ETL, Estatística, Banco de Dados e Serie temporal

Oscar J. O. Ayala

Introducción

El conjunto de datos analizado en este estudio fue puesto a disposición por la empresa *Millenium s.a.*, que se refiere a 5 (cinco) hojas de cálculo de Excel no relacionadas, etiquetadas como **BBDD**, **Param_Horas**, **Param_Fechas**, **Param_Colas** y **Param_Canales**. El objetivo es descubrir conocimientos a partir de los datos que permitan obtener indicios sobre la situación de una determinada organización. Se utiliza la metodología de *Data Mining* para realizar las validaciones de calidad y los análisis estadísticos pertinentes, junto con el enfoque de base de datos relacional y ETL (Extracción, Transformación e Carga). Los soportes computacionales utilizados para el correcto análisis, creación de **data base** y presentación de los resultados fueron R, SQLite (por ser de código abierto), Excel, Markdown y Power BI, así mismo todos los códigos em lenguaje SQL y R se encuentran en el script *02_scriptAnalysis_Oscar.R*.

Validación de calidad de datos

La base de datos **BBDD** presenta problemas que pueden afectar el análisis. La Tabla 1, muestra un resumen de los campos que requieren mayor atención. En general, Valores faltantes, posibles campos redundantes y formatos de variables incorrectos son los que más de destacan. Así, se realiza un preprocesamiento para obtener datos de calidad y luego se comentan los problemas encontrados junto a solución adoptada:

Tabla 1: Resumen estadístico relevante.

	canal	tiempo de respuesta	tiempo de conversación
Mínimo		0,0	0
cuartil 1		1,0	373
Mediana		6,0	877
Média		192,7	3177
Cuartil 3		78,0	3920
Máximo		54380,0	79230
Valores faltantes	28588	577	577

- El campo *canal* es una columna de valores faltantes, por lo que se elimina.
- Los registros 26304 y 27406 son líneas de valores faltantes, por lo que se excluyen.

- Los campos *tiempo_respuesta* y *tiempo_conversacion* tienen un total de 577 valores faltantes en los mismos registros, incluidos los registros 26304 y 27406. Sin embargo, eliminarlos supondría una pérdida significativa de datos, por lo que se decide mantenerlos. La imputación no parece adecuada dado el número de valores ausentes.
- La columna *tiempo de abandono* tiene 25120 entradas desconocidas, que por su gran número se mantienen.
- El campo *campana* tiene un solo nivel, es conveniente excluirlo.
- La columna *fecha* está en formato general, por lo que se convierte al formato día/mes/año.

Así mismo, el análisis de valores atípicos es de interés. En las Figuras 1 - 3 se crean *Boxplots* por grupos de variables numéricas según el orden de grandeza similar, para tener una mejor visualización de los datos.

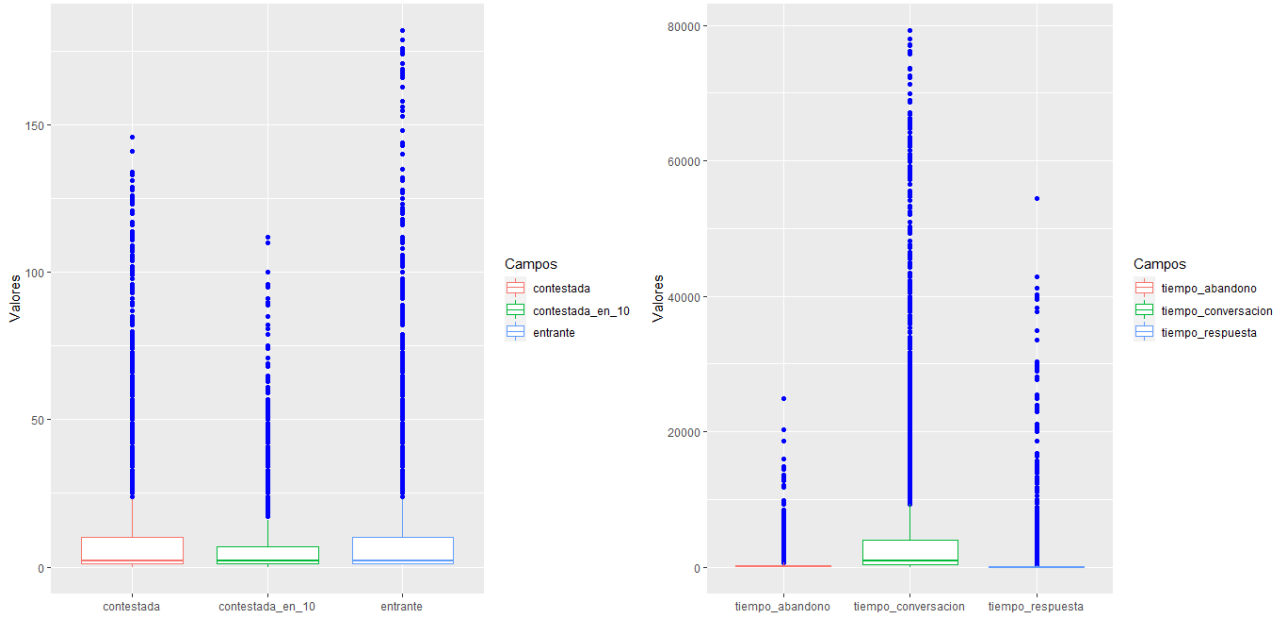

Figura 1: Box plot de variáveis numéricas. Fonte: *elaboración propia*.

Los Boxplot's presentan outlier (puntos azules), sin embargo, no son valores inconsistentes y dado su número, se decide dejarlos. En estas representaciones, se observa que, en general, la mayoría de las llamadas son atendidas en 10 unidades de tiempo y contienen los valores atípicos más grandes dentro deste grupo (Figura 1 - 2). El número de clientes que abandonan el servicio parece hacerlo mayormente en 5 unidades de tiempo y de manera similar a las 10, 15 y 20 unidades de tiempo (Figura 2). Además, se observa que hay indicios de clientes que han esperado una respuesta durante mucho tiempo (Figura 1).

También se sospecha que, dada esta cantidad de valores extremos, la distribución de las variables puede presentar asimetría positiva en los mismos. Esta forma es muy común en series de conteo. Sin embargo, para no hacer demasiado extenso este informe, se ponen a disposición los histogramas como anexo de imagenes.

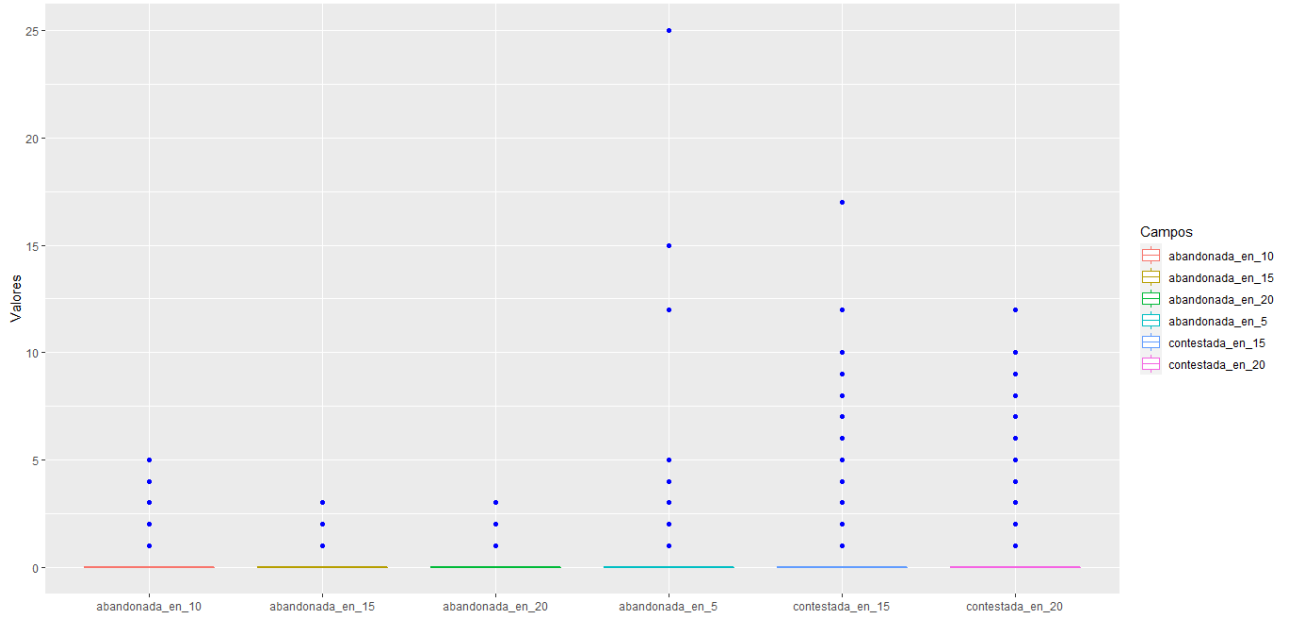

Figura 2: Box plot de variáveis numéricas. Fonte: *Elaboración propia*.

Normalización de tablas

A través de SQLite y Rstudio fue posible crear una base de datos relacional, adjuntando su archivo con el nombre *03_dbseletivo.sqlite3*. La relación entre tablas fue dada con las siguientes consideraciones:

- Tabla *Colas* Primary key `Id_Cola` -> Tabla *BBDD* Foreign key`Id_Cola`.
- Tabla *Canales* Primary key `Id_Canal` -> Tabla *Colas* Foreign key`Id_Canal`.
- Tabla *Fechas* Primary key `Fecha` -> Tabla *BBDD* Foreign key`fecha`.
- Tabla *Horas* Primary key `Hora_Num` -> Tabla *BBDD* Foreign key `Tiempo_conversación`.

En la Figura 3, se muestra el modelo relacional de esta base de datos, el cual fue realizado utilizando Power BI (Ver Anexo).

Porcentaje de llamadas entrantes contestadas en 15 segundos

La Figura 4, muestra la proporción de llamadas recibidas contestadas. Curiosamente, alrededor del 95,8 por ciento se contestan en 10 segundos, mientras que el porcentaje de llamadas contestadas en 15 y 20 segundos se acerca al 2 por ciento. Sin embargo, a partir de la Figura 1, se sospecha que esto no implica que se den respuestas rápidas.

Comportamiento de las interacciones entrantes

En la Figura 5, se puede observar que las interacciones *entrantes* tienen un marcado componente estacional semanal con picos de martes a viernes. Alcanza su punto más alto la última semana de marzo

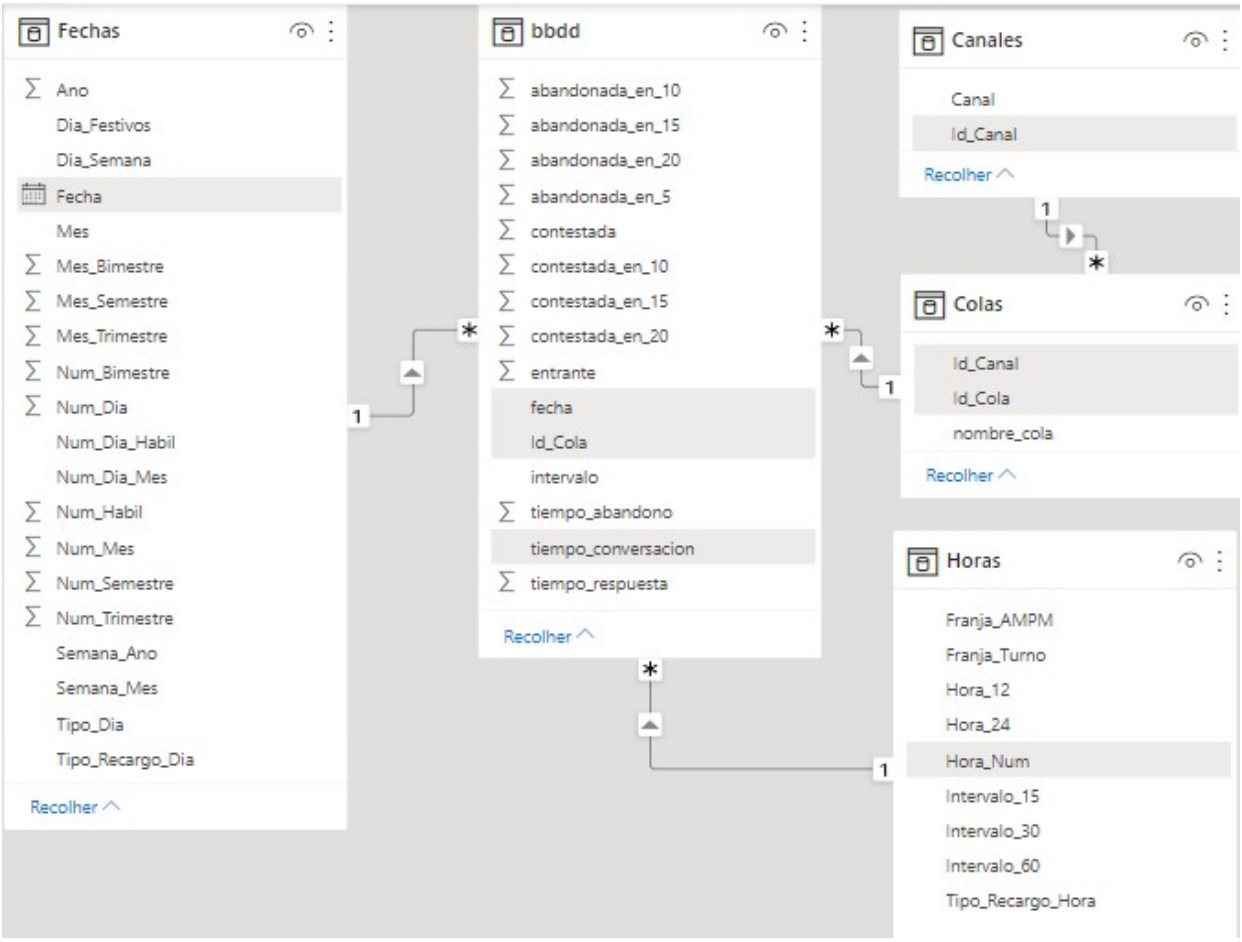

Figura 3: Modelo de tablas normalizadas. *Fuente: elaboración propia*

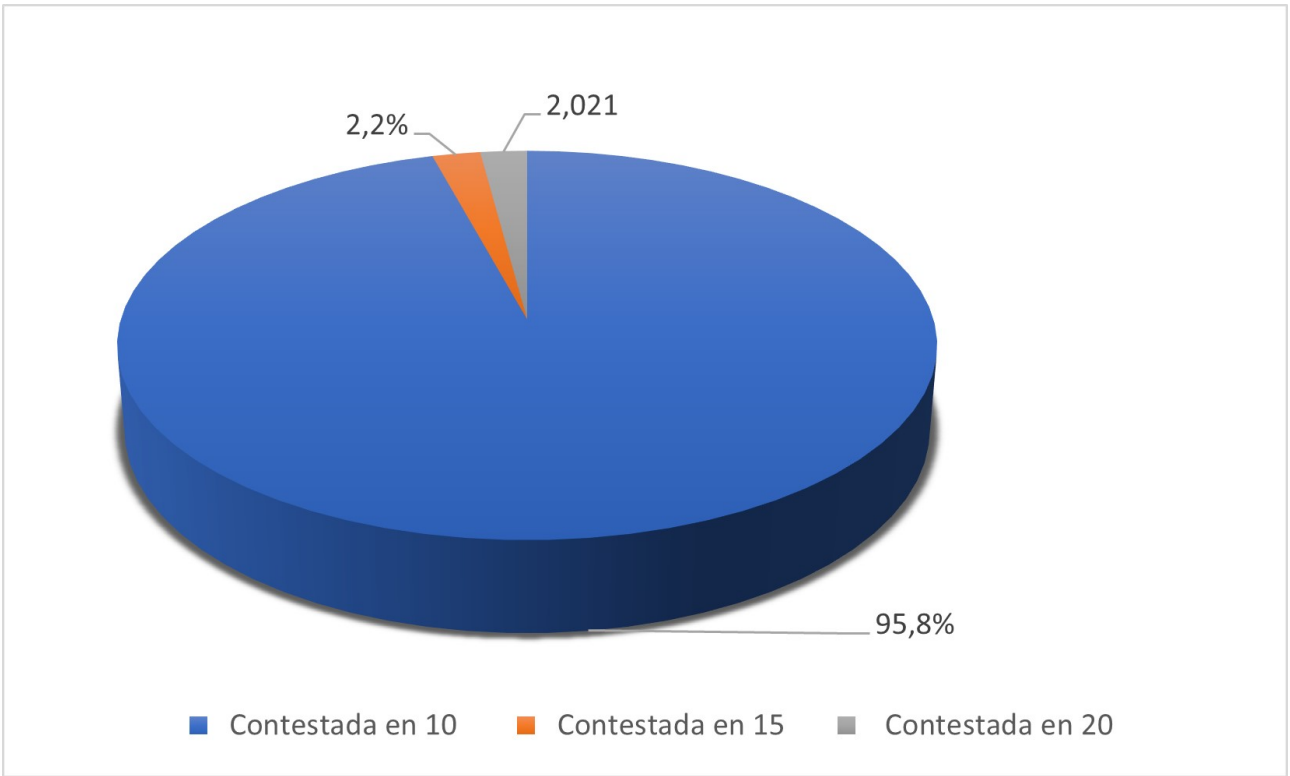

Figura 4: Porcentaje de participación de llamadas entrandes contesta. *Fuente: elaboración propia*

que corresponde al área sombreada de la Figura 5, específicamente el día 31. Se puede proponer un modelo de pronóstico que admita estos cambios que ocurren con el calendario para representar el comportamiento estacional. También parece haber un componente de tendencia, aunque menos marcante, ascendente hasta el último día de marzo y luego cae ligeramente hasta estabilizarse.

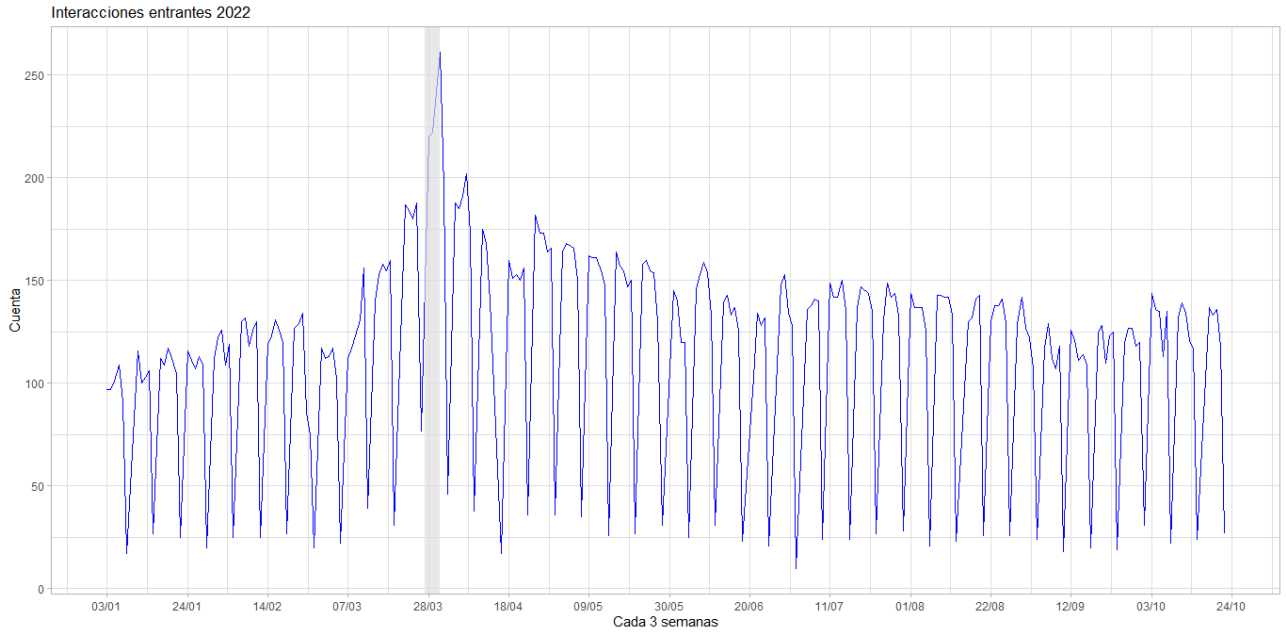

Figura 5: Serie temporal interacciones entrantes *Fuente: elaboración propia*