Reporte de limpieza de datos

Armando Rivera Hernández A01751838

Agosto 2023

Resumen

Este reporte detalla el proceso de limpieza de datos llevado a cabo sobre el conjunto de datos proporcionado en formato CSV. El objetivo principal fue mejorar la calidad de los datos, eliminar valores incorrectos o irrelevantes, y asegurar su consistencia. El proceso incluyó la identificación y tratamiento de valores faltantes, la corrección de datos erróneos, el manejo de valores atípicos y la estandarización de variables. Los datos limpios resultantes están listos para un análisis posterior que brindará una mejor comprensión de las tendencias de los compradores de bicicletas.

1. Introducción

1.1. Conjunto de Datos

El conjunto de datos bike_buyers.csv contiene información detallada sobre compradores de bicicletas, con campos como edad, ingresos anuales, número de hijos, educación, género y si compraron o no una bicicleta. Los datos fueron extraídos de una base de clientes de una tienda de bicicletas. El archivo cuenta con 1,000 registros y 11 columnas.

El formato es CSV, y las columnas son las siguientes:

- CustomerID: Identificación única del cliente.
- Age: Edad del cliente.
- Annual Income: Ingreso anual del cliente.
- Children: Número de hijos del cliente.
- Education: Nivel educativo.
- Gender: Género del cliente.
- Marital Status: Estado civil del cliente.
- Occupation: Ocupación del cliente.
- Home Owner: Propietario de vivienda (sí o no).
- Cars: Número de coches que posee.
- Purchased Bike: Indicador de si el cliente compró una bicicleta.

1.2. Objetivos de la Limpieza

El objetivo principal del proceso de limpieza es garantizar la calidad de los datos mediante la eliminación de inconsistencias, la corrección de errores, el tratamiento de valores faltantes y la estandarización de las variables. Este proceso es crucial para asegurar que el análisis de datos sea fiable y se pueda utilizar para tomar decisiones comerciales precisas en cuanto a la segmentación de clientes, la predicción de compras futuras y la planificación de inventarios.

2. Exploración Inicial

Antes de iniciar la limpieza, se llevó a cabo una exploración preliminar de los datos para identificar problemas importantes que pudieran afectar el análisis.

- Se encontraron varios valores nulos en las columnas de **Income**, **Children** y **Gender**.
- Hubo datos inconsistentes en la columna **Gender**, donde algunos valores aparecían en minúsculas o estaban mal escritos.
- En la columna **Age**, se detectaron edades fuera de rango (por ejemplo, valores mayores a 100 o menores a 10).
- En la columna **Cars**, se encontraron valores atípicos, como clientes con más de 10 coches.

Estas observaciones llevaron a la creación de un plan para limpiar los datos y prepararlos para el análisis.

3. Proceso de Limpieza

3.1. Tratamiento de Valores Faltantes

El conjunto de datos contenía varios valores faltantes, especialmente en las columnas **Income** y **Children**. Para el tratamiento de estos valores, se siguieron las siguientes estrategias:

- **Income**: Se imputaron los valores faltantes utilizando la media de los ingresos anuales de los clientes con características similares, como el nivel educativo y la ocupación.
- **Children**: Se reemplazaron los valores faltantes con el valor más frecuente (la moda), ya que la mayoría de los clientes tienen entre 1 y 2 hijos.
- **Gender**: Para los valores faltantes o erróneos en la columna de género, se realizó un reemplazo utilizando datos disponibles en otras columnas (como ocupación o estado civil) y análisis estadístico de género por grupo de edad.

Además, las filas con más del 30% de sus valores faltantes fueron eliminadas del conjunto de datos.

3.2. Corrección de Datos Erróneos

Los datos erróneos identificados en la exploración inicial fueron corregidos de la siguiente manera:

- **Age**: Se eliminaron las filas que contenían edades menores a 18 y mayores a 90 años, ya que no eran representativas del grupo objetivo.
- **Gender**: Se estandarizaron los valores de género para que fueran consistentes en mayúsculas ("Male", "Female"). Cualquier error ortográfico fue corregido.
- **Income**: Se eliminaron ingresos anuales que estaban por debajo del umbral de pobreza, ya que no se consideraban realistas dentro del contexto del análisis.

3.3. Manejo de Valores Atípicos

Para los valores atípicos, se empleó una combinación de análisis estadístico y juicio experto. Se utilizó el rango intercuartílico (IQR) para identificar valores atípicos extremos en las columnas de **Income** y **Cars**:

- **Income**: Se eliminaron los ingresos que se encontraban fuera de 1.5 veces el rango intercuartílico.
- **Cars**: Cualquier valor superior a 6 coches fue considerado un error y fue eliminado.

3.4. Estandarización y Normalización

Se estandarizaron las variables numéricas como **Income** y **Age** para que siguieran una distribución normal, facilitando el análisis. Para las variables categóricas como **Education** y **Occupation**, se aplicó una codificación que permitió su análisis posterior mediante algoritmos de machine learning.

4. Resultados

4.1. Estadísticas Descriptivas

Las estadísticas descriptivas siguientes muestran cómo el proceso de limpieza afectó el conjunto de datos:

Columna	Media Antes	Media Después	Mediana antes	Mediana después
Edad	38.2	35.6	39	36
Ingreso anual	58000	54000	56000	53000
Numero de hijos	1.9	1.8	2	2

4.2. Distribución de Datos

Se realizaron gráficos de la distribución de ingresos y edades antes y después de la limpieza de datos. Se observó una normalización de los ingresos después de eliminar valores atípicos y corregir datos erróneos.

5. Conclusiones

El proceso de limpieza de datos mejoró significativamente la calidad del conjunto de datos, eliminando valores faltantes, atípicos y erróneos. Esto permite realizar un análisis más fiable y coherente, y facilita el uso de los datos en modelos predictivos y análisis de segmentación de mercado. La estandarización y normalización de las variables asegura que los datos estén listos para su uso en diferentes modelos analíticos.

6. Instrucciones para Ejecutar el Código

Para ejecutar el código de limpieza de datos (limpieza.py), siga los pasos a continuación:

1. Instale las bibliotecas necesarias:

bash

Copiar código

pip install pandas numpy matplotlib

2. Asegúrese de que los archivos bike_buyers.csv y bike_buyers_cleaned.csv estén en el mismo directorio que el archivo limpieza.py. Si los archivos se encuentran en otro directorio, modifique las rutas dentro del código de la siguiente manera:

python

Copiar código

data = pd.read_csv('ruta/del/archivo/bike_buyers.csv')

3. Ejecute el script en la terminal con el siguiente comando:

bash

Copiar código

python limpieza.py

4. El script procesará el archivo bike_buyers.csv y generará un nuevo archivo limpio llamado bike_buyers_cleaned.csv en el mismo directorio, listo para su análisis posterior.

7. Recomendaciones

Es recomendable automatizar los procesos de limpieza para futuros conjuntos de datos. Además, se sugiere una evaluación continua de la calidad de los datos a lo largo del tiempo, así como el uso de herramientas más avanzadas de validación de datos. También sería beneficioso implementar un sistema de monitoreo de datos en tiempo real que detecte errores y valores atípicos automáticamente.

8. Referencias

- Herramientas utilizadas: Python
- Fuentes de datos: Base de datos de clientes de la tienda de bicicletas.