Statistical analysis of factors that affect the answer of general questions in spoken Chinese

Hypothesis

In spoken Chinese, there are two ways to answer general question (yes-no question) such as "Are you feeling good?" "Do you play games?": simply "Yes" or "No", or detailed answer that reflects the question ("Not very well", "I play Dota2").

In this project, we suppose that there are two factors that affect the choices, which are:

- Gender of the respondent:

Differences in the way male and female think may lead to different ways of answering.

- Personality of the respondent:

An active voluble extrovert and an introvert of few words may answer the same question in different ways.

- The positive/negative answer of the question:

There is a possibility that when positive answers are given, people will become more talkative.

So the hypotheses are:

- Hypothesis 1: Female prefers detailed answer than male.

- Hypothesis 2: Extrovert prefers detailed answer than introvert.

- Hypothesis 3: People are more inclined to give positive answer in detail, compared to negative answer.

Data Collection

In order to restore the spoken language application environment to a greater extent, I apply face-to-face or telephone interview to collect the data.

First of all, we need to clarify the gender ("M" as male, "F" as female) and personality ("EX" as extrovert, "IN" as introvert) of the respondent;

Then the respondent needs to answer 7 general questions. These questions will be presented to the respondent as shown below:

1) Are you feeling good during self-isolation?

2) Has your university resumed class？

3) Do you want to go back to university?

4) Do you order takeaway at home?

5) Do you play games?

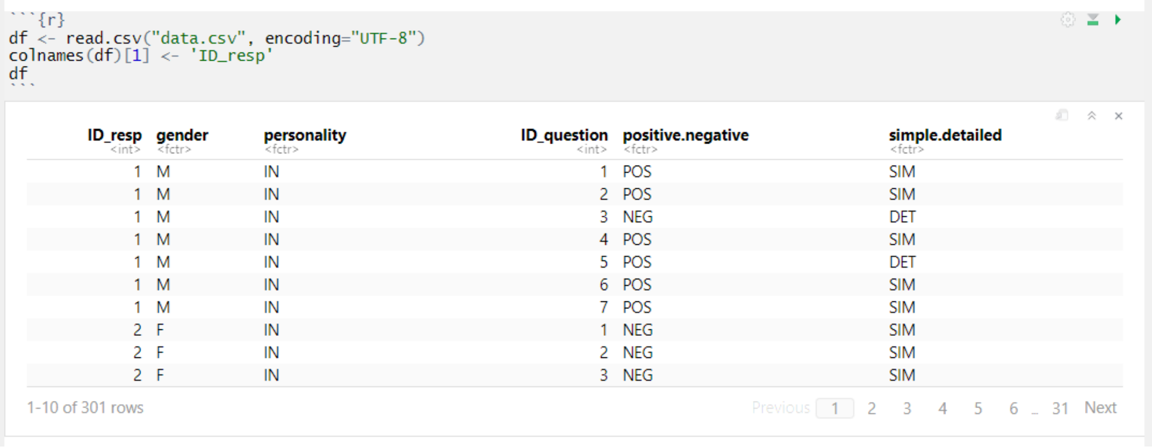
6) Do you like Coca-Cola better than Pepsi?

7) Do you buy clothes online more often than in physical stores?

After that, the answers need to be marked as positive/negative ("POS" as positive, "NEG" as negative) and simple/detailed ("SIM" as simple, "DET" as detailed).

Data Visualization

43 people were interviewed and the answers are recorded in the table below:

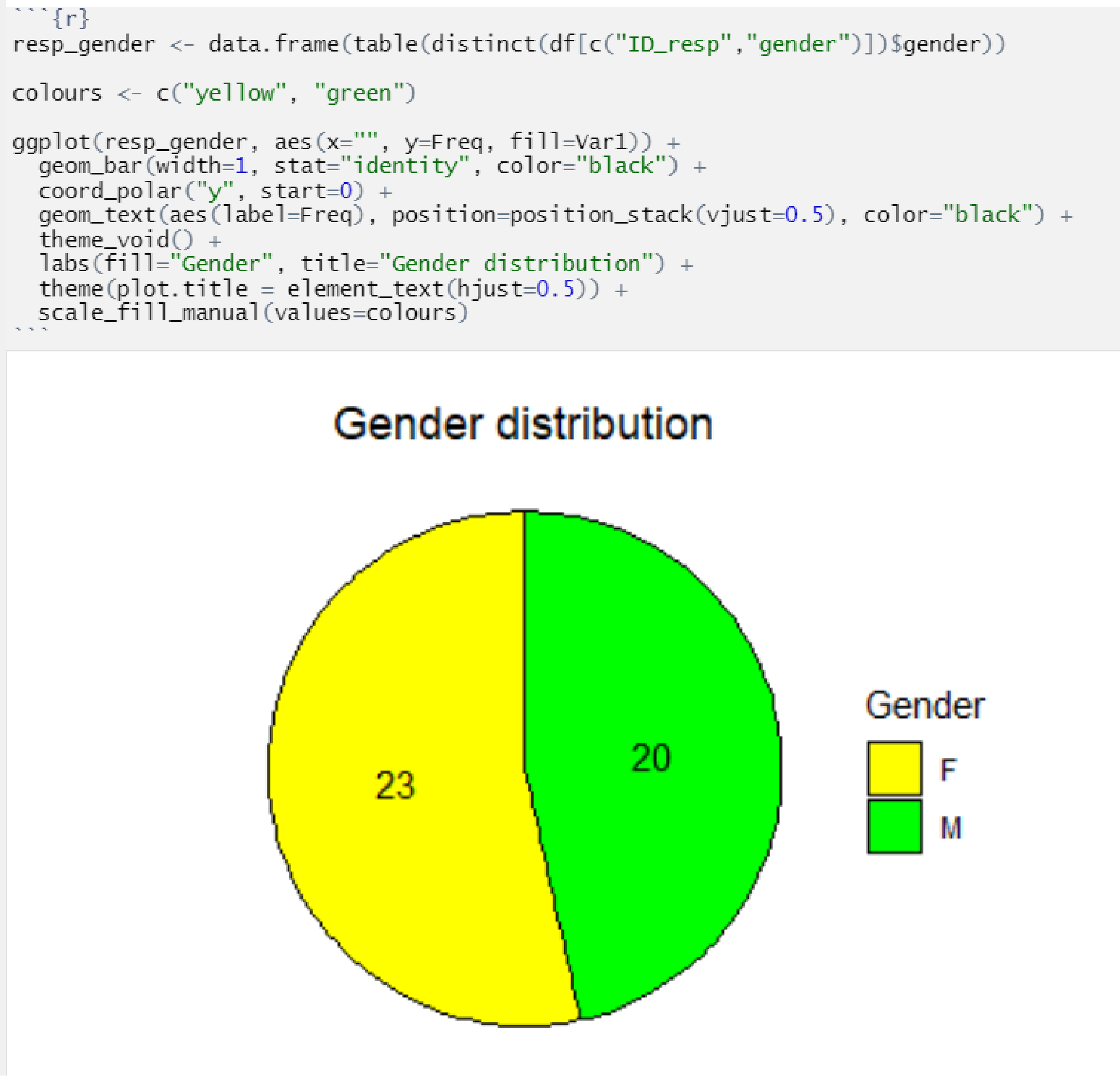


~~~
```{r}
df <- read.csv("data.csv", encoding="UTF-8")
colnames(df)[1] <- 'ID_resp'
df
```
~~~

ID_resp <int>	gender <fctr>	personality <fctr>	ID_question <int>	positive.negative <fctr>	simple.detailed <fctr>
1	M	IN	1	POS	SIM
1	M	IN	2	POS	SIM
1	M	IN	3	NEG	DET
1	M	IN	4	POS	SIM
1	M	IN	5	POS	DET
1	M	IN	6	POS	SIM
1	M	IN	7	POS	SIM
2	F	IN	1	NEG	SIM
2	F	IN	2	NEG	SIM
2	F	IN	3	NEG	SIM

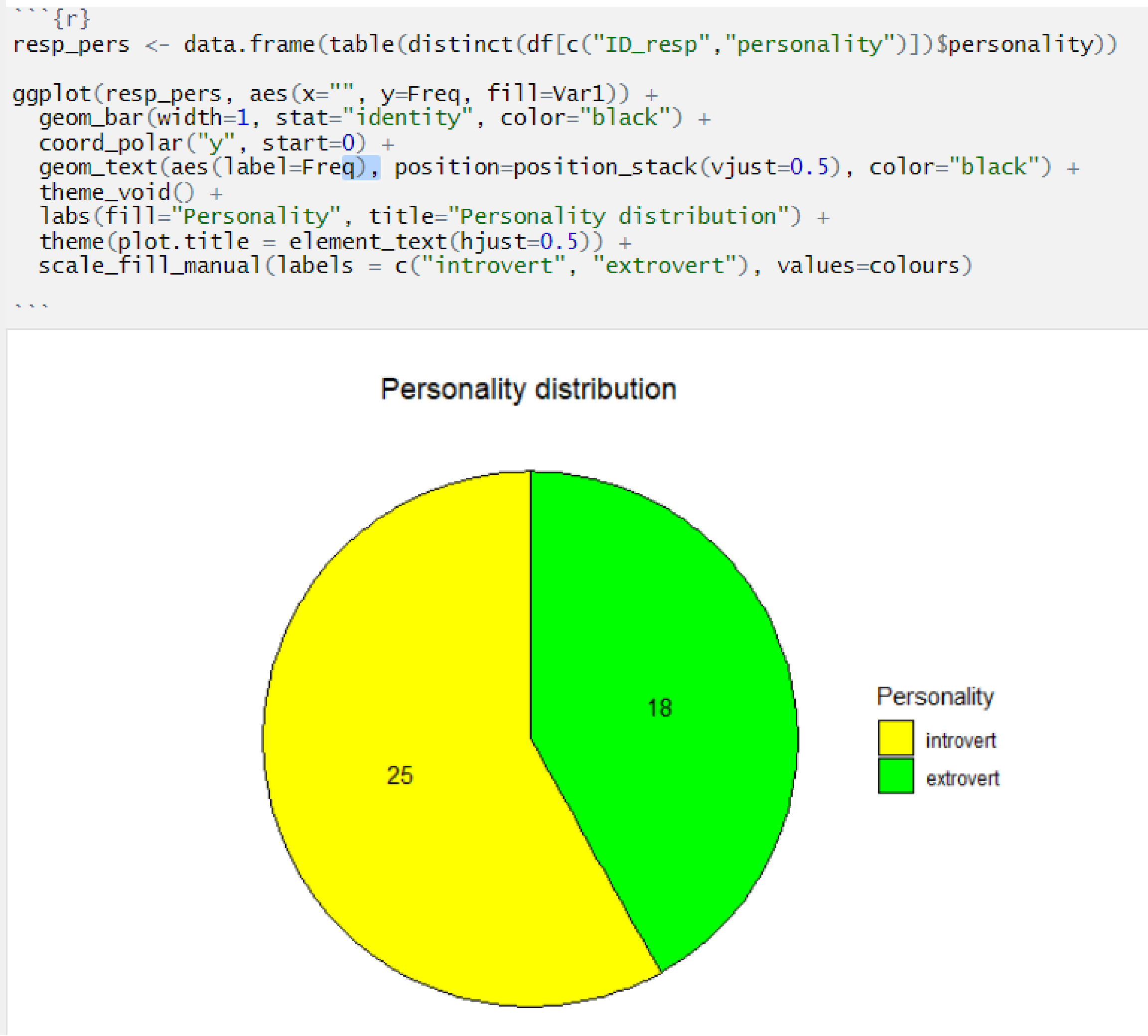
1-10 of 301 rows

First, take a look at distributions of respondents' gender & personality:



~~~
```{r}
resp_gender <- data.frame(table(distinct(df[c("ID_resp","gender")])$gender))

colours <- c("yellow", "green")

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

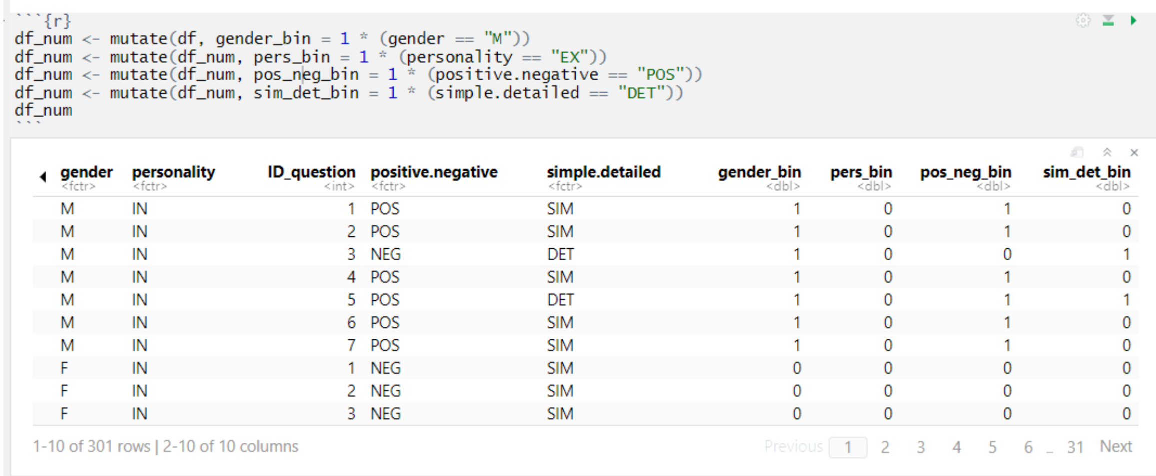
Female : Male is nearly 5:5.



~~~
```{r}
resp_pers <- data.frame(table(distinct(df[c("ID_resp","personality")])$personality))

ggplot(resp_pers, aes(x="", y=Freq, fill=Var1)) +
  geom_bar(width=1, stat="identity", color="black") +
  coord_polar("y", start=0) +
  geom_text(aes(label=Freq), position=position_stack(vjust=0.5), color="black") +
  theme_void() +
  labs(fill="Personality", title="Personality distribution") +
  theme(plot.title = element_text(hjust=0.5)) +
  scale_fill_manual(labels = c("introvert", "extrovert"), values=colours)
```
~~~

introvert : extrovert is nearly 6:4.

To analyze the data, we should at first transform it to numeric binary form:



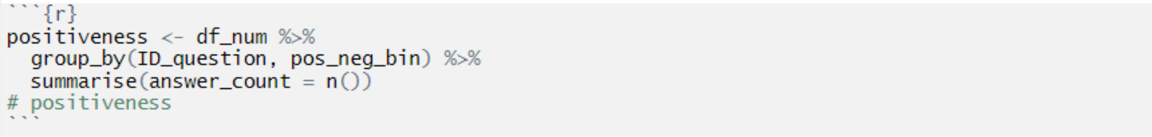
~~~
```{r}
df_num <- mutate(df, gender_bin = 1 * (gender == "M"))
df_num <- mutate(df_num, pers_bin = 1 * (personality == "EX"))
df_num <- mutate(df_num, pos_neg_bin = 1 * (positive.negative == "POS"))
df_num <- mutate(df_num, sim_det_bin = 1 * (simple.detailed == "DET"))
df_num
```
~~~

gender <fctr>	personality <fctr>	ID_question <int>	positive.negative <fctr>	simple.detailed <fctr>	gender_bin <dbl>	pers_bin <dbl>	pos_neg_bin <dbl>	sim_det_bin <dbl>
M	IN	1	POS	SIM	1	0	1	0
M	IN	2	POS	SIM	1	0	1	0
M	IN	3	NEG	DET	1	0	0	1
M	IN	4	POS	SIM	1	0	1	0
M	IN	5	POS	DET	1	0	1	1
M	IN	6	POS	SIM	1	0	1	0
M	IN	7	POS	SIM	1	0	1	0
F	IN	1	NEG	SIM	0	0	0	0
F	IN	2	NEG	SIM	0	0	0	0
F	IN	3	NEG	SIM	0	0	0	0

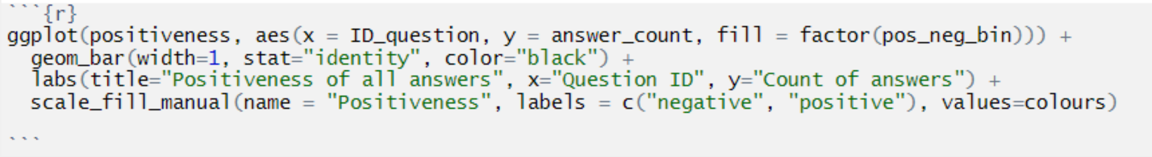
1-10 of 301 rows | 2-10 of 10 columns

Previous 1 2 3 4 5 6 ... 31 Next

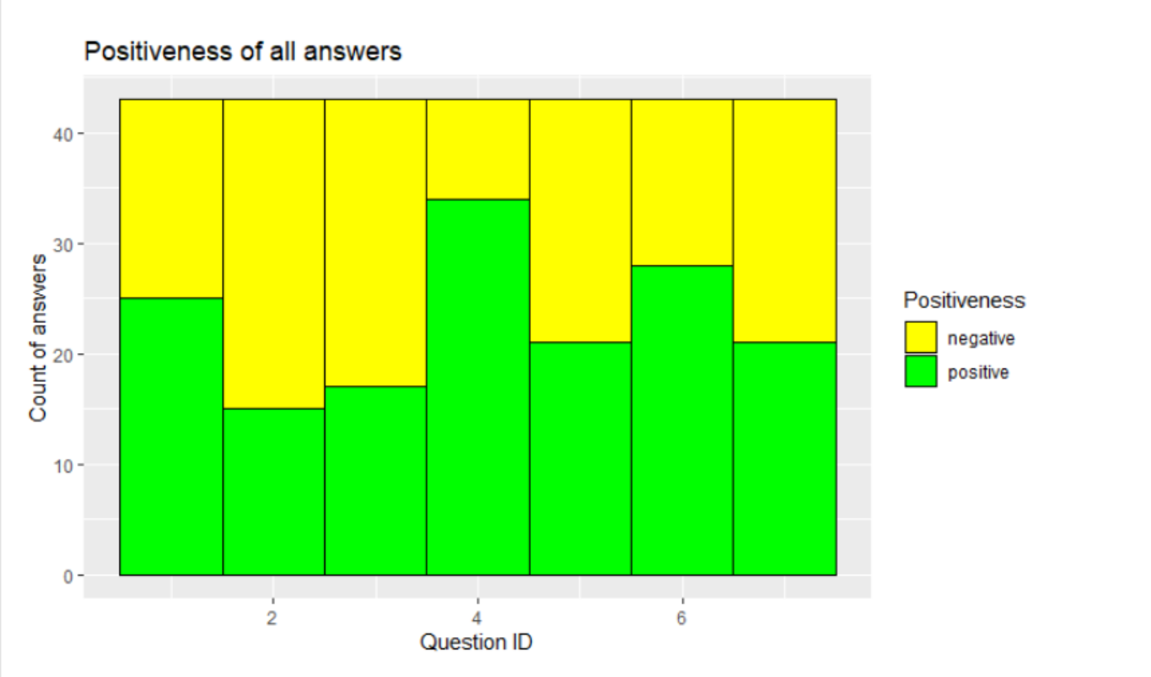
Let's see the positiveness of all answers:



~~~
```{r}
positiveness <- df_num %>%
  group_by(ID_question, pos_neg_bin) %>%
  summarise(answer_count = n())
# positiveness
```
~~~

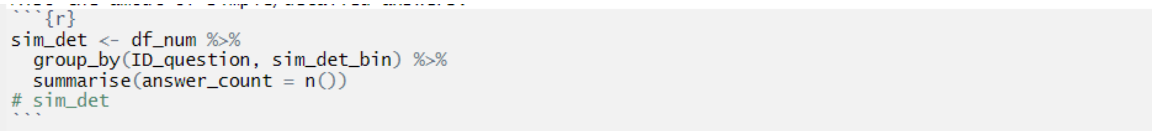


~~~
```{r}
ggplot(positiveness, aes(x = ID_question, y = answer_count, fill = factor(pos_neg_bin))) +
  geom_bar(width=1, stat="identity", color="black") +
  labs(title="Positiveness of all answers", x="Question ID", y="Count of answers") +
  scale_fill_manual(name = "Positiveness", labels = c("negative", "positive"), values=colours)

```
~~~

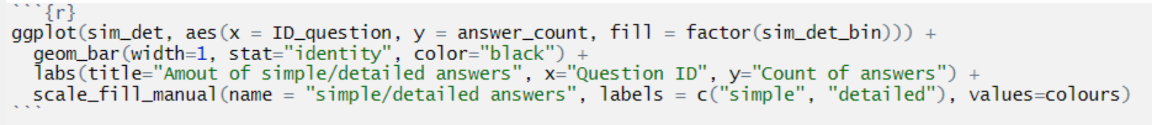


Except that the 4th question get mostly positive answers, the positiveness of other questions is basically near 1:1.

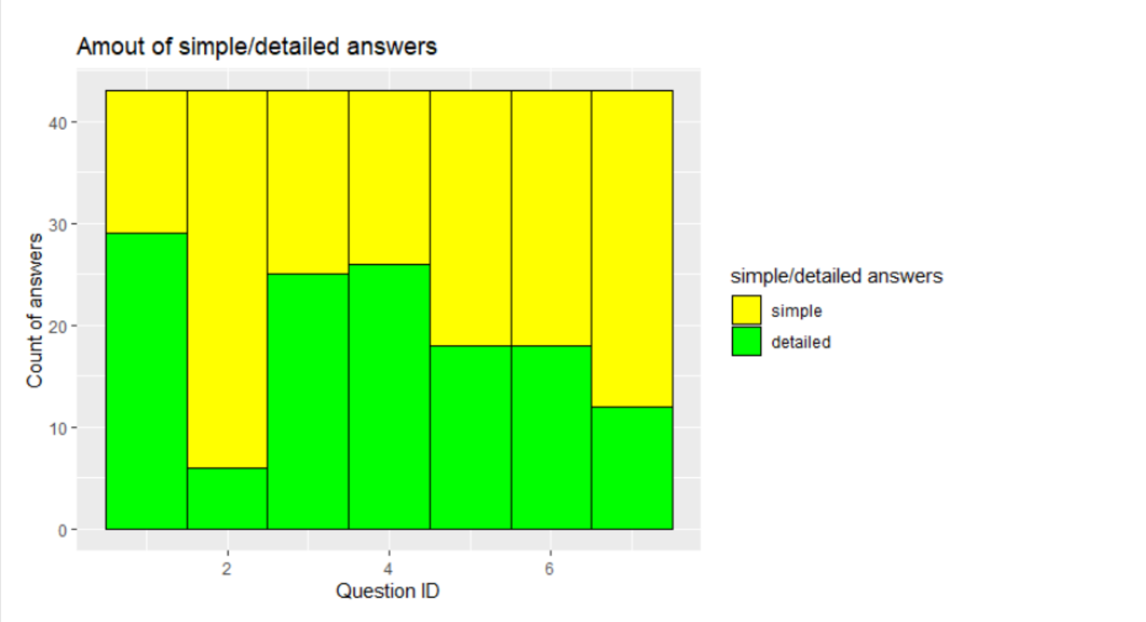
Also the amount of simple/detailed answers:



~~~
```{r}
sim_det <- df_num %>%
  group_by(ID_question, sim_det_bin) %>%
  summarise(answer_count = n())
# sim_det
```
~~~



~~~
```{r}
ggplot(sim_det, aes(x = ID_question, y = answer_count, fill = factor(sim_det_bin))) +
  geom_bar(width=1, stat="identity", color="black") +
  labs(title="Amout of simple/detailed answers", x="Question ID", y="Count of answers") +
  scale_fill_manual(name = "simple/detailed answers", labels = c("simple", "detailed"), values=colours)
```
~~~



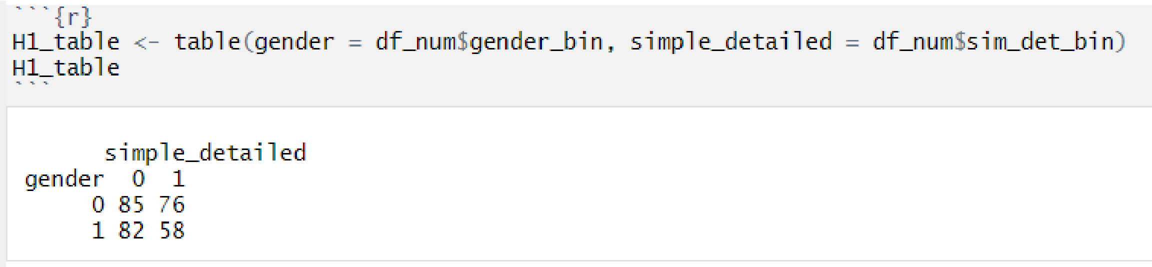
Mostly people prefer answering general questions in simple ways.

The 2nd & the 7th questions get mostly simple answers.

Testing hypotheses

Hypothesis 1

Female prefers detailed answer than male.

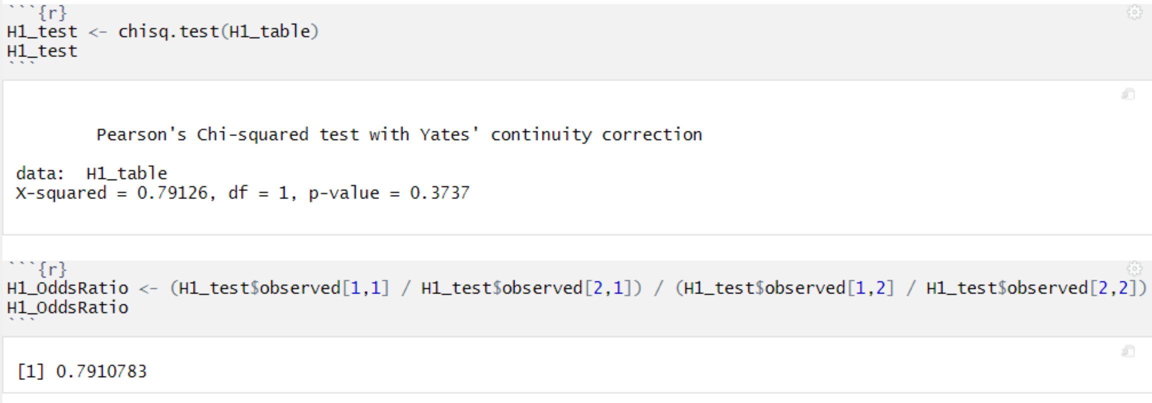


~~~
```{r}
H1_table <- table(gender = df_num$gender_bin, simple_detailed = df_num$sim_det_bin)
H1_table
```
~~~

```
       simple_detailed
gender  0  1
     0 85 76
     1 82 58
```

Data here meets the applicable conditions of Pearson's Chi-squared Test.

H0: Gender of respondent and the choice of simple/detailed answer are independent.

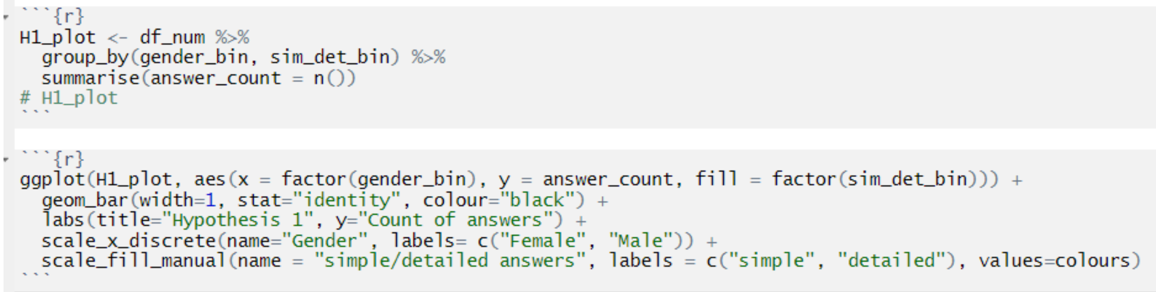


~~~
```{r}
H1_test <- chisq.test(H1_table)
H1_test
```
~~~

```

	Pearson's Chi-squared test with Yates' continuity correction

data:  H1_table
X-squared = 0.79126, df = 1, p-value = 0.3737
```

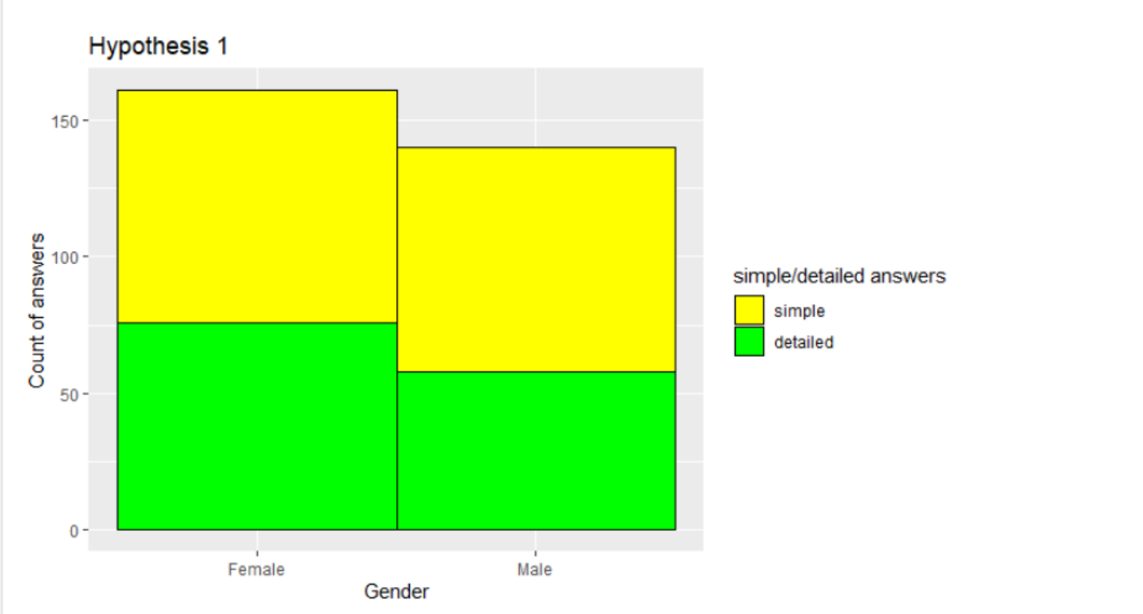
~~~
```{r}
H1_OddsRatio <- (H1_test$observed[1,1] / H1_test$observed[2,1]) / (H1_test$observed[1,2] / H1_test$observed[2,2])
H1_OddsRatio
```
~~~

```
[1] 0.7910783
```



~~~
```{r}
H1_plot <- df_num %>%
  group_by(gender_bin, sim_det_bin) %>%
  summarise(answer_count = n())
# H1_plot
```
~~~

~~~
```{r}
ggplot(H1_plot, aes(x = factor(gender_bin), y = answer_count, fill = factor(sim_det_bin))) +
  geom_bar(width=1, stat="identity", colour="black") +
  labs(title="Hypothesis 1", y="Count of answers") +
  scale_x_discrete(name="Gender", labels= c("Female", "Male")) +
  scale_fill_manual(name = "simple/detailed answers", labels = c("simple", "detailed"), values=colours)
```
~~~



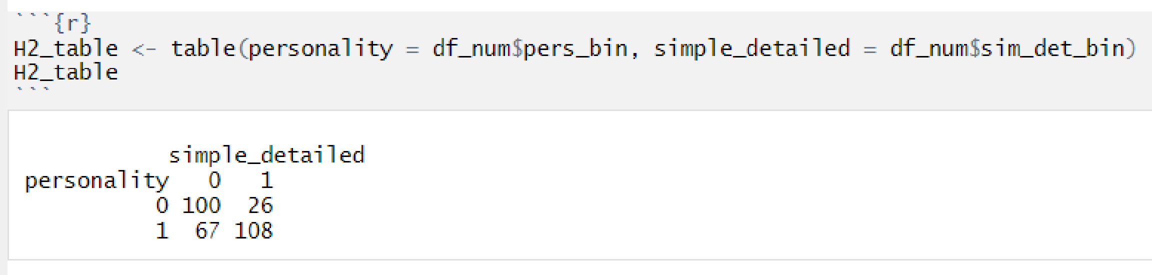
The p-value > 0.05, so we do not reject null hypothesis: Gender of respondent and the choice of simple/detailed answer are independent.

The Odds Ratio = 0.79, which means male who choses detailed answer are 0.79 times less often than expected.

Hypothesis 1 is incorrect.

Hypothesis 2

Extrovert prefers detailed answer than introvert.

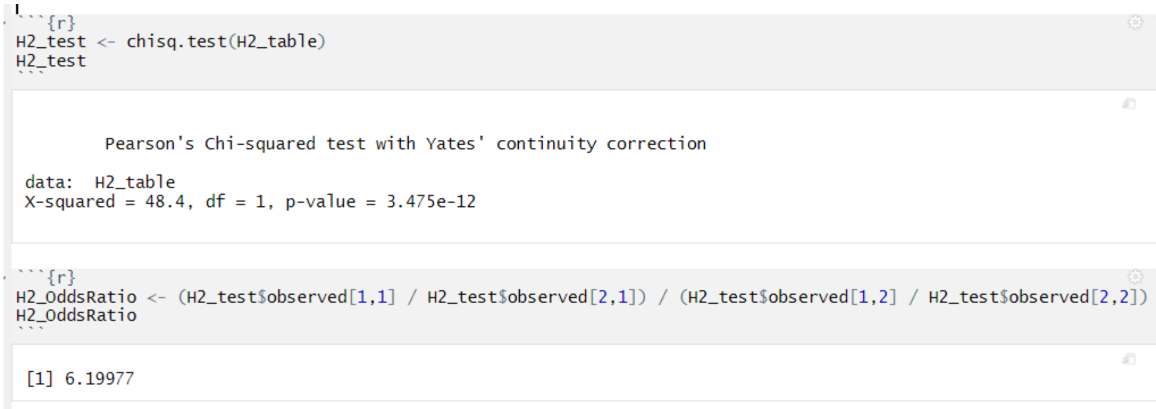


~~~
```{r}
H2_table <- table(personality = df_num$pers_bin, simple_detailed = df_num$sim_det_bin)
H2_table
```
~~~

```
           simple_detailed
personality   0   1
          0 100  26
          1  67 108
```

Data here meets the applicable conditions of Pearson's Chi-squared Test.

H0: Personality of respondent and the choice of simple/detailed answer are independent.

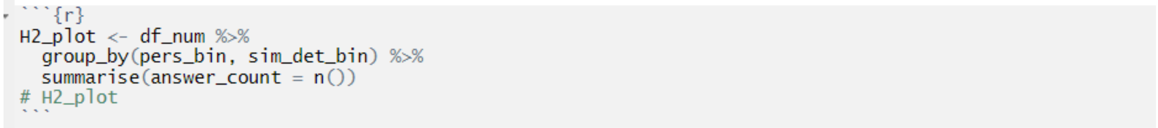


~~~
```{r}
H2_test <- chisq.test(H2_table)
H2_test
```
~~~

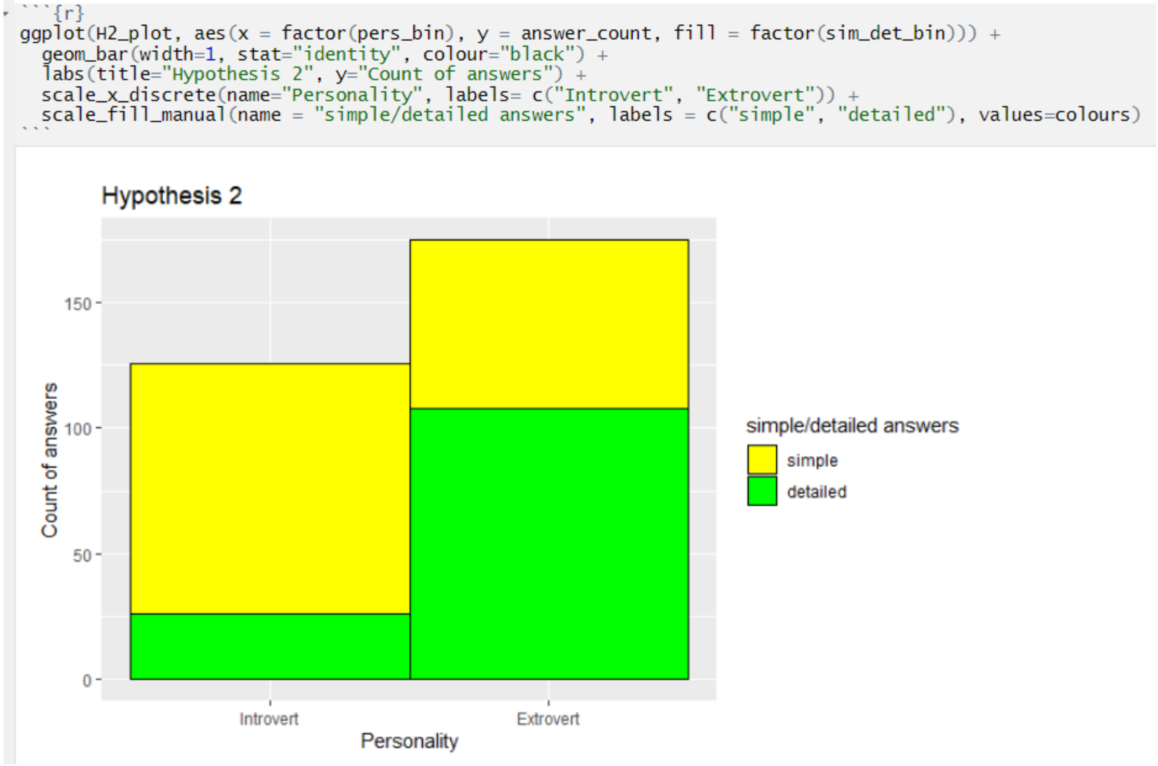
```

	Pearson's Chi-squared test with Yates' continuity correction

data:  H2_table
X-squared = 48.4, df = 1, p-value = 3.475e-12
```

~~~
```{r}
H2_OddsRatio <- (H2_test$observed[1,1] / H2_test$observed[2,1]) / (H2_test$observed[1,2] / H2_test$observed[2,2])
H2_OddsRatio
```
~~~

```
[1] 6.19977
```



~~~
```{r}
H2_plot <- df_num %>%
  group_by(pers_bin, sim_det_bin) %>%
  summarise(answer_count = n())
# H2_plot
```
~~~



~~~
```{r}
ggplot(H2_plot, aes(x = factor(pers_bin), y = answer_count, fill = factor(sim_det_bin))) +
  geom_bar(width=1, stat="identity", colour="black") +
  labs(title="Hypothesis 2", y="Count of answers") +
  scale_x_discrete(name="Personality", labels= c("Introvert", "Extrovert")) +
  scale_fill_manual(name = "simple/detailed answers", labels = c("simple", "detailed"), values=colours)
```
~~~

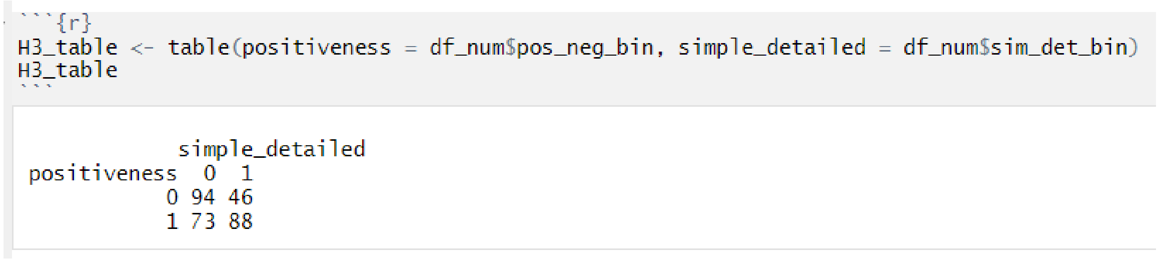
The p-value < 0.05, so we reject null hypothesis: Personality of respondent and the choice of simple/detailed answer are independent.

The Odds Ratio = 6.20, which means introvert who choses simple answer and extrovert who choses detailed answer are 6.20 times more often than expected.

Hypothesis 2 is correct.

Hypothesis 3

People are more inclined to give positive answer in detail, compared to negative answer.

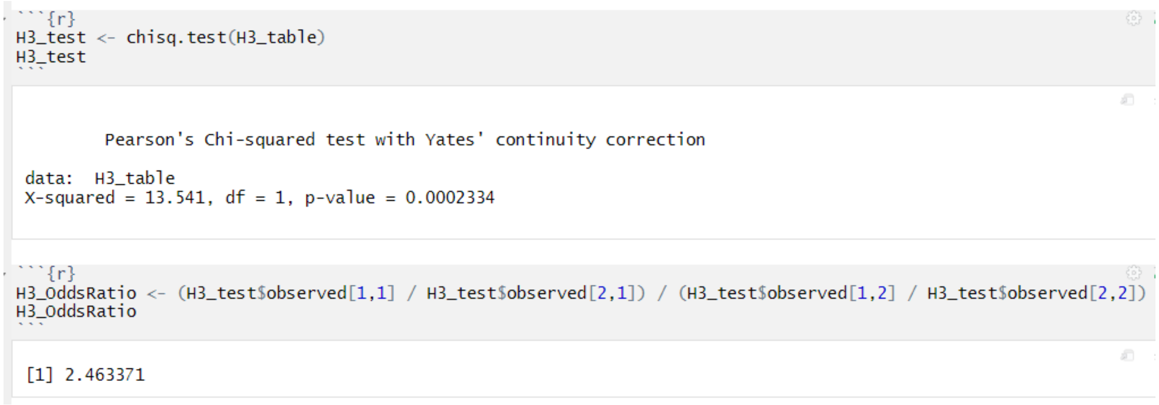


~~~
```{r}
H3_table <- table(positiveness = df_num$pos_neg_bin, simple_detailed = df_num$sim_det_bin)
H3_table
```
~~~

```
            simple_detailed
positiveness  0  1
           0 94 46
           1 73 88
```

Data here meets the applicable conditions of Pearson's Chi-squared Test.

H0: Positiveness of answer and the choice of simple/detailed answer are independent.

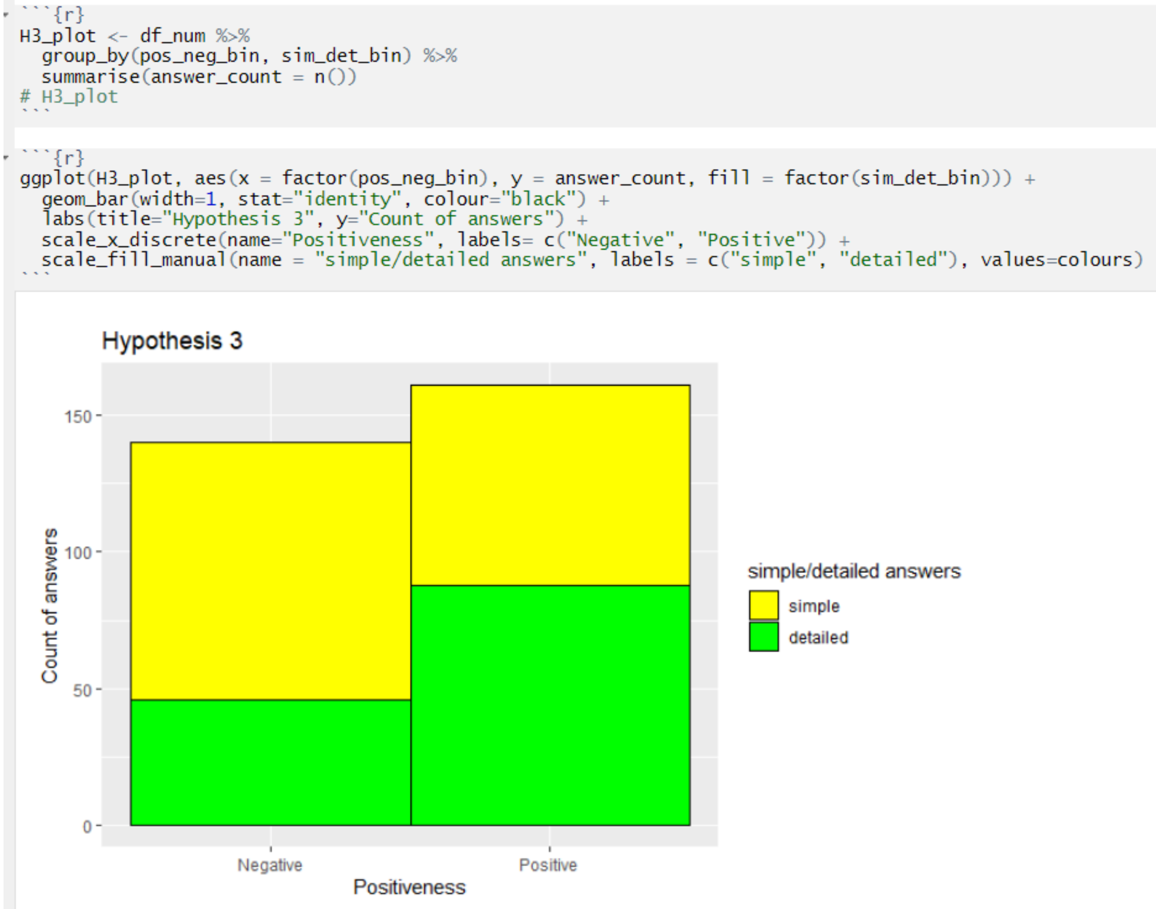


~~~
```{r}
H3_test <- chisq.test(H3_table)
H3_test
```
~~~

```

	Pearson's Chi-squared test with Yates' continuity correction

data:  H3_table
X-squared = 13.541, df = 1, p-value = 0.0002334
```

~~~
```{r}
H3_OddsRatio <- (H3_test$observed[1,1] / H3_test$observed[2,1]) / (H3_test$observed[1,2] / H3_test$observed[2,2])
H3_OddsRatio
```
~~~

```
[1] 2.463371
```



~~~
```{r}
H3_plot <- df_num %>%
  group_by(pos_neg_bin, sim_det_bin) %>%
  summarise(answer_count = n())
# H3_plot
```
~~~

~~~
```{r}
ggplot(H3_plot, aes(x = factor(pos_neg_bin), y = answer_count, fill = factor(sim_det_bin))) +
  geom_bar(width=1, stat="identity", colour="black") +
  labs(title="Hypothesis 3", y="Count of answers") +
  scale_x_discrete(name="Positiveness", labels= c("Negative", "Positive")) +
  scale_fill_manual(name = "simple/detailed answers", labels = c("simple", "detailed"), values=colours)
```
~~~

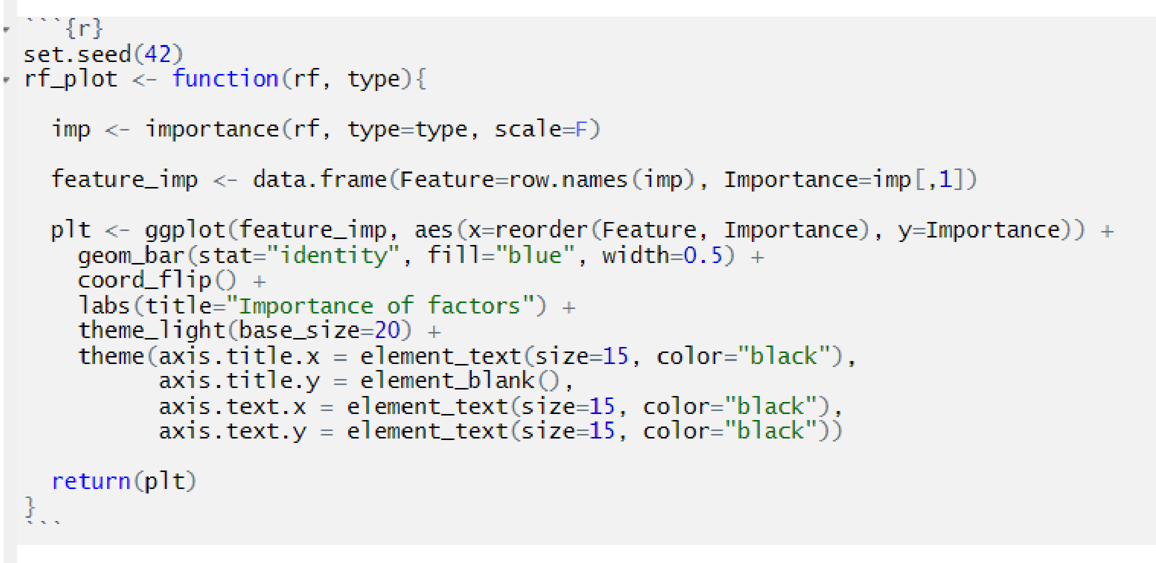
The p-value < 0.05, so we reject null hypothesis: Positiveness of answer and the choice of simple/detailed answer are independent.

The Odds Ratio = 2.46, which means negative answer in simple way and positive answer in detail are 2.46 times more often than expected.

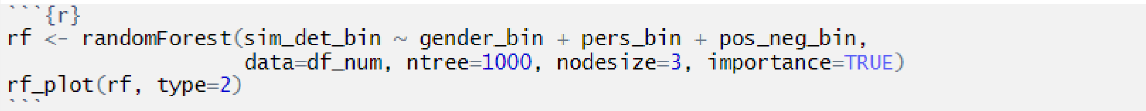
Hypothesis 3 is correct.

Random forest

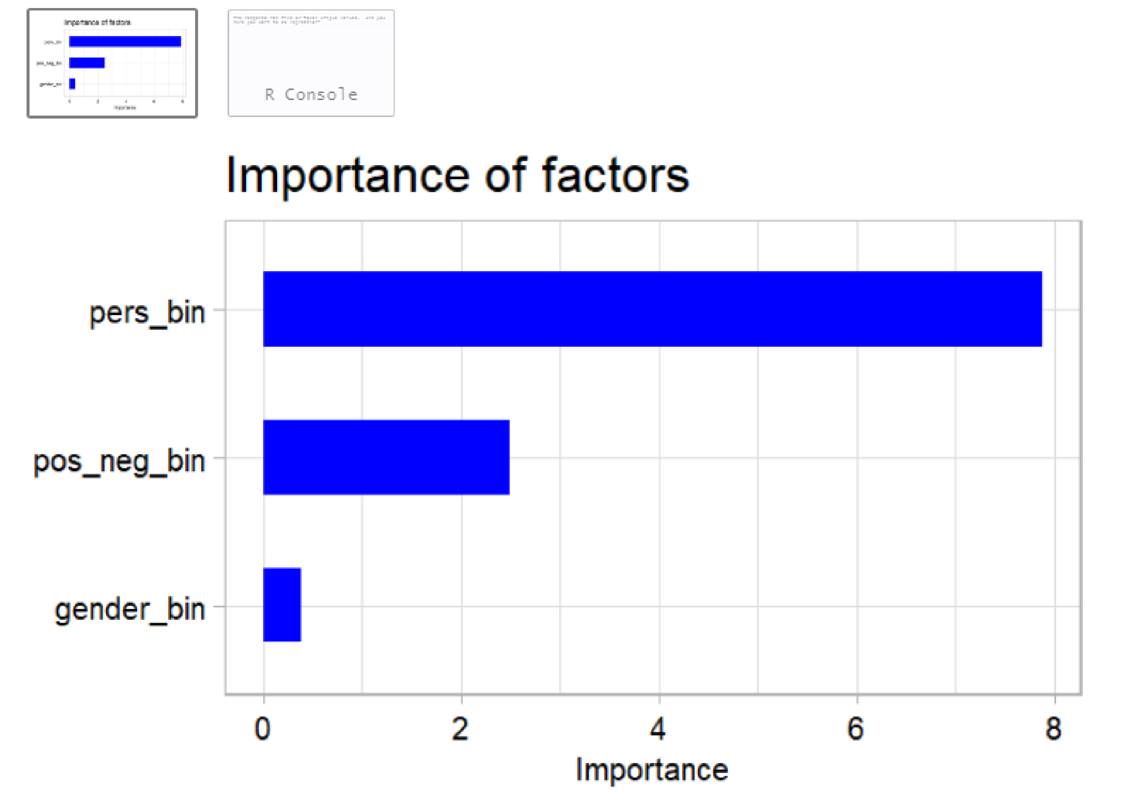
Let's take a look at the significance of these factors with Random forest:



~~~
```{r}
set.seed(42)
rf_plot <- function(rf, type){

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

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

  plt <- ggplot(feature_imp, aes(x=reorder(Feature, Importance), y=Importance)) +
    geom_bar(stat="identity", fill="blue", width=0.5) +
    coord_flip() +
    labs(title="Importance of factors") +
    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(plt)
}
```
~~~



~~~
```{r}
rf <- randomForest(sim_det_bin ~ gender_bin + pers_bin + pos_neg_bin,
                   data=df_num, ntree=1000, nodesize=3, importance=TRUE)
rf_plot(rf, type=2)
```
~~~



As we can see from the plot, personality is the most significant factor and occupies a large proportion; The positiveness of the answer follows; Gender has minimal impact on results.

Conclusion & future works

Results show that two factors -- personality of respondent & positiveness of answer -- affect the answer of general questions in spoken Chinese. Personality of respondent affects mostly. We can use these factors to predict whether the answer is simple/detailed. Though the results may not be accurate, because the 4th question get mostly positive answers, and the 2nd & the 7th questions get mostly simple answers. The problem may have been caused due to the chosen questions. In the future, we can continue to improve this project: Design questions in form with balanced positive/negative & simple/detailed answers; Test whether the results can be applied to other languages; etc.