UNIVERSITY OF SANTO TOMAS

Data Preparation

using

Altair® AI Studio

Make sure that you have the two files:

CustomerDetails.xls

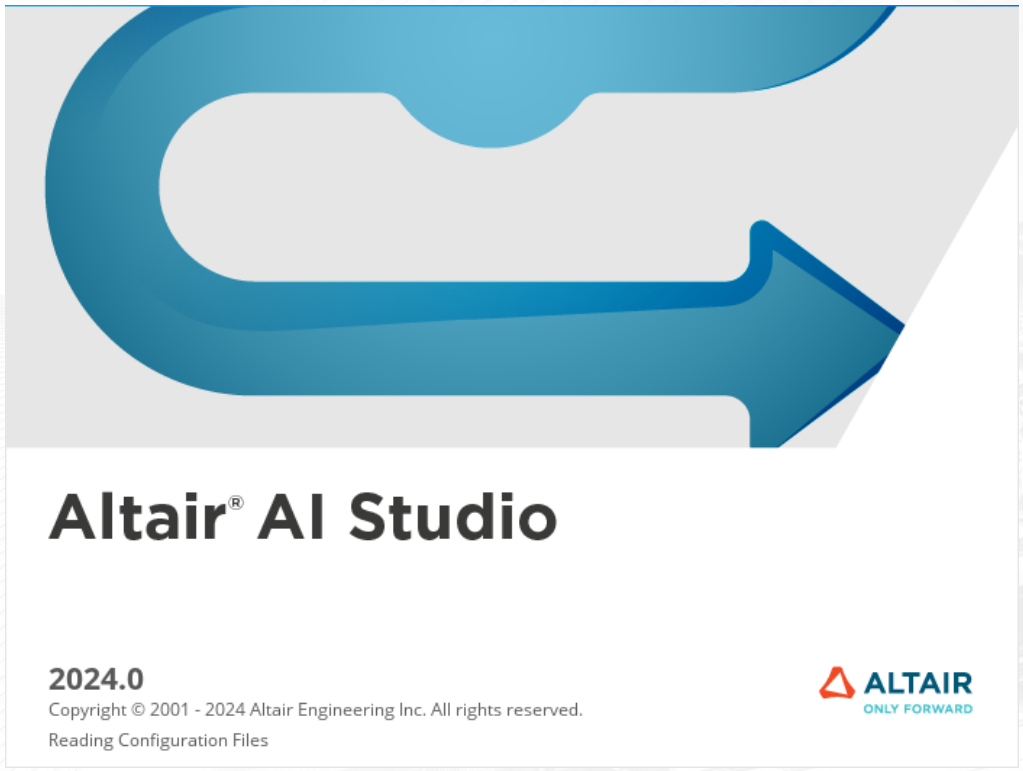
	A	B	C	D	E	F	G
1	Customer ID	Responder	First Name	Last Name	Sex	Z9_Latitude	Z9_Longitude
2	2	No	JEAN	SMITH	female	14.55780655	121.0791117
3	3	No	JULIA	CARRERA	female	14.50814636	121.1537586
4	6	No	H	MACK	male	14.72748316	121.0204381
5	9	No	VIVIAN	GAULDEN	female	14.60740905	120.9738968
6	10	No	PAMELA	WRIGHT	female	14.63883519	121.0540112
7	11	No	MARIA	GONZALEZ	FEMALE	14.54300548	120.9336061
8	12	No	WANDA	MAYBERRY	f	14.51419589	121.1002092
9	14	No	KATHLEEN	KIRBACH	female	14.46343411	121.017539
10	15	Yes	BERNADETTE	MCKALE	female	14.59736204	120.9634643
11	16	No	JEANNINE	QUINLAN	female	14.65834207	121.0245017
12	17		CYNTHIA	LOUY	female	14.5625494	121.068564

OrderDetails.xlsx

	A	B	C	D	E	F	G	H	I
1	Custom	Order_Il	Order_Dat	Store Number	Product_ID	Unit_Price	Discount	Quantity	Response
2	101	88209	4/26/14	100	16474	145.45	0.09	13	1
3	102	16710	4/26/14	103	16474	145.45	0.09	53	1
4	129	86698	11/30/15	101	16124	111.03	0.10	43	1
5	180	90784	11/4/15	100	16238	291.73	0.09	21	1
6	180	90785	12/14/15	100	16566	140.98	0.07	21	1
7	1025	89013	11/9/14	108	16182	284.98	0.08	20	1
8	1027	89016	12/29/15	103	15956	449.99	0.08	20	1
9	1030	89622	4/22/14	100	15653	175.99	0.09	11	1
10	1043	87851	9/8/15	100	16083	270.97	0.06	12	1
11	1056	90215	1/15/13	106	16244	113.98	0.06	12	1
12	1060	58628	3/25/12	105	15872	138.75	0.09	23	1
13	1074	86424	1/27/14	107	16474	145.45	0.06	11	1
14	1106	45824	12/8/12	103	16253	140.81	0.08	81	1
15	1118	86773	11/15/13	102	16361	209.84	0.09	26	1
16	1123	87016	12/22/12	100	15675	175.99	0.09	22	1
17	1129	32037	2/20/12	107	15948	699.99	0.07	15	1
18	1135	87942	12/11/13	106	16197	500.98	0.09	14	1
19	1136	87940	7/4/12	104	16083	270.97	0.09	15	1
20	1138	86577	2/4/13	105	16676	107.53	0.07	12	1
21	1139	86585	5/8/15	108	15775	152.48	0.10	12	1
22	1143	86583	8/23/14	103	16194	300.98	0.10	25	1
23	1182	86916	3/6/14	107	16886	125.99	0.10	11	1
24	1211	88601	10/25/13	109	15694	195.99	0.06	23	1
25	1211	88608	7/3/15	109	16149	165.98	0.10	17	1
26	1217	54595	5/1/12	101	16454	130.98	0.09	41	1

the process of preparing data for analysis by removing or modifying **incorrect**, **incomplete**, **irrelevant**, **duplicated**, or **improperly formatted** data.

Open the **Altair® AI Studio** app installed on your computer.



Terminologies used in Altair AI Studio:

attributes

variables
(represented by columns)
of a given data set

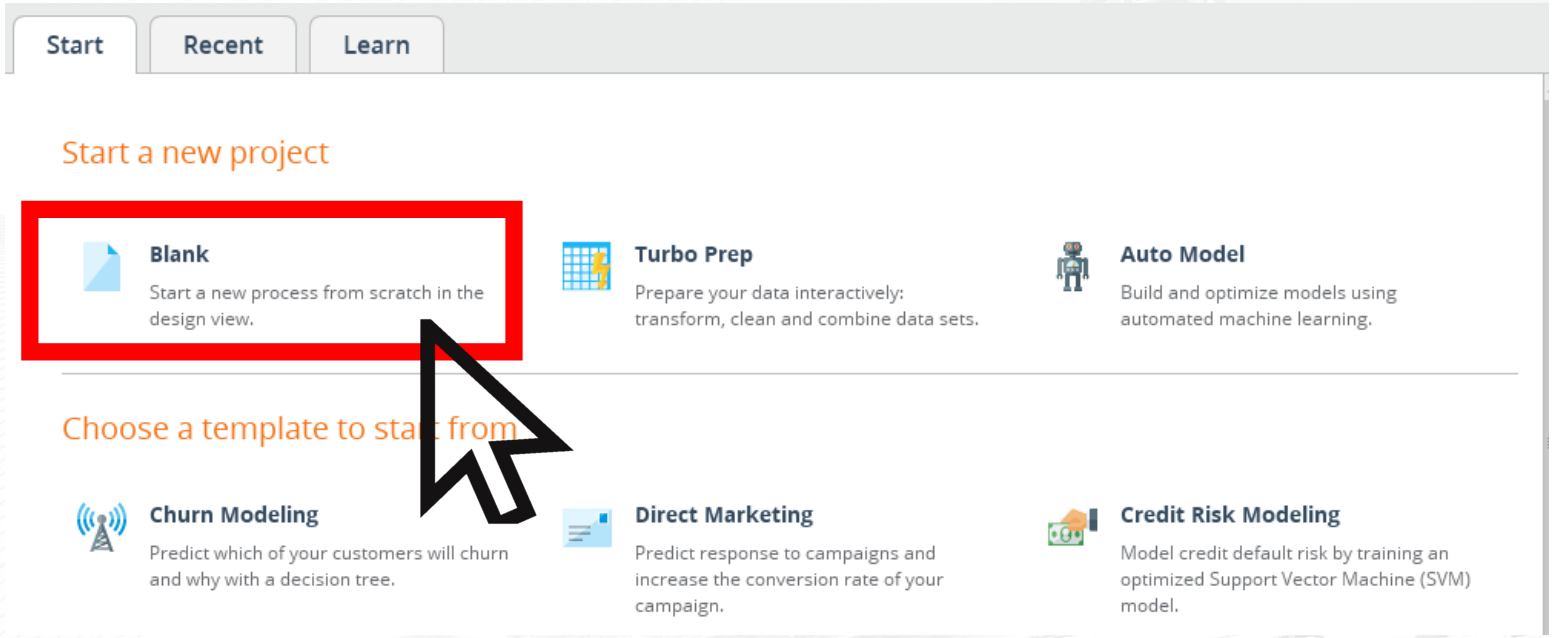
examples

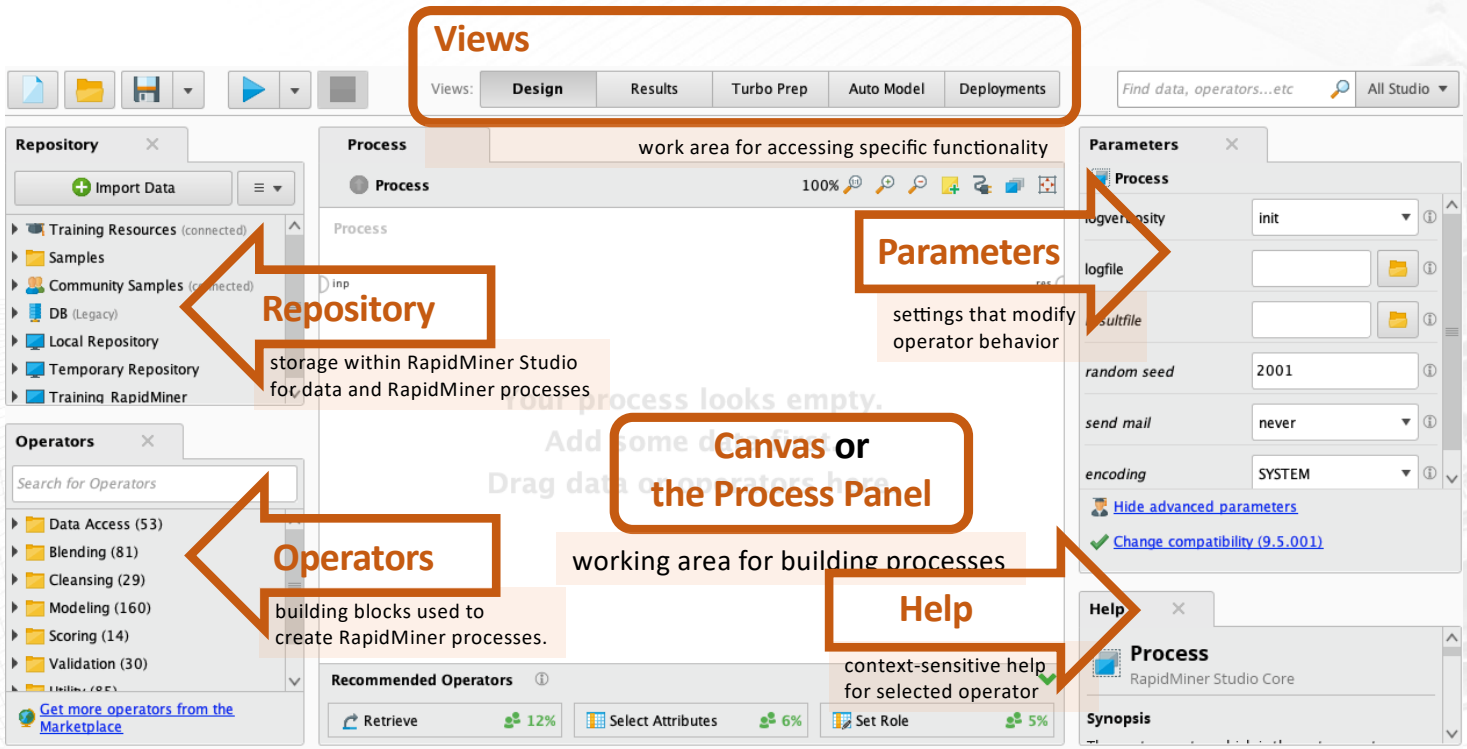
cases or observations
(represented by rows)
of a given data set

operators

functions or
building blocks
that creates
processes

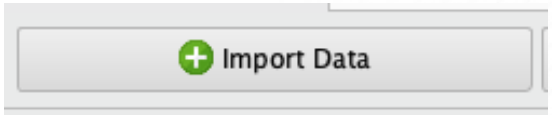
Open the Altair® AI Studio app installed on your computer.



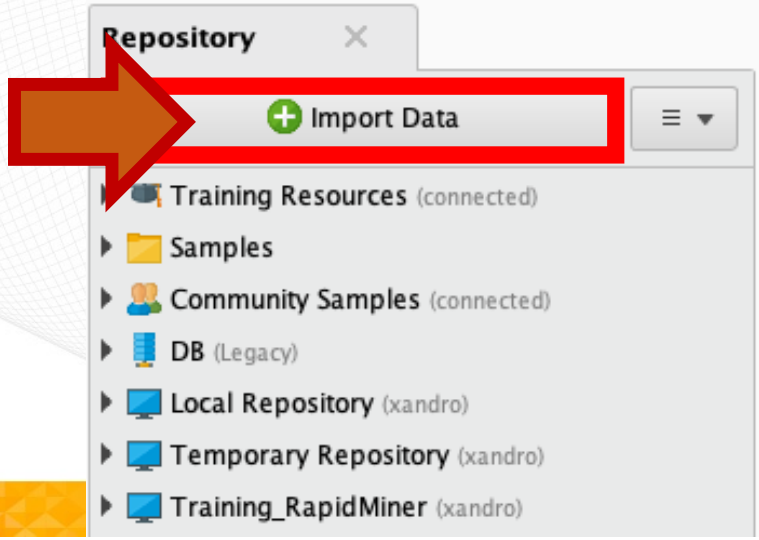

Views
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples
DB (Legacy)
Local Repository
Temporary Repository
Training RapidMiner
Repository
storage within RapidMiner Studio for data and RapidMiner processes
Process
work area for accessing specific functionality
Process
100%
Your process looks empty.
Add some data first.
Drag data or operators here.
Canvas or the Process Panel
working area for building processes
Parameters
Process
logverbosity
init
logfile
resultfile
random seed
2001
send mail
never
encoding
SYSTEM
Hide advanced parameters
Change compatibility (9.5.001)
Parameters
settings that modify operator behavior
Operators
Search for Operators
Data Access (53)
Blending (81)
Cleansing (29)
Modeling (160)
Scoring (14)
Validation (30)
Get more operators from the Marketplace
Operators
building blocks used to create RapidMiner processes.
Recommended Operators
Retrieve 12%
Select Attributes 6%
Set Role 5%
Help
Process
RapidMiner Studio Core
Synopsis
Help
context-sensitive help for selected operator

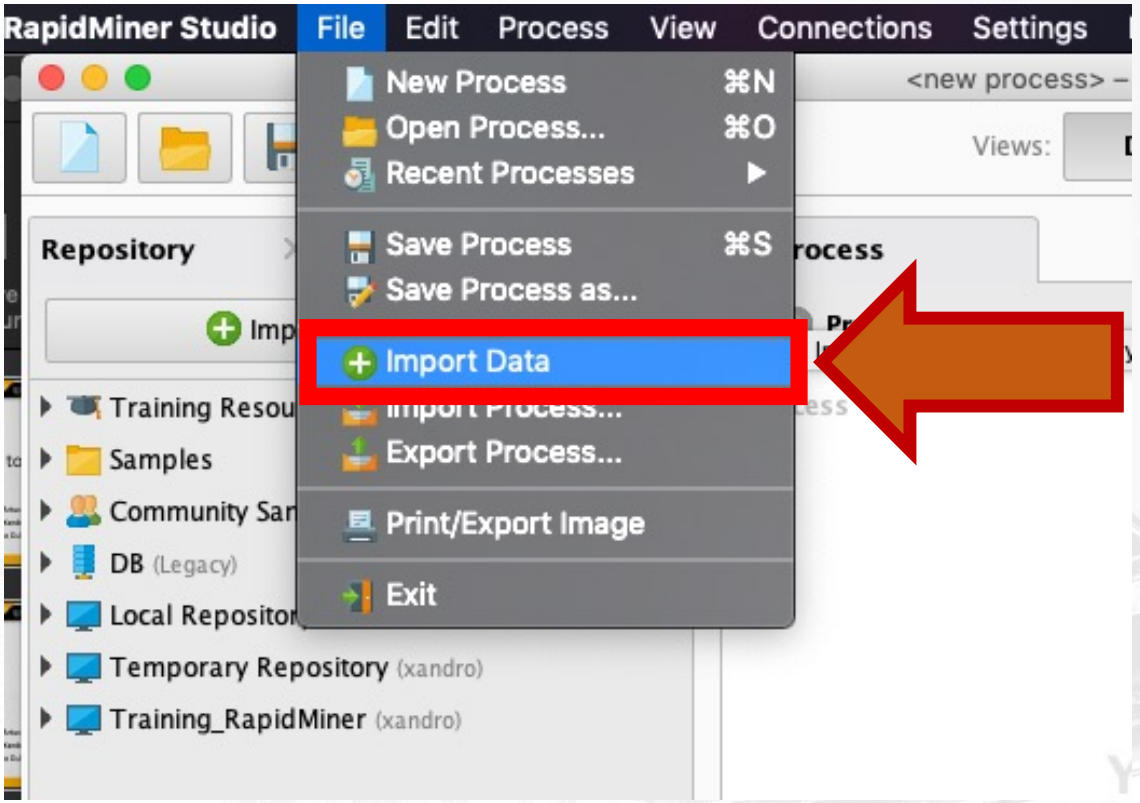
Importing Data

- Click **File** then **Import Data**,

or click 

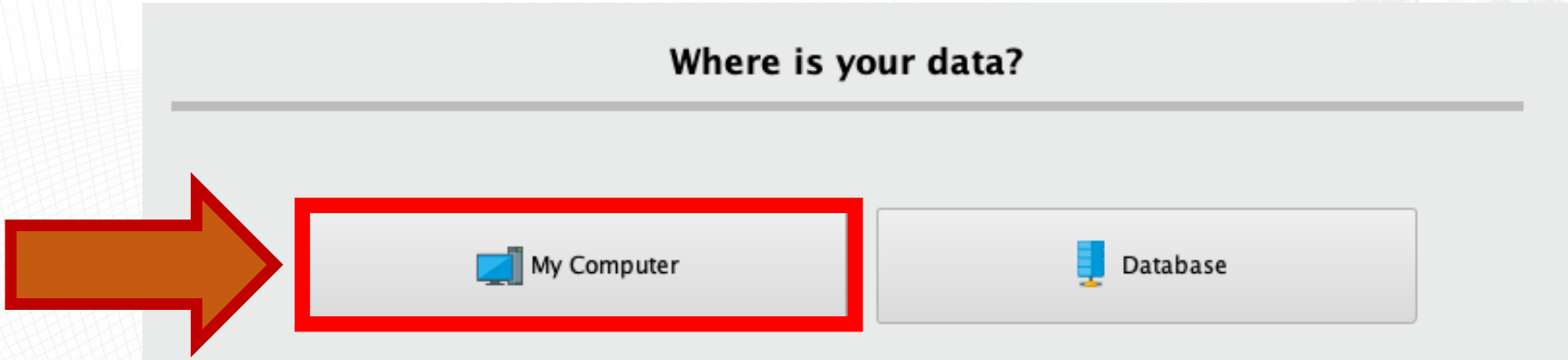
in the Repository tab





Importing Data

- Choose the source of your data set

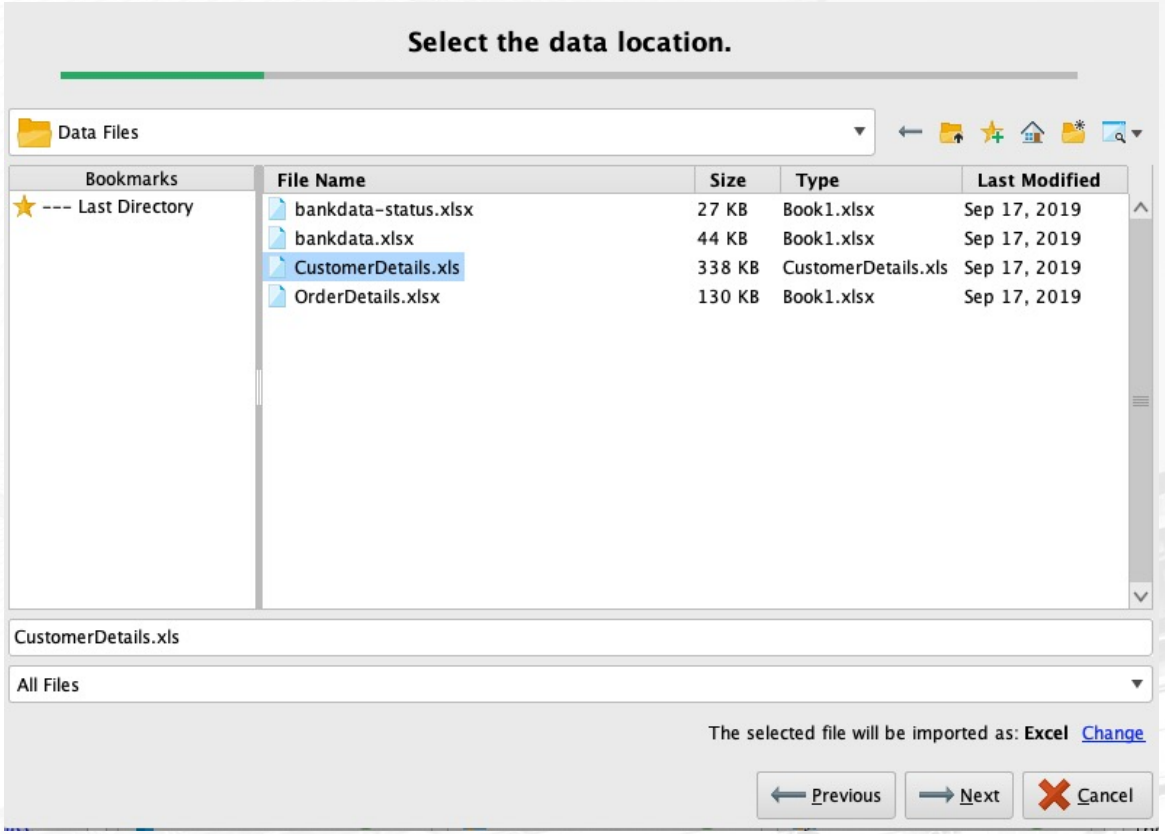


Importing Data

- Locate the data then click **Next**.

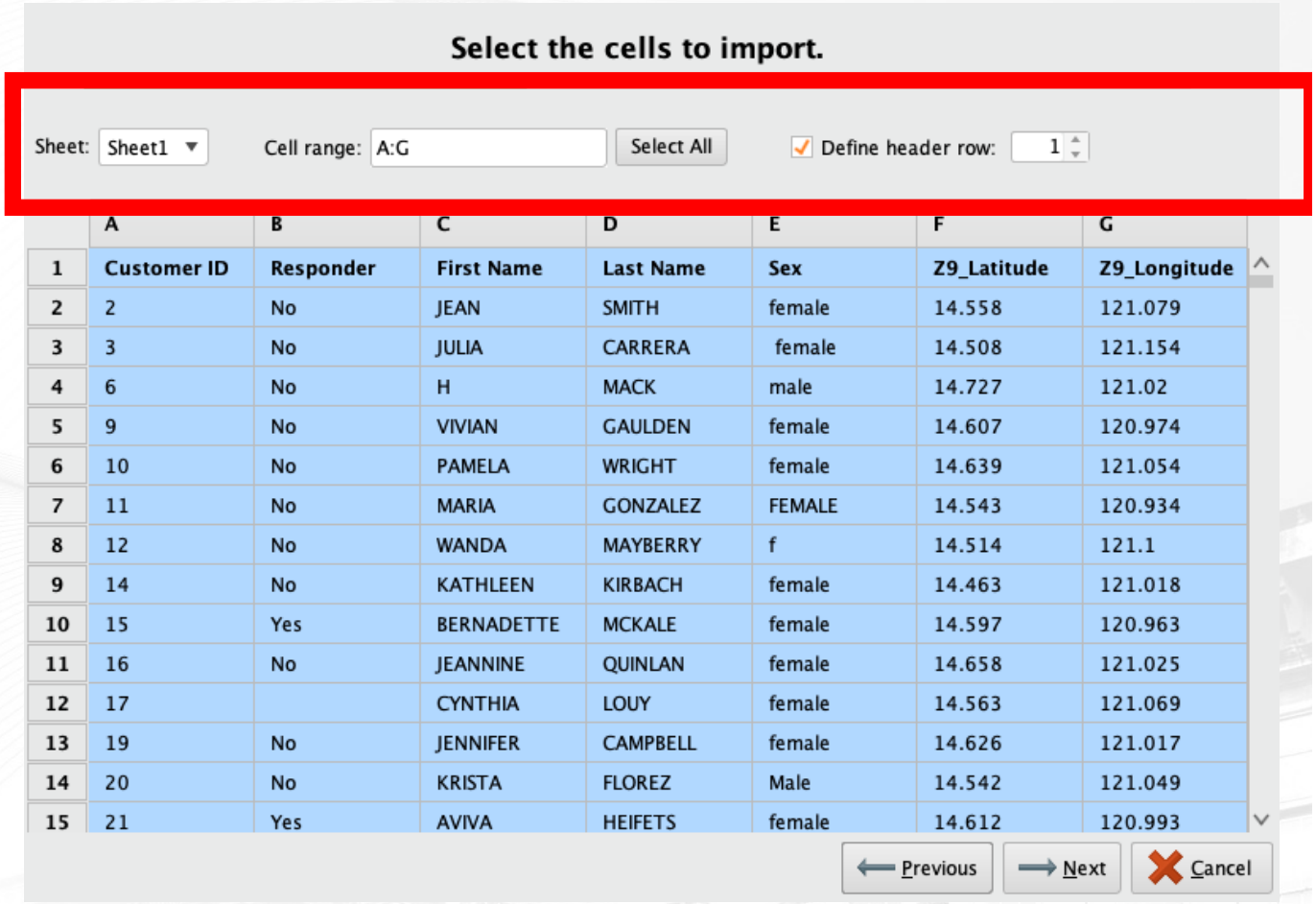
  For this lecture,

  *choose **CustomerDetails.xls**.*



Importing Data

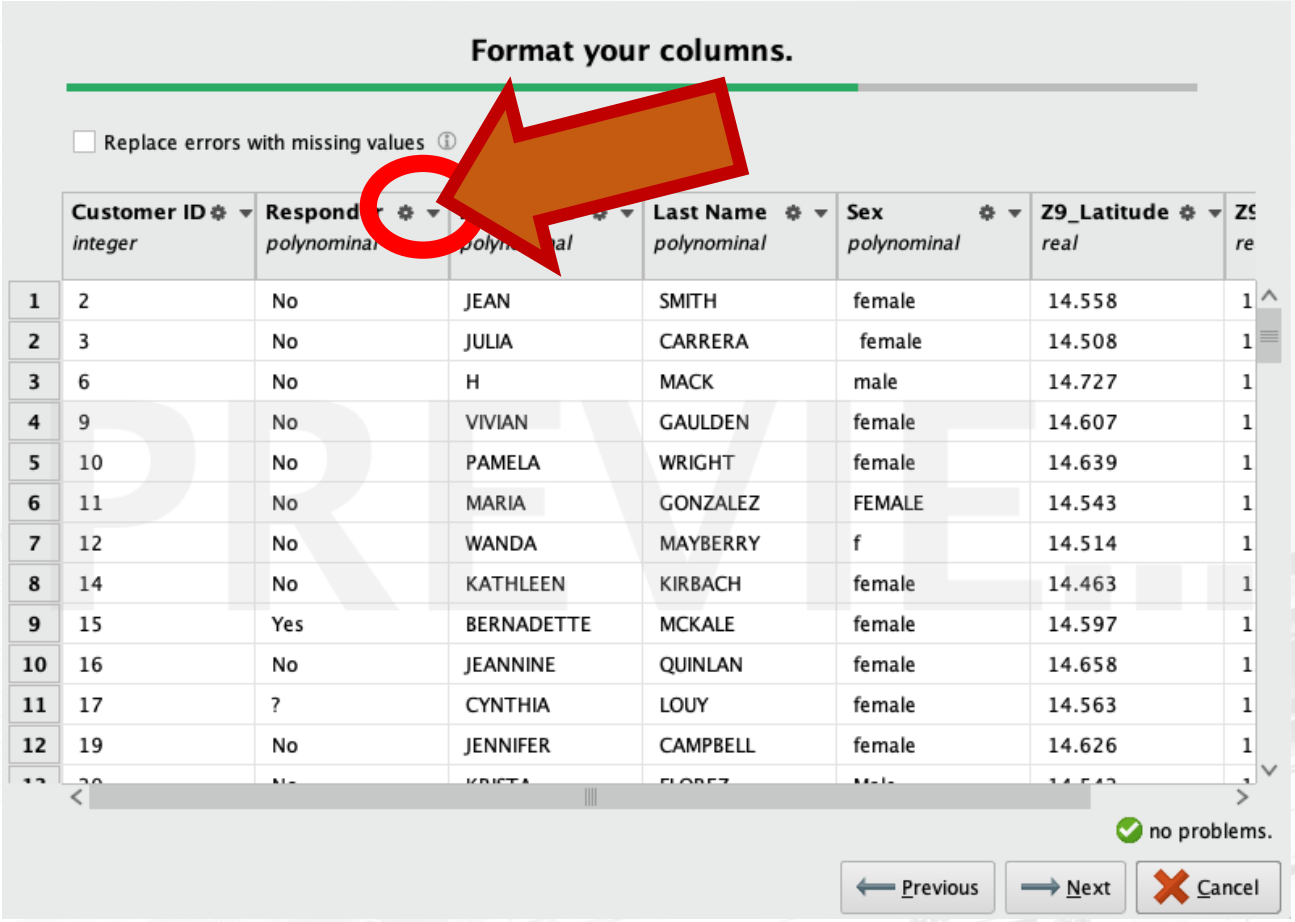
- Verify the cells you want to import and click **Next**.



	A	B	C	D	E	F	G
1	Customer ID	Responder	First Name	Last Name	Sex	Z9_Latitude	Z9_Longitude
2	2	No	JEAN	SMITH	female	14.558	121.079
3	3	No	JULIA	CARRERA	female	14.508	121.154
4	6	No	H	MACK	male	14.727	121.02
5	9	No	VIVIAN	GAULDEN	female	14.607	120.974
6	10	No	PAMELA	WRIGHT	female	14.639	121.054
7	11	No	MARIA	GONZALEZ	FEMALE	14.543	120.934
8	12	No	WANDA	MAYBERRY	f	14.514	121.1
9	14	No	KATHLEEN	KIRBACH	female	14.463	121.018
10	15	Yes	BERNADETTE	MCKALE	female	14.597	120.963
11	16	No	JEANNINE	QUINLAN	female	14.658	121.025
12	17		CYNTHIA	LOUY	female	14.563	121.069
13	19	No	JENNIFER	CAMPBELL	female	14.626	121.017
14	20	No	KRISTA	FLOREZ	Male	14.542	121.049
15	21	Yes	AVIVA	HEIFETS	female	14.612	120.993

Importing Data

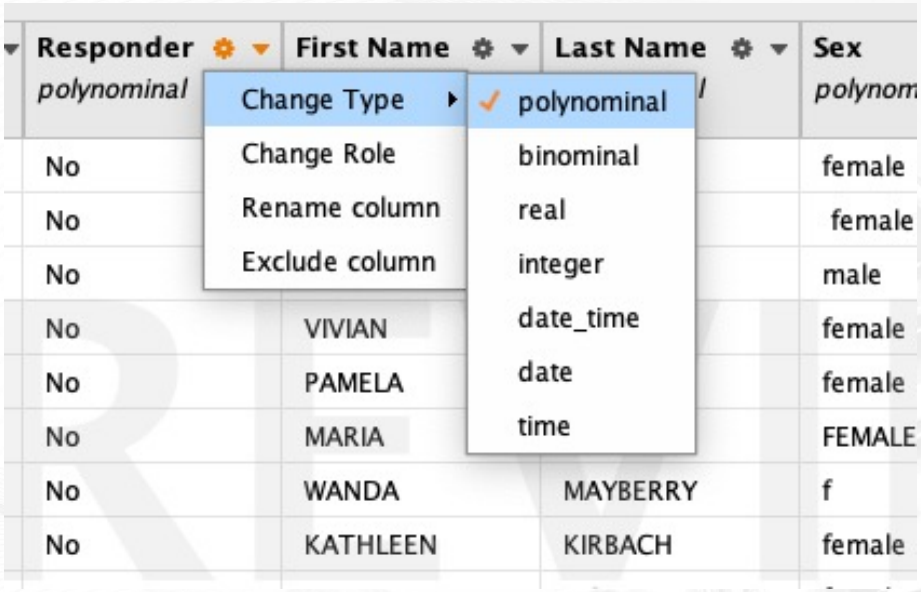
- Format the columns with your specifications.



	Customer ID *integer*	Respond *polynominal*	Last Name *polynominal*	Sex *polynominal*	Z9_Latitude *real*
1	2	No	JEAN	SMITH	female	14.558
2	3	No	JULIA	CARRERA	female	14.508
3	6	No	H	MACK	male	14.727
4	9	No	VIVIAN	GAULDEN	female	14.607
5	10	No	PAMELA	WRIGHT	female	14.639
6	11	No	MARIA	GONZALEZ	FEMALE	14.543
7	12	No	WANDA	MAYBERRY	f	14.514
8	14	No	KATHLEEN	KIRBACH	female	14.463
9	15	Yes	BERNADETTE	MCKALE	female	14.597
10	16	No	JEANNINE	QUINLAN	female	14.658
11	17	?	CYNTHIA	LOUY	female	14.563
12	19	No	JENNIFER	CAMPBELL	female	14.626

Importing Data

- Format the columns with your specifications.
- You may change the type, role, and name of each attribute.



Importing Data

- Types:

polynomial

many different string values (for example: red, green, blue, yellow)

binomial

exactly two values (for example: true/false, yes/no)

real

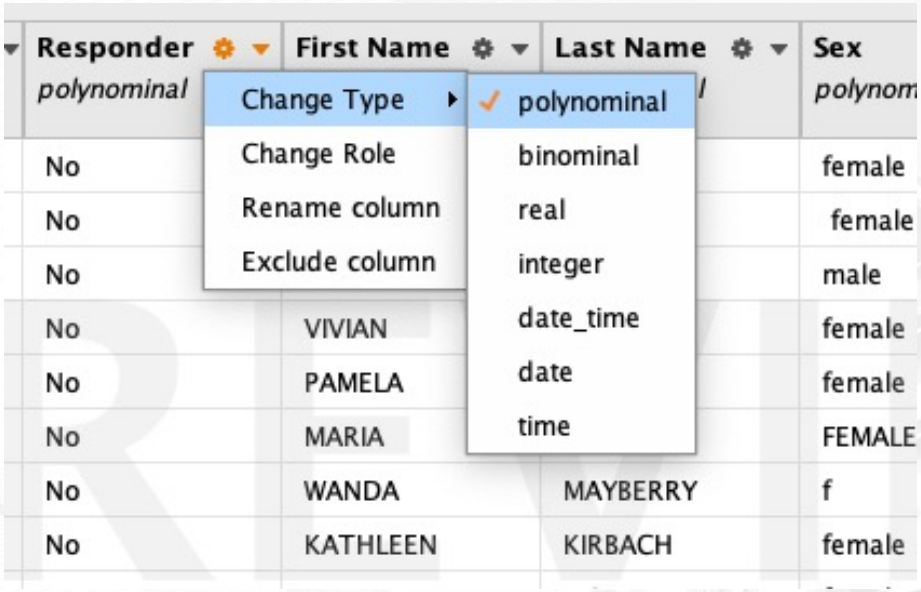
a fractional number (for example: 11.23 or -0.0001).

integer

a whole number (for example: 23, -5, or 11,024,768).

date_time

- both date and time (for example: 23.12.2014 17:59).



date

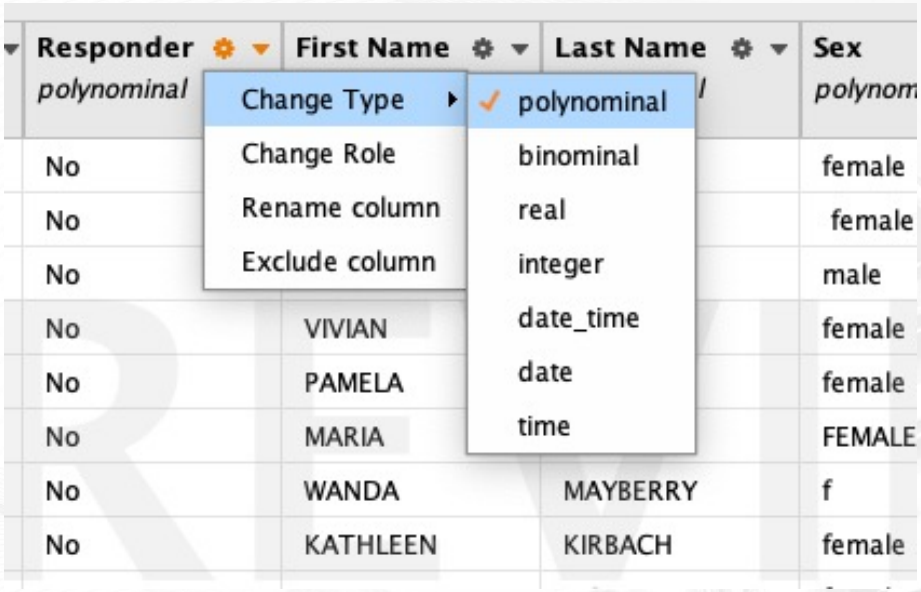
date without time (for example 23.12.2014).

time

time without date (for example 17:59).

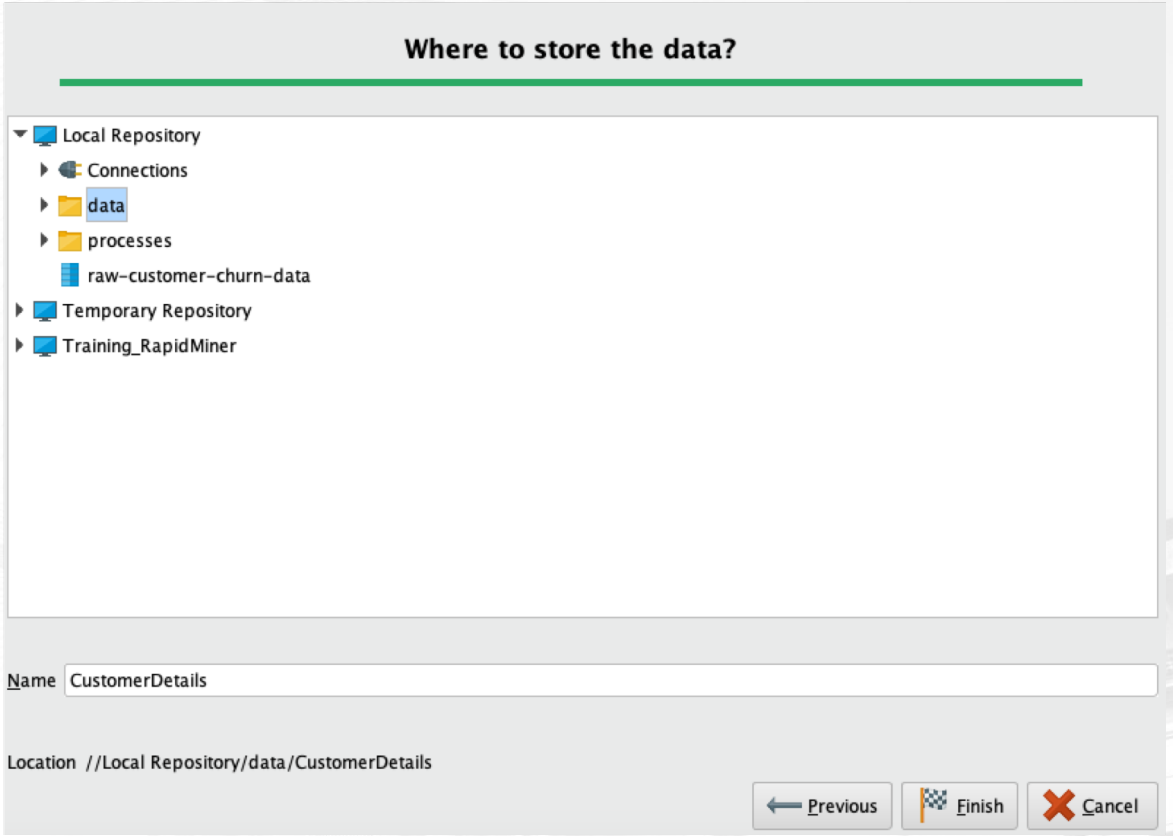
Importing Data

- Format the columns with your specifications.
- You may change the type, role, and name of each attribute.
- Click **Next.**



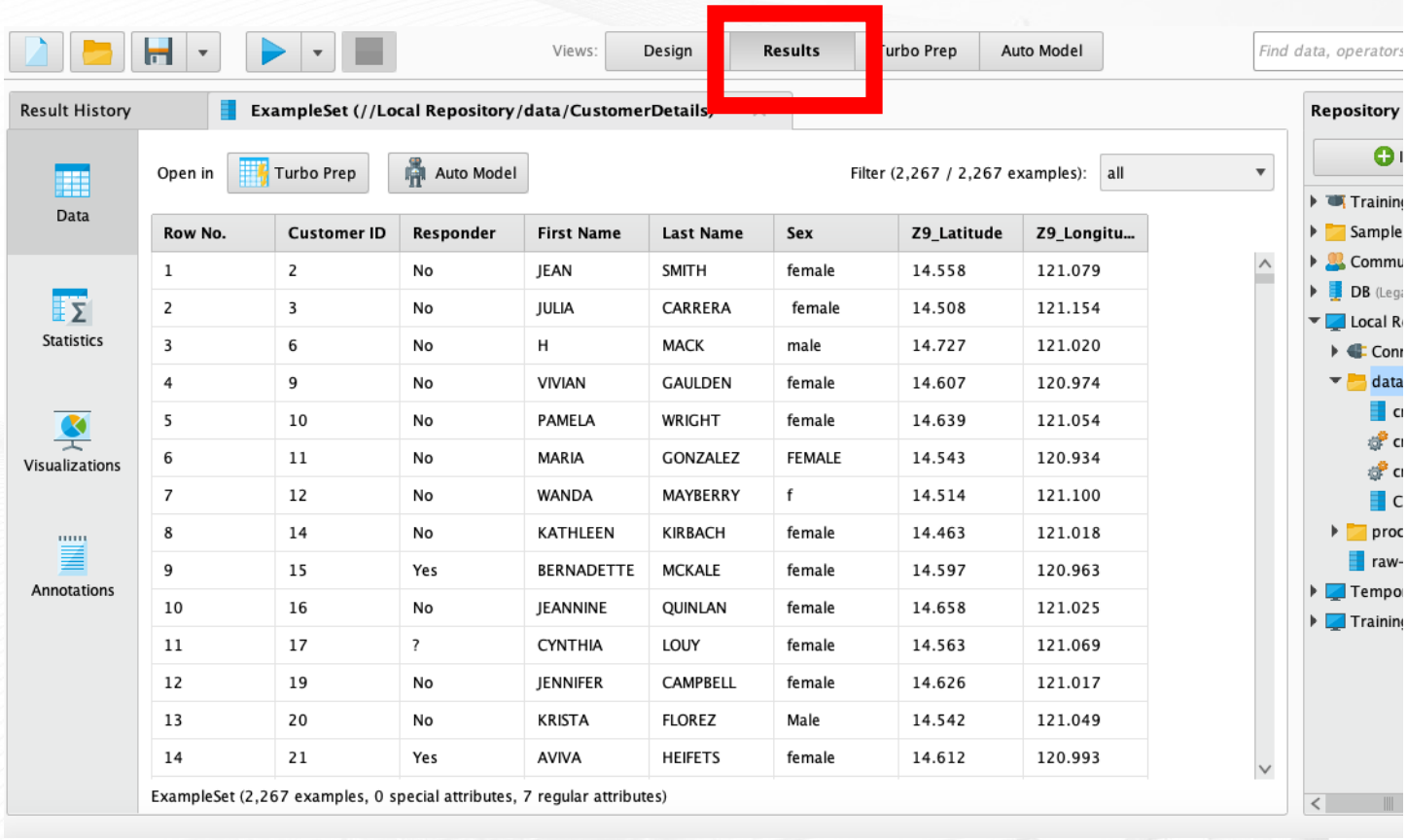
Importing Data

- Choose the folder where the data will be stored.
- Type the file name.
- Click **Finish**.
- The data will appear in the result view.



Importing Data

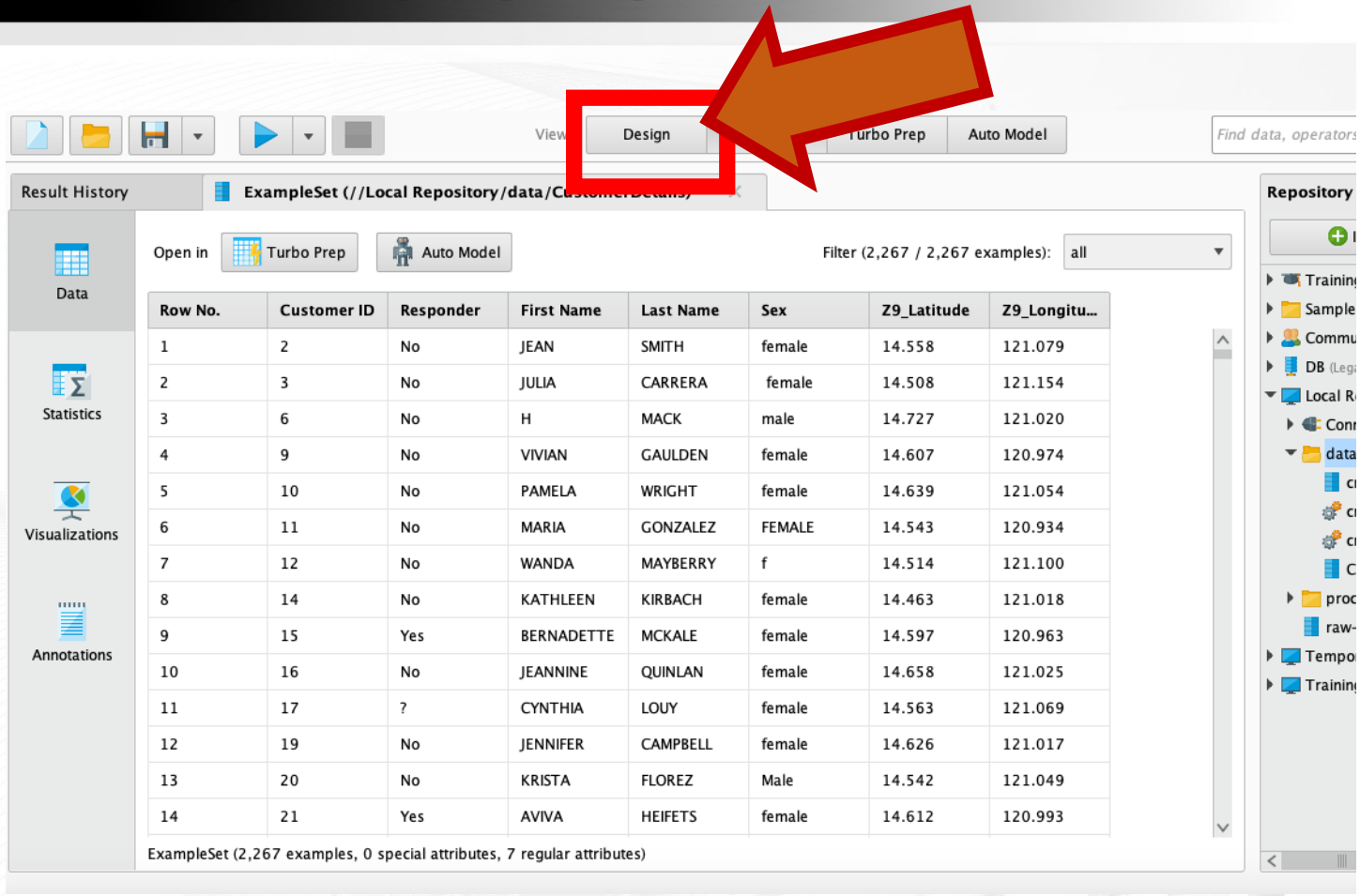
- The data will appear in the **Results** tab.



This time, using a RapidMiner operator.

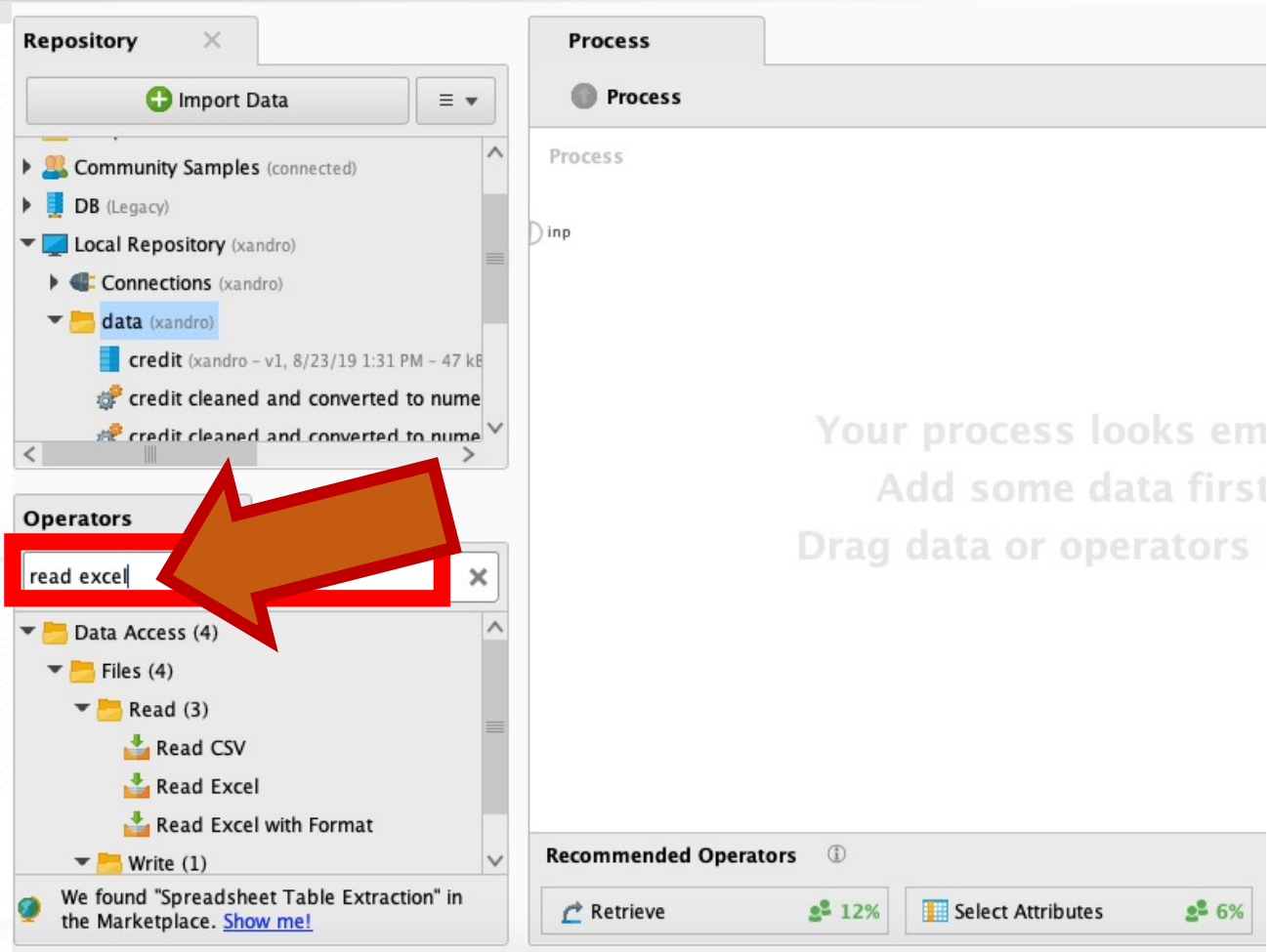
Importing Data

- In the Views tab, click **Design**.



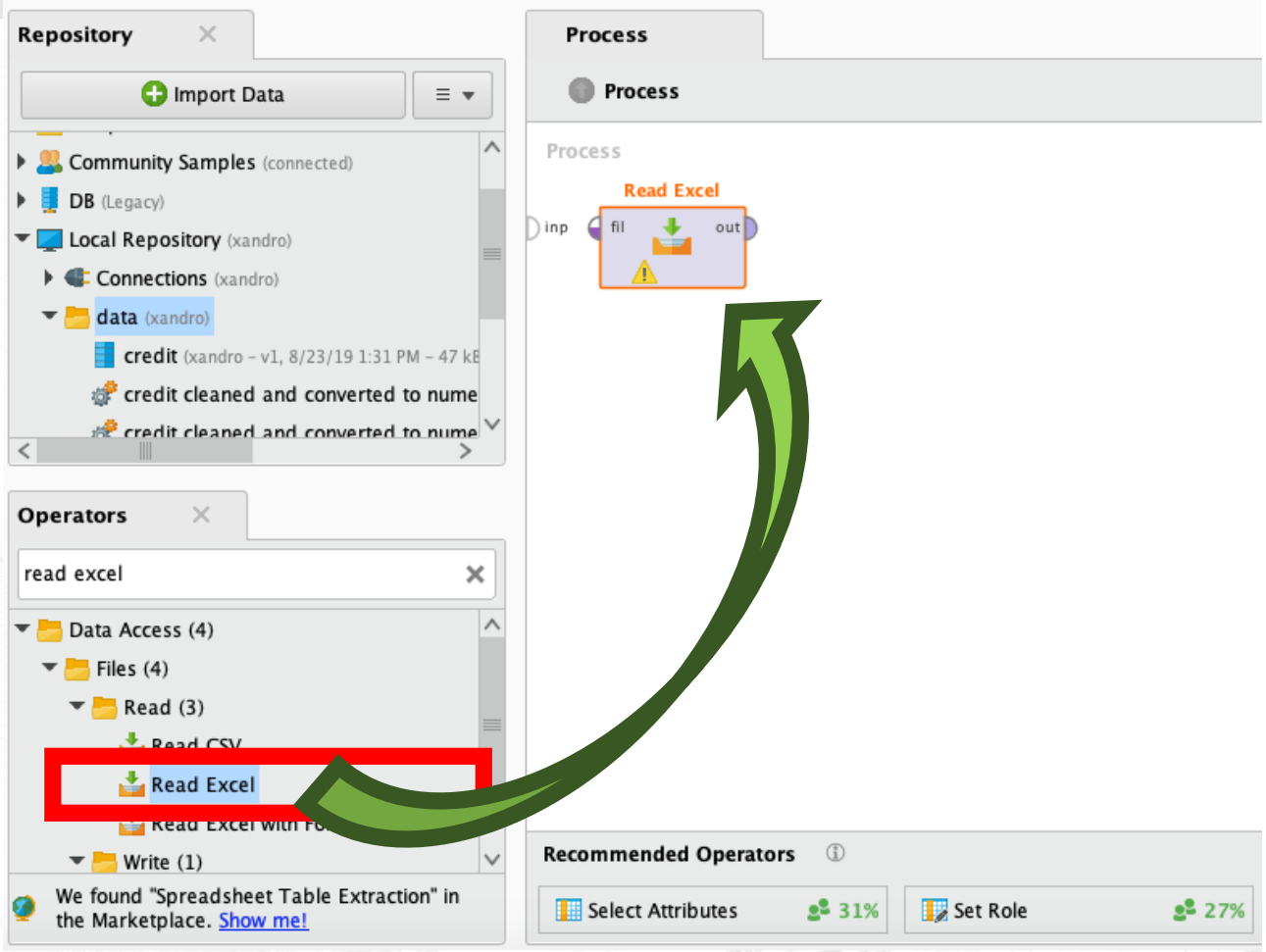
Importing Data

- Search for **Read Excel** in the operator tab.



Importing Data

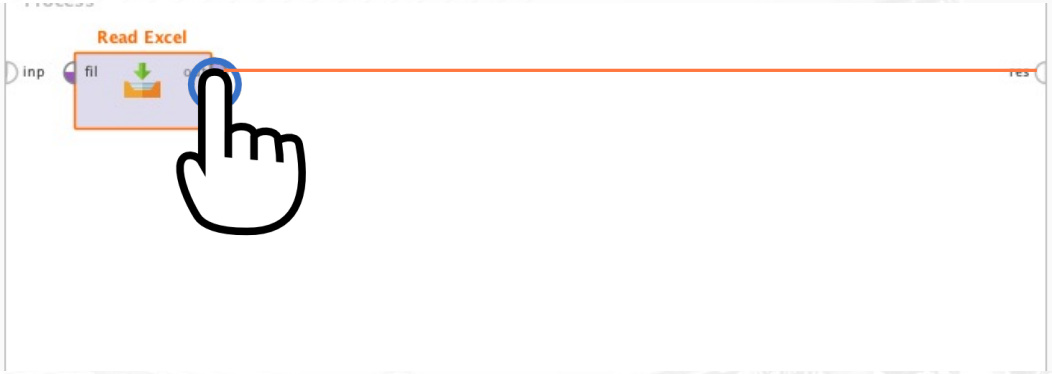
- Search for **Read Excel** in the operator tab.
- Drag and drop it to the canvas.



Data Preparation

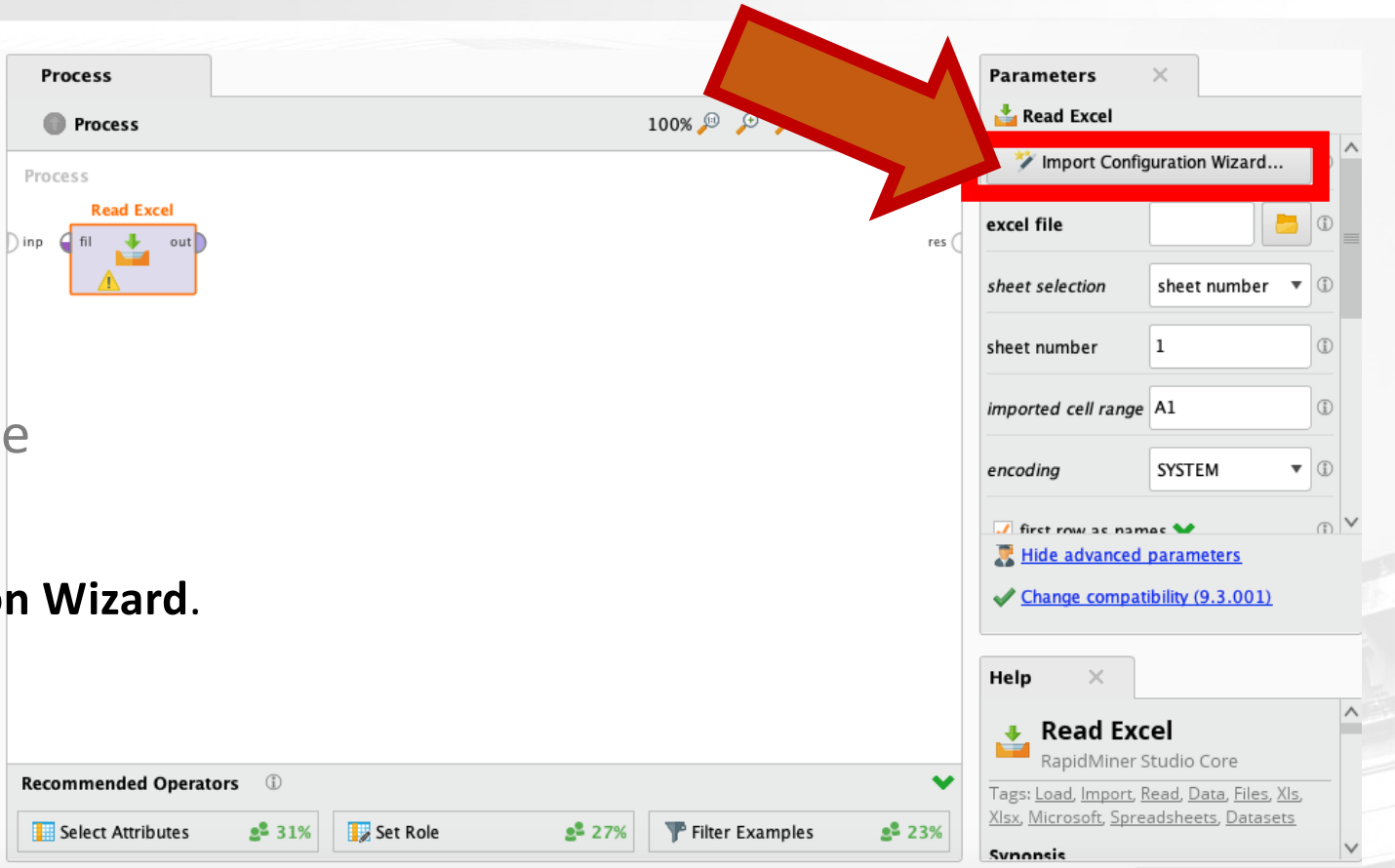
Connect the **Out** node of the **Read Excel** operator and **res** of the result knob.

This will put the **out**put to the **res**ults tab.



Importing Data

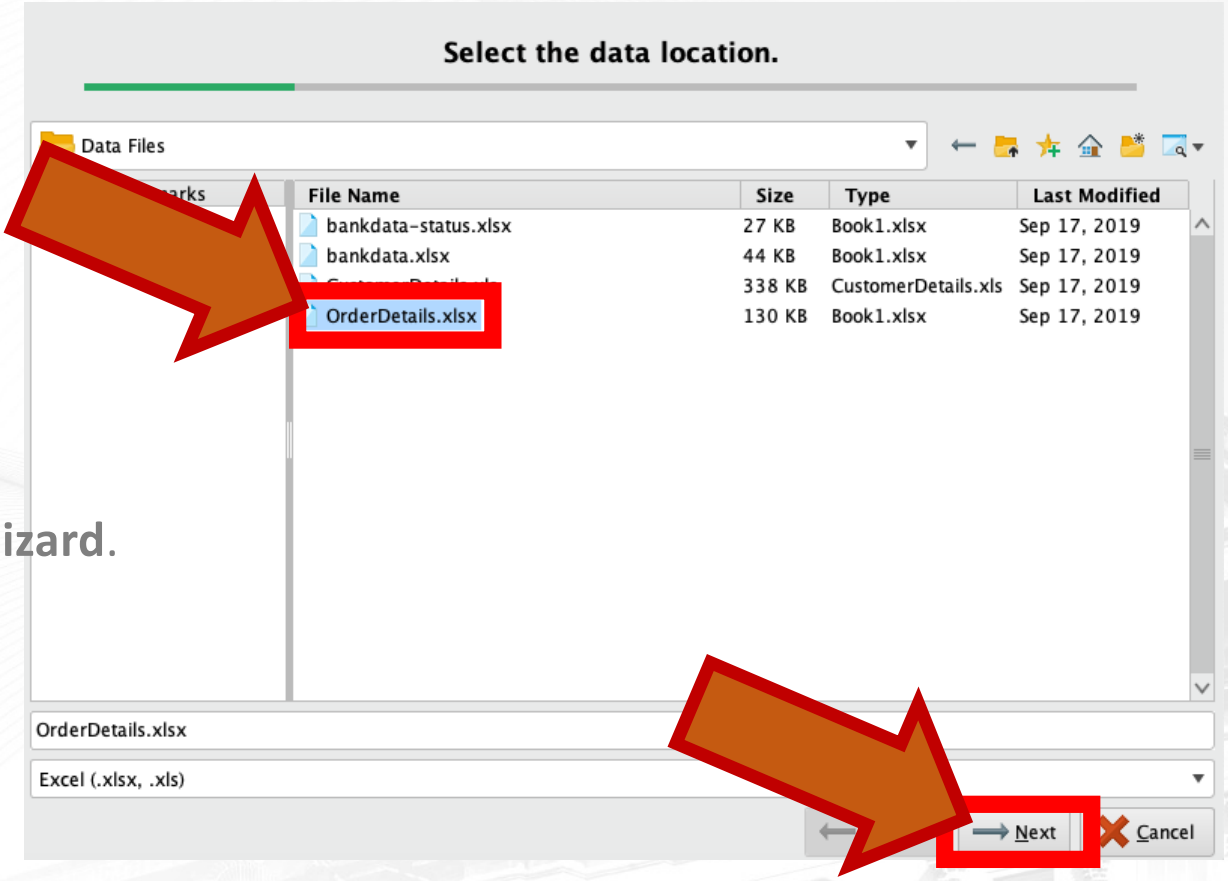
- Search for **Read Excel** in the operator tab.
- Drag and drop it to the canvas.
- Click **Import Configuration Wizard**.



Importing Data

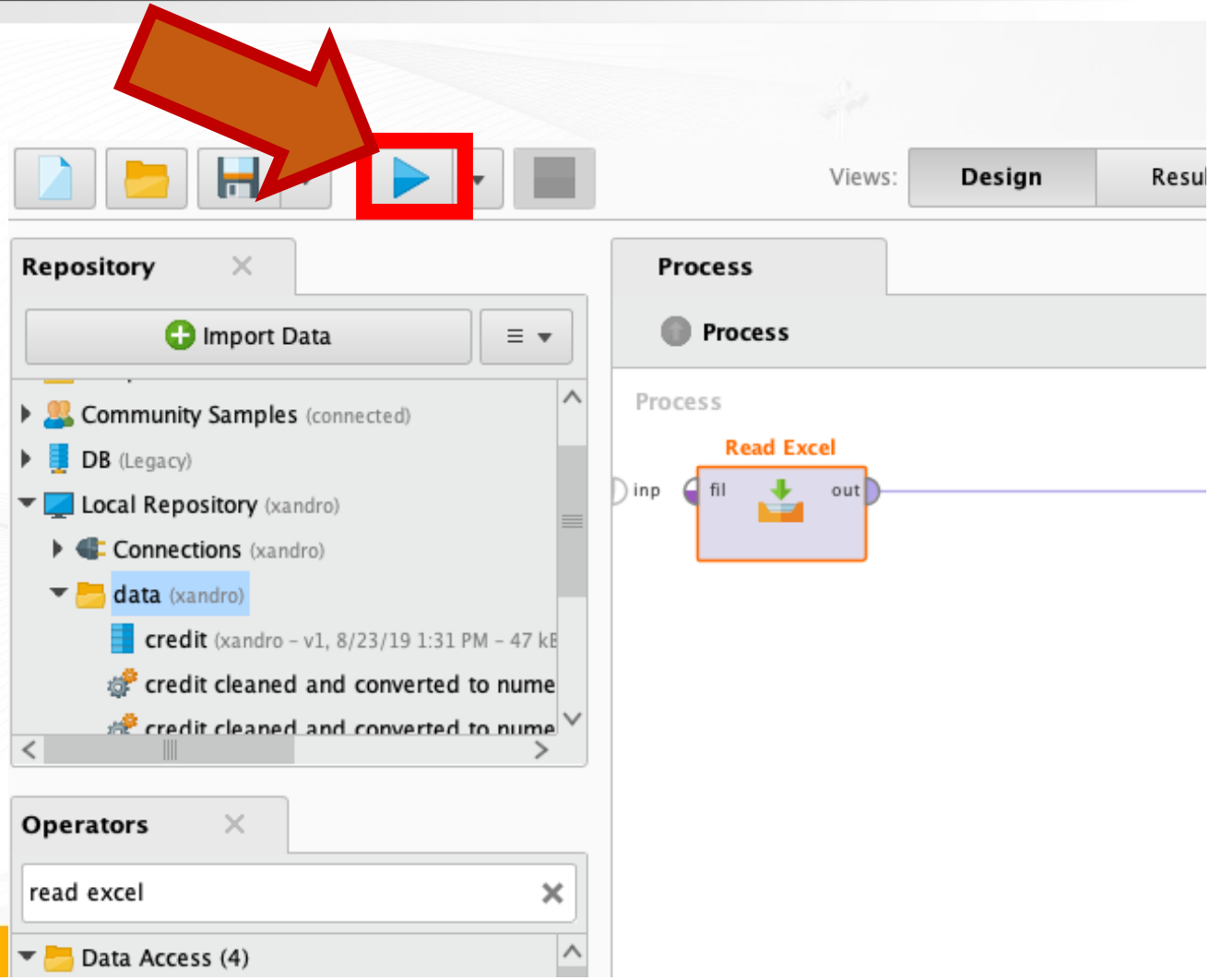
- Search for **Read Excel** in the operator tab.
- Drag and drop it to the canvas.
- Click **Import Configuration Wizard**.
- Locate and open the file.

  For this lecture, choose ***OrderDetails.xls*****.**
- Click **Next, Next,** and **Finish**.

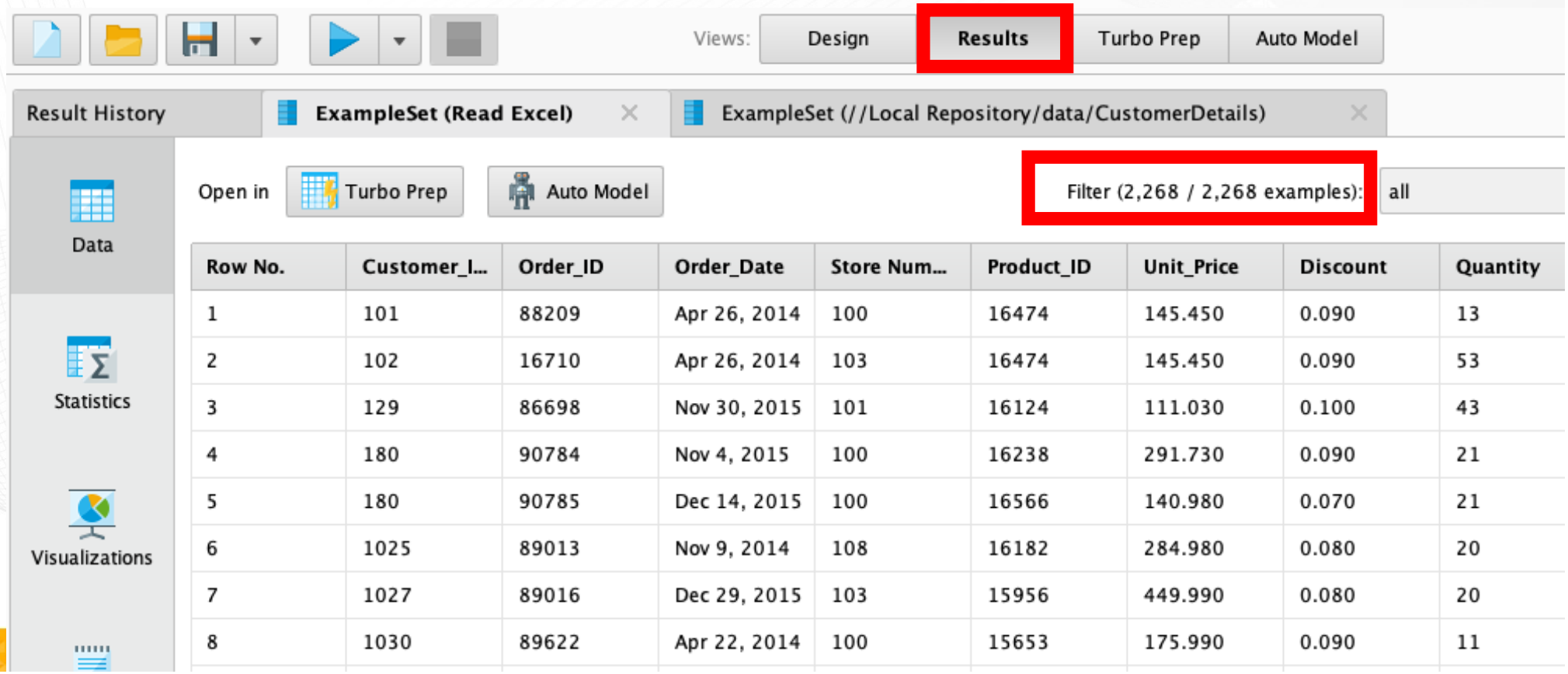


Data Preparation

Click the **Play** button to execute the process.



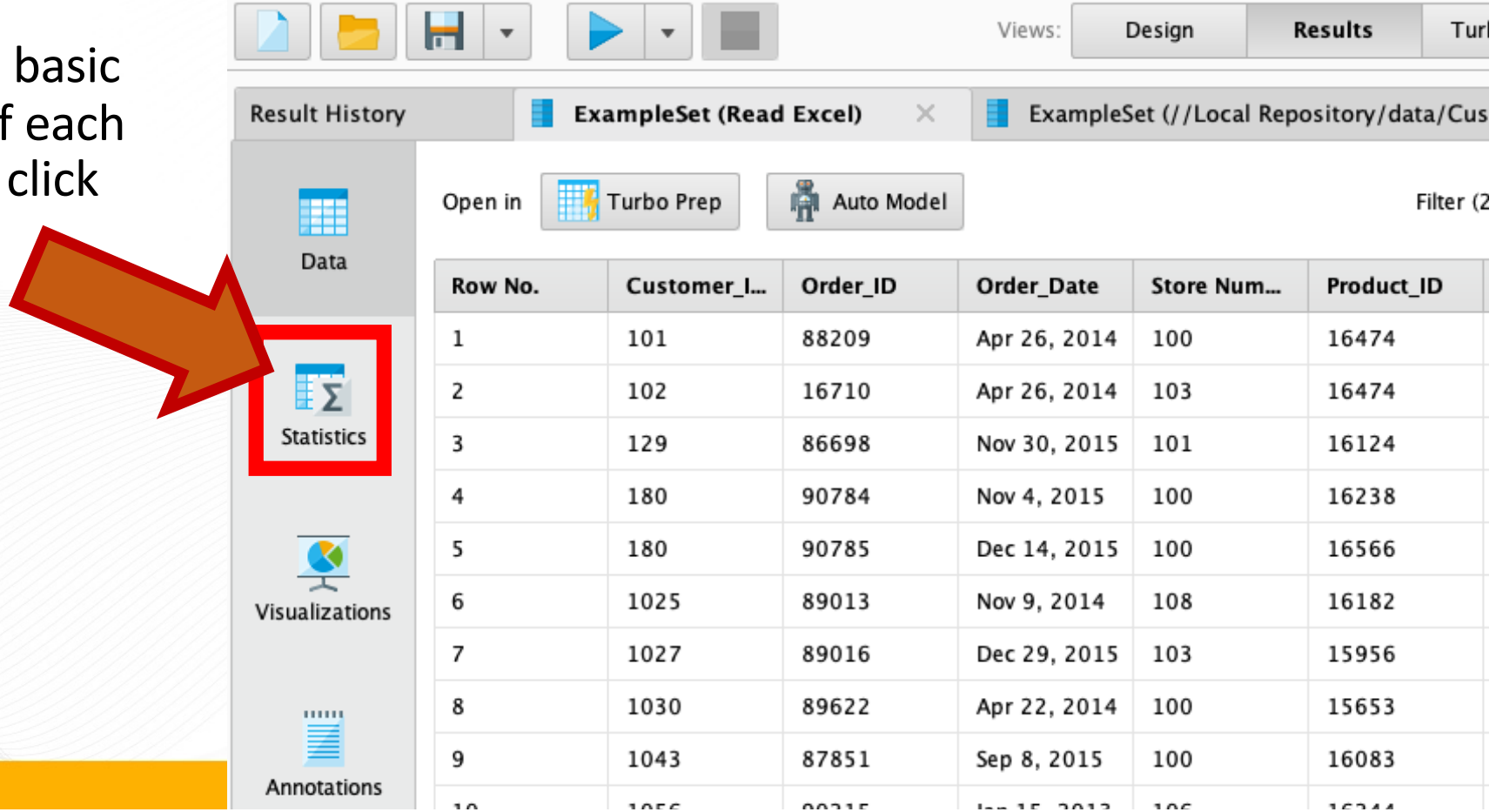
Importing Data



Row No.	Customer_I...	Order_ID	Order_Date	Store Num...	Product_ID	Unit_Price	Discount	Quantity
1	101	88209	Apr 26, 2014	100	16474	145.450	0.090	13
2	102	16710	Apr 26, 2014	103	16474	145.450	0.090	53
3	129	86698	Nov 30, 2015	101	16124	111.030	0.100	43
4	180	90784	Nov 4, 2015	100	16238	291.730	0.090	21
5	180	90785	Dec 14, 2015	100	16566	140.980	0.070	21
6	1025	89013	Nov 9, 2014	108	16182	284.980	0.080	20
7	1027	89016	Dec 29, 2015	103	15956	449.990	0.080	20
8	1030	89622	Apr 22, 2014	100	15653	175.990	0.090	11

Exploratory Analysis

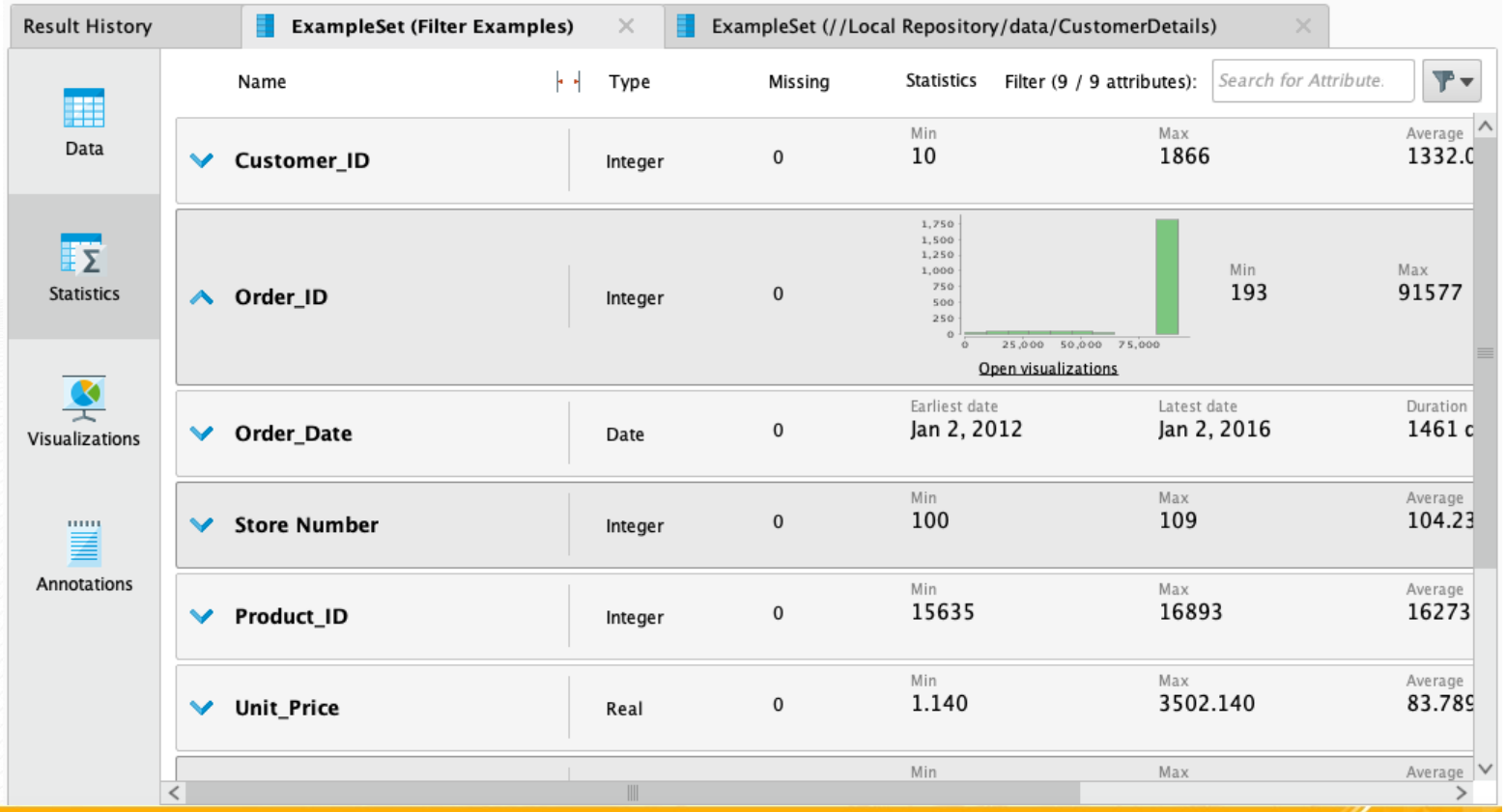
- To find the basic statistics of each attributes, click **Statistics**.



Row No.	Customer_I...	Order_ID	Order_Date	Store Num...	Product_ID
1	101	88209	Apr 26, 2014	100	16474
2	102	16710	Apr 26, 2014	103	16474
3	129	86698	Nov 30, 2015	101	16124
4	180	90784	Nov 4, 2015	100	16238
5	180	90785	Dec 14, 2015	100	16566
6	1025	89013	Nov 9, 2014	108	16182
7	1027	89016	Dec 29, 2015	103	15956
8	1030	89622	Apr 22, 2014	100	15653
9	1043	87851	Sep 8, 2015	100	16083

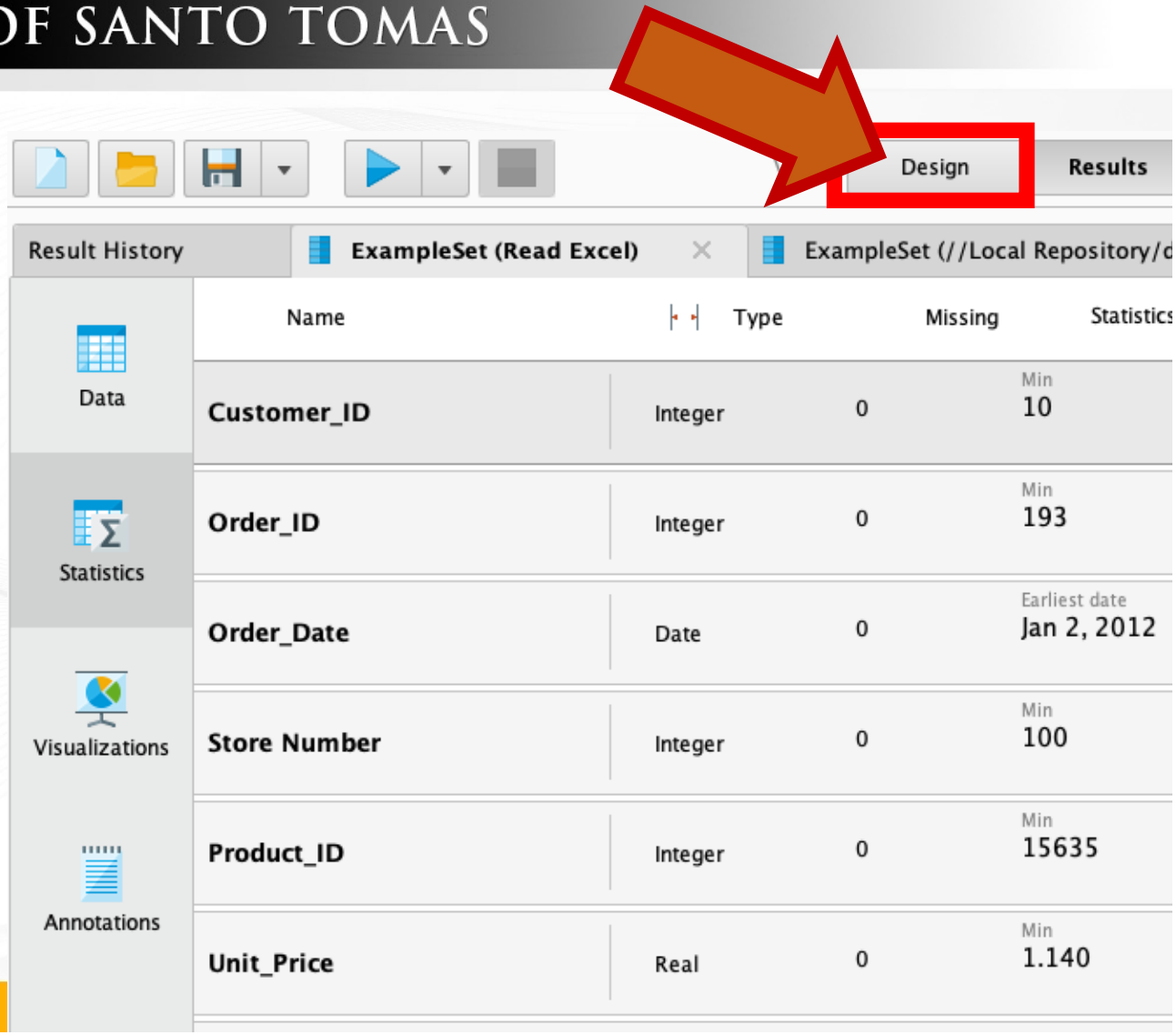
Exploratory Analysis

- To find the basic statistics of each attribute, click **Statistics**.



Data Preparation

Go back to **Design** view.



Data Filtering

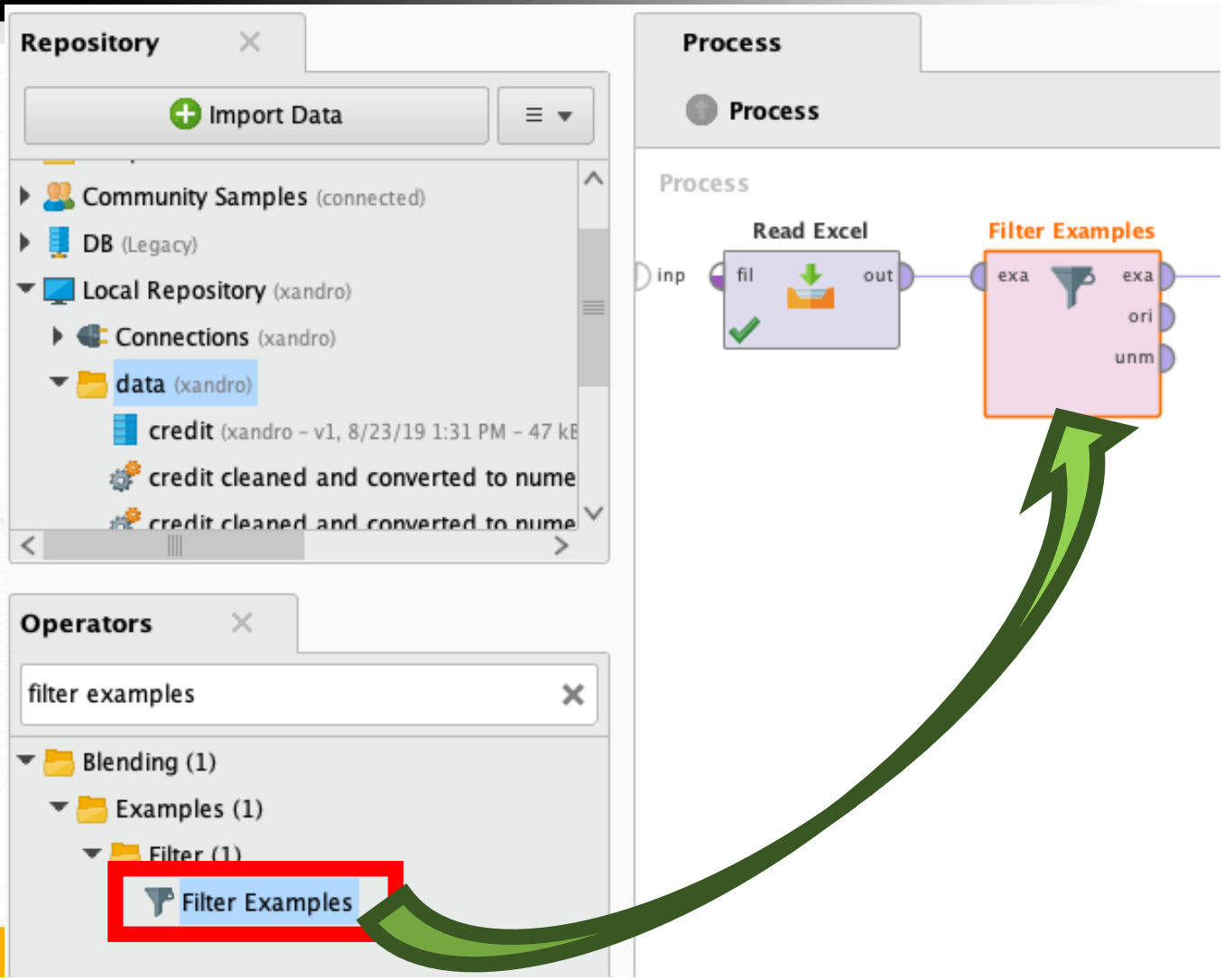
using

Altair® AI Studio

the process performed to decide which examples are **kept** and which are removed

Data Preparation

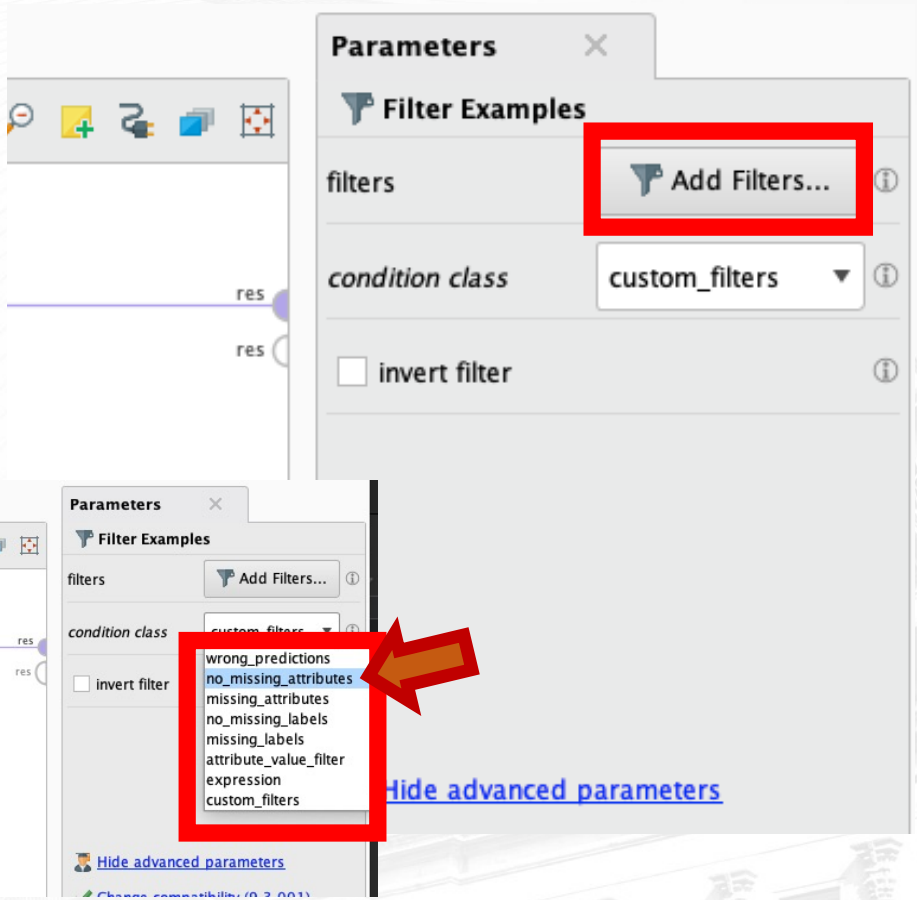
1. **Filtering cases**.
   - In the operator tab, search for **Filter Examples,** then drag and drop on the line connecting the *Read Excel* and the *res* knob.



Data Preparation

1. **Filtering cases**.
   - In the parameter tab, choose **Add Filter** in the condition class.

Note: Instead of filtering, you may **remove all cases with missing values** using the condition class, instead of Add Filters.

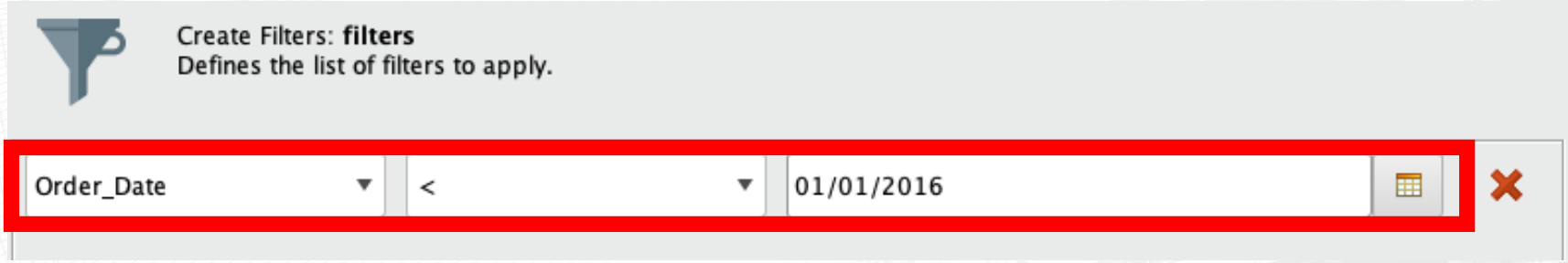


Data Preparation

1. **Filtering cases**.
   - Choose the attribute's filtering criteria.

Data Preparation

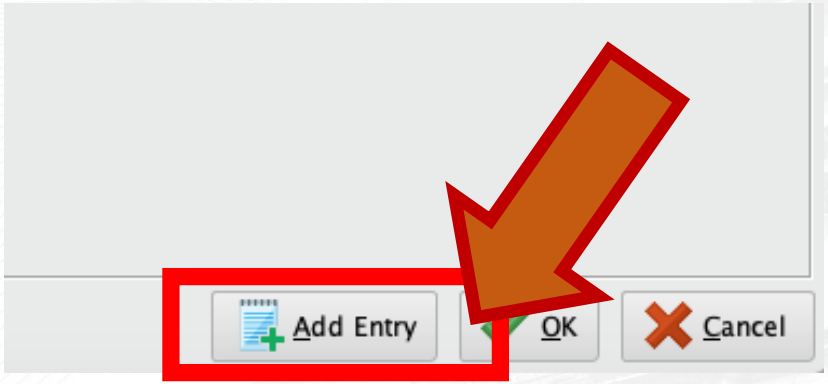
1. **Filtering cases**.
   - Choose the attribute's filtering criteria.
   - Example, retaining only the orders before 2016.



This will remove case(s) ordered from 2016 and beyond.

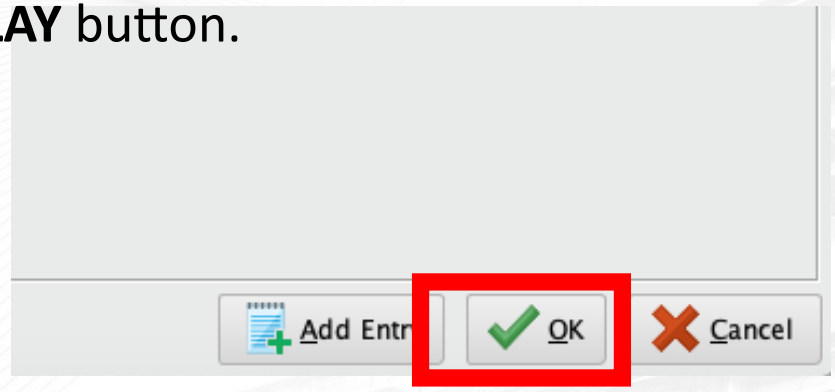
Data Preparation

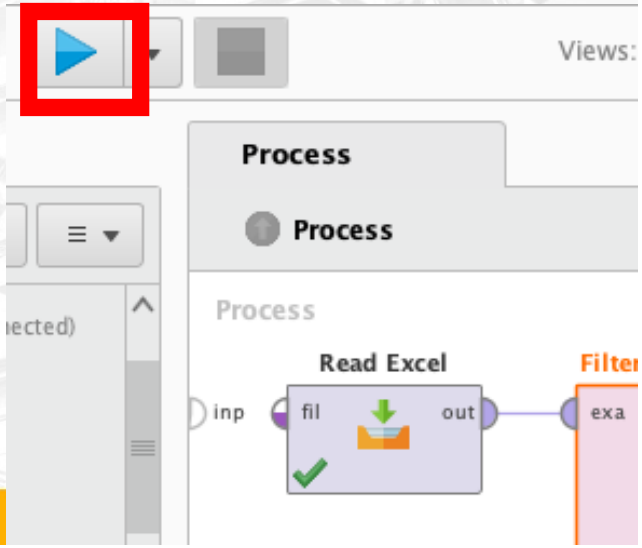
1. **Filtering cases**.
   - Choose the attribute's filtering criteria.
   - Example, retaining only the orders before 2016.
   - You may add more criteria by clicking **Add Entry**.



Data Preparation

1. **Filtering cases**.
   - Choose the attribute's filtering criteria.
   - Example, retaining only the orders before 2016.
   - You may add more criteria by clicking **Add Entry**.
   - Once all criteria have been set, click OK
     then click the **PLAY** button.





Data Preparation

RapidMiner removed 1 case, an order taken from 2016 onwards.

Design | Results | Turbo Prep | Auto Model

Set (//Local Repository/data/CustomerDetails)

Filter (2,268 / 2,268 examples):

Store Num...	Product_ID	Unit_Price	Discount	Quantity
100	16474	145.450	0.090	13
103	16474	145.450	0.090	53
101	16124	111.030	0.100	43
100	16238	291.730	0.090	21
100	16566	140.980	0.070	21

npleSet (//Local Repository/...)

Filter (2,267 / 2,267 examples):

Store Num...	Product_ID	Unit_Price	Discount	Quantity
100	16474	145.450	0.090	13
103	16474	145.450	0.090	53
101	16124	111.030	0.100	43
100	16238	291.730	0.090	21
100	16566	140.980	0.070	21
108	16182	284.980	0.080	20

Missing Value Imputation

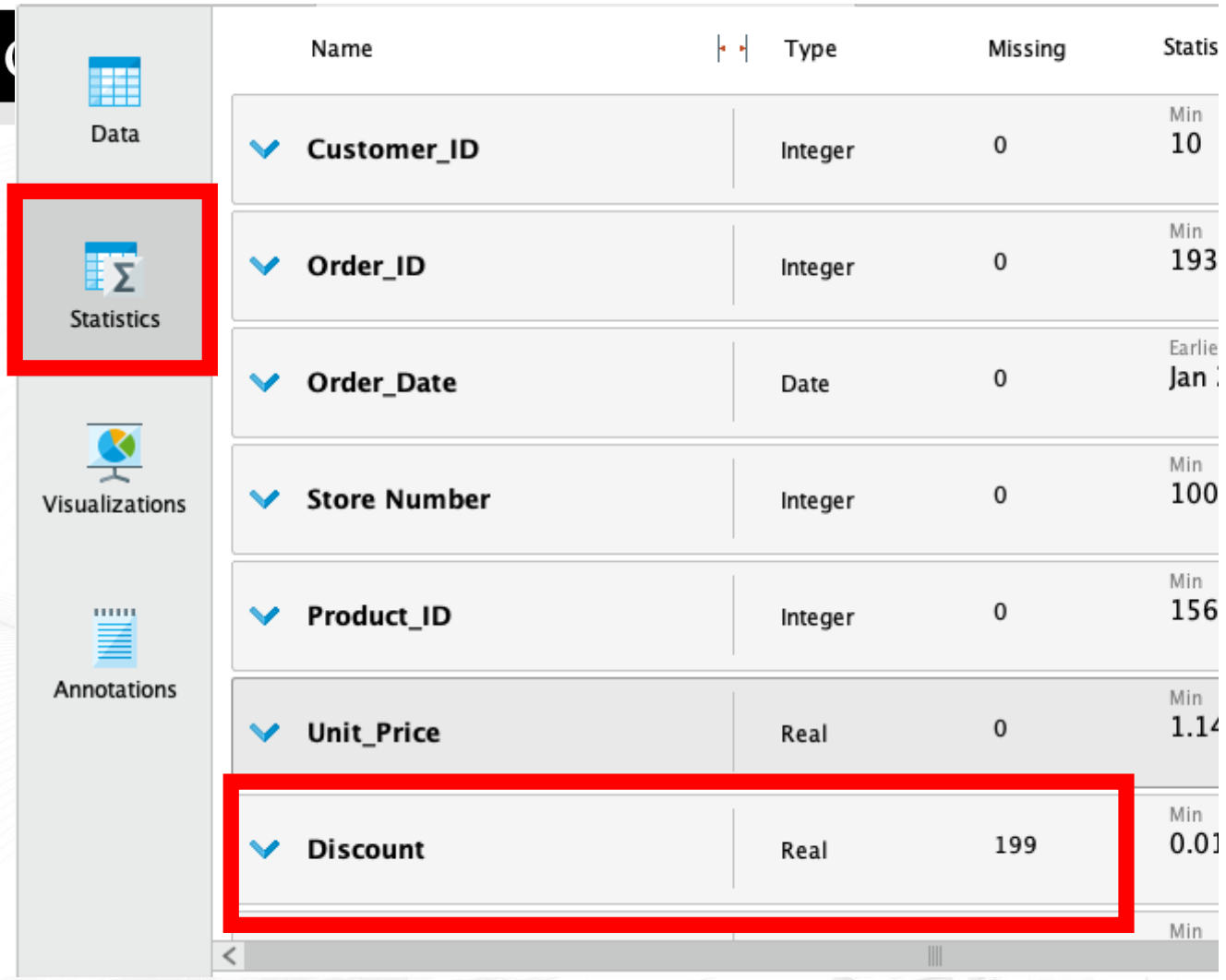
using

Altair® AI Studio

replaces missing values by the attribute's minimum, maximum or average value; zero may also be a replacement

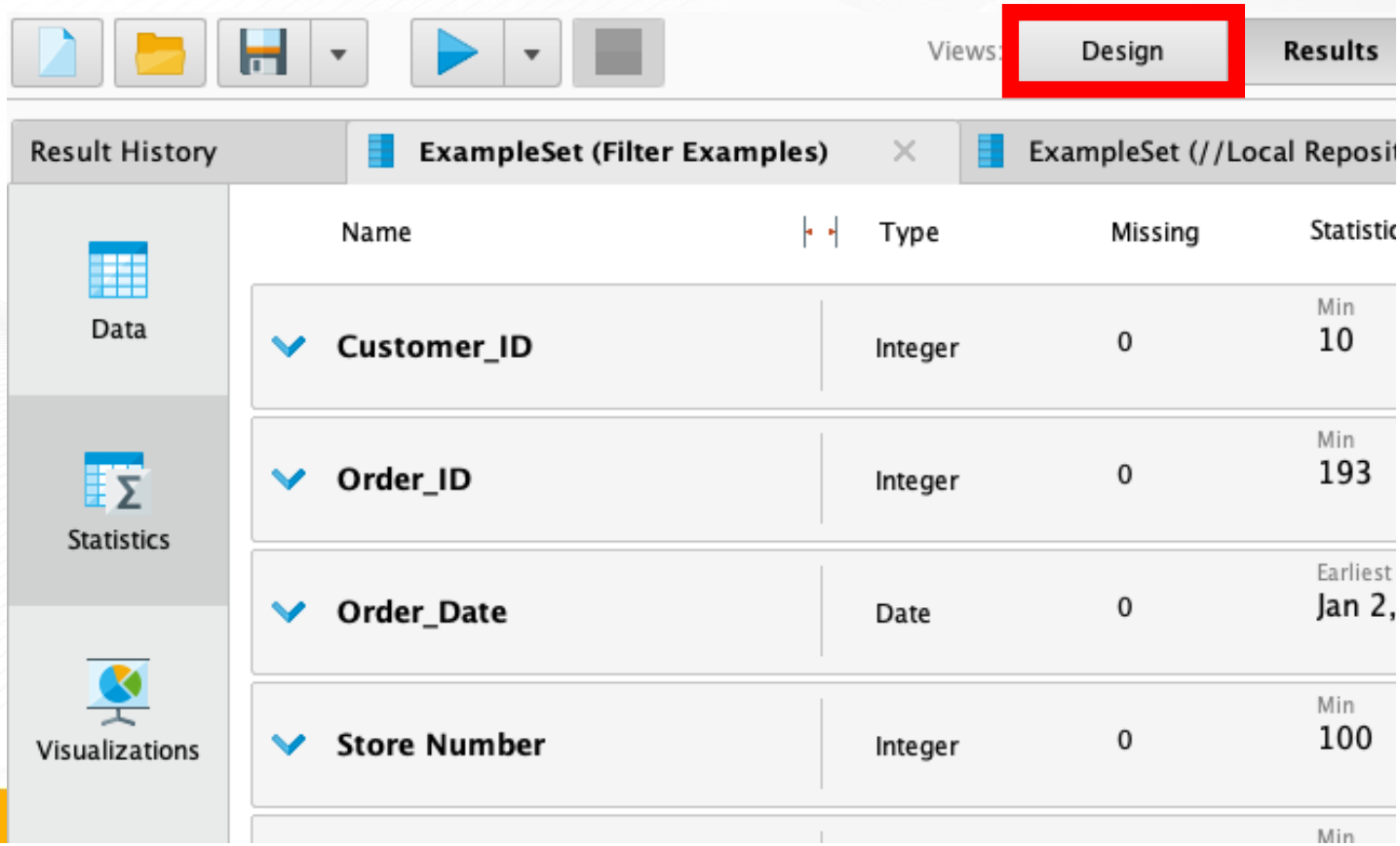
Data Preparation

As seen in the statistics of the data, ***199 examples*** *have missing values in the* ***Discount*** *attribute.*



Data Preparation

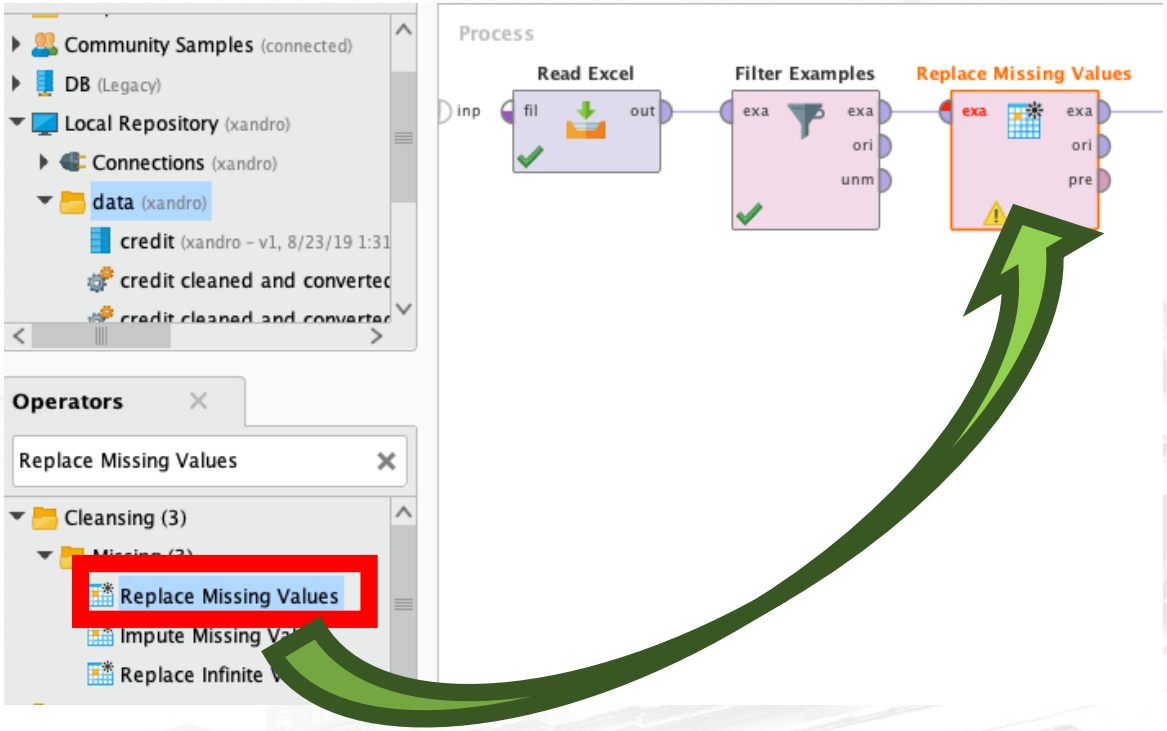
Go back to **Design** view.



Data Preparation

2. Imputing Missing Data

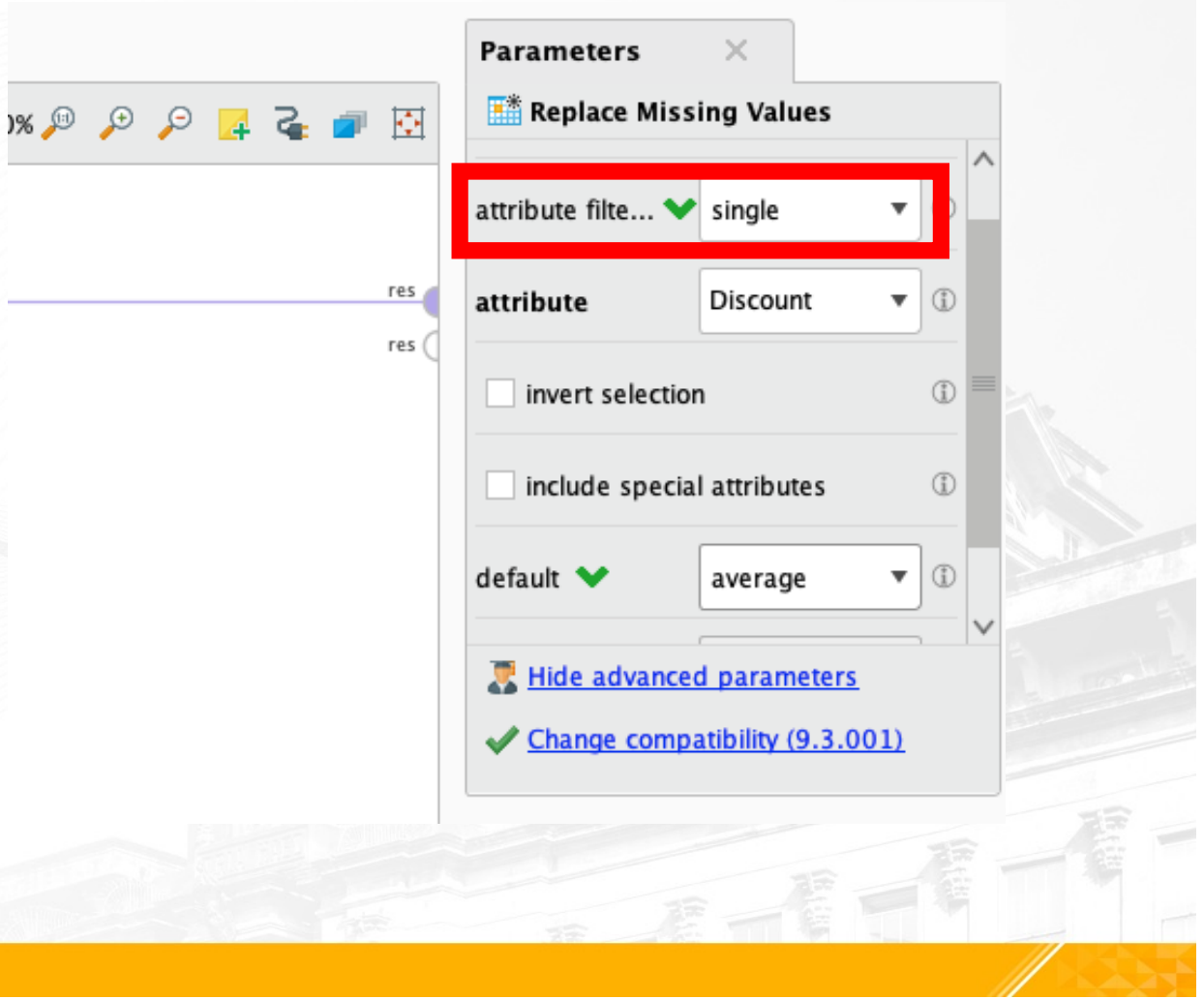
- In the operator tab, search for **Replace Missing Values,** then drag and drop on the line connecting the *Filtering Examples* and the *res* knob.



Data Preparation

2. Imputing Missing Data

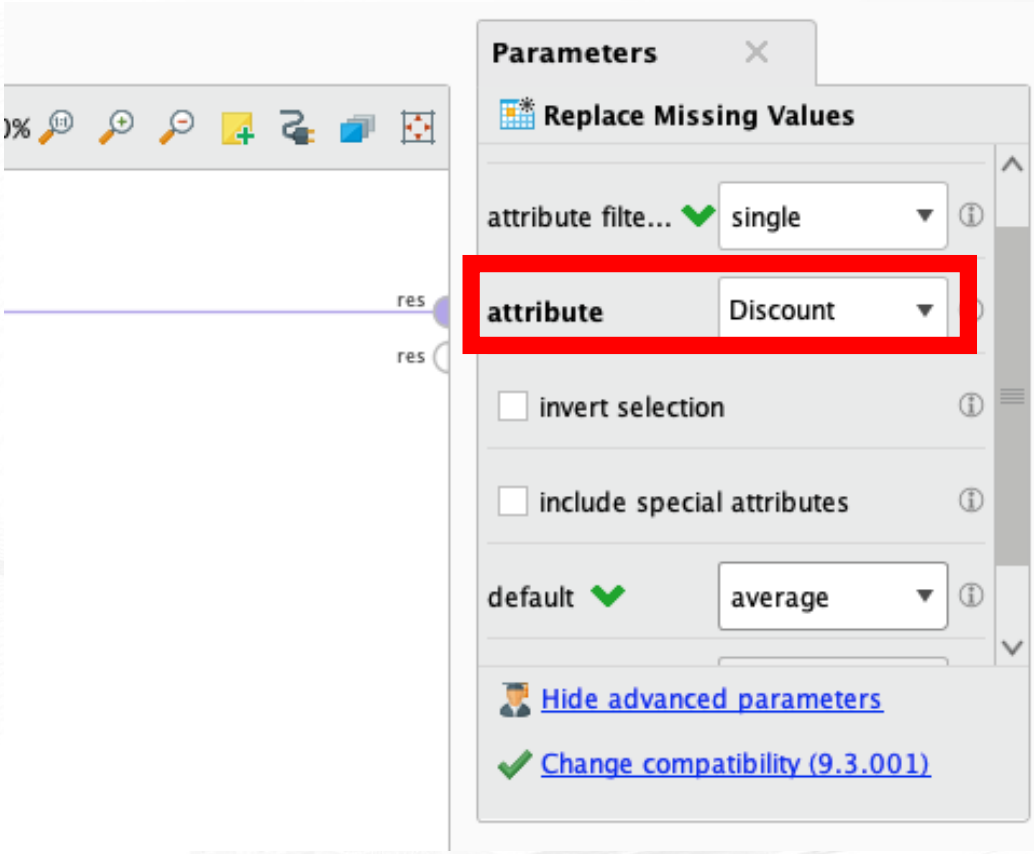
- In the parameter tab, select how many **attribute filter.** Choose **single** if the imputation will apply to a single attribute.



Data Preparation

2. Imputing Missing Data

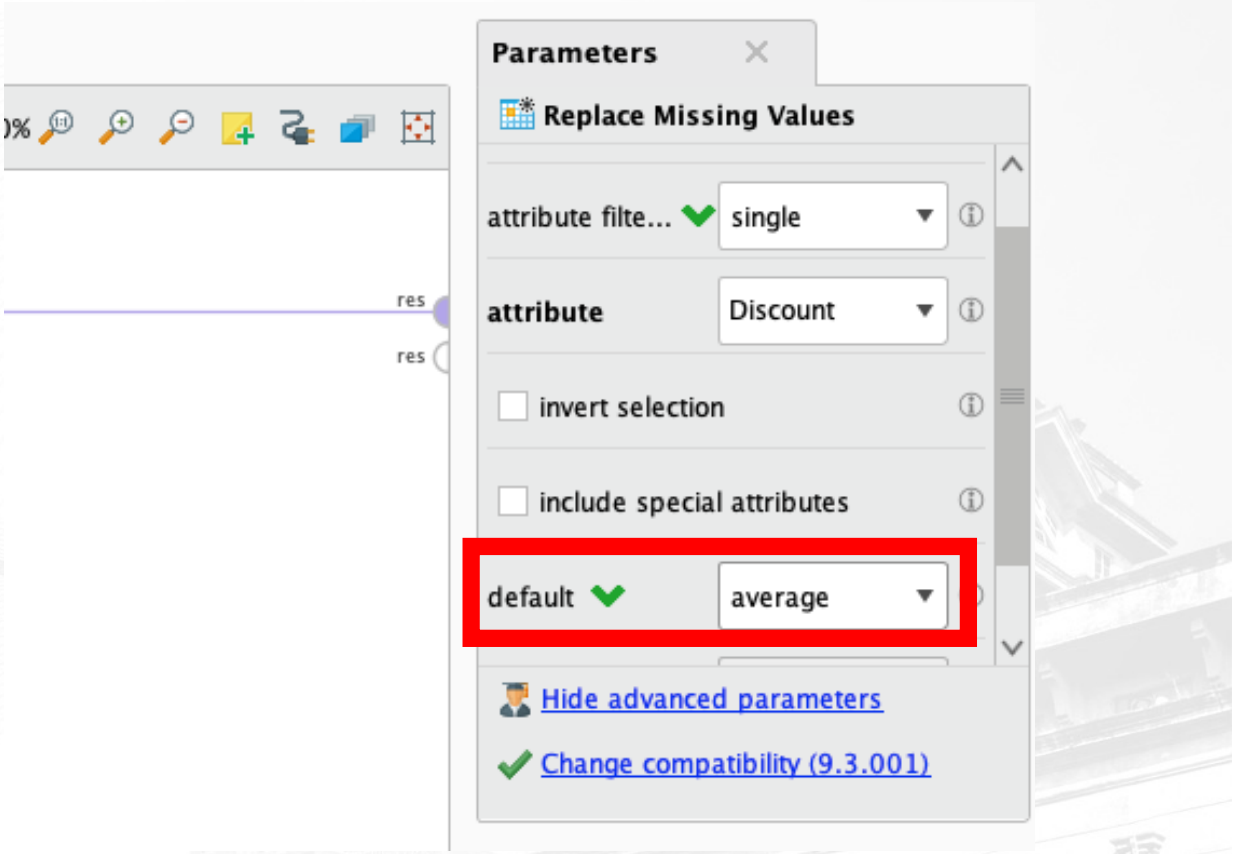
- In the parameter tab, select how many **attribute filter.** Choose **single** if the imputation will apply to a single attribute.
- Select the attribute where the imputation be applied.



Data Preparation

2. Imputing Missing Data

- In the parameter tab, select how many **attribute filter.** Choose **single** if the imputation will apply to a single attribute.
- Select the attribute where the imputation be applied.
- Select the imputation method in the **Default.**



Data Preparation

2. Imputing Missing Data

- In the parameter tab, select how many **attribute filter.** Choose **single** if the imputation will apply to a single attribute.
- Select the attribute where the imputation be applied.
- Select the imputation method in the **Default.**
- Click the **Play** button to see result.