Reporte de Limpieza de Datos

Allan Mauricio Brenes Castro - A01750747

Octubre 2024

Resumen

El objetivo de limpiar los datos es verificar que la integridad de los datos esté correcta para poder obtener resultados adecuados a la hora de analizarlos. Se borraron los datos duplicados y aquellas filas que tenían datos nulos. Al final se convirtió la columna de “Income” a un dato numérico para poder realizar operaciones estadísticas en ellos.

1. Introducción

Los datos almacenan diferentes datos de clientes que compraron motocicletas. Almacenan datos como género, ingreso, edad, cuántos hijos tienen, máximo nivel de educación, entre otros varios datos. El objetivo de limpiar estos datos es de poder realizar un análisis correcto y evitar datos redundantes o nulos que podrían comprometer la identidad de los análisis. También es importante checar que el formato

1.1. Conjunto de Datos

Los datos se obtuvieron al hacer una compra en la tienda. No se encuentran ordenados. Una peculiaridad que tiene el archivo base es que “income” se encontraba como un campo string, lo cual haría muy complicado hacer cálculos matemáticos con ellos.

1.2. Objetivos de la Limpieza

La limpieza se realizó porque se encontraron filas con datos nulos, habían datos duplicados, y el campo de ingreso se encontraba como texto. Primero se va a obtener información del tipo de dato que almacena cada columna. Una vez hecho eso, se busca si hay datos duplicados y/o datos nulos con la ayuda de la librería de Python Pandas. Una vez hecho eso, se cambiará el tipo de dato de “income” a float.

2. Exploración Inicial

Como primer vistazo, se encontró que, efectivamente, “income” almacenaba datos de tipo string, esto porque incluía la coma para la separación de los números para una mayor legibilidad. También se encontraron varios datos duplicados y filas con datos nulos.

3. Proceso de Limpieza

Para la limpieza se usó en el dataframe el método dropna() y dropduplicate() y se almacenó en la memoria, esto para actualizar el dataframe con el que se trabaja y para el uso posterior.

3.1. Tratamiento de Valores Faltantes

Las filas con valores faltantes simplemente se eliminaron para tener un set de datos más puro.

3.2. Corrección de Datos Erróneos

No se encontraron datos erróneos.

3.3. Manejo de Valores Atípicos

Los valores atípicos se mantuvieron.

3.4. Estandarización y Normalización

Se eliminaron los espacios en blanco de todas las celdas de texto que incluían espacios en blanco al principio y/o al final. Una vez hecho esto, se cambiaron todos los datos de "income" a que sean numéricos.

4. Resultados

Una vez se limpiaron los datos, se realizó un análisis de los datos resultantes para obtener información de valor.

4.1. Estadísticas Descriptivas

	ID	Children	Cars	Age
count	1251.000000	1238.000000	1242.000000	1238.000000
mean	20030.208633	1.929725	1.479066	44.058966
std	5331.451777	1.638977	1.121885	11.271138
min	11000.000000	0.000000	0.000000	25.000000
25%	15465.000000	0.000000	1.000000	35.000000
50%	19731.000000	2.000000	1.000000	43.000000
75%	24549.000000	3.000000	2.000000	52.000000
max	29447.000000	5.000000	4.000000	89.000000

	ID	Income	Children	Cars	Age
count	952.000000	952.000000	952.000000	952.000000	952.000000
mean	19979.940126	55903.361345	1.898109	1.452731	44.256303
std	5334.000279	30845.483596	1.620426	1.111962	11.428167
min	11000.000000	10000.000000	0.000000	0.000000	25.000000
25%	15310.250000	30000.000000	0.000000	1.000000	35.000000
50%	19747.500000	60000.000000	2.000000	1.000000	43.000000
75%	24531.500000	70000.000000	3.000000	2.000000	52.000000
max	29447.000000	170000.000000	5.000000	4.000000	89.000000

Arriba se encuentra los datos sucios, mientras que abajo se muestran los datos ya limpios. Antes había un número diferentes de datos, mientras que ahora todos los datos numéricos están una misma cantidad de veces. También se puede apreciar que "Income" ahora se encuentra como dato numérico. Se puede apreciar que se eliminaron 299 filas duplicadas o con datos nulos.

4.2. Distribución de Datos

La distribución de los datos no cambió tras la limpieza.

5. Conclusiones

La limpieza de datos permitió liberar memoria que estaba siendo desperdiciada por la existencia de los datos duplicados y las filas con datos nulos. Eso sigue siendo una pequeña muestra de lo complicado que se podría volver con una cantidad de datos todavía más grande.

6. Recomendaciones

En el futuro es posible que se haga un script con los métodos dropna() y dropduplicated(), ya que es algo que se debería de hacer como limpieza inicial.

7. Referencias

Herramientas utilizadas:

Librería Pandas para Python

Librería Matplotlib para Python

Referencias:

Pandas documentation. (2024, 20 septiembre). pandas.pydata.org. Recuperado 23 de octubre de 2024, de https://pandas.pydata.org/docs/