Informe de resultados y análisis estadístico

1. Antecedentes y objetivos

El proyecto involucra tres fuentes de datos:

1. **customers.csv**, con información de clientes (ID, nombre, dirección, etc.).
2. **orders.csv**, con detalles sobre las órdenes (fecha de pedido, fecha requerida, envío, etc.).
3. **order_items.csv**, relación de cada orden con productos adquiridos, precio de lista, cantidad y descuentos.

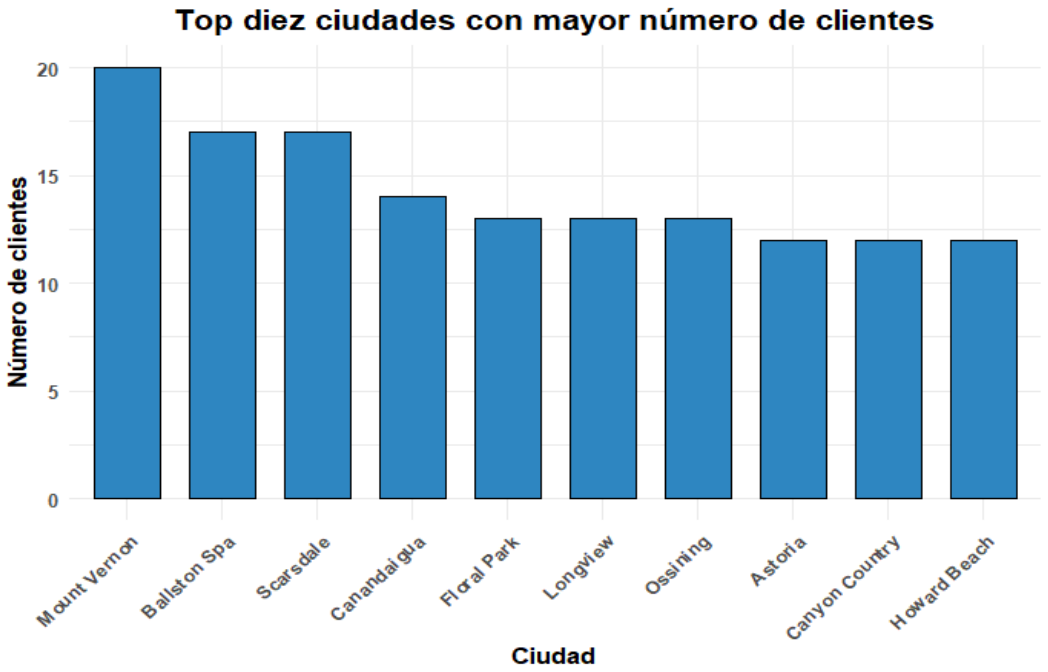
Los objetivos principales son:

- Conocer el número total de clientes, cuántos no han realizado órdenes y cómo se distribuyen por ciudad.
- Analizar el volumen de órdenes por fecha (particularmente mensual).
- Calcular el descuento promedio y las ventas totales por fecha.
- Realizar un análisis de series de tiempo (descomposición y pronóstico ARIMA).
- Construir un modelo de regresión lineal múltiple para predecir ventas.
- Efectuar pruebas estadísticas adicionales (t-test y test de normalidad de residuos) para robustecer la inferencia.

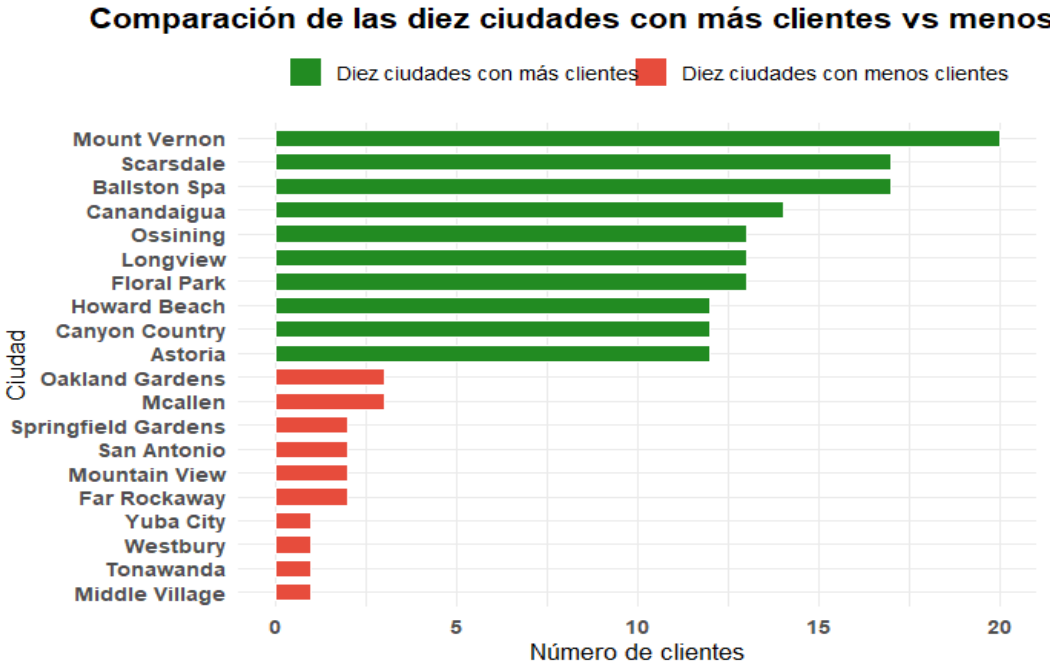
2. Resumen de resultados

2.1 Análisis de clientes

1. Se cuenta con **1445** clientes únicos (después de verificar y descartar registros duplicados en customers).
2. El número de clientes que no han realizado ninguna orden resultó ser cero **0** tras comparar contra la tabla de órdenes.
3. Al agrupar por ciudad, se identificaron las diez ciudades con mayor número de clientes; es decir, Mount Vernon, Ballston Spa, Scarsdale, Canandaigua, Fidral Park, Longview, Ossining, Astoria, Canyon Country, HowardBeach. En el gráfico de barras se visualiza claramente esta concentración.



4. Asimismo, se compararon las diez ciudades con más clientes vs diez ciudades con menos clientes en otra visualización, para entender la disparidad de clientes por región.

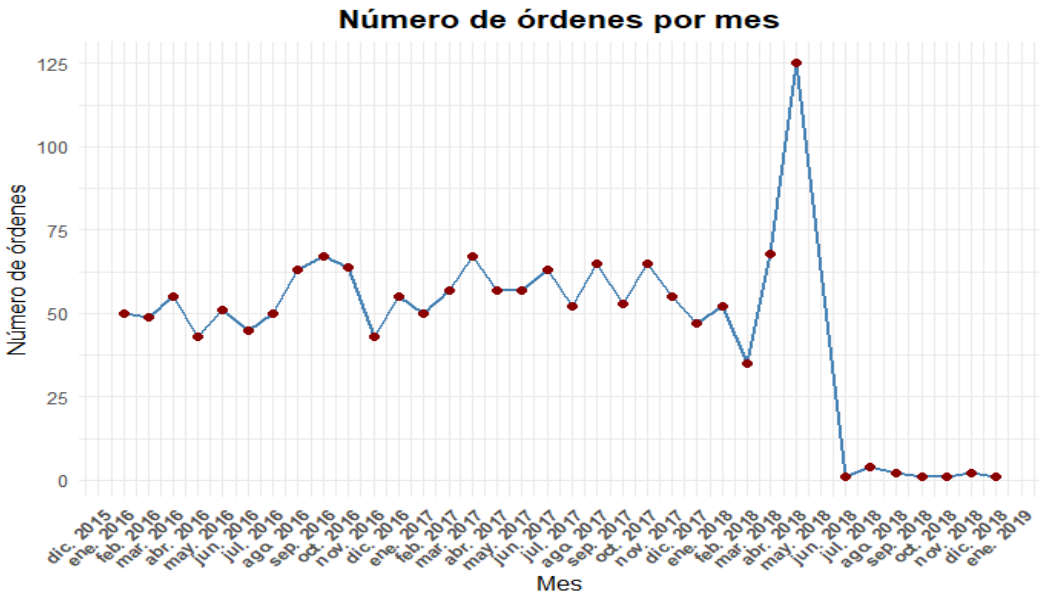


A continuación, se realiza la siguiente interpretación:

- Las ciudades con mayor número de clientes se concentran en New York (por ejemplo Mount Vernon) y algunas en California (por ejemplo Ballston Spa).
- El hecho de que 0 clientes queden fuera de órdenes indica que, en este *dataset*, todos tuvieron al menos una compra registrada.

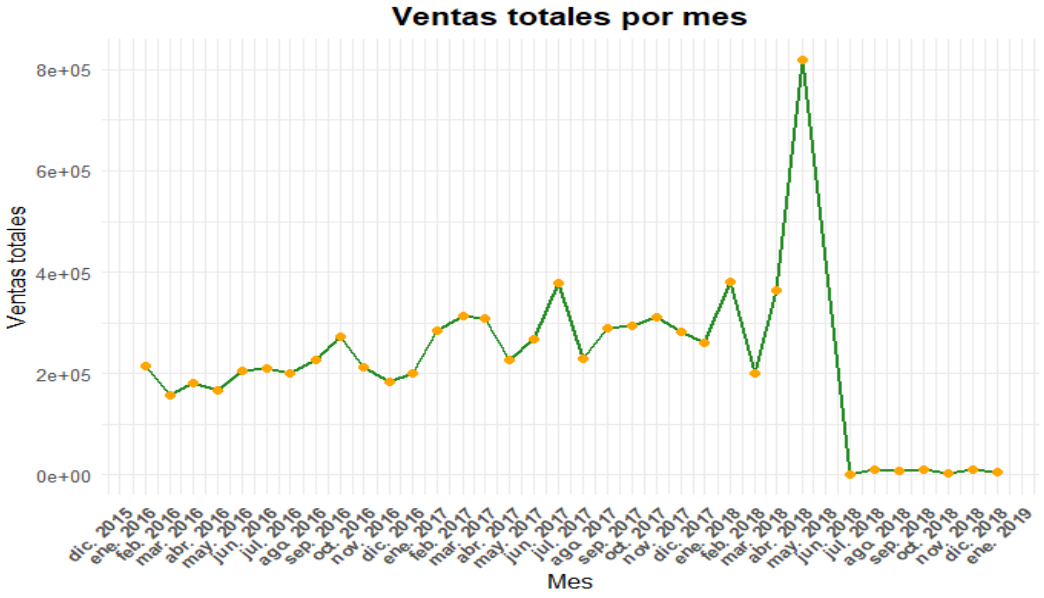
2.2 Análisis de órdenes

1. Al agrupar las órdenes por mes, observamos que en promedio se realizan entre 40 y 60 órdenes mensuales, con picos específicos en ciertos meses.



2. El descuento promedio en los ítems de la tabla **order_items** ronda el 10.54%, lo cual sugiere un nivel de promoción/ rebaja relativamente sustancial en ciertos productos.

3. Las ventas totales por mes presentan un comportamiento creciente (más o menos estable) hasta llegar a un pico muy alto en torno a mediados de 2018 y luego una caída abrupta. Lo anterior se ilustra en el gráfico de línea con ventas totales.



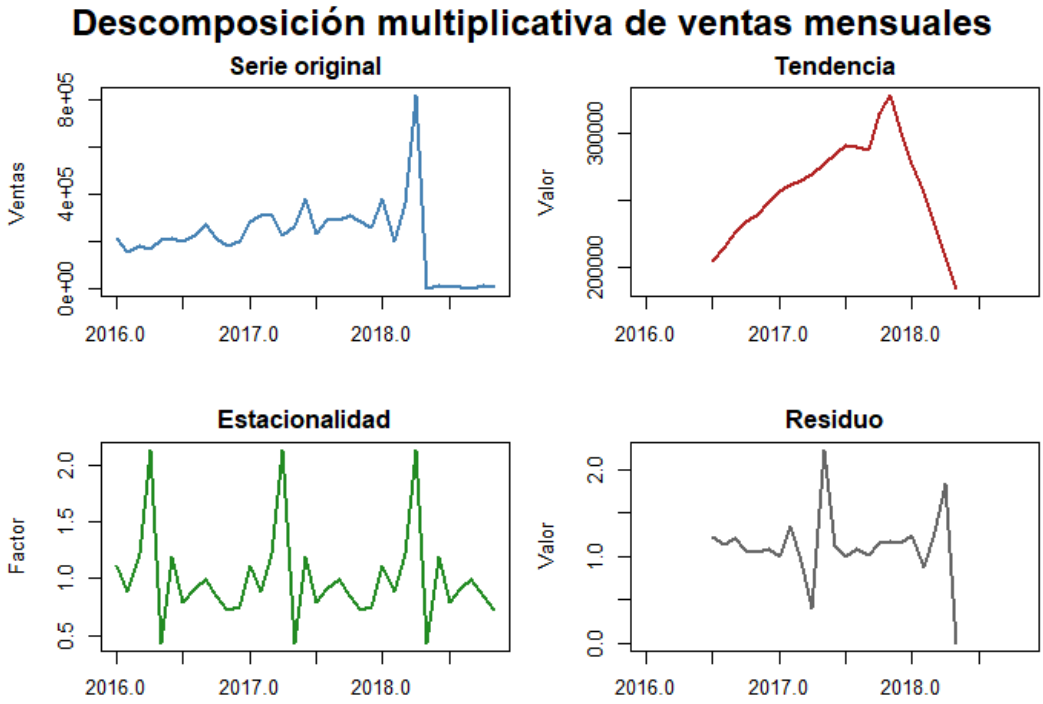
A continuación, se realiza la siguiente interpretación:

- La volatilidad observada en las ventas totales podría responder a campañas específicas y cambios de temporada.
- El descuento promedio de ~10.54% incide directamente en los márgenes de utilidad, por lo que conviene monitorizarlo a futuro.

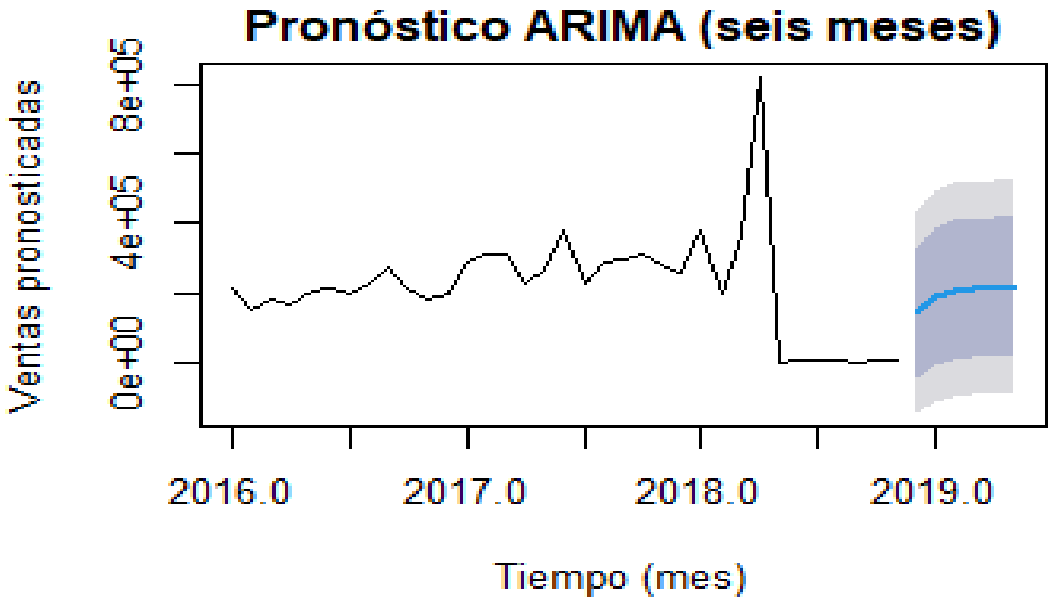
2.3 Análisis de series de tiempo y ARIMA

1. Se tomó la serie mensual de ventas totales (**ts_ventas_mensual**) para realizar una descomposición multiplicativa. El análisis gráfico muestra:

   - Tendencia (**trend**): Crecimiento durante 2016–2017, con un pico pronunciado en 2018 seguido de un descenso.
   - Estacionalidad (**seasonal**): Se aprecia un patrón repetitivo a lo largo de los meses con variaciones cíclicas marcadas.
   - Residuo (**random**): Variaciones no explicadas por tendencia ni estacionalidad.



2. Se ajustó un modelo ARIMA de forma automática con **auto.arima()**, obteniendo un **ARIMA(1,0,0)** (con media distinta de cero). Aunque la precisión puede mejorarse, se realizó un pronóstico a 6 meses que apunta a un posible rango de ventas totales entre ~150K y ~400K en el horizonte de predicción.



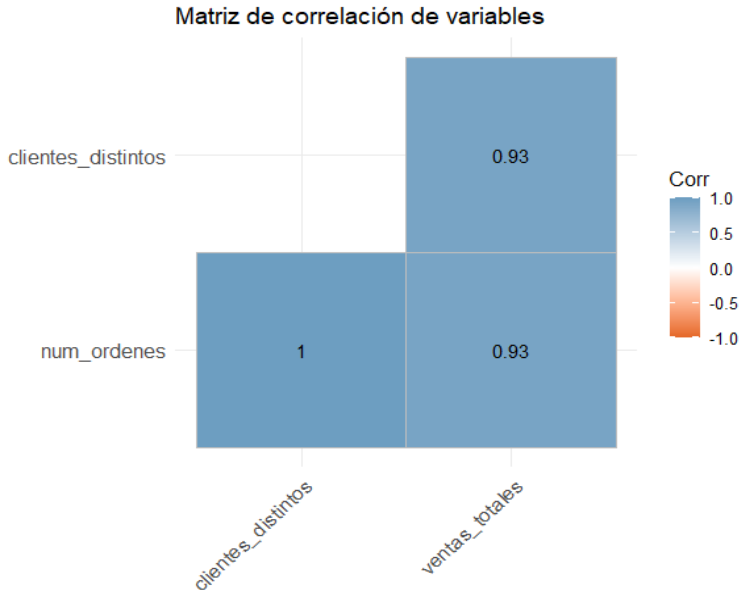
A continuación, se realiza la siguiente interpretación:

- La descomposición ilustra cómo gran parte del comportamiento global puede escindirse en componentes de tendencia y estacionalidad; sin embargo, la fuerte caída cercana a 2018 sugiere un evento extraordinario o cambio brusco de mercado.
- El **ARIMA(1,0,0)** no siempre será el mejor modelo para series con quiebres estructurales tan marcados. Por lo tanto, se deben revisar datos adicionales o usar modelos con cambios de régimen/intervención (por ejemplo, ARIMA con covariables o Prophet).

2.4 Regresión lineal múltiple

Se construyó un modelo para predecir las ventas (**ventas_totales**) a partir de:

- **num_ordenes** (número de órdenes mensuales).
- **clientes_distintos** (número de clientes distintos que compraron en ese mes).



A partir del **lm()** se obtiene lo siguiente:

- El coeficiente de determinación (R^2) es de ~0.86, lo que sugiere un ajuste alto en términos de correlación global. Sin embargo, los p-valores para ***num_ordenes*** y ***clientes_distintos*** (~0.302 y ~0.249, respectivamente) no indican significancia estadística al nivel 0.05 debido a la colinealidad entre ambas variables.
- La matriz de correlación corrobora que ***num_ordenes*** y ***clientes_distintos*** están fuertemente correlacionadas con ***ventas_totales*** (valores cercanos a 0.93), así como entre sí (perfectamente correlacionadas = 1.0 en el subset de datos). Lo anterior genera problemas de multicolinealidad en el modelo.
- Ahora bien, se deben incluir más variables (estacionales, promociones, día festivo, entre otros) para mejora la explicación de **ventas_totales**. Asimismo, se deben validar supuestos adicionales, como homocedasticidad (Breusch-Pagan test) y autocorrelación (Durbin-Watson).

A continuación, se realiza la siguiente interpretación:

- Aunque el R^2 sea elevado, la baja significancia (p-valores altos) indica que una gran parte de la variabilidad está capturada, pero no de forma estadísticamente diferenciable entre las variables. Por lo tanto, se deben introducir más predictores (descuentos, canales, estacionalidad, entre otros) o aplicar regularización (Ridge, Lasso) usando más variables.

3. Análisis Estadístico Adicional

3.1 T-test para diferencia de medias (importe) entre las diez ciudades con más vs diez con menos

Se comparó el importe promedio (columna ***importe = list_price * quantity * (1 – discount)***) entre las diez ciudades con más clientes (**Top_10**) y las diez con menos clientes (**Bottom_10**). El resultado es:

- El p-value ~ 0.2676, superior a 0.05, por lo que no se rechaza la hipótesis nula de que ambas medias sean iguales en la población.
- El intervalo de confianza al 95% incluye el cero ([-970.3151, 273.7785]), reforzando la conclusión de que no existe evidencia estadística de que difieran las medias de importe de compra en los dos grupos.

3.2 Test de normalidad (Shapiro-Wilk) de los residuos de la regresión

- Con un W = 0.94 y un p-value = 0.06305, no existe suficiente evidencia para rechazar la normalidad de los residuos al nivel de significancia de 5%.
- Los residuos se comportan aproximadamente de manera normal, cumpliendo uno de los supuestos centrales del modelo lineal.

A continuación, se realiza la siguiente interpretación:

- El T-test no mostró diferencias significativas en el gasto promedio según “tipo de ciudad” (top vs bottom), lo que podría indicar que la localización no es un factor determinante en el valor de las compras, al menos en esta muestra.
- La aproximada normalidad de los residuos es positiva para la validez general de la regresión, aunque se necesitan examinar otros supuestos (independencia, homocedasticidad, entre otros).