



Tecnológico de Monterrey

CAMPUS TOLUCA

Laboratorio de diseño y optimización de
operaciones

Proyecto final

Alumnos:

Roberto Arturo Gómez Mercado	A01365947
Miguel Padrón Vences	A01362804
Marcela Martínez y Martínez	A01381548

Profesora: Ana Luisa Masetto Herrera.

20 de noviembre de 2020.

Etapa 1: Comprensión del Negocio

1. Samsung en México

Debido a la crisis que nos encontramos actualmente, la venta de smartphones ha decaído en los últimos meses del año al no tener como primera necesidad en la población de México el comprar un teléfono inteligente. Sin importar lo mencionado, Samsung es una marca de teléfonos inteligentes quien se ha posicionado en ser marca líder en ventas en México durante los últimos años después de haber lanzado el Galaxy S3 donde tomó una buena posición en el mercado Mexicano. Según Forbes, Samsung sigue siendo la compañía que se mantiene a la cabeza del mercado en un 35%.

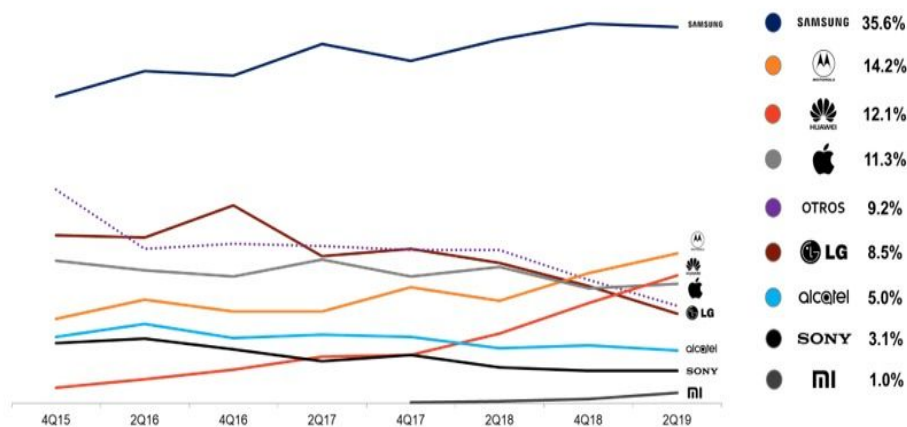


Figura 1. Comportamiento de las marcas en venta de smartphones (Forbes, 2019).

Aunque hay demasiadas cifras buenas, la marca no puede llevar gran ventaja competitiva dentro del mercado tecnológico, ya que hay algunas noticias donde se han presentado algunos problemas con ciertos modelos tal es el caso con el Galaxy Note 7, teniendo mala calidad, etc. Esto conlleva a perder usuarios y total abandono de la marca. Por esta y más razones es importante realizar un análisis de datos para averiguar los factores por los que cambia la demanda a través del tiempo, motivo para necesitar apoyo de un ingeniero industrial para comprender los datos y de forma pueden ser usados como en este caso cuyo objetivo es hacer un pronóstico de ventas mediante el estudio de ciencia de datos, también un ingeniero industrial puede ayudar analizando la calidad, comportamiento de las demandas a través del tiempo, entre mas cosas. A través de ello podremos ver qué comportamientos tiene la demanda de la marca, si presenta algún patrón o en qué puntos del tiempo hay irregularidades y esto se puede hacer con la ayuda de un software conocido como R-studio donde en el desarrollo del mismo presentaremos gráficos o los posibles datos afectados por algún factor.

2. Entender y describir la problemática (en términos del negocio).

La industria de las telecomunicaciones ha alcanzado grandes avances tecnológicos. El sector de los teléfonos celulares también ha presentado un aumento considerable en la demanda y esta a su vez genera mayor cantidad de necesidades en el mercado, ya que cada año los consumidores y/o usuarios se vuelven más exigentes cada marca hace lo posible por seguir innovando y mantener al usuario dentro de ella.

Es por eso que Samsung se ha caracterizado por su innovación tecnológica en ser la marca que vende celulares con chips rápidos igualando la velocidad a la de una computadora. Así mismo al fabricar pantallas OLED con gran nitidez, hablando de ello genera mayor consumo de energía. Por la cual, creó baterías inteligentes capaces de soportar cargas rápidas en tiempos mucho menores a los normales, dando esto a su fallo en calidad donde desplomó sus ventas en el año 2017 por la mala calidad en algunos dispositivos; creando desconfianza y haciendo que algunos de sus consumidores abandonaran la marca. Viendo esto reflejado hubo malas recomendaciones de los mismos. Añadiendo ese y más factores uno de los retos es que la demanda se vuelve cada vez más incierta y por lo tanto crea desconfianza al hacer un pronóstico de ventas, por medio del análisis de demandas anteriores se puede predecir qué número de ventas se esperan en periodos futuros pero debido a lo mencionado anteriormente, no sabes que tanto pueden bajar las ventas al que los usuarios abandonen la marca, haciendo esto es de tal importancia que no produzcas menos cantidad ya que genera mal servicio al cliente y a su vez inconformidad por parte de ellos, o una sobreproducción igual genera pérdidas ya sea niveles altos de inventarios y obsolescencia del producto. Es por esto que el reto es dar una predicción con muy cercana a lo real con mayor credibilidad.

3. Entender y describir la problemática (en términos de ciencia de datos).

El reto es desarrollar un proyecto de ciencia de datos para ayudar a la compañía enfocada a la distribución de estos productos a pronosticar el **número de unidades a vender de cada dispositivo móvil en sus diversos puntos de venta**; esto para evitar ciertos problemas como costos logísticos excesivos o insatisfacción. Se pretende crear un **modelo de regresión**, transformando los datos en información, para pronosticar las ventas en los distintos puntos y de esa manera, evitar riesgos.

4.Objetivos.

El objetivo de este proyecto es reforzar los conocimientos en Estadística y Programación para poder aplicar los métodos más avanzados de este proyecto de Ciencia de Datos.

También, con los conocimientos a adquirir en la materia, estructurar un proyecto de Ciencia de Datos y lograr manejar todos los datos posibles siendo una manera útil y fácil de entender, en este caso estructurar de una mejora manera los datos de Samsung.

Con ayuda de herramientas de programación como R-studio, aprender a usarlas de la mejor manera para mostrar o dar resultados precisos y de una manera más efectiva, de tal modo, será creando un modelo de predicción o pronóstico de ventas confiable con el menor error estadístico; ya que las demandas son inciertas y presentan diversos comportamientos a lo largo del tiempo.

5. Estructurar el proyecto y hacer un plan preliminar.

Las actividades realizadas a lo largo del desarrollo del proyecto, así como las fechas de inicio y terminación, se encuentran en la Tabla 1 de los anexos. A continuación, se muestra el diagrama de Gantt, figura X, que muestra el proceso del proyecto.

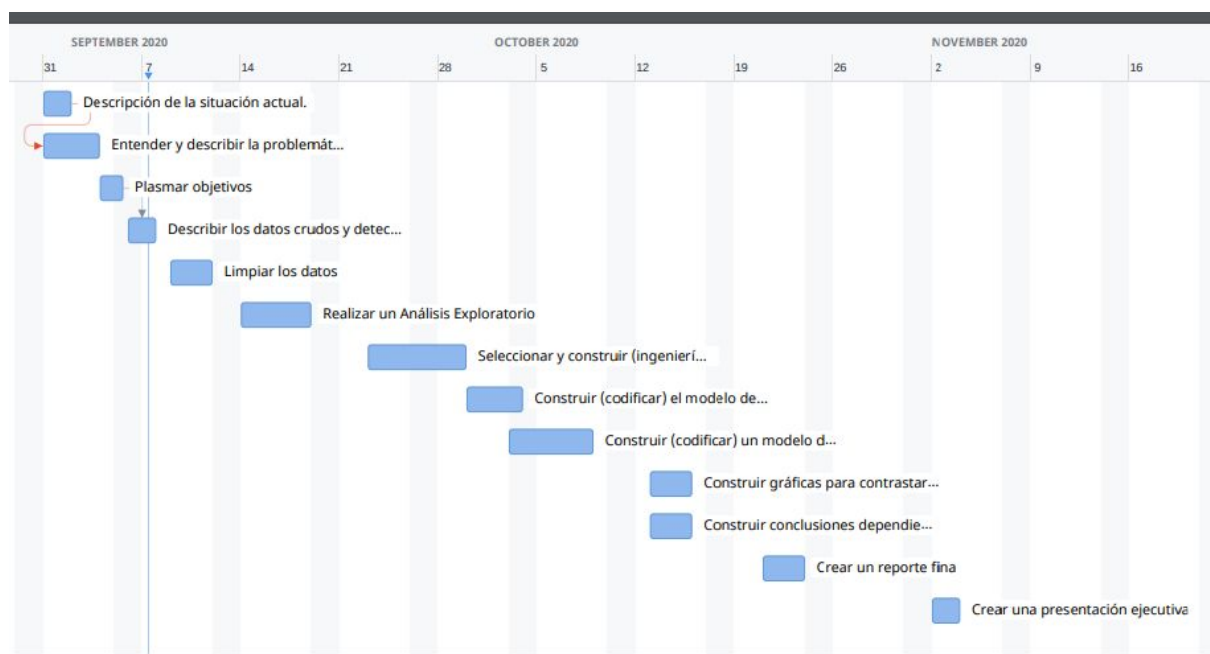


Figura 2. Diagrama de Gantt.

Etapas 2: Comprensión de los datos

Los datos mostrados a continuación son las ventas de teléfonos celulares de la marca Samsung en México, en los años 2018 y 2019. Dentro de los mismos hay errores de calidad, algunos son como faltas de ortografía, mala distinción entre ciudades y estados, variables numéricas como mes y año escritas en diferentes formatos o con letras y así sucesivamente con cada una de las variables; esto puede ser debido a que al capturador de datos no se le explica de qué manera deben capturarse, y para ahorrar tiempo no revisa el contenido, o la mala programación al crear formularios donde deba verificar que sea la información solicitada, por ejemplo, el año o precio son variables numéricas lo cual debe verificar que sean dígitos ingresados solamente. De igual forma en otros que solo sean tipo caracter o factores. Revisar y limpiar estos errores es necesario para que todos los datos sigan un solo formato. De igual manera el uso de letras mayúsculas o minúsculas.

- Punto de venta: Diferentes sucursales donde la empresa opera (lugares).
- Fecha: Fecha del día mes y año que fue vendido algún dispositivo.
- Mes: El mes que se está haciendo esa venta pero solo puede ser numérico, comenzando en 0 con junio 2018 y terminando en 10 con marzo 2019.
- Año: Año de la venta de ese dispositivo.
- Número de ventas: Es el número de celulares o unidades que están siendo vendidas ese día.
- Sku: Modelo del celular vendido, como un folio del modelo.
- Marca: Marca del teléfono, en este caso todos pertenecen a la misma.
- Gamma: Es como son filtrados los teléfonos a través de su calidad, rendimiento y costo.
- Costo: El precio pagado del teléfono vendido.
- Zona: Zona en la que se encuentra localizado algún punto de venta.
- Estado: Estados de la república donde están los distintos puntos de venta.
- Ciudad: Ciudad donde se ubica el centro de venta.
- Latitud: Coordenadas geográficas para precisar el punto de venta.
- Longitud: Coordenadas geográficas para precisar el punto de venta.

Etapla 3: Preparación de los datos.

Al resolver un problema sobre ciencia de datos, se busca responder a ciertas preguntas que ayuden a reducir costos o hacer más eficiente la cadena de suministro. Por tal razón se busca el apoyo de herramientas como “*Machine Learning*”, que ayuden a entender mejor la situación llevando a cabo una serie de pasos previos a la respuesta, uno de ellos es conocer y entender los datos sobre los que se va a trabajar. A continuación se muestra la preparación de datos sobre las ventas de smartphones de la marca Samsung en México durante los años 2018 y 2019.

Limpieza de datos

En el momento de visualizar y familiarizarse con los datos, existen muchos problemas a resolver antes de desarrollar tu modelo y obtener un pronóstico certero, por ello la primera fase es limpiarlos. Existen casos en donde los capturadores no agregan los datos de la forma adecuada, se usan diferentes formatos o llegan a tener faltas de ortografía generando inconsistencia o mala interpretación de los mismos. Para hacer dicha limpieza se utilizó R-Studio.

En el proceso de limpieza encontramos los siguientes errores:

- Combinación de las palabras con letras mayúsculas y minúsculas.
- Faltas de ortografía en los puntos de venta. Asimismo, puntos de venta mal escritos.
- Los meses estaban en diferentes formatos (numérico y texto).
- El año, diferente número de dígitos.
- Marca “*Samsung*”, escrita incorrectamente varias veces.
- Zona, algunas con mayúsculas.
- Algunas ciudades fueron registradas como estados.
- La longitud y latitud, valores fuera de rango que no se ubican en el territorio mexicano.

Para corregir los errores anteriores se aplicaron varias estrategias. La primera fue cambiar todas las mayúsculas a minúsculas para que se mantuviera un formato estándar. Se corrigieron las faltas de ortografía en los puntos de venta y en la marca “*Samsung*”, y se cambiaron todos los meses a formato numérico.

En la variable “Año”, se estandarizó a 4 dígitos. Por último, las ciudades mal registradas se reescribieron en el estado al que pertenecen, además de que los valores de longitud y latitud se adecuaron a los valores más cercanos de otros puntos de venta existentes en el país.

Análisis exploratorio

El análisis exploratorio permite tener una visualización de lo que está ocurriendo. Para que no existan incongruencias, el análisis se realiza después de la limpieza de datos.

La manera más fácil de lograr el entendimiento de los datos es mediante preguntas que guíen la investigación; cada pregunta formulada revela nuevos aspectos de los datos.

Para profundizar en los datos, se plantearon y respondieron las siguientes preguntas:

1. ¿Qué rango de fechas abarca el data set?

En los datos existen registros desde junio 2018 hasta marzo 2019.

2. ¿Cuántas y cuáles son las gamas de los teléfonos?

Se encuentran 4 gamas: Premium, Alta, Media y Baja.

3. En total, ¿cuántos puntos de venta existen?

1788 puntos de venta alrededor del país.

4. ¿Cuál gama de equipos es la más vendida?

En la figura 3, se muestra que la gama “Baja” de la marca Samsung es la más vendida con una amplia brecha con la siguiente más vendida que es la gama alta. Del total de equipos vendidos, tan solo el 3.6% son de gama premium.

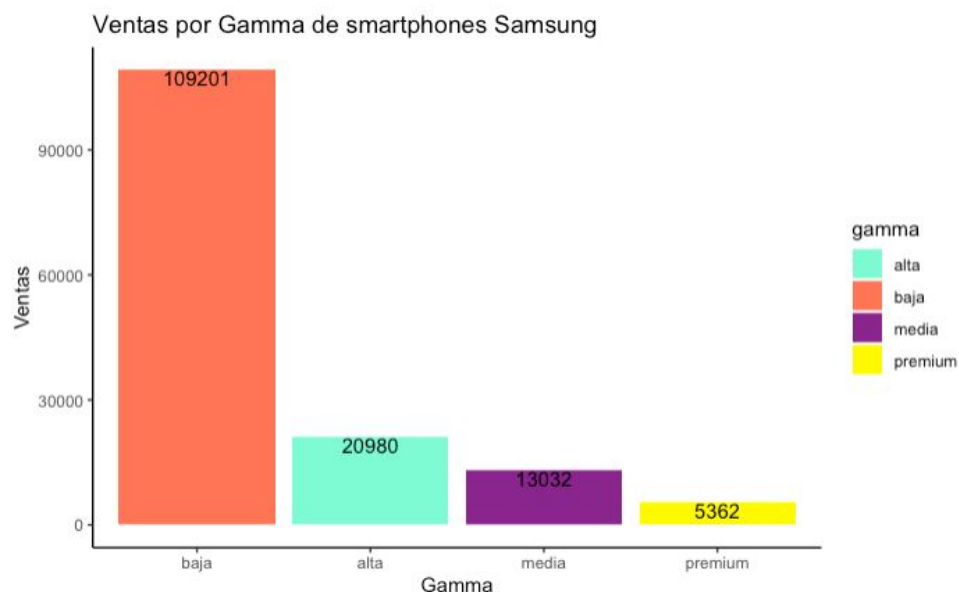


Figura 3. Ventas frente a Gama de smartphones.

5. ¿Cuál es el precio de venta promedio de cada gama?

Algo que influye directamente con las ventas de cada gama es el precio de éstas.

En la siguiente figura se exhiben los precios de cada gama. Se observa que la gama “Baja” es la de menores precios, oscilando entre los \$1,800 y \$5,000, con una media aproximada de \$2,800. La siguiente, en cuanto a precios, difiere de la segunda gama más vendida, ya

que se venden más celulares de gama “Alta” que de gama “Media” a pesar de que el precio de venta medio de la primera es de casi el doble.

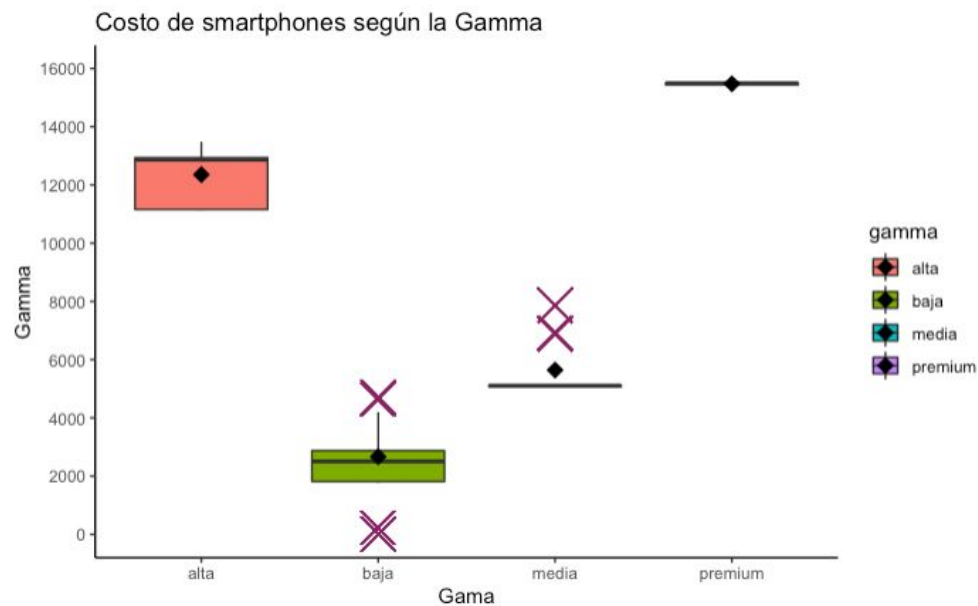


Figura 4. Precio de smartphones según la gama.

6. ¿En qué mes se registraron más ventas?

Diciembre es el mes en el que se registra el mayor número de ventas. Se observa que los últimos meses del año se compran más celulares. Noviembre y Diciembre ocupan los primeros lugares en ventas, mientras que marzo, es el mes en el que menos celulares se venden, esto pudiera ser debido a que a finales del año salen promociones y muchas personas reciben su aguinaldo usandolo para comprar dichos productos y por lo tanto, durante los primeros meses en México regularmente se hacen pagos anuales de servicios como agua, predial, etc (Figura 5).

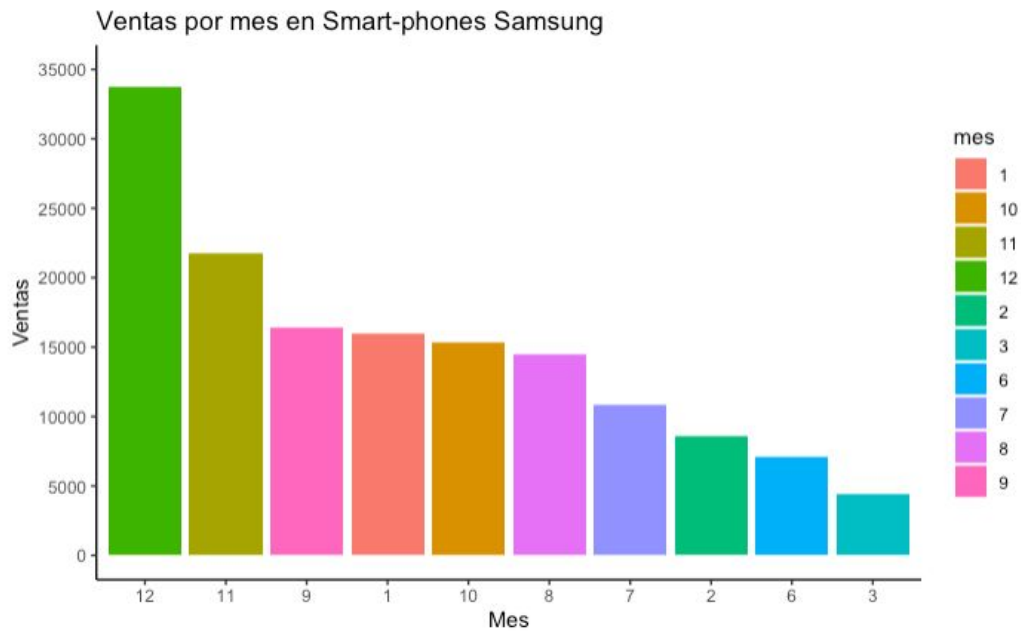


Figura 5. Ventas según los meses.

7. ¿En qué estados se presentan las mayores ventas?

Aunque en todos los estados hay ventas de teléfonos, la Ciudad de México, que cuenta con 270 puntos de venta, registra el mayor número de ventas. El Estado de México ocupa el segundo lugar en ventas, con poco más de 20,000 ventas, y le siguen Jalisco, Nuevo León, Guanajuato y Veracruz, 10,000 ventas por debajo del segundo estado. Los estados con menos ventas son Nayarit (746), Zacatecas (686) y Durango (629). (Figura 6).

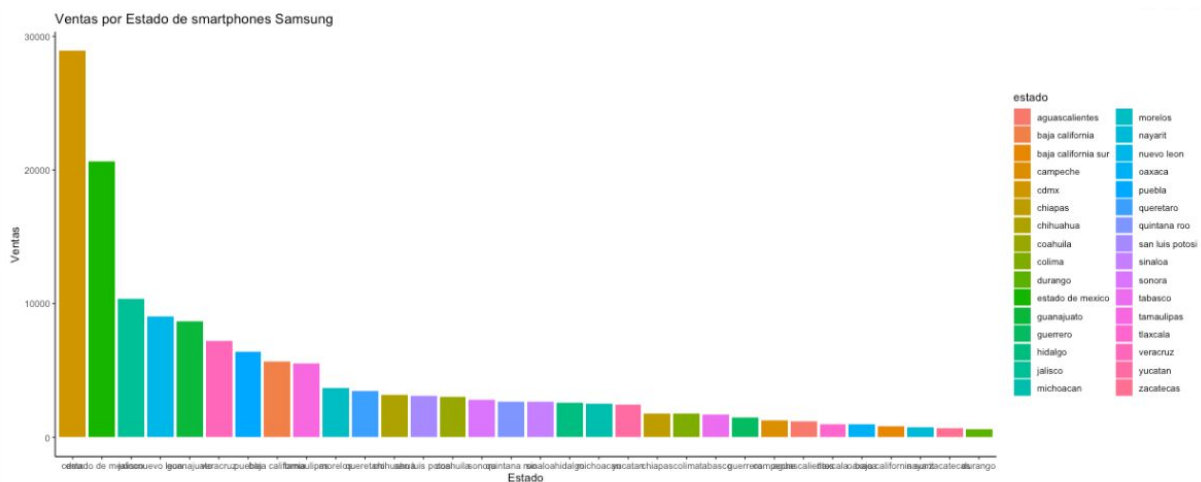


Figura 6. Ventas por estado

8. Dentro de la Ciudad de México, ¿cuál es el punto de venta más popular?

Dentro del estado con más ventas, el punto de venta más popular es un centro de telefonía en la plaza Fórum Buenavista.

Ingeniería de características

Este proceso es uno de los más importantes y de donde se puede obtener la información más valiosa. El objetivo es eliminar las variables que no son relevantes para hacer la predicción y también crear nuevas variables a partir de los datos limpios para aumentar el poder predictivo del modelo y agregarle valor.

Las variables que se van a trabajar son Punto de venta, Mes¹ y SKU (Stock Keeping Unit), esto, porque son variables importantes para responder a la pregunta del proyecto de cuántas unidades se van a vender en el siguiente periodo en todos los puntos de venta. Son variables cualitativas y para facilitar su manejo, se les colocó un índice numérico.

Estas variables se unieron (*left join*) con el *data frame* original, creando en éste columnas nuevas.

En cuanto a las ventas totales, en el *data frame* original, se encontraban ventas únicas por equipo en cada punto de venta, las cuales se agruparon para tener las unidades vendidas en total en cada punto y en qué mes se realizaron las ventas.

Enseguida, para tener los datos existentes en su totalidad, incluso con los faltantes o de los que no hay registro, se multiplicaron los números de filas de los *data frames* creados (punto de venta, mes y SKU). El resultado, 643.680, son las filas en total que se encuentran dentro de los registros. Se prosiguió a combinar los *data frames* y solo se muestra cuántas unidades se vendieron, en qué punto de venta de cada mes registrado, todo plasmado con índices numéricos.

Como se mencionó anteriormente, la variable respuesta que se predice en este proyecto es el número de unidades que se venderán en el siguiente mes en cada punto de venta, por lo que, con base en los procesos de pronósticos, se aplicó un desfase de una fila a los datos, por lo que el pronóstico se fija en las unidades vendidas registradas del siguiente mes.

En la creación de nuevas características, se tomó como base el ejercicio realizado en clase, sin embargo, se realizaron algunos cambios, los cuales se enlistan a continuación:

1. En cuanto a la variable de ventas totales de cada SKU, se tomó la decisión de simplemente trabajar con los valores promedio para facilitar el análisis.
2. En los rezagos también se mantuvieron las ventas totales de 4 meses anteriores, el promedio por tienda de 4 meses pasados, y las ventas promedio por SKU de 4 meses.

Al finalizar el proceso, se cuenta con **579,312 filas** y **20 columnas** en el dataframe.

¹ Se utiliza la variable “Mes” y no “Fecha” porque el periodo base del pronóstico es el mes.

Etapla 4: Modelado.

Modelo promedios móviles.

Se eligió este modelo porque se busca pronosticar el número de ventas para el próximo mes en cada punto de venta, así como de cada tipo de celular, por lo que se creará un modelo de estadística tradicional que nos entregará como respuesta final: “Unidades vendidas al siguiente mes”. Un promedio móvil es un indicador de tendencias, siendo utilizado para analizar los datos anteriores y formar una serie de medias provenientes de subconjuntos de datos, por lo tanto, tienen la capacidad de examinar las medias que convergen en un periodo de tiempo y así finalmente son suavizados los datos o la tendencia. Tales medidas se pueden clasificar como conductores o rezagos, que son manifestadas como tendencias porque se utilizan datos históricos.

Por sí solos, no son predictores muy precisos del movimiento de los datos. Deben combinarse, es decir, una media móvil de corto plazo y una media móvil de largo plazo para producir una señal, o combinarse con otros indicadores que puedan medir el impulso de la acción de los precios, como los osciladores y así dando origen a 2 tipos de promedios móviles.

- Promedio móvil simple.
- Promedio móvil ponderado.

Ambos tipos comparten características importantes como:

- ❑ Logran estimar valores futuros extraídos de datos históricos.
- ❑ Diversidad de técnicas → suavizamiento.
- ❑ Facilitan información que tienden a enmascarar por una medida.
- ❑ Cuando los periodos usados son pequeños, el pronóstico puede responder con mayor rapidez a cualquier cambio que se presente.

Es evidente que en todo método existen riesgos y se procura tener una respuesta certera aunque a veces no todo es positivo en un método fácil de diseñar como es este caso, por esa razón presentamos alguna lista de ventajas y desventajas.

VENTAJAS	DESVENTAJAS
Permite la aproximación del futuro, lo que puede llegar a facilitar la toma de decisiones.	Pronostican solo un periodo más, el siguiente.

Los pronósticos planificados tienen mucho más valor y exactitud que los que son intuitivos.	Suelen ser simplificaciones reales y no garantizan las variables influyentes en el futuro de los pronósticos que se encuentren incluidos en el modelo de dicho pronóstico.
---	--

Tabla 1. Ventajas y desventajas de los promedios móviles.

A pesar de que este método puede mejorar la tendencia mediante diferentes técnicas de suavizamiento para acercarse a lo más real posible de nuestro evento futuro, existen algunas métricas que nos ayudan reducir esa incertidumbre que nos genera al pensar que tan bueno podría ser nuestro pronóstico. Estas métricas, conocidas como errores, son bastante útiles evitando grandes desperdicios o problemas de manera anticipada.

Entre ellos, los más conocidos son:

- **CFE.**
- **MAD.**
- **MSE.**
- **MAPE.**
- **MAE.**

Para el desarrollo del pronóstico sobre las ventas de la compañía “Samsung”, el primer paso fue conocer qué variables serían las que nos ayudarán a encontrar nuestro resultado, de las cuales solo seleccionamos: punto de venta, mes, sku, ventas totales, ventas del siguiente mes. Las mencionadas anteriormente nos ayudarían a encontrar nuestro pronóstico en diferentes temporalidades, es decir, usamos promedios móviles para 1 mes, 2 meses y 3 meses de acuerdo a nuestros datos históricos como se muestra en el Anexo 2.

Se hizo un cálculo de errores para cada mes obteniendo un error conocido como MAE “Mean Absolute Error”, error absoluto medio de cada mes de acuerdo al tipo de modelo (media de 1, 2 o 3 meses), y así finalmente medir la precisión de nuestros pronósticos para obtener nuestras conclusiones como ver cual modelo es más estable y cuál tiene un error menor para así identificar el mejor. Anexo 3.

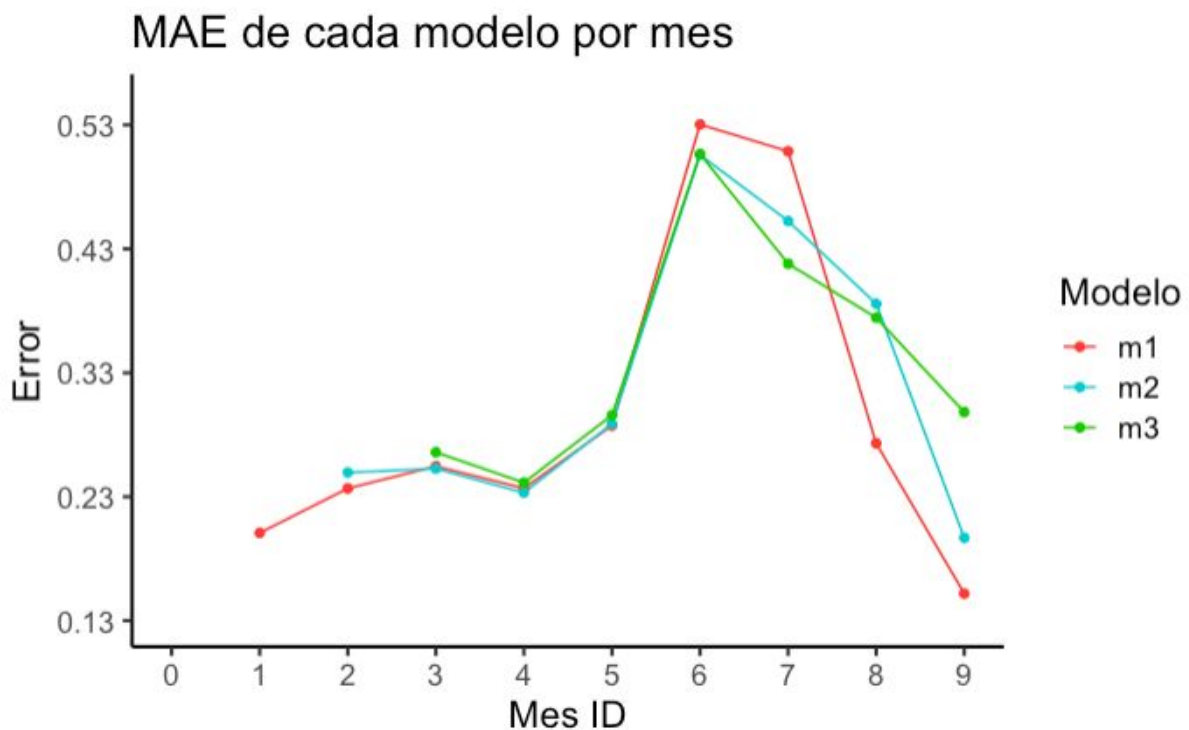


Figura 7. Gráfica Error MAE de cada modelo por mes.

Comparando el error absoluto medio de cada modelo de promedios móviles (1 mes, 2 meses y 3 meses), se concluye que el modelo de predicción con un parámetro de dos meses de anterioridad, ya que es el de menor variación.

Modelo de aprendizaje de máquina.

Para finalizar, cabe resaltar la gran cantidad de modelos para resolver problemas en ciencia de datos desde modelos con estadística tradicional hasta redes neuronales; de tal modo en este caso se resolvió con un método de aprendizaje de máquina supervisado, en otras palabras para entender de mejor forma un modelo de ML es usado para aprender de los datos generando reglas y mejorar el desempeño aprendiendo por sí mismos el comportamiento de dichos y así generar una respuesta que ayuda a tomar una decisión lo más certera posible y a su vez estos datos aprenden ya que se necesitó de nuestra ayuda para alimentar el modelo y entrenarlo como se explicó anteriormente.

Por lo tanto, una mejor manera de explicar cómo se llegaron a los resultados fue mediante el uso de modelos de árboles de decisión, y bosques aleatorios en Jupyter Notebook. Para

entender mejor la predicción del número de unidades que se venderán en el siguiente mes de cada tipo de producto en los diferentes puntos de venta; un árbol de decisión analiza todas las variables de cualquier tipo aunque no tengan relación entre ellas para dar un resultado final y hacer diferentes suposiciones con ellas o también se pueden elegir las que afectan mejor a nuestra variable respuesta, cabe resaltar que en algunos casos son ineficientes al crear árboles con profundidades largas haciendo sesgos o outliers; dichos árboles son formados por una raíz, es decir, una pregunta a la que se busca dar respuesta, condición de la que se despliegan nodos y finalmente las hojas que son las respuesta a la pregunta. Una vez realizado para medir su desempeño qué tan bueno es o preciso se usan errores como en la estadística tradicional que mencionamos anteriormente y uno de ellos fue **MAE** “Error medio absoluto” con 2 parámetros diferentes los cuales fueron la profundidad del árbol uno fue 1 y el otro con profundidad de 5 ramas.

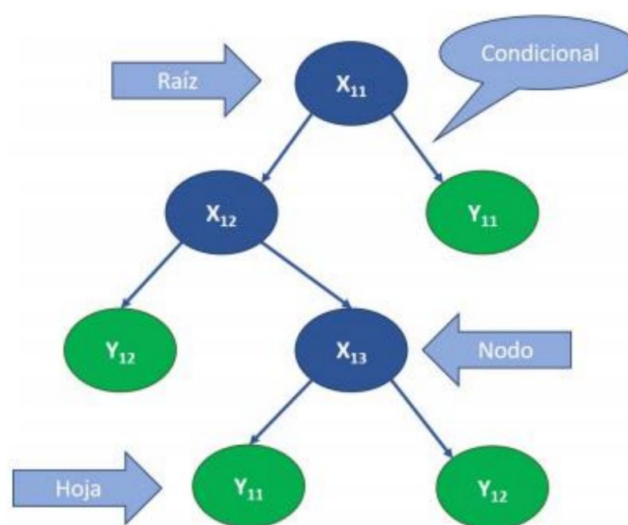


Figura 8. Estructura de un árbol de decisión.

Añadiendo otro modelo para dar una respuesta más concisa fue uno similar a los árboles y esta vez hablamos de los “bosques aleatorios”, dicho método toma de forma aleatoria los datos haciendo muchas iteraciones para generar una cantidad indicada de árboles de decisión, con este modelo también usamos 2 parámetros el de generar 100 y 500 árboles, a diferencia del anterior no sobreajuste del modelo y puede usar una dimensión grande de datos dando un pronóstico más robusto, aunque no se pueda tener mucho control de los datos como en los árboles de decisión ya que éste los toma de manera aleatoria, por esta razón, igual medimos el desempeño de la misma con un error “MAE” quien nos arrojó un error mas grande a diferencia de los demás donde se explica detalladamente más adelante.

Etapa 5. Evaluación

Para concluir, cabe mencionar una gran problema económico al que nos enfrentamos actualmente, pudiera ser el decremento de ventas en smartphones, debido al desempleo o que las necesidades cambian en algunas personas por las que comprar un nuevo smartphone no sea de prioridad, por lo tanto pudiera incrementar la incertidumbre para Samsung. Al pasar por este y más problemas es importante como estrategia en Samsung conocer con precisión el número de unidades que se venderán en periodos de tiempo posteriores en cada punto de venta, ya que las ubicaciones en cada punto de venta los consumidores pudieran tener diferentes necesidades o preferencias, ya sea por las clases socioeconómicas, características, recomendaciones de boca en boca, etc. Por tal motivo, como hemos mencionado anteriormente se desarrollaron 4 modelos diferentes en el que tienen parámetros diferentes para tener la respuesta a nuestra pregunta ¿Cuál será la venta de los distintos tipos de artículos para el o los siguientes meses dependiendo en cada punto de venta?

Modelo	Parametros	
Arbol de decisión	1	Profundidad
	5	
Bosques Aleatorios	100	Árboles
	500	

Tabla 2. Modelos con especificación de parámetros

Con los 4 modelos diferentes obtuvimos en total el número de unidades que venderemos el próximo mes en este caso marzo, gracias al aprendizaje hacen los modelos analizando el comportamiento de los datos, como afectan entre ellas las variables y haciendo varias suposiciones, entrenamiento de los mismos para evitar sobreajuste a continuación en la próxima tabla se muestra la cantidad en total de unidades a vender en México .

Modelo	Pronóstico
Árbol de decisión con profundidad 1.	1,947 unidades.
Árbol de decisión con profundidad 5.	6,978 unidades.
Bosques aleatorios de 100 árboles.	13,016 unidades.
Bosques aleatorios de 500 árboles.	13,394 unidades.

Tabla 3. Pronóstico de cada modelo.

Al ver los resultados, a simple vista no sabemos cual es más certero y viéndolo así son cantidades completamente diferentes que no convergen en algún punto; anteriormente, mencionamos algunas herramientas para medir su desempeño de los mismos y de igual forma se explicó que para esta ocasión usamos el error medio absoluto por sus siglas “MAE”, obteniendo el error de cada mes de las ventas tanto en el conjunto de entrenamiento como en el de prueba.

	Metrica	Conjunto	Mes	MB_pm	DT_1	DT_5	RF_100	RF_500
0	MAE	Entrenamiento	Julio	NaN	0.202228	0.176174	0.162348	0.161944
1	MAE	Entrenamiento	Agosto	NaN	0.195175	0.182327	0.133436	0.132962
2	MAE	Entrenamiento	Septiembre	NaN	0.206718	0.191337	0.119925	0.119853
3	MAE	Entrenamiento	Octubre	NaN	0.208714	0.198022	0.103036	0.102718
4	MAE	Entrenamiento	Noviembre	NaN	0.230872	0.216117	0.101920	0.101706
5	MAE	Entrenamiento	Diciembre	NaN	0.276866	0.267009	0.297221	0.296799
6	MAE	Entrenamiento	Enero	NaN	0.276786	0.251744	0.098365	0.097457
7	MAE	Entrenamiento	Febrero	NaN	0.262850	0.246957	0.089606	0.088786
8	MAE	Prueba	Agosto	0.249557	0.202228	0.176174	0.162348	0.161944
9	MAE	Prueba	Septiembre	0.252703	0.237385	0.279564	0.270025	0.267478
10	MAE	Prueba	Octubre	0.233260	0.214703	0.237773	0.270989	0.270538
11	MAE	Prueba	Noviembre	0.288901	0.301345	0.280497	0.269078	0.269093
12	MAE	Prueba	Diciembre	0.505873	0.506836	0.522496	0.517649	0.515707
13	MAE	Prueba	Enero	0.452500	0.335213	0.556099	0.658775	0.656848
14	MAE	Prueba	Febrero	0.385727	0.165300	0.216738	0.360707	0.361872
15	MAE	Prueba	Marzo	0.196891	0.081376	0.123182	0.188215	0.192471

Figura 9. Error MAE de los 4 modelos.

A simple vista es difícil entender los datos o visualizar el comportamiento de ellos, ¿Cómo elegir el mejor? Con la ayuda de Júpiter fue fácil graficar los errores mostrados antes y ahora se muestran de la siguiente forma durante los meses de julio a marzo, en donde hace una comparación de las predicciones a los datos reales de esas mismas fechas en ambos conjuntos de datos.

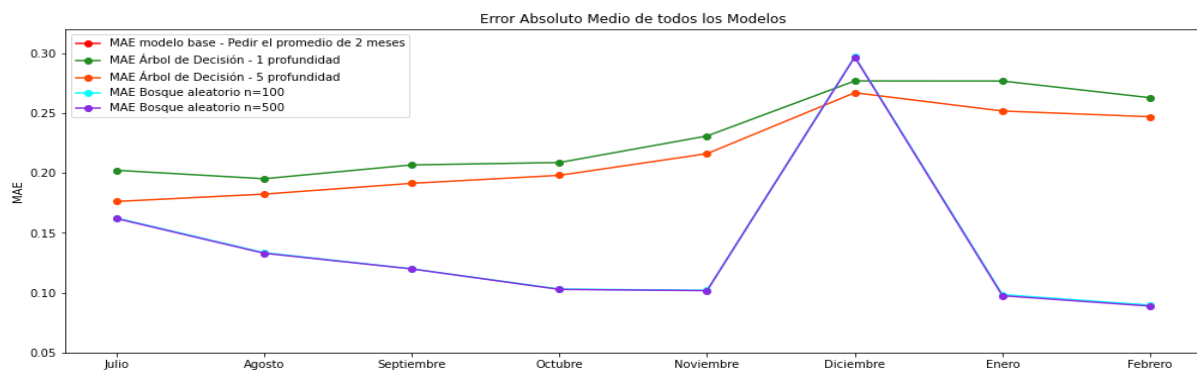


Figura 10. Error MAE para el conjunto de entrenamiento.

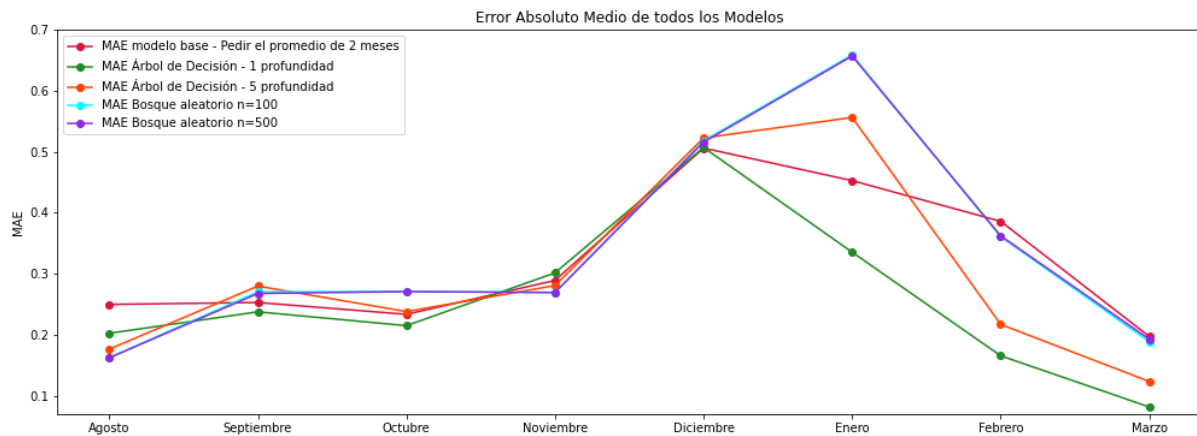


Figura 11. Error MAE para el conjunto de prueba.

Finalmente, podemos decir que de los 4 modelos realizados el más confiable es el bosque aleatorio de 100 árboles debido a que sigue de mejor manera la estacionalidad de cada periodo y teniendo estabilidad, los árboles pueden mostrar de momento mejor desempeño pero cuando tienen más información tienden a crear sesgos, y pensando a futuro al ser una empresa líder en ventas en México alimentaremos con más datos en ventas posteriores para que el programa aprenda a manejarlos para mejorar las predicciones y reducir esos picos con el paso del tiempo, añadiendo que genera respuestas más robustas ya que tiene la capacidad de utilizar más datos que un árbol al hacerlo más inteligente en casos como este y evita que haya sesgos como sucede con los árboles; por lo tanto le recomendamos a Samsung preparar su producción para la venta de 13,016 unidades en México solamente, a pesar de la pandemia es una cantidad buena para seguir estando en la cabeza como marca líder en ventas e innovación con productos con la mejor calidad.

Anexos.

Tabla 1.

Tareas	Fecha de inicio	Fecha de Terminación
Plasmar objetivos	31/08/2020	01/09/2020
Descripción de la situación actual.	31/08/2020	03/09/2020
Entender y describir la problemática (en términos del negocio y ciencia de datos)	04/09/2020	04/09/2020
Describir los datos crudos y detectar problemas de calidad	06/09/2020	07/09/2020
Limpiar los datos	09/09/2020	11/09/2020
Realizar un Análisis Exploratorio	14/09/2020	18/09/2020
Seleccionar y construir (ingeniería de características) variables para la etapa de modelado.	23/09/2020	29/09/2020
Construir (codificar) el modelo de promedios móviles, cambiando 2 veces el periodo móvil.	30/09/2020	03/10/2020
Construir (codificar) un modelo de aprendizaje de máquina.	03/10/2020	08/10/2020
Construir gráficas para contrastar el desempeño de los modelos	13/10/2020	15/10/2020
Construir conclusiones dependiendo de los resultados obtenidos.	13/10/2020	15/10/2020
Crear un reporte final	21/10/2020	23/10/2020

Anexo 2.

```
datos_ma <- samsung_completos %>% select(pdv_id, mes_id, sku_id, ventas_totales, y_ventas_siguiente_mes)
datos_ma
```

pdv_id <int>	mes_id <int>	sku_id <int>	ventas_totales <int>	y_ventas_siguiente_mes <int>
1	0	1	1	0
1	1	1	0	0
1	2	1	0	0
1	3	1	0	0
1	4	1	0	0
1	5	1	0	0
1	6	1	0	0
1	7	1	0	0
1	8	1	0	0
1	0	2	1	0

1-10 of 579,312 rows

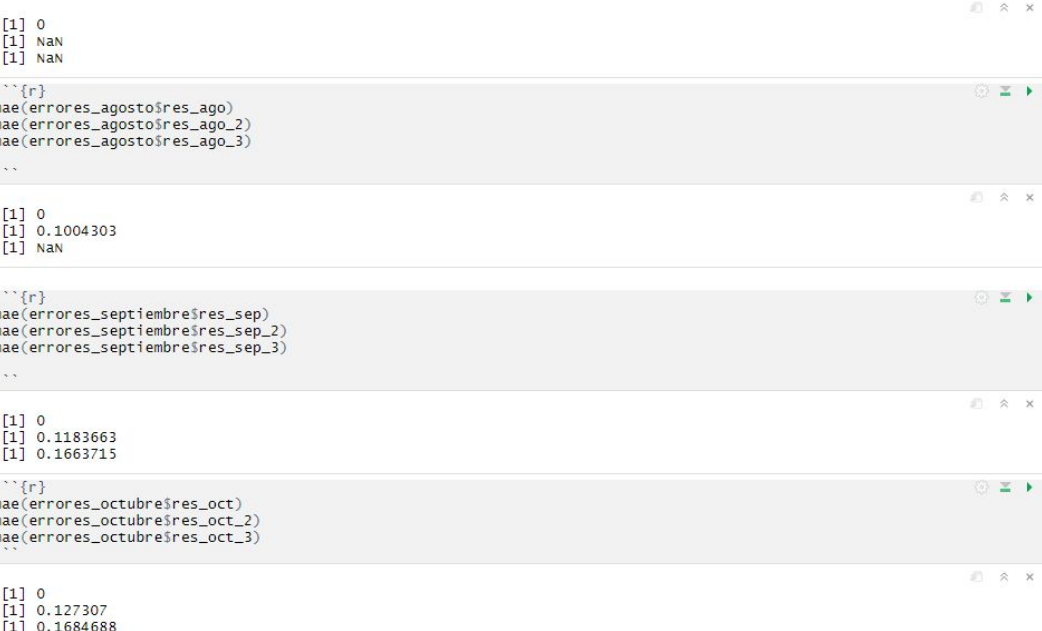
```
{r}
datos_ma <- datos_ma %>% group_by(pdv_id,sku_id) %>%
  mutate(pedir_mes_pasado = rollmean(ventas_totales, k= 1, fill = NA, align = "right")) %>%
  mutate(pedir_2meses_pasado = rollmean(ventas_totales, k= 2, fill = NA, align = "right")) %>%
  mutate(pedir_3meses_pasado = rollmean(ventas_totales, k= 3, fill = NA, align = "right"))
head(datos_ma,50)
```

pdv_id <int>	mes_id <int>	sku_id <int>	ventas_totales <int>	y_ventas_siguiente_mes <int>	pedir_mes_pasado <dbl>	pedir_2meses_pasado <dbl>
1	0	1	1	0	1	NA
1	1	1	0	0	0	0.5
1	2	1	0	0	0	0.0
1	3	1	0	0	0	0.0
1	4	1	0	0	0	0.0
1	5	1	0	0	0	0.0
1	6	1	0	0	0	0.0
1	7	1	0	0	0	0.0
1	8	1	0	0	0	0.0
1	0	2	1	0	1	NA

1-10 of 50 rows | 1-7 of 8 columns

Anexo 3.

```
104 mae <- function(error)
105 {
106   mean(abs(error), na.rm = TRUE)
107 }
108
109
110
111
112
113 maeerrores_juliosres_jul)
114 maeerrores_juliosres_jul_2)
115 maeerrores_juliosres_jul_3)
116
117
118
119 maeerrores_agosto$res_ago)
120 maeerrores_agosto$res_ago_2)
121 maeerrores_agosto$res_ago_3)
122
123
124
125
126 maeerrores_septiembre$res_sep)
127 maeerrores_septiembre$res_sep_2)
128 maeerrores_septiembre$res_sep_3)
129
130
131
132 maeerrores_octubre$res_oct)
133 maeerrores_octubre$res_oct_2)
134 maeerrores_octubre$res_oct_3)
135
```



Month	MAE (res)	MAE (res_2)	MAE (res_3)
July	0	NaN	NaN
August	0	0.1004303	NaN
September	0	0.1183663	0.1663715
October	0	0.127307	0.1684688

Referencias:

- Statistics, A. (2020). Top Android phones and tablets in Mexico | AppBrain. Recuperado el 6 Septiembre 2020, de: <https://www.appbrain.com/stats/top-android-phones-tablets-by-country?country=MX>.
- Guisado, G. (2019). *Introducción a la ciencia de datos en la Ingeniería Industrial* [Ebook]. Leganés: Universidad Carlos III de Madrid. Recuperado de: <https://core.ac.uk/download/pdf/288502311.pdf>.
- Erik Peñaloza y Regina Zúñiga. (2019). Samsung seguirá reinando en México pese a esfuerzos de Apple y Huawei. 05/09/2020, de FORBES Sitio web: <https://www.forbes.com.mx/samsung-seguira-reinando-en-mexico-pese-a-esfuerzos-de-apple-y-huawei/>.
- ANTONIO CAHUN. Huawei es la segunda marca que más smartphones vendió en México a finales de 2019, pero Samsung y Motorola mantienen su dominio. 05/09/2020, de Xataka Sitio web: <https://www.xataka.com.mx/telecomunicaciones/huawei-segunda-marca-que-smartphones-vendio-mexico-a-finales-2019-samsung-motorola-mantienen-su-dominio>.
- SHERISSE PHAM. (2019). Samsung advierte sobre la caída en sus ganancias al arranque del 2019. 05/09/2020, de EXPANSION Sitio web: <https://expansion.mx/tecnologia/2019/03/27/samsung-advierte-sobre-la-caida-en-sus-ganancias-al-arranque-del-2019>.