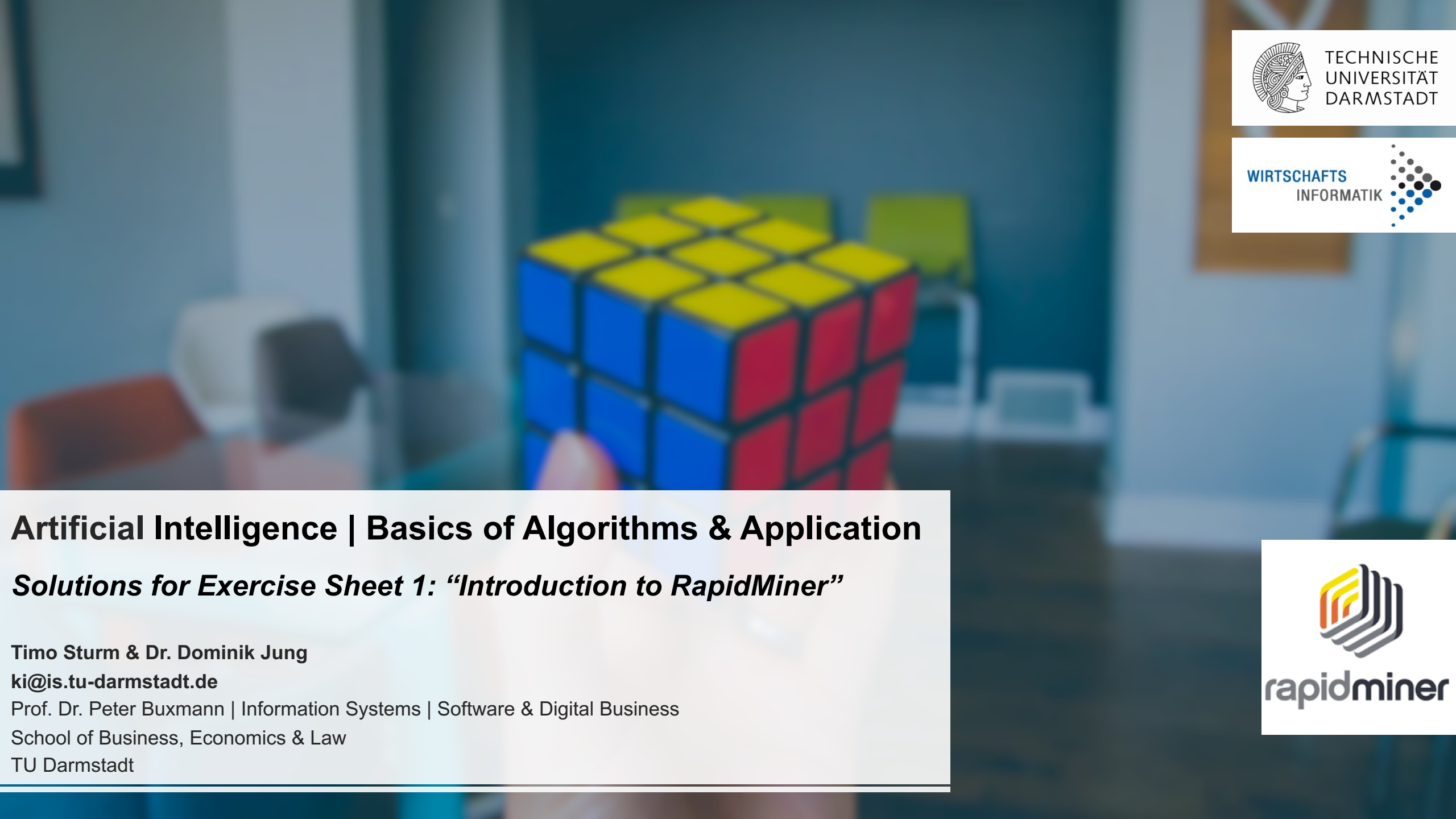


Artificial Intelligence | Basics of Algorithms & Application

Solutions for Exercise Sheet 1: "Introduction to RapidMiner"

Timo Sturm & Dr. Dominik Jung
ki@is.tu-darmstadt.de
Prof. Dr. Peter Buxmann | Information Systems | Software & Digital Business
School of Business, Economics & Law
TU Darmstadt

Exercise 1.1: Iris Flowers (1/2)

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *Please explore the iris dataset visually in RapidMiner Studio and indicate the value ranges for the different attributes that can be used to distinguish the different species.*

Solution:

1. You can find the Iris data in your data repository under `Samples/data/Iris`.
2. Simply **drag** the data into the process design area and **connect** the resulting operator with the *res* port.

3. **Click** on the *Start button* to load the data into your result view.

Exercise 1.1: Iris Flowers (2/2)

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *Please explore the iris dataset visually in RapidMiner Studio and indicate the value ranges for the different attributes that can be used to distinguish the different species.*

Solution:

4. **Open** the *Visualization subview* and **select** “Scatter” as plot type.
5. Explore the data by **varying** the values of the *X-Axis* and *Value.*
6. By **looking** at the scatter plots, you can identify potential clusters. Now, you can manually define a feature combination with approx. values to differentiate the clusters from each other.
7. For example, the following feature/value combination can be used:
   - **Iris-Setosa**: *a3 = [0.0, 2.5]* and *a4 = [0.0, 0.75]*
   - **Iris-Versicolor**: *a3 = [3.0, 5.0]* and *a4 = [1.0, 1.75]*
   - **Iris-Virginica**: *a3 = (5.0, 7.0]* and *a4 = (1.75, 2.5]*

   *(*yes, a few data points are in the wrong cluster but that’s okay as the selection of values/features aimed to minimize this wrongly clustered data points*)*

Legend:
- **Square bracket** *(“[]“)*: Include min/max value of indicated value range
- **Round bracket** *(“()“)*: Exclude min/max value of indicated value range

Exercise 1.2: Restaurant Orders (1/7)

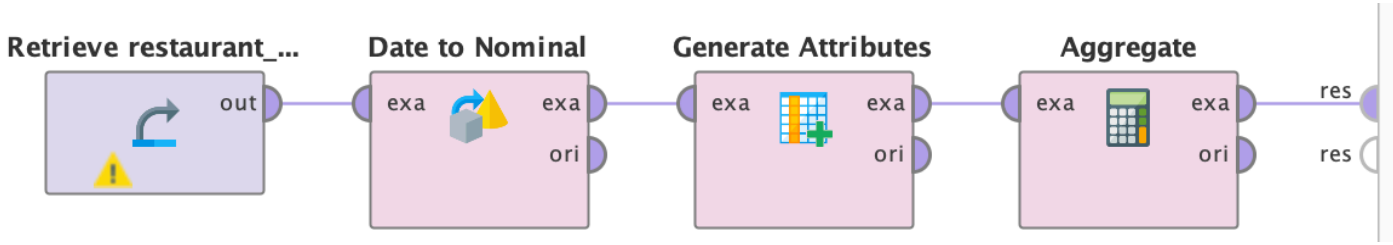
WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *a) Which was the day with the highest revenue?*

Solution:

- Use the “Date to Nominal” operator to remove the time from the date variable
- Use the “Generate Attributes” operator to create a variable that produces Quantity * Product Price and give it some name (e.g., “Profit”)
- Use the “Aggregate” operator to aggregate the generated variable by the nominalized date using “sum”
- After running the process, you can sort the data by the aggregated variable to determine the date with the highest revenue



Output:

Row No.	Order Date	sum(Profit) ↓
892	09/18/2017	2098.750
420	05/05/2018	1884.750
1207	12/31/2018	1459.400
1116	11/30/2018	1411.850

Exercise 1.2: Restaurant Orders (2/7)

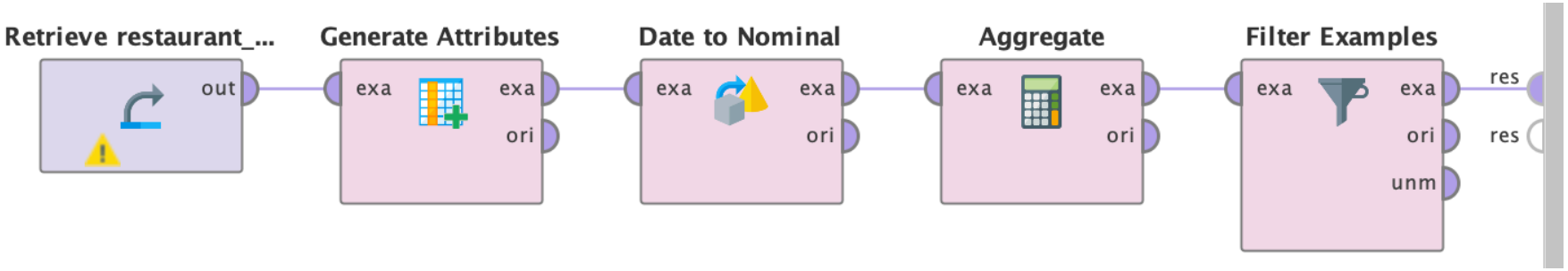
WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *b) Which items had the highest revenue in 2019?*

Solution:

- Use the “Generate Attributes” operator to create a variable that produces Quantity * Product Price and give it some name (e.g., “Profit”)
- Use the “Date to Nominal” operator to transform the date-time variable to a variable indicating only the year (*yyyy*)
- Use the “Aggregate” operator to aggregate the generated variable by the nominalized date and the product names using “sum”
- Filter the data using the “Filter Examples” operator to select only examples of 2019
- After running the process, you can sort the data by the aggregated variable to determine the date with the highest revenue



Output:

Row No.	Item Name	Order Date	sum(Pro... ↓
50	Chicken Tikka Masala	2019	5101.500
90	Korma - Chicken	2019	4824.050
168	Pilau Rice	2019	4422.050

Exercise 1.2: Restaurant Orders (3/7)

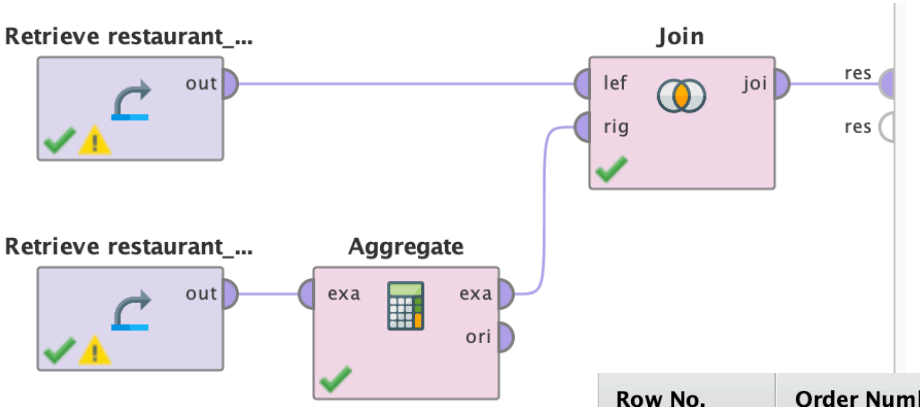
WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *c) You add the sum of ordered items per order to the result set*

Solution:

- Use the “Aggregate” operator to aggregate the Quantity attribute by the Order Number using “sum”
- Use a second instance of your initial data without applying any preprocessing
- Use the “Join” operator to perform an inner join between the non-preprocessed data and the aggregated data based on the Order Number variable



Output:

Row No.	Order Number	Order Date	Item Name	Quantity	Product Price	Total prod...	sum(Quantity)
1	16118	Aug 3, 201...	Plain Papad...	2	0.800	6	7
2	16118	Aug 3, 201...	King Prawn ...	1	12.950	6	7
3	16118	Aug 3, 201...	Garlic Naan	1	2.950	6	7
4	16118	Aug 3, 201...	Mushroom R...	1	3.950	6	7
5	16118	Aug 3, 201...	Paneer Tikk...	1	8.950	6	7
6	16118	Aug 3, 201...	Mango Chut...	1	0.500	6	7

Exercise 1.2: Restaurant Orders (4/7)

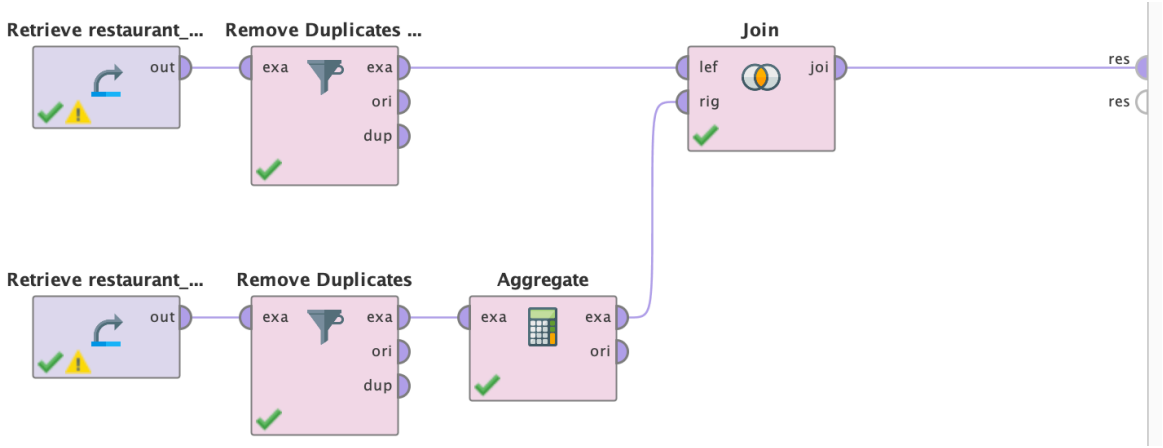
WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *d) You noticed that for some orders, the same product is listed multiple times but the "total products" feature is filled with a 1 which is wrong based on the restaurant's process for creating this data. You exclude respective orders.*

Solution:

- Simply add the "Remove Duplicates" operator to your process of subtask c)



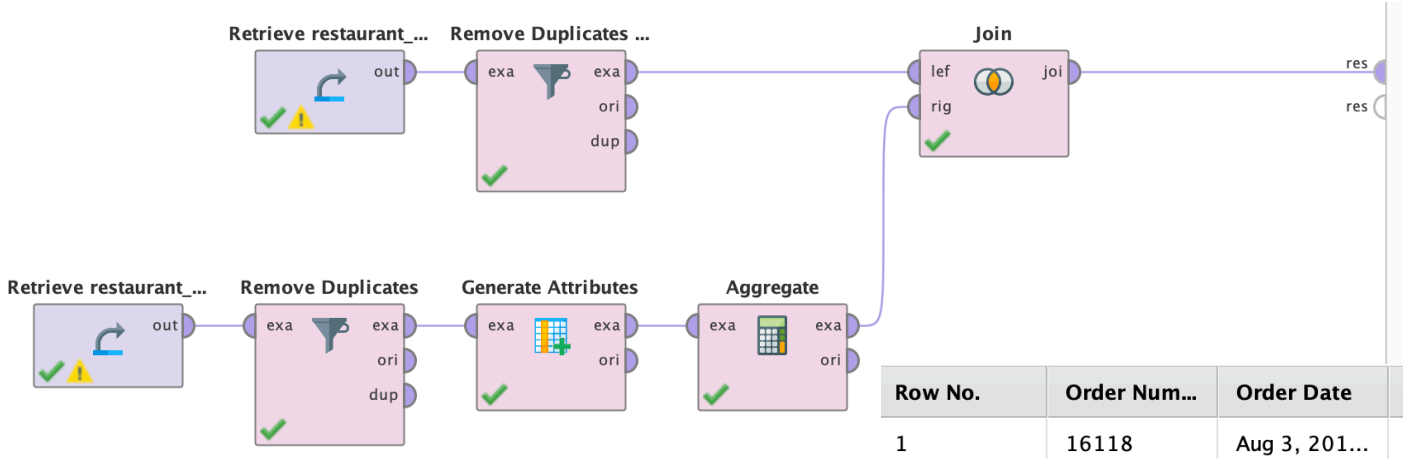
Result: Number of data points is reduced from **74,818** to **74,721**

Exercise 1.2: Restaurant Orders (5/7)

Task: *e) Afterwards, you create a new feature which indicates the full price for each order.*

Solution:

- Use the process of the previous subtask as a basis
- Use the “Generate Attributes” operator to create a variable that produces Quantity * Product Price and give it some name (e.g., “Profit”)
- Extend the “Aggregate” operator with the generated variable as a further “aggregation attribute” using also sum



Row No.	Order Num...	Order Date	Item Name	Quantity	Product Price	Total prod...	sum(Quant...	sum(Profit)
1	16118	Aug 3, 201...	Plain Papad...	2	0.800	6	7	30.900
2	16118	Aug 3, 201...	King Prawn ...	1	12.950	6	7	30.900
3	16118	Aug 3, 201...	Garlic Naan	1	2.950	6	7	30.900
4	16118	Aug 3, 201...	Mushroom R...	1	3.950	6	7	30.900
5	16118	Aug 3, 201...	Paneer Tikk...	1	8.950	6	7	30.900
6	16118	Aug 3, 201...	Mango Chut...	1	0.500	6	7	30.900

Exercise 1.2: Restaurant Orders (6/7)

WIRTSCHAFTS INFORMATIK

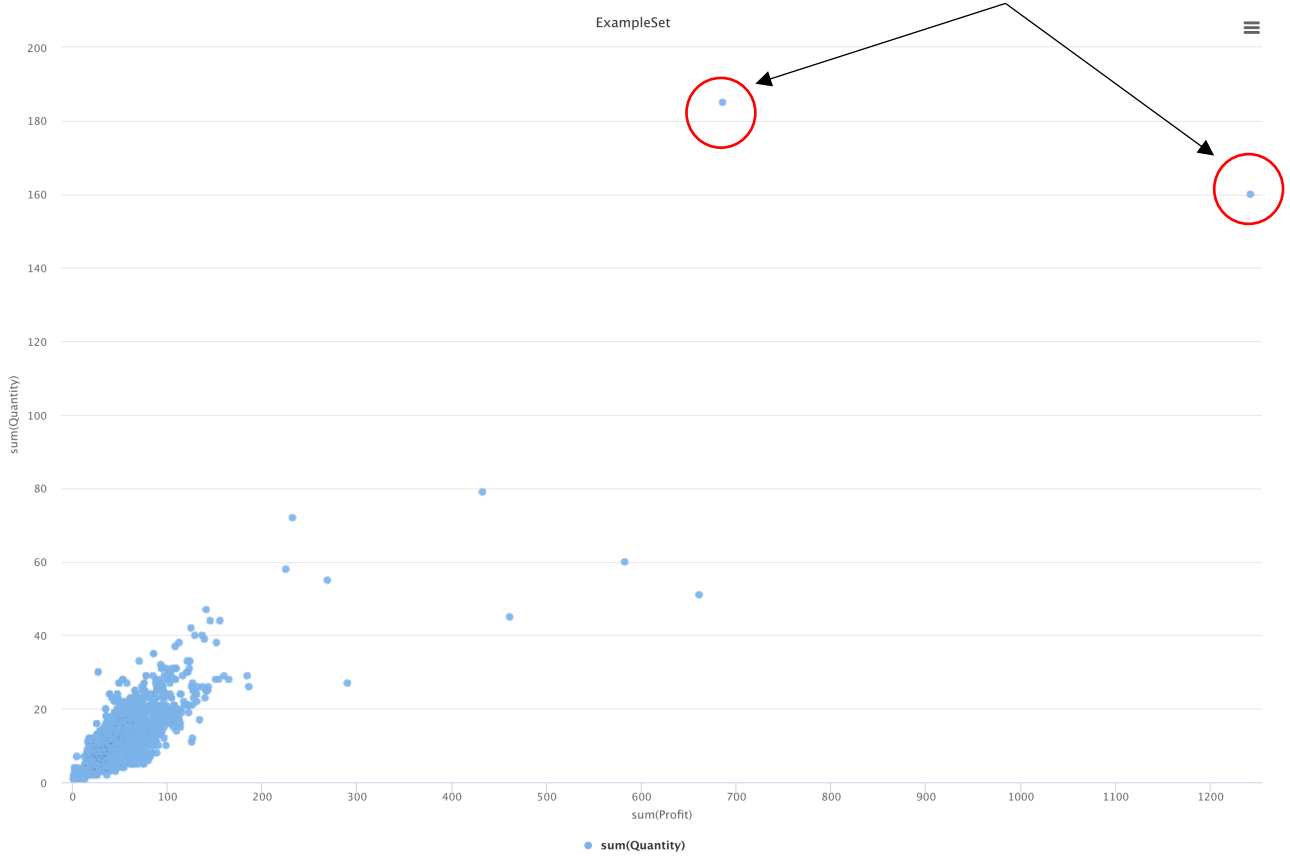
TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *f) Then, you have a look at different scatter plots of the data. Can you identify any outliers? If yes, how could you exclude them?*

Solution:

- Use the process of the previous subtask as a basis and switch to the Visualizations View
- Simply explore different variable combinations for the Scatter Plots
- You may find something like this visualization

Potential outliers



- To identify outliers, you can use the "Detect outliers" operator. However, you should be careful with excluding data that may look like outliers but are still representative for the problem you want to solve

  This operator may take a lot of time in a large dataset as it may include distance computations between all data points resulting in a high complexity

Exercise 1.2: Restaurant Orders (7/7)

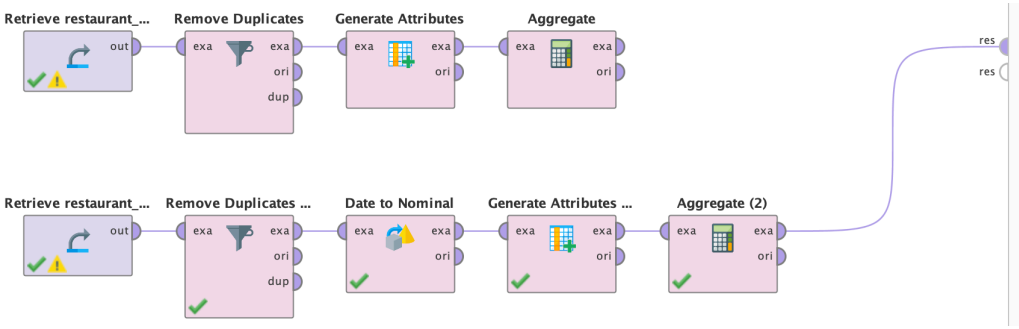
WIRTSCHAFTS INFORMATIK

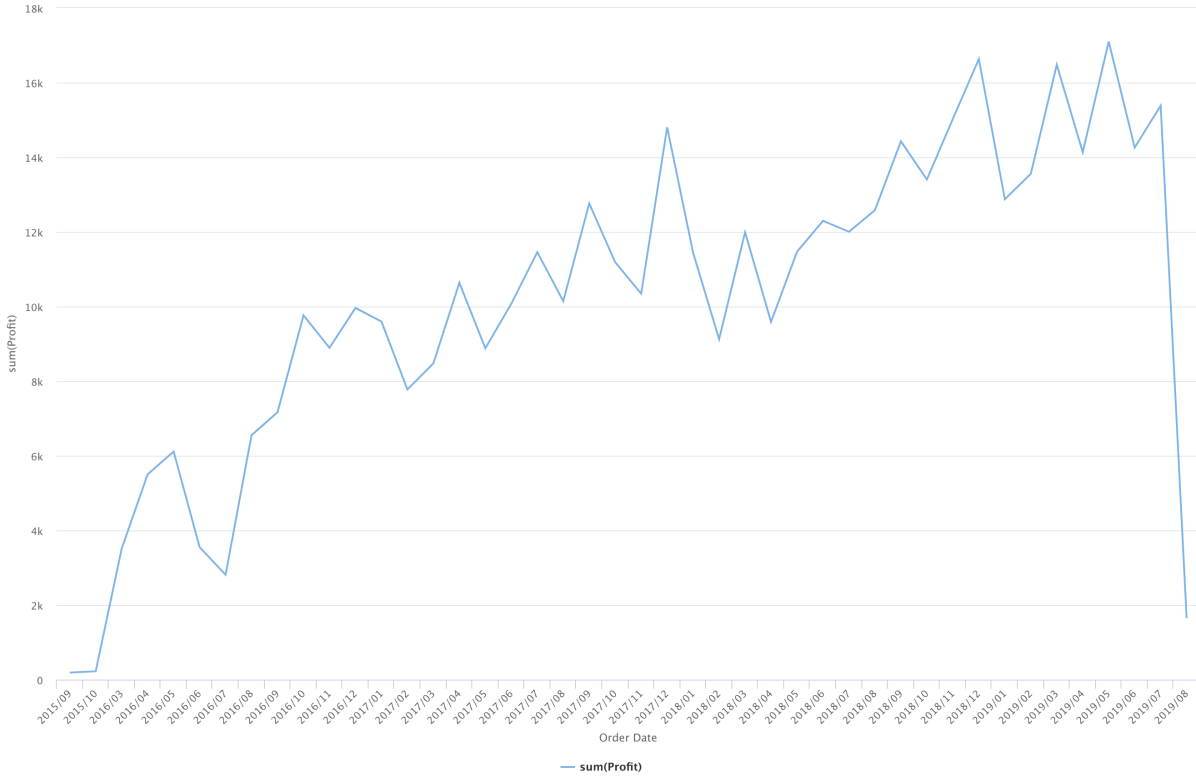
TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *g) Finally, you visualize the following KPIs to understand the performance of the current business: Total price per order, Average price per order, and Total Price per Month. Can you identify any trends?*

Solution:

- Use the “Generate Attributes” operator to create a variable that produces Quantity * Product Price and give it some name (e.g., “Profit”)
- **Total Price per Order / Average Price per Order:**
  - Use the “Aggregate” operator to aggregate the generated attribute by the Order Number using “sum” or “average”
- **Total Price per Month:**
  - Use the “Date to Nominal” operator to transform the date-time variable to a variable indicating only the year and month (*yyyy/MM*)
  - Use the “Aggregate” operator to aggregate the generated attribute by Year/Month using “sum”
- **Trends:** On the right, you can see for example, that the profit increase over the last months and that there are some seasonal trends





Exercise 1.3: Titanic (1/3)

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *a. Is there any missing data in the dataset? If yes, how much?*

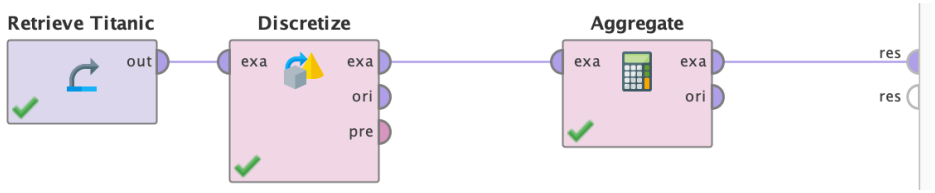
➢ **Solution:**

1. You can use the *Filter Examples* operator to only derive data points that hold missing information. Simply select *missing_attributes* in the settings of the *Filter Examples* operator.
2. There is quite a lot of incomplete data: **1,129 data points**. However, most of it is missing in the *Cabin* and *Life Boat* features.

Task: *b. How many people are 1-19, 20-45, 46-65 and >65 years old?*

➢ **Solution:**

- Use the “Discretize by user specification” operator to define the ranges
- Use the “Aggregate” operator to count the age classes



Output:

Row No.	Age	count(Age)
1	1-19	225
2	20-45	666
3	46-65	145
4	>65	10
5	?	0

Exercise 1.3: Titanic (2/3)

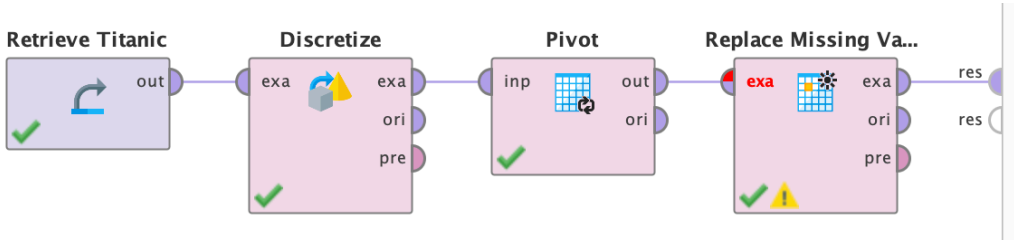
WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *c. Furthermore, you want to understand how many people died and survived. Thus, you create a tabular output that indicates for each of the above age categories the actual number of people who died and who survived as well as the survival ratio of each age category. Please explore and explain the result.*

➢ **Solution:**

- Use the “Pivot” operator with Age and Survived as group by attributes, Age as column grouping attribute, and as aggregation attributes Age with count(percentage) and Age with count
- To remove the resulting missing values, you can use the ”Replace Missing Values” operator to insert zeros instead



Output:

Row No.	Age	Survived	percentage...	percentage...	percentage...	percentage...	percentage...	count(Age)...	count(Age)...	count(Age)...	count(Age)...	count(Age)_?
1	20-45	Yes	0	38.889	0	0	0	0	259	0	0	0
2	20-45	No	0	61.111	0	0	0	0	407	0	0	0
3	1-19	Yes	47.111	0	0	0	0	106	0	0	0	0
4	1-19	No	52.889	0	0	0	0	119	0	0	0	0
5	46-65	Yes	0	0	41.379	0	0	0	0	60	0	0
6	46-65	No	0	0	58.621	0	0	0	0	85	0	0
7	>65	No	0	0	0	80	0	0	0	0	8	0
8	>65	Yes	0	0	0	20	0	0	0	0	2	0
9	?	No	0	0	0	0	0	0	0	0	0	0
10	?	Yes	0	0	0	0	0	0	0	0	0	0

Exercise 1.3: Titanic (3/3)

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *d. Which features appear to explain the chances of survival best? Modify your operators to test changes in the survival rates.*

- **Solution:** E.g., the Passenger class has a relatively high explanatory power as it splits the persons who survived and died quite well

Row No.	Survived	Passenger ...	percentage...	percentage...	percentage...	count(Age)...	count(Age)...	count(Age)...
1	Yes	First	63.732	0	0	181	0	0
2	Yes	Second	0	44.061	0	0	115	0
3	Yes	Third	0	0	26.148	0	0	131
4	No	First	36.268	0	0	103	0	0
5	No	Second	0	55.939	0	0	146	0
6	No	Third	0	0	73.852	0	0	370

Task: *e.* Which model would you choose to predict the chances of a person's survival? Why?

- **Solution:**
  - A high-transparent model like, e.g., a decision tree, as we are interested in understanding the reasons for survival and death

Exercise 1.4: Customer Credit Risk Analysis (1/5)

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *a. Into which type of machine learning problem may this be translatable into?*

- **Solution:** Classification

Task: *b. What kind of data do you need to create such an AI?*

- **Solution:** Labeled data that describes credit assignments for customers who received and not received a credit

Task: *c. How many examples are contained in the dataset?*

- **Solution:** 988

Task: *d. Does the dataset contain a feature which you could use as a label for training your AI? If yes, which one is it?*

- **Solution:** creditworthy

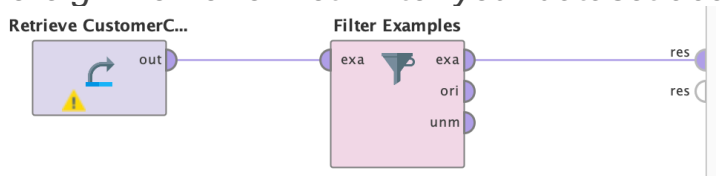
Task: *e. How many requests of non-foreign worker are contained in the data? How many are unknown?*

- **Solution:**
  - **Non-foreign:** 36
  - **Unknown**: 11

Exercise 1.4: Customer Credit Risk Analysis (2/5)

Task: *f. You decide to exclude examples for which it is not known whether the person is a foreign worker or not. Filter your dataset accordingly.*

- **Solution:** Filter with “does not equal” for “Unknown” in Foreign Workers variable



Task: *g. How many requests for granting a credit were made to buy a car? How many were successfully granted for buying a used car?*

- **Solution:** Buying a new car (230) + Buying a used car (101) = 431

Exercise 1.4: Customer Credit Risk Analysis (3/5)

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *h-i. First, you split your filtered data into a training set (containing 70% of the data) and a test set (containing 30% of the data) using stratified sampling. What does stratified sampling do?*

➢ **Solution:**

- *Stratified Sampling* builds random subsets and ensures that the class distribution in the subsets is the same as in the whole example set
- Use the “Split Data” operator to create 2 subsets with “Stratified Sampling” with a ratio of 0.7 and 0.3

Exercise 1.4: Customer Credit Risk Analysis (4/5)

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

Task: *h-ii. Then, you* train a *decision tree* based on the training data set. Why would be a good idea to apply an algorithm like a *decision tree* for this use case?

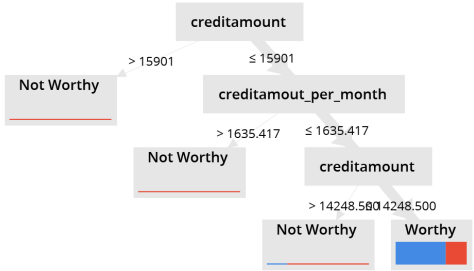
➢ **Solution:**

- A Decision Tree is a highly transparent algorithm as it explicitly shows the rules of how the output is computed which is required for comprehending why certain people got credits and why others did not get any
- Use the 70% dataset (i.e., your training data) for training a Decision Tree. To achieve this, you can use the “Decision Tree” operator

Task: *h-iii. Finally, you apply your trained decision tree to your test data and compute its performance. Is your trained model a good model?*

➢ **Solution:**

- Use the “Apply Model” operator to apply the trained Decision Tree to the remaining 30% sample (i.e., your test data)
- Use the “Performance” operator to compute, e.g., its accuracy and the confusion matrix
- It is not really a good model, as it does not have a high accuracy and creates a lot of false positives



accuracy: 69.62%

	true Worthy	true Not Worthy	class precision
pred. Worthy	204	88	69.86%
pred. Not Worthy	1	0	0.00%
class recall	99.51%	0.00%	

Exercise 1.4: Customer Credit Risk Analysis (5/5)

WIRTSCHAFTS INFORMATIK

TECHNISCHE UNIVERSITÄT DARMSTADT

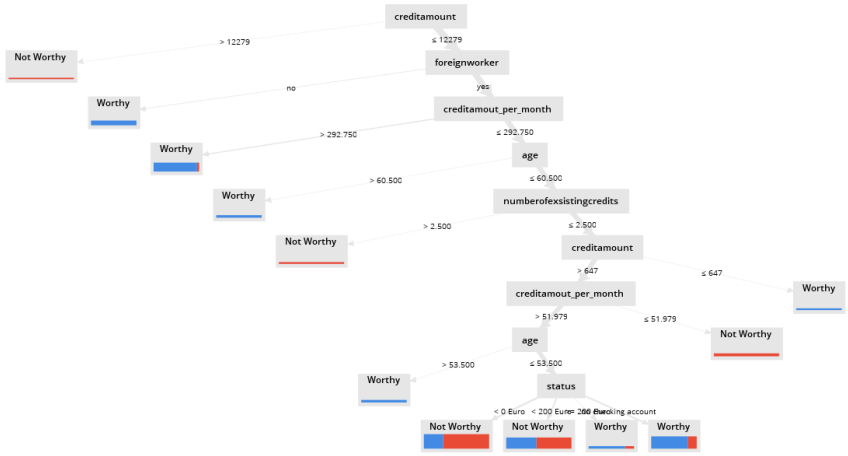
Task: *i. You* decide to train the same AI that only focuses on credit requests for buying new cars. Is the resulting decision tree better or worse than the previous one?

➢ **Solution:**

- **No, the model is even worse! It not only has a worse accuracy, the resulting tree is even way more complex than the previous one.**

accuracy: 63.24%

	true Worthy	true Not Worthy	class precision
pred. Worthy	25	8	75.76%
pred. Not Worthy	17	18	51.43%
class recall	59.52%	69.23%	



Important Take-Away:

You should always choose the simpler model if its performance is comparable to the more complex alternative to avoid possible overfitting!
(See "*Occam's Razor*")
(e.g.: here)