DATA 100 UwU

Sara Imam, Shirley Wang, Sierra Dean

INVESTIGATING CONTRACEPTIVES

TABLE OF CONTENTS

Abstract

The use (or lack thereof) of contraceptive methods can be strongly affected by demographic and socio-economic factors. Being able to predict the use from the factors, allows for better marketing or advertising to targeted groups. With this goal, we attempted to be able to predict the use of contraceptives (no use, short term, long term) from looking at certain demographic and socio-economic features. Upon the creation of many models (including LogisticRegression, Cross-Validation, Random Forests, etc) plus analysis to find the most highly correlated features and the addition of hyperparameters, we were able to achieve a training accuracy of 72%. From there, we simplified our model to become a binary classification, predicting solely the use of contraception (1 for uses contraception or 0 for no use). Utilizing the same procedure as before we were able to achieve an 82% training accuracy and a 67% test accuracy.

Introduction

Contraception is the use of deliberate methods to prevent or reduce the chance of pregnancy. Methods can be anywhere from short term, one-time use through long term, more permanent use, as well as opting to use no form of contraceptive methods. Short term methods include condoms and spermicides and long term methods include IUDS, and “the Pill”, among other methods. Overall, the use of different birth control methods can be affected by factors including religion, education, media perception, among others. The ability to predict such use could aid in marketing or educational efforts. By analyzing the use of contraception as described in a 1987 data set from the National Indonesia Contraception Prevalence Survey could help to answer some of these questions.

Description of Data & Data Cleaning


Columns:

1. Wife's age (numerical)
2. Wife's education (categorical) 1=low, 2, 3, 4=high
3. Husband's education (categorical) 1=low, 2, 3, 4=high
4. Number of children ever born (numerical)
5. Wife's religion (binary) 0=Non-Islam, 1=Islam
6. Wife's now working? (binary) 0=Yes, 1=No
7. Husband's occupation (categorical) 1, 2, 3, 4
8. Standard-of-living index (categorical) 1=low, 2, 3, 4=high
9. Media exposure (binary) 0=Good, 1=Not good
10. Contraceptive method used (class attribute) 1=No-use, 2=Long-term, 3=Short-term

Figure 1. *Data presented in the Survey for use in our analysis.*

Most of the data from this data set was categorical, so we utilized One-Hot Encoding for the features Wife Education, Husband Education, Husband Occupation, and Standard of Living. By using this method, we could effectively use these features in our analysis. Additionally, we checked and found no null values in our data set to account for. We chose not to normalize our numerical values because both had similar ranges.

Visualizations

To gain a better understanding of the relationship between the features, we created several visualizations to reveal trends and correlations between contraceptives, the number of children, and other interesting features.

One of the insights found is between the wife's education and the use of contraceptives (Figure B.2). As illustrated in the bar graph, as the wife's education level increases, the proportion of couples who do not use any contraceptive decreases almost linearly. Moreover, the long term use of contraceptives shows a clear increase as the wife's education increases. In the short term, the proportion increases again, but then decreases as the wife's education reaches the highest level. There's a similar trend when looking at the graph depicting husband_education and contraceptive use, where as the husband's education increases, the proportion of couples who do not use contraceptives clearly decrease (Figure B.3). However, in the long term, the trend seems to dip at level 2, then increase from there. Another strong relation can be found between the standard of living and the use of contraceptives, where the trends are the same as the trends described previously between wife education and contraceptive use.

Interestingly, when looking at the relationship between contraceptives and the wife's religion, there is a higher proportion of couples who do not use any contraceptives when the wife's religion is Islam compared with non-Islam (Figure B.7.a). The proportion of long-term contraceptive use is also lower when the wife's religion is Islam and is slightly higher for short-term contraceptive use compared with non-Islam.

Although the next visualization may not be as clear or useful for the model, when inspecting contraceptive use vs media exposure, there seems to be almost a 30% difference in the proportion of couples who have good media exposure and do not use contraceptives (40%) and those who do not have good media exposure and do not use contraceptives (70%) (Figure B.9). Additionally, both long-term and short-term contraceptive use is lower for not good media exposure versus good media exposure.

Using the information deducted from the visualizations, there seems to be a clear trend in some features compared to others.

Description of Methods + Model

Since we wanted to try to accurately predict contraception use levels 1-3, we decided to try 3 different types of models that all supported multiclass classifying: Logistic Regression with Cross Validation, Decision Tree, and Random Forest.

We started by picking the most relevant features based on our observations made through visualizations. We picked the features that showed the strongest correlation with contraceptive use: whether the wife worked or not, the standard of living, media exposure, the level of wife education, and the level of husband education. The strong correlations are outlined in the visualizations portion of this report (Figure E.1).

We then used our selected features to create the first 3 models. Our Logistic Regression with Cross validation performed the poorest with a 50% training accuracy. The models that better fit our data was the Decision Tree and Random Forest model. They both produced a training accuracy of 71.6% and 71.5% respectively. However, we found that the Decision Tree model was slightly overfitting our training data, as It's testing accuracy (~43.7%) was lower compared to our Random Forest Model's testing accuracy (~45.8%). More so, in general, while both models overfit, the Decision Tree always tends to overfit more. We thus concluded that the Random Forest was the best model for our data.

However, the accuracy for our model was low overall. So, to ensure that the features we chose were the best possible features, we fit multiple more random forest models with different features to try to improve our accuracy. However, all the models we tried only lowered our accuracy. The only one model that improved was using every feature available, however this was grossly overfitting our training data. As such, we determined that the features we used for our model were the best features to use.

In a last attempt to improve our accuracy, we attempted to tune the hyper parameters of the Random Forest model. After trying many hyper parameters, we found that limiting max_features to be 10 improved our model. The implemented changes resulted in a 71.6% training accuracy, and a testing accuracy of 47.1%. Thus, we concluded that the best model we could create was a Random Forest model, with the features outlined above, and a max_feature set to 10. It had both the best training and testing accuracy of all models created.

Analyzing Model & Summarizing Results

Our final model was a Random Forest, and it had a 71.6% training accuracy and a 47.1% testing accuracy. We realized that the Random Forest was best as the Decision Tree model overfit, and for Linear Regression, there was likely no linear decision boundary, making it a poor classifier. However, while this is better than a completely random model (with expected accuracy of 33%), the accuracy is still very low.

One of the reasons our accuracy was low was because of the limitation of the data itself. We plotted a heatmap of the data, which revealed that there was nothing relevant correlated to each other (Figure E.1). The only things in our dataset that were highly correlated were wife education and husband education, and wife age and number of children. These were the only 2 correlations above 0.5. As such, our classifiers did not have much information to work off of.

We next analyzed class imbalance to further understand our model. It turns out that class 1 had 629 people, Class 2 had 333 people, and class 3 had 511 people. This shows that both class 1 and class 3 were more populated than class 2. Our classifier reflects this, as our confusion matrix shows that we predicted more 1's than anything else, and more accurately predicted 1's. More so, it shows that we had less overall class 2 prediciations and correctly predicted class 2 less accurately than either class 1 or class 3, as class 2 was accurately predicted only 60% of the time, but class 1 was predicted correctly 80% of the time, and class 3 was predicted correctly 70% of the time, as seen by the calculated precision. This adds to our lack of accuracy.

Overall, our average precision came to be about 70%, which means our predictions for the positive class were actually correctly 70% of the time. Our average recall was also about 70% , which means 70% of total relevant results were correctly classified by our model. This shows that our precision and recall was not balanced. Our model algorithm has classified an equal amount of users as false positives, as it classified false negatives. This further explains lack of accuracy in the model.

So, in order to try to improve our model, we tried asking a new question: Could we more accurately classify binary Data? We changed our Y values from 1-3 to 0-1, where 0 is no contraceptive use (corresponding to the previous 1), and 1 is contraceptive is used

(corresponding to the previous 2 and 3). We then tried to refit a Logistic Regression with cross validation model, Decision Tree model, and Random Forest model with the new Y, but with the same features and hyperparameters as previously used. The model which was the most accurate was the Random Forest model, with an 82.3% training accuracy and a 66.8% testing accuracy. So, a binary classifier did indeed more accurately classify the data, with a 10.8% higher training accuracy and a 19.7% higher testing accuracy, a very large improvement.

Discussion & Conclusion

In conclusion, our best model for the non-binary model was a Random Forest, with a 71.6% training accuracy and a 47.1% testing accuracy. However, we decided that this model was not accurate enough to successfully classify the levels of contraceptives. Comparatively, our binary model was also a Random Forest model with an 82.3% training accuracy and a 66.8% testing accuracy. We conclude that this is a much better model for accurately classifying whether contraceptive is used or not.

In the creation and analysis of our model, we were able to pinpoint certain features of interest. Specifically, the features of husband and wife education. The education features had a very strong trend as well as very highly correlated with one another (Figure E.1), a trend not reflected by many of the other features. Upon the creation of the heat map, we saw there was a high correlation between these two and one of the features we initially thought was going to be useful but turned out not to be was Wife Religion. We initially made the assumption that muslim woman were more likely to be against contraceptive use, as the Qu'ran states that "You should not kill your children for fear of want" (17:31, 6:151), which is implied to mean no contracpetive use in more traditionally religious countries like Indonesia, where the data was collected.

Additionally, as referenced in the analysis, we struggled with the lack of correlation in our data set. We were stuck at this point for a while, searching for ways to improve the over accuracy, ultimately trying different features and adding hyperparameters, as well as investigating a slightly different question to improve the model. This was also one of the limitations we faced. Attempting to overcome the lack of correlation was one of the bigger challenges of the data set.

One of the ethical questions raised with this data set was the consent of use of information. When the survey was given, was the scope of expansion made known to the surveyees? Were the women who were taking this survey in 1987 made aware their personal information could be analyzed by some Berkeley students nearly 40 years later? This is the main ethical question for use of the data set.

In future investigations, it would be helpful to have access to the rest of the survey results, as we were only presented with a portion of the results. This could help us find more correlated features to better our models. Otherwise, it would be interesting to re-do this survey in modern times (2020 as compared to 1987) which would allow for the testing of different hypotheses including trends in contraceptive use over time.

Lastly, in studying the use and prediction of contraceptives there arises the ethical question of invasion of privacy. For many people, especially in these countries of interest, the use of contraceptive devices is often kept within the household and not publicly discussed. The ability to use this data to predict whether or not certain households use contraception could be considered an invasion of privacy for families.

Appendix

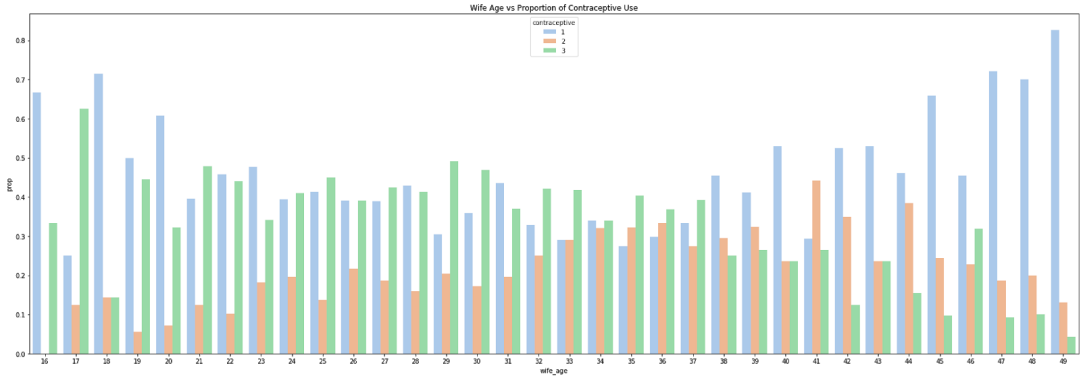

Figure B.1. *The proportion of contraceptive use for each age in Wife Age normalized over each age.*

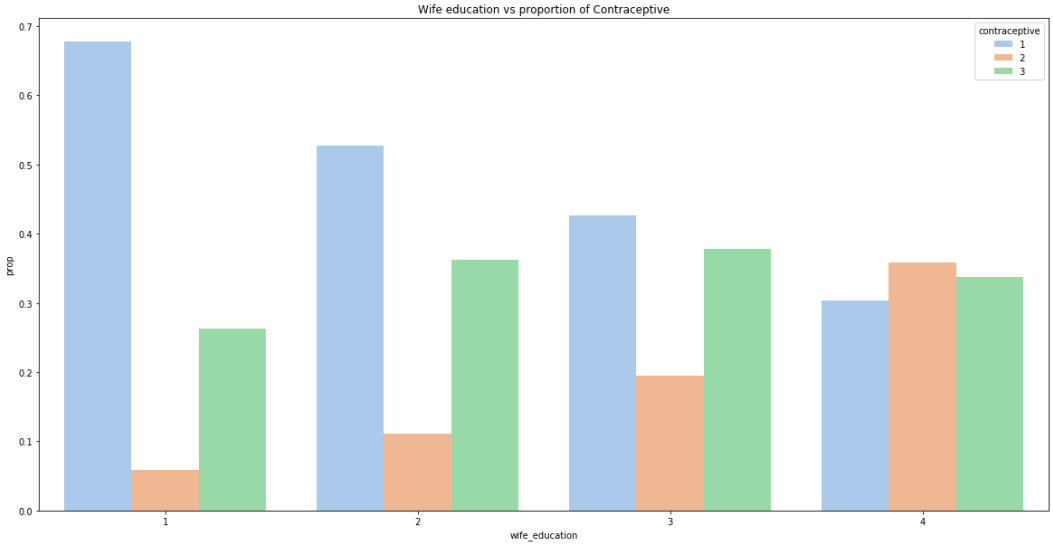

Figure B.2. *The proportion of contraceptive use for each Wife Education level, normalized for each level.*

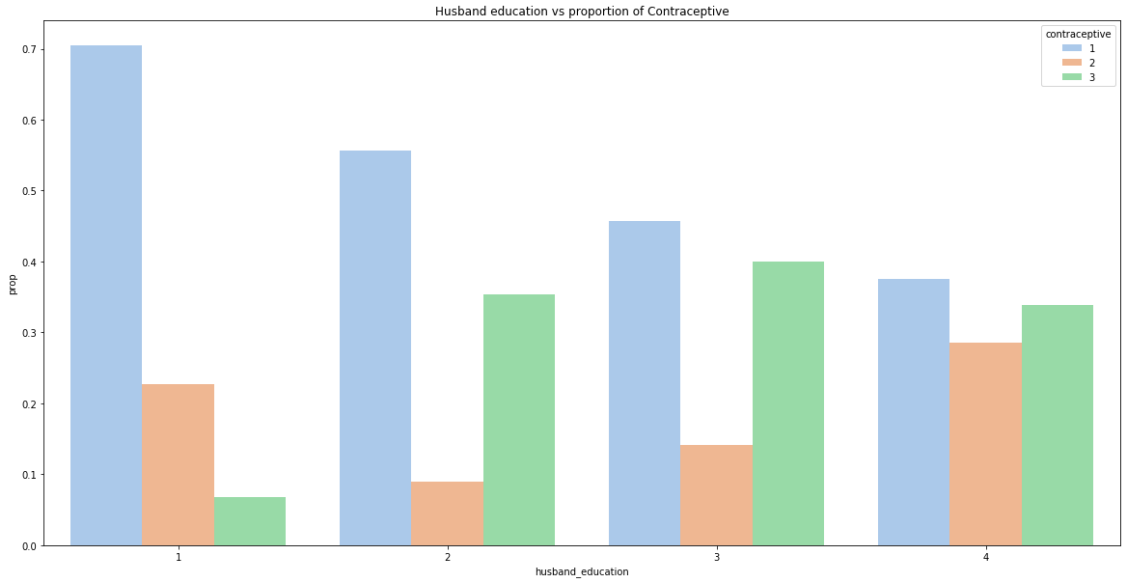

Figure B.3. *The proportion of contraceptive use for each Husband Education level, normalized over each level.*

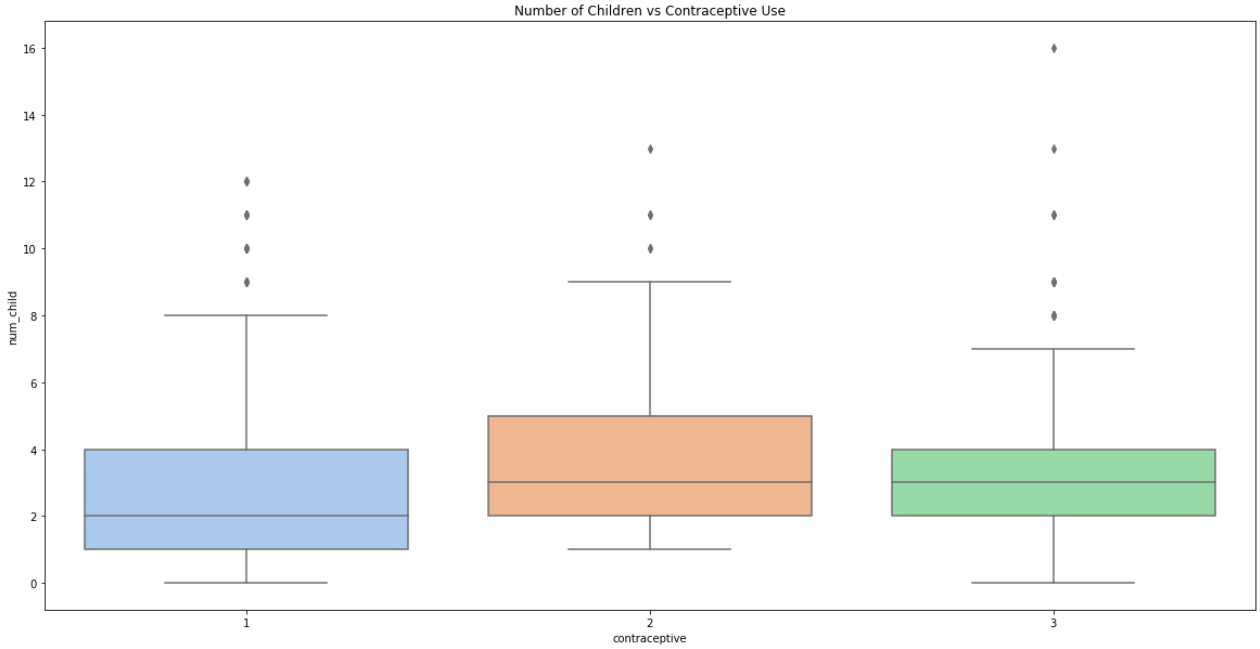

Figure B.4. *Distribution of number of children for each given contraceptive type.*

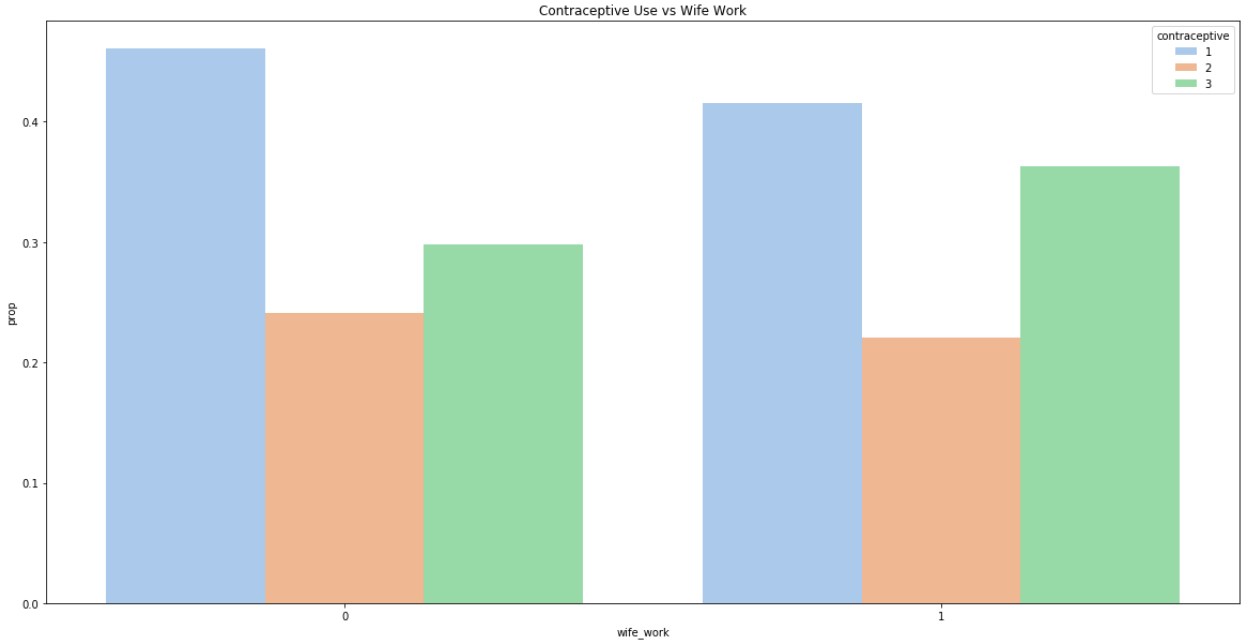

Figure B.5. *The proportion of contraceptive use for each Wife Work, normalized over each category of Wife Work.*

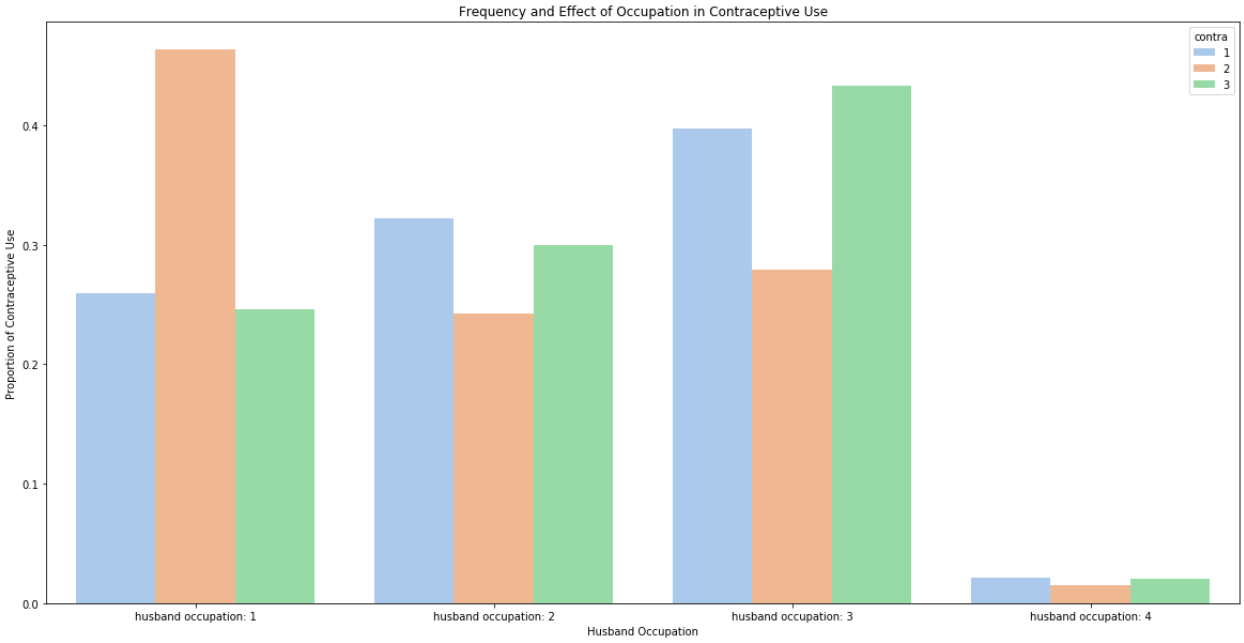

Figure B.6. *The proportion of contraceptive use for each Husband Occupation type, normalized over total.*

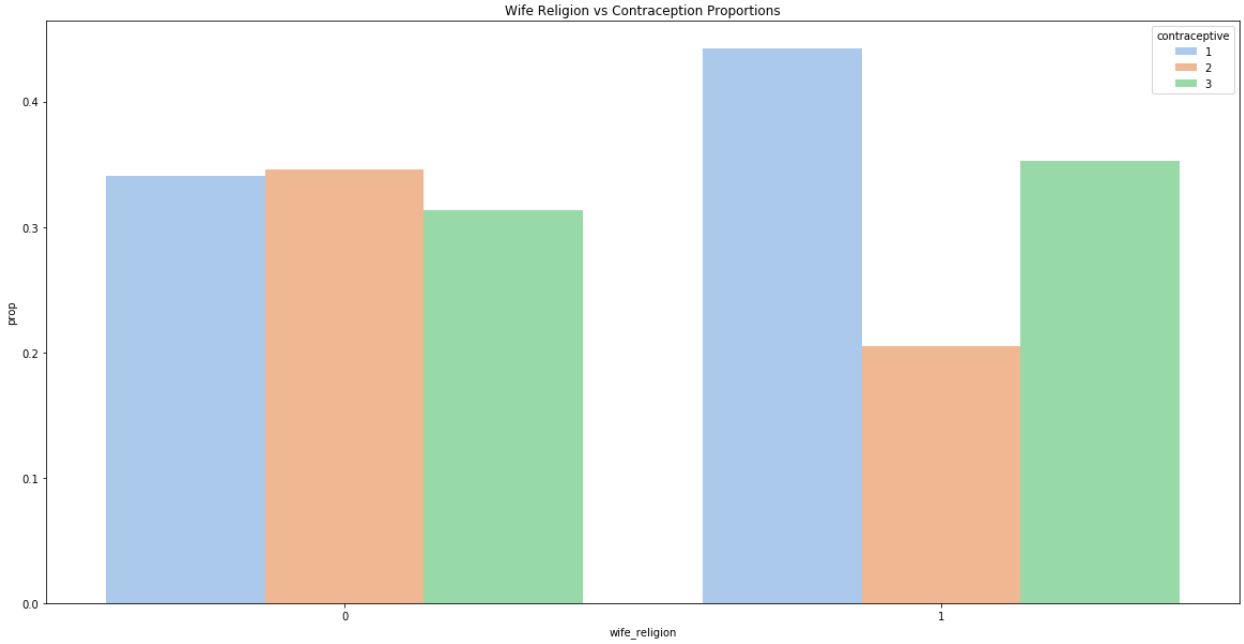

Figure B.7.a. *The proportion of contraceptive use for each Wife Religion, normalized for each Wife Religion category.*

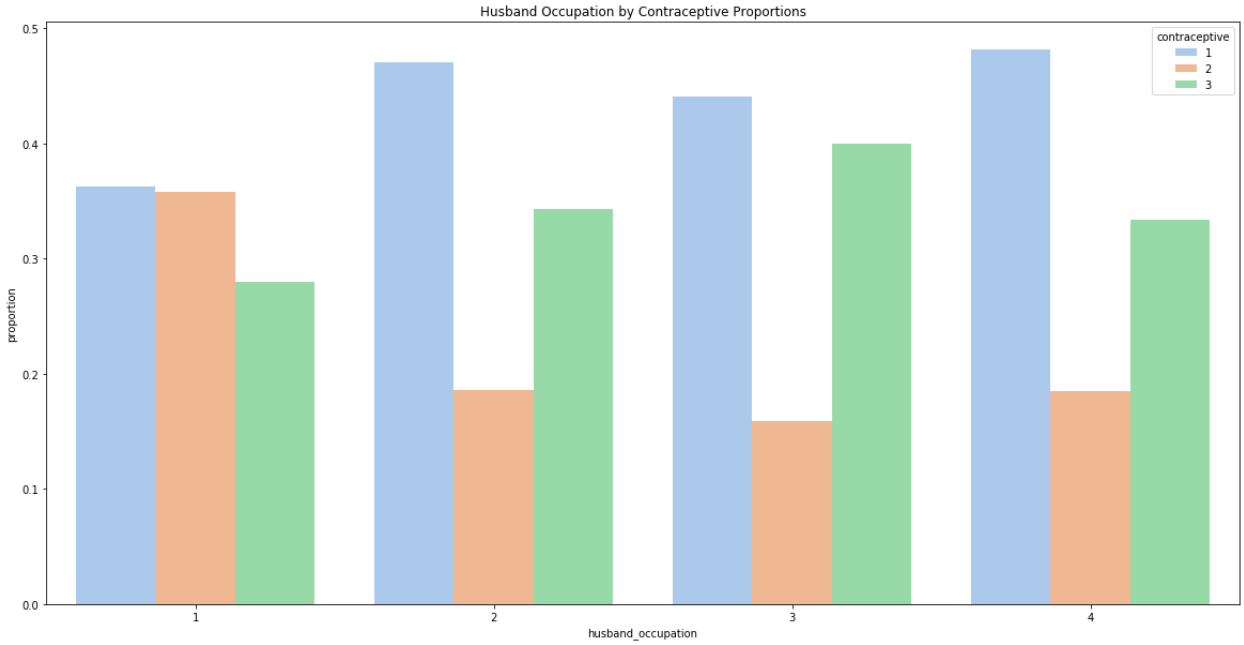

Figure B.7.b. *The proportion of contraceptive use for each Husband Occupation category, normalized over each occupation category.*

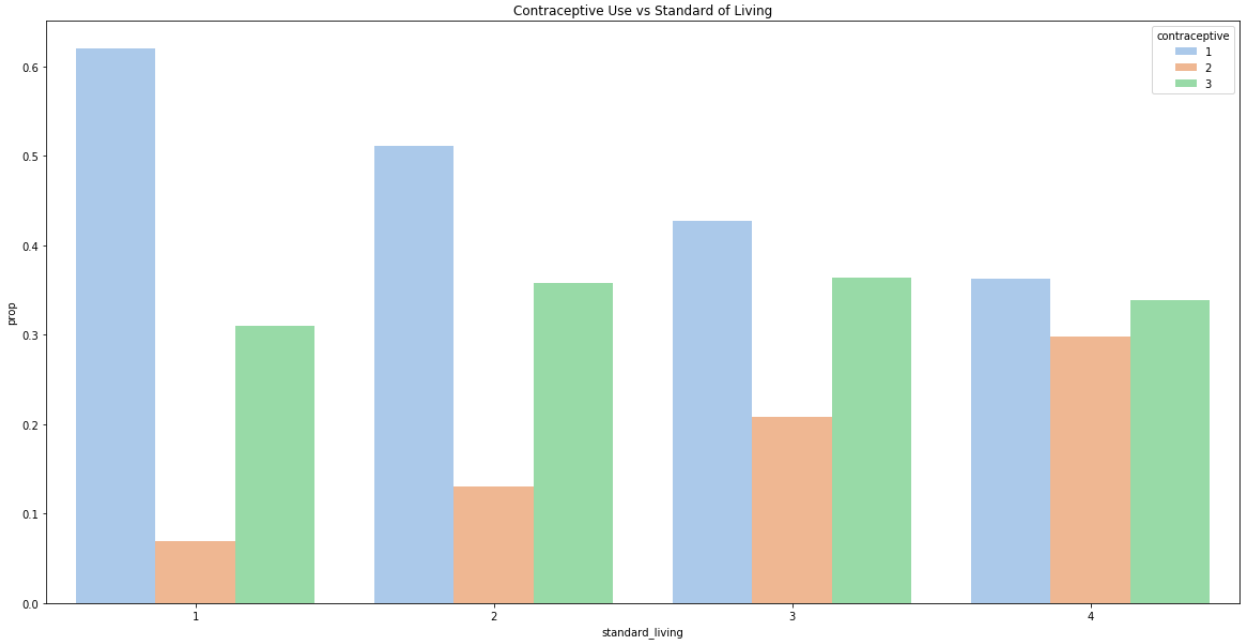

Figure B.8. *The proportion of contraceptive use types for each level of Standard Living, normalized over each level of living.*

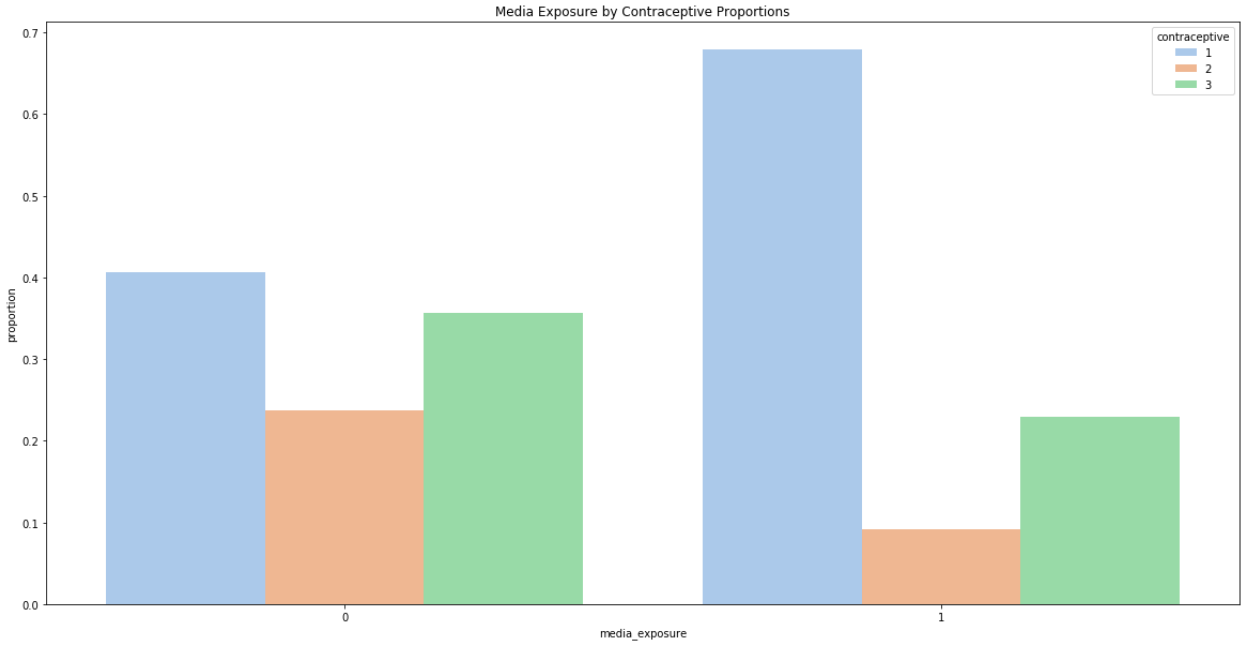

Figure B.9. *The proportion of contraceptive use types for each type of media exposure, normalized over each type of media exposure.*

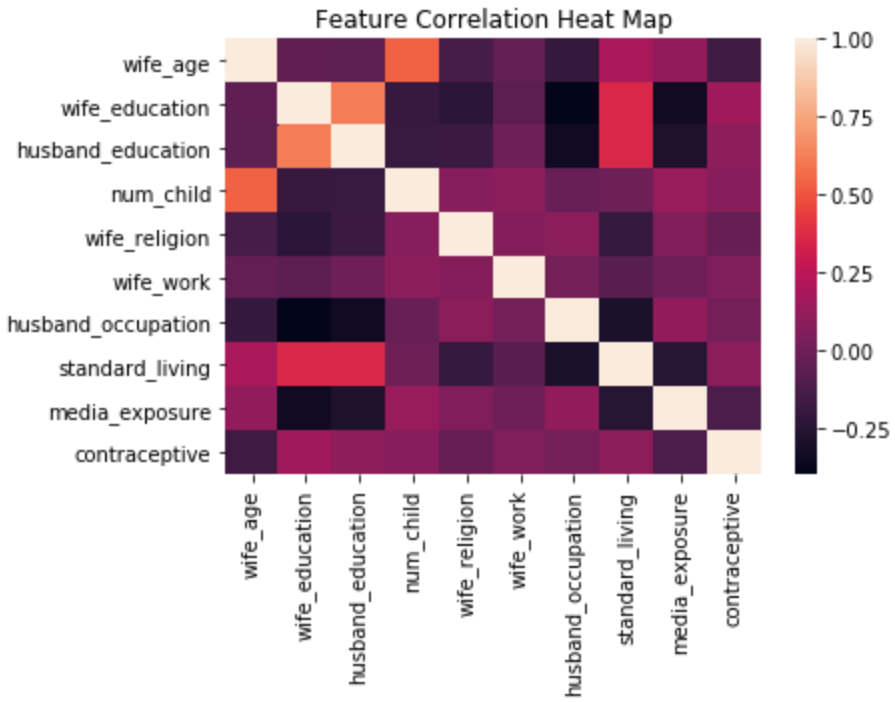

Figure E.1. *Heat Map showcasing the correlation between the different features.*