Федеральное государственное автономное образовательное учреждениевысшего образования

«Национальный исследовательский ядерный университет «МИФИ»

ОТЧЕТ

по лабораторной работе №2 (вариант 54) по дисциплине:
«Статистика в экономике»

Выполнил: *студент группы Б20-902* *Цуканов Л. О.*

(подпись) (Фамилия И.О.)

Проверил: *Смирнов Д. С.*

(оценка) (подпись) (Фамилия И.О.)

Москва 2022 г

Условие

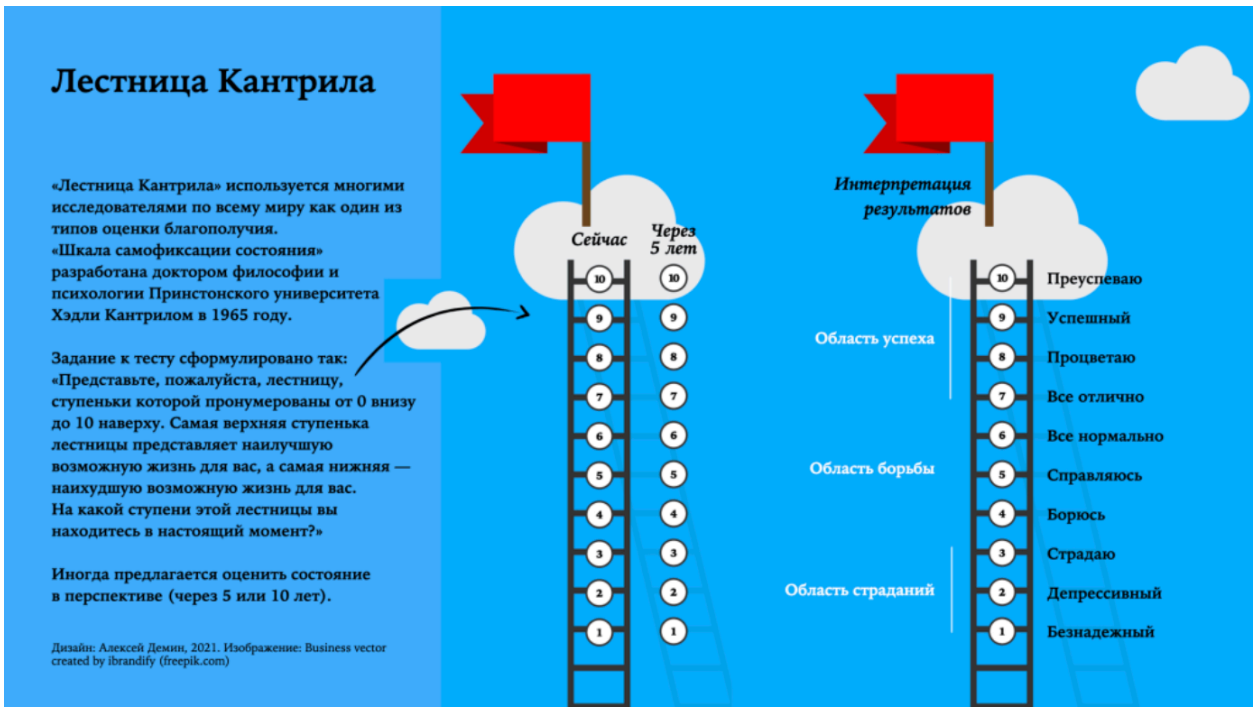
В Файле содержатся результат опроса населения о его условиях существования. Переменные разбиты на 2 класса - "Признаки состояния" - это субъективная оценка населения своего бытия и "Признаки причины" - объектные количественные признаки оценивающие жизнедеятельность индивида и социума в котором он проживает.

К признакам состояния относятся:
- Оценка благополучия
- Оценка социальной поддержки
- Ожидаемая продолжительность здоровой жизни
- Свобода граждан самостоятельно принимать жизненно важные решения
- Индекс Щедрости
- Индекс отношения к коррупции
- Оценка риска безработицы
- Индекс кредитного оптимизма
- Индекс страха социальных конфликтов
- Индекс семьи
- Индекс продовольственной безопасности
- Чувство технологического прогресса
- Чувство неравенства доходов в обществе

К индивидуальным признакам причины относятся:
- Среднегодовой доход, тыс. $
- Объем потребленного алкоголя в год, л.
- Количество членов семьи
- Количество лет образования
- Доля от дохода семьи, которая тратится на продовольствие, %

К общественным признакам причины относятся:
- Коэффициент Джини сообщества
- Издержки сообщества на окружающую среду, млн. $
- Охват беспроводной связи в сообществе, %
- Количество смертей от вирусных и респираторных заболеваний в сообществе, тыс. человек
- Волатильность потребительских цен в сообществе

Индивидуальные показатели характеризуют непосредственно индивида, общественные - сообщество, в котором он проживает.
В выборке могут присутствовать по несколько человек из одного сообщества. Все их общественные характеристики таким образом будут совпадать.
Также в данных присутствует интегральная характеристика удовлетворенности человека жизнью - для ее описания используется шкала Кантрила («Рис. 1»)



«Рис. 1»

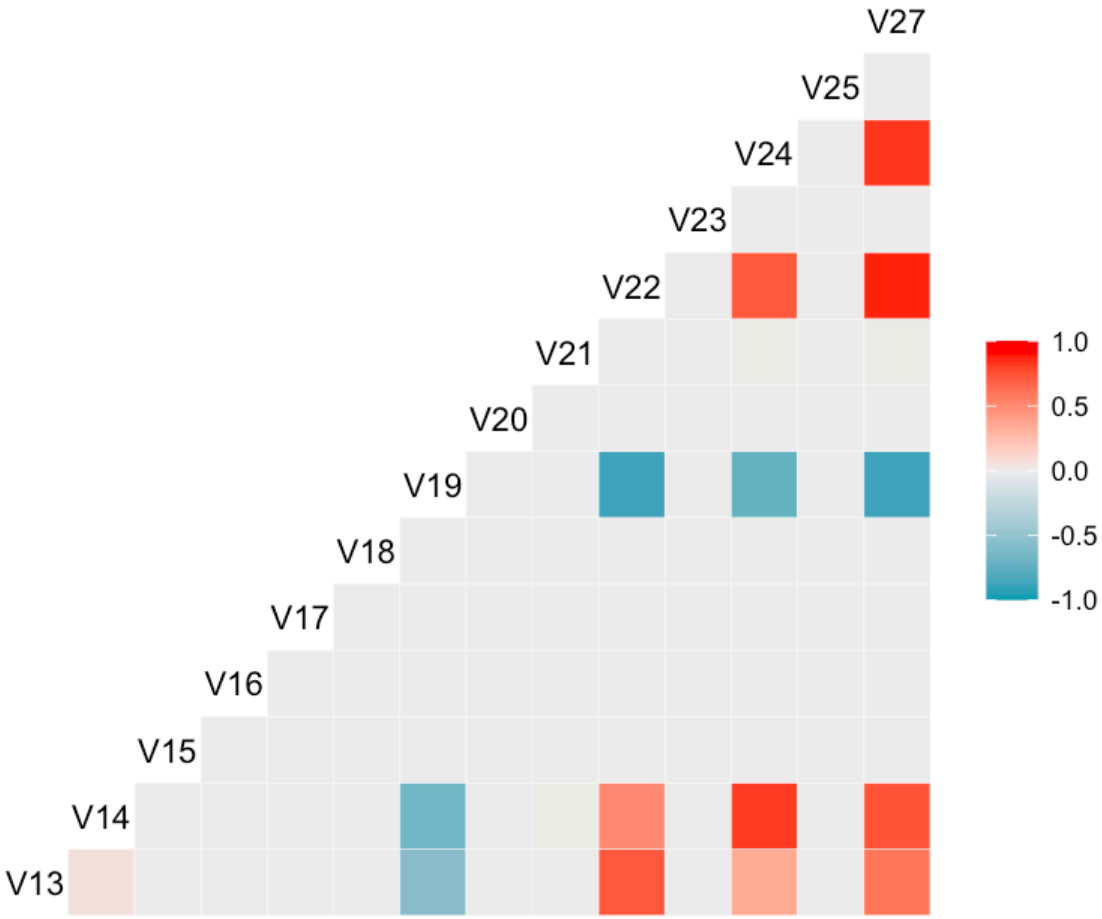
Задание

- Определить какие из признаков состояния наиболее сильно связаны с интегральной оценкой счастья (благополучия) респондента
- Определить, как влияют признаки причины на наиболее важные признаки состояния
- Пользуясь найденными закономерностями спрогнозировать попадание респондентов, у которых интегральная характеристика отмечена как "Неизвестно" в укрупненные группы шкалы Кантрила

* Модель должна иметь точность больше 60%

Выявление зависимостей

Зависимости между признаками состояния и интегральной оценкой счастья

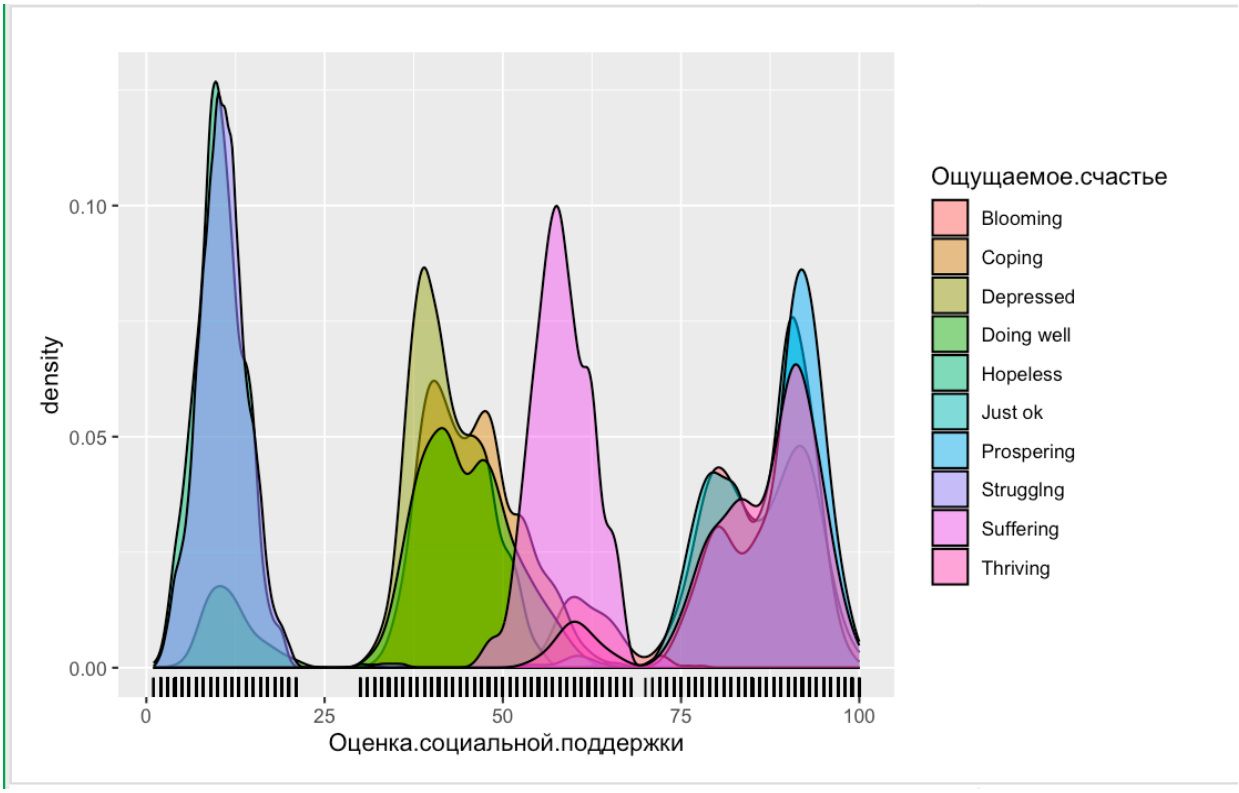


«Рис. 2» Корреляционная матрица

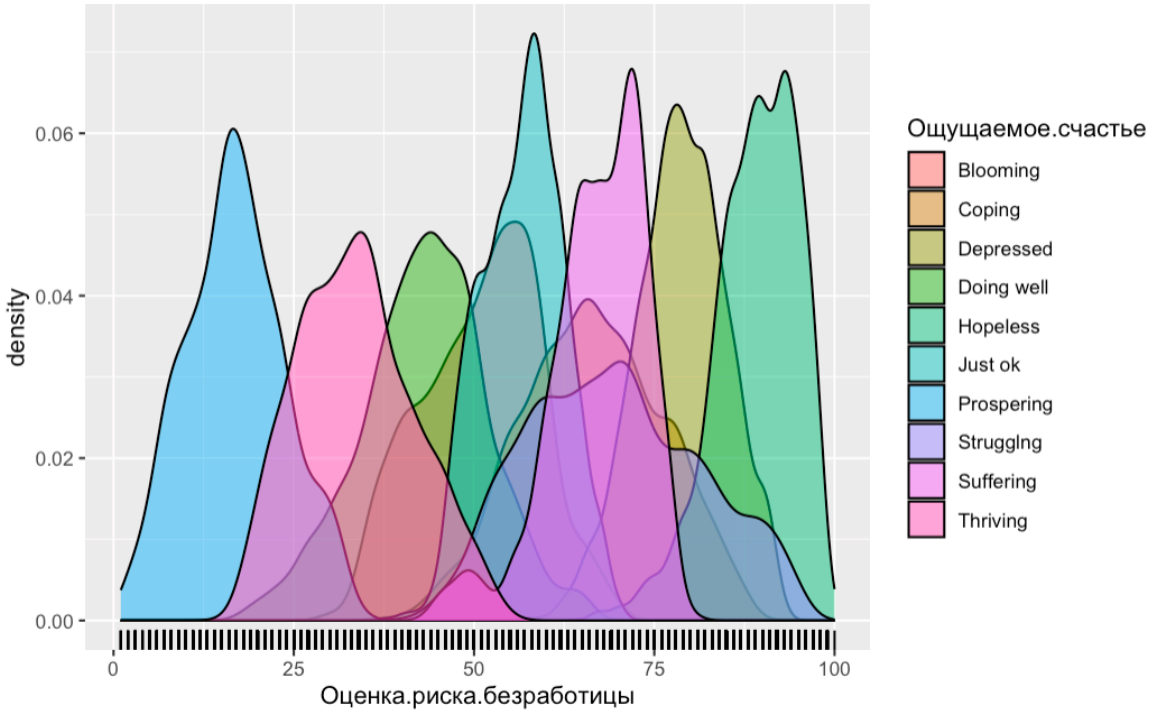
Наиболее важные признаки состояния: Оценка благополучия, Оценка социальной поддержки, Оценка риска безработицы, Индекс семьи, Чувство технологического прогресса

[V13] "Оценка благополучия"
[V14] "Оценка социальной поддержки"
[V15] "Ожидаемая продолжительность здоровой жизни"
[V16] "Свобода граждан самостоятельно принимать жизненно важные решения"
[V17] "Индекс Щедрости"
[V18] "Индекс отношения к коррупции"
[V19] "Оценка риска безработицы"
[V20] "Индекс кредитного оптимизма"
[V21] "Индекс страха социальных конфликтов"
[V22] "Индекс семьи"
[V23] "Индекс продовольственной безопасности"
[V24] "Чувство технологического прогресса"
[V25] "Чувство неравенства доходов в обществе"
[V27] "Ощущаемое счастье"

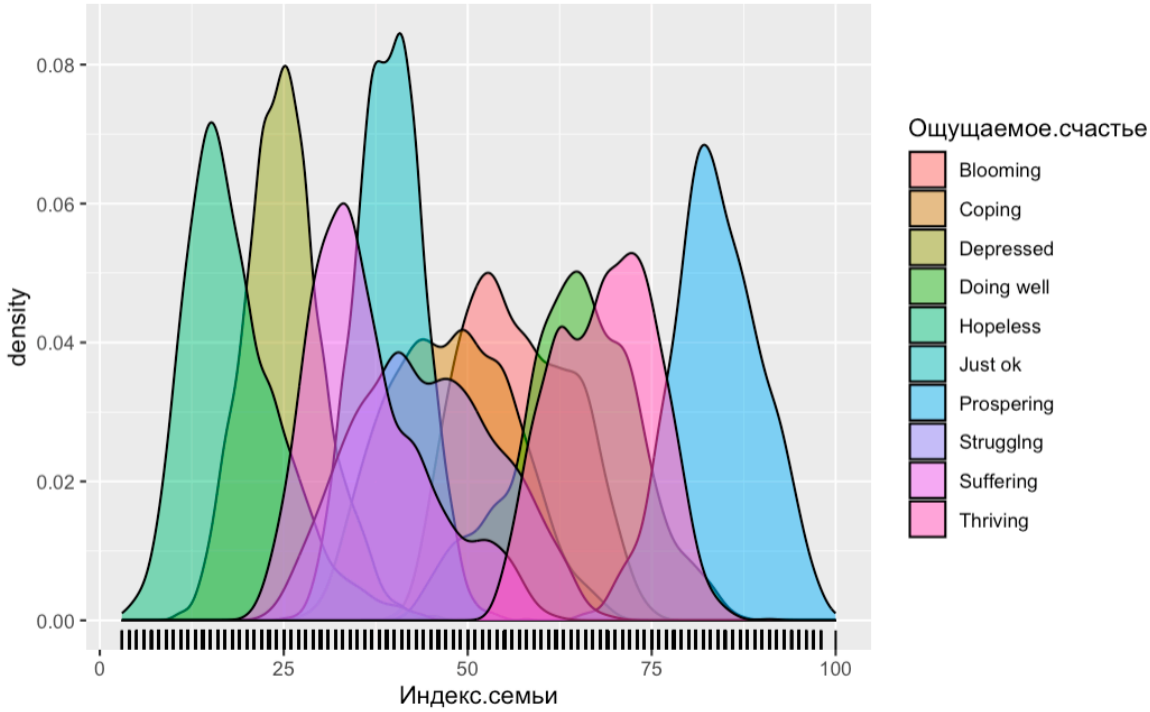
Визуализация зависимостей между признаками состояния и интегральной оценкой счастья



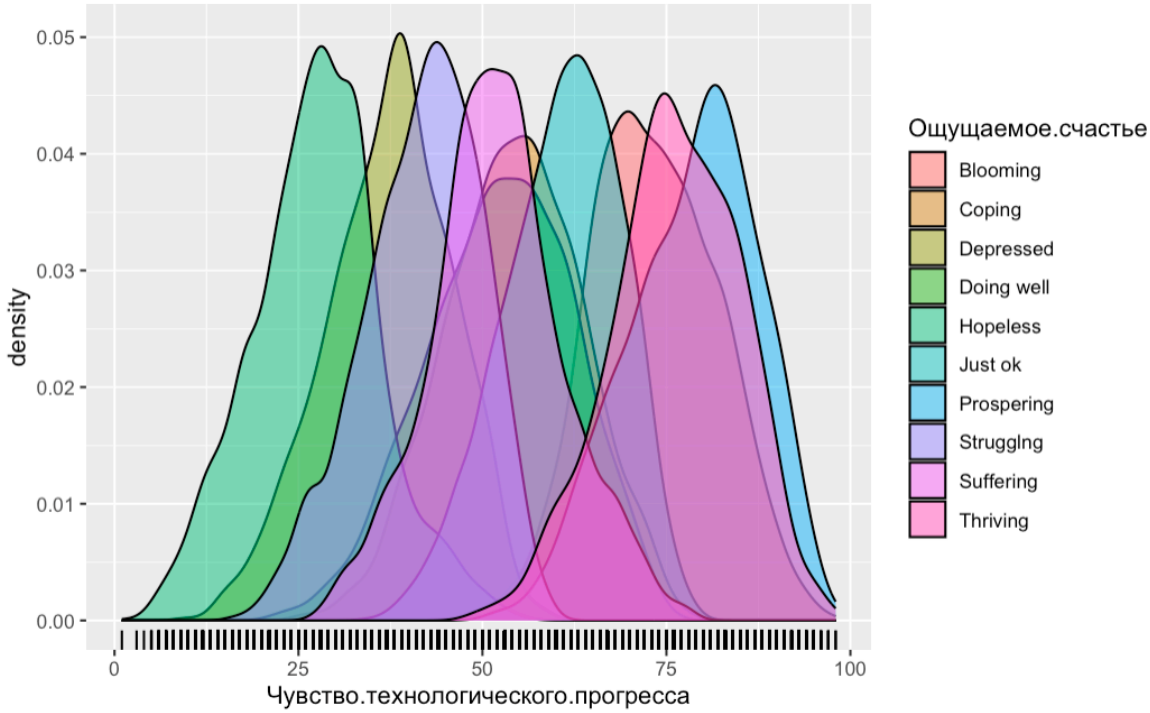
«Рис. 3»



«Рис. 4»

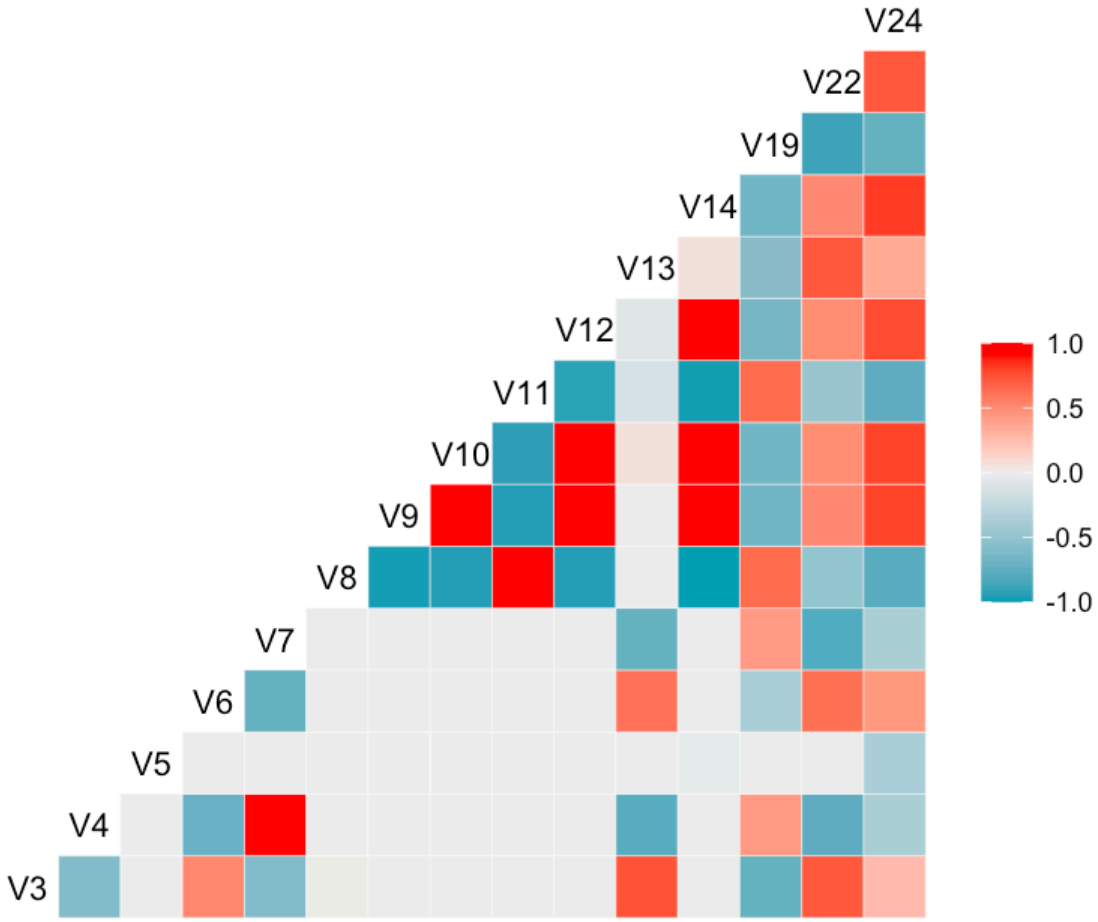

density
0.08
0.06
0.04
0.02
0.00
0
25
50
75
100
Индекс.семьи
Ощущаемое.счастье
Blooming
Coping
Depressed
Doing well
Hopeless
Just ok
Prospering
Strugglng
Suffering
Thriving

«Рис. 5»


density
0.05
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
Чувство.технологического.прогресса
Ощущаемое.счастье
Blooming
Coping
Depressed
Doing well
Hopeless
Just ok
Prospering
Strugglng
Suffering
Thriving

«Рис. 6»

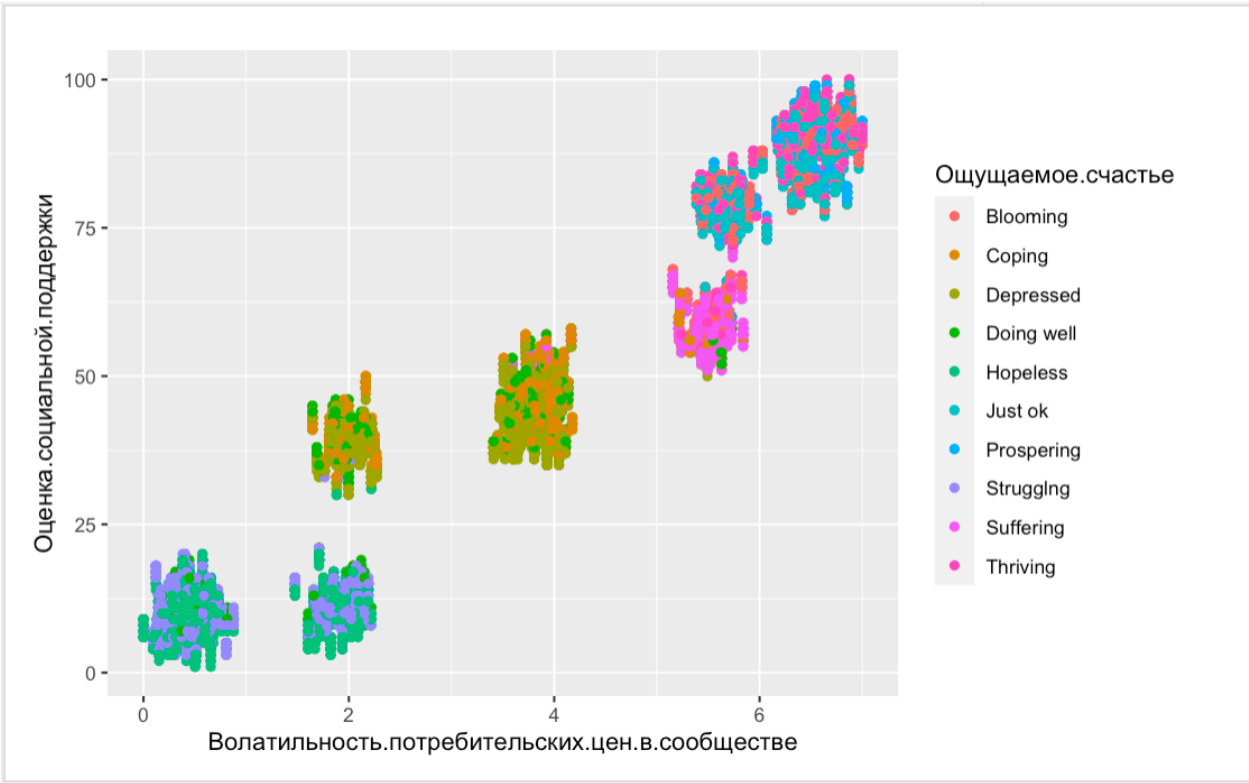
Зависимости между основными признаками состояния и признаками причины



«Рис. 7» Корреляционная матрица

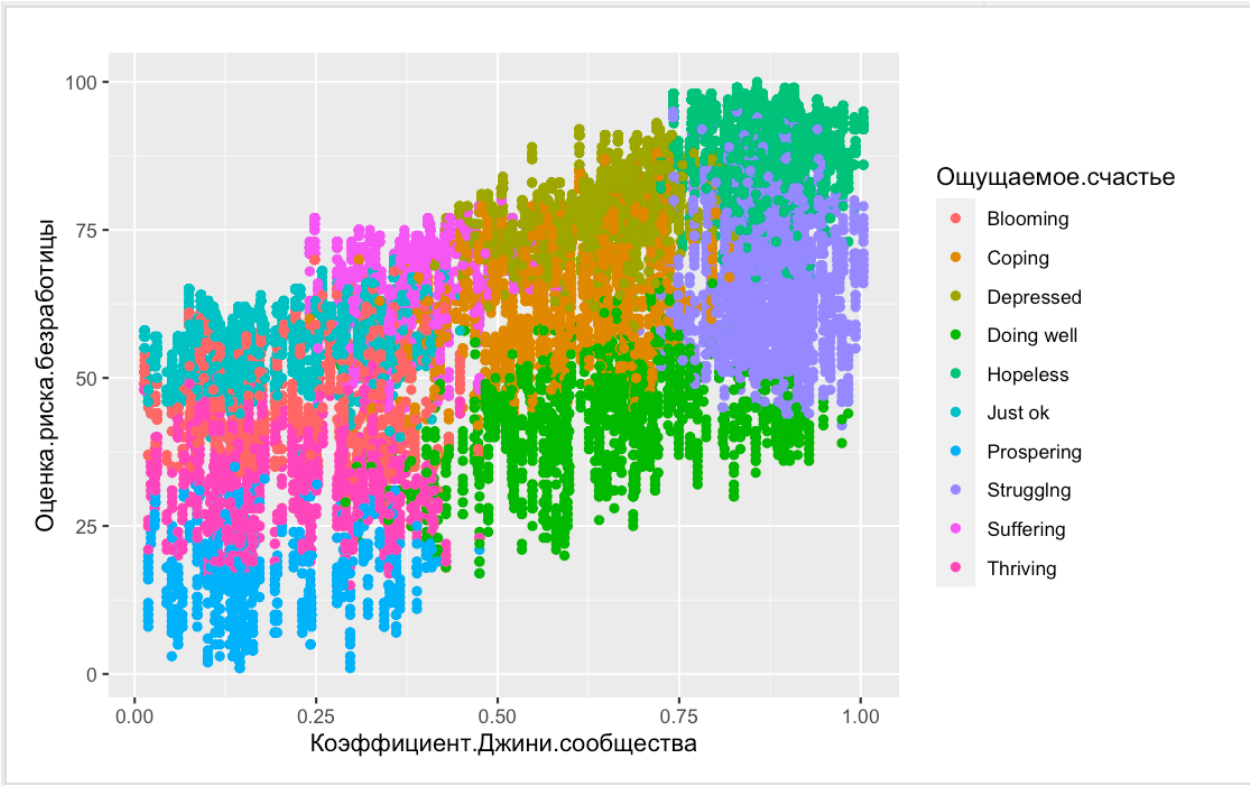
[V3] "Среднегодовой доход тыс. $"
[V4] "Объем потребленного алкоголя в год л."
[V5] "Количество членов семьи"
[V6] "Количество лет образования"
[V7] "Доля от дохода семьи, которая тратится на продовольствие"
[V8] "Коэффициент Джини сообщества"
[V9] "Издержки сообщества на окружающую среду млн. $"
[V10] "Охват беспроводной связи в сообществе"
[V11] "Количество смертей от вирусных и респираторных заболеваний в сообществе тыс. человек"
[V12] "Волатильность потребительских цен в сообществе"
[V13] "Оценка благополучия"
[V14] "Оценка социальной поддержки"
[V19] "Оценка риска безработицы"
[V22] "Индекс семьи"
[V24] "Чувство технологического прогресса"

Визуализация зависимостей между основными признаками состояния и признаками причины

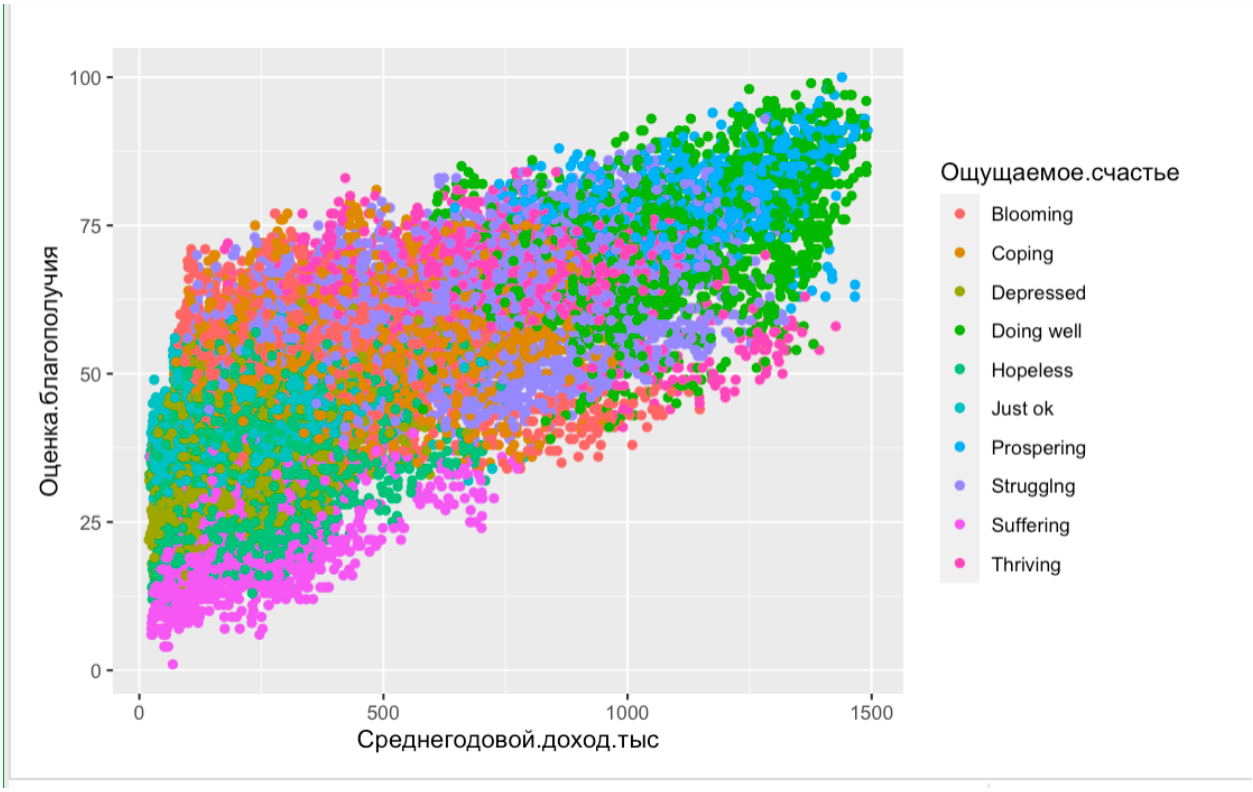


«Рис. 8»

«Рис. 9»


Оценка.риска.безработицы
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
Коэффициент.Джини.сообщества
Ощущаемое.счастье
Blooming
Coping
Depressed
Doing well
Hopeless
Just ok
Prospering
Strugglng
Suffering
Thriving

«Рис. 10»


Оценка.благополучия
100
75
50
25
0
0
500
1000
1500
Среднегодовой.доход.тыс
Ощущаемое.счастье
Blooming
Coping
Depressed
Doing well
Hopeless
Just ok
Prospering
Strugglng
Suffering
Thriving

«Рис. 11»

Модели предсказания оценок состояния

Строим модели предсказания оценок состояния, основываясь на признаках причины. Для создания моделей будем использовать линейную регрессию.

1) Модель для предсказания ***оценки благополучия***

```
Call:
lm(formula = formula2_optimal, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50377 -0.24774 -0.00277  0.24670  0.51175

Coefficients:
                                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                                     1.084e+02  2.093e-02    5179   <2e-16 ***
Среднегодовой.доход.тыс                                                         2.007e-02  6.678e-06    3005   <2e-16 ***
Объем.потребленного.алкоголя.в.год.л.                                          -5.608e-01  1.466e-04   -3826   <2e-16 ***
Охват.беспроводной.связи.в.сообществе                                           4.836e+01  2.754e-02    1756   <2e-16 ***
Количество.смертей.от.вирусных.и.респираторных.заболеваний.в.сообществе.тыс..человек -1.221e+00  4.744e-04   -2574   <2e-16 ***
Волатильность.потребительских.цен.в.сообществе                                  -1.143e+01  2.767e-03   -4131   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2881 on 22494 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9997
F-statistic: 1.535e+07 on 5 and 22494 DF,  p-value: < 2.2e-16
```

2) Модель для предсказания ***оценки социальной поддержки***

```
Call:
lm(formula = formula3_optimal, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51489 -0.25257  0.00329  0.25014  0.51062

Coefficients:
                                                                                 Estimate Std. Error   t value Pr(>|t|)
(Intercept)                                                                    1.041e+02  9.500e-03 10958.701   <2e-16 ***
Объем.потребленного.алкоголя.в.год.л.                                         -2.299e-02  3.149e-04   -73.018   <2e-16 ***
Количество.членов.семьи                                                       -2.670e-01  6.580e-04  -405.812   <2e-16 ***
Доля.от.дохода.семьи.которая.тратится.на.продовольствие                       -6.005e-04  3.665e-04    -1.639    0.101
Коэффициент.Джини.сообщества                                                  -5.786e+01  1.759e-02 -3289.596   <2e-16 ***
Количество.смертей.от.вирусных.и.респираторных.заболеваний.в.сообществе.тыс..человек -1.235e+00  4.117e-04 -2999.133   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2883 on 22494 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 5.188e+07 on 5 and 22494 DF,  p-value: < 2.2e-16
```

3) Модель для предсказания ***оценки риска безработицы***

```
Call:
lm(formula = formula4_optimal, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51216 -0.24941  0.00073  0.25029  0.50691

Coefficients:
                                          Estimate Std. Error   t value Pr(>|t|)
(Intercept)                             1.076e+02  2.973e-02  3619.442   <2e-16 ***
Среднегодовой.доход.тыс                -4.227e-02  5.411e-06 -7810.651   <2e-16 ***
Коэффициент.Джини.сообщества           -4.689e-02  2.847e-02    -1.647   0.0997 .
Издержки.сообщества.на.окружающую.среду.млн. -7.461e-02  7.699e-05  -969.140   <2e-16 ***
Охват.беспроводной.связи.в.сообществе  -1.989e+01  3.037e-02  -655.145   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2887 on 22495 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 2.793e+07 on 4 and 22495 DF,  p-value: < 2.2e-16
```

4) Модель для предсказания ***индекса семьи***

```
Call:
lm(formula = formula5_optimal, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50717 -0.25064  0.00503  0.25107  0.50134

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                            5.883e+01  1.175e-02    5006   <2e-16 ***
Среднегодовой.доход.тыс                                2.118e-02  6.617e-06    3200   <2e-16 ***
Доля.от.дохода.семьи.которая.тратится.на.продовольствие -7.896e-01  1.690e-04   -4671   <2e-16 ***
Издержки.сообщества.на.окружающую.среду.млн.           8.340e-02  1.622e-05    5141   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2891 on 22496 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 3.298e+07 on 3 and 22496 DF,  p-value: < 2.2e-16
```

5) Модель для предсказания ***чувства технологического прогресса***

```
Call:
lm(formula = formula6_optimal, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51121 -0.25116 -0.00149  0.25188  0.50592

Coefficients:
                                                                    Estimate Std. Error   t value Pr(>|t|)
(Intercept)                                                        5.393e+01  3.087e-02  1746.920   <2e-16 ***
Объем.потребленного.алкоголя.в.год.л.                             -1.709e-01  1.702e-04 -1003.728   <2e-16 ***
Количество.членов.семьи                                           -2.281e+00  6.616e-04 -3448.021   <2e-16 ***
Количество.лет.образования                                         9.649e-01  4.257e-04  2266.816   <2e-16 ***
Коэффициент.Джини.сообщества                                      -2.371e+01  2.420e-02  -979.805   <2e-16 ***
Охват.беспроводной.связи.в.сообществе                              3.151e+01  2.916e-02  1080.782   <2e-16 ***
Количество.смертей.от.вирусных.и.респираторных.заболеваний.в.сообществе.тыс..человек  7.603e-04  4.739e-04     1.605    0.109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2899 on 22493 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9997
F-statistic: 1.487e+07 on 6 and 22493 DF,  p-value: < 2.2e-16
```

- Коэффициент детерминации почти равен 1;
- t-value для предикатов в большинстве случаев значительно больше 2;
- p-value стремится к 0.

Данные факторы говорят нам, что модели, предсказывающие признаки состояния, достаточно точны и мы можем предсказать интегральную оценку счастья.

Прогноз ощущаемого счастья

Модель полиномиальный регрессии для оценки счастья

Для предсказания будем использовать полиномиальную регрессию, чтобы учесть не линейную связь между предикатами.

```
Residual standard error: 0.5847 on 22347 degrees of freedom
Multiple R-squared:  0.954,     Adjusted R-squared:  0.9537
F-statistic:  3052 on 152 and 22347 DF,  p-value: < 2.2e-16
```

Коэффициент детерминации -> 1

Точность модели составила **69%**

Модель случайного леса

Модель случайного леса основана на признаках причины

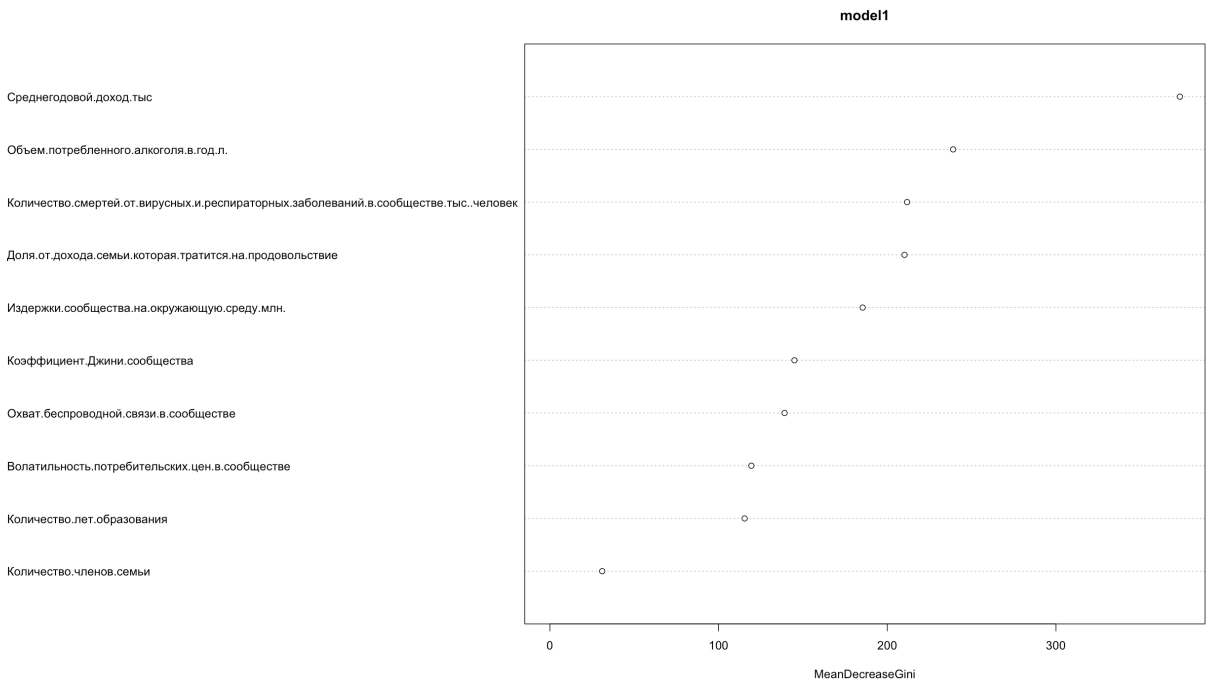


График отражает важность каждого фактора.
По оси X отображается среднее увеличение частоты узлов деревьев регрессии на основе разделения по различным предикторам, отображаемым по оси Y.

Точность модели составила **92%**