UNIVERSIDAD DE BUENOS AIRES

BIG DATA & MACHINE LEARNING

Tercer Trabajo Práctico

Grupo #27

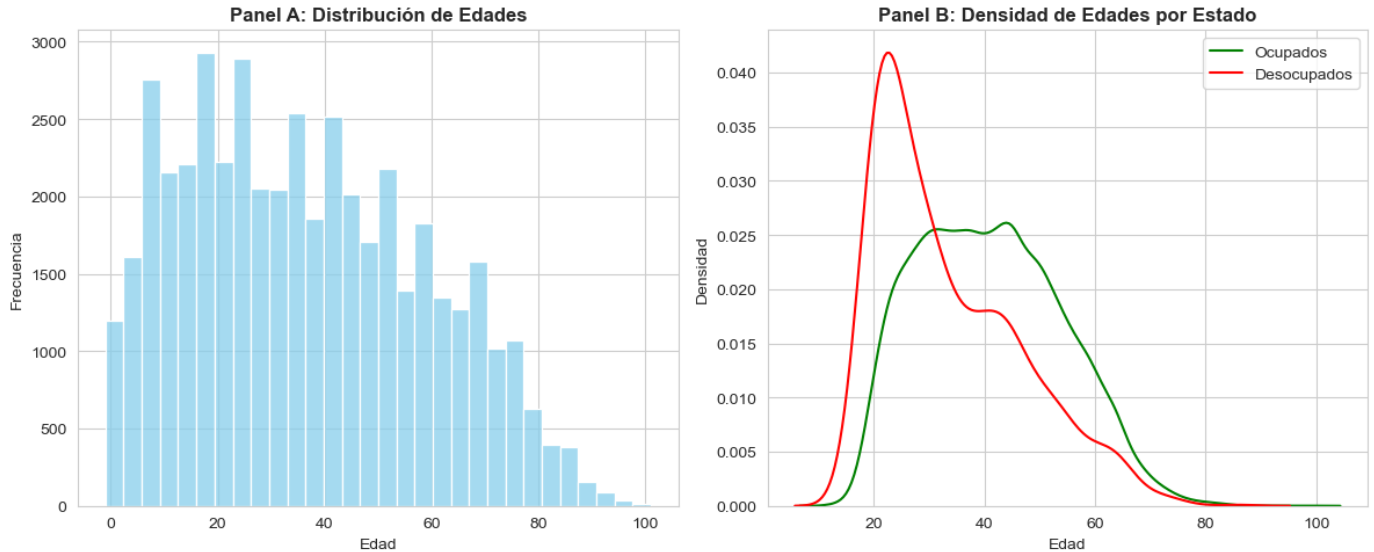
Integrantes:

Ramiro Sanchez

Francisco Barrial

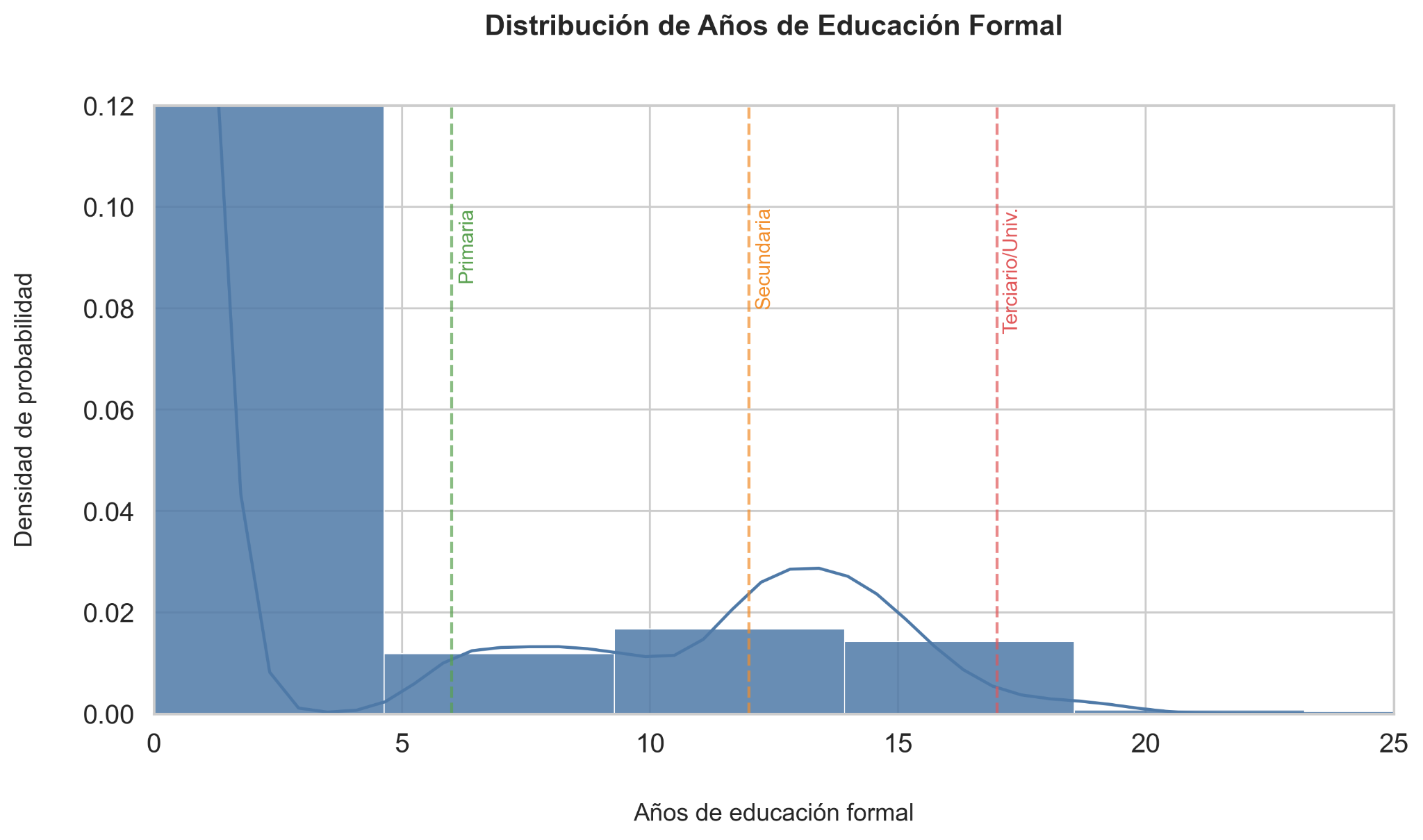
Parte 1

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.

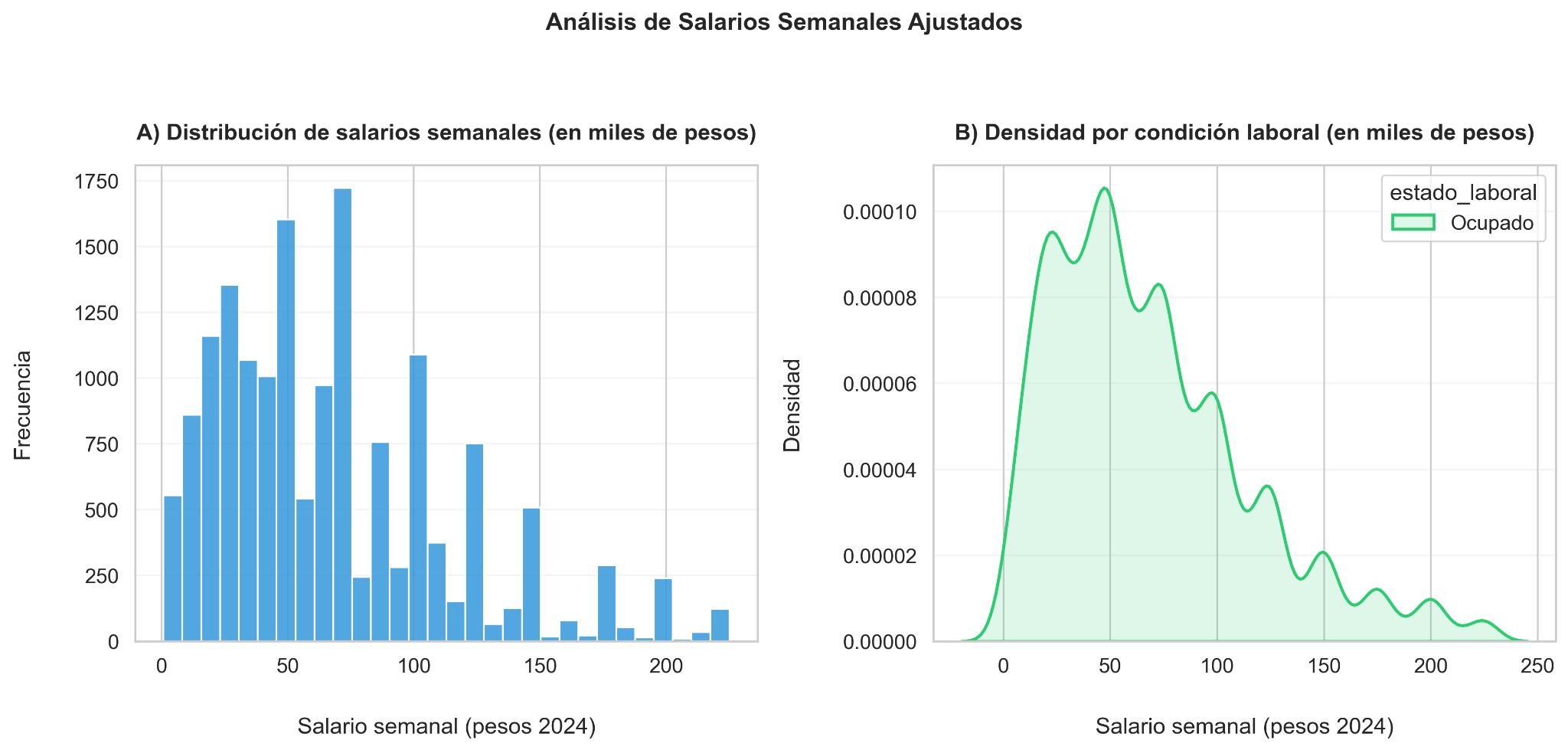
Como se puede ver, la edad más frecuente son los 20 años con unos casi 3000 registros, quienes a su vez rondan la edad con mayor nivel de desocupación, representando un 4% de la población total. Observamos que la curva de ocupación contiene una meseta en el rango de edad de 30-40 años y alcanza su pico 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:



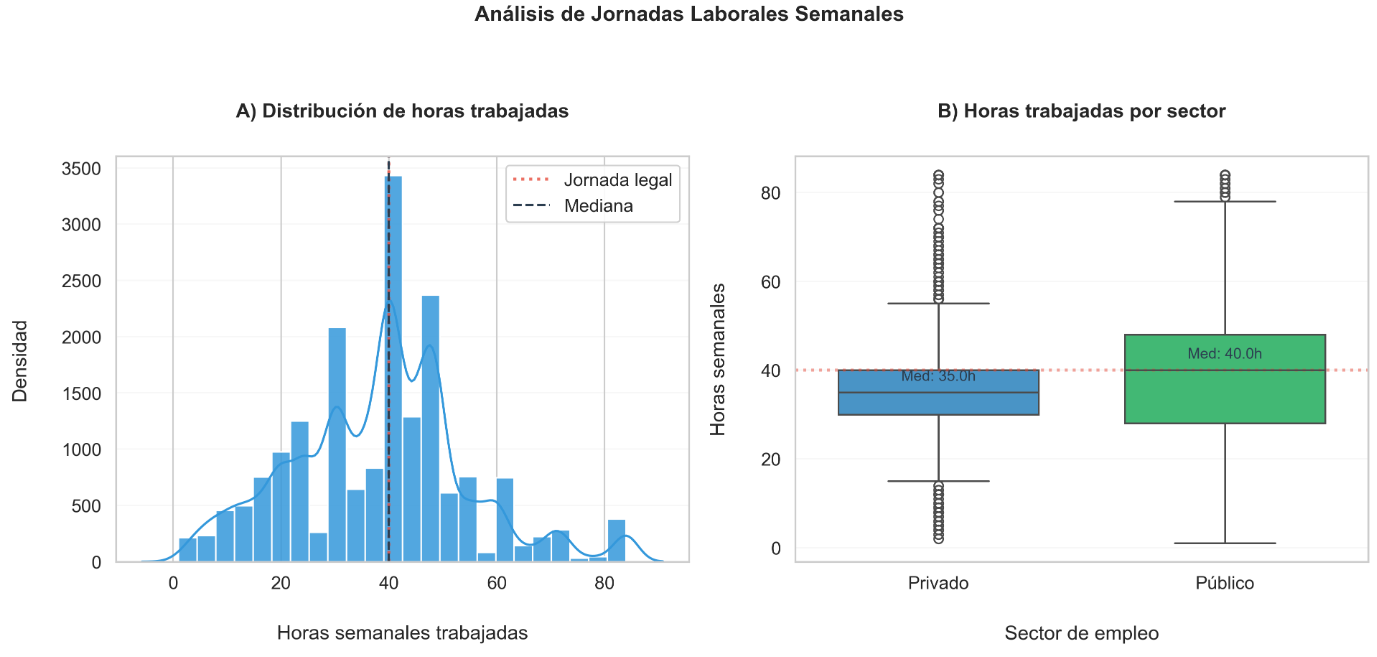
El gráfico revela que la mayoría de las personas encuestadas gana menos de $5000 por semana, con muy pocos casos que superan esta cifra. Quienes tienen trabajo suelen percibir ingresos medios entre $30,000 y $80,000 semanales, mientras que aquellos sin empleo aparecen concentrados cerca de $0, como es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