Binary answers to OR-questions

Hypothesis

There is quite a common pattern when one person ask another one a question like “Would you like A or B” and gets an answer “Yes” or “No”; the issue is that it’s usually unclear which argument was meant by the answering person. Thus, the quesion is: can we predict which was the chosen argument? What does this choice depend on?
We can suppose that there are some factors influencing the answer, which may be:
- gender of the respondent;
- positiveness / negativeness of the answer;
- the existance of some context before the question which is connected to one of the arguments.

Gender has been chosen due to life experience; it seemed like male and female mean different arguments by their “Yes” answer.
Positiveness has been chosen due to the same reason; seemed like people tend to answer “Yes” meaning the first argument.
The context has been chosen because it seemed to be expected for one to answer “Yes” meaning that argument which was mentioned in context; so if we have a situation like “It’s now the FIFA World Cup which takes place in Russia; since childhood you like to play football. You’re currently returning after watching a football match with your best friend. He suggests to meet the next day: ‘Let’s meet and play something! Would you like to play football or volleyball?’” it won’t matter if he asked “Would you like to play volleyball or football”, the answer still would be “Yes”.

Assumptions :
There are cases when one has to choose from two options and one is much more favorable for the respondent; in this case if that argument is in a comfortable-to-define-by-yes position (let’s assume it is the first position), the answer will be “Yes” meaning something like “Yes, the first option”; otherwise it will be “No” meaning “No, not the second option”.
It’s also worth mentioning that even though sometimes the answer mean “Any of these” or “None of this”, we won’t research such answers, we’re currenlty interested in answers referring to one of the arguments.

Now let’s formulate some hypotheses:
- the existence of context makes people mean the mentioned-in-context argument by the positive answer no matter which place it occupies more often;
- positive answer more often refers to the first argument;
- male and female tend to make emphasis on different arguments of the question no matter if the answer is positive or negative;
- the chosen factors are statistically significant so that they can be used to predict the argument emphasized by the respondent.

Research design

Testing of the mentioned hypotheses requires some data which has to be collected manually. The process will be described further. The exemplary form of the collected data can be found in Table 1.

Table 1. Exemplary form of the collected data.

reviewer	gender	question	answer	chosen_position	context_provided
1	m	1	Y	1	N
1	m	2	N	2	Y
2	f	1	Y	2	N

Table 1 description:
- reviewer — number of the respondent;
- gender — gender of the respondernt, female or male;
- question — question number;
- answer — the answer of the respondent, Y stands for “Yes, A(B)”, N stands for “No, not B(A)”;
- chosen_position — the argument to be chosen;
- context_provided — if there is some context before the question.

In order to test the first three hypotheses either Pearson’s Chi-squared Test or Fisher’s Exact Test will be used.

The fourth hypotheses will be tested using several models:
- logistic regression;
- decision tree;
- random forest.

Data collection

To collect the data to run the models based on a google form with questions will be made. The gender of the reviewer will be asked firstly and then there will be questions of this kind:
- Your friend asked you: “What’d you like on your sandwich: cheese or ham?” Assuming you wanted cheese, how would you answer: “Yes” (meaning “Yes, cheese”) or “No” (meaning “No, not ham”)?
Or this one:
- You’ve discussed the recent news about your favourite football team. Then a friend yours asked: “Would you like to play football or volleyball?” Assuming you wanted to play football, how would you answer: “Yes” (meaning “Yes, football”) or “No” (meaning “No, not volleyball”)?

This form of questions has been chosen instead of "… Assuming that you answered 'Yes', which one did you mean?" so that personal preferences won't affect the results.

There will be 8 questions asked — 2 questions for each combination of the researched factors. The questions can be seen in Table 2; the factors' values — Table 3.

Table 2. Questions asked.

Question	Text	Answer 1	Answer 2
1	Ваш одногруппник спрашивает вас: "Пойдем на пару или прогуляем?" С учетом того, что вы бы хотели посетить пару, как бы вы ответили?	"Да" (в смысле "Да, пойдем на пару")	"Нет" (в смысле "Нет, не будем прогуливать")
2	Ваш друг интересуется у вас как-то вечером: "Давай посмотрим что-нибудь. Ты хочешь глянуть какую-нибудь драму или комедию?" Как бы вы ответили, если бы хотели посмотреть комедию?	"Да" (в смысле "Да, комедию")	"Нет" (в смысле "Нет, не драму")
3	Самоизоляция идет уже который месяц, заказывать еду надоедает. Теперь уже домашняя еда становится желаннее. И вот в один вечер родня интересуется у вас: "Закажем еды или приготовим сами?" Как бы вы ответили, чтобы дать понять, что хотите домашней еды?	"Да" (в смысле "Да, приготовим")	"Нет" (в смысле "Нет, не будем заказывать")
4	С детства вы ужасно хотели завести собаку, однако родители были против. Прошло много лет, и одним вечером партнер внезапно спрашивает у вас: "Давай заведем животное! Хочешь кошку или собаку?" Как вы ответите с учетом того, что мечтаете именно о собаке?	"Да" (в смысле "Да, собаку")	"Нет" (в смысле "Нет, не кошку")
5	Ваш одногруппник, с которым вы обыкновенно вместе делаете домашки, спрашивает вас: "Сделаем математику до дедлайна или забьём и сдадим после срока?" С учетом того, что вы решили взяться за голову и больше не просрочивать дедлайны и хотите заняться математикой, как бы вы ответили?	"Да" (в смысле "Да, сделаем и сдадим до дедлайна")	"Нет" (в смысле "Нет, не будем забивать")
6	Вы с друзьями поехали на море, и вот настал момент, ради которого вы ходили на занятия по волейболу — друзья спрашивают вас: "Сыграем в волейбол или пойдем поплаваем в море?" С учетом того, что вы бы очень хотели показать им, как вы хороши в волейболе, как бы вы ответили?	"Да" (в смысле "Да, давайте в волейбол")	"Нет" (в смысле "Нет, не пойдем плавать")
7	Представим, что в детстве вас бы спросили родители: "Ты хочешь пойти учиться играть на пианино или на скрипке?" Как бы вы ответили, если бы хотели учиться играть на скрипке?	"Да" (в смысле "Да, я бы хотел играть на скрипке")	"Нет" (в смысле "Нет, мне бы не хотелось учиться играть на пианино")
8	Вам нравится заполнять различные тесты и опросники, и ваши друзья часто просят вам помочь им, ответив на вопросы, чтобы они могли сдавать курсовые работы и тд. И вот в очередной раз вам пишет друг и просит: "Ты сейчас занят или можешь помочь мне и пройдешь небольшой опросник?" С учетом того, что вы не особо против пройти небольшой тест, как бы вы ответили?	"Да" (в смысле "Да, я пройду")	"Нет" (в смысле "Нет, я не занят сейчас")

Table 3. Questions and factor values.

Context exists?	Argument to choose	Questions
No	A	1,5
No	B	2,7
Yes	A	3,6
Yes	B	4,8

Collected data

A total of 40 people have completed the form:

```
library(lme4)
library(tidyverse)
library(ggplot2)
library(rpart)
library(rpart.plot)
library(randomForest)
```

```
form_results <- read_csv2("https://raw.githubusercontent.com/lwahomura/LingData-CourseWork/master/Form%20Results2.csv")
```

```
form_results
```

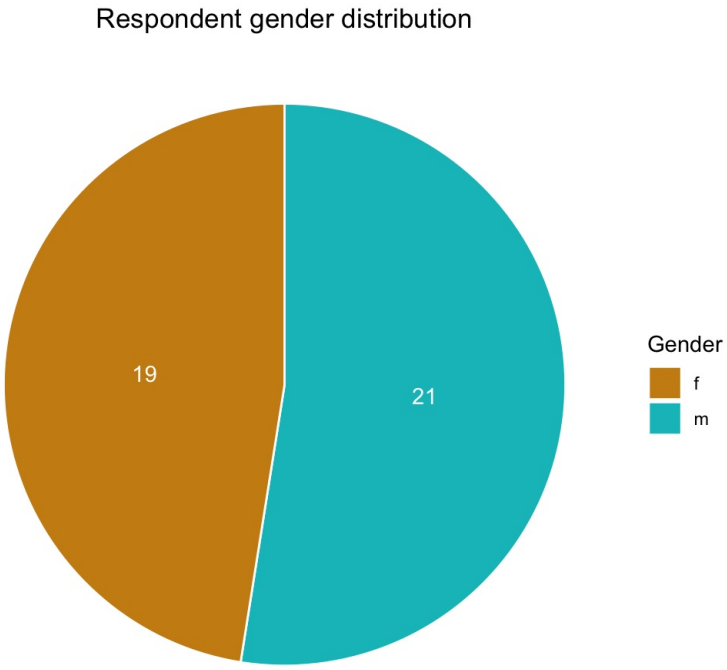
```
## # A tibble: 320 x 6
##    Respondent Respondent_gend… Question Context_exists Answer_positive…
##         <dbl> <chr>                <dbl>          <dbl>            <dbl>
##  1          1 f                        1              0                0
##  2          1 f                        2              0                1
##  3          1 f                        3              1                1
##  4          1 f                        4              1                1
##  5          1 f                        5              0                0
##  6          1 f                        6              1                0
##  7          1 f                        7              0                1
##  8          1 f                        8              1                1
##  9          2 f                        1              0                1
## 10          2 f                        2              0                0
## # … with 310 more rows, and 1 more variable: Chosen_argument <chr>
```

As we can see, the distibution of respondents' genders is nearly 50:50 :

```
respondents <- data.frame(table(distinct(form_results[,1:2])$Respondent_gender))

main_colours <- c("#cc8c12", "#00bfc4")

ggplot(respondents, aes(x="", y=Freq, fill = Var1)) +
  geom_bar(width = 1, stat = "identity", color = "white")+
  coord_polar("y", start=0) +
  geom_text(aes(label = Freq), position = position_stack(vjust = 0.5), color = "white") +
  theme_void() +
  labs(fill = "Gender", title = "Respondent gender distribution") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_manual(values=main_colours)
```



Data analysis

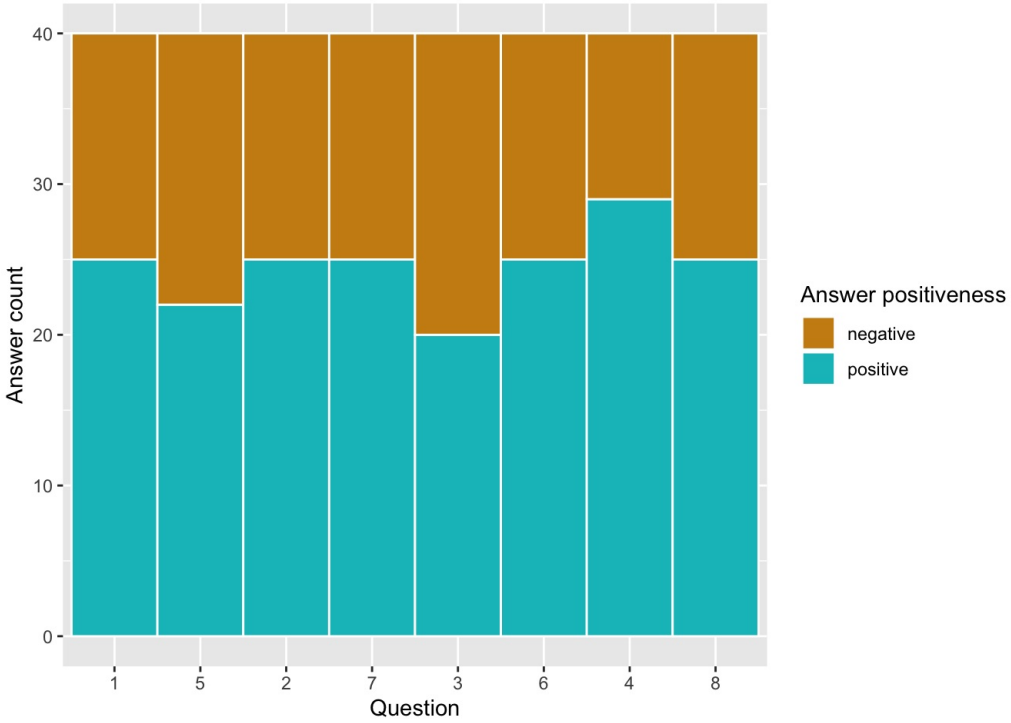
In order to work with data we should at first transform the collected data so that all the columns we'll work with will have only numeric values; we also need a column with the empasized argument which is calculated in this way: if the respondent needs to choose argument A and answers positively, argument A has been emphasized; if negatively — argument B (the same for choosing argument B).

```
form_results_trans <- mutate(form_results, Gender_binary = 1 * (Respondent_gender == 'm'))
form_results_trans <-  mutate(form_results_trans, Argument_binary = 1 * (Chosen_argument == 'B'))
form_results_trans <- mutate(form_results_trans, Emphasized_argument = (Answer_positiveness == Argument_binary) *
1 )
form_results_trans
```

```
## # A tibble: 320 x 9
##    Respondent Respondent_gend… Question Context_exists Answer_positive…
##         <dbl> <chr>                <dbl>          <dbl>             <dbl>
##  1          1 f                        1              0                 0
##  2          1 f                        2              0                 1
##  3          1 f                        3              1                 1
##  4          1 f                        4              1                 1
##  5          1 f                        5              0                 0
##  6          1 f                        6              1                 0
##  7          1 f                        7              0                 1
##  8          1 f                        8              1                 1
##  9          2 f                        1              0                 1
## 10          2 f                        2              0                 0
## # … with 310 more rows, and 4 more variables: Chosen_argument <chr>,
## #   Gender_binary <dbl>, Argument_binary <dbl>, Emphasized_argument <dbl>
```

Let's see how've the people answered the questions; as we have two questions for each combination of the factors the bars will be oreder by that combination.

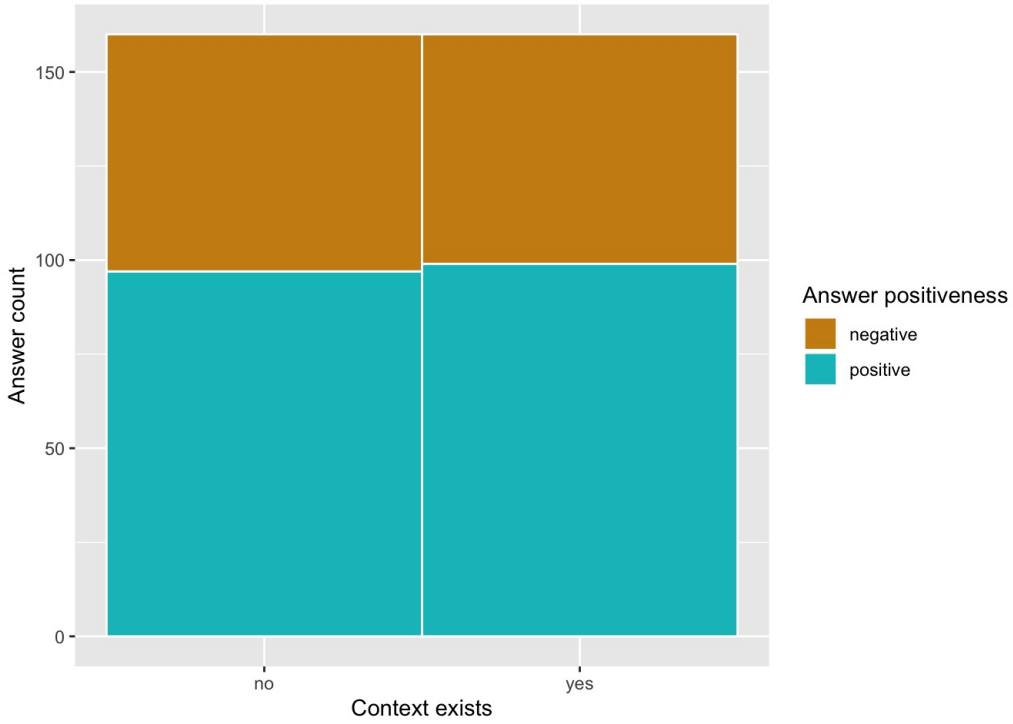
```
answers_pos <- form_results_trans %>% group_by(Context_exists, Argument_binary, Question, Answer_positiveness) %>%
summarise(ans_count = n())
answers_pos <- mutate(answers_pos, Q_order = 2 * Context_exists + Argument_binary)
ggplot(answers_pos, aes(x = reorder(factor(Question), Q_order), y=ans_count, fill = factor(Answer_positiveness)))
+
  geom_bar(width = 1, stat = "identity", color = "white") +
  labs(x = "Question", y = "Answer count") +
  scale_fill_manual(name = "Answer positiveness", labels = c("negative", "positive"), values=main_colours)
```



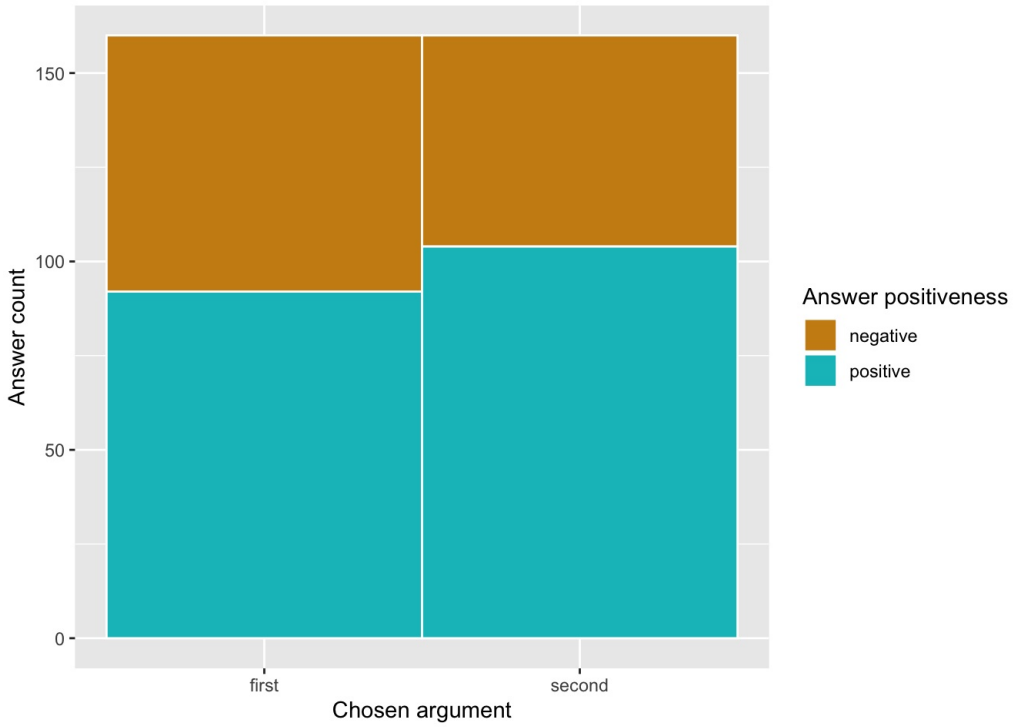
As we can see, people tend to answer "Yes" more often no matter which argument they mean — first or second.

Now let's see how the context existance and the meant argument effect the results:

```
answers_con <- form_results_trans %>% group_by(Context_exists, Answer_positiveness) %>% summarise(ans_count = n())
ggplot(answers_con, aes(x = factor(Context_exists), y=ans_count, fill = factor(Answer_positiveness))) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  labs(y = "Answer count") +
  scale_x_discrete(name = "Context exists", labels = c("no", "yes")) +
  scale_fill_manual(values=main_colours, name = "Answer positiveness", labels = c("negative", "positive"))
```



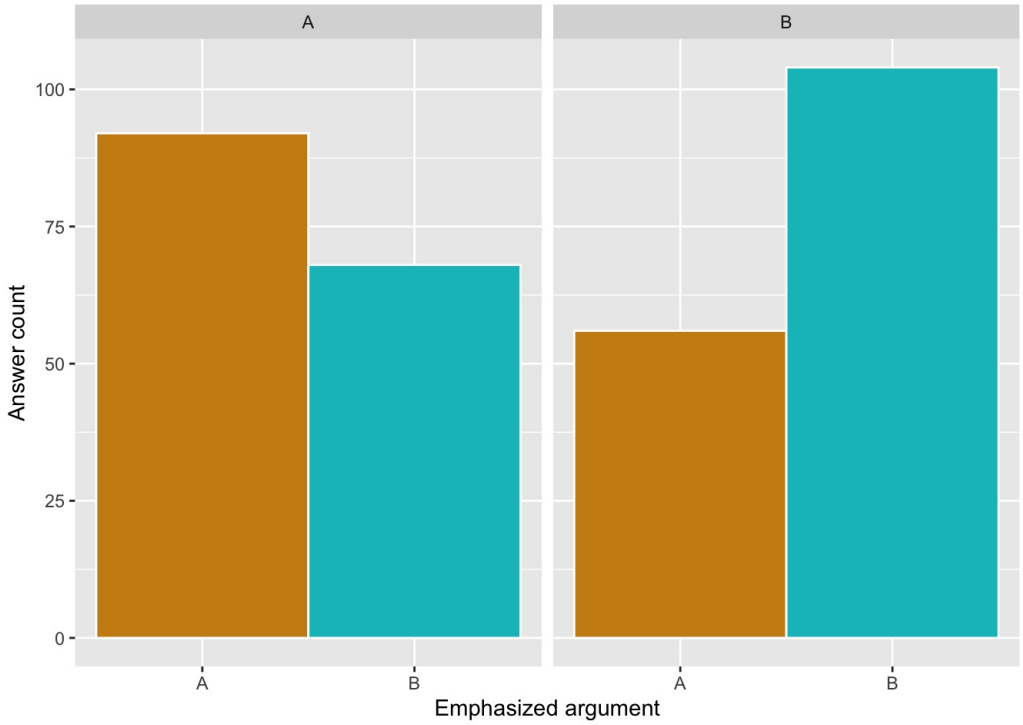

```
answers_arg <- form_results_trans %>% group_by(Chosen_argument, Answer_positiveness) %>% summarise(ans_count = n()
)
ggplot(answers_arg, aes(x = factor(Chosen_argument), y=ans_count, fill = factor(Answer_positiveness))) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  labs(y = "Answer count") +
  scale_x_discrete(name = "Chosen argument", labels = c("first", "second")) +
  scale_fill_manual(name = "Answer positiveness", labels = c("negative", "positive"), values=main_colours)
```



As we can see, people tend to answer “Yes” in most cases.

Let’s also inspect how the chosen argument is presented:

```
answers_emp <- form_results_trans %>% group_by(Emphasized_argument, Chosen_argument) %>% summarise(ans_count = n()
)
ggplot(answers_emp, aes(x = factor(Emphasized_argument), y=ans_count, fill=factor(Emphasized_argument))) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  labs(y = "Answer count") +
  scale_x_discrete(name = "Emphasized argument", labels = c("A", "B")) +
  facet_grid(cols = vars(Chosen_argument)) +
  scale_fill_manual(values=main_colours) +
  theme(legend.position = "none")
```



As we can see, people more often put emphasis on argument A to choose it; for choosing argument B we can see the opposite situation.

The initial data analysis shows that people usually answer positively to the OR-questions; though we should test the hypothesis by more accurate methods than visualization.

Testing hypotheses

First hypothesis

The existence of context makes people mean the mentioned-in-context argument by the positive answer no matter which place it occupies more often

The first hypothesis states that if any context've been provided before the question, people will emphasize the mentioned-in-context argument more often that another one.

```
first_hyp_table <- table(context = form_results_trans$Context_exists, arg = form_results_trans$Emphasized_argument
)
first_hyp_table
```

```
##        arg
## context  0  1
##       0 77 83
##       1 71 89
```

We can use Pearson's Chi-squared Test because the limitations of its usage are not violated; the null hypothesis is that two factors are independent.

```
first_hyp_test <- chisq.test(first_hyp_table)
first_hyp_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  first_hyp_table
## X-squared = 0.31427, df = 1, p-value = 0.5751
```

```
first_hyp_odds_ratio <- (first_hyp_test$observed[1,1] / first_hyp_test$observed[2,1]) / (first_hyp_test$observed[1
,2] / first_hyp_test$observed[2,2])
first_hyp_odds_ratio
```

```
## [1] 1.162905
```

The p-value is greater than 0.05, so we do not reject the null hypothesis. The existence of context and the emphasized argument seem to be independent; the odds ration tells us that the existence of context make people emphasize the second argument only 1.16 times more often than expected. Thus, our first hypothesis is incorrect.

Positive answer more often refers to the first argument

```
second_hyp_table <- table(positiveness = form_results_trans$Answer_positiveness, arg = form_results_trans$Chosen_argument)
second_hyp_table
```

```
##             arg
## positiveness   A   B
##            0  68  56
##            1  92 104
```

We can use Pearson's Chi-squared Test because the limitations of its usage are not violated; the null hypothesis is that two factors are independent.

```
second_hyp_test <- chisq.test(second_hyp_table)
second_hyp_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  second_hyp_table
## X-squared = 1.5932, df = 1, p-value = 0.2069
```

```
second_hyp_odds_ratio <- (second_hyp_test$observed[1,1] / second_hyp_test$observed[2,1]) / (second_hyp_test$observed[1,2] / second_hyp_test$observed[2,2])
second_hyp_odds_ratio
```

```
## [1] 1.372671
```

The p-value is greater than 0.05, so we do not reject the null hypothesis. The answer positiveness and the chosen argument seem to be independent; the odds ration tells us that people choose to answer "Yes" to mean the second argument 1.37 times more often than expected. Thus, our second hypothesis is incorrect.

Male and female tend to make emphasis on different arguments of the question no matter if the answer is positive or negative

```
third_hyp_table <- table(gender = form_results_trans$Gender_binary, arg = form_results_trans$Emphasized_argument)
third_hyp_table
```

```
##       arg
## gender  0  1
##      0 59 93
##      1 89 79
```

We can use Pearson's Chi-squared Test because the limitations of its usage are not violated; the null hypothesis is that two factors are independent.

```
third_hyp_test <- chisq.test(third_hyp_table)
third_hyp_test
```

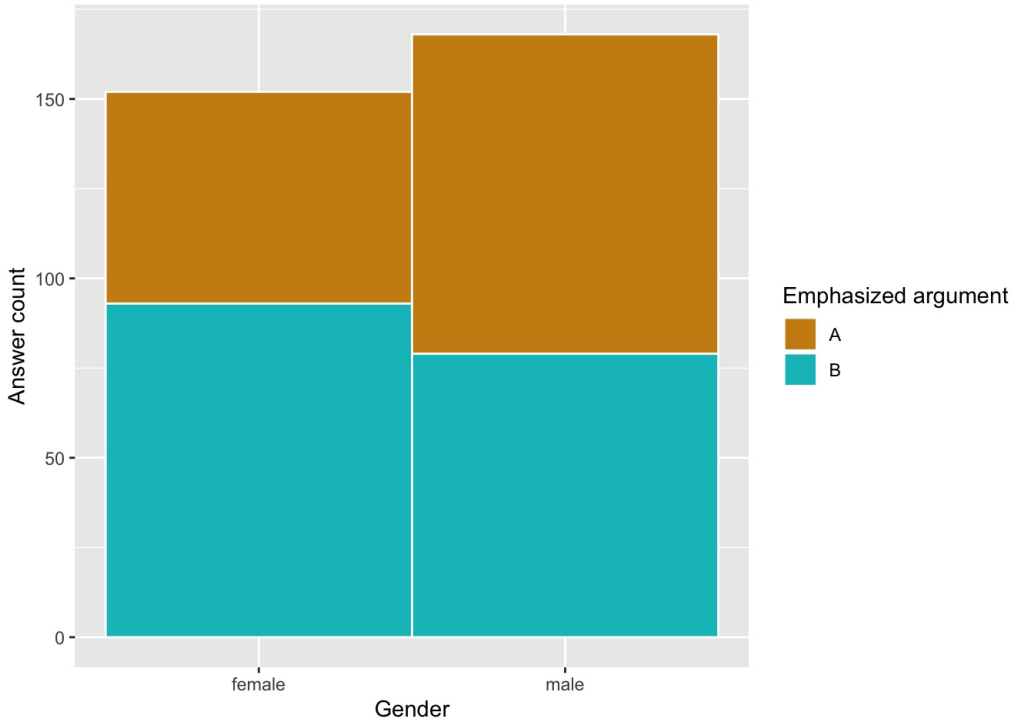
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  third_hyp_table
## X-squared = 5.8797, df = 1, p-value = 0.01532
```

```
third_hyp_odds_ratio <- (third_hyp_test$observed[1,1] / third_hyp_test$observed[2,1]) / (third_hyp_test$observed[1,2] / third_hyp_test$observed[2,2])
third_hyp_odds_ratio
```

```
## [1] 0.5631267
```

The p-value is less than 0.05, so we do reject the null hypothesis. The gender of the respondent and the emphasized argument are not independent; the odds ration tells us that female make emphasis on the first argument and male make emphasis on the second argument both 0.56 times less often than expected. Thus, our third hypothesis is correct.
Let's build a plot to illustrate this:

```
answers_gender_emph <- form_results_trans %>% group_by(Emphasized_argument, Gender_binary) %>% summarise(ans_count
= n())
ggplot(answers_gender_emph, aes(x = factor(Gender_binary), y=ans_count, fill=factor(Emphasized_argument))) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  labs(y = "Answer count") +
  scale_x_discrete(name = "Gender", labels = c("female", "male")) +
  scale_fill_manual(name = "Emphasized argument", labels = c("A", "B"), values=main_colours)
```



The chosen factors are statistically significant so that they can be used to predict the argument emphasized by the respondent

logistic regression

In order to test the hypothesis we need to build some models predicting the emphasized argument using the chosen factors:

```
glm1 <- glm(Emphasized_argument ~ Context_exists + Answer_positiveness + Gender_binary, form_results_trans, family
= "binomial")
summary(glm1)
```

```
##
## Call:
## glm(formula = Emphasized_argument ~ Context_exists + Answer_positiveness +
##     Gender_binary, family = "binomial", data = form_results_trans)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.441  -1.142   0.935   1.037   1.281
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.4455     0.2495   1.786   0.0741 .
## Context_exists        0.1554     0.2268   0.685   0.4931
## Answer_positiveness  -0.1054     0.2333  -0.452   0.6515
## Gender_binary        -0.5806     0.2278  -2.549   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 441.81  on 319  degrees of freedom
## Residual deviance: 434.68  on 316  degrees of freedom
## AIC: 442.68
##
## Number of Fisher Scoring iterations: 4
```

As we can see, the only significant factor is the gender; if the respondent is male, the chance he emphasizes on the second argument drops by 0.58. Let's see if we can use only gender to predict the argument. In order to do it we need to build a new model and make an anova test for the models.

```
glm2 <- glm(Emphasized_argument ~ Gender_binary , form_results_trans, family = "binomial")
summary(glm2)
```

```
##
## Call:
## glm(formula = Emphasized_argument ~ Gender_binary, family = "binomial",
##     data = form_results_trans)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3758  -1.1272   0.9912   0.9912   1.2284
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.4551     0.1664   2.734  0.00625 **
## Gender_binary  -0.5743     0.2271  -2.528  0.01147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 441.81  on 319  degrees of freedom
## Residual deviance: 435.35  on 318  degrees of freedom
## AIC: 439.35
##
## Number of Fisher Scoring iterations: 4
```

```
anova(glm1, glm2, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Emphasized_argument ~ Context_exists + Answer_positiveness +
##     Gender_binary
## Model 2: Emphasized_argument ~ Gender_binary
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       316     434.68
## 2       318     435.35 -2 -0.66623   0.7167
```

As we can see, gender is still significant in the second model. Moreover, the anova test stated that the models are equal.

Let's also see if there is a random effect of respondent:

```
glm3 <- glmer(Emphasized_argument ~ Gender_binary + (1|Respondent), data = form_results_trans, family = "binomial"
)
summary(glm3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Emphasized_argument ~ Gender_binary + (1 | Respondent)
##    Data: form_results_trans
##
##      AIC      BIC   logLik deviance df.resid
##    431.9    443.2   -212.9    425.9      317
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7024 -0.8732  0.5874  0.7969  1.5400
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  Respondent (Intercept) 0.5312   0.7288
## Number of obs: 320, groups:  Respondent, 40
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.5050     0.2430   2.078   0.0377 *
## Gender_binary  -0.6393     0.3344  -1.912   0.0559 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## Gender_bnry -0.728
```

As we can see, gender is still a significant factor; aic is also slightly better, so the third model is better, which means that there is some effect of the respondent — he or she can have a habit of responding.

Decision tree

```
fit4 <- rpart(Emphasized_argument ~ Context_exists + Answer_positiveness + Gender_binary, data = form_results_tran
s, method = 'class')
fit4$variable.importance
```

```
## Gender_binary
##      3.200251
```

As we see, decision tree also tell us that only the gender is significant.

Random forest

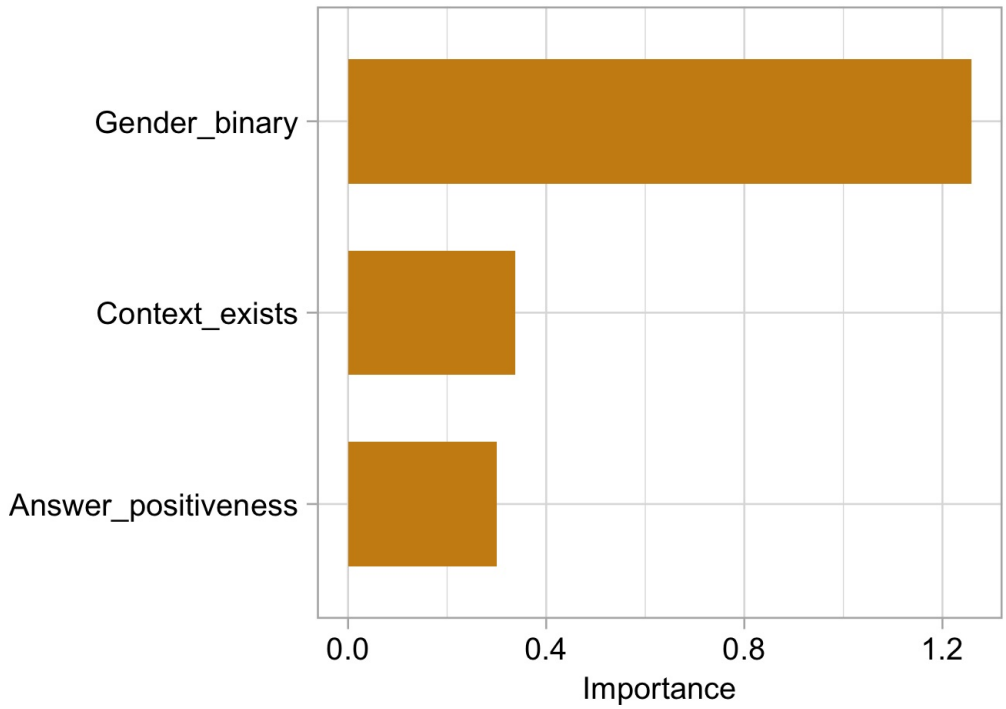
```
set.seed(42)
create_rfplot <- function(rf, type){

  imp <- importance(rf, type = type, scale = F)

  featureImportance <- data.frame(Feature = row.names(imp), Importance = imp[,1])

  p <- ggplot(featureImportance, aes(x = reorder(Feature, Importance), y = Importance)) +
       geom_bar(stat = "identity", fill = "#cc8c12", width = 0.65) +
       coord_flip() +
       theme_light(base_size = 20) +
       theme(axis.title.x = element_text(size = 15, color = "black"),
             axis.title.y = element_blank(),
             axis.text.x  = element_text(size = 15, color = "black"),
             axis.text.y  = element_text(size = 15, color = "black"))
  return(p)
}

fit5 <- randomForest(Emphasized_argument ~ Context_exists + Answer_positiveness + Gender_binary, data = form_resul
ts_trans, ntree = 1000, nodesize=3, importance= TRUE)
create_rfplot(fit5, type = 2)
```



We've built a plot for factor importance measured by decreasing the node impurity, and as we see, the most important factor is gender.

Results

The research has been done in order to find out if some factor influences the choice of the answer on OR-question. Turned out that the only significant factor is gender of the respondent - male make emphasis on the second word less often than women, so the next time a male answers "Yes" he can immediately be asked if he meant the first argument, and he will probably be impressed that his thought has been interpretted correctly. It also turned out that people may have their favourable way to answer on OR-question, which should also be considered. The research can be continued because there are even more things to test: are there more significant factors? how often do people mean "any of the provided choices" while answering positively? can the results be implied to other languages?