UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

Tesis para optar al Título de Licenciado en Economía.

¿Cómo consumen los uruguayos en el supermercado?

Ramiro Almada
Pablo Moya
Marcos Rivero

Tutor: Juan Dubra

Montevideo, Uruguay
Diciembre 2010

PÁGINA DE APROBACIÓN

FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

El tribunal docente integrado por los abajo firmantes aprueba la Tesis de Investigación:

Título

¿Cómo consumen los uruguayos en el supermercado?

Autor/es

Ramiro Almada

Pablo Moya

Marcos Rivero

Tutor

Juan Dubra

Carrera

Licenciado en Economía.

Puntaje.....

Tribunal

Profesor.....(Nombre y firma)

Profesor.....(Nombre y firma)

Profesor.....(Nombre y firma)

Fecha……..

AGRADECIMIENTOS

A Juan Dubra por su dedicación.
A Daniel Gramoso por su invalorable ayuda.
A Gabriel Camaño, Inés Urrestarazú, Henry Willebald y Leandro Zipitría por sus comentarios.
A nuestras familias por el constante apoyo que nos dieron.

TABLA DE CONTENIDO

RESUMEN

El trabajo identifica patrones de comportamiento de los consumidores de una cadena de supermercados de Montevideo y el este del país. Para ello se analizaron las interrelaciones entre las diferentes categorías de productos de consumo doméstico utilizando datos de ventas para el semestre julio-diciembre de 2004.

Para el análisis de la información se definieron variables que describen el consumo en términos de frecuencia, gasto, peso relativo y distribución en el mes de las compras de cada categoría de bienes y se utilizaron técnicas de análisis multivariado: componentes principales y conglomerados (*clusters*).

Estos procedimientos permitieron identificar cuatro grupos o canastas de productos que poseen características homogéneas a su interior y bien diferenciadas entre ellas: canasta básica, bienes perecederos, fiestas y carne vacuna.

Las diferencias entre las canastas están determinadas, principalmente, por la cantidad y monto promedio de los *tickets* y por el gasto total en los productos de cada conglomerado.

Respecto a las preferencias intertemporales no se encontró evidencia de una tendencia decreciente del gasto a lo largo del mes.

Palabras clave:

clusters o conglomerados, componentes principales, patrones de consumo, canasta de bienes, supermercados

1. Introducción

El presente trabajo identifica patrones de comportamiento de los consumidores de una cadena de supermercados de Montevideo y el este del país[1]. Para ello se analizaron las interrelaciones entre las diferentes categorías de productos de consumo doméstico utilizando datos para el semestre julio-diciembre de 2004.

Las ventas de los supermercados en 2004 representaban el 32% del total de ventas en comercios minoristas según un estudio de la consultora internacional AC Nielsen, siendo la participación de los comercios tradicionales el 39% y el de los autoservicios el 29%.

El consumo es uno de los determinantes más importante de la actividad económica. Como variable macroeconómica es un componente fundamental de la demanda agregada por su importancia cualitativa y por su incidencia sobre el resto de las variables. En 2009 el consumo privado representó el 69,2% del Producto Bruto Interno (PBI) del país y el 52,9% de la demanda agregada, según la información de Cuentas Nacionales del Banco Central del Uruguay (BCU).

Por otro lado, en su dimensión microeconómica el consumo es el soporte básico de la demanda de bienes y servicios.

El comportamiento de los individuos frente a las decisiones de consumo ha sido extensamente estudiado en economía. Desde los pioneros análisis de Keynes (1936) acerca del consumo como una proporción estable del ingreso (propensión a consumir), la teoría de la Hipótesis de Ciclo de Vida de Modigliani y Brumberg (1954), o la teoría del Ingreso Permanente de Friedman (1957), el estudio del gasto de los consumidores y su relación respecto a otras variables como el ingreso, otros bienes, etc., ha sido un tema central de la teoría económica.

[1] Este trabajo analiza únicamente el gasto en la cadena de supermercados Disco.

La teoría del consumidor en su versión más tradicional considera que las decisiones de consumo de los individuos dependen del ingreso, los precios y las preferencias individuales, y que los individuos toman decisiones con el objetivo de maximizar su utilidad individual.

Stigler y Becker (1977) formulan una "nueva" teoría de la elección del consumidor, en la cual la unidad doméstica (o el agente individual) se involucra en la maximización de la utilidad de los bienes que compra a través de su transformación en productos de consumo. A partir de esta teoría se puede analizar comportamientos que a priori se entenderían como consecuencia de cambios en las preferencias. Estos comportamientos corresponden a los que son consecuencia de los hábitos, la publicidad y las modas, entre otros.

Como antecedentes más recientes, Durán y Souto (2009) encontraron evidencia para Uruguay de un efecto positivo del momento de cobro del ingreso sobre el consumo, estudiando la misma base de datos que en este trabajo y desarrollando un modelo econométrico que sigue las especificaciones de Tobin (1956) para una canasta particular de bienes.

El presente trabajo se plantea dos objetivos para el estudio de la conducta de los consumidores: i) analizar la existencia de un patrón de consumo a lo largo del mes, en donde los individuos suavicen su consumo para maximizar su utilidad, y ii) caracterizar el gasto de los consumidores en función de las similitudes o disimilitudes de las compras tratando de detectar la existencia de reglas de asociación entre las compras de los distintos productos.

El análisis de la tendencia del gasto en el tiempo es relevante, pues si se encontrara un patrón decreciente, permitiría descartar los modelos más tradicionales de la teoría del consumo, a favor de modelos de corte más "irracional" (ver Laibson (1998) y su hipótesis del descuento hiperbólico, y Durán y Souto (2009) y sus referencias para un marco teórico sobre esta pregunta).

Respecto a las preferencias intertemporales se concluyó que de las cuatro canastas de bienes definidas, una presenta una tendencia creciente y en las otras tres los resultados no son concluyentes (gráficamente se observa que dos son decrecientes, aunque la estimación por mínimos cuadrados ordinarios (MCO) no fue significativa).

En cuanto al segundo objetivo, del análisis de la base de datos permite concluir que el consumo de los uruguayos en los supermercados se divide en cuatro grupos de productos denominados *clusters*, conformados por categorías de bienes similares al interior de cada uno y claramente diferenciadas de las categorías de los demás. Las diferencias entre ellos surgen principalmente de las variables “Número de *ticket* que incluyen productos de la categoría j ”, “Gasto total de productos de la categoría j”, “Gasto total en productos del supermercado que efectúan los compradores de productos de la categoría j” y “Monto promedio de los *tickets* que incluyen productos de la categoría j”.

El desarrollo del trabajo prosigue con la descripción de la base de datos utilizada en el capítulo 2.

En el capítulo 3 se desarrolla la metodología empleada y se presentan los resultados. Incluye la definición de las variables que describen el consumo en términos de frecuencia, gasto, peso relativo y distribución en el mes de compras de cada categoría de bienes, la aplicación de técnicas de análisis multivariado como componentes principales y conglomerados o *clusters* y la caracterización de los *clusters* resultantes.

En el capítulo 4 se presentan las conclusiones del trabajo.

2. Descripción de los Datos

2.1. Caracterización de la base de datos

La información utilizada corresponde a una base de datos suministrada por la cadena de supermercados Disco que incluye la totalidad de las compras realizadas por los clientes asociados al programa de comprador frecuente (tarjeta Más), en todos los locales de la cadena durante el período comprendido entre el 1° de Julio y el 31 de Diciembre de 2004 (184 días).

La base incluye la totalidad de las compras realizadas por los compradores frecuentes que se hayan identificado como tales mediante la presentación de la referida tarjeta.

La tarjeta genera beneficios para el usuario sin ningún costo asociado (salvo el completar un formulario por única vez), por lo cual está incentivado a presentarla cada vez que hace una compra. Las cajeras siempre preguntan al cliente si posee la tarjeta. Respecto a la representatividad de la muestra, Duran y Souto señalan: *"Dado que los beneficios son proporcionales al gasto que realiza, existe la posibilidad de que la muestra posea un sesgo si el individuo se olvida de su tarjeta con mayor probabilidad cuando va a realizar compras de pequeño monto"*. A ello hay que agregar que eso no distorsionaría nuestro análisis a menos que las compras pequeñas tengan algún sesgo particular.

Cada registro de la base incluye el número de tarjeta, la fecha, el producto comprado, la cantidad, el gasto y el local de la compra. La base de datos consta de 412.794 registros correspondientes a más de 63.025 compras (*tickets*) realizadas por 7.297 compradores frecuentes diferentes. Las compras incluyen 14.602 productos distintos.

La base de datos contiene toda la información recogida en los locales de la cadena de supermercados, los cuales se distribuyen en Montevideo (16 locales), Canelones (cuatro locales) y Maldonado (cuatro locales).

A continuación se detalla la ubicación de los diferentes locales:

Cuadro 1: Locales de la cadena Disco

Local	Barrio / Zona	Departamento
GEANT	Ciudad de la Costa	Canelones
Súper-Uno	P. del Este	Maldonado
Disco 23 - Médanos de Solymar	Médanos de Solymar	Canelones
Disco 22 - Barrios Amorín 859	Palermo	Montevideo
Disco 21 – Centro	Centro	Montevideo
Disco 20 - Parque Batlle	Parque Batlle	Montevideo
Disco 19 – Atlántida	Atlántida	Canelones
Disco 18 – Solymar	Solymar	Canelones
Disco 17 - Maldonado 1024	Centro	Montevideo
Disco 16 - Ayacucho 3370	Villa Dolores	Montevideo
Disco 15 - Sucursal Goes	Goes	Montevideo
Disco 14 - Maldonado - P. del Este	P. del Este	Maldonado
Disco 13 - Chucarro 1320	Pocitos	Montevideo
Disco 12 - Av. 8 de Octubre 4786	Curva de Maroñas	Montevideo
Disco 11 - La Blanqueada-*Fresh Market*	La Blanqueada	Montevideo
Disco 09 - P. Carretas Shopping-Fresh Market	P. Carretas	Montevideo
Disco 08 - Maldonado-*Fresh Market*	Maldonado	Maldonado
Disco 07 - Dr. Francisco Soca 1318	Pocitos	Montevideo
Disco 06 – Centro	Centro	Montevideo
Disco 05 - Maldonado - Pta. del Este	P. del Este	Maldonado
Disco 04 - Av. Legrand	Malvín	Montevideo
Disco 03 – Cordón	Cordón	Montevideo
Disco 02 - Agraciada 2986-*Fresh Market*	Bella Vista	Montevideo
Disco 01 – Pocitos	Pocitos	Montevideo

La distribución de la cantidad de *tickets* por departamento es:

Cuadro 2: Distribución de locales por área geográfica

Montevideo	74%
Canelones	18%
Maldonado	8%
Total	**100%**

2.2. Modelización

Inicialmente se agruparon todos los bienes incluidos en los *tickets* de la base de datos según la clasificación que realiza el Instituto Nacional de Estadística (INE) para relevar el Índice de Precios al Consumo (Base 1997) (Ver cuadro 3).

Seguidamente se eliminaron de la base de datos los bienes de menor frecuencia de compra. La exclusión se fundamentó en que estos bienes no serían incluidos en la conformación de una canasta de consumo promedio de bienes de uso frecuente y perecederos (asimilable a una "canasta de consumo tipo") tratando de evitar la distorsión de un análisis de comportamiento.
Los bienes excluidos totalizan 4.939 (de un total de 14.164), están incluidos en el 11,5% del total de *tickets* de la base de datos y representan el 17,3% del gasto, por lo que las conclusiones a las que arribamos no se verán distorsionados por el tratamiento realizado.

En el cuadro 3 se detallan los Rubros, Agrupaciones y Subrubros de bienes incluidos y excluidos del análisis.

Cuadro 3: Agrupaciones de bienes incluidos y excluidos del análisis

Rubros, Agrupaciones y Subrubros del IPC utilizados	Rubros, Agrupaciones y Subrubros del IPC que se eliminaron
Aceites y grasas	Aparatos de audio, video, TV
Azucar, cafe, te, yerba, cacao	Artefactos y electrodomésticos
Bebidas alcohólicas	Combustible
Bebidas no alcohólicas	Cuid.médicos y conserv.de la salud
Carnes y derivados	LIBROS Y MATERIALES DE ENSEÑANZA
Comidas elaboradas	LIBROS, PERIODICOS Y REVISTAS
Comidas semielaboradas	MUEBLES, ACC. FIJOS Y REPARACIONES
Cristalería, vajilla, utensillos	Otros alimentos
CUIDADOS Y EFECTOS PERSONALES	Otros artíc. recreativos no duraderos
Frutas	Servicios y materiales p/reparación
Lácteos y huevos	TEJIDOS PARA EL HOGAR Y OTROS ACC.
Panes y cereales	Transporte y comunicaciones
TABACO	Vestimenta y calzado
Verduras, legumbres y tubérculos	

A partir de los rubros, agrupaciones y subrubros seleccionados, que incluían 9.225 productos, se definieron 55 categorías de productos que se detallan en el cuadro 4.

Cuadro 4: Categorías de productos definidas.

Nº	RubCort	Cantidad de Tarjetas MAS	Cantidad de Tickets	GastoTotal	Peso relativo	Peso relativo acumulado
1	Cvacuna	3,765	14,731	1,116,379	11.5%	11.5%
2	Bazar	4,401	13,791	801,252	8.2%	19.7%
3	Verduras	4,892	23,731	755,541	7.8%	27.4%
4	CuidPers	4,020	10,986	578,629	5.9%	33.4%
5	Fiambres	3,968	13,409	487,075	5.0%	38.4%
6	Frutas	3,849	15,536	403,656	4.1%	42.5%
7	Queso	3,552	11,112	401,018	4.1%	46.6%
8	Leche	4,292	23,442	372,595	3.8%	50.5%
9	Panes	4,060	20,517	366,380	3.8%	54.2%
10	ComElab	2,402	6,070	361,997	3.7%	57.9%
11	Gaseosa	2,430	7,087	336,908	3.5%	61.4%
12	Pollo	1,737	4,018	334,371	3.4%	64.8%
13	Galletitas	3,467	9,937	285,541	2.9%	67.7%
14	Postres	2,674	7,630	207,038	2.1%	69.9%
15	Agua	2,114	7,764	195,504	2.0%	71.9%
16	Aceite	2,124	4,131	181,941	1.9%	73.7%
17	Vino	927	2,267	169,514	1.7%	75.5%
18	Yogur	2,105	5,856	169,073	1.7%	77.2%
19	Wsky	337	791	165,901	1.7%	78.9%
20	Pescado	1,556	3,070	152,202	1.6%	80.5%
21	Tabaco	1,006	3,998	145,255	1.5%	82.0%
22	Pasta	2,230	4,707	132,841	1.4%	83.3%
23	Yerba	1,897	3,645	117,187	1.2%	84.5%
24	Café	1,554	2,927	115,762	1.2%	85.7%
25	Harinas	2,274	4,940	100,064	1.0%	86.8%
26	ComSemiEl	1,214	2,301	97,346	1.0%	87.8%
27	Azucar	2,237	4,483	86,306	0.9%	88.6%
29	Helados	716	1,088	85,964	0.9%	89.5%
29	Huevos	1,621	3,839	84,257	0.9%	90.4%
30	Arroz	1,928	3,697	82,535	0.8%	91.2%
31	Cerveza	684	1,487	80,477	0.8%	92.1%
32	Cerdo	509	855	76,138	0.8%	92.8%
33	Manteca	1,596	3,299	61,281	0.6%	93.5%
34	Te	1,257	2,117	60,015	0.6%	94.1%
35	Hambur	721	1,397	59,449	0.6%	94.7%
36	Otpanes	1,319	2,337	52,575	0.5%	95.2%
37	Achuras	701	1,319	47,016	0.5%	95.7%
38	OtBebAlc	394	556	46,964	0.5%	96.2%
39	Oliva	265	372	45,626	0.5%	96.7%
40	JugosNat	686	1,333	41,563	0.4%	97.1%
41	JugosInst	1,017	2,609	40,775	0.4%	97.5%
42	OtLacteos	791	1,337	40,096	0.4%	97.9%
43	Cereales	532	873	30,377	0.3%	98.2%
44	PostresLact	605	1,266	28,678	0.3%	98.5%
45	CacaoChoc	695	1,062	28,448	0.3%	98.8%
46	Edulcor	303	468	26,645	0.3%	99.1%
47	LecheSabor	431	942	22,550	0.2%	99.3%
48	Margarina	495	761	21,571	0.2%	99.6%
49	LecheLV	274	734	20,702	0.2%	99.8%
50	Covina	98	110	12,208	0.1%	99.9%
51	OtBebNoAlc	105	173	3,622	0.0%	99.9%
52	LechePolvo	49	67	3,200	0.0%	100.0%
53	Grasa	102	118	2,352	0.0%	100.0%
54	OtCarnes	29	45	1,424	0.0%	100.0%
55	OtCafe	2	2	162	0.0%	100.0%
	Total		**267,140**	**9,743,945**	**100.0%**	

De estas cincuenta y cinco categorías se eliminaron las seis de menor peso relativo, tanto en cantidad de *tickets* como en gasto total (en conjunto representaban el 0,19% de los *tickets* y el 0,24% del gasto total).

Al realizar el análisis de componentes principales y luego de determinar los distintos *cluster*, se obtuvo uno de ellos integrado únicamente por Aceite de Oliva. Este producto se caracteriza por representar una proporción muy baja de las ventas totales del supermercado y que está presente en un número mínimo de *tickets*. Sin embargo, los *tickets* que incluyen Aceite de Oliva tienen un monto promedio alto (máximo de todas las categorías) e incluyen un gran número de categorías diferentes. En consecuencia, se eliminó de la base de datos original los datos correspondientes a la categoría Aceite de oliva y se volvieron a calcular los componentes principales y los *clusters*.

Finalmente, llegamos a una serie de cuarenta y ocho categorías que conforman la canasta de consumo promedio o más frecuente de la cadena de supermercados que agrupa 9.181 productos.

2.3. Definición de variables

Las variables definidas para el presente trabajo fueron, para cada categoría j:

CanTick: Número de *tickets* que incluyen productos de j
GsTotal: Gasto total de productos de j
GsToTik: Gasto total en productos del supermercado que efectúan los compradores de productos de j
MonTick: Monto promedio de los *tickets* que incluyen productos de j (GsToTik/CanTick)
PorCatg: Razón entre el gasto de j y el gasto total de los compradores de productos de j (GsTotal/GsToTik)
CategDf: Número promedio de categorías diferentes incluidas en los *tickets* de compradores de productos de j

Perce10: Día del mes en que se acumula 10% del gasto mensual total en productos de j
Perce50: Día del mes en que se acumula 50% del gasto mensual total en productos de j
Perce90: Día del mes en que se acumula 90% del gasto mensual total en productos de j

Básicamente son solo siete variables, ya que las variables MonTick y PorCatg son, por definición, combinaciones de otras variables. Igualmente fueron incluidas pues facilitan el análisis descriptivo de las categorías y los *clusters*.

No solo se consideró el consumo (gasto) en los distintos bienes sino también las características del gasto en términos de frecuencia y momento del mes, así como la asociación con otras categorías al momento de la compra.

Las primeras cinco variables definidas anteriormente permiten realizar una descripción cuantitativa de las características de consumo de la base de datos.

La inclusión de la variable CategDf. permite la búsqueda de asociaciones de compras (o patrones) entre los distintos artículos.

Las últimas tres variables incorporan la dimensión temporal al análisis, lo cual permitirá estudiar si el consumo de algunos bienes se ve influenciado por el día en que se realiza la compra.

Los valores de estas variables para cada una de las cuarenta y ocho categorías se detallan en el cuadro 5.

Cuadro 5: Valores de las variables para cada categoría.

Nº	Categoría	CanTick	GsTotal	GsToTik	MonTick	PorCatg	CategDf	Perce10	Perce50	Perce90
1	AceitCom	4,131	181,941	1,774,128	429	0.10	8.77	4	15	28
2	Achuras	1,319	47,016	409,573	311	0.11	6.71	3	16	29
3	Agua	7,764	195,504	2,225,583	287	0.09	7.29	4	17	28
4	Arroz	3,697	82,535	1,425,883	386	0.06	8.59	4	15	28
5	Azucar	4,483	86,306	1,712,309	382	0.05	8.50	4	15	28
6	Cacao	1,062	28,448	480,162	452	0.06	9.61	3	15	28
7	Café	2,927	115,762	1,216,342	416	0.10	8.73	3	14	28
8	Cvacuna	14,731	1,116,379	4,028,545	273	0.28	6.36	4	16	28
9	Cerdo	855	76,138	389,188	455	0.20	8.21	4	18	30
10	Cereales	873	30,377	431,654	494	0.07	9.28	3	16	28
11	Cerveza	1,487	80,477	576,490	388	0.14	7.61	4	19	29
12	ComElab	6,070	361,997	2,114,586	348	0.17	7.24	4	17	29
13	ComSemElab	2,301	97,346	1,076,484	468	0.09	9.12	4	17	28
14	Cristal	13,791	801,252	4,692,582	340	0.17	7.22	4	16	28
15	CuiEfePers	10,986	578,629	3,979,313	362	0.15	7.61	4	15	28
16	Edulcor	468	26,645	254,528	544	0.10	9.09	4	15	28
17	Fiambres	13,409	487,075	4,088,836	305	0.12	7.34	4	16	28
18	Frutas	15,536	403,656	4,213,255	271	0.10	6.95	4	17	28
19	Galletitas	9,937	285,541	3,156,462	318	0.09	7.44	4	16	28
20	Gaseosas	7,087	336,908	2,481,556	350	0.14	7.51	4	18	29
21	Hambur	1,397	59,449	449,478	322	0.13	7.87	4	17	29
22	Harinas	4,940	100,064	1,751,557	355	0.06	8.28	4	15	28
23	Helados	1,088	85,964	528,606	486	0.16	7.98	4	19	29
24	Huevo	3,839	84,257	1,498,846	390	0.06	9.00	4	16	28
25	JugInstant	2,609	40,775	763,412	293	0.05	7.86	3	17	28
26	JugNatur	1,333	41,563	609,446	457	0.07	8.65	4	17	28
27	Leche	23,442	372,595	4,970,305	212	0.07	6.13	4	16	28
28	LecheLV	734	20,702	368,038	501	0.06	8.99	4	17	29
29	LecheSabor	942	22,550	348,208	370	0.06	8.41	4	17	28
30	Manteca	3,299	61,281	1,171,654	355	0.05	8.55	4	16	28
31	Margarina	761	21,571	375,161	493	0.06	9.41	4	15	29
32	Postres	7,630	207,038	2,061,756	270	0.10	6.54	4	17	28
33	OtBebAlc	556	46,964	294,368	529	0.16	8.05	5	20	29
34	OtrosLact	1,337	40,096	564,723	422	0.07	8.66	4	16	28
35	OtrosPanes	2,337	52,575	973,112	416	0.05	8.80	4	15	28
36	Panes	20,517	366,380	4,876,324	238	0.08	6.41	4	16	28
37	Pasta	4,707	132,841	1,660,984	353	0.08	8.16	4	15	28
38	Pescado	3,070	152,202	1,282,627	418	0.12	8.07	4	17	28
39	Pollo	4,018	334,371	1,490,948	371	0.22	7.39	3	14	28
40	PostresLact	1,266	28,678	450,938	356	0.06	8.57	4	17	29
41	Queso	11,112	401,018	3,668,561	330	0.11	7.87	4	16	28
42	Tabaco	3,998	145,255	804,813	201	0.18	5.26	4	16	28
43	Te	2,117	60,015	912,114	431	0.07	8.94	4	15	28
44	VerdLegum	23,731	755,541	6,206,403	262	0.12	6.64	4	16	28
45	Vino	2,267	169,514	923,976	408	0.18	7.52	4	17	29
46	Whisky	791	165,901	427,362	540	0.39	7.41	4	18	29
47	Yerba	3,645	117,187	1,268,621	348	0.09	7.66	4	15	28
48	Yogur	5,856	169,073	1,897,400	324	0.09	7.58	3	16	28
	Suma	**266,253**	**9,675,352**	**83,327,203**	**18,031**	**5.39**	**379.85**	**186**	**781**	**1358**
	Máximo	**23,731**	**1,116,379**	**6,206,403**	**544**	**0.39**	**9.61**	**5**	**20**	**30**
	Promedio	**5,547**	**201,570**	**1,735,983**	**376**	**0.11**	**7.91**	**4**	**16**	**28**
	Mínimo	**468**	**20,702**	**254,528**	**201**	**0.05**	**5.26**	**3**	**14**	**28**
		6,013	**230,262**	**1,533,110**	**85**	**0.07**	**0.95**	**0.4**	**1.3**	**0.5**

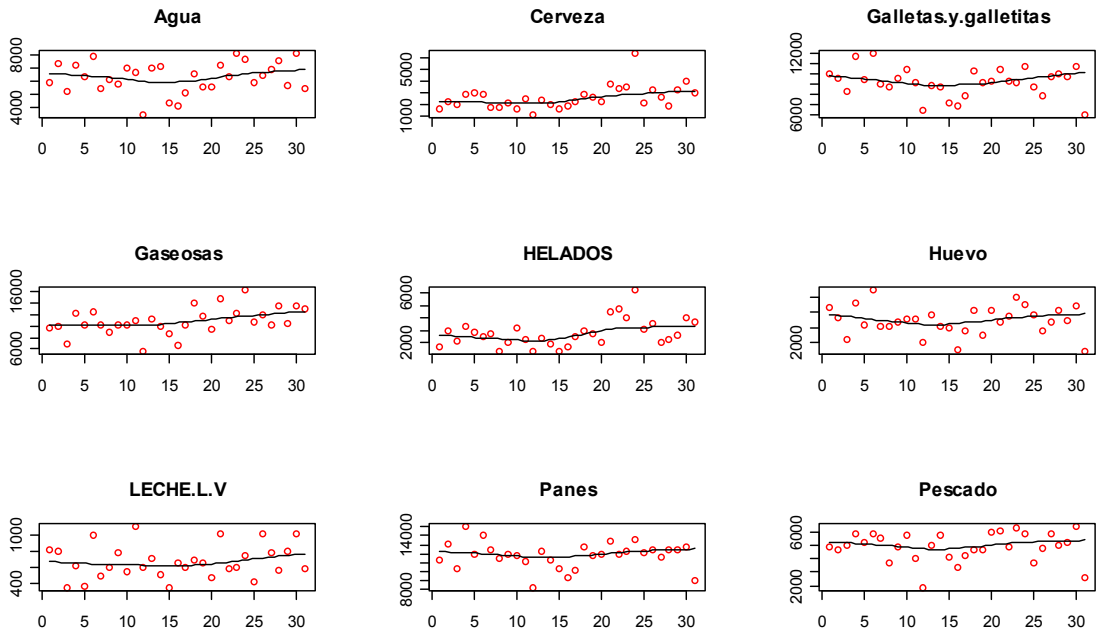
3. Metodología y resultados

3.1. Existencia de un patrón de consumo a lo largo del mes

Con el fin de analizar el primer objetivo del trabajo - la existencia de un patrón de consumo mensual- se regresó el gasto total por día de cada una de las categorías seleccionadas contra el tiempo. Para ello se utilizó una regresión no lineal Loess[2].

En consecuencia, nuestro punto de partida fue un análisis gráfico de las relaciones entre el gasto total y el tiempo para las 48 categorías de productos o bienes.

Grafico 1: Gasto total en el mes



Como ilustración, el gráfico 1 presenta el comportamiento de algunas categorías de bienes relevantes que muestran una leve tendencia creciente.

[2] Local Linear Regression; Cleveland (1979) y Cleveland y Devlin (1988).

3.1.1. El método Loess

El método Loess (o Lowess) es uno de los muchos métodos modernos de modelización que se basan en los métodos "clásicos", tales como las regresiones lineales y no lineales por mínimos cuadrados.

Para un caso general $y = m(x) + \varepsilon$, con $E(y/x) = m(x)$, la estimación no paramétrica m se obtiene mediante técnicas de suavizado aplicadas localmente a los pares de observaciones (xi,yi), *con* $i=1,2,....,n$. La estimación local se hace por MCO ponderados.

El procedimiento es similar al utilizado en la estimación de funciones de densidad univariada: el valor medio condicional para un intervalo pequeño de x se estima, no sólo con las observaciones de dicho intervalo, sino con las de intervalos adyacentes. Esta información se pondera en forma decreciente a medida que es mayor la distancia de la observación respecto al centro del intervalo.

Para una observación puntual X_0 consideraremos no sólo esa observación sino también las adyacentes. Las observaciones X_i demasiado alejadas de X_0 reciben una ponderación $w(X_i) = 0$, ya que dichas observaciones no suministran mayor información sobre $m(X_0)$. Observaciones X_i para las cuales la diferencia $X_0 - X_i$ es menor que cierto umbral prefijado, se incluyen en la regresión local ponderadas según la proximidad de los valores al punto explicado X_0.

Una función de ponderación tradicionalmente utilizada para Loess es la tricúbica[3]:

$$w(x) = \begin{cases} (1 - |x|^3)^3 & para\ |x| < 1 \\ 0 & para\ |x| \geq 1 \end{cases}$$

Donde $|x| = \frac{|Xi - Xo|}{Max|Xj - Xo|}$ para todo j que pertenece al intervalo local.

[3] Se utilizó el paquete estadístico R, un software libre disponible en www.r-project.org La función de ponderación tricúbica es la utilizada por dicho programa.

El uso de la ponderación se basa en la idea de que los puntos cercanos entre sí en el espacio de las variables explicativas son más propensos a estar relacionados entre sí, que de los puntos más separados.

El objetivo principal de esta parte del estudio era estudiar la robustez del análisis de Duran y Souto (2009), que trabajaron con una sola canasta. Nuestro análisis muestra que su resultado es sensible a la especificación de la misma.

3.2. Análisis de Componentes Principales

En esta sección realizamos un análisis de la base utilizando la metodología de componentes principales (ver Rencher, A. (2002): *Methods of Multivariate Analysis,* Capítulo 12*)* para obtener un conjunto menor de variables correlacionadas con las iniciales e incorrelacionadas entre si.

Para evitar el problema de medir las varianzas en diferentes unidades de medida, y que unas variables oculten a otras es necesario que las mismas estén expresadas en una base comparable (normalizadas). En tal sentido, en este trabajo se normalizaron (z-estandarización) las variables restando la media de cada una y dividiendo por su desvío estándar.

Luego de normalizadas, a las nueve variables iniciales se les aplicó la técnica de componentes principales.

Cuadro 6: Matriz de correlaciones de las variables originales

	CanTick	GsTotal	GsToTik	MonTick	PorCatg	CategDf	Perce10	Perce50	Perce90
CanTick	1	0.787	0.975	-0.635	-0.695	-0.044	0.137	-0.137	-0.359
GsTotal	0.787	1	0.843	-0.585	-0.498	0.323	0.115	-0.098	-0.228
GsToTik	0.975	0.843	1	-0.573	-0.649	-0.045	0.142	-0.182	-0.389
MonTick	-0.635	-0.585	-0.573	1	0.771	-0.332	-0.093	-0.180	0.022
PorCatg	-0.695	-0.498	-0.649	0.771	1	0.174	0.078	0.156	0.340
CategDf	-0.044	0.323	-0.045	-0.332	0.174	1	0.144	0.316	0.361
Perce10	0.137	0.115	0.142	-0.093	0.078	0.144	1	0.411	0.188
Perce50	-0.137	-0.098	-0.182	-0.180	0.156	0.316	0.411	1	0.607
Perce90	-0.359	-0.228	-0.389	0.022	0.340	0.361	0.188	0.607	1

Como se observa en el cuadro 6, las variables originales presentan una alta correlación entre ellas, con lo cual aplicar la técnica de componentes principales resulta muy útil.

Dada una serie de variables ($x_1,x_2,\ldots,x_p$) de un grupo de objetos o individuos, se propone calcular a partir de ellas un nuevo conjunto de variables ($y_1,y_2,\ldots,y_p$) incorrelacionadas entre sí, cuyas varianzas vayan decreciendo progresivamente.

Cada y_j (donde j = 1, …, p) es una combinación lineal de las $x_1,x_2,\ldots,x_p$ originales, es decir que:

$$y_1 = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jp}x_p = a`_j x$$

Donde $a`_j = (a_{1j},a_{2j},\ldots,a_{pj})$ es un vector de constantes, y $x = (x_1,x_2,\ldots,x_p)`$.

Si lo que buscamos es maximizar la varianza explicada, como veremos luego, una manera sencilla sería aumentar los coeficientes $a_{ij.}$ Por ello, para mantener la ortogonalidad de la transformación se impone que la norma de $a`_j = (a_{1j},a_{2j},\ldots,a_{pj})$ sea 1. Es decir,

$$a`_j a_j = \sum_{k=1}^{p} a_{kj}^2 = 1$$

El primer componente se calcula eligiendo a_1 de modo que y_1 explique la mayor varianza posible, sujeta a la restricción de que $a`_1 a_1 = 1$. El segundo componente principal se calcula obteniendo a_2 de modo que la variable obtenida, y_2 esté incorrelacionada con y_1.

Del mismo modo se eligen $y_1,y_2,\ldots,y_p$, incorrelacionados entre sí, de manera que la combinación lineal de las variables originales acumulen el mayor porcentaje de inercia inicial.

Queremos elegir a_1 de manera de maximizar la varianza de y_1 sujeto a la restricción $a`_1 a_1 = 1$.

$$\mathrm{Var}\,(y_1) = \mathrm{Var}\,(a`_1 x) = a`_1 \Sigma\, a_1$$

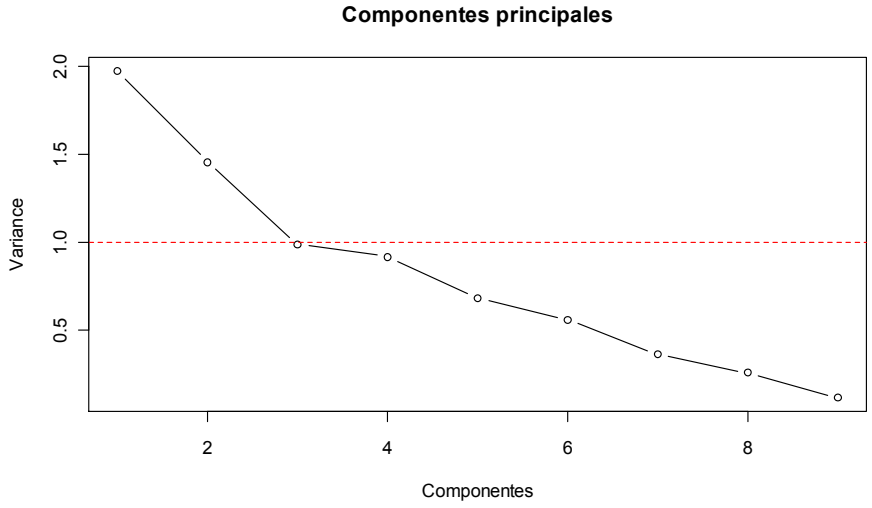
siendo Σ la matriz de covarianzas

El segundo componente principal, digamos $y_2 = a´_2x$, se obtiene mediante un argumento similar. Tendremos que maximizar la varianza de y_2, sujeto a las restricciones $a`_2a_2 = 1$ y $a`_2a_1 = 0$ (esto último garantiza la incorrelación de los componentes principales).

Para la determinación del número óptimo de nuevas variables o componentes principales se puede aplicar el análisis del gráfico de sedimentación de Cattell o analizar el nivel de explicación de la varianza total.

El gráfico de sedimentación de Cattell (Cattell 1966, Cattell y Vogelman 1977) es un método visual que analiza la representación gráfica de los valores propios de la matriz de correlaciones de las variables originales estandarizadas. El valor propio indica la cantidad de varianza explicada por un componente principal. Se consideran las q primeras componentes (con q<p siendo p el número total de variables) hasta que los descensos de pendiente sean poco significativos. Esta representación gráfica indica con claridad dónde terminan los valores propios “altos” (que explican gran parte de la varianza) y donde empiezan los “bajos”. Los valores propios residuales se encuentran a la derecha formando una planicie de poca pendiente, en contraposición tenemos en la parte izquierda del gráfico los valores propios relevantes que explican la mayor parte de la varianza original.

Gráfico 2: Gráfico de sedimentación de Cattell.



En nuestro caso, observando el gráfico 2 se determina en tres el número de componentes principales, ya que en tres se da el punto de quiebre de la pendiente, o sea el punto a partir del cual los nuevos componente contribuyen limitadamente a la explicación de la varianza total.

Por su parte en el cuadro 7 se refleja la determinación del número de componentes principales según qué porcentaje explican de la varianza total original de las variables. En este caso, los tres primeros componentes explican más del 79% de la varianza original, nivel considerado razonable de explicación.

Cuadro 7: Varianza explicada por los componentes principales.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Devío estandar	1.975	1.454	0.986	0.917	0.680	0.559	0.360	0.256	0.115
Proporción de varianza	0.443	0.240	0.110	0.095	0.052	0.035	0.015	0.007	0.002
Proporción aumulada	0.443	0.683	0.793	0.888	0.941	0.976	0.991	0.998	1.000

3.2.1. Relaciones entre los componentes principales y las variables originales

Caracterizamos los 3 componentes principales en función de las variables originales utilizando los vectores propios asociados a los tres primeros valores propios (cuadro 8).

Cuadro 8: Matriz de vectores propios.

	Comp.1	Comp.2	Comp.3
CanTick (norm.)	-0.475	0.013	-0.152
GsTotal (norm.)	-0.424	0.142	0.120
GsToTik (norm.)	-0.473	-0.007	-0.164
MonTick (norm.)	0.410	0.065	-0.064
PorCatg (norm.)	-0.015	0.474	0.527
CategDf (norm.)	0.379	-0.274	-0.249
Perce10 (norm.)	-0.035	0.374	-0.738
Perce50 (norm.)	0.101	0.552	-0.191
Perce90 (norm.)	0.217	0.481	0.106

De la matriz se desprende que el componente 1 está asociado al número de *tickets* que incluyen productos de la categoría j (CanTick), por el gasto total en productos del supermercado que efectúan los compradores de productos de la categoría j (GsToTik), por el gasto total de productos de la categoría j (GsTotal) y por el Monto promedio de los *tickets* que incluyen productos de la categoría j (MonTick). En los tres primeros casos la relación con el componente es negativa y en el cuarto es positiva, o sea que un valor alto del componente 1 indica baja cantidad de *tickets*, bajo monto total de los *tickets* que incluyen los bienes de esa categoría pero con un monto promedio alto de los *tickets*. En menor medida el componente 1 se explica por el número promedio de categorías diferentes incluidas en los *tickets* de compradores de productos de la categoría j (CategDf), con una relación positiva.

En los gráficos 3 y 4 se observa que categorías de bienes como la leche larga vida, los cereales, la margarina y los helados, entre otros, presentan altos valores en el componente 1. Esto implica que son bienes incluidos en pocos *tickets*, con baja participación en el gasto total del supermercado y con un gasto total de los clientes que compran esos bienes también bajo. Sin embargo, el monto promedio de los *tickets* y la cantidad de categorías diferentes incluidas en el *ticket* son relativamente altos.

En el otro extremo encontramos categorías como las verduras y legumbres, la carne vacuna o la leche. Estas categorías presentan niveles muy bajos en el componente uno y por lo tanto agrupan bienes con muchos *tickets* y que representan un porcentaje alto del gasto total, tanto a nivel de la categoría individual como de total de gasto de los tickets que los incluyen. Por otro lado, se compran junto con pocas categorías diferentes y el monto promedio del ticket es bajo.

Grafico 3 Representación de las categorías según componentes.

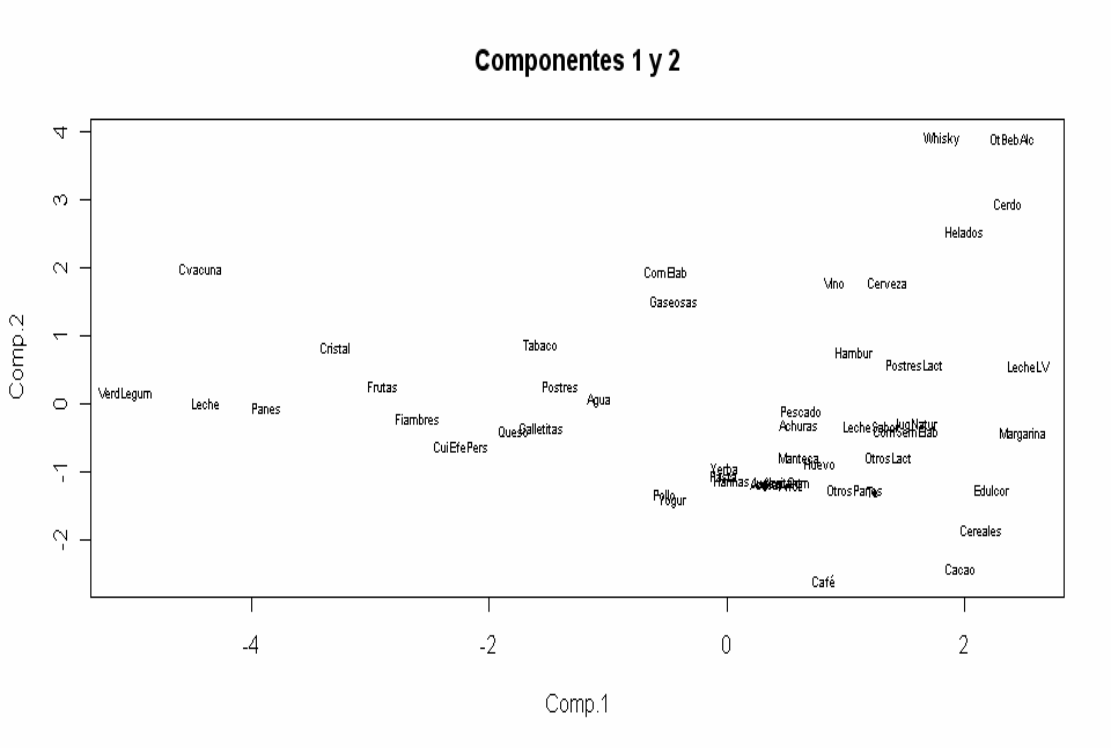
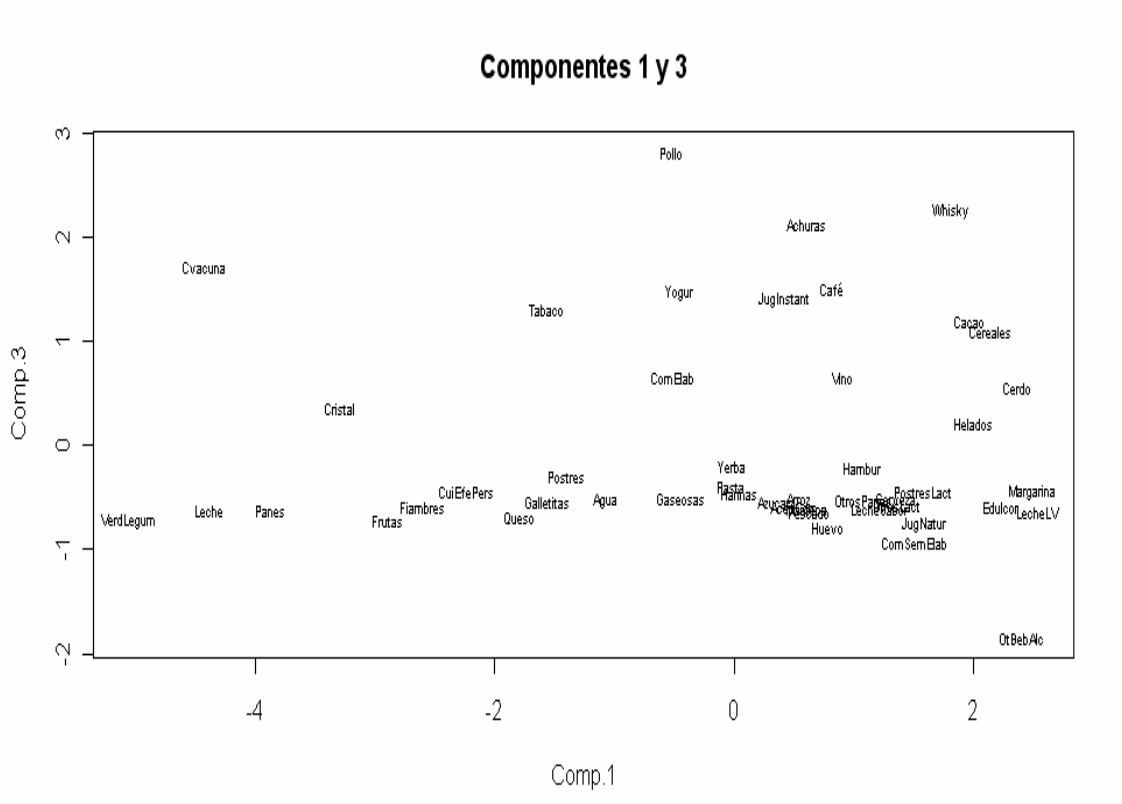


Grafico 4 Representación de las categorías según componentes.



El componente 2 se explica mayoritariamente por el día del mes en que se acumula 50% del gasto mensual total en productos de la categoría j (Perce50),

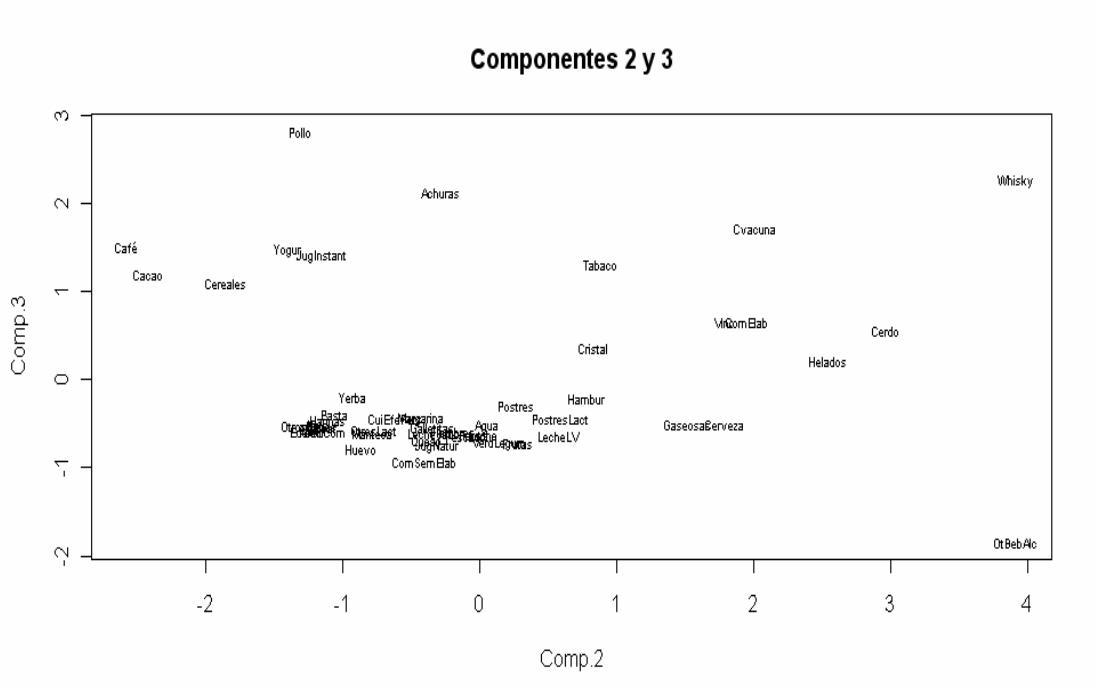
el día del mes en que se acumula 90% del gasto mensual total en productos de la categoría j (Perce90) y la razón entre las ventas de la categoría j y el gasto total de los compradores de productos de la categoría j (PorCatg).

En tal sentido, el componente 2 indica con que frecuencia se realizan las compras y la ponderación del gasto de la categoría respecto al total.

En los tres casos la relación entre el componente 2 y las variables originales es positiva, lo que implica que valores altos del componente 2 marcan una compra lenta a lo largo del mes (se demora más en llegar el percentil respectivo) y un peso relativo alto del gasto de la categoría en el total del gasto de los *tickets* que incluyen a esa categoría.

En menor medida el componente se explica por el día del mes en que se acumula 10% del gasto mensual total en productos de la categoría j (Perce10) en una relación positiva y por el número promedio de categorías diferentes incluidas en los *tickets* de compradores de productos de la categoría j (CategDf) con una relación negativa.

Gráfico 5 Representación de las categorías según componentes.



Las categorías con alto nivel de componente 2 incluyen bienes que se compran tardíamente en el mes, que tienen un fuerte peso relativo en los *tickets* en que figuran e incluyen pocas categorías diferentes dentro del *ticket*. Los gráficos 3 y 5 permiten observar que whisky, cerdo, helados y otras bebidas alcohólicas son ejemplos de este tipo de categorías. Por otro lado, también se observa que café y cacao por ejemplo, muestran bajos niveles en el componente 2, y por lo tanto son categorías que se compran rápidamente en el mes y que tienen baja incidencia en los *tickets* que los incluyen.

Finalmente, el componente 3 se explica principalmente por el día del mes en que se acumula 10% del gasto mensual total en productos de la categoría j (Perce10) con una relación negativa y por la razón entre las ventas de la categoría j y el gasto total de los compradores de productos de la categoría j (PorCatg) con una relación positiva.

Los gráficos 4 y 5 permiten observar que el Pollo es la principal categoría en este componente. Esto implica que este producto se compra rápidamente en el mes y tiene un peso relativo alto en los *tickets* que lo incluyen. Por otro lado, las otras bebidas alcohólicas se compran lentamente en el mes y tiene un bajo peso relativo en los *tickets* pues tienen un nivel mínimo en el componente 3.

3.3. Análisis de *Cluster* o Conglomerados

Con el objetivo de clasificar una población en un número determinado de conglomerados -agrupando aquellos que presenten similitudes - se aplican las técnicas de análisis de *cluster* para revelar las agrupaciones naturales dentro de un conjunto de datos.

Es así que se conforman grupos homogéneos según una medida de similitud o proximidad, de tal forma que cada grupo esté integrado por unidades con características relativamente homogéneas entre sí y diferenciadas respecto a las de los otros grupos. Estos grupos se llaman conglomerados o *clusters*. Si la clasificación realizada es óptima, los objetos dentro de cada *cluster* estarán

cercanos unos de otros y muy separados de los objetos de los *cluster* diferentes. La similitud (ó disimilitud) entre los grupos puede ser medida a través de la distancia, por ejemplo la distancia euclidiana[4], que es la utilizada en este trabajo. El resultado final de los *clusters* depende en forma importante de la medida de distancia o similitud que se utilice.

Es importante destacar que la técnica de *clusters* no tiene propiedades de inferencia; los resultados sólo se aplican a la muestra que los originó.

Por otro lado, se debe aplicar algún método para conformar los *clusters*. El resultado final de los *clusters* también depende en forma importante del algoritmo de aglomeración que se utilice. En este trabajo se utilizó el método "*bagged clustering*".

A partir de la aplicación de la técnica de "*bagged clustering*" a los 3 componentes principales definidos se generaron los primeros *clusters*.

Luego de ello se debió determinar la cantidad de *clusters* óptimos existentes en la muestra, para lo cual se utilizó el algoritmo KGS.

El algoritmo KGS permitió determinar cuatro *clusters*. El *cluster* 1 está integrado por categorías de bienes asimilados a una canasta de consumo básica conformadas por productos en su mayoría no perecederos. El *cluster* 2 está integrado por bienes fundamentalmente perecederos. El *cluster* 3 está conformado por bienes que no son de primera necesidad y el *cluster* 4 está integrado únicamente por la carne vacuna. En función de la conformación de cada uno de los *clusters* les hemos asignado una denominación que engloba las características más representativas de cada uno: "canasta básica", "bienes perecederos", "fiestas" y "carne vacuna".

El siguiente cuadro muestra las categorías de bienes incluidas en cada *cluster*:

[4] Otras medidas de distancia comúnmente utilizadas son la función de la distancia absoluta (o Manhattan o City-Block), la Formulación general de Power (s,r), el D^2 de Mahalanobis, etc

Cuadro 9 Categorías por *cluster*.

Básicos				Perecederos		Fiestas	Carne Vacuna
AceitCom	ComSemElab	LecheLV	Pescado	Agua	Gaseosas	Cerdo	Cvacuna
Achuras	Edulcor	LecheSabor	Pollo	ComElab	Leche	Cerveza	
Arroz	Hambur	Manteca	PostresLact	Cristal	Panes	Helados	
Azucar	Harinas	Margarina	Te	CuiEfePers	Postres	OtBebAlc	
Cacao	Huevo	OtrosLact	Yerba	Fiambres	Queso	Vino	
Café	JugInstant	OtrosPanes	Yogur	Frutas	Tabaco	Whisky	
Cereales	JugNatur	Pasta		Galletitas	VerdLegum		

3.3.1. El método "*bagged clustering*".

El método de "*bagged clustering*" combina dos de los algoritmos de aglomeración más frecuentemente utilizadas, el Método Particionado (o no jerárquico) y el Método Jerárquico.

El Método Particionado esta diseñado para encontrar *cluster* convexos en los datos de forma tal que cada uno pueda ser representado por un centroide del mismo. El método de los *clusters* convexos (o de segmentación de datos) está estrechamente relacionado con el método de clasificación de cuantificación de vectores (*vector quantization*), donde cada vector de entrada se asocia a su respectiva representación. En particular, el *centroide* de los datos segmentados y los vectores pueden verse como idénticos, y tienen una relación uno a uno, donde uno define al otro y viceversa (en condiciones generales).

Sea $X_N = (x_1,x_2,\ldots,x_N)$ una base de "*N*" datos, $C_K = (c_1,c_2,\ldots,c_N)$ el conjunto de centroides de los "*K*" *clusters* y $c(x) \in C_K$ el centroide más cercano a *X* en función de alguna relación de distancia medida por *d*. La solución al problema de *clusters* convexos surge de encontrar un conjunto de *centroides* tal que la distancia media entre cada punto y el *centroide* mas cercano sea mínima.

$$\sum_{n=1}^{N} d(x_n, c(x_n)) \rightarrow \min_{C_k}$$

Donde $d(X_n,c(X_n))$ es la distancia euclideana entre X_n y el centroide mas cercano y min c_k es el valor mínimo en función del numero de centroides (equivalente al numero de *clusters*).

El Método Jerárquico no trata de encontrar una segmentación con un número fijo de *clusters*, sino que crea soluciones para los *K clusters*, *con K = 1,…., N*. Para *K = 1* la única solución posible es un gran *cluster* que incluye todos los datos. Por otro lado, para *K = N* la solución son *N clusters* que contienen un único punto (o dato). Entre ambos extremos la jerarquización de *clusters* se van generando agrupando de a dos aquellos *clusters* más próximos.

En este trabajo se utiliza el método "*bagged clustering*", que combina los dos procedimientos o técnicas anteriores para determinar la cantidad de *clusters* óptima.

El algoritmo de "*bagged clustering*" funciona construyendo *B* muestras $X^1_{N,}$ $X^2_{N,\ldots,}X^B_N$ a partir de extracciones con reposición de la muestra original y luego se ejecuta un método no jerárquico de *clusters* (K-medias[5], aprendizaje competitivo, etc.) sobre cada una de las muestras. Esto resulta en *BxK* centroides (c_{11}, $c_{12},\ldots,c_{1K},c_{2K},\ldots,c_{BK}$) donde K es el número de los centroides utilizados en el método de base y c_{ij} es el *j*-ésimo centroide hallado usando X^i_N.

Todos los centroides se combinan para conformar un nuevo conjunto de datos $C^B = C^B(K) = (c_{11},\ldots,c_{BK})$.

Al nuevo conjunto de datos C^B se le aplica un algoritmo de agrupamiento jerárquico para llegar a la estructura de grupos final.

Una forma de determinar la cantidad de grupos es a través de la observación del dendrograma[6]. Sea c(x) Є C^B el centroide más cercano a *x*. Una partición de los datos originales puede obtenerse cortando el dendrograma en un cierto

[5] El algoritmo de k medias (*k-means*) es el referente principal entre los diversos métodos para seleccionar grupos representativos entre los datos. Sus diferentes variantes se basan fundamentalmente en la forma de medir distancias entre los datos y los grupos, el criterio para definir la pertenencia de los datos a cada grupo y la forma de actualizar dichos grupos.

[6] El dendrograma es la representación gráfica que mejor ayuda a interpretar el resultado de un análisis de clusters. Es un tipo de representación gráfica o diagrama de datos en forma de árbol (Dendro=árbol) que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado (asemejándose a las ramas de un árbol que se van dividiendo en otras sucesivamente).

nivel, lo que resulta en una partición $C^B_1,\ldots,C^B_m$ de la muestra C^B con $1 \leq m \leq BK$. Cada punto x Є X_N está asignado al *cluster* que contiene a c(x).

3.3.2. Determinación de la cantidad de *clusters*: algoritmo KGS

Para determinar la cantidad de *clusters* óptimos existentes en la muestra se utilizó el estadístico KGS (Kelley L.A., Gardner, S.P. 1996), el cual es un algoritmo que promedia las distancias entre los distintos *clusters* y aplica una función de penalización que busca obtener el número óptimo de los mismos.

Primeramente, el algoritmo necesita como insumo una matriz de las distancias existentes entre los *N* componentes del conjunto de datos iniciales, una matriz simétrica de *NxN*.

En segundo lugar se aplica un algoritmo de aglomeración promedio para un análisis de *clusters* de tipo jerárquico. El método de aglomeración promedio toma las distancias entre dos *clusters m* y *n* calculada de la siguiente forma:

$$dist(m,n) = \frac{\left(\sum_{i=1}^{X} \sum_{j=1}^{Y} \text{dist}(i,j)\right)}{X\,Y}$$

Donde *Y* es la cantidad de miembros de *cluster n, X* es la cantidad de miembros de cluster *m y dist(i,j)* es la distancia euclideana entre dos miembros de los *clusters m* y *n* respectivamente.

Los elementos que presentan una menor distancia entre ellos se van agrupando y generando un nuevo grupo ó *cluster*, calculando nuevamente la distancia con el resto, y agrupándose con el de menor distancia. Este cálculo se realiza sucesivamente hasta que resulte un solo grupo (y su distancia sea cero).

En cada etapa de agrupamiento se calcula el *spread* del *cluster* de la siguiente manera:

$$spread_m = \frac{\left(\sum_{k=1}^{N} \sum_{i=1, i<k}^{N} dist(i,k)\right)}{N\,(N-1)/2}$$

Donde *m* es el *cluster*, *N* la cantidad de miembros del *cluster*, e *i* y *k* son miembros del *cluster*

Luego se calculan los *spreads* promedio ($AvSp_i$) de cada *cluster*:

$$AvSp_i = \frac{\sum_{m=1}^{cnum_i} spread_m}{cnum_i}$$

Dónde $AvSp_i$ es el *spread* promedio de todos los *cluster*s
y $cnum_i$ es el numero de *cluster*s en la etapa *i*

Una vez finalizado el proceso de agrupamiento, los valores de los *spreads* promedios son normalizados para que estén en el rango entre *1* y *N-1*, donde *N* es el número de grupos en el conjunto inicial de datos.

$$AvSp(norm)_i = \left(\frac{N-2}{Max(AvSp) - Min(AvSp)}\right)(AvSp_i - Min(AvSp)) + 1$$

Donde $AvSp(norm)_i$ es el spread promedio normalizado en la etapa i, N número de grupos en el conjunto inicial de datos, $AvSp_i$ es el *spread* promedio de todos los clusters, y $Max(AvSp_i)$ y $Min(AvSp_i)$ son el máximo y mínimo spread promedio respectivamente.

En cada instancia de agrupamiento *i* se calcula el valor de penalidad P_i de la siguiente manera:

$$P_i = AvSp(norm)_i + nclus_i$$

Donde $nclus_i$ es el número total de *cluster*s en la instancia *i* de agrupación.

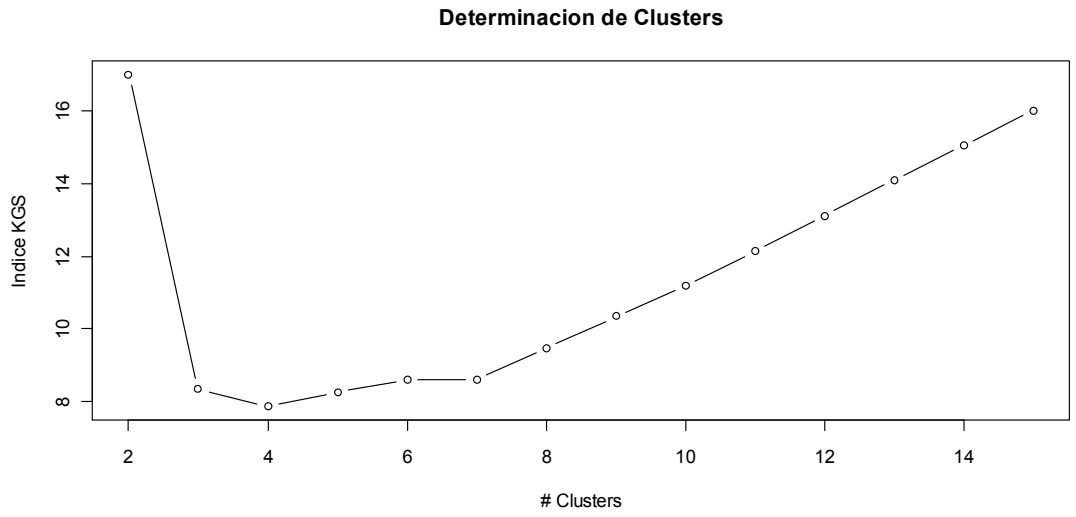
El valor mínimo de penalidad en el set (P_1,P_2,…,P_{N-1}) se elige como el nivel de corte.

$$P_{icut} = Min(P)$$

Así, la etapa icut representa el estado donde los grupos son tan altamente poblados como sea posible, manteniendo al mismo tiempo la menor dispersión (spread) entre sus miembros.

La observación gráfica del mínimo de la función determina la cantidad óptima de *cluster*s. En este caso, en el gráfico 6 se observa que el número óptimo de *clusters* es 4.

Gráfico 6



3.4. Caracterización de los *clusters*

A partir del procedimiento de *bagged clustering* y aplicando el algoritmo KGS determinamos cuatro *clusters*: “canasta básica”, “bienes perecederos”, “fiestas” y “carne vacuna”.

Cluster 1 – Canasta básica

El *Cluster* 1 incluye los bienes considerados del “surtido mensual o semanal”. Se caracteriza por estar integrado principalmente por bienes de consumo no perecedero como arroz, aceites, cereales, entre otros, y algunos perecederos pero de consumo no instantáneo como huevos, hamburguesas y margarina.

Las categorías del *cluster* presentan niveles altos en los componentes uno y dos, y niveles dispersos en el componente tres. Esto implica que el gasto total de cada uno de estos bienes es bajo en relación al resto de los bienes, que el gasto en el bien es bajo en relación al total de gasto del *ticket* y que presentan una baja frecuencia de compra, ya que están presentes en pocos *tickets*. Por otro lado, los *tickets* que incluyen estos bienes tienen un valor alto.

A su vez, son bienes que se compran en los primeros días del mes, ya que su gasto rápidamente acumula porcentajes altos del gasto total del bien en el mes.

En menor medida, se caracteriza el *cluster* por estar integrado por bienes que se compran con varios otros a la vez.

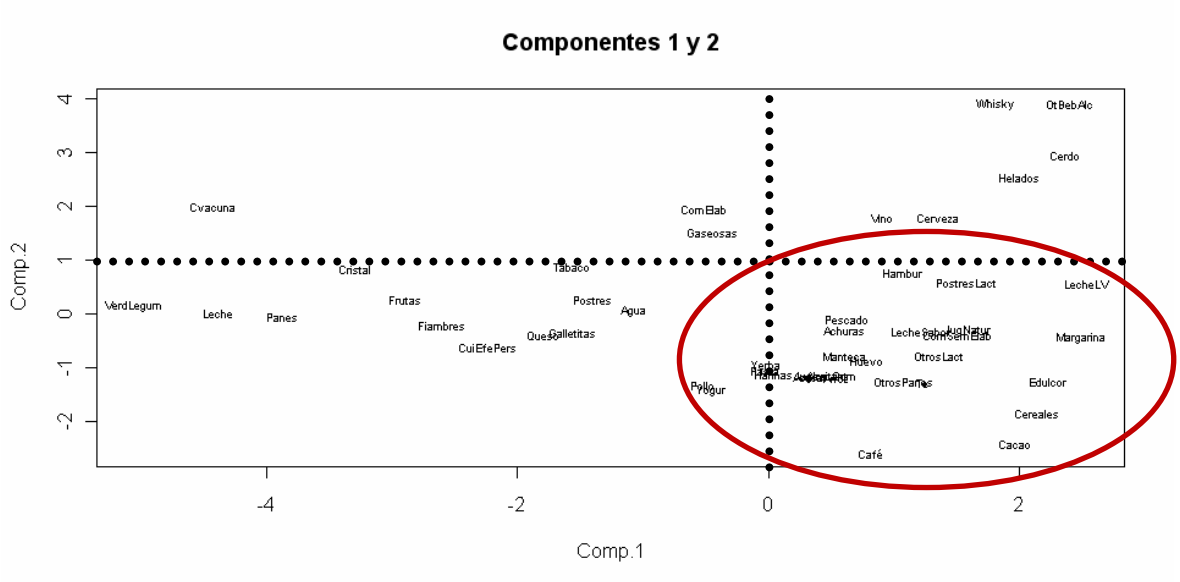
En resumen, son bienes de bajo valor relativo que se compran al inicio del mes, en pocas ocasiones y conjuntamente con muchos otros bienes, todo lo cual caracteriza los bienes que se compran en los "surtidos" semanales o mensuales.

En el cuadro 10 presentamos los valores descriptivos principales de las nueve variables del *cluster* y de la base de datos completa para poder compararlos fácilmente.

Cuadro 10 Datos por variable del *cluster* 1 y de la base total.

Nº de Cat.	Categoría	CanTick	GsTotal	GsToTik	MonTick	PorCatg	CategDf	Perce10	Perce50	Perce90
27	Suma	69,468	2,235,625	26,618,332						
Cluster 1	Maximo	5,856	334,371	1,897,400	544	0	10	4	17	29
	Promedio	2,573	82,801	985,864	402	0	8	4	16	28
	Mínimo	468	20,702	254,528	293	0	7	3	14	28
48	Suma	266,253	9,675,352	83,327,203						
Base total	Máximo	23,731	1,116,379	6,206,403	544	0.39	9.61	5	20	30
	Promedio	5,547	201,570	1,735,983	376	0.11	7.91	4	16	28
	Mínimo	468	20,702	254,528	201	0.05	5.26	3	14	28

Gráfico 7 Identificación del *cluster* 1 según componentes 1 y 2.



Cluster 2 - Productos perecederos

En este *cluster* encontramos principalmente a los bienes perecederos o de consumo instantáneo. Son bienes que se compran diariamente, ya sea por sus características intrínsecas (son perecederos) como la leche, el pan y las verduras o por ser de consumo instantáneo o impulsivo, como las gaseosas, tabaco y postres. Si bien estos bienes son los de mayor peso relativo dentro del *cluster*, también engloban algunas categorías de bienes que no presentan estas características como las galletitas, cuidados y efectos personales, y cristales.

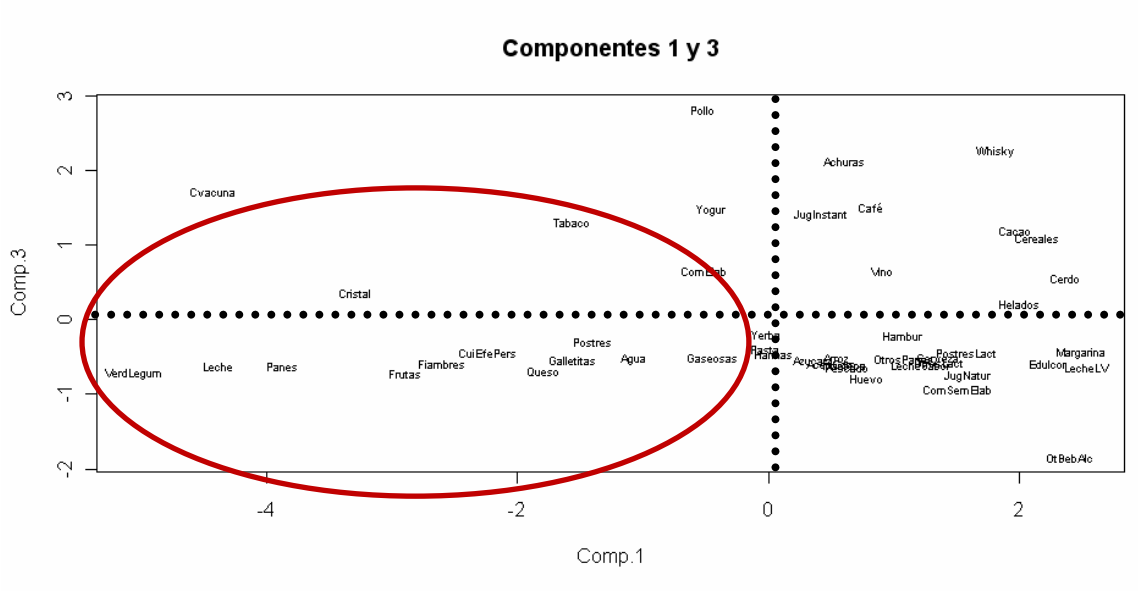
Cuadro 11 Datos por variable del *cluster* 2 y de la base total.

Nº de Cat.	Categoría	CanTick	GsTotal	GsToTik	MonTick	PorCatg	CategDf	Perce10	Perce50	Perce90
14	Suma	175,010	5,698,389	49,540,336						
Cluster 2	Maximo	23,731	801,252	6,206,403	362	0	8	4	18	29
	Promedio	12,501	407,028	3,538,595	292	0	7	4	16	28
	Mínimo	3,998	145,255	804,813	201	0	5	4	15	28
48	Suma	266,253	9,675,352	83,327,203						
Base total	Máximo	23,731	1,116,379	6,206,403	544	0.39	9.61	5	20	30
	Promedio	5,547	201,570	1,735,983	376	0.11	7.91	4	16	28
	Mínimo	468	20,702	254,528	201	0.05	5.26	3	14	28

Las categorías del *cluster* se diferencian del resto en el componente uno. Se caracterizan principalmente por tener un gasto total alto, un peso relativo alto

dentro de los *tickets* que los incluyen, una alta frecuencia de compra (están presente en muchos *tickets*) y por estar incluidos en *tickets* de bajo valor. Además, son bienes que se compran de manera regular durante todo el mes.

Gráfico 8 Identificación del *cluster* 2 según componentes 1 y 3.



Son bienes que se va a comprar expresamente, de alto consumo y que se compran muy frecuentemente.

Cluster 3 – Fiestas

El *cluster* 3 agrupa las categorías de bienes que se compran para las fiestas navideñas. Su comportamiento está influenciado por las compras del 25 y 31 de diciembre[7].

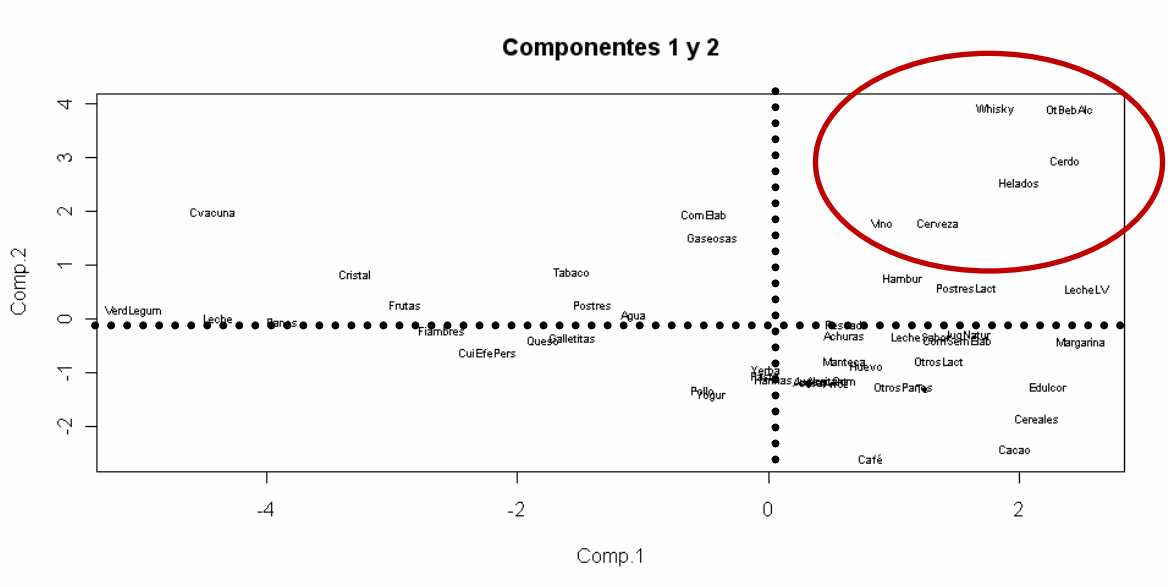
Se diferencia del resto tanto en el componente uno como en el dos.

[7] Se realizó el mismo ejercicio excluyendo el mes de diciembre. Solo el whisky mantenía claramente estas características. Aún así se decidió mantener el escenario base con diciembre debido al peso relativo de ese mes en el total de ventas del supermercado (mayor a un tercio del total de ventas del semestre).

Cuadro 12 Datos por variable del *cluster* 3 y de la base total.

Nº de Cat.	Categoría	CanTick	GsTotal	GsToTik	MonTick	PorCatg	CategDf	Perce10	Perce50	Perce90
6	Suma	7,044	624,959	3,139,990						
Cluster 3	Maximo	2,267	169,514	923,976	540	0	8	5	20	30
	Promedio	1,174	104,160	523,332	468	0	8	4	19	29
	Mínimo	556	46,964	294,368	388	0	7	4	17	29
48	Suma	266,253	9,675,352	83,327,203						
Base total	Máximo	23,731	1,116,379	6,206,403	544	0.39	9.61	5	20	30
	Promedio	5,547	201,570	1,735,983	376	0.11	7.91	4	16	28
	Mínimo	468	20,702	254,528	201	0.05	5.26	3	14	28

Gráfico 9 Identificación del *cluster* 3 según componentes 1 y 2.



Está integrado por categorías de bienes que incluyen pocos *tickets* pero con un valor promedio alto y con una alta participación en el valor total. El gasto total en estos bienes es bajo al igual que el gasto total de los *tickets* que los incluyen. Son bienes que se compran tardíamente en el mes, principalmente por el efecto de las fiestas navideñas. Están incluidos en este *cluster* las bebidas alcohólicas, los helados y el cerdo.

Cluster 4 – Carne Vacuna

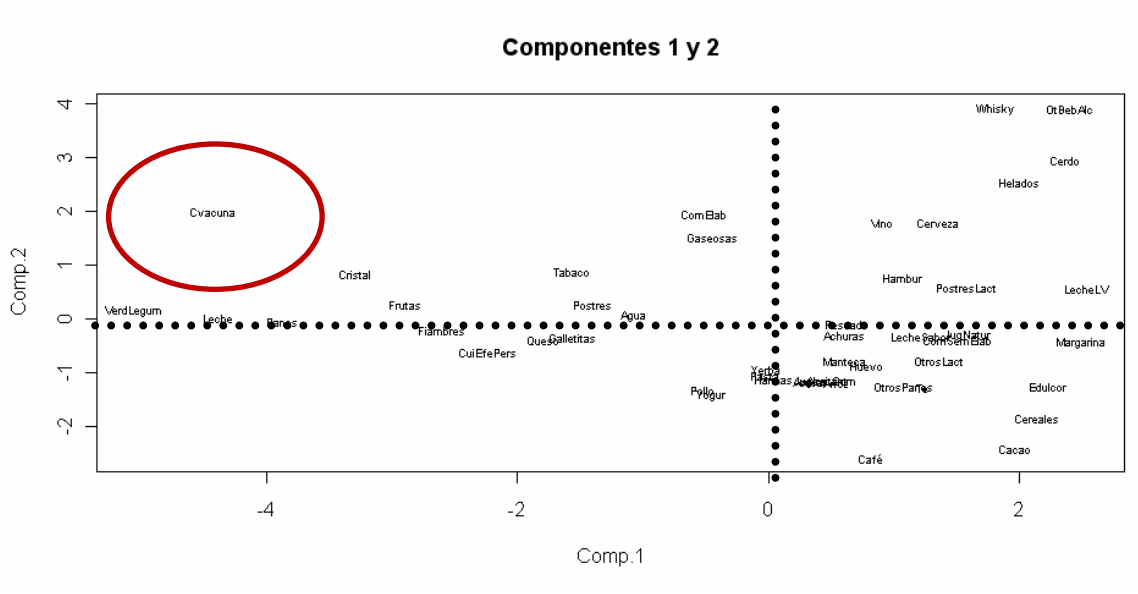
La carne vacuna es un bien tan relevante en las ventas de la cadena de supermercados que conforma un *cluster* por si misma.

El *cluster* queda diferenciado del resto en cualquiera de los tres componentes.

Cuadro 13 Datos por variable del *cluster* 4 y de la base total.

Nº de Cat.	Categoría	CanTick	GsTotal	GsToTik	MonTick	PorCatg	CategDf	Perce10	Perce50	Perce90
1	Suma	14,731	1,116,379	4,028,545						
Cluster 4	Maximo	14,731	1,116,379	4,028,545	273	0	6	4	16	28
	Promedio	14,731	1,116,379	4,028,545	273	0	6	4	16	28
	Mínimo	14,731	1,116,379	4,028,545	273	0	6	4	16	28
48	Suma	266,253	9,675,352	83,327,203						
Base total	Máximo	23,731	1,116,379	6,206,403	544	0.39	9.61	5	20	30
	Promedio	5,547	201,570	1,735,983	376	0.11	7.91	4	16	28
	Mínimo	468	20,702	254,528	201	0.05	5.26	3	14	28

Gráfico 10 Identificación del *cluster* 4 según componentes 1 y 2.

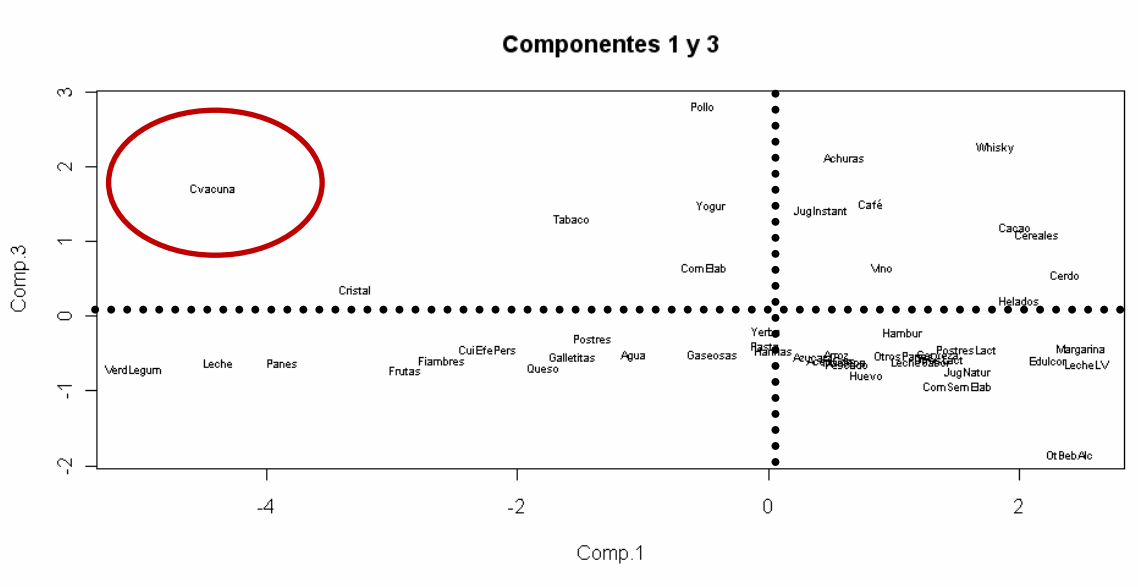


La carne vacuna es un bien que está presente en una importante cantidad de *tickets* y su gasto total es muy alto. De hecho es la categoría individual con mayor participación en el gasto total del supermercado.

Además, el peso relativo de la carne en el gasto total de los *tickets* que la incluyen es muy alto.

Por otro lado, es un producto que se compra más sobre fin de mes. Esto posiblemente responda al alto consumo de carne vacuna durante las fiestas.

Gráfico 11 Identificación del *cluster* 4 según componentes 1 y 3.



3.5. Significación estadística de los resultados

La representación gráfica de las categorías en los ejes de los componentes principales, junto con las correlaciones entre estos componentes principales y las variables originales permitió realizar un detallado análisis de las características de cada *cluster*.

Para verificar si las diferencias entre clusters eran significativas, realizamos tests de medias. Previo al test de medias se realizó un test de homogeneidad de varianzas[8] para aplicar el test de medias con el estadístico adecuado.

Para el test de homogeneidad de varianzas las hipótesis son:

Ho: Las varianzas de las variables es igual en los dos *cluster* que se comparan.

H1: Las varianzas no son iguales.

El estadístico de prueba es el siguiente:

[8] Test de Bartlett.

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i-1)\ln s_i^2}{1+(1/(3(k-1)))((\sum_{i=1}^{k}1/(N_i-1))-1/(N-k))}$$

Donde s_i^2 es la varianza del grupo i, N es el tamaño de la muestra total, N_i es el tamaño de la muestra del grupo *i-ésimo, k* es el número de grupos, y s_p^2 es la varianza agrupada.

La varianza combinada es una media ponderada de las varianzas de cada grupo y se define como:

$$s_p^2 = \sum_{i=1}^{k}(N_i-1)s_i^2/(N-k)$$

El estadístico T se distribuye χ_α^2 con k-1 grados de libertad bajo Ho cierta. Se rechaza Ho toda vez que $T > \chi^2_{(\alpha,k-1)}$ con α el grado de significación.

La primera parte del cuadro 14 detalla los resultados del test para las nueve variables en cada par de *clusters*. No se aplica el test al *cluster* 4 (Carne vacuna) por tener un solo elemento.

En base a los resultados del test de homogeneidad de varianzas se aplica el test de igualdad de medias[9], cuyo estadístico incorpora el hecho de que las varianzas sean iguales o no. En este caso las hipótesis son:

Ho: La media de la variable es igual en los dos *cluster* que se comparan.

H1: Las medias no son iguales.

Para la prueba de igualdad de medias (cuando no se rechaza la hipótesis nula de igualdad de varianzas) se utiliza el siguiente estadístico que, bajo Ho, se distribuye t- *student* con n+m -2 grados de libertad:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n-1)\hat{S}_1^2 + (m-1)\hat{S}_2^2}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

donde Sx y Sy corresponden a las estimaciones de las varianzas muestrales de los correspondientes grupos y “n” y “m” el tamaño de las respectivas poblaciones.

[9] T-test

Cuando se rechaza la hipótesis nula de igualdad de varianzas se sustituye en el denominador del estadístico anterior la varianza combinada por la siguiente estimación:

$$\sqrt{\frac{\hat{S}_1^2}{n}+\frac{\hat{S}_2^2}{m}}$$

Bajo Ho, el estadístico se distribuye *t-student* con los siguientes grados de libertad:

$$\mathrm{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}.$$

La segunda parte del cuadro 15 muestra los resultados del test para las nueve variables en cada par de *clusters.* Se observa que las diferencias de las medias son significativas en la gran mayoría de los casos. Esto implica que las medias de las nueve variables son significativamente diferentes en cada uno de los cuatro *clusters* considerados, excepto en tres casos: el gasto total entre los *clusters* 1 y 3, el percentil 10 entre los *clusters* 2 y 3 y el percentil 90 entre los *clusters* 1 y 2.

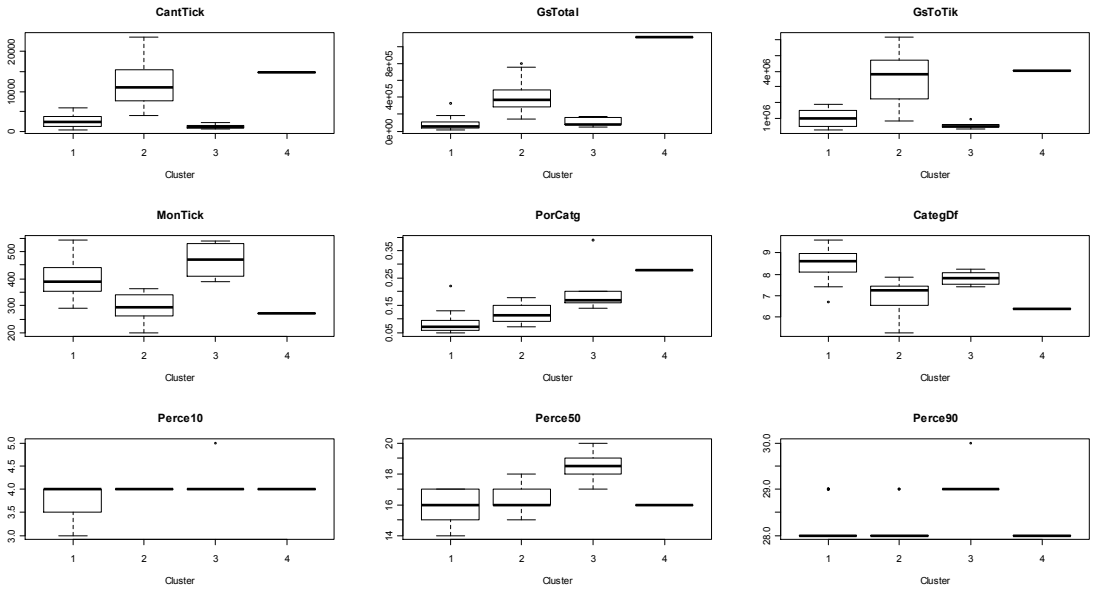
Cuadro 14 Resultados del test de homogeneidad de varianzas y de medias

Homogeneidad de Varianzas									
Cluster	**CanTick**	**GsTotal**	**GsToTik**	**MonTick**	**PorCatg**	**CategDf**	**Perce10**	**Perce50**	**Perce90**
1 con 2	***	***	***	NRH0	NRH0	NRH0	***	NRH0	NRH0
1 con 3	**	NRH0	**	NRH0	***	*	NRH0	NRH0	NRH0
2 con 3	***	***	***	NRH0	***	*	***	NRH0	NRH0
Igualdad de Medias									
1 con 2	***	***	***	***	***	***	***	**	NRH0
1 con 3	***	NRH0	***	**	**	***	**	***	***
2 con 3	***	***	***	***	*	***	NRH0	***	***
Significación al: * 10%, ** 5%, *** 1%									

Cuadro 14

El gráfico 12 muestra la distribución de las variables en cada uno de los clusters.

Gráfico 12: Distribución de las variables según *cluster*.



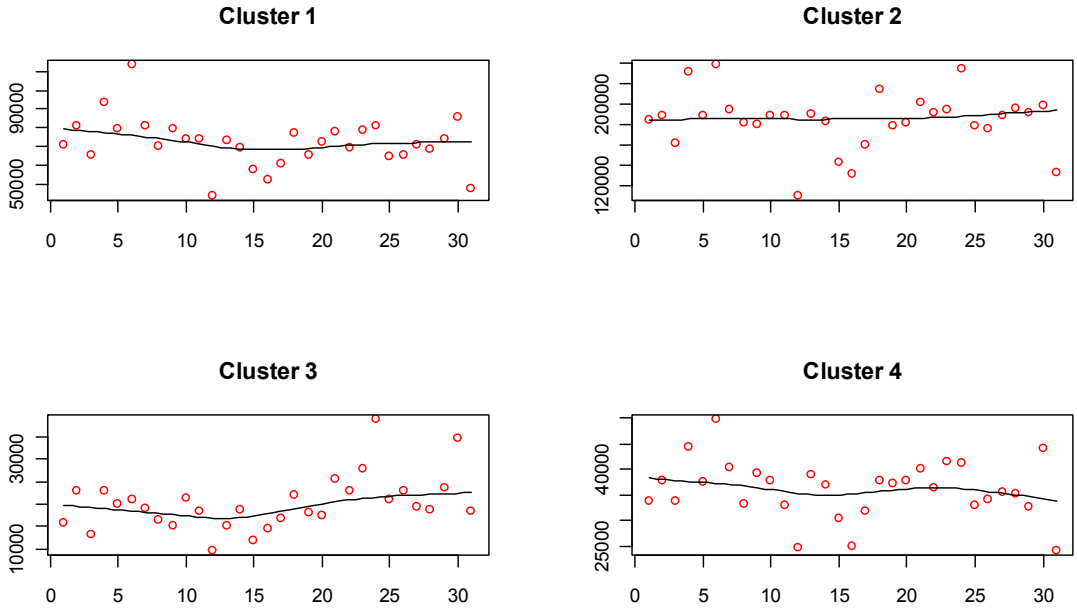
3.6. Regresiones sobre los *clusters* para observar si el gasto decrece durante el mes

Con el objetivo de analizar el comportamiento de los *clusters* y en búsqueda de identificar algún patrón de comportamiento del consumidor en términos temporales, realizamos un análisis gráfico a partir del modelo de regresión no paramétrico *Loess* que utilizamos al inicio para estudiar el comportamiento de las 48 categorías de bienes.

También corrimos una regresión por mínimos cuadrados ordinarios (MCO) al gasto total (en logaritmos) contra el día del mes para los cuatro *clusters*.

En el gráfico 13 se presenta la evolución del gasto total en categorías del *cluster* a lo largo del mes. Los *clusters* no presentan un comportamiento homogéneo en el tiempo. No todos presentan una trayectoria decreciente a lo largo del mes. Específicamente, el *cluster* 3 (Fiestas) muestra una trayectoria creciente a los largo del mes. Esto puede estar explicado por las compras para las fiestas de fin de año.

Gráfico 13 Gasto total por día.



A su vez, en las regresiones por MCO del gasto total (en logaritmos) contra el día del mes para los cuatro *clusters* observamos que sólo el coeficiente del *cluster* 3 (Fiestas) resultó significativo y positivo. Los coeficientes de las regresiones de los otros tres *cluster* no fueron significativos.

***Cluster* 1 – Canasta básica**

log(Total) ~ Dia

Residuals:

Min	1Q	Median	3Q	Max
-0.50249	-0.05716	0.01293	0.08651	0.41540

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	11.254950	0.067003	167.976	<2e-16	***
Dia	-0.005322	0.003655	-1.456	0.156	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.182 on 29 degrees of freedom
Multiple R-squared: 0.06813, Adjusted R-squared: 0.036
F-statistic: 2.12 on 1 and 29 DF, p-value: 0.1561

***Cluster* 2 – Bienes perecederos**

log(Total) ~ Dia

Residuals:

Min	1Q	Median	3Q	Max
-0.49847	-0.01522	0.02768	0.07019	0.26101

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	12.126809	0.061106	198.456	<2e-16	***
Dia	-0.001082	0.003334	-0.325	0.748	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.166 on 29 degrees of freedom
Multiple R-squared: 0.003618, Adjusted R-squared: -0.03074
F-statistic: 0.1053 on 1 and 29 DF, p-value: 0.7479

***Cluster* 3 – Fiestas**

log(Total) ~ Dia

Residuals:

Min	1Q	Median	3Q	Max
-0.6399725	-0.1648695	-0.0001770	0.1732783	0.5931471

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.668484	0.097080	99.593	<2e-16	***
Dia	0.012718	0.005296	2.401	0.0230	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2637 on 29 degrees of freedom
Multiple R-squared: 0.1659, Adjusted R-squared: 0.1371
F-statistic: 5.767 on 1 and 29 DF, p-value: 0.02296

***Cluster* 4 – Carne vacuna**

log(Total) ~ Dia

Residuals:

Min	1Q	Median	3Q	Max
-0.38219	-0.09243	0.02820	0.07655	0.30247

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	10.528686	0.060514	173.987	<2e-16	***
Dia	-0.003101	0.003301	-0.939	0.355	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1644 on 29 degrees of freedom
Multiple R-squared: 0.02953, Adjusted R-squared: -0.003938
F-statistic: 0.8823 on 1 and 29 DF, p-value: 0.3553

Esto permite concluir que de las cuatro canastas de bienes definidas solo dos presentan una tendencia del gasto decreciente a lo largo del mes que se puede ver gráficamente pero no se verifica estadísticamente. En los otros dos casos la tendencia es creciente pero solo en uno es significativa en el modelo MCO. En resumen, en un caso la tendencia es significativa y creciente mientras que los tres casos restantes los resultados no son concluyentes.

El análisis del gasto a lo largo del mes realizado sobre los clusters confirma el análisis realizado en la sección 3.1.

4. Conclusiones

Los resultados del análisis del comportamiento del consumidor respecto de la muestra analizada nos permitió identificar cuatro canastas de bienes que presentan similitudes a su interior y disimilaridades entre si. Las cuatro canastas definidas son: “Canasta básica”, “Productos perecederos”, “Fiestas” y “Carne Vacuna”.

Generamos asociaciones entre productos en base a criterios objetivos aplicando técnicas con muy pocos antecedentes en Uruguay.

Las diferencias entre los *clusters* surgen principalmente de las variables “Número de *tickets* que incluyen productos de la categoría j” (CanTick), “Gasto total de productos de la categoría j” (GsTotal), “Gasto total en productos del supermercado que efectúan los compradores de productos de la categoría j” (GsToTik) y “Monto promedio de los *tickets* que incluyen productos de la categoría j “(MonTick).

Las variables referidas al tiempo (momento del mes en que se realiza la compra: Perce10, Perce50 y Perce90) tienen menor incidencia en la definición de los *clusters*, al igual que el “Número promedio de categorías diferentes incluidas en los *tickets* de compradores de productos de la categoría j” (CategDf).

El *cluster* al que llamamos “Canasta básica” incluye bienes de bajo valor relativo, que se compran en pocas ocasiones y conjuntamente con muchos otros bienes, todo lo cual caracteriza los bienes que se compran en los “surtidos” semanales o mensuales. Está integrado principalmente por bienes de consumo no perecedero como arroz, aceites, cereales, entre otros, y algunos perecederos pero de consumo no instantáneo como huevos, hamburguesas, margarina, etc.

En el *cluster* de “Productos perecederos” encontramos principalmente bienes perecederos o de consumo instantáneo. Son bienes que se compran

diariamente, ya sea por sus características intrínsecas (son perecederos) como la leche, el pan y las verduras o por ser de consumo instantáneo o impulsivo, como las gaseosas, tabaco y postres. Son bienes a los cuales generalmente se los va a comprar expresamente, de alto consumo y que se compran muy frecuentemente.

El *cluster* "Fiestas" agrupa las categorías de bienes cuyo consumo aumenta notoriamente para las fiestas de fin de año. Su comportamiento está influenciado por las compras del 24 y 31 de diciembre. Se compran poco en términos relativos pero dentro de *tickets* de alto valor. Incluye todas las bebidas alcohólicas, el cerdo y los helados.

Finalmente, el último *cluster* esta integrado únicamente por la carne vacuna. Esta categoría de productos es tan relevante en las ventas de la cadena de supermercados que se diferencia del resto y conforma un *cluster* por si sola. De hecho es la categoría individual con mayor participación en el gasto total del supermercado.

Las diferencias entre los cuatro *clusters* fueron verificadas en su mayoría por tests de medias.

La importancia relativa de cada canasta en relación al gasto total en las ventas del supermercado es la siguiente[10]: canasta básica 19,1 %, bienes perecederos 48,7 %, carne vacuna 9,5% y fiestas 5,3%.

Tanto el gasto total promedio que realizan los consumidores de cada categoría, como el gasto total promedio que realizan solamente en esa categoría son mayores al promedio general para los *clusters* de productos perecederos y de carne vacuna y menores para los *clusters* de la canasta básica y la de fiestas.

Un tema no menor en esta estructura de comportamiento se observa al analizar el monto promedio de los *tickets* que incluyen productos de alguna de estas

[10] El 100% de las ventas del supermercado se completa con un 17,3% del gasto que corresponde a las categorías de productos eliminadas del análisis.

cuatro canastas y la cantidad de productos de otras categorías que integran la compra de un determinado producto. El monto promedio de los *tickets* que incluyen productos perecederos ($292) y carne vacuna ($273) registran montos por debajo del promedio de la muestra ($376), y a su vez se compran conjuntamente con una cantidad de productos de otras categorías también por debajo de la cantidad media de la muestra. Esto nos indica que son bienes que se van a comprar expresamente y en solitario o con muy pocos productos asociados.

Por su parte, las categorías de productos de las otras dos canastas presentan características bien diferentes en el sentido de que son bienes que se compran con un número importante de productos diferentes y el monto promedio de los *tickets* que los incluye está por encima del valor promedio de la muestra. Esto convalida la percepción de que son bienes que se compran en el surtido semanal o mensual dado que sus características intrínsecas así lo permiten (no perecederos).

Respecto a las preferencias intertemporales mostramos que de las cuatro canastas de bienes definidas sólo dos presentan una tendencia del gasto decreciente a lo largo del mes, que se puede ver gráficamente pero no se verifica estadísticamente. En los otros dos casos la tendencia es creciente pero solo en un caso es significativa en el modelo MCO. Cabe destacar que el único *cluster* que presenta una tendencia creciente y significativa es el de “Fiestas". Esa tendencia puede estar influenciada por las compras navideñas y de fin de año que se dan solo en el mes de diciembre pero tienen una ponderación importante en el total de gasto.

Debe considerarse que los resultados aquí enunciados sólo son válidos para la muestra de datos del presente trabajo. No se utilizó información de otros comercios de la ciudad de Montevideo y por lo tanto no se pueden generalizar los resultados al comportamiento de todos los consumidores. El consumidor de los supermercados en general podría tener un patrón de consumo diferente al resto de los consumidores, por ejemplo por posibles diferencias de poder adquisitivo o por la variedad de productos disponibles.

Bibliografía

- **Cattell, R.B. (1966):** *The Scree Test for the Number of Factors.* Multivariable Behavioral Research.
- **Cattell, R.B. y Vogelman, S. (1977):** *A comprehensive trial of the scree and KG criteria for determining the number of factors.* Multivariate Behavioral Research.
- **Cleveland, W.S. (1979):** *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, Vol. 74, pp. 829-836.
- **Cleveland, W.S. and Devlin, S.J. (1988):** *Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting.* Journal of the American Statistical Association, Vol. 83, pp. 596-610.
- **Durán, N. y Souto, G. (2009):** *Llegando a fin de mes: un análisis de las preferencias intertemporales de los uruguayos*. Trabajo Monográfico presentado ante la Facultad de Ciencias Económicas y Administración de la Universidad de la República para obtener el título de Licenciado en Economía
- **Friedman, M. (1957):** *A Theory of the Consumption Function,* Princeton University Press, Princeton.
- **Guerrero, F., Ramirez, J. (2002):** *El análisis de escalamiento multidimensional: Una alternativa y complemento a otras técnicas multivariantes*
- **Huffman y Berenstein (2004):** *Riches to Rags Every Month, The Fall in Consumption Expenditures Between Paydays*
- **Instituto Nacional de Estadística**; Índice de Precios al Consumo, Metodología Base Marzo 1997=100; http://www.ine.gub.uy
- **Kelley, L.**, **Gardner, S. y Sutcliffe, M. (1996):** *An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies.*
- **Kelmansky, D. (2006):** Análisis exploratorio y confirmatorio de Datos de Experimentos de Microarrays
- **Keynes, J.M. (1936):** *The General Theory of Employment, Interest and Money.*

- **Kotler, P. (1990):** *Principles of Marketing.*
- **Laibson, D**. (1998): *Life-cycle consumption and hyperbolic discount functions*, European Economic Review 42, p.862-871
- **Leisch, F. (1999):** *Bagged Clustering*
- **Linares, G. (2001):** *Escalamiento multidimensional: Conceptos y enfoques.*
- **Mastrobuoni, G. y Weinberg, M. (2007):** *Heterogeneity in Intra-Monthly Consumption Patterns.*
- **Modigliani, F. y Brumberg, R. (1954):** *Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data*, in K. Kurihara, ed., Post Keynesian Economics, Rutgers University Press, New Brunswick.
- **NIST/SEMATECH (2010):** *e-Handbook of Statistical Methods* http://www.itl.nist.gov/div898/handbook/.
- **Rencher, A. (2002):** *Methods of Multivariate Analysis*
- **Shapiro, M. y Slemrod, J. (1995):** *Consumer Response to the Timing of Income, Evidence from a Change in Tax Withholding.*
- **Stephens, M. (2002):** *Paycheck receipt and the timing of Consumption*
- **Stephens, M. (2002):** *3rd of tha month: Do Social Security Recipients Smooth Consumption Between Checks?*
- **Stigler G.J. y Becker G.S. (1977):** *Gustibus non est disputandum*. American Economic Review, 67, 76-90.
- **Tobin, J. (1958):** *Estimation of Relationship for Limited Dependent Variables* Econométrica 26(1), p.24-36.

Anexo I

Relaciones entre Componentes principales y variables originales.

En el cuadro 15 se presenta la relación entre cada uno de los tres componentes principales y las variables originales.

Las variables cuya correlación con el componente es mayor tienen fondo naranja o sombreado oscuro $(|\rho| > 0{,}4)$, correlación media con fondo amarillo o sombreado claro $(|\rho| > 0{,}3)$ y correlación menor con fondo blanco $(|\rho| > 0{,}2)$.

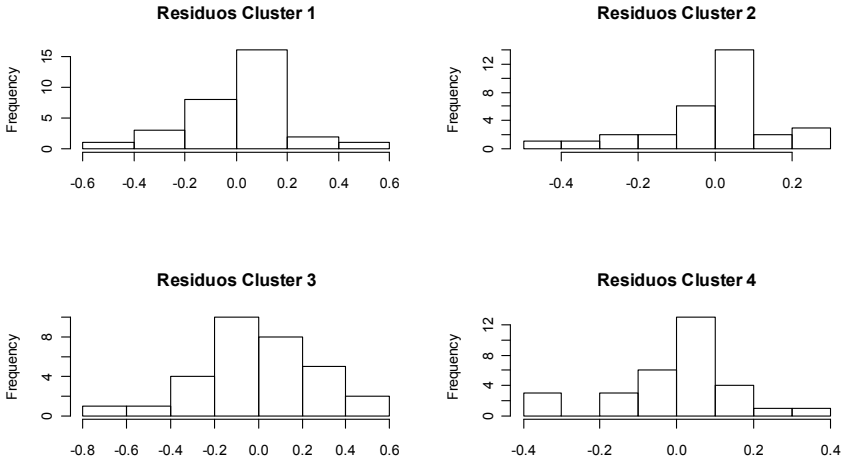
La columna de la derecha indica el signo de la correlación.

Cuadro 15: Correlación entre componentes principales y variables originales

-	**Componente 1**		+
+	V1	Cantidad de tickets que incluyen la categoría j	-
+	V3	Gasto total de todos los tickets que incluyen la categoría j	-
+	V2	Gasto total en la categoría j	-
-	V4	Monto promedio de los tk que incluyen la categoría j (V3/V1)	+
-	V6	Categoría diferentes que se incluyen en los tk que incluyen j	+
-	**Componente 2**		+
-	**p50**	**Día del mes que acumula el 50% del gasto de la categoría j**	+
-	p90	Día del mes que acumula el 90% del gasto de la categoría j	+
-	V5	Peso relativo del gasto de la categoría j en el total del gasto d elos tk que incluyen j (V2/V3)	+
-	p10	Día del mes que acumula el 10% del gasto de la categoría j	+
+	V6	Categoría diferentes que se incluyen en los tk que incluyen j	-
-	**Componente 3**		+
+	**p10**	**Día del mes que acumula el 10% del gasto de la categoría j**	-
-	**V5**	**Peso relativo del gasto de la categoría j en el total del gasto d elos tk que incluyen j (V2/V**	+

Anexo II

Análisis de los residuos de las regresión MCO del gasto total en categorías de productos del cluster contra los días del mes.



Shapiro-Wilk normality test

***Cluster* 1 (Basicos):**

> shapiro.test(lm(log(Gasto)~Dia,datos[datos$Clust==1,])$resid)
Shapiro-Wilk normality test
data: lm(log(Gasto) ~ Dia, datos[datos$Clust == 1,])$resid
W = 0.9513, **p-value = 0.1693**

***Cluster* 2 (Perecederos):**

> shapiro.test(lm(log(Gasto)~Dia,datos[datos$Clust==2,])$resid)
Shapiro-Wilk normality test
data: lm(log(Gasto) ~ Dia, datos[datos$Clust == 2,])$resid
W = 0.8828, **p-value = 0.002748**

***Cluster* 3 (Fiestas)**

> shapiro.test(lm(log(Gasto)~Dia,datos[datos$Clust==3,])$resid)
Shapiro-Wilk normality test
data: lm(log(Gasto) ~ Dia, datos[datos$Clust == 3,])$resid
W = 0.9878, **p-value = 0.9723**

***Cluster* 4 (Carne Vacuna):**

> shapiro.test(lm(log(Gasto)~Dia,datos[datos$Clust==4,])$resid)
Shapiro-Wilk normality test
data: lm(log(Gasto) ~ Dia, datos[datos$Clust == 4,])$resid
W = 0.9365, **p-value = 0.06624**

De las cuatro regresiones hay tres que tienen residuos con comportamiento normal. Para esas regresiones sí es posible extraer conclusiones sobre los coeficientes.

Anexo III

Tests de igualdad de varianzas *test* de medias

Tests de Igualdad de Varianzas

CanTick
statistic 34.26194
parameter 1
p.value 4.817109e-09
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
GsTotal
statistic 19.33485
parameter 1
p.value 1.096862e-05
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
GsToTik
statistic 18.00948
parameter 1
p.value 2.198073e-05
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
MonTick
statistic 0.6631297
parameter 1
p.value 0.4154572
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
PorCatg
statistic 5.371111e-05
parameter 1
p.value 0.9941525
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
CategDf
statistic 0.03999247
parameter 1
p.value 0.8414953
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
Perce10
statistic Inf
parameter 1
p.value 0
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
Perce50

statistic 1.215648
parameter 1
p.value 0.2702165
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
Perce90
statistic 0.1227011
parameter 1
p.value 0.7261232
data.name "x[datos$cl == 1 | datos$cl == 2] and datos$cl[datos$cl == 1 | datos$cl == 2]"
method "Bartlett test of homogeneity of variances"
CanTick
statistic 4.244617
parameter 1
p.value 0.03937494
data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
GsTotal
statistic 0.6035637
parameter 1
p.value 0.4372215
data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
GsToTik
statistic 4.173496
parameter 1
p.value 0.041061
data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
MonTick
statistic 0.005706343
parameter 1
p.value 0.9397848
data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
PorCatg
statistic 10.09087
parameter 1
p.value 0.001490061
data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
CategDf
statistic 2.895452
parameter 1
p.value 0.08882988
data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
Perce10
statistic 0.06066496
parameter 1
p.value 0.805448

data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
Perce50
statistic 0.0326373
parameter 1
p.value 0.856636
data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
Perce90
statistic 0.007571725
parameter 1
p.value 0.930659
data.name "x[datos$cl == 1 | datos$cl == 3] and datos$cl[datos$cl == 1 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
CanTick
statistic 16.23062
parameter 1
p.value 5.608032e-05
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
GsTotal
statistic 7.429557
parameter 1
p.value 0.006416114
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
GsToTik
statistic 12.33227
parameter 1
p.value 0.0004451949
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
MonTick
statistic 0.2197901
parameter 1
p.value 0.6391999
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
PorCatg
statistic 7.075572
parameter 1
p.value 0.007814186
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
CategDf
statistic 3.002176
parameter 1
p.value 0.08315274
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
Perce10

statistic Inf
parameter 1
p.value 0
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
Perce50
statistic 0.8580359
parameter 1
p.value 0.3542889
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"
Perce90
statistic 0.0953248
parameter 1
p.value 0.7575139
data.name "x[datos$cl == 2 | datos$cl == 3] and datos$cl[datos$cl == 2 | datos$cl == 3]"
method "Bartlett test of homogeneity of variances"

***Tests* de Igualdad de Medias**

Welch Two Sample t-test

data: datos$CanTick[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
t = -5.7768, df = 13.794, p-value = 5.082e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-13618.940 -6236.711
sample estimates:
mean in group 1 mean in group 2
2572.889 12500.714

Welch Two Sample t-test

data: datos$GsTotal[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
t = -6.0518, df = 14.71, p-value = 2.409e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-438616.1 -209837.6
sample estimates:
mean in group 1 mean in group 2
82800.96 407027.79

Welch Two Sample t-test

data: datos$GsToTik[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
t = -6.3, df = 14.846, p-value = 1.493e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3417159 -1688304

```
sample estimates:
mean in group 1 mean in group 2
         985864         3538595

                Two Sample t-test

data:  datos$MonTick[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
t = 5.5206, df = 39, p-value = 2.397e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  69.4963 149.8688
sample estimates:
mean in group 1 mean in group 2
       402.1111         292.4286

                Two Sample t-test

data:  datos$PorCatg[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
t = -3.3598, df = 39, p-value = 0.001754
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06344571 -0.01576064
sample estimates:
mean in group 1 mean in group 2
     0.08111111       0.12071429

                Two Sample t-test

data:  datos$CategDf[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
t = 6.8657, df = 39, p-value = 3.278e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.079268 1.980785
sample estimates:
mean in group 1 mean in group 2
       8.490741         6.960714

                Welch Two Sample t-test

data:  datos$Perce10[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
t = -3.0166, df = 26, p-value = 0.005652
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.43591887 -0.08259964
sample estimates:
mean in group 1 mean in group 2
       3.740741         4.000000

                Two Sample t-test

data:  datos$Perce50[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
```

```
t = -2.0534, df = 39, p-value = 0.04678
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.223574778 -0.009229455
sample estimates:
mean in group 1 mean in group 2
       15.74074        16.35714

            Two Sample t-test

data:  datos$Perce90[datos$cl == 1 | datos$cl == 2] by datos$cl[datos$cl == 1 | datos$cl == 2]
t = 0.3336, df = 39, p-value = 0.7405
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2143100  0.2989661
sample estimates:
mean in group 1 mean in group 2
       28.18519        28.14286

            Welch Two Sample t-test

data:  datos$CanTick[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = 3.5969, df = 20.453, p-value = 0.001752
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  588.7735 2209.0042
sample estimates:
mean in group 1 mean in group 3
       2572.889        1174.000

            Two Sample t-test

data:  datos$GsTotal[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = -0.7152, df = 31, p-value = 0.4799
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -82269.57  39552.17
sample estimates:
mean in group 1 mean in group 3
       82800.96        104159.67

            Welch Two Sample t-test

data:  datos$GsToTik[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = 3.3701, df = 20.211, p-value = 0.003011
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 176434.9 748629.9
sample estimates:
mean in group 1 mean in group 3
       985864.1        523331.7
```

Two Sample t-test

data: datos$MonTick[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = -2.2793, df = 31, p-value = 0.02969
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-124.214920 -6.896191
sample estimates:
mean in group 1 mean in group 3
402.1111 467.6667

Welch Two Sample t-test

data: datos$PorCatg[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = -3.2141, df = 5.334, p-value = 0.02155
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.2211359 -0.0266419
sample estimates:
mean in group 1 mean in group 3
0.08111111 0.20500000

Welch Two Sample t-test

data: datos$CategDf[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = 3.7613, df = 15.959, p-value = 0.001713
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.302801 1.085347
sample estimates:
mean in group 1 mean in group 3
8.490741 7.796667

Two Sample t-test

data: datos$Perce10[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = -2.1418, df = 31, p-value = 0.04018
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.83151849 -0.02033336
sample estimates:
mean in group 1 mean in group 3
3.740741 4.166667

Two Sample t-test

data: datos$Perce50[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = -6.1448, df = 31, p-value = 8.132e-07
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
 -3.675078 -1.843440
sample estimates:
mean in group 1 mean in group 3
       15.74074   18.50000

            Two Sample t-test

data: datos$Perce90[datos$cl == 1 | datos$cl == 3] by datos$cl[datos$cl == 1 | datos$cl == 3]
t = -5.4656, df = 31, p-value = 5.651e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.3477261 -0.6152369
sample estimates:
mean in group 1 mean in group 3
       28.18519   29.16667

            Welch Two Sample t-test

data: datos$CanTick[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = 6.6161, df = 13.572, p-value = 1.352e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7643.997 15009.431
sample estimates:
mean in group 2 mean in group 3
       12500.71   1174.00

            Welch Two Sample t-test

data: datos$GsTotal[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = 5.4136, df = 16.419, p-value = 5.249e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 184514.6 421221.6
sample estimates:
mean in group 2 mean in group 3
       407027.8   104159.7

            Welch Two Sample t-test

data: datos$GsToTik[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = 7.502, df = 14.305, p-value = 2.517e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2154938 3875590
sample estimates:
mean in group 2 mean in group 3
      3538595.4   523331.7

            Two Sample t-test

```
data:  datos$MonTick[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = -6.5067, df = 18, p-value = 4.06e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -231.8199 -118.6563
sample estimates:
mean in group 2 mean in group 3
      292.4286        467.6667

            Welch Two Sample t-test

data:  datos$PorCatg[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = -2.1548, df = 5.649, p-value = 0.07744
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.18145831  0.01288689
sample estimates:
mean in group 2 mean in group 3
     0.1207143       0.2050000

            Welch Two Sample t-test

data:  datos$CategDf[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = -3.6483, df = 17.699, p-value = 0.001881
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.3179366 -0.3539681
sample estimates:
mean in group 2 mean in group 3
      6.960714        7.796667

            Welch Two Sample t-test

data:  datos$Perce10[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = -1, df = 5, p-value = 0.3632
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5950970  0.2617636
sample estimates:
mean in group 2 mean in group 3
      4.000000        4.166667

            Two Sample t-test

data:  datos$Perce50[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = -5.2253, df = 18, p-value = 5.716e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.004436 -1.281278
sample estimates:
```

```
mean in group 2 mean in group 3
      16.35714        18.50000

            Two Sample t-test

data:  datos$Perce90[datos$cl == 2 | datos$cl == 3] by datos$cl[datos$cl == 2 | datos$cl == 3]
t = -5.5772, df = 18, p-value = 2.71e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4094799 -0.6381391
sample estimates:
mean in group 2 mean in group 3
      28.14286        29.16667
```