UNIVERSIDAD DE BUENOS AIRES

BIG DATA & MACHINE LEARNING

Tercer Trabajo Práctico

Grupo #27

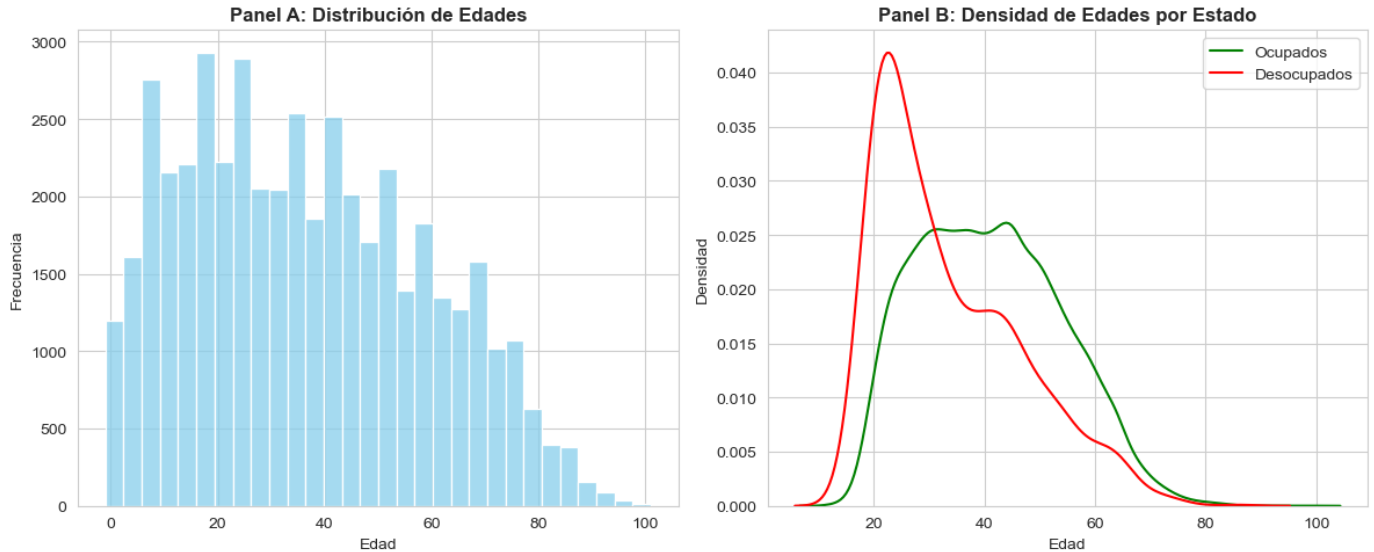
Integrantes:

Ramiro Sanchez

Francisco Barrial

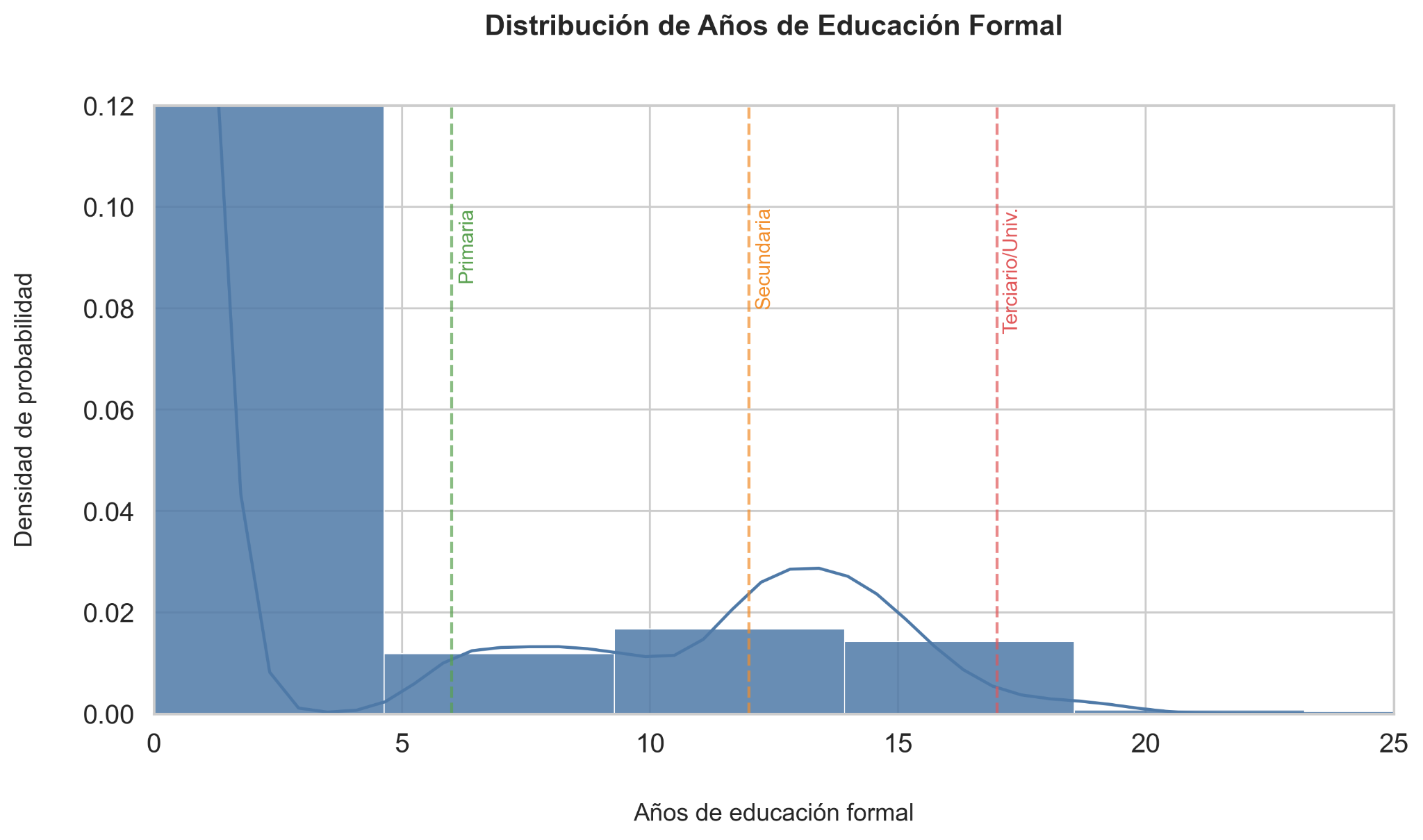
Parte 1

1. En el primer punto de este trabajo, fue solicitado crear gráficos sobre la distribución de edad y de ocupación/desocupación de nuestra región seleccionada.

Para esto, unificamos las bases de los archivos de datos de 2004 y 2024 con los siguientes criterios, para la variable de edad, se usó “CH06” que hacía referencia a la misma y ya estaba estandarizada. La variable “Estado” (ocupación/desocupación), la redefinimos como categórica siendo el número 1 para ocupados y el 0 para desocupados. Por último, creamos la variable edad2 y obtuvimos los siguientes gráficos.

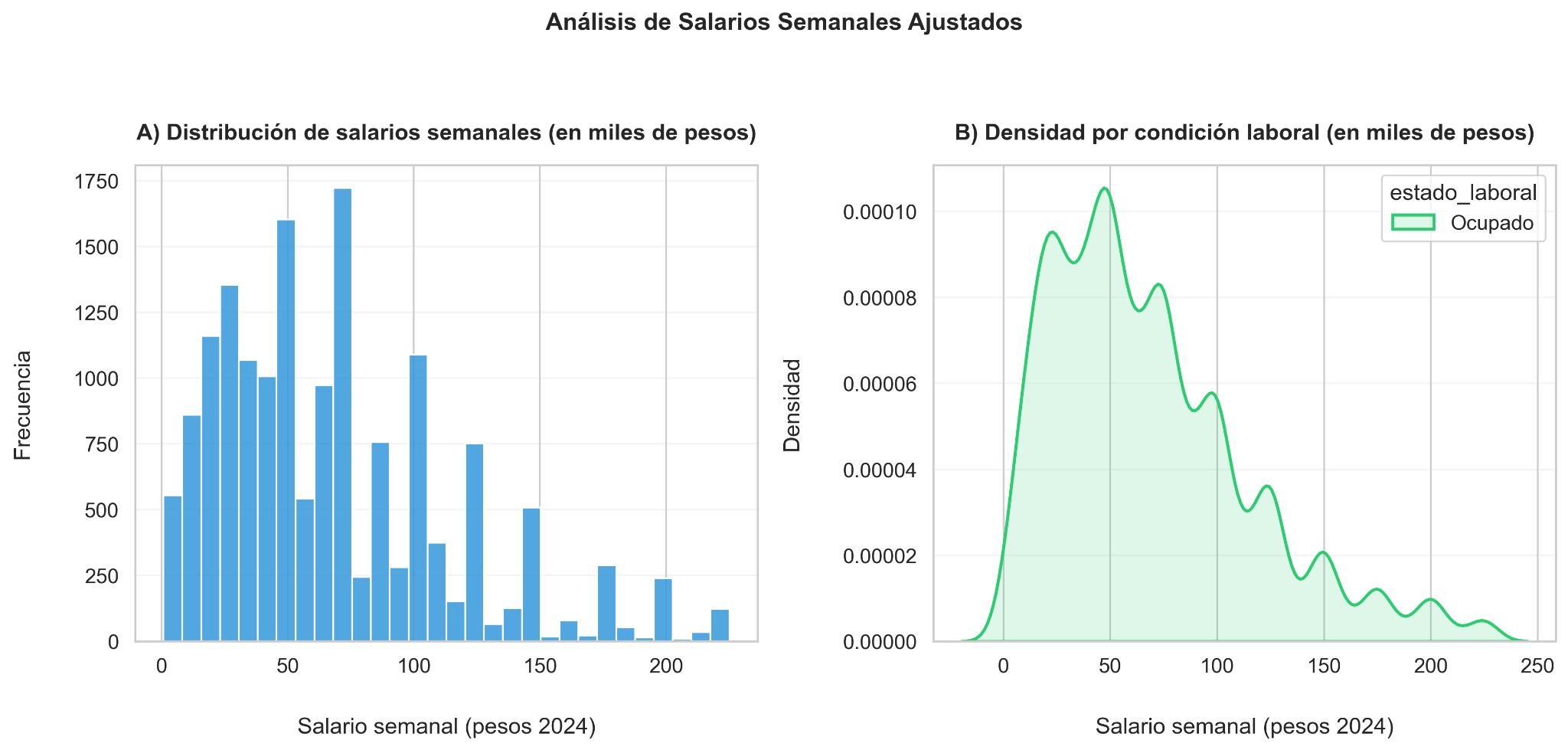
Como se puede ver, la edad más frecuente son los 20 años con unos casi 3000 registros, quienes a su vez rondan la edad con mayor nivel de desocupación, representando un 4% de la población total. Observamos que la curva de ocupación contiene una meseta en el rango de edad de 30-40 años y alcanza su pico máximo rondando los 45 con algo mas del 2.5% de la población. Por otra parte en el gráfico de edad podemos ver que si bien hay una sólida cantidad de registros de edades mayores o iguales a 80, no es en conjunto, para nada envejecida.

1. Para el segundo punto, hubo que confeccionar un gráfico sobre la cantidad de años de educación formal recibida y procedimos transformamos las variables CH12 (nivel educativo), CH13 (finalización) y CH14 (años aprobados) en años totales de educación, asignando: 6 años por primaria completa, 12 años por secundaria completa y 17+ años por educación superior y obtuvimos el siguiente gráfico.



El gráfico muestra una distribución cuyos picos están marcados en los 6, 12 y 17 años, correspondientes a la finalización de primaria, secundaria y educación superior. Se observa una mayor concentración en el nivel secundario (a los 12 años), seguido por un segundo grupo significativo en educación superior (17+ años), lo que sugiere que esta población tiene una tendencia creciente hacia la profesionalización. La esperable disminución abrupta después de los 20 años refleja la menor proporción de población con posgrados.

1. Para el tercer punto analizamos los salarios, primero convertimos los ingresos a pesos de 2024 usando factores de inflación y calculé el salario semanal dividimos los ingresos entre 40 horas. Filtramos valores extremos para evitar outliers (percentil 99) e hicimos los siguientes gráficos:



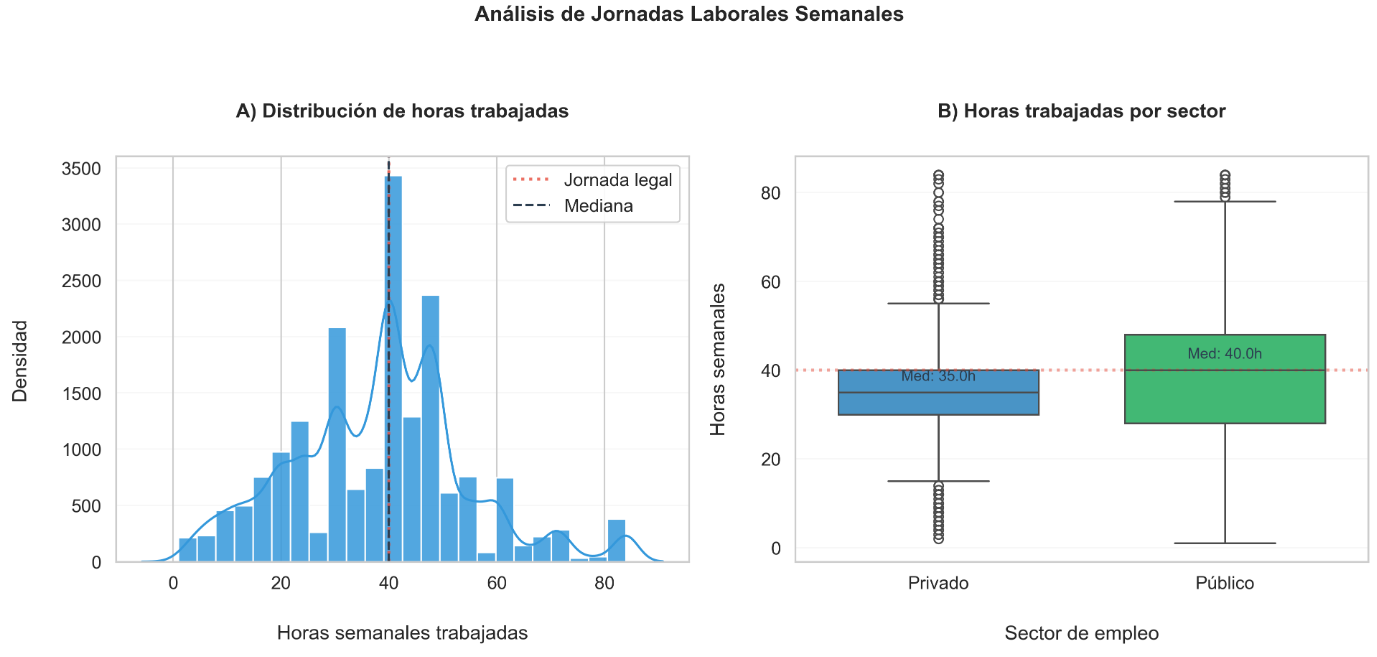
El gráfico revela que la mayoría de las personas encuestadas gana menos de $5000 por semana, con muy pocos casos que superan esta cifra. Quienes tienen trabajo suelen percibir ingresos medios entre $30,000 y $80,000 semanales, mientras que aquellos sin empleo aparecen concentrados cerca de $0, como es de esperar.

Se observan pequeños picos en montos exactos (como $20,000 o $50,000), lo que podría indicar que muchas personas redondean sus respuestas al dar información sobre sus ingresos.

1. Para el cuarto punto, calculamos el total de horas sumando las dedicadas a la ocupación principal y otras ocupaciones. Limitamos el máximo a 84 horas (12h/día × 7 días) para evitar distorsiones. Filtramos solo personas ocupadas con horas positivas para evitar incoherencias e hicimos dos paneles, el primero muestra la concentración alrededor de la jornada estándar de 40 horas (línea roja), con un segundo pico cerca de 45-50 horas

y agregamos un segundo Panel para mostrar el promedio de horas trabajadas en los sectores público y privado

Estos mismos son los siguientes:



La distribución de horas trabajadas muestra que la mayor cantidad de los ocupados trabaja entre 35 y 45 horas semanales, aunque una gran porción supera las 45 horas. En definitiva, la mayoría de la gente trabaja más de 40 horas. Por otra parte en otro gráfico podemos ver cómo en promedio trabajan más horas los empleados del sector público que del privado, probablemente dado por cuestiones de informalidad en el mismo.

1. Tamaño de la base:

**Resumen de la base final para la región 17**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **2004** | **2024** | **Total** |
| Cantidad de observaciones | 0 | 1160 | 1160 |
| Cantidad de observaciones con Nas en la variable “Estado” | 0 | 0 | 0 |
| Cantidad de Ocupados | 0 | 559 | 559 |
| Cantidad de Desocupados | 0 | 40 | 40 |
| Cantidad de variables limpias y homogeneizadas | 31 | 31 | 31 |

Parte 2

1)

Gráfico, Gráfico de rectángulos

El contenido generado por IA puede ser incorrecto.

Este gráfico muestra una matriz de correlaciones entre variables sociodemográficas y laborales de Neuquén. Los colores indican la fuerza y dirección de la relación: rojo para correlaciones positivas y azul para negativas. Por ejemplo, CH06 y edad2 tienen una correlación muy alta (0.96), mientras que educ no se correlaciona fuertemente con las otras variables. Además, salario\_semanal y horastrab tienen una correlación moderada (0.52), lo que sugiere que a mayor cantidad de horas trabajadas, mayor es el salario semanal.

2)

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Este gráfico muestra los resultados de un Análisis de Componentes Principales (PCA) para Neuquén, graficando el primer componente principal contra el segundo. Cada punto representa un individuo o muestra, proyectado en el nuevo espacio reducido de dos dimensiones. El objetivo es visualizar la dispersión y agrupamiento de los datos, facilitando la identificación de patrones o tendencias. Se observa que la mayoría de los datos se concentran cerca del origen, con algunos valores atípicos dispersos.

3) Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Este gráfico combina los resultados del PCA mostrando tanto los scores (puntos azules) como los loadings (flechas rojas) para Neuquén. Los puntos representan las observaciones proyectadas en los dos primeros componentes principales. Las flechas indican la dirección y el peso de cada variable original en el nuevo espacio, mostrando cuáles variables influyen más en cada componente. Por ejemplo, CH06 y edad2 están alineadas con el primer componente, mientras que salario\_semanal y horastrab lo están con el segundo. Esto ayuda a interpretar qué variables explican la mayor variabilidad en los datos.

4) Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

La variable mostrada, varianza\_explicada, indica el porcentaje de la variabilidad total de los datos que explica cada componente principal en el análisis PCA. Cada valor representa la importancia de un componente para resumir la información original. Por ejemplo, si el primer componente tiene un valor alto, significa que captura gran parte de la estructura de los datos. Esto permite reducir la dimensionalidad manteniendo la mayor cantidad de información posible.

5) Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Este gráfico muestra el resultado de un agrupamiento K-means con k=2, utilizando las variables CH06 y edad2. Cada punto representa una observación, coloreada según el grupo al que fue asignada por el algoritmo. El objetivo es identificar patrones o diferencias entre dos grupos dentro de los datos. Se observa una clara separación en la distribución, lo que sugiere que existen dos agrupaciones principales basadas en estas variables.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

En este caso, los colores de los puntos representan el estado de cada observación según la variable "ESTADO". Así, el gráfico permite visualizar cómo se distribuyen los diferentes estados (por ejemplo, azul para uno y rojo para otro) en el espacio definido por CH06 y edad2, facilitando la identificación de patrones o diferencias entre los grupos según su estado.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

En este gráfico se muestra el resultado del agrupamiento K-means con k=10 usando las variables CH06 y edad2. Cada color representa uno de los diez grupos identificados por el algoritmo. Esto permite observar una segmentación más detallada de los datos, revelando subgrupos o patrones más específicos en la relación entre CH06 y edad2 dentro de la muestra analizada.

B) Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Este gráfico muestra los resultados de un K-means con k=2, donde el eje X representa la edad y el eje Y el cluster asignado. Los puntos están coloreados según la ocupación: azul para ocupados y rojo para desocupados. Se observa que el algoritmo separa a los individuos en dos grupos principalmente según la edad, y dentro de cada grupo hay tanto ocupados como desocupados, aunque con diferente proporción.

6) Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

Un dendograma es un diagrama en forma de árbol que se utiliza para representar la agrupación jerárquica de elementos. Muestra cómo se van uniendo los datos en grupos (clústeres) a diferentes niveles de similitud o distancia. Cada rama representa una fusión de grupos, permitiendo visualizar la estructura y relaciones entre los datos. Es muy útil para identificar cuántos grupos naturales existen en un conjunto de datos.