Guía de Análisis Exploratorio.
Proyecto

INTRODUCCIÓN

Para hacer una investigación formal es necesario basarse en una situación problemática y por consiguiente un problema que justifique la investigación. Es importante revisar la teoría que rodea la problemática y los antecedentes de investigaciones similares. La investigación puede ser aplicada a diversos temas incluyendo finanzas, economía y negocios.

El Instituto Nacional de Estadística (INE) tiene numerosas Bases de Datos. Algunas pueden usarse para explorar cómo se está comportando la sociedad guatemalteca, por ejemplo, las bases de datos de estadísticas vitales y las de violencia. En la de estadísticas vitales, se pueden encontrar 5 conjuntos de datos por año desde 2009 hasta 2022 (https://www.ine.gob.gt/ine/vitales/):

- Nacimientos.
- Matrimonios.
- Divorcios.
- Defunciones.
- Defunciones Fetales

En la de violencia, hay 5 conjuntos de datos, pero particularmente 3 resultan interesantes:

- Hechos delictivos
- Violencia intrafamiliar
- Violencia en contra de la mujer y delitos sexuales.

Deben trabajar con uno o varios de los conjuntos de datos antes mencionados. Tengan en cuenta que deben trabajar con más de 10 años de datos, así que es posible que tengan que hacer transformaciones para unir los archivos de cada año. Tenga cuidado con los datos que se encuentran en la página del INE, debe trabajar con los datos crudos, generalmente en formato .sav.

El objetivo principal es explorar los datos para obtener preguntas interesantes. Si tienen acceso a otros conjuntos de datos que creen puedan servirles, son libres de utilizarlos, respetando siempre las condiciones de quien los publica.

ACTIVIDADES

1. Exploren los datos para encontrar preguntas interesantes de investigación. Para esto:
    a. Describan el conjunto de datos: cuantas variables y observaciones hay y el tipo de cada una de las variables.
    b. Realicen un resumen de las variables numéricas e investiguen si siguen una distribución normal, si no es así expliquen la distribución que pueden presentar. Para las variables categóricas obtengan una tabla de frecuencia, documenten lo que vayan encontrando.

   c. Crucen las variables que consideren sean las más importantes para hallar los elementos clave que permitan comprender lo que está causando el problema encontrado.
   d. Formulen al menos 5 preguntas de investigación basadas en supuestos, creencias o hipótesis preliminares que tengan sobre el fenómeno estudiado. Estas preguntas deben reflejar ideas que a priori podrían considerarse verdaderas (por ejemplo, que ciertos hechos delictivos se concentran en zonas específicas, que existen patrones temporales marcados, o que determinadas variables están relacionadas, etc.). Cada supuesto deberá ser validado o refutado mediante análisis de datos, utilizando código, tablas y gráficos. Se espera que discutan explícitamente si los resultados confirman o contradicen la creencia inicial. Se valorará especialmente la capacidad de cuestionar intuiciones iniciales y de permitir que los datos guíen las conclusiones, incluso cuando estas contradigan expectativas previas.
   e. Realicen gráficos exploratorios que les dé ideas del estado de los datos.
   f. Hagan un agrupamiento "clustering" e interpreten los resultados.
2. Una vez hayan explorado los datos
   a. Describan la situación problemática que los lleva a plantear un problema a resolver.
   b. Enuncien un problema científico y unos objetivos preliminares.
   c. Describan los datos que tienen para responder el problema planteado. Esto incluye el estado en que se encontró el o los conjuntos de datos y las operaciones de limpieza que realizaron, en caso de que hayan sido necesarias.
   d. Escriban unas conclusiones con los hallazgos encontrados durante el análisis exploratorio

EVALUACIÓN

Notas: Para tener derecho a calificación deben mostrar evidencias de contribuciones significativas tanto en el repositorio como en el documento. El uso de paquetes y módulos para hacer Análisis Exploratorio de forma automática, no está permitido.

- **(10 puntos) Situación Problemática**: Describe la situación problemática que da lugar al problema.
- **(10 puntos). Problema científico:** Enuncia el problema científico que se desprende de la situación planteada. Comprende bien cuál es el problema.
- **(10 puntos). Objetivos:** Plantea los objetivos a cumplir para darle solución al problema planteado. Enuncia al menos un objetivo general y 2 específicos. Los objetivos deben ser medibles y alcanzables durante la investigación.
- **(15 puntos). Descripción de los datos:** Describe los datos, tanto las variables y observaciones como las operaciones de limpieza que se hicieron si fueron necesarias.
- **(35 puntos). Análisis Exploratorio:**
  - Estudian las variables cuantitativas mediante técnicas de estadística descriptiva
  - Presentan gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión, que ayudan a explicar los datos.

  - Determina la distribución de las variables numéricas.
  - Analizan las correlaciones entre las variables, tratan de explicar los datos atípicos "outliers" y toman decisiones acertadas ante la presencia de valores faltantes.
  - Estudian las variables categóricas.
  - Elaboran gráficos de barra, tablas de frecuencia y de proporciones
  - Explican muy bien todos los procedimientos y los hallazgos que van haciendo.
  - Plantea preguntas interesantes basadas en supuestos, creencias e hipótesis que le permiten explorar los datos más a fondo. Se discuten los hallazgos, si se cumplen o se rechazan esas hipótesis.
  - Determina la tendencia al agrupamiento y el mejor número de "clusters" a utilizar.
  - Hace el agrupamiento con cualquiera de los algoritmos estudiados.
  - Verifica la calidad del agrupamiento, incluya el método de la silueta.
  - Interpreta los grupos, usando para eso las variables numéricas y categóricas dentro de cada grupo.
- **(20 puntos). Hallazgos y conclusiones:**
  - Realiza un resumen de los hallazgos en el análisis exploratorio
  - Le pone un nombre a los grupos que reflejen sus características principales
  - Presenta un plan de los siguientes pasos a seguir.

MATERIAL A ENTREGAR

- Vínculo de Google docs con el informe de análisis exploratorio. Se debe poder verificar el historial de cambios
- Informe de análisis exploratorio formal en formato .pdf (sin código)
- Script de R (.r o .rmd) o de Python que utilizaron para responder las preguntas.
- Vínculo del repositorio de github que utilizaron