Universidad Nacional Autónoma de México
Facultad de Ciencias

Profesores:
Dra. Amparo López Gaona
M. I. Gerardo Avilés Rosas

Alumnos:
Vázquez Lázaro José Luis (411067432)

Introducción

El presente trabajo se basa en un escenario ficcticio, en el cual, el cliente que contrata nuestros servicios profesionales (evidentemente en minería de datos) tiene un alto cargo en el gobierno de los Estados Unidos, específicamente en el departamento de agricultura. Este está interesado en las ventas de productos procedentes del campo (frutas, verduras, cereales, quesos, vinos, etc.) en todo el territorio estadounidense a través de supermercados y mercados sobre ruedas; pues está preocupado por el creciente aumento y demanda de productos sintéticos y/o preprocesados, que por cierto, son más baratos. Para realizar el análisis, el cliente nos ha proporcionado un *dataset* que contiene datos acerca de productos del campo vendidos en supermercados y en mercados sobre ruedas en todo el territorio estadounidense.

Al establecer este escenario, tendremos un contexto más específico para la utilización de la metodología ***CRISP-DM***. El entendimiento del negocio reflejará los intereses del cliente; una vez definido, nos proporcionará una guía para reducir el dataset de trabajo. La parte de entendimiento de los datos se trabajará como un diccionario de datos, el cual incluirá una exploración estadística que permita verificar la calidad de estos. Este punto junto con el entendimiento de los datos, definirán el preprocesamiento de los datos. Con respecto al modelado, se eligieron a las reglas de asociación (técnica para tarea de minería de datos predictiva) y al clustering (técnica para tarea de minería de datos descriptiva) como las tareas de minería de datos que se aplicarán al dataset. La elección de las reglas de asociación se debe a su antecedente inherente en el contexto de compras de productos en supermercados; mientras que la elección del clustering se debe a que éste nos permite trabajar con un número relativamente pequeño de grupos de datos, cada uno de los cuales tiene un conjunto de característica singulares que proporciona patrones útiles, quizás interesantes e inesperados, que ayudan a formular conclusiones bastante directas.

Las herramientas de software que se utiizarán para llevar a cabo el ánalisis de los datos, el preprocesamiento y las tareas de minería de datos serán ***R*** y ***RapidMiner***.

Metodología *CRISP-DM*

1. Entendimiento del negocio

A medida que la venta de productos sintéticos y/o preprocesados en supermercados y mercados sobre ruedas ha aumentado por sus bajos costos, el departamento de agricultura de los Estados Unidos ha detectado una disminución en las ventas de productos procedentes directamente del campo. Al afrontar la realidad de que los productos del campo están siendo desplazados por los productos sintéticos y/o preprocesados, el secretario de agricultura debe encontrar fórmulas para aumentar la rentabilidad de los productos agricolas, sin aumentar el coste de la adquisición de estos por parte de los clientes.

(a) Objetivos del negocio.

Los siguientes son los objetivos deseables:

- Mejorar las ventas de los productos agrícolas realizando mejores recomendaciones.
- Fomentar el consumo de productos del campo con una mejor distribución regional de estos.

Se considerarán un éxito si:

- Las ventas aumentan al menos un 10%.
- Se consigue una distribución regional adecuada de los productos agrícolas de al menos un 50%.

(b) Objetivos de la minería de datos.

Hemos traducido los objetivos del negocio en términos de minería de datos, dando como resultado los siguientes objetivos:

- Utilizar la información proporcionada por el *dataset* para generar un modelo que enlace elementos "relacionados": **análisis de la cesta de compras**.
- Utilizar la información proporcionada por el dataset para generar un modelo que permita vizualizar la distribución regional de los productos agrícolas: **análisis de la geografía de la oferta**.

Se considerarán un éxito si:

- El **análisis de la cesta de compras** proporciona al menos 2 patrones útiles e interesantes.
- El **análisis de la greografia de la oferta** proporciona al menos 3 distribuciones regionales distintas.

2. Entendimiento de los datos.

Conjunto de datos *MercadoProductos*.

I. Información general.

Propietario y/o donador.

Nombre: *Dra. Amparo López Gaona*
Procedecia: *Universidad Nacional Autónoma de México, Facultad de Ciencias, Departamento de Matemáticas.*
Correo electrónico: alg@ciencias.unam.mx

Descripción.

Los datos corresponden a productos del campo vendidos en supermercados y mercados sobre ruedas en todo el territorio estadounidense.

Número de atributos.

45 atributos que describen los distintos supermercados y mercados sobre ruedas en los Estados Unidos y los productos del campo que estos venden.

Número de registros.

8144 registros.

Características de los atributos.

1 atributo identificador, 2 atributos numéricos y 42 atributos nominales.

II. Diccionario de datos.

✓ ***FMID***. Identificador del supermercado ó del mercado sobre ruedas.

Tipo de atributo*: Entero no aritmético*
Dominio: *{ 20001, ..., 1008929 }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *no aplica*
Desviación estándar: *no aplica*

✓ ***MarketName***. Nombre del supermercado ó del mercado sobre ruedas.

Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas de caracteres finitas que resulta inconveniente listar*

Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Winter Farmers Market and Meal for Hope*
Desviación estándar: *no aplica*

✓ ***Website***. Página de internet del supermercado o mercado sobre ruedas.
Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas de caracteres finitas y válidas como direcciones web que resulta inconveniente listar*
Valores ausentes: *hay 3610 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *44.32%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *http://www.grownyc.org*
Desviación estándar: *no aplica*

✓ ***street***. Calle dónde se ubica el supermercado o mercado sobre ruedas.
Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas de caracteres finitas que resulta inconveniente listar*
Valores ausentes: *hay 203 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *2.49%*
Valores atípicos: *{ www.stonecounty.locallygrown.net (1), www.twincitieslocalfood.com (1) }*
Porcentaje de valores atípicos con respecto al total: *0.02%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Main Street*
Desviación estándar: *no aplica*

✓ ***city***. Ciudad dónde se ubica el supermercado o mercado sobre ruedas.
Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas de caracteres finitas que resulta inconveniente listar*
Valores ausentes: *hay 20 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0.24%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Chicago*
Desviación estándar: *no aplica*

✓ ***Country***. Condado dónde se ubica el supermercado o mercado sobre ruedas.
Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas de caracteres finitas que resulta inconveniente listar*
Valores ausentes: *hay 643 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *7.89%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Los Angeles*
Desviación estándar: *no aplica*

✓ ***State***. Estado dónde se ubica el supermercado o mercado sobre ruedas.
Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas de caracteres finitas que resulta inconveniente listar*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *{ Calafornia (5), Miinesota (1), Virgin Islands (4), Viriginia (1) }*
Porcentaje de valores atípicos con respecto al total: *0.13%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *California*
Desviación estándar: *no aplica*

✓ ***zip***. Código postal correspondiente a la dirección del supermercado o mercado sobre ruedas.
Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas alfanuméricas finitas que resulta inconveniente listar*
Valores ausentes: *hay 962 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *11.81%*
Valores atípicos: *{ MA (1), n/a (1) }*
Porcentaje de valores atípicos con respecto al total: *0.02%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *60602*
Desviación estándar: *no aplica*

✓ ***Season1Date***. Fechas de la primera temporada de ventas de productos del campo en el supermercado o mercado sobre ruedas.
Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas finitas que describen un periodo en meses que resulta inconveniente listar*
Valores ausentes: *hay 4161 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *51.09%*
Valores atípicos: *no hay valores atípicos*

Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *May to October*
Desviación estándar: *no aplica*

✓ ***Season1Time***. Horario de la primera temporada de ventas de productos del campo en el supermercado o mercado sobre ruedas.

Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas finitas que describen un periodo en horas que resulta inconveniente listar*
Valores ausentes: *hay 3901 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *47.90%*
Valores atípicos: *No hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Sat: 9:00 AM-1:00 PM;*
Desviación estándar: *no aplica*

✓ ***Season2Date***. Fechas de la segunda temporada de ventas de productos del campo en el supermercado o mercado sobre ruedas.

Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas finitas que describen un periodo en meses que resulta inconveniente listar*
Valores ausentes: *hay 7915 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *97.18%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *November to April*
Desviación estándar: *no aplica*

✓ ***Season2Time***. Horario de la segunda temporada de ventas de productos del campo en el supermercado o mercado sobre ruedas.

Tipo de atributo*: Polinominal*
Dominio: *conjunto de cadenas finitas que describen un periodo en horas que resulta inconveniente listar*
Valores ausentes: *hay 7911 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *97.13%*
Valores atípicos: *No hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Sat: 9:00 AM-1:00 PM;*
Desviación estándar: *no aplica*

✓ ***Season3Date***. Fechas de la tercera temporada de ventas de productos del campo en el supermercado o mercado sobre ruedas.

Tipo de atributo: *Polinominal*
Dominio: *conjunto de cadenas finitas que describen un periodo en meses que resulta inconveniente listar*
Valores ausentes: *hay 8102 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *99.48%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *01/02/2014 to 03/31/2014*
Desviación estándar: *no aplica*

✓ ***Season3Time***. Horario de la tercera temporada de ventas de productos del campo en el supermercado o mercado sobre ruedas.

Tipo de atributo: *Polinominal*
Dominio: *conjunto de cadenas finitas que describen un periodo en horas que resulta inconveniente listar*
Valores ausentes: *hay 8102 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *99.48%*
Valores atípicos: *No hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Sat: 10:00 AM-12:00 PM;*
Desviación estándar: *no aplica*

✓ ***Season4Date***. Fechas de la cuarta temporada de ventas de productos del campo en el supermercado o mercado sobre ruedas.

Tipo de atributo: *Polinominal*
Dominio: *conjunto de cadenas finitas que describen un periodo en meses que resulta inconveniente listar*
Valores ausentes: *hay 8138 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *99.92%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *12/14/2013 to 12/14/2013*
Desviación estándar: *no aplica*

✓ ***Season4Time***. Horario de la tercera temporada de ventas de productos del campo en el supermercado o mercado sobre ruedas.

Tipo de atributo: *Polinominal*

Dominio: *conjunto de cadenas finitas que describen un periodo en horas que resulta inconveniente listar*
Valores ausentes: *hay 8138 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *99.92%*
Valores atípicos: *No hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Fri: 11:00 AM-2:00 PM;*
Desviación estándar: *no aplica*

✓ ***Longitude***. Longitud de las coordenadas geográficas del supermercado o mercado sobre ruedas.
Tipo de atributo: *Real*
Dominio: *{ -159.718, ..., -35.544 }*
Valores ausentes: *hay 24 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0.29%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *-91.273*
Mediana: *-86.524*
Moda: *-71.577*
Desviación estándar: *17.599*

✓ ***Latitude***. Latitud de las coordenadas geográficas del supermercado o mercado sobre ruedas.
Tipo de atributo: *Real*
Dominio: *{ 17.710, ..., 77.058 }*
Valores ausentes: *hay 24 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0.29%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *39.334*
Mediana: *40.177*
Moda: *43.685*
Desviación estándar: *5.056*

✓ ***Location***. Tipo de lugar dónde se encuentra el supermercado o mercado sobre ruedas.
Tipo de atributo: *Polinominal*
Dominio: *{ Local government building grounds; Private business parking lot; Other, Closed-off public street; Faith-based institution (e.g., church, mosque, synagogue, temple); Educational institution; Healthcare Institution; On a farm from: a barn, a greenhouse, a tent, a stand, etc; Federal/State government building grounds; Co-located with wholesale market facility }*
Valores ausentes: *hay 3647 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *44.78%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*

Media: *no aplica*
Mediana: *no aplica*
Moda: *Local government building grounds*
Desviación estándar: *no aplica*

✓ ***Credit***. Indica si el supermercado o mercado sobre ruedas da crédito.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Y*
Desviación estándar: *no aplica*

✓ ***WIC***. Indica si el supermercado o mercado está incorporado al Programa Especial de Nutrición Suplementaria para Mujeres, Infantes y Niños; que es un programa gubernamental.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***WICcash***. Indica si el supermercado o mercado acepta vales proporcionados por el programa *WIC*.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***SFMNP***. Indica si el supermercado o mercado está incorporado al Programa de Nutrición del Mercado de Granjeros para Personas de la Tercera Edad; que es un programa gubernamental.

Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***SNAP***. Indica si el supermercado o mercado está incorporado al Programa de Asistencia Nutricional Suplementaria; que es un programa gubernamental.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *hay 24 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0.29%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Bakedgoods***. Indica si el supermercado o mercado vende productos horneados artesanales.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Cheese***. Indica si el supermercado o mercado vende quesos artesanales.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*

Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Crafts***. Indica si el supermercado o mercado sobre ruedas vende artesanías.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Flowers***. Indica si el supermercado o mercado sobre ruedas vende flores cultivadas artesanalmente.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Eggs***. Indica si el supermercado o mercado sobre ruedas vende huevo fresco.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Seafood***. Indica si el supermercado o mercado vende mariscos frescos.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*

Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Herbs***. Indica si el supermercado o mercado sobre ruedas vende hiervas frescas para el consumo humano.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Vegetables***. Indica si el supermercado o mercado sobre ruedas vende verdura fresca.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *Y*
Desviación estándar: *no aplica*

✓ ***Honey***. Indica si el supermercado o mercado vende miel extraída de forma artesanal.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Jams***. Indica si el supermercado o mercado sobre ruedas vende mermeladas artesanales.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Maple***. Indica si el supermercado o mercado sobre ruedas vende miel de maple artesanal.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Meat***. Indica si el supermercado o mercado sobre ruedas vende carne fresca.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Nursery***. Indica si el supermercado o mercado posee un vivero.
Tipo de atributo: *Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*

Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Nuts***. Indica si el supermercado o mercado sobre ruedas vende nueces frecas.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Plants***. Indica si el supermercado o mercado sobre ruedas vende plantas cultivadas artesanalmente.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Poultry***. Indica si el supermercado o mercado vende aves de corral.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Prepared***. Indica si el supermercado o mercado vende abono preparado, es decir, composta.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*

Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Soap***. Indica si el supermercado o mercado sobre ruedas vende jabón artesanal.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Tree***. Indica si el supermercado o mercado vende aves de árboles.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***Wine***. Indica si el supermercado o mercado vende vinos artesanales.
Tipo de atributo*: Binominal*
Dominio: *{ Y (Yes), N (No) }*
Valores ausentes: *no hay valores ausentes*
Porcentaje de valores ausentes con respecto al total: *0%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *N*
Desviación estándar: *no aplica*

✓ ***UpdateTime***. Fecha y/o hora del registro de la información, quizás sólo el año.

Tipo de atributo*: Binominal*
Dominio: *Conjunto de cadenas de caracteres finitas que desriben una fecha y/o hora o sólo el año que resulta inconveniente listar*
Valores ausentes: *hay 381 valores ausentes*
Porcentaje de valores ausentes con respecto al total: *4.67%*
Valores atípicos: *no hay valores atípicos*
Porcentaje de valores atípicos con respecto al total: *0%*
Media: *no aplica*
Mediana: *no aplica*
Moda: *2009*
Desviación estándar: *no aplica*

3. Preparación de los datos.

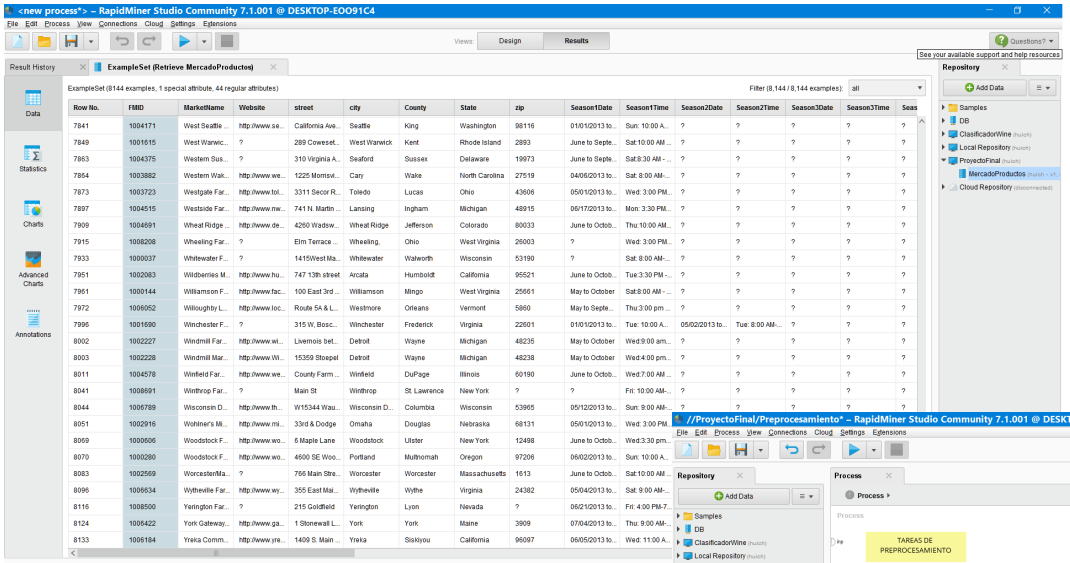
En base a los dos puntos anteriores, se realizarán las siguientes tareas de preprocesamiento:

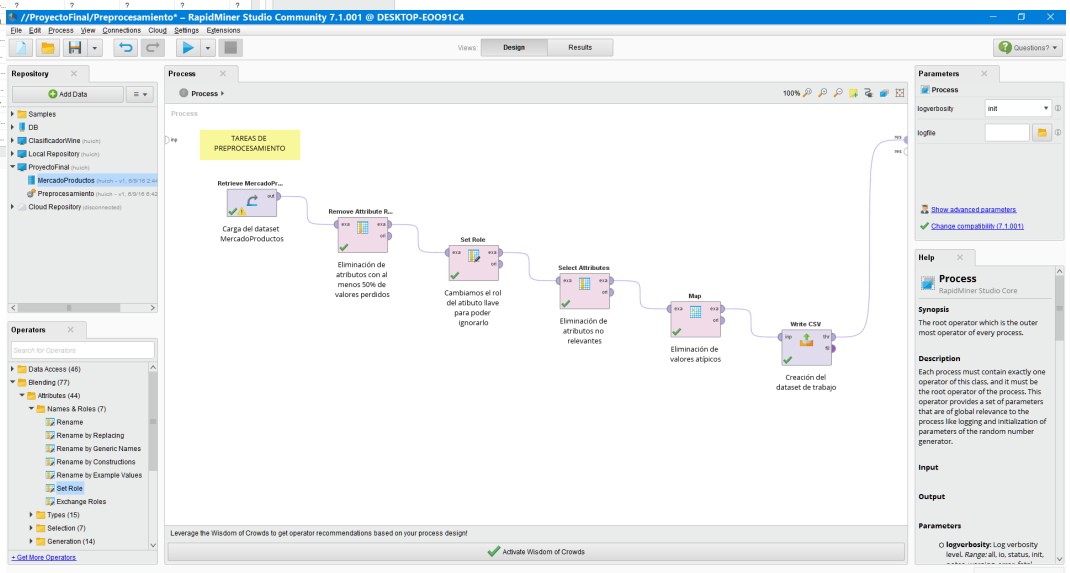
- Limpieza.
  - ✓ Se eliminarán aqellos atributos que tengan al menos *50%* de valores ausentes o perdidos; para evitar que "contaminen" el conocimiento que se va a descubrir.
  - ✓ Los valores atípicos del atributo *State* se deben a errores en la captura de los datos; por ejemplo, se registró "Calafornia" en lugar de "California". Como solamente hay 11 valores atípicos, estos serán corregidos manualmente.

- Reducción.
  - ✓ El conjunto de atributos *{ FMID, MarketName, Website, street, city, County, zip, Season1Time, Longitude, Latitude, Location, Credit, WIC, WICcash, SFMNP, SNAP, updateTime }* no es relevante para los objetivos del negocio, por lo que este conjunto de atributos será eliminado.

Estas tareas de preprocesamiento las realizaremos utilizando la herramienta ***RapidMiner***. Nuestro objetivo será la creación de un nuevo *dataset* como resultado de la aplicación de las tareas de preprocesamiento al *dataset* proporcionado por el cliente. Después del preprocesamiento, ocuparemos los *8144* registro resultantes (el total), pues **la regla de oro** para la cantidad adecuada de datos indica que *5000* registros o más son deseables.

Carga de datos y creación de un proceso para realizar el preprocesamiento en ***RapidMiner***:



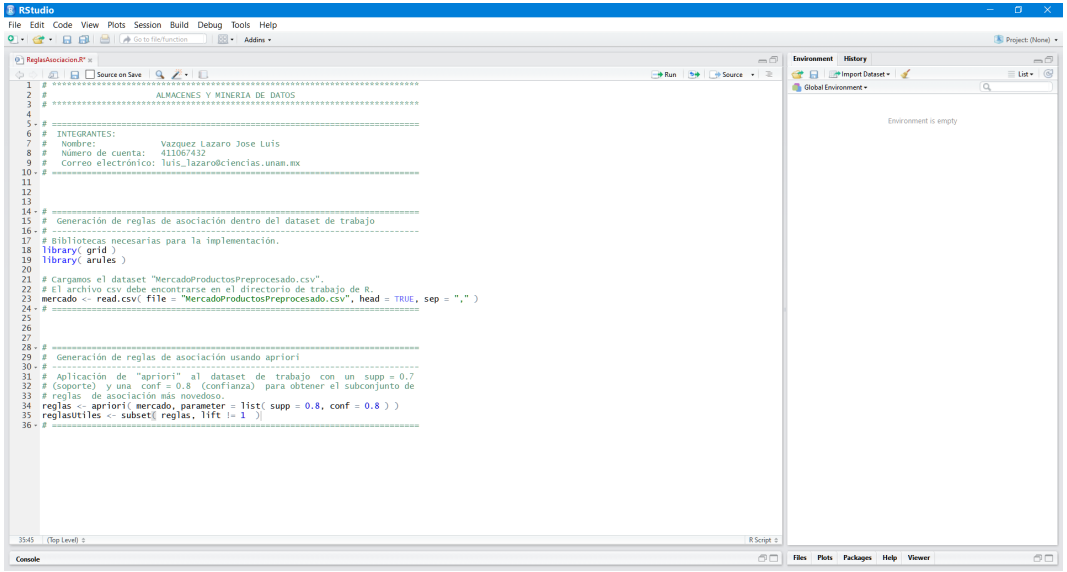


Reporte elaborado por José Luis Vázquez Lázaro (411067432) / 01 de junio de 2016.

4. Modelado.

El primer objetivo planteado es utilizar la información proporcionada por el *dataset* para generar un modelo que enlace elementos "relacionados": **análisis de la cesta de compras**. De esta descripción, es natural pensar en aplicar tareas de minería de datos predictivas al *dataset* de trabajo para alcanzar este objetivo; de hecho, esto es lo que haremos. Para lograr esta tarea, aplicaremos al *dataset* de trabajo la técnica de reglas de asociación; pues consideramos que es la técnica más apropiada por su fuerte conexión con las compras de productos en supermercados.

Usaremos ***R*** para generar reglas de asociación útiles e interesantes dentro del *dataset* de trabajo, específicamente ocuparemos la función ***apriori***. La siguiente imagen muestra el script creado para esta labor:

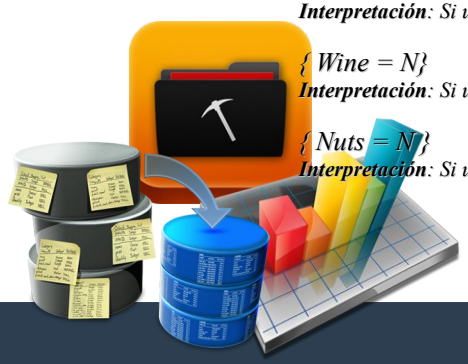



```
# ********************************************************************************
#                              ALMACENES Y MINERIA DE DATOS
# ********************************************************************************

# ================================================================================
#  INTEGRANTES:
#    Nombre:              Vazquez Lazaro Jose Luis
#    Número de cuenta:    411067432
#    Correo electrónico: luis_lazaro@ciencias.unam.mx
# ================================================================================




# ================================================================================
#  Generación de reglas de asociación dentro del dataset de trabajo
# --------------------------------------------------------------------------------
# Bibliotecas necesarias para la implementación.
library( grid )
library( arules )

# Cargamos el dataset "MercadoProductosPreprocesado.csv".
# El archivo csv debe encontrarse en el directorio de trabajo de R.
mercado <- read.csv( file = "MercadoProductosPreprocesado.csv", head = TRUE, sep = "," )
# ================================================================================




# ================================================================================
#  Generación de reglas de asociación usando apriori
# --------------------------------------------------------------------------------
#  Aplicación  de  "apriori"  al  dataset  de  trabajo  con  un  supp = 0.7
# (soporte)  y una  conf = 0.8  (confianza)  para obtener el subconjunto de
# reglas  de asociación más novedoso.
reglas <- apriori( mercado, parameter = list( supp = 0.8, conf = 0.8 ) )
reglasUtiles <- subset( reglas, lift != 1  )
# ================================================================================
```

La función ***apriori*** se aplicó con un soporte y una confianza iguales a *0.8*, pues consideramos que estos valores proporcionan una alta seguridad en la veracidad las reglas generadas y una buena cantidad de atributos involucrados.

Las reglas de asociación generadas con esta configuración y sus interpretaciones son las siguientes:

lhs		***rhs***	***support***	***confidence***
{ Seafood = N }	⇒	*{ Nursery = N }*	*0.8059921*	*0.9222987*
***Interpretación**: Si un supermercado o mercado sobre ruedas tiene un vivero entonces es probable que venda mariscos.*				
{ Nursery = N }	⇒	*{ Seafood = N }*	*0.8059921*	*0.8926969*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende mariscos entonces es probable que tenga un vivero.*				
{ Seafood = N }	⇒	*{ Wine = N }*	*0.8267436*	*0.9460447*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende vinos entonces es probable que venda mariscos.*				
{ Wine = N }	⇒	*{ Seafood = N }*	*0.8267436*	*0.8985720*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende mariscos entonces es probable que venda vinos.*				
{ Seafood = N }	⇒	*{ Nuts = N }*	*0.8332515*	*0.9534916*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende nueces entonces es probable que venda mariscos.*				
{ Nuts = N }	⇒	*{ Seafood = N }*	*0.8332515*	*0.8936002*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende mariscos entonces es probable que venda nueces.*				
{ Nursery = N }	⇒	*{ Wine = N }*	*0.8444253*	*0.9352645*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende vinos entonces es probable que tenga un vivero.*				
{ Wine = N }	⇒	*{ Nursery = N }*	*0.8444253*	*0.9177899*
***Interpretación**: Si un supermercado o mercado sobre ruedas tiene un vivero entonces es probable que venda vinos.*				
{ Nursery = N }	⇒	*{ Nuts = N }*	*0.8746316*	*0.9687203*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende nueces entonces es probable que tenga un vivero.*				
{ Nuts = N }	⇒	*{ Nursery = N }*	*0.8746316*	*0.9379774*
***Interpretación**: Si un supermercado o mercado sobre ruedas tiene un vivero entonces es probable que venda nueces.*				
{ Wine = N}	⇒	*{ Nuts = N }*	*0.8638261*	*0.9388763*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende nueces entonces es probable que venda vinos.*				
{ Nuts = N }	⇒	*{ Wine = N }*	*0.8638261*	*0.9263893*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende vinos entonces es probable que venda nueces.*				



{ Nursery = N, Wine = N } ⇒ *{ Nuts = N }* *0.8175344* *0.9681547*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende nueces entonces es probable que tenga un vivero o que venda vinos.*

{ Nursery = N, Nuts = N } ⇒ *{ Wine = N }* *0.8175344* *0.9347185*
***Interpretación**: Si un supermercado o mercado sobre ruedas vende vinos entonces es probable que tenga un vivero o que venda nueces.*

{ Nuts = N, Wine = N } ⇒ *{ Nursery = N }* *0.8175344* *0.9464108*
***Interpretación**: Si un supermercado o mercado sobre ruedas tiene un vivero entonces es probable que venda nueces o que venda vinos.*

Ahora procederemos a eliminar las reglas redundantes usando el siguiente criterio:

$$X_2 \Rightarrow Y_2 \text{ es redundante con respecto a } X_1 \Rightarrow Y_1 \text{ si y sólo si } X_1 \subseteq X_2 \text{ y } (X_2 \cup Y_2) \subseteq (X_1 \cup Y_1)$$

Bajo este criterio ninguna de las reglas de asociación anteriores es redundante.

Las reglas de asociación que nos resultan más interesantes y novedosas son las siguientes:

lhs		***rhs***	***support***	***confidence***
{ Wine = N }	⇒	*{ Seafood = N }*	*0.8267436*	*0.8985720*

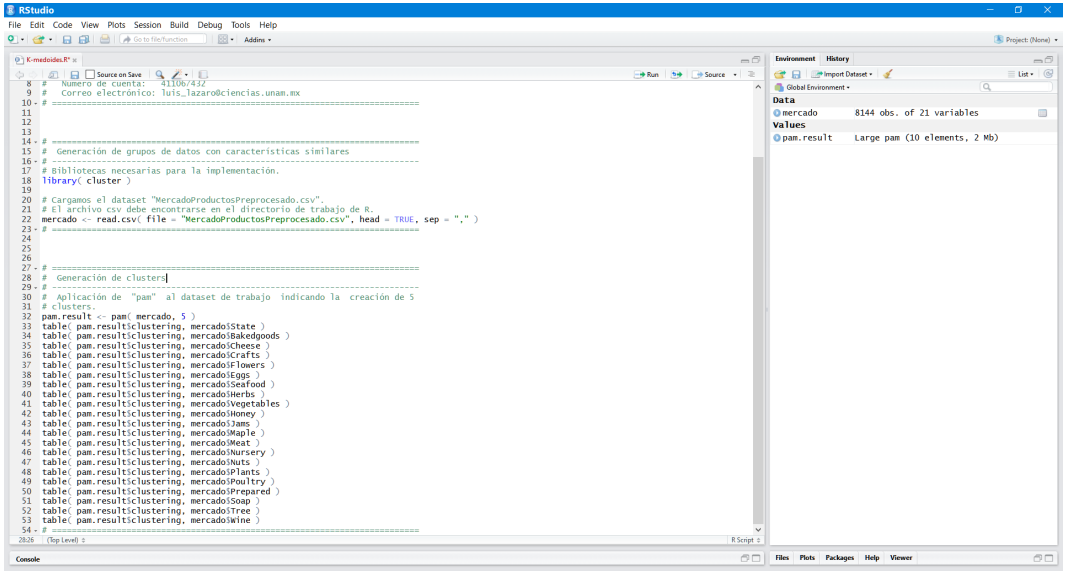
***Interpretación**: Si un supermercado o mercado sobre ruedas vende mariscos entonces es probable que venda vinos.*

{ Nuts = N, Wine = N } ⇒ *{ Nursery = N }* *0.8175344* *0.9464108*
***Interpretación**: Si un supermercado o mercado sobre ruedas tiene un vivero entonces es probable que venda nueces o que venda vinos.*

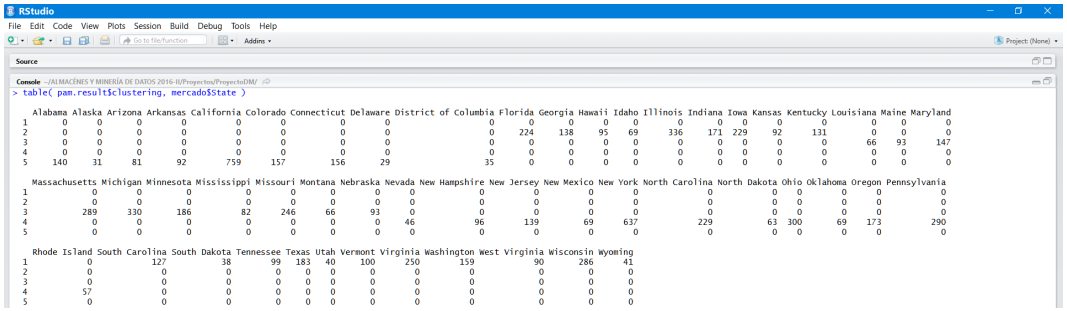
Nótese que estás reglas son ciertas en el *90%* de los casos.

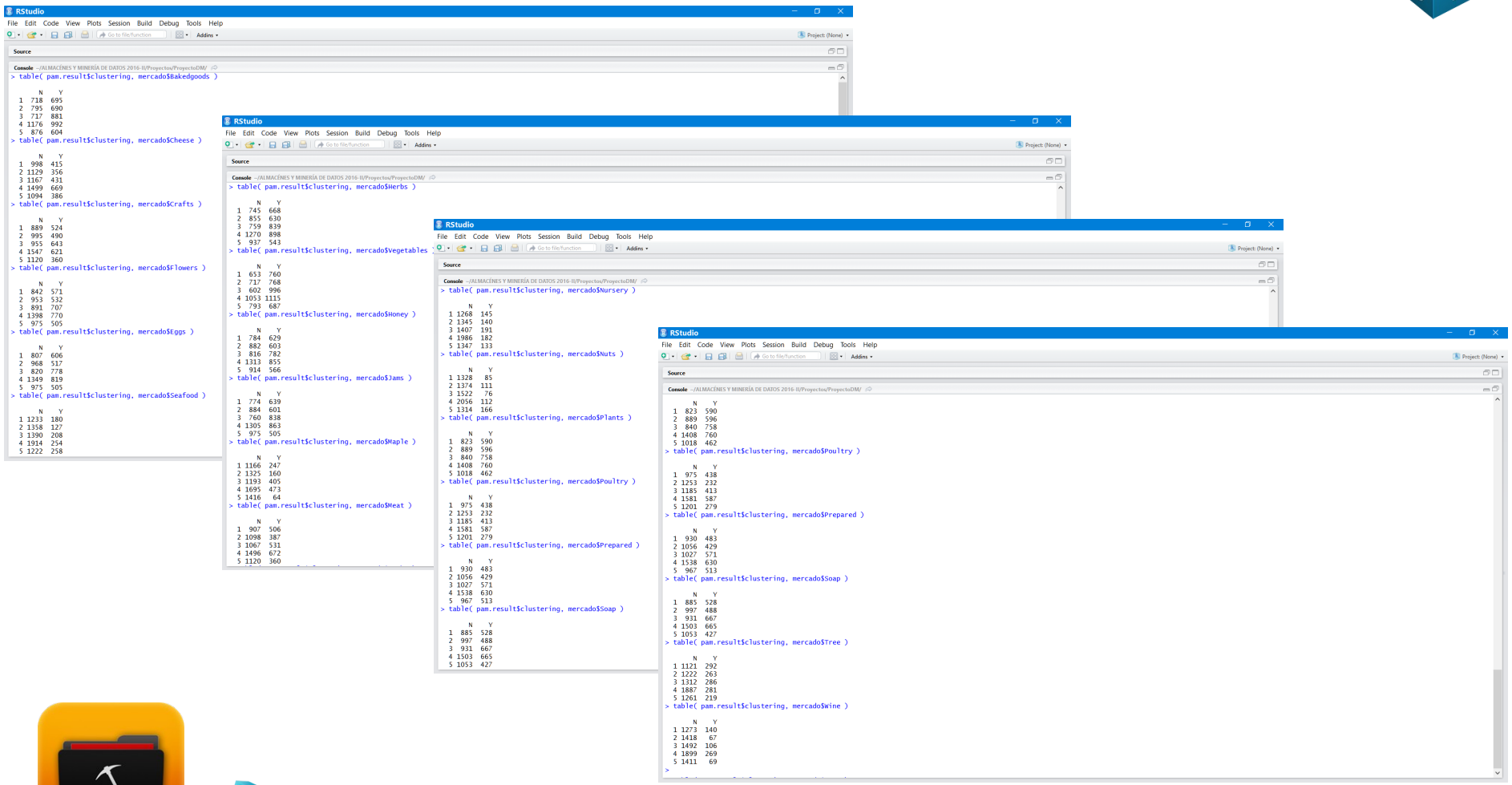
El segundo objetivo planteado es utilizar la información proporcionada por el *dataset* para generar un modelo que permita vizualizar la distribución regional de los productos agrícolas: **análisis de la geografía de las compras**. De esta descripción, es natural pensar en aplicar tareas de minería de datos descriptivas al *dataset* de trabajo para alcanzar este objetivo; de hecho, esto es lo que haremos. Para lograr esta tarea, aplicaremos al *dataset* de trabajo la técnica de k-medóides; pues éste método nos permitirá trabajar con un número relativamente pequeño de grupos de datos, cada uno de los cuales tiene un conjunto de característica singulares que proporciona patrones útiles, quizás interesantes e inesperados, que ayudan a formular conclusiones bastante directas.

Usaremos ***R*** para crear al menos 3 colecciones o grupos de datos con características similares dentro del *dataset* de trabajo, específicamente usaremos la biblioteca ***cluster*** y la función ***pam***. La siguiente imagen muestra el script creado en ***R*** para llevar a cabo esta labor:



La función ***pam*** se aplicó con un número de clusters igual a *5*, pues consideramos que este valor proporciona un conjunto de clusters manejable y bastante descriptivo. Los clusters con esta interpretación se muestran en las siguientes imágenes:.





La interpretación correspondiente a esa información, en terminos de grupos de estados y la mayor oferta de productos, es la siguiente:

Reporte elaborado por José Luis Vázquez Lázaro (411067432) / 01 de junio de 2016.

- ***Cluster 1***. Está conformado por el conjunto de estados *{ South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming }*, este bloque de estados tiene la mayor oferta de *Tree*.

- ***Cluster 2***. Está conformado por el conjunto de estados *{ Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky }*, este bloque de estados tiene la menor oferta de cualquier producto del campo registrado.

- ***Cluster 3***. Está conformado por el conjunto de estados *{ Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska }*, este bloque de estados tiene la mayor oferta de *Crafts, Nursery* y *Soap*.

- ***Cluster 4***. Está conformado por el conjunto de estados *{ Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island }*, este bloque de estados tiene la mayor oferta de *Bakedgoods, Cheese, Flowers, Eggs, Herbs, Vegetables, Honey, Jams, Maple, Meat, Plants, Poultry, Prepared* y *Wine*.

- ***Cluster 5***. Está conformado por el conjunto de estados *{ Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, District of Columbia }*, este bloque de estados tiene la mayor oferta de *Seafoods* y *Nuts*.

5. Evaluación.

El modelo basado en reglas de asociación arrojó un conjunto de 15 reglas útiles. Como se mencionó en el punto anterior, las reglas de asociación que nos resultaron más interesantes y novedosas son las siguientes:

lhs		***rhs***	***support***	***confidence***
{ Wine = N }	⇒	*{ Seafood = N }*	*0.8267436*	*0.8985720*

Interpretación: *Si un supermercado o mercado sobre ruedas vende mariscos entonces es altamente probable que venda vinos.*

{ Nuts = N, Wine = N }	⇒	*{ Nursery = N }*	*0.8175344*	*0.9464108*

Interpretación: *Si un supermercado o mercado sobre ruedas tiene un vivero entonces es altamente probable que venda nueces o que venda vinos.*

La interpretación de estas reglas, en el contexto de las ventas, proporciona una posible cesta de compras, pues es fácil traducirlas a recomendaciones de compras. Por ejemplo, la primera regla que puede traducirse en la siguiente recomendación: *sugerir a los clientes que compran mariscos, comprar vino, indicando la buena combinación que resulta de estos dos productos*. Nótese además, que estas reglas satisfacen el primer objetivo de la minería de datos.

Por otro lado, estas reglas son poco intuitivas, por lo que el conocimiento descubierto a través de este modelo es nuevo. Además, dado que estas son ciertas en al menos un *90%* de los casos, el conocimiento descubierto es bastante confiable. De esta manera, este modelo resulta bastante útil para alcanzar el primer objetivo del negocio.

El modelo basado en k-medóides genero 5 clusters con características bastante interesantes, pues inicialmente no se conocía la distribución regional de la oferta de los productos. Por lo que el conocimiento descubierto a través de este modelo es nuevo.

La interpretación del conocimiento proporcionado por este modelo indica que:

- El bloque formado por los estados: *South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin* y *Wyoming*, tiene acaparada la oferta de *árboles*.

- El bloque formado por los estados: *Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas* y *Kentucky* tiene la menor oferta de productos agrícolas de todos los bloques, por lo que sus ventas no son significativas.

- El bloque formado por los estados: *Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana* y *Nebraska*, tiene acaparada la oferta de *artesanías* y *jabones artesanales*, y tiene la mayor presencia de *viveros*.

- El bloque formado por los estados: *Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania* y *Rhode Island*, tiene acaparada la oferta de *productos horneados (galletas, tartas, etc.), quesos artesanales, flores, huevo fresco, hierbas frescas para el consumo humano, vegetales, miel artesanal, mermeladas artesanales, miel de maple, carne fresca, plantas, aves de corral, abono preparado (composta)* y *vinos*; de hecho, este bloque ofrece la mayor diversidad de productos agricolas.

- El bloque formado por los estados: *Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware* y *District of Columbia*, tiene acaparada la oferta de *mariscos* y *nueces*.

Nótese que estos cluster satisfacen el segundo objetivo de la minería de datos.

Este conocimiento es bastante útil, pues con este número de cluters se ha logrado crear una partición de los estados de Estados Unidos, bastante homogénea con respecto a la cardinalidad de cada bloque de estados, que refleja la distribución regional de la oferta de los productos agrícolas. Esta "homogeneidad" da confiabilidad al conocimiento encontrado, pues ningún cluster está cargado en relación a la cantidad de estados que lo conforman. De esta manera, este modelo resulta bastante útil para alcanzar el segundo objetivo del negocio.

6. Explotación de la utilidad de los modelos.

Del conocimiento encontrado a través del modelo basado en reglas de asociación, damos las siguientes sugerencias que pueden ayudar a alcanzar el primer objetivo del negocio (tener éxito):

- Dado que en el *90%* de los casos es cierto que: *si un supermercado o mercado sobre ruedas vende mariscos entonces es probable que venda vinos*, recomendamos que *aquellos supermercados o mercados sobre ruedas que vendan mariscos sugieran a sus clientes consumidores de mariscos, comprar vino, indicando la buena combinación que resulta de estos dos productos*.

- Dado que en el *90%* de los casos es cierto que: *si un supermercado o mercado sobre ruedas tiene un vivero entonces es probable que venda nueces o que venda vinos*, recomendamos *que aquellos supermercados o mercados sobre ruedas que tengan un vivero sugieran a sus clientes comprar nueces y/o vino*.

Del conocimiento encontrado a través del modelo basado k-medóides, damos las siguientes sugerencias que pueden ayudar a alcanzar el segundo objetivo del negocio (tener éxito):

- Dado que el bloque formado por los estados: *Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas* y *Kentucky* tiene la menor oferta de productos agrícolas, recomendamos aplicar los apoyos gubernamentales necesarios para aumentar la producción agricola en este conjunto de estados; esto permitirá que la oferta en el bloque aumente.

- Para el bloque formado por los estados: *South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin* y *Wyoming*, para el bloque formado por los estados: *Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana* y *Nebraska*, y para el bloque formado por los estados: *Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware* y *District of Columbia*, recomendamos aplicar los apoyos gubernamentales necesarios para diversificar su oferta de productos agricolas.

Experiencias adquiridas en el desarrollo del proyecto

La experiencia adquirida durante el desarollo del proyecto nos mostró que la columna vertebral de este recayó en la creación de un escenario ficticio adecuado para darle contexto; de hecho, fue fácil crear el escenario gracias a la simpleza en la descripción del *dataset* que se nos proporcionó para trabajar. Una vez creada las hipótesis de trabajo, la aplicación de la metodología ***CRISP-DM*** resultó más natural; pues no tuvimos que forzar cada parte de esta. Por otro lado, durante el proceso de desarrollo de esta metodología, nos dimos cuenta que esta no es modular, es decir, cada una de la etapas que la conforman dependen de las nteriores; siendo la primera la más importante y la más difícil de definir sino se ha establecido un contexto útil con respecto al material que se tiene. En nuestro caso, la creación de un escenario ficticio facilitó la definición de la primera etapa. Este es otro punto importante en la experiencia adquirida.