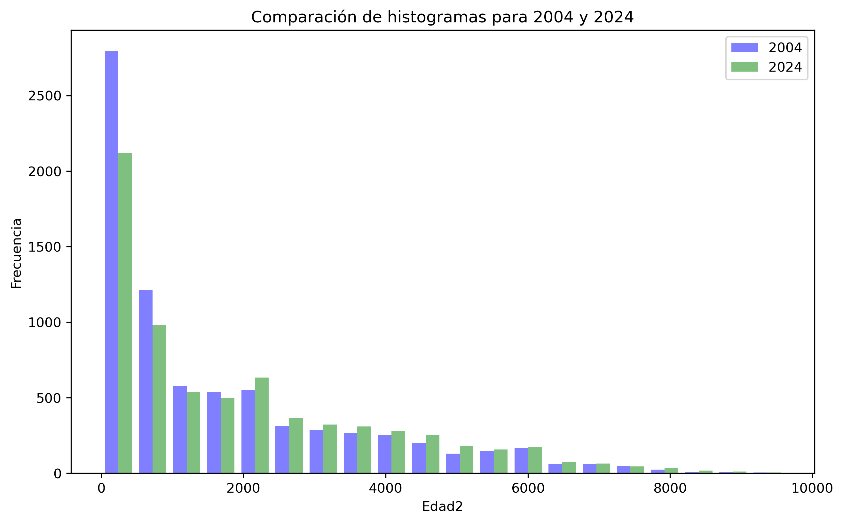
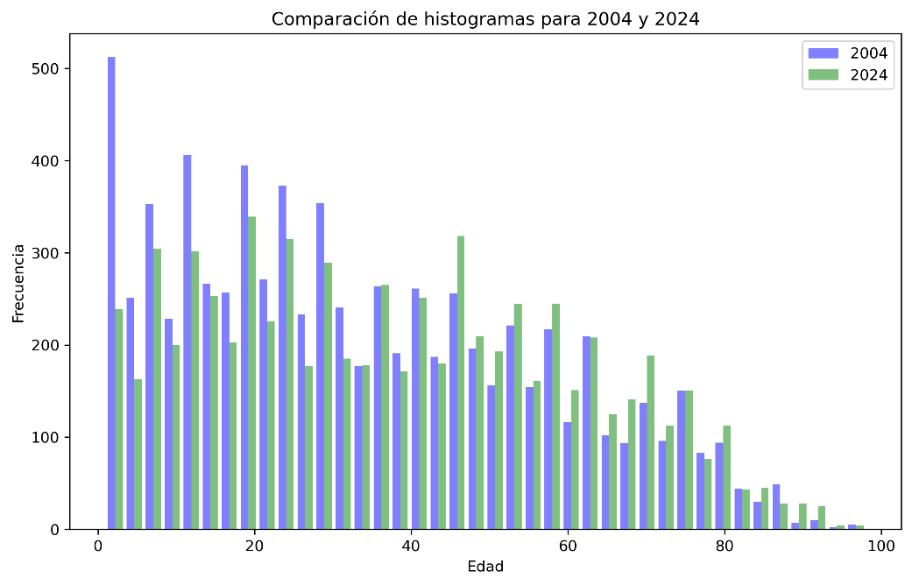
TP3 Big data and Machine Learning  
Franco Bustelo  
Juan Pablo Tomasello  
David Agudelo.

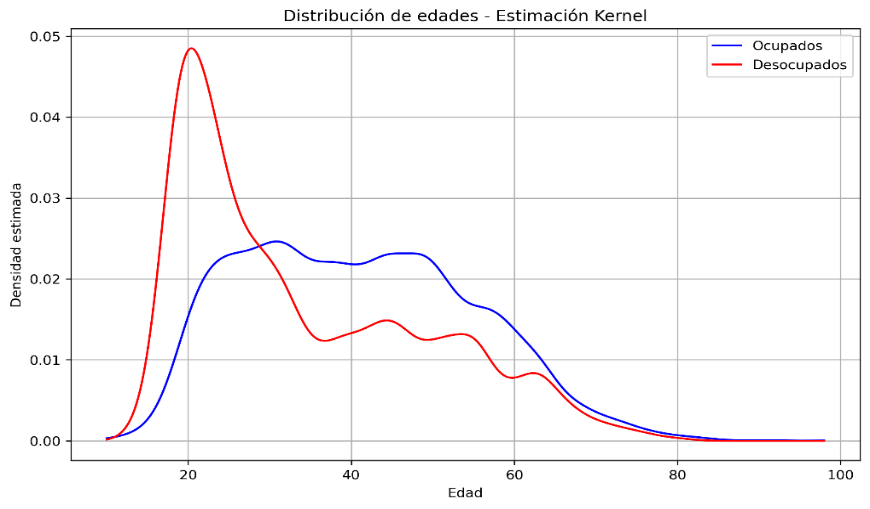
**PARTE 1**

**Punto 1**

El gráfico muestra un histograma comparativo de la variable **Edad2** (edad elevada al cuadrado) para los años 2004 (en azul) y 2024 (en verde).

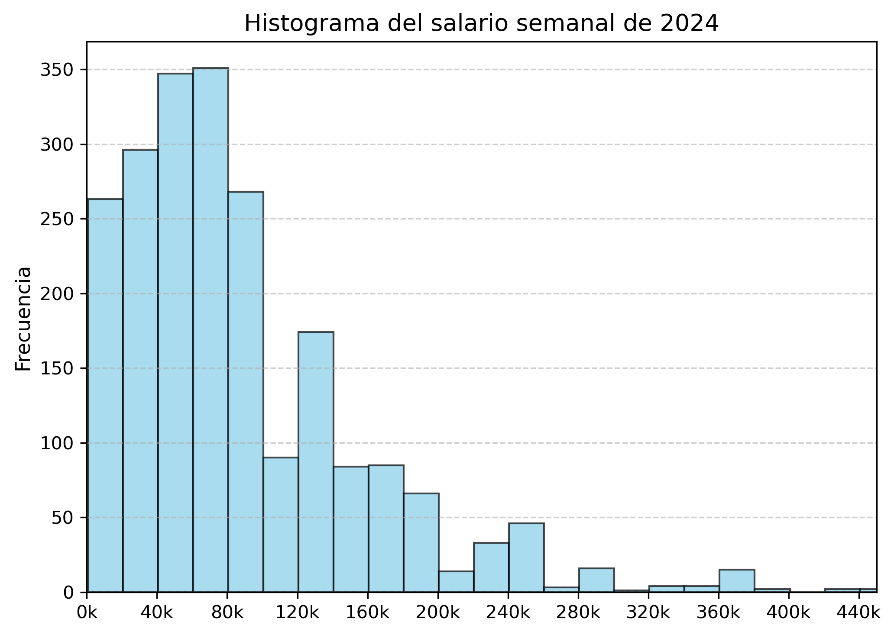
El gráfico compara la distribución de la **edad al cuadrado** entre 2004 y 2024. En ambos años, la mayor frecuencia se da en valores bajos, lo que indica predominancia de personas jóvenes. Sin embargo, en 2024 hay un leve aumento en los valores más altos, lo que sugiere un envejecimiento de la población en ese período.

  
El gráfico muestra la distribución de la **edad** entre 2004 y 2024. En 2004 predominan los grupos más jóvenes, especialmente menores de 30 años. En 2024 se observa una mayor frecuencia en edades medias y altas (de 40 años aproximadamente), lo que indica un claro **envejecimiento de la población** en este periodo. No tiene sentido analizarlas al cuadrado, así se entiende mucho mejor.

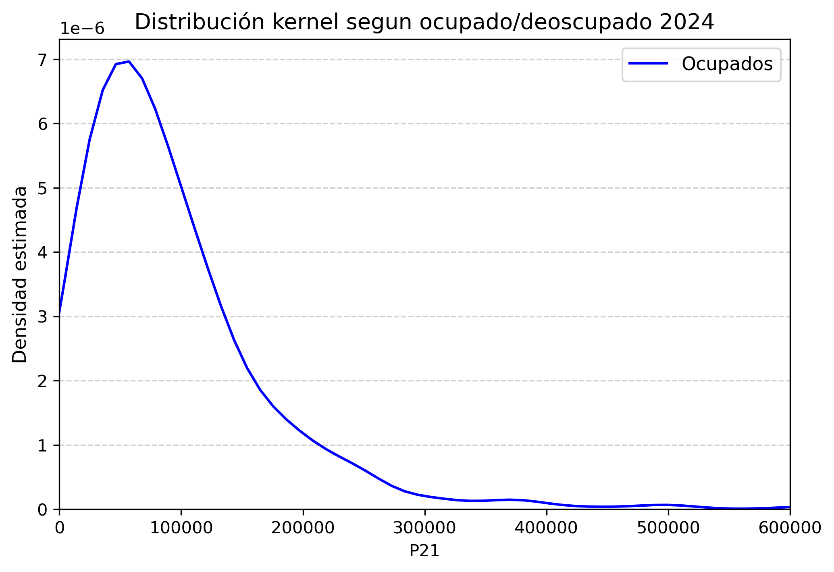
****

La mayor densidad de desocupados se concentra en edades **jóvenes de alrededor de 20 años,** mientras que los ocupados presentan una distribución **que abarca más edades**, con mayor densidad entre los **22 y 60 años aproximadamente**. Esto podría tratarse de un indicador de que el desempleo afecta principalmente a personas jóvenes.

**Punto 3**



**Histograma del salario semanal**   
Muestra que la mayoría de los ocupados (donde más frecuencia hay) ganan entre 20000 y 100000 pesos semanales. Teniendo valores más bajos en cuanto a frecuencia en salarios por encima de los 200000.



**Distribución kernel según ocupado o desocupado**   
Se ve una gran concentración hasta los salarios de 200000 y muestra, con una larga cola a la derecha los valores no tan recurrentes. Este histograma nos está indicando que hay pocos casos con ingresos altos.

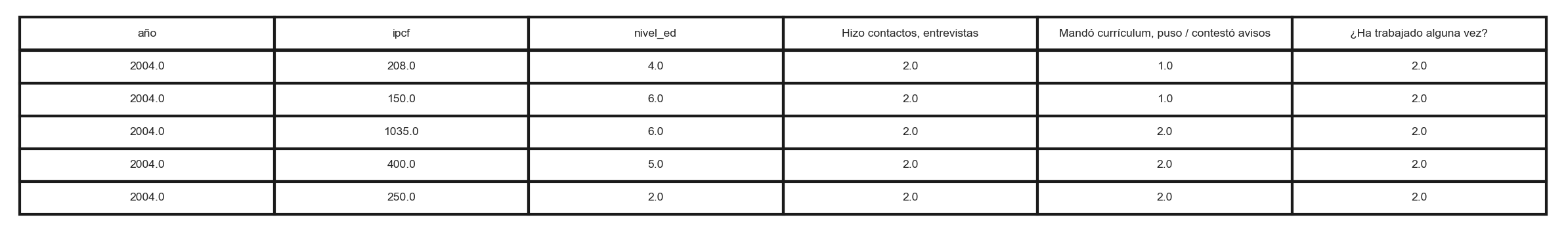
Gráfico, Histograma

El contenido generado por IA puede ser incorrecto. En la consigna se pide hacerlo con desocupados también, pero, como mostramos en los códigos, no hay ningún desocupado que tenga algún salario ya que todos son 0.

En el primer panel se ve que el salario semanal de 2004 expresado en pesos de 2024 (multiplicado por 120, que es la inflación acumulada de 2004 a 2024 basándose en el IPC) tiene su mayor concentración de frecuencia entre los 5000 y los 40000 pesos semanales. Contando con una mucha menor frecuencia por encima de dichos 40000 pesos.   
 En el otro panel se puede ver la distribución de kernel de dichos salarios semanales pero esta vez teniendo en cuenta el estado, es decir, si están ocupados o desocupados. Se puede ver como los ocupados rondan dentro del rango entre los mismos 5000 y los 40000 pesos una vez más. Esto quiere decir que la mayoría de ocupados tienen su salario semanal rondado por dichos valores. En la consigna también se pide tener en cuenta a los desocupados pero estos, como se explicó y demostró en los códigos no presentan ningún salario. Todo aquel que sea desocupado tiene un p21 de valor 0, lo cual tendría sentido ya que un desempleado no debería porque tener un salario semanal.  
 Aclaración: Dividimos por 4, no por 40, ya que no tenía sentido dividir por 40 si queríamos hacer semanal un ingreso mensual.

**Punto 5**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2004 | 2024 | Total |
| Cant. de observaciones | 7647 | 7051 | 14698 |
| Can. De observaciones con nas en la variable “Estado” | 10 | 41 | 51 |
| Cantidad de ocupados | 3079 | 3224 | 6303 |
| Cantidad de desocupados | 528 | 311 | 839 |
| Cantidad de variables limpias y homogeneizadas | 311 | 232 | 543 |

 Para la cantidad de variables limpias y homogeneizadas decidimos escoger una base especial para desocupados, la cual muestra 6 variables, entre ellas el salario y el nivel educativo. Nos pareció interesante homogeneizar solo estos datos para desocupados ya que nos parecía más relevante y preciso a la hora de analizar las causas por las que una persona puede estar en dicha situación. Dejamos de referencia los primeros 5 elementos de la tabla que usamos.

Para su mejor entendimiento recordamos el significado de los valores de la columna “nivel\_ed”:

1 = Primario incompleto (incluye educación especial)

2 = Primario completo

3 = Secundario incompleto

4 = Secundario completo

5 = Superior universitario incompleto

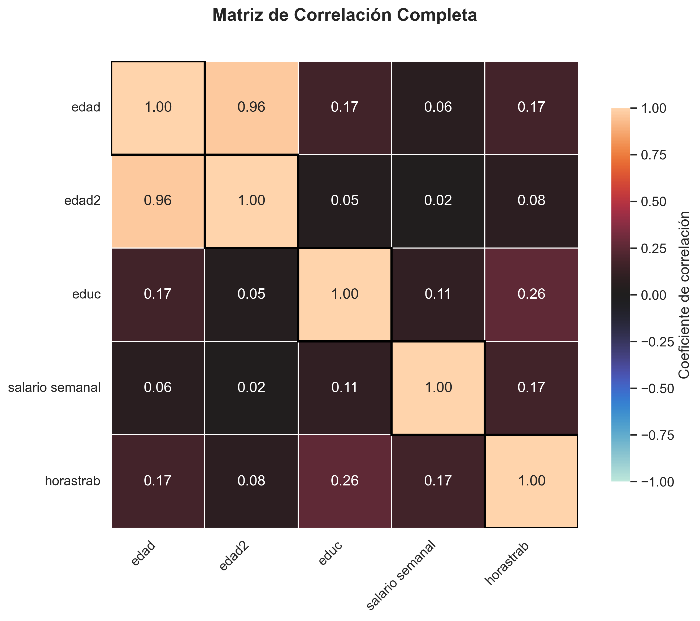
6 = Superior universitario completo

7 = Sin instrucción

9 = Ns/Nr

**PARTE 2**

**Punto 1**



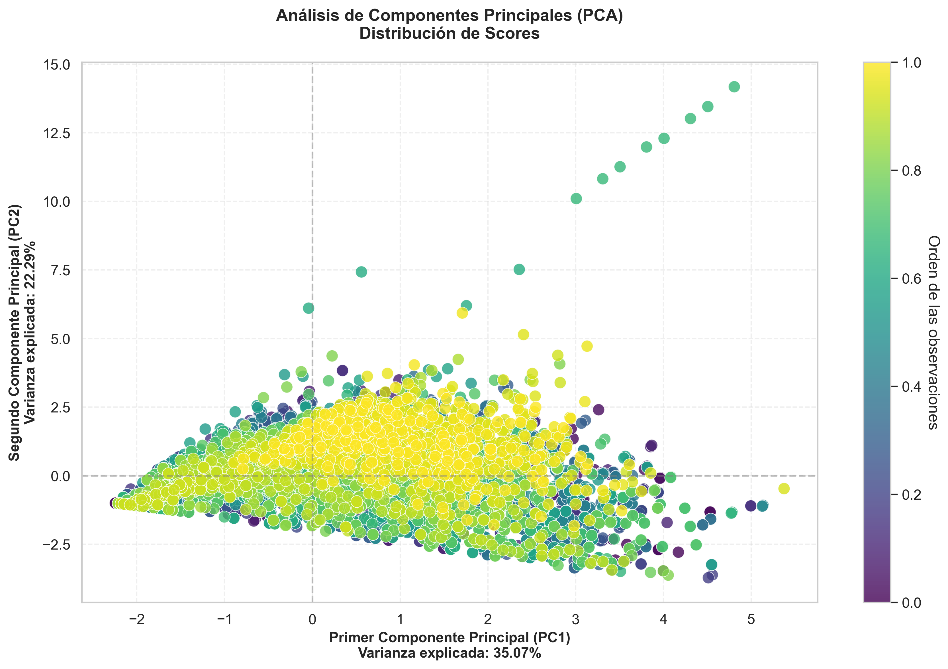
La matriz correlación nos muestra, obviando la correlación entre ‘edad’ y ‘edad2’ ( de **0.96** que es bastante obvia ya que son lo mismo pero elevada al cuadrado), que la más interesante es la que se observa entre ‘educ’ y ‘horastrab’, con un valor de **0.26**. Esto quiere decir que las personas con mayor nivel educativo tiene una muy chica correlación positiva.

Por otro lado, la correlación entre ‘educ’ y ‘salario\_semanal’ es más baja (de un **0.11**), lo cual indica que al tener un mayor nivel educativo no necesariamente se traduce en un salario más alto.

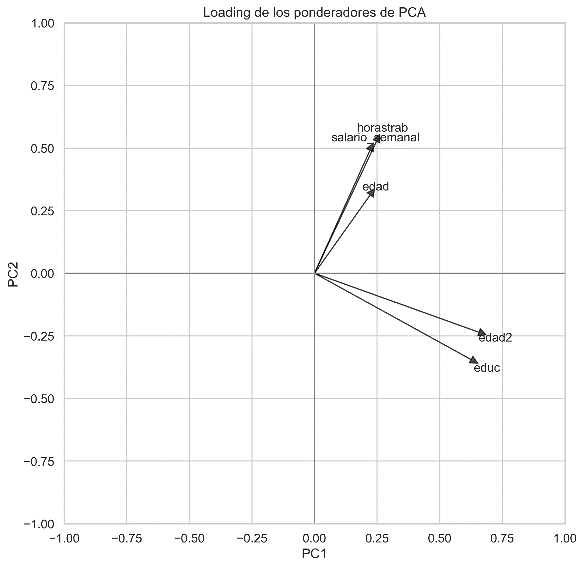
Otra correlación que se tiene en cuenta es la de ‘salario\_semanal’ con ‘horastrab’ la cual es de **0.17**. Esto quiere decir que hay una relación muy baja entre la cantidad de horas que una persona trabaja y lo que gana por semana. Es decir, trabajar más horas no necesariamente significa tener un salario más alto.

Finalmente, se observa que ni ‘edad’ ni ‘edad2’ tienen una relación significativa con el salario semanal (entre **0.06** y **0.02** respectivamente) ni con las horas trabajadas (en torno a **0.17 y 0.08**). Es decir, la edad no parece ser un factor determinante ni en el ingreso semanal ni en la carga horaria laboral de las personas encuestadas

**Punto 2**

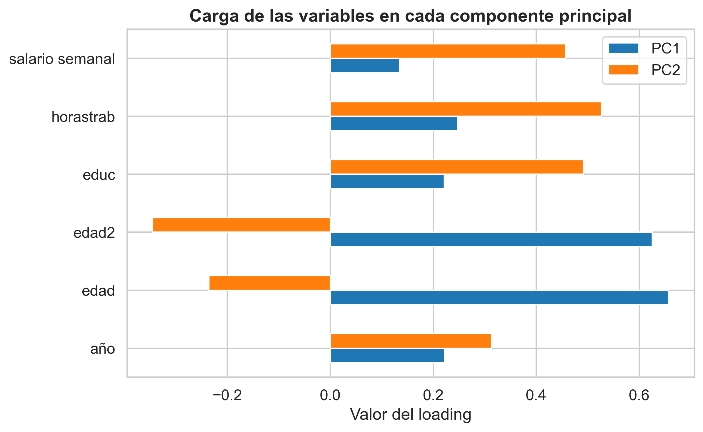
Primero podemos ver que está ordenado por colores según el orden en la tabla, las más amarillas pertenecen a 2024 y las más violetas a 2004. Podemos notar que las variables que se encuentran en los valores más altos son aquellas que poseen valores demasiado altos en todas las variables, por ejemplo: salarios demasiado grandes, muchas horas de trabajo, muy viejos, etc. Como veremos luego, en el PC2 lo que más peso tiene es el salario y las horas trabajadas así que por eso se pueden encontrar tan arriba algunos puntos. Mientras que los puntos más a la derecha se pueden deber a edades muy grandes (ya que es el factor con más peso en PC1). Todo esto se verá más claro en el punto 3. (Que sea el que más peso tiene no quiere decir que sea el único factor importante, por lo que puede ocurrir que su valor no se deba solamente a su edad, aunque si en alguna gran parte.)

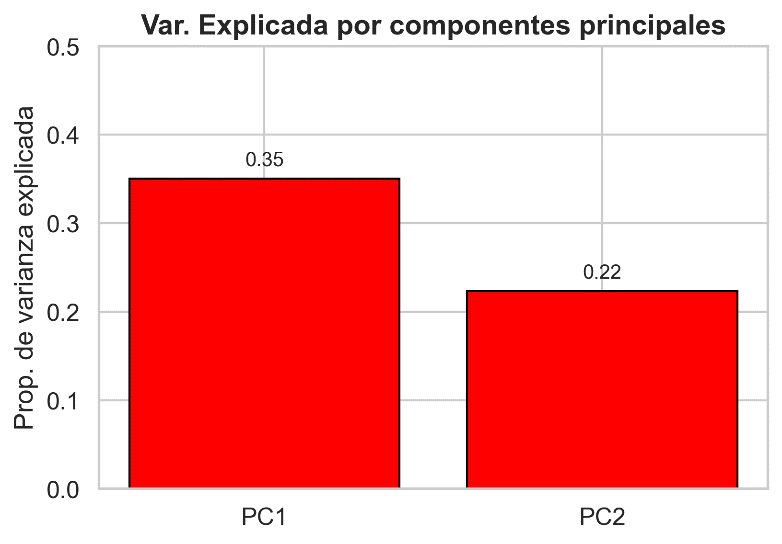
**Punto 3**

**‘Educ’** y ‘**edad2’** tienen una mayor carga sobre el eje **PC1**, o sea que influyen más en ese componente. Como ambas tienen que ver con el nivel educativo y la edad, podríamos decir que hace un “índice” creado por ellas 2.

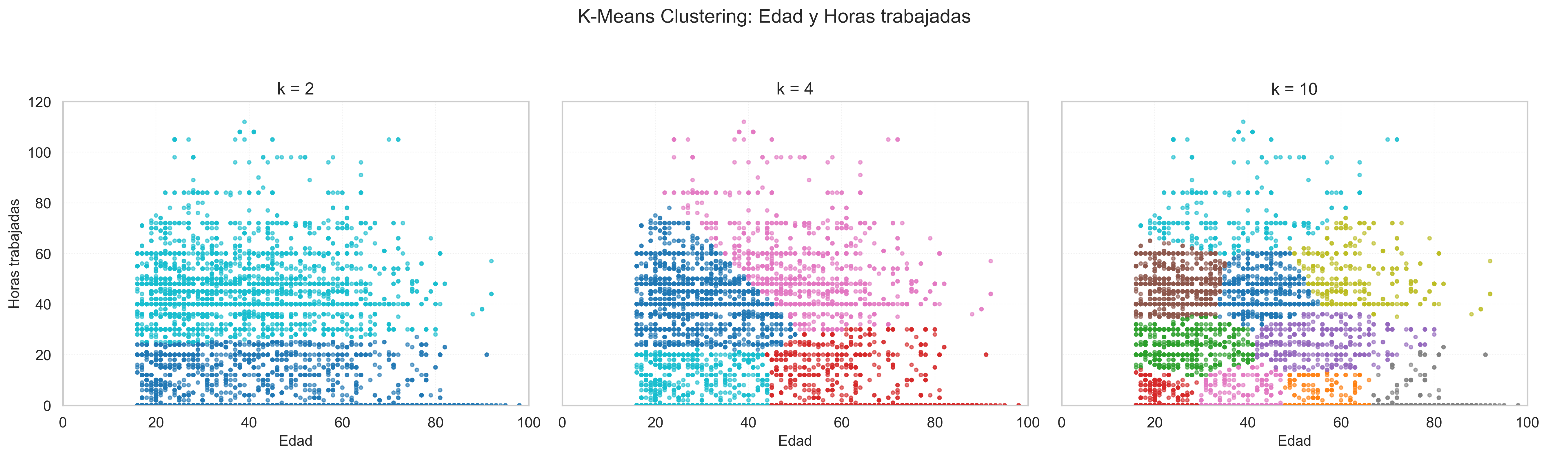
Por otro lado, **horastrab**, **salario\_semanal** y **edad** apuntan más hacia arriba (positivo para el PC2), lo que indica que tienen más peso en ese eje. Como todas están relacionadas al trabajo, es posible que **PC2** represente algo así como la responsabilidad con el trabajo.

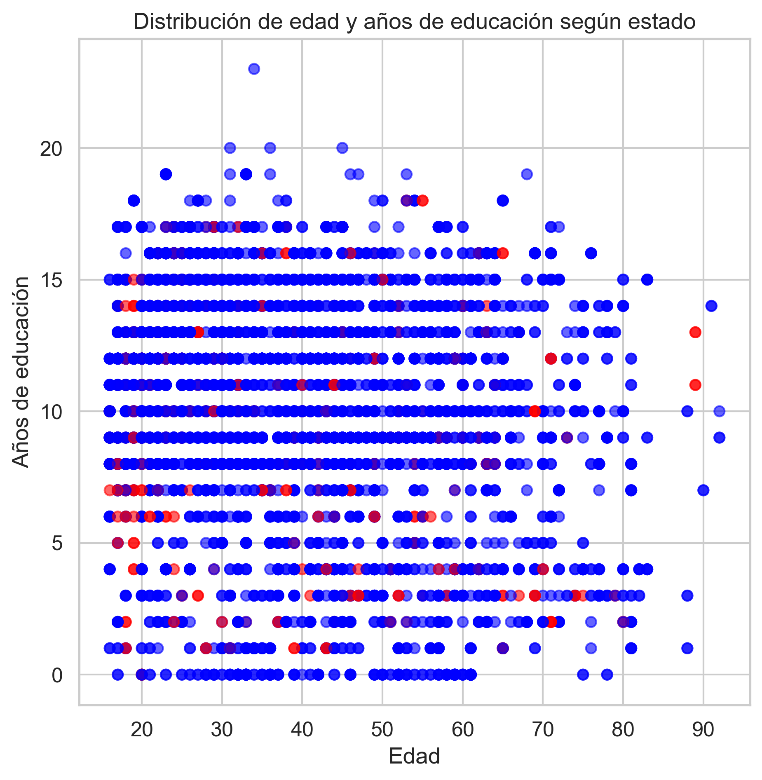
También es interesante ver que las flechas no están todas en la misma dirección. Eso nos dice que las variables están captando cosas diferentes, lo cual puede ser bueno para el análisis porque ayuda a que los componentes principales separen bien la información.

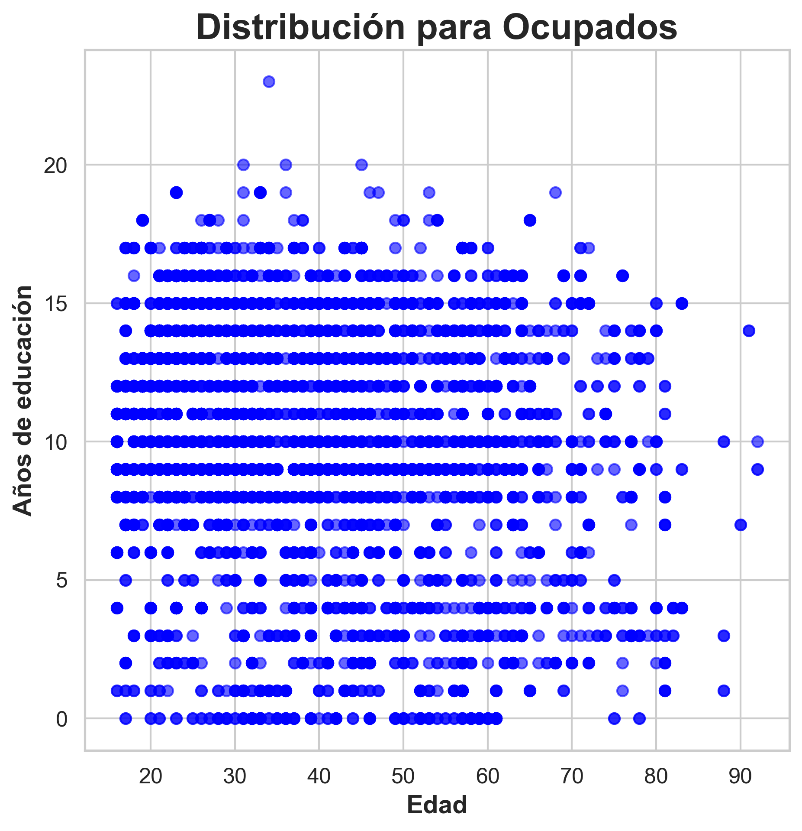
En este gráfico se ve que el primer componente principal (PC1) está fuertemente influenciada por la edad, mientras que la segunda (PC2) representa más el aspecto laboral, con carga alta en salario semanal, educación y horas trabajadas. Esto muestra que las variables se agrupan en dos ejes principales: uno más personal y otro más laboral.

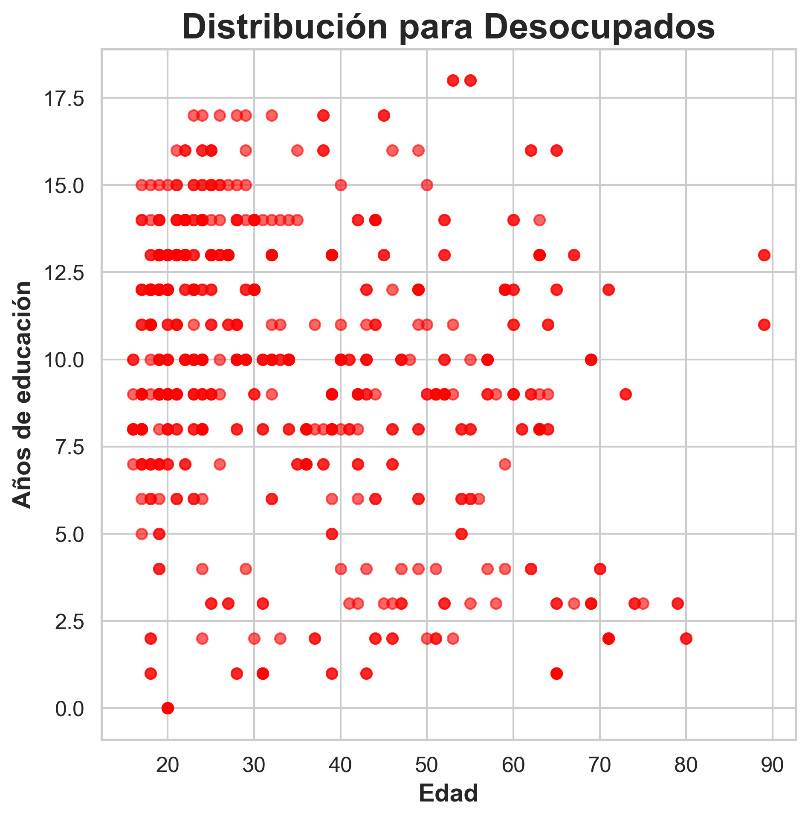
**Punto 4**  
  
  
  
PC1 explica el 35% de la varianza total del conjunto de datos. PC2 explica otro 22%.

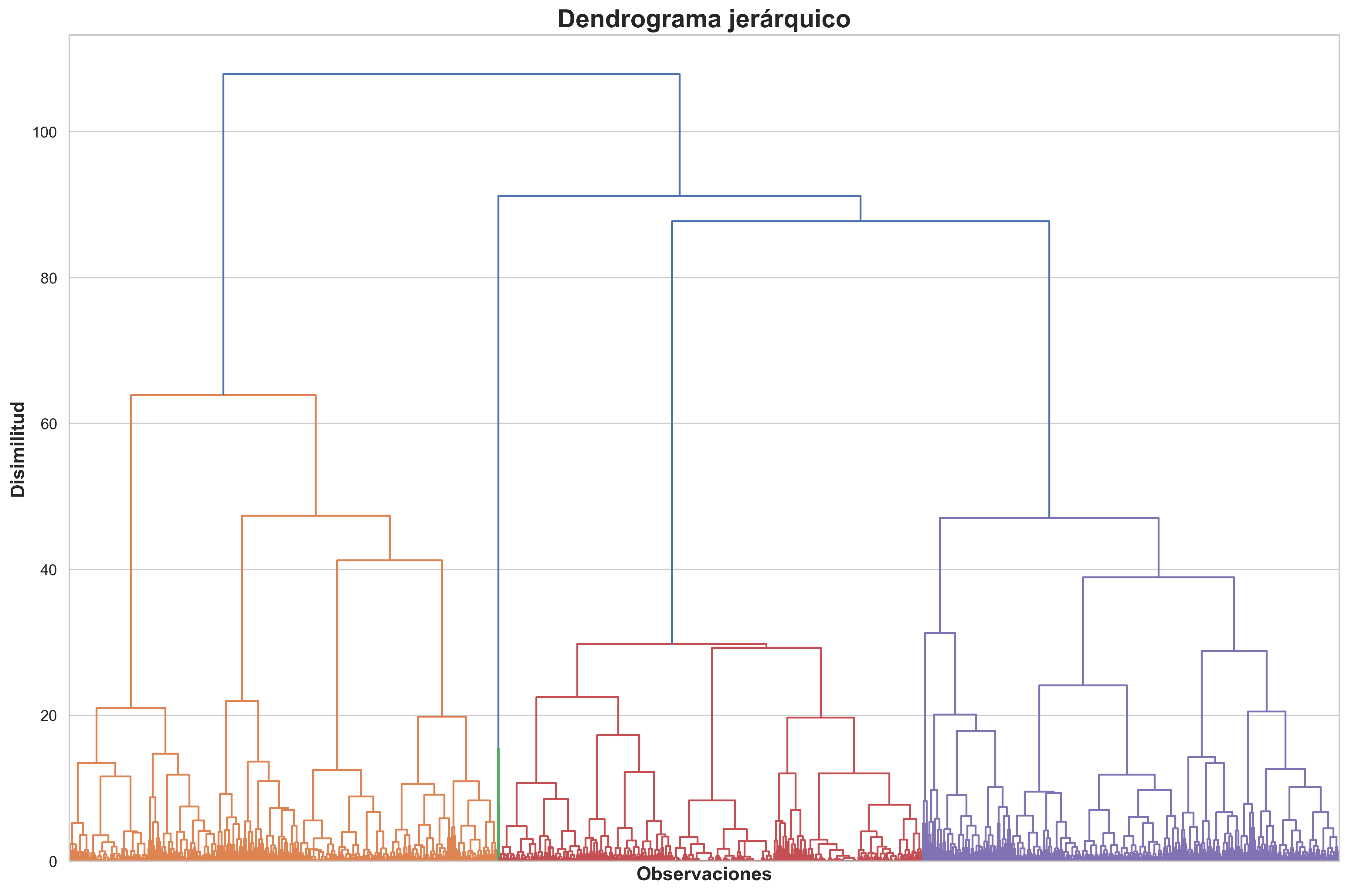
Juntas, explican un 57% de la varianza total, lo que quiere decir que más de la mitad de la información original ya está representada en esas dos componentes.

**Punto 5A**

Esta imagen muestra cómo cambia la agrupación de personas según su edad y horas trabajadas al aplicar k-means con distintos valores de **k** (número de grupos que tomamos). Con **k=2** no se distinguen bien las diferencias de horas trabajadas y los grupos de edad. A medida que se aumenta **k a 4 y luego a 10**, los grupos capturan más detalles, separando mejor a las personas según combinaciones específicas de edad y horas trabajadas.  
  
**Punto 5B**  
  
  
Este gráfico, de esta distribución, nos muestra como los ocupados son aquellos que tienen los mayores años de educación y los que a mayor edad, todavía con sus años de educación, se mantienen ocupados. Se puede ver en el grafico como los mayores valores de años de educación forman parte de los ocupados.

Mientras que los desocupados están más concentrados en bajos años de educación y en bajas edades, aunque esto no quiere decir que no haya desocupados con muchos años de educación o con edades más grandes.  
  
  
  
  
  
  
  
Acá podemos ver la gran masa de ocupados que se encuentra entre los 20 y los 60 años de edad, y entre los 5 y los 15 años de educación, en su mayoría.

 Aca podemos ver la distribución por separado con más precisión, podemos notar que en edades más bajas se concentran los desocupados e incluso la gran parte se concentran en años de educación más altos.

**Punto 6**

Un dendograma es una herramienta que utiliza el cluster jerárquico, no se necesita un número de clusters. Sino que mediante la similitud de cada observación se van juntando clusters, y luego mide la distancia entre clusters utilizando linkage, para esto existen distintos métodos: average linkage, simple linkage o complete linkage.

El complete linkage utiliza la distancia máxima entre los dos elementos más lejanos de un cluster A y un cluster B.  
El simple linkage usa la distancia mínima entre dos elementos más cercanos de un cluster A y un cluster B.  
El average linkage usa la distancia promedio (como era de esperarse) entre los elementos de un cluster A y un cluster B.  
Pero en este caso usamos ward linkage (no lo conocíamos), que es bastante útil para disminuir el aumento de la varianza al unir clusters. Busca unir clusters solo si estos aumentan lo menos posible la varianza adentro del cluster.   
Por último, algo importante es elegir el nivel de disimilitud que queremos, lo cual corta el gráfico con una línea horizontal. Las líneas que alcance serán la cantidad de clusters que se consiguen con dicho nivel de disimilitud. Por ejemplo, en este caso, si corto en 80, tendré 5 clusters.