Lab 5 - Data Wrangling using Knime (Basics)

Table of Contents

Learning Outcome

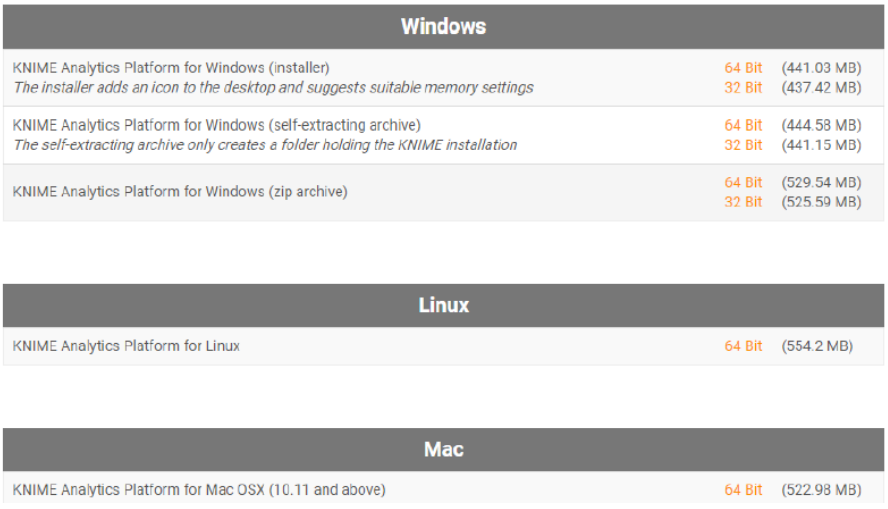
At the end of this lab, you will be able to clean your dataset with the data processing nodes in Knime Analytics Platform.

Task 1: Download and Install Knime

1. Go to the download page on the KNIME.com website to start installing KNIME Analytics Platform.

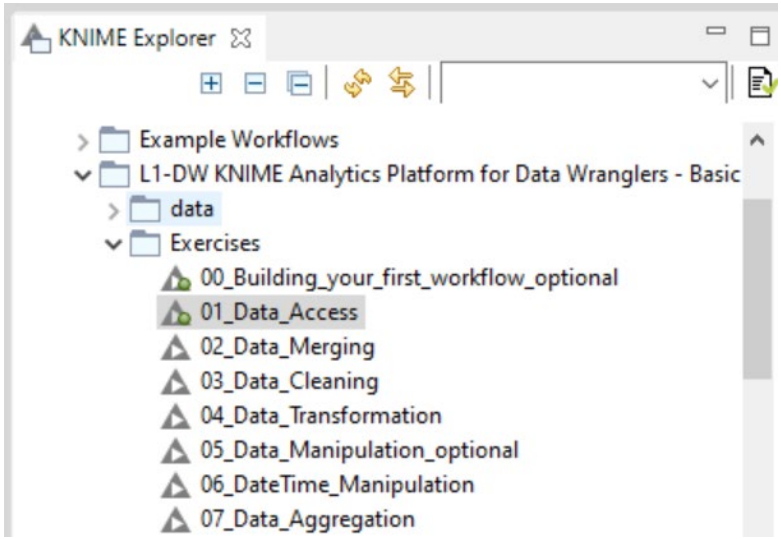
2. The download page shows three tabs which can be opened individually:
   - *Register for Help and Updates*: here you can optionally provide some personal
   - information and sign up to our mailing list to receive the latest KNIME news
   - *Download KNIME*: this is where you can download the software
   - *Getting Started*: this tab gives you information and links about what you can do after you have installed KNIME Analytics Platform
3. Now open the Download KNIME tab and click the installation option that fits your operating system. KNIME Analytics Platform can be installed on Windows, Linux, or macOS.

   Notes on the different options for Windows:
   - *The Windows installer extracts the compressed installation folder, adds an icon to your desktop, and suggests suitable memory settings.*
   - *The self-extracting archive simply creates a folder containing the KNIME*
   - *installation files. You don't need any software to manage archiving.*
   - *The zip archive can be downloaded, saved, and extracted in your preferred*
   - *location on a system to which you have full access rights.*



Task 2: Import Knime Archive File

Download and import L1-DW_KNIME_Analytics_Platform_for_Data_Wranglers_-_Basics_2021-07 into your Knime workspace



Task 3: Data Access

Open the workflow 01_Data_Access and read the following data files:

1. Customer information
   - CustomerInfoSystem1.csv
   - CustomerInfoSystem2.table
2. Online shop transactions, and product number & price information
   - TransactionOnline from Transations.sqlite
   - ProductNrAndPrice from Transations.sqlite
3. Store transactions and information
   - Store.xlsx
   - TransactionsStore.table
4. Try to use workflow relative-paths

Task 4: Data Merging

Open the workflow 02_Data_Merging:

1. **Concatenate** the customer information from the two systems
2. Add the price information to each online product purchase (DB Joiner) and read the table into KNIME (DB Reader)
3. Add the location information to each purchase in a store based on the StoreID (Joiner node)

4. Create three metanodes to clean up your workflow
   - Customer data
   - Online transactions & product+price (two output ports)
   - Onsite purchases in stores

Task 5: Data Cleaning

Open the workflow 03_Data_Cleaning:

1. Explore the data using the Data Explorer node
2. Replace numeric outliers in the "Age" column with missing values
3. Correct the spelling mistakes in the "Country" column
   - Extract the values with spelling mistakes
   - Manually define the correct spelling for the lookup table
   - Optional: Create the lookup table automatically, using a similarity search

4. If the age of a customer is missing, replace the birthday with a missing value
   Hint: Use the expression NOT MISSING Age => $Birthday$
5. Impute the missing values in the age column with the column mean
6. Remove rows for duplicate CustomerIDs

Task 6: Data Transformation

1. Change the structure of the table with the onsite purchases so that each purchased product is in a separate row and not the whole purchase event
   - Unpivot the columns that show the products ordered in one purchase event. Retain other columns in the table.
   - Remove rows that have missing values
   - Rename the "ColumnValues" column to "ProductNr" and "ShoppingNumber" to "OrderNumber" and remove unnecessary columns
2. Optional:
   - Standardize the Product Numbers

Task 7: Data Manipulation

1. Add the price to the onsite product purchase data
2. Add transaction types to each product purchase
   - "Store - no CC" if the customer ID is not available in the onsite transaction
   - "Store - CC" if the customer ID is available in the onsite transaction
   - "OnlineStore" for the orders coming from the online store

3. Concatenate the data of online and onsite purchases
4. Add the customer information to each transaction

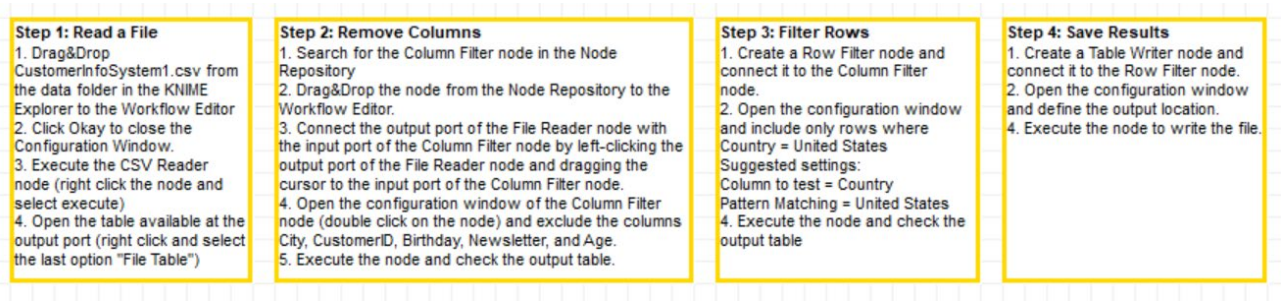
Task 8: Data Time Manipulation

1. Convert order dates from string to Date&Time
2. Extract the product purchases that were submitted in 2019
3. Extract the remaining product purchases into a separate table
4. Extract quarter and year of each product purchase into separate columns

Task 9: Data Aggregation

1. Calculate the total purchase amount by a customer ID both in 2019 and earlier
2. Calculate the total purchase amount by quarter and transaction type
3. Calculate the numbers of orders by basket size and transaction type (optional)
4. Convert the dates of births of the customers to Date&Time and extract the birth year into a separate column (optional)

Task 10: Building Your First Workflow (Optional)



~The End~