TÉCNICO LISBOA

Data Analysis and Integration

Data Profiling

Motivation

Data Profiling

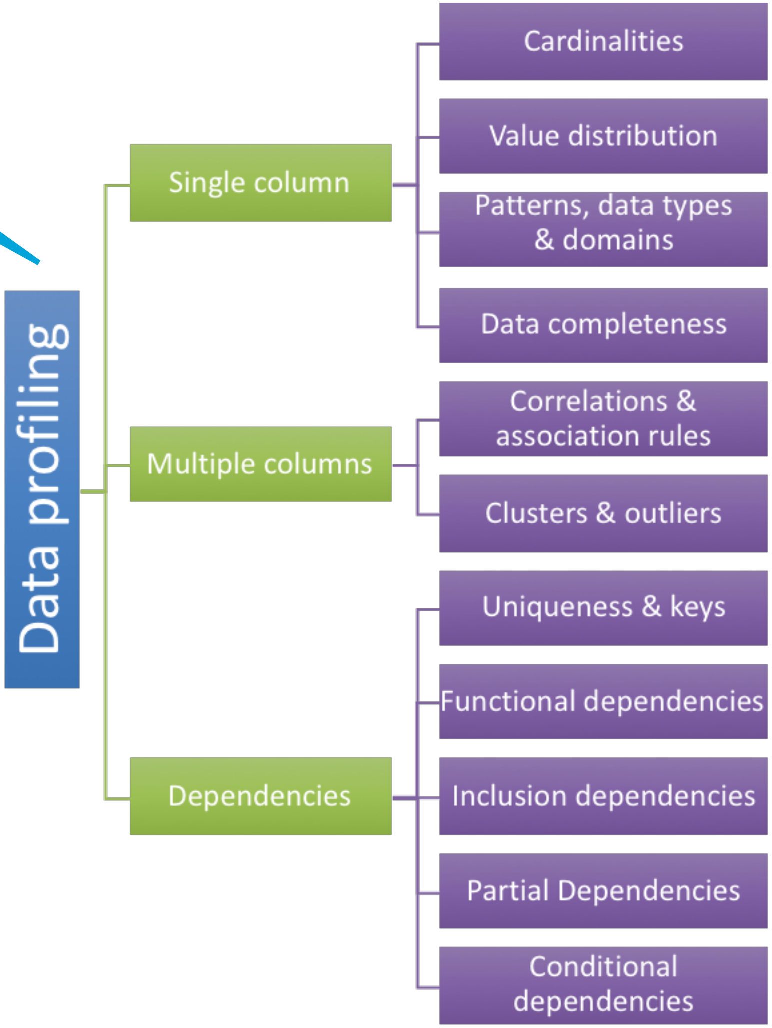
- What is data profiling?
  - Analyze the contents of a data source
  - Gather statistics about the data contained therein
    - minimum, maximum, average, range, value distribution, etc.
  - Identify data quality problems
    - missing or incomplete data, errors, inconsistencies, etc.
  - Understand the logic and relationships between data
    - unique values, keys, foreign keys and other constraints

Data Profiling

- These tasks (and many more) are instances of data profiling
  - number of rows, number of null values, number of distinct values
  - minimum value, maximum value, minimum length, maximum length
  - single- and multi-column frequency histogram
  - precision of numeric values, length of string values
  - data type discovery
  - uniqueness and constancy
  - single- and multi-column primary key discovery
  - single- and multi-column foreign key discovery
  - value overlap (cross-domain analysis)
  - text field profiling
  - pattern discovery (e.g. phone number patterns)
  - soundex frequency histogram
  - etc.

Data Profiling

Taxonomy of Data Profiling Operations



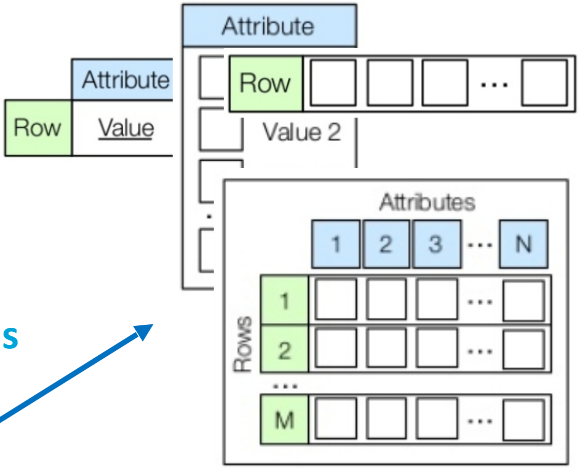
Z. Abedjan, L. Golab, F. Naumann
Profiling relational data: a survey
VLDB Journal, vol. 24, no. 4, 2015

Data Profiling

- Data Profiling tools

Data profiling

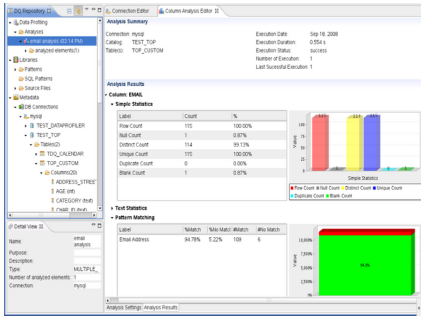
data quality problems



metadata

- data types
- keys
- dependencies
- …

data statistics



Detail View

Name: email analysis
Purpose:
Description:
Type: MULTIPLE_
Number of analyzed elements: 1
Connection: mysql

Analysis Summary

Connection: mysql
Catalog: TEST_TOP
Table(s): TOP_CUSTOM

Execution Date: Sep 18, 2008
Execution Duration: 0.554 s
Execution Status: success
Number of Execution: 1
Last Sucessful Execution: 1

Analysis Results

Column: EMAIL

Simple Statistics

Label	Count	%
Row Count	115	100.00%
Null Count	1	0.87%
Distinct Count	114	99.13%
Unique Count	115	100.00%
Duplicate Count	0	0.00%
Blank Count	1	0.87%

Text Statistics

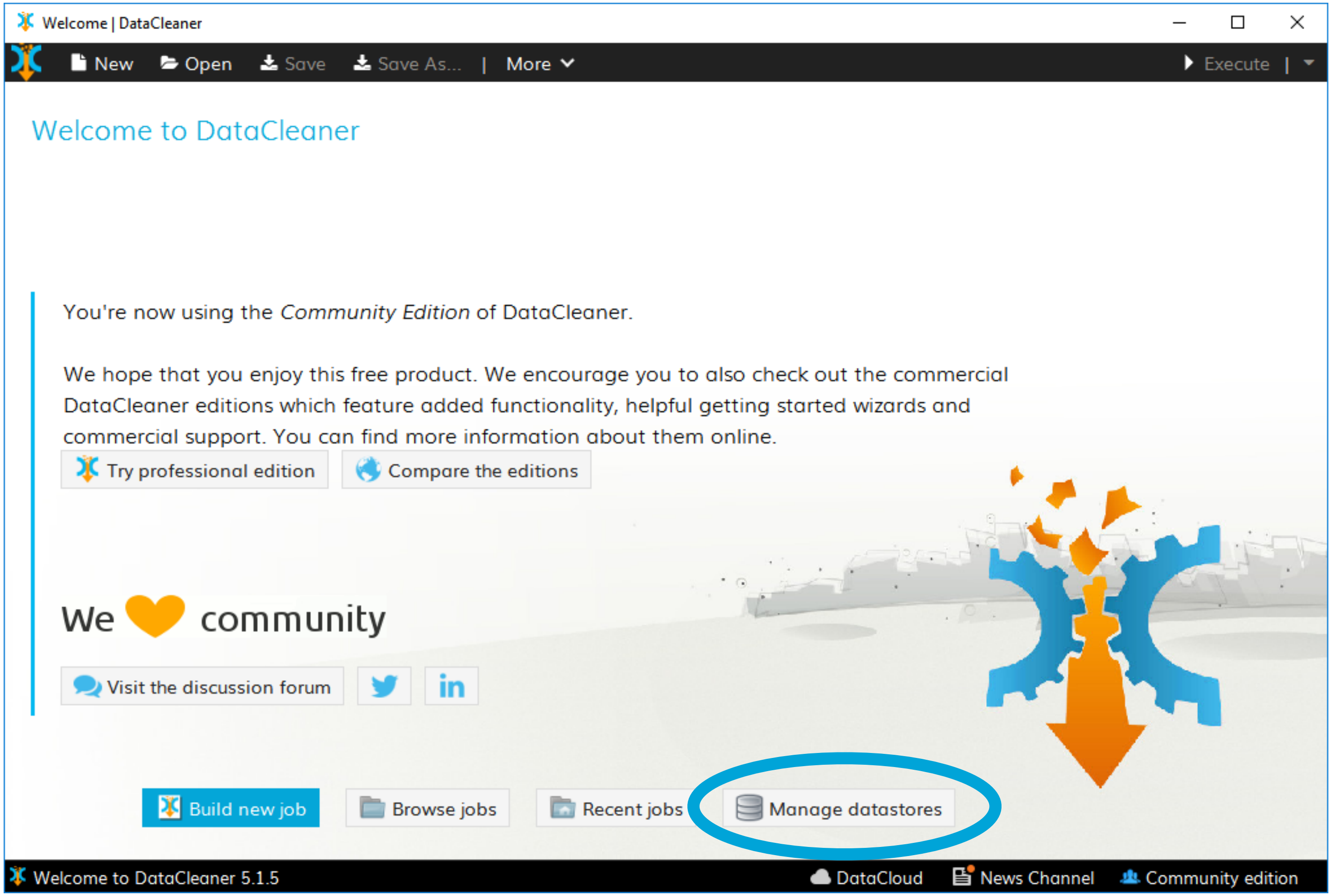
Pattern Matching

Label	%Match	%No Matcl	#Match	#No Match
Email Address	94.78%	5.22%	109	6

Data Profiling Tool Concepts

Data Profiling

- Data profiling tools

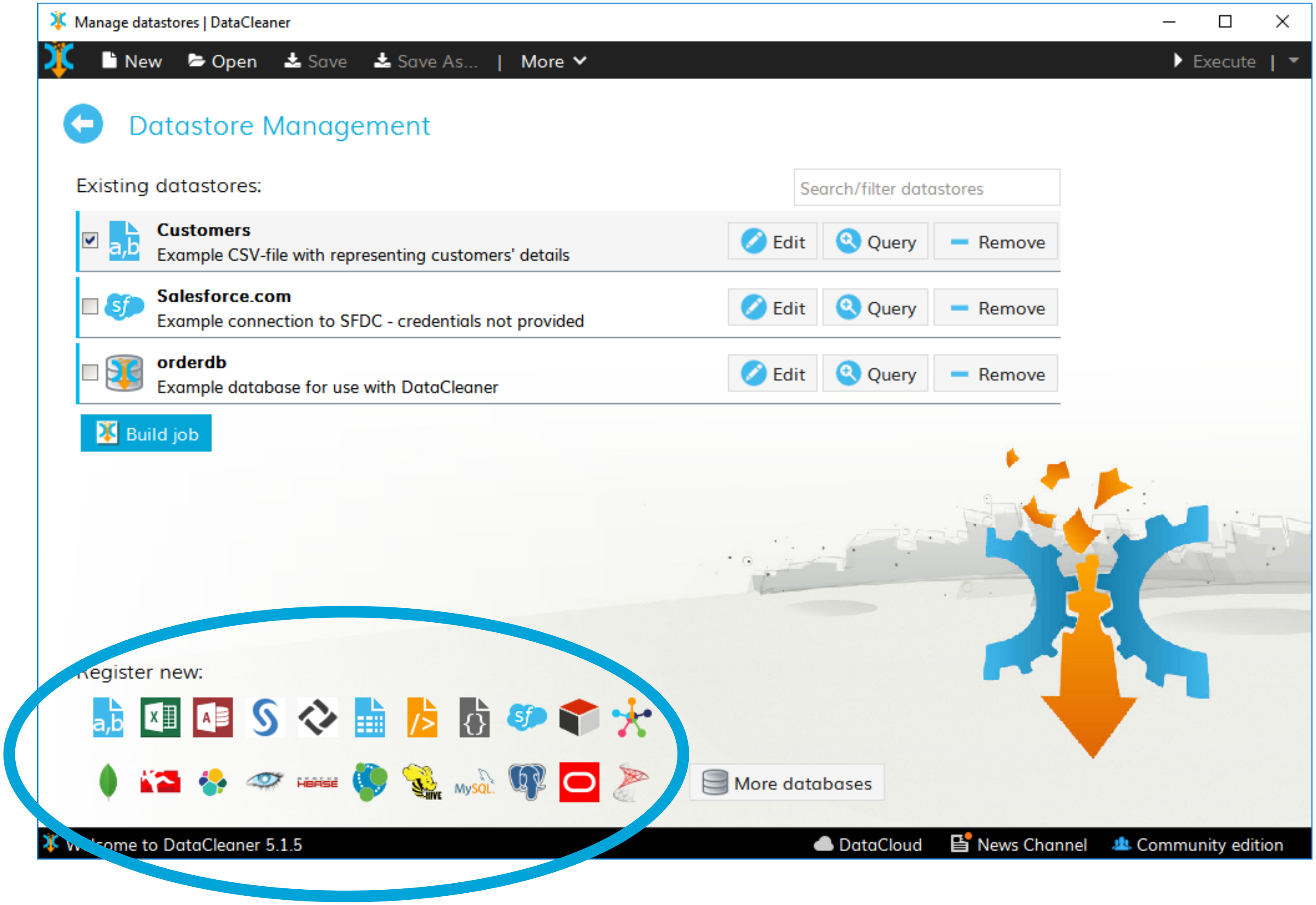


You're now using the *Community Edition* of DataCleaner.

We hope that you enjoy this free product. We encourage you to also check out the commercial DataCleaner editions which feature added functionality, helpful getting started wizards and commercial support. You can find more information about them online.

Data Profiling

- A datastore can be a file, database, etc.



Manage datastores | DataCleaner

Datastore Management

Existing datastores:

Search/filter datastores

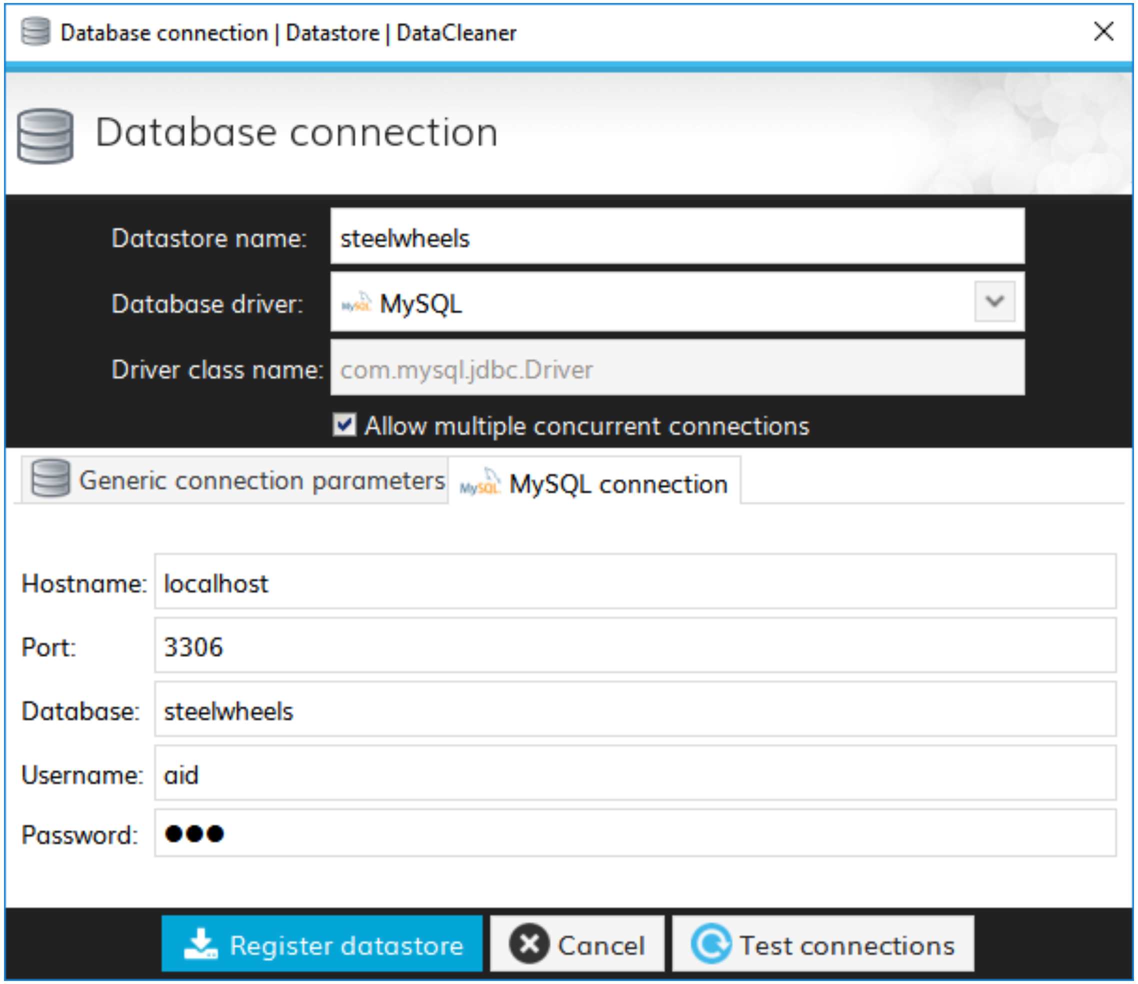
☑	**Customers** Example CSV-file with representing customers' details	Edit	Query	Remove
□	**Salesforce.com** Example connection to SFDC - credentials not provided	Edit	Query	Remove
□	**orderdb** Example database for use with DataCleaner	Edit	Query	Remove

Register new:

Welcome to DataCleaner 5.1.5

Data Profiling

- Creating a new datastore



Database connection | Datastore | DataCleaner

Database connection

Datastore name: steelwheels

Database driver: MySQL

Driver class name: com.mysql.jdbc.Driver

☑ Allow multiple concurrent connections

Hostname: localhost

Port: 3306

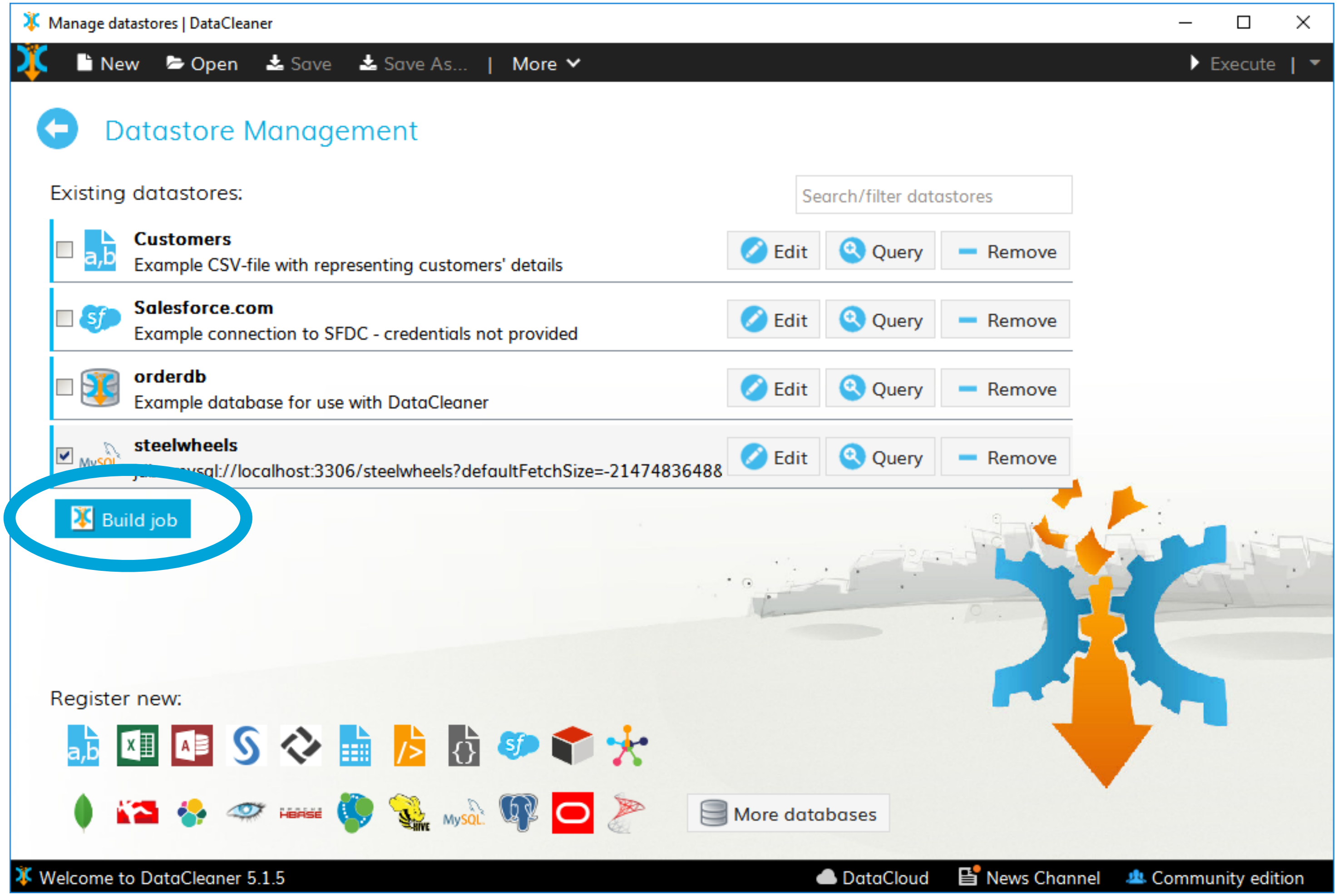
Database: steelwheels

Username: aid

Password: ●●●

Data Profiling

- Creating a new datastore



Datastore Management

Existing datastores:

☐ **Customers**
Example CSV-file with representing customers' details

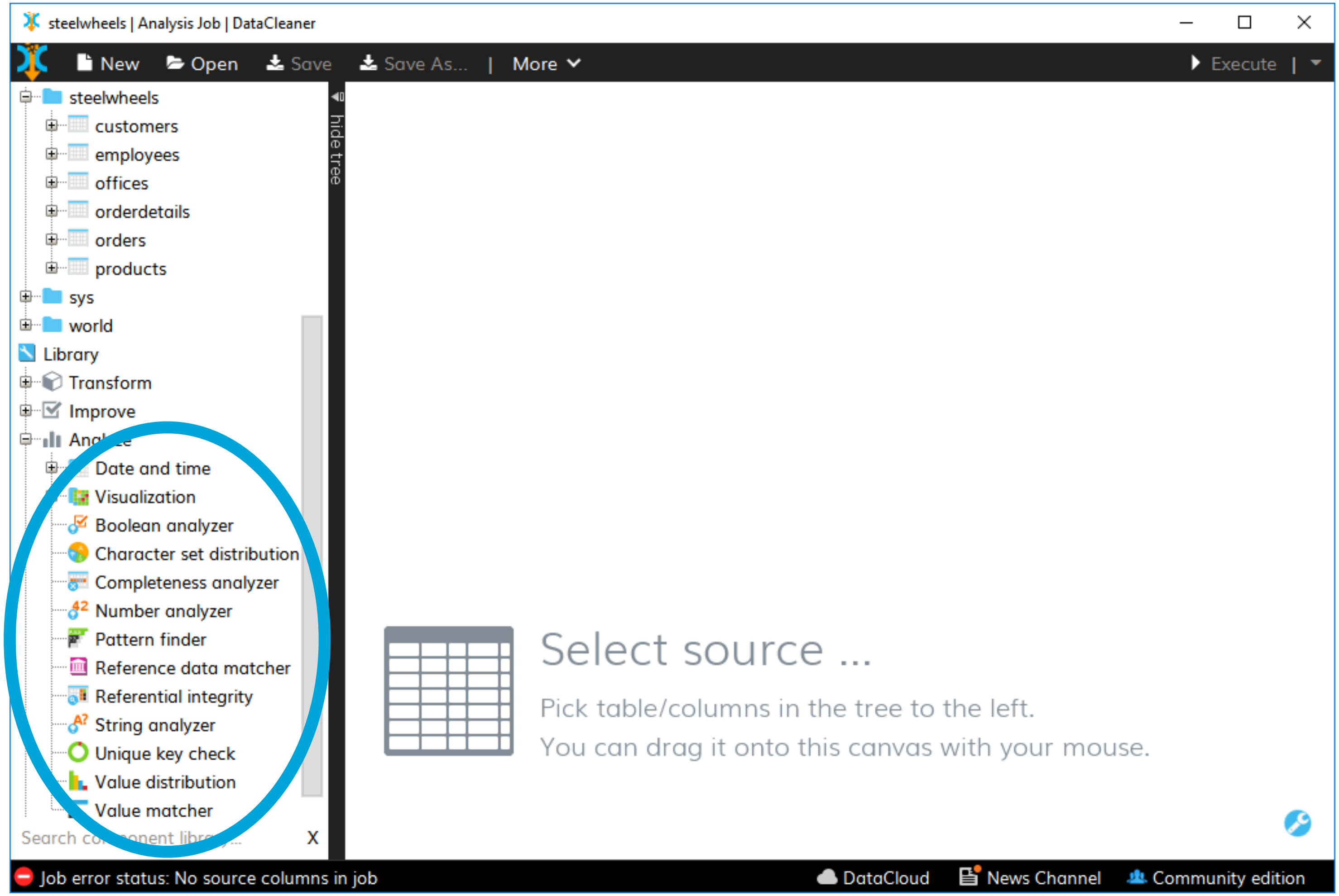
☐ **Salesforce.com**
Example connection to SFDC - credentials not provided

☐ **orderdb**
Example database for use with DataCleaner

☑ **steelwheels**
jdbc:mysql://localhost:3306/steelwheels?defaultFetchSize=-2147483488

Data Profiling

- Several options for data profiling tasks



Select source ...

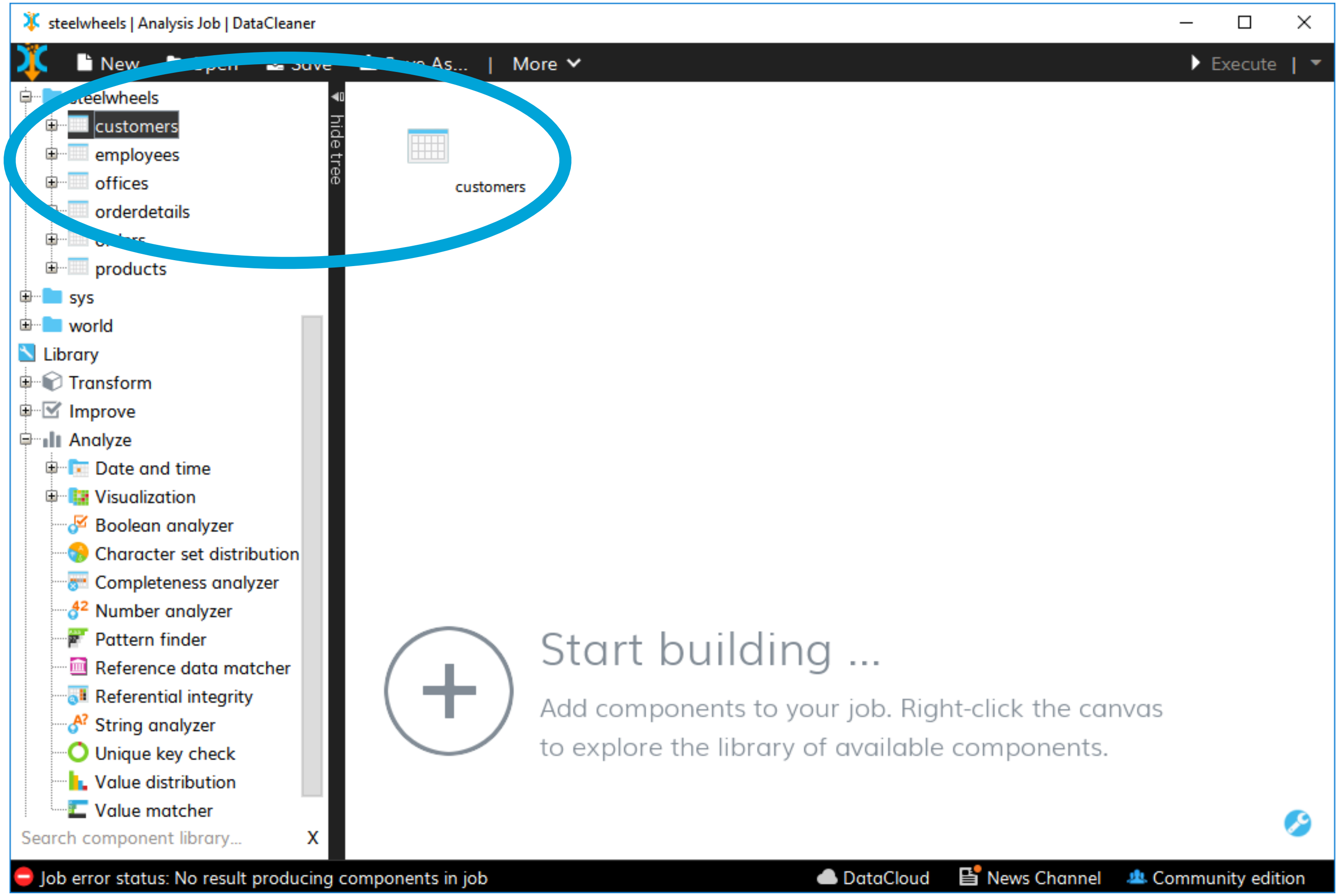
Pick table/columns in the tree to the left.

You can drag it onto this canvas with your mouse.

Job error status: No source columns in job

Data Profiling

- Selecting the data source



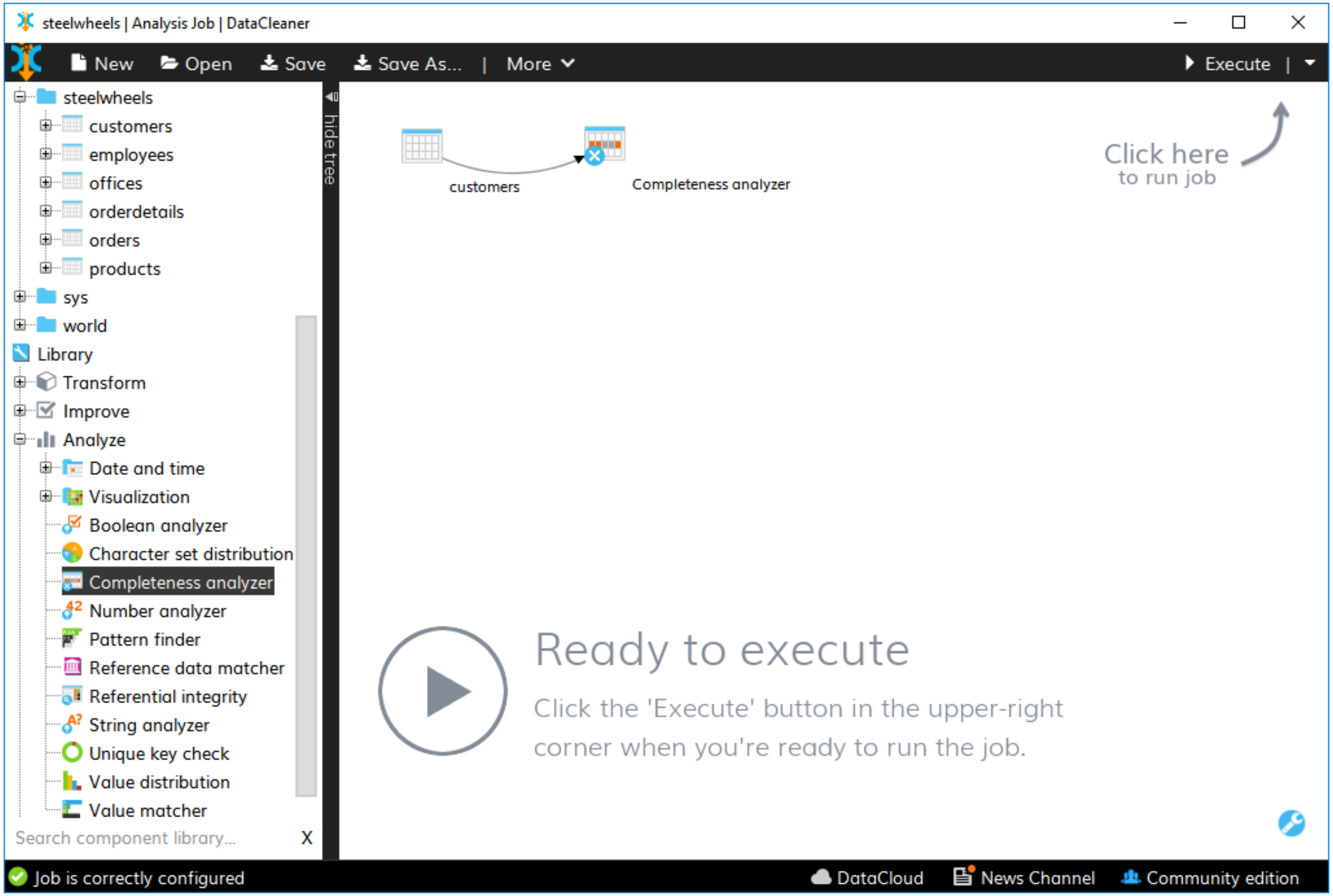
Start building ...

Add components to your job. Right-click the canvas to explore the library of available components.

Job error status: No result producing components in job

Data Profiling

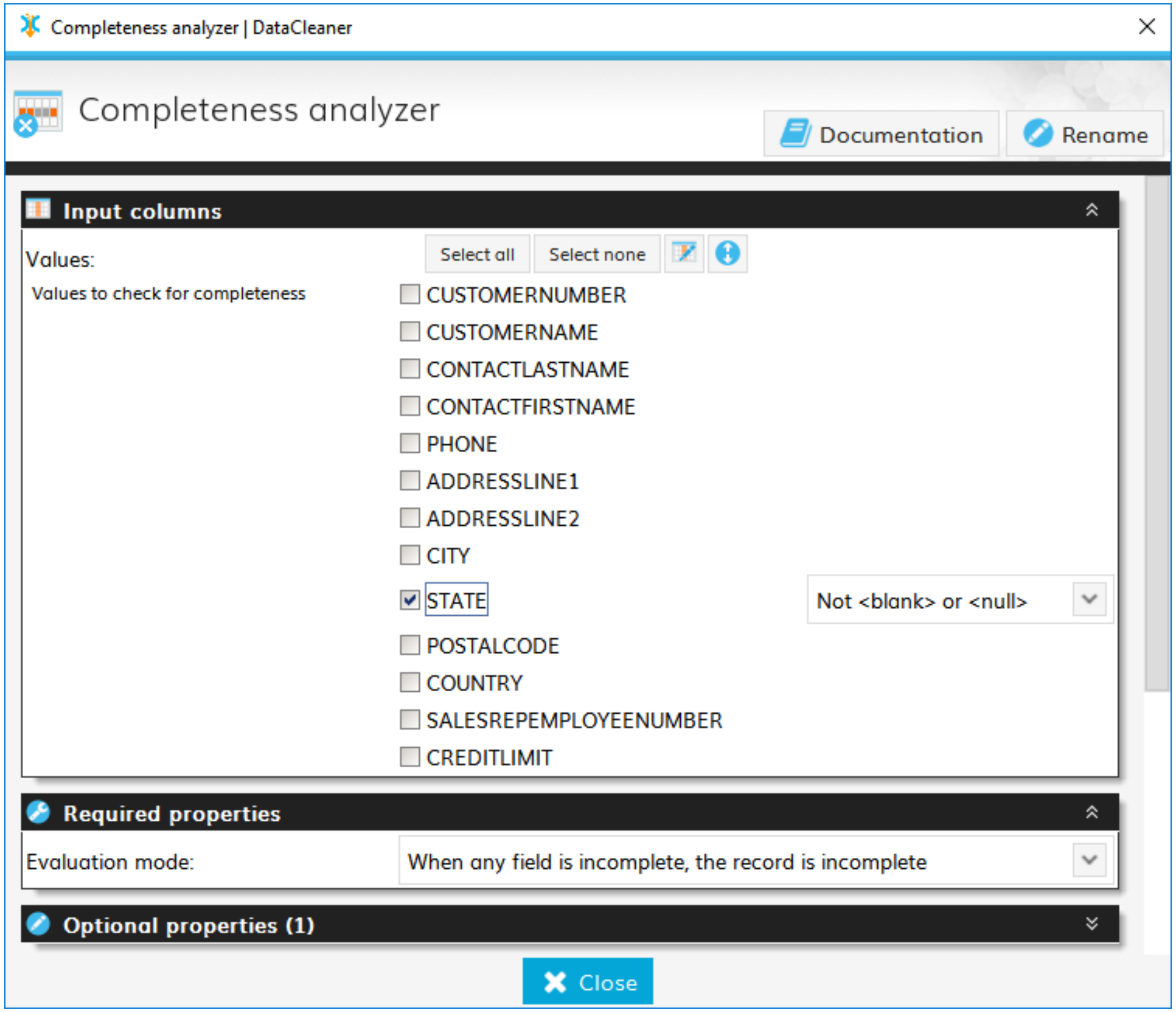
- Completeness analysis



Click the 'Execute' button in the upper-right corner when you're ready to run the job.

Data Profiling

- Completeness analysis



Completeness analyzer | DataCleaner

Completeness analyzer

Input columns

Values:

Values to check for completeness

□ CUSTOMERNUMBER
□ CUSTOMERNAME
□ CONTACTLASTNAME
□ CONTACTFIRSTNAME
□ PHONE
□ ADDRESSLINE1
□ ADDRESSLINE2
□ CITY
☑ STATE Not <blank> or <null>
□ POSTALCODE
□ COUNTRY
□ SALESREPEMPLOYEENUMBER
□ CREDITLIMIT

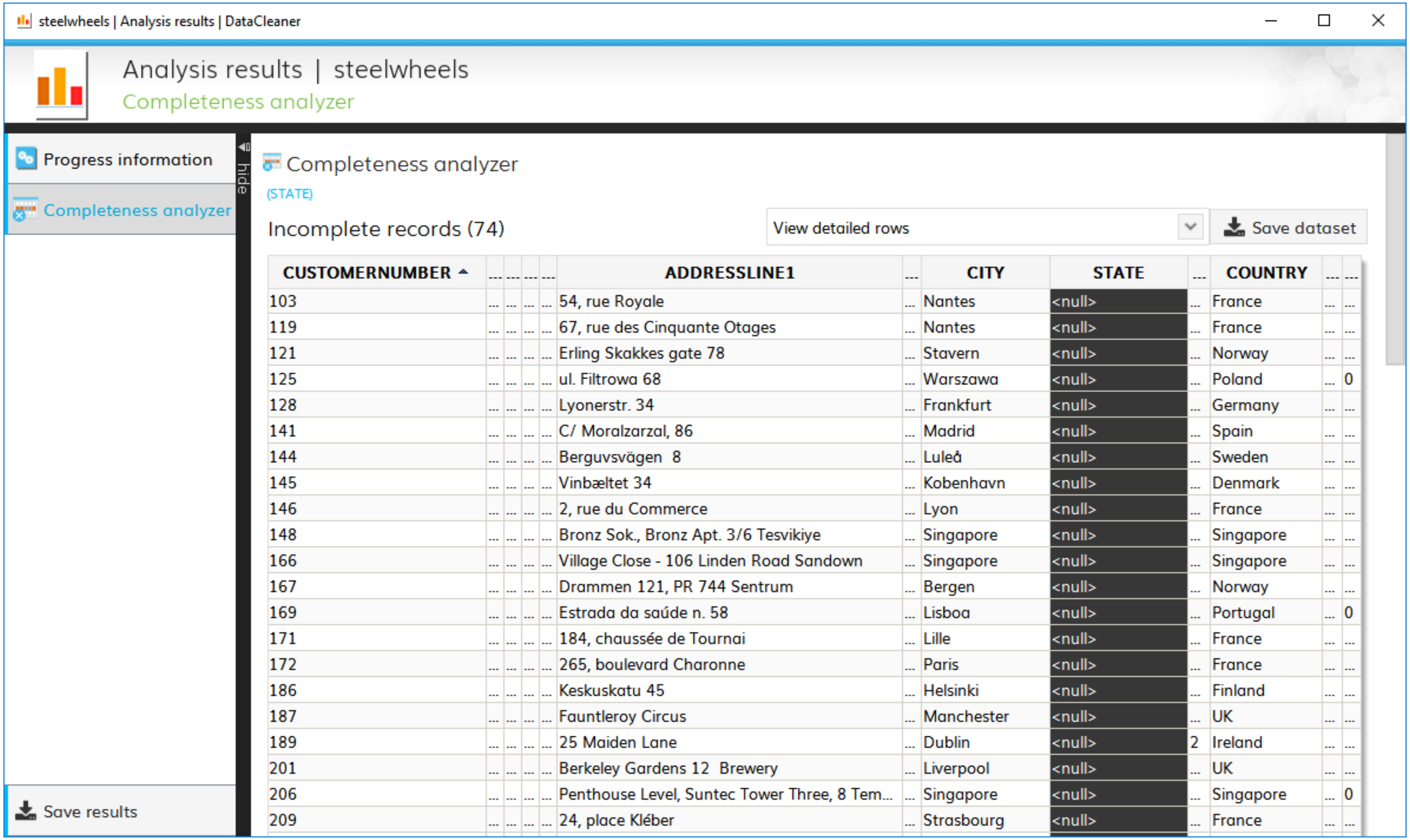
Required properties

Evaluation mode: When any field is incomplete, the record is incomplete

Optional properties (1)

Data Profiling

- Completeness analysis



Completeness analyzer

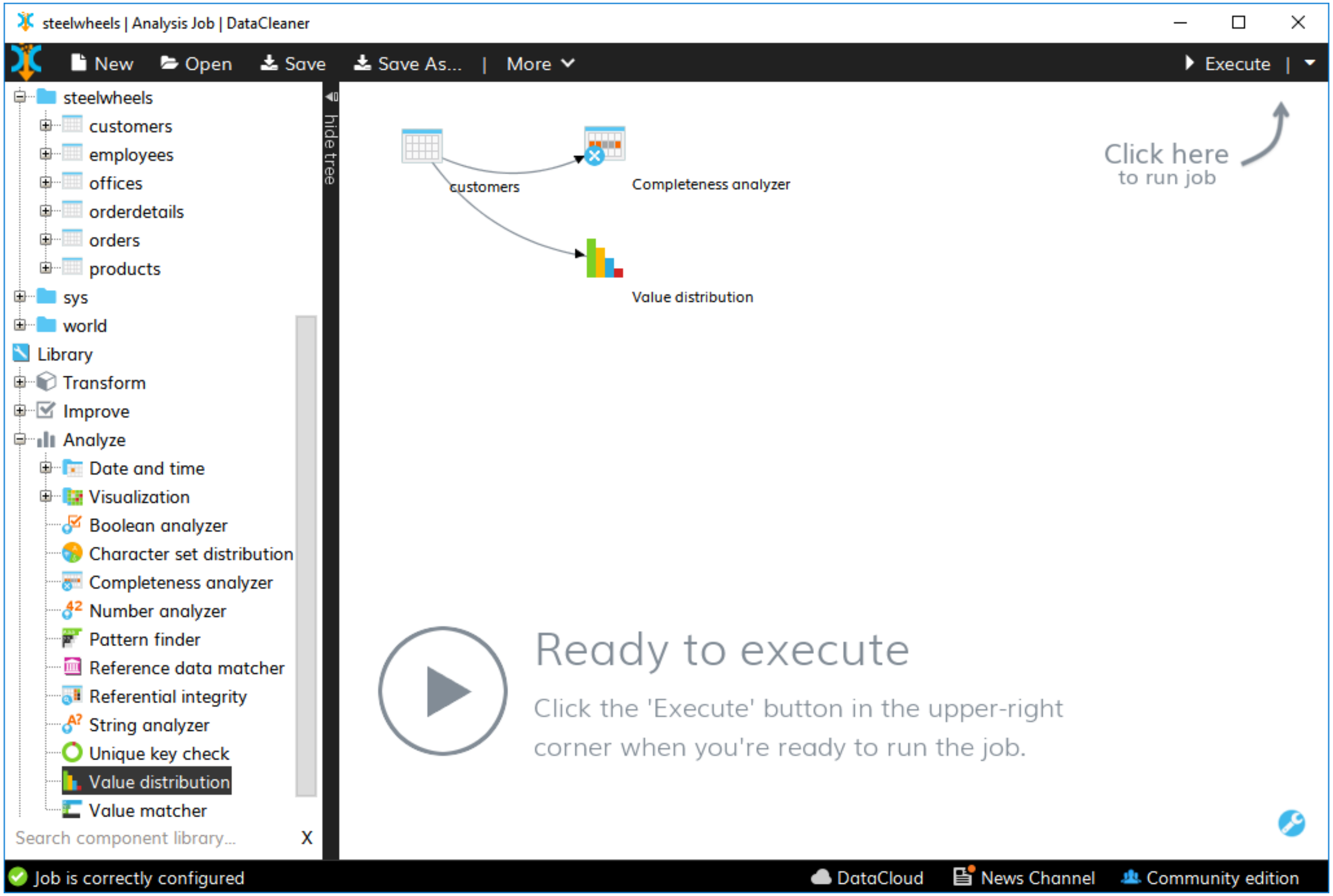
(STATE)

Incomplete records (74)

CUSTOMERNUMBER ▲	...	...	...	...	ADDRESSLINE1	...	CITY	STATE	...	COUNTRY	...	...
103	...	...	...	...	54, rue Royale	...	Nantes	<null>	...	France	...	...
119	...	...	...	...	67, rue des Cinquante Otages	...	Nantes	<null>	...	France	...	...
121	...	...	...	...	Erling Skakkes gate 78	...	Stavern	<null>	...	Norway	...	...
125	...	...	...	...	ul. Filtrowa 68	...	Warszawa	<null>	...	Poland	...	0
128	...	...	...	...	Lyonerstr. 34	...	Frankfurt	<null>	...	Germany	...	...
141	...	...	...	...	C/ Moralzarzal, 86	...	Madrid	<null>	...	Spain	...	...
144	...	...	...	...	Berguvsvägen 8	...	Luleå	<null>	...	Sweden	...	...
145	...	...	...	...	Vinbæltet 34	...	Kobenhavn	<null>	...	Denmark	...	...
146	...	...	...	...	2, rue du Commerce	...	Lyon	<null>	...	France	...	...
148	...	...	...	...	Bronz Sok., Bronz Apt. 3/6 Tesvikiye	...	Singapore	<null>	...	Singapore	...	...
166	...	...	...	...	Village Close - 106 Linden Road Sandown	...	Singapore	<null>	...	Singapore	...	...
167	...	...	...	...	Drammen 121, PR 744 Sentrum	...	Bergen	<null>	...	Norway	...	...
169	...	...	...	...	Estrada da saúde n. 58	...	Lisboa	<null>	...	Portugal	...	0
171	...	...	...	...	184, chaussée de Tournai	...	Lille	<null>	...	France	...	...
172	...	...	...	...	265, boulevard Charonne	...	Paris	<null>	...	France	...	...
186	...	...	...	...	Keskuskatu 45	...	Helsinki	<null>	...	Finland	...	...
187	...	...	...	...	Fauntleroy Circus	...	Manchester	<null>	...	UK	...	...
189	...	...	...	...	25 Maiden Lane	...	Dublin	<null>	2	Ireland	...	...
201	...	...	...	...	Berkeley Gardens 12 Brewery	...	Liverpool	<null>	...	UK	...	...
206	...	...	...	...	Penthouse Level, Suntec Tower Three, 8 Tem...	...	Singapore	<null>	...	Singapore	...	0
209	...	...	...	...	24, place Kléber	...	Strasbourg	<null>	...	France	...	...

Data Profiling

- Value distribution



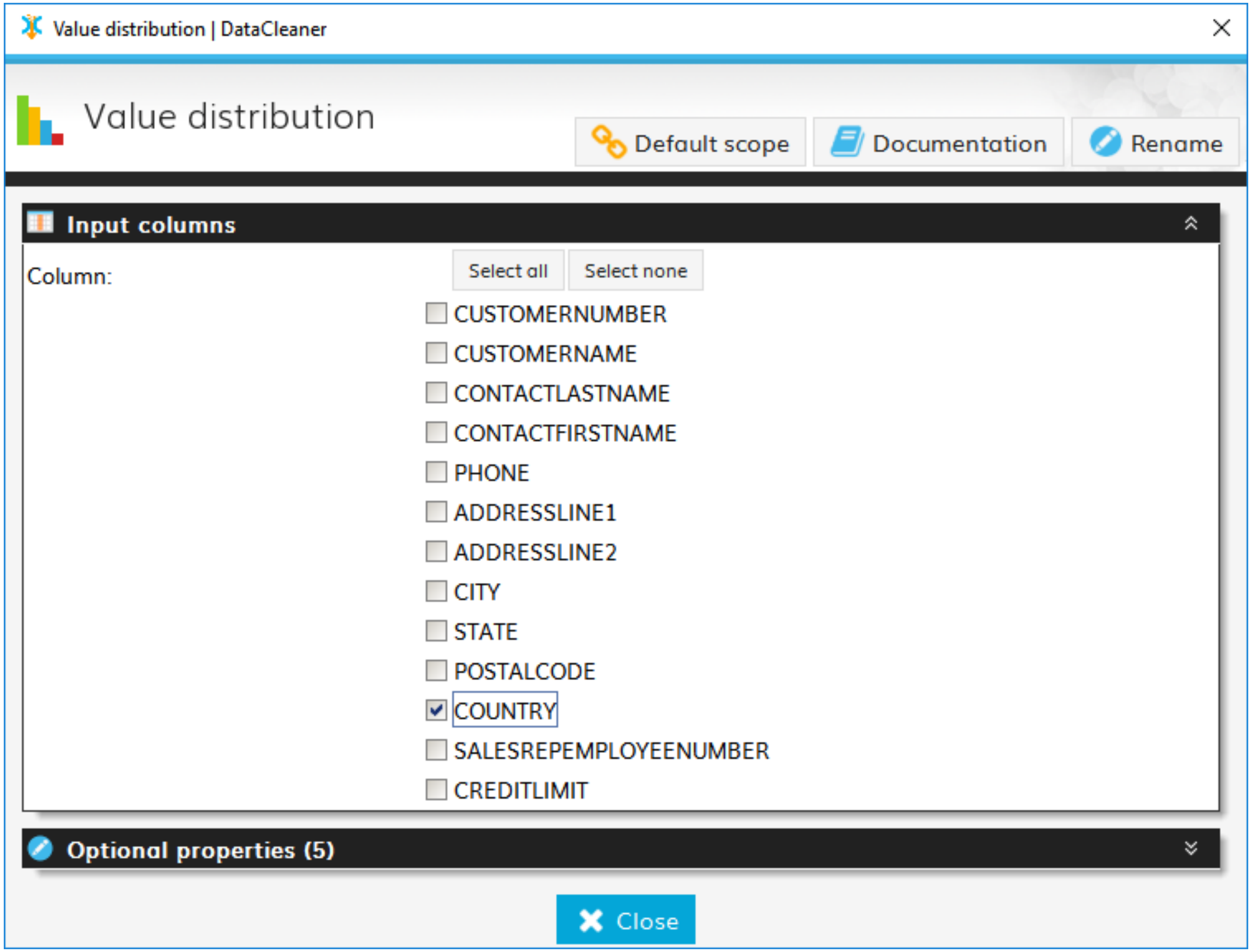
Ready to execute

Click the 'Execute' button in the upper-right corner when you're ready to run the job.

Job is correctly configured

Data Profiling

- Value distribution



Value distribution | DataCleaner

Value distribution

Input columns

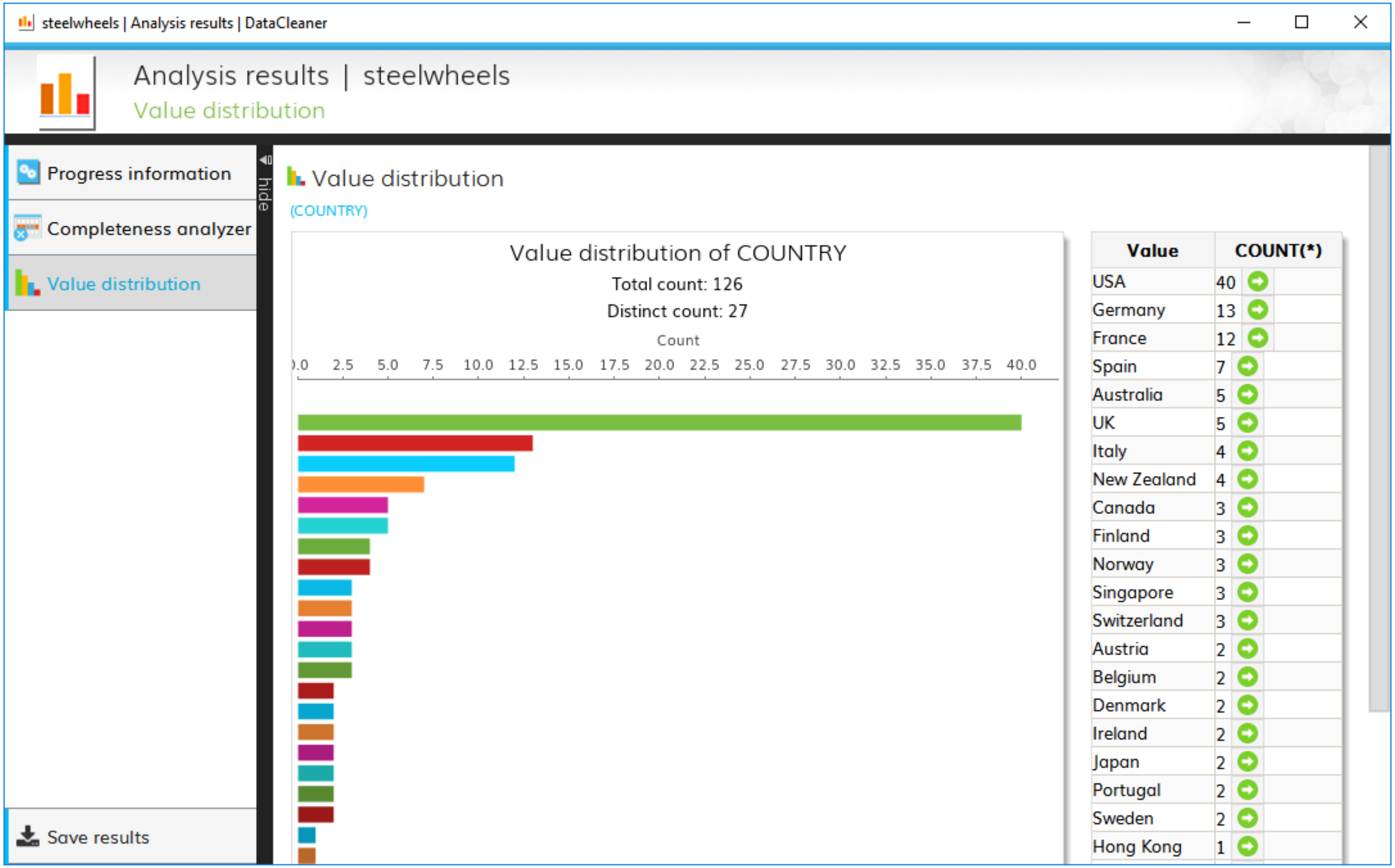
Column:

□ CUSTOMERNUMBER
□ CUSTOMERNAME
□ CONTACTLASTNAME
□ CONTACTFIRSTNAME
□ PHONE
□ ADDRESSLINE1
□ ADDRESSLINE2
□ CITY
□ STATE
□ POSTALCODE
☑ COUNTRY
□ SALESREPEMPLOYEENUMBER
□ CREDITLIMIT

Optional properties (5)

Data Profiling

- Value distribution



Value distribution of COUNTRY

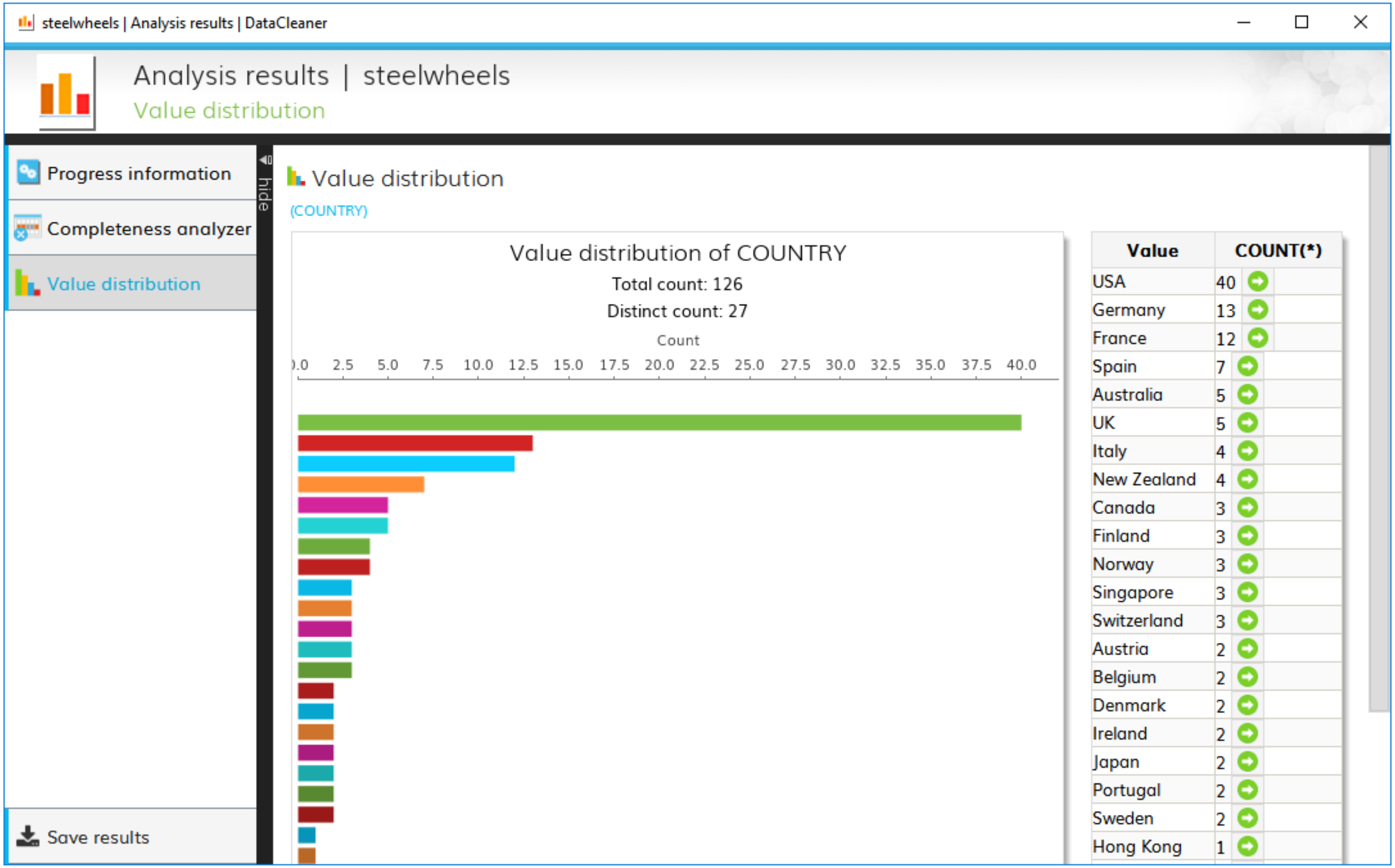
Total count: 126

Distinct count: 27

Value	COUNT(*)
USA	40
Germany	13
France	12
Spain	7
Australia	5
UK	5
Italy	4
New Zealand	4
Canada	3
Finland	3
Norway	3
Singapore	3
Switzerland	3
Austria	2
Belgium	2
Denmark	2
Ireland	2
Japan	2
Portugal	2
Sweden	2
Hong Kong	1

Data Profiling

- Value distribution



steelwheels | Analysis results | DataCleaner

Analysis results | steelwheels

Value distribution

Value distribution

(COUNTRY)

Value distribution of COUNTRY

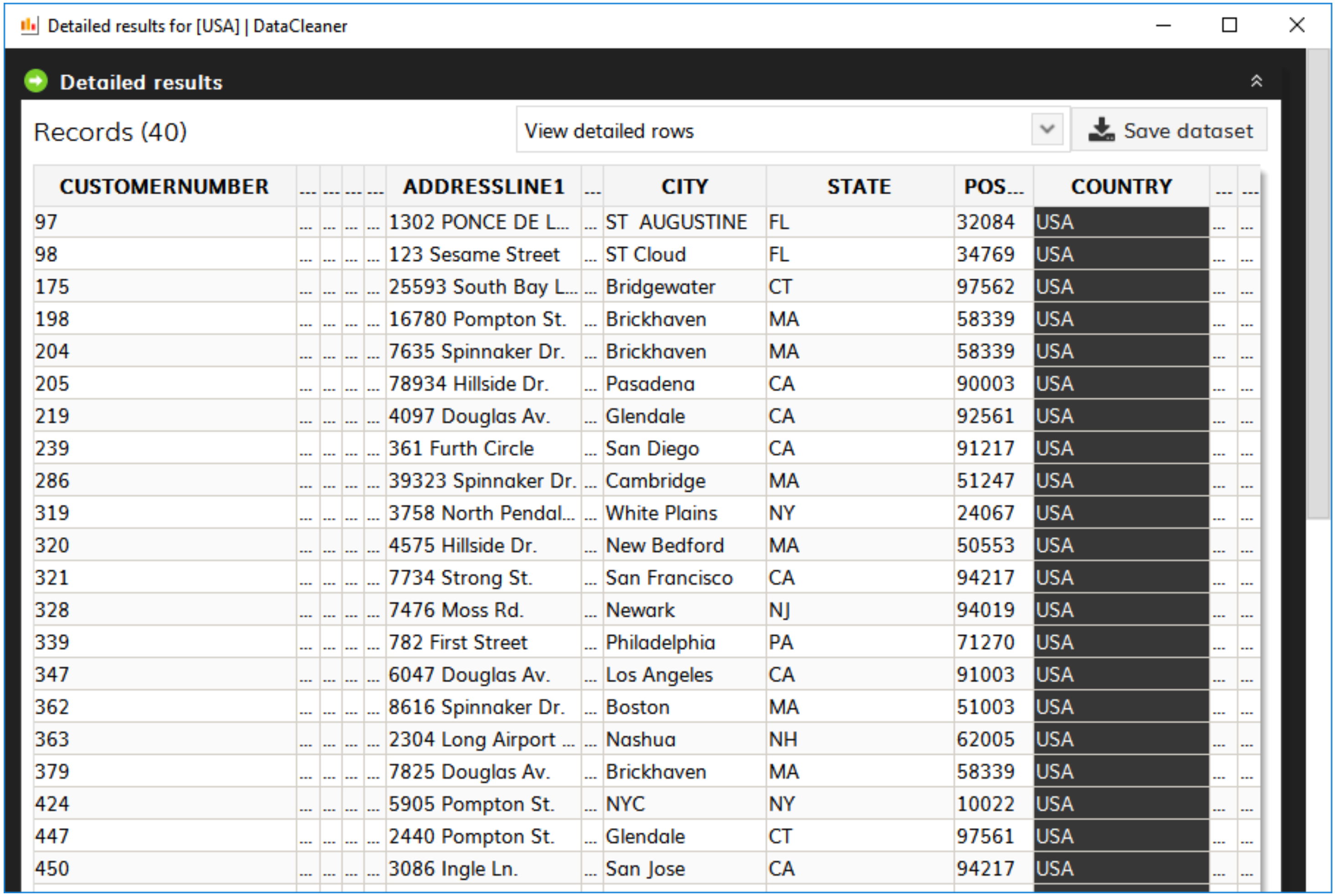
Total count: 126

Distinct count: 27

Value	COUNT(*)
USA	40
Germany	13
France	12
Spain	7
Australia	5
UK	5
Italy	4
New Zealand	4
Canada	3
Finland	3
Norway	3
Singapore	3
Switzerland	3
Austria	2
Belgium	2
Denmark	2
Ireland	2
Japan	2
Portugal	2
Sweden	2
Hong Kong	1

Data Profiling

- Value distribution – detailed results

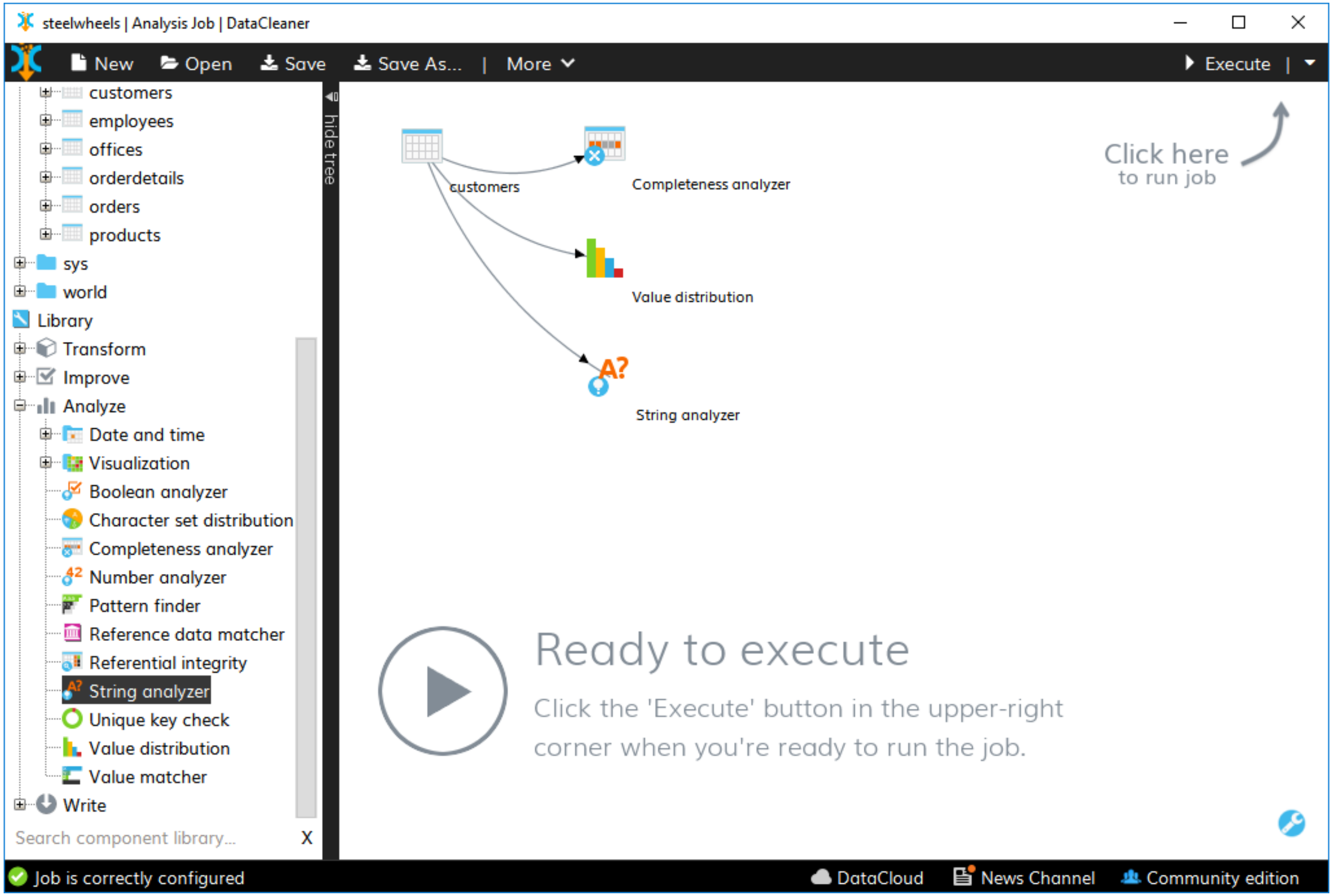


Records (40)

CUSTOMERNUMBER	...	...	...	...	ADDRESSLINE1	...	CITY	STATE	POS...	COUNTRY	...	...
97	...	...	...	...	1302 PONCE DE L...	...	ST AUGUSTINE	FL	32084	USA	...	...
98	...	...	...	...	123 Sesame Street	...	ST Cloud	FL	34769	USA	...	...
175	...	...	...	...	25593 South Bay L...	...	Bridgewater	CT	97562	USA	...	...
198	...	...	...	...	16780 Pompton St.	...	Brickhaven	MA	58339	USA	...	...
204	...	...	...	...	7635 Spinnaker Dr.	...	Brickhaven	MA	58339	USA	...	...
205	...	...	...	...	78934 Hillside Dr.	...	Pasadena	CA	90003	USA	...	...
219	...	...	...	...	4097 Douglas Av.	...	Glendale	CA	92561	USA	...	...
239	...	...	...	...	361 Furth Circle	...	San Diego	CA	91217	USA	...	...
286	...	...	...	...	39323 Spinnaker Dr.	...	Cambridge	MA	51247	USA	...	...
319	...	...	...	...	3758 North Pendal...	...	White Plains	NY	24067	USA	...	...
320	...	...	...	...	4575 Hillside Dr.	...	New Bedford	MA	50553	USA	...	...
321	...	...	...	...	7734 Strong St.	...	San Francisco	CA	94217	USA	...	...
328	...	...	...	...	7476 Moss Rd.	...	Newark	NJ	94019	USA	...	...
339	...	...	...	...	782 First Street	...	Philadelphia	PA	71270	USA	...	...
347	...	...	...	...	6047 Douglas Av.	...	Los Angeles	CA	91003	USA	...	...
362	...	...	...	...	8616 Spinnaker Dr.	...	Boston	MA	51003	USA	...	...
363	...	...	...	...	2304 Long Airport ...	...	Nashua	NH	62005	USA	...	...
379	...	...	...	...	7825 Douglas Av.	...	Brickhaven	MA	58339	USA	...	...
424	...	...	...	...	5905 Pompton St.	...	NYC	NY	10022	USA	...	...
447	...	...	...	...	2440 Pompton St.	...	Glendale	CT	97561	USA	...	...
450	...	...	...	...	3086 Ingle Ln.	...	San Jose	CA	94217	USA	...	...

Data Profiling

- String analysis



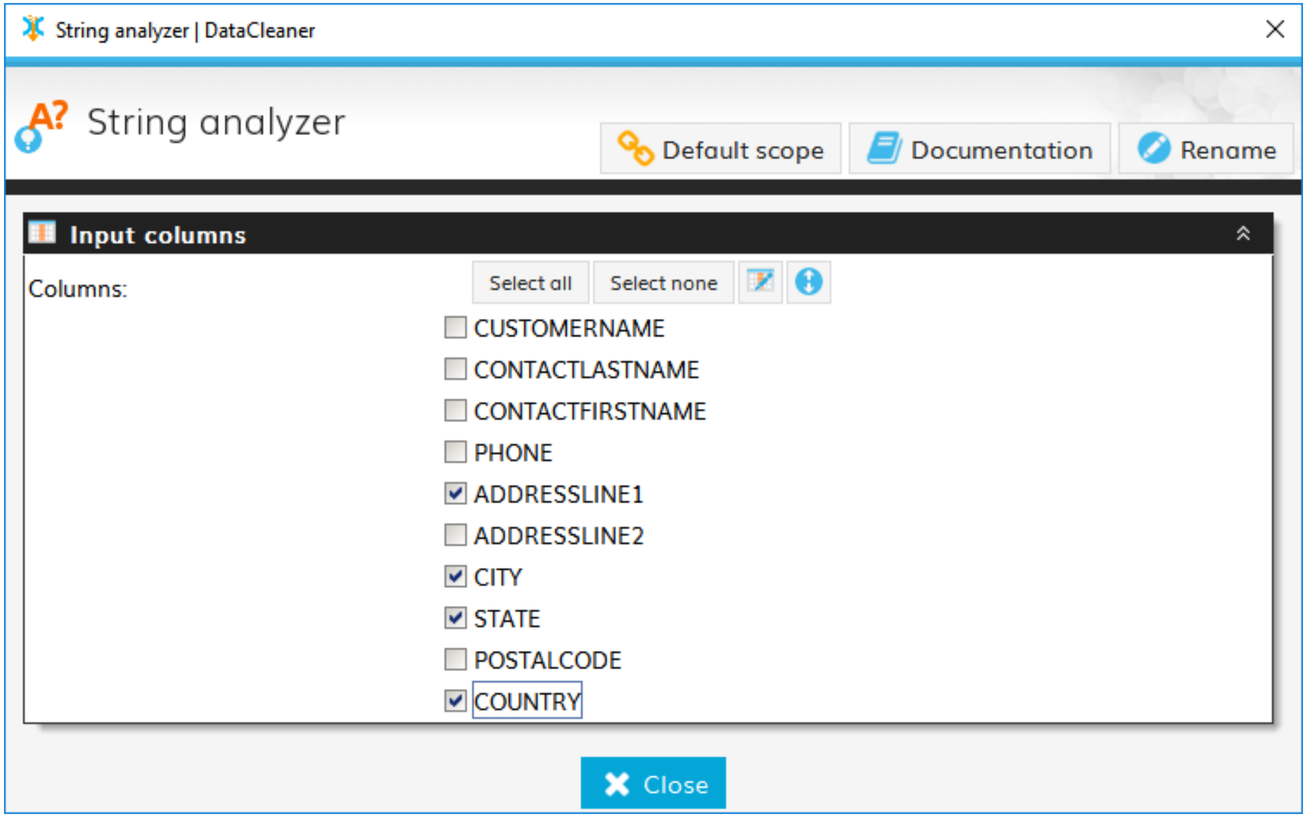
Ready to execute

Click the 'Execute' button in the upper-right corner when you're ready to run the job.

Job is correctly configured

Data Profiling

- String analysis



String analyzer

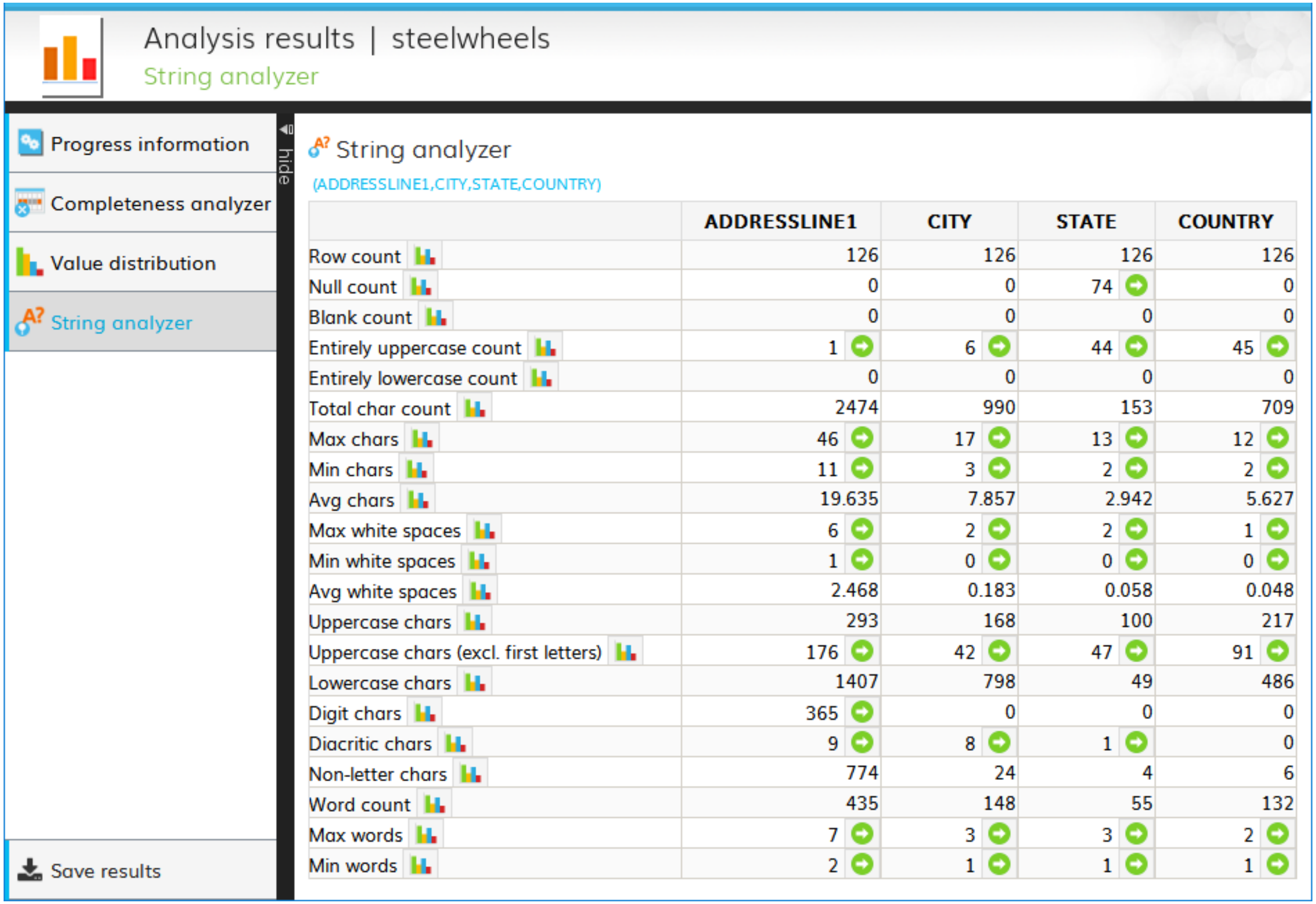
Input columns

Columns:

□ CUSTOMERNAME
□ CONTACTLASTNAME
□ CONTACTFIRSTNAME
□ PHONE
☑ ADDRESSLINE1
□ ADDRESSLINE2
☑ CITY
☑ STATE
□ POSTALCODE
☑ COUNTRY

Data Profiling

- String analysis



Analysis results | steelwheels

String analyzer

String analyzer

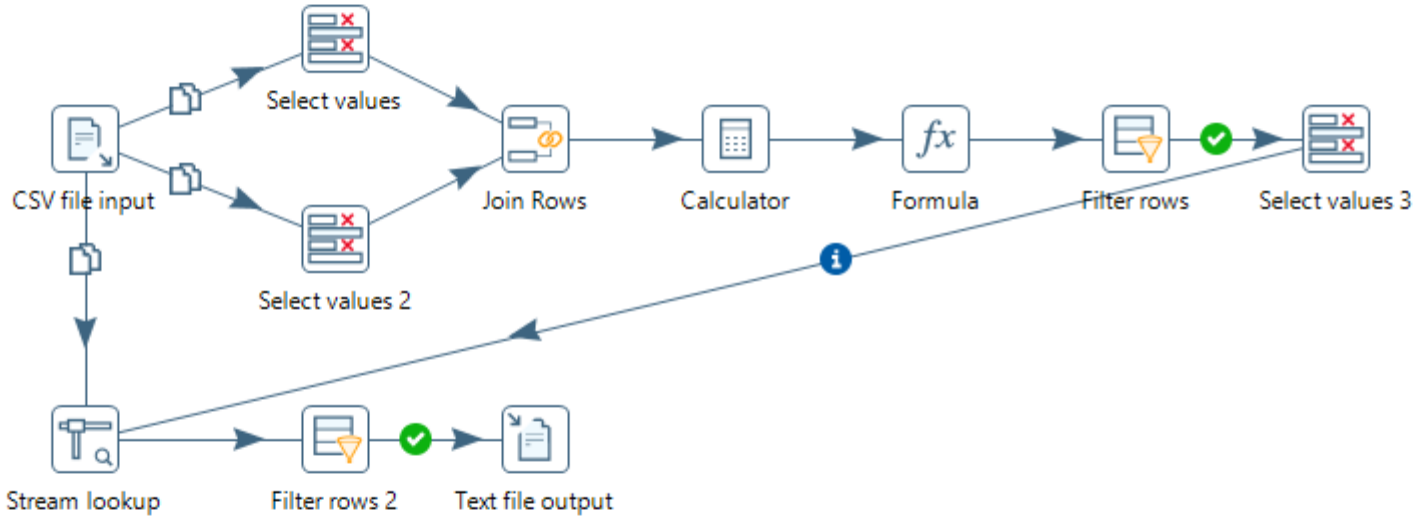
(ADDRESSLINE1,CITY,STATE,COUNTRY)

	ADDRESSLINE1	CITY	STATE	COUNTRY
Row count	126	126	126	126
Null count	0	0	74	0
Blank count	0	0	0	0
Entirely uppercase count	1	6	44	45
Entirely lowercase count	0	0	0	0
Total char count	2474	990	153	709
Max chars	46	17	13	12
Min chars	11	3	2	2
Avg chars	19.635	7.857	2.942	5.627
Max white spaces	6	2	2	1
Min white spaces	1	0	0	0
Avg white spaces	2.468	0.183	0.058	0.048
Uppercase chars	293	168	100	217
Uppercase chars (excl. first letters)	176	42	47	91
Lowercase chars	1407	798	49	486
Digit chars	365	0	0	0
Diacritic chars	9	8	1	0
Non-letter chars	774	24	4	6
Word count	435	148	55	132
Max words	7	3	3	2
Min words	2	1	1	1

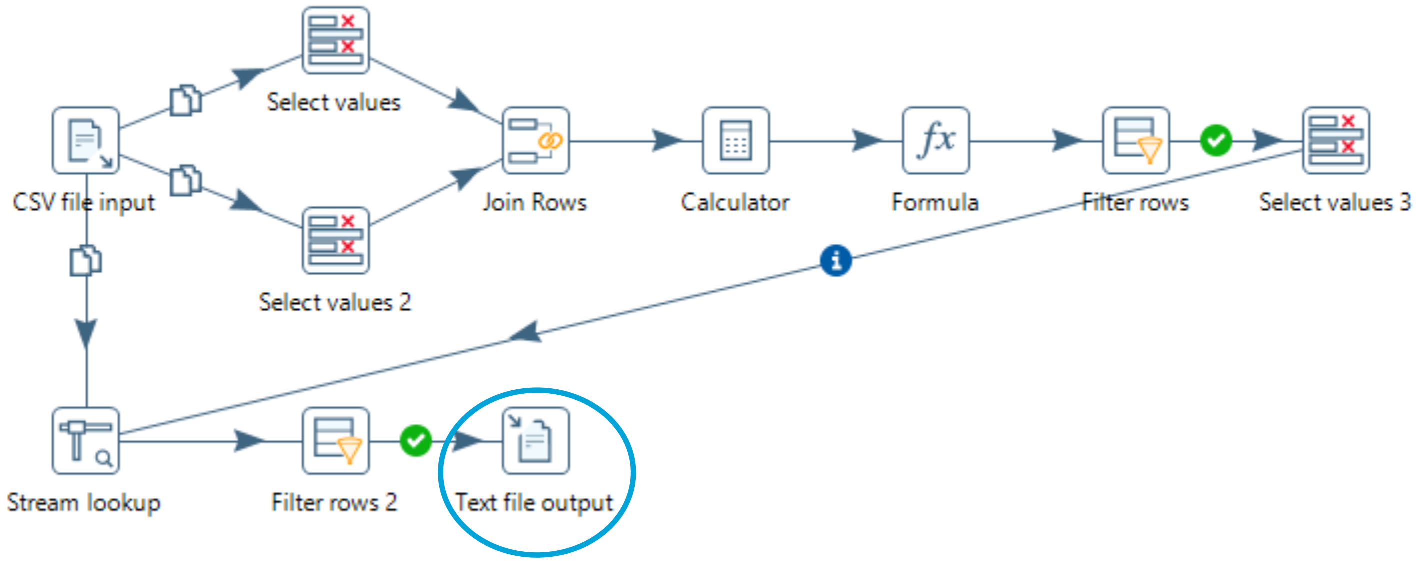
Integration

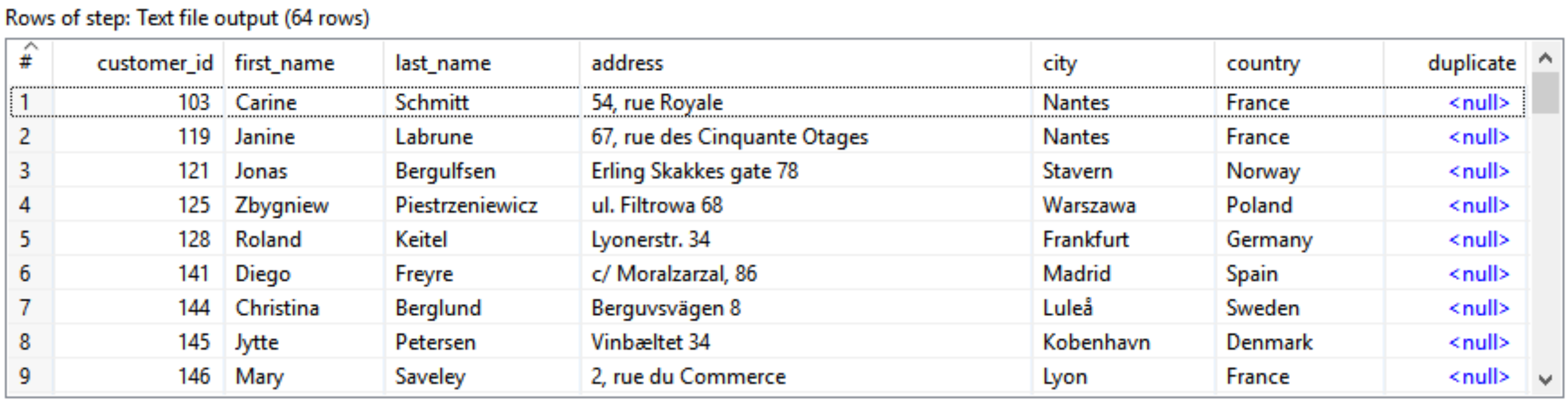
Data Profiling Integration

- Integration between tools
  - Pentaho Data Integration (PDI) with DataCleaner plugin
  - The output of any transformation step can be a data source for data profiling



Data Profiling Integration



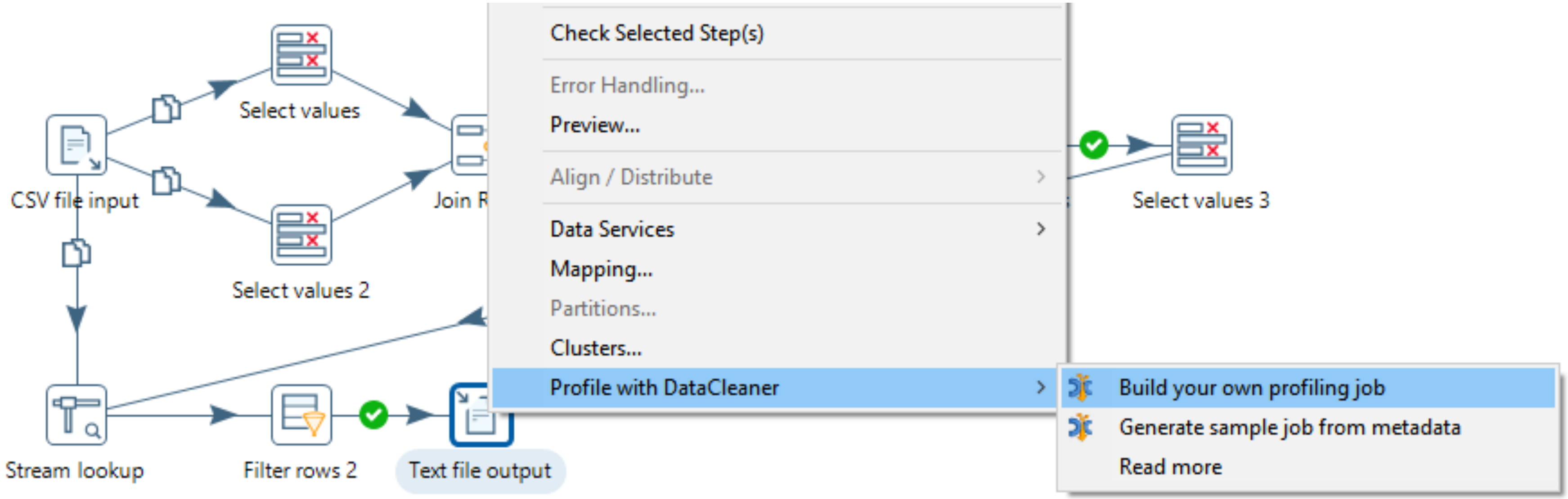


Rows of step: Text file output (64 rows)

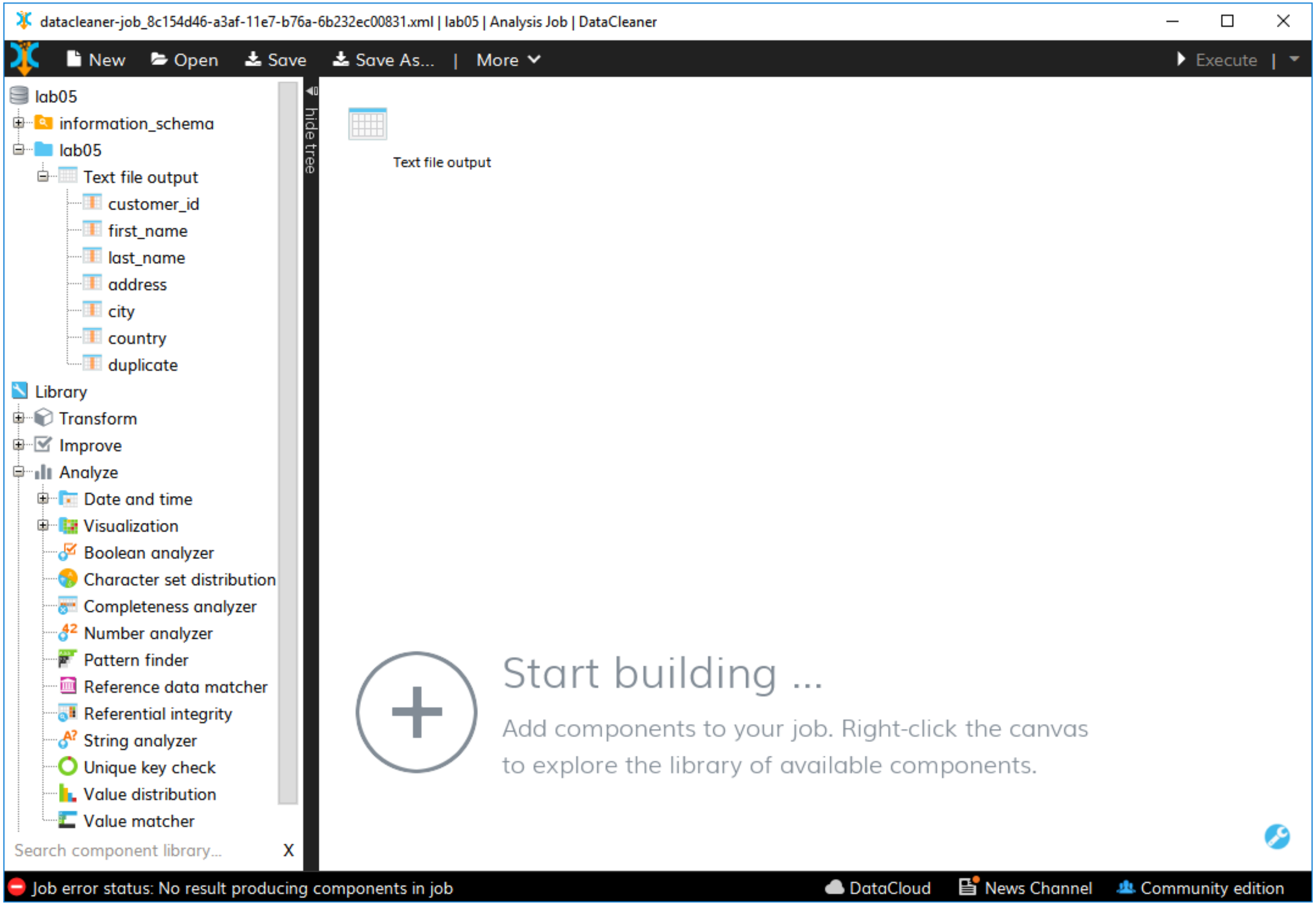
#	customer_id	first_name	last_name	address	city	country	duplicate
1	103	Carine	Schmitt	54, rue Royale	Nantes	France	<null>
2	119	Janine	Labrune	67, rue des Cinquante Otages	Nantes	France	<null>
3	121	Jonas	Bergulfsen	Erling Skakkes gate 78	Stavern	Norway	<null>
4	125	Zbygniew	Piestrzeniewicz	ul. Filtrowa 68	Warszawa	Poland	<null>
5	128	Roland	Keitel	Lyonerstr. 34	Frankfurt	Germany	<null>
6	141	Diego	Freyre	c/ Moralzarzal, 86	Madrid	Spain	<null>
7	144	Christina	Berglund	Berguvsvägen 8	Luleå	Sweden	<null>
8	145	Jytte	Petersen	Vinbæltet 34	Kobenhavn	Denmark	<null>
9	146	Mary	Saveley	2, rue du Commerce	Lyon	France	<null>

Data Profiling Integration

- Example
  - Perform data profiling after duplicate elimination



Data Profiling Integration



datacleaner-job_8c154d46-a3af-11e7-b76a-6b232ec00831.xml | lab05 | Analysis Job | DataCleaner

- lab05
  - information_schema
  - lab05
    - Text file output
      - customer_id
      - first_name
      - last_name
      - address
      - city
      - country
      - duplicate
- Library
  - Transform
  - Improve
  - Analyze
    - Date and time
    - Visualization
    - Boolean analyzer
    - Character set distribution
    - Completeness analyzer
    - Number analyzer
    - Pattern finder
    - Reference data matcher
    - Referential integrity
    - String analyzer
    - Unique key check
    - Value distribution
    - Value matcher

Search component library... X

Text file output

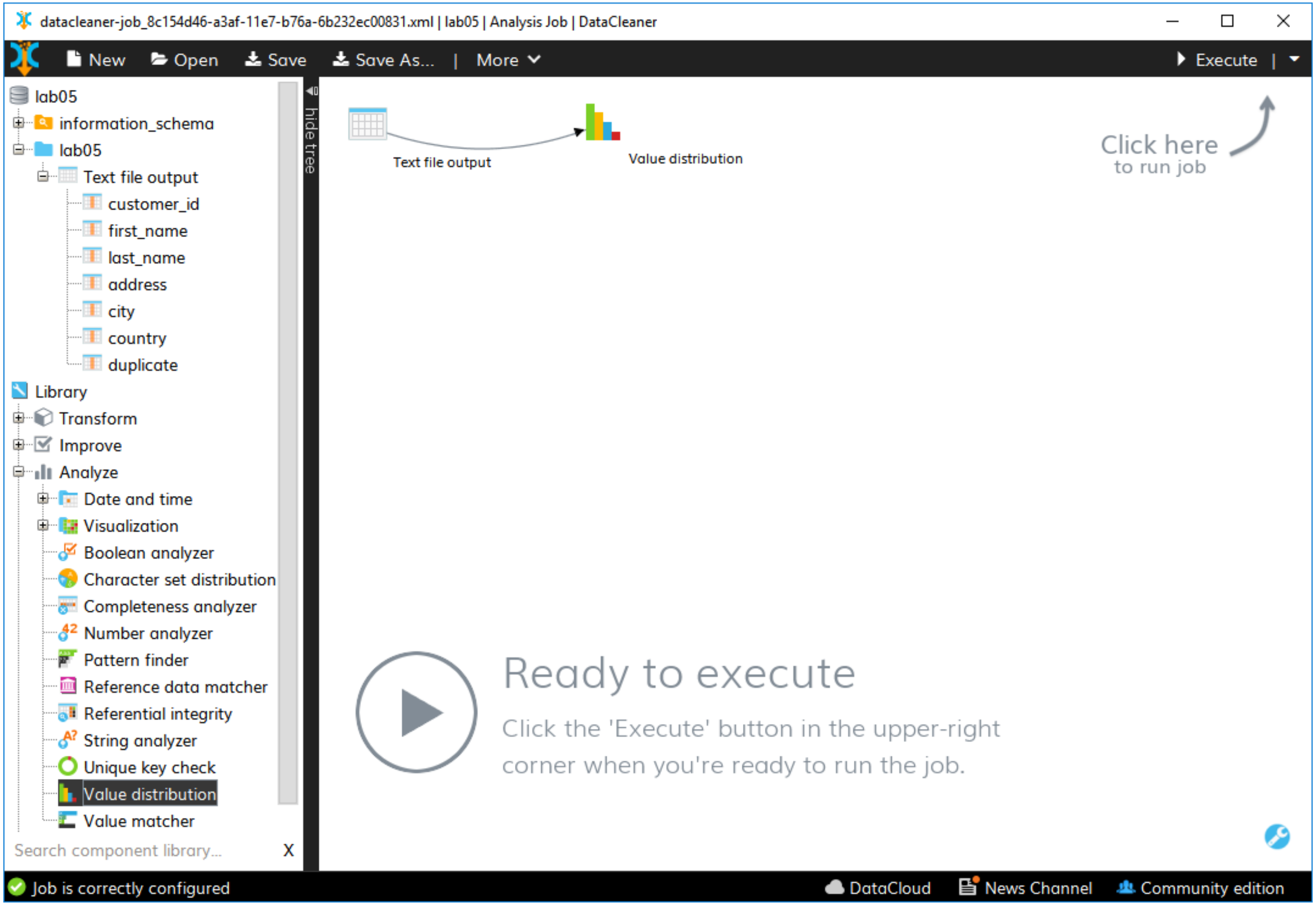
Start building ...

Add components to your job. Right-click the canvas to explore the library of available components.

Job error status: No result producing components in job

DataCloud News Channel Community edition

Data Profiling Integration



datacleaner-job_8c154d46-a3af-11e7-b76a-6b232ec00831.xml | lab05 | Analysis Job | DataCleaner

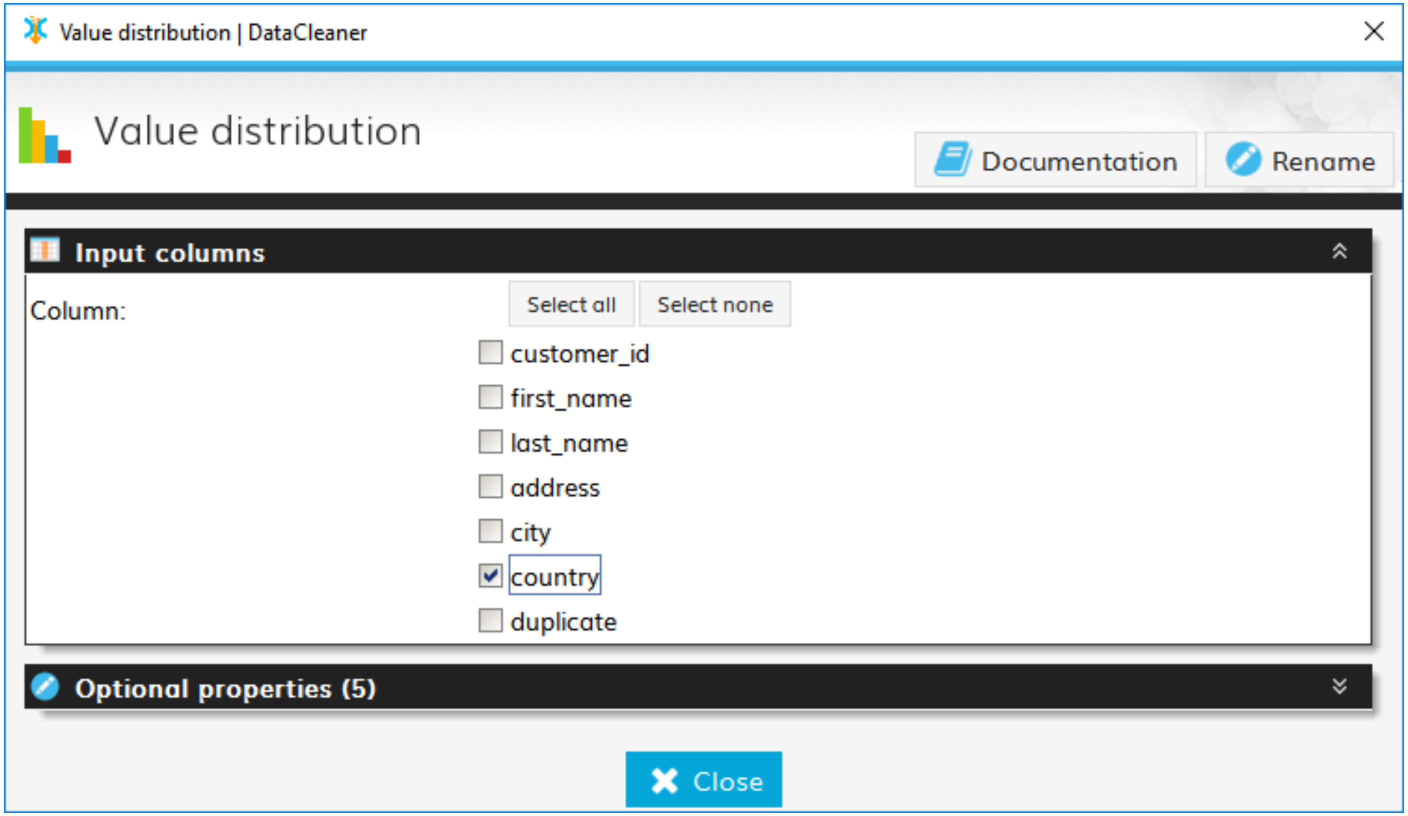
Ready to execute

Click the 'Execute' button in the upper-right corner when you're ready to run the job.

Job is correctly configured

Data Profiling Integration

- Value distribution



Value distribution | DataCleaner

Value distribution

Input columns

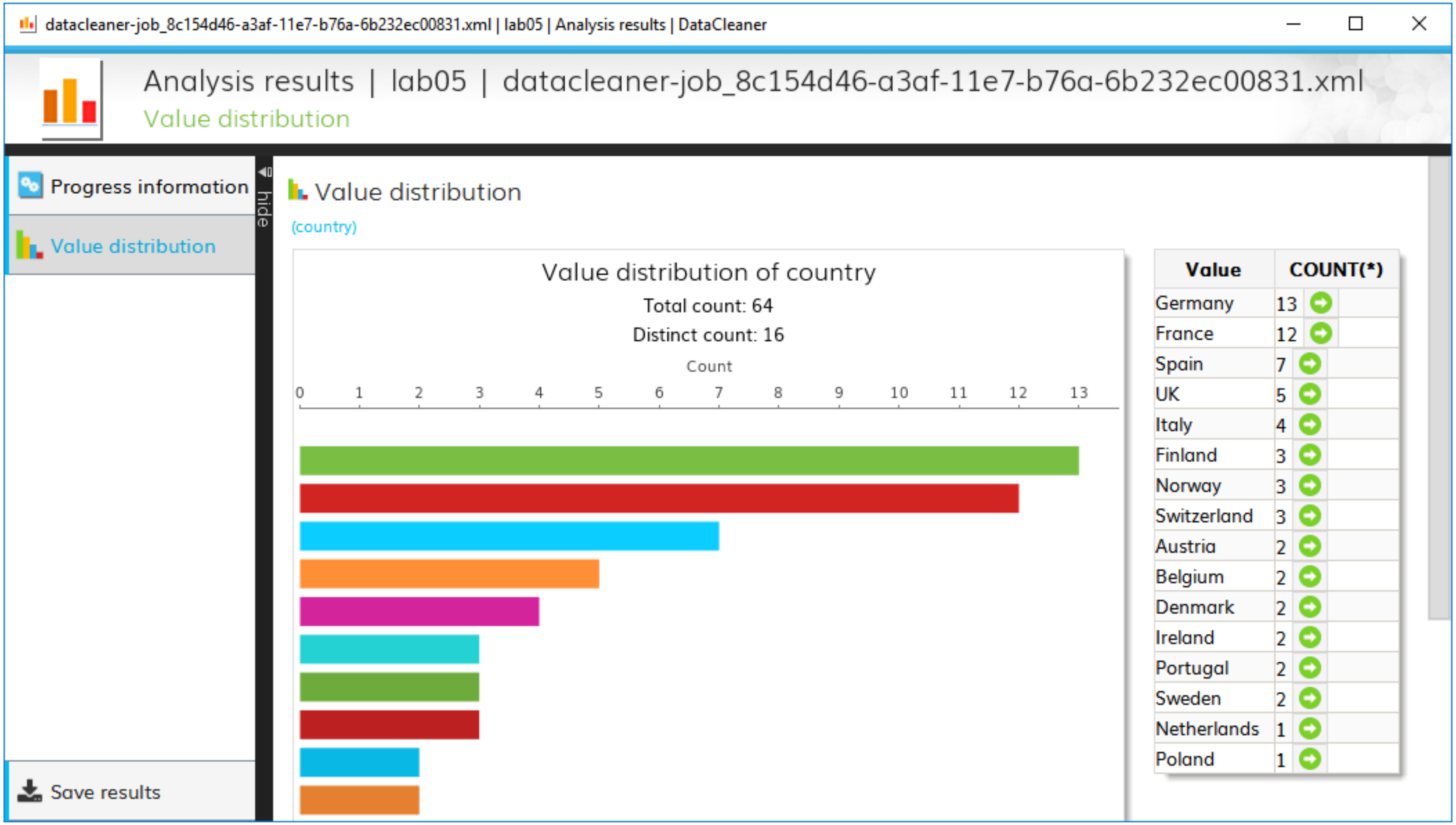
Column:

□ customer_id
□ first_name
□ last_name
□ address
□ city
☑ country
□ duplicate

Optional properties (5)

Data Profiling Integration

- Value distribution



datacleaner-job_8c154d46-a3af-11e7-b76a-6b232ec00831.xml | lab05 | Analysis results | DataCleaner
Analysis results | lab05 | datacleaner-job_8c154d46-a3af-11e7-b76a-6b232ec00831.xml
Value distribution
Progress information
Value distribution
Save results
hide
Value distribution
(country)
Count
0
1
2
3
4
5
6
7
8
9
10
11
12
13

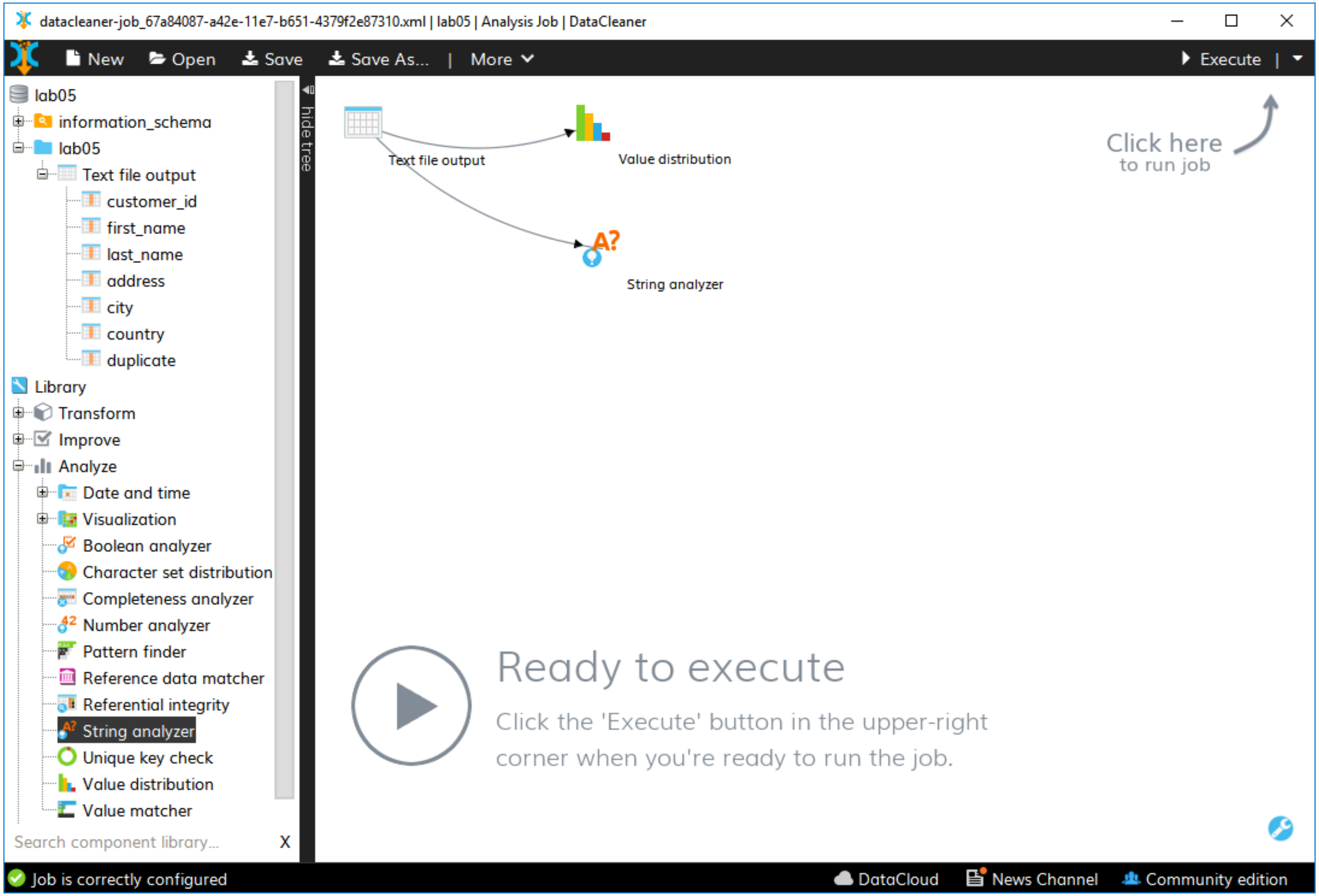
Value distribution of country

Total count: 64

Distinct count: 16

Value	COUNT(*)
Germany	13
France	12
Spain	7
UK	5
Italy	4
Finland	3
Norway	3
Switzerland	3
Austria	2
Belgium	2
Denmark	2
Ireland	2
Portugal	2
Sweden	2
Netherlands	1
Poland	1

Data Profiling Integration



datacleaner-job_67a84087-a42e-11e7-b651-4379f2e87310.xml | lab05 | Analysis Job | DataCleaner

- lab05
  - information_schema
  - lab05
    - Text file output
      - customer_id
      - first_name
      - last_name
      - address
      - city
      - country
      - duplicate
- Library
  - Transform
  - Improve
  - Analyze
    - Date and time
    - Visualization
    - Boolean analyzer
    - Character set distribution
    - Completeness analyzer
    - Number analyzer
    - Pattern finder
    - Reference data matcher
    - Referential integrity
    - String analyzer
    - Unique key check
    - Value distribution
    - Value matcher

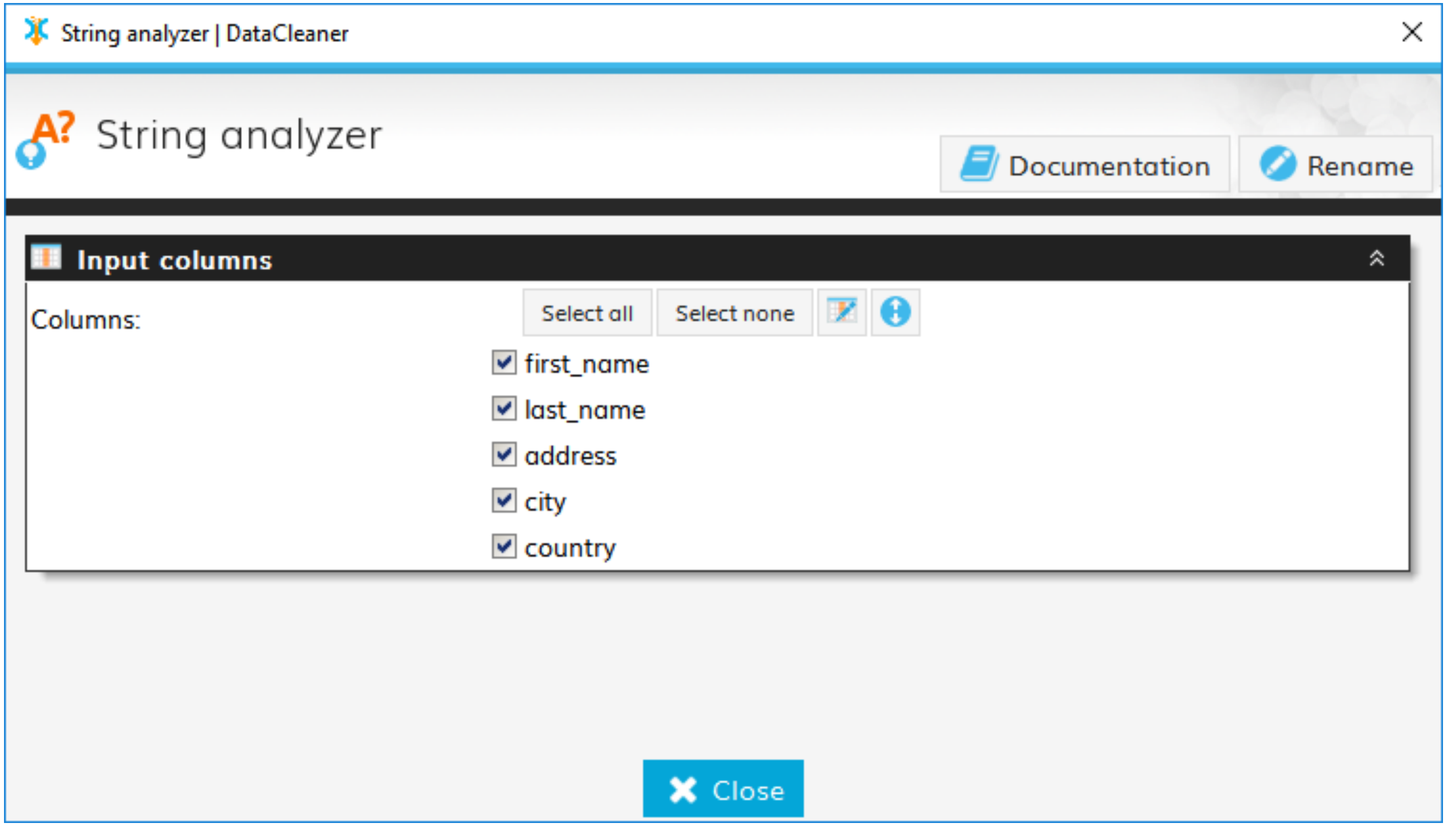
Ready to execute

Click the 'Execute' button in the upper-right corner when you're ready to run the job.

Job is correctly configured

Data Profiling Integration

- String analysis



String analyzer | DataCleaner

String analyzer

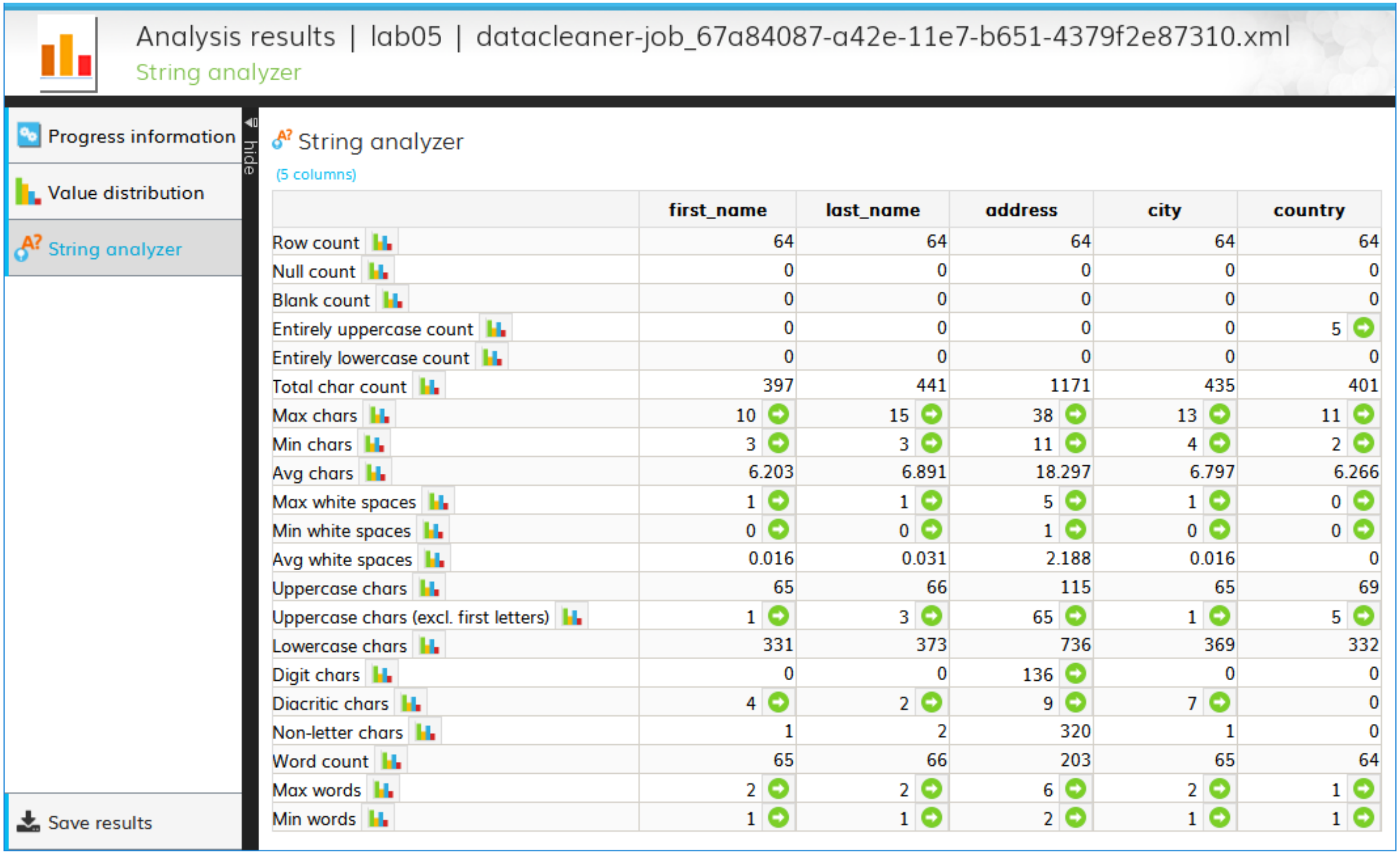
Input columns

Columns:

☑ first_name
☑ last_name
☑ address
☑ city
☑ country

Data Profiling Integration

- String analysis



Analysis results | lab05 | datacleaner-job_67a84087-a42e-11e7-b651-4379f2e87310.xml

String analyzer

String analyzer

(5 columns)

	first_name	last_name	address	city	country
Row count	64	64	64	64	64
Null count	0	0	0	0	0
Blank count	0	0	0	0	0
Entirely uppercase count	0	0	0	0	5
Entirely lowercase count	0	0	0	0	0
Total char count	397	441	1171	435	401
Max chars	10	15	38	13	11
Min chars	3	3	11	4	2
Avg chars	6.203	6.891	18.297	6.797	6.266
Max white spaces	1	1	5	1	0
Min white spaces	0	0	1	0	0
Avg white spaces	0.016	0.031	2.188	0.016	0
Uppercase chars	65	66	115	65	69
Uppercase chars (excl. first letters)	1	3	65	1	5
Lowercase chars	331	373	736	369	332
Digit chars	0	0	136	0	0
Diacritic chars	4	2	9	7	0
Non-letter chars	1	2	320	1	0
Word count	65	66	203	65	64
Max words	2	2	6	2	1
Min words	1	1	2	1	1