Can machine learning predict risky sexual behavior?

Alonso Quijano

1/2/2020

Introduction

Motivation

This project was developed as part of the **edX Data Science Professional Programm** capstone project.

Latin America and the Caribbean are the regions with the second largest ratio of adolescent pregnancy in the world (Pan American Health Organization, 2016). Particularly in Ecuador, the adolescent birth rate is much higher than the regional average. While the rate in the region stands at 66.5 births per 1000 girls (15 to 19 years old), in Ecuador it is estimated at 76.5. (Minsiterio de Salud Publica del Ecuador, 2018). This is a major issue as many girls and adolescents must drop out of school to become young mothers, which has long-term repercussions in their health and the course of their lives. As a consequence, their educational and employment opportunities are affected, not to mention their psychological health and wellbeing (Pan American Health Organization, 2016).

Are parents aware of the sexual behavior of their children? At least in Ecuador, the answer is generally no. In fact, talking about sexuality is something many parents in Ecuador feel embarrassed about (Rodriguez, 2015). Additionally, they feel reluctant to leave the role of educating their children in these matters to schools and the government as demonstrated by the 2017 protests against education that promoted gender identity, which were supported by religious groups and were held in various cities of Ecuador. Their campaign slogan was: Don't meddle with my children (El Comercio, 2017).

Project's objective

Although the use of condom can highly reduce the probability of getting STDs and avoid unwanted pregnancies, its use among teenagers, as well as the use of other contraception methods, is rare in Ecuador. How do teenagers decide when to start their sexual life, and how do they choose to use or not to use contraception? Although we may not be able to answer these questions robustly, we will deep dive into the factors that are correlated with a riskier sexual behavior, some of which are more obvious and some of which are less.

The objective of this project is to develop a machine-learning algorithm to predict risky sexual behavior on female adolescents in Ecuador. We want to identify the factors that can best predict sexual behavior and how they relate to the outcome. The goal is to create an algorithm that can be applied in the real world to identify the population more at risk of pregnancy. Therefore, by using strategic decision-making, we could help the government and organizations better direct programs and policies aimed to reduce teenage pregnancy, focusing on those who need them the most and saving time and other resources.

The idea of this project was inspired by the Poverty Probability Index (PPI), which is a poverty measurement tool created using machine-learning techniques. It is statistically-sound yet simple to use. By answering ten questions about a household's characteristics and asset ownership, one can compute the likelihood that the household is living below the poverty line. Although it is not part of the scope of this project, our goal is also to design an index similar to the PPI, by creating a questionnaire that can be easily administered in schools or households to know the risk of pregnancy.

The data

To elaborate the algorithm, we will use data from the Maternal and Child Health; and Sexual and Reproductive Health Survey (ENSANUT) in Ecuador collected in 2018. The goal of this survey was to know the status of: fertility, maternal and child health, child nutrition, food consumption, access to food programs, physical activity, among other topics. The sample consists of approximately 48,700 women of childbearing age. Only adolescents from 12 to 18 and their mothers were included in our sample.

The ENSANUT dataset consists of several sub-datasets, which are described as follows:

- **f1_personas:** Demographic and economic data of each of the members of the household;
- **f1_hogar:** Data about the house the household lives in;
- **f2_mef:** Data about the sexual health of women from 10 to 49;
- **f4_fact_riesgo:** Data about behavioral risk factors of people from 5 to 18

The data is stored in different .dta document files. Along with the datasets, there are the questionnaires and the manuals for each survey. We used the questionnaires to identify the variables we were interested in and the manuals to know how they were categorized.

The model

We used a logistic regression model. Logistic regression is the most commonly used generalized linear model (GLM). It is an extension of linear regression that assures that the estimate of $Pr(Y = 1|X = x)$ is between 0 and 1 (Irizarry, 2019).

In this model, we define Y as 1 for having or having had risky sexual behavior, and 0 for safe sexual behavior. X are the explanatory variables or predictors that will be used to predict the likelihood of risky sexual behavior.

The variables

The predicted variable Y

We want to predict if a teenager is at risk of pregnancy or is already pregnant

1. The woman has ever been pregnant: **risk**
2. The woman has ever had intercourse and never used contraception: **risk**
3. The woman did not use contraception at first intercourse: **risk**
4. The woman did not use contraception at last intercourse: **risk**
5. The woman wouldn't use contraception because does not like it: **risk**
6. The woman wouldn't use contraception because is afraid of side effects: **risk**
7. The woman wouldn't use contraception because the partner doesn't like it: **risk**
8. The woman wouldn't use contraception because she feels embarrassed: **risk**
9. The woman wouldn't use contraception because of economic reasons: **risk**
10. The wouldn't use contraception because she has no knowledge about contraception: **risk**
11. The woman wouldn't use contraception because of other reasons or simply does not know why: **risk**
12. The rest: **no risk**

The explanatory variables or predictors X

We considered variables from different dimensions: demographic, economic and social, educational (including sexual education), and behavioral

Demographic variables:

From now on, we will define the female teenagers as the "daughter" and their mothers as "mother" to make it clear who we are referring to.

- **age :** age of the daughter;
- **area :** area where the household is located (urban or rural);
- **ethnicity :** ethnicity of the daughter (indigenous, black, mestizo, white, etc.);
- **m.age :** age of the mother;
- **d.m.age.diff :** age difference between the mother and the daughter;
- **m.num.children :** number of children of the mother

Economic and social variables:

- **num.bedrooms :** number of bedrooms in the house;
- **internet :** whether there is internet access in the house;
- **cellphone :** whether the daughter has an activated cellphone;
- **transfer :** whether someone in the household receives an economic transfer from the government;
- **m.job :** whether the mother has a job or not;
- **f.live.house:** whether the father lives in the house

Educational variables (including those about sexual education):

- **attend.school :** whether the daughter attends school;
- **m.education :** mother's education attainment (none, primary school, secondary school, university)

Variables related to sexual education:

- **contraception.info :** whether the daughter has ever received information about contraception;
- **contraception.info.family :** whether the daughter has learned about contraception mainly from her family;
- **contraception.info.school :** whether the daughter has learned about contraception mainly from school;
- **period.info :** whether the daughter knew about menstruation when she had her first period;
- **pregnancy.info :** whether the daughter can correctly answer the question: can a women get pregnant at first intercourse?;
- **aids.info :** whether the daughter can correctly answer the question: can AIDS be transmited through handshake?;
- **m.pregnancy.info :** whether the mother can correctly answer the question: can a women get pregnant at first intercourse?;
- **m.aids.info :** whether the mother can correctly answer the question: can AIDS be transmited through handshake?

Behavioral variables:

- **intercourse :** whether the daughter has ever had intercourse;
- **alcohol :** whether the daughter has ever drunk alcohol;
- **alcohol.recent :** whether the daughter has drunk alcohol in the past 30 days;
- **smoke :** whether the daughter has ever smoked;
- **smoke.recent :** whether the daughter has smoked in the past 30 days;
- **m.age.first.intercourse :** mother's age at first intercourse;
- **m.contraception:** mother's use of contraception (is using, has ever used, has never used)

Analyzing the data

Firstly, we need to load the libraries we will use to analyze the data and create the algorithm.

```
library(tidyverse)
library(gridExtra)
library(caret)
library(tibble)
```

We will load the data from Github. Note that we have already transformed the data from "raw" data to "tidy" data. We can check the appendix of this report for more detail about the data wrangling process.

```
# We will load the data from Github and check the variables using the glimpse option.
githubURL <- "https://github.com/aquijanoruiz/HarvardX_capstone/raw/master/SexRisk/pregnancy_risk.rds"
preg.risk <- readRDS(url(githubURL))
glimpse((preg.risk))
```

```
## Observations: 8,607
## Variables: 38
## $ id.subject                  <chr> "01015000020103105", "010150000201...
## $ id.household                <chr> "010150000201031", "01015000020104...
## $ id.mother                   <chr> "01015000020103102", "010150000201...
## $ contraception               <fct> no sex, has never used, is using, ...
## $ contraception.no.use.reason <fct> NA, feels embarrassed, NA, NA, no ...
## $ contraception.first         <fct> no sex, no, no, no sex, no sex, no...
## $ contraception.last          <fct> no sex, no, no, no sex, no sex, no...
## $ pregnant                    <fct> no, yes, no, no, no, no, no, no, n...
## $ age                         <int> 18, 18, 15, 12, 16, 18, 13, 13, 13...
## $ area                        <fct> urban, urban, urban, urban, urban,...
## $ ethnicity                   <fct> mestizo, mestizo, mestizo, mestizo...
## $ m.age                       <int> 42, 38, 38, 38, 42, 40, 38, 38, 40...
## $ d.m.age.diff                <int> 24, 20, 23, 26, 26, 22, 25, 25, 27...
## $ num.bedrooms                <int> 3, 4, 4, 4, 2, 3, 3, 3, 3, 3, 1, 4...
## $ internet                    <fct> yes, yes, yes, yes, no, yes, yes, ...
## $ cellphone                   <fct> yes, no, yes, no, yes, yes, no, no...
## $ transfer                    <fct> no, no, no, no, si, no, si, si, no...
## $ f.live.house                <fct> yes, yes, yes, yes, no, no, no, no...
## $ attend.school               <fct> yes, yes, yes, yes, yes, yes, yes,...
## $ contraception.info          <fct> yes, yes, yes, no, yes, yes, no, n...
## $ contraception.info.family   <fct> no, no, no, no, no, no, no, no, no...
## $ contraception.info.school   <fct> yes, yes, yes, no, yes, yes, no, n...
## $ period.info                 <fct> yes, yes, yes, NA, no, yes, NA, NA...
## $ pregnancy.info              <fct> yes, yes, no, yes, yes, yes, no, n...
## $ aids.info                   <fct> no, no, no, no, no, no, yes, yes, ...
## $ intercourse                 <fct> no, yes, yes, no, no, no, no, no, ...
## $ alcohol                     <fct> NA, NA, NA, no, NA, NA, NA, NA, no...
## $ alcohol.recent              <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ smoke                       <fct> NA, NA, NA, no, NA, NA, NA, NA, no...
## $ smoke.recent                <fct> NA, NA, NA, no, NA, NA, NA, NA, no...
## $ m.num.children              <int> 3, 5, 5, 5, 4, 2, 2, 2, 2, 2, 3, 4...
## $ m.job                       <fct> yes, yes, yes, yes, no, yes, no, n...
## $ m.education                 <fct> secondary school, primary school, ...
## $ m.pregnancy.info            <fct> NA, yes, yes, yes, yes, no, yes, y...
```

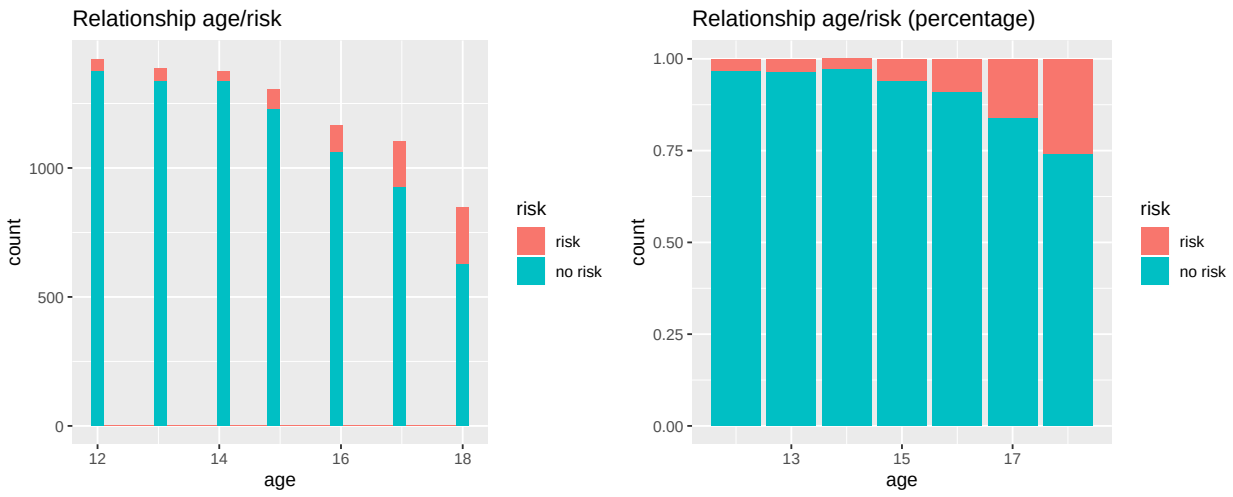
```
## $ m.aids.info                <fct> NA, no, no, no, yes, NA, no, no, n...
## $ m.age.first.intercourse    <int> NA, 16, 16, 16, 16, 17, 21, 21, 22...
## $ m.contraception            <fct> NA, is using, is using, is using, ...
## $ risk                       <fct> no risk, risk, risk, no risk, no r...
```

Demographic variables

- **age**

We can see the risk of pregnancy starts appearing as the age increases.

```
age.plot1 <- preg.risk %>% ggplot(aes(age, fill = risk)) +
  geom_histogram() + ggtitle("Relationship age/risk")

age.plot2 <- preg.risk %>% ggplot(aes(age, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship age/risk (percentage)")

grid.arrange(age.plot1,age.plot2,ncol=2)
```



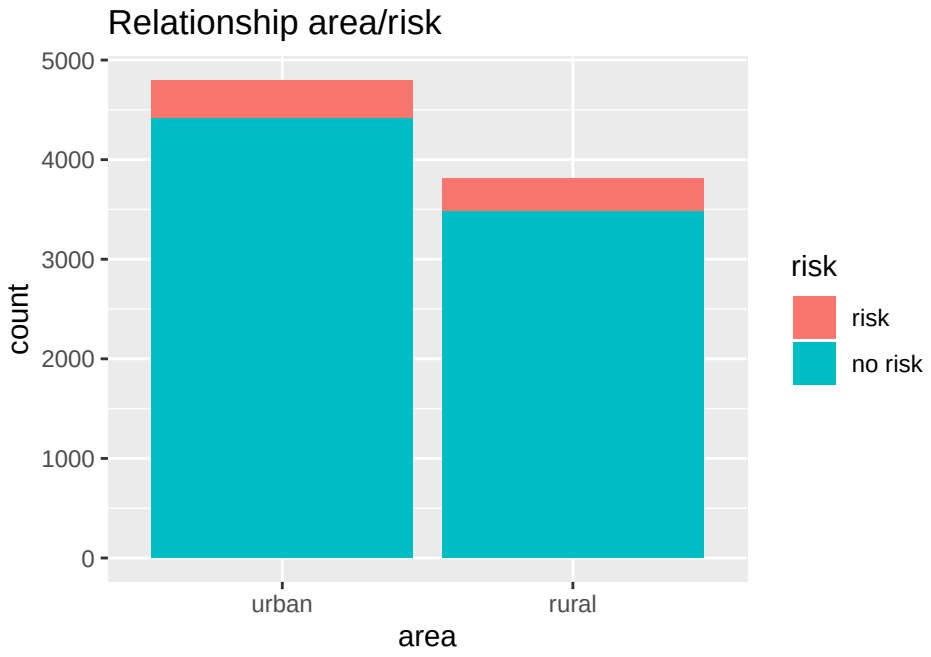

```
chisq.test(preg.risk$age, preg.risk$risk) # the chi square test shows it's true
```

```
FALSE
FALSE   Pearson's Chi-squared test
FALSE
FALSE data:  preg.risk$age and preg.risk$risk
FALSE X-squared = 587.91, df = 6, p-value < 2.2e-16
```

- **area**

We are not sure if the area has a significant correlation with the risk of pregnancy. There doesn't seem to be one.

```
preg.risk %>% ggplot(aes(area, fill = risk)) +
  geom_bar() + ggtitle("Relationship area/risk")
```



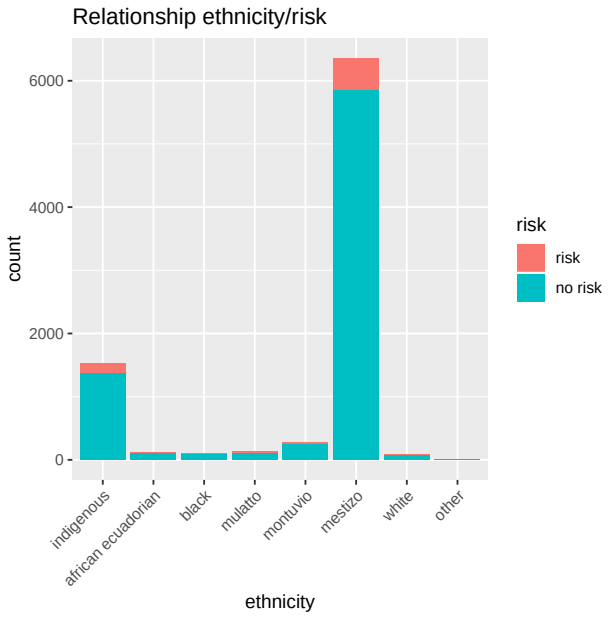

```
chisq.test(preg.risk$area, preg.risk$risk) # we can't reject the null
```

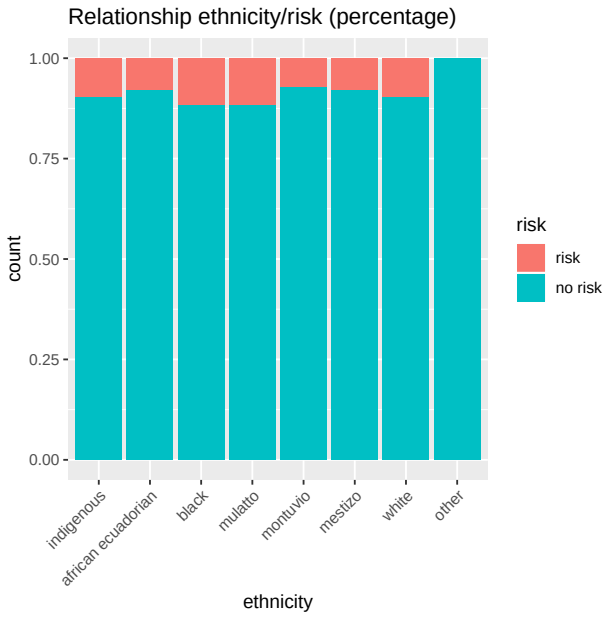
```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$area and preg.risk$risk
FALSE X-squared = 1.4422, df = 1, p-value = 0.2298
```

- **ethnicity**

It also doesn't seem to be a problem of ethnicity.

```
ethnicity.plot1 <- preg.risk %>% ggplot(aes(ethnicity, fill = risk)) +
  geom_bar() + ggtitle("Relationship ethnicity/risk") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

ethnicity.plot2 <- preg.risk %>% ggplot(aes(ethnicity, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship ethnicity/risk (percentage)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

grid.arrange(ethnicity.plot1,ethnicity.plot2,ncol=2)
```





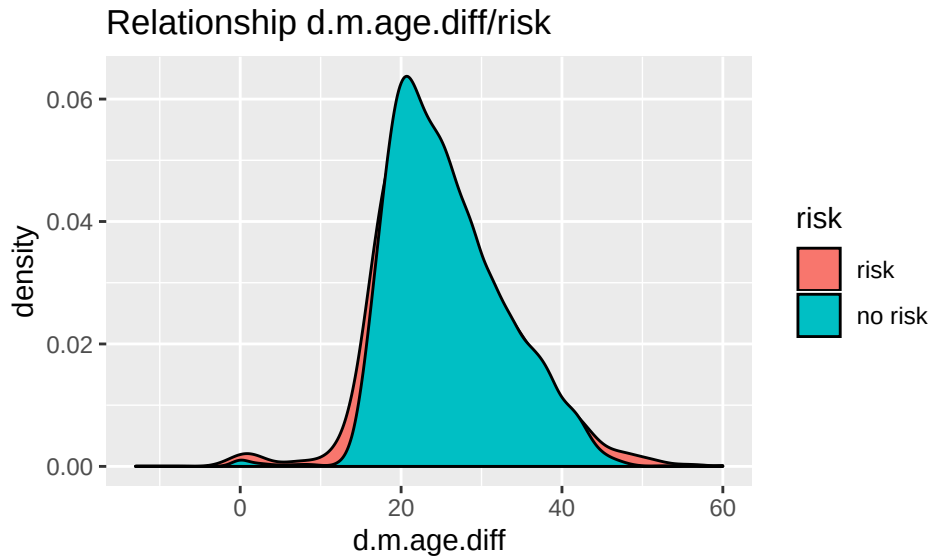

```
chisq.test(preg.risk$ethnicity, preg.risk$risk) # we can't reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test
FALSE
FALSE data:  preg.risk$ethnicity and preg.risk$risk
FALSE X-squared = 10.492, df = 7, p-value = 0.1624
```

- **d.m.age.diff**

It looks like very small and very large age differences between the daughter and the mother affect the risk.

```
preg.risk %>% ggplot(aes(d.m.age.diff, fill = risk)) +
  geom_density() + ggtitle("Relationship d.m.age.diff/risk")
```

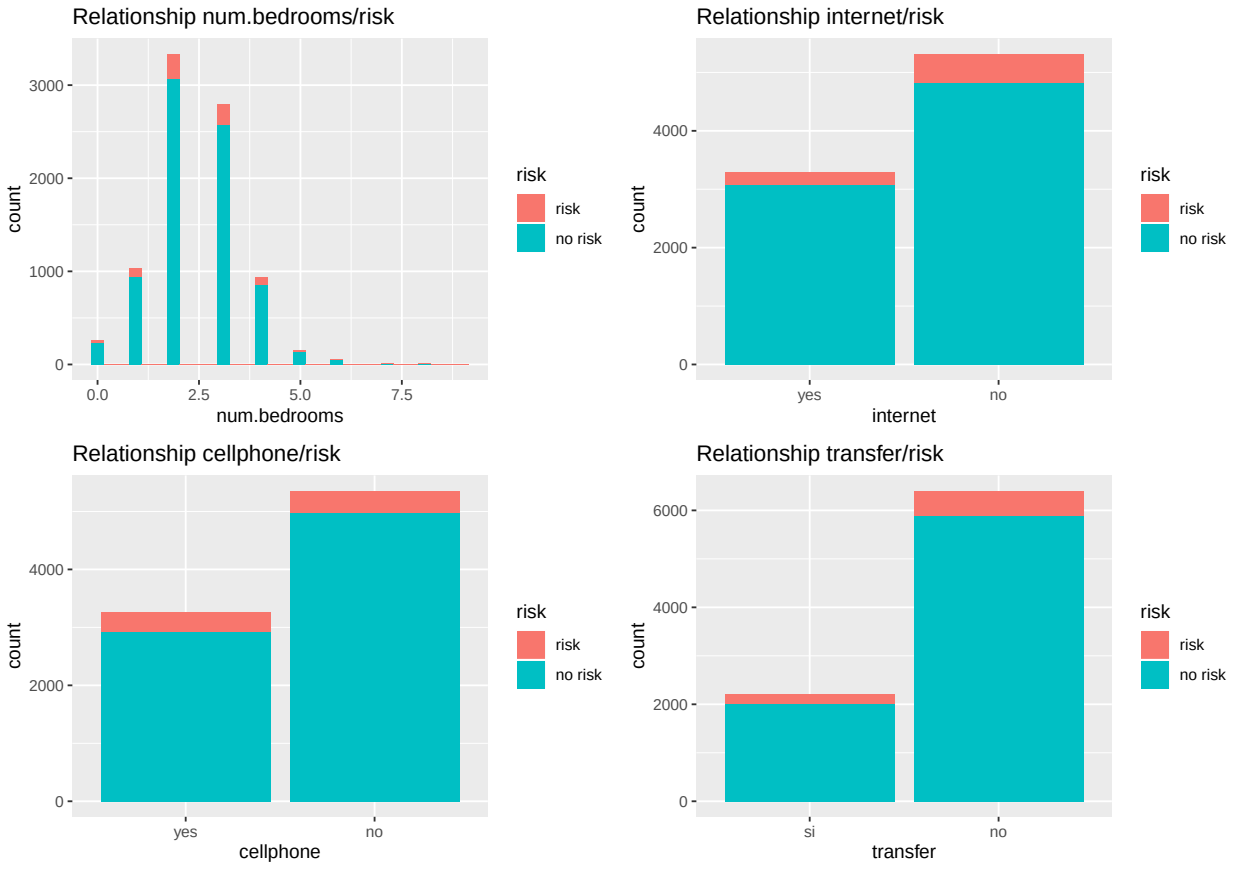


Economic and social variables

- **n.bedrooms, internet, cellphone, transfer**

The risk of pregnancy seesms to be correlated to income level.

```
num.bedrooms.plot <- preg.risk %>% ggplot(aes(num.bedrooms, fill = risk)) +
  geom_histogram() + ggtitle("Relationship num.bedrooms/risk")

internet.plot <- preg.risk %>% ggplot(aes(internet, fill = risk)) +
  geom_bar() + ggtitle("Relationship internet/risk")

cellphone.plot <- preg.risk %>% ggplot(aes(cellphone, fill = risk)) +
  geom_bar() + ggtitle("Relationship cellphone/risk")

transfer.plot <- preg.risk %>% ggplot(aes(transfer, fill = risk)) +
  geom_bar() + ggtitle("Relationship transfer/risk")

grid.arrange(num.bedrooms.plot, internet.plot, cellphone.plot, transfer.plot, nrow=2, ncol=2)
```




```
chisq.test(preg.risk$internet, preg.risk$risk)$p.value # we reject the null
```

```
FALSE [1] 2.33266e-05
```

```
chisq.test(preg.risk$cellphone, preg.risk$risk)$p.value  # we reject the null
```

```
FALSE [1] 1.360221e-08
```

```
chisq.test(preg.risk$transfer, preg.risk$risk)$p.value  # there seems to be a correlation
```

```
FALSE [1] 0.04004721
```

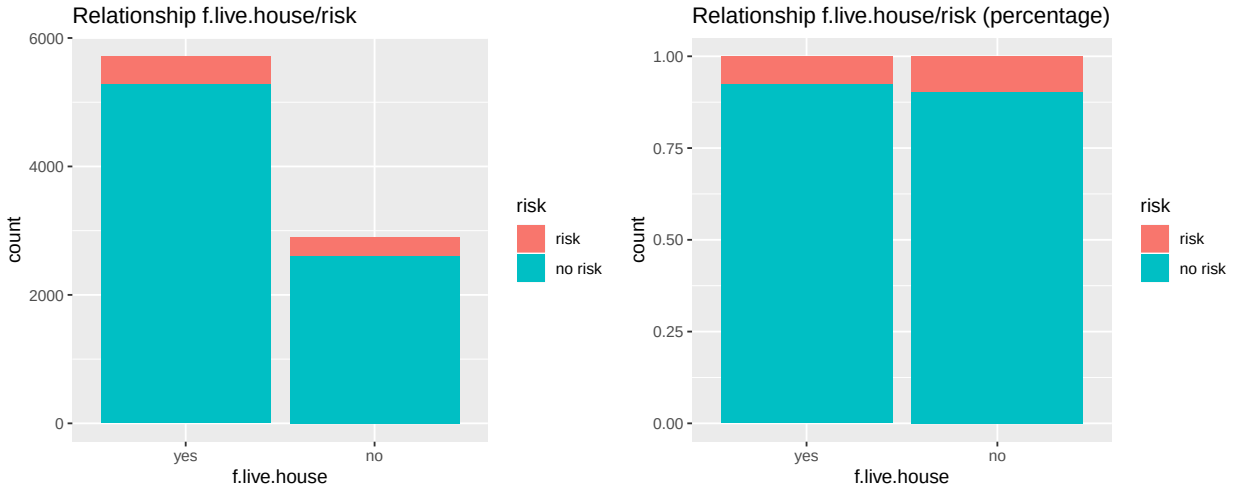
```
chisq.test(preg.risk$num.bedrooms, preg.risk$risk)$p.value  # the approximation may be incorrect
```

```
FALSE [1] 0.1506078
```

- **f.live.house**

Whether the father lives in the house or not (having a single-parent household) seems to be correlated with the risk of pregnancy. Although it is not so clear from the graph, the chi-square test shows there appears to be an association.

```
f.live.house.plot1 <- preg.risk %>% ggplot(aes(f.live.house, fill = risk)) +
  geom_bar() + ggtitle("Relationship f.live.house/risk")

f.live.house.plot2 <- preg.risk %>% ggplot(aes(f.live.house, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship f.live.house/risk (percentage)")

grid.arrange(f.live.house.plot1, f.live.house.plot2, ncol=2)
```



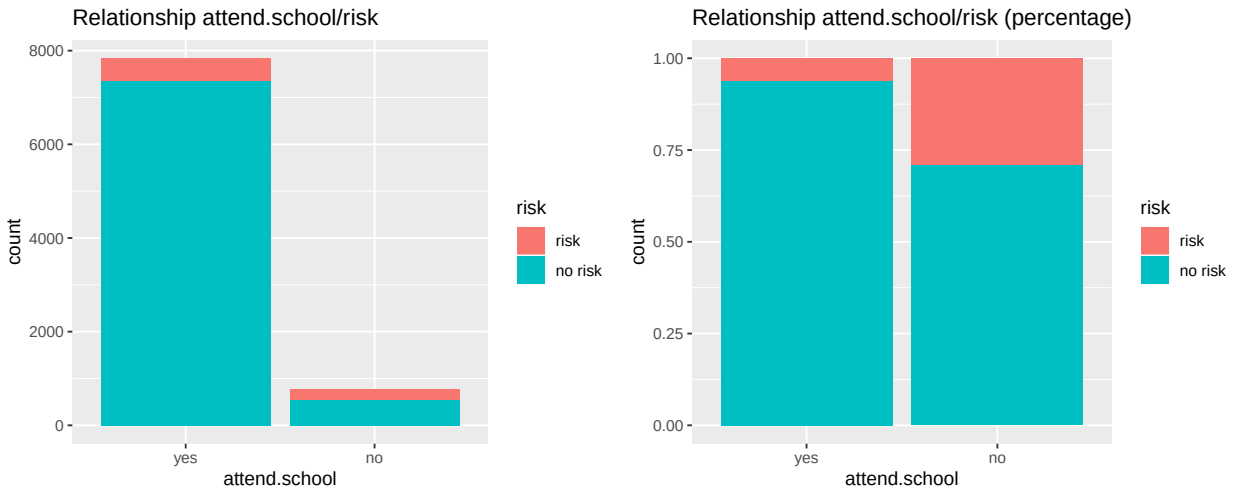

```
chisq.test(preg.risk$f.live.house, preg.risk$risk) # we reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$f.live.house and preg.risk$risk
FALSE X-squared = 14.41, df = 1, p-value = 0.0001471
```

Educational variables

It is evident that teenagers who do not attend school are more likely to be at risk of pregnancy. But a lot of our sample comes from people who were pregnant. So, maybe they left school after they got pregnant, and there may be reverse causality.

```
attend.school.plot1 <- preg.risk %>% ggplot(aes(attend.school, fill = risk)) +
  geom_bar() + ggtitle("Relationship attend.school/risk")

attend.school.plot2 <- preg.risk %>% ggplot(aes(attend.school, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship attend.school/risk (percentage)")

grid.arrange(attend.school.plot1, attend.school.plot2, ncol=2)
```



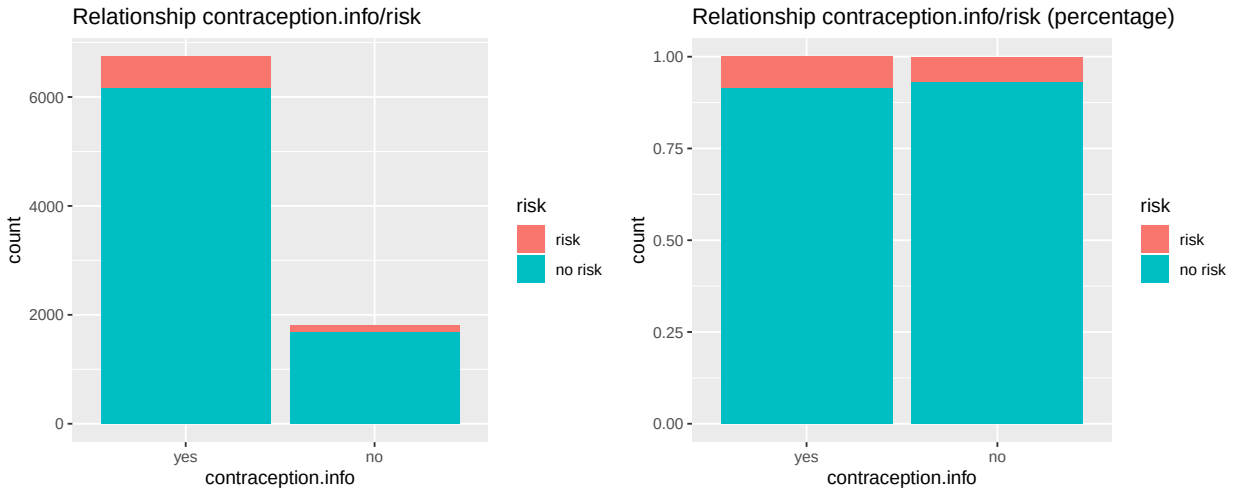

```
chisq.test(preg.risk$attend.school, preg.risk$risk) # we reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$attend.school and preg.risk$risk
FALSE X-squared = 482.09, df = 1, p-value < 2.2e-16
```

- **contraception.info**

The correlation is a little bit obscure. Having received information about contraception seems to increase the likelihood of pregnancy risk.

```
contraception.info.plot1 <- preg.risk %>% filter(!is.na(contraception.info)) %>%
  ggplot(aes(contraception.info, fill = risk)) +
  geom_bar() + ggtitle("Relationship contraception.info/risk")

contraception.info.plot2 <- preg.risk %>% filter(!is.na(contraception.info)) %>%
  ggplot(aes(contraception.info, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship contraception.info/risk (percentage)")

grid.arrange(contraception.info.plot1, contraception.info.plot2, ncol=2)
```



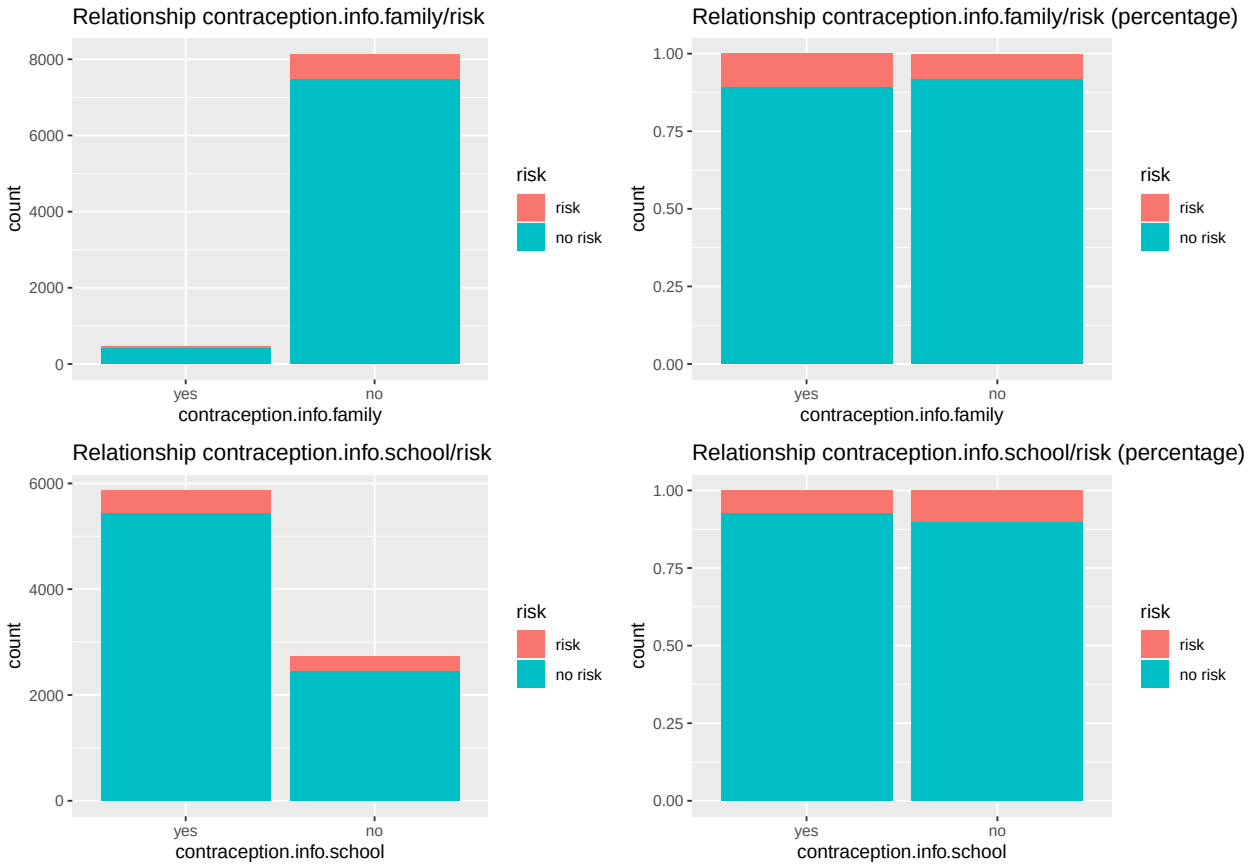

```
chisq.test(preg.risk$contraception.info, preg.risk$risk) # we reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$contraception.info and preg.risk$risk
FALSE X-squared = 6.1381, df = 1, p-value = 0.01323
```

- **contraception.info.family and contraception.info.school**

Those who learned more about sexuality at school are less likely to be at risk of pregnancy, while those who learned from their family are more likely to be at risk of pregnancy. Maybe parents are not so good at talking about sexuality with their kids.

```
contraception.info.family.plot1 <- preg.risk %>% ggplot(aes(contraception.info.family, fill = risk)) +
  geom_bar() + ggtitle("Relationship contraception.info.family/risk")

contraception.info.family.plot2 <- preg.risk %>% ggplot(aes(contraception.info.family, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship contraception.info.family/risk (percentage)")

contraception.info.school.plot1 <- preg.risk %>% ggplot(aes(contraception.info.school, fill = risk)) +
  geom_bar() + ggtitle("Relationship contraception.info.school/risk")

contraception.info.school.plot2 <- preg.risk %>% ggplot(aes(contraception.info.school, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship contraception.info.school/risk (percentage)")

grid.arrange(contraception.info.family.plot1, contraception.info.family.plot2,
             contraception.info.school.plot1, contraception.info.school.plot2, nrow=2, ncol=2)
```




```
chisq.test(preg.risk$contraception.info.family, preg.risk$risk)$p.value # we reject the null
```

```
FALSE [1] 0.05646993
```

```
chisq.test(preg.risk$contraception.info.school, preg.risk$risk)$p.value # we can't reject the null
```

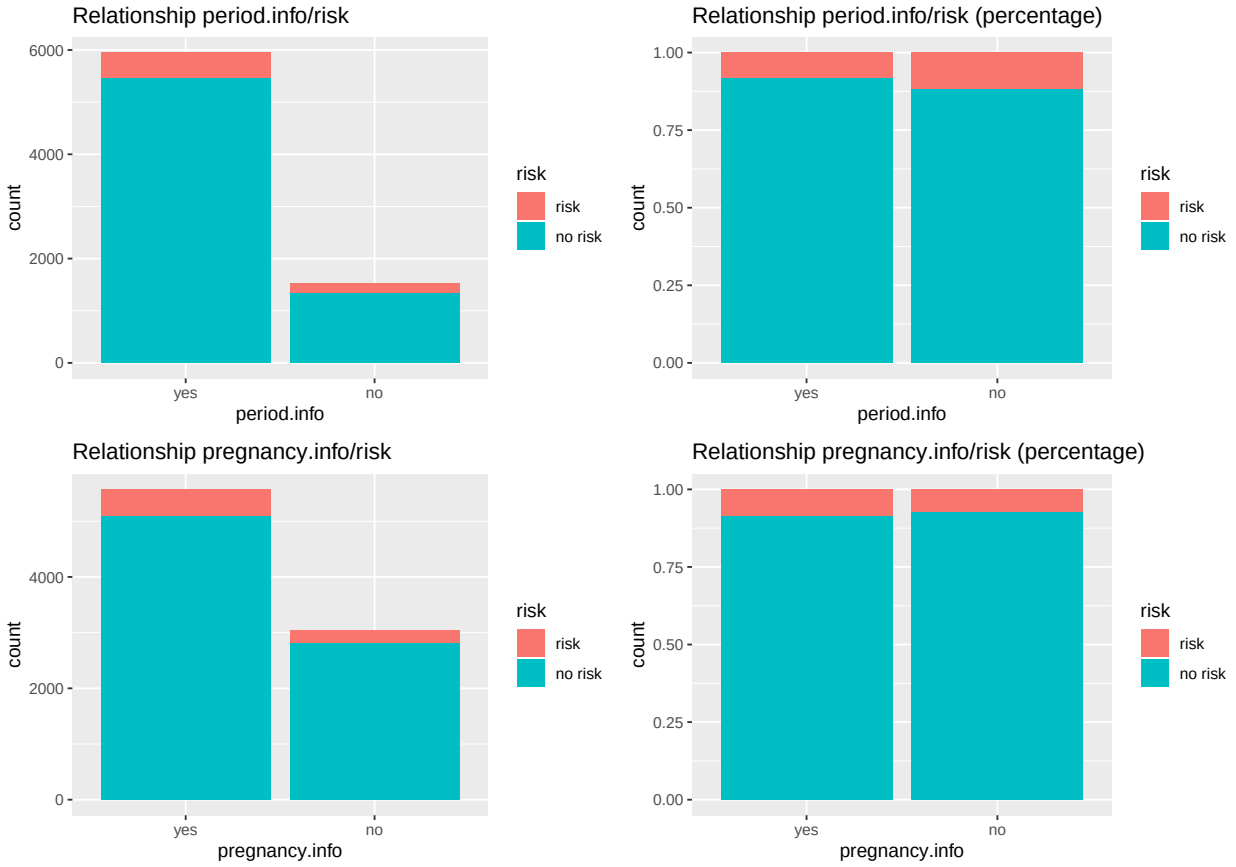
```
FALSE [1] 7.546633e-06
```

- **period.info and pregnancy.info**

Those who did not know what menstruation was when they had their first period are more likely to be at risk of pregnancy.

```
period.info.plot1 <- preg.risk %>% filter(period.info == "yes" | period.info == "no") %>%
  ggplot(aes(period.info, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar() + ggtitle("Relationship period.info/risk")

period.info.plot2 <- preg.risk %>% filter(period.info == "yes" | period.info == "no") %>%
  ggplot(aes(period.info, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar(position = "fill") + ggtitle("Relationship period.info/risk (percentage)")

pregnancy.info.plot1 <- preg.risk %>% ggplot(aes(pregnancy.info, fill = risk)) +
  geom_bar() + ggtitle("Relationship pregnancy.info/risk")
```

```
pregnancy.info.plot2 <- preg.risk %>% ggplot(aes(pregnancy.info, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship pregnancy.info/risk (percentage)")

grid.arrange(period.info.plot1, period.info.plot2,
             pregnancy.info.plot1, pregnancy.info.plot2, nrow=2, ncol=2)
```




```
chisq.test((preg.risk %>% filter(period.info == "yes" | period.info == "no"))$period.info,
           (preg.risk %>% filter(period.info == "yes" | period.info == "no"))$risk)$p.value # we reject
```

```
FALSE [1] 3.132794e-05
```

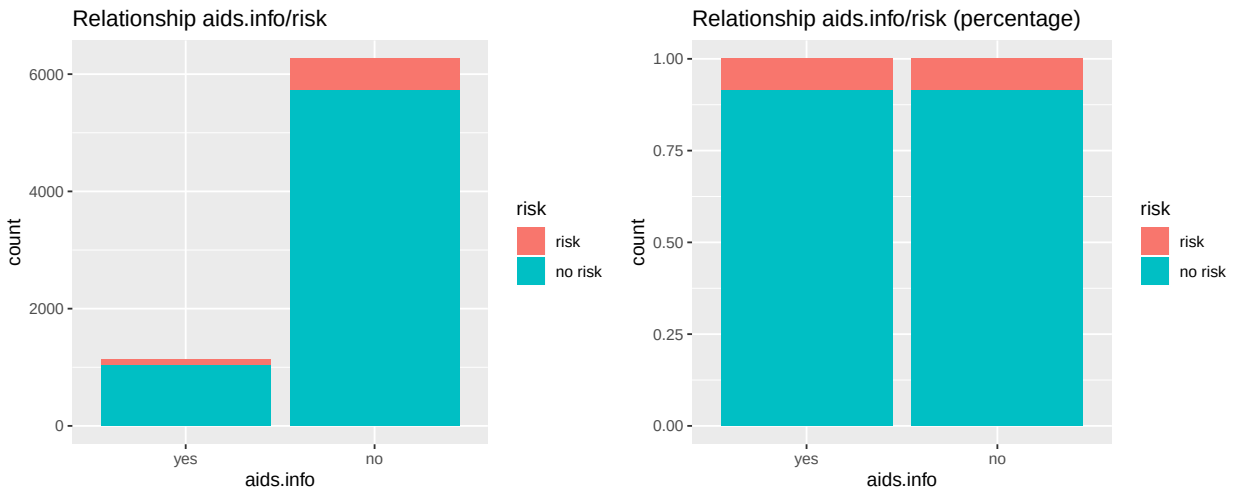
```
chisq.test(preg.risk$pregnancy.info, preg.risk$risk)$p.value # we reject the null
```

```
FALSE [1] 0.04817493
```

- **aids.info**

Although it is not very clear from the plot, there seems to be a correlation between the risk of pregnancy and the information about aids.

```
aids.info.plot1 <- preg.risk %>% filter(!is.na(aids.info)) %>% ggplot(aes(aids.info, fill = risk)) +
  geom_bar() + ggtitle("Relationship aids.info/risk")

aids.info.plot2 <- preg.risk %>% filter(!is.na(aids.info)) %>% ggplot(aes(aids.info, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship aids.info/risk (percentage)")

grid.arrange(aids.info.plot1, aids.info.plot2, ncol=2)
```




```
chisq.test(preg.risk$aids.info, preg.risk$risk) # we reject the null
```

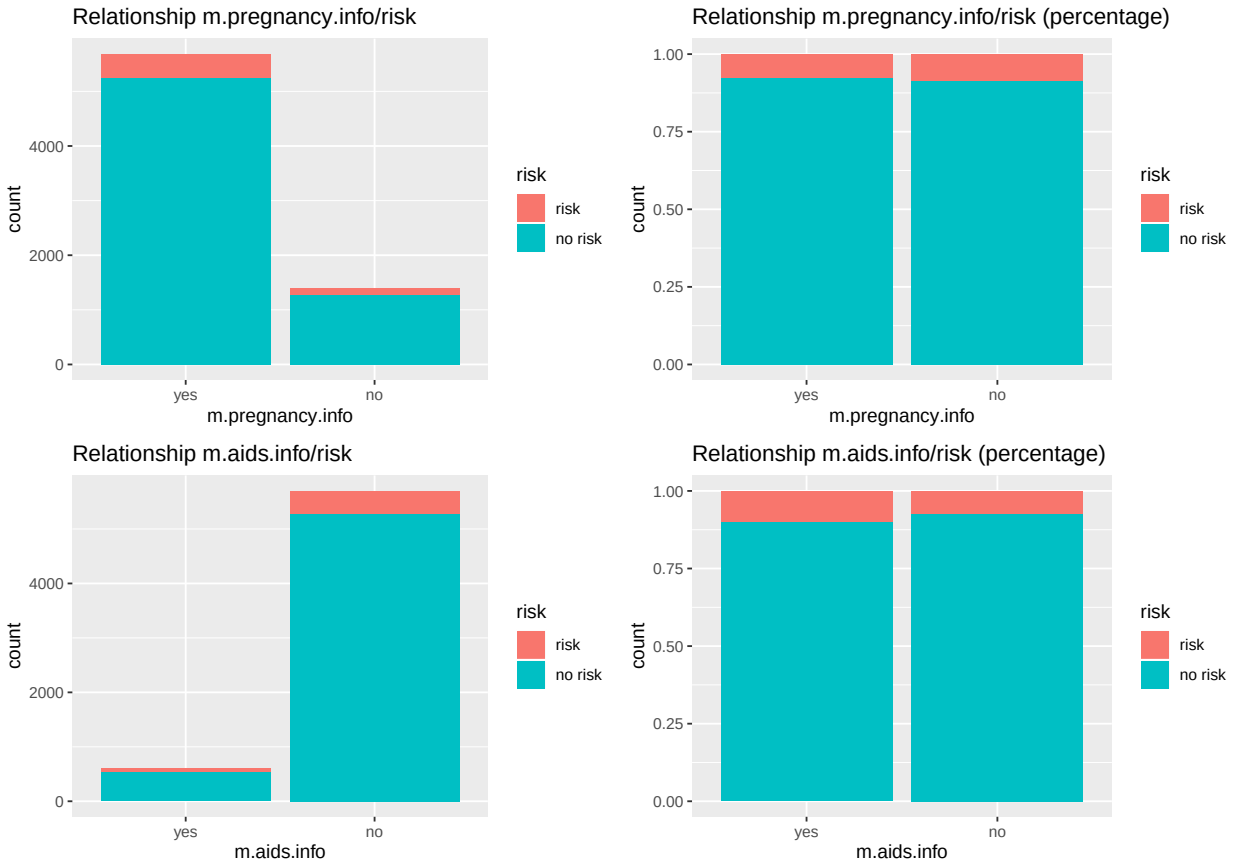
```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$aids.info and preg.risk$risk
FALSE X-squared = 3.2607e-29, df = 1, p-value = 1
```

- **m.pregnancy.info and m.aids.info**

There seems to be a correlation between the risk of pregnancy and the information about AIDS of the mother, but we can't say the same about the information about pregnancy.

```
m.pregnancy.info.plot1 <- preg.risk %>% filter(!is.na(m.pregnancy.info)) %>%
  ggplot(aes(m.pregnancy.info, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar() + ggtitle("Relationship m.pregnancy.info/risk")

m.pregnancy.info.plot2 <- preg.risk %>% filter(!is.na(m.pregnancy.info)) %>%
  ggplot(aes(m.pregnancy.info, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar(position = "fill") + ggtitle("Relationship m.pregnancy.info/risk (percentage)")

m.aids.info.plot1 <- preg.risk %>% filter(!is.na(m.aids.info))%>%
  ggplot(aes(m.aids.info, fill = risk)) +  # we remove those who do not remember and the NAs
  geom_bar() + ggtitle("Relationship m.aids.info/risk")

m.aids.info.plot2 <- preg.risk %>% filter(!is.na(m.aids.info)) %>%
```

```
  ggplot(aes(m.aids.info, fill = risk)) +  # we remove those who do not remember and the NAs
  geom_bar(position = "fill") + ggtitle("Relationship m.aids.info/risk (percentage)")

grid.arrange(m.pregnancy.info.plot1, m.pregnancy.info.plot2,
             m.aids.info.plot1, m.aids.info.plot2, nrow=2, ncol=2)
```




```
chisq.test(preg.risk$m.pregnancy.info, preg.risk$risk) # we can't reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$m.pregnancy.info and preg.risk$risk
FALSE X-squared = 1.3883, df = 1, p-value = 0.2387
```

```
chisq.test(preg.risk$m.aids.info, preg.risk$risk) # we reject the null
```

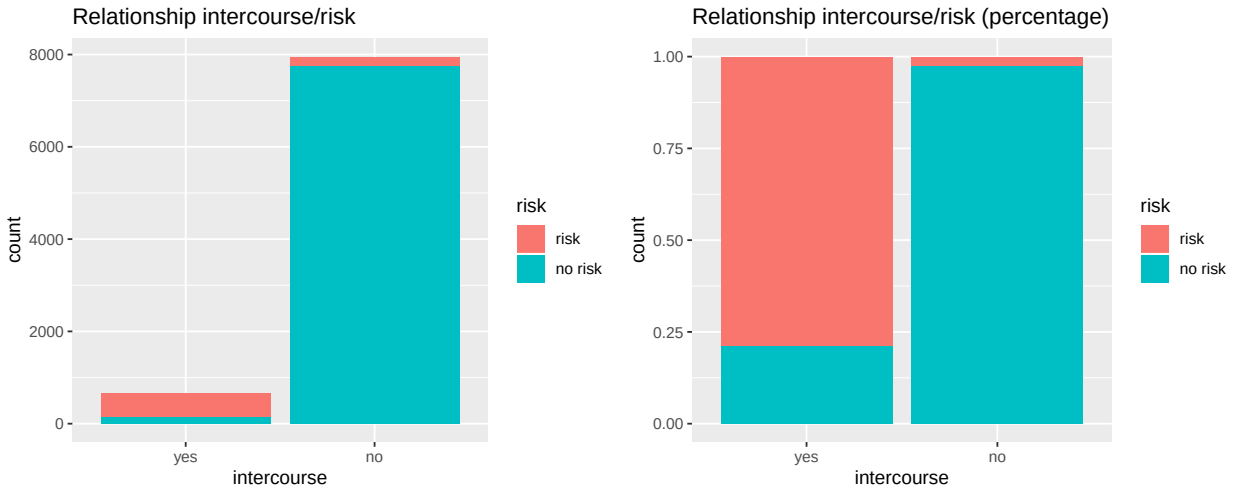
```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$m.aids.info and preg.risk$risk
FALSE X-squared = 4.6045, df = 1, p-value = 0.03189
```

Behavioral variables

- **intercourse**

This may seem obvious, but those who do not have a sexual life are not likely to be at risk of pregnancy.

```
intercourse.plot1 <- preg.risk %>% ggplot(aes(intercourse, fill = risk)) +
  geom_bar() + ggtitle("Relationship intercourse/risk")

intercourse.plot2 <- preg.risk %>% ggplot(aes(intercourse, fill = risk)) +
  geom_bar(position = "fill") + ggtitle("Relationship intercourse/risk (percentage)")

grid.arrange(intercourse.plot1, intercourse.plot2, ncol=2)
```




```
chisq.test(preg.risk$intercourse, preg.risk$risk) # we reject the null
```

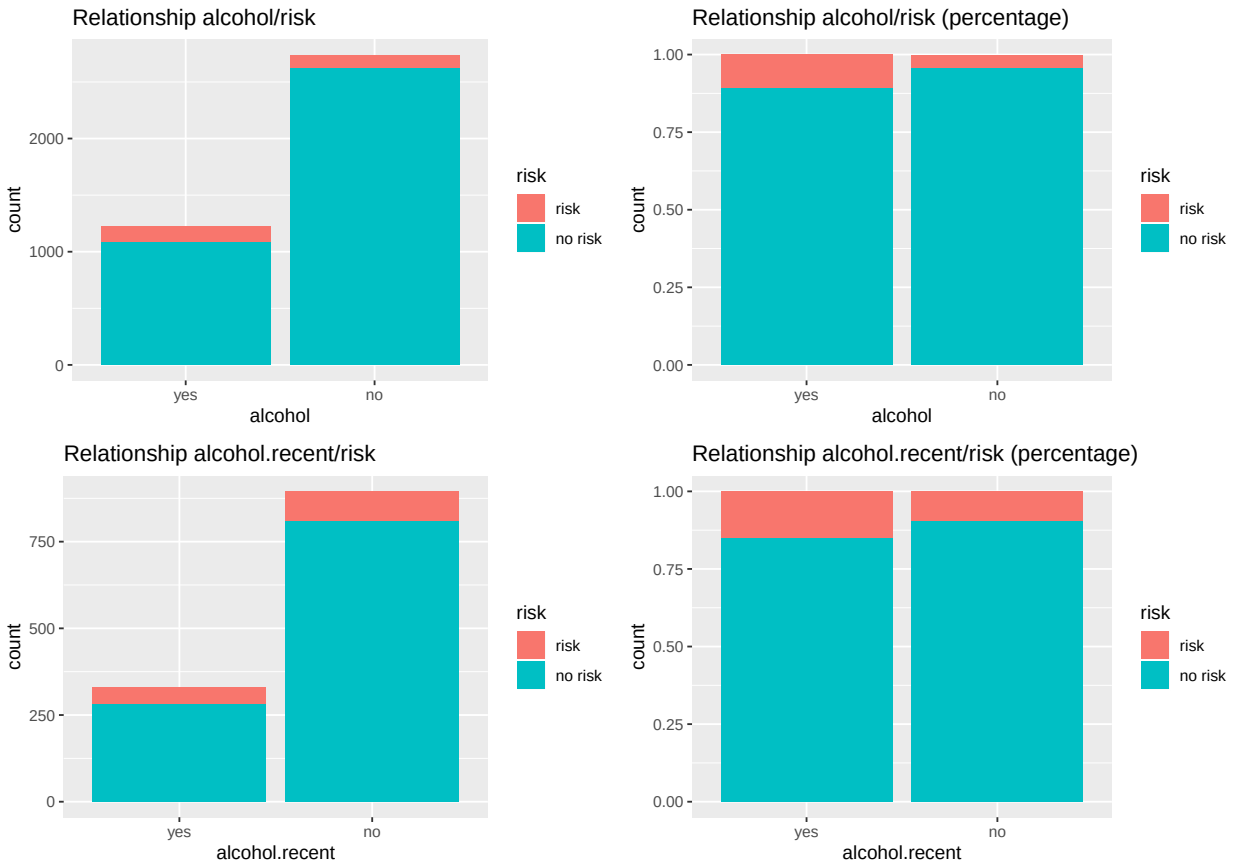
```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$intercourse and preg.risk$risk
FALSE X-squared = 4651, df = 1, p-value < 2.2e-16
```

- **alcohol and alcohol.recent**

This may not be surprising for some, but it is interesting to prove it statistically. Those who like drinking may like socializing more and may be more likely to have a partner and have intercourse.

```
alcohol.plot1 <- preg.risk %>% filter(!is.na(alcohol)) %>%
  ggplot(aes(alcohol, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar() + ggtitle("Relationship alcohol/risk")

alcohol.plot2 <- preg.risk %>% filter(!is.na(alcohol)) %>%
  ggplot(aes(alcohol, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar(position = "fill") + ggtitle("Relationship alcohol/risk (percentage)")
```

```r
alcohol.recent.plot1 <- preg.risk %>% filter(!is.na(alcohol.recent))%>%
  ggplot(aes(alcohol.recent, fill = risk)) +  # we remove those who do not remember and the NAs
  geom_bar() + ggtitle("Relationship alcohol.recent/risk")

alcohol.recent.plot2 <- preg.risk %>% filter(!is.na(alcohol.recent)) %>%
  ggplot(aes(alcohol.recent, fill = risk)) +  # we remove those who do not remember and the NAs
  geom_bar(position = "fill") + ggtitle("Relationship alcohol.recent/risk (percentage)")

grid.arrange(alcohol.plot1, alcohol.plot2,
             alcohol.recent.plot1, alcohol.recent.plot2, nrow=2, ncol=2)
```




```r
chisq.test(preg.risk$alcohol, preg.risk$risk) # we reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$alcohol and preg.risk$risk
FALSE X-squared = 63.949, df = 1, p-value = 1.277e-15
```

```r
chisq.test(preg.risk$alcohol.recent, preg.risk$risk) # we reject the null
```

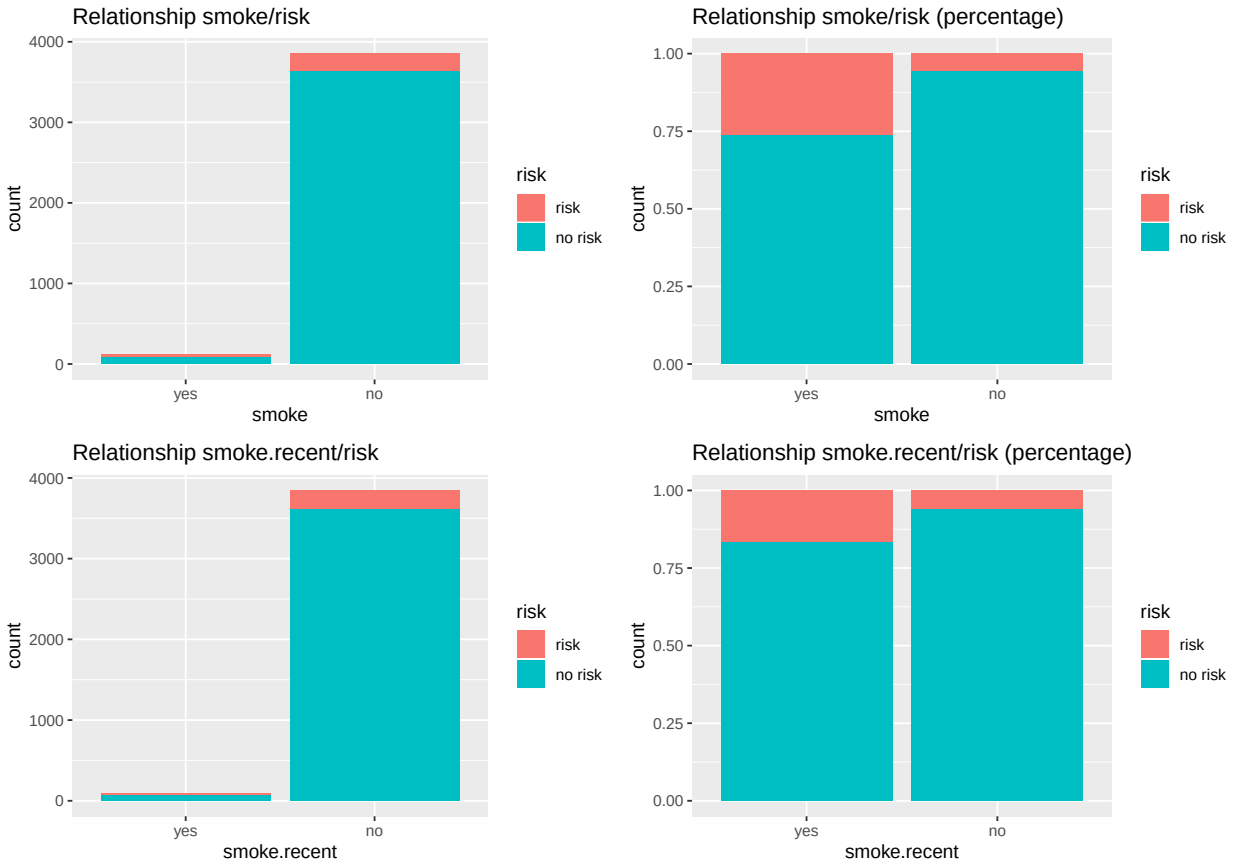
```
FALSE
```

```
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$alcohol.recent and preg.risk$risk
FALSE X-squared = 6.8809, df = 1, p-value = 0.008712
```

- **smoke and smoke.recent**

The findings here are very similar to those we found when analyzing the correlation between drinking alcohol and the risk of pregnancy.

```
smoke.plot1 <- preg.risk %>% filter(!is.na(smoke)) %>%
  ggplot(aes(smoke, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar() + ggtitle("Relationship smoke/risk")

smoke.plot2 <- preg.risk %>% filter(!is.na(smoke)) %>%
  ggplot(aes(smoke, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar(position = "fill") + ggtitle("Relationship smoke/risk (percentage)")

smoke.recent.plot1 <- preg.risk %>% filter(!is.na(smoke.recent))%>%
  ggplot(aes(smoke.recent, fill = risk)) +  # we remove those who do not remember and the NAs
  geom_bar() + ggtitle("Relationship smoke.recent/risk")

smoke.recent.plot2 <- preg.risk %>% filter(!is.na(smoke.recent)) %>%
  ggplot(aes(smoke.recent, fill = risk)) +  # we remove those who do not remember and the NAs
  geom_bar(position = "fill") + ggtitle("Relationship smoke.recent/risk (percentage)")

grid.arrange(smoke.plot1, smoke.plot2,
             smoke.recent.plot1, smoke.recent.plot2, nrow=2, ncol=2)
```




```
chisq.test(preg.risk$smoke, preg.risk$risk) # we reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$smoke and preg.risk$risk
FALSE X-squared = 77.627, df = 1, p-value < 2.2e-16
```

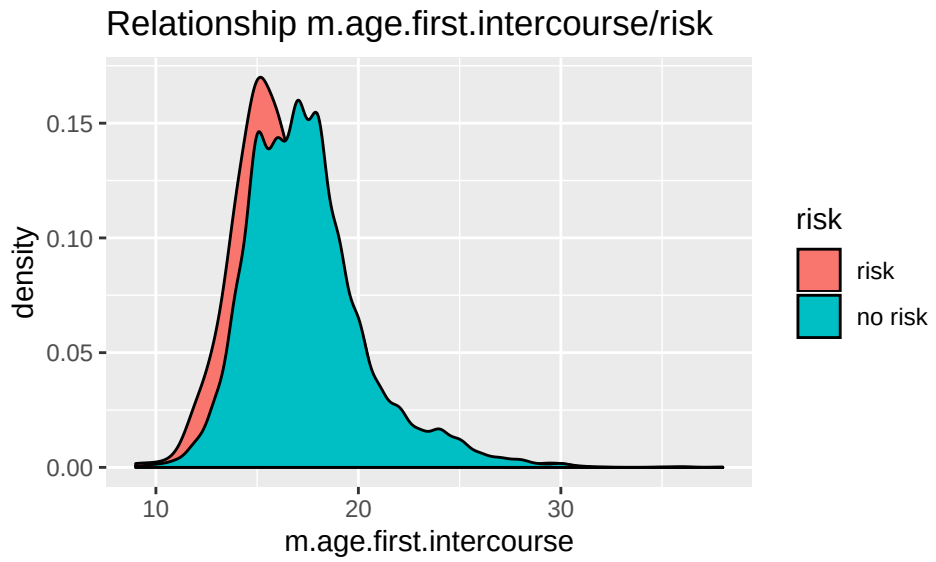
```
chisq.test(preg.risk$smoke.recent, preg.risk$risk) # we reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test with Yates' continuity correction
FALSE
FALSE data:  preg.risk$smoke.recent and preg.risk$risk
FALSE X-squared = 15.459, df = 1, p-value = 8.431e-05
```

- **m.age.first.intercourse**

This may be a striking fact for some of us. There is a correlation between early sexual initiation of the mothers and the risk of pregnancy of their daughters. Why is this?

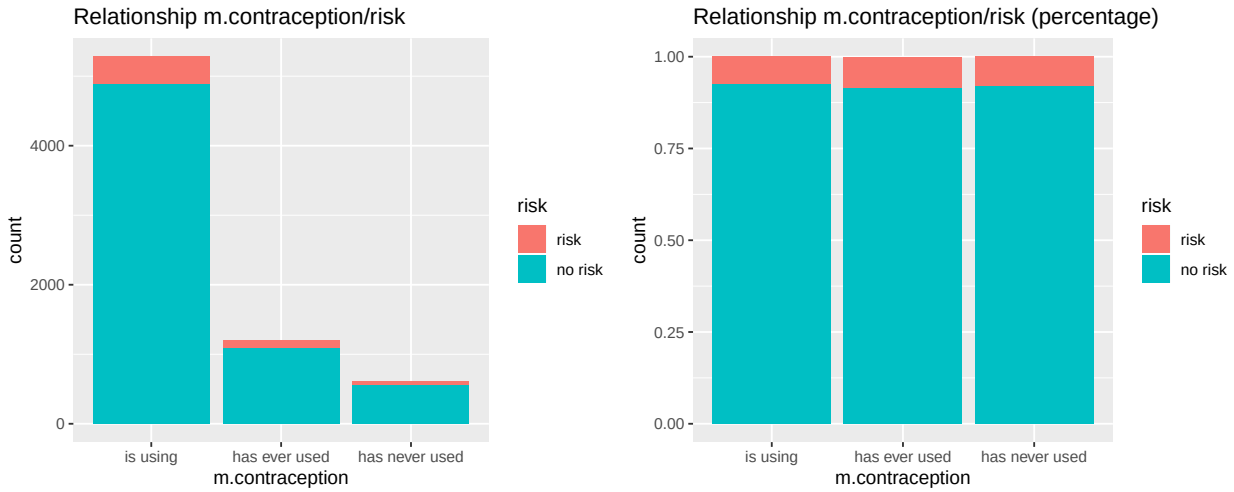
```
preg.risk %>% ggplot(aes(m.age.first.intercourse, fill = risk)) +
  geom_density() + ggtitle("Relationship m.age.first.intercourse/risk")
```



- **m.contraception**

There doesn't seem to be a correlation between the use of contraception of the mother and the risk of pregnancy of the daughter.

```
m.contraception.plot1 <- preg.risk %>% filter(!is.na(m.contraception)) %>%
  ggplot(aes(m.contraception, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar() + ggtitle("Relationship m.contraception/risk")

m.contraception.plot2 <- preg.risk %>% filter(!is.na(m.contraception)) %>%
  ggplot(aes(m.contraception, fill = risk)) + # we remove those who do not remember and the NAs
  geom_bar(position = "fill") + ggtitle("Relationship m.contraception/risk (percentage)")

grid.arrange(m.contraception.plot1, m.contraception.plot2, ncol=2)
```



```r
chisq.test(preg.risk$m.contraception, preg.risk$risk) # we can't reject the null
```

```
FALSE
FALSE   Pearson's Chi-squared test
FALSE
FALSE data:  preg.risk$m.contraception and preg.risk$risk
FALSE X-squared = 1.5082, df = 2, p-value = 0.4704
```

The algorithm

As mentioned before, we will apply a logistic regression model to predict the sexual behavior of teenagers. We need to select the predictors that can maximize sensitivity without sacrificing too much specificity. We care about specificity as much as we care about sensitivity, so we will look at balanced accuracy and not the overall accuracy to evaluate the efficacy of the algorithm. We want to achieve the highest sensitivity while keeping false positives as low as possible.

We also want to create an algorithm as practical as possible that can be easily applied in the real world. In practice, we will ask teenagers certain information that can help us predict their sexual behavior while disguising our goal from the subject. For this reason, we will not consider **intercourse**, **m.age.first.intercourse**, and **m.contraception** as predictors, as it may be difficult in practice to get this information. Additionally, if we manage to get this information, it may be biased as some tend to lie about their sexual behavior more than they do on other issues.

Partitioning the data

We will randomly split the data into two sets: the train set and the test set. Each set will contain 80% and 20% of the data, respectively.

```r
# We first create a partition of the data
set.seed(25)
test.index <- createDataPartition(preg.risk$risk, times = 1, p = 0.2, list = FALSE)

train.set <- preg.risk[-test.index,]
test.set <- preg.risk[test.index,]

# We need to transform the categorical variable into a numeric class variable
train.set <- mutate(train.set, y = as.numeric(risk == "risk"))
```

A first model

We have a lot of explanatory variables. Which should be included in the model? We need to check how well each variable can do on its own to increase the sensibility.

As we will only include one predictor in this first algorithm, it will be difficult for it to conclude a person at risk of pregnancy if we set the **p.hat** to be higher than 0.5 when applying the predict function on the test set. Therefore, we will set **p.hat** to be higher than 0.2 in this first algorithm.

```r
# we do not want to add intercourse, m.age.first.intercourse, and m.contraception as predictors
variables <- (9:37)[!(9:37 == 26 | 9:37 == 36 | 9:37 == 37)]
```

```
predictors <-
  sapply(variables, function(x1){
    train.set$x1 <- train.set[,x1]
    test.set$x1 <- test.set[,x1]
    glm.fit <- glm(y ~ x1, data = train.set, family = "binomial")
    p.hat.logit <- predict(glm.fit, newdata = test.set, type = "response")
    y.hat.logit <- ifelse(p.hat.logit > 0.20, "risk", "no risk") %>% factor(levels = c("risk", "no risk
    conf.matrix <- confusionMatrix(y.hat.logit, test.set$risk)
    return(unlist(data.frame(Sensitivity = conf.matrix$byClass[["Sensitivity"]],
                             Specificity = conf.matrix$byClass[["Specificity"]],
                             Balanced.Accuracy = conf.matrix$byClass[["Balanced Accuracy"]])))
  })

cbind(data.frame(Variable = colnames(test.set[variables])), t(as.data.frame(predictors)))
```

```
##                       Variable Sensitivity Specificity Balanced.Accuracy
## V1                         age   0.3120567   0.9203036         0.6161802
## V2                        area   0.0000000   1.0000000         0.5000000
## V3                   ethnicity   0.0000000   1.0000000         0.5000000
## V4                       m.age   0.0000000   1.0000000         0.5000000
## V5                 d.m.age.diff   0.0000000   1.0000000         0.5000000
## V6                 num.bedrooms   0.0000000   1.0000000         0.5000000
## V7                     internet   0.0000000   1.0000000         0.5000000
## V8                    cellphone   0.0000000   1.0000000         0.5000000
## V9                     transfer   0.0000000   1.0000000         0.5000000
## V10               f.live.house   0.0000000   1.0000000         0.5000000
## V11              attend.school   0.3191489   0.9196711         0.6194100
## V12         contraception.info   0.0000000   1.0000000         0.5000000
## V13 contraception.info.family   0.0000000   1.0000000         0.5000000
## V14 contraception.info.school   0.0000000   1.0000000         0.5000000
## V15                 period.info   0.0000000   1.0000000         0.5000000
## V16              pregnancy.info   0.0000000   1.0000000         0.5000000
## V17                   aids.info   0.0000000   1.0000000         0.5000000
## V18                     alcohol   0.0000000   1.0000000         0.5000000
## V19              alcohol.recent   0.0000000   1.0000000         0.5000000
## V20                       smoke   0.1636364   0.9774236         0.5705300
## V21                smoke.recent   0.0000000   1.0000000         0.5000000
## V22              m.num.children   0.0000000   1.0000000         0.5000000
## V23                       m.job   0.0000000   1.0000000         0.5000000
## V24                 m.education   0.0000000   1.0000000         0.5000000
## V25            m.pregnancy.info   0.0000000   1.0000000         0.5000000
## V26                 m.aids.info   0.0000000   1.0000000         0.5000000
```

The variables that manage to increase sensitivity above zero are **age**, **attend.school**, and **smoke**. We will use these variables to create a second model and try to find out how well they work together and what other variables we can include as well.

```
glm.fit.1 <- glm(y ~ age + attend.school + smoke, data = train.set, family = "binomial")
p.hat.logit.1 <- predict(glm.fit.1, newdata = test.set, type = "response")
y.hat.logit.1 <- ifelse(p.hat.logit.1 > 0.20, "risk", "no risk") %>% factor(levels = c("risk", "no risk
conf.matrix.1 <- confusionMatrix(y.hat.logit.1, test.set$risk)
glm.fit.1.sensitivity <- conf.matrix.1$byClass[["Sensitivity"]]
```

```r
glm.fit.1.accuracy <- conf.matrix.1$byClass[["Balanced Accuracy"]]
conf.matrix.1
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction risk no risk
##    risk      16      35
##    no risk   39     718
##
##                Accuracy : 0.9084
##                  95% CI : (0.8864, 0.9274)
##     No Information Rate : 0.9319
##     P-Value [Acc > NIR] : 0.9955
##
##                   Kappa : 0.253
##
##  Mcnemar's Test P-Value : 0.7273
##
##             Sensitivity : 0.29091
##             Specificity : 0.95352
##          Pos Pred Value : 0.31373
##          Neg Pred Value : 0.94848
##              Prevalence : 0.06807
##          Detection Rate : 0.01980
##    Detection Prevalence : 0.06312
##       Balanced Accuracy : 0.62221
##
##        'Positive' Class : risk
##
```

The three combined variables produce a sensitivity of 0.2909091 and a balanced accuracy of 0.6222142. Not bad! Let's see if we can improve that.

The final model

We need to run several models to see how much our algorithm can improve by adding a variable. Then, we can select the variables that best fit our model.

```r
variables <- (9:37)[!(9:37 == 26 | 9:37 == 36 | 9:37 == 37 | 9:37 == 9 | 9:37 == 19 | 9:37 == 29)]

predictors <-
  sapply(variables, function(x4){
    train.set$x4 <- train.set[,x4]
    test.set$x4 <- test.set[,x4]
    glm.fit <- glm(y ~ age + attend.school + smoke + x4, data = train.set, family = "binomial")
    p.hat.logit <- predict(glm.fit, newdata = test.set, type = "response")
    y.hat.logit <- ifelse(p.hat.logit > 0.20, "risk", "no risk") %>% factor(levels = c("risk", "no risk
    conf.matrix <- confusionMatrix(y.hat.logit, test.set$risk)
    data.frame(Sensitivity = conf.matrix$byClass[["Sensitivity"]],
               Specificity = conf.matrix$byClass[["Specificity"]],
               Balanced.Accuracy = conf.matrix$byClass[["Balanced Accuracy"]])
```

```
  })

cbind(data.frame(Variable = colnames(test.set[variables])), t(as.data.frame(predictors))) %>%
  mutate(Sen.Improv = as.numeric(Sensitivity) - glm.fit.1.sensitivity,
         Acc.Improv = as.numeric(Balanced.Accuracy) - glm.fit.1.accuracy)
```

```
##                     Variable Sensitivity Specificity Balanced.Accuracy
## 1                       area   0.2545455   0.9601594         0.6073524
## 2                  ethnicity   0.2909091   0.9548473         0.6228782
## 3                      m.age   0.2727273   0.9574468          0.615087
## 4                d.m.age.diff   0.2727273   0.9574468          0.615087
## 5                num.bedrooms   0.2727273   0.9601594         0.6164433
## 6                    internet   0.2909091   0.9561753         0.6235422
## 7                   cellphone   0.2909091   0.9561753         0.6235422
## 8                    transfer   0.2727273   0.9561753         0.6144513
## 9                f.live.house   0.2545455   0.9601594         0.6073524
## 10        contraception.info   0.2909091   0.9530831         0.6219961
## 11 contraception.info.family   0.2909091   0.9535193         0.6222142
## 12 contraception.info.school   0.2909091   0.9561753         0.6235422
## 13                period.info   0.3018868   0.9472868         0.6245868
## 14             pregnancy.info   0.2727273   0.9561753         0.6144513
## 15                  aids.info   0.3061224   0.9555556          0.630839
## 16                    alcohol   0.2727273   0.9535193         0.6131233
## 17             alcohol.recent         0.5   0.9026549         0.7013274
## 18               smoke.recent   0.2545455   0.9597855         0.6071655
## 19             m.num.children   0.2545455   0.9574468         0.6059961
## 20                      m.job   0.2909091    0.954727         0.6228181
## 21                m.education   0.2909091   0.9597315         0.6253203
## 22          m.pregnancy.info   0.2432432   0.9633333         0.6032883
## 23                m.aids.info   0.2941176   0.9650092         0.6295634
##      Sen.Improv    Acc.Improv
## 1  -0.036363636 -0.0148617651
## 2   0.000000000  0.0006640106
## 3  -0.018181818 -0.0071271330
## 4  -0.018181818 -0.0071271330
## 5  -0.018181818 -0.0057708560
## 6   0.000000000  0.0013280212
## 7   0.000000000  0.0013280212
## 8  -0.018181818 -0.0077628878
## 9  -0.036363636 -0.0148617651
## 10  0.000000000 -0.0002180732
## 11  0.000000000  0.0000000000
## 12  0.000000000  0.0013280212
## 13  0.010977702  0.0023726335
## 14 -0.018181818 -0.0077628878
## 15  0.015213358  0.0086248287
## 16 -0.018181818 -0.0090909091
## 17  0.209090909  0.0791132600
## 18 -0.036363636 -0.0150486849
## 19 -0.036363636 -0.0162180421
## 20  0.000000000  0.0006038872
## 21  0.000000000  0.0031061437
## 22 -0.047665848 -0.0189258853
```

```
## 23  0.003208556  0.0073492540
```

We found several variables that can improve our model substantially: **alcohol**, **d.m.age.diff**, **transfer**, **f.live.house**, **aids.info**, **period.info**. Let's add them to our model.

```
# we can see how well they work together
glm.fit.2 <- glm(y ~ age + attend.school + smoke + alcohol.recent + d.m.age.diff +
                   transfer + f.live.house + aids.info + period.info, data = train.set, family = "binom
p.hat.logit.2 <- predict(glm.fit.2, newdata = test.set, type = "response")
y.hat.logit.2 <- ifelse(p.hat.logit.2 > 0.20, "risk", "no risk") %>% factor(levels = c("risk", "no risk
conf.matrix.2 <- confusionMatrix(y.hat.logit.2, test.set$risk)
glm.fit.2.sensitivity <- conf.matrix.2$byClass[["Sensitivity"]]
glm.fit.2.accuracy <- conf.matrix.2$byClass[["Balanced Accuracy"]]
conf.matrix.2
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction risk no risk
##    risk      17      29
##    no risk   11     170
##
##                Accuracy : 0.8238
##                  95% CI : (0.7679, 0.871)
##     No Information Rate : 0.8767
##     P-Value [Acc > NIR] : 0.99203
##
##                   Kappa : 0.3616
##
##  Mcnemar's Test P-Value : 0.00719
##
##             Sensitivity : 0.60714
##             Specificity : 0.85427
##          Pos Pred Value : 0.36957
##          Neg Pred Value : 0.93923
##              Prevalence : 0.12335
##          Detection Rate : 0.07489
##    Detection Prevalence : 0.20264
##       Balanced Accuracy : 0.73071
##
##        'Positive' Class : risk
##
```

Now, we have a sensitivity of 0.6071429 and a balanced accuracy of 0.7307071. Not bad, knowing that the portion of teenagers at risk of pregnancy is really small.

Conclusions

We used a logistic regression model a the predicting method of our machine-learning algorithm. The final model consisted of nine predictors: **age**, **attend.school**, **smoke**, **alcohol**, **d.m.age.diff**, **transfer**, **f.live.house**, **aids.info**, **period.info**. We can accurately predict around 60.71% of the teenagers at risk of pregnancy, without sacrificing too much specificity. Of course, this algorithm may not be the most optimal

one. There may be more variables we did not consider that could have a significant impact on our prediction. However, this was an excellent start in the exploration of new machine-learning applications in social sciences.

One of the most significant limitations I faced in this project was the lack of data for all the variables we wanted to use as predictors. For example, in the case of **alcohol.recent** and **smoke**, we had a lot of missing information. This may be because the database is very recent, and the Ecuadorian government will continue publishing updates of it with the missing observations. As it is a health survey, it does not collect all the behavioral variables that may also be interesting for our model. Additionally, in this project, I only explored one machine-learning technique, logistic regression. However, there are many more techniques that could be applied and may lead to better results.

I want to thank the team at HarvardX, who created this data science program. The goal of this study was not only to serve as the capstone project, but also to be used outside in the field, to be able to improve decision making and, hopefully, improve the fate of the adolescents who every year have to become mothers. By no mean, I think this algorithm is flawless. In fact, I believe many things that can be done to improve it. I kindly appreciate any feedback that can help this model become better.

Appendix (Data wrangling)

In this section, we will describe the data wrangling process used to create the final dataset used to develop the algorithm. It is worth mentioning that this was the most time-consuming part of the whole project. I hope you appreciate the time put into this process.

```
##############################################################################
####                          1 Loading the data                          ####
##############################################################################

####                               WARNING                                ####

# During my work on this project, the government updated the URL to
# the databases several times. I have uploaded the datasets to my
# Github account. However, I have also left the URL of the statistics
# bureau website in case you want to check it. Thank you!

##############################################################################
####                        1.1 The packages                              ####
##############################################################################

if(!require(readstata13)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(gridExtra)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(plyr)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")

##############################################################################
####                     1.2 Downloading the data                         ####
##############################################################################

# We will download the files and save them on our system with the names: people, women, house, and beha
# 1) The people dataset will contain the demographic and economic data of each of the members of the
# household, contained in the "1_BDD_ENS2018_f1_personas.dta" file
# 2) The house dataset will contain the data about the house the household lives in, contained in the
```

```
# "2_BDD_ENS2018_f1_hogar.dta" file
# 3) The women dataset will contain the data about the sexual health of women from 10 to 49, contained
# "4_BDD_ENS2018_f2_mef.dta" file
# 4) The behavior dataset will contain the data about behavioral risk factors of people from 5 to 18,
# contained in the "8_BDD_ENS2018_f4_fact_riesgo.dta" file

# This is the original url. However, as we are not sure whether the url will be stil available by the t
# the assessment of this project. We wil use the Github url.
org.url <- "https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANU
# We give the url a name
url <- "https://github.com/aquijanoruiz/HarvardX_capstone/raw/master/SexRisk/BDD_ENSANUT_2018_STATA_.zi
# We create a temporary directory
td <- tempdir()
# We create the placeholder file
tf <- tempfile(tmpdir=td, fileext = ".zip")
# We download the data into the placeholder file
download.file(url,tf)

# We get the name of the file inside the zip file that contains the demographic and economic data,
# unzip it, get the full path name of it, and finally load it
people.f.name <- unzip(tf, list=TRUE)$Name[2] # The people dataset is the number 2
unzip(tf, files=people.f.name, exdir=td, overwrite=TRUE)
people.f.path <- file.path(td, people.f.name)
people <- read.dta13(people.f.path)

# Now, we need to do the same for the house, women, and behavior datasets
house.f.name <- unzip(tf, list=TRUE)$Name[3] # The house dataset is the number 3
women.f.name <- unzip(tf, list=TRUE)$Name[5] # The women dataset is the number 5
behavior.f.name <- unzip(tf, list=TRUE)$Name[9] # The behavior dataset is the number 9

unzip(tf, files=c(house.f.name, women.f.name, behavior.f.name), exdir=td, overwrite=TRUE)
house.f.path <- file.path(td, house.f.name)
women.f.path <- file.path(td, women.f.name)
behavior.f.path <- file.path(td, behavior.f.name)

# Now, we can load the three files
house <- read.dta13(house.f.path)
women <- read.dta13(women.f.path)
behavior <- read.dta13(behavior.f.path)

# As these are STATA files, the label of each of the variables is stored inside the datasets,
# we can extract them using the following code
data.key.people <- data.frame(variable = names(people),
                              label = attr(people,"var.labels"))

data.key.house <- data.frame(variable = names(house),
                             label = attr(house,"var.labels"))

data.key.women <- data.frame(variable = names(women),
                             label = attr(women,"var.labels"))

data.key.behavior <- data.frame(variable = names(behavior),
                             label = attr(behavior,"var.labels"))
```

```r
# If you know how to read Spanish, you can check what each variable means. Just use the View()
# fucntion like this: View(data.key.people)
# For example, let's look at the first 20 variables of the people set and their labels
head(data.key.people, 20) # Can you read Spanish?

# The name of each variable is assigned according to its code in the survey. For example,
# whether a woman between 12 and 49 has ever had  intercourse can be found in variable f2_s8_803
# of the women set, which means form 2, section 8, question 803
summary(women$f2_s8_803)

# We can see that the questions and the answers are all provided in Spanish. This should be
# no surprise as the official language in Ecuador is Spanish. For your better understanding,
# I will recode every variable of interest from Spanish into English

##############################################################################
####                        2 Data wrangling                              ####
####                 2.1 Extracting the mothers' Ids                      ####
##############################################################################

# We have loaded four set: people, house, women, and behavior. Our variables of interest
# are scattered in all these datasets. We need to create a singe dataset with all these
# variables. For now, we will create two datsets: the daughters and the mothers set.

# Who is whose mother? Daughters and mothers are in the same dataset in the people set, but
# as a different observarion. Let's extract the Ids of each person surveyed and the variables
# we will use to link the data.
people <- people %>% mutate(id.household = id_hogar, # Let's first change the names into English
                            id.subject = id_per,
                            persona = as.integer(persona),
                            sex = revalue(sexo, c("hombre"="male", "mujer"="female")),
                            mother.code = f1_s2_15_1,
                            father.code = f1_s2_14_1,
                            age = f1_s2_3_1)

people.id <- people %>% select(id.household, id.subject, persona, sex, age, mother.code, father.code)

head(people.id, 10)

# Now, let's look for example at people in household 010150000201011
people.id %>% filter(id.household == "010150000201011")

# We can see that there are six members, 1 is the father, 2 is the mother, and 3 to 6 are the
# children (all males...). As the numebrs 1, 2, 3, etc. are repeated all over the dataset
# and not unique Ids, we cannot work with this data set simply as it is. To be able to work with
# it, we need to add two columns to each observation with the fathers' and the mothers' unique Ids.
# We actually only need the mothers's unique Ids, but we will take both for explanatory purposes.

# We can write the code like this:
id.mothers <- people.id %>% group_by(id.household) %>% slice(mother.code) %>%
  distinct(persona, .keep_all = TRUE) %>% ungroup() %>%
  select(-mother.code) %>% mutate(id.mother = id.subject, mother.code = persona) %>%
  select(id.household, id.mother, mother.code)
```

```r
id.fathers <- people.id %>% group_by(id.household) %>% slice(father.code) %>%
  distinct(persona, .keep_all = TRUE) %>% ungroup() %>%
  select(-father.code) %>% mutate(id.father = id.subject, father.code = persona) %>%
  select(id.household, id.father, father.code)

# Combining the datasets:
people.id <- left_join(people.id, id.mothers, by = c("id.household" = "id.household",
                                                      "mother.code" = "mother.code")) %>%
             left_join(id.fathers, by = c("id.household" = "id.household",
                                          "father.code" = "father.code"))

# Now, let's look again at people in household 010150000201011. We have added the mothers' and the
# fathers' Ids.
people.id %>% filter(id.household == "010150000201011")

########################################################################
####               2.2 Combining all the datasets                   ####
########################################################################

# The dataset we will mostly work with is the women dataset. However, There are some variables in the w
# set that also appear in the people, house, and behavior sets. We will delete them so that we do not h
# repeated variables when combining the datasets. Note that you do not need to know now what each of th
# following variables represent.

women <- women %>% dplyr::rename(id.subject = id_per,
                                 id.household = id_hogar)

people <- people %>% select(-c("area", "prov", "upm", "id_viv", "persona", "fecha_anio",
                               "fecha_mes", "fecha_dia", "region", "etnia", "edadanios",
                               "gedad_anios", "nivins", "nbi_1", "nbi_2", "escolaridad",
                               "fexp", "estrato", "id_hogar"))

house <- house %>% select(-c("area", "prov", "upm", "id_viv", "fecha_anio", "fecha_mes",
                             "fecha_dia", "region", "fexp", "estrato")) %>%
                   dplyr::rename(id.household = id_hogar)

behavior <- behavior %>% select(-c("area", "prov", "upm", "id_viv", "id_hogar", "persona",
                                   "sexo", "region", "etnia", "edadanios", "gedad_anios",
                                   "nivins", "fexp", "estrato")) %>%
                         dplyr::rename(id.subject = id_per)

people.id <- people.id %>% select(-c("id.household", "persona", "sex", "age", "mother.code",
                                     "father.code"))

# We will called data to the commbined dataset. It is constructed by combining the women, people, house
# behavior, and people.id sets. Why people.id? Because we later need to create another dataset with the
# mothers' data, and we need to filter according the daughters included in our analysis.

data <- women %>% left_join(people %>% select(-id.household) , by = "id.subject") %>%
  left_join(people.id, by = "id.subject") %>% # we add id.mother and id.father
  left_join(house, by = "id.household") %>% # we add the data about the house
  left_join(behavior, by = "id.subject") # we add the behavioral variables
```

```r
dim(data) # Let's look at how many observarions and variables we have (a lot of variables...)
# At the end, we won't need most of them so we will delete them.

########################################################################
####              2.3 Creating the daughters set                  #####
########################################################################

# We have joined all the datasets. Now, we will filter only the women from 12 to 18
daughters <- women %>% left_join(people %>% select(-id.household) , by = "id.subject") %>%
  left_join(people.id, by = "id.subject") %>% # we add id.mother and id.father
  left_join(house, by = "id.household") %>% # we add the data about the house
  left_join(behavior, by = "id.subject") %>% # we add the behavioral variables
  filter(between(age,12,18))

# Note that this won't be the final set we will use. We still need to work a lot on it!
nrow(daughters) # We can look at how many subjects we will use to build the algorithm

########################################################################
####              2.4 Creating the mothers set                    #####
########################################################################

# We filter the data including only all those poeple who are the mothers of those subjects in our analy
mothers <- people %>% left_join(women %>% select(-id.household) , by = "id.subject") %>%
  left_join(people.id, by = "id.subject") %>% # we add id.mother and id.father
  left_join(house, by = "id.household") %>% # we add the data about the house
  left_join(behavior, by = "id.subject") %>%
  semi_join(daughters, by = c("id.subject" = "id.mother"))

nrow(mothers) # We can look at how many mothers there are

########################################################################
####                          3. The variables                    #####
########################################################################

# The next step is to open the questionnaires and check what variables we want to use to build the algo
# look at the code of the question we want to add, and translate it into English.

# We previosly explained how the variables' names were constructed. For example, the variable f2_s8_803
# form 2, section 8, question 803, and it referes to wether a woman has ever had intercourse. We will q
# rename each variable of interest from English into Spanish, without explaining each detail of the que
# If you want to know more about the census used to contruct these datasets, please go to the following
#https://www.ecuadorencifras.gob.ec/salud-salud-reproductiva-y-nutricion/ and check the methodology.

########################################################################
####                  3.1 The explanatory variables               #####
####                  3.1.1 Demographic variables                 #####
########################################################################

# age ------------------------------
summary(daughters$age)

# area -----------------------------
daughters$area <- factor(revalue(daughters$area, c("urbano" = "urban")),
```

```
                          levels = c("urban", "rural"))
summary(daughters$area)

# ethnicity ------------------------------
daughters$ethnicity <- daughters$f1_s2_9
levels(daughters$ethnicity) <- c("indigenous","african ecuadorian", "black", "mulatto",
                                 "montuvio", "mestizo", "white", "other")
summary(daughters$ethnicity)

# m.age and d.m.diff.age -----------------------------
# These are the age of the mother and the age difference between the mother and the daughter.
# We first create the m.age variable in the mothers set and then move it into the daughters set to comp
# the age difference between the mother and the daughter

m.age <- mutate(mothers, m.age = mothers$age) %>%
         select(id.subject, m.age)

daughters <- left_join(daughters, m.age, by = c("id.mother" = "id.subject")) %>%
  mutate(d.m.age.diff = m.age - age)

hist(daughters$m.age)
hist(daughters$d.m.age.diff)

# m.num.children -------------------------------
mothers$m.num.children <- mothers$f2_s2_217_4
hist(mothers$m.num.children)

# The question f2_s2_217_4 asks for all the children given birth by a woman (curently dead or alive,
# living with the household or not). However, not all women are included in the form 2. Hence, all the
# missing values. For the missing mothers, we approximate to the number of children living in the house
# which is given in form 1.
sum(is.na(mothers$m.num.children)) # We can look at the number of NAs

m.num.children.house <- people.id %>% group_by(id.mother) %>% dplyr::summarize(m.num.children.house = n
mothers <- mothers %>% left_join(m.num.children.house, by = c("id.subject" = "id.mother"))
mothers <- mothers %>% mutate(m.num.children = ifelse(is.na(m.num.children), m.num.children.house,
                                                      m.num.children))
sum(is.na(mothers$m.num.children))
hist(mothers$m.num.children) # Now we have no NAs

#############################################################################
####          3.1.2 Economic and social variables                      #####
#############################################################################

# num.bedrooms -------------------------------
# Number of bedrooms in the house
daughters$num.bedrooms <- daughters$f1_s1_8
hist(daughters$num.bedrooms)

# internet -------------------------------
# Whether there is internet access in the house
daughters$internet <- revalue(daughters$f1_s1_42, c("si"="yes"))
summary(daughters$internet)
```

```r
# cellphone ------------------------------
# Whether the daughter has an activated cellphone
daughters$cellphone <- revalue(daughters$f1_s2_23, c("si"="yes"))
summary(daughters$cellphone)

# transfer ------------------------------
# Whether someone in the household receives an economic transfer from the government, known in Ecuador
# as "bono de desarrollo humano"
transfer <- people %>% select(id.household, f1_s3_27) %>% filter(f1_s3_27 == "si") %>%
  dplyr::rename(transfer = f1_s3_27)

daughters <- daughters %>% left_join(transfer, by = "id.household")
daughters$transfer[which(is.na(daughters$transfer))] <- "no"
summary(daughters$transfer)

# m.job ------------------------------
# Whether the mother has a job or no
mothers$m.job <- mothers$f1_s3_1
levels(mothers$m.job) <- c("yes", "no")
summary(mothers$m.job)
summary(mothers$f1_s3_1)

# m.live.house ------------------------------
# Whether the mother lives in the house
daughters$m.live.house <- revalue(daughters$f1_s2_15, c("si" = "yes"))
summary(daughters$m.live.house)

# f.live.house ------------------------------
# Whether the father lives in the house
daughters$f.live.house <- revalue(daughters$f1_s2_14, c("si" = "yes"))
summary(daughters$f.live.house)

#############################################################################
####               3.1.3 Educational variables                       #####
#############################################################################

# attend.school ------------------------------
# Whether the daughter attends school
daughters$attend.school <- revalue(daughters$f1_s2_17, c("si"="yes"))
summary(daughters$attend.school)

table(daughters$attend.school, daughters$age) # a huge portion of people at age 18 do not go to school
# maybe it is because the already finished it

# m.education ------------------------------
# Mother's education attainment (none, primary school, secondary school, university)
mothers$m.education <- mothers$f1_s2_19_1
levels(mothers$m.education) <- c("none", "none", "none", "primary school", "primary school", "secondary
                                 "secondary school", "university", "university", "university")
summary(mothers$m.education)

# contraception.info ------------------------------
# Whether the daughter has ever received information about contraception
```

```
daughters$contraception.info <- factor(revalue(daughters$f2_s8_800f,c("si"="yes")),
                                       levels = c("yes","no"))
summary(daughters$contraception.info)

# contraception.info.family ------------------------------
# Whether the daughter has learned about contraception mainly from herfamily
daughters$contraception.info.family <- ifelse(daughters$f2_s8_801f == "familiar?", "yes", "no")
daughters$contraception.info.family[which(is.na(daughters$contraception.info.family))] <- "no"
daughters$contraception.info.family <- factor(daughters$contraception.info.family, levels = c("yes","no
summary(daughters$contraception.info.family)

# contraception.info.school ------------------------------
# Whether the daughter has learned about contraception mainly fromschool
daughters$contraception.info.school <- ifelse(daughters$f2_s8_801f == "escuela/colegio?", "yes", "no")
daughters$contraception.info.school[which(is.na(daughters$contraception.info.school))] <- "no"
daughters$contraception.info.school <- factor(daughters$contraception.info.school, levels = c("yes","no
summary(daughters$contraception.info.school)

# period.info ------------------------------
# Whether the daughter knew about menstruation when she had her first period
daughters$period.info <- factor(revalue(daughters$f2_s8_841,c("si"="yes",
                                                              "no sabe/ no responde"="doesn't know")),
                                levels = c("yes","no", "doesn't know"))
table(daughters$period.info)

# pregnancy.info ------------------------------
# Whether the daughter can correctly answer the question: can a women get pregnantat first intercourse?
# (We consider as "no" people who do not know the answer to this question)
daughters$pregnancy.info <- factor(revalue(daughters$f2_s8_845,c("si"="yes",
                                                                 "no sabe/ no responde" = "no")),
                                   levels = c("yes","no"))
summary(daughters$pregnancy.info)
# aids.info ------------------------------
# Whether the daughter can correctly answer the question: can AIDS be transmited throughhandshake?
daughters$aids.info <- factor(revalue(daughters$f2_s10_1011_1,
                                      c("si"="yes","no sabe/ no responde"="yes")),
                              levels = c("yes","no"))
table(daughters$aids.info)

# m.pregnancy.info ------------------------------
# Whether the mother can correctly answer the question: can a women get pregnantat first intercourse?
mothers$m.pregnancy.info <- factor(revalue(mothers$f2_s8_845,c("si"="yes",
                                                               "no sabe/ no responde" = "no")),
                                   levels = c("yes","no"))

summary(mothers$m.pregnancy.info)

# m.aids.info  ------------------------------
# Whether the daughter can correctly answer the question: can AIDS be transmited throughhandshake?
mothers$m.aids.info <- factor(revalue(mothers$f2_s10_1011_1,
                                      c("si"="yes","no sabe/ no responde"="yes")),
                              levels = c("yes","no"))
summary(mothers$m.aids.info)
```

```r
#########################################################################
####               3.1.4 Behavioral variables                      #####
#########################################################################

# alcohol  ------------------------------
# Whether the daughter has ever drunk alcohol
daughters$alcohol <- revalue(daughters$f4_s5_500, c("si" = "yes"))
summary(daughters$alcohol)

# alcohol.recent ------------------------------
# Whether the daughter has drunk alcohol in the past 30 days
daughters$alcohol.recent <- factor(ifelse(daughters$f4_s5_502 >=1, "yes", "no"),
                                   levels = c("yes", "no"))
summary(daughters$alcohol.recent)

# smoke ------------------------------
# Whether the daughter has ever smoked
daughters$smoke <- revalue(daughters$f4_s6_600, c("si" = "yes"))
summary(daughters$smoke)

# smoke.recent ------------------------------
# Whether the daughter has smoked in the past 30 days
daughters$smoke.recent <- revalue(daughters$f4_s6_611, c("si" = "yes"))
daughters$smoke.recent[which(daughters$smoke.recent == "no sabe/no responde")] <- NA
daughters$smoke.recent <- factor(daughters$smoke.recent, levels = c("yes", "no"))
summary(daughters$smoke.recent)

# m.age.first.intercourse ------------------------------
# Mother's age at first intercourse
mothers$m.age.first.intercourse <- ifelse(mothers$f2_s8_831 == 88 |
                                            mothers$f2_s8_831 == 77 |
                                            mothers$f2_s8_831 == 99, NA, mothers$f2_s8_831)
hist(mothers$m.age.first.intercourse)

# m.contraception ------------------------------
# Mother's use of contraception (is using, has ever used, has never used)
mothers$m.contraception <- mothers$f2_s6_604
levels(mothers$m.contraception) <- c("is using", "has never used", "has ever used")
summary(mothers$m.contraception)


#########################################################################
####               3.1.4 The outcome variable                      #####
#########################################################################

# The outcome variable will be the type of sexual behavior of the daughter, a categorical variable
# defined as: risky or not risky. To obtain the outcome variable we first need to analyze the behaviora
# characteristics of the daughter

# intercourse ------------------------------
# Whether the daughter has ever had intercourse
daughters$intercourse <- factor(revalue(daughters$f2_s8_803,c("si"="yes")),
                                   levels = c("yes","no"))
```

```r
summary(daughters$intercourse)

# contraception ------------------------------
# Daughter's use of contraception (is using, has ever used, has never used)
daughters$contraception <- daughters$f2_s6_604
levels(daughters$contraception) <- c("is using", "has never used", "has ever used")
daughters$contraception <- as.character(daughters$contraception)

# Many women who have never used contraception have never had done so because they have never had
# intercourse. However, some of them have never used contraception despite having had intercourse.
table(daughters$intercourse,daughters$contraception,
      dnn = c("intercourse", "contraception"))

# We want to differentiate those who have had intercourse but never used contraception and those who
#have neverused contraception because they have never needed to
daughters <- daughters %>% mutate(contraception = factor(ifelse(contraception == "has never used" & inte
                                                                "no sex", contraception),
                                                         levels = c("is using", "has ever used", "has ne
table(daughters$intercourse,daughters$contraception,
      dnn = c("intercourse", "contraception")) # no inconsistencies

# contraception.no.use.reason ----------------------------
# Reason why the daughter does not use contraception (wants to get pregnant, postnatal, no sex life,
# doesn't like it, afraid of side effects, had side effects, partner doesn't like it, feels embarrassed
# because of economic reasons, no knowledge about contraception, religious reasons, others, doesn't kno
daughters$contraception.no.use.reason <- daughters$f2_s6_613
levels(daughters$contraception.no.use.reason) <-
  c("wants to get pregnant", "postnatal", "no sex life", "because of age", "doesn't like it",
    "afraid of side effects", "had side effects", "partner doesn't like it", "feels embarrassed",
    "because of economic reasons", "no knowledge about contraception", "religious reasons",
    "others" , "doesn't know")
summary(daughters$contraception.no.use.reason)

# contraception.first ----------------------------
# Whether the daughter used contraception at first intercourse
daughters$contraception.first <- as.character(revalue(as.factor(daughters$f2_s8_808),
                                                      c("1"="yes","2"="no")))

# This is a conditional question. People who were asked this question were the ones who claimed having
# had intercourse. We have lots of NAs, some belong to those who have never had intercourse, and some
# to those who claimed having had intercourse but did not provide information about contraception
# We need to fix those NAs
daughters$contraception.first <- ifelse (daughters$intercourse == "no" &
                                         is.na(daughters$contraception.first),
                                         "no sex", daughters$contraception.first)
daughters$contraception.first <- factor(daughters$contraception.first,levels = c("yes","no","no sex"))
summary(daughters$contraception.first)

# contraception.last ----------------------------
# Whether the daughter used contraception at last intercourse
daughters$contraception.last <- as.character(revalue(as.factor(daughters$f2_s8_834),
                                                     c("si"="yes")))
# We do the same as we did with contraception.first. We change the NAs into "no sex"
```

```r
daughters$contraception.last <- ifelse (daughters$intercourse == "no" &
                                        is.na(daughters$contraception.last),
                                        "no sex", daughters$contraception.last)
daughters$contraception.last <- factor(daughters$contraception.last,levels = c("yes","no","no sex"))
summary(daughters$contraception.last)

# pregnant ------------------------------
# Whether the daughter has ever been pregnant
daughters$pregnant <- revalue(daughters$f2_s2_207,c("si"="yes"))
summary(daughters$pregnant)

#########################################################################
####                    4. Putting it all together              #####
####       4.1 Checking for inconsistencies and errors          #####
#########################################################################

# Now, we need to combine the two data sets (daughters and mothers set) into one. Before that, we
# need to check for inconsistencies and erros in the data

# Some women were not willing to provide information their sexual behavior. We need to remove their dat
sum(is.na(daughters$intercourse))

# We have missing contraception data of many of the women who claimed having had intercourse, and they
# were erroneously marked as not having had intercourse.
table(daughters$contraception.first,daughters$contraception.last,
      dnn = c("first","last"))

# Some women claimed not having had intercourse, but claimed being using contraception or having
# used contraception.
table(daughters$intercourse, daughters$contraception,
      dnn = c("intercourse", "contraception"))

# These is missing data about pregnancy.
sum(is.na(daughters$pregnant))

# A few declared being pregnant but never had sexual intercourse
table(daughters$intercourse, daughters$pregnant,
      dnn = c("intercourse", "pregnant"))

#########################################################################
####          4.2 Removing inconsistencies and errors           #####
#########################################################################

# Now, we need to filter the data to keep only what we want:

daughters <- daughters %>%
  filter(m.live.house == "yes") %>% # we select the daughters whose mother lives in the house
  filter(!is.na(intercourse)) %>% # we remove missing data about intercourse
  # we remove those who have had intercourse but do not want to provide information whether they used
  # contraception at first or last intercourse
  filter(!(intercourse == "yes" & is.na(contraception.first))) %>%
  filter(!(intercourse == "yes" & is.na(contraception.last))) %>%
  # we also remove other inconsistencies such as those who claimed not having had intercourse
```

```r
  # but claimed being using contraception or having used contraception
  filter(!(intercourse == "no" & contraception == "is using")) %>%
  filter(!(intercourse == "no" & contraception == "has ever used")) %>%
  # we remove data of those who claimed having been pregnant but never had sexual intercourse
  filter(!(intercourse == "no" & pregnant == "yes")) %>%
  filter(!is.na(pregnant)) # we remove missing data about pregnancy

# We can check now on the data:
sum(is.na(daughters$intercourse)) # no NAs

table(daughters$contraception.first,daughters$contraception.last,
      dnn = c("first","last")) # no inconsistencies

table(daughters$intercourse, daughters$contraception,
      dnn = c("intercourse", "contraception")) # no inconsistencies

sum(is.na(daughters$pregnant)) # no NAs

table(daughters$intercourse, daughters$pregnant,
      dnn = c("intercourse", "pregnant")) # no inconsistencies

#############################################################################
####              4.3 Keeping only what we want                  #####
#############################################################################

# daughters data ------------------------------
daughters <- daughters %>%
  select(id.subject, id.household, id.mother, contraception, contraception.no.use.reason,
         contraception.first, contraception.last, pregnant, age, area, ethnicity, m.age,
         d.m.age.diff, num.bedrooms, internet, cellphone, transfer, f.live.house, attend.school,
         contraception.info, contraception.info.family, contraception.info.school,
         period.info, pregnancy.info, aids.info, intercourse, alcohol, alcohol.recent,
         smoke, smoke.recent)

# mothers data -------------------------------
mothers <- mothers %>%
  select(id.subject, m.num.children, m.job, m.education, m.pregnancy.info, m.aids.info,
         m.age.first.intercourse, m.contraception)

#############################################################################
####                4.4 Putting it all together                  #####
#############################################################################

preg.risk <- daughters %>% left_join(mothers, by = c("id.mother" = "id.subject"))

#############################################################################
####             5 Constructing the outcome variable             #####
#############################################################################

# We want to predict if the woman is at risk of pregnancy or is already pregnant

# 1. The woman has ever been pregnant:risk
# 2. The woman has ever had intercourse and never used contraception:risk
```

```r
# 3. The woman did not use contraception at first intercourse:risk
# 4. The woman did not use contraception at last intercourse:risk
# 5. The woman wouldn't use contraception because does not like it:risk
# 6. The woman wouldn't use contraception because is afraid of side effects:risk
# 7. The woman wouldn't use contraception because the partner doesn't like it:risk
# 8. The woman wouldn't use contraception because she feels embarrassed:risk
# 9. The woman wouldn't use contraception because of economic reasons:risk
# 10. The wouldn't use contraception because she has no knowledge about contraception:risk
# 11. The woman wouldn't use contraception because of other reasons or simply does not know why:risk
# 12. The rest:no risk

# We can observe why women wouldn't be willing to use contraception
table(preg.risk$intercourse, preg.risk$contraception.no.use.reason,
      dnn = c("has had inercourse", "reason why wouldn't use contraception"))

# People who used contraception at first and last intercourse and claim to be currently using
# contraception are considered not at risk
table(preg.risk$contraception.first, preg.risk$contraception.first,
      dnn = c("used contraception at first intercourse", "used contraception at last intercourse"))

preg.risk <- preg.risk %>%
  mutate(risk = ifelse(pregnant == "yes", "risk", # 1
                       ifelse(contraception == "has never used", "risk", # 2
                              ifelse(contraception.first == "no", "risk", # 3
                                     ifelse(contraception.last == "no", "risk", # 4
                                            ifelse(contraception.no.use.reason == "doesn't like it" | #
                                                     contraception.no.use.reason == "afraid of side eff
                                                     contraception.no.use.reason == "partner doesn't li
                                                     contraception.no.use.reason == "feels embarrassed"
                                                     contraception.no.use.reason == "because of economi
                                                     contraception.no.use.reason == "no knowledge about
                                                     contraception.no.use.reason == "others" | # 11
                                                     contraception.no.use.reason == "doesn't know", "ri

preg.risk$risk[which(is.na(preg.risk$risk))] <- "no risk" # we change the NAs into "no risk"
preg.risk$risk <- factor(preg.risk$risk, levels = c("risk", "no risk"))
summary(preg.risk$risk)
```

References

- El Comercio. (2017, October 17). Fact checking a las declaraciones de ???Con mis hijos no te metas??? en Ecuador. Retrieved from El Comercio: https://www.elcomercio.com/tendencias/factchecking-marcha-genero-conmishijosnote
- Irizarry, R. (2019). Introduction to Data Science: Data Analysis and Prediction Algorithms with R
- Minsiterio de Salud P??blica del Ecuador. (2018, June). La Pol??tica Intersectorial de Prevenci??n del Embarazo en Ni??as y Adolescentes 2018???2025.
- Pan American Health Organization. (2016). Accelerating progress toward the reduction of adolescent pregnancy in Latin America and the Caribbean. Washington D.C.: Pan American Health Organization, United Nations Population Fund, and United Nations Children???s Fund.