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 2021 (https://www.ine.gob.gt/ine/vitales/):

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

En la de violencia, hay 5 conjuntos de datos que resultan interesantes también, 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. 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 y guías 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 y, 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. Realicen gráficos exploratorios que les dé ideas del estado de los datos.
   e. 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.

- **(10 puntos) Situación Problemática**: Describen la situación problemática que da lugar al problema.
- **(10 puntos). Problema científico:** Enuncian el problema científico que se desprende de la situación planteada. Comprenden bien cuál es el problema.
- **(10 puntos). Objetivos:** Plantean los objetivos a cumplir para darle solución al problema planteado. Enuncian al menos un objetivo general y 2 específicos. Los objetivos deben ser medibles y alcanzables durante la investigación.
- **(20 puntos). Descripción de los datos:** Describen los datos, tanto las variables y observaciones como las operaciones de limpieza que se hicieron si fueron necesarias.
- **(30 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.
  - 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.
  - Determinan la tendencia al agrupamiento y el mejor número de "clusters" a utilizar.
  - Hacen el agrupamiento con cualquiera de los algoritmos estudiados.
  - Verifican la calidad del agrupamiento, incluyan el método de la silueta.
  - Interpretan los grupos, usando para eso las variables numéricas y categóricas dentro de cada grupo.
- **(20 puntos). Hallazgos y conclusiones:**
  - Realizan un resumen de los hallazgos en el análisis exploratorio
  - Le ponen un nombre a los grupos que reflejen sus características principales

- Presentan 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
- Script de R (.r o .rmd) o de Python que utilizaron para responder las preguntas.
- Vínculo del repositorio de github que utilizaron