针对性别与社会经济地位等变量之间的关系的研究

姓名：吴桐

班级：22计算广告

学号：202218093013

1、问题概述

本研究旨在探讨性别与其他变量（如宗教信仰、总收入、政治面貌、身高、体重以及半年内上网情况）之间的关系，特别是与个人社会经济地位的关系。通过分析CGSS2015中国综合调查数据，本研究将采用适当的统计方法，包括逻辑回归和相关系数分析，来探究性别对于这些变量的影响，并进一步研究性别与个人社会经济地位之间的联系。

具体而言，我们将运用逻辑回归模型来预测性别，并考察性别对宗教信仰、总收入、政治面貌、身高、体重以及半年内上网情况等变量的关联。此外，我们还将运用相关系数分析来评估这些变量之间的相关程度。

除了统计分析，我们还将进行可视化分析，通过绘制图表和可视化图形，直观展示性别与个人社会经济地位之间的关系。这有助于更好地理解性别在各个变量上的差异以及与社会经济地位的关联。

通过对性别与其他变量之间关系的深入研究，我们旨在为性别平等和包容的社会发展提供支持，并提供关于性别与个人社会经济地位之间关系的新见解。这将有助于促进公平、可持续和包容的社会目标的实现，并为决策者提供参考依据，以制定相应的政策和措施。

2、模型构建方法以及变量确定

2.1 模型构建方法

本研究除了采用构建相关系数矩阵的方法还要构建逻辑回归模型来探究性别与其他变量之间的关系。逻辑回归是一种广泛应用于分类问题的统计方法，可以预测因变量（性别）的取值，并分析自变量（社会经济地位、宗教信仰、总收入、政治面貌、身高、体重和上网情况等）对因变量的影响程度。在逻辑回

归模型中，我们将使用二元变量（0和1）来表示性别，0代表女性，1代表男性。通过拟合逻辑回归模型，我们可以得到自变量对于性别的权重系数，进而判断不同自变量与性别之间的关联程度。

2.2 自变量与因变量

在本研究中，自变量包括宗教信仰、总收入、政治面貌、身高、体重和半年内是否上网等变量。这些自变量将被用来预测性别（因变量）。例如，在逻辑回归模型中，我们可以通过分析各个自变量的系数来确定它们与性别之间的相关程度。具体而言，我们可以观察到宗教信仰、总收入、政治面貌、身高、体重和半年内是否上网等变量对于性别的预测能力。通过分析这些关系，我们可以深入了解性别与这些自变量之间的联系。

3、数据分析

3.1 对自变量和因变量进行描述

我们可以对各个变量进行描述：

因变量：

1.性别（Gender）：该变量表示被调查者的性别，取值为1和2，分别代表男性和女性，是数值型变量。

自变量：

2.省份（Province）：该变量表示被调查者所在的省份，为字符型变量，共有8148个观测值。

2.性别（Gender）：该变量表示被调查者的性别，取值为1和2，分别代表男性和女性，是数值型变量。

3.宗教信仰（Religion）：该变量表示被调查者的宗教信仰，取值范围从1到21，是数值型变量。

4.总收入（Total Income）：该变量表示被调查者去年全年的总收入，单位为人民币（RMB），是数值型变量。

5.政治面貌（Political Status）：该变量表示被调查者的政治面貌，取值范围从1到5，还有98、98，是数值型变量。

6.身高（Height）：该变量表示被调查者的身高，单位为厘米（cm），是数值型变量。

7.体重（Weight）：该变量表示被调查者的体重，单位为千克（kg），是数值型变量。

8.半年内是否上网（Internet Usage）：该变量表示被调查者在最近半年是否上过互联网，包括使用电脑、手机、智能设备等，是数值型变量。

9.认为男性比女性能力更强（Belief in Male Superiority）：该变量表示被调查者对于男性与女性能力的看法，取值范围从1到5，分别代表完全不同意、比较不同意、无所谓同意不同意、比较同意和完全同意，是数值型变量。

10.本人的社会经济地位（Self-assessed Socioeconomic Status）：该变量表示被调查者对自己在目前社会上的社会经济地位的评估，取值范围从1到5，分别代表上层、中上层、中层、中下层和下层，是数值型变量。

总结以上数据，我们可以看到参与调查的人群中，男女比例相对平衡，宗教信仰较多样，收入分布较广，政治面貌多样化，身高和体重有一定的分布范围，大部分人在半年内上过互联网，关于男性与女性能力的认知分布较为均匀，而对社会经济地位的大多集中在3位置上。

3.2进一步的描述分析和相关分析

代码

```
############################

# 导入所需的库

library(ggplot2)

library(dplyr)

library(corrplot)


table(data1$本人的社会经济地位)

# 将数据框中的98替换为8

data1$本人的社会经济地位[data1$本人的社会经济地位 == 98] <- 8
```

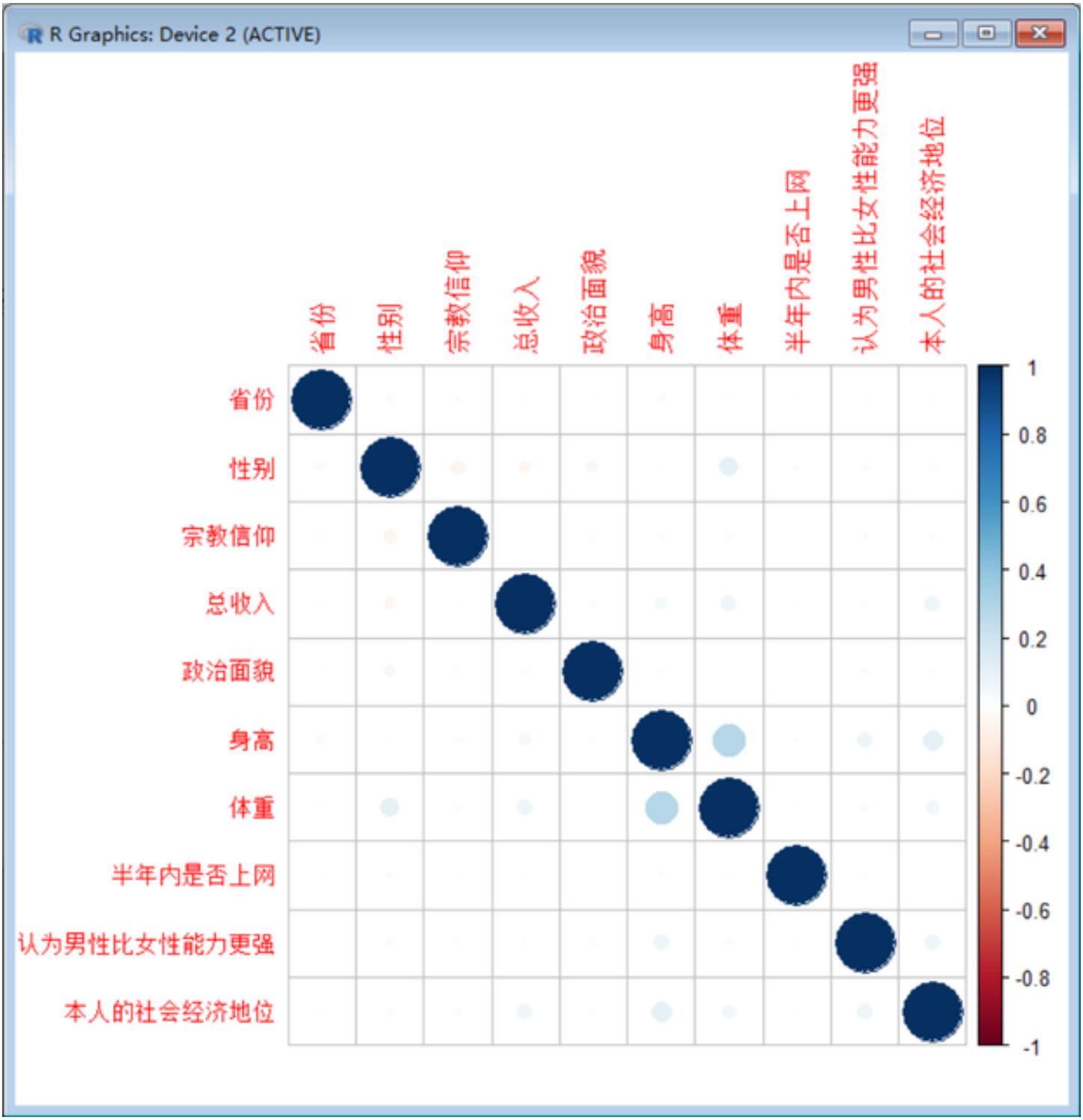
```
# 将数据框中的99替换为9
data1$本人的社会经济地位[data1$本人的社会经济地位 == 99] <- 9
summary(data1)
str(data1)
cor(data1[9],data1[10])


#######性别与经济地位的描述性统计分析
# 频数和比例
gender_counts <- table(data1$性别)
gender_prop <- prop.table(gender_counts)

social_status_counts <- table(data1$本人的社会经济地位)
social_status_prop <- prop.table(social_status_counts)

# 平均值
average_income <- mean(data1$总收入)

# 输出结果
cat("性别频数：\n")
print(gender_counts)

cat("\n性别比例：\n")
print(gender_prop)

cat("\n社会经济地位频数：\n")
```

```
print(social_status_counts)


cat("\n社会经济地位比例：\n")

print(social_status_prop)


cat("\n平均收入：", average_income, "\n")



###########相关性

# 计算相关系数

correlation <- cor(data1$性别, data1$本人的社会经济地位, method = "pearson")


# 检验相关系数的显著性

cor_test <- cor.test(data1$性别, data1$本人的社会经济地位)


# 输出结果

cat("p-value：", cor_test$p.value, "\n")

#可以看出性别与经济地位并没有什么太大的关系


data1$省份 = as.factor(data1$省份)


data1$省份 = as.numeric(data1$省份)

#看一下与什么关系比较大

corrplot(cor(data1))

##这么看来本人的社会经济地位居然和身高有一点点弱关系。
```

结果

```
> summary(data1)
      省份            性别          宗教信仰          总收入            政治面貌          身高            体重          半年内是否上网   认为男性比女性能力更强 本人的社会经济地位
 Min.   : 1.00   Min.   :0.0000   Min.   : 1.000   Min.   :      0   Min.   : 1.000   Min.   : 98.0   Min.   : 40.0   Min.   : 1.000   Min.   : 1.000   Min.   : 1.0
 1st Qu.: 5.00   1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.:   3000   1st Qu.: 1.000   1st Qu.:158.0   1st Qu.:109.0   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 3.0
 Median :10.00   Median :0.0000   Median : 1.000   Median :  30000   Median : 1.000   Median :163.0   Median :120.0   Median : 1.000   Median : 3.000   Median : 4.0
 Mean   :10.27   Mean   :0.4515   Mean   : 1.904   Mean   :1046361   Mean   : 1.586   Mean   :178.5   Mean   :130.8   Mean   : 1.421   Mean   : 4.063   Mean   : 5.8
 3rd Qu.:16.00   3rd Qu.:1.0000   3rd Qu.: 1.000   3rd Qu.:  70000   3rd Qu.: 1.000   3rd Qu.:170.0   3rd Qu.:140.0   3rd Qu.: 2.000   3rd Qu.: 4.000   3rd Qu.: 5.0
 Max.   :19.00   Max.   :1.0000   Max.   :21.000   Max.   :9999999   Max.   :99.000   Max.   :999.0   Max.   :999.0   Max.   :99.000   Max.   :99.000   Max.   :99.0
```

```
>
> # 平均值
> average_income <- mean(data1$总收入)
>
> # 输出结果
> cat("性别频数：\n")
性别频数：
> print(gender_counts)

   0    1
4469 3679
>
> cat("\n性别比例：\n")

性别比例：
> print(gender_prop)

         0         1
0.5484782 0.4515218
>
> cat("\n社会经济地位频数：\n")

社会经济地位频数：
> print(social_status_counts)

   1    2    3    4    5   98   99
  41  445 3046 2553 1884  162   17
>
> cat("\n社会经济地位比例：\n")

社会经济地位比例：
> print(social_status_prop)

           1           2           3           4           5          98          99
0.005031910 0.054614629 0.373834070 0.313328424 0.231222386 0.019882180 0.002086402
>
> cat("\n平均收入：", average_income, "\n")

平均收入： 1046361
>
```



分析

我们可以得到以下分析：

1、性别频数和比例：

男性个体数为 3679，女性个体数为 4469。

男性占总样本的 45.15%，女性占总样本的 54.85%。

2、社会经济地位频数和比例：

社会经济地位为1的个体数为41，地位为2的个体数为445，地位为3的个体数为3046，地位为4的个体数为2553，地位为5的个体数为1884。

其他地位（98和99）的个体数较少，分别为162和17。

社会经济地位为3的个体最多，占总样本的37.38%。

3、平均收入：

样本中的个体平均收入为1,046,361。

根据以上分析，我们可以进一步考察不同变量之间的关系，例如性别与省份、宗教信仰、政治面貌、认为男性比女性能力更强等之间的关系。此外，还可以通过进一步的统计分析和可视化手段，深入了解变量之间的相互影响，并对模型进行建立和预测

根据相关系数矩阵，我们可以得到以下分析和建议：

1、性别与其他变量之间的关系：

2、性别与宗教信仰之间的相关系数为-0.056，表示两者之间略微负相关。这可能意味着不同性别在宗教信仰上存在一定差异。

3、性别与总收入之间的相关系数为-0.040，表示两者之间略微负相关。然而，相关性较低，不足以说明性别对总收入的决定性影响。

4、社会经济地位与其他变量之间的关系：

5、社会经济地位与身高、体重以及认为男性比女性能力更强之间的相关系数都较高，分别为0.107、0.053和0.065。这可能意味着社会经济地位较高的个体在身高、体重以及认为男性比女性能力更强方面可能具有一定优势。

此外，社会经济地位与总收入之间的相关系数为0.061，表示两者之间存在一定的正相关关系。

基于以上分析，我们可以向企业或政策制定者提出以下建议：

1、在进行市场划分时，可以考虑性别和宗教信仰之间的关系，以更好地理解不同群体之间的差异和需求。

2、在进行社会经济调查时，可以关注身高、体重以及认为男性比女性能力更强等指标，以更全面地了解不同社会经济地位群体的特征。

3、对于企业来说，可以根据社会经济地位与总收入的正相关关系，改进定价策略，提供针对不同社会经济地位群体的产品和服务。

4、政策制定者可以关注社会经济地位较低的群体，提供针对性的支持措施，以促进社会公平和经济发

展。

3.3 对变量数据运用Logistic回归模型

使用Logistic回归模型对性别进行预测。自变量包括省份、宗教信仰、总收入、政治面貌、身高、体重、半年内是否上网、认为男性比女性能力更强和本人的社会经济地位。模型使用二项式逻辑回归（binomial）的形式进行拟合。通过拟合模型，可以得到每个自变量的系数，从而了解每个自变量对性别的影响程度。

代码

```
#################逻辑回归

#选取变量

data1 = dataf[,c(2,7,12,14,17,30,31,62,81,89)]

head(data1)

str(data1)

#预处理

colnames(data1) = c("省份","性别","宗教信仰","总收入","政治面貌","身高","体重","半年内是否上网","认为男性比女性能力更强","本人的社会经济地位")

data1$省份 = as.factor(data1$省份)

data1$省份 = as.numeric(data1$省份)


head(data1)


data1$性别 <- ifelse(data1$性别 >1 , 0, 1)

# 将数据集分为训练集和测试集

set.seed(123) # 设置随机种子，以便结果可重现

train_indices <- sample(1:nrow(data1), 0.7 * nrow(data1)) # 70% 的数据作为训练集

train_data <- data1[train_indices, ]
```

```
test_data <- data1[-train_indices, ]


head(data1)


# 建立逻辑回归模型

model <- glm(性别 ~ 本人的社会经济地位 + 宗教信仰 + 总收入+省份 + 政治面貌+身高+体重+半年内是否上网 +认为男性比女性能力更强, data = train_data, family = binomial())

summary(model)


# 在测试集上进行预测

predictions <- predict(model, newdata = test_data, type = "response")


# 计算准确率

accuracy <- sum(round(predictions) == test_data$性别) / nrow(test_data)


# 输出结果

cat("预测模型准确率：", accuracy, "\n")


# 分析系数或特征重要性

coef <- coef(model)

cat("模型系数：\n")

print(coef)
```

结果

```
> #预处理
> colnames(data1) = c("省份","性别","宗教信仰","总收入","政治面貌","身高","体重","半年内是否上网","认为男性比女性能力更强","本人的社会经济地位")
> data1$省份 = as.factor(data1$省份)
> data1$省份 = as.numeric(data1$省份)
>
> head(data1)
  省份 性别 宗教信仰 总收入 政治面貌 身高 体重 半年内是否上网 认为男性比女性能力更强 本人的社会经济地位
1    2    2        1  96000        1  160  118              1                  1                  4
2    2    2        1  60000        4  160  120              1                  1                  2
3    2    1        1  96000        4  172  156              2                  2                  4
4    2    1        1 160000        1  172  130              1                  2                  4
5    2    2        1  48000        1  158  115              1                  2                  3
6    2    1       12 600000        1  180  150              1                  4                  5
>
```

```
>
> data1$性别 <- ifelse(data1$性别 >1 , 0, 1)
> # 将数据集分为训练集和测试集
> set.seed(123) # 设置随机种子，以便结果可重现
> train_indices <- sample(1:nrow(data1), 0.7 * nrow(data1)) # 70% 的数据作为训练集
> train_data <- data1[train_indices, ]
> test_data <- data1[-train_indices, ]
>
> head(data1)
  省份 性别 宗教信仰 总收入 政治面貌 身高 体重 半年内是否上网 认为男性比女性能力更强 本人的社会经济地位
1    2    0        1  96000        1  160  118              1                  1                  4
2    2    0        1  60000        4  160  120              1                  1                  2
3    2    1        1  96000        4  172  156              2                  2                  4
4    2    1        1 160000        1  172  130              1                  2                  4
5    2    0        1  48000        1  158  115              1                  2                  3
6    2    1       12 600000        1  180  150              1                  4                  5
>
```

```
> # 建立逻辑回归模型
> model <- glm(性别 ~ 本人的社会经济地位 + 宗教信仰 + 总收入+省份 + 政治面貌+身高+体重+半年内是否上网 +认为男性比女性能力更强, data = train_data,
> summary(model)

Call:
glm(formula = 性别 ~ 本人的社会经济地位 + 宗教信仰 +
    总收入 + 省份 + 政治面貌 + 身高 + 体重 + 半年内是否上网 +
    认为男性比女性能力更强, family = binomial(), data = train_data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.570e-01  9.171e-02  -3.893 9.92e-05 ***
本人的社会经济地位 -2.466e-03  2.028e-03  -1.216  0.22406
宗教信仰          -4.187e-02  8.922e-03  -4.693 2.70e-06 ***
总收入            -2.780e-08  9.088e-09  -3.059  0.00222 **
省份              -6.965e-03  4.643e-03  -1.500  0.13357
政治面貌           1.896e-02  8.051e-03   2.355  0.01851 *
身高              -6.867e-04  2.831e-04  -2.426  0.01528 *
体重               3.216e-03  4.936e-04   6.515 7.28e-11 ***
半年内是否上网     6.326e-03  7.913e-03   0.799  0.42401
认为男性比女性能力更强 -3.488e-03  2.651e-03  -1.316  0.18834
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7831.1  on 5702  degrees of freedom
Residual deviance: 7722.1  on 5693  degrees of freedom
AIC: 7742.1

Number of Fisher Scoring iterations: 4
```

```
>
> # 在测试集上进行预测
> predictions <- predict(model, newdata = test_data, type = "response")
>
> # 计算准确率
> accuracy <- sum(round(predictions) == test_data$性别) / nrow(test_data)
>
> # 输出结果
> cat("预测模型准确率：", accuracy, "\n")
预测模型准确率： 0.5541922
>
> # 分析系数或特征重要性
> coef <- coef(model)
> cat("模型系数：\n")
模型系数：
> print(coef)
          (Intercept)      本人的社会经济地位              宗教信仰              总收入              省份              政治面貌
        -3.569952e-01       -2.466194e-03       -4.186845e-02       -2.780429e-08       -6.964799e-03        1.896371e-02
                  身高                  体重 半年内是否上网 认为男性比女性能力更强
        -6.866716e-04        3.215804e-03        6.326234e-03       -3.487645e-03
>
```

分析

根据此逻辑回归结果可知：

1、模型中，变量的系数代表了对应变量对性别的影响程度。系数的正负表明了变量与性别之间的关系方向，而绝对值的大小表示了影响的强度。

2、在这个模型中，宗教信仰、总收入、政治面貌和体重的系数是显著的（P值小于0.05），因此它们对性别有相关影响。

3、宗教信仰的系数为-0.042，表示信仰宗教与女性性别之间存在负相关。可能是因为某些宗教对性别角色有特定的规定。

4、总收入的系数为-2.78e-08，虽然非常小，但在统计意义上仍然是显著的。这意味着总收入可能与性别存在微弱的负相关关系。

5、政治面貌的系数为0.019，表示政治面貌与女性性别之间存在正相关。这可能是由于特定的政治观点或社会文化背景导致的。

6、体重的系数为0.0032，表示体重与女性性别之间存在正相关。这可能是由于生理差异或社会期望导致的。

7、其他变量（本人的社会经济地位、省份、半年内是否上网、认为男性比女性能力更强）的系数在统计意义上不显著，即它们与性别的关系不够明确。

基于以上分析结果，我们可以提出以下实用性建议：

1、在社会、文化和政治环境中，要重视宗教信仰对性别角色和性别认同的影响。特别是在制定政策和开展社会活动时，需要尊重和包容不同宗教信仰的个体。

2、虽然总收入对性别的影响很小，但仍然值得关注。在劳动力市场和工资制定中，要确保性别不成为决定总收入的因素之一，促进性别平等。

3、政治面貌和体重与性别的相关性可能涉及到深层次的社会观念和偏见。这提示我们需要加强性别教育，消除性别歧视和刻板印象。

3.4 数据可视化分析

代码

```
library(ggplot2)


# 可视化分析省份与性别的关系
ggplot(data1, aes(x = 省份, fill = factor(性别))) +
 geom_bar() +
 labs(x = "省份", y = "人数", fill = "性别") +
 theme_minimal()


# 可视化分析宗教信仰与性别的关系
ggplot(data1, aes(x = factor(宗教信仰), fill = factor(性别))) +
 geom_bar() +
 labs(x = "宗教信仰", y = "人数", fill = "性别") +
 theme_minimal()


# 可视化分析政治面貌与性别的关系
ggplot(data1, aes(x = factor(政治面貌), fill = factor(性别))) +
 geom_bar() +
 labs(x = "政治面貌", y = "人数", fill = "性别") +
 theme_minimal()


# 可视化分析认为男性比女性能力更强与性别的关系
```

```
ggplot(data1, aes(x = factor(认为男性比女性能力更强), fill = factor(性别))) +
 geom_bar() +
 labs(x = "认为男性比女性能力更强", y = "人数", fill = "性别") +
 theme_minimal()

# 可视化分析本人的社会经济地位与性别的关系
ggplot(data1, aes(x = factor(本人的社会经济地位), fill = factor(性别))) +
 geom_bar() +
 labs(x = "本人的社会经济地位", y = "人数", fill = "性别") +
 theme_minimal()
```

结果

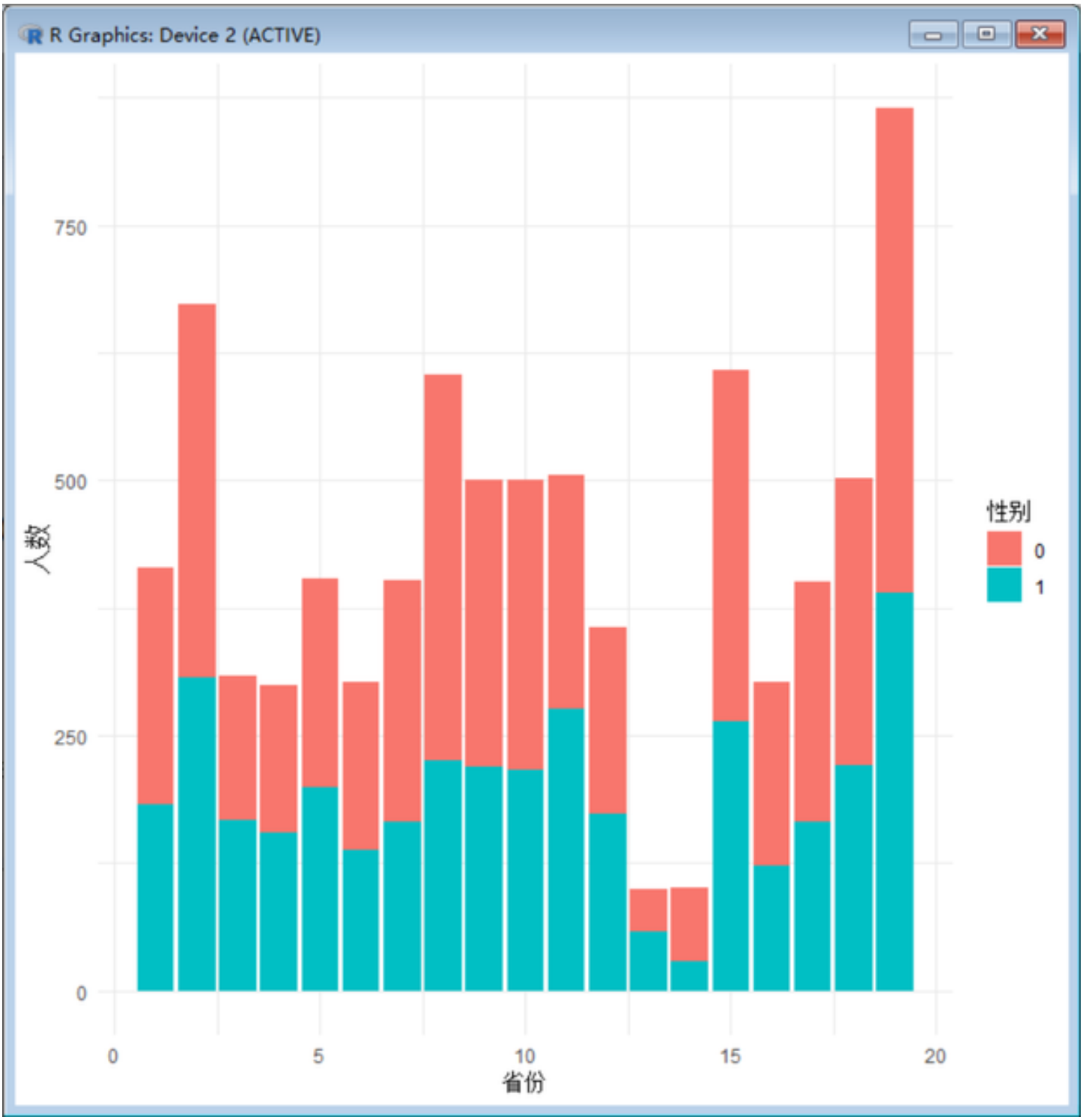


图1 省份与性别的关系

由此图可知在各个省中参加问卷调查的男女比例其实并不偏向某一方。

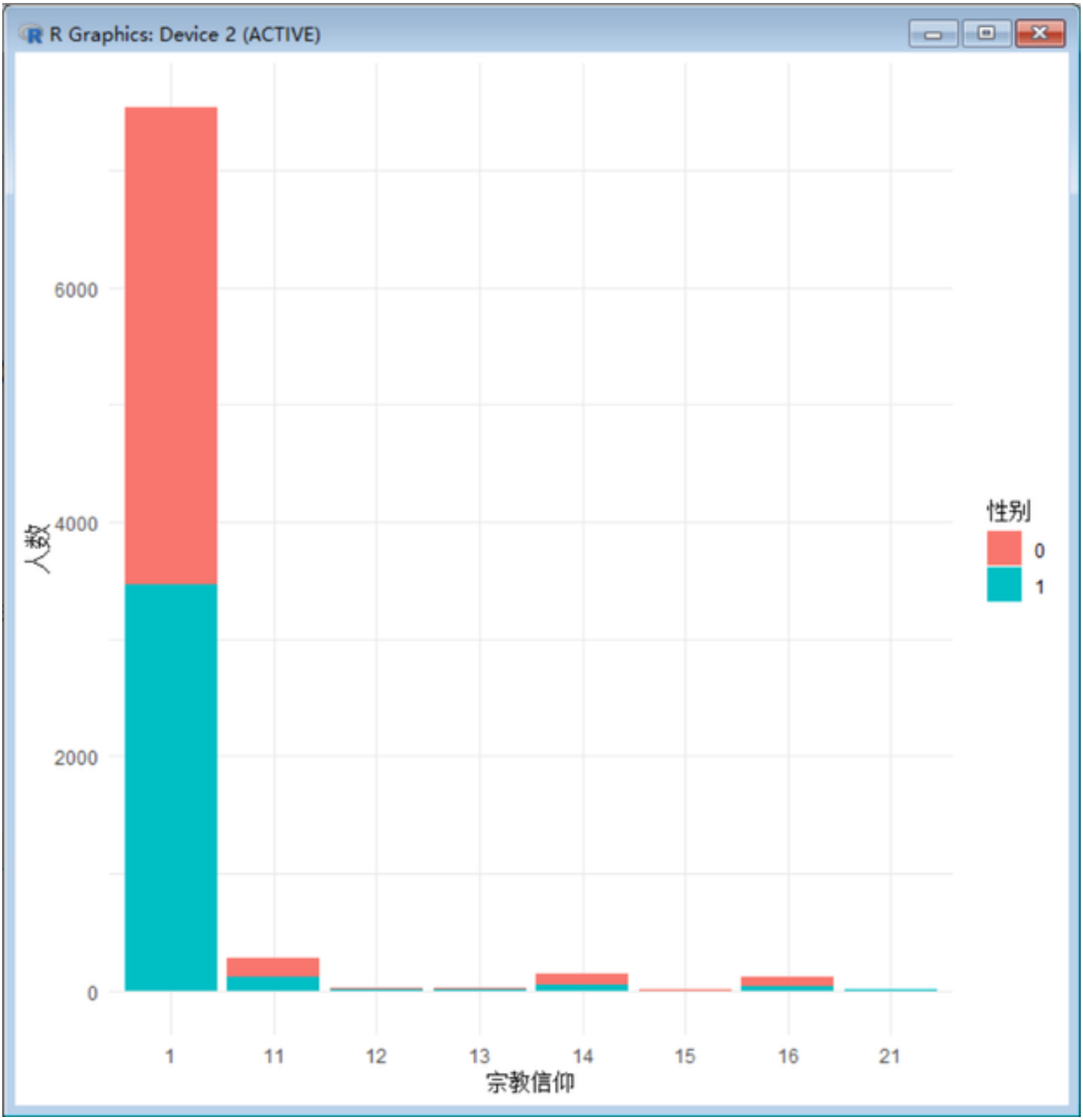


图2 宗教信仰与性别的关系

由此图可知参加问卷中的人数在最多人不信奉总教。部分人信奉佛教等。

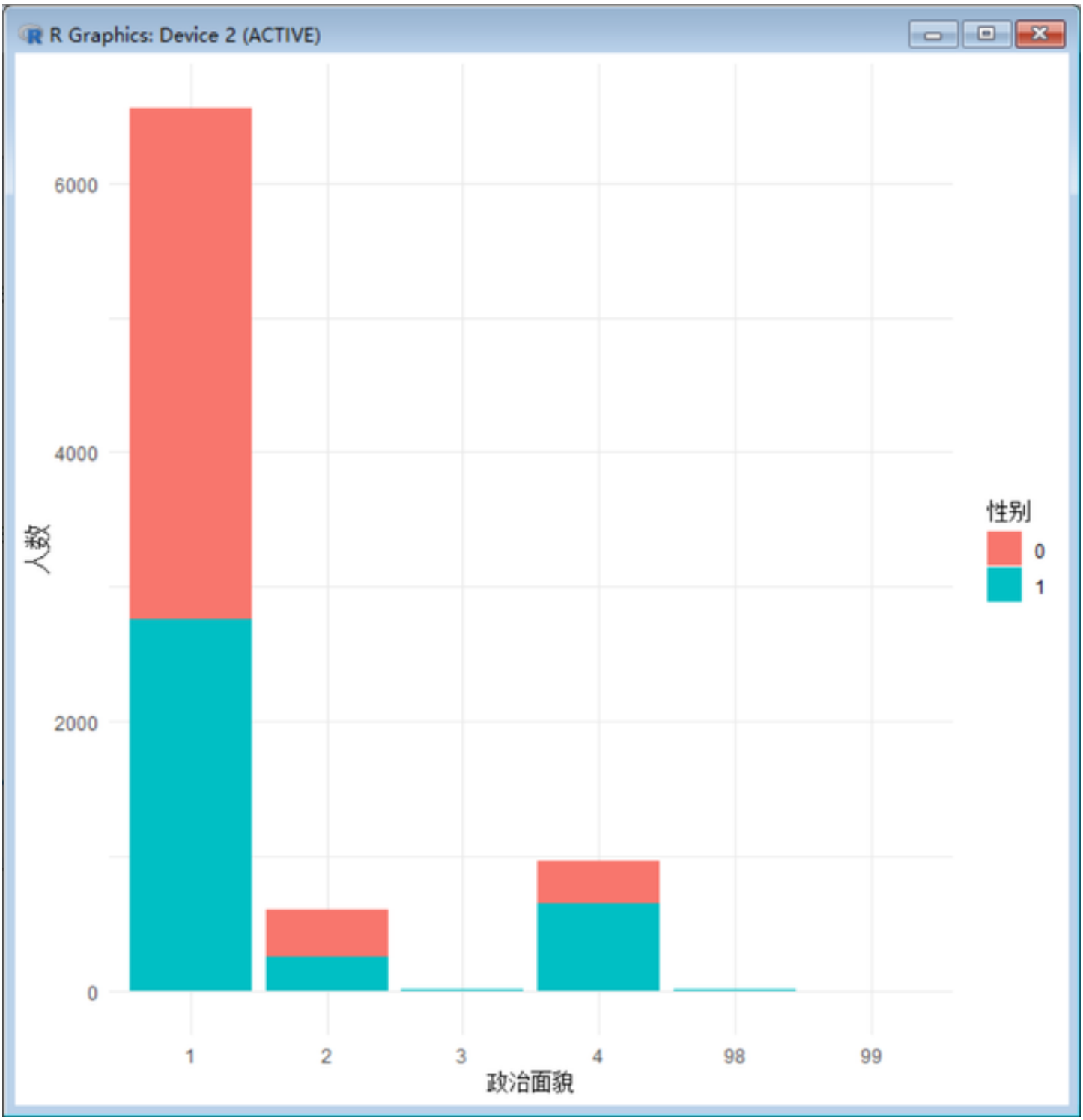


图3 政治面貌与性别的关系

由此图可知参加问卷中的人中群众、共青团员和党员居多，男女几乎均衡。

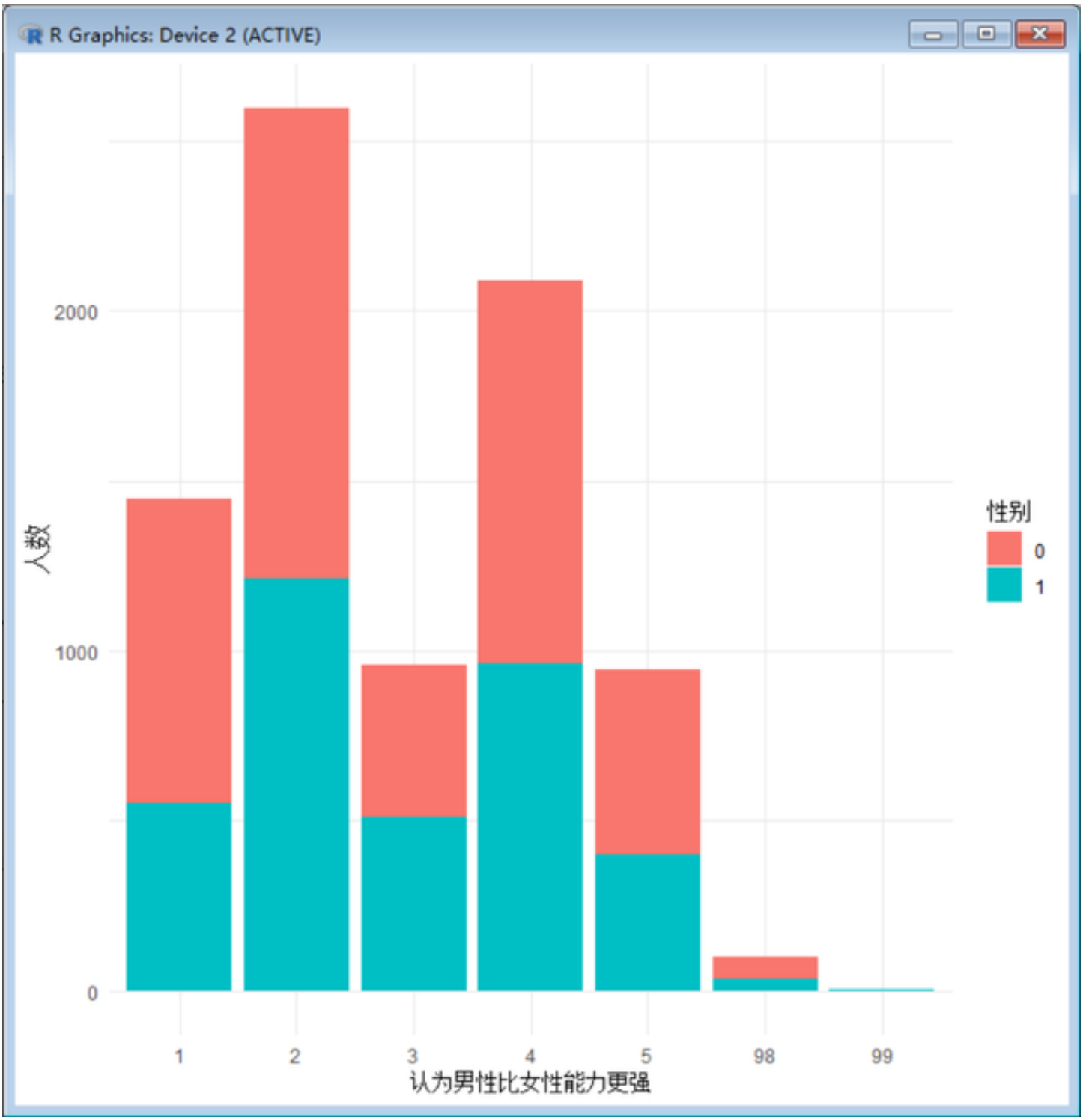


图4 认为男性比女性能力更强与性别的关系

由此图可知认为比较不同意的人数最多，其中男女均衡。

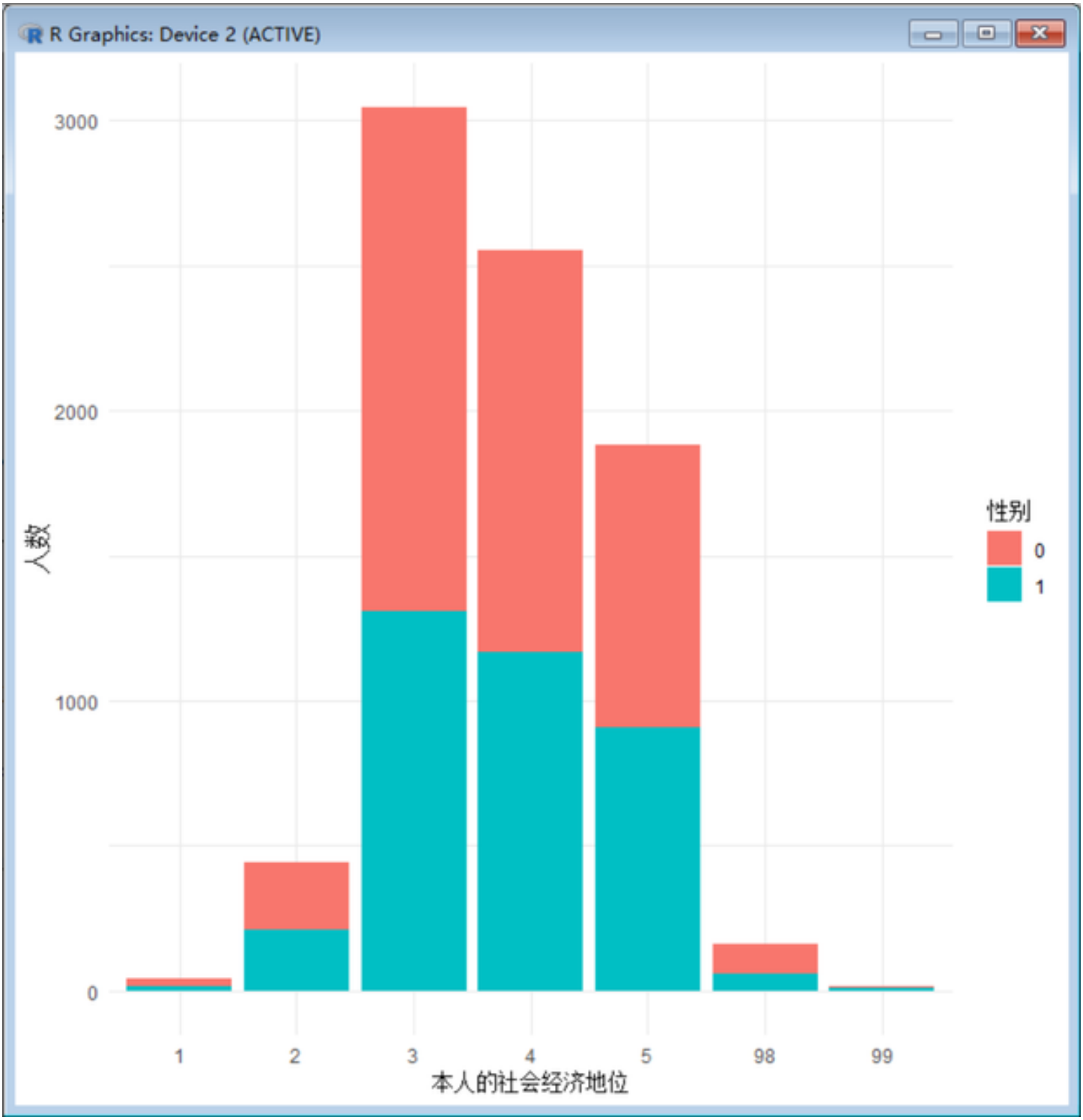


图5 本人的社会经济地位与性别的关系

由此图可知社会经济地位处于中层人数最多，然后是中下层等，其中处于中层来看，男性居多。

4、总结

1. 性别比例：参与调查的男女比例相对平衡，男性占45.15%，女性占54.85%。

2. 宗教信仰：参与调查的人群宗教信仰较多样，其中不信奉任何宗教的人数最多。
3. 收入分布：参与调查的人群收入分布较广，个体平均收入为1,046,361。
4. 政治面貌：参与调查的人群的政治面貌比较多样化，其中群众、共青团员和党员占多数。
5. 身高和体重：身高和体重在参与调查的人群中有一定的分布范围，没有明显的偏倚。
6. 上网情况：大部分人在半年内上过互联网。
7. 男女能力认知：对于男性与女性能力的认知分布较为均匀。
8. 社会经济地位：参与调查的人群中，社会经济地位集中在3位置上，处于中层的人数最多。

基于以上分析，我们可以向企业或政策制定者提出以下建议：

1. 市场划分时要考虑性别和宗教信仰之间的关系，以更好地满足不同群体的需求。
2. 进行社会经济调查时应关注身高、体重和认为男性比女性能力更强等指标，以更全面地了解不同社会经济地位群体的特征。
3. 针对不同社会经济地位群体提供针对性的产品和服务，改进定价策略。
4. 政策制定者可以关注社会经济地位较低的群体，提供针对性的支持措施，促进社会公平和经济发展。

另外，根据逻辑回归结果，我们发现宗教信仰、总收入、政治面貌和体重对性别有相关影响。因此，我们提出以下实用性建议：

1. 在社会、文化和政治环境中要重视宗教信仰对性别角色和性别认同的影响，尊重和包容不同宗教信仰的个体。
2. 在劳动力市场和工资制定中要确保性别不成为决定总收入的因素之一，促进性别平等。
3. 加强性别教育，消除性别歧视和刻板印象，以改变政治面貌和体重与性别的相关性。

综上所述，以上分析结果为企业和政策制定者提供了有关性别、宗教信仰、社会经济地位和其他相关因素之间关系的信息和建议，以便更准确地了解人群特征并采取适当的措施。