Determing Contraception Use in Indonesian Women from 1987

Rafid Akhyara Agung

11/9/2021

Introduction

In this project for the HarvardX Data Science course, the CMC data set was investigated. This data set contains the types of contraception used by a group of Indonesian women in 1987, and the goal of this project was to create a method to predict the contraception used based on the other present variables like the wife's age, education, religion and occupation, the husband's education and occupation, as well as the number of children they have, their media exposure and standard of living.

In order to do so, the data set was split into a training and test set. The training set was exclusively used to train the model, while the test set was untouched during training and only used at the end for validation. To calculate the effect of each variable, two models were used, a basic linear model, and a regularized linear model.

Analysis: Data Exploration

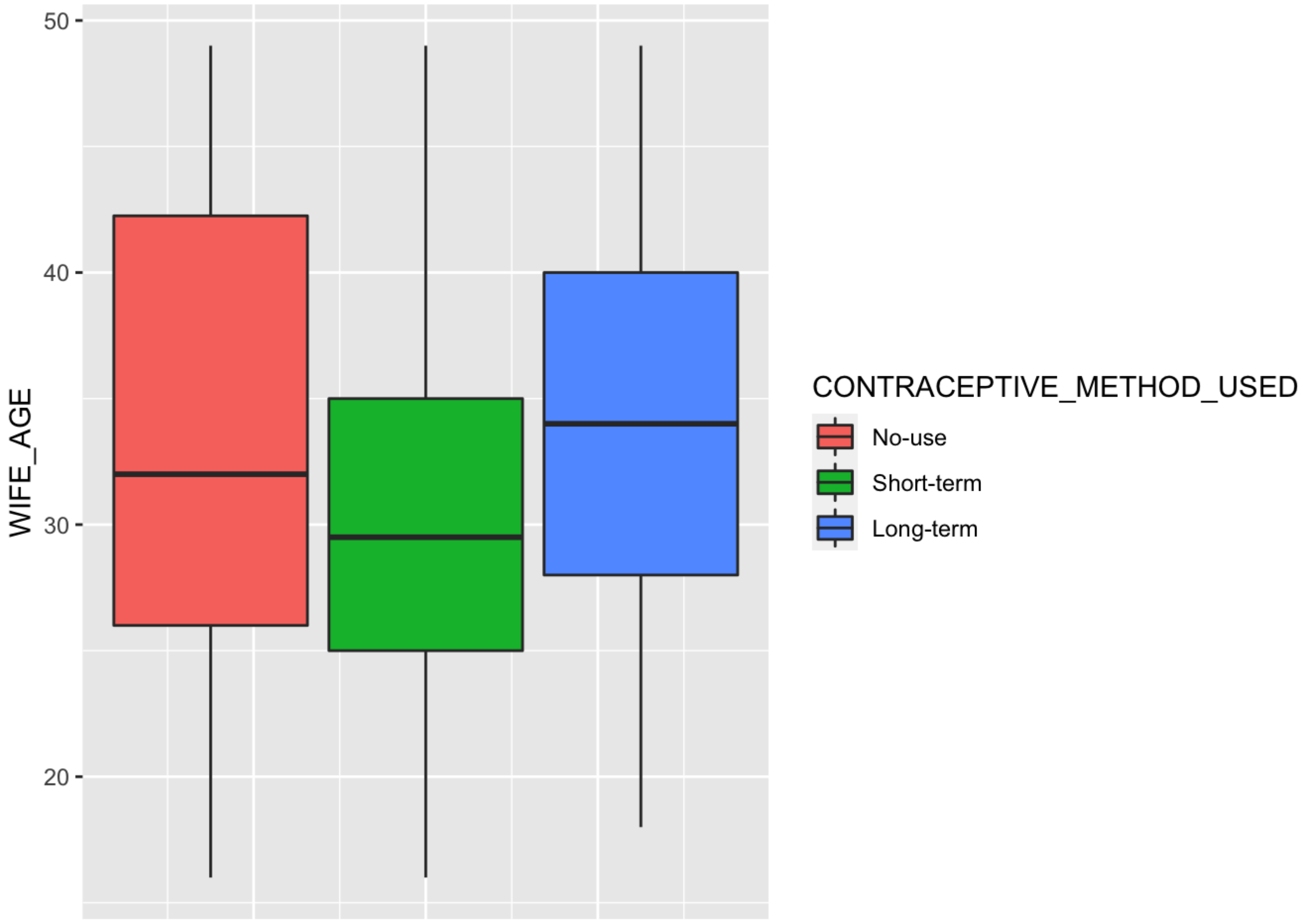
First we obtain the data and create the training and test sets. We also load the required packages.

```r
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(dplyr)) install.packages("dplyr")
if(!require(ggplot2)) install.packages("ggplot2")
if(!require(graphics)) install.packages("graphics")
if(!require(gridExtra)) install.packages("gridExtra")
if(!require(glmnet)) install.packages("glmnet")
if(!require(ggpubr)) install.packages("ggpubr")
if(!require(raster)) install.packages("raster")
if(!require(gtools)) install.packages("gtools")

library(ggplot2)
library(dplyr)
library(graphics)
library(gridExtra)
library(caret)
library(glmnet)
library(tidyverse)
library(ggpubr)
library(raster)
library(gtools)

dl <- tempfile()
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/cmc/cmc.data", dl)

#Downloading and importing the data
cmc <- read.csv(dl)

#Setting the variable names
colnames(cmc) <- c("WIFE_AGE", "WIFE_EDUCATION", "HUSB_EDUCATION",
                   "CHILDREN_NUMBER", "WIFE_RELIGION",
                   "WIFE_WORKING", "HUSBAND_OCCUPATION", "STANDARD_OF_LIVING",
                   "MEDIA_EXPOSURE", "CONTRACEPTIVE_METHOD_USED")

#Creating the training and test sets
set.seed(1, sample.kind = "Rounding")
testindex <- sample.int(n = nrow(cmc),
                        size = floor(0.3 * nrow(cmc)),
                        replace = FALSE)
testset <- cmc[testindex,]
trainset <- cmc[-testindex,]

#Setting the variables as characters for visualization purposes
for(x in c(1:10)[-c(1,4)]){
  trainset[,x] <- as.character(trainset[,x])
}

#Converting the numbers in 'CONTRACEPTIVE_METHOD_USED' into labels
trainset$CONTRACEPTIVE_METHOD_USED <- trainset$CONTRACEPTIVE_METHOD_USED %>%
  str_replace_all("1", "No-use") %>%
  str_replace_all("2", "Long-term") %>%
  str_replace_all("3", "Short-term") %>%
  factor(levels = c("No-use", "Short-term", "Long-term"))
```

In order to explore the training set and choose which variables to analyze, we will create some visualizations.

First, we visualize the effect of the wife's age.

```
#Creating boxplots for age for each type contraceptive
ageplot <- trainset %>%
    ggplot()+
    geom_boxplot(aes(y = WIFE_AGE,
                     fill = CONTRACEPTIVE_METHOD_USED))+
    theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
ageplot
```

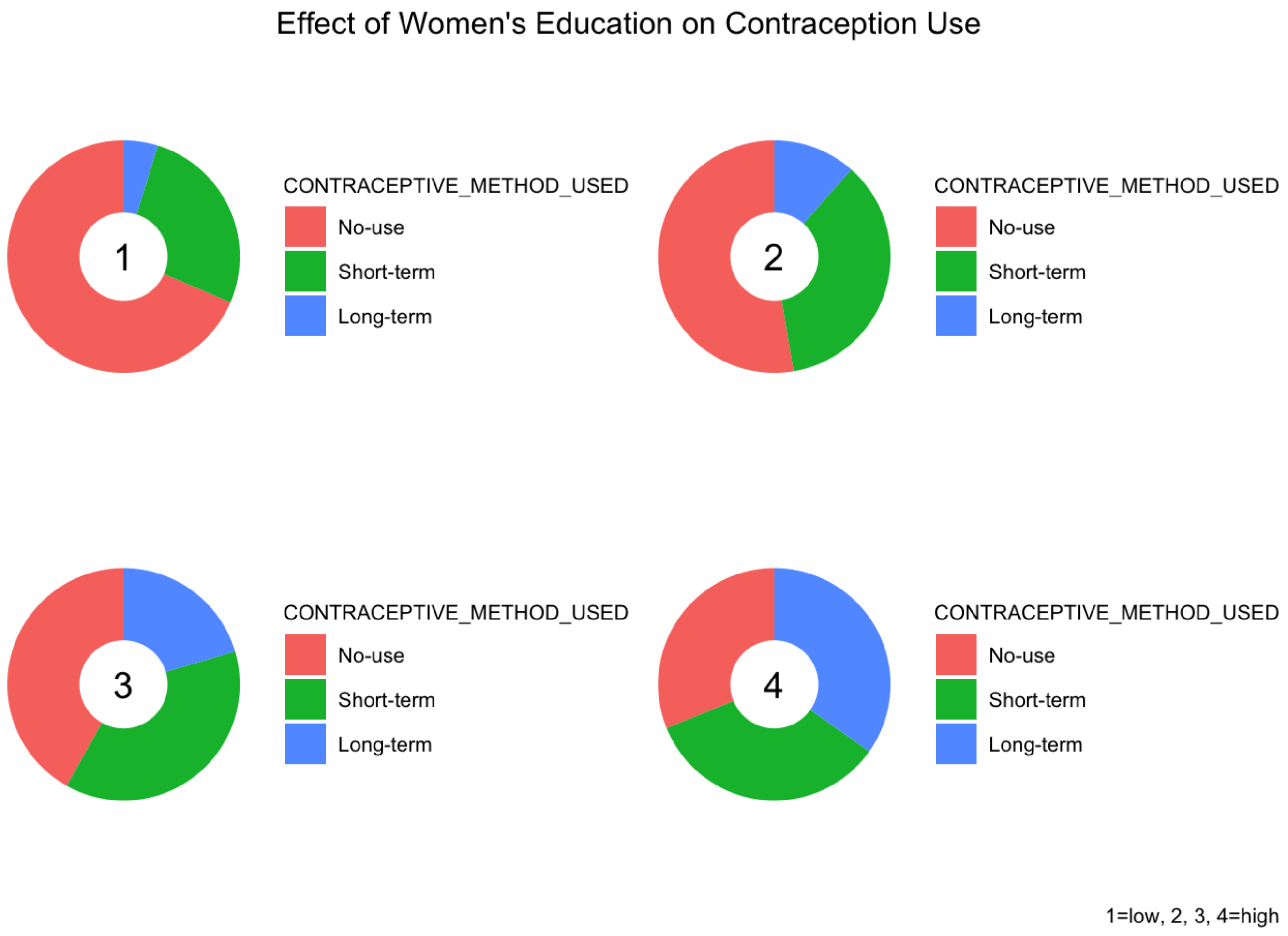


Here we see an effect from age, in that the median age differs between groups of women using different contraceptives. Younger women tend towards short-term contraception, whereas older women either don't use contraception or use long-term contraception.

Next, we visualize the effect of the wife's education.

```
#Creating donut charts for each level of education
weduplot1 <- trainset %>%
  filter(WIFE_EDUCATION == 1) %>%
  ggplot()+
  geom_bar(aes(x = WIFE_EDUCATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "1",
           size = 5)

weduplot2 <- trainset %>%
  filter(WIFE_EDUCATION == 2) %>%
  ggplot()+
  geom_bar(aes(x = WIFE_EDUCATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "2",
           size = 5)

weduplot3 <- trainset %>%
  filter(WIFE_EDUCATION == 3) %>%
  ggplot()+
  geom_bar(aes(x = WIFE_EDUCATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "3",
           size = 5)

weduplot4 <- trainset %>%
  filter(WIFE_EDUCATION == 4) %>%
  ggplot(aes(x = WIFE_EDUCATION,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "4",
           size = 5)
```

```
#Combining all charts
arrangeGrob(weduplot1, weduplot2,
            weduplot3, weduplot4) %>%
  annotate_figure(top = text_grob("Effect of Women's Education on Contraception Use"
),
                  bottom = text_grob("1=low, 2, 3, 4=high",
                                     hjust = 1, x = 1, size = 8))
```

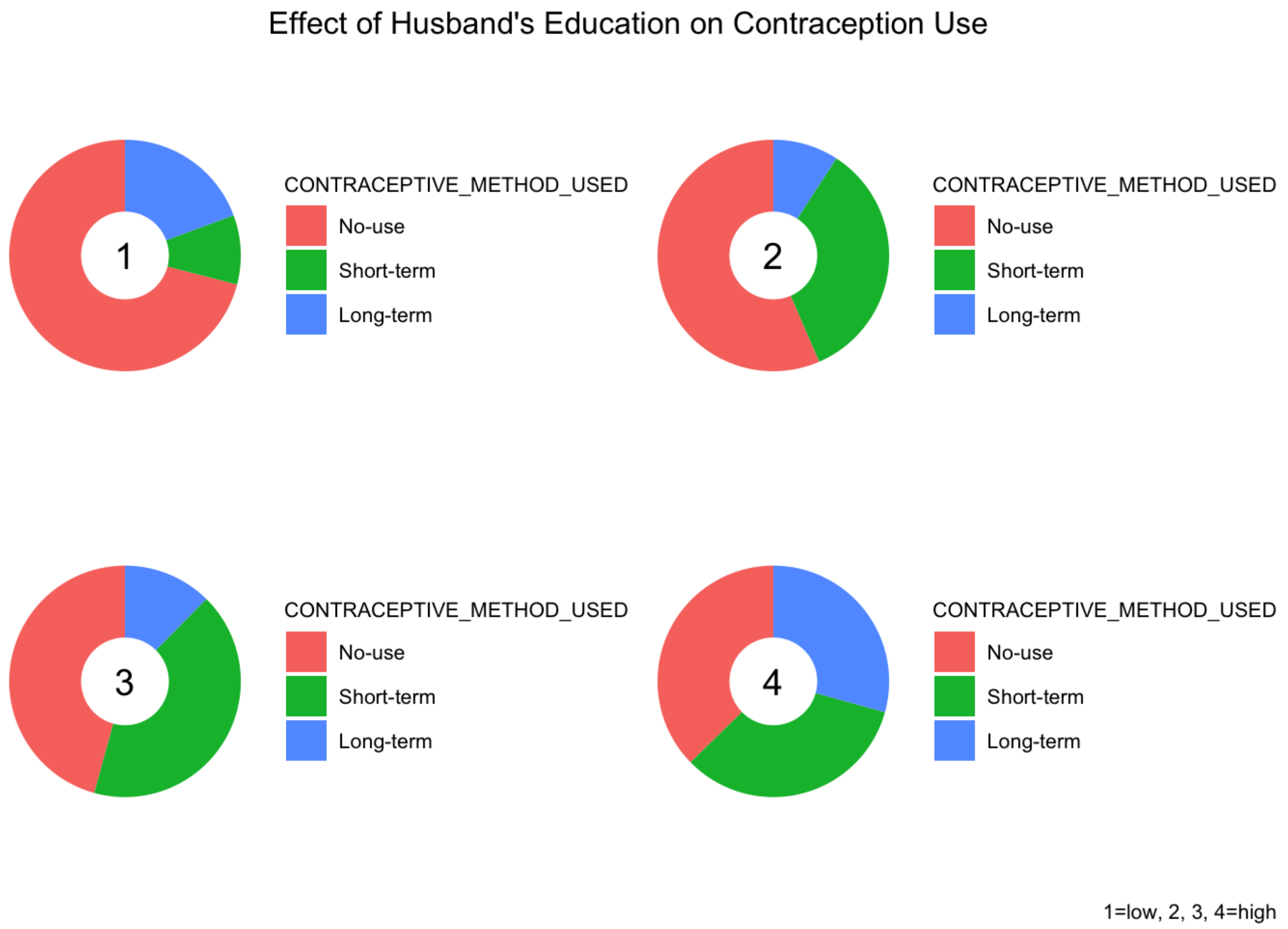


Here, we see an effect as the proportion of women who don't use contraception decrease as education increases. The inverse is true for the proportion of women who use contraception. For short-term contraception, the increase seems marginal, but for long-term contraception the increase is significant.

Next, we visualize the effect of the husband's education.

```
#Creating donut charts for each level of education
heduplot1 <- trainset %>%
  filter(HUSB_EDUCATION == 1) %>%
  ggplot()+
  geom_bar(aes(x = HUSB_EDUCATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "1",
           size = 5)

heduplot2 <- trainset %>%
  filter(HUSB_EDUCATION == 2) %>%
  ggplot()+
  geom_bar(aes(x = HUSB_EDUCATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "2",
           size = 5)

heduplot3 <- trainset %>%
  filter(HUSB_EDUCATION == 3) %>%
  ggplot()+
  geom_bar(aes(x = HUSB_EDUCATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "3",
           size = 5)

heduplot4 <- trainset %>%
  filter(HUSB_EDUCATION == 4) %>%
  ggplot(aes(x = HUSB_EDUCATION,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "4",
           size = 5)
```

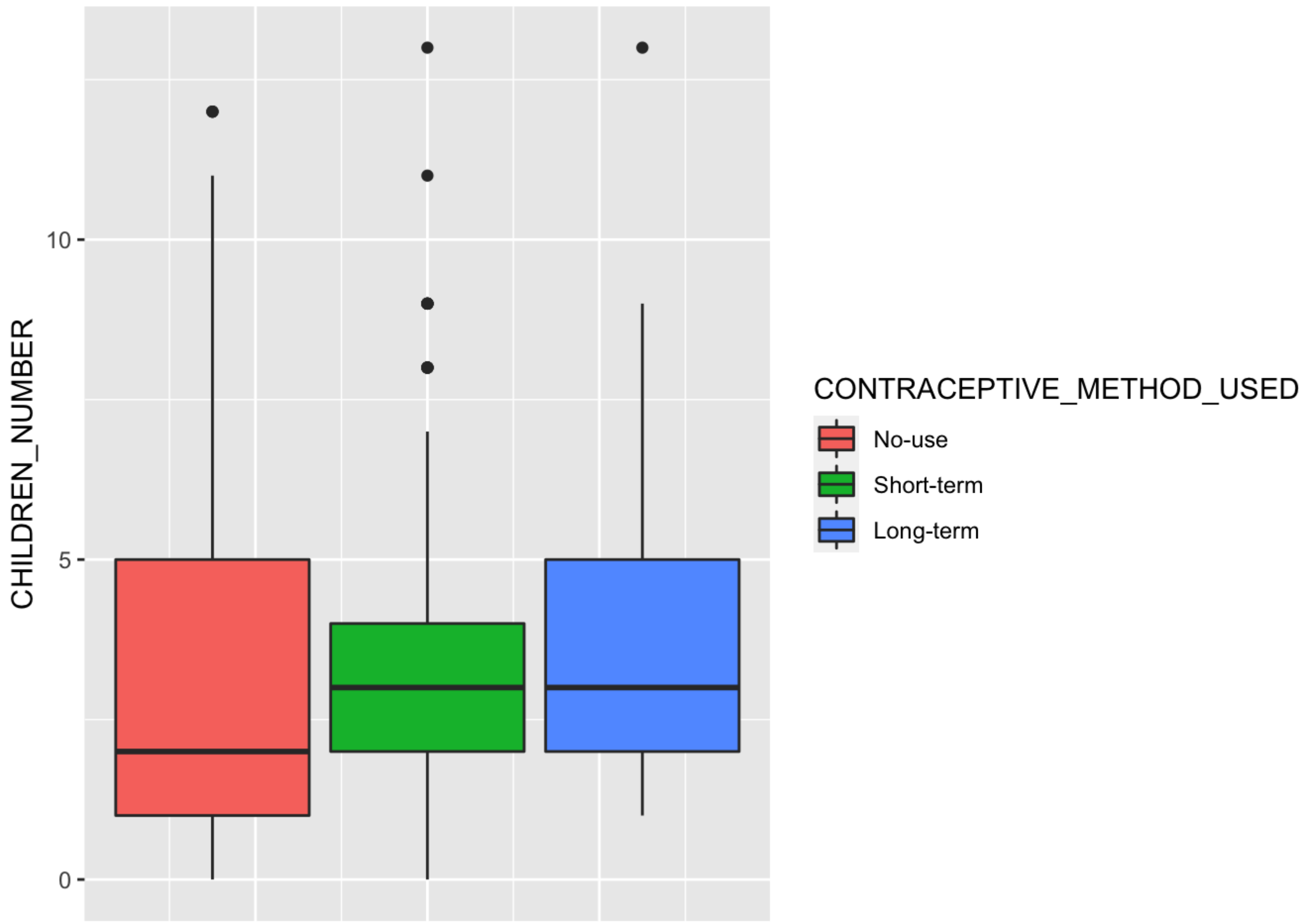
```
#Combining all charts
arrangeGrob(heduplot1, heduplot2,
            heduplot3, heduplot4)%>%
  annotate_figure(top = text_grob("Effect of Husband's Education on Contraception Us
e"),
                  bottom = text_grob("1=low, 2, 3, 4=high",
                                     hjust = 1, x = 1, size = 8))
```



Here, we see a similar trend with the previous variable. The increase in use of short-term contraception is however more drastic and for both types the proportions don’t seem to always strictly increase with higher husband education but there is an increase overall.

Next, we visualize the effect of the number of children a couple has.

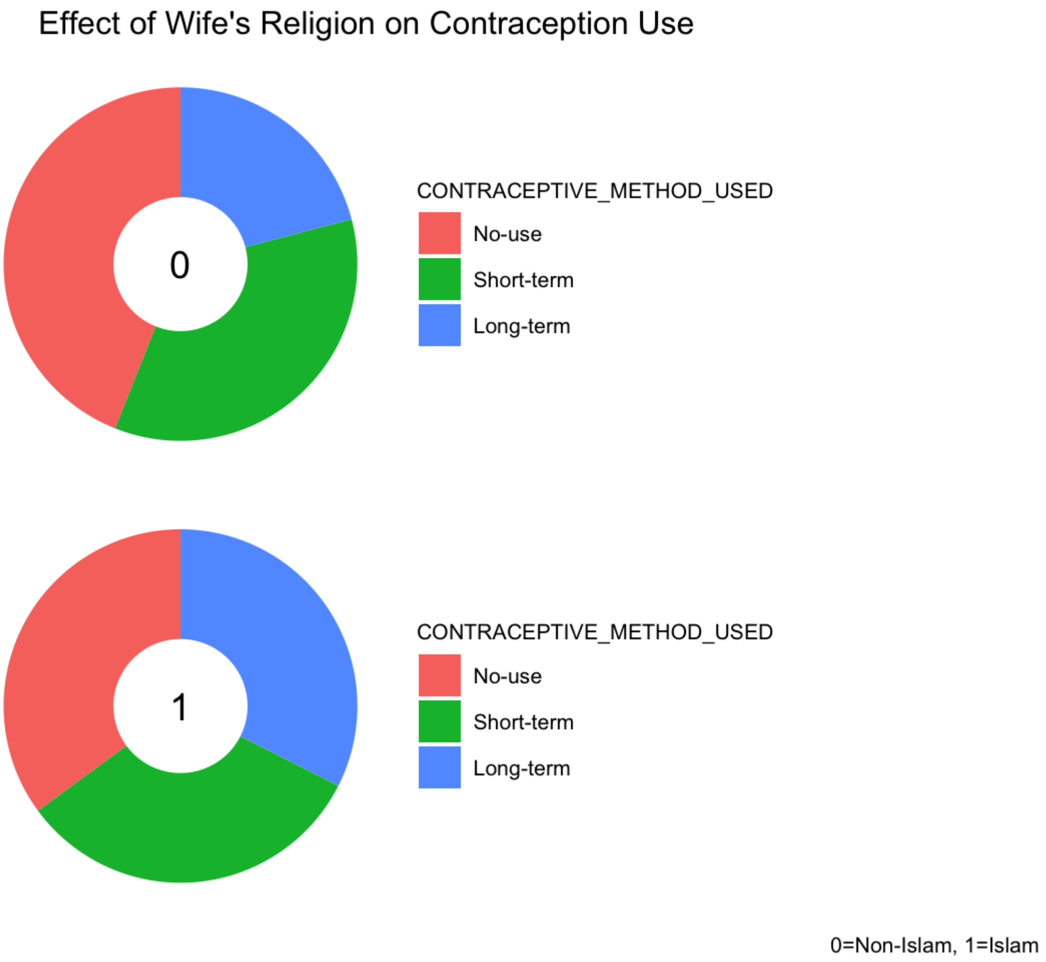
```
#Creating boxplots for number of children for each type of contraceptive
childplot <- trainset %>%
  ggplot()+
    geom_boxplot(aes(y = CHILDREN_NUMBER,
                     fill = CONTRACEPTIVE_METHOD_USED))+
    theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
childplot
```



Here, we see that women that don't use contraception have less median children. However, the variance for that group is also much larger.

Next, we visualize the effect of the wife's religion.

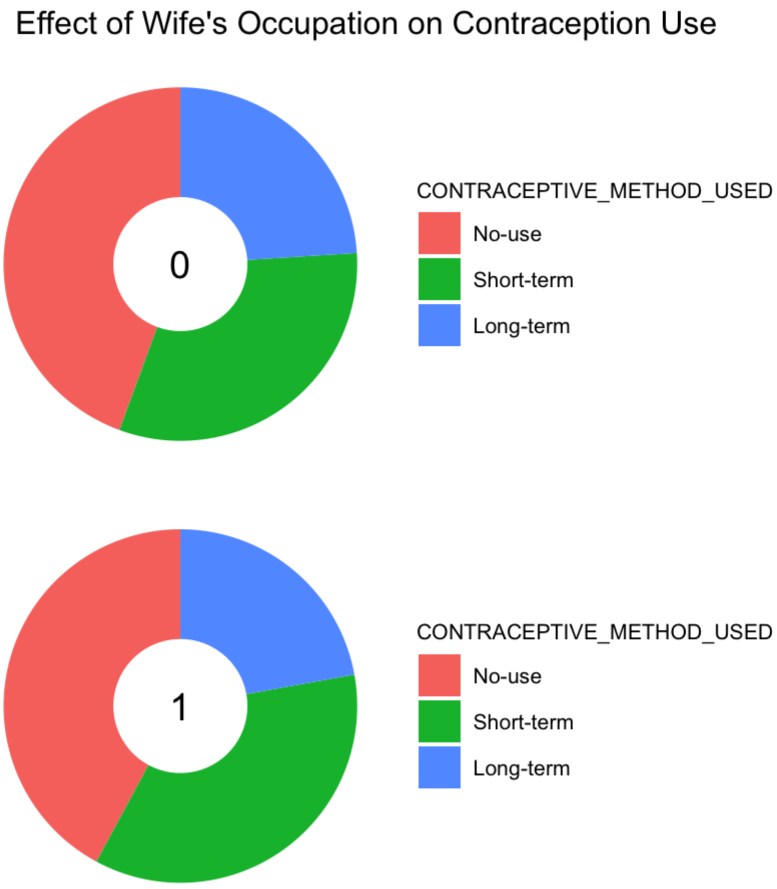
```
#Creating donut charts for each religion
religplot1 <- trainset %>%
  filter(WIFE_RELIGION == 1) %>%
  ggplot(aes(x = WIFE_RELIGION,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "0",
           size = 5)

religplot2 <- trainset %>%
  filter(WIFE_RELIGION == 0) %>%
  ggplot(aes(x = WIFE_RELIGION,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "1",
           size = 5)

#Combining all charts
arrangeGrob(religplot1, religplot2)%>%
  annotate_figure(top = text_grob("Effect of Wife's Religion on Contraception Use"),
                  bottom = text_grob("0=Non-Islam, 1=Islam",
                                     hjust = 1, x = 1, size = 8))
```



Here, we see the proportions of the types of contraception differing between Muslim and non-Muslim women. Non-Muslims tend use contraception slightly less and prefer short-term contraception more. Muslim women tend to use contraception slightly more and have less of a preference towards either long- or short-term contraception.

Next, we visualize the effect of the wife’s working status.

```
#Creating donut charts for working status
wifeworkplot1 <- trainset %>%
  filter(WIFE_WORKING == 0) %>%
  ggplot(aes(x = WIFE_WORKING,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "0",
           size = 5)

wifeworkplot2 <- trainset %>%
  filter(WIFE_WORKING == 1) %>%
  ggplot(aes(x = WIFE_WORKING,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "1",
           size = 5)

#Combining all charts
arrangeGrob(wifeworkplot1, wifeworkplot2)%>%
  annotate_figure(top = text_grob("Effect of Wife's Occupation on Contraception Use"
),
                  bottom = text_grob("0=Working, 1=Not working",
                                     hjust = 1, x = 1, size = 8))
```



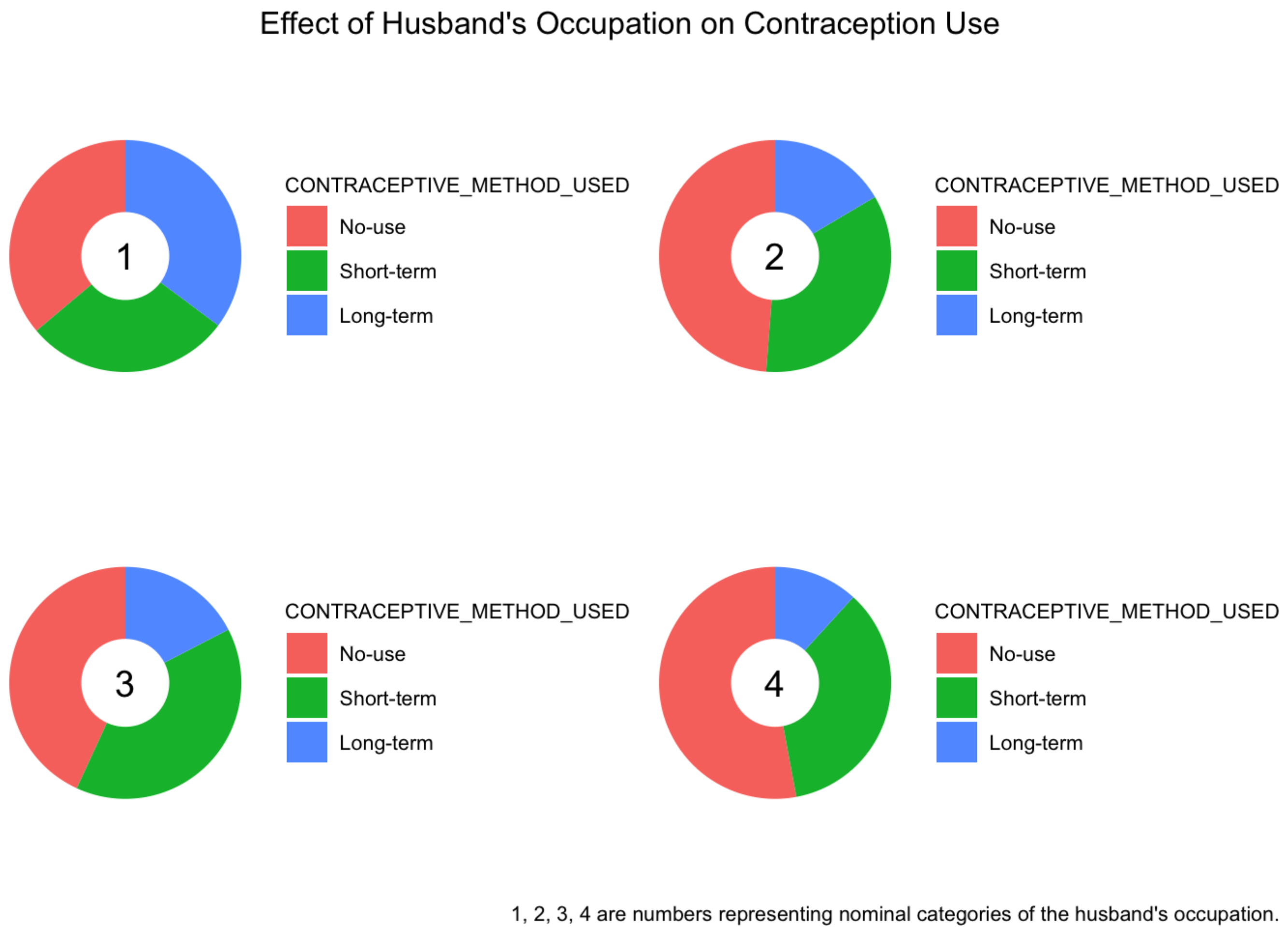
0=Working, 1=Not working

Here, the effect seems marginal between working and non-working women.

Next, we visualize the effect of the husband’s occupation.

```
#Creating donut charts for each category of job
husbworkplot1 <- trainset %>%
  filter(HUSBAND_OCCUPATION == 1) %>%
  ggplot()+
  geom_bar(aes(x = HUSBAND_OCCUPATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "1",
           size = 5)

husbworkplot2 <- trainset %>%
  filter(HUSBAND_OCCUPATION == 2) %>%
  ggplot()+
  geom_bar(aes(x = HUSBAND_OCCUPATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "2",
           size = 5)

husbworkplot3 <- trainset %>%
  filter(HUSBAND_OCCUPATION == 3) %>%
  ggplot()+
  geom_bar(aes(x = HUSBAND_OCCUPATION,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "3",
           size = 5)

husbworkplot4 <- trainset %>%
  filter(HUSBAND_OCCUPATION == 4) %>%
  ggplot(aes(x = HUSBAND_OCCUPATION,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "4",
           size = 5)
```

```
#Combining all charts
arrangeGrob(husbworkplot1, husbworkplot2,
             husbworkplot3, husbworkplot4)%>%
  annotate_figure(top = text_grob("Effect of Husband's Occupation on Contraception Us
e"),
                  bottom = text_grob("1, 2, 3, 4 are numbers representing nominal cat
egories of the husband's occupation.",
                                     hjust = 1, x = 1, size = 8))
```


Effect of Husband's Occupation on Contraception Use

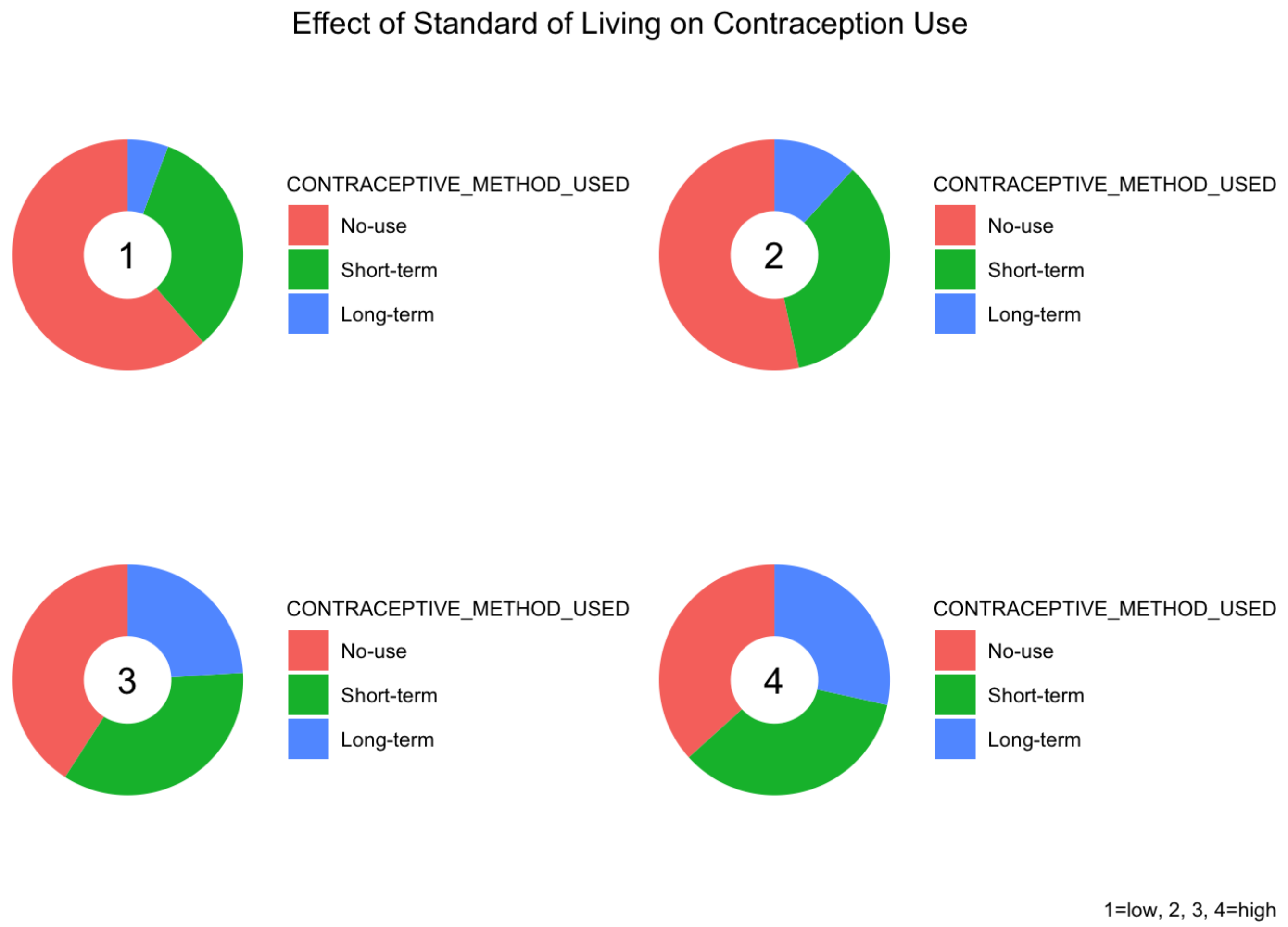
1, 2, 3, 4 are numbers representing nominal categories of the husband's occupation.

Here, we see that job type 1 is the odd one out as the husbands in other categories’ wives tend to use less contraception.

Next, we visualize the effect of the standard of living.

```r
#Creating donut charts for each standard of living index
solplot1 <- trainset %>%
  filter(STANDARD_OF_LIVING == 1) %>%
  ggplot()+
  geom_bar(aes(x = STANDARD_OF_LIVING,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "1",
           size = 5)

solplot2 <- trainset %>%
  filter(STANDARD_OF_LIVING == 2) %>%
  ggplot()+
  geom_bar(aes(x = STANDARD_OF_LIVING,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "2",
           size = 5)

solplot3 <- trainset %>%
  filter(STANDARD_OF_LIVING == 3) %>%
  ggplot()+
  geom_bar(aes(x = STANDARD_OF_LIVING,
               fill = CONTRACEPTIVE_METHOD_USED))+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "3",
           size = 5)

solplot4 <- trainset %>%
  filter(STANDARD_OF_LIVING == 4) %>%
  ggplot(aes(x = STANDARD_OF_LIVING,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "4",
           size = 5)
```

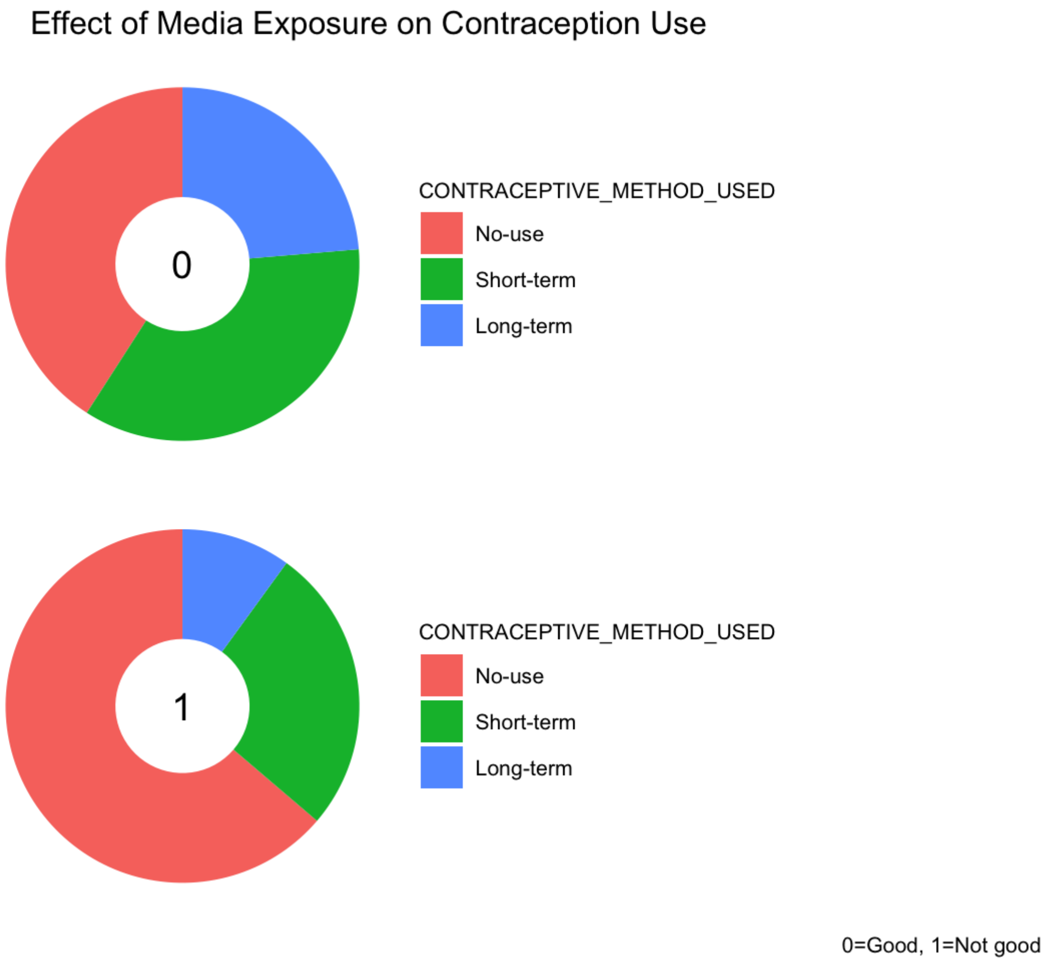
```
#Combining all charts
arrangeGrob(solplot1, solplot2,
             solplot3, solplot4)%>%
  annotate_figure(top = text_grob("Effect of Standard of Living on Contraception Use"
),
                  bottom = text_grob("1=low, 2, 3, 4=high",
                                     hjust = 1, x = 1, size = 8))
```



Here, we see as standard of living increases, then contraception use increases and most notably, the use of long-term contraceptives.

Next, we visualize the effect of the media exposure.

```r
#Creating donut charts for each level of media exposure
mediaplot1 <- trainset %>%
  filter(MEDIA_EXPOSURE == 0) %>%
  ggplot(aes(x = MEDIA_EXPOSURE,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "0",
           size = 5)

mediaplot2 <- trainset %>%
  filter(MEDIA_EXPOSURE == 1) %>%
  ggplot(aes(x = MEDIA_EXPOSURE,
             fill = CONTRACEPTIVE_METHOD_USED))+
  geom_bar()+
  coord_polar(theta = "y")+
  theme_void()+
  theme(legend.title = element_text(size=8),
        legend.text = element_text(size=8))+
  annotate("text",
           x = 0, y = 0,
           label = "1",
           size = 5)

#Combining all charts
arrangeGrob(mediaplot1, mediaplot2)%>%
  annotate_figure(top = text_grob("Effect of Media Exposure on Contraception Use"),
                  bottom = text_grob("0=Good, 1=Not good",
                                     hjust = 1, x = 1, size = 8))
```



Here, we see that as media exposure increases, contraception use, especially, again, long-term contraceptive use, increases.

Analysis: Forming the Model

For model evaluation, we will be using the Confusion Matrix as this is categorical nominal data.

From our visualization, we are going to include all variables in the model, despite some of them seeming insignificant, but the data is not that large so we can afford to run a model with them and can cull them later on. We will be using a linear model, then we'll be modifying to improve results.

After each modification, we will look at the confusion matrix to evaluate the effectiveness of the algorithm.

First though, we will check the results we can achieve by random guessing.

```r
#Setting to numeric for analysis purposes
trainset$CONTRACEPTIVE_METHOD_USED <- trainset$CONTRACEPTIVE_METHOD_USED %>%
  str_replace_all("No-use", "1") %>%
  str_replace_all("Long-term", "2") %>%
  str_replace_all("Short-term", "3") %>%
  as.factor()

for(n in c(1:10)){
  trainset[,n] <- as.numeric(trainset[,n])
}

#Simulating guessing with corresponding probabilities
set.seed(69, sample.kind = "Rounding")
```

```
## Warning in set.seed(69, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
GUESS <- as.factor(sample(c(1,2,3),
       length(testset$CONTRACEPTIVE_METHOD_USED),
       replace = TRUE,
       prob = c(sum(trainset$CONTRACEPTIVE_METHOD_USED == 1)/length(trainset$CONTRACE
PTIVE_METHOD_USED),
                sum(trainset$CONTRACEPTIVE_METHOD_USED == 2)/length(trainset$CONTRACE
PTIVE_METHOD_USED),
                sum(trainset$CONTRACEPTIVE_METHOD_USED == 3)/length(trainset$CONTRACE
PTIVE_METHOD_USED))))

data.frame(Model = "Guessing",
                     Accuracy = confusionMatrix(GUESS, as.factor(testset$CONTRACEPTIV
E_METHOD_USED))$overall["Accuracy"]) %>%
  knitr::kable()
```

	Model	Accuracy
Accuracy	Guessing	0.3446712

Now we will see when we use all the seemingly significant variables from visualization.

```
#Creating a basic linear model.

linear <- lm(CONTRACEPTIVE_METHOD_USED ~ .,
             data = trainset)

predictlm <- as.factor(clamp(round(predict(linear, testset)),1,3))

accuracy <- data.frame(Model = "Wife's Age + Wife's Education + Husband's Education +
Children Number + Wife's Religion + Husband's Occupation + Standard of Living + Media
Exposure + Wife Working",
                     Accuracy = confusionMatrix(predictlm, as.factor(testset$CONTRACE
PTIVE_METHOD_USED))$overall["Accuracy"])

accuracy %>% knitr::kable()
```

	Model	Accuracy
Accuracy	Wife’s Age + Wife’s Education + Husband’s Education + Children Number + Wife’s Religion + Husband’s Occupation + Standard of Living + Media Exposure + Wife Working	0.3219955

The results actually show a drop, so we need to modify the model to improve performance. First, we’ll perform statistical tests on the variables to see their significance.

```
summary(linear)[4]
```

```
## $coefficients
##                       Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)         2.12590390 0.254327916  8.3589090 2.052631e-16
## WIFE_AGE           -0.03873201 0.004045387 -9.5743639 7.459376e-21
## WIFE_EDUCATION      0.12496871 0.035177584  3.5525097 3.990088e-04
## HUSB_EDUCATION      0.00748893 0.041110726  0.1821649 8.554895e-01
## CHILDREN_NUMBER     0.10948027 0.014078208  7.7765769 1.814857e-14
## WIFE_RELIGION      -0.10907693 0.076759392 -1.4210239 1.556152e-01
## WIFE_WORKING        0.02362581 0.060950418  0.3876234 6.983755e-01
## HUSBAND_OCCUPATION  0.04096376 0.033643620  1.2175789 2.236654e-01
## STANDARD_OF_LIVING  0.09490598 0.030774731  3.0838929 2.098015e-03
## MEDIA_EXPOSURE     -0.08631494 0.105059500 -0.8215815 4.115069e-01
```

Here we can see from the t and p values provided, there are 5 significant variables, Wife's Age, Education, Religion, The Number of Children and Standard of Living. So for the next model we will only be using those variables.

```
#Linear model using only significant variables
lineartested <- lm(CONTRACEPTIVE_METHOD_USED ~ WIFE_AGE + WIFE_EDUCATION + CHILDREN_N
UMBER + WIFE_RELIGION + STANDARD_OF_LIVING,
                   data = trainset)

predictlm2 <- as.factor(clamp(round(predict(lineartested, testset)),1,3))

accuracy <- rbind(accuracy,
data.frame(Model = "Tested Linear Model",
                     Accuracy = confusionMatrix(predictlm2, as.factor(testset$CONTRAC
EPTIVE_METHOD_USED))$overall["Accuracy"])
)

accuracy %>% knitr::kable()
```

	Model	Accuracy
Accuracy	Wife's Age + Wife's Education + Husband's Education + Children Number + Wife's Religion + Husband's Occupation + Standard of Living + Media Exposure + Wife Working	0.3219955
Accuracy1	Tested Linear Model	0.3219955

The results still have not improved by much so we need to further modify. In our previous linear models, we used the round function to coerce the numbers to integers. So we will try to improve the rounding.

We do this by creating a list of permutations of numbers from 0 to 1. We'll be using those as boundaries for rounding. (i.e. If number is 2-x, we round down. 2+y, we round up. We will be trying to find the x and y.) We find those numbers by using 5-fold cross validation.

When optimizing the cut-off point, we use specificity as well since we can get a similar accuracy by predicting majority of "1" as that is the most common category.

```
predictlm3 <- predict(lineartested, testset)

#Creating the permutations
deci <- permutations(length(seq(0,1,0.05)),2,seq(0,1,0.05))

#Finding optimal cut-off point for rounding.
b <- 5

set.seed(69, sample.kind = "Rounding")
bestdeci <- replicate(b,{
  #Creating sub- training and test sets
  testindex <- sample.int(n = nrow(trainset),
                        size = floor(0.3 * nrow(trainset)),
                        replace = FALSE)
  subtestset <- trainset[testindex,]
  subtrainset <- trainset[-testindex,]

  for(n in c(1:10)){
    subtrainset[,n] <- as.numeric(subtrainset[,n])
  }

  lineartested <- lm(CONTRACEPTIVE_METHOD_USED ~ WIFE_AGE + WIFE_EDUCATION + CHILDREN_NUMBER + WIFE_RELIGION + STANDARD_OF_LIVING,
                     data = subtrainset)

  predictlm3 <- predict(lineartested, subtestset)
  a <- data.frame()

  #Testing the cutoff points
  decitest <- sapply(c(1:nrow(deci)),function(x){
    for(n in c(1:length(predictlm3))){
      predictlm3[n] <- ifelse(predictlm3[n]<2-deci[x,1], floor(predictlm3[n]), ifelse(predictlm3[n]>2+deci[x,2], ceiling(predictlm3[n]), 2))
    }
    predictlm3 <- as.factor(clamp(predictlm3,1,3))

    a <- rbind(a, data.frame(
      Class_1 = as.numeric(confusionMatrix(predictlm3, as.factor(subtestset$CONTRACEPTIVE_METHOD_USED))$byClass[1,1]),
      Class_2 = as.numeric(confusionMatrix(predictlm3, as.factor(subtestset$CONTRACEPTIVE_METHOD_USED))$byClass[2,1]),
      Class_3 = as.numeric(confusionMatrix(predictlm3, as.factor(subtestset$CONTRACEPTIVE_METHOD_USED))$byClass[3,1]),
      x1 = as.numeric(deci[x,1]),
      x2 = as.numeric(deci[x,2]),
      acc = confusionMatrix(predictlm3, as.factor(subtestset$CONTRACEPTIVE_METHOD_USED))$overall["Accuracy"]))
  })
  decitest
})

bestdeci <- as.data.frame(t(matrix(bestdeci, 6, 420*b)))

colnames(bestdeci) <- c("C1", "C2", "C3", "X1", "X2", "Acc")

#Determining the best cutoff points
bestdeci <- bestdeci %>%
  mutate(across(.fns = as.numeric)) %>%
  dplyr::group_by(X1,X2) %>%
  summarize(C1bar = mean(C1),
            C2bar = mean(C2),
            C3bar = mean(C3),
            Acc = mean(Acc)) %>%
  filter(C1bar >= 0.4 & C2bar >= 0.4 & C3bar >= 0.4)

predictlm3 <- predict(lineartested, testset)

#Applying cutoff points in rounding procedure
for(n in c(1:length(predictlm3))){
  predictlm3[n] <- ifelse(predictlm3[n]<2-bestdeci$X1[which.max(bestdeci$Acc)], floor
(predictlm3[n]), ifelse(predictlm3[n]>2+bestdeci$X2[which.max(bestdeci$Acc)], ceiling
(predictlm3[n]), 2))
}

predictlm3 <- as.factor(clamp(predictlm3,1,3))

accuracy <- rbind(accuracy,
data.frame(Model = "Tested Linear Model, Custom Rounding",
                     Accuracy = confusionMatrix(predictlm3, as.factor(testset$CONTRAC
EPTIVE_METHOD_USED))$overall["Accuracy"])
)

accuracy %>% knitr::kable()
```

	Model	Accuracy
Accuracy	Wife’s Age + Wife’s Education + Husband’s Education + Children Number + Wife’s Religion + Husband’s Occupation + Standard of Living + Media Exposure + Wife Working	0.3219955
Accuracy1	Tested Linear Model	0.3219955
Accuracy2	Tested Linear Model, Custom Rounding	0.4535147

As we can see accuracy increases drastically.

To try and further improve results, we will try regularization. We use the same approach as the Tested Linear Model with Custom Rounding, but we will regularize values.

Regularization penalizes effects of the variables when the number of observations are smaller. We will try it to see if there is an improvement in our model.

```
#Converting to matrix for regularization
matrix_trainset <- as.matrix(trainset)
x <- matrix_trainset[,-c(3,6,7,9,10)]
y <- matrix_trainset[,10]

set.seed(1, sample.kind = "Rounding")
reg <- function(a,b){
  cv.glmnet(a,b, alpha = 0,
            nfolds = 10)
}

regularize <- reg(x,y)

z <- as.matrix(testset)[,-c(3,6,7,9,10)]

predictreg <- predict(regularize, z,
                      s = "lambda.min")

#Same process as above in picking cutoff point
set.seed(69, sample.kind = "Rounding")
bestdecireg <- replicate(b,{
  testindex <- sample.int(n = nrow(trainset),
                          size = floor(0.3 * nrow(trainset)),
                          replace = FALSE)
  subtestset <- trainset[testindex,]
  subtrainset <- trainset[-testindex,]

  for(n in c(1:10)){
    subtrainset[,n] <- as.numeric(subtrainset[,n])
  }

  matrix_subtrainset <- as.matrix(subtrainset)
  x <- matrix_subtrainset[,-c(3,6,7,9,10)]
  y <- matrix_subtrainset[,10]

  regularize <- reg(x,y)

  z <- as.matrix(subtestset)[,-c(3,6,7,9,10)]

  predictreg <- predict(regularize, z,
                        s = "lambda.min")

  a <- data.frame()

  decitest <- sapply(c(1:nrow(deci)),function(x){
    for(n in c(1:length(predictreg))){
      predictreg[n] <- ifelse(predictreg[n]<2-deci[x,1], floor(predictreg[n]), ifelse
(predictreg[n]>2+deci[x,2], ceiling(predictreg[n]), 2))
    }
    predictreg <- as.factor(clamp(as.numeric(predictreg), 1, 3))

    a <- rbind(a, data.frame(
      Class_1 = as.numeric(confusionMatrix(predictreg, as.factor(subtestset$CONTRACEP
TIVE_METHOD_USED))$byClass[1,1]),
      Class_2 = as.numeric(confusionMatrix(predictreg, as.factor(subtestset$CONTRACEP
TIVE_METHOD_USED))$byClass[2,1]),
      Class_3 = as.numeric(confusionMatrix(predictreg, as.factor(subtestset$CONTRACEP
TIVE_METHOD_USED))$byClass[3,1]),
      x1 = as.numeric(deci[x,1]),
      x2 = as.numeric(deci[x,2]),
      acc = confusionMatrix(predictreg, as.factor(subtestset$CONTRACEPTIVE_METHOD_USE
D))$overall["Accuracy"]))
  })
  decitest
})

bestdecireg <- as.data.frame(t(matrix(bestdecireg, 6, 420*b)))

colnames(bestdecireg) <- c("C1", "C2", "C3", "X1", "X2", "Acc")

bestdecireg <- bestdecireg %>%
  mutate(across(.fns = as.numeric)) %>%
  dplyr::group_by(X1,X2) %>%
  summarize(C1bar = mean(C1),
            C2bar = mean(C2),
            C3bar = mean(C3),
            Acc = mean(Acc)) %>%
  filter(C1bar >= 0.4 & C2bar >= 0.4 & C3bar >= 0.4)
```

```
## `summarise()` has grouped output by 'X1'. You can override using the `.groups` arg
ument.
```

```
predictreg <- predict(lineartested, testset)

for(n in c(1:length(predictreg))){
  predictreg[n] <- ifelse(predictreg[n]<2-bestdecireg$X1[which.max(bestdecireg$Acc)],
floor(predictreg[n]), ifelse(predictreg[n]>2+bestdecireg$X2[which.max(bestdecireg$Ac
c)], ceiling(predictreg[n]), 2))
}

predictreg <- as.factor(clamp(as.numeric(predictreg), 1, 3))
```

Results

```
accuracy <- rbind(accuracy,
data.frame(Model = "Regularized Tested Linear Model, Custom Rounding",
                     Accuracy = confusionMatrix(predictreg, as.factor(testset$CONTRAC
EPTIVE_METHOD_USED))$overall["Accuracy"])
)

accuracy %>% knitr::kable()
```

	Model	Accuracy
Accuracy	Wife’s Age + Wife’s Education + Husband’s Education + Children Number + Wife’s Religion + Husband’s Occupation + Standard of Living + Media Exposure + Wife Working	0.3219955
Accuracy1	Tested Linear Model	0.3219955
Accuracy2	Tested Linear Model, Custom Rounding	0.4535147
Accuracy3	Regularized Tested Linear Model, Custom Rounding	0.4535147

Unfortunately, the results did not improve by much through our use of regularization. However, nonetheless our models show a clear improvement towards higher and higher accuracies.

Conclusion

In this project, a model was designed for the CMC data set. Through consideration of the available variables and regularization of these factors, the model was able to better than guessing in predicting type of contraception used by women. But it can still be improved.

In order to improve the performance, more factors could be included such as the specific living places (urban/rural) or the data of their parents. And these factors should be statistically significant.

Other ways to improve may be to use another modeling approach that would better predict the outcomes.