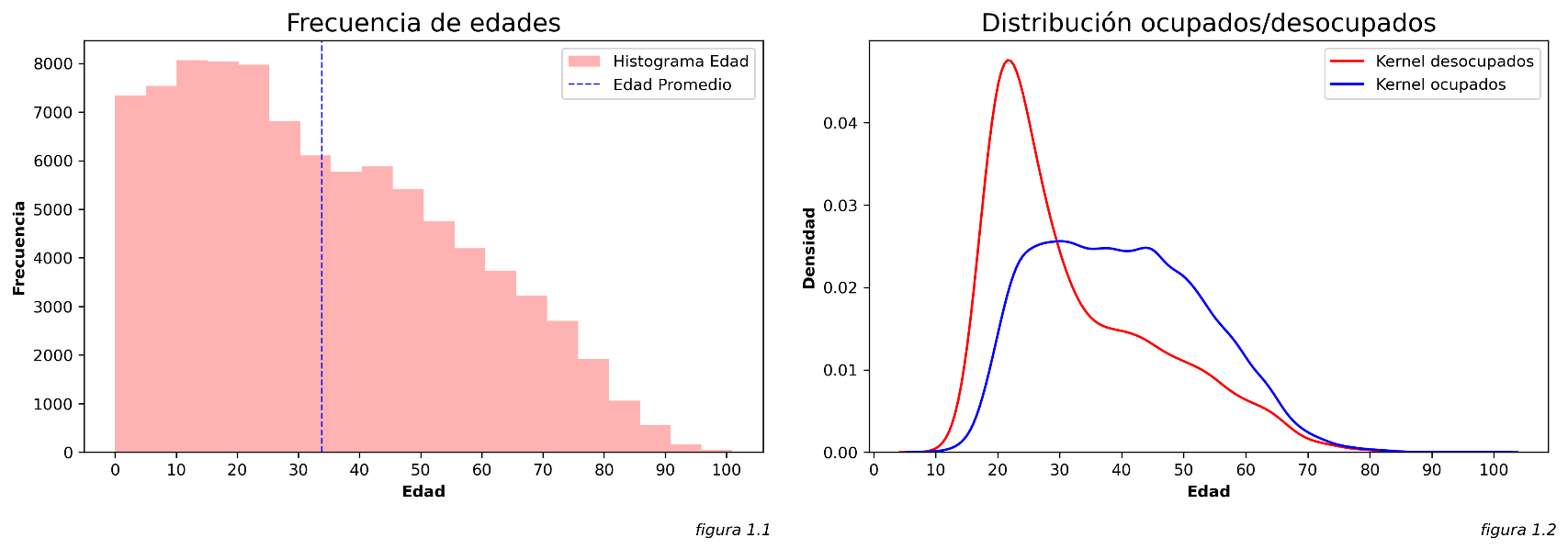
**Parte I**

**Limpieza final**

Un requerimiento previo para concatenar ambas bases de datos es que estas tengan las mismas variables (columnas), la diferencia en ambas bases es simplemente que una está en minúscula (2004) y la otra en mayúscula (2024). Se optó por poner ambas en mayúsculas ya que corresponde al formato de la base más reciente.

La limpieza se dividió en tres partes, la que había quedado del trabajo práctico anterior, una general donde aquellas variables que tienen el mismo código para sus respuestas (Sí = 1, No = 2, Ns/Nr = 9) fueron todas incluidas en una sola limpieza general. Y por último el resto de las variables que comparten ambas bases de datos fueron limpiadas de manera “manual” al igual que se hizo con el trabajo práctico anterior.

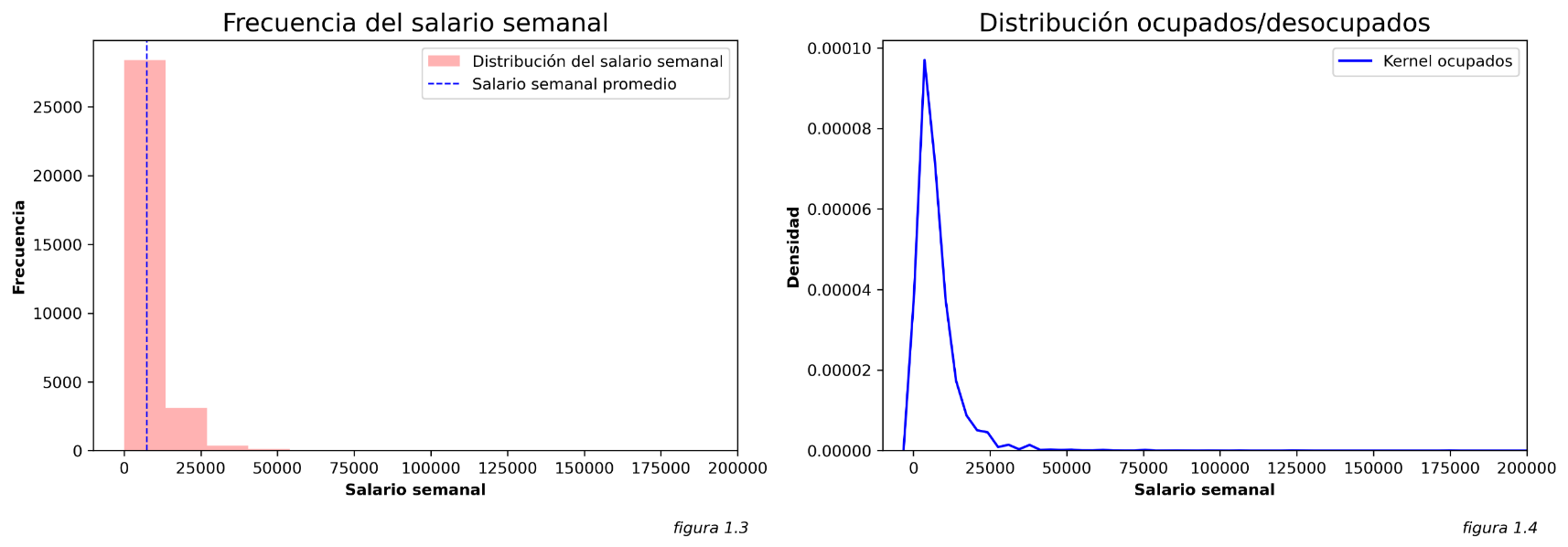
**Punto 1**

Podemos ver que las líneas de ambos kernels juntos (figura 1.2), repiten la distribución de la edad en la muestra (figura 1.2), esto se debe a que la mayoría de los usuarios que reportaron edad, se identificaron como **ocupados/desocupados**. En la frecuencia de edades lo mas notorio es una disminución, prácticamente lineal, en la cantidad de personas una vez pasados los 40 años. Mientras que lo que destaca en la figura 1.2 es, en primer lugar, como el grafico de ocupados tiene sus valores máximos en el intervalo correspondiente a la población en edad laboral. En cambio, los desocupados muestran una clara asimetría hacia la derecha, siendo el pico de densidad en aproximadamente 25 años.

**Punto 2**



El valor que más llama la atención es el promedio, ya que este nos muestra que la media de la población no termina el secundario, tan solo terminando el primario o apenas terminando la EGB (Educación General Básica), que hoy en día no existe.

**Punto 3**

Como podemos ver en ambos gráficos, la media de las observaciones reporta números mas cercanos a 0, como se espera de este tipo de distribuciones. Particularmente llama la atención en la figura 1.4 la **ausencia de un kernel de desocupados**, el mismo se encuentra en el código, pero no en el gráfico ya que el mismo representa gente desocupada, la cual **no tiene un salario**.

**Punto 4**

Breve explicación del armado de la variable:

Hay personas que tienen una ocupación principal pero no otras (PP3F\_TOT= NaN). Para aquellas, la suma entre ambas variables no va a ocurrir, por lo que se procedió a cambiar los NaN de **PP3F\_TOT** por 0s. Así se le suma a la cantidad de horas en la ocupación principal, las horas en ocupaciones extra (aunque este valor sea 0).

Por otra parte también se limitó la suma de ambas variables a 168, que es la cantidad de horas que hay en una semana (aunque no es posible trabajar cada una de esas horas)



El valor **promedio** está muy cerca del valor **mediano**, lo cual nos habla de **simetría** (el valor que promedio coincide con el más frecuente)

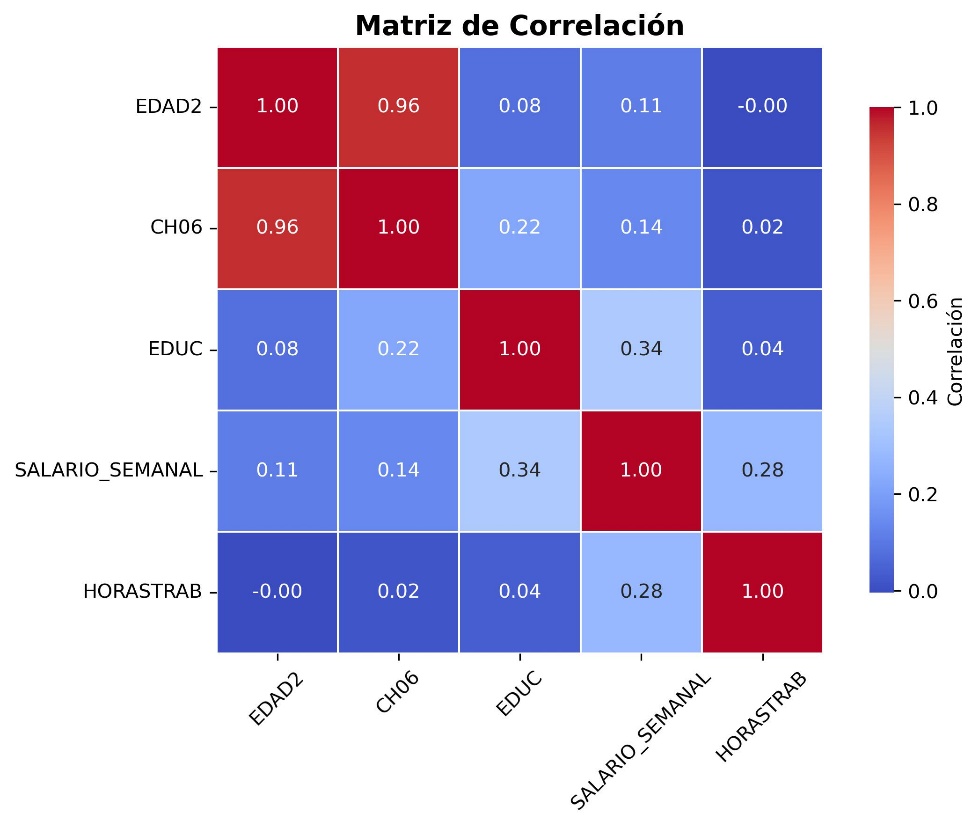
El valor que nos dio en promedio da a interpretar a nivel nacional la persona promedio trabaja 8 horas al día por cinco días a la semana, lo cual sería de esperar

**Punto 5**



El dato más significativo que se deduce de esta tabla es que en la región pampeana **la cantidad de desocupados bajó a la mitad**, y que **la cantidad de** **ocupados subió en hasta un 15%.**

A pesar de que las variables en común entre ambas bases de datos sean 172, el número final queda en 176 por las cuatro nuevas variables.

**PARTE 2  
Punto 1**

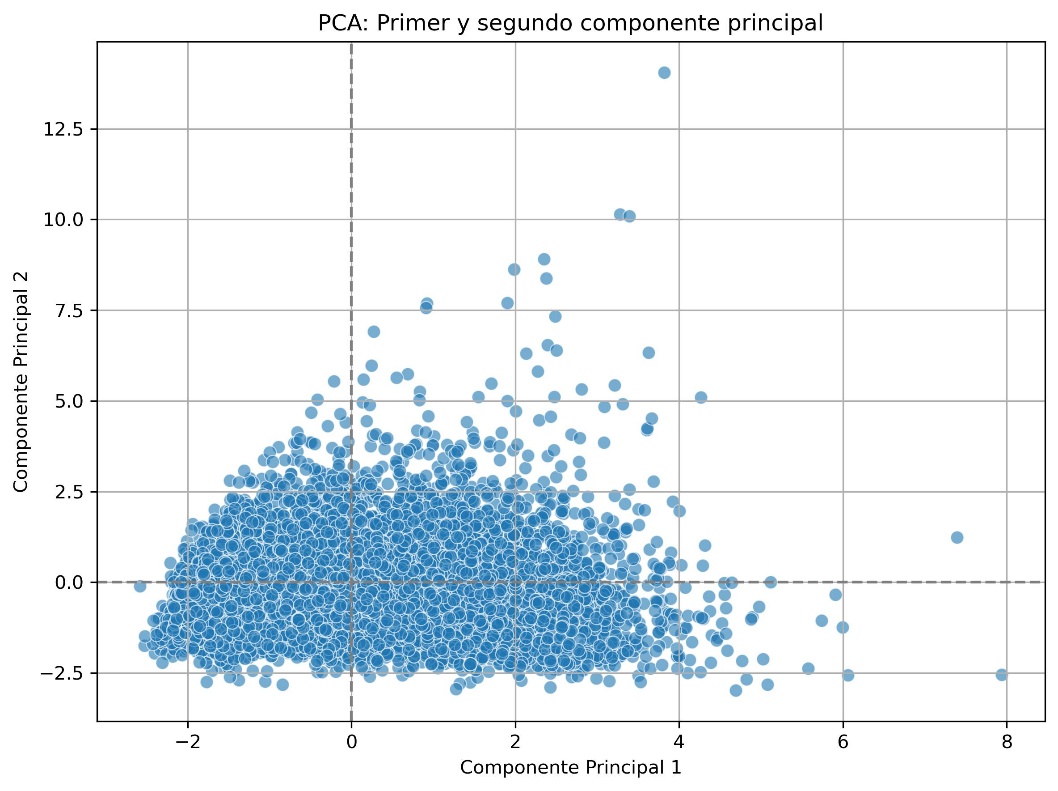
*figura 2.1*

En la matriz de correlación se observa una **alta correlación positiva entre EDAD2 y CH06** (0.96), lo cual es esperable si ambas variables capturan características relacionadas con la edad o la estructura del hogar.

La variable **EDUC** muestra una correlación moderada con el **salario semanal** (0.34), lo que sugiere que un mayor nivel educativo está parcialmente asociado con mayores ingresos.

Las variables **HORASTRAB** y **SALARIO\_SEMANAL** presentan una correlación baja (0.28), lo cual podría indicar que no todos los trabajadores son remunerados proporcionalmente a sus horas trabajadas, posiblemente por diferencias en tipo de empleo o informalidad.

En general, las correlaciones no son muy altas con excepción del par EDAD2–CH06, lo que sugiere que cada variable aporta información distinta y útil para el análisis mediante componentes principales.

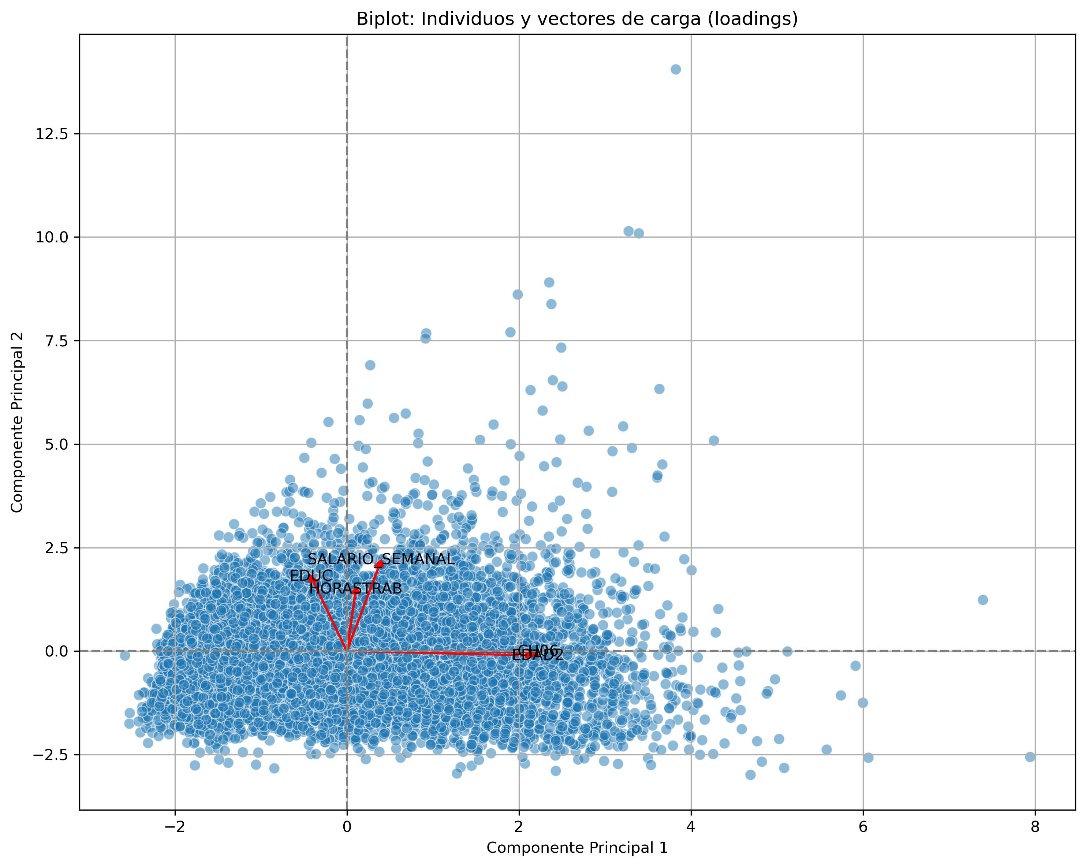
**Puntos 2 y 3**

*figura 2.2*

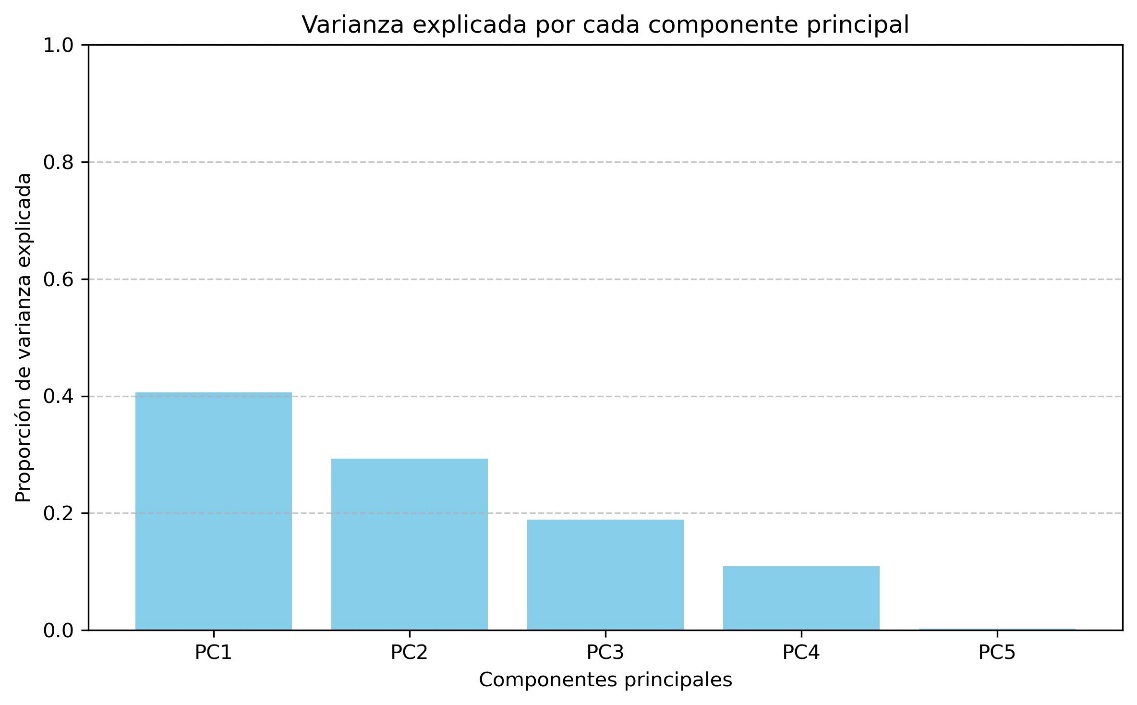
El gráfico de dispersión muestra cómo se distribuyen los individuos en el espacio definido por el **primer** y **segundo componente principal** del PCA.

Se observa una **concentración importante de puntos cercanos al eje vertical (Componente Principal 1 ≈ 0)** y una **asimetría hacia la derecha**, lo que indica que el primer componente capta una mayor variabilidad que el segundo.

Este tipo de distribución sugiere que hay **una dimensión dominante** que explica la mayor parte de las diferencias entre individuos (CP1), mientras que el segundo componente (CP2) representa una fuente de variabilidad secundaria.

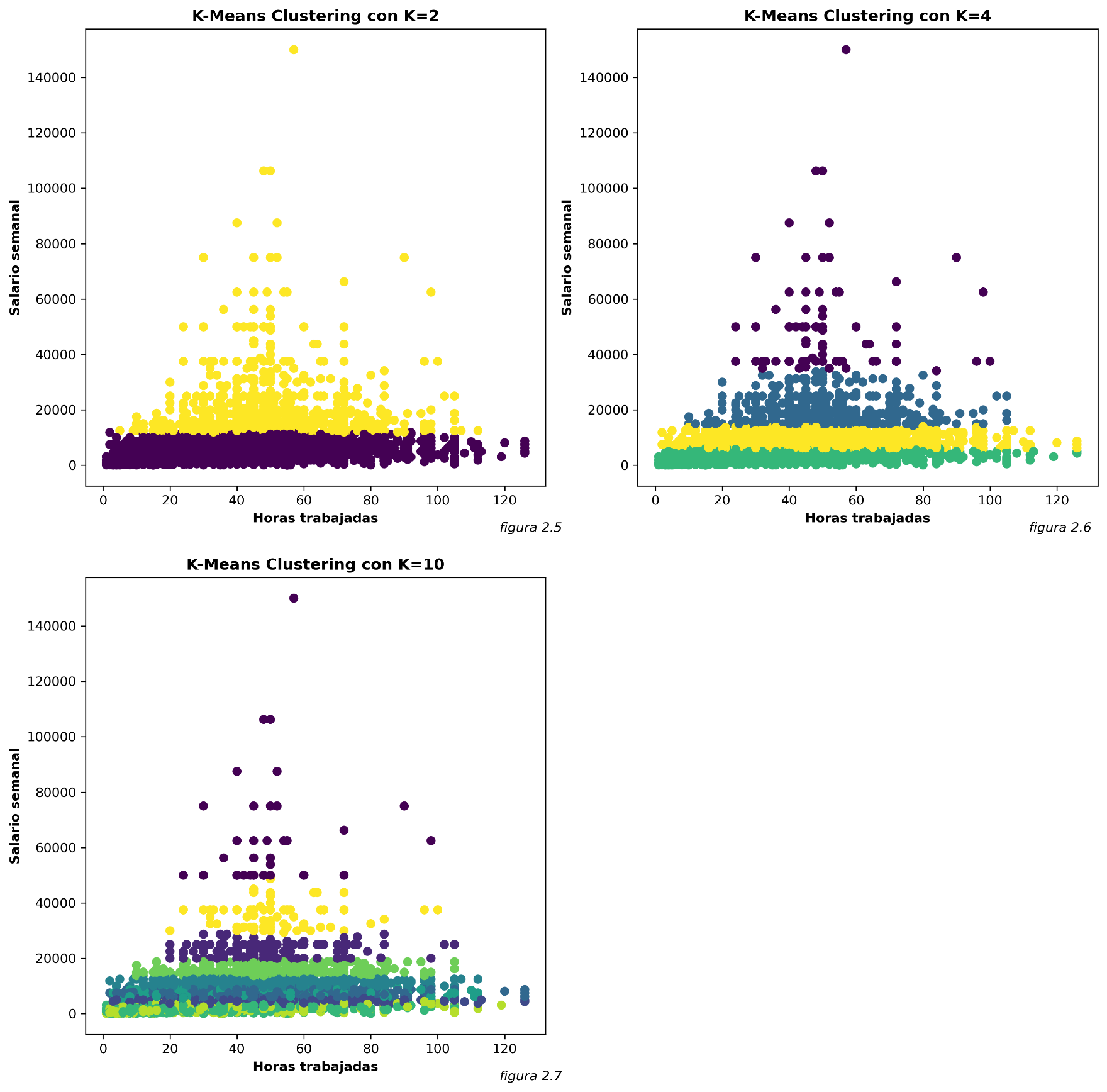
 *figura 2.3*

**Punto 4**

*figura 2.4*

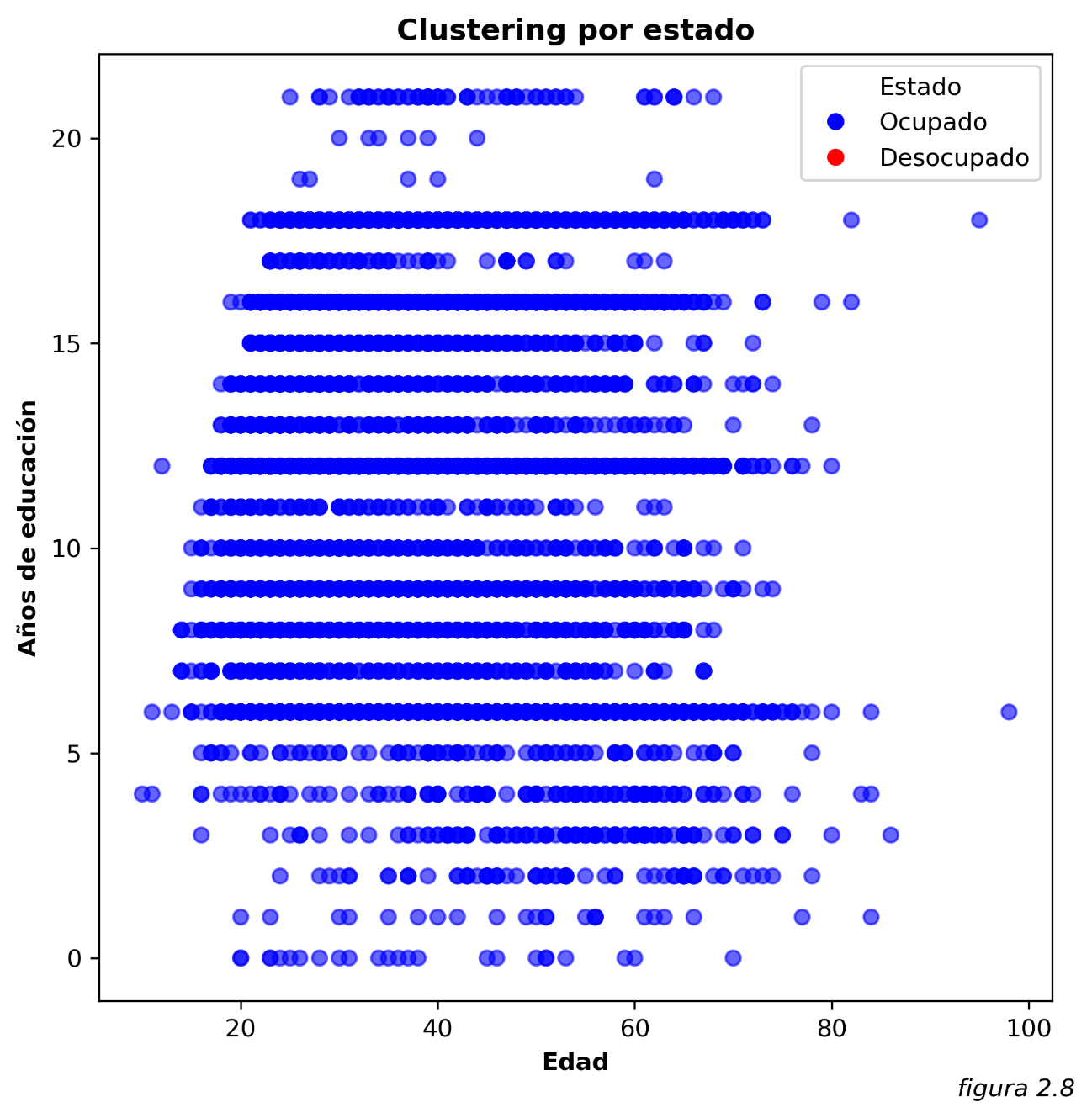
En conjunto, **los dos primeros componentes explican aproximadamente un 70% de la variabilidad** en los datos originales, lo que justifica su uso para representar visualmente las observaciones sin perder demasiada información.

Este resultado indica que el PCA ha sido efectivo para reducir la dimensionalidad: **con solo dos componentes se preserva buena parte de la estructura de los datos**, lo que simplifica el análisis y permite una interpretación más clara de las relaciones entre las variables.

**Punto 5 a**

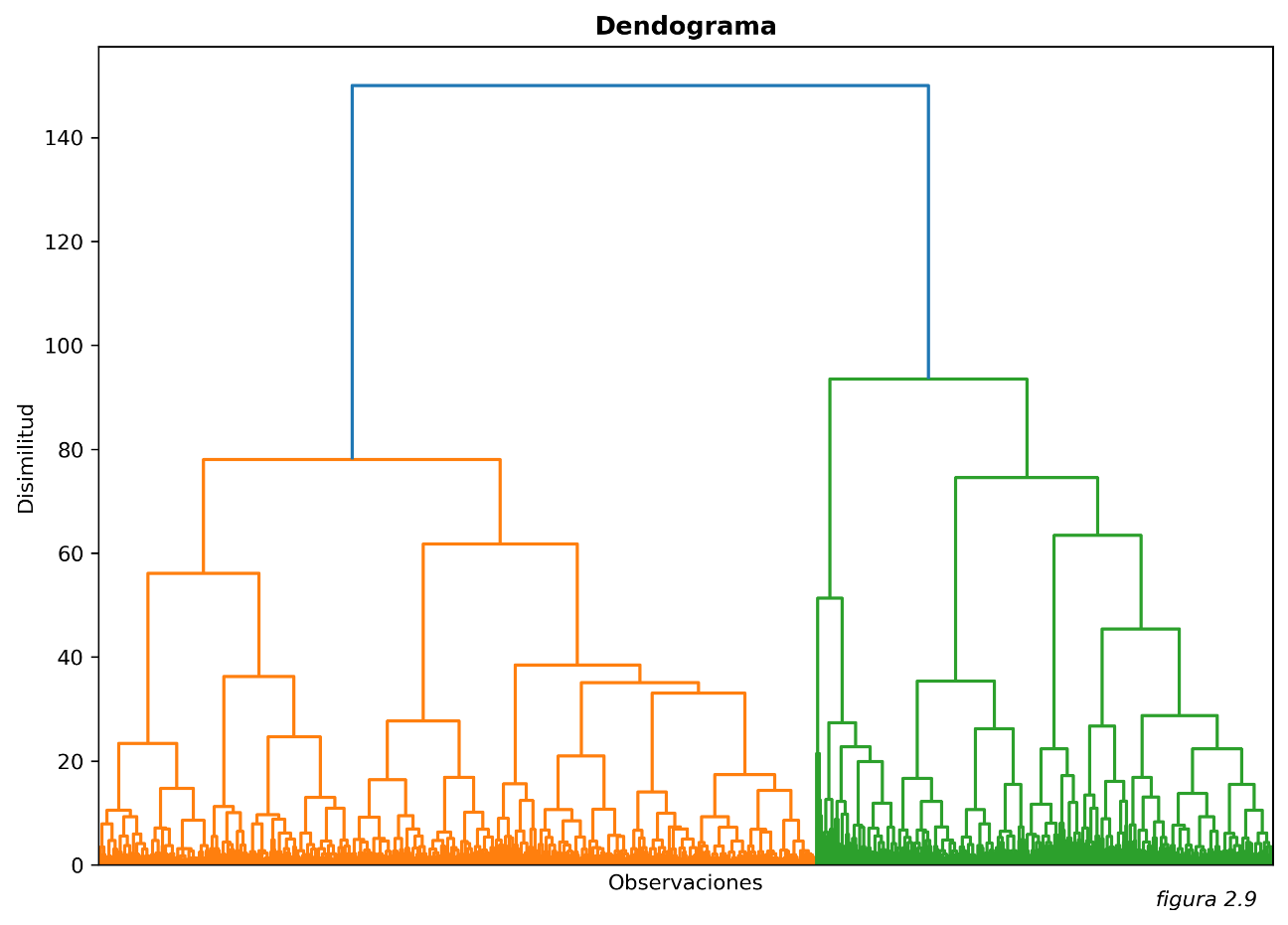
Se eligieron las variables **HORASTRAB** y **SALARIO\_SEMANAL** para no repetir las variables que se usarán en el punto siguiente.

Se puede ver como el algoritmo **agrupa a la población según su ingreso semanal**. A medida que aumentan los K-medias, los grupos se vuelven más homogéneos, al punto tal de que **en la** **figura 2.7 no se aprecian grupos** que podamos asociar a diferencias en el salario, como si ocurre en las figuras 2.5 y 2.6

**Punto 5**

Cuando se le pide al algoritmo que agrupe por valores que toma la variables estado **no logra efectuarlo**, los datos salen todos como “ocupados”

**Punto 6**

Se trata de un diagrama en forma de árbol que muestra las relaciones jerárquicas entre un grupo de objetos o elementos, representando como se agrupan o dividen en función de su similitud o distancia. Allí agrupamos ‘las hojas’ de abajo hacia arriba.