Guía de Análisis Exploratorio.
Proyecto

INTRODUCCIÓN:

Para hacer una investigación formal es necesario que esta se base en una situación problemática y por consiguiente un problema que justifique la investigación. 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 tiene numerosas Bases de Datos, pero una en particular puede usarse para explorar como se está comportando la sociedad guatemalteca. Se trata de la base de datos de estadísticas vitales. Podremos encontrar 4 conjuntos de datos por año desde 2009 hasta 2019 (https://www.ine.gob.gt/ine/vitales/):

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

Debe trabajar con uno o varios de los conjuntos de datos antes mencionado. Tenga en cuenta que debe trabajar con los 10 años así que es posible que tenga que hacer transformaciones para unir los archivos de cada año. El objetivo principal es explorar los datos para obtener preguntas interesantes. Si tiene acceso a otros conjuntos de datos que cree que puedan servirle, es libre de utilizarlos, respetando siempre las condiciones de quien los publica.

ACTIVIDADES

1. Explore los datos para encontrar preguntas interesantes y guías de investigación. Para esto:
    a. Comience describiendo cuantas variables y observaciones tiene disponibles, el tipo de cada una de las variables.
    b. Haga un resumen de las variables numéricas e investigue si siguen una distribución normal y tablas de frecuencia para las variables categóricas, escriba lo que vaya encontrando.
    c. Cruce las variables que considere que son las más importantes para hallar los elementos clave que lo pueden llevar a comprender lo que está causando el problema encontrado.
    d. Haga gráficos exploratorios que le de ideas del estado de los datos.
    e. Haga un agrupamiento (clustering) e interprete los resultados.
2. Una vez que haya explorado los datos
    a. Describa la situación problemática que lo lleva a acotar un problema a resolver.
    b. Enuncie un problema científico y unos objetivos preliminares.
    c. Describa los datos que tiene para responder el problema planteado. Esto incluye el estado en que encontró el o los conjuntos de datos y las operaciones de limpieza que le realizó, en caso de que hayan sido necesarias.

d. Escriba unas conclusiones con los hallazgos encontrados durante el análisis exploratorio

EVALUACIÓN

- **(10 puntos) Situación Problemática**: Describe la situación problemática que da lugar al problema.
- **(10 puntos). Problema científico:** Se enuncia el problema científico que se desprende de la situación planteada. Se comprende bien cuál es el problema.
- **(10 puntos). Objetivos:** Se plantean los objetivos a cumplir para darle solución al problema planteado. Se enuncia al menos un objetivo general y 2 específicos. Los objetivos deben ser medibles y alcanzables durante la investigación.
- **(20 puntos). Descripción de los datos:** Se describen los datos, tanto las variables y observaciones como las operaciones de limpieza que se le hicieron si fueron necesarias.
- **(30 puntos). Análisis Exploratorio:**
  - Estudia las variables cuantitativas mediante técnicas de estadística descriptiva
  - Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión, que ayudan a explicar los datos.
  - Analiza las correlaciones entre las variables, trata de explicar los outliers (puntos atípicos) y toma decisiones acertadas ante la presencia de valores faltantes.
  - Estudia las variables categóricas.
  - Elabora gráficos de barra, tablas de frecuencia y de proporciones
  - Explica muy bien todos los procedimientos y los hallazgos que va haciendo.
  - Determina el mejor número de clusters a utilizar.
  - Hace el agrupamiento con cualquiera de los algoritmos estudiados.
  - Verifica la calidad del agrupamiento usando 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:**
  - Hace un resumen de los hallazgos en el análisis exploratorio
  - Le pone un nombre a los grupos que reflejen sus características principales
  - Llega a conclusiones sobre 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 utilizó para responder las preguntas con el código utilizado o archivo de flujo de trabajo de KNime.
- Vínculo de repositorio de github