

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего
образования

Национальный исследовательский университет ИТМО
Факультет инфокоммуникационных технологий

Отчет

Проектирование интеллектуальных систем в управлении Практическая работа №2

Студент:

Катаева Вероника Алексеевна

Группа: К3342

Преподаватель:

Бережков Андрей Вячеславович

Санкт-Петербург

2021

Цель работы: Провести анализ датасета «Титаник» в KNIME.

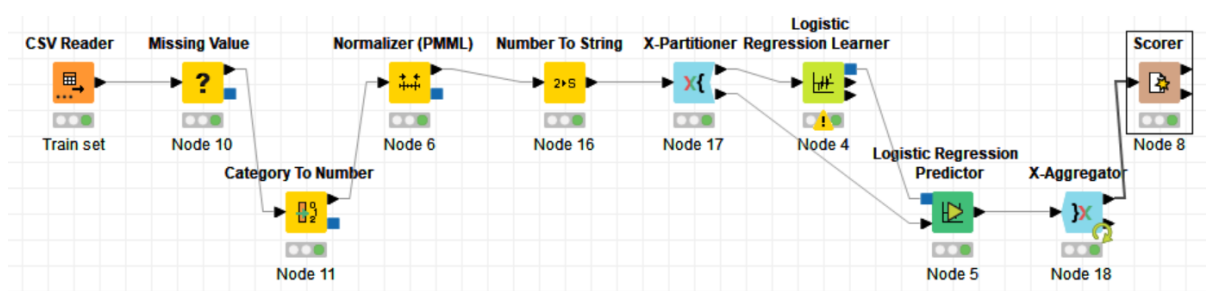
Задачи:

1. Необходимо выстроить процессы аналогично тетрадке (<https://www.kaggle.com/pramodsivakumar/titanic-simple-eda-and-predictions>).

Выполнение работы:

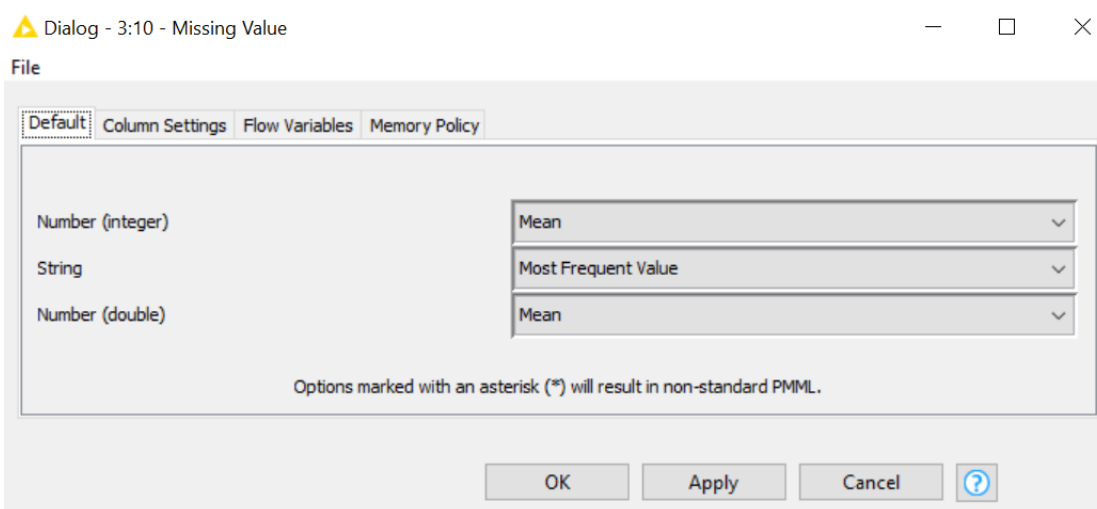
Построение моделей:

Поток работ линейной регрессии:



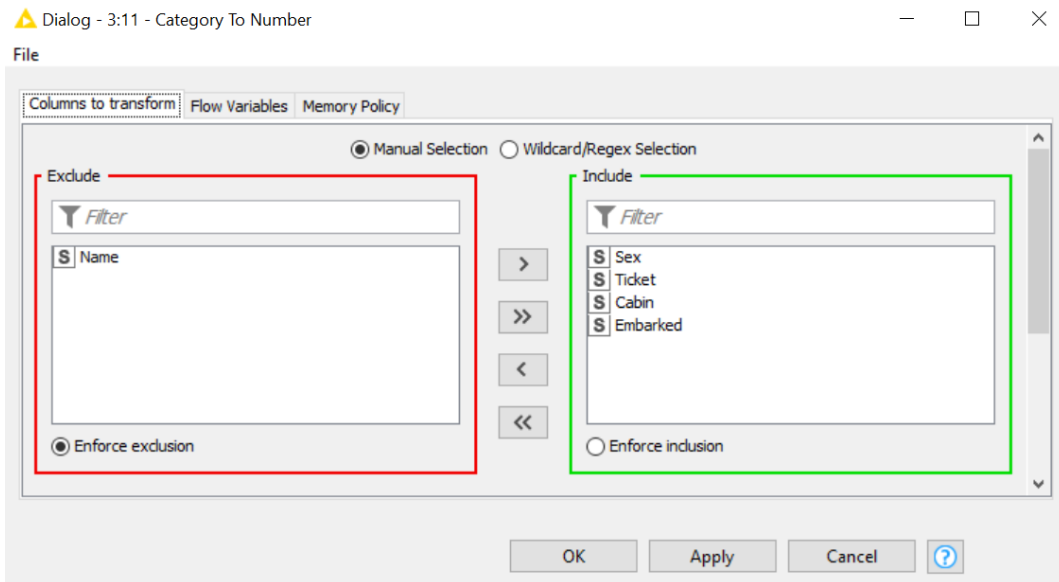
Состоит из следующих узлов:

- 1) Чтение csv файла – данных о Титанике.
- 2) Обработка недостающих значений.

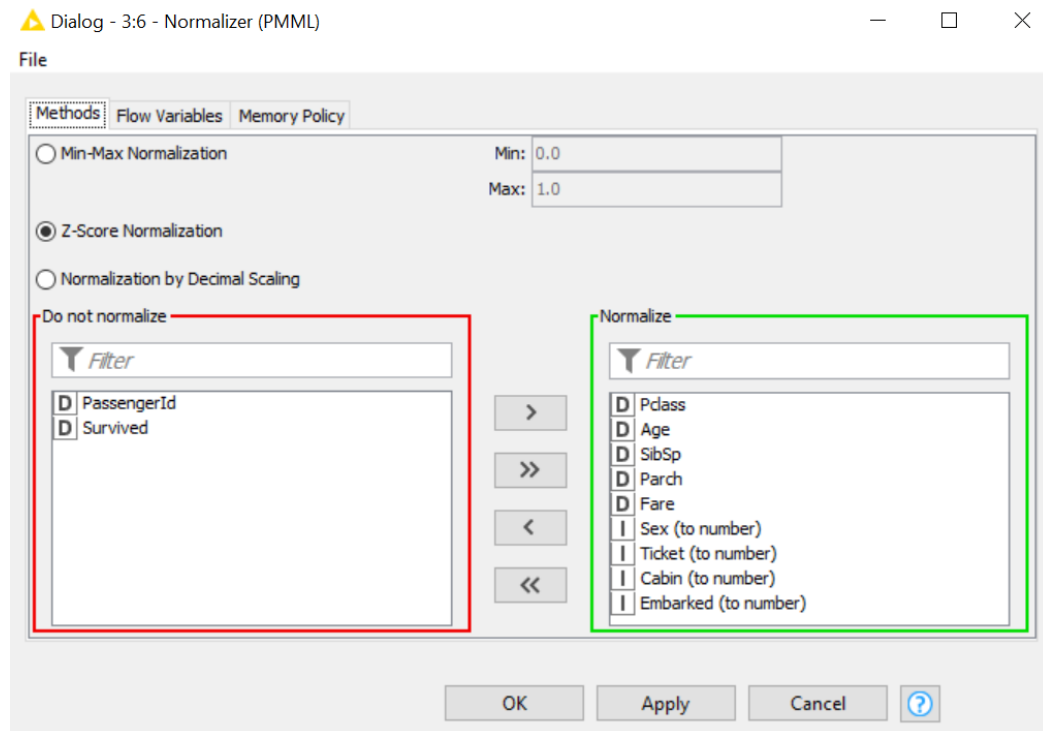


Числа заменяются на среднее значение, строки – на самое популярное.

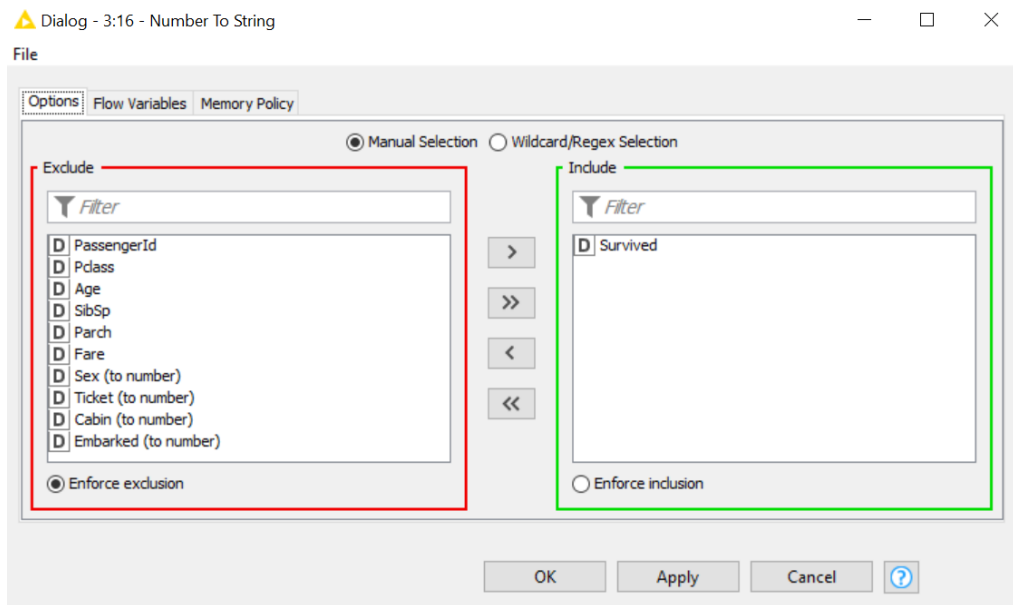
- 3) Замена категориальных данных на числовые (за исключением имени пассажира).



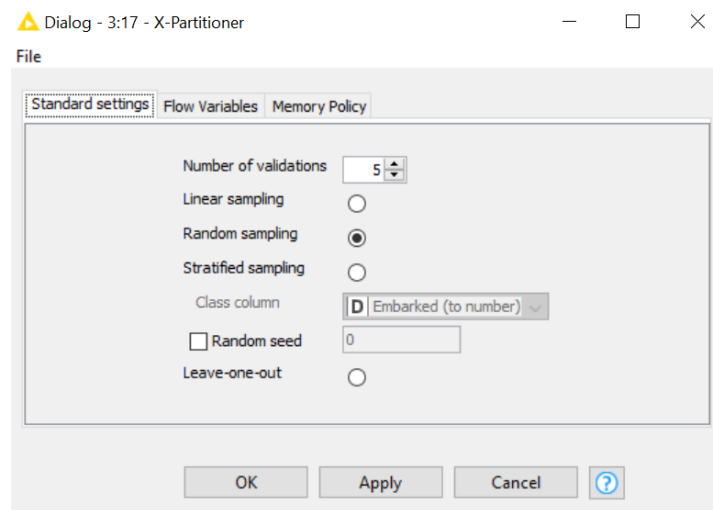
- 4) Нормализация данных (за исключением ID пассажира и целевую переменную).



- 5) Конвертация числа в строку (для создания номинальных данных для предсказания).



6, 9) 5-кратная кросс-валидация.



7, 8) Логистическая регрессия.

File

Settings | Advanced | Flow Variables | Memory Policy

Target

Target column: ▼

Reference category: ▼

☐ Use order from target column domain (only relevant for output representation)

Solver

Select solver: ▼

Feature selection

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude

- ☒ PassengerId
- ☒ Name

☒ Enforce exclusion

Include

- ☒ Pclass
- ☒ Sex
- ☒ Age
- ☒ SibSp
- ☒ Parch
- ☒ Ticket
- ☒ Fare
- ☒ Cabin

☐ Enforce inclusion

☐ Use order from column domain (applies only to nominal columns). First value is chosen as reference for dummy variables.

OK Apply Cancel ?

Гиперпараметры логистической регрессии:

File

Settings | **Advanced** | Flow Variables | Memory Policy

Solver options

☒ Perform calculations lazily (more memory expensive but often faster)

☒ Calculate statistics for coefficients

Termination conditions

Maximal number of epochs: ▼

Epsilon:

Learning rate / step size

Learning rate strategy: ▼

Step size:

Regularization

Prior: ▼

Variance: ▼

Data handling

☒ Hold data in memory

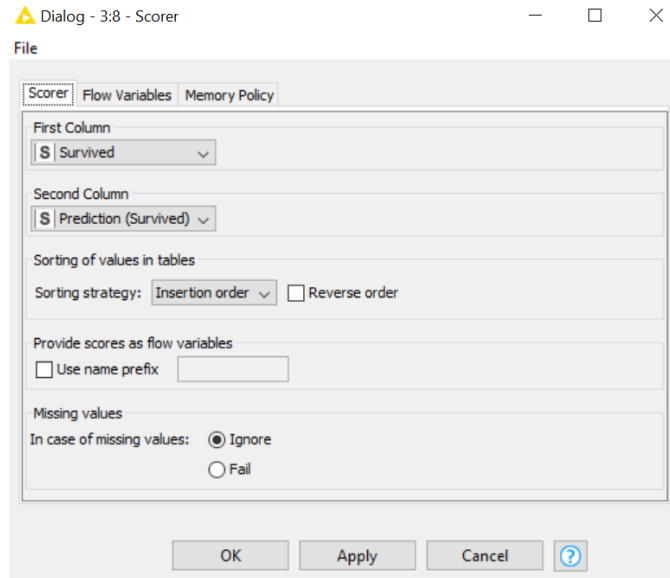
Chunk size: ▼

☐ Use seed

Seed: New

OK Apply Cancel ?

10) Счетчик.



Результаты логистической регрессии:

Confusion Matrix - 3:8 - Scorer

File Hilite

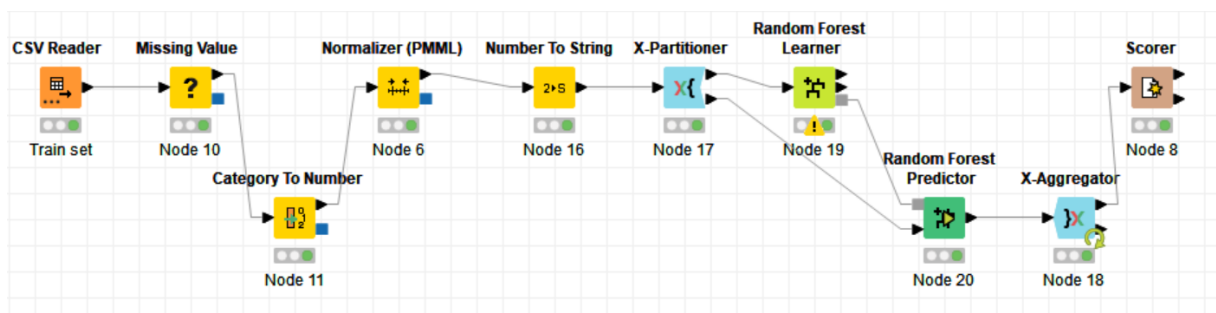
Survived \ ...	0	1
0	466	83
1	104	238

Correct classified: 704 Wrong classified: 187

Accuracy: 79,012 % Error: 20,988 %

Cohen's kappa (κ) 0,551

Поток работ классификации с помощью случайного леса:



Поток работ содержит схожие узлы – вместо линейной регрессии используется модель случайного леса.

Конфигурация и гиперпараметры случайного леса:

File

Options
Flow Variables
Memory Policy

Target Column
S Survived

Attribute Selection

☐ Use fingerprint attribute

☒ Use column attributes

☒ Manual Selection
☐ Wildcard/Regex Selection

Exclude

Filter

D PassengerId
S Name

☒ Enforce exclusion

Include

Filter

D Pclass
S Sex
D Age
D SibSp
D Parch
S Ticket
D Fare
S Cabin

☐ Enforce inclusion

Tree Options
Split Criterion
Information Gain Ratio

☐ Limit number of levels (tree depth)

☐ Minimum node size

Forest Options
Number of models
100

☒ Use static random seed
161609242231
New

OK
Apply
Cancel
?

Результаты случайного леса:

Confusion Matrix - 3:8 - Scorer

— □ ×

File Hilite		
Survived \ ...	0	1
0	496	53
1	102	240

Correct classified: 736

Wrong classified: 155

Accuracy: 82,604 %

Error: 17,396 %

Cohen's kappa (κ) 0,622

Поток работ метода ближайших соседей:

```

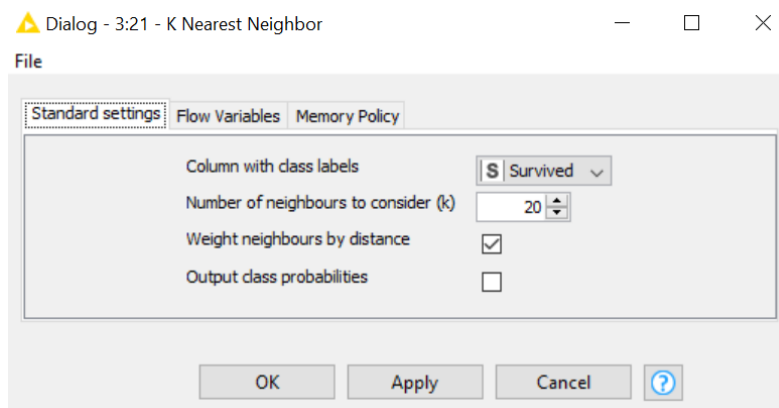
graph LR
    CSV[CSV Reader] --> Missing[Missing Value]
    Missing --> Normalizer[Normalizer PMML]
    Missing --> Category[Category To Number]
    Normalizer --> Number[Number To String]
    Number --> Partition[X-Partitioner]
    Partition --> Neighbor[K Nearest Neighbor]
    Neighbor --> Aggregator[X-Aggregator]
    Aggregator --> Scorer[Scorer]
    
```

The flowchart illustrates the workflow for the K-Nearest Neighbor method. It starts with a 'CSV Reader' (Node 8) loading the 'Train set'. The data then passes through a 'Missing Value' node (Node 10). From there, it splits into two paths: one through a 'Normalizer (PMML)' (Node 6) and another through a 'Category To Number' node (Node 11). Both paths converge and then pass through a 'Number To String' node (Node 16), an 'X-Partitioner' (Node 17), a 'K Nearest Neighbor' node (Node 21), and an 'X-Aggregator' (Node 18), finally reaching the 'Scorer' (Node 8).

7

Поток работ содержит схожие узлы – вместо линейно регрессии используется модель k ближайших соседей.

Конфигурация и гиперпараметры метода ближайших соседей:



Результаты метода ближайших соседей:

Confusion Matrix - 3:8 - Scorer

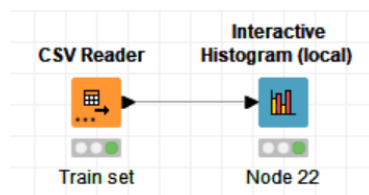
Survived \ ...	0	1
0	505	44
1	305	37

Correct classified: 542	Wrong classified: 349
Accuracy: 60,831 %	Error: 39,169 %
Cohen's kappa (κ) 0,033	

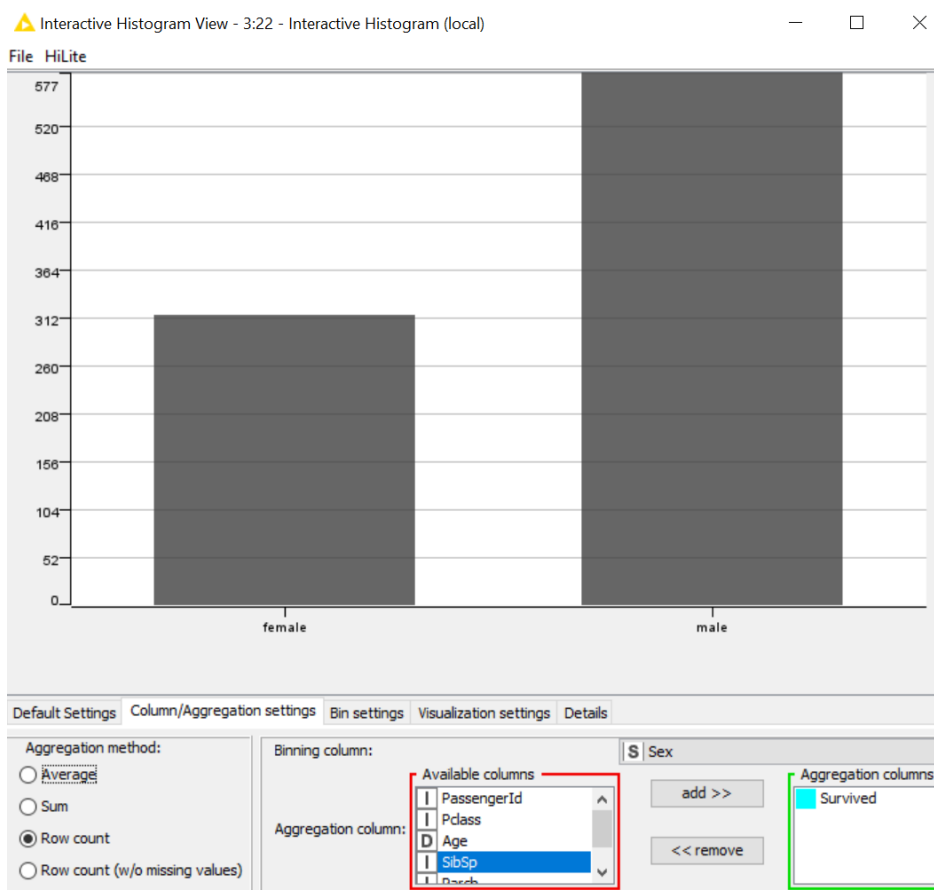
Вывод: таким образом, модель случайного леса в большей степени подходит для представленных данных, так как обладает наивысшей в сравнении с другими моделями долей правильных ответов и лучшими показателями матрицы ошибок.

Визуализация:

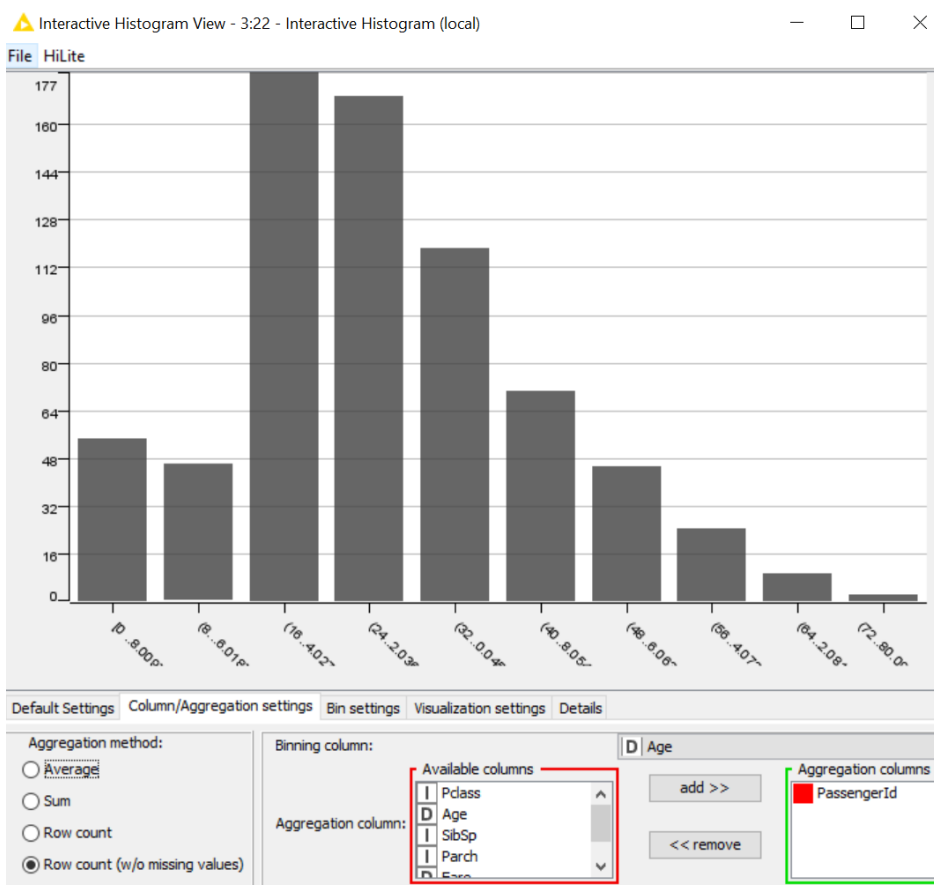
Поток работ гистограммы:



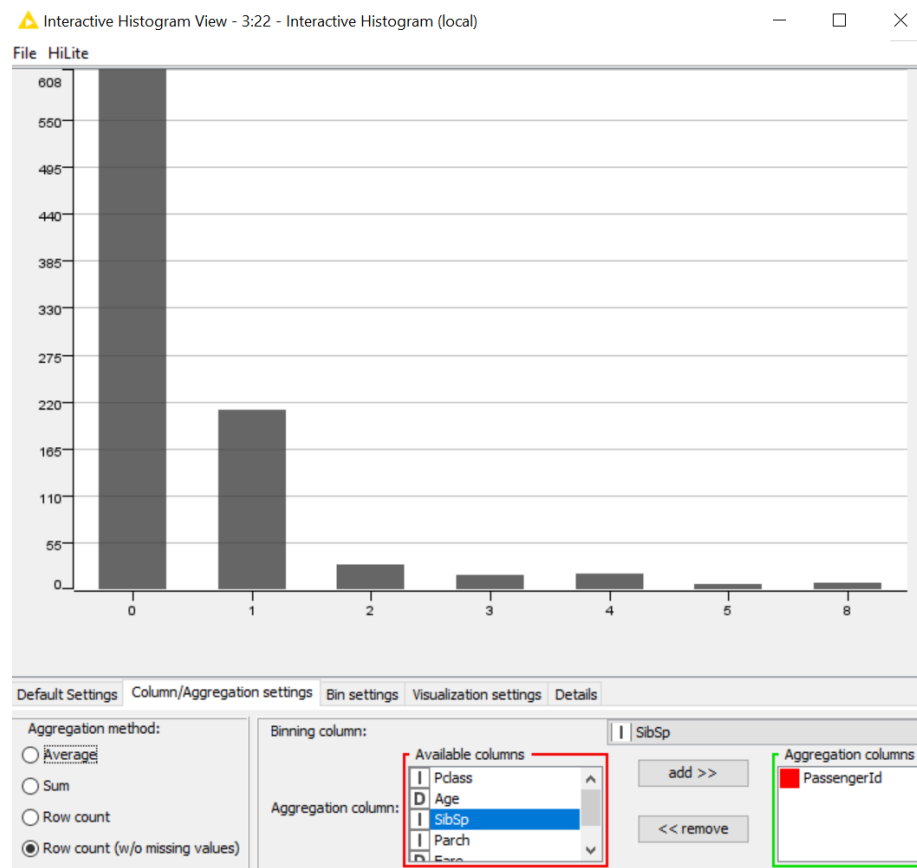
Гистограмма распределения пола в зависимости от того, что человека не спасли:



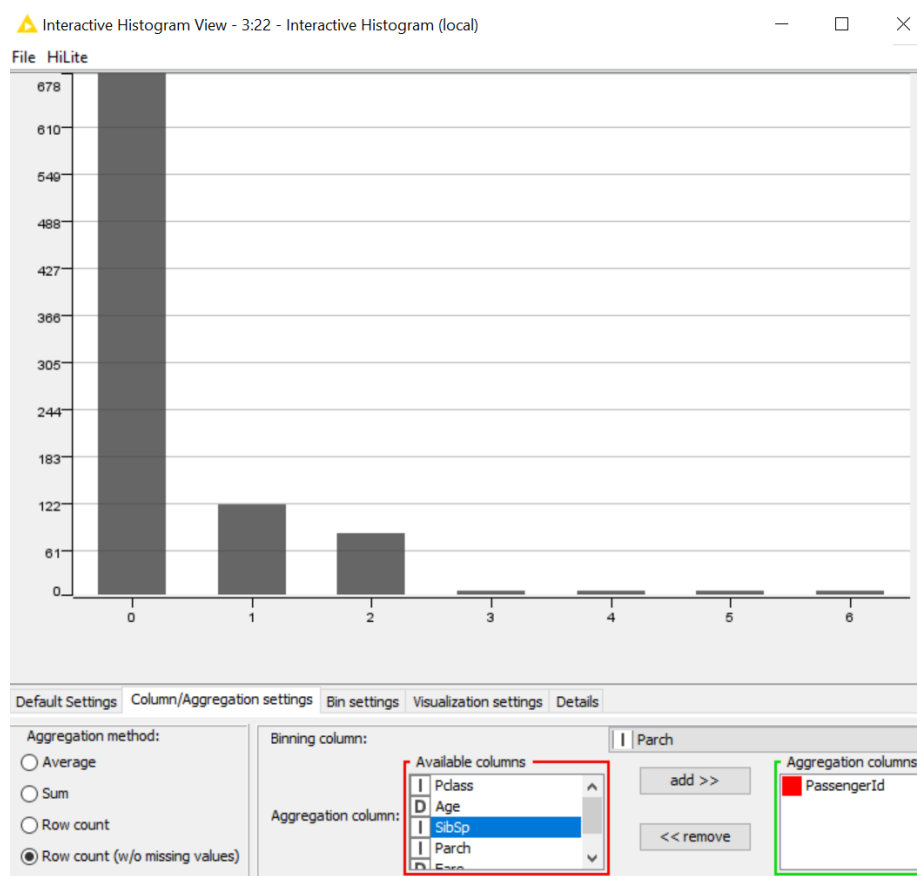
Гистограмма распределения возрастов:



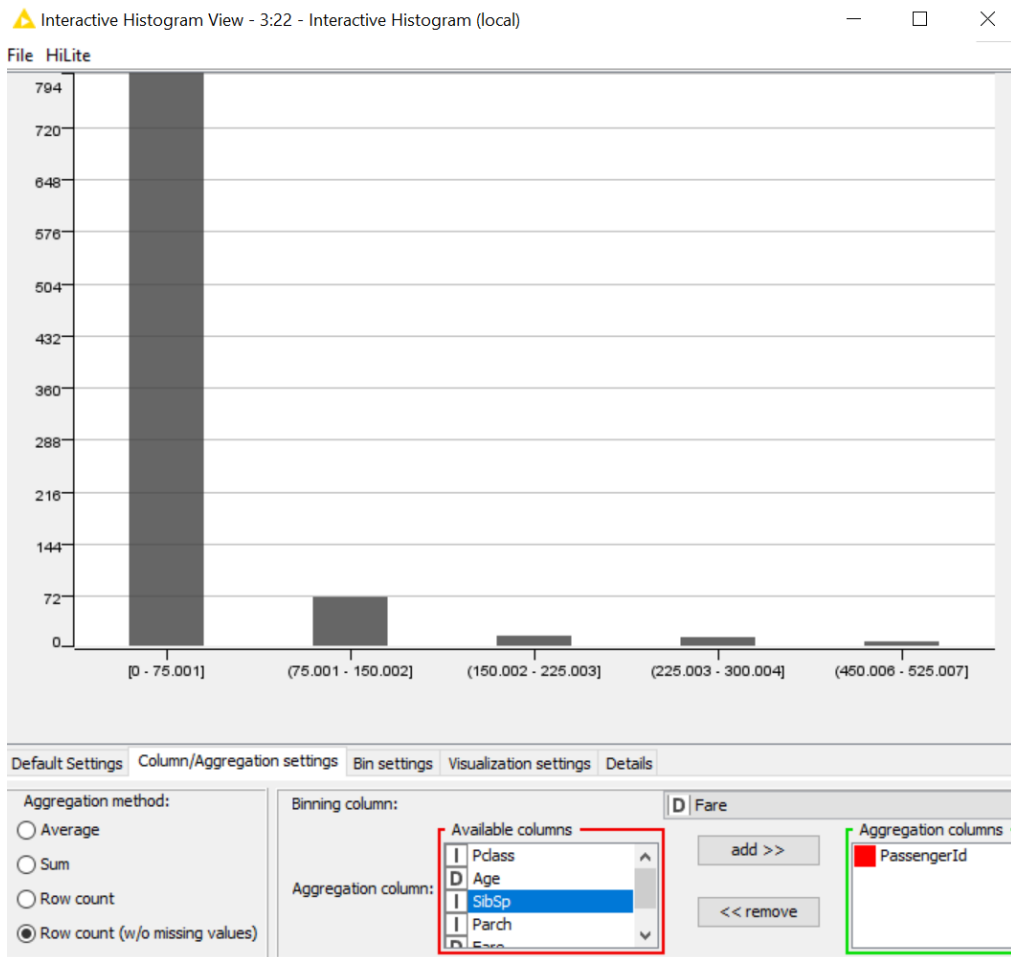
Распределение количества братьев, сестёр или наличия супруга на борту:



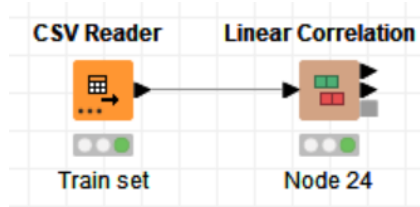
Распределение количества родителей, детей на борту:



Пассажирский тариф в британских фунтах стерлингов:

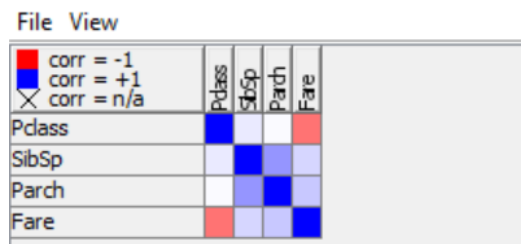


Поток работ линейной корреляции:

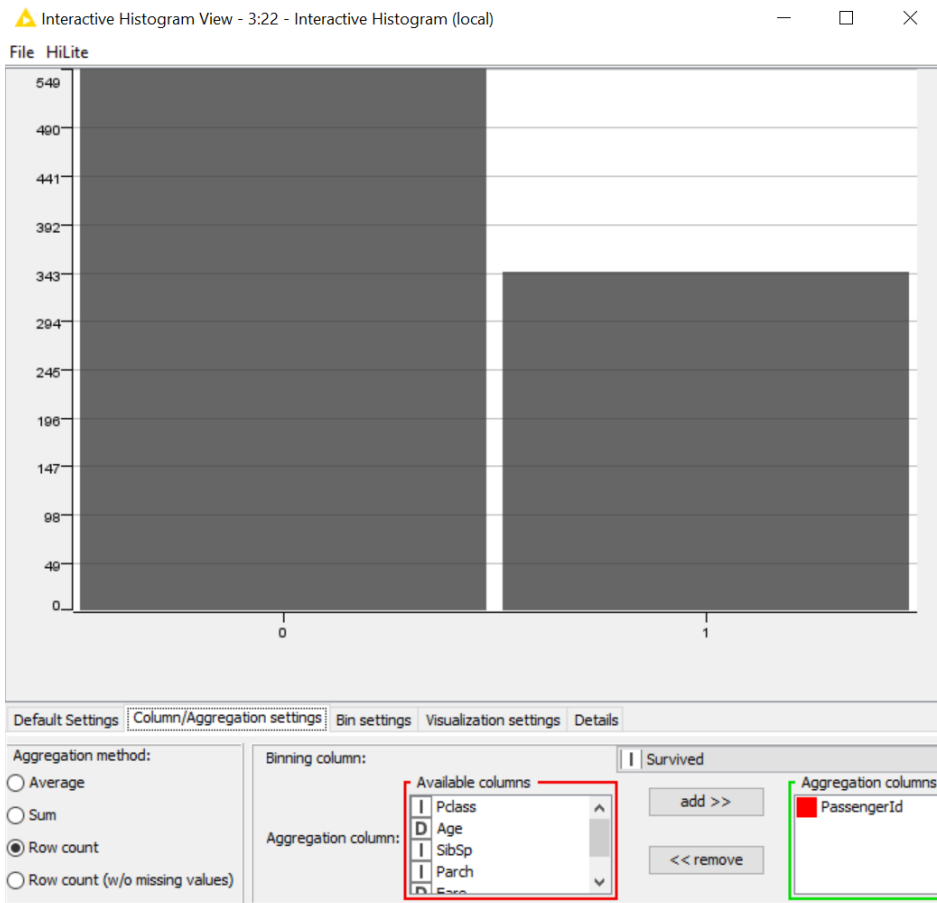


Матрица корреляции между классом пассажира, наличием его братьев/сестер/супруга на борту, наличием его родителей/детей, пассажирским тарифом:

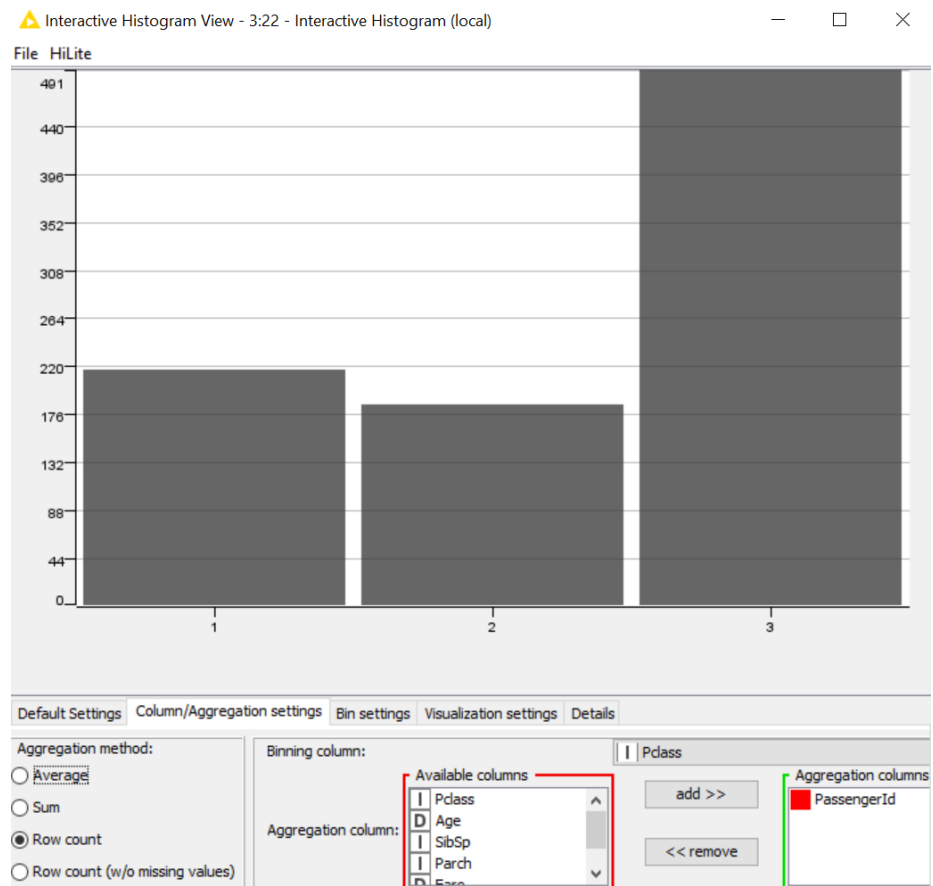
Correlation Matrix - 3:24 - Linear Correlation



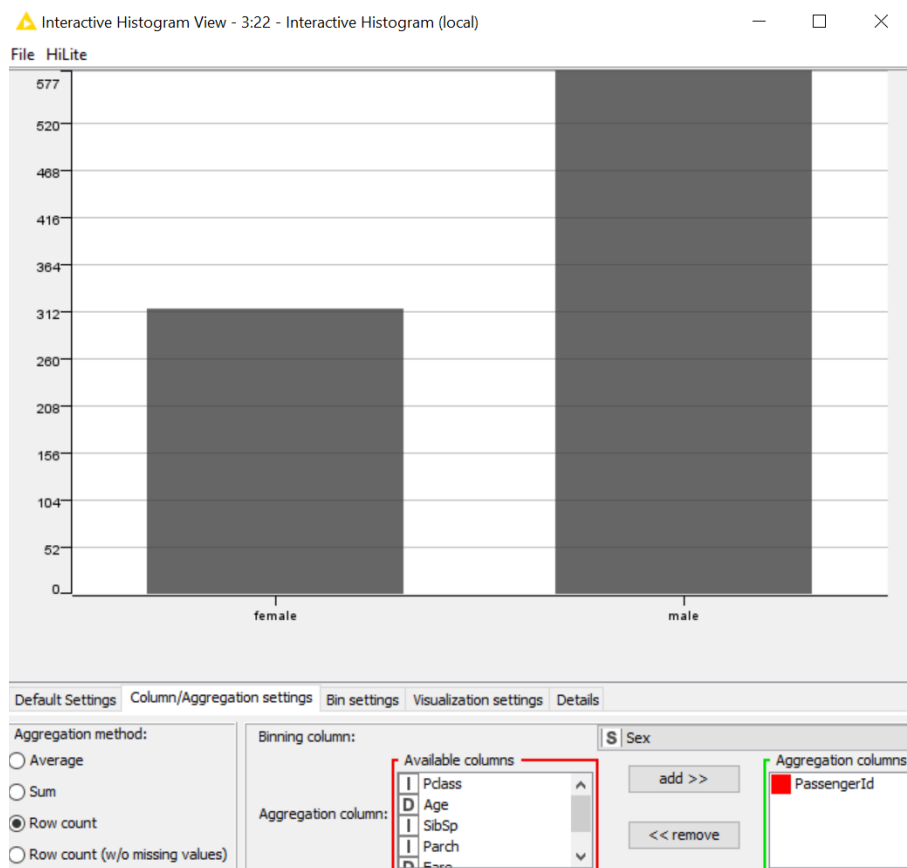
Распределение выживших:



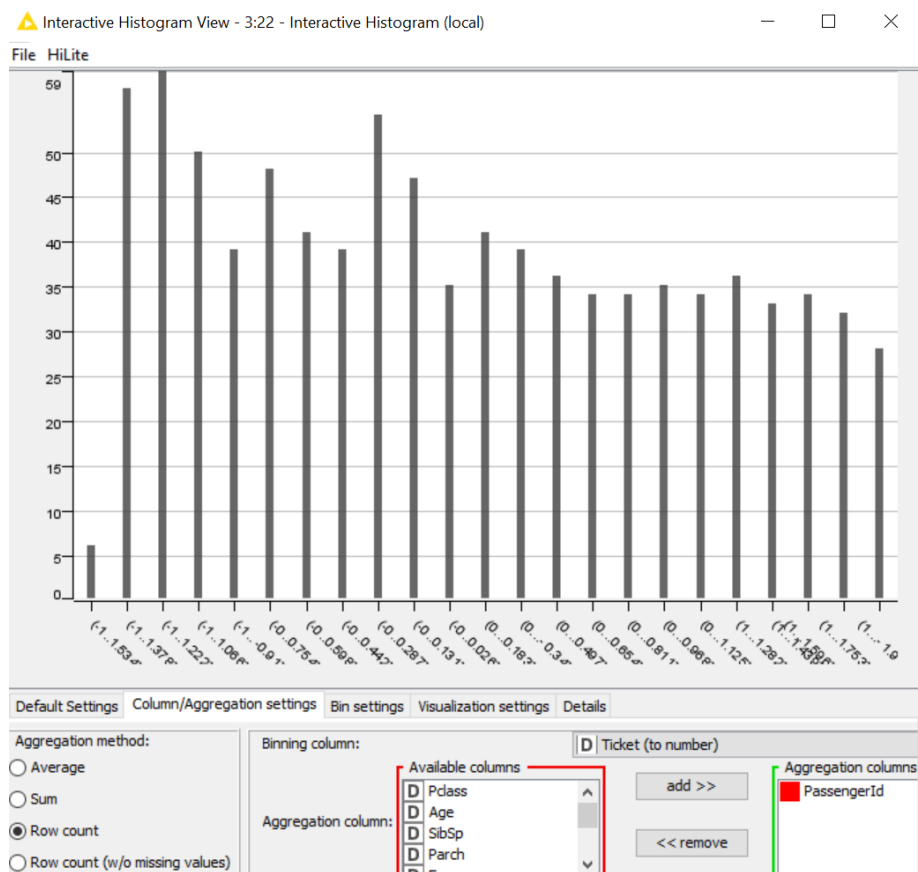
Распределение пассажиров по значению класса:



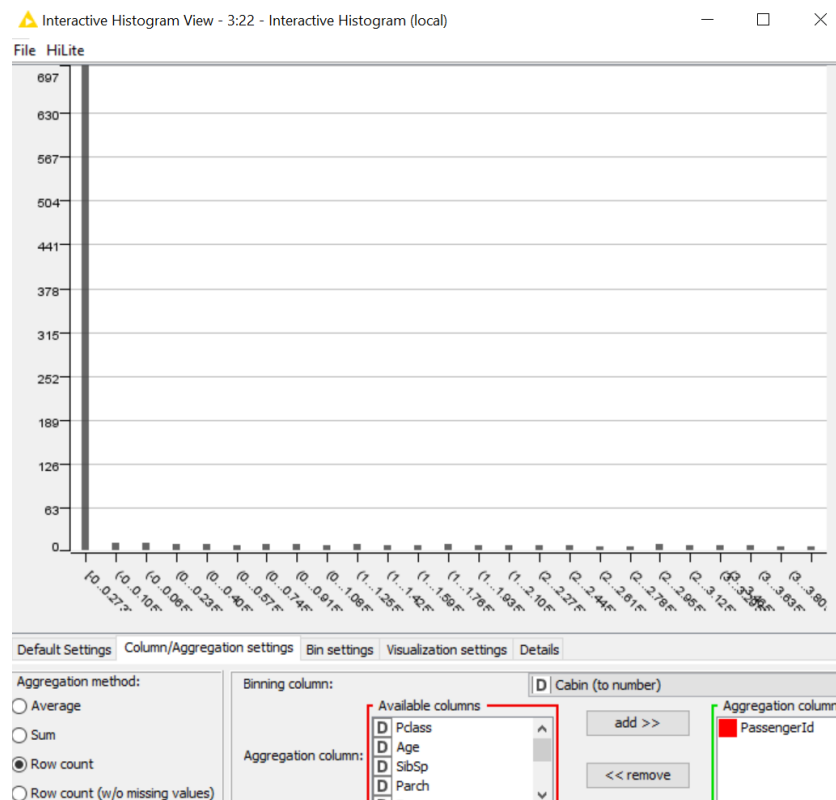
Распределение пассажиров по половому признаку:



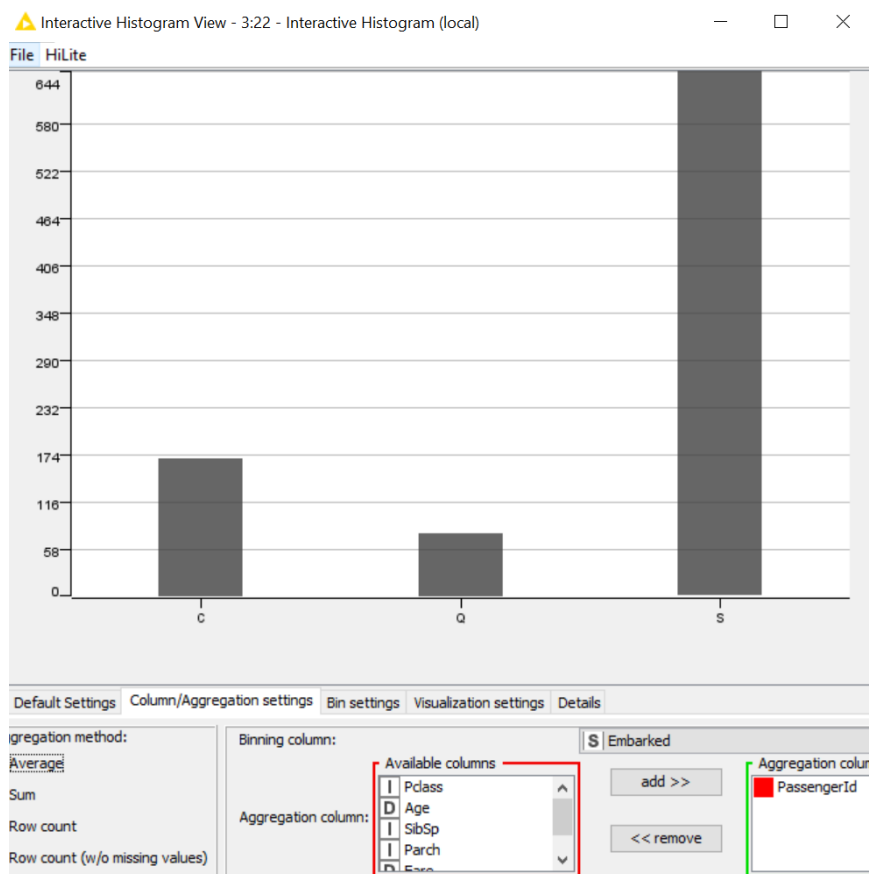
Распределение пассажиров в зависимости от билета (нормализованные значения):



Распределение пассажиров в зависимости от каюты (нормализованные значения):



Распределение пассажиров в зависимости от порта посадки (C = Шербур; Q = Квинстаун; S = Саутгемптон):



Вывод: в ходе лабораторной работы я познакомилась с инструментарием KNIME, а в частности с моделями для анализа и визуализации данных.