Babu Banarasi Das University

Case Study: Cleanup and Analysis of Health Insurance Company Data

SUBMITTED TO:

Mr. Robin Tyagi Sir

SUBMITTED BY:

Name: Devansh Soni

Roll No: 1230258161

Class: BCADS 33

Definition:

In this practical, the goal is to clean, reclassify, and correct inconsistencies within an insurance dataset to prepare it for modelling. The process involves redefining categorical variables, standardizing field values, and correcting anomalies in numeric data fields such as *age* and *children*. This ensures the dataset is consistent, accurate, and ready for predictive analysis.

Outcomes/Learning:

- Learned how to use the **Reclassify Node** to modify and standardize categorical data values.
- Understood the application of the **Type Node** for assigning appropriate field roles and measurement levels.
- Gained practical experience in correcting and validating numeric fields using the **Derive Node**.
- Learned to apply data cleansing techniques for improving data quality and accuracy.
- Enhanced understanding of data preparation as a prerequisite for effective data modelling.

Required Tool:

IBM SPSS Modeler

Working:

SPSS Modeler follows the **CRISP–DM (Cross-Industry Standard Process for Data Mining)** framework with six stages — *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation,* and *Deployment*.
In this practical, the focus is on the **Data Preparation** phase, where raw data is cleansed and standardized for analytical use.

The dataset *(insurance.csv)* is first imported using the **Var. File Node**, and field roles are defined using the **Type Node**. Next, the **Reclassify Nodes** are applied to standardize categorical variables — *smoker* and *region* — ensuring consistent value representation.

Subsequently, **Filler Nodes** are used to correct missing values in *age* and *children* fields by applying conditional expressions and logical transformations. Each transformation step is validated with a **Table Node** to confirm the accuracy of applied corrections.

Finally, the prepared dataset is ready for further analysis or predictive modelling.

Steps:

Step 1: Importing the Dataset and Viewing Initial Records

- Open IBM SPSS Modeler and create a New Stream.
- From the Sources tab, double-click **Var. File** — a Var. File Source Node will appear.
- Browse and select the dataset *(insurance.csv)* using the browse button.
- Click **Apply**, then **OK** to load the dataset.
- Attach a **Table Node** to display and review the initial data.

Step 2: Defining Field Roles Using the Type Node

- Select the Var. File Node, go to the **Field Ops** tab, and double-click **Type**.
- A Type Node will be added to the stream.
- Open the Type Node and click **Read Values** to automatically detect data properties.
- Adjust measurement levels and field roles (e.g., Nominal, Continuous) as needed.
- Click **Apply**, then **OK** to confirm.
- Use a **Table Node** to verify updated field settings.

Step 3: Reclassifying the "Smoker" Field

- Select the Type Node, go to the **Field Ops** tab, and double-click **Reclassify**.
- A Reclassify Node will appear — rename it **smoker reclassified**.
- Open the Node and choose the **smoker** field for reclassification.
- Modify inconsistent values (e.g., "yes", "Yes", "Y") into a single category ("yes").
- Click **Apply**, then **OK**.
- Connect a **Table Node** to verify the transformation.

Step 4: Reclassifying the "Region" Field

- Select the **smoker reclassified** Node, go to the **Field Ops** tab, and double-click **Reclassify**.
- Rename the Node **region reclassified**.
- Select the **region** field for reclassification.
- Standardize inconsistent entries (e.g., "Northwest", "NW" → "northwest").
- Click **Apply**, then **OK**.
- Attach a **Table Node** and run it to view the results.

Step 5: Correcting Age Field Using Filler Node

- Select the **region reclassified** Node, go to the **Field Ops** tab, and double-click **Filler**.
- Rename the Node **age correction**.
- In the formula box, enter the filler logic/value (e.g., Replace Black and Null Values -> 39[median]).
- Click **Apply**, then **OK**.
- Use a **Table Node** to confirm corrected age values.

Step 6: Correcting Children Field Using Filler Node

- Select the **age correction** Node, go to the **Field Ops** tab, and double-click **Filler**.
- Rename the Node **children correction**.
- In the formula box, add the correction logic (e.g., Replace Black and Null Values -> 1[median]).
- Click **Apply**, then **OK**.
- Connect a **Table Node** to review the final output.

✅ Final Output:

A cleansed and standardized insurance dataset, free from inconsistent categorical entries and invalid numeric values, ready for analysis or modelling.

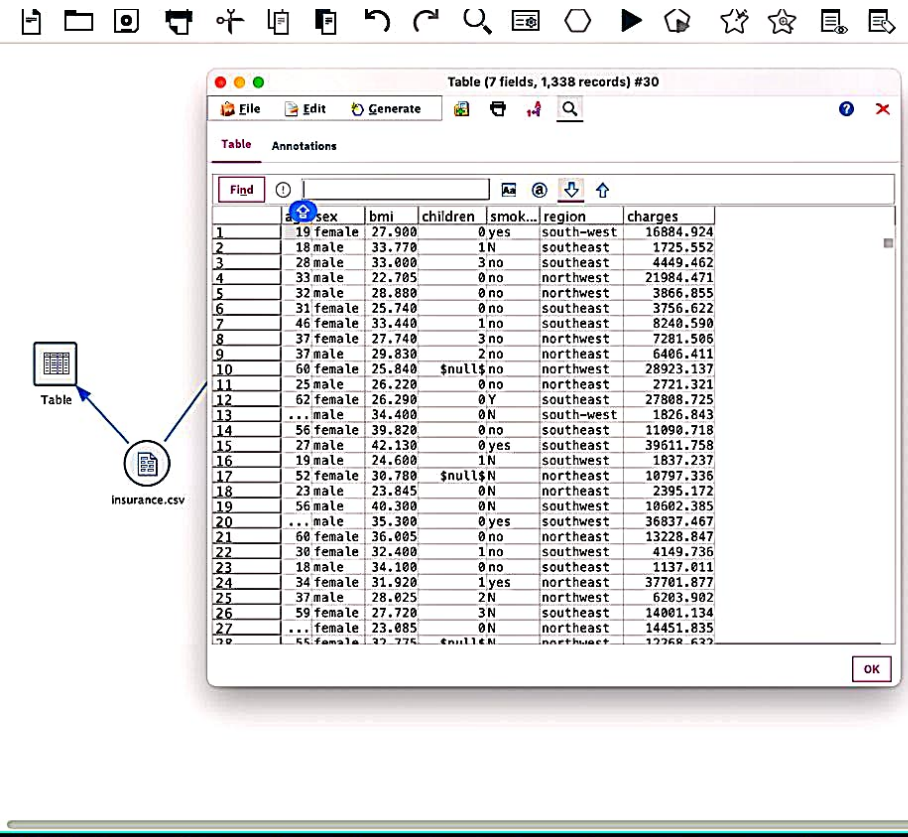
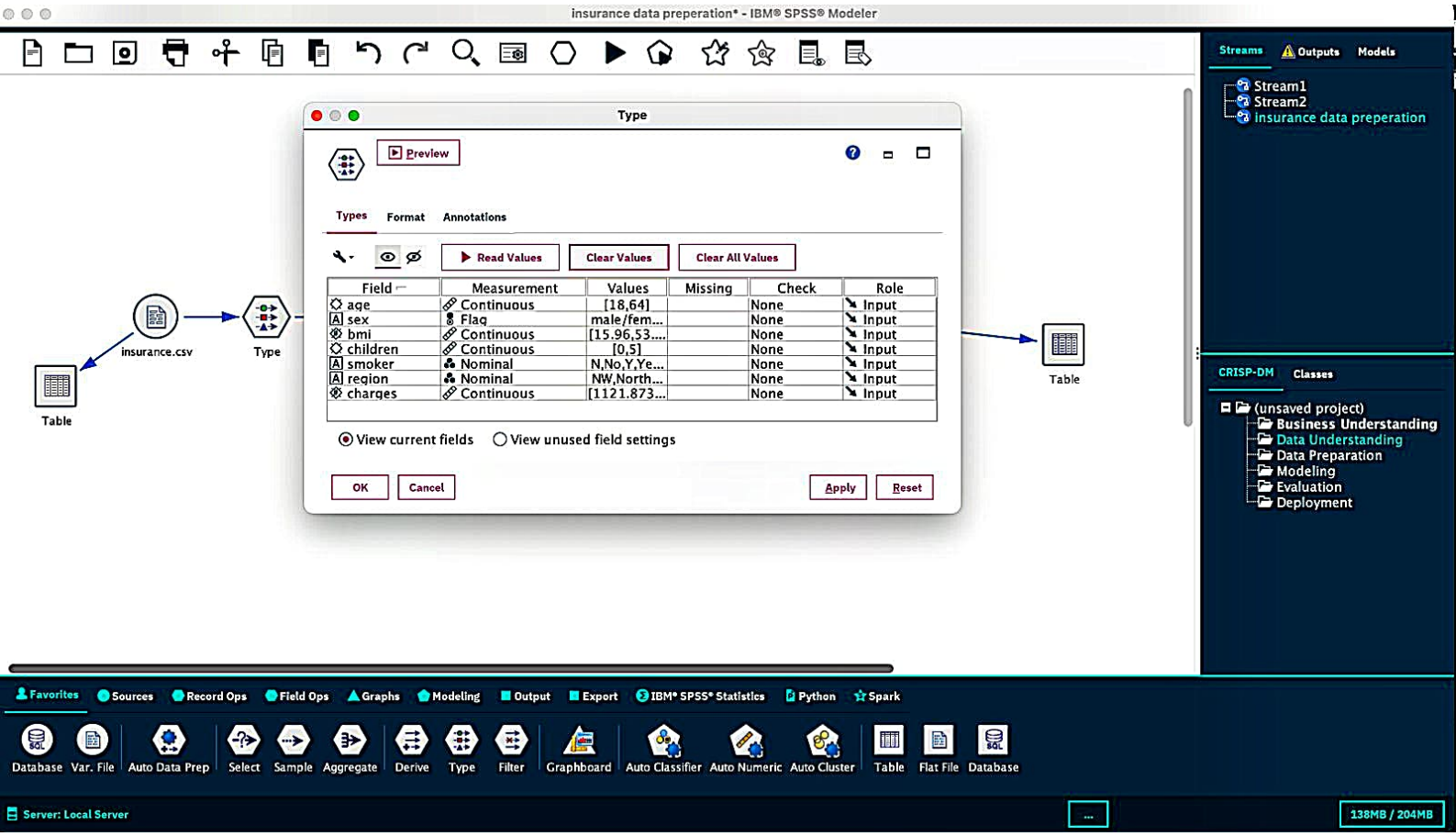
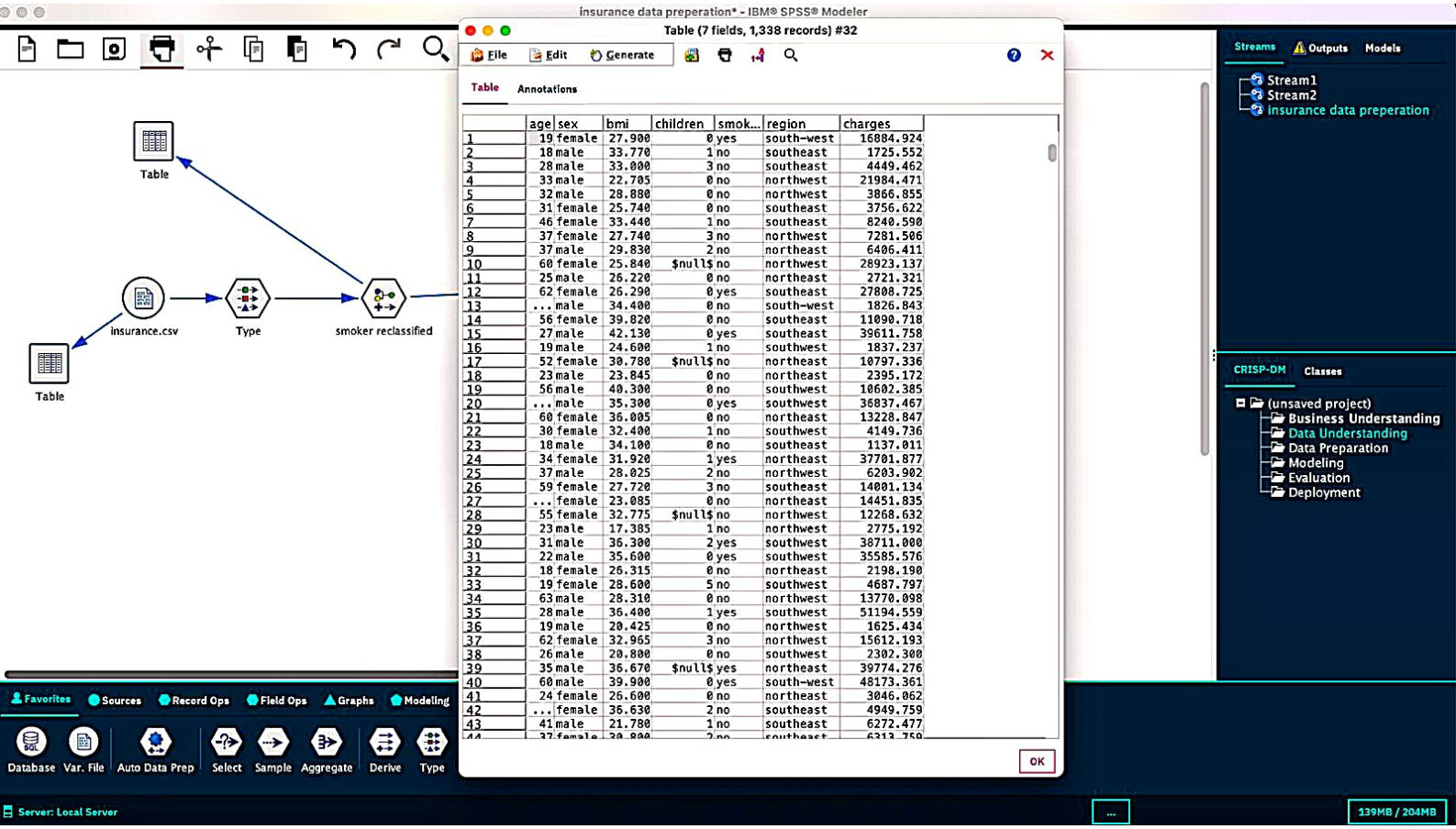
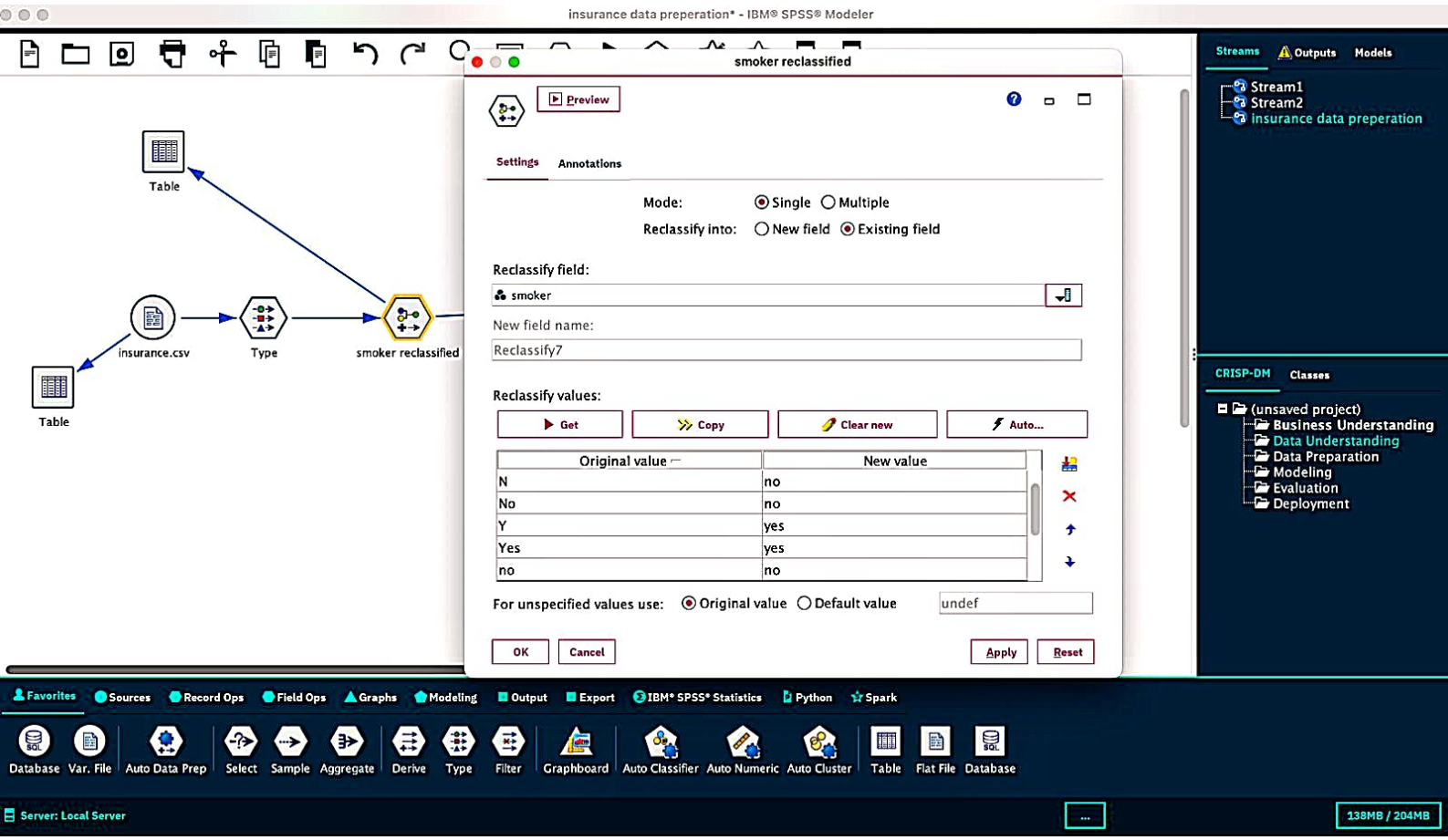


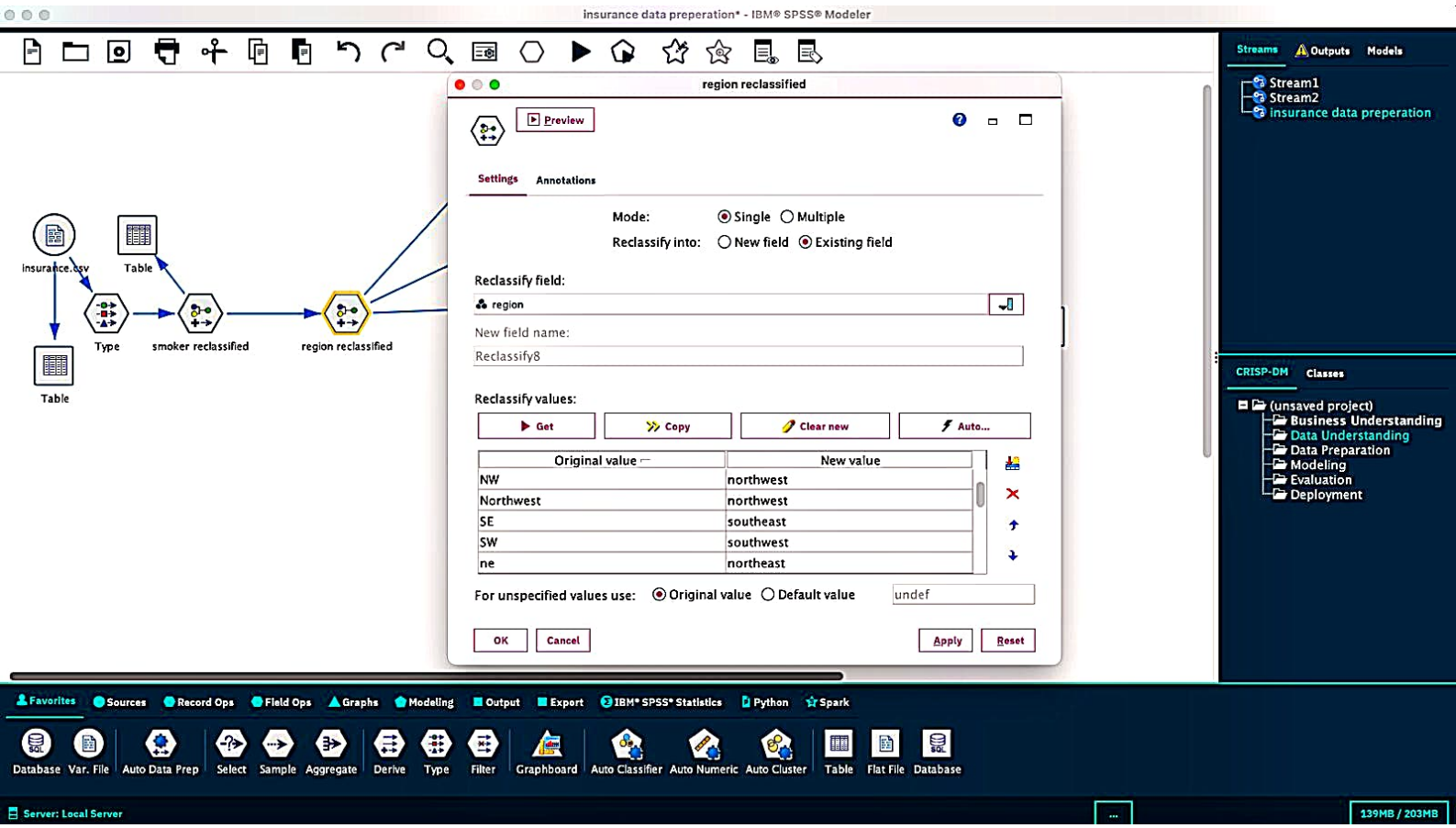
Table (7 fields, 1,338 records) #30
File
Edit
Generate
Table
Annotations
Find
age sex bmi children smok... region charges
1 19 female 27.900 0 yes south-west 16884.924
2 18 male 33.770 1 N southeast 1725.552
3 28 male 33.000 3 no southeast 4449.462
4 33 male 22.705 0 no northwest 21984.471
5 32 male 28.880 0 no northwest 3866.855
6 31 female 25.740 0 no southeast 3756.622
7 46 female 33.440 1 no southeast 8240.590
8 37 female 27.740 3 no northwest 7281.506
9 37 male 29.830 2 no northeast 6406.411
10 60 female 25.840 $null$ no northwest 28923.137
11 25 male 26.220 0 no northeast 2721.321
12 62 female 26.290 0 Y southeast 27808.725
13 ... male 34.400 0 N south-west 1826.843
14 56 female 39.820 0 no southeast 11090.718
15 27 male 42.130 0 yes southeast 39611.758
16 19 male 24.600 1 N southwest 1837.237
17 52 female 30.780 $null$ N northeast 10797.336
18 23 male 23.845 0 N northeast 2395.172
19 56 male 40.300 0 N southwest 10602.385
20 ... male 35.300 0 yes southwest 36837.467
21 60 female 36.005 0 no northeast 13228.847
22 30 female 32.400 1 no southwest 4149.736
23 18 male 34.100 0 no southeast 1137.011
24 34 female 31.920 1 yes northeast 37701.877
25 37 male 28.025 2 N northwest 6203.902
26 59 female 27.720 3 N southeast 14001.134
27 ... female 23.085 0 N northeast 14451.835
OK
Table
insurance.csv
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
CRISP-DM
Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
...
138MB / 204MB

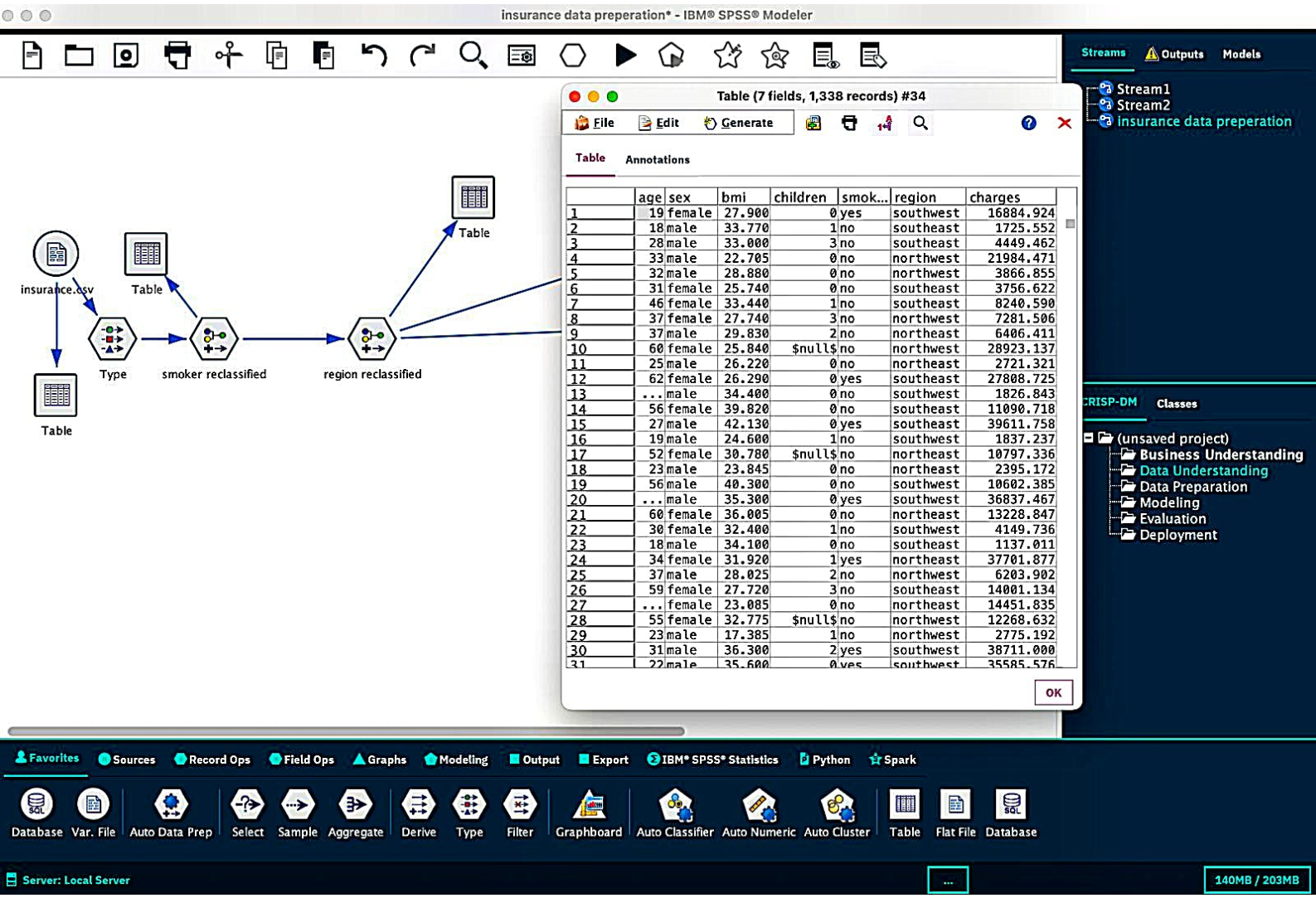

insurance data preperation* - IBM® SPSS® Modeler
Type
Preview
Types
Format
Annotations
Read Values
Clear Values
Clear All Values
Field
Measurement
Values
Missing
Check
Role
age Continuous [18,64] None Input
sex Flag male/fem... None Input
bmi Continuous [15.96,53.... None Input
children Continuous [0,5] None Input
smoker Nominal N,No,Y,Ye... None Input
region Nominal NW,North... None Input
charges Continuous [1121.873... None Input
View current fields
View unused field settings
OK
Cancel
Apply
Reset
insurance.csv
Type
Table
Table
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
CRISP-DM
Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
138MB / 204MB

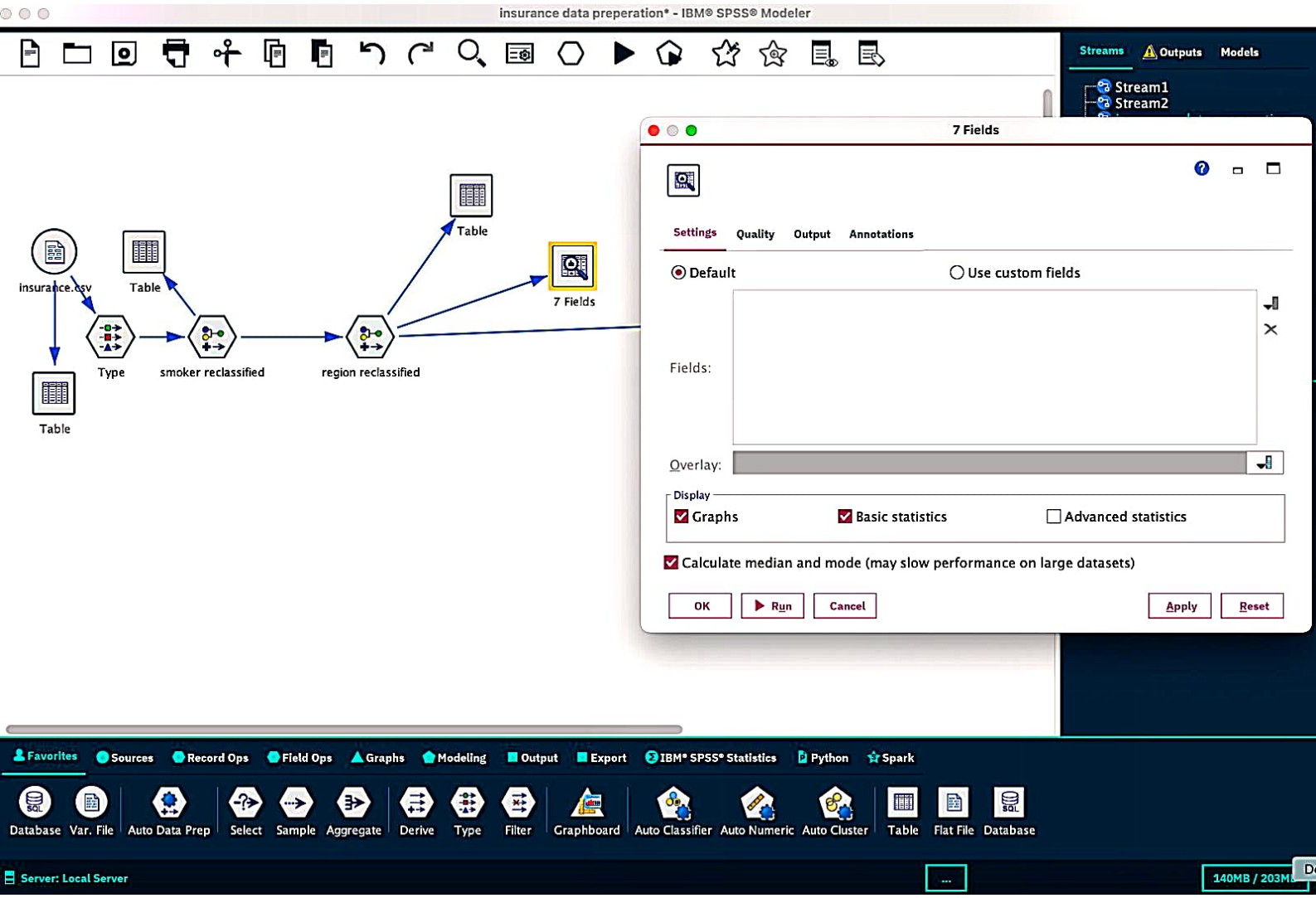


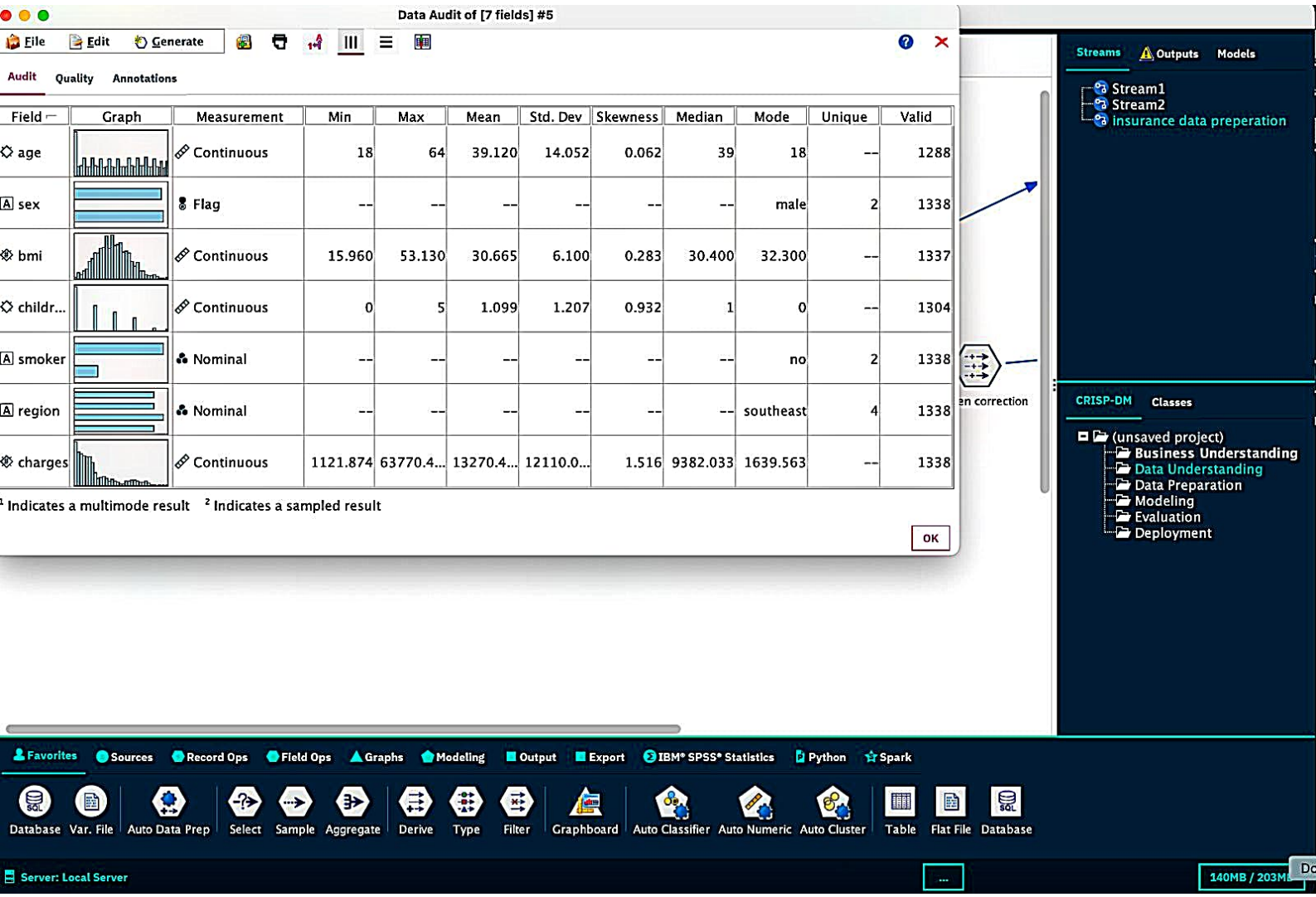
	age	sex	bmi	children	smok...	region	charges
1	19	female	27.900	0	yes	south-west	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622
7	46	female	33.440	1	no	southeast	8240.590
8	37	female	27.740	3	no	northwest	7281.506
9	37	male	29.830	2	no	northeast	6406.411
10	60	female	25.840	$null$	no	northwest	28923.137
11	25	male	26.220	0	no	northeast	2721.321
12	62	female	26.290	0	yes	southeast	27808.725
13	...	male	34.400	0	no	south-west	1826.843
14	56	female	39.820	0	no	southeast	11090.718
15	27	male	42.130	0	yes	southeast	39611.758
16	19	male	24.600	1	no	southwest	1837.237
17	52	female	30.780	$null$	no	northeast	10797.336
18	23	male	23.845	0	no	northeast	2395.172
19	56	male	40.300	0	no	southwest	10602.385
20	...	male	35.300	0	yes	southwest	36837.467
21	60	female	36.005	0	no	northeast	13228.847
22	30	female	32.400	1	no	southwest	4149.736
23	18	male	34.100	0	no	southeast	1137.011
24	34	female	31.920	1	yes	northeast	37701.877
25	37	male	28.025	2	no	northwest	6203.902
26	59	female	27.720	3	no	southeast	14001.134
27	...	female	23.085	0	no	northeast	14451.835
28	55	female	32.775	$null$	no	northwest	12268.632
29	23	male	17.385	1	no	northwest	2775.192
30	31	male	36.300	2	yes	southwest	38711.000
31	22	male	35.600	0	yes	southwest	35585.576
32	18	female	26.315	0	no	northeast	2198.190
33	19	female	28.600	5	no	southwest	4687.797
34	63	male	28.310	0	no	northwest	13770.098
35	28	male	36.400	1	yes	southwest	51194.559
36	19	male	20.425	0	no	northwest	1625.434
37	62	female	32.965	3	no	northwest	15612.193
38	26	male	20.800	0	no	southwest	2302.300
39	35	male	36.670	$null$	yes	northeast	39774.276
40	60	male	39.900	0	yes	south-west	48173.361
41	24	female	26.600	0	no	northeast	3046.062
42	...	female	36.630	2	no	southeast	4949.759
43	41	male	21.780	1	no	southeast	6272.477
44	37	female	30.800	2	no	southeast	6313.759

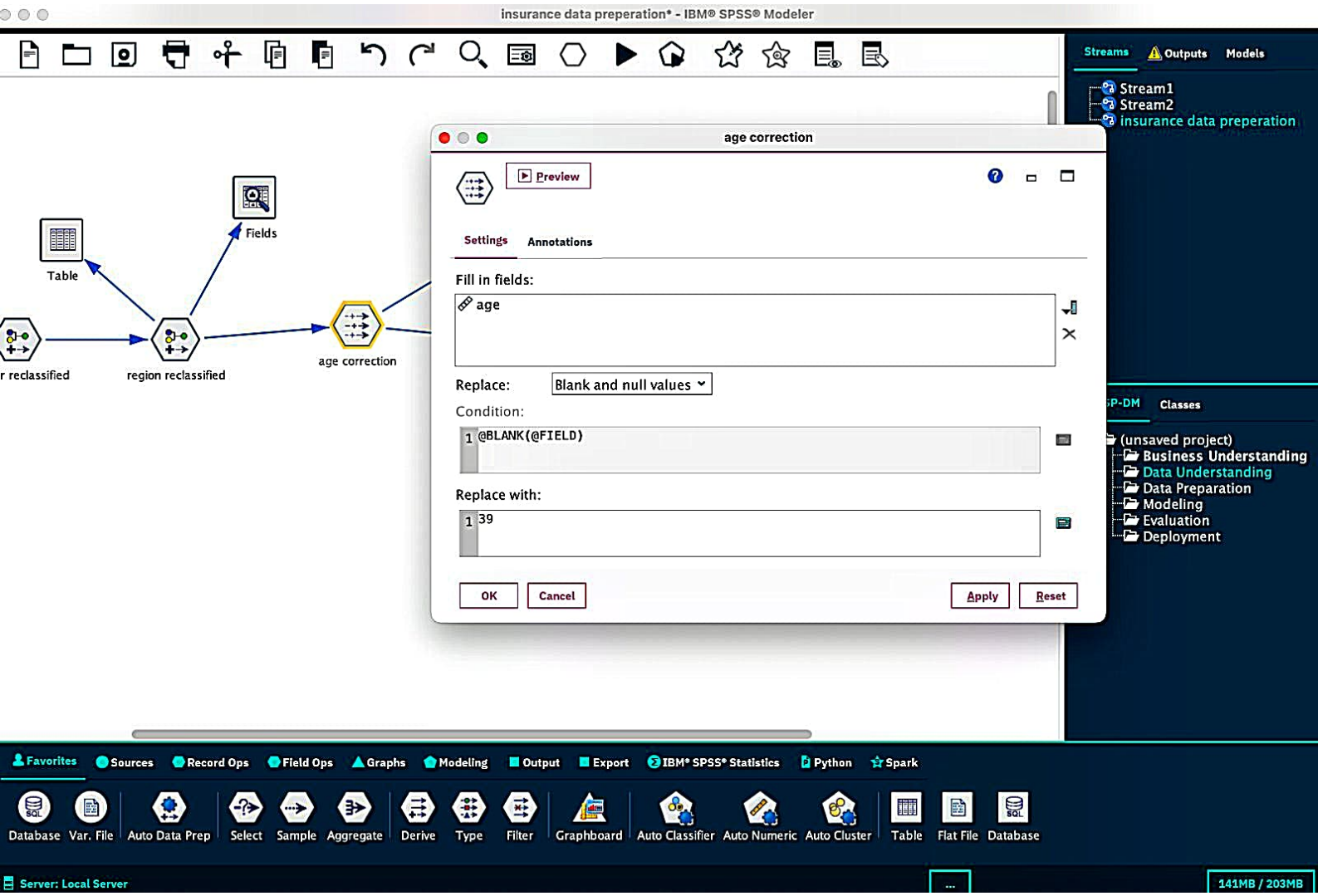

insurance data preperation* - IBM® SPSS® Modeler
smoker reclassified
Preview
Settings
Annotations
Mode: Single Multiple
Reclassify into: New field Existing field
Reclassify field:
smoker
New field name:
Reclassify7
Reclassify values:
Get
Copy
Clear new
Auto...
Original value
New value
N
no
No
no
Y
yes
Yes
yes
no
no
For unspecified values use: Original value Default value
undef
OK
Cancel
Apply
Reset
Table
insurance.csv
Type
smoker reclassified
Table
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
CRISP-DM
Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
138MB / 204MB

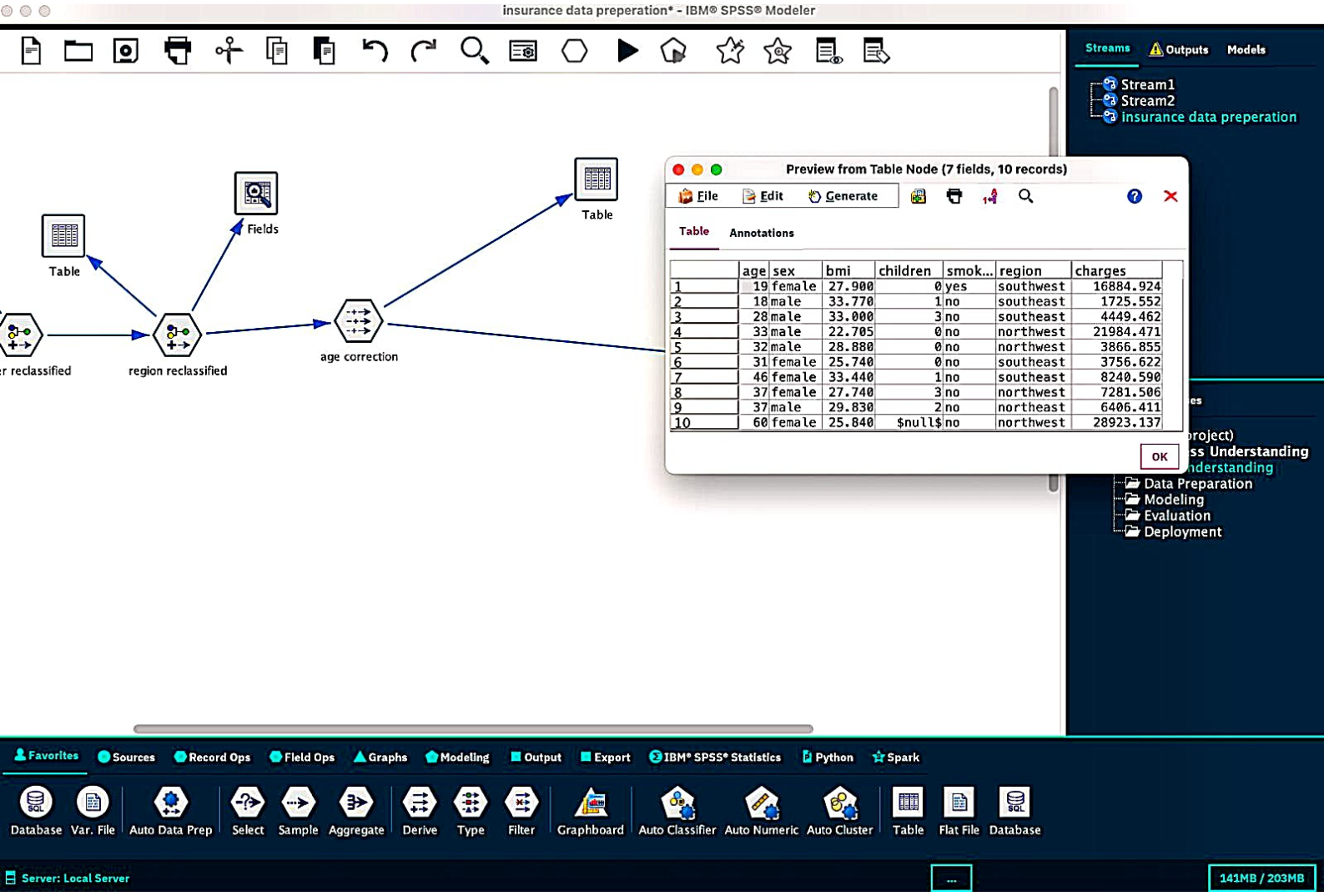

insurance data preperation* - IBM® SPSS® Modeler
region reclassified
Preview
Settings
Annotations
Mode: Single Multiple
Reclassify into: New field Existing field
Reclassify field:
region
New field name:
Reclassify8
Reclassify values:
Get
Copy
Clear new
Auto...
Original value
New value
NW
northwest
Northwest
northwest
SE
southeast
SW
southwest
ne
northeast
For unspecified values use: Original value Default value
undef
OK
Cancel
Apply
Reset
insurance.csv
Table
Type
smoker reclassified
region reclassified
Table
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
CRISP-DM
Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
139MB / 203MB

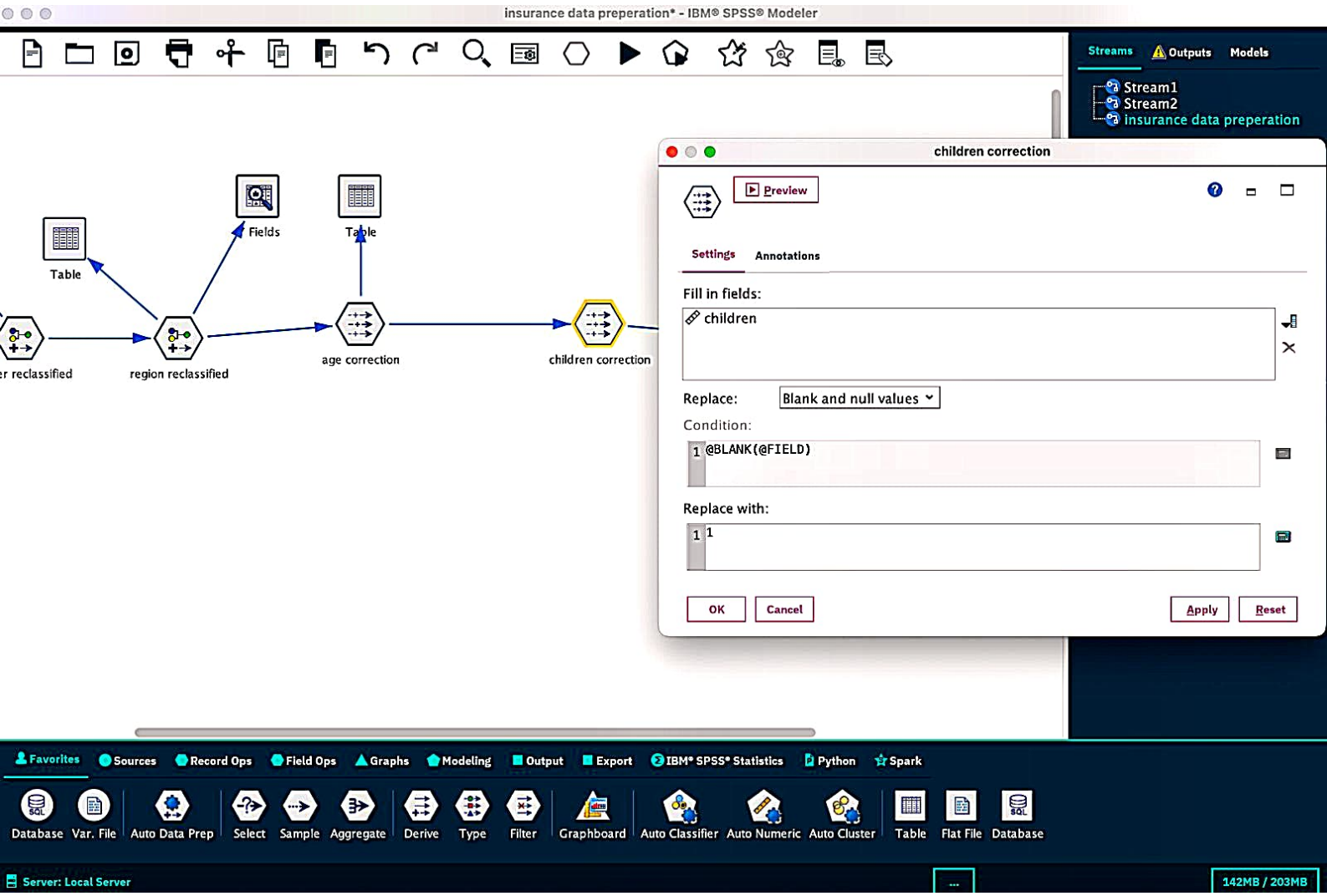

insurance data preperation* - IBM® SPSS® Modeler
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
Table (7 fields, 1,338 records) #34
File
Edit
Generate
Table
Annotations
age sex bmi children smok... region charges
1 19 female 27.900 0 yes southwest 16884.924
2 18 male 33.770 1 no southeast 1725.552
3 28 male 33.000 3 no southeast 4449.462
4 33 male 22.705 0 no northwest 21984.471
5 32 male 28.880 0 no northwest 3866.855
6 31 female 25.740 0 no southeast 3756.622
7 46 female 33.440 1 no southeast 8240.590
8 37 female 27.740 3 no northwest 7281.506
9 37 male 29.830 2 no northeast 6406.411
10 60 female 25.840 $null$ no northwest 28923.137
11 25 male 26.220 0 no northeast 2721.321
12 62 female 26.290 0 yes southeast 27808.725
13 ... male 34.400 0 no southwest 1826.843
14 56 female 39.820 0 no southeast 11090.718
15 27 male 42.130 0 yes southeast 39611.758
16 19 male 24.600 1 no southwest 1837.237
17 52 female 30.780 $null$ no northeast 10797.336
18 23 male 23.845 0 no northeast 2395.172
19 56 male 40.300 0 no southwest 10602.385
20 ... male 35.300 0 yes southwest 36837.467
21 60 female 36.005 0 no northeast 13228.847
22 30 female 32.400 1 no southwest 4149.736
23 18 male 34.100 0 no southeast 1137.011
24 34 female 31.920 1 yes northeast 37701.877
25 37 male 28.025 2 no northwest 6203.902
26 59 female 27.720 3 no southeast 14001.134
27 ... female 23.085 0 no northeast 14451.835
28 55 female 32.775 $null$ no northwest 12268.632
29 23 male 17.385 1 no northwest 2775.192
30 31 male 36.300 2 yes southwest 38711.000
31 22 male 35.600 0 yes southwest 35585.576
OK
Table
insurance.csv
Table
Type
smoker reclassified
region reclassified
Table
CRISP-DM
Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
140MB / 203MB


insurance data preperation* - IBM® SPSS® Modeler
Streams
Outputs
Models
Stream1
Stream2
7 Fields
Settings
Quality
Output
Annotations
Default
Use custom fields
Fields:
Overlay:
Display
Graphs
Basic statistics
Advanced statistics
Calculate median and mode (may slow performance on large datasets)
OK
Run
Cancel
Apply
Reset
insurance.csv
Table
Type
smoker reclassified
region reclassified
Table
7 Fields
Table
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
140MB / 203MB


Data Audit of [7 fields] #5
File
Edit
Generate
Audit
Quality
Annotations
Field
Graph
Measurement
Min
Max
Mean
Std. Dev
Skewness
Median
Mode
Unique
Valid
age
Continuous
18
64
39.120
14.052
0.062
39
18
--
1288
sex
Flag
--
--
--
--
--
--
male
2
1338
bmi
Continuous
15.960
53.130
30.665
6.100
0.283
30.400
32.300
--
1337
childr...
Continuous
0
5
1.099
1.207
0.932
1
0
--
1304
smoker
Nominal
--
--
--
--
--
--
no
2
1338
region
Nominal
--
--
--
--
--
--
southeast
4
1338
charges
Continuous
1121.874
63770.4...
13270.4...
12110.0...
1.516
9382.033
1639.563
--
1338
[1] Indicates a multimode result
[2] Indicates a sampled result
OK
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
CRISP-DM
Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
140MB / 203MB


insurance data preperation* - IBM® SPSS® Modeler
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
Table
Fields
r reclassified
region reclassified
age correction
age correction
Preview
Settings
Annotations
Fill in fields:
age
Replace:
Blank and null values
Condition:
1 @BLANK(@FIELD)
Replace with:
1 39
OK
Cancel
Apply
Reset
Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
141MB / 203MB


insurance data preperation* - IBM® SPSS® Modeler
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
Table
Fields
Table
region reclassified
age correction
Preview from Table Node (7 fields, 10 records)
File
Edit
Generate
Table
Annotations
age sex bmi children smok... region charges
1 19 female 27.900 0 yes southwest 16884.924
2 18 male 33.770 1 no southeast 1725.552
3 28 male 33.000 3 no southeast 4449.462
4 33 male 22.705 0 no northwest 21984.471
5 32 male 28.880 0 no northwest 3866.855
6 31 female 25.740 0 no southeast 3756.622
7 46 female 33.440 1 no southeast 8240.590
8 37 female 27.740 3 no northwest 7281.506
9 37 male 29.830 2 no northeast 6406.411
10 60 female 25.840 $null$ no northwest 28923.137
OK
Data Preparation
Modeling
Evaluation
Deployment
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
141MB / 203MB


insurance data preperation* - IBM® SPSS® Modeler
Streams
Outputs
Models
Stream1
Stream2
insurance data preperation
Table
Fields
Table
er reclassified
region reclassified
age correction
children correction
children correction
Preview
Settings
Annotations
Fill in fields:
children
Replace:
Blank and null values
Condition:
1 @BLANK(@FIELD)
Replace with:
1 1
OK
Cancel
Apply
Reset
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Python
Spark
Database
Var. File
Auto Data Prep
Select
Sample
Aggregate
Derive
Type
Filter
Graphboard
Auto Classifier
Auto Numeric
Auto Cluster
Table
Flat File
Database
Server: Local Server
142MB / 203MB

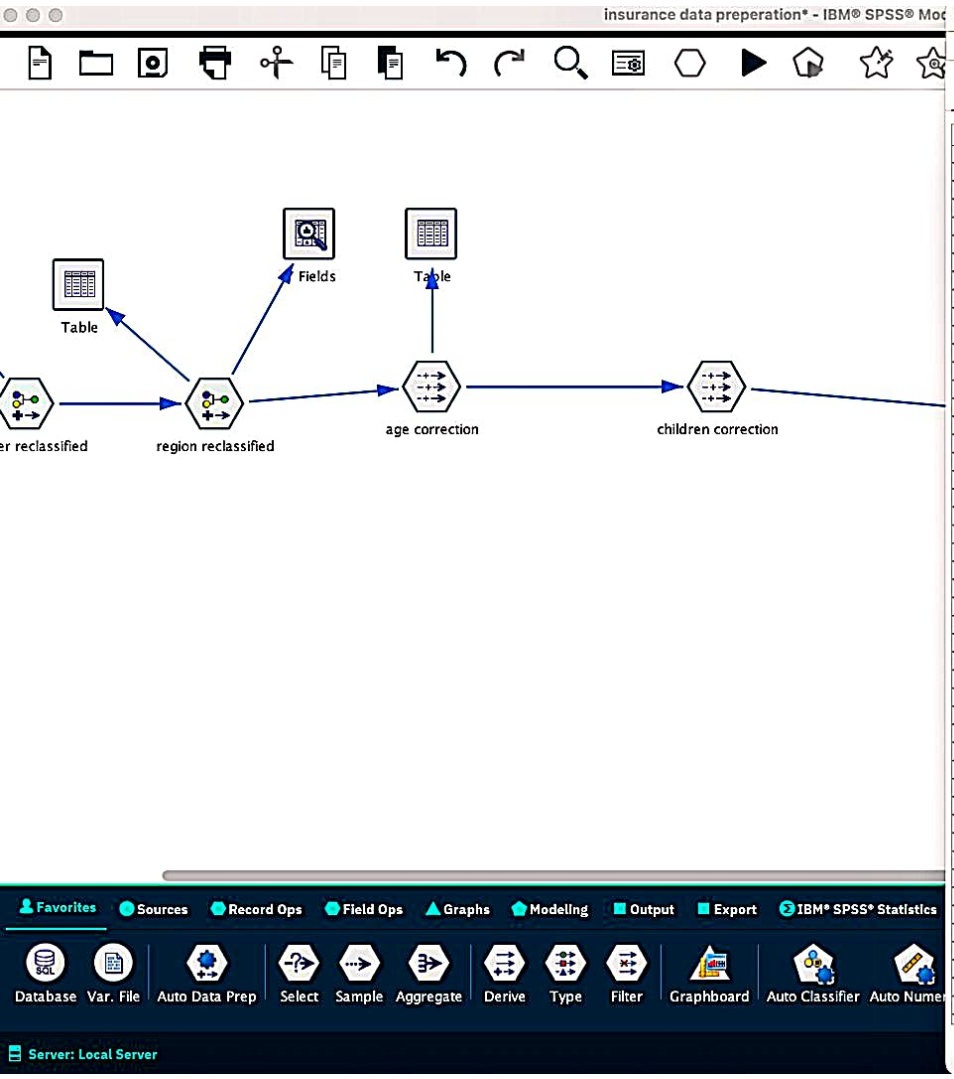


Table (7 fields, 1,338 records) #36

File Edit Generate

Table Annotations

	age	sex	bmi	children	smok...	region	charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622
7	46	female	33.440	1	no	southeast	8240.590
8	37	female	27.740	3	no	northwest	7281.506
9	37	male	29.830	2	no	northeast	6406.411
10	60	female	25.840	1	no	northwest	28923.137
11	25	male	26.220	0	no	northeast	2721.321
12	62	female	26.290	0	yes	southeast	27808.725
13	39	male	34.400	0	no	southwest	1826.843
14	56	female	39.820	0	no	southeast	11090.718
15	27	male	42.130	0	yes	southeast	39611.758
16	19	male	24.600	1	no	southwest	1837.237
17	52	female	30.780	1	no	northeast	10797.336
18	23	male	23.845	0	no	northeast	2395.172
19	56	male	40.300	0	no	southwest	10602.385
20	39	male	35.300	0	yes	southwest	36837.467
21	60	female	36.005	0	no	northeast	13228.847
22	30	female	32.400	1	no	southwest	4149.736
23	18	male	34.100	0	no	southeast	1137.011
24	34	female	31.920	1	yes	northeast	37701.877
25	37	male	28.025	2	no	northwest	6203.902
26	59	female	27.720	3	no	southeast	14001.134
27	39	female	23.085	0	no	northeast	14451.835
28	55	female	32.775	1	no	northwest	12268.632
29	23	male	17.385	1	no	northwest	2775.192
30	31	male	36.300	2	yes	southwest	38711.000
31	22	male	35.600	0	yes	southwest	35585.576
32	18	female	26.315	0	no	northeast	2198.190
33	19	female	28.600	5	no	southwest	4687.797
34	63	male	28.310	0	no	northwest	13770.098
35	28	male	36.400	1	yes	southwest	51194.559
36	19	male	20.425	0	no	northwest	1625.434
37	62	female	32.965	3	no	northwest	15612.193
38	26	male	20.800	0	no	southwest	2302.300
39	35	male	36.670	1	yes	northeast	39774.276
40	60	male	39.900	0	yes	southwest	48173.361
41	24	female	26.600	0	no	northeast	3046.062
42	39	female	36.630	2	no	southeast	4949.759
43	41	male	21.780	1	no	southeast	6272.477
44	37	female	30.800	2	no	southeast	6313.759
45	38	male	37.050	1	no	northeast	6079.672
46	55	male	37.300	0	no	southwest	20630.284
47	18	female	38.665	2	no	northeast	3393.356
48	28	female	34.770	1	no	northwest	3556.922

OK