Ejercicio 2: EDA

instituto tecnológico de morelia

Ejercicios de práctica

M.I.A. WILFRIDO CORTES OROZCO

2023

|  |  |  |
| --- | --- | --- |
| Nombre del Alumno | Número de control | Número de actividad |
|  |  |  |

# Descripción y objetivos de la actividad.

Objetivos:

* Que el estudiante pueda aplicar el análisis exploratorio, elimine las anomalías y los datos nulos de manera efectiva y adecuada.
* Que el estudiante pueda visualizar los datos.
* Que el estudiante pueda responder preguntas específicas sobre su dataset.

Descripción de la actividad:

Esta actividad es continuación de la primera actividad, se requiere que el estudiante ya haya cargado e iniciado el set de datos llamado “**books\_simple\_dataset.csv**”.

En el ejercicio anterior, se contabilizaron los datos nulos, se contabilizaron las posibles variaciones de las columnas categóricas y se mencionó cuál(es) se podría(n) categorizar.

Ahora, se va a dar tratamiento a los datos nulos, se van a generar gráficas de visualización, se van a tratar las anomalías numéricas y se van a convertir a numéricas las columnas categóricas.

1. Muestre la información del dataset, muestre qué datos son numéricos y cuáles son de tipo objeto.
2. Nuevamente contabilice la cantidad de datos nulos por columna.
3. Haga un límite y muestre qué columnas superan ese límite (para el tratamiento de nulos).
4. Para las columnas que no superen el límite, elimine los datos nulos.
5. Si alguna columna sigue con datos nulos después de la eliminación, proceda a hacer la imputación usando la moda; si ninguna columna sigue con datos nulos, puede regresar al estado anterior y hacer la imputación con una columna de su agrado, principalmente para que pruebe cómo funciona (opcional).
6. Para las columnas categóricas, muestre un diagrama de distribución (tipo histograma, usando displot) y haga su análisis al respecto.
7. Para las columnas numéricas, muestre un diagrama tipo boxplot y haga su análisis al respecto.
8. Si una columna se puede categorizar, hágalo y explique el porqué se puede (o no) hacer.
9. Con las columnas que hayan sido categorizadas, use el one-hot encoder de pandas para obtener las columnas numéricas (pd.get\_dummies), luego concatene ese dataframe al dataframe original.
10. Haga un diagrama de la matriz de correlaciones y responda:
    1. ¿Qué columnas están más correlacionadas a rating?
    2. ¿Qué columnas están menos correlacionadas a rating?
    3. ¿Hay alguna correlación inversa?
11. Siguiendo el análisis de boxplot, identifique si hay datos anómalos, si los hay, haga el procedimiento adecuado para eliminarlos, use el rango intercuartílico.
12. Muestre nuevamente el boxplot y un displot para las columnas que hayan sido tratadas y haga su análisis.

En la sección de ejemplos de Teams, se agregó un documento con el ejemplo visto en clase, ahí vienen los comandos necesarios para el desarrollo de esta actividad.

Entregue sobre este documento o sobre un documento de Jupyter-Notebook (preferido), no olvide incluir sus datos como son:

1. Nombre completo.
2. Número de control.
3. Número de actividad.

Como recomendación, vaya subiendo sus modificaciones a GitHub, para ello solo necesita una cuenta y crear un repositorio, esta plataforma será utilizada para el proyecto, entonces puede irse familiarizando.