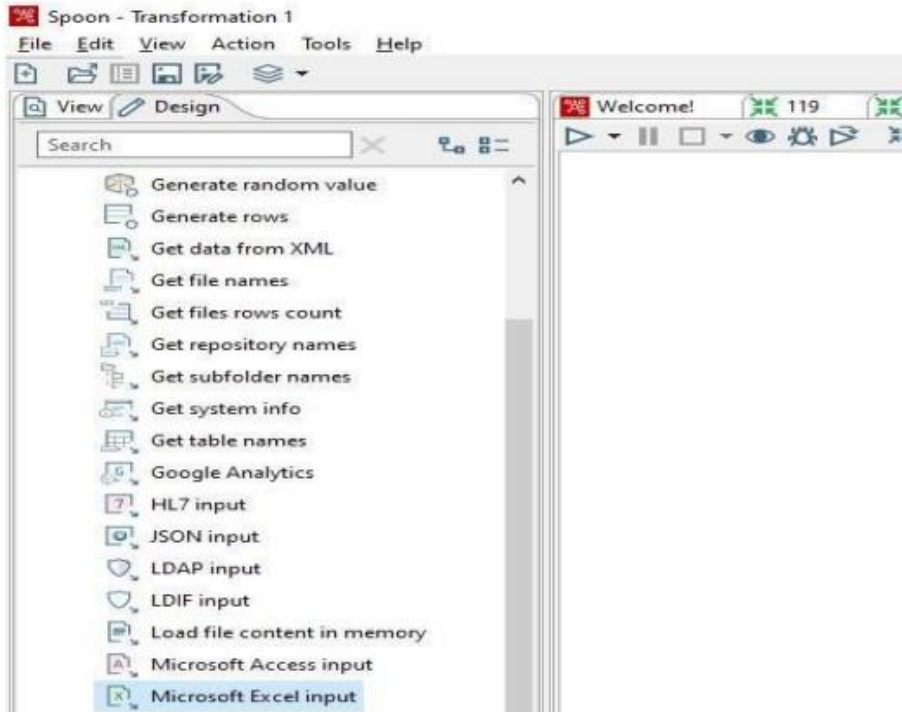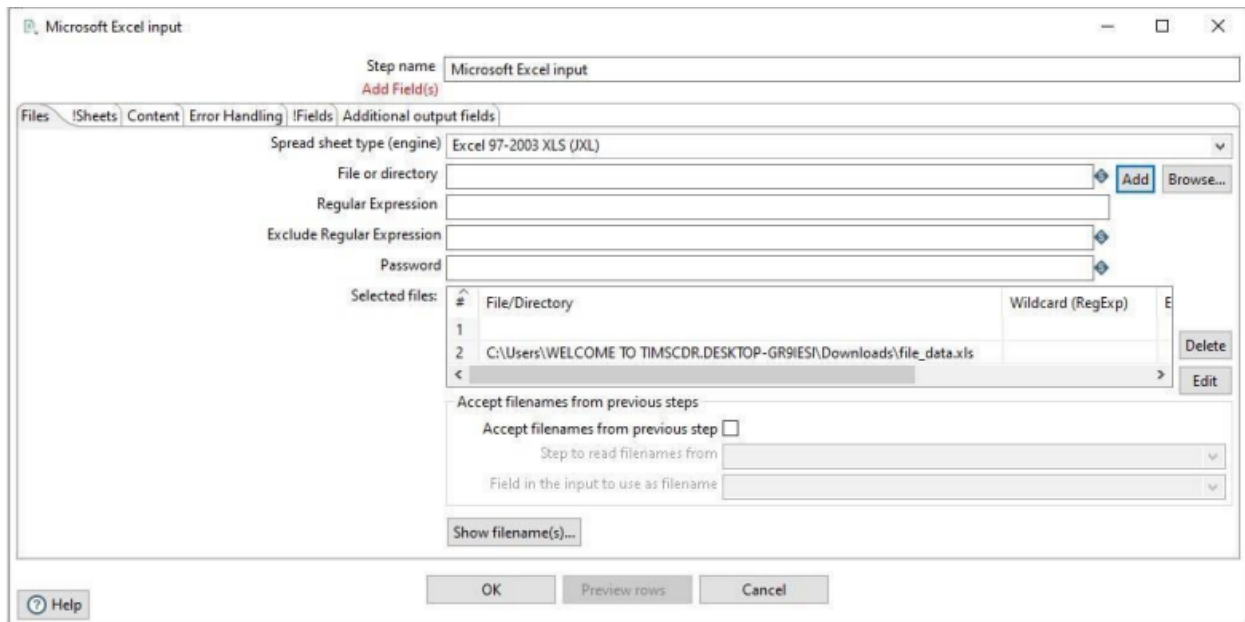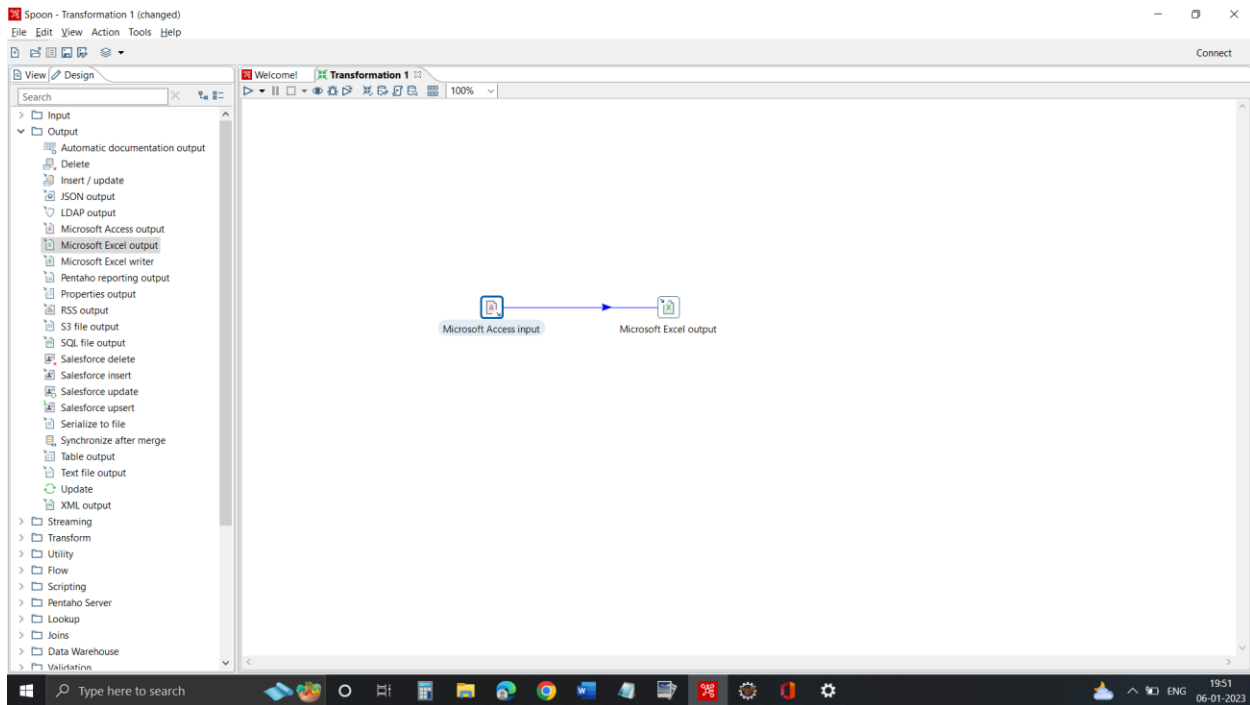# Pentaho Assignment

Pentaho is business intelligence (BI) software that provides data integration, OLAP services, reporting, information dashboards, data mining and extract, transform, load (ETL) capabilities. Pentaho Reporting is a suite (collection of tools) for creating relational and analytical reporting. Using Pentaho, we can transform complex data into meaningful reports and draw information out of them. Pentaho supports creating reports in various formats such as HTML, Excel, PDF, Text, CSV, and xml.

Pentaho can accept data from different data sources including SQL databases, OLAP data sources, and even the Pentaho Data Integration ETL tool.

Assignment on input/output:

1. Retrieving and load Data from a dbms file (MS Access) to data warehouse

2. Retrieving and load Data from a Flat File (Excel) to data warehouse

3. Retrieving and load Data from a RDBMS (MS SQL Server / ORACLE) to data warehouse (If you have oracle installed at home)

Spoon - Transformation 1 (changed)

File Edit View Action Tools Help

Connect

View Design

Search

- Input
- Output
  - Automatic documentation output
  - Delete
  - Insert / update
  - JSON output
  - LDAP output
  - Microsoft Access output
  - Microsoft Excel output
  - Microsoft Excel writer
  - Pentaho reporting output
  - Properties output
  - RSS output
  - S3 file output
  - SQL file output
  - Salesforce delete
  - Salesforce insert
  - Salesforce update
  - Salesforce upsert
  - Serialize to file
  - Synchronize after merge
  - Table output
  - Text file output
  - Update
  - XML output
- Streaming
- Transform
- Utility
- Flow
- Scripting
- Pentaho Server
- Lookup
- Joins
- Data Warehouse
- Validation

Welcome! Transformation 1

100%

Microsoft Access input → Microsoft Excel output

Microsoft Excel input

Step name: Microsoft Excel input

Add Field(s)

Files !Sheets Content Error Handling !Fields Additional output fields

Spread sheet type (engine): Excel 97-2003 XLS (JXL)

File or directory: [   ] Add Browse...

Regular Expression: [   ]

Exclude Regular Expression: [   ]

Password: [   ]

Selected files:

#	File/Directory	Wildcard (RegExp)	E
1			
2	C:\Users\WELCOME TO TIMSCDR.DESKTOP-GR9IESI\Downloads\file_data.xls		

Delete
Edit

Accept filenames from previous steps

Accept filenames from previous step ☐

Step to read filenames from [   ]

Field in the input to use as filename [   ]

Show filename(s)...

Help   OK   Preview rows   Cancel

Microsoft Excel input

Step name: Microsoft Excel input

Files | Sheets | Content | Error Handling | Fields | Additional output fields

#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	Teacher ID	String			none	N				
2	First Name	String			none	N				
3	Last Name	String			none	N				
4	Address	String			none	N				
5	City	String			none	N				
6	State	String			none	N				
7	Zip	String			none	N				
8	Phone	String			none	N				
9	Age	Number			none	N				
10										

Get fields from header row...

OK | Preview rows | Cancel

Help

Assignment on transformation:

1. Retrieve data from file staff.accdb and concat two columns firstname and lastname as fullname of teacher table and load data to data warehouse

Concat fields

Step name: Concat fields
Target Field Name: fullname
Length of Target Field: 0
Separator: ;
Enclosure: "

Fields | Advanced

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim Type	Null
1	Teacher ID	String								
2	First Name	String								
3	Last Name	String								

Get Fields | Minimal width

OK | Cancel

Help

Microsoft Excel input → Concat fields

2. Sort column fullname in descending order and load data to data warehouse.

Sort rows dialog:

Field	Value
Step name	Sort rows
Sort directory	%%java.io.tmpdir%%
TMP-file prefix	out
Sort size (rows in memory)	1000000
Free memory threshold (in %)	
Compress TMP Files?	☐
Only pass unique rows? (verifies keys only)	☐

Fields :

#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
1	Last Name	Y	N	N	0	N

Transformation flow: Microsoft Excel input → Concat fields → Microsoft Excel output → Sort rows

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

◉ First rows ○ Last rows ○ Off

#	Teacher ID	First Name	Last Name	Address	City	State	Zip	Phone	Age	fullname
1	42655	Edward	Alter	1513 Fourth Ave W	Seattle	WA	98119	206-555-0126	5.0	42655;Edward;Alter
2	53196	Jennifer	Buckingham	3612 32nd Ave W	Seattle	WA	98199	206-555-7455	25.0	53196;Jennifer;Buckingham
3	41233	Lily	Edwards	13010 Corliss Ave S	Seattle	WA	98133	206-555-3633	18.0	41233;Lily;Edwards
4	82774	Maddie	Hayes	3813 48th Ave NE	Seattle	WA	98125	206-555-0123	18.0	82774;Maddie;Hayes
5	11599	Shelley	Nelson	701 17th Ave S	Seattle	WA	98122	206-555-8761	5.0	11599;Shelley;Nelson
6	12910	Marcella	Riley	13545 27th Ave SE	Seattle	WA	98125	206-555-3650	15.0	12910;Marcella;Riley
7	52786	Johnny	Roth	912 E Pike St	Seattle	WA	98122	206-555-9410	15.0	52786;Johnny;Roth
8	21606	Ollie	Sherman	1714 E Pike Ave	Seattle	WA	98122	206-555-8301	25.0	21606;Ollie;Sherman
9	30840	Marty	Smith	3826 NW 100th St	Seattle	WA	98125	206-555-7316	11.0	30840;Marty;Smith
10	61689	J.M.	Smith	3826 NW 100th St	Seattle	WA	98125	206-555-7316	11.0	61689;J.M.;Smith

3. Retrieve data from file and descritize age column of class table with <5, 5-10,10-15,15-20, 20-25, >25 and load data to data warehouse

Number range

Step name:	Number range
Input field:	Age
Output field:	Age_Range
Default value(if no range matches):	unknown

Ranges (min <= x< max):

#	Lower Bound	Upper Bound	Value
1	5.0	10.0	Young
2	11.0	20.0	Middle
3	21.0	30.0	Adult

OK   Cancel

? Help

Microsoft Excel input → Concat fields → Microsoft Excel output → Sort rows → Microsoft Excel output 2 → Number range

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows ○ Last rows ○ Off

#	Teacher ID	First Name	Last Name	Address	City	State	Zip	Phone	Age	fullname	Age_Range
1	42655	Edward	Alter	1513 Fourth Ave W	Seattle	WA	98119	206-555-0126	5.0	42655;Edward;Alter	Young
2	53196	Jennifer	Buckingham	3612 32nd Ave W	Seattle	WA	98199	206-555-7455	25.0	53196;Jennifer;Buckingham	Adult
3	41233	Lily	Edwards	13010 Corliss Ave S	Seattle	WA	98133	206-555-3633	18.0	41233;Lily;Edwards	Middle
4	82774	Maddie	Hayes	3813 48th Ave NE	Seattle	WA	98125	206-555-0123	18.0	82774;Maddie;Hayes	Middle
5	11599	Shelley	Nelson	701 17th Ave S	Seattle	WA	98122	206-555-8761	5.0	11599;Shelley;Nelson	Young
6	12910	Marcella	Riley	13545 27th Ave SE	Seattle	WA	98125	206-555-3650	15.0	12910;Marcella;Riley	Middle
7	52786	Johnny	Roth	912 E Pike St	Seattle	WA	98122	206-555-9410	15.0	52786;Johnny;Roth	Middle
8	21606	Ollie	Sherman	1714 E Pike Ave	Seattle	WA	98122	206-555-8301	25.0	21606;Ollie;Sherman	Adult
9	30840	Marty	Smith	3826 NW 100th St	Seattle	WA	98125	206-555-7316	11.0	30840;Marty;Smith	Middle
10	61689	J.M.	Smith	3826 NW 100th St	Seattle	WA	98125	206-555-7316	11.0	61689;J.M.;Smith	Middle

4. Retrieve data from file and add sequence in the staff table.

Add sequence — □ ×

Step name	Add sequence
Name of value	Seq1

Use a database to generate the sequence
- Use DB to get sequence? ☐
- Connection [ ] Edit... New... Wizard...
- Schema name [ ] Schemas...
- Sequence name SEQ_ Sequences...

Use a transformation counter to generate the sequence
- Use counter to calculate sequence? ☑
- Counter name (optional) [ ]
- Start at value 1
- Increment by 1
- Maximum value 999999999

[? Help]   [OK]   [Cancel]

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data
◉ First rows ○ Last rows ○ Off

#	Teacher ID	First Name	Last Name	Address	City	State	Zip	Phone	Age	fullname	Age_Range	Seq1
1	42655	Edward	Alter	1513 Fourth Ave W	Seattle	WA	98119	206-555-0126	5.0	42655;Edward;Alter	Young	1
2	53196	Jennifer	Buckingham	3612 32nd Ave W	Seattle	WA	98199	206-555-7455	25.0	53196;Jennifer;Buckingham	Adult	2
3	41233	Lily	Edwards	13010 Corliss Ave S	Seattle	WA	98133	206-555-3633	18.0	41233;Lily;Edwards	Middle	3
4	82774	Maddie	Hayes	3813 48th Ave NE	Seattle	WA	98125	206-555-0123	18.0	82774;Maddie;Hayes	Middle	4
5	11599	Shelley	Nelson	701 17th Ave S	Seattle	WA	98122	206-555-8761	5.0	11599;Shelley;Nelson	Young	5
6	12910	Marcella	Riley	13545 27th Ave SE	Seattle	WA	98125	206-555-3650	15.0	12910;Marcella;Riley	Middle	6
7	52786	Johnny	Roth	912 E Pike St	Seattle	WA	98122	206-555-9410	15.0	52786;Johnny;Roth	Middle	7
8	21606	Ollie	Sherman	1714 E Pike Ave	Seattle	WA	98122	206-555-8301	25.0	21606;Ollie;Sherman	Adult	8
9	30840	Marty	Smith	3826 NW 100th St	Seattle	WA	98125	206-555-7316	11.0	30840;Marty;Smith	Middle	9
10	61689	J.M.	Smith	3826 NW 100th St	Seattle	WA	98125	206-555-7316	11.0	61689;J.M.;Smith	Middle	10

5. Split fullname column into two columns Fname and Lname using staff table

Split fields — □ ×

Step name	Split fields
Field to split	fullname
Delimiter	.
Enclosure	

Fields

#	New field	ID	Remove ID?	Type	Length	Precision	Format	Group	Decimal	Currency	Nullif	Default	Trim type
1	Fname		N	String	15								
2	Lname		N	String	15								

[? Help]   [OK]   [Cancel]

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data
◉ First rows ○ Last rows ○ Off

#	Teacher ID	First Name	Last Name	Address	City	State	Zip	Phone	Age	Fname	Lname
1	12910	Marcella	Riley	13545 27th Ave SE	Seattle	WA	98125	206-555-3650	15.0	12910	Marcella
2	30840	Marty	Smith	3826 NW 100th St	Seattle	WA	98125	206-555-7316	11.0	30840	Marty
3	42655	Edward	Alter	1513 Fourth Ave W	Seattle	WA	98119	206-555-0126	5.0	42655	Edward
4	53196	Jennifer	Buckingham	3612 32nd Ave W	Seattle	WA	98199	206-555-7455	25.0	53196	Jennifer
5	82774	Maddie	Hayes	3813 48th Ave NE	Seattle	WA	98125	206-555-0123	18.0	82774	Maddie

6. Retrieve data from file staff.accdb and set value of zip column of staff table in staffid column

#	Field name	Replace by value from field	
1	Zip	Address	

Set field value

Step name : Set field value

Fields:

OK   Get Fields   Cancel

Help

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

First rows Last rows Off

#	Teacher ID	First Name	Last Name	Address	City	State	Zip	Phone	Age	fullname	Age_Range	Seq1
1	41233	Lily	Edwards	13010 Corliss Ave S	Seattle	WA	13010 Corliss Ave S	206-555-3633	18.0	41233;Lily;Edwards	Middle	1
2	11599	Shelley	Nelson	701 17th Ave S	Seattle	WA	701 17th Ave S	206-555-8761	5.0	11599;Shelley;Nelson	Young	2
3	52786	Johnny	Roth	912 E Pike St	Seattle	WA	912 E Pike St	206-555-9410	15.0	52786;Johnny;Roth	Middle	3
4	21606	Ollie	Sherman	1714 E Pike Ave	Seattle	WA	1714 E Pike Ave	206-555-8301	25.0	21606;Ollie;Sherman	Adult	4
5	61689	J.M.	Smith	3826 NW 100th St	Seattle	WA	3826 NW 100th St	206-555-7316	11.0	61689;J.M.;Smith	Middle	5

7. Load data transformed in step 1-6 in another MS – Access database.
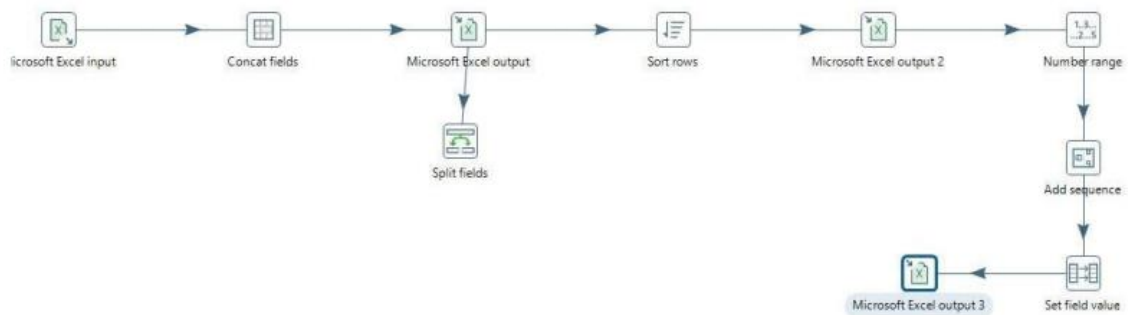
8. Retrieve data from file staff.accdb and apply following string operation

    a. Convert city column in upper case

    b. Convert state column in lower case

    c. Apply left padding in city and fix width 20 and padding character should be "*"

    d. Apply left padding in state and fix width 20 and padding character should be "#"

String operations — □ ✕

Step name | String operations

The fields to process:

#	In stream field	Out stream field	Trim type	Lower/Upper	Padding	Pad char	Pad Length	InitCap	Escape	Digits	Remove Special character
1	City		none	upper	left	*	20	N	None	none	none
2	State		none	lower	left	#	20	N	None	none	none
3											
4											

⑦ Help   OK   Get fields   Cancel

Execution Results

Logging ⊘ Execution History Step Metrics Performance Graph Metrics ⊙ Preview data

⦿ First rows ○ Last rows ○ Off

#	Teacher ID	First Name	Last Name	Address	City	State	Zip	Phone
1	41233	Lily	Edwards	13010 Corliss Ave S	************SEATTLE	################wa	13010 Corliss Ave S	206-555-3633
2	11599	Shelley	Nelson	701 17th Ave S	************SEATTLE	################wa	701 17th Ave S	206-555-8761
3	52786	Johnny	Roth	912 E Pike St	************SEATTLE	################wa	912 E Pike St	206-555-9410
4	21606	Ollie	Sherman	1714 E Pike Ave	************SEATTLE	################wa	1714 E Pike Ave	206-555-8301
5	61689	J.M.	Smith	3826 NW 100th St	************SEATTLE	################wa	3826 NW 100th St	206-555-7316

9. Retrieve data from file staff.accdb and calculate total salary using which is sum of salary and bonus column

Calculator — □ ×

Step name
Calculator

☑ Throw an error on non existing files

Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove	Conversion mask	Decimal symbol	Grouping sym
1	addvalue	A + B	Age	Seq1		Number			N			

? Help   OK   Cancel

	Zip	Phone	Age	fullname	Age_Range	Seq1	addvalue
a	13010 Corliss Ave S	206-555-3633	18.0	41233;Lily;Edwards	Middle	1	19.0
a	701 17th Ave S	206-555-8761	5.0	11599;Shelley;Nelson	Young	2	7.0
a	912 E Pike St	206-555-9410	15.0	52786;Johnny;Roth	Middle	3	18.0
a	1714 E Pike Ave	206-555-8301	25.0	21606;Ollie;Sherman	Adult	4	29.0
a	3826 NW 100th St	206-555-7316	11.0	61689;J.M.;Smith	Middle	5	16.0

10. Retrieve data from file staff.accdb and validate Age field where age should be greater or equal to 18

Number range — □ ×

Step name: Number range 2
Input field: Age
Output field: range
Default value(if no range matches): unknown

Ranges (min <= x< max):

#	Lower Bound	Upper Bound	Value
1	18.0	18.0	Equal to 18
2	19.0		Greater than 18
3	0.0	17.0	Less than 18

OK   Cancel

? Help

Zip	Phone	Age	fullname	Age_Range	Seq1	addvalue	range
13010 Corliss Ave S	206-555-3633	18.0	41233;Lily;Edwards	Middle	1	19.0	unknown
701 17th Ave S	206-555-8761	5.0	11599;Shelley;Nelson	Young	2	7.0	Less than 18
912 E Pike St	206-555-9410	15.0	52786;Johnny;Roth	Middle	3	18.0	Less than 18
1714 E Pike Ave	206-555-8301	25.0	21606;Ollie;Sherman	Adult	4	29.0	Greater than 18
3826 NW 100th St	206-555-7316	11.0	61689;J.M.;Smith	Middle	5	16.0	Less than 18