# Mapping HIV/AIDs

Fatima Dineh, Jainaba Jawara, Simon Wu

## Research Question Summary

1. How do cases of HIV/AIDS compare across countries, especially comparing continents? We wanted to see how the cases differ in each continent. People stereotype specific continents and say that's where HIV/AIDS cases mainly occur. We wanted to show people data if those stereotypes were true or not.

**Results:** Africa was really high compared to the other continents, and we really can't compare the results due to a big difference. The Seven Seas had the lowest number of people living with HIV_median.

2. How do cases of HIV/AIDS compare across different demographic categories (race, sexuality, gender identity) within the United States? Specifically, what does an intersecting analysis of multiple identity categories reveal about the extent and impact of the HIV/AIDS pandemic? We want to investigate if there is a positive correlation between marginalized identity categories and rates of HIV/AIDS transmission/infection.

**Results:** The percentage representation of different genders living with HIV/AIDS in the US changed very little from 2011 to 2017. In terms of racial categories, the percentage representation of white people notably decreased over time, as opposed to Latin and Asian populations which increased. Male-to-male sexual contact steadily remained by far the most common method of HIV transmission, but notably, high-risk heterosexual contact decreased for both male and female populations over time, whereas injection drug use transmission increased.

3. How does access to ART (HIV/AIDS treatment) compare across different countries? We want to see if there is an ART accessibility imbalance in different continents, which would lead us to see if those areas have trends of risk of higher or lower transmission, morbidity, and death rates.

**Results:** Our results showed us that Africa had the highest ART coverage among people living with HIV_median and Europe came in second. The Seven seas had the lowest ART coverage among people living with HIV_median within the continents.

4. How does the overall deaths, incidence, and prevalence compare across continents? We wanted to see if what was higher in each continent was it deaths, incidence, or prevalence.

**Results:** Africa had the highest deaths, incidence, and prevalence, and Asia came in second. Africa and Asia had a massive difference in deaths, incidence, and prevalence. When you look at the chart, you will be able to see that.

### Motivation

The past few years have raised discussions of global health to the forefront of public consciousness. In particular, there has been greater awareness raised about the intersections between disparities in health/healthcare and the contexts and needs of different communities. Yet, despite the deserved weight and timely significance we have placed on Covid-19, it is important to remember that the unequal impact of epidemics is not a new phenomenon, but rather laced in historical systems of oppression and injustice. HIV/AIDs are an incredible example of this; not only has the HIV/AIDS epidemic historically affected large populations of people in the developing world (particularly on the African continent), but even in countries like the United States with more healthcare infrastructure, HIV/AIDs have disproportionately afflicted and killed queer communities (particularly queer communities of color), and were even politicized as the “gay disease” to further stigmatize research and relief. Thus, by attempting to answer the research questions surrounding HIV/AIDs, we might be able to dismantle racial/sexual stereotypes that persist even today.

### Dataset

We will be using two different datasets to answer our problems. One of them is from the Kaggle website, and the other is from the CHHS website. The Kaggle datasets talk about the different countries and the estimated number of cases of HIV/AIDS that have been reported, the number of people receiving antiretroviral therapy (ART), and the estimated ART coverage among people living with HIV. Also, the min, max, and mean of the estimated number of people living with HIV and estimated ART coverage among people living with HIV. In the second dataset, CHHS talks about the year they got infected with HIV/AIDS. The category column explains the group so that the category will say the age at the end of the year, current gender, race/ethnicity, transmission category: male adult or adolescent, transmission category: female adult or adolescent, and transmission category: child (<12 Years Old at the End of Year); the group will answer the category type. The count column is the number of people living in California when they were diagnosed with HIV during the reported year. The third dataset is from our world in data and it talks about the different countries and the years in which the number of deaths, incidence, and prevalence occurred in all ages and in female and male. For our last dataset we used the world health organization dataset and we used the location and estimated number of people (all ages) living with HIV from 2000 to 2020.

The four datasets are linked below:

https://www.kaggle.com/datasets/imdevskp/hiv-aids-dataset

https://data.chhs.ca.gov/dataset/hiv-aids-cases/resource/ffc84784-957e-486f-ad34-9b14d07583c1

https://ourworldindata.org/hiv-aids

https://www.who.int/data/gho/data/indicators/indicator-details/GHO/estimated-number-of-people--living-with-hiv

### Method

For the first question, we will take all the countries and put them into categories based on continents so that way we have seven continents. Once we have the seven continents, we will take all the estimated number of people living with HIV from our datasets. We will create a couple of charts to see which charts have a better data visualization trend. That will allow us to answer our questions by comparing and seeing if these stereotypes are true or not.

For our second question, we will begin by establishing the different demographics we want to consider across various races, then gender identities, then sexualities. For example, our dataset has categories for cisgender men/women, transgender men/women, and nonbinary individuals, or data for racial categories. Within each of these demographic buckets, we will graph infected counts for the different groups across each year. To visualize this data, we can create 4 separate multivariable line graphs, with each line representing a different group within the demographic category, where the x axis would be year and the y axis would represent HIV/AIDs cases. By analyzing how HIV/AIDS affects different demographics, we can hopefully better understand the extent of the disparate impact of the epidemic, and maybe even draw intersectional conclusions between these demographic bucket categories.

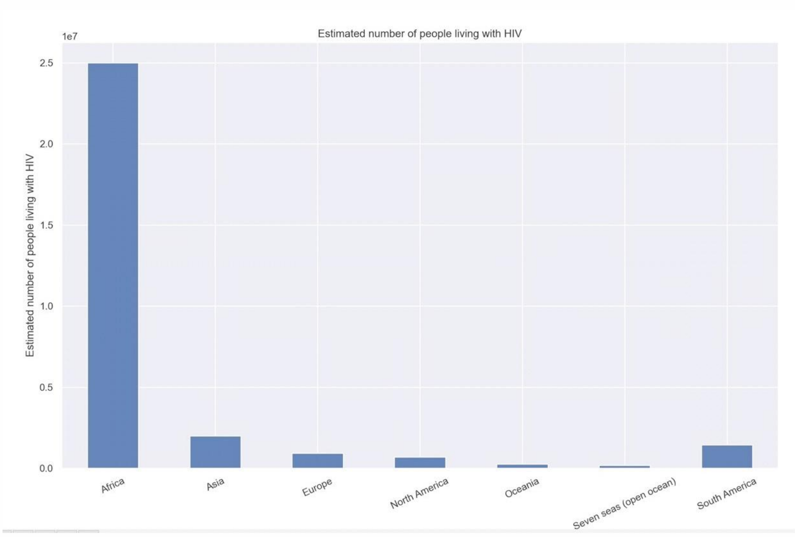
For the third question, “How does access to ART (HIV/AIDS treatment) compare across different countries?” We plan to merge the two datasets and generate data visualizations to see if the amount of ART accessibility in North America differs from that in a region such as Africa. Key columns we would be looking to answer those questions would be shows that different countries and the estimated number of cases of HIV/AIDS that have been reported, the number of people receiving antiretroviral therapy (ART), and the estimated ART coverage among people living with HIV. Also looking at current gender, race/ethnicity with HIV/AIDS.  This will allow us to investigate if there is a link between ART access and transmission and mortality rates on different continents. We would also add a model that could predict future ART coverage.

For our fourth question we will take all the countries and put them into categories based on continents so that way we have seven continents. Once we have the seven continents, we will take the overall deaths, incidence, and prevalence from our datasets. We will create a couple of charts to see which charts have a better data visualization trend.

## Results

### How do cases of HIV/AIDS compare across countries, especially comparing continents?

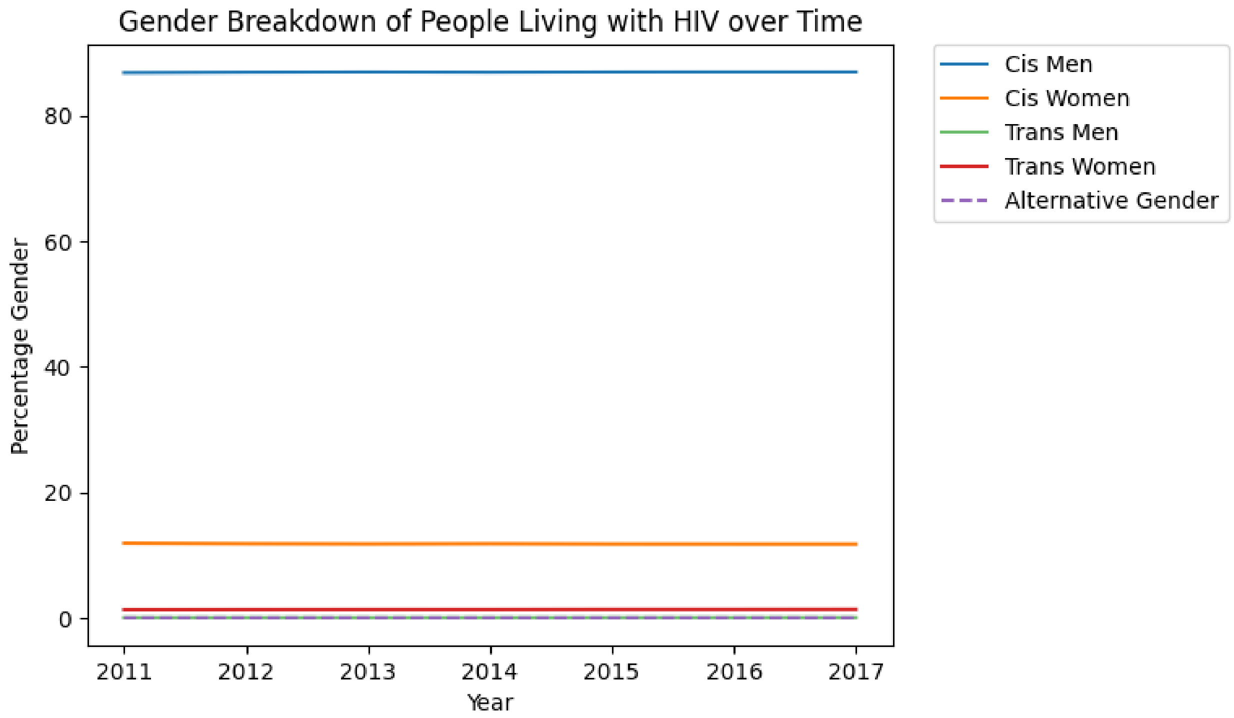
For our first question, we wanted to answer: How do cases of HIV/AIDS compare across countries, especially comparing continents? With that, we found a dataset that had all the countries and the number of cases in each country. First, we grouped the countries into continents then we added the cases based on which country lies on which continent. We created a bar chart for our visualizations, and we were able to see that Africa had the highest cases, and it was high compared to the other continents. We can't compare it to the different continents due to the big difference. The Seven Seas had the lowest number of people living with HIV.

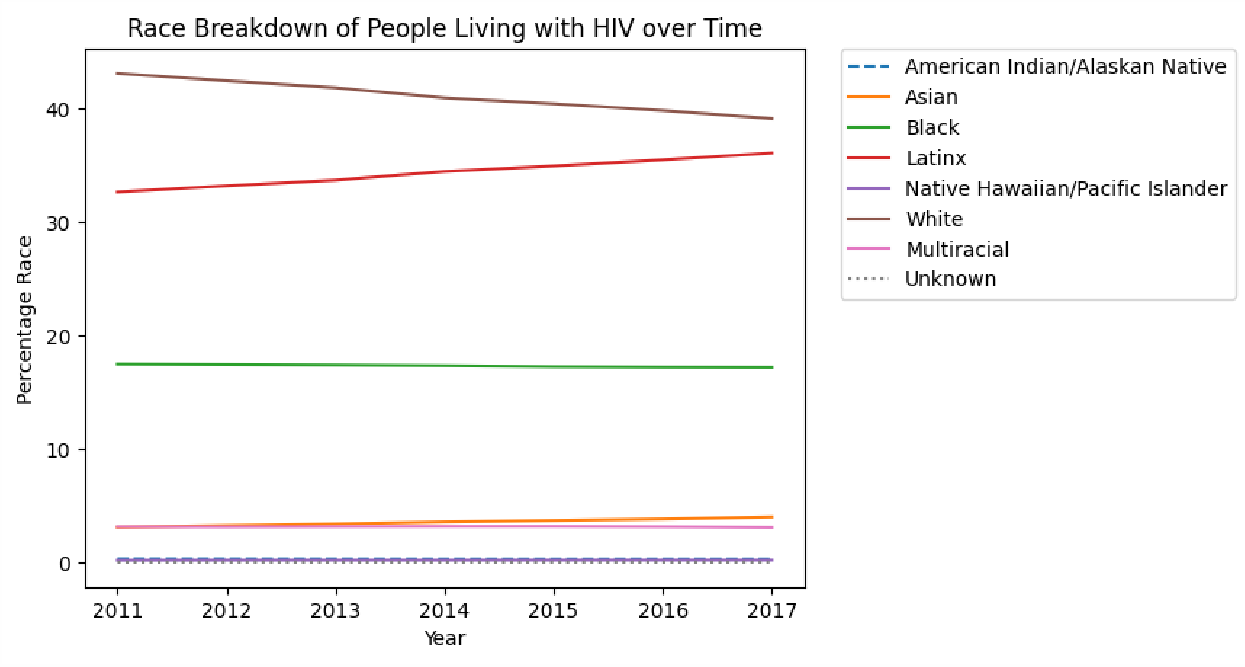


**Impacts and Limitations:** We were surprised by the results because we didn't expect to see that big difference. One of the harmful consequences of the graph is that people can say that Africa has a high HIV infection rate and say that every African has HIV. Even though Africa has a really high infection rate, we aren't taking into consideration the population. If we considered the population, maybe Africa would have a high cases rate but not as high as now.

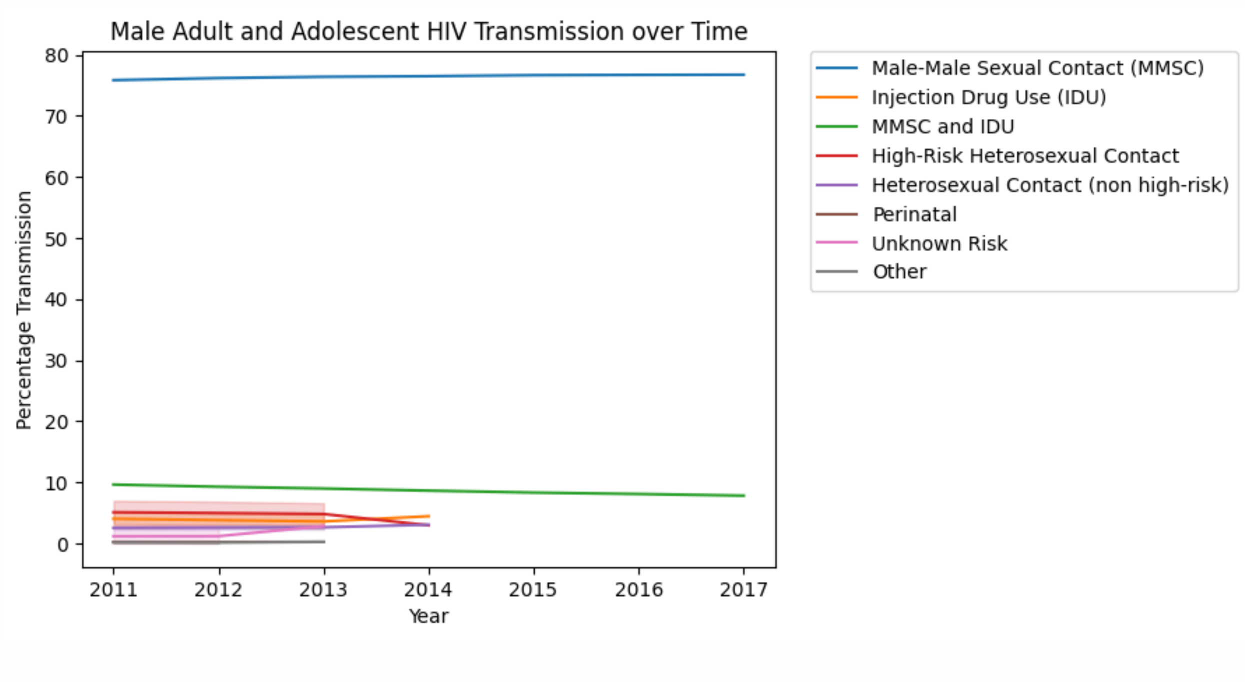
### How do cases of HIV/AIDS compare across different demographic categories (race, sexuality, gender identity) within the United States?

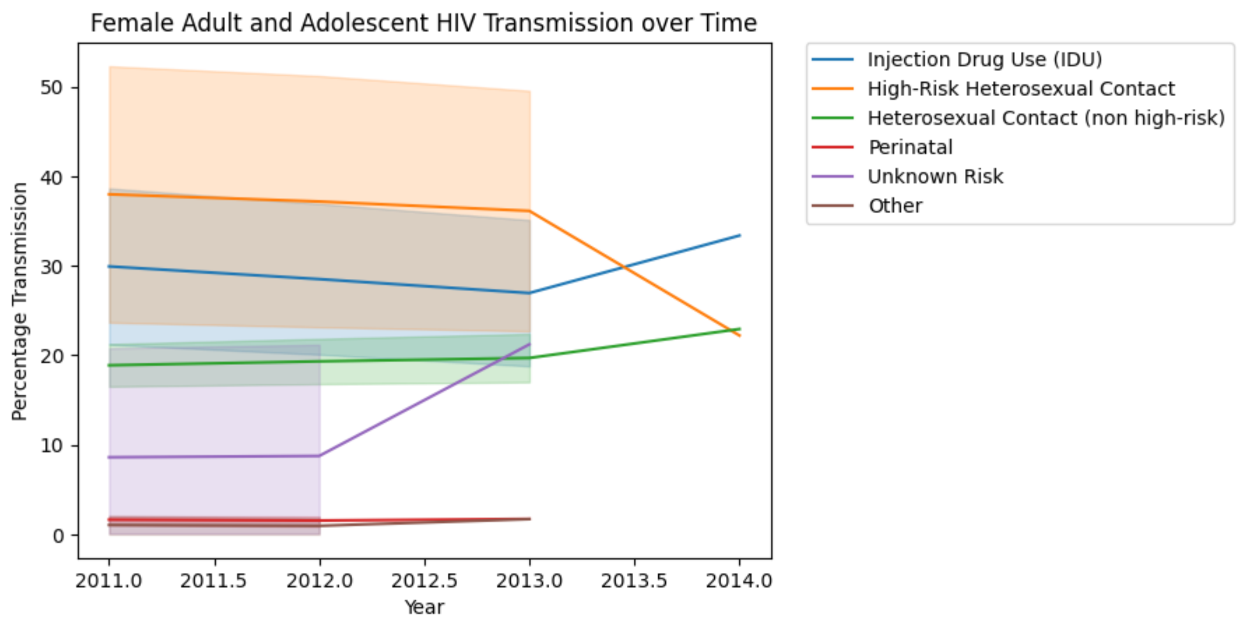
For this question, the biggest conclusions we could draw were about how the HIV/AIDS case percentage representation changed over time across categories within a demographic bucket. We immediately noticed that in terms of raw numbers, the percentage distributions for the gender and race demographic bucket was usually weighted the heaviest by the most populous / most dominant categories. For example, Out of all the genders, cisgender men most heavily overrepresent HIV/AIDS cases in the US, followed by cis women, trans women, trans men, then gender nonbinary individuals. Out of all the races, white people most heavily overrepresent cases of HIV/AIDS, followed by Latine, Black, Asian, multiracial, American Indians and Alaskan Natives, and Pacific Islanders. However, if we look at the change in percentage over time, we can draw more interesting conclusions. For example, we noticed that for the race percentages, the only category with a noticeable decrease in percent representation was white people, whereas Latine and Asian percentages trended upwards. This suggests that over time, less and less white people were infected with HIV as opposed to other races. However, gender distribution stayed relatively stagnant over time, suggesting that race is a stronger predictive and determining factor for changes in HIV cases over time.





As for transmission categories, which were split into male and female transmission (unfortunately, there was no data gathered on other genders), we noticed some similarities and differences. For one, out of all male adults/adolescents infected with HIV/AIDS, the vast majority contracted the disease from male-to-male sexual contact. However, for both male and female adult/adolescent transmission rates, the percentage for high-risk heterosexual contact (defined as “Heterosexual contact with a high risk partner HIV infected partner or a partner who is an IDU or MSM”, from the dataset website) noticeably decreased over time. This suggests that there may have been more effort to curb the spread of HIV in heterosexual communities than queer communities.





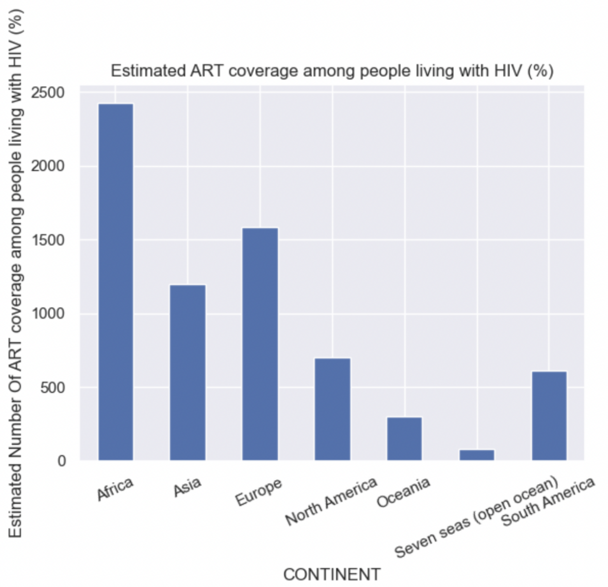
**Impacts and Limitations:**

For a future analysis, if we wanted to get a better understanding of the true weight and impact of the pandemic, we might instead try to compare percent infected within each category and compare those to others in the demographic bucket (e.g.. find out what percentage of cisgender men have HIV/AIDs compared to what percentage trans women have HIV/AIDS). This would give us more specific information for if certain demographics are more predisposed to have cases of HIV/AIDs – a big problem we noticed with our results for this question was that by raw numbers, demographics that had overall greater base populations (e.g. cis-men over non-binary individuals, or white people over American Indian/Alaskan Native people) were the most representative category within each demographic bucket. By weighting each category with their overall population, we can get a better sense for inequities between categories while taking into account skewed data and representations.

Unfortunately, this was extremely difficult to do with our limited dataset, since it did not include the total distribution of respondent answers (i.e. it did not have data for how many people did NOT have HIV/AIDS). It was possible for us to find aggregate census data of total US demographic breakdown for this time period, however, we feared that our own dataset did not collect its data in a wide enough scope for us to feel comfortable combining datasets, as it would be extremely likely for our dataset to be less extensively representative than the census data.

**How does access to ART (HIV/AIDS treatment) compare across different countries?**

We wanted to answer our third question: How does access to ART (HIV/AIDS treatment) compare across different countries? We found a dataset with all the countries and the number of people who got ART treatment. First, we grouped the countries into continents then we added the ART based on which country lies on which continent. For our data visualizations, we created a bar chart, and it showed us that Africa had the highest ART coverage, which isn't surprising considering that Africa has a high HIV number of people living with HIV.

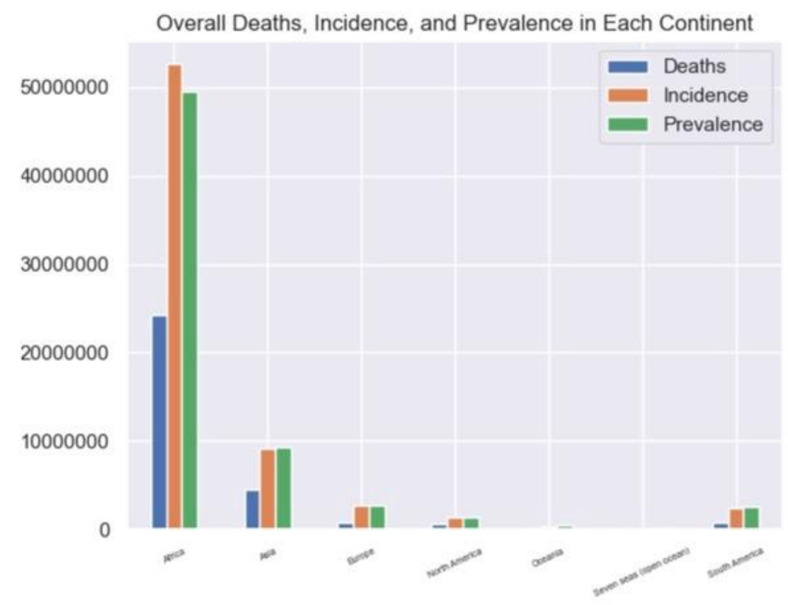


**Impacts and Limitations:** One thing that we found interesting was that Europe had a high ART coverage for people living with HIV. Still, the cases in Europe are low, so why is the estimated ART coverage among people living with HIV high compared to the estimated number of people living with HIV. One of the harmful consequences of the graph is that even though it shows that Africa has a high ART, it still doesn't factor in the population, which would mean the ART would be lower. There isn't enough ART coverage with high cases and lower ART, and there needs to be more.

### How does the overall deaths, incidence, and prevalence compare across continents?

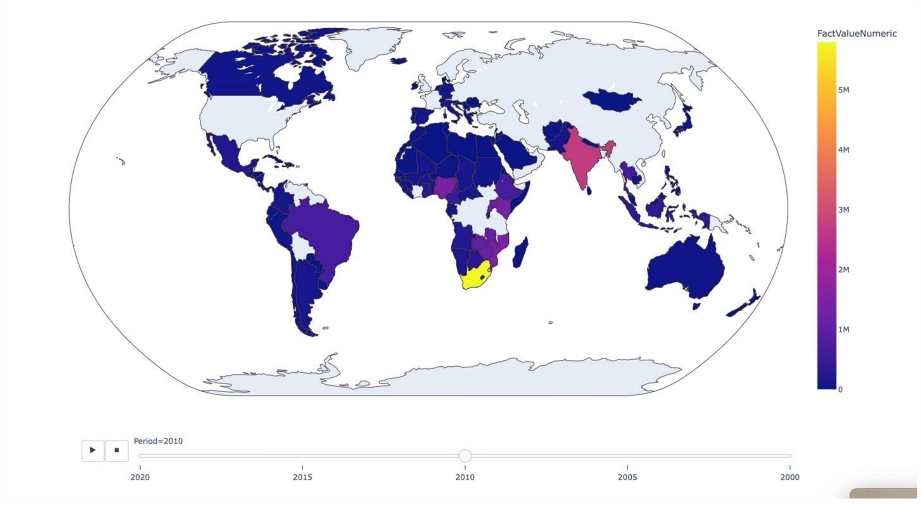
For our fourth question, we wanted to answer: How do the overall deaths, incidence, and prevalence compare across continents? We found a dataset with all the countries and the number of overall deaths, incidence, and prevalence in each country. First, we grouped the countries into continents, then we added the widespread deaths, incidence, and prevalence based on which country lies on which continent. We created a bar chart for our visualizations of deaths, incidence, and prevalence of all ages in three different colors. Blue is for deaths, orange is for incidence, and green is for prevalence. Africa had the highest deaths, incidence, and prevalence, and Asia second. Africa and Asia had a massive difference in deaths, incidence, and prevalence.

**Impacts and Limitations**: One of the harmful consequences of the graph is that people can say that Africa has a really high overall death, incidence, and prevalence of HIV/AIDS. We also need to understand that we didn't consider the population when making that chart. Africa has the second highest population, and maybe if we thought about the population, the difference wouldn't be that big.



## Challenge Goals

We changed our original goal from the beginning, and instead, we created a slider. We wanted to create a map of the world to see if HIV/AIDS is similar or different in all parts of the world. We also wanted to see if the infection rate had increased or decreased from 2020 to 2000. Our first big problem was finding a suitable countries and geometries dataset that would work with our library, and then merging our existing data to fit with the new geometries dataset. Another problem we ran into was that some countries didn't have data on the dataset, and it said no data, which meant that we couldn't have HIV/AIDS cases for the entire world. Overall, the new library threw a lot of curveballs at us, including different keyerrors or behaviors that we needed to resolve through re-reading documentation or looking at forum questions.



## Work Plan Evaluation

We quickly diverged from our initial work plan after discussing it with our TA mentor and reflecting/changing some of our research questions and challenge goals. Mainly, we decided that a machine learning component might not make sense or even be responsible given the limited amount of data we have, the limited scope that machine learning results might produce, or just the nature of data collected itself. As we've learned in class, machine learning has a history of making inaccurate and potentially harmful predictions/representations, especially when dealing with data collected in a context shaped by different racial and gendered layers (for example, AI sentencing for criminal convictions or crime-predictive facial recognition technology disproportionately harming Black populations). Thus, we changed our research plan to focus more on the visualization component instead.

During our first meeting after the TA check-in, we quickly set a new timeline for work that reflected our revised goals. The overall structure of our original work plan was still solid; we just tweaked our challenge goal timeline a bit. We stuck to our work plan relatively diligently, since we made sure to give ourselves some extra time buffers in case any issues or difficulties arose, and also maintained strong communication and consistent check-ins to hold ourselves and each other accountable.

## Testing

Given that our project mainly centered around data visualization, the bulk of our testing work centered on making sure that the cleaning and processing of the data was handled correctly. To do this, we decided to pare down our data (randomly) into a smaller test file, where we would run the same data cleaning/processing code and hand-verify the results. Paring the data down made the hand-verification process much more manageable, and we also couldn’t think of any relevant way to use assertEquals tests with our method or goals. Similarly, we were then able to more easily verify our data visualization results by hand on the smaller dataset and the processed test data, especially when comparing similar values across test cases and viewing graph output. After our testing process, we were much more confident that our results accurately and responsibly represented our datasets and research questions.

Some datasets could not be pared down due to interference with our code (e.g. the geometries or countries datasets we used for our sliders).

**Collaboration**

Aside from talking with our TA mentor Dylan Stockard, we were pretty self-sufficient in completing this project. We had to read some documentation files and some StackOverflow posts to figure out how to use the new library we found, but we felt very prepared by the 163 course content and everything else went pretty smoothly.