

UNIVERSITY OF CAPE TOWN



Addressing Data Missingness in Risky Sexual Behaviour Analysis: A Comparison of Traditional and Machine Learning Imputation Methods

Students (alphabetical order):

Akullo Angura - ANGZUK001

Rockette Ngoepe - NGPROC001

Supervisor:

A/Prof Freedom Gumedze

DEPARTMENT OF STATISTICAL SCIENCES

October 24, 2024

Acknowledgements

We would like to acknowledge and thank Associate Professor Freedom Gumedze for guiding us as our supervisor for this research project.

Plagiarism Declaration

I, Rockette Ngoepe student number NGPROC001 and I Akullo Angura student number ANGZUK001 undertake that:

1. This work, submitted for an honours degree in Statistics at the University of Cape Town, is our original work and has not previously been submitted at any institution of higher learning.
2. All sources cited or quoted are indicated and acknowledged by means of a comprehensive list of references.
3. Copyright cedes to the University of Cape Town.

Abstract

In medical research, the problem of nonresponse in surveys has a profound effect on analyses, especially if the missing values are not handled appropriately. While deleting missing values is the simplest approach, it can result in substantial loss of information. This study revisits Khati's (2012) research on factors associated with risky sexual behaviour, aiming to improve it by addressing the issue of missing data through various imputation methods.

Using an incomplete survey data on factors affecting risky sexual behaviour among women in urban Gauteng and rural Western Cape, we compare multiple imputation – a traditional approach – with three machine-learning methods: K-Nearest Neighbours, Self-Organising Maps, and Random Forest imputation. Logistic regression analyses were conducted on the incomplete dataset and each of the imputed datasets. Results shows that when the missing values are deleted, fewer significant factors are identified compared to when imputation methods are applied. Among the imputation methods, machine-learning-based imputation generally outperforms multiple imputation, with Random Forest yielding the best results.

Key findings show that for women in urban Gauteng, their first language, marital status, and attitude towards condom usage with their spouses or regular partners are significant factors influencing their risky sexual behaviour when the analysis is conducted on the incomplete dataset. Whereas, the analysis on the data imputed using multiple imputation show that age, racial group, binge drinking, and contraceptive use are all additional significant factors. When the same analysis is conducted on the machine-learning imputed datasets, in contrast to the significant factors highlighted by the analysis conducted on the incomplete data, our results show partners' employment status, lifetime contraceptive use, and engagement in risky drinking also influence significantly risky sexual behaviour.

Contents

1	Introduction	7
1.1	Study Objectives	7
1.2	Report Layout	8
2	Literature Review	9
3	Data	13
3.1	Study Areas	13
3.2	Ethics	13
3.3	Data Description	14
3.3.1	Outcome variable	14
3.3.2	Covariates	14
3.3.3	Description of factors for urban Gauteng and rural Western Cape woman.	17
3.4	Missing Data Patterns and Mechanisms	25
4	Methodology	29
4.1	Multiple Imputation	29
4.1.1	MICE Algorithm	29
4.2	Machine Learning Imputation	31
4.2.1	KNN Imputation	32
4.2.2	Self-Organising Maps Imputation	34
4.2.3	Random Forest Imputation	36
4.3	Logistic Regression Analysis	37
5	Results	39
5.1	Multivariate Analysis of Factors Associated with Risky Sexual Behaviour	40
5.1.1	Urban Gauteng	40
5.1.2	Rural Western Cape	41
5.2	Multiple Imputation multivariate analyses	43
5.2.1	Urban Gauteng	43
5.2.2	Rural Western Cape	44
5.3	Machine Learning Cross-Validation Results	46
5.4	Machine Learning Multivariate Analyses	48
5.5	Model Evaluation	54
5.5.1	Urban Gauteng Region	54
5.5.2	Rural Western Cape Region	57
5.6	Performance of Different Imputation Methods	59

6	Discussion	61
6.1	Factors Associated with Risky Sexual Behaviour	61
6.2	Comparison between Traditional and Machine-Learning Imputation Tech- niques	62
6.3	Limitations	63
7	Conclusion	65
8	References	66
A	Appendices	71
A.1	Univariate analyses of factors associated with risky sexual behaviour for the urban Gauteng and rural Western Cape regions	71
A.2	Performance of different imputation methods across factors with different proportions of missing observations	74

List of Tables

1	Description of how the covariates were re-coded.	14
2	Description of demographic variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites.	18
3	Description of Socio-economic and household hunger variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites. . . .	19
4	Description of Partner characteristics variables for women interviewed in the urban Gauteng) and rural (Western Cape) sites.	20
5	Description of psycho-social variables for women interviewed in Gauteng and Western Cape.	21
6	Description of community- and social-support variables for women in Gauteng Western Cape.	22
7	Description of substance use variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites.	23
8	Description of General health, contraceptive/pregnancy variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites	24
9	Description of sex related variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites	25
10	Missingness proportion for variables in the Gauteng dataset.	27
11	Missingness proportion for variables in the Western Cape dataset.	28
15	Multivariate logistic regression analysis of factors associated with risky sexual behaviour for women in the urban Gauteng region.	41
16	Multivariate logistic regression analysis of factors associated with risky sexual behaviour for women in the rural Western Cape region.	42
17	Multivariate logistic regression analysis of factors associated with risky sexual behaviour for women in the urban Gauteng region using the MICE imputed dataset.	44
18	Multivariate logistic regression analysis of factors associated with risky sexual behaviour for women in the rural Western Cape region using the MICE imputed dataset.	45
19	Comparison of different imputation methods and their optimal parameters, along with misclassification errors.	48
20	Multivariate Logistic Regression results for KNN, SOM, and RF Models for urban Gauteng	50
21	Multivariate Logistic Regression results for KNN, SOM, and RF Models for rural Western Cape	52
22	Model Performance Metrics for Multivariate Regression Models in the Gauteng Region.	54
23	Model Performance Metrics for Multivariate Regression Models in the Western Cape Region.	57

24	The strength of association between demographic variables and risky sexual behaviour (univariate logistic regression analysis)	71
25	The strength of association between socio-economic/household hunger and risky sexual behaviour (univariate logistic regression analysis)	71
26	The strength of association between psycho-social variables and risky sexual behaviour (univariate logistic regression analysis)	72
27	The strength of association between partner characteristics and risky sexual behaviour (univariate logistic regression analysis)	72
28	The strength of association between community/social support variables and risky sexual behaviour (univariate logistic regression analysis)	73
29	The strength of association between substance use variables and risky sexual behaviour (univariate logistic regression analysis)	73
30	The strength of association between general health, contraceptive/pregnancy variables and risky sexual behaviour (univariate logistic regression analysis)	74
31	The strength of association between sex related variables and risky sexual behaviour (univariate logistic regression analysis)	74

List of Figures

1	Missingness proportion (left panel) and missingness patterns (right panel) for the Gauteng data.	26
2	Missingness proportion (left panel) and missingness patterns (right panel) for the Western Cape data.	27
3	Summary of Multiple Imputation Procedure.	31
4	KNN training errors	46
5	SOM training errors	46
6	Random Forest Imputation training errors	47
7	Binned residual plots for Gauteng Models	56
8	Binned residual plots for Western Cape Models	58
9	Proportions of risk drinking classes predicted by imputation methods against original data proportions in the Gauteng dataset.	59
10	Proportions of employment status classes predicted by imputation methods against original data proportions in the Gauteng dataset.	60
11	Comparison of proportions for the psycho-social factor "Children are a sign of a worthy woman" classes predicted by imputation methods against original data proportions in the urban Gauteng region, which had 80% missing observations	75
12	Comparison of proportions for the current contraceptive use classes predicted by imputation methods against original data proportions in the rural Western Cape region, which had 45.6% missing observations	75

13	Comparison of proportions for the age of alcohol onset classes predicted by imputation methods against original data proportions in the urban Gauteng region, which had 59.9% missing observations	76
14	Comparison of proportions for the age of alcohol onset classes predicted by imputation methods against original data proportions in the rural Western Cape region, which had 28.2% missing observations.	76

Definitions

Binge drinking	Consumption of more than five units of alcoholic drinks per day for males and more than three units of alcoholic drinks per day for females within the previous 12 months.
Current use of contraceptive	Currently using any method, including traditional herbs/remedies or any other unproven methods to delay or avoid pregnancy.
Effective contraceptive	Any method that has been empirically shown to be effective in preventing pregnancy, such as a pill, intrauterine device, injections, diaphragm, condom, female sterilization, male sterilization, or abstinence.
Household hunger	Sometimes or often going hungry, or having no food to eat.
Lifetime use	Having ever used or tried anything such as alcohol, cigarettes, drugs, sex, contraceptives, etc.
Native language	Any of the following languages: IsiNdebele, siXhosa, IsiZulu, SeSotho, SeTswana, SePedi, SiSwati, Tshivenda, and Xitsonga.
Risky (risk) drinking	A pattern of alcohol consumption that is above the recognized sensible drinking levels, represented by the 10-item AUDIT score of above 8 [5].
Risky sexual behaviour	Having more than one sexual partner within three months, or not always using a condom with a casual or regular partner within three months, or not having used a condom during the last time one had sex.
Working	Self-employed or doing a part-time or full-time paid job.

1 Introduction

The HIV epidemic, exacerbated by poor economic development and lack of access to adequate healthcare, has since caught Southern African countries by storm. In 2023, for example, Eswatini was reported to be having the highest prevalence of HIV as 26 percent of the population was living with HIV, whereas South Africa is the third-highest prevalence country with 18 percent of the population living with HIV [53]. Among those living with HIV, more women than men are infected [9]. In South Africa’s Kwa-Zulu Natal province, for example, risky sexual behaviour among women in Umlazi Township leave them highly susceptible to risk of exposure to STIs such as HIV [19]. Findings such as these are important as they not only help decision makers understand factors exacerbating HIV rates, but also helps them identify target groups where intervention is required to help minimise and control the spread of STIs such as HIV. As such, the importance of medical research dedicated to guide policy-making in healthcare remains undisputed. Even more important is the accuracy of models used to report these results.

Model accuracy is influenced, among other things, by the data on which the model is trained. In medical research, it is a common problem that the data on which models are built is incomplete. This is often due to non-response in surveys as individuals are often reluctant to share sensitive and personal information. This problem was common in Khati’s [27] *Analysis of Factors Associated with Risky Sexual Behaviour amongst Women in Rural Western Cape and Urban Gauteng provinces*, where 16.3% of the observations were missing. According to Khati [27], risky sexual behaviour includes having multiple partners, having risky casual or unknown partners, failure to discuss the risk topic before intercourse, and failure to use effective prophylactics such as condoms. Using this data, Khati [27] concludes that women in the urban Gauteng region exhibiting risky sexual behaviour are risk drinkers whereas easy access to recreational activities and perceived importance of condom usage are associated with lower tendencies of risky sexual behaviour among these women. On the other hand, having a current partner who is working is the only significant factor contributing to risky sexual behaviour amongst rural Western Cape women [27]. A key limitation of Khati’s study is that the missing observations were deleted when conducting the analysis.

1.1 Study Objectives

In this research, we revisit Khati’s (2012) study which aimed to assess various factors associated with risky sexual behaviour among the women. Specifically, we revisit the study with the following objectives:

1. First, we will account for the missingness in the data by imputing the missing values through the widely-used imputation method, multiple imputation. For our purposes, multiple imputation is seen as a traditional method for imputation.

2. Secondly, we address the missingness problem using machine-learning imputation techniques, namely, K-nearest neighbours (KNN) imputation, Self-Organising Maps (SOM) imputation, and Random Forest (RF) imputation.
3. After imputing the missing values, we will then run logistic regression analyses on the incomplete and complete datasets. A comparison between the results from the regressions based on the multiple-imputation complete dataset and the results from the machine learning-imputed datasets is drawn. This comparison is used to compare the effectiveness of the traditional imputation method against the machine-learning imputation methods, as well as to compare between the three machine-learning imputation methods.
4. Finally, we compare our results from the complete-data analyses with initial results from the incomplete-data analysis, focusing on factors associated with risky sexual behaviour.

1.2 Report Layout

The rest of this research report is structured as followed: The next section (section 2) presents a literature review on factors associated with risky sexual behaviour. Next, in section 3, the data utilised for our research is discussed. This section commences by discussing how Khati's data was collected and used for our purposes, and puts to rest any ethical concerns around the data. We then present the data description for both the Gauteng and Western Cape regions. Evident from Khati's study is the problem of missing data, which motivates our discussion on missing data mechanisms and patterns in this section.

To address the missingness, our methodology discussion in section 4 then introduces multiple imputation and how it is done through the Multiple Imputation using Chained Equations (MICE) algorithm, as well as introduces machine learning-based imputation methods, namely, K-Nearest Neighbour (KNN) Imputation, Self-Organising Maps (SOM) Imputation, and Random Forest (RF) Imputation. The imputation procedures - both traditional and machine learning techniques - are simply pre-processing techniques, and once carried out, the analysis of factors affecting risky sexual behaviour can begin. Thus, we describe the univariate and multivariate logistic regression analysis in our methodology section.

Section 5 of this report will table the results from our analyses, and section 6 will discuss these results. Our discussions has two layers: the first layer is the pre-processing analysis, where we compare how traditional imputation techniques perform relative to machine-learning techniques in the face of real-world data. This comparison is heavily

reliant on the logistic regression results. The second layer of our discussion emanates from Khati's (2012) study, where we compare regression results in the case of the original incomplete data and when the data has been imputed. Following these discussions, section 7 will finally conclude, showing that the analysis of factors associated with risky sexual behaviour improves significantly when missingness has been appropriately accounted for.

2 Literature Review

Sexually transmitted infections (STIs), including HIV/AIDS, present a significant public health challenge in South Africa, particularly within low-income communities [3]. The transmission of these infections is accelerated by risky sexual behaviours (RSB), such as promiscuity, unprotected sex, and a lack of communication regarding one's sexual health status. Risky sexual behaviour takes on many definitions across the existing literature making it difficult to draw direct comparisons. To address this, our paper uses the occurrence of STIs and HIV as indicators of risky sexual behaviour. In this literature review, we will explore the associations between RSB and demographic, socio-economic, and cultural factors.

Age has been found to have significant associations with the occurrence of STIs. A cross-sectional survey-based study in Kazakhstan found a statistically significant association between specific age groups and the number of Human papillomavirus (HPV) infections, with the highest prevalence of high-risk HPV infections occurring among women aged 26–35 and 36–45 years [6]. This corroborates with a South African cross-sectional study that found repeat genital symptoms to be associated with the age group of 25–34 years [32]. However, a similar study conducted in Greece identified that the highest prevalence of HPV occurred amongst the youngest age group of 16–20 years [49]. Additionally, among Bangladeshi women, it was found that the percentage of women with STIs decreased with age from 10.3% (25–34 years) to 8.6% (35–49 years) [39].

Education emerges as a significant mediator of STI and HIV risks, where higher levels of education have been found to decrease the occurrence of STIs ([47], [48], [2]). Shabnam et al found that 11.7% of women with no education reported a recent contraction of STIs, compared to 6.8% of women who had at least a secondary education [47]. The association between higher education and reduced STI/HIV risks suggests that education may act as a "social vaccine", promoting safer sexual behaviours [2]. However, contradictory results were found by the Nelson Mandela/Human Sciences Research Council (NM/HSRC). The study showed that among black South Africans, 21.1% of those with a matric certificate were HIV positive, compared to 8.7% of those without any schooling [48]. These studies highlight the complexities of the relationship between education and STI risks.

Population-level studies from low-income countries have consistently shown that women with lower social status, characterised by lower levels of education and income, have an increased risk for STIs ([40],[2], [46]). The prevalence of STIs in the Couple's File data

was twice as high in the poorest group (13.4%) compared to the wealthiest group (7%) [46]. The socio-economic gradient of STI/HIV risk is not uniform across all contexts. In adjusted analyses at the individual level, being in the richest wealth quintile was associated with a 14% lower risk of STIs in Uganda. Conversely, the middle wealth quintile was associated with a higher risk of STIs compared to the poorest quintile.[2]. Moreover, the NM/HSRC study showed that among black South Africans, the chance of being infected with HIV was similar across socio-economic strata as measured by self-reported income levels [48].

Poverty, in particular, exposes women to heightened risks for STIs and HIV, partly due to its impact on one's access to healthcare and basic living conditions [40]. Research has shown that an economically vulnerable women is more likely to exchange sex for money or other favours [18]. Ostrach (2012) coins this as 'survival sex' where rural women engage in sexual activities in exchange for money to satisfy their basic needs [37].

Experiences of intimate partner violence (IPV) have seen to increase women's risks of contracting STIs ([46], [35], [7]). Shabnam found that spousal violence significantly increased women's likelihood of experiencing STIs, with those enduring both physical and sexual assault by their husbands being 4 times more likely to contract STIs [46]. This trend was also seen among Bangladeshi women [39]. Furthermore, women in a Ugandan cross-sectional study who had experienced sexual violence were two times more likely to contract and STIs compared to women who experienced no violence[35]. These studies highlight the significant influence that relationship violence has on the occurrence of STIs.

The power dynamics within sexual relationships significantly influence condom usage and, by extension, the risk of STIs. Literature suggests that women in relationships characterized by power imbalances are less likely to succeed in negotiating protection [18]. This relative powerlessness in negotiating condom use is evidenced by studies showing that condom use is significantly higher in male HIV-negative, female HIV-positive couples compared to male HIV-positive, female HIV-negative couples [7]. Additionally, women with older partners often face greater challenges in negotiating condom use compared to those with same-age partners [30]. This difficulty can be attributed to the traditional gender roles and expectations that frequently arise in relationships with significant age disparities. Furthermore, women with no control over their own earnings are less likely to initiate HIV-risk negotiations [4]. This lack of financial autonomy can further diminish a woman's ability to negotiate safer sex practices.

On the other hand, research has shown that women with high levels of relationship power are more likely to report consistent condom use compared to women with low levels of power [38]. Sexually empowered women, defined as those who have the agency to make decisions about their sexual health, are at a lower risk of contracting STIs. Also, women who participate in decision-making concerning their own health, either individually or jointly with their partners, have lower odds of reporting STIs [35]. These findings suggest

that female empowerment and autonomy act as protective factors against STIs.

Cultural norms surrounding gender roles and sexual behaviour significantly influence women’s vulnerabilities to STIs. The cultural construct of a ‘good’ woman being ignorant about sex, passive in sexual interactions, and uncomfortable discussing sexual matters increases her vulnerability to STIs, including HIV/AIDS [24]. The traditional norm of virginity for unmarried girls that exists in many societies, paradoxically, increases young women’s risk of infection because it restricts their ability to ask for information about sex out of fear that they will be thought to be sexually active [59]. In largely religious and patriarchal societies, women are often unable to initiate safer-sex negotiations. In Bangladesh, despite women’s awareness to STIs and their risks, they cannot discuss contraception with their partners [24].

Multiple sexual partners are a well-documented risk factor for STIs. This, coupled with unsafe sexual practices, such as irregular condom usage, put women at an increased risk of contracting STIs. Fethers et al revealed that bacterial vaginosis (BV) is significantly associated with sexual contact with new and multiple partners [12]. Similarly, a study done using nationally representative survey data from Uganda found that having multiple lifetime sexual partners was associated with a higher risk of contracting STIs [2]. A Ugandan cross-sectional study found complementary results [35]. Women with multiple partners were over two times more likely to contract STIs.

Shirin et al (2009) found that inconsistent condom usage had a positive association with STIs [47]. However, Shabnam (2017) found there to be a higher prevalence of STIs among women who use condoms or other contraceptive methods, compared to those who do not [46]. This counter intuitive finding may be attributed to inconsistent use, or related to the frequency of sexual activities..

According to many studies, alcohol usage and risky sexual behaviour are strongly related. For instance, a US national survey of over 17,000 American college students revealed that those engaging in heavy episodic (HE ¹) drinking were almost three times more likely to have multiple sexual partners within a month compared to their non-HE drinking counterparts [58]. This association is further supported by findings that the frequency of alcohol consumption before intercourse is directly linked to the frequency of unprotected sex [36]. Graves (1995) corroborated these findings, noting that increased HE drinking among young adults was associated with decreased condom usage [14]. These observations are nuanced by the fact that the occurrence of alcohol use and risky sexual behaviour depend on the frequency of intercourse [10].

In South Africa, qualitative research among adults aged 25 to 44 in Gauteng uncovered a strong correlation between alcohol consumption and risky sexual practices [34]. Partici-

¹Heavy episodic (HE) drinkers are defined as having five or more drinks on a single occasion during a specified period.

pants reported high levels of alcohol use and unprotected sex with the latter occurring mainly among casual sexual partners. Alcohol consumption was perceived to enhance sexual arousal and desire, especially in the context of casual sexual encounters, while potentially diminishing arousal in the context of marital or regular partnerships [34]. This research also highlighted a common South African trend of weekend risky drinking and transactional sex, where younger women frequent drinking venues to engage in sexual activities with older men, colloquially referred to as "sugar daddies", in exchange for material benefits [34].

A prospective cohort study, utilizing self-reported daily diaries, tracked 58 HIV-positive women and 24 HIV-positive men over 42 days. The study found that participants consumed an average of 6.13 drinks per session and reported 4,297 sex events, with 80.17% of them occurring without condom use [28]. Despite acknowledged limitations, the study suggests that moderate to high-risk drinking before sex significantly increases the likelihood and rate of unprotected sex among HIV-positive individuals [28].

Evidently, existing literature suggests many factors have associations with STIs/HIV, and, by extension, risky sexual behaviour. These include demographic, socio-economic, cultural, and alcohol abuse factors.

Our study looks to find similar factors that have associations with risky sexual behaviour for women in the rural Western Cape and urban Gauteng regions. We also aim to address the issue of inadequate handling of missing values when identifying the associated factors. We look to demonstrate how properly accounting for missingness enhances the understanding of these factors.

3 Data

This project utilises the same anonymised dataset used in Khati's (2012) study. The original data was previously collected through a multi-institutional collaboration, including the University of Cape Town (UCT), the University of Pretoria (UP), and the Medical Research Council (MRC). The data was initially intended for a study titled *Comprehensive foetal alcohol syndrome prevention programme in Western Cape and Gauteng Provinces*, conducted in 2006.

The study population comprised of women of child-bearing age, between 18 and 44 years. In the Gauteng region, a cluster sampling approach was used to collect data from 820 women. In the rural Western Cape site, spanning three municipal areas, a stratified cluster random sampling approach was deployed to reach 650 women. A structured questionnaire was used to obtain information on the independent and dependent variables. Overall, 606 responses were recorded for the Gauteng site and 412 responses from the Western Cape site. This was a response rate of 73,9% and 63,4% respectively.

3.1 Study Areas

The study areas were comprised of two sites: a densely populated urban area in the Gauteng Province, and a sparsely populated rural area in the Western Cape. The Gauteng site is located in a highly industrialised region within the City of Tshwane Metropolitan Municipality over an area of 2 199 km². This region had a population of 1.98 million at the time of the data collection [1]. The black population made up the majority for this site (78.3%). The Western Cape site is set between the Atlantic Ocean on the west and agricultural land on the east spanning over 15 311 km² of land. With a cumulative population of around 160 000 people, it spans over the Cederberg, Bergrivier, and Swartland municipalities. The Coloured population make up the majority and are primarily Afrikaans speaking.

3.2 Ethics

The previous data collection was approved by the Faculty of Health Sciences Research Ethics Committees (REC) of the Universities of Pretoria (121/2005) and Cape Town (381/2005; renewal number 001/2007). The data was originally collected via interviews conducted in the participants' households by trained fieldworkers. The participants provided informed consent and signed informed consent forms before the interviews. The data contains anonymised responses with no personal identifiers present. The data is used solely for the purpose of addressing our research question and will not be used in any way that could potentially compromise the privacy or well-being of the previous participants.

3.3 Data Description

3.3.1 Outcome variable

The outcome variable for this study is risky sexual behaviour (RSB). A participant was seen to exhibit RSB if they met at least one of the three criteria:

- (a) The participant had more than one sexual partner three months prior to the interview.
- (b) The participant did not always use a condom with a casual or regular partner within three months prior to the interview.
- (c) The participant did not use a condom the last time they had sex.

Participants who were classified to exhibit risky sexual behaviour were coded '1', and those who did not were coded '0'. Note that this criteria for risky sexual behaviour is not identical to the one used by Khati (2012), and hence different proportions of risky sexual behaviour may result in this research.

3.3.2 Covariates

The covariates in this data are all categorical, with the majority of them being binary responses. This study considered 39 covariates for both urban Gauteng and rural Western Cape regions. These covariates fall within the following eight domains: demographic factors, socio-economic and household hunger, psycho-social factors, current partner characteristics, community and social support factors, substance and alcohol use factors, general health and reproductive health factors, and sexual behaviour related factors including contraception usage. These covariates were factorised and re-coded as shown in Table 1

Table 1: Description of how the covariates were re-coded.

Variable	Code: level	Variable	Code: level
Demographic factors			
Age	0: 18-24 years 1: 25-34 years 2: 35-44 years	Education	0: Primary or below 1: Above primary
Marital status	0: Legally married 1: Traditionally married 2: Cohabiting 3: Never married 4: Divorced or Separated or Widow	First language	0: Native Language 1: Afrikaans 2: English 3: Other

Variable	Code: level	Variable	Code: level
Racial classification	0: Black/African 1: Coloured 2: White		
Socio-economic factors and household hunger			
Paid Work in Last 12 Months	0: No 1: Yes	Currently working	0: No 1: Yes
High SES	0: No 1: Yes	Household Hunger	0: No 1: Yes
Employment status	0: Unemployed 1: Employed		
Partner characteristics variables			
Current partner age	0: <30 years 1: \geq 30 years	Current Partner Employed	0: No 1: Yes
Current partner satisfaction	0: Disagree 1: Agree	Serious arguments	0: Disagree 1: Agree
Psycho-social variables			
Male fertility entitlement	0: Disagree 1: Agree	Childless choice is right	0: Not Wrong 1: Wrong
Children are a sign of worthy Woman	0: Untrue 1: True	Children are a sign of worthy Man	0: Untrue 1: True
Community and social Support variables			
Available recreation	0: Disagree 1: Agree	Easy to use recreational facilities	0: Disagree 1: Agree
Easy to buy alcohol in the community	0: Disagree 1: Agree	Significant heavy drinking in the community	0: Disagree 1: Agree

Variable	Code: level	Variable	Code: level
Community accepts the abuse of alcohol	0: Disagree 1: Agree	Helpful neighbours	0: Disagree 1: Agree
Substance use variables			
Lifetime alcohol Usage	0: No 1: Yes	Current alcohol asage	0: No 1: Yes
Binge drinking	0: No 1: Yes	Risk drinking	0: No 1: Yes
Lifetime cigarette Usage	0: No 1: Yes	Age of alcohol onset	0: ≤ 18 years 1: > 18 years
General health, contraceptive and pregnancy variables			
Lifetime con- traceptive Usage	0: No 1: Yes	Current con- traceptive Usage	0: No 1: Yes
Effective contracep- tive usage	0: No 1: Yes	Children	0: No 1: Yes
Sex related variables			
Importance of condom usage with regular partner or spouse	0: Unimportant 1: Important	Importance of condom usage with a casual partner	0: Unimportant 1: Important
Current sexual partner	0: Husband/boyfriend is not the most recent sexual partner 1: Husband/boyfriend is the most recent sexual partner		

3.3.3 Description of factors for urban Gauteng and rural Western Cape woman.

3.3.3.1 Demographic factors

The 25-34 age group is the largest in both regions, with Gauteng having 38% and the Western Cape having 30.1% (Table 2). There are no women in Gauteng who have an education level below primary while 59.2% of women from the Western Cape have only received education up to a primary level or lower. The marital landscape between the two regions differs vastly. Women in Gauteng are all married, either legally or traditionally. In contrast, 34.7% of women in the Western Cape are Cohabiting but are unmarried. Afrikaans is the predominant language in the Western Cape (93%) whilst the native languages (78.2%) are mainly spoken in Gauteng. Similarly, black women (81%) make up the largest racial group in Gauteng whilst Coloured women (90.5%) are the prominent racial group in the Western Cape.

Table 2: Description of demographic variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites.

Category	Total N=1018	Gauteng N=606	Western Cape N=412
Age			
18-24	284 (27.9%)	182 (30.0%)	102 (24.8%)
25-34	391 (38.4%)	230 (38.0%)	161 (39.1%)
35-44	343 (33.7%)	194 (32.0%)	149 (36.2%)
Education			
Above primary	167 (16.4%)	0 (0%)	167 (40.5%)
Primary or lower	848 (83.3%)	604 (99.7%)	244 (59.2%)
Missing	3 (0.3%)	2 (0.3%)	1 (0.2%)
Marital status			
Cohabiting	143 (14.0%)	0 (0%)	143 (34.7%)
Divorced/separated/widow	17 (1.7%)	0 (0%)	17 (4.1%)
Legally married	344 (33.8%)	221 (36.5%)	123 (29.9%)
Never married	125 (12.3%)	0 (0%)	125 (30.3%)
Traditionally married	388 (38.1%)	384 (63.4%)	4 (1.0%)
Missing	1 (0.1%)	1 (0.2%)	0 (0%)
Language			
Afrikaans	479 (47.1%)	96 (15.8%)	383 (93.0%)
English	22 (2.2%)	20 (3.3%)	2 (0.5%)
Native Language ^a	501 (49.2%)	474 (78.2%)	27 (6.6%)
Other	13 (1.3%)	13 (2.1%)	0 (0%)
Missing	3 (0.3%)	3 (0.5%)	0 (0%)
Race			
Black/African	526 (51.7%)	491 (81.0%)	35 (8.5%)
Coloured	442 (43.4%)	68 (11.2%)	374 (90.8%)
White	48 (4.7%)	45 (7.4%)	3 (0.7%)
Missing	2 (0.2%)	2 (0.3%)	0 (0%)

^a Native language includes any of the following: IsiNdebele, siXhosa, IsiZulu, SeSotho, SeTswana, SePedi, SiSwati, Tshivenda, and Xitsonga.

3.3.3.2 Socio-economic and household hunger variables

As evident in Table 3, 82% of women in the Western Cape do not do paid work but only 19.9% are unemployed. 42.1% of employment status information is missing for the Gauteng region. 35.3% of Gauteng women are working compared to 57.3% of women in the Western Cape. 65.2% of women in Gauteng are considered to high a high socio-economic status while 35.7% of women have this same status in the Western Cape. There is no household hunger in Gauteng whilst 12.1% of Western Cape women experience household hunger.

Table 3: Description of Socio-economic and household hunger variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites.

Category	Total N=1018	Gauteng N=606	Western Cape N=412
Paid work			
No	396 (38.9%)	322 (53.1%)	74 (18.0%)
Yes	622 (61.1%)	284 (46.9%)	338 (82.0%)
Employment status			
Employed	330 (32.4%)	0 (0%)	330 (80.1%)
Unemployed	433 (42.5%)	351 (57.9%)	82 (19.9%)
Missing	255 (25.0%)	255 (42.1%)	0 (0%)
Working^a			
Not Working	430 (42.2%)	345 (56.9%)	85 (20.6%)
Working	450 (44.2%)	214 (35.3%)	236 (57.3%)
Missing	138 (13.6%)	47 (7.8%)	91 (22.1%)
High socio-economic status^b			
No	463 (45.5%)	200 (33.0%)	263 (63.8%)
Yes	542 (53.2%)	395 (65.2%)	147 (35.7%)
Missing	13 (1.3%)	11 (1.8%)	2 (0.5%)
Household hunger^c			
No	965 (94.8%)	604 (99.7%)	361 (87.6%)
Yes	50 (4.9%)	0 (0%)	50 (12.1%)
Missing	3 (0.3%)	2 (0.3%)	1 (0.2%)

^a Working means self-employed or doing part-time or full-time paid job.

^b High socio-economic status (SES) was assigned to those who possessed five or more out of eight specified assets or amenities (e.g. electricity, TV set, radio, landline, cellular phone, computer, fridge, washing machine) and could pay for household essentials. Low SES was assigned to those with fewer than five assets and who sometimes or never could pay for household essentials.

^c Household hunger is defined as sometimes or often going hungry or having no food to eat.

3.3.3.3 Partner characteristics variables

Of the 606 respondents in Gauteng, 431 women responded to having a current partner. Of these respondents, 51% of their current partners were over the age of 30 years old as depicted table 4 below. There was no data on this covariate for the Western Cape region. However, 76.2% of women from the Western Cape reported having an employed partner. 68.6% of women in Gauteng reported being satisfied with their current partner, compared to only 3.9% responding to the same question in the Western Cape.

Table 4: Description of Partner characteristics variables for women interviewed in the urban Gauteng) and rural (Western Cape) sites.

Category	Total N=1018	Gauteng N=606	Western Cape N=412
Current partner Age			
<30 years	122 (12.0%)	122 (20.1%)	0 (0%)
≥30 years	309 (30.4%)	309 (51.0%)	0 (0%)
Missing	587 (57.7%)	175 (28.9%)	412 (100%)
Current partner Employed			
No	114 (11.2%)	100 (16.5%)	14 (3.4%)
Yes	674 (66.2%)	360 (59.4%)	314 (76.2%)
Missing	230 (22.6%)	146 (24.1%)	84 (20.4%)
Current partner satisfaction			
Agree	353 (34.7%)	41 (6.8%)	312 (75.7%)
Disagree	432 (42.4%)	416 (68.6%)	16 (3.9%)
Missing	233 (22.9%)	149 (24.6%)	84 (20.4%)
Serious arguments			
Agree	447 (43.9%)	245 (40.4%)	202 (49.0%)
Disagree	330 (32.4%)	205 (33.8%)	125 (30.3%)
Missing	241 (23.7%)	156 (25.7%)	85 (20.6%)

3.3.3.4 Psycho-Social variables

Depicted in table 5 below are statistics of responses on psycho-social variables. Of the 606 and 412 women in Gauteng and Western Cape respectively, 38.4% and 25.2% in the respective provinces agree that according to their culture, males are entitled to have as many children as they want (male fertility entitlement). However, for both regions, over 50% of respondents did not answer this question. More women in the Western Cape (74%) believe that it is wrong to choose not to have a child compared to Gauteng women (58.7%). Overall, 79% of women did not respond to whether children are a sign of a worthy woman. Similarly, 77.7% did not respond to whether children are a sign of a worthy man.

Table 5: Description of psycho-social variables for women interviewed in Gauteng and Western Cape.

Category	Total N=1018	Gauteng N=606	Western Cape N=412
Male fertility entitlement			
Agree	337 (33.1%)	233 (38.4%)	104 (25.2%)
Disagree	146 (14.3%)	61 (10.1%)	85 (20.6%)
Missing	535 (52.6%)	312 (51.5%)	223 (54.1%)
Childless choice is right			
Not Wrong	354 (34.8%)	247 (40.8%)	107 (26.0%)
Wrong	661 (64.9%)	356 (58.7%)	305 (74.0%)
Missing	3 (0.3%)	3 (0.5%)	0 (0%)
Children are a sign of worthy woman			
True	130 (12.8%)	69 (11.4%)	61 (14.8%)
Untrue	84 (8.3%)	52 (8.6%)	32 (7.8%)
Missing	804 (79.0%)	485 (80.0%)	319 (77.4%)
Children are a sign of worthy man			
True	142 (13.9%)	76 (12.5%)	66 (16.0%)
Untrue	85 (8.3%)	53 (8.7%)	32 (7.8%)
Missing	791 (77.7%)	477 (78.7%)	314 (76.2%)

3.3.3.5 Community and social support variables

In the Western Cape, 81.3% of women feel that they do not have access to recreational facilities, as shown in table 6. Whereas, 67% of women in Gauteng share the same sentiment. When it comes to alcohol accessibility, 82.2% of women in the Gauteng region agree that it is easy to buy alcohol in the community compared to only 31.3% in the Western Cape. However, both regions agree on the significant heavy drinking in the community. Overall, 67.7% of women would agree to having helpful neighbours.

Table 6: Description of community- and social-support variables for women in Gauteng Western Cape.

Category	Total N=1018	Gauteng N=606	Western Cape N=412
Available recreation			
Agree	279 (27.4%)	202 (33.3%)	77 (18.7%)
Disagree	739 (72.6%)	404 (66.7%)	335 (81.3%)
Easy to use recreational facilities			
Agree	276 (27.1%)	194 (32.0%)	82 (19.9%)
Disagree	741 (72.8%)	411 (67.8%)	330 (80.1%)
Missing	1 (0.1%)	1 (0.2%)	0 (0%)
Easy to buy alcohol in the community			
Agree	627 (61.6%)	498 (82.2%)	129 (31.3%)
Disagree	387 (38.0%)	107 (17.7%)	280 (68.0%)
Missing	4 (0.4%)	1 (0.2%)	3 (0.7%)
Significant heavy drinking in the community			
Agree	813 (79.9%)	499 (82.3%)	314 (76.2%)
Disagree	197 (19.4%)	104 (17.2%)	93 (22.6%)
Missing	8 (0.8%)	3 (0.5%)	5 (1.2%)
Community accepts abuse of alcohol			
Agree	547 (53.7%)	347 (57.3%)	200 (48.5%)
Disagree	467 (45.9%)	255 (42.1%)	212 (51.5%)
Missing	4 (0.4%)	4 (0.7%)	0 (0%)
Helpful neighbours			
Agree	689 (67.7%)	367 (60.6%)	322 (78.2%)
Disagree	325 (31.9%)	237 (39.1%)	88 (21.4%)
Missing	4 (0.4%)	2 (0.3%)	2 (0.5%)

3.3.3.6 Substance use variables

In Gauteng, 40.3% of the women reported lifetime alcohol use compared to 72.3% in the Western Cape, as depicted in table 7. About 25.9% of the women in urban Gauteng and 45.6% in the rural Western Cape were current users of alcohol. About 14.4% and 7.6% women in Gauteng were binge and risky drinkers. This is compared to 37.4% and 32% in the rural Western Cape. However, over 71.9% and 54.9% of responses were missing for urban Gauteng and rural Western Cape, respectively.

Only 11.5% women in urban Gauteng started drinking alcohol before the age of 18

compared to 30.3% in rural Western Cape. However, over 47.1% of total responses for this variable were missing.

Table 7: Description of substance use variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites.

Category	Total N=1018	Gauteng N=606	Western Cape N=412
Lifetime alcohol usage ^a			
No	475 (46.7%)	361 (59.6%)	114 (27.7%)
Yes	542 (53.2%)	244 (40.3%)	298 (72.3%)
Missing	1 (0.1%)	1 (0.2%)	0 (0%)
Current alcohol usage ^b			
No	456 (44.8%)	449 (74.1%)	7 (1.7%)
Yes	345 (33.9%)	157 (25.9%)	188 (45.6%)
Missing	217 (21.3%)	0 (0%)	217 (52.7%)
Binge drinking ^c			
No	560 (55.0%)	519 (85.6%)	41 (10.0%)
Yes	241 (23.7%)	87 (14.4%)	154 (37.4%)
Missing	217 (21.3%)	0 (0%)	217 (52.7%)
Risk drinking ^d			
No	178 (17.5%)	124 (20.5%)	54 (13.1%)
Yes	178 (17.5%)	46 (7.6%)	132 (32.0%)
Missing	662 (65.0%)	436 (71.9%)	226 (54.9%)
Lifetime cigarette use			
No	613 (60.2%)	488 (80.5%)	125 (30.3%)
Yes	402 (39.5%)	115 (19.0%)	287 (69.7%)
Age of alcohol onset			
< 18 years	195 (19.2%)	70 (11.6%)	125 (30.3%)
≥ 18 years	344 (33.8%)	173 (28.5%)	171 (41.5%)
Missing	479 (47.1%)	363 (59.9%)	116 (28.2%)

^a Lifetime use means having ever had a drink containing alcohol, smoked a cigarette, or used other drugs, respectively.

^b Currently drinking at least one drink on a typical day.

^c Consumption of over 3 alcoholic beverages on a typical day for females.

^d AUDIT score of over 8.[5]

3.3.3.7 General health, contraceptive/pregnancy variables

Looking at both regions, 55.7% of women have used a form of contraception to delay or

avoid getting pregnant (Table 8). Current contraceptive usage sits at 50.2% for urban Gauteng and 44.9% for rural Western Cape. However, 33.3% and 45.6% of responses are missing in Gauteng and Western Cape, respectively. The effective contraceptive usage category tells a similar story, including the proportions of missing observations. In urban Gauteng 48.5% of women have children whilst 58% of women have children in the rural Western Cape site. Overall, this category has a missingness of 28.7%.

Table 8: Description of General health, contraceptive/pregnancy variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites

Category	Total N=1018	Gauteng N=606	Western Cape N=412
Lifetime Contraceptive Usage ^a			
No	432 (42.4%)	244 (40.3%)	188 (45.6%)
Yes	567 (55.7%)	354 (58.4%)	213 (51.7%)
Missing	19 (1.9%)	8 (1.3%)	11 (2.7%)
Current Contraceptive Usage			
No	139 (13.7%)	100 (16.5%)	39 (9.5%)
Yes	489 (48.0%)	304 (50.2%)	185 (44.9%)
Missing	390 (38.3%)	202 (33.3%)	188 (45.6%)
Effective Contraceptive Usage ^b			
No	59 (5.8%)	33 (5.4%)	26 (6.3%)
Yes	502 (49.3%)	312 (51.5%)	190 (46.1%)
Missing	457 (44.9%)	261 (43.1%)	196 (47.6%)
Child(ren)			
No	193 (19.0%)	143 (23.6%)	50 (12.1%)
Yes	533 (52.4%)	294 (48.5%)	239 (58.0%)
Missing	292 (28.7%)	169 (27.9%)	123 (29.9%)

^a Lifetime contraceptive use means having ever used anything in any way to delay or avoid getting pregnant.

^b Effective contraceptive means using any method, which has been empirically shown to be effective in preventing pregnancy such as a pill, intrauterine device (IUD), injections, diaphragm, condom, female sterilisation, male sterilization or abstinence

3.3.3.8 Sex related variables

The importance of condom usage with a regular partner differs greatly between the two regions. Table 9 below shows that in urban Gauteng, 71% of women deem it important to use a condom with a regular partner, whilst 33.3% of women in rural Western Cape

share the same sentiment. In contrast, 89.1% of women in rural Western Cape deem it important to use a condom with a casual partner. 91.9% of women in urban Gauteng share the same sentiment. For Gauteng women, 83.3% of their current sexual partners are their husbands or boyfriends. Similarly, 85.7% of the Western Cape women’s current sexual partners are their husbands or boyfriends.

Table 9: Description of sex related variables for women interviewed in the urban (Gauteng) and rural (Western Cape) sites

Category	Total N=1018	Gauteng N=606	Western Cape N=412
Risky sexual behaviour			
Yes	265 (26%)	240 (39.6%)	25 (6.1%)
No	753 (74.0%)	366 (60.4%)	387 (93.9%)
Importance of condom usage:			
Regular Partner			
Important	567 (55.7%)	430 (71.0%)	137 (33.3%)
Unimportant	415 (40.8%)	154 (25.4%)	261 (63.3%)
Missing	36 (3.5%)	22 (3.6%)	14 (3.4%)
Importance of condom usage:			
Casual partner			
Important	924 (90.8%)	557 (91.9%)	367 (89.1%)
Unimportant	62 (6.1%)	32 (5.3%)	30 (7.3%)
Missing	32 (3.1%)	17 (2.8%)	15 (3.6%)
Current sexual partner			
Husband/Boyfriend	858 (84.3%)	505 (83.3%)	353 (85.7%)
most recent sex partner			
Other most recent sex	59 (5.8%)	27 (4.5%)	32 (7.8%)
partner			
Missing	101 (9.9%)	74 (12.2%)	27 (6.6%)

3.4 Missing Data Patterns and Mechanisms

As evidenced by the data description above, the data used for our analysis has 16.3% missing observations in the combined dataset: 15.3% from the Gauteng region and 17.7% from the Western Cape region. The nature of the missing data can be described through missing data patterns and mechanisms. A missing data pattern refers to the arrangement in which data is missing in the dataset [31]. It can be monotone or non-monotone: a

monotone missingness pattern is a type of missingness where if a value is missing for an observation in one variable, then all subsequent variables have missing values for that observation [31]. Whereas, a non-monotone missingness pattern is one where the arrangement in which the data is missing is not predictable; that is, there is no discernible pattern in the missing data [31].

Figure 1 below shows the missingness pattern present among several variables in the Gauteng data (right panel) alongside proportions of the missing observations in these variables (left panel). The accompanying Table 10 shows the variables in question and their respective percentages of missing observations. The figure shows that different variables have different missingness proportions, depending possibly on how invasive a questions might have been to respondents. For example, risk drinking (variable m) has the highest missing proportion among Gauteng respondents, probably due to respondents being reluctant to admit to irresponsible drinking behaviour. The missingness map on the right panel of figure 1 shows that there's no obvious pattern followed by the missing values in the Gauteng data, thus depicting a non-monotone missingness pattern present in the data.

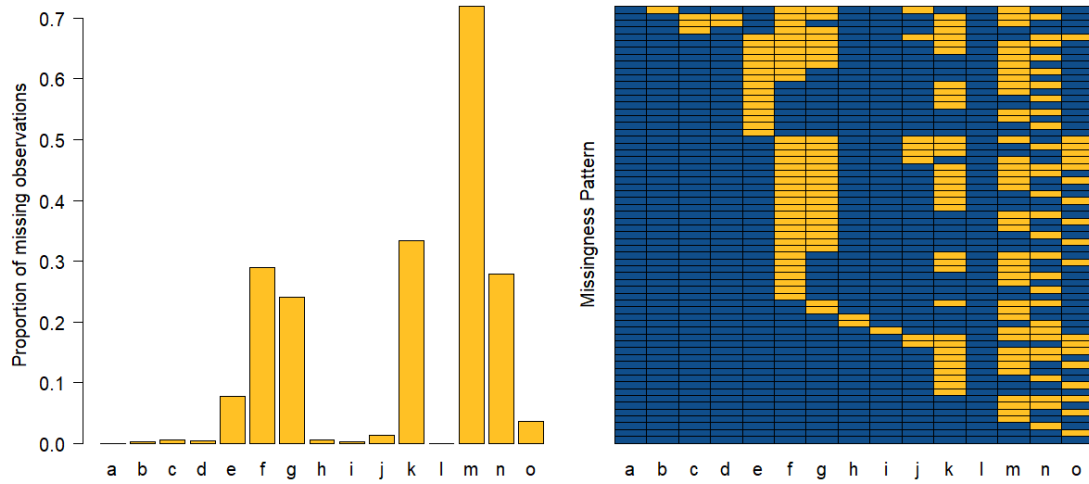


Figure 1: Missingness proportion (left panel) and missingness patterns (right panel) for the Gauteng data.

Table 10: Missingness proportion for variables in the Gauteng dataset.

Variable	%	Variable	%
a: Age	0.0	i: Easy access to alcohol	0.2
b: Marital status	0.2	j: Lifetime contraceptive use	1.3
c: Language	0.5	k: Current contraceptive use	33.3
d: Race	0.3	l: Binge drinking	0.0
e: Working	7.8	m: Risk drinking	71.9
f: Current partner Age	28.9	n: Children	27.9
g: Current partner Employed	24.1	o: Importance of condom usage with spouse/regular partner	3.6
h: Children choice is right	0.50		

Similar to Figure 1, Figure 2 below shows the missingness proportions of variables in the Western Cape dataset (left panel) and the missingness pattern of this data (right panel). Accompanying the figure is Table 11 which shows the variable names and their missingness percentages. From the figure and the table, is it clear that the *Current contraceptive use* variable (variable k) has the highest missingness proportion among all variables in the Western Cape data. Similar to the Gauteng data, the Western Cape data exhibits a non-monotone missingness pattern as the pattern is not obvious.

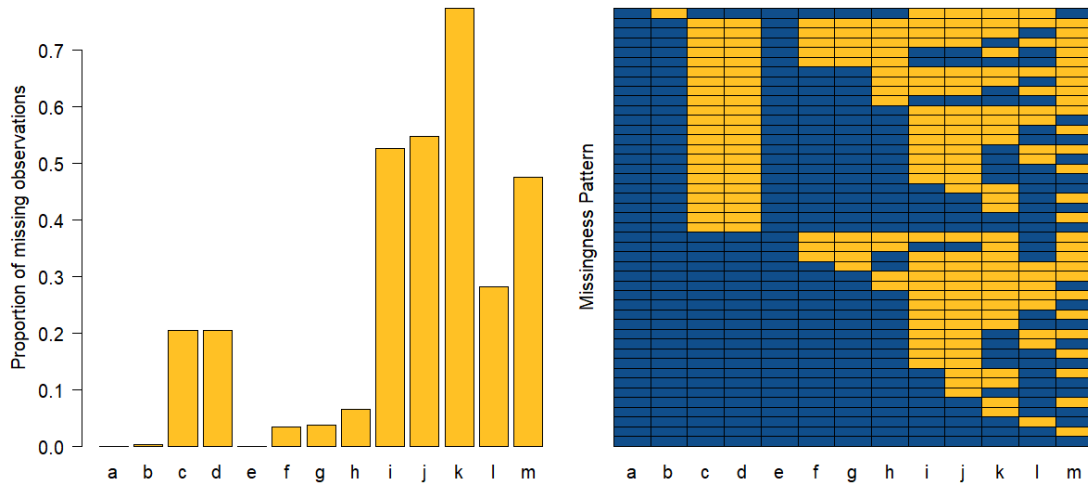


Figure 2: Missingness proportion (left panel) and missingness patterns (right panel) for the Western Cape data.

Table 11: Missingness proportion for variables in the Western Cape dataset.


Variable	%	Variable	%
a: Language	0.0	h: Current sex partner	6.6
b: Household hunger	0.2	i: Binge drinking	52.7
c: Satisfied with current partner	20.4	j: Risk drinking	54.9
d: Current partner employed	20.4	k: Children are a sign of a worthy woman	77.4
e: Lifetime cigarette usage	69.7	l: Age of alcohol onset	28.2
f: Importance of condom usage with spouse/regular partner	28.88	m: Effective contraceptive use	47.6
g: Importance of condom usage with casual partner	24.09		

The underlying process generating this missingness can be explained through missingness mechanisms [42]. Missing-data mechanisms can be classified into three groups, namely: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [31]. Missing data follows an MCAR mechanism if the probability that subjects provide data on a particular variable does not depend on the observed or unobserved variables for that subject. That is, for the indicator matrix \mathbf{R} indicating whether values are observed or missing and given the matrix \mathbf{Y} of complete data as well as unknown parameters θ , $\Pr[\mathbf{R}|\mathbf{Y}, \theta] = \Pr[\mathbf{R}|\theta] \forall \mathbf{Y}, \theta$ [33].

Meanwhile, missing data follows a MAR mechanism if the probability of a variable missing, after accounting for the observed data, does not depend on the unobserved data. That is, if \mathbf{Y}_{obs} represents the observed components of \mathbf{Y} and \mathbf{Y}_{mis} denotes the missing components of \mathbf{Y} , then $\Pr[\mathbf{R}|\mathbf{Y}, \theta] = \Pr[\mathbf{R}|\mathbf{Y}_{obs}, \theta] \forall \mathbf{Y}_{mis}, \theta$ [33].

Finally, a missing-data mechanism is MNAR if the probability of a missing variable depends only on the missing values in \mathbf{Y} , i.e., $\Pr[\mathbf{R}|\mathbf{Y}, \theta] = \Pr[\mathbf{R}|\mathbf{Y}_{mis}, \theta] \forall \mathbf{Y}_{obs}, \theta$ [33]. For our research, we assume that the data is missing at random (MAR). This allows us to use the multiple imputation and the Machine-Learning imputation techniques to address this missingness problem before running our analysis.

4 Methodology

In this section, the methodology followed in conducting our analysis is explained. First, we discuss the traditional multiple imputation technique, followed by the machine-learning KNN, SOM, and RF-imputation techniques. After discussing these pre-processing techniques, the section provides a brief overview of the logistic regression model used to analyse factors associated with risky sexual behaviour, and subsequently concludes by discussing performance metrics used for model comparison. We implement our methods using version 4.3.2 of the  programming language.

4.1 Multiple Imputation

Imputation methods for handling missing data include simple imputation techniques, such as mean or mode imputation, and multiple imputation [56]. While simple imputation is straightforward, it fails to account for the uncertainty inherent in the imputed values [43]. This limitation can be overcome through Rubin's (1986) multiple imputation method. Multiple Imputation replaces missing values with multiple plausible estimates, capturing the uncertainty associated with the imputation [43]. That is, for t iterations, a vector of t imputed values results in t complete datasets, each providing a different estimation for the same missing value. These datasets undergo statistical analysis separately, and the results are combined using Rubin's [44] pooling rules.

Multiple imputation can be carried out using software packages such as MICE or Amelia in R. MICE (Multiple Imputation by Chained Equations) employs chained equations for imputation [54], while Amelia uses an expectation-maximisation with bootstrapping (EMB) algorithm [20]. Both methods assume that the data are missing at random (MAR), but Amelia additionally requires that the data follow a multivariate normal distribution. For this research, we prefer MICE over Amelia because our data do not meet Amelia's distributional requirements [45].

4.1.1 MICE Algorithm

The MICE procedure consists of two main steps: creating M complete datasets using Gibbs sampling and applying statistical analyses to these datasets, followed by pooling the results with Rubin's rules.

Step 1: Generating M complete datasets (Imputation)

The MICE technique starts with an initial simple imputation, such as mean imputation, for each missing value to create a complete dataset. The procedure is then carried out as follows [54]:

1. **Initialisation** ($t = 0$): Perform a simple imputation for all missing values by

imputing missing values to be the mean of the variable, $\bar{\mathbf{X}}$:

$$\mathbf{X}_0 \sim N(\bar{\mathbf{X}}, \sigma^2).$$

2. Conditional Imputation for Each Variable (\mathbf{X}_1^*):

- (a) For each variable with missing data, set the missing values back to their original state.
- (b) Estimate the missing values using a linear regression approach:

$$\theta_1^{*(t)} \sim P\left(\theta_1 \mid \mathbf{X}_1^{\text{obs}}, \mathbf{X}_2^{(t-1)}, \dots, \mathbf{X}_p^{(t-1)}\right)$$

- (c) Impute new values for \mathbf{X}_1^* based on the conditional distribution:

$$\mathbf{X}_1^{*(t)} \sim P\left(\mathbf{X}_1 \mid \mathbf{X}_1^{\text{obs}}, \mathbf{X}_2^{(t-1)}, \dots, \mathbf{X}_p^{(t-1)}, \theta_1^{*(t)}\right)$$

- 3. Iterate Through All Variables (p):** Repeat step 2 for each variable with missing data to create a complete dataset.
- 4. Repeat the Procedure for t Iterations:** Use the imputed dataset from the first iteration as the starting point for subsequent iterations.
- 5. Stop:** After the maximum number of iterations $t = M$ is reached, stop.

The result is M imputed datasets resulting from the M iterations.

Step 2: Analysis and Pooling

After generating the M complete datasets, statistical analysis is performed on each dataset independently. The results are then pooled using Rubin's rules to produce a unified estimate [44]. Suppose that the the estimated parameters are θ_i for $i = 1, 2, \dots, M$ datasets, then the final (pooled) estimate is given by:

$$\bar{\theta} = \frac{1}{M} \left(\sum_{i=1}^M \theta_i \right).$$

The variability of the estimate on the other hand consists of the within-imputation variance and between-imputation variance. The within-imputation variance is the mean of the squared standard error of the estimates in each completed dataset, here denoted by W_M , i.e.,

$$W_M = \frac{1}{M} \sum_{i=1}^M SE(\theta_i)^2,$$

whereas, the between-imputation variance, denoted B_M , is the variance due to missing data. This is the deviation of the estimated parameter for the i -th dataset from the mean parameter of all datasets:

$$B_M = \frac{1}{M-1} \left(\sum_{i=1}^M (\theta_i - \bar{\theta}) \right).$$

Consequently, the total (pooled) variance of the pooled estimate is the sum of the within- and between-imputation variances:

$$VAR_{pooled} = W_M + B_M + \frac{B_M}{M}.$$

For our purposes, the pooling step is done using the *pool()* function in the MICE R package. Figure 3 below shows a summary flow chart of the multiple imputation procedure. In the figure, there are five main stages; the first stage is the incomplete data set, whose values are imputed using M missing data models in the second stage. This yields M imputed datasets shown in stage 3. Subsequently, analyses are run on all the M datasets in stage 4, and finally in stage 5, the results from these M analyses are pooled together to form an output result.

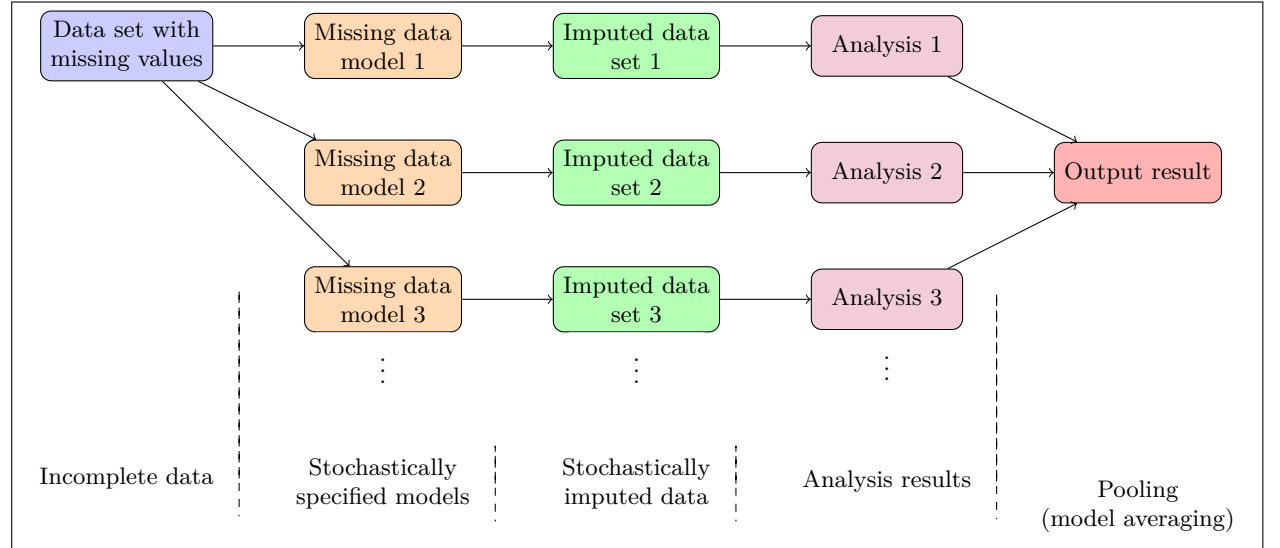


Figure 3: Summary of Multiple Imputation Procedure.

4.2 Machine Learning Imputation

An alternative to the traditional multiple imputation technique is imputation through machine-learning methods. Specifically, we consider the K-nearest neighbourhood (KNN) imputation technique, the Random Forest (RF) imputation technique, and the Self-Organising Maps (SOM) imputation technique.

Recent studies show that machine-learning methods generally handle missing values better than traditional imputation methods. For example, in their comparative study of

the two methods using medical data of patients with spontaneous intracerebral hemorrhage, Wang *et al.* [56] find that machine learning has better imputation performance compared to traditional imputation methods such as mean/mode imputation and multiple imputation. Another similar comparative study using breast cancer data also concludes that machine learning methods outperform traditional imputation methods in the prediction of patient outcome [25]. On the other hand, Emmanuel *et al.* [11] show that KNNs and random-forest-based algorithms can successfully handle missing values. Random forest imputation methods are found to be generally preferable as they can handle different types of data and are also able to capture non-linear structures much more accurately [51]. However, the performance of these machine-learning methods depends on the type of missing-data mechanism present in the data being used [11, 56].

Machine-learning based imputation involves building some predictive model that uses supervised or unsupervised learning techniques to handle missing data [11]. Much like regression imputation, the models used will predict a value for the missing value using complete cases in the data [13]. Once the missing values have been imputed with the predicted values, the training imputed data is then used to train a classification model [13].

For each method in our study, a cross-validation procedure is employed to choose hyperparameters. This is a type of Monte Carlo resampling techniques used to evaluate how well the predictive models generalises [16]. To conduct a cross-validation analysis, we take the subset of all complete cases in the entire dataset and split the data into the training and test set. For the training set, we simulate a 10% missingness rate to mimic that of the whole data, and train the machine-learning models on this incomplete dataset - using the methods to impute the simulated missing values. We then calculate the training misclassification error made by comparing the predicted values with the actual values in the dataset. For our purpose, a k-fold cross-validation is necessary for hyperparameter tuning. Particularly, it is necessary in choosing plausible number of nearest neighbours k in KNN imputation, choosing the number of optimal trees in Random Forest imputation, and choosing the grid dimensions, learning rate, and number of iterations in SOM imputation. Below we discuss how these machine-learning methods impute missing data, followed by each method's cross-validation procedure.

4.2.1 KNN Imputation

In KNN imputation, the algorithm identifies missing values in the incomplete dataset and uses the k -nearest data points to the missing values to fill in the blank using a similarity measure [22, 25]. Once the k -nearest neighbours have been identified, the algorithm uses a weighted mean of these observed data points as the estimate for the missing value [13]. In determining the k -nearest neighbours using the observed data points, several dissimilarity measures including Manhattan distance, Minkowski distance, and Euclidean distance can be used; however, the Euclidean distance is recommended [52, 25, 11]

If x_{ai} represents the value of the i -th attribute containing missing data (case a) and x_{bi} represents the value of the i -th attribute containing complete data (case b), for attributes $i \in D$ where D is the set of non-missing values in cases a and b , then the Euclidean distance between the two cases, denoted $Dist_{a,b}$, is defined as follows [60, 26]:

$$Dist_{a,b} = \sqrt{\sum_{i \in D} (x_{ai} - x_{bi})^2}.$$

The KNN algorithm will identify the k -nearest neighbours by minimising this distance function, with the value of k determined through a cross-validation. Once the k -nearest neighbours have been identified, the algorithm will then estimate a values for the missing values, depending on whether data points are discrete or numerical [25]. For discrete data points, the mode of the k -nearest neighbours is often selected as the imputation value whereas for numerical data, the mean is used [25]. For our research, the variables for which values we wish to impute are either categorical or discrete, in which case the categorical variables are re-coded to be discrete, and the estimated values are the modal class of the k -nearest neighbours. We implement KNN imputation using the VIM package in R.

KNN Cross-Validation

As mentioned above, the KNN imputation is computed using the `kNN()` function of the VIM package. The algorithm imputes data using an aggregation of k neighbours which, for categorical data, the category of the most occurrences in the k neighbours is used as the imputation value for the missing value [29]. The cross-validation procedure attempts to search for that k value which will yield the smallest error. A complete subset of the original dataset was used for the cross-validation procedure which is here below referred to as the complete data \mathbf{X} . Since the whole dataset is categorical, the error of interest is the misclassification error.

Given the complete data \mathbf{X} , a set of possible k values, for $i = 1, \dots, 10$ folds:

1. Introduce a 10% MAR missingness to create the incomplete data \mathbf{X}^{mis} .
2. At each i -th iteration:
 - a. Impute \mathbf{X}^{mis} using KNN imputation, over different values of k . Call the imputed dataset \mathbf{X}^{Imp} .
 - b. Calculate the misclassification error, ε :

$$\varepsilon = \frac{1}{n} \sum_{j=1}^n I(X_j \neq X_j^{\text{Imp}})$$

3. Select the best value of k to be the one yielding the minimum classification error.
 4. Use the optimal k value to impute the original data set.
-

Some key advantages of KNN imputation are that it is flexible between discrete and continuous data; it is robust procedure for estimating missing values; and it can be used to handle multiple missing values at a time [13, 25, 11]. A limitation of the KNN algorithm is that it can be computationally expensive, especially in large datasets [25]. Another limitation of the KNN imputation algorithm is that it tends to infer associations where they do not exist, thereby leading to low precision in the estimation of the imputation values [8].

4.2.2 Self-Organising Maps Imputation

Self-organising maps (SOMs) is an unsupervised learning technique used to reduce high-dimension data into a low-dimension grid while preserving the topological structure in the data [17]. SOMs impute missing data by finding the most similar pattern on a grid of neurons and using that pattern to fill in the gaps. The SOM consists of a grid of neurons each represented by weight vector \mathbf{w}_i (for $i = 1, \dots, n$ neurons) on the same dimension as the input vector \mathbf{X} , where the input vector is the data with missing values [15]. The training phase consists of three main steps: first, the input vector \mathbf{X} is presented and the weights initialised; secondly, each neuron is examined to determine which of the neurons' weights best matches the input vector through minimising some distance metric – the best weight is called the best matching unit (BMU); finally, the weights are updated through an update equation so that they best match the input vectors [15].

In the second step of the training phase, the distance between each i -th input vector x_i and the i -th neuron weight w_i is calculated often using the Euclidean distance [15]:

$$E = \sqrt{\sum_{i=1}^n (x_i - w_i)^2}$$

The BMU \mathbf{w}_c is the weight that has the smallest distance from the observation. During the calculation of the distances, the missing values in the input vector \mathbf{X} are ignored [25]. Once the BMU has been found, each weight vector is then updated using the updating rule:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \eta(t)h_{c,i}(t)(\mathbf{x} - \mathbf{w}_i(t)),$$

where t is the iteration step, $\eta(t) \in (0, 1)$ is the learning rate and $h_{c,i}(t)$ is the neighbourhood function that decreases over time [15, 25]. Most authors including Harp, Samad, and Villano (1995) [15] and Jerez *et al.* (2010) [25] take the neighbourhood function to be the Gaussian:

$$h_{c,i}(t) = \exp \left\{ -\frac{d^2(i, c)}{2\sigma^2(t)} \right\},$$

where $d^2(i, c)$ is the distance between the i -th neuron \mathbf{r}_i and the BMU's neuron \mathbf{r}_c , i.e., $d^2(i, c) = \|\mathbf{r}_i - \mathbf{r}_c\|^2$ and $\sigma^2(t)$ is the neighbourhood width parameter. For imputation, the missing values in the imputation vector \mathbf{X} are estimated by finding the BMU and using its corresponding weight vector to fill in the gaps [13, 25]. This ensures that the imputed values are consistent with the learned structure and relationships within the data. For our purposes, the SOM imputation is implemented using the missSOM package in R [41]. The problem with the SOM algorithm is that it handles only numeric data, and often the categorical variables are recoded to binary values prior to the training phase [23].

SOM Cross-Validation

For SOMs, model accuracy is affected by the grid dimensions, the initial learning rate, and the neighbourhood radius [55]. In particular, the grid of the SOM has a profound impact on the model accuracy. A map that is too small for the dataset has fewer neurons, which forces each weight vector to act as a representative for a larger group of training data points. Since there's less homogeneity in these data points, the best-matching units (BMUs) will not represent them accurately, leading to a higher quantisation error – the difference between the input data and the BMU [55]. In our cross-validation procedure, we search over different grid sizes to find the one that yields the minimum error. The procedure is described below:

Given the complete subset of the data \mathbf{X} , a grid of possible values of the dimensions (x_{dim} , y_{dim}), learning rate (α), and number of iterations t , over $k = 1, \dots, 10$ folds:

1. Introduce a 10% MAR missingness to create the incomplete data \mathbf{X}^{mis} .
2. At each k -th iteration:
 - a. Impute \mathbf{X}^{mis} using missSOM – searching over the grid of possible values of x_{dim} , y_{dim} , α , and t . Denote by \mathbf{X}^{Imp} the imputed dataset.
 - b. Calculate the misclassification error, ε :

$$\varepsilon = \frac{1}{n} \sum_{j=1}^n I(X_j \neq X_j^{\text{Imp}})$$

3. The optimal parameters are the set of x_{dim} , y_{dim} , α , and t that yield the minimum classification error.
 4. Using the optimal set of parameters from 3) above, impute the original dataset.
-

4.2.3 Random Forest Imputation

The random forest imputation algorithm imputes missing values by using the complete cases in the data to train a random forest model, and then uses the model to predict values for the missing values [21]. The missing data are initially imputed as the mean of the variable (for numeric values) or the modal class of the variable (for categorical data). After this pre-processing step, the algorithm then imputes the missing values as follows [21]: Given the data matrix $\mathbf{X} : n \times p$, the data is divided into complete \mathbf{X}^C and incomplete cases \mathbf{X}^{miss} , with the \mathbf{X}^C acting as the dependent variables and \mathbf{X}^{miss} as the regressors. During the training phase, the \mathbf{X}^C are used to create a random forest model by finding optimal splits that make decision rules in the trees. The trained model is then used to predict values for the missing values. In this research, the missForest algorithm is used to impute the data. The missForest algorithm imputes data in the same way in that each of the p variables are regressed in turn against all other variables, resulting in p random forest models being fitted for each iteration [50]. For random forests, the number of trees grown by the algorithm is a hyperparameter, as in the absence of a stopping criterion the algorithm can grow a large and complex forest whose decision rules are unclear. For this purpose, a cross-validation procedure is used to choose optimal number of trees.

Random Forest Cross-Validation

The missForest package takes as a parameter the number of trees (*ntree*) to grow in each forest. The algorithm stores out-of-bag (OOB) error and the true imputation

error after the imputation process, with the true imputation error calculated on the condition that the true dataset is provided [50]. The OOB error is the average prediction error calculated from the observations that were not included (out-of-sample) in the training of each individual tree [16]. Whereas, the imputation error is the proportion of imputed values that differ from the actual values in the original dataset. For our cross-validation procedure, we track not only these errors but include the misclassification error which was used for the KNN and SOM imputation methods.

Given the complete data \mathbf{X} , a set of possible number of trees, $ntree$, and $k = 1, \dots, 10$ folds:

1. Introduce 10% MAR missingness to \mathbf{X} to create the missing dataset, \mathbf{X}^{mis} .
2. At each k-th iteration:
 - a. Impute \mathbf{X}^{mis} using missForest, searching across different values of $ntree$ at each iteration.
 - b. Calculate the misclassification error, ε :

$$\varepsilon = \frac{1}{n} \sum_{j=1}^n I(X_j \neq X_j^{\text{Imp}})$$

- c. Calculate the OOB error and the imputation error.
 3. Choose the optimal value of $ntree$ to be the one corresponding to the minimum classification error, OOB error, and imputation error.
 4. Impute the original dataset using this optimal value of $ntree$.
-

4.3 Logistic Regression Analysis

After imputing the missing values using the traditional and machine-learning imputation techniques, logistic regression analyses were performed on the imputed datasets and the incomplete data. For the risky sexual behaviour response variable Y , defined as

$$Y = \begin{cases} 1 & \text{if risky sexual behaviour occurs} \\ 0 & \text{if no risky sexual behaviour occurs} \end{cases},$$

given the covariates X_1, X_2, \dots, X_p , the logisitic regression model is defined as:

$$\text{logit}(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

First, univariate logistic regression models are fit for 39 factors and risky sexual behaviour to identify potential relationships. A p-value equal to or below 0.05 indicates a significant univariate association. The multivariate logistic regression models were built by including all the univariately significant variables. This was termed by Khati as a "forced" multivariate logistic regression model. This is due to the lack of proper model-building techniques or consideration of any interactions in favour of a more straightforward approach.

The performances of the multivariate logistic regression models are assessed by their sensitivity, specificity, negative predictive value (NPV), and the area under the curve (AUC). Binned residual plots are used to examine the homoscedasticity and potential outliers within each model,

5 Results

This section tables the results obtained from implementing our methodology above. The section commences by discussing logistic regression results from the complete-case analyses of the original data when no imputation is conducted. Next, the results from the logistic regression analyses on data imputed using multiple imputation are tabled. Before similar analyses can be conducted for the three machine-learning based imputation techniques, the results from the cross-validation procedure used to choose optimal parameters are reported, after which selection of the optimal parameters are the results from the logistic regression analyses from these methods tabled. In all cases, results for the Gauteng and Western Cape provinces are reported separately. Finally, the section concludes by reporting model performance metrics for the original-data analysis, the multiple-imputation data analysis, and the machine-learning-imputation data analyses - where the comparison between imputation methods is conducted.

To start off, univariate logistic regression analyses were performed to identify associations with RSB and for Gauteng and the Western Cape, separately. The unadjusted odds ratios used in the regressions characterise the strength of association between the two variables. Next, multivariate logistic regression analysis was performed for both regions separately. All variables with univariate significance ($p < 0.05$) were granted automatic entry into their respective multivariate logistic regression models. To account for the relatively small datasets and the sizable number of missing observations, additional variables were considered for entry into the final model. A combination of the strength of a variable's marginal significance as well as its prevalence in the literature could grant it access to the multivariate model.

The process of univariate and multivariate analyses was repeated for each imputation method. The aim was to fit models with the same covariates to enable like for like comparisons. The results of the multivariate logistic regression models - the original model and the models obtained after each imputation method - were compared to each other. The performance of the machine-learning based techniques was compared, firstly, during the cross-validation stage, where the training and test classification errors were compared for the three methods. Those methods which perform poorly will generally have a lower training error and a high test error, thereby indicating over-fitting in the training phase. This comparison allows us to assess the imputation accuracy among the machine-learning imputation methods. However, since multiple imputation does not have a cross-validation procedure, we do not rely on the cross-validation results to draw a comparison between the traditional imputation method and the machine learning imputation methods. Instead, we assess the results of the logistic regression model trained on the data imputed by these methods. Particularly, we performance metrics such as Akaike Information Criterion (AIC), sensitivity, specificity, negative predictive value (NPV), and area under the curve (AUC) values arising from logistic regressions to compare how the different imputation techniques influence the results.

5.1 Multivariate Analysis of Factors Associated with Risky Sexual Behaviour

The following factors were found to have univariate significances in the Gauteng region: Age ($p < 0.01$), Marital status ($p < 0.01$), Language (Afrikaans $p < 0.01$), Racial classification ($p < 0.01$), Currently working ($p = 0.02$), Childless choice is right ($p = 0.01$), Current partner age ($p < 0.01$), Current partner employed ($p = 0.03$), Easy access to alcohol ($p = 0.06$), Binge drinking ($p = 0.04$), Risk drinking ($p < 0.01$), Importance of condom use with spouse/regular partner ($p < 0.01$), Current contraceptive use ($p < 0.01$), Have children ($p = 0.01$), Lifetime contraceptive use ($p < 0.01$).

For the rural Western Cape region, the following variables were found to have univariate significance: Language (Afrikaans $p = 0.04$), Racial classification (Coloured $p < 0.01$), Current partner satisfaction ($p = 0.06$), Current partner employed ($p = 0.04$), lifetime cigarette use ($p = 0.05$), current sex partner ($p = 0.04$). The factor pertaining to the perceived importance of condom usage with casual partner ($p = 0.11$) was included into the multivariate model due to significant results in the literature review.

5.1.1 Urban Gauteng

Pre-analysis, 562 observations were deleted due to missingness in the Gauteng data. For the urban Gauteng women, fourteen factors were found to be univariately significant A.1. The race factor, despite having a significant association, was omitted from the final multivariate logistic regression model to improve the interpretability of the results.

Only three factors retained their significance in this multivariate regression model (Table 15): Language, Marital status, and Importance of condom usage with a spouse/regular partner. Women in Gauteng who spoke Afrikaans as their first language did not engage in risky sexual behaviour ($OR = 0$, $p = 0.05$). Similarly, women who were traditionally married were not identified to engage in risky sexual behaviour ($OR = 0$, $p = 0.04$). The factor "Importance of condom use with spouse/regular partner" produced a very high odds ratio ($OR = 94.31$, $p = 0.04$).

Many factors retained some marginal significance, though. Multivariately, there is a marginal association between risky sexual behaviour and age, lifetime contraceptive use, binge drinking, risk drinking, and the choice to be childless. The odds ratios (OR) for the age factor for women aged 25-34 ($p = 0.09$) and 35-44 ($p = 0.06$) are both zero indicating that no observations over the age of 24 engaged in any risky sexual behaviour. Similarly, with the binge drinking ($p = 0.06$) and Lifetime contraceptive use ($p = 0.07$) factors, the OR s of zero indicate that women who partook in binge drinking or had used any form of contraception did not engage in any risky sexual behaviour. The Risk

drinking (OR=875.08, p=0.07) and "Childless choice is wrong" (OR=1.76e+6, p=0.08) variables have very high odds ratios indicating a high likelihood of risky sexual behaviour.

Table 15: Multivariate logistic regression analysis of factors associated with risky sexual behaviour for women in the urban Gauteng region.

Variable	Gauteng	
	OR \pm SE (95% CI)	p-value
Age (years)		
18-24	Naturally coded	-
25-34	0.00 \pm 14.49 (0.00-0.01)	0.09
35-44	0.00 \pm 26.86 (0.00-0.00)	0.06
First Language		
Native Language	Naturally coded	-
Afrikaans	0.00 \pm 12.46 (0.00-0.00)	0.04
English	0.00 \pm 7.38 (0.00-0.47)	0.12
Race	-	-
Marital Status		
Legally married	Naturally coded	-
Traditionally married	0.00 \pm 14.29 (0.00-0.00)	0.04
Currently working	0.24 \pm 1.36 (0.01-3.69)	0.29
Older partner (≥ 30 years)	2.15e+4 \pm 6.92 (0.35-9.04e+11)	0.15
Current Partner employed	1.60 \pm 2.13 (0.02-2.06e+2)	0.83
Lifetime contraceptive use	0.00 \pm 17.53 (0.00-0.00)	0.06
Binge drinking	0.00 \pm 12.02 (0.00-0.00)	0.06
Risk drinking	875.08 \pm 3.67 (3.04-1.85e+7)	0.07
Importance of condom use with spouse/regular partner	1043.09 \pm 3.30 (5.65-3.54e+6)	0.04
Current contraceptive use	0.06 \pm 2.01 (0.00-2.62)	0.17
Have child(ren)	18.96 \pm 3.02 (0.08-4.49e+4)	0.33
Childless choice is wrong	1.76e+6 \pm 8.28 (20.01-4.22e+16)	0.08
AIC	55.93	-

5.1.2 Rural Western Cape

Before the analysis for the Western Cape region can be conducted, 96 observations were deleted due to having missing values for some of the factors. Subsequently, in the complete-case analysis for the rural Western Cape region, six variables were found to be univariately significant A.1. Race was omitted from the regression due to its confounding effects with language. Additionally, the factor pertaining to the importance of condom usage with a casual partner was included in the multivariate regression due to its marginal

univariate significance, as well as its significance in the literature.

In the multivariate regression, the "Afrikaans" language , "Importance of condom usage with spouse/regular partner" and "Importance of condom usage with a casual partner" (OR=0.23, p=0.04) emerged as the only significant variables. Women who speak Afrikaans in the Western Cape are 79% less likely to engage in RSB compared to their Native Language speaking counterparts. Women who see the importance of condom usage with their spouses or regular partners were over five times more likely to engage in RSB (OR=5.19, p<0.01). Conversely, women who emphasize the usage of condoms with casual partners were 77% less likely to engage in RSB compared to women who did not see the importance of condom usage (OR=0.21, p=0.02).

The variable pertaining to the employment status of a woman's current partner has a marginal association with risky sexual behaviour (OR=0.27, p=0.10). Women with employed partners are 73% less likely to engage in RSB compared to women with unemployed partners.

Table 16: Multivariate logistic regression analysis of factors associated with risky sexual behaviour for women in the rural Western Cape region.

Variable	Western Cape	
	OR \pm SE (95% CI)	p-value
(Intercept)	7.23 \pm 1.56 (0.31-150.36)	0.20
Language		
Native Language	Naturally coded	-
Afrikaans	0.21 \pm 0.66 (0.06-0.79)	0.02
English	4.54e+6 \pm 1455.40 (0.00-Inf)	0.99
Household Hunger	1.43 \pm 0.60 (0.40-4.42)	0.55
Satisfied with partner	0.52 \pm 0.89 (0.11-4.01)	0.46
Current partner employed	0.27 \pm 0.81 (0.06-1.52)	0.10
Lifetime cigarette use	0.81 \pm 0.53 (0.29-2.37)	0.69
Husband/Boyfriend most recent sex partner	0.39 \pm 0.69 (0.10-1.67)	0.17
Importance of condom use with spouse/regular partner	5.19 \pm 0.56 (1.77-16.77)	<0.01
Importance of condom use with casual partner	0.23 \pm 0.71 (0.06-1.02)	0.04
AIC	142.47	-

5.2 Multiple Imputation multivariate analyses

Tables 17 and 18 show the multivariate logistic regression results obtained using the pooled dataset created using the MICE algorithm for urban Gauteng and rural Western Cape, respectively.

5.2.1 Urban Gauteng

Univariately, a similar list of variables were significant when using the MICE imputed dataset. The variable pertaining to the perceived ease of accessing alcohol is included in the multivariate regression, due to its univariate significance (OR=0.64, $p=0.04$) (Table 17).

Of the 15 variables, five of them emerged to have multivariate significance. Older women between the ages of 35-44 had a significant lower odds of engaging in any forms of RSB (OR=0.32, $p=0.04$) compared to younger women aged 18-24 years. Similarly, Coloured women were also found to have a lower odds of engaging in RSB compared to women belonging to other racial classifications (OR=0.13, $p=0.03$). The act of binge drinking increased the likelihood of RSB twofold (OR=2.16, $p=0.01$). Women who see the importance of condom usage with their spouses or regular partners also had a higher odds of engaging in RSB at a 1% significance level (OR=2.07, $p<0.01$). Similarly, women who indicated current contraceptive use were two times more likely to engage in risky sexual behaviour compared to women who had not recently used any contraception (OR=2.01, $p=0.03$).

Table 17: Multivariate logistic regression analysis of factors associated with risky sexual behaviour for women in the urban Gauteng region using the MICE imputed dataset.

Variable	Gauteng	
	OR \pm SE (95% CI)	p-value
Age (years)		
18-24	Naturally coded	-
25-34	0.49 \pm 0.42 (0.21-1.11)	0.10
35-44	0.32 \pm 0.50 (0.12-0.86)	0.04
Marital Status		
Legally married	Naturally coded	-
Traditionally married	0.95 \pm 0.24 (0.59-1.51)	0.82
Racial Classification		
Black/African	Naturally coded	-
White	0.14 \pm 1.08 (0.02-1.16)	0.07
Coloured	0.13 \pm 0.96 (0.02-0.85)	0.03
First Language		
Native Language	Naturally coded	-
Afrikaans	2.67 \pm 0.99 (0.38-18.60)	0.32
English	4.10 \pm 0.89 (0.72, 23.46)	0.11
Currently working	1.04 \pm 0.24 (0.65-1.67)	0.86
Older partner (≥ 30 yrs)	0.83 \pm 0.42 (0.36-1.88)	0.66
Childless choice is wrong	1.38 \pm 0.22 (0.90-2.12)	0.14
Current partner employed	0.81 \pm 0.26 (0.48-1.34)	0.41
Lifetime contraceptive use	1.49 \pm 0.22 (0.97-2.29)	0.08
Binge drinking	2.16 \pm 0.28 (1.25-3.74)	0.01
Risk drinking	1.58 \pm 0.49 (0.60-4.12)	0.40
Importance of condom use with spouse/regular partner	2.07 \pm 0.24 (1.29-3.31)	<0.01
Current contraceptive use	2.01 \pm 0.29 (1.14-3.55)	0.03
Have child(ren)	1.32 \pm 0.37 (0.64-2.72)	0.46
Easy access to alcohol	1.36 \pm 0.21 (0.90-2.06)	0.15
AIC	740.42	-

5.2.2 Rural Western Cape

Only two variables emerged to have any multivariate significance with risky sexual behaviour; The Afrikaans language and the perceived importance of condom use with spouse or regular partner.

The Afrikaans language is a potential protective factor against RSB. Women who speak

Afrikaans as their first language were 79% less likely to engage in risky sexual behaviour compared to women who primarily spoke any of the native languages (OR=0.21, p=0.01). Similar to the original multivariate logistic regression (Table 16), Women who deemed it important to use a condom with their spouse or regular partner were 4.69 times more likely to engage in RSB compared to women who did not see condom usage with their regular partner as important.

Table 18: Multivariate logistic regression analysis of factors associated with risky sexual behaviour for women in the rural Western Cape region using the MICE imputed dataset.

Variable	Western Cape	
	OR \pm SE (95% CI)	p-value
Language		
Native Language	Naturally coded	-
Afrikaans	0.21 \pm 0.61 (0.06-0.68)	0.01
English	1.69 \pm 1.83 (0.05-60.95)	0.77
Household Hunger	1.72 \pm 0.54 (0.60-4.97)	0.32
Satisfied with partner	0.98 \pm 0.91 (0.17-5.86)	0.99
Current partner employed	0.47 \pm 0.72 (0.11-1.92)	0.30
Lifetime cigarette use	0.81 \pm 0.48 (0.32-2.08)	0.67
Husband/Boyfriend most recent sex partner	0.41 \pm 0.63 (0.12-1.43)	0.17
Importance of condom use with spouse/regular partner	4.69 \pm 0.54 (1.63-13.51)	<0.01
Importance of condom use with casual partner	0.33 \pm 0.70 (0.08-1.29)	0.11
AIC	175.83	-

5.3 Machine Learning Cross-Validation Results

Before the machine-learning imputation methods can be implemented, the parameters of the models were chosen through cross-validation. Figure 4 shows the KNN imputation training errors across different values of k , with the optimal value of k indicated on the cross-validation plot. For SOM cross-validation in Figure 5, the minimum misclassification error occurs at various dimensions of the grid. In the figure, the same (x, y) dimensions yield different errors due to the different k folds and different combinations of the learning rate and number of iterations t . The cross-validation plot shows that the (10,5) grid dimensions yields the lowest classification training error. However, this is not a unique solution, as the same lowest error is produced elsewhere in the grid search. We choose the optimal grid dimensions as (5,5) so that fewer neurons are used as this increases computational efficiency. The accompanying values of t ($rlen$) and α , shown in Table 19 below, are also used.

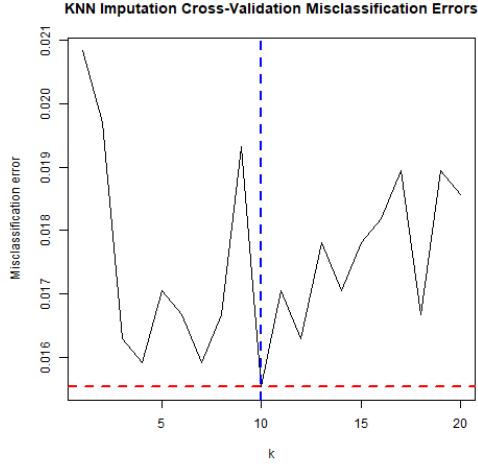


Figure 4: KNN training errors

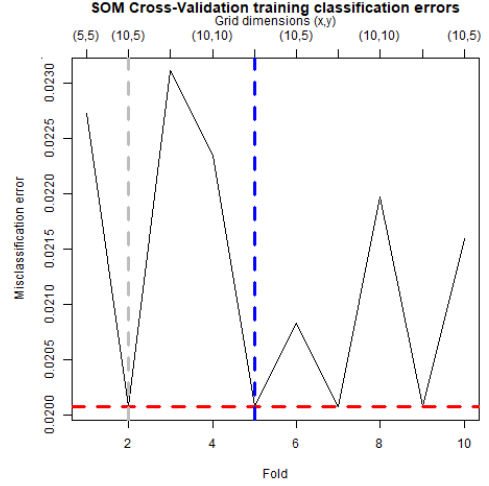


Figure 5: SOM training errors

For random forest cross-validation in Figure 6, the imputation errors in panel (a) and the OOB errors in panel (b) were tracked by the missForest package at every iteration of the k -fold cross validation. The imputation-error plot shows that the minimum error occurs at 12 trees in the random forest, whereas the misclassification-error plot in panel (b) suggests that the minimum misclassification error occurs at 16 trees in the random forest. On the other hand, the OOB-error plot in panel (c) suggests the optimal number of trees to be 64. The number of optimal trees suggested by the missForest imputation errors is not very far from the one suggested by our manually-tracked classification errors, and so we choose the optimal number of trees from the misclassification-error plot ($ntree=16$) as this is consistent with the criteria defined for KNN and SOM imputations.

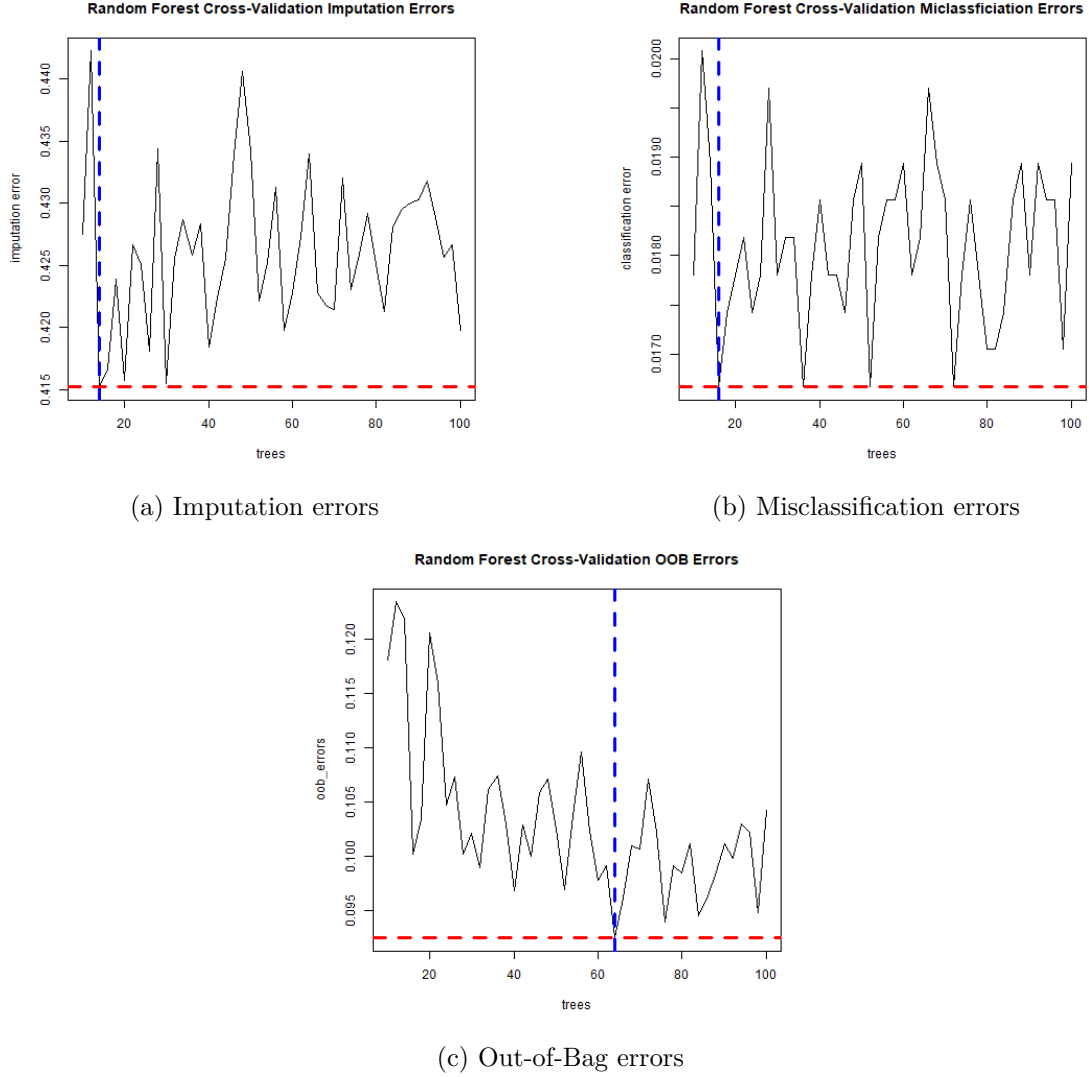


Figure 6: Random Forest Imputation training errors

After choosing the optimal parameters for the different methods in Table 19, they were used to impute the unseen test set with the test misclassification errors being recorded for each method. The KNN imputation method is outperformed by the random forest and SOM methods, which yield the same misclassification error of 0.0273, while this misclassification error for KNN imputation is 0.0333. The KNN imputation method tends to overfit during the training phase in comparison to the other methods, as it is best-forming during the training stage with the lowest training misclassification error of 0.0155. Overall, the imputation techniques seem to impute the data very well, as the errors made are consistently lower than 4% in both the training and testing phase.

Table 19: Comparison of different imputation methods and their optimal parameters, along with misclassification errors.

Method	Package	Optimal parameters	Training Error	Test Error
KNN imputation	VIM package	$k = 10$	0.0155	0.0333
Random Forest (RF) Imputation	missForest	$ntrees = 16$	0.0167	0.0273
Self-organising Maps (SOM) imputation	missSOM	$x_{dim} = 5$ $y_{dim} = 5$ $r_{len} = 100$ $alpha = 0.05$	0.0201	0.0273

These optimal parameters arising from the cross-validation procedures are used in the imputation of the whole dataset.

5.4 Machine Learning Multivariate Analyses

The multivariate logistic regression results for the KNN-imputed, SOM-imputed, and RF-imputed Gauteng datasets in Table 20 below shows that significant factors explaining variation in risky sexual behaviour among Gauteng women include age (34-44 years), the current-partner-employed factor, current- and lifetime-contraceptive use, risk drinking, and importance in condom usage.

For the *age* factor, the model based on the KNN-imputed dataset implies that there is no significance of the 25-34 age category (p-value = 0.18), whereas the models based on the SOM-imputed and RF-imputed datasets suggest a marginal significance at the 10% level of significance (p-value = 0.09) and the 5% level of significance (p-value = 0.06) respectively. These models based on the SOM-imputed and RF-imputed datasets respectively shows that women aged 25-34 could potentially have a 44% and 47% reduction in the odds of engaging in risky sexual behaviour compared to those aged between 18-24. Whereas, the models based on the three datasets all suggest that there is strong significance in the age 34-44 category, as the models built on the KNN-imputed, SOM-imputed, and RF-imputed datasets show that the factor is highly significant at the 5% (for KNN) and 1% levels of significance (for SOM and RF). All models show that there is a significant decrease in the odds of women aged between 34-44 years engaging in risky sexual behaviour compared to those aged between 18-24, with the model based on the KNN-imputed data suggesting a 54% decrease and the model based on the SOM-imputed data suggesting a 65% decrease whereas the model based on the RF-imputed data suggests a 66% decrease in the odds.

On the other hand, the *current-partner-employed* factor is significant in the models based

on the KNN-imputed data (p-value = 0.03) and the SOM-imputed data (p-value = 0.02). Both models suggest that having a partner that is employed is associated with lower odds of engaging in risky sexual behaviour, with the model on the KNN-imputed data suggesting a 41% reduction in the odds while the model on the SOM-imputed data suggests a 44% reduction in the odds of engaging in risky sexual behaviour compared to those women whose current partners are not employed.

Lifetime contraceptive use is a significant factor across all models built on the machine-learning-imputed datasets (p-value < 0.01), with odds ratios indicating increased odds of engaging in risky sexual behaviour (KNN:1.76, SOM:2.00, RF:1.89). The results show that women who have used contraceptives in their lifetime have higher odds of engaging in risky sexual behaviour than those who haven't. Similarly, the *current contraceptive use* factor is significant in the models based on the KNN- and RF-imputed datasets at the 5% (p-value = 0.04) and 1% (p-value <0.01) levels of significance respectively. The models estimate odds ratios of 1.65 for KNN and 2.34 for RF, suggesting that there are higher odds of engaging in risky sexual behaviour for those who are currently on contraceptives. This factor is not significant in the model based on the SOM-imputed dataset however (p-value = 0.22).

Risk Drinking is also a highly significant factor across all models built on the machine-learning-imputed datasets (p-value < 0.01). The odds ratios of 4.33 (KNN), 2.3 (SOM), and 4.4 (RF) suggest that risk drinking is highly associated with increased odds of engaging in risky sexual behaviour; that is, women who engage in risky drinking are highly likely to engage in risky sexual behaviour. While *binge drinking* also suggests increased odds of engaging in risky sexual behaviour, this factor is only marginally significant for the model based on the SOM-imputed dataset (p-value = 0.04). Much like *Risk Drinking*, the *Importance of Condom Usage* factor is highly significant across all the models built on the KNN-, SOM-, and RF-imputed datasets, all with a p-value of less than 0.01. The odds ratios (KNN:2.10, SOM: 2.02, RF: 2.40) indicate that the *importance of condom usage* is associated with increased odds of engaging in risky sexual behaviour. That is, women who value the importance of using condoms with their spouses and/or regular partner tend to have higher odds of engaging in risky sexual behaviour.

Overall, the models show that key factors significantly associated with risky sexual behaviour include *age* (34-44 years), *current partner employment*, *lifetime contraceptive use*, *risk drinking*, and *importance of condom usage*. These factors are significant at the 1% level and have narrower confidence intervals – thereby depicting more certainty around odds ratio estimates. Another factor which could also explain risky sexual behaviour among women in urban Gauteng is *current contraceptive use*, as it is significant at the 5% level. The model shows some variability across the different machine-learning-imputed datasets, with models built on SOM- and RF-imputed datasets mostly capturing slightly different aspects than the model based on KNN-imputed data.

Table 20: Multivariate Logistic Regression results for KNN, SOM, and RF Models for urban Gauteng

	KNN		SOM		RF	
Variable	OR±SE (95% CI)	p - value	OR±SE (95% CI)	p - value	OR±SE (95% CI)	p - value
Age (years)						
18-24	Naturally coded	-	Naturally coded	-	Naturally coded	-
25-34	0.68 ± 0.29 (0.38, 1.21)	0.18	0.56 ± 0.34 (0.28, 1.09)	0.09	0.53 ± 0.34 (0.27, 1.02)	0.06
34-44	0.46 ± 0.35 (0.23, 0.90)	0.02	0.35 ± 0.39 (0.16, 0.74)	0.01	0.34 ± 0.39 (0.16, 0.73)	0.01
Marital Status						
Traditionally married	Naturally coded	-	Naturally coded	-	Naturally coded	-
Legally married	1.09 ± 0.22 (0.70, 1.68)	0.71	0.98 ± 0.22 (0.64, 1.50)	0.91	0.99 ± 0.22 (0.64, 1.55)	0.98
Language						
Native Language	Naturally coded	-	Naturally coded	-	Naturally coded	-
Afrikaans	1.22 ± 0.93 (0.19, 7.60)	0.83	1.77 ± 1.01 (0.24, 13.91)	0.57	1.79 ± 1.02 (0.24, 14.25)	0.57
English	2.60 ± 0.83 (0.48, 13.80)	0.25	3.23 ± 0.87 (0.57, 19.57)	0.57	2.96 ± 0.88 (0.52, 18.05)	0.22
Racial Classification						
Black/African	Naturally coded	-	Naturally coded	-	Naturally coded	-
White	0.29 ± 1.02 (0.04, 2.19)	0.22	0.19 ± 1.09 (0.02, 1.57)	0.13	0.21 ± 1.10 (0.02, 1.81)	0.16
Coloured	0.22 ± 0.90 (0.04, 1.35)	0.10	0.17 ± 0.97 (0.02, 1.11)	0.07	0.20 ± 0.98 (0.03, 1.31)	0.10
Currently Working	0.99 ± 0.21 (0.66, 1.49)	0.95	0.86 ± 0.21 (0.57, 1.29)	0.48	0.99 ± 0.21 (0.66, 1.49)	0.96
Older partner (≥ 30yrs)	0.69 ± 0.29 (0.39, 1.21)	0.19	1.26 ± 0.32 (0.68, 2.37)	0.47	0.92 ± 0.36 (0.45, 1.90)	0.83

	KNN		SOM		RF	
Current partner employed	0.59 ± 0.24 (0.37, 0.95)	0.03	0.56 ± 0.25 (0.35, 0.91)	0.02	0.71 ± 0.24 (0.44, 1.13)	0.15
Easy access to alcohol	0.74 ± 0.25 (0.46, 1.20)	0.22	0.77 ± 0.24 (0.48, 1.24)	0.29	0.82 ± 0.25 (0.51, 1.34)	0.44
Lifetime contraceptive use	1.76 ± 0.20 (1.19, 2.62)	0.00	2.00 ± 0.20 (1.36, 2.96)	0.00	1.89 ± 0.20 (1.28, 2.81)	0.00
Importance of condom use	1.41 ± 0.32 (0.84, 2.37)	0.20	1.32 ± 0.32 (0.76, 2.29)	0.32	1.16 ± 0.32 (0.66, 2.04)	0.59
Number of Previous Pregnancies	1.11 ± 0.30 (0.66, 1.88)	0.69	1.12 ± 0.30 (0.65, 1.90)	0.69	1.06 ± 0.30 (0.61, 1.83)	0.82
Current contraceptive use	1.46 ± 0.30 (0.86, 2.50)	0.15	1.51 ± 0.30 (0.88, 2.61)	0.13	1.47 ± 0.30 (0.85, 2.54)	0.17
AIC	722.84	-	736.79	-	171.33	-

Table 21: Multivariate Logistic Regression results for KNN, SOM, and RF Models for rural Western Cape

	KNN		SOM		RF	
Variable	OR±SE (95% CI)	p - value	OR±SE (95% CI)	p - value	OR±SE (95% CI)	p - value
Language						
Native Language	Naturally coded	-	Naturally coded	-	Naturally coded	-
Afrikaans	0.21 ± 0.62 (0.06, 0.76)	0.01	0.22 ± 0.61 (0.07, 0.79)	0.01	0.21 ± 0.62 (0.06, 0.75)	0.01
English	1.34 ± 1.67 (0.04, 46.45)	0.86	1.41 ± 1.67 (0.04, 48.73)	0.84	1.34 ± 1.67 (0.04, 46.73)	0.86
Household	1.50 ± 0.55 (0.47, 4.21)	0.46	1.52 ± 0.55 (0.48, 4.26)	0.45	1.52 ± 0.55 (0.48, 4.27)	0.44
Hunger						
Satisfied with current partner	0.33 ± 0.79 (0.08, 1.94)	0.17	0.34 ± 0.79 (0.08, 1.95)	0.17	0.33 ± 0.79 (0.08, 1.93)	0.16
Current partner employed	0.21 ± 0.80 (0.05, 1.18)	0.05	0.21 ± 0.79 (0.05, 1.16)	0.05	0.21 ± 0.80 (0.05, 1.17)	0.05
Lifetime cigarette use	0.90 ± 0.49 (0.35, 2.45)	0.83	0.87 ± 0.49 (0.34, 2.36)	0.78	0.90 ± 0.49 (0.35, 2.45)	0.83
Husband/boyfriend is current sex partner	0.44 ± 0.62 (0.14, 1.60)	0.18	0.44 ± 0.62 (0.14, 1.58)	0.18	0.43 ± 0.61 (0.14, 1.57)	0.17
Importance of condom usage: casual partner	0.26 ± 0.68 (0.07, 1.12)	0.05	0.26 ± 0.68 (0.07, 1.11)	0.05	0.27 ± 0.68 (0.07, 1.12)	0.05
Importance of condom usage: spouse/regular partner	5.29 ± 0.53 (1.94, 16.16)	0.00	5.12 ± 0.53 (1.88, 15.63)	0.00	5.15 ± 0.53 (1.89, 15.74)	0.00
AIC	171.01	-	171.42	-	171.33	-

Table 21 above shows the multivariate logistic regression results built on the KNN-, SOM- and RF-imputed datasets for rural Western Cape. The results show that for women in rural Western Cape, risky sexual behaviour is significantly influenced by *Language (Afrikaans)*, *current partner's employment status*, *the importance of condom usage with casual partners* and the *importance of condom usage with spouses/regular partners*.

The *Language* (Afrikaans) factor is significant across all models built on the different machine-learning-imputed datasets, with a p-value of less than 0.01 for all the models. The odds ratios are consistently below 1 (KNN: 0.21, SOM: 0.22, RF: 0.21), suggesting a protective effect: That is, Afrikaans speakers have lower odds of engaging in risky sexual behaviour compared to those with a native language.

Whereas, the *current-partner-employed* factor is consistently marginally significant at the 5% level (p-value = 0.05) across all models built on the different machine-learning-imputed datasets. The models from the datasets all estimate the odds ratio of 0.21, suggesting that women whose current partners are employed are 79% less likely to engage in risky sexual behaviour than those whose partners are not currently employed. Also consistently marginally significant at the 5% level, the *importance of condom usage with casual partner* factor has an odds ratio of around 0.26 for all models from the different dataset, suggesting that there is a 74% reduction in the odds of engaging in risky sexual behaviour for women who emphasise use of condoms with their casual partners, than those who don't.

On the contrary, the results across the different datasets are highly significantly agreed (at the 1% level) that emphasising the importance of condom usage with spouse/regular partner increases the odds of engaging in risky sexual behaviour. The odds ratios greater than five (KNN: 5.29, SOM: 5.12, RF: 5.15) specifically suggest that women who use condoms with their spouses/regular partners are 5 times more likely to engage in risky sexual behaviour than those who don't.

Overall, women who are Afrikaans speakers and whose partners are currently employed, as well as those who use condoms with their casual partners have lower odds of engaging in risky sexual behaviour. Whereas those women who use condoms with their spouses/regular partners are more likely to engage in risky sexual behaviour.

5.5 Model Evaluation

In this section we evaluate the performance of all our models built from the MICE and machine learning imputation methods. By comparing metrics such as sensitivity, specificity, negative predictive value (NPV), and area under the curve (AUC), we aim to identify which methods provide the most reliable insights into the underlying patterns in the data.

The binned residuals are used to assess the goodness of fit of the models. The data is first divided into categories (bins) based on their fitted values. Then the binned residuals plot the average residual versus the average fitted value for each bin [57].

5.5.1 Urban Gauteng Region

Table 22 presents a comparison of performance metrics, including sensitivity, specificity, negative predictive value (NPV), and area under the curve (AUC), across various imputation methods (MICE, KNN, SOM, and RF) used in multivariate regression analyses for the Gauteng region.

The Random Forest (RF) model has the highest area under the curve (AUC) value of 0.764. This is followed closely by the KNN and SOM models which have AUC values of 0.758 and 0.748, respectively. The MICE model has the lowest AUC at 0.738, indicating a weaker ability to capture the data's underlying patterns than the machine learning approaches. Regarding sensitivity, KNN achieves the highest score of 0.853, indicating that it correctly identifies a larger proportion of true positives, while MICE shows a notably lower sensitivity of 0.812. MICE also has the lowest specificity (0.026), which measures the true negative rate. The negative predictive value (NPV) is highest for the RF model at 0.750, suggesting it is the most reliable in predicting negative outcomes amongst all models. Overall, these results highlight the effectiveness of particularly the RF and KNN machine learning imputation methods, in modeling risky sexual behaviour in the Gauteng region.

Table 22: Model Performance Metrics for Multivariate Regression Models in the Gauteng Region.

Performance Metric	MICE	KNN	SOM	RF
Sensitivity	0.812	0.853	0.831	0.836
Specificity	0.026	0.525	0.563	0.567
NPV	0.042	0.700	0.685	0.750
AUC	0.738	0.758	0.748	0.764

The Original model's binned residual plot depicted in Figure 7 below shows a diamond shape centered around zero, indicating consistent residuals with low variance. This

suggests that the model has a stable performance across different expected values. The MICE model has a scattered distribution of residuals with noticeable variation, especially in the upper and lower bounds. The residuals fluctuation indicates potential bias or variance in prediction. The KNN plot shows a wide range of residuals with some clustering. The variability suggests that the KNN model might not be capturing patterns as well as desired, leading to inconsistent predictions. The SOM Model plot also exhibits scattered residuals but with slightly less variability than KNN. The RF Model shows a distribution of residuals similar to those of the MICE and KNN models, with evident fluctuations.

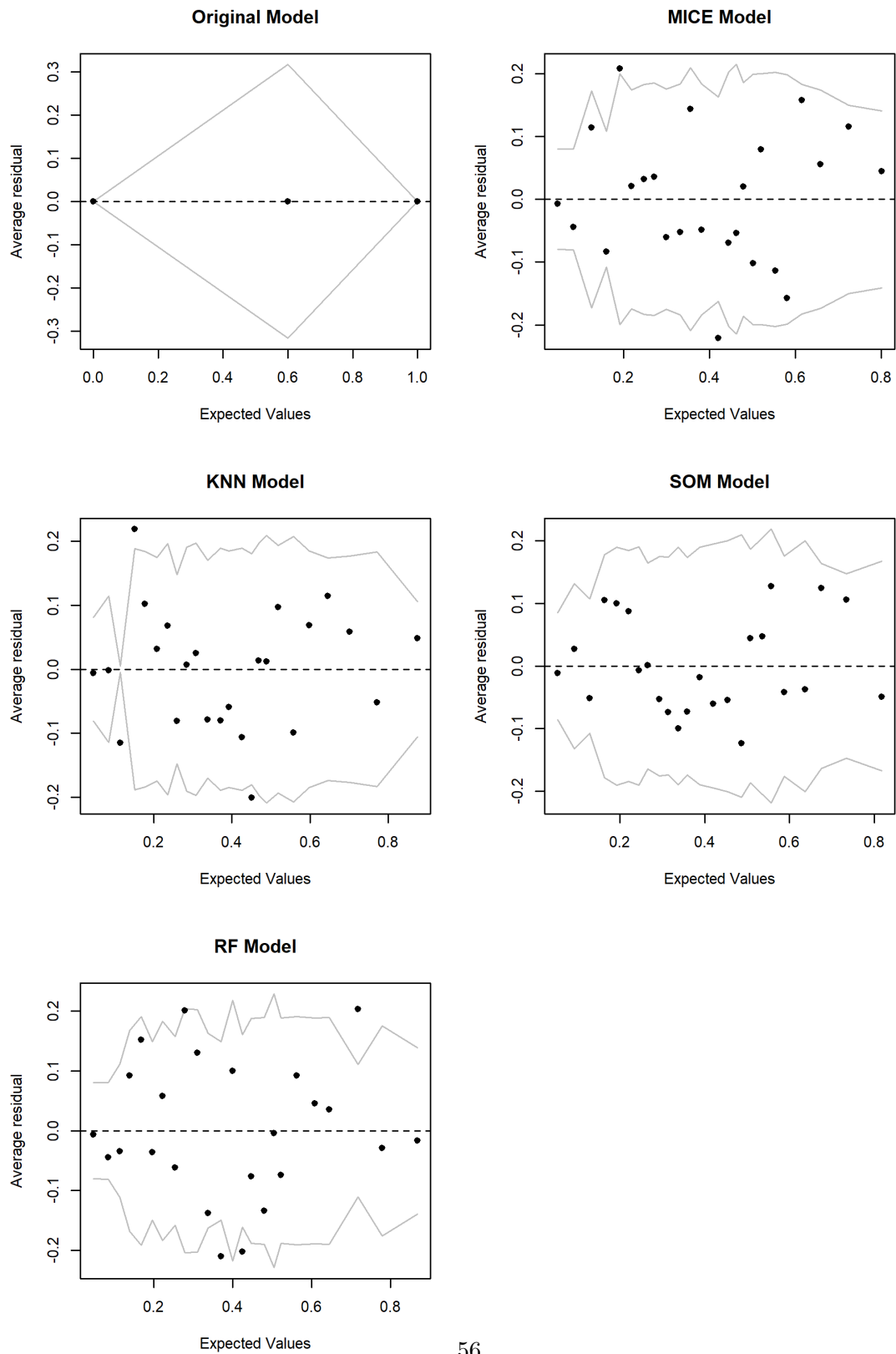


Figure 7: Binned residual plots for Gauteng Models

5.5.2 Rural Western Cape Region

The sensitivity across all imputation methods is high and consistent at 0.997. This suggests that all models can correctly identify true positive cases of risky sexual behaviour. However, the specificity is notably low for all methods, with MICE being the lowest at 0.040. These models struggle to identify the true negative cases of risky sexual behaviour. Regarding the negative predictive value (NPV), the KNN model stands out with a value of 0.992, indicating that it is highly reliable in predicting negative outcomes when the model indicates no risky behaviour. Whereas, the MICE model remains the worst performing on this metric, with an NPV value of 0.5. The AUC values for all models are quite similar, with the RF model achieving the highest AUC of 0.786, followed closely by the other methods (MICE: 0.785, KNN: 0.784, SOM: 0.783). This indicates a good level of predictive accuracy across all models, with the RF model slightly outperforming the others. Overall, while the models demonstrate strong sensitivity and reasonable AUC values, the low specificity and varying NPVs indicate potential difficulties in accurately predicting non-risky sexual behaviour in the Western Cape region.

Table 23: Model Performance Metrics for Multivariate Regression Models in the Western Cape Region.

Performance Metric	MICE	KNN	SOM	RF
Sensitivity	0.997	0.997	0.997	0.997
Specificity	0.040	0.120	0.120	0.120
NPV	0.500	0.992	0.750	0.750
AUC	0.785	0.784	0.783	0.786

In Figure 8, the residuals of the original model are closely clustered around zero, with some variability seen at the extremes of the expected values. The spread suggests the model maintains reasonable accuracy, with minor deviations throughout the range. The MICE model shows residuals concentrated around zero, indicating accurate predictions for most expected values. However, there is a slight increase in variability at higher expected values, which could suggest some bias or overfitting in those areas. The KNN model residuals are tightly grouped around the zero line across the range of expected values. Some slight outliers exist at the extremes. The residuals of the SOM model are largely centered around zero. Similar to the MICE model, there is an increase in spread at higher expected values. The RF model follows a similar trend with increasing variability at the higher expected values indicating potential inaccuracies or uncertainties in those predictions.

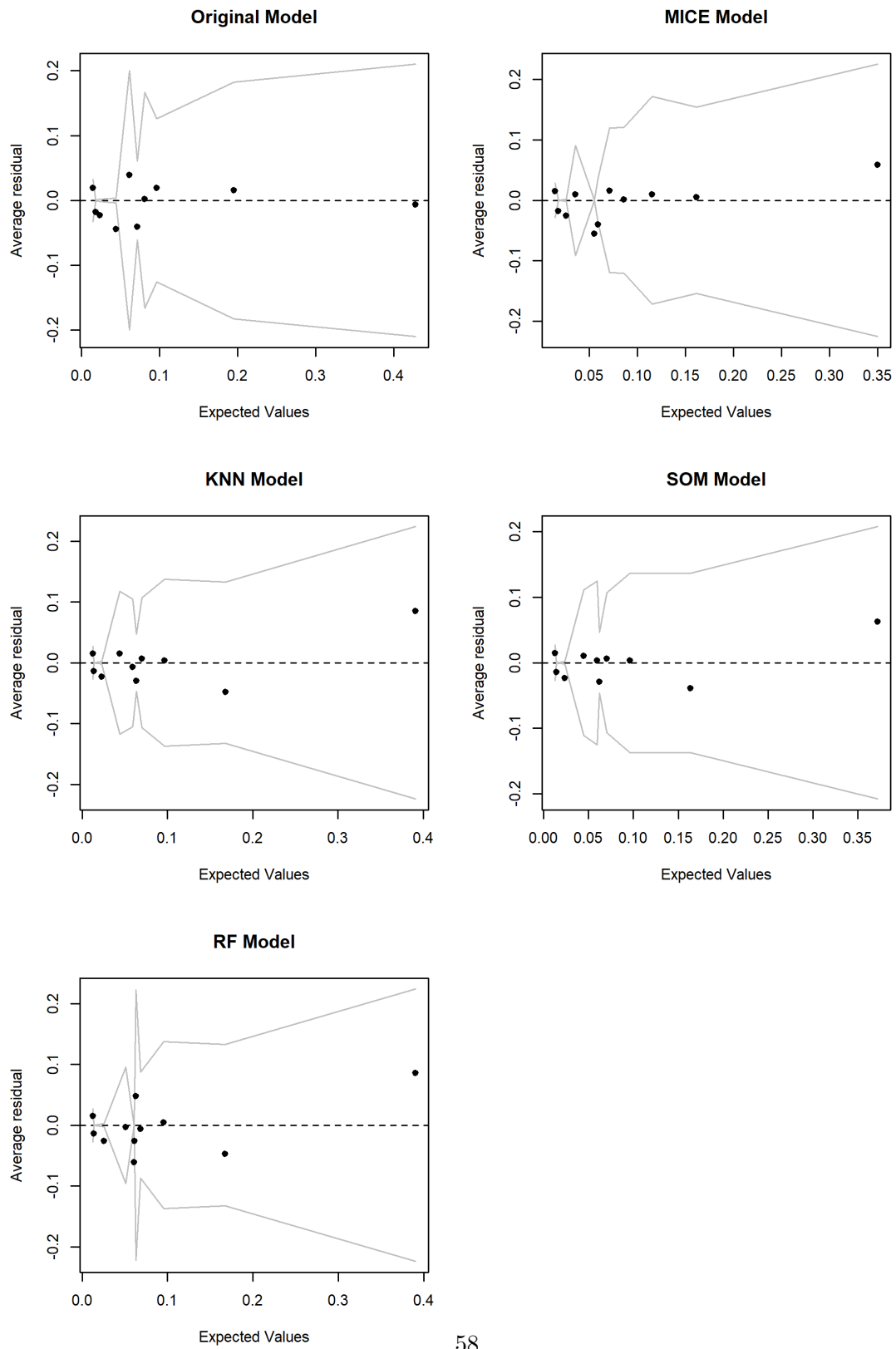


Figure 8: Binned residual plots for Western Cape Models

5.6 Performance of Different Imputation Methods

The model performance results above suggest that the different imputation methods affect the results differently, with grave differences between the traditional imputation method and the machine-learning imputation methods. Because imputation performance differs by proportions of missingness in a variable, this analysis wonders how severe are the differences in how the imputation methods impute missing values when there's high missingness rates compared to low missingness rates. The *Risk Drinking* factor from the Gauteng dataset which has a missingness proportion of 71.9% in the original dataset and the *Current Partner Employed* factor from the same dataset which has a 24.1% missingness proportion are considered herein as representatives of the high and low missingness proportions.

Figures 9 and 10 below shows the proportions of *Risk Drinking* and *Current-partner-employed* values predicted by the different imputation methods compared to existing proportions in the original incomplete data. Figure 9 shows that when there's high missingness rate, all the machine-learning imputation methods are able to preserve the ratio of classes in the *Risk Drinking* factor, with all of them predicting more observations to be non-risk drinkers (class 0) than risk drinkers (class 1). In contrast, the MICE multiple imputation algorithm does not preserve this ratio, as it predicts the two classes more or less equally. Figure 11 in the Appendix shows a similar pattern for the *Children are a sign of a worthy woman* factor with a high missingness rates of 80%. Figure 13 of the Appendix demonstrates this issue even with a smaller missingness rate of 59.9%.

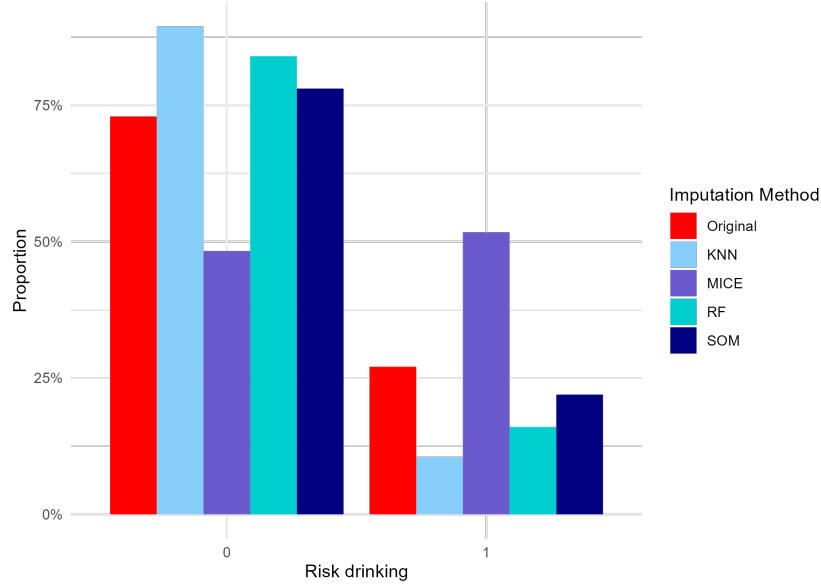


Figure 9: Proportions of risk drinking classes predicted by imputation methods against original data proportions in the Gauteng dataset.

However, Figure 10 shows that when there's a low missingness rate, both the traditional and machine-learning imputation methods preserve the ratio of classes in the original variable. That is, for the *Current-partner-employed* factor, the MICE, KNN, RF, and SOM imputation methods all predict more women to have reported their current partners to be employed (class 1) than not (class 0). Figures 12, , and 14 in the Appendix shows a similar pattern for the *Current contraceptive use*, and *Age of alcohol consent (Western Cape)* factors with low missingness rates of 45.6% and 28.2% respectively.

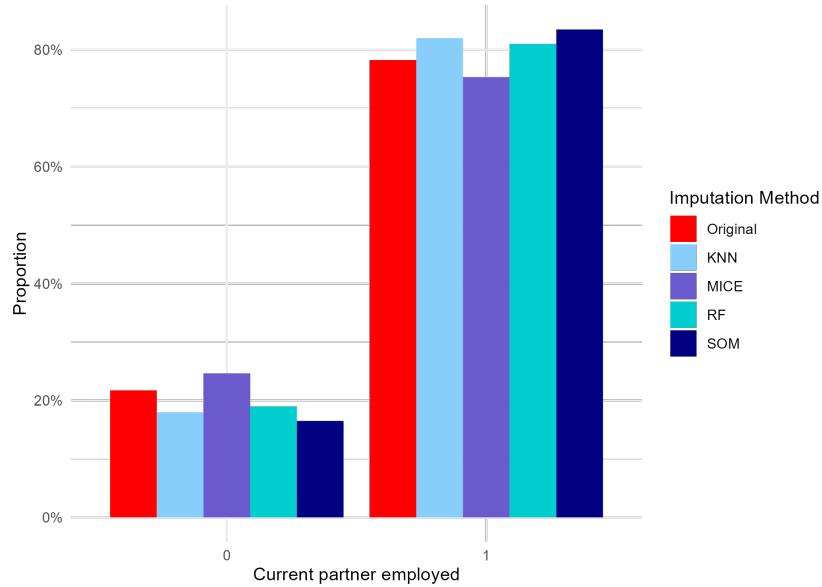


Figure 10: Proportions of employment status classes predicted by imputation methods against original data proportions in the Gauteng dataset.

6 Discussion

This section discusses the results tabled in the preceding section. The discussion is twofold: first, we discuss results on factors associated with risky sexual behaviour (RSB) and secondly, we draw a comparison between the traditional and machine-learning imputation methods.

6.1 Factors Associated with Risky Sexual Behaviour

Our results show that appropriately accounting for missingness in the data has enhanced Khati's study on the factors associated with risky sexual behaviour. By means of imputation we were able to obtain a larger number of significant variables and robust results from our multivariate logistic regressions to better understand the associations between certain factors and RSB in the urban Gauteng and rural Western Cape study areas.

As expected, the women from the two regions differed significantly in terms of their characteristics and the factors associated with risky sexual behaviour. Khati attributes these discrepancies to the difference in social and economic activities between the locations: Gauteng is cosmopolitan and is economic hub of South Africa [27]. In contrast, the rural Western Cape is provincial with truncated networks and limited economic activity. These regions also differed in the occurrence of risky sexual behaviour, as Gauteng had 39.6% occurrence of the outcome variable while the western cape had an occurrence of only 6.1%

For the Gauteng region, the regression models suggest that the occurrence of risky sexual behaviour decreases with age. The SOM and RF regression models found women aged between 25-34 to be less likely to engage in risky sexual behaviour compared to those aged 18-24. The strong significance and small odds ratios across all imputation methods for the age category 35-44 further amplify this storyline. These results are similar to those from a South African cross-sectional study conducted in 2020 [32]; However, they do not completely agree with all international studies which found the highest incidence of STI's, contracted through RSB, to be in the middle age range [6, 39].

The KNN and SOM models suggest that having an employed partner acts as a protective factor against risky sexual behaviour. This is especially true for women living in poverty, as having an employed partner who is able to financially care for the woman means they are not required to perform sexual favours in exchange for money [37].

Across all imputed models, the analysis revealed that women who have used any form of contraception in their lifetime are nearly twice as likely to engage in risky sexual behaviour. Similarly, women who had recently used contraception within three months of the questionnaire were more likely to engage in risky sexual behavior, as indicated by the KNN and RF models. This pattern is also observed among women who consider

it important to use a condom with their spouse or regular partner. These findings are corroborated by Shabnam (2017), who found that women using contraception are more likely to be sexually active, thereby increasing their odds of engaging in risky sexual behavior compared to women who are not sexually active. Additionally, Shirin *et al.* (2014) suggest that this heightened likelihood could be attributed to inconsistent condom usage, which is a significant factor in risky sexual behavior.

Risk drinking was found to greatly increase the odds of a women engaging in risky sexual behaviour. This variable was highly significant across all models. Binge drinking, although to a lesser extent, suggests a similar relationship with risky sexual behaviour. These results were consistent with a previous Gauteng study, thus highlighting the importance of the risk drinking factor [34].

For the Western Cape Region, the Afrikaans language, the employment status of one's current partner, and the perceived importance of condom usage all were all found to be associated with risky sexual behaviour. Afrikaans speaking women in the Western Cape were less likely to engage in risky sexual behaviour compared to women who spoke a native language, thus suggesting a protective factor. Women with employed partners shared similar odds as native Afrikaans speaking women, although to a marginal significance.

The perceived importance of condom usage yielded differing results based on the type of partner involved. Women who considered it important to use condoms with casual partners had a reduced odds of engaging in risky sexual behavior compared to those who did not prioritize condom usage. This finding aligns with studies on relationship power dynamics, which suggest that women with greater agency in negotiating condom use are less likely to engage in risky sexual activities [38]. Conversely, women who emphasized the importance of condom use with their regular partners were approximately five times more likely to engage in risky sexual behavior than those who did not, as consistently shown across all models.

6.2 Comparison between Traditional and Machine-Learning Imputation Techniques

On the other hand, the results suggest that in general, KNN-, SOM- and RF-imputation techniques are able to capture the structure in the data more accurately than multiple imputation, and thus improve results better this traditional approach. This is seen in model comparison metrics shown in tables 22 and 23 for the Gauteng and Western Cape models respectively, where MICE is consistently outperformed by the machine-learning methods across majority of the metrics. In particular, it can be deduced from these results that RF imputation improves the quality of the data better than multiple imputation.

Figures 9 and 10 in the results show that the rate of missing values in a variable significantly influences how the traditional and machine-learning imputation methods impute missing values. In particular, the results suggest that when there's a high missingness rate

in the data, the machine-learning imputation methods are able to capture the underlying structure in the data and thereby preserve this structure. This is seen in the case of the *Risk Drinking* binary variable where the proportion of classes in the variable are preserved by the machine-learning methods compared to multiple imputation which fails to preserve this structure. This is not surprising, as machine-learning techniques have the advantage of being able to capture complex non-linear patterns in the data while multiple imputation, which relies on linear regressions to impute data, cannot capture non-linear patterns. Even so, the preservation of class proportions by the machine-learning methods could be due to bias arising from imbalance of the *Risk Drinking* classes. It could be that the more frequent class of non-risk drinkers (class 0) is favoured over the less frequent class of risky drinkers (class 1), which may explain why these methods retain the proportions existing in the original data. As such, readers must exercise caution.

Among machine-learning imputation techniques, the cross-validation results suggest that overall, at the 10% missingness rate, the KNN-, RF-, and SOM-imputation techniques generally have smaller misclassification errors. Even so, the results initially suggest that RF- and SOM-imputation methods perform similarly compared to KNN imputation. This is consistent with the multivariate logistic regression results in table 20, which shows that overall, models built on RF-imputed and SOM-imputed datasets find similar significant factors affecting risky sexual behaviour among women in Gauteng than those built on KNN-imputed datasets. However, when these factors are studied for Western Cape women, all the three machine-learning imputation methods highlight the same significant factors, which contradicts the assertion that KNN imputation will impute differently from the RF and SOM imputations. When models based on the three machine-learning-imputed datasets are compared, it is clear that RF imputation performs better than KNN and SOM imputation, indicating the robustness of this method among the three.

It is clear, therefore, that machine-learning techniques impute data better than multiple imputation, as suggested by Jerez *et al.* (2010) [25] and Wang *et al.* (2022) [56]. Among the machine-learning methods, our study shows that RF imputation performs slightly better than KNN and SOM imputation methods, which explains why Tang and Ishwaran (2017) [51] find that RF imputation generally preferred as an imputation technique that can handle missing values more accurately. Future research should focus on examining closely the conditions under which the RF, SOM, and KNN imputation techniques are most suitable, and conduct a comparison of these methods to examine.

6.3 Limitations

This study has several limitations that should be acknowledged. Firstly, the focus on specific regions and the inclusion of only women from lower economic standings restricts the generalizability of the findings. By only reporting on women in urban Gauteng and rural Western Cape, the results may not apply to other regions of South Africa or to women of differing economic groups, like those belonging to upper-middle-class.

Consequently, we were unable to capture any broader patterns of risky sexual behaviour across diverse demographics and regions within South Africa.

Secondly, issues related to the study design may affect the validity of the findings. The original data was collected cluster sampling in the urban Gauteng region and stratified cluster random sampling in the rural Western Cape. However, Khati (2012) did not account for the complexities in the data collection process when conducting his univariate, bivariate, and multivariate analyses. This oversight includes the failure to utilise survey weights or implement cluster analysis, which may have led to biased results and would have diminished the robustness of the conclusions drawn from the data.

Additionally, Khati's approach to building the multivariate logistic regression model building was not comprehensive. The lack of a systematic model-building approach means that there are potentially significant factors that may have been excluded from the multivariate logistic regression model due to their non-significance in the univariate sense. These omitted factors could have provided deeper insights into the factors associated with risky sexual behaviour. Additionally, no interactions between factors were utilised in the multivariate model despite the literature suggesting potential interactions between age and alcohol use, amongst other potential interactions. This limitation highlights the necessity of utilising rigorous statistical techniques to ensure that all relevant factors are considered in the analysis.

Overall, these limitations indicate that while the study contributes valuable insights into the factors associated with risky sexual behaviour, caution should be exercised when interpreting these findings and their applicability in a broader South African context.

7 Conclusion

In this paper, we sought to address the missingness problem in Kathi's 2012 study by employing traditional and machine-learning imputation techniques to improve the analysis of factors affecting risky sexual behaviour. Following the imputation, we conducted logistic regression analyses on the complete datasets to evaluate how the various methods influenced the results. The results indicate that machine-learning imputation techniques outperform the traditional multiple imputation technique, as they are able to capture better the underlying non-linear structure in the data compared to multiple imputation. Notably, among the machine-learning methods, random forest imputation performs better than k-nearest neighbours imputation and self-organising-maps imputation.

The logistic regression models uncovered more significant factors associated with risky sexual behaviour when imputation was applied, compared to when the observations with missing data were deleted as done in Khati's study. For risky sexual behaviour among women in urban Gauteng, this paper finds that in addition to Khati's findings that age, language, marital status, and importance of condom usage with spouse/regular partner are significant factors affecting risky sexual behaviour, we also find that in the presence of complete data, whether the women's current partner is employed, whether they're on lifetime contraceptives or currently on contraceptives, and engaging in risky drinking are all significant factors affecting the women's risky sexual behaviour. For women in rural Western Cape, this study finds that not only are Language and importance of condom usage with spouses and casual partners significant factors affecting their risky sexual behaviour (as does Khati's study), but whether a current partner is employed is also a significant factor.

8 References

- [1] Statistics South Africa. *Census 2001: Census in brief*. 3. Statistics South Africa, 2003.
- [2] Godwin Anguzu et al. “Relationship between socioeconomic status and risk of sexually transmitted infections in Uganda: multilevel analysis of a nationally representative survey”. In: *International journal of STD & AIDS* 30.3 (2019), pp. 284–291.
- [3] Sevgi O Aral et al. “Sexually transmitted infections”. In: (2011).
- [4] Kim Ashburn, Deanna Kerrigan, and Michael Sweat. “Micro-credit, women’s groups, control of own money: HIV-related negotiation among partnered Dominican women”. In: *AIDS and Behavior* 12 (2008), pp. 396–403.
- [5] AUDIT Screening. *Check Your Drinking: AUDIT Score*. Accessed: 2024-10-23. 2024. URL: <https://auditscreen.org/check-your-drinking/>.
- [6] Aisha Babi et al. “Prevalence of high-risk human papillomavirus infection among Kazakhstani women attending gynecological outpatient clinics”. In: *International Journal of Infectious Diseases* 109 (2021), pp. 8–16.
- [7] Stan Becker. “Couples and reproductive health: a review of couple studies”. In: *Studies in family planning* (1996), pp. 291–306.
- [8] L. Beretta and A. Santaniello. “Nearest neighbor imputation algorithms: a critical evaluation”. In: *BMC Med Inform Decis Mak* 16.3 (2016), p. 74. DOI: 10.1186/s12911-016-0318-z.
- [9] A. Berhan and Y. Berhan. “A Meta-Analysis on Higher-Risk Sexual Behaviour of Women in 28 Third World Countries”. In: *World Journal of AIDS* 2.2 (2012), pp. 78–88. DOI: 10.4236/wja.2012.22011.
- [10] M Lynne Cooper. “Alcohol use and risky sexual behavior among college students and youth: evaluating the evidence.” In: *Journal of Studies on Alcohol, supplement* 14 (2002), pp. 101–117.
- [11] T. Emmanuel, T. Maupong, D. Mpoeleng, et al. “A survey on missing data in machine learning”. In: *J Big Data* 8 (2021), p. 140. DOI: 10.1186/s40537-021-00516-9.
- [12] Katherine A Fethers et al. “Sexual risk factors and bacterial vaginosis: a systematic review and meta-analysis”. In: *Clinical Infectious Diseases* 47.11 (2008), pp. 1426–1435.
- [13] Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. “Pattern classification with missing data: a review”. In: *Neural Computing and Applications* 19 (2010), pp. 263–282.
- [14] Karen L Graves. “Risky sexual behavior and alcohol use among young adults: Results from a national survey”. In: *American Journal of Health Promotion* 10.1 (1995), pp. 27–36.

- [15] S. Harp, T. Samad, and M. Villano. “Modeling Student Knowledge with Self-Organizing Feature Maps”. In: *IEEE Transactions on Systems, Man and Cybernetics* 25 (1995), pp. 727–737. DOI: 10.1109/21.376487.
- [16] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. 2nd. New York/Berlin/Heidelberg: Springer, 2008.
- [17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Available: <https://books.google.co.za/books?id=tVIjmNS30b8C> [Accessed: June 28, 2024]. New York, USA: Springer, 2009.
- [18] Lori L Heise and Christopher Elias. “Transforming AIDS prevention to meet women’s needs: a focus on developing countries”. In: *Social science & medicine* 40.7 (1995), pp. 931–943.
- [19] M. Hlongwa, K. Peltzer, and K. Hlongwana. “Risky sexual behaviours among women of reproductive age in a high HIV burdened township in KwaZulu-Natal, South Africa”. In: *BMC Infectious Diseases* 20 (2020), p. 563. DOI: 10.1186/s12879-020-05302-1.
- [20] James Honaker, Gary King, and Matthew Blackwell. “Amelia II: A program for missing data”. In: *Journal of statistical software* 45 (2011), pp. 1–47.
- [21] S. Hong and H.S. Lynn. “Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction”. In: *BMC Medical Research Methodology* 20 (2020), p. 199. DOI: 10.1186/s12874-020-01080-1.
- [22] E. Hruschka, E. Hruschka, and N. Ebecken. “Towards Efficient Imputation by Nearest-Neighbors: A Clustering-Based Approach”. In: *Data Mining and Knowledge Discovery Handbook*. Rockette, 2004, pp. 513–525. DOI: 10.1007/978-3-540-30549-1_45.
- [23] C.C. Hsu. “Generalizing self-organizing map for categorical data”. In: *IEEE Transactions on Neural Networks* 7.2 (2006), pp. 294–304. DOI: 10.1109/TNN.2005.863415.
- [24] Pinar Ilkcaracan and Susie Jolly. “Gender and sexuality”. In: *UK: IDS* (2007).
- [25] José M Jerez et al. “Missing data imputation using statistical and machine learning methods in a real breast cancer problem”. In: *Artificial intelligence in medicine* 50.2 (2010), pp. 105–115.
- [26] Per Jonsson and Claes Wohlin. “An evaluation of k-nearest neighbour imputation using likert data”. In: *10th International Symposium on Software Metrics, 2004. Proceedings*. IEEE. 2004, pp. 108–118.
- [27] Makobetsa Khati. “An analysis of alcohol use and possible confounding risk factors for risky sexual behaviour amongst women in the rural Western Cape and urban Gauteng provinces”. In: (2013).

- [28] Susan M Kiene et al. “High rates of unprotected sex occurring among HIV-positive individuals in a daily diary study in South Africa: the role of alcohol use”. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 49.2 (2008), pp. 219–226.
- [29] A. Kowarik and M. Templ. “Imputation with the R Package VIM”. In: *Journal of Statistical Software* 74.7 (2016), pp. 1–16. DOI: 10.18637/jss.v074.i07.
- [30] Amy G Lam and James E Barnhart. “It takes two: The role of partner ethnicity and age characteristics on condom negotiations of heterosexual Chinese and Filipina American college women”. In: *AIDS Education & Prevention* 18.1 (2006), pp. 68–80.
- [31] Roderick JA Little and Donald B Rubin. “Statistical analysis with missing data”. In: *Statistical analysis with missing data* (2002).
- [32] RC Mathebula et al. “Factors associated with repeat genital symptoms among sexually transmitted infection service attendees in South Africa, 2015-2016”. In: *South African Medical Journal* 110.7 (2020), pp. 661–666.
- [33] Alessandra Mattei, Fabrizia Mealli, and Donald B Rubin. “Missing data and imputation methods”. In: *Modern Analysis of Customer Surveys: with Applications using R* (2011), pp. 129–154.
- [34] Neo K Morojele et al. “Alcohol use and sexual behaviour among risky drinkers and bar and shebeen patrons in Gauteng province, South Africa”. In: *Social science & medicine* 62.1 (2006), pp. 217–227.
- [35] Olivia Nankinga, Cyprian Misinde, and Betty Kwagala. “Gender relations, sexual behaviour, and risk of contracting sexually transmitted infections among women in union in Uganda”. In: *BMC public health* 16 (2016), pp. 1–11.
- [36] Ann O’leary et al. “Predictors of safer sex on the college campus: A social cognitive theory analysis”. In: *Journal of American College Health* 40.6 (1992), pp. 254–263.
- [37] Bayla Ostrach and Merrill Singer. “At special risk: Biopolitical vulnerability and HIV/STI syndemics among women”. In: *Health Sociology Review* 21.3 (2012), pp. 258–271.
- [38] Julie Pulerwitz et al. “Relationship power, condom use and HIV risk among women in the USA”. In: *AIDS care* 14.6 (2002), pp. 789–800.
- [39] Mosiur Rahman et al. “Intimate partner violence and symptoms of sexually transmitted infections: are the women from low socio-economic strata in Bangladesh at increased risk”. In: *International journal of behavioral medicine* 21 (2014), pp. 348–357.
- [40] D Rajaraman, S Russell, and J Heymann. “HIV/AIDS, income loss and economic survival in Botswana”. In: *AIDS care* 18.7 (2006), pp. 656–662.
- [41] S. Rejeb, C. Duveau, and T. Rebafka. “Self-organizing maps for exploration of partially observed data and imputation of missing values”. In: *Chemometrics and Intelligent Laboratory Systems* 231 (2022), p. 104653.

- [42] Donald B Rubin. “Inference and missing data”. In: *Biometrika* 63.3 (1976), pp. 581–592.
- [43] Donald B Rubin. “Basic ideas of multiple imputation for nonresponse”. In: *Survey Methodology* 12.1 (1986), pp. 37–47.
- [44] Donald B Rubin. “An overview of multiple imputation”. In: *Proceedings of the survey research methods section of the American statistical association*. Vol. 79. Citeseer. 1988, p. 84.
- [45] Seema Sangari and Herman E Ray. “Evaluation of imputation techniques with varying percentage of missing data”. In: *arXiv preprint arXiv:2109.04227* (2021).
- [46] Shewli Shabnam. “Sexually transmitted infections and spousal violence: The experience of married women in India”. In: *Indian Journal of Gender Studies* 24.1 (2017), pp. 24–46.
- [47] Tahmina Shirin et al. “Prevalence of sexually transmitted diseases and transmission of HIV in Dhaka, Bangladesh”. In: *Bangladesh Journal of Medical Microbiology* 3.1 (2009), pp. 27–33.
- [48] Olive Shisana and Leickness Chisamu Simbayi. *Nelson Mandela/HSRC study of HIV/AIDS: South African national HIV prevalence, behavioural risks and mass media: household survey 2002*. HSRC Press, 2002.
- [49] Petroula Stamataki et al. “Prevalence of HPV infection among Greek women attending a gynecological outpatient clinic”. In: *BMC infectious diseases* 10 (2010), pp. 1–6.
- [50] D.J. Stekhoven and P. Bühlmann. “MissForest—nonparametric missing value imputation for mixed-type data”. In: *Bioinformatics* 28.1 (2012), pp. 112–118. DOI: 10.1093/bioinformatics/btr597.
- [51] F. Tang and H. Ishwaran. “Random Forest Missing Data Algorithms”. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10.6 (2017), pp. 359–447. DOI: 10.1002/sam.11348.
- [52] Olga Troyanskaya et al. “Missing value estimation methods for DNA microarrays”. In: *Bioinformatics* 17.6 (2001), pp. 520–525.
- [53] UNAIDS. *UNAIDS Global AIDS Update 2023*. Available online: <https://www.unaids.org/en/resources/documents/2023/2023-report-global-aids> [Last accessed: October 14, 2024]. 2023.
- [54] Stef Van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate imputation by chained equations in R”. In: *Journal of statistical software* 45 (2011), pp. 1–67.
- [55] W.S. Van Heerden. “Self-Organizing Feature Maps for Exploratory Data Analysis and Data Mining: A Practical Perspective”. PhD thesis. Gauteng, South Africa: University of Pretoria, 2017.

- [56] Huimin Wang et al. “Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example”. In: *BMC Medical Informatics and Decision Making* 22 (2022), pp. 1–14.
- [57] Jeff Webb. *Logistic Regression*. Accessed: 2024-10-23. 2023. URL: <https://bookdown.org/jefftemplewebb/IS-6489/logistic-regression.html>.
- [58] Henry Wechsler et al. “Correlates of college student binge drinking.” In: *American journal of public health* 85.7 (1995), pp. 921–926.
- [59] Ellen Weiss, Daniel Whelan, and Geeta Rao Gupta. “Gender, sexuality and HIV: making a difference in the lives of young women in developing countries”. In: *Sexual and Relationship Therapy* 15.3 (2000), pp. 233–245.
- [60] D. R. Wilson and T. R. Martinez. “Improved Heterogeneous Distance Functions”. In: *Journal of Artificial Intelligence Research* 6 (1997), pp. 1–34.

A Appendices

A.1 Univariate analyses of factors associated with risky sexual behaviour for the urban Gauteng and rural Western Cape regions

Demographic factors

Table 24: The strength of association between demographic variables and risky sexual behaviour (univariate logistic regression analysis)

Variable	Gauteng		Western Cape	
	OR \pm SE (95% CI)	p-value	OR \pm SE (95% CI)	p-value
Age (years)				
18-24	Naturally coded	-	Naturally coded	-
25-34	0.51 \pm 0.20 (0.34-0.75)	< 0.01	1.03 \pm 0.47 (0.42 - 2.70)	0.95
35-44	0.28 \pm 0.22 (0.18-0.43)	< 0.01	0.32 \pm 0.63 (0.08-1.06)	0.07
Education				
Primary	Naturally coded	-	Naturally coded	-
Above Primary	NA	NA	1.16 \pm 0.42 (0.5-2.61)	0.72
Marital Status				
Legally married	Naturally coded	-	Naturally coded	-
Traditionally married	1.75 \pm 0.18	< 0.01	NA	-
First Language				
Native Language	Naturally coded	-	Naturally coded	-
Afrikaans	0.24 \pm 0.29 (0.13-0.41)	< 0.01	0.05 \pm 1.43 (0.00-1.28)	0.04
English	0.53 \pm 0.50 (0.18-1.34)	0.20	0.29 \pm 1.49 (0.01-7.94)	0.40
Racial Classification				
Black/African	Naturally coded	-	Naturally coded	-
Coloured	0.26 \pm 0.33 (0.13-0.49)	< 0.01	0.13 \pm 0.46 (0.05-0.33)	< 0.01
White	0.23 \pm 0.42 (0.09-0.49)	< 0.01	NA	0.99
Asian/Other	NA	-	NA	-

Socio-economic/household hunger variables

Table 25: The strength of association between socio-economic/household hunger and risky sexual behaviour (univariate logistic regression analysis)

Variable	Gauteng		Western Cape	
	OR \pm SE (95% CI)	p-value	OR \pm SE (95% CI)	p-value
Paid work done in the last 12 months	0.79 \pm 0.17 (0.57-1.10)	0.16	1.65 \pm 0.63 (0.48-5.66)	0.43
Currently working	0.65 \pm 0.18 (0.46-0.93)	0.02	1.72 \pm 0.65 (0.48-6.15)	0.40
High SES	0.85 \pm 0.18 (0.60-1.21)	0.37	0.55 \pm 0.48 (0.21-1.40)	0.21
Household Hunger	NA	-	2.45 \pm 0.49 (0.93-6.48)	0.07

Psycho-social variables

Table 26: The strength of association between psycho-social variables and risky sexual behaviour (univariate logistic regression analysis)

Variable	Gauteng		Western Cape	
	OR \pm SE (95% CI)	p-value	OR \pm SE (95% CI)	p-value
Male fertility entitlement	1.06 \pm 0.29 (0.60-1.89)	0.83	1.52 \pm 0.58 (0.49-4.71)	0.47
Childless choice is right	1.54 \pm 0.17 (1.10-2.16)	0.01	1.12 \pm 0.48 (0.43-2.88)	0.82
Children are a sign of a worthy woman	0.97 \pm 0.37 (0.47-2.01)	0.94	0.00 \pm 3134.32	1.00
Children are a sign of a worthy man	0.95 \pm 0.37 (0.47-1.95)	0.89	0.00 \pm 1901.06	0.99

Partner characteristic variables

Table 27: The strength of association between partner characteristics and risky sexual behaviour (univariate logistic regression analysis)

Variable	Gauteng		Western Cape	
	OR \pm SE (95% CI)	p-value	OR \pm SE (95% CI)	p-value
Older partner (≥ 30 yrs)	0.35 \pm 0.22 (0.23-0.55)	<0.001	Dropped	-
Currently employed	0.62 \pm 0.23 (0.39-0.96)	0.03	0.24 \pm 0.69 (0.06-0.92)	0.04
Satisfied with partner	1.10 \pm 0.33 (0.58-2.09)	0.77	0.28 \pm 0.68 (0.07-1.07)	0.06
Serious disagreements	0.95 \pm 0.19 (0.66-1.38)	0.80	0.81 \pm 0.46 (0.33-1.99)	0.65

Community/social support variables

Table 28: The strength of association between community/social support variables and risky sexual behaviour (univariate logistic regression analysis)

Variable	Gauteng		Western Cape	
	OR \pm SE (95% CI)	p-value	OR \pm SE (95% CI)	p-value
Availability of recreational facilities	0.94 \pm 0.18 (0.67-1.33)	0.72	0.71 \pm 0.49 (0.27-1.85)	0.48
Easy to use recreational facilities	1.00 \pm 0.18 (0.71-1.42)	0.99	1.01 \pm 0.52 (0.37-2.77)	0.99
Easy to buy alcohol	0.67 \pm 0.21 (0.44-1.02)	0.06	0.84 \pm 0.46 (0.34-2.05)	0.69
Heavy drinking in the community	1.11 \pm 0.22 (0.72-1.72)	0.62	2.04 \pm 0.63 (0.59-7.03)	0.26
Community accepts alcohol abuse	0.87 \pm 0.17 (0.63-1.22)	0.42		
Helpful neighbours	0.86 \pm 0.17 (0.62-1.20)	0.37	0.86 \pm 0.48 (0.33-2.22)	0.75

Substance use variables

Table 29: The strength of association between substance use variables and risky sexual behaviour (univariate logistic regression analysis)

Variable	Gauteng		Western Cape	
	OR \pm SE (95% CI)	p-value	OR \pm SE (95% CI)	p-value
Lifetime alcohol use	1.13 \pm 0.17 (0.81-1.57)	0.48	0.98 \pm 0.46 (0.40-2.42)	0.97
Current alcohol use	1.07 \pm 0.19 (0.74-1.55)	0.73	Dropped	-
Binge drinking	1.60 \pm 0.23 (1.01-2.52)	0.04	1.35 \pm 0.80 (0.28-6.44)	0.70
Risk drinking	4.80 \pm 0.37 (2.31-9.98)	<0.01	1.24 \pm 0.69 (0.32-4.78)	0.75
Alcohol onset age: minor	0.71 \pm 0.29 (0.40-1.26)	0.24	1.10 \pm 0.49 (0.42-2.87)	0.84
Lifetime cigarette use	0.81 \pm 0.22 (0.53-1.24)	0.33	0.45 \pm 0.42 (0.20-1.01)	0.05

General health, contraceptive use and pregnancy Sex related variables

Table 30: The strength of association between general health, contraceptive/pregnancy variables and risky sexual behaviour (univariate logistic regression analysis)

Variable	Gauteng		Western Cape	
	OR \pm SE (95% CI)	p-value	OR \pm SE (95% CI)	p-value
Lifetime contraceptive use	1.89 \pm 0.17 (1.34-2.66)	<0.01	1.35 \pm 0.42 (0.59-3.08)	0.48
Current contraceptive use	1.97 \pm 0.24 (1.23-3.15)	<0.01	0.83 \pm 0.67 (0.22-3.10)	0.78
Effective contraceptive use	0.79 \pm 0.37 (0.38-1.61)	0.51	0.47 \pm 0.69 (0.12-1.81)	0.27
Have child(ren)	0.60 \pm 0.21 (0.40-0.91)	0.01	0.35 \pm 0.58 (0.11-1.10)	0.07
Lifetime miscarriage	1.07 \pm 0.22 (0.70-1.65)	0.75	0.76 \pm 0.56 (0.25-2.29)	0.63

Table 31: The strength of association between sex related variables and risky sexual behaviour (univariate logistic regression analysis)

Variable	Gauteng		Western Cape	
	OR \pm SE (95% CI)	p-value	OR \pm SE (95% CI)	p-value
Husband/boyfriend most recent sex partner	0.75 \pm 0.40 (0.34-1.62)	0.46	0.32 \pm 0.54 (0.11-0.93)	0.04
Importance of condom use with spouse/regular partner	2.49 \pm 0.21 (1.65-3.76)	<0.01	5.49 \pm 0.46 (2.23-13.50)	<0.01
Importance of condom use with casual partner	1.16 \pm 0.38 (0.55-2.41)	0.70	0.39 \pm 0.58 (0.13-1.23)	0.11

A.2 Performance of different imputation methods across factors with different proportions of missing observations

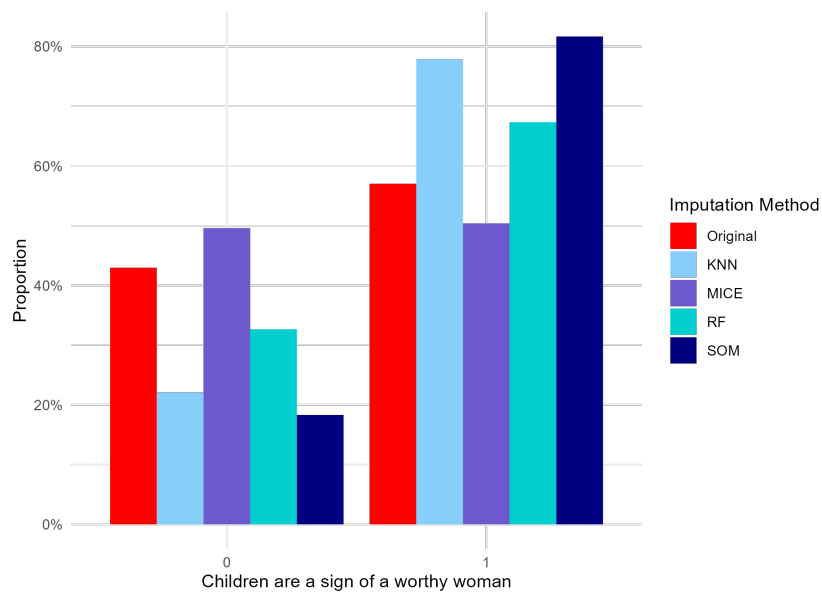


Figure 11: Comparison of proportions for the psycho-social factor "Children are a sign of a worthy woman" classes predicted by imputation methods against original data proportions in the urban Gauteng region, which had 80% missing observations

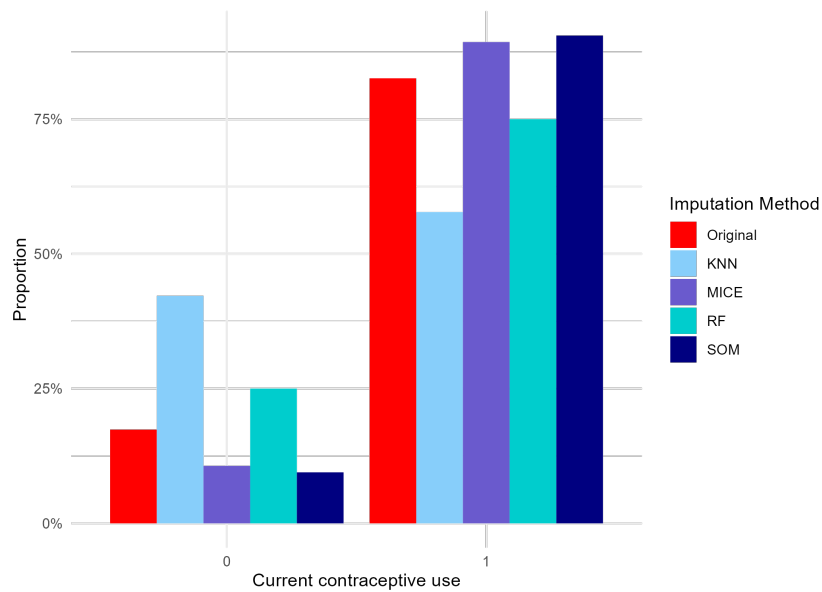


Figure 12: Comparison of proportions for the current contraceptive use classes predicted by imputation methods against original data proportions in the rural Western Cape region, which had 45.6% missing observations

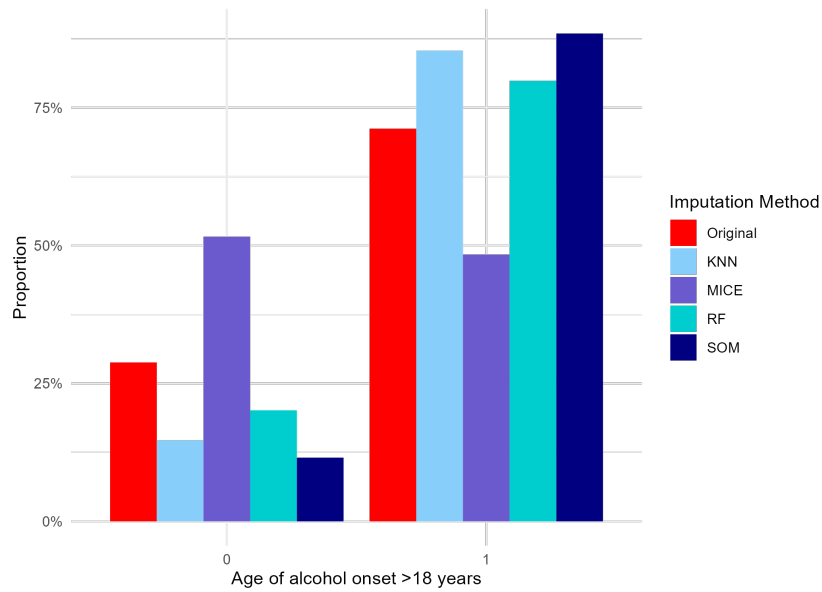


Figure 13: Comparison of proportions for the age of alcohol onset classes predicted by imputation methods against original data proportions in the urban Gauteng region, which had 59.9% missing observations

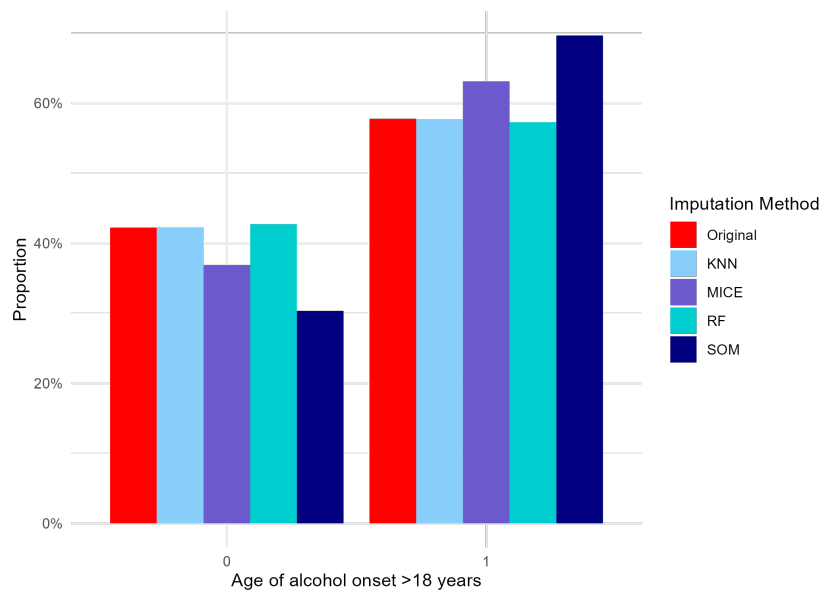


Figure 14: Comparison of proportions for the age of alcohol onset classes predicted by imputation methods against original data proportions in the rural Western Cape region, which had 28.2% missing observations.