Actividad 4 (Extracción de Características)

Andrés Salmerón García - A01731809

Programación:

1.- Lectura de datos: Utilicé la librería Pandas para cargar un archivo Excel llamado "microretailer\_mit\_lift\_lab.xlsx" en un DataFrame de Pandas. Igualmente, usé Numpy y Matplotlib para las diferentes funciones matemáticas.

2.- Limpieza de datos: Apliqué métodos para reemplazar valores nulos en el DataFrame, utilizando la media de las columnas.

3.- Eliminación de outliers: Intenté eliminar outliers utilizando tanto el método de Rango Intercuartílico como la Desviación Estándar. Sin embargo, hubo problemas con la implementación. Aquí se necesito revisar y ajustar la lógica de eliminación de outliers.

4.- Análisis univariado: Realicé un análisis univariado en columnas categóricas seleccionadas, contando las ocurrencias de cada valor y creando gráficas de barras.

5.- Generación de tablas: Utilicé groupby para realizar análisis agregados en las columnas categóricas seleccionadas y se almacenaron los resultados en un DataFrame.

Resultados:

* Obtuve un DataFrame llamado resultados que contiene análisis agregados (media, desviación estándar, recuento) para cada columna categórica seleccionada.
* Generé gráficas de barras para visualizar el conteo de valores en cada columna categórica.
* Intenté eliminar outliers, pero se encontraron problemas en la implementación que necesitan corrección.

Análisis Univariado: A través del análisis univariado de las columnas categóricas, he podido visualizar la distribución de frecuencias de diferentes categorías. Esto puede proporcionar información sobre la variabilidad en los datos y resaltar las categorías dominantes.

Análisis Agregado: Los análisis agregados que realicé (media, desviación estándar, recuento) para cada columna categórica pueden ayudar a identificar tendencias o patrones en los datos. Por ejemplo, puedo observar la media y la desviación estándar para comprender la dispersión de valores en una categoría en particular.

Reemplazo de Valores Nulos: El reemplazo de valores nulos con la media de las columnas puede ayudarme a mantener la integridad de los datos, pero también puede afectar la distribución de los mismos. Es importante considerar si este enfoque es adecuado para mi análisis y si los valores nulos tienen un significado específico en el contexto de tu estudio.

Eliminación de Outliers: Aunque intenté eliminar outliers, enfrenté problemas en la implementación. La eliminación de outliers puede influir en las estadísticas resumidas y en la interpretación de los datos. Es importante entender por qué existen outliers y cómo manejarlos adecuadamente.

Gráficas de Barras Apiladas: Las gráficas de barras apiladas pueden proporcionar información sobre la distribución de categorías a lo largo del tiempo o en diferentes grupos. La inclinación hacia la izquierda en las gráficas puede indicar una acumulación de valores en las opciones más fáciles de aplicar.

Para obtener conclusiones más sólidas y específicas, es importante conocer el contexto del conjunto de datos y el objetivo del análisis. Preguntas como "¿Qué estamos tratando de entender o lograr con estos datos?" y "¿Qué hipótesis queremos probar?" pueden guiar el proceso de análisis de datos de manera más efectiva; y aunque sí se podía interpretar como una base de datos de negocios (en especial su manera de operar), seguía sin ser del todo claro su propósito para mi. Además, debo considerar realizar análisis multivariados y modelos estadísticos para obtener insights más profundos si es necesario.