GUI для предобработки данных

Егор Суворов

Руководитель: Тимофей Брыксин

Практика, весна 2017

Среда, 21 июня 2017 года

Задача

Ситуация:

- Имеется набор данных
- Хочется быстро применить методы «машинного обучения» для поиска закономерностей
- Желания разбираться в теории нет
- Важно получить хоть какой-то результат быстро

Пример: хочется проверить, нет ли каких-то простых закономерностей в данных пациентов.

Возможные решения

- Отдать анализ данных внешним исполнителям
  - Медленно
  - Теряется знание о контексте данных
- Ручной анализ в Excel
  - Надо знать формулы
  - Сложно вносить изменения
- Python/R/Java с библиотеками для ML
  - Требуется уметь программировать
- Популярные специализированные решения:
  - Локальные: Weka, Orange, KNIME, RapidMiner, Elki
  - SaaS: Azure ML Studio, BigML, MLJAR

Частые проблемы

Было протестировано восемь специализированных решений.

- Присутствующий GUI является оболочкой поверх «низкоуровневых» действий.
- Требуется понимание основных шагов ML и их необходимость.
- В текстовых языках надо помнить названия команд.
- В блок-схемах требуется соблюдать типы подключений.
- Очень легко что-то неправильно подключить и получить непонятную ошибку.
- Ошибки выдаются без руководства к действию
  - Input example set must have special attribute 'label'

Желаемый результат

Проверка осуществимости требований:

- Набор программ для применения простого ML к данным
- Фокус на получение хоть какого-нибудь результата, а не наилучшего
- Программы направляют действия пользователя
- Не требуется предварительных знаний об обработке данных

Подзадачи:

1. Установка инструментария
2. **Предобработка данных**
3. Применение методов ML (работа Екатерины Малютиной, 5 курс)

Используемые инструменты

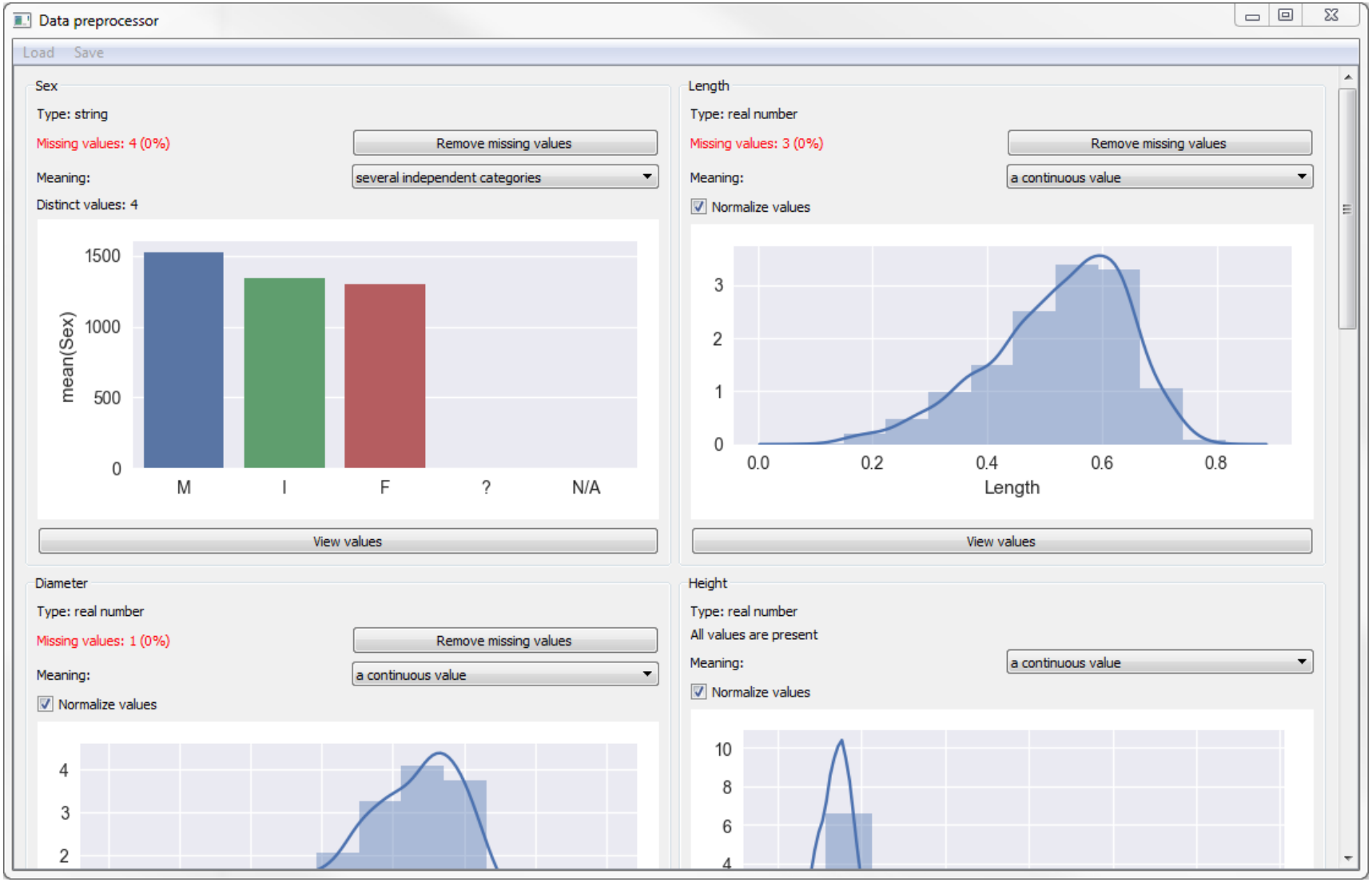
- Язык программирования — Python
- Обработка и хранение данных:
  - Numpy — быстрая обработка данных для Python
  - Pandas — более удобный интерфейс для Numpy
  - Seaborn — визуализация данных
- Интерфейс — PyQt

Что получилось

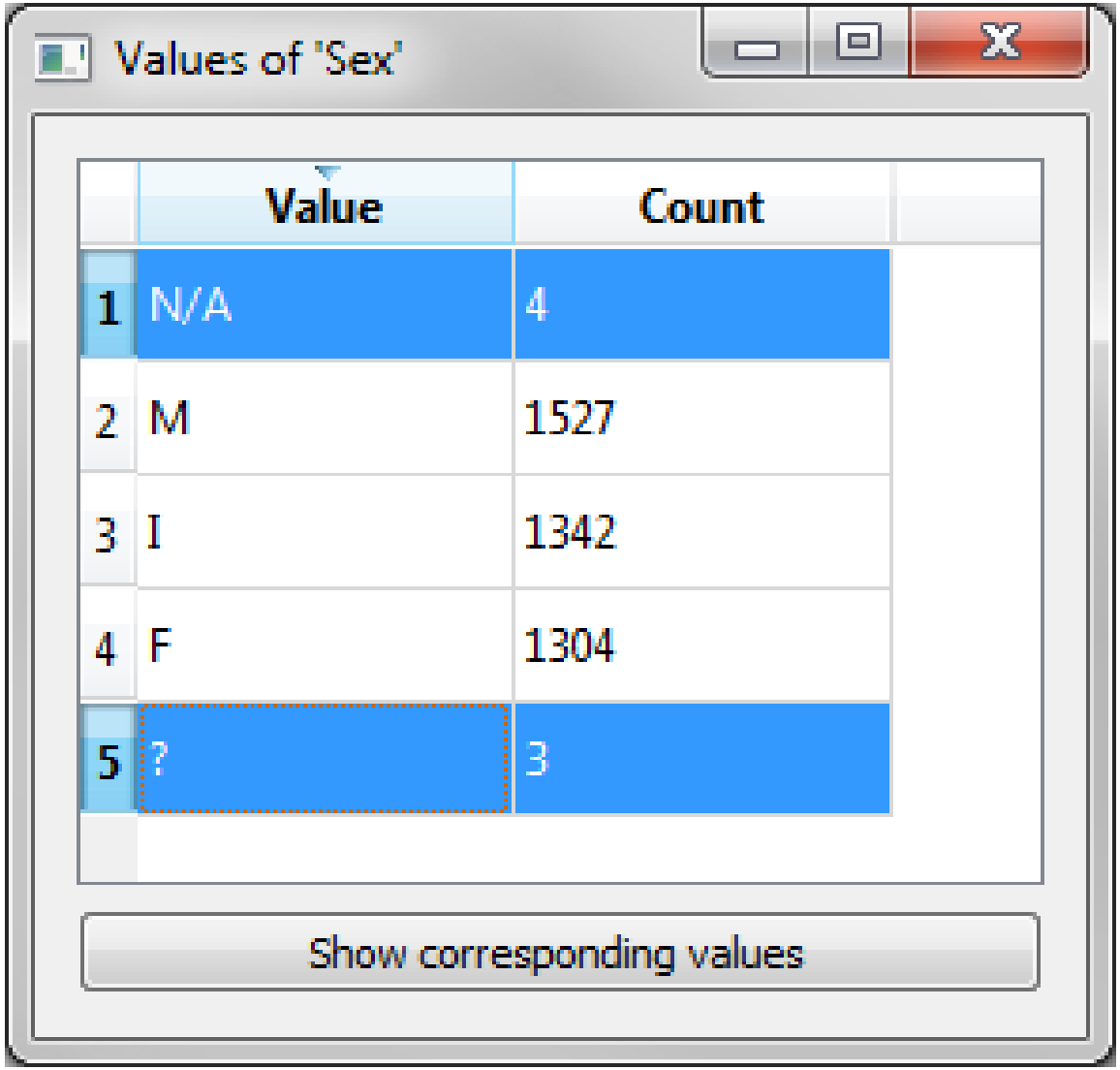
Прототип:

- Загрузка и сохранение данных в формате CSV
- Построение гистограмм по столбцам данных
- Фильтрация данных по значениям
- Удаление строк из набора
- Кодирование столбцов методом one-hot
- Нормализация числовых столбцов

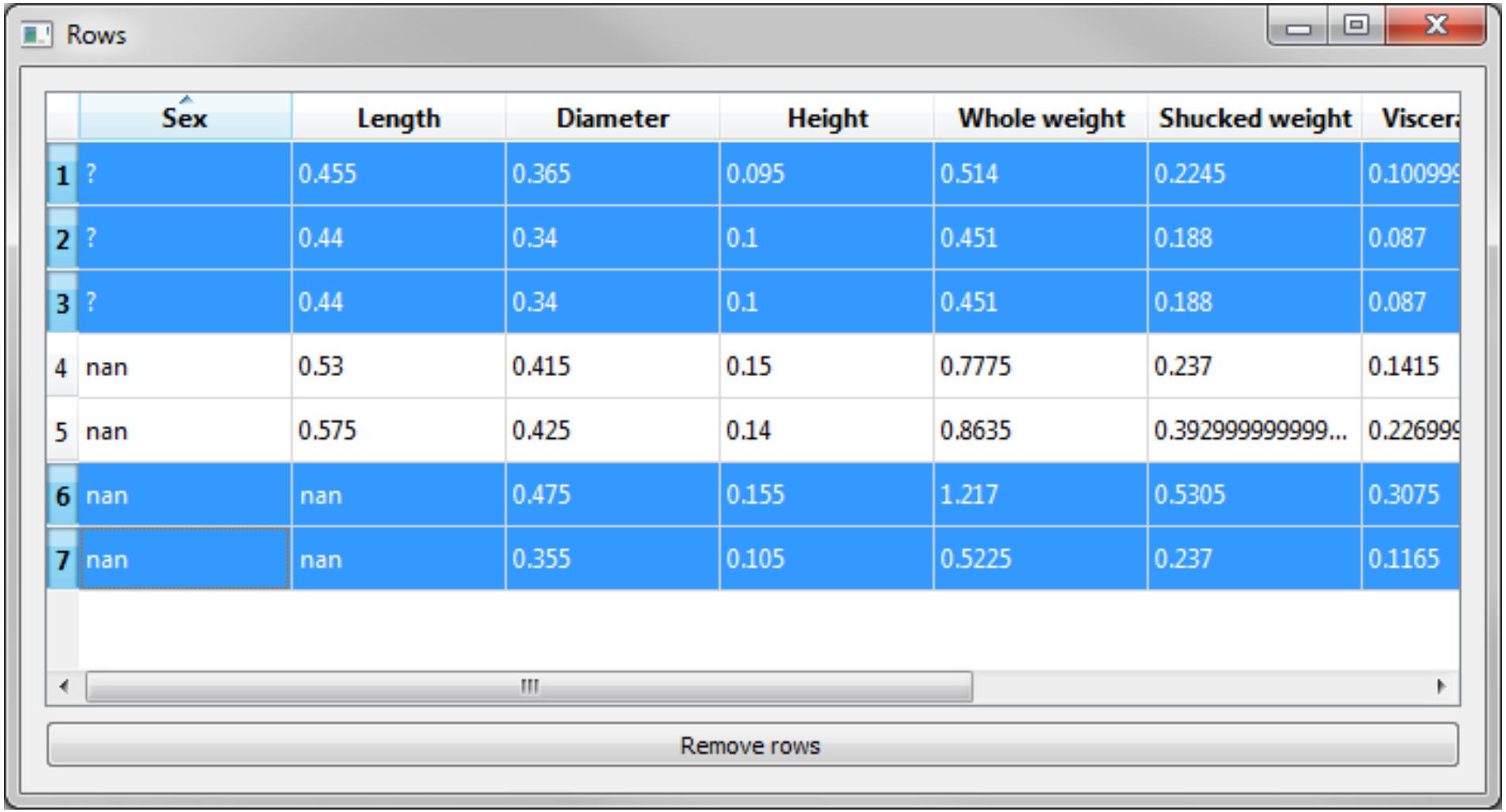
Скриншоты


Data preprocessor
Load
Save
Sex
Type: string
Missing values: 4 (0%)
Remove missing values
Meaning:
several independent categories
Distinct values: 4
1500
1000
500
0
mean(Sex)
M
I
F
?
N/A
View values
Length
Type: real number
Missing values: 3 (0%)
Remove missing values
Meaning:
a continuous value
Normalize values
3
2
1
0
0.0
0.2
0.4
0.6
0.8
Length
View values
Diameter
Type: real number
Missing values: 1 (0%)
Remove missing values
Meaning:
a continuous value
Normalize values
4
3
2
Height
Type: real number
All values are present
Meaning:
a continuous value
Normalize values
10
8
6
4

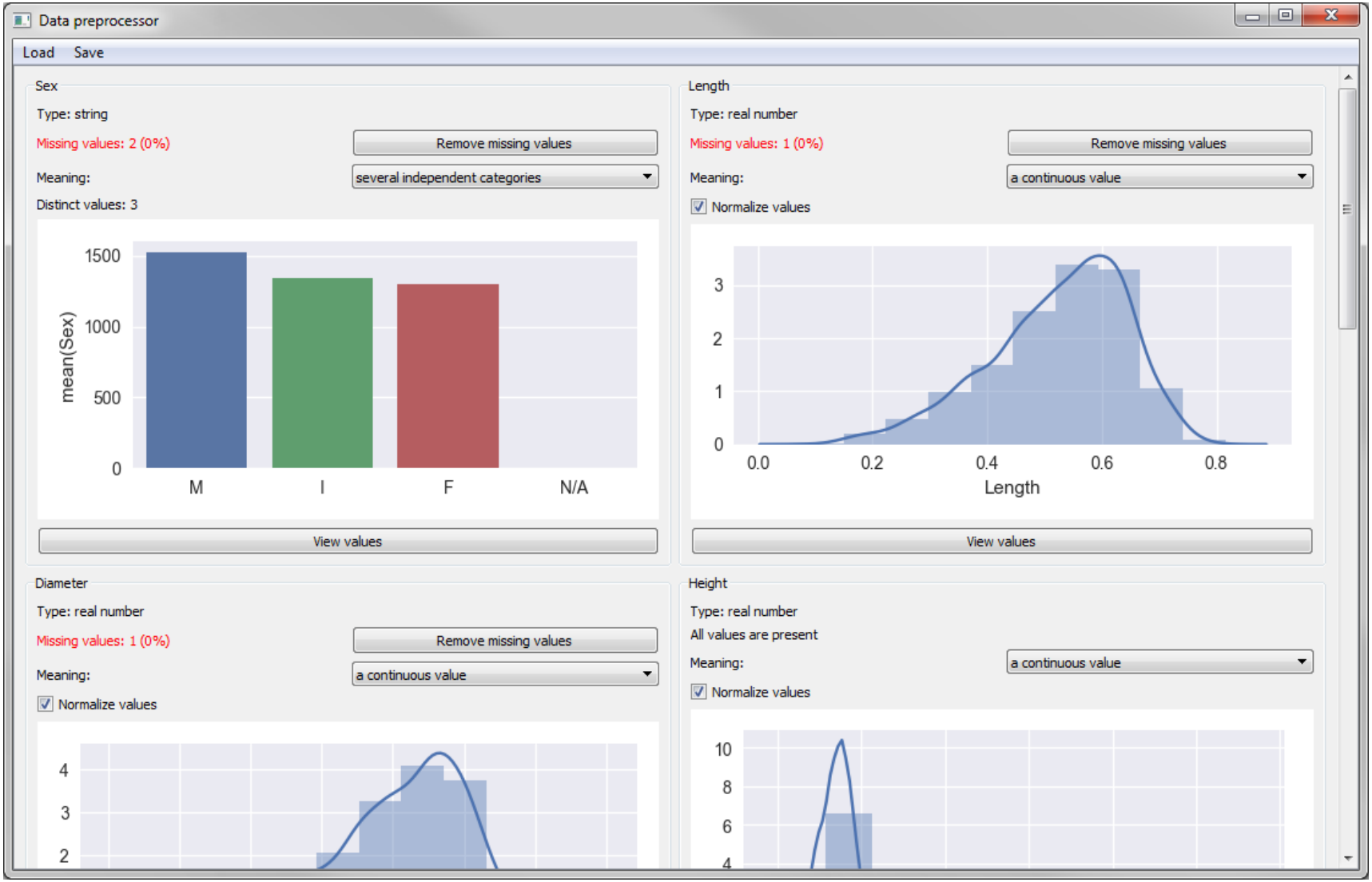
Скриншоты


Values of 'Sex'
Value
Count
1 N/A 4
2 M 1527
3 I 1342
4 F 1304
5 ? 3
Show corresponding values

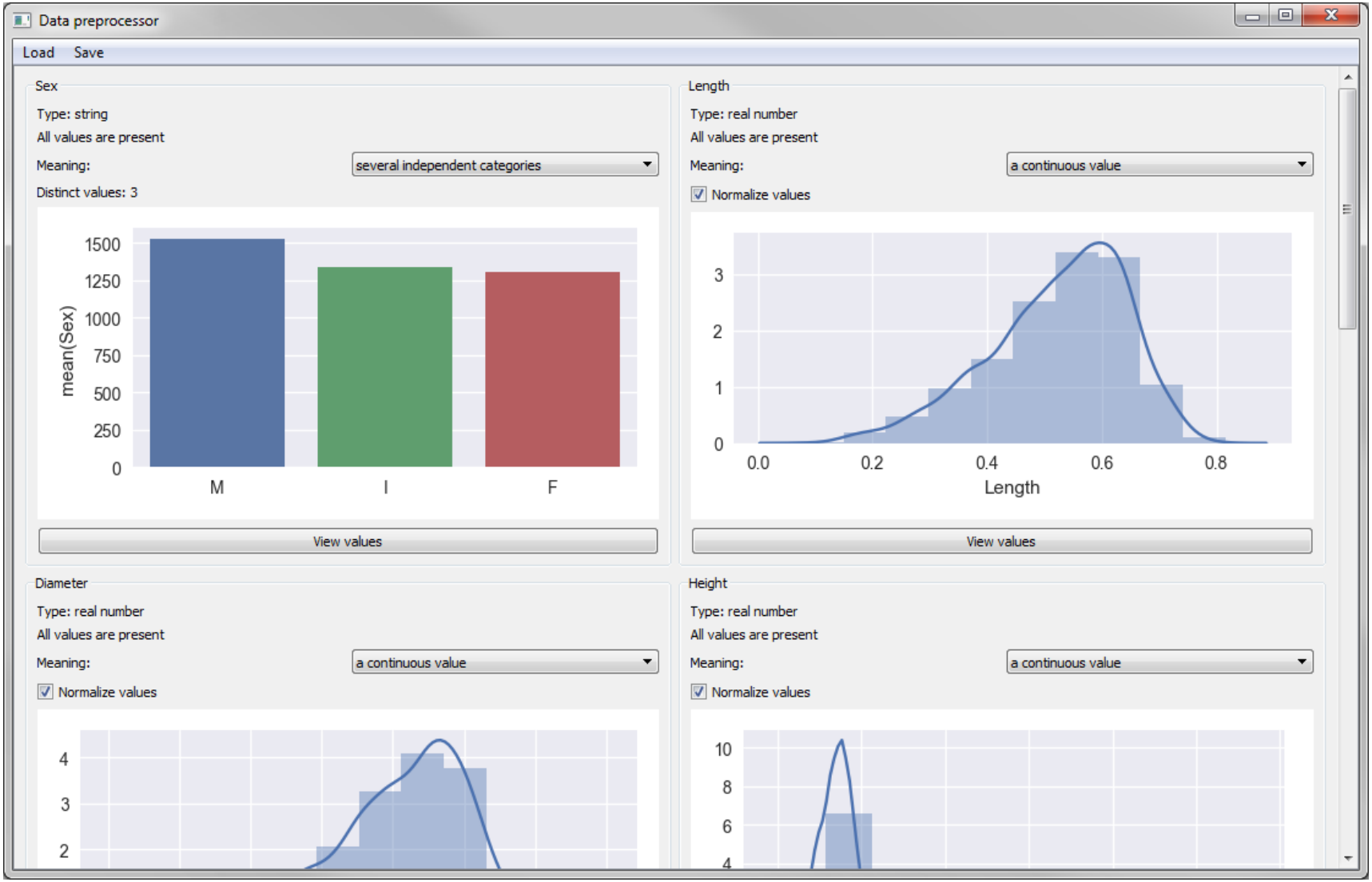
Скриншоты


Rows
Sex
Length
Diameter
Height
Whole weight
Shucked weight
Viscer
1 ? 0.455 0.365 0.095 0.514 0.2245 0.10099
2 ? 0.44 0.34 0.1 0.451 0.188 0.087
3 ? 0.44 0.34 0.1 0.451 0.188 0.087
4 nan 0.53 0.415 0.15 0.7775 0.237 0.1415
5 nan 0.575 0.425 0.14 0.8635 0.392999999999... 0.22699
6 nan nan 0.475 0.155 1.217 0.5305 0.3075
7 nan nan 0.355 0.105 0.5225 0.237 0.1165
Remove rows

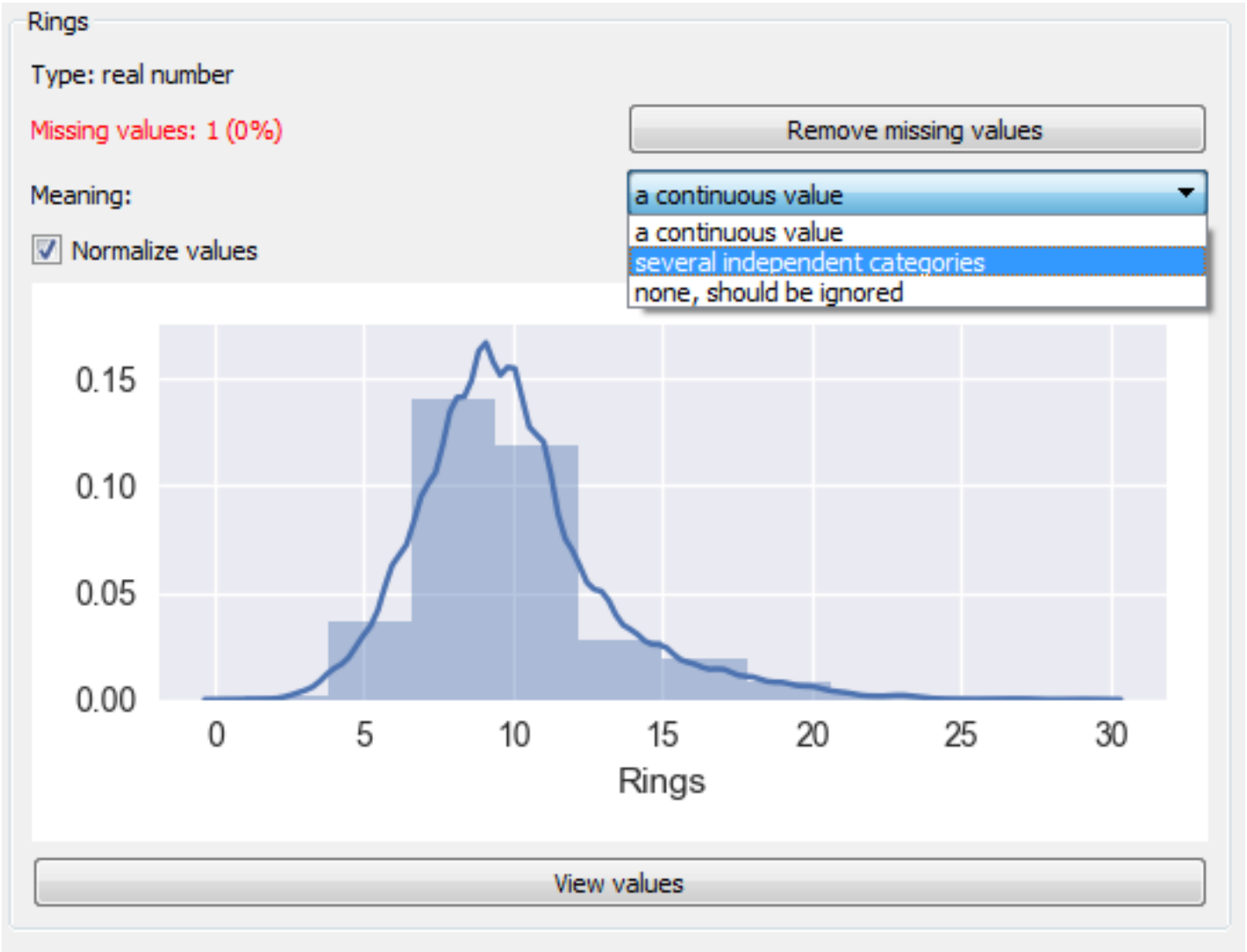
Скриншоты


Data preprocessor
Load
Save
Sex
Type: string
Missing values: 2 (0%)
Remove missing values
Meaning:
several independent categories
Distinct values: 3
1500
1000
500
0
mean(Sex)
M
I
F
N/A
View values
Length
Type: real number
Missing values: 1 (0%)
Remove missing values
Meaning:
a continuous value
Normalize values
3
2
1
0
0.0
0.2
0.4
0.6
0.8
Length
View values
Diameter
Type: real number
Missing values: 1 (0%)
Remove missing values
Meaning:
a continuous value
Normalize values
4
3
2
Height
Type: real number
All values are present
Meaning:
a continuous value
Normalize values
10
8
6
4

Скриншоты


Data preprocessor
Load
Save
Sex
Type: string
All values are present
Meaning:
several independent categories
Distinct values: 3
mean(Sex)
1500
1250
1000
750
500
250
0
M
I
F
View values
Length
Type: real number
All values are present
Meaning:
a continuous value
Normalize values
3
2
1
0
0.0
0.2
0.4
0.6
0.8
Length
View values
Diameter
Type: real number
All values are present
Meaning:
a continuous value
Normalize values
4
3
2
Height
Type: real number
All values are present
Meaning:
a continuous value
Normalize values
10
8
6
4

Скриншоты


Rings
Type: real number
Missing values: 1 (0%)
Remove missing values
Meaning:
a continuous value
a continuous value
several independent categories
none, should be ignored
Normalize values
0.15
0.10
0.05
0.00
0
5
10
15
20
25
30
Rings
View values

Чему научился

- Использование Qt в Python
- Проанализированы существующие системы предобработки данных
- Интеграция PyQt и Seaborn для отрисовки графиков
- Разработка относительно большого проекта на Python

Репозиторий

- github.com/ml-in-programming/ml-tool/
- Ветка data-preprocessing