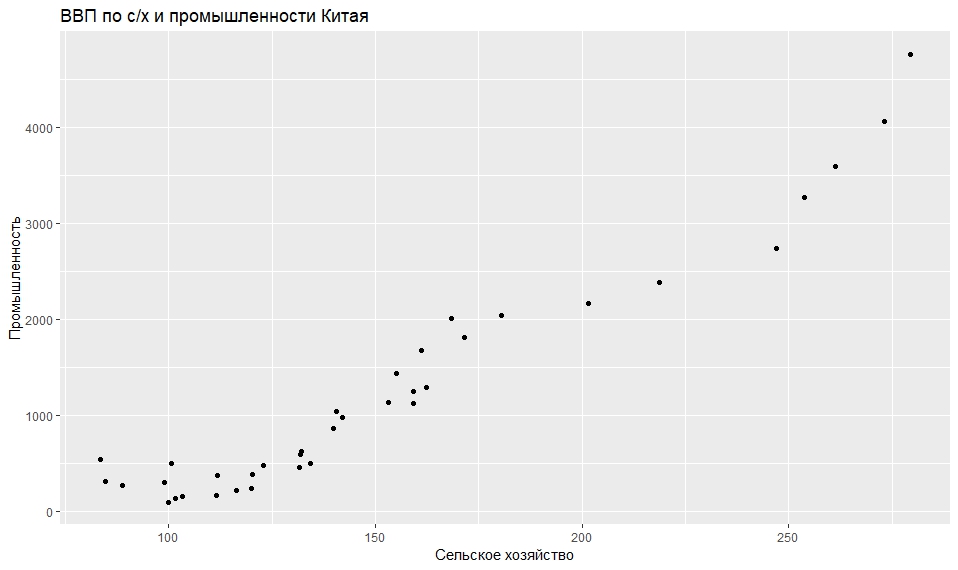
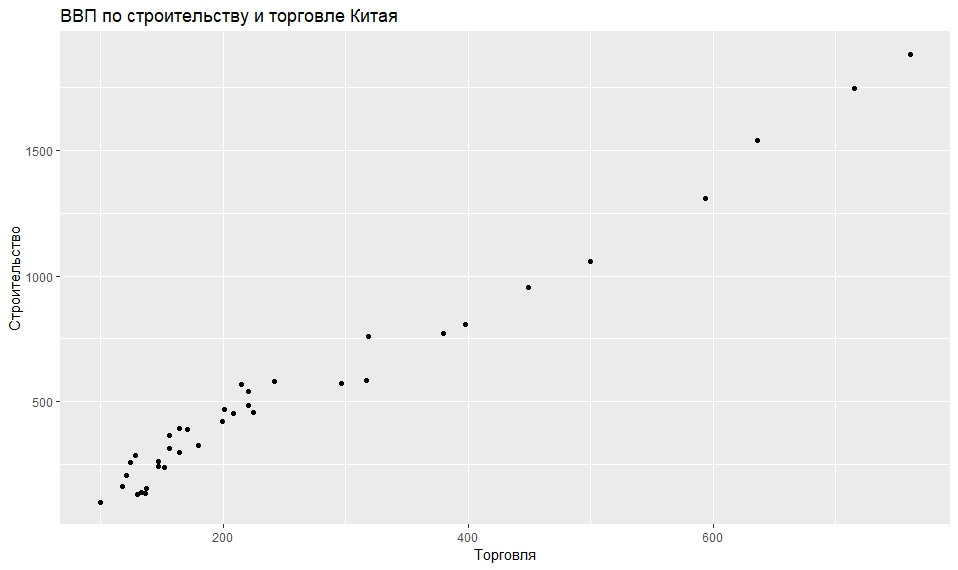
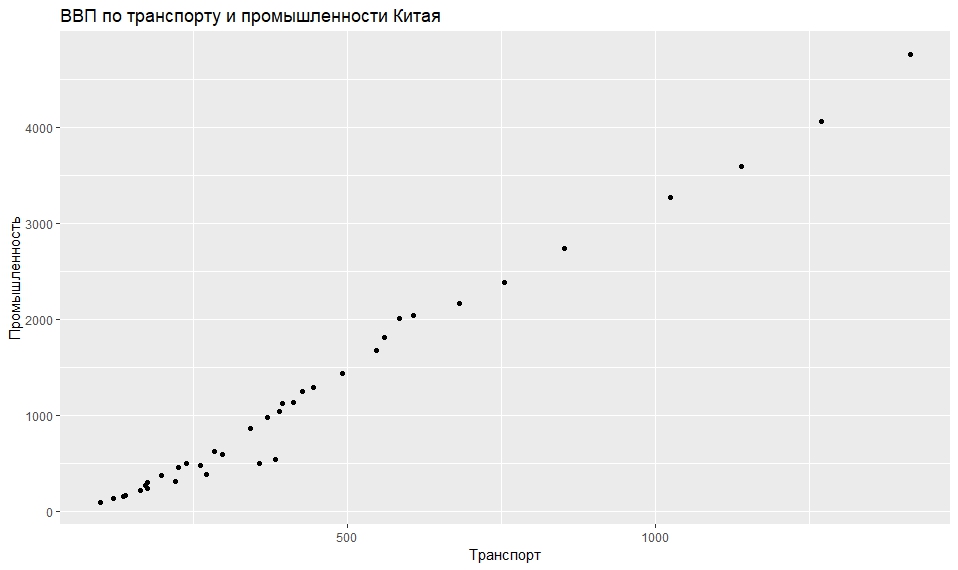
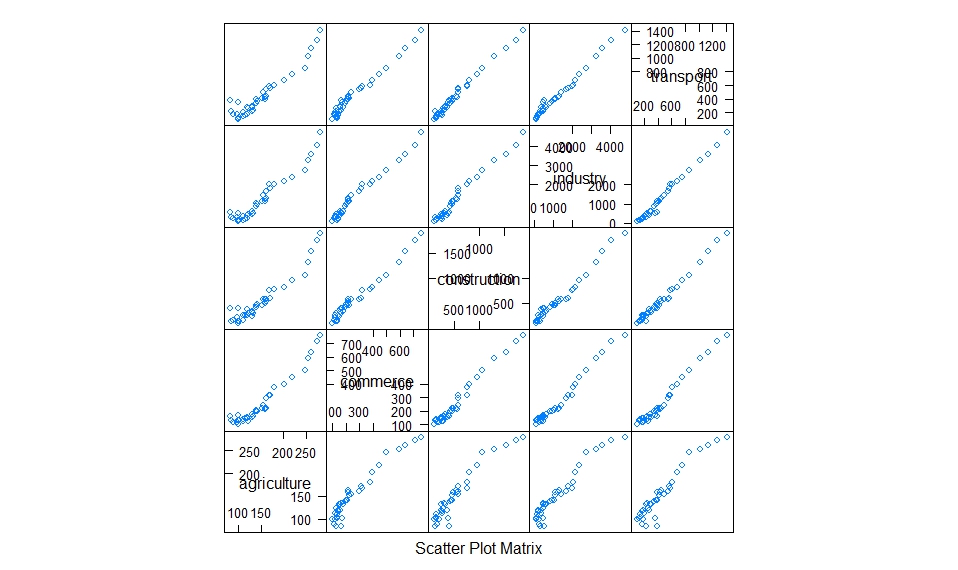
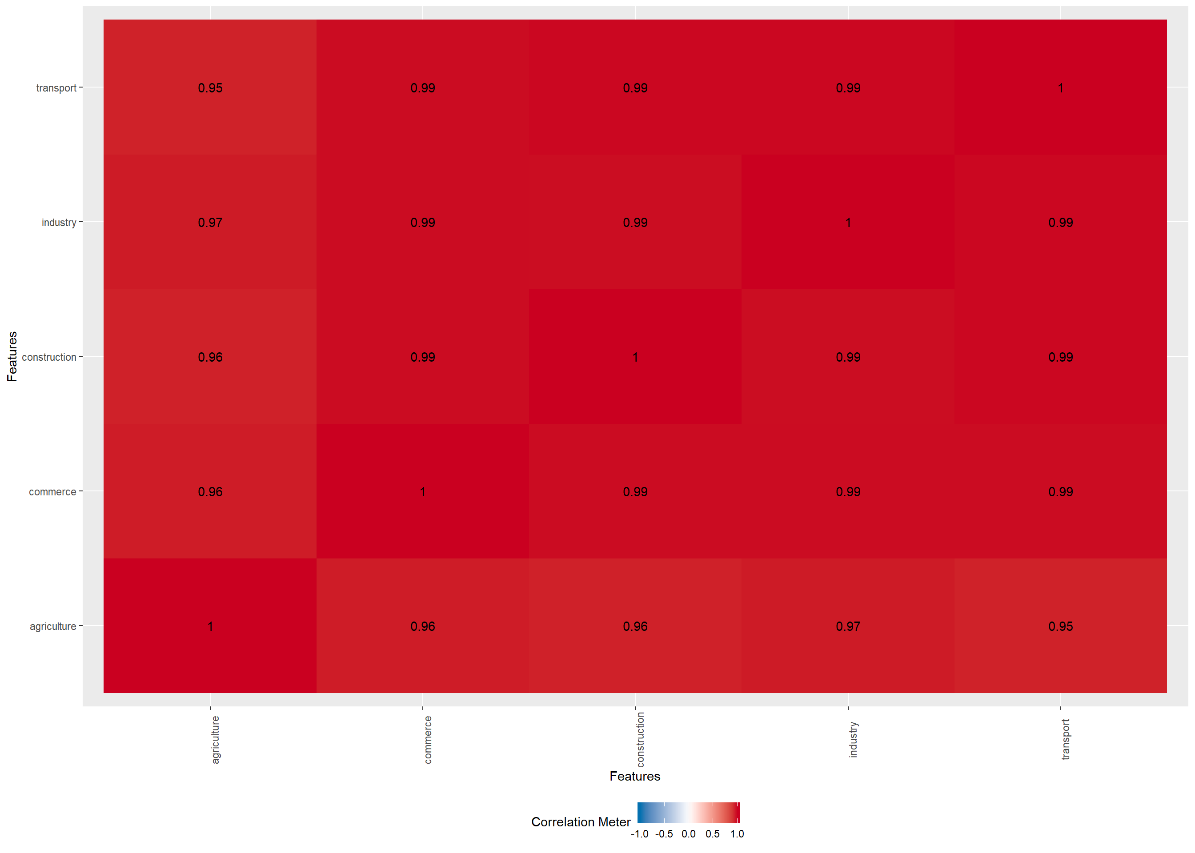
1. Набор данных ”China Income – Chinese Real National Data” (пакет- AER)
2. Описательный анализ данных
   1. Количество переменных – 5; наблюдений – 37.
   2. Все переменные – количественные (dbl).
   3. –











agriculture commerce construction industry transport

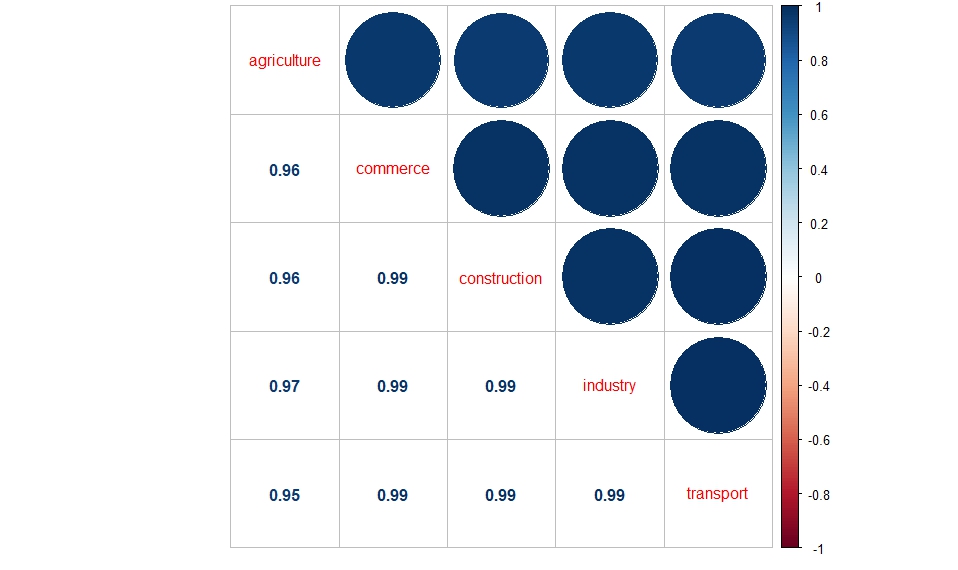
agriculture 1.0000000 0.9641547 0.9550280 0.9670784 0.9521588

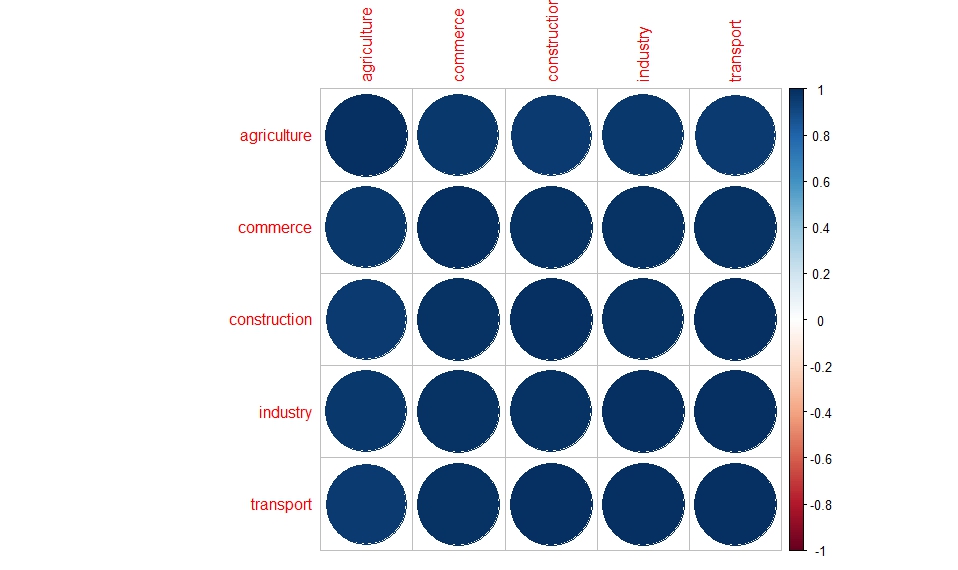
commerce 0.9641547 1.0000000 0.9880994 0.9883226 0.9881633

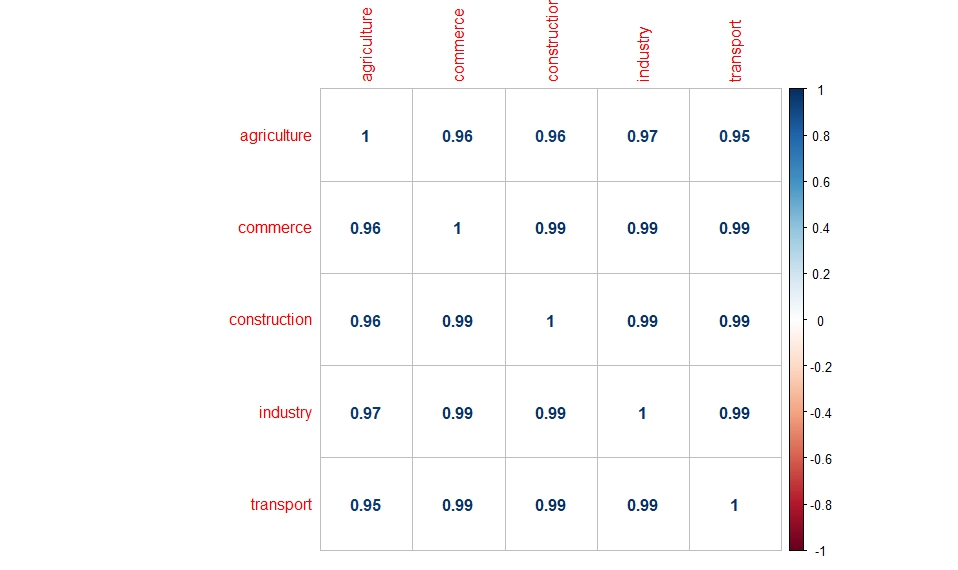
construction 0.9550280 0.9880994 1.0000000 0.9873144 0.9936743

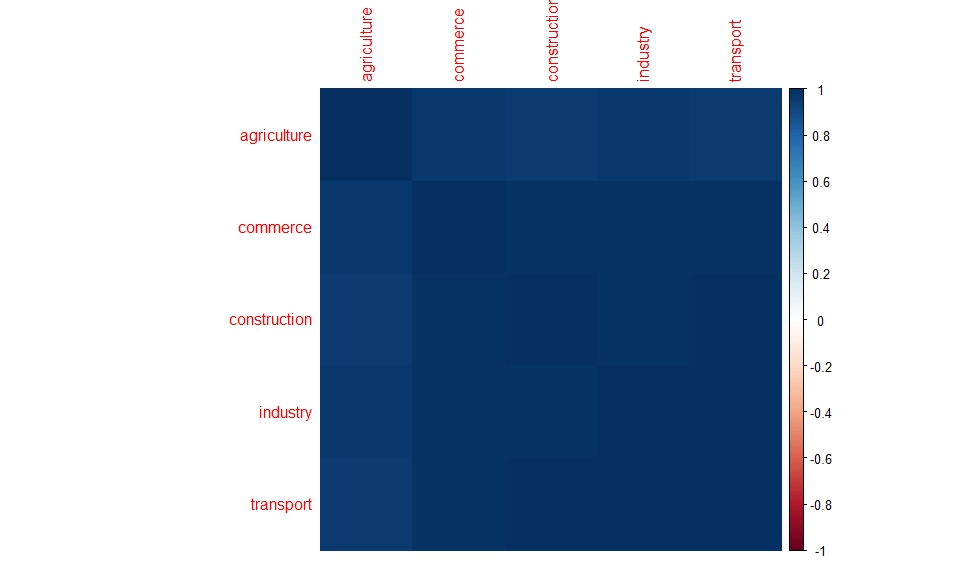
industry 0.9670784 0.9883226 0.9873144 1.0000000 0.9927794

transport 0.9521588 0.9881633 0.9936743 0.9927794 1.0000000



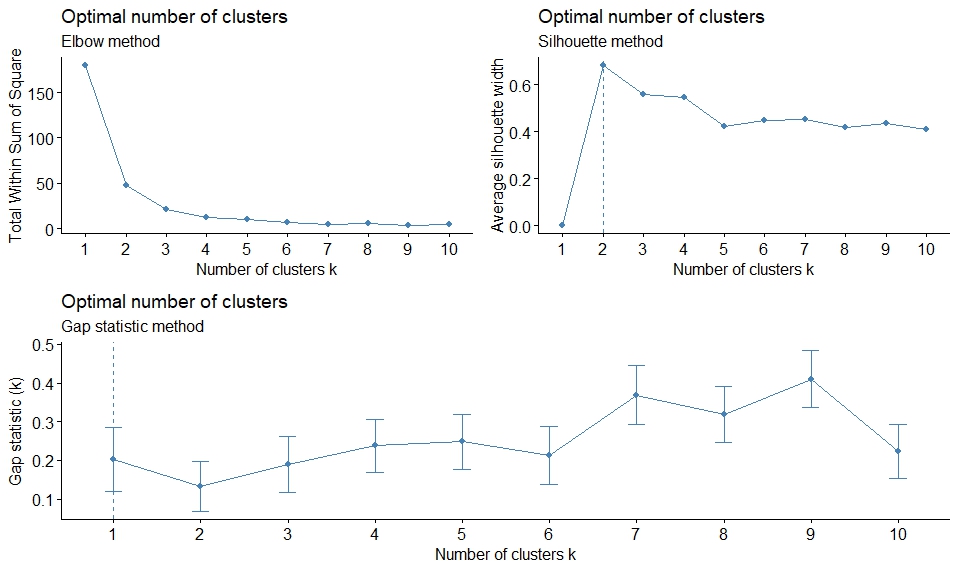




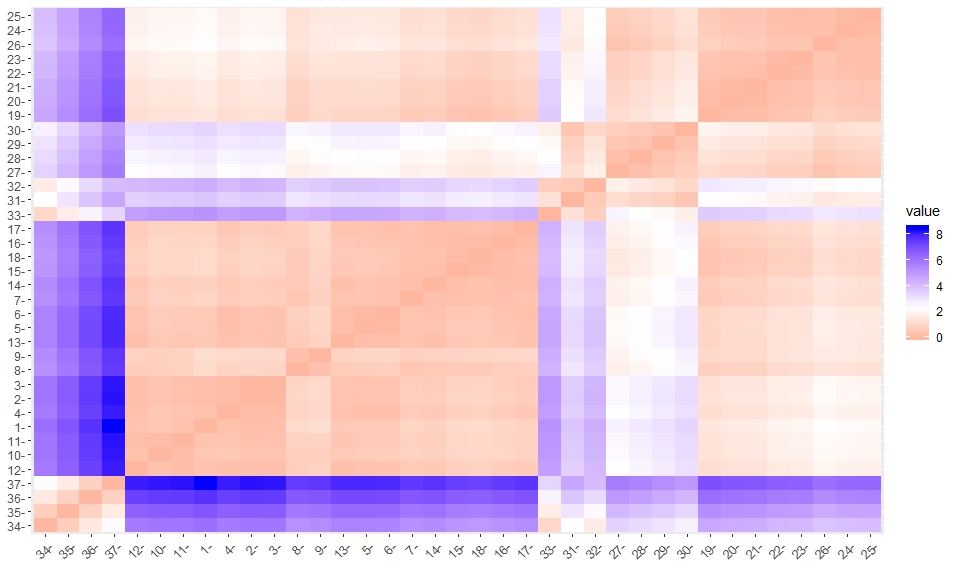


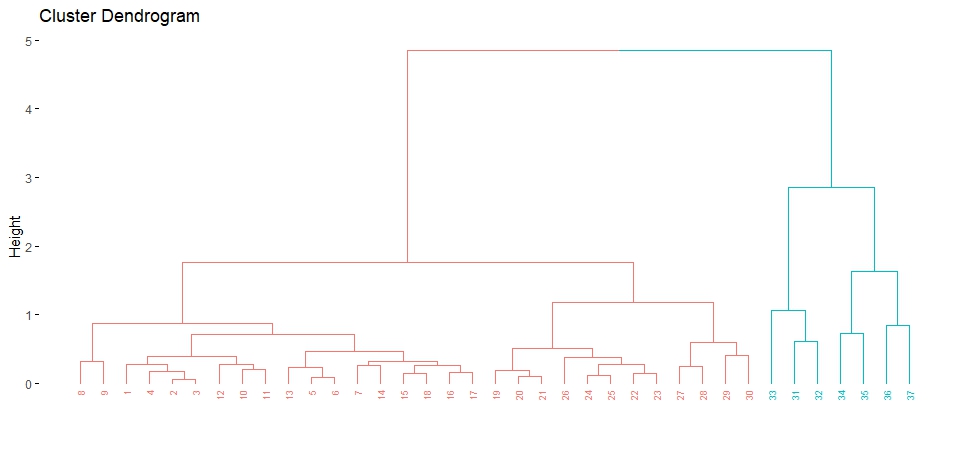
1. Кластеризация

3.1. a) Оптимальное число кластеров – 2 (k-means)

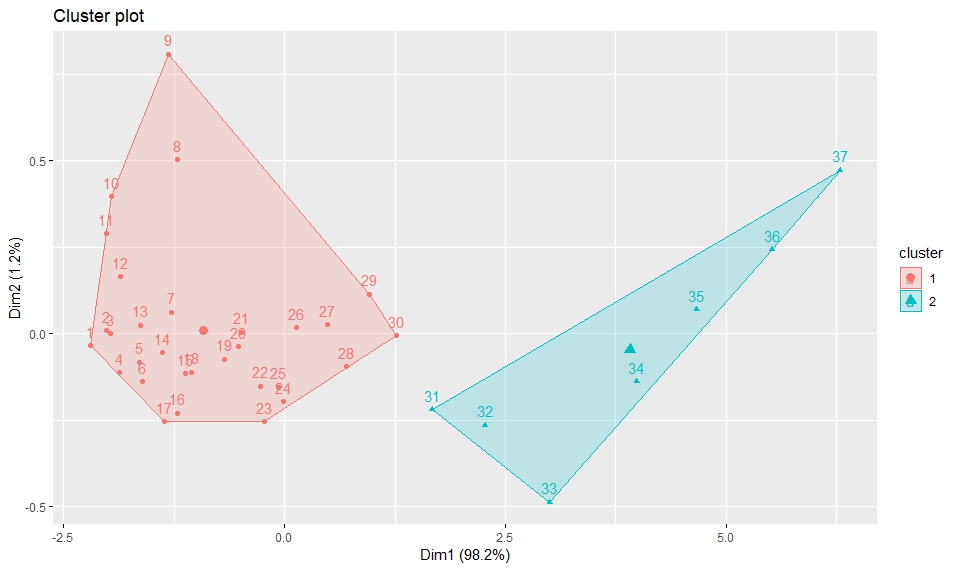


b) Оптимальное число кластеров – 2 (иерархическая кластеризация)





3.2. Карта кластеров (k-means)



На карте выведена кластеризация данных по 2-м кластерам (разбиение по средним), где по осям указаны 2 главные компоненты, одна их которых объясняет 98.2% общей дисперсии (остальные 3 компоненты – еще менее значимы, нежели Dim2. Сами главные компоненты формируются по стандартизированным значениям.

Фактически – разделение на 2 кластера вполне логично, т.к. данные кластеризуются по годам, и во второй кластер выходят данные последний лет анализируемой статистики (с 1981 по 1988 гг.). Это период – когда Китай находился в стадии экономических реформ, начало которым было положено в 1978 г., и соответствующего роста.

1. Модель линейной регрессии

4.1. Значимые коэффициенты

1) Call:

lm(formula = commerce ~ transport + agriculture + industry +

construction, data = mydata\_fct)

Residuals:

Min 1Q Median 3Q Max

-49.782 -15.256 2.454 16.311 38.394

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.03012 36.17423 0.111 0.9120

transport 0.12308 0.15185 0.810 0.4236

agriculture 0.50588 0.30525 1.657 0.1072

industry 0.03694 0.03434 1.076 0.2901

construction 0.14331 0.08123 1.764 0.0872 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.51 on 32 degrees of freedom

Multiple R-squared: 0.9842, Adjusted R-squared: 0.9822

F-statistic: 497.6 on 4 and 32 DF, p-value: < 2.2e-16

Ни одно из значений не отвергается в данной модели, все выше порога 0,05 (0,01).

В данной модели значим “Construction” (однако нулевая гипотеза не отвергается, т.к. р-значения выше минимального порога в 0,05/0,01. Значим коэффициент на уровне 0,05 – 0,1.

Если изменить модель:

Call:

lm(formula = commerce ~ transport, data = mydata\_fct)

Residuals:

Min 1Q Median 3Q Max

-61.957 -16.876 -3.183 24.266 46.665

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 22.72899 7.70759 2.949 0.00565 \*\*

transport 0.53005 0.01391 38.108 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.42 on 35 degrees of freedom

Multiple R-squared: 0.9765, Adjusted R-squared: 0.9758

F-statistic: 1452 on 1 and 35 DF, p-value: < 2.2e-16

Здесь ‘transport’ значим – на уровне 0 – 0,001.

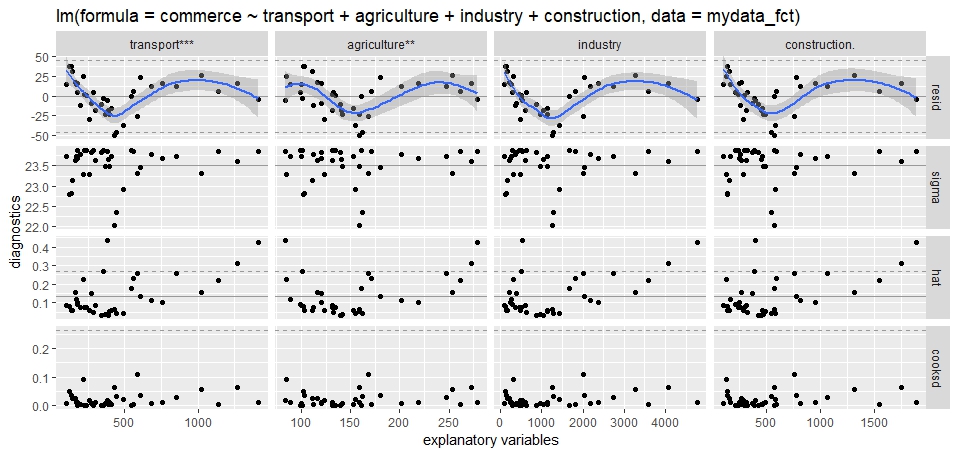
4.2. Сравнение моделей

Сравнивая две приведенные выше модели, можно сказать, что незначительно выигрывает первая (длинная модель), т.к. Adjusted R-squared выше при равных p-value.

Хотя, и та и другая модель – хороши, т.к. значения Adjusted R2 близки к единице, что характеризует малую условную дисперсию, т.е. с большей долей уверенности обе модели отражают реальное положение вещей.

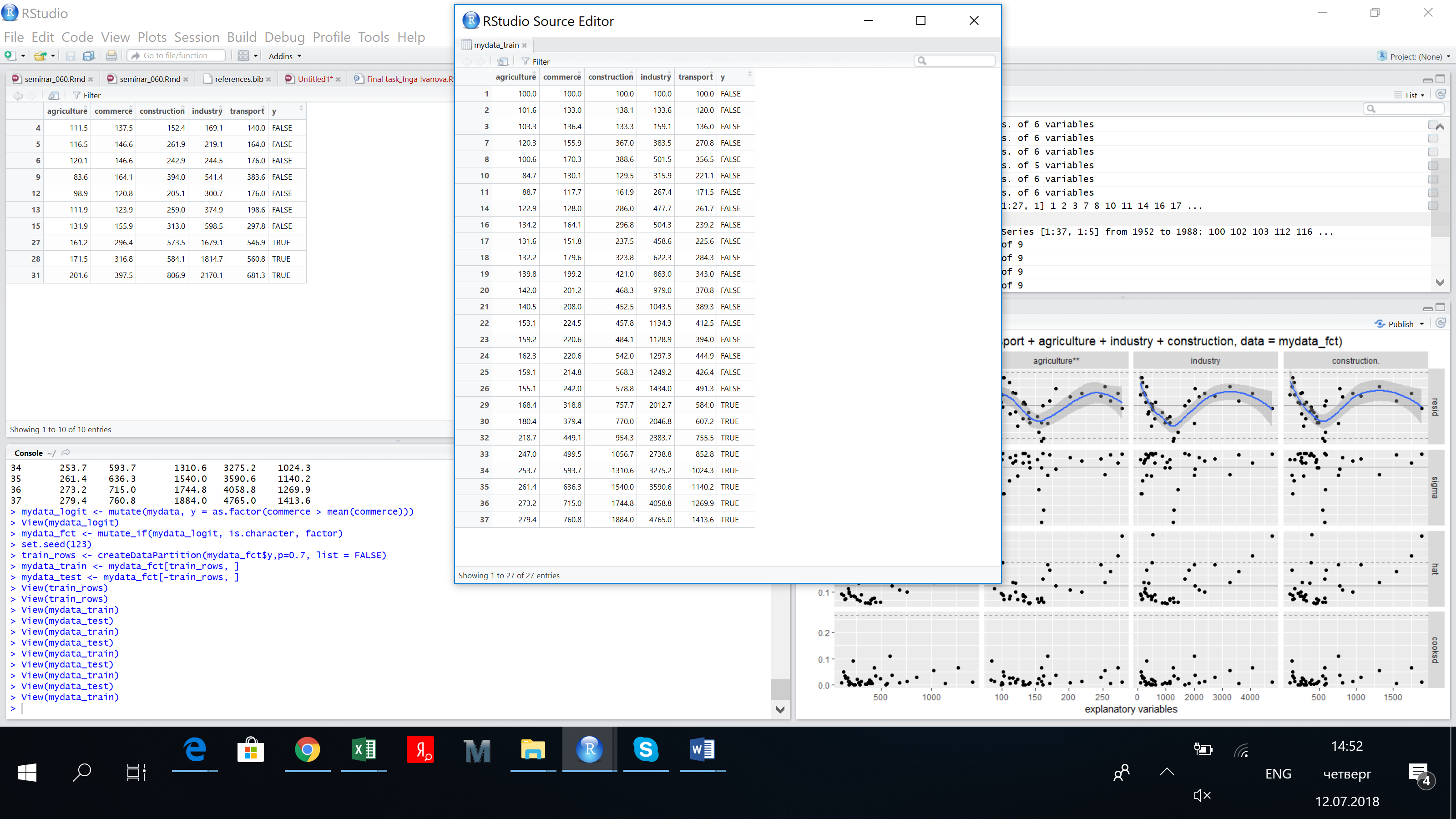
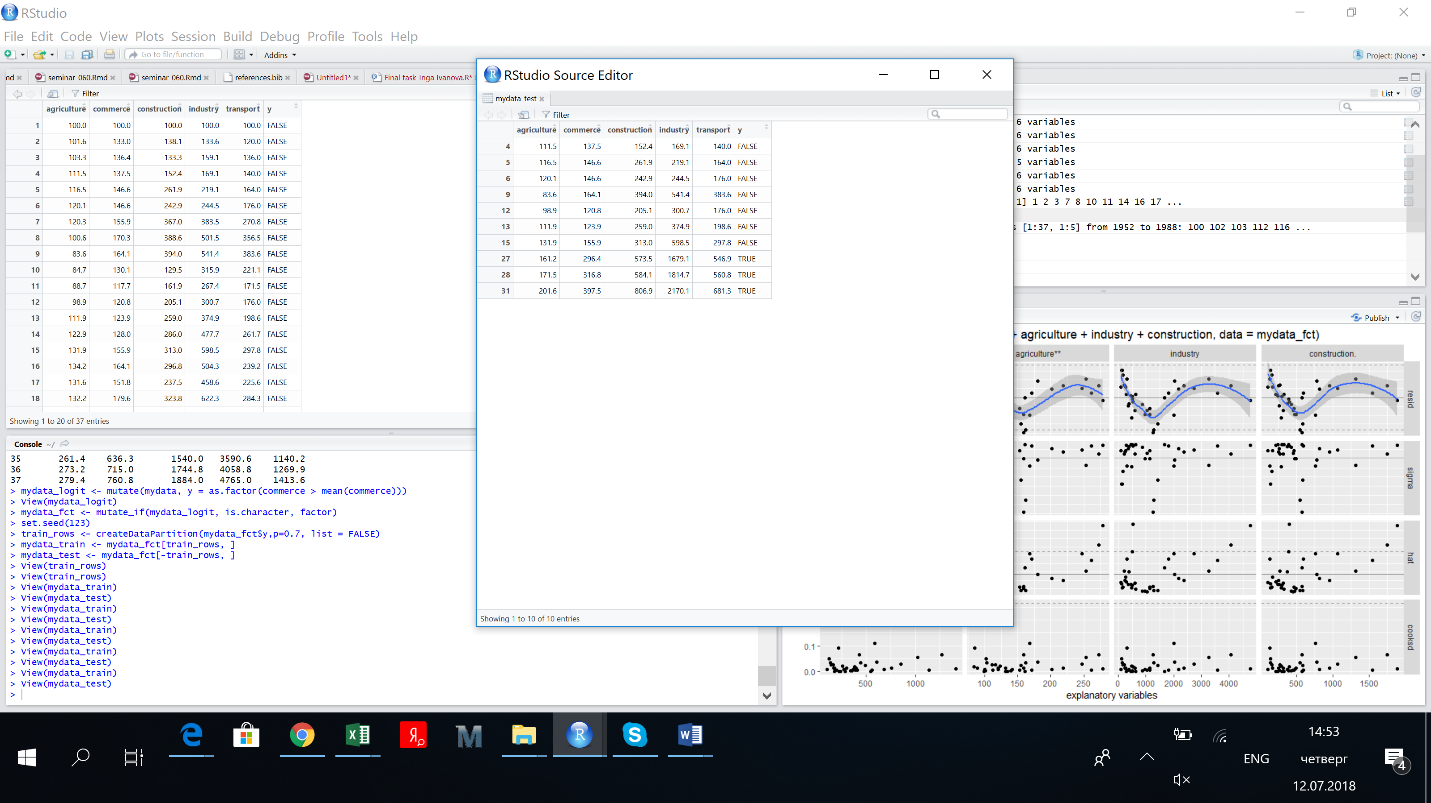
Примечание: Здесь лучше смотреть на Adjusted R2, т.к. R2 имеет недостаток – при увеличении количества предикторов, его значение только возрастает.

Выбросы (для информации, серьезных погрешностей нет):



1. Классификация

5.1. разделение выборки на две части



5.2. Прогноз и оценка качества

> mydata\_pred <- predict(mydata\_lmodel, newdata = mydata\_test)

> head(mydata\_pred)

[1] FALSE FALSE FALSE FALSE FALSE FALSE

Levels: FALSE TRUE

> mydata\_pred <- predict(mydata\_lmodel, newdata = mydata\_test, type = 'prob')

> head(mydata\_prob)

Error in head(mydata\_prob) : object 'mydata\_prob' not found

> mydata\_prob <- predict(mydata\_lmodel, newdata = mydata\_test, type = 'prob')

> head(mydata\_prob)

FALSE TRUE

3 1.000 0.000

4 1.000 0.000

5 1.000 0.000

8 0.998 0.002

11 1.000 0.000

13 1.000 0.000

Модель имеет высокую степень вероятности точности прогнозных значений.

Confusion Matrix and Statistics

Reference

Prediction FALSE TRUE

FALSE 7 0

TRUE 0 3

Accuracy : 1

95% CI : (0.6915, 1)

No Information Rate : 0.7

P-Value [Acc > NIR] : 0.02825

Kappa : 1

Mcnemar's Test P-Value : NA

Sensitivity : 1.0

Specificity : 1.0

Pos Pred Value : 1.0

Neg Pred Value : 1.0

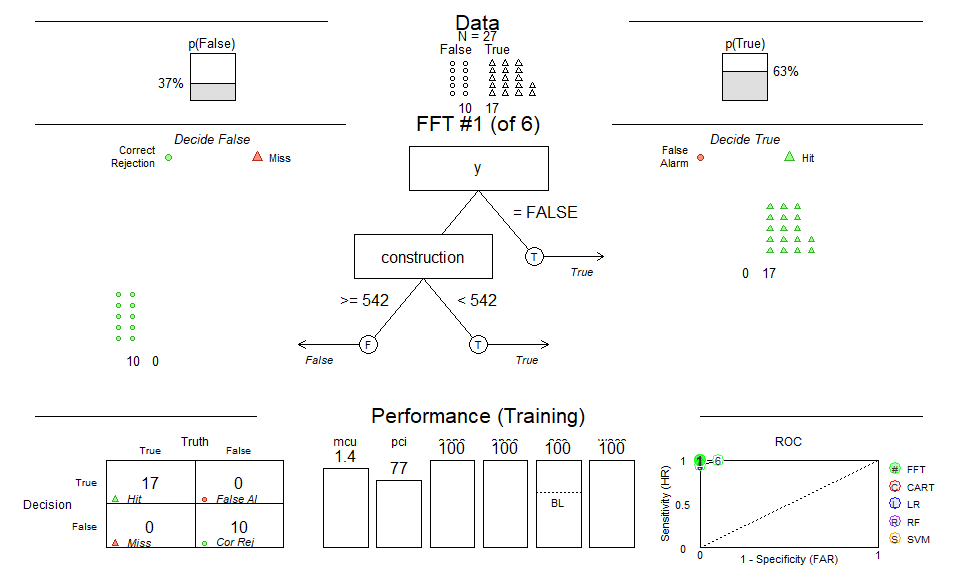
Prevalence : 0.7

Detection Rate : 0.7

Detection Prevalence : 0.7

Balanced Accuracy : 1.0

'Positive' Class : FALSE



5.3. ROC-кривая

