# Impacts of Social Backgrounds on People's Perception of the Risk of Getting AIDS*

An Analysis Based on the Kenya's Demographic and Health Survey in 1998

Charles Lu Mahak Jain Yujun Jiao

14 April 2022

**Abstract**

Using data extracted from Kenya's Demographic and Health Surveys in 1998, this analysis aims to study the Kenyans' perception of their risk of getting AIDS. By analyzing the age, marital status, number of sex partners, residence, province, and education of the survey's participants, results show that these factors all have different levels of impact on people's risk of being infected. R (R Core Team 2020) is used in this report to create visualizations for the results. In the future, this report can help scholars to dive deeper into the history of the sexually-transmissible disease, study the correlation between the transmission rate and social backgrounds, and aid the authorities in making governmental or health-related plans.

## Contents



*The data and code that support this report can be found in the Github repository: https://github.com/R300G/Kenya_Demographic_1998

# 1 Introduction

Being a sub-Saharan African country, Kenya bears one of the largest AIDS epidemics in the world. Despite the use of active antiretroviral therapy (ART) and other effective prevention methods, AIDS is still a major health problem that attracts the public's attention(Hartmann et al. 2021) in Kenya. As the best way to stop the spread of AIDS is to enhance people's knowledge about the disease itself, Demographic and Health Surveys were implemented to gather socio-demographic, behavioral, and ideological data on sexually transmissible diseases (Population et al. 1999).

While advocating changes to future health-related plans, it's also important to look back into the past to study the people's reactions, perceptions, and responses to similar matters in the past century. In the Kenya Demographic and Health Survey conducted in 1998, survey answers were collected to reflect the demographic, health, and living conditions of the average Kenyans(Population et al. 1999). More precisely, this report extracts data from chapter 10 of the survey report, where AIDs and other sexually transmitted diseases were discussed in detail. By using a data set taken and cleaned from two tables titled "AIDS-related knowledge," this report aims to analyze the impacts of genders, age groups, marital status, number of sexual partners, residence regions, provinces and education levels on people's perception of the risk of getting AIDS. In other words, this report analyzes the question: what are the chances of getting AIDS for people from different backgrounds? This report utilizes analytic programming language R (R Core Team 2020). R packages tidyverse (Wickham et al. 2019), janitor (Firke 2021) and dplyr (Wickham et al. 2021) are used to clean, organize and manipulate the data. Furthermore, R packages ggplot2 (Wickham 2016), scales (Wickham and Seidel 2020) and knitr (Xie 2021b) are used to create figures and tables. R packages bookdown (Xie 2021a) and tinytex (Xie 2021c) are used to generate the R markdown Report.

By studying the many ways in which the knowledge of AIDS can be influenced by the participants' personal backgrounds, this report came to the conclusion that all the variables presented in the report have an impact on people's perception of their chances of getting AIDS. As controlling the spread of AIDS is one of the most compelling topics in modern society, it's exceedingly important to deconstruct the risk of catching the virus for people with different lifestyles and backgrounds.

This report is divided into three sections. The first section outlines the data that has been used to support the later analysis. Next, the second section discusses the trends observed from the data. Finally, the last section explores the progress that has been made in this report, the weakness of the analysis, and the future directions. The data and code that support this report can be found in the Github repository: Kenya_Demographic_1998.

# 2 Data

## 2.1 Data Description

### 2.1.1 Variables

Data used in this paper is extracted from table 10.9.1: Perception of the risk of getting AIDS: women and table 10.9.2 Perception of the risk of getting AIDS: men from the Kenya Demographic and Health Survey 1998 report (Population et al. 1999). These two tables illustrate the female and male populations' perceptions of their risk of getting the AIDS virus. Based on the tables, the independent variables include genders, age groups, marital status, number of sexual partners other than spouse, residence, province, and education. This report focuses on all the variables from the data set: the gender, age, marital status, number of sexual partners, residence region, province and education. To sum up, all these variables are analyzed in this report, and their impacts on the response variable, people's perception of the risk of getting AIDS, are discussed.

Figure 1 and Figure 2 are screen-shots of the tables in the orginal DHS report. These two tables are made based on the same survey question: What is your risk to be infected by AIDS? Figure 1 shows the distribution of answers for female participants and Figure 2 shows the distribution of answers for male participants.

Table 10.9.1 Perception of the risk of getting AIDS: women

Percent distribution of women who have heard of AIDS by their perception of their risk of getting AIDS, according to background characteristics, Kenya 1998

| Background characteristic | Chances of getting AIDS | | | | | Total | Number of women |
|---|---|---|---|---|---|---|---|
| | No risk at all | Small | Moderate | Great | Don't know | | |
| **Age** | | | | | | | |
| 15-19 | 45.9 | 34.1 | 13.2 | 6.6 | 0.1 | 100.0 | 1,827 |
| 20-24 | 31.7 | 34.4 | 23.1 | 10.5 | 0.3 | 100.0 | 1,537 |
| 25-29 | 25.1 | 33.3 | 28.8 | 12.7 | 0.2 | 100.0 | 1,356 |
| 30-39 | 26.2 | 33.2 | 29.5 | 11.1 | 0.1 | 100.0 | 1,962 |
| 40-49 | 29.3 | 36.8 | 27.0 | 6.7 | 0.1 | 100.0 | 1,122 |
| **Marital status** | | | | | | | |
| Currently married | 27.1 | 33.8 | 28.2 | 10.7 | 0.2 | 100.0 | 4,785 |
| Formerly married | 30.9 | 36.9 | 22.2 | 9.5 | 0.3 | 100.0 | 668 |
| Never married | 42.7 | 34.2 | 15.7 | 7.2 | 0.1 | 100.0 | 2,351 |
| **No. of sexual partners other than husband in last 12 months** | | | | | | | |
| 0 | 33.3 | 34.2 | 23.2 | 9.1 | 0.2 | 100.0 | 6,523 |
| 1 | 28.1 | 34.4 | 26.2 | 11.1 | 0.2 | 100.0 | 1,014 |
| 2-3 | 20.0 | 24.5 | 36.3 | 19.2 | 0.0 | 100.0 | 155 |
| 4+ | (19.6) | (30.0) | (22.1) | (28.3) | (0.0) | 100.0 | 29 |
| Don't know/missing | 20.0 | 48.6 | 27.1 | 4.3 | 0.0 | 100.0 | 83 |
| **Residence** | | | | | | | |
| Urban | 30.4 | 33.1 | 22.4 | 13.8 | 0.3 | 100.0 | 1,821 |
| Rural | 32.7 | 34.5 | 24.4 | 8.3 | 0.1 | 100.0 | 5,983 |
| **Province** | | | | | | | |
| Nairobi | 27.3 | 32.5 | 24.9 | 15.1 | 0.2 | 100.0 | 768 |
| Central | 33.1 | 49.0 | 13.7 | 4.2 | 0.0 | 100.0 | 830 |
| Coast | 35.1 | 41.5 | 18.8 | 4.1 | 0.0 | 100.0 | 597 |
| Eastern | 45.4 | 28.2 | 22.9 | 3.5 | 0.1 | 100.0 | 1,378 |
| Nyanza | 20.8 | 33.4 | 25.8 | 19.8 | 0.1 | 100.0 | 1,687 |
| Rift Valley | 36.0 | 32.2 | 23.2 | 8.2 | 0.4 | 100.0 | 1,648 |
| Western | 27.3 | 31.3 | 35.3 | 5.9 | 0.1 | 100.0 | 896 |
| **Education** | | | | | | | |
| No education | 30.0 | 34.5 | 28.2 | 7.1 | 0.1 | 100.0 | 860 |
| Primary incomplete | 34.4 | 33.5 | 23.1 | 8.9 | 0.2 | 100.0 | 2,871 |
| Primary complete | 32.7 | 34.5 | 23.1 | 9.5 | 0.1 | 100.0 | 1,773 |
| Secondary+ | 29.7 | 34.7 | 24.1 | 11.4 | 0.2 | 100.0 | 2,300 |
| Total | 32.1 | 34.2 | 23.9 | 9.6 | 0.2 | 100.0 | 7,804 |

Note: Total includes 2 women who reported that they have AIDS. Figures in parentheses are based on 25-49 cases.

Figure 1: Table on Page 164 of DHS Report

Table 10.9.2 Perception of the risk of getting AIDS: men

Percent distribution of men who have heard of AIDS by their perception of their risk of getting AIDS, according to background characteristics, Kenya 1998

| Background characteristic | Chances of getting AIDS | | | | | Total | Number of men |
|---|---|---|---|---|---|---|---|
| | No risk at all | Small | Moderate | Great | Don't know | | |
| **Age** | | | | | | | |
| 15-19 | 40.0 | 44.2 | 11.9 | 3.8 | 0.1 | 100.0 | 805 |
| 20-24 | 28.3 | 45.1 | 20.9 | 5.7 | 0.1 | 100.0 | 581 |
| 25-29 | 26.8 | 48.2 | 19.6 | 5.3 | 0.0 | 100.0 | 462 |
| 30-39 | 25.0 | 48.6 | 20.6 | 5.6 | 0.2 | 100.0 | 789 |
| 40-49 | 24.6 | 50.9 | 19.5 | 4.8 | 0.1 | 100.0 | 565 |
| 50-54 | 30.2 | 55.2 | 10.6 | 4.0 | 0.0 | 100.0 | 183 |
| **Marital status** | | | | | | | |
| Currently married | 26.3 | 49.3 | 19.2 | 5.2 | 0.1 | 100.0 | 1,784 |
| Formerly married | 18.3 | 53.8 | 19.1 | 8.4 | 0.4 | 100.0 | 126 |
| Never married | 34.6 | 45.2 | 15.8 | 4.4 | 0.1 | 100.0 | 1,476 |
| **No. of sexual partners other than wife (wives) in last 12 months** | | | | | | | |
| 0 | 34.8 | 46.6 | 14.4 | 4.1 | 0.2 | 100.0 | 2,099 |
| 1 | 21.6 | 53.5 | 19.9 | 5.0 | 0.0 | 100.0 | 652 |
| 2-3 | 20.8 | 44.6 | 27.7 | 6.7 | 0.1 | 100.0 | 436 |
| 4+ | 23.0 | 41.4 | 24.2 | 11.4 | 0.0 | 100.0 | 166 |
| Don't know/missing | (8.4) | (70.2) | (21.4) | (0.0) | (0.0) | 100.0 | 32 |
| **Residence** | | | | | | | |
| Urban | 31.4 | 44.3 | 19.1 | 5.1 | 0.1 | 100.0 | 909 |
| Rural | 28.9 | 48.9 | 17.2 | 4.9 | 0.1 | 100.0 | 2,477 |
| **Province** | | | | | | | |
| Nairobi | 35.3 | 41.3 | 18.6 | 4.8 | 0.0 | 100.0 | 429 |
| Central | 25.9 | 53.3 | 16.2 | 4.6 | 0.0 | 100.0 | 338 |
| Coast | 48.8 | 32.7 | 15.3 | 2.0 | 1.1 | 100.0 | 240 |
| Eastern | 40.8 | 49.0 | 8.7 | 1.6 | 0.0 | 100.0 | 633 |
| Nyanza | 11.5 | 61.0 | 16.1 | 11.3 | 0.2 | 100.0 | 640 |
| Rift Valley | 34.6 | 35.3 | 25.2 | 4.9 | 0.0 | 100.0 | 753 |
| Western | 15.4 | 59.9 | 23.0 | 1.7 | 0.0 | 100.0 | 354 |
| **Education** | | | | | | | |
| No education | 34.3 | 45.1 | 15.6 | 3.3 | 1.6 | 100.0 | 128 |
| Primary incomplete | 34.9 | 43.6 | 16.9 | 4.5 | 0.0 | 100.0 | 1,037 |
| Primary complete | 27.8 | 48.5 | 17.5 | 6.2 | 0.0 | 100.0 | 833 |
| Secondary+ | 26.3 | 50.4 | 18.6 | 4.6 | 0.1 | 100.0 | 1,388 |
| Total | 29.6 | 47.6 | 17.7 | 4.9 | 0.1 | 100.0 | 3,386 |

Note: Figures in parentheses are based on 25-49 cases

Figure 2: Table on Page 165 of DHS Report

### 2.1.2 Similar Data Sets

The Kenya DHS 1998 provided many other interesting datasets that focused on various intriguing survey questions. For instance, table 10.8.1 in the DHS report outlined five AIDS-related questions and people's general knowledge about the AIDS disease. Nevertheless, having too many questions leads to the difficulty of coming up with a concrete topic and conclusion. Indeed, it's extremely redundant to have many similar questions studied in one report. Thus, the authors of this report have decided to use tables 10.9.1 and 10.9.2 for simplicity, practicability, and clarity purposes. By focusing on only one specific question, the report has an opportunity to explore the question and the variables in a more detailed, explicit, and professional manner.

### 2.1.3 The cleaning process

Variables from tables 10.9.1 and 10.9.2 were cleaned and reconstructed during data gathering and cleaning. In order to create and maintain a tidy dataset, the percent distribution of survey participants' answers "no risk at all", "small", "moderate" and "great" to the survey question "chances of getting AIDS" were combined into one column called "chances_of_getting_AIDS". A new column named "survey_answer" was also created to provide additional information on which survey answer the percent distribution under "chances_of_getting_AIDS" falls under. To continue maintaining a tidy dataset, the same treatment has been applied to background characteristics. Information on "age", "No. of sexual partners other than husband/wife in last 12 months", "province","education","marital status", and "residence" in the column was selected and renamed to "demographic_info", and a new column named "demographic_type" was also created to provide additional information on which type of "demographic_info" falls under. A table with the first 10 rows of the filtered and cleaned data was shown in table 1.

## 2.2 Methodology

The DHS program collects demographic data from countries across the world using model questionnaires. Compared to other surveys, the data collecting process for the DHS program is lengthy and selective. First, household questionnaires are distributed to the households in the studied region, and the specific information of the family members were collected. For example, the questionnaire collects the height, weight, age and other characteristic information from the children, men and women in the household. ("The Dhs Program," n.d.) Then, the program will select the qualified individuals for individual interviews. The qualified individuals are usually the reproductive age groups of women ranging in age of 15-49 years old and men of 15-59 years old ("The Dhs Program," n.d.). Moreover, the questionnaires are not always the same. Other types of questionnaires are also utilized by the DHS program to serve other topics such as education, health care providers, young adults, etc. ("The Dhs Program," n.d.) It's also important to note that the questionnaire provided by the DHS program differs from one country to another, as each country has unique

Table 1: The first 10 Rows of cleaned Dataset

| demographic_info | demographic_type | gender | survey_answer | chances_of_getting_aids |
|---|---|---|---|---|
| 15-19 | age group | male | no risk at all | 40.0 |
| 20-24 | age group | male | no risk at all | 28.3 |
| 25-29 | age group | male | no risk at all | 26.8 |
| 30-39 | age group | male | no risk at all | 25.0 |
| 40-49 | age group | male | no risk at all | 24.6 |
| 50-54 | age group | male | no risk at all | 30.2 |
| 15-19 | age group | male | small | 44.2 |
| 20-24 | age group | male | small | 45.1 |
| 25-29 | age group | male | small | 48.2 |
| 30-39 | age group | male | small | 48.6 |

systems, culture and laws. Hence different model questionnaires are employed in different countries.

## 2.3 Visualization

The following figures represent the visualizations for the percent distribution of each survey answer to the question "What are your chances of getting AIDS" across the selected survey participants' backgrounds of age, number of sexual partners, province, education, and gender. Due to the nature of the data, gender is the only variable that has cross data with other demographic information, which allows us to study the effect of its interactions with other demographic information. Thus, the visualizations for age, number of sexual partners, province, education, marital status, and residence were all split into males and females. The report aims to study the relationship between survey participants' perception of the risk of AIDS and their demographic background, and how each demographic information can influence their perception of receiving AIDS in the near future. Due to the large amount of figures, this section will showcase only the percentage distribution of survey answers by provinces. The data visualization for other demographic backgrounds are available in the apendix.

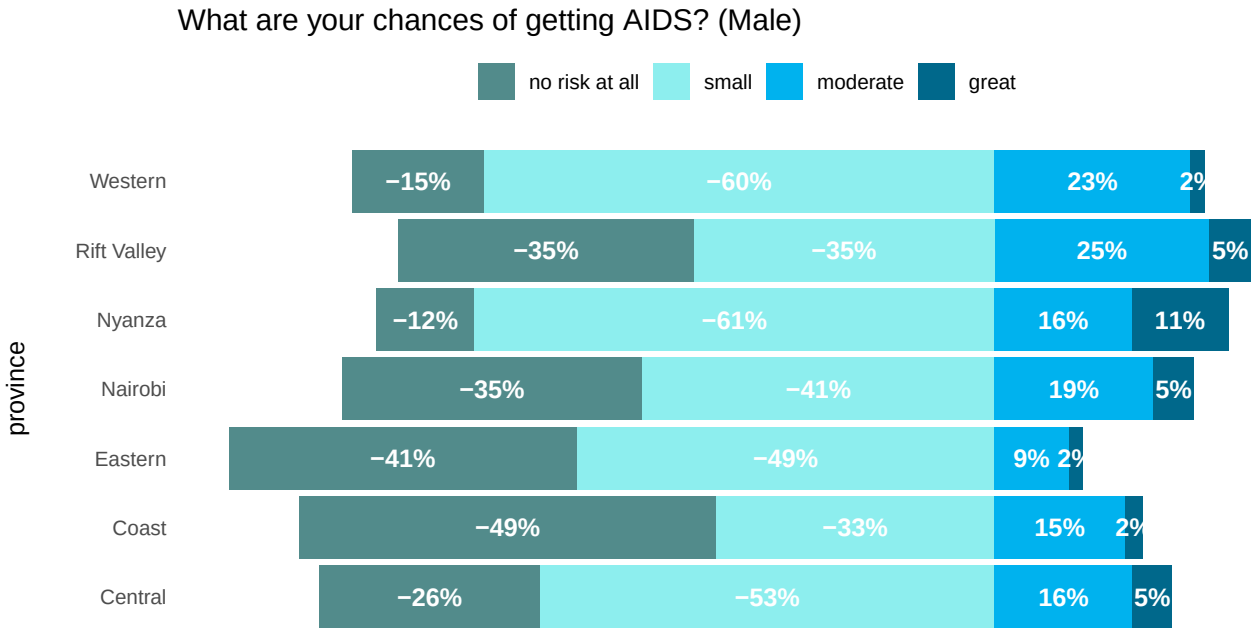


Figure 3: Percentage Distribution of Survey Answers by Provinces

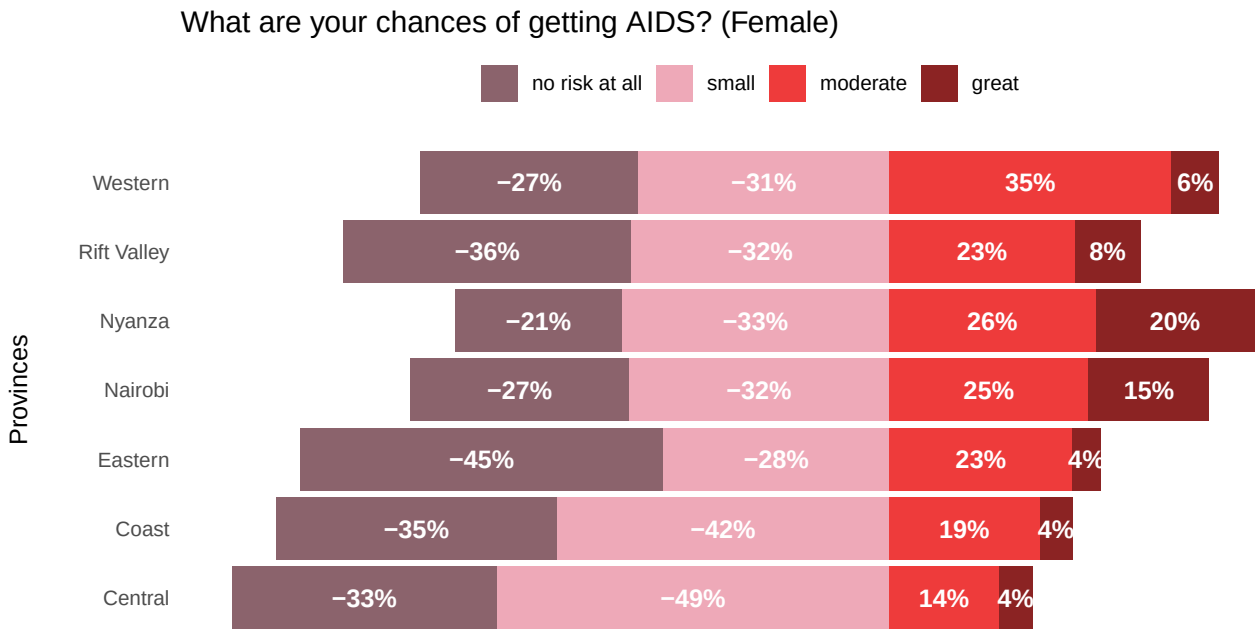


Figure 4: Percentage Distribution of Survey Answers by Provinces

# 3 Results

## 3.1 Age

Figure 5 shows the response distribution to "chances of getting AIDS" from survey participants of all ages. This figure has shown whether age has an influence on participants' selected survey answers. To start off, most participants aged 15-19 answered small or no risk at all, while only around 20% of them answered moderate or great risks. However, the distribution of answers changes as age increases. Firstly, the percentage of participants selecting the option "no risk at all" drops significantly as age increases and bounces back up a bit as age reaches 50+. On the other hand, the distribution of option "moderate" displayed the exact opposite trend compared to "no risk." There are two ways to interpret the trends, the occurrence of sexual activities and the perception of AIDS-related knowledge. The higher rate of "no risk" answers among aged 15 to 19 participants may be due to them being so young; sex and AIDS are new concepts to them. Thus, it is reasonable to assume that the young population has a lower occurrence of sexual activities and possesses a minimal amount of AIDS-related knowledge. When participants get older, their sexual activities and AIDS-related knowledge are likely to increase as they gain more experiences, which can be explained in the plot by both a downwards trend of "no risk at all" and an upwards trend of "moderate." As participants reach the age of 50 +, the occurrence of sexual activities is likely to decrease due to their maturity and or physical capacity, which explains in the plot as "no risk at all" curves up and "moderate" curves down when participants age further increases. However, it is difficult to interpret late adulthood's perception of AIDS as some may be well informed on AIDS-related knowledge while some may be too old or isolated to receive such knowledge. Thus, it is unclear whether the trend displayed for elders can represent the overall perception of AIDS-related knowledge.

The distribution of "small risk" shares a similar trend with "no risk at all" due to the same interpretations and reasons, and the same case applies to "great risk" with "moderate risk." However, both "small" and "great" appear to have flatter curves, as age seems to have a lesser influence when participants select these options.

When comparing the plots between females and males, the distribution of answering "no risk at all" is almost identical. However, females believe they have a higher risk of receiving AIDS as the plot shows that they are more likely to answer "moderate risk" and "great risk" than "small risk" when compared to males.

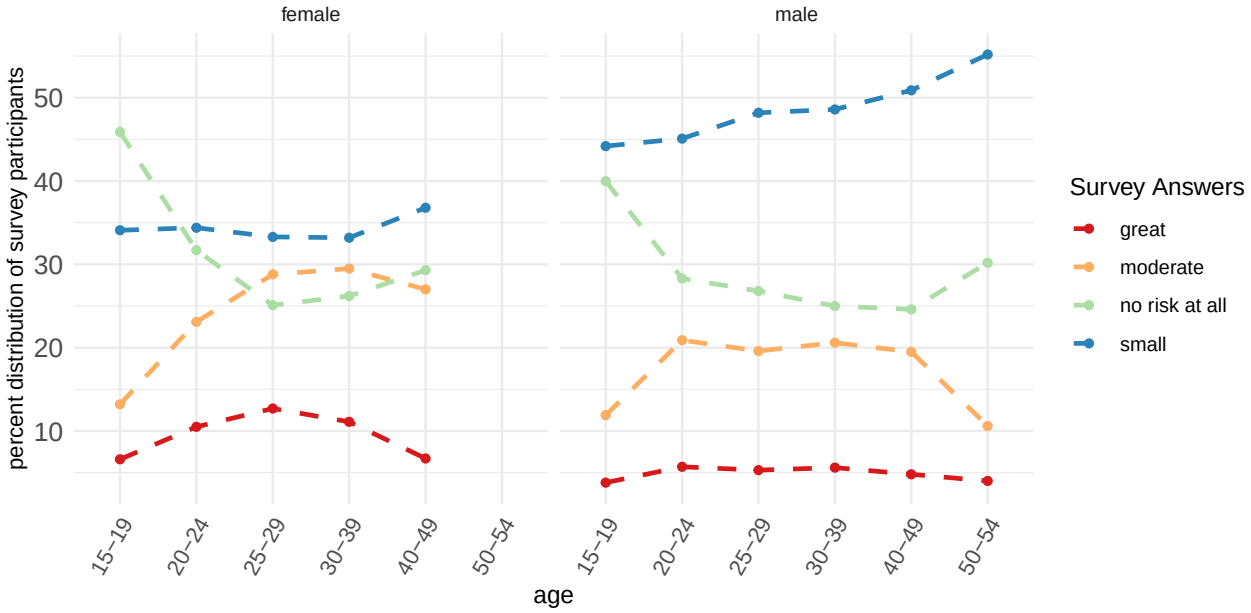


Figure 5: Survey Participants Distribution for Age

### 3.2 Number of Sexual Partners other than Husband/Wife in the past Year

Figures 6 and 7 show the correlation between participants' number of sexual partners other than husband/wife in the past 12 months, and participants' selected survey answers through linear and non-linear lines. Logically, having more sexual partners other than the husband/wife increases the frequency of sexual activities and makes the survey participants vulnerable to more potential sources of AIDS. Thus, it is reasonable to assume that the number of sexual partners other than the husband/wife positively correlates with the risk of receiving AIDS. Thus, there should be a downward trend for both answers "no risk at all" and "small," and an upward trend for both survey answers "moderate" and "great." Overall, 6 displayed the same trends as expected.

However, through non-linear lines in 7, the participants with 4 or more sexual partners other than husband/wife responded to the survey differently than expected. Although the percentage of "great risk" continues to rise, both females and males participants have a significant drop in the percentage of selecting "moderate risk" when it is expected to continue rising like "great risk." There is even an around 3% increase for male participants who have 4 or more sexual partners to select the option "no risk at all." It is clear that survey participants with 4+ sexual partners generally have a lower perception of the risk of getting AIDS, and the importance of prevention measures was neglected.

Similar to the age groups, when comparing the number of sexual partner plots between females and males, females seem to have higher awareness and stronger beliefs that they have a higher risk of receiving AIDS as the plot shows that they are more likely to answer "moderate risk" and "great risk" than "small risk" when compared to males.

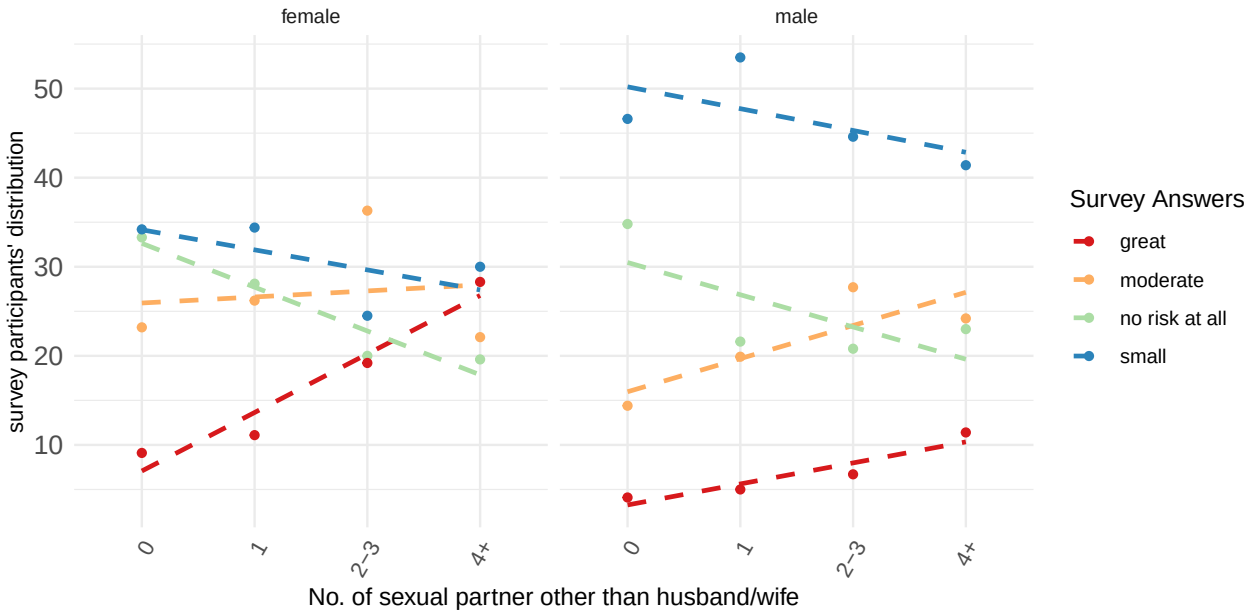


Figure 6: Survey Participants Distribution for Number of Sexual Partner (linear)

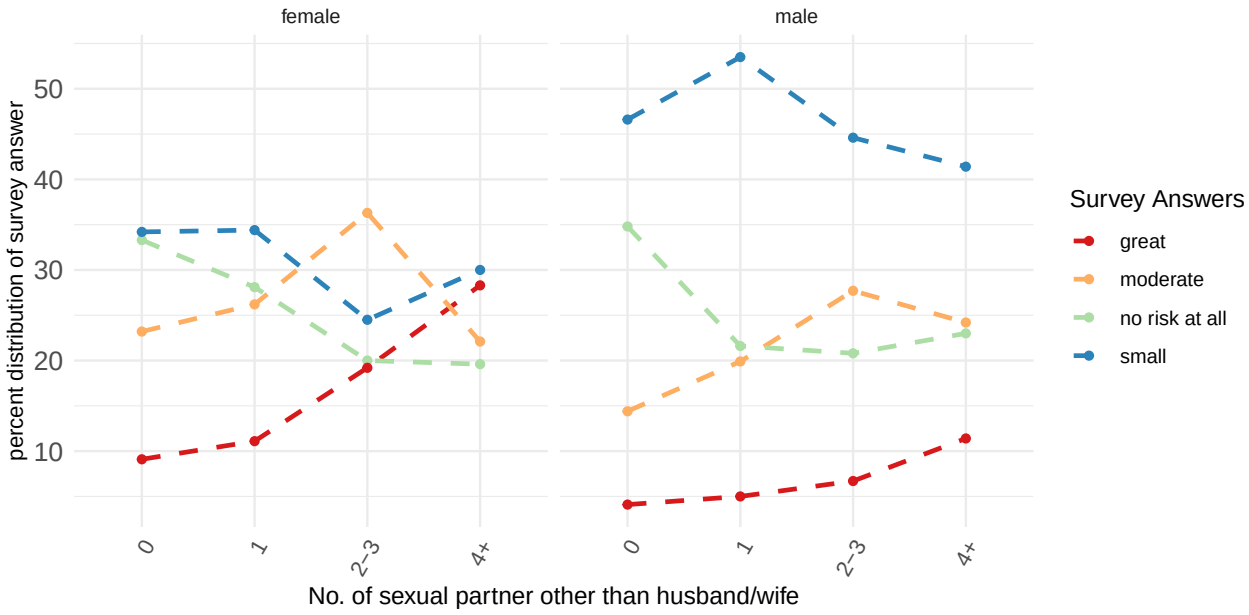


Figure 7: Survey Participant Distribution for Age (Non-Linear)

### 3.3 Province

Figure 3 and Figure 4 depict people's perception of their risk of getting AIDS from different provinces. Figure 3 demonstrates results from male survey participants, whereas Figure 4 shows the results from female participants. First of all, according to Figure 3, the results of the survey question were significantly affected by the region factor. In other words, people from different regions are very likely to have different perspectives regarding their chances of being infected. For instance, the number of people who answered "no risk at all" to the survey question in Coast Kenya(49%) is much higher than those who lived in Nyanza (12%). This feature is also visible in other categorical options such as "small," "moderate," and "great." In contrast, in Figure 4, women from different regions tend to have similar answers to their risk of getting AIDS. Even though there are small differences among the numbers of answers, their answers were consistent throughout the provinces. The only interesting point is that the number of women who thought they were very likely to catch the virus in Nyanza and Nairobi is much higher than in the other provinces. By combining the conclusions made in Figure 3, it's fair to conclude that both the male and female population living in Nyanza felt like they had a higher chance of getting the AIDS virus than the people living in other provinces.

Furthermore, Figure 8 is a multi-variable bar chart that shows the perception of the risk of getting AIDS from survey participants of both male and female living in different provinces. In the scatter-plot, the answers of each group of populations were illustrated. In the plot, the number of people who stated their chances of getting AIDS were "great" was highest in Kenya. Similarly, in Kenya, the number of people who claimed that they have "no risk at all" was also lowest in comparison to other provinces. In fact, Nyanza has an average AIDS rate of 14.7%, which is much higher than Kenya's national average (9%)(Bondo 2015). This social trend is primarily caused by the Luo ethnic group in Kenya, in which the women have low social status and cannot refuse to have sex with their husbands (Bondo 2015). In the Luo community, it's also common for men to have extramarital affairs. Thus, the dominant role of sex within the Luo community and the low status of women leads to the high incidence of AIDS in the Luo community.

In contrast, according to the survey participants, the risks of being infected were low in Coast, Western and Eastern Kenya. In Figure 8, the number of people who answered "great" was low, and the number of people who answered "no risk at all" was high in these three provinces. These features can be explained by the different cultures and communities existing within these areas. In a word, people may have contrasting views about sex, marriage, and sexual hygiene due to different traditions, customs, and ideologies.

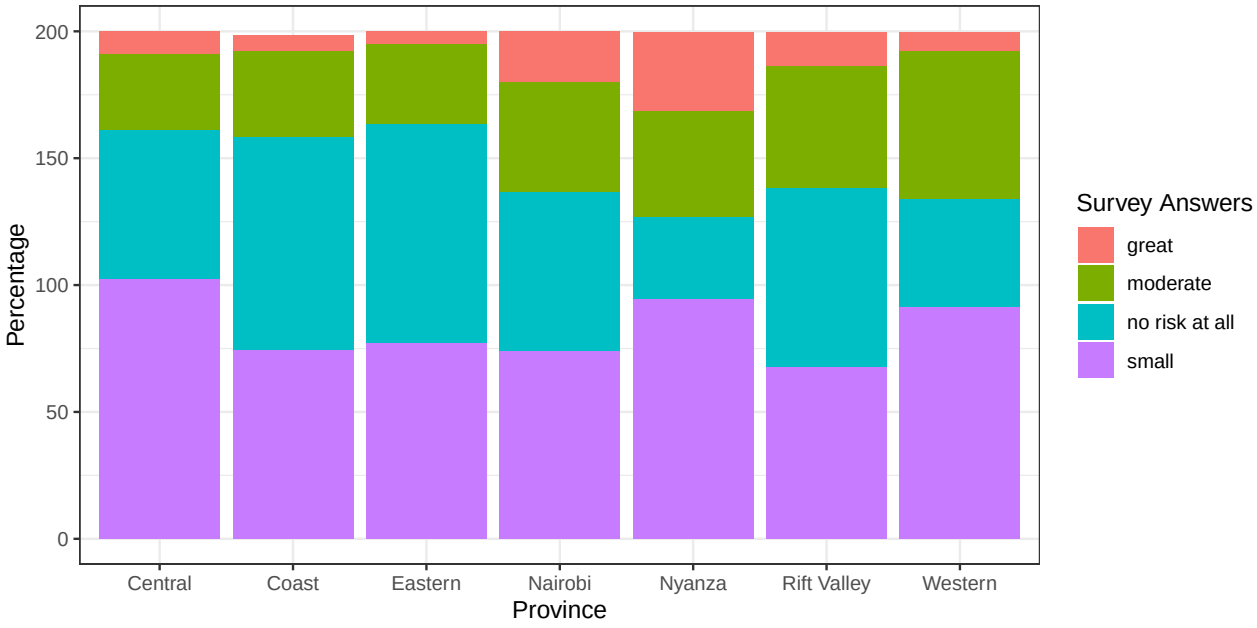


Figure 8: Survey Participant Distribution for Provinces

## 3.4 Education

Figure 16 shows the male participants' perceptions about their chances of getting the AIDS virus. In the graph, a great number of people thought that their chances of getting AIDS were small, and only a small portion of the population stated that their chances of being infected were "great." More precisely, for those with at least a secondary school degree, 26% of them claimed that they faced no risk at all. Similarly, 50% of them stated that their chances of being infected are small; 19% said their infected chance is moderate, and only 5% of them stated that their chances are great. To sum up, this bar chart shows that the data remains constant among different groups of people with different education levels. The small fluctuations that exist in the numbers can hardly explain the influence of different education levels on people's perceptions about getting infected. Besides, Figure 17 demonstrates the answers from female participants who held different levels of education degrees. Similar to Figure 16, the female participants with different educational backgrounds tend to have homogeneous answers in terms of their chances of getting AIDS. Nevertheless, a small difference existed in the graph: the number of women with at least a secondary degree said that their chances of getting infected (11%) were slightly higher than those with no education (7%).

The answers of the participants with different education backgrounds are also demonstrated in Figure 9. This scatterplot has indicated people with four types of school degrees: no education, primary incomplete, primary complete, and secondary. From the plot, the number of men who answered "small" when being asked about their risk of getting AIDS has a proportional relationship with their degree levels. That is to say, the higher their degree levels are, the higher chance that they state to have "small" chances of being infected. On the other hand, the curves observed in the plot for women are flat in comparison to those of the men. In addition, it's interesting to point out that in the women's plot, the chances of answering "great" increases with their education levels. Hence, the higher the degree is, the higher chance that the women think that they will have a great chance of getting the disease. It's fair to conclude that the education factor has a small impact on people's perception of the risk of getting AIDS.

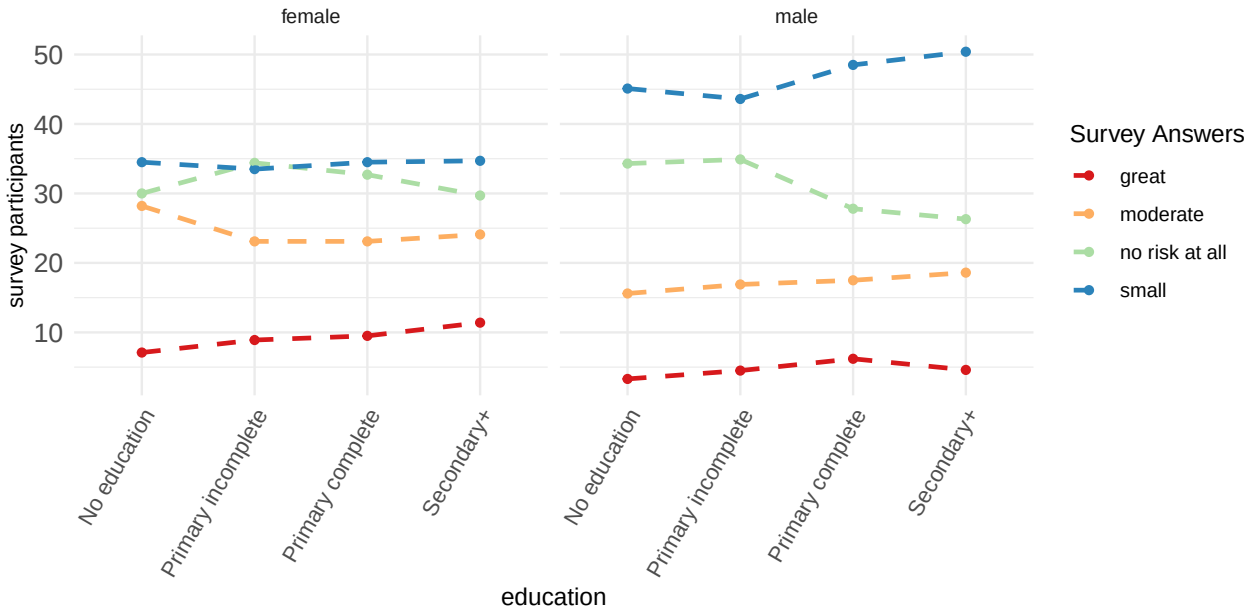


Figure 9: Survey Participant Distribution for Education

## 3.5 Marital Status

Figure 10 showcases the correlation between participants' marital status past 12 months and participants' selected survey answers through both linear and non-linear lines. For female and male participants, the line depicts that the "no risk at all" to the chances of getting AID increases once the participant is not married or is formerly married. Overall, the perception of getting AIDS amongst male and female participants is positively

correlated to the marital status, in terms of it increases when they're currently married, to never married as expected. The relationship between the perception of getting AIDS and it being a "small", "moderate" and "great" risk almost show a similar pattern for male and female participants both. Their perception of AIDS being a "small" risk is lower for those currently married, increases once formerly married, and decreases when never married, this pattern is similar to the male participants' "great" and "moderate" perception of getting AIDS as well. The pattern for "great" and "moderate" chances for female participants has a similar linear trend as they're decreasing.

The difference between the perception of male and female participants as shown in the graph Figure 18 and Figure 19 are based on several factors, some of which may include differences in societal expectations for men versus women, education, customs, and ideologies.

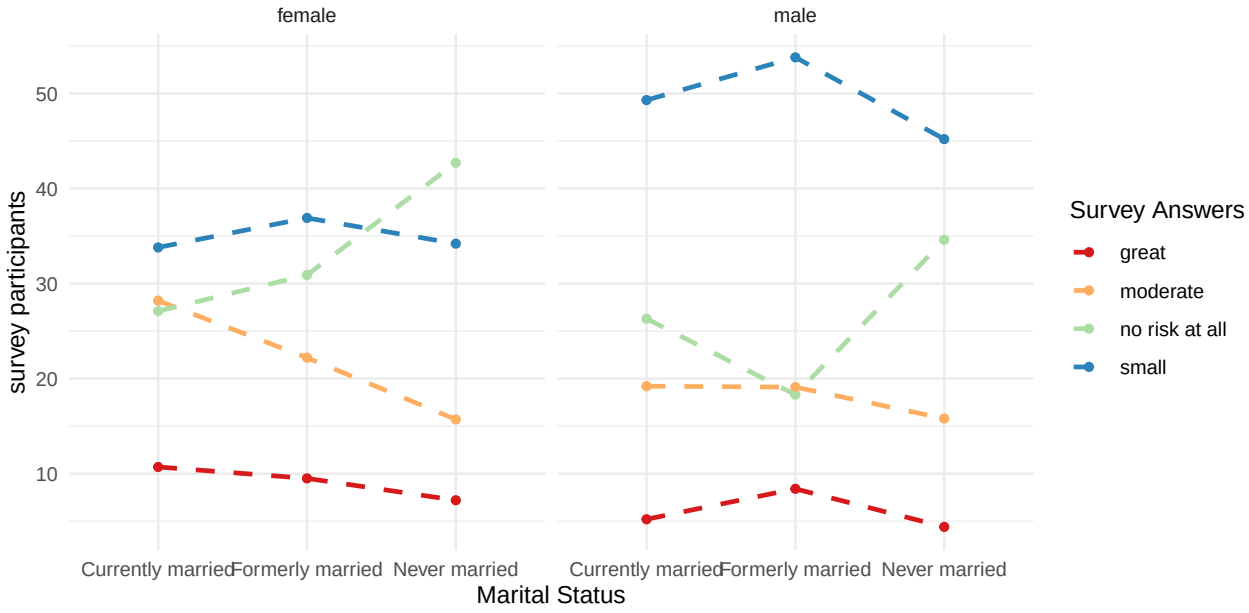


Figure 10: Survey Participant Distribution for Marital Status

## 3.6 Residence Region

Figure 11 showcases the correlation between participants' region of residence and participants' selected survey answers through both linear and non-linear lines. From societal understanding, it would be safe to assume that residents living in the urban area would have more access to information and the safety of diseases as there would be more sources of accessing that information. Thus, it is reasonable to assume that residents who reside in urban areas would have more behavioral changes in terms of perception of the chances of getting AIDS, due to the comparitve influx of information. Thus, we should expect downwards trends for both survey answers "no risk at all", "moderate" and "small" for urban compared to rural. Overall, the figure for female participants showcased a similar trend as expected.

The graphs below depict the answers of participants on rural and urban sides about their chances of getting the AIDS virus. There may be several factors affecting their perception of their chances of getting AIDS on both sides, some of which may include education, differences in societal expectations, marriage, customs, and ideologies.

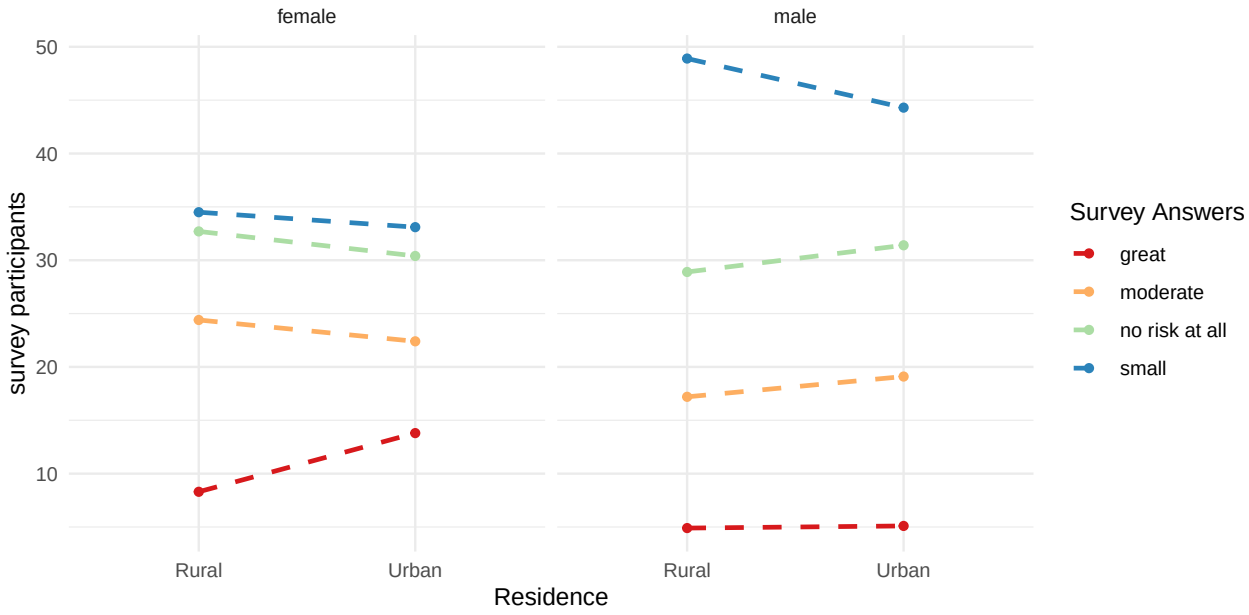


Figure 11: Survey Participant Distribution for Marital Status

# 4 Discussion

## 4.1 Summary

The perception of the risk of getting AIDS is considered to be the first stage toward behavioral change from risk-taking to safer behavior (Akwara, Madise, and Hinde 2003). From the findings above, the paper and results observe the perception and attitude of residents in Kenya towards AIDS and their chances of getting AIDS. The paper explores the relationship between education, marital status, province, number of sex partners, residence, and age with the perception of chances of getting AIDS. HIV/AIDS has been and continues to cause major concerns due to its devastating impact at present, there exists no cure or vaccine for HIV/AIDS and the only way to stop its spread is through behavioral and attitudinal change (Ndambuki et al. 2006) This paper helps in understanding how different socioeconomic factors affect the perception of the risk of getting HIV/AIDS and therefore helps the Government of Kenya could use this information to form comprehensive target policies for 1. Imparting skills for self-protection from vulnerability to sex and HIV/AIDS infection. 2. Provide life skills that build and enhance self-esteem, self-worth, and self-confidence. 3. Targeted guidance and spread of awareness for different socio-economic demographics.

## 4.2 HIV/AIDS in Kenya

There is a high rate of poverty and illiteracy in Kenya, which accounts for the difference in the residents with the respect to access to living regions and residence, access to different levels of education, and the perception of the chances of getting AIDS which is not accounted for by the Demographic and Health Surveys in 1998. Therefore there is no information on the income or economic status of the resident in this survey which could help eliminate bias.

The fact that HIV/AIDS is a multi-faceted, multi-sectoral problem was not understood and appreciated in the initial responses to the pandemic in Kenya (Ndambuki et al. 2006). It was much later that the need for a comprehensive policy was recognized and addressed seriously (Ndambuki et al. 2006). Due to the delay in the comprehensive policymaking, the awareness gradually took time penetrating to different provinces of Kenya. Along with this and the societal beliefs, ideologies and expectations, and cultural values in terms of hygienic sexual behavior, there were different rates of awareness in provinces. The province of Nyanza has the highest number of participants who think that they have a great risk of getting infected. The first AIDS case in Kenya was diagnosed in 1984 and after a year, in 1985, the government of Kenya established a

National AIDS Committee to advise on matters related to HIV/AIDS (Ndambuki et al. 2006). HIV/AIDs awareness was included in the education curriculum of schools in Kenya in 2000. As shown in Figure 9, a higher education level results in a greater chance of having the perception of a "small" risk of AIDS amongst male participants. This could be due to a higher awareness rate with education but since HIV/AIDS-related topics weren't included until after this survey, it is ambiguous.

Due to the lack of awareness and vaccines, the perception of the chances plays a huge role in behavioral changes toward the disease. There exists a higher awareness rate amongst female participants in general, as compared to the male participants. There is a lower awareness and perception of the risk of getting AIDS among the youth of Kenya, which is disadvantageous as it increases the risk and decreases the chances of behavioral changes. Economically, this low awareness comes with a disadvantage due to a higher male-to-female ratio in the labor force. Due to the low perception and risk amongst youth, risky behavior could translate to a higher chance of getting infected, lowering the participation in the labor force, and leading to a plummeting labor force productivity. The economic effects of the lack of implementation of policies resulted in a loss of labor supply and costs such as the direct costs of AIDS include expenditures for medical care, drugs, and funeral expenses, and indirect costs include lost time due to illness, recruitment, and training costs to replace workers, and care of orphans (Mwangi 2022).

## 4.3 Weakness

Weaknesses exist in this report as the data set used can be biased and incomplete. First, the data used in the original DHS report was collected from several types of questionnaires ("The Dhs Program," n.d.). People first need to complete a household questionnaire, and based on the responses; some qualified individuals will be selected to be interviewed using an individual questionnaire. For instance, most eligible survey participants are people of reproductive age (15-19 years old for women and 15-59 for men).("The Dhs Program," n.d.) Thus, a selection bias occurs during the data collecting process as only a certain group of the whole population was selected to contribute to the DHS program. With the issues discussed in this report, it's important to acknowledge that not only the reproductive group is exposed to this deadly disease. The children and the aged population also face the risk of getting AIDS through mother-to-child transmission and blood transmission(Zainiddinov and Habibov 2016). Moreover, the non-response bias and voluntary bias also contribute to the weakness of the data. As some people may not have the opportunity to answer the questionnaires and some people only answer the sections that they want to, the answers to the survey questions can also be biased.

In addition, the data used in this report was extracted from only two tables from the original DHS report: one table for men and one for women. Thus, the information regarding people's reactions to AIDS is very limited, and some of the critical features of the issue remain missing. For instance, we only have the percentage distribution of the whole population but not the actually recorded observation. This particular nature of the data made the dataset difficult to clean in order to maintain a tidy format, it also blocked us from studying the interaction effects of survey participants' demographic backgrounds on their choices of survey answer.

Indeed, understanding the perception of the risk of getting AIDS alone is not sufficient to make perfectly significant and powerful statements about the conditions of AIDS in the country. As this disease has bothered the Kenyans for decades and continues to threaten lives worldwide, its obligatory to focus on all aspects of the disease before making decisive answers to the problem.

It's always important for the readers to always acknowledge the flaws in this analysis, as they may potentially affect the effectiveness, clarity, and correctness of the statements made in the report.

## 4.4 Future Directions

AIDS is a transmissible disease that cannot be cured. AIDS/HIV has infected more than 35 million people worldwide, it can only be controlled with active therapy and medications(Deeks et al. 2016). In this report, by analyzing the answers from people from various backgrounds, the chances of them being infected with AIDS were analyzed in detail. However, studying people's perceptions alone is not enough to make fair

conclusions about the nature, the trends, and the effective prevention methods of the disease. Other factors such as people's knowledge about AIDS, the universalization rate of effective prevention methods, and the availability of healthcare in the studied region are also crucial to the impacts of the disease. In addition, the DHS report studied in this analysis was published in 1998, signifying that it has lost some relevance to modern society. Hence, in the future, more up-to-date reports will be reviewed to achieve a better understanding of the different aspects of the disease. In hopes of reducing the negative influence of AIDS on people's daily life, this analysis aims to educate the public about the seriousness of sexually transmissible diseases, outline the factors that increase the infection rate, and promote future governmental and health plans.

# 5 Apendix

## 5.1 Data Visualization

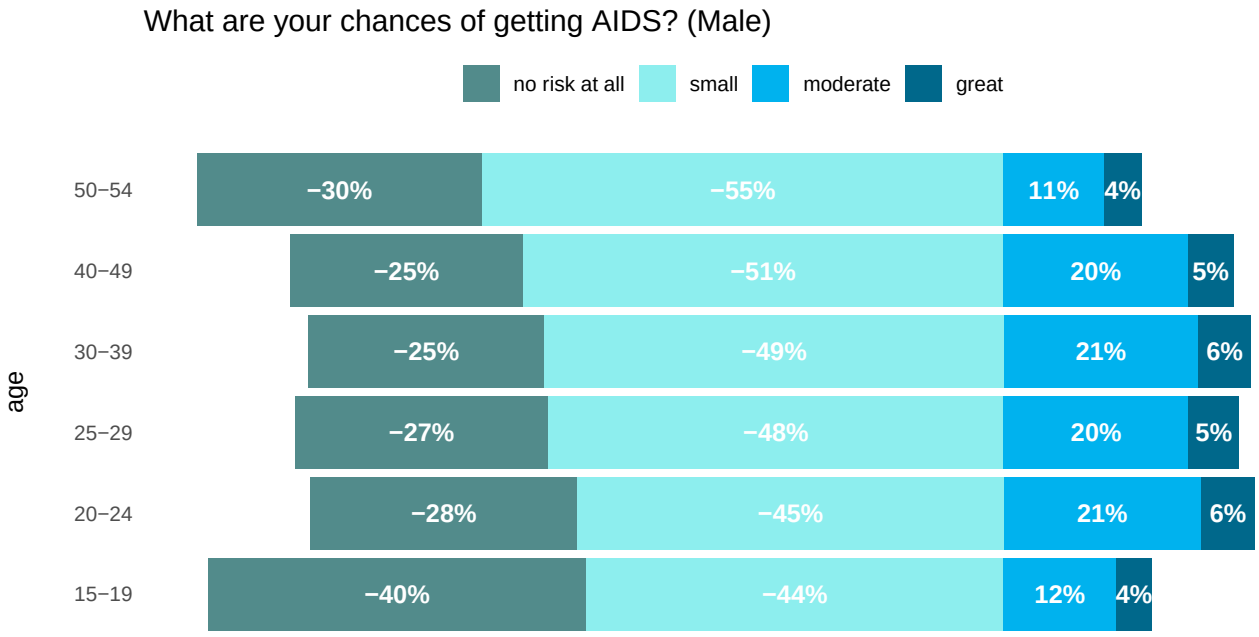


Figure 12: Percentage Distribution of Survey Answers by Age

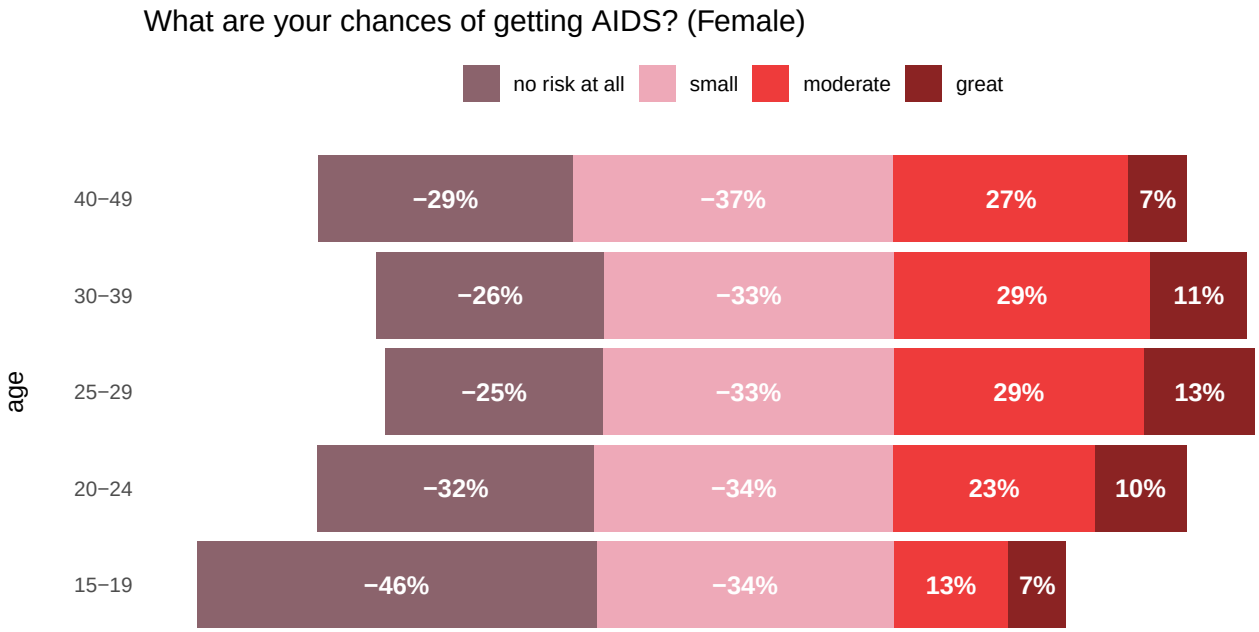


Figure 13: Percentage Distribution of Survey Answers by Age

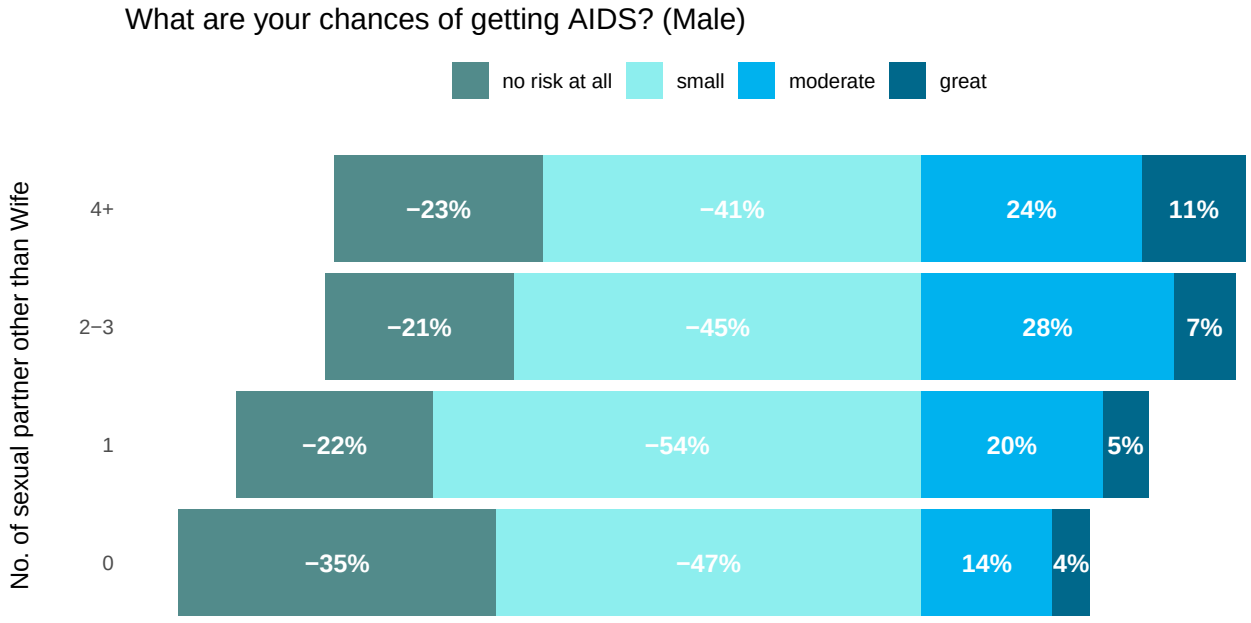


Figure 14: Percentage Distribution of Survey Answers by Age

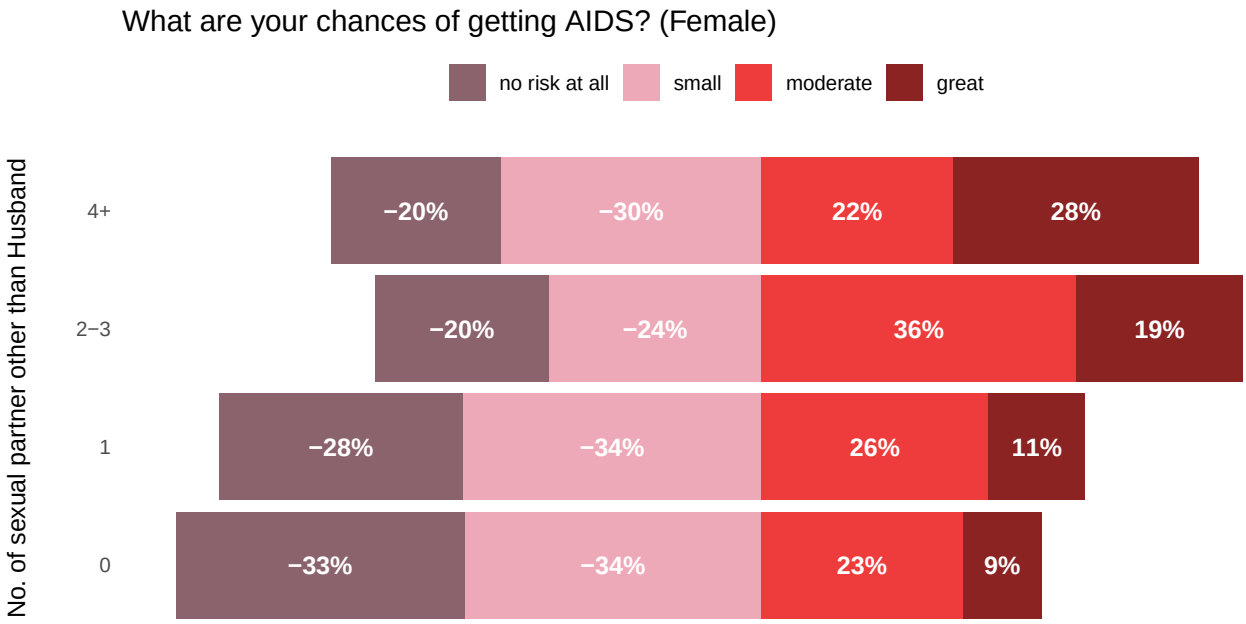


Figure 15: Percentage Distribution of Survey Answers by Number of Sexual Partners

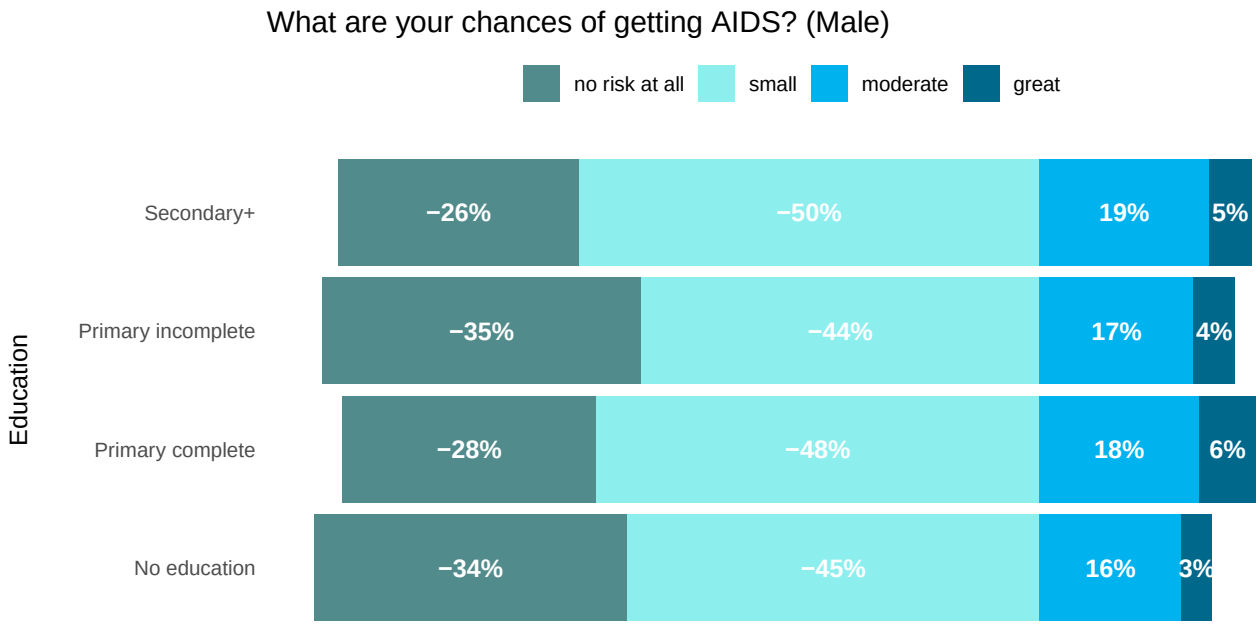


Figure 16: Percentage Distribution of Survey Answers by Education Level

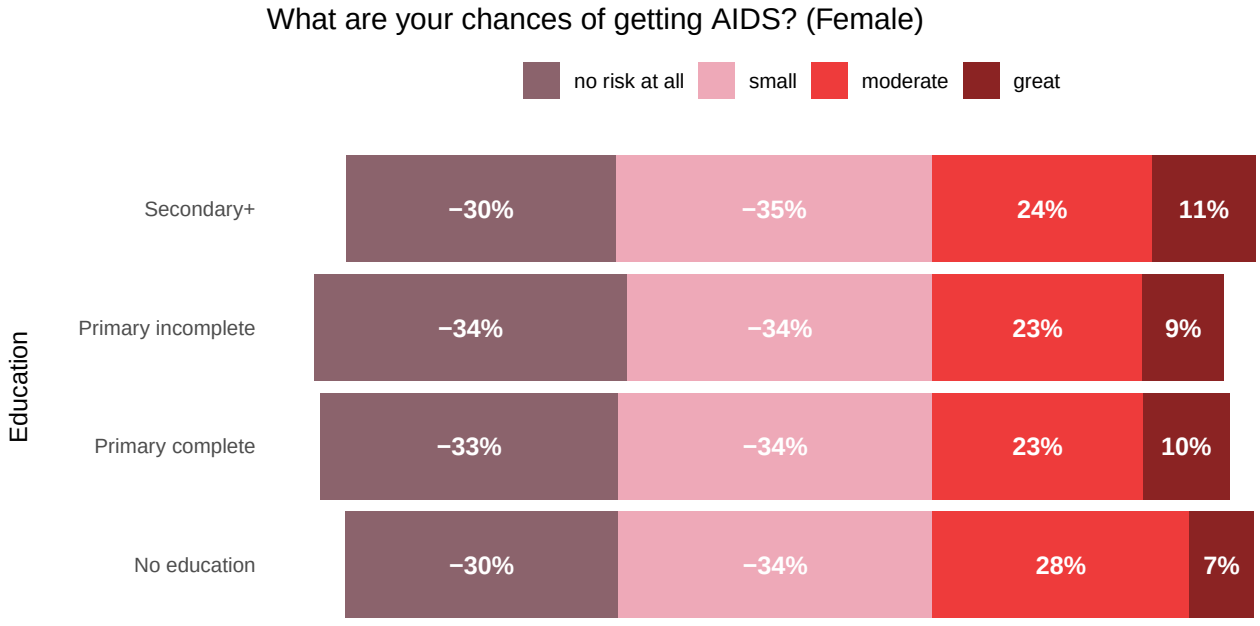


Figure 17: Percentage Distribution of Survey Answers by Education Level

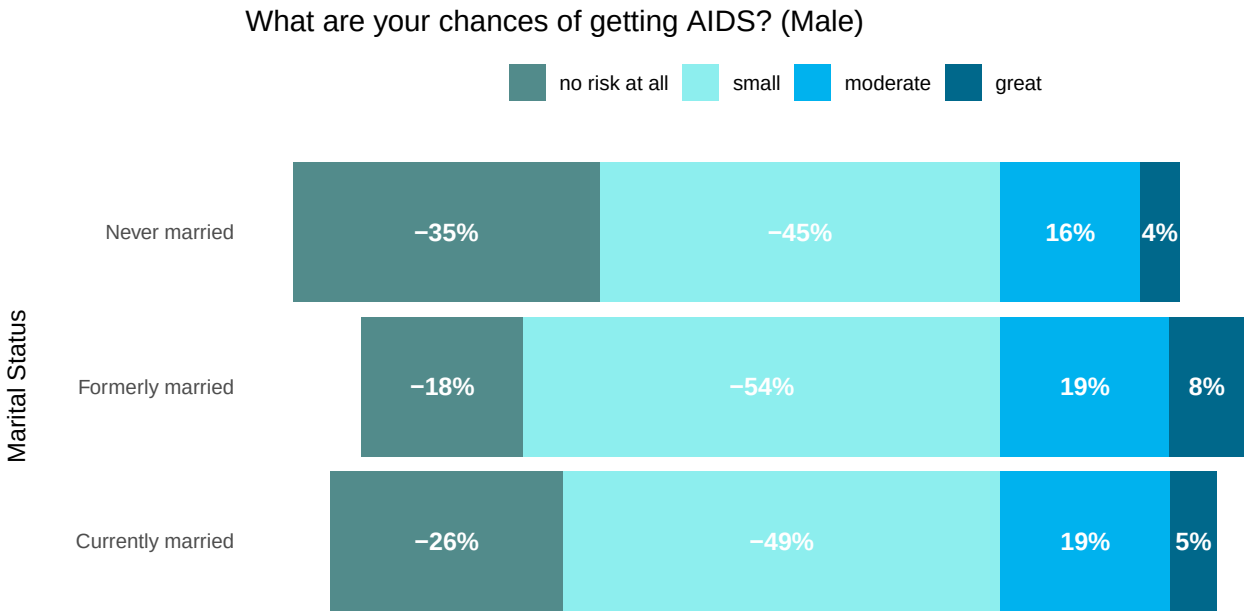


Figure 18: Percentage Distribution of Survey Answers by Marital Status

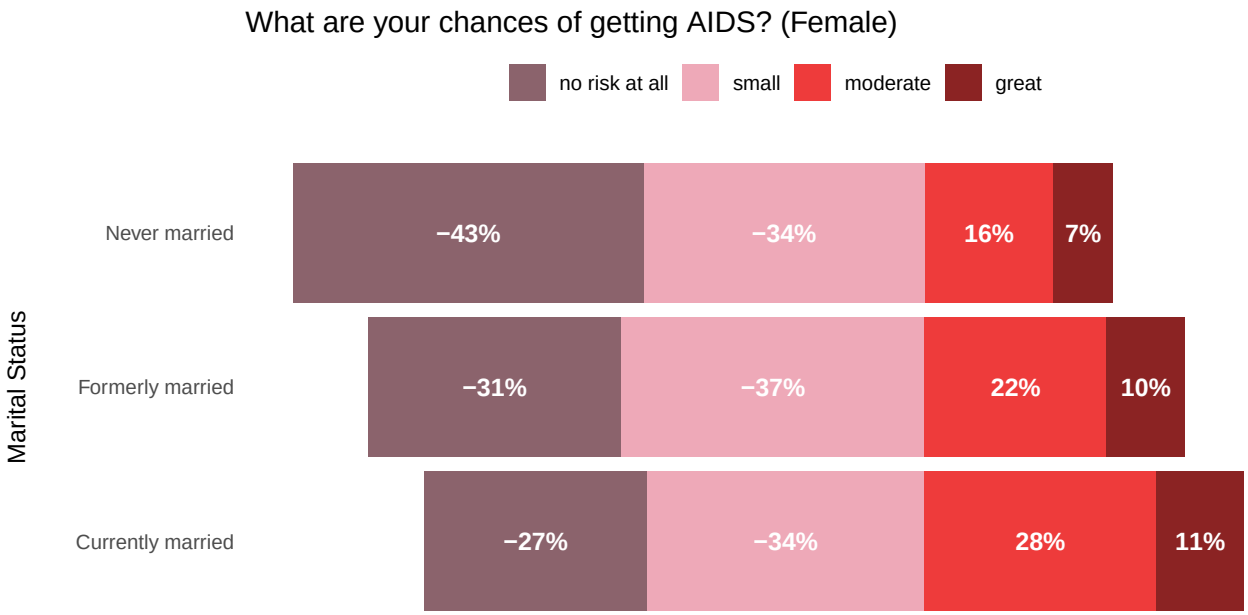


Figure 19: Percentage Distribution of Survey Answers by Marital Status

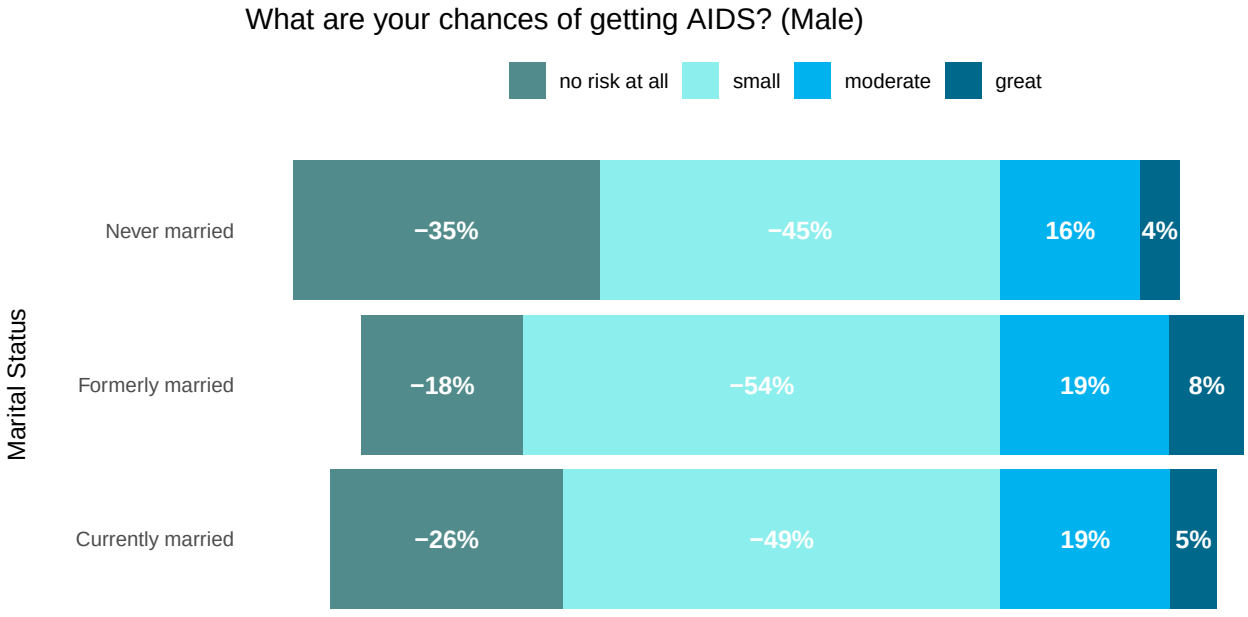


Figure 20: Percentage Distribution of Survey Answers by Residence

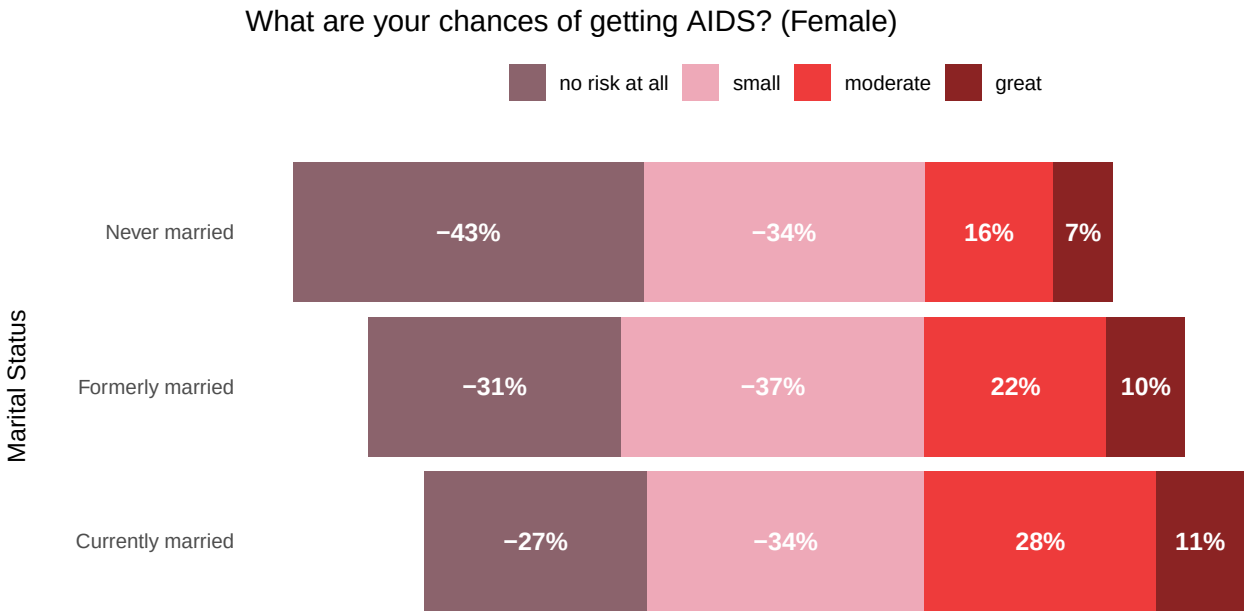


Figure 21: Percentage Distribution of Survey Answers by Residence

## 5.2 Datasheet

**Motivation**

1. For what purpose was the dataset created?

- Using data extracted from Kenya's Demographic and Health Surveys in 1998, this analysis aims to study the Kenyans' perception of their risk of getting AIDS. By analyzing the age, marital status, number of sex partners, residence, province, and education of the survey's participants, results show that these factors all have different levels of impact on people's risk of being infected. In the future, this report can help scholars to dive deeper into the history of the sexually-transmissible disease, study the correlation between the transmission rate and social backgrounds, and aid the authorities in making governmental or health-related plans.

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

- The Republic of Kenya, Central Bureau of Statistics

3. Who funded the creation of the dataset?

- U.S. Agency for International Development (USAID/Nairobi) and the Department for International Development (DFID/U.K.)

**Composition**

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

- Respondent information: age, education, marital status, residence, province, number of sexual partners.

2. How many instances are there in total (of each type, if appropriate)?

- Three types of questionnaires were used in the 1998 KDHS: the Household Questionnaire, the Women's Questionnaire, and the Men's Questionnaire.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

- The larger set was the 1998 KDHS consisting of 3 datasets. For the dataset used in the paper, a smaller sample of the larger sample is taken into account. It is representative in terms of regional coverage as it includes the same respondents but focuses on a question which is: Chances of getting AIDS? The raw dataset consists of two tables. The table is on Kenya's Demographic and Health Surveys in 1998, Chapter 10, Table 10,9.1 and 10.9.2

4. Is there a label or target associated with each instance?

- Table 10.9.1 Perception of the risk of getting AIDS: women: Percent distribution of women who have heard of AIDS by their perception of their risk of getting AIDS, according to background characteristics, Kenya 1998 and Table 10.9.2 Perception of the risk of getting AIDS: men: Percent distribution of men who have heard of AIDS by their perception of their risk of getting AIDS, according to background characteristics, Kenya 1998

5. Is any information missing from individual instances?

- Information on the income level of respondents is not included in the table.

6. Are there any errors, sources of noise, or redundancies in the dataset?

- There are no errors, sources of noise or redundancies in the dataset.

7. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

- The dataset is self-contained.

8. Does the dataset contain data that might be considered confidential?

- There is no confidential data, the dataset is publicly available.

9. Does the dataset identify any subpopulations (e.g., by age, gender)?

- The dataset comprises men and women aged 15 years or above.

10. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

- It is not possible to identify individuals.

11. Does the dataset contain data that might be considered sensitive in any way?

- Sensitive information may include but is not limited to: Number of sexual partners for men and women before 18, The Percent of women aged 15-19 and 20-24 married before 18. Percent of men aged 15-19 and 20-24 married before 18.

13. Any other comments?

- None.

**Collection Process**

1. How was the data associated with each instance acquired?

- The data set was collected by the DHS program in Kenya. Details can be found in the Data Methodology section of the report.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

- Household questionnaires, in-person interviews, and other types of model questionnaires were employed.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

- The enumeration areas are drawn from census files, and samples of households were selected from the enumeration areas.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

- The DHS program members were involved in the data collection process.

5. Over what timeframe was the data collected?

- The data used in this report was collected in 1998.

6. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

- The data is obtained from the official website of the DHS program. The link is https://dhsprogram.com/

7. Were the individuals in question notified about the data collection?

- The individuals of the question are well aware of the survey as well as the data collection.

8. Did the individuals in question consent to the collection and use of their data?

- The individuals in question have consented to the collection and use of their data.

9. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

- There's no open information available to tell if the consent for revokes were granted to the survey participants.

10. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

- The analysis of the original survey data was conducted by the National Council for Population and 11. Development of the Republic of Kenya. The link to the report is: https://dhsprogram.com/pubs/pdf/FR102/FR102.pdf

12. Any other comments?

- None.

**Preprocessing/ cleaning/ labeling**

1. Was any preprocessing/cleaning/labeling of the data done?

- The data was originally obtained in PDF format. The table from the survey PDF was converted to a usable data frame in R using the library pdftools for R.

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

- The raw data obtained is saved in outputs/data/raw_data_male.csv and outputs/data/raw_data_female.csv

3. Is the software that was used to preprocess/clean/label the data available?

- R software is available at https://www.R-project.org/

**Uses**

1. Has the dataset been used for any tasks already?

- The dataset has not been used for other tasks yet.

2. Is there a repository that links to any or all papers or systems that use the dataset?

- https://github.com/R300G/Kenya_Demographic_1998.git

3. What (other) tasks could the dataset be used for?

- The dataset would be used for understanding behavior to disease amongst different socio-economic demographic

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

- The cleaning process is very specific to the way this table was formatted in the original PDF and may not work on other tables.

**Distribution**

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

- No. The data set is for personal use only.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

- The dataset can be found in the Github repository mentioned in the report.

3. When will the dataset be distributed?

- The dataset has already been distributed in April 2022.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

- The report is under the MIT license.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

- There are no IP-based or other restrictions on the data.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

- No.

7. Any other comments?

- None.

**Maintenance**

1. Who will be supporting/hosting/maintaining the dataset?

- All three authors of the report will be maintaining the dataset. They are Mahak Jain, Yujun Jiao and Charles Lu

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

- They can be contacted via the email addresses listed in the README file on Github.

3. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

- No dataset updates are scheduled yet.

4. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?

- The data set was collected by the DHS program in 1998. There are no application limits on the retention of the data since it was collected voluntarily.

5. Will older versions of the dataset continue to be supported/hosted/maintained?

- No. Only the current data is available for now and future.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

- The extensions, augmentations and contributions from other users are not supported.

# References


Akwara, Priscilla A., Nyovani Janet Madise, and Andrew Hinde. 2003. "Perception of Risk of Hiv/Aids and Sexual Behaviour in Kenya: Journal of Biosocial Science." *Cambridge Core*. Cambridge University Press. https://www.cambridge.org/core/journals/journal-of-biosocial-science/article/abs/perception-of-risk-of-hivaids-and-sexual-behaviour-in-kenya/530D86938267A12D5160591D4000EC36.

Bondo. 2015. "Cultural Traditions Fuel the Spread of Hiv/Aids." *The New Humanitarian*. https://www.thenewhumanitarian.org/report/57401/kenya-cultural-traditions-fuel-spread-hivaids.

Deeks, Steven G, Sharon R Lewin, Anna Laura Ross, Jintanat Ananworanich, Monsef Benkirane, Paula Cannon, Nicolas Chomont, et al. 2016. "International Aids Society Global Scientific Strategy: Towards an Hiv Cure 2016." *Nature Medicine* 22 (8): 839–50.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Hartmann, Miriam, Alexandra M Minnis, Emily Krogstad, Sheily Ndwayana, Siyaxolisa Sindelo, Millicent Atujuna, Shannon O'Rourke, Linda-Gail Bekker, and Elizabeth T Montgomery. 2021. "IPrevent: Engaging Youth as Long-Acting Hiv Prevention Product Co-Researchers in Cape Town, South Africa." *African Journal of AIDS Research* 20 (4): 277–86.

Mwangi, Paul. 2022. "Economic Burden and Mental Health of Primary Caregivers of Perinatally Hiv Infected Adolescents from Kilifi, Kenya." *Academia.edu*. Research Square. https://www.academia.edu/74285150/Economic_burden_and_mental_health_of_primary_caregivers_of_perinatally_HIV_infected_adolescents_from_Kilifi_Kenya.

Ndambuki, Dr. Jacinta K, Elena McCretton, Nigel Rider, Mary Gichuru, and Janet Wildish. 2006. "HIV-Aids Policy Formulation in Kenya - Cdn.odi.org." *AN ANALYSIS OF HIV/AIDS POLICY FORMULATION AND IMPLEMENTATION STRUCTURES, MECHANISMS AND PROCESSES IN THE EDUCATION SECTOR IN KENYA*. https://cdn.odi.org/media/documents/3685.pdf.

Population, National Council for, Development (Kenya), Kenya. Central Bureau of Statistics, Macro International. Demographic, and Health Surveys. 1999. *Kenya Demographic and Health Survey, 1998*. National Council for Population; Development, Central Bureau of ….

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

"The Dhs Program." n.d. *The DHS Program - Data Variables and Definitions*. https://dhsprogram.com/data/data-variables-and-definitions.cfm.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2021a. *Bookdown: Authoring Books and Technical Documents with R Markdown*. https://github.com/rstudio/bookdown.

———. 2021b. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

———. 2021c. *Tinytex: Helper Functions to Install and Maintain Tex Live, and Compile Latex Documents*. https://github.com/yihui/tinytex.

Zainiddinov, Hakim, and Nazim Habibov. 2016. "Trends and Predictors of Knowledge About Hiv/Aids and Its Prevention and Transmission Methods Among Women in Tajikistan." *The European Journal of Public Health* 26 (6): 1075–9.