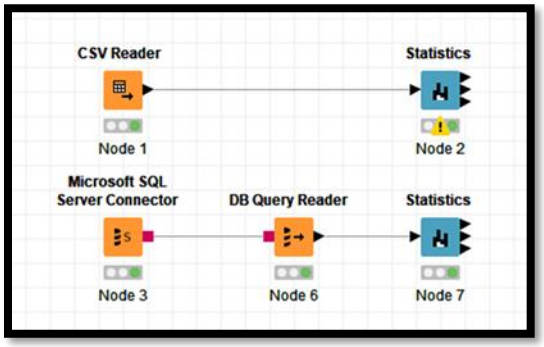
KNIME

Data analytics, profiling, reporting and integration platform



Links

- ✓ https://www.knime.com

License GNU [GPL-3.0]+ license
Version 4.1.2
Last Update 3/5/2020
OS Linux, macOS, Windows

System Requirements Sophisticated graphics hardware is not needed, multi core systems a plus as KNIME makes use of multiple cores. The available hard drive space (NOT main memory) limits the amount of processable data - several tens GB free space are recommended. Main memory should be 1GB or above, on 32-bit systems up to 1.5GB can be used, more on 64-bit systems

Description KNIME Analytics Platform is an open solution for data-driven innovation, designed for discovering the potential hidden in data, mining for fresh insights, or predicting new futures. Organizations can take their collaboration, productivity and performance to the next level with a robust range of commercial extensions to our open source platform.
KNIME Analytics Platform provides the tools to connect to a host of databases and data warehouses, access a variety of file formats (*formatted text files, binary files, SOAP and REST web services, databases, big data platforms, files from other proprietary software tools, and more*), retrieve data from cloud resources or external services, and more. The broad set of out-of-the-box functionality, allows you to seamlessly integrate and transform the data in one uniform, visual environment on your own - no dependencies on central IT. If there's a functionality you're missing, simply integrate the tools you like or take advantage of the many integrations we have with other open source projects.
Workflows created with KNIME Analytics Platform automatically document each step of your data wrangling process. Meaning, if you share workflows or results with your colleagues, they can easily understand the individual steps of your workflow and provide feedback.
KNIME Software covers all kinds of data analytics functionality - for example classification, regression, dimension reduction, or clustering, using advanced algorithms including deep learning, tree-based methods, and logistic regression. Among these, are integrations with other large, open source projects such as Keras or Tensorflow for deep learning, H2O for high performance machine learning, R and Python for coding, and various implementations for model interpretability and validation.
From integrations with Apache Spark for big data processing, to KNIME Server distributed executors for handling concurrent workflow execution, KNIME Software ensures data science is created and deployed quickly and efficiently.

Features

- Powerful Analytics
- Data and Tool Blending
- Over 1000 modules and growing
- Connectors for all major file formats and databases
- Supports multiple data types: XML, JSON, images, documents etc.
- Native and in-database data blending and transformation
- Math and statistical functions
- Advanced predictive and machine learning algorithms
- Workflow control
- Tool blending for Python, R, SQL, Java, Weka and others
- Interactive data views and reporting

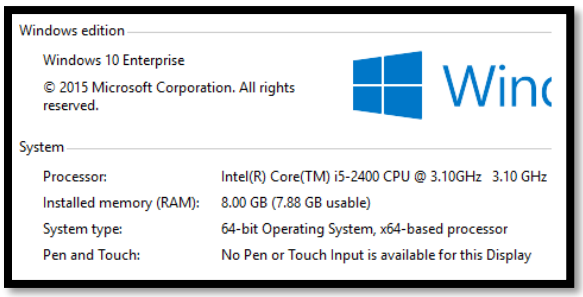
Connectivity / Supported Data Sources & Formats

- Simple text formats (CSV, PDF, XLS, JSON, XML, etc.)
- Unstructured data types (images, documents, networks, molecules, etc.)
- Time series data
- Connect to a host of databases and data warehouses to integrate data from Oracle, Microsoft SQL, Apache Hive, and more
- Load Avro, Parquet, or ORC files from HDFS, S3, or Azure
- Access and retrieve data from sources such as Twitter, AWS S3, Google Sheets, and Azure and extended via pandas

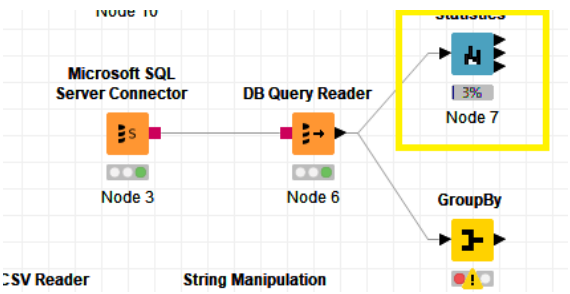
Limitations

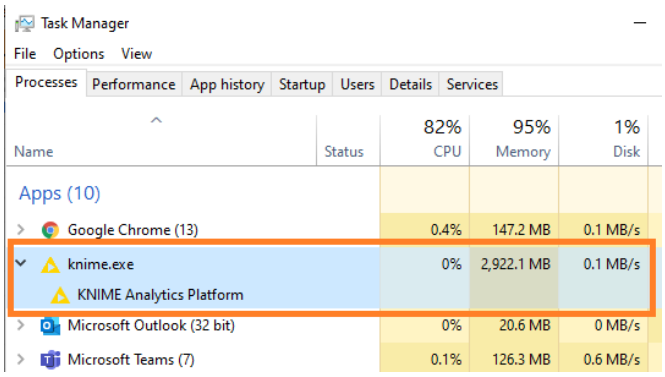
The number of rows is unlimited, but the number of columns shouldn't get much larger than ~10 000. In case all your columns have the same type, you may want to have a look at the "KNIME Nodes for Wide Data" extension. It allows to read all columns into a single "array" column.

Performance



KNIME works with 1.3M records but the performance is slow and takes good amount of memory from the system as shown in Pic 1 and 2. You will need a machine with appropriate capacity.



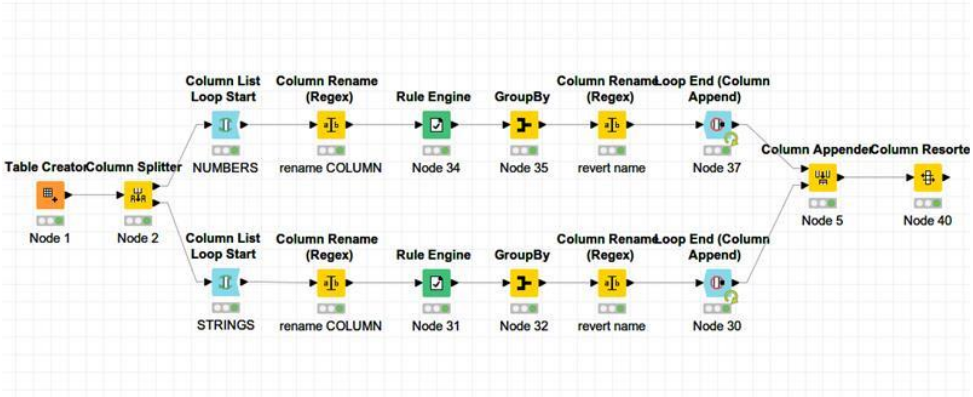


Feedback from Cystic Fibrosis Team:

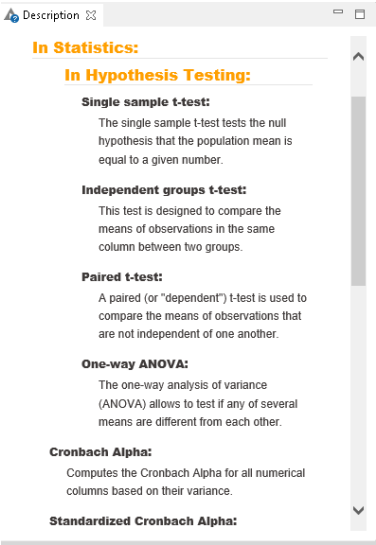
Kieran Earlam and Rebecca Cosgriff from the Cystic Fibrosis team used Knime to analyze data. Neither Kieran nor Rebecca have technical roles and despite this Kieran was able to get Knime working within 3 days. While they toiled valiantly to some degree of success, it should not and would not be that hard for the appropriate staff. If one considers what it takes to create a trusted high quality dataset, if the resources are unable to run data profiling tools which report data quality gaps, it seems unlikely they would be able to run other data quality tools and processes which contribute to creating a high quality dataset.

The feedback was as follows:

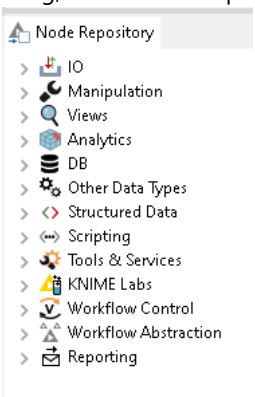
- KNIME is quiet resource intensive for system resources, it took a while to interrogate data with i5 Coffeelake processor overclocked at 4.2ghz and 16gb of ram.
- We store our verified data sets in .dta files for Stata and therefore had to convert to .csv for use in KNIME.



- Complex workflows could be created and run; however, it took substantial time to get to grips with the tool.
- The node repository seemed vast with detailed explanations on how to work a node into the workflow. However as previously mentioned complex workflows were hard to get to grips with and lots of time had to be spent debugging and learning the basics.



- The in-window explanations (show above) were very detailed and helpful, and given the days and weeks it would take to learn the tool I'm sure it would effectively interrogate most data.
- The huge number of extensions and functionality meant that when searching for solutions I was given multiple routes to achieve my goal, which can be both an advantage and a disadvantage
- KNIME had the ability to function with R coding, our stats team primarily use STATA.



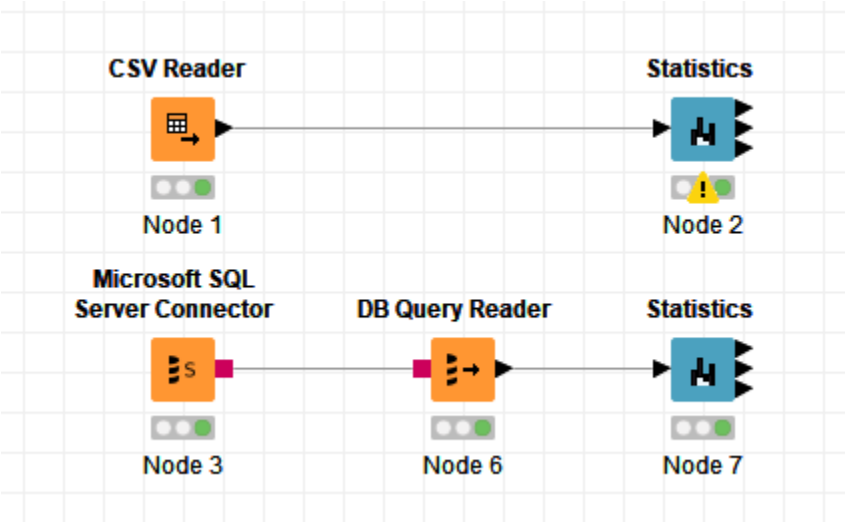
- The node repository was vast with 100s of tools to integrate into workflows. With adequate training KNIME looks like it could handle huge data sets.

The supporting document "**Data Quality Tool Assessment Request - Data Custodian Cystic Fibrosis.docx**" contains all feedback provided by the Cystic Fibrosis team.

KNIME DATA ANALYSIS

1. Download Knime from "https://www.knime.com/downloads/download-knime".
2. Install Knime and from "Workflow Coach" tool box drag and drop the "Microsoft SQL Server Connector" and configure with the "HostName" port and DB Names.
3. Right click and select "Execute".

4. Drag and drop the “DB Query Reader” and configure it by writing the SQL Queries and make a link to the MS SQL Server Connector. Right click and execute.
5. Drag and Drop the “Statistics” node and link to the DB Query reader and “Execute and Open Views”. Sample screen shot of this tool is below,



6. The file extension for a KNIME workflow, is .knwf (KNIME workflow file).
   Importing and exporting workflows are also introduced in this video: Import/Export Workflows (https://youtu.be/4GiwmM-qcC4).

PERCENTAGE OF REQUISITE INFORMATION AVAILABLE:

File

Numeric | Nominal | Top/bottom

Column	No. missings	Histogram
BIRTHDATE	0	
DEATHDATE	1,000	
SSN	0	
DRIVERS	213	

ROW COUNTS:

Statistics Table - 0:2 - Statistics

File Hilite Navigation View

Table "default" - Rows: 5 | Spec - Columns: 16 | Properties | Flow Variables

Row ID	S Column	D Min	D Max	D Mean	D Std. de...	D Variance	D Skewness	D Kurtosis	D Overall ...	I No. mis...	I No. NaNs	I No. +∞s	I No. -∞s	D Median	I Row co...	Histogram
ZIP	ZIP	1,001	2,861	2,009.645	451.871	204,187.831	-0.593	-0.077	1,262,057	543	0	0	0	2,119	1171	1.001
LAT	LAT	41.338	42.896	42.256	0.289	0.083	-0.656	0.16	49,482.214	0	0	0	0	42.305	1171	41
LON	LON	-73.382	-69.941	-71.36	0.63	0.397	-1.034	1.054	-83,562.341	0	0	0	0	-71.152	1171	-73
HEALTHCARE...	HEALTHCAR...	1,822.16	2,145,924.4	764,940.902	546,474.868	298,634,78...	0.164	-1.321	895,745,79...	0	0	0	0	718,679.1	1171	1.822
HEALTHCARE...	HEALTHCAR...	0	927,873.53	12,932.737	43,882.481	1,925,672,1...	12.854	211.86	15,144,235.59	0	0	0	0	5,741.8	1171	0

HIGHEST AND LOWEST VALUE OF KEY ELEMENTS:

Occurrences Table - 0:2 - Statistics

File

Table "default" - Rows: 1171 | Spec - Columns: 72 | Properties | Flow Variables

Columns: 72	Column Type	Column Index	Color Handler	Size Handler	Shape Han...	Filter Handler	Lower Bound	Upper Bound	Value 0	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6
RACE	String	33					?	?	white	native	asian	black	other	?	?
Count (RACE)	Number (int...	34					2	965	?	?	?	?	?	?	?
Relative Frequency (RACE)	Number (do...	35					0	1	?	?	?	?	?	?	?
ETHNICITY	String	36					?	?	hispanic	nonhispanic	?	?	?	?	?
Count (ETHNICITY)	Number (int...	37					113	1,058	?	?	?	?	?	?	?
Relative Frequency (ETHNICITY)	Number (do...	38					0	1	?	?	?	?	?	?	?
GENDER	String	39					?	?	M	F	?	?	?	?	?
Count (GENDER)	Number (int...	40					562	609	?	?	?	?	?	?	?
Relative Frequency (GENDER)	Number (do...	41					0	1	?	?	?	?	?	?	?
BIRTHPLACE	String	42					?	?	?	?	?	?	?	?	?
Count (BIRTHPLACE)	Number (int...	43					1	94	?	?	?	?	?	?	?
Relative Frequency (BIRTHPLACE)	Number (do...	44					0	1	?	?	?	?	?	?	?
ADDRESS	String	45					?	?	?	?	?	?	?	?	?
Count (ADDRESS)	Number (int...	46					1	1	?	?	?	?	?	?	?
Relative Frequency (ADDRESS)	Number (do...	47					0	1	?	?	?	?	?	?	?
CITY	String	48					?	?	?	?	?	?	?	?	?
Count (CITY)	Number (int...	49					1	116	?	?	?	?	?	?	?
Relative Frequency (CITY)	Number (do...	50					0	1	?	?	?	?	?	?	?
STATE	String	51					?	?	Massachusetts	?	?	?	?	?	?
Count (STATE)	Number (int...	52					1,171	1,171	?	?	?	?	?	?	?
Relative Frequency (STATE)	Number (do...	53					0	1	?	?	?	?	?	?	?
COUNTY	String	54					?	?	Hampden C...	Middlesex C...	Suffolk County	Plymouth Co...	Franklin Cou...	Bristol County	Norfolk Co
Count (COUNTY)	Number (int...	55					2	255	?	?	?	?	?	?	?
Relative Frequency (COUNTY)	Number (do...	56					0	1	?	?	?	?	?	?	?
ZIP	Number (int...	57					1,001	2,861	?	?	?	?	?	?	?
Count (ZIP)	Number (int...	58					1	543	?	?	?	?	?	?	?
Relative Frequency (ZIP)	Number (do...	59					0	1	?	?	?	?	?	?	?
LAT	Number (do...	60					41.338	42.896	?	?	?	?	?	?	?
Count (LAT)	Number (int...	61					1	1	?	?	?	?	?	?	?
Relative Frequency (LAT)	Number (do...	62					0	1	?	?	?	?	?	?	?
LON	Number (do...	63					-73.382	-69.941	?	?	?	?	?	?	?
Count (LON)	Number (int...	64					1	1	?	?	?	?	?	?	?
Relative Frequency (LON)	Number (do...	65					0	1	?	?	?	?	?	?	?
HEALTHCARE_EXPENSES	Number (do...	66					1,822.16	2,145,924.4	?	?	?	?	?	?	?
Count (HEALTHCARE_EXPENSES)	Number (int...	67					1	1	?	?	?	?	?	?	?
Relative Frequency (HEALTHCARE_EXPENSES)	Number (do...	68					0	1	?	?	?	?	?	?	?
HEALTHCARE_COVERAGE	Number (do...	69					0	927,873.53	?	?	?	?	?	?	?
Count (HEALTHCARE_COVERAGE)	Number (int...	70					1	28	?	?	?	?	?	?	?
Relative Frequency (HEALTHCARE_COVERAGE)	Number (do...	71					0	1	?	?	?	?	?	?	?

MISSING VALUES (UNUSABLE VALUES):

Nominal Histogram Table - 0:2 - Statistics

File Hilite Navigation View

Table "default" - Rows: 24 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	S Column	I No. mis...	Histogram
RACE	RACE	0	white, black, asian, native, other
COUNTY	COUNTY	0	Middlesex County, Worcester County, Suffolk County, Norfolk County, Essex County, Bristol County, Hampden County, Plymouth County, Barnstable County, Hampshire County, Berkshire County, Franklin County, Dukes County, Nantucket County
DEATHDATE	DEATHDATE	1000	
ZIP	ZIP	543	

REAL WORLD VALUES AND DUPLICATES:

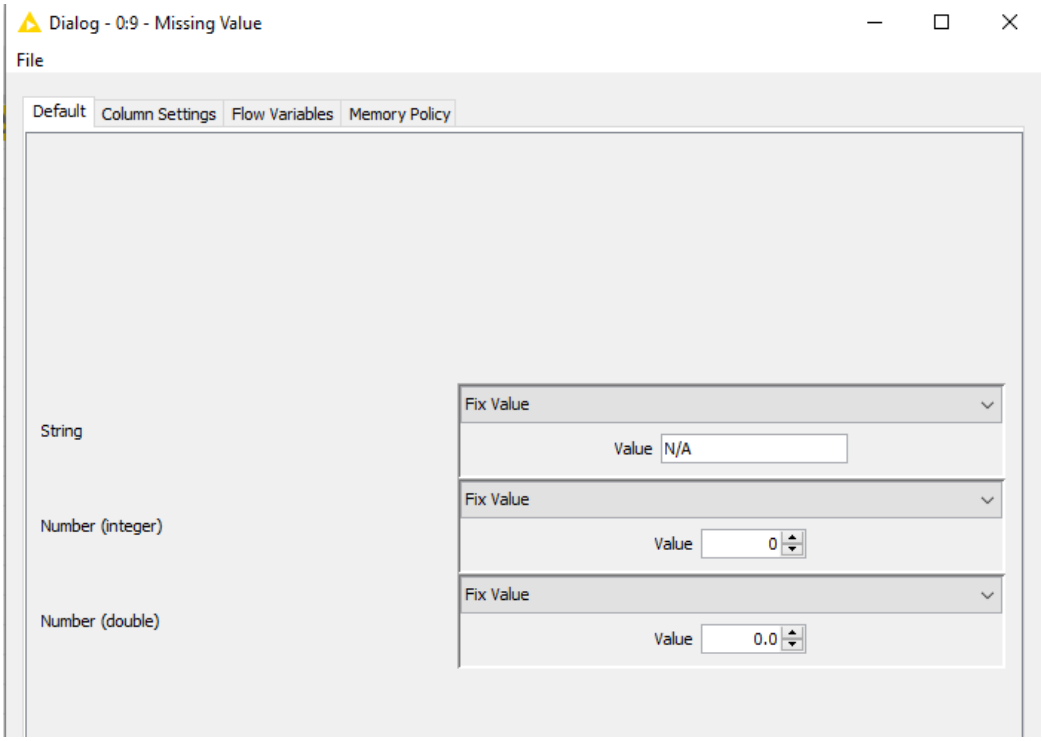
Group table - 0:8 - GroupBy

File Hilite Navigation View

Table "default" - Rows: 1171 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	S SSN	S FIRST	S LAST	S ADDRESS
Row0	999-10-1847	Mario764	Waters156	326 Ritchie Trafficway
Row1	999-10-3134	Reyna401	Shanahan202	824 Schuppe Gate U...
Row2	999-10-3892	Della552	Feil794	321 Wisozk Green S...
Row3	999-10-6031	Colby655	Gleichner915	408 Dicki Corner Uni...
Row4	999-10-8228	Dan465	Walker122	858 Prosacco Boulev...
Row5	999-10-8649	Larisa542	Fritsch593	1094 Herman Strave...
Row6	999-10-9322	Aubrey96	Lehner980	546 Mills Rue Suite 58
Row7	999-10-9858	Dean966	Heathcote539	674 Botsford Orchard
Row8	999-11-1235	Corinna386	Boehm581	915 Keeling Esplanade
Row9	999-11-1805	Boris111	Marks830	686 Buckridge Bypass
Row10	999-11-3069	Armand155	Schmeler639	663 O'Connell Wall
Row11	999-11-3190	Lindsy319	Cummings51	714 Lemke Crossing
Row12	999-11-3275	Brooks264	Weimann465	1033 Ledner Way
Row13	999-11-3522	Jon665	Zboncak558	914 Hyatt Loaf
Row14	999-11-5243	Oliver401	Davis923	861 Roberts Terrace...
Row15	999-11-5642	Elliott563	Towne435	551 Quigley Crossing

FILL MISSING VALUES:



MIN AND MAX DATES:

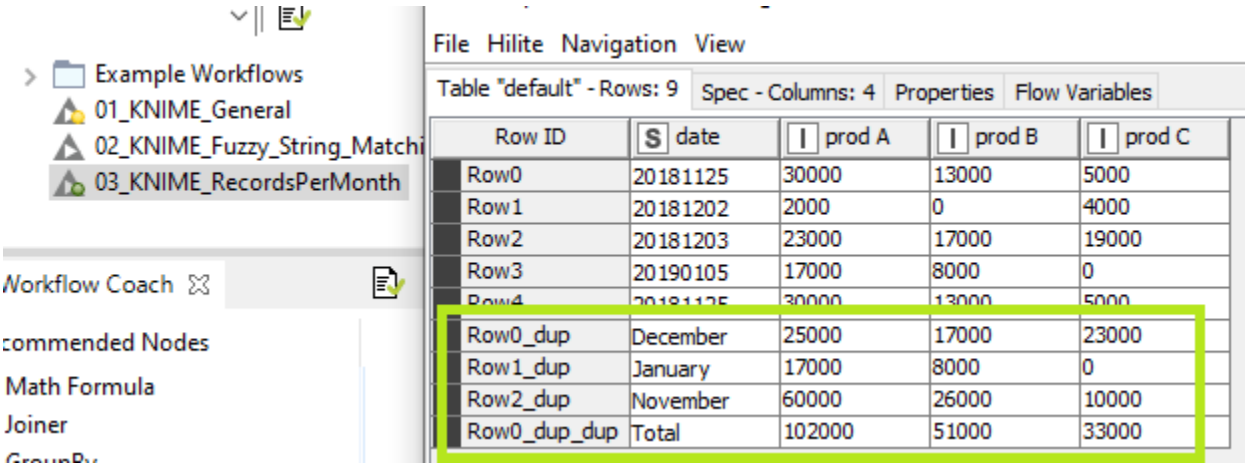
Group table - 0:12 - GroupBy

File Hilite Navigation View

Table "default" - Rows: 1171 | Spec - Columns: 3 | Properties | Flow Variables

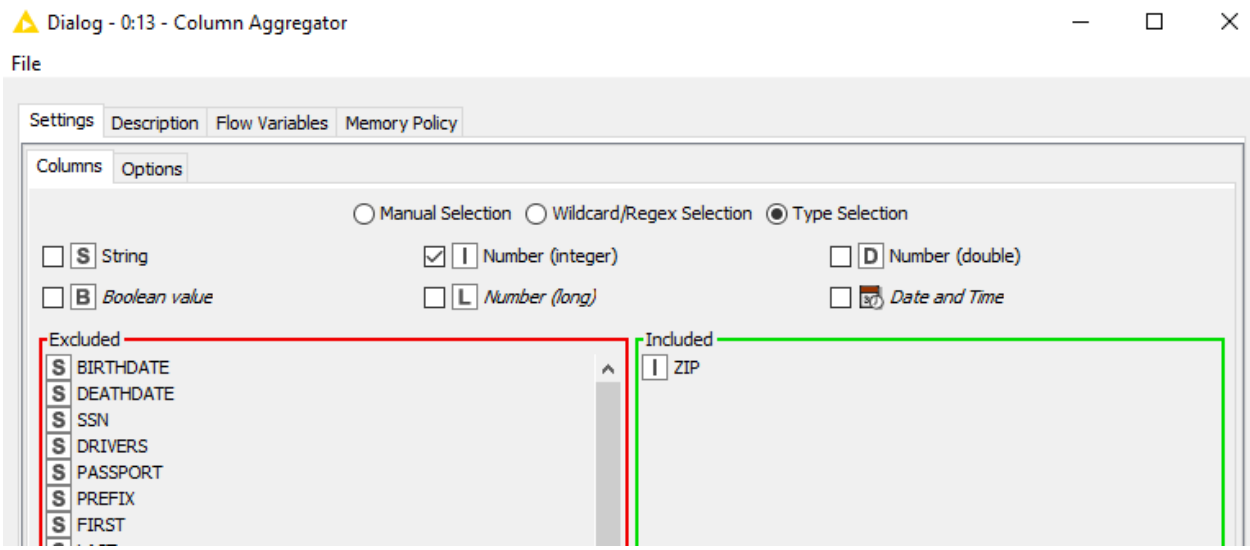
Row ID	S PATIENT	S Max(START)	S Min(START)
Row0	00185faa-2760-4218-9bf5-db301acf8274	2020-04-14T15:06:37Z	2010-11-09T15:06:37Z
Row1	0042862c-9889-4a2e-b782-fac1e540ecb4	2019-12-05T23:31:38Z	2010-05-06T23:31:38Z
Row2	0047123f-12e7-486c-82df-53b3a450e365	2020-03-19T00:15:54Z	2010-09-25T00:15:54Z
Row3	010d4a3a-2316-45ed-ae15-16f01c611674	2019-08-11T04:32:14Z	2010-06-20T04:32:14Z
Row4	01207ecd-9dff-4754-8887-4652eda231e2	2020-04-22T06:27:24Z	2019-05-15T06:27:24Z
Row5	0149d553-f571-4e99-867e-fcb9625d07c2	2019-06-27T13:13:20Z	1988-06-23T13:13:20Z
Row6	01e1f394-7219-4189-bceb-3cbd90cff90b	2019-07-26T14:39:24Z	2004-10-01T14:39:24Z
Row7	023a7d29-32b3-4db5-89c8-b88bd7582...	2020-03-03T13:04:22Z	1959-05-05T13:04:22Z
Row8	0288abb6-633c-40c3-ba0c-66c7d957727e	2019-11-06T04:07:36Z	1952-02-22T04:07:36Z
Row9	02ea2f1a-ddcf-4809-8279-dde7a62e0318	2019-04-11T12:49:12Z	2012-03-16T12:49:12Z
Row10	02f9aadd-72de-4b20-b381-f4c3b1cf7aa3	2005-10-01T20:06:53Z	1982-05-13T20:06:53Z
Row11	03172f6e-fb21-4770-8eef-513730174ab7	1968-01-23T09:16:22Z	1961-03-14T09:16:22Z
Row12	0325261f-61eb-46f8-acc6-89d15053fecd	2007-09-01T13:04:22Z	1926-05-04T13:04:22Z
Row13	034e9e3b-2def-4559-bb2a-7850888ae...	2018-01-29T17:45:28Z	2010-01-23T17:45:28Z
Row14	03612a7e-6460-4ef6-9528-59d60f970b93	2014-09-16T12:21:33Z	1957-08-03T12:21:33Z
Row15	03963166-b49f-4440-a80d-30abb90b4...	2019-08-25T10:41:23Z	1993-09-19T10:41:23Z
Row16	03c5e223-c016-4477-947f-22c691d6a62c	2020-03-17T15:05:51Z	2011-04-14T15:05:51Z
Row17	0447625b-b860-483c-9f30-17ed375b1493	2020-04-26T09:16:22Z	1968-09-06T09:16:22Z
Row18	04630e85-e9f5-4a9b-be75-97f2c3346037	2019-12-21T23:46:24Z	1958-03-13T23:46:24Z
Row19	04a29a39-c12f-480b-9521-f2d20559089f	2020-01-26T05:41:47Z	1973-12-16T05:41:47Z
Row20	04a849f4-1aaf-4112-a62f-d44df4325773	2020-02-28T20:13:17Z	2010-08-13T20:13:17Z
Row21	04db6603-0017-4cc6-a46d-6df577b0a10d	2020-04-13T08:54:26Z	1989-10-23T08:54:26Z
Row22	04dff6e5-123a-4c13-bd08-ad690d287173	2018-11-28T00:51:06Z	2010-05-04T00:51:06Z
Row23	0522e580-6775-49f5-b471-b39624280...	2016-04-27T08:39:57Z	1970-11-26T08:39:57Z
Row24	052c405b-ad28-4411-a81e-18cd811d1ca1	2018-09-24T04:56:03Z	2009-11-23T04:56:03Z

RECORDS PER MONTH TIMELINES:

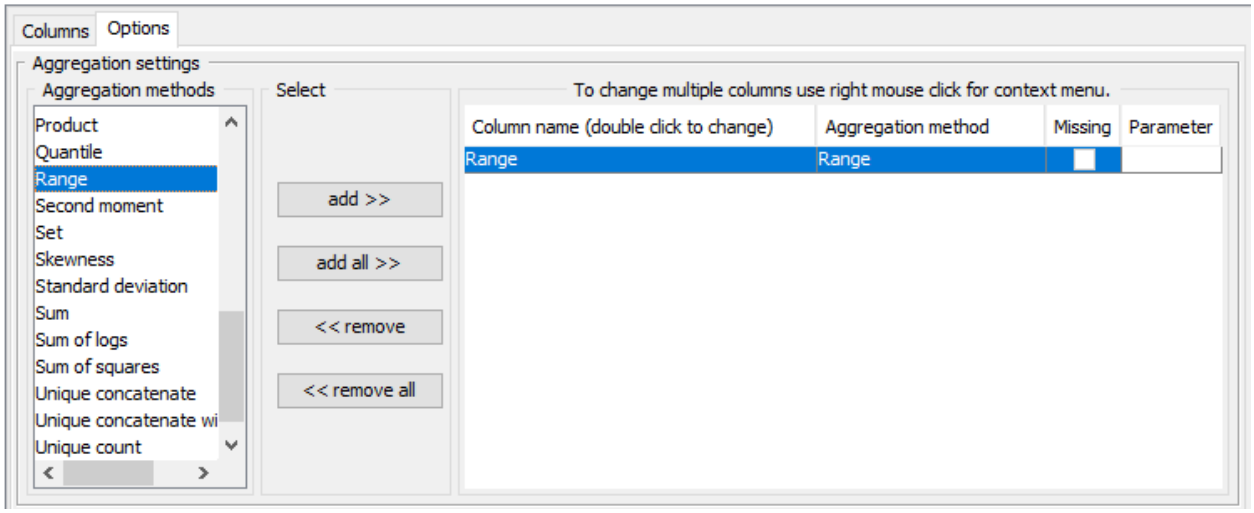


Row ID	S date	I prod A	I prod B	I prod C
Row0	20181125	30000	13000	5000
Row1	20181202	2000	0	4000
Row2	20181203	23000	17000	19000
Row3	20190105	17000	8000	0
Row4	20181125	30000	13000	5000
Row0_dup	December	25000	17000	23000
Row1_dup	January	17000	8000	0
Row2_dup	November	60000	26000	10000
Row0_dup_dup	Total	102000	51000	33000

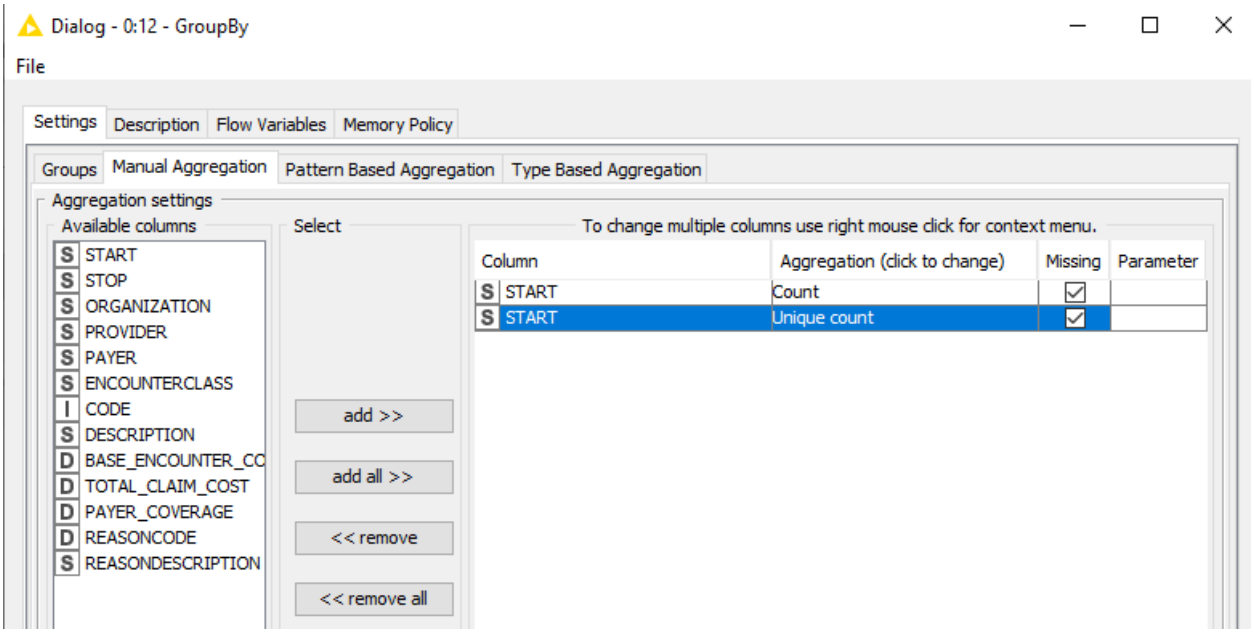
COLUMN FORMATS AND RANGE:



RANGE:



COUNTS OF REAL WORLD AND DUPLICATES:



OUTLIERS:

Robust Statistics - 0:14 - Box Plot (local)

File Hilite Navigation View

Table "default" - Rows: 7 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	D CODE	D BASE_E...	D TOTAL...	D PAYER...	D REASO...
Minimum	22,298,006	77.49	77.49	0	6,072,007
Smallest	22,298,006	129.16	129.16	0	6,072,007
Lower Quartile	162,673,000	129.16	129.16	17.49	55,822,004
Median	185,347,001	129.16	129.16	69.16	72,892,002
Upper Quartile	390,906,007	129.16	129.16	89.16	195,967,001
Largest	702,927,004	129.16	129.16	129.16	403,191,005
Maximum	702,927,004	129.16	129.16	129.16	124,171,00...

PERCENTAGE OF COLUMNS:

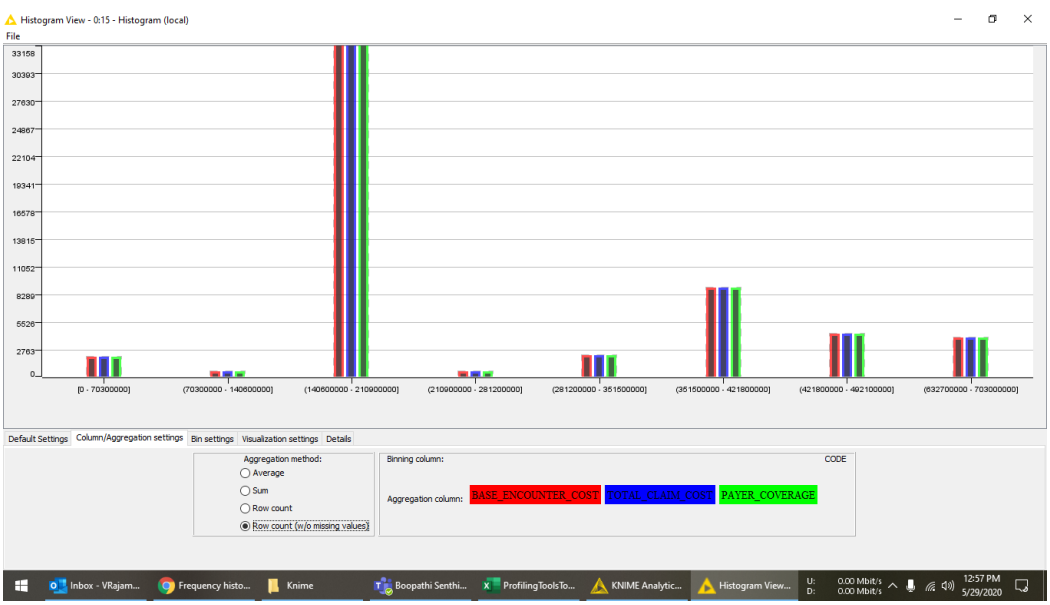
Transposed Table - 2:32 - Transpose

File Hilite Navigation View

Table "default" - Rows: 4 | Spec - Columns: 3 | Properties | Flow Varia

Row ID	? column1	? column2	? column3
Row0	a	b	c
Sum	10	10	10
Max	6	4	3
Percentage	60.0	40.0	30.0

HISTOGRAM SAMPLES:



LIMITATIONS:

- Works for 1.3M records, there is a performance drop and it can able to handle it by taking more system memory. A system with more memory will be recommended for KNIME while operating with huge datasets. Example shown below,