

**Laboratorio de Diseño y Optimización de Operaciones**

**Modelos tradicionales vs. Modelos de Aprendizaje de Máquina para predicción de demanda de productos de telefonía celular de la empresa ZTE**

**Entregable 2**

“LIMPIEZA DE DATOS”

**Integrantes:**

Carlos Raúl Muciño Angeles A01104864

César Iván Trejo Cerón A01368390

Nancy Escamilla Sánchez A01366872

**Docente:**

Ana Luisa Masetto.

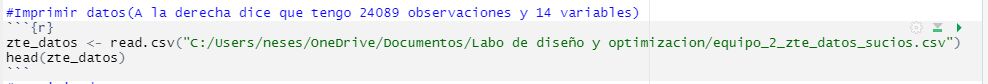
**Introducción**

Hoy en día el uso de los teléfonos celulares se ha vuelto parte de la vida cotidiana de todas las personas, por ello el giro de estas grandes empresas es uno de los más importantes en todo el mundo, estas empresas deben ser lo suficiente capaces para innovar productos de alta calidad, además que estos cumplan con gustos, necesidades e intereses diferentes de los consumidores y tener estos disponibles en tiempo y forma para los clientes.

Es por ello que en este proyecto nos enfocaremos en la predicción de la demanda de la empresa ZTE, la cual resulta ser importante para el mercado debido a que ayudará a comprender a gran escala el negocio y adecuar ciertas decisiones para ser una empresa eficiente. Para lograr ello se utilizará la herramienta RStudio que nos permitirá realizar el procesamiento de una base de datos la cual presenta información importante de la compañía ZTE, esto con el objetivo de comprender la predicción de las demandas de ZTE.

**Desarrollo**

Para poder leer correctamente los datos del archivo CSV utilizamos el siguiente comando:



Para analizar este archivo utilizaremos la función de read.csv, la cual nos permitirá ver la dimensión de los datos que tenemos; con ello descubrimos que nuestra base de datos cuenta con 24089 entradas divididas en 14 variables, donde se detallan las ventas que se tuvieron durante los años 2019 y 2020 de equipos en diferentes lugares del país. El significado de las variables se detalla en el siguiente diccionario:

**punto\_de\_venta**: Especifica la tienda en donde se vendió el dispositivo, detallado con diferentes acrónimos o simplemente el nombre de la plaza o centro en donde la tienda se encuentra.

En los puntos de venta encontramos que son en total 1463 observaciones y además esta variable debe ser factor porque es cualitativa.

**fecha**: Variable que detalla la fecha en la que se realizó la venta con el formato de día/mes/año.

En la fecha se encontró que hay 289 registros.

**mes**: Variable que detalla el mes en el que se realizó la venta.

Los meses deben de ser 12, de los cuales se deben de realizar las adecuadas correcciones porque al leer los datos nos indica que son 15, lo cual es incorrecto, además esta es una variable entera por lo que los datos deben coincidir.

**anio**: Variable que detalla el año en el que se realizó la venta.

En los años observamos que únicamente deben haber dos años en la base de datos 2019 y 2020. También se deben hacer las correcciones pertinentes y esta es una variable entera.

**num\_ventas**: Variable que detalla la cantidad de ventas registradas.

Se puede observar únicamente 1 observación, la cual debe ser también una variable entera debido a que es numérica.

**sku**: Variable que detalla el número que identifica el dispositivo vendido.

En esta variable se observan 29 datos distintos.

**marca**: Variable que detalla la marca del dispositivo vendido.

En la marca observamos distintas marcas, pero se debe corregir este dato debido a que la marca solamente debe de ser única, porque solamente estamos manejando la marca ZTE.

**gamma**: Variable que detalla la categoría a la que pertenece el dispositivo vendido.

Se observa solo una observación la cual resulta ser “baja”, además esta es una variable tipo character.

**costo\_promedio**: Variable que detalla el precio en el que se vendió el dispositivo.

Se observan 28 costos promedios diferentes en total en toda la base de datos.

**zona**: Variable que detalla la parte del país en donde se vendió el dispositivo.

Se observan 9 diferentes zonas en la base de datos de la cual 1 esta mal escrita, por lo que debe corregirse para únicamente tener 8 zonas correctas.

**estado**: Variable que detalla el territorio del país donde se vendió el dispositivo.

En total se observan 35 estados en total en toda la base de datos, de los cuales 3 están incorrectos porque en vez de ser un estado es una capital, los cuales deben ser corregidos.

**ciudad**: Variable que detalla la zona urbana en donde se vendió el dispositivo.

Se observan 208 ciudades en toda la base de datos.

**latitud**: Variable que detalla la distancia a la que se encuentra el punto de venta del ecuador.

En la variable latitud se observan 1437 tipos de latitudes en toda la base de datos, la cual debe ser una variable tipo numérica porque tiene puntos decimales.

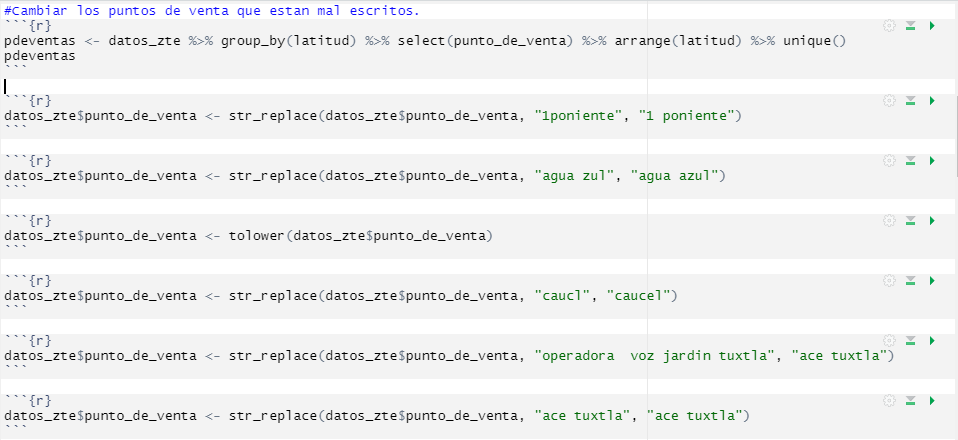
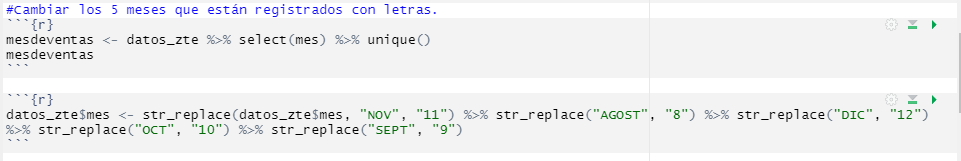
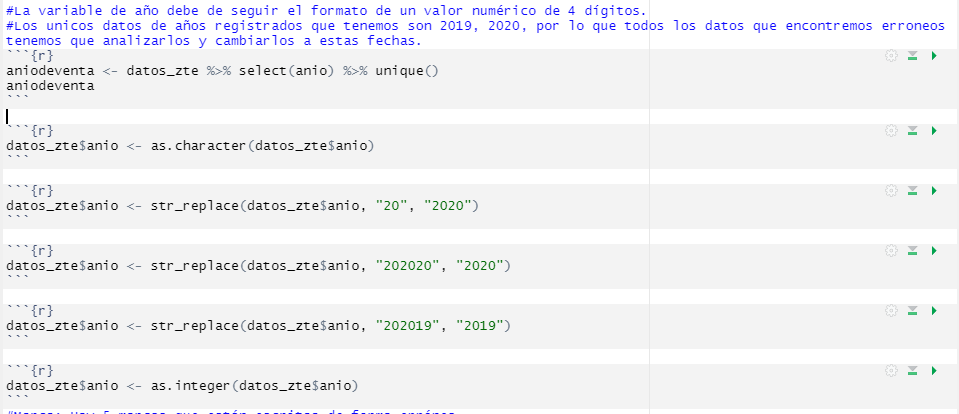
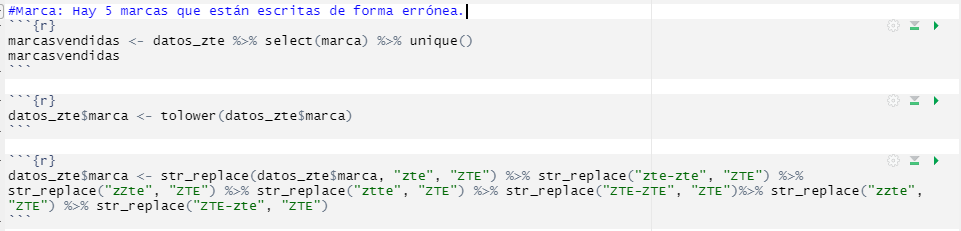
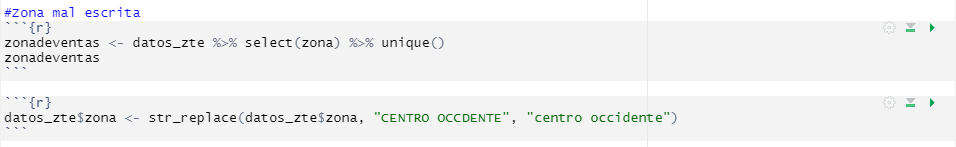
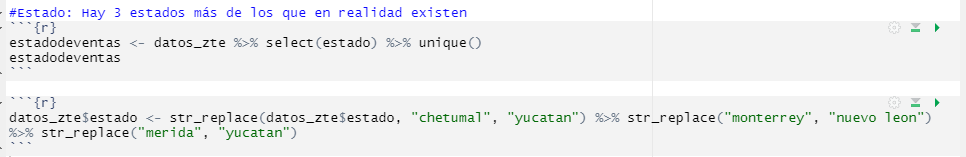
**longitud**: Variable que detalla la distancia angular sobre el ecuador donde se encuentra el punto de venta.

En la variable longitud se observan 1409 tipos de longitud en toda la base de datos, la cual debe ser una variable tipo numérica porque tiene puntos decimales.

**Problemas de calidad. (¿Qué?, ¿Dónde?, ¿Por qué?)**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Nombre del tipo de variable*** | ***¿Qué problema de calidad?*** | ***¿Dónde está el problema?*** | ***¿Por qué es un***  ***problema de calidad?*** |
| Punto de Venta | Hay 5 puntos de venta escritos de manera errónea. | Los problemas están en los siguientes puntos de venta: "1poniente"  "agua zul"  “bca PABELLON valle”  "caucl"  "operadora voz jardin tuxtla" | Los puntos de venta deben de estar escritos de forma correcta porque si no podría provocar errores en las interpretaciones. |
| Fecha | No hay problemas de calidad, los registros están limpios. |  |  |
| Mes | Mes: Esta variable es numérica. Hay valores mal registrados (en lugar de numero, son letras). | Los problemas en los meses son los siguientes:  "NOV"  "AGOST"  "DIC"  "OCT"  "SEPT" | Porque si tengo letras en vez de números no se puede definir la variable correcta, por lo que no se pueden procesar y utilizar los datos para el análisis. |
| Año | Año: La variable del año debe de seguir el formato de un valor numérico de 4 dígitos. | Los problemas están en el siguiente número:  "20" | Porque si tenemos fechas erróneas no podremos realizar pronósticos adecuados. |
| Número de ventas | No hay problemas de calidad, los registros están limpios. |  |  |
| SKU (Producto vendido) | No hay problemas de calidad, los registros están limpios. |  |  |
| Marca | Marca: Hay 5 marcas que están escritas de forma errónea | Los problemas se encuentran en las siguientes marcas:  “ztte”  “ZTE-ZTE”  “zte-zte”  “zZte”  “zte” | Solo debe existir una marca para todos porque solamente estamos manejando la marca ZTE. |
| Gamma (Nivel del producto vendido) | No hay problemas de calidad, los registros están limpios. |  |  |
| Costo promedio | No hay problemas de calidad, los registros están limpios. |  |  |
| Zona | Zona: Hay 1 zona que está mal escrita | La zona que está mal escrita es la siguiente:  "CENTRO OCCDENTE" | Debemos corregir la zona porque si no se corrige se estaría tomando como otra zona y eso sería incorrecto porque es un dato mal escrito que pertenece a una zona existente el cual es el "centro occidente". |
| Estado | Estado: Hay 3 estados más de los que en realidad existen. | Los estados que no existen son los siguientes:  "chetumal"  "monterrey"  "merida" | Estos deben ser reemplazados por los estados correctos de la República Mexicana debido a que las que tenemos son capitales. |
| Ciudad | No hay problemas de calidad, los registros están limpios. |  |  |
| Latitud | Latitud: Hay 1 valor fuera de rango. | El valor fuera de rango es el siguiente:  "1903991" | El valor de la latitud es incorrecta, debido a que las latitudes tienen un punto decimal después de los primeros dos dígitos, por lo que este dato de latitud si se procesa de esa manera podría afectar demasiado en el procesamiento de datos. |
| Longitud | Longitud: Hay 1 valor fuera de rango. | El valor fuera de rango es el siguiente:  "-9824871" | El valor de longitud es incorrecto, debido a que las longitudes tienen un punto decimal después de los dos dígitos, se debe de hacer la corrección porque ese dato afectará demasiado si se realizan análisis con datos erróneos. |

**¿Cómo se corrigieron los problemas de calidad?**

* **Punto de Venta**
* **Mes**
* **Año**
* **Marca **
* **Zona**
* **Estado **
* **Latitud **
* **Longitud**

**Conclusión**

Para poder tomar decisiones que tengan un impacto positivo y considerable en un proyecto es necesario tener información clara y correcta, para eso los datos que manejamos deben ser los adecuados. Uno no puede hacer un gran platillo sin tener ingredientes de calidad.

Es por eso que la limpieza de datos puede ser el paso más importante en el desarrollo de un proyecto de ciencia de datos. Ubicar los errores en los registros, determinar las causas que pudieron provocar esos errores y poder establecer la manera en que se deben de corregir son actividades prioritarias al inicio de nuestro análisis.

En este proyecto, a pesar de que la cantidad de datos erróneos parecía ser poca, la tarea fue laboriosa y tomó un tiempo mayor al considerado. Reconocer las fallas fue el paso más sencillo pues en muchas variables solo se ocupaba un comando, en cambio, el corregirlas fue más complicado pues se tuvo que establecer la manera correcta los datos que tendría que estar el registro y posteriormente utilizar varios comandos (detallados en el documento) para llegar al resultado deseado.