Contraceptive prevalence rate

Nevil Abraham Elias

July 28, 2021

Abstract

This report was prepared as part of the capstone project for HarvardX's Data Science Professional Certificate Program.

Contents

1 Introduction

The United States Agency for International Development (USAID) selected Westinghouse Health Systems to carry out contraceptive surveys. The primary objectives of the Contraceptive Prevalence Surveys (CPS) are to determine periodically the levels of contraceptive use in the country; to examine the correlates of and differentials in these levels in order to assess the impact of various types of governmental and non governmental programs; to identify factors that will facilitate an increase in contraceptive use, particularly factors involved in program planning activities; and to institutionalize in each country the capability to design and implement studies of contraceptive prevalence, to be undertaken at regular intervals by an in-country agency.[1]

This document is structured as follows:
Chapter 1 - Introduction - Dataset, goal of the project and key steps involved.
Chapter 2 - Methods/Analysis - Data cleaning, exploration, visualisation.
Chapter 3 - Results - Modelling performance and results.
Chapter 4 - Conclusion - Summary, limitations and future works.

1.1 CPR Dataset

This data set[2] is from a 1987 National Indonesia Contraceptive Prevalence Survey. All observations are married women who were definitely not pregnant or did not know yet. Questions on the survey covered topics regarding socio-economic status and general demographics.

Variables in the dataset:
1. Age - age of the woman
2. Education - level of education woman has received (1=low, 4=high)
3. Partner Education - level of education partner has received (1=low, 4=high)
4. Number of Children - number of kids mothered by woman
5. Religion=Islam - woman that identify as Muslim (0=No, 1=Yes)
6. Currently Working - woman is currently employed (0=Yes, 1=No)
7. Husbands Occupation - Not specified (categorical 1-4)
8. Standard of Living - based on the standard of living index (1=low, 4=high)
9. Media exposure - quality of media exposure (0=Good, 1=Not good)
10. Contraceptive Method Used - 1=No-use, 2=Long-term, 3=Short-term.

Note: In the proceeding sections, some of the variable names will be changed and the 0's and 1's in the columns *Currently Working* and *Media Exposure* will be interchanged to make the data convenient and intuitive.

1.2 Goal

The goal of this project is to analyse the given data and understand the contraceptive use of couples. This information can aid the government in population control programs, to create more awareness among specific categories of people, as well private companies that make contraceptives to make necessary changes to their product.

1.3 Process

Key steps involved in this project include:

1. Project Understanding: understanding the project's goals and creating a workflow of key steps.

[1] https://pubmed.ncbi.nlm.nih.gov/12338999/
[2] https://www.kaggle.com/joelzcharia/contraceptive-prevalence-survey

2. Data Preparation: downloading, importing and preparing the dataset for analysis.
3. Data Exploration: to gain insights from the data and identify key features or predictors.
4. Data Preprocessing: involves cleaning and transforming the data, such as feature selection, scaling, removing unnecessary information, etc.
5. Modelling Methods: researching and selecting modelling approaches that work best for the type of dataset.
6. Data Modelling: Creating, training and testing the selected models, including any fine tuning of parameters.
7. Model Evaluation: Evaluating the results and model's performance.
8. Communication: Present the final results along with any limitations or recommendations for future work.

For any machine learning project, it is essential that the model performs well for both the available data and the real-world data. To ensure this, the dataset is initially split into two, a training set and a validation set. Steps 3 through 6 are performed on the training set (by the method of cross-validation) to select a model. This model is then trained on the entire training set and evaluated using the validation set.

2 Methods/Analysis

2.1 Data Preparation

As previously mentioned this step involves all the activities required to make the data ready for analysis ranging from downloading to splitting the data.

We start off by loading the necessary libraries.

```r
# Load libraries
library(tidyverse)
library(ggthemes)
library(caret)
library(MASS) #for polr() function
library(class) #for k-nearest neighbour
library(rpart) # for decision tree
library(knitr) #A General-Purpose Package for Dynamic Report Generation in R
library(kableExtra)
library(tinytex)
library(latexpdf)
# set global options
options(timeout=10000, digits=3)
```

2.1.1 Downloading

Because the dataset can't be directly downloaded into RStudio using the Kaggle link, it is first uploaded to the GitHub Repo and then downloaded from the repo.

```r
# Download and import dataset
urlfile<-"https://raw.githubusercontent.com/nevillio/Harvardx-capstone2/main/cpr_dataset.csv"
cpr<-read_csv(urlfile,col_types = cols())
```

2.1.2 Data Transformation

We'll start our exploration with these functions

```
class(cpr) # type of dataset
```

```
# [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
dim(cpr) # no. of rows and columns in the dataset
```

```
# [1] 1473   10
```

The cpr dataset is of class type "data frame" and there are 1473 observations and 10 features (columns).

```
str(cpr) # structure of the dataset
```

```
# spec_tbl_df [1,473 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
#  $ Age                      : num [1:1473] 24 45 43 42 36 19 38 21 27 45 ...
#  $ Education                : num [1:1473] 2 1 2 3 3 4 2 3 2 1 ...
#  $ Partner Education        : num [1:1473] 3 3 3 2 3 4 3 3 3 1 ...
#  $ Number of Children       : num [1:1473] 3 10 7 9 8 0 6 1 3 8 ...
#  $ Religion = Islam         : num [1:1473] 1 1 1 1 1 1 1 1 1 1 ...
#  $ Currently working        : num [1:1473] 1 1 1 1 1 1 1 0 1 1 ...
#  $ Husband Occupation       : num [1:1473] 2 3 3 3 3 3 3 3 3 2 ...
#  $ Standard of Living       : num [1:1473] 3 4 4 3 2 3 2 2 4 2 ...
#  $ Media Exposure           : num [1:1473] 0 0 0 0 0 0 0 0 0 1 ...
#  $ Contraceptive Method Used: num [1:1473] 1 1 1 1 1 1 1 1 1 1 ...
#  - attr(*, "spec")=
#   .. cols(
#   ..   Age = col_double(),
#   ..   Education = col_double(),
#   ..   `Partner Education` = col_double(),
#   ..   `Number of Children` = col_double(),
#   ..   `Religion = Islam` = col_double(),
#   ..   `Currently working` = col_double(),
#   ..   `Husband Occupation` = col_double(),
#   ..   `Standard of Living` = col_double(),
#   ..   `Media Exposure` = col_double(),
#   ..   `Contraceptive Method Used` = col_double()
#   .. )
```

```
head(cpr) # first few rows of the dataset
```

Age	Education	Partner Education	Number of Children	Religion = Islam	Currently working	Husband Occupat
24	2	3	3	1	1	
45	1	3	10	1	1	
43	2	3	7	1	1	
42	3	2	9	1	1	
36	3	3	8	1	1	
19	4	4	0	1	1	

Before splitting the dataset into training and test sets, we will make some basic changes to the variables. This won't be considered data snooping/leaking as we're not modifying any of the variables based on information that could be found only in the test set. The following changes will be made at this stage.

1. Replacing white spaces in the column names with "_". (Eg. Media Exposure to Media_exposure).
2. Renaming columns *Religion = Islam* to 'Religion' and *Number_of_Children* to 'n_children'
3. Renaming the categories in *Contraceptive Method Used* as "None","Long-term" and "Short-term" and those in *Religion* as "Muslim" and "Non-muslim".
4. Swapping the values in the columns *Current Occupation* and *Media Exposure* since 1 being the favourable scenario is quite intuitive.
5. Converting all categorical variables from numeric to factors.

The following code executes these changes:

```
#Remove whitespaces from column names
names(cpr) <- str_replace_all(names(cpr)," ","_")

#Change name and values of 'Religion = Islam' Column
cpr <- rename(cpr, Religion = 'Religion_=_Islam',n_children = Number_of_Children)

#view the modified column names
names(cpr)

#  [1] "Age"                       "Education"
#  [3] "Partner_Education"         "n_children"
#  [5] "Religion"                  "Currently_working"
#  [7] "Husband_Occupation"        "Standard_of_Living"
#  [9] "Media_Exposure"            "Contraceptive_Method_Used"

#Renaming categories of required columns
cpr <- cpr %>%
# Renaming categories in Contraceptive_Method_Used
    mutate(Contraceptive_Method_Used =
               factor(
                   case_when(
                       Contraceptive_Method_Used == 1 ~ "None",
                       Contraceptive_Method_Used == 2 ~ "Long Term",
                       Contraceptive_Method_Used == 3 ~ "Short Term",
                   ),
                   levels = c("None","Short Term","Long Term")
               ),
# Renaming categories in Religion
           Religion = ifelse(Religion == 1, "Muslim", "Non_muslim")
)

# Swapping values
## Creating a swapping function
swap <- function(x) {
    x = x-1
    x = ifelse(x==-1,-x,x)
    x
}

## Swapping 0's and 1's in required columns
cpr$Media_Exposure <- swap(cpr$Media_Exposure)
cpr$Currently_working <- swap(cpr$Currently_working)
```

```
# Converting categorical variables to factors, these variables have at most 4 unique values
cpr[sapply(cpr, function(x) n_distinct(x)<5)] <-
        lapply(cpr[sapply(cpr,function(x)n_distinct(x)<5)],as.factor)

#Result of all Transformations
str(cpr)

# spec_tbl_df [1,473 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
#  $ Age                      : num [1:1473] 24 45 43 42 36 19 38 21 27 45 ...
#  $ Education                : Factor w/ 4 levels "1","2","3","4": 2 1 2 3 3 4 2 3 2 1 ...
#  $ Partner_Education        : Factor w/ 4 levels "1","2","3","4": 3 3 3 2 3 4 3 3 3 1 ...
#  $ n_children               : num [1:1473] 3 10 7 9 8 0 6 1 3 8 ...
#  $ Religion                 : Factor w/ 2 levels "Muslim","Non_muslim": 1 1 1 1 1 1 1 1 1 1 ...
#  $ Currently_working        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
#  $ Husband_Occupation       : Factor w/ 4 levels "1","2","3","4": 2 3 3 3 3 3 3 3 3 2 ...
#  $ Standard_of_Living       : Factor w/ 4 levels "1","2","3","4": 3 4 4 3 2 3 2 2 4 2 ...
#  $ Media_Exposure           : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 1 ...
#  $ Contraceptive_Method_Used: Factor w/ 3 levels "None","Short Term",..: 1 1 1 1 1 1 1 1 1 1 ...
#  - attr(*, "spec")=
#   .. cols(
#   ..   Age = col_double(),
#   ..   Education = col_double(),
#   ..   `Partner Education` = col_double(),
#   ..   `Number of Children` = col_double(),
#   ..   `Religion = Islam` = col_double(),
#   ..   `Currently working` = col_double(),
#   ..   `Husband Occupation` = col_double(),
#   ..   `Standard of Living` = col_double(),
#   ..   `Media Exposure` = col_double(),
#   ..   `Contraceptive Method Used` = col_double()
#   .. )
```

Now, we're ready to split the data.

2.1.3 Splitting data

The dataset is split into a training set containing 70% of data and test set containing 30% data, called cpr_train and cpr_test respectively. The cpr_train dataset will be used to explore the dataset and train our models. The cpr_test dataset acts as the final hold-out test set (data not seen by the model) to evaluate the model's performance.

The reason for a 70-30 split is the small size of data. As the size of the training set becomes smaller, it is subjected to large variations in prediction accuracy for relatively small changes. Hence a size ratio closer to 2:1 is selected. The following code achieves the splitting of data:

```
# Split the dataset into training and test sets
set.seed(1, sample.kind = "Rounding")
test_index <- createDataPartition(y=cpr$Contraceptive_Method_Used, times = 1, p = 0.3, list = FALSE)
cpr_train <- cpr[-test_index,]
cpr_test <- cpr[test_index,]
rm(test_index,cpr)
nrow(cpr_train) #check number of observations of train set

# [1] 1030
```

```
nrow(cpr_test) #check number of observations of test set
```

```
# [1] 443
```

The train set has 1030 observations and test set has 443 observations.

Since our dataset is not very large, we won't be doing a further split to get a validation set. Instead we will use repeated k-fold cross-validation when training our models, which will be covered later. This works better when dealing with small to medium sized datasets.

2.2 Data Exploration

In this step, we will be exploring each variable and identifying possible predictors. For easier plotting, we'll begin with a summary of each variable.

```
sapply(cpr_train, n_distinct) #No.of unique values in each variable
```

```
#                       Age                 Education         Partner_Education
#                        34                         4                         4
#                n_children                  Religion         Currently_working
#                        14                         2                         2
#        Husband_Occupation        Standard_of_Living            Media_Exposure
#                         4                         4                         2
# Contraceptive_Method_Used
#                         3
```

```
summary(cpr_train) #Summary of distribution of each variable
```

```
#       Age       Education Partner_Education   n_children         Religion
#  Min.   :16.0   1:102     1: 27             Min.   : 0.00   Muslim    :868
#  1st Qu.:25.2   2:236     2:124             1st Qu.: 1.00   Non_muslim:162
#  Median :32.0   3:289     3:255             Median : 3.00
#  Mean   :32.3   4:403     4:624             Mean   : 3.21
#  3rd Qu.:38.0                               3rd Qu.: 4.00
#  Max.   :49.0                               Max.   :13.00
#  Currently_working Husband_Occupation Standard_of_Living Media_Exposure
#  0:774             1:300              1: 95              0: 71
#  1:256             2:283              2:148              1:959
#                    3:424              3:316
#                    4: 23              4:471
#
#
#  Contraceptive_Method_Used
#  None      :440
#  Short Term:357
#  Long Term :233
#
#
#
```

Some preliminary observations from the summary are listed below:

1. Majority of the women are:

    a. Muslims.
    b. Unemployed.
    c. Having higher standards of living.
    d. Exposed to media.

2. Women do not pursue the higher levels of education as much as men do, that is, most women may not be as educated as their husbands.
3. Only a little higher than 40% of couples use contraceptives.
4. Among the couples using contraceptives, short-term contraceptives are prevalent.

Since the summary provides sufficient details of all categorical variables, we will proceed with detailed exploration of numeric variables only.

2.2.1 Age

This column shows the age of the surveyed woman. Since a summary was provided in the last section, we'll directly move into plotting the distribution of age.

```
cpr_train %>%
  ggplot(aes(Age)) +
  geom_histogram(col = "black",binwidth = 1) +
  ggtitle("Age Distribution of Surveyed Women") +
  xlab("Age") +
  ylab("Count of Women")
```

We'll also look at the distribution by religion.

```
cpr_train %>%
  ggplot(aes(Age, col = Religion)) +
  geom_density() +
  ggtitle("Age Distribution by Religion") +
  xlab("Age") +
  ylab("Count of Women") +
  theme(legend.position = "top")
```

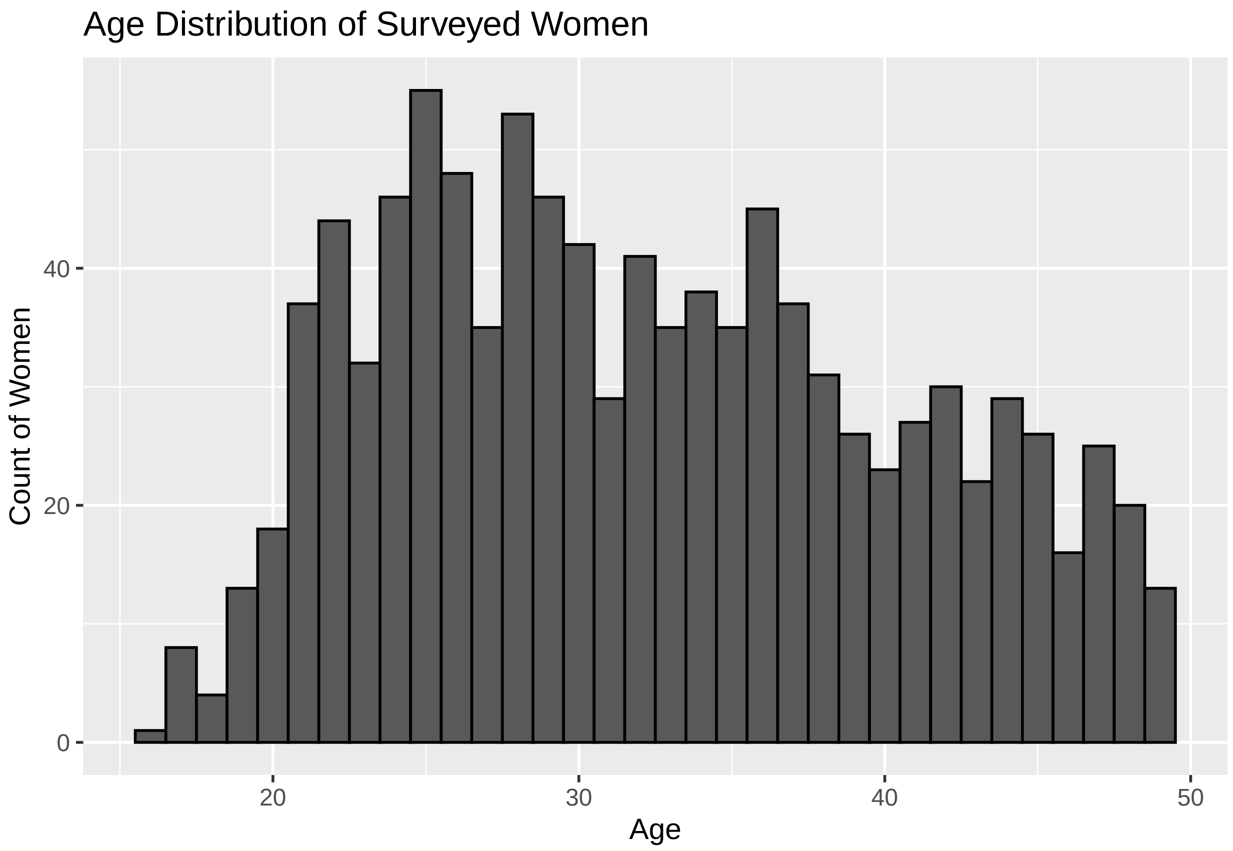
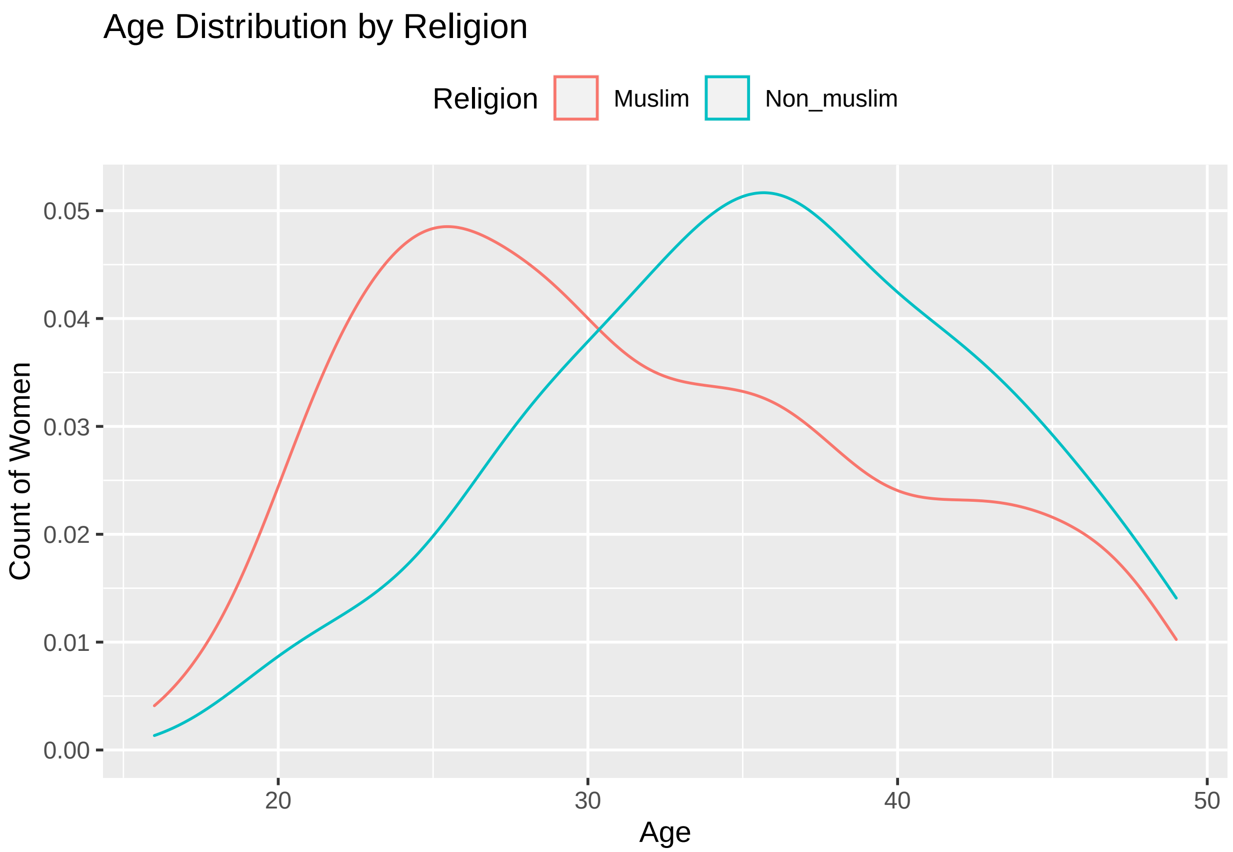


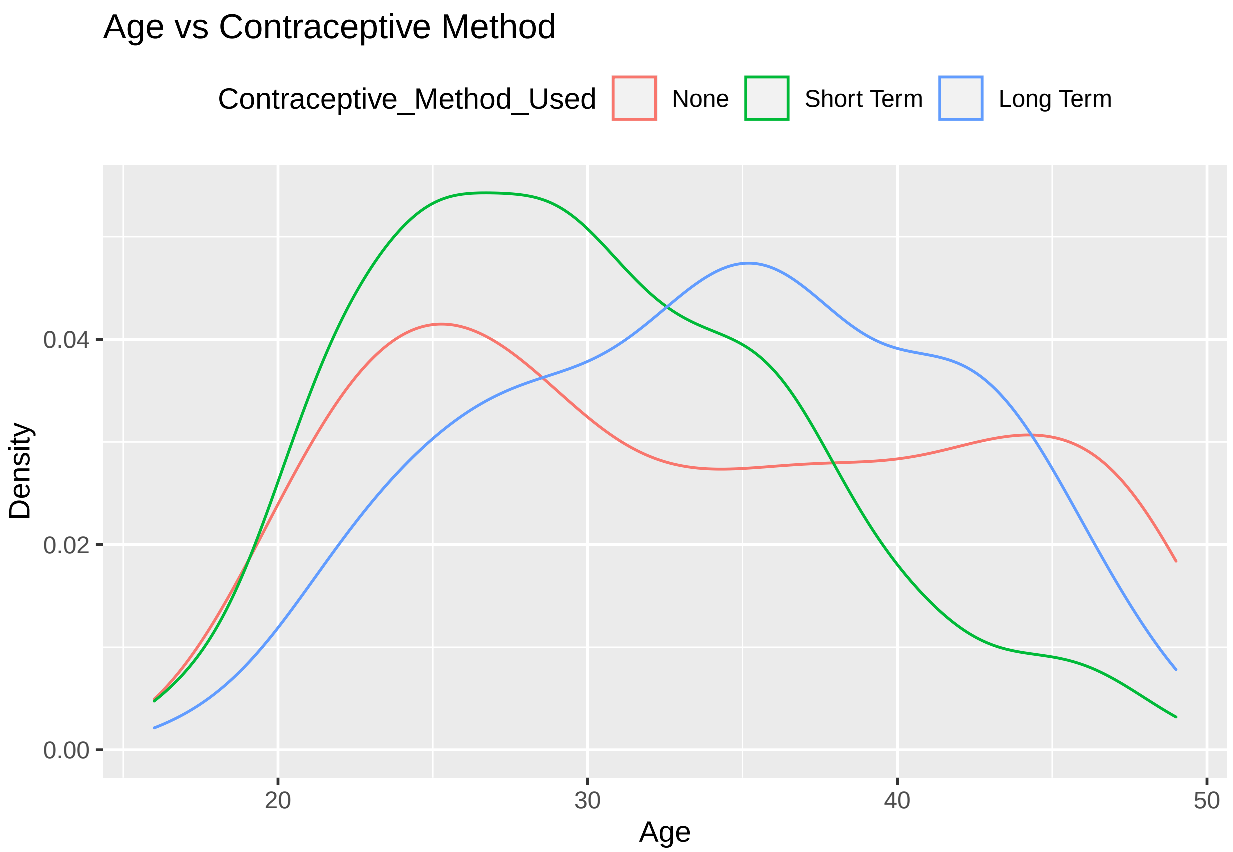
Figure 1: Distribution of Age of Surveyed women



From the above plot we infer that Muslim women are married off earlier than women of other religion.

Now we'll look at the distribution of Age by contraceptive method used.

```
cpr_train %>%
  ggplot(aes(Age, col = Contraceptive_Method_Used)) +
  geom_density() +
  ggtitle("Age vs Contraceptive Method") +
  xlab("Age") +
  ylab("Density") +
  theme(legend.position = "top")
```

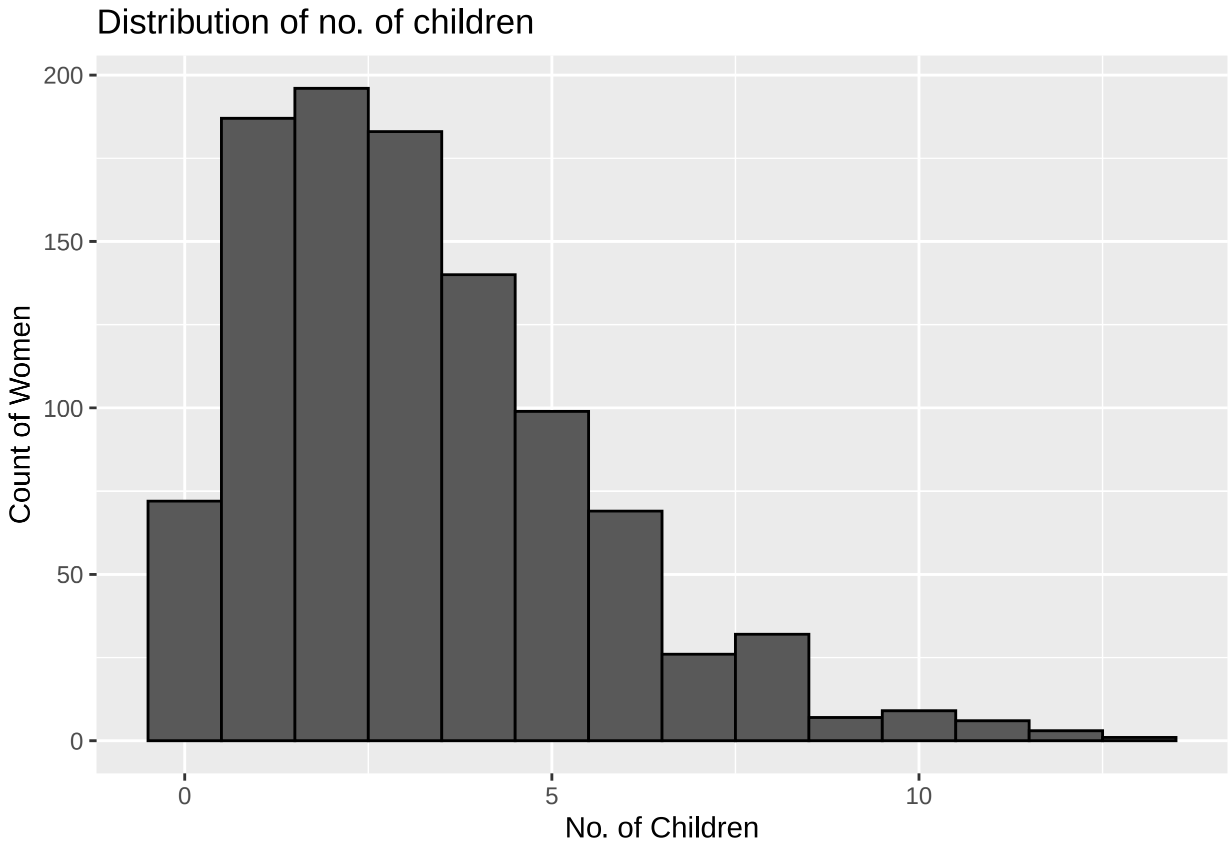


The plot shows that married couples that use contraceptives resort to short term methods in early years of marriage and long term methods in later years but stop altogether at mid forties due to menopause.

2.2.2 n_children

We know intuitively that if a couple wants to limit the number of children they'll employ some birth control mechanism. We'll validate our intuition in this subsection.

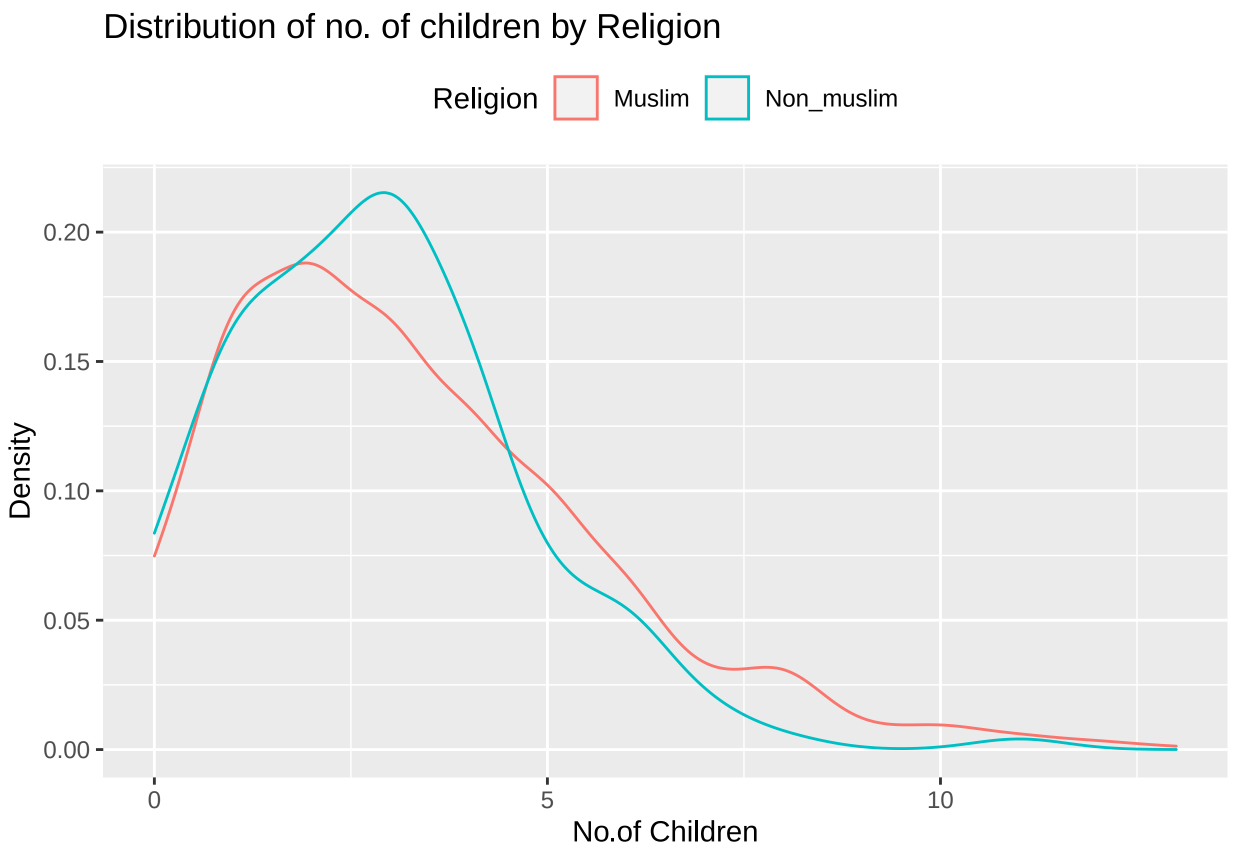
```r
cpr_train %>%
  ggplot(aes(n_children)) +
  geom_histogram(col = "black", binwidth = 1) +
  ggtitle("Distribution of no. of children ") +
  xlab("No. of Children") +
  ylab("Count of Women")
```



We see that a majority of the surveyed women have between 1 to 5 children.

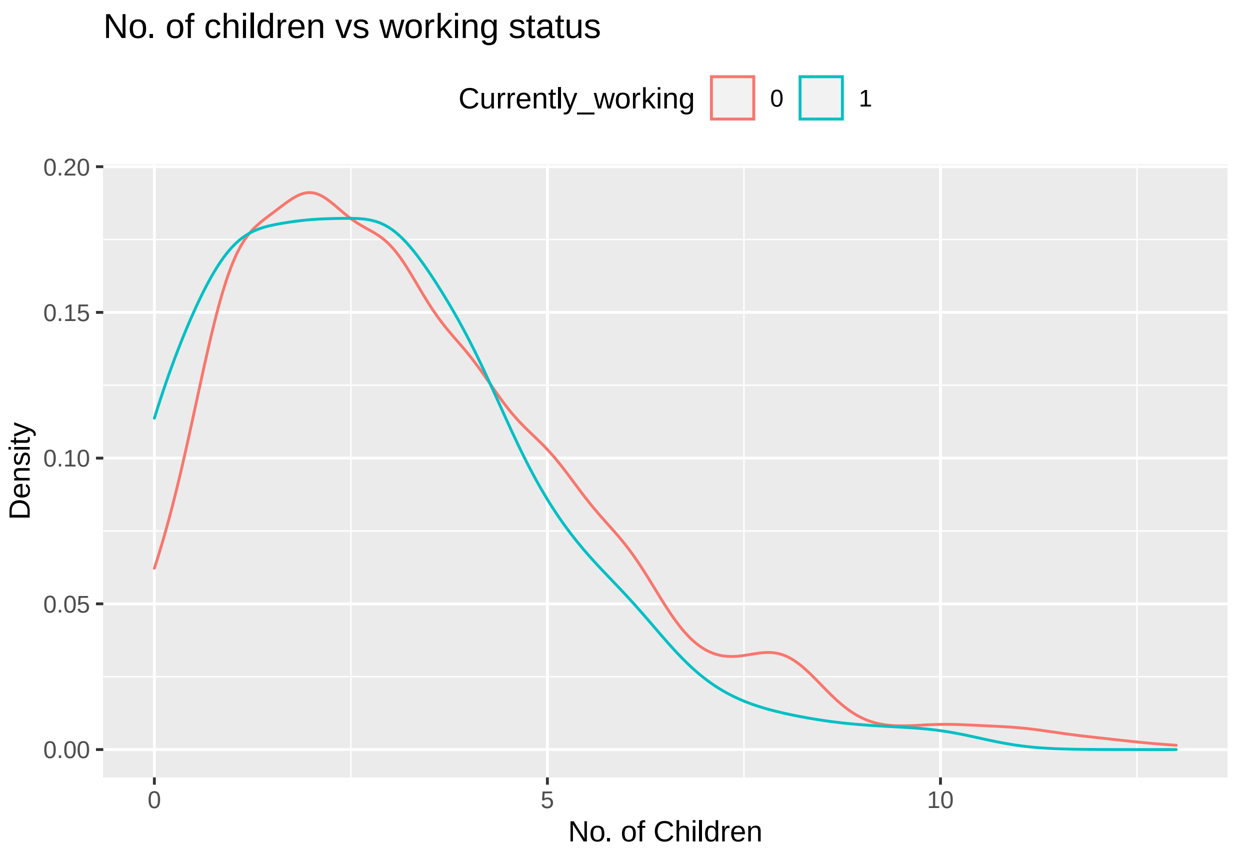
Now, we'll look at the same distribution grouped by categorical variables

```
cpr_train %>%
  ggplot(aes(n_children, col = Religion)) +
  geom_density() +
  ggtitle("Distribution of no. of children by Religion") +
  xlab("No.of Children") +
  ylab("Density") +
  theme(legend.position = "top")
```



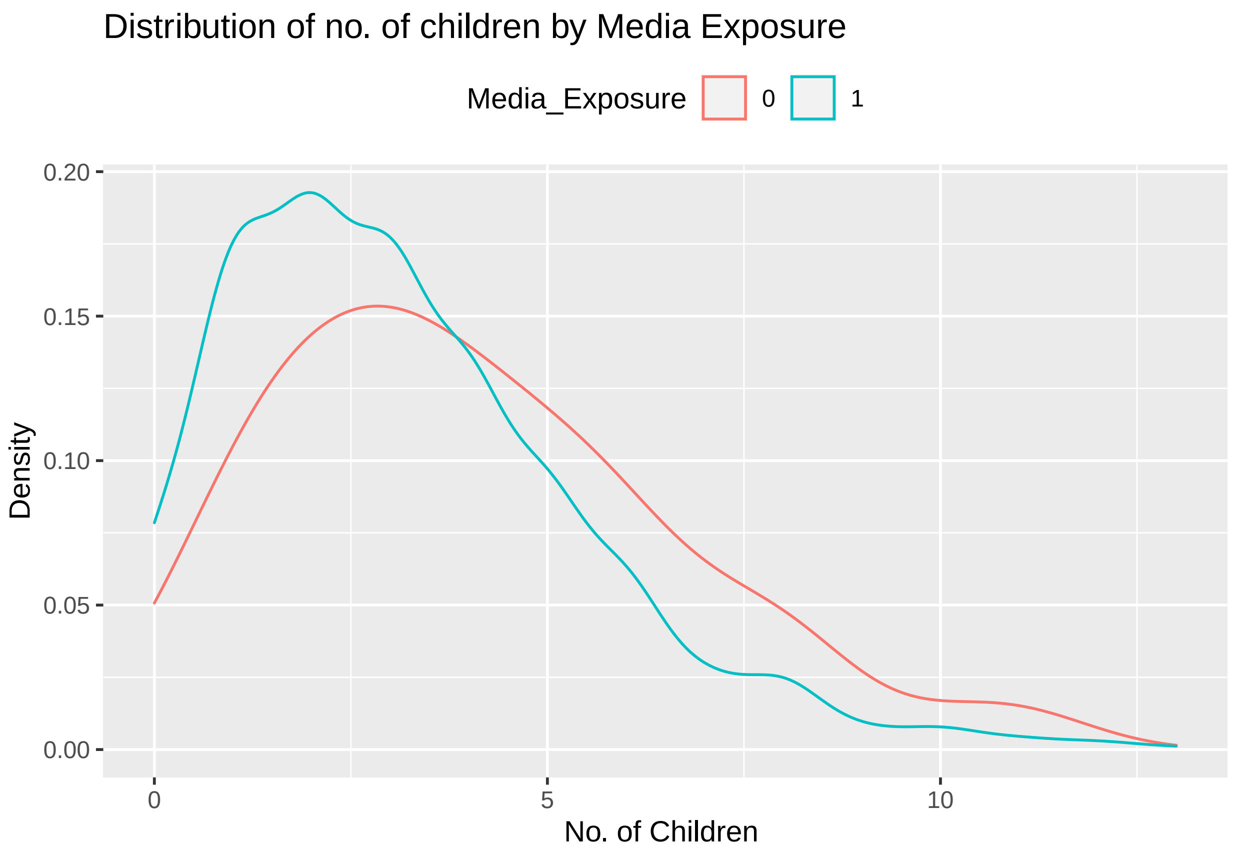
We see that Non-muslim women tend to have less number of children than Muslim women on average.

```
cpr_train %>%
  ggplot(aes(n_children, col = Currently_working)) +
  geom_density() +
  ggtitle("No. of children vs working status") +
  xlab("No. of Children") +
  ylab("Density") +
  theme(legend.position = "top")
```



The difference is less significant compared to the previous one.

```
cpr_train %>%
  ggplot(aes(n_children, col = Media_Exposure)) +
  geom_density() +
  ggtitle("Distribution of no. of children by Media Exposure") +
  xlab("No. of Children ") +
  ylab("Density") +
  theme(legend.position = "top")
```



The media clearly creates awareness about population control.

```
# distribution by education level
cpr_train %>%
  ggplot(aes(n_children, col = Education)) +
  geom_density() +
  ggtitle("Distribution of no. of children by Wife's Education level") +
  xlab("No. of Children") +
  ylab("Density") +
  theme(legend.position = "top")

# distribution by partner's education level
cpr_train %>%
  ggplot(aes(n_children, col = Partner_Education)) +
  geom_density() +
  ggtitle("Distribution of no. of children by Husband's Education level") +
  xlab("No. of Children") +
  ylab("Density") +
  theme(legend.position = "top")
```

With increased education level couples participate more in population control.

```
cpr_train %>%
  ggplot(aes(n_children, col = Standard_of_Living)) +
  geom_density() +
  ggtitle("Distribution of no. of children by Standard of Living") +
```

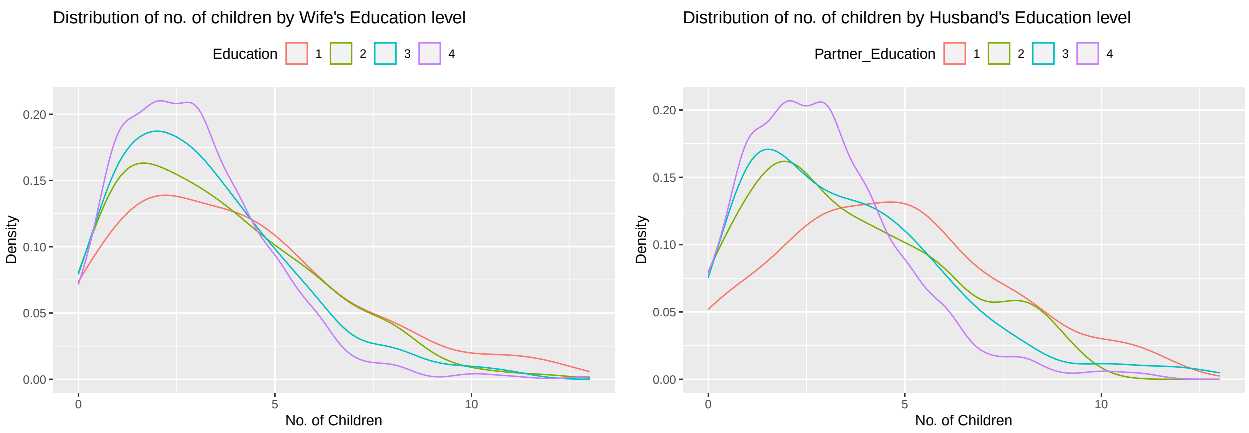
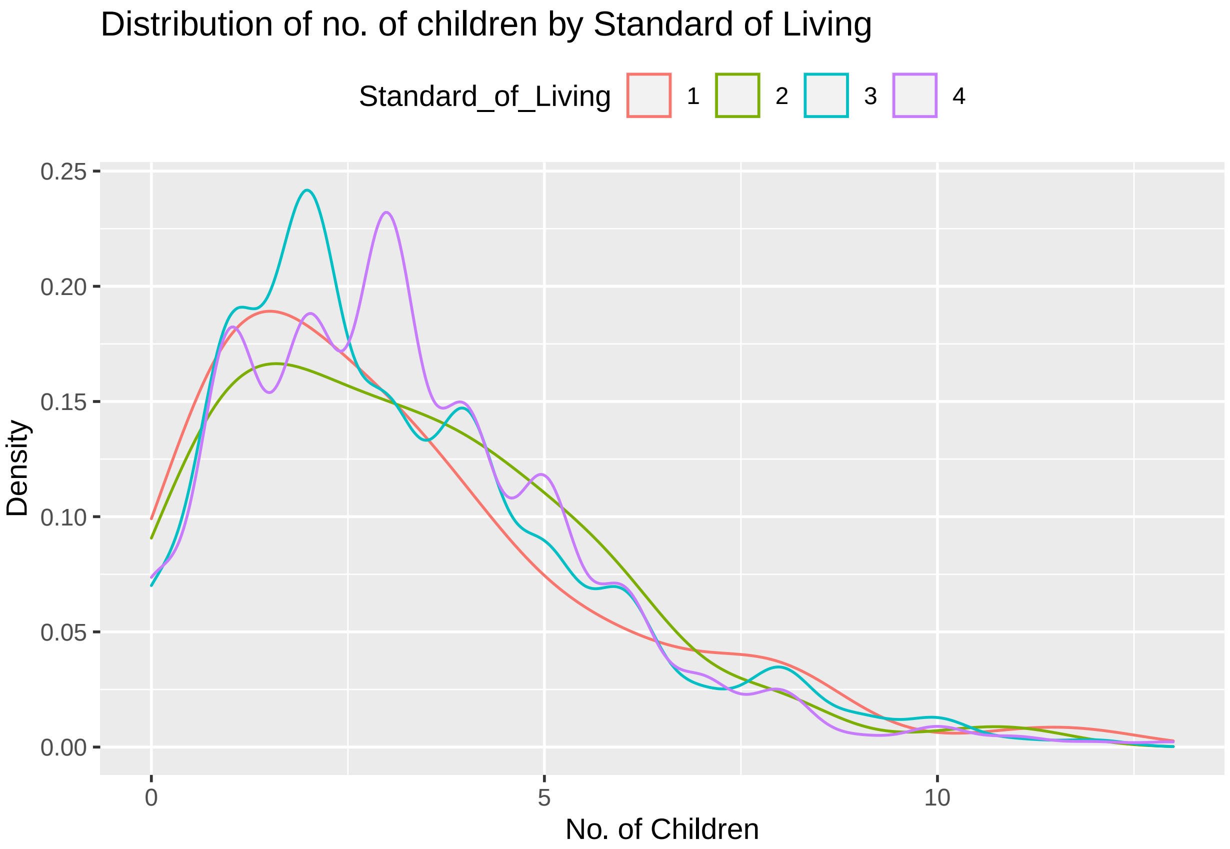


Figure 2: Distribution of no. of children by Education level

```
  xlab("No. of Children") +
  ylab("Density") +
  theme(legend.position = "top")
```



We loosely make out that people living in higher standards also tend to have less children on average.

```
cpr_train %>%
  ggplot(aes(n_children, col = Husband_Occupation)) +
  geom_density() +
  ggtitle("No. of children vs Husband Occupation") +
  xlab("No.of children") +
  ylab("Density") +
  theme(legend.position = "top")
```

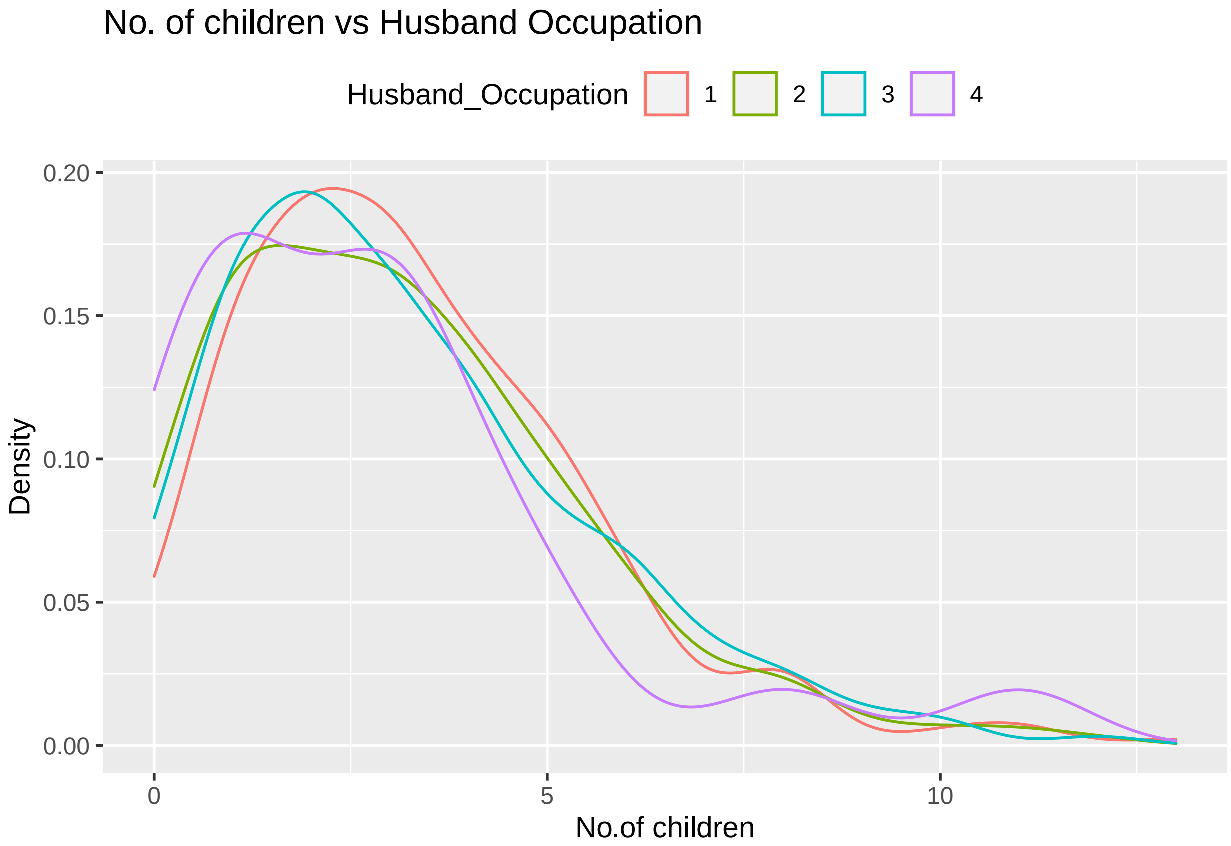


Figure 3: No.of Children by Husband Occupation

Again we see almost the same pattern.

```
# Here we are contrasting the use or non-use of contraceptives
# Hence both the use categories are first combined
cpr_train %>%
  mutate(Use_of_Contraceptive =
                ifelse(Contraceptive_Method_Used == "None",
                           "No", "Yes")) %>%
  ggplot(aes(n_children,
             fill = Use_of_Contraceptive)) +
  geom_density() +
  geom_histogram(col = "black", position = "dodge", binwidth = 1) +
  ggtitle("Distribution of no. of children by use of Contraceptives") +
  xlab("No. of Children") +
  ylab("Count of Women") +
theme(legend.position = "top")
```

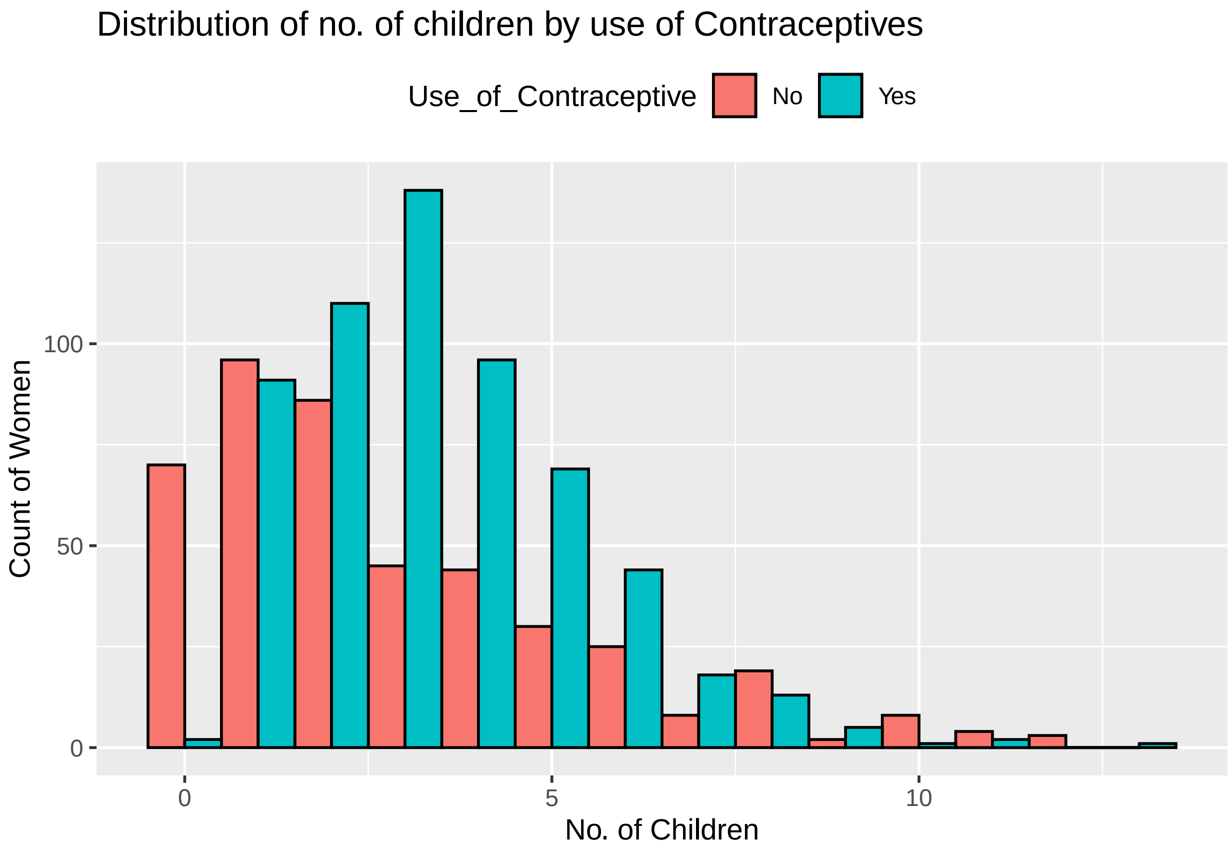


Figure 4: Distribution of no. of children by use of contraceptive

We see that use of contraceptives start only after the birth of the first child. Our intuition that birth control is associated with limiting the number of children is also validated.

```
# Here we are contrasting the use categories
# Hence we filter out the non-use category first
cpr_train %>%
  dplyr::filter(Contraceptive_Method_Used != "None") %>%
  ggplot(aes(n_children,
             fill = Contraceptive_Method_Used)) +
  geom_histogram(binwidth = 1, position = "dodge",col = "black") +
  ggtitle("Distribution of no. of children by use of Contraceptives") +
  xlab("No. of Children") +
  ylab("Count of Women") +
  theme(legend.position = "top")
```

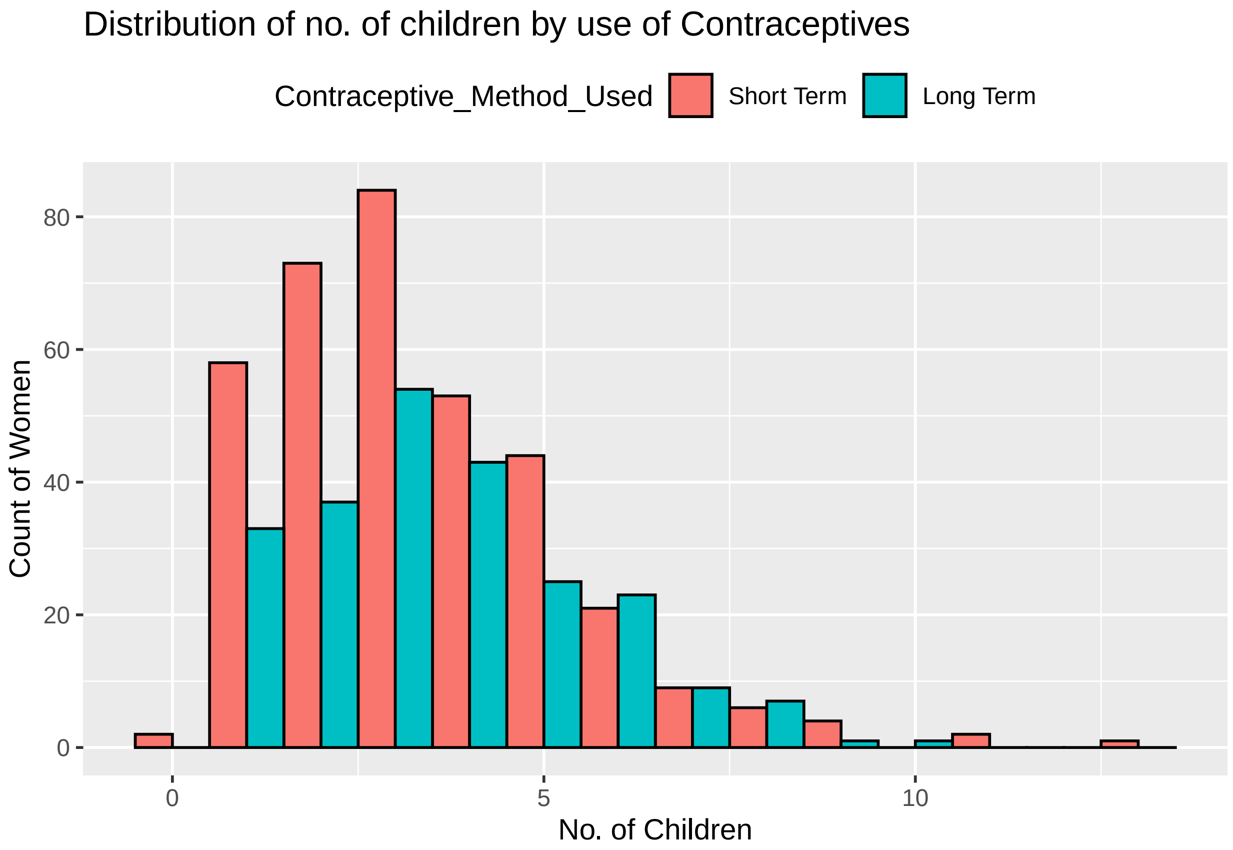


Figure 5: Distribution of no. of children by Contraceptive Method

Some couples are more decisive about the number of children they want and are quicker to resort to permanent methods.

2.2.3 Contraceptive_Method_Used

This is the target variable of our project. Now, we'll explore the correlation of this variable with other categorical variables.

```
cpr_train %>%
  ggplot(aes(Contraceptive_Method_Used, fill = Education)) +
  geom_bar(position = "dodge",col = "black") +
  ggtitle("Contraceptive Method vs Education level") +
  xlab("Contraceptive Method Employed") +
  ylab("Count of Women") +
  theme(legend.position = "top") +
  theme_economist_white()
```

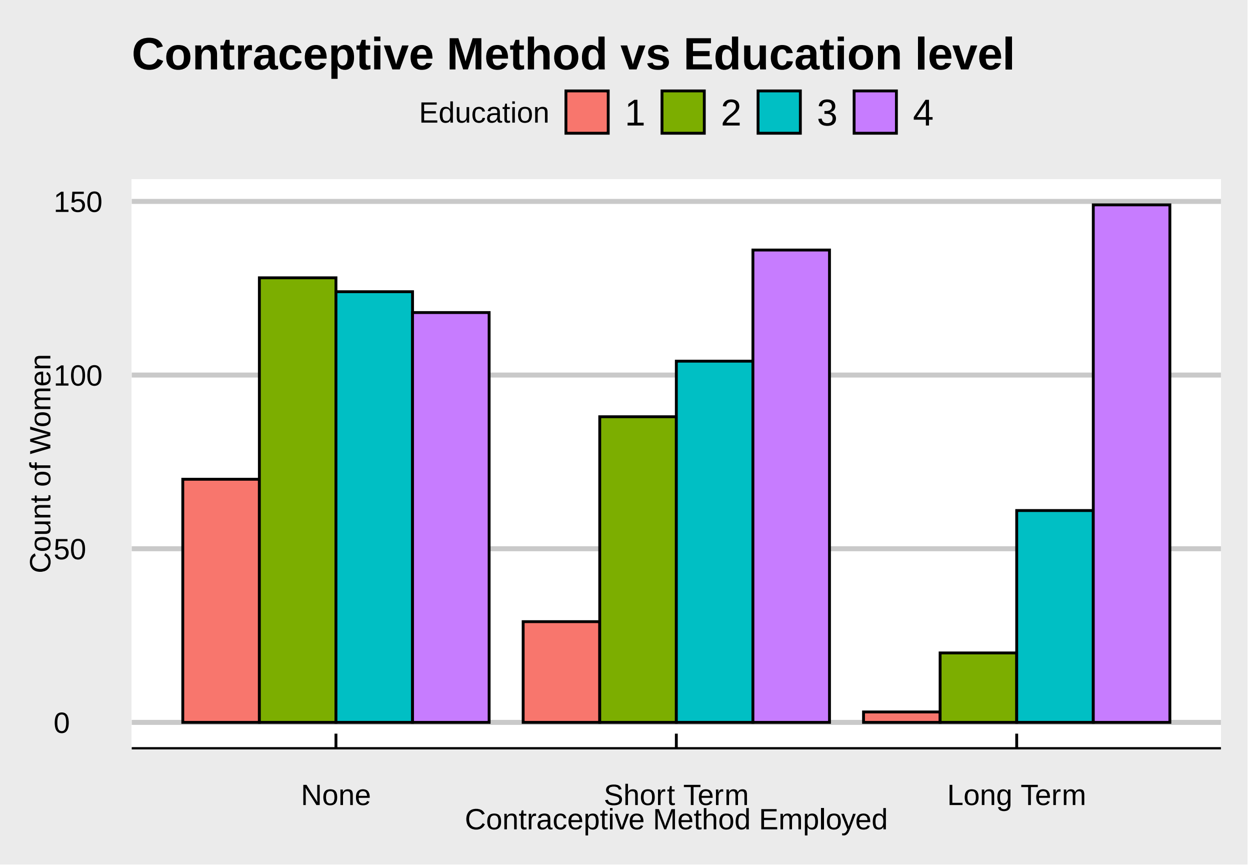


Figure 6: Contraceptive Method used vs education level

With increased education level, couples tend to use contraceptives more.

```
cpr_train %>%
    ggplot(aes(Contraceptive_Method_Used, fill = Partner_Education)) +
    geom_bar(position = "dodge",col = "black") +
    ggtitle("Contraceptive Method vs Partner's Education level") +
    xlab("Contraceptive Method Employed") +
    ylab("Count of Women") +
    theme(legend.position = "top") +
    theme_economist_white()
```

Interestingly enough the husband's education level does not necessarily mean increased use of contraceptives, but again people using long-term contraceptives tend to be the highest level. This maybe because of the high cost of long term methods.

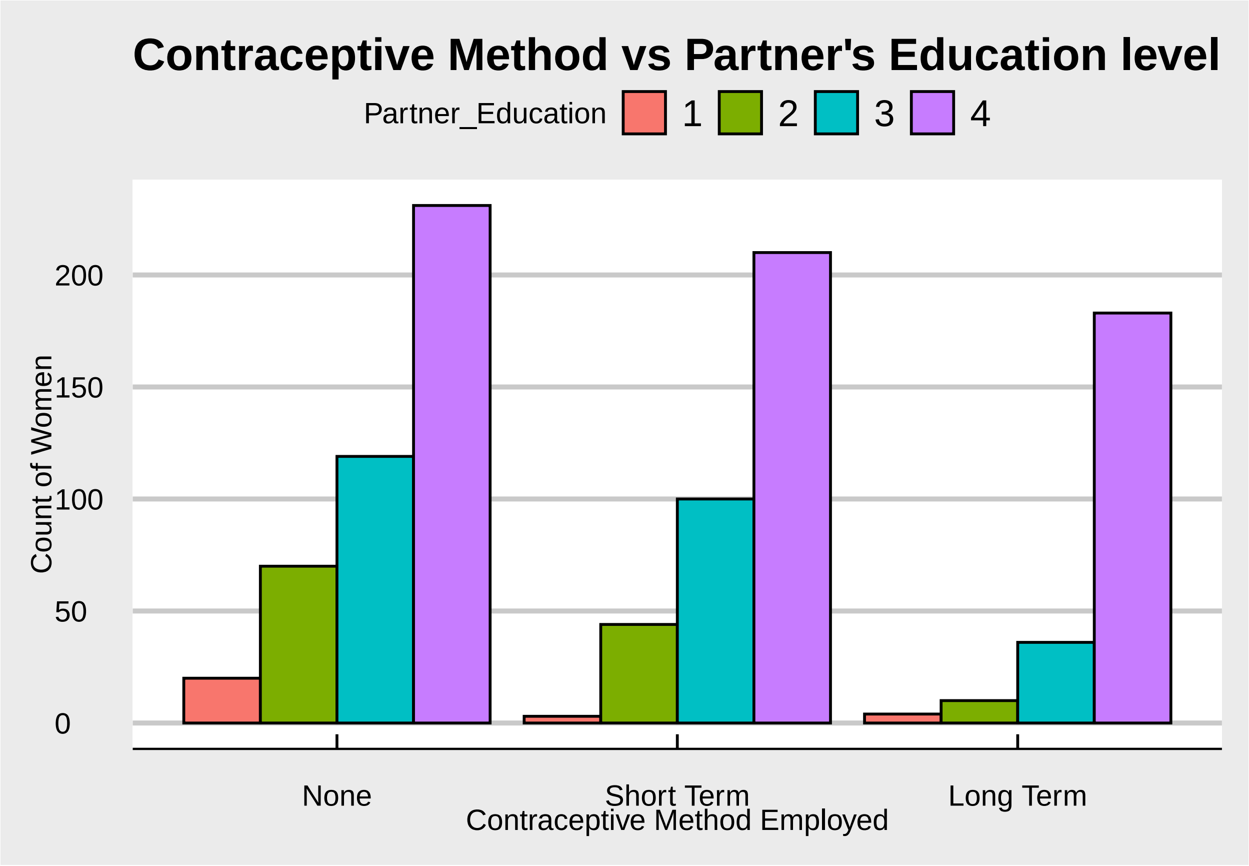


Figure 7: Contraceptive Method by Partner's Education Level

```
cpr_train %>%
    ggplot(aes(Contraceptive_Method_Used, label = ..count..,
               fill = Religion)) +
    geom_bar(position = "dodge", col = "black") +
    geom_text(stat =  "count", vjust = 1.2,
              position = position_dodge(1))+
    ggtitle("Contraceptive Method vs Religion") +
    xlab("Contraceptive Method Employed") +
    ylab("Count of Women") +
    theme(legend.position = "top") +
    theme_economist_white()
```

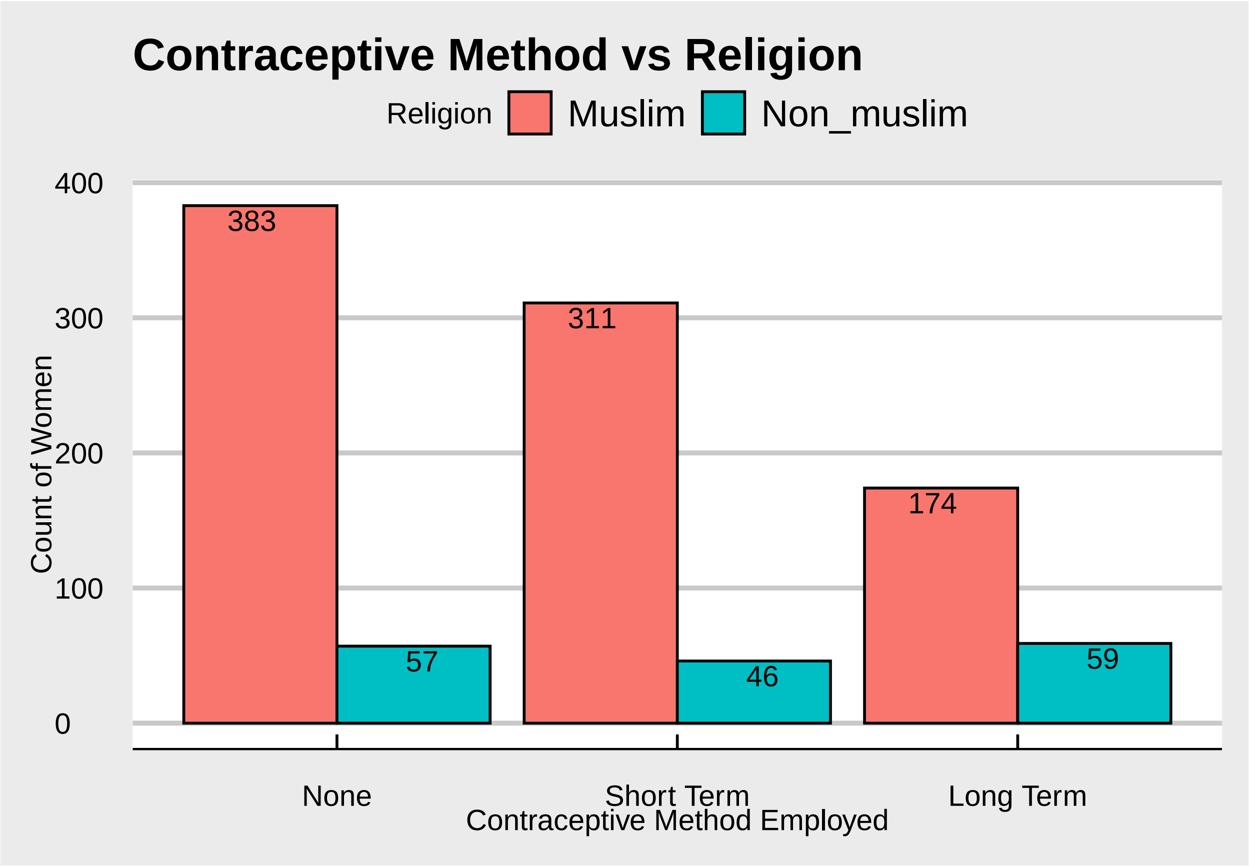


Figure 8: Contraceptive Method by Religion

We see that Muslim couples are comparatively less likely to use contraceptives

```
cpr_train %>%
    ggplot(aes(Contraceptive_Method_Used, fill =
                 Currently_working, label = ..count..)) +
    geom_bar(position = "dodge",col = "black") +
    geom_text(stat =  "count", vjust = 1.2,
              position = position_dodge(1))+
    ggtitle("Contraceptive Method by Working Status of women") +
    xlab("Contraceptive Method Employed") +
    ylab("Count of Women") +
    theme(legend.position = "top") +
    theme_economist_white()
```

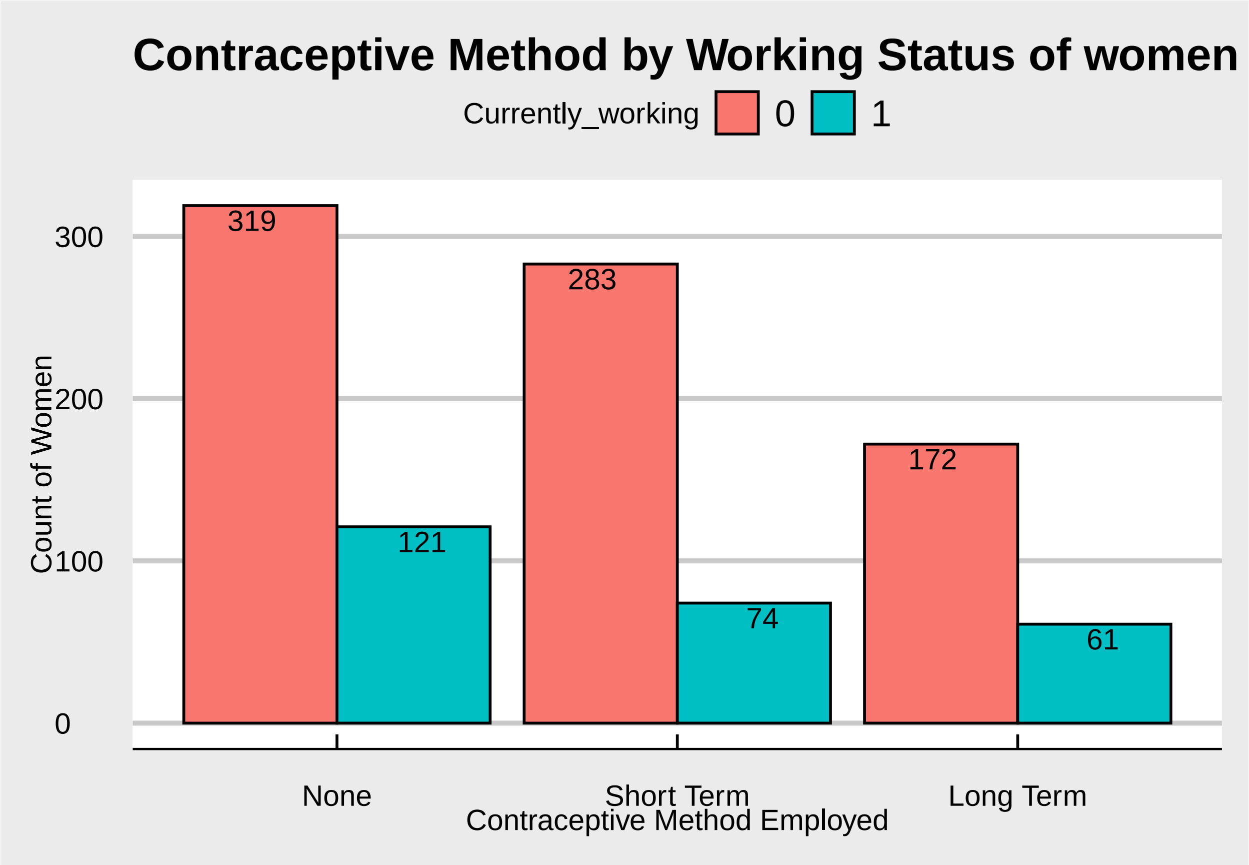


Figure 9: Contraceptive Method by Working Status

Strangely we see that working women are less likely to use contraceptives, but those women using contraceptives prefer Long term methods.

```
cpr_train %>%
  ggplot(aes(Contraceptive_Method_Used, fill = Husband_Occupation)) +
  geom_bar(col = "black", position = "dodge") +
  ggtitle("Contraceptive Method used vs Husband Occupation") +
  xlab("Contraceptive Method Employed") +
  ylab("Count of Women") +
  theme(legend.position = "top") +
  theme_economist_white()
```

Here we cannot makeout a definite pattern because the description is not clearly specified.

```
cpr_train %>%
  ggplot(aes(Contraceptive_Method_Used, fill =Standard_of_Living)) +
  geom_bar(col = "black", position = "dodge") +
```

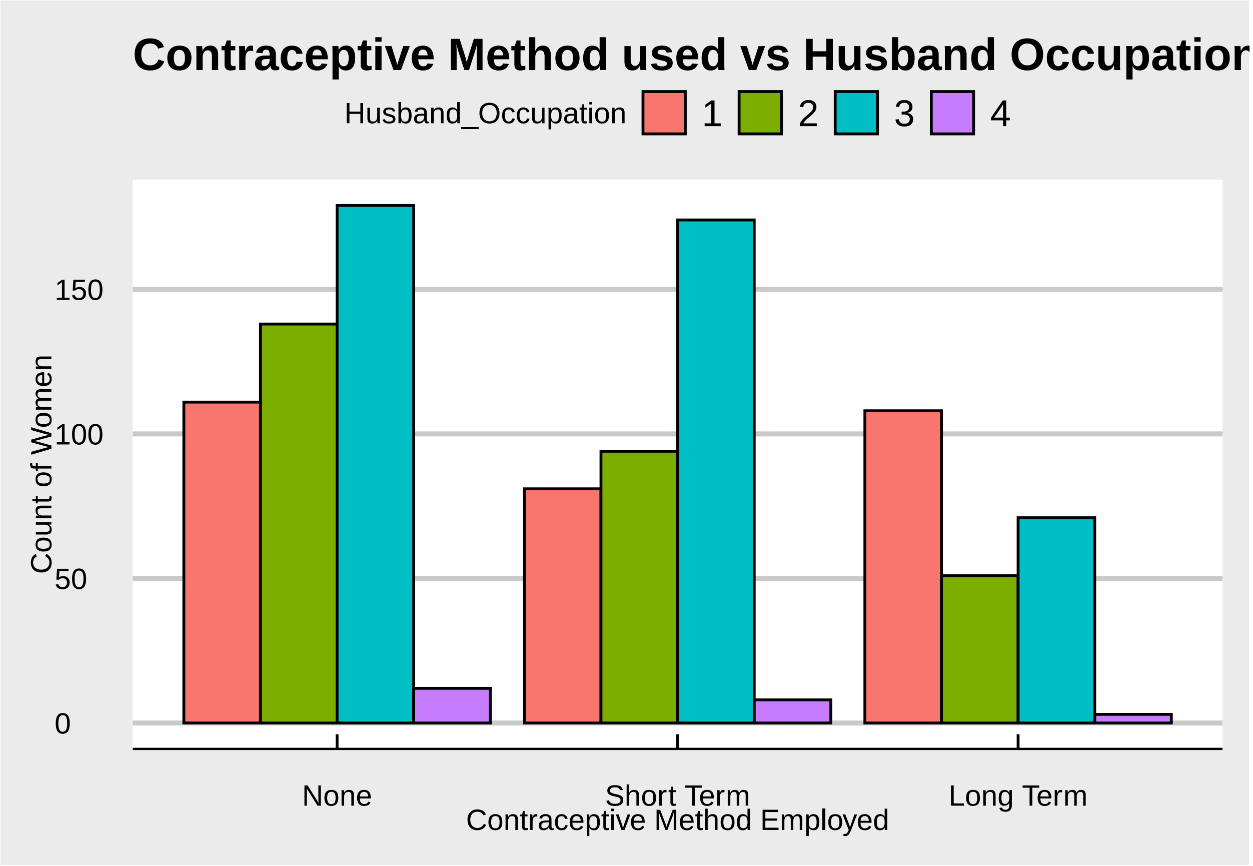


Figure 10: Contraceptive Method by Husband Occupation

```
  ggtitle("Contraceptive Method used vs Living Standard") +
  xlab("Contraceptive Method Employed") +
  ylab("Count of Women") +
  theme(legend.position = "top") +
  theme_economist_white()
```

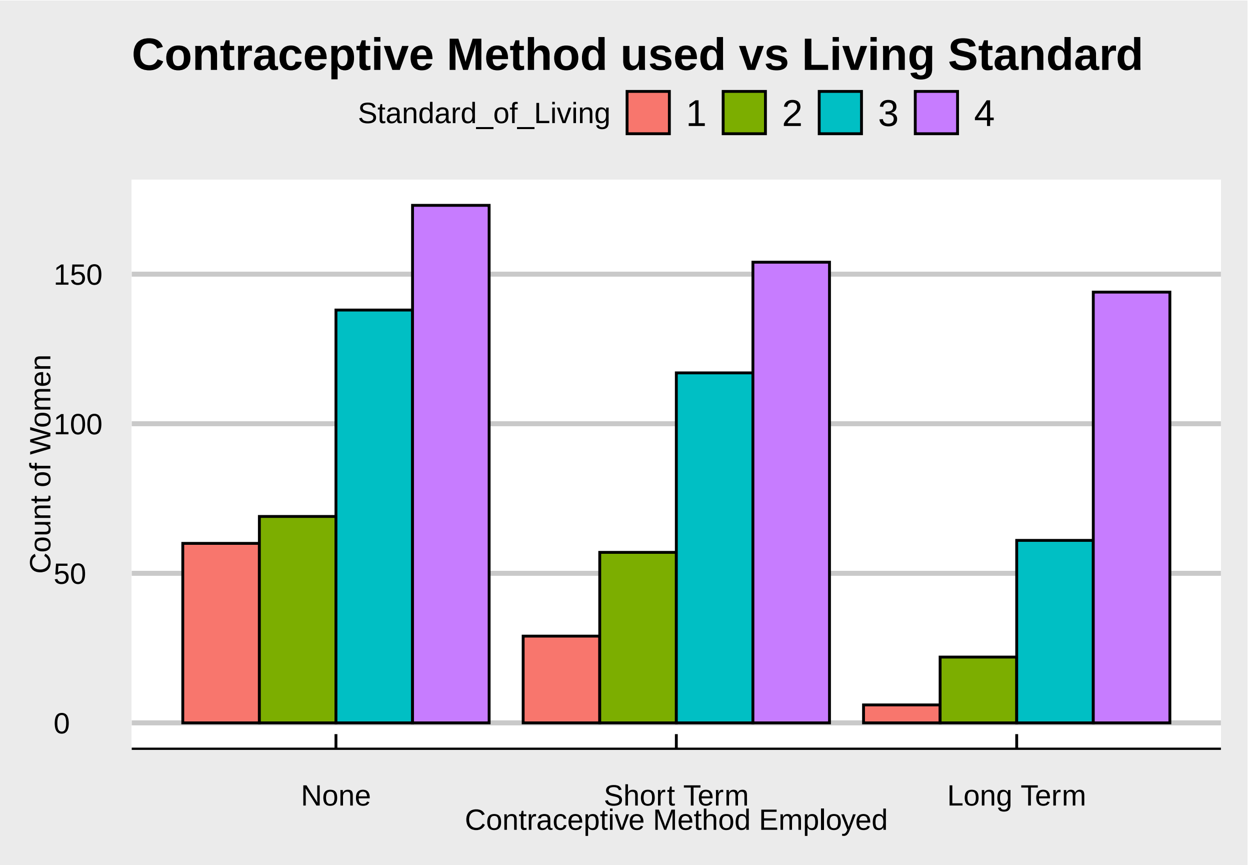


Figure 11: Contraceptive Method by Standard of Living

We see that women living under low standards are less likely to opt for long term contraceptive methods probably because of high cost.

```
cpr_train %>%
    ggplot(aes(Contraceptive_Method_Used, label = ..count..,
               fill = Media_Exposure)) +
    geom_bar(position = "dodge", col = "black") +
    geom_text(stat =  "count", vjust = 1.2,
              position = position_dodge(1))+
    ggtitle("Contraceptive Method by Media Exposure") +
    xlab("Contraceptive Method Employed") +
    ylab("Count of Women") +
    theme(legend.position = "top") +
    theme_economist_white()
```

Though the number of people without media exposure is quite small we observe a positive correlation.

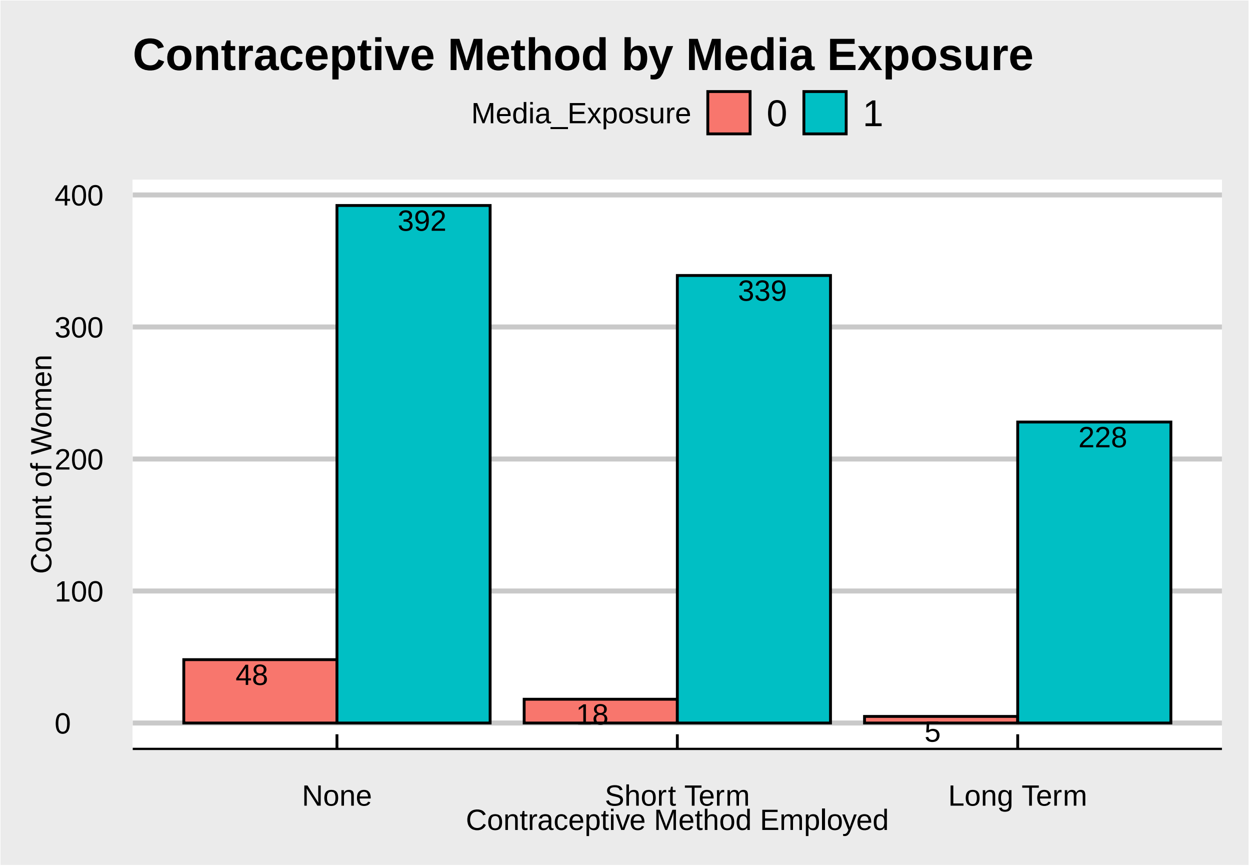


Figure 12: Contraceptive Method by Media Exposure

We've seen that all the variables can be used as predictors. Now we are ready to develop our predicting models.

3 Modelling and Results

3.1 Modelling

In this section we will use the algorithms learned in the previous courses to build predictive models and evaluate their performances in order to select the best model.

3.1.1 Control Parameters

Before we can build our models, we have to set control parameters which will be implemented when we train the models on our dataset.

For the purposes of this project, we will be tuning the "Method" parameter which is used to evaluate the accuracy of our models. In other words, how well our model generalizes to unseen data. There are several methods to do this, including splitting the dataset into train/validation/test sets, but because we don't have a large dataset to work with, we have chosen Repeated K-Fold Cross Validation which builds on k-fold cross validation.

K-fold cross validation is a robust technique that splits the data into k-subsets and for each iteration, the model holds back a subset (as a validation set) and trains on all the other data. The overall accuracy is then calculated by aggregating the accuracies produced by each iteration. Repeated CV repeats this process several times and calculates the model's accuracy by taking the mean of the repeats.

We will set the number of k-folds at 10, the number of repeats at 3, and save only the final predictions. The following code achieves these steps:

```
# Set control parameters - run algorithms using 10-fold cross validation repeated 3 times
trControl <- trainControl(method = "repeatedcv",
                          number = 10,
                          repeats=3,
                          savePredictions="final")
```

3.1.2 Ordinal Logistic Regression

We start with the simplest of our models – logistic regression. Regression is a type of predictive modelling algorithm that focuses on "the relationship between a dependent (target) variable and an independent variable(s) (predictors)", under the assumption that the independent variable(s) cause(s) the dependent variable.

Simply Logistic Regression is used only when there are two categories in the target variable. When there are more than two variables, we use:

1. Multinomial Logistic Regression when the categories have no intrinsic order.
2. Ordinal Logistic Regression when the categories are ordered.
   Both methods essentially estimates separate binary logistic regression model for each variable.[3]

[3] https://www.analyticsvidhya.com/blog/2016/02/multinomial-ordinal-logistic-regression/

From our exploration we know that the target variable has a definite order (descending from “None” to “Long Term”). Hence we will use the Ordinal Logistic Regression Method.

This model converts each categorical variable in the model to separate variables, so instead of *Eduacation* we will have *Education1*, *Education2* etc corresponding to each level of education with the values as either 1 or 0.

```
# Ordinal Logistic Regression Model
set.seed(4,sample.kind = "Rounding")
fit_olr<-train(Contraceptive_Method_Used~.,
              data=cpr_train,
              method="polr",
              trControl=trControl)

# Predict on training set
train_olr<-predict(fit_olr,cpr_train)

# Accuracy
accuracy <- mean(train_olr == cpr_train$Contraceptive_Method_Used)
accuracy

# [1] 0.504

# Variable Importance
varImp(fit_olr)

# polr variable importance
#
#                      Overall
# n_children           100.00
# Age                   85.74
# Education4            60.48
# ReligionNon_muslim    39.17
# Standard_of_Living4   36.86
# Standard_of_Living2   26.25
# Standard_of_Living3   25.43
# Husband_Occupation2   24.11
# Education3            22.93
# Media_Exposure1       18.43
# Partner_Education2    13.17
# Currently_working1    12.55
# Partner_Education3    11.62
# Partner_Education4    10.39
# Husband_Occupation3    6.13
# Education2             3.19
# Husband_Occupation4    0.00
```

The number of children, age and level 4 education status are the most important predictors.

We will create a table to store the result from each model so we can compare and select the best performing one.

```
# Create a table of results to compare models' performance
results_table <- tibble(Method = "Ordinal Logistic Regression",
                        Accuracy = accuracy)
kable(results_table) %>% kable_styling()
```

Method	Accuracy
Ordinal Logistic Regression	0.504

3.1.3 K-Nearest Neighbour

The k-nearest neighbour uses a similarity measure (Euclidean distance, cosine similarity etc) between the existing and new data to classify the new data. The k in KNN denotes the number of values to be considered.[4] With the help of class library and train workflow knn algorithm can be implemented as follows:

```
set.seed(7,sample.kind = "Rounding")
#train the model
fit_knn<-train(Contraceptive_Method_Used~.,
               data=cpr_train,
               method="knn",
               trControl=trControl,
               tuneGrid = data.frame(k = seq(3,30,2)))

#predict the result
train_knn<-predict(fit_knn,cpr_train)

# Accuracy
accuracy <- mean(train_knn == cpr_train$Contraceptive_Method_Used)
accuracy

# [1] 0.586

# Optimum value of k
best_k <- fit_knn$bestTune$k
best_k

# [1] 23

# Add Results
results_table <- bind_rows(results_table,
                           tibble(Method = "knn",
                                  Accuracy = accuracy))
```

Thus we find an improvements from our previous model.

3.1.4 Decision Tree

Decision Trees are versatile Machine Learning algorithm that can perform both classification and regression tasks. They are very powerful algorithms, capable of fitting complex datasets.[5]

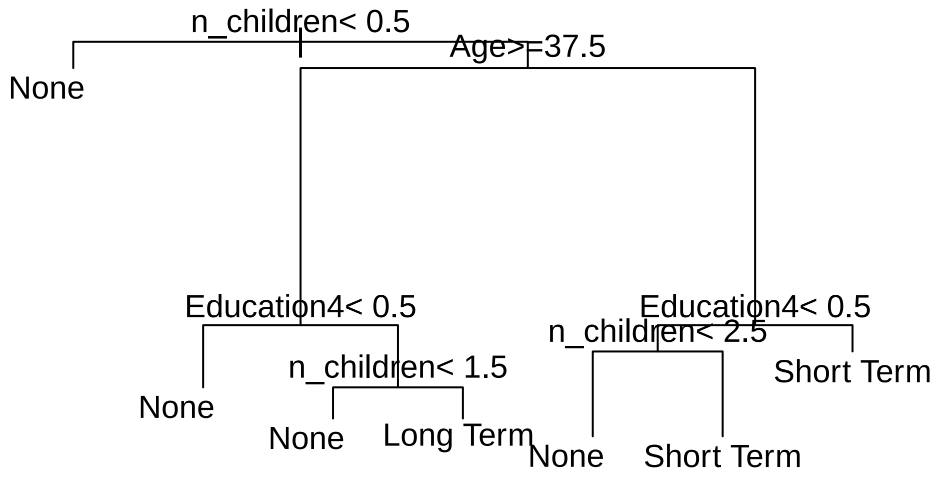
The parameter for this model is the pruning parameter cp.[6]

[4] https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[5] https://www.guru99.com/r-decision-trees.html

[6] http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/

```
set.seed(14,sample.kind = "Rounding")
#train the model
fit_rpart<-train(Contraceptive_Method_Used~.,
                 data=cpr_train,
                 method="rpart",
                 trControl=trControl,
                 tuneGrid = data.frame(cp = seq(0, 0.05, 0.002)),
)
# Predict on training set
train_rpart<-predict(fit_rpart,cpr_train)

# Accuracy
accuracy <- mean(train_rpart == cpr_train$Contraceptive_Method_Used)
accuracy

# [1] 0.577

# Optimum value of cp
best_cp <- fit_rpart$bestTune$cp
best_cp

# [1] 0.018

# plot the decision tree model
plot(fit_rpart$finalModel,margin = .1)
text(fit_rpart$finalModel)
```



```
# Check the variable importance
varImp(fit_rpart)

# rpart variable importance
#
#                      Overall
# Education4            100.00
# n_children             85.40
# Age                    61.80
# Partner_Education4     56.54
# Education2             39.20
# Standard_of_Living4    25.34
# Media_Exposure1         8.95
# Education3              5.17
# Currently_working1      3.11
# ReligionNon_muslim      2.65
# Husband_Occupation3     1.82
# Partner_Education3      1.60
# Partner_Education2      0.00
# Standard_of_Living3     0.00
# Husband_Occupation2     0.00
# Standard_of_Living2     0.00
# Husband_Occupation4     0.00

# Add Results
results_table <- bind_rows(results_table,
                           tibble(Method = "Decision Tree",
                                  Accuracy = accuracy))
```

In the plot the true conditions are represented by the left side branches. Thus this model predicts that Long term contraceptives are used by women 38 years or older, having level 4 education and 3 or more kids. The model determines Level 4 education status of the woman as the most important predictor.

Even though the accuracy of decision tree model is slightly less than the knn model, we will use it for final validation, since the model provides more information for making a decision and chance of overfitting is low.

3.1.5 Final Validation

Now we'll use the entire train set to predict the test set using the random forest model. The parameters of the model will be the parameters obtained in the last section.

```
set.seed(14,sample.kind = "Rounding")
fit_final<-train(Contraceptive_Method_Used~.,
                 data=cpr_train,
                 method = "rpart",
                 tuneGrid = data.frame(cp = best_cp)
                 )

# Predict on test set
test_rpart<-predict(fit_final,cpr_test)

# Accuracy
accuracy <- mean(test_rpart == cpr_test$Contraceptive_Method_Used)
accuracy
```

Table 1: RMSE Results

Method	Accuracy
Ordinal Logistic Regression	0.504
knn	0.586
Decision Tree	0.577
Final Validation	0.542

```
# [1] 0.542
```

```
# Add Results
results_table <- bind_rows(results_table,
                           tibble(Method = "Final Validation",
                                  Accuracy = accuracy))
```

The decision tree model performs well with the test data. Hence we can be confident that it well with new data. Now, we'll move to the results and discussion.

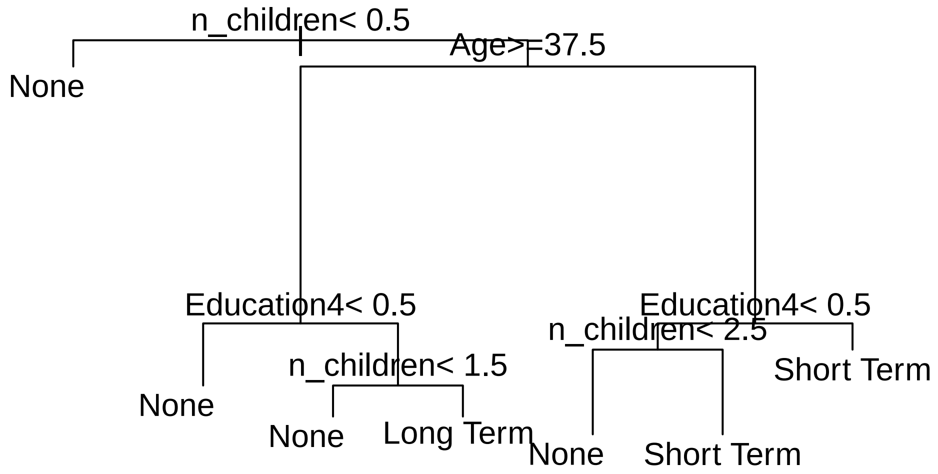
3.2 Result

```
knitr::kable(results_table, caption="RMSE Results") %>%
  kable_styling(bootstrap_options = c("striped", "hover","condensed"))
```

```
#Investigate the final model
## Check the variable importance
varImp(fit_final)
```

```
# rpart variable importance
#
#                      Overall
# Education4            100.00
# n_children             85.40
# Age                    61.80
# Partner_Education4     56.54
# Education2             39.20
# Standard_of_Living4    25.34
# Media_Exposure1         8.95
# Education3              5.17
# Currently_working1      3.11
# ReligionNon_muslim      2.65
# Husband_Occupation3     1.82
# Partner_Education3      1.60
# Husband_Occupation2     0.00
# Standard_of_Living3     0.00
# Partner_Education2      0.00
# Standard_of_Living2     0.00
# Husband_Occupation4     0.00
```

```
## plot the final model
plot(fit_final$finalModel,margin = .1)
text(fit_final$finalModel)
```



Our prediction model can be summarised as follows:

1. Contraceptives are used only after the birth of a child.
2. Short term methods are used by women till their late thirties. women not having level 4 education in this category tend to not use contraceptives.
3. Long term contraceptives are used by women from their mid-thirties. Here too women not having level 4 education tend to not use contraceptives.

4 Conclusion

The aim of this project was to analyse the contraceptive use of married women to identify the factors that promote contraceptive use and propose solution for both the government and private firms. Through this project we discovered that Level 4 education of the woman, Age and number of children are the most influential factors and that short term contraceptives are common till late-thirties and long term contraceptives are common from mid-thirties.

Lack of education is the main reason for not using both these contraceptive methods. Hence governmental agencies can make sure that women are educated upto level 4.

Private firms that sell contraceptives can actively advertise their product through media to the youth population to create more awareness. The older women who can't afford long term methods can also be encouraged to use short term contraceptives.

5 Limitations

The limitations of this project include:

1. The dataset used is from 1987 and women are now more educated.
2. Only limited information is available. Cost of contraceptive methods, average income etc are not available and could have been very useful.
3. Only few predictive models could be tried upon and the accuracy achieved was 54%.