

Data Science Unsupervised Learning Demonstration

Segmentation Using Consumer Expenditure Data

Spectrum™ Technology Platform
July 2018



Table of Contents

Introduction	4
Scenario	5
Process Overview	5
Prepare Data	5
Combine Datasets	5
Explore Income Data	5
Explore Expenditure Data	5
Create Variables	6
Create Models and Categorize Consumers.....	6
Score New Data.....	6
Consumer Expenditure Survey Data	6
Data Download	7
Prepare Data	7
Import the 2015 Q1 Dataset	7
Import the Remaining Datasets	8
Combine Datasets	8
Explore Income Data.....	9
Sort Data by CUID.....	9
Export Data to a Comma-Delimited Text File.....	10
Inspect Data	10
Explore Expenditure Data	11
Create Variables	11
Set Past Quarter Expenditure to 0	11
Set current quarter expenditure to 0	12
Calculate Group Statistics	12
Calculate yearly expenditure	13
Save Cleansed Data.....	14
Create Training and Test Data Files	15
Add a Transformer Stage.....	15
Add a Conditional Router Stage.....	15
Add a Write to File Stage for Training	16
Add a Write to File Stage for Testing.....	16

Create Models and Segment Consumers.....	17
K-Means Clustering Modeling.....	17
Principle Component Analysis (PCA) Modeling	19
Score New Data Using Java Model Scoring.....	22
Subflow Substitution	24

Introduction

The purpose of this workbook is to demonstrate the functionality of the Spectrum™ Technology Platform Data Science Solution in Enterprise Designer. More specifically, it demonstrates unsupervised learning using the Machine Learning Module and how to score new data using the Analytics Scoring Module. It works in conjunction with the provided dataflows. You can create your own dataflows by following the step-by-step instructions in this workbook, or you can use the included dataflows as a reference to confirm what the individual completed stages and output should look like. If you prefer the latter option and are running Spectrum™ Technology Platform 12.2, use the following dataflows:

- **ConsumerExpenditure_v122_sampleandcluster.df**— creates K-Means Clustering and Principal Component Analysis models
- **ConsumerExpenditure_v122_Subflow.df**—performs part of the data preparation task by creating a reusable subflow containing the Group Statistics and Transformer stages
- **ConsumerExpenditure_v122_sampleandcluster_subflow.df**—identical to ConsumerExpenditure_v122_sampleandcluster.df but contains subflow in place of Group Statistics and Transformer stages
- **ConsumerExpenditure_v122_score.df**—scores the K-Means Clustering model using newer data

If you are running Spectrum™ Technology Platform 12.1, use the similarly named dataflows but with “v121” instead of “v122”. If you are creating your own dataflows while following these instructions, you will first create a job in the Prepare Data section and will continue in that job for the duration.

Also included in the zip file is a folder called “PythonBased,” which contains dataflows, a process flow, a batch script, and a Python script. Used together, these files achieve the same result as the model creation dataflows and subflow listed above; in this case, the Python script conducts the transformations. Note that any third-party program can be utilized within Spectrum™ Technology Platform using the techniques in this example.

Stages covered in this workbook include the following; click a stage to access its webhelp for more information:

- Sources: [Read from File](#)
- Sinks: [Write to File](#)
- Control Stages: [Transformer](#), [Stream Combiner](#), [Sorter](#), [Broadcaster](#), [Group Statistics](#), and [Conditional Router](#)
- Inputs and Outputs: [Input](#) and [Output](#)
- Analytics Scoring: [Java Model Scoring](#)
- Machine Learning: [K-Means Clustering](#) and [Principal Component Analysis](#)

If you refer to the provided dataflows, you may notice that some stages have labels that are different from their original (e.g., “Read from File” is now “Read from fmli151x”). Renaming the stages has no effect on the dataflow’s processing or output so feel free to change stage labels or leave them as-is.

Scenario

You are a data analyst at a manufacturing company and have been asked to prepare a yearly summary dataset using Consumer Expenditure Survey quarterly data published by the Bureau of Labor Statistics. You have also been asked to segment consumers based on their income and expenditure patterns using your created file. The result will be used for future product marketing.

Process Overview

We complete the following steps in this workbook:

Prepare Data

Data preparation is the construction of datasets for the purpose of carrying out customer analytics. In practice, data preparation may be undertaken in different ways, depending on the nature and location of stored data and the facilities available to build customer datasets.

Combine Datasets

Working with data in this manner often requires multiple analyses that use different raw data. A repeat analysis might be based on a different time period or a different group of customers. In such cases, the structure of the dataset required for the analysis might be exactly the same as the original, but it may need to be updated or “refreshed” with new raw data. For example, a dataset built using June 2016 data may require updating with July 2016 data when it becomes available. The amount of effort required to carry out an update will vary depending on the processes used.

When working with data from multiple sources or with multiple vintages, it is sometimes necessary to combine the files. In our job, we use a Stream Combiner to bring together all the data and then sort it before performing additional processing.

Explore Income Data

In this section we calculate annual income for each Consumer Unit Identification Number (CUID). First we sort the data by CUID, then export it to a comma-delimited file, and then we analyze the data using inspection points in Enterprise Designer.

Explore Expenditure Data

In the Interview Survey, each family in the sample is interviewed every quarter over the course of a year. The sample for each quarter is divided into three monthly panels, with CUs being interviewed in the same panel of every quarter. Expenditure information that is based on three months of respondents' recall is also collected during this interview. A list of each of the expenditure items we are interested is provided in this demonstration.

Create Variables

We observe that CU income numbers vary by interview; therefore, we will select the highest number among them as the annual income number for that CU. We will then summarize annual expenditures and determine the yearly expenditures for each CUID by performing a variety of calculations.

Create Models and Categorize Consumers

In this demonstration we will create a K-Means Clustering model and a Principal Component Analysis (PCA) model. The K-Means Clustering algorithm creates models based on analytical clustering, which segments a set of records into clusters of similar records based on data values. Principal Component Analysis is a statistical process that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables known as principal components.

Score New Data

We will use the models created in the original dataflow to categorize people into like groups based on a variety of parameters, such as expenditure behavior, house size, number of earners in household, and so on. We accomplish this in a dataflow that uses Java Model Scoring`.

Consumer Expenditure Survey Data

The Consumer Expenditure Survey provides continuous and comprehensive information on the buying habits of U.S. consumers. The spending shares for items in the survey are used as weights for the Consumer Price Index.

Consumer Expenditure (CE) collects data through two surveys: the Quarterly Interview Survey (Interview) and the Diary Survey (Diary). The Interview generally tracks consumer units' (CU) large expenditures, such as major appliances and cars, while the Diary tracks smaller, everyday expenditures that might be easily forgotten after a few days, such as a cup of coffee. The Interview is conducted quarterly with each consumer unit and the Diary is conducted over two consecutive one-week periods with each respondent.

An Interview "quarter" refers to the calendar quarter in which the interview occurred. For example, any Consumer Unit (CU) interviewed in April, May, or June would have their data stored in the quarter 2 (YYQ2) datasets. During an interview, the CU is asked to report expenditures for a reference period of three months. So, for a CU interviewed in April, their expenditures in the YYQ2 file are for January, February, and March. (This is important to remember when calculating calendar year estimates.) Five quarters of interviews are included in each public-use microdata (PUMD) release, YYQ1 of <release year> through YYQ1 of <release year + 1>.

The Census Bureau selects a sample of approximately 12,000 addresses per year. The Interview Survey is a rotating panel survey in which approximately 12,000 addresses are contacted each calendar quarter of the year for the survey. One-fourth of the addresses that are contacted each quarter are new to the survey. Usable interviews are obtained from approximately 6,900 households at those addresses each quarter of the year.

Data Download

Note: The files discussed below are provided in the **C:\ConsumerExpenditureData\data.zip** folder of the zip file you downloaded for this demonstration. The **data.zip\input** folder contains subfolders for 2015 ("intrvw15") and 2016 ("intrvw16"). For your reference and use in the scoring dataflow, we have also provided output files that are generated by this demonstration's dataflows in the **data.zip\output** folder.

The Consumer Expenditure Survey (CE) public use data files and documentation (file structure, data dictionary, sample code, etc.), are available here:

<http://www.bls.gov/cex/pumd.htm>

For this demonstration, we are interested in the contents of the 2015 and 2016 Interview zip files in CSV format. These files contain quarterly family data FMLI files, which are files with characteristics, income, weights, and summary-level expenditures for the CU.

https://www.bls.gov/cex/pumd_data.htm

Prepare Data

In these instructions, we assume that the Spectrum™ Technology Platform server, the Data Science bundle, and Enterprise Designer are installed on your local machine. Additionally, we assume that you either are using the aforementioned Consumer Expenditure data provided with this demonstration or have downloaded the data from the links above.

Import the 2015 Q1 Dataset

The starting point for our work is the 2015 Quarter 1 dataset, a comma-separated values (CSV) file. After this, we will import other quarters' data.

1. Open Enterprise Designer.
2. Select File > New > Dataflow > Job.
3. Drag a Read from File stage onto the canvas and double-click that stage.
4. In the **File name** field, locate and specify the file *fml151x.csv*.
5. In the **Record type** field, choose Delimited.
6. In the **Field separator** field, select Comma (,).
7. In the **Text qualifier** field, select Double quote (").
8. In the **Record separator** field, select Unix (U+000A).
9. Check **First row is header record**.
10. Uncheck **Treat records with fewer fields than defined as malformed**.

- Click the Fields tab and drag the bar down as far as it will go. Modify the last two fields to be in all capital letters with no quotation marks and check the **Trim** box.

Read from fmli151x Options

File Properties Fields Sort Fields Runtime

Name	Type	Position	<input checked="" type="checkbox"/>	Trim
FINATXE1	string	793	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FINATXE2	string	794	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FINATXE3	string	795	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FINATXE4	string	796	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FINATXE5	string	797	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
TOTXEST	string	798	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
BUSCREEN	string	799	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
BUSC_EEN	string	800	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
INT_HOME	string	801	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
INT_PHON	string	802	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
DIVISION	string	803	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
DESFLG	string	804	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FFTAXOWE	string	805	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FSTAXOWE	string	806	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

- Click OK.
- Change the stage name to “Read from fmli151x.”
- Select Edit > Job options and check **Do not terminate the job on a malformed record** and click OK.
- Select Edit > Type Conversion Options and check **Override system default options with the following values**. Select **Initialize the field using default values**, uncheck all **Fail Null** boxes, and then click OK.
- Save the job.

Import the Remaining Datasets

Repeat steps 2-13 for the remaining 2015 datasets and the 2016 Q1 dataset, changing the stage names as follows:

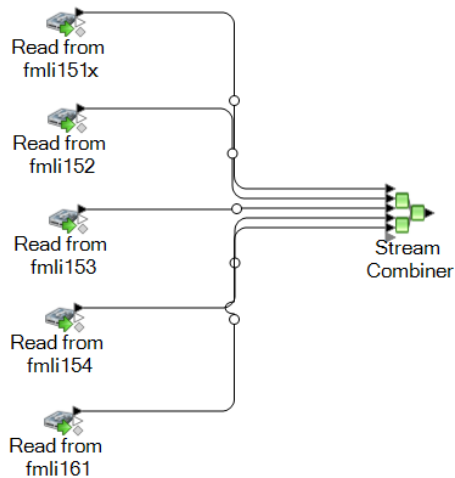
- Read from fmli152
- Read from fmli153
- Read from fmli154
- Read from fmli161

Combine Datasets

You have now imported data from both of your sources. Your next task is to combine these into one dataset.

- Drag a Stream Combiner stage onto the canvas.
- Click the solid black triangle on the right side of the Read from File stage (the output port) and drag it to the gray triangle on the left side of the Stream Combiner stage to create a channel connecting the two stages.
- Repeat step 2 for the remaining datasets.

Your dataflow should look like this:



Explore Income Data

Income was asked at Second or Fifth Interview or New Consumer Units interview. Interviewer asked for the entire CU as a group. Now we want to calculate annual income for each Consumer Unit Identification Number (CUID). These are the variables we are interested in for this purpose:

- **FINATXEM**—Mean of imputed income before taxes minus estimated taxes:
[FINCBTXM - (FFTAXOWE + FSTAXOWE + MISCTAXX)]
- **FINCBTXM**—Amount of pre-tax CU income in previous 12 months:
[FSALARYM, FSMPFRXM, FRRETIRM, FSSIXM, RETSURVM, INTRDVXM, ROYESTXM, NETRENTM, WELFAREM, JFS_AMTM, OTHREGXM, OTHRINCM]
- **INC_RNKM**—Weighted cumulative percent income ranking based on total current income
- **FSMPFRMX**—Family-level summation for new variables SEMPFRMX and SMPFRMBX, where SEMPFRMX represents the amount of self-employment income or loss and SMPFRMBX represents the median value of the bracket range for SMPFRMB (the range that best reflects the income or loss from self-employment during the previous 12 months).

To explore income data, we want to first sort the data and then either output the data to a file or through an inspection point in the dataflow.

Sort Data by CUID

1. Drag a Sorter stage onto the canvas and double-click that stage.
2. Click Add and in the **Field Name** column select CUID. Then click OK.


Export Data to a Comma-Delimited Text File

Now that we have sorted the data, the next step is to export the data.

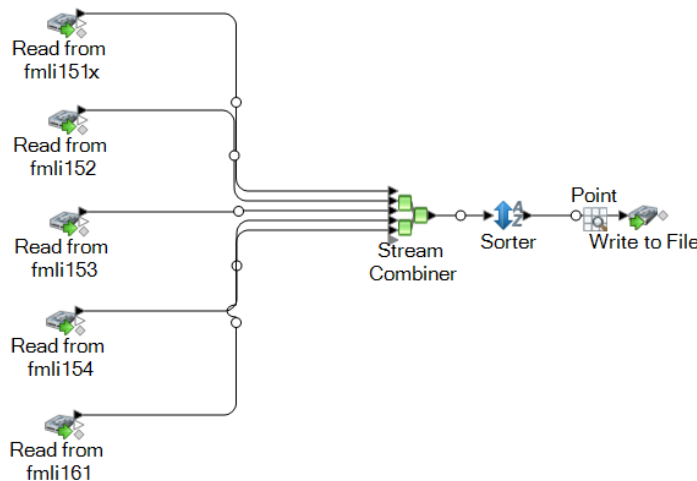
1. Drag a Write to File stage onto the canvas.
2. Connect the Sorter stage and the Write to File stage.
3. Double-click the Write to File stage.
4. In the **File name** field, save the data as "Combined.csv".
5. In the **Record type** field, choose Delimited.
6. In the **Field separator** field, select Comma (,).
7. In the **Text qualifier** field, select Double quote (").
8. In the **Record separator** field, select Unix (U+000A).
9. Check **First row is header record**.
10. Click the Fields tab.
11. Click **Quick Add**, then **Select All**.
12. Click OK twice and save the job.

Inspect Data

We can also explore data by adding an inspection point to the dataflow. The inspection tool can process a maximum of 50 records, which by default is the first 50 records of the input file or database. The Consumer Expenditure data has about 50% missing data, so the first 50 records may not contain much useful information. Therefore, exporting the data to a text file is a more effective way for us to explore this data.

1. Right-click the channel that connects the Sorter stage and the Write to File stage.
2. Select **Add Inspection Point**.
3. Click the Inspect Current Flow button () on the toolbar.

Your dataflow should look like this:



Explore Expenditure Data

In the Interview Survey, each family in the sample is interviewed every quarter over the course of a year. The sample for each quarter is divided into three monthly panels, with CUs being interviewed in the same panel of every quarter. In other words, CUs are always interviewed in the first month of every quarter, the second month of every quarter, or the third month of every quarter. Expenditure information that is based on three months of respondents' recall is also collected during this interview.

These are the expenditure items we are interested in:

- **Expenditures last quarter:** [FOODPQ + ALCBEVPQ + HOUSPQ + APPARPQ + TRANSPQ + HEALTHPQ + ENTERTPQ + PERSCAPQ + READPQ + EDUCAPQ + TOBACCPQ + MISCPQ + CASHCOPQ + PERINSPQ]
- **Expenditures current quarter:** [FOODCQ + ALCBEVCQ + HOUSCQ + APPARCQ + TRANSCQ + HEALTHCQ + ENTERTCQ + PERSCACQ + READCQ + EDUCACQ + TOBACCCQ + MISCCQ + CASHCOCQ + PERINSCQ]

Create Variables

As depicted above, expenditure information for each variable is separated into current quarter and last quarter. To summarize annual expenditures for 2015, we will first set past quarter expenditures for *fmli151x.csv* to 0 because those would've taken place in 2014 (the last quarter prior to Q1 2015). We then set current quarter expenditure for *fmli161.csv* to 0 because those took place in 2016 (current quarter of Q1 2016). We then add the sum of past quarter and current quarter expenditures to determine the yearly expenditures for each CUID.

We observe that CU income numbers vary by interview; therefore, we will use a Transformer stage to select the highest number among them as the annual income number for that CU. The Transformer stage has predefined transforms that perform a variety of common data transformations. If the predefined transforms do not meet our needs, we can write a custom transform script using Groovy. Click [here](#) for more information on creating custom transforms and [here](#) for more information on using Groovy scripts.

Set Past Quarter Expenditure to 0

1. Drag the Transformer stage onto the canvas between the Read from fmli15x stage and the Stream Combiner stage. Open the stage.
2. Click Add. Under the **General** heading, click "Custom".
3. Enter a name for the transform you are creating in the **Custom transform name** field. The name must be unique.

- Click **Script Editor** and then enter the following code in the script editor pane:

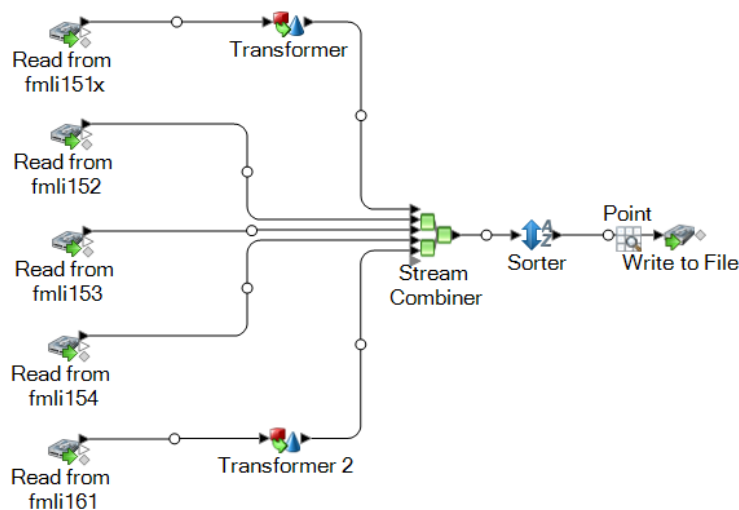
```
data['FOODPQ'] = '0';
data['ALCBEPQ'] = '0';
data['HOUSPQ'] = '0';
data['APPARPQ'] = '0';
data['TRANSPQ'] = '0';
data['HEALTHPQ'] = '0';
data['ENTERTPQ'] = '0';
data['PERSCAPQ'] = '0';
data['READPQ'] = '0';
data['EDUCAPQ'] = '0';
data['TOBACCPQ'] = '0';
data['MISCPQ'] = '0';
data['CASHCOPQ'] = '0';
data['PERINSPQ'] = '0';
```

- Click Close, then Add, then Close, then OK.

Set current quarter expenditure to 0

Repeat steps 1-5 above except this time drag this Transformer stage between the Read from fmli161 stage and the Stream Combiner stage.

Your dataflow should look like this:



Calculate Group Statistics

Note: This step and the next step, “Calculate Yearly Expenditure,” may be replaced with the subflow titled “ConsumerExpenditure_v122_Subflow.df”. See “Subflow Substitution” on page 13 for more information.

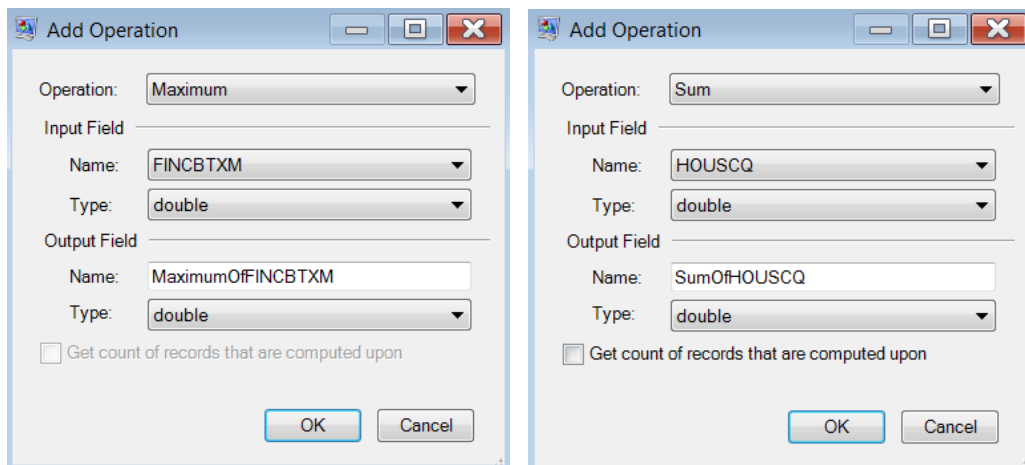
In this dataflow, the Group Statistics stage is used to calculate the maximum salary and yearly expenditure for each Consumer Unit.

1. Insert a Broadcaster stage between the Sorter and Write to file stage.
2. Drag a Group Statistics stage onto the canvas and attach it to the Broadcaster stage.
Open the stage.
3. Select **CUID** as the Rows variable.
4. In the Operations field, assign Maximum and Summation to the following fields, all with a type of double:

Operations:

```
> Assign Maximum of FINCBTXM to MaximumOfFINCBTXM
Assign Maximum of FINATXEM to MaximumOfFINATXEM
Assign Maximum of INC_RNKM to MaximumOfINC_RNKM
Assign Maximum of FSMPFRMX to MaximumOfFSMPFRMX
Assign Sum of FOODPQ to SumOfFOODPQ
Assign Sum of FOODCQ to SumOfFOODCQ
Assign Sum of ALCBEVPQ to SumOfALCBEVPQ
Assign Sum of ALCBEVCQ to SumOfALCBEVCQ
Assign Sum of HOUSPQ to SumOfHOUSPQ
Assign Sum of HOUSCQ to SumOfHOUSCQ
```

Your screen should look similar to one of the following for each field:



Calculate yearly expenditure

1. Drag a Transformer stage onto the canvas and attach it to the Group Statistics stage.
Open the stage.
2. Click Add. Under the **General** heading, click "Custom".
3. Enter a name for the transform you are creating in the **Custom transform name** field.
The name must be unique.

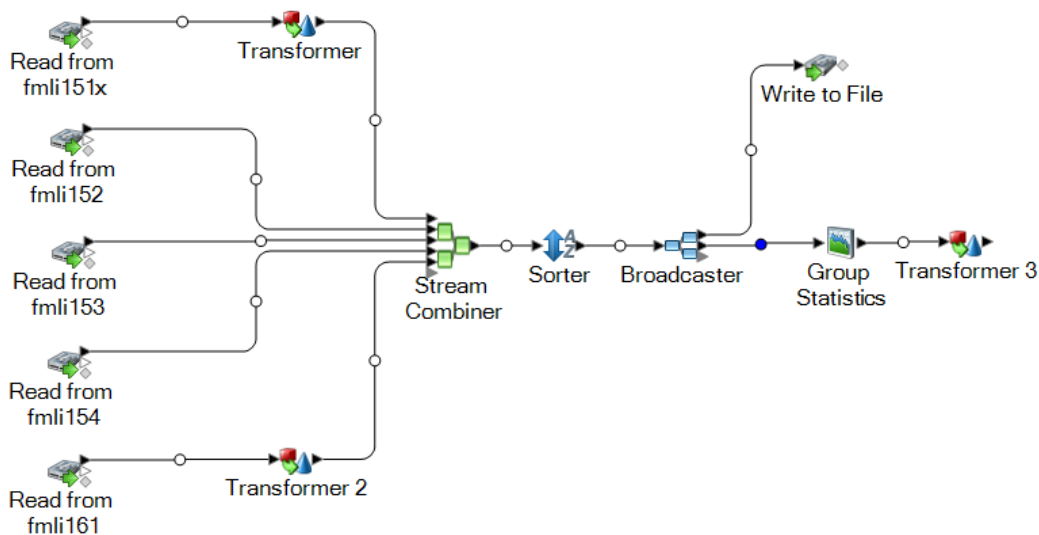
4. Click **Script Editor** and then enter the following code in the script editor pane:

```
data['FOODP_CQ'] = data['SumOfFOODPQ'] + data['SumOfFOODCQ'] ;
data['ALCBEP_CQ'] = data['SumOfALCBEPQ'] + data['SumOfALCBECQ'] ;
data['HOUSP_CQ'] = data['SumOfHOUSPQ'] + data['SumOfHOUSCQ'] ;
data['TRANSP_CQ'] = data['SumOfTRANSPQ'] + data['SumOfTRANSCQ'] ;
data['APPARP_CQ'] = data['SumOfAPPARPQ'] + data['SumOfAPPARCQ'] ;
data['HEALTHP_CQ'] = data['SumOfHEALTHPQ'] + data['SumOfHEALTHCQ'] ;
data['ENTERTP_CQ'] = data['SumOfENTERTPQ'] + data['SumOfENTERTCQ'] ;
data['PERSCAP_CQ'] = data['SumOfPERSCAPQ'] + data['SumOfPERSCACQ'] ;
data['READP_CQ'] = data['SumOfREADPQ'] + data['SumOfREADCQ'] ;
data['EDUCAP_CQ'] = data['SumOfEDUCAPQ'] + data['SumOfEDUCACQ'] ;
data['TOBACCP_CQ'] = data['SumOfTOBACCPQ'] + data['SumOfTOBACCCQ'] ;
data['MISCP_CQ'] = data['SumOfMISCPQ'] + data['SumOfMISCCQ'] ;
data['CASHCOP_CQ'] = data['SumOfCASHCOPQ'] + data['SumOfCASHCOCQ'] ;
data['PERINSP_CQ'] = data['SumOfPERINSPQ'] + data['SumOfPERINSCQ'] ;
```

Note that you can use the Python script included in this zip file to complete the same tasks performed by the Group Statistics and Transformer stages just added to the dataflow.

5. Click Close, then Add, then Close, then OK.

Your dataflow should look like this:



Save Cleansed Data

The next step is to save all the data in its current state.

1. Drag a Write to File stage onto the canvas.
2. Connect the Transformer 3 stage and the Write to File stage.
3. Double-click the Write to File stage.
4. In the **File name** field, save the data as "Output_Expenditure2015.csv".
5. In the **Record type** field, choose Delimited.
6. In the **Field separator** field, select Comma (,).
7. In the **Text qualifier** field, select Double quote (").

8. In the **Record separator** field, select Unix (U+000A).
9. Check **First row is header record**.
10. Click the Fields tab.
11. Click **Quick Add**, then **Select All**.
12. Click OK twice and save the job.

Create Training and Test Data Files

No we will divide the data into Training data (95%) and Test data (5%) by sending the output to those two respective Write to File stages.

Add a Transformer Stage

1. Insert a Broadcaster stage between the Transformer 3 and Write to file stages.
2. Drag a Transformer stage onto the canvas and attach it to the open Broadcaster stage. Open the stage.
3. Click Add. Under the **General** heading, click "Custom".
4. Enter a name for the transform you are creating in the **Custom transform name** field. The name must be unique.
5. Click **Script Editor** and then enter the following code in the script editor pane to generate random number indicator for each record:

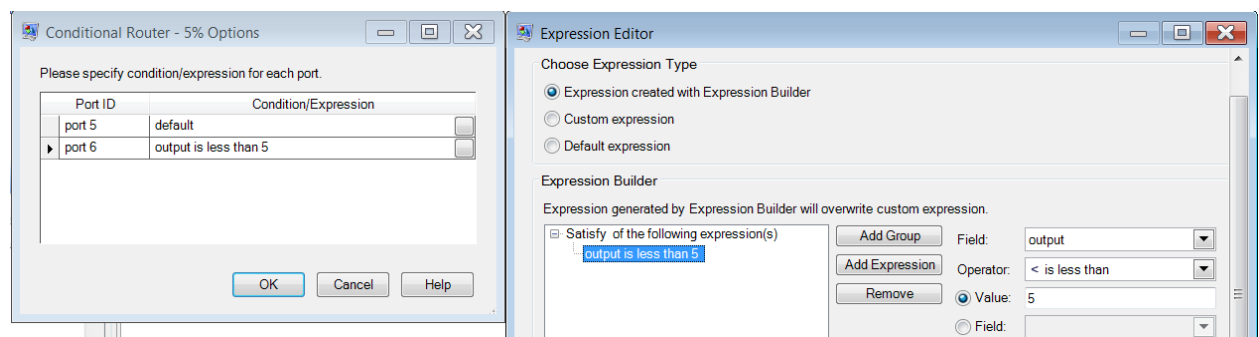
```
//data['RandNum'] =random(100)
import java.util.Random;
import groovy.transform.Field;

@Field Random rand = new Random(147852); //where seed will be specified
by the user.
data['output'] = rand.nextInt(100); //where max will be specified by the
user.
```

6. Click Close, then Add, then Close, then OK.

Add a Conditional Router Stage

1. Drag a Conditional Router stage onto the canvas and connect it to the open Transformer stage.
2. Configure the stage to split the data into Training (95%) and Test (5%):



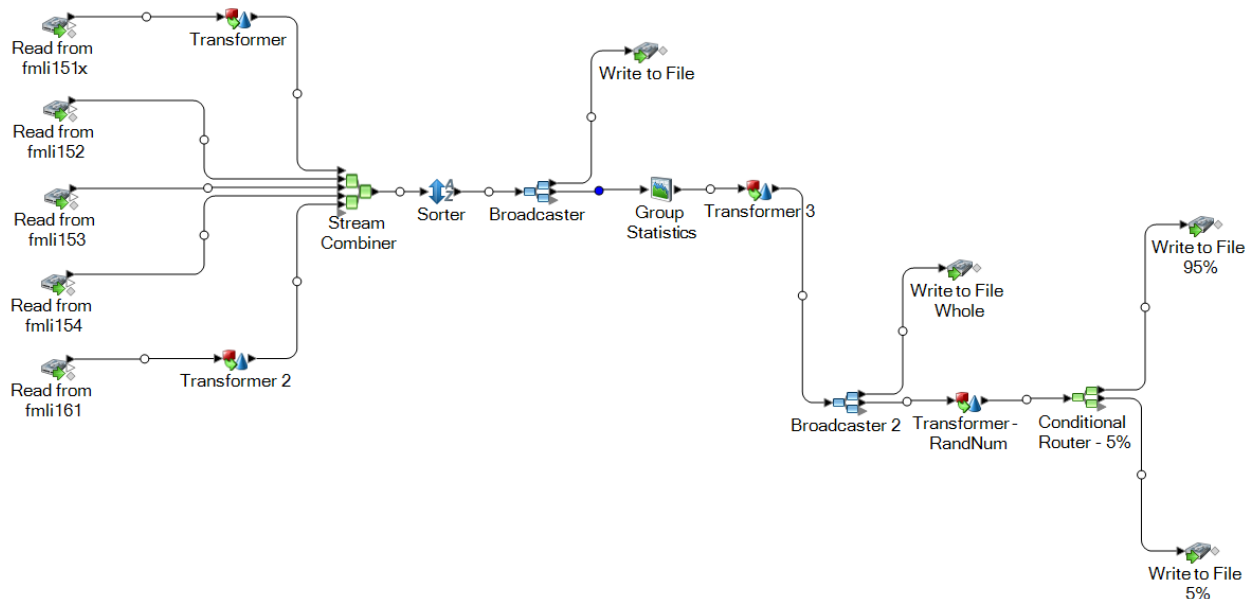
Add a Write to File Stage for Training

1. Drag a Write to File stage onto the canvas and connect it to the Conditional Router.
2. Double-click the Write to File stage.
3. In the **File name** field, save the data as "Output_Expenditure2015_95perc.csv".
4. In the **Record type** field, choose Delimited.
5. In the **Field separator** field, select Comma (,).
6. In the **Text qualifier** field, select Double quote (").
7. In the **Record separator** field, select Unix (U+000A).
8. Check **First row is header record**.
9. Click OK.

Add a Write to File Stage for Testing

1. Drag a Write to File stage onto the canvas and connect it to the Conditional Router.
2. Double-click the Write to File stage.
3. In the **File name** field, save the data as "Output_Expenditure2015_5perc.csv".
4. In the **Record type** field, choose Delimited.
5. In the **Field separator** field, select Comma (,).
6. In the **Text qualifier** field, select Double quote (").
7. In the **Record separator** field, select Unix (U+000A).
8. Check **First row is header record**.
9. Click OK and save the job.

Your dataflow should look like this:



10. Click the Run button (▶) to run your dataflow. The Execution Details window appears and shows the status of the job.
11. Click Refresh. Once the status shows Succeeded, click Close.

Create Models and Segment Consumers

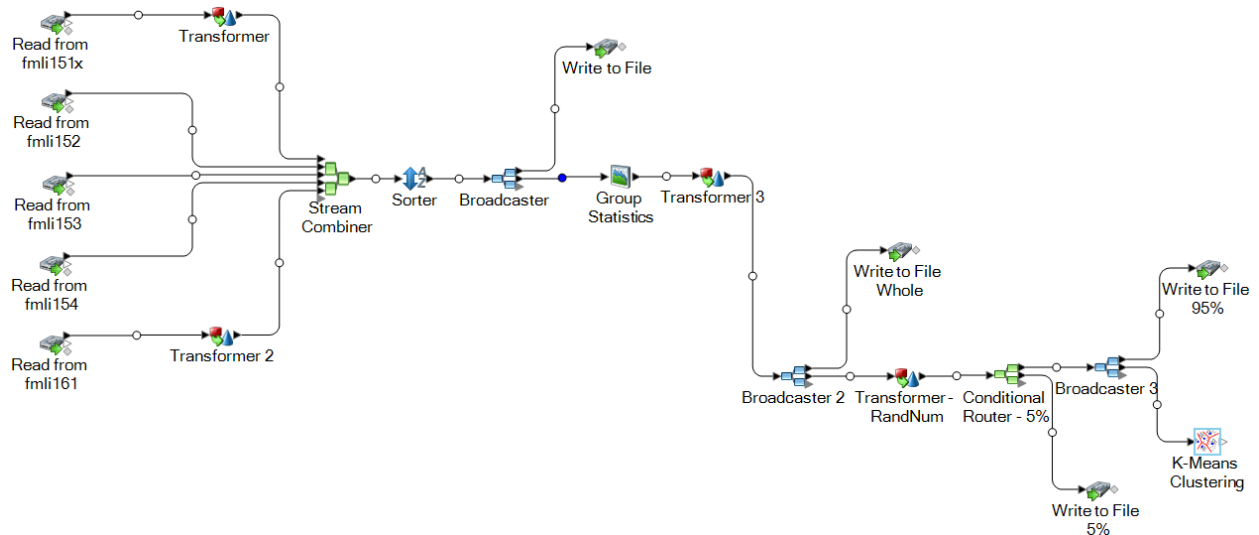
Now we will create K-Means Clustering and Principal Component Analysis (PCA) models to segment consumers.

K-Means Clustering Modeling

1. Drag a Broadcaster stage onto canvas and connect it to the Conditional Router stage.
2. Connect the Write to File Training stage to the Broadcaster stage.
3. Drag a K-Means Clustering stage onto the canvas and connect it to the Broadcaster stage.
4. Double-click the K-Means Clustering stage; the Model Properties tab appears.
5. In the **Model name** field, enter "Consumer Exp KMeans original".
6. Check **Overwrite** to overwrite the existing model.
7. Enter the **Number of clusters** you want in your model if you do not want the default number (5).
8. Enter a **Description** of the model.
9. Click **Include** for each field whose data you want added to the model.
10. Use the **Model Data Type** drop-down to specify whether the input field is to be used as a numeric, categorical, or datetime field.
11. Continue to the Basic Options tab.
 - a. Leave **Standardize input fields** checked to standardize the numeric columns to have zero mean and unit variance.
 - b. Leave **Score input data** unchecked.
 - c. Check **Estimate number of clusters** to have the K-Means algorithm determine the number of clusters that your model will contain. Even though you designate the number of desired clusters on the Model Properties tab, the routine may discover in its processing that a different number of clusters is more appropriate given the data.
 - d. Leave the value 100 as the **Percentage for training data**.
 - e. Leave the value 0 as the **Percentage for test data**.
 - f. Leave **Seed for sampling** checked with the default value or enter a seed number. The data will be split into test and train data and it will occur the same way each time you run the dataflow. Leave "0" in this field to get a random split each time you run the flow.
12. Continue to the Advanced Options tab.
 - a. Leave **Ignore constant fields** checked to skip fields that have the same value for each record.
 - b. Leave **Seed for algorithm** checked with the default value or enter a seed number. The data will be split for cross-validation and it will occur the same way each time you run the dataflow. Leave "0" in this field to get a random split each time you run the flow.
 - c. Leave the **Init** (initialization mode) dropdown set to Random. It will choose K clusters from the set of N observations at random so that each observation has an equal chance of being chosen.
 - d. Check **N fold** and use the default number of folds (5) to perform cross-validation.
 - e. Leave **Fold assignment** at its default setting of Auto. This selection allows the algorithm to automatically choose an option from Random and Modulo; currently it uses Random.
 - f. Leave Fold field unchecked.
 - g. Leave Maximum iterations unchecked.

13. Click OK and save the job.

Your dataflow should look like this:



14. Click the Run button (▶) to run your dataflow. The Execution Details window appears and shows the status of the job.

15. Click Refresh. Once the status shows Succeeded, click Close.

16. Double-click the K-Means Clustering stage, then click the Model Output tab, followed by the **Output** button. A limited version of the resulting model metrics appears:

K-Means Clustering Options

Model Properties Basic Options Advanced Options **Model Output**

Outputs Output


Output Metrics	Training	Test	N Fold
▶ Number of rows	14219		14219
Number of cluste	5		5
Number of categ	0		0
Within cluster su	345072.926852		349551.205159
Total sum of squ	454976.000000		454975.998708
Between cluster	109903.073148		105424.793549

[Model details](#)







17. Click the Model details link; it will lead you to the Machine Learning Model Management page, where you can see all the metrics for the model you just created. You may need to log on to this page using a provided user name and password.

18. Click the **Model Assessment** tab. (Note that if you are using Spectrum™ Technology Platform 12.1, this is called the Model Analysis tab.)



19. Check the box next to Consumer Exp KMeans original, and then click the View Detail () button.

Model Assessment

     						Filter
Model Name	Status	Model Type	User	Dataflow Name	Creation Time	
<input type="checkbox"/> Consumer Exp PCA original	Unexposed	PCA	admin	ConsumerExpenditure_v122_sampleandcluster	22 May 2018 10:29:	
<input checked="" type="checkbox"/> Consumer Exp KMeans original	Unexposed	K-Means clustering	admin	ConsumerExpenditure_v122_sampleandcluster	22 May 2018 10:29:	

The details webpage will appear:

Model Detail

Model Name:	Consumer Exp KMeans original	Status:	Unexposed
Model Type:	K-Means clustering	Dataflow Name:	ConsumerExpenditure_v122_sampleandcluster
User:	admin	Creation Time:	22/05/2018 10:29 PM
Description:			

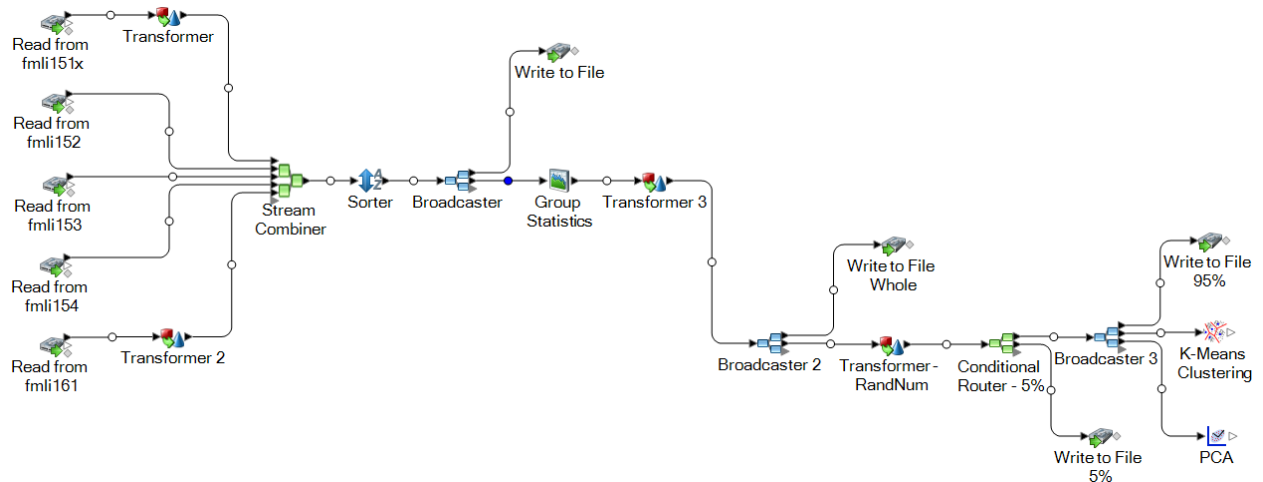
Model Summary	Metrics	Centroid Statistics	Cluster Means	Standardized Cluster Means																
<table><tr><th>Parameter</th><th>Value</th></tr><tr><td>Number of rows</td><td>14,219</td></tr><tr><td>Number of clusters</td><td>5</td></tr><tr><td>Number of categorical columns</td><td>0</td></tr><tr><td>Number of iterations</td><td>42</td></tr><tr><td>Within cluster sum of squares</td><td>345,072.926852</td></tr><tr><td>Total sum of squares</td><td>454,976.000000</td></tr><tr><td>Between cluster sum of squares</td><td>109,903.073148</td></tr></table>					Parameter	Value	Number of rows	14,219	Number of clusters	5	Number of categorical columns	0	Number of iterations	42	Within cluster sum of squares	345,072.926852	Total sum of squares	454,976.000000	Between cluster sum of squares	109,903.073148
Parameter	Value																			
Number of rows	14,219																			
Number of clusters	5																			
Number of categorical columns	0																			
Number of iterations	42																			
Within cluster sum of squares	345,072.926852																			
Total sum of squares	454,976.000000																			
Between cluster sum of squares	109,903.073148																			

Principle Component Analysis (PCA) Modeling

1. Drag a PCA stage onto the canvas and connect it to the Broadcaster stage.
2. Double-click the PCA stage; the Model Properties tab appears.
3. In the **Model name** field, enter "Consumer Exp PCA original".
4. Check **Overwrite** to overwrite the existing model.
5. Enter the **Principal Components** you want in your model if you do not want the default number (1).
6. Enter a **Description** of the model.

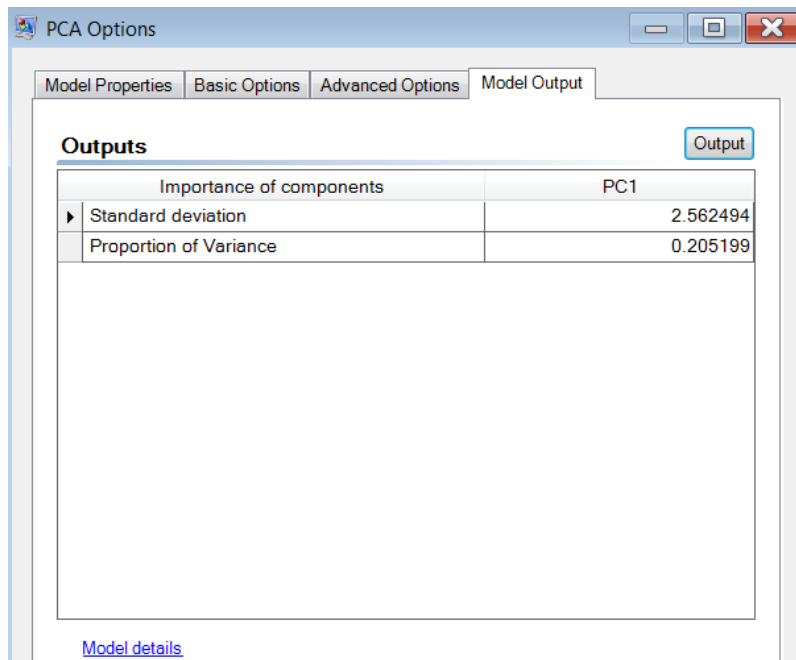
7. Click **Include** for each field whose data you want added to the model.
8. Use the **Model Data Type** drop-down to specify whether the input field is to be used as a numeric, categorical, or datetime field.
9. Continue to the Basic Options tab.
 - a. Leave **Use all factor level** unchecked.
 - b. Leave **Score input data** unchecked.
 - c. Leave the **Transform** field set to Standardize. This parameter specifies the transformation method for the training data. The Standardize method subtracts the mean of each column and divides each column by its range (maximum - minimum).
 - d. Set **Missing data** to Impute means.
10. Continue to the Advanced Options tab.
 - a. Leave **Ignore constant fields** checked.
 - b. Leave **PCA method** set to GramSVD. This algorithm computes the principal components using a distributed computation of the Gram matrix followed by a local SVD using the JAMA package.
 - c. Leave **Maximum iterations** blank.
11. Click OK and save the job.

Your dataflow should look like this:



12. Click the Run button (▶) to run your dataflow. The Execution Details window appears and shows the status of the job.
13. Click Refresh. Once the status shows Succeeded, click Close.

14. Double-click the PCA stage, then click the Model Output tab, followed by the **Output** button. A limited version of the resulting model metrics appears:









15. Click the Model details link; it will lead you to the Machine Learning Model Management page, where you can see all the metrics for the model you just created. You may need to log on to this page using a provided user name and password.
16. Click the **Model Assessment** tab.



17. Check the box next to Consumer Exp PCA original, and then click the View Detail

() button.

Model Assessment

<div>       </div> <div>Filter</div>					
<input type="checkbox"/> Model Name	Status	Model Type	User	Dataflow Name	Creation Time
<input checked="" type="checkbox"/> Consumer Exp PCA original	Unexposed	PCA	admin	ConsumerExpenditure_v122_sampleandcluster	22 May 2018 10:29:
<input type="checkbox"/> Consumer Exp KMeans original	Unexposed	K-Means clustering	admin	ConsumerExpenditure_v122_sampleandcluster	22 May 2018 10:29:

The details webpage will appear:

Model Detail

Model Name:	Consumer Exp PCA original	Status:	Unexposed
Model Type:	PCA	Dataflow Name:	ConsumerExpenditure_v122_sampleandcluster
User:	admin	Creation Time:	22/05/2018 10:29 PM
Description:			

Importance of components		Rotation
		PC1
Standard deviation		2.562494
Proportion of Variance		0.205199
Cumulative Proportion		0.205199

Score New Data Using Java Model Scoring

To score the new data we first need to expose the model(s).

1. Return to the Machine Learning Model Management page and click the Model Assessment tab. (Note that if you are using Spectrum™ Technology Platform 12.1, this is called the Model Analysis tab.)
2. Check the model you intend to expose. You can only expose one model at a time.

Model Assessment

Filter

<input type="checkbox"/>	Model Name	Status	Model Type	User	Dataflow Name	Creation Time
<input checked="" type="checkbox"/>	Consumer Exp KMeans original	Unexposed	K-Means clustering	admin	ConsumerExpenditure_v122_sampleandcluster	22 May 2018 10:29:
<input type="checkbox"/>	Consumer Exp PCA original	Unexposed	PCA	admin	ConsumerExpenditure_v122_sampleandcluster	22 May 2018 10:29:

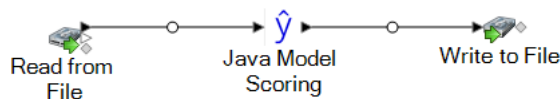
3. Click the Expose button () and click Yes at the prompt. The model is now exposed.

Second, we need to obtain new summarized yearly data. For the purposes of this demonstration, we will use the 5% test data that we created in **Create Variables > Divide Data and Write to File** with file name **Output_Expenditure2015_5perc.csv**.

Third, we need to create a new dataflow that scores the model.

1. Open Enterprise Designer.
2. Select File > New > Dataflow > Job.
3. Drag a Read from File stage onto the canvas and double-click that stage.
 - a. In the **File name** field, locate and specify Output_WithoutSubflow_Expenditure2015_5perc.csv.
 - b. In the **Record type** field, choose Delimited.
 - c. In the **Field separator** field, select Comma (,).
 - d. In the **Text qualifier** field, select Double quote (").
 - e. In the **Record separator** field, select Unix (U+000A).
 - f. Check **First row is header record**.
 - g. Uncheck **Treat records with fewer fields than defined as malformed** and click OK.
4. Drag a Java Model Scoring stage onto the stage. Double-click the stage to show the Model Scoring Options dialog box.
 - a. Select "K-Means clustering" as the **Type filter**.
 - b. Select either "Consumer Exp KMeans original" or "Consumer Exp KMeans from subflow" in the **Model name** dropdown.
 - c. The **Inputs** table shows information for the model's input fields. These fields and their data types automatically map to Spectrum fields and data types.
 - d. Click **Model Output** tab.
 - e. Click **Include** for each field whose data you want included in the model's output.
 - f. Click OK to save the model.
5. Drag a Write to File stage onto the canvas and connect it to the Java Model Scoring stage.
6. Save the file in CSV format and call it "Output_KMeans_Expenditure2015_5perc_scored.csv".
7. Click the Fields tab. Click the Quick Add button and select all of the fields.
8. Click OK and save the job.

Your dataflow should look like this:



After running the job, go to where the output file saved to and have a look at the resultant scored data file.

Subflow Substitution

Subflows are dataflows that can be reused in other dataflows as sources, sinks, or middle stages, which is helpful when you want to create a reusable process that can be easily incorporated into other jobs. In effect, the subflow becomes a custom stage in your dataflow.

These steps demonstrate how to create a subflow that generates test data by reusing the Group Statistics stage and subsequent Transformer stage that are added in the Create Variables section of this document. You can create a subflow on your own with these instructions, or you can use the provided subflow, `ConsumerExpenditure_122_Subflow.df`, in lieu of the Group Statistics and Transformer stages.

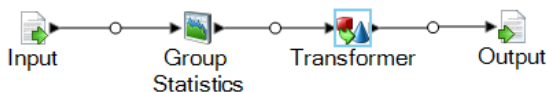
1. Select File > New > Dataflow > Subflow.
2. Drag an Input stage onto the canvas and open that stage.
3. Click Add and enter **Field name** "ALCBEVCQ" with **Data type** of double and press OK.
4. Repeat step 3 until you have added all of the following fields and data types.

ALCBEVCQ	FINATXEM	MISCPQ
ALCBEVPQ	FINCBTXM	PERINSCQ
APPARCQ	FOODCQ	PERINSPQ
APPARPQ	FOODPQ	PERSCACQ
CASHCOCQ	FSMPFRMX	PERSCAPQ
CASHCOPQ	HEALTHCQ	READCQ
CUID	HEALTHPQ	READPQ
EDUCACQ	HOUSCQ	TOBACCCQ
EDUCAPQ	HOUSPQ	TOBACCPQ
ENTERTCQ	INC_RNKM	TRANSCQ
ENTERTPQ	MISCCQ	TRANSPQ

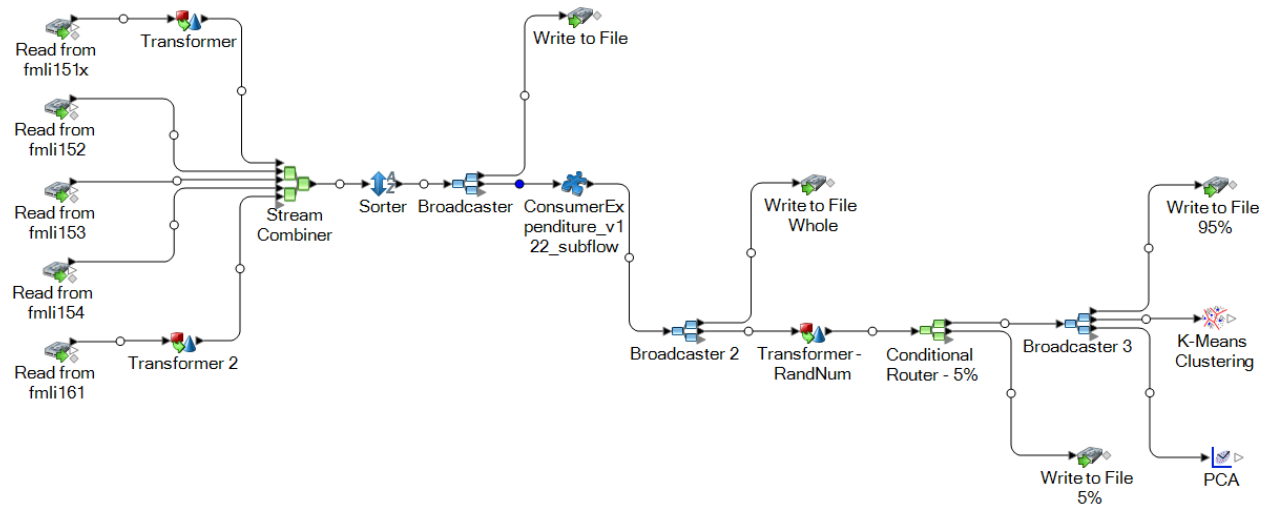
5. Press OK twice to return to the dataflow.
6. Copy the Group Statistics and Transformer stages just added to the dataflow and paste those stages onto the subflow canvas.
7. Attach the Group Statistics stage to the Input stage and the Transformer stage.
8. Drag an Output stage onto the canvas and connect it to the Transformer stage. Open the Output stage.
9. Click the "Expose" checkbox to expose all fields in the subflow's output and click OK to return to the dataflow.
10. Expose the created subflow by clicking the Expose icon (💡).

The subflow is now ready to be used in another dataflow.

Your subflow should look like this:



Your dataflow using the subflow should look like this:



Congratulations! You have finished your project!