Tecnológico de Monterrey

Materia: Analítica de Datos y Herramientas de Inteligencia Artificial I

Actividad Evaluable: Mapas de calor y boxplots

Equipo #2. Estudiantes:

Abner Palma García - A01735312

Edwin Nepomuceno Rivera - A01174706

Luis Alberto Mirón Toruño - A01735489

Paula Gabriela Armenta Nazario - A01735618

Fecha: 02 de Junio del 2023

Archivo: Antigüedad de saldos 2023

1. **¿Hay alguna variable que no aporte información?**
   No, en este caso todas las variables son relevantes para la comprensión de la antigüedad de saldos: 'No. CLIENTE', 'NOMBRE', 'FACTURA', 'FECHA_FACTURA','FECHA_VENCIMIENTO', 'MONTO ADEUDADO'.

2. **Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?**
   En este conjunto de datos, no eliminaría ninguna variable. Todas me parecen relevantes para el análisis de la antigüedad de saldos. Todas las variables pueden aportar información significativa para el análisis.

3. **¿Existen variables que tengan datos extraños?**
   No, todas las variables contienen información lógica relativa al nombre de su columna.

4. **Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?**
   Debido a que sólo se tiene una variable de tipo numérico en este conjunto de datos, no es posible responder esa pregunta, ya que existe únicamente un rango para la columna MONTO ADEUDADO.

5. **¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos?**
   En este conjunto de datos es posible identificar grupos de clientes por el monto adeudado a calor y control. Por ejemplo, es posible agrupar aquellos clientes que adeudan un monto mayor a 40,000 pesos. También podemos agrupar a aquellos clientes que deben menos de 20,000 a la empresa.

Archivo: Datos de Facturación

1. **¿Hay alguna variable que no aporte información?**
   En este caso las variables que son relevantes para la comprensión de las facturas son la siguientes: CVE_DOC, CVE_CLPV, STATUS, CVE_VEND, FECHAELAB, FECHA_DOC, FECHA_ENT, FECHA_VEN, FECHA_CANCELA, CAN_TOT y DES_TOT.
   Sin embargo, la de DES_FIN no tiene información, todos los valores están en 0. Es por ello que en el primer Boxplot se ve esta variable, sin embargo, en los demás gráficos ya no aparece debido a que me di cuenta que todos los valores están en 0 y por eso opté por no utilizarla en los demás graáficos.

2. **Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?**
   En este conjunto de datos, después de que se eliminó la de DES_FIN como se menciona y explica en la pregunta anterior, no eliminaría ninguna variable. Todas me

parecen relevantes para el análisis de las facturas, todas aportan la información que se necesita, la clave del documento, su status, que vendedor emite la factura, en que fecha se elaboró la factura, su fecha de vencimiento, si fue cancelada en que fecha fue y el monto por el cuál se emite la factura y si tuvo algún descuento.

3. **¿Existen variables que tengan datos extraños?**
   Basándonos en las variables proporcionadas en el archivo de facturas, todas parecen contener información lógica relacionada con el nombre de su columna. Estas variables representan diferentes aspectos de las facturas, como identificadores, estados, fechas y montos.

4. **Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?**
   5. CVE_DOC: Esta variable representa el código de documento de la factura. No es una variable numérica, por lo que no se puede evaluar en términos de rangos similares.
   6. CVE_CLPV: Representa el código del cliente/proveedor. Al igual que CVE_DOC, no es una variable numérica, por lo que no se puede evaluar en términos de rangos similares.
   7. STATUS: Esta variable indica el estado de la factura, por ejemplo, si está pagada o pendiente de pago. No es una variable numérica, por lo que no se puede evaluar en términos de rangos similares.
   8. CVE_VEND: Representa el código del vendedor asociado a la factura. Al igual que las variables anteriores, no es una variable numérica, por lo que no se puede evaluar en términos de rangos similares.
   9. FECHAELAB, FECHA_DOC, FECHA_ENT, FECHA_VEN, FECHA_CANCELA: Estas variables representan fechas asociadas a la factura, como la fecha de elaboración, fecha del documento, fecha de entrega, fecha de vencimiento y fecha de cancelación. Estas variables pueden ser evaluadas en términos de rangos de fechas similares para determinar si las facturas se generan, entregan o cancelan en períodos similares. Si estas fechas difieren significativamente, puede indicar problemas en la gestión de facturas.
   10. CAN_TOT: Esta variable representa el monto total de la factura. Es una variable numérica y puede ser evaluada en términos de rangos similares. Si las facturas tienen montos totales muy diferentes, esto puede indicar discrepancias o irregularidades en las transacciones.
   11. DES_TOT, DES_FIN: Estas variables representan los descuentos totales y financieros aplicados a la factura. Al igual que CAN_TOT, son variables numéricas y pueden ser evaluadas en términos de rangos similares.

   En resumen, algunas variables como las fechas y los montos pueden ser evaluadas en términos de rangos similares para determinar si hay alguna variabilidad significativa. Sin embargo, las variables no numéricas como los códigos y estados de las facturas no pueden ser evaluadas en términos de rangos similares. La afectación dependerá del

contexto y de cómo se utilicen estas variables en el análisis o procesamiento de las facturas.

12. **¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos?**
    En este conjunto de datos es posible identificar grupos de acuerdo al status de la factura, si es emitida o cancelada. También para agrupar las facturas según la fecha, se puede aplicar un análisis de series temporales o utilizar técnicas de agrupamiento basadas en la proximidad de las fechas. Por ejemplo, se pueden crear grupos de facturas según el trimestre o mes en que se generaron o según la cercanía de las fechas de elaboración, vencimiento o entrega.

    Respecto al monto total de las facturas, se puede aplicar una técnica de agrupamiento basada en el valor numérico de CAN_TOT. Por ejemplo, se pueden crear grupos de facturas en función de rangos específicos de montos, como facturas de bajo valor, facturas de valor medio y facturas de alto valor.

Archivo: Gastos y costos

1. **¿Hay alguna variable que no aporte información?**

    Considero que toda la información que se nos presenta en el documento en Excel es relevante sin embargo después de realizar la limpieza y eliminar algunas filas que no ocupamos que están vacías nos quedamos con la información pura que nos ofrece el socio formador y respecto a la información que nos queda considero que todo es relevante aunque existen algunas variables que son más relevantes que otras como por ejemplo la fecha el tipo de gasto y el importe que son variables muy importantes considerando el contexto de la base de datos analizada sin embargo existen variables como la variable "GASTO" y la variable "TC" que tiene muchos valores nulos por lo que aportan menor información al análisis realizado.

2. **Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?**

    Eliminaría las columnas o las variables que tuviesen o aportadas en menos información como se mencionó anteriormente la columna de gasto y tc no tienen la información completa ya que tiene muchos valores no los mientras que la variable de póliza únicamente tiene 21 valores de los 3342 filas que existen en el el conjunto de datos por lo que en primera instancia eliminaría esas tres columnas ya que no aportan suficiente información y hace muy difícil categorizar valores que no tienen valores en esa categoría.

3. **¿Existen variables que tengan datos extraños?**

Como podemos observar en la gráfica de cajas hay variables que tienen muchos valores nulos y algunos de estas variables que tienen valores nulos esos valores son muy altos en comparación a la mayoría de los datos que existen en la variable en cuestión como por ejemplo observamos que la variable de importe y total además del total MX tienen muchos valores nulos y algunos de estos tienen valores que son negativos y que y otros son muy grandes por mucho.

4. **Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?**

   hay algunas columnas que tienen rasgos similares como por ejemplo el importe el "RET ISR" y el "RET IVA" tienen valores con rangos muy similares que van de 0 a 3000 considero que estos rangos no afectan ya que depende mucho el contexto de la información a la que se está analizando ya que estas variables se relacionan fuertemente por lo que no considero que afecte que tengan un Rango similar ya que su interpretación es distinta sin embargo considero que también hay que tener mucho cuidado en esa cuestión para no repetir información que no sea relevante o que ya se haya analizado previamente, sin embargo podemos observar que existen columnas dentro de nuestro conjunto de datos que no tiene los mismos rangos como el "IVA" etc.

5. **¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos?**
   En este conjunto de datos se pueden agrupar por distintos valores como por ejemplo existe una variable que tiene el tipo de gasto en el cual existen categorías que representan el contexto en el que se gastó el importe realizado de esa manera podemos realizar una agrupación por el tipo de gasto que se realiza siendo comisión ,sueldos, honorarios etcétera sin embargo también se pueden realizar categorizaciones o agrupar los valores de acuerdo a la fecha ya que en esta podemos agruparlo por año mes etcétera además de esto también existen variables de estatus por lo que también podemos agrupar la información del conjunto de datos por el tipo de estatus ya sea vigente o no vigente y por último también los podemos agrupar o diferenciar por el importe que se realiza ya que al ser una variable numérica podemos identificar o agrupar los gastos que se realizan por el importe siendo rangos de valores ya sea los gastos más caros los gastos más pequeños etcétera.

Archivo: Precios y Productos 2022

1. **¿Hay alguna variable que no aporte información?**
   No, dentro de este documento todas las columnas muestran algún tipo de información relevante que nos permite, correlacionar los datos mostrados, y de esta manera poder seleccionar aquellos que puedan ser de mayor utilidad, para el proyecto.

2. **Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?**

Existen columna que tienen los siguientes nombres por dar un ejemplo “COSTO UNITARIO” y “COSTO UNITARIO CALCULADO” y dentro de la información de cada columna el 95% de sus datos son iguales, siendo la excepción el otro 5% en donde varía, para este caso, sería conveniente analizar cuál columna es más representativa o tener en cuenta por que ocurre esta variación, y así poder quedarnos con la columna que sea más conveniente para los objetivos del proyecto y de esta manera poder tener los resultados más acertados.

3. **¿Existen variables que tengan datos extraños?**
   No, en el caso de este archivo todos los datos contenidos están correctamente establecidos con el nombre de la columna al que pertenecen.

4. **Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?**
   Si, hay columnas que contienen datos similares, pero que tienen una correlación en común y lógica, por lo tanto esto no tiene ningún efecto negativo en los objetivos que se quieren alcanzar con el proyecto.

5. **¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos?**
   En este caso podemos encontrar grupos principalmente ocupando la columna NOMBRE_VENDEDOR, ya que por ejemplo podemos encontrar una relación con la cantidad de artículos vendidos, esto nos permite crear una correlación entre el vendedor y la cantidad de productos que ha venido, además esta misma correlación puede servir y ser aplicada al nombre del cliente, ya que podemos ver el número de artículos vendidos y a que cliente se le ha vendido más, o que vendedor tiene más injerencia en la compra de un cliente o producto.

Archivo: Facturación, Notas de Crédito, Devoluciones y Clientes

En este archivo solo trabajaremos con DEVOLUCIONES, NOTAS DE CRÉDITO, CLIENTES ya que FACTURACIÓN se hizo con el dataset de DATOS DE FACTURACIÓN.

1. **¿Hay alguna variable que no aporte información?**
   Tanto para las hojas de Notas de Crédito y Devoluciones:
   En la hoja de clientes las variables disponibles son "CLAVE", "RFC" y "NOMBRE", las cuales son variables categóricas. Las variables categóricas representan características o atributos que no pueden ser cuantificados o medidos de forma numérica. Estas variables no tienen un orden inherente ni una relación numérica entre sus categorías.

   En este caso las variables que son relevantes para la comprensión de las notas de crédito son la siguientes: TIP_DOC, CVE_DOC, CVE_CLPV, STATUS, CVE_VEND, CVE_PEDI, FECHA_DOC, FECHA_ENT, FECHA_VEN, FECHA_CANCELA, CAN_TOT, FECHAELAB, RFC, SERIE, FOLIO

Las variables que son relevantes para la comprensión de las devoluciones son la siguientes: CVE_DOC, CVE_CLPV, STATUS, CVE_VEND, FECHAELAB, FECHA_DOC, FECHA_ENT, FECHA_VEN, FECHA_CANCELA, CAN_TOT, DES_TOT, DES_FIN, RFC

Sin embargo, la de DES_FIN no tiene información, todos los valores están en 0. Es por ello que en los gráficos con variables numéricas ya no aparece debido a que me di cuenta que todos los valores están en 0 y por eso opté por no utilizarla.

2. **Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?**
   Tanto para las hojas de Notas de Crédito y Devoluciones:
   En este conjunto de datos, después de que se eliminó la de DES_FIN como se menciona y explica en la pregunta anterior, no eliminaría ninguna variable. Todas me parecen relevantes para el análisis de las notas de crédito y devoluciones, todas aportan la información que se necesita, la clave del documento, su status, que vendedor emite la factura, en que fecha se elaboró la factura, su fecha de vencimiento, si fue cancelada en que fecha fue y el monto por el cuál se emite la factura y si tuvo algún descuento.

3. **¿Existen variables que tengan datos extraños?**
   Basándonos en las variables proporcionadas en el archivo de Facturación, Notas de Crédito, Devoluciones y Clientes, todas parecen contener información lógica relacionada con el nombre de su columna. Estas variables representan diferentes aspectos de las Notas de Crédito y Devoluciones, como identificadores, status, fechas y montos.

4. **Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?**
   De las variables en común entre las hojas de Notas de Crédito y Devoluciones:

- CVE_DOC: Esta variable representa el código de documento de la factura. No es una variable numérica, por lo que no se puede evaluar en términos de rangos similares.
- CVE_CLPV: Representa el código del cliente/proveedor. Al igual que CVE_DOC, no es una variable numérica, por lo que no se puede evaluar en términos de rangos similares.
- STATUS: Esta variable indica el estado de la factura, por ejemplo, si está pagada o pendiente de pago. No es una variable numérica, por lo que no se puede evaluar en términos de rangos similares.
- CVE_VEND: Representa el código del vendedor asociado a la factura. Al igual que las variables anteriores, no es una variable numérica, por lo que no se puede evaluar en términos de rangos similares.
- FECHAELAB, FECHA_DOC, FECHA_ENT, FECHA_VEN, FECHA_CANCELA: Estas variables representan fechas asociadas a la factura, como la fecha de elaboración, fecha del documento, fecha de entrega, fecha de vencimiento y fecha de cancelación. Estas variables pueden ser evaluadas en términos de rangos de fechas

similares para determinar si las facturas se generan, entregan o cancelan en períodos similares. Si estas fechas difieren significativamente, puede indicar problemas en la gestión de facturas.

- CAN_TOT: Esta variable representa el monto total de la factura. Es una variable numérica y puede ser evaluada en términos de rangos similares. Si las facturas tienen montos totales muy diferentes, esto puede indicar discrepancias o irregularidades en las transacciones.
- DES_TOT, DES_FIN: Estas variables representan los descuentos totales y financieros aplicados a la factura. Al igual que CAN_TOT, son variables numéricas y pueden ser evaluadas en términos de rangos similares.
- RFC: Esta variable representa el Registro Federal de Contribuyentes del cliente o proveedor asociado a la factura. Esta no es una variable numérica y no se puede evaluar en términos de rangos similares.

En resumen, las variables comunes entre las hojas de Notas de Crédito y Devoluciones, como CVE_DOC, CVE_CLPV, STATUS, CVE_VEND y RFC, no son variables numéricas y, por lo tanto, no se pueden evaluar en términos de rangos similares. Sin embargo, variables como las fechas (FECHAELAB, FECHA_DOC, FECHA_ENT, FECHA_VEN, FECHA_CANCELA) y los montos (CAN_TOT, DES_TOT, DES_FIN) pueden ser evaluadas en términos de rangos similares para identificar variabilidad significativa. La falta de rangos similares en las variables no numéricas puede limitar el tipo de análisis y visualizaciones que se pueden realizar, y debe considerarse al interpretar los datos y tomar decisiones basadas en ellos.

5. **¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos?**
   De las variables en común entre las hojas de Notas de Crédito y Devoluciones:
   En este conjunto de datos es posible identificar grupos de acuerdo al status de la nota de crédito o devolución, si es emitida o cancelada. También para agruparlas según la fecha, se puede aplicar un análisis de series temporales o utilizar técnicas de agrupamiento basadas en la proximidad de las fechas. Por ejemplo, se pueden crear grupos de facturas según el trimestre o mes en que se generaron o según la cercanía de las fechas de elaboración, vencimiento o entrega. Respecto al monto total, se puede aplicar una técnica de agrupamiento basada en el valor numérico de CAN_TOT. Por ejemplo, se pueden crear grupos en función de rangos específicos de montos, como facturas de bajo valor, facturas de valor medio y facturas de alto valor.