Tecnológico de Monterrey

Campus Toluca

Segundo Avance del Proyecto
Reporte 1

Materia: IN3038.1

Laboratorio de Diseño y Optimización de Operaciones

Profesora: M. en C. Ana Luisa Masetto Herrera

Equipo número 3:

Alejandra Velazquez Bastida	A01368039
Arlin San Juan Pezat	A01368226
Yessica Vidal Castellanos	A01366760
Alejandro Gabriel Hernandez	A01367757
Juan Carlos Robles Guicho	A01368398

Fecha de entrega: *Lunes 18 de octubre del 2021*

Semestre Agosto- Diciembre 2021

1. Etapa 3- Preparación de los datos

1.1 Limpieza de datos.

Limpiar los datos es uno de los pasos más relevantes para comenzar con un proyecto de ciencia de datos, se debe efectuar con el propósito de crear una cultura en torno a la toma de decisiones de datos de calidad. En este primer reporte describiremos los pasos que nuestro equipo siguió para realizar el proceso de limpieza de datos.

1.1.1 Comprensión de los datos.

Antes de comenzar a depurar nuestro conjunto de datos, un paso importante es comprender con qué variables estamos trabajando, esto nos ayudará a realizar una mejor limpieza de los mismos. En este caso nuestros datos son el cotejo del número de ventas de celulares de la marca Hisense con la información específica de cada variable.

Una vez comprendidos los datos que se nos brindaron, comenzamos a utilizar R studio para su limpieza. Para ello empleamos algunos comandos y librerías, como requerimientos iniciales, dichos comandos fueron los siguientes: **library(tverse)**: Llamar a la librería tidyverse / **getwd()**: Obtener la dirección donde estamos trabajando en nuestra PC. / **datosE3<- read.csv()**:Leer el archivo CSV, con el conjunto de datos.

1.1.2 Análisis de datos.

Con el fin de profundizar en la comprensión de nuestro datos, se utilizaron tres comandos los cuales nos ayudan a visualizar las dimensiones, y un sumario de cada variable. Para lo anterior los comandos utilizados fueron los siguientes: **dim(datosE3)**: Muestra las dimensiones de nuestra base de datos (Renglones y columnas) / **str(datosE3**): Nos brinda una lista de los tipos de datos con los que estamos trabajando. / **summary(datosE3)**: Nos brinda un resumen de cada variable, valores mínimos, máximos, medias, etc.

1.1.3 Detección de problemas de calidad.

Después de hacer un primer análisis de nuestros datos y de leerlos en la plataforma de R studio, el paso siguiente es detectar problemas de calidad. Para ello utilizamos las métricas universales del data quality, las cuales buscan eliminar problemas de: unicidad, validez, precisión, entre otros. Tomando esto en consideración, los problemas de calidad detectados fueron los siguientes:

- #1. En la variable punto de venta hay 5 puntos de venta escritos de manera errónea, este es un error de calidad de tipo coherencia.
- #2. En la variable mes hay valores mal registrados (en lugar de número, son letras). Cambiar los 5 meses que están registrados con letras, es un problema de tipo coherencia.
- #3. La variable de año no sigue un formato de valor numérico de 4 dígitos, es un error de tipo coherencia.

Entre otro errores que se describieron en el entregable.

1.1.4 Corrección de errores.

Debido a que cada error presentaba requerimientos diferentes, fue necesario emplear herramientas diferentes en cada problema. Algunos de los comandos utilizados para limpiar nuestro datos fueron: **datosE3[_,_]<- "_____"**: Se empleó para reemplazar algún dato de una columna y renglon especificos. / **datosE3$___<- tolower(datosE3$____)**: Se empleó para convertir en minúsculas los nombres de variables que se encontraban en mayúsculas. / **datosE3$mes<- str_replace(datosE3$__, "__", "_")**: Se utilizó para reemplazar todos los datos de un valor o carácter específico de una columna. /**datosE3 %>% select(___) %>% unique()**: Con este comando observamos los valores unicos de cada variable.

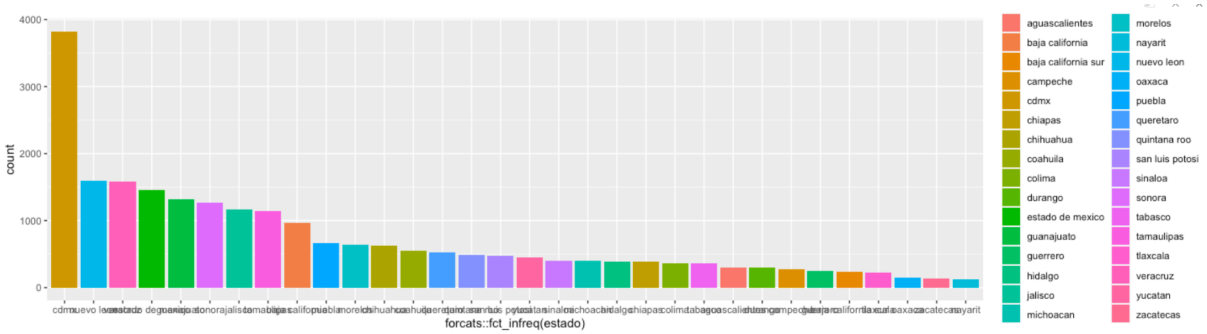
1.1.5 Exportar nueva base de datos.

Una vez que realizamos todas las correcciones de nuestra base de datos, el paso siguiente era escribirlos en un nuevo archivo CSV, con el fin de tenerlos actualizados con respecto a la base de datos inicial. Para este último paso se empleó el comando: **write.csv(datosE3, file= "datoslimpioshisense.csv")**: Con este comando exportamos nuestra nueva y actualizada base de datos.

1.2 Análisis exploratorio.

Posterior a la limpieza de los datos, el análisis de los datos para observar qué es lo que los datos representan es indispensable, para ello se recurrió al uso de herramientas gráficas, para que podamos visualizar la relación entre ellos y al mismo tiempo confirmar que no existan inconsistencias. Se plantearon las siguientes preguntas y se usaron los códigos a continuación:

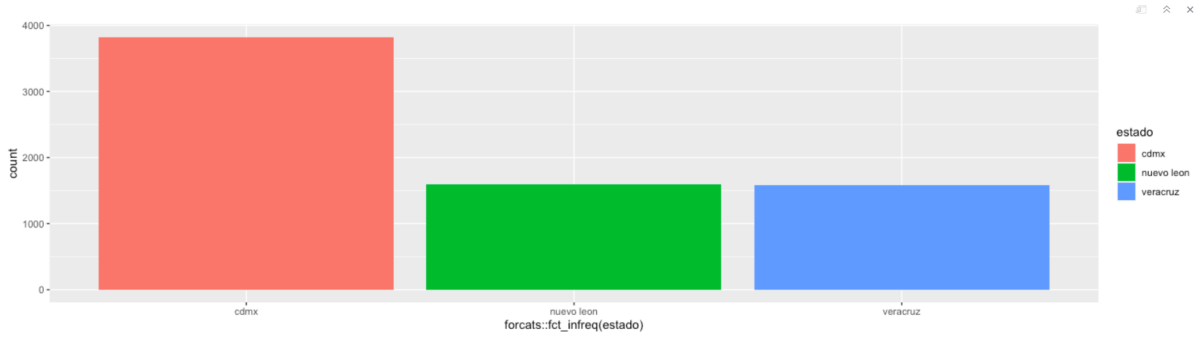
1) ¿Cómo están las ventas distribuidas por los estados de la república?
   Código: ggplot(**datosE3**, aes(x=forcats::fct_infreq(**estado**),fill = **estado**)) +
   geom_bar()



   Se observa que se tienen ventas en las 32 entidades del país, sin embargo estás son en diferente cantidad para cada una.

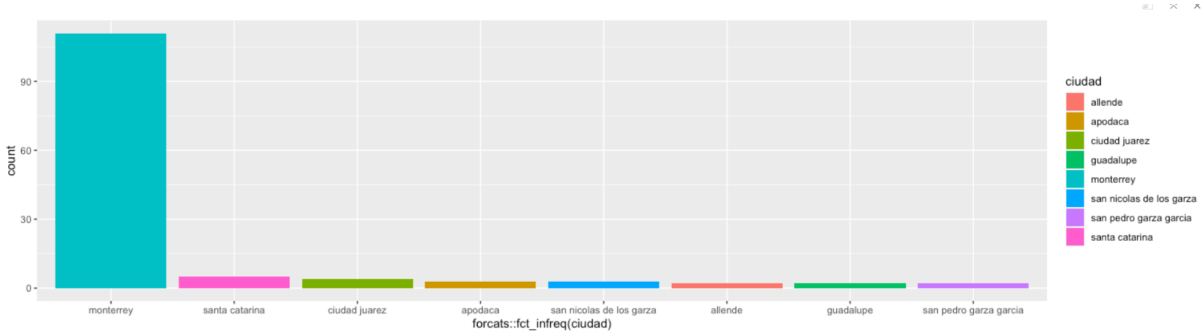
2) ¿Cuáles son los estados con mayores ventas a 1500?
   Código: **filter1** <- **datosE3 ni** %>%group_by(**estado**)%>% summarise(ventas=sum (**num_ventas**))

graficaventas <- **filter1** %>% arrange(**-ventas**)%>%filter(**ventas** >=1500)
graficaventas
filtro2 <- datosE3%>%filter(estado %in% c(**"veracruz", "cdmx", "nuevo leon"**))
ggplot(**filtro2**, aes(x = forcats::fct_infreq(**estado**),fill= **estado**))+geom_bar()



Los estados con mayores ventas son: Cdmx, Veracruz y Nuevo León considerando ventas arriba de 1500 u.

3) Divide las ciudades del estado que más ventas tiene.
caso1 <- **datosE3**%>%filter(estado=="**nuevo leon**")%>%filter(**anio**=="2020")
ggplot(**caso1**, aes(x=forcats::fct_infreq(**ciudad**),fill = **ciudad**)) + geom_bar()



Las ciudades con mayores ventas se enlistan a continuación, 8 ciudades pertenecientes al estado de Nuevo León, segundo estado con mayores ventas.

1.3 Ingeniería de características.

La ingeniería de características es el proceso por el cual a partir de nuestro conjunto de datos limpios, podemos generar nuevas variables que impactan o nos ayuden a predecir de una mejor manera lo que buscamos en nuestro proyecto.

1.3.1 Leer los datos.

Como paso inicial para esta etapa de ingeniería de características, volvemos a leer nuestros datos en R studio, para ello utilizamos el comando **datos <- read.csv("___")**, para corroborar sus dimensiones y un pequeño resumen de los mismos, utilizamos los comandos **dim(datos), str(datos) y summary(datos).**

Como paso adicional en esta fase inicial, asignamos algunas de nuestras variables al tipo de variable específica, que haga más fácil su manejo; ejemplo de ello es convertir las variables de carácter a numéricas si es necesario. Para esta etapa el comando empleado fue el siguiente:

- datos$VARIABLE <- as.NUEVA VARIABLE(datos$VARIABLE)

1.3.2 Índices para variables consideradas.

Para un mejor manejo, convertimos nuestras variables cualitativas en índices. Este proceso se realizó ordenando por orden alfabético algunas variables para después, asignar una columna extra de ID con una secuencia de número enteros, todo esto para las variables consideradas. Los comandos para este proceso fueron:

- pdv_id <- datos%>%select(**VARIABLE**)%>%unique()%>%arrange() head(pdv_id)
- pdv_id$pdv_id <- as.character(seq.int(nrow(**VARIABLE**_id))) head(pdv_id)

Estos comandos se ejecutaron para las variables cualitativas que deseamos transformar en índices. Las variables consideradas fueron: Punto de Venta, Mes y SKU.

1.3.3 Agrupar conjunto de datos.

Como segundo paso, se realizará la agrupación de datos de nuestro nuevo data frame con nuestros datos originales basándonos en la columna de ID. Este proceso se realizó con la función left join, para todas las variables consideradas en los puntos anteriores. Todo lo anterior se describe en el siguiente código: **datos <- left_join(datos, VARIABLE_id, by="VARIABLE") head(datos)**

1.3.4 Agrupas ventas totales.

Con el fin de observar las ventas adicionales que hay en los puntos de venta, en la misma fecha, se realizó la agrupación de ventas totales. Para realizar esto se tomó en cuenta la sugerencia de quitar la información adicional implícita en el punto de venta. Lo anterior se realizó con el comando: **datos <- datos %>% #quitamos fecha porque vamos a hacer el análisis por mes group_by(pdv_id, sku_id, mes_id)%>% summarise(ventas_totales = sum(num_ventas))**

1.3.5 Completar serie de tiempo.

Con el paso anterior podemos observar que nuestra serie de tiempo está incompleta, pues hay periodos en donde no hay ventas registradas. Para resolver esto se construyeron 3 conjuntos nuevos de índices con las variables designadas. Posteriormente se realizó la combinación de estos conjuntos de datos empleando la función merge ().

Una vez realizada la combinación, se empleó nuevamente la función left join () para obtener las ventas totales en cada punto de venta. Con esta combinación se puede ver los meses en donde no hubo ventas de productos, para los cuales se colocó un índice 0.

1.3.6 Construcción de variable de respuesta (Y).

Para nuestro modelo de pronósticos de ventas, y considerando la información histórica que tenemos, se decidió emplear un código, el cual desplaza la información una columna creando una nueva variable (Y), esto nos permitirá saber las ventas para el siguiente mes.

1.3.7 Crear nuevas características.

Este paso nos permite crear nuevas variables que agreguen valor a nuestro análisis. Para nuestro proyecto se realizaron agrupaciones y conteos que permitan realizar las predicciones de una mejor manera. En esta etapa creamos las características de ventas promedio por mes, tienda y producto y ventas totales con las cuáles se crean las características que necesitamos de manera rezagada más adelante para nuestro modelo de predicción.

Realizado el paso anterior, se agruparon con nuestra base de datos inicial, en complemento con las fases anteriores.

1.3.8 Rezagos.

Como último paso con los siguientes comandos: **library(zoo) datos_completos<-na.locf(datos_completos, fromLast = TRUE) head(datos_completos)** se realizaron los valores faltantes NA con 0, con el fin de tener una base de datos completa. Para finalizar este paso, se escribió nuestra nueva base de datos en un nuevo archivo de CSV, con el siguiente código: **write.csv(datos_completos, file="datos_completosE3.csv", row.names = FALSE)**

1.3.9 Conclusiones de ingeniería de características.

Después de realizar el proceso con ingeniería de características, podemos observar que el cambio de dimensiones de nuestro nuevo archivo son: 386316 renglones con 24 columnas. En cuanto al archivo trabajado anteriormente antes del proceso de ingeniería de características donde las dimensiones eran de 23032 renglones con 14 columnas.

2. Referencias Bibliográficas:

[1] (2020, noviembre 13). Data Cleansing. Todo lo que debes saber sobre la 'limpieza de datos'. Se recuperó el octubre 12, 2021 de https://bigdatamagazine.es/data-cleansing-todo-lo-que-debes-saber-sobre-la-limpieza-de-datos

[2] (n.d.). Calidad de Datos. Cómo impulsar tu negocio con los datos.. Se recuperó el octubre 14, 2021 de https://www.powerdata.es/calidad-de-datos

Tecnológico de Monterrey

Campus Toluca

Segundo Avance del Proyecto
Reporte 2

Materia: IN3038.1

Laboratorio de Diseño y Optimización de Operaciones

Profesora: M. en C. Ana Luisa Masetto Herrera

Equipo número 3:

Alejandra Velazquez Bastida | A01368039
Arlin San Juan Pezat | A01368226
Yessica Vidal Castellanos | A01366760
Alejandro Gabriel Hernandez | A01367757
Juan Carlos Robles Guicho | A01368398

Fecha de entrega: *Lunes 18 de octubre del 2021*

Semestre Agosto- Diciembre 2021

1. Etapa 4- Modelado.

1.1 Promedios móviles.

El promedio móvil es un indicador de tendencias que se usan para realizar análisis de datos anteriores con la finalidad de formar una serie de medidas que provengan de diversos subconjuntos de datos de precios, por lo tanto, tienen la capacidad de examinar las medidas de precios que disminuyen en un período de tiempo.

Los promedios móviles se deben calcular después de las observaciones consecutivas de los subgrupos artificiales. Estos se pueden utilizar en las gráficas de control para crear gráficas de promedios para los datos en determinados tiempos programados. Cuando se realizan los análisis de series de tiempo, se usa el promedio móvil y de esa forma se pueden suavizar los datos y disminuir las dudas aleatorias en una determinada serie de tiempo.

1.1.1 Tipos de promedios móviles.

1.1.1.1 Promedio móvil simple.

El modelo de promedio móvil funciona mejor con datos horizontales (datos sin tendencia). Un promedio móvil se obtiene encontrando la media de un conjunto específico de valores y aplicándolo después para pronosticar el siguiente periodo.

1.1.1.2 Promedio móvil ponderado.

En general se difiere que los diversos puntos de datos se pueden ponderar o asignar a un punto concreto de gran importancia. La media móvil ponderada tiene la capacidad de agregarle importancia a los puntos de datos que estén más recientes. Dentro del período estipulado, a cada uno de los puntos se le asignará un multiplicador de datos reciente para que luego vaya descendiendo ordenadamente. Después cuando se le añade al principio un nuevo punto, se eliminará el punto de datos que contenga mayor antigüedad.

1.1.2 Métricas: MAE,RMSE,MSE

Las métricas nos ayudan a medir el desempeño del modelo, además de ayudarnos a contrastar el Modelo actual vs el Modelo propuesto. Mediante las métricas podemos obtener resultados interpretables y medibles.

- MAE: Error absoluto medio
- MSE: Error cuadrado medio
- RMSE: Error medio

2.1 Construcción del modelo de promedios móviles

A continuación se presenta de manera breve la construcción del modelo de promedios móviles utilizado para la base de datos de la compañía Hisense, empresa bajo estudio durante este proyecto.

2.1.1 Lectura de datos

El código del modelo se realizó en Jupyter, descargando las librerías correspondientes y cargando el archivo con nuestros datos.

2.1.2 Descartar columnas de datos

A través del comando *datos.drop (),* se descartaron algunas columnas de nuestros datos, dejando la columna ***ventas_totales*** y ***y_ventas_siguiente_mes*** como nuestras variables x y y respectivamente para realizar el modelo de promedios móviles.

```
In [10]: datos_E3.head(10)
```

Out[10]:

	pdv_id	mes_id	sku_id	ventas_totales	y_ventas_siguiente_mes
0	1	0	1	0	1
1	1	1	1	1	1
2	1	2	1	1	1
3	1	3	1	1	0
4	1	4	1	0	0
5	1	5	1	0	0
6	1	6	1	0	0
7	1	7	1	0	0
8	1	8	1	0	0
9	1	0	2	0	0

Fig 1. Datos con columnas descartadas

2.1.3 Pedir datos

Se procede a construir una columna que muestre lo vendido el mes anterior como base para la predicción de ventas posterior. Se calcula otra columna que muestre el promedio de ventas de los dos meses anteriores. De esta forma cambiamos dos veces el periodo móvil.

	pdv_id	mes_id	sku_id	ventas_totales	y_ventas_siguiente_mes	m1_pedir_lo_del_mes_pasado	m2_promedio_de_dos_meses_anteriores
0	1	0	1	0	1	0	NaN
1	1	1	1	1	1	1	0.5
2	1	2	1	1	1	1	1.0
3	1	3	1	1	0	1	1.0
4	1	4	1	0	0	0	0.5
5	1	5	1	0	0	0	0.0
6	1	6	1	0	0	0	0.0
7	1	7	1	0	0	0	0.0
8	1	8	1	0	0	0	0.0
9	1	0	2	0	0	0	NaN
10	1	1	2	0	2	0	0.0
11	1	2	2	2	0	2	1.0
12	1	3	2	0	0	0	1.0
13	1	4	2	0	0	0	0.0
14	1	5	2	0	0	0	0.0
15	1	6	2	0	0	0	0.0

Fig 2. Datos con 2 cambios de periodo móvil

2.1.4 Cálculo de MAE

Se realiza el cálculo del indicador MAE respecto al error del promedio calculado con los datos reales de las ventas por mes. Se realiza el cálculo manual mostrando el nombre del mes y el MAE tanto para el promedio de 2 meses anteriores como para 3. Finalmente, se muestra de manera gráfica el error absoluto total en el modelo.

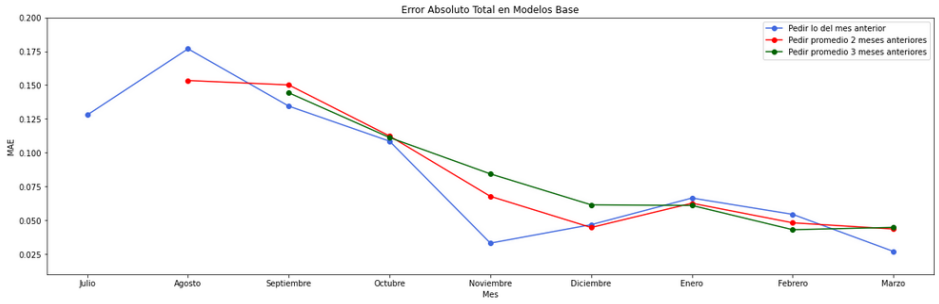


Fig 3. Gráfica del Error Absoluto Total en el modelo

2. Referencias Bibliográficas:

[1] Minitab.(2019). ¿Qué es un promedio móvil?Disponible en: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/time-series/supporting-topics/moving-average/what-is-a-moving-average/#:~:text=Los%20promedios%20m%C3%B3viles%20son%20promedios%20calculados%20a%20partir,suavizar%20los%20datos%20y%20reducir%20las%20fluctuaciones%20

[2] Pacheco, J. (2021). *¿Qué es el Promedio Móvil?*. Web y Empresas. Disponible en: https://www.webyempresas.com/promedio-movil/

[3] Avilés, E.G. (2021), *Ingeniería Estadística*, [presentación], Disponible en: Mis clases: Ingeniería estadística (Gpo 1) (tec.mx)