Reporte de Limpieza de Datos

Lizbeth Islas Becerril
Sebastián Antonio Almanza

Octubre 2024

Resumen

Análisis de un dataset que contiene datos demográficos sobre las personas que han comprado o no una bicicleta

Índice

1. Introducción

En este documento se realizará la limpieza y posteriormente un análisis de datos sobre un dataset que contiene datos demográficos sobre las personas que han comprado o no una bicicleta. En las secciones posteriores se detallarán los tipos de datos del conjunto de datos dado seguido de los pasos de la limpieza de los mismos. Finalmente se darán conclusiones derivadas de los resultados obtenidos.

Conjunto de Datos

Dentro de este dataset encontramos los siguientes datos:

Datos	**Tipo de Datos**
ID	Int
Marital Status	String
Gender	Sting
Income	Int
Children	Int
Education	String
Occupation	String
Home Owner	String
Cars	Int
Commute Distance	String
Región	String
Age	Int
Purchased bike	Boolean

Objetivos de Limpieza

El objetivo de la limpieza es eliminar cualquier dato no dado o no útil dentro del dataset, ya que este puede contener valores no útiles como por ejemplo datos duplicados o de valor nulo. Es importante limpiar los datos antes de realizar un análisis de los datos para obtener una estimación más cercana a la realidad.

Posterior a la limpieza, se espera lograr un análisis de datos que nos permita entender el comportamiento de los compradores de bicicletas.

2. Exploración Inicial

Los primeros hallazgos encontrados fueron los siguientes:

- Se realizó una inspección rápida de los primeros 10 valores y los últimos 10 valores, para observar el comportamiento de los datos.
- Se investigó cómo están constituidos los datos.
- Se investigó el tamaño del dataset y se encontró que contiene (1251, 13).
- Se investigó si existían duplicidades
- Se investigó si el archivo contenía datos vacíos.

Proceso de limpieza

Como proceso de limpieza se realizó lo siguiente:

Se observó el comportamiento de los datos para poder identificar los datos anormales o los datos que sean propensos a obstruir la claridad de los datos.

- Se identificaron los valores nulos.
- Se corrigieron los datos “erróneos” (mala sintaxis)
- Se normalizaron los datos

Tratamiento de valores faltantes

Los valores nulos hace que el análisis se entorpezca, por ende se decidió buscar el número de valores nulos dentro del dataset; al encontrar que eran pocos se decidió eliminar los registros que contenían los valores nulos, ya que no afectan significativamente al análisis.

Corrección de datos erróneos y normalización

En el caso de el dataset asignado se tiene como valor erróneo que las cantidades grandes como el ingreso de la persona incluya coma, por lo tanto se elimina esta para tener un manejo correcto del valor.

Manejo de valores atípicos

Para detectar valores atípicos dentro del conjunto de datos realizamos gráficos de caja para poder visualizar si es que existen datos que se salgan del rango común. En el caso de la edad pudimos notar que en el gráfico hay 3 valores que se salen del rango común. Sin embargo, estos valores podrían significar que hay un sector de personas mayores interesadas en comprar bicicletas, por ende se decidió conservar los valores.

Resultados

Como resultado de la exploración de los datos obtuvimos lo siguiente:

	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

- Edad
    - La edad máxima es de 89
    - La edad mínima es de 25
    - La edad promedio es de 44
- Hijos
    - La cantidad de hijos máxima es de 5
    - La cantidad de hijos mínima es de 0.
    - La cantidad promedio de hijos es de 1.9
- Autos
    - La cantidad de autos máxima es de 4
    - La cantidad de autos mínima es de 0
    - La cantidad de autos promedio es de 1.4

Conclusiones

Contar con estos conocimientos nos permitirá mejorar nuestras competencias como ingenieros en tecnología, ampliando nuestra perspectiva para comprender cómo limpiar datos en un dataset y tomar decisiones fundamentadas en la información proporcionada.

Recomendaciones

Como trabajo futuro sugerimos mejorar la recopilación de datos para obtener menores valores nulos (faltantes) y revisar a fondo el dataset para evitar duplicados dentro de él. En cuestiones de herramientas de análisis de datos sugerimos utilizar el software de SAS ya que permite un análisis más ágil y conciso de los datos a comparación de python, además de poder automatizar la limpieza de los datos