Exploring Factors Affecting Contraception Use & Standard of Living in 1987 Indonesia

Authors: Rohan Chilukuri, Margaret Misyutina, Parham Rouzbahani

I. Abstract:

We wanted to explore the most significant factors influencing contraceptive use and standard of living in 1987 Indonesia. Our data is taken from a subset of the 1987 National Indonesia Contraceptive Prevalence Survey, and our exploratory data analysis lead us to believe that wife age and number of children were redundant features for predicting contraceptive use, and that husband occupation and husband education were the most important features to consider for predicting standard of living. Building a variety of models from the data, we found that wife age and number of children were not entirely redundant features in predicting contraceptive use, and that husband occupation and education were largely sufficient for predicting standard of living. Finally, we note the issues surrounding the dataset and our methods, and end with a suggestion for further avenues for analysis.

II. Introduction

The 1987 National Indonesia Contraceptive Prevalence Survey was conducted on married women who were either not pregnant or unaware of pregnancy at the time. Questions dealt with a variety of features, including education levels of both husband and wife, wife's religion, husband's occupation, media exposure, and standard of living. The goal of the survey was to predict contraceptive use based on these features, utilizing the woman's socio-economic and demographic characteristics. We chose this dataset to do our own analysis of how different features contribute to a couple's decision to use contraceptive as well as how a husband's education level and occupation affects the couple's overall standard of living.

III. Description of Data

Key Observations

Although the data includes numerical, categorical, and binary features, all attributes are labeled with integers. Numerical features such as wife age and number of children must be greater than or equal to 0, meanwhile others are required to be in a specific range (binary, 1-3, or 1-4). Wife religion, wife working, and media exposure fall into the binary category:

- Wife Religion: 0 = Non-Islam, 1 = Islam
- Wife Working: 0 = Yes, 1 = No
- Media Exposure: 0 = Good, 1 = Not Good

Wife education, husband education, husband occupation, and standard of living are categorical features from 1 to 4; the education levels and standard of living improve as the numbers increase, with 1 = low and 4 = high, meanwhile the occupations correspond as follows:

- 1 = professional, technical, and clerical, 2 = sales and services, 3 = manual, 4 = agricultural

Husband occupation is the only feature that decreases in quality while integer labels increase.
Finally, contraceptive is also categorical, labeled from 1-3:

- 1 = no use, 2 = long-term use, 3 = short-term use

Exploratory Data Analysis

We carried out EDA to inform and finetune the project's scope and exploratory questions - we were specifically interested in the effect of each feature on contraceptive use and standard of living. It is important to note that before we conducted any EDA, we cleaned our data in accordance with the data cleaning procedures outlined below. Our initial thought was to explore education as an influence on contraceptive use, since education has historically had a profound effect on making important life decisions. In this endeavour, we could not determine whether we should prioritize the husband's education, or the wife's education, and how they relate to each other and to contraceptive use. We therefore decided to plot the husband's education counts and wife's education counts to see any trends.

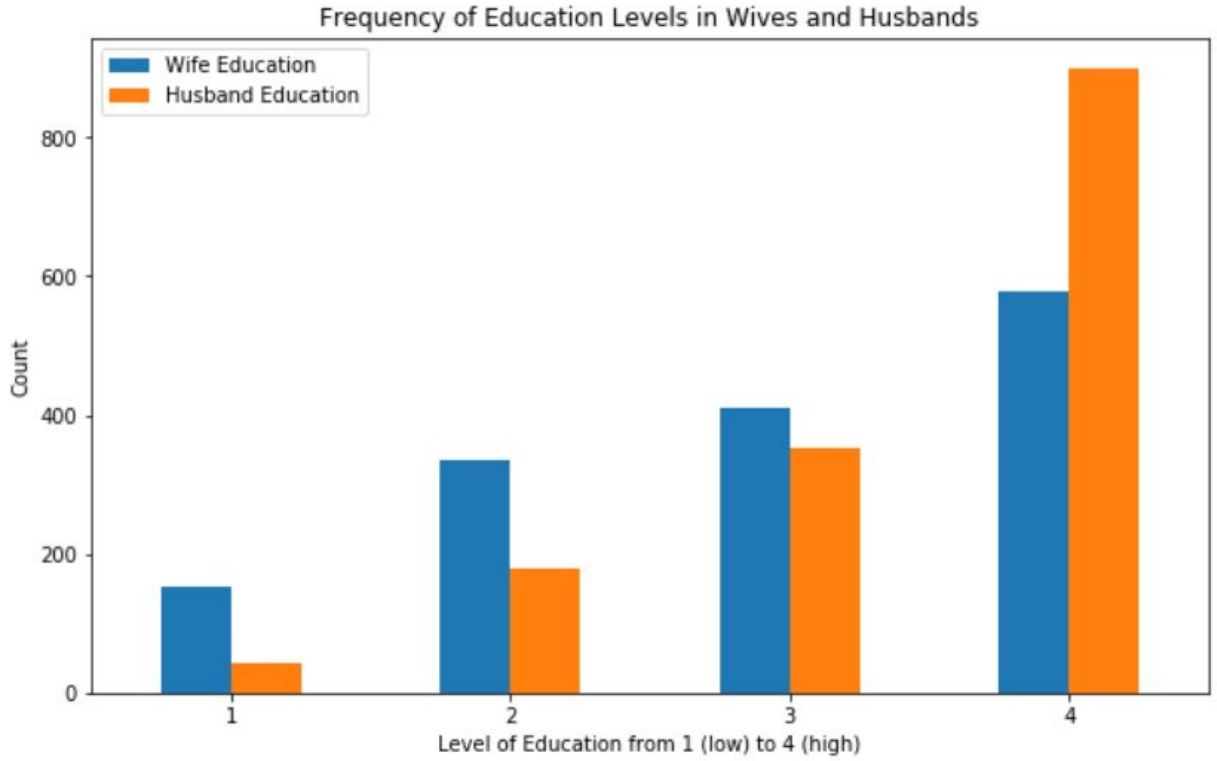

Fig 1. Barplot Displaying Husband Education Counts against Wife Education Counts

Note: There are an equal number of husbands and wives in this survey (1473).

From this plot, we immediately noticed an upwards trend in both cases, for wife education and for husband education, indicating that the counts within the level increase as the level of education increases. This means that higher education is more common in both roles. However, it is also apparent that husband education levels are overall much higher than the wives; over half of the husbands fall within the education level = 4 category and very few (~50/1473 = ~3%) of husbands fall within the education level = 1 category.

Moving forward, we decided to see if different education levels were associated with different kinds of contraceptive use. Our expectation was that a higher level of education would mean more knowledge of and information available about contraceptives. This could potentially lead to some use of contraceptive,

whether it be short-term or long-term. To visualize this, we made violin plots for both the wives and husbands education levels and their corresponding usage of contraceptives.

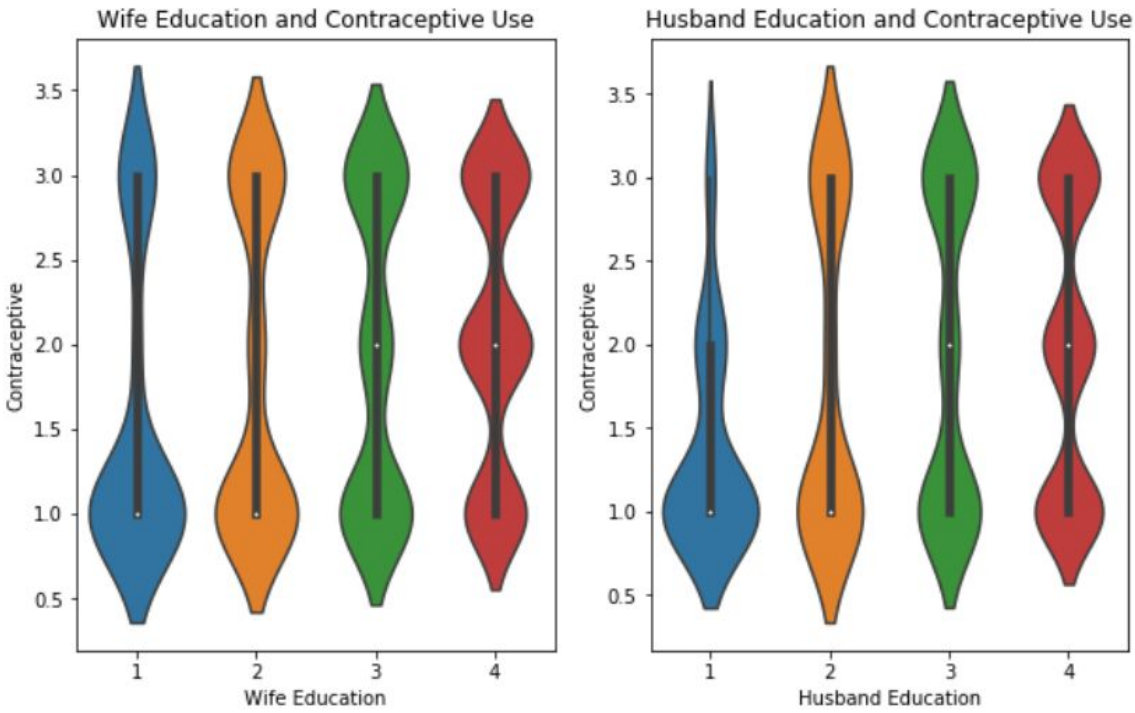

Fig 2. Violin Plots Displaying Education Levels in Wives and Husbands Against Use of Contraceptive

We can see that regardless of gender, an education level of 1 or 2 has a median contraceptive use of 1 (no use), denoted by the white dot, meanwhile an education level of 3 or 4 has a median contraceptive use of 2 (long-term use). This supports our initial prediction that low education levels would be associated with no contraceptive use and higher education levels would be more likely to use some sort of contraceptive.

Next we wanted to see how well different features correlated with one another, and particularly with contraceptive. We excluded education levels and occupation levels since we already touched on education levels above and we will elaborate on education levels and occupation against contraceptive below.

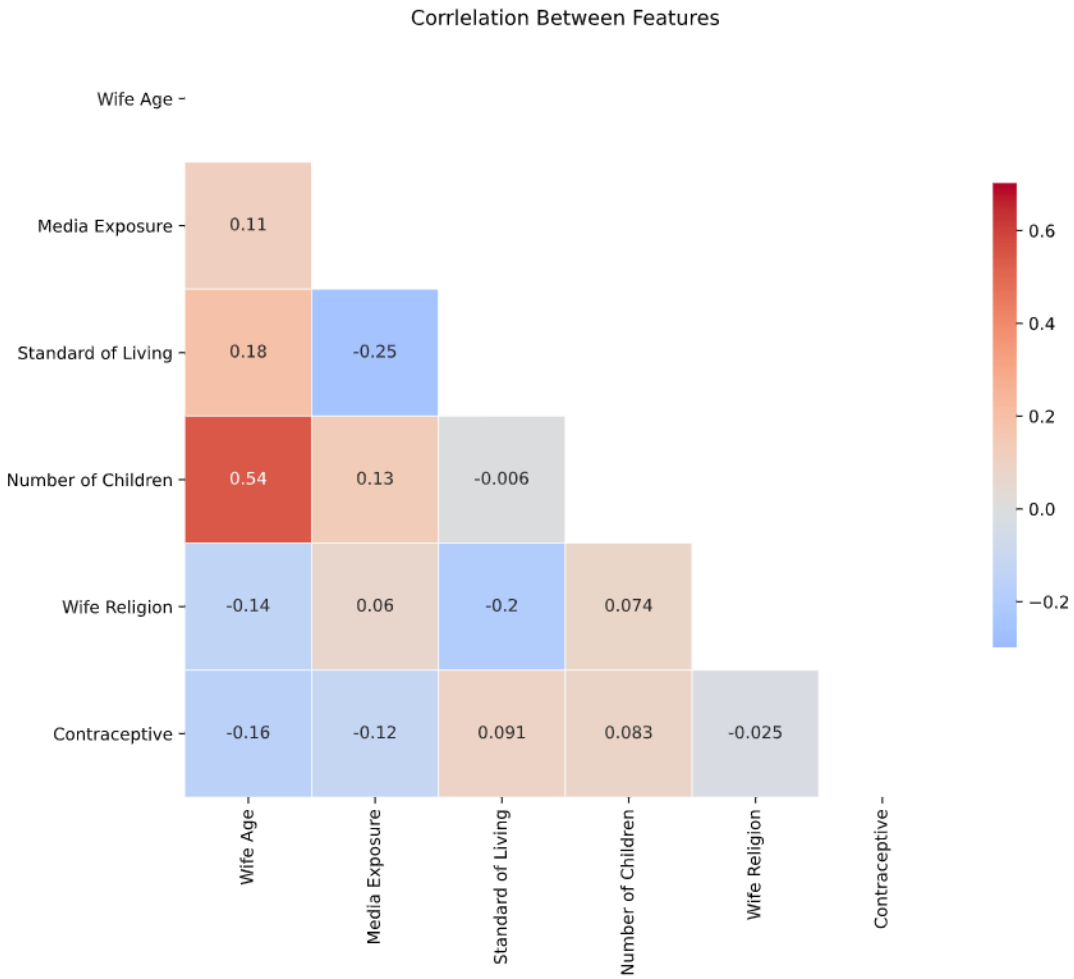

Fig 3. Heatmap Displaying Correlation Between Different Features

From this heatmap, we immediately notice that there is a high correlation between a wife's age and the number of children she has; this makes sense as typically the older you are, the more kids you have. Otherwise, none of the features have significantly strong correlations with one another.

Lastly, we wanted to approach the question of how a husband's education level and occupation affects the couple's overall standard of living. To do this, we decided to create a point plot of the husband's education against the standard of living and the husband's occupation against the standard of living.

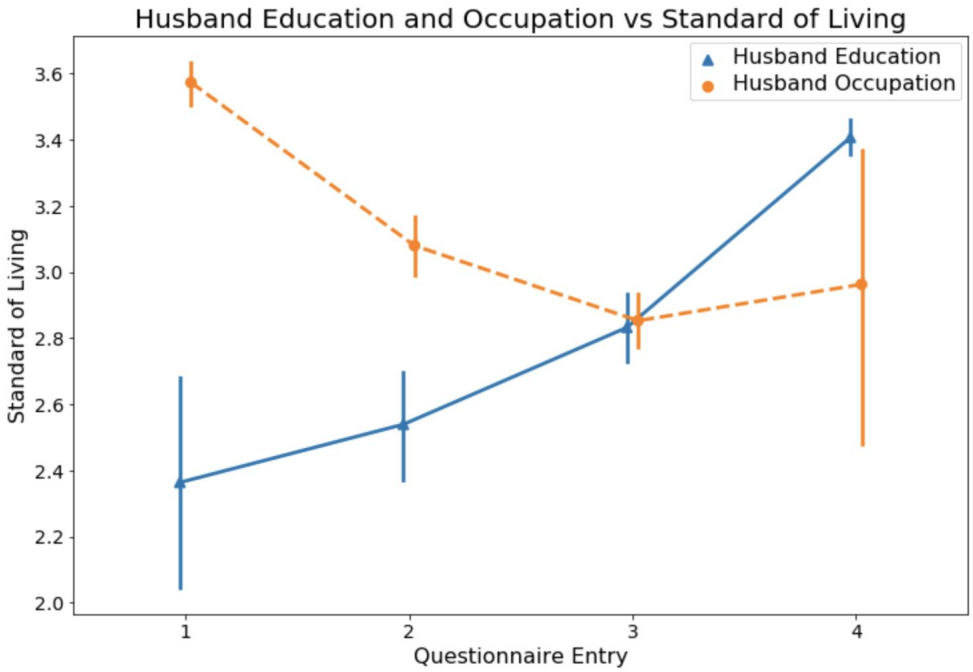

Fig 4. Pointplot of Husband Education and Occupation Against Standard of Living

This plot was particularly interesting due to the counterintuitive nature of the husband occupation labels (recall that an increase in husband occupation label means a decrease in job quality), resulting in an inverse relationship between the two in the graph. However, we can still see that as the husband education and occupation qualities increase, so does the standard of living. We also notice that the range of standard of living values across education levels is larger than across occupation levels, indicating that education has a bigger effect on the standard of living. Lastly, we see that there is a wider confidence interval for low education (~[2.1,2.7]) and narrower for high education (~3.4), as well as a narrower confidence interval for low occupation (~3.6) and wider for high occupation (~[2.5, 3.4]).

IV. Description of Methods

Data Cleaning and Transformations

The data was already in a workable form with no missing values. In case such missing values exist, we replaced them with the invalid value -1. We also converted all entries of the table to integers to make the data easier to work with in the future. Thus, we simply checked that each of the entries in each column were of the right format according to the valid values (given in the specification of the dataset). Finally, we renamed columns to make them more readable.

In order to prepare the data for numerical analysis (modeling), we one-hot-encoded each of the categorical variables, and standardized each of the numerical variables (these are performed for the usual reasons - magnitude considerations, numerical computation, etc.). Transformations were applied to variables as needed for each modeling task (different models require some subset of columns in some form: response variables are not changed, but other variables involved in computation are). This is

performed based on the results of the exploratory data analysis. We found that only a subset of variables seemed significant in predicting a category for the task of classification, and thus we incorporated such findings into the processing of the data.

Prediction Tasks

1. What effect does a variety of conditions (husband and wife education, wife age, media exposure, standard of living, number of children, and religion) have on contraceptive use?

The conclusion of the exploratory data analysis was that wife age and number of children were possibly redundant features. We want to test this hypothesis. As such, we trained three models associated with this question. One model includes all variables in the dataset as a control. The other two models remove one of wife age or number of children to determine the difference that excluding one of the features makes on the predictive capabilities of the model.

2. What is the relationship between husband education and occupation, and how does that affect standard of living?

The conclusion of the exploratory data analysis was that standard of living and husband education are correlated, and that standard of living and husband occupation are correlated as well. Thus, we generated two models. One model includes all variables in the dataset as a control. The other model includes only husband education and husband occupation to determine how effective these two factors are at determining the standard of living.

Modeling

Data is split into a train and test set (90-10 split). We train models on the training set and determine the effectiveness of the model based on the test set accuracy.

Based on our questions (predicting contraceptive use and standard of living, both categorical variables), we used appropriate models for categorical classification. The models included multi-class logistic regression, decision trees, random forests, K nearest neighbors, and support vector machines since these are all typically used for the classifying categorical data.

Hyper-parameters are chosen through sklearn's built-in LogisticRegressionCV for the logistic regression model, or through 5-fold cross-validation with sklearn's GridSearchCV. A set of possible hyper-parameters are selected for each model: regularization strength for logistic regression (note: automatic), maximum depth for decision trees and random forests, number of neighbors for k-nearest neighbors, and regularization strength and kernel coefficient for the support vector classifier.

The final model for a given prediction task is chosen by performing 10-fold cross validation (scored by accuracy) on each of the models, and selecting the model with the highest average cross validation score across the 10 folds. The chosen model is then trained on the entire training set, and is then evaluated by predicting labels of the test set and calculating the accuracy.

V. Summary of Results

Train/Test Accuracy:

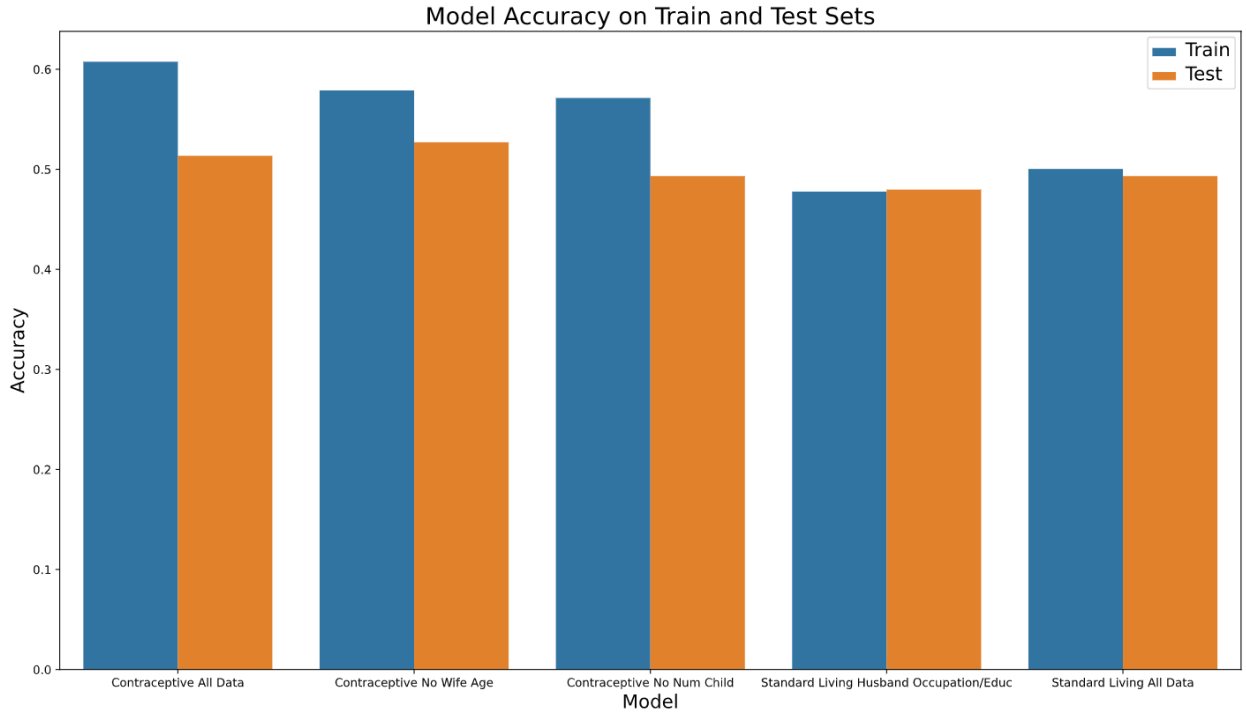

Fig 5. Barplot comparing train and test set accuracy for each model

Chosen Models:

Contraceptive All Data: Support Vector Classifier
Contraceptive No Wife Age: Support Vector Classifier
Contraceptive No Num Child: Decision Tree Classifier
Standard Living Husband Occupation/Education: Logistic Regression Classifier
Standard Living All Data: Logistic Regression Classifier

Note: Chosen models and accuracy are based on the train-test split. Models could have performed better or worse based on the data included on each side of the split, and different models could have been selected as the "best" for the classification task. Cross validation is meant to reduce this possibility, but the randomness of the split can result in a subset that fails to accurately reflect the underlying structure of the data.

VI. Discussion

Conclusions

For the first question, based on the accuracy, it appears that wife age and number of children (the features that interested us the most based on the EDA for this question) are not completely correlated with each other. The accuracy number of removing either wife age or number of children are not the same.

Surprisingly, and likely due to randomness in the model selection, the accuracy of predicting contraceptive use without wife age was slightly more accurate than the accuracy with all variables. The accuracy of predicting contraceptive use without the number of children was much lower than the other two models for this task. This suggests that the number of children is more indicative of contraceptive use than wife age, and that wife age cannot fully explain the number of children.

For the second question, based on the small difference in accuracy between the models, it appears that much of standard of living can be explained by the husband occupation and husband occupation (the features that interested us the most based on the EDA for this question). The accuracy of the model with only two features was within three percent of the accuracy of the model with all features. This indicates that the two features were significant in predicting standard of living.

Challenges

While working through the exploratory data analysis, we believed that religion would play a large role in determining contraceptive use. However, it turned out to have a very low correlation coefficient compared to other features. Additionally, finding graphs that worked with our data proved to be more challenging than expected since all points were integer values. This meant that there was no fluidity between the points, only gaps between them.

Evaluation of Approach/Limitations/Assumptions:

Our approach was to test our hypotheses by training models and comparing accuracy scores. The models were formed based on the exploratory data analysis. In general, this approach gave us rough conclusions for our questions. However, accuracy scores are limited in their explanatory scope and only provide rough estimates of the efficacy of the models. This is a minor consideration compared to the fact that the data is from a survey where response bias could severely influence the results of the models. Also, the data is from a subset of the survey, and we are unsure if it accurately reflects the overall dataset, which is another limitation of the data. Finally, the dataset is from 1987, so any analysis may not be an accurate reflection of current Indonesia.

Clearly, we assumed that the dataset was a representative sample of the overall population of Indonesia. We also assumed that raw accuracy scores were a sufficient metric to evaluate the capabilities of our models. Obviously, these assumptions could be incorrect.

Ethics

One major ethical concern that we encountered was in the wording of the survey itself. While reading specific questions, we realized that several of them came across as blunt and inconsiderate of the audience, which could lead to response bias as the respondent will want to avoid embarrassment. An example of this was question 207 of the survey, which bluntly asks “how many [of your] boys have died?” A fix to this would be to be considerate of the audience’s perspectives while designing the survey. Another similar ethical issue was the lack of an option not to answer a question, which explains no missing values. Some questions probe sensitive subjects, such as religion, and may have severe implications for the respondent. The fix to this would be to include an option to skip a question.

VII. Next Steps

Some further considerations of this dataset include combining different factors than the ones we used to predict contraceptive use and standard of living, as well as potentially predicting other categories. Since this is a subset of the wider survey that includes more variables and persons surveyed, one could also use such additions to create more robust, predictive models. Finally, comparing this dataset to another similar one from another country to compare contraceptive use across countries, or to a similar dataset from a different time period to compare contraceptive use over time are both possible considerations for further analysis.