Herramientas computacionales: el arte de la analítica

Grupo: 201

Aram Baruch

Reporte de Limpieza de datos

Octubre 2024

Jocelyn Ileana Balderas Sánchez - A01798528 Emiliano Alberto Celis Montero - A01799348

Santiago Chevez Trejo - A01749887

Resumen

Este reporte describe el proceso de limpieza de datos aplicado al archivo *"bike_buyers.csv"* utilizando Python en un cuaderno de Jupyter. El objetivo de esta limpieza fue identificar y corregir anomalías en los datos, asegurando su calidad y consistencia para su posterior análisis en el proyecto final.

Índice

1. Introducción

El presente reporte describe el procedimiento de limpieza de datos aplicado al archivo *"bike_buyers.csv"* utilizando el lenguaje de programación Python en un cuaderno de Jupyter.

Es importante comprender que la limpieza de datos es una etapa crítica en cualquier análisis, ya que asegura que el conjunto de datos esté libre de errores, inconsistencias y valores faltantes, lo que permite obtener resultados confiables y precisos en los análisis posteriores. Trabajar con datos sucios o incompletos puede llevar a conclusiones erróneas y comprometer la calidad de las decisiones basadas en los resultados. Por ello, realizar una correcta limpieza de los datasets es esencial para garantizar la validez y consistencia de los datos antes de aplicar cualquier modelo analítico o predictivo.

Durante esta actividad, se identificaron anomalías en los datos del archivo *"bike_buyers.csv"*, por lo que se implementaron diversas técnicas para corregirlas, y se realizaron análisis adicionales que revelaron algunos fenómenos interesantes que se describen más a detalle en el punto 4.4

1.1. Conjunto de Datos

El conjunto de datos *"bike_buyers.csv"* contiene 13 columnas sobre la información de los compradores de bicicletas.

Estas columnas son: "ID", "Marital Status", "Gender", "Income", "Children", "Education", "Occupation", "Home Owner", "Cars", "Commute Distance", "Region", "Age", "Purchased Bike".

Cada una proporciona información relevante para conocer más a los compradores de bicicletas.

1.2. Objetivos de la Limpieza

El objetivo de la limpieza fue asegurar la consistencia, exactitud y utilidad del dataset, eliminando valores atípicos y faltantes, y corrigiendo datos erróneos, para optimizar el análisis de los datos.

2. Exploración Inicial

Se utilizaron data frames y algunas funciones como:
df = pd.read_csv('bike_buyers.csv')
df.head(10)
df.info()
para cargar el conjunto de datos, así como observar las primeras 10 filas y obtener una descripción general de las columnas con los tipos de datos y valores faltantes, y se identificaron varias anomalías, como valores nulos en varias columnas, otro tipo de datos en ciertas columnas y duplicación de algunas filas.

3. Proceso de Limpieza

3.1. Tratamiento de Valores Faltantes

Se observó la cantidad de datos nulos en cada columna al usar la función ***df.isna().sum().***

Se analizaron las columnas con valores nulos y se aplicaron medidas estadísticas para imputarlos:

- Moda: para las columnas categóricas "Marital Status", "Gender" y "Home Owner".
    - La moda es la opción más adecuada para este tipo de datos porque representa el valor más frecuente en una columna categórica, asegurando que el dato imputado sea coherente con los valores existentes y no introduzca distorsiones en las categorías.
- Mediana: para las columnas numéricas "Income", "Children", "Cars" y "Age".
    - La mediana es una medida robusta de tendencia central que no se ve afectada por valores atípicos, lo cual la hace más confiable que el promedio en columnas que pueden tener outliers. Al usar la mediana, se garantiza que los valores imputados reflejen mejor el centro de la distribución de los datos, sin ser influenciados por extremos.

En seguida de esta imputación de datos, se verificó nuevamente que no quedaran valores faltantes: ***df.isna().sum()***

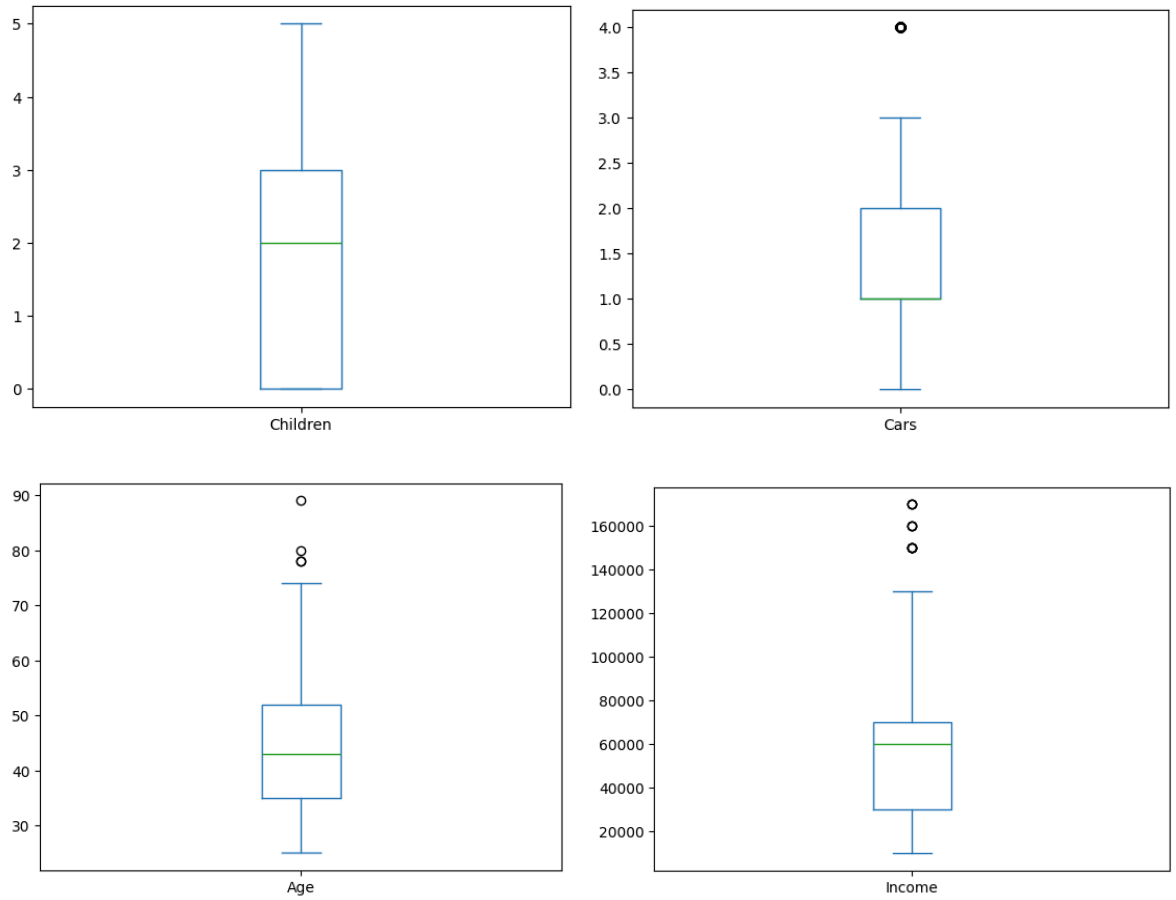
3.2. Corrección de Datos Erróneos

Se detectaron valores en la columna "Income" que contenían comas y espacios en blanco, dificultando el análisis numérico. Estos problemas se corrigieron eliminando los caracteres adicionales y convirtiendo los valores a formato numérico porque estaban como string.

df["Income"] = pd.to_numeric(df["Income"].str.replace(",", ""))

3.3. Manejo de Valores Atípicos

Se generaron gráficos de caja para las columnas "Children", "Cars", "Age" e "Income". Los gráficos revelaron la presencia de valores atípicos, especialmente en "Income" y "Cars", donde algunos valores extremadamente altos sugieren que ciertos compradores tienen características distintas que podrían requerir análisis adicional.



3.4. Estandarización y Normalización

Se estandarizó el formato de las columnas categóricas eliminando espacios en blanco y saltos de línea, y convirtiendo el texto a minúsculas. Esto se aplicó a columnas como "Marital Status", "Gender", "Education", y otras, para asegurar la uniformidad en el análisis.

columnas = ["Marital Status", "Gender", "Education", "Occupation", "Home Owner", "Commute Distance", "Region", "Purchased Bike"]

for col in columnas:

df[col] = df[col].str.strip().str.replace("\n", " ").str.lower()

4. Resultados

4.1. Estadísticas Descriptivas

Las estadísticas descriptivas mostraron una distribución más homogénea en los datos tras la corrección de errores y la imputación en los valores nulos.

Antes de la limpieza

```
df.describe()
✓ 0.0s
```

	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

Después de la limpieza

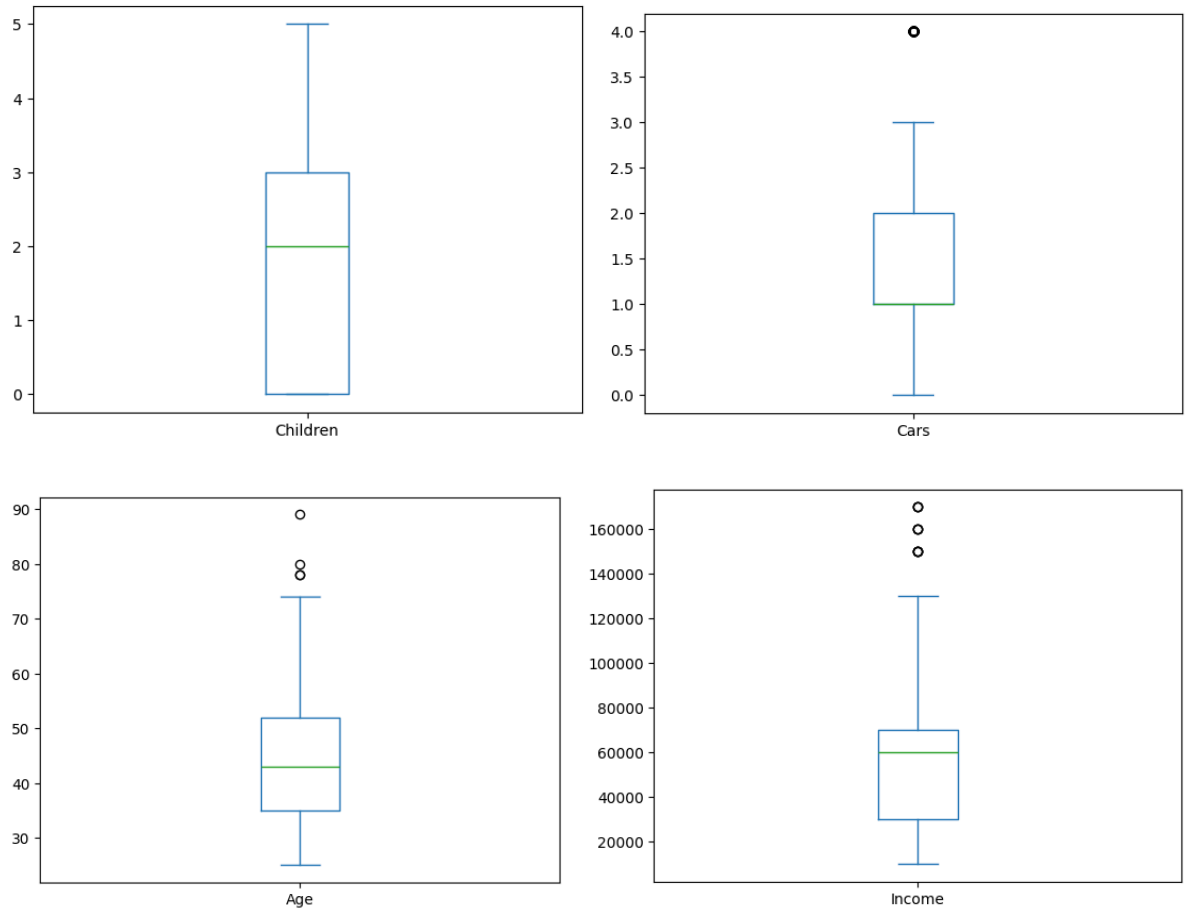
```
df.describe()
✓ 0.0s
```

	ID	Income	Children	Cars	Age
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	19965.992000	56290.000000	1.911000	1.451000	44.172000
std	5347.333948	30975.722678	1.620403	1.117519	11.316912
min	11000.000000	10000.000000	0.000000	0.000000	25.000000
25%	15290.750000	30000.000000	0.000000	1.000000	35.000000
50%	19744.000000	60000.000000	2.000000	1.000000	43.000000
75%	24470.750000	70000.000000	3.000000	2.000000	52.000000
max	29447.000000	170000.000000	5.000000	4.000000	89.000000

4.2. Distribución de Datos

Se generaron visualizaciones de la distribución de valores antes y después de la limpieza, destacando la reducción de outliers en “Cars” e “Income”.

Después de la limpieza



4.3. Anomalías detectadas y soluciones implementadas

- Los valores en "Income" tenían comas y/o espacios en blanco extras, lo que impedía el análisis numérico. Esto se corrigió eliminando las comas y convirtiendo la columna a formato numérico.

- Se encontraron y eliminaron filas duplicadas para evitar sesgos en los análisis.

- Las columnas tenían valores nulos que se modificaron al imputar medidas estadísticamente adecuadas (moda y mediana).

4.4 Fenómenos interesantes

- Outliers: los diagramas de cajas en las columnas "Cars" e “Income” mostraron la presencia de posibles valores atípicos (personas con un número de autos significativamente alto), lo cual podría ser una subpoblación interesante para estudiar más a fondo.

5. Conclusiones

El proceso de limpieza de datos del archivo csv *"Bike Buyers"* fue fundamental para asegurar la calidad y consistencia del dataset. A lo largo de este procedimiento, se corrigieron anomalías importantes, como la conversión de ingresos con formato incorrecto, la eliminación de duplicados, y la imputación de valores no nulos mediante técnicas apropiadas. También se realizaron análisis exploratorios para detectar patrones de distribución y outliers, lo cual proporcionó un mejor entendimiento de las características de los compradores de bicicletas.

Además, se observaron fenómenos interesantes, como la distribución sesgada de los ingresos y la presencia de valores atípicos en las columnas "Cars" e “Income”, que podrían ser objeto de estudios más detallados en futuras fases de análisis.

Gracias a estas acciones, se obtuvo un conjunto de datos limpio y preparado para análisis más profundos, lo que permitirá obtener insights relevantes y confiables sobre los factores que influyen en la compra de bicicletas.

6. Recomendaciones

Se sugiere un seguimiento continuo para mantener la consistencia de los datos y, de ser posible, automatizar procesos de limpieza y asegurar la recopilación precisa de los mismos en fases futuras.

7. Referencias

- GeeksforGeeks. (2018). *Different ways to create Pandas Dataframe.* GeeksforGeeks. Recuperado el 23 de octubre de 2024 del URL: https://www.geeksforgeeks.org/different-ways-to-create-pandas-dataframe/
- Tate, A. (2023). *Comprehensive Guide to Visualizing Data in Jupyter.* Hex.tech. Recuperado el 23 de octubre de 2024 del URL: https://hex.tech/blog/visualizing-data-in-jupyter/