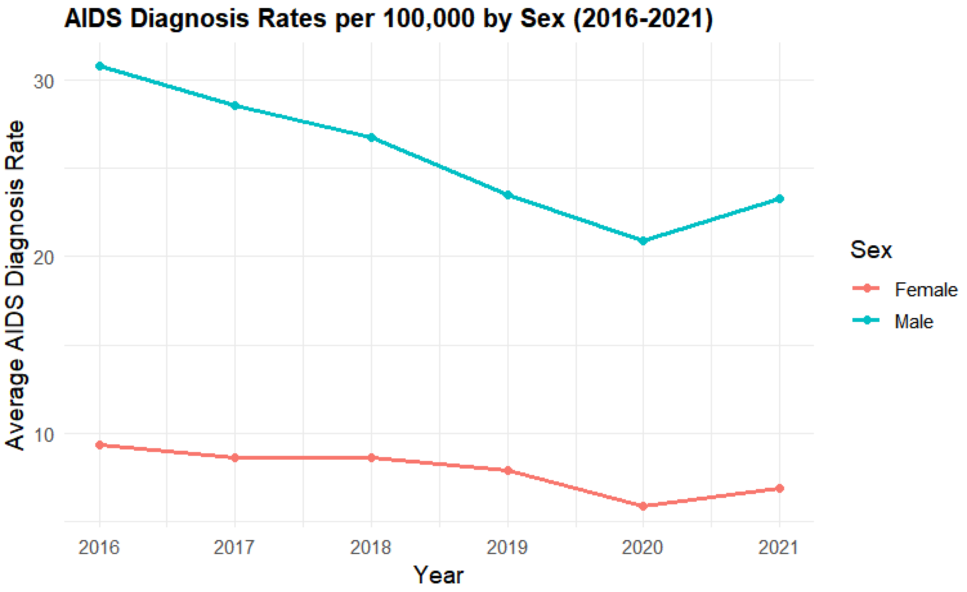
Across 2016-2021, the lines show clear differences in HIV rates by race/ethnicity. Groups at the top of the plot maintain higher burdens throughout the period, while lower lines indicate consistently smaller rates. Year-to-year changes are visible as small rises/dips; the relative ordering between groups remains the main pattern to note.



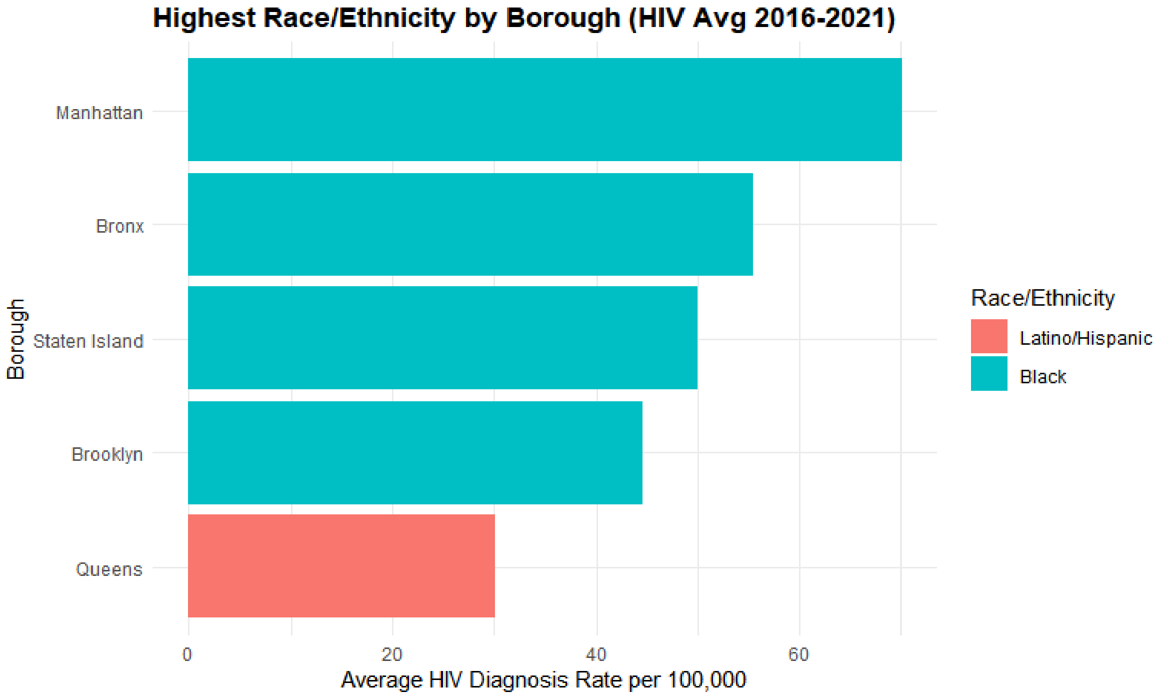
The map plot shows strong racial differences in HIV diagnosis rates across NYC. Black residents have the highest rates in every borough, especially in the Bronx, Manhattan, and Brooklyn. Latino/Hispanic residents show moderately elevated rates, also highest in the Bronx and Manhattan. Asian/Pacific Islander and White groups consistently have the lowest rates, with minimal variation across boroughs. Overall, the visualization highlights a clear disparity, with Black and Latino/Hispanic populations experiencing a disproportionately higher HIV burden across the city.



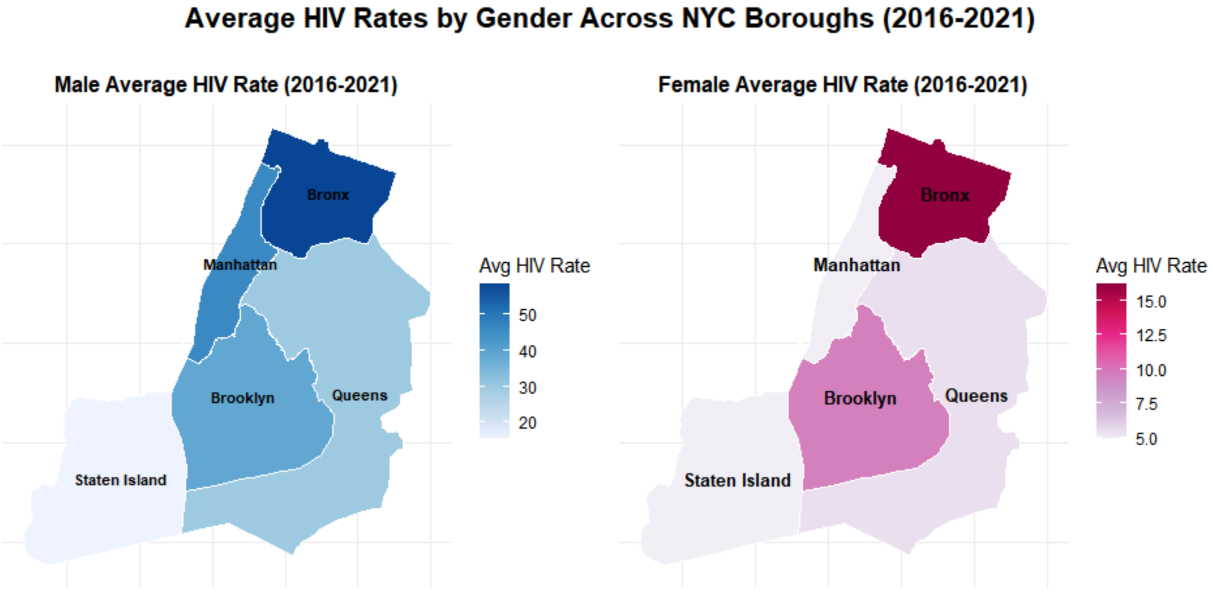
The sex-based lines reveal how HIV rates change year by year for Females and Males. The higher line indicates which sex carries more burden in a given year; the gap shows the size of difference. Parallel movement suggests similar trends; diverging lines indicate changing differences over time.



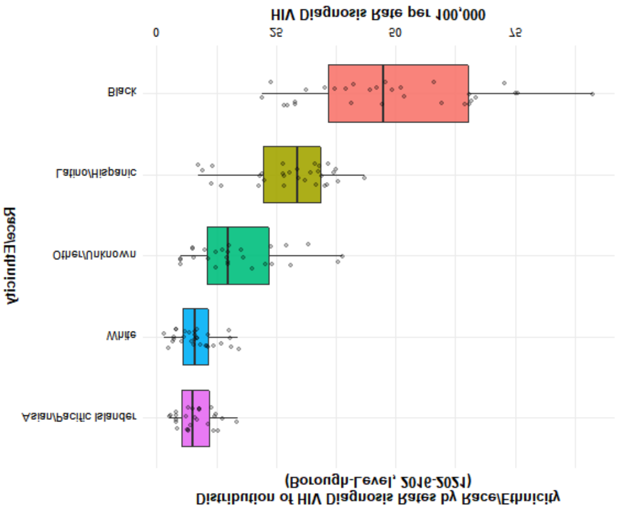
AIDS rates by sex follow their own paths across the period. Compare the distance between the lines to assess the sex gap and whether it widens or narrows. Differences from the HIV trend can suggest variation in progression from HIV to AIDS.

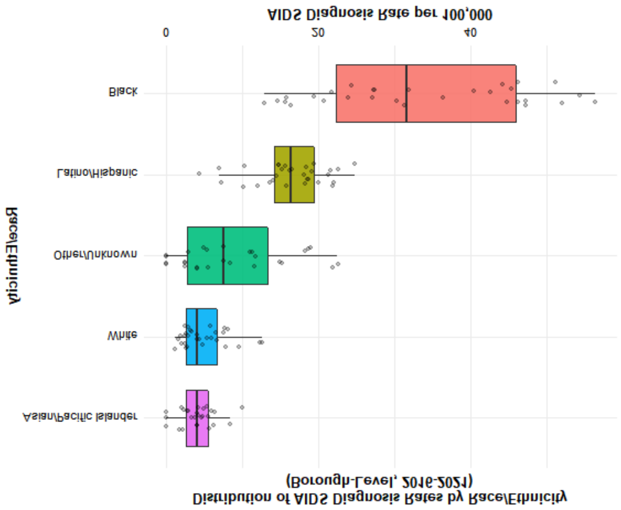


From 2016 to 2021, Black residents had the highest average HIV diagnosis rates in Manhattan, the Bronx, Staten Island, and Brooklyn, while Latino/Hispanic residents had the highest rate in Queens. The chart highlights clear borough-level differences, showing that the highest HIV rates are concentrated within specific demographic groups in each borough.

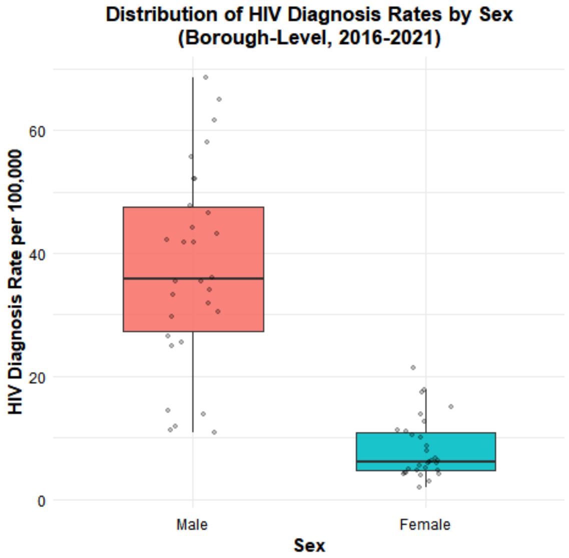


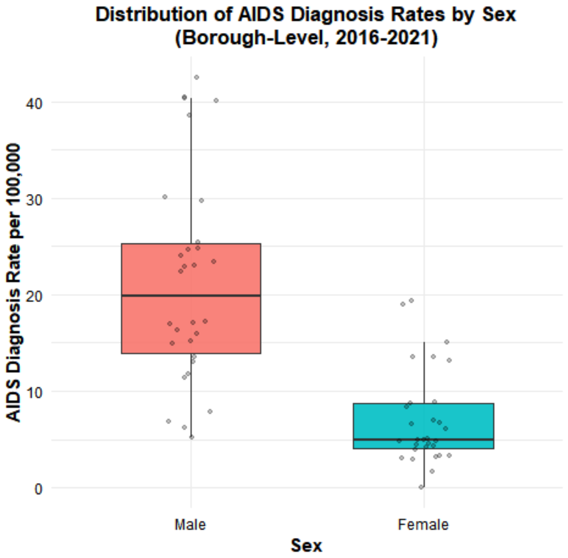
Male HIV rates are consistently higher than female rates across all NYC boroughs. The Bronx shows the highest rates for both men and women, followed by Manhattan and Brooklyn, indicating these areas carry the greatest HIV burden. Queens and Staten Island have the lowest average rates for both genders. This shows a clear gender disparity, with men experiencing substantially higher HIV rates while geographic differences across boroughs remain similar for both groups.



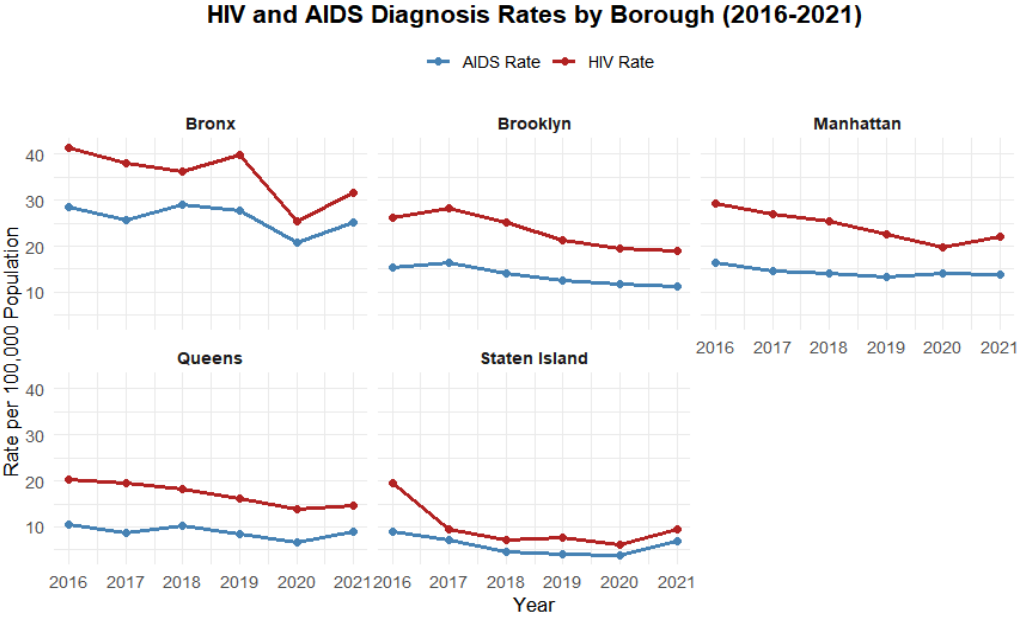


The boxplots reveal clear disparities across race/ethnicity groups. Groups with higher medians and wider IQRs show both greater overall burden and more variability across boroughs. Comparing HIV vs AIDS panels shows whether the same groups lead on both measures or if progression patterns differ.

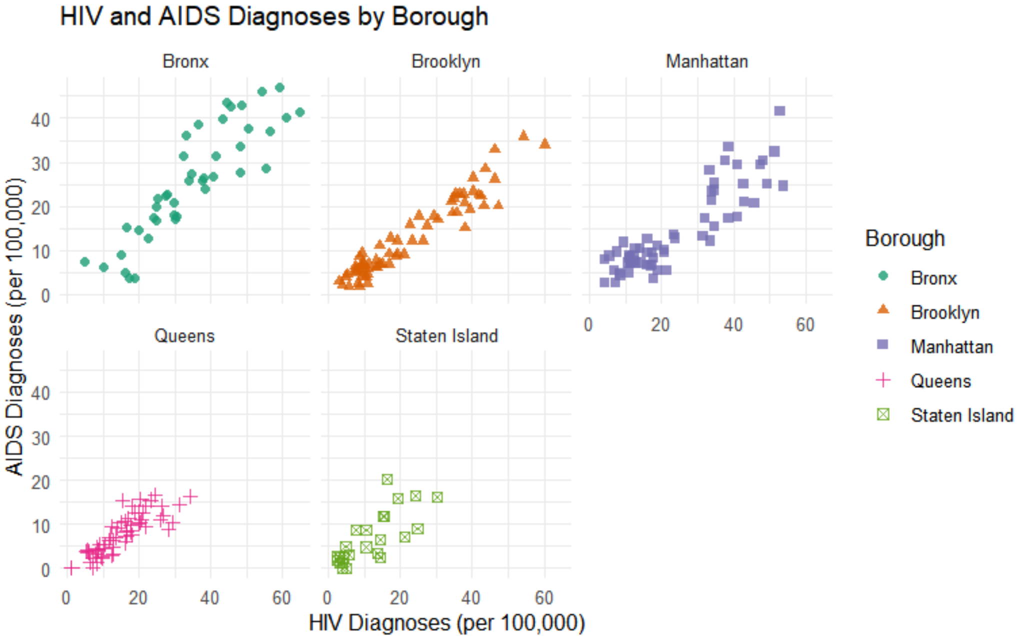




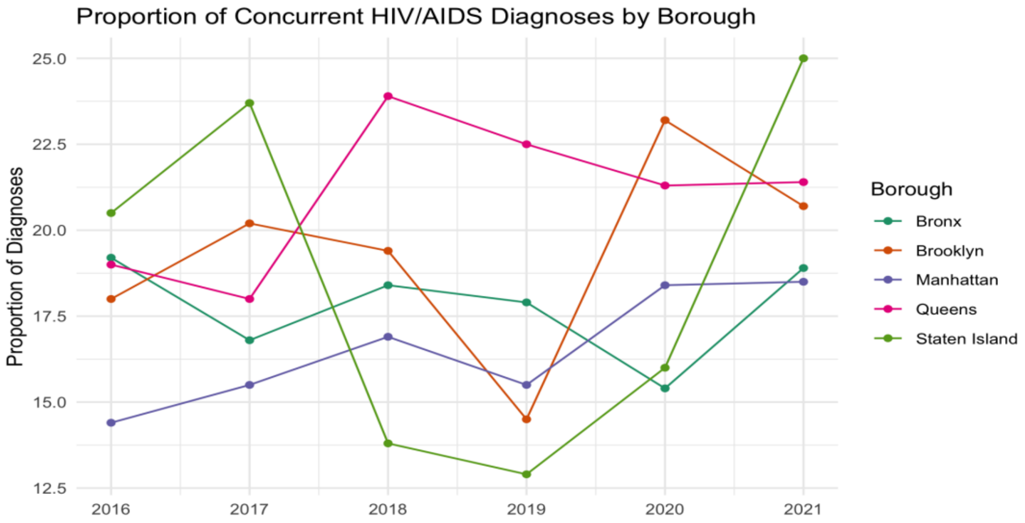
Distributions by sex show median differences and spread across boroughs. Wider boxes/longer whiskers signal greater variability across borough-year observations. Side-by-side HIV vs AIDS panels help assess whether the sex gap is consistent at diagnosis and at AIDS stage.



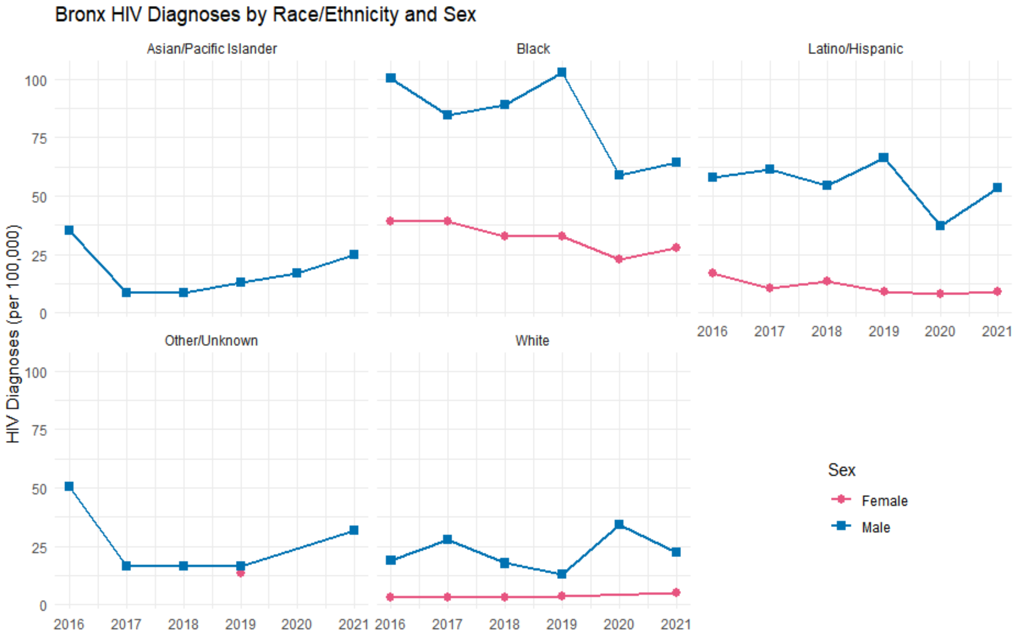
Overall, a general decline in HIV diagnoses until 2020, where a spike can be seen from 2020-2021 (except in Brooklyn). The Bronx is the borough with the highest diagnoses, and Staten Island with the lowest. Again, the Bronx has the highest AIDS diagnoses with a spike from 2020-2021. Manhattan has been relatively stable, and Brooklyn has even seen a slight decline. Queens and Staten Island have the lowest diagnoses, but have seen an increase from 2020-2021.



The plot shows a clear positive relationship between HIV and AIDS diagnosis rates across all five boroughs, meaning that areas with higher HIV rates also tend to have higher AIDS rates. The Bronx stands out with the highest overall rates, while Queens and Staten Island have noticeably lower values. This suggests that the disease burden is not evenly distributed across the city.

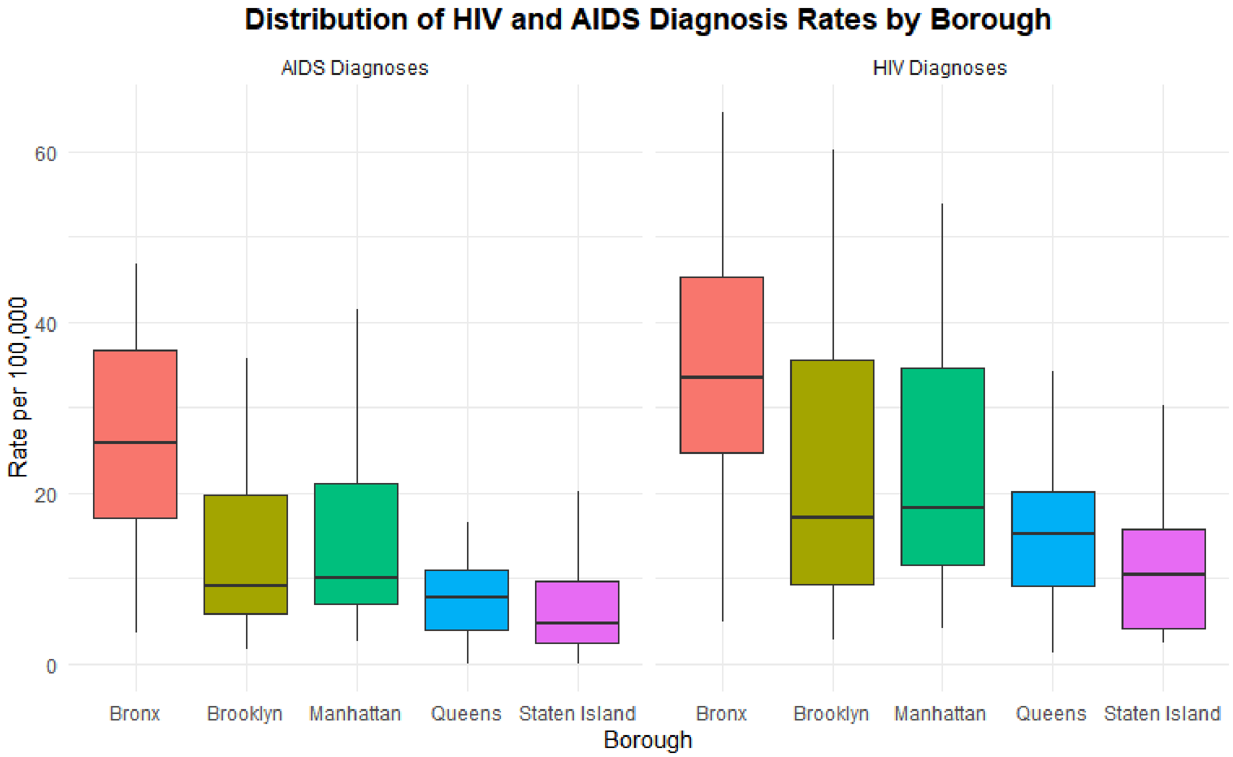


The proportion of concurrent HIV/Aids diagnoses is a measure of late diagnoses of HIV, where the patient is diagnosed for HIV, then AIDS at the same time or shortly after. Manhattan has seen an increase in late diagnoses.



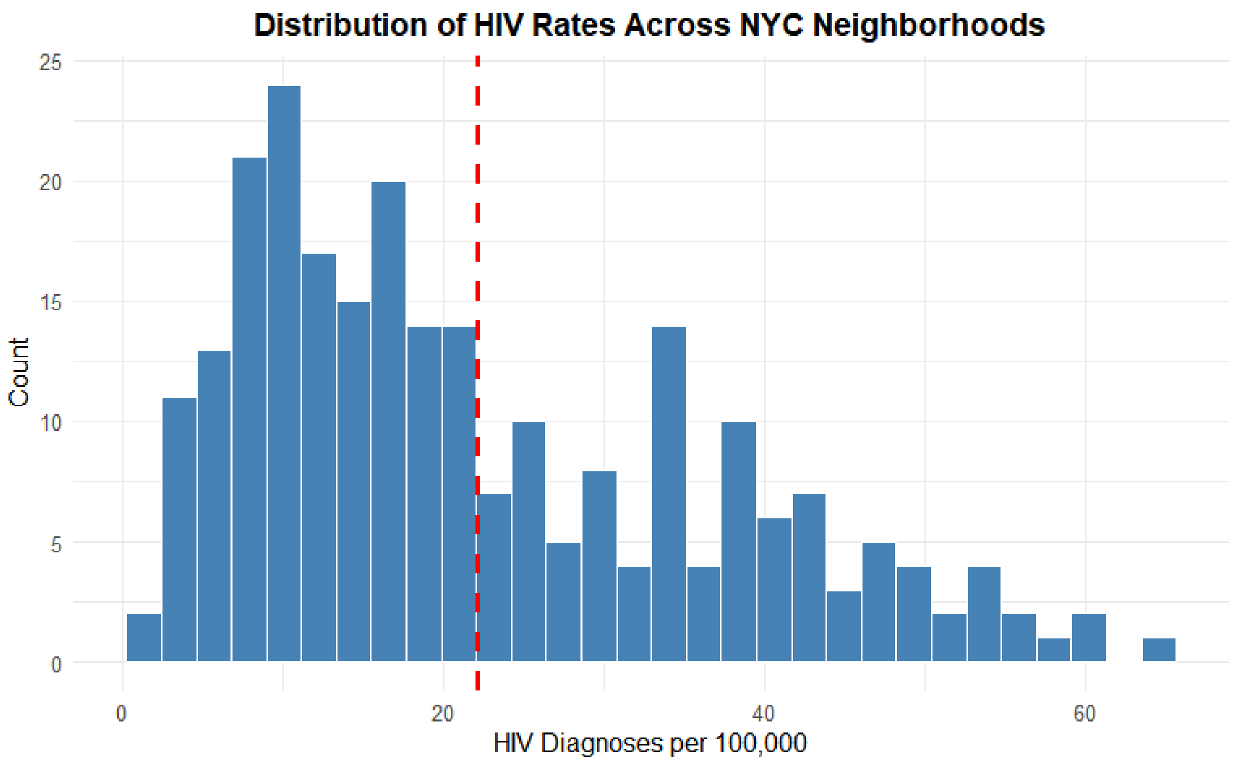
Taking a more in-depth look at the Bronx, males have a higher rate of HIV diagnoses than females, when there is data (missing some female data). Black males have the highest overall HIV diagnoses, with Latino/Hispanic males having the second most. Males in the other races have similar diagnoses to one another.

Note: Some female data points (Asian and Other) are missing because their reported values were either zero or marked as NA in the dataset. The Other/Unknown race female has one data point, which is for 2019.



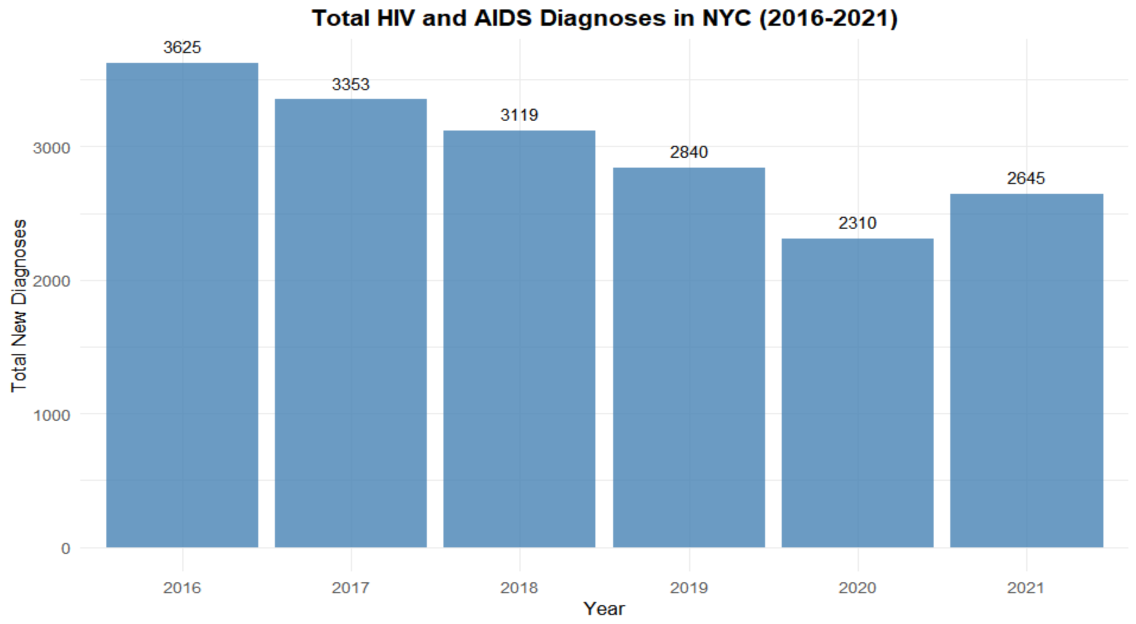
This boxplot compares the distribution of HIV and AIDS diagnosis rates across New York City's five boroughs. In both panels, the Bronx stands out with the highest median rates and widest spread, indicating higher and more variable neighborhood-level burden. Brooklyn and Manhattan follow with moderately high and overlapping ranges, while Queens and Staten Island consistently show lower rates and narrower distributions.

The pattern suggests that HIV and AIDS diagnoses are not evenly distributed across the city. The Bronx remains the most affected borough, while outer boroughs like Queens and Staten Island experience relatively lower and more stable rates.

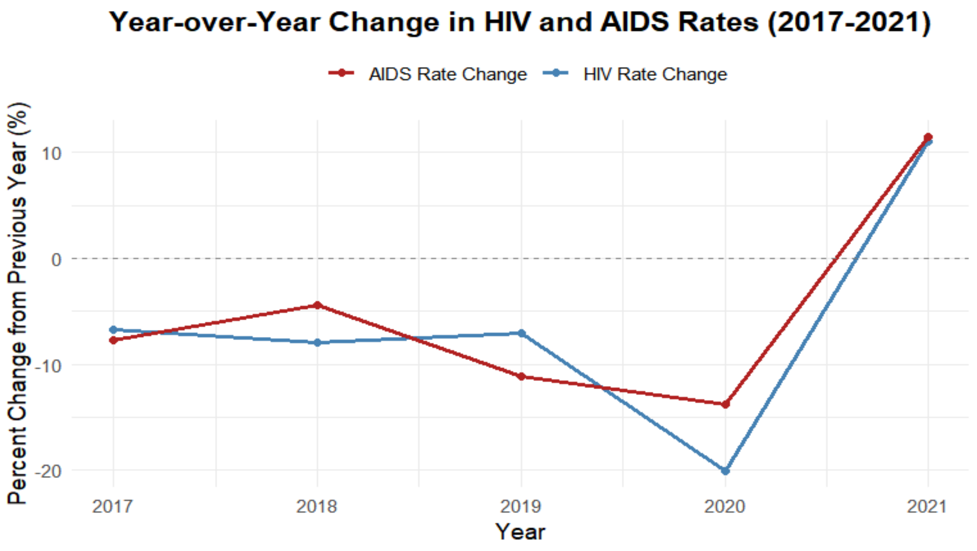


This histogram shows how HIV diagnosis rates are distributed across New York City neighborhoods. Most neighborhoods have relatively low rates, clustered below the citywide average. The distribution is right-skewed, meaning only a few neighborhoods experience much higher rates, which raises the overall mean.

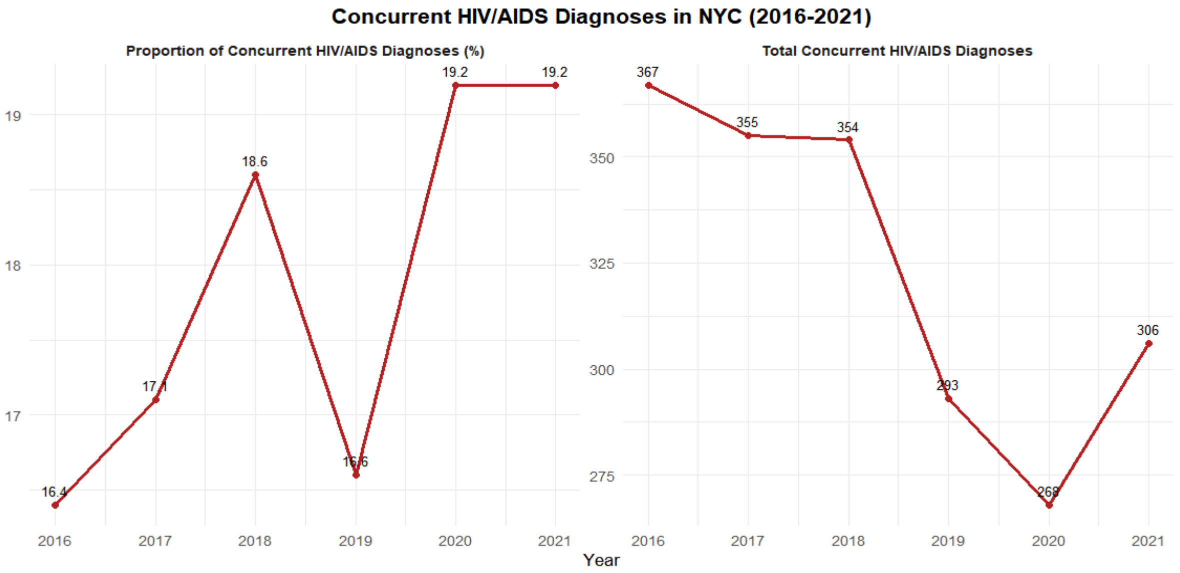
This pattern further confirms that the HIV burden in NYC is unevenly distributed, concentrated in a limited number of high-incidence neighborhoods while the majority of areas maintain lower rates. It highlights persistent geographic disparities, where targeted prevention and testing efforts may be most needed in those higher-rate communities.



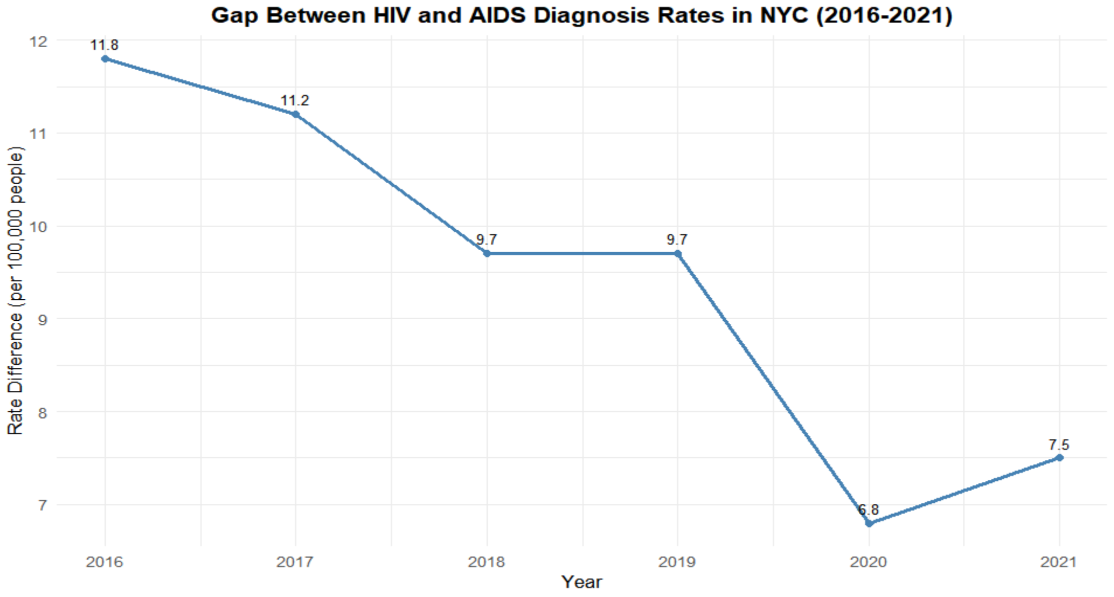
This chart shows the total burden of HIV and AIDS in New York City from 2016 to 2021, measured by the number of new diagnoses each year. The overall downward trend indicates that fewer people are entering the HIV care system, reflecting progress in prevention, early detection, and long-term treatment efforts. The dip in 2020 suggests a temporary decline in testing or reporting during the pandemic, while the slight increase in 2021 points to a return to normal healthcare activity. Overall, the data highlights a steady reduction in the city's HIV/AIDS burden over time.



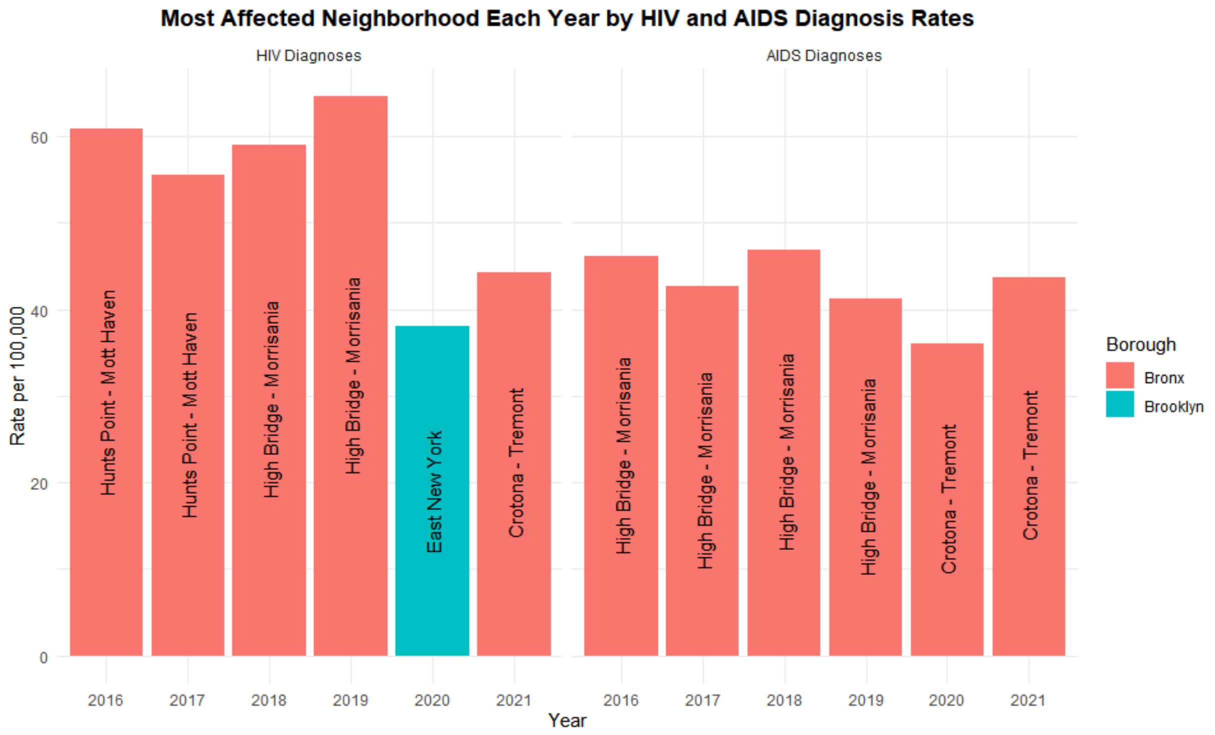
This chart shows the year-over-year percentage change in HIV and AIDS diagnosis rates in New York City from 2017 to 2021. Both rates declined consistently through 2019, reflecting steady progress in reducing new infections. A sharp drop in 2020 suggests significant disruptions in testing and care access during the pandemic. The strong rebound in 2021 likely represents resumed testing and diagnosis rather than a true increase in new cases. Overall, the pattern shows long-term improvement with short-term fluctuations linked to external healthcare disruptions.



This figure shows how often people in New York City were diagnosed with HIV and AIDS at the same time between 2016 and 2021. The left chart shows the share of new HIV cases that were already AIDS at diagnosis, while the right chart shows the total number of such cases each year. Together, these trends suggest that while the number of concurrent cases dropped, the share of HIV diagnoses that were already AIDS at detection grew slightly. This means fewer people were diagnosed overall, but a larger portion of them had advanced infection, indicating slower or delayed testing in recent years. The small rise in both metrics in 2021 likely reflects the rebound of HIV testing and diagnosis after the pandemic's disruptions.



This chart shows the difference between HIV and AIDS diagnosis rates in New York City from 2016 to 2021. The gap between the two rates narrows over time. The shrinking gap aligns with the earlier finding that a larger share of new HIV diagnoses were already at an advanced stage, suggesting the two measures are moving closer together over time.



Across the years, the Bronx dominates as the borough with the most affected neighborhoods for both HIV and AIDS diagnoses. High Bridge-Morrisania and Crotona-Tremont appear repeatedly, showing that certain Bronx neighborhoods continue to carry the highest burden. Brooklyn's East New York is the only neighborhood outside the Bronx that stands out, appearing once in 2020 for HIV diagnoses. This pattern suggests that while rates vary slightly by year, the Bronx consistently faces the highest concentration of new and advanced HIV cases citywide.

Summary of Key Findings

Overall, the analysis shows that HIV and AIDS diagnoses in New York City have declined over time, reflecting progress in prevention and care, though disparities remain across groups and locations. The Bronx continues to show the highest rates, while Queens and Staten Island report the lowest. Black and Latino/Hispanic residents experience the greatest overall burden, and males consistently have higher diagnosis rates than females. The temporary dip in 2020 and rebound in 2021 likely reflect testing disruptions during the pandemic. Together, these findings highlight steady citywide improvement but persistent demographic and geographic gaps that remain important for public health planning

Hypothesis Testing

1. Was the 2020 HIV Diagnosis Total Significantly Lower Than Expected?

To assess whether the drop in HIV diagnoses observed in 2020 was unusually large compared to pre-pandemic levels, a one-sample t-test was performed. The goal is to compare the 2020 total number of HIV diagnoses to the average annual total from the four pre-COVID years (2016 to 2019).

Hypotheses

Let μ represent the true average annual number of HIV diagnoses in the pre-COVID period.

- **H_0:** μ = HIV_2020
  The 2020 total is consistent with the pre-COVID average.
- **H_1:** μ > HIV_2020
  The pre-COVID average is higher than the 2020 value, indicating a meaningful decline in 2020.

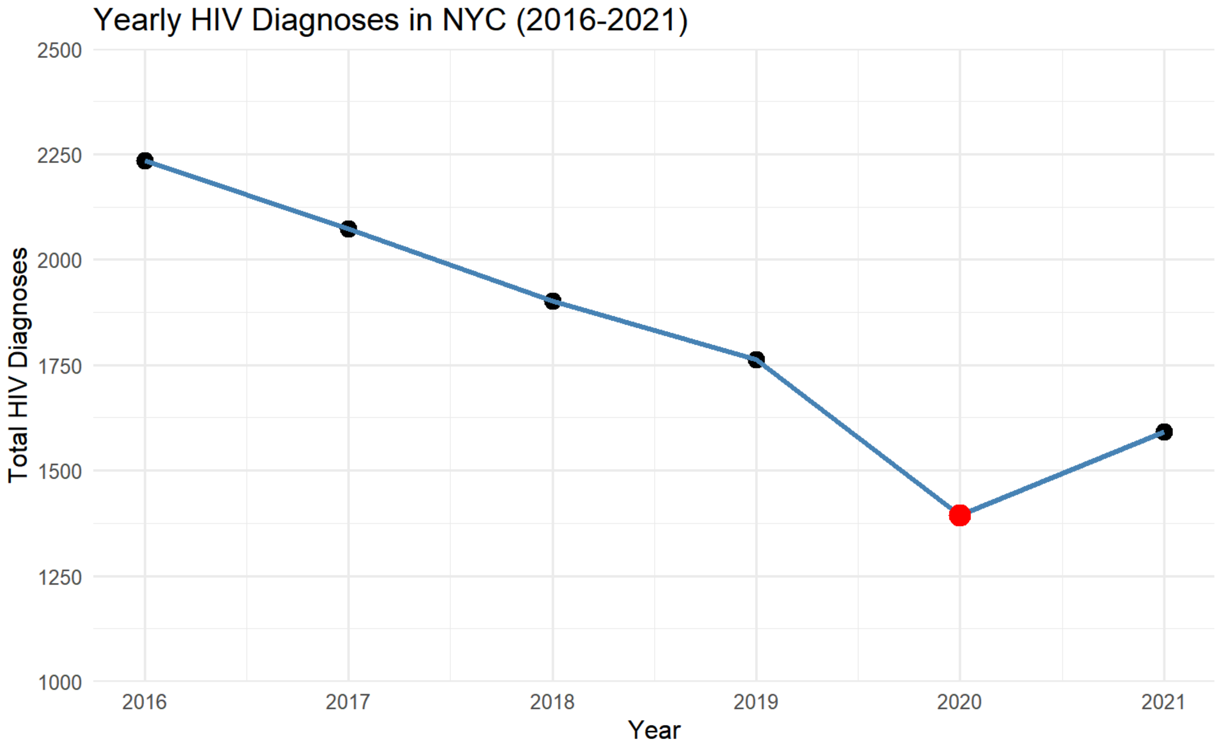
Assumptions

- The yearly totals from 2016 to 2019 are independent.
- The sample size is small (n = 4), so the t-distribution is appropriate.
- No extreme outliers are present among the pre-COVID yearly totals.

```
##  One Sample t-test
## 
## data:  pre_2020
## t = 5.8316, df = 3, p-value = 0.005022
## alternative hypothesis: true mean is greater than 1394
## 95 percent confidence interval:
##  1751.273      Inf
## sample estimates:
## mean of x 
##      1993
```

The one-sample t-test indicates that HIV diagnoses in 2020 were much lower than what we typically saw before the pandemic. The average number of cases from 2016-2019 was clearly higher than the 2020 value of 1,394, and the p-value (0.005) shows that this drop is unlikely to be due to normal year-to-year variation.

Based on this result, we conclude that the decline in 2020 was statistically significant and likely reflects the impact of COVID-19 on HIV testing and reporting.



2. Are Bronx HIV Rates Consistently Higher Than the NYC Average?

To examine whether the Bronx consistently reports higher HIV diagnosis rates than the overall New York City average, a paired t-test was conducted. For each year from 2016 to 2021, the Bronx HIV rate was matched with the corresponding citywide rate, creating yearly paired observations.

Hypotheses

Let d represent the difference between the Bronx HIV rate and the NYC-wide rate for each year.

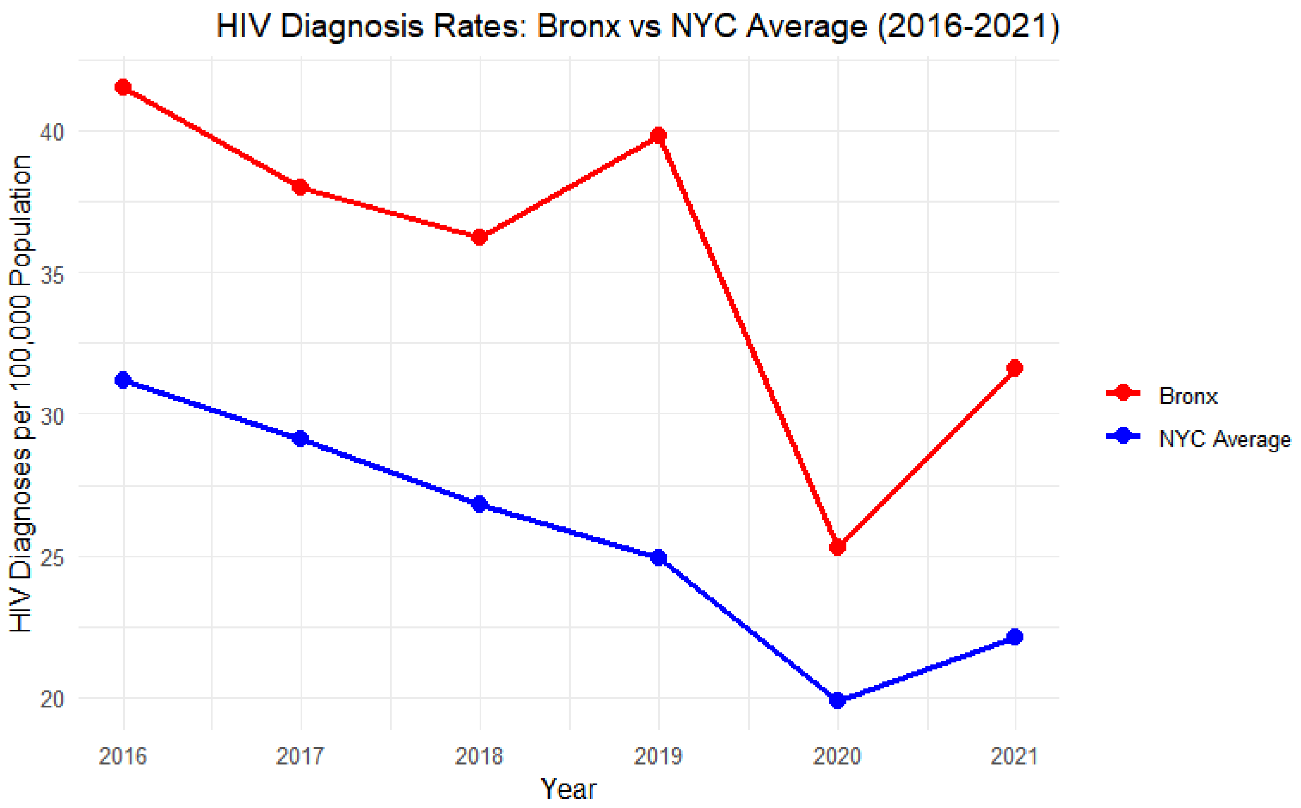
- **H_0:** mean(d) = 0
  On average, Bronx HIV rates are equal to the citywide rates.
- **H_1:** mean(d) > 0
  Bronx HIV rates are higher on average, indicating a consistent elevation relative to the NYC average.

Assumptions

- Each year's Bronx-NYC rate difference is independent.
- The distribution of the differences is approximately normal.
- Each year forms a valid matched pair for comparison.

```
##  Paired t-test
##
## data:  bronx_rates and nyc_rates
## t = 7.8051, df = 5, p-value = 0.0002766
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  7.220486      Inf
## sample estimates:
## mean difference
##        9.733333
```

The paired t-test shows that HIV diagnosis rates in the Bronx were consistently higher than the NYC-wide rate from 2016 to 2021. On average, the Bronx rate was about 9.7 cases per 100,000 above the citywide level. The test result (t = 7.81, p = 0.00028) and the one-sided 95% confidence interval (7.22, ∞) both indicate a clear and statistically significant difference. Overall, the evidence strongly supports that the Bronx remains a higher-burden area compared to the rest of New York City.



3. Does Queens have a higher proportion of concurrent HIV/AIDS diagnoses than the high-burden boroughs, even though it has fewer total HIV cases??

To examine whether Queens experiences higher late-stage HIV diagnoses despite having fewer total HIV cases, a two-sample t-test was conducted. Each year from 2016 to 2021, the concurrent diagnosis proportion for Queens was compared to the combined concurrent proportions of the high-burden boroughs (Bronx, Brooklyn, and Manhattan). This produced six yearly observations for Queens and eighteen for the high-burden group.

Hypotheses

Let μ_Q represent the mean concurrent HIV/AIDS diagnosis proportion in **Queens**, and μ_H represent the mean concurrent diagnosis proportion in the **high-burden boroughs**.

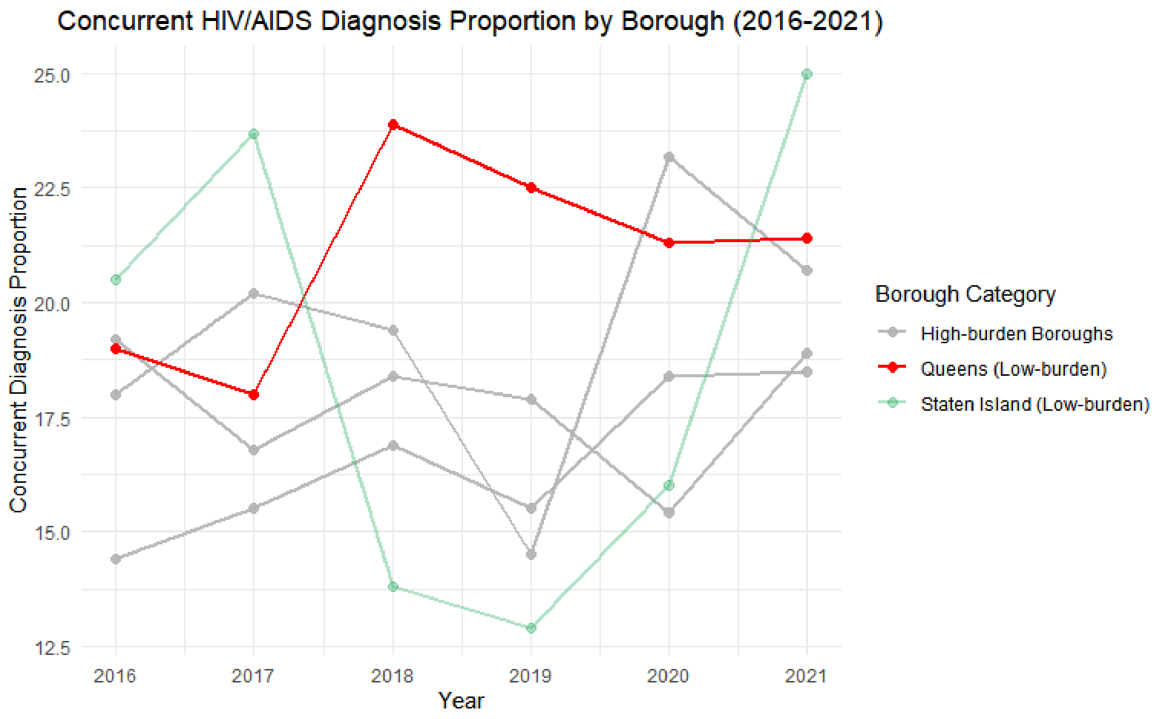
- **H_0:** μ_Q ≤ μ_H
  On average, Queens has concurrent proportions that are the same or lower than the high-burden boroughs.
- **H_1:** μ_Q > μ_H
  Queens has a higher concurrent proportion, indicating more frequent late-stage diagnoses despite being a low-burden borough.

Assumptions

- Borough-year observations are independent.
- The yearly concurrent proportions for Queens (n = 6) and the high-burden boroughs (n = 18) are approximately normally distributed.
- Unequal sample sizes and variances are acceptable under Welch’s t-test.

```
##  Welch Two Sample t-test
##
## data:  queens_vals and high_vals
## t = 3.0014, df = 9.0481, p-value = 0.007417
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.222967      Inf
## sample estimates:
## mean of x mean of y
##  21.01667  17.87778
```

The one-sided t-test shows strong evidence that Queens has a higher proportion of concurrent HIV/AIDS diagnoses than the high-burden boroughs (p = 0.0074). Since the p-value is well below 0.05, we reject the null hypothesis and conclude that Queens experiences significantly more late-stage diagnoses. This suggests that, despite having fewer overall HIV cases, Queens faces greater challenges with early detection and timely testing compared to the high-burden boroughs.



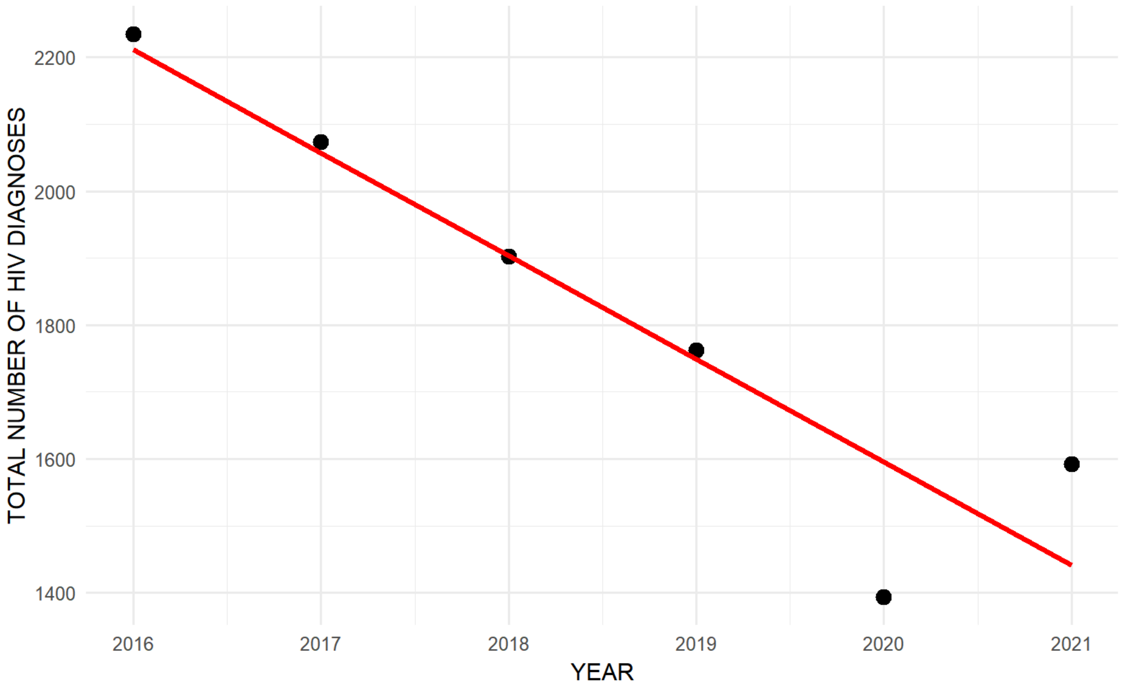
Modeling & Cross-validation

Yearly Trends: We fit a linear model to examine how total HIV diagnoses changed from 2016 to 2021 and to quantify whether the downward trend seen in the data is statistically meaningful.

```
## Call:
## lm(formula = `TOTAL NUMBER OF HIV DIAGNOSES` ~ YEAR, data = df_years)
##
## Residuals:
##        1        2        3        4        5        6
##  150.810 -201.248   12.695   -1.362   15.581   23.524
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 312790.68   61130.58   5.117  0.00690 **
## YEAR          -154.06      30.29  -5.087  0.00705 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 126.7 on 4 degrees of freedom
## Multiple R-squared:  0.8661, Adjusted R-squared:  0.8326
## F-statistic: 25.88 on 1 and 4 DF,  p-value: 0.007047
```

The linear model shows a clear and statistically significant downward trend in total HIV diagnoses from 2016 to 2021. The estimated slope is −154.06, indicating that diagnoses decreased by roughly

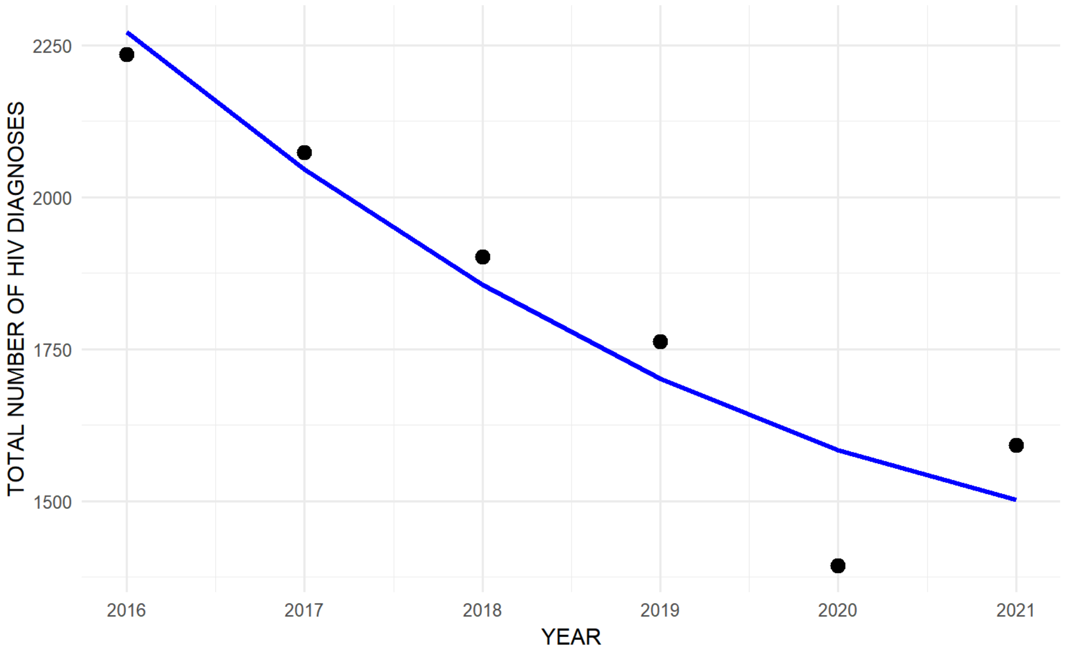
154 cases per year on average. This decline is statistically significant (p ≈ 0.007), meaning it is unlikely to be due to chance. The model explains a large portion of the year-to-year variation (R² = 0.866), suggesting that the overall decline is strong and consistent across the observed period.



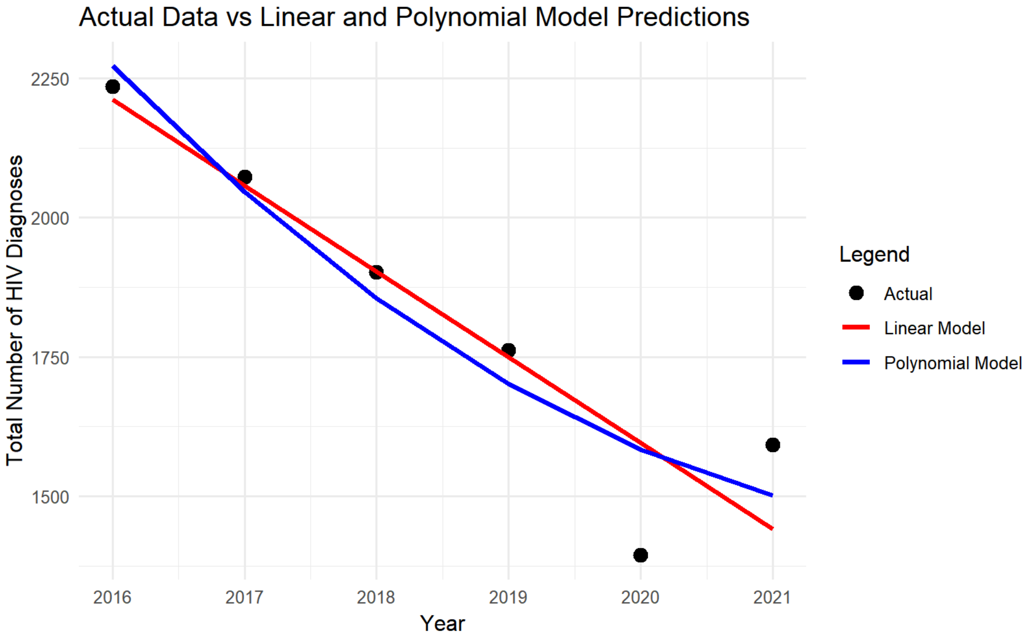
To check whether the trend in HIV diagnoses followed a curved pattern rather than a straight-line decline, we fit a quadratic model that includes a YEAR² term. This helps us test whether a nonlinear pattern provides a better explanation of the data.

```
## Call:
## lm(formula = `TOTAL NUMBER OF HIV DIAGNOSES` ~ YEAR + I(YEAR^2),
##     data = df_years)
##
## Residuals:
##       1       2       3       4       5       6
##   90.57 -189.20   60.89   46.83   27.63  -36.71
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.394e+07  8.780e+07   0.842    0.462
## YEAR        -7.311e+04  8.700e+04  -0.840    0.462
## I(YEAR^2)    1.807e+01  2.155e+01   0.839    0.463
##
## Residual standard error: 131.7 on 3 degrees of freedom
## Multiple R-squared:  0.8915, Adjusted R-squared:  0.8192
## F-statistic: 12.33 on 2 and 3 DF,  p-value: 0.03572
```

The quadratic model does not provide evidence of a curved trend. Neither the YEAR term nor the YEAR² term is significant (p ≈ 0.46), meaning the data do not support any meaningful nonlinear pattern. Even though the R² is high, the added quadratic term does not improve the model. Overall, the trend remains essentially linear.



The fitted curve from the quadratic model looks nearly identical to a straight line, reflecting the fact that the quadratic term is not significant. The predicted values follow the same steady downward pattern seen in the data, with no noticeable curvature. This visual pattern reinforces the model result: the trend in HIV diagnoses is effectively linear over this period.



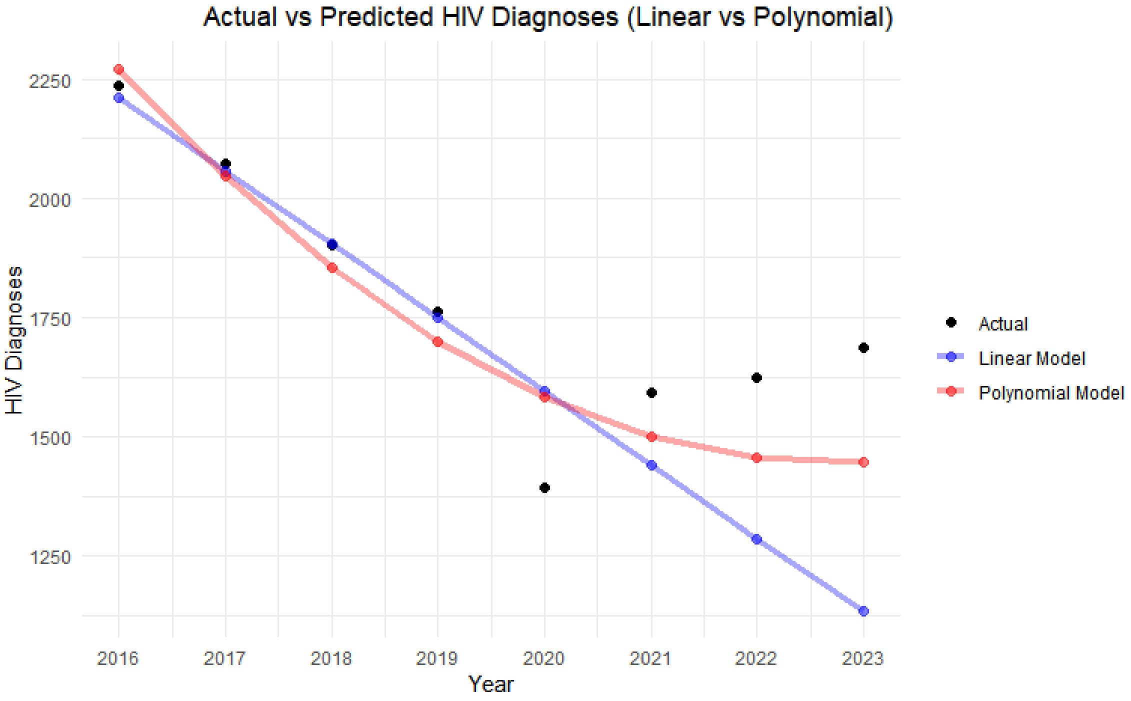
The plot compares the actual HIV diagnoses with predictions from the linear and quadratic models. When viewed side by side, the polynomial curve shows a slight curvature, but it does not meaningfully change the overall trend. Both fitted lines closely follow the steady downward pattern in the data, reinforcing that the decline is essentially linear.

Testing Extrapolation for 2022-2024

To explore how well our model extends beyond the observed data, we checked the most recent NYC HIV Surveillance Annual Reports for 2022 and 2023 and compared their published totals with our model's extrapolated predictions. The 2024 report is not yet available.

It is important to note that the annual reports do not perfectly match the values in our dataset. For example, the reports list 1,594 cases for 2021 (our dataset: 1,592), 1,396 cases for 2020 (dataset: 1,394), and 1,772 cases for 2019 (dataset: 1,762). These small discrepancies likely reflect data revisions or reporting differences between sources.

Nonetheless, we were curious to see whether the overall trend observed from 2016 to 2021 continued in recent years, and whether our fitted model could reasonably approximate the direction of that trend.



This plot compares actual HIV diagnoses with the linear and polynomial model predictions extended through 2023. While both models stay close to the observed data from 2016 to 2021, their extrapolated predictions diverge afterward. The linear model continues a sharp downward decline, whereas the polynomial model flattens out and predicts a slower decrease. The actual values for 2022 and 2023 fall above both prediction lines, which is somewhat expected given the small sample size and the disruptions caused by COVID-19 during the recent years.

Geographic Exploration

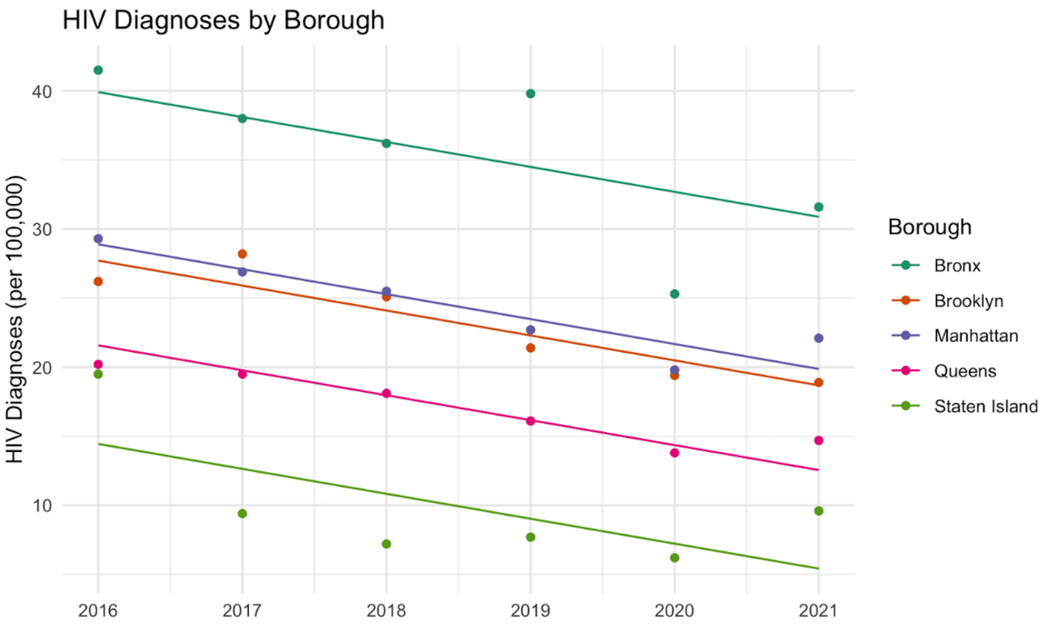
We also explored the trends geographically, examining how HIV diagnoses varied across New York City's boroughs to understand whether the overall pattern differed by location.

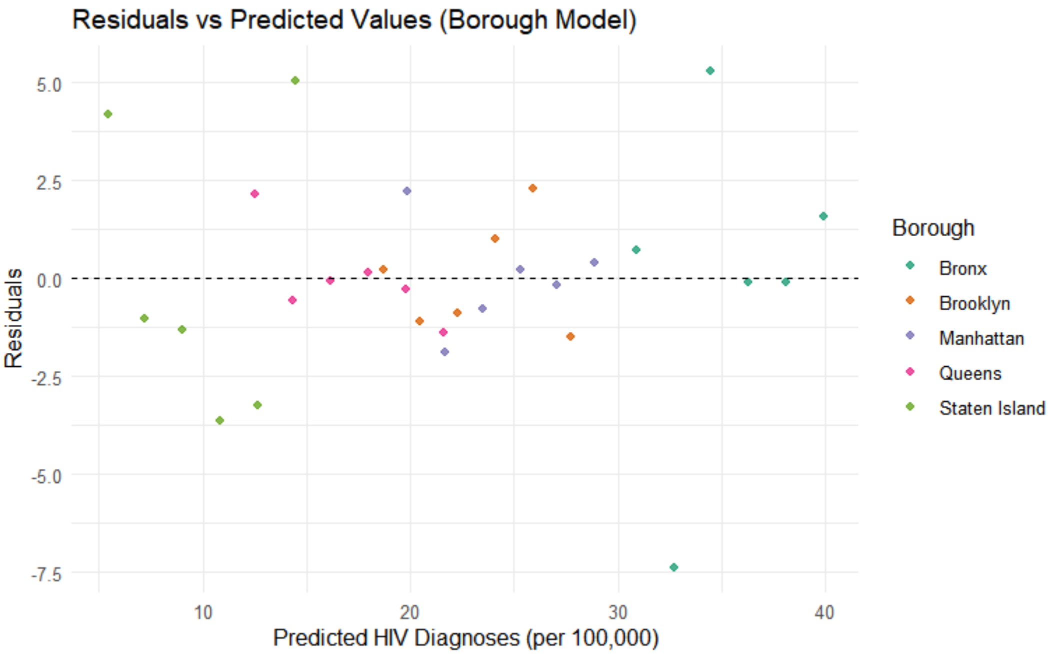
Borough-level Model

We also fit a regression model using borough and year to examine how HIV diagnosis rates differ across locations and how those borough-level trends have changed over time.

```
## Call:
## lm(formula = `HIV DIAGNOSES PER 100,000 POPULATION` ~ YEAR +
##     Borough, data = borough_df_agg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3923 -1.0756 -0.1051  0.9263  5.3026
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3679.0809   594.5785   6.188 2.15e-06 ***
## YEAR                    -1.8051     0.2946  -6.128 2.49e-06 ***
## BoroughBrooklyn        -12.2000     1.5908  -7.669 6.61e-08 ***
## BoroughManhattan       -11.0167     1.5908  -6.925 3.67e-07 ***
## BoroughQueens          -18.3333     1.5908 -11.524 2.87e-11 ***
## BoroughStaten Island  -25.4667     1.5908 -16.008 2.61e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.755 on 24 degrees of freedom
## Multiple R-squared:  0.9301, Adjusted R-squared:  0.9155
## F-statistic: 63.88 on 5 and 24 DF,  p-value: 4.458e-13
```

The model fits the borough-level data very well. The low residual standard error (2.755) indicates strong accuracy, and the high R-squared (0.9301) and adjusted R-squared confirm that year and borough together explain most of the variation in HIV diagnosis rates. The small p-value for the overall F-statistic shows that these predictors jointly contribute significantly to the model. Each borough also has a very low p-value, meaning their average rates differ significantly from one another. Combined with the negative year coefficient, the results show two clear patterns: HIV diagnoses are declining over time, and each borough follows its own distinct rate level.





The residual plot shows no strong pattern, indicating that the borough model generally meets linearity assumptions. Most residuals are close to zero, though Staten Island and the Bronx show larger positive or negative values, indicating the model tends to underpredict or overpredict in those boroughs. This suggests that while the model captures the overall trend, it struggles more with boroughs that have unusually low or high HIV rates.

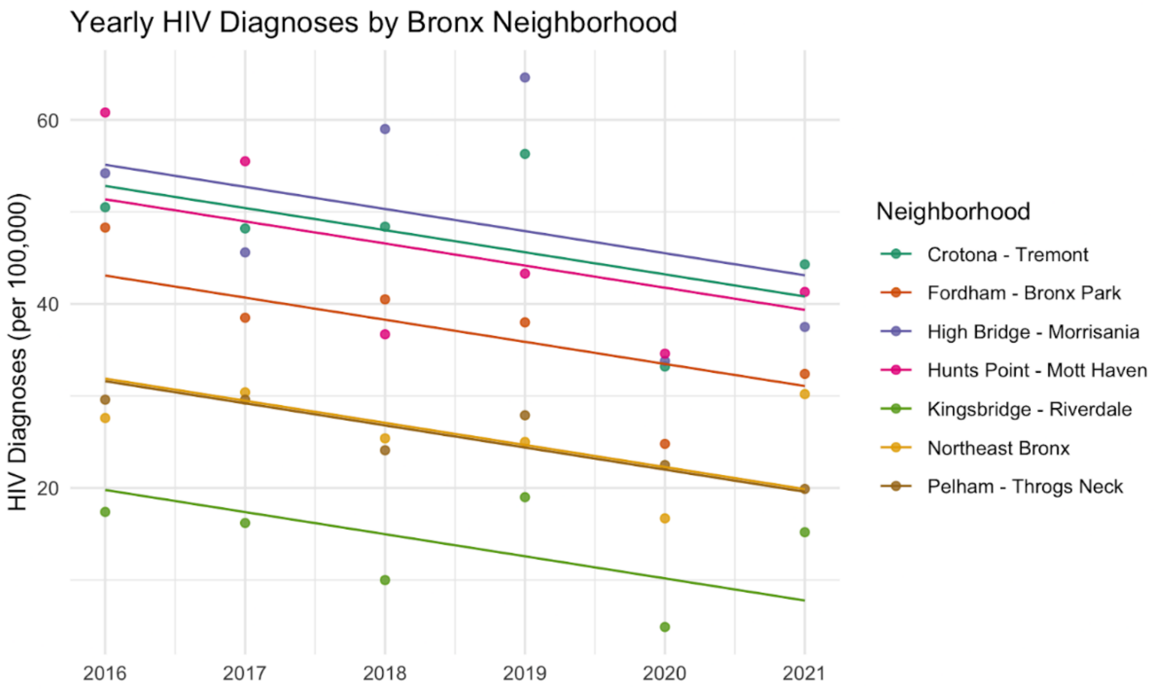
Neighborhood-level Model

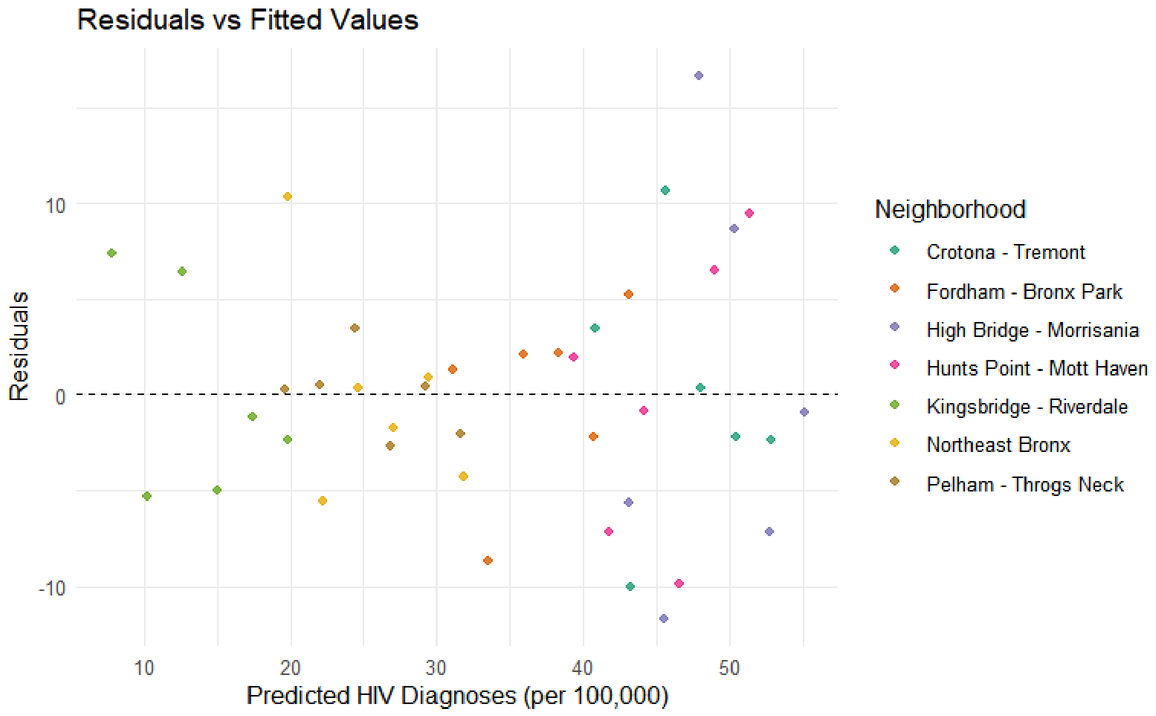
We also fit a neighborhood-level model within the Bronx to examine how HIV diagnosis rates vary across different areas of the borough and how those rates have changed over time.

```
## Call:
## lm(formula = `HIV DIAGNOSES PER 100,000 POPULATION` ~ YEAR +
##     Neighborhood, data = bronx_neighborhood_simple)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7136  -3.8916  -0.2803   3.1702  16.6844
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          4895.3361  1233.0498   3.970 0.000353 ***
## YEAR                                   -2.4020     0.6109  -3.932 0.000393 ***
## NeighborhoodFordham - Bronx Park       -9.7333     3.9035  -2.493 0.017678 *
## NeighborhoodHigh Bridge - Morrisania    2.3000     3.9035   0.589 0.559616
## NeighborhoodHunts Point - Mott Haven   -1.4500     3.9035  -0.371 0.712601
## NeighborhoodKingsbridge - Riverdale   -33.0333     3.9035  -8.462 6.98e-10 ***
## NeighborhoodNortheast Bronx           -20.9333     3.9035  -5.363 5.81e-06 ***
## NeighborhoodPelham - Throgs Neck      -21.2167     3.9035  -5.435 4.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.761 on 34 degrees of freedom
## Multiple R-squared:  0.8212, Adjusted R-squared:  0.7844
## F-statistic: 22.31 on 7 and 34 DF,  p-value: 5.611e-11
```

The model indicates a clear downward trend in HIV diagnosis rates within the Bronx, with the year coefficient showing a significant decline over time ($p < 0.001$). The overall model also fits well, reflected by a high R-squared value (0.8212), meaning year and neighborhood together explain a substantial share of the variation. The small p-value for the F-statistic confirms that the predictors collectively contribute significantly to the model. Several neighborhoods have very low p-values, indicating their average rates are statistically different from the reference neighborhood.

Overall, Bronx neighborhoods do not follow a single uniform pattern. Some areas have consistently lower rates than others, even after accounting for the overall citywide decline.





The neighborhood-level residual plot shows a wider spread than the borough model, which means the linear model has more difficulty capturing variation across individual neighborhoods. While most points hover around zero, several neighborhoods show large positive or negative residuals, indicating noticeable underprediction or overprediction. This suggests that the relationship between predictors and HIV rates is more complex at the neighborhood level than a simple linear model can capture.

Why We Moved Beyond Linear Models

After running these models, we realized that this dataset is not a great fit for standard linear modeling. The dataset contains too few time points for a linear trend to be dependable. Borough and neighborhood data also follow patterns driven more by population differences and reporting rules than by smooth numerical trends. Because of this, the linear models did not capture the data's structure very well. So, we decided to treat the data differently and try a different type of regression that better fits how this dataset behaves.

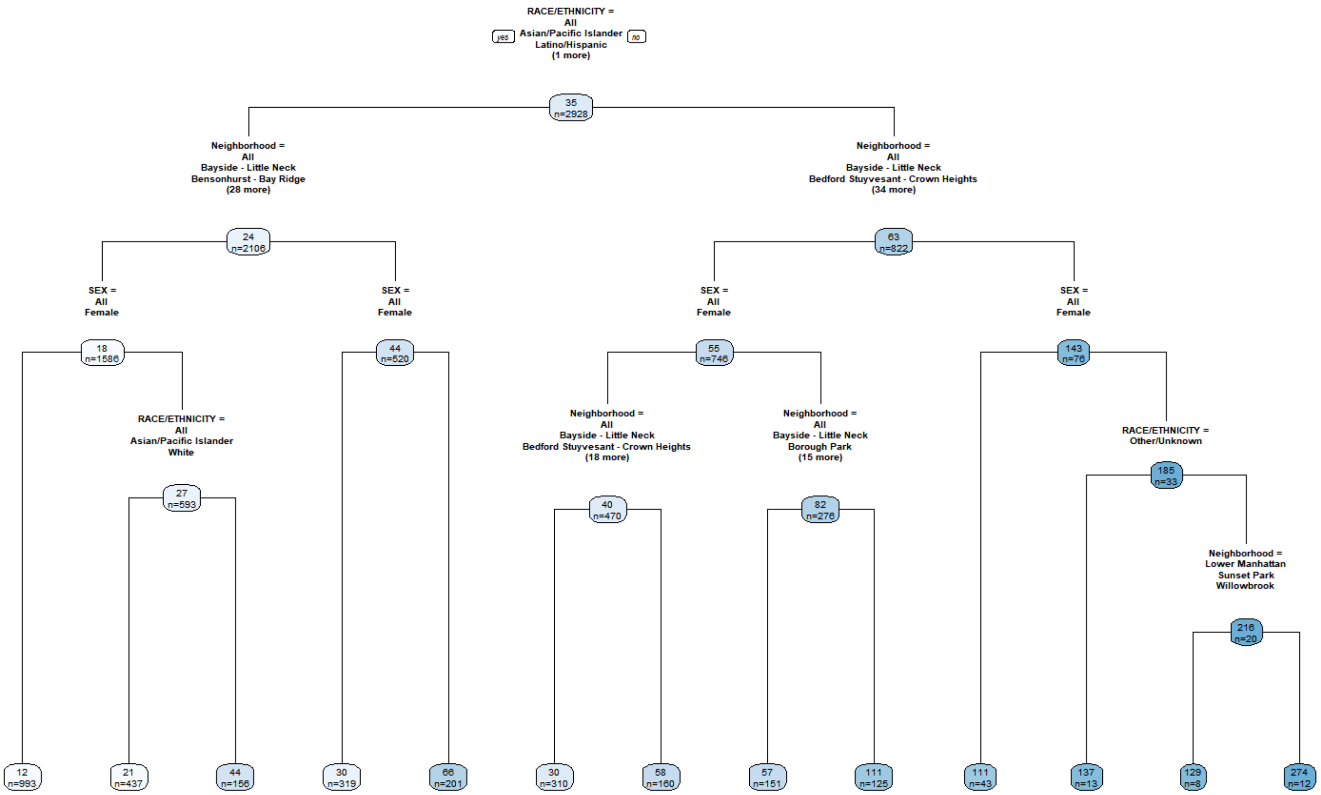
A Different Approach - Tree-based Regression

Because linear models were not giving us meaningful results, we moved to tree-based methods, starting with decision tree regression and then random forest regression. These models work differently: instead of trying to fit one line to all the data, they split the data into smaller, more similar groups and make predictions within those groups. This makes them more flexible and better suited for patterns that are not linear. Random forests in particular average many trees together, which usually gives more stable and accurate predictions. These methods allowed us to explore borough, race, and demographic factors in a way that better matched the structure of the dataset.

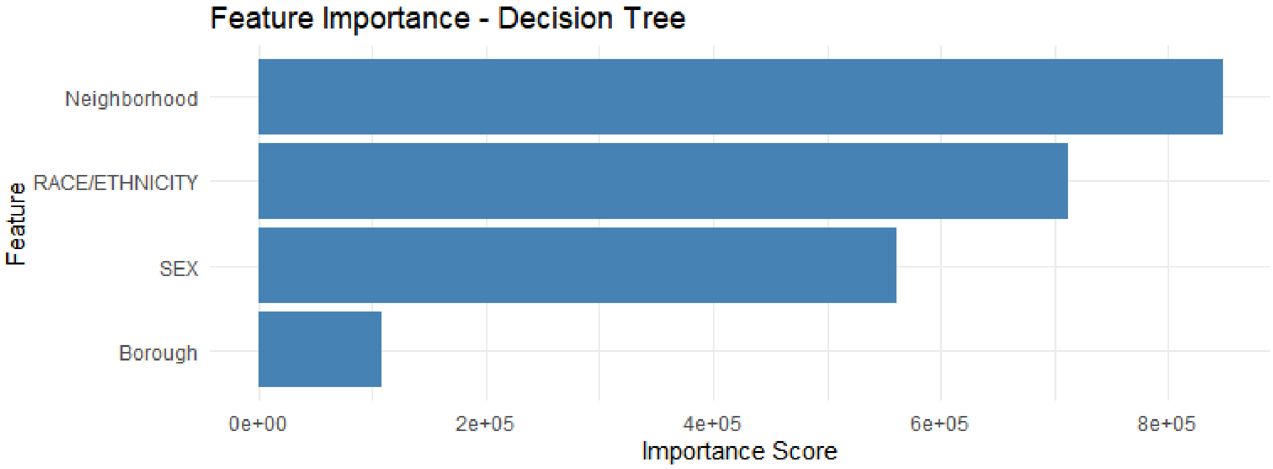
To prepare the data for tree-based regression, we first created a modeling dataset containing the HIV diagnosis rate along with key predictors such as year, borough, neighborhood, sex, race/ethnicity, and the total number of diagnoses. All categorical variables were converted into factors so that the tree model could correctly treat them as groups rather than numeric values. We then split the data into training and testing sets using an 80-20 split to allow for model evaluation on unseen data. After setting a random seed for reproducibility, we fit a decision tree regression model using the rpart package.

Decision Tree Regressor

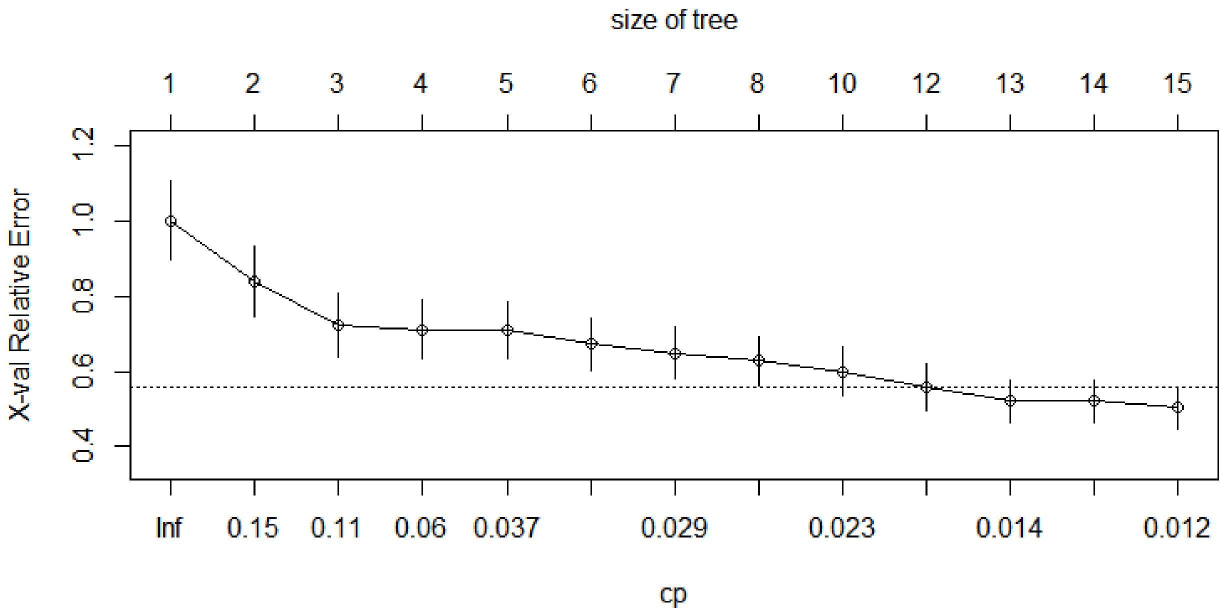
We fit a decision tree regression model using the *rpart* function, with the HIV diagnosis rate as the outcome and year, borough, neighborhood, sex, and race/ethnicity as predictors. The model uses the "anova" method, which is designed for continuous outcomes. This approach lets the tree automatically find splits in the data that best explain differences in HIV rates across groups and locations. By doing so, it identifies which factors and combinations of factors contribute most to variations in diagnosis rates.



The decision tree begins with race and ethnicity because this variable creates the strongest separation in the data. As the tree moves downward, neighborhood and sex appear often because they help refine the groups even further. Each final leaf represents a subgroup with its own average predicted HIV rate, showing how combinations of demographic and geographic factors shape the overall pattern. The structure of the tree makes it clear that the differences in HIV rates are not driven by any single factor but by how these characteristics interact with one another.



The feature importance results show that Neighborhood contributes the most to predicting HIV diagnosis rates in the decision tree. Race/ethnicity is the next most important factor, followed by sex. Borough has the smallest impact, likely because neighborhood information already captures most of the geographic variation. Overall, this tells us that local neighborhood differences explain more of the variation in HIV rates than broader borough-level or demographic factors alone.



This plot shows how the tree's cross-validated error changes as the model becomes more complex. When the tree is very small, the error is high because the model cannot capture much of the variation. As the tree grows, the error drops steadily, meaning additional splits help improve prediction accuracy. After a certain point, the curve levels off, which tells us that making the tree larger provides only small improvements. This pattern helps identify a reasonable tree size that improves accuracy without becoming unnecessarily complex.

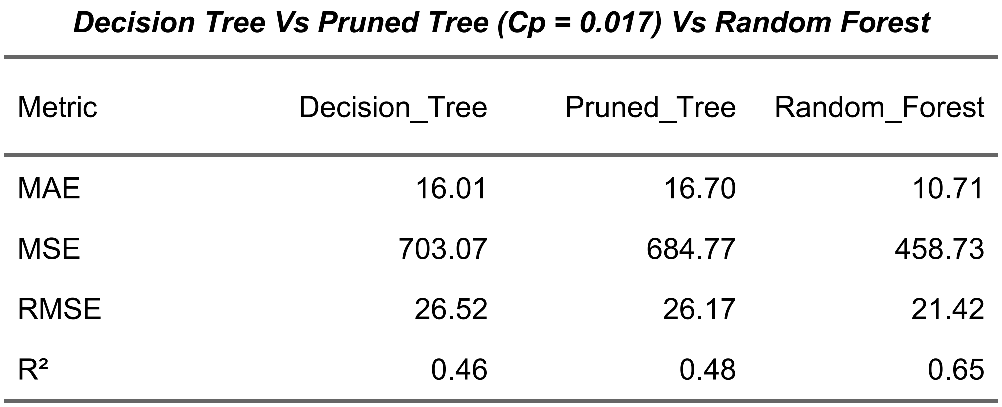
Decision Tree Vs Pruned Tree (Cp = 0.017)

Metric	Decision_Tree	Pruned_Tree
MAE	16.01	16.70
MSE	703.07	684.77
RMSE	26.52	26.17
R²	0.46	0.48

The full decision tree and the pruned tree perform very similarly. The pruned tree shows a small increase in MAE but achieves slightly lower MSE and RMSE, indicating marginally better average error overall. Its R² increases from 0.46 to 0.48, meaning the pruned model explains a little more variance in HIV diagnosis rates. In this case, pruning produces a simpler model without hurting performance and may even offer a modest improvement.

Random Forest Regressor

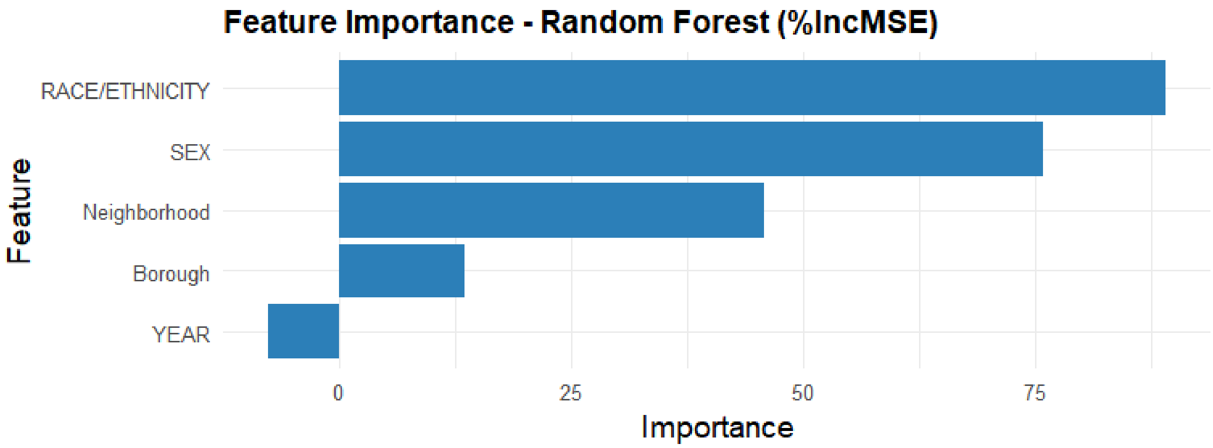
We trained a random forest model to improve prediction accuracy. Unlike a single decision tree, a random forest builds many trees using random subsets of the data and then averages their predictions, which helps reduce overfitting and produce more stable results. We used 500 trees and set *mtry* to 3, allowing the model to consider three predictors at each split. This approach lets us capture more complex patterns in HIV rates across neighborhoods and demographic groups.



Decision Tree Vs Pruned Tree (Cp = 0.017) Vs Random Forest

Metric	Decision_Tree	Pruned_Tree	Random_Forest
MAE	16.01	16.70	10.71
MSE	703.07	684.77	458.73
RMSE	26.52	26.17	21.42
R²	0.46	0.48	0.65

The random forest model performs noticeably better than both the full and pruned decision trees. It has the lowest MAE, MSE, and RMSE values, meaning its predictions are closer to the actual HIV rates. Its R² value is also much higher at 0.65, indicating that it explains a larger share of the variation in the data. These results show that combining many trees produces a more accurate and stable model than relying on a single tree, which is expected given the complexity and variability of HIV rates across neighborhoods and demographic groups.



The random forest feature importance results show that race and ethnicity is the strongest predictor of HIV diagnosis rates, followed closely by sex. Neighborhood still plays an important role but contributes less than the first two variables. Borough and year have much smaller impacts once the other predictors are included. This pattern highlights demographic characteristics as the main drivers of HIV rate variation in the random forest model. In the decision tree, neighborhood appeared as the most important factor, which reflects how a single tree often relies on large geographic splits. The random forest spreads its importance across many trees, which brings demographic variables like race and sex to the forefront.

Random Forest Cross-Validation

To tune the random forest model, we used 10-fold cross-validation and tested different combinations of the number of trees (ntree) and the number of predictors considered at each split (mtry). For each ntree value, the model evaluated five possible mtry settings and selected the one that produced the lowest RMSE. This process allows the model to choose the tuning parameters that give the best predictive performance rather than relying on defaults. After collecting the results across all ntree values, we compared them and identified the combinations with the highest R² and lowest errors. This cross-validation step helps ensure that the final random forest model is well-optimized and generalizes better to unseen data.

Top 5 Cross-Validated Scores - Random Forest

ntree	mtry	RMSE	Rsquared	MAE
1,500	4	28.93	0.61	16.74
200	4	29.14	0.60	16.74
800	4	29.19	0.60	16.78
2,000	4	29.05	0.60	16.76
500	4	29.32	0.59	16.82

The cross-validation results show that the random forest performs consistently well across different numbers of trees, with RMSE values around 29 and R² values in the range of 0.59 to 0.61. This suggests that, on average, the model can explain a little more than half of the variation in HIV diagnosis rates when evaluated on different subsets of the training data.

The initial random forest model trained with 500 trees and mtry = 3 achieved a higher R² of about 0.65 on the test set, which reflects the performance on that specific data split. This difference can occur when the test data happen to follow patterns that the model learned especially well, making that particular split slightly easier to predict than the folds used during cross-validation.

Can a Random Forest Trained on Past Data Predict the Future?

To explore this question, we trained the model using data from 2016-2020 and evaluated its performance on the unseen year 2021. Although Random Forests are not traditional forecasting models and YEAR itself was not a strong predictor, this temporal split allows us to assess whether the demographic and geographic patterns learned in earlier years remain stable when applied to a new period.

The goal is not to forecast future values in the strict time-series sense, but to test the model’s ability to generalize across years.

RF Model Evaluation on Unseen Year (2021)

mtry	ntree	MAE	MSE	RMSE	R2
2	500	11.383	602.318	24.542	0.612

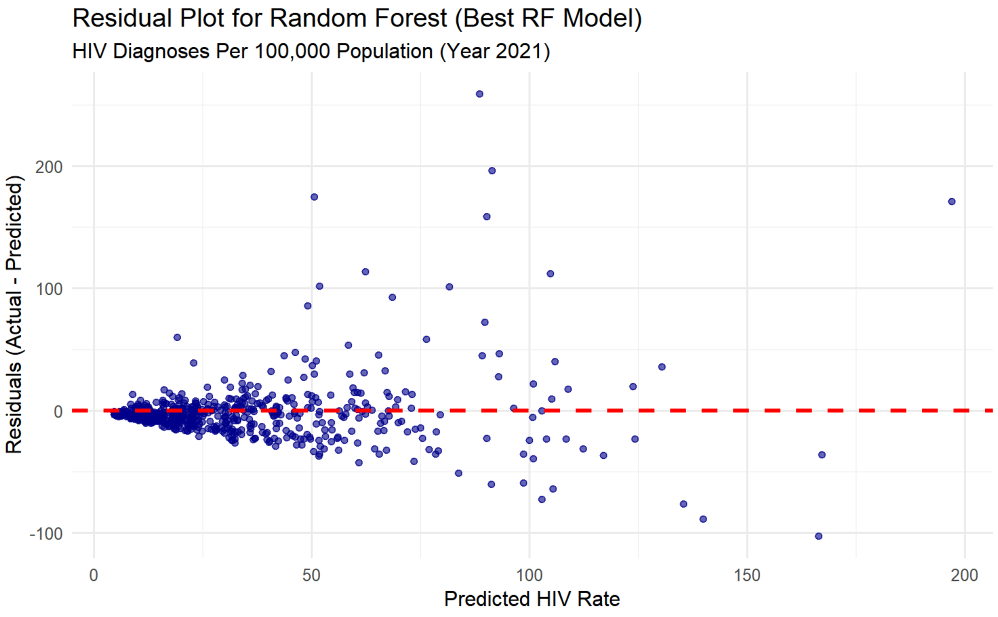
The model performs well on the held-out data. With its chosen parameters, it reaches an MAE of about 11 and an RMSE of about 24, meaning the forecasts for 2021 are generally close to the true HIV rates. The R² value of 0.612 shows that the model captures over 60 percent of the variation, which indicates strong year-to-year generalization.

Random Forest Hyperparameter Search

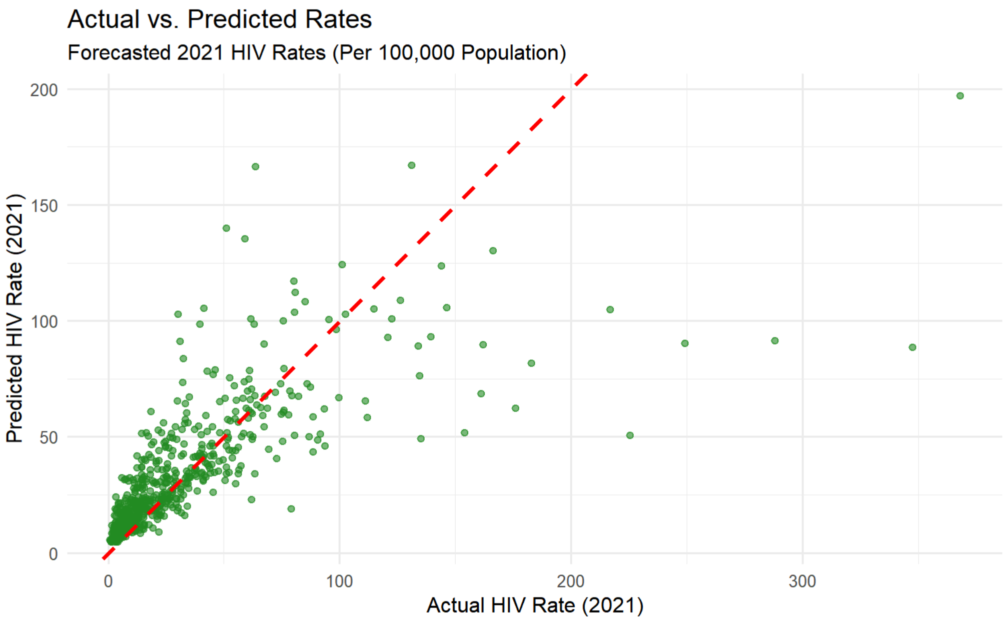
We ran a grid-search experiment by tuning four random forest parameters: mtry (1 to 4), ntree (200 to 1000), nodesize (3 to 15), and maxnodes (15 to 50). For each combination we trained a model and recorded the out-of-bag RMSE and R² to assess performance. Sorting the results by R² helped identify the strongest parameter settings.

mtry	ntree	nodesize	maxnodes	RMSE	R2
3	800	10	50	25.09064	0.6126113
3	1,000	3	50	25.09942	0.6123402
3	800	5	50	25.11305	0.6119190

We then trained a final random forest model using the best parameter combination identified in the grid search. The selected settings were mtry = 3, ntree = 800, nodesize = 10, and maxnodes = 50. After fitting this model on the training years, we used it to generate predictions for the 2021 test data to assess how well the tuned model performs on unseen observations.



The residual plot shows that the model predicts lower HIV diagnosis rates fairly accurately, with most points near zero for predicted values under 50. This happens because most observations fall in this range, giving the model more training examples. At higher predicted rates, the residuals spread out and become more extreme, meaning the model is less reliable for neighborhoods or groups with very high HIV diagnosis rates, which are much less common in the data.



The plot shows that as the actual HIV rate increases, the model's predicted rate also increases, so it is capturing the overall relationship in the data. For most observations under about 50 cases per 100,000, the predictions stay close to the diagonal reference line. At higher actual rates, many points fall below the line, indicating that the model underpredicts these extreme cases. This aligns with what we observed in the residual plot, where larger positive residuals appeared mostly at higher predicted values, confirming that the model struggles with the rare high-rate observations and tends to shrink them toward the center of the distribution.

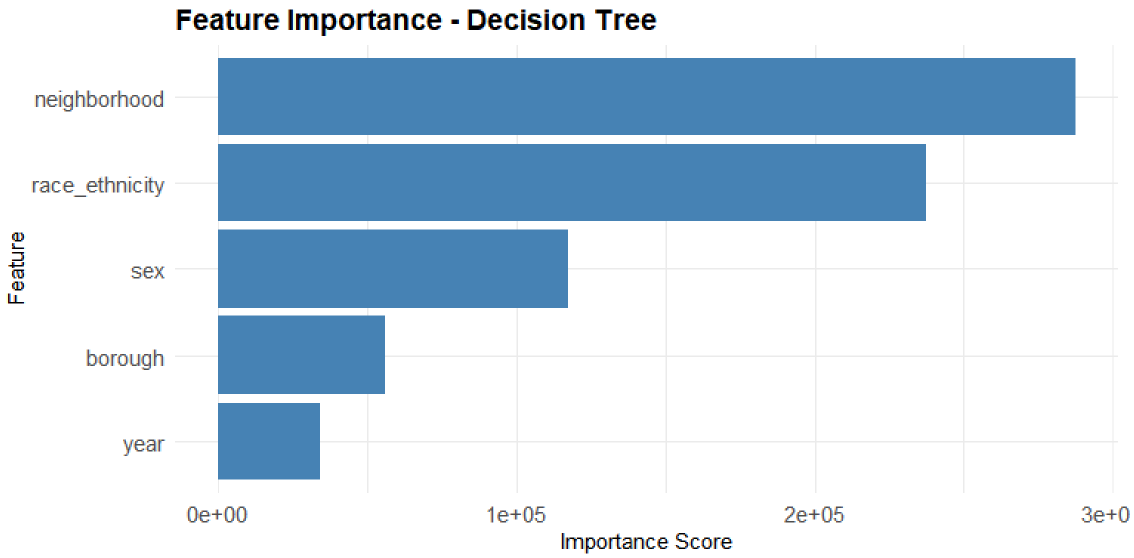
Cross-Validation - RF Model with Best Parameters

mtry	ntree	nodesize	maxnodes	MAE	MSE	RMSE	R2
3	800	10	50	13.908	640.728	24.611	0.615

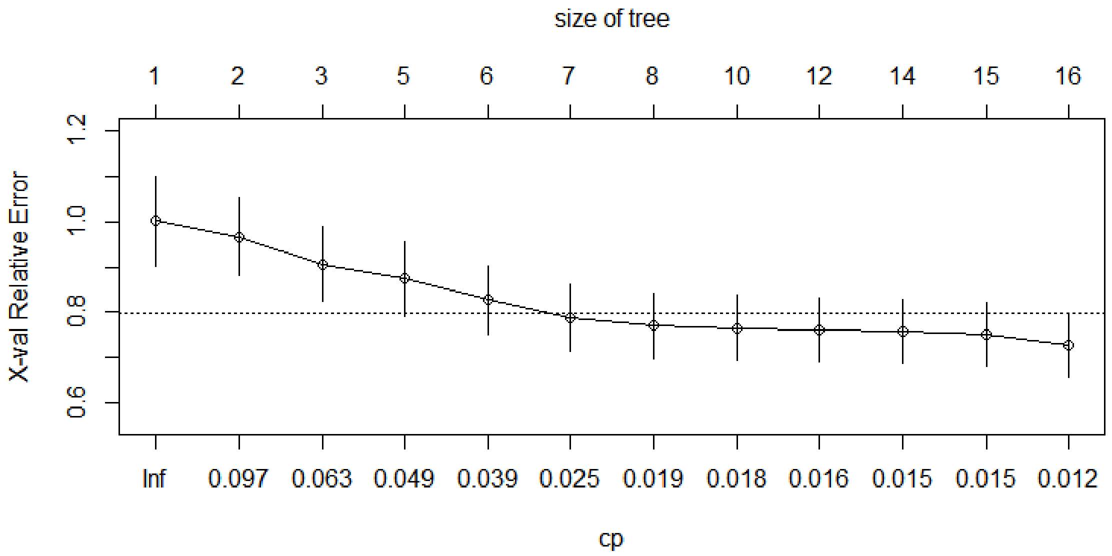
The cross-validation result for the model with mtry = 3, ntree = 800, nodesize = 10, and maxnodes = 50 gives an R² of about 0.615. This is very close to the training R² of about 0.612 from the initial model with mtry = 2 and ntree = 500, which we used to predict on unseen 2021 data. The similarity between these values suggests that the model stays stable during tuning and does not show clear signs of overfitting.

AIDS Rate Modeling

We also extended our analysis by trying to predict AIDS diagnoses per 100,000 population using both decision trees and random forests. The models followed the same setup as our HIV rate experiments.



The decision tree importance plot shows that neighborhood is the strongest predictor of AIDS diagnoses, followed by race/ethnicity and sex. Borough contributes less, and year has very little influence. This means the model relies most on local neighborhood-level differences when making predictions.

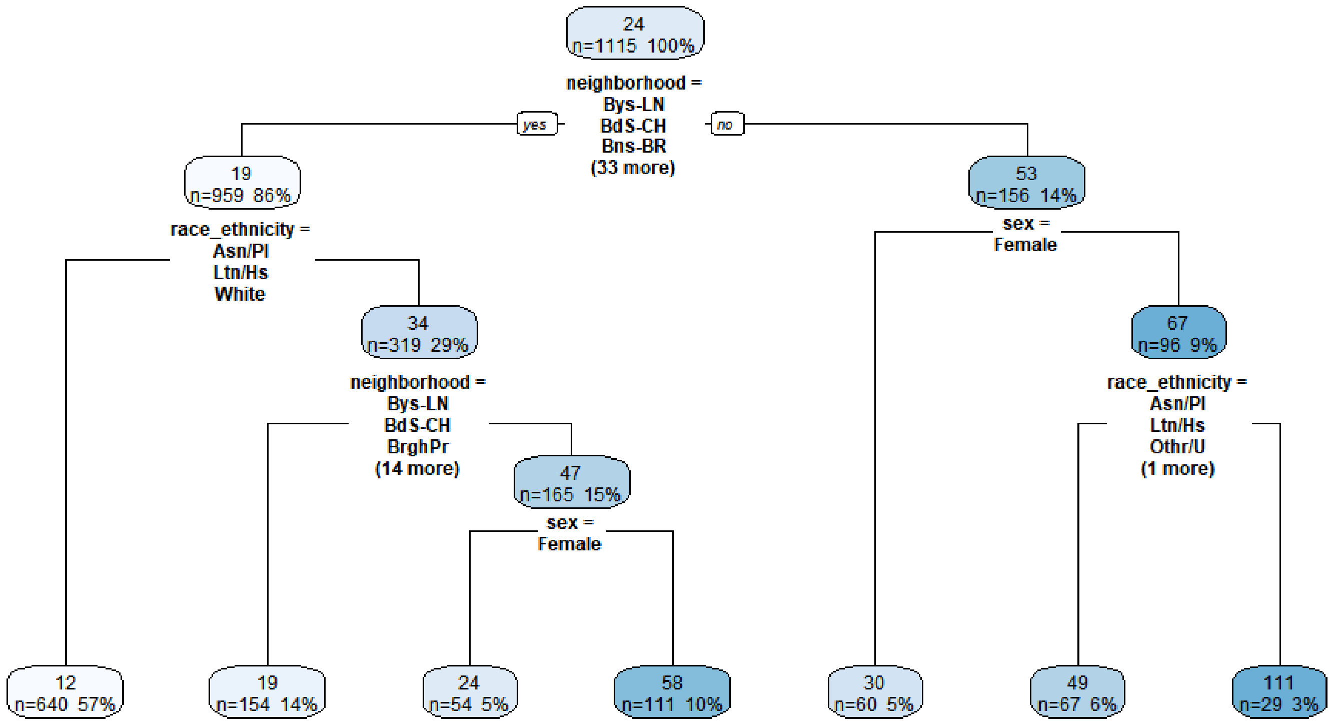


The cross-validation curve shows a steady reduction in relative error as cp decreases, and the minimum point occurs around cp ≈ 0.012. This is where the tree achieves its lowest cross-validated error, indicating the best balance between model complexity and predictive performance for this dataset.

Decision Tree Vs Pruned Tree (Cp = 0.025)

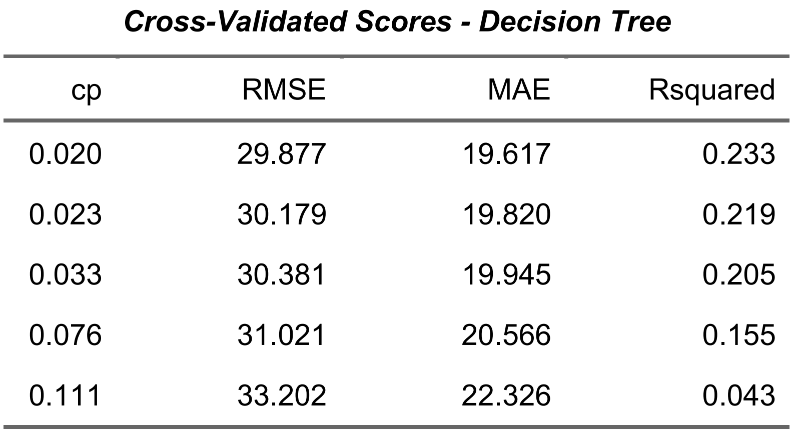
Metric	Decision_Tree	Pruned_Tree
MAE	17.88	19.16
MSE	859.98	842.49
RMSE	29.33	29.03
R²	0.24	0.25

Even though the cross-validation curve identifies cp ≈ 0.012 as the best value, the tree pruned at cp = 0.025 achieved slightly better RMSE, and R² on our evaluation set. This happens because rpart optimizes cross-validated relative error, which does not always align perfectly with external performance metrics, especially in smaller or noisier datasets. A slightly simpler tree can generalize better in practice.



Pruned Decision Tree

The pruned tree shows that neighborhood is the dominant factor separating HIV diagnosis rates, with race and ethnicity and sex providing additional refinements within each group. The simplified structure still captures the same hierarchy seen earlier, and it further reinforces the feature importance results, where neighborhood and demographic variables were the strongest predictors.



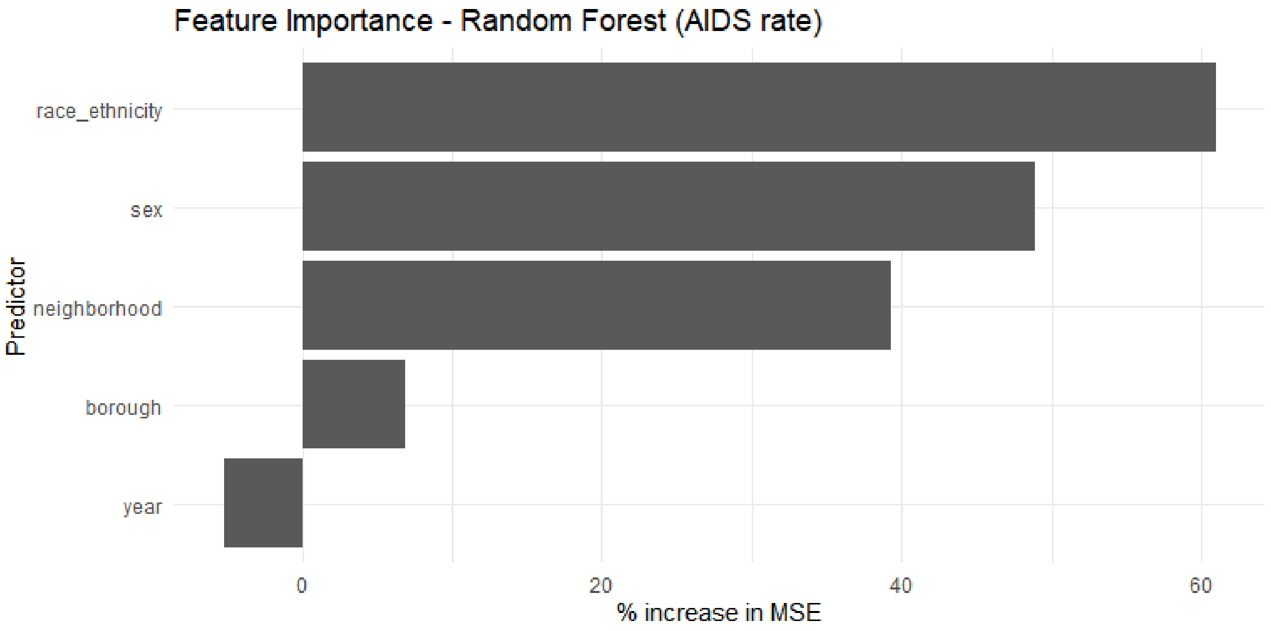
Cross-Validated Scores - Decision Tree

cp	RMSE	MAE	Rsquared
0.020	29.877	19.617	0.233
0.023	30.179	19.820	0.219
0.033	30.381	19.945	0.205
0.076	31.021	20.566	0.155
0.111	33.202	22.326	0.043

The cross-validation results show that the decision tree performs best at cp = 0.020, where RMSE and MAE are lowest and R² is highest at about 0.223. As cp increases, the model becomes more aggressively pruned and performance worsens, with R² dropping and errors rising. This indicates that lightly pruned trees capture the AIDS-rate patterns better than more simplified versions, though overall predictive power remains modest.

Random Forest

We also fit a random forest model to predict AIDS diagnoses per 100,000, using 500 trees and mtry = 3. This model provides more stable predictions than a single tree and again highlights neighborhood and demographic variables as the strongest predictors.



The random forest feature importance results show that race and ethnicity is the strongest predictor of AIDS rates, followed closely by sex. Neighborhood also has a meaningful impact but contributes less than the demographic variables. Borough and year add very little once the model accounts for race, sex, and neighborhood.

Decision Tree Vs Pruned Tree Vs Random Forest

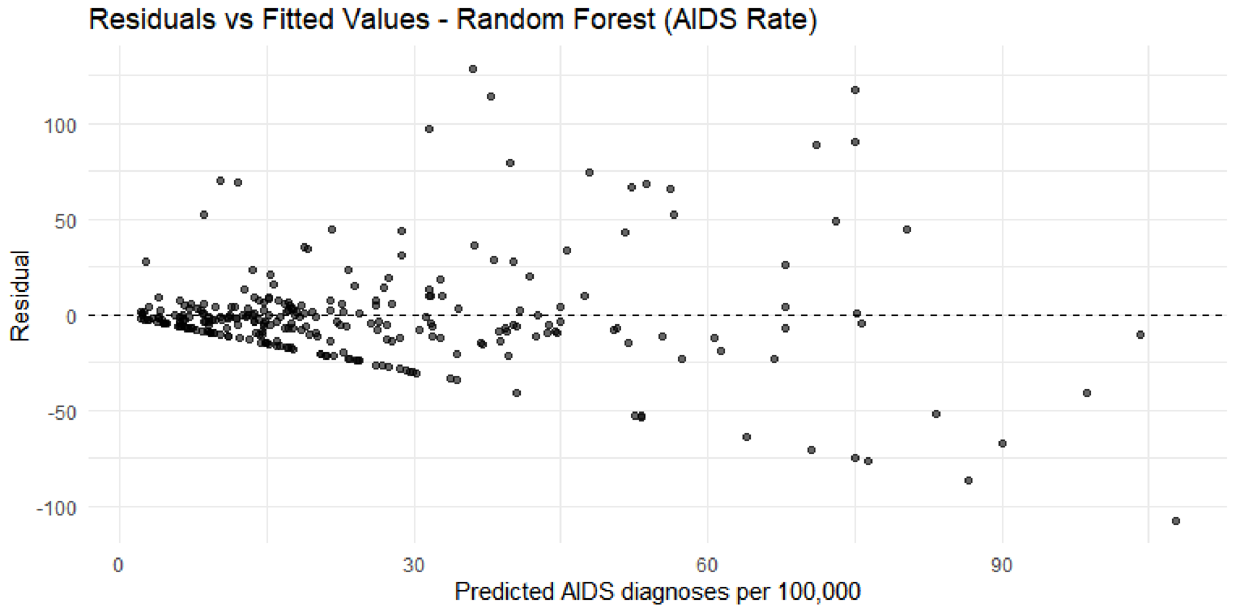
Metric	Decision_Tree	Pruned_Tree	Random_Forest
MAE	17.88	19.16	16.73
MSE	859.98	842.49	788.04
RMSE	29.33	29.03	28.07
R^2	0.24	0.25	0.30

The results show that predicting AIDS diagnoses is more challenging than predicting HIV rates. All three models have relatively low R^2 values, meaning they explain only a small portion of the variation in AIDS rates. The random forest performs the best, with the lowest MAE and RMSE and the highest R^2 at 0.30, but the improvement over the decision tree models is modest. Overall, the models capture some structure in the data, but AIDS rates appear more difficult to model accurately.

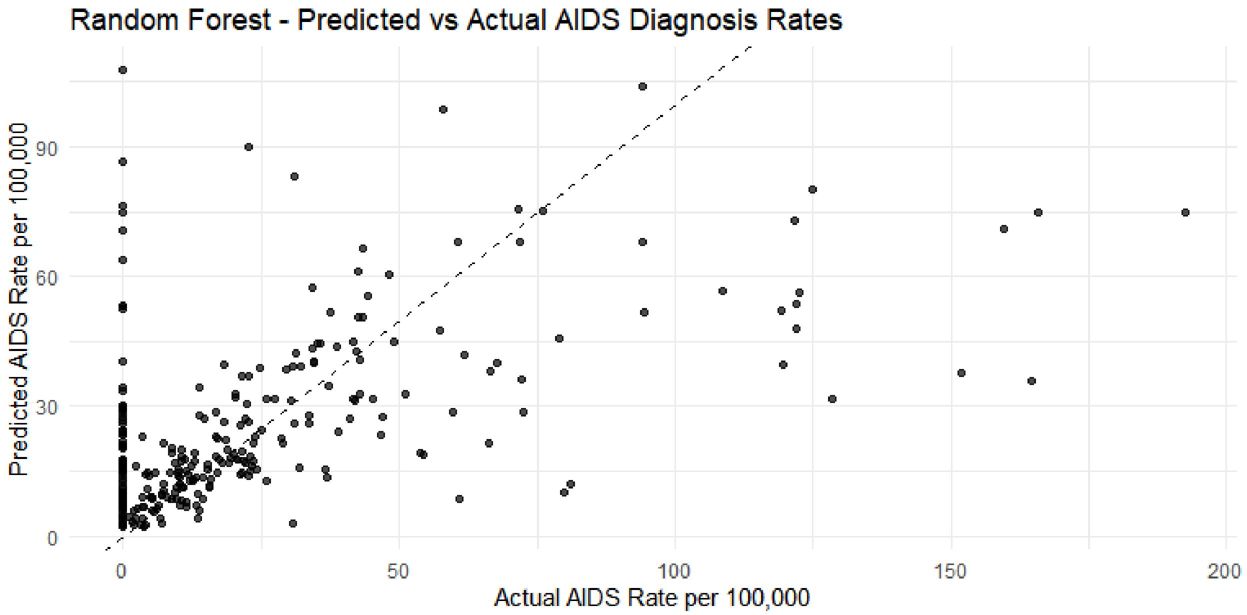
Cross-Validated Scores - Random Forest

mtry	RMSE	MAE	Rsquared
15	26.049	15.318	0.403
28	26.337	15.260	0.399
55	26.590	15.307	0.394
41	26.566	15.281	0.392
2	29.858	19.622	0.321

The cross-validation results for the random forest model show that performance improves noticeably as mtry increases. The best scores appear around mtry = 15-28, where RMSE is roughly 26 and R^2 reaches about 0.40. This means the model explains around 40 percent of the variation during cross-validation, which is stronger than the decision tree results.



The residual plot shows that the random forest predicts lower AIDS rates fairly well, but the errors get larger as the predicted values increase. Most points near 0-30 diagnoses per 100,000 cluster close to the horizontal line, meaning good accuracy in the common range. At higher predicted values, the residuals spread out more and often fall below zero, which indicates underprediction for the highest-rate neighborhoods. This matches what we saw earlier with HIV rates: the model performs best where the data is dense and struggles with rare, extreme cases.



The predicted vs actual plot shows that the random forest model captures the general increase in AIDS rates but consistently underestimates the highest values. Most low-rate observations (near zero) are predicted reasonably well, but as the actual AIDS rate rises, the predicted values fall below the diagonal reference line. This underprediction at the upper end matches what we saw in the residual plot, where large positive residuals appeared for high-rate neighborhoods.

Limitations

Our analysis has several limitations that should be considered when interpreting the results. First, the dataset contains only six years of consistently reported values for many of the variables, which limits the reliability of trend modeling and reduces the statistical power of time-based comparisons. Many neighborhood and demographic categories also have sparse data, particularly for smaller populations, which restricts model performance and contributes to the underprediction of extreme values. In addition, the structure of the surveillance dataset reflects reporting practices rather than a randomly sampled process, meaning some patterns arise from data availability rather than underlying epidemiology. Finally, tree-based models capture group differences well but are less effective for forecasting rare, high-burden observations, since the data are highly imbalanced and dominated by low-rate neighborhoods.

Future Works

- Incorporate additional years of data and bring in external variables such as socioeconomic indicators, access-to-care metrics, or viral suppression rates to improve explanatory power.
- Explore more advanced models such as gradient boosting or spatial models that account for geographic relationships between neighborhoods.
- Expand and refine the feature set to reduce missing information and improve model robustness.
- Build a more formal forecasting framework using rolling or expanding training windows to better assess how models generalize to future years.
- Evaluate model stability across different boroughs and demographic groups to understand where errors are highest and why.

Wrap Up

Overall, the project provides a detailed view of how HIV and AIDS diagnoses vary across time, boroughs, neighborhoods, and demographic groups in New York City. The EDA confirms persistent geographic and demographic disparities, with the Bronx, Black and Latino/Hispanic populations, and males showing consistently higher burdens.

Linear models captured broad trends but were not well suited for the structure of the dataset, leading us to adopt tree-based methods. Decision trees offered interpretable groupings, while random forests delivered stronger predictive performance and more stable generalization across years.

Even with their limitations, these models highlight the key factors shaping HIV and AIDS rates and help identify which communities carry the greatest burden. Together, these findings reinforce the need for targeted public health interventions and continued monitoring of neighborhood-level disparities.