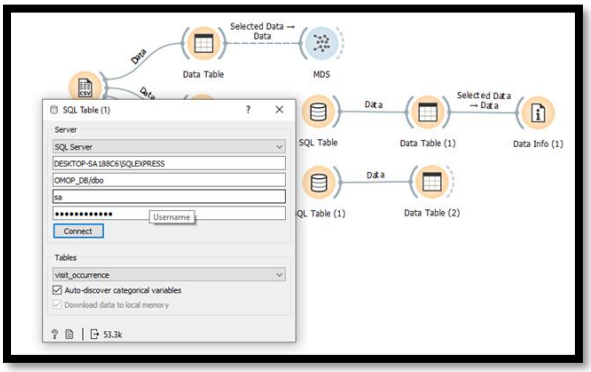
ORANGE

Linux, MacOS - preferred

Data visualization, machine learning, data profiling and mining toolkit



Links

- ✓ http://orange.biolab.si/
- ✓ https://orange.biolab.si/getting-started/
- ✓ https://orange-data-mining-library.readthedocs.io/en/latest/index.html
- ✓ https://www.javatpoint.com/orange-data-mining

License GNU [GPL-3.0]+ license
Version 3.24.1
Last Update 3/5/2020
OS Linux (preferred), macOS, Windows (not recommended)

Description Open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox. Orange consists of a canvas interface onto which the user places widgets and creates a data analysis workflow. Widgets offer basic functionalities such as reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualizing data elements, etc. The user can interactively explore visualizations or feed the selected subset into other widgets.
Data mining is done through visual programming or Python scripting. Python scrips can run in a terminal window, integrated environments like PyCharm and PythonWin, or shells like iPython.
The tool has components for machine learning, add-ons for bioinformatics and text mining and it is packed with features for data analytics.

Orange consists of a canvas interface onto which the user places widgets and creates a data analysis workflow. Widgets off basic functionalities such as reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualizing data elements, etc. The user can interactively explore visualizations or feed the selected subset into other widgets.
In Orange, data analysis process can be designed through visual programming. Orange remembers the choices, suggests most frequently used combinations. Orange has features for different visualization, such as scatterplots, bar charts, trees, to dendorograms, networks and heatmaps. There are over 100 widgets for standard data analysis and specialized add-ons for Bioorange for bioinformatics.
Core Orange supports Excel, comma- and tab-delimited files (.xlsx, .csv, .tab). It also reads online data, such as Google Spreadsheets. SQL widget supports PostgreSQL and MSSQL databases. Add-ons can load additional formats. For example, Orange3-ImageAnalytics add-on can import images (.jpg, .tiff, .png) and Orange3-Text add-on can import text files (.txt, .docx, .pdf).

Features

- Canvas: graphical front-end for data analysis
- Widgets:
  - Data: widgets for data input, data filtering, sampling, imputation, feature manipulation and feature selection
  - Visualize: widgets for common visualization (box plot, histograms, scatter plot) and multivariate visualization (mosaic display, sieve diagram).
  - Classify: a set of supervised machine learning algorithms for classification
  - Regression: a set of supervised machine learning algorithms for regression
  - Evaluate: cross-validation, sampling-based procedures, reliability estimation and scoring of prediction methods
  - Unsupervised: unsupervised learning algorithms for clustering (k-means, hierarchical clustering) and data projection techniques (multidimensional scaling, principal component analysis, correspondence analysis).
- Add-ons:
  - Associate: widgets for mining frequent item sets and association rule learning
  - Bioinformatics: widgets for gene set analysis, enrichment, and access to pathway libraries
  - Data fusion: widgets for fusing different data sets, collective matrix factorization, and exploration of latent factors
  - Educational: widgets for teaching machine learning concepts, such as k-means clustering, polynomial regression, stochastic gradient descent
  - Geo: widgets for working with geospatial data
  - Image analytics: widgets for working with images and ImageNet embeddings
  - Network: widgets for graph and network analysis
  - Text mining: widgets for natural language processing and text mining
  - Time series: widgets for time series analysis and modeling
  - Spectroscopy: widgets for analyzing and visualization of (hyper)spectral datasets

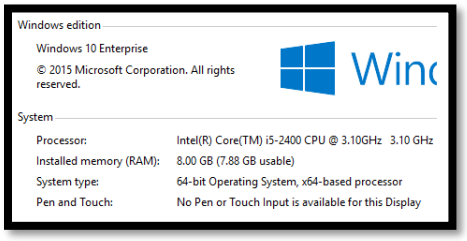
Connectivity / Supported Data Sources & Formats

- Excel (.xlsx), simple tab-delimited (.txt), comma-separated files (.csv) or Google Sheets document
- distance matrix: Distance File
- predictive model: Load Model
- network: Network File from Network add-on
- images: Import Images from Image Analytics add-on
- several spectroscopy files: Multifile from Spectroscopy add-on
- PostgreSQL, SQL, online repository, and extended via pandas

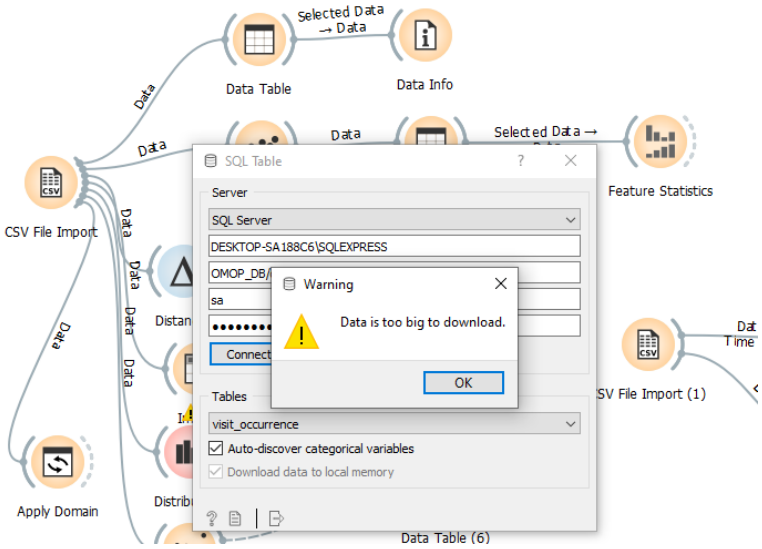
Limitations

Although document indicates that Orange supports Windows it was built predominantly for Linux and Mac OS using PostgreSQL. Orange doesn't support connection to MySQL.
For large datasets, Orange requires an additional module, quantile, in order to ensure computation is done on the database. This is in order to avoid a "data too big to download" error.

Performance



As indicated in limitations Orange does not work with this volume of data (1.3 Million records) and shows a warning as shown in Pic and no data is loaded. For this to work, an additional module, quantile, needs to be used in order to compute on the database, alternatively a machine with more memory and CPU capacity should be used.



Feedback from Neonatal Team:

Victor L Banda, a Data Analyst from the Neonatal Data Analysis Unit, Neonatal Medicine Research Group, Imperial College London, Chelsea and Westminster Hospital provided the following feedback:

Orange could not directly interact with the MySQL database, data migration was done, first to MS SQL Server. Since data was too big to download to local memory on an SQL Server instance on SQL Table widget, we later migrated to postgres SQL as it had support to uncheck the "download to local memory" option. However to adequately interact with the Postgres database, an installation of two modules: quantile and sm_system_time was required. Installing these two on a windows machine is not fully supported documentation wise, when compared to Mac OS and Linux. We failed to setup a connection of the pgxn client to both standalone orange or the anaconda instance. I believe this exercise would have been best carried via Linux machine. We could set up a virtual machine if we are given additional time for this. If need be.

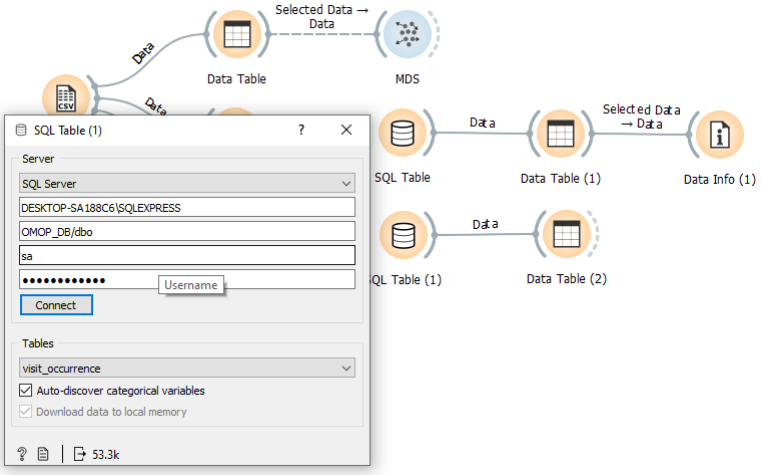
With these challenges, we still connected the NNRD but through the MS SQL instance using pymssql.

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

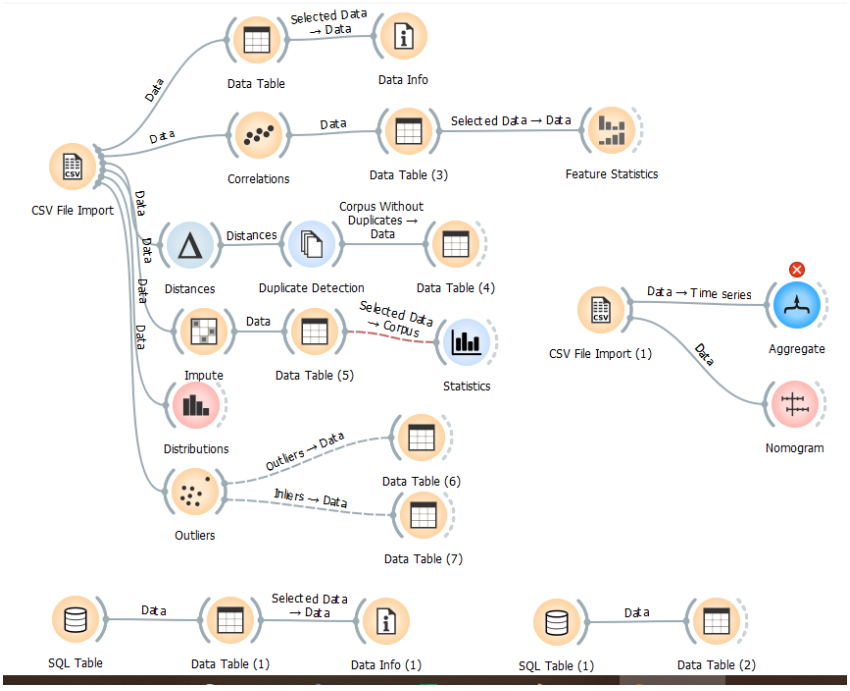
ORANGE DATA ANALYSIS

1. Download Orange from "https://orange.biolab.si/download/#windows"
2. Install Orange and from the tools, drag and drop the "SQL Table" tool.
3. Give the required information like Hostname, username and password.
4. Press the connect button and verify the connection is successful.
5. Drag and drop the "Data Table" and make a link from "SQL Table" to "Data Table".

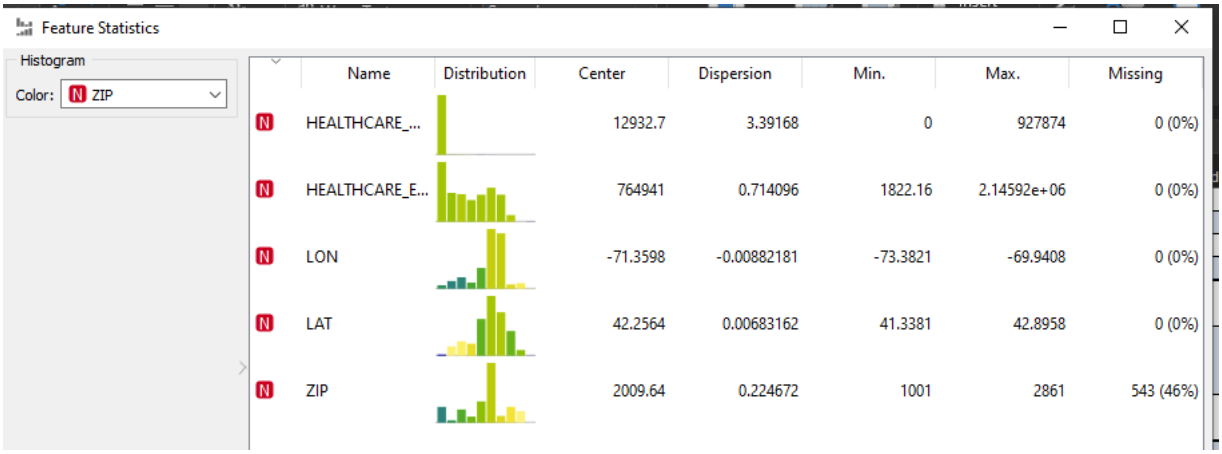
6. View the data and the statistics, sample widgets shown below,



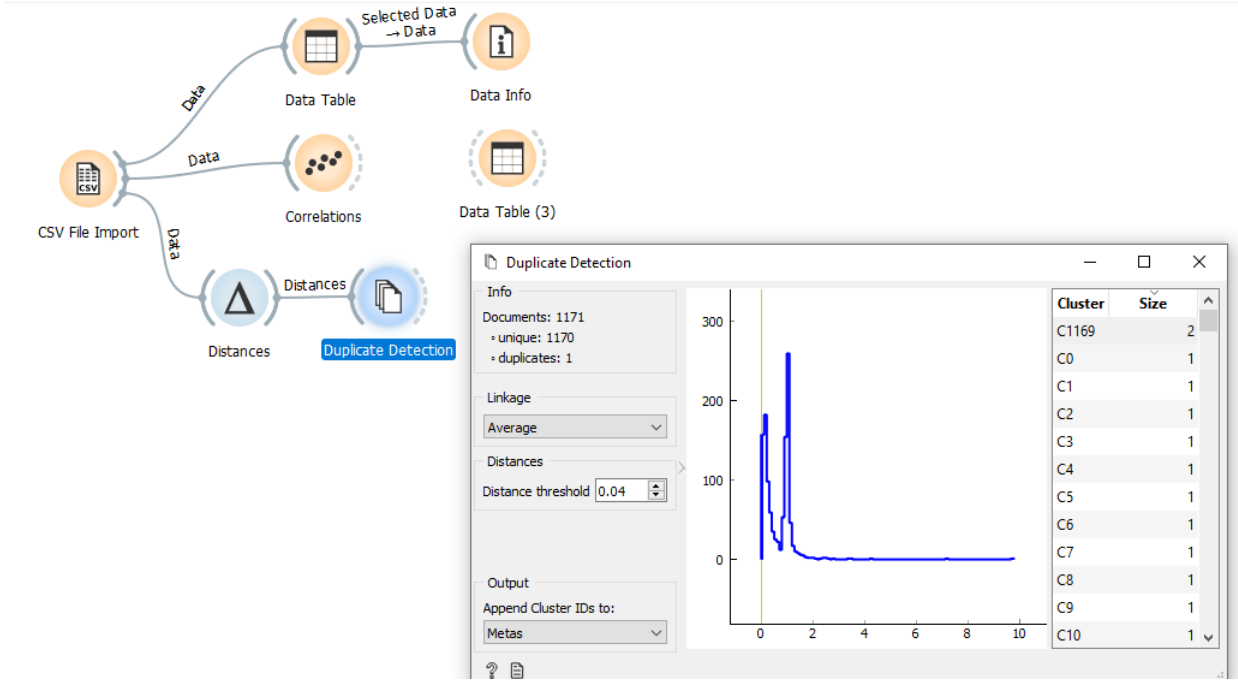
SAMPLE ORANGE WIDGETS:



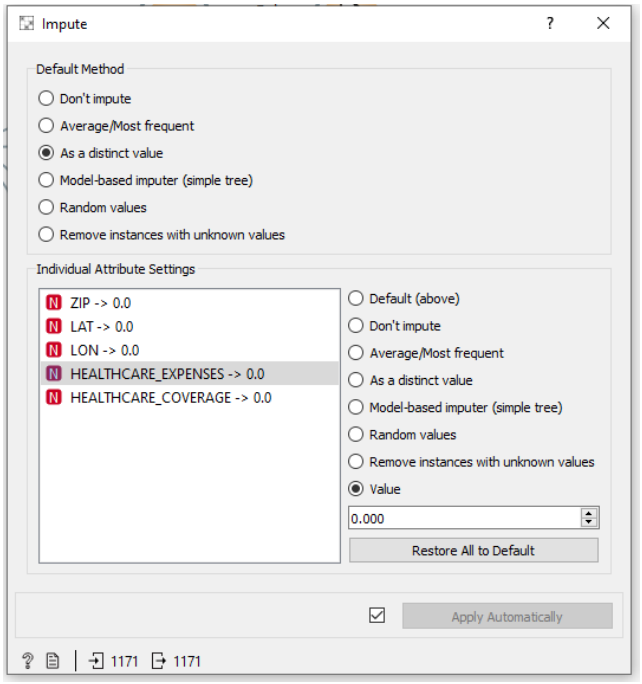
MIN AND MAX VALUES:



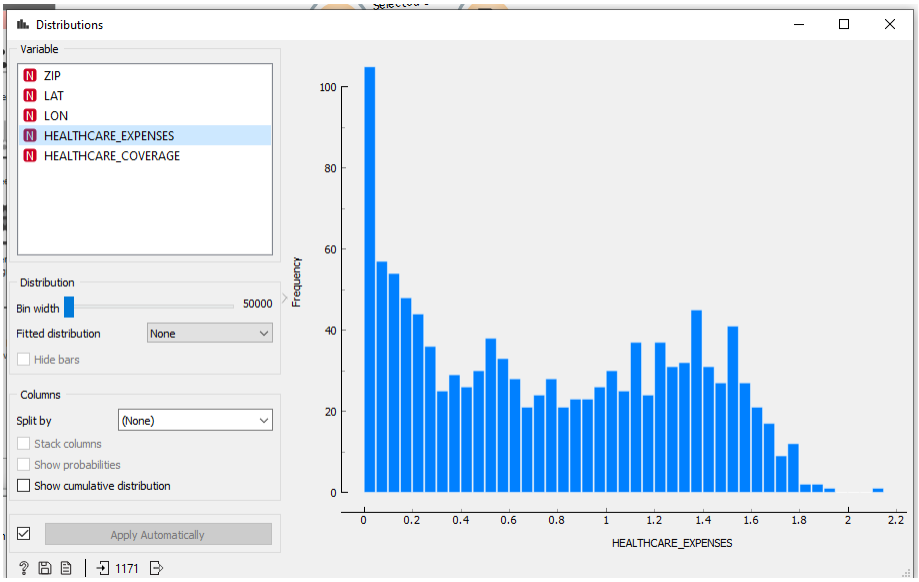
DUPLICATE VALUES ANALYSIS:



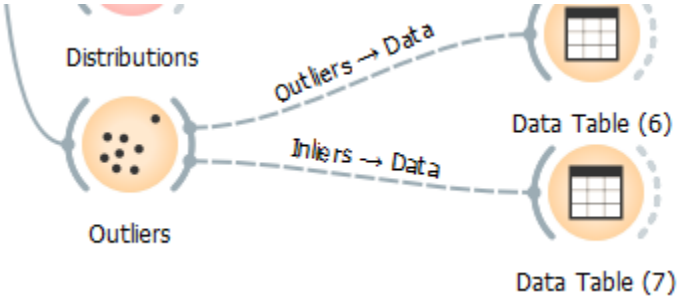
FILL MISSING VALUES:



DISTRIBUTIONS:



OUTLIERS:



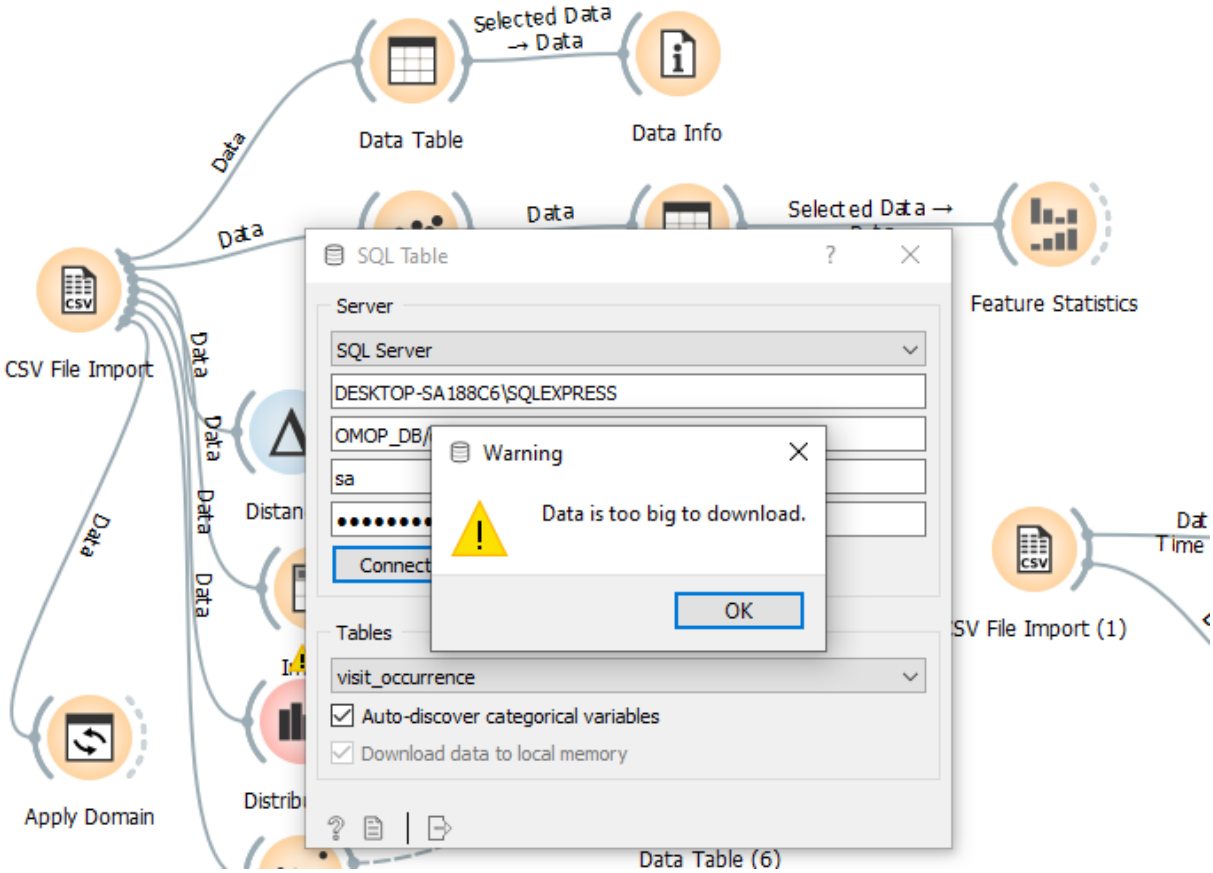
PERCENTAGE OF REQUISITE INFORMATION AVAILABLE:

File
Numeric | Nominal | Top/bottom

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

LIMITATIONS:

- Orange can handle up to 500K records, when tried with 1.3M records the tool was not able to load and process and freezes as shown below,



How to load .ows files:

Load the .ows file to the Orange IDE, from “File” menu, select “Open” menu item and select the *.ows file located.