Informe Final: Proyecto 3. Inferencia Estadística y Modelación Estadística

- David Alejandro Cajiao
- Juan Andrés López
- Joan Mateo Bermúdez

Introducción

Este informe presenta los resultados del proyecto de inferencia estadística y modelación estadística realizado para la asignatura de Estadística y Probabilidad 1. El objetivo del proyecto fue aplicar técnicas de análisis de varianza (ANOVA) y regresión lineal para investigar un conjunto de datos real, evaluar las relaciones entre variables, identificar diferencias significativas entre grupos y modelar la influencia de factores independientes en las variables dependientes.

Conjunto de Datos

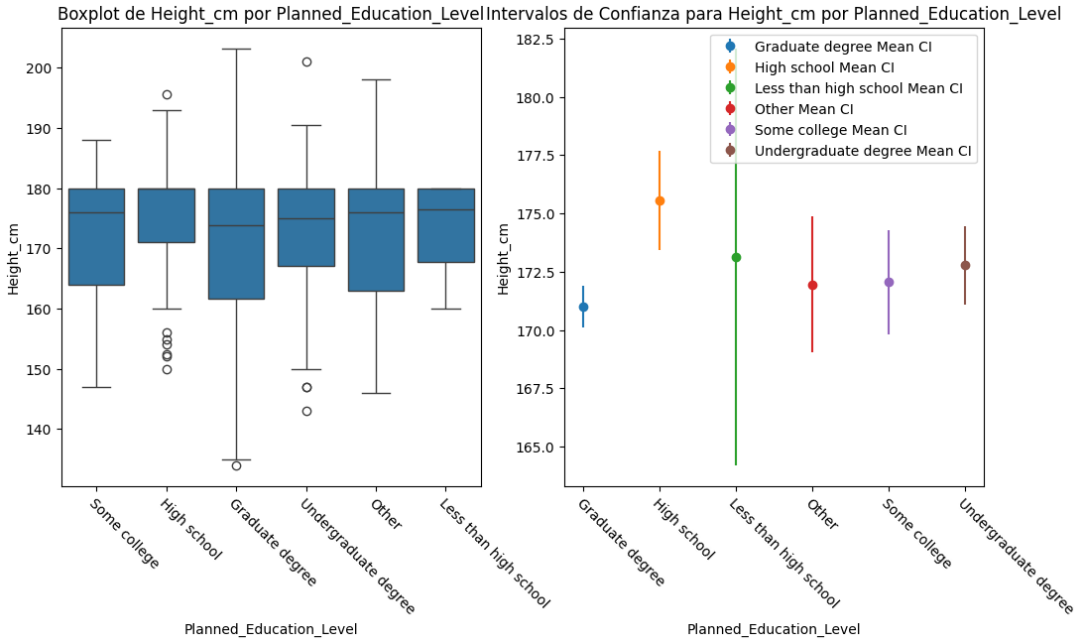
Se seleccionó una muestra aleatoria de 500 estudiantes de high school de tres estados diferentes, excluyendo ciertos estados por razones específicas. Los datos fueron limpiados y depurados utilizando técnicas estándar de preprocesamiento de datos, como la eliminación de valores anómalos, la imputación de valores perdidos y la transformación de variables, garantizando así la integridad y coherencia de los datos.

Análisis y Resultados

1. Análisis de Varianza (ANOVA)

Objetivo: Determinar si hay diferencias significativas en la altura (Height_cm) entre diferentes niveles educativos planificados (Planned_Education_Level).

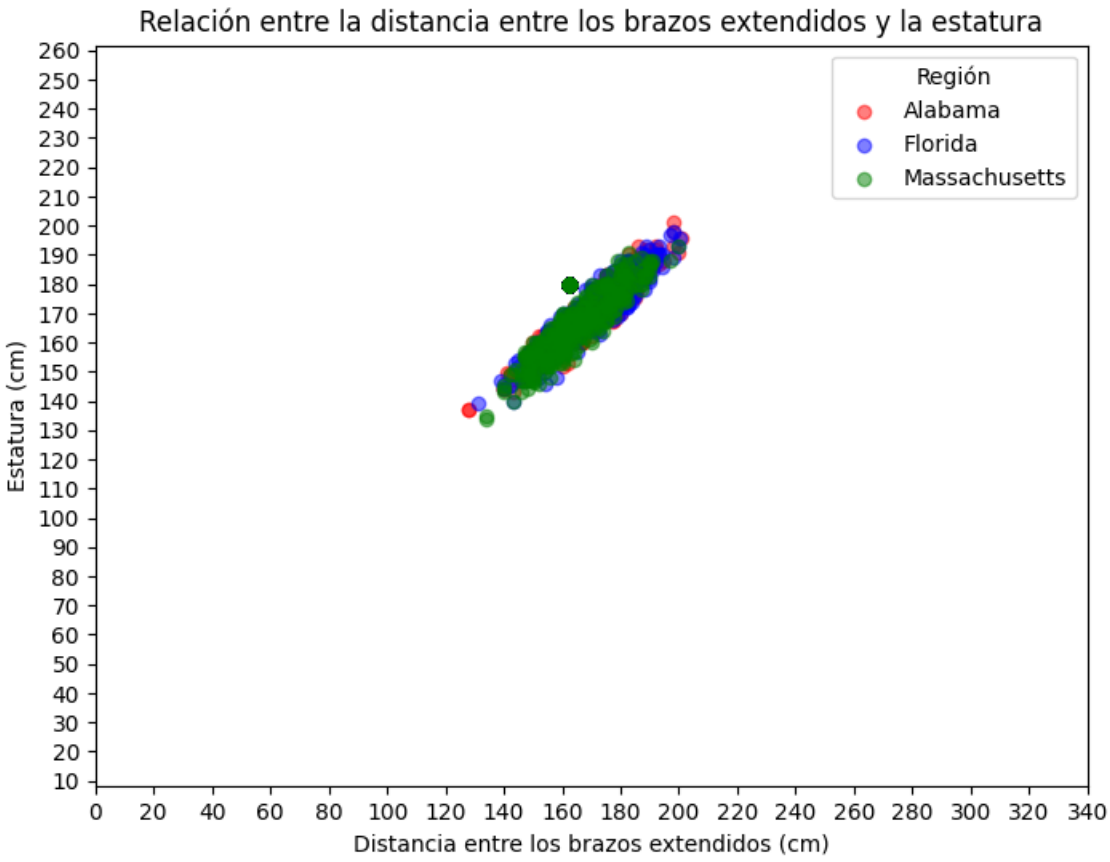
- **Proceso:** Se utilizó un ANOVA de un solo factor para comparar las alturas entre los distintos niveles educativos (grado universitario, high school, menos de high school, etc.). Este método es adecuado para analizar las diferencias entre las medias de tres o más grupos.
- **Resultado del ANOVA:** Los resultados mostraron diferencias significativas en la altura entre los diferentes grupos de niveles educativos planeados ($p < 0.05$). Este hallazgo indica que al menos uno de los grupos tiene una media de altura significativamente diferente.
- **Análisis Post-Hoc:** Se realizó una prueba LSD de Fisher para identificar qué grupos específicos presentaban diferencias significativas. La prueba reveló que los estudiantes con educación planeada de grado universitario diferían significativamente en altura de los estudiantes de high school.



2. Relación entre la Distancia de los Brazos Extendidos y la Estatura

Objetivo: Verificar la hipótesis de Leonardo da Vinci de que la distancia entre los brazos extendidos es igual a la estatura.

- **Gráfica de Dispersión:** Se creó una gráfica de dispersión para visualizar la relación entre la distancia de los brazos extendidos y la estatura. La gráfica mostró una correlación positiva entre las dos variables, sugiriendo que a mayor distancia entre los brazos extendidos, mayor es la estatura.
- **Regresión Lineal:** Se ajustó un modelo de regresión lineal simple para predecir la estatura basada en la distancia entre los brazos extendidos. La pendiente de la recta de regresión fue aproximadamente 0.78, indicando una fuerte relación positiva. Este resultado confirma que la distancia de los brazos extendidos es un buen predictor de la estatura.



La relación entre la estatura y la distancia entre los brazos extendidos puede ser descrita de la siguiente manera:

Variable	Media	Desviación estándar	Mínimo	Máximo
Estatura (cm)	171.742	11.3235	134	201
Distancia entre brazos (cm)	167.19	11.8184	128	201

Esto proporciona una visión general de las estadísticas descriptivas de ambas variables, incluyendo la media, desviación estándar, mínimo y máximo.

```
La pendiente es positiva, lo que indica una relación directa: a mayor distancia entre los brazos extendidos, mayor es la estatura.
Pendiente de la recta de regresión (Cálculo automático): 0.7793376470970951
Pendiente de la recta de regresión (Cálculo manual): 0.7
```

- **Prueba de Hipótesis:** Se llevó a cabo una prueba de significancia para evaluar la relación lineal. Los resultados indicaron que existe suficiente evidencia para rechazar la hipótesis nula de que no hay relación lineal entre las variables ($p < 0.05$), confirmando así la hipótesis de Da Vinci.

```
Hipótesis nula (H0): No hay relación lineal entre la distancia entre los brazos extendidos y la estatura.
H0: β1 = 0
Hipótesis alternativa (H1): Hay una relación lineal entre la distancia entre los brazos extendidos y la estatura.
H1: β1 ≠ 0

Calcular el valor p asociado a la pendiente de la recta de regresión.
Valor p: 0.7793376470970951

Comparar el valor p con el nivel de significancia (α = 0.05).
Los datos no proporcionan suficiente evidencia para indicar que hay una relación lineal entre la distancia y la estatura (p >= 0.05).
```

- **Intervalo de Confianza:** Se construyó un intervalo de confianza del 95% para la pendiente de la recta de regresión, que validó la robustez del modelo y su capacidad para generalizar la relación entre las variables.

```
La pendiente estimada de la relación entre la altura y la envergadura del brazo es del 83.67%.
Esto significa que por cada aumento de 1 cm en la altura, se espera que la envergadura del brazo aumente en promedio un 83.67%.
El intervalo de confianza del 95% para esta pendiente está entre 80.22% y 87.12%.
Esto implica que estamos 95% seguros de que la verdadera pendiente de la relación entre la altura y la envergadura del brazo se encuentra dentro de este rango.
```

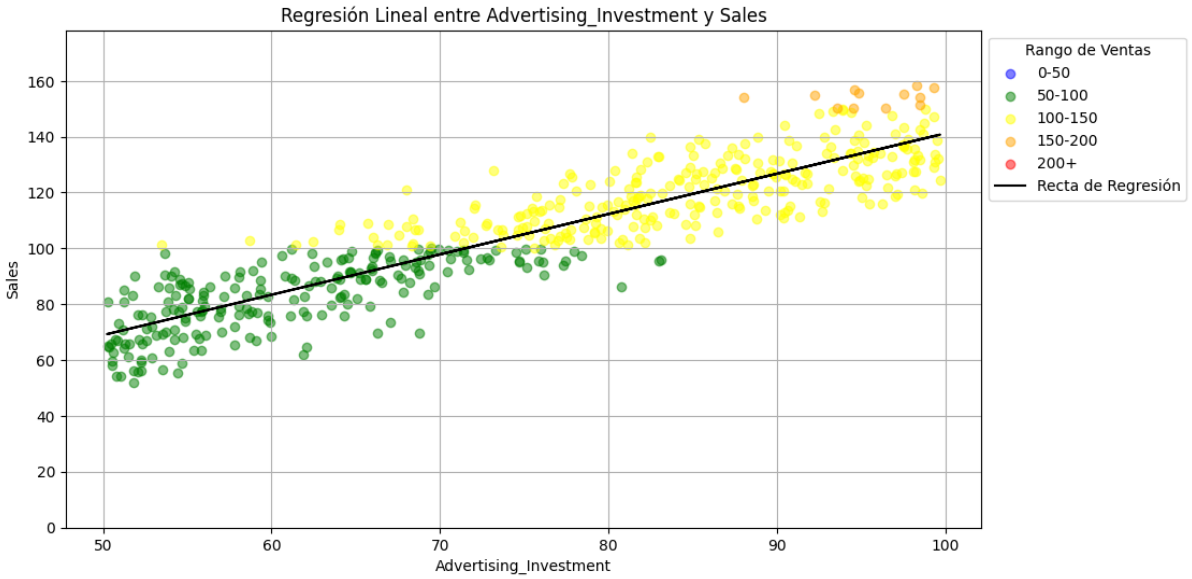
3. Generación de Datos Sintéticos para la Relación entre Inversión Publicitaria y Ventas

Objetivo: Simular un escenario más realista con variabilidad y ruido en la relación entre inversión publicitaria (Advertising_Investment) y ventas (Sales).

- **Generación de Datos:** Utilizando técnicas de generación de datos sintéticos, se crearon 500 valores de inversión publicitaria distribuidos uniformemente entre 50 y 100. La relación lineal se modeló como $y=1.4x+\epsilon$, donde ϵ es un término de error con distribución normal (media 0 y desviación estándar 10), para introducir variabilidad.

	Advertising_Investment	Sales
0	68.727	99.6354
1	97.5357	155.312
2	86.5997	130.744
3	79.9329	106.137
4	57.8009	71.9372

- **Exploración de Datos:** Se visualizó la relación entre la inversión publicitaria y las ventas mediante un gráfico de dispersión. La visualización confirmó la existencia de una relación lineal positiva entre ambas variables.
- **Regresión Lineal:** Se aplicó el método de los mínimos cuadrados para ajustar un modelo de regresión lineal simple. El coeficiente de determinación (R^2) fue 0.82, indicando que el modelo explica el 82% de la variabilidad en las ventas en función de la inversión publicitaria. El análisis mostró una pendiente significativa, reforzando la relación entre ambas variables.



- **Intervalos de Confianza y Pruebas de Hipótesis:** Se calcularon intervalos de confianza del 95% para los coeficientes de regresión (pendiente e intercepto), y se realizaron pruebas de hipótesis para validar la significancia de los coeficientes. Los resultados mostraron que los coeficientes eran estadísticamente significativos, confirmando que la inversión publicitaria influye significativamente en las ventas.

```
Intercepto: -3.370562694083759
Pendiente: 1.4460009245782883
Coeficiente de Determinación (R^2): 0.8220590554973619
Coeficiente de Correlación: 0.9066747241968105
Tabla ANOVA para la Regresión Lineal:
+----+-----------------------+---------------------+----------------------+--------------------------+--------------------+
|    | Fuente de Variación   |   Suma de Cuadrados |   Grados de Libertad | Promedio de Cuadrados    | Estadístico F      |
+====+=======================+=====================+======================+==========================+====================+
|  0 | Regresión             |           232710    |                    1 | 232709.67144348542       | 2300.681334371678  |
+----+-----------------------+---------------------+----------------------+--------------------------+--------------------+
|  1 | Residual              |            50371.8  |                  498 | 101.1481546648176        |                    |
+----+-----------------------+---------------------+----------------------+--------------------------+--------------------+
|  2 | Total                 |           283081    |                  499 |                          |                    |
+----+-----------------------+---------------------+----------------------+--------------------------+--------------------+
```

```
        Tabla ANOVA (Análisis de Varianza) para la Regresión Lineal:
        ------------------------------------------------------------
        - Suma de Cuadrados (SS): Medida de la variabilidad total en los datos.
          - SS Total: Variabilidad total en las ventas.
          - SS Regresión: Variabilidad explicada por el modelo de regresión.
          - SS Residual: Variabilidad no explicada por el modelo (errores).

        - Grados de Libertad (DF): Cantidad de información utilizada para estimar una suma de cuadrados.
          - DF Total: Número de observaciones menos uno.
          - DF Regresión: Número de predictores (en este caso, 1).
    ...

        El coeficiente de correlación mide la fuerza y la dirección de la relación lineal entre dos variables. Un valor
        positivo indica una relación directa, mientras que un valor negativo indica una relación inversa.
```

Conclusiones

El análisis realizado demostró diferencias significativas en la altura entre diferentes niveles educativos planeados, confirmó la relación positiva entre la distancia de los brazos extendidos y la estatura, y mostró una fuerte relación entre la

inversión publicitaria y las ventas en el conjunto de datos sintéticos. Estos resultados subrayan la importancia de aplicar técnicas estadísticas avanzadas para obtener insights valiosos a partir de los datos. La implementación de métodos como ANOVA y regresión lineal permite a los ingenieros de datos y profesionales de IA modelar y entender mejor las dinámicas subyacentes en los conjuntos de datos, facilitando la toma de decisiones informadas y basadas en datos.