第 3 章
データの前処理

概要

ジェリーはインターネットのデザインと広告の小さな会社のマーケティングのマネージャーです。ある日、上司からインターネットを使用しているユーザーに関する情報を新たに取得するように依頼がありました。この情報を使用して、企業はどのような人がインターネットを使用しているかを確認し、どのようにすれば企業のサービスを顧客へ販売できるかどうかを判断したいと考えていました。

ジェリーはオンライン調査を作成し、いくつかの人気のある Web サイト上へオンライン調査へのリンクを配置しました。分析を始めるために、２週間以内に十分なデータを収集しましたが、それらのデータを非正規化する必要があることや、データセットの中に欠損値や無効な値が含まれていることを発見し、ジェリーは分析を開始する前に、データを整備する必要があることに気づきました。

学習目標

- データの前処理の概念と目的を説明する
- 欠損値の対処方法を列挙する
- 一貫性のないデータを明確にし、対処する
- データ行の削除の目的を説明し、データ行の削除を実施する
- 変数の削減の目的を説明し、変数の削減を実施する

ハンズオン

これから実際に PC 上で操作していきますので、RapidMiner がインストールされていることが前提です。また、この本の関連サイトへアクセスするためには、インターネット接続が必要です。下記の関連サイトからは、各章で使用されるすべてのデータセットのコピーをダウンロード出来ます。

https://sites.google.com/site/dataminingforthemasses3e/

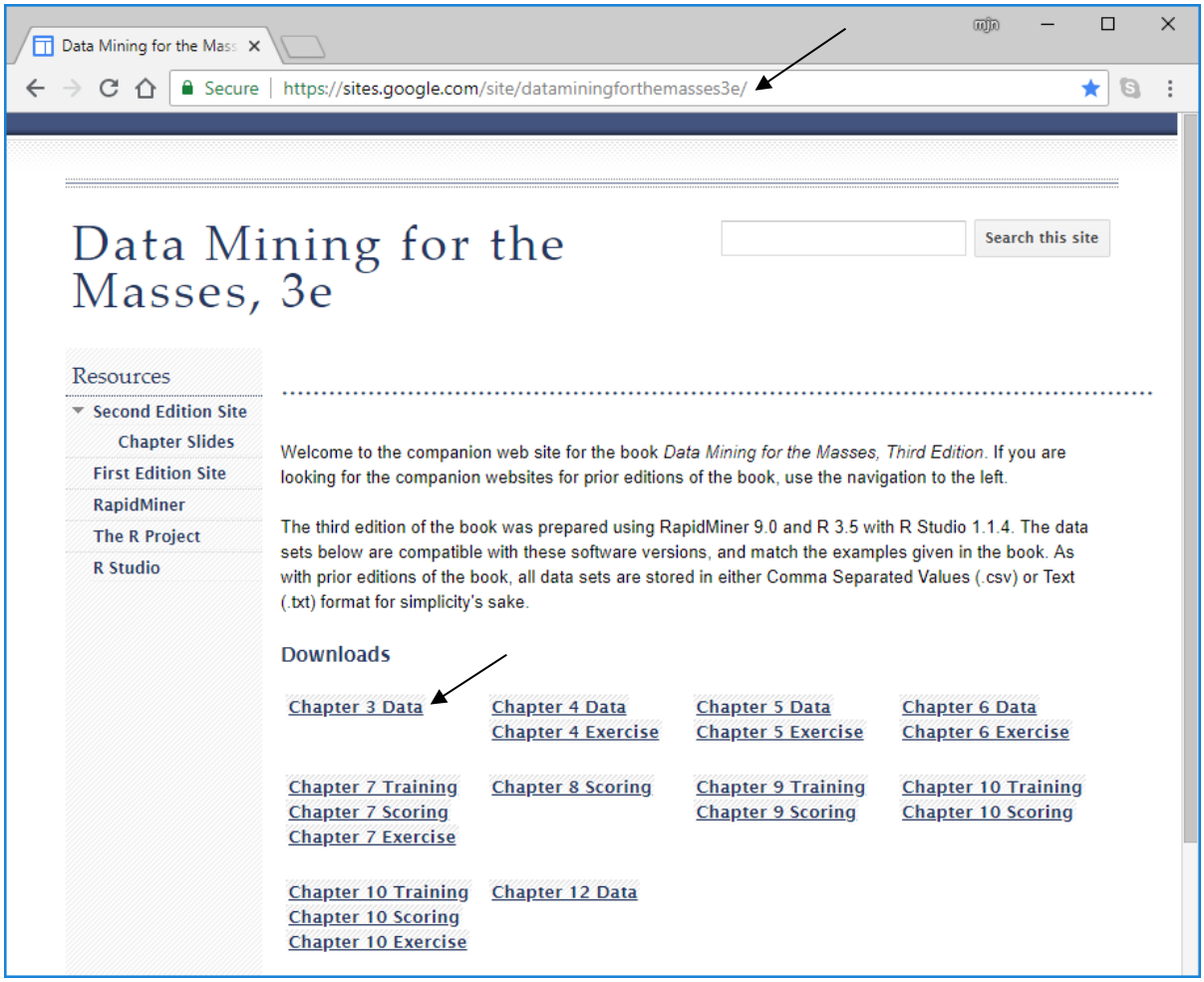
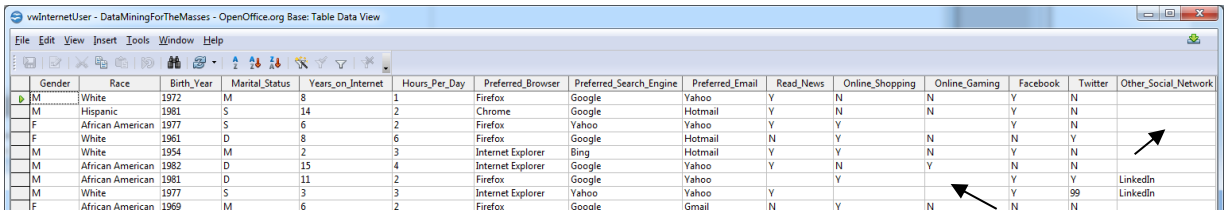


図 3-4. *Data Mining for the Masses, Third Edition* 関連サイト

上記サイトにて、「Chapter 3 data」のダウンロードリンクをクリックすることで第 3 章のデータセット（Chapter03DataSet.csv）をダウンロードすることが可能です。

データの読み込みと欠損値の対処について

データの前処理の最初のタスクは**欠損値（Missing data）**の除去です。図 3-9 をご覧ください。欠損値はゼロや他の値と違い空白になっています。値は不明、未定義、または未定です。欠損値はデータベースの世界では、**null** と呼ばれることもあります。データマイニングの目的に応じて、欠損値をそのままにするか、他の値に置き換えることが可能です。



Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N		Y	N	
M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	
F	African American	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y	N	
F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	
M	White	1954	M	2	3	Internet Explorer	Bing	Hotmail	Y	Y	N	Y	N	
M	African American	1982	D	15	4	Internet Explorer	Google	Yahoo	Y	N	Y	N	N	
M	African American	1981	D	11	2	Firefox	Google	Yahoo		Y		Y	Y	LinkedIn
M	White	1977	S	3	3	Internet Explorer	Yahoo	Yahoo	Y			Y	99	LinkedIn
F	African American	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	

図 3-9: 調査データ内の欠損値

例えば、Other Social Network 変数は、一般的な SNS 以外でユーザーが使用している SNS についての自由記述欄ですが、欠損値となっているユーザーは単に回答していないだけと考えられるため、欠損値はそのままにしておきます。一方で、Online Gaming 変数には、オンラインゲームを利用”Y”もしくは未利用”N”のどちらかの回答がありますが、欠損値となっているものもありますので対処する必要があります。

下記チュートリアルでは、RapidMiner での欠損値の対処方法について学びます。以下の手順に従ってデータにアクセスし、欠損値を修正します。

1) RapidMiner Studio を起動していない場合は、デスクトップにあるアイコンをダブルクリックして起動して下さい。起動すると次の画面が表示されますので、新しくプロセスを作成するために、Blank ボタンをクリックして下さい。

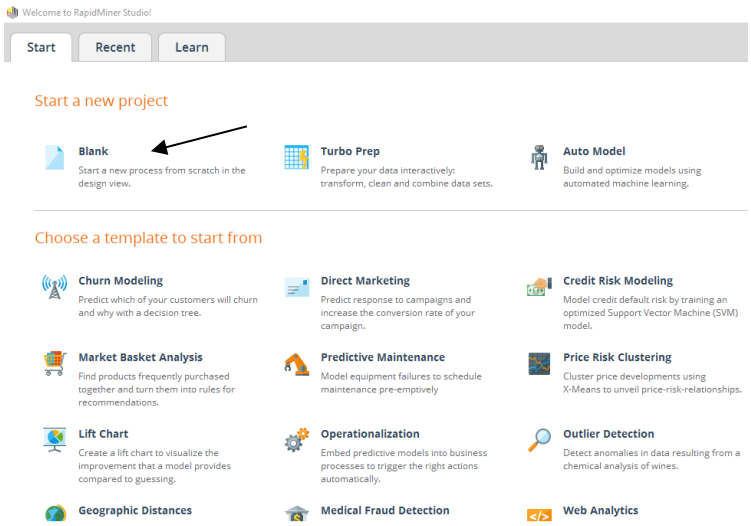


図 3-10. 新しいプロセスを作成

2) RapidMiner は、様々な種類のデータソースへアクセスすることが可能ですが、Studio の Free 版を使用している場合は、レコード数が 10,000 件以下に制限されます。本チュートリアルの目的としては 10,000 件以下で十分ですが、ビジネスにおいては、RapidMiner ライセンスをアップグレードすることをお勧めします。RapidMiner の画面上の左側にリポジトリがあります。これから、リポジトリを新たに作成する方法をご紹介します。まず、下記のようにリポジトリタブの右側にあるアイコン ≡ をクリックし、Create repository メニューをクリックします。

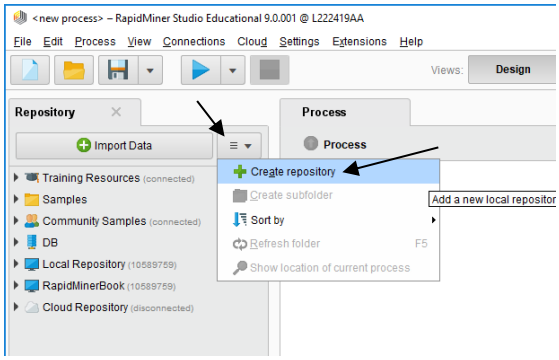


図 3-11. リポジトリを作成する

3) New local repository が選択された状態で Next ボタンをクリックすると、下記(図 3-12)の画面が表示されますので、「RapidMinerBook」と入力し、ディレクトリのパスはそのままで Finish ボタンをクリックします。

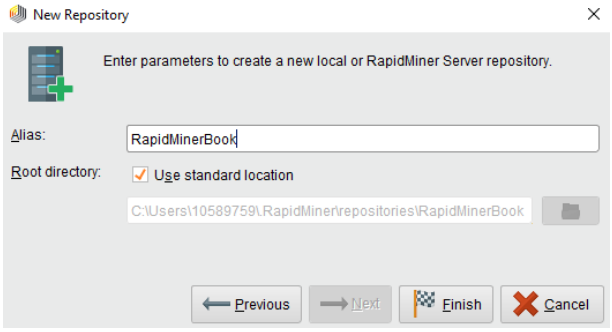


図 3-12. リポジトリの名前とディレクトリのパスを設定

4) クリックすると、下記のように RapidMinerBook リポジトリが作成されていることを確認出来ます。

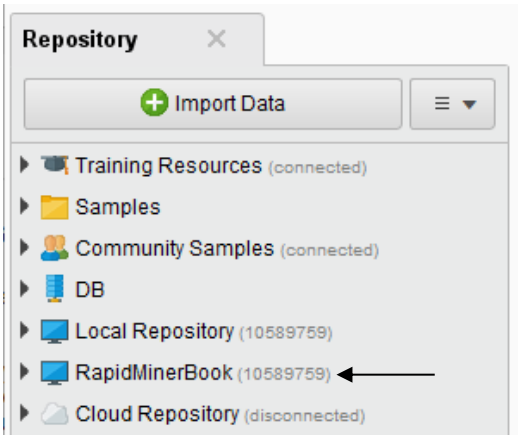


図 3-13. 新しいリポジトリが作成された

5) 次の画像内の矢印で示されている**オペレータ**と**リポジトリ**と**パラメータ** はよく使用されます。リポジトリは取り込んだデータを保存する場所です。オペレータはデータマイニングを実施するためのツールが配置されており、パラメータ

を変更することでオペレータの動作を変更することが可能です。 中央の領域はデザイン画面と呼ばれており、ここでモデルを作成します。

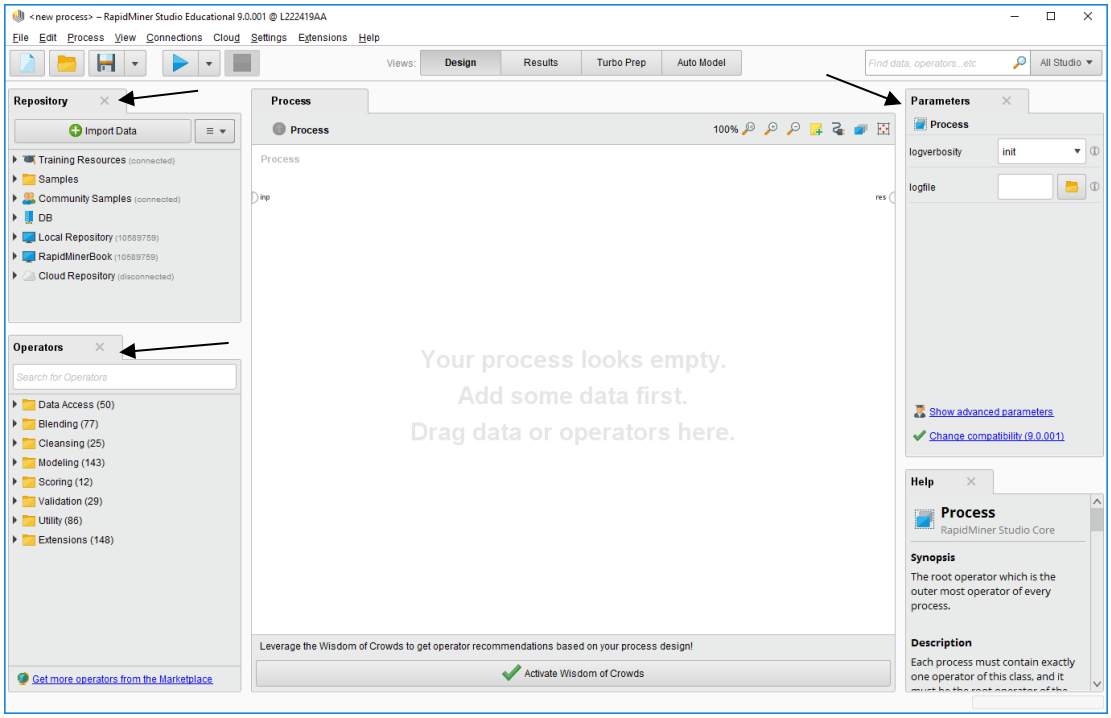


図 3-14. RapidMiner でよく使われるツール

6) 関連サイトからダウンロードした Chapter03DataSet.csv の変数は１５個で、１１レコードあります。これからこのデータを取り込みますが、もし関連サイトから Chapter03DataSet.csv をダウンロードされていない場合は、 関連サイトからデータをダウンロードします。次に、上記の図 3-14 のリポジトリタブの下にある Import Data ボタンをクリックします。そして、下記の図 3-15 の画面が表示されたら My Computer ボタンをクリックします。

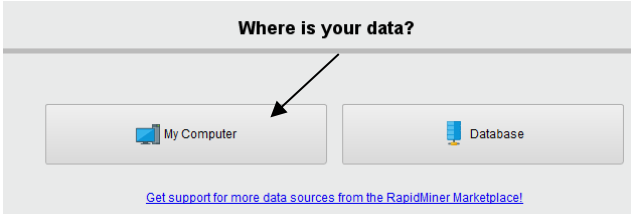


図 3-15. CSV ファイルを読み込む

7) 図 3-15 に表示されている‘My Computer’ボタンを押して、第 3 章のデータセットをダウンロードして保存したディレクトリへ移動します。正しいディレクトリへ移動するために小さな矢印のついたフォルダのアイコンを使う必要があるかもしれません。

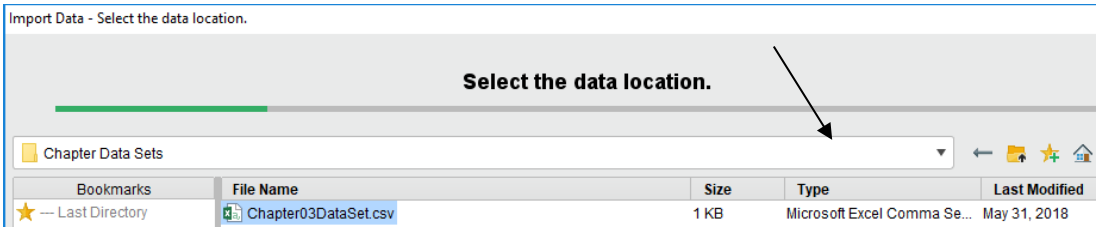


図 3-16. インポートするデータセットの場所

8) 図 3-16 のように、Chapter03DataSet.csv をクリックし、Next を選択します。

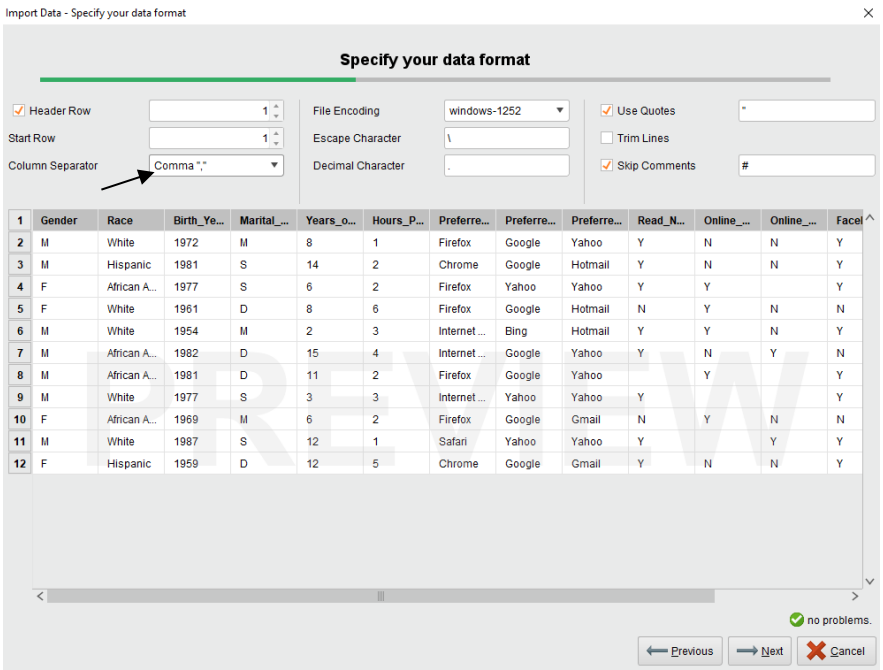


図 3-17. データ形式の指定

9) デフォルトでは、カンマ区切りとして認識します。もし、データセットがカンマ以外を区切り文字として使用している場合には、これを変更することができ

ます。変更する場所は、図 3-17 の黒い矢印で示されています。もし、データにカンマが含まれている場合は、データにない区切り文字、例えばタブ、セミコロン、パイプ(|)、または他の記号などを使えば、意図しない場所で列が区切られることを回避することができるでしょう。それでもデータを正しく区切れない場合は、正規表現(Regular Expression)を使いましょう。データの両端に"（クォーテーションマーク）を使用することもできます。これにより、"（クォーテーションマーク）内にあるカンマはデータの一部として認識され、"（クォーテーションマーク）の外にあるカンマは列の区切り文字として扱われます。もし、CSV ファイルがこのようなデータであれば、'Use Quotes'のボックスをチェックすることを忘れないようにしましょう。Chapter03DataSet.csv ファイル内のデータは列内のデータに"（クォーテーションマーク）も含まれておらず、また、列内のデータにカンマも含まれていないので、'Use Quotes'ボックスのチェックの有無による影響はありません。

デフォルトの取込設定では、CSV データセットの一行目をヘッダー行として指定しており、一行目が列名として認識されます。もし、CSV データセットの一行目の列が列名を含んでいない場合には、データセットをインポートする時に、図 3-18 の左端の矢印の Header Row のチェックが外れていることを確認しましょう。そうすれば、一行目のレコードがヘッダー行として扱われることはありません。

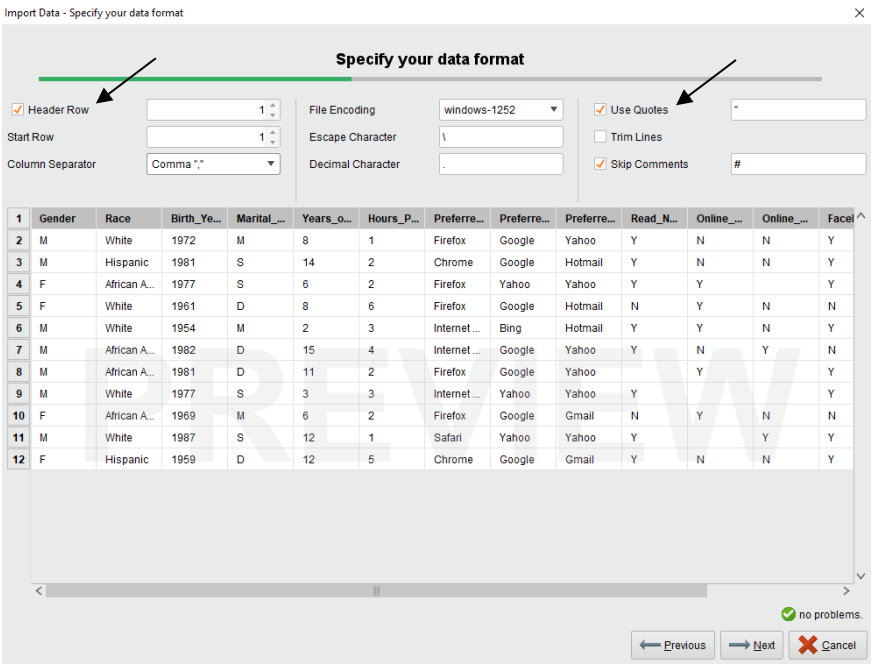
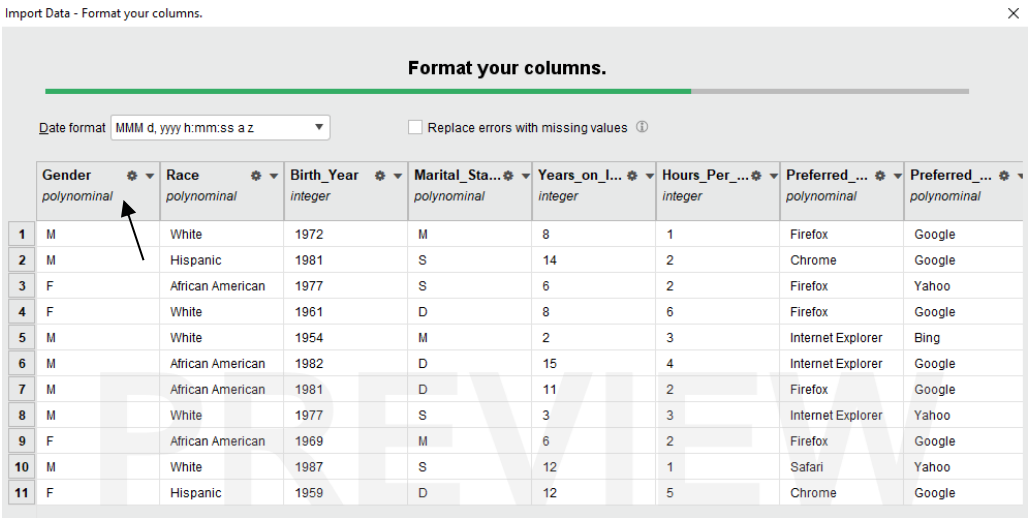


図 3-18. CSV データの列のプレビュー

10) データのプレビューで、それぞれの列が正しく区切られていることを確認したら、Next をクリックします。



	Gender	Race	Birth_Year	Marital_Sta...	Years_on_l...	Hours_Per_...	Preferred_...	Preferred_...
	polynominal	polynominal	integer	polynominal	integer	integer	polynominal	polynominal
1	M	White	1972	M	8	1	Firefox	Google
2	M	Hispanic	1981	S	14	2	Chrome	Google
3	F	African American	1977	S	6	2	Firefox	Yahoo
4	F	White	1961	D	8	6	Firefox	Google
5	M	White	1954	M	2	3	Internet Explorer	Bing
6	M	African American	1982	D	15	4	Internet Explorer	Google
7	M	African American	1981	D	11	2	Firefox	Google
8	M	White	1977	S	3	3	Internet Explorer	Yahoo
9	F	African American	1969	M	6	2	Firefox	Google
10	M	White	1987	S	12	1	Safari	Yahoo
11	F	Hispanic	1959	D	12	5	Chrome	Google

図 3-19. 列名を設定

11) 図 3-19 の画面で、RapidMiner はそれぞれの列のデータ型を推測します。データ型とは数値、文字列、または日付のように列が持つデータの種類のことです。

これらはこの画面の各列の名前の右側にある歯車のアイコンの横の下三角▼をクリックすると変更できます。歯車アイコンによって列の名前やロールを変えることも、また、除外する列を選択することもできますが、今回はデフォルトの設定のままにします。後の章で、さまざまな種類のデータやデータマイニングのタスクに適したデータ型やロールについての設定をさらに扱います。もし、データが日付を含んでいる場合、必要であれば、Date format（日付フォーマット）を変更して、RapidMiner にデータの日付のフォーマットをどのように読ませるのか必ず教えましょう。これらの各機能は図 3-20 の黒い矢印で示してあります。設定はデフォルトのままで、Next をクリックし、先に進みます。

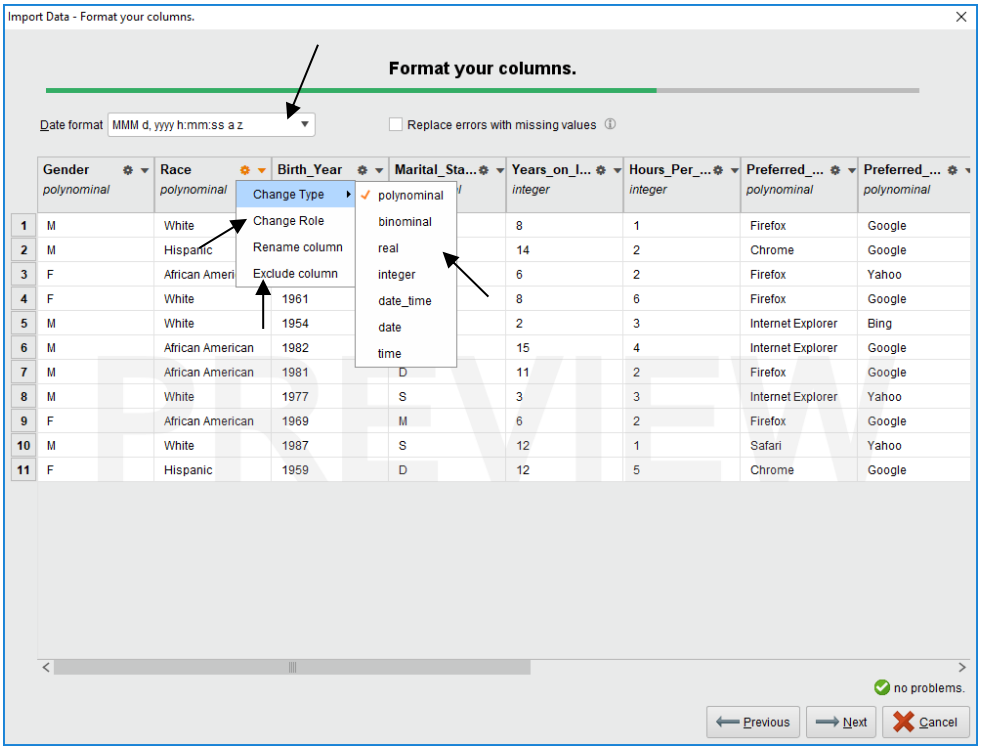


図 3-20. データ型、ロール、日付フォーマットの設定と列のインポート

12) 図 3-21 のように、最後のステップでは RapidMinerBook リポジトリをデータの保存場所として選択し、データセットに Chapter03DataSet と名前をつけます。

Finish をクリックすると、私たちが作りたい様々なタイプのデータマイニングプロセスでこのデータセットを利用できるようになります。

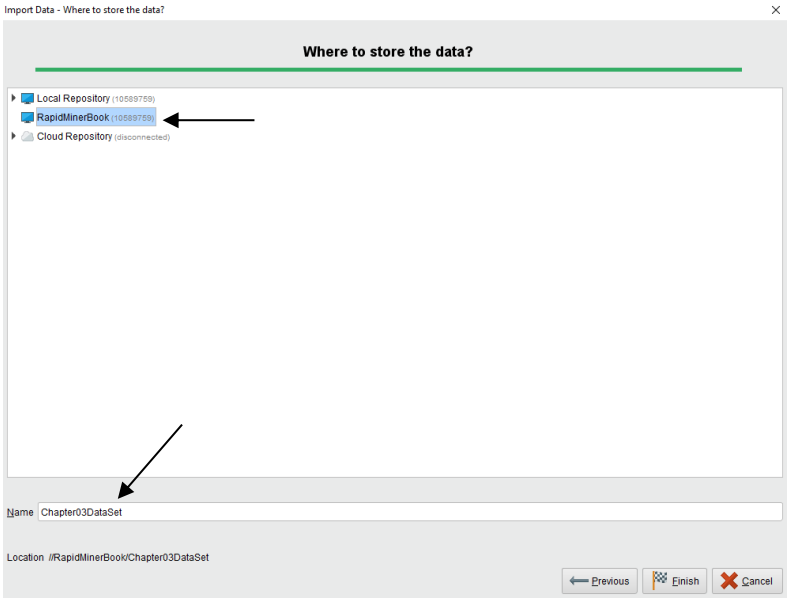
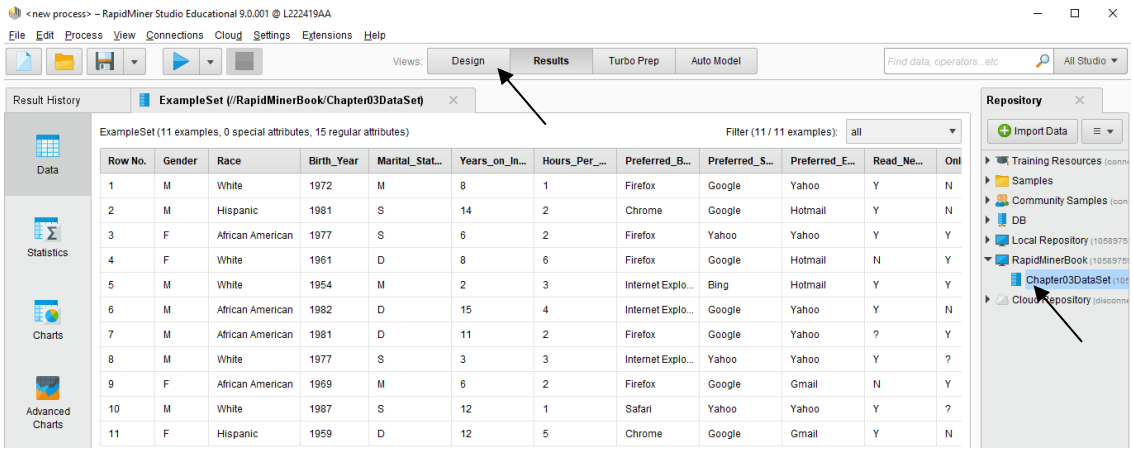


図 3-21. リポジトリの選択とデータセット名の設定

13) Finish をクリックすると、結果画面(Result Perspective)、または結果ビュー(Result view)と呼ばれるビューにデータが表示されます。



Row No.	Gender	Race	Birth_Year	Marital_Stat...	Years_on_In...	Hours_Per_...	Preferred_B...	Preferred_S...	Preferred_E...	Read_Ne...	Onl
1	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N
2	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N
3	F	African American	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y
4	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y
5	M	White	1954	M	2	3	Internet Explo...	Bing	Hotmail	Y	Y
6	M	African American	1982	D	15	4	Internet Explo...	Google	Yahoo	Y	N
7	M	African American	1981	D	11	2	Firefox	Google	Yahoo	?	Y
8	M	White	1977	S	3	3	Internet Explo...	Yahoo	Yahoo	Y	?
9	F	African American	1969	M	6	2	Firefox	Google	Gmail	N	Y
10	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	?
11	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N

図 3-22. 結果画面(Result Perspective)または結果ビュー(Result view)

14) 図 3-23 の画面の左側では、データの記述統計量(Σ アイコン)を見たり、データのチャートやアノテーションを作れることがわかるでしょう。例えば、Statistics

をクリックしたら、すぐにインターネット歴の平均が 8.818 年であるとか（Years_on_Internet 変数をクリックすると、標準偏差(Deviation)が 4.332 であるなども見られます）、Preferred_Search_Engine 変数の統計から Google(7)が最も人気のサーチエンジンであることなどがすぐにわかるでしょう。

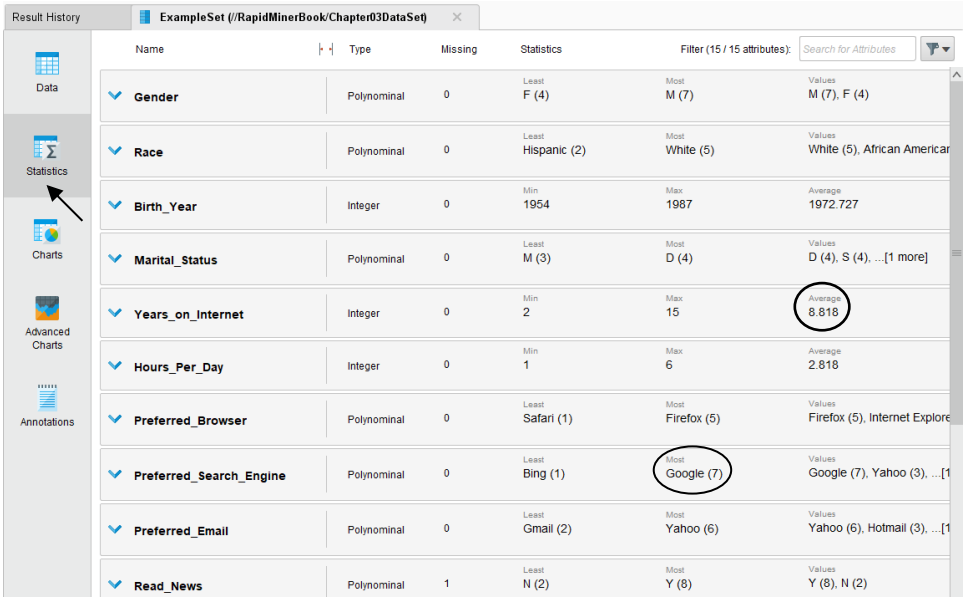


図 3-23. 記述統計量

15) 画面上部にある Design ビューボタンをクリックし、デザイン画面(またはビュー)へ切り替えます。そして、RapidMinerBook レポジトリ下の Chapter03DataSet を選択して、プロセスウィンドウへドラッグ&ドロップしましょう。

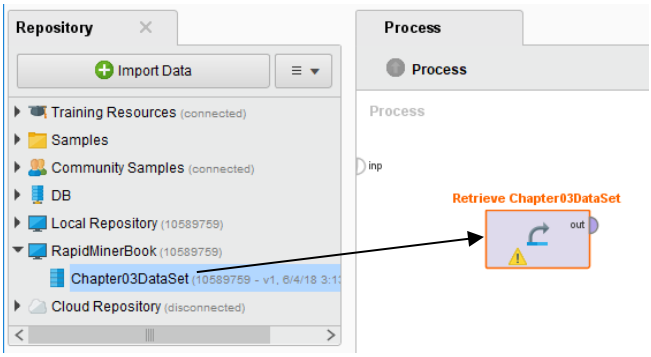


図 3-24. リポジトリ下のデータをプロセスへ追加

16) プロセスウィンドウにある長方形のものをオペレーターと言います。 プロセスウィンドウにデータセットをドラッグ＆ドロップすることによって、Chapter03DataSet にアクセスする Retrieve オペレータを追加しました。Retrieve オペレータをクリックして選択すると、このオペレータが Chapter03DataSet にアクセスされていることが右側にあるパラメータパネルでわかります。オペレータが選択されている時は、オペレータの名前と外枠がオレンジに変わります。Retrieve オペレータは単純にデータセットをリポジトリから取得し、プロセスウィンドウで利用できるようにしてくれます。オペレータの端とプロセスウィンドウの端にある小さな半円は、ポートと呼ばれています。Retrieve オペレータの右側にアウトプット(out)ポートが、プロセスウィンドウの右側に結果(res)ポートが見えるでしょう。この 2 つのポートを線でつなぐことができます。out ポートをクリックしてから、res ポートをクリックするとポート同士が線でつながります。

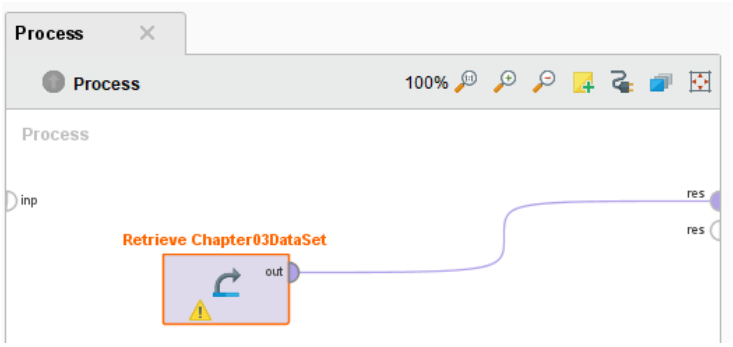


図 3-25. *out* ポートを *res* ポートへ接続

17) オペレータを組み合わせて線でつなぎ、データマイニングの流れを構築します。データマイニングプロセスを実行してその結果を見るには、図 3-26 の矢印にあ

る青い三角の実行ボタンをクリックします。これにより、先ほどのデータセットのインポートが完了した時に見られたような結果画面に切替わります。

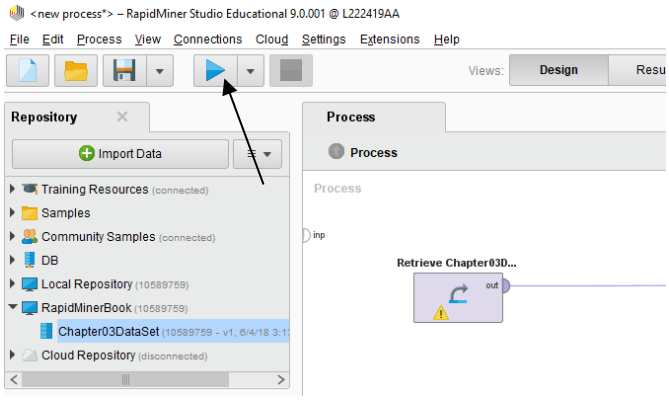


図 3-26. プロセスを実行し結果を確認

18) 画面の最上部にある Design と Results ボタンを使うことにより、デザイン画面と結果画面を切り替えることができます。結果画面は、どこに欠損値があり、それをどう処理するかを決めるのに非常に役立ちます。Online Game 変数を例にあげましょう。結果画面の Statistics タブでは、'N'が 6 件、'Y'が 2 件、そして欠損値（Missing）が 3 件あることがわかります。

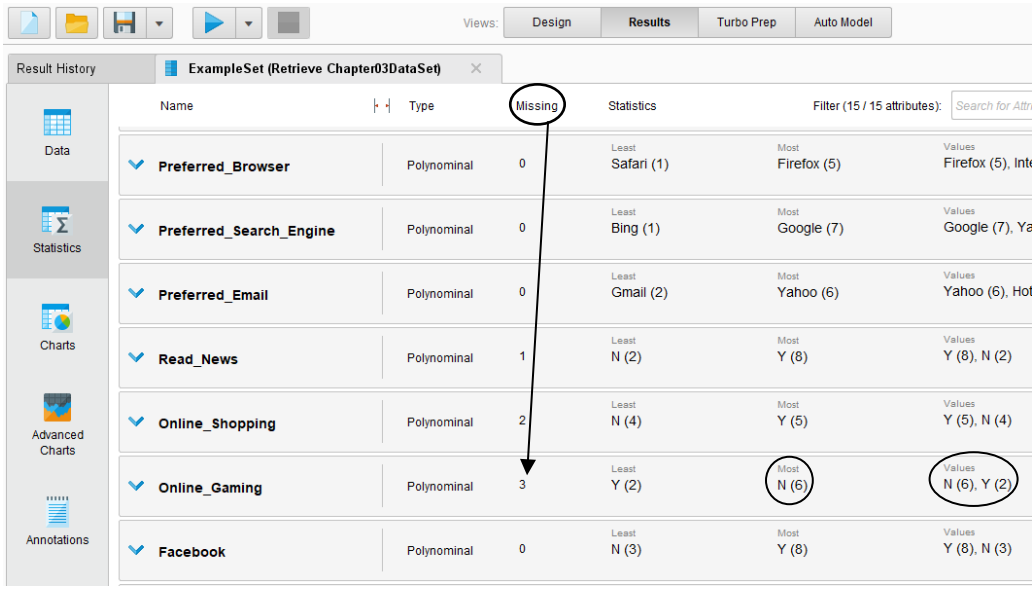


図 3-27. 結果画面の記述統計量

中心的傾向の測定値 — 平均、中央値、最頻値（モード） —　はデータ内に見つかる欠損値を置き換えるのに使用されることがあります。最頻値（モード）、または最も一般的な値で欠損値を置き換えることもできます。もちろん、これは最頻値（モード）や最も一般的な値がすべての観測の正確な値であることを前提としていて、これは間違っている場合があります。ですから、これから RapidMiner でどのように欠損値を扱うかを、現在の例のデータ中の欠損値を変更して見せますが、それが欠損値を扱う上で常に適切な方法ではないということに注意して下さい。RapidMiner で Online_Gaming 変数内の 3 件の欠損値を'N'に変更するためには、画面上部の Design ボタンを選択し、デザイン画面へ戻ります。

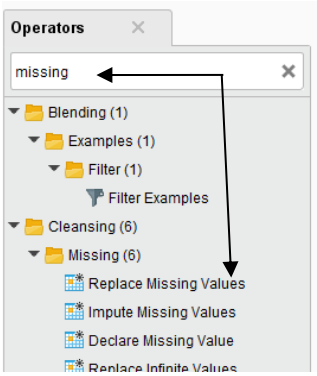


図 3-28. 欠損値（missing values）を処理するオペレータを検索

19) 図 3-28 のオペレータパネルで該当のオペレータを見つけるために、フォルダの階層を移動したり、検索することができます。RapidMiner はたくさんのオペレータを提供しているため、時々、自分が欲しい物を探すのが難しい場合があります。図 3-28 の中で黒い矢印で示されている便利な検索ボックスへキーワードを入力することで、該当のオペレータを見つけやすくするができます。この検索ボックスに'missing'とタイプしてみましょう(図 3-28 参照)。RapidMiner が自動

でオペレータ群の中からオペレータの名前または概要(description)にこのワードが含まれるものを探します。検索するといくつか見つかりましたが、私たちは、欠損値を置き換えたいと思っていますので、Cleansing フォルダーの中の Missing サブフォルダにある Replace Missing Values オペレータを選択し、図 3-29 のようにプロセスウィンドウの中の線の上にドラッグしましょう。マウスのカーソルを線上に合わせると、線が少し太く変わります。そして、マウスボタンを離すと、オペレータが接続されます。Replace Missing Values オペレータを離しても接続されない場合には、手動で線をつなぐことが可能です。Retrieve オペレータの out ポートをクリックしてから、Replace Missing Values オペレータの exa ポートをクリックするだけでオペレータ同士が接続されます。exa は Example Set のことを表しています。'examples'というワードは RapidMiner の中では、データセットにおける行として使われています。プロセスを実行したときに、結果が表示されるように、Replace Missing Values オペレータの右側にある exa ポートから結果(res)ポートに接続されていることを確認しましょう（図 3-29）。

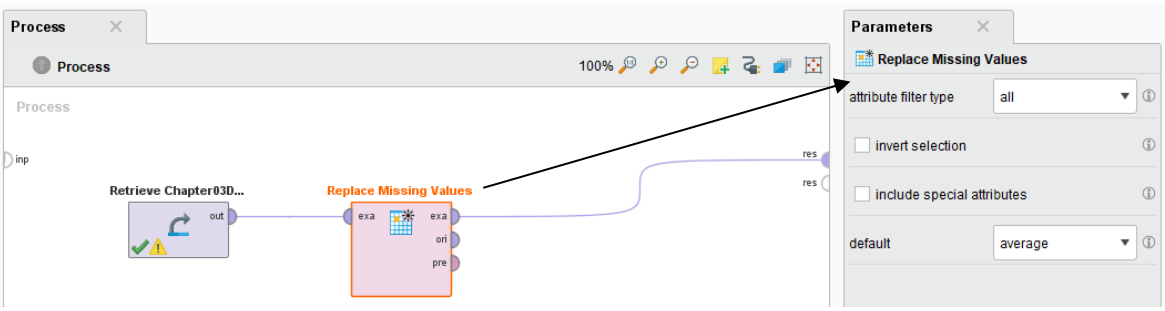


図 3-29. Replace Missing Values オペレータを追加

20) RapidMiner の中で、オペレーターが 選択されているときには、オペレータの外枠の長方形がオレンジ色になります(図 3-29 で Replace Missing Values オペレータが選択されているのが見られます)。この時に、選択されたオペレータのパラメ

ータやプロパティを変更することができます。今回のチュートリアルでは、Online_Gaming 変数の欠損値 3 つ全てを'N'で置き換えることに決めました。なぜならこれがこの変数における最も多い回答(モード)だからです。これを行うために、図 3-30 のように attribute filter type を 'single' に変更します。すると attribute とラベルのついたドロップダウンリストが現れますので、'Online_Gaming' 変数を選択します。次に default とついたドロップダウンリストの候補の中から'value'を選択します。これにより、replenishment value ボックスが表示されます。図 3-30 のように'N'をこのボックスにタイプします。選択したオプションに応じて使用可能なオプションが変わるため、全てのオプションを表示するには、画面を広げるか、パラメータパネルの右側にある縦方向のスクロールバーを使用する必要がある場合があるので、注意しましょう。

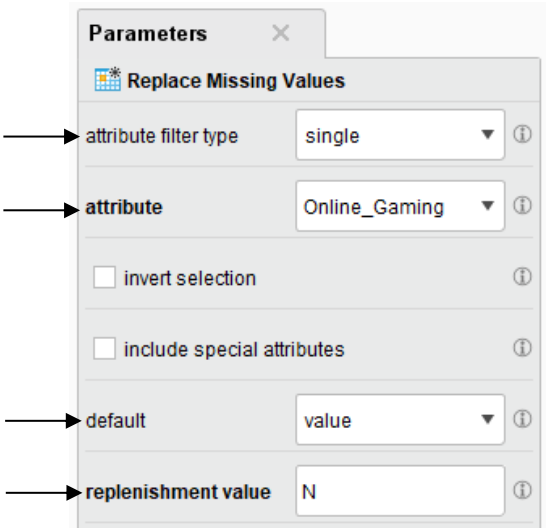


図 3-30. 欠損値のパラメータ設定

21) パラメータパネルの中には、他にもたくさんのオプションが用意されています。ここでは、全てを見てみることはしませんが、自由に試して見てください。例えば、'attribute filter type'の 'subset'を使用することで、データセット内の複数

(subset)の変数を変更することができますが、今回は‘single’のままにしておいて下さい。パラメータをセットしたら、実行ボタンをクリックしましょう。プロセスが実行され、結果画面に切り替わりますので、次に Statistics アイコンをクリックします。結果画面が図 3-31 のように見えるはずです。

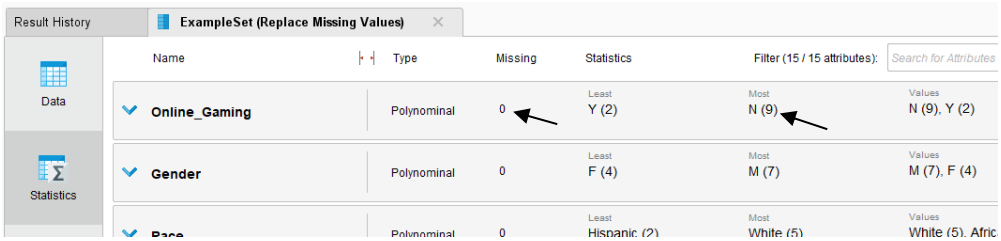


図 3-31. 欠損値を変更した結果

22) Online_Gaming 変数がリストの一番上に移動し、欠損値（Missing）が 0 になったことが見えるでしょう。次に Data アイコン(statistics アイコンのすぐ上にある)をクリックしましょう。Online_Gaming 変数には‘Y’と‘N’の値のみが入力されていることが分かります。この変数の欠損値を全て置換することに成功しました。一方で、Online_Shopping 変数には 2 つの欠損値があります。疑問符(?)は、欠損値を表しています。この変数では、nulll 値を最頻値（モード）で置き換えず、レコードごとデータを削除します。

レコードの削除

デザイン画面に切り替えます。 次は、データセット内のレコードを削除する方法について説明します。

1) オペレータパネル内の検索ボックスに、’ filter’という単語を入力します。これは、この例で使用する Filter Example オペレータを見つけるのを助けてくれます。Filter Examples オペレータをドラッグして、Replace Missing Values オペレータのすぐ後に接続します。プロセスウィンドウは図 3-32 のようになります。

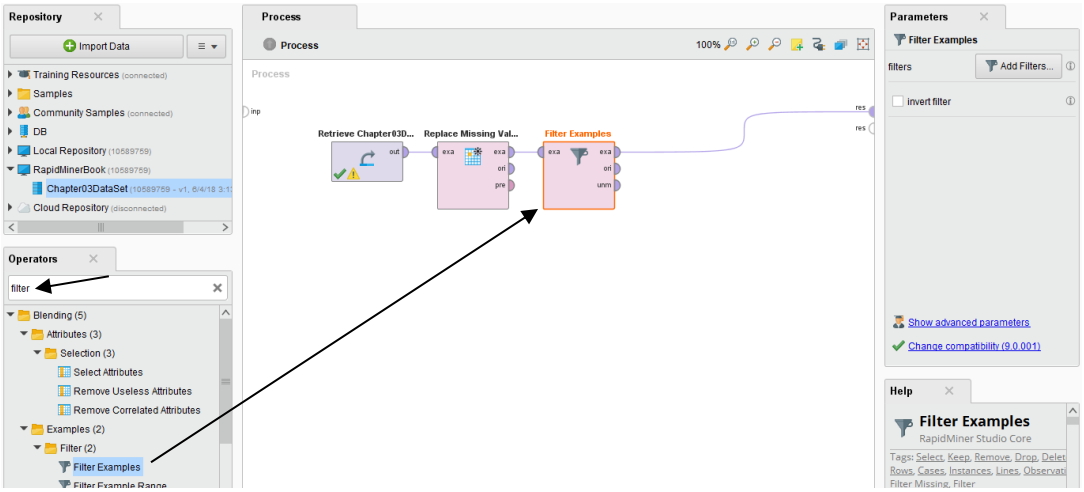


図 3-32. フィルターを追加

2) Filter Examples オペレーターを選択した状態で、パラメータパネルを確認します。Add Filters…ボタンをクリックします。すると、図 3-33 のように Create Filters 画面が表示されますので、最初のドロップダウンリストで Online_Shopping 変数を選択します。次に 2 番目のドロップダウンリストで ’is missing’ を選択します。

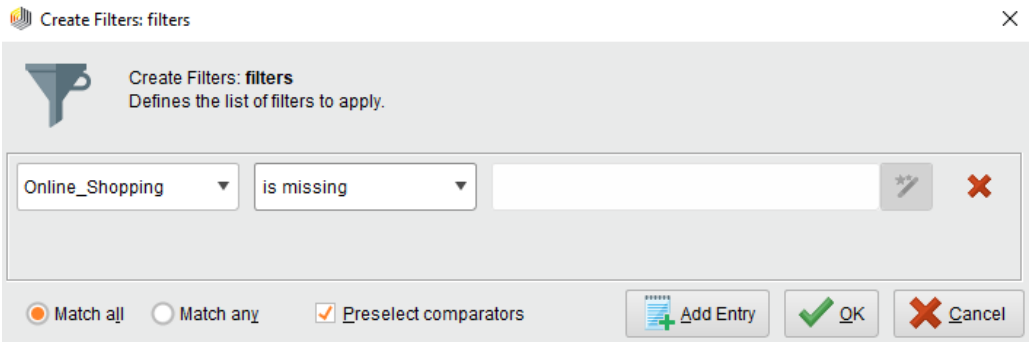


図 3-33. フィルターを作成

3) OK をクリックして実行し、結果(Result)画面に切り替わると、Online_Shopping に欠損値があった 2 件しか表示されていないことがわかります。他のすべてのレコードは削除されているため、これは意図したものではありません。 私たちは Online_Shopping の値が欠損している 2 つのレコードを削除し、他のレコードを残したいと考えています。これを直す方法は 2 つあります。 デザインビューに切り替えます。 パラメータパネルに、'invert filter'というチェックボックスが表示されています。

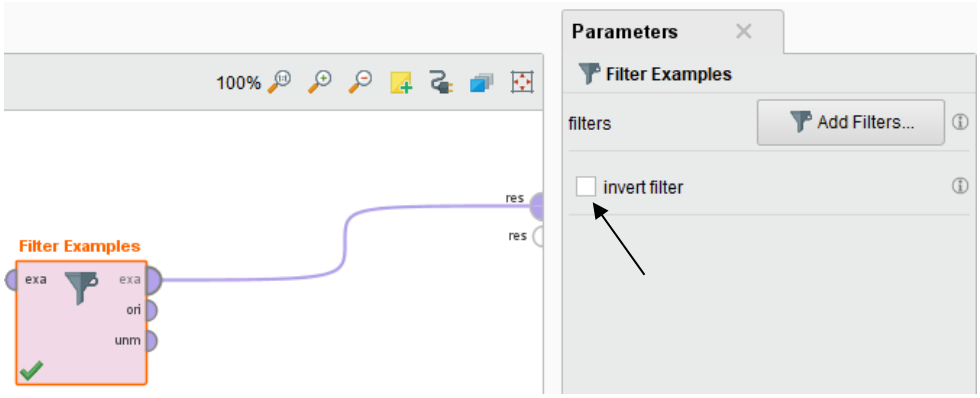


図 3-34. invert filter オプション

4) これをクリックしてプロセスを再実行すると、Online_Shopping 変数の欠損値以外の値を持つ 9 件のレコードが保持され、2 件がフィルターにより削除されてい

ることがわかります。または、Add Filters… ボタンをもう一度クリックして Create Filter ウィンドウを再度開き、今度は ‘is missing’ ではなく ‘is not missing’ を選択して、問題を修正することもできます。これでも同じ結果が得られます。1 つの Filter Examples オペレータで複数のフィルターを追加できることも覚えておいてください。Create Filters ウィンドウを開いた状態で、Add Entry ボタンをクリックすれば、他の変数のフィルターを設定するための 2 行目が作成されます。 さまざまな条件でフィルタリングを試してみてください。

フィルタリング以外の方法でもデータを削減することができます。 以下の手順に従って、RapidMiner でデータセットのサンプリングを実施しましょう。

1) すぐ前に紹介した検索のテクニックを使用して、オペレータ検索機能を使用して Sample というオペレータを見つけ、これをプロセスへ追加します。パラメータパネルで、サンプルを‘relative’ (相対) サンプルになるように設定し、サンプル比率(sample ratio) フィールドに.5 と入力して、結果のデータセットが全体の 50%を保持するように指定します。 画面は図 3-35 のようになります。

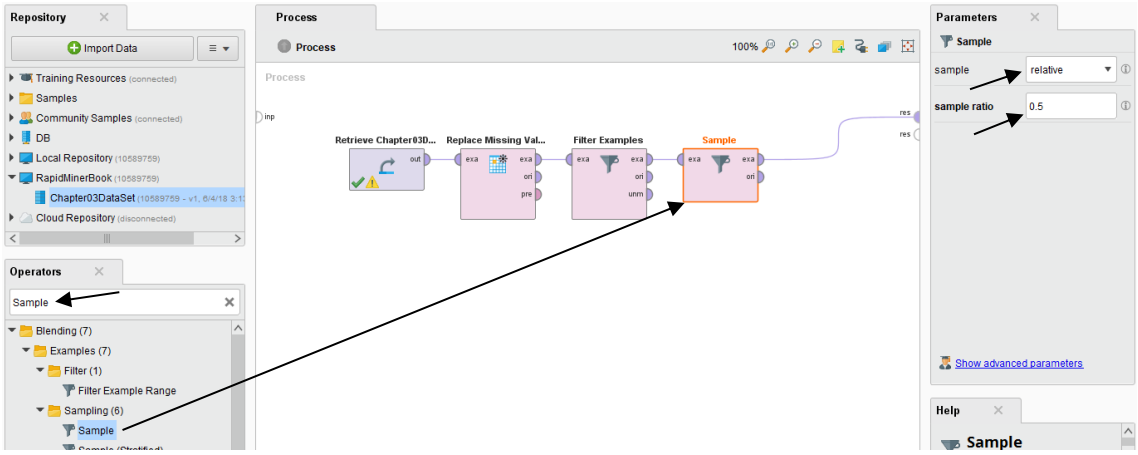


図 3-35. データセットのランダムサンプリング（50%）

2) ここでプロセスを実行すると、結果画面には Filter Examples オペレーターが Online_Shopping の値を持たないレコードを削除した後に残った 9 件のデータから、ランダムに選択された 4 件または 5 件のデータのみがあるのがわかるでしょう。

このように、データセット内のレコード数を減少させてデータを削減させる方法は多く存在し、さまざまな理由があることがわかります。次に、不整合(inconsistent)データの処理に移りますが、フィルタリングを実施した際に不整合データを削除してしまったので元の状態に戻す方法について説明します（図 3-37 のような状態にします）。デザイン画面に戻り、Sample オペレータをクリックします。 次に、右クリックして[削除(Delete)]を選択するか、キーボードの Delete キーを押します。次に Filter Examples オペレータをクリックします。 削除するのではなく、これを無効にします。無効にするには、右クリックして Enable Operator オプションを選択するか（トグルスイッチでオン/オフを切り替える仕様）、またはキーボードで Ctrl + E を押します（こちらもトグル）。無効にした Filter Examples オペレータを、メインプロセスの隅にドラッグして、邪魔にならないようにします。結果(res)ポートに繋がっていた線が消えている場合は、Replace Missing Values オペレータの exa ポートから res ポートに再接続します。プロセスが図 3-37 のような状態になっていることを確認して下さい

不整合データの対処

不整合データは欠損値とは違います。不整合データとは、値は存在しても、それが妥当な値または意味のある値ではないことを指します。Twitter 変数の中には不整合データが存在します（図 3-36）。

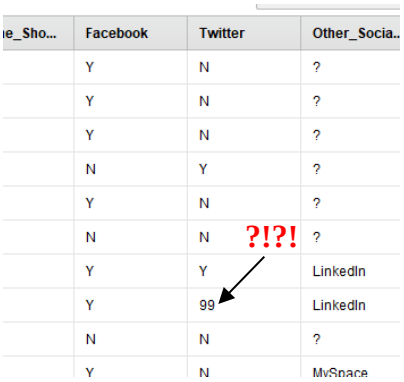


図 3-36. Twitter 変数の不整合データ

99 は何をあらわしているのでしょう？Twitter 変数の有効な値は「Y」と「N」の 2 つだけのようです（つまり、回答者は Twitter を使用しているか、使用していないかのどちらかです）。これらの値は統一されていないので、意味がありません。 データマイニングを行う者として、Online_Shopping に欠損値があったレコードの場合と同様に、レコードをフィルターで除外するか、特定の値を他の値で置き換えられるオペレータを使用するかを選択することができます。

1) もし他の画面を開いていたら、デザイン画面へ戻りましょう。図 3-37 のような画面になるように、Sample オペレータを削除（Delete）し、Filter Examples オペレータを削除もしくは無効（Ctrl + E）にしてあることを確認しましょう。

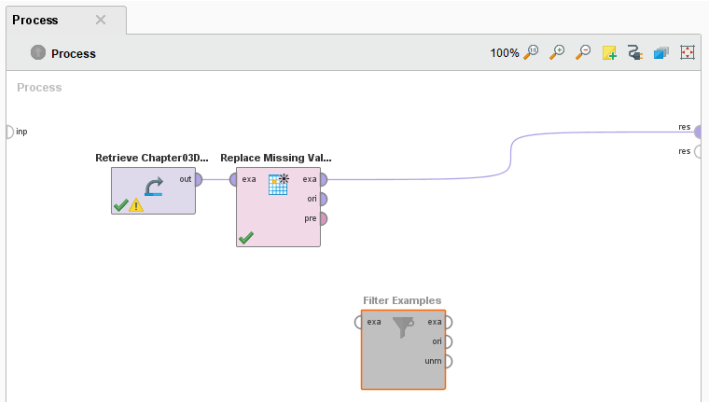


図 3-37. Filter Examples オペレータを無効にし、Sample オペレータを削除

2) Replace Missing Value オペレータは、Online_Gaming 変数の値のみを変更するだけですので、データセットのレコードを削除することはないため、削除する必要はありません。図 3-38 のようにオペレータパネルの検索機能を使用して、Replace(置換)と呼ばれるオペレータを検索し、プロセスへ追加します。

3) パラメータパネルで、attribute filter type を ‘single’ に変更し、attribute へ ’Twitter’ を指定します。 replace what のフィールドへ、これから私たちが置換したいと思っている値、99 を入力しましょう。そして、回答者の約 80%が Twitter を使用していないと回答しているので、今回は、‘replace by’フィールドへ最頻値（モード）の「N」を入力します。

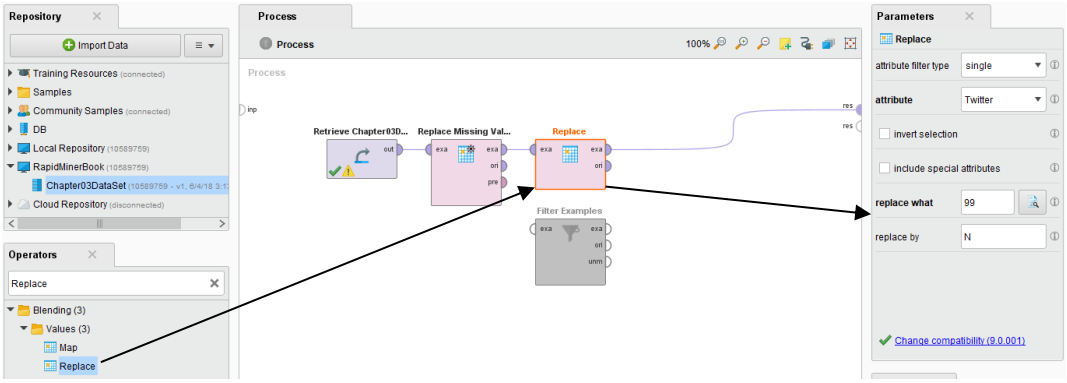


図 3-38. Replace オペレータのパラメータ設定

4)　プロセスを実行し、Statistics(統計)タブをクリックします。 図 3-39 を見ると、Twitter 変数には「N」の値が 9 つ、「Y」の値が 2 つあることがわかります。

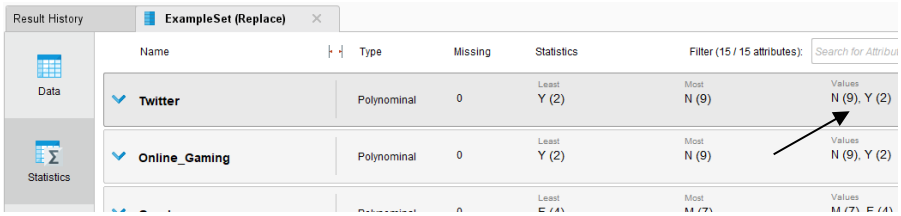


図 3-39. 不整合な値を整合性のある値で置換

変数の削減

多くのデータセットにおいて、与えられた質問への回答を出すのに無関係な変数があることがわかるでしょう。第 4 章では、相関関係または特定の変数間の関係の強さを評価する方法について説明します。データセット内の特定の変数が、特定の質問に答えるのに関係ないという理由だけで、それらの変数が絶対に関心のある変数にならないということではないということを覚えておいてください。 このような理由から、この第 3 章の前半でデータセットをインポートするときにすべての変数を取り込みました。興味のない変数や無関係な変数は、次の手順で簡単に除外できます。

1)　デザイン画面に戻ります。オペレータの検索フィールドで、'Select Attributes' と入力します。Select Attributes オペレータがオペレータのエリアの中のフォルダの階層の中に現れますので、ドラッグして Replace オペレータと結果ポート（res）の間に収まるようにします。画面は図 3-40 のようになります。

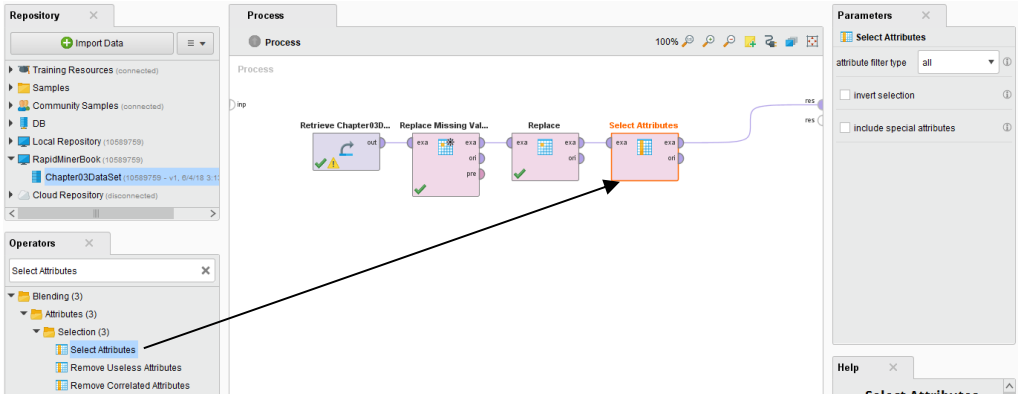


図 3-40. データセットの変数を選択

2) パラメータパネルで、attribute filter type を 'subset' に設定し、Select Attributes ボタンをクリックします。 図 3-41 のようなウィンドウが表示されます。

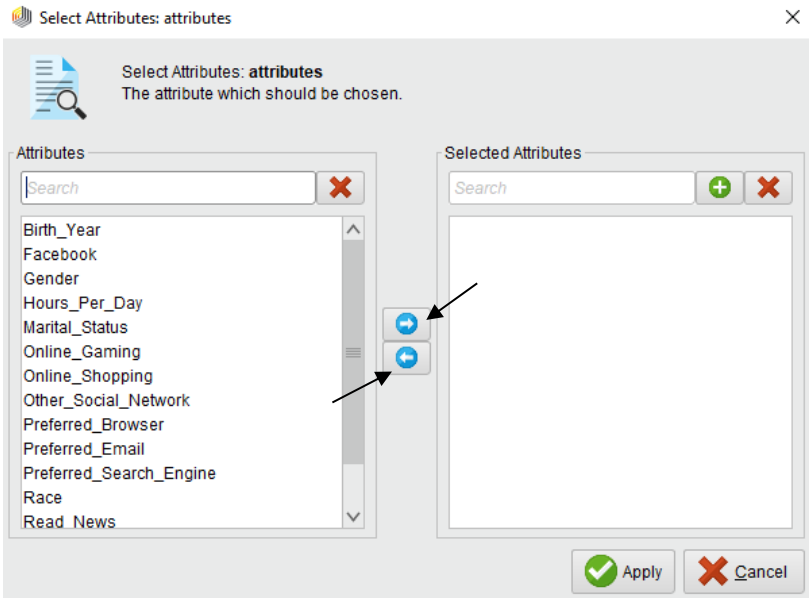


図 3-41. 変数選択画面

3) 青い右矢印と左矢印を使用して、変数を選択できます。この例では、Birth_Year、Gender、Marital_Status、Race、そして Years_on_Internet を選択し、右矢印を使用して右側の Selected Attributes (選択した変数)エリアへ移動させます。一度に複数の変数を選択するには、コントロール(Ctrl)またはシフト(Shift)キーを押しなが

ら、選択したい、または除外したい変数をクリックします。Apply ボタンを選択しプロセスを実行すると、選択した変数のみが結果画面に表示されます。

その他のデータ操作

この章ではこれまで、RapidMiner リポジトリにインポートした第 3 章のデータセットを使用して作業してきました。しかし、時間の経過とともに変化する可能性のあるデータセットがあり、それが増大したり変化したりする場合ではどうでしょうか。この例では、データセットを RapidMiner リポジトリへインポートせずに、直接 CSV ファイルを読み取る方法についてご紹介します。

1) オペレータの検索ボックスで、「Read」と入力します。

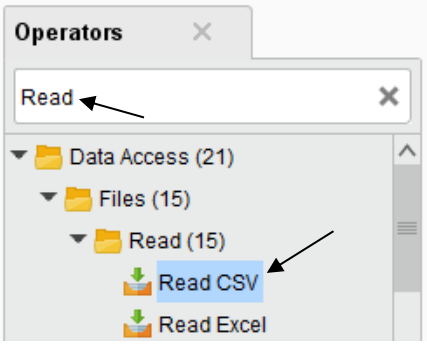


図 3-42. RapidMiner で Read 系のオペレータを検索

2) Read CSV オペレータをプロセスウィンドウにドラッグ&ドロップします。この例では、前に使用した全てのオペレータをプロセスから削除しています(Retrieve、Replace Missing Values、Replace、Select Attributes オペレータなど)。

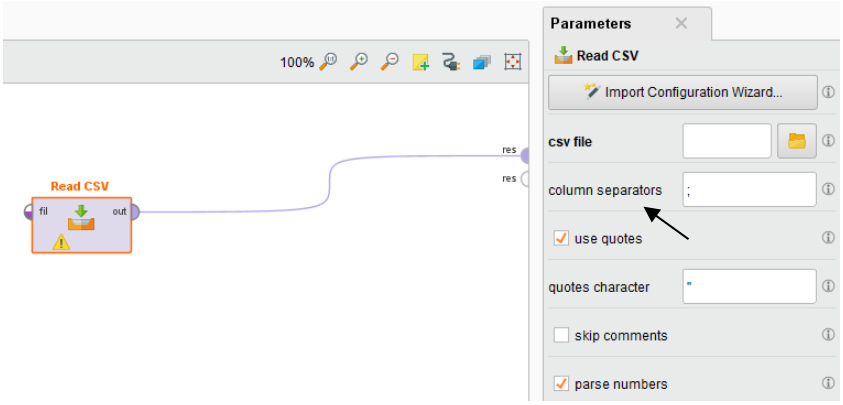


図 3-43. Read CSV オペレータのパラメータ

3)　パラメーターを表示できるように、Read CSV オペレータが選択されていることを確認してください。データファイルの区切り記号が第 3 章のデータセットのようにカンマである場合は、column separators パラメーターをカンマ（,）に変更します。次に csv file ボックスを無視して、代わりに Import Configuration Wizard ボタンをクリックしてください。これにより、データインポートウィザードが開きます。 既存のリポジトリに追加するでも、新しいリポジトリを作成するのでもなく、コンピューター上またはネットワーク上のどこかにある.csv ファイルへアクセスすることが可能です。まず、最初のステップで第 3 章のデータファイルを選択し Next ボタンをクリックします。そして、Column Separator がカンマとなっていることを確認し、データファイルの一行目が列名として正しく認識されていることを確認後に Next ボタンをクリックします。データ型とロールの設定は何も変更せずに Finish ボタンをクリックします。図 3-43 のように、Read CSV オペレータの出力ポート(out)が res ポートへ接続されていることを確認してから、プロセスを実行します。結果画面に表示される内容が、以前に

リポジトリへインポートしたデータとまったく同じであることが確認できるでしょう。

この先の章のチュートリアルでは、データファイルのインポート(import)をするのではなく読み取り(read)をします。

新機能について

2018 年の夏に、RapidMiner は RapidMiner Studio アプリケーションの一部として新機能 TurboPrep をリリースしました。TurboPrep を使用すると、データを見ながら前処理ができます。また、TurboPrep は、データの異常またはエラーを自動的に修正する自動クレンジング機能も備えています。さらに、ピボットテーブルを作成したり、データを可視化することができます。

TurboPrep 機能の詳細は RapidMiner の Web サイトで確認することができます：
https://www.rapidminer.jp/rapidminer-studio/turboprep-automodel/
RapidMiner Studio をダウンロードしてインストールすると、TurboPrep を試すこともできます。RapidMiner Studio のインストール方法は下記ブログで紹介されています。

RapidMiner の始め方～10step でできる簡単インストール方法～
https://www.ksk-anl.com/blog/rapidminer%E3%81%AE%E5%A7%8B%E3%82%81%E6%96%B9%EF%BD%9E10step%E3%81%A7%E3%81%A7%E3%81%8D%E3%82%8B%E7%B0%A1%E5%8D%98%E3%82%A4%E3%83%B3%E3%82%B9%E3%83%88%E3%83%BC%E3%83%AB%E6%96%B9%E6%B3%95%EF%BD%9E/

※本チュートリアルは　2020 年 04 月 28 日　時点のものです。