Aplicación de Machine Learning a Predicción de Demanda

Primer Informe de Análisis y Visualización

6 DE JUNIO DE 2021

Integrantes:

Gutiérrez Montecino, Denise

Nieva, Sofía

Rodríguez, Alfredo Manuel

Contenido

1 INTRODUCCIÓN

En el siguiente informe se detallan los resultados obtenidos del análisis y visualización de la base de datos suministrada como parte de la Mentoría ***Predicciones de Demanda de Producto*** de la Diplomatura de Ciencia de Datos de la Facultad de Matemáticas, Astronomía y Física (FaMAF) de la Universidad Nacional de Córdoba, Argentina.

El conjunto de datos de partida contiene productos vendidos de los últimos 5 años de una compañía de venta de alimentos congelados en distintos países de la región. El objetivo final de la mentoría es poder predecir la demanda de los productos elaborados en los centros de elaboración mes a mes en los diferentes países y zonas en donde opera esta compañía.

El presente entregable contiene la presentación del dataset, la documentación del análisis exploratorio, resultados y visualizaciones de las consignas presentadas en la sección ***Respuestas a Preguntas Disparadoras***.

2 PRESENTACIÓN DEL DATASET

El set de datos corresponde al registro de ventas de la casa central y centros de producción de una empresa de alimentos congelados a los distintos puntos de venta de su franquicia en Argentina y otros países de la región.

Variables

El set de datos se compone a partir de 5 set de datos, que contienen las siguientes variables:

Productos: información de los productos vendidos. Contiene:

- sku: identificador de producto
- descripción: descripción del producto enmascarado
- marca: marca del producto
- id categoría: identificador de categoría a la cual pertenece el producto
- presentación: forma de presentación de producto (pack, kilogramos, etc)
- unidadcm3: cm3 del producto
- unidadkg: kg del producto
- id proveedor: identificador del proveedor del producto

Categorías: información de categoría de productos vendidos. Contiene:

- id categoría: identificador de categoría.
- nombre: nombre de la categoría

Puntos de venta: información de puntos de ventas. Contiene:

- id_punto_venta: identificador de punto de venta
- nombre: nombre de punto de venta enmascarado
- id_provincia: identificador de la provincia a la que pertenece el punto de venta
- id_localidad: identificador de la localidad a la que pertenece el punto de venta
- id_pais: identificador del pais al que pertenece el punto de venta

Ventas: información de ventas. Contiene:

- dia: dia de la venta
- mes: mes de la venta
- año: año de la venta
- hora: hora en que se produzco la venta

- sku: identificador de producto
- cantidad pedida: cantidad comprada por punto de venta
- id_punto_venta: identificador de punto de venta

Países: información de países

- id pais: identificador de país
- nombre: nombre de país
- Provincias: información de provincias. Contiene:
- id provincia: identificador de provincia
- nombre: nombre de provincia
- Localidades: información de localidades. Contiene:
- id localidad: identificador de localidad
- nombre: nombre de la localidad

El dataset final es un objeto tipo pandas.DataFrame: **'ventas_producto_pdv'**. Se obtiene por unión [.merge()] de los datasets mencionados anteriormente, y contiene las siguientes variables (25 en total):

dia	int64
mes	int64
anio	int64
hora	object
sku	int64
cantidad_pedida	float64
id_punto_venta	object
fecha	datetime64[ns]
descripcion	object
marca	object
id_categoria	float64
presentacion	object
unidadcm3	float64
unidadkg	float64
id_proveedor	float64
Categoria	object
Punto_Venta	object
id_Provincia	float64
id_Localidad	float64

id_Pais	float64
Localidad	object
Provincia	object
Pais	object

Para el análisis de predicción de demanda se incorpora la variable **‘totalkg’** como el producto entre ‘**cantidad_pedida’** y ‘**unidadkg**’. Esta se establece como una de las variables relevantes como parte de los objetivos específicos.

Se crea la variable ‘**time: %y.%m**’ para resumir los datos temporales en una única variables. Dado que el el objetivo es una predicción de forma mensual solo se incorporan los datos de Año y Mes.

Análisis Exploratorio

El dataset contiene 5,666,365 entradas. Si hacemos una detección de los valores NaN, se puede ver rápidamente que existe todo un subset de datos que tiene valores faltantes para las variable de ubicación ***‘id_Provincia’ ‘id_Localidad’ ‘id_Pais’ ‘Localidad’ ‘Provincia’*** y ***‘Pais’*** (Fig 1, Fig 2). Existe una correlación de estos datos faltantes, de tal manera que las pérdidas son sistemáticas: donde hay una pérdida en una de las variables de ubicación también están faltantes el resto. Según la información de dominio los datos faltantes corresponden a las ventas fuera de Argentina y no hay forma de imputar las pérdidas por lo que se eliminan del set de datos. En total las pérdidas afectan a 667,024 entradas (11.78%), quedando el dataset sin datos faltantes con 4,999,341 filas.

Para la variable ‘***totalkg***’ existen:

- 2 valores negativos
- 269,923 valores nulos

La ocurrencia de valores nulos está asociada a productos que tienen valor nulo para ***‘unidadkg’*** o ***‘cantidad_pedida’***. Tanto valores negativos como valores nulos son no informativos a los fines del análisis por lo que se eliminan del dataset (4.76%). Los datos no afectados son 4,760,848 entradas de ventas.

En línea con el análisis anterior muchos productos que tienen valores nulos, sumamente pequeños o extremadamente grandes en ***‘unidadkg’*** estuvieron asociadas a ciertas categorías de productos. Analizando la variable ***‘Categoria’*** se detectaron categorías de productos no comestibles. Dichos productos presentan una alta densidad de dichos datos anómalos en ***‘unidadkg’***. Las categorías mencionadas son: ***'0010 - PRODUCTOS PROMOCIONALES','0001 - GRIDO MARKET','0006 - PRODUCTOS COMPOSTABLES /BIODEGRADABLES' ,'07 - ACCESORIOS, UTENSILIOS Y REPUESTOS ','0005 - ENVASES TERMICOS Y VASOS', '06 - PRODUCTOS DE LIMPIEZA ', '0006 - EQUIPOS', '0007 - ACCESORIOS Y REPUESTOS PARA EQUIPOS DE FRIO', '0008 - MUEBLES', '09 - DISPENSER', '0005 - CARTELERIA, INSTITUCIONALES Y PLOTEOS', '0010 - PACKAGING', '0001 - INDUMENTARIA ', '0009 - DESCARTABLES ', '0012 - MATERIAL MKT', '19 - MATERIAS PRIMAS'***.

Dado que no aportan al objetivo del estudio, ya que no son productos elaborados en los centros de producción, se quitaron del dataset, lo que afecta a 453,302 (8%). Quedan por tanto 4,307,546 entradas de ventas.

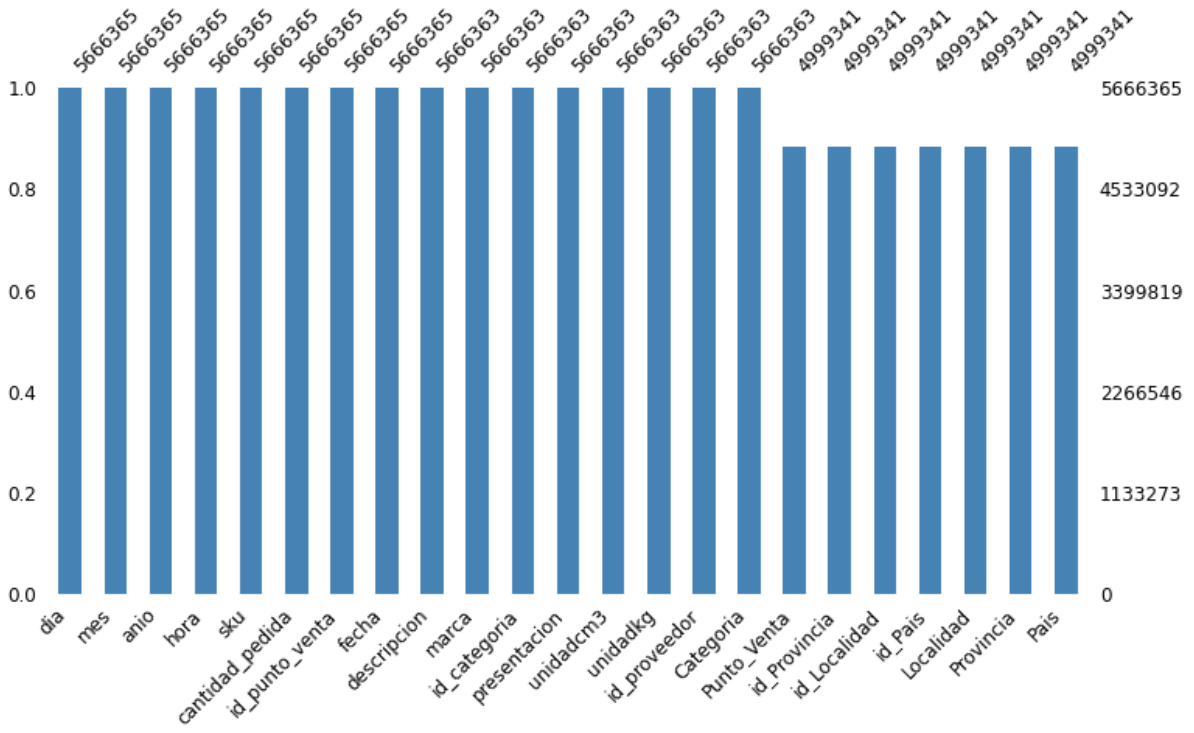


Figura 1 - Detección de Valores Faltantes: missingno.bar()

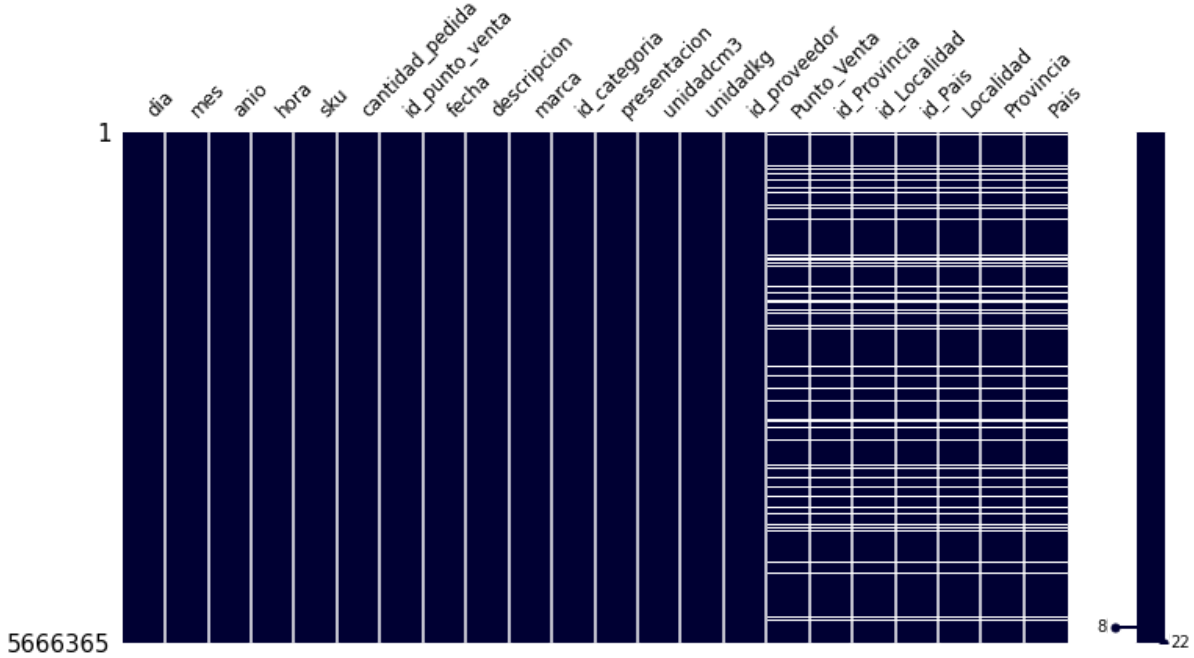


Figura 2 - Detección de Datos Faltantes: missingno.matrix()

Valores Extremos y Anómalos

Tomando la variable ***'totalkg'*** se probaron diferentes métodos de eliminación de valores extremos considerados anómalos, mejor conocidos como "*outliers*". Los métodos ensayados son Z-score, Z-score modificado, por rango intercuartílico (IQR), y el 2% de valores más extremos por percentiles.

Teniendo en cuenta la gran asimetría de la distribución de los datos se descartó el empleo de los dos primeros métodos al estar basados en una distribución gaussiana. Optando por los métodos no paramétricos la proporción de muestra que queda fuera del dataset una vez eliminados valores extremos es:

- IQR: 8,7%
- Rango Percentil 1%-99% : 2%

Es difícil determinar si los valores extremos hacia la derecha de la distribución pueden considerarse o no anómalos ya que estamos tratando con compras realizadas por puntos de ventas de la franquicia que bien pueden hacerse por cantidades pequeñas o grandes dependiendo del tipo de movimiento de stock que tenga cada uno y la estación del año. Por esto consideramos que el primero de estos métodos elimina datos que deberían quedar dentro del dataset (Fig 3).

El método para eliminación de outliers seleccionado implica eliminar todos los valores que se encuentren por debajo del percentil 1% y por encima del percentil 99%. De esta forma controlamos la proporción de la muestra que decidimos eliminar que corresponde a un 2% del dataset filtrado[1].

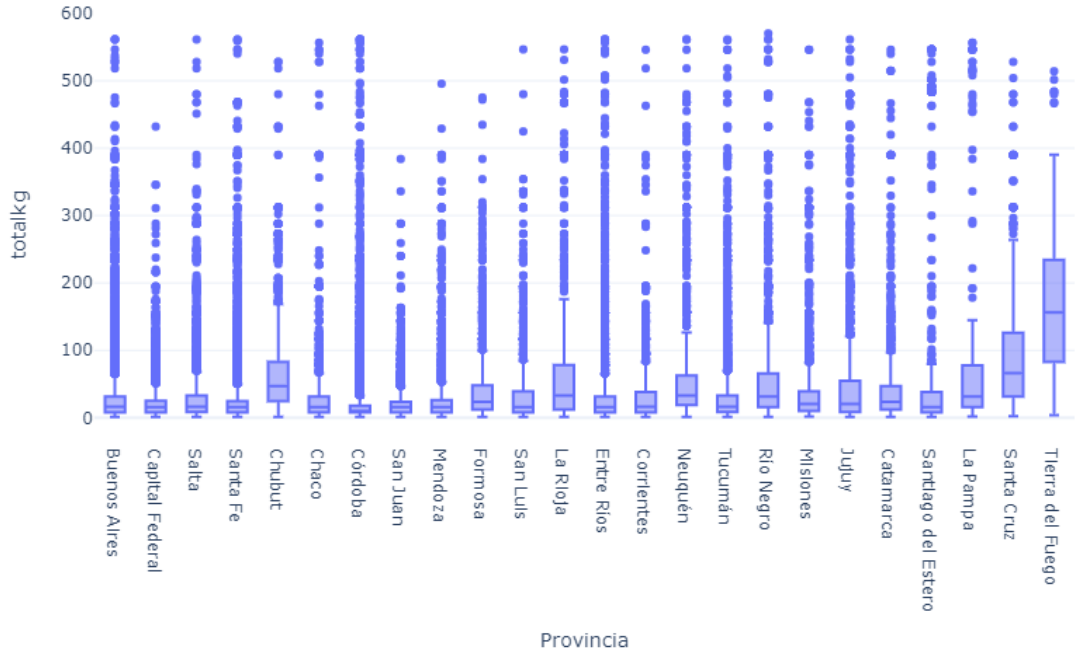


Figura 3 - Distribución de los datos ***'totalkg'*** después de la eliminación de valores extremos.

La estadística descriptiva del nuevo set de datos es:

- Tamaño: 4,219,560
- Media: 26.32kg
- DesvEst: 38.91kg
- Min: 1.04kg
- Q1: 7.8kg
- Mediana: 1.56kg
- Q3: 30.00kg
- Max: 574.46kg

[1] Este método se propuso aplicar sobre la distribución de ***'totalkg'*** agrupada por mes y producto ya que el análisis posterior se hace a escala mensual. Queda pendiente revisar esta hipótesis.

Variables de Interés

El set de datos final cuenta con 4,219,560 entradas de ventas sobre las que se realizan los análisis de la siguiente sección. En total se eliminó un 25.51% de la muestra original.

Dentro de las columnas que se seleccionan como variables de interés son [***'time','Localidad','Provincia','sku','totalkg'***]. Las métricas de estas variables son:

'time' : Rango 2018.01-2019.12

'Localidad' : recuento 648

'sku': recuento 217

'Provincia' : frecuencias absolutas (Fig 4)

Tierra del Fuego	2535
La Pampa	4088
Santa Cruz	10751
Santiago del Estero	13616
La Rioja	21903
Formosa	33442
Neuquén	33674
Catamarca	34438
Chubut	34781
Río Negro	39147
Jujuy	48993
Chaco	51177
Misiones	61086
San Luis	65687
Corrientes	73059
San Juan	88874
Salta	93501
Tucumán	121545
Capital Federal	151182
Entre Ríos	175780
Mendoza	188540
Santa Fe	458002
Córdoba	1091309
Buenos Aires	1323551

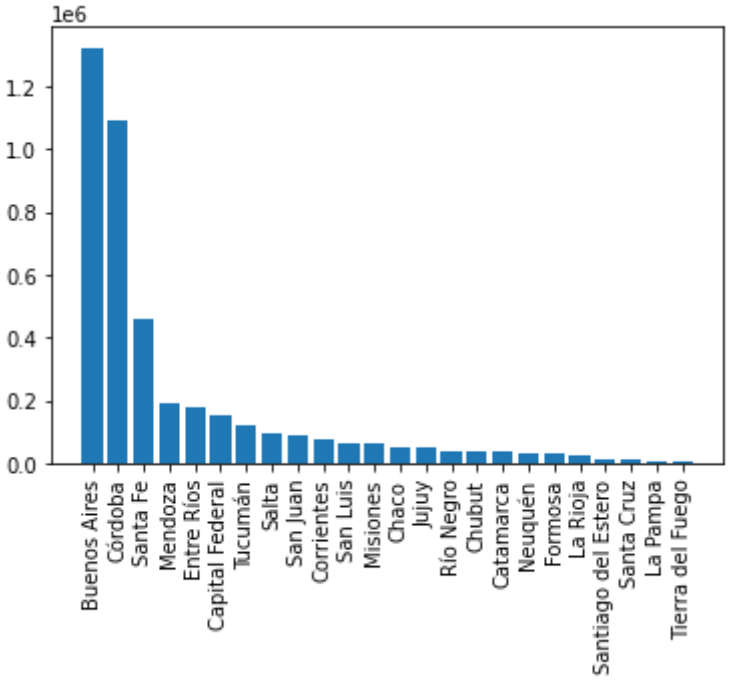


Figura 4 - Distribución de ventas entre las provincias de Argentina

3 RESPUESTAS A PREGUNTAS DISPARADORAS

En esta sección se desarrollarán las consignas dadas para el informe con sus respectivas observaciones y comentarios de los resultados obtenidos así como también de las decisiones que se fueron realizando, si fuese el caso. A continuación se describe cada una de ellas.

1) ¿Cuáles son las 3 provincias que más kilogramos solicitan por mes?

En primer lugar se realizó un análisis global de cuáles eran las tres provincias que solicitaron más kilogramos para todo el período de estudio (años 2018 y 2019). Como resultado se obtuvo que las provincias de Buenos Aires, Córdoba y Santa Fe, en el orden mencionado, fueron las provincias que más solicitaron. Luego, se procedió a realizar el análisis de aquellas provincias que más solicitaron para cada mes (Fig 5).

Como primer hallazgo se puede apreciar una clara tendencia estacionaria, en que los meses de verano o próximos a dicha estación hay una demanda de productos mayor en relación a los meses de invierno o próximos a este. También se observa que en promedio la demanda de productos fue mayor para el año 2019 en relación a lo demandado para el año 2018.

Como se observa las provincias que más solicitan mes a mes corresponden a Buenos Aires, Córdoba y Santa Fe coincidiendo con el análisis global realizado para todo el período comprendido. En que las proporciones de representación se mantienen a lo largo de los dos años de análisis, en otras palabras, la provincia de Buenos Aires es la que más solicita independiente de la estación en la que se encuentra.

De los resultados obtenidos se encuentra coherencia dado que estas tres provincias son las que mayor población de personas concentran en el país y hay una clara relación entre la demanda y la cantidad de personas.

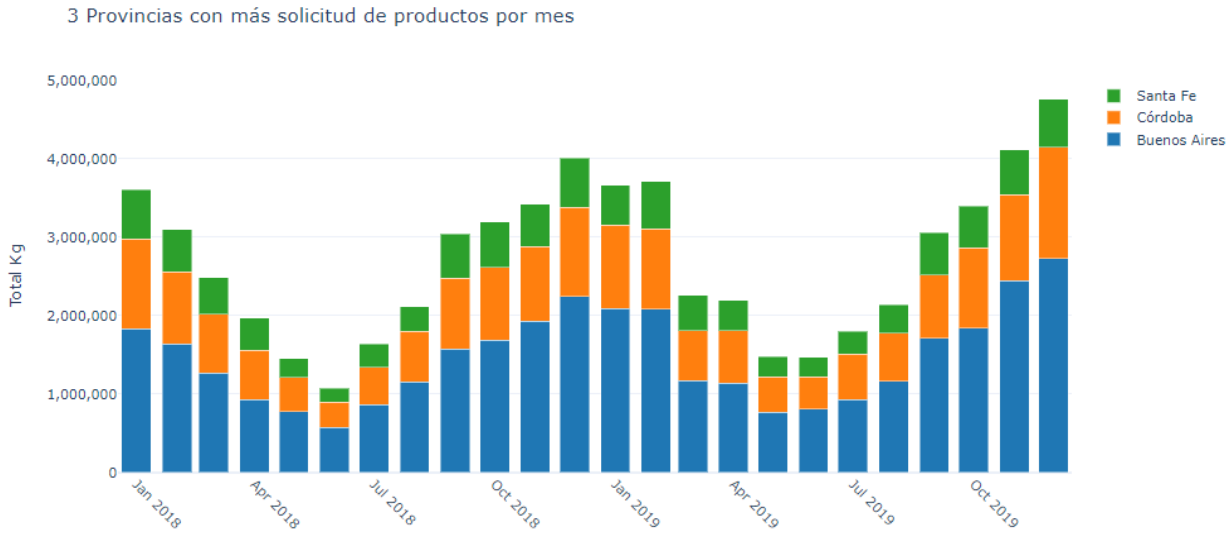


Figura 5 - Tres provincias que más ***'totalkg'*** solicitaron por mes

2) ¿Cuáles son las 3 localidades que más kilogramos solicitan por mes?

Las localidades de Córdoba, Salta y Santa Fe fueron las localidades que máyor valor acumulado de totalkg solicitaron en los dos años de análisis. Como siguiente paso se analizó cuáles fueron las tres localidades que más solicitaron para cada mes (Fig 6).

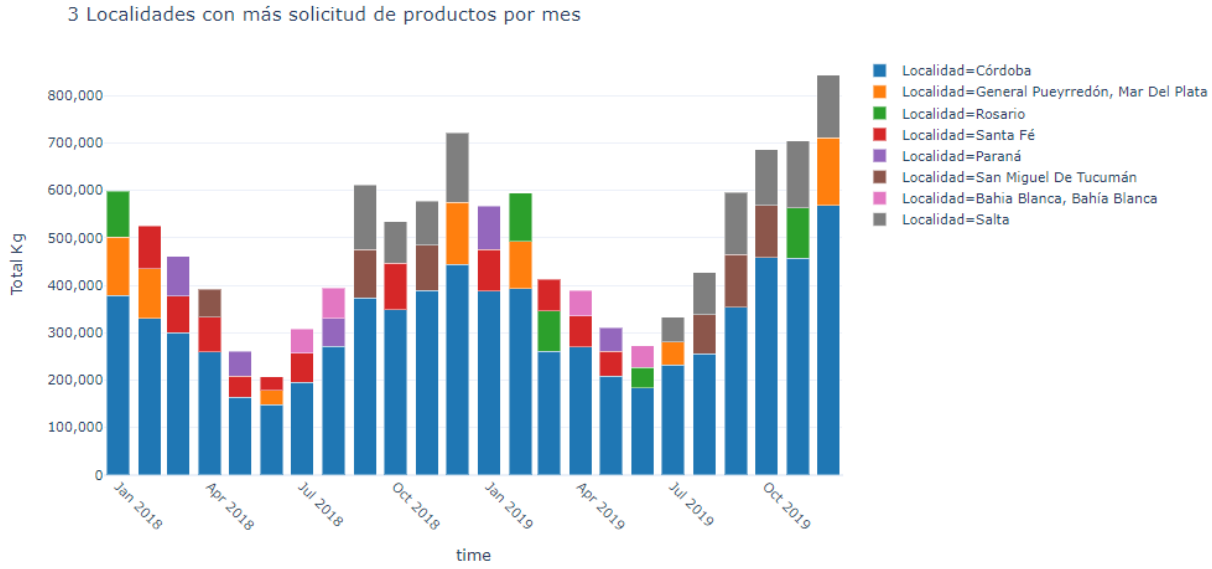


Figura 6 - Tres localidades que más ***'totalkg'*** solicitaron por mes

Al igual que en el punto anterior se observa una clara tendencia estacionaria en que los meses de veranos o próximos a este poseen una demanda mayor de productos en relación a los meses de invierno o próximos al mismo. También se aprecia que en promedio el año 2019 tuvo una mayor demanda de productos en relación al año 2018.

A diferencia de lo que sucedía con las provincias se puede apreciar que las tres principales localidades solicitantes de productos no son las mismas para todo el período de análisis comprendido en el estudio, con excepción de la localidad de Córdoba que no sólo se mantiene para todos los meses sino que también es la localidad que mayor demanda de productos posee en todos los meses respecto a las otras dos localidades. Otra observación

en relación con el punto anterior es que a pesar de que Córdoba es la localidad que mayor demanda posee, en el punto anterior la provincia de Córdoba ocupaba el segundo lugar detrás de la provincia de Buenos Aires como la más demandante. Es un resultado coherente dado que la provincia de Buenos Aires es mayor a nivel poblacional que la provincia de Córdoba, pero esta población está distribuida entre más localidades.

En cuanto a las demás localidades que completan la terna de los primeros lugares no existen claras tendencias para ninguna de ellas. A continuación se describen algunos comentarios de la demanda de cada localidad:

- General Pueyrredón (Mar del Plata): aparece en algunos meses de verano y de invierno siendo la primera estación la que concentra mayor demanda en relación con la de invierno cumpliendo con la estacionalidad ya mencionada.
- Rosario: sólo posee presencia en enero para el año 2018 mientras que para el año 2019 se observa una fuerte demanda en los meses de verano y en junio del mismo, también siguiendo la estacionalidad descrita anteriormente.
- Paraná: presenta demanda en los meses de otoño e invierno para el 2018 y verano y otoño para el 2019.
- Santa Fe: tiene un mayor peso en los últimos meses de verano y los meses de otoño.
- San Miguel de Tucumán: tiene una aparición más marcada en los meses de primavera y en menor medida en abril del año 2018.
- Bahía Blanca: aparece en los meses de invierno para ambos años y en abril para el año 2019.
- Salta: aparece con mayor preponderancia en los meses de primavera y los últimos meses de invierno para el año 2019.

Se observan sólo dos localidades: San Miguel de Tucumán y Salta, que no pertenecen a las tres provincias con mayor demanda de productos según el punto anterior.

3) ¿Cuáles son las 3 categorías que más kilogramos solicitan los clientes por mes?

Las categorías resultantes fueron: 07 - Sabores Comunes, 08 - Sabores Especiales y 17 - Pote 1 lts. Luego se procedió a realizar el análisis a nivel mensual (Fig 7).

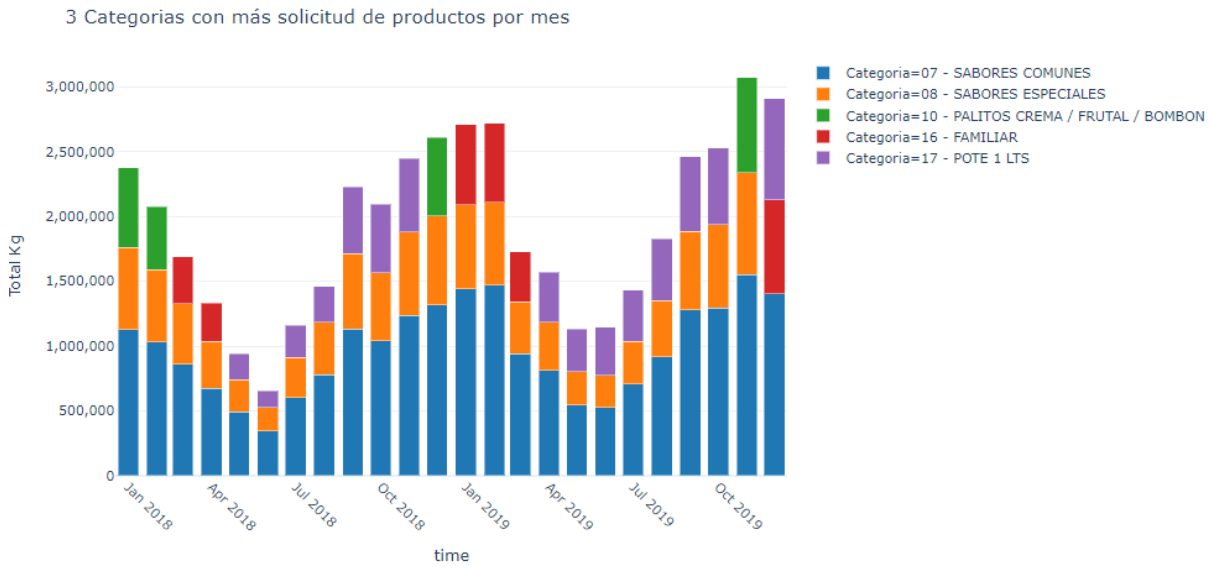


Figura 7 - Tres categorías que más ***'totalkg'*** solicitaron por mes

De la misma manera que sucedía con los dos puntos anteriores se aprecia una clara estacionalidad y una mayor demanda en promedio para el año 2019. Se observa que la terna de las tres categorías que obtuvieron más demanda para los años 2018 y 2019 no se mantiene constante para todos los meses.

La categoría que aparece en todos los meses y que ocupa el primer lugar de las tres categorías que más solicitudes tuvieron por parte de los clientes corresponde a la categoría 07 - Sabores Comunes, como habíamos obtenido en el análisis global realizado.

La categoría 08 - Sabores especiales ocupa el segundo lugar para todo el año 2018, mientras que para el año 2019 esta categoría ocupa el tercer puesto detrás de la categoría 17 - Pote 1 lts para los meses comprendidos entre mayo y agosto (inclusive) y en diciembre no figura dentro de las tres categorías con más solicitud.

En general la categoría 17 - Pote 1 lts figura principalmente a partir de abril en adelante. La categoría 10 - Palitos Crema/ Frutal/ Bombón y 16 - Familiar aparece principalmente en los meses de verano, en donde el tercer lugar se va alternando entre ambas categorías.

4) ¿Existen productos que no se venden en todas las localidades?

Sí, existen productos que no se venden en todas las localidades. La localidad que más productos vende o, en otras palabras, que posee mayor variedad de productos corresponde a la Ciudad de Córdoba que posee 213 tipos de productos para comercializar. Mientras que la localidad de Hipólito Yrigoyen sólo comercializa 1 producto de un total de 217 productos.

5) ¿Existen productos que no se venden en todas las provincias?

Sí, existen productos que no se venden en todas las provincias. La provincia que más productos vende o, en otras palabras, que posee mayor variedad de productos corresponde a la provincia de Córdoba que posee 214 tipos de productos para comercializar. Mientras que la provincia de Tierra del Fuego es la que menos variedad posee con 115 productos de un total de 217 productos.

6) ¿Cuáles son los 3 productos que más kilogramos solicitan los clientes por mes? Realice este análisis por localidad y provincia agrupando los pedidos por mes.

En primer lugar, se realizó el análisis global a nivel país de cuáles son los tres productos que más kg se solicitan mes a mes (Fig 8)

Se pueden hacer varias observaciones en base al gráfico. Como siempre es clara la estacionalidad, hay muchos más pedidos en verano que en invierno. Sin embargo, para el producto mas pedido a nivel, el 334, se ve que se pide una cantidad importante a lo largo de todo el año, incluso en invierno no disminuye tanto como disminuye la cantidad del segundo producto mas pedido a nivel global, el 1689. También se puede observar que hay productos que logran entrar entre los tres más vendidos solo en verano, como el 320 (que pertenece a la categoría 10- PALITOS CREMA/FRUTAL/BOMBON) y otros solo en invierno, como el 1872 (pertenece a la categoría 01-FRIZZIO, que corresponde a las pizzas congeladas que no tienen tanta estacionalidad como el helado). Se puede ver también que en la mayoría de los meses correspondientes a otoño y primavera, cuando la demanda es intermedia, la composición de los tres productos más vendidos suele ser 334, 1689, 888, que son justamente los tres más vendidos a nivel global. Luego, se procedió a realizar el mismo análisis, pero ahora por provincia y por localidad (Fig 9, Fig 10).

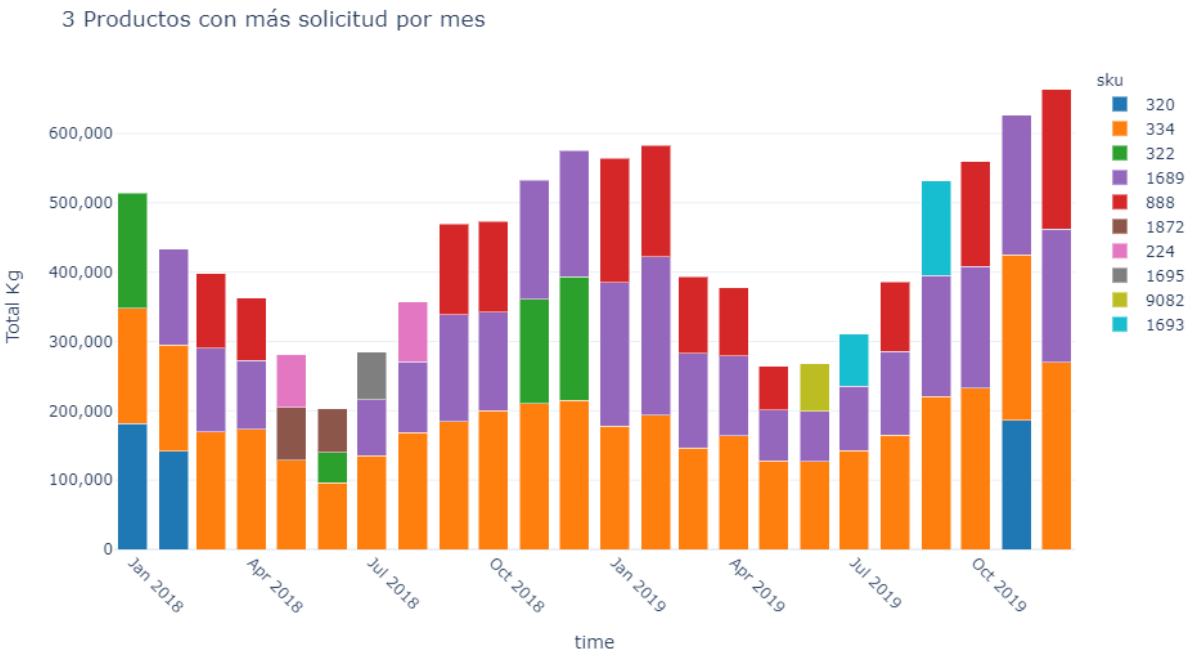


Figura 8 - Tres productos que más ***'totalkg'*** se solicitaron por mes

Se puede apreciar que las composición de los tres productos más vendidos mes a mes no es la misma para todas las provincias. Sin embargo el producto más pedido, el 334, si está presente en todas las provincias en casi todos los meses (principalmente no está presente en el verano en Córdoba). El segundo producto más vendido, el 1689, también está presente todos los meses en Córdoba y Santa Fe, y en Buenos Aires mayormente en el 2019. El tercer producto más vendido varía mucho de acuerdo a la estación y a la provincia.

Se puede observar que es bastante distinta la composición de los tres productos más vendidos de cada localidad. O sea, cuando miramos al nivel de localidades, éstas parecen ser bastante heterogéneas, no siguen todas la misma tendencia. En Salta por ejemplo, no está presente el producto más vendido y en varios meses ganan los productos 888, 889, 890, que pertenecen todos a la misma categoría 16 - FAMILIAR (baldes de helado). En cambio en Santa Fe, en la mayoría de los meses ganan los productos 1689, 1695 que pertenecen a la categoría más vendida 07 - SABORES COMUNES. Se puede decir entonces, que no solo cambian los productos, sino que también las categorías preferidas por los clientes de acuerdo a la localidad.

En Córdoba Capital la distribución es muy similar a la de la provincia de Córdoba, claramente tiene mucho peso porque concentra a una parte importante de la población. Con Santa Fe ocurre lo mismo y probablemente sea así en general para las capitales de cada provincia ya que estas suelen concentrar gran parte de la población

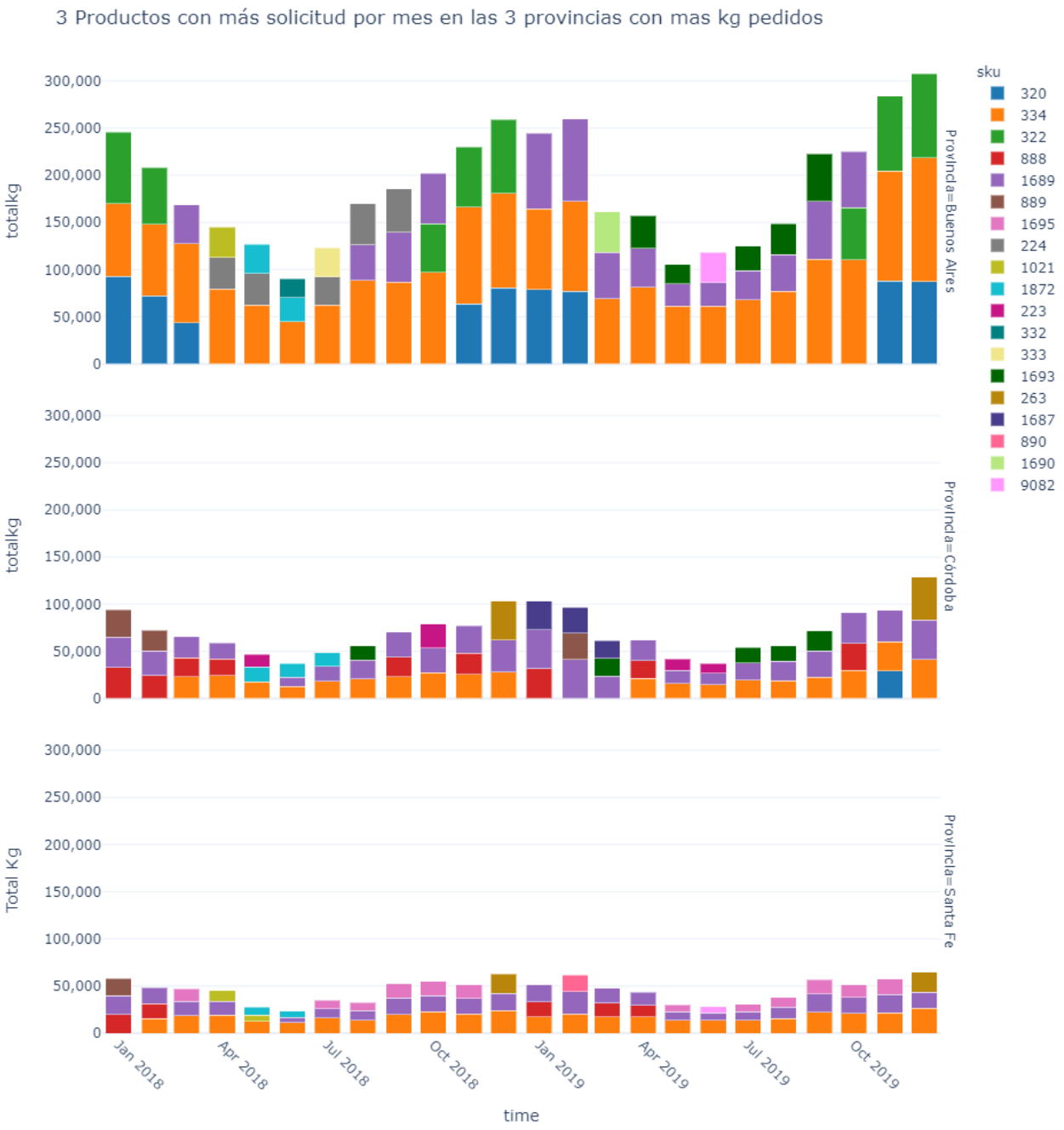


Figura 9 - Tres productos que más ***'totalkg'*** se solicitaron por mes en las tres provincias con mayor cantidad de pedidos.

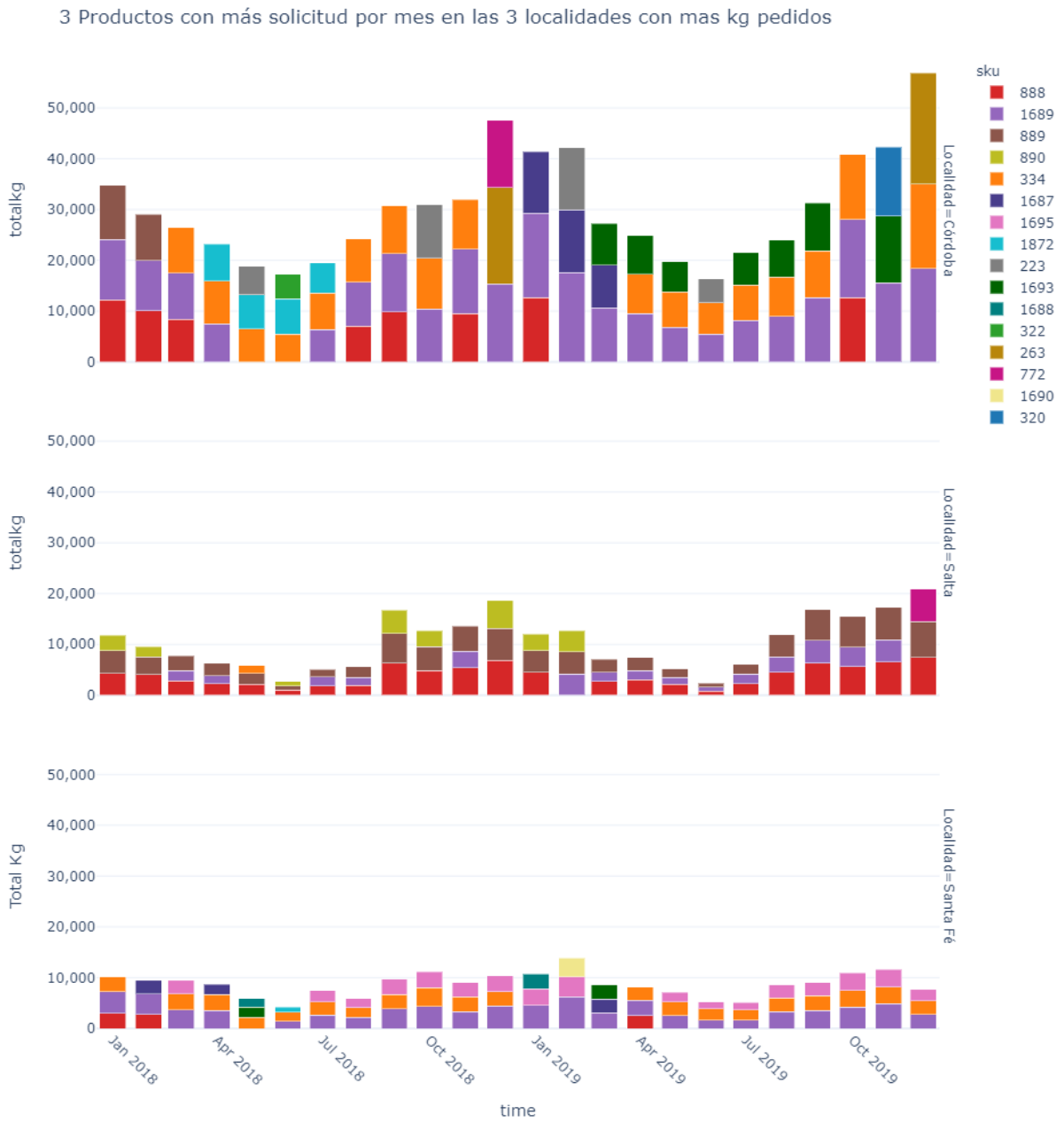


Figura 10 - Tres productos que más ***'totalkg'*** se solicitaron por mes en las tres localidades con mayor cantidad de pedidos.

7) ¿El producto que más se pide pertenece a la categorías que más se vende?

Tal como se anticipó en el punto 3 se conoce que las categorías más vendidas son: 07 - Sabores Comunes, 08 - Sabores Especiales y 17 - Pote 1 lts, siendo la primera categoría mencionada la correspondiente a la más vendida.

Por otro lado, el producto que posee más ventas para el período de análisis en estudio corresponde al sku 334. Dicho producto pertenece a la categoría 13 - Impulsivos.

Por lo tanto, se concluye que el producto que más se pide no pertenece a la categoría que más se vende.

8) ¿El producto que más se pide se vende en todas las localidades?

Siguiendo la línea del punto anterior, es de conocimiento que el producto que más se vende corresponde al sku 334. Dicho producto figura en 621 localidades de un total de 657 localidades presente en la base de datos.

Por lo tanto, el producto que más se pide no se vende en todas las localidades ya que existen 36 localidades en las que no se vende este producto.

9) Seleccionar 4 variables que consideren de interés para el objetivo del proyecto. Una de las variables debe ser "totalkg".

Las cuatro variables que se seleccionaron para el análisis son las siguientes:'totalkg','sku', 'Ubicación' siendo la combinación de las antiguas variables de 'Localidad' y 'Provincia' y 'time'. Con este set de variables es posible dar respuesta a los objetivos planteados en la mentoría.

Se necesita al menos una variable de tiempo para poder generar una predicción mes a mes. Una variable de ubicación geográfica para darle un contexto regional y poder atenderlo desde los centros de elaboración correspondientes. Esta última se decidió construirla a partir de los datos de localidad y provincia para evitar redundancias: existen distintas localidades con el mismo nombre. Y dos variables de identificación de ventas, que indican los kilogramos vendidos y a que producto corresponden, que funcionan como el núcleo de información respecto de cómo se comporta la demanda de los productos elaborados por la empresa.

10) Para las variables de interés seleccionadas indicar qué tipo de variable es cada una.

Para el análisis la variable 'time' se toma como una variable numérica discreta que toma valores [1,12] asociados a los meses del año, siendo los años considerados 2018-2019.

La variable 'Ubicación' es una variable categórica nominal. No se tuvieron en cuenta datos de geolocalización.

La variable 'totalkg' es una variable numérica discreta que dado el rango de valores que abarca puede suponerse contínua.

La variable 'sku' es una variable categórica nominal.

11) De las 3 categorías más pedidas de productos, analizar la dispersión de totalkg de dichas categorías. Realizar este análisis también no solo a nivel nacional sino también por provincia (las 3 que más piden) y por localidad (las 3 localidades que más piden).

La cantidad de kilogramos pedidos a nivel nacional para los productos de las tres categorías con mayor prevalencia en la muestra ('07 - SABORES COMUNES','08 - SABORES ESPECIALES','16 - FAMILIAR') sigue un patrón estacional teniendo valores mayores en los primeros y últimos meses, que corresponden a la época estival. La categoría 07 tiene valores marcadamente mayores comparado con las otras dos (Fig 11).

El mismo análisis realizado para las primeras tres provincias en ventas, Buenos Aires, Córdoba y Santa Fe, muestra un patrón similar. En el caso de Buenos Aires se observa al igual que a nivel nacional que la línea para la categoría 07 se mantiene por encima de las demás de forma más marcada que en las otras provincias. Dado que Buenos Aires es la provincia que más kilogramos demanda por mes de este tipo de productos es posible sea la demanda de esta provincia la principal responsable del patrón que se observa a nivel nacional (Fig 12).

Haciendo el análisis a nivel de localidad es interesante marcar que la estacionalidad de la demanda parece ser mucho menos marcada en las localidad de Salta, que en la ciudad de Córdoba. Esto puede tener una correlación con las condiciones climáticas y las temperaturas de la época invernal vs época estival, o algún aspecto cultural. De nuevo solo en Córdoba se logra ver como la línea que corresponde a la categoría 07 se desprende por encima de las otras dos (Fig 13).

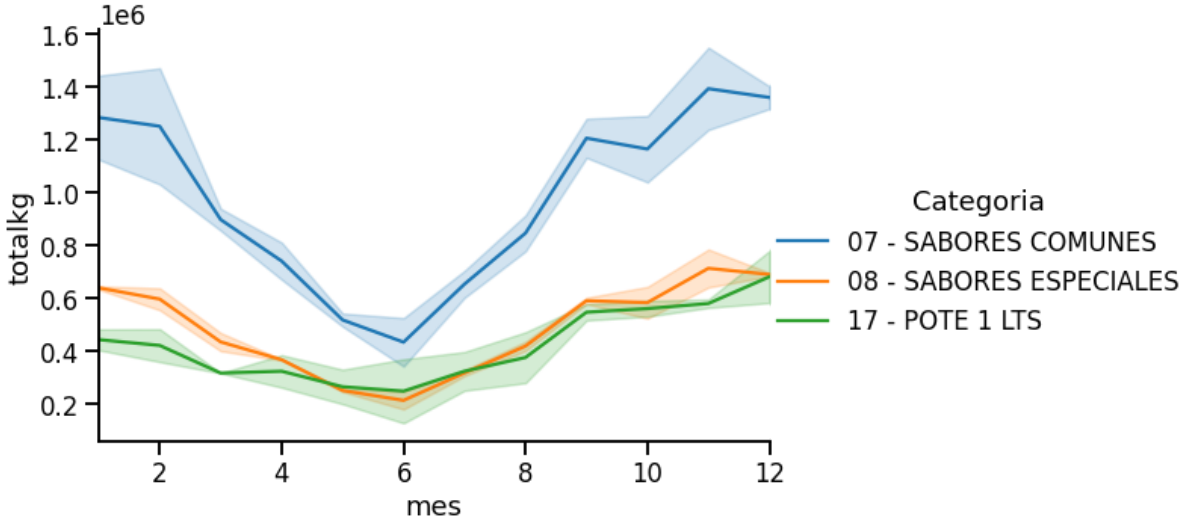


Figura 11 - Pedidos por mes a nivel nacional para las categorías más prevalentes

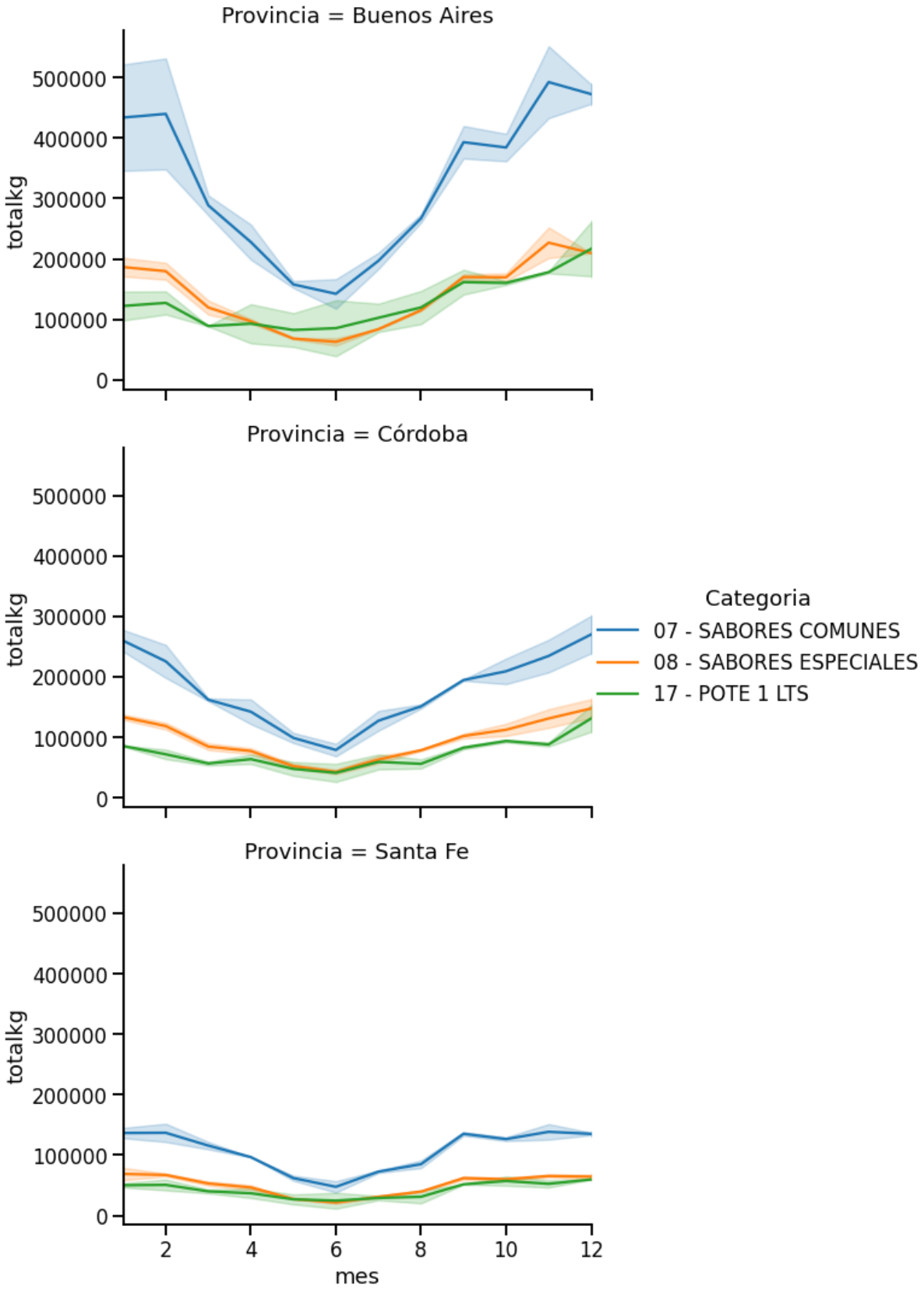


Figura 12 - Pedidos por mes para las categorías más prevalentes en las 3 provincias más con mayor volumen de pedidos

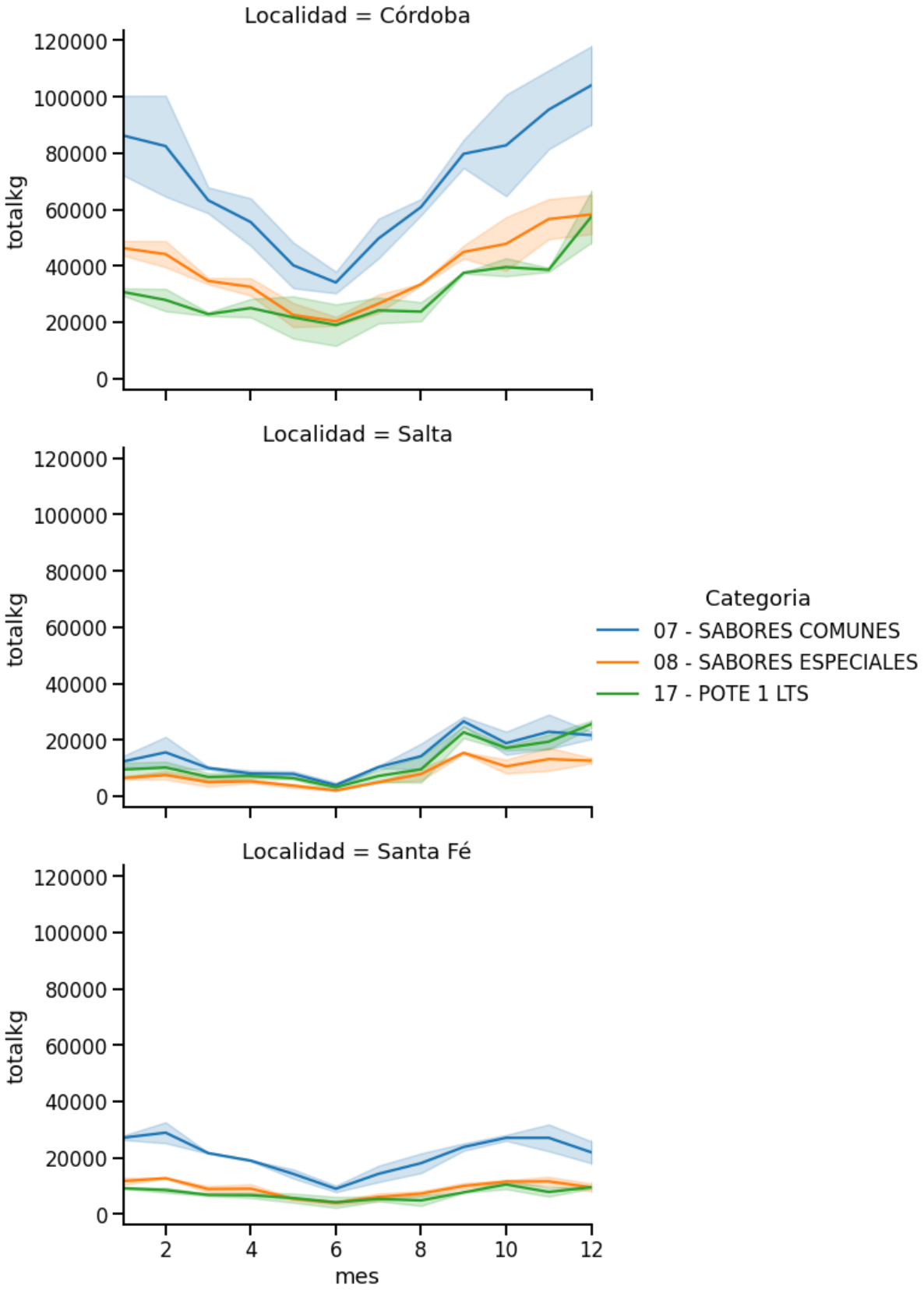


Figura 12 - Pedidos por mes para las categorías más prevalentes en las 3 localidades más con mayor volumen de pedidos

12) Determine si hay outliers en las variables seleccionadas. Realice este análisis por distintas variables, como por ejemplo totalkg por provincia cada mes.

En primer lugar, se realizó un análisis a nivel Provincial para determinar la existencia o no de outliers para cada una de estas. Como resultado se obtuvo la siguiente figura que se muestra a continuación (Fig 14):

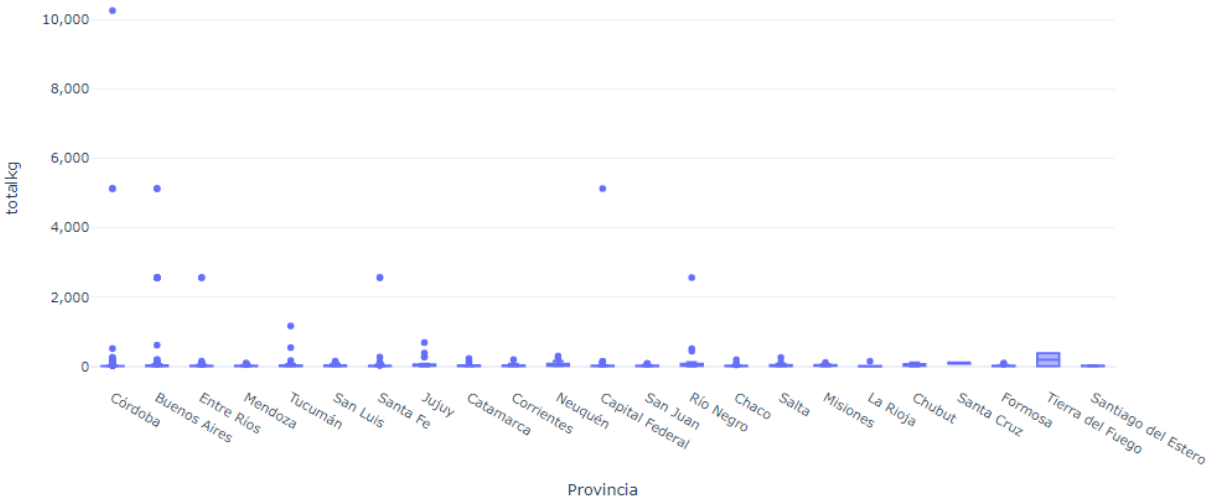


Figura 14 - Boxplot de ***'totalkg'*** por Provincia (con outliers)

En el boxplot presentado se puede observar a grandes rasgos varios outliers para algunas provincias como Córdoba, Buenos Aires, Entre Rios, Tucumán, Santa Fe, Capital Federal y Rio Negro. En que los valores de totalkg alcanzaban cifras por encima de los 2,000 para cada provincia mencionada.

Tras la eliminación de los outliers según el criterio ya explicado en la sección anterior se puede observar los siguientes cambios en la figura, tal como se muestra a continuación (Fig 15):

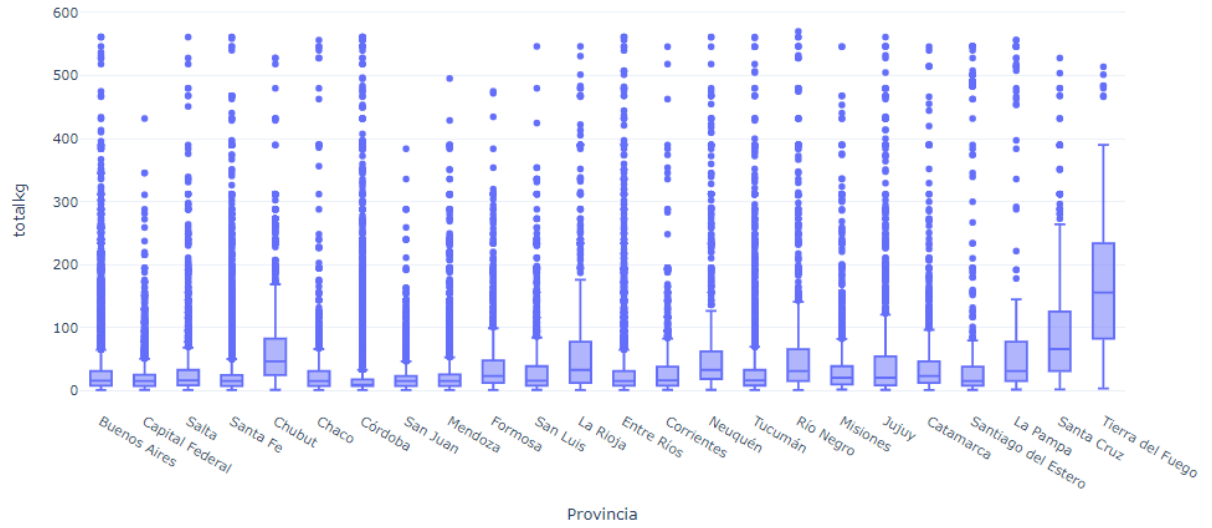


Figura 15 - Boxplot de ***'totalkg'*** por Provincia (sin outliers)

Se percibe que los outliers que se encontraban por encima de 600 ya no figuran y por lo tanto se aprecia un boxplot para cada provincia con más representatividad que la figura anterior.

También se realizó un análisis por provincia para los meses de cada año por separado. Los resultados se presentaron en gráficos confeccionados con la librería Plotly dado que estos permiten realizar interacciones de distintas maneras, por ejemplo seleccionar una provincia en particular (Córdoba) y ver su boxplot para cada mes del año seleccionado (2018) (Fig 16).

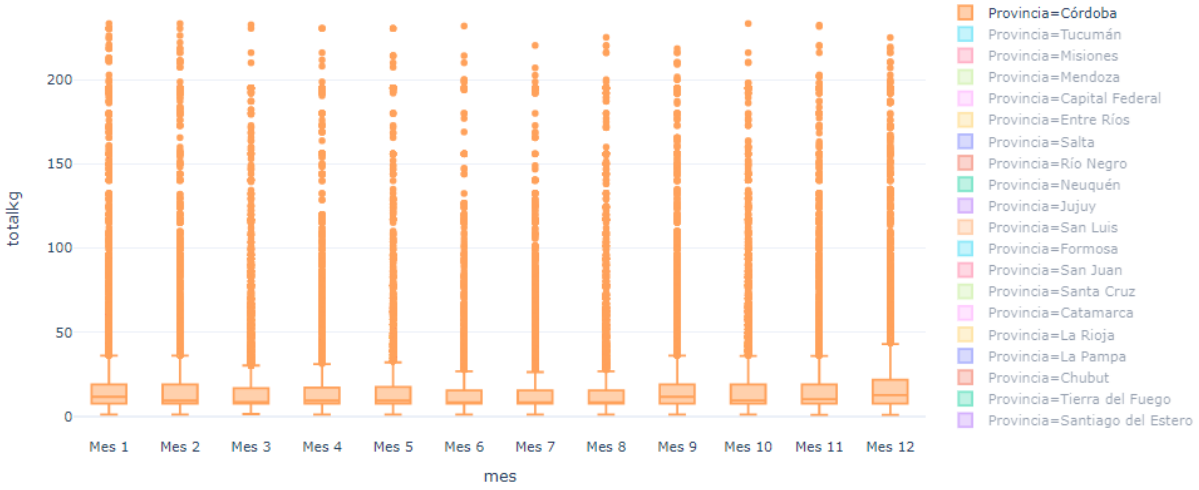


Figura 16 - Boxplot de ***'totalkg'*** por Provincia (sin outliers), versión interactiva

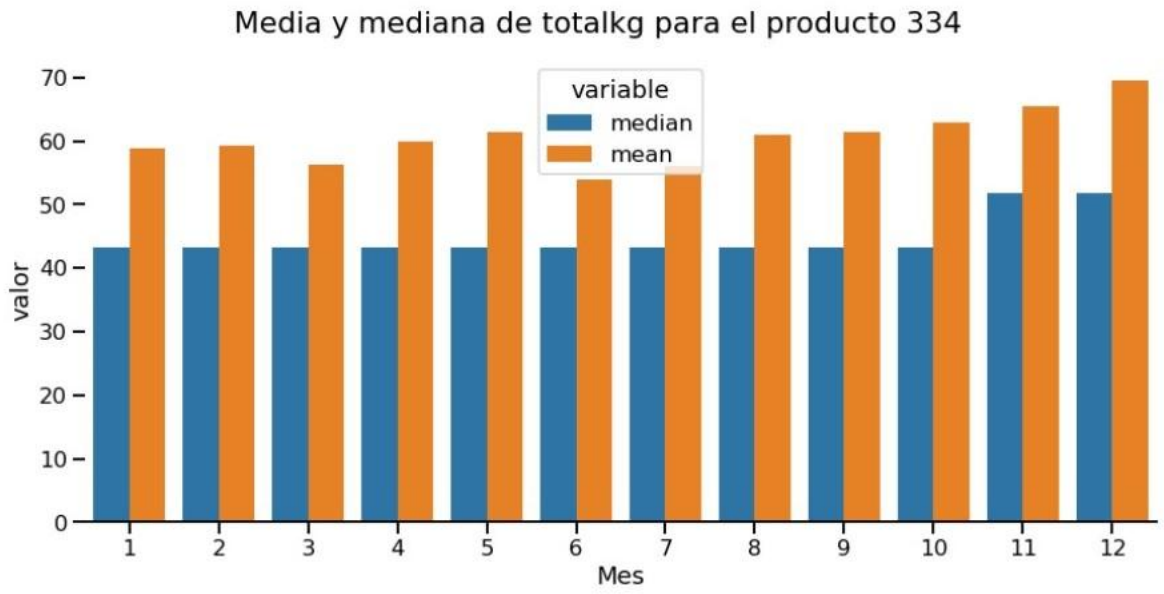
Por razones de visibilidad este tipo de gráfica no se presentará en el informe dado que sin las interacciones proporcionadas por la librería Plotly resulta difícil interpretarlos. Aún así los resultados que se visualizan siguen una interpretación similar a los gráficos explicados anteriormente.

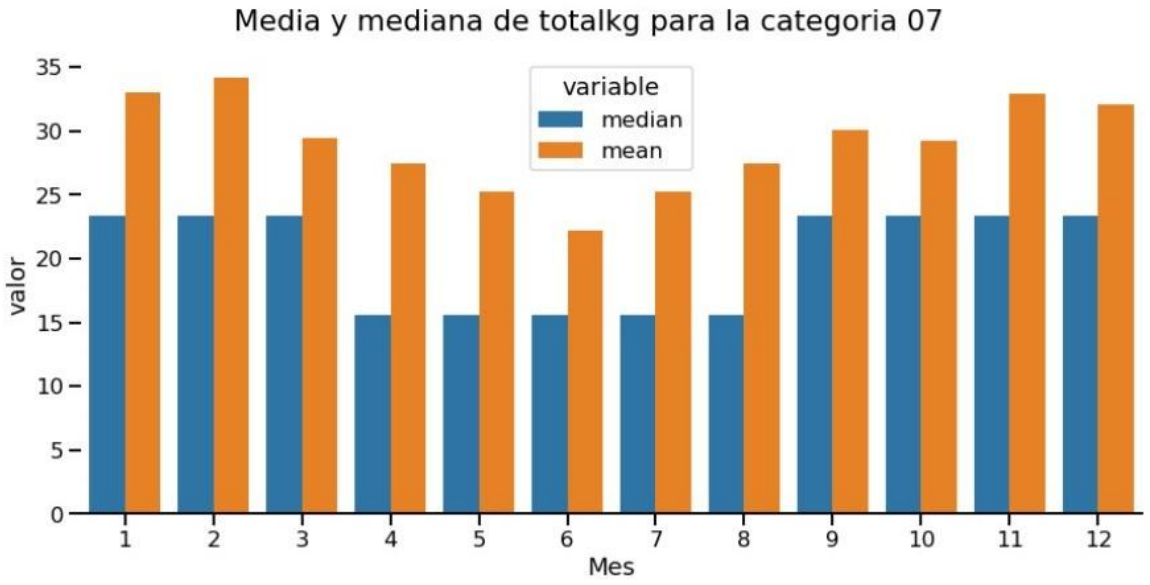
13) Determinar si hay valores faltantes e indicar qué tratamiento darle en tal caso. Indicar cuáles son los features con más valores faltantes. ¿Conviene descartarlos o completarlos con un valor particular?

Ver sección Análisis Exploratorio

14) Calcule la media y la mediana de totalkg por mes del producto más vendido y de la categoría más vendida. Realice este análisis por localidad y por provincia agrupando los pedidos por mes.

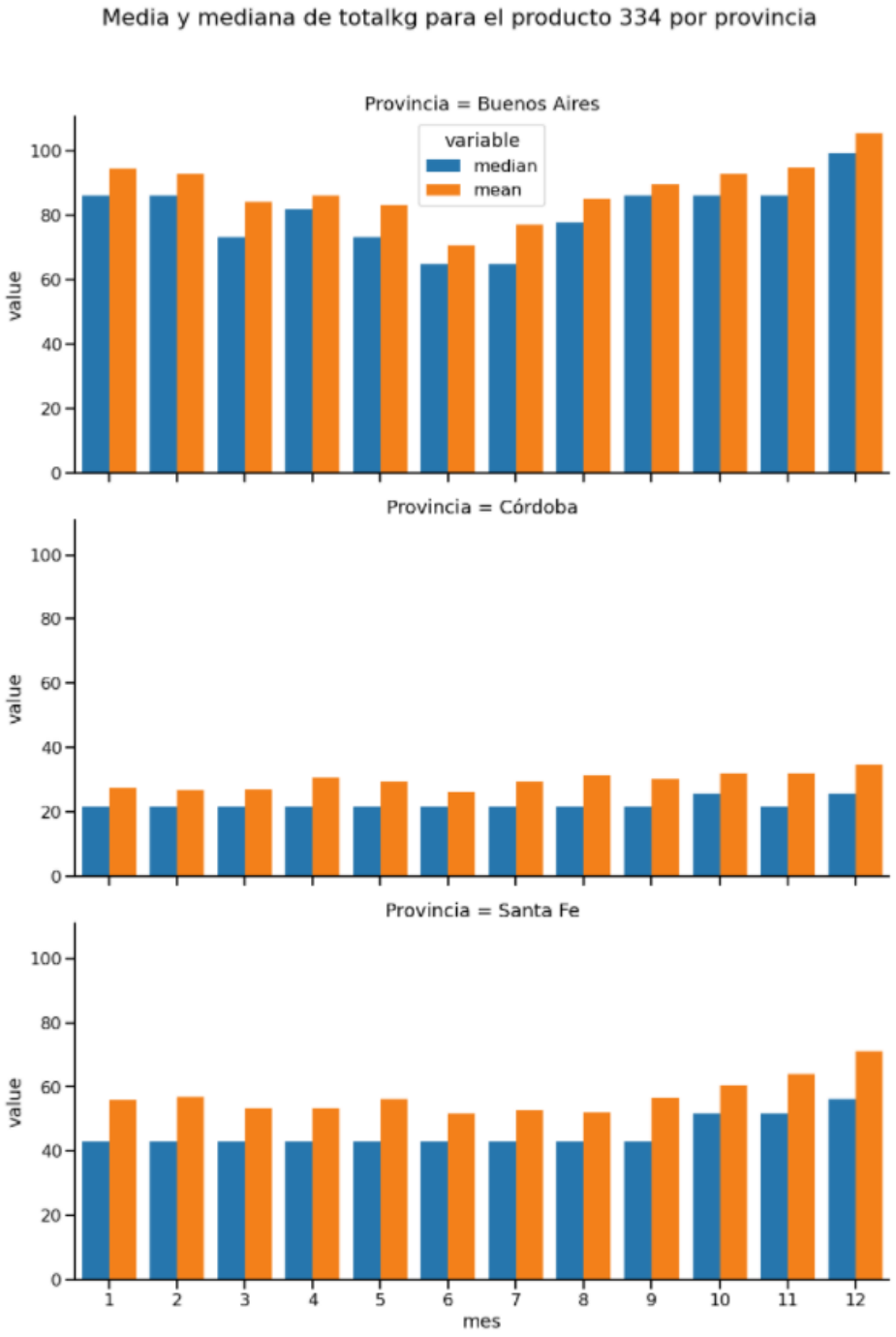
El producto más vendido es el 334 y la categoría más vendida la 07 - SABORES COMUNES. Recordemos que el producto 334 no pertenece a la categoría más vendida, por lo tanto no necesariamente van a mostrar las mismas tendencias. Nuevamente, se realizó primero un análisis a nivel país y luego para las tres provincias y tres localidades con más kg pedidos (Fig 17, Fig 18, Fig 19).



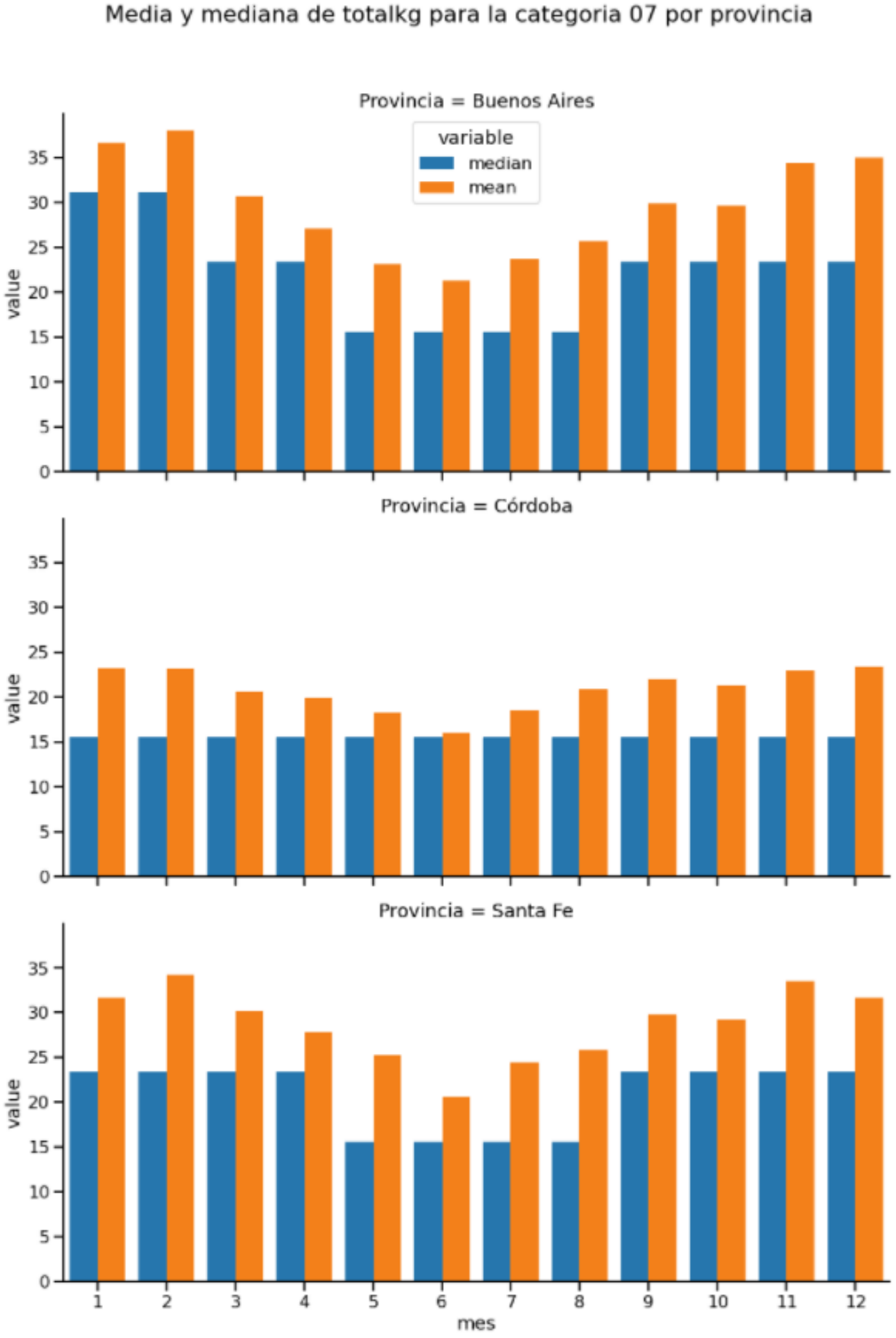


Se puede observar que efectivamente es distinto el comportamiento del producto más vendido y el de la categoría más vendida. Para el promedio de la categoría 7 por mes se puede ver claramente la estacionalidad: aumenta de forma escalonada cuando se acerca el verano y disminuye también de forma escalonada al acercarse el invierno. La mediana por otro lado es bastante más estable, pero pueden verse claramente dos temporadas: una temporada baja en los meses de abril a agosto y una alta en los meses de septiembre a marzo. Con el producto 334 en cambio, no hay una estacionalidad evidente, tanto el promedio como la mediana se mantienen más o menos estables a lo largo del año. Quizás sí haya un leve aumento en los meses de noviembre a diciembre, pero puede deberse a que las ventas de todos los productos aumentan en esas fechas porque aumenta el tráfico en las heladerías en general. No se pudo determinar exactamente qué es el producto 334 con la información provista para encontrar una explicación a este comportamiento que es muy distinto al de la categoría con más ventas, pero probablemente este comportamiento sí explica porque es el producto más vendido: se vende mucho en todos los meses

Por otro lado, en ambos casos, la media es siempre mayor a la mediana, probablemente porque se hacen algunos pedidos excepcionalmente grandes que no modifican la mediana porque son pocos pero si la media.

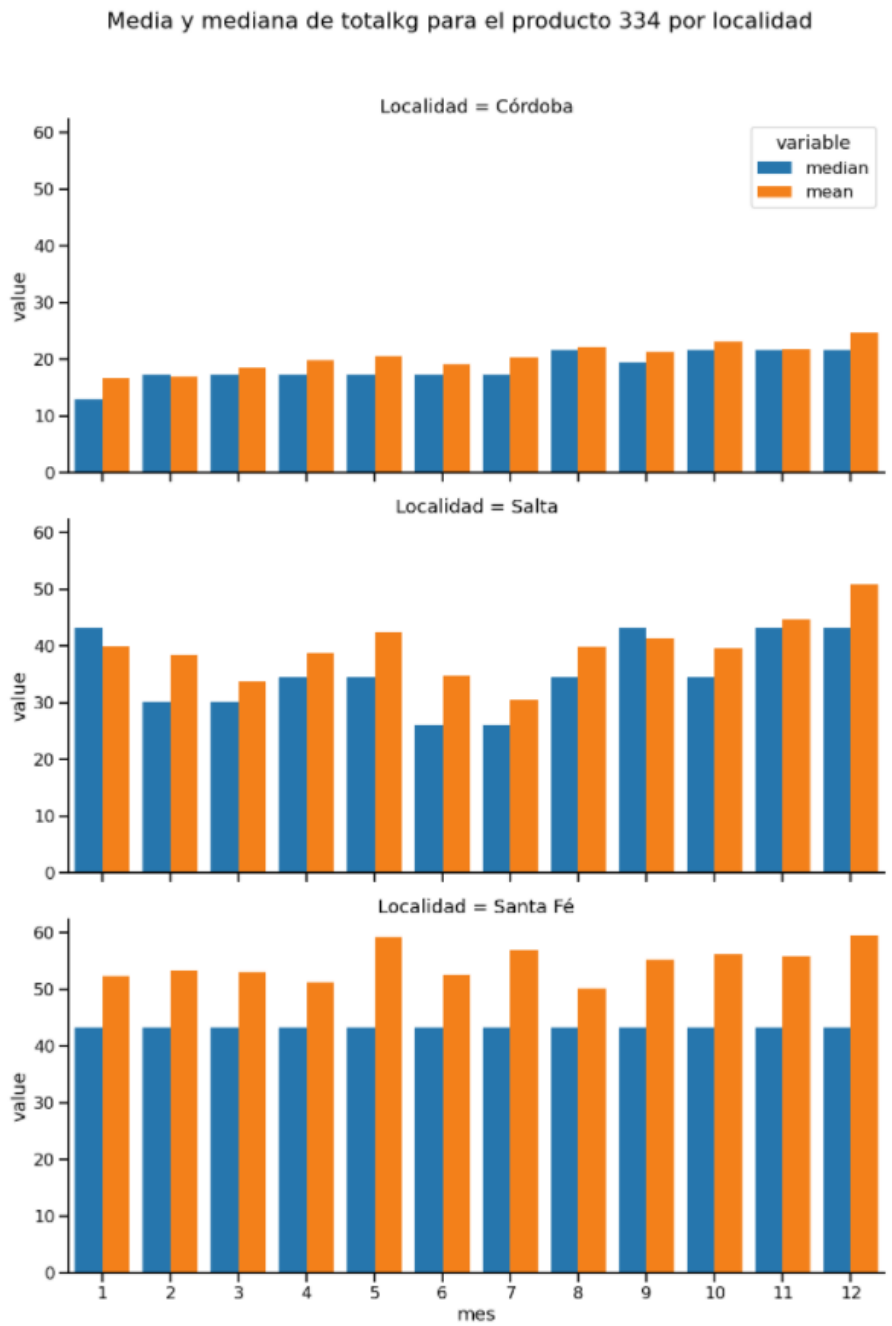


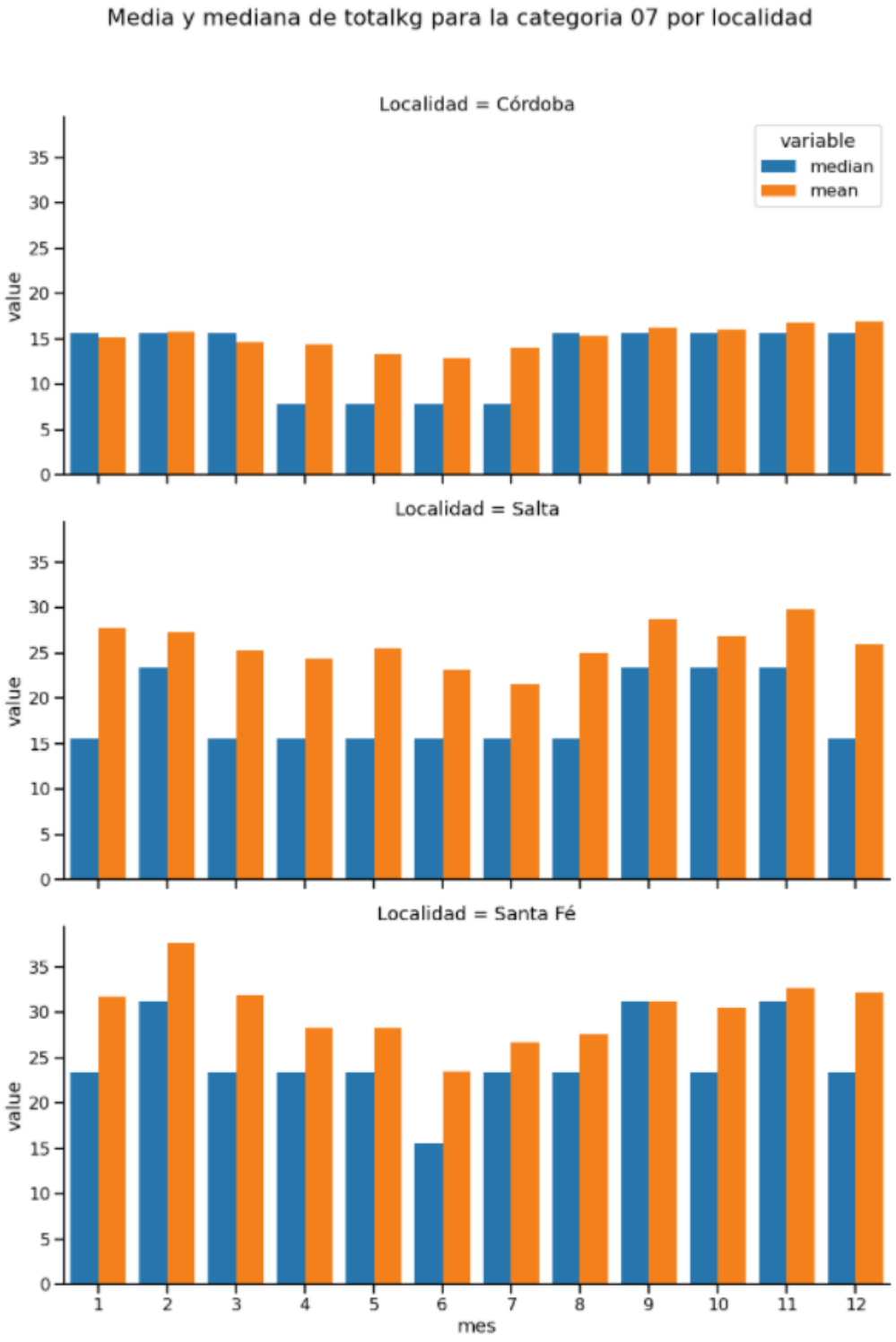
Para el producto 334 se observa una tendencia similar a la nacional en Córdoba y Santa Fe. En Buenos Aires baja un poco más el promedio y la mediana en invierno, pero podría deberse a que disminuyen en general las ventas en las heladerías.


Media y mediana de totalkg para la categoria 07 por provincia
Provincia = Buenos Aires
variable
median
mean
value
Provincia = Córdoba
value
Provincia = Santa Fe
value
mes
1
2
3
4
5
6
7
8
9
10
11
12

Para la categoría 07 se observa una tendencia similar a la nacional en Buenos Aires y Santa Fe. En Córdoba hay una cierta estacionalidad en el promedio, pero la mediana se mantiene igual en todos los meses analizados.

Cuando vemos a nivel Localidad hay más variabilidad y no hay tendencias tan claras, tanto para el producto como para la categoría, probablemente porque son menos los datos disponibles.




Media y mediana de totalkg para la categoria 07 por localidad
Localidad = Córdoba
Localidad = Salta
Localidad = Santa Fé
variable
median
mean
value
mes
0
5
10
15
20
25
30
35
1
2
3
4
5
6
7
8
9
10
11
12

15) ¿Cuál es la provincia con mayor promedio de totalkg por mes? ¿Y la de menor promedio por mes?

En primer lugar se realiza un análisis global para todo el período de análisis, se obtiene como resultado que la provincia de Tierra del Fuego es la que posee el mayor promedio de totalkg mientras que la provincia de Córdoba posee el menor promedio de totalkg para los dos años de estudio.

Para el análisis mensual de las provincias que poseen el mayor promedio en totalkg se identifican sólo dos provincias: Tierra del Fuego y Santa Cruz (Fig 20).

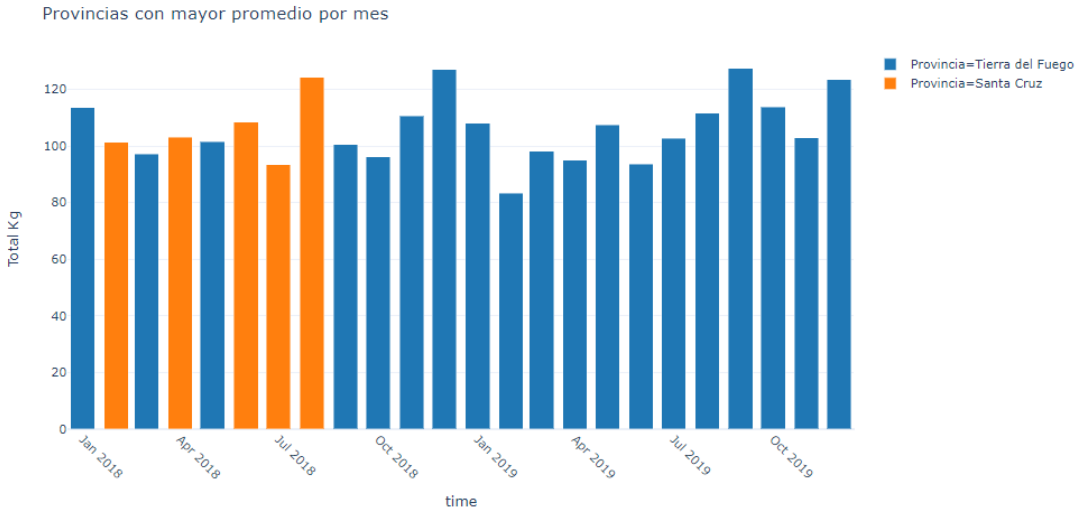


Figura 20 - Provincias con mayor promedio de ***'totalkg'*** por mes

Para la mayoría de los meses la provincia de Tierra del Fuego es la que posee el mayor promedio, mientras que Santa Cruz prevalece sólo en 5 meses de los 24 meses comprendidos en el estudio estos corresponden todos al año 2018.

Por otro lado las provincias que poseen el menor promedio de solicitud de productos por mes son: Córdoba y San Juan (Fig 21).

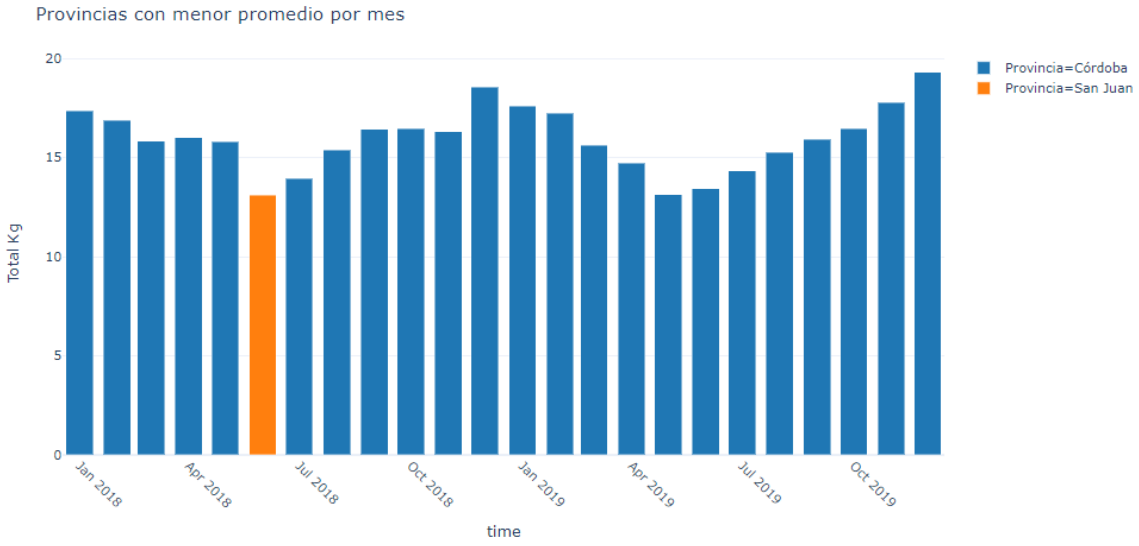


Figura 21 - Provincias con menor promedio de ***'totalkg'*** por mes

Sólo para junio del 2018 el menor promedio corresponde a San Juan mientras que para el resto de los meses corresponde a la provincia de Córdoba.

16) ¿Qué distribución tiene la variable totalkg? ¿Qué implicancias tiene la distribución de dicha variable?

La distribución es asimétrica hacia la izquierda, si lo vemos en un histograma con diferente grado de resolución (bin 100, 50, 20). En baja resolución se asemeja a una distribución unimodal sin embargo aumentando el grado de resolución se evidencian varios picos separados. Es probable que la mejor forma de modelar esta distribución sea como una combinación lineal de distribuciones gaussianas, más que como una distribución unimodal asimétrica a hacia la izquierda (Fig 22).

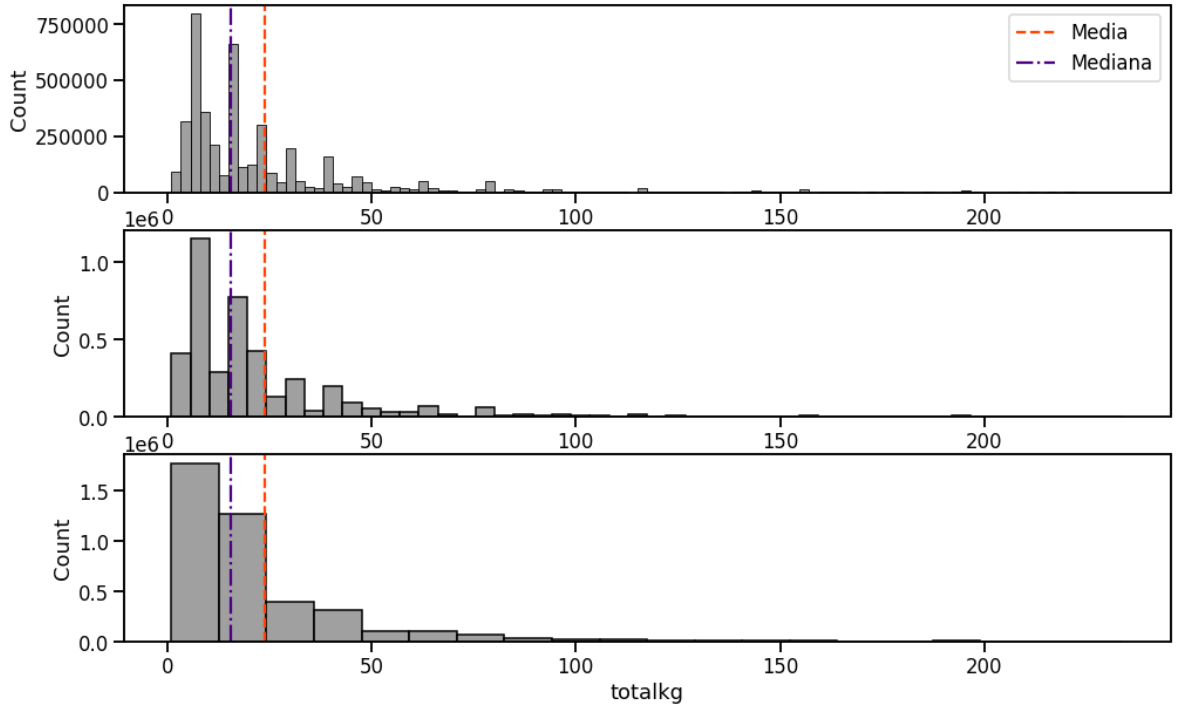


Figura 22 - Histograma de frecuencias de la variable '***totalkg***' con distinto grado de resolución

17) ¿Cuál es la frecuencia de las variables categóricas que seleccionaron?

De las variables que se seleccionaron, las categóricas son: sku, time y ubicación. Se realizaron histogramas para cada una de ellas para estudiar la frecuencia.

- Variable sku

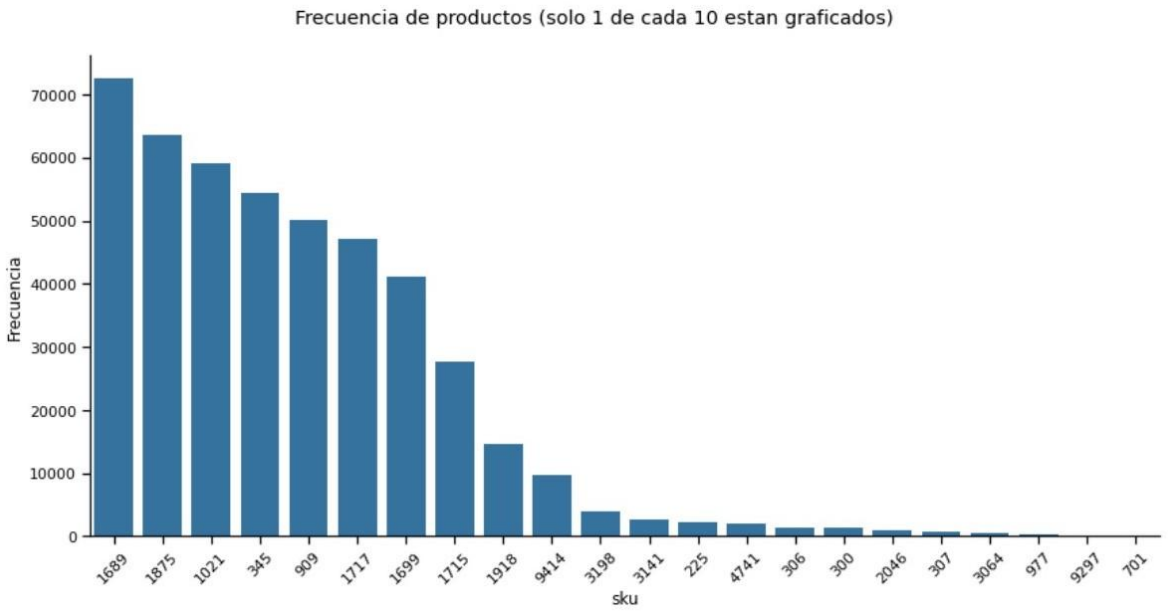


Figura 23 - Diagrama de Pareto para la variable '***sku***'

Como eran 217 sku's distintos se decidió ordenarlos por frecuencia de mayor a menor e ir graficando 1 de cada 10, para tener una idea de la distribución y que el gráfico fuera legible. Así se puede observar que, por ejemplo, hay un 54% de los productos que sólo representan aproximadamente el 3,4% de la cantidad de pedidos por parte de las sucursales a la central (sin tener en cuenta la cantidad de ese producto en cada pedido). Por otro lado, hay un 27,5% de los productos que representan 80% de los pedidos. Esto va a ser de ayuda para saber en qué productos centrarse para predecir la mayor parte de la demanda.

- Variable 'time'

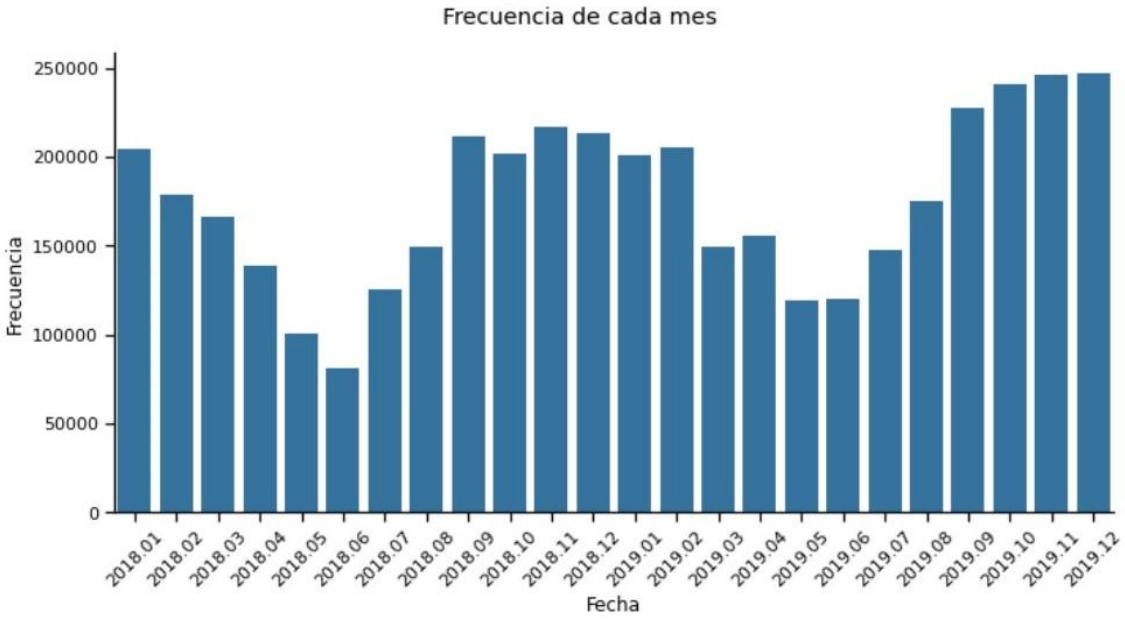


Figura 24 - Frecuencia de pedidos por mes para los años 2018-2019

Acá se puede ver la misma estacionalidad que para la suma y el promedio de la variable totalkg. Es decir, en el invierno disminuyen tanto la cantidad de pedidos como la cantidad de

unidades por pedido, en verano aumentan y el cambio de una a otra estación es escalonado (Fig 24).

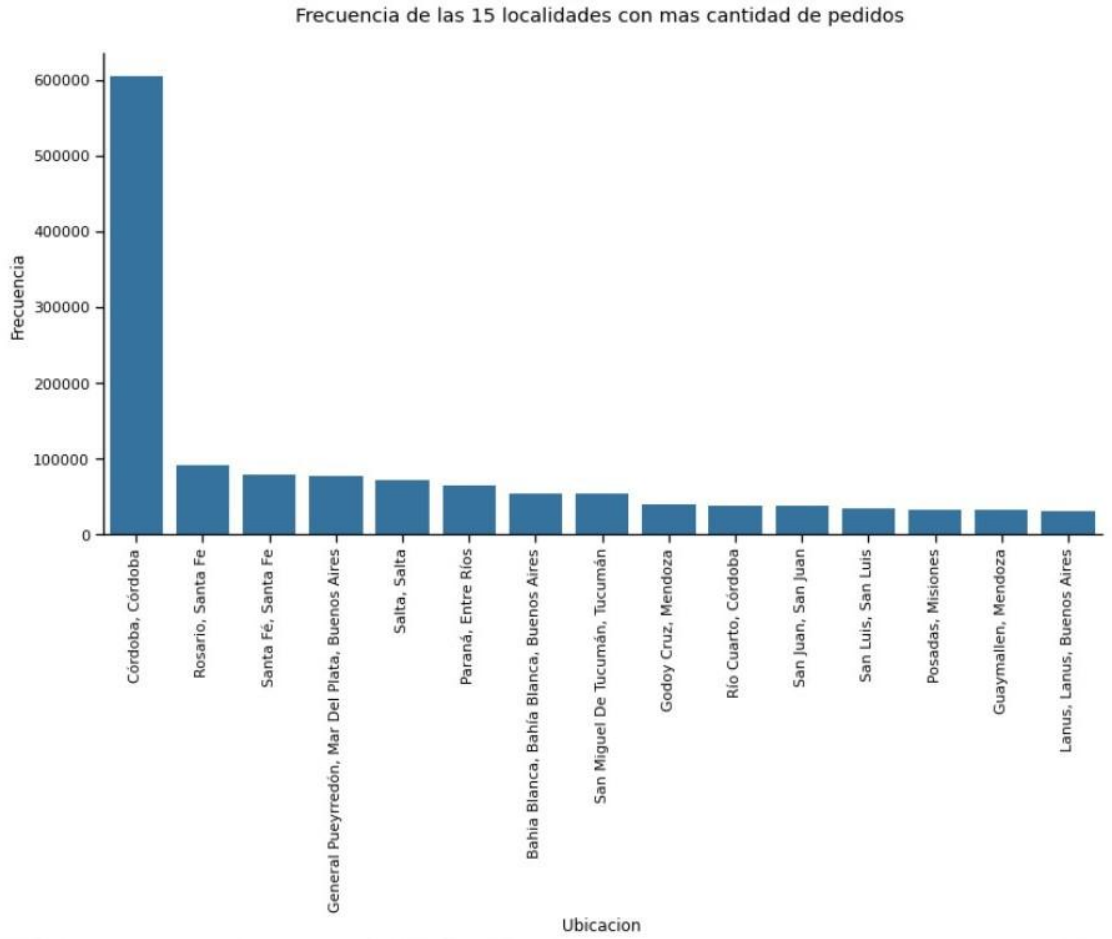
- Variable ubicación

La cantidad de valores de la variable ubicación, es la cantidad de localidades distintas presentes en el dataset (ya que al incluir también la provincia nos permite diferenciar entre localidades del mismo nombre pertenecientes a distintas provincias). Éstas son más de 600 y al ordenarlas por su frecuencia de mayor a menor se puede observar que las primeras tienen valores bastante distintos al resto. Por eso se decidió hacer dos gráficos (en la siguiente página): uno que muestra la frecuencia de las primeras 15 localidades y otro que muestra las restantes graficando 1 de cada 20 para que sea legible (Fig 25).

En el primer gráfico se puede observar que Córdoba presenta una diferencia muy importante en frecuencia respecto al resto, tiene más de 6 veces la frecuencia de la segunda localidad, Santa Fe, y sola representa el 14,3% de las cantidad de ventas, cuando solo representa el 0,15% de la cantidad de localidades. Esto ayuda a explicar porque la provincia de Córdoba tiene el promedio de kg más bajo, a pesar de ser de las que más kg pide.

Mirando también el segundo gráfico y la tabla con la que fue generado, se puede ver que:

- Un 5,3% de las localidades representan 42,4% de los pedidos
- Un 50% de las localidades representan sólo un 15% de los pedidos
- Un 50% las localidades representan tiene una frecuencia entre 2500 y 6500


Frecuencia de las 15 localidades con mas cantidad de pedidos
Frecuencia
600000
500000
400000
300000
200000
100000
0
Córdoba, Córdoba
Rosario, Santa Fe
Santa Fé, Santa Fe
General Pueyrredón, Mar Del Plata, Buenos Aires
Salta, Salta
Paraná, Entre Ríos
Bahía Blanca, Bahía Blanca, Buenos Aires
San Miguel De Tucumán, Tucumán
Godoy Cruz, Mendoza
Río Cuarto, Córdoba
San Juan, San Juan
San Luis, San Luis
Posadas, Misiones
Guaymallen, Mendoza
Lanus, Lanus, Buenos Aires
Ubicacion

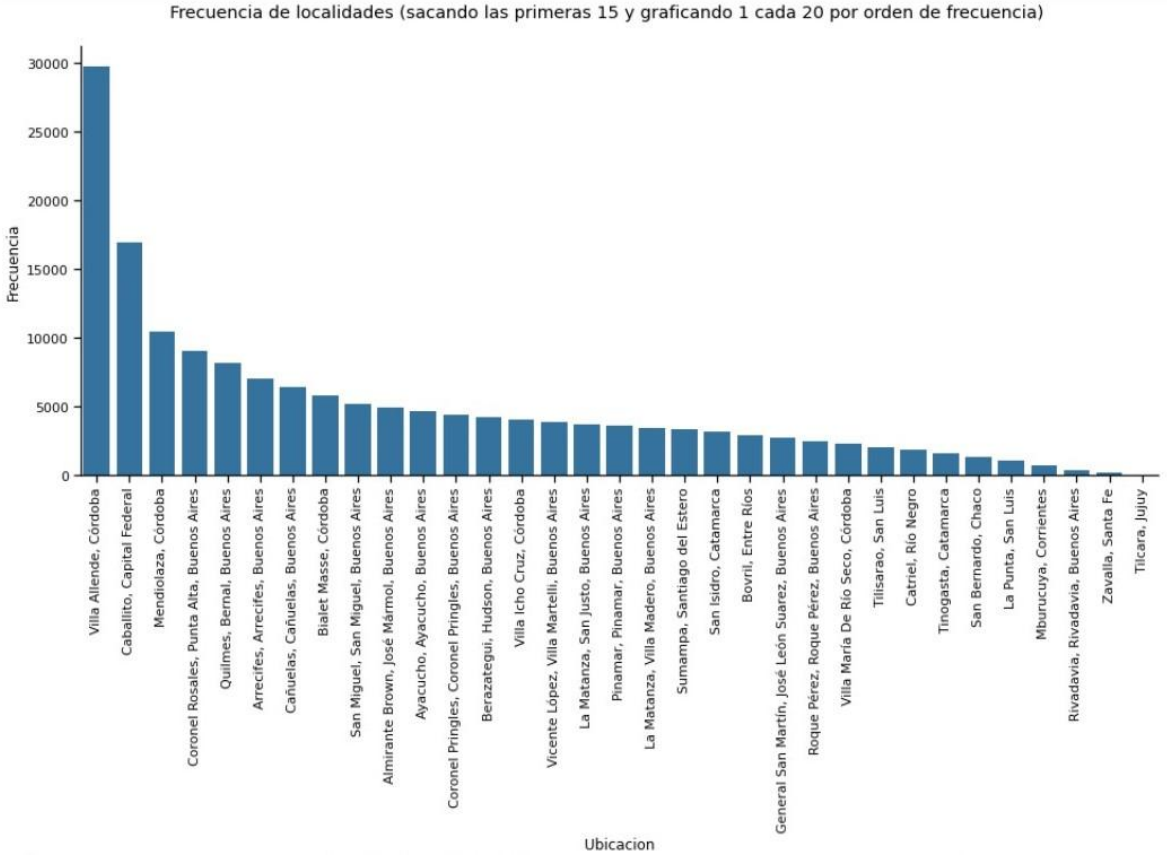

Frecuencia de localidades (sacando las primeras 15 y graficando 1 cada 20 por orden de frecuencia)
Frecuencia
30000
25000
20000
15000
10000
5000
0
Villa Allende, Córdoba
Caballito, Capital Federal
Mendiolaza, Córdoba
Coronel Rosales, Punta Alta, Buenos Aires
Quilmes, Bernal, Buenos Aires
Arrecifes, Arrecifes, Buenos Aires
Cañuelas, Cañuelas, Buenos Aires
Bialet Masse, Córdoba
San Miguel, San Miguel, Buenos Aires
Almirante Brown, José Mármol, Buenos Aires
Ayacucho, Ayacucho, Buenos Aires
Coronel Pringles, Coronel Pringles, Buenos Aires
Berazategui, Hudson, Buenos Aires
Villa Icho Cruz, Córdoba
Vicente López, Villa Martelli, Buenos Aires
La Matanza, San Justo, Buenos Aires
Pinamar, Pinamar, Buenos Aires
La Matanza, Villa Madero, Buenos Aires
Sumampa, Santiago del Estero
San Isidro, Catamarca
Bovril, Entre Ríos
General San Martín, José León Suarez, Buenos Aires
Roque Pérez, Roque Pérez, Buenos Aires
Villa María De Río Seco, Córdoba
Tilisarao, San Luis
Catriel, Río Negro
Tinogasta, Catamarca
San Bernardo, Chaco
La Punta, San Luis
Mburucuya, Corrientes
Rivadavia, Rivadavia, Buenos Aires
Zavalla, Santa Fe
Tilcara, Jujuy
Ubicacion

Figura 25 - Diagrama de Pareto para la variable '***Ubicación***'

18) ¿Cómo es la distribución de totalkg condicionada a algunas otras variables que decida seleccionar?

Si vemos la distribución con respecto al mes del año que consideremos vemos que la forma de la distribución se repite, siendo únicamente menor la densidad para los meses invernales. Lo opuesto es verdad para los meses estivales (Fig 26).

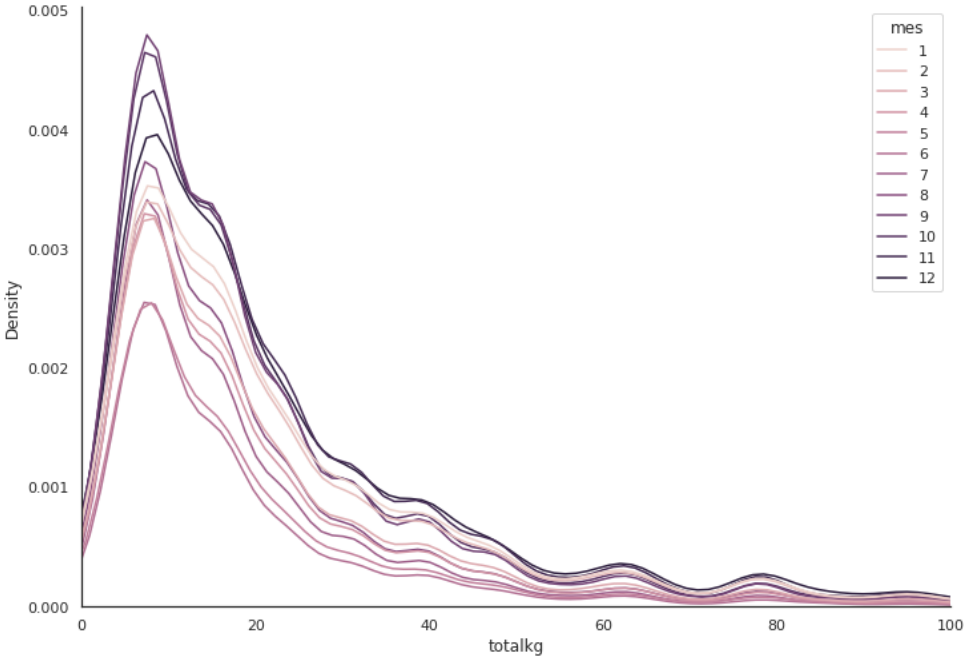
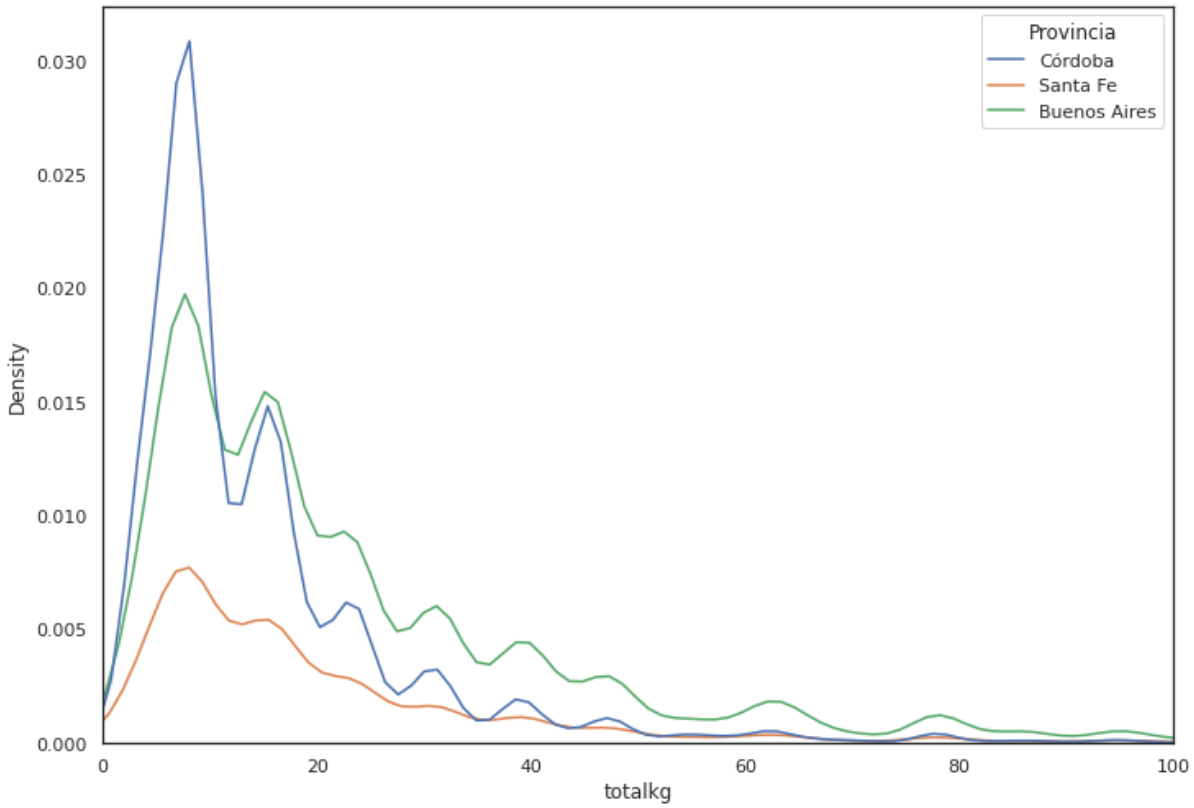


Figura 26 - Diagrama de densidad condicional de '***totalkg***'con respecto a '***mes***'

En general la distribución condicionada a las provincias parece tener el mismo patrón sin tener en cuenta las diferencias de densidad. Si comparamos las primeras tres provincias, Buenos Aires, Córdoba y Santa Fe, replican la misma forma. Si vemos la distribución tomando provincias de distintas regiones con totalkg similares como Mendoza, Tucuman y Capital Federal vemos que el patrón se mantiene. Para menores densidades y manteniendo el factor de suavizado si es posible notar que los picos son menos marcados para provincias con menor demanda pero difícil definir si es una cuestión del patrón de consumo o de una menor cantidad de datos (Fig 27).

Algunas categorías presentan una distribución que se modifica con la época del año. A modo de ejemplo de muestra la distribución de 'totalkg' de la Categoría "16 - FAMILIAR" donde se ve que en meses estivales hay una mayor densidad en valores de totalkg más grandes, mientras que en época invernal se concentra en valores pequeños (Fig 28).



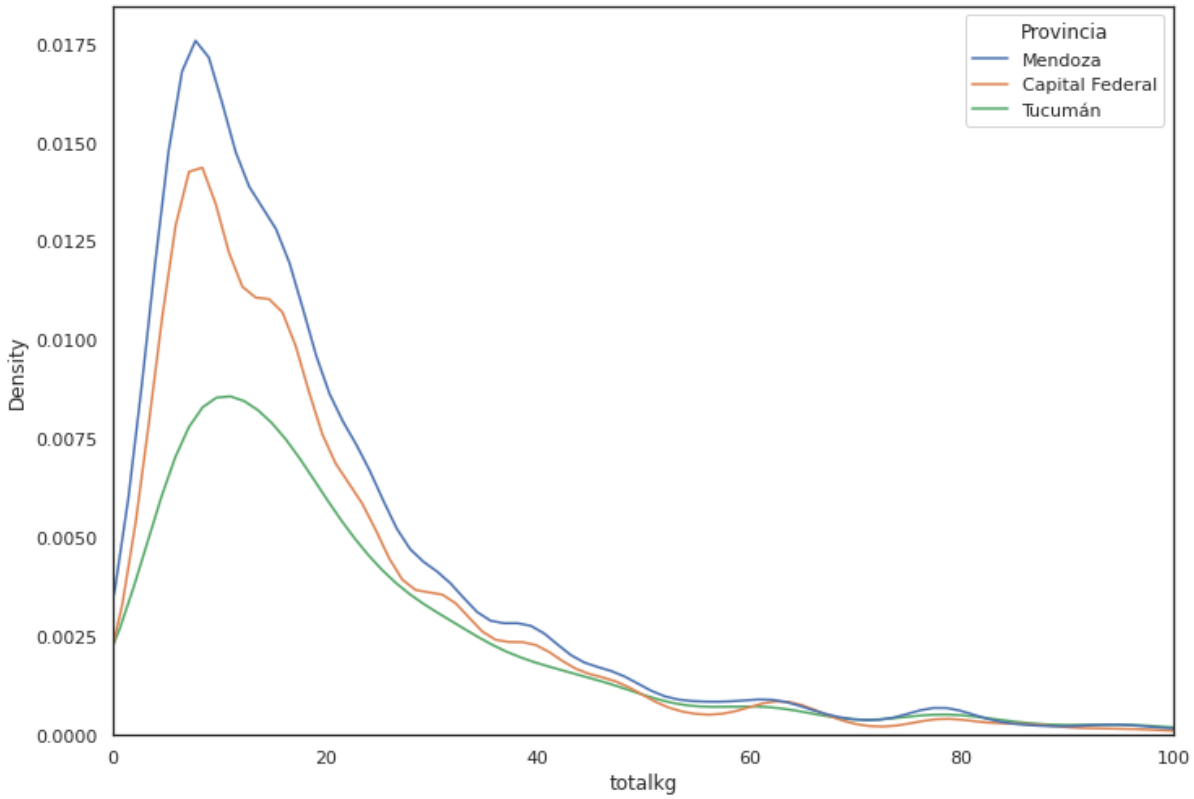


Figura 27 - Diagrama de densidad condicional de '***totalkg***'con respecto a '***Provincia***'

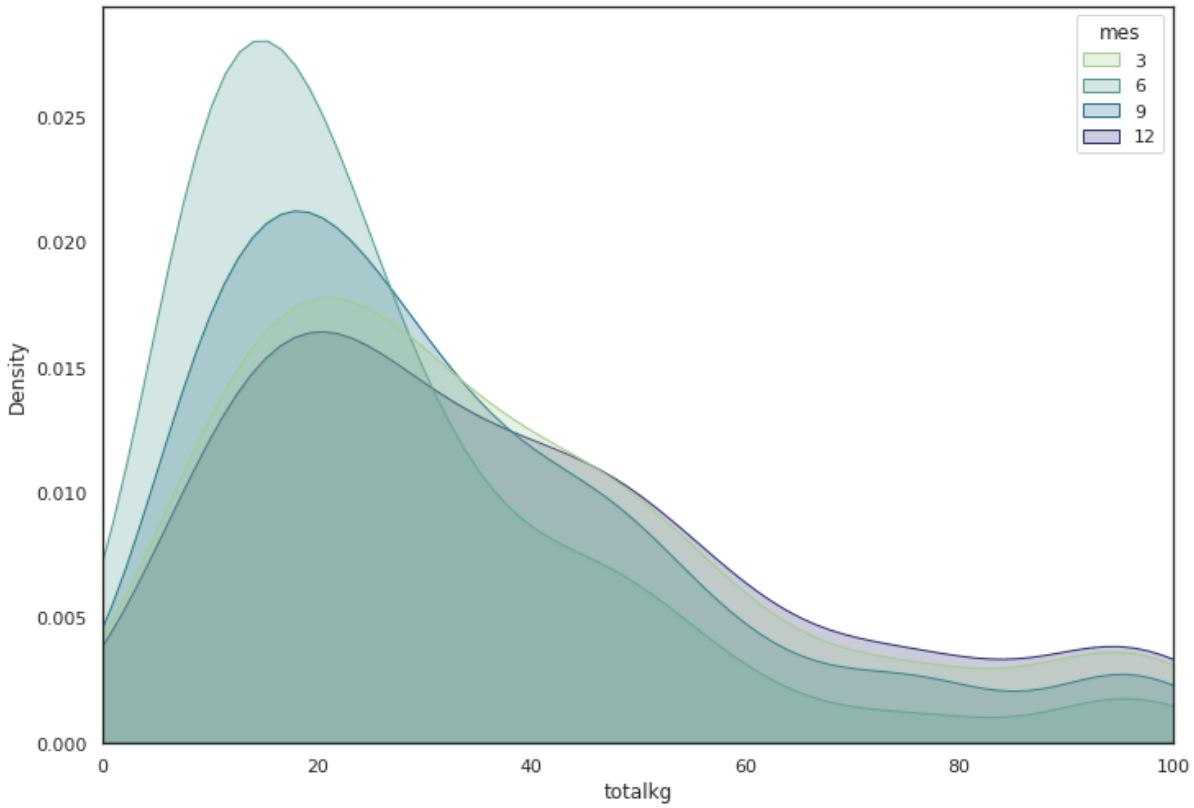


Figura 27 - Diagrama de densidad condicional de '***totalkg***'con respecto a '***mes***' para la categoría 16 - FAMILIAR

19) Indicar cuales son las variables que tienen mayor correlación

Ver Sección 4

20) ¿Son estadísticamente distintas las medias o medianas (lo que indique que corresponde) de totalkg entre dos provincias (compare las dos provincias que mayor promedio de totalkg anual tienen)?

Las dos provincias que mayor promedio de totalkg anual tienen son Tierra del Fuego y Santa Cruz. Podemos observar que el promedio de Tierra del Fuego es un poco mayor al de Santa Cruz, y la mediana es bastante mayor, casi el doble. Por otro lado en Tierra del Fuego, la mediana es mayor que el promedio, o sea, que la gran mayoría de los pedidos que se hacen son de muchos kilos y quizás hay algunos excepcionalmente chicos que hacen que baje el promedio pero no se modifique la mediana. Esto probablemente se deba a que Tierra del Fuego es la provincia más alejada de la central y por lo tanto se hacen pocos pedidos, pero cada uno de muchos kilos para ahorrar en envío. Se seguirá discutiendo un poco más este fenómeno en la siguiente sección. Para Santa Cruz, la media es levemente menor que el promedio, pero igual ambos son altos, probablemente por la misma razón, ya que Santa Cruz también es una provincia alejada.

Provincia	Promedio	Mediana
Tierra del Fuego	103.41	117.00
Santa Cruz	74.95	62.40

Tabla I - Media y mediana de las dos Provincias con mayor promedio anual

21) ¿Son estadísticamente distintas las medias o medianas (lo que indique que corresponde) de totalkg entre los promedios de los 3 puntos de venta que más venden y los 3 puntos de ventas que menos venden?

Para desarrollar este inciso se procedió a determinar cuáles eran los tres puntos de venta que más piden, los id correspondientes a estos son: 100069,100308 y 100640.

A continuación se muestran las sumas, medias y medianas de cada una de ellas en la siguiente tabla:

ID PDV	Suma	Promedio	Mediana
100069	358,431.82	125.63	144.0
100308	324,143.90	75.77	62.4
100640	309,895.99	98.72	82.8

Tabla II - Suma, media y mediana de los tres Puntos de Ventas (PDV) que más piden

Se observan diferencias en las medias y medianas para los tres PDV que más piden. Para el primer punto de venta (100069) se observa que la mediana es mayor al promedio, es decir, que este PDV posee una distribución asimétrica derecha. Mientras que para los dos puntos de ventas restantes (100308 y 100640) las medianas son menores al promedio, es decir, poseen una distribución asimétrica izquierda. Esto último puede tener explicación si es que existe algún pedido de gran cantidad que se realizó generando que el promedio aumente, dado que es una medida estadística muy sensible a este tipo de valores.

En cuanto a la relación de las medias entre los tres puntos de ventas se puede observar que no presentan gran diferencia entre ellos. Mientras que en el caso de la mediana del primer (100069) y segundo (10038) punto de venta se observa una diferencia de más del doble.

Mientras que los tres puntos de venta que menos piden son: 100107-3, 100107-5 y 101086-1. A continuación se muestran las sumas, medias y medianas de cada una de ellas en la siguiente tabla:

ID PDV	Suma	Promedio	Mediana
100107-3	1.45	1.45	1.45
100107-5	1.45	1.45	1.45
101086-1	7.35	3.68	3.68

Tabla III - Suma, media y mediana de los tres Puntos de Ventas (PDV) que menos piden

Como principal observación que se arroja de la tabla presentada es que los primeros dos puntos de ventas son iguales entre ellos y sólo registran un sólo pedido por eso la suma, promedio y media coinciden. Para el tercer punto de venta se realizaron dos pedidos de la misma cuantía es por eso que la media y mediana coinciden y la suma es el doble de estas. Por lo tanto, los tres puntos de venta que menos piden responden a una distribución simétrica.

Si bien los dos primeros puntos de ventas son distintos al tercero por una diferencia de más del doble en la media y mediana, podemos concluir que no son distintos a gran escala considerando los valores posibles que puede asumir.

22) Establecer la probabilidad de que el promedio de los pedidos realizados por los puntos de venta pertenecientes a Córdoba se encuentren por encima de la media nacional. Realice este análisis tomando los promedios mensuales.

Si tomamos un horizonte de tiempo mensual de los 3,070 pedidos realizados por puntos de venta radicados en la provincia de Córdoba 1,047 de dichos pedidos estuvieron por encima de la media de pedidos mensuales de puntos de venta a nivel nacional, lo que equivale a una probabilidad estimada de 35%.

23) ¿En qué época del año el promedio de totalkg por provincia y categoría (tome sólo las 3 categorías más pedidas) es más alto y cuál es más bajo? ¿Qué comportamiento se observa en los pedidos?

Como se mencionó anteriormente el comportamiento es estacional, si vemos la distribución temporal del promedio de pedidos por provincia y categoría, tomando las primeras 3 categorías por cantidad de pedidos vemos un patrón similar al que se describió en el punto 11, con un marcado descenso en la época invernal (ej. Fig 28)

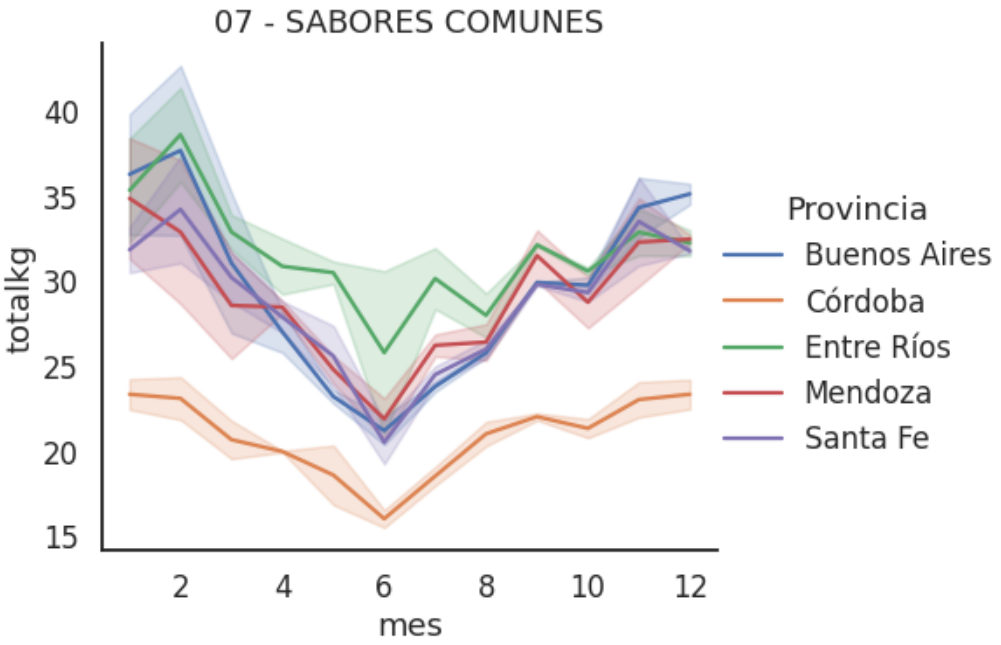


Figura 28 - Promedio de '***totalkg***' en pedidos mensuales para las primeras 5 provincias con mayor volumen de pedidos

4 RELACIONES INTERESANTES ENTRE VARIABLES

En esta sección se desarrollan las relaciones importantes entre las variables que fueron encontradas a lo largo del análisis. A continuación se describe cada una de estas relaciones.

Una de las relaciones más importante que se percibe a lo largo de todo el análisis es la presencia de estacionalidad para la variable ***'totalkg'***. En que la demanda en los meses de verano aumenta mientras que para los meses de invierno la demanda cae. También se percibe que en promedio la demanda para el año 2019 es mayor a la del 2018. La relación de estacionalidad encontrada es una situación esperada dado que el estudio refiere a los productos de una compañía de venta de alimentos congelados.

Otra relación importante se refiere a la demanda de productos en las provincias. En primer lugar, hay que distinguir dos tipos de análisis que arroja resultados distintos. Por un lado se encuentra el análisis de aquellas provincias que más ***'totalkg'*** solicitan por mes (o también a nivel anual). Y, por otro lado, el análisis del promedio de ***'totalkg'*** por mes (o a nivel anual). Un claro ejemplo de las diferencias que pueden presentar estos dos tipos de análisis es la provincia de Córdoba, en que a pesar de ser una de las principales provincias que solicitan más productos por mes (ver inciso 1) posee el más bajo promedio de pedido de productos por mes (ver inciso 15). Estas particularidades pueden encontrar causa en los costos de transporte asociadas, en que Córdoba al tener una cercanía a la central poseen menor costo de transporte y por lo tanto la frecuencia de pedidos puede ser mayor por una menor proporción de pedidos en relación por ejemplo con Tierra del Fuego que posee un alto costo de transporte y por lo tanto es de esperarse que la frecuencia de pedido sea menor pero en una gran cuantía.

Por último, existe una relación positiva entre las variables **'totalkg'** tanto con la variable '**cantidad_pedida'** y con la variable **'unidadkg'**, dado que la variable '**totalkg'** es el producto de estas dos variables mencionadas.

5 PRINCIPALES CONCLUSIONES

Algunas de las principales conclusiones que se obtienen en este informe se describen a continuación.

La distribución de los pedidos es en general asimétrica teniendo una mayor densidad de pedidos pequeños.

A pesar de que en ciertas épocas del año y para ciertas categorías de productos el volumen de pedidos de mayor tamaño aumenta, en general siempre los pedidos más pequeños son más prevalentes.

Excepciones a esta tendencia se dan en algunas provincias, como Tierra del Fuego, donde el tamaño de los pedidos en promedio son grandes. Esto puede explicarse atendiendo a las complicaciones asociadas al transporte.

Existe una marcada estacionalidad en todas las combinaciones de variables estudiadas, dándose menor cantidad de pedidos en épocas invernales con respecto a épocas estivales. La estacionalidad se cumple en menor medida para las provincias de regiones subtropicales o tropicales donde hay una variación menos marcada de las condiciones climáticas entre distintas épocas del año. Lo opuesto no se observó para provincias del Sur donde la estacionalidad observada es similar a la de provincias de la región centro.