*Para la tarea se les pide:*

1. Indicar qué problemas detecta en la base de datos y tome medidas para corregirlos. Debe indicar el procedimiento que siguió para resolverlos de forma clara y precisa

***Respuesta:*** Se encontraron valores faltantes en 6 variables:

|  |  |
| --- | --- |
| Variables | NA |
| GENERO | 11 |
| EDAD | 2 |
| NIV\_EDUC | 11 |
| E\_CIVIL | 11 |
| COD\_COM | 3 |
| CIUDAD | 3 |

Los valores faltantes son aquellos campos vacíos que se encuentran en las filas correspondiente a una o varias variables. Estos campos resultan ser importantes, ya que ellos pueden ayudar a extraer información valiosa, por lo que es necesario **imputar** esos valores mediante estimaciones.

En secuencia a esto, para las variables cualitativas se utilizará la Moda que tenga mayor frecuencia y para las variables cuantitativas se utilizará la Media (obviando los NA), quedando estas medidas de la siguiente manera:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Imputación | | |
| Variables | NA | Media | Moda | |
| GENERO | 11 |  | M | 1385 datos |
| EDAD | 2 | 44 |  |  |
| NIV\_EDUC | 11 |  | UNV | 947 datos |
| E\_CIVIL | 11 |  | CAS | 1429 datos |
| COD\_COM | 3 | 108 |  |  |
| CIUDAD | 3 |  | SANTIAGO | 1389 datos |

Por otra parte, el campo EDAD posee caracteres especiales (en este caso el signo menos) en algunos registros, por la cual se consideró eliminarlos. Algo similar ocurrió con la variable MONTO; esta tenía valores con 2 notaciones distintas (notación científica y notación decimal), por la cual fue necesario cambiar dicha notación antes de importar el conjunto de datos, y posteriormente darle un formato numérico para no tener problemas más adelante.

1. Realizar un análisis descriptivo de los datos. En este punto deben calcular las principales métricas estadísticas para una mejor descripción de los datos, por ejemplo, estadísticas descriptivas, tablas de frecuencia, etc. Apoye su análisis utilizando gráficos.

* ***Variables estadísticas***

|  |  |  |
| --- | --- | --- |
| VARIABLES | TIPO | |
| CUALITATIVA | CUANTITATIVA |
| ID |  | discreta |
| GENERO | nominal |  |
| RENTA |  | continua |
| EDAD |  | discreta |
| NIV\_EDUC | ordinal |  |
| E\_CIVIL | nominal |  |
| COD\_OFI |  | discreta |
| COD\_COM |  | discreta |
| CIUDAD | nominal |  |
| D\_Marzo |  | continua |
| D\_Abril |  | continua |
| D\_Mayo |  | continua |
| D\_Junio |  | continua |
| D\_Julio |  | continua |
| D\_Agosto |  | continua |
| D\_Septiembre |  | continua |
| M\_MOROSO |  | discreta |
| MONTO |  | continua |
| SEGURO | nominal |  |
| FUGA | nominal |  |
| Total: 20 variables | | |

* ***Medidas de posición central***
* ***medidas de posición no central***

EN EL SCRIPT ESTÁN

* ***Medidas de dispersión***
* ***Tablas de frecuencias***

Se escogieron las variables cuantitativas RENTA y MONTO para representar las tablas de frecuencia. En el caso de las variables cualitativas, se escogieron NIV\_EDUC y FUGA.

Para RENTA

1. Transformación de datos y generación de indicadores: transforme las variables disponibles de forma de encontrar nuevas variables, generadas a partir de las originales, que puedan tener un buen desempeño predictivo a partir de sus conocimientos del negocio

A veces es necesario utilizar variables categóricas en métodos analíticos que solo permiten el ingreso de variables numéricas, por lo que a veces es necesario crear variables indicadoras. Con el paquete dummies podemos hacer estas conversiones.

Al hacer la aplicación del modelo dummies podemos observar que pasa de 21 variables a 105 variables posibles indicadores, por lo consiguiente haciendo una observación, las variables más relevantes fueron:

**GENERO M:** género masculino

**GENERO F:** género femenino

**NIVEDUECUN:** nivel educacional universitario

**NIVEDUBAS:** nivel educacional básico

**NIVEDUUTEC:** nivel educacional técnico

**EDAD:**  edad de la persona

**TOTAL DEUDA:** la sumatoria de las deudas individuales de cada mes

1. Selección de atributos: analice sus variables de la base. ¿Discriminan entre clientes fugados y no fugados?

Para responder esta pregunta describimos los atributos iniciales del dataset y a continuación los atributos finales, escogidas las variables indicadoras.

|  |
| --- |
| 1. ID: identificador del cliente. |
| 2. Género: Genero del cliente. |
| 3. Renta: Renta en pesos. |
| 4. Edad: Edad en años. |
| 5. NIV\_Educ: Nivel educacional. |
| 6. E\_Civil: Estado civil. |
| 7. COD\_Ofi: Código de la oficina |
| 8. COD\_Com: Código de la comuna. |
| 9. Ciudad: Ciudad de la oficina. |
| 10. D\_Marzo: Deuda de Marzo. |
| 11. D\_Abril: Deuda de Abril. |
| 12. D\_Mayo: Deuda de Mayo. |
| 13. D\_Junio: Deuda de Junio. |
| 14. D\_Julio: Deuda de Julio. |
| 16. D\_Agosto: Deuda de Agosto. |
| 17. D\_Septiembre: Deuda de Septiembre. |
| 18. M\_Moroso: Meses en Mora. |
| 19. Monto: Monto crédito preaprobado. |
| 20. Seguro: Seguro de gravamen. |
| 21. Fuga: Variable objetivo. |

\* **VARIABLES INDICADORAS QUE DISCRIMINAN ENTRE CLIENTES FUGADOS Y NO FUGADOS**

|  |
| --- |
| Género: Genero del cliente. |
| Renta: Renta en pesos. |
| Edad: Edad en años. |
| NIV\_Educ: Nivel educacional. |
| M\_Moroso: Meses en Mora. |
| Monto: Monto crédito preaprobado. |
| Fuga: Variable objetivo. |
| Deuda total: sumatoria de todas las deudas en los diferentes meses |