Application of Data Analysis and Modelling to Predict Contraceptive Usage

Edward Han, Bryan Ho, Robert Jiang

Abstract

Utilizing data analysis techniques, we were able to create an adequate model that can predict contraceptive usage based on a variety of features and attributes of a married woman. We investigated the data itself to study any potential relationships between the listed features and then incorporated the data into various models, of which logistic regression was the best with a test accuracy of 52.88%, specifically 63.97% for no contraceptive usage, 30.88% for short-term contraceptive usage, and 52.75% for long-term contraceptive usage. Further analysis, however, showed that it is challenging to improve our accuracy due to each contraceptive response uniquely favoring certain features over others as well as the general messiness of the dataset. In the end, we determine that logistic regression is the most suitable model and that feature reduction/selection would be very challenging based on Lasso and PCA analysis.

Introduction

For the final project, we will be analyzing a dataset from the 1987 National Indonesia Contraceptive Prevalence Survey conducted on married women about their contraceptive method of choice and their household, demographic, and socio-economic background. We will attempt to isolate any statistically significant features and construct a model of our own with the features in the dataset to predict contraceptive method usage.

Beforehand, we used one-hot-encoding to convert categorical variables into binary variables for a usable statistic in our methods. To find our features, we did some exploratory data analysis using pair plots and heat maps. Pair plots are a collection of subplots that compare several subsets of data to each other with a distribution plot to reveal any explicit relationship between data subsets. For heat maps, we considered 2 variations: Cramer's V, which is based off of chi-square tests and outputs a nonnegative association value between [0,1] for each pair of variables, and Pearson's correlation, which is based off of covariance and returns a magnitude and direction of correlation.

Afterwards, we implemented the bulk of calculations by generating models using logistic regression, linear regression, decision tree, and random forest models with varying degrees of accuracy. For the regression models, we primarily used Lasso regularization to best improve our accuracy and to appropriately scale our features. The Lasso feature selection originates from the Linear Regression class and forces certain features to have a coefficient value of zero; useful for

simplifying the model to prevent overfitting the data and ranking the features by relationship to the outcome. We then observe the effects of all features on our models in order to definitively narrow down which features are essential. From there, we expanded upon our Lasso feature selection with a PCA analysis to further determine the importance of each of the features with respect to each contraceptive. Finally we used our best model, the logistic regression, to predict labels for the type of contraceptive used per respondent.

Description of Data

The dataset was originally used for predicting the contraceptive method using only the limited background info provided by respondents. A study on the dataset states that the survey was recorded over a four month period from mid-September to the third week of December and that the number of people selected was based on the size of the province. In order to maximize data, people were notified beforehand resulting in 98.5% of people selected being successfully interviewed. The dataset consists of a total of 10 features for each respondent sampled from the total populus of Indonesia across all of its provinces. These features consist of 2 numerical variables: wife's age and number of children born, 3 binary variables: Wife's religion, Wife's current employment status, and media exposure, as well as 5 categorical variables: wife's education, husband's education, husband's occupation, standard-of-living index, and contraceptive method. For the rest of the paper, we will regard both the binary and categorical variables as categorical variables. These are further defined in **Table I**.

Table I: Features of Our Data Set

Feature	**Type**	**Description**
Wife's age	Numerical	Wife's age
Wife's education	Categorical	1=low, 2, 3, 4=high
Husband's education	Categorical	1=low, 2, 3, 4=high
Number of Children	Numerical	Number of Children
Wife's Religion	Binary	0=Non-Islam, 1=Islam
Wife's Employment	Binary	0=Yes, 1=No
Husband's Occupation	Categorical	1=Professional, Technical, Clerical; 2=Sales, Services; 3=Manual, 4=Agriculture
Standard-of-living Index	Categorical	1=low, 2, 3, 4=high
Media Exposure	Binary	0=Good, 1=Not good
Contraceptive Method	Categorical	1=No-use, 2=Long-term, 3=Short-term

Before analyzing the data, we made sure that there were no nonsensical data values by removing all entries with either a negative or null value in any category (and luckily, there were none). We then standardized our data by first normalizing the numerical variables so that each set had a mean of 0 and standard deviation of 1 and one-hot-encoding the categorical variables. The former is necessary because it permits us to use Lasso regression for feature selection as well as make each feature more interpretable since they are now on the same scale. The latter is necessary because, although many of the categorical variables are "numbered", our model should not weigh them based on the magnitude of each variable, but rather whether or not that variable is present. For instance, the Husband's Occupation variable consists of four different occupations; each would have to be weighed individually based on the label, since a sales job would differ from an agricultural job.

Exploratory Data Analysis

We first examined the demographic of our population, and noted the following about our dataset from **Appendix I, Figure 1**:

1. Contraceptive 1 is most popular, followed by contraceptive 3 and then 2.
2. Wife age population peaks around age 20-30, although there are a significant number of women outside of this range.
3. The husbands tended to be more highly educated. The wives also tended to be more highly educated, but this distribution was slightly more uniform.
4. The number of children tended to be in the smaller range, 0-5.
5. The wife population mostly consisted of Muslim women.
6. The wives tended to not work.
7. The husbands tended to have a certain job (3= Manual)
8. The standard of living tended to be more high.
9. Media exposure tended to be present.

Because we investigated first how these variables affected our response variable, contraceptives, we also examined the demographic of each sub-population (sorted by contraceptive type, found in **Appendix I, Figures 2, 3a, and 3b**). While most of the trends of the overall population held true within the sub-populations, there were some noticeable differences:

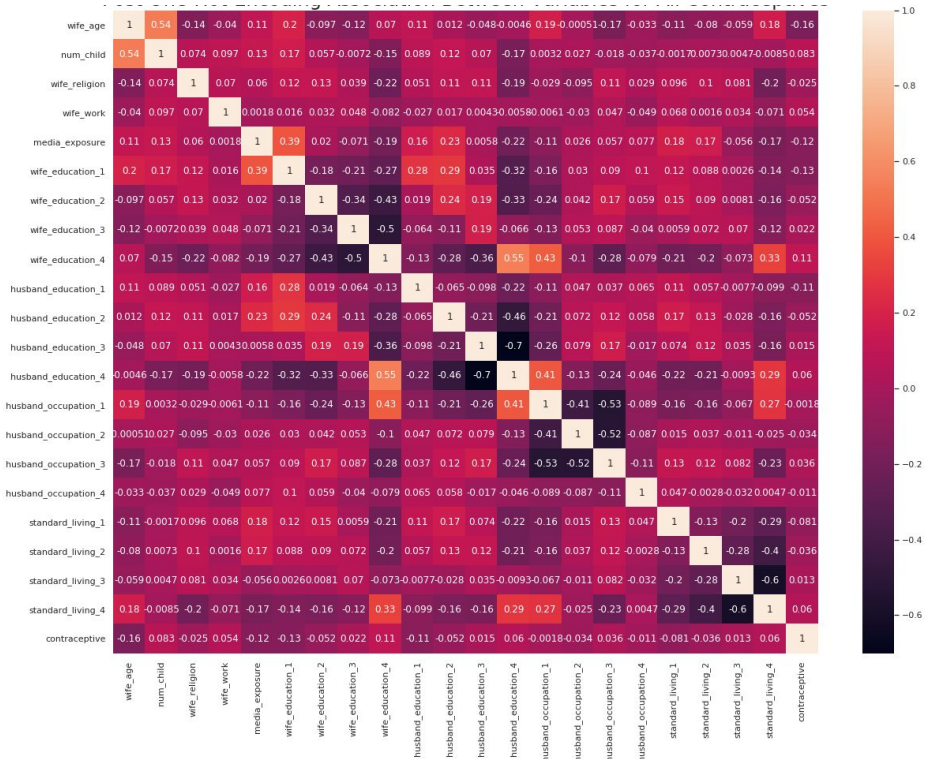
1. Wife age for contraceptives 2 and 3 were unimodally distributed, with modes at about ages 30 and 34 respectively, whereas contraceptive 1 was bimodally distributed, with modes at about ages 25 and 45 respectively.
2. num_child for contraceptives 1, 2, and 3 were unimodally distributed, with modes at about 1, 2.5, and 3 children. However, all contraceptives were right-skewed for this feature.
3. Contraceptive 1 had a much more balanced wife education distribution compared to contraceptives 2 & 3.
4. Contraceptives 1 & 3 had the most husbands belonging to husband occupation 3, then 2, 1, and 4, but contraceptive 2 had the most husbands belonging to husband occupation 1, then 2, 3, and 4.

To investigate further trends in the data, we examined possible relationships and correlations between numerical variables. We first modeled the linear relationship between number of children and wife age (**Appendix I, Figure 4**) for each contraceptive, and found that the number of children women have was on average slightly higher for all ages 15-50 for women who used contraceptives 2 & 3, while women who used no contraceptive (contraceptive 1) tended to have less children for their age group.

We then did the same for categorical variables (using our original data before one-hot-encoding). Naively, we began by creating pairplots of each pair of features for each contraceptive type, with a small amount of noise mixed in to each value for better visual clarity (**Appendix I, Figure 5a**). However, as is evident from our investigation of wife work vs. wife religion (**Appendix I, Figure 5b**) these plots were overall very similar and hard to contrast; therefore, we elected to compare heat maps of each feature pair correlation instead.

To begin, we first plotted the correlational heat map of all 10 features, resulting in **Figure 1**. While we also used correlational heat maps based on Cramer's V (**Appendix I, Figure 8**) instead of Pearson's correlation, we found that Pearson's correlation provided the most insight and thus continued with one-hot-encoded, regularized, Pearson's correlation heat maps.

Figure 1: Heat Map of Correlations Between All Variables



We noticed that many of these variables do not have a strong correlation, in particular with contraceptives. However, this does not at all mean that these features are useless in building a model - it simply indicates that linear regression, which correlation applies, cannot properly

capture the relationship between contraceptive type and numerical/categorical variables. This gives us a good reason to use another model, such as logistical regression.

Nonetheless, capturing the differences in feature correlations between each contraceptive type may provide insight in how well a model may distinguish between contraceptives. We therefore plotted the correlational heat map for each contraceptive type in (**Appendix I, Figure 6b, c, and d)**. Examining the heat maps, we concluded that contraceptive 2 & 3 feature correlations are remarkably similar, while contraceptive 1 feature correlations are noticeably different than 2 & 3. We can therefore expect a model to somewhat distinguish between contraceptive 1 vs. contraceptives 2 & 3, and have a harder time distinguishing between contraceptive 2 vs. contraceptive 3.

Methods

Due to the nature of this dataset, we are facing a multi-classification problem, as there are three possible choices each of the wives had--contraceptives 1, 2, and 3. We can establish the contraceptive method choice as a response variable changing in accordance with the other features and use classification models to predict the response variable numerically.

Looking at our data, we saw that there were two primary variable types, categorical and numerical. To ensure that every feature was on the same scale, we kept the standardization and one-hot-encoding transformations we had performed prior, so as to ensure that each feature was measured appropriately.

For our first model, we decided that logistic regression would be most appropriate for numerically estimating the probabilities of each contraceptive response based on the features. Since we concluded so far that each feature had some contribution to the response (ie. none of the data was “useless”), the logistic regression would take into account all the separate features and calculate multiclass probabilities for each of the response labels. For our specific model, we chose to utilize “one vs. rest” calculations, since doing so would isolate each contraceptive response. Additionally, we conducted 5-fold cross-validation and utilized varying regularization strengths (Lasso) to improve our overall accuracy and enhance variable selection.

The next models we constructed were decision trees and random forests. These will similarly estimate the response variables by individually comparing the attributes of each of the features with one another. Accumulating these comparisons, probabilities for each of the responses are used to predict the resulting label. For the random forest, we utilized 100 estimators.

Finally, our last model was a linear regression model. Since contraceptive method choice can not be linearly estimated (it is a categorical variable), we instead changed our focus to the secondary response variable, the number of children each wife had. Going this route, the contraceptive variable would then be used as a feature and thus have to be one-hot encoded. The number of children can appropriately be left standardized, since we want to use it as a numerical response measurement.

For each of our four models, we used training-test splits to best calibrate and test our models, using a 80-train-20-test configuration. This allows us to use our data to create the models and subsequently test its accuracy.

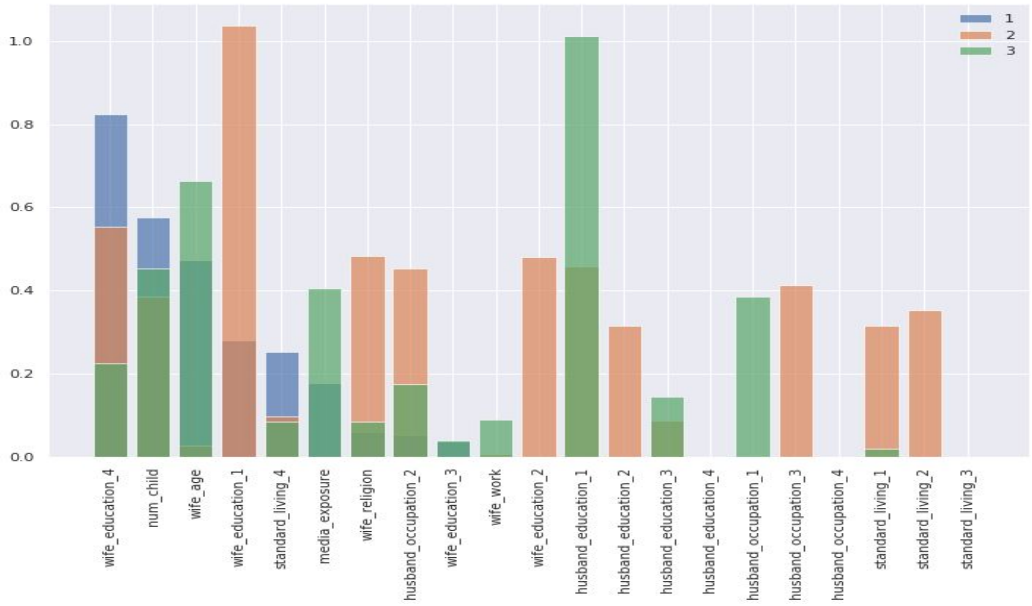
Table 2: Accuracy of Each Model

	Model	Training Accuracy	Test Accuracy
0	Logistic Regression	0.516978	0.528814
1	Decision Tree	0.962649	0.481356
2	Random Forest	0.962649	0.525424
3	Linear Regression	0.405522	0.404393

Judging by the overall accuracies for each model (**Table 2**), it is clear that all of them seem to have their faults. Both the decision tree and random forests seem to heavily overfit the training data, while only adequately performing on the test data. Both the logistic and linear regression models performed adequately in both areas, with the linear regression doing worse, indicating that the number of children may not be a suitable response variable, at least for this dataset. This may indicate that our current dataset and feature selection may be flawed, so we will investigate further using our logistic regression model, which we had assumed and concluded to be the most suitable working model due to its highest test accuracy. Using this model will also allow us to examine the relative importance of each feature.

First, we will examine the relative importance of each feature for each contraceptive choice based on the (absolute) numerical values of the coefficients used in their respective "one vs. rest" regressions.

Figure 2: Numerical Importance of each Feature by Contraceptive Choice



A cursory overview of the coefficients (**Figure 2**) seems to indicate that some features may not contribute to the overall prediction (coefficient = 0) for a particular contraceptive. These "zeroed" features, however, are not consistent between the three responses. For example, media exposure contributes for contraceptive 1 and 3 responses, but is zeroed out for contraceptive 2. This makes it difficult to create a consistent set of "contributing features," since each response seems to have unique affinities for each of the features.

To further support this, we conducted PCA on our variables. While we found that there only seemed to be 4 primary principal components (**Appendix I, Figure 8**), we also noted that there were no evident clusters among our first 3 principal components (**Appendix I, Figure 9a, b**). Upon analyzing the linear coefficients of our data, we see a similar trend as the Lasso coefficients (**Appendix I, Figure 10**); each principal component is defined uniquely by multiple features, and it's impossible to reduce our dimensions by a significant amount using these features. We can further examine the idiosyncrasies of each contraceptive response by examining the logistic regression model's accuracy for each contraceptive type. This is done by dividing the original data set into the three contraceptive responses and individually scoring them with the model to see how accurately the model predicts each of the labels.

Table 3: Contraceptive-Respective Accuracies

```
Contraceptive 1 accuracy: 0.6397058823529411
Contraceptive 2 accuracy: 0.3088235294117647
Contraceptive 3 accuracy: 0.5274725274725275
```

As we can see, the accuracy for each label varies greatly, with the model performing poorly on contraceptive 2 and adequately on contraceptive 1 and 3.

Analysis and Conclusion

As mentioned earlier, we ruled out using Decision Tree and Random Forest due to their optimistically high training accuracy but rather low test accuracy, suggesting that they overfit the data. Moving forward with our logistic (Lasso) regression model with 5-fold cross validation, we obtained accuracies labelled in (**Table 3)**; such differences in accuracy suggest that this model may be able to distinguish between no contraceptive users somewhat accurately, but fail in distinguishing between contraceptives 2 and 3 in particular. These results are further supported by our EDA, as we identified that in our correlational heat maps that contraceptives 2 & 3 were quite similar whereas contraceptive 1 noticeably differed. Therefore, we conclude that, using this dataset and measuring the features of an individual from this dataset's population, we can somewhat predict if an individual is likely to use a contraceptive or not (contraceptive 1).

In order to best improve this model, future research should look into addressing more features related to contraceptives, such as race, income, etc., although many may prove to be collinear to features already measured. In addition, contraceptives could be measured either on a binary (no usage/usage) or numerical scale in order to improve accuracy, as our results showed that measuring for contraceptive 1 was most accurate.

Addressing the 7 Questions:

(i) Wife education, husband education, and number of children seemed to be the most influential factors in our model. Wife education was perhaps the most interesting, since, not only did this feature's coefficient vary widely between the contraceptives, wife education 4, 1, and 2 also each had a very high coefficient value for at least one of the contraceptives. Husband education operated similarly, although to a lesser extent. Finally, each contraceptive type was strongly influenced in our model by the numerical feature, number of children.

(ii) One feature we were quite surprised to see had a small coefficient among all 3 contraceptive types was the employment status of the wife (wife_work). Our group expected an employed wife to undergo either a long-term or short-term contraceptive, but the employment status seemed to have little effect in determining what contraceptive type a woman used.

(iii) We faced serious trouble in finding features to determine our model. This is because, while our features seem to be tied to our response variable of contraceptives, many of the features did not differ too much from contraceptive to contraceptive - this is especially evident in our EDA in Figures 2-6, and its results show in our model accuracy as well.

(iv) From our modelling analysis, we saw that it was very difficult to create any predictive model for the contraceptive method choice, meaning that our reliance on the logistic regression model (in terms of analysing accuracy and feature selection) may not be the most appropriate or accurate. It was, however, our best model, so we were limited to using it. We assumed that, for the most part, a majority of features contributed at least slightly to at least one response type. This assumes that the dataset we utilized is representative of our population of interest, however, so an inclusion of more data or better dataset could entirely negate our conclusions.

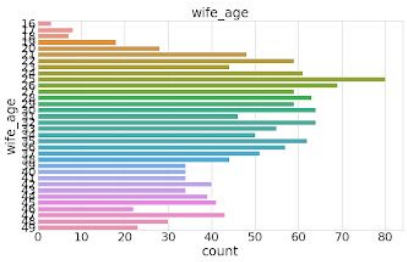
(v) There didn't appear to be any ethical dilemmas that resulted from obtaining data from the survey as each family was notified beforehand of what the information was used for and asked for their consent to be interviewed. However, when it came to reading and analyzing the data, especially the questionnaire, it asks for personal details which thankfully were not included with the dataset.

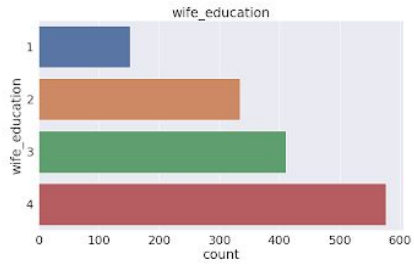
(vi) For our analyses, the presence of more data would help us better determine any relationships between the various features, since our models and analysis seemed to be pretty confounded and messy with the data we were given. Additionally, more concrete, numerical data would most likely help us find clearer patterns. Our given dataset had an abundance of very broad, categorical features, which are still useful but does not provide much to help us select and separate key features and patterns.

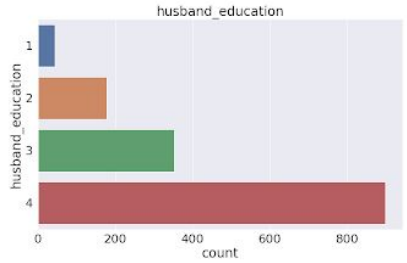
(vii) A potential ethical concern would be if during the data collection process, people were not notified beforehand of what they will be interrogated for and what the information will be used for, as well as not obtaining their legal consent to be interrogated. Luckily, the National Family Planning Coordinating Board appointed people that attended a training session involving learning the language and customs of local ethnic groups and provided all necessary, legal information to the village staff with public acceptance to contribute to the study.

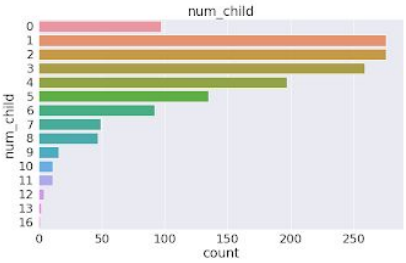
Appendix I

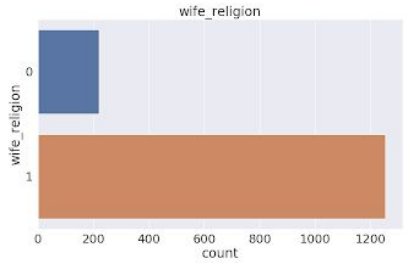
Figure 1: Population Count of Each Feature by Feature Value

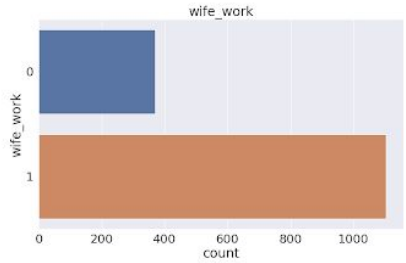


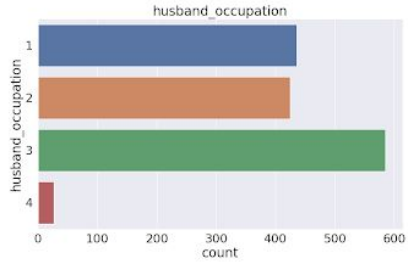


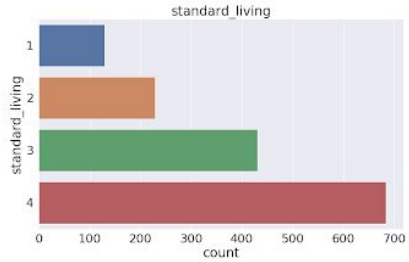


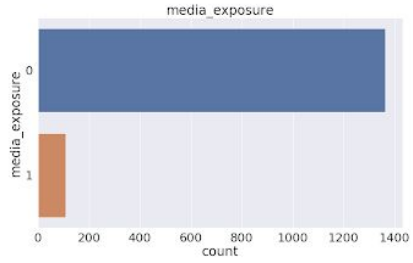












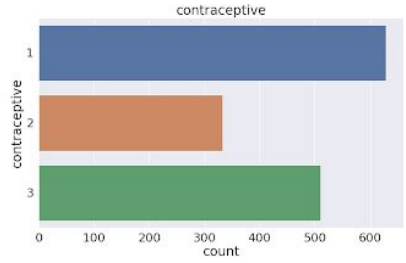
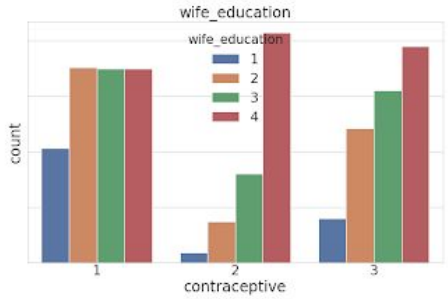
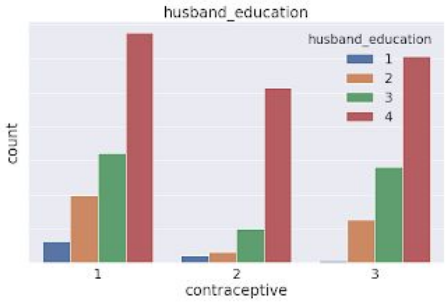
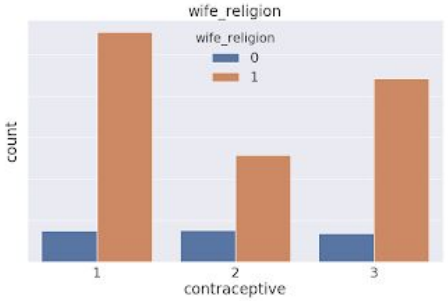
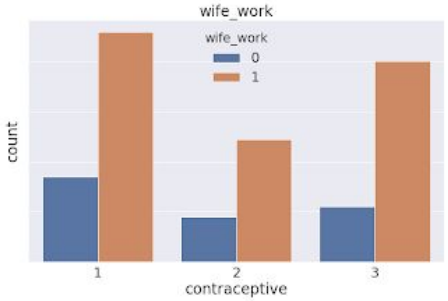


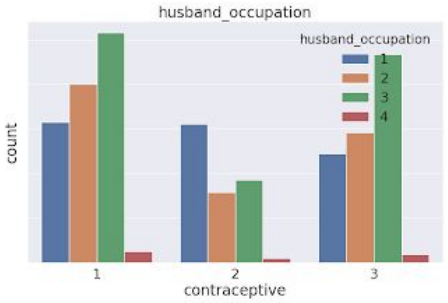
Figure 2: Population Count of Each Feature by Value For Each Contraceptive Type

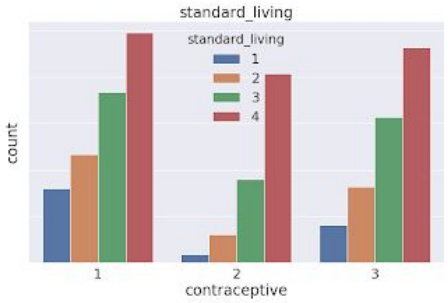












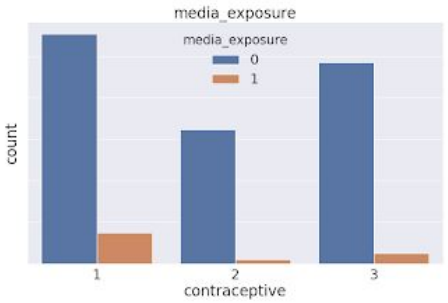


Figure 3a: Distribution of Women by Age for Each Contraceptive

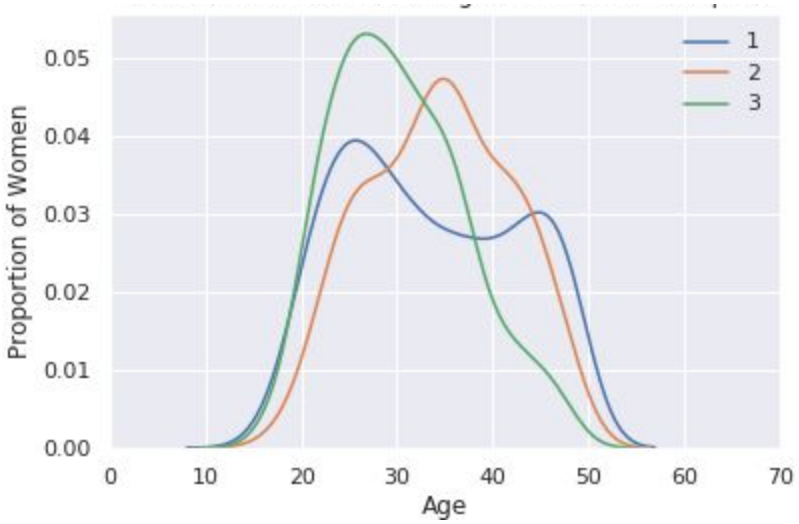


Figure 3b: Distribution of Women by Number of Children For Each Contraceptive

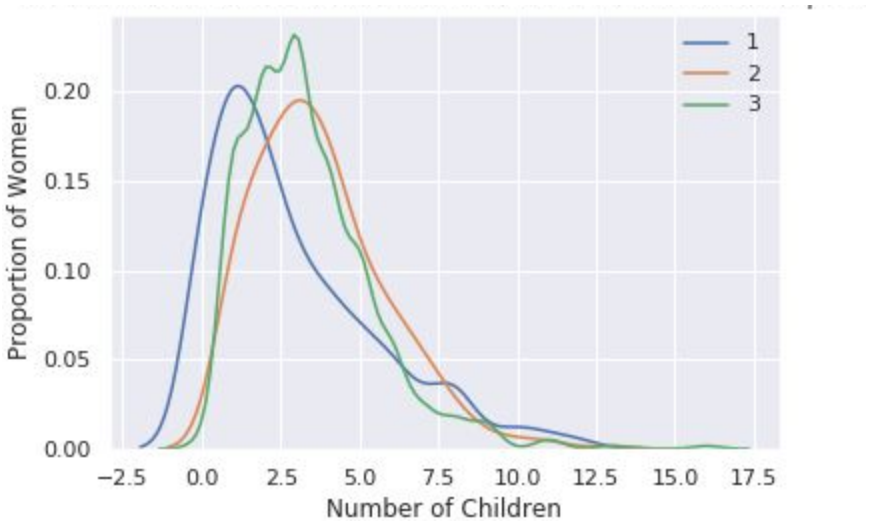


Figure 4: Number of Children vs. Wife Age For Each Contraceptive Type

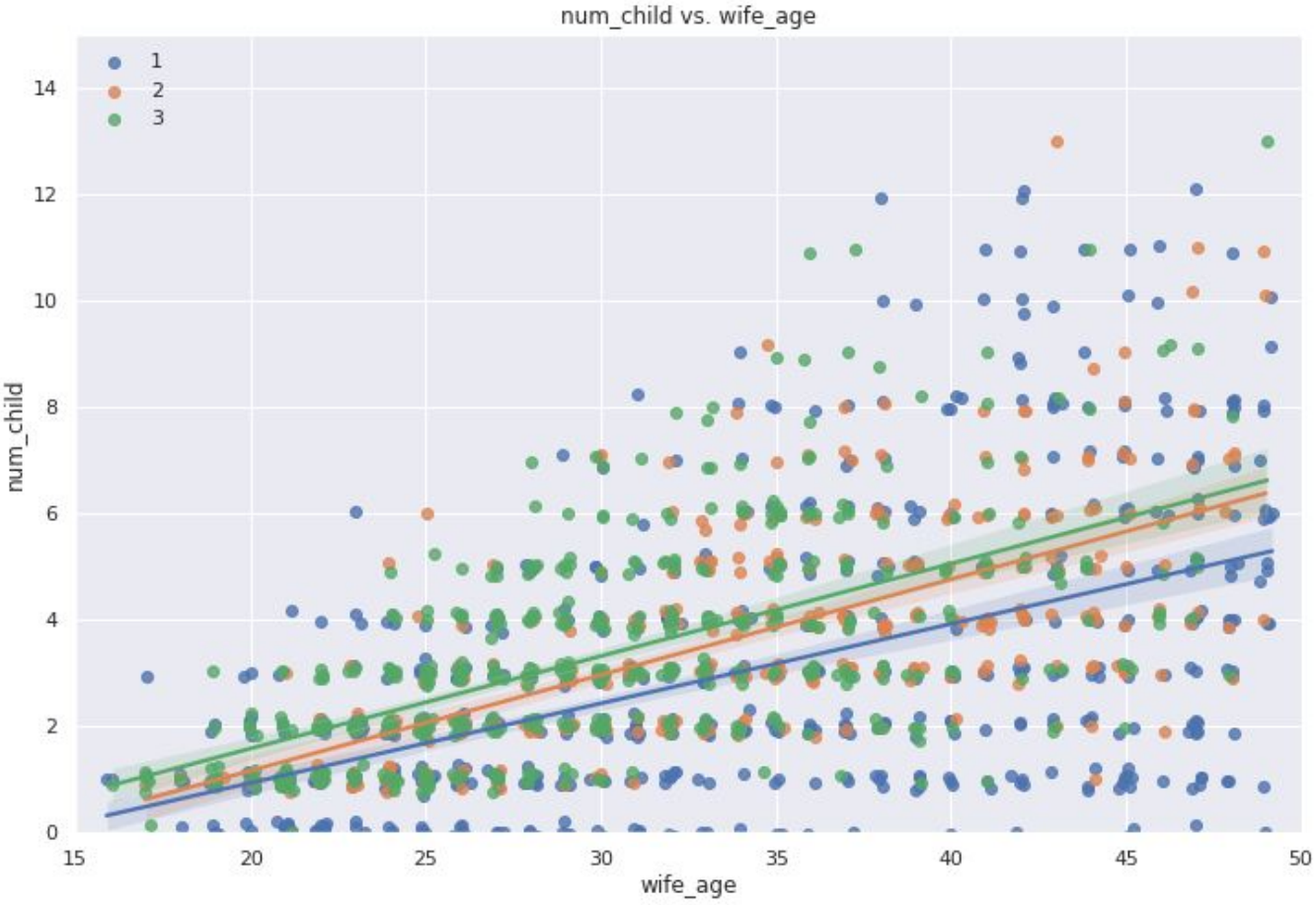


Figure 5a: Pairplots of All Categorical Features

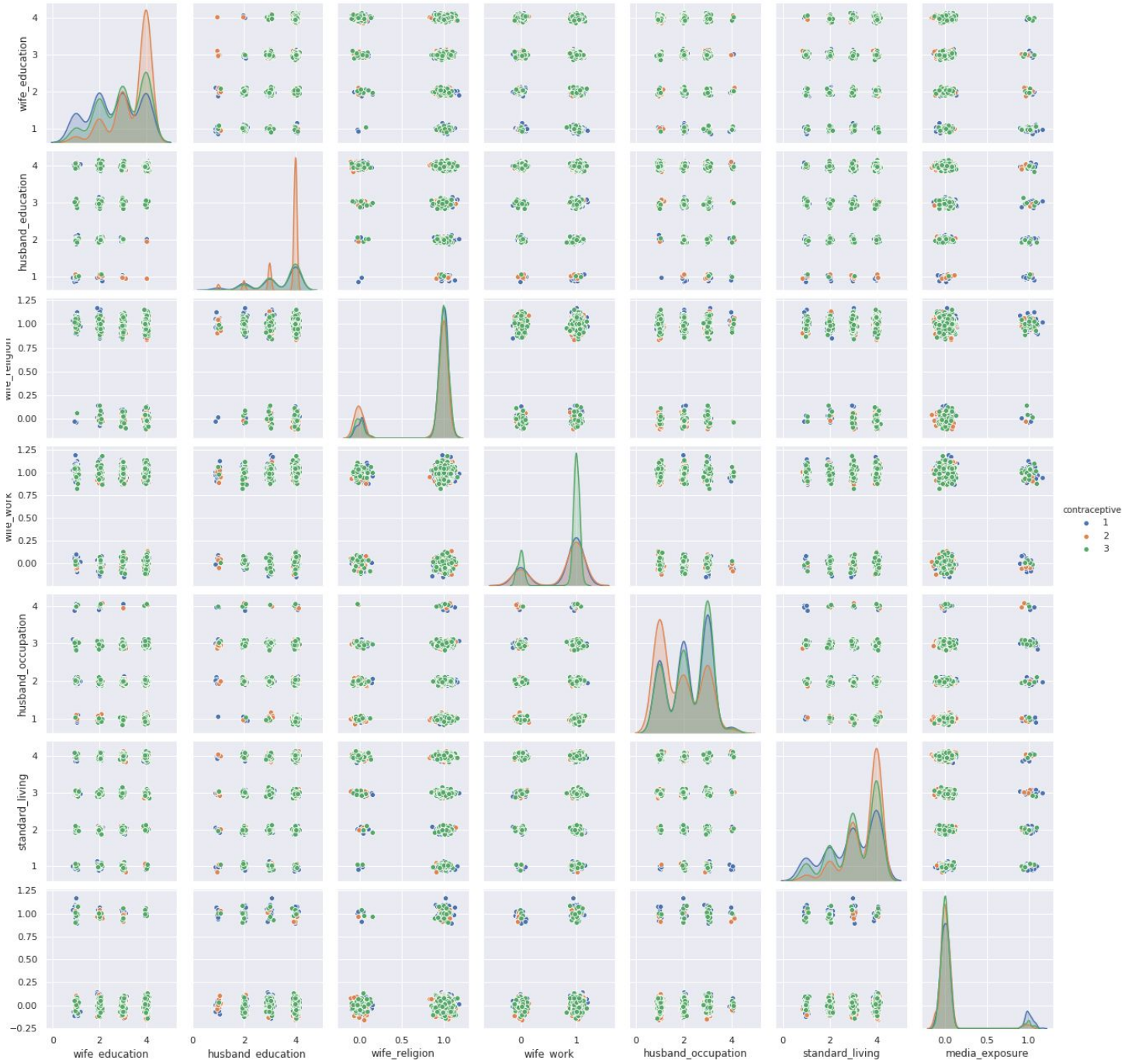


Figure 5b: Pairplots of Wife Work vs Wife Religion For Each Contraceptive Type

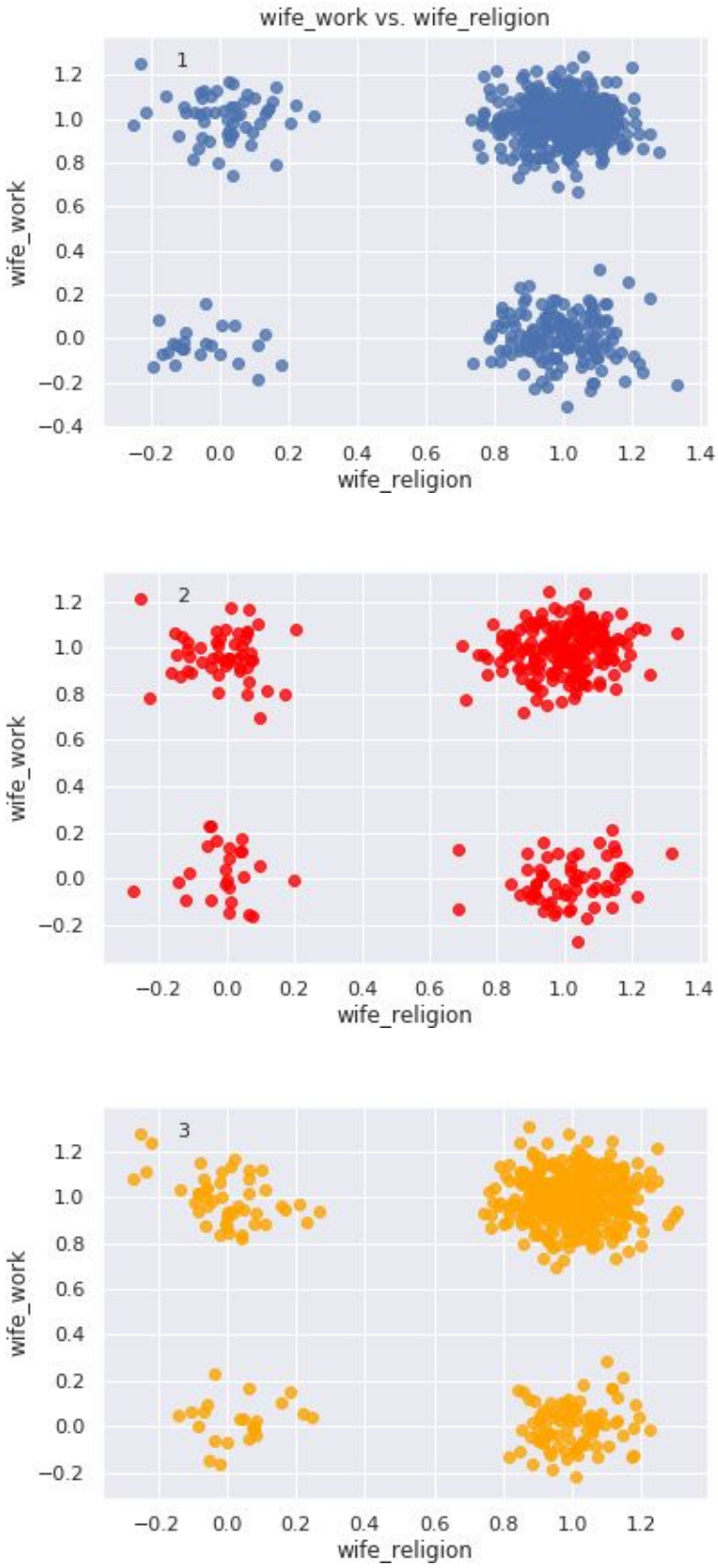


Figure 6a: Heat Map of Correlations Between All Variables (Enlarged Figure 1)

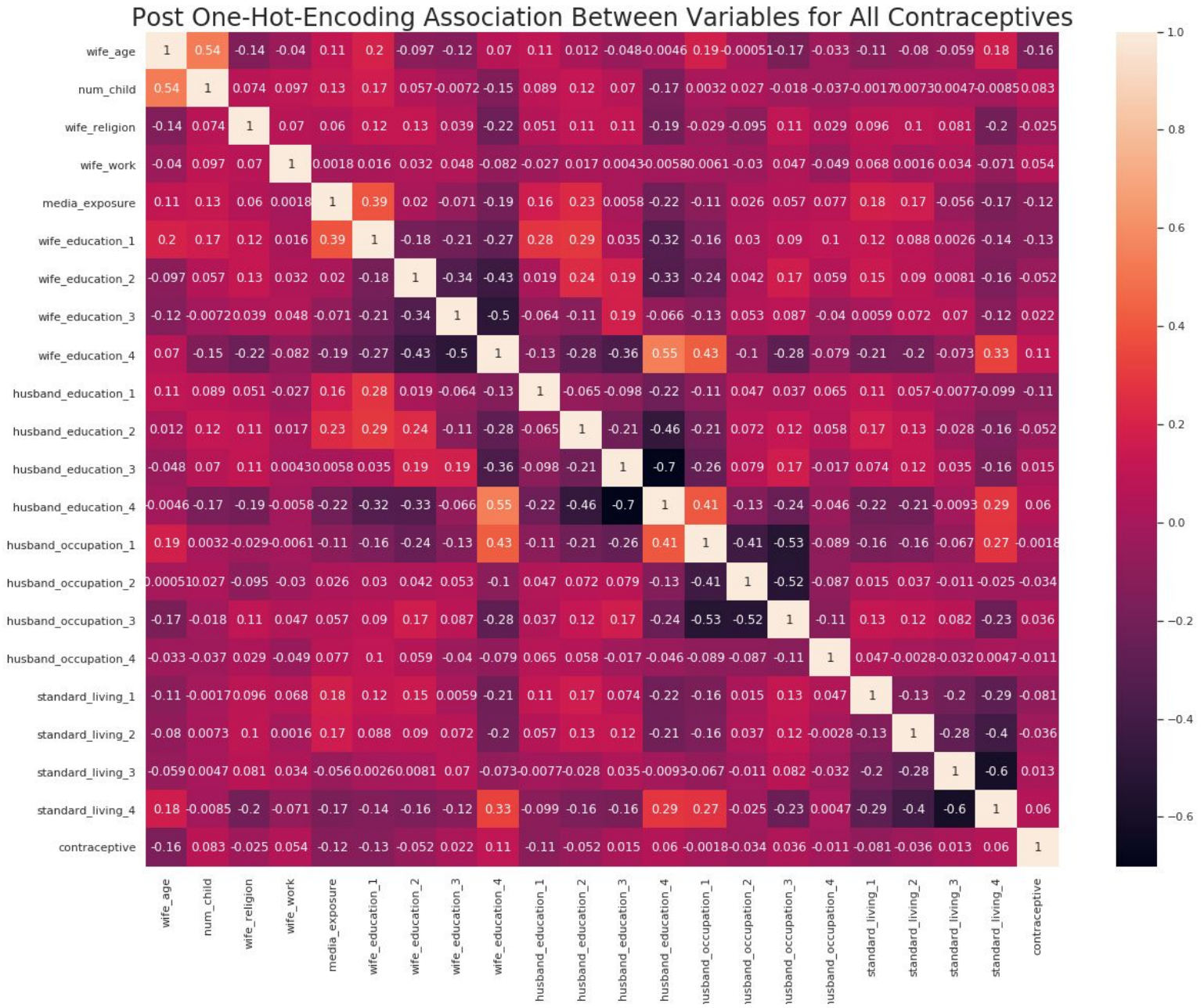


Figure 6b: Heat Map of Correlations Between All Variables For Contraceptive 1

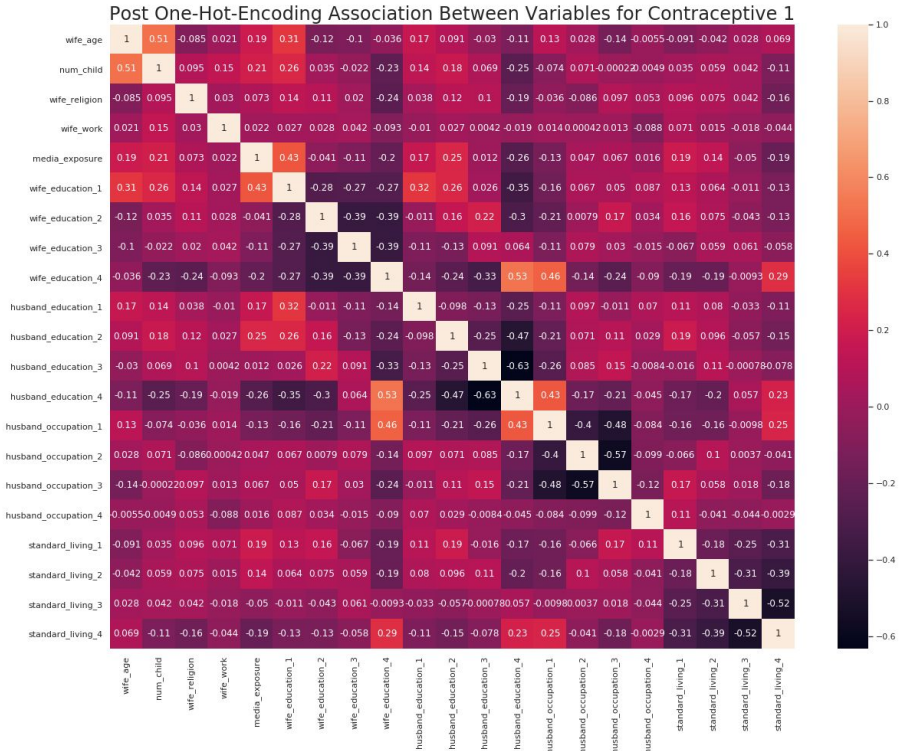


Figure 6c: Heat Map of Correlations Between All Variables For Contraceptive 2

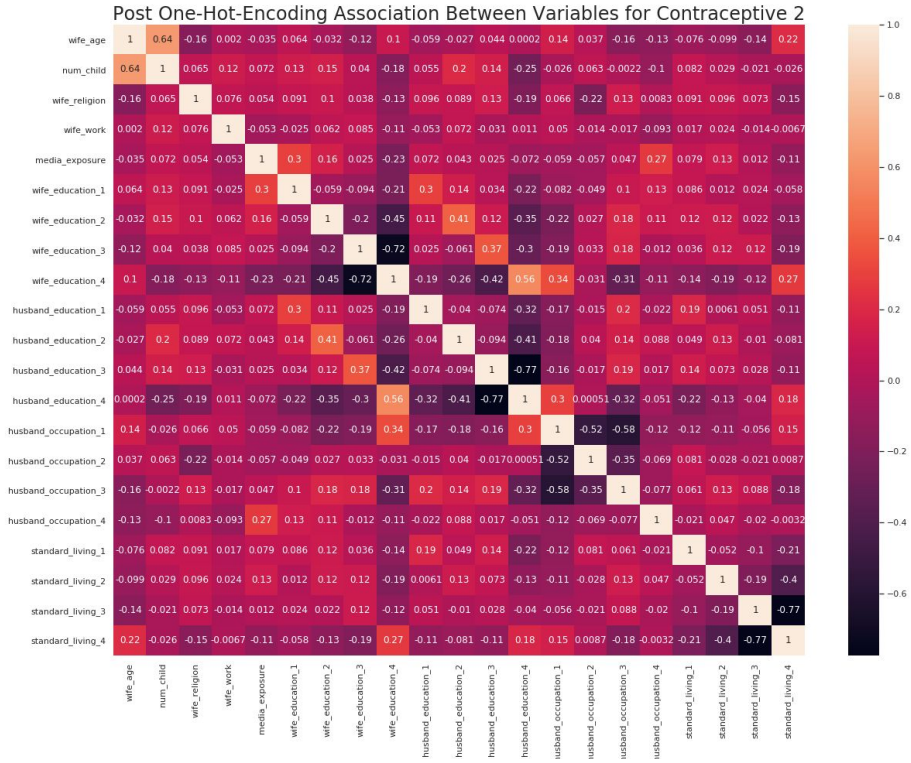


Figure 6d: Heat Map of Correlations Between All Variables For Contraceptive 3

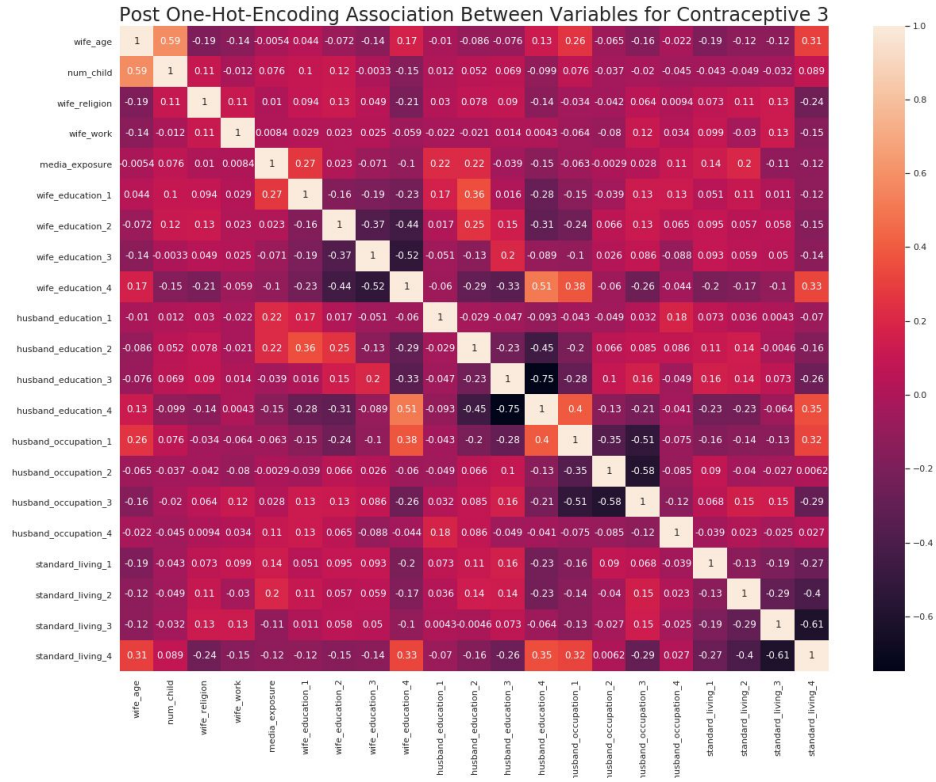


Figure 7: Heat Map of Correlations Between All Variables Using Cramer's V

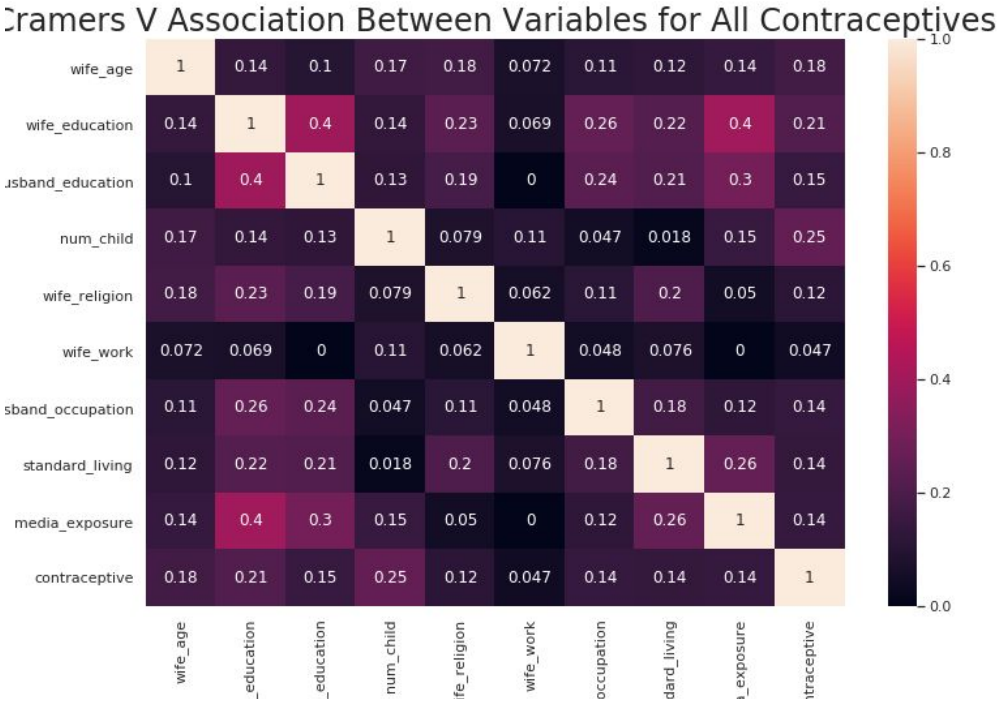


Figure 8: Scree Plot of PCA

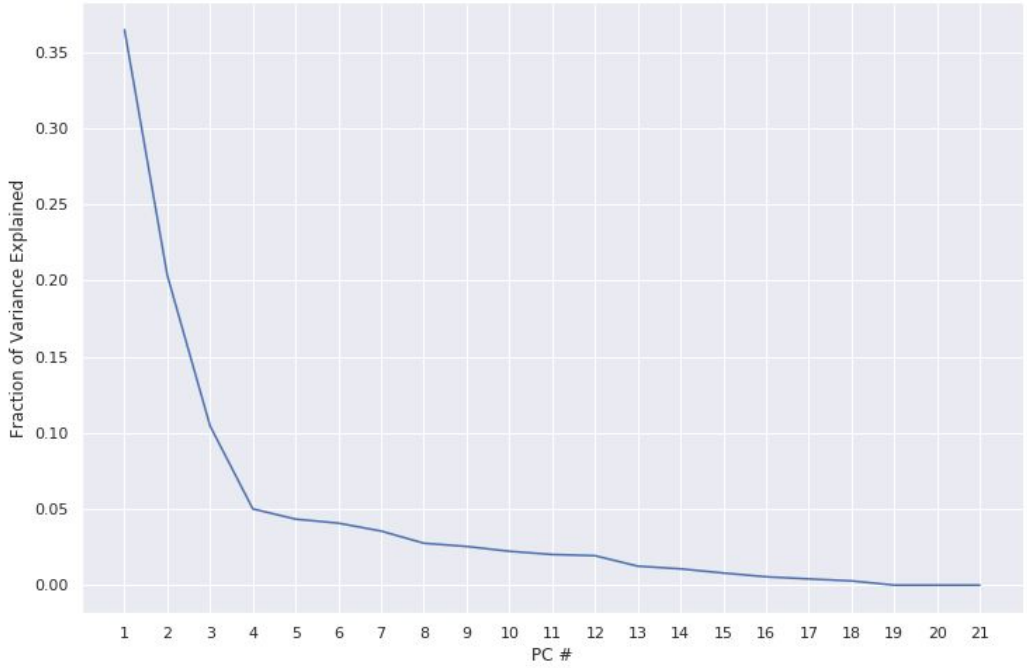


Figure 9a: Plot of Principal Components 1 and 2

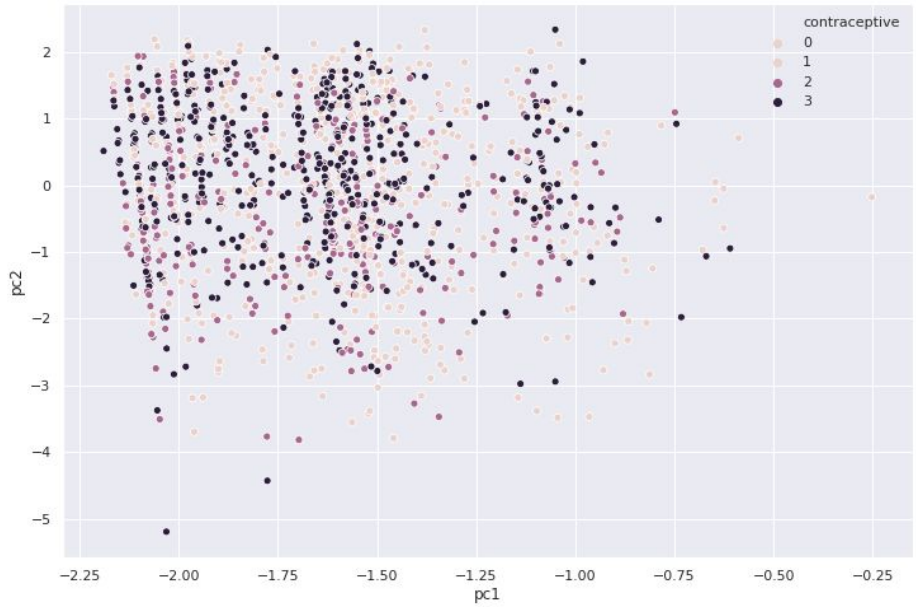


Figure 9b: Plot of Principal Components 1 and 3

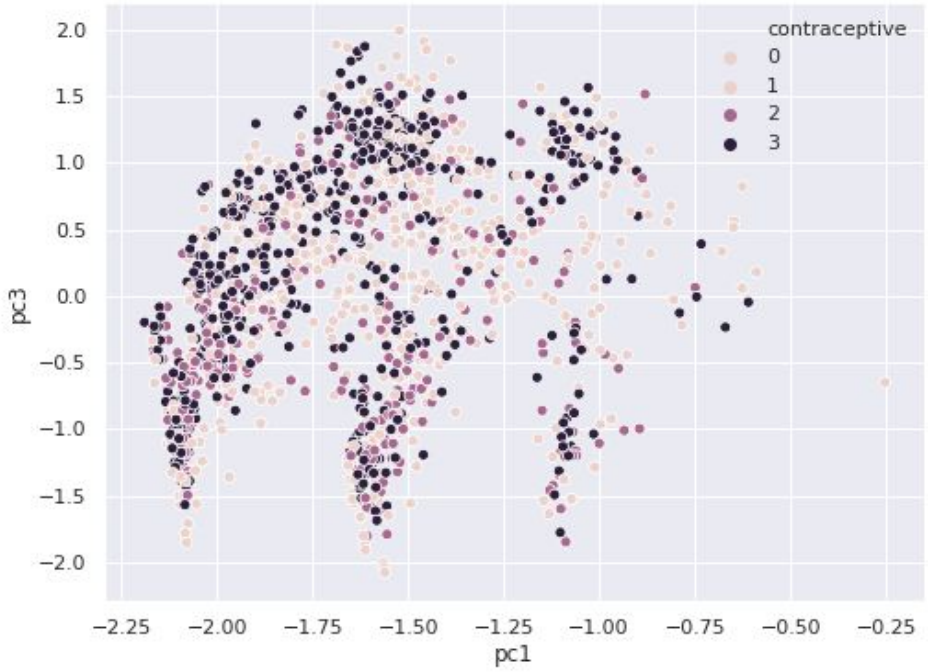


Figure 10: Linear Coefficients of the First 3 Principal Components