Code

HW 2: Предсказание отклика различных групп респондентов на новую макркетинговую кампанию

Пясецкая Вероника, vvpyasetskaya

Задача

Спрогнозировать, какие группы респондентов окажутся наиболее восприимчивы к новой маркетинговой кампании, проводимой в сети магазинов. Иными словами, определить целевую аудиторию предстоящей кампании, выявив закономерности в реакции респондентов и предсказав их отклик на неё. Выводы данного исследования могут быть использованы для более точечного и предметного воздействия на клиента с помощью маркетинговых инструментов.

Загрузка данных и преобразование

Code

В процессе первичной обработки данных перемнные типа character и ряд числовых переменных были преобразованы в факторные. На дальнейших этапах анализа была удалена переменная ID.Так же были выявлены и удалены наблюдения имеющие неизвестные значения в важной для исследования переменной Income. Из наблюдений по переменным Age и Income были удалены выбросы, так же по переменной Marital_Status были удалены ответы “YOLO” и “Absurd”. Переменная Dt_Customer была преобразована в дату вида: день-месяц-год, и с её помощью была создана новая переменная Age (возраст респондентов). Кроме того, при расмотрении Вопроса №6 была создана переменная food_percent, представляющая долю дохода, которую респондент тратит на продукты питания.

Code

Исследовательские вопросы и тесты

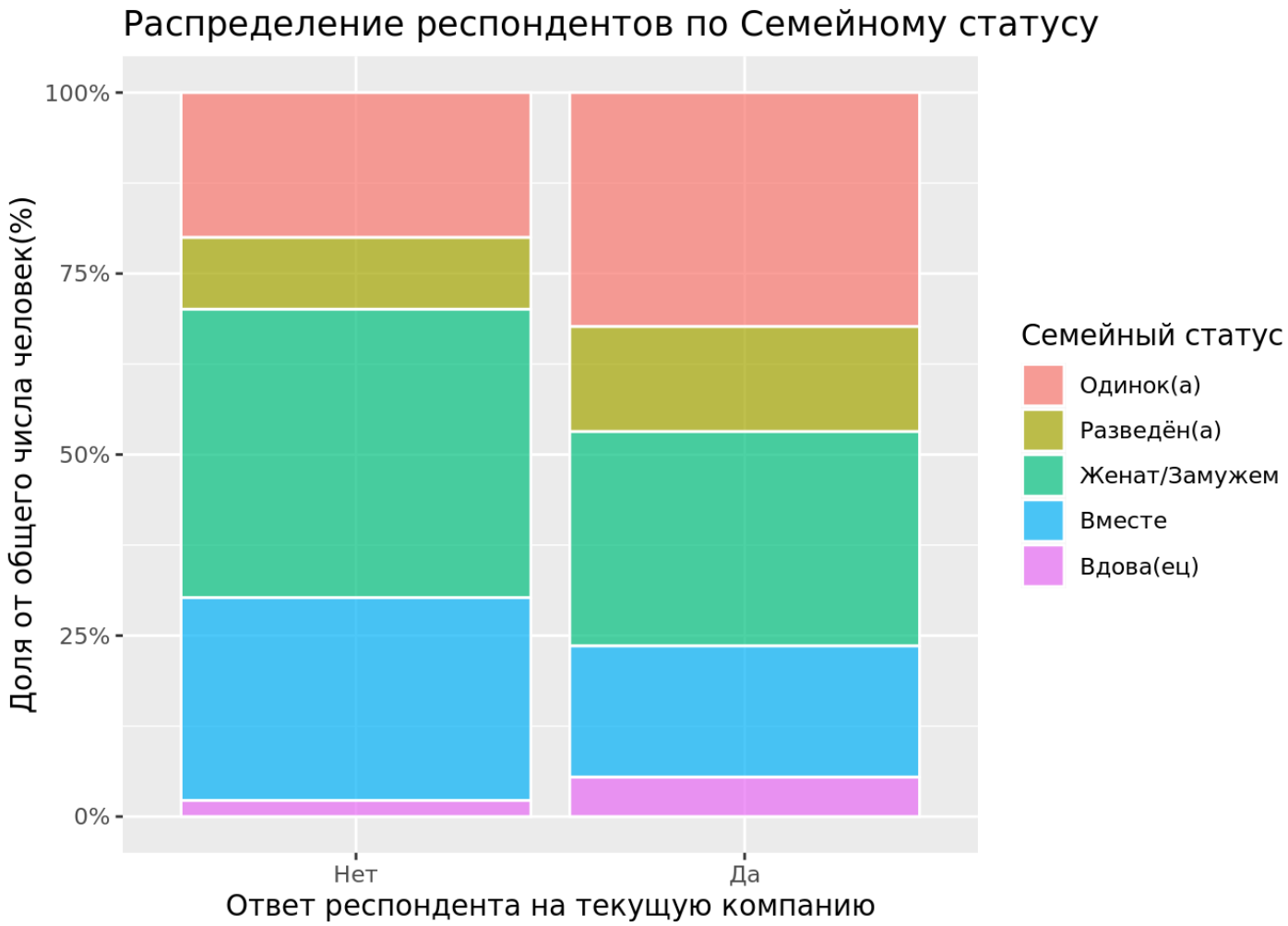
Проведём разведовательный анализ

Выясним, как распределены клиенты по возрасту

Code

По семейному статусу?

Code

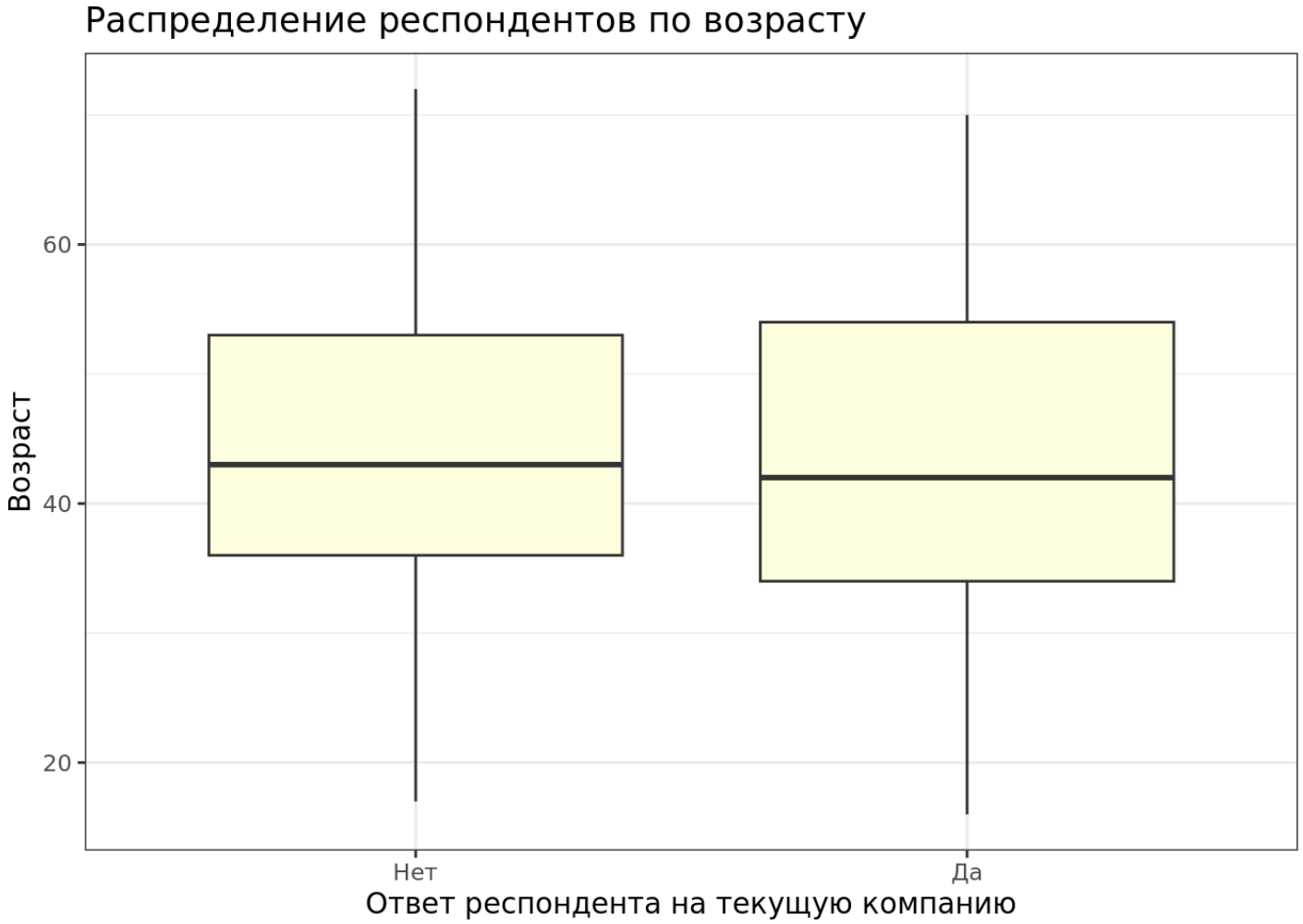


Из графика видно, что распределение по семейному статусу неравномерно. Среди откликнувшихся на кампанию больше Одиноких, Разведённых и Вдовцов и меньше Женатых/Состоящих в отношениях (Вместе) людей.

Теперь мы имеем представление о демографическом распределении респондентов. Сформулируем вопросы и изучим зависимости между переменными подробней.

Вопрос №1: Зависит ли Ответ на кампанию от Возраста респондента?

Code



Code

Code

Итог: По графику можно сказать, что разница между средним возрастом обеих групп, составляющая меньше 1 года, визуально незначительна.

Сформулируем гипотезы и убедимся с помощью теста

`H0`: Средний возраст людей откликнувшихся на текущую компанию респондентов не отличается от среднего возраста не откликнувшихся

`H1`: Существует разница в среднем возрасте людей откликнувшихся и не откликнувшихся на компанию

Переменная Age имеет распределение близкое к нормальному, используем t-test

Code

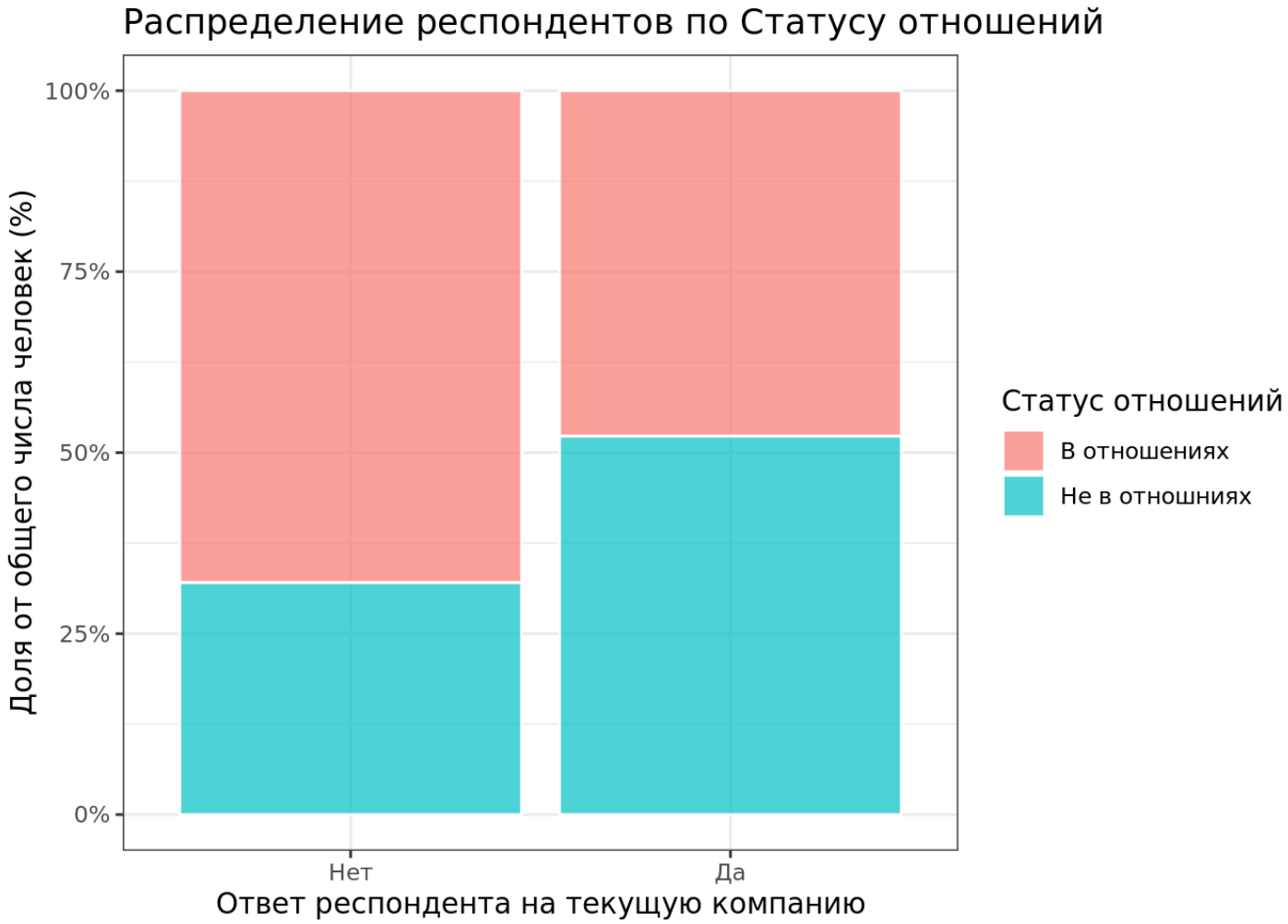
```
##
##  Welch Two Sample t-test
##
## data:  Age by Response
## t = 0.76868, df = 539.52, p-value = 0.4424
## alternative hypothesis: true difference in means between group 0 and group 1 is
not equal to 0
## 95 percent confidence interval:
##  -0.9187522  2.1000272
## sample estimates:
## mean in group 0 mean in group 1
##        43.92901        43.33837
```

Вывод: На уровне значимости 0.05 статистически значимой разницы не обнаружено. Мы не можем отвергнуть H0.

Вопрос №2: Действительно ли люди, состоящие в отношениях, реже отвечают на кампанию, чем те, у кого нет партнёра?

При демографическом анализе было выявлено, что среди откликнувшихся на кампанию больше Одиноких, Разведённых и Вдовцов и меньше Женатых/Состоящих в отношениях (Вместе) людей. Таким образом, попробуем разделить их на две группы: В отношениях / Не в отношениях и посмотреть на разницу:

Code



Различие между группами стало намного заметней, посчитаем разницу

Code

Итог: Выявленный размер эффекта составляет 20%

`H0`: Разницы между Статусом отношений положительно и отрицательно ответивших на компанию респондентов не сущесвтует

`H1`: Существует разница между Статусом отношений положительно и отрицательно ответивших на компанию респондентов

Для двух категориальных переменных используем тест Хи-квадрат:

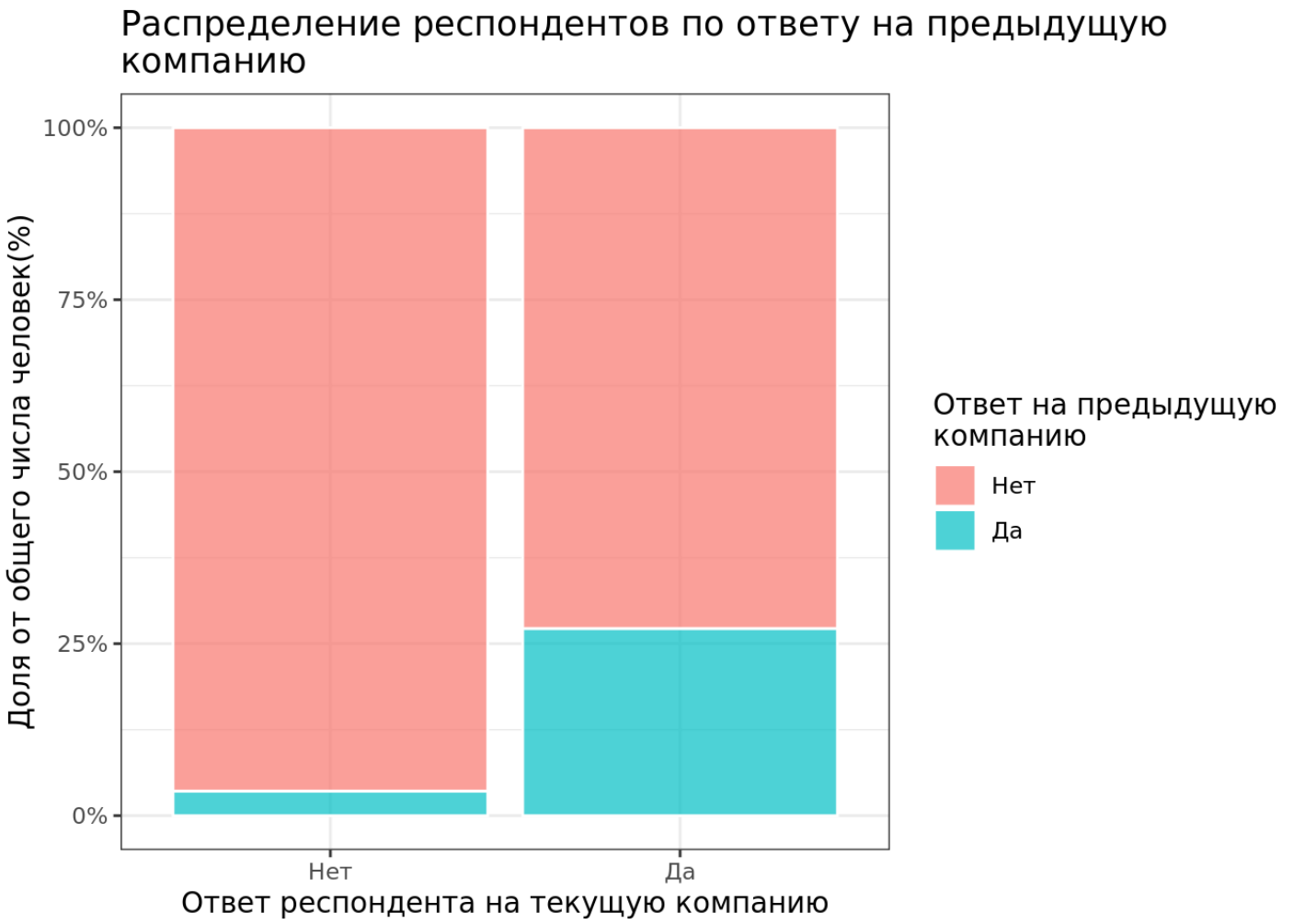
Code

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  marketing$Relationships and marketing$Response
## X-squared = 42.529, df = 1, p-value = 6.966e-11
```

Вывод: p-value достаточно мало. Разница между долями респондентов статистически значима (на уровне значимости 0.05). Мы можем отвергнуть H0 и использовать переменную Relationships подбирая разбиения при построении модели.

Вопрос №3: Правда ли, что люди положительно ответившие на предыдущую кампанию охотней откликаются на текущую?

Code



Code

Итог: Вычисленная явным образом разница составляет 24 %

Проверим с помощью теста

`H0`: Разница между долями респондентов принявших предложение в предыдущую компанию, среди положительно и отрицательно ответивших на текущую компанию равна 0

`H1`: Разница между долями респондентов принявших предложение в предыдущую компанию, среди положительно и отрицательно ответивших на текущую компанию не равна 0

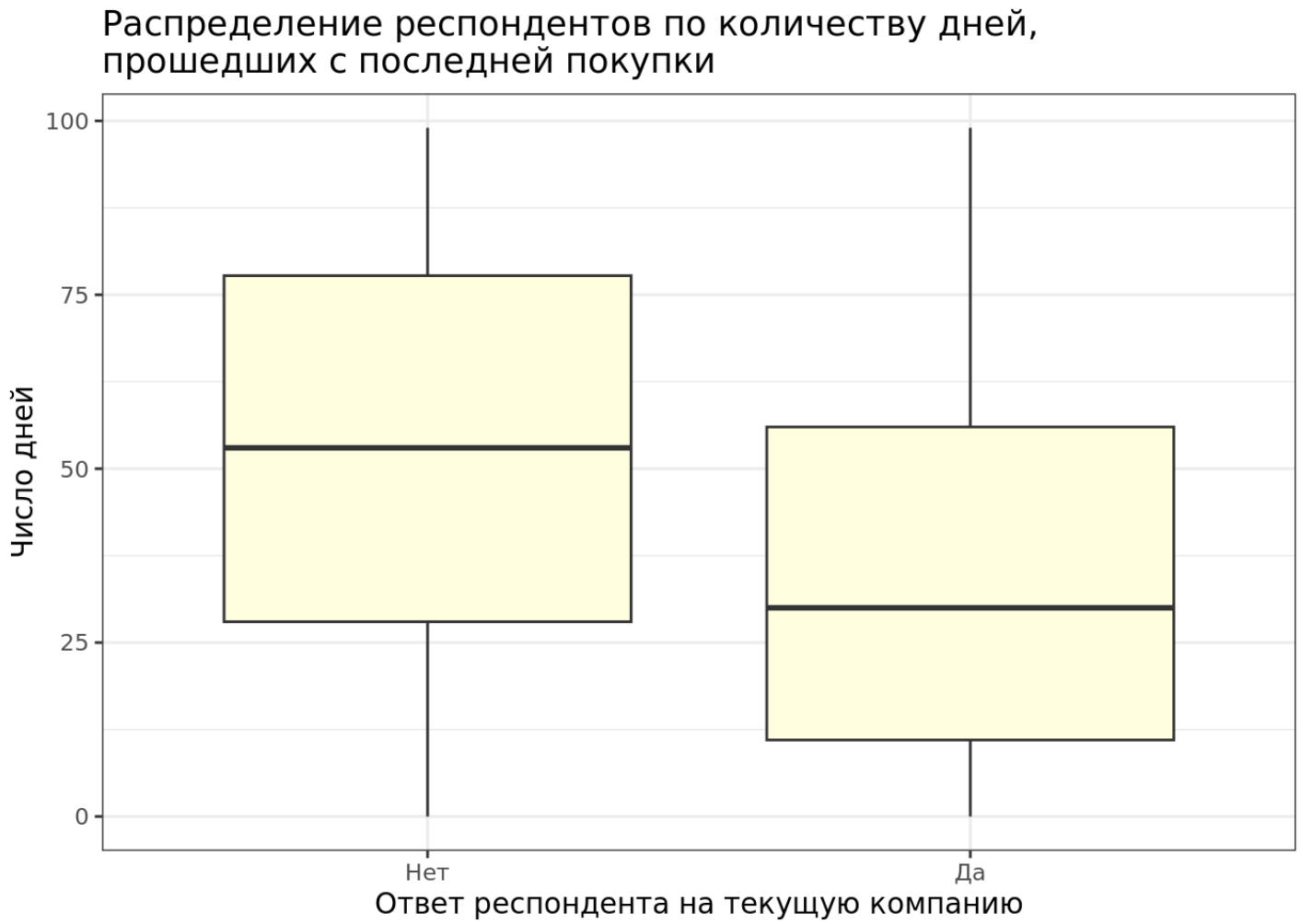
Code

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  marketing$Response and marketing$AcceptedCmp
## X-squared = 158.48, df = 1, p-value < 2.2e-16
```

Вывод: p-value очень мало. На уровне значимости 0.05 существует статистически значимая разница между долями респондентов. Мы можем отвергнуть H0 и использовать переменную AcceptedCmp подбирая разбиения при построении модели.

Вопрос №4: Действительно ли чем меньше времени прошло с последней покупки, совершённой респондентом, тем охотней он принимает участие в текущей кампании?

Code



Визуально разница достаточно заметна. Если точнее:

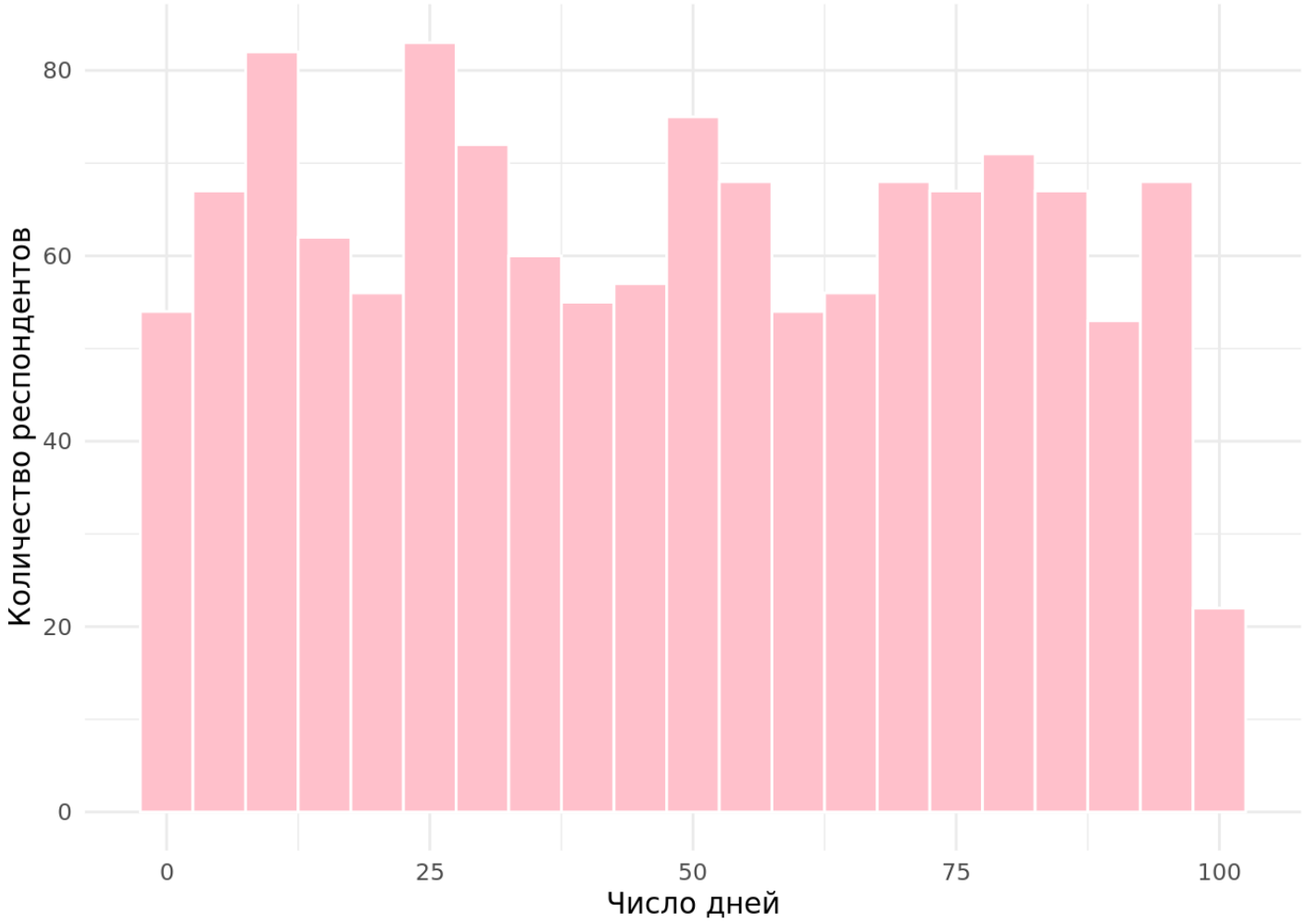
Code

Итог: Разница в среднем количестве дней, прошедших с последней покупки составляет 17 дней. Разница медианных значений составила 23 дня.

`H0`: Существует разница между количестовом дней, прошедших с последней покупки среди согласившихся и отказавшихся от участия в компании репондентов

`H1`: Разницы между количестовом дней, прошедших с последней покупки среди согласившихся и отказавшихся от участия в компании репондентов не существует

Code



Распределение не нормальное, используем тест перестановок

Code

```
##
##  Asymptotic General Independence Test
##
## data:  Recency by Response (0, 1)
## Z = 9.1058, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Вывод: Между количеством прошедших с последней покупки дней существует статистически значимая разница (на уровне значимости 0.05). Мы можем отвергуть H0 и учесть переменную Recency при построении модели.

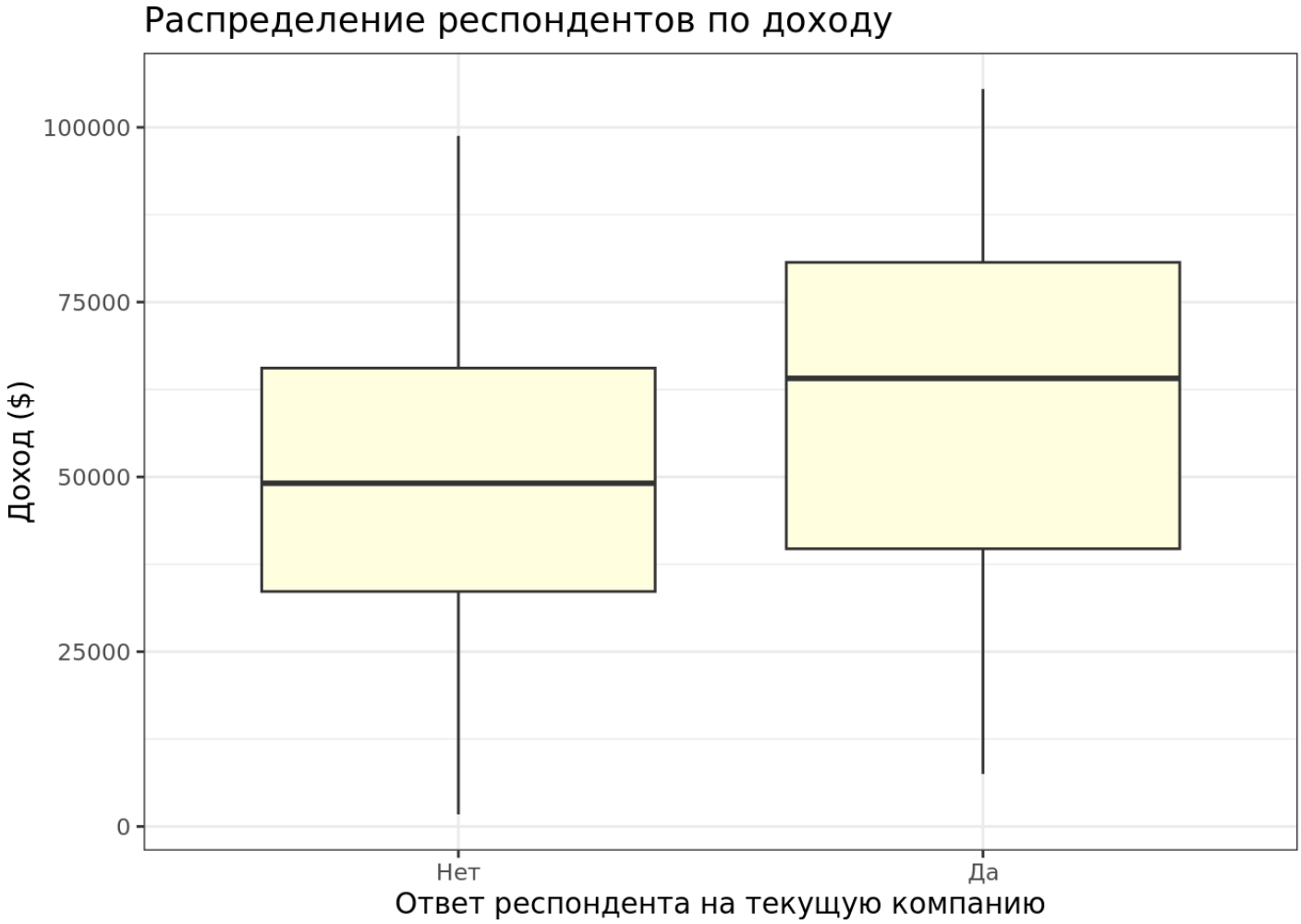
Вопрос №5: Участвуют ли чаще в текущей кампании клиенты с бОльшим доходом?

Удалим выбросы сильно влияющие на результат

Code

Посмотрим на распределение

Code



Code

Итог: Выявлена разница среднего и медианного дохода, составляющая 10 847$ и 14 998$ соответственно

`H0`: Разница в размере дохода согласившихся и отказавшихся от участия в компании респондентов равна 0

`H1`: Разница в размере дохода согласившихся и отказавшихся от участия в компании респондентов не равна 0

Code

Распределение визуально неочевидно, используем тест перестановок

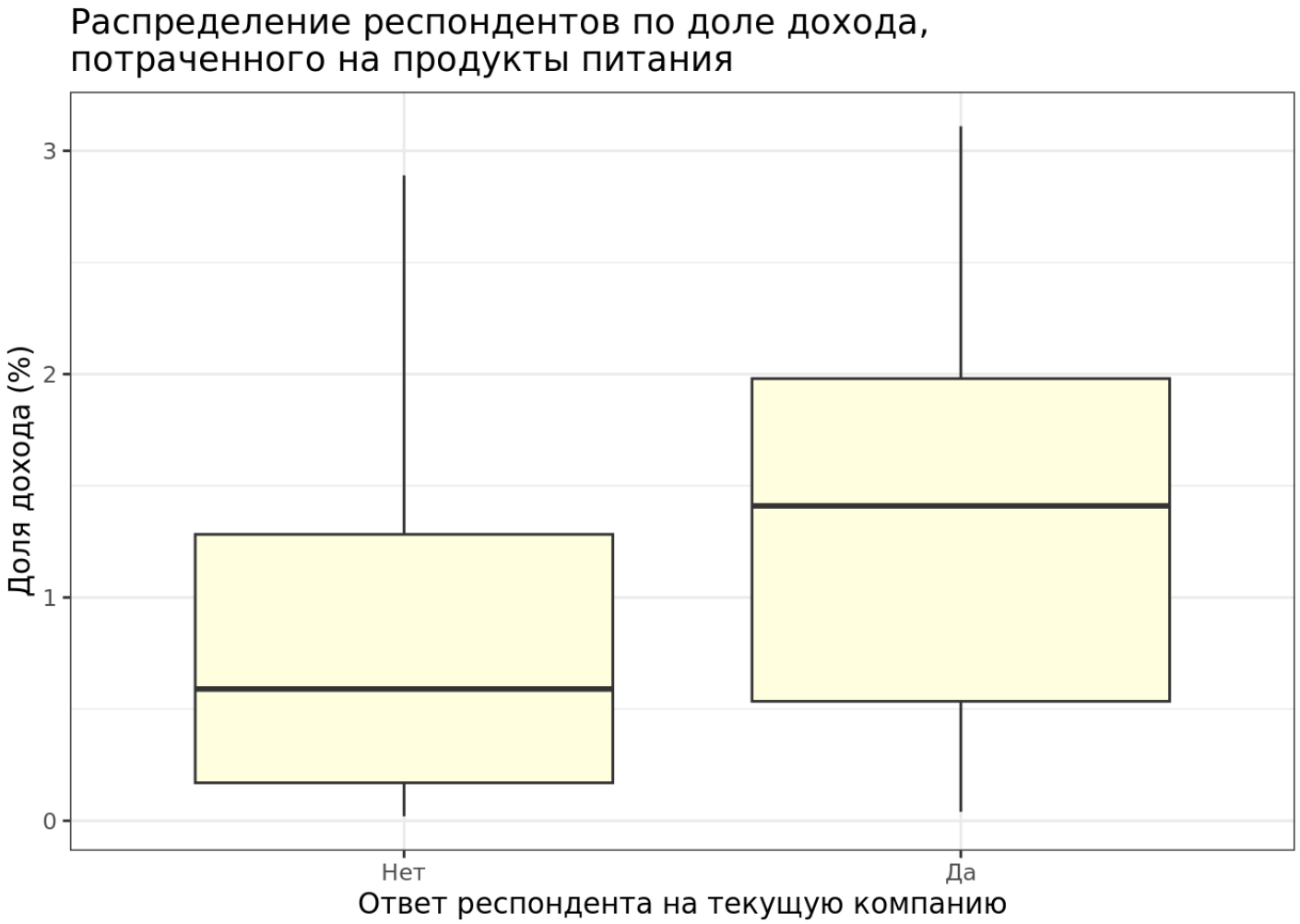
Code

```
##
##  Asymptotic General Independence Test
##
## data:  Income by Response (0, 1)
## Z = -7.9581, p-value = 1.747e-15
## alternative hypothesis: two.sided
```

Вывод: Разница в размере дохода между двумя группами существет и является статистически значимой (на уровне значимости 0.05). Мы можем отвернуть H0 и использовать переменную Income при построении модели. Однако, возможно более справедливой для оценки переменной была бы доля дохода, потраченного непосредственно на продукты питания. Проверим предположение

Вопрос №6: Связана ли доля дохода которую репсондент тратит на продукты питания с его ответом на предложение в текущей кампании?

Code



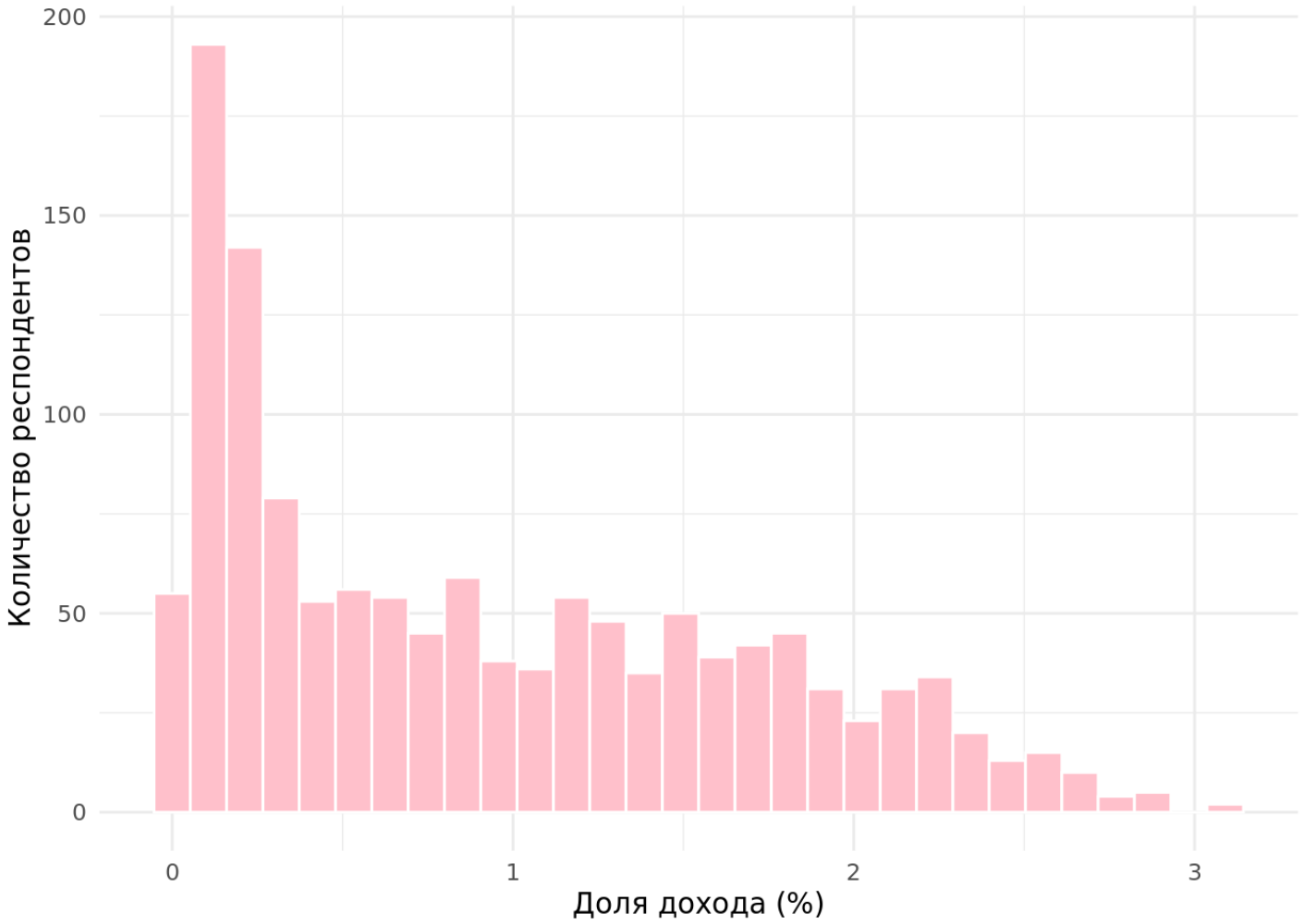
Code

Итог: Разница в средней и медианной доле дохода, уходящей на продукты составила 0.52% и 0.82% соответственно

`H0`: Существует рзница в доле дохода, потраченного на продукты питания для респондентов ответивших и не ответивших на текщую компанию

`H1`: Разницы в доле дохода, потраченного на продукты питания для респондентов ответивших и не ответивших на текщую компанию не существует

Code



Используем тест перестановок

Code

```
##
##  Asymptotic General Independence Test
##
## data:  food_percent by Response (0, 1)
## Z = -10.765, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Вывод: Выявленна статистически знаимая разница (на уровне значимости 0.05) в доле дохода, потраченного на продукты питания между двумя группами. Следовательно, мы можем отвергнуть H0 и использовать эту переменную, как критерий разбиения.

Посмотрим на данные иначе

Code

```
##
## --------Summary descriptives table by 'Response'---------
##
## _____________________________________________________________
##                            0             1       p.overall
##                         N=980         N=331
## -------------------------------------------------------------
## ID                   5630 (3299)   5398 (3150)    0.252
## Year_Birth           1969 (11.5)   1969 (12.3)    0.724
## Education:                                       <0.001
##     2n Cycle         89 (9.08%)    22 (6.65%)
##     Basic            29 (2.96%)     2 (0.60%)
##     Graduation       509 (51.9%)   151 (45.6%)
##     Master           153 (15.6%)   56 (16.9%)
##     PhD              200 (20.4%)   100 (30.2%)
## Marital_Status:                                  <0.001
##     Одинок(а)        197 (20.1%)   107 (32.3%)
##     Разведён(а)      97 (9.90%)    48 (14.5%)
##     Женат/Замужем    392 (40.0%)   98 (29.6%)
##     Вместе           272 (27.8%)   60 (18.1%)
##     Вдова(ец)        22 (2.24%)    18 (5.44%)
## Income              49341 (20085) 60188 (23232)  <0.001
## Kidhome              0.48 (0.55)   0.34 (0.49)   <0.001
## Teenhome             0.55 (0.55)   0.31 (0.49)   <0.001
## Recency              52.3 (28.7)   35.3 (27.6)   <0.001
## MntWines              268 (307)     503 (429)    <0.001
## MntFruits            23.0 (36.4)   37.9 (45.9)   <0.001
## MntMeatProducts       141 (194)     295 (288)    <0.001
## MntFishProducts      33.7 (51.5)   51.4 (61.1)   <0.001
## MntSweetProducts     24.8 (39.2)   38.4 (46.2)   <0.001
## NumDealsPurchases    2.31 (1.72)   2.34 (2.11)    0.857
## NumWebPurchases      3.88 (2.67)   5.07 (2.57)   <0.001
## NumStorePurchases    5.71 (3.24)   6.08 (3.08)    0.063
## AcceptedCmp:                                     <0.001
##     0                945 (96.4%)   241 (72.8%)
##     1                35 (3.57%)    90 (27.2%)
## Complain:                                         1.000
##     0                971 (99.1%)   328 (99.1%)
##     1                 9 (0.92%)     3 (0.91%)
## Age                  43.9 (11.6)   43.3 (12.3)    0.437
## Relationships:                                   <0.001
##     В отношениях     664 (67.8%)   158 (47.7%)
##     Не в отношниях   316 (32.2%)   173 (52.3%)
## food_percent         0.79 (0.69)   1.31 (0.82)   <0.001
## -------------------------------------------------------------
```

Перекосы в данных в основном соотствуют выявленным переменным, что позволяет подтвердить заключения сделанные выше

Предсказание отклика на кампанию

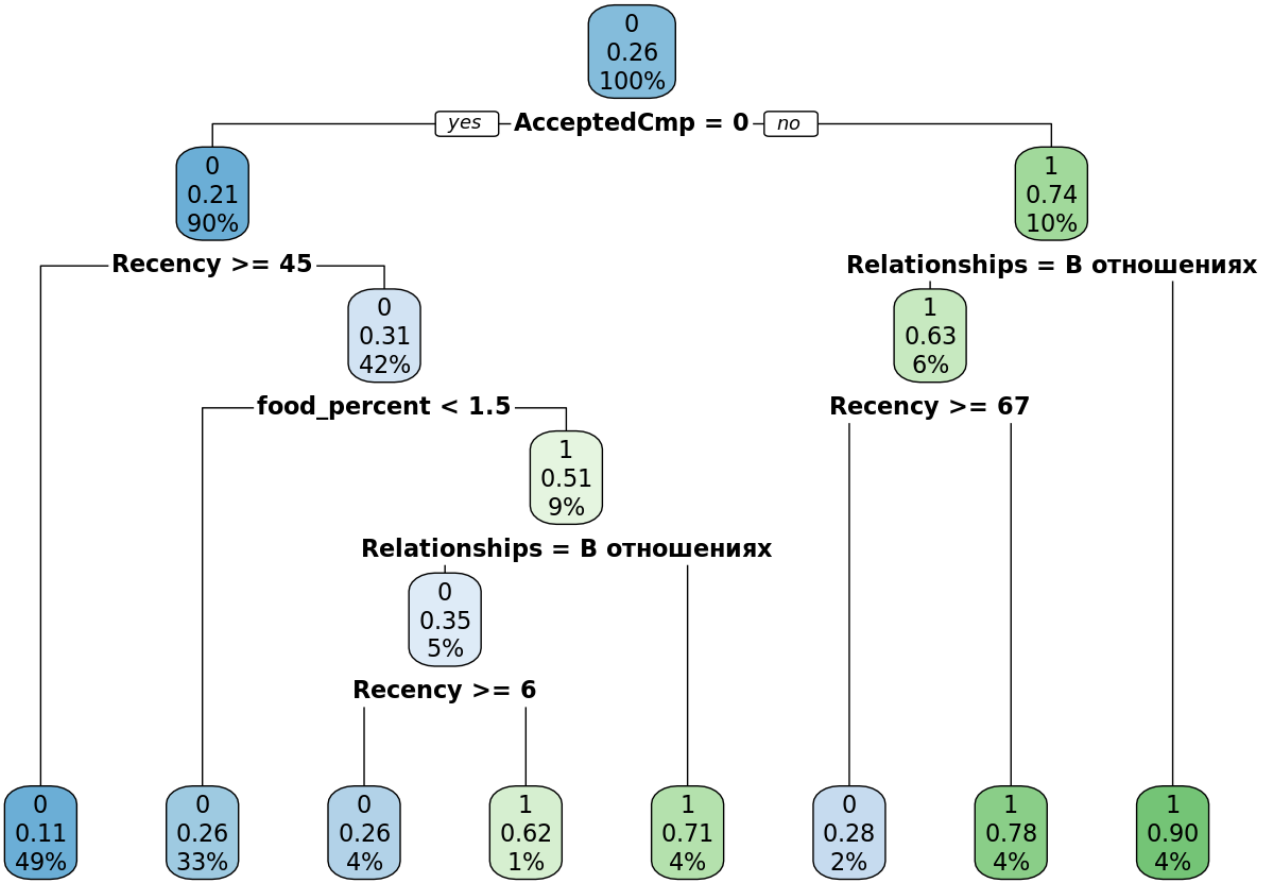
Посчитаем насколько однородны по Ответам на компанию наши исходные данные

Code

Исходный Gini index: 0.37

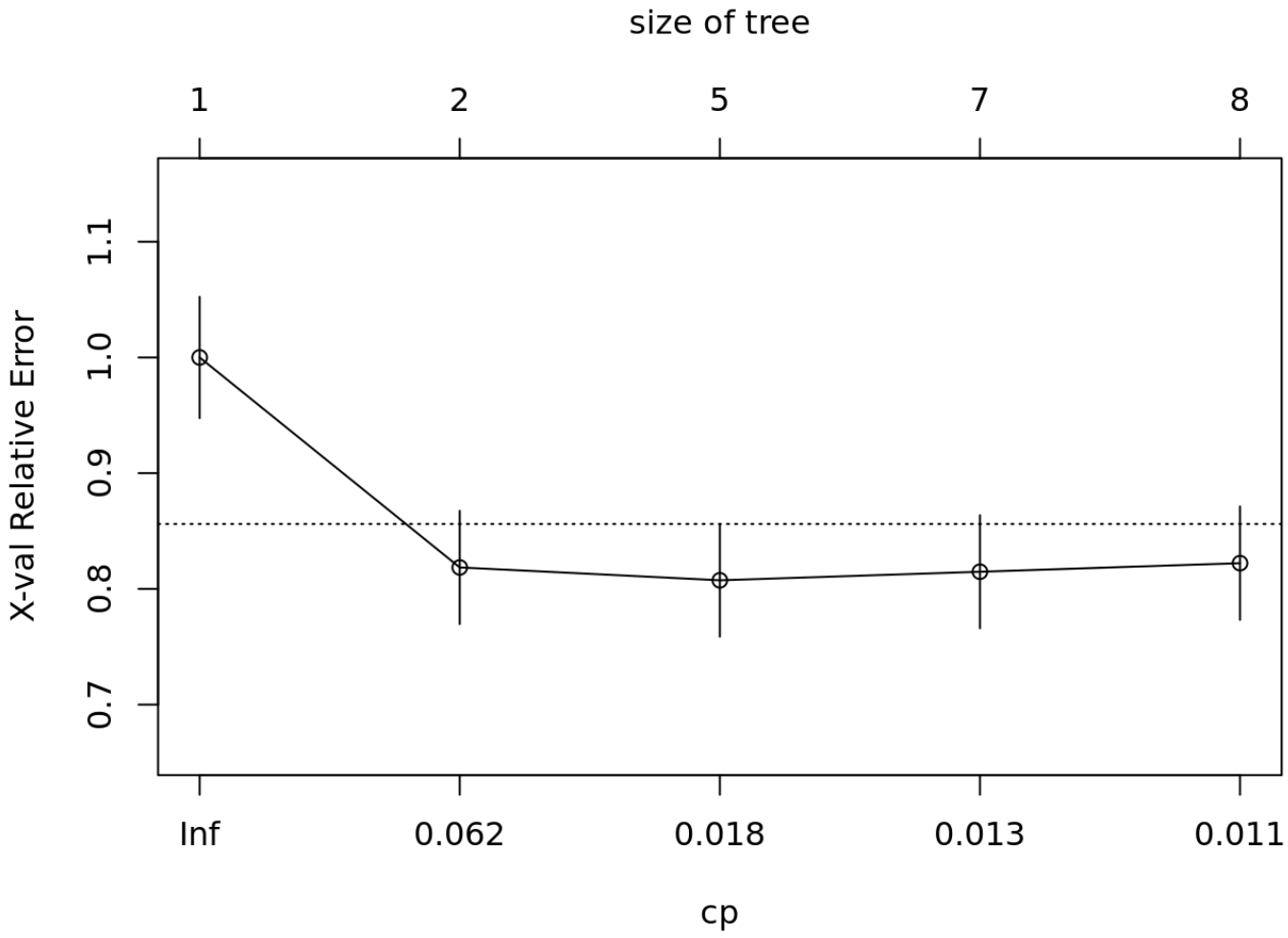
Для того, чтобы понять, насколько наши предсказания генерализируемы и распространимы на популяцию, разделим наблюдения на тренировочную и тестовую выборки. Построим разбиение по выявленным переменным на тренировочной выборке и проверим качество предсказания на тестовой

Code



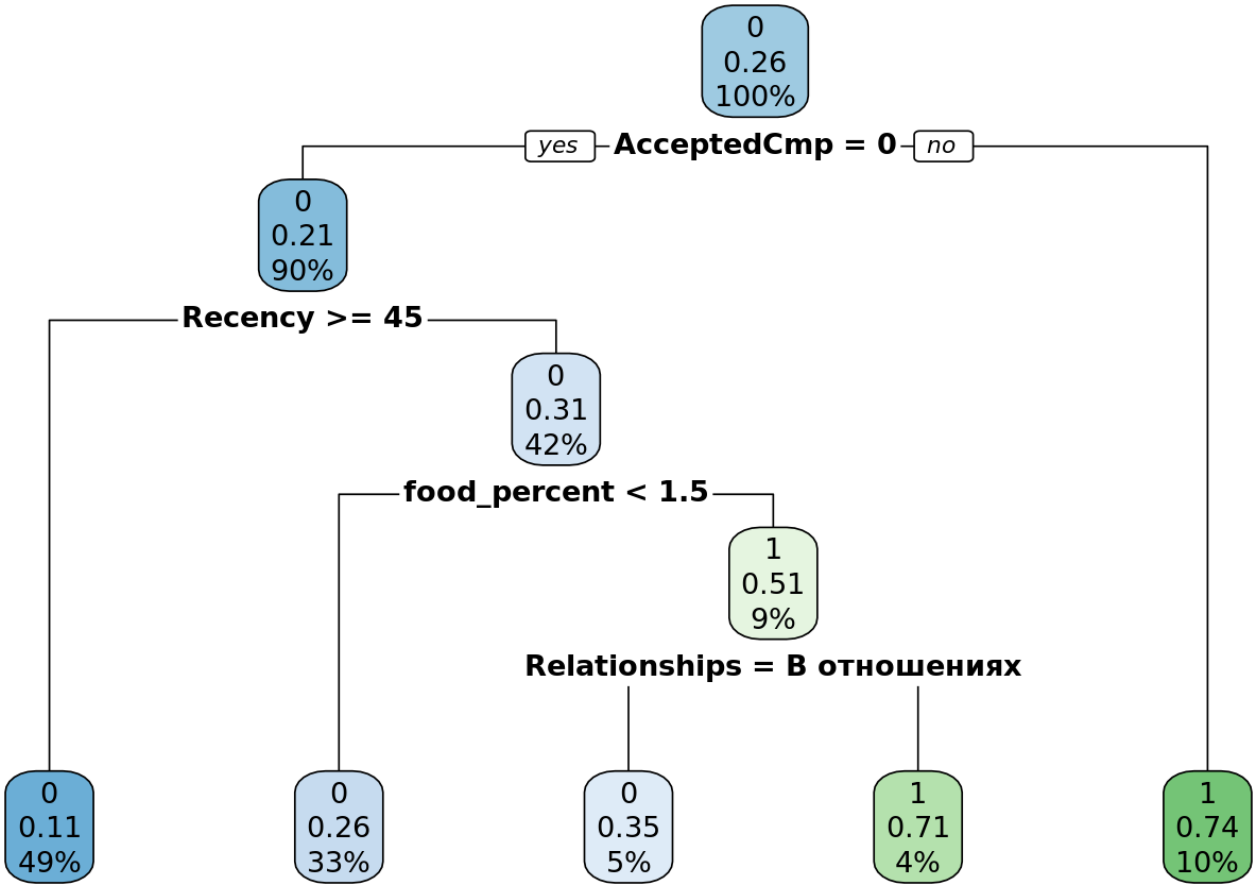
Обрежем дерево по ср

Code


size of tree
1
2
5
7
8
X-val Relative Error
0.7
0.8
0.9
1.0
1.1
Inf
0.062
0.018
0.013
0.011
cp

Code

Code



Проверим качество нашей модели на тренировочной выборке

Code

Точность на тренировочной модели составила: 0.81 *Gini impurity:* = 0.31

Посчитаем качество насколько применима построенная модель для тестовой выборки

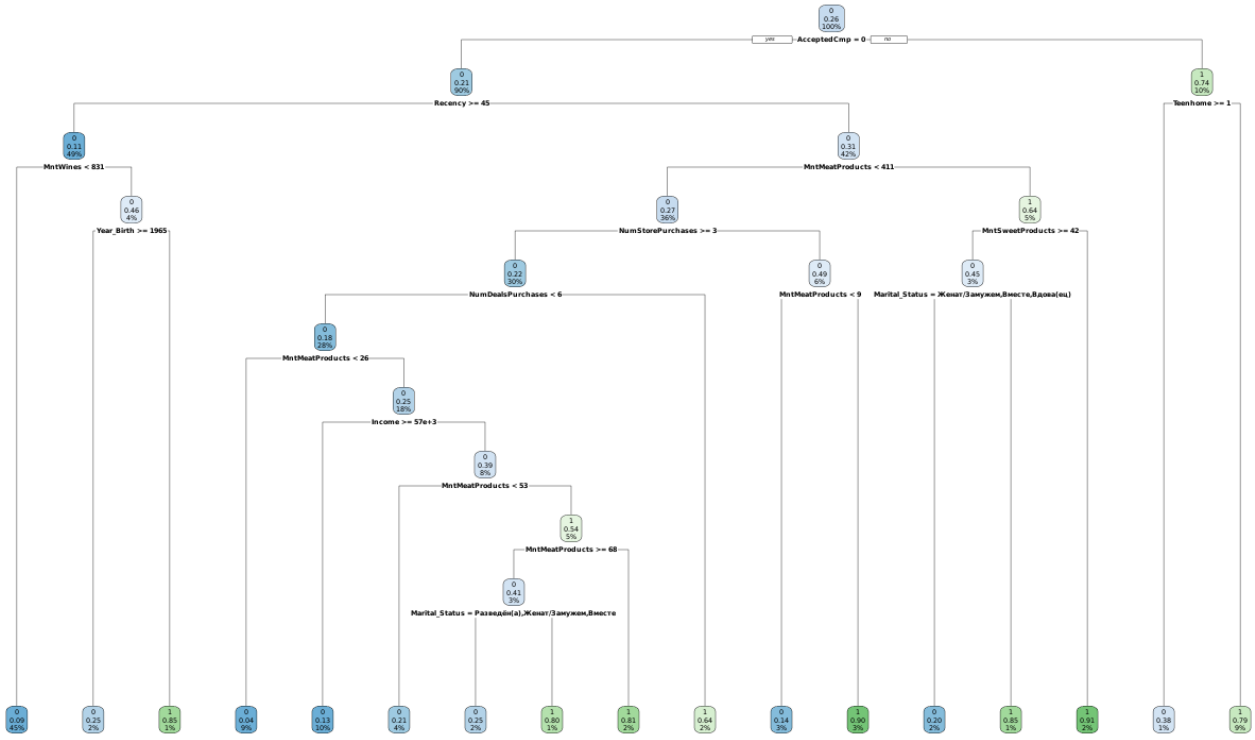
Code

Точность на тестовой модели составила: 0.79 *Gini impurity:* = 0.32

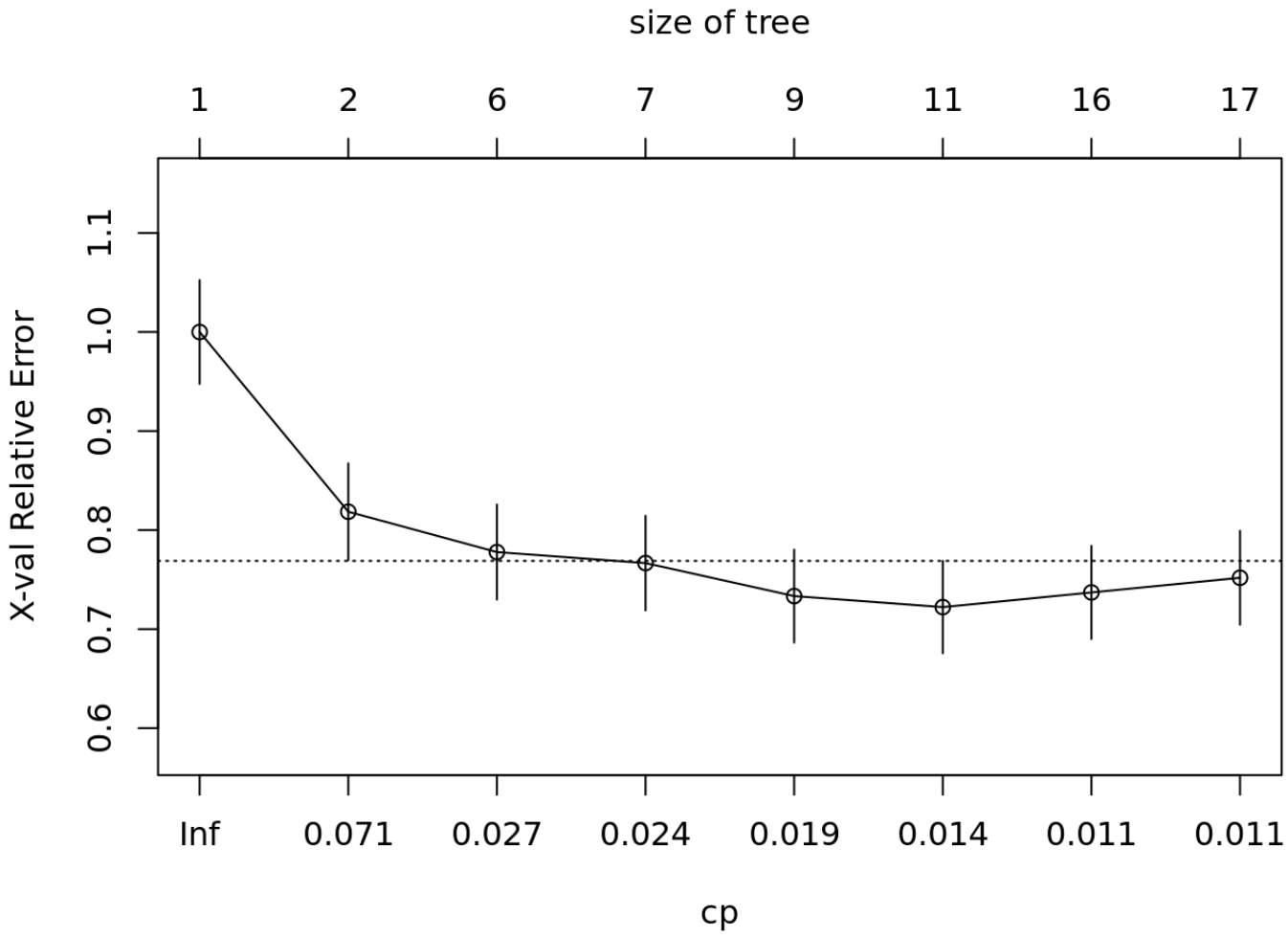
Вывод: Дерево решений произвело разбиение по заданным переменным, несколько из которых были созданы вручную. Таким образом модель предсказала положительный отклик на кампанию двум группам: тем, кто откликнулся на предыдущую кампанию и тем, кто: не откликнулся на предыдущую кампанию, но совершил последнюю покупку менее 45 дней назад, тратит на продукты питания более 1.5 % от дохода и при этом не состоит в отношениях. Тем, кто не удовлетворил данным критериям модель предсказала отрицательный отклик на текущую кампанию. Итоговая модель демонстрирует достаточно высокую точность, как на тренировочной, так и на тестовой выборке, так же снижая исходный Gini Impurity.

Можно ли сделать нашу модель более точной и ещё больше снизить Gini Impurity? Определим наилучшее автоматическое разбиение по всем переменным для тенировочной выборки

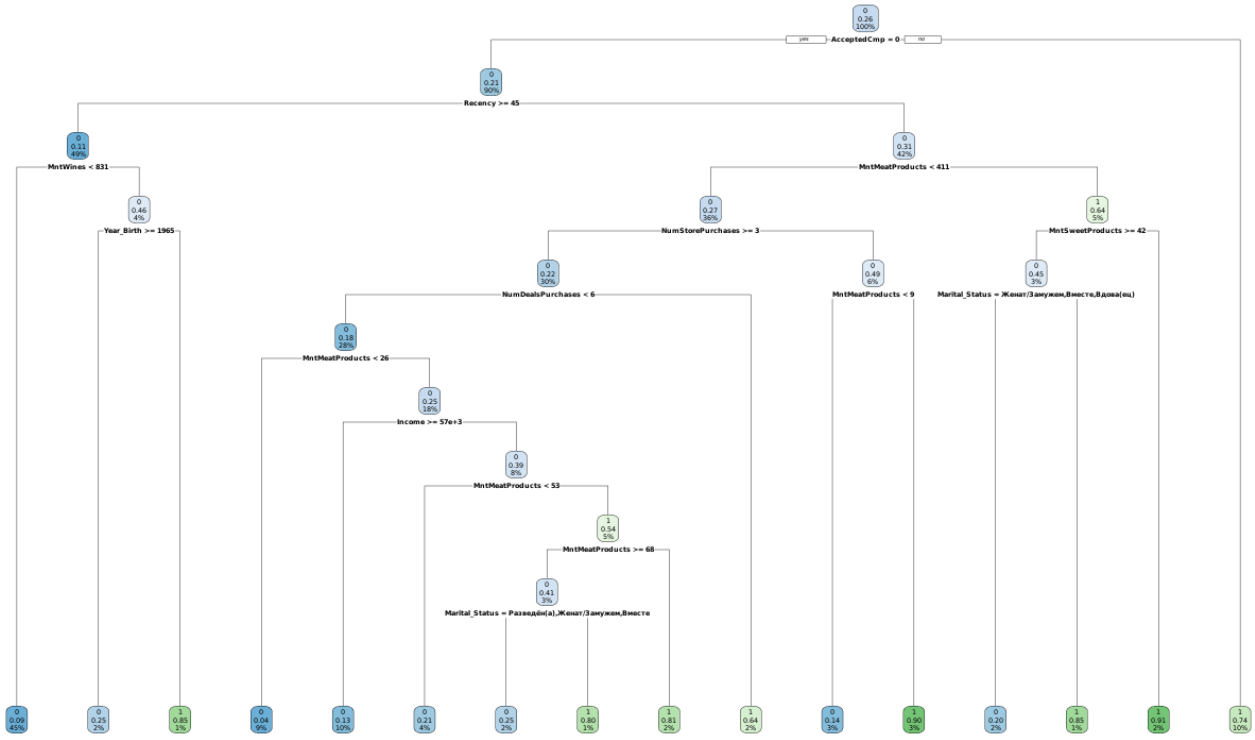
Code


0
0.26
100%
yes
AcceptedCmp = 0
no
0
0.21
90%
Recency >= 45
1
0.74
10%
Teenhome >= 1
0
0.11
49%
MntWines < 831
0
0.31
42%
MntMeatProducts < 411
0
0.46
4%
Year_Birth >= 1965
0
0.27
36%
NumStorePurchases >= 3
1
0.64
5%
MntSweetProducts >= 42
0
0.22
30%
NumDealsPurchases < 6
0
0.49
6%
MntMeatProducts < 9
0
0.45
3%
Marital_Status = Женат/Замужем,Вместе,Вдова(ец)
0
0.18
28%
MntMeatProducts < 26
0
0.25
18%
Income >= 57e+3
0
0.39
8%
MntMeatProducts < 53
1
0.54
5%
MntMeatProducts >= 68
0
0.41
3%
Marital_Status = Разведён(а),Женат/Замужем,Вместе
0
0.09
45%
0
0.25
2%
1
0.85
1%
0
0.04
9%
0
0.13
10%
0
0.21
4%
0
0.25
2%
1
0.80
1%
1
0.81
2%
1
0.64
2%
0
0.14
3%
1
0.90
3%
0
0.20
2%
1
0.85
1%
1
0.91
2%
0
0.38
1%
1
0.79
9%

Code


size of tree
1
2
6
7
9
11
16
17
X-val Relative Error
0.6
0.7
0.8
0.9
1.0
1.1
Inf
0.071
0.027
0.024
0.019
0.014
0.011
0.011
cp

Code



Code

Точность на тренировочной модели составила: 0.87 *Gini impurity:* = 0.22

Проверим её качество на тестовой:

Code

Точность на тестовой модели составила: 0.79 *Gini impurity:* = 0.3

Вывод: Дерево построенное по всем переменным вышло достаточно глубоким, использовало многие из выявленных переменных, а так же несколько других. Точность на тренировочной выборке выше, а gini impurity ниже, чем у предыдущей модели. Однако, на тестовой выборке разница между двумя моделями в точности и чистоте практически незаметна. В то время как глубина дерева, построенного по заданным переменным заметно меньше, что говорит в пользу первой модели. Её мы и мудем использовать для итогового предсказания.

Общие выводы

В результате проведённого анализа были выявлены группы респондентов, которые с наибольшей вероятностью откликнуться на новую маркетинговую компанию. Моя рекомендация при дальнейшем проведении кампании в других магазинах сети: уделить особое внимание целевой аудитории, состоящей в первую очередь из респондентов, откликнувшихся на предыдущую

маркетинговую кампанию, а так же из одиноких клиентов, выделяющих больше 1.5 % своего дохода на продукты питания и совершивших в нашем магазине покупки не позже, чем за 45 дней до проведения кампании.