Pré-processamento de Dados: *pyspark* na prática

MBI - Manufatura avançada
Indústria 4.0

Robespierre Pita

Roteiro

- Introdução
- Análise Descritiva
- Pré-processamento
  - Eliminação manual
  - Integração
  - Amostragem
  - Balanceamento
  - Limpeza
  - Transformação de dados

Roteiro

- **Introdução**
- Análise Descritiva
- Pré-processamento
  - Eliminação manual
  - Integração
  - Amostragem
  - Balanceamento
  - Limpeza
  - Transformação de dados

Introdução

```
In [4]: 1 data = pd.read_csv('credit_aproval.txt')
```

```
In [5]: 1 data
```

Out[5]:

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	class
0	b	30.83	0.000	u	g	w	v	1.250	t	t	1	f	g	202.0	0	+
1	a	58.67	4.460	u	g	q	h	3.040	t	t	6	f	g	43.0	560	+
2	a	24.50	0.500	u	g	q	h	1.500	t	f	0	f	g	280.0	824	+
3	b	27.83	1.540	u	g	w	v	3.750	t	t	5	t	g	100.0	3	+
4	b	20.17	5.625	u	g	w	v	1.710	t	f	0	f	s	120.0	0	+
5	b	32.08	4.000	u	g	m	v	2.500	t	f	0	t	g	360.0	0	+
6	b	33.17	1.040	u	g	r	h	6.500	t	f	0	t	g	164.0	31285	+
7	a	22.92	11.585	u	g	cc	v	0.040	t	f	0	f	g	80.0	1349	+
8	b	54.42	0.500	y	p	k	h	3.960	t	f	0	f	g	180.0	314	+
9	b	42.50	4.915	y	p	w	v	3.165	t	f	0	t	g	52.0	1442	+

Introdução

```
In [4]: 1 data = pd.read_csv('credit_aproval.txt')

In [5]: 1 data
```

Out[5]:

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	class
0	b	30.83	0.000	u	g	w	v	1.250	t	t	1	f	g	202.0	0	+
1	a	58.67	4.460	u	g	q	h	3.040	t	t	6	f	g	43.0	560	+
2	a	24.50	0.500	u	g	q	h	1.500	t	f	0	f	g	280.0	824	+
3	b	27.83	1.540	u	g	w	v	3.750	t	t	5	t	g	100.0	3	+
4	b	20.17	5.625	u	g	w	v	1.710	t	f	0	f	s	120.0	0	+
5	b	32.08	4.000	u	g	m	v	2.500	t	f	0	t	g	360.0	0	+
6	b	33.17	1.040	u	g	r	h	6.500	t	f	0	t	g	164.0	31285	+
7	a	22.92	11.585	u	g	cc	v	0.040	t	f	0	f	g	80.0	1349	+
8	b	54.42	0.500	y	p	k	h	3.960	t	f	0	f	g	180.0	314	+
9	b	42.50	4.915	y	p	w	v	3.165	t	f	0	t	g	52.0	1442	+

Atributos

Introdução

```
In [4]: 1 data = pd.read_csv('credit_aproval.txt')

In [5]: 1 data
```

Out[5]:

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	class
0	b	30.83	0.000	u	g	w	v	1.250	t	t	1	f	g	202.0	0	+
1	a	58.67	4.460	u	g	q	h	3.040	t	t	6	f	g	43.0	560	+
2	a	24.50	0.500	u	g	q	h	1.500	t	f	0	f	g	280.0	824	+
3	b	27.83	1.540	u	g	w	v	3.750	t	t	5	t	g	100.0	3	+
4	b	20.17	5.625	u	g	w	v	1.710	t	f	0	f	s	120.0	0	+
5	b	32.08	4.000	u	g	m	v	2.500	t	f	0	t	g	360.0	0	+
6	b	33.17	1.040	u	g	r	h	6.500	t	f	0	t	g	164.0	31285	+
7	a	22.92	11.585	u	g	cc	v	0.040	t	f	0	f	g	80.0	1349	+
8	b	54.42	0.500	y	p	k	h	3.960	t	f	0	f	g	180.0	314	+
9	b	42.50	4.915	y	p	w	v	3.165	t	f	0	t	g	52.0	1442	+

Atributos

Valor

Introdução

```
In [4]: 1 data = pd.read_csv('credit_aproval.txt')

In [5]: 1 data
```

Out[5]:

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	class
0	b	30.83	0.000	u	g	w	v	1.250	t	t	1	f	g	202.0	0	+
1	a	58.67	4.460	u	g	q	h	3.040	t	t	6	f	g	43.0	560	+
2	a	24.50	0.500	u	g	q	h	1.500	t	f	0	f	g	280.0	824	+
3	b	27.83	1.540	u	g	w	v	3.750	t	t	5	t	g	100.0	3	+
4	b	20.17	5.625	u	g	w	v	1.710	t	f	0	f	s	120.0	0	+
5	b	32.08	4.000	u	g	m	v	2.500	t	f	0	t	g	360.0	0	+
6	b	33.17	1.040	u	g	r	h	6.500	t	f	0	t	g	164.0	31285	+
7	a	22.92	11.585	u	g	cc	v	0.040	t	f	0	f	g	80.0	1349	+
8	b	54.42	0.500	y	p	k	h	3.960	t	f	0	f	g	180.0	314	+
9	b	42.50	4.915	y	p	w	v	3.165	t	f	0	t	g	52.0	1442	+

Atributos

Instâncias

Valor

Introdução

```
In [4]: 1 data = pd.read_csv('credit_aproval.txt')

In [5]: 1 data
```

Out[5]:

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	class
0	b	30.83	0.000	u	g	w	v	1.250	t	t	1	f	g	202.0	0	+
1	a	58.67	4.460	u	g	q	h	3.040	t	t	6	f	g	43.0	560	+
2	a	24.50	0.500	u	g	q	h	1.500	t	f	0	f	g	280.0	824	+
3	b	27.83	1.540	u	g	w	v	3.750	t	t	5	t	g	100.0	3	+
4	b	20.17	5.625	u	g	w	v	1.710	t	f	0	f	s	120.0	0	+
5	b	32.08	4.000	u	g	m	v	2.500	t	f	0	t	g	360.0	0	+
6	b	33.17	1.040	u	g	r	h	6.500	t	f	0	t	g	164.0	31285	+
7	a	22.92	11.585	u	g	cc	v	0.040	t	f	0	f	g	80.0	1349	+
8	b	54.42	0.500	y	p	k	h	3.960	t	f	0	f	g	180.0	314	+
9	b	42.50	4.915	y	p	w	v	3.165	t	f	0	t	g	52.0	1442	+

Atributos

Instâncias

Valor

Variável dependente

Introdução

Credit Approval Data Set

Download: Data Folder, Data Set Description

Abstract: This data concerns credit card applications; good mix of attributes

Data Set Characteristics:	Multivariate	**Number of Instances:**	690	**Area:**	Financial
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	15	**Date Donated**	N/A
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	317922

Source:

(confidential source)

Submitted by quinlan '@' cs.su.oz.au

Data Set Information:

This file concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data.

This dataset is interesting because there is a good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values.

Attribute Information:

A1: b, a.
A2: continuous.
A3: continuous.
A4: u, y, l, t.
A5: g, p, gg.
A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.
A7: v, h, bb, j, n, z, dd, ff, o.
A8: continuous.
A9: t, f.
A10: t, f.
A11: continuous.
A12: t, f.
A13: g, p, s.
A14: continuous.
A15: continuous.
A16: +,- (class attribute)

Introdução

```
In [8]:  1 data_ = spark.read.csv('credit_aproval.txt', header=True)

In [10]: 1 data_.show()
```

```
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
| A1|   A2|    A3| A4| A5| A6| A7|   A8| A9|A10|A11|A12|A13|  A14|  A15|class|
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
|  b|30.83|     0|  u|  g|  w|  v| 1.25|  t|  t| 01|  f|  g|00202|    0|    +|
|  a|58.67|  4.46|  u|  g|  q|  h| 3.04|  t|  t| 06|  f|  g|00043|  560|    +|
|  a|24.50|   0.5|  u|  g|  q|  h|  1.5|  t|  f|  0|  f|  g|00280|  824|    +|
|  b|27.83|  1.54|  u|  g|  w|  v| 3.75|  t|  t| 05|  t|  g|00100|    3|    +|
|  b|20.17| 5.625|  u|  g|  w|  v| 1.71|  t|  f|  0|  f|  s|00120|    0|    +|
|  b|32.08|     4|  u|  g|  m|  v|  2.5|  t|  f|  0|  t|  g|00360|    0|    +|
|  b|33.17|  1.04|  u|  g|  r|  h|  6.5|  t|  f|  0|  t|  g|00164|31285|    +|
|  a|22.92|11.585|  u|  g| cc|  v| 0.04|  t|  f|  0|  f|  g|00080| 1349|    +|
|  b|54.42|   0.5|  y|  p|  k|  h| 3.96|  t|  f|  0|  f|  g|00180|  314|    +|
|  b|42.50| 4.915|  y|  p|  w|  v|3.165|  t|  f|  0|  t|  g|00052| 1442|    +|
|  b|22.08|  0.83|  u|  g|  c|  h|2.165|  f|  f|  0|  t|  g|00128|    0|    +|
|  b|29.92| 1.835|  u|  g|  c|  h|4.335|  t|  f|  0|  f|  g|00260|  200|    +|
|  a|38.25|     6|  u|  g|  k|  v|    1|  t|  f|  0|  t|  g|00000|    0|    +|
|  b|48.08|  6.04|  u|  g|  k|  v| 0.04|  f|  f|  0|  f|  g|00000| 2690|    +|
|  a|45.83|  10.5|  u|  g|  q|  v|    5|  t|  t| 07|  t|  g|00000|    0|    +|
|  b|36.67| 4.415|  y|  p|  k|  v| 0.25|  t|  t| 10|  t|  g|00320|    0|    +|
|  b|28.25| 0.875|  u|  g|  m|  v| 0.96|  t|  t| 03|  t|  g|00396|    0|    +|
|  a|23.25| 5.875|  u|  g|  q|  v| 3.17|  t|  t| 10|  f|  g|00120|  245|    +|
|  b|21.83|  0.25|  u|  g|  d|  h|0.665|  t|  f|  0|  t|  g|00000|    0|    +|
|  a|19.17| 8.585|  u|  g| cc|  h| 0.75|  t|  t| 07|  f|  g|00096|    0|    +|
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
only showing top 20 rows
```

Roteiro

- Introdução
- **Análise Descritiva**
- Pré-processamento
  - Eliminação manual
  - Integração
  - Amostragem
  - Balanceamento
  - Limpeza
  - Transformação de dados

Análise descritiva

Descritiva simples para atributos continuos

In [22]:
```
1 data_.select(['A2']).describe().show()
```

```
+-------+-----------------+
|summary|               A2|
+-------+-----------------+
|  count|              678|
|   mean|31.56817109144546|
| stddev|11.95786249827088|
|    min|            13.75|
|    max|            80.25|
+-------+-----------------+
```

In [21]:
```
1 data_.select(['A2']).summary().show()
```

```
+-------+-----------------+
|summary|               A2|
+-------+-----------------+
|  count|              678|
|   mean|31.56817109144546|
| stddev|11.95786249827088|
|    min|            13.75|
|    25%|            22.58|
|    50%|            28.42|
|    75%|            38.25|
|    max|            80.25|
+-------+-----------------+
```

Análise descritiva

Descritiva simples para atributos continuos

In [22]:
```
1 data_.select(['A2']).describe().show()
```

```
+-------+-----------------+
|summary|               A2|
+-------+-----------------+
|  count|              678|
|   mean|31.56817109144546|
| stddev|11.95786249827088|
|    min|            13.75|
|    max|            80.25|
+-------+-----------------+
```

In [21]:
```
1 data_.select(['A2']).summary().show()
```

```
+-------+-----------------+
|summary|               A2|
+-------+-----------------+
|  count|              678|
|   mean|31.56817109144546|
| stddev|11.95786249827088|
|    min|            13.75|
|    25%|            22.58|
|    50%|            28.42|
|    75%|            38.25|
|    max|            80.25|
+-------+-----------------+
```

In [45]:
```
1 # Variancia
2 data_.agg(F.variance('A2')).show()
```

```
+-----------------+
|    var_samp(A2)|
+-----------------+
|142.9904755275531|
+-----------------+
```

Análise descritiva

Descritiva simples para atributos continuos

In [22]:
```
1 data_.select(['A2']).describe().show()
```

```
+-------+-----------------+
|summary|               A2|
+-------+-----------------+
|  count|              678|
|   mean|31.56817109144546|
| stddev|11.95786249827088|
|    min|            13.75|
|    max|            80.25|
+-------+-----------------+
```

In [21]:
```
1 data_.select(['A2']).summary().show()
```

```
+-------+-----------------+
|summary|               A2|
+-------+-----------------+
|  count|              678|
|   mean|31.56817109144546|
| stddev|11.95786249827088|
|    min|            13.75|
|    25%|            22.58|
|    50%|            28.42|
|    75%|            38.25|
|    max|            80.25|
+-------+-----------------+
```

In [45]:
```
1 # Variancia
2 data_.agg(F.variance('A2')).show()
```

```
+-----------------+
|     var_samp(A2)|
+-----------------+
|142.9904755275531|
+-----------------+
```

Duplicados

In [154]:
```
1 # Todos os registros do dataset
2 print data_.count()
3 # Todos os registros únicos do dataset
4 print data_.distinct().count()
5 # Todos os registros únicos do dataset,
6 # levando em consideração os atributos: 'A1', 'A2' e 'A5'
7 print data_.select(['A1', 'A2', 'A5']).distinct().count()
```

```
690
690
529
```

Análise descritiva

Descritiva simples para atributos continuos

In [22]:
```
1 data_.select(['A2']).describe().show()
```

```
+-------+-----------------+
|summary|               A2|
+-------+-----------------+
|  count|              678|
|   mean|31.56817109144546|
| stddev|11.95786249827088|
|    min|            13.75|
|    max|            80.25|
+-------+-----------------+
```

In [21]:
```
1 data_.select(['A2']).summary().show()
```

```
+-------+-----------------+
|summary|               A2|
+-------+-----------------+
|  count|              678|
|   mean|31.56817109144546|
| stddev|11.95786249827088|
|    min|            13.75|
|    25%|            22.58|
|    50%|            28.42|
|    75%|            38.25|
|    max|            80.25|
+-------+-----------------+
```

In [45]:
```
1 # Variancia
2 data_.agg(F.variance('A2')).show()
```

```
+-----------------+
|    var_samp(A2)|
+-----------------+
|142.9904755275531|
+-----------------+
```

Duplicados

In [154]:
```
1 # Todos os registros do dataset
2 print data_.count()
3 # Todos os registros únicos do dataset
4 print data_.distinct().count()
5 # Todos os registros únicos do dataset,
6 # levando em consideração os atributos: 'A1', 'A2' e 'A5'
7 print data_.select(['A1', 'A2', 'A5']).distinct().count()
```

```
690
690
529
```

Descritiva simples para atributos categóricos

In [18]:
```
1 data_.groupby('A1').count().show()
```

```
+----+-----+
|  A1|count|
+----+-----+
|null|   12|
|   b|  468|
|   a|  210|
+----+-----+
```

Análise descritiva

Analise de nulos

```
In [51]: 1 data_.filter(data_['A2'].isNull()).count()
Out[51]: 12

In [52]: 1 data_.filter(data_['A2'].isNotNull()).count()
Out[52]: 678
```

Filtros

```
In [56]: 1 data_.filter(data_['A2'] > 11).limit(5).show()
```

```
+---+-----+-----+---+---+---+---+----+---+---+---+---+---+-----+---+-----+
| A1|   A2|   A3| A4| A5| A6| A7|  A8| A9|A10|A11|A12|A13|  A14|A15|class|
+---+-----+-----+---+---+---+---+----+---+---+---+---+---+-----+---+-----+
|  b|30.83|    0|  u|  g|  w|  v|1.25|  t|  t| 01|  f|  g|00202|  0|    +|
|  a|58.67| 4.46|  u|  g|  q|  h|3.04|  t|  t| 06|  f|  g|00043|560|    +|
|  a|24.50|  0.5|  u|  g|  q|  h| 1.5|  t|  f|  0|  f|  g|00280|824|    +|
|  b|27.83| 1.54|  u|  g|  w|  v|3.75|  t|  t| 05|  t|  g|00100|  3|    +|
|  b|20.17|5.625|  u|  g|  w|  v|1.71|  t|  f|  0|  f|  s|00120|  0|    +|
+---+-----+-----+---+---+---+---+----+---+---+---+---+---+-----+---+-----+
```

```
In [57]: 1 data_.select(['A1', 'A2', 'A3', 'class']).filter(data_['A2'] > 11).limit(5).show()
```

```
+---+-----+-----+-----+
| A1|   A2|   A3|class|
+---+-----+-----+-----+
|  b|30.83|    0|    +|
|  a|58.67| 4.46|    +|
|  a|24.50|  0.5|    +|
|  b|27.83| 1.54|    +|
|  b|20.17|5.625|    +|
+---+-----+-----+-----+
```

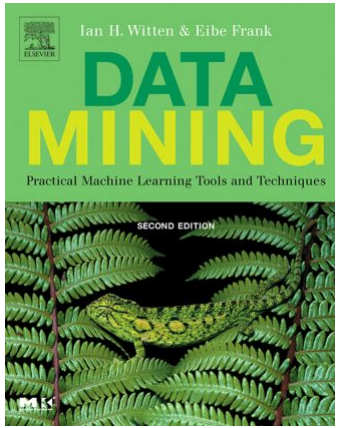
Roteiro

- Introdução
- Análise Descritiva
- **Pré-processamento**
  - Eliminação manual
  - Integração
  - Amostragem
  - Balanceamento
  - Limpeza
  - Transformação de dados

Pré-processamento

Preprocessing, as well as data cleaning, are time-consuming and labor-intensive procedures, but that is absolutely necessary for successful data mining. With a large dataset, people often give up - how can they possibly check it all? Instead, you should sample a few instances and examine them carefully. You'll be surprised at what you find. Time looking at your data is always well spent

(Witten et. Al, 2011).



Roteiro

- Introdução
- Análise Descritiva
- Pré-processamento
  - **Eliminação manual**
  - Integração
  - Amostragem
  - Balanceamento
  - Limpeza
  - Transformação de dados

Pré-processamento: Eliminação manual

Eliminação manual

In [106]:

```
data_temp = data_.select(['A1', 'A2', 'A3', 'A4', 'A5','class'])\
            .filter((data_['A5']=='g') | (data_['A1']\
            .isNull()))
print "Tamanho do extrato: {}".format(data_temp.count())
data_temp.show()
```

```
Tamanho do extrato: 523
+---+-----+------+---+---+-----+
| A1|   A2|    A3| A4| A5|class|
+---+-----+------+---+---+-----+
|  b|30.83|     0|  u|  g|    +|
|  a|58.67|  4.46|  u|  g|    +|
|  a|24.50|   0.5|  u|  g|    +|
|  b|27.83|  1.54|  u|  g|    +|
|  b|20.17| 5.625|  u|  g|    +|
|  b|32.08|     4|  u|  g|    +|
|  b|33.17|  1.04|  u|  g|    +|
|  a|22.92|11.585|  u|  g|    +|
|  b|22.08|  0.83|  u|  g|    +|
|  b|29.92| 1.835|  u|  g|    +|
|  a|38.25|     6|  u|  g|    +|
|  b|48.08|  6.04|  u|  g|    +|
|  a|45.83|  10.5|  u|  g|    +|
|  b|28.25| 0.875|  u|  g|    +|
|  a|23.25| 5.875|  u|  g|    +|
|  b|21.83|  0.25|  u|  g|    +|
|  a|19.17| 8.585|  u|  g|    +|
|  b|25.00| 11.25|  u|  g|    +|
|  b|23.25|     1|  u|  g|    +|
|  a|47.75|     8|  u|  g|    +|
+---+-----+------+---+---+-----+
only showing top 20 rows
```

Pré-processamento: Eliminação manual

Eliminação manual

In [106]:

```
data_temp = data_.select(['A1', 'A2', 'A3', 'A4', 'A5','class'])\
            .filter((data_['A5']=='g') | (data_['A1']\
            .isNull()))
print "Tamanho do extrato: {}".format(data_temp.count())
data_temp.show()
```

```
Tamanho do extrato: 523
+---+-----+------+---+---+-----+
| A1|   A2|    A3| A4| A5|class|
+---+-----+------+---+---+-----+
|  b|30.83|     0|  u|  g|    +|
|  a|58.67|  4.46|  u|  g|    +|
|  a|24.50|   0.5|  u|  g|    +|
|  b|27.83|  1.54|  u|  g|    +|
|  b|20.17| 5.625|  u|  g|    +|
|  b|32.08|     4|  u|  g|    +|
|  b|33.17|  1.04|  u|  g|    +|
|  a|22.92|11.585|  u|  g|    +|
|  b|22.08|  0.83|  u|  g|    +|
|  b|29.92| 1.835|  u|  g|    +|
|  a|38.25|     6|  u|  g|    +|
|  b|48.08|  6.04|  u|  g|    +|
|  a|45.83|  10.5|  u|  g|    +|
|  b|28.25| 0.875|  u|  g|    +|
|  a|23.25| 5.875|  u|  g|    +|
|  b|21.83|  0.25|  u|  g|    +|
|  a|19.17| 8.585|  u|  g|    +|
|  b|25.00| 11.25|  u|  g|    +|
|  b|23.25|     1|  u|  g|    +|
|  a|47.75|     8|  u|  g|    +|
+---+-----+------+---+---+-----+
only showing top 20 rows
```

In [110]:

```
print "Quantidade de NAs no extrato: {}"\
    .format(data_temp.filter(data_temp['A1']\
    .isNull()).count())

data_temp.filter(data_temp['A1']\
        .isNull()).show()
```

```
Quantidade de NAs no extrato: 12
+----+-----+-----+---+---+-----+
|  A1|   A2|   A3| A4| A5|class|
+----+-----+-----+---+---+-----+
|null|24.50|12.75|  u|  g|    +|
|null|40.83|  3.5|  u|  g|    -|
|null|32.25|  1.5|  u|  g|    -|
|null|28.17|0.585|  u|  g|    -|
|null|29.75|0.665|  u|  g|    -|
|null|26.50| 2.71|  y|  p|    -|
|null|45.33|    1|  u|  g|    -|
|null|20.42|  7.5|  u|  g|    +|
|null|20.08|0.125|  u|  g|    +|
|null|42.25| 1.75|  y|  p|    -|
|null|33.17| 2.25|  y|  p|    -|
|null|29.50|    2|  y|  p|    -|
+----+-----+-----+---+---+-----+
```

Pré-processamento: Eliminação manual

Eliminação manual

In [106]:
```
1 data_temp = data_.select(['A1', 'A2', 'A3', 'A4', 'A5','class'])\
2             .filter((data_['A5']=='g') | (data_['A1']\
3             .isNull()))
4 print "Tamanho do extrato: {}".format(data_temp.count())
5 data_temp.show()
```

```
Tamanho do extrato: 523
+---+-----+------+---+---+-----+
| A1|   A2|    A3| A4| A5|class|
+---+-----+------+---+---+-----+
|  b|30.83|     0|  u|  g|    +|
|  a|58.67|  4.46|  u|  g|    +|
|  a|24.50|   0.5|  u|  g|    +|
|  b|27.83|  1.54|  u|  g|    +|
|  b|20.17| 5.625|  u|  g|    +|
|  b|32.08|     4|  u|  g|    +|
|  b|33.17|  1.04|  u|  g|    +|
|  a|22.92|11.585|  u|  g|    +|
|  b|22.08|  0.83|  u|  g|    +|
|  b|29.92| 1.835|  u|  g|    +|
|  a|38.25|     6|  u|  g|    +|
|  b|48.08|  6.04|  u|  g|    +|
|  a|45.83|  10.5|  u|  g|    +|
|  b|28.25| 0.875|  u|  g|    +|
|  a|23.25| 5.875|  u|  g|    +|
|  b|21.83|  0.25|  u|  g|    +|
|  a|19.17| 8.585|  u|  g|    +|
|  b|25.00| 11.25|  u|  g|    +|
|  b|23.25|     1|  u|  g|    +|
|  a|47.75|     8|  u|  g|    +|
+---+-----+------+---+---+-----+
only showing top 20 rows
```

In [110]:
```
1 print "Quantidade de NAs no extrato: {}"\
2     .format(data_temp.filter(data_temp['A1']\
3     .isNull()).count())
4
5 data_temp.filter(data_temp['A1']\
6         .isNull()).show()
```

```
Quantidade de NAs no extrato: 12
+----+-----+-----+---+---+-----+
|  A1|   A2|   A3| A4| A5|class|
+----+-----+-----+---+---+-----+
|null|24.50|12.75|  u|  g|    +|
|null|40.83|  3.5|  u|  g|    -|
|null|32.25|  1.5|  u|  g|    -|
|null|28.17|0.585|  u|  g|    -|
|null|29.75|0.665|  u|  g|    -|
|null|26.50| 2.71|  y|  p|    -|
|null|45.33|    1|  u|  g|    -|
|null|20.42|  7.5|  u|  g|    +|
|null|20.08|0.125|  u|  g|    +|
|null|42.25| 1.75|  y|  p|    -|
|null|33.17| 2.25|  y|  p|    -|
|null|29.50|    2|  y|  p|    -|
+----+-----+-----+---+---+-----+
```

In [111]:
```
1 data_temp.dropna('any').count()
```

Out[111]: 503

Pré-processamento: Eliminação manual

Eliminação manual: excluindo colunas

In [113]:
```
1 data_temp.show()
```

```
+---+-----+------+---+---+-----+
| A1|   A2|    A3| A4| A5|class|
+---+-----+------+---+---+-----+
|  b|30.83|     0|  u|  g|    +|
|  a|58.67|  4.46|  u|  g|    +|
|  a|24.50|   0.5|  u|  g|    +|
|  b|27.83|  1.54|  u|  g|    +|
|  b|20.17| 5.625|  u|  g|    +|
|  b|32.08|     4|  u|  g|    +|
|  b|33.17|  1.04|  u|  g|    +|
|  a|22.92|11.585|  u|  g|    +|
|  b|22.08|  0.83|  u|  g|    +|
|  b|29.92| 1.835|  u|  g|    +|
|  a|38.25|     6|  u|  g|    +|
|  b|48.08|  6.04|  u|  g|    +|
|  a|45.83|  10.5|  u|  g|    +|
|  b|28.25| 0.875|  u|  g|    +|
|  a|23.25| 5.875|  u|  g|    +|
|  b|21.83|  0.25|  u|  g|    +|
|  a|19.17| 8.585|  u|  g|    +|
|  b|25.00| 11.25|  u|  g|    +|
|  b|23.25|     1|  u|  g|    +|
|  a|47.75|     8|  u|  g|    +|
+---+-----+------+---+---+-----+
only showing top 20 rows
```

Pré-processamento: Eliminação manual

Eliminação manual: excluindo colunas

In [113]:
```
1 data_temp.show()
```

```
+---+-----+------+---+---+-----+
| A1|   A2|    A3| A4| A5|class|
+---+-----+------+---+---+-----+
|  b|30.83|     0|  u|  g|    +|
|  a|58.67|  4.46|  u|  g|    +|
|  a|24.50|   0.5|  u|  g|    +|
|  b|27.83|  1.54|  u|  g|    +|
|  b|20.17| 5.625|  u|  g|    +|
|  b|32.08|     4|  u|  g|    +|
|  b|33.17|  1.04|  u|  g|    +|
|  a|22.92|11.585|  u|  g|    +|
|  b|22.08|  0.83|  u|  g|    +|
|  b|29.92| 1.835|  u|  g|    +|
|  a|38.25|     6|  u|  g|    +|
|  b|48.08|  6.04|  u|  g|    +|
|  a|45.83|  10.5|  u|  g|    +|
|  b|28.25| 0.875|  u|  g|    +|
|  a|23.25| 5.875|  u|  g|    +|
|  b|21.83|  0.25|  u|  g|    +|
|  a|19.17| 8.585|  u|  g|    +|
|  b|25.00| 11.25|  u|  g|    +|
|  b|23.25|     1|  u|  g|    +|
|  a|47.75|     8|  u|  g|    +|
+---+-----+------+---+---+-----+
only showing top 20 rows
```

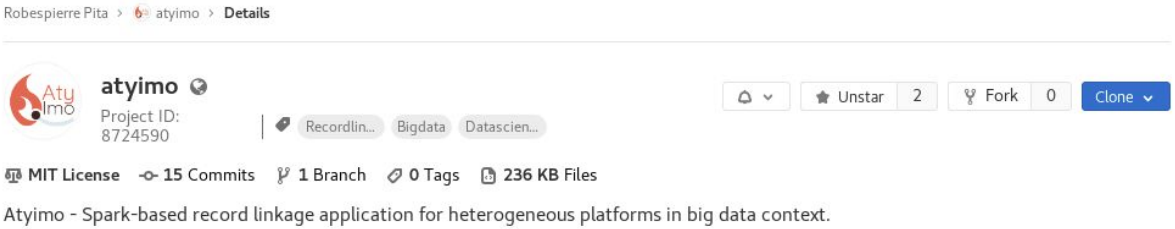
In [119]:
```
1 # Retirando as variaveis categóricas
2 data_temp.drop(F.col('A1'))\
3         .drop(F.col('A4'))\
4         .drop(F.col('A5')).show()
```

```
+-----+------+-----+
|   A2|    A3|class|
+-----+------+-----+
|30.83|     0|    +|
|58.67|  4.46|    +|
|24.50|   0.5|    +|
|27.83|  1.54|    +|
|20.17| 5.625|    +|
|32.08|     4|    +|
|33.17|  1.04|    +|
|22.92|11.585|    +|
|22.08|  0.83|    +|
|29.92| 1.835|    +|
|38.25|     6|    +|
|48.08|  6.04|    +|
|45.83|  10.5|    +|
|28.25| 0.875|    +|
|23.25| 5.875|    +|
|21.83|  0.25|    +|
|19.17| 8.585|    +|
|25.00| 11.25|    +|
|23.25|     1|    +|
|47.75|     8|    +|
+-----+------+-----+
only showing top 20 rows
```

Roteiro

- Introdução
- Análise Descritiva
- Pré-processamento
  - Eliminação manual
  - **Integração**
  - Amostragem
  - Balanceamento
  - Limpeza
  - Transformação de dados

Pré-processamento: Integração



https://gitlab.com/pierrepita/atyimo

Roteiro

- Introdução
- Análise Descritiva
- Pré-processamento
  - Eliminação manual
  - Integração
  - **Amostragem**
  - Balanceamento
  - Limpeza
  - Transformação de dados

Pré-processamento: Amostragem

Sampling: Test and Training

In [132]:
```
train = data_.sample(withReplacement=False, fraction=0.8, seed=27)
```

In [133]:
```
print train.count()
train.limit(10).show()
```

```
567
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+----+-----+
| A1|   A2|    A3| A4| A5| A6| A7|   A8| A9|A10|A11|A12|A13|  A14| A15|class|
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+----+-----+
|  b|30.83|     0|  u|  g|  w|  v| 1.25|  t|  t| 01|  f|  g|00202|   0|    +|
|  a|24.50|   0.5|  u|  g|  q|  h|  1.5|  t|  f|  0|  f|  g|00280| 824|    +|
|  b|27.83|  1.54|  u|  g|  w|  v| 3.75|  t|  t| 05|  t|  g|00100|   3|    +|
|  b|32.08|     4|  u|  g|  m|  v|  2.5|  t|  f|  0|  t|  g|00360|   0|    +|
|  a|22.92|11.585|  u|  g| cc|  v| 0.04|  t|  f|  0|  f|  g|00080|1349|    +|
|  b|54.42|   0.5|  y|  p|  k|  h| 3.96|  t|  f|  0|  f|  g|00180| 314|    +|
|  b|42.50| 4.915|  y|  p|  w|  v|3.165|  t|  f|  0|  t|  g|00052|1442|    +|
|  b|22.08|  0.83|  u|  g|  c|  h|2.165|  f|  f|  0|  t|  g|00128|   0|    +|
|  a|38.25|     6|  u|  g|  k|  v|    1|  t|  f|  0|  t|  g|00000|   0|    +|
|  b|48.08|  6.04|  u|  g|  k|  v| 0.04|  f|  f|  0|  f|  g|00000|2690|    +|
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+----+-----+
```

In [134]:
```
test = data_.subtract(train)
```

In [135]:
```
print test.count()
test.limit(10).show()
```

```
123
+---+-----+-----+---+---+---+---+-----+---+---+---+---+---+-----+----+-----+
| A1|   A2|   A3| A4| A5| A6| A7|   A8| A9|A10|A11|A12|A13|  A14| A15|class|
+---+-----+-----+---+---+---+---+-----+---+---+---+---+---+-----+----+-----+
|  b|39.50|1.625|  u|  g|  c|  v|  1.5|  f|  f|  0|  f|  g|00000| 316|    -|
|  b|28.25|5.125|  u|  g|  x|  v| 4.75|  t|  t| 02|  f|  g|00420|   7|    +|
|  a|27.25| 0.29|  u|  g|  m|  h|0.125|  f|  t| 01|  t|  g|00272| 108|    -|
|  b|23.08|    0|  u|  g|  k|  v|    1|  f|  t| 11|  f|  s|00000|   0|    -|
|  b|29.92|1.835|  u|  g|  c|  h|4.335|  t|  f|  0|  f|  g|00260| 200|    +|
|  b|34.08| 0.08|  y|  p|  m| bb| 0.04|  t|  t| 01|  t|  g|00280|2000|    +|
|  b|34.92|    5|  u|  g|  x|  h|  7.5|  t|  t| 06|  t|  g|00000|1000|    +|
|  b|19.17|  9.5|  u|  g|  w|  v|  1.5|  t|  f|  0|  f|  g|00120|2206|    +|
|  a|27.67|  1.5|  u|  g|  m|  v|    2|  t|  f|  0|  f|  s|00368|   0|    -|
|  b|33.17|3.165|  y|  p|  x|  v|3.165|  t|  t| 03|  t|  g|00380|   0|    +|
+---+-----+-----+---+---+---+---+-----+---+---+---+---+---+-----+----+-----+
```

Roteiro

- Introdução
- Análise Descritiva
- Pré-processamento
  - Eliminação manual
  - Integração
  - Amostragem
  - **Balanceamento**
  - Limpeza
  - Transformação de dados

Pré-processamento: Amostragem

In [136]:
```
1 data_.groupby('class').count().show()
```

```
+-----+-----+
|class|count|
+-----+-----+
|    +|  307|
|    -|  383|
+-----+-----+
```

Pré-processamento: Amostragem

In [136]:
```
data_.groupby('class').count().show()
```

```
+-----+-----+
|class|count|
+-----+-----+
|    +|  307|
|    -|  383|
+-----+-----+
```

In [144]:
```
data_positive = data_.filter(data_['class'] == '+')\
                .limit(300)

data_negative = data_.filter(data_['class'] == '-')\
                .limit(300)

balanced_data = data_positive.union(data_negative)
```

Pré-processamento: Amostragem

In [136]:
```
1 data_.groupby('class').count().show()
```

```
+-----+-----+
|class|count|
+-----+-----+
|    +|  307|
|    -|  383|
+-----+-----+
```

In [144]:
```
1 data_positive = data_.filter(data_['class'] == '+')\
2                  .limit(300)
3
4 data_negative = data_.filter(data_['class'] == '-')\
5                  .limit(300)
6
7 balanced_data = data_positive.union(data_negative)
```

In [146]:
```
1 balanced_data.groupby('class').count().show()
```

```
+-----+-----+
|class|count|
+-----+-----+
|    +|  300|
|    -|  300|
+-----+-----+
```

Roteiro

- Introdução
- Análise Descritiva
- Pré-processamento
  - Eliminação manual
  - Integração
  - Amostragem
  - Balanceamento
  - **Limpeza**
  - Transformação de dados

Pré-processamento: Limpeza

Data Cleaning

In [148]:

```
# Deduplicação
print data_.count()

print data_.drop_duplicates(subset=['A1', 'A2', 'A5']).count()
```

```
690
529
```

Roteiro

- Introdução
- Análise Descritiva
- Pré-processamento
  - Eliminação manual
  - Integração
  - Amostragem
  - Balanceamento
  - Limpeza
  - **Transformação de dados**

Pré-processamento: Transformação de dados

Transformação de dados

In [163]:
```
data_.limit(10).show()
```

```
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
| A1|   A2|    A3| A4| A5| A6| A7|   A8| A9|A10|A11|A12|A13|  A14|  A15|class|
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
|  b|30.83|     0|  u|  g|  w|  v| 1.25|  t|  t| 01|  f|  g|00202|    0|    +|
|  a|58.67|  4.46|  u|  g|  q|  h| 3.04|  t|  t| 06|  f|  g|00043|  560|    +|
|  a|24.50|   0.5|  u|  g|  q|  h|  1.5|  t|  f|  0|  f|  g|00280|  824|    +|
|  b|27.83|  1.54|  u|  g|  w|  v| 3.75|  t|  t| 05|  t|  g|00100|    3|    +|
|  b|20.17| 5.625|  u|  g|  w|  v| 1.71|  t|  f|  0|  f|  s|00120|    0|    +|
|  b|32.08|     4|  u|  g|  m|  v|  2.5|  t|  f|  0|  t|  g|00360|    0|    +|
|  b|33.17|  1.04|  u|  g|  r|  h|  6.5|  t|  f|  0|  t|  g|00164|31285|    +|
|  a|22.92|11.585|  u|  g| cc|  v| 0.04|  t|  f|  0|  f|  g|00080| 1349|    +|
|  b|54.42|   0.5|  y|  p|  k|  h| 3.96|  t|  f|  0|  f|  g|00180|  314|    +|
|  b|42.50| 4.915|  y|  p|  w|  v|3.165|  t|  f|  0|  t|  g|00052| 1442|    +|
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
```

In [164]:
```
# Renomeando categorias
data_.withColumn('A1', F.when(data_['A1'] == 'a', 'classe_a').otherwise('class_b'))\
      .limit(10)\
      .show()
```

```
+--------+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
|      A1|   A2|    A3| A4| A5| A6| A7|   A8| A9|A10|A11|A12|A13|  A14|  A15|class|
+--------+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
| class_b|30.83|     0|  u|  g|  w|  v| 1.25|  t|  t| 01|  f|  g|00202|    0|    +|
|classe_a|58.67|  4.46|  u|  g|  q|  h| 3.04|  t|  t| 06|  f|  g|00043|  560|    +|
|classe_a|24.50|   0.5|  u|  g|  q|  h|  1.5|  t|  f|  0|  f|  g|00280|  824|    +|
| class_b|27.83|  1.54|  u|  g|  w|  v| 3.75|  t|  t| 05|  t|  g|00100|    3|    +|
| class_b|20.17| 5.625|  u|  g|  w|  v| 1.71|  t|  f|  0|  f|  s|00120|    0|    +|
| class_b|32.08|     4|  u|  g|  m|  v|  2.5|  t|  f|  0|  t|  g|00360|    0|    +|
| class_b|33.17|  1.04|  u|  g|  r|  h|  6.5|  t|  f|  0|  t|  g|00164|31285|    +|
|classe_a|22.92|11.585|  u|  g| cc|  v| 0.04|  t|  f|  0|  f|  g|00080| 1349|    +|
| class_b|54.42|   0.5|  y|  p|  k|  h| 3.96|  t|  f|  0|  f|  g|00180|  314|    +|
| class_b|42.50| 4.915|  y|  p|  w|  v|3.165|  t|  f|  0|  t|  g|00052| 1442|    +|
+--------+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
```

In [176]:
```
# Vamos supor que 'A3' significa 'o numero de salarios'
data_.select(['A3']).summary().show()
# Categorizando variaveis
data_.withColumn('A3', F.when(F.col('A3') <= 2.75, 'pobre').otherwise('rico')).show()
```

```
+-------+------------------+
|summary|                A3|
+-------+------------------+
|  count|               690|
|   mean|4.7587246376811585|
| stddev| 4.978163248528541|
|    min|                 0|
|    25%|               1.0|
|    50%|              2.75|
|    75%|              7.25|
|    max|              9.96|
+-------+------------------+

+---+-----+-----+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
| A1|   A2|   A3| A4| A5| A6| A7|   A8| A9|A10|A11|A12|A13|  A14|  A15|class|
+---+-----+-----+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
|  b|30.83|pobre|  u|  g|  w|  v| 1.25|  t|  t| 01|  f|  g|00202|    0|    +|
|  a|58.67| rico|  u|  g|  q|  h| 3.04|  t|  t| 06|  f|  g|00043|  560|    +|
|  a|24.50|pobre|  u|  g|  q|  h|  1.5|  t|  f|  0|  f|  g|00280|  824|    +|
|  b|27.83|pobre|  u|  g|  w|  v| 3.75|  t|  t| 05|  t|  g|00100|    3|    +|
|  b|20.17| rico|  u|  g|  w|  v| 1.71|  t|  f|  0|  f|  s|00120|    0|    +|
|  b|32.08| rico|  u|  g|  m|  v|  2.5|  t|  f|  0|  t|  g|00360|    0|    +|
|  b|33.17|pobre|  u|  g|  r|  h|  6.5|  t|  f|  0|  t|  g|00164|31285|    +|
|  a|22.92| rico|  u|  g| cc|  v| 0.04|  t|  f|  0|  f|  g|00080| 1349|    +|
|  b|54.42|pobre|  y|  p|  k|  h| 3.96|  t|  f|  0|  f|  g|00180|  314|    +|
|  b|42.50| rico|  y|  p|  w|  v|3.165|  t|  f|  0|  t|  g|00052| 1442|    +|
|  b|22.08|pobre|  u|  g|  c|  h|2.165|  f|  f|  0|  t|  g|00128|    0|    +|
|  b|29.92|pobre|  u|  g|  c|  h|4.335|  t|  f|  0|  f|  g|00260|  200|    +|
|  a|38.25| rico|  u|  g|  k|  v|    1|  t|  f|  0|  t|  g|00000|    0|    +|
|  b|48.08| rico|  u|  g|  k|  v| 0.04|  f|  f|  0|  f|  g|00000| 2690|    +|
|  a|45.83| rico|  u|  g|  q|  v|    5|  t|  t| 07|  t|  g|00000|    0|    +|
|  b|36.67| rico|  y|  p|  k|  v| 0.25|  t|  t| 10|  t|  g|00320|    0|    +|
|  b|28.25|pobre|  u|  g|  m|  v| 0.96|  t|  t| 03|  t|  g|00396|    0|    +|
|  a|23.25| rico|  u|  g|  q|  v| 3.17|  t|  t| 10|  f|  g|00120|  245|    +|
|  b|21.83|pobre|  u|  g|  d|  h|0.665|  t|  f|  0|  t|  g|00000|    0|    +|
|  a|19.17| rico|  u|  g| cc|  h| 0.75|  t|  t| 07|  f|  g|00096|    0|    +|
+---+-----+-----+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
only showing top 20 rows
```

Pré-processamento: Transformação de dados

Transformação de dados

In [163]:
```
1 data_.limit(10).show()
```

```
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
| A1|   A2|    A3| A4| A5| A6| A7|   A8| A9|A10|A11|A12|A13|  A14|  A15|class|
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
|  b|30.83|     0|  u|  g|  w|  v| 1.25|  t|  t| 01|  f|  g|00202|    0|    +|
|  a|58.67|  4.46|  u|  g|  q|  h| 3.04|  t|  t| 06|  f|  g|00043|  560|    +|
|  a|24.50|   0.5|  u|  g|  q|  h|  1.5|  t|  f|  0|  f|  g|00280|  824|    +|
|  b|27.83|  1.54|  u|  g|  w|  v| 3.75|  t|  t| 05|  t|  g|00100|    3|    +|
|  b|20.17| 5.625|  u|  g|  w|  v| 1.71|  t|  f|  0|  f|  s|00120|    0|    +|
|  b|32.08|     4|  u|  g|  m|  v|  2.5|  t|  f|  0|  t|  g|00360|    0|    +|
|  b|33.17|  1.04|  u|  g|  r|  h|  6.5|  t|  f|  0|  t|  g|00164|31285|    +|
|  a|22.92|11.585|  u|  g| cc|  v| 0.04|  t|  f|  0|  f|  g|00080| 1349|    +|
|  b|54.42|   0.5|  y|  p|  k|  h| 3.96|  t|  f|  0|  f|  g|00180|  314|    +|
|  b|42.50| 4.915|  y|  p|  w|  v|3.165|  t|  f|  0|  t|  g|00052| 1442|    +|
+---+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
```

In [164]:
```
1 # Renomeando categorias
2 data_.withColumn('A1', F.when(data_['A1'] == 'a', 'classe_a').otherwise('class_b'))\
3       .limit(10)\
4       .show()
```

```
+--------+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
|      A1|   A2|    A3| A4| A5| A6| A7|   A8| A9|A10|A11|A12|A13|  A14|  A15|class|
+--------+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
| class_b|30.83|     0|  u|  g|  w|  v| 1.25|  t|  t| 01|  f|  g|00202|    0|    +|
|classe_a|58.67|  4.46|  u|  g|  q|  h| 3.04|  t|  t| 06|  f|  g|00043|  560|    +|
|classe_a|24.50|   0.5|  u|  g|  q|  h|  1.5|  t|  f|  0|  f|  g|00280|  824|    +|
| class_b|27.83|  1.54|  u|  g|  w|  v| 3.75|  t|  t| 05|  t|  g|00100|    3|    +|
| class_b|20.17| 5.625|  u|  g|  w|  v| 1.71|  t|  f|  0|  f|  s|00120|    0|    +|
| class_b|32.08|     4|  u|  g|  m|  v|  2.5|  t|  f|  0|  t|  g|00360|    0|    +|
| class_b|33.17|  1.04|  u|  g|  r|  h|  6.5|  t|  f|  0|  t|  g|00164|31285|    +|
|classe_a|22.92|11.585|  u|  g| cc|  v| 0.04|  t|  f|  0|  f|  g|00080| 1349|    +|
| class_b|54.42|   0.5|  y|  p|  k|  h| 3.96|  t|  f|  0|  f|  g|00180|  314|    +|
| class_b|42.50| 4.915|  y|  p|  w|  v|3.165|  t|  f|  0|  t|  g|00052| 1442|    +|
+--------+-----+------+---+---+---+---+-----+---+---+---+---+---+-----+-----+-----+
```

Exercício de fixação

Cumpra o seguinte roteiro de pré-processamento dos dados de aprovação de crédito:

- Encontre o(s) atributo(s) mais esparsos e transforme o dataset para que ele contenha apenas valores até o terceiro quartil para esses atributos.
- Separe apenas os dados contínuos do dataset.
- Apague todos os registros com qualquer atributo nulo.
- Balanceie a base de dados.
- Escreva o resultado num arquivo csv, no computador.