ADVANCING STATISTICAL MODELS FOR ADDRESSING MISSING DATA IN HIV/AIDS SURVEYS: A COMPREHENSIVE APPROACH

by

Chinonso Franklin Okoroafor

Thesis submitted to University of Plymouth
in partial fulfilment of the requirements for the degree of

MSc Data Science and Business Analytics

University of Plymouth
Faculty of Science & Engineering

Supervised by
Dr. Malgorzata Wojtys

September 2024

Copyright statement

Abstract

The development of HIV therapy, which has shown promise in postponing the onset of AIDS and mortality, depends on early HIV testing. According to current guidelines, antiretroviral medication should be started before AIDS symptom’s manifest. Missing values are a prevalent issue with data quality in HIV databases. When this frequent problem arises during the data integration process, low accuracy and high bias are usually the outcome. Three categories exist for the missing data: missing at random (MAR), missing completely at random (MCAR), and not missing at random (NMAR). Different data imputation algorithms that are either unrelated to the ML model, not designed for HIV-EHR data, or that only utilize univariate information on the observed characteristics are some of the ways that have been explored to address this issue. We compare the imputation techniques used in machine learning and statistics for this research. The Heckman model, random forest, decision tree, k-means clustering, and simple imputation were used to assess the models. The models' results show that the random forest model is superior, with scores in accuracy, sensitivity, specificity, and precision of 96.70%, 96.68%, 97.22%, and 99.69%, respectively, while the Heckman model fared the lowest.

Word Count: 7951

Table of Contents

List of Tables

List of Figures

1 Introduction

Early human immunodeficiency virus (HIV) testing is necessary for the development of HIV treatment which has demonstrated promise in delaying the onset of AIDS and death. Current standards recommend initiating antiretroviral therapy before the onset of AIDS. However, the cost of medical treatment for individuals living with HIV is significantly higher when diagnosed when they meet the clinical criteria for AIDS than it is for those who are not. Despite these established benefits, late HIV diagnoses still occur (Schwarcz *et al.*, 2006).

A personal healthcare records (PHR) is an electronic health record that the patient, not their healthcare practitioner, controls. Analyzing with low data quality in PHR datasets leads to diminished accuracy and high bias in the data integration process. The data collection process is impacted by the particular situation for acquisition, respondents' interests, the device's status and the environment (Kim and Chung, 2020). Analyzing and interpreting health data with missing data is a major concern. This creates an avenue for data validity and credibility of research outcomes to be questioned, therefore, voiding an entire study (Agbo *et al.*, 2023).

The data collection process makes missing data unavoidable. This is attributed to faulty equipment, human errors, respondents' refusal to answer etc. Although most clinical trials will have some missing data, it can be minimized through careful study design, strict implementation, and meticulous data collection. Nevertheless, certain assumptions about the missing information must be made during analysis (Alabadla *et al.*, 2022). A great deal of complicated data regarding patients, hospital resources, sickness diagnoses, electronic patient records, and medical equipment are gathered by the healthcare organization. The medical device sector can be improved, diseases can be predicted and diagnosed, treatment efficacy can be evaluated, healthcare can be managed to minimize time and money waste and poor treatment decisions which can be fatal for patients (Ramadhan *et al.*, 2024).

Three categories exist for missing data: missing at random (MAR), missing completely at random (MCAR), and not missing at random (NMAR). When it comes to MCAR, missing data happens at random and has no bearing on the results because it is unrelated to either seen or unseen data. The MAR pattern identifies a relationship between missing data and detected data, suggesting that the missing value may be inferred from the available data. On the other hand, with NMAR, unspotted data is associated with missing values. Additionally, the missing data can be divided into missing values from single and multiple imputations. A single, unique piece of data, such as the mean, median, or median of a given dataset, is used to replace the incomplete data in the single imputation missing values technique. Using a combination of machine learning and statistical methods, the multiple imputation value technique replaces missing data with numerous values (Al-Jamali *et al.*, 2023).

Many strategies have been proposed to address missing data problems. Removing missing records is one method used that reduces the number of samples and may result in the loss of important information. Utilizing mean, mode, or hot-deck imputation to impute missing data disregards the relationships between the attributes. When more than 5% of data is missing, model-based strategies like the expectation-maximization (EM) algorithm have demonstrated better for imputing more accurate missing data. Nonetheless, a significant limitation of the EM method is its inability to automatically generate standard errors (Khan *et al.*, 2022). Machine learning (ML) can analyze data, recognize patterns among data attributes, and obtain valuable insight from datasets. Because machine learning models are data-dependent, they demand a substantial and balanced amount of training inputs. This is particularly challenging as disease samples are noticeably less represented than healthy ones (Sultan *et al.*, 2023).

The proposed research study is a comparative analysis of different imputation models for handling missing data. This research is important as analysis of individual features significantly impacts the works of domain expert and ML engineers who have to design healthcare assistants and guidelines. We implement and compare the performance of several machine models to predict the most accurate HIV/AIDS missing data. Hyper-parameter tuning is implemented to assess its impact on both imbalanced and balanced datasets.

1.1 Research Question

Personal Health records (PHRs) differ in type and scope and are being managed by individuals. Because they are obtained under different circumstances, some data may be missing which adversely impacts the result of the data analysis. Missing data needs be replaced with appropriate values. However, basic imputation methods no longer meet the require standard which begs the question "how effective are machine learning techniques in handling missing data?"

1.2 Research Aims and Objectives

The primary aim of this study is to improve the quality and reliability of HIV/AIDS survey data by employing advanced machine learning techniques, including the Heckman Selection Model, Random Forest, Decision Tree, and K-Means Clustering, to effectively manage and reduce the impact of missing data. A thorough comparison is conducted against traditional statistical imputation methods. Through this investigation, the research seeks to provide more accurate and representative data to the global health community, ultimately enhancing the effectiveness of HIV prevention, treatment, and care programs. The following objectives have been set to achieve the main goal of this research:

1.2.1 Evaluate Prevalence and Characteristics of Incomplete Data in HIV/AIDS Research

The first step in this research project will involve a thorough examination of HIV/AIDS survey data from the Population-based HIV Impact Assessment (PHIA) Project in Zimbabwe. It involves seeking out the patterns, regularities, and underlying causes of missing information. It hinges on identifying and evaluating the many forms of missing data, including missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR), and how they affect statistical analysis and interpretations.

1.2.2 Apply Sample Selection Correction Techniques

Following a comprehensive analysis of the datasets, any biases in sample selection will be found and corrected in the data using the a basic statistical, supervised and unsupervised machine learning models.

1.2.3 Build and Validate Sophisticated Models for Data Imputation

The goal is to create sophisticated data imputation models and strategies that can handle different kinds of missing data by utilising the latest developments in machine learning and statistics along with Python and R Studio. To ensure that these models are dependable and efficient at imputing missing data, they undergo a rigorous validation process.

1.2.4 Assess the Performance of Innovative Models Versus Conventional Approaches

Quantifying the effectiveness of the implemented models for handling missing data is imperative. The objective of this analysis is to offer verifiable proof of the increases in accuracy and dependability that sophisticated machine learning models bring to the analysis of HIV/AIDS survey data.

1.2.5 Formulate and Distribute Guidelines for Managing Missing Data in HIV/AIDS Studies

Based on the findings and advancements from this project, this research aims to create and disseminate detailed guidelines and best practices for addressing missing data in HIV/AIDS research. These resources may help to set a standard for data handling, enhancing the quality of research and its impact on public health policies and programs.

1.3 Report Structure

Outlined is the structure of this report. An examination of missing data is presented, encompassing an overview of the fundamental concepts, the various types of missing data, the various methods employed in missing data handling, and a critical review of state-of-the art techniques is given in Chapter 2. Chapter 3 explains the research methodology applied in this research, demonstrates the data collection process and gives a comprehensive report of the models' implementations and assessments. Chapter 4 presents a clear description of the research results. Chapter 5 interprets and explains the results. Finally, Chapter 6 concludes the research and suggests recommendations for further exploration within this domain.

2 Review

Many study disciplines, including medical research, data reduction, and phrase creation, have focused attention on the ubiquitous problem of missing values in datasets. Modern machine learning methods have been developed as a result of these cooperative efforts to address this problem in datasets with distinctive features. The categorization of missing data mechanisms has significance as it facilitates in the appropriate strategy selection for addressing various missing data issues. There are three significant factors that account for missing data.

2.1 Classification of Missing Data

2.1.1 Missing Completely at Random

A missing value in a data collection is called MCAR if it has no relationship to any other value. Survey questions or any other event, visible or not, are often not linked to data that are MCAR. Let's say, for illustration purposes, that one question is related to income and is denoted by the letter X1, while another question is related to profession and is denoted as X2. Not X1 (income) or X2 (occupation) is the cause of the missing value in MCAR; that is, the survey question is not the origin of the missing value, nor is it another confounder. When MCAR is suspected, the Little's Test of Missingness can be used to ascertain whether the missing data satisfy MCAR requirements. We presume that there is a pattern to the missing data (not MCAR) and reject the null hypothesis based on a significant p-value finding (Mirzaei *et al.*, 2022).

2.1.2 Missing at Random

When respondents fail to answer questions due to an underlying or confusing reason, for example, data that are MAR are absent because of another observable event. For a variety of reasons, certain groups could choose not to answer questions. People who have high-paying occupations, for example, might not be motivated to respond to queries about finances. Theoretically and philosophically, this is accurate because studies show that those with greater incomes are more inclined to not answer queries about their income. Using the MCAR example above, where X1 represents income and X2 represents occupation. Occupation (X2): People in better paid jobs are less likely to respond, which is one of the reasons why X1 (income) may not be reported. Therefore, in the instance of MAR, X2, a different variable, provides the basis for X1's missing answer (Al-Jamali *et al.*, 2023).

2.1.3 Missing Not at Random

Non-ignorable missingness (MNAR) data are those that don't fit into the MCAR or MAR categories. To determine if data are MNAR, subjective analysis is necessary, in contrast to MCAR and the use of an objective statistical test. In MAR, there could not be a clear explanation for data missing, but rather a link between an observed behavior and the reason. On the other hand, MNAR data can be directly linked to an unobservable component that influences the reason why the data values are absent. This might be due to underlying presumptions or the question itself not providing the required information. Assume, for the sake of another example, that X1 is a question about depression and X2 is gender in a survey on general health. Men are less inclined than women to discuss depression, therefore X1 (depression) may not receive as much attention as it should depending on X2 (gender). This would be a MAR case. Conversely, if the individual's null reaction is being caused by their level of sadness (X1), then the missingness is MNAR. Here, the phenomena that the item itself is evaluating—in this example, X1—is the reason for the missingness (Lang and Little, 2018).

2.2 Methods for Handling Missing Data

2.2.1 Complete Case Analysis

After eliminating all missing values, complete case analysis, also known as "listwise deletion," analyzes only the data of the variables seen at each time point. This approach is achieved by omitting the cases with missing data and analyzing the remnant data. While the arguments for simplicity of analysis can be made and advantageous for unbiased analysis when the data is MCAR, reduced sample size and lower statistical power are arguments that can be made against it. This is due to the fact that it becomes challenging to make statistical conclusions during processing, particularly when patients who have missing data differ consistently from those who have complete data (i.e., data that is not MCAR). It is the most often used technique for handling missing values in statistical analysis software like SAS and SPSS. (Wells *et al.*, 2013; Kwak and Kim, 2017).

2.2.2 Available Case Analysis

While sample sizes vary between variables, this approach allows for a greater sample size than whole case analysis since it leverages accessible data for each study. Power is increased by imputed missing data under the premise of Missing Completely at Random (MCAR) without changing point estimations from full case analysis. A model that incorporates missing data is required to provide unbiased estimates for data that is Not Missing at Random (NMAR). If enough variables are supplied, imputations employing Electronic Health Record (EHR) data often yield correct findings since imputations in most cases assume Missing at Random (MAR). To improve imputation accuracy, factors such as illness severity and health care usage might be included. The inclusion of follow-up time and outcome information further improves the accuracy of results obtained through multiple imputations (Mirzaei *et al.*, 2022).

2.2.3 Imputation Method

Imputation is a technique used to address missing data by replacing incomplete entries with estimated values, allowing the dataset to be fully analyzed. Imputation can be performed through two primary approaches: explicit and implicit modeling. Explicit modeling involves generating imputed values by estimating the parameters of a predictive distribution, based on the assumption that the variables follow a specific distribution. This approach commonly employs probability-based methods such as regression, mean, and median imputation. In contrast, implicit modeling focuses on deriving assumed values through algorithms without explicitly defining a distribution. Common implicit techniques include substitution, cold-deck imputation, and hot-deck imputation. Both explicit and implicit methods are often used together to enhance the accuracy of the imputation process.

A widely used technique for handling missing data is Multiple Imputation using Chained Equations (MICE). MICE is highly effective because it can manage different types of variables, such as continuous and categorical data, by imputing a distinct regression model for each variable. The method involves iteratively regressing each variable with missing data on all other variables in the dataset until the missing values are replaced with estimated values. This process is repeated multiple times, producing several complete datasets. The procedure concludes when stable estimates are obtained across these iterations. Due to its versatility and effectiveness, MICE is frequently used to address missing data, even when more specialized alternatives may be available (Psychogyios et al., 2023).

2.3 Related Works

To address the challenge of imputing missing values using non-linear and multivariate information across patients, Bernardini et al. (2023) introduced a machine learning-based imputation method known as Clinical Conditional Generative Adversarial Networks (ccGAN). This model employs a discriminator to distinguish between real and imputed data and a generator to predict missing data. Both networks were trained through a minimax game, wherein the discriminator progressively enhances its ability to detect synthetic data, and the generator improves its accuracy in imputing missing values. To stabilize the training process, particularly for data missing completely at random (MCAR), a "masked reconstruction loss" term was incorporated to ensure statistically consistent and accurate imputations. However, the study has limitations, including its untested generalizability in more complex scenarios and the omission of certain EHR variables relevant to patient care.

Figueroa-García, Neruda, and Hernandez-Pérez (2023) introduced MIGA to address the multivariate and multiple missing data problem. MIGA integrates various objectives by applying dimensionless transformations, creating a single fitness function based on the Minkowski distance of the means, variances, covariances, and skewness between the available and completed datasets. The model was evaluated using a continuous/discrete dataset and compared with the EM algorithm and auxiliary regressions on seven benchmark datasets. MIGA demonstrated superior performance in datasets with complex structures, particularly those involving integer or binary variables, where traditional imputation techniques struggled. However, it performed less effectively on datasets with stronger correlations.

Saroja and Kalpana (2023) proposed a Deep Learning Tensor (DLT) model for early identification of chronic diseases, using tensor factorizations for continuous data and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) for imputing discrete missing values. The most significant predictive variables were first identified using the Adaptive Weighted Differential Biogeography-based Optimization Algorithm (AWDBOA), and a

Normalized Weighted Convolutional Neural Network (NWCNN) classifier was then applied for diagnosis. The approach was demonstrated using two datasets: one from the UCI CKD repository, preprocessed with tensor factorization and ANFIS, and the other analyzing gene expression patterns associated with CKD. The DLT model outperformed previous techniques in terms of sensitivity, specificity, precision, recall, F1-score, and accuracy.

Xu et al. (2020) developed a restricted Boltzmann machine (RBM) within a deep autoencoder framework, employing a novel loss function for error calculation and regularization. This model accounted for temporal patterns prevalent in patient records and incorporated critical relationships within patient data, along with patterns of missingness. When compared with MICE, KNN, and deep autoencoder (DAE) methodologies, their approach demonstrated superior performance, achieving imputation errors as low as 15.5%.

Psychogyios et al. (2023) introduced an innovative approach to pre-imputation missing value handling that combined k-nearest neighbors (kNN) with denoising autoencoders (DAE). The kNN algorithm was reapplied to the missing data after 10 epochs, and the training approach was optimized by adjusting the value of k at each iteration to enhance the accuracy of the imputation. Additionally, the authors revised a generative adversarial network (GAN)-based method for missing data imputation, presenting improvements in both architecture and training processes.

In a related study, Yoon, Jordon, and van der Schaar (2018) developed the Generative Adversarial Imputation Network (GAIN) by modifying the original GAN architecture. Their results showed that GAIN outperforms traditional robust imputation methods, including autoencoder-based approaches.

Finally, Gupta et al. (2021) evaluated machine learning algorithms for heart attack prediction using the Heart dataset from the UCI Machine Learning Repository and the Framingham Heart Study dataset. For missing data imputation, they employed mean or median values, favoring the median for features with skewed distributions. For categorical data, they created an additional "missing" category. Other imputation techniques applied in their analysis included kNN and Multilayer Perceptron (MLP).

3 Methodology/Procedure

The use of machine learning techniques to impute missing data in electronic health records (EHR) has been extensively validated by numerous studies (Bernardini et al., 2023; Psychogyios et al., 2023). This process involves approximating missing data points through sophisticated algorithms. For this project, several key steps are required, including data collection, data cleaning, pattern analysis, imputing missing values using various machine learning methods, and model evaluation. It is essential that data collection is conducted using a reliable and approved source to ensure the integrity of the information. This measure is crucial for maintaining the accuracy of the dataset and preventing any potential biases or inconsistencies that could undermine the validity of the study's findings.

3.1 Data Collection

To assess the effectiveness of Zimbabwe's HIV response, the Zimbabwe Population-based HIV Impact Assessment (ZIMPHIA 2020) was conducted among adults, defined as individuals aged 15 years and older, between November 2019 and March 2020. ZIMPHIA 2020 aimed to evaluate the uptake of HIV care and treatment services, while also providing participants with HIV counseling and testing, with results shared directly with those involved. Of the 11,707 households that met the inclusion criteria, 89.1% successfully completed the household interviews.

During the survey process, 22,751 individuals (13,290 women and 9,461 men) were eligible for HIV testing. Of these, 19,535 individuals (11,871 women and 7,664 men) underwent HIV testing. Adult respondents comprised 76.5% of the total participants, with 79.6% of women and 72.2% of men taking part.

The dataset used in this analysis was sourced from the Population-based HIV Impact Assessment (PHIA) project, which conducts surveys to capture the state of the HIV epidemic in the countries most affected by the disease. We applied for and were granted access to this dataset to support our analysis. This comprehensive data collection serves as a critical tool for evaluating national HIV responses and improving public health outcomes.

3.2 Data Cleaning

Following the data collection process, the next critical step is data cleaning. This involves using logical assumptions to ensure data consistency and addressing missing values, ultimately refining the data frame for analysis. As part of this procedure, certain imputed values in specific columns are adjusted based on logical relationships present in other columns. For instance, it can be logically assumed that men cannot become pregnant, and thus the "everpregnant" category for male records is updated to "Not Applicable." This modification helps maintain the accuracy and integrity of the dataset.

Additionally, the process includes tracking a designated missing_label variable to identify and retrieve columns containing a specific value, such as "Missing Data." The function iterates through each column of the data frame, excluding NA values, and searches for entries matching the pre-assigned missing_label. The outcome is a logical vector indicating which columns contain the specified value. The function then extracts and returns the column names that meet this condition, facilitating further handling of missing or mislabeled data points.

3.3 Exploratory Data Analysis

This stage counts and prints the frequency of each unique value for each variable in the dataset by going over each column iteratively. This makes it easier to comprehend the structure of the dataset by indicating how frequently certain categories occur inside each non-numeric variable. After that, the function separates the dataset according to gender and produces charts for each column, as well as distinct summaries for each subset that is gender-specific. The current column is used to categorize the data, and the count and proportion of each category are determined in relation to the overall number of rows. For columns that are not gender specific, a bar chart representing the general distribution is shown; for columns that are gender neutral, a bar chart representing the gender distribution is constructed. Following the use of the suggested procedures, column names with missing data points are tracked, which is then used to explore and display rectified datasets.

3.4 Modelling

3.4.1 Basic Statistical Imputation

During this step, the function that handles missing data in a data frame is called by applying imputation approaches based on fundamental statistical measurements such as, the mean, median and mode. A data frame (df) is taken as input, loops through each column and looks for values that tally with the preset missing label. In order to generate a subset of the new data, it filters away rows for columns that have missing values that contain the value kept in the missing label's variable or "Not Applicable." After that, the function determines the column's mean, median, and mode statistics. Using an imputation value, the function fills in the missing values in a particular column of a data frame. The imputation value, the column name (col), and the data frame (data) are the three factors required. In the designated column, the function substitutes the given imputation value for each instance of the value kept in the missing label's variable, updating the data frame appropriate. It evaluates the existence of outliers in numerical columns. In the event that outliers are discovered, the mean is used for imputation; otherwise, the median is used. It uses the mode for imputation in factor or character columns. Once the missing values have been imputed, it returns the new data frame.

3.4.2 K-Means Cluster

Each column in the dataframe is iterated over in order to use k-means clustering to manage missing data. If any missing values (NA) are found in the current column, the function handles such mistakes by going through a sequence of stages. Initially, the dataframe is divided into two subsets: rows in the target column that include no missing values (training data) and rows in the target column that have missing values (testing data). Depending on how many levels there are in the target column, the function calculates the number of clusters (k). Following the transformation of the training data, k-means clustering is done to create k clusters. The training data is supplemented by the group assignments that are produced. By changing its category characteristics to numerical values, the testing data is ready for prediction. Each cluster centroid's Euclidean distance is computed for every row in the testing data. By using the most prevalent value from the training data in that cluster to impute the missing value, the row is allocated to the nearest cluster. The next column is reached once the cluster assignments are eliminated and the training and imputed testing data are merged (Patil, Joshi and Toshniwal, 2010).

The Euclidean distance between many cluster centers and a particular data point may be determined using the Euclidean distance function. Centers, a matrix of cluster centroids, and x, the data point (usually a row from the testing data), are the two parameters it takes. The Euclidean distance for each centroid is obtained by taking the square root of the sum of the squared differences between the coordinates of the centroid and the data point. Finding the nearest cluster in the main function is made easier by the vector return of the distances.

3.4.3 Random Forest

To deal with missing data in the data frame, the random forest is employed. Iteratively reviewing each column in the data frame, the function changes each one to character type. The model substitutes (NA) for any entry that corresponds to the missing label and then returns the columns to factors. The process locates rows that include missing values and divides the data frame into training and testing sets, which consist of rows without missing values in the target column and rows with missing values. Next, the training and test datasets divide the predictor variables from the target column (col). The model is trained using the target column and predictor variables from the training data with 500 trees. The model predicts the missing values in the test data after it has been trained. The original missing values from the data frames are filled using these projections (Saroja and Kalpana, 2023).

3.4.4 Decision Tree

A decision tree function is employed to address missing data in the data frame. The data frame's columns are all changed to character types and fields that have missing labels are recorded as NA are then changed back to factors, and so on. With the help of this process, the data is standardized, and missing values are reliably represented as NA. To search for any missing values, the function loops over each column in the data frame. In columns where values are absent, the function finds the rows that have NA values and divides the data into training and test sets, which contain rows with and without missing values in the target column, respectively. Further, all other columns are assigned as predictors and the target column as the response variable. Based on the data type of the target variable—"anova" for numeric data and "class" for categorical data—it chooses the best approach for the decision tree. After that, the decision tree is trained on the training set of data in order to predict the values that are absent from the testing set. The "anova" approach is used to get predictions for numerical columns, whereas the "class" method is used to get predictions for categorical columns. The original data frame's missing items are then filled up using the predicted values (Beaulac and Rosenthal, 2020).

3.4.5 Heckman Model

The training data, testing data, sel col, out col, and levels target are the five parameters that the Heckman selection model defined. Using a correlation matrix to calculate and exclude highly correlated variables over 0.7, the model determines if the predictor variables in the training data are multicollinear. After that, it creates selection and result formulae by picking the pertinent columns from the training and testing datasets. Using the provided formulae and the training data, the function uses the selection function to train a Heckman selection model. The model predicts the missing values in the testing data once it has been fitted. The model predicts the missing values in the testing data once it has been fitted. Subsequently, the predicted values are classified according to the target levels. The final values that were imputed are given back. The predicted value for a categorical class is then assigned using a function that uses established ranges. In addition to a list of category ranges, it requires two parameters for the predicted value. In order to determine if the predicted value falls inside any of the category ranges, the function iterates across the ranges. It returns the matching category if a match is discovered, and the initial projected value if not. Calculations are made about the target levels and imputed values to determine the range for each category class. The parameters it requires are the predicted values and the target levels. The predicted values' lowest and maximum values, as well as their whole range, are calculated. According to the number of target levels, it splits this range into equal pieces and adds a little unit to make sure the ranges don't overlap. Every category is given a list of ranges, which the function then returns. Every column in the data frame is iterated over by the Heckman model, which creates factors out of each column and substitutes NA for missing labels. The function populates the missing data and determines the column's target levels. After that, it adds these imputed values to the original data frame. For columns with missing data, the method generates a new data frame with imputed values (Bärnighausen *et al.*, 2011).

The Heckman model consists of two equations:

Selection Equation (binary response):

$$\pi(z) = \Phi(Z\beta) = \Phi(X\beta + \varepsilon)$$

where $\pi(z)$ is the probability of selection (1 for selected, 0 for not selected), Z is a vector of covariates, X is a subset of Z, β is a vector of coefficients, Φ is the cumulative distribution function of the standard normal distribution, and ε is an error term.

Outcome Equation (dependent variable):

$$y^* = X\gamma + v$$

where y^* is the true outcome variable, X is a vector of covariates, γ is a vector of coefficients, and v is an error term.

3.5 Evaluation

For this research purpose, we consider accuracy, sensitivity, specificity and precision as the evaluation metrics. The following defines what the evaluation metrics used are

Accuracy: The proportion of correctly classified data that a machine learning model that has been trained can accomplish is called its accuracy. The confusion matrix, which is made up of true positives, false positives, true negatives, and false negatives, is strongly connected to accuracy. Accuracy is a proportional measure of the number of correct predictions over all predictions (Bernardini *et al.*, 2023).

Sensitivity: Sensitivity, also referred to as the true positive rate, quantifies how well a model can detect positive events among all real positive instances. The percentage of true positives—or positively anticipated outcomes—to the total of true positives and false negatives is how it is computed. A high sensitivity suggests that the model successfully catches the majority of positive cases, reducing the false negative rate and the number of real positives that the model misses (Saroja and Kalpana, 2023).

Specificity: The capacity of a model to accurately identify negative occurrences out of all the real negative examples in a dataset is measured by a parameter called specificity. Another name for it is the genuine negative rate. Put another way, specificity is the degree to which a model can precisely forecast the lack of a condition or class—that is, the negative class—that it is attempting to uncover. With a high specificity, the model can effectively reduce false positives and provide an accurate prediction about whether a case actually falls into the negative group (Saroja and Kalpana, 2023).

Precision: Accurately classifying positive events is measured by a model's precision. The ratio of properly detected positive cases, or true positives, to the total number of positive predictions made by the model (true positives + false positives) is used to compute it. Presence of false positives is reflected in a reduced accuracy rate, which is essentially a measure of the model's dependability when it makes a positive prediction about an occurrence. A high precision model, then, ensures that the majority of its positive predictions are accurate by making fewer mistakes when recognizing positives (Xu *et al.*, 2020).

Confusion Matrix: A summary of a machine learning model's performance on a set of test data is provided via a confusion matrix. Based on the model's predictions, it is a way to show the proportion of accurate and inaccurate cases.

The number of occurrences that the model generated on the test data is shown in the matrix.

True Positive (TP): When a positive outcome is accurately predicted by the model, the actual result is also positive.

True Negative (TN): When a negative result is accurately predicted by the model, the real result is also negative.

False Positive (FP): When a positive result is predicted by the model but the actual result is negative.

False Negative (FN): When a positive result occurs instead of the expected negative one, the model predicted the wrong thing.

4 Results

This section presents the result obtained after carrying out a comprehensive exploratory data analysis, it also presents results of the various machine learning models in handling missing data.

Table 1 below gives a description of the chosen variables and their importance.

Table 1. Variables List and Description

Demographic Information		
province	Regional location	These variables provide general background about individuals, which helps understand the distribution and diversity of the sample population.
urban	Urban or rural setting	
gender	Gender	
age_group	Age group	
Socio-economic Information		
wealthquintile	Economic status	These variables represent people's social and economic circumstances, which might affect HIV risk factors, healthcare access, and preventive initiatives.
education	Educational level	

School-Related Factors		
schlat	School attendance history	These factors may affect early understanding of HIV prevention techniques since they are related to childhood education.
schlcur	Current school attendance	
Marital and Relationship Status		
evermar	Previous marital status	These factors investigate how HIV risk, transmission dynamics, and care-seeking behavior are impacted by marital and relationship status.
curmar	Current marital status	
Male circumcision		
mcstatus	Male circumcision status	Male circumcision is a key prevention strategy for reducing the risk of HIV transmission.
Reproductive Health and Sexual Behavior		
sexever	History of sexual activity	These factors center on sexual and reproductive health which is closely linked to the spread of HIV and the efficacy of preventative measures.
everpregnant	Pregnancy history	
avoidpreg	Steps to avoid pregnancy	
pregnant	Current pregnancy status	
HIV Testing and Care		
hivtstever	Ever done HIV test	

hivtstrslt	HIV test result	These variables relate to the detection, care, and treatment of HIV, providing insight into healthcare access and the effectiveness of testing and treatment programs.
hivcare	Have HIV care	
hivstatusfinal	Final HIV status	

4.1 Data Visualization

Figures 1 and 2 below show the distribution of males and females across 10 provinces. It also demonstrates that the distribution of men and women are relatively evenly distributed across the provinces. However, more women are represented in the dataset than men.

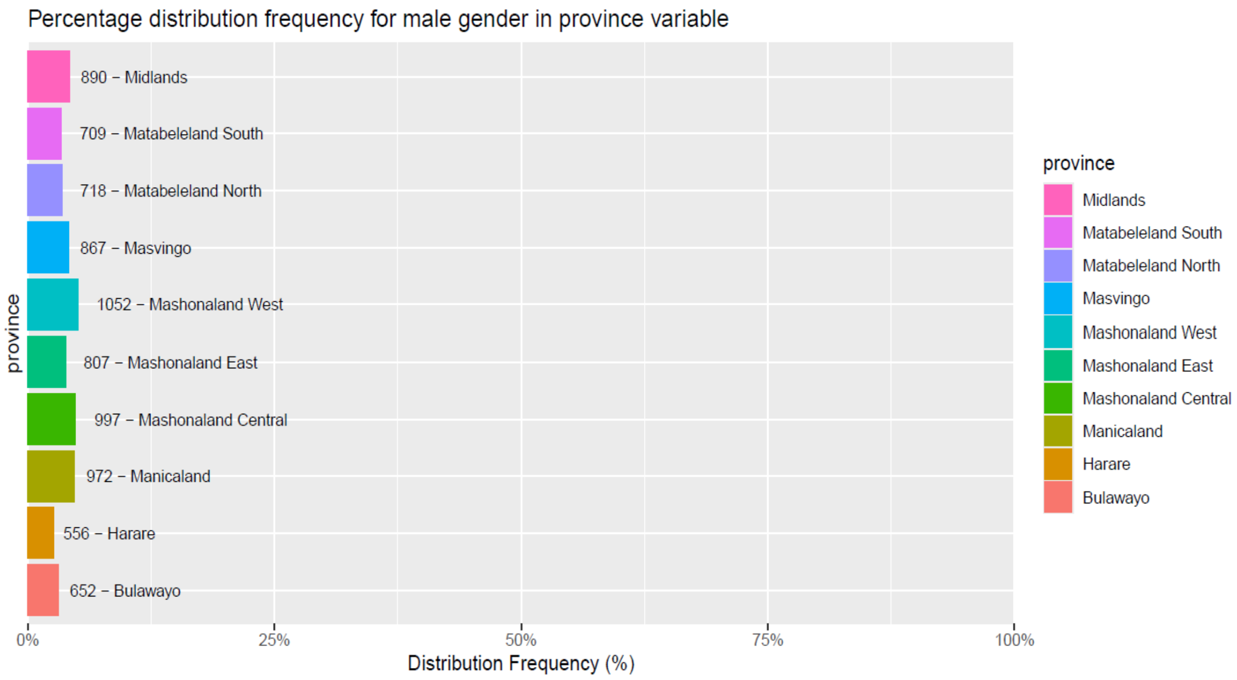


Figure 1. Percentage distribution frequency for male gender in province variable

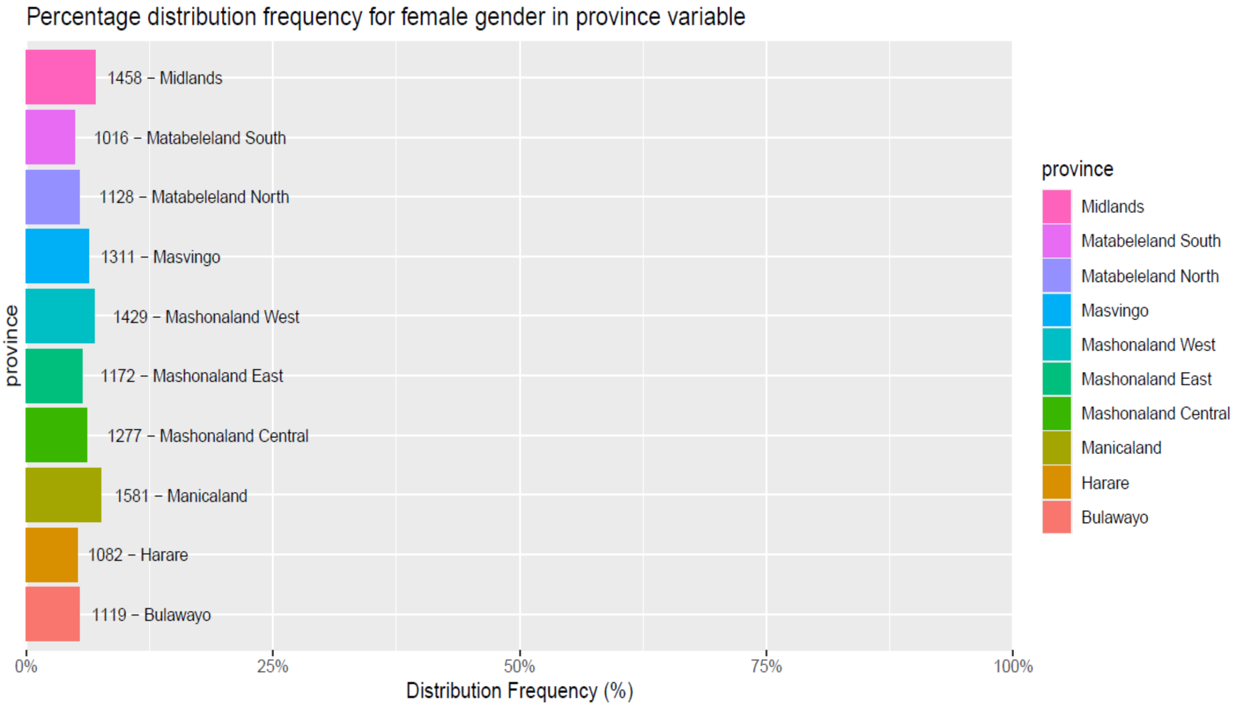


Figure 2. Percentage distribution frequency for female gender in province variable

Figures 3 and 4 below illustrate the distribution of men and women across urban and rural areas, respectively. Specifically, Figure 3 shows the ratio of men in urban regions compared to those in rural regions, while Figure 4 presents the ratio of women in rural areas relative to those in urban provinces. The data reveals a significant disparity in the representation of individuals from rural versus urban settings. In particular, the number of men residing in rural areas is approximately three times greater than that of men in urban areas. Similarly, the data indicates that the population of women in rural areas is about 2.4 times higher than that of women in urban areas. These findings suggest a pronounced rural bias in the dataset, which could influence subsequent analyses and interpretations, particularly in terms of demographic representation and its impact on health outcomes in the context of HIV/AIDS surveys. Understanding this uneven distribution is crucial for ensuring that the survey data accurately reflects the population and for developing appropriate interventions targeting both urban and rural communities.

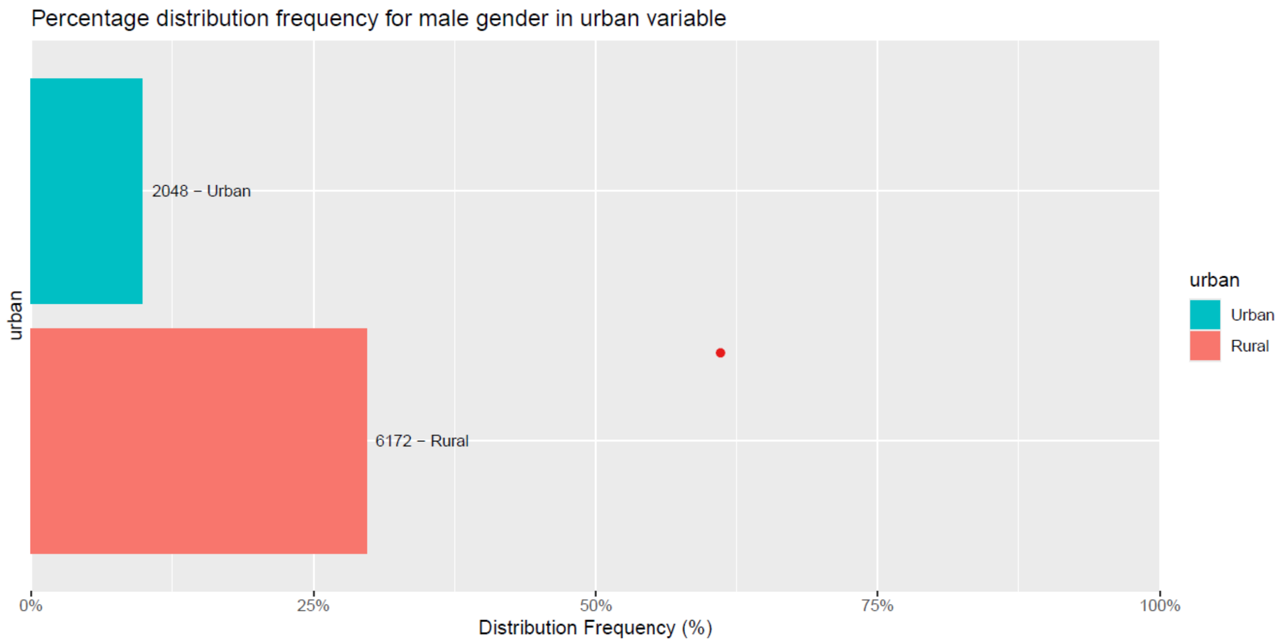


Figure 3. Percentage distribution frequency for male gender in urban variable

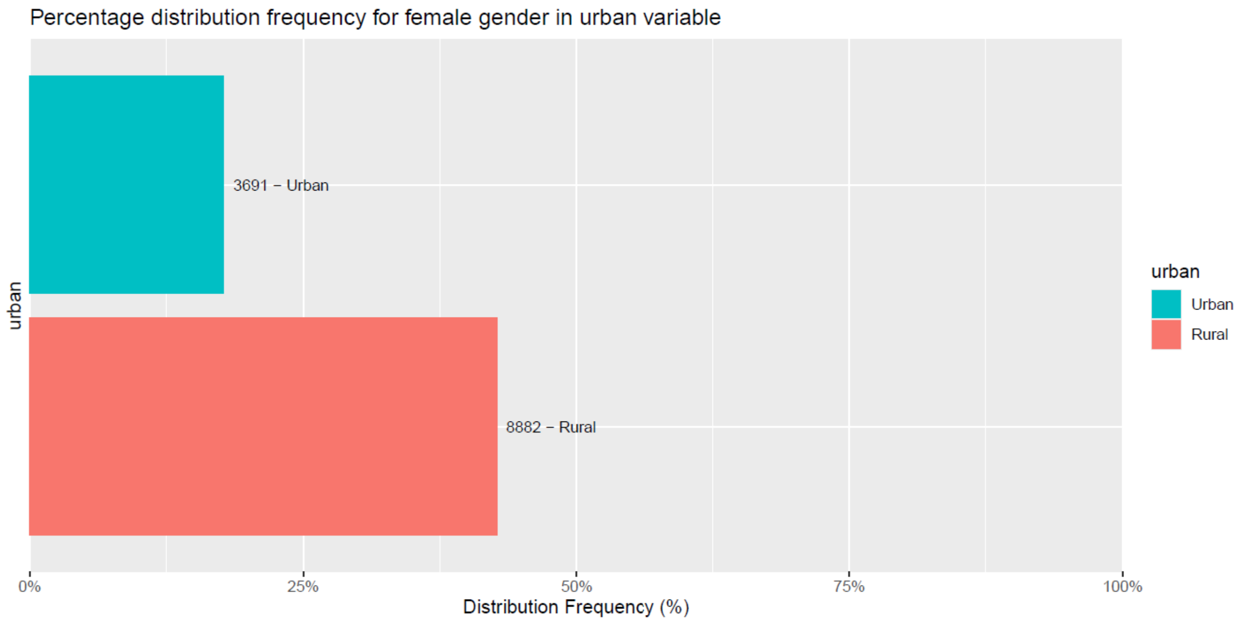


Figure 4. Percentage distribution frequency for female gender in urban variable

Figure 5 highlights that among men, the highest representation falls within the "Lowest" wealth quintile, followed by the "Second" quintile, which presents an intriguing pattern. This is followed by those in the "Middle" wealth quintile, with the "Highest" earners and individuals in the "Fourth" quintile showing the lowest representation. Similarly, Figure 6 demonstrates that for women, the "Lowest" wealth quintile also has the largest representation, followed by the "Second", "Middle", "Highest", and "Fourth" quintiles.Both figures exhibit a consistent trend across genders, where individuals in the "Lowest" and "Second" wealth quintiles constitute the majority of the surveyed population. This trend underscores a socio-economic disparity within the dataset, with a disproportionate representation of individuals from lower wealth quintiles. The prevalence of lower-income individuals in the sample may influence the generalizability of findings, particularly in understanding how wealth impacts health outcomes, access to healthcare, and HIV/AIDS interventions. It suggests that socio-economic status, as captured by wealth quintiles, plays a critical role in shaping the demographic distribution of the study population, which may require targeted strategies to address the specific needs of lower-income groups in both the prevention and treatment of HIV/AIDS. The underrepresentation of higher wealth quintiles further indicates the need for careful interpretation when applying these findings to wealthier populations.

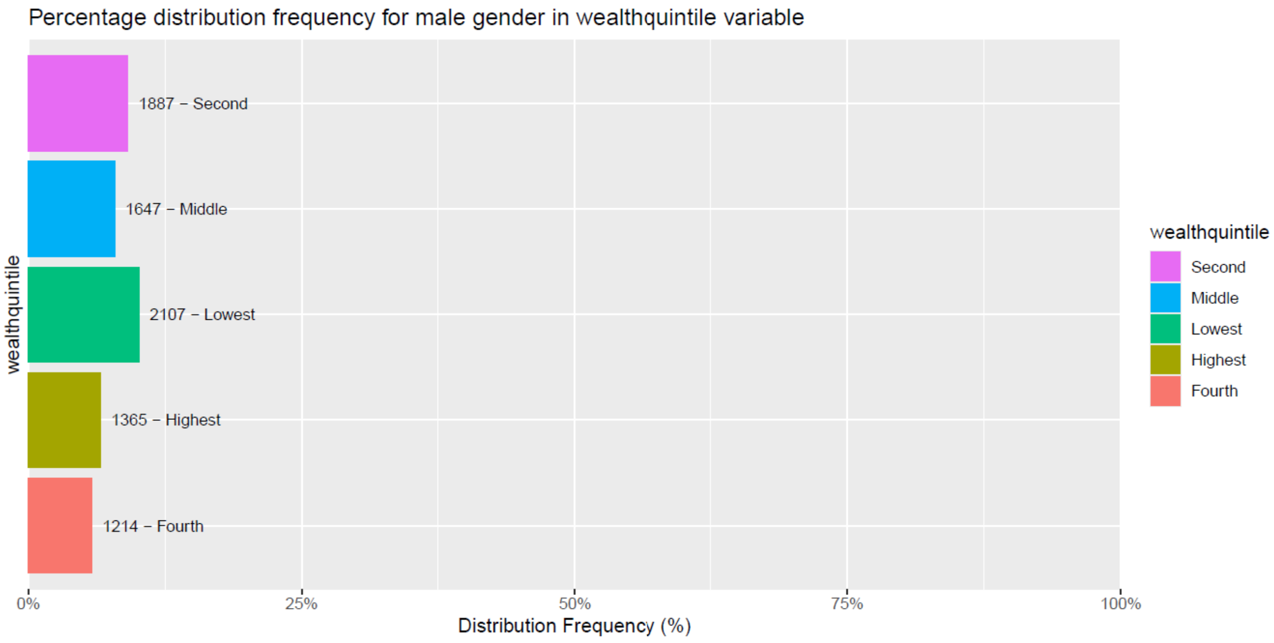


Figure 5. Percentage distribution frequency for male gender in wealthquintile variable

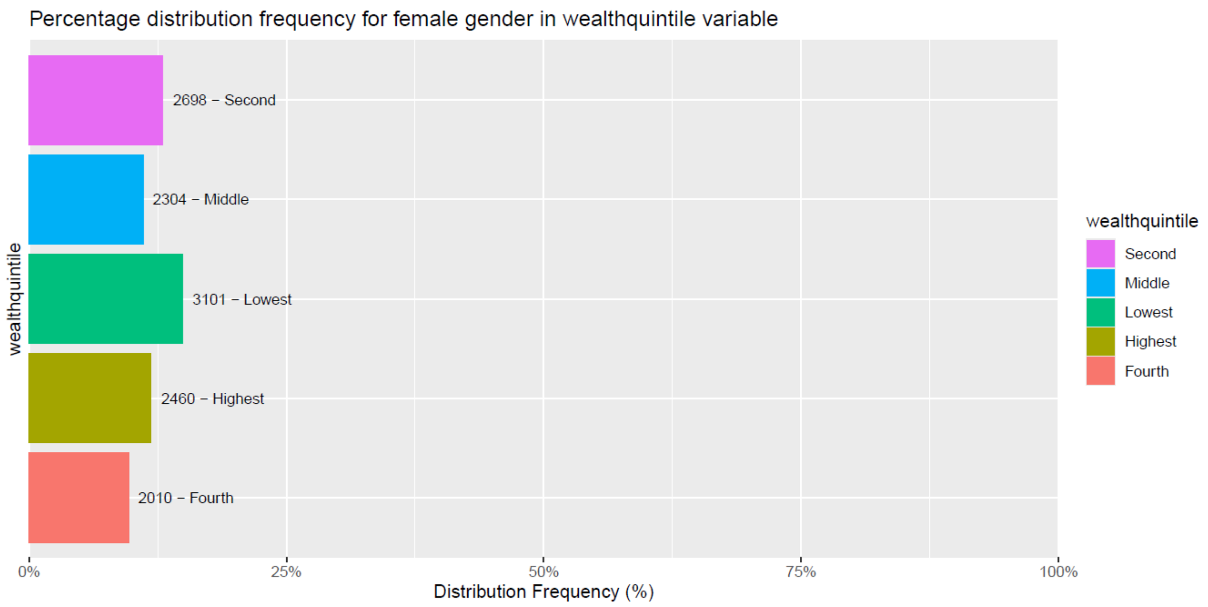


Figure 6. Percentage distribution frequency for female gender in wealthquintile variable

Figures 7 and 8 show that 97% of men and 90% of women responded favourably to the “schlat” variable, respectively.

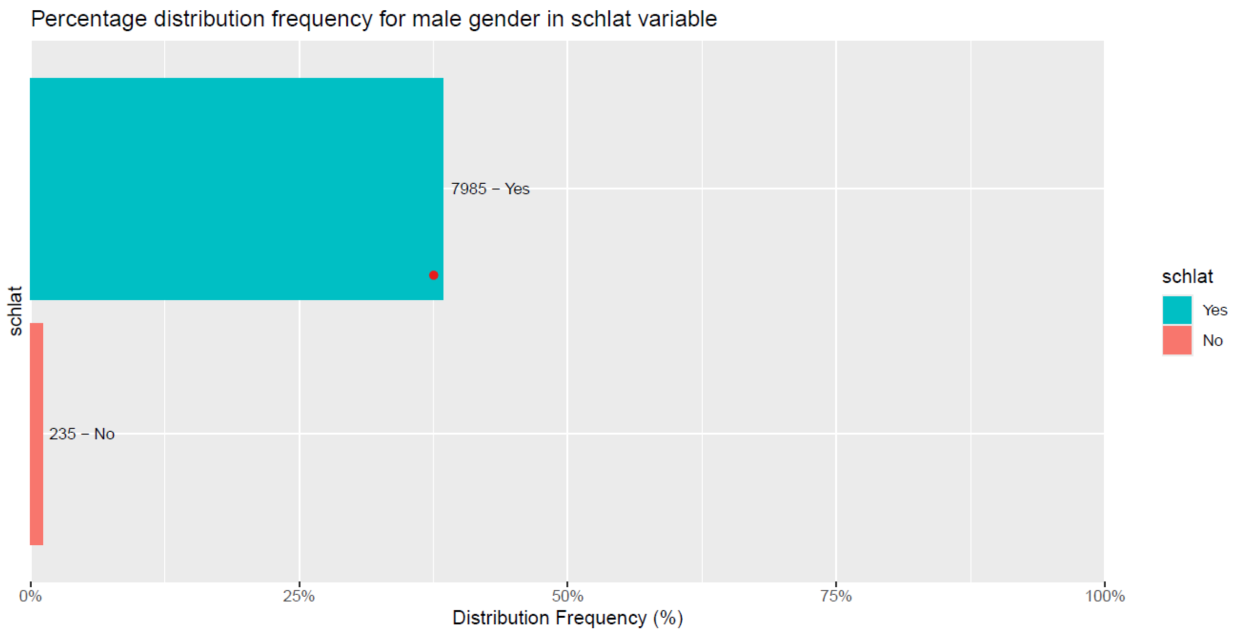


Figure 7. Percentage distribution frequency for male gender in schlat variable

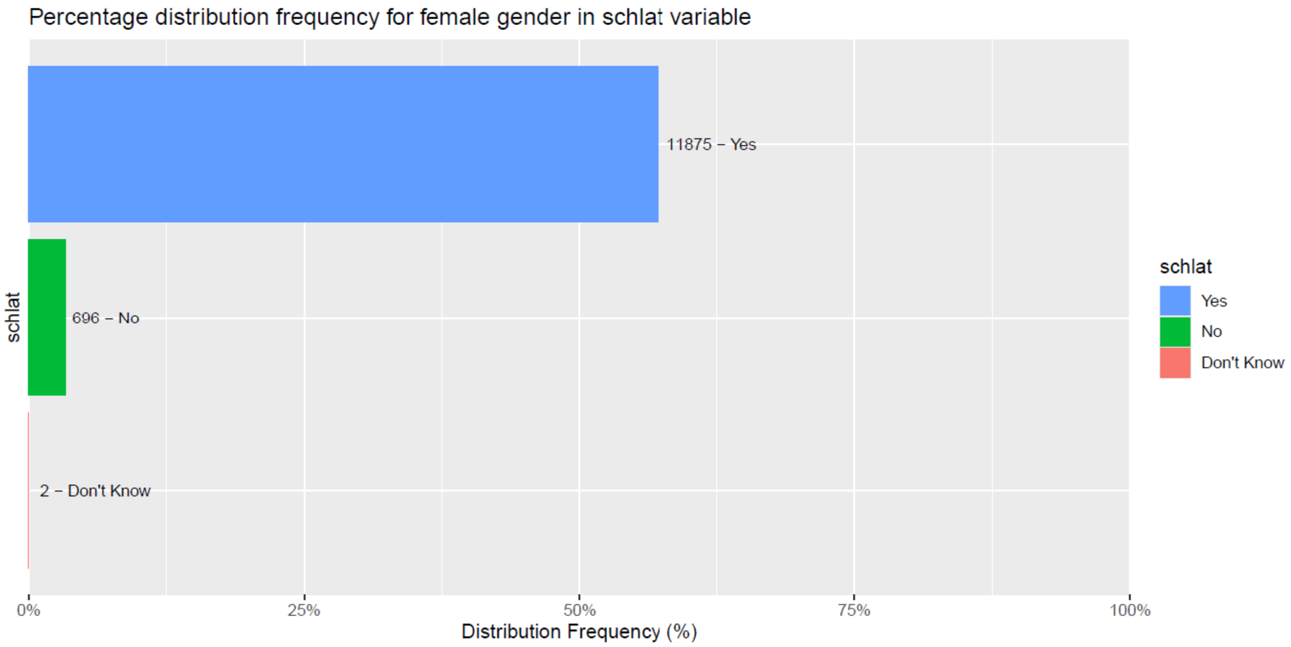


Figure 8. Percentage distribution frequency for male gender in schlat variable

Figure 9 and 10 show that just 2% of the data is missing for the female gender in the “schlcur” variable, with the vast majority giving the “No” value.

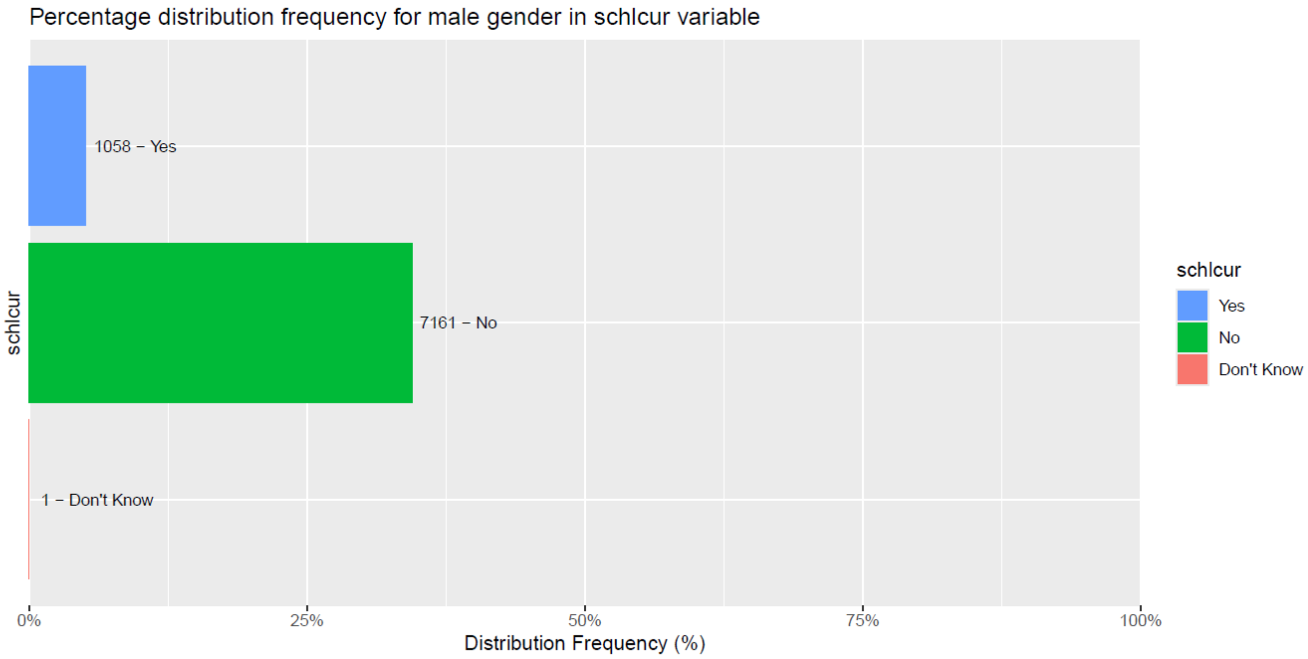


Figure 9. Percentage distribution frequency for male gender in schlcur variable

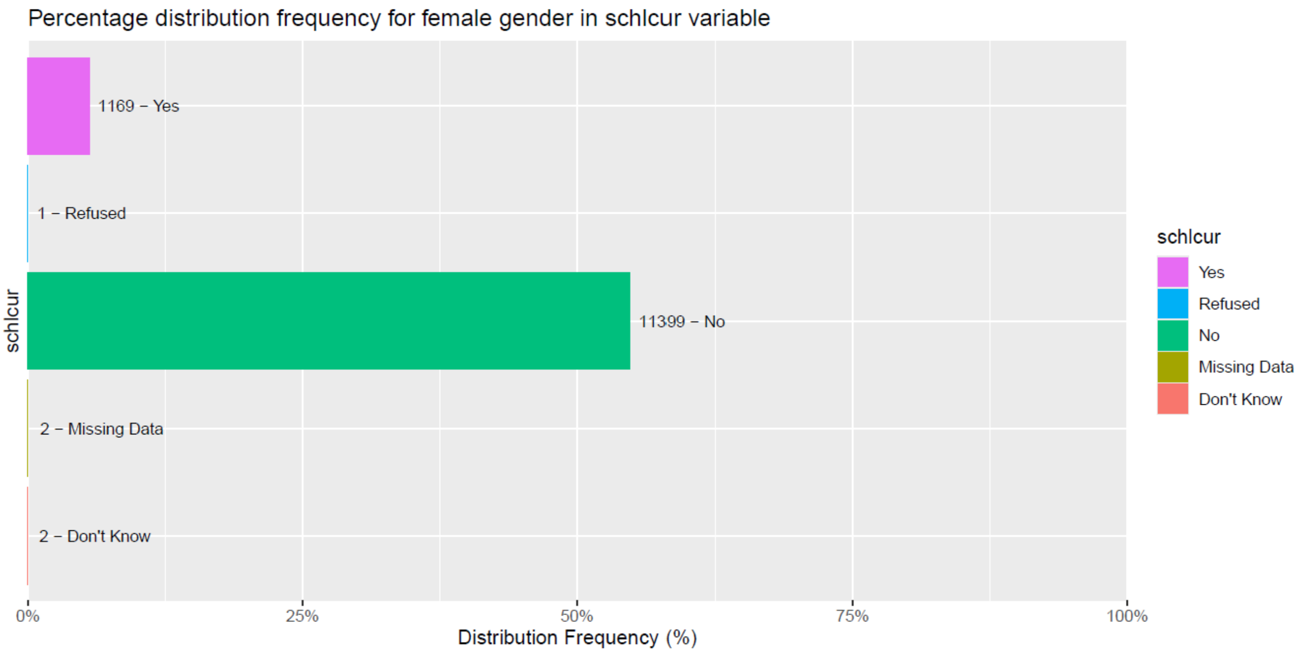


Figure 10. Percentage distribution frequency for female gender in schlcur variable

Figures 11 and 12 illustrates the educational background of the respondents. The distribution for the education column is similar for both genders where the missing data is negligible and does not significantly affect the dataset.

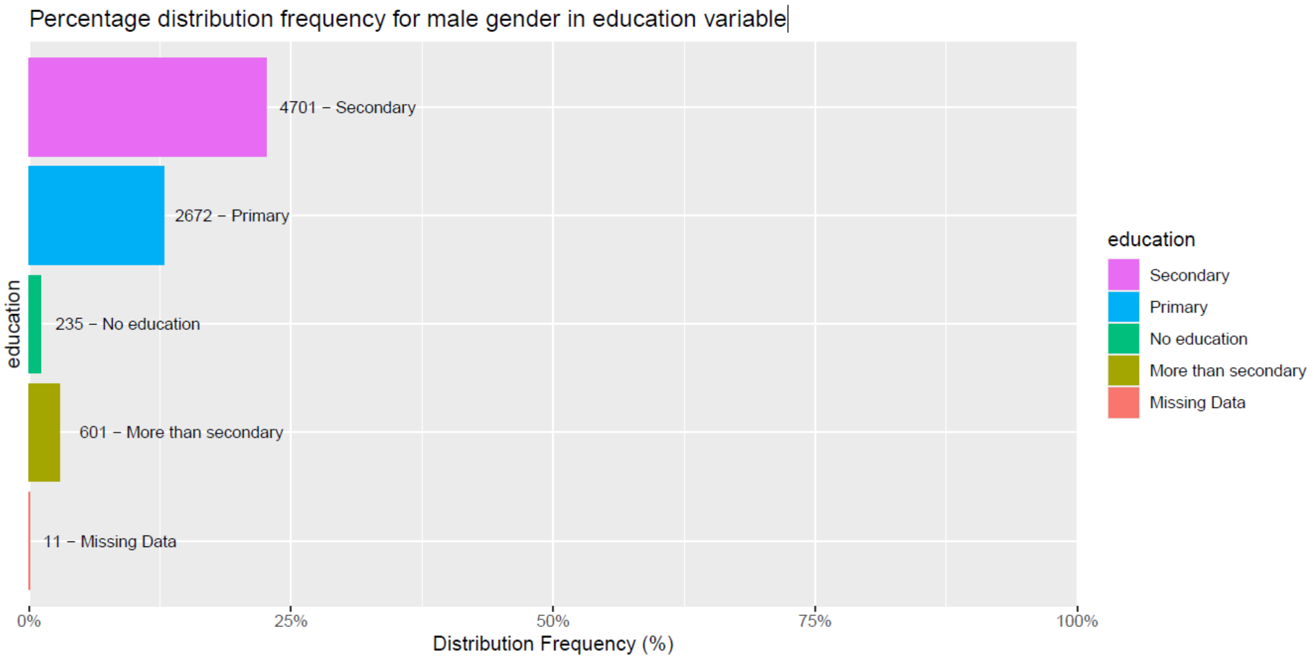


Figure 11. Percentage distribution frequency for male gender in education variable

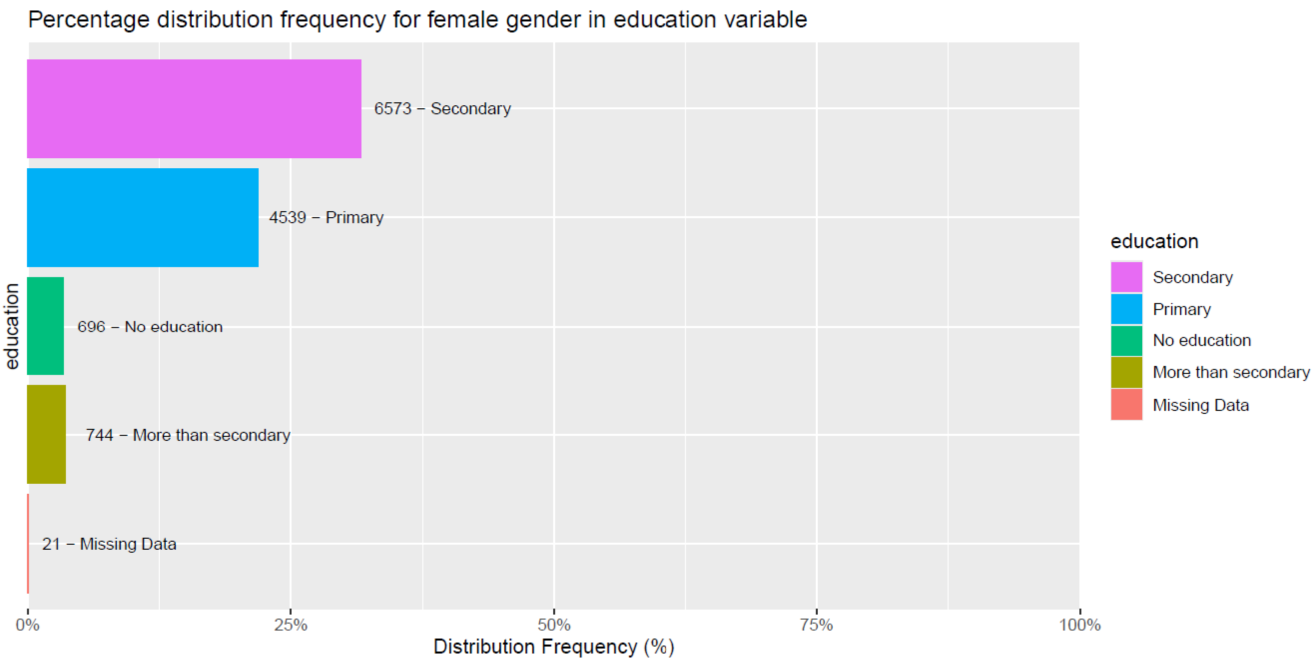


Figure 12. Percentage distribution frequency for male gender in education variable

Figures 13 and 14 show that there is no missing data in the “evermar” variable. However, figure 14 depicting the distribution shows a more noticeable skewness than the other group.

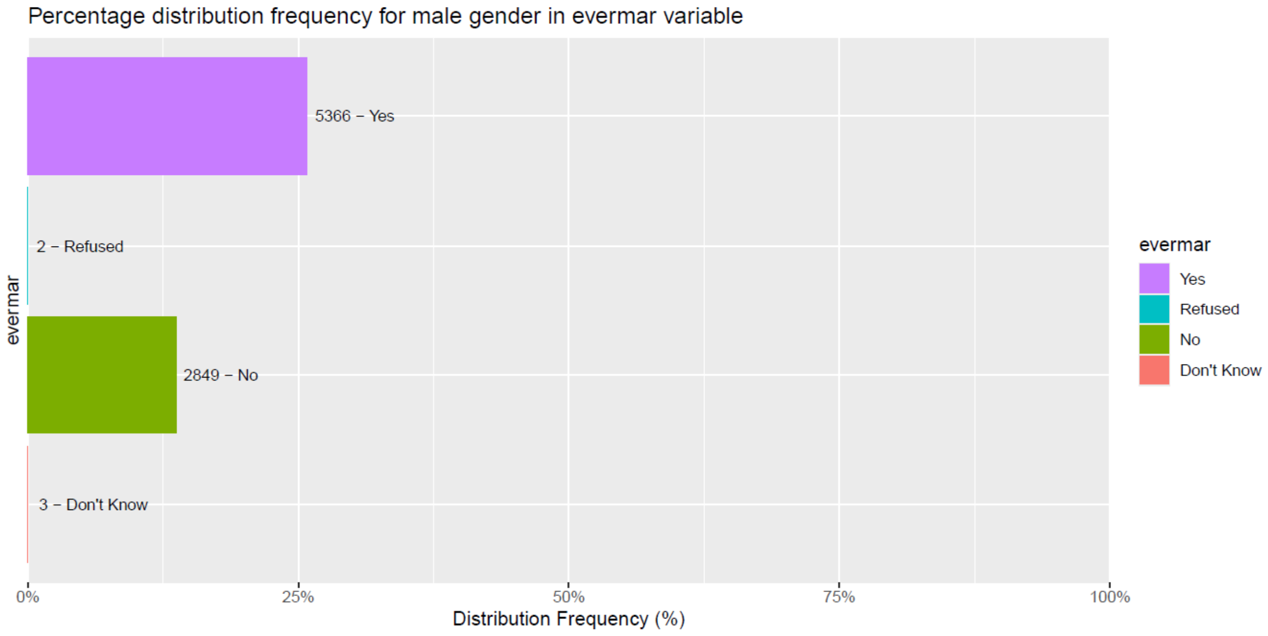


Figure 13. Percentage distribution frequency for male gender in evermar variable

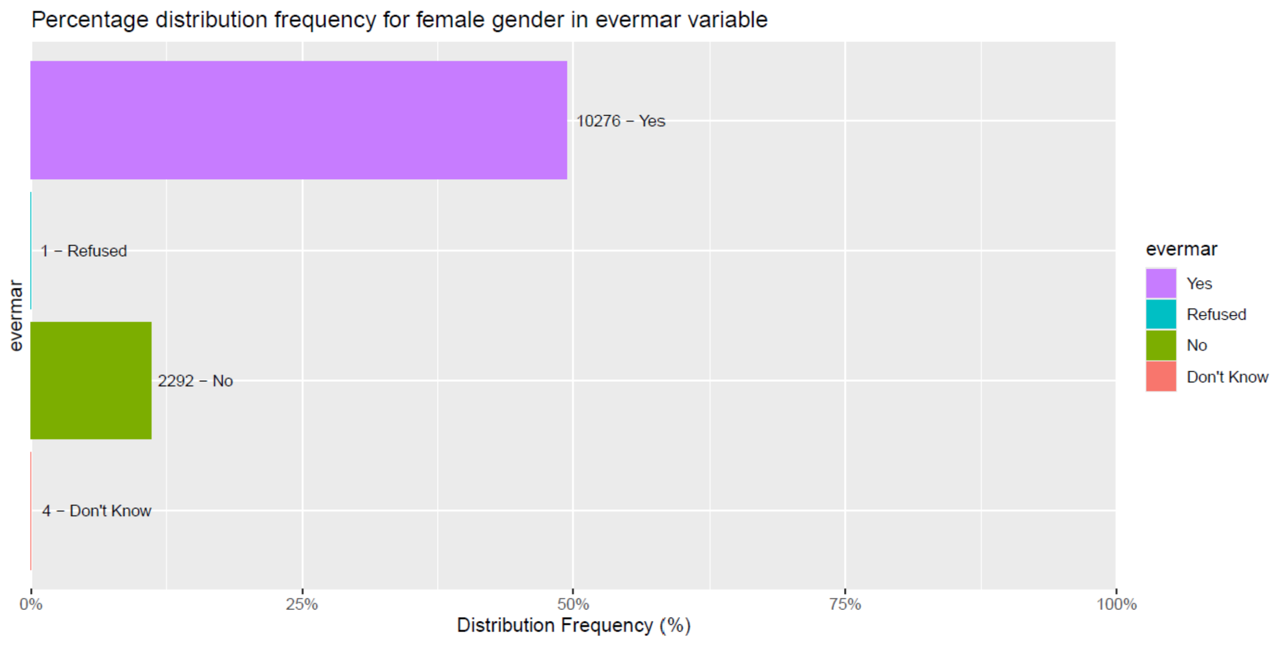


Figure 14. Percentage distribution frequency for female gender in evermar variable

For the “curmar” variable, the missing data for both genders are about the same as demonstrated in figures 15 and 16

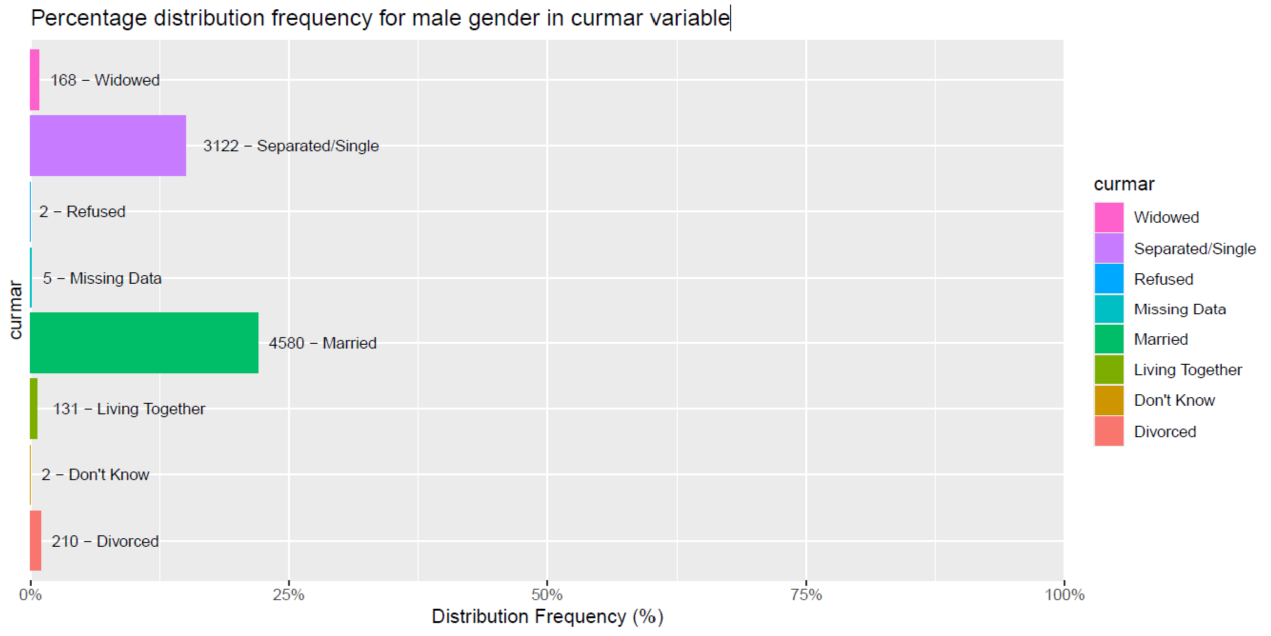


Figure 15. Percentage distribution frequency for male gender in curmar variable

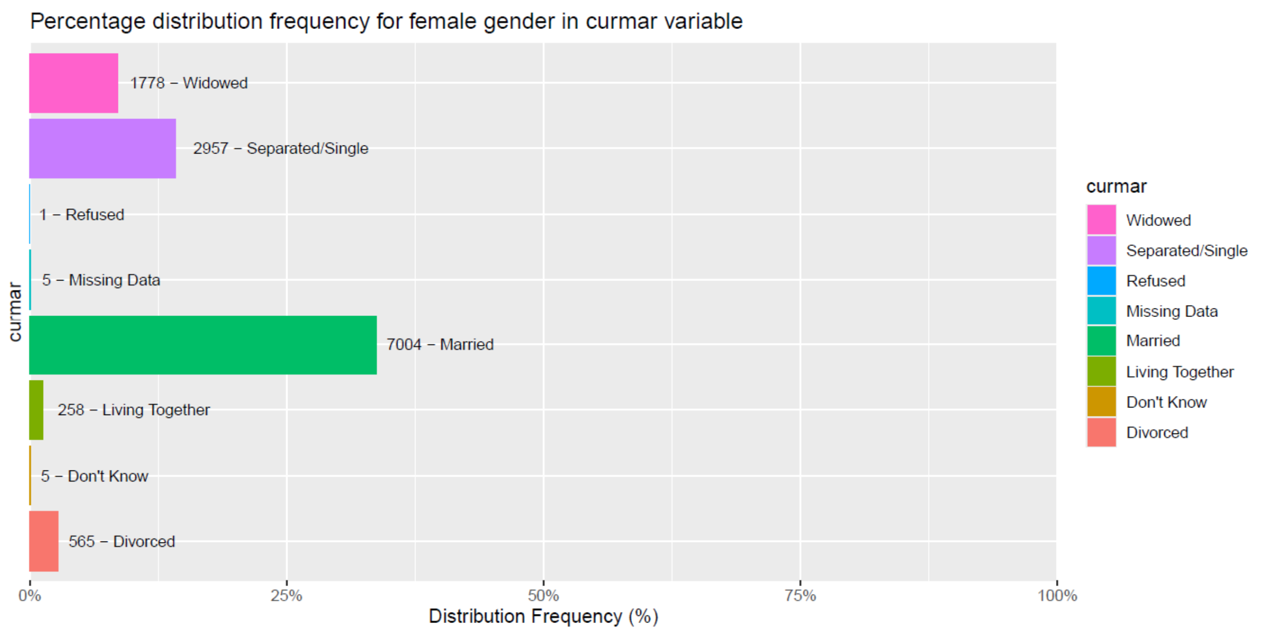


Figure 16. Percentage distribution frequency for female gender in curmar variable

Figures 17 and 18 shows at >0.5% of this data is missing for this group. However, the mcstatus variable does not apply to the female group, thus no data is regarded as missing.

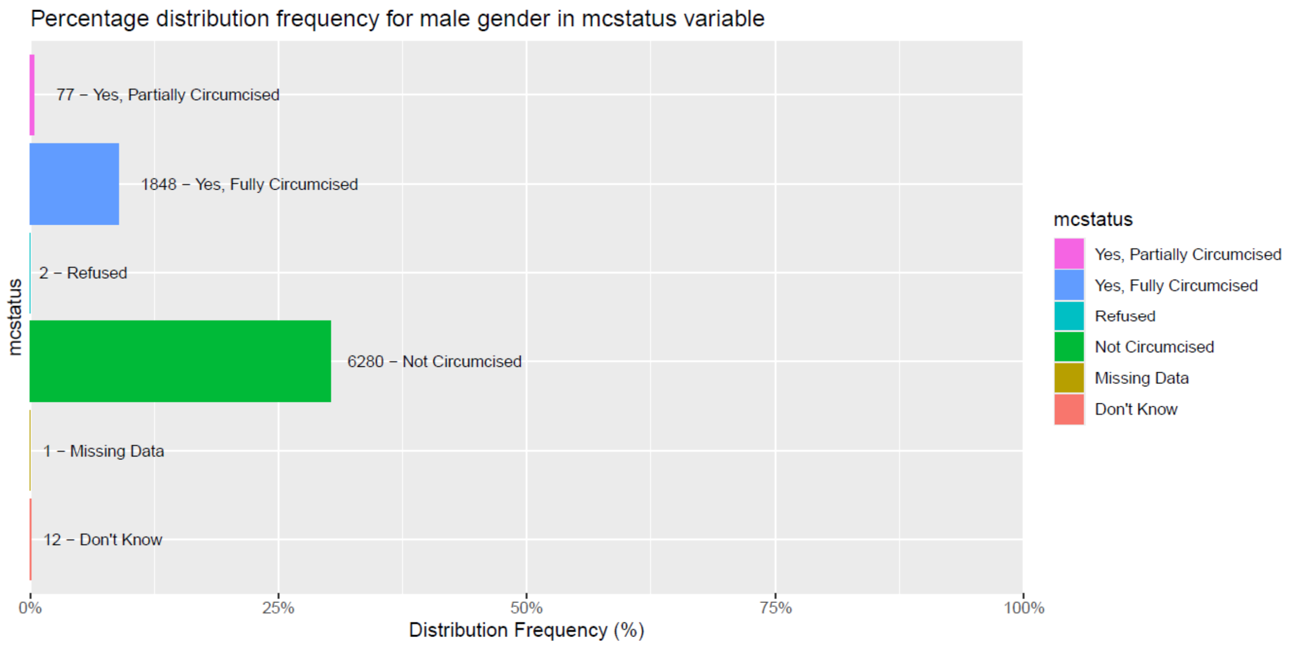


Figure 17. Percentage distribution frequency for male gender in mcstatus variable

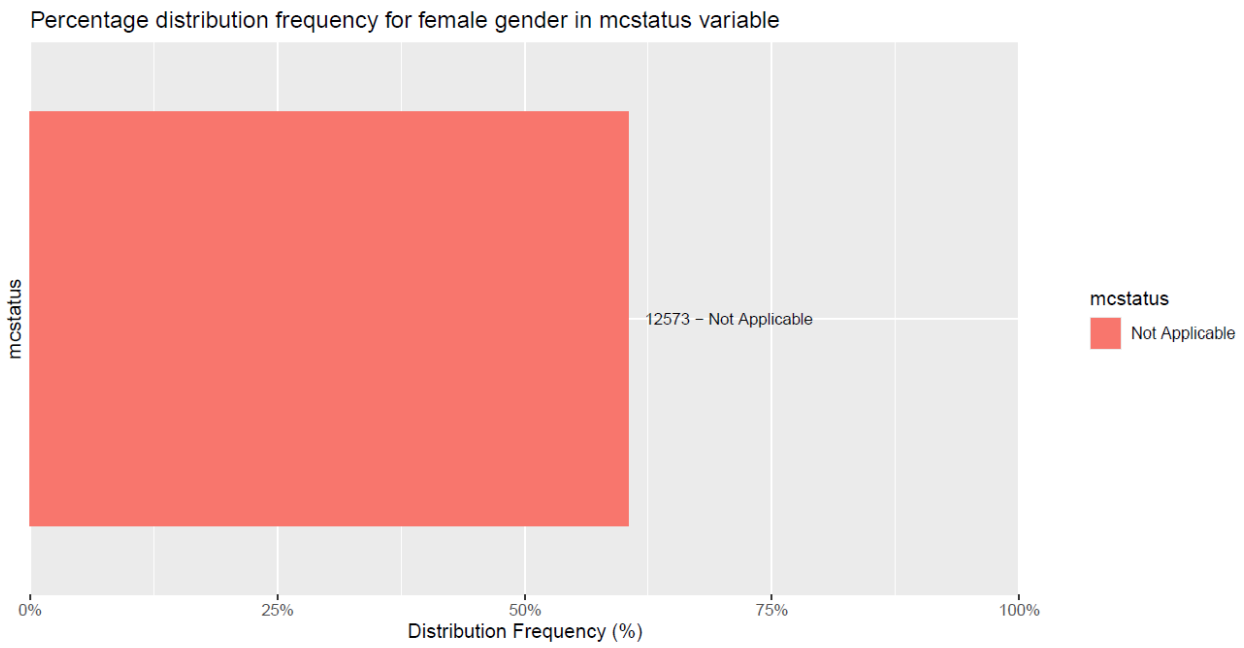


Figure 18. Percentage distribution frequency for female gender in mcstatus variable

Figures 19 and 20 show that in the “sexever” variable, the larger group has fewer missing data than the smaller group which is interesting.

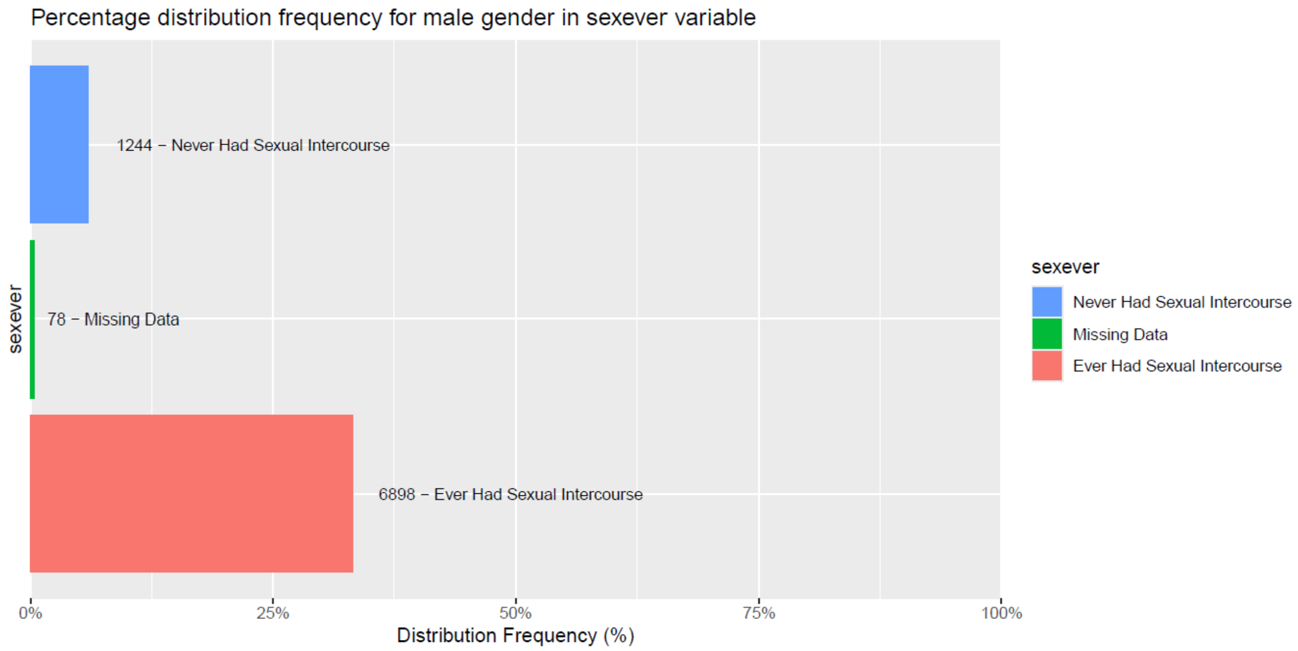


Figure 19. Percentage distribution frequency for male gender in sexever variable

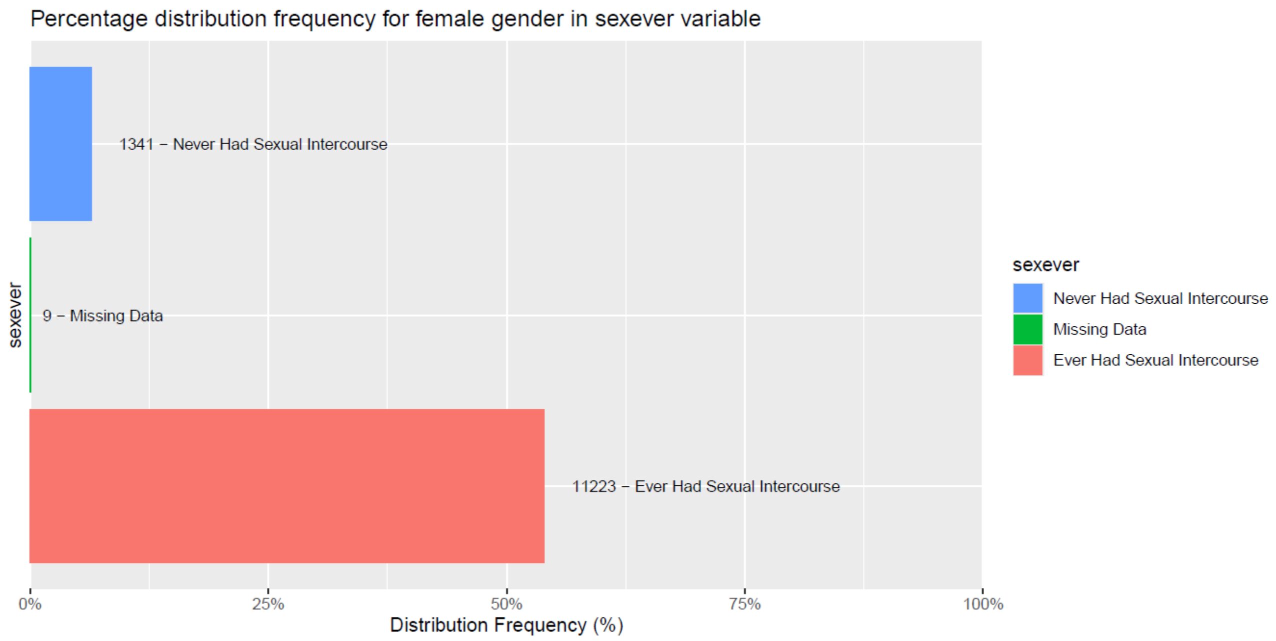


Figure 20. Percentage distribution frequency for female gender in sexever variable

Similar to figures 17 and 18 but the inverse, 21 and 22 show that the "everpregnant" value is not applicable to the male group. However, some data was missing for the group in question.

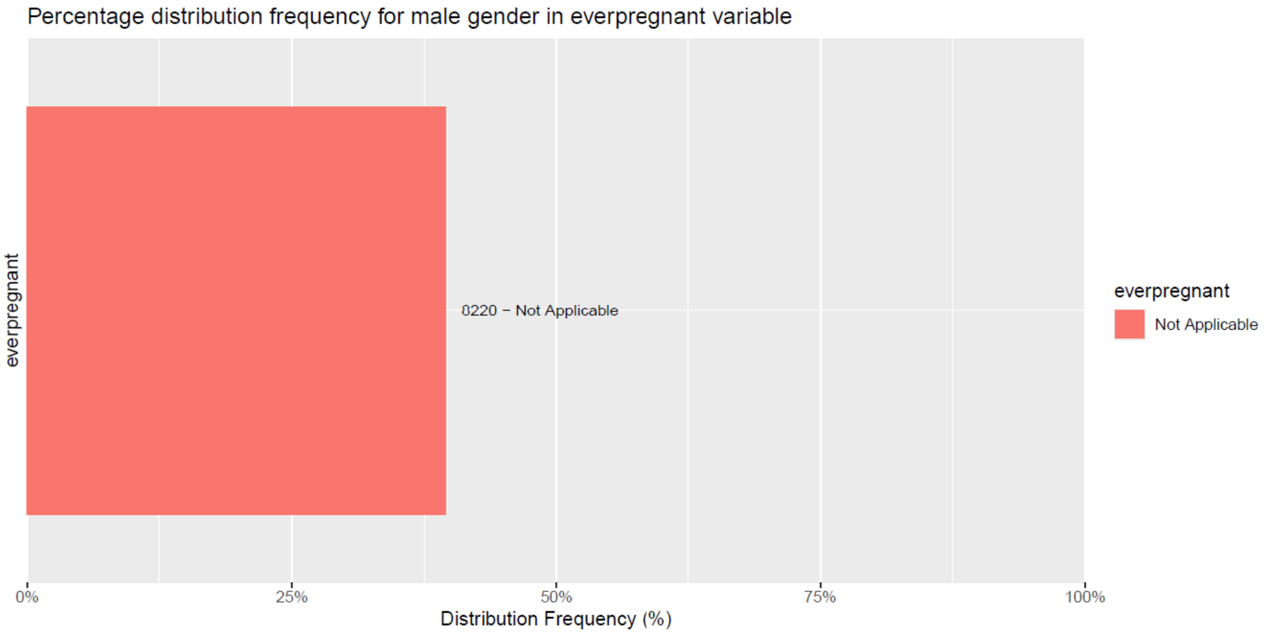


Figure 21. Percentage distribution frequency for male gender in everpregnant variable

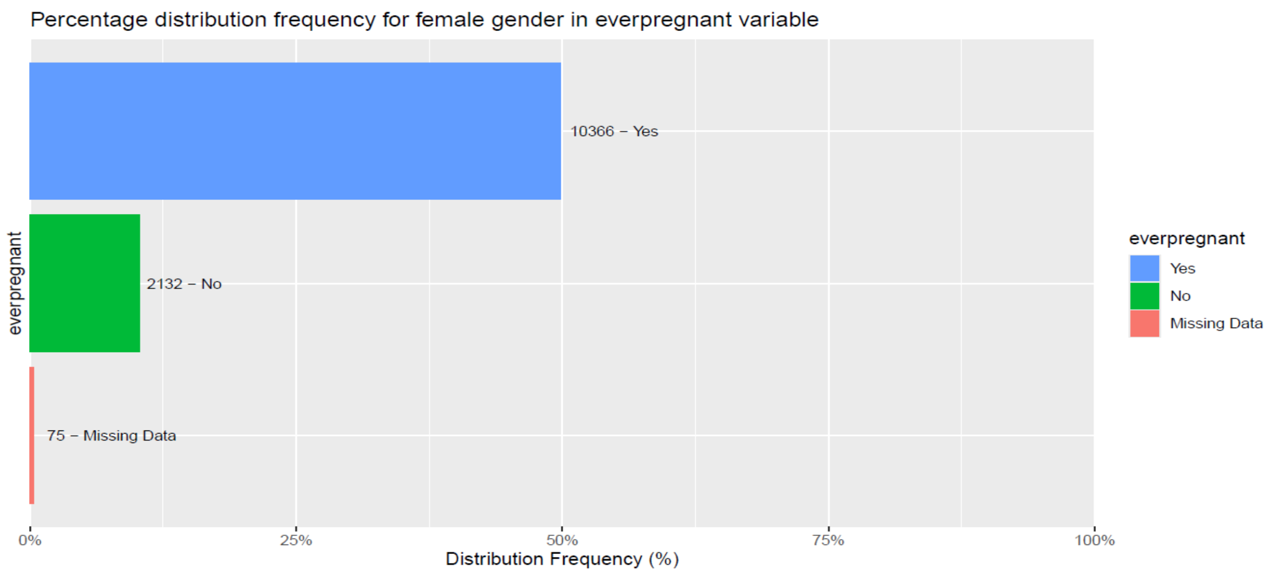


Figure 22. Percentage distribution frequency for female gender in everpregnant variable

Figures 23 and 24 do not demonstrate any missing data but exhibits similar distribution between both groups.

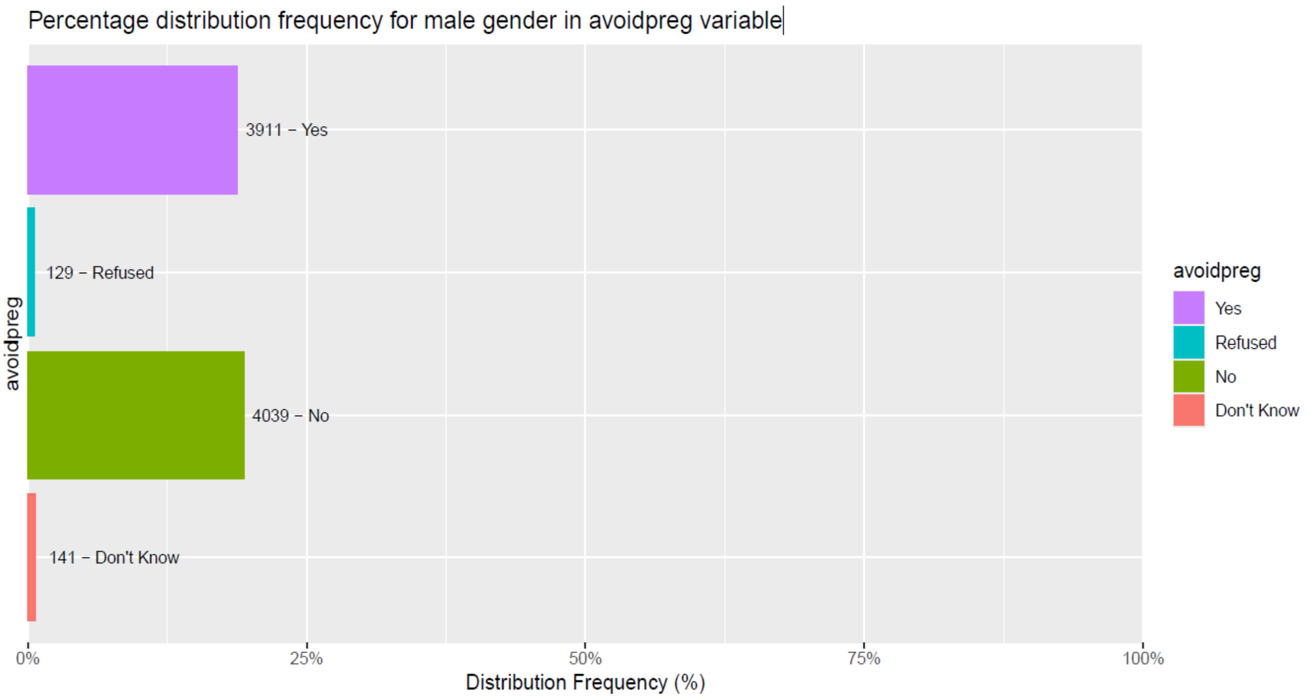


Figure 23. Percentage distribution frequency for male gender in avoidpreg variable

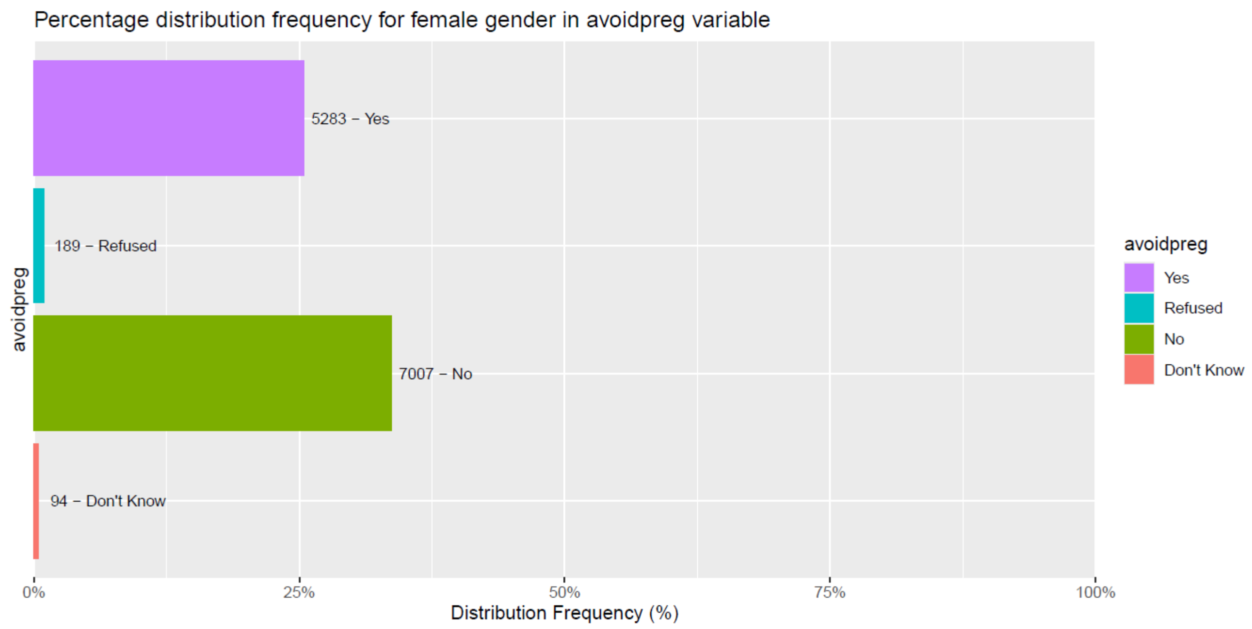


Figure 24. Percentage distribution frequency for female gender in avoidpreg variable

Figures 25 and 26 do not demonstrate and missing data, however the target value is not applicable to the male group.

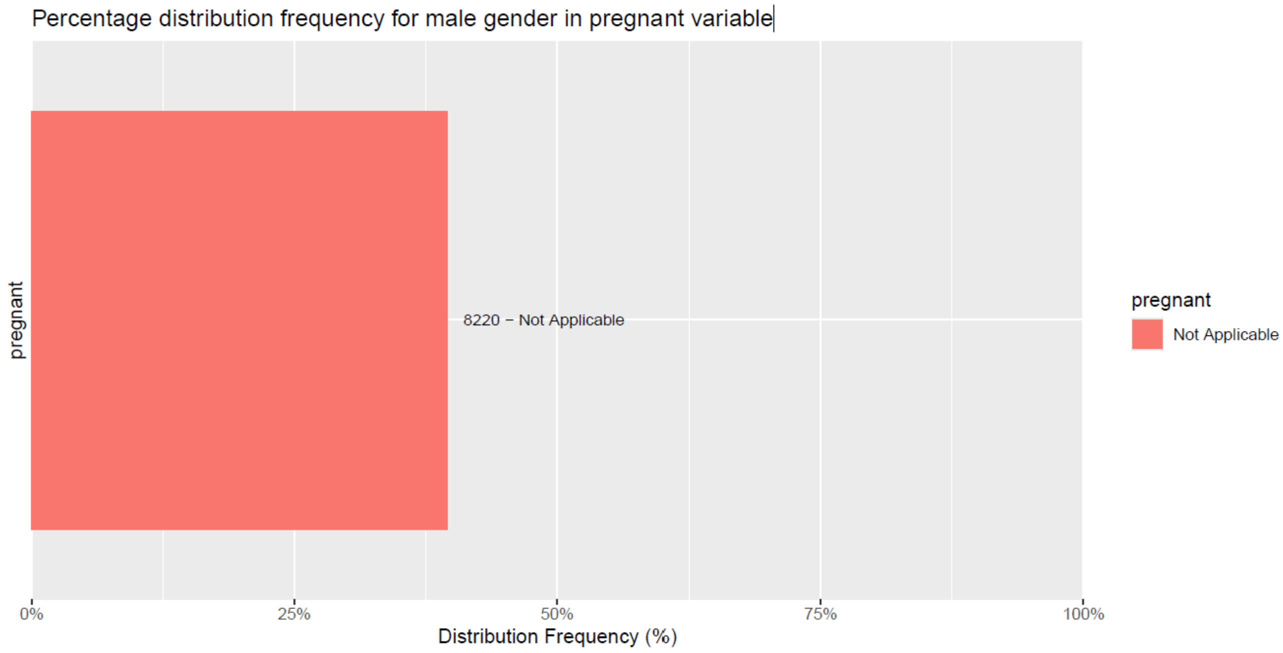


Figure 25. Percentage distribution frequency for male gender in pregnant variable

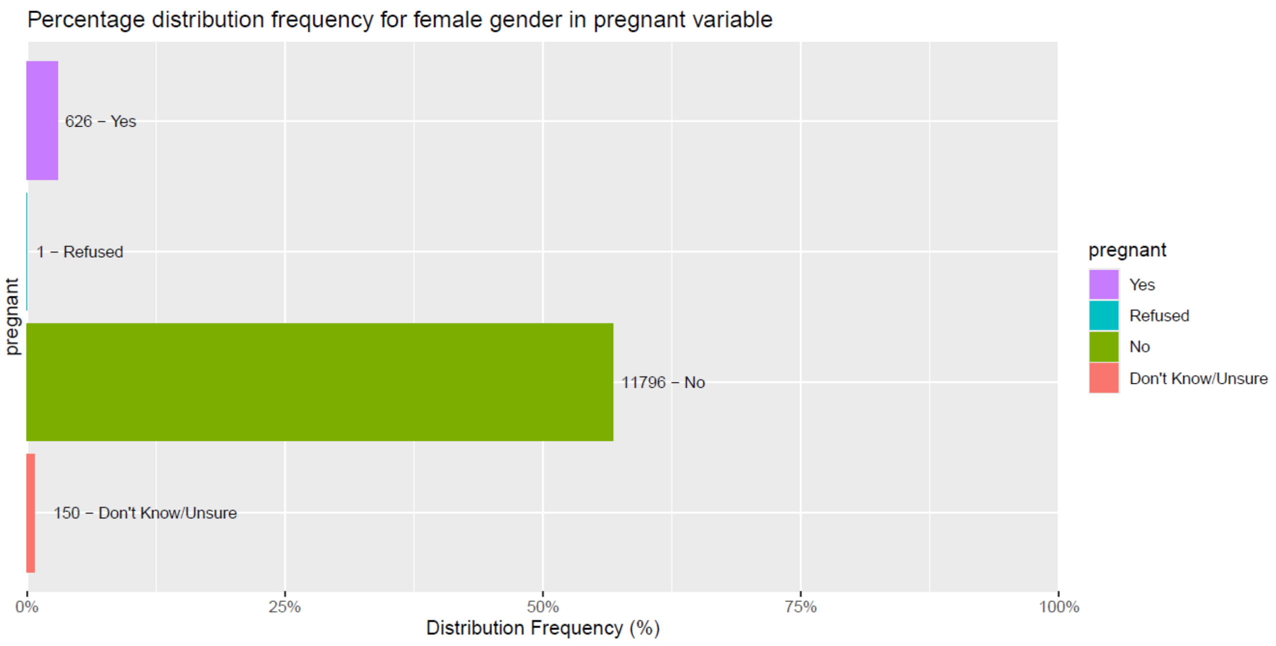


Figure 26. Percentage distribution frequency for female gender in pregnant variable

Figures 27 and 28 demonstrate that there is no missing data for both groups in the “hivtstever” variable.

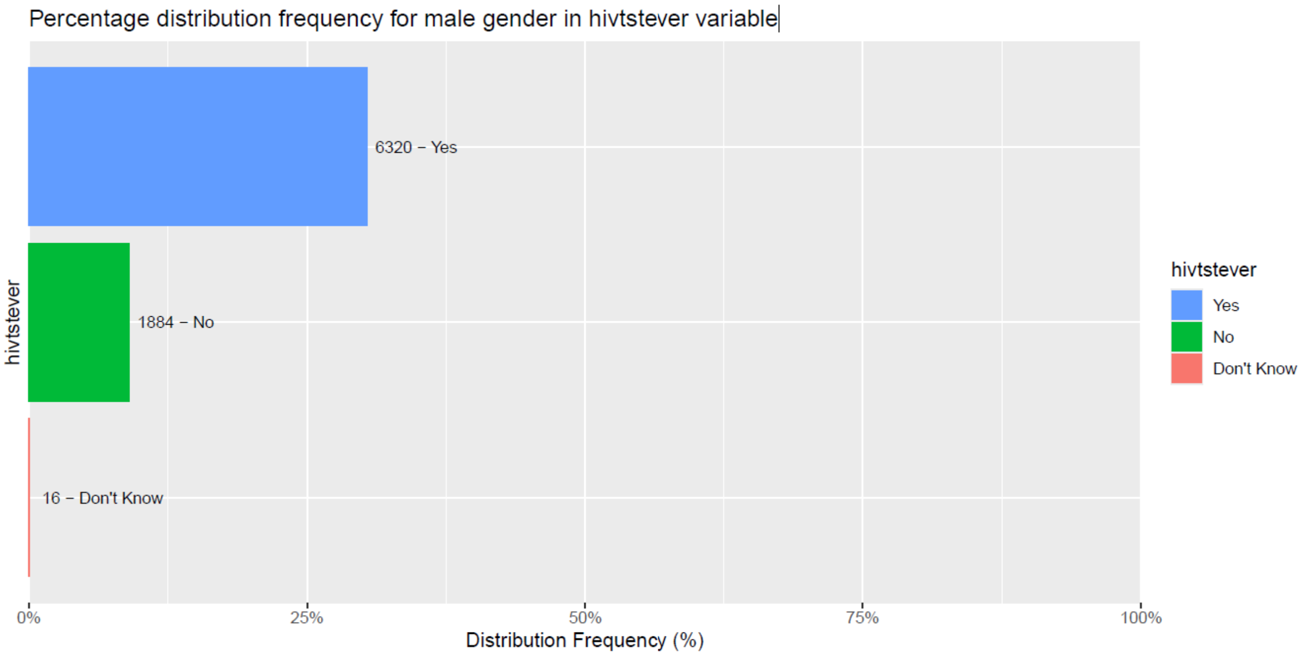


Figure 27. Percentage distribution frequency for male gender in hivtstever variable

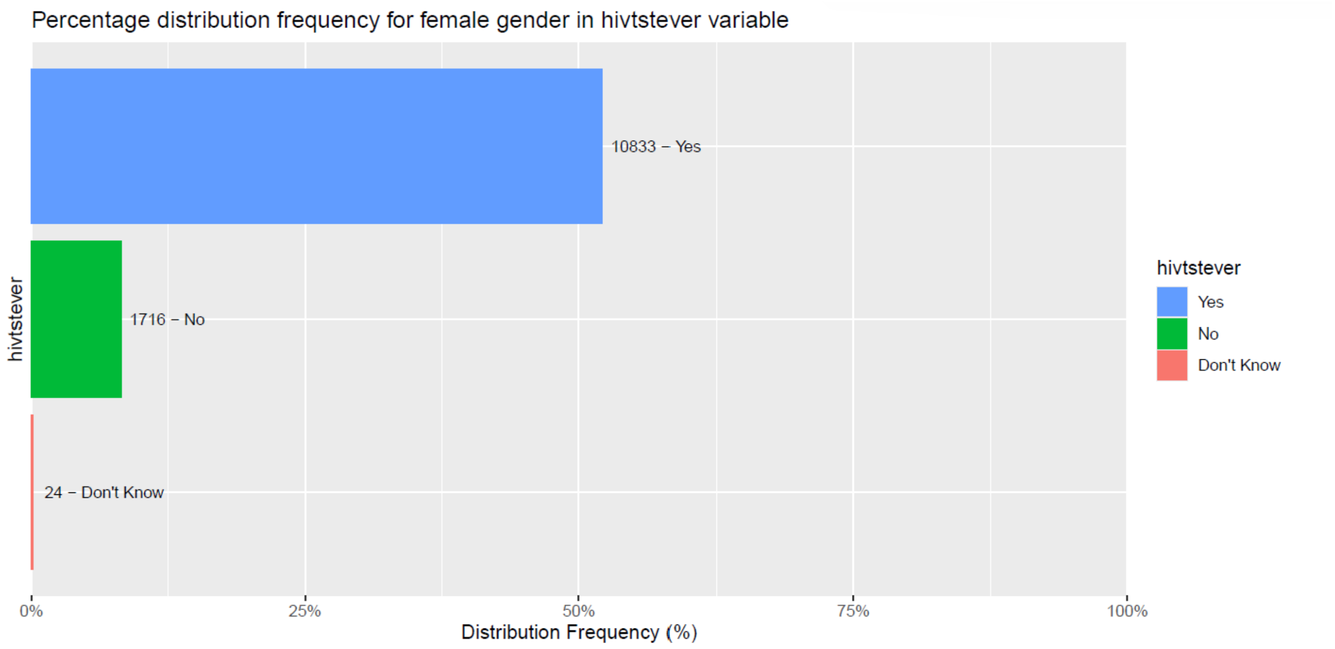


Figure 28. Percentage distribution frequency for female gender in hivtstever variable

Figure 29 and 30 show that both groups have missing data to the same degree, with the same pattern of distribution.

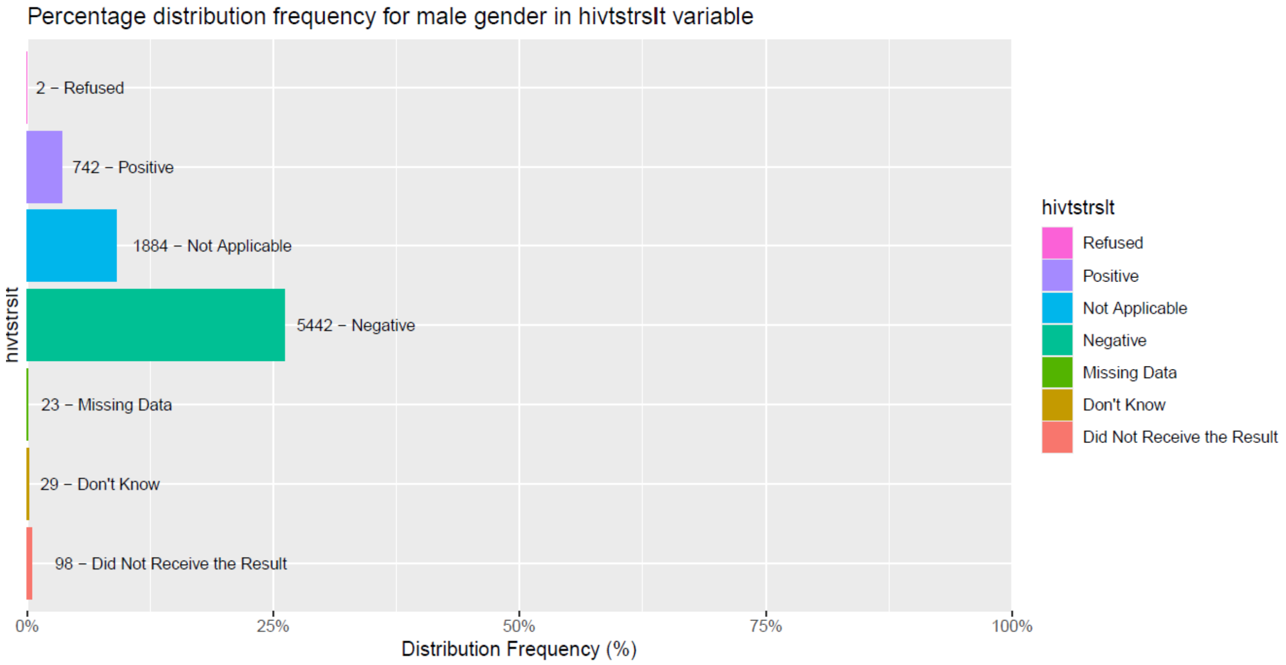


Figure 29. Percentage distribution frequency for male gender in hivtstrslt variable

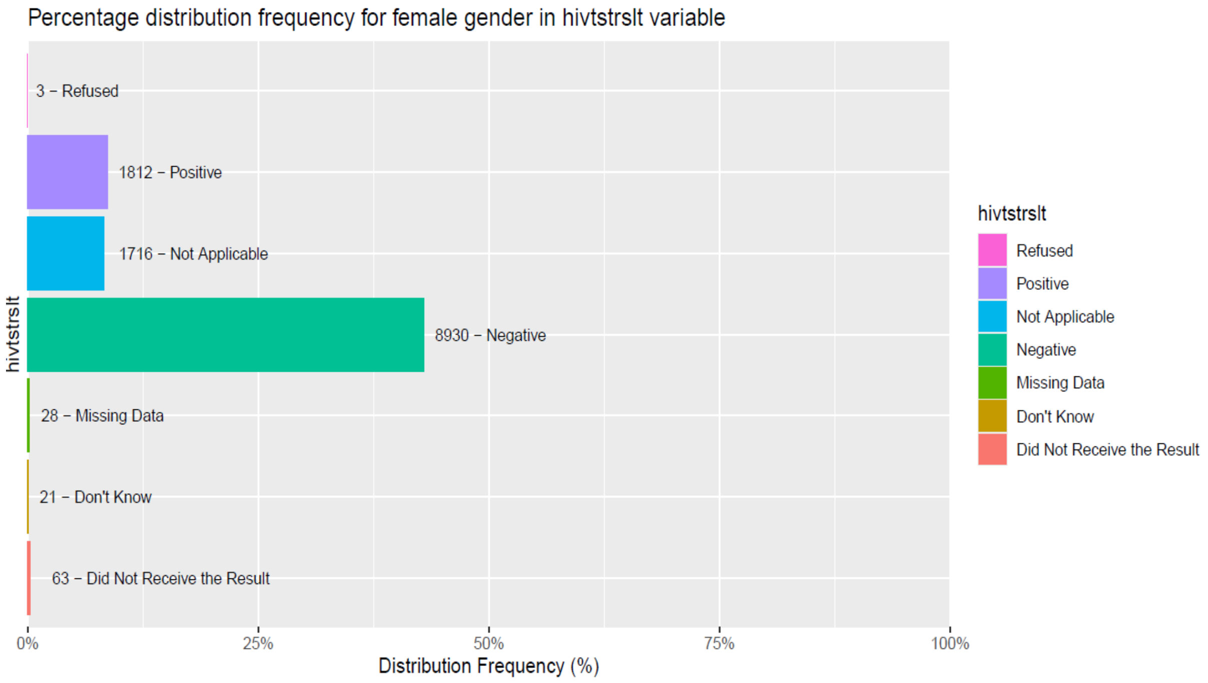


Figure 30. Percentage distribution frequency for female gender in hivtstrslt variable

The “hivcare” variable for both groups illustrated in figures 31 and 32, do not demonstrate any data missingness.

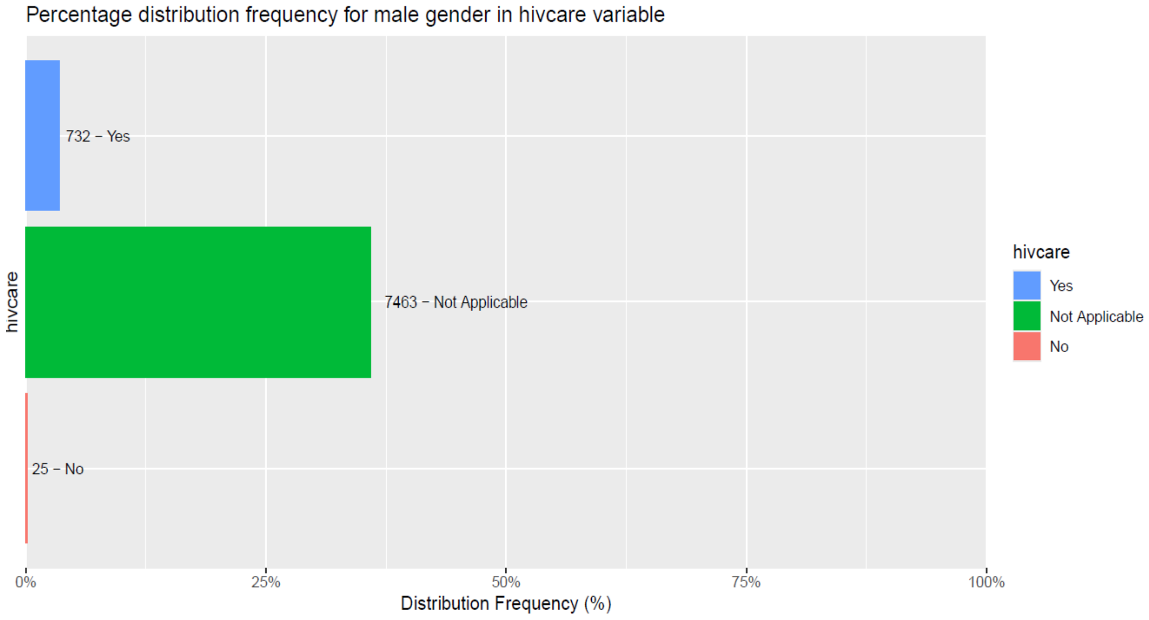


Figure 31. Percentage distribution frequency for male gender in hivcare variable

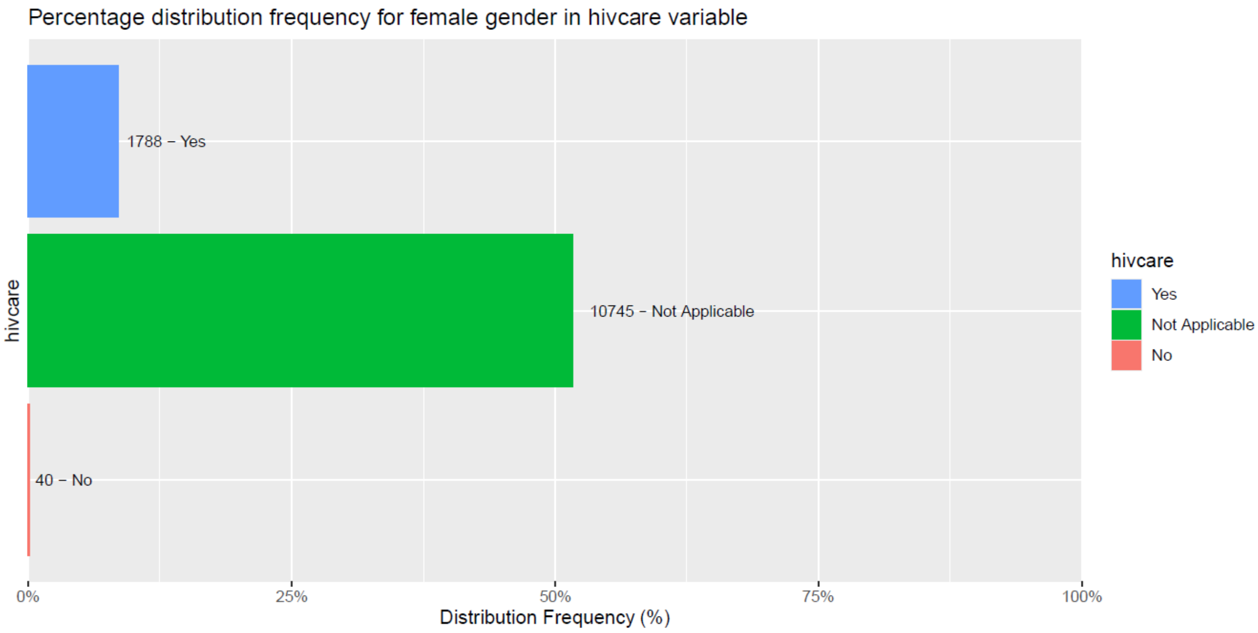


Figure 32. Percentage distribution frequency for female gender in hivcare variable

The “hivstatusfinal” for both groups depicted in figures 33 and 34 demonstrate the highest degree of missing data.

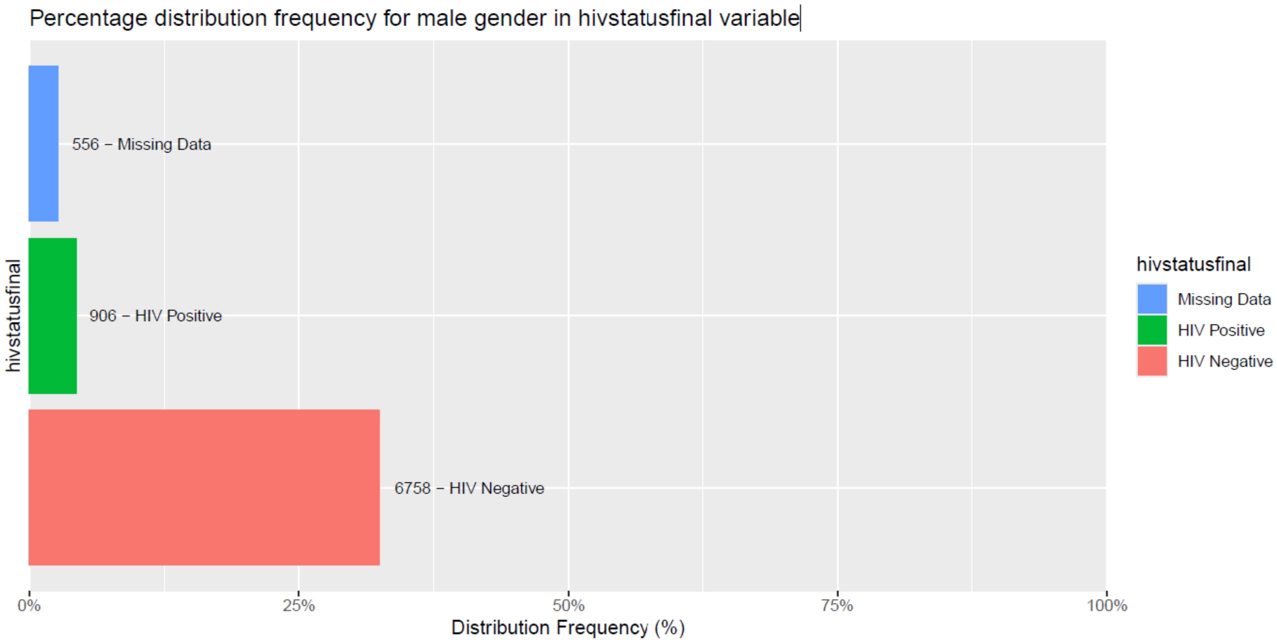


Figure 33. Percentage distribution frequency for male gender in hivstatusfinal variable

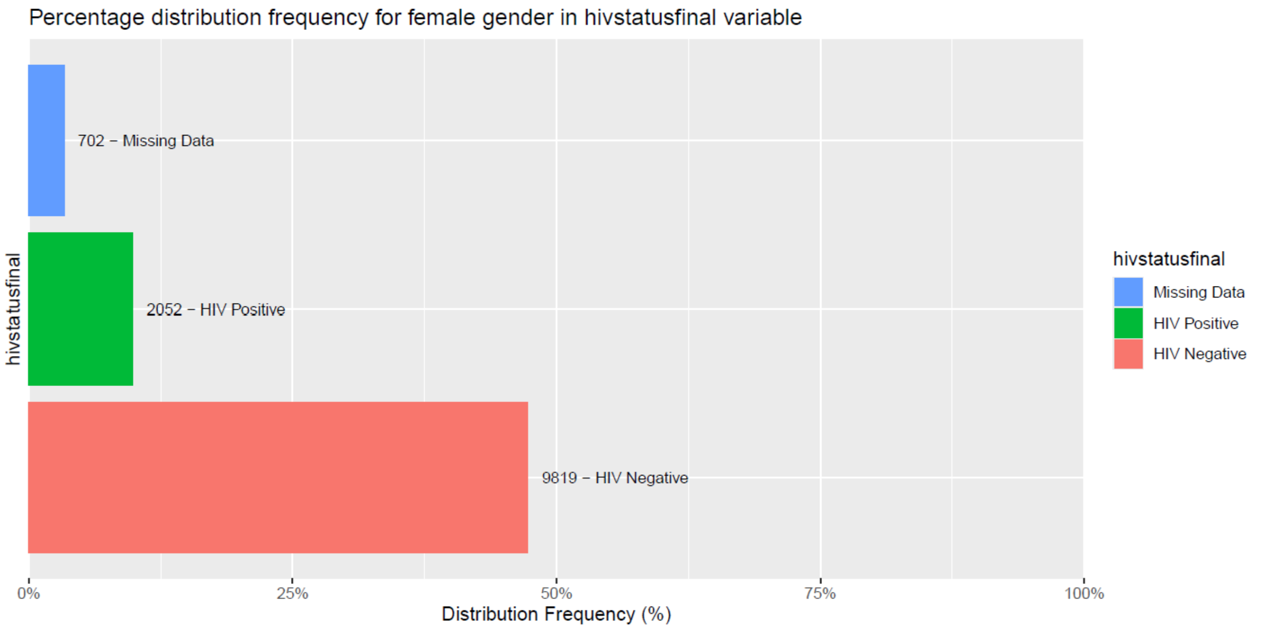


Figure 34. Percentage distribution frequency for female gender in hivstatusfinal variable

Figures 35 and 36 which represents the distribution of the age groups show similar pattern for both groups, with no missing data.

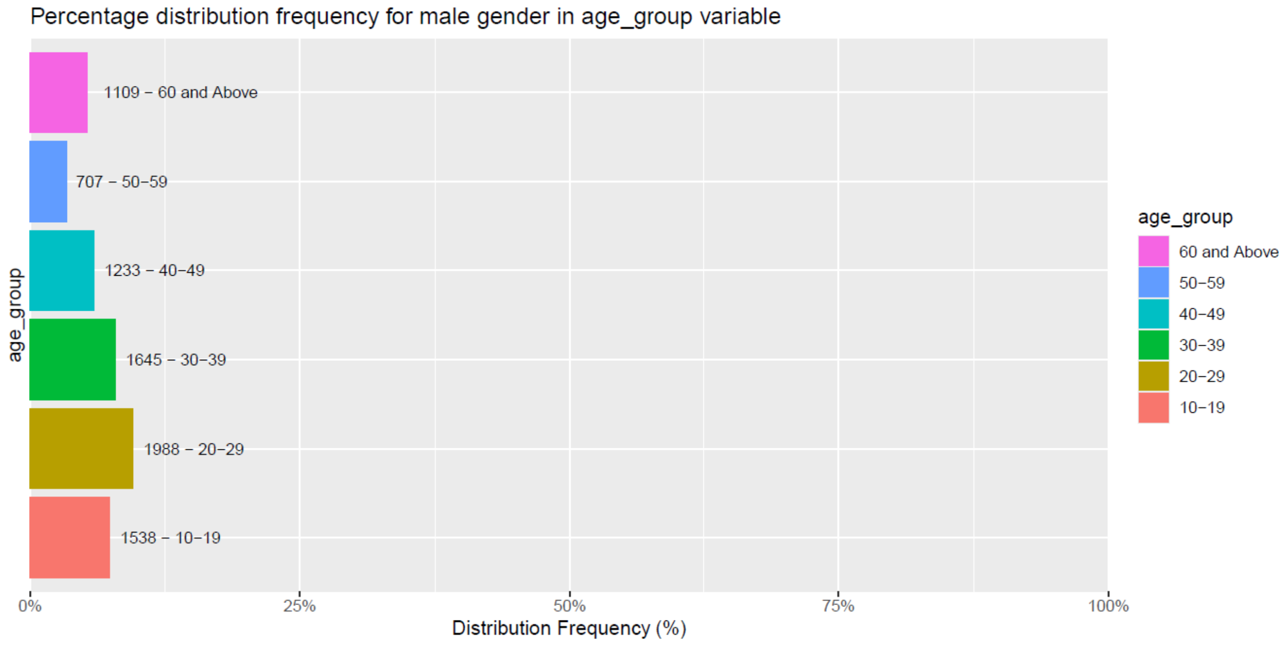


Figure 35. Percentage distribution frequency for male gender in age_group variable

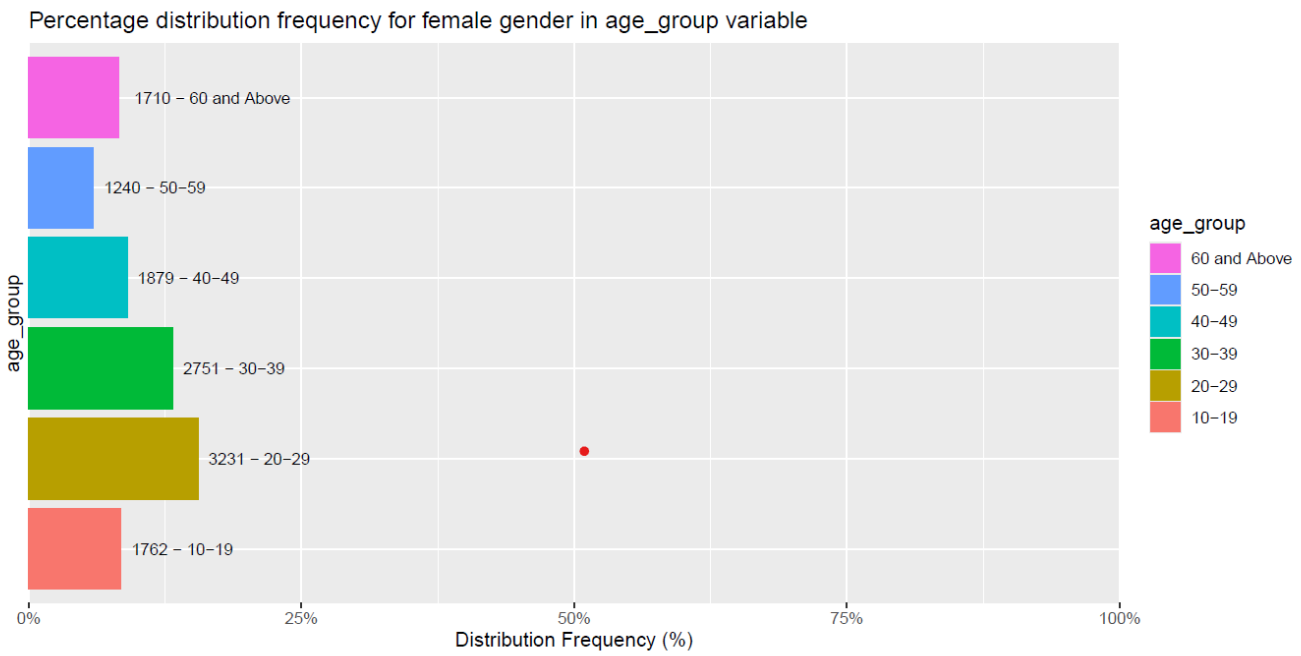
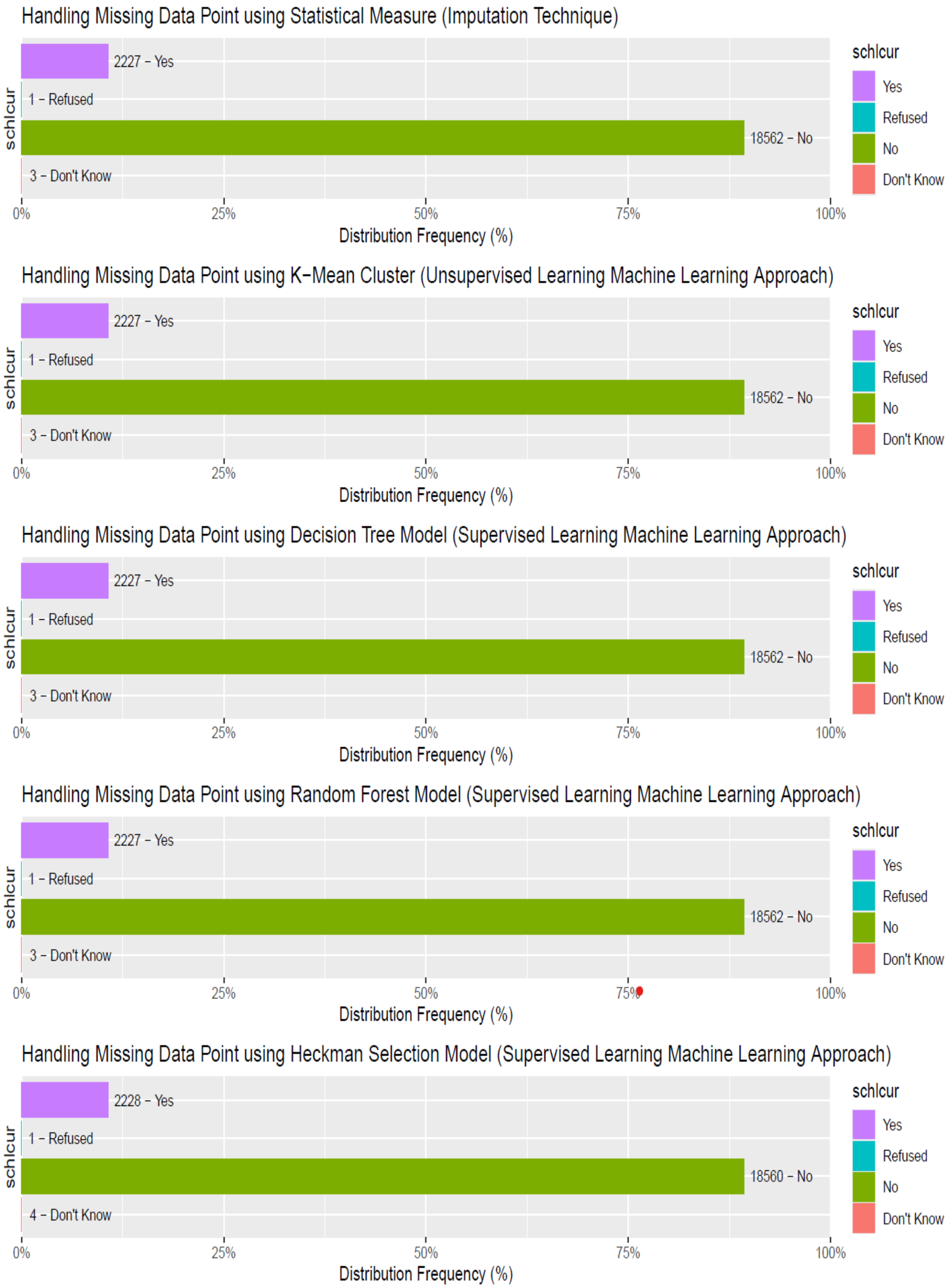


Figure 36. Percentage distribution frequency for female gender in age_group variable

4.1.1 Visualizing the Columns with Missing Data Points Corrected Using Different Techniques

The figures below (37 – 44) show the columns with missing data with imputed values. On the “schlcur” variable data was imputed for the female group. For the variables’ “education”, “curmar”, “sexever”, “hivtstrslt” and “hivstatusfinal” which have missing data in both groups, as illustrated in figures 38, 39, 41, 43 and 44, respectively. The missing data in “mcstatus” variable shown in figure 40 is only applicable to the male gender group. Missing data in “everpregnant” illustrated in figure 42 is only applicable to the female group. The images demonstrate the comparison in the imputed data between the models.



Handling Missing Data Point using Statistical Measure (Imputation Technique)

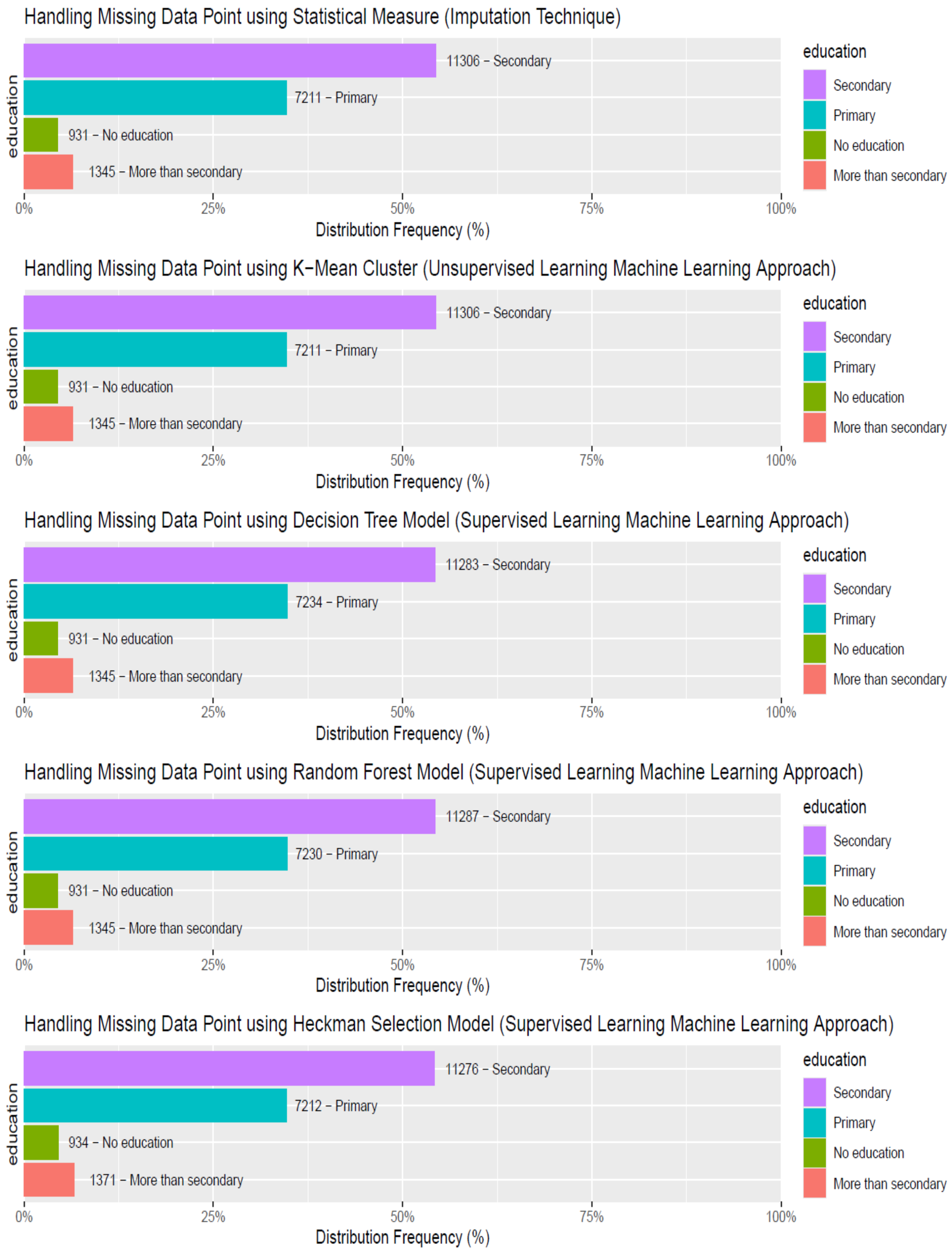
Handling Missing Data Point using K−Mean Cluster (Unsupervised Learning Machine Learning Approach)

Handling Missing Data Point using Decision Tree Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Random Forest Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Heckman Selection Model (Supervised Learning Machine Learning Approach)

Figure 37. Corrected Missing Data Point in schlcur Variable



Handling Missing Data Point using Statistical Measure (Imputation Technique)

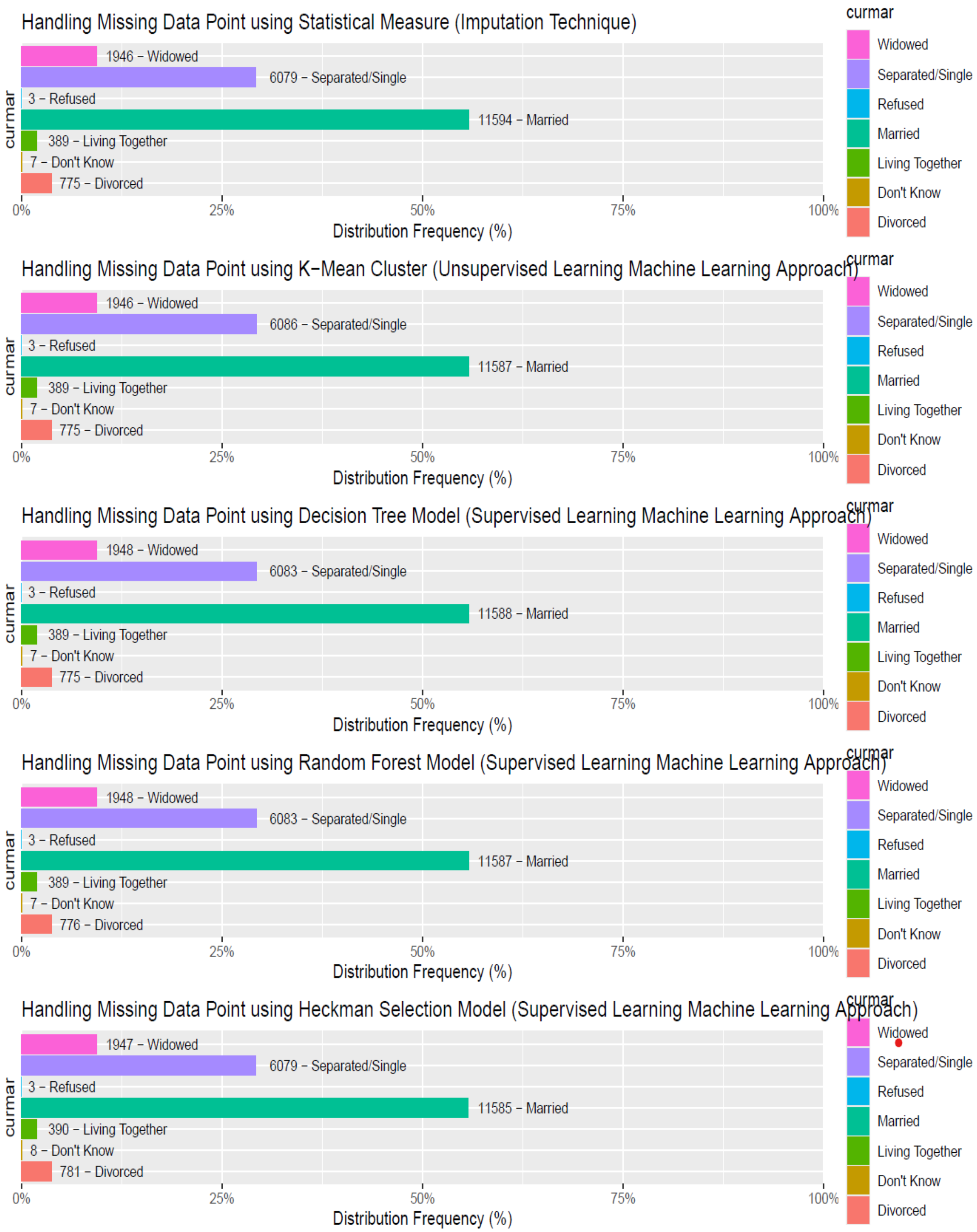
Handling Missing Data Point using K−Mean Cluster (Unsupervised Learning Machine Learning Approach)

Handling Missing Data Point using Decision Tree Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Random Forest Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Heckman Selection Model (Supervised Learning Machine Learning Approach)

Figure 38. Corrected Missing Data Point in education Variable



Handling Missing Data Point using Statistical Measure (Imputation Technique)

Handling Missing Data Point using K−Mean Cluster (Unsupervised Learning Machine Learning Approach)

Handling Missing Data Point using Decision Tree Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Random Forest Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Heckman Selection Model (Supervised Learning Machine Learning Approach)

Figure 39. Corrected Missing Data Point in curmar Variable

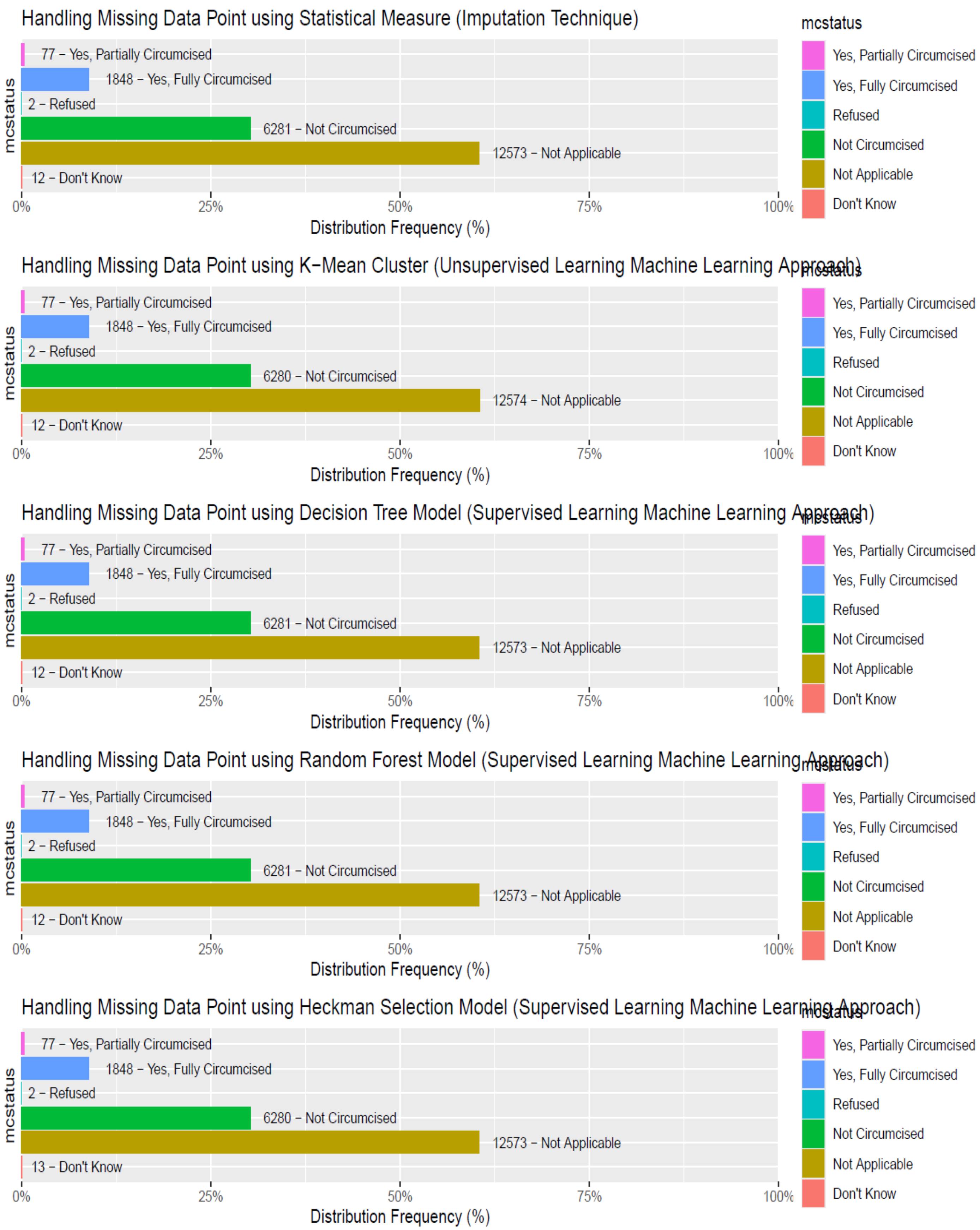
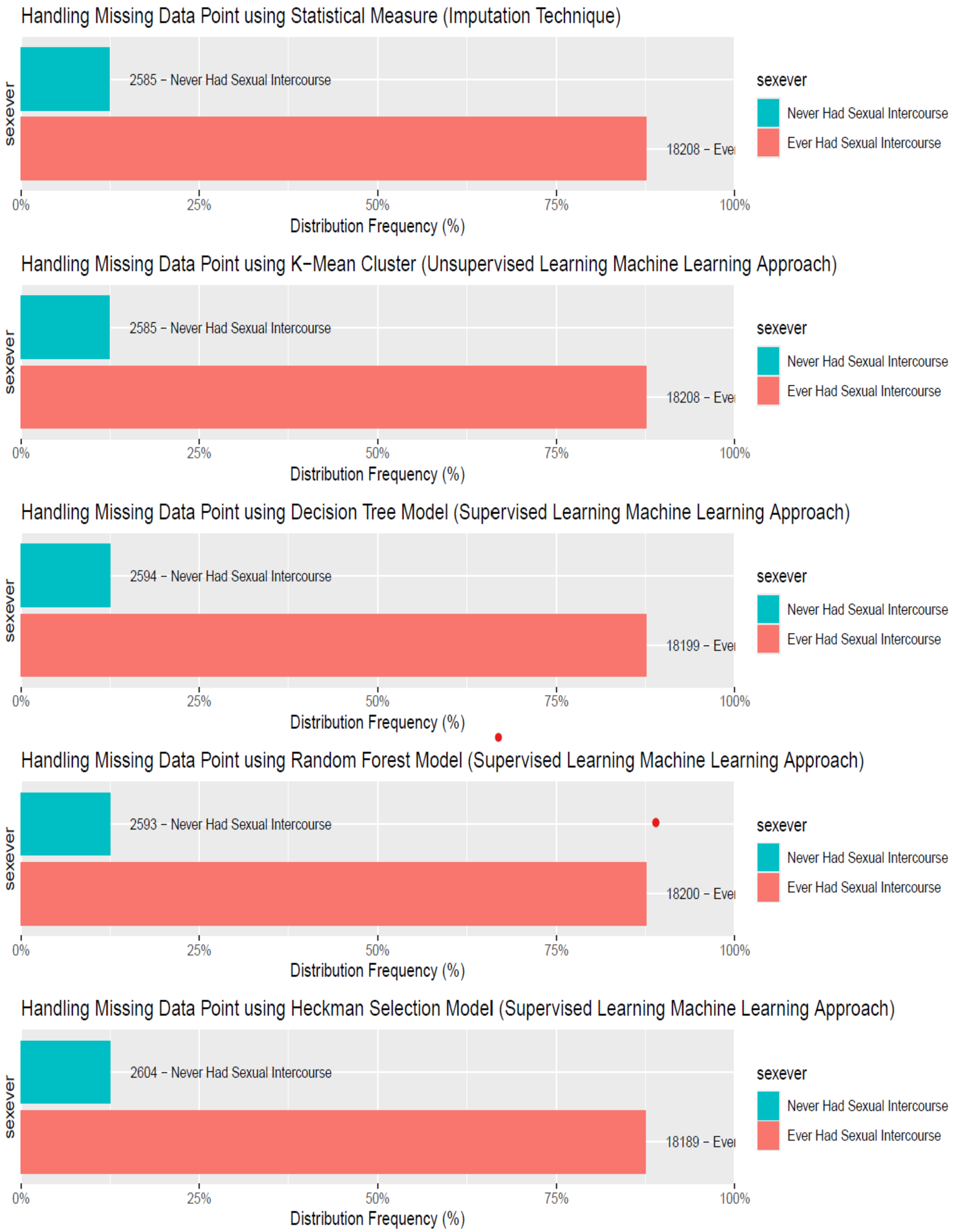


Figure 40. Corrected Missing Data Point in mcstatus Variable



Handling Missing Data Point using Statistical Measure (Imputation Technique)

Handling Missing Data Point using K−Mean Cluster (Unsupervised Learning Machine Learning Approach)

Handling Missing Data Point using Decision Tree Model (Supervised Learning Machine Learning Approach)

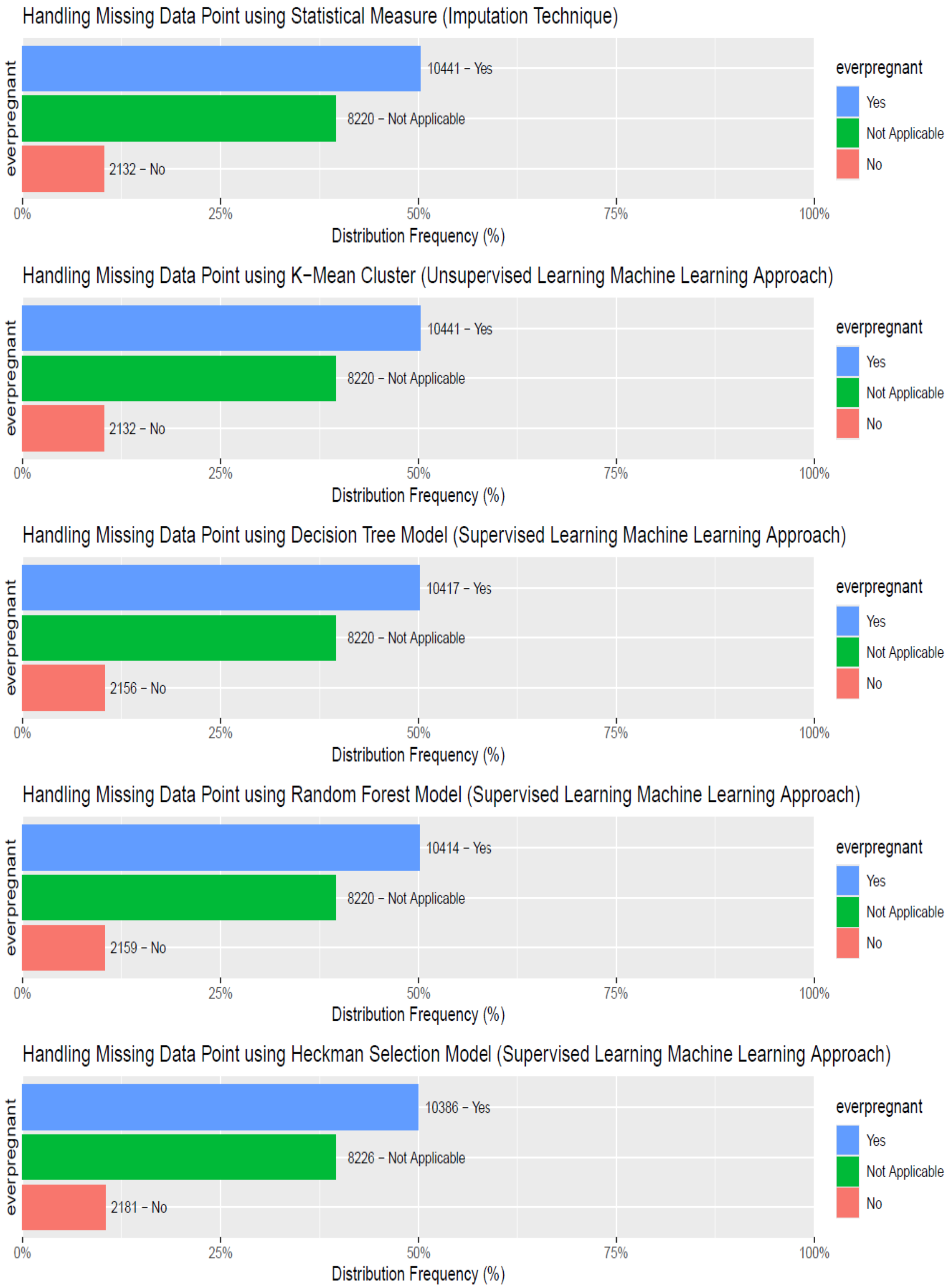
Handling Missing Data Point using Random Forest Model (Supervised Learning Machine Learning Approach)

sexever
2593 − Never Had Sexual Intercourse
18200 − Ever
0%
25%
50%
75%
100%
Distribution Frequency (%)
sexever
Never Had Sexual Intercourse
Ever Had Sexual Intercourse

Handling Missing Data Point using Heckman Selection Model (Supervised Learning Machine Learning Approach)

sexever
2604 − Never Had Sexual Intercourse
18189 − Ever
0%
25%
50%
75%
100%
Distribution Frequency (%)
sexever
Never Had Sexual Intercourse
Ever Had Sexual Intercourse

Figure 41. Corrected Missing Data Point in sexever Variable


Handling Missing Data Point using Statistical Measure (Imputation Technique)

Handling Missing Data Point using K−Mean Cluster (Unsupervised Learning Machine Learning Approach)

Handling Missing Data Point using Decision Tree Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Random Forest Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Heckman Selection Model (Supervised Learning Machine Learning Approach)

Figure 42. Corrected Missing Data Point in everpregnant Variable

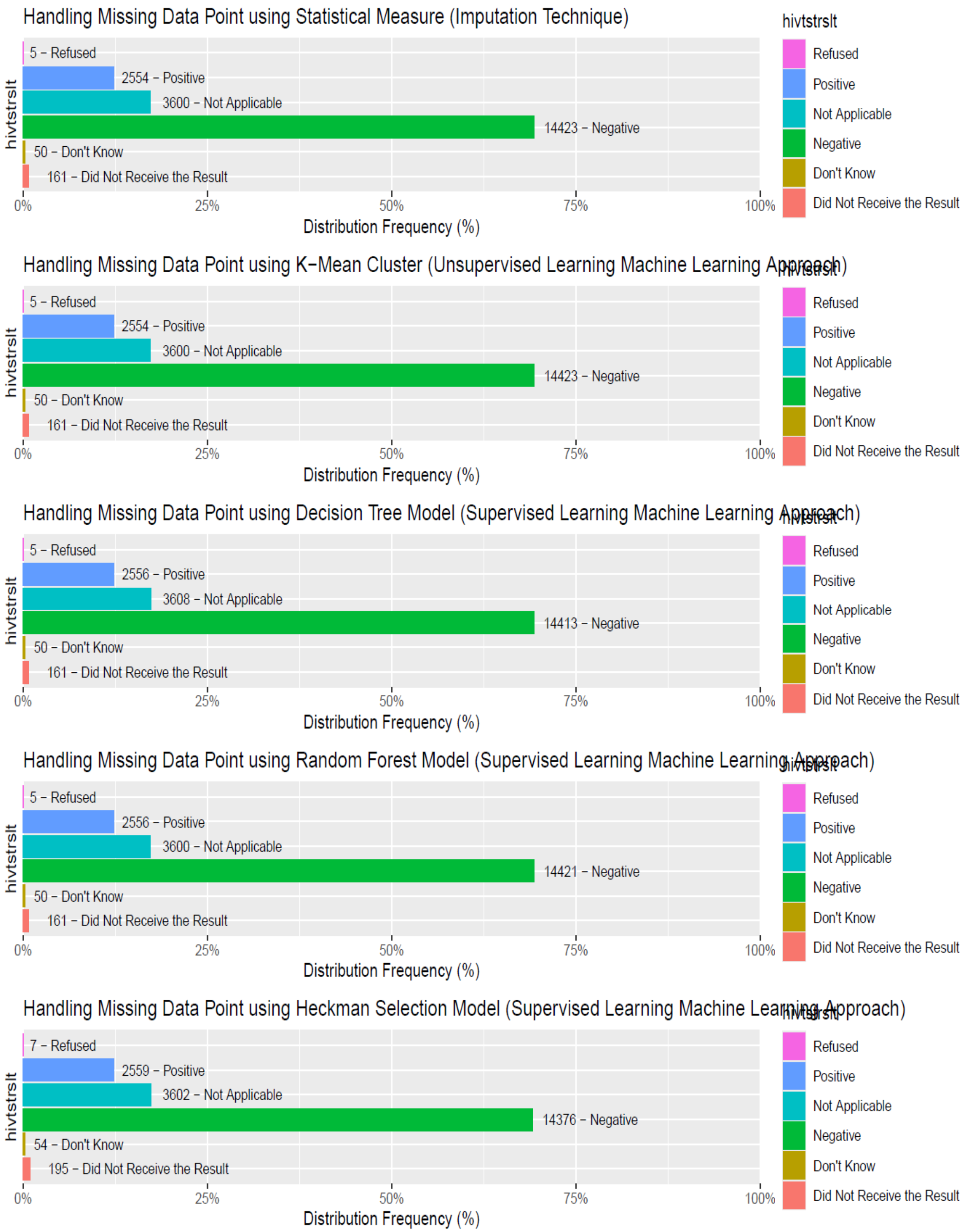
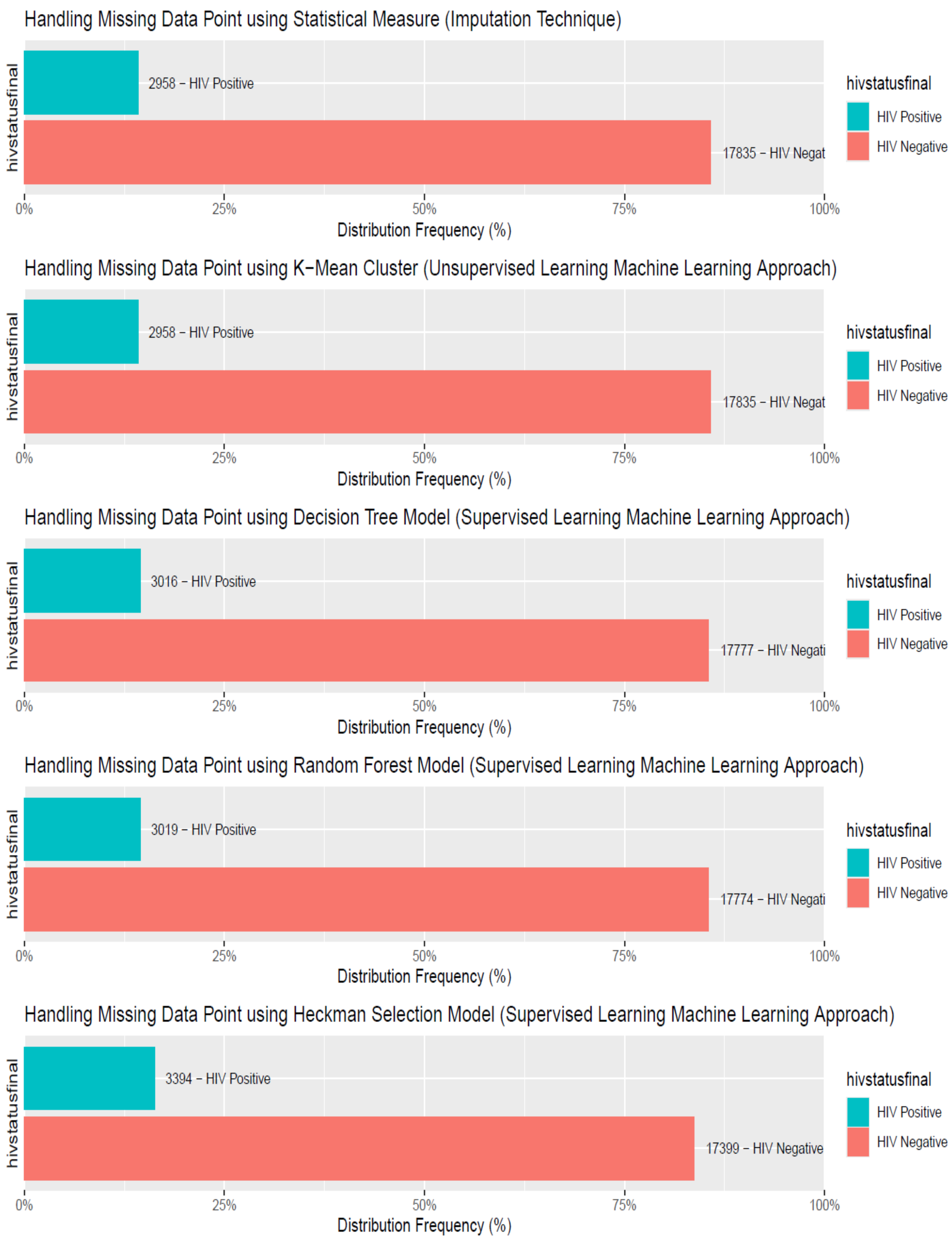


Figure 43. Corrected Missing Data Point in hivtstrslt Variable



Handling Missing Data Point using Statistical Measure (Imputation Technique)

Handling Missing Data Point using K−Mean Cluster (Unsupervised Learning Machine Learning Approach)

Handling Missing Data Point using Decision Tree Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Random Forest Model (Supervised Learning Machine Learning Approach)

Handling Missing Data Point using Heckman Selection Model (Supervised Learning Machine Learning Approach)

Figure 44. Corrected Missing Data Point in hivstatusfinal Variable

5 Discussion

The goal of the project is to predict missing data as close to its actual value as possible using machine learning. We employed random forest, K-Means cluster, the Heckman, Simple Imputation and Decision Tree. With a score of 97.70%, random forest is the most accurate model according to the results, demonstrating its greater capacity to categorize data with better accuracy even when some values are missing. This dominance is further exercised when evaluated on precision, sensitivity and specificity; recording scores of 99.69%, 97.68% and 97.82%, respectively. The Decision Tree approach comes in close behind at 97.65%, indicating that, though, not as successful as Random Forest in handling missing data, it is a reliable alternative. The Heckman model performed the least in all measures except in precision, against the simple imputation method. Its precision was 99.36% compared to simple imputation's 99.32%. K-Means cluster performed better than Heckman model and simple imputation method. This is illustrated in the evaluation summary in Table 2.

Table 2. Machine Learning Models' Evaluation Summary

	Accuracy	Sensitivity	Specificity	Precision
Decision Tree	97.65	97.68	97.42	99.63
Heckman	95.54	95.93	95.66	99.36
Imputation	97.38	97.68	95.23	99.32
K-Means Cluster	97.41	97.51	96.95	99.52
Random Forest	**97.70**	**97.68**	**97.82**	**99.69**

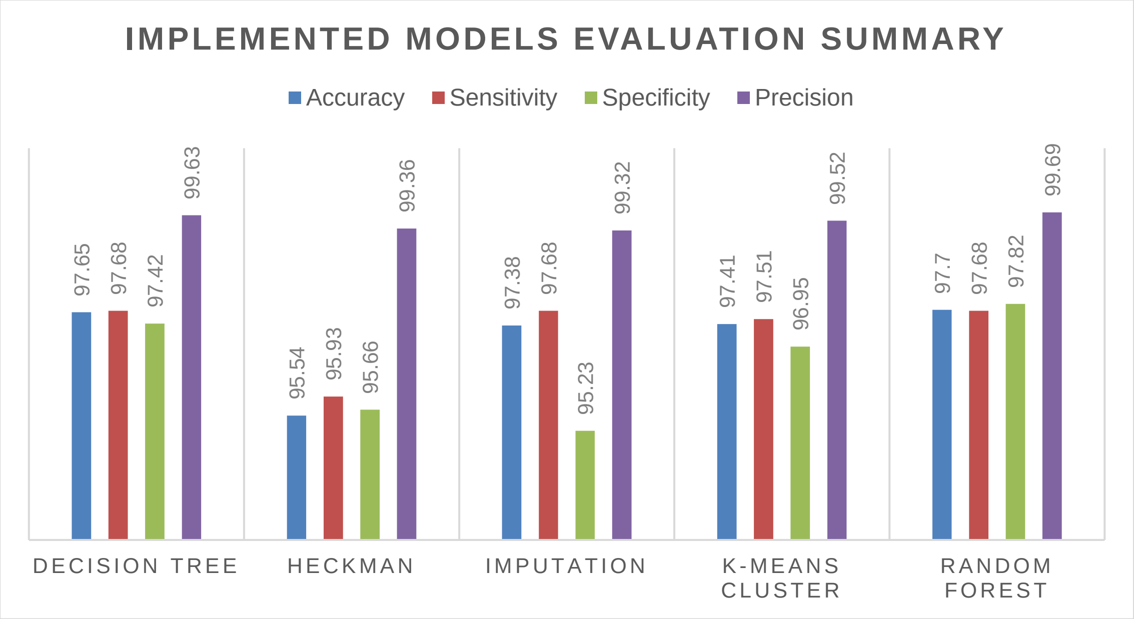


Figure 45. Implemented Model Evaluation Plot

In summary, the percentage difference between random forest and decision tree models is most pronounced, recording just 0.41% enhancement in specificity; suggesting that random forest has a little advantage over decision tree in preventing false positives. On the other hand, the random forest performs palpably better than the Heckman model, with gains of about 2.26% in specificity and accuracy, 1.82% in sensitivity, and 0.33% in precision. Suggesting that although random forest and decision tree are almost equal, random forest outperforms Heckman in more significant ways, especially when it comes to overall accuracy and having less errors with imputing missing data.

5.1 Recommendation for Handling Missing Data

Based on the outcome of this study, we believe that several steps could be adopted to lessen the effects of missing data.

- Iterative Process: Managing missing data is a continuous process. Evaluate your imputation strategy iteratively, make necessary adjustments, and compare the effects of various approaches on the performance of your model.

  To make sure the appropriate strategy is selected, experiment with different imputation techniques and verify them against baseline models.

- Transparency and Documentation: A thorough record of the methods utilized to address missing data and the underlying assumptions. Keep the workflow transparent, including any dataset transformations that are made. Keep a data processing report in which the steps used to address missing data are well-detailed.
- Special Considerations for Time-Series Data: Irregular intervals or missed records are frequently the cause of missing values. Make use of temporal pattern-accounting model-based algorithms, forward/backward fill, and interpolation. To fill in the gaps while maintaining trends, use time-based techniques like autoregressive models or spline interpolation.

6 Conclusion

In conclusion, the primary aim of this project was to enhance the accuracy and reliability of HIV/AIDS survey data by applying advanced machine learning algorithms to address and mitigate the effects of missing data. To achieve this, a comprehensive exploratory data analysis was conducted to gain a deeper understanding of the HIV dataset. The analysis revealed varying proportions of missing data, some of which could be accurately imputed. It also highlighted an imbalance in the dataset, particularly regarding gender distribution, although the variable distributions were largely similar across groups. Sample selection correction techniques were applied to manage the missing data.

The machine learning models used for imputing missing values included Random Forest, Decision Tree, K-Means Clustering, the Heckman selection model, and statistical imputation methods. These models were developed and applied to predict missing data. The study found that the dataset exhibited complexity at various levels, with missing data present in all but one column. The models were evaluated using metrics such as accuracy, specificity, sensitivity, and precision.

Empirical results demonstrated that the Random Forest model outperformed other supervised, unsupervised, and statistical techniques in terms of imputation accuracy, sensitivity, specificity, and precision under both MCAR (Missing Completely at Random) and MAR (Missing at Random) conditions. Random Forest proved particularly effective for clinical research dealing with classification and prediction problems where varying rates of missing data are present. The Decision Tree yielded the second-best performance, followed by K-Means Clustering, while the Heckman model and statistical imputation performed the worst.

A limitation of this study is the reliance on a single HIV dataset for training, testing, and optimizing the machine learning models. Finally, the study offers several recommendations to help reduce the occurrence of missing data and mitigate its effects on future analyses.

6.1 Future Work

Future research could explore the applicability of the proposed machine learning model, along with other deep learning techniques, in more complex settings (such as general practice) where infectious diseases are the focus, and a higher proportion of missing data is prevalent. Investigating this area would provide valuable insights into the robustness and effectiveness of machine learning approaches for missing data imputation when applied to more intricate and challenging datasets.

List of References

Agbo, B. *et al.* (2023) ‘Imputation of Missing Clinical Covariates for Downstream Classification Problems’, *IEEE Access*, 11, pp. 102935–102943. Available at: https://doi.org/10.1109/ACCESS.2023.3317775.

Alabadla, M. *et al.* (2022) ‘Systematic Review of Using Machine Learning in Imputing Missing Values’, *IEEE Access*. Institute of Electrical and Electronics Engineers Inc., pp. 44483–44502. Available at: https://doi.org/10.1109/ACCESS.2022.3160841.

Al-Jamali, N.A.S. *et al.* (2023) ‘A New Imputation Technique Based a Multi-Spike Neural Network to Handle Missing Data in the Internet of Things Network (IoT)’, *IEEE Access*, 11, pp. 112841–112850. Available at: https://doi.org/10.1109/ACCESS.2023.3323435.

Bärnighausen, T. *et al.* (2011) ‘Correcting HIV prevalence estimates for survey nonparticipation using heckman-type selection models’, *Epidemiology*, 22(1), pp. 27–35. Available at: https://doi.org/10.1097/EDE.0b013e3181ffa201.

Beaulac, C. and Rosenthal, J.S. (2020) ‘BEST: a decision tree algorithm that handles missing values’, *Computational Statistics*, 35(3), pp. 1001–1026. Available at: https://doi.org/10.1007/s00180-020-00987-z.

Bernardini, M. *et al.* (2023) ‘A novel missing data imputation approach based on clinical conditional Generative Adversarial Networks applied to EHR datasets’, *Computers in Biology and Medicine*, 163. Available at: https://doi.org/10.1016/j.compbiomed.2023.107188.

Figueroa-García, J.C., Neruda, R. and Hernandez-Pérez, G. (2023) ‘A genetic algorithm for multivariate missing data imputation’, *Information Sciences*, 619, pp. 947–967. Available at: https://doi.org/10.1016/j.ins.2022.11.037.

Gupta, S.K. *et al.* (2021) ‘A Machine Learning Approach for Heart Attack Prediction’, *International Journal of Engineering and Advanced Technology*, 10(6), pp. 124–134. Available at: https://doi.org/10.35940/ijeat.F3043.0810621.

Khan, W. *et al.* (2022) ‘Mixed Data Imputation Using Generative Adversarial Networks’, *IEEE Access*, 10, pp. 124475–124490. Available at: https://doi.org/10.1109/ACCESS.2022.3218067.

Kim, J.C. and Chung, K. (2020) ‘Multi-Modal Stacked Denoising Autoencoder for Handling Missing Data in Healthcare Big Data’, *IEEE Access*, 8, pp. 104933–104943. Available at: https://doi.org/10.1109/ACCESS.2020.2997255.

Kwak, S.K. and Kim, J.H. (2017) ‘Statistical data preparation: Management of missing values and outliers’, *Korean Journal of Anesthesiology*. Korean Society of Anesthesiologists, pp. 407–411. Available at: https://doi.org/10.4097/kjae.2017.70.4.407.

Lang, K.M. and Little, T.D. (2018) ‘Principled missing data treatments’, *Prevention Science*, 19(3), pp. 284–294. Available at: https://doi.org/10.1007/s11121-016-0644-5.

Mirzaei, A. *et al.* (2022) ‘Missing data in surveys: Key concepts, approaches, and applications’, *Research in Social and Administrative Pharmacy*, 18(2), pp. 2308–2316. Available at: https://doi.org/10.1016/j.sapharm.2021.03.009.

Patil, B.M., Joshi, R.C. and Toshniwal, D. (2010) *CCIS 94 - Missing Value Imputation Based on K-Mean Clustering with Weighted Distance*. Springer-Verlag.

Psychogyios, K. *et al.* (2023) ‘Missing Value Imputation Methods for Electronic Health Records’, *IEEE Access*, 11, pp. 21562–21574. Available at: https://doi.org/10.1109/ACCESS.2023.3251919.

Ramadhan, N.G. *et al.* (2024) ‘Chronic Diseases Prediction Using Machine Learning With Data Preprocessing Handling: A Critical Review’, *IEEE Access*, 12, pp. 80698–80730. Available at: https://doi.org/10.1109/ACCESS.2024.3406748.

Saroja, T. and Kalpana, Y. (2023) ‘Hybrid missing data imputation and novel weight convolution neural network classifier for chronic kidney disease diagnosis’, *Measurement: Sensors*, 27. Available at: https://doi.org/10.1016/j.measen.2023.100715.

Schwarcz, S. *et al.* (2006) ‘Late Diagnosis of HIV Infection’, *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 43(4), pp. 491–494. Available at: https://doi.org/10.1097/01.qai.0000243114.37035.de.

Sultan, N. *et al.* (2023) ‘Cesarean Section Classification Using Machine Learning With Feature Selection, Data Balancing, and Explainability’, *IEEE Access*, 11, pp. 84487–84499. Available at: https://doi.org/10.1109/ACCESS.2023.3303342.

Wells, B.J. *et al.* (2013) ‘Strategies for Handling Missing Data in Electronic Health Record Derived Data’, *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 1(3), p. 7. Available at: https://doi.org/10.13063/2327-9214.1035.

Xu, D. *et al.* (2020) ‘A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management’, *Journal of Biomedical Informatics*, 111. Available at: https://doi.org/10.1016/j.jbi.2020.103576.

Yoon, J., Jordon, J. and van der Schaar, M. (2018) ‘GAIN: Missing Data Imputation using Generative Adversarial Nets’. Available at: http://arxiv.org/abs/1806.02920.