



Asignatura: Estadística y probabilidad 1

Facultad: Ingeniería y Ciencias Básicas

Núcleo: Matemáticas y Estadística

Proyecto 1. **Análisis Exploratorio de Datos**

1. Objetivo

En este proyecto, se aplicarán los conceptos de análisis descriptivo para analizar un conjunto de datos real. El objetivo es identificar patrones, tendencias y obtener insights significativos a partir de los datos.

2. Conjunto de Datos

Cada estudiante es responsable de:

1. Seleccionar una muestra aleatoria de 500 estudiantes de un estado en el siguiente enlace:

<https://ww2.amstat.org/censusatschool/RandomSampleForm.cfm>

Nota: No escoger alguno de los siguientes estados: Canal Zone, Hawaii, Louisiana, Maine, Mississippi, Montana, Nevada, North and South Dakota, Puerto Rico, Virgin Islands, Wyoming.

2. Descargar el cuestionario en el siguiente enlace:

<http://ww2.amstat.org/censusatschool/pdfs/C@SQuestionnaire.pdf>

A partir del cuestionario revisa los siguientes conceptos donde se explica el origen y relevancia de los datos para el análisis:

- a) Unidad de estudio.
 - b) Población objeto de estudio.
 - c) Muestra de estudio.
 - d) Variable de análisis.
 - e) Tipo de variable y su escala de medición.
3. Depurar la base de datos de datos anómalos (error de entrada de datos) o faltantes.

3. Pasos del Proyecto

Conformar grupos de 4 estudiantes y unir los datos de los 4 estados. A partir de esta base de datos con 2000 muestras, realizar lo siguiente:

3.1. Primer paso: Preparación y Exploración de Datos

- Carga el conjunto de datos en Python o R.
- Inspecciona las primeras filas, dimensiones y tipos de variables.
- Identifica las variables clave para el análisis descriptivo.

3.2. Segundo paso: Análisis Descriptivo, Visualización e Interpretación

Responder a las siguientes preguntas:

1. ¿Cuántos estudiantes son diestros, zurdos o ambidiestros?
2. ¿Cuál es el mes donde nacieron más estudiantes de los que participan en el Censo escolar en cada estado?
3. ¿Cuál es la bebida favorita de los estudiantes que participan en el Censo escolar en cada estado?
4. ¿Cuál es el deporte / actividad favorita de los estudiantes que participan en el Censo escolar en cada estado?
5. ¿Cuál es el nivel más alto de educación que planean alcanzar los estudiantes que participan en el Censo escolar en cada estado?
6. Construya un histograma con la variable “tiempo que trabaja con pago el estudiante en la semana (medido en horas)” y a partir de ésta gráfica responda las siguientes preguntas:
 - a) ¿Qué porcentaje de los estudiantes de la muestra trabajan máximo 13 horas a la semana con pago?
 - b) ¿Qué porcentaje de los estudiantes de la muestra trabajan mínimo 3 horas a la semana con pago?
 - c) ¿El 20 % de los estudiantes trabajan mínimo cuantas horas con pago?
 - d) ¿El 20 % de los estudiantes trabajan máximo cuantas horas con pago?
 - e) ¿En cuál intervalo se halla la moda?
 - f) ¿El 50 % de los estudiantes trabajan máximo cuantas horas con pago?
 - g) ¿En promedio cuantas horas trabajan con pago a la semana los estudiantes? ¿La desviación estándar?

- h) ¿Qué tipo de asimetría presenta esta variable?
7. Construya una Ojiva para la variable “Tiempo que dedican los estudiantes para ayudar a los quehaceres del Hogar” y a partir de esta gráfica responda las siguientes preguntas:
- a) ¿Qué porcentaje de los estudiantes de la muestra dedican mínimo 15 horas a la semana para ayudar en sus casas?
 - b) ¿Qué porcentaje de los estudiantes de la muestra dedican máximo 5 horas a la semana para ayudar en sus casas?
 - c) ¿Cuál es el máximo número de horas que el 15 % de los estudiantes ayudan en sus casas?
 - d) ¿Qué porcentaje de los estudiantes ayudan en sus casas al menos 5 horas a la semana?
 - e) ¿Cuál es el mínimo número de horas que el 35 % de los estudiantes ayudan en sus casas?
 - f) ¿En cuál intervalo se halla la moda?
 - g) ¿En promedio cuántas horas ayudan en sus casas en la semana los estudiantes? ¿Cuál es la desviación estándar?
 - h) ¿Qué tipo de asimetría presenta esta variable?
8. Construya una gráfica de Diagrama de cajas para la comparar la opinión en los diferentes Estados de hombres y mujeres (Gender) con respecto a la importancia de reducir la contaminación (pollution) a partir de ésta grafica responda las siguientes preguntas:
- a) ¿En términos generales para quienes (hombres o mujeres) es más importante reducir la contaminación?
 - b) ¿En términos generales en cuál Estado es más importante reducir la contaminación?
 - c) ¿En términos generales en cuál Estado es más importante para las mujeres reducir la contaminación?
 - d) ¿En términos generales en cuál Estado es más importante para los hombres reducir la contaminación?
 - e) ¿En términos generales en qué género la importancia de reducir la contaminación es más homogénea?
 - f) ¿En términos generales en cuál Estado la importancia de reducir la contaminación es más heterogénea?
 - g) ¿En cuál género se presenta mayor asimetría en las calificaciones de la importancia de reducir la contaminación?

- h)* ¿En cuál Estado se presenta mayor asimetría en las calificaciones de la importancia de reducir la contaminación?
- i)* ¿En cuál Estado las mujeres presentan mayor simetría en las calificaciones de la importancia de reducir la contaminación?
9. Construya una tabla cruzada de indicadores entre la variable “¿Cuántas horas de sueño por noche suele dormir cuando no tiene clases al día siguiente?” versus “Género”. A partir de esta tabla responda las siguientes preguntas:
- a)* ¿Qué porcentaje de estudiantes de la muestra son mujeres?
- b)* ¿Cuántas horas de sueño en promedio suelen dormir cuando no tiene clases al día siguiente los estudiantes de la muestra?
- c)* ¿En términos generales que género duermen más tiempo?
- d)* ¿En términos generales qué género los tiempos son más homogéneos?
- e)* ¿En términos generales qué género los tiempos son más simétricos?
- f)* ¿Cuál género tiene un histograma más puntiagudo? ¿Cuál género tiene un histograma más aplanado o chato?
10. Construya una tabla cruzada entre la variable “Materia preferida en el Colegio” versus la variable “Género”. A partir de ésta tabla responda las siguientes preguntas:
- a)* ¿A cuántos estudiantes hombres les gusta la historia?
- b)* ¿Qué porcentaje de estudiantes son hombres?
- c)* ¿Qué porcentaje de estudiantes prefieren música?
- d)* ¿De las mujeres, qué porcentaje de estudiantes prefieren otras actividades?
- e)* ¿De los estudiantes que prefieren Ciencias, qué porcentaje son hombres?
- f)* ¿Qué porcentaje son mujeres y prefieren otras actividades?
- g)* ¿Qué porcentaje son mujeres o prefieren música?
- h)* ¿De las mujeres, qué porcentaje le gusta la música o las Ciencias?

3.3. Tercer paso: Entrega del Proyecto

- Elaborar (y compartir) en, por ejemplo “Colab”, el documento completo que les permitió escribir y ejecutar el código donde incluya las respuestas a las preguntas anteriores.
- Elaborar (y compartir) un informe final, en “PDF”, que resuma los hallazgos que considere más significativos y presente los resultados de manera clara y coherente.