Nabeel Hingun, Urjil Sanghvi

May 13th 2020

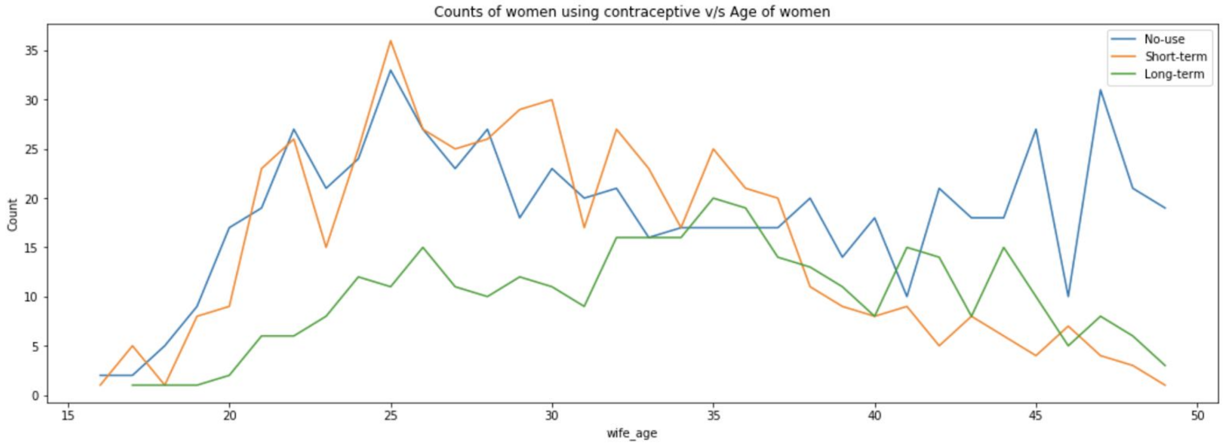
Data 100 Final Project Report

In the following paper, we investigate the Contraceptives Dataset. The purpose of this research is to develop a classification model to accurately predict the form of contraception used by a married woman based on other individual features such as age, education, standard of living, etc. In this report, we explain our analysis of the data, the models developed, and conclude with some ethical issues encountered. The interest in the type of contraceptives used by Indonesian women in 1987 seems to have stemmed from a decline in the fertility rate of the country which started in the 1970s and the government tried to understand what was happening.

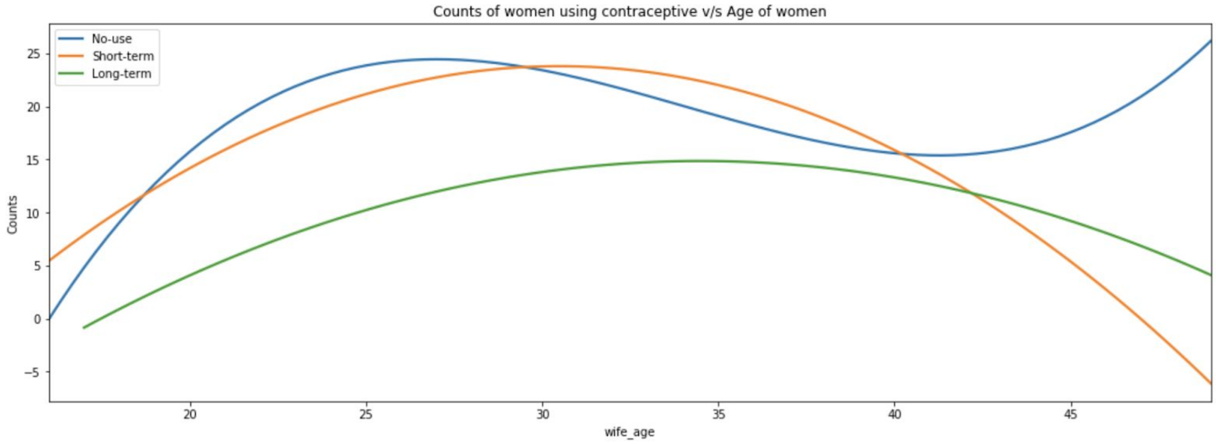
Some of the questions we had included:

- What are the best indicators to help predict what kind of contraception (if any) a married Indonesian woman used? Could her choice be best determined by the number of children she bore, her education, her religion, the characteristics of her spouse or any external sources such as media exposure and her working status.
- Are there any co-founding factors that affect this dataset? Such as husband education and wife education, or wife education and media exposure?
- What could be the least useful indicator to predict the kind of contraception a woman uses?

To understand the dataset, we plotted various graphs to understand how different features affected a woman's choice of contraceptive. This plot visualizes the relationship between the age of a woman and the use of contraceptives.

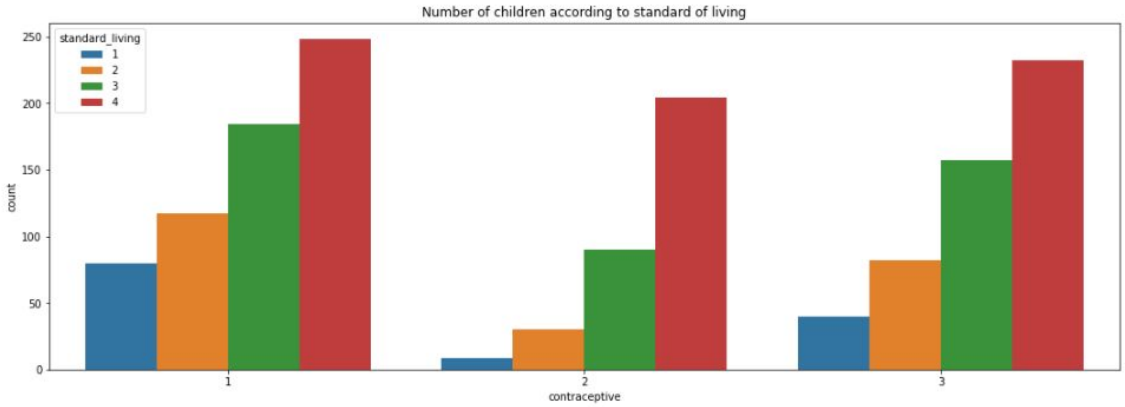


We want to try and capture a general trend in the use of contraceptives, so we fit minimum degree polynomials to the points and smooth the jagged lines:

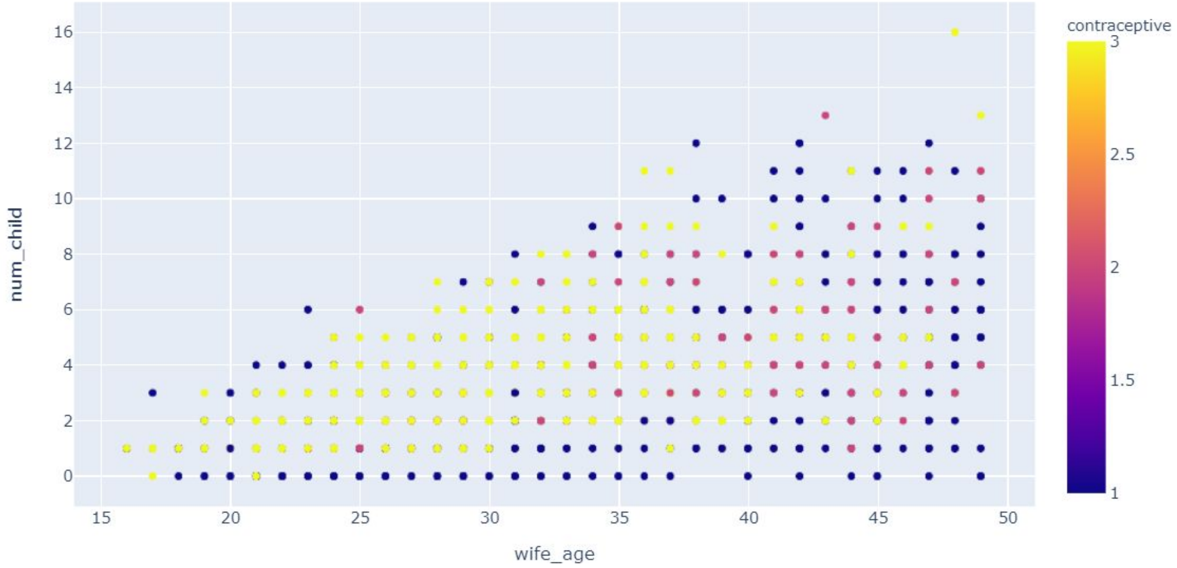


Even if the degrees of the lines are arbitrarily chosen, there still are some interesting relationships. For instance, the general trend in the use of long-term contraceptives indicates that women around 35 years old tend to use them more than women of other ages. Interestingly, all 3 lines have a decreasing trend past age 35 which could simply be because there is less data about people above that age. Simultaneously, the absence of women beyond the age of 48 makes intuitive sense since most women reach menopause by this age eliminating the need for contraception. An important aspect of this graph is that it captures only a single snapshot in time. Hence, this is not a longitudinal study and we are not observing how the same women's use of contraceptives change over time.

Another interesting observation is that amongst those using contraceptive, there seems to be a pattern between type of contraception and standard of living.



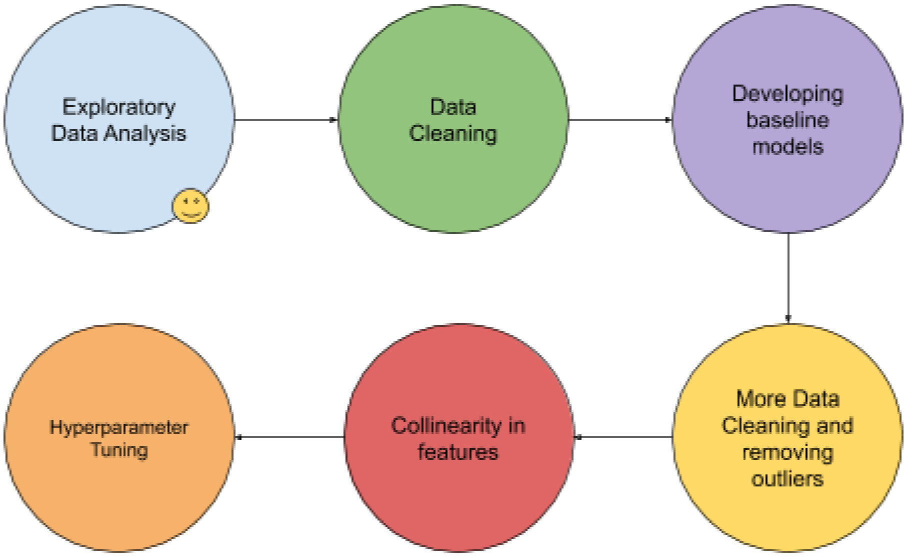
It seems that the majority of people in every standard of living bracket prefer shorter term contraception (such as pills, and condoms) over longer term contraception (such as intra-uterine devices or vasectomy). Under the simplifying assumption that standard of living is a proxy for family income, this could point to confounding factors such as the cost of long term contraception and associated surgeries which could possibly explain low participation from people with the lowest standard of living. However a deeper probe into the use of contraceptives shows a more interesting trend among different age groups.



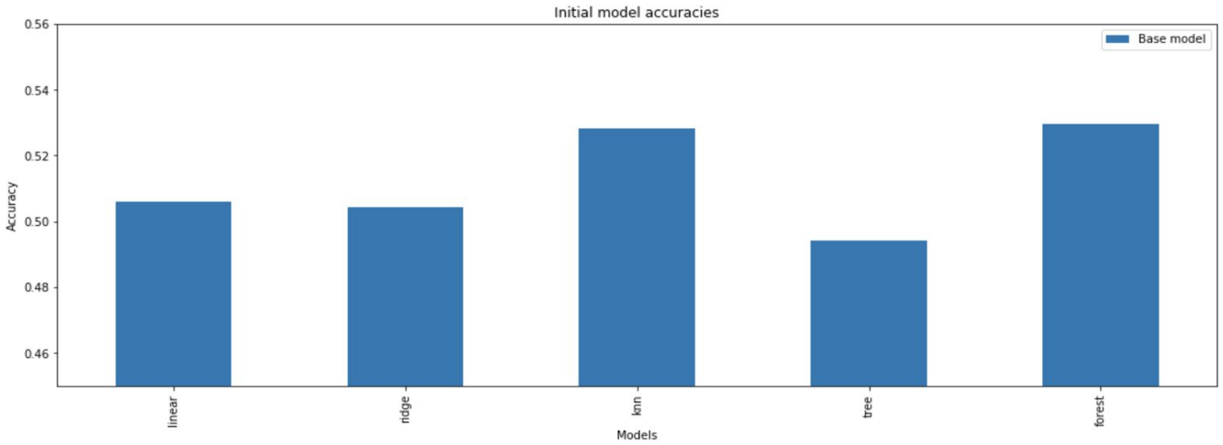
Keeping in mind the nature of the data (a single snapshot) a preliminary analysis shows some age specific trends. Contraceptive use is far more popular in those less than 35-36 year in age, as opposed to those above. In fact, the majority of the women who have 4 or more children (in the 25-35 age group) are already using contraceptive, with most women preferring short term contraceptive. In fact long term contraceptive is extremely uncommon in the less than 30 age group, especially since they are more permanent techniques that often require procedure to reverse.

Data Cleaning and Method and Experiments

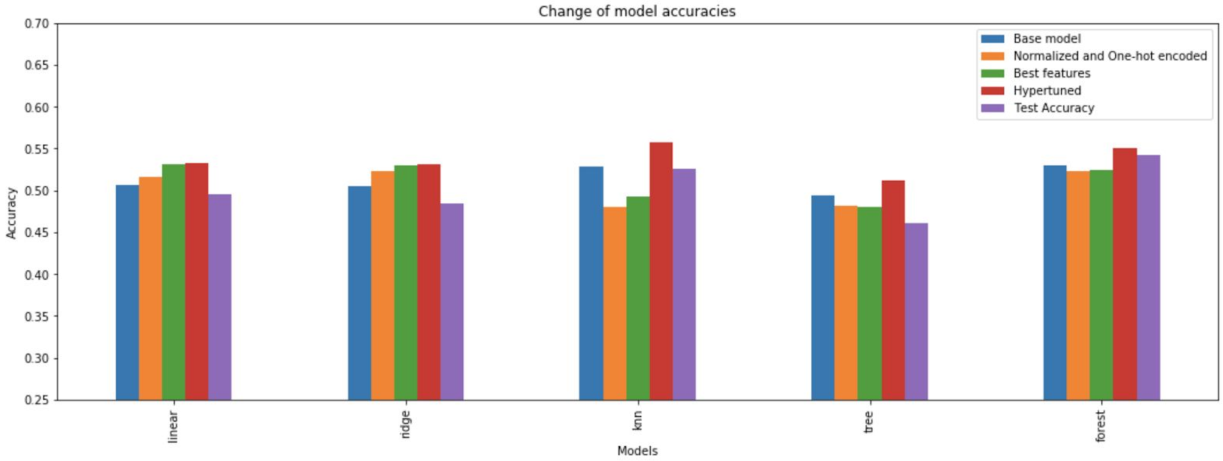
In our Jupyter notebook, we adopted a particular workflow for model improvement and experimentation:


Exploratory Data Analysis
Data Cleaning
Developing baseline models
More Data Cleaning and removing outliers
Collinearity in features
Hyperparameter Tuning

At this point, we have already conducted preliminary Exploratory Data Analysis. The dataset that was provided had no missing values. For **data cleaning**, we one-hot encoded all of our categorical variables and standardized our numerical variables. We also decided to remove outliers from our train set which we defined as being 3 standards deviations or more away from the mean for numerical variables. We chose 5 different models to try and predict the use of contraceptives: a linear model, a ridge classifier model (converts the target values into {-1, 1} and then treats the problem as a regression task), a k-nearest-neighbor classifier, a decision tree, and a random forest. To get a fair ground of comparison for our models, we will calculate their current accuracies. Note that in this step and in subsequent iterations, we are using 5-fold cross validation to get a sense of how well our models will perform on the test set. Here are how our initial models compare:



Our goal is to improve those models as much as possible. Through more data cleaning, selection of features, removal of outliers and hypertuning (all of which we go more into detail in the Jupyter notebook), we achieve the following results:



The purple bar represents the accuracy on the test set for each of our models. Our best classifier on the test set is a Random Forest with a 54% accuracy. We realized through experimentation, what improved the accuracy of the models. For instance, removing relatively highly associated features, such as the husband education or husband occupation, helped (most of the green bars are taller than the yellow bars) and hypertuning(regularization, number of neighbors,) our models also increased accuracy scores(all red bars are taller than the green bars). Some of the things that did not work, particularly for tree-based models, was using one-hot encoding as they tend to assign more importance to the numerical variables then. Using Pearson's correlation also did not work because we were working with a lot of categorical variables. Hence, we had to use a new statistic: Cramer's V. (a nominal variation of Pearson's Chi-Square test).

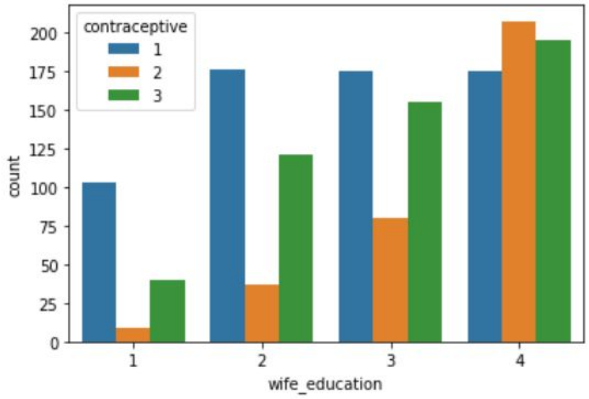
Model	Initial Parameters	Hypertuned Parameters
Linear	C: 1.0, penalty: L2	C: 0.80, penalty: L2
Ridge	alpha: 1.0, fit_intercept: True	alpha: 2.20, fit_intercept: True Alpha corresponds to 1 / (2C)
K-NN	5 Nearest Neighbors	151 Nearest Neighbors
Tree	criterion: gini, max_depth: None, max_leaf_nodes: None	criterion: gini, max_depth: 3, max_leaf_nodes: 4
Random Forest	criterion: gini, max_depth: None, max_leaf_nodes: None, n_estimators: 100	criterion: gini, max_depth: 3, max_leaf_nodes: 6, n_estimators: 110

The majority of the tuned hyperparameters prevent the models from overfitting. For instance, the linear and ridge models have higher L2 regularization terms and the tree based models have limited depth and leaf nodes as compared to the initial models with no limit to depth or number or leaves.

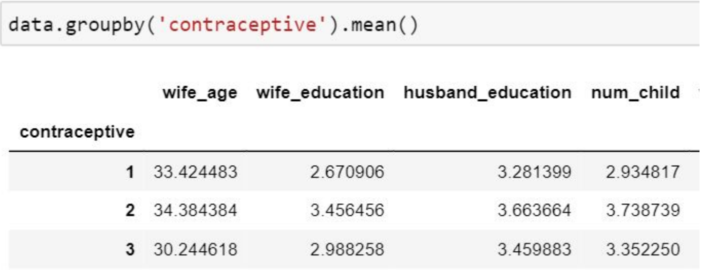
Analysis and Conclusion

- **What were two or three of the most interesting features you came across for your particular question?**

While trying to understand the factors that affected the prediction of the type of contraceptive used, we realized a couple of interesting facts. We observed that the wife's working status did not offer much insight into the use of contraceptives. This was surprising since we expected working women to be more likely to use contraceptives as compared to their homemaker counterparts due to assumptions about career path, however this result is a reminder than the Indonesian economy and the jobs were not as robust in the 1980's as they are right now.



On the other hand, a woman's education level seemed to be a good indicator of the type of contraceptive used. There seems to be a clear trend where use of contraceptives (short and long term) increases steadily with increase in education level, and interestingly, at the highest education level, long term contraceptive seems to be more popular than short term contraceptive. Therefore education, and literacy could have helped generate higher awareness about family planning.



```
data.groupby('contraceptive').mean()
```

	wife_age	wife_education	husband_education	num_child
contraceptive				
1	33.424483	2.670906	3.281399	2.934817
2	34.384384	3.456456	3.663664	3.738739
3	30.244618	2.988258	3.459883	3.352250

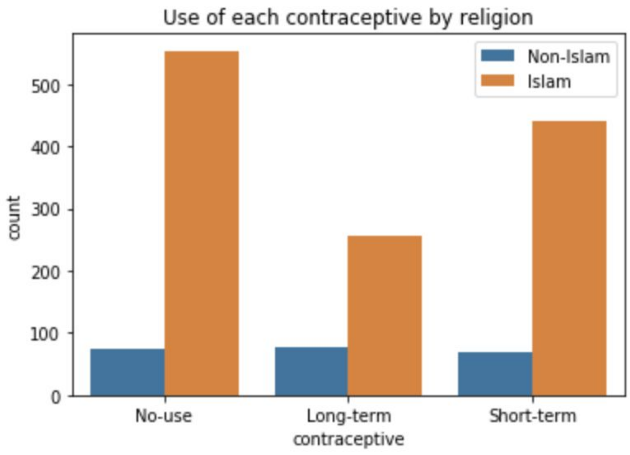
Another interesting phenomena is the fact that amongst all the groups of women, on average the ones using no contraceptives have the lowest number of average children (3.28) whereas people using long-term contraceptives have the highest number of children on average of the three groups. However, though these facts seem counterintuitive, we believe that there are a couple of possible explanations.

First explanation: there is a different adoption pattern in Indonesia where long term contraceptive use could be most prevalent amongst women who have already given birth to too many children and are now trying to control this issue by adopting the most cautious techniques.

Second explanation: the intimate nature of the questions, and the setting of the interview (which may happen in the presence of other family members) which might affect the accuracy of the participants' answers resulting is skewed instances of contraceptive uses at the extremes.

- **Describe one feature you thought would be useful, but turned out to be ineffective.**

We thought that the wife’s religion would be a relatively important determinant of the type of contraceptive used. In our EDA, we noticed that about 85% of the women identified as Muslims, which is close to the official data on the percentage of Muslims in Indonesia. Knowing that Indonesia is the world's most populous Muslim-majority nation, we expected a higher rate of non-use in this group, due to the banning of contraceptives in Islam. We generated the following statistics:



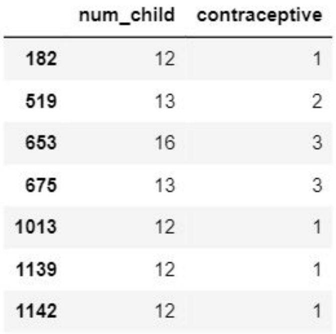
Type of contraceptive	General Population (%)	Islamic Population (%)	Non-Islamic Population (%)
No use	42.7	44.2	34.1
Long Term	22.6	20.5	34.5
Short Term	34.7	35.3	31.4

Only 44% of the Muslim women do use any form of contraceptives which is contrary to our prior expectations. In actuality, a majority of Muslim women in Indonesia use contraceptives. We also observe a significant difference in the use of contraceptives between the Muslim and non-Muslim population. In fact, the non-Islamic population seems to be indifferent to the type of contraceptive, with almost the same amount of people using both short term and long term contraceptives. By the end of our analysis, a bar plot of the feature importances for our random forest model also supported this viewpoint, as it assigned the ‘wife religion’ feature the lowest relative importance.

- **What challenges did you find with your data? Where did you get stuck?**

One of the challenges we faced was that our dataset was imbalanced — not necessarily in terms of contraceptive use, but of the participant’s standard of living. About 50% of the women claimed to have a high standard of living , which might have stemmed from a sampling bias; since the surveyors were going "door to door", they may have preferred traveling to 'nicer' neighborhoods. Another challenge inherent to the dataset being used was the presence of a lot of categorical variables. Hence, we were not able to use Pearson's correlation to determine the degree of association between variables, and instead used another statistic (Cramer's V).

- **What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?**



	num_child	contraceptive
182	12	1
519	13	2
653	16	3
675	13	3
1013	12	1
1139	12	1
1142	12	1

The biggest challenge is that though the survey collects information on the use of contraceptives, it provides no information about when the participant started using contraceptives, such as if they started using contraceptives after the first child, second, or third child. Furthermore, it also does not say if the participant is still using the contraceptive. For example, the image shows the choice of contraception for women with more than 11 children. Having a large number of children while not using contraceptives is more likely than having a dozen children while using contraception, therefore it's more plausible that that participant simply stopped her use of contraceptives, or only recently began using them. Therefore, our initial instinct was to get rid of outliers since they could hurt our overall prediction. Moreover, we made the simplifying assumption that the survey captured the woman's most recent status surrounding the use of contraceptives.

- **What ethical dilemmas did you face with this data?**

Predicting the use of contraceptives is in itself a sensitive issue. While the survey was conducted by the government to get a better understanding of the changing demographics of Indonesia, the way in which such a survey was conducted was quite questionable. The surveyors performed door-to-door to ask people the questions. Thus it is understandable that there might have been a response bias, as the women might have felt embarrassed to talk about the type of contraceptives they use, reproduction, and family planning. A better alternative would have been to mail the questionnaire anonymously or interview them privately to generate an atmosphere of trust. As a data scientist analyzing this dataset now, it is important to bear in mind these considerations.

- **What additional data, if available, would strengthen your analysis, or allow you test some other hypotheses?**

  We know that a major part of the dataset is missing from the original survey and we only have a partial view of the bigger picture. The survey itself consisted of multiple questions which have not been included in the dataset. For instance, some interesting questions asked by the surveyors included "*Have you ever used anything or tried in any way to delay or avoid getting pregnant?*", "*In the last month, have you heard or seen a message about family planning on the radio or the television?*" or "*Have you experienced any problem using a certain type of contraceptive?*". We think the answers to these questions would have provided some more insight into the usage of contraceptives. A single feature like the price of contraceptives would have been beneficial in predicting their use. Another possible source of information that would have strengthened our analysis would be access to data from similar types of study in different years. This would have allowed us to get more data points to observe adoption patterns of contraceptives. Similarly, data from different countries and their use of contraceptives could help test other hypotheses pertaining to the use of certain types of contraceptives and understanding if Indonesia has a peculiar consumption pattern. Therefore though the survey tries to gather maximum possible information, the dataset does not contain information about certain elements that need to be taken into consideration when choosing contraceptives, such as cost of surgery (for LT contraceptive), individual's medical conditions, access to family planning officers.

- **What ethical concerns might you encounter in studying this problem? How might you address those concerns?**

  We were pleasantly surprised that the dataset handed to us had no Personally Identifying information such as the names of the women, their address or their phone numbers. We hope that the participants in the survey agreed to have their data used and that written informed consent was obtained from each individual. This would allow them to know how their data is being used and ensure that their answers will be kept anonymous. This is even more applicable to a study that was conducted in 1987 and which is being analyzed now.

Through this project, we were able to explore the whole Data Science life cycle and answer some of our own open ended questions. We do realize that we analyzed only a few data points and that some of our conclusions might be flawed by the assumptions we made and by the lack of further information. According to our models, the most important indicators to predict the use of contraceptives is the number of children and the level of education of the woman whereas the least influential features are religion and wife working status.