Instituto Profesional INACAP - Sede Renca

Informe de Análisis de Clientes y Visualización de Datos
par **Supermercado Lidl - Segmentación y Rentabilidad**

par Carrera: Ingeniería en Informática

Asignatura: Visualización de Datos

Integrantes:
Jaime Herrera
Benjamin Valenzuela
Paulo Brito

Profesor: Pablo Andrés Constancio Navarro

Fecha de Entrega: Octubre, 2025

1. Resumen Ejecutivo

El presente informe detalla el análisis de datos de clientes de un supermercado chileno, realizado con miras a su posible adquisición por parte de Lidl. Utilizando RStudio, se procesó una muestra de 500 clientes para identificar perfiles rentables, detectar comportamientos atípicos y segmentar la base mediante **K-Means Clustering**. Los resultados destacan que el cliente **Planificado** es el más rentable en términos de gasto promedio total. Además, se identificaron cuatro grupos de clientes, siendo el **Cluster VIP (Gasto Alto)** el foco estratégico principal para campañas de fidelización y valor.

2. Introducción

Este estudio tiene como objetivo proporcionar a la cadena de supermercados Lidl una visión analítica del comportamiento de compra de su futura base de clientes. Se aplicaron técnicas de **Análisis Exploratorio de Datos (EDA)**, **Detección de Outliers** y **Clustering** para transformar los datos crudos en inteligencia de negocio aplicable. Las tres preguntas centrales del estudio guían la metodología para determinar la rentabilidad, la consistencia de los patrones de compra y los segmentos de mercado existentes.

3. Metodología y Herramientas

El análisis se desarrolló íntegramente en el entorno **RStudio**, utilizando el lenguaje de programación R.

- **Herramientas:** RStudio versión 4.x, librerías `dplyr`, `ggplot2`, `cluster`, `factoextra`, `tidyr`.
- **Dataset:** `clientes_lidl.csv`, con 500 observaciones y 13 variables de comportamiento.
- **Técnicas:**
    1. EDA y `aggregate()` para cálculo de rentabilidad.

2. Método del Rango Intercuartil (**IQR**) para detección de outliers.
3. Algoritmo **K-Means** para segmentación de clientes.

4. Análisis Exploratorio de Datos (EDA) e Insights

El análisis preliminar reveló una base de clientes con una edad promedio de 36.11 años. Las comunas de **Santiago, Ñuñoa y Lo Barnechea** concentran la mayor parte de la muestra. En cuanto a la tipología, los clientes **Entusiastas** (148) y **Planificados** (140) dominan en cantidad.

4.1. Perfil de Rentabilidad (Pregunta 1)

El perfil más rentable se mide por el **Gasto Promedio Total de Compra**. La tabla 1 muestra que el cliente **Planificado** lidera este indicador.

Cuadro 1: Perfil de Rentabilidad por Tipo de Cliente

Tipo_Cliente	**Gasto_Promedio ($)**	**Frecuencia_Promedio**	**Nº Clientes**
Planificado	472.062	5.36	140
Entusiasta	457.226	5.58	148
Compulsivo	428.210	5.79	82
Organizado	420.499	5.46	130

El cliente Planificado gasta en promedio $14.836 más que el Entusiasta.

Insights Clave: El cliente **Planificado** es el más valioso en términos monetarios. Sin embargo, el **Entusiasta** y el **Compulsivo** presentan mayor **Ratio de Frecuencia**. La estrategia debe enfocarse en aumentar la frecuencia del Planificado o el gasto del Entusiasta, concentrando esfuerzos en comunas con alto poder adquisitivo donde el gasto promedio es consistentemente más alto (e.g., Lo Barnechea y Vitacura).

5. Detección de Outliers (Pregunta 2)

Se utilizó el método del **Rango Intercuartil (IQR)** para identificar valores atípicos en la variable `Promedio_Total_Compra`.

El umbral superior de detección fue de **$1.096.068**. Se identificaron **6 clientes** cuyo gasto promedio supera significativamente este valor.

Cuadro 2: Outliers detectados en Gasto Promedio (Top 3)

RUN	Tipo_Cliente	Comuna	Promedio_Total_Compra ($)
16598453	Entusiasta	La Reina	**1.478.940**
8636530	Planificado	Las Condes	1.394.120
14961030	Entusiasta	Lo Barnechea	1.382.570

Estos clientes representan la cúspide del gasto en la muestra.

El gráfico Boxplot (Figura 1) visualiza la alta dispersión y la presencia de estos valores.

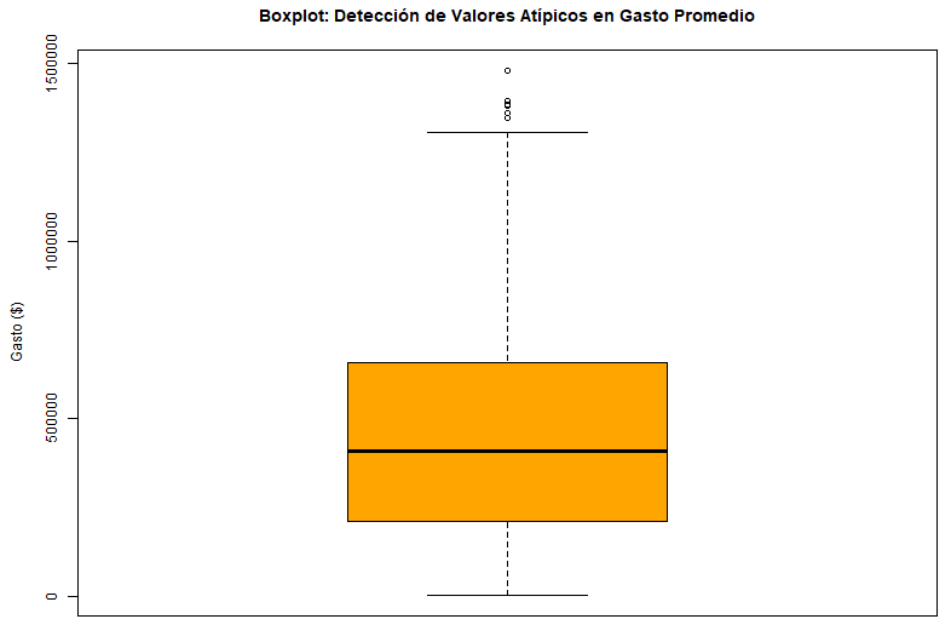

Figura 1: Detección de valores atípicos en Gasto Promedio (Método IQR).

Implicancias: Estos **Clientes VIP** provienen de comunas de altos ingresos y su comportamiento es clave para el volumen de ventas. Deben ser segmentados para re-

cibir tratamiento exclusivo (ej., atención personalizada, preventas, beneficios de lealtad elevados).

6. Clustering o segmentación (Pregunta 3)

El análisis de segmentación se realizó mediante el algoritmo **K-Means**, utilizando $k = 4$ clusters. Las variables empleadas fueron: `Numero_Compras`, `Promedio_Total_Compra`, `Ratio_Frecuencia` y `Puntos_Socio`. Los datos fueron previamente estandarizados con la función `scale()`.

6.1. Interpretación de Centroides

Los centroides, que representan el perfil promedio de cada cluster, permiten la siguiente caracterización:

Cuadro 3: Centroides (Valores Promedio) de los 4 Clusters

Cluster	Nº Compras	Gasto Promedio ($)	Frecuencia	Puntos Socio
1	292	386.921	5.13	2184
2	**770**	282.768	5.75	4522
gray!20 **3 (VIP)**	605	**879.383**	5.63	5366
4	314	365.266	5.56	**8095**

Perfiles Identificados:

- **Cluster 3 (VIP):** Clientes de **Alto Gasto Extremo** ($879k), con una frecuencia y número de compras altos. Es el segmento de mayor valor.
- **Cluster 2 (Frecuentes de Bajo Valor):** Se distingue por el **mayor número de compras** (770), pero el **menor gasto promedio** ($282k). Oportunidad para estrategias de venta cruzada (cross-selling).
- **Cluster 4 (Orientados a Puntos):** Clientes que compran poco (314) pero tienen la mayor cantidad de puntos de fidelidad (**8.095**). Muy sensibles a programas de

lealtad.

- **Cluster 1 (Compradores de Frecuencia Media):** Perfil equilibrado, con valores medios en todas las variables.

El gráfico 2 muestra la separación de estos grupos.

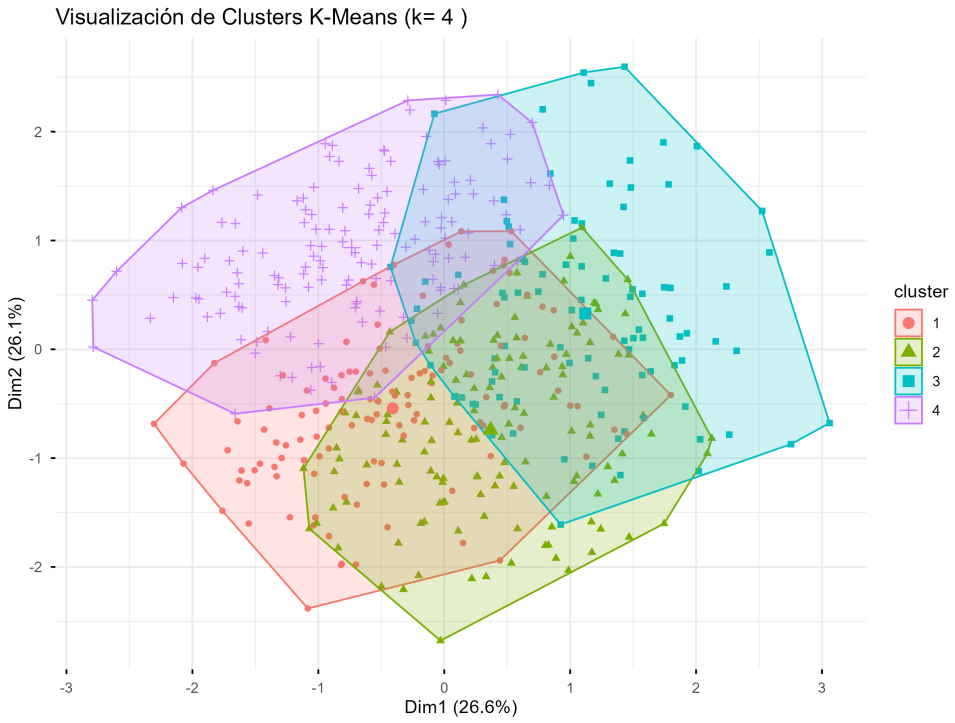

Figura 2: Visualización de clusters K-Means.

Ejemplo de código R utilizado para Clustering:

```
# Escalamiento
variables_cluster <- datos_lidl %>%
  select(Numero_Compras, Promedio_Total_Compra,
         Ratio_Frecuencia, Puntos_Socio)
datos_scaled <- scale(variables_cluster)


# Aplicación de K-Means
set.seed(42)
k.final <- kmeans(datos_scaled, centers = 4, nstart = 25)
```

```
# Visualización
fviz_cluster(k.final, data = datos_scaled)
```

7. Principales Insights y Conclusiones

El análisis proporciona una hoja de ruta clara para la estrategia de Lidl en Chile:

- **Grupos de Clientes Identificados:** Se definieron cuatro segmentos, siendo el **Cluster 3 (VIP)** el segmento de mayor valor. El **Cluster 2 (Frecuentes de Bajo Valor)** es la principal oportunidad para aumentar los ingresos por transacción.
- **Rentabilidad por Perfil:** Los clientes **Planificados** son los que generan el mayor gasto promedio, siendo el perfil más rentable a nivel de transacción, superando ligeramente al Entusiasta.
- **Variables Más Influyentes:** La variable **Promedio_Total_Compra** es el factor más determinante en la identificación de segmentos VIP (Cluster 3), mientras que **Numero_Compras** define la alta actividad del Cluster 2.
- **Recomendaciones Estratégicas:**
  1. **Fidelización VIP:** Crear un programa de lealtad de nivel "Platinum.enfocado en el Cluster 3, basado en experiencias exclusivas y no solo en descuentos.
  2. **Maximización del Gasto:** Diseñar estrategias de venta cruzada (cross-selling) dirigidas al Cluster 2 (Frecuentes de Bajo Valor) para aumentar el tamaño de su canasta promedio sin afectar su alta frecuencia.
  3. **Enfoque Geográfico:** Priorizar campañas de marketing y expansión en comunas de alto gasto promedio como Lo Barnechea, Vitacura y Providencia, donde residen los clientes de mayor valor.