UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO
CENTRO UNIVERSITARIO UAEM ZUMPANGO
INGENIERÍA EN COMPUTACIÓN

Ciencia de los datos

Laboratorio: Visualización de datos

Profesor: Dr. Asdrúbal López Chau, Dr. Rafael Rojas Hernández

Fecha: septiembre 2024


Este símbolo significa que deberás de incluir lo que se solicita en el reporte final de la práctica.


Este símbolo significa que deberás de escribir código fuente comentado y que realice exactamente lo que se solicita.


Este símbolo significa que deberás de realizar una actividad por tu cuenta, pero puedes solicitar ayuda al profesor

Este símbolo significa un consejo, guía o ayuda que se proporciona

Nota: El plagio se penaliza con la anulación de la calificación a todos los involucrados

Objetivos

Objetivo general: Realizar una narrativa de datos eligiendo los tipos de gráficos más adecuados.

Objetivos específicos:

1. Recuperar y procesar datos utilizando Python.
2. Realizar un análisis exploratorio a los datos (conocido como EDA: *Explorative Data Analysis*)
3. Procesar los datos para identificar y dar tratamiento a casos anómalos, datos faltantes, etc.
4. Elegir los tipos de gráficos adecuados dependiendo el tipo de dato y el propósito de la narrativa.
5. Escribir una narrativa de datos interesante con el apoyo de gráficos adecuados.

Introducción

La visualización de datos es una de las herramientas más poderosas en el campo de la ciencia de los datos, permitiendo convertir conjuntos de datos complejos y abstractos en representaciones visuales comprensibles. Esta práctica facilita la comprensión, el análisis y la comunicación de patrones, tendencias y relaciones dentro de los datos. El objetivo principal es presentar la información de manera intuitiva y accesible para ayudar en la toma de decisiones basadas en datos.

En este laboratorio, exploraremos diferentes técnicas y herramientas de visualización de datos, desde gráficos básicos como histogramas y diagramas de dispersión hasta visualizaciones más avanzadas como gráficos interactivos y mapas de calor. Aprenderemos a utilizar bibliotecas populares como matplotlib y Seaborn en Python para crear visualizaciones efectivas y atractivas, así como la biblioteca pandas para la manipulación de los datos.

La visualización de datos no solo se trata de estética, sino de funcionalidad. Una visualización bien diseñada puede revelar información oculta en los datos, mientras que una mala visualización puede inducir al error. Por tanto, este laboratorio también se centrará en las mejores prácticas para diseñar visualizaciones claras, precisas y efectivas.

Requerimientos Previos

- Conocimiento las bibliotecas pandas, matplotlib, seaborn.

Instrucciones:

 Explicar para qué sirven y mostrar un ejemplo de cómo se usan las siguientes bibliotecas para crear distintos tipos de gráficos: pandas, matplotlib, seaborn

  Carga de los datos

- Carga el conjunto de datos **heart_disease_uci** y asígnalo a una variable llamada df, los datos son proporcionados con este laboratorio.
- Realiza una exploración inicial de los datos para identificar: número de variables, tipo de dato de cada variable, tamaño del conjunto de datos, estadísticas básicas.
-  En el reporte, realiza una descripción clara de los resultados de esta primera exploración.

 Realiza un EDA a los datos.

- Muestra un resumen estadístico de los datos, pero ahora separando por tipo de variable, es decir, un resumen para numéricas y otro para no numéricas
- Para las variables categóricas, indica los niveles de cada una y sus frecuencias. Busca en la documentación, el significado de cada variable.
- Usando gráficos adecuados muestra las frecuencias de las variables categóricas.
- Usando gráficos adecuados muestra la distribución de las variables numéricas.
- Revisa visualmente si existen datos anómalos en las variables numéricas, usando gráficos adecuados.
- Identifica si hay datos faltantes en los datos, o datos incorrectos, todavía no les apliques algún un tratamiento.

 En el reporte, realiza una descripción clara de los resultados de esta segunda exploración.

 Realiza un procesamiento o limpieza a los datos

- Elimina los datos faltantes en los valores, o datos incorrectos identificados anteriormente, guarda los datos resultantes en una variable llamada df_drop.
- Imputa un valor fijo a los valores faltantes en los datos, o datos incorrectos identificados anteriormente, guarda los datos resultantes en una variable llamada df_fill. Elige un valor fijo adecuado, explica en el reporte porqué usaste ese valor.

 Realiza una narrativa de los datos

- Escribe una narrativa apoyada por gráficos necesarios, cuidadosamente elegidos, para convencer, y mantener el interés la explicación de los datos.

  Aquí algunas ideas sobre la narrativa.

- Usa la edad de las personas que aparecen en los datos, indica edad promedio, mínima y máximas, de manera general y por género, luego explica cómo se distribuye el tipo de dolor de pecho con respecto al género y edad.
- Muestra si hay relaciones entre los tipos de dolor de pecho, colesterol, frecuencia cardiaca, edad, género, etc.
- Relaciona los defectos (reversibles o no reversibles, con la presión arterial, colesterol, etc.)

Rúbrica
Considera estos para tu reporte

1. Claridad y Concisión del Mensaje (25%)
- **Objetivo:** ¿La visualización transmite claramente el mensaje principal?
- **Simplicidad:** ¿Es la visualización lo suficientemente simple para ser entendida rápidamente?
- **Evita el exceso de información:** ¿Se han eliminado los elementos visuales innecesarios?

2. Elección del Tipo de Gráfico (25%)
- **Adecuación:** ¿El tipo de gráfico elegido es el más apropiado para los datos y el mensaje?
- **Variedad:** ¿Se han utilizado diferentes tipos de gráficos cuando es necesario para mostrar múltiples aspectos de los datos?
- **Evita la redundancia:** ¿Se han evitado gráficos que repiten la misma información?

3. Diseño Estético (20%)
- **Colores:** ¿La paleta de colores es coherente y facilita la distinción entre las diferentes categorías?
- **Etiquetas:** ¿Las etiquetas son claras y fáciles de leer?
- **Formato:** ¿El tamaño de las fuentes, los espaciados y la disposición general son adecuados?

4. Corrección de los Datos (10%)
- **Precisión:** ¿Los datos se representan correctamente?
- **Escalas:** ¿Las escalas de los ejes son apropiadas para los datos?
- **Cálculos:** ¿Los cálculos utilizados para generar la visualización son correctos?

5. Interactividad (10%)
- **Relevancia:** ¿La interactividad agrega valor a la visualización?
- **Facilidad de uso:** ¿La interacción es intuitiva y fácil de controlar?
- **Funcionalidad:** ¿Las opciones de interacción permiten explorar los datos de manera efectiva?

6. Código (10%)
- **Legibilidad:** ¿El código es bien estructurado y fácil de entender?
- **Eficiencia:** ¿El código es eficiente y se ejecuta rápidamente?
- **Comentarios:** ¿El código está bien comentado para explicar las decisiones de diseño?

Calificación:
- **Excelente:** Cumple con todos los criterios de manera sobresaliente.
- **Bueno:** Cumple con la mayoría de los criterios de manera satisfactoria.
- **Regular:** Cumple con algunos de los criterios, pero hay áreas de mejora.
- **Pobre:** No cumple con la mayoría de los criterios.