STA141A Group 17 Final Project

By: Brady Horton, Rachel Chan, Kotoha Togami, and Jasper Dong

2023-12-12

Figure 1: Picture of Numerous Types of Contraceptives

I. Introduction

For this project we are using the Contraceptive Method Choice data set from the UC Irvine Machine Learning repository. The data set we are utilizing in our statistical analysis is a subset of greater data acquired in Indonesia in 1978. The samples and observations of this set consists of only married women who are currently not pregnant or did not have knowledge that they were pregnant at the time of data collection. In this data set, our response variable is contraceptive use which is a binary variable of 0 or 1. 0 represents no contraceptive use, while 1 represents contraceptive use. We have 9 predictor variables that will be used to analyze the relationship between the response variable of contraceptive use. These 9 predictors are: wife's age, wife's education, husband's education, number of children, Muslim, if the wife works, husband's occupation, standard of living, and type of media exposure. The majority of our predictor variables are binary, categorical variables, which includes: wife's education (low or high), husband's education (low or high), Muslim (0 for no, 1 for yes), whether wife works (0 for no, 1 for yes), husband's occupation (low-income or high income), standard of living (low or high), and media exposure (bad or good). Using this data, we hope to determine which factors contribute to a married woman's choice in using contraceptives or not, and how accurately we can predict their choices based on the predictive variables in this data set.

II. How We Cleaned the Data

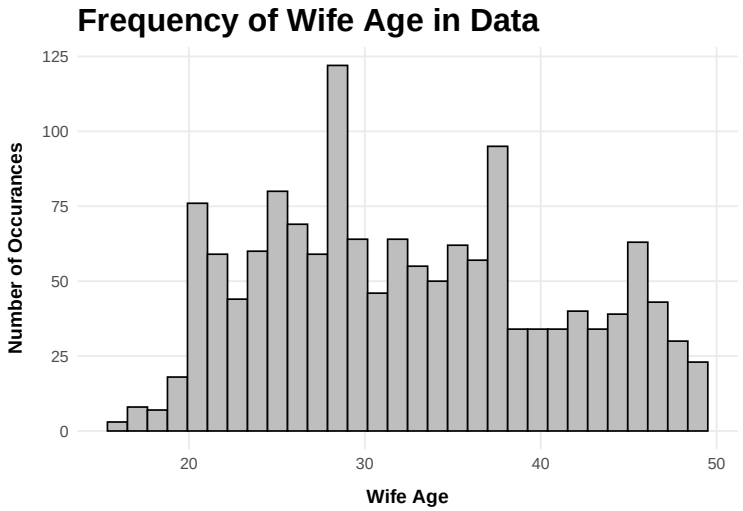
The raw data from the UC Irvine Machine Learning Repository was relatively messy and would have been hard to use for analysis. Therefore, we had to do extensive data cleaning prior to using the data for classification purposes. Firstly, many of the categorical variables were initially numbered 1-4, with 1 being a "low" and 4 being a "high" measurement for the given data category. In order to make our analysis more practical and the subsequent models easier to interpret, we decided to group all of the categorical variables into two categories, with readings of 1 or 2 being "low" for a category and 3 or 4 being "high" for that same category. This would make our models easier to interpret, since one of those readings on the outcome would be absorbed into the intercept, while the other could be easily read as a coefficient to a given predictor. We also made a modification to the outcome variable of contraceptive use. In the original data, the outcome variable of contraceptive use was a tertiary variable of 1-3, with 1 being no contraceptive use, 2 being long term contraceptive use, and 3 being short term contraceptive use. Instead of having this tertiary response variable, we instead opted to morph the categories of 2 and 3 into a single category of "uses contraceptives" with a response of 1, and the category of 1 into "doesn't use contraceptives" with a response of 0. This change will make classification both easier, and allow us to use both logistic regression and LDA to classify respondents' birth control usage.

III. Descriptive statistics

Below we will show the proportions and the visualizations of our data. This is important to know and keep in mind when determining which factors are significant in predicting whether contraceptives were used in our statistical analysis later.

a) Predictor Variable: Wife's Age

The histogram visualizes the frequency so we can clearly see the age range where the majority of the observations lie: Age 26 is where most of our observations are from, with 80 observations from this age group. The data of age seems to show a right skewed distribution of data with the peak being in the age range 23-30. We also found the average age out of all our observations which is age 32.542. The standard deviation of age is 8.227.



b) Predictor Variable: Wife's Education

The frequency table shows the number of observations for the level of wife's education in our data set. This data set contains 485 married women with low education, and 987 married women with high education. It can be seen that there are significantly more married women with a high education level in this data set compared to women with a low education level.

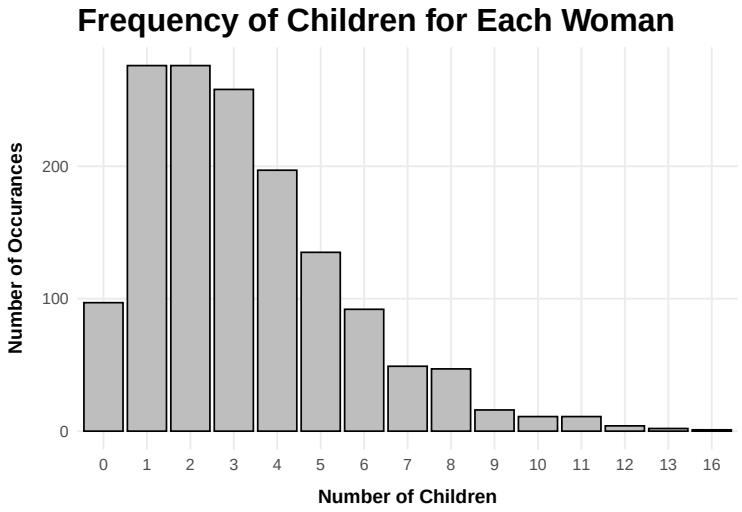
Education Level	Count
low	485
high	987

c) Predictor Variable: Number of Children

The frequency table shows the number of occurrences of the amount of children each married woman has. It can be seen from the table that most of the married women have 1 child, 2 children, and 3 children. 1 child and 2 children have the highest frequency in our data set with 276 occurrences each, while the frequency of married women having 3 children had 258 occurrences in our data set. The histogram visualizes our data and is able to show us the distribution of our data. As it can be seen, our data of the number of children each married woman has, has a right skew distribution. The average amount of children a woman has is 3.2615, or 3 children. The standard deviation of the number of children a married woman has is 2.3593.

Number of Children	Count
0	97
1	276
2	276
3	258
4	197
5	135
6	92
7	49
8	47
9	16
10	11

Number of Children	Count
11	11
12	4
13	2
16	1



d) Predictor Variable: Religious Choice (Muslim)

The table shows the frequency of the religious choice of the married women in our data set. 220 married women in our data set are not Muslim. 1215 married women in our data set are Muslim. As seen, there is a significantly larger number of Muslim women compared to women that are not Muslim. The proportion of the married women in this data that are Muslim is 0.8505.

Muslim (0 = No, 1 = Yes)	Count
0	220
1	1252

e) Predictor Variable: Husband's Education

The frequency table shows the number of observations for the level of husband's education in our data set. As seen, we have 1250 husbands with a high education level and 222 husbands with a low education level. We can see that the number of husbands with a high education level is much higher than the number of husbands with a low education level.

Education Level	Count
low	222
high	1250

f) Predictor Variable: If The Women Works

The table shows the frequency of women in this data set that do not work and the amount of women in the data set that do work. 369 married women do not work and 1103 do work. As it can be seen from the frequency table, the amount of women who do work is significantly higher than women who do not work. The proportion of married women who work in the data is 0.7493.

Working(0 = No, 1 = Yes)	count
0	369
1	1103

g) Predictor Variable: Husband Occupation

After looking at the descriptive attributes, and looking more closely at the description of the variable, we found that it was defined very vaguely by the original source of the data. We concluded that the information we would be able to obtain from this predictor variable will not be useful. Hence, we decided not to include this variable in any of our models.

h) Predictor Variable: Standard of Living

The table shows the frequency of each type of standard of living. There are 358 women with a low standard of living and 1114 with a high standard of living. As seen, the frequency of women with a high standard of living is significantly greater than women with low standard of living.

Standard of Living	Count
low	358
high	1114

i) Predictor Variable: Type of Media Exposure

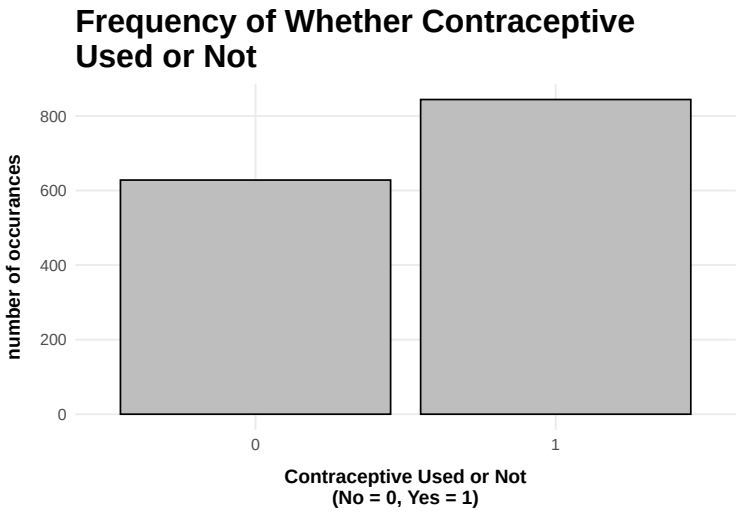
The table shows the numerical frequency of each type of media exposure the women in the sample have. 109 have bad media exposure, and 1363 have good media exposure. As it can be seen there is a much higher frequency of women with good media exposure in this sample compared to women with bad media exposure.

Media Exposure	Count
bad	109
good	1363

j) Response Variable: Contraceptive Use

The table shows the numerical frequency of whether women used contraceptives or not. 628 married women do not use contraceptives, and 844 married women do use contraceptives. Through the bar plot, it can be visualized that a greater amount of married women in this data set use contraceptives than married women who do not use contraceptives. The proportion of married women who use contraceptives in the data is 0.5734.

Contraceptive Use(0 = No, 1 = Yes)	Count
0	628
1	844



We keep this descriptive information in mind when applying supervised statistical analysis to our data to discover which predictors are significant in predicting whether women choose to use contraceptives or not. We utilize logistic regression, linear discriminant analysis, K-fold cross validation with K = 10, and analyzing visualizations.

IV. Research Questions

This report seeks to answer the following questions:

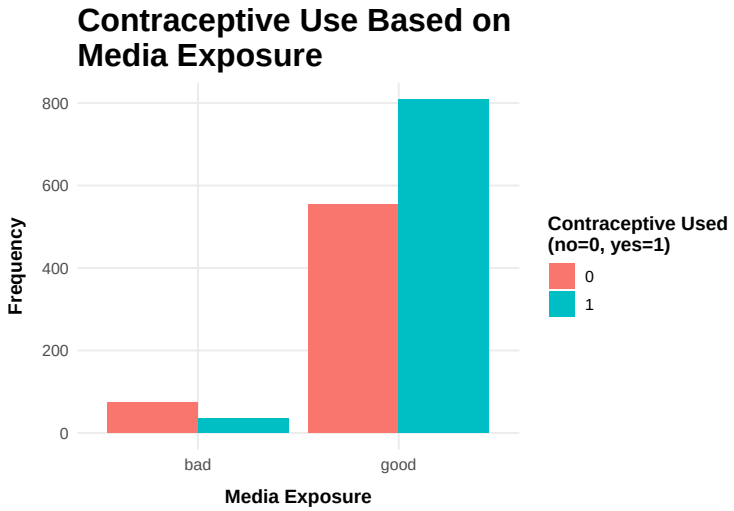
1. Which factors contribute to the choice of married women using contraceptives or not?

2. How accurately can we predict contraceptive use with these variables?

V. Visualizations

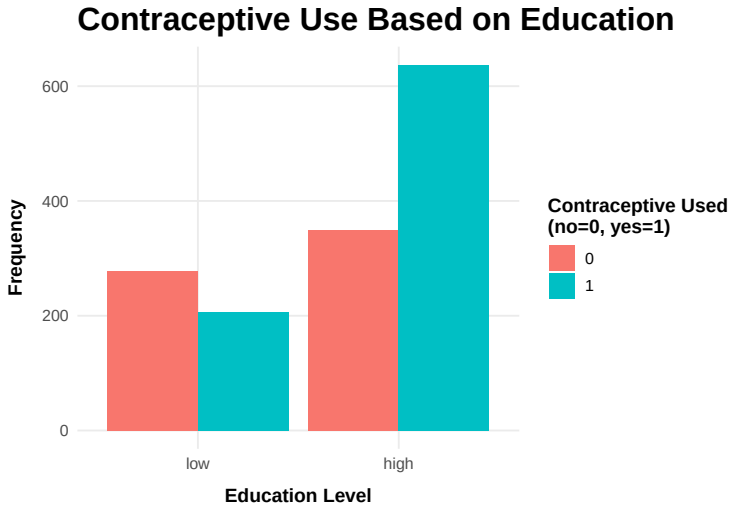
To get a better idea of which variables may have an effect on contraceptive use, we constructed a few visualizations to see how contraceptive use may vary across categories.

Good/Bad Media Consumption Graph



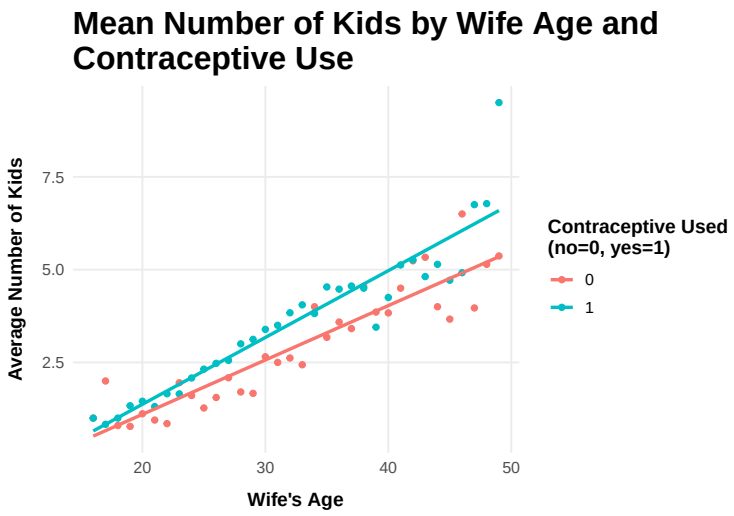
The bar graph compares contraceptive use of wives with good media exposure to the contraceptive use of wives with bad media exposure. We are unsure how the UC Irvine Machine Learning repository categorized 'good' and 'bad' media exposure. Wives with good media exposure chose to use contraceptives more often than not, whereas wives with bad media exposure were more likely to choose not to use contraceptives. This may be an indicator that the type of media exposure is significant to predicting whether or not wives choose to use contraceptives.

High/Low Wife Education Plot



This bar graph compares contraceptive use of wives with low education level to the contraceptive use of wives with high education level. Wives with low education chose not to use contraceptives more often, whereas wives with a high education level were more likely to use contraceptives than not. There is a higher discrepancy of contraceptive use for wives with high education, which may indicate significance when predicting contraceptive use.

Mean Number of Kids by Wife Age and Contraceptive Use



The scatter plot graphs the average number of kids for each age of the wives, separated by contraceptive use. Here, we can see that the wife's age and the corresponding number of kids have a linear relationship. Interestingly, we can also observe that the more kids that a woman has, the more likely she is to use a contraceptive at some point in her life. This may indicate significance in contraceptive use for both age and number of kids.

VI. Modeling

Our main objective when exploring this data set was to determine which factors may affect whether or not a woman uses contraceptives. This is a question of classification, and as such, linear regression is not an appropriate option to analyze the data. Therefore, we employed both logistic regression and linear discriminant analysis to fit models that may be able to predict whether or not a woman chooses to use contraceptives.

a) Logistic Regression

We started by using logistic regression to predict contraceptive usage among the observations in our data set. We began by fitting a large model that used most of the variables to try to predict whether or not a woman used contraceptives by: a woman's age, her education, the education of her husband, whether or not she was Muslim, her standard of living, her media exposure, the number of kids she has, and whether or not she works. Here is the model summary for the large model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.4108765	0.4104225	1.0011061	0.3167755
wife_age	-0.0766365	0.0090635	-8.4555471	0.0000000
wife_educationhigh	0.8249148	0.1396700	5.9061687	0.0000000
husband_educationhigh	0.2103563	0.1847936	1.1383313	0.2549822
factor(muslim)1	-0.5032121	0.1682791	-2.9903416	0.0027867
standard_of_livinghigh	0.5269829	0.1419216	3.7131962	0.0002047
media_exposuregood	0.6490773	0.2446343	2.6532553	0.0079720
number_of_children	0.3170932	0.0334343	9.4840586	0.0000000
factor(wife_works)1	0.0829613	0.1311247	0.6326898	0.5269362

As seen, many of the observations were significant at the 0.05 alpha level. However, the education level of the husband and whether or not the wife works were insignificant in this model. As such, we decided to drop these predictors in our second model, which was fit with the same predictors otherwise. The summary of the smaller model can be seen below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.5745499	0.3873219	1.483391	0.1379705
wife_age	-0.0771089	0.0090174	-8.551123	0.0000000
wife_educationhigh	0.8818497	0.1300919	6.778665	0.0000000
standard_of_livinghigh	0.5447293	0.1406507	3.872923	0.0001075
media_exposuregood	0.6949742	0.2417025	2.875329	0.0040361
factor(muslim)1	-0.5043208	0.1681466	-2.999293	0.0027061
number_of_children	0.3169668	0.0330824	9.581130	0.0000000

All of these predictors are significant at the 0.05 alpha level, and as such will all be statistically significant in predicting contraceptive choice.

Since there is no testing data available to us, we wanted to see how accurate our model was using the training data. We used a standard probability cutoff of greater than 0.5 to assign a class of 1 to the data and 0 otherwise. We then formed a confusion matrix based on this data to assess how well our model performed compared to the actual data. The confusion matrix for the first model is as follows, with the reference data as the columns and the predicted data as the rows:

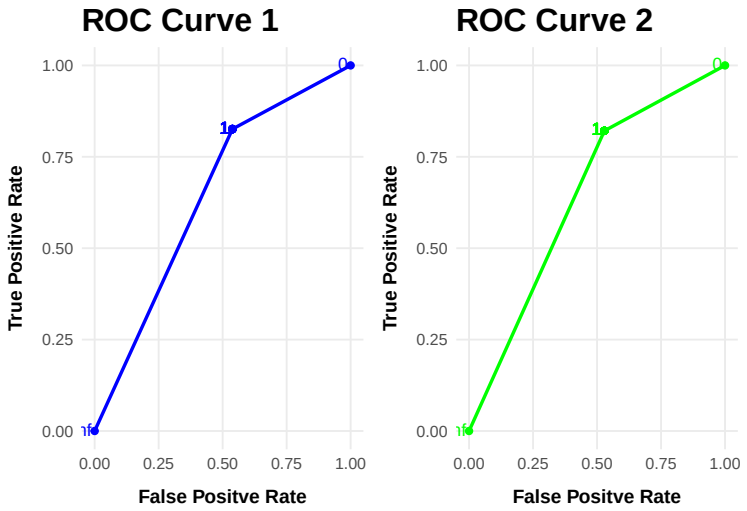
	0 - Reference	1 - Reference
0 - Predicted	290	147
1 - Predicted	338	697

This model has an accuracy of 0.6705. Unfortunately, this model had a fairly low true positive rate with 0.4618, but it had a higher true negative rate with 0.8258.The second model's confusion matrix is as follows:

	0 - Reference	1 - Reference
0 - Predicted	296	151
1 - Predicted	332	693

This model had a slightly higher accuracy of 0.6719. It also had a fairly low true positive rate of 0.4713, but was higher than the previous model. The true negative rate in this model was 0.8211, which was slightly lower than the true negative rate in the previous model. However, this model was overall slightly more accurate than the last model, so this model was a better predictor for our data.

The ROC curves for the two models are shown below:



We can see from the ROC curves that the second model has a slightly bigger area under the curve than the first model. This means that the probability that the second model classifies each category correctly is higher than the first model.

b) Linear Discriminant Analysis

Next, we repeated this process with linear discriminant analysis. Since our data was presumably collected via a random sample, we can assume that our observations are independent, and therefore use linear discriminant analysis. We fit 2 LDA models with the same predictors in our two logistic regression models above. The confusion matrix for the first linear discriminant model is as follows:

	0 - Reference	1 - Reference
0 - Predicted	281	144
1 - Predicted	347	700

This model had an accuracy of 0.6664, which was lower than the accuracies of both of our logistic regression models. It had a true negative rate of 0.4475 and a true positive rate of 0.8294, which means that the probability of the model assigning the class of "no contraceptive use" correctly is 0.4475, while the probability of assigning the class of "contraceptive use" correctly is 0.8294. The second LDA model performed slightly better, and the confusion matrix is as follows:

	0 - Reference	1 - Reference
0 - Predicted	288	147
1 - Predicted	340	697

This model had an accuracy of 0.6692, which is slightly higher than the first linear discriminant model. It had a true negative rate of 0.4586 and a true positive rate of 0.8258. This model performed better than the first LDA model, but other methods to assess the goodness of fit for these models are necessary to truly assess which ones perform better than others. One such method is cross-validation.

VII. K-fold Cross Validation

We can assess model accuracy through cross-validation methods. We decided to use the k-fold cross-validation method with K=10 because we determined the number of observations in our data set to be too large (n = 1472) to use a more computationally expensive method such as leave-one-out cross-validation.

When we performed the k-fold cross-validation method on our logistic regression models, we obtained an accuracy of approximately 0.6679 for our first model, and then obtained an accuracy of approximately 0.6684 for our second model. When we performed the k-fold cross-validation method on our LDA models, we obtained an accuracy of approximately 0.6685 for our first model, and then obtained an accuracy of approximately 0.6697 for our second model. Based on our findings, our second model performs slightly better than the first model, which is consistent with our earlier validation methods.

VIII. Conclusion

With our analytical findings, we were able to find variables that provide statistical significance to predicting the choice of married women using contraceptives or not. Through conducting logistic regression, LDA models, and K-fold cross validation methods, with K=10, we were able to determine that the statistically significant predictor variables were: a woman's age, her education, whether or not she was Muslim, her standard of living, her media exposure, and the number of kids she has. Our models with these predictors were consistently giving a high accuracy compared to the models we made with all the predictor variables provided in the data set. After seeing that the model with only statistically significant predictors is more accurate at predicting the choice of a married woman using contraceptives, we were able to look more closely at which model of the two statistical analysis models we utilized was best. Our model obtained using logistic regression gives us an accuracy of 0.6719, while the model obtained using LDA gave the accuracy of 0.6692. Additionally, using k-fold cross validation at K=10, we found the accuracy for the logistic model was 0.6684, and the accuracy for the LDA model was 0.6697.

Out of two models created with the same statistically significant predictor variables, we can conclude that the LDA model was able to give the highest predicted accuracy at 0.6697 using k-fold cross-validation. While this figure was lower than the accuracy obtained with the second logistic model running the training data against our predictors, using k-fold cross validation is a much more statistically sound validation approach. As such, our best approximation for how accurately we can predict contraceptive use with these variables is 0.6697. However, based on the conclusions from the ROC curve and looking at the highest possible accuracies we were able to obtain using k-fold cross validation at K=10 on the logistic regression models and the LDA models, we concluded our accuracies were relatively low. Despite the fact that these variables were statistically significant when classifying contraceptive use, further information may be necessary to increase the accuracy of our classification methods.

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
library(plotROC)
library(tidyverse)
library(caret)
library(MASS)
library(pastecs)
library(knitr)
library(gridExtra)

#factored_case_when function
factored_case_when <- function(...) {
  args <- rlang::list2(...)
  rhs <- map(args, rlang::f_rhs)

  cases <- case_when(
    !!!args
  )

  purrr::exec(fct_relevel, cases, !!!rhs)
}

#Theme nice
theme_nice = function() {
  theme_minimal(base_size = 12) +
    theme(panel.grid.minor = element_blank(),
          plot.title = element_text(face = "bold", size = rel(1.7)),
          plot.subtitle = element_text(face = "plain", size = rel(1.3), color = "grey70"),
          plot.caption = element_text(face = "italic", size = rel(0.7),
                                      color = "grey70", hjust = 0),
          legend.title = element_text(face = "bold"),
          strip.text = element_text(face = "bold", size = rel(1.1), hjust = 0),
          axis.title = element_text(face = "bold"),
          axis.title.x = element_text(margin = margin(t = 10), hjust = 0.5),
          axis.title.y = element_text(margin = margin(r = 10), hjust = 0.5),
          strip.background = element_rect(fill = "grey90", color = NA),
          panel.border = element_rect(color = NA, fill = NA))
}

theme_set(theme_nice())
df = read.delim("C:/Users/herob/Documents/GitHub/sta141a-proj/Data/cmc.data", sep = ",")

#Rename columns
names(df) = c("wife_age", "wife_education", "husband_education", "number_of_children", "muslim",
              "wife_works", "husband_occupation", "standard_of_living", "media_exposure",
              "contraceptive_used")

#convert Wife education into factors
df = df |>
  mutate(wife_education = factored_case_when(
    wife_education == 1 | wife_education == 2 ~ "low",
    wife_education == 3 | wife_education == 4 ~ "high"
```

```
  ))

#convert husband education into factors
df = df |>
  mutate(husband_education = factored_case_when(
    husband_education == 1 | husband_education == 2 ~ "low",
    husband_education == 3 | husband_education == 4 ~ "high"
  ))

#convert husband occupation into factors
df = df |>
  mutate(husband_occupation = factored_case_when(
    husband_occupation == 1 | husband_occupation == 2 ~ "low-income",
    husband_occupation == 3 | husband_occupation == 4 ~ "high-income"
  ))

#convert standard-of-living into factors
df = df |>
  mutate(standard_of_living = factored_case_when(
    standard_of_living == 1 | standard_of_living == 2 ~ "low",
    standard_of_living == 3 | standard_of_living == 4 ~ "high"
  ))

#convert media exposure into factors
df = df |>
  mutate(media_exposure = factored_case_when(
    media_exposure == 1 ~ "bad",
    media_exposure == 0 ~ "good"
  ))

#combine contraceptive use into two outcomes
df = df |>
  mutate(contraceptive_used = case_when(
    contraceptive_used == 1 ~ 0,
    contraceptive_used == 2 | contraceptive_used == 3 ~ 1
  ))

clean_data = df
mean(clean_data$wife_age)
sd(clean_data$wife_age)
ggplot(clean_data, aes(x = wife_age)) +
  geom_histogram(color = "black", fill = "grey")+
  ggtitle("Frequency of Wife Age in Data")+
  xlab("Wife Age")+
  ylab("Number of Occurances")
kable(table(clean_data$wife_education), col.names = c("Education Level", "Count"))
#finding the frequency wife with low education and wife with high education levels
mean(clean_data$number_of_children)
sd(clean_data$number_of_children)
kable(table(clean_data$number_of_children), col.names = c("Number of Children", "Count"))
# finding the frequency of the number of children each wife has
ggplot(clean_data, aes(x = factor(number_of_children))) +
  geom_histogram(stat = "count", color = "black", fill = "grey")+
```

```
  ggtitle("Frequency of Children for Each Woman")+
  xlab("Number of Children")+
  ylab("Number of Occurances")
kable(table(clean_data$muslim), col.names = c("Muslim (0 = No, 1 = Yes)", "Count"))
#finding frequency of whether the wife is Muslim or not
kable(table(clean_data$husband_education), col.names = c("Education Level", "Count"))
#finding the frequency of husband with low education and husband with high education levels
kable(table(clean_data$wife_works), col.names = c("Working(0 = No, 1 = Yes)", "count"))
#finding the frequency of whether the wife works or not
mean(clean_data$wife_works)
kable(table(clean_data$standard_of_living), col.names = c("Standard of Living", "Count"))
# finding the frequency of low and high standard of living
kable(table(clean_data$media_exposure), col.names = c("Media Exposure", "Count"))
# finding the frequency of bad and good media exposure
kable(table(clean_data$contraceptive_used), col.names = c("Contraceptive Use(0 = No, 1 = Yes)", "Count"
ggplot(clean_data, aes(x = factor(contraceptive_used))) +
  geom_histogram(stat = "count",color = "black", fill = "grey")+
  ggtitle("Frequency of Whether Contraceptive\nUsed or Not")+
  xlab("Contraceptive Used or Not\n(No = 0, Yes = 1)")+
  ylab("number of occurances")
#putting the frequency of whether contraceptive was used or not into a plot to visualize the data
mean(clean_data$contraceptive_used)      #finding the proportion
media_count = clean_data |>
  group_by(media_exposure, contraceptive_used) |>
  tally()

combined.plot2 <- ggplot(data=media_count, aes(x = media_exposure, y = n, fill=factor(contraceptive_use
  geom_col(position = "dodge") +
  labs(title="Contraceptive Use Based on\nMedia Exposure",
       x='Media Exposure',
       y='Frequency',
       fill='Contraceptive Used\n(no=0, yes=1)')
combined.plot2
#count by wife education and contraceptive used
education_count = clean_data |>
  group_by(wife_education, contraceptive_used) |>
  tally()

combined.plot <- ggplot(data=education_count, aes(x = wife_education, y = n, fill=factor(contraceptive_
  geom_col(position = "dodge") +
  labs(title="Contraceptive Use Based on Education",
       x='Education Level',
       y='Frequency',
       fill='Contraceptive Used\n(no=0, yes=1)')
combined.plot
#find average number of kids by age
average_kids = clean_data |>
  group_by(wife_age, contraceptive_used) |>
  summarise(mean_kids = mean(number_of_children))

kids_plot = ggplot(data = average_kids, aes(x = wife_age, y = mean_kids, color = factor(contraceptive_us
  geom_point() + geom_smooth(method = "lm", se = FALSE) +
  labs(x="Wife's Age", y="Average Number of Kids",
```

```
        color ="Contraceptive Used\n(no=0, yes=1)",
       title = "Mean Number of Kids by Wife Age and\nContraceptive Use")
kids_plot
lrm_1 = glm(contraceptive_used ~ wife_age + wife_education + husband_education + factor(muslim) +
              standard_of_living + media_exposure + number_of_children + factor(wife_works),
            data = clean_data, family = "binomial")
lrm_2 = glm(contraceptive_used ~ wife_age + wife_education + standard_of_living + media_exposure +
              factor(muslim) + number_of_children,
            data = clean_data, family = "binomial")
kable(summary(lrm_1)$coefficients)
kable(summary(lrm_2)$coefficients)
lrm_predict_1 = ifelse(lrm_1$fitted.values > .5, 1, 0)
lrm_predict_2 = ifelse(lrm_2$fitted.values > .5, 1, 0)
actual_values = clean_data$contraceptive_used

#make confusion matricies for logistic regression
lrm_cm_1 = confusionMatrix(data = factor(lrm_predict_1), reference = factor(actual_values))$table
rownames(lrm_cm_1) = c("0 - Predicted", "1 - Predicted")
colnames(lrm_cm_1) = c("0 - Reference", "1 - Reference")
kable(lrm_cm_1)
lrm_cm_2 = confusionMatrix(data = factor(lrm_predict_2), reference = factor(actual_values))$table
rownames(lrm_cm_2) = c("0 - Predicted", "1 - Predicted")
colnames(lrm_cm_2) = c("0 - Reference", "1 - Reference")
kable(lrm_cm_2)
roc_1 = ggplot(clean_data, aes(m = lrm_predict_1, d = contraceptive_used)) +
  geom_roc(n.cuts = 20,color = "blue") +
  labs(x = "False Positve Rate",
       y = "True Positive Rate",
       title = "ROC Curve 1")

#ROC Curve for model 2
roc_2 = ggplot(clean_data, aes(m = lrm_predict_2, d = contraceptive_used)) +
  geom_roc(n.cuts = 20,color = "green") +
  labs(x = "False Positve Rate",
       y = "True Positive Rate",
       title = "ROC Curve 2")

grid.arrange(roc_1, roc_2, nrow = 1)
lda_1 = lda(contraceptive_used ~ wife_age + wife_education + husband_education + factor(muslim) +
              standard_of_living + media_exposure + number_of_children + factor(wife_works),
            data = clean_data)
lda_2 = lda(contraceptive_used ~ wife_age + wife_education + standard_of_living + media_exposure +
              factor(muslim) + number_of_children, data = clean_data)

#make lda predictions
lda_predict_1 = predict(lda_1, clean_data)
lda_predicted_1 = lda_predict_1$class

lda_predict_2 = predict(lda_2, clean_data)
lda_predicted_2 = lda_predict_2$class

#confusion matrix for lda
lda_cm_1 = confusionMatrix(data = lda_predicted_1, reference = factor(actual_values))$table
```

```r
rownames(lda_cm_1) = c("0 - Predicted", "1 - Predicted")
colnames(lda_cm_1) = c("0 - Reference", "1 - Reference")
kable(lda_cm_1)
lda_cm_2 = confusionMatrix(data = lda_predicted_2, reference = factor(actual_values))$table
rownames(lda_cm_2) = c("0 - Predicted", "1 - Predicted")
colnames(lda_cm_2) = c("0 - Reference", "1 - Reference")
kable(lda_cm_2)
set.seed(12)

#using k = 10 for k-fold
#performing k-fold on logistic regression models
kfold_lrm_1 = train(factor(contraceptive_used) ~ wife_age + wife_education + husband_education +
                      factor(muslim) + standard_of_living + media_exposure + number_of_children +
                      factor(wife_works),
                    data = clean_data,
                    method = "glm",
                    family = "binomial",
                    trControl = trainControl(method = "cv", number = 10))
kfold_lrm_2 = train(factor(contraceptive_used) ~ wife_age + wife_education + standard_of_living +
                      media_exposure + factor(muslim) + number_of_children,
                    data = clean_data,
                    method = "glm",
                    family = "binomial",
                    trControl = trainControl(method = "cv", number = 10))

#output results
kfold_lrm_1
kfold_lrm_2

#performing k-fold on lda models
kfold_lda_1 = train(factor(contraceptive_used) ~ wife_age + wife_education + husband_education +
                      factor(muslim) + standard_of_living + media_exposure + number_of_children +
                      factor(wife_works),
                    data = clean_data,
                    method = "lda",
                    trControl = trainControl(method = "cv", number = 10))
kfold_lda_2 = train(factor(contraceptive_used) ~ wife_age + wife_education + standard_of_living +
                      media_exposure + factor(muslim) + number_of_children,
                    data = clean_data,
                    method = "lda",
                    trControl = trainControl(method = "cv", number = 10))

#output results
kfold_lda_1
kfold_lda_2
```