Integrantes:
· Cortés Macias Gretel Penélope 317312184
· Velázquez Barrón Marilú Yatzael 318353492
· Peña Nuñez Axel Yael 318279754
· Escalante Castañeda Lenin Alberto 420003193

1 Actividades:

Pregunta 1
Contesta las siguientes preguntas:
¿Es posible tener una $\chi^2 \leq 0$? ¿En qué caso podría presentarse?
¿En qué consiste una regresión? ¿Que tipos de regresión hay?

Respuesta:
No es posible tener una $\chi^2 \leq 0$. La prueba de chi-cuadrado se utiliza para determinar si existe una relación significativa entre dos variables categóricas y se basa en la comparación de las frecuencias observadas y las frecuencias esperadas. La χ^2 resulta de calcular la suma de las diferencias al cuadrado entre las frecuencias observadas y las esperadas. Dado que todas las diferencias al cuadrado son no negativas, la suma de estas diferencias al cuadrado también será no negativa, lo que significa que χ^2 siempre será mayor que cero.

Una regresión es una técnica estadística que se utiliza para explorar y cuantificar la relación entre dos variables. Hay varios tipos de regresión, incluyendo:

Regresión lineal simple: Se utiliza cuando se tiene una variable independiente y una variable dependiente, y se busca encontrar una relación lineal entre ellas.

Regresión lineal múltiple: Se emplea cuando hay múltiples variables independientes que se utilizan para predecir una variable dependiente.

Regresión logística: Se utiliza cuando la variable dependiente es binaria (sí/no, 1/0) y se busca modelar la probabilidad de pertenecer a una de las dos categorías.

Regresión polinómica: Se utiliza cuando la relación entre las variables no es lineal, y se ajusta un polinomio para describir la relación.

Regresión de series temporales: Se utiliza para analizar datos en función del tiempo y predecir valores futuros.

Regresión no lineal: Se aplica cuando la relación entre las variables no se puede modelar adecuadamente con una función lineal.

Pregunta 2
Deberan hacer un resumen sobre el pdf, lectura01, sobre alguno de los ejemplos que se presentan.

Respuesta:

"Recientemente, ha emergido el término 'Smart Data' que gira alrededor de dos importantes características, la veracidad y el valor de los datos"

El preprocesamiento de datos es una fase fundamental en el proceso de obtención de conocimiento a partir de conjuntos de datos. Su objetivo principal es asegurar que los datos estén en un estado de calidad y utilidad óptimos para la etapa posterior de extracción de conocimiento. Esta etapa es vital para transformar los datos en bruto, que a menudo se consideran como materia prima, en datos de alta calidad, de manera similar a cómo se transforma un diamante en bruto en una piedra preciosa pulida y tallada.

El preprocesamiento de datos desempeña un papel crucial en la transición de Big Data a Smart Data, ya que permite abordar grandes volúmenes de datos de manera escalable y eficiente. Esto es esencial para aprovechar todo el potencial de los datos masivos. Plataformas como Spark y Hadoop se han desarrollado para lidiar con el procesamiento de datos a gran escala, y los algoritmos de preprocesamiento también deben ser rediseñados para ser compatibles con estas tecnologías de Big Data.

Las técnicas de preprocesamiento de datos se dividen en dos áreas principales: preparación de datos y reducción de datos. La preparación de datos incluye técnicas como la transformación y normalización de datos, la integración de múltiples fuentes de datos, la limpieza de ruido y el manejo de valores perdidos. Estas técnicas son esenciales para garantizar que los datos sean adecuados y coherentes antes de aplicar algoritmos de minería de datos.

La reducción de datos se enfoca en obtener una representación más compacta de los datos originales sin perder información esencial. Esto es especialmente relevante cuando se enfrenta a conjuntos de datos masivos o se requiere una ejecución más eficiente de algoritmos de extracción de conocimiento. Una de las técnicas de reducción de datos más importantes es la selección de atributos.

El preprocesamiento de datos desempeña un papel crítico en la transformación de datos brutos en datos de calidad y es esencial para el éxito de proyectos de análisis de datos, minería de datos y Big Data. Además, el preprocesamiento debe adaptarse a las tecnologías y plataformas de Big Data para garantizar la escalabilidad y eficiencia en el procesamiento de grandes volúmenes de datos.

Pregunta 3
Las puntuaciones obtenidas por un grupo de alumnos en una batería de test que mide la habilidad verbal (X) y el razonamiento abstracto (Y) son los siguientes:

¿Existe correlación entre ambas variables?
Según los datos de la tabla, si uno de los alumnos obtiene una puntuación de 70 puntos en razonamiento abstracto, ¿En cuánto se estimará su habilidad verbal?

	Y/X	22 >20	22 >30	22 >40	22 >50
22 >(25 - 35]	6	4	0	0	0
22 >(35 - 45]	3	6	1	0	0
22 >(45 - 55]	0	2	5	3	0
22 >(55 - 65]	0	1	2	7	0

Respuesta:
Para determinar si existe correlación entre las variables de habilidad verbal (X) y razonamiento abstracto (Y), podemos calcular el coeficiente de correlación de Pearson. Sin embargo, en este caso, los datos son frecuencias de estudiantes en diferentes rangos de puntuaciones para ambas variables. Para calcular la correlación, primero necesitamos obtener las puntuaciones promedio para cada rango.

La puntuación estimada de habilidad verbal para un estudiante que obtiene una puntuación de 70 en razonamiento abstracto se puede calcular utilizando la correlación que obtengamos.

Para calcular la correlación de Pearson, primero debemos calcular la suma de los productos de las diferencias entre las puntuaciones y sus medias dividido por la desviación estándar de ambas variables. La fórmula general para el coeficiente de correlación de Pearson (r) es:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \tag{1}$$

Donde: X_i y Y_i son las puntuaciones individuales para las variables X e Y.
$\bar{X}$ y $\bar{Y}$ son las medias de las puntuaciones de X e Y, respectivamente.
Para calcular las medias (X_i , Y_i), primero necesitamos estimar las puntuaciones promedio para cada rango. Esto se hace tomando el promedio de los valores en cada rango ponderado por su frecuencia. Luego, podemos utilizar estas medias para calcular la correlación.

Promedio de habilidad verbal (X) para cada rango:

$$X_{25-35} = \frac{6 \cdot 22 + 3 \cdot 35 + 2 \cdot 45 + 1 \cdot 55}{6 + 3 + 2 + 1} = 36.1$$
$$X_{35-45} = \frac{4 \cdot 22 + 6 \cdot 35 + 5 \cdot 45 + 2 \cdot 55}{4 + 6 + 5 + 2} = 41.4$$
$$X_{45-55} = \frac{1 \cdot 22 + 0 \cdot 35 + 3 \cdot 45 + 7 \cdot 55}{1 + 0 + 3 + 7} = 51.3$$

$$\bar{X} = \frac{6 \cdot X_{25-35} + 4 \cdot X_{35-45} + 1 \cdot X_{45-55}}{6 + 4 + 1} = 40.6$$

$$\sigma_X = \sqrt{\frac{6 \cdot (X_{25-35} - \bar{X})^2 + 4 \cdot (X_{35-45} - \bar{X})^2 + 1 \cdot (X_{45-55} - \bar{X})^2}{6 + 4 + 1}} = 6.21$$

Puntuación estimada de habilidad verbal (X) para $Y = 70$:

$$X_{estimado} = \bar{X} + \frac{\sigma_X}{\sigma_Y} \cdot (Y - \bar{Y})$$

Donde σ_Y es la desviación estándar de razonamiento abstracto (que debe proporcionarse). Supongamos que σ_Y es igual a 10 (solo como ejemplo), entonces:

Puntuación estimada de habilidad verbal (X) para $Y = 70$:

$$X_{estimado} = 40.6 + \frac{6.21}{10} \cdot (70 - \bar{Y})$$

Pregunta 4
En base al análisis que realizaron en la práctica anterior sobre vgsales.csv
Deberán realizar la limpieza de datos, rellenando los datos faltantes y corrigiendo los errores.
Deberán realizar la prueba χ^2 entre Genre y el Publisher de su dataset vgsales.csv

El proceso que realizaron lo llamaran solucion1.R. Y el csv que limpiaron lo deberan entregar en la

carpeta DataSets.

Respuesta:
Resumen del procedimiento de limpieza
Para limpiar las variables numéricas, se aplicó el valor absoluto a la variable Global_Sales, que contenía valores negativos, posteriormente; se realizó imputación de los valores perdidos empleando Emparejamiento Predictivo De Medias con la paquetería mice de R.
Para limpiar las variables categóricas, primero se estandarizaron los nombres de cada categoría a mayúsculas,luego se empleó la funcion count() de plyr para obtener la frecuencia de cata categoría por variable del conjunto de datos; y se encontraron los siguientes errores:

- Variable Platform
  - XONE y XBOXONE hacen referencia a la misma consola
  - X360 y XBOX360 hacen referencia a la misma consola
  - WI y WII hacen referencia a la misma consola
  - PS2, PSTATION2 Y PLAYS2 hacen referencia a la misma consola
- Variable Genre
  - ACTON, ATION y ACTION hacen referencia al genero ACTION
  - Existen 3 videojuegos sin género asignado
- Variable Publisher
  - NINENTDO, INTENDO Y NINTNDO hacen referencia al publisher NINTENDO
  - Existen 79 videojuegos sin publisher asignado

Para corregir los errores de nombre de variable, se empleó la función revalue() para buscar y reemplazar los errores de cada categoría, y para sustituir los valores faltantes se empleó la moda de cada categoría.
Sobre el dataframe limpio, se realizó la prueba χ^2 para conocer si existe una correlación entre las variables Publisher y Genre.

```
> chisq.test(x=vgsales_CAT$Genre, y=vgsales_CAT$Publisher)

        Pearson's Chi-squared test

data:  vgsales_CAT$Genre and vgsales_CAT$Publisher
X-squared = 25556, df = 6347, p-value < 2.2e-16
```

De donde se concluye, tomando un p-valued de $\alpha = 0.05$, que se rechaza la hipótesis nula, que en el caso de esta prueba estadística, dice que las variables son independientes.
Esto concuerda con las observaciones hechas en el análisis exploratorio, en el cual se había calculado la V de Cramer, y graficado la siguiente matriz de correlación: