**Proyecto final:** Regresión Logística

1. Identificar las variables en español y su tipología (continua, dummy, etc.). Explicar de qué se trata la base de datos.

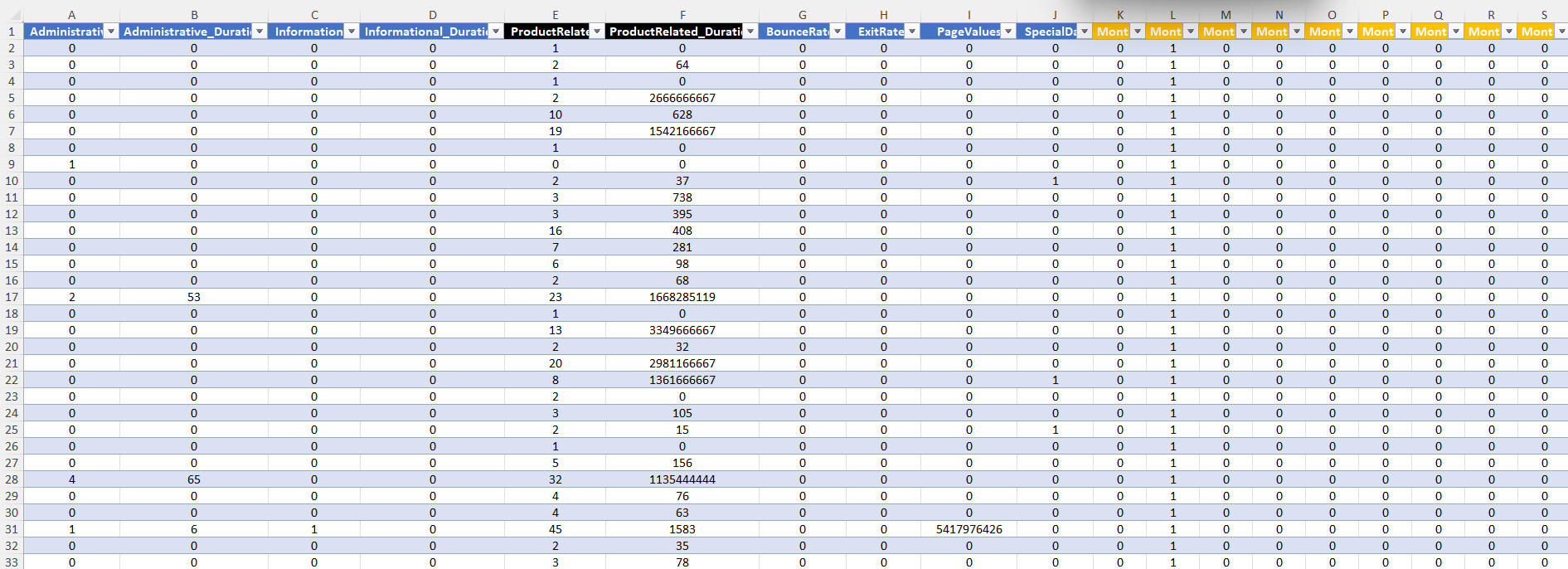
|  |  |  |  |
| --- | --- | --- | --- |
| VARIABLE | DEFINICION | TIPO | EXPLICACION |
| Weekend | Fin de semana | DUMMY | Valor booleano que indica si la fecha de la visita es un fin de semana. |
| Revenue | Transacción | DUMMY | Etiqueta de clase que indica si la visita ha sido finalizada con una transacción. |
| Visitortype | Tipo visitante | Escala Likert | Tipo de visitante como 'Nuevo visitante', 'Visitante recurrente' y 'Otro' |
| TrafficType | Tipo de trafico | Escala Likert | Fuente de tráfico por la cual el visitante ha llegado al sitio web (por ejemplo, banner, SMS, directo). |
| Region | Región | Escala Likert | Región geográfica desde la cual el visitante ha iniciado la sesión. |
| Browser | Navegador | Escala Likert | Navegador del visitante. |
| OperatingSystems | Sistema Operativo | Escala Likert | Sistema operativo del visitante. |
| Month | Mes | Escala Likert | Valor del mes de la fecha de visita. |
| Administrative | # Paginas | Numéricas/Discretas | Número de páginas visitadas por el visitante sobre gestión de cuentas. |
| Administrative\_Duration | Duración en paginas | Numéricas/Continuas | La cantidad total de tiempo (en segundos) que el visitante ha pasado en páginas relacionadas con la gestión de cuentas. |
| Informational | Numero de páginas visitadas | Numéricas/Discretas | El número de páginas visitadas por el visitante en el sitio web, así como la información de comunicación y dirección del sitio de compras. |
| Informational\_Duration | Duración páginas visitadas | Numéricas/Continuas | La cantidad total de tiempo (en segundos) que el visitante ha pasado en páginas informativas. |
| ProductRelated | Producto relacionado | Numéricas/Discretas | Número de páginas visitadas por el visitante sobre páginas relacionadas con productos. |
| ProductRelated\_Duration | Duración con el producto | Numéricas/Continuas | La cantidad total de tiempo (en segundos) que el visitante ha pasado en páginas relacionadas con productos. |
| BounceRates | Tasa de rebote | Numéricas/Continuas | Valor promedio de la tasa de rebote de las páginas visitadas por el visitante. |
| ExitRates | Tasa de salida | Numéricas/Continuas | Valor promedio de la tasa de salida de las páginas visitadas por el visitante. |
| PageValues | Valores de página | Numéricas/Continuas | Valor promedio de la página de las páginas visitadas por el visitante. |
| SpecialDay | Duración en día especial | Numéricas/Continuas | Cercanía del tiempo de visita al sitio a un día especial. |

EXPLICACION:

Según las variables esta base de datos parece ofrecer ofrecer información sobre el flujo de visitas e información relacionada de la misma a un sitio web de comercio electrónico para entender a mas profundidad cuales son los factores que puedes influir en la estadía de esta.

1. Limpiar y preparar la base de datos

Transformación de la variables tipo Likert para evitar el error de la escala y usarlo como variables independiente dummy para una mejor manipulación de estas:



Tabla, Excel

Descripción generada automáticamente

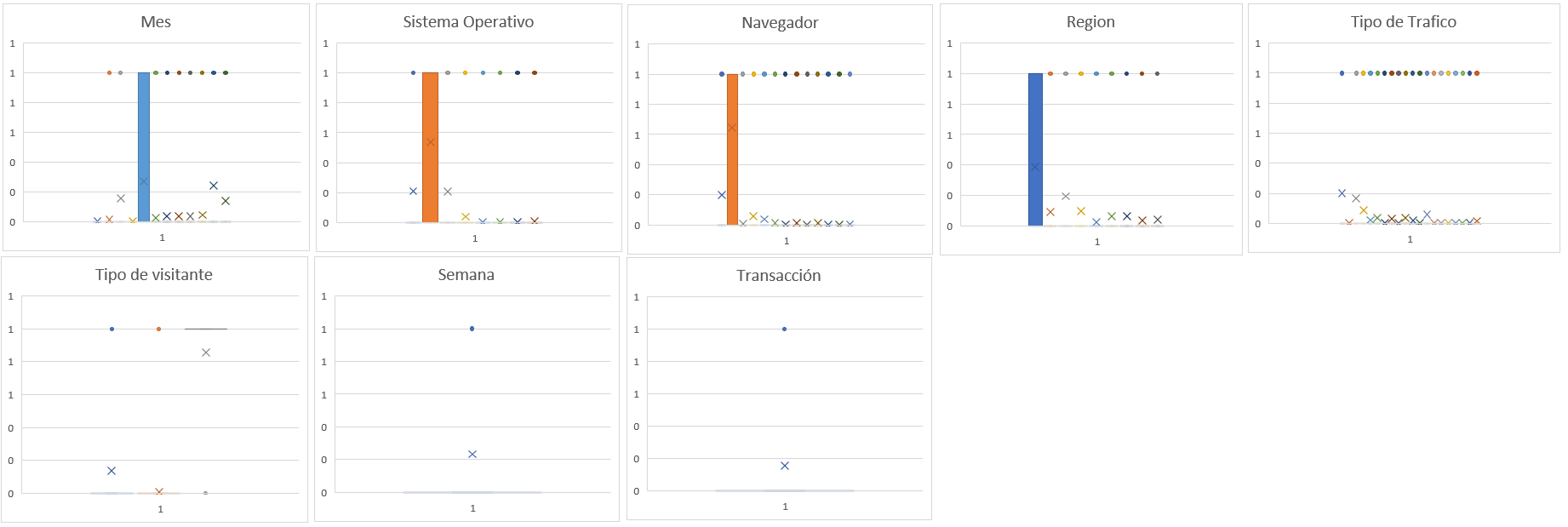
Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

1. Eliminar los “outliers” aplicando un EDA en Excel.

Se realiza un grafico de caja para revisar cuales son los Outliers mas y menos relevantes para el modelo:

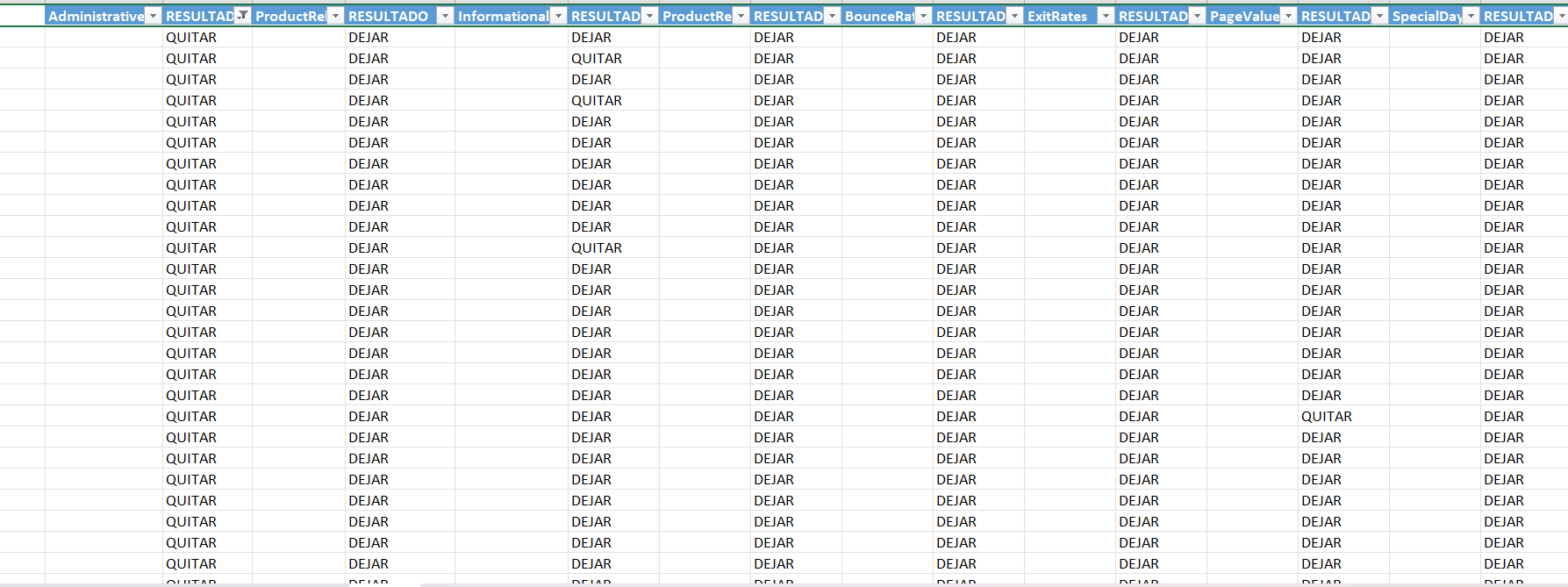




Al obtener estos grafico se puede evidenciar datos atípicos que están demasiado lejano a la mediana por lo que pueden estar generando ruido, estas variables son:

* Duración en pagina / Administrative\_Duration
* Producto relacionado / ProductRelated
* Duración paginas visitadas / Informational\_Duration
* Duración con el producto / ProductRelated\_Duration
* Tasa de rebote / BounceRates
* Tasa de salida / ExitRates
* Valores de pagina / PageValues
* Duración en día especial / SpecialDay

Se procede a investigar y eliminar los valores que estén por encima del último percentil:



1. Identificar un modelo de relaciones con la respectiva pregunta de investigación (eliminar variables no relevantes según el EDA y un check-face validity, definir las variables independientes y la dependiente). Explicar por qué las variables independientes inciden en la dependiente (hipótesis)
2. Realizar regresión logística usando Realstatistics o excel análisis y evaluar los resultados.
3. Realizar regresión logística con datos balanceados y evaluar los resultados (el balanceo de datos hacerlo por fuera de excel)
4. Comparar los resultados de 5 con 6, y establecer el mejor resultado. Determinar si los porcentajes de la variable objetivo son adecuados para el análisis que se hizo en 5.
5. Tener en cuenta para el análisis: coeficientes de las variables independientes, significancias, coeficiente de determinación, evaluación de la precisión del modelo (matriz de confusión), curva ROC, entre otros.
6. Presentar el 20 de mayo los resultados, haciendo énfasis en el paso a paso realizado (presentación por grupo). Identificar las variables que tienen mayor influencia en la variable objetivo (contextualizar esto según el negocio al que pertenece la base de datos).
7. Si sale todo bien, les pondré como definitiva 5.