

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ  
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Інститут Комп'ютерних наук та інформаційних технологій

Кафедра Програмної інженерії та інтелектуальних технологій управління

Спеціальність 122 Комп'ютерні науки

Освітня програма Комп'ютерні науки та інтелектуальні системи

**ЛАБОРАТОРНА РАБОТА №4 за курсом**

**«Інтелектуальний аналіз даних та видобування знань»**

Тема лабораторної роботи Побудова моделі кластеризації великих даних за допомогою інструментів SAP Predictive Analytics

Виконав студент 5 курсу, групи КН-М422

Захар ПАРАХІН

(підпис, прізвище та ініціали)

Перевірила Оксана ІВАЩЕНКО

(підпис, прізвище та ініціали)

Харків 2022

## ЗМІСТ

Вступ	3
1 Хід виконання роботи	4
1.1 Кластеризація	4
1.2 SAP Predictive Analytics	5
1.2.1 Аналіз набору даних	6
1.2.2 Побудова моделі	8
Висновки	21
Список джерел інформації	22

## Вступ

Кластеризація — це завдання поділу генеральної сукупності або точок даних на кілька груп, щоб точки даних у тих же групах були більш схожими на інші точки даних у тій самій групі та відрізнялися від точок даних в інших групах. Це в основному сукупність об'єктів на основі подібності та несхожості між ними.

В основному це тип методу навчання без нагляду. Метод неконтрольованого навчання — це метод, у якому ми створюємо посилення з наборів даних, що складаються з вхідних даних без позначених відповідей. Як правило, він використовується як процес пошуку значущої структури, пояснювальних базових процесів, генеративних ознак і групувань, властивих набору прикладів.

## 1 Хід виконання роботи

### 1.1 Кластеризація

Методи кластеризації:

Методи на основі щільності: ці методи розглядають кластери як щільну область, що має деякі подібності та відмінності від нижчої щільної області простору. Ці методи мають хорошу точність і здатність об'єднувати два кластери. Приклад DBSCAN (просторова кластеризація програм на основі щільності з шумом), OPTICS (упорядкування точок для ідентифікації структури кластеризації) тощо.

Ієрархічні методи: кластери, сформовані в цьому методі, утворюють деревовидну структуру на основі ієрархії. Нові кластери формуються з використанням раніше сформованого. Він розділений на дві категорії  
Агломерація (підхід знизу вгору)

Розбійний (підхід зверху вниз)

приклад CURE (кластеризація з використанням представників), BIRCH (збалансована ітеративна редукція кластеризації та використання ієрархій) тощо.

Методи поділу: ці методи поділяють об'єкти на  $k$  кластерів, і кожен поділ утворює один кластер. Цей метод використовується для оптимізації функції подібності об'єктивного критерію, наприклад, коли відстань є основним параметром, наприклад K-середні, CLARANS (кластеризація великих програм на основі рандомізованого пошуку) тощо.

Методи на основі сітки: у цьому методі простір даних формулюється у вигляді кінцевої кількості клітинок, які утворюють сіткову структуру. Усі операції кластеризації, що виконуються на цих сітках, є швидкими та не залежать від кількості об'єктів даних, наприклад STING (Statistical Information Grid), хвильовий кластер, CLIQUE (CLustering In Quest) тощо.

## 1.2 SAP Predictive Analytics

SAP Predictive Analytics — це рішення для статистичного аналізу та інтелектуального аналізу даних, яке дає змогу створювати прогнозні моделі для виявлення прихованої інформації та взаємозв'язків у ваших даних, на основі яких можна робити прогнози щодо майбутніх подій.

SAP Predictive Analytics поєднує в собі SAP InfiniteInsight і SAP Predictive Analysis в одній інсталяції на робочому столі. SAP Predictive Analytics містить два інтерфейси користувача: Automated Analytics і Expert Analytics.

Automated Analytics містить такі модулі:

- Data Manager — це інструмент семантичного рівня, який використовується для полегшення підготовки даних.
- Modeler допомагає створювати такі моделі, як класифікація, регресія, кластеризація, часові ряди та правила асоціації. Моделі можна експортувати в різні формати, щоб ви могли легко застосувати їх у своєму виробничому середовищі.
- Social витягує та використовує неявну структурну реляційну інформацію, що зберігається в різних типах наборів даних, покращуючи здатність моделей приймати рішення та прогнозувати. Він може представляти дані у вигляді графіків, які показують, як різні дані пов'язані між собою. Спеціалізовані робочі процеси допомагають створювати спільне розміщення та аналізувати часті шляхи на основі даних із геоприв'язкою.
- Recommendation генерує рекомендації продуктів для ваших клієнтів на основі аналізу соціальних мереж.

### 1.2.1 Аналіз набору даних

SAP Predictive Analytics® (Automated Analytics) - New Clustering Model

File

Help

</

### Рисунок 1 - Опис даних

Стовпець	Опис
Index	Цей стовпець є просто числовим індексом, який дає ідентифікаційний номер кожній змінній
Name	У цьому стовпці відображається ім'я змінної, знайдене в наборі даних
Storage	У цьому стовпці відображається тип змінної (число, ціле число, рядок, дата і т.д.)
Value	У цьому стовпці відображається тип вмісту змінної (безперервний, номінальний, порядковий, або текстовий)
Key	Цей стовпець показує, чи є змінна первинним ключем набору даних (значення 1) або вторинний ключ (значення 2). Змінні, які не є ключами, мають значення 0.
Order	Цей стовпець показує, чи має змінна порядок (значення 1) і чи може бути таким використовується в пункті Order By
Missing	Надає значення, яке буде використано, коли змінна має значення null
Group	У цьому стовпці відображається група, до якої належить

	змінна, якщо така є
Description	Цей стовпець містить текстовий опис змінної
Structure	Ця кнопка відкриває сторінку визначення структури, на якій можна перевизначити автоматичне кодування змінних

Таблиця 1 - Опис стовпців в датасеті

Зразок файл складається з даних про витрати клієнтів оптового дистриб'ютора. Набір даних містить 6 безперервних цілих полів, що показують річні витрати клієнта в різних сферах (Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicatessen), а також є автоматично створений атрибут індекс. Поле «Channel» показує, чи відбулася покупка через канал роздрібної торгівлі (значення 1) або в готелі чи ресторані (значення 2). Номінальне поле «Region» показує, чи відбувся продаж у Лісабоні (значення 1), у Порту (значення 2), або десь ще (значення 3).

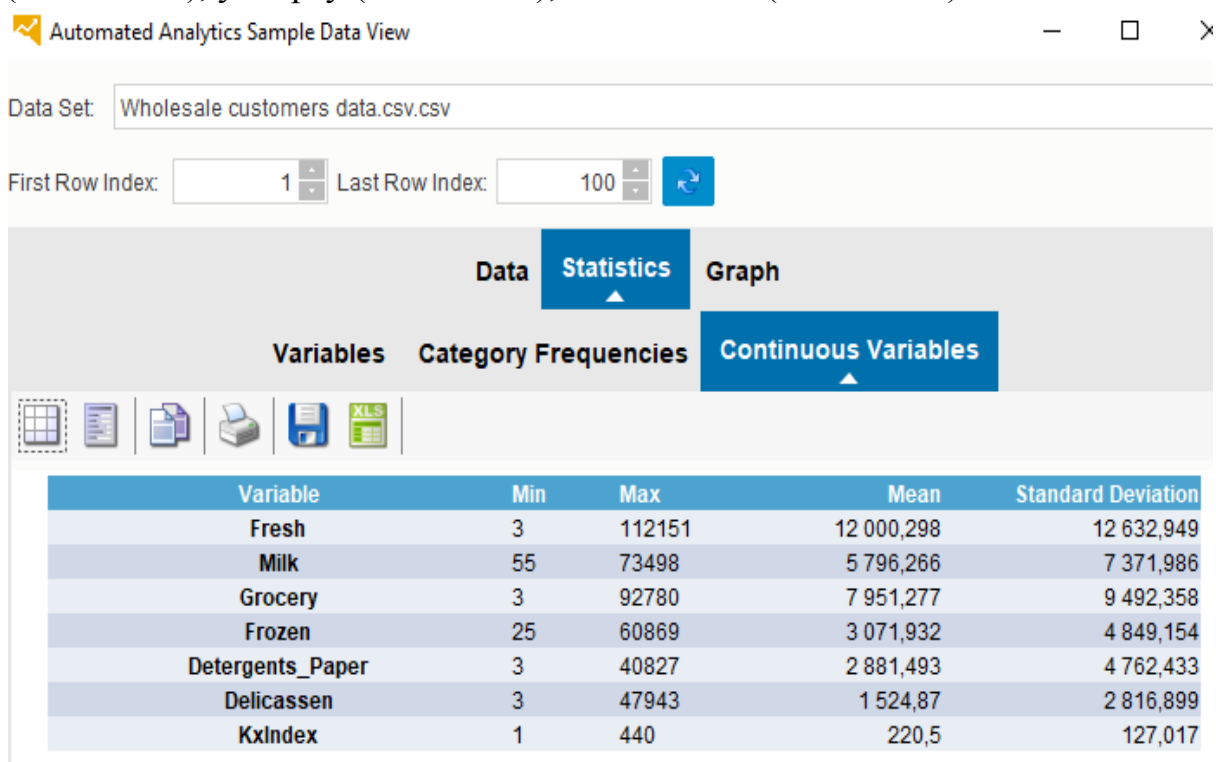


Рисунок 2 - Статистика безперервних полів

На Рисунку 2, бачимо максимум і мінімум, середнє і стандартне відхилення вказаних змінних.

Одразу переглядаємо в Statistics частоти всіх змінних, та їх статичні змінні як середнє.

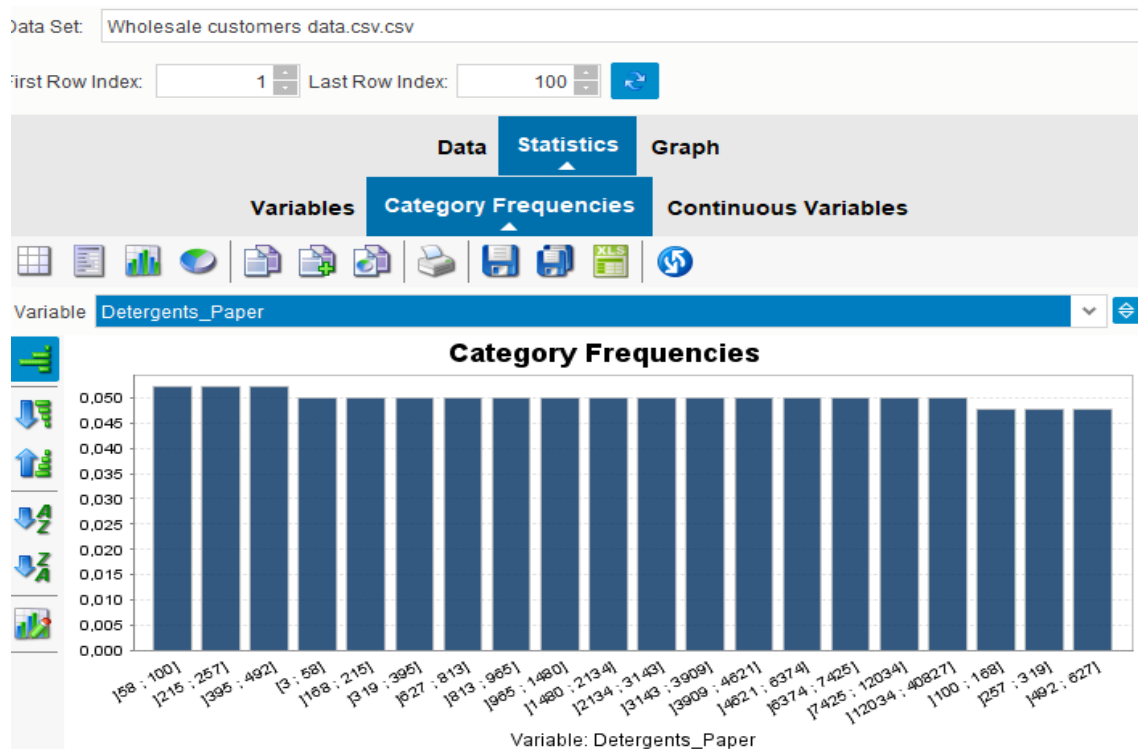


Рисунок 3 - Приклад виводу частот по змінній Detergents\_Paper

## 1.2.2 Побудова моделі

Для подальшої роботи необхідно визначити змінні (атрибути) для побудови моделі далі і розмістити їх за групами.

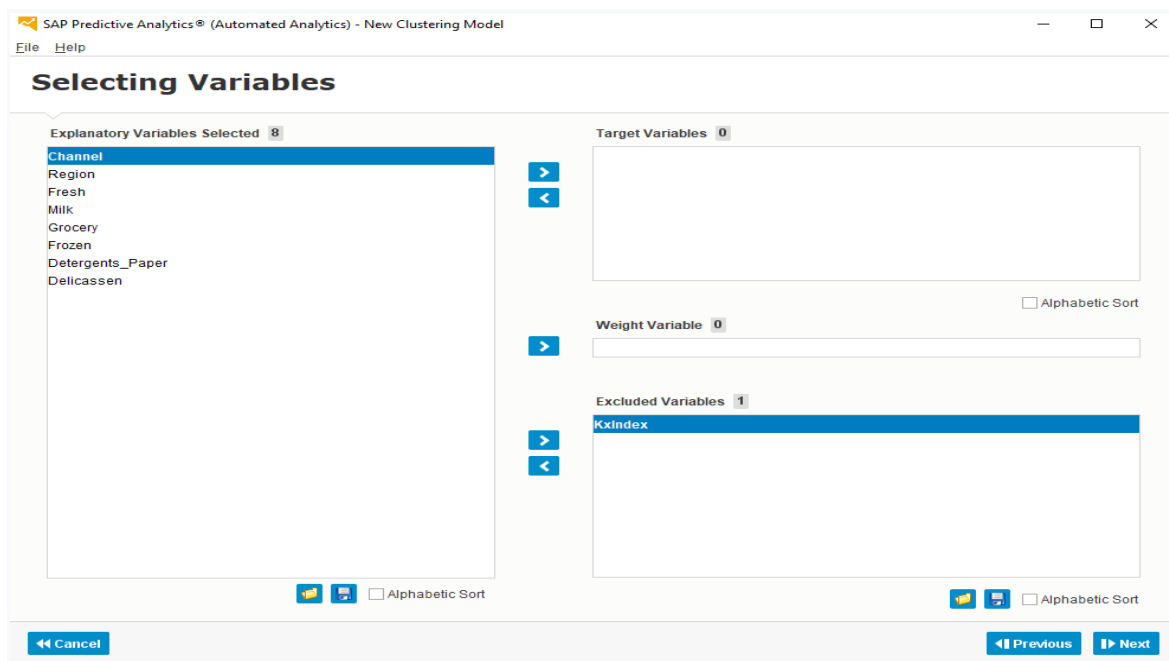


Рисунок 4 - Вікно вибору змінних



Група	Опис
Explanatory Variables Selected	Змінні, які розміщують в цій групі, описують і пояснюють модель. Для кластеризації потрібно розмістити усі змінні, які розглядаються як параметри для аналізу цієї групи.
Target Variables	Змінні, які розміщують в цій групі, використовуватимуться як цілі для контрольованої кластеризації. Кожна змінна створює окрему модель. Відсутність змінних на цій панелі означає, що інструмент використовуватиме неконтрольоване кластеризування.
Weight Variables	Змінні в цій групі, надають різні ваги для запису, до якого належать. Можна використовувати цю змінну, щоб змінювати визначення моделі, використовуючи записи, які вважаєте більш важливими. Фактично вказує скільки разів розглядати записи.
Excluded Variables	Змінні, що не будуть включені до моделі. Зазвичай індекси.

Таблиця 2 - Групи при виборі змінних

SAP Predictive Analytics® (Automated Analytics) - \_Wholesale customers data.csv

File Help

## Summary of Modeling Parameters

Model Name: \_Wholesale customers data.csv

Description:

### Kxen.SmartSegmenter

Data to be Modeled: C:\Users\999\Downloads\Wholesale customers data.csv.csv  
Cutting Strategy: Random without test  
Target Variable: None  
Weight Variable (Optional): None

Find the best number of clusters in this range: [ 10 ; 10 ]

Calculate SQL Expressions: ☒

Autosave... Export KxShell Script... Advanced...

Cancel Previous Generate

Рисунок 5 - Підсумок вибору параметрів

Можна змінити специфічні параметри моделі в Advanced, та вказується можливість автозбереження, Calculate SQL Expressions - визначається чи буде не використовуватися найкращій математичний підхід.

Find the best number of clusters in this range - надає змогу обрати кількість кластерів від мінімальної до максимальної, та однакове значення вказує на фіксацію кількості кластерів.

SAP Predictive Analytics® (Automated Analytics) - \_Wholesale customers data.csv

File
Help

# Training the Model

Stop
 View Type
 Copy
 Print
 Save
 Export to PowerPoint

Current Report

All Reports

Report Type:
Model Overview

## Model Overview

### Overview

Model: \_Wholesale customers data.csv

Data Set: Wholesale customers data.csv.csv

Initial Number of Variables: 9

Number of Selected Variables: 8

Number of Records: 440

Building Date: 2022-12-04 17:52:25

Learning Time: 2 s

Engine Name: Kxen.SmartSegmenter

Author: 999

Minimum Requested Number of Clusters: 5

Maximum Requested Number of Clusters: 10

Percentage of Unassigned Records: 6,35%

SQL Expressions: enabled

### Cluster Counts

clusterId

Initial Number of Clusters: 5

Final Number of Clusters: 5

Overlap: 23,12%

Percentage of Unassigned Records: 6,35%

### Cluster Count

Target: clusterId

Model	Winner	Initial Number of Clusters	Final Number of Clusters	Overlap	Percentage of Unassigned Records
Engine0	true	5	5	23,12%	6,35%
Engine1	false	6	6	29,32%	6,67%
Engine2	false	7	7	25,96%	4,76%
Engine3	false	8	8	29,47%	4,76%
Engine4	false	9	9	26,15%	2,54%
Engine5	false	10	10	29,02%	4,76%

Cancel

Previous

Next

Рисунок 6 -Огляд моделі після генерації

Відсоток не взятих до кластерів 6,35, також межі кластерів змінив до цього з 7 до 10 як ренж. Overlap - вказує, що 23,12 % увійшли до декількох кластерів.

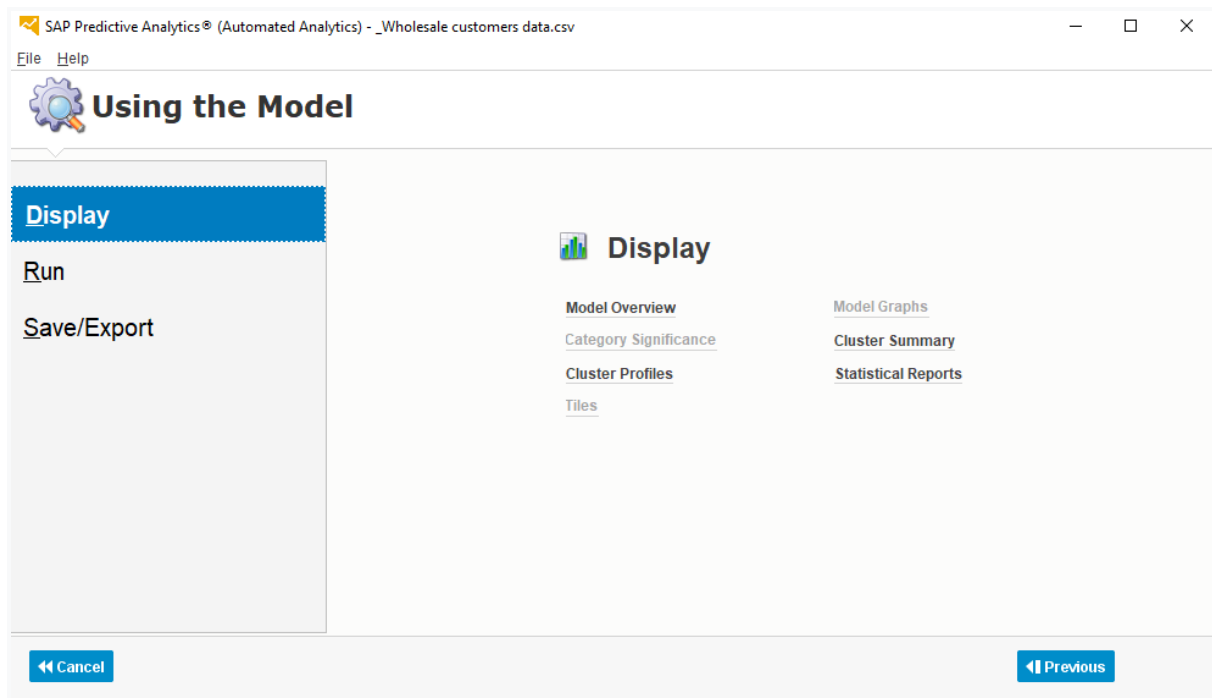


Рисунок 7 - Інтерфейс для використання моделі після генерації

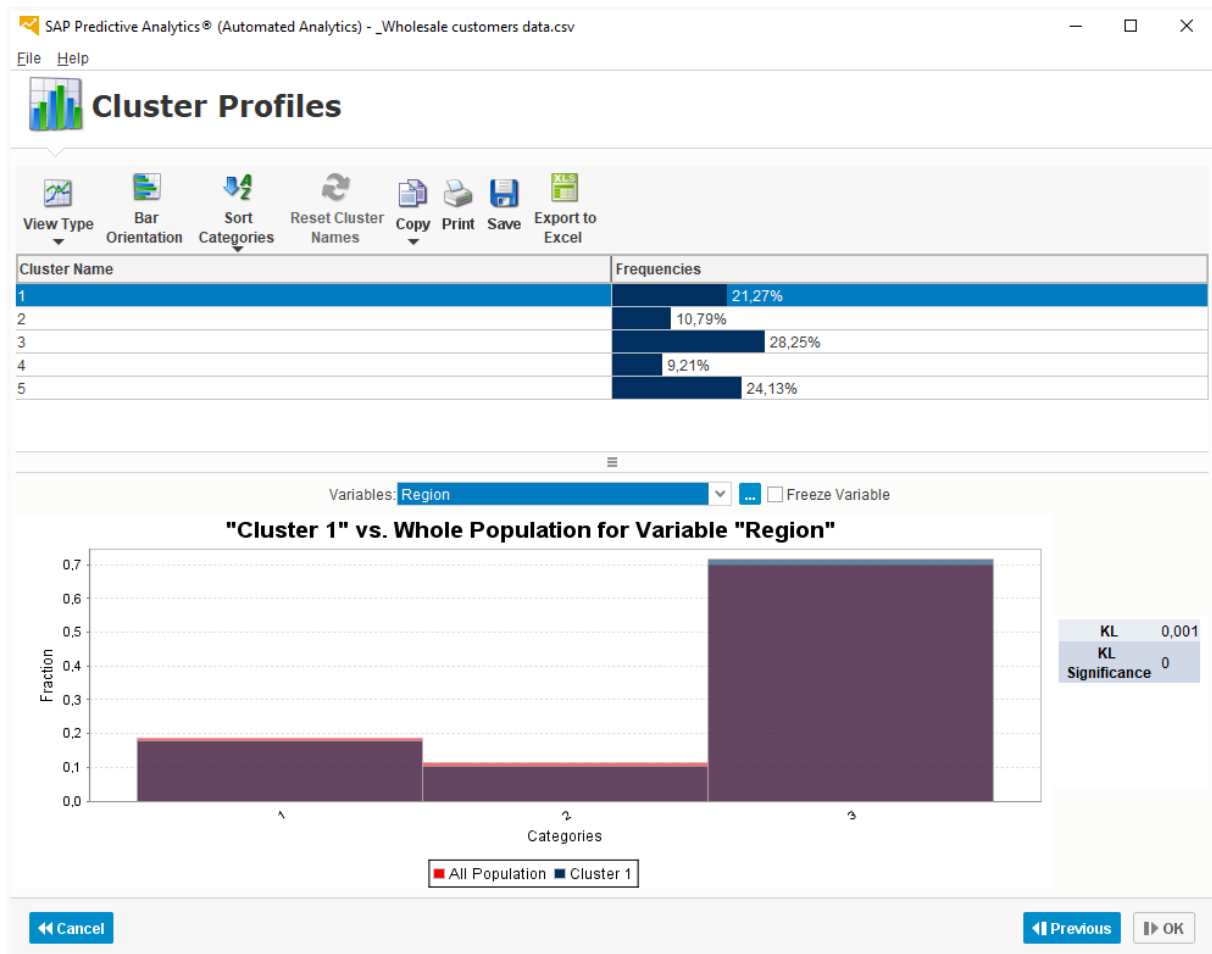


Рисунок 8 - Профіль Кластерний

У верхній таблиці показано список усіх кластерів і відсотки, які вони представляють. У нашому прикладі кластер 3 містить 28,25% записів набору даних.

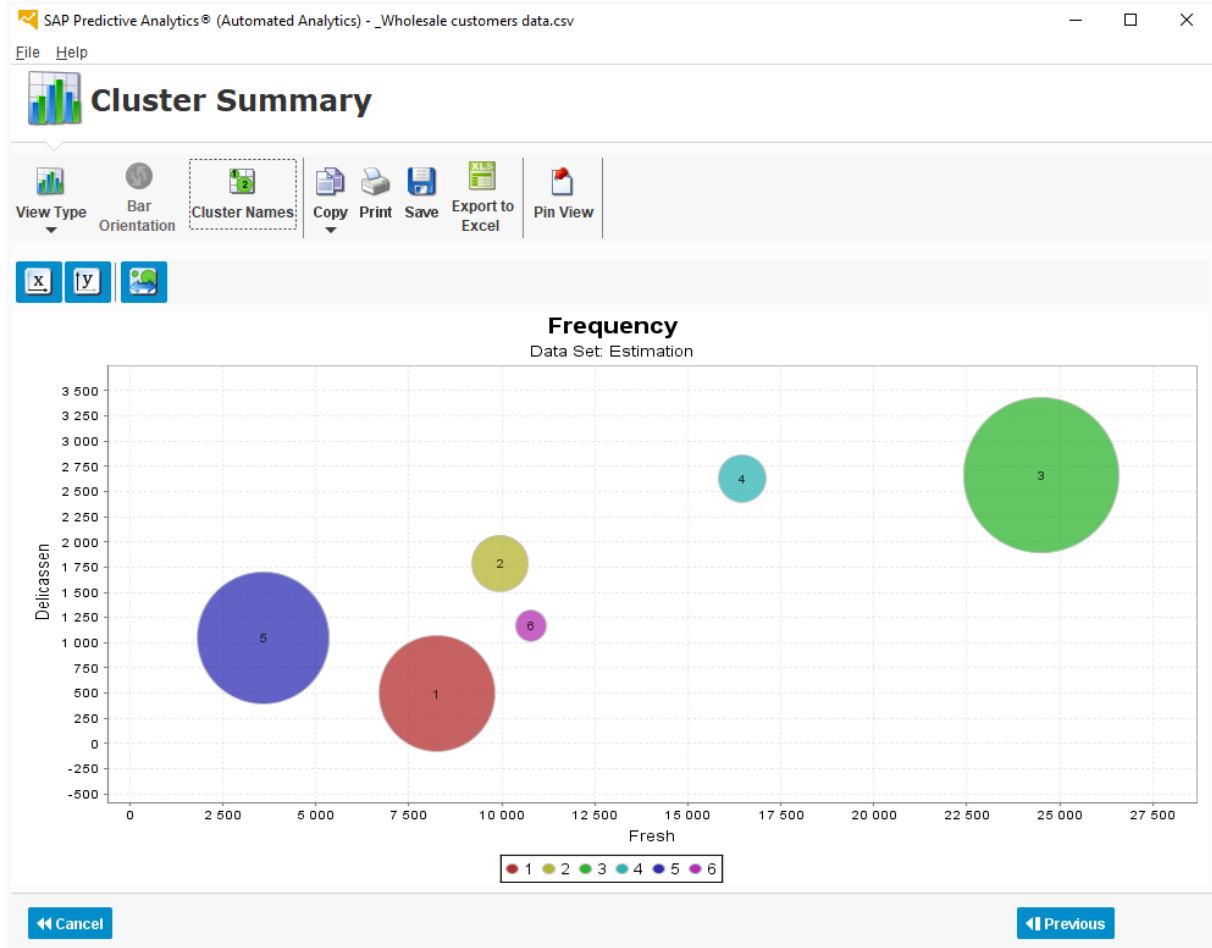


Рисунок 9 - Підсумок Кластеризації

У підсумку видно графічне відображення розмірів кластерів, звідси їх розміщення корелюється з вибором змінних для осей, так на Fresh 1,2 і 6 перетинаються.

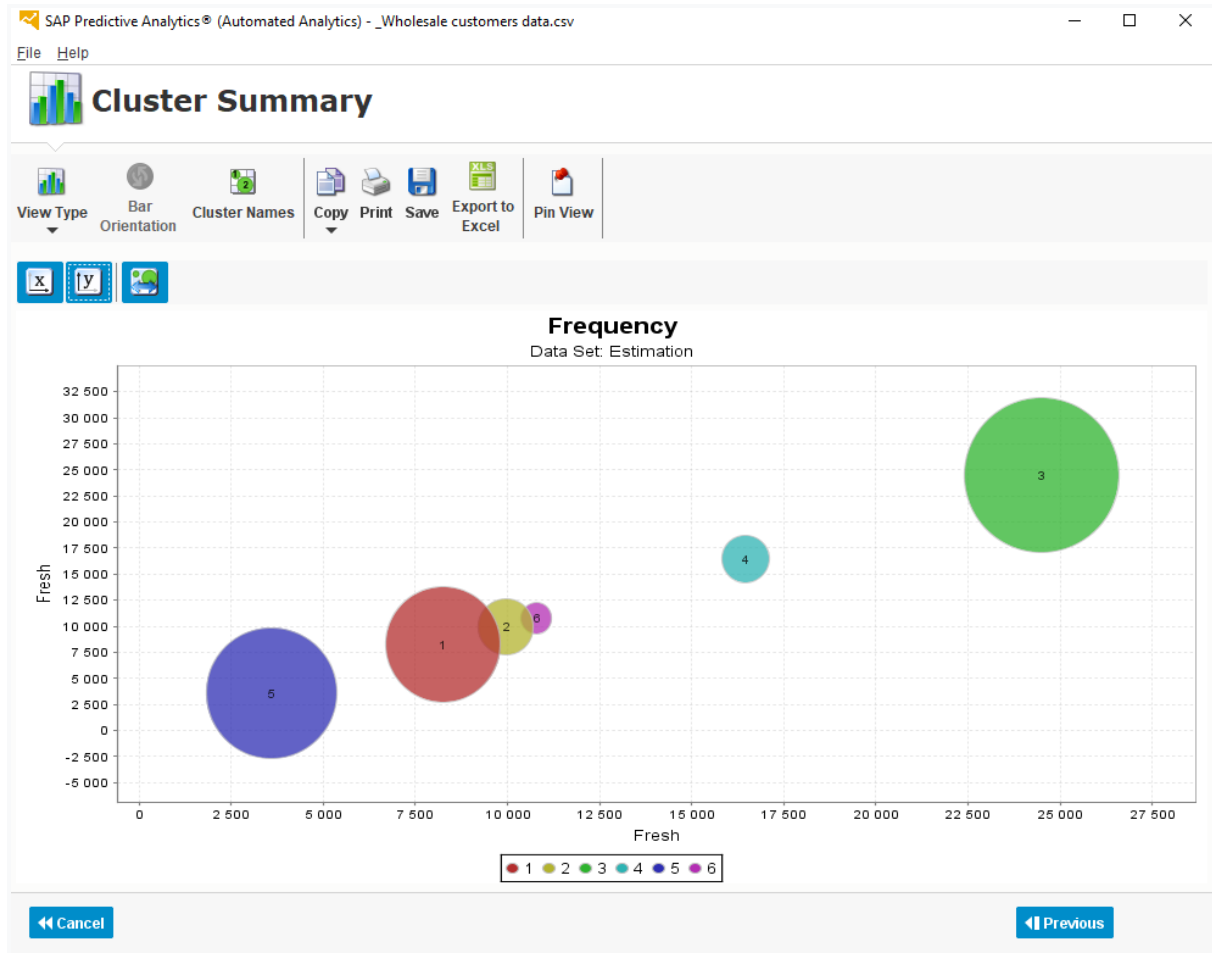


Рисунок 10 - зміна змінної на Fresh

Тепер проведемо кластеризацію з обраними змінними.

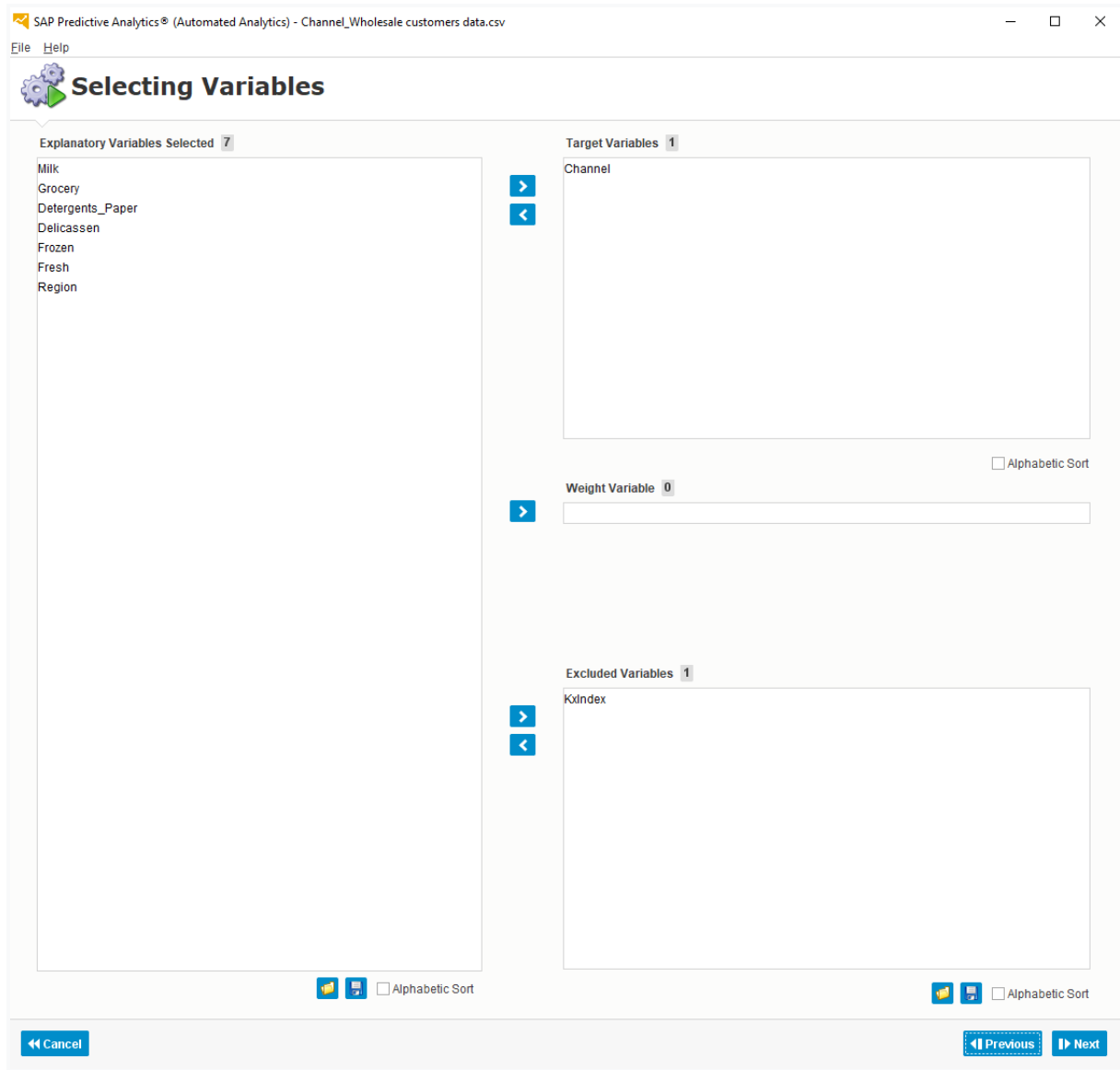



Рисунок 11 - Вибір змінних для кластеризації

Пояснення цього вибору просте, візьмемо Channel для цільової змінної, щоб дізнатися кластеризацію клієнтів.

SAP Predictive Analytics® (Automated Analytics) - Channel\_Wholesale customers data.csv

FileHelp



Summary of Modeling Parameters

Model Name: Channel\_Wholesale customers data.csv

Description:

Kxen.SmartSegmenter

Data to be Modeled: C:\Users\999\Downloads\Wholesale customers data.csv.csv

Cutting Strategy: Random without test

Target Variable: Channel

Weight Variable (Optional): None

Find the best number of clusters in this range: [ 2 : 5 ]

Calculate SQL Expressions: ☒

Autosave...Export KxShell Script...

Advanced...

Cancel

PreviousGenerate

Рисунок 12 - Обрані параметри



**Training the Model**

Current Report

All Reports

Report Type: Model Overview

**Overview**

Model: Channel_Wholesale customers data.csv	
Data Set:	Wholesale customers data.csv.csv
Initial Number of Variables:	9
Number of Selected Variables:	7
Number of Records:	440
Building Date:	2022-12-05 00:55:24
Learning Time:	1 s
Engine Name:	Kxen.SmartSegmenter
Author:	999
Minimum Requested Number of Clusters:	2
Maximum Requested Number of Clusters:	5
SQL Expressions:	enabled

**Nominal Targets**

Channel	Target Key	2
	1 - Frequency	68,25%
	2 - Frequency	31,75%

**Performance Indicators**

Target: Channel

kc_Channel	Predictive Power (KI)	0,8741
	Prediction Confidence (KR)	0,9562

**Cluster Counts**

Channel	Initial Number of Clusters	5
	Final Number of Clusters	5
	Overlap	10%
	Percentage of Unassigned Records	1,9%

**Cluster Count**

Target: Channel

Model	Winner	KI	KR	Initial Number of	Final Number of	Overlap	Percentage of Unassigned
-------	--------	----	----	-------------------	-----------------	---------	--------------------------

Cancel

Previous

Next

Рисунок 13 - Результати генерації

Power Predictable вказує (87%) на достатньо високий рівень підтримки точності за цільовою змінною.

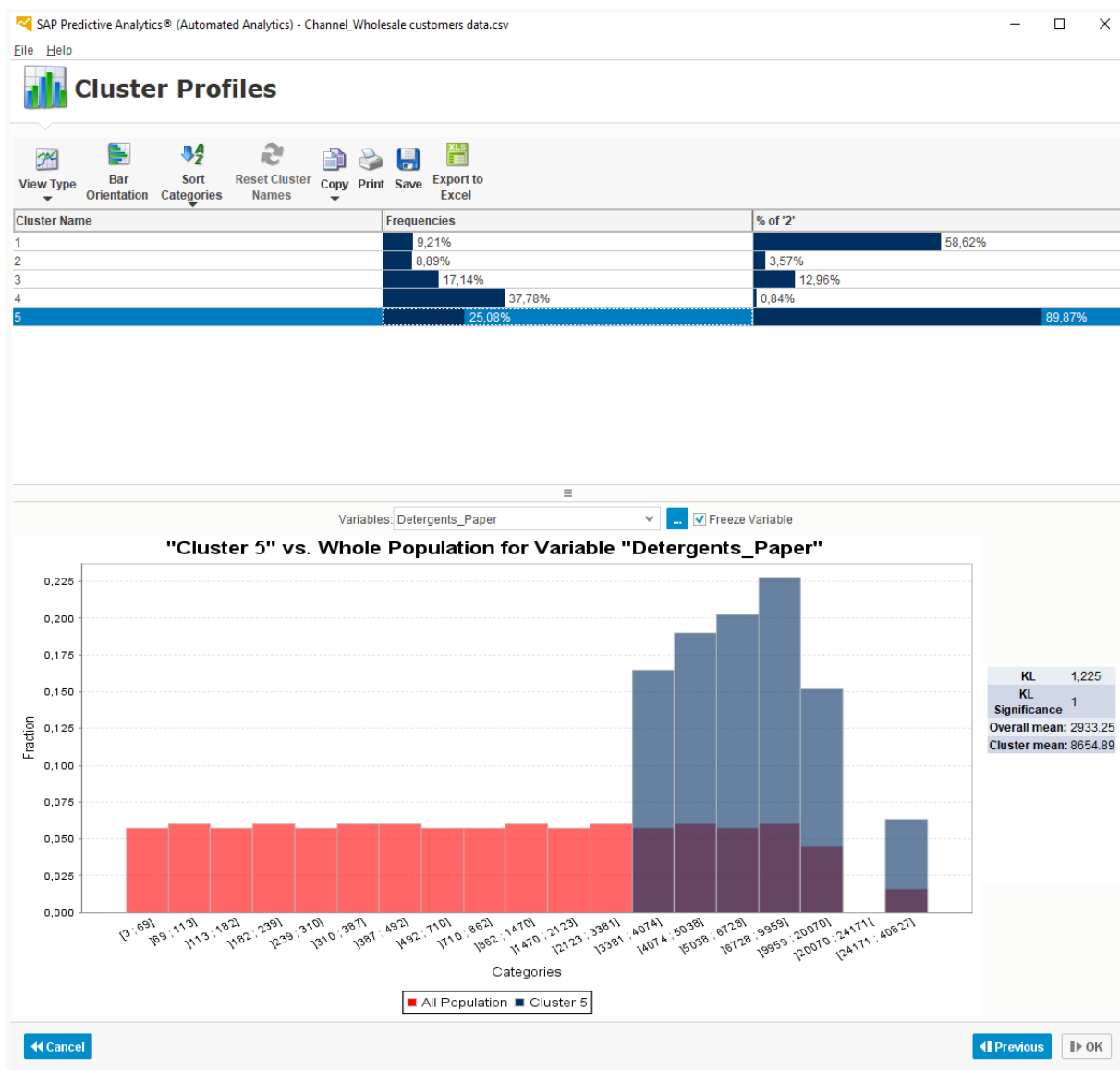


Рисунок 14 - Кластерний профіль

Кластери 4, 5 мають разом більшість серед усіх, до цього відрізняються типом Channel так як 4 кластер відноситься до значення 1, а п'ятий до 2. Також у них відрізняються інші категорії, так кластер 5 має більшу частоту в Milk, Detergents\_Paper, Groceries, а четвертий в усіх інших.

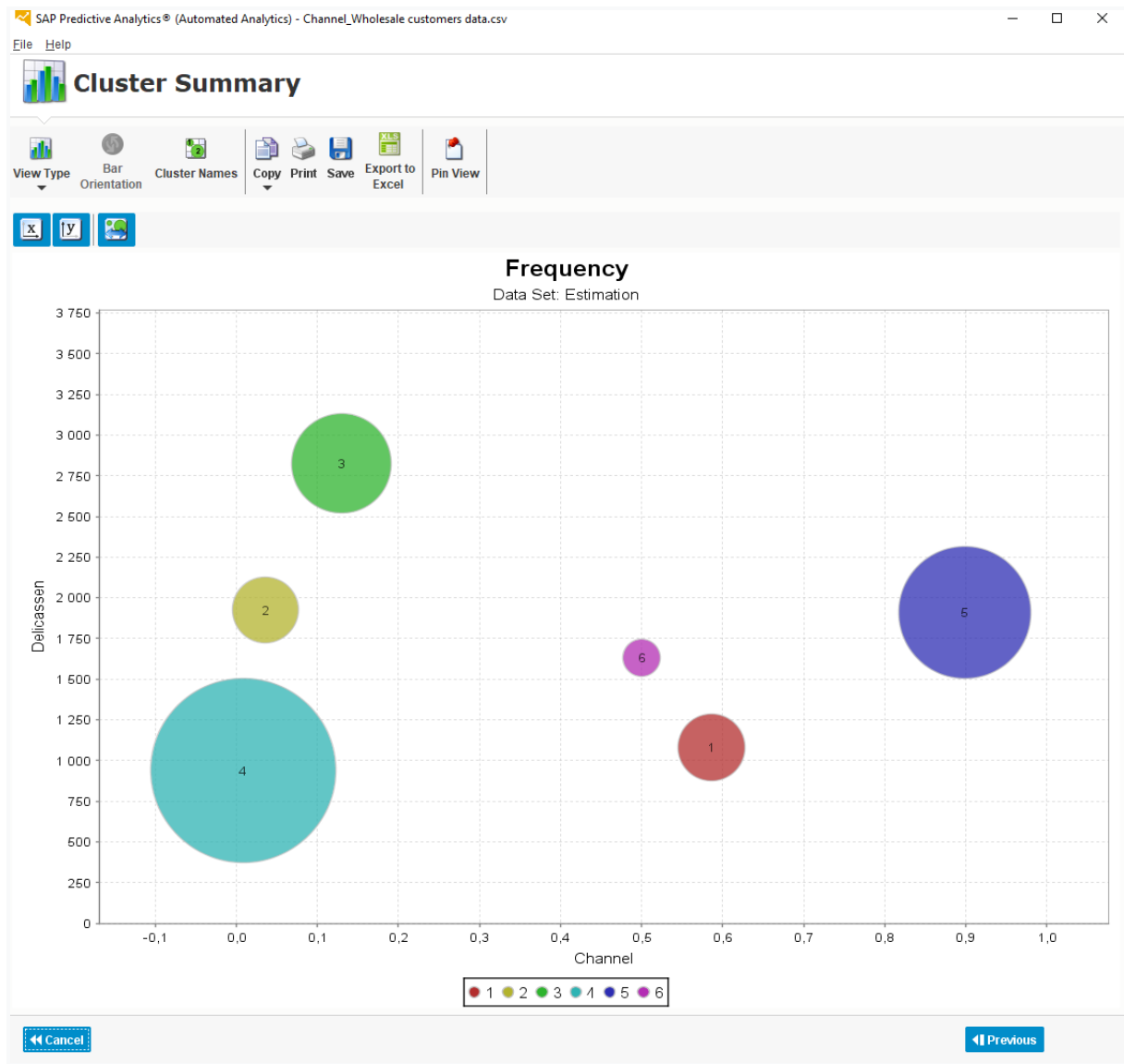


Рисунок 15 - підсумок змінної

Це створює підставу для орієнтації певної діяльності на ці два кластера. За допомогою контрольованої кластеризації інструмент генерує всю статистичну інформацію, пов'язану з значущістю категорії

SAP Predictive Analytics® (Automated Analytics) - Fresh\_Frozen\_Wholesale customers data.csv

File Help

## Applying the Model

**Application Data Set**

Data Type: Text Files

Folder: C:\Users\999\Downloads [Browse](#)

Data: [Browse](#) [Define Mapping](#)

**Generation Options**

Generate: Predicted Value Only [Advanced Apply Settings...](#)

Mode: Apply

**Results Generated by the Model**

Data Type: Text Files

Folder: C:\Users\999\Downloads [Browse](#)

Data: [Browse](#) [Define Mapping](#)

☐ Use Direct Apply in the Database

[Cancel](#) [The input file is missing.](#) [Previous](#) [Apply](#)

Рисунок 16 - Інтерфейс для використання сформованої моделі

## **Висновки**

Було вивчено, що таке кластеризація та як успішно визначити модель кластеризації за допомогою Automated Analytics у SAP Predictive Analytics. Також закріплено знання про більшість варіантів налаштування моделі відповідно до певних потреб і розуміння різниці між контрольованим і не контрольованим кластеризуванням.

## СПИСОК ДЖЕРЕЛ ІНФОРМАЦІЇ

- 1 Data Mining: Practical Machine Learning Tools and Technique \ Ian Witten, Eibe Frank, Mark Hall
- 2 Інтелектуальний аналіз даних: Навчальний посібник \ А. О. Олійник, С. О. Субботін, О. О. Олійник