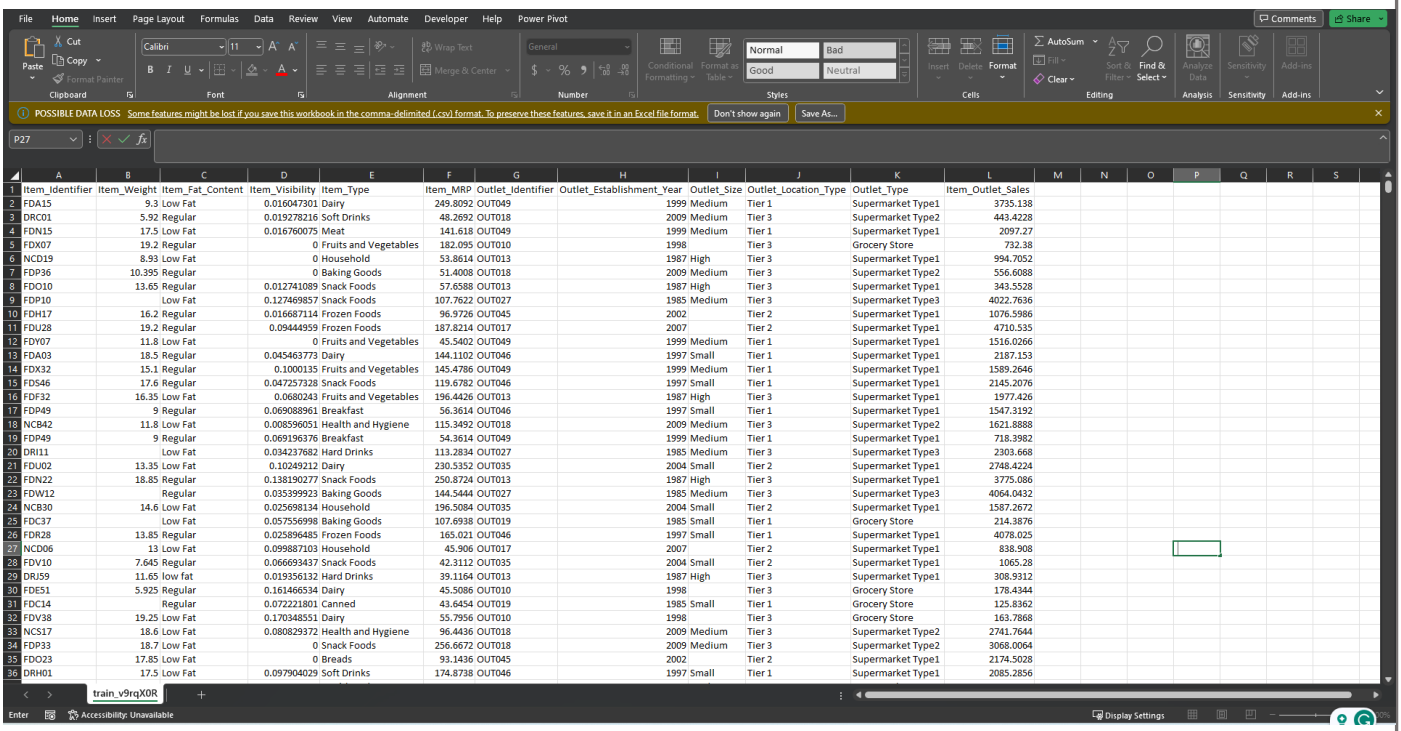
# Predicting Big Mart Sales Using Machine Learning Models in KNIME

## Overview:

In this project, machine learning workflows were built to forecast Big Mart's product sales by utilizing the visual interface of KNIME. The retail sales dataset was used to create models such as linear regression and random forest. Data import, exploration, preprocessing, modeling, testing, and result generation were all covered by the workflows, which were based on the precise actions taken during the project.

**Training data set:**



| Item_Identifier | Item_Weight | Item_Fat_Content | Item_Visibility | Item_Type | Item_MRP | Outlet_Identifier | Outlet_Establishment_Year | Outlet_Size | Outlet_Location_Type | Outlet_Type | Item_Outlet_Sales |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FDA15 | 9.3 | Low Fat | 0.016047301 | Dairy | 249.8092 | OUT049 | 1999 | Medium | Tier 1 | Supermarket Type1 | 3735.138 |
| DRC01 | 5.92 | Regular | 0.019278216 | Soft Drinks | 48.2692 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 | 443.4228 |
| FDN15 | 17.5 | Low Fat | 0.016760075 | Meat | 141.618 | OUT049 | 1999 | Medium | Tier 1 | Supermarket Type1 | 2097.27 |
| FDX07 | 19.2 | Regular | 0 | Fruits and Vegetables | 182.095 | OUT010 | 1998 | | Tier 3 | Grocery Store | 732.38 |
| NCD19 | 8.93 | Low Fat | 0 | Household | 53.8614 | OUT013 | 1987 | High | Tier 3 | Supermarket Type1 | 994.7052 |
| FDP36 | 10.395 | Regular | 0 | Baking Goods | 51.4008 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 | 556.6088 |
| FDO10 | 13.65 | Regular | 0.012741089 | Snack Foods | 57.6588 | OUT013 | 1987 | High | Tier 3 | Supermarket Type1 | 343.5528 |
| FDP10 | | Low Fat | 0.127469857 | Snack Foods | 107.7622 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 4022.7636 |
| FDH17 | 16.2 | Regular | 0.016687114 | Frozen Foods | 96.9726 | OUT045 | 2002 | | Tier 2 | Supermarket Type1 | 1076.5986 |
| FDU28 | 19.2 | Regular | 0.09444959 | Frozen Foods | 187.8214 | OUT017 | 2007 | | Tier 2 | Supermarket Type1 | 4710.535 |
| FDY07 | 11.8 | Low Fat | 0 | Fruits and Vegetables | 45.5402 | OUT049 | 1999 | Medium | Tier 1 | Supermarket Type1 | 1516.0266 |
| FDA03 | 18.5 | Regular | 0.045463773 | Dairy | 144.1102 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 | 2187.153 |
| FDX32 | 15.1 | Regular | 0.1000135 | Fruits and Vegetables | 145.4786 | OUT049 | 1999 | Medium | Tier 1 | Supermarket Type1 | 1589.2646 |
| FDS46 | 17.6 | Regular | 0.047257328 | Snack Foods | 119.6782 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 | 2145.2076 |
| FDF32 | 16.35 | Low Fat | 0.0680243 | Fruits and Vegetables | 196.4426 | OUT013 | 1987 | High | Tier 3 | Supermarket Type1 | 1977.426 |
| FDP49 | 9 | Regular | 0.069088961 | Breakfast | 56.3614 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 | 1547.3192 |
| NCB42 | 11.8 | Low Fat | 0.008596051 | Health and Hygiene | 115.3492 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 | 1621.8888 |
| FDP49 | 9 | Regular | 0.069196376 | Breakfast | 54.3614 | OUT049 | 1999 | Medium | Tier 1 | Supermarket Type1 | 718.3982 |
| DRI11 | | Low Fat | 0.034237682 | Hard Drinks | 113.2834 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 2303.668 |
| FDU02 | 13.35 | Low Fat | 0.10249212 | Dairy | 230.5352 | OUT035 | 2004 | Small | Tier 2 | Supermarket Type1 | 2748.4224 |
| FDN22 | 18.85 | Regular | 0.138190277 | Snack Foods | 250.8724 | OUT013 | 1987 | High | Tier 3 | Supermarket Type1 | 3775.086 |
| FDW12 | | Regular | 0.035399923 | Baking Goods | 144.5444 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 4064.0432 |
| NCB30 | 14.6 | Low Fat | 0.025698134 | Household | 196.5084 | OUT035 | 2004 | Small | Tier 2 | Supermarket Type1 | 1587.2672 |
| FDC37 | | Low Fat | 0.057556998 | Baking Goods | 107.6938 | OUT019 | 1985 | Small | Tier 1 | Grocery Store | 214.3876 |
| FDR28 | 13.85 | Regular | 0.025896485 | Frozen Foods | 165.021 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 | 4078.025 |
| NCD06 | 13 | Low Fat | 0.099887103 | Household | 45.906 | OUT017 | 2007 | | Tier 2 | Supermarket Type1 | 838.908 |
| FDV10 | 7.645 | Regular | 0.066693437 | Snack Foods | 42.3112 | OUT035 | 2004 | Small | Tier 2 | Supermarket Type1 | 1065.28 |
| DRJ59 | 11.65 | low fat | 0.019356132 | Hard Drinks | 39.1164 | OUT013 | 1987 | High | Tier 3 | Supermarket Type1 | 308.9312 |
| FDE51 | 5.925 | Regular | 0.161466534 | Dairy | 45.5086 | OUT010 | 1998 | | Tier 3 | Grocery Store | 178.4344 |
| FDC14 | | Regular | 0.072221801 | Canned | 43.6454 | OUT019 | 1985 | Small | Tier 1 | Grocery Store | 125.8362 |
| FDV38 | 19.25 | Low Fat | 0.170348551 | Dairy | 55.7956 | OUT010 | 1998 | | Tier 3 | Grocery Store | 163.7868 |
| NCS17 | 18.6 | Low Fat | 0.080829372 | Health and Hygiene | 96.4436 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 | 2741.7644 |
| FDP33 | 18.7 | Low Fat | 0 | Snack Foods | 256.6672 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 | 3068.0064 |
| FDO23 | 17.85 | Low Fat | 0 | Breads | 93.1436 | OUT045 | 2002 | | Tier 2 | Supermarket Type1 | 2174.5028 |
| DRH01 | 17.5 | Low Fat | 0.097904029 | Soft Drinks | 174.8738 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 | 2085.2856 |

The Big Mart sales dataset provided contains retail transaction data for predicting product sales at specific store outlets. It has been split into a training and test set for developing and evaluating machine learning models.
source: [https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/#ProblemStatement](https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/#ProblemStatement)

The training dataset is available as a CSV file with 8523 rows and 12 columns. Each row represents a product sold at a particular store and the columns provide attributes of the item, outlet details, and the target sales variable to predict. The columns are:

Item_Identifier (Numeric): Unique ID for each product

Item_Weight (Numeric): Weight of the product

Item_Fat_Content (Categorical): Whether the product is low fat or not.

Item_Visibility (Numeric): Display area percentage for product

Item_Type (Categorical): Category of the product

Item_MRP (Numeric): Maximum retail price of the product

Outlet_Identifier (Numeric): Unique ID for each outlet

Outlet_Establishment_Year (Numeric): Year when the outlet was established.

Outlet_Size (Categorical): Size of the outlet in sq. ft.

Outlet_Location_Type (Categorical): Type of city where the outlet is located.

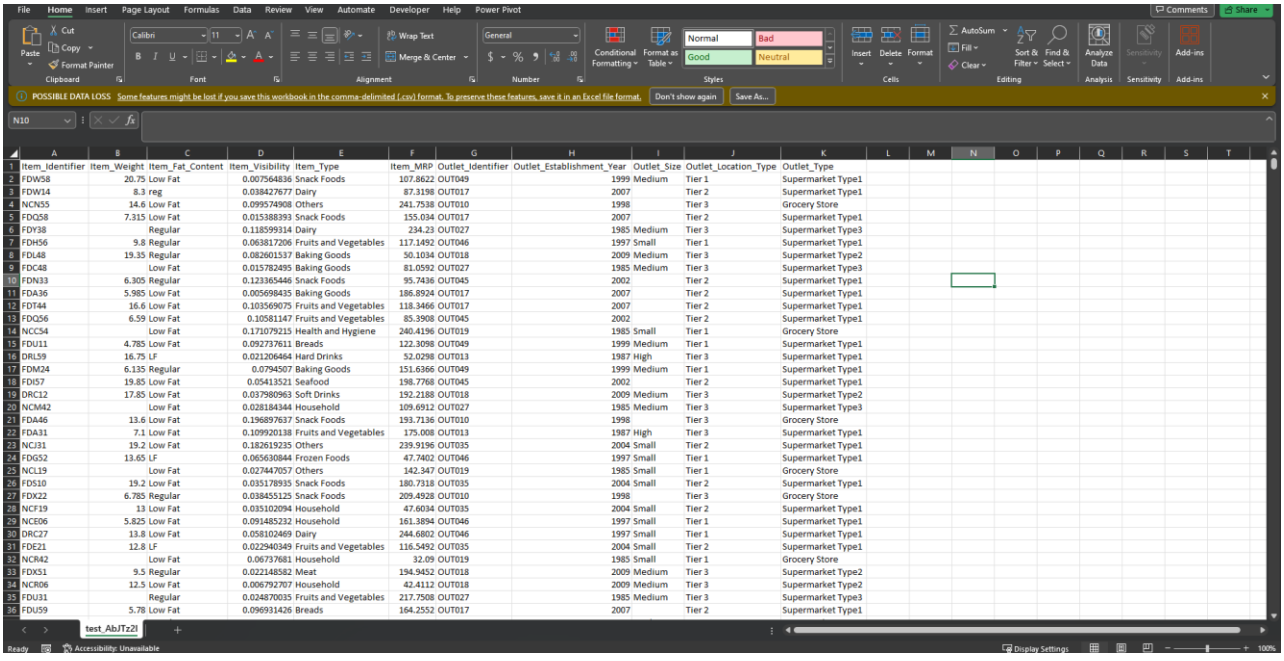
Outlet_Type (Categorical): Grocery or supermarket outlet

Item_Outlet_Sales (Numeric): Sales of the product at the outlet. This is the target variable.

The test dataset contains 5681 rows and 11 columns. It has the same 11 predictor columns as training without the target variable.

This data can be used to develop a machine learning model to predict the sales for the test products at given outlets based on their attributes. The model's performance can then be evaluated by comparing predictions to actual sales.

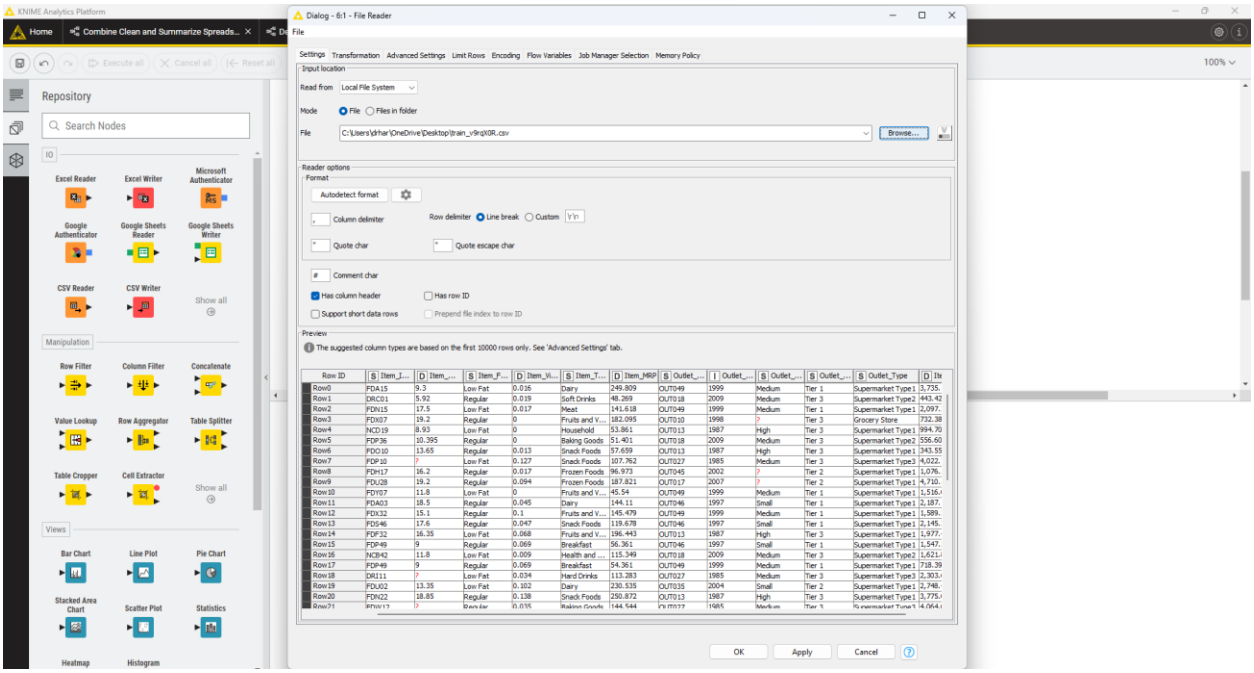
**Testing Data set:**



| Item_Identifier | Item_Weight | Item_Fat_Content | Item_Visibility | Item_Type | Item_MRP | Outlet_Identifier | Outlet_Establishment_Year | Outlet_Size | Outlet_Location_Type | Outlet_Type |
|---|---|---|---|---|---|---|---|---|---|---|
| FDW58 | 20.75 | Low Fat | 0.007564836 | Snack Foods | 107.8622 | OUT049 | 1999 | Medium | Tier 1 | Supermarket Type1 |
| FDW14 | 8.3 | reg | 0.038427677 | Dairy | 87.3198 | OUT017 | 2007 | | Tier 2 | Supermarket Type1 |
| NCN55 | 14.6 | Low Fat | 0.099574908 | Others | 241.7538 | OUT010 | 1998 | | Tier 3 | Grocery Store |
| FDQ58 | 7.315 | Low Fat | 0.015388393 | Snack Foods | 155.034 | OUT017 | 2007 | | Tier 2 | Supermarket Type1 |
| FDY38 | | Regular | 0.118599314 | Dairy | 234.23 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 |
| FDH56 | 9.8 | Regular | 0.063817206 | Fruits and Vegetables | 117.1492 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 |
| FDL48 | 19.35 | Regular | 0.082601537 | Baking Goods | 50.1034 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 |
| FDC48 | | Low Fat | 0.015782495 | Baking Goods | 81.0592 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 |
| FDN33 | 6.305 | Regular | 0.123365446 | Snack Foods | 95.7436 | OUT045 | 2002 | | Tier 2 | Supermarket Type1 |
| FDA36 | 5.985 | Low Fat | 0.005698435 | Baking Goods | 186.8924 | OUT017 | 2007 | | Tier 2 | Supermarket Type1 |
| FDT44 | 16.6 | Low Fat | 0.103569075 | Fruits and Vegetables | 118.3466 | OUT017 | 2007 | | Tier 2 | Supermarket Type1 |
| FDQ56 | 6.59 | Low Fat | 0.10581147 | Fruits and Vegetables | 85.3908 | OUT045 | 2002 | | Tier 2 | Supermarket Type1 |
| NCC54 | | Low Fat | 0.171079215 | Health and Hygiene | 240.4196 | OUT019 | 1985 | Small | Tier 1 | Grocery Store |
| FDU11 | 4.785 | Low Fat | 0.092737611 | Breads | 122.3098 | OUT049 | 1999 | Medium | Tier 1 | Supermarket Type1 |
| DRL59 | 16.75 | LF | 0.021206464 | Hard Drinks | 52.0298 | OUT013 | 1987 | High | Tier 3 | Supermarket Type1 |
| FDM24 | 6.135 | Regular | 0.0794507 | Baking Goods | 151.6366 | OUT049 | 1999 | Medium | Tier 1 | Supermarket Type1 |
| FDI57 | 19.85 | Low Fat | 0.05413521 | Seafood | 198.7768 | OUT045 | 2002 | | Tier 2 | Supermarket Type1 |
| DRC12 | 17.85 | Low Fat | 0.037980963 | Soft Drinks | 192.2188 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 |
| NCM42 | | Low Fat | 0.028184344 | Household | 109.6912 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 |
| FDA46 | 13.6 | Low Fat | 0.196897637 | Snack Foods | 193.7136 | OUT010 | 1998 | | Tier 3 | Grocery Store |
| FDA31 | 7.1 | Low Fat | 0.109920138 | Fruits and Vegetables | 175.008 | OUT013 | 1987 | High | Tier 3 | Supermarket Type1 |
| NCJ31 | 19.2 | Low Fat | 0.182619235 | Others | 239.9196 | OUT035 | 2004 | Small | Tier 2 | Supermarket Type1 |
| FDG52 | 13.65 | LF | 0.065630844 | Frozen Foods | 47.7402 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 |
| NCL19 | | Low Fat | 0.027447057 | Others | 142.347 | OUT019 | 1985 | Small | Tier 1 | Grocery Store |
| FDS10 | 19.2 | Low Fat | 0.035178935 | Snack Foods | 180.7318 | OUT035 | 2004 | Small | Tier 2 | Supermarket Type1 |
| FDX22 | 6.785 | Regular | 0.038455125 | Snack Foods | 209.4928 | OUT010 | 1998 | | Tier 3 | Grocery Store |
| NCF19 | 13 | Low Fat | 0.035102094 | Household | 47.6034 | OUT035 | 2004 | Small | Tier 2 | Supermarket Type1 |
| NCE06 | 5.825 | Low Fat | 0.091485232 | Household | 161.3894 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 |
| DRC27 | 13.8 | Low Fat | 0.058102469 | Dairy | 244.6802 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 |
| FDE21 | 12.8 | LF | 0.022940349 | Fruits and Vegetables | 116.5492 | OUT035 | 2004 | Small | Tier 2 | Supermarket Type1 |
| NCR42 | | Low Fat | 0.06737681 | Household | 32.09 | OUT019 | 1985 | Small | Tier 1 | Grocery Store |
| FDX51 | 9.5 | Regular | 0.022148582 | Meat | 194.9452 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 |
| NCR06 | 12.5 | Low Fat | 0.006792707 | Household | 42.4112 | OUT018 | 2009 | Medium | Tier 3 | Supermarket Type2 |
| FDU31 | | Regular | 0.024870035 | Fruits and Vegetables | 217.7508 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 |
| FDU59 | 5.78 | Low Fat | 0.096931426 | Breads | 164.2552 | OUT017 | 2007 | | Tier 2 | Supermarket Type1 |

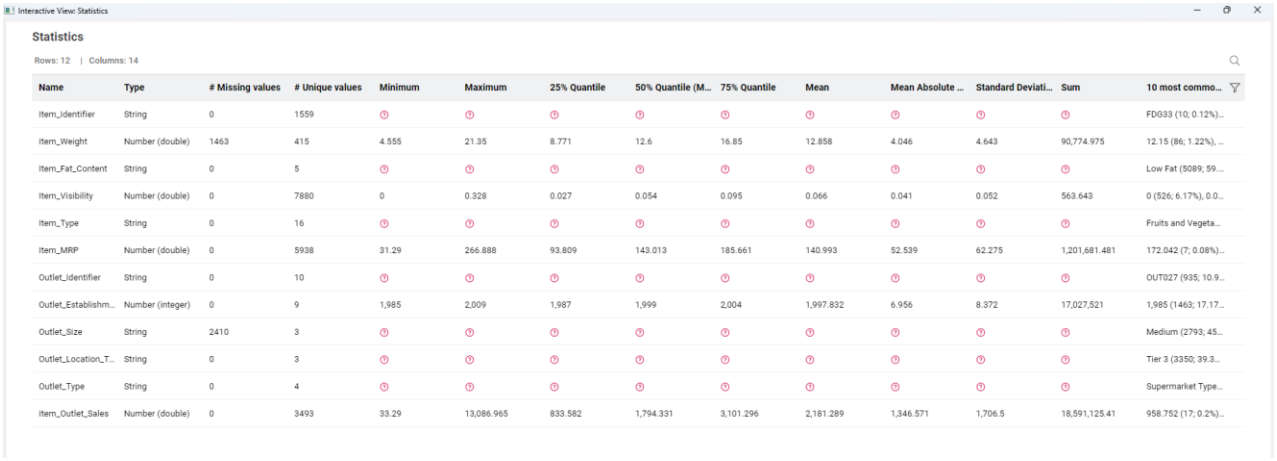
**Data Import**

The first step was importing the Big Mart sales training and testing dataset CSV files into KNIME using the File Reader nodes. This node reads the data from a specified file path and outputs a KNIME table. Two File Reader nodes were added, configured to the respective file paths, and executed to import the data.

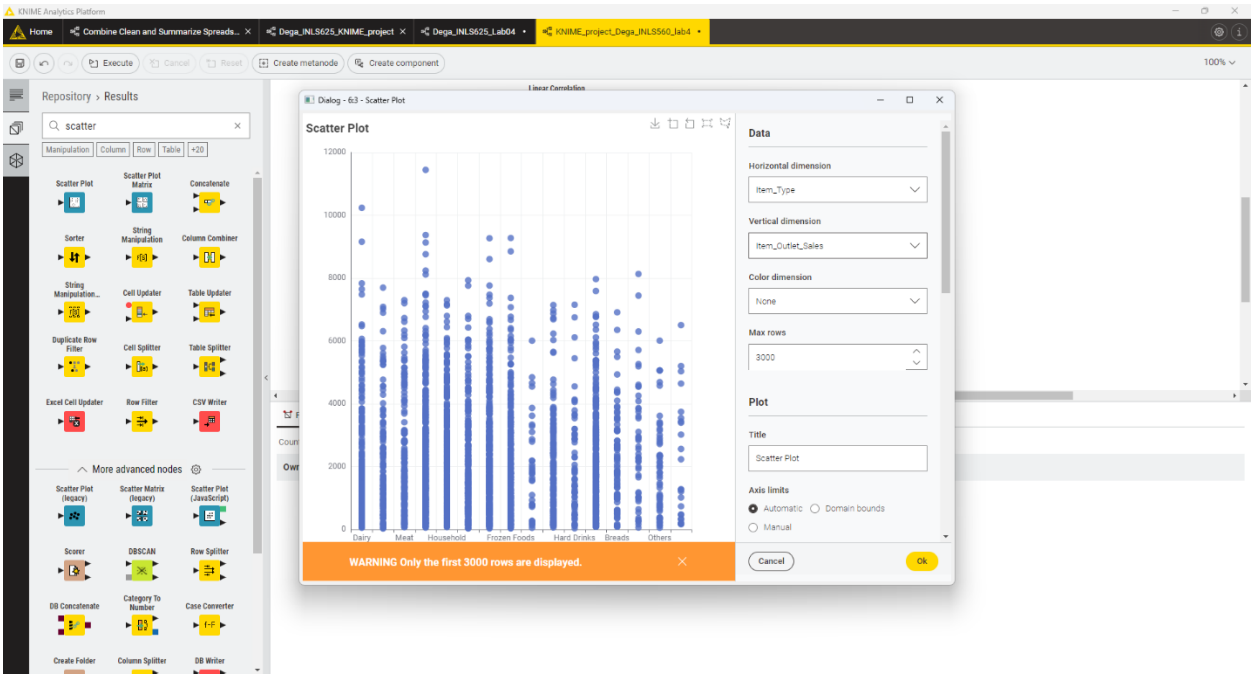


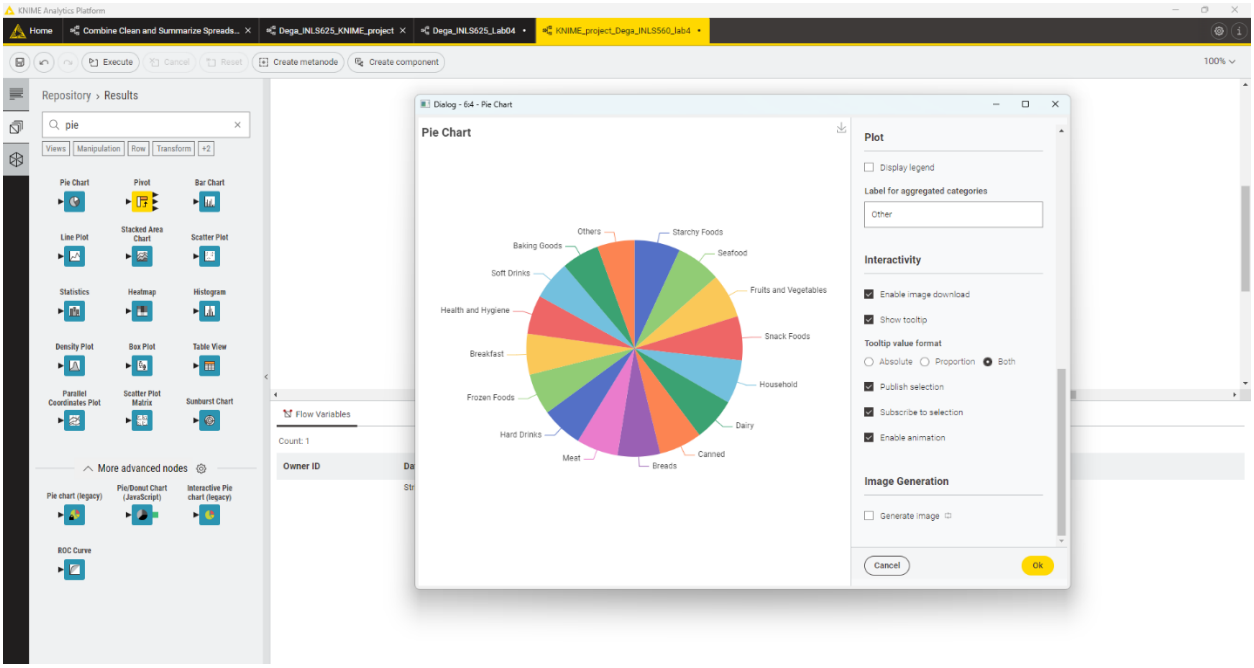

**Exploratory Data Analysis**

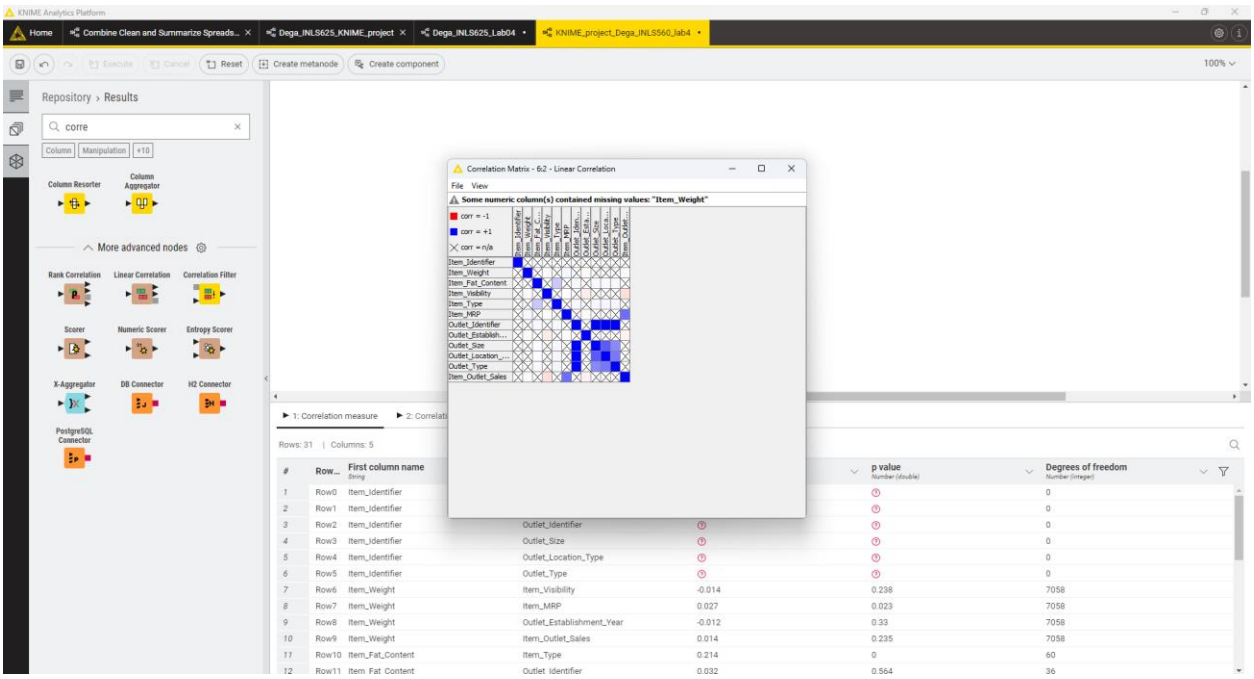
Some initial exploratory analysis was done on the training data to understand relationships between columns**. The Statistics node** provided summary metrics like mean, min, and max for numerical columns. **The Scatter Plot** visualized relationships between attributes like Item_Type and Item_Outlet_Sales. **The correlation matrix** was viewed to identify strongly correlated variables. This exploration provided insights into higher-selling outlet types and items.




| Name | Type | # Missing values | # Unique values | Minimum | Maximum | 25% Quantile | 50% Quantile (M... | 75% Quantile | Mean | Mean Absolute ... | Standard Deviati... | Sum | 10 most commo... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item_Identifier | String | 0 | 1559 | | | | | | | | | | FDG33 (10; 0.12%)... |
| Item_Weight | Number (double) | 1463 | 415 | 4.555 | 21.35 | 8.771 | 12.6 | 16.85 | 12.858 | 4.046 | 4.643 | 90,774.975 | 12.15 (86; 1.22%), ... |
| Item_Fat_Content | String | 0 | 5 | | | | | | | | | | Low Fat (5089; 59.... |
| Item_Visibility | Number (double) | 0 | 7880 | 0 | 0.328 | 0.027 | 0.054 | 0.095 | 0.066 | 0.041 | 0.052 | 563.643 | 0 (526; 6.17%), 0.0... |
| Item_Type | String | 0 | 16 | | | | | | | | | | Fruits and Vegeta... |
| Item_MRP | Number (double) | 0 | 5938 | 31.29 | 266.888 | 93.809 | 143.013 | 185.661 | 140.993 | 52.539 | 62.275 | 1,201,681.481 | 172.042 (7; 0.08%)... |
| Outlet_Identifier | String | 0 | 10 | | | | | | | | | | OUT027 (935; 10.9... |
| Outlet_Establishm... | Number (integer) | 0 | 9 | 1,985 | 2,009 | 1,987 | 1,999 | 2,004 | 1,997.832 | 6.956 | 8.372 | 17,027,521 | 1,985 (1463; 17.17... |
| Outlet_Size | String | 2410 | 3 | | | | | | | | | | Medium (2793; 45... |
| Outlet_Location_T... | String | 0 | 3 | | | | | | | | | | Tier 3 (3350; 39.3... |
| Outlet_Type | String | 0 | 4 | | | | | | | | | | Supermarket Type... |
| Item_Outlet_Sales | Number (double) | 0 | 3493 | 33.29 | 13,086.965 | 833.582 | 1,794.331 | 3,101.296 | 2,181.289 | 1,346.571 | 1,706.5 | 18,591,125.41 | 958.752 (17; 0.2%)... |

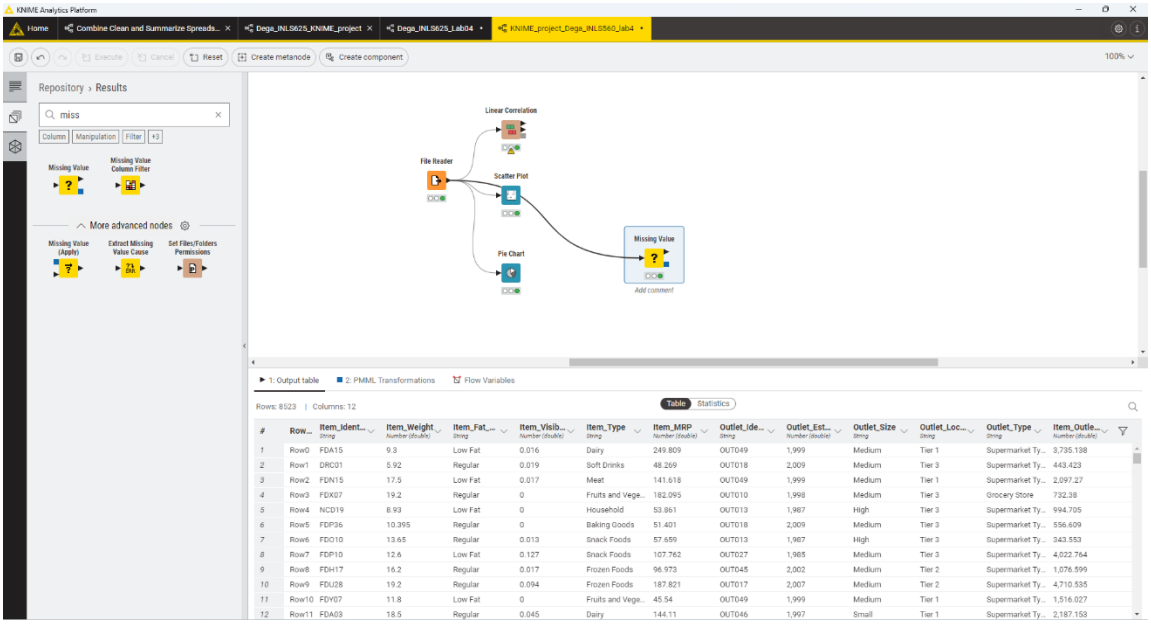




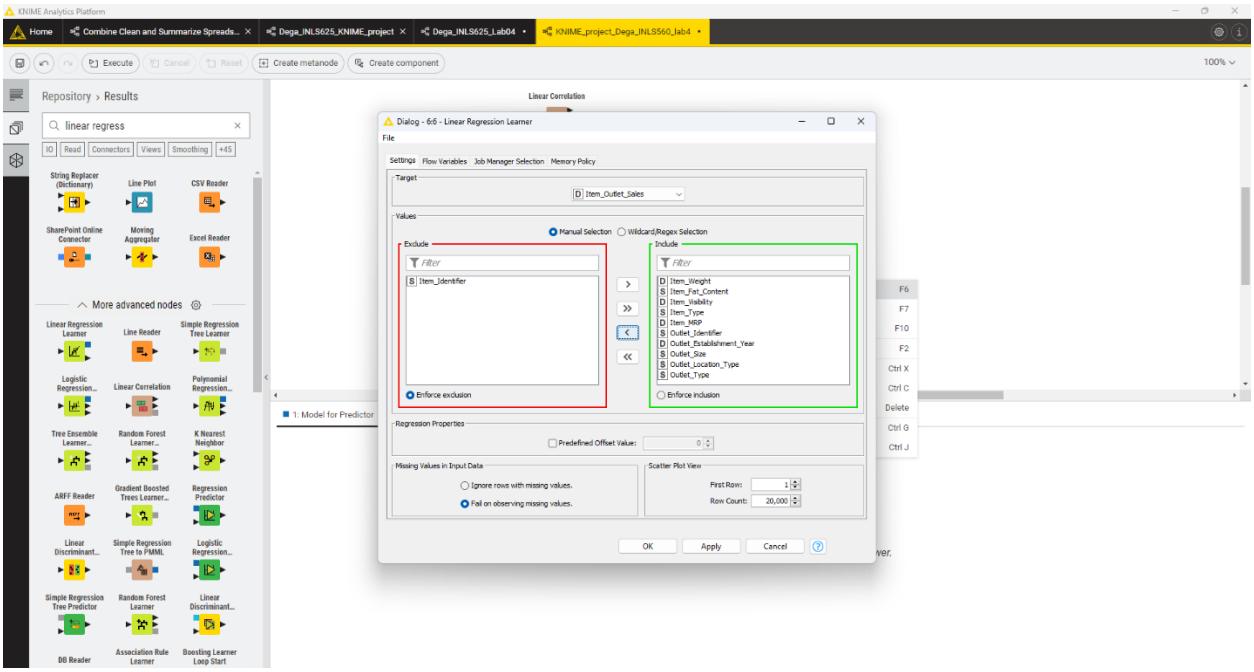

**Data Preprocessing**

The raw training data had issues like missing values and non-numeric columns which were addressed before modeling. The Missing Value node helped identify and replace missing values in numerical columns with a median. These steps output clean, normalized training data ready for ML modeling.

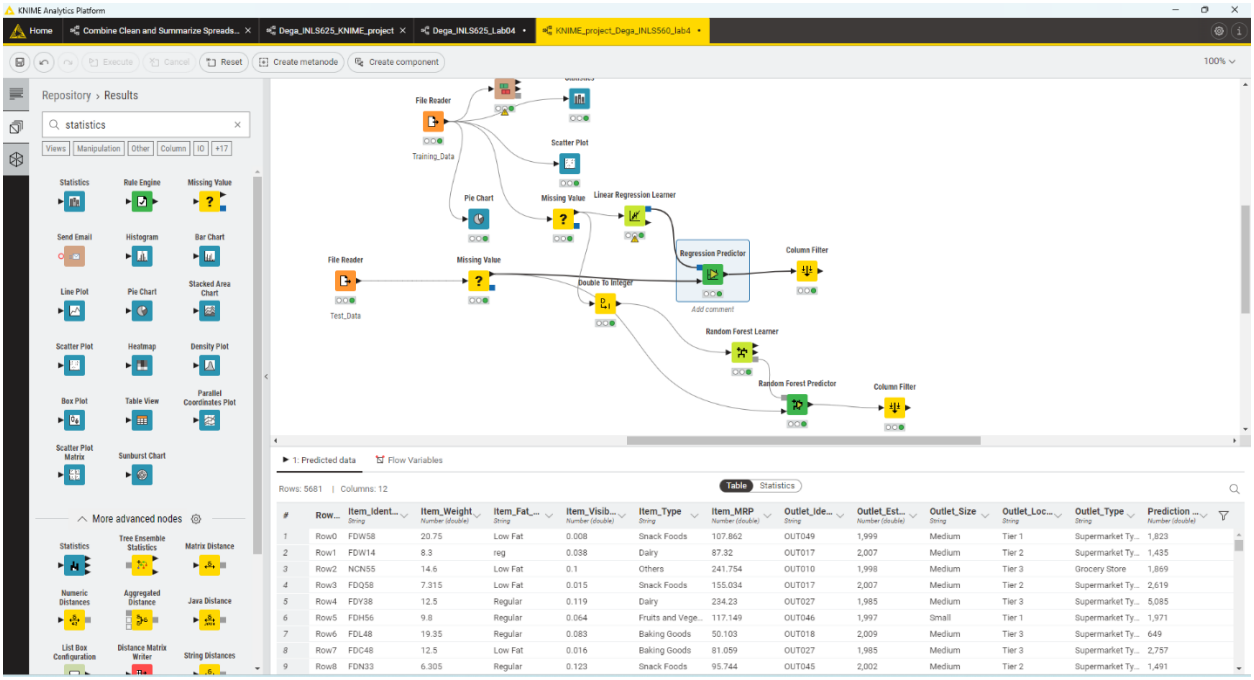



**Linear Regression Model**

A linear regression model was built to predict the Item_Outlet_Sales. The Linear Regression Learner node was added, and the cleaned training data was connected to its input port. In the node configuration, Item_Outlet_Sales was set as the target column while numeric columns like Item_MRP, Outlet_Establishment_Year, etc. were chosen as predictors. The node was executed to train the regression model.

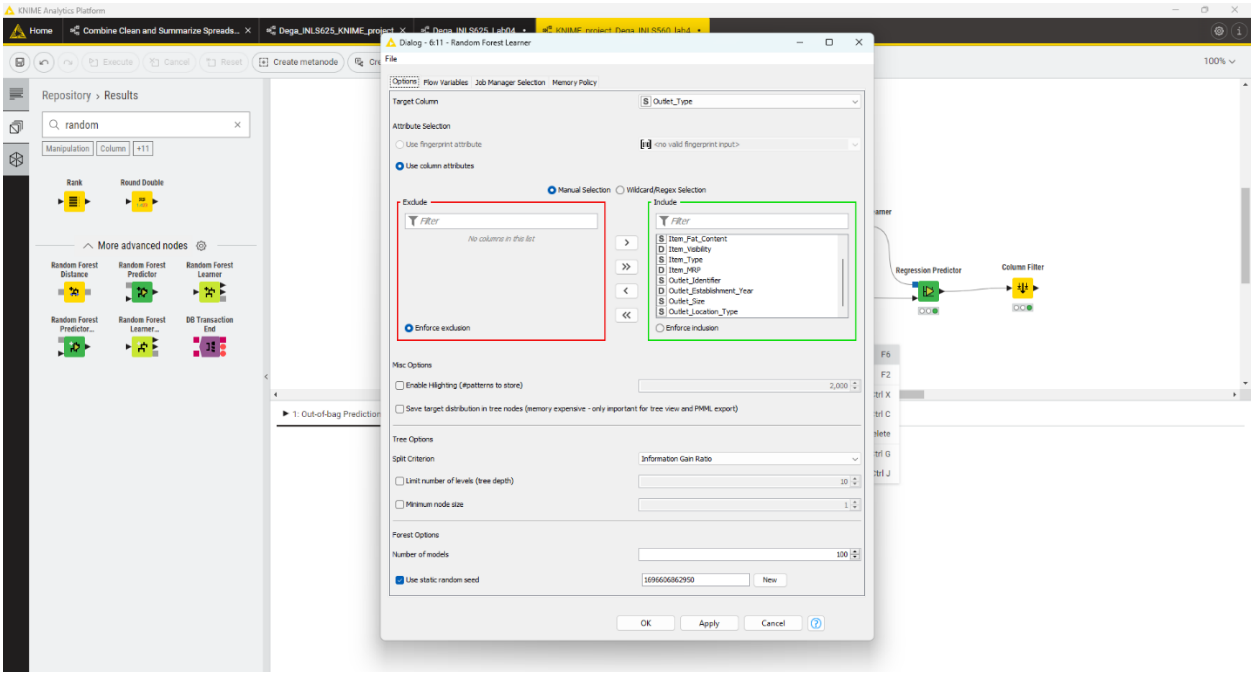


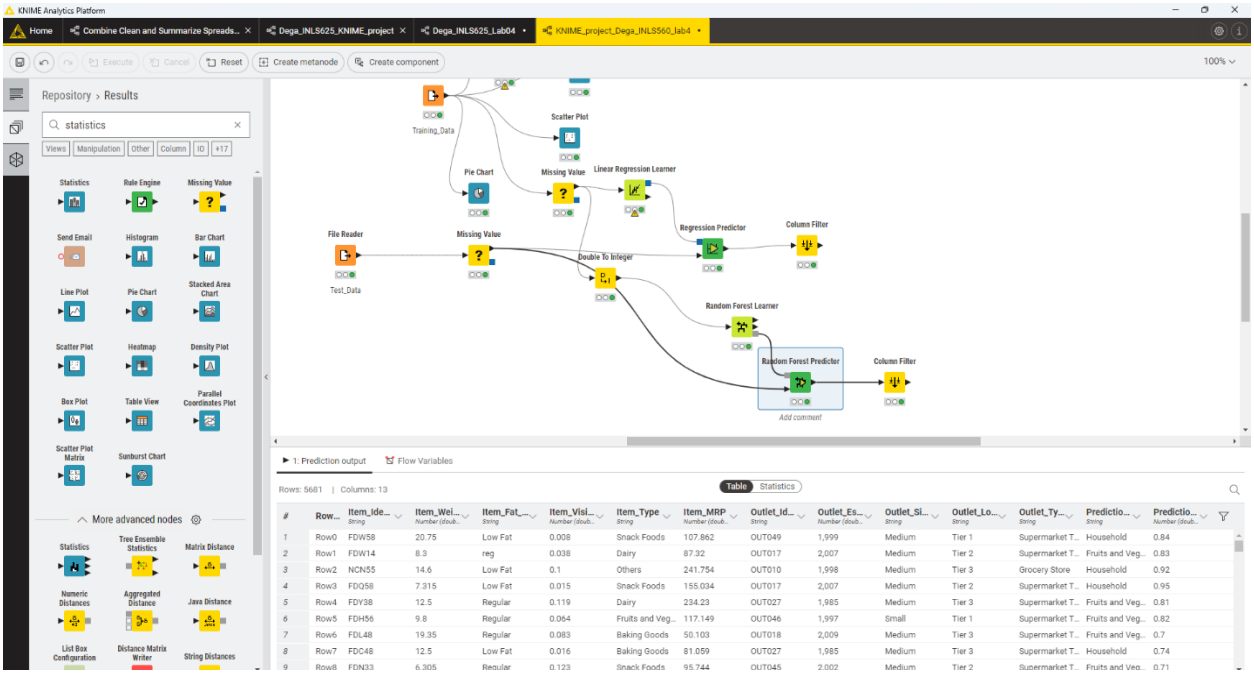
The trained model was connected to a Linear Regression Predictor node. The clean test data was connected to the predictor's second input port. After executing this workflow, the predicted sales values were compared to actuals to evaluate performance on the test set.




**Random Forest Model**

A random forest classifier was built with Outlet_type as the target variable. The Random Forest Learner node was configured to use the preprocessed training data. Outlet_type was selected as the target column while columns like item_fat_content, outlet_size, etc. were chosen as predictors after converting Double to an integer using the Double to integer node.

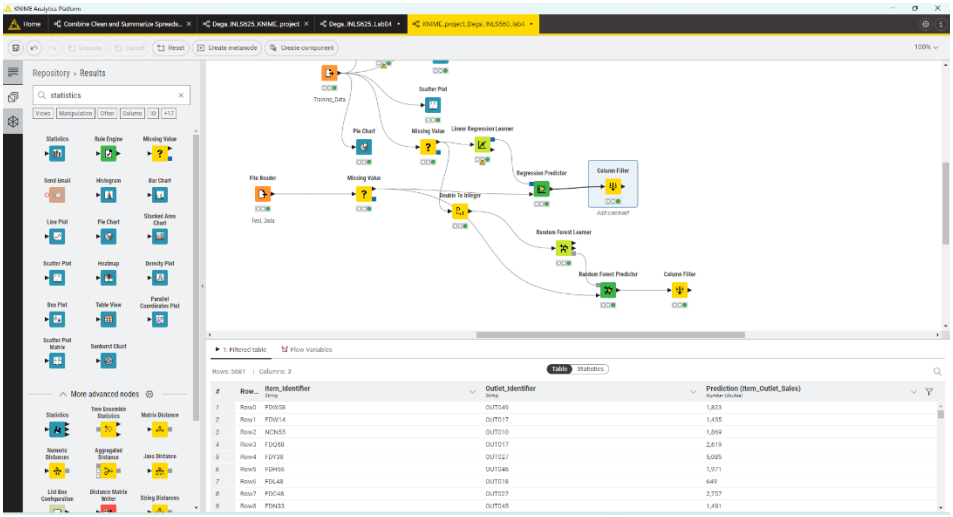


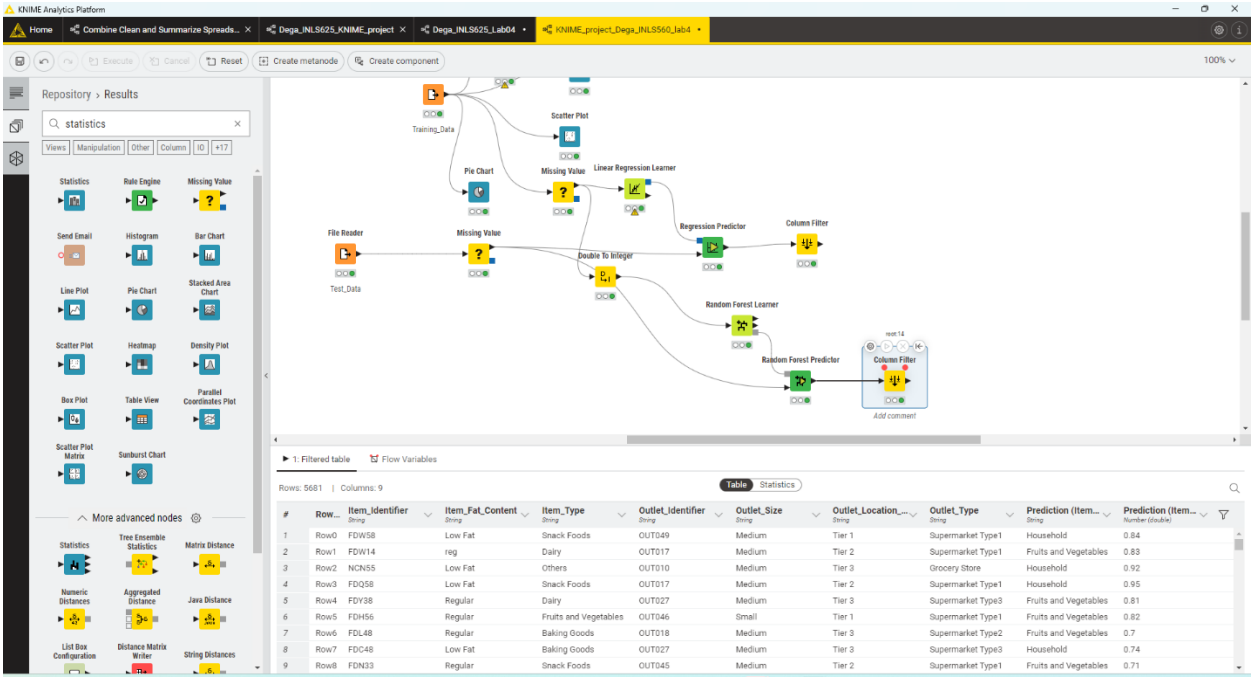




The trained random forest model was connected to the Random Forest Predictor node. The preprocessed test data was fed to the second input port. The predicted item types were compared to actual values to assess classification accuracy on the test data.

**Result Analysis**

The Column Filter node was used to output only the required columns from the predictors. Summary statistics and scoring metrics were computed to evaluate model performance. The predictions were reasonably accurate and can be further improved by tuning models and adding more relevant features.






**Conclusion:**

This project provided hands-on experience with KNIME's capabilities for building end-to-end ML workflows visually without coding. The linear regression model performed decently for sales prediction. The random forest model also had reasonable accuracy in classifying item types. The workflows can be extended by adding more data, features, and tuning techniques.

List of other project sites that you searched and explored before choosing your project (at least 5):

1. Healthcare fraud Source: https://www.kaggle.com/datasets/tamilsel/healthcare-providers-data

2. Medical Practitioner and Prescription Behavior-National Survey of Physician Characteristics and Prescription Practices in the USA
   https://www.kaggle.com/datasets/tubmak/dataset
3. Titanic - ML Project https://hub.knime.com/augustinejoseph/spaces/Public/latest/Titanic%20-%20ML%20Project~TsA5hTL4pnFIfDHg
4. ECG Heartbeat Categorization Dataset https://www.kaggle.com/datasets/shayanfazeli/heartbeat
   https://hub.knime.com/knime/spaces/Digital%20Healthcare/latest/ECG%20Arrythmia%20Detection/ecg_cnn_mit~bWv0UtH6sTLAgcnn

5. Triage Score Prediction
   https://hub.knime.com/knime/spaces/Digital%20Healthcare/latest/Triage%20Score%20Prediction/Triage%20Score%20Prediction~-wdfGA8UecAI42L8