Fernández Montes Mariel Quetzali / A01707797

Actividad 4 (Extracción de características)

Del documento de Excel “microretailer\_mit\_lift\_lab” se realizó una limpieza de datos, eliminando valores nulos y datos atípicos para posteriormente realizar un análisis univariado de las variables cualitativas y poder obtener conclusiones. A continuación, se presenta a modo de instrucciones los pasos que se llevaron a cabo para poder realizar el presente análisis de datos:

Etapa 1: Procesamiento de datos

1. Instalamos/ importamos las siguientes librerías:

* Instalamos funpymodeling:
  + Obtenemos la librería “freq\_tbl”: Nos ayudará a graficar los valores univariados de las columnas.
* Importamos librería “pandas” y la nombramos “pd”: Ayuda a visualizar el análisis de las variables.
* Importamos librería “numpy” y la nombramos “pd”: Ayuda a realizar operaciones sencillas con las variables que requiramos.
* Importamos librería “matplotlib.pyplot” y la nombramos “plt”: Ayuda a graficar las variables que requiramos.

1. Importamos el archivo y lo convertimos en Data Frame
2. Con la función “info()” obtenemos la información del Data Frame
3. Realizamos el conteo de valores nulos por columna del Data Frame
4. Generamos una copia del Data Frame, copia que se utilizará para manipularla posteriormente y no manipular el Data Frame original.
5. Con la función “select\_dtypes” agrupamos los datos cualitativos en una sola variable llamada “cuali” y los datos cuantitativos en otra variable llamada “cuanti”. Esto con la finalidad de manipular los datos por separado.
6. Eliminamos las columnas que no cuentan con ningún tipo de información ya que solo utilizan espacio innecesario, en el caso de este documento son 4 columnas.
7. Sustituimos los valores nulos de la variable “cuali” por el texto “Sin datos”
8. Sustituimos los valores nulos de la variable “cuanti” por el número “0”. No se utilizaron diferentes métodos de sustitución de valores nulos pues no son datos que se puedan relacionar entre sí, por lo que asignar el promedio de la columna o los valores siguientes o anteriores al valor nulo, afectaría el análisis de los datos.
9. Una vez limpia la información concatenamos las variables cualitativas y cuantitativas en un solo Data Frame para tener una sola tabla con toda la información.
10. Realizamos nuevamente el conteo de los valores nulos para asegurarnos que todos fueron sustituidos por los datos que asignamos.
11. Para realizar el análisis de los valores atípicos, se utilizó la variable que solo contenía datos cuantitativos y se utilizó el método de desviación estándar que ayuda a limitar las 3 posiciones a la derecha y a la izquierda para identificar los valores que se encuentran dentro de este rango (inliers) y aquellas que no se encuentren dentro de este rango serán considerados como valores atípicos.
12. Imprimimos como NA en un nuevo Data Frame, los valores que son outliers para identificarlos rápidamente.
13. En otro Data Frame imprimimos los valores que no son outliers para identificarlos rápidamente.
14. Sustituimos los outliers por el promedio de la columna en el que se encuentra cada uno de los outliers.
15. Volvemos a concatenar la variable que cuenta con valores cuantitativos con la variable que tiene valores cualitativos para tener toda la información en un solo Data Frame.

Etapa 2: Extracción de características y visualización de datos

1. El nuevo Data Frame lo dividimos en 2 variables, una de valores cualitativos y uno de valores cuantitativos para una manipulación más sencilla.
2. Realizamos el análisis univariado y se ejecutó una gráfica de barras para las siguientes columnas del Data Frame

* \_updated\_by
* 184\_store\_devices
* 214\_customer\_relationship\_tools
* 157\_frequency\_organize\_shelves
* 158\_charact\_organize\_products\_on\_shelves

1. Realizamos el análisis univariado y se ejecutó una gráfica de pie para las siguientes columnas del Data Frame

* 108\_does\_the\_micro\_retailer\_has\_a\_barred\_window\_
* 99\_does\_the\_micro\_retailer\_exhibits\_products\_outside\_
* 102\_does\_the\_micro\_retailer\_sells\_fresh\_products\_
* hola\_somos\_estudiantes\_del\_tec\_de\_monterrey\_estamos\_realiz
* le\_comento\_que\_toda\_esta\_informacin\_es\_confidencial

1. Realizamos el análisis univariado y se ejecutó una gráfica de área para las siguientes columnas del Data Frame

* 268\_number\_fridges
* 2\_current\_permanent\_employees
* 4\_number\_permanent\_employees\_last\_year
* 145\_number\_direct\_competitors
* 163\_number\_high\_perishable\_products

1. Realizamos el análisis univariado y se ejecutó una gráfica de dispersión para las siguientes columnas del Data Frame

* pictures\_of\_shelves\_if\_possible
* 5\_change\_store\_space\_last\_year
* 49\_inventory\_records
* 189\_payment\_methods
* 18\_sales\_records